TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Estudi del rendiment i evolució dels estudiants del grau de Matemàtiques (UPC) des d'un punt de vista del gènere

Autor: Judith Pascual Terrón

Director: Lourdes Rodero De Lamo

Departament: Estadística i Investigació Operativa

Convocatòria: Juny 2020

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Resum

En aquest treball s'analitzaran diverses variables referents als estudiants del grau de Matemàtiques de la Universitat Politècnica de Catalunya, incidint de forma rellevant en l'efecte del sexe, ja que actualment existeix una preocupació general pel que fa a la perspectiva de gènere en aquesta carrera. Concretament, s'estudiarà de què depèn el valor de la nota final dels alumnes, la probabilitat d'aquests d'aprovar la fase inicial del grau i com ha estat la seva evolució al llarg dels anys respecte a diverses dades, com per exemple el nombre d'estudiants de cada sexe o el nombre d'aquests que aconsegueix superar la fase inicial. A més a més, com ja s'ha dit, s'incidirà de forma rellevant en les diferències que puguin existir entre sexes, és a dir, es pretén comprovar si realment existeixen discrepàncies entre els nois i les noies en aquests estudis. Per a fer-ho s'aplicaran entre d'altres mètodes, models lineals, models lineals generalitzats, tests estadístics per a la comparació de proporcions i anàlisi de sèries temporals, mitjançant els programes *R* i *Excel*.

Paraules clau

Diferències de gènere en el grau de Matemàtiques, Nota final, Probabilitat d'aprovar la fase inicial, Comparació de proporcions, Evolució dels estudiants.

Classificació AMS:

62-07	Data Analysis
62J05	Linear regression
62J12	Generalized linear models

Title

Study of the performance and evolution of the students of the Mathematics degree (UPC) from the point of view of gender.

Summary

In this paper I will analyse several variables belonging to students of the Mathematics degree at the Technical University of Catalonia. For this purpose, data provided by the same institution are available. On the one hand, I will study which of the variables have an influence on the final mark of the students and on the probability that they will pass the initial phase of the degree, and on the other hand, I will also analyse how the students have evolved over the years with respect to various data, such as the number of students of each sex or the number of students who manage to pass the initial phase. In addition, the differences that may exist between the two sexes will be highlighted, in other words, the aim is to check whether there are really significant discrepancies from a statistical point of view between boys and girls in these studies and, more specifically, which areas of these studies are affected. In order to do so, among other methods, linear models, generalized linear models, statistical tests for the comparison of proportions and time series analysis will be applied with the help of the *R* and *Excel* programs.

Keywords

Differences between men and women, Final mark, Probability of passing the initial phase, Comparison of proportions, Evolution of students.

Índex

1. Introducció

En els últims anys la presència de les dones a les universitats ha augmentat considerablement. Les últimes dades de l'Institut Nacional d'Estadística (Curs 2016) mostren que aproximadament un 53.3% dels alumnes que es van graduar a les universitats espanyoles són noies, per tant, pel que fa a aquest aspecte es pot dir que actualment no hi ha desigualtat en el nombre d'estudiants en funció del sexe.

En canvi, sí es troben diferències destacables pel que fa als graus que els alumnes cursen. Avui en dia les dones encara segueixen preferint carreres relacionades amb l'àmbit de l'educació, la sanitat o les ciències socials, per altra banda, els nois es decanten més per estudis tecnològics o enginyeries. En l'any 2016 el percentatge de dones graduades en Educació, Arts, Humanitats, Llengües, Ciències Socials, Administració, Dret, Salut i Serveis Socials va representar un 42.9% del total de graduats d'aquell any. Pel que fa a les carreres de Ciències Naturals, Matemàtiques, Tecnologies i Enginyeries, les noies només van representar un 7% dels graduats. Com a conseqüència, cada vegada més s'està creant la necessitat d'estudiar els motius que porten a l'elecció dels estudis per part dels alumnes i comprovar si aquesta diferència entre sexes està provocada per l'existència de trets psicològics o en canvi si està causada per altres raons.

En aquest treball s'analitzarà la diferència existent entre sexes en el grau de Matemàtiques de la Universitat Politècnica de Catalunya, de forma més concreta es vol conèixer si realment el fet de ser home o dona té un impacte rellevant en l'èxit en aquesta carrera, a partir de l'estudi de diverses dades proporcionades per la mateixa institució. Per a la seva realització s'ha comptat amb una beca proporcionada per l'ICE (Institut de Ciències de l'Educació) per a l'estudi de la perspectiva de gènere.

Els tres objectius principals de la investigació són: per una banda, veure quines variables tenen un efecte més important en la nota final dels estudiants i en la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres (la fase inicial està composta per les assignatures del primer curs), i per altra banda, analitzar l'evolució dels alumnes en la carrera. A l'hora d'aconseguir aquests objectius es tindrà molt en compte, com ja s'ha dit, la incidència del sexe i es comprovarà si actualment es troben diferències entre ells. Per poder-los assolir s'aplicaran diverses tècniques estadístiques amb l'ajuda de programes informàtics.

Cal destacar que la hipòtesi inicial de la qual es parteix és que actualment sí existeixen diferències entre sexes en el grau, més concretament, el nombre d'homes supera al de dones, per tant, amb la realització d'aquest treball es vol saber, per una banda si aquesta afirmació és certa o no i com ha evolucionat cap a aquí.

Aquest treball inclourà dues anàlisis principals. La primera consistirà en estudiar com és el desenvolupament dels alumnes en la carrera, nois i noies, segons diverses variables mitjançant l'estudi de la *nota final dels estudiants* i la *probabilitat d'aprovar la fase inicial en dos quadrimestres*. També, es durà a terme una comparació de les proporcions d'homes i dones que aconsegueixen graduar-se dels estudis de Matemàtiques, per tal de veure si realment hi ha diferències en el nombre d'alumnes al acabar la carrera entre sexes i a més a més, si aquestes proporcions es mantenen similars al llarg dels anys. En la segona s'analitzarà l'evolució dels estudiants que accedeixen al grau durant els últims anys en funció de diferents dades. Finalment, es farà una conclusió sobre els resultats extrets.

2. Marc teòric

En aquest apartat s'especificaran totes les tècniques estadístiques emprades per a realitzar les següents parts de l'anàlisi. Per altra banda, les dades que s'usaran són totalment anònimes i han estat proporcionades per la Universitat Politècnica de Catalunya.

2.1. *Models Lineals*

Per començar, per a analitzar les variables que influeixen en la nota final dels estudiants s'aplicaran *Models Lineals*. Aquesta tècnica permet ajustar un model a unes variables amb l'objectiu de veure quin efecte tenen un conjunt d'aquestes, que s'anomenaran explicatives (o regresores), en la variable resposta (o independent). Seguidament, s'explicaran les bases dels *Models Lineals*.

L'expressió més habitual d'un model lineal és:

$$y_i = \beta_0 + \beta_1 * x_{i1} + \cdots + \beta_m * x_{im} + \varepsilon_i, i = 1, \ldots, n$$

En el qual, *y* és la variable resposta, *x* són les variables explicatives, β són els paràmetres del model, ε indica el terme d'error, n fa referència al nombre total d'observacions de la base de dades i m al nombre de variables regresores.

En forma matricial:

$$\begin{pmatrix} y_1 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x1m \\ 1 & \ldots & \ldots & \ldots \\ 1 & x_{n1} & \ldots & x_{nm} \end{pmatrix} * \begin{pmatrix} \beta_0 \\ \cdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \cdots \\ \varepsilon_n \end{pmatrix} \rightarrow Y = X\beta + \varepsilon$$

Pel que fa als paràmetres β, s'estimen mitjançant el mètode dels Mínims Quadrats Ordinaris. Aquest es basa en trobar el conjunt de valors pels paràmetres que minimitzen la suma dels errors (diferència entre el valor de la variable independent real i l'estimació segons el model) al quadrat, ja que entre més petits siguin aquests millor serà l'ajust.

És a dir, es vol trobar $\hat{\beta} = (\widehat{\beta_1}, \ldots, \widehat{\beta_n})$, que minimitzen l'equació:

$$\varepsilon'\varepsilon = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - \widehat{y_i} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 * x_{i1} - \cdots - \beta_m * x_{im})^2,$$

$\widehat{y_i}$ indica els valors estimats pel model per a la variable resposta.

En forma matricial, trobar l'estimació dels paràmetres implica resoldre l'equació:

$$X'X\beta = X'Y \rightarrow \hat{\beta} = (X'X)^{-1} * X'Y, \text{ on } (X'X)^{-1} \text{ és la inversa de la matriu } X'X$$

Per a valorar l'ajust del model estimat a les dades reals s'usarà l'indicador $R^2\ ajustat.$ Aquest parteix del valor d'una altra mesura de bondat de l'ajust, l'R^2 (indica el percentatge de variabilitat de la resposta que es troba explicat pel model), però aquesta té un inconvenient, i és que augmenta si s'augmenta el nombre de variables incloses en el model, en canvi l'R^2 $ajustat$ penalitza la inclusió de variables que no siguin rellevants per explicar la resposta.

- $R^2\ ajustat = 1 - \left[\frac{n-1}{n-k-1}\right] * (1 - R^2),\ R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

Per tant, l'R^2 $ajustat$ s'utilitza per veure l'efectivitat que tenen les variables regresores per explicar la variable independent. Aquest indicador ajudarà a comparar diversos models lineals amb l'objectiu d'escollir el que millor s'adapti a les dades reals, és a dir, es triarà el model que tingui un valor de l'R^2 $ajustat$ més elevat.

Cal esmentar, que es important tenir en compte que el model no ha de patir problemes de col·linealitat. La col·linealitat apareix quan hi ha un elevat grau de relació entre algunes de les variables explicatives. Aquest problema provocarà que:

$$|X'X| \cong 0$$

Com a conseqüència, hi haurà un augment en les variàncies dels estimadors dels coeficients i per tant, aquestes estimacions seran poc precises. Per altra banda, també provoca inestabilitats en el model, és a dir, al fer variar lleugerament la resposta, els coeficients β varien molt.

Per a poder detectar-la es pot calcular el que s'anomena *Factor d'Inflació de la Variància* (*VIF*) per a les variables que formen part del model:

$$VIF_i = \frac{1}{1 - R_i}, on\ R_i\ fa\ referència\ al\ coefcient\ R^2\ d'una\ regressió\ en\ la\ qual\ la\ variable\ i\ és\ la\ resposta\ i\ les\ variables\ regressores\ són\ la\ resta\ d'explicatives.$$

Si el VIF pren un valor elevat (major a 5) és considera que hi ha col·linealitat i per tant, la variabilitat de la variable en concret ja es troba explicada per les altres variables del model. A causa d'això, per a reduir la col·linealitat i solucionar el problema es pot eliminar les variables que es troben correlacionades amb les altres (la pèrdua d'ajust a les dades que aquesta eliminació provocarà serà petita).

Un cop escollit el millor model cal validar-lo. El mètode dels *Mínims Quadrats Ordinaris* suposa que els errors o residus del model, ε, són variables aleatòries que verifiquen les tres condicions següents:

- $E(\varepsilon_i) = 0, \quad i = 1, \dots, n$
- $Var(\varepsilon_i) = \sigma^2, \quad i = 1, \dots, n$
- $E(\varepsilon_i * \varepsilon_j) = 0, \quad i = 1, \dots, n, \ \ j = 1, \dots, n, \quad \forall i \neq j$

A més a més, per a que el model es consideri model lineal normal també s'ha de verificar la condició:

- $e_i \sim N(0, \sigma),\ \ i = 1, \dots, n \quad i \quad e_1, \dots, e_n\ \ són\ independents\ estocàsticament$

Com a conseqüència per a poder validar un model lineal s'ha de comprovar que aquestes suposicions es compleixen. Per a aconseguir-ho es crearan els gràfics:

- Residus vs Valors ajustats $\widehat{y_i}$: aquesta representació permetrà observar si es compleix que $E(\varepsilon_i) = 0$. Aquesta suposició serà certa si els valors del gràfic es troben situats al voltant de l'abscissa igual a 0.
  Per altra banda, aquest gràfic també permetrà comprovar si $Var(\varepsilon_i) = \sigma^2$, és a dir, si la variància dels residus és constant (homoscedasticitat). Això, es validarà si els errors es troben situats aleatòriament i de forma més o menys homogènia per a tots els valors ajustats de la variable resposta al voltant de l'abscissa igual a 0.
- QQ-Plot dels residus del model: Amb aquest gràfic es podrà comprovar si els residus es distribueixen segons una Normal.

Cal destacar, que per a comprovar la hipòtesi de normalitat dels residus també es farà el test de *Shapiro-Wilks*. Les hipòtesis d'aquest són:

- H_0: $Les\ dades\ segueixen\ una\ distribució\ normal: X \sim N(\mu, \sigma^2)$
- H_1: $Les\ dades\ no\ segueixen\ una\ distribució\ normal: X \nsim N(\mu, \sigma^2)$

Seguidament, per a poder dur a terme test estadístics sobre els paràmetres del model, és a dir, poder comprovar si són estadísticament significatius caldrà definir les distribucions estadístiques que segueixen.

Suposant que el model lineal és normal, és a dir, es compleix la hipòtesi de normalitat dels residus, llavors: $Y \sim N(X\beta, \sigma^2 * I_n)$, on I és la matriu identitat i com a conseqüència es verifiquen les següents propietats:

- $\hat{\beta} \sim N(\beta, \sigma^2 * (X'X)^{-1})$
- $\frac{(\hat{\beta}-\beta)' X' X\, (\hat{\beta}-\beta)}{\sigma^2} \sim \chi^2_m$
- $\frac{\sum_{i=1}^{n} e_i^2}{\sigma} \sim \chi^2_{n-m}$

Cal esmentar que l'estimador de σ^2 és:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{(n-r)},$$

r indica el nombre de paràmetres del model.

A partir d'aquestes propietats es podran contrastar les hipòtesis següents sobre els paràmetres:

- $H_0: \beta_i = 0$
- $H_1: \beta_i \neq 0, \quad i = 0, \ldots, m$

L'estadístic del test serà: $t = \frac{\widehat{\beta}_i}{se(\widehat{\beta}_i)} \sim t - student_{n-r}, \; se(\hat{\beta}_i) = \sqrt{Var(\hat{\beta}_i)}$

Amb aquest contrast es podrà conèixer quins dels paràmetres són estadísticament diferents de 0 i per tant tenen una influència en la variable resposta.

Per a saber si la regressió és globalment significativa, és a dir, si almenys un dels paràmetres que acompanyen a les variables regresores és diferent de 0 es farà el test:

- $H_0: \beta_i = 0, \quad \forall\, i = 1, \ldots, m$
- $H_1: algun\; \beta_i \neq 0, \quad i = 1, \ldots, m$

L'estadístic del contrast serà: $F = \frac{SCR_{H0} - SCR/q}{SCR/(n-r)} \sim F_{q,n-r}$, SCR_{H0} és la suma de quadrats dels errors sota el model de la hipòtesi nul·la SCR és la suma de quadrats dels errors sota el model de la hipòtesi alternativa i q és la diferència de paràmetres entre els dos models, en aquest cas degut a que es compara el model creat amb el model nul (només un paràmetre, β_0) q serà igual al nombre de variables explicatives.

Aquest contrast també serveix per a comparar models encaixats (s'anomenarà test *ANOVA*), és a dir, un model amb un altre el qual inclou les mateixes variables explicatives i una més (pot ser més d'una), per tant, es podrà veure si la nova variable inclosa es significativa, és a dir, estadísticament diferent de 0. Les hipòtesis en aquest cas seran (suposant que només s'ha inclòs la variable k com a nova regresora):

- $H_0: \beta_k = 0$
- $H_1: \beta_k \neq 0$

Tot i així, cal esmentar que en el cas que només s'afegeixi una nova variable explicativa aplicant aquest contrast s'arriba a la mateixa conclusió que aplicant el explicat anteriorment en el qual es comprova la significació de cada paràmetre del model, ja que es compleix la relació: $F_{q,n-r}, si\; q = 1\; és\; igual\; a\;\; t - student_{n-r}{}^2$. Però, en canvi pot ser de molta utilitat a l'hora de comparar models en els quals s'afegeixin més d'una variable.

Per altra banda, amb aquestes propietats també es podran construir intervals de confiança pels paràmetres del model:

$$Interval\; de\; Confiança_{(1-\alpha)\%} = \hat{\beta}_i \pm\; t - student_{n-r}\left(\frac{\alpha}{2}\right) * se(\hat{\beta}_i), \qquad i = 0, \ldots, m$$

Finalment, cal destacar que en el model s'inclouran variables categòriques o factors com a variables regresores, com per exemple el sexe dels estudiants per tant, serà interessant dur a terme una anàlisi de la covariància, *ANCOVA*. En aquesta anàlisi es tracta d'estudiar les diferències entre els diversos nivells pels factors sobre una variable quantitativa inclosa com a explicativa en el model, d'aquesta manera es podrà veure si hi ha alguna regresora que té un efecte diferent en funció del nivell que prengui una variable qualitativa.

Per a fer aquesta anàlisi s'inclouran en el model interaccions entre variables explicatives categòriques i quantitatives i es farà el test que s'ha especificat anteriorment per a veure si la regressió era globalment significativa, però en aquest cas el model sota la hipòtesi nul·la suposarà que tots els paràmetres sorgits a causa de la inclusió d'aquesta interacció (serà igual al nombre de nivells del factor menys 1, ja que un d'ells és considera la categoria de referència) són 0 i per altra banda el model de la hipòtesi alternativa considerarà que almenys un d'aquests paràmetres és estadísticament diferent de 0. Per tant, s'aconseguirà conèixer si la interacció entre un factor i una variable quantitativa s'ha d'incloure en el model ja que té un efecte en la variable independent.

Per acabar, cal remarcar que per a fer tots els càlculs i test explicats s'utilitzaran diverses funcions del programa *R*.

2.2. *Models Lineals Generalitzats*

A continuació, per a realitzar l'apartat del treball en el qual s'estudiarà la probabilitat d'aprovar la fase inicial del grau de Matemàtiques en dos quadrimestres s'aplicaran els *Models Lineals Generalitzats,* ja que es tracta d'una variable resposta que només pot prendre valors entre 0 i 1, per tant, les tècniques anteriors no són aplicables. Concretament, s'hauran d'aplicar *Models Lineals Generalitzats per a Resposta Binària (model logístic)*. S'ha d'esmentar que les dades de les quals es disposa són desagregades, és dir, cada fila de la base de dades representa una sola observació. Seguidament, s'explicaran les bases d'aquest tipus de models aplicats a aquestes dades.

En els *Models Lineals Generalitzats*, igual que en els *Models Lineals* també es té un vector anomenat Y, que conté les observacions de la variable independent, una matriu X que conté les dades de les variables regresores i el vector β que conté els paràmetres del model.

Però en aquest cas hi haurà una component aleatòria definida per la següent expressió:

$$f_{y_i} = e^{\frac{\theta_i y - b(\theta_i)}{a(\emptyset)} + c(y,\emptyset)}, \quad independents,$$

s'anomenarà $\phi = a(\emptyset)$ paràmetre de dispersió i $\sqrt{\phi}$ paràmetre d'escala.

Per altra banda, es troba la component determinista que s'anomena predictor lineal: $\eta_i = X\beta$ i la funció d'enllaç (*link*) que és aquella que relaciona el valor esperat μ_i (variable que es vol estudiar) amb el predictor lineal η_i, $g(\mu_i) = \eta_i$, és a dir,

$$g(\mu) = \eta = X\beta$$

Per tant, la variable independent del model no serà la variable resposta real, sinó que serà una transformació d'aquesta, com a conseqüència, per a saber el valor real de la variable s'haurà de desfer la transformació aplicant la inversa de la funció *link*. La funció d'enllaç s'escollirà en funció de com sigui l'esperança del model que es vol descriure.

Els paràmetres β del model s'estimaran aplicant el mateix mètode que en el cas dels *Models Lineals,* però a l'hora d'interpretar l'efecte que tenen en la variable resposta real caldrà tenir en compte la funció d'enllaç escollida.

En aquests models la variància no cal que sigui constant, sinó que aquesta es trobarà definida pel que s'anomenarà funció de variància: $V(\mu_i)$.

Cal destacar, que en els *Models Lineals Generalitzats* existeixen diverses famílies de distribucions en les quals es pot aplicar aquest mètode. En el cas d'estudiar una variable de resposta binària com és el fet de que un alumne aprovi o no la fase inicial del grau de Matemàtiques en 2 quadrimestres caldrà definir un *Model Lineal Generalitzat* en el qual la família serà la distribució Binomial (aquesta distribució indica la probabilitat de que uns individus tinguin o no una característica concreta).

Per a aquesta distribució:

- $\theta_i = q(\mu_i) = \log\frac{p_i}{1-p_i}$, p indica la probabilitat de resposta positiva.
- $\emptyset = 1$
- $\phi = 1$
- $b(\theta_i) = N * \log\left(1 + e^{\theta_i}\right), N$ indica el nombre total d'observacions.
- $\mu_i = b'(\theta_i) = \frac{N * e^{\theta_i}}{1+e^{\theta_i}}$
- $V(\mu_i) = N * p_i * (1 - p_i)$
- $Var\left(Y_i, (\theta_i, \emptyset)\right) = \phi * V(\mu_i)$

L'objectiu dels *Models de Resposta Binària* es analitzar la relació entre la resposta Y i les variables explicatives mitjançant la probabilitat de resposta positiva, s'anomenarà π.

La funció d'enllaç més usada per a la família Binomial és el *link logit*, aquesta es defineix de la manera següent:

$$g(\pi) = \eta = logit\ (\pi) = \log\left(\frac{\pi}{1-\pi}\right), \qquad per\ tant, \qquad g(\eta)^{-1} = \frac{e^{\eta}}{1+e^{\eta}} = \pi$$

$\frac{\pi}{1-\pi}$ s'anomena *odds* i representa la relació entre la probabilitat de resposta positiva i resposta negativa. Si l'*odds* pren un valor major a 1 indica que la probabilitat de resposta positiva és major a la de negativa, al contrari succeeix per a valors menor a la unitat. Com a conseqüència l'exponencial dels paràmetres β representarà l'*odds-rati*, és a dir, la relació entre dues *odds* en funció del valor que pren la variable explicativa que acompanya al paràmetre corresponent.

Per a avaluar l'ajust dels *Models Lineals Generalitzats de Resposta Binària* s'usa principalment el que s'anomena:

- Deviança Residual: $D(y,\hat{\pi}) = 2 * l(y,y) - 2 * l(\hat{\pi},y)$ on $l(y,y)$ és la funció de versemblança del model saturat (aquell que té tantes variables com observacions, es considera l'ajust *perfecte*) i $l(\hat{\pi},y)$ la funció de versemblança del model creat.

Cal esmentar, que a partir d'aquest estadístic es pot dur a terme un test per a comparar dos models encaixats amb l'objectiu de conèixer si les variables incloses en un dels models són necessàries per a explicar la variabilitat de la resposta. Suposant que es tenen dos models (A amb q paràmetres i B amb p paràmetres, p > q) les hipòtesis seran:

- $H_0: El\ model\ B\ no\ millora\ l'ajust\ del\ model\ A$
- $H_1: El\ model\ B\ sí\ millora\ l'ajust\ del\ model\ A$

L'estadístic del contrast serà: $\Delta D_{AB} = D(y,\hat{\pi}_A) - D(y,\hat{\pi}_B) \sim \chi^2_{p-q}$.

Per altra banda, per a comparar models també es poden usar els següents indicadors:

- $AIC\ (Criteri\ d'Informació\ d'Akaike) = 2 * \left(p - l(\hat{\pi},y)\right), p$ indica el nombre de paràmetres del model.
- $BIC\ (Criteri\ d'Informació\ Bayesià) = p * \log(n) - 2 * \left(p - l(\hat{\pi},y)\right),\ n$ indica el nombre total d'observacions.
- $Pseudo - R^2 = 1 - \frac{D(y,\hat{\pi}_A)}{D(y,\hat{\pi}_0)}$, on $D(y,\hat{\pi}_A)$ es la deviança residual del model creat i $D(y,\hat{\pi}_0)$ és la deviança del model nul (aquell que només té un paràmetre, β_0).

Pel que fa a l'AIC i al BIC s'escollirà el model que tingui un valor més baix d'aquests indicadors, en canvi, pel $Pseudo - R^2$ es preferible el model que el tingui més elevat. Cal destacar, que tot i que l'AIC i el BIC són molt similars, aquest últim penalitza de forma més elevada la inclusió de paràmetres.

Seguidament, per a validar el model és necessari calcular els residus d'aquest. Pels *Models Lineals Generalitzats* existeixen principalment dos tipus de residus:

- Residus de Pearson:

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}},$$

$\hat{\mu}$ indica el valor predit pel model per a la probabilitat de resposta positiva.

$$\sum_{i=1}^{N} (r_i^P)^2 = X^2\ \ Estadístic\ de\ Pearson$$

L'Estadístic de Pearson és molt útil per a valorar l'ajust d'un model. Pel cas de dades desagregades un model és bo si el seu valor és semblant al nombre d'observacions de la base de dades menys el de paràmetres d'aquest.

- Residus Deviança:

$$r_i^D = sign\,(y_i - \widehat{\mu_i}) * d_i, \quad d_i = \sqrt{2\left(\left(\widetilde{\theta_i}y - b(\widetilde{\theta_i})\right) - \left(\widehat{\theta_i}y - b(\widehat{\theta_i})\right)\right)}$$

$\widetilde{\theta_i}$ és el valor de θ_i pel model saturat i $\widehat{\theta_i}$ és el del model creat.

Aquests residus compleixen que: $\sum_{i=1}^{N}(r_i^D)^2 = D(y, \hat{\pi})$.

Un cop calculats els residus per a donar un model per correcte cal representar-los gràficament per observar com es distribueixen:

- Gràfic residus de Person/Deviança vs Predictor Lineal: En aquesta representació s'ha de veure que la distribució dels residus per als diferents valors del predictor lineal és semblant. A causa de que la resposta d'aquest tipus de models són 0 o 1 aquesta distribució és complicada d'observar. Un model correcte requereix que la funció mitjana condicional sigui constant per a tots els valors del predictor lineal, per això s'usa la ajuda del programa *R*, ja que aplicant una funció mostra el gràfic i li afegeix una representació d'una línia suavitzada, com a conseqüència si aquesta és prou recta (és a dir, és constant) es donarà el model per validat en aquest aspecte.
- Gràfic residus de Person/Deviança vs Variables explicatives: En aquests gràfics s'ha d'observar que els residus es distribueixen de forma més o menys homogènia per a tots els possibles valors de les variables explicatives. En aquest cas, també s'observarà que la línia suavitzada representada pel programa *R* sigui constant al llarg de tots els valors de les variables regresores.

A continuació, per a mesurar la capacitat predictiva del model s'han de calcular quatre mesures, a més a més aquestes també ajudaran a conèixer la precisió de les prediccions futures:

- Sensibilitat: proporció d'observacions que s'han predit mitjançant el model com positives ($\hat{y}_i = 1$) dins dels resultats positius observats ($Y = 1$).
- Especificitat: proporció d'observacions que s'han predit mitjançant el model com negatives ($\hat{y}_i = 0$) dins dels resultats negatius observats ($Y = 0$).
- Valor Predictiu Positiu: proporció de resultats positius observats ($Y = 1$) dins dels que s'han predit com a positius ($\hat{y}_i = 1$).
- Valor Predictiu Negatiu: proporció de resultats negatius observats ($Y = 0$) dins dels que s'han predit com a negatius ($\hat{y}_i = 0$).

A partir d'aquestes mesures es representa el que s'anomena corba ROC. L'anàlisi de la corba ROC serveix per a comprar i descriure la precisió de les prediccions. Aquesta representa per a cada possible llindar (és a dir, el valor de la probabilitat que prediu el model a partir del qual es considera que la resposta es positiva o negativa) el valor de la $Sensibilitat$ vs la taxa de falsos positius ($1 - Especificitat$). Si la corba ROC augmenta ràpidament es considera que el model funciona bé. Per això, es calcularà l'estadístic AUC (Àrea Sota la Corba ROC) i segons el seu valor es pot saber com és el model:

- $0.9 - 1$: $El\ model\ és\ excel \cdot lent.$
- $0.8 - 0.9$: $El\ model\ és\ molt\ bo.$
- $0.7 - 0.8$: $El\ model\ és\ bo.$
- $0.6 - 0.7$: $El\ model\ és\ dolent.$
- $0.5 - 0.6$: $El\ model\ és\ molt\ dolent.$

En aquest cas, igual que pel cas dels Models Lineals s'usarà el programa *R* per a fer els càlculs i tests explicats.

2.3. Comparació de Proporcions

A més a més, per a verificar que realment existeixen diferències significatives entre sexes al grau de Matemàtiques es durà a terme un test per a comparar les proporcions de nois i noies que han aconseguit finalitzar aquests estudis.

S'ha d'esmentar, que si les proporcions dels dos sexes fossin iguals serien cada una 0.5, és a dir, hi hauria un 50% de nois i un 50% de noies al grau, com a conseqüència només serà necessari comparar una de les dues proporcions amb aquest valor.

Les premissa que s'ha de complir per a poder aplicar el contrast és:

- La variable observada correspon a una variable Binomial (mesura el nombre de vegades que es produeix un succés en una mostra).

En aquest cas, es comprarà una mostra formada per tots els alumnes que han aconseguit graduar-se dels estudis de Matemàtiques i la variable d'estudi serà quants d'ells són nois. La proporció d'estudiants del sexe masculí s'anomenarà p i la mida mostral es denotarà per n.

Les hipòtesis del test són:

- $H_0: p = 0.5$
- $H_1: p > 0.5$

Cal destacar, que s'ha decidit aquesta hipòtesi alternativa, en comptes de: $H_1: p \neq 0.5$, a causa de que l'objectiu principal de l'anàlisi es conèixer si la proporció d'homes és major a la de dones o si per el contrari aquestes són iguals.

L'estadístic per a resoldre aquest contrast és:

$$Z = \frac{\hat{p} - 0.5}{\sqrt{\frac{\hat{p} * \hat{q}}{n}}} \sim N(0,1)$$

$$on \quad \hat{p} = \frac{nombre\ d'alumnes\ nois}{total\ d'alumnes\ graduats} \qquad i \qquad \hat{q} = 1 - \hat{p}$$

Per tant, es rebutjarà la hipòtesi nul·la amb un nivell de significació α si: $Z \geq Z_{\alpha}$ (valor crític unilateral de la distribució Normal (0, 1) associat a una probabilitat α $(P(Z \geq Z_{\alpha}) = \alpha)$.

S'ha d'esmentar que l'estadístic Z es distribueix de forma asimptòtica com una distribució Normal tipificada, com a conseqüència per a que aquesta aproximació sigui adequada es requereix que:

- $n \geq 30$
- $n * \hat{p} \geq 5$

Finalment, calcular intervals de confiança per a cada una de les proporcions, d'aquesta manera es podrà conèixer entre quins valors es situarà la proporció de nois i la de noies al grau amb un nivell de confiança seleccionat.

Tot i que el contrast es farà unilateral pot ser interessant calcular un interval de confiança bilateral, ja que aquest indica tant una cota inferior com superior per a les proporcions, en canvi l'interval unilateral només indica, en aquest cas, una cota inferior pel cas dels nois o una superior pel cas de les noies (a causa de la definició de la hipòtesi alternativa).

- $Interval\ de\ Confiança_{(1-\alpha)\%}\ bilateral = \left(\hat{p} \pm z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}\ (1-\hat{p})}{n}}\right)$

- $Interval\ de\ Confiança_{(1-\alpha)\%}\ unilateral\ nois =$

$$\left(\left(\hat{p} + z_{\alpha} * \sqrt{\frac{\hat{p}\ (1-\hat{p})}{n}}\right);\ 1\right)$$

- $Interval\ de\ Confiança_{(1-\alpha)\%}\ unilateral\ noies =$

$$\left(0; \left(\hat{p} + z_{\alpha} * \sqrt{\frac{\hat{p}\ (1-\hat{p})}{n}}\right)\right)$$

Per a realitzar tots els càlculs comentats s'usarà el programa *R*.

Per altra banda, un cop realitzada la comparació de la proporció de nois i noies per a cada any, també serà interessant conèixer si aquestes es comporten de forma similar en els diferents períodes.

Per aconseguir aquest objectiu s'aplicarà una prova Chi-quadrat d'homogeneïtat. Aquesta es basa en comprovar si un conjunt de poblacions son homogènies respecte una característica categòrica concreta, és a dir, si la distribució de les categories de la variable d'estudi és la mateixa per a totes de les poblacions.

En aquest cas, la variable d'estudi és el sexe, per tant, hi ha dues categories, home i dona, i les poblacions són els alumnes del grau de Matemàtiques per a cada any.

Les premisses que s'han de complir per a poder aplicar el test és:

- La grandària mostral ha de ser prou gran: $N \geq 25$.
- La freqüència observada per a cada categoria no pot ser massa petita, almenys 3.

Les hipòtesi del contrast d'homogeneïtat són:

- $H_0: P_{1H} = \cdots = P_{RH} = P_H,$
  $P_{1D} = \cdots = P_{RD} = P_D$
- $H_1: P_{ij} \neq P_j \quad per\ algun\ i = 1, \dots, R \quad j = H, D$

$$R\ indica\ el\ nombre\ total\ d'anys,$$
$$P_{ij}\ fa\ referència\ a\ la\ probabilitat\ de\ pertànyer\ a\ la\ categòria\ j\ (H = Home,$$
$$D = Dona)\ a\ la\ població\ i.$$

Per tant, es vol comprovar si la probabilitat d'home i dona són les mateixes per a tots els anys de la base de dades o si per el contrari alguna d'aquestes probabilitats és diferent.

Per a calcular l'estadístic del test es crearan les dues taules de contingència següents:

- Taula de contingència dels efectius observats: Cada fila de la taula de contingència correspondrà a una població, que en aquest cas, seran els alumnes del grau per a un any concret. Les columnes representaran les categories de la variable d'estudi (home o dona), i finalment, els valors de cada fila estaran formats per les freqüències observades.

	Home	Dona	Total
Població 1	O_{1H}	O_{1D}	n_1
...	...	...	...
Població R	O_{RH}	O_{RD}	n_R
Total	n_H	n_D	N

$$O_{ij} = Efectius\ observats\ per\ a\ l'any\ i, i\ la\ categòria\ j.$$
$$n_i = Efectius\ totals\ de\ l'any\ i$$
$$n_j = Efectius\ totals\ de\ la\ categòria\ j$$
$$N = Nombre\ total\ d'efectius$$

- Taula de contingència dels efectius esperats: Aquesta taula té la mateixa forma que l'anterior però cada fila estarà formada pels valors que s'esperarien si la hipòtesi nul·la fos certa. La freqüència esperada sota H_0 es calcularà:

$$E_{ij} = \frac{n_i * n_j}{N}$$

Finalment, es calcularà l'estadístic del test amb la formula següent:

$$X^2 = \sum_{j\,\in\,\{H,D\}} \sum_{i=1}^{R} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

$$\text{Sota } H_0: X^2 \sim \chi^2_{(R-1)x(k-1)},$$

$$k = 2 \ (nombre\ de\ categòries), \qquad R\ indica\ el\ nombre\ total\ d'anys.$$

Com a conseqüència, es rebutjarà la hipòtesi nul·la amb un nivell de significació α si:

$X^2 > \chi^2_{(R-1)x(2-1)}(\alpha)$ (valor crític unilateral de la distribució Chi-quadrat amb $(R-1)x(2-1)$ graus de llibertat associat a una probabilitat α.

Per a realitzar tots els càlculs necessaris per a dur a terme el contrast s'usarà el programa *R*.

2.4. Anàlisi de sèries temporals

Per a observar com ha estat l'evolució dels homes i les dones al grau de Matemàtiques durant els diferents anys s'usarà el programa Excel, mitjançant el qual s'analitzaran les dades i es crearan diversos gràfics. Cal destacar que les dades formen el que s'anomena una sèrie temporal, és a dir, són una seqüencia d'observacions ordenades de forma cronològica.

A partir de les representacions es podrà veure quina ha estat la tendència dels alumnes en aquests estudis i es podran identificar discrepàncies entre sexes, per a poder comprovar això també es calcularà el test de Daniel, les hipòtesis del qual són:

- H_0: *La sèrie temporal no té tendència*
- H_1: *La sèrie temporal sí té tendència*

Gràcies al resultat del contrast es podrà conèixer si realment el nombre de noies o nois al grau tendeix a augmentar o disminuir amb el temps.

L'estadístic del test és:

$$Z = \sqrt{T-1} * T_s \sim N(0,1), \quad on \quad T_s = 1 - \frac{6 * \sum_{t=1}^{T} d_t^2}{T\,(T^2-1)}$$

$$d_t = Rang(y_t) - t, \quad T = nombre\ total\ d'observacions,$$

$$t = variable\ que\ pren\ valors\ des\ de\ 1\ fins\ a\ T\ ordenats\ de\ forma\ temporal.$$

$$Rang(y_t) = variable\ que\ indica\ el\ lloc\ que\ ocupa\ cada\ observació\ en\ la\ sèrie\ ordenada\ de\ menor\ a\ major$$

Per tant, per un nivell de significació α ($N_{\alpha/2} = valor\ de\ la\ distribució\ de\ la\ N(0,1)\ que\ deixa\ per\ sobre\ i\ per\ sota\ una\ probabilitat\ igual\ a\ \alpha/2$):

- $Si\ |Z| > N_{\alpha/2}\ es\ rebutjarà\ H_0,\ la\ sèrie\ té\ tendència$
- $Si\ |Z| < N_{\alpha/2}\ no\ es\ rebutjarà\ H_0,\ la\ sèrie\ no\ té\ tendència$

Seguidament, si s'obté, a causa del contrast anterior, que les dades tenen una tendència s'aplicarà un mètode d'anàlisi determinista adequat per a aquest tipus de sèries temporals anomenat *Mètode de la tendència lineal*.

S'ha escollit aquest mètode degut a que és un dels més senzills i ràpids d'aplicar dels que existeixen i per altra banda, és suficient per a poder assolir l'objectiu d'extreure com es la tendència de les dades. A més a més, tot i que no és l'objectiu principal, aquest mètode també permetrà fer noves prediccions sobre la variable d'estudi.

El *Mètode de la tendència lineal* es basa en aproximar la tendència mitjançant una funció lineal:

$$y_t = T_t + u_t \;\rightarrow\; T_t = \beta_0 + \beta_1 * t \;\rightarrow\; y_t = \beta_0 + \beta_1 * t + u_t,$$

$$u_t \text{ indica l'error en l'estimació de les dades de la sèrie temporal}$$

El mètode consisteix en estimar els valors dels paràmetres β_0 i β_1 utilitzant tota la informació de la mostra aplicant el mètode dels *Mínims Quadrats Ordinaris*. Per tant, l'estimació de la tendència i de les dades de la sèrie seran:

$$\widehat{T_t} = \widehat{\beta_0} + \widehat{\beta_1} * t \;\rightarrow\; \widehat{y_t} = \widehat{\beta_0} + \widehat{\beta_1} * t, \quad t = 1, \dots, T$$

Per a mesurar la capacitat predictiva del mètode es dividirà la sèrie en dos conjunts d'observacions, el primer definirà el període mostral, i s'usarà per a aplicar el *Mètode de la tendència Lineal*, és a dir, calcular els paràmetres β_0 i β_1, i l'altre formarà el període extra-mostral i servirà per a conèixer si les prediccions que fa el mètode són bones. Per això es calcularà l'*Error Percentual Absolut Mitjà (EPAM):*

$$EPAM\ (H) = \frac{100}{H} * \sum_{l=1}^{H} \left| \frac{e_T(l)}{y_{T+l}} \right|, \quad e_T(l) = y_{T+l} - \hat{y}_{T+l}$$

$$y_{T+l} \text{ i } \hat{y}_{T+l} \text{ són el valor real i estimat pel model, respectivament, de l'observació}$$
$$T + l, H \text{ indica el nombre total d'observacions del període extra} - \text{mostral,}$$
$$T \text{ indica el nombre total d'observacions del període mostral.}$$

La interpretació de l'*EPAM* pel que fa a la capacitat predictiva del mètode és la següent:

- $EPAM \leq 1\%$: *Molt bona capacitat predictiva*
- $1\% < EPAM \leq 3\%$: *Bona capacitat predictiva*
- $3\% < EPAM \leq 5\%$: *Capacitat predictiva regular*
- $EPAM > 5\%$: *Baixa/molt baixa capacitat predictiva*

Finalment, un cop valorat la capacitat predictiva del mètode es faran prediccions de la variable d'estudi per als pròxims anys. Però, en aquest cas, per a aplicar el mètode de la tendència lineal s'usaran totes les dades de les que es disposen per a que els resultats siguin els més pròxims a la realitat possible.

3. Estudi de la base de dades

3.1. *Anàlisi descriptiva univariant de la base de dades*

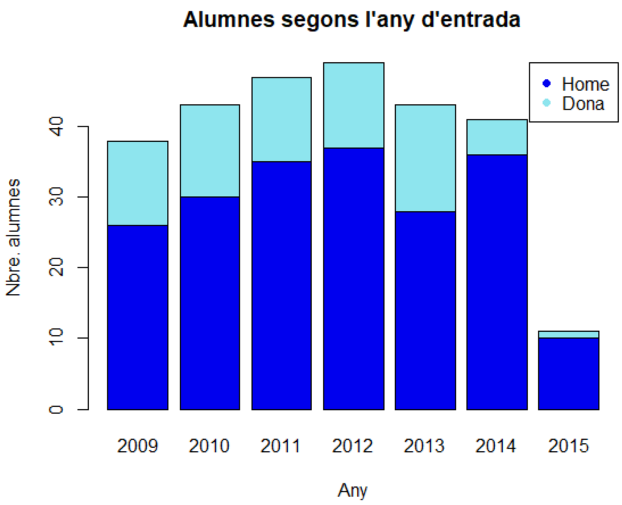
Abans de dur a terme les diverses regressions és important observar com es distribueixen cada una de les variables de la base de dades per separat, a més a més, a l'hora d'analitzar-les es tindrà en compte el factor *sexe*, d'aquesta manera es podrà veure en quines d'elles aquest aspecte té més influència. Les variables[1] que es troben a la base de dades són:

- Any d'entrada: fa referència a l'any d'entrada de l'estudiant.
- CFIS: indica si l'alumne pertany al Centre de Formació Interdisciplinària Superior o no. Els estudis que el centre ofereix són d'una dificultat més alta i per tant, aquest duu a terme un procés de selecció propi que garanteix la idoneïtat dels estudiants. Dins d'aquest procés s'inclou, addicionalment als procediments establerts per accedir a la universitat, la superació d'unes proves específiques de coneixements de matemàtiques i física.
- Nota d'accés a la universitat i nota mínima: aquestes dues dades indiquen la nota amb la qual un alumnes va accedir a la universitat i la nota mínima d'accés dels estudiants que ho van fer aquell any, respectivament. Cal destacar, que la nota d'accés no es pot valorar igual l'any que la nota mínima va ser alta o en canvi, va ser baixa (és a dir, per exemple, no és el mateix que un alumnes entri a la universitat amb una nota de 9 l'any que la nota mínima d'accés va ser un 8, que ho faci quan aquesta nota és un 6), per tant, es creu més convenient estudiar la diferència entre aquestes dues dades.
- Prova de nivell: aquesta dada indica la nota que un estudiant va obtenir a una prova de nivell voluntària que es fa a l'inici del curs.
- Sexe: aquesta variable fa referència al sexe dels alumnes, és a dir, si són homes o dones.
- Nombre de quadrimestres que triguen en aprovar la fase inicial del grau.
- Assignatures obligatòries: aquest conjunt de dades indiquen les notes de les 25 assignatures obligatòries dels alumnes.
- Nombre assignatures aprovades en la primera matricula: fa referència al nombre d'assignatures obligatòries que cada estudiant aconsegueix aprovar el primer cop que les matricula.

A continuació, es farà una anàlisi descriptiu usant diverses funcions del programa *R* [2], de totes les variables mencionades i de la variable resposta (*nota final*).

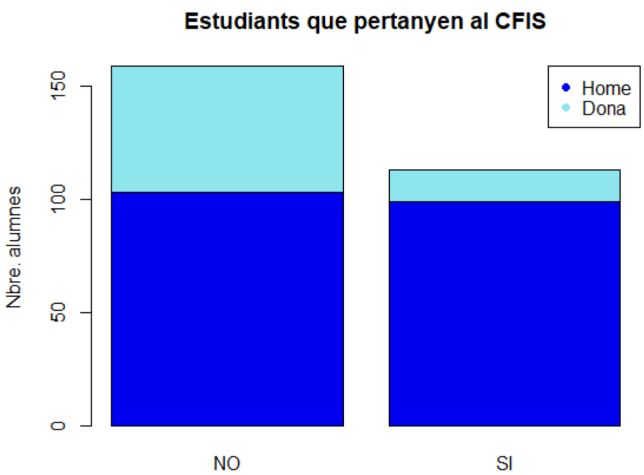
[1] A l'annex es pot veure una mostra de la base de dades: Dades dels apartats 4. Anàlisi de la nota final dels estudiants del grau de Matemàtiques i 5. Anàlisi de la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres.

[2] Els codis usats es poden trobar a l'annex: Codi d'R per a l'anàlisi descriptiu de la base de dades dels apartats 4. Anàlisi de la nota final dels estudiants del grau de Matemàtiques i 5. Anàlisi de la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres.



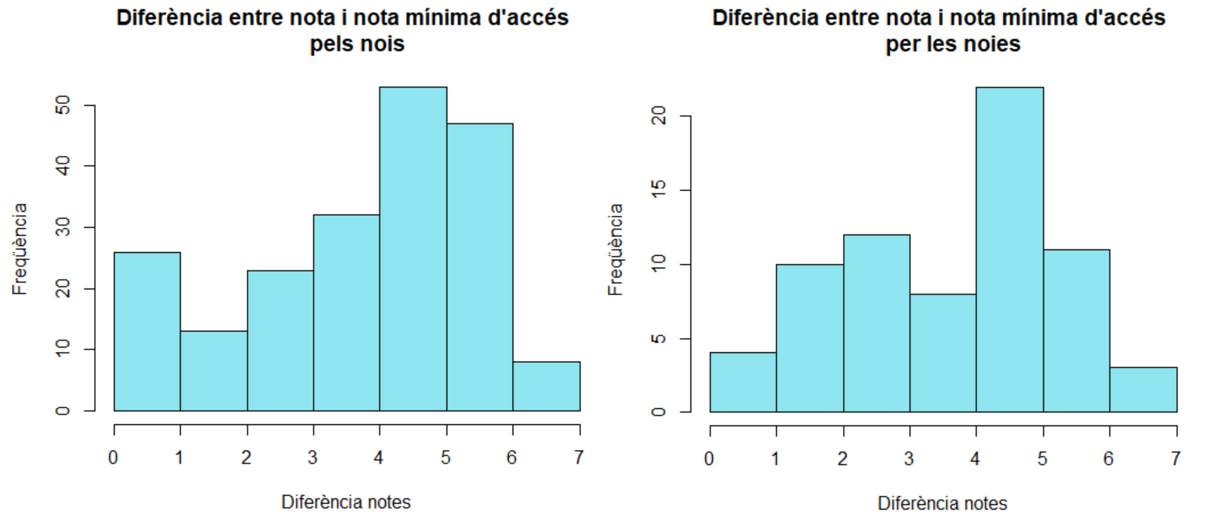
Gràfic 3.1.1. Nombre d'alumnes del grau de Matemàtiques segons l'any d'entrada del 2009 al 2015.

En el gràfic 3.3.1 s'observa que el nombre d'alumnes graduats del sexe masculí va augmentant des del 2009 al 2012, seguidament pels de l'any 2013 es produeix una disminució, pels del 2014 va tornar a augmentar i finalment, pel nombre d'alumnes graduats que van accedir a la universitat l'any 2015 s'ha produït una disminució destacable. Pel que fa a les noies, el nombre d'alumnes graduades s'ha mantingut similar per les que van entrar durant els anys 2009-2012, al 2013 va augmentar lleugerament, en l'any 2014 ha haver una disminució i per acabar l'any 2015 representen un percentatge molt petit del total de graduats. Però, cal destacar que del 2015 només s'han graduat 10 alumnes, aproximadament, i segurament durant els pròxims anys es continuïn graduant estudiants que van entrar aquell període, ja que aquest grau, tot i que està previst per a cursar-lo en 4 anys, és molt exigent.



Gràfic 3.1.2. Nombre estudiants que pertanyen al CFIS segons el sexe.

En el gràfic 3.1.2 es pot veure clarament que el nombre de nois que pertany al CFIS supera de forma molt detacable al de noies que hi pertanyen. A més a més, pel que fa al nombre total d'estudiants masculins, aquest es troba prou ben repartit entre els que cursen CFIS i els que no ho fan, en canvi, pel que fa a les dones predominen aquelles que no formen part d'aquest centre.



Gràfic 3.1.3. Distribució de la diferència entre la nota d'accés i la nota mínima de l'any corresponent en funció del sexe pels estudiants del grau de Matemàtiques.

En el gràfic 3.1.3 s'observa que els nois acostumen a accedir a la universitat amb una diferència entre la seva nota d'accés i la nota mínima d'aquell any d'entre 4 i 6, en canvi, pel que fa a les noies aquestes ho fan amb una diferència d'entre 4 i 5, majoritàriament. Com a conseqüència es pot dir, que, de mitjana, els homes tenen notes d'accés més altes (en comparació amb la nota mínima) que les noies.

Pel que fa a la nota de la prova de nivell voluntària que es fa a principi del curs, en la següent taula es poden observar els valors dels principals estadístics descriptius d'aquesta variable.

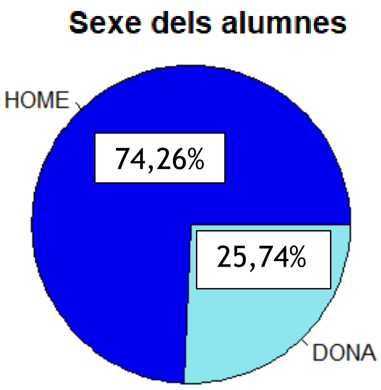
	Mín.	Mediana	Màx.	Mitjana	Alumnes que no fan la prova	Total
HOMES	3,790	7,240	10	7,194	80	202
DONES	3,970	6,295	8,790	6,369	12	70
T-test	Estadístic T = 4.1389 (124.23 g.ll.)				p-valor = 3.194e-05	

Taula 3.1.1. Principals estadístics descriptius de la variable nota de la prova voluntària del grau de Matemàtiques en funció del sexe i test T-Student per a la comparació de mitjanes.

En la taula 3.1.1 es pot veure que, de mitjana, els nois obtenen notes més altes en la prova de nivell que les noies. Però, per altra banda, es troben 80 homes dels 202 totals (40% aproximadament) que no la duen a terme i en canvi, només gairebé un 20% de les dones escull no realitzar-la. Per altra banda, gràcies al test de comparació de mitjanes, (les hipòtesis del qual són:

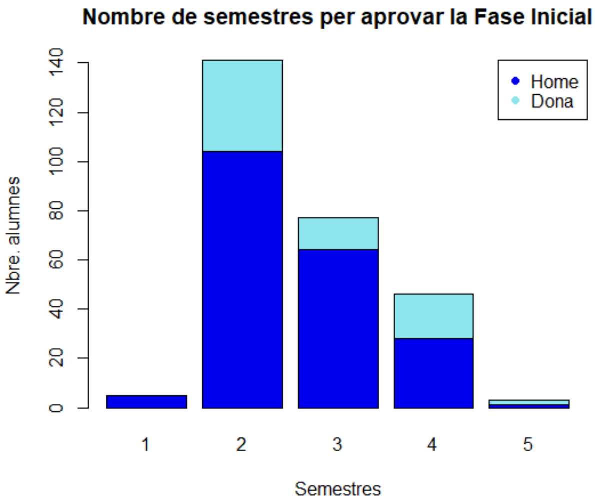
- $H_0: \mu_{Homes} = \mu_{Dones}$
- $H_1: \mu_{Homes} > \mu_{Dones}$)

es pot concloure que, pel que fa a les dades disponibles, sí existeixen diferències entre sexes, concretament les noies tenen de mitjana notes més baixes en comparació amb els nois. Tot i així, causa de que aquesta prova no és obligatòria i per tant, no tots els estudiants la realitzen no es poden extreure conclusions fiables en relació als valors d'aquesta dada.



En el gràfic 3.1.4 s'observa clarament, que sí hi ha diferències destacables entre el nombre de noies i el de nois graduats de Matemàtiques pels anys 2009-2015, concretament, es troben més homes que dones. Aproximadament, les alumnes del sexe femení representen un quart (25%) del total de graduats.

Gràfic 3.1.4. Nombre d'alumnes del grau segons el sexe.



En el gràfic 3.1.5 es pot veure que no es troben grans diferències entre el nombre de semestres que els alumnes triguen en aprovar la fase inicial en funció del sexe, sinó que tant pels nois com per les noies el més habitual es fer-ho en 2 semestres.

Gràfic 3.1.5. Nombre de semestres que els alumnes triguen en aprovar la fase inicial del grau en funció del sexe.

Seguidament, es farà una taula en la qual es pot observar el valor dels principals estadístics descriptius per a les variables que indiquen les notes de les assignatures obligatòries en funció del sexe.

	Mín.	1r Quartil	Mediana	3r Quartil	Màx.	Mitjana
Fonaments de la Matemàtica						
HOMES	5	6	7	8,5	10	7,201
DONES	5	5,5	6,15	7,275	10	6,464
Àlgebra Lineal						
HOMES	5	5,9	7,10	8,5	10	7,194
DONES	5	5,325	6,05	7	10	6,327
Càlcul en una Variable						
HOMES	5	5,5	7	7,9	10	6,945
DONES	5	5,2	6	6,7	10	6,15
Informàtica						
HOMES	5	6	7,2	8,6	10	7,411

DONES	5	5,2	5,95	7,2	10	6,357
			Càlcul Diferencial			
HOMES	5	5,425	7	8	10	6,945
DONES	5	5,225	5,8	7,15	10	6,276
			Geometria Afí i Euclidiana			
HOMES	5	5,7	7	8	10	6,903
DONES	5	5,225	6	7,4	10	6,474
			Àlgebra Lineal Numèrica			
HOMES	5	6,4	7,2	8,6	10	7,486
DONES	5	6,025	7,1	7,8	9,5	7,006
			Matemàtica Discreta			
HOMES	5	6,3	7,3	8,7	10	7,457
DONES	5	5,4	6,1	7,25	10	6,42

Taula 3.1.2. Principals estadístics descriptius de les variables nota de les assignatures obligatòries del primer curs del grau de Matemàtiques.

En la taula 3.1.2 s'observa que en gran part de les assignatures que formen el primer curs del grau de Matemàtiques hi han diferències destacables entre els dos sexes (excepte, en Àlgebra Lineal Numèrica, en la qual es troben valors prou semblants). Pel que fa a les notes mitjanes, aquestes acostumen a estar al voltant de 6 i 7, a més a més entre els homes i les dones hi ha diferències d'aproximadament 1 punt en la majoria de casos, tenint els nois la nota més alta. Per altra banda, si es tenen en compte els valors de la mediana i tercer quartil es pot concloure, també, que els estudiants del sexe masculí arriben a puntuacions més altes que les del sexe femení.

	Mín.	1r Quartil	Mediana	3r Quartil	Màx.	Mitjana
			Càlcul Integral			
HOMES	5	5,7	6,5	7,6	10	6,779
DONES	5	5,4	6	7,075	9,8	6,384
			Àlgebra Multilineal i Geometria			
HOMES	5	5,5	6,6	8,2	10	6,904
DONES	5	5,225	6	7	9,6	6,323
			Algorísmia			
HOMES	5	5,4	6,3	7,8	10	6,737
DONES	5	5	5,65	6,7	9	5,99
			Programació Matemàtica			
HOMES	5	6	7	8,175	10	7,107
DONES	5	5,8	6,45	7,2	10	6,637
			Funcions de variable complexa			
HOMES	5	6,4	7,6	9,175	10	7,651
DONES	5	5,6	6,4	7,35	10	6,641
			Topologia			
HOMES	5	6	7,3	8,925	10	7,363
DONES	5	5,325	6	7,3	9,5	6,439
			Física			
HOMES	5	6,125	7,25	8,675	10	7,386

DONES	5	5,15	6,05	7	10	6,314
			Anàlisi Real			
HOMES	5	6	7	8,175	10	7,096
DONES	5	5,325	5,9	6,950	9,3	6,176

Taula 3.1.3. Principals estadístics descriptius de les variables nota de les assignatures obligatòries del segon curs del grau de Matemàtiques.

En la taula 3.1.3 es pot veure que en diverses de les assignatures obligatòries del segon curs hi han diferències significatives entre sexes (només en l'assignatura Càlcul Integral s'observen valors similars). En aquest cas, igual que succeïa en el primer curs, les notes mitjanes de nois i noies es diferencien per 1 punt aproximadament, en la majoria de casos. A més a més, també es segueixen mantenint els valors més habituals de les notes mitjanes, entre 6 i 7. Finalment, observant els valors de les medianes i dels tercers quartils es pot concloure que els homes acostumen a arribar a notes més elevades que les dones, per tant, es manté la mateixa tendència que en el primer curs.

	Mín.	**1r Quartil**	**Mediana**	**3r Quartil**	**Màx.**	**Mitjana**
			Equacions Diferencials Ordinàries			
HOMES	5	6,4	7,5	8,6	10	7,551
DONES	5	5,7	6,65	7,525	10	6,737
			Estructures Algebraiques			
HOMES	5	5,3	6,55	8	10	6,865
DONES	5	5	6	7	9,8	6,191
			Càlcul Numèric			
HOMES	5	6,3	7,6	8,6	10	7,528
DONES	5	5,7	6,55	7,8	10	6,823
			Teoria de la Probabilitat			
HOMES	5	5,925	7	8,5	10	7,192
DONES	5	5,2	6	6,65	9,7	6,206
			Equacions en Derivades Parcials			
HOMES	5	5,3	6,50	7,8	10	6,748
DONES	5	5,0	5,35	6,5	10	5,866
			Geometria Diferencial			
HOMES	5	6,4	7,5	8,5	10	7,438
DONES	5	6	6,5	7,3	9,5	6,639
			Models Matemàtics de la Física			
HOMES	5	6,4	7,55	8,6	10	7,535
DONES	5	6,325	7,2	8	10	7,134
			Estadística			
HOMES	5	6,025	7,3	8,45	10	7,291
DONES	5	5,925	6,5	7,5	9,7	6,786

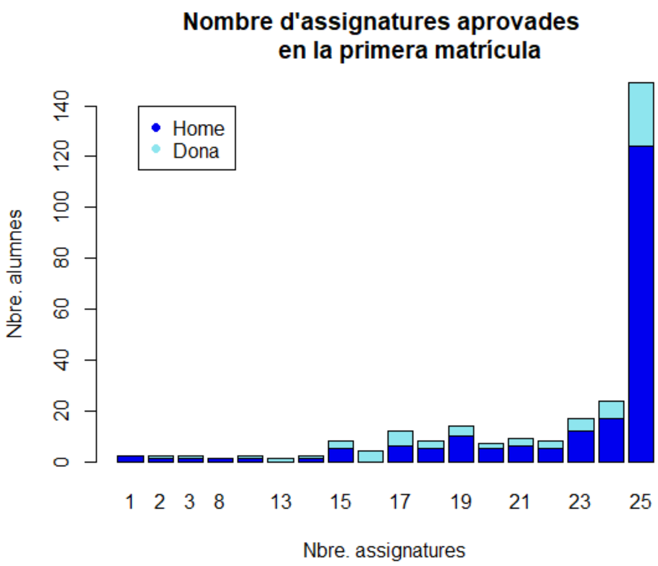
Taula 3.1.4. Principals estadístics descriptius de les variables nota de les assignatures obligatòries del tercer curs del grau de Matemàtiques.

En la taula 3.1.4 s'observa que, igual que pels dos cursos anteriors, en aquet cas també hi ha diferències destacables entre els dos sexes (excepte en l'assignatura Models Matemàtics de la Física, en la qual es troben valors pels estadístics lleugerament semblants). Pel que fa a les notes mitjanes de nois i noies, aquestes s'acostumen a trobar entre 6 i 7, a més a més també cal destacar que entre sexes existeix una diferència d'aproximadament un punt en aquesta nota per a la majoria de casos. Per acabar, a causa dels valors que prenen les medianes i tercers quartils per a les assignatures, es pot dir que els alumnes del sexe masculí arriben a obtenir notes més altes en comparació amb les dones, igual que succeïa en el primer i segon curs.

	Mín.	1r Quartil	Mediana	3r Quartil	Màx.	Mitjana
			Models Matemàtics de la Tecnologia			
HOMES	5	6,5	7,6	9	10	7,678
DONES	5	6	7,0	7,5	9,5	7,03

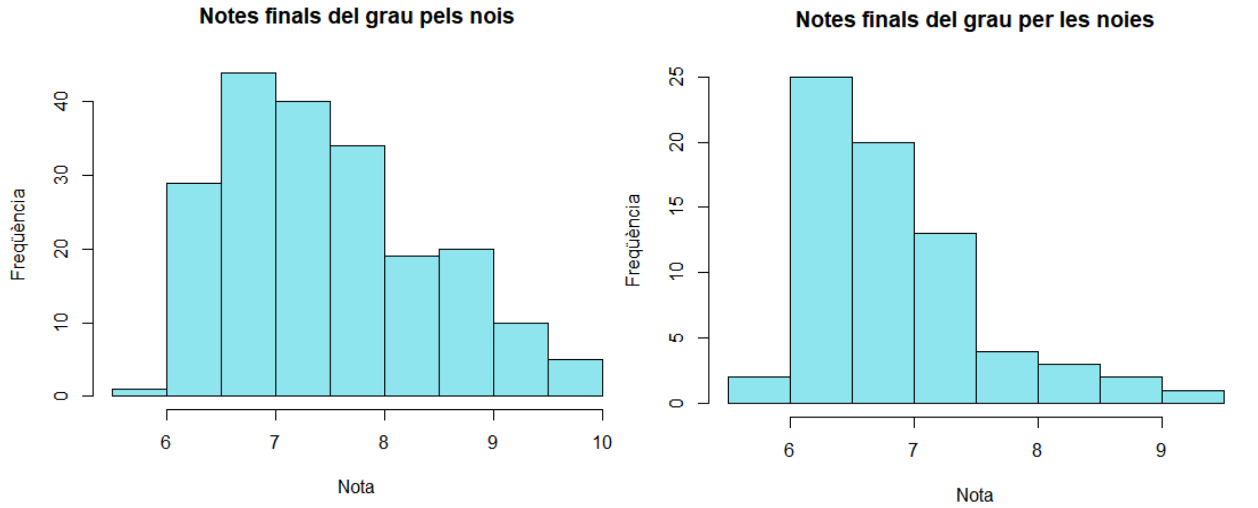
Taula 3.1.5. Principals estadístics descriptius de les variables nota de les assignatures obligatòries del quart curs del grau de Matemàtiques.

En la taula 3.1.5 també es segueixen veient les diferències entre els nois i les noies que es trobaven en els tres cursos anteriors. Tot i així, en aquesta variable no es troben grans discrepàncies pel que fa als valors del mínim, primer quartil, mediana, mitjana i màxim, en canvi, sí se'n troben en el tercer quartil, com a conseqüència es pot dir que els estudiants del sexe masculí arriben a obtenir notes més altes de manera més freqüent en comparació amb les del sexe femení.



Gràfic 3.1.6. Nombre d'assignatures que els alumnes aproven en la primera matrícula en funció del sexe.

En el gràfic 3.1.6 es pot veure clarament que no hi ha gaire diferències pel que fa al nombre d'assignatures que els alumnes aproven la primera vegada que les matriculen en funció del sexe, sinó que el més habitual tant pels nois com per les noies és aprovar les 25 assignatures obligatòries en la primera matrícula.



Gràfic 3.1.7. Distribució de la nota final del grau de Matemàtiques en funció del sexe.

En el gràfic 3.1.7 s'observa que sí existeixen diferències en la nota final del grau de Matemàtiques entre els dos sexes, segurament, a causa de les diferències que s'han vist en les notes de les assignatures obligatòries. Pel cas dels nois, el més habitual és obtenir una nota final del grau entre 6 i 8, en canvi, pel cas de les noies, aquestes acostumen a estar entre 6 i 7,5 aproximadament. A més a més, els estudiants del sexe masculí arriben de forma més freqüent a notes altes (entre 8 i 10) en comparació amb les dones que ho fan, ja que aquestes representen una petita part del total d'estudiants d'aquest sexe. Com a conseqüència es pot dir, que els homes tenen, de mitjana, una nota final més elevada que les alumnes del sexe femení.

4. Anàlisi de la nota final dels estudiants del grau de Matemàtiques

En aquest apartat s'estudiarà la nota final dels alumnes graduats que van entrar durant els anys 2009-2015 amb l'objectiu d'observar quines variables de la base de dades tenen un efecte en aquesta i si aquest varia en funció del sexe, és a dir, es vol saber si el fet de ser home o dona és un aspecte rellevant per aconseguir l'èxit en el grau de Matemàtiques.

Per assolir aquest objectiu es farà una regressió lineal usant la variable *nota final* com a independent o resposta i les altres variables de la base de dades com a explicatives.

4.1. Anàlisi bivariant amb la variable resposta

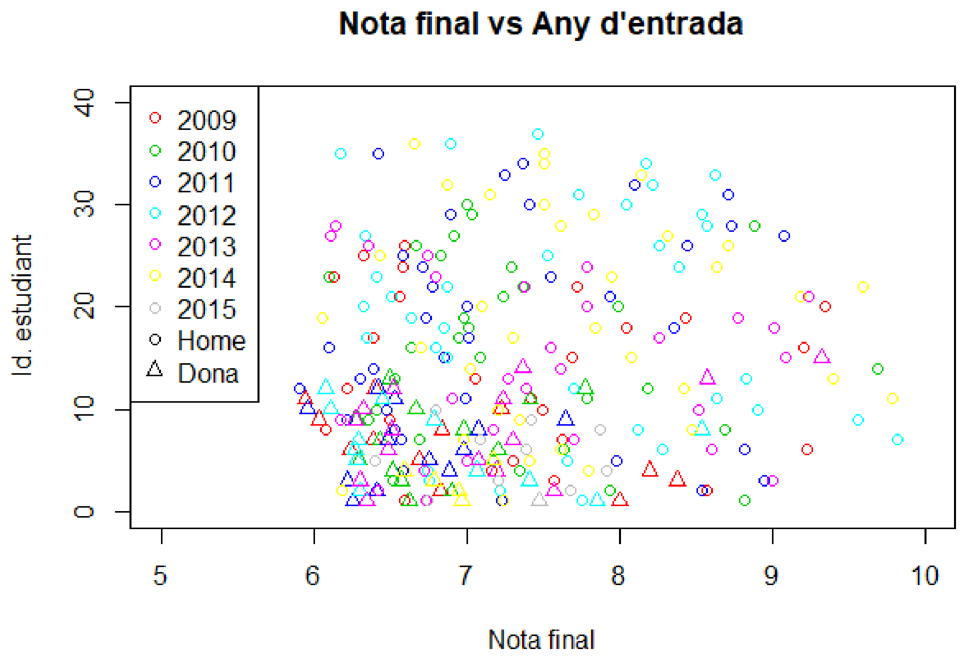
És necessari, abans de construir el model lineal, fer una anàlisi descriptiva de les variables de la base de dades amb la variable independent del model (*nota final del grau*).

L'objectiu principal de l'anàlisi descriptiva és poder observar quines variables es troben més relacionades amb la resposta i com és aquesta relació, d'aquesta manera després d'haver creat el model lineal es podrà veure si els paràmetres estimats mostren la mateixa relació que es veurà en aquesta anàlisi.

Per a dur a terme l'anàlisi es crearan diversos gràfics bivariants amb el programa *R* [3], ja que gràcies a aquests es pot observar fàcilment la relació entre dues variables.

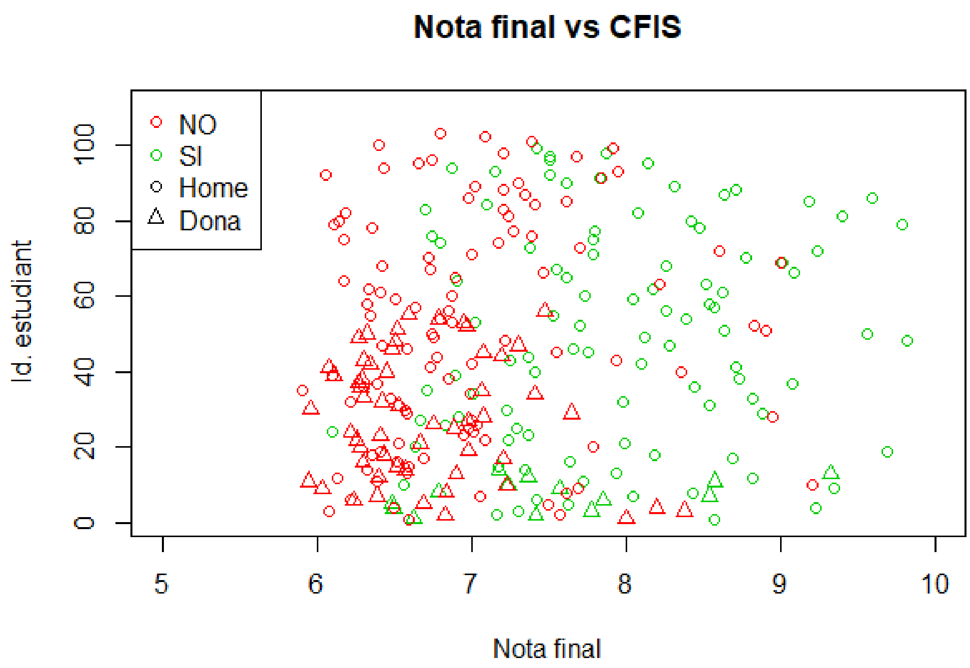
Cal destacar, que a l'hora de fer aquesta anàlisi es tindrà en compte el factor sexe, per tal de veure quines de les variables tenen una relació amb la resposta que varia en funció del fet de ser home o dona.

[3] Els codis que s'han fet servir es troben a l'annex: Codi d'R per a l'anàlisi descriptiu bivariant de les variables de la base de dades dels apartats 4. Anàlisi de la nota final dels estudiants del grau de Matemàtiques i 5. Anàlisi de la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres amb la variable nota final.



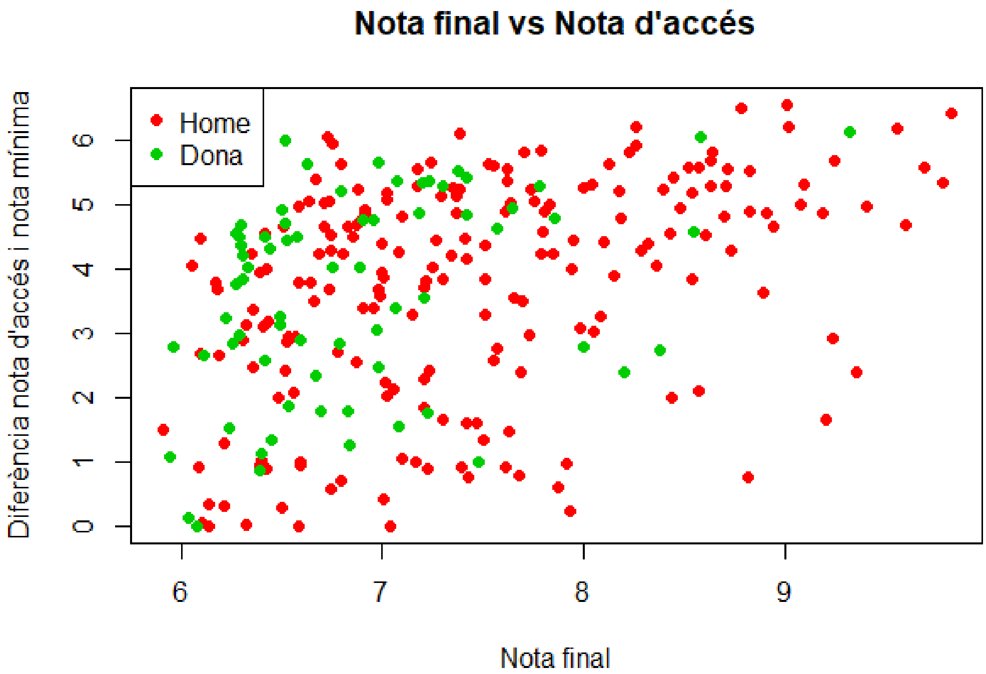
Gràfic 4.1.1. Distribució de la nota final dels alumnes del grau de Matemàtiques en funció de l'any d'entrada i el sexe.

En el gràfic 4.1.1 es pot veure que la nota final dels estudiants del grau no té una relació clara amb l'any d'entrada, sinó que els seus valors es dristribueixen de forma aleatoria (majoritàriament entre 6 i 9) per a cada un dels anys. Seguidament, pel que fa al sexe dels alumnes, tampoc es veu que afecti a la relació entre aquestes dues variables, però sí s'observa que per a gairebé tots els anys les notes de les dones s'acostumen a trobar per sota de les dels nois.



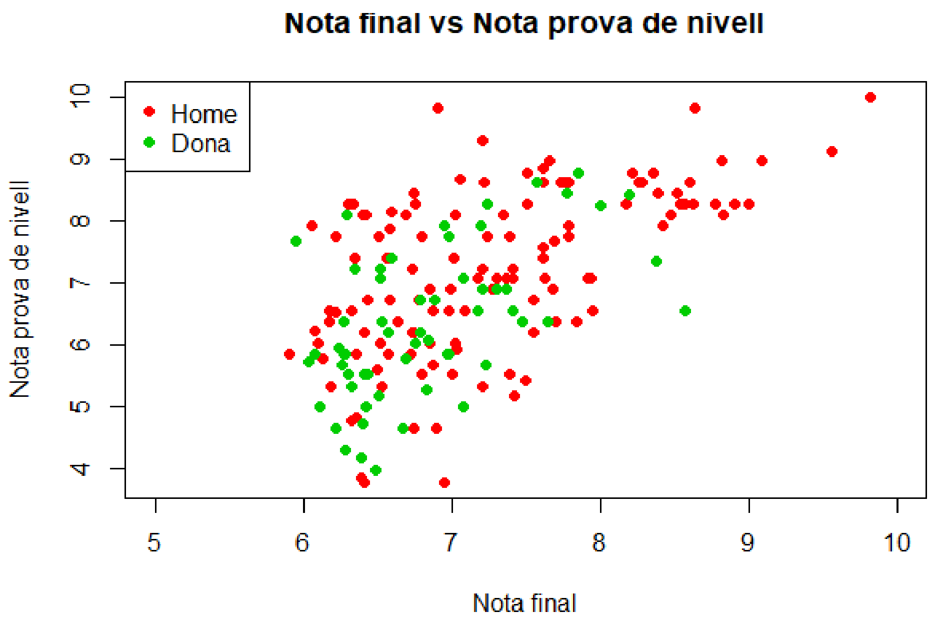
Gràfic 4.1.2. Distribució de la nota final dels alumnes del grau de Matemàtiques en funció de si pertanyen o no al CFIS i el sexe.

En el gràfic 4.1.2 s'observa que sí hi ha una relació entre pertànyer o no al CFIS i la nota final del grau de Matemàtiques. Tal i com es veu en aquest, els alumnes amb notes finals més altes són aquells que sí formen part del CFIS, en canvi, els que no hi pertanyen acostumen a tenir notes més baixes. Pel que fa al sexe, s'arriba a la mateixa conclusió tant pels nois com per les noies, és a dir, l'efecte de la variable CFIS a la nota final no depèn del valor que prengui aquest factor, ja que, pel cas de les estudiants del sexe femení, la major part de les que tenen notes més elevades pertanyen al CFIS.



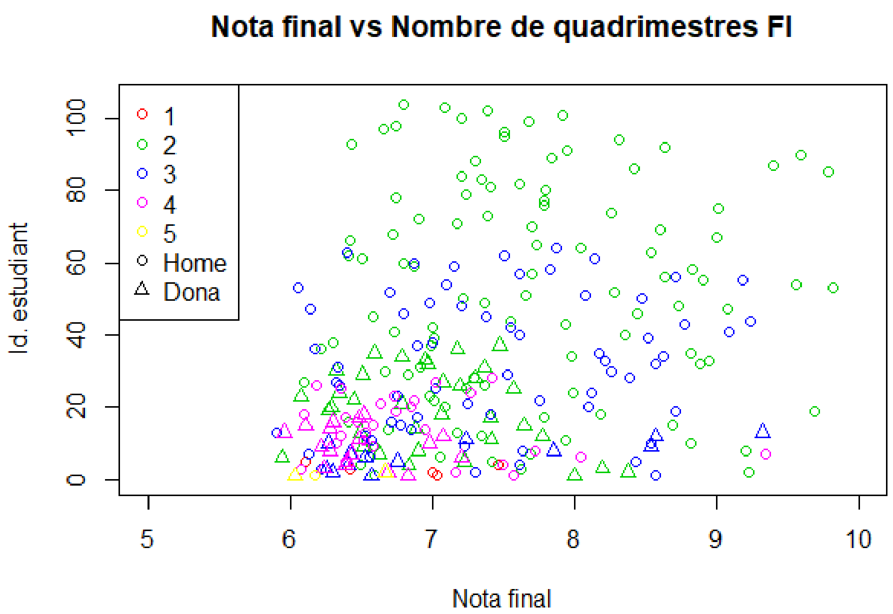
Gràfic 4.1.3. Gràfic bivariant entre la nota final i la diferència entre la nota d'accés i la nota mínima de l'any corresponent pels alumnes del grau de Matemàtiques.

En el gràfic 4.1.3 es pot veure que existeix una lleugera relació positiva entre la diferència entre la nota d'accés i la nota mínima i la nota final, és a dir, aquells alumnes que tenen notes finals elevades acostumen a accedir a la universitat amb notes altes (en relació amb la nota mínima per a l'any corresponent). Per altra banda, pel que fa al sexe s'observa que no afecta a la relació entre aquestes dues variables, ja que tant pel cas dels nois com pel de les noies aquells estudiants amb notes finals altes tenen notes d'accés també elevades. A causa d'això, es pot concloure que és molt poc habitual que els estudiants que accedeixen a la carrera amb notes baixes arribin a obtenir una nota final alta.



Gràfic 4.1.4. *Gràfic bivariant entre la nota final i la nota de la prova de nivell voluntària pels alumnes del grau de Matemàtiques.*

En el gràfic 4.1.4 es pot veure clarament que sí hi ha una relació entre la nota final i la nota de la prova de nivell voluntària que es realitza a l'inici del curs al grau de Matemàtiques. De manera més concreta, la relació existent entre aquestes dues variables és positiva, és a dir, si la nota de la prova de nivell és elevada la nota final també ho és, en canvi, es pot habitual que hi hagi alumnes amb notes finals elevades i notes a la prova de nivell baixes. Seguidament, pel que fa al sexe, aquest factor no influeix en la relació entra la nota final i la nota a la prova de nivell, sinó que s'observa la mateixa entre aquestes tant pels nois com per les noies. Finalment, cal destacar, que aquesta prova es voluntària, a causa d'això les conclusions que s'extreuen a conseqüència dels seus valors no són del tot fiables.



Gràfic 4.1.5. Distribució de la nota final dels alumnes del grau de Matemàtiques en funció del nombre de quadrimestres que van trigar en aprovar la fase inicial i el sexe.

En el gràfic 4.1.5 s'observa que sí hi ha una lleugera relació inversa entre el nombre de quadrimestres que els alumnes triguen en aprovar la fase inicial del grau i la seva nota final, és a dir, si el nombre de quadrimestres augmenta, la nota final disminueix. Tot i que, aquesta relació només es veu per a un nombre de quadrimestres major a 3. El més comú, tant pels nois com per les noies és aprovar aquesta fase en dos quadrimestres. A més a més, pel cas dels estudiants masculins, també és habitual fer-ho en 3, en aquests casos, la nota final varia, majoritàriament, entre 6 i 9.

A continuació, es farà una taula en la qual es pot observar el valor de la correlació entre la nota final del grau i la nota de les assignatures obligatòries en funció del sexe.

	Correlació
Fonaments de la Matemàtica	
HOMES	0,6099423
DONES	0,5881106
Àlgebra Lineal	
HOMES	0,7237505
DONES	0,6638744
Càlcul en una variable	
HOMES	0,6412803
DONES	0,4334937
Informàtica	
HOMES	0,6050491
DONES	0,6196410

	Correlació
Càlcul Diferencial	
HOMES	0,7570398
DONES	0,6169207
Geometria Afí i Euclidiana	
HOMES	0,7713132
DONES	0,6900299
Àlgebra Lineal i Numèrica	
HOMES	0,6576368
DONES	0,6839102
Matemàtica Discreta	
HOMES	0,7248519
DONES	0,7468511

Taula 4.1.1. Correlació entre les variables nota de les assignatures obligatòries del primer curs i la nota final del grau de Matemàtiques.

En la taula 4.1.1 es pot veure que sí existeix una relació entre les notes de les assignatures obligatòries del primer curs i la nota final del grau de Matemàtiques. Concretament, es veu que la relació entre aquestes dues variables és positiva per a tots els casos, és a dir, si la nota de les assignatures és elevada, la nota final del grau també ho acostuma a ser.

Tot i així, no s'observa que hi hagi el mateix grau de relació per a totes les assignatures i pels dos sexes, només en l'assignatura Matemàtica Discreta, hi ha una relació elevada tant pels nois com per les noies. Seguidament, pels homes Àlgebra Lineal, Càlcul Diferencial i Geometria Afí i Euclidiana són les que tenen major relació amb la nota final, és a dir, el fet de treure una nota alta en aquestes assignatures fa que sigui més possible treure una nota final elevada.

Finalment, pel que fa al sexe, per a les assignatures Fonaments de la Matemàtica, Informàtica, Àlgebra Lineal i Numèrica i Matemàtica Discreta no es veuen diferències destacables entre nois i noies. Per altra banda, a les altres les noies tenen una relació entre la nota final i la nota de les assignatures menor als nois. Com a conseqüència es pot dir que, pel cas dels estudiants del sexe masculí és més habitual que si la nota que s'ha obtingut de les assignatures obligatòries és elevada la nota final també ho sigui.

	Correlació
Càlcul Integral	
HOMES	0,6707248
DONES	0,7257218
Àlgebra Multilineal i Geometria	
HOMES	0,7224202
DONES	0,4798365
Algorísmia	
HOMES	0,6675188
DONES	0,5940980
Programació Matemàtica	
HOMES	0,7434386
DONES	0,6658044

	Correlació
Funcions de Variable Complexa	
HOMES	0,7740404
DONES	0,7330253
Topologia	
HOMES	0,7030846
DONES	0,6353674
Física	
HOMES	0,6760402
DONES	0,6152482
Anàlisi Real	
HOMES	0,7564369
DONES	0,5910188

Taula 4.1.2. Correlació entre les variables nota de les assignatures obligatòries del segon curs i la nota final del grau de Matemàtiques.

En la taula 4.1.2 s'observa que sí hi ha una relació entre les notes de les assignatures obligatòries del segon curs i la nota final del grau de Matemàtiques. Exactament, aquesta relació és positiva per a tots els casos, és a dir, si la nota de les assignatures és alta, la nota final del grau també acostuma a ser-ho.

Per altra banda, no es veu que hi hagi el mateix grau de relació pels dos sexes entre totes les assignatures i la nota final, només en el cas de Funcions de Variable Complexa es troba una relació elevada pels dos sexes. Seguidament, pel cas dels nois, hi ha una relació més destacable per a Àlgebra Multilineal i Geometria, Programació Matemàtica, Topologia i Anàlisi Real. En canvi, pel cas de les noies, es troba una major relació amb la nota final en l'assignatura Càlcul Integral. És a dir, el fet de treure una nota elevada en aquestes assignatures fa que sigui més possible treure una nota final elevada.

Per acabar, en totes les assignatures (excepte Càlcul Integral) els nois tenen una relació major entre la nota d'aquestes i la nota final, en comparació amb les noies. A causa d'això, es pot concloure que pels homes és més comú que si la nota que s'ha obtingut a les assignatures obligatòries és elevada, la nota final també ho sigui.

	Correlació
Equacions Diferencials Ordinàries	
HOMES	0,7360891
DONES	0,7004947
Estructures Algebraiques	
HOMES	0,7058827
DONES	0,6100205
Càlcul Numèric	
HOMES	0,7541300

	Correlació
Equacions en Derivades Parcials	
HOMES	0,6615483
DONES	0,6451447
Geometria Diferencial	
HOMES	0,7531051
DONES	0,6929037
Models Matemàtics de la Física	
HOMES	0,5893793

DONES	0,7320403
Teoria de la Probabilitat	
HOMES	0,7868954
DONES	0,6945797

DONES	0,5974391
Estadística	
HOMES	0,7836101
DONES	0,7263289

Taula 4.1.3. Correlació entre les variables nota de les assignatures obligatòries del tercer curs i la nota final del grau de Matemàtiques.

En la taula 4.1.3 es pot veure que sí existeix una relació entre les notes de les assignatures obligatòries del tercer curs i la nota final del grau de Matemàtiques. Concretament, es troba una relació positiva en tots els casos, és a dir, si la nota de les assignatures és elevada, és habitual que la nota final del grau també ho sigui.

Tot i així, no s'observa que hi hagi el mateix grau de relació en totes les assignatures pels dos sexes. En aquest cas, per a Equacions Diferencials Ordinàries, Càlcul Numèric i Estadística es troba una relació elevada tant pels nois com per les noies. Per altra banda, en Estructures Algebraiques, Teoria de la Probabilitat i Geometria Diferencial només hi ha una relació destacable pel cas dels alumnes del sexe masculí, és a dir, treure una nota alta en aquestes assignatures fa que sigui més habitual aconseguir una nota final també elevada.

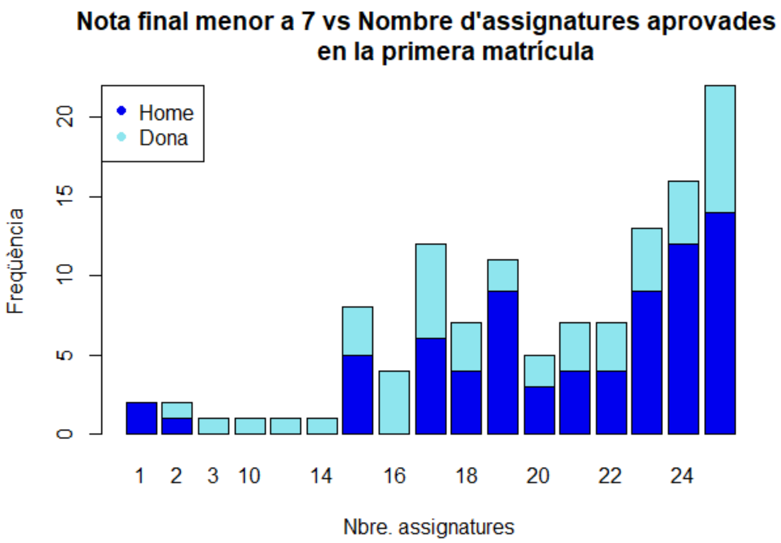
Finalment, en les assignatures Equacions Diferencials Ordinàries, Càlcul Numèric, Equacions en Derivades Parcials, Models Matemàtics de la Física no es veuen grans diferències entre nois i noies. En canvi, a les altres les dones tenen una relació entre la nota de les assignatures i la nota final menor als homes. Com a conseqüència, es pot dir que en el cas dels alumnes del sexe masculí, si la nota de les assignatures obligatòries és elevada la nota final també ho acostuma a ser.

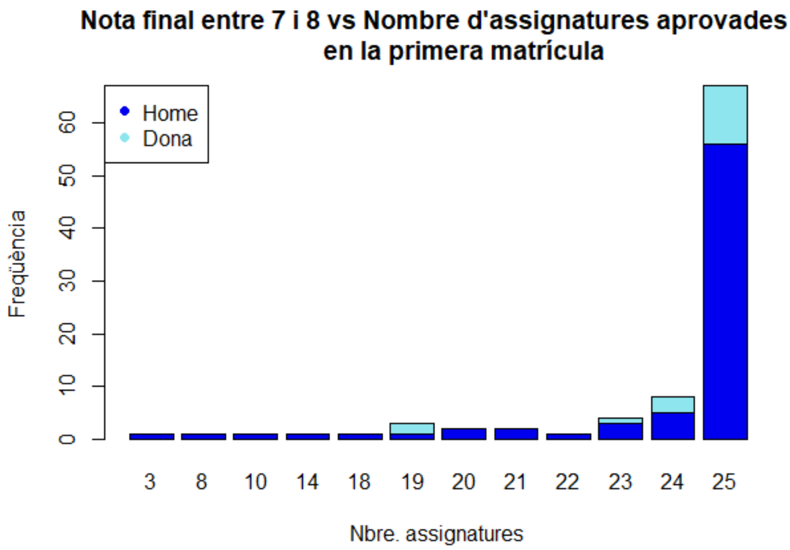
	Correlació
Models Matemàtics de la Tecnologia	
HOMES	0,4770393
DONES	0,4058938

Taula 4.1.4. Correlació entre les variables nota de les assignatures obligatòries del quart curs i la nota final del grau de Matemàtiques.

En la taula 4.1.4 s'observa que no hi ha gaire relació entre la nota de l'assignatura obligatòria del quart curs i la nota final del grau. Tot i així, aquesta relació és positiva, és a dir, si la nota d'aquesta assignatura és alta, és més possible que la nota final també ho sigui.

Pel que fa al sexe, en els dos casos la relació que es troba no és gaire destacable. Però, pel cas dels nois aquesta és lleugerament més elevada, com a conseqüència es pot dir que si aquests tenen una nota en aquesta assignatura alta la nota final també acostuma a ser-ho de forma més habitual que per a les noies.





Gràfic 4.1.6. Distribució del nombre d'assignatures aprovades en la primera matricula dels alumnes del grau de Matemàtiques en funció de la nota final i el sexe.

En el gràfic 4.1.6 es pot veure que sí existeix una relació entre el nombre d'assignatures aprovades en la primera matricula i la nota final. En aquest cas, la relació és positiva, és a dir, si la nota final augmenta el nombre d'assignatures aprovades en la primera matricula també tendeix a fer-ho, ja que tal i com s'observa en el gràfic, per a notes finals menors 7 hi ha molta variabilitat pel que fa al nombre d'assignatures, en canvi, per a notes majors a 7, es troba que la major part dels estudiants aproven les 25 assignatures obligatòries a la primera matrícula.

Per altra banda, cal destacar que aquesta relació no depèn del sexe, sinó que s'observa la mateixa relació entre les variables tant pels nois com per les noies.

A causa d'això, es pot concloure que, pels dos sexes és poc comú que si s'han aprovat menys de 25 assignatures a la primera matricula la nota final sigui superior a 7.

Nota final major a 8 vs Nombre d'assignatures aprovades en la primera matricula		
	Assignatures	**Freqüència**
HOMES	25	54
DONES	25	6

Taula 4.1.5. Distribució del nombre d'assignatures aprovades en la primera matricula dels alumnes del grau de Matemàtiques en funció la nota final i el sexe.

4.2. Model Lineal

Després de veure com és la distribució de les variables de la base de dades que s'utilitzaran com a possibles regresores i la relació entre aquestes i la resposta (la nota final), es crearan diversos models lineals [4] per tal de trobar quin és el que millor explica la variabilitat de la variable independent. De forma més concreta, el model ajudarà a conèixer quins factors influeixen més en la nota final dels estudiats del grau de Matemàtiques.

Per a la creació dels models s'usarà la funció *lm* del paquet *stats* del programa *R* [5] especificant quina serà la resposta del model i quines seran les variables que s'inclouran com a explicatives. A continuació, s'obtindrà com a resultat de la funció anterior els valors dels paràmetres del model i a més a més, també altres indicadors que seran molt útils per a valorar l'ajust d'aquest a les dades reals.

Per començar, a causa de que es té una gran quantitat de dades referent a les notes de les assignatures obligatòries i la mitjana d'aquestes defineix la nota final (majoritàriament), es crearà un model sense tenir-les en compte i seguidament, s'afegiran algunes d'elles per a poder millorar l'ajust del model final. Per a poder crear aquest primer model, es partirà d'un en el s'inclouran totes les variables de les quals es disposa, excepte la nota de la prova de nivell voluntària, ja que és una variable de la qual es troben molts valors faltants, i les interaccions entre les variables numèriques i els factors sexe i CFIS, per tal de veure si alguna de les regresores té un efecte en la resposta que depèn del fet de ser home o dona, o de pertànyer o no al CFIS.

A continuació, per a crear els models successius s'aniran eliminat del primer aquelles interaccions o variables que no siguin significatives (és a dir, que es poden considerar estadísticament iguals a 0). Finalment, l'objectiu en l'últim model és que totes les variables incloses siguin significatives, és a dir, que no s'hagin inclòs variables irrellevants que no ajuden a explicar la variabilitat de la variable independent.

Un cop s'hagi trobat el millor model sense les variables que indiquen les notes de les assignatures obligatòries, es crearà un altre model només amb aquestes com a explicatives i la nota final com a resposta. Després, usant la funció *step* d'*R,* a la qual se li especificarà que parteixi del model nul i vagi afegint les variables més rellevants referents a les notes de les assignatures, es seleccionaran les cinc assignatures que tinguin més influència en la nota final, i per acabar, s'afegiran aquestes al primer model trobat. Cal esmentar, que la funció *step* parteix d'un model inicial i afegeix o elimina variables amb l'objectiu d'obtenir un valor per a l'indicador AIC millor, és a dir, més baix (ja que són preferibles models amb valors per a l'AIC menors).

[4] Les bases d'aquest tipus de models es poden veure en l'apartat: 2.1 Models Lineals

[5] A l'annex es troben tots els codis usats: Codi d'R per a la creació del model lineal de l'apartat 4. Anàlisi de la nota final dels estudiants del grau de Matemàtiques.

En la taula següent es troben tots els models lineals creats sense les variables referents a les notes de les assignatures. En tots ells la variable resposta és la mateixa, és a dir, la nota final dels alumnes.

Model	Variables explicatives
MODEL 1	CFIS, diferència entre la nota d'accés i la nota mínima de l'any corresponent, diferència entre la nota d'accés i la nota mínima de l'any corresponent*CFIS, diferència entre la nota d'accés i la nota mínima de l'any corresponent*sexe, sexe, nombre de quadrimestres per a aprovar FI, nombre de quadrimestres per a aprovar FI*CFIS, nombre de quadrimestres per a aprovar FI*sexe, nombre d'assignatures aprovades a la primera, nombre d'assignatures aprovades a la primera*CFIS, nombre d'assignatures aprovades a la primera*sexe [6]
MODEL 2	Mateixes variables que el MODEL 1 excepte: diferència entre la nota d'accés i la nota mínima de l'any corresponent*CFIS
MODEL 3	Mateixes variables que el MODEL 2 excepte: nombre d'assignatures aprovades a la primera*sexe
MODEL 4	Mateixes variables que el MODEL 3 excepte: nombre de quadrimestres per a aprovar FI*sexe
MODEL 5	Mateixes variables que el MODEL 4 excepte: nombre de quadrimestres per a aprovar FI*CFIS

Taula 4.2.1. Models lineals creats per a explicar la variabilitat de la nota final dels estudiants del grau de Matemàtiques (excloent les variables referents a les notes de les assignatures obligatòries).

Un cop creats tots els models es necessari dur a terme el test ANOVA per a poder-los comparar dos a dos i decidir quin és l'escollit. Per a fer-ho s'ha usat la funció *anova* del paquet *stats* d'*R* [7].

En la taula mostrada a continuació es troben els p-valors dels diferents contrasts. Cal tenir en compte que s'utilitzarà un nivell de significació del 5%, com a conseqüència, un p-valor major a 0.05 indicarà que els models comparats són equivalents, per tant, no es pot rebutjar la hipòtesi nul·la del test.

[6] El signe "*" indica la interacció entre dues variables.

[7] Els codis utilitzats es poden observar a l'annex: Codi d'R per a la creació del model lineal de l'apartat 3. Anàlisi de la nota final dels estudiants del grau de Matemàtiques.

Models comparats	P-valor del test ANOVA
MODEL 1 VS MODEL 2	0,8617
MODEL 2 VS MODEL 3	0,3943
MODEL 3 VS MODEL 4	0,1912
MODEL 4 VS MODEL 5	0,08636

Taula 4.2.2. P-valors dels test ANOVA per a comparar dos models lineals encaixats.

A la taula 4.2.2 es pot veure que a causa de que els 5 models creats són equivalents s'escollirà com a millor el més simple, és a dir, el model 5.

Abans de validar el model triat és important comprovar que no existeix col·linealitat entre les variables explicatives. Per a fer-ho es calcularà el VIF per a les variables que formen part del model usant la funció *vif* del paquet *car* d'*R*. Si aquest pren un valor superior a 10 es considera que hi ha problemes de col·linealitat i per tant, la variabilitat que explica la variable en concret ja es troba explicada per les altres variables del model. Com a conseqüència, per a reduir la col·linealitat s'eliminarà una de les variables que es troba correlacionada amb les altres.

	VIF
Diferència entre la nota d'accés i la nota mínima de l'any corresponent	1,653967
Nombre de quadrimestres per a aprovar FI	1,075426
Nombre d'assignatures aprovades a la primera	1,552079
CFIS	500,72687
Sexe	6,738573
Nombre d'assignatures aprovades a la primera*CFIS	504,88513
Diferència entre la nota d'accés i la nota mínima de l'any corresponent*sexe	6,760020

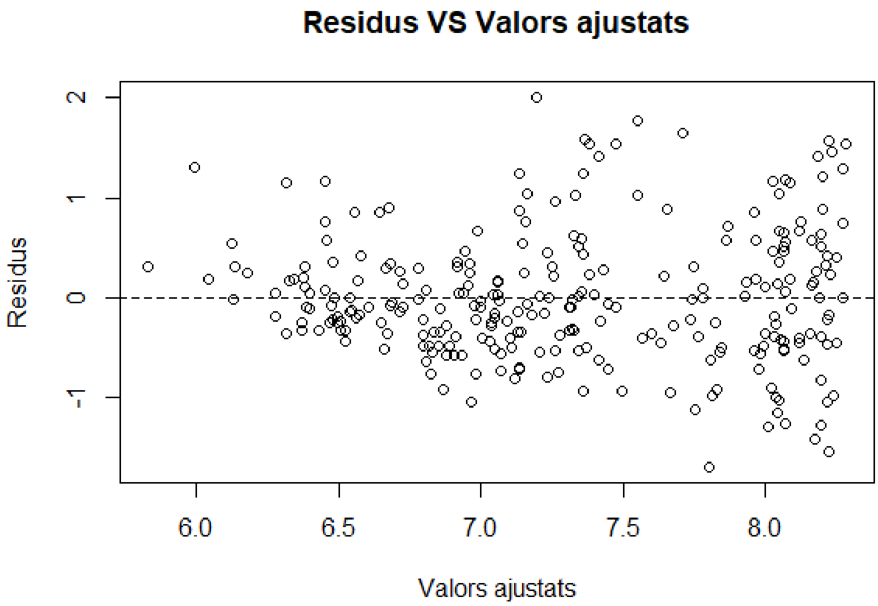
Taula 4.2.3. VIF per a les variables explicatives del model 5.

Tal i com s'observa a la taula 4.2.3 les variables numèriques principals no pateixen col·linealitat. Pel que fa a la interacció entre el nombre d'assignatures aprovades a la primera i CFIS i el factor CFIS, i la interacció entre la diferència entre la nota d'accés i la nota mínima i sexe i el factor sexe, tenen un elevat grau de multicol·linealitat, tot i així, al ser una variable categòrica i la interacció entre aquesta i una numèrica és habitual trobar aquests valors tan elevats.

Un cop s'ha triat el millor model, en aquest cas el model 5, cal validar-lo. Per a fer-ho es crearan dos gràfics, i a més a més es farà el test de *Shapiro-Wilks*, per a comprovar la hipòtesi de normalitat dels residus[8].

[8] A l'annex es poden observar els codis usats: Codi d'R per a la creació del model lineal de l'apartat 4. Anàlisi de la nota final dels estudiants del grau de Matemàtiques.

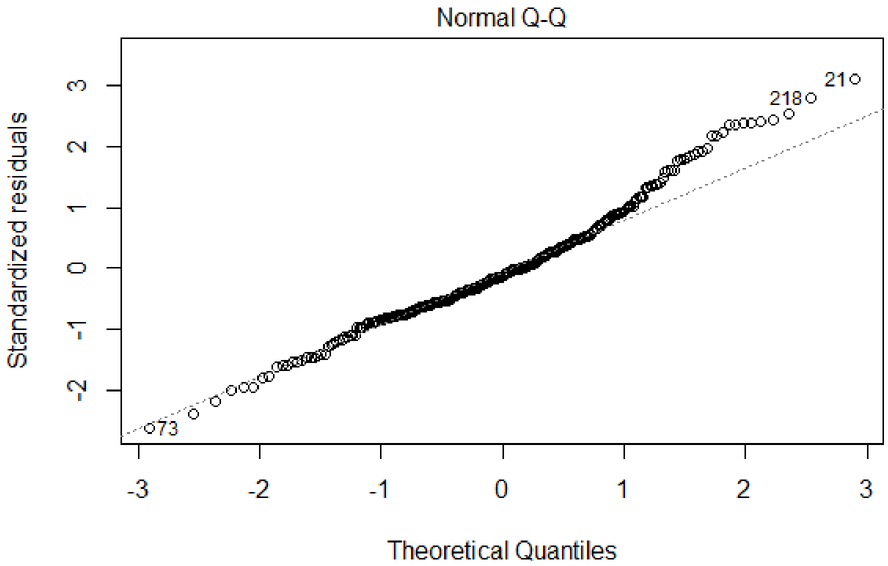
Per començar, en el següent gràfic es representaran els residus en front dels valors ajustats ($\hat{y}_i$). En aquesta representació cal observar que per una banda, els valors es situen al voltant de l'abscissa igual a 0, d'aquesta manera es complirà la hipòtesi $E(\varepsilon_i) = 0$, i per altra banda, que aquests es troben situats aleatòriament i de forma més o menys homogènia per a tots els valors ajustats de la variable resposta al voltant de l'abscissa igual a 0, per tant, es complirà la hipòtesi $Var(\varepsilon_i) = \sigma^2$ (homoscedasticitat).



Gràfic 4.2.1. Representació dels residus vs Valors ajustats pel model 5.

En el gràfic 4.2.1 es pot veure que els valors dels residus sí es situen al voltant de 0, per tant, es valida la hipòtesi $E(\varepsilon_i) = 0$, però en canvi, la seva variabilitat no és constant, ja que aquesta augmenta a mesura que els valors ajustats són majors, com a conseqüència es pot dir que la hipòtesi $Var(\varepsilon_i) = \sigma^2$ no es compleix, és a dir, hi ha heteroscedasticitat.

Per altra banda, pel que fa a la normalitat, a continuació es troba el gràfic *qq-plot* pels residus representat usant el programa *R* i el p-valor resultant d'aplicar el test de normalitat de *Shapiro-Wilks,* gràcies a la funció *shapiro.test* del paquet *stats* d'*R*.



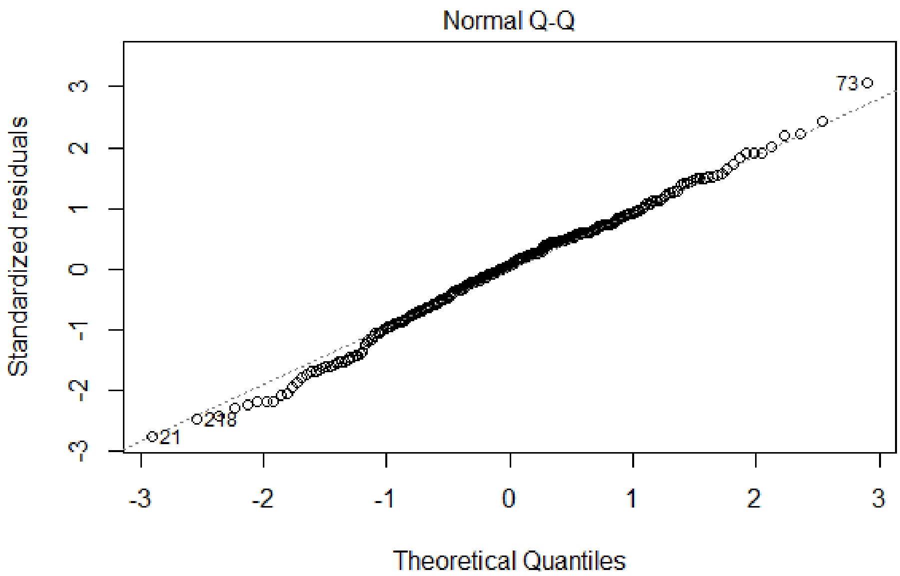
P-valor del test de normalitat de *Shapiro-Wilks*
0,0003369

Gràfic 4.2.2. Gràfic qq-plot i p-valor pel test de normalitat pels residus del model triat.

Gràcies al gràfic 4.2.2 es pot conclure que els residus no compleixen la hipòtesi de normalitat, a més a més s'obté un p-valor pel test de *Shapiro-Wilks* menor a 0.05, per tant, amb un nivell de significació del 5% es rebutja la hipòtesi nul·la, és a dir, en aquest cas els residus del model no segueixen una distribució normal.

Per intentar solucionar aquests problemes es transformarà la variable resposta aplicant la inversa, és a dir, es crearà un model lineal amb les mateixes variables explicatives que el model triat però en aquest cas, la resposta serà el la inversa de la nota final dels alumnes.

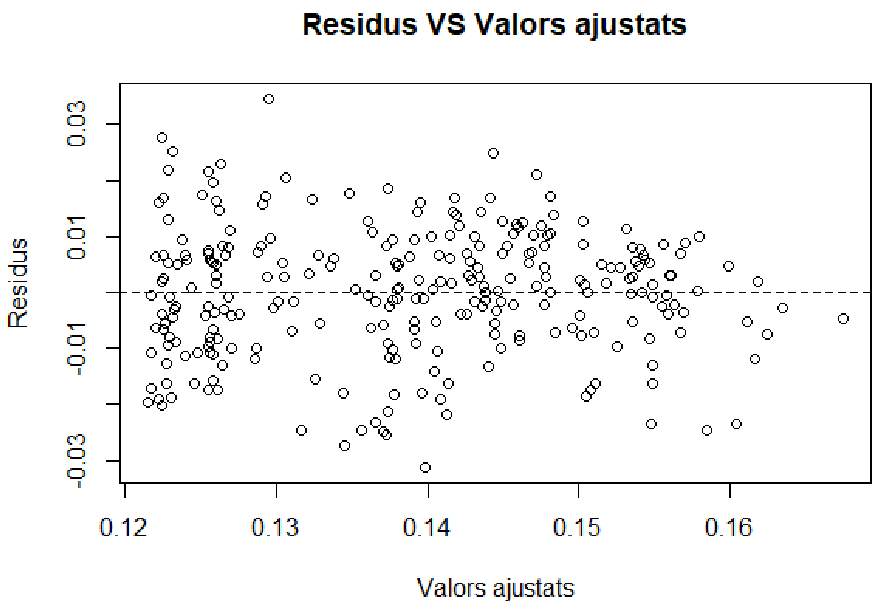
En el gràfic que es mostra a continuació es pot veure que gràcies a la transformació realitzada, els residus del nou model sí compleixen la hipòtesi de normalitat. A més a més, el p-valor del test de *Shapiro-Wilks* és major al 5% (nivell de significació).



P-valor del test de normalitat de *Shapiro-Wilks*
0,334

Gràfic 4.2.3. Gràfic qq-plot i p-valor pel test de normalitat pels residus del model triat en el qual la resposta s'ha transformat fent la inversa.

Per altra banda, per a poder comprovar si en el nou model es verifiquen les dues hipòtesis següents: $E(\varepsilon_i) = 0$ i $Var(\varepsilon_i) = \sigma^2$ s'ha tornat a representar en un gràfic els residus en front dels valors ajustats.



Gràfic 4.2.4. Representació dels residus vs Valors ajustats del model triat en el qual la resposta s'ha transformat fent la inversa.

En el gràfic 4.2.4 s'observa que tot i que s'ha transformat el model la hipòtesi de variància constant no es compleix a causa de que, igual que succeïa amb el model sense transformació hi ha heteroscedasticitat, en aquest cas la variabilitat dels errors disminueix a mesura que els valors ajustats són més elevats, en canvi, l'altre hipòtesi ($E(\varepsilon_i) = 0$) sí que es verifica, ja que els valors dels residus es troben situats al voltant de 0.

Com a conseqüència, es pot concloure que aquest model no queda validat a causa de que no compleix les hipòtesi dels model lineals. Tot i així, s'analitzaran els valors dels seus coeficients.

	Valor estimat	Error estàndard	Valor t	Pr(>$\|t\|$)
Intercept	0,1595861	0,0048524	32,888	< 2e-16
Diferència entre la nota d'accés i la nota mínima	-0,0008610	0,0005304	-1,623	0,105706
Sexe: Dona	-0,0033412	0,0041120	-0,813	0,417206
CFIS: SÍ	0,1181864	0,0314454	3,758	0,000210
Nombre de quadrimestres en aprovar la fase inicial	0,0033106	0,0008756	3,781	0,000193
Nombre d'assignatures aprovades a la primera matrícula	-0,0009963	0,0001893	-5,262	2,95e-07
Diferència entre la nota d'accés i la nota mínima*Sexe	0,0020819	0,0010128	2,056	0,040812
Nombre d'assignatures aprovades a la primera matrícula*CFIS	-0,0052980	0,0012780	-4,145	4,57e-05
Significació global del model	Estadístic F: 39,7 (7 i 264 graus de llibertat), p-valor: < 2.2e-16, $R^2 ajustat$: 0,4999			

Taula 4.2.4. Valors, significació individual pels coeficients dels paràmetres i significació global del model.

Com es pot veure a la taula 4.2.4 la regressió és globalment significativa, ja que el p-valor pren un valor menor al nivell de significació (5%), com a conseqüència almenys un dels seus coeficients és estadísticament diferent de 0. En aquest cas, tots els coeficients del model excepte el de les variables sexe i la diferència entre la nota d'accés i la nota mínima, són estadísticament diferent de 0, ja que els p-valors dels contrasts de significació individual són menors a 0,05.

Per començar, cal esmentar que en el model s'ha realitzat una transformació en la resposta com a conseqüència no es pot considerar que els coeficients de les variables indiquin el valor en el qual variarà la nota final, sinó la inversa d'aquesta, a més a més, cal tenir en compte que el signe positiu indica que un augment en la variable corresponent provocarà una disminució en la resposta, en canvi, pel cas del signe negatiu succeeix el contrari. Per començar, pel que fa a la variable que indica el nombre de quadrimestres en aprovar la fase inicial, es pot dir, que si aquest augmenta disminueix el valor de la nota final. Per altra banda, si el nombre d'assignatures aprovades en la primera matrícula és major augmenta la nota final. Per acabar, pel que fa referència a la interacció entre la diferència entre la nota d'accés i la nota mínima i el factor sexe, indica que pel cas de les dones accedir a la carrera amb una diferència entre aquestes dues notes més elevada provoca una disminució en la nota final en comparació amb la dels nois, i pel que fa al factor CFIS, cal esmentar que interacciona amb el nombre d'assignatures aprovades a la primera matrícula, per tant, en funció d'aquesta variable tindrà un efecte o un altre sobre la nota final, ja que per una banda, pertànyer al CFIS fa disminuir la nota final, però, pertànyer i a més a més aprovar més assignatures a la primera fa que aquesta augmenti. Finalment, cal destacar que el coeficient $R^2 ajustat$ pren un valor baix, és a dir, només s'aconsegueix explicar un 50% aproximadament de la variabilitat total de la resposta mitjançant el model.

A continuació es crearan altres models en els quals s'inclouran a més de les variables que s'han inclòs en els models anteriors, aquelles que fan referència a les cinc assignatures obligatòries que tenen més influència en la nota final. Per a seleccionar-les s'ha aplicat la funció *step* del paquet *stats* d'*R* [9], a la qual se li ha especificat que parteixi del model nul i afegeixi o elimini les variables que indiquen les notes de les assignatures obligatòries de forma que s'obtingui un millor valor de l'indicador AIC. Un cop s'ha executat aquesta funció s'han seleccionat les primeres cinc assignatures que aquesta afegeix al model, i que per tant, són les més rellevants. Aquestes són:

- Teoria de la probabilitat
- Càlcul numèric
- Programació matemàtica
- Estructures algebraiques
- Estadística

Cal esmentar que a l'hora d'incloure la informació referent a les variables que indiquen les notes de les assignatures obligatòries, també existeix un altre mètode molt útil que s'anomena *Anàlisi de Components Principals*. L'objectiu d'aquest és reduir el nombre de variables explicatives mitjançant combinacions lineals de les mateixes variables que prenen el nom de *Components Principals*. Tot i així, aquest mètode no ha estat escollit, ja que com s'ha dit es basa en fer una combinació de les variables, com a conseqüència inclou informació de totes aquestes i la nota final es calcula en major part, a partir de les notes d'aquestes assignatures, a causa d'això si s'incloïa en el model algun dels components principals calculats que més variabilitat explicaven, no es considerava com a significatives a altres regresores, en canvi, si s'incloïen components que explicaven menys percentatge de variabilitat s'obtenia un valor per a l'indicador $R^2\ ajustat$ molt baix, és a dir, el model explicava una part reduïda de la variabilitat total de la resposta. Per altra banda, mitjançant el mètode de selecció de variables *step* s'ha aconseguit explicar gran part de la variabilitat de la nota final amb un nombre reduït de variables explicatives.

Tot i així, s'ha de comentar que les assignatures escollides formen part de les que es troben més correlacionades amb la primera i segona dimensió, aquestes expliquen aproximadament un 55% de la variabilitat total. Concretament, la primera separa aquells alumnes que tenen més nota dels que en tenen menys, en canvi, la segona crea dos grups d'assignatures, és a dir, els estudiants que tenen més nota en les que pertanyen a un grup acostumen a tenir-ne menys en les de l'altre, en aquest cas, de les assignatures escollides, els que tenen major nota en càlcul numèric i estadística, acostumen a tenir-ne menys en teoria de la probabilitat, programació matemàtica i estructures algebraiques.

A continuació, s'ha creat un model en el qual s'inclouen les variables de la base de dades que no fan referència a les notes de les assignatures (excepte la nota de la prova de nivell voluntària) i les notes de les cinc assignatures triades, a més a més també s'afegiran les interaccions entre les variables numèriques i els factors sexe i CFIS.

[9] Els codis utilitzats es poden trobar a l'annex: Codi d'R per a la creació del model lineal de l'apartat 4. Anàlisi de la nota final dels estudiants del grau de Matemàtiques.

Seguidament, s'aniran eliminant del primer model totes les variables o interaccions que no siguin significatives (és a dir, que es poden considerar estadísticament iguals a 0). L'objectiu final és aconseguir que en l'últim model totes les variables incloses siguin significatives.

En la taula següent es mostren dos dels models lineals creats. Cal esmentar, que la variable resposta en tots ells és la mateixa, és a dir, la nota final dels estudiants. A causa del gran nombre de variables només s'especificaran aquelles que inclou el primer model i l'últim, pel que fa als altres s'han creat eliminant del primer les variables o interaccions no significatives una a una [10].

Model	Variables explicatives
MODEL 1	CFIS, diferència entre la nota d'accés i la nota mínima de l'any corresponent, diferència entre la nota d'accés i la nota mínima de l'any corresponent*CFIS, diferència entre la nota d'accés i la nota mínima de l'any corresponent*sexe, sexe, nombre de quadrimestres per a aprovar FI, nombre de quadrimestres per a aprovar FI*CFIS, nombre de quadrimestres per a aprovar FI*sexe, nombre d'assignatures aprovades a la primera, nombre d'assignatures aprovades a la primera*CFIS, nombre d'assignatures aprovades a la primera*sexe, nota de càlcul numèric, nota de càlcul numèric*CFIS, nota de càlcul numèric*sexe, nota de teoria de la probabilitat, nota de teoria de la probabilitat*CFIS, nota de teoria de la probabilitat*sexe, nota de programació matemàtica, nota de programació matemàtica*CFIS, nota de programació matemàtica*sexe, nota de estructures algebraiques, nota de estructures algebraiques*CFIS, nota de estructures algebraiques*sexe, nota d'estadística, nota d'estadística*CFIS, nota d'estadística*sexe [11]
...	...
MODEL 19	Inclou les següents variables: CFIS, diferència entre la nota d'accés i la nota mínima de l'any corresponent, sexe, nota de càlcul numèric, nota de teoria de la probabilitat, nota de programació matemàtica, nota de estructures algebraiques, nota de estructures algebraiques*CFIS, nota d'estadística

Taula 4.2.5. Models lineals creats per a explicar la variabilitat de la nota final dels alumnes del grau de Matemàtiques incloent les notes de les assignatures obligatòries.

Un cop s'han creat tots els models es necessari dur a terme un test ANOVA per a poder-los comparar dos a dos i decidir quin és l'escollit. Per a realitzar-ho s'ha usat la funció *anova* del paquet *stats* d'*R* [12]. En la taula que es mostra a continuació es poden veure els p-valors dels contrasts. S'ha de tenir en compte que s'usarà un nivell de significació dels 5%, per tant, si el p-valor pren un valor major a 0.05 indicarà que els models comparats són equivalents, és a dir, no es pot rebutjar la hipòtesi nul·la del test.

[10] A l'annex es poden veure tots els models creats amb R: Codi d'R per a la creació del model lineal de l'apartat 4. Anàlisi de la nota final dels estudiants del grau de Matemàtiques.

[11] El signe "*" indica la interacció entre dues variables.

[12] A l'annex es poden veure els codis usats: Codi d'R per a la creació del model lineal de l'apartat 4. Anàlisi de la nota final dels estudiants del grau de Matemàtiques.

Models comparats	P-valor del test ANOVA
MODEL 1 VS MODEL 2	0,9977
MODEL 2 VS MODEL 3	0,9883
MODEL 3 VS MODEL 4	0,973
MODEL 4 VS MODEL 5	0,8956
MODEL 5 VS MODEL 6	0,885
MODEL 6 VS MODEL 7	0,7199
MODEL 6 VS MODEL 8	0,781
MODEL 8 VS MODEL 9	0,7827
MODEL 9 VS MODEL 10	0,7114
MODEL 10 VS MODEL 11	0,7361
MODEL 11 VS MODEL 12	0,6052
MODEL 12 VS MODEL 13	0,6394
MODEL 13 VS MODEL 14	0,5539
MODEL 14 VS MODEL 15	0,4229
MODEL 15 VS MODEL 16	0,1711
MODEL 16 VS MODEL 17	0,07672
MODEL 17 VS MODEL 18	0,0724
MODEL 18 VS MODEL 19	0,1483

Taula 4.2.6. P-valors dels test ANOVA per a comparar dos models lineals encaixats en els quals s'han inclòs les notes de les assignatures obligatòries.

A causa dels p-valors de la taula 4.2.6 es pot concloure que tots els models creats són equivalents, per tant, s'escollirà com a millor el més simple, és a dir, el model 19.

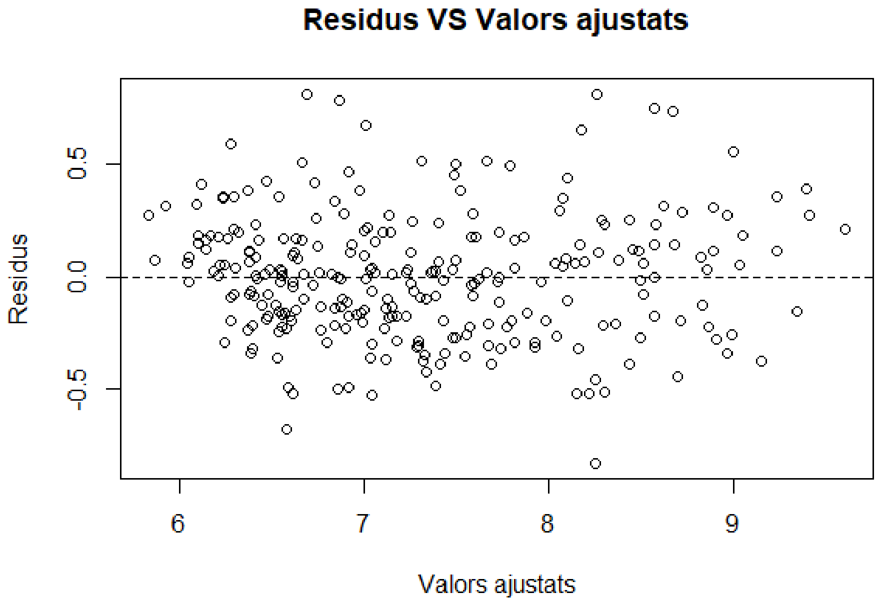
Abans de validar el model triat cal comprovar que no existeix col·linealitat entre les variables regresores. Per a fer-ho es calcularan els VIF de les variables que formen part del model utilitzant la funció *vif* del paquet *car* d'*R*.

	VIF
Diferència entre la nota d'accés i la nota mínima de l'any corresponent	1,320417
CFIS	24,468238
Sexe	1,135379
Nota de càlcul numèric	2,121292
Nota de teoria de la probabilitat	2,192337
Nota de programació matemàtica	1,806659
Nota d'estadística	2,123139
Nota d'estructures algebraiques	3,196326
Nota d'estructures algebraiques*CFIS	28,135521

Taula 4.2.7. VIF per a les variables explicatives del model 19.

Tal i com es pot veure a la taula 4.2.7 cap de les variables numèriques principals pateix col·linealitat, només es troba que la interacció entre la nota de l'assignatura estructures algebraiques i CFIS i el factor CFIS tenen un elevat grau de multicol·linealitat, tot i així, a causa de que fa referència a una variable categòrica és habitual que apareguin aquests valors.

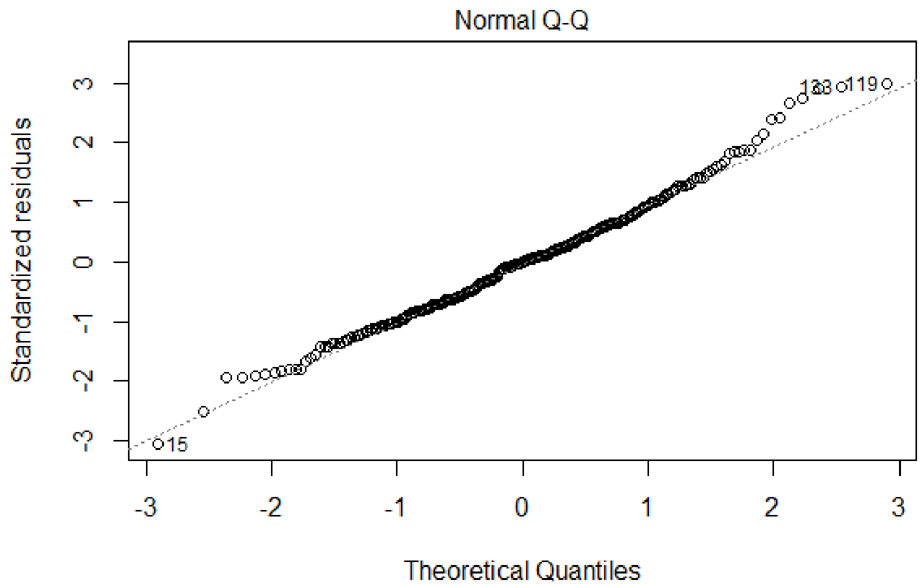
A continuació, després de triar el millor model, en aquest cas el model 19, cal validar-lo. Per a fer-ho es crearan dos gràfics i també es farà el test de normalitat de *Shapiro-Wilks* per a comprovar que els residus segueixen una distribució normal.



Gràfic 4.2.5. Representació dels residus vs Valors ajustats pel model 19.

En el gràfic 4.2.5 s'observa que els valors dels residus sí es distribueixen de forma aleatòria i més o menys constant al voltant de l'abscissa igual a 0, per tant, es validen les dues hipòtesi següents: $E(\varepsilon_i) = 0$ *i* $Var(\varepsilon_i) = \sigma^2$ (homoscedasticitat).

Per altra banda, pel que fa a la normalitat, en el gràfic següent es troba representat el *qq-plot* pels residus i el p-valor resultant d'aplicar el test de normalitat de *Shapiro-Wilks* amb el programa *R* [13].



[13] Els codis aplicats es poden veure a l'annex: Codi d'R per a la creació del model lineal de l'apartat 4. Anàlisi de la nota final dels estudiants del grau de Matemàtiques.

P-valor del test de normalitat de *Shapiro-Wilks*
0,06109

Gràfic 4.2.6. Gràfic qq-plot i p-valor pel test de normalitat pels residus del model 19.

Gràcies al gràfic 4.2.6 i al p-valor obtingut del test de normalitat, es pot concloure que els residus del model triat sí segueixen una distribució normal amb un nivell de significació del 5%, és a dir, a causa de que el p-valor pren un valor major a 0.05 no es pot rebutjar la hipòtesi nul·la de normalitat.

Com a conseqüència dels resultats anteriors, el model triat com a millor queda validat, és a dir, compleix les hipòtesi dels models lineals. A continuació, es calcularà l'indicador $R^2\ ajustat$ per a mesurar l'ajust del model a les dades. Aquest indica l'efectivitat que tenen les variables explicatives per a explicar la variabilitat de la variable independent. Per tant, es considerarà que l'ajust a les dades és millor si aquest coeficient és elevat, s'ha de destacar que l'$R^2\ ajustat$ pot prendre com a màxim el valor 1, a causa d'això com més proper es trobi a 1 millor serà considerat el model.

$\boldsymbol{R^2\ ajustat}$
0,9041334

Taula 4.2.8. Valor del coeficient $R^2\ ajustat$ pel model escollit.

En aquest cas, el coeficient $R^2\ ajustat$ pren un valor molt proper a la unitat, aproximadament de 0,90, per tant, es pot concloure que el model construït explica prou bé la variabilitat de la variable resposta, és a dir, les variables incloses com a explicatives tenen un efecte rellevant en la nota final dels estudiants del grau de Matemàtiques. A causa d'això, el model també podria ser útil per a fer prediccions sobre les notes finals de futurs alumnes si es disposés dels valors de les variables regresores.

Per acabar, s'interpretaran els valors dels coeficients dels paràmetres del model triat.

	Valor estimat	Error estàndard	Valor t	Pr(>|*t*|)
Intercept	2,54587	0,14063	18,103	< 2e-16
Diferència entre la nota d'accés i la nota mínima	0,03434	0,01162	2,956	0,00340
Sexe: Dona	-0,11615	0,04138	-2,807	0,00538
CFIS: SÍ	-0,33760	0,17042	-1,981	0,04864
Nota de càlcul numèric	0,12529	0,01684	7,439	1,46e-12
Nota de teoria de la probabilitat	0,14034	0,01623	8,646	5,42e-16
Nota de programació matemàtica	0,13863	0,01619	8,563	9,54e-16
Nota d'estructures algebraiques	0,09194	0,01994	4,610	6,30e-06
Nota d'estructures algebraiques*CFIS	0,06241	0,02403	2,596	0,00995
Nota d'estadística	0,15953	0,01787	8,926	< 2e-16
Significació global del model	Estadístic F: 285 (9 i 262 graus de llibertat), p-valor: < 2.2e-16			

Taula 4.2.9. Valors, significació individual pels coeficients dels paràmetres i significació global del model.

Com s'observa a la taula 4.2.9 la regressió és globalment significativa, ja que el p-valor pren un valor menor al nivell de significació (5%), per tant, es pot considerar que almenys un dels seus coeficients és estadísticament diferent de 0. En aquest cas, tots els coeficients del model són estadísticament diferent de 0, ja que els p-valors dels contrasts de significació individual són tots menors a 0.05.

Per començar, pel que fa a la variable que indica la diferència entre la nota mínima i la nota d'accés de l'any corresponent s'obté pel coeficient del seu paràmetre un valor de 0,03434, a causa d'això, es pot concloure que si aquesta diferència augmenta en un punt la nota final de l'estudiant augmentarà en aquesta dada, seguidament pel que fa al sexe, cal tenir en compte que s'ha escollit com a categoria base els homes, com a conseqüència, el fet de ser dona fa disminuir la nota final en 0,11615. Per acabar, pel que fa a les variables que fan referència a les notes de les assignatures, totes tenen valors pels coeficients dels seus paràmetres positius, és a dir, al obtenir major nota en aquestes augmenta la nota final. Tot i això, s'ha de destacar que l'assignatura d'estadística és la que té un coeficient més elevat, com a conseqüència, treure major nota en aquesta fa augmentar més la nota final, en comparació amb les altres. Pel que fa al CFIS, cal esmentar que interacciona amb la nota de l'assignatura estructures algebraiques, per tant, en funció d'aquesta variable tindrà un efecte o un altre sobre la nota final, ja que per una banda, pertànyer al CFIS fa disminuir la nota final en 0,3376, però, pertànyer i a més a més treure una nota més alta en estructures algebraiques fa que aquesta augmenti per cada punt de la nota de l'assignatura 0,06241 punts la nota final.

Per altra banda, també és necessari comentar que quatre de les cinc assignatures incloses en el model (excepte programació matemàtica) pertanyen al tercer curs del grau, per tant, es pot considerar que aquest és un dels més rellevants.

El model final és:

$$y = 2{,}54587 + 0{,}03434 * diferència\ entre\ la\ nota\ mínima\ i\ la\ nota\ d'accés$$
$$-0{,}11615 * sexe - 0{,}3376 * CFIS + 0{,}12529 * Nota\ de\ càlcul\ numèric + 0{,}14034$$
$$* Nota\ de\ teoria\ de\ la\ probabilitat + 0{,}13863 * Nota\ de\ programació\ matemàtica$$
$$+ 0{,}09194 * Nota\ d'estructures\ algebraiques + 0{,}15953 * Nota\ d'estadística$$
$$+ 0{,}06241 * Nota\ d'estructures\ algebraiques * CFIS$$

$$on\ sexe\ i\ CFIS\ són\ variables\ dicotòmiques\ que\ prenen\ el\ valor\ 1\ per\ les\ dones\ i\ 0$$
$$pels\ homes, i\ 1\ pels\ alumnes\ que\ pertànyen\ al\ CFIS\ i\ 0\ pels\ que\ no\ ho\ fan.$$

Pel cas dels alumnes nois i que no pertanyen al CFIS:

$$y = 2{,}54587 + 0{,}03434 * diferència\ entre\ la\ nota\ mínima\ i\ la\ nota\ d'accés$$
$$+0{,}12529 * Nota\ de\ càlcul\ numèric + 0{,}14034 * Nota\ de\ teoria\ de\ la\ probabilitat$$
$$+0{,}13863 * Nota\ de\ programació\ matemàtica + 0{,}09194$$
$$* Nota\ d'estructures\ algebraiques + 0{,}15953 * Nota\ d'estadística$$

Pel cas dels alumnes nois i que sí pertanyen al CFIS:

$$y = 2{,}20827 + 0{,}03434 * diferència\ entre\ la\ nota\ mínima\ i\ la\ nota\ d'accés$$
$$+0{,}12529 * Nota\ de\ càlcul\ numèric + 0{,}14034 * Nota\ de\ teoria\ de\ la\ probabilitat$$
$$+0{,}13863 * Nota\ de\ programació\ matemàtica + 0{,}15435$$
$$* Nota\ d'estructures\ algebraiques + 0{,}15953 * Nota\ d'estadística$$

Pel cas de les alumnes noies i que no pertanyen al CFIS:

$$y = 2{,}42972 + 0{,}03434 * diferència\ entre\ la\ nota\ mínima\ i\ la\ nota\ d'accés$$
$$+0{,}12529 * Nota\ de\ càlcul\ numèric + 0{,}14034 * Nota\ de\ teoria\ de\ la\ probabilitat$$
$$+0{,}13863 * Nota\ de\ programació\ matemàtica + 0{,}09194$$
$$* Nota\ d'estructures\ algebraiques + 0{,}15953 * Nota\ d'estadística$$

Pel cas de les alumnes noies i que sí pertanyen al CFIS:

$$y = 2{,}09212 + 0{,}03434 * diferència\ entre\ la\ nota\ mínima\ i\ la\ nota\ d'accés$$
$$+0{,}12529 * Nota\ de\ càlcul\ numèric + 0{,}14034 * Nota\ de\ teoria\ de\ la\ probabilitat$$
$$+0{,}13863 * Nota\ de\ programació\ matemàtica + 0{,}15435$$
$$* Nota\ d'estructures\ algebraiques + 0{,}15953 * Nota\ d'estadística$$

Seguidament, es calcularan els intervals de confiança al 95% pels coeficients del model utilitzant la funció *confint* d'*R* del paquet *stats*.

	2,5%	97,5%
Intercept	2,26895503	2,822782003
Diferència entre la nota d'accés i la nota mínima	0,01146554	0,057221646
Sexe: Dona	-0,19763613	-0,034669792
CFIS: SÍ	-0,67316879	-0,002024845
Nota de càlcul numèric	0,09212503	0,158454928
Nota de teoria de la probabilitat	0,10837832	0,172303243
Nota de programació matemàtica	0,10675307	0,170513813
Nota d'estructures algebraiques	0,05266846	0,131212875
Nota d'estadística	0,12434137	0,194724483
Nota d'estructures algebraiques*CFIS	0,01508026	0,109732113

Taula 4.2.10. Intervals de confiança pels coeficients dels paràmetres del model triat

En la taula 4.2.10 es pot veure que cap dels intervals al 95% de confiança pels coeficients dels paràmetres del model inclou el 0, ja que tal i com s'ha vist a la taula anterior tots aquests són significatius estadísticament. Pel cas de la diferència entre la nota mínima i la nota d'accés, si aquesta augmenta en un punt, la nota final augmentarà entre un 0,011 i un 0,0572 amb un 95% de confiança. Per altra banda, pel que fa al sexe, les dones tenen unes notes finals entre un 0,198 i un 0,0347 inferiors a les dels nois, a més a més pertànyer al CFIS fa que la nota final disminueixi entre un 0,67 i un 0,002. Tot i així, cal tenir en compte que per altra banda, si es pertany a aquest centre per cada punt que augmenti la nota de l'assignatura d'estructures algebraiques la nota final creix entre un 0,015 i 0,11. Per acabar, pel que fa a les notes de les assignatures totes fan augmentar la nota final, però com ja s'ha dit, la nota d'estadística és la que ho fa de forma més destacable, entre un 0,1243 i un 0,1947.

Finalment, un cop s'han realitzat tots els passos per a la creació d'un model lineal amb l'objectiu d'explicar la variabilitat de la nota final dels alumnes del grau de Matemàtiques s'ha arribat a diverses conclusions. Per una banda, el model creat té un ajust a les dades molt elevat, per tant, les variables que s'han escollit com a predictores tenen un gran efecte en la nota final.

També s'ha vist que les notes són les variables que més influència tenen en aquesta i en canvi, altres com el nombre d'assignatures aprovades a la primera o el nombre de quadrimestres en aprovar la fase inicial no hi tenen influència. A més a més, les interaccions entre el factor sexe i les variables numèriques no són significatives estadísticament, a causa d'això, es pot dir que les variables quantitatives afecten de forma igual als alumnes dels dos sexes. Per altra banda, la diferència entre la nota d'accés i la nota mínima sí influeix en el valor de la variable resposta, en aquest cas, alumnes que accedeixen al grau amb diferències entre aquestes dues notes elevades acostumen a tenir notes més altes. Pel cas dels estudiants que pertanyen al CFIS cal esmentar, que per una banda, el coeficient d'aquesta variable és negatiu, però per altra banda, la nota de l'assignatura d'estructures algebraiques té un major efecte, en aquest cas fa augmentar més la nota final en comparació amb la d'aquells que no hi pertanyen, com a conseqüència aquestes dues dades no es poden interpretar per separat sinó que l'efecte de pertànyer al CFIS en la nota final dependrà de la nota de l'estudiant en aquesta assignatura. Pel que fa al factor sexe, aquest sí té una influència en la nota final, és a dir, les noies acostumen a tenir notes més baixes que els nois, tot i així, l'efecte que aquesta variable té no és gaire destacable, ja que com es pot veure en la taula 4.2.10 amb un 95% de confiança la disminució de la nota pel fet de ser dona pot arribar fins a 0,2, per tant, es pot dir que la diferència entre homes i dones al grau a causa d'aquest factor és molt lleugera.

Com a conseqüència, l'èxit en aquests estudis no es troba marcat de forma específica pel fet de pertànyer al sexe femení o masculí, sinó que són les notes que s'obtenen a les assignatures les que hi tenen un major efecte.

5. Anàlisi de la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres

En aquest apartat s'estudiarà la probabilitat d'aprovar la fase inicial pels alumnes del grau de Matemàtiques en dos quadrimestres. Cal destacar, que la fase inicial està composta per les assignatures del primer i segon quadrimestre de la carrera, és a dir, el primer curs.

L'objectiu de l'anàlisi consistirà en observar quines de les variables de la base de dades tenen un efecte en aquesta probabilitat, concretament es vol conèixer si el fet de ser home o dona afecta de forma rellevant a la variable d'estudi.

La base de dades que s'usarà és la mateixa que en l'apartat 3, però en aquest cas, s'ha creat una nova variable dicotòmica a partir de la que indicava el nombre de quadrimestres que els alumnes graduats que van entrar durant els anys 2009-2015 havien trigat en aprovar la fase inicial, aquesta pren el valor 1, per aquells que l'han aprovat en dos quadrimestres i 0 pels que no ho han fet.

Per tant, per assolir l'objectiu es crearà un model lineal generalitzat usant la variable que indica si s'ha aprovat la fase inicial en dos quadrimestres o no com a resposta i com a regresores s'utilitzaran les altres variables de la base de dades, però en aquest cas, no s'inclouran aquelles que fan referència a les notes de les assignatures, ni el nombre d'assignatures aprovades en la primera matricula, ja que fan referència a dades de períodes posteriors.

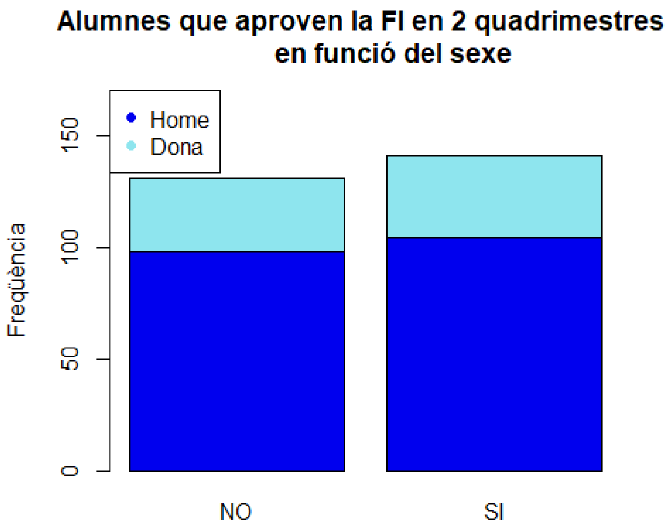
Pel que fa a les dades de les assignatures obligatòries del primer i segon quadrimestre s'ha decidit no incloure-les en el model, a causa que no es coneix si aquestes van ser aprovades en la primera matricula o no per a tots els estudiants, és a dir, no té el mateix efecte en aquesta variable resposta una nota que ha estat aprovada en la primera o per el contrari en successives matricules.

5.1. *Anàlisi bivariant amb la variable resposta*

Abans de construir el model lineal generalitzat és molt útil realitzar un anàlisi bivariant de les variables explicatives amb la resposta, amb l'objectiu de poder veure quines d'aquestes es troben més relacionades amb la variable independent (*si s'ha aprovat la fase inicial del grau en dos quadrimestres o no*) i com és aquesta relació.

A més a més, alhora de fer l'anàlisi es tindrà en compte el factor sexe, per tal d'observar quines de les variables tenen una relació amb la resposta que varia en funció del fet de ser home o dona. Per a dur a terme aquest es faran diversos gràfics i taules bivariants amb el programa *R* [14], ja que gràcies a aquests es pot veure fàcilment la relació entre dues variables.

[14] Els codis que s'han fet servir es troben a l'annex: Codi d'R per a l'anàlisi descriptiu bivariant de les variables de la base de dades dels apartats 4. Anàlisi de la nota final dels estudiants del grau de Matemàtiques i 5. Anàlisi de la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres amb la variable dicotòmica que indica si s'ha aprovat la fase inicial del grau en dos quadrimestres o no.



Gràfic 5.1.1. Nombre d'alumnes que aproven la fase inicial del grau de Matemàtiques en dos quadrimestres en funció del sexe.

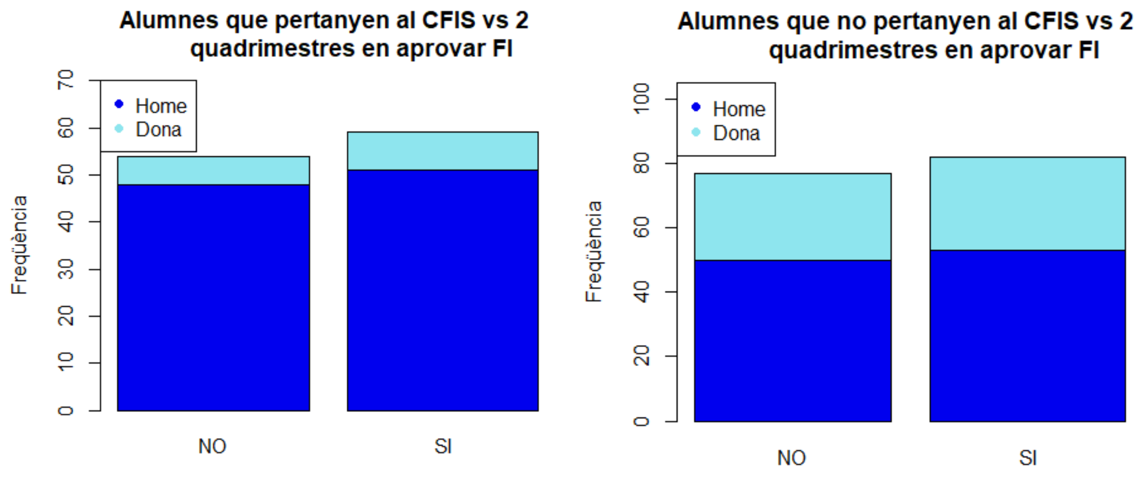
En el gràfic 5.1.1 es pot veure que no hi ha una relació entre el nombre d'alumnes que aproven la fase inicial del grau en dos quadrimestres amb el factor sexe, sinó que pels dos casos el nombre d'estudiants que l'aconsegueix aprovar representa aproximadament la meitat del total. Tot i així, cal esmentar que existeix un nombre lleugerament més elevat d'alumnes, tant nois com noies, que sí aproven aquesta fase en dos quadrimestres.

Aprovar la fase inicial en dos quadrimestres		
	NO	**SÍ**
2009		
HOMES	17	9
DONES	6	6
2010		
HOMES	7	23
DONES	6	7
2011		
HOMES	18	17
DONES	10	2
2012		
HOMES	21	16
DONES	4	8

	NO	**SÍ**
2013		
HOMES	15	13
DONES	7	8
2014		
HOMES	17	19
DONES	0	5
2015		
HOMES	3	7
DONES	0	1

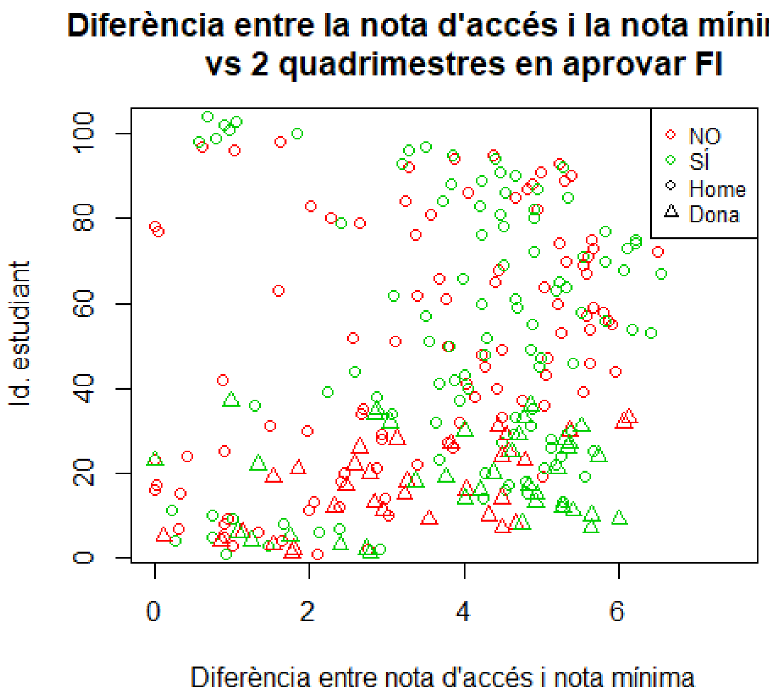
Taula 5.1.1 Nombre d'alumnes que aproven la fase inicial del grau de Matemàtiques en dos quadrimestres en funció de l'any d'entrada i el sexe.

En la taula 5.1.1 s'observa que no existeix una relació clara entre el nombre d'alumnes que aprova la fase inicial del grau en dos quadrimestres i l'any d'entrada. Per altra banda, pel que fa al sexe tampoc es veu que tingui una influència en l'efecte d'aquesta variable. Tot i així, cal destacar que pel que fa als homes, els anys 2010, 2014 i 2015 hi va haver més que van aprovar aquesta fase en dos quadrimestres, en canvi, pel que fa a les noies hi va haver més els anys 2010, 2012, 2013, 2014, 2015.



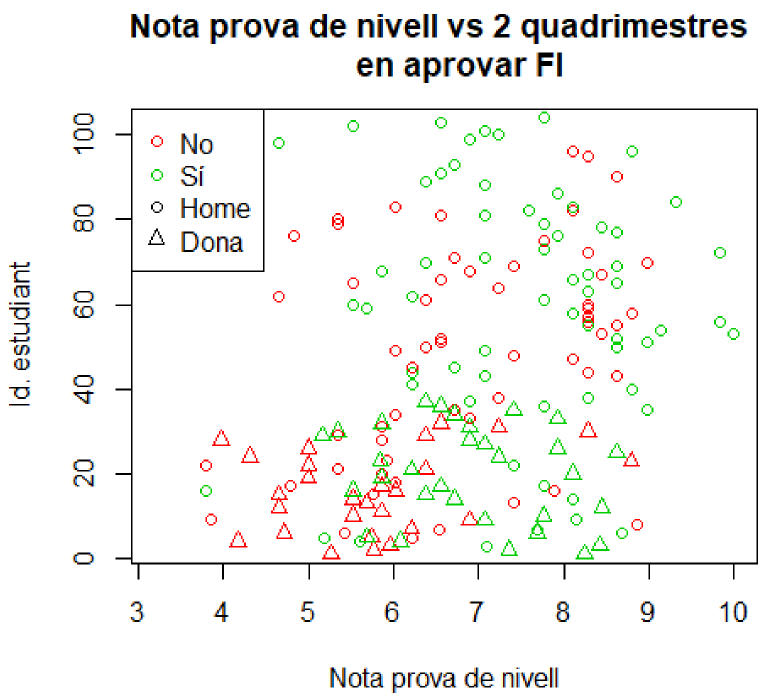
Gràfic 5.1.2. Nombre d'alumnes que aproven la fase inicial del grau de Matemàtiques en dos quadrimestres en funció de si pertanyen o no al CFIS i el sexe.

En el gràfic 5.1.2 es pot veure que no existeix relació entre el nombre d'alumnes que aproven la fase inicial del grau de Matemàtiques en dos quadrimestres i el fet de si pertanyen o no al CFIS, a més a més en aquesta relació tampoc té influència el sexe, ja que, tant pels estudiants nois i noies que pertanyen al CFIS com pels que no ho fan s'observa que existeix un nombre lleugerament major d'alumnes que aproven la fase inicial en dos quadrimestres.



Gràfic 5.1.3. Distribució de la diferència entre la nota d'accés i la nota mínima de l'any corresponent dels alumnes del grau de Matemàtiques en funció de si aproven la fase inicial de la carrera en dos quadrimestres i el sexe.

En el gràfic 5.1.3 s'observa que sí existeix una lleugera relació entre la diferència entre la nota d'accés i la nota mínima i el fet d'aprovar la fase inicial del grau en dos quadrimestres. Pel que fa als alumnes que aproven aquesta fase en dos quadrimestres, acostumen a accedir a la universitat amb una diferència entre la seva nota d'accés i la nota mínima de l'any corresponent d'entre 4 i 6 aproximadament, en canvi, per a aquells estudiants que no aproven la fase inicial en aquest període aquesta diferència varia entre 0 i 6, majoritàriament. Per altra banda, pel que fa al sexe, es troba que hi ha la mateixa relació, tant pels nois com per les noies, és a dir, per aquells que aproven la fase inicial en dos quadrimestres predominen diferències entre les dues notes més elevades.



Gràfic 5.1.4. Distribució de la nota de la prova de nivell voluntària dels alumnes del grau de Matemàtiques en funció de si aproven la fase inicial de la carrera en dos quadrimestres i el sexe.

En el gràfic 5.1.4 es pot veure que sí hi ha una relació entre la nota de la prova de nivell voluntària i el fet d'aprovar la fase inicial del grau de Matemàtiques en dos quadrimestres. Pel que fa als estudiants que tenen notes de la prova de nivell elevada predominen aquells que sí aproven aquesta fase en el període indicat, en canvi, els que tenen notes més baixes acostumen a no aprovar la fase inicial en dos quadrimestres. Per altra banda, pel que fa al sexe, s'observa que no té una influència destacable en la relació entre aquestes variables. Tot i així, s'ha d'esmentar, que pel cas dels nois i les noies que aproven la fase inicial en dos quadrimestres acostumen a tenir notes d'entre 6 i 9, en canvi, pels alumnes que no aproven la fase inicial en aquest període pel cas dels homes es pot veure que les seves notes varien entre 4 i 9, i per les dones es troben entre 4 i 7, majoritàriament. Però, cal tenir en compte que aquestes conclusions no són gaire fiables, ja que la prova de nivell es voluntària i per tant, hi ha un nombre elevat d'estudiants que decideix no realitzar-la.

5.2. *Model Lineal Generalitzat*

Un cop s'ha observat com es distribueixen les diverses variables de la base de dades que s'usaran com a possibles predictores i la relació entre aquestes i la resposta (*si s'ha aprovat la fase inicial en dos quadrimestres o no*), es crearan diversos models lineals generalitzats[15] amb l'objectiu de trobar quin és el més útil per a poder explicar la variabilitat de la variable independent.

Per a la creació dels models s'usarà la funció *glm* del paquet *stats* del programa *R* [16] especificant com valor per a la família la *distribució Binomial* i el link *logit*, a més a més també caldrà determinar quines seran les variables que s'inclouran en cada un dels possibles models. Seguidament, s'obtindrà com a resultat de la funció anterior els valors dels paràmetres del model així com altres indicadors que seran molt útils per a valorar l'ajust de cada un dels models a les dades reals.

Per començar, es crearà un model en el que s'inclouran totes les variables de les quals es disposa, excepte la nota de la prova de nivell voluntària, ja que es una variable de la qual es troben molts valors faltants, i les interaccions entre les variables numèriques i els factors sexe i CFIS, per tal de veure si alguna de les explicatives té un efecte en la resposta que depèn del fet de ser home o dona, o de pertànyer o no al CFIS.

A continuació, per a crear els models successius s'aniran eliminant del primer aquelles interaccions o variables que no siguin significatives, és a dir, aquelles que es poden considerar estadísticament iguals a 0. Finalment, l'objectiu en l'últim model és que totes les variables incloses siguin significatives, és a dir, que no hi hagi variables irrellevants que no ajuden a explicar la variabilitat de la resposta.

En la taula següent es troben tots els models lineals generalitzats creats i les variables de cadascun. Cal esmentar, que la variable resposta en tots ells és una dicotòmica que pren el valor 1 si l'estudiant ha aprovat la fase inicial del grau de Matemàtiques en dos quadrimestres o 0 en cas contrari.

Model	Variables explicatives
MODEL 1	CFIS, diferència entre la nota d'accés i la nota mínima de l'any corresponent, diferència entre la nota d'accés i la nota mínima de l'any corresponent*CFIS, diferència entre la nota d'accés i la nota mínima de l'any corresponent*sexe, sexe[17]
MODEL 2	CFIS, diferència entre la nota d'accés i la nota mínima de l'any corresponent, diferència entre la nota d'accés i la nota mínima de l'any corresponent*CFIS, sexe

[15] Les bases d'aquests models es troben a l'apartat: 2.2 Models Lineals Generalitzats.

[16] A l'annex es troben els codis usats: Codi d'R per a la creació del model lineal generalitzat de l'apartat 5. Anàlisi de la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres.

[17] El signe "*" indica la interacció entre dues variables.

MODEL 3	CFIS, diferència entre la nota d'accés i la nota mínima de l'any corresponent, diferència entre la nota d'accés i la nota mínima de l'any corresponent*CFIS
MODEL 4	CFIS, diferència entre la nota d'accés i la nota mínima de l'any corresponent
MODEL 5	Diferència entre la nota d'accés i la nota mínima de l'any corresponent

Taula 5.2.1. Models lineals generalitzats creats per a explicar la variabilitat de la probabilitat d'aprovar la fase inicial del grau de Matemàtiques en dos quadrimestres.

Un cop creats tots els models es necessari dur a terme el test de la deviança per a poder-los comparar dos a dos i decidir quin és l'escollit. Per a fer-ho s'ha aplicat la funció *anova* del paquet *stats* d'*R* [18].

En la taula que es mostra a continuació es troben el p-valors dels diferents contrasts. Cal tenir en compte que s'usarà un nivell de significació del 5%, per tant, un p-valor major a 0.05 indicarà que els models comparats són equivalents, és a dir, no es pot rebutjar la hipòtesi nul·la del test.

Models comparats	P-valor del test de la deviança
MODEL 1 VS MODEL 2	0,8197
MODEL 2 VS MODEL 3	0,9752
MODEL 3 VS MODEL 4	0,2343
MODEL 4 VS MODEL 5	0,2735

Taula 5.2.2. P-valors dels test de la deviança per a comparar dos models lineals generalitzats encaixats.

Per tant, a causa de que tots els models són equivalents s'escollirà com a millor el més simple, en aquest cas el model 5, ja que només inclou com a variable explicativa la diferència entre la nota d'accés i la nota mínima de l'any corresponent.

Seguidament, per assegurar que el model 5 és el més òptim es tronaran a comparar aquests segons els valors dels altres indicadors: AIC, BIC i Pseudo-R^2. Per a obtenir els valors de l'AIC i el BIC s'han utilitzat les funcions *AIC* i *BIC* del paquet *stats* d'*R* (aquesta calcula el valor de l'indicador corresponent a partir de l'especificació d'un model lineal generalitzat). Per al Pseudo-R^2 s'ha especificat la seva fórmula en el mateix programa.

En la següent taula es mostren els valors de l'AIC, BIC i Pseudo-R^2 pels 5 models creats.

[18] A l'annex es poden veure els codis d'R que s'han utilitzat: Codi d'R per a la creació del model lineal generalitzat de l'apartat 5. Anàlisi de la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres.

Model	AIC	BIC	Pseudo-R^2
MODEL 1	378,2282	399,863	0,02780988
MODEL 2	376,2802	394,3092	0,02767193
MODEL 3	374,2812	388,7044	0,02766937
MODEL 4	373,6955	384,5129	0,02391484
MODEL 5	372,8947	380,1063	0,02073145

Taula 5.2.3. Valors de l'AIC, el BIC i el Pseudo-R^2 pels 6 models lineals generalitzats.

Tal i com s'observa a la taula 5.2.3 el millor model segons el valor de l'AIC és el 5, ja que són preferibles models amb un valor baix per a aquest indicador, pel que fa al BIC també es segueix el mateix criteri de preferència, per tant, el millor model també és el 5. Finalment, segons el Pseudo-R^2 és millor que sigui més elevat, a causa d'això seria escollit el model 1.

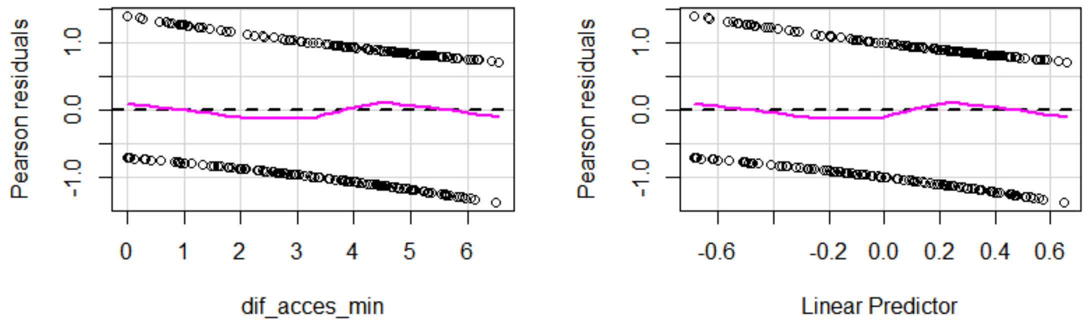
Cal destacar que el coeficient Pseudo-R^2 indica la proporció de la variabilitat de la resposta que es troba explicada pel model, per tant, és habitual que models amb més variables (com és el model 1) tinguin un valor d'aquest indicador més elevat que els que en tenen menys, és a dir, s'aconseguirà explicar més proporció de la variabilitat incloent un major nombre d'explicatives. Però, en aquest cas, l'augment de la proporció de la variabilitat que s'aconsegueix explicar amb el model 1 en comparació amb el 5 és molt poc significativa, ja que ambdós models aconsegueixen explicar una proporció molt petita, com a conseqüència no compensa el gran nombre de variables que té aquest.

Per acabar, es pot concloure que dels cinc models lineals generalitzats creats per a conèixer quines variables tenen més efecte en la probabilitat d'aprovar la fase inicial del grau de Matemàtiques en dos quadrimestres el millor és el que inclou com a variable explicativa només la diferència entre la nota d'accés de l'alumne i la nota mínima d'accés per a cada un dels anys.

A continuació, cal validar el model escollit. Per aconseguir-ho, es representaran els residus de *Pearson* en diversos gràfics amb l'ajuda de la funció *residualPlots* del paquet *car* d'*R* [19].

Per a donar un model per vàlid cal que en aquestes representacions es vegi que la distribució dels residus pels diferents valors del predictor lineal o de les variables explicatives sigui semblant, però, a causa de que la resposta d'aquest tipus de models són 0 o 1 aquesta distribució és complicada d'observar. Un model correcte necessita que la funció mitjana condicional sigui constant per a tots els valors del predictor lineal o de les variables regresores. Per això, aquesta funció d'*R* no només representa els residus del model, sinó que afegeix una representació d'una línia suavitzada, com a conseqüència si aquesta es prou recta (per tant és constant) el model es donarà per validat.

[19] Els codis d'R usats es troben a l'annex: Codi d'R per a la creació del model lineal generalitzat de l'apartat 5. Anàlisi de la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres.



Gràfic 5.2.1. Respresentació dels residus de Pearson del model lineal generalitzat escollit vs el predictor lineal i la variable esxplicativa.

En el gràfic 5.2.1 es pot veure que la línia suavitzada que representa *R* a partir de la funció *residualPlots* es comporta prou constant al voltant de 0 tant per la variable que indica la diferència entre la nota d'accés i la nota mínima com pels valors del predictor lineal, per tant, el model escollit es pot donar per vàlid.

Seguidament, s'avaluarà la capacitat predictiva del model per tal de conèixer la precisió de les estimacions que aquest du a terme. Per això, es calcularan els indicadors següents: sensibilitat, especificitat, valor predictiu positiu i valor predictiu negatiu [20]. A més a més, a partir d'aquestes mesures es representarà la corba ROC i es calcularà l'estadístic AUC per a valorar com de bo és el model pel que fa a la capacitat de discriminació entre aquells alumnes que aproven la fase inicial en dos quadrimestres i els que no ho fan, a partir del valor de la variable explicativa. En la taula que es mostra a continuació es troben els valors dels indicadors mencionats i la proporció d'encert del model, cal esmentar que per al seu càlcul s'ha usat com a llindar de la probabilitat predita a partir de la qual es considera que la resposta és positiva el valor 0.5.

Sensibilitat	0,7234043
Especificitat	0,4656489
Valor predictiu positiu	0,5930233
Valor predictiu negatiu	0,61
Proporció d'encert	0,5992647

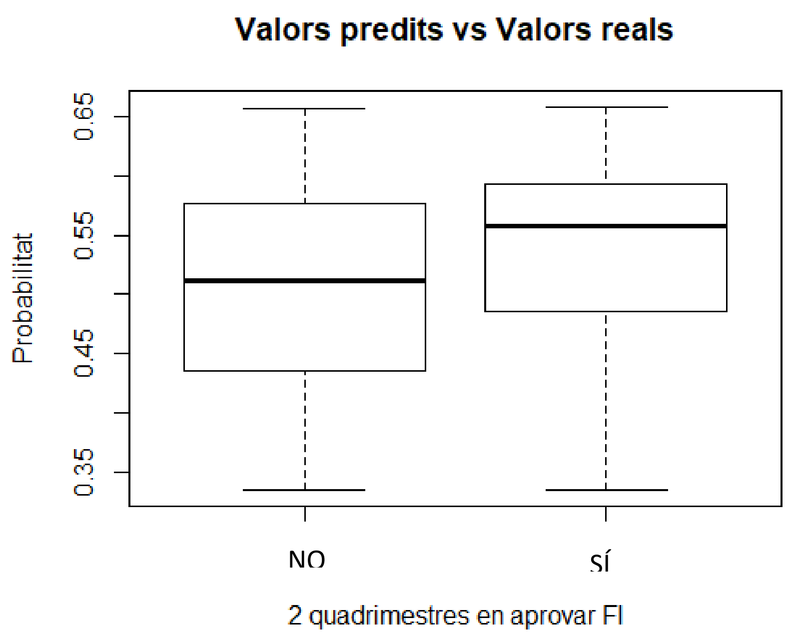
Taula 5.2.4. Valors dels indicadors: Sensibilitat, Especificitat, Valor predictiu positiu i Valor predictiu negatiu i la proporció d'encert pel model triat.

En la taula 5.2.4 es pot veure que el model té una proporció d'encert del 0,60 aproximadament, és a dir, un 60% de les vegades diferència correctament entre els estudiants que aproven la fase inicial en dos quadrimestres i els que no ho fan. Per altra banda, la sensibilitat pren un valor del 0,72, aquesta indica la proporció d'observacions que

[20] Per al càlcul d'aquests indicadors s'ha utilitzat R: Codi d'R per a la creació del model lineal generalitzat de l'apartat 5. Anàlisi de la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres.

s’han predit correctament com a positives mitjançant el model, l’especificitat indica la mateixa proporció, però pels resultats negatius, és a dir, la proporció d’observacions que s’han predit correctament com a negatives és gairebé d’un 0,47. Pel que fa al valor predictiu positiu, fa referència a la proporció de resultats positius observats dins dels que s’han predit com a positius, un 0,59 i finalment el valor predictiu negatiu és d’un 0,61 (proporció de resultats negatius observats dins dels que s’han predit com a negatius).

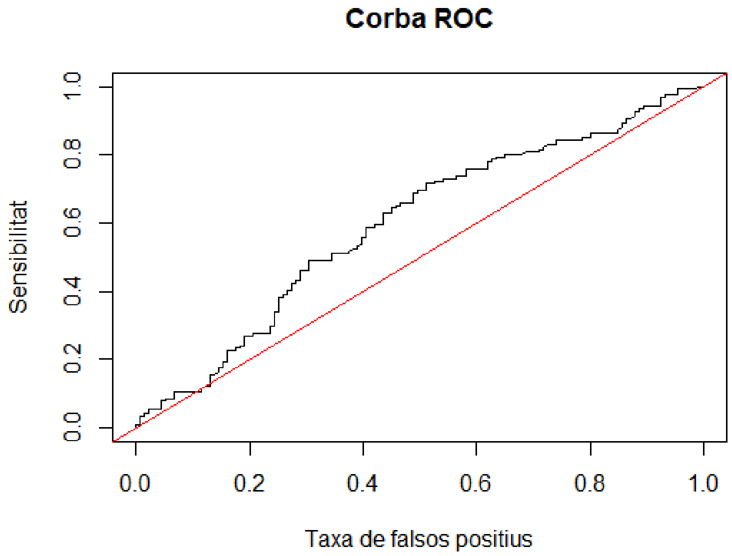
A causa d’aquests resultats es pot concloure que, el model té una capacitat més elevada per a detectar els casos realment positius (és a dir, aquells estudiants que han aprovat la fase inicial en dos quadrimestres) que els casos verdaderament negatius, però en canvi, si aquest prediu que l’alumne aprovarà la fase inicial en dos quadrimestres aproximadament només un 60% de les vegades realment ho farà.



Gràfic 5.2.3. Gràfic dels valors de la probabilitat predits pel model vs el valor real de la variable resposta (si s’ha aprovat o no la fase inicial del grau en dos quadrimestres).

En el gràfic 5.2.3 s’observa que dels valors que prediu el model per a la probabilitat d’aprovar la fase inicial del grau de Matemàtiques en dos quadrimestres gran part d’ells es troben compartits pels que sí l’aproven realment i els que no ho fan, és a dir, el model no discrimina força bé entre els dos grups, cosa que ja es veia amb el valors dels indicadors calculats anteriorment. Cal destacar, que pels estudiants amb verdaderament resposta positiva el model prediu majoritàriament probabilitats elevades (majors a 0,5), però, en canvi, pels quals la resposta és negativa hi ha molta més variabilitat, o sigui, les probabilitats no es concentren en valors baixos, sinó que la major part es troben entre 0,45 i 0,60.

Per acabar d'avaluar la capacitat predictiva del model es calcularà l'estadístic AUC[21]. Aquest indica l'àrea sota la corba ROC. L'anàlisi de la corba ROC serveix per a conèixer la precisió de les estimacions. Aquesta representa per a cada possible llindar per a la probabilitat predita el valor de la $Sensibilitat$ en front de la taxa de falsos positius ($1 - Especificitat$). Si la corba ROC augmenta ràpidament es considera que el model funciona bé, per tant, entre més elevat sigui l'estadístic AUC[22] millor serà el model.



Estadístic AUC
0,5965297

Gràfic 5.2.2. Representació de la corba ROC i valor de l'estadístic AUC pel model escollit.

Pel que fa a l'estadístic AUC, pren un valor d'aproximadament 0,60 per tant, es pot dir que el model no és gaire útil per conèixer quins alumnes aprovaran la fase inicial del grau de Matemàtiques en dos quadrimestres. Això, pot ser degut a causa de que aquesta probabilitat no depèn estrictament de les variables que s'han usat en el model com a predictores, sinó que altres factors no coneguts també tenen influència. Per tant, tal i com s'havia vist en els gràfics de l'anàlisi descriptiva bivariant, tot i que la variable que indica la diferència entre la nota d'accés i la nota mínima tenia relació amb el fet d'aprovar la fase inicial en dos quadrimestres, no era una relació gaire forta (hi havia molta variabilitat) i això ha provocat que la capacitat predictiva del model sigui baixa. Tot i així, cal tenir en compte que la finalitat d'aquest model no és fer prediccions sinó explicar la variabilitat de la resposta.

Finalment, s'interpretarà el coeficient del model per a la variable que indica la diferència entre la nota d'accés i la nota mínima de l'any corresponent. Per a fer-ho cal tenir en compte la funció d'enllaç usada, que en aquest cas ha estat el *link logit*, aquest es defineix de la següent manera:

$$g(\pi) = \eta = logit\ (\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

[21] Per al càlcul de l'AUC s'ha utilitzat el programa R: Codi d'R per a la creació del model lineal generalitzat de l'apartat 5. Anàlisi de la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres.

[22] Els criteris per a conèixer com de bo és un model segons el valor de l'estadístic AUC es troben a l'apartat: 2.2 Models Lineals Generalitzats.

$$Per\ tant,\ \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 * diferència\ entre\ la\ nota\ mínima\ i\ la\ nota\ d'accés$$

$\frac{\pi}{1-\pi}$ s'anomena *odds* i representa la relació entre la probabilitat de resposta positiva i resposta negativa. Com a conseqüència, si l'*odds* pren un valor major a 1 indica que la probabilitat de resposta positiva és major a la de negativa, al contrari succeeix per a valors menor a la unitat. Per tant, l'exponencial dels paràmetres ß representarà l'*odds-rati*, és a dir, la relació entre dues *odds* en funció del valor que pren la variable explicativa que acompanya al paràmetre corresponent.

Els valors obtinguts pels paràmetres del model escollit es mostren a la taula que es pot veure a continuació:

	Valor estimat	Error estàndard	Valor Z	Pr(>|Z|)
Intercept	-0,68566	0,30260	-2,266	0,02346
Diferència entre la nota d'accés i la nota mínima	0,20484	0,07443	2,752	0,00592

Taula 5.2.5. Valors i significació individual pels coeficients dels paràmetres del model triat.

Com s'observa a la taula 5.2.5 els dos coeficients del model són estadísticament significatius, és a dir, diferents de 0. A més a més, el coeficient del paràmetre de la variable que fa referència a la diferència entre la nota d'accés i la nota mínima de l'any corresponent pren un valor de 0,20484, és a dir, l'increment d'un punt en la diferència de notes provoca un augment del *logaritme d'odds* de 0,20484. Per a poder desfer la transformació logarítmica cal calcular l'exponencial d'aquest valor:

$$e^{0{,}20484} = 1{,}227323 = \frac{Odd_{diferencia\ notes+1}}{Odd_{diferència\ notes}}$$

Per tant, l'*odds* d'aprovar la fase inicial en dos quadrimestres és 1,227323 vegades més gran en augmentar la diferència de notes en un punt (és a dir, és un 22,73% major). Com a conseqüència, es pot concloure que la probabilitat d'aprovar la fase inicial del grau de Matemàtiques en dos quadrimestres és major per aquells alumnes que tenen una diferència entre les dues notes elevada. A causa d'aquests resultats, cal esmentar que no es troben diferències en la probabilitat d'aprovar la fase inicial en funció del sexe.

El model final és:

$$\log\left(\frac{\pi}{1-\pi}\right) = -0{,}68566 + 0{,}20484 * diferència\ entre\ la\ nota\ mínima\ i\ la\ nota\ d'accés$$

$$\pi = \frac{e^{(-0{,}68566+0{,}20484*diferència\ entre\ la\ nota\ mínima\ i\ la\ nota\ d'accés)}}{1 + e^{(-0{,}68566+0{,}20484*diferència\ entre\ la\ nota\ mínima\ i\ la\ nota\ d'accés)}}$$

$$on\ \pi\ indica\ la\ probabilitat\ de\ resposta\ positiva, és\ a\ dir, la\ probabilitat\ d'aprovar\ la\ fase\ incial\ del\ grau\ en\ dos\ quadrimestres$$

A continuació, es calcularan els intervals de confiança al 95% pels dos coeficients del model. Per això, s'usarà la funció *confint* d'*R* [23] del paquet *stats*. A més, també es farà l'exponencial dels valors de l'interval per a poder-los interpretar més fàcilment, en termes d'*odds*.

	2,5%	97,5%	Exponencial (2,5%)	Exponencial (97,5%)
Intercept	-1,29061459	-0,100065 3	0,2751017	0,9047784
Diferència entre la nota d'accés i la nota mínima	0,06068457	0,3532238	1,0625637	1,4236497

Taula 5.2.6. Intervals de confiança pels coeficients dels paràmetres del model triat.

En la taula 5.2.6 es pot veure que cap dels dos intervals pels coeficients inclou el 0 (ni pel cas dels exponencials l'1), el que indica que aquests són significatius estadísticament, tal i com s'havia vist anteriorment. Pel cas, de la diferència entre la nota d'accés i la nota mínima es pot concloure, que l'*odds* d'aprovar la fase inicial en dos quadrimestres serà entre 1,062563 i 1,4236497 vegades més gran (és a dir, entre un 6,2563% i un 42,3697% major), amb un 95% de confiança, en augmentar la diferència de notes en un punt. Per tant, la probabilitat d'aprovar la fase inicial del grau de Matemàtiques en dos quadrimestres és més elevada pels alumnes que tenen una major diferència entre la seva nota d'accés i la nota mínima de l'any corresponent.

Per acabar, un cop s'han realitzat tots els passos per a la creació del model s'ha arribat a diverses conclusions. Per una banda, s'ha vist que el sexe no té influència en el fet d'aprovar la fase inicial del grau en dos quadrimestres, a més a més de totes les variables disponibles per a usar com a regresores només la diferència entre la nota d'accés i la nota mínima té una incidència en aquesta probabilitat. Per altra banda, el model creat té una capacitat predictiva baixa, és a dir, la variabilitat de la resposta depèn, en gran part, d'altres factors no coneguts, com pot ser la situació personal de l'estudiant, la seva adaptació al món universitari, etc., tot i així, es pot dir que per a aquells alumnes que accedeixen a la carrera amb un nivell més alt (és a dir, amb una nota més elevada) la probabilitat d'aprovar la fase inicial en dos quadrimestres és major.

5.3. Comparació de les proporcions d'alumnes graduats de cada sexe

En aquest apartat es compararan les proporcions de nois i noies per a cada un dels anys de la base de dades, és a dir, del 2009-2015, amb l'objectiu de verificar si realment existeixen diferències entre sexes pel que fa als alumnes graduats, concretament es vol conèixer si el nombre d'homes és major al de dones o si per el contrari són iguals.

[23] A l'annex es poden veure els codis usats: Codi d'R per a la creació del model lineal generalitzat de l'apartat 5. Anàlisi de la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres.

Cal destacar, que si no hi hagués diferències entre sexes les dues proporcions serien iguals a 0.5, per tant, per a aconseguir l'objectiu només serà necessari comparar una de les dues proporcions amb aquest valor. En aquest cas, s'ha escollit comparar la proporció d'estudiants del sexe masculí amb el valor 0.5.

Les hipòtesis del test són:

- $H_0: p = 0.5$
- $H_1: p > 0.5, \; p \; indica \; la \; porporció \; d'alumnes \; nois.$

L'estadístic per a resoldre aquest contrast és:

$$Z = \frac{\hat{p} - 0.5}{\sqrt{\frac{\hat{p} * \hat{q}}{n}}} \sim N(0,1)$$

$$on \;\; n = \; total \; d'alumnes \; graduats, \hat{p} = \frac{nombre \; d'alumnes \; nois}{total \; d'alumnes \; graduats} \;\; i \;\; \hat{q} = 1 - \hat{p}$$

Es rebutjarà la hipòtesi nul·la amb un nivell de significació α si: $Z \geq Z_\alpha$ (valor crític unilateral de la distribució Normal (0, 1) associat a una probabilitat α $(P(Z \geq Z_\alpha) = \alpha)$.

Per a que l'estadístic Z es distribueixi de forma asimptòtica com una distribució Normal tipificada cal verificar que es compleixen aquestes condicions:

- $n \geq 30$
- $n * \hat{p} \geq 5$

Per a dur a terme tots els càlculs s'usarà el programa *R* [24].

En la taula que es mostra a continuació es troben el nombre de nois, noies i el total de graduats per a cada any.

Any d'entrada	Nois	Noies	Total
2009	26	12	38
2010	30	13	43
2011	35	12	47
2012	37	12	49
2013	28	15	43
2014	36	5	41
2015	10	1	11

Taula 5.3.1. Nombre d'alumnes graduats per sexe pels anys 2009 - 2015.

Tal i com s'observa a la taula 5.3.1 n és major a 30, excepte pel 2015 i $n * \hat{p}$, és a dir, el nombre de nois, és major a 5 per a tots els anys. Com a conseqüència, el test es podrà aplicar correctament pels anys 2009-2014. Per altra banda, el nombre total d'estudiants es situa al voltant de 45 per a la majoria d'anys. A més a més, també es pot veure que el de

[24] A l'annex es poden trobar els codis usats: Codis d'R per a la comparació de proporcions de *nois i noies al grau de Matemàtiques* i test d'homogeneïtat.

noies està habitualment al voltant de 13 (excepte per a l'any 2014 que n'hi va haver només 5), en canvi, el nombre de nois es troba a prop de 35, el que representa gairebé el doble del de dones, per tant, segurament es rebutjarà la hipòtesi nul·la del contrast.

Seguidament, es calcularà el valor de l'estadístic Z [25] pels anys 2009-2014, ja que són els que compleixen les condicions de validesa del test. En la taula següent es mostren els resultats.

Any d'entrada	Proporció de nois	Z
2009	0,68421053	2,44293854
2010	0,69767442	2,82241486
2011	0,74468085	3,84699945
2012	0,75510204	4,15256623
2013	0,65116279	2,07980654
2014	0,87804878	7,39754088

Taula 5.3.2. Proporció d'alumnes nois i càlcul de l'estadístic Z pels anys 2009 – 2014.

Per un nivell de significació $\alpha = 0.05$, $Z_{\alpha} = 1{,}644853627$, per tant, a causa de que el valor de l'estadístic calculat per a cada any és major a aquest es pot dir amb un nivell de confiança del 5% que la proporció de nois pels anys 2009-2014 és major al 50%. Com a conseqüència, la proporció d'homes i dones al grau de Matemàtiques són diferents, exactament la del sexe masculí és major a la del femení.

Per acabar, es calcularan els intervals de confiança al 95% bilaterals i unilaterals [26] per a les proporcions de nois i noies pels anys 2009, 2010, 2011, 2012, 2013 i 2014. A les taules que es poden veure a continuació es troben els seus valors:

		Bilateral		Unilateral	
Any d'entrada	Proporció de nois	Límit inf.	Límit sup.	Límit inf.	Límit sup.
2009	0,68421053	0,53641884	0,83200221	0,56017984	1
2010	0,69767442	0,56040376	0,83494508	0,58247325	1
2011	0,74468085	0,62002119	0,86934051	0,64006316	1
2012	0,75510204	0,63469678	0,8755073	0,65405476	1
2013	0,65116279	0,5087103	0,79361528	0,53161289	1
2014	0,87804878	0,7778855	0,97821207	0,7939891	1

Taula 5.3.3. Proporció d'alumnes nois i càlcul dels intervals de confiança al 95% pels anys 2009 – 2014.

A la taula 5.3.3 es pot observar que pel que fa als intervals de confiança tant bilaterals com unilaterals cap inclou el valor 0.5 sinó que inclouen valors més elevats, a causa d'això s'arriba al mateix resultat que anteriorment, és a dir, la proporció d'homes al grau és major a 0.5 amb una confiança del 95%.

[25] Els càlculs realitzats per trobar els valors de l'estadístic Z es poden veure a l'annex: Codis d'R per a la comparació de proporcions de *nois i noies al grau de Matemàtiques* i test d'homogeneïtat.

[26] A l'annex es troben els càlculs pels intervals de confiança: Codis d'R per a la comparació de proporcions de *nois i noies al grau de Matemàtiques* i test d'homogeneïtat.

Per altra banda, s'ha de destacar que la majoria d'intervals bilaterals inferiors prenen valors propers als 0.6, excepte per a l'any 2014 que pren un valor de 0.78, aproximadament, com a conseqüència, es pot dir que per a aquell any hi va haver molta més diferència entre el nombre d'alumnes de cada sexe en comparació amb els altres. Aquesta diferència també es pot veure en el valor que pren la proporció de nois de l'any 2014, 0.87, per tant, quasi tots els estudiants d'aquell any van ser homes, en canvi, pel que fa als altres aquesta proporció pren valors propers al 0.7.

		Bilateral		Unilateral	
Any d'entrada	Proporció de noies	Límit inf.	Límit sup.	Límit inf.	Límit sup.
2009	0,31578947	0,16799779	0,46358116	0	0,43982016
2010	0,30232558	0,16505492	0,43959624	0	0,41752675
2011	0,25531915	0,13065949	0,37997881	0	0,35993684
2012	0,24489796	0,1244927	0,36530322	0	0,34594524
2013	0,34883721	0,20638472	0,4912897	0	0,46838711
2014	0,12195122	0,02178793	0,2221145	0	0,2060109

Taula 5.3.4. Proporció d'alumnes noies i càlcul dels intervals de confiança al 95% pels anys 2009 – 2014.

A la taula 5.3.4 s'observa que pel que fa a les noies, s'arriba a la conclusió contraria a la del nois, és a dir, la seva proporció es menor a 0.5 amb un nivell de confiança del 95%, ja que tant els intervals de confiança bilaterals com els unilaterals inclouen valors menors a aquest.

Igual que pel cas dels nois, l'any 2014 s'observa una gran diferència en la proporció de noies en comparació amb els altres, de forma més concreta aquesta proporció es va reduir en aproximadament la meitat. A més a més, també es pot veure aquesta desigualtat en el valor dels límits superiors bilaterals que es troben habitualment al voltant de 0.4, i en canvi, per a aquest any és de 0.22.

Finalment, un cop s'han realitzat tots els càlculs s'ha verificat que existeix una diferència estadísticament significativa entre la proporció dels estudiants del sexe masculí i la del sexe femení al grau de Matemàtiques, concretament, la proporció dels alumnes nois és major a la de les noies, és a dir, hi ha més homes que dones en aquesta carrera.

5.3.1. Comparació de l'evolució de les proporcions de nois i noies graduats

En l'apartat anterior s'ha pogut comprovar que efectivament hi ha diferències pel que fa als alumnes graduats de cada sexe del grau de Matemàtiques pels diferents anys. A continuació, es calcularà un test per a veure si aquestes diferències entre el nombre d'estudiants nois i noies es mantenen constants al llarg dels anys, és a dir, es comprovarà si les proporcions d'homes i dones són similars entre el 2009-2015. Per aconseguir aquest objectiu s'aplicarà una prova Chi-quadrat d'homogeneïtat. Per a poder aplicar-la primer es necessari comprovar que es compleixen les dues premisses del test:

- La mida mostral ha de ser prou gran: $N \geq 25$.
- La freqüència observada per a cada categoria no pot ser massa petita (almenys 3).

Les dades que s'usaran pel contrast, i que per tant han de complir aquestes condicions són:

Any d'entrada	Nois	Noies	Total
2009	26	12	38
2010	30	13	43
2011	35	12	47
2012	37	12	49
2013	28	15	43
2014	36	5	41
2015	10	1	11
Total	202	70	272

Taula 5.3.1.1. Nombre d'alumnes graduats per sexe pels anys 2009 – 2015.

En la taula 5.3.1.1 es pot comprovar que les dades dels anys 2009-2014 compleixen la segona condició, és a dir, les freqüències observades per a cada categoria són majors a 3, en canvi, les del 2015 no es poden usar per a aplicar el test, ja que per a una de les seves categories la freqüència és 1 (menor a 3). Per altra banda, la mida mostral total sí és prou gran (major a 25).

Les hipòtesi del test d'homogeneïtat són:

- $H_0: P_{2009H} = \cdots = P_{20141H} = P_H,$
  $P_{2009D} = \cdots = P_{2014D} = P_D$
- $H_1: P_{ij} \neq P_j \quad per\ algun\ i = 2009, \ldots, 2014 \quad j = H, D$
  $P_{ij}\ fa\ referència\ a\ la\ probabilitat\ de\ pertànyer\ a\ la\ categòria\ j\ (H = Home,$
  $D = Dona)\ a\ la\ població\ i.$

És a dir, es vol comprovar si la probabilitat d'home i dona són les mateixes per a totes les poblacions (cada una està formada per les dades d'un dels anys, 2009-2014) o si per el contrari alguna d'aquestes probabilitats és diferent.

Seguidament, es calcularan les dues taules de contingència per a construir l'estadístic del test. Cal destacar, que la taula de contingència de les freqüències esperades es calcularà amb ajuda de la funció *chisq.test* del paquet *stats* del programa *R* [27]. Taula de contingència dels efectius observats:

Any d'entrada	Nois	Noies	Total
2009	26	12	38
2010	30	13	43
2011	35	12	47
2012	37	12	49
2013	28	15	43
2014	36	5	41
Total	192	69	261

Taula 5.3.1.2. Taula de contingència dels efectius observats per a les dades del nombre d'alumnes graduats per sexe pels anys 2009 – 2014.

[27] A l'annex es poden veure tots els codis utilitzats: Codis d'R per a la comparació de proporcions de *nois i noies al grau de Matemàtiques* i test d'homogeneïtat.

Taula de contingència dels efectius esperats:

Any d'entrada	Nois	Noies	Total
2009	27,95402299	10,045977	38
2010	31,63218391	11,3678161	43
2011	34,57471264	12,4252874	47
2012	36,04597701	12,954023	49
2013	31,63218391	11,3678161	43
2014	30,16091954	10,8390805	41
Total	192	69	261

Taula 5.3.1.3. Taula de contingència dels efectius esperats per a les dades del nombre d'alumnes graduats per sexe pels anys 2009 – 2014.

Tal i com s'observa en aquestes dues taules, el nombre d'efectius observats i esperats és molt similar pels anys 2009 fins al 2012, tant pels nois com per les noies. En canvi, pel 2013 són lleugerament diferents i finalment, per a l'any 2014 són molt discrepants, exactament, si la hipòtesi nul·la fos certa i per tant, les proporcions dels dos sexes haurien de ser iguals per a tots els anys, s'esperaria que el nombre de nois pel 2014 fos de 30 i en canvi, realment hi va haver 36, pel cas de les noies s'esperarien 11, aproximadament, i només hi va haver 5.

Finalment, es calcularà l'estadístic del test amb el programa *R* [28]:

$$X^2 = 6.8041$$

$$\text{Sota } H_0\text{: } X^2 \sim \chi^2_{(6-1)x(2-1)},$$

Per tant, per a un nivell de significació $\alpha = 0.05$ es rebutjarà la hipòtesi nul·la si:

$$X^2 > \chi^2_5(\alpha) = 11.0705$$

En aquest cas, $X^2 = 6.8041 < \chi^2_5(\alpha) = 11.0705$, com a conseqüència es pot dir amb una confiança del 5% que la proporció d'homes i dones es manté similar durant els anys 2009-2014.

Per acabar, es pot dir que tot i que la proporció de nois i noies al grau de Matemàtiques és diferent estadísticament, exactament hi ha més alumnes del sexe masculí que del femení, aquesta es prou constant al llarg dels anys. Tot i així, cal esmentar que durant els dos últims anys les discrepàncies entre sexes han variat respecte als períodes anteriors, per tant, es possible que en un futur aquesta diferència no sigui similar a la que hi va haver durant el 2009-2012.

[28] Els codis usats es troben a l'annex: Codis d'R per a la comparació de proporcions de *nois i noies al grau de Matemàtiques* i test d'homogeneïtat.

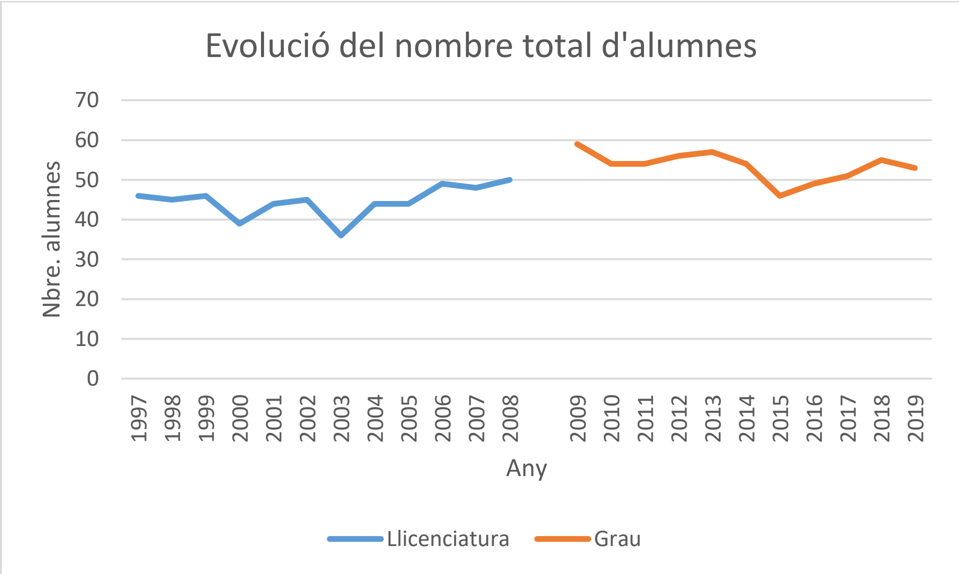
6. Evolució dels nois i noies al grau de Matemàtiques

En aquest apartat es farà una anàlisi descriptiu de les dades, seguidament es calcularà el test de Daniel per a comprovar si tenen una tendència i finalment s'aplicarà el mètode de la tendència lineal i es faran prediccions pel futur.

Cal destacar, que s'estudiaran dades des de l'any 1997 fins al 2019 del nombre de noies, nois i el total d'alumnes del grau de Matemàtiques, també del nombre d'estudiants que aconsegueixen superar la Fase Inicial del grau i per altra banda, del nombre de nois, noies i total d'estudiants que pertanyen al CFIS des de l'any 2001 al 2019.[29]

6.1. Anàlisi del nombre total d'alumnes

Per començar s'analitzarà la variable que indica el nombre total d'alumnes del grau. En el gràfic que es mostra a continuació es troba representada la seva evolució en el temps.



Gràfic 6.1.1. Evolució del nombre total d'alumnes a la llicenciatura i al grau de Matemàtiques durant els anys 1997-2019.

En el gràfic 6.1.1 s'observa que el nombre total d'alumnes ha anat variant durant els anys però, no es veu que hi ha una tendència clara durant el temps, sinó que aquest es manté més o menys constant en un interval entre 40 i 60 alumnes. Tot i així, durant els anys 1997-2003 el nombre d'estudiants va anar disminuint, fins a situar-se en el seu valor mínim al 2003 (36 alumnes), seguidament, fins al 2012 aquest va tendir a augmentar. A continuació, des del 2013 al 2015 el nombre d'estudiants va patir una lleugera disminució i finalment, ha augmentat fins a situar-se al voltant de 50 alumnes.

Tot i que en la representació 6.1 no es veu que aquestes dades tinguin una tendència es calcularà el test de Daniel per a comprovar-ho de forma analítica.

[29] Les dades es troben a l'annex: Dades apartat *6. Evolució dels nois i noies al grau de Matemàtiques.*

Cal esmentar, que es calcularan dos test, un pels anys de llicenciatura i un altre pels de grau, ja que com s'observa en el gràfic aquest fet va produir un canvi en la variable d'estudi, ja que el nombre d'alumnes va passar de 50 l'any 2008 a gairebé 60 al 2009.

Els càlculs necessaris s'han fet amb l'ajuda del programa Excel i a continuació es mostra el valor de l'estadístic del contrast:

Pel cas de la llicenciatura [30]:

$$T_s = 0{,}311188811, \quad Z = 1{,}032096526, \quad |Z| = 1{,}032096526$$

Com a conseqüència per a un nivell de significació $\alpha = 0.05$, $N_{\alpha/2} = 1{,}959963985$:

$$|Z| = 1{,}032096526 < N_{\alpha/2} = 1{,}959963985$$

Per tant, no es pot rebutjar la hipòtesi nul·la del test, és a dir, la sèrie temporal que formen els valors del nombre total d'alumnes pels anys 1997-2008 no té tendència.

Per altra banda, pel cas del grau [31]:

$$T_s = -0{,}477272727, \quad Z = -1{,}509268883, \quad |Z| = -1{,}509268883$$

A causa d'això, per a un nivell de significació $\alpha = 0.05$, $N_{\alpha/2} = 1{,}959963985$:

$$|Z| = 1{,}509268883 < N_{\alpha/2} = 1{,}959963985$$

Com a conseqüència, en aquest cas tampoc es pot rebutjar la hipòtesi nul·la del test, per tant, la sèrie temporal que formada pels valors del nombre total d'alumnes pels anys 2009-2019 no té tendència.

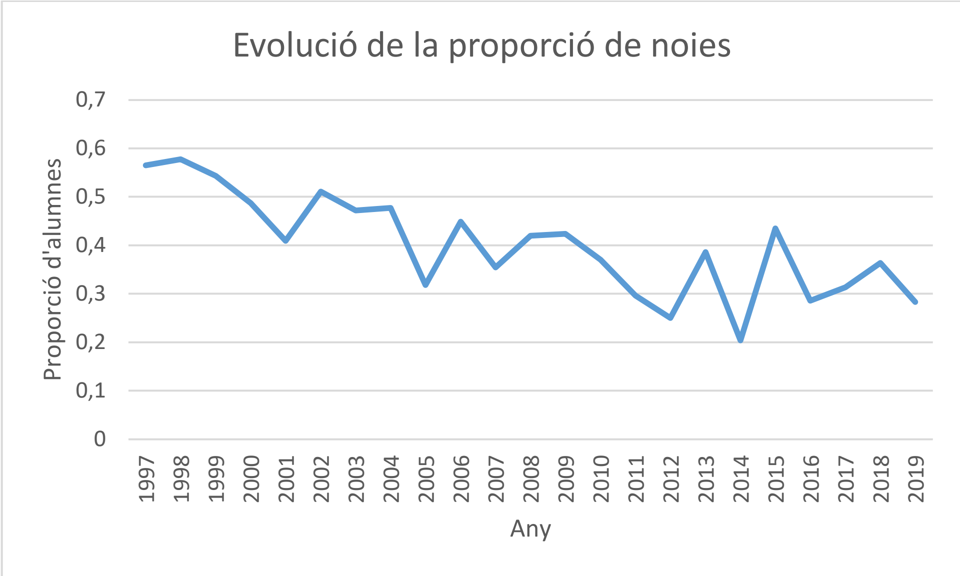
A causa d'aquests resultats no s'aplicarà el mètode de la tendència lineal, ja que no hi ha cap tendència per a estimar.

6.2. *Anàlisi de la proporció de noies*

Seguidament, s'analitzarà la sèrie temporal formada per les dades del nombre de noies al grau de Matemàtiques durant els anys 1997-2019, però, a causa de que les dades del nombre total d'alumnes no es mantenen totalment constants al llarg dels anys es creu que es més convenient estudiar-les en forma de proporció, és a dir, la proporció que els estudiants del sexe femení representen sobre el total. En el següent gràfic es mostra la seva evolució.

[30] Els càlculs es poden veure a l'annex: Test de Daniel per a la variable *nombre total d'alumnes a la llicenciatura de Matemàtiques.*

[31] A l'annex es troben els càlculs: Test de Daniel per a la variable *nombre total d'alumnes al grau de Matemàtiques.*



Gràfic 6.2.1. Evolució de la proporció d'alumnes del sexe femení al grau de Matemàtiques durant els anys 1997-2019.

En el gràfic 6.2.1 s'observa que pel que fa a la proporció de noies al grau sí existeix una tendència clara, concretament aquesta tendeix a anar disminuint durant els anys. Per començar, l'any 1997 hi havia un 56,52% de persones del sexe femení, seguidament fins l'any 2014 aquesta variable va anar baixant, al 2015 va tornar a augmentar, fins un 43,47%, i finalment, ha seguit disminuint fins que l'últim any el percentatge de dones pren un valor del 30% aproximadament. Per tant, aquesta variable disminueix en gairebé un 30% en el període d'estudi.

A continuació, es realitzarà el test de Daniel per tal de comprovar si hi ha una tendència de forma analítica. El valor de l'estadístic del contrast és [32]:

$$T_s = -0{,}793478261, \quad \mathrm{Z} = -3{,}72174294, \quad |\mathrm{Z}| = 3{,}72174294$$

A causa d'aquests resultats, per a un nivell de significació $\alpha = 0.05$:

$$|\mathrm{Z}| = 3{,}72174294 > N_{\alpha/2} = 1{,}959963985$$

Es pot concloure que la sèrie temporal formada per la proporció de noies al grau de Matemàtiques durant els anys 1997-2019 sí té tendència. Per tant, s'aplicarà el mètode de la tendència lineal per a poder estudiar com és aquesta i a més a més es faran noves prediccions pel futur.

Per aplicar aquest mètode s'han dividit les dades en dos grups, amb l'objectiu d'utilitzar una part d'elles per a valorar la capacitat predictiva del model. El període mostral el formaran les dades des del 1997 fins al 2015, i per altra banda el període extra-mostral per les del 2016 fins al 2019. S'ha escollit que el període extra-mostral estigués format només

[32] A l'annex es troben els càlculs realitzats amb Excel: Test de Daniel per a la variable *proporció d'alumnes noies al grau de Matemàtiques.*

per quatre observacions a causa de que es disposen de molt poques dades, i la majoria d'aquestes és millor usar-les per a construir el mètode.

Tal i com s'ha explicat en l'apartat *2.4 Anàlisi de sèries temporals* aquest mètode es basa en dur a terme una regressió lineal en la qual la variable independent és, en aquest cas, la proporció d'alumnes noies al grau durant els diferents anys i la variable dependent està formada per valors des de l'1 fins al 19 (nombre de dades del període mostral).

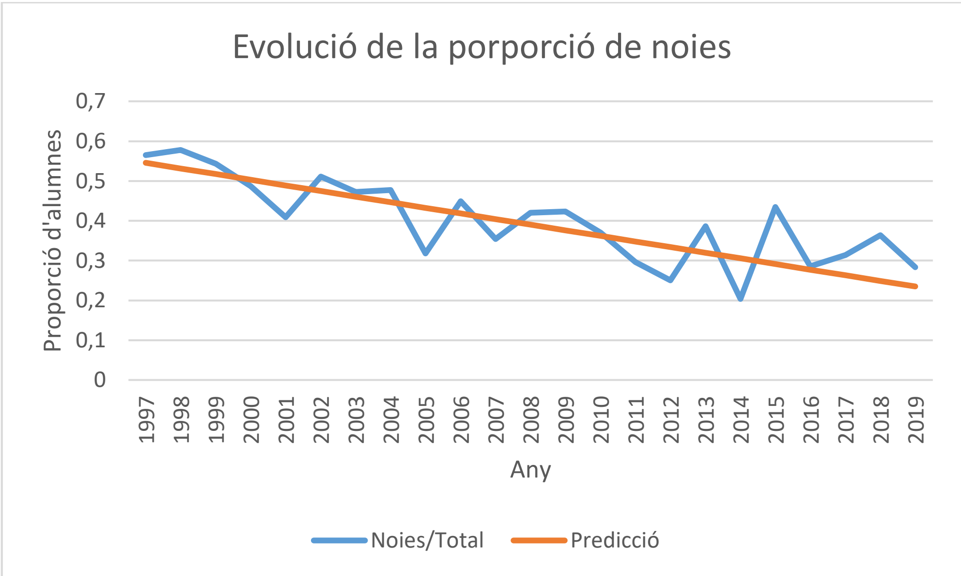
Els valors calculats amb el programa Excel[33], pels paràmetres del model són:

$\widehat{\beta_0} = 0{,}55950398$ $\widehat{\beta_1} = -0{,}014110794$

Pel que fa al valor de $\widehat{\beta_0}$ aquest indica la proporció d'alumnes noies estimada per a un any anterior a l'inicial (1996), i $\widehat{\beta_1}$ fa referència a la tendència de la sèrie, que en aquest cas, a causa de que pren un valor negatiu, és pot dir que de forma aproximada, cada any es perd un 1,41% d'estudiants del sexe femení en aquests estudis.

A partir d'aquests valors estimats es calcularan les prediccions del model per a tots els anys des del 1997 al 2019. A més a més, gràcies a les prediccions des del 2016 fins al 2019 i als valors reals d'aquestes dades es podrà calcular l'indicador anomenat Error Percentual Absolut Mitjà (*EPAM)*[34] que servirà per a conèixer quina és la capacitat predictiva del model.

En el gràfic que es mostra a continuació es troben representats els valors reals i les prediccions per a cada any.



Gràfic 6.2.2. Evolució de la proporció d'alumnes del sexe femení al grau de Matemàtiques i prediccions pel mètode de la tendència lineal durant els anys 1997-2019.

[33] A l'annex es pot veure el resultat d'aplicar la funció *regressió* sobre les dades d'estudi: Resultat d'aplicar la funció regressió sobre les dades de la *proporció de noies al grau de Matemàtiques* dels anys 1997-2015.

[34] La fórmula de l'EPAM es troba a l'apartat 2.4 Anàlisi de sèries temporals.

En el gràfic 6.2.2 es pot veure que la tendència estimada mitjançant el model és negativa, és a dir, la proporció de noies al grau Matemàtiques tendeix a anar disminuint amb els anys, mateixa conclusió a la qual s'havia arribat gràcies al valor estimat pel paràmetre β_1. Concretament, a partir del model es prediu que el percentatge de dones disminueixi en un 30% des del 1997 al 2019.

A continuació, es calcularà el valor de l'*EPAM* [35] per a poder valorar com de fiables són aquestes prediccions:

$$EPAM = 16{,}8876086\%$$

A causa de que l'*EPAM* és més elevat que el 5% es pot considerar que la capacitat predictiva del model és dolenta. Aquest valor segurament estigui causat pel fet de disposar de molt poques dades el que produeix que el model ajustat no sigui gaire precís, com a conseqüència, les prediccions que fa no són molt fiables, tot i així, aquest sí és útil per a aconseguir l'objectiu principal, és a dir, estimar la tendència de les dades.

Seguidament, es faran prediccions per a la proporció de noies pels períodes 2020-2025, però en aquest cas, per a que siguin el més precises possibles s'usaran per a aplicar el mètode de la tendència lineal totes les dades (des del 1997 fins al 2019). Els valors estimats [36] pels paràmetres són:

$$\widehat{\beta_0} = 0{,}543090824 \qquad \widehat{\beta_1} = -0{,}011940106$$

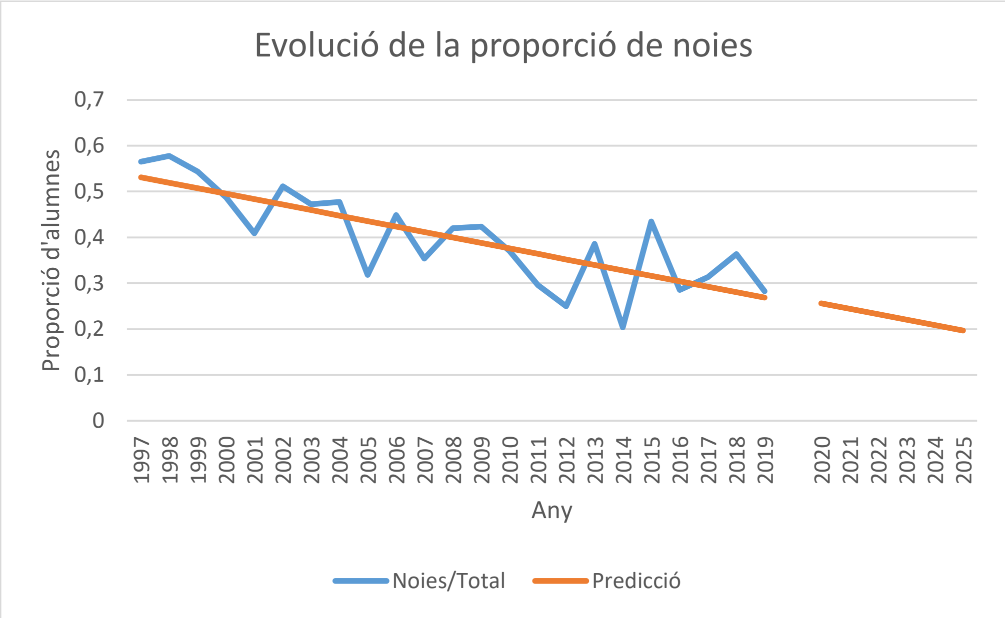
Per començar, tot i que els valors estimats pels paràmetres $\widehat{\beta_0}$ i $\widehat{\beta_1}$ són prou similars als que s'obtenien usant menys dades, s'ha d'esmentar que, per una banda, pel que fa a $\widehat{\beta_0}$ aquest és lleugerament inferior i en canvi, $\widehat{\beta_1}$ és superior, però, alhora d'analitzar aquests valors s'arriba a una conclusió semblant, és a dir, la proporció de noies disminueix a mesura que passen els anys, en aquest cas, en un 1,2% aproximadament.

En el gràfic següent es mostren els valors de les dades reals des del 1997 fins al 2019 i les prediccions pels anys 1997-2025 [37].

[35] Els càlculs per a l'EPAM es troben a l'annex: Prediccions resultants d'aplicar el mètode de la tendència lineal considerant les dades des del 1997-2015 com a període mostral i les del 2016-2019 com a extra-mostral i càlcul de l'EPAM per a la proporció de noies.

[36] A l'annex es pot veure el resultat d'aplicar la funció *regressió* d'Excel: Resultat d'aplicar la funció regressió sobre les dades de la *proporció de noies al grau de Matemàtiques* dels anys 1997-2019.

[37] Els càlculs es troben a l'annex: Prediccions pels anys 1997-2025 resultants d'aplicar el mètode de la tendència lineal usant les dades des del 1997-2019 de la proporció de noies.



Gràfic 6.2.3. Evolució de la proporció d'alumnes del sexe femení al grau de Matemàtiques i prediccions pel mètode de la tendència lineal durant els anys 1997-2025.

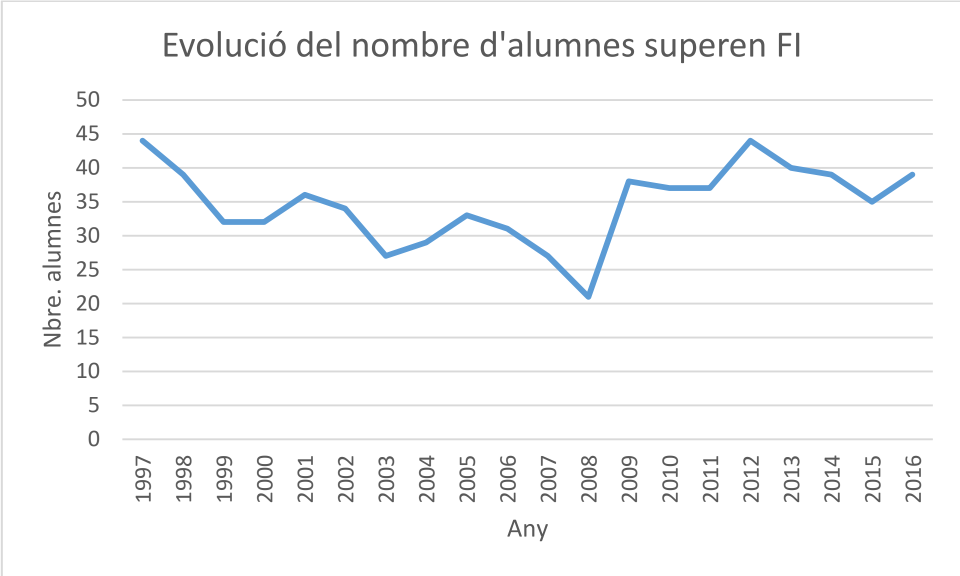
Com s'observa en el gràfic 6.2.3 s'espera que la proporció de noies disminueixi en els pròxims anys si la tendència que hi hagut durant el 1997 fins al 2019 es manté. De forma més concreta, s'espera que l'any 2025 les noies representin un 20% del total d'estudiants del grau de Matemàtiques, valor que és quasi un 30% menys al que el model prediu per l'any 1997. Tot i així, cal tenir en compte que degut al valor de l'*EPAM* no es pot considerar que aquestes prediccions realitzades siguin fiables.

Per acabar, s'ha destacar que l'evolució del percentatge de noies al grau de Matemàtiques és molt diferent a la dels nois, ja que aquest creix al llarg del temps, en canvi pel que fa a les dones a mesura que passen els anys van perdent presència en aquests estudis, és a dir, mentre que a l'inici de la sèrie temporal les noies representaven el mateix percentatge que els nois, a mesura que augmenten els anys, aquests passen a ser la major part del total d'estudiants.

6.3. Anàlisi de la proporció d'alumnes que superen la fase inicial

Seguidament, s'analitzarà la variable que fa referència al nombre d'alumnes que aconsegueixen superar la fase inicial del grau de Matemàtiques de l'any 1997 al 2016.

Tot i que es disposen de dades fins al 2019 s'ha escollit considerar només fins al 2016, ja que tal i com es mostra en la normativa del grau superar aquesta fase vol dir aprovar les assignatures corresponents al primer curs en un màxim de quatre anys, per tant, els alumnes dels anys 2017 en endavant encara tenen temps per a poder superar-la. En el següent gràfic es troba representada la seva evolució.



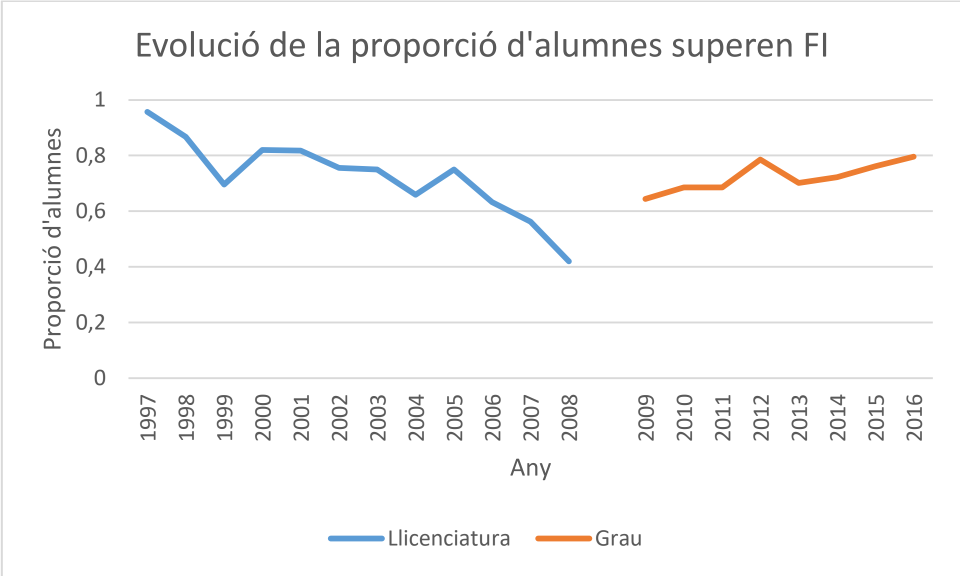
Gràfic 6.4.1. Evolució del nombre d'alumnes que superen la fase inicial de la llicenciatura i el grau de Matemàtiques durant els anys 1997-2016.

En el gràfic 6.4.1 no s'observa que hi hagi una tendència clara pel que fa al nombre d'alumnes que superen la fase inicial. Per començar, des de l'any 1997 al 2008 aquesta variable tendeix a anar disminuint amb el temps, exactament, va passar d'haver-hi 44 alumnes que superen aquesta fase, al 1997 a només 21, al 2008, en canvi, al 2009 aquesta va patir un augment destacable (de 21 a 38 estudiants). Durant els anys 2009-2011 el nombre d'alumnes que supera la fase inicial es manté prou constant, a continuació creix fins al 2010, període en el qual comença a disminuir, tot i que en el 2016 s'ha produït un lleuger augment.

S'ha d'esmentar que l'any 2008 va haver un canvi en aquests estudis, ja que van passar de ser una llicenciatura a ser un grau, per tant, possiblement aquesta ha sigut la causa principal del augment d'alumnes que superen la fase inicial que es va produir al 2009. A més a més, a partir d'aquest any es trenca la tendència a disminuir que tenia la sèrie temporal i passa a créixer, com a conseqüència d'aquest efecte es tindrà en compte el fet de ser llicenciatura o grau a l'hora d'analitzar les dades.

Cal destacar, que aquesta variable depèn del nombre d'alumnes que té el grau, com a conseqüència els anys que hi ha hagut menys estudiants el nombre dels que superen aquesta fase també ha disminuït. A causa d'això, és més interessant estudiar la proporció d'alumnes que superen la fase inicial sobre el total.

En el gràfic que es pot veure a continuació es troba representada l'evolució de la proporció d'alumnes que superen la fase inicial sobre el total per a cada any entre 1997 i 2016.



Gràfic 6.3.2. Evolució de la proporció d'alumnes que superen la fase inicial de la llicenciatura i el grau de Matemàtiques durant els anys 1997-2016.

En el gràfic 6.3.2 es veu clarament que existeixen dues tendències ben diferenciades pel que fa a la proporció d'alumnes que superen la fase inicial. Primer, des de l'any 1997 al 2008 el nombre d'alumnes que superava aquesta fase sobre el total va anar disminuint amb el temps i seguidament, a partir d'aquest any fins al 2016 passa a augmentar lleugerament.

En el gràfic 6.3.1 també es veia aquest canvi però no era tan marcat, sinó que semblava que la sèrie a partir de l'any 2009 seguia tenint petites variacions.

A continuació, per a comprovar l'existència d'una tendència de forma analítica es faran dos test de Daniel, un per a cada grup de dades: el primer pels anys de llicenciatura i el segon pels anys de grau, ja que degut a la influència d'aquest fet no tindria sentit estudiar les dades de forma conjunta.

El valor de l'estadístic del contrast pel primer cas és [38]:

$$T_s = -0{,}884615385, \quad Z = -2{,}933937315, \quad |Z| = 2{,}933937315$$

Per tant, per a un nivell de significació $\alpha = 0.05$:

$$|Z| = 2{,}933937315 > N_{\alpha/2} = 1{,}959963985$$

A causa d'aquest resultat es pot dir que la sèrie temporal formada per la proporció d'alumnes que superen la fase inicial del grau de Matemàtiques durant els anys 1997-2008 sí té tendència.

Pel segon cas, el valor de l'estadístic del test és [39]:

$$T_s = 0{,}845238095, \quad Z = 2{,}236289799, \quad |Z| = 2{,}236289799$$

[38] A l'annex es poden veure els càlculs realitzats amb Excel: Test de Daniel per a la variable *proporció d'alumnes que superen la fase inicial de la llicenciatura de Matemàtiques dels anys 1997-2008.*

[39] Els càlculs realitzats amb Excel es troben a l'annex: Test de Daniel per a la variable *proporció d'alumnes que superen la fase inicial del grau de Matemàtiques dels anys 2009-2016.*

Com a conseqüència, per a un nivell de significació $\alpha = 0.05$:

$$|Z| = 2{,}236289799 > N_{\alpha/2} = 1{,}959963985$$

Degut al resultat anterior es pot concloure que la sèrie temporal formada per la proporció d'alumnes que superen la fase inicial del grau de Matemàtiques durant els anys 2009-2016 també té tendència.

Seguidament, s'aplicarà el mètode de la tendència lineal per estimar la tendència de les dades dels anys 1997 fins al 2008, i per a la del 2009 al 2016. És a dir, es faran dues regressions lineals, una en la qual la variable independent serà la proporció d'alumnes que superen la fase inicial durant els anys 1997-2008 i la variable dependent estarà formada per valors des de l'1 fins al 12 (nombre d'observacions d'aquest grup), i en l'altre l'independent serà la proporció d'aquests pels anys 2009-2016 i la dependent la formaran valors de l'1 al 8. En el primer cas, no es necessari valorar la capacitat predictiva del mètode, ja que no s'usarà per a fer prediccions per a valors futurs a causa de que les dades fan referència a un període passat que en l'actualitat no existeix, per tant s'usaran totes per a aplicar el mètode, és a dir, totes formaran el període mostral i no hi haurà període extra-mostral, pel segon cas tampoc hi haurà període extra-mostral ja que no es faran prediccions a causa de que es disposa de molt poques dades.

Un cop s'ha aplicat el mètode s'han obtingut els següents valors estimats pels paràmetres del model, pel cas de la llicenciatura, amb l'ajuda d'Excel [40]:

$\widehat{\beta_0} = 0{,}96778632$ $\widehat{\beta_1} = -0{,}040374357$

Pel que fa al valor de $\widehat{\beta_0}$ fa referència a la proporció d'alumnes sobre el total que superen la fase inicial per a un any anterior a l'inicial, el 1996. El paràmetre $\widehat{\beta_1}$ indica la tendència de la sèrie, en aquest cas és negativa, és a dir, a mesura que passa el temps disminueix la proporció d'estudiants que superen aquesta fase, aproximadament en un 4% per any de llicenciatura.

Per altra banda, pel cas del grau, els paràmetres estimats pel model són[41]:

$\widehat{\beta_0} = 0{,}604637726$ $\widehat{\beta_1} = 0{,}013840923$

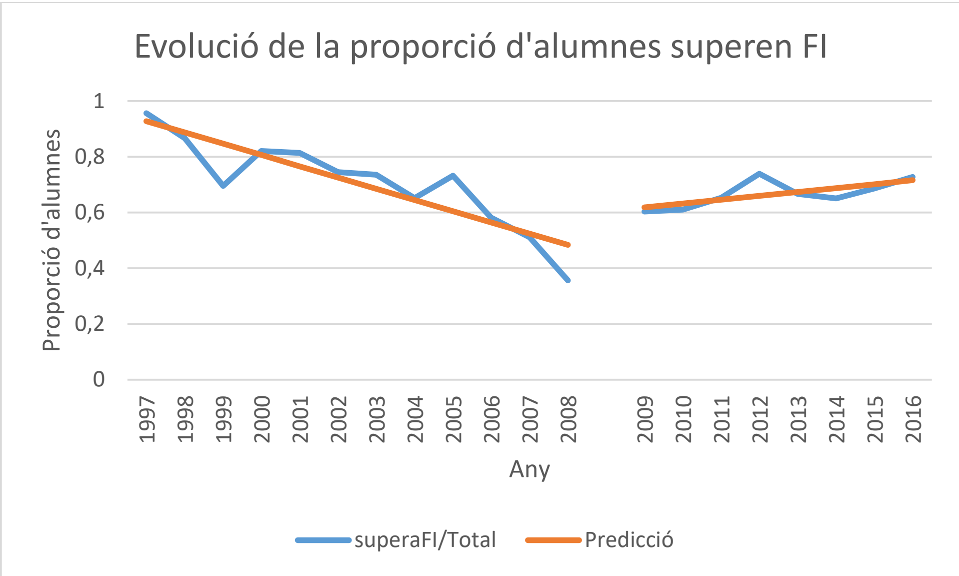
En aquest cas, $\widehat{\beta_0}$ indica la proporció d'alumnes estimada sobre el total que superen la fase inicial per a l'any 2008 i $\widehat{\beta_1}$ fa referència a la tendència de la sèrie, que a diferència del cas anterior, aquesta pren un valor positiu, per tant, la proporció d'alumnes que superen aquesta fase augmenta a mesura que passa el temps, concretament, en gairebé un 1,38% per any de grau.

[40] A l'annex es troba el resultat d'aplicar la funció *regressió* sobre les dades del 1997 al 2008: Resultat d'aplicar la funció regressió sobre les dades de la *proporció d'alumnes que superen la fase inicial de la llicenciatura de Matemàtiques* dels anys 1997-2008.

[41] A l'annex es troba el resultat d'aplicar la funció *regressió* sobre les dades del 2009 al 2015: Resultat d'aplicar la funció regressió sobre les dades de la *proporció d'alumnes que superen la fase inicial del grau de Matemàtiques* dels anys 2009-2016.

Cal comentar, que un cop realitzat l'anàlisi s'ha vist clarament que el fet d'haver passat de la llicenciatura al grau ha produït un canvi en l'evolució de la sèrie temporal estudiada. A més a més, aquest canvi ha estat beneficiós pels alumnes, ja que s'ha trencat amb la tendència que hi havia a cada cop fossin més els estudiants que no aconseguien superar la fase inicial i per tant eren obligats a abandonar forçosament els estudis de Matemàtiques. Per altra banda, actualment gràcies al nou pla d'estudis el nombre d'alumnes que supera aquesta fase sobre el total augmenta de forma suau al llarg del temps.

A continuació, es mostra un gràfic en el que es troba representada l'evolució de la proporció d'alumnes que superen la fase inicial pels anys 1997 fins al 2016 i les prediccions pel 1997 fins al 2016 [42].



Gràfic 6.3.3. Evolució de la proporció d'alumnes que superen la fase inicial i prediccions pel mètode de la tendència lineal durant els anys 1997-2016.

En el gràfic 6.3.3 es pot veure que la tendència estimada pel model és negativa pels primers 12 anys, és a dir, la proporció d'alumnes tendeix a anar disminuint amb el temps fins al 2008, en l'any següent passa a augmentar i finalment, a partir del 2009 es manté lleugerament creixent, tal i com s'havia vist gràcies als dos valors estimats pel paràmetre β_1, concretament segons el model es prediu una disminució del 45% aproximadament entre els anys 1997 i 2008, i un augment del 10% entre el 2009 i el 2016.

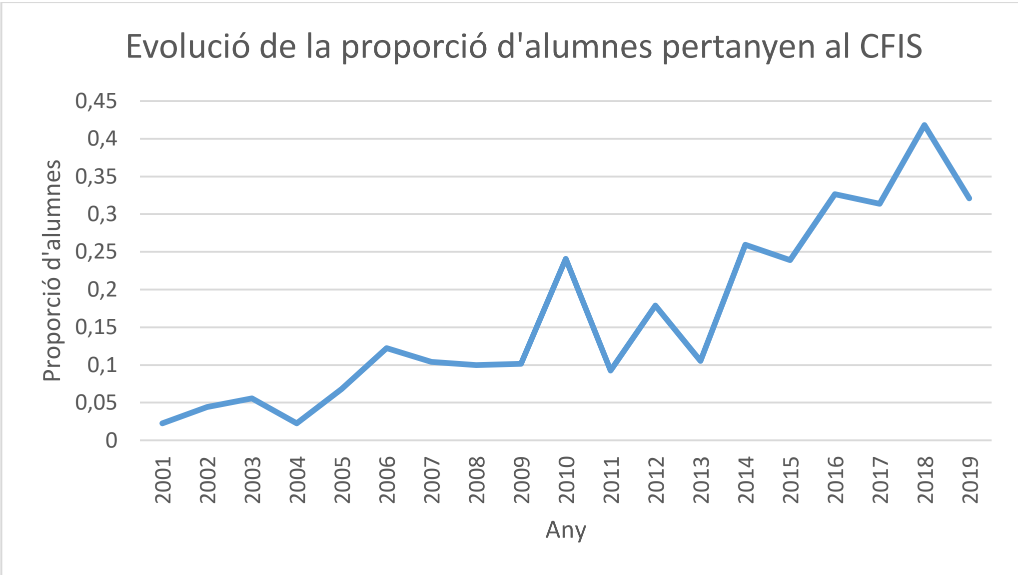
Per acabar, cal esmentar que tot i els resultats obtinguts, pot ser que en el futur es torni a produir un canvi en el pla d'estudis i com s'ha vist aquesta variable està fortament relacionada amb aquest fet, com a conseqüència es podria tornar a produir, també, un canvi en la tendència. Tot i així, és molt possible que durant el futur més pròxim la proporció

[42] Els càlculs realitzats es troben a l'annex: Prediccions pels anys 1997-2016 resultants d'aplicar el mètode de la tendència lineal usant les dades des del 1997-2016 de la proporció d'alumnes que supera la fase inicial.

d’alumnes sobre el total que supera aquesta fase es mantingui creixent si la tendència que hi hagut a partir de l’any 2009 es manté.

6.4. Anàlisi de la proporció d’alumnes que pertany al CFIS

A continuació, s’analitzaran les dades referents al nombre d’alumnes que pertany al CFIS, però a causa de que aquesta variable es troba relacionada amb el nombre total d’estudiants del grau es creu més convenient calcular la proporció que aquests representen sobre el total per a cada un dels anys. En el gràfic següent es mostra la seva evolució.



Gràfic 6.5.1. Evolució de la proporció d’alumnes que pertany al CFIS del grau de Matemàtiques durant els anys 2001-2019.

En el gràfic 6.5.1 es pot veure que clarament existeix una tendència pel que fa a l’evolució de la proporció d’alumnes que pertany al CFIS. Concretament, aquesta tendeix a anar augmentant a mesura que passen els anys. Per començar, des de l’any 2001 al 2009 el creixement en la proporció d’estudiants que pertanyen al CFIS va ser molt suau, a continuació, al 2010 es va produir un augment destacable, però seguidament va tornar a disminuir, fins que a partir de l’any 2013 el creixement d’aquesta variable s’ha accelerat. S’ha de destacar que en el període d’estudi, aquesta proporció passa de representar un 2,3% aproximadament, al 2001, a un 32,1% al 2019.

Cal esmentar que, possiblement, la causa principal de l’augment en la proporció d’alumnes que pertany al CFIS durant els últims anys ha estat causat pel fet de que cada cop més els estudis que proporciona aquest centre són més coneguts entre els estudiants.

Seguidament, es calcularà el test de Daniel per poder comprovar si existeix una tendència de forma analítica. El valor de l’estadístic del contrast és [43]:

$$T_s = 0{,}908333333, \quad Z = 3{,}853731957, \quad |Z| = 3{,}853731957$$

[43] A l’annex es troben els càlculs realitzats amb Excel: Test de Daniel per a la variable *proporció d’alumnes que pertany al CFIS del grau de Matemàtiques.*

Per tant, per a un nivell de significació $\alpha = 0.05$:

$$|Z| = 3{,}853731957 > N_{\alpha/2} = 1{,}959963985$$

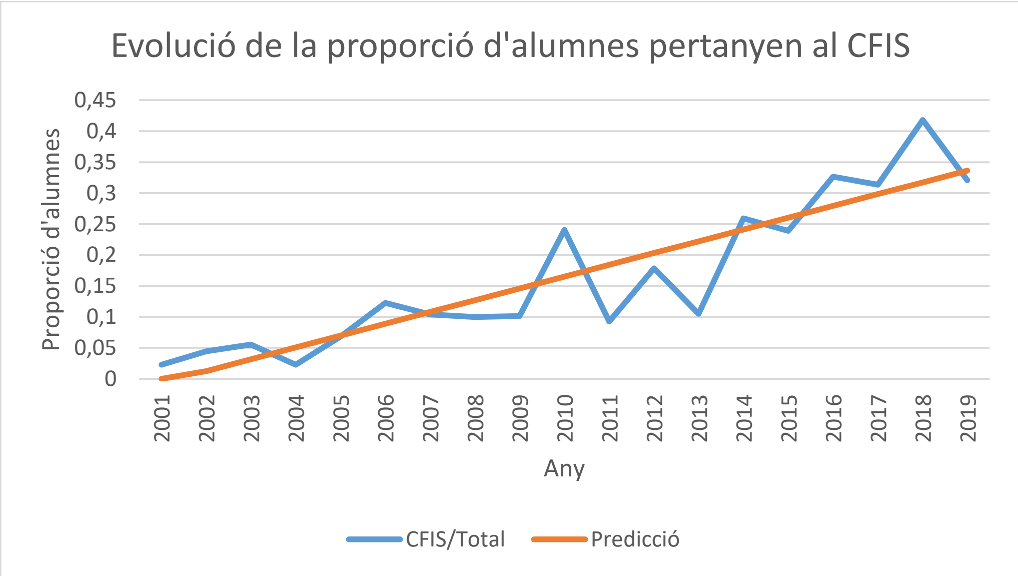
A causa d'aquests resultats es pot dir que la sèrie temporal que formen els valors de la proporció d'estudiants que pertany al CFIS durant els anys 2001 al 2019 sí té tendència. Com a conseqüència s'aplicarà el mètode de la tendència lineal per a poder estimar-la, i a més a més poder fer prediccions de valors futurs.

En aquest cas, s'aplicarà també el mètode de la tendència lineal, per a poder estimar la tendència de la sèrie temporal estudiada. Els paràmetres obtinguts aplicant una regressió lineal amb ajuda d'Excel, en la qual la variable independent és la proporció d'alumnes que pertany al CFIS i la dependent són valors de l'1 al 19, són [44]:

$\widehat{\beta_0} = -0{,}025357513$ $\quad\quad\quad\quad$ $\widehat{\beta_1} = 0{,}019044684$

Pel que fa referència a $\widehat{\beta_1}$ es pot dir que la proporció d'alumnes que pertany al CFIS augmenta a mesura que passen els anys, ja que pren un valor positiu. De forma més concreta en un 1,9% per any.

A continuació, es calcularan les prediccions del model pels anys des del 2001 al 2019 [45]. En el següent gràfic es troben representades les prediccions i els valors reals per a cada un dels períodes.



Gràfic 6.5.2. Evolució de la proporció d'alumnes que pertany al CFIS del grau de Matemàtiques i prediccions pel mètode de la tendència lineal durant els anys 2001-2019.

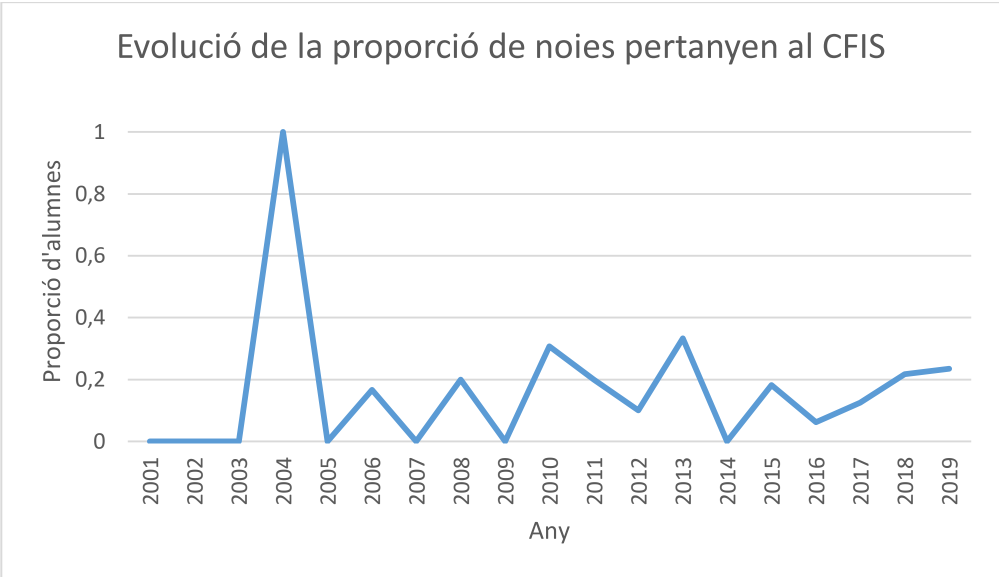
[44] A l'annex es pot veure el resultat d'aplicar la funció *regressió* sobre les dades d'estudi: Resultat d'aplicar la funció regressió sobre les dades de la *proporció d'alumnes que pertany al CFIS del grau de Matemàtiques* dels anys 2001-2019.

[45] Els càlculs es troben a l'annex: Prediccions pels anys 2001-2019 resultants d'aplicar el mètode de la tendència lineal usant les dades des del 2001-2019 del nombre d'alumnes que pertany al CFIS.

En el gràfic 6.5.2 s'observa que la tendència estimada pel model és positiva, per tant, la proporció d'alumnes que pertany al CFIS augmenta a mesura que passa el temps, tal i com s'havia conclòs gràcies al valor estimat pel paràmetre β_1. Concretament, mitjançant el model es prediu que aquesta proporció creixi en gairebé un 30% del 2001 al 2019. A causa d'aquests resultats, s'espera que la proporció d'alumnes que pertany al CFIS augmenti en els pròxims anys si la tendència que hi hagut durant el 2001-2019 es manté.

6.5. Anàlisi de la proporció de noies que pertany al CFIS

Finalment, s'estudiarà la sèrie temporal formada per les dades del nombre d'alumnes del sexe femení que pertany al CFIS durant els anys 2001-2019, però, a causa de que aquesta variable es troba fortament relacionada amb el nombre total d'alumnes que forma part del CFIS, i aquest no es manté constant al llarg del temps es creu que és millor estudiar-les en forma de proporció sobre el total. En el gràfic següent es pot observar la seva evolució.



Gràfic 6.5.1. Evolució de la proporció d'alumnes del sexe femení que pertany al CFIS del grau de Matemàtiques durant els anys 2001-2019.

Tal i com es pot veure en el gràfic 6.5.1 l'evolució de la proporció de noies que pertany el CFIS no té una tendència clara al llarg del temps, sinó que tendeix a mantenir-se constant en un interval entre 0% i 30%, aproximadament. Tot i això, cal destacar que l'any 2004, es produeix un augment molt rellevant, ja que les estudiants del sexe femení que pertanyen al CFIS passen de representar un 0% al 2003, a un 100% l'any següent.

Per tant, es pot concloure que l'evolució de les dones que formen part d'aquest centre és molt diferent a la del nois, ja que aquests acostumen a representar la major part dels estudiants, en canvi, les noies només representen gairebé un terç d'aquests. Tot i així, com s'observa en la representació 6.7.1 durant els últims períodes, 2017-2019 la proporció de noies està començant a augmentar lleugerament, a causa d'això és possible que en els pròxims anys les alumnes del sexe femení representin un nombre més elevat, en comparació amb l'actual.

Per acabar, es realitzarà el test de Daniel per a verificar de forma analítica l'existència d'una tendència. El valor de l'estadístic del contrast, en aquest cas, és [46]:

$$T_s = 0{,}422368421, \quad \mathrm{Z} = 1{,}791957448, \quad |\mathrm{Z}| = 1{,}791957448$$

A causa d'aquests resultats, per a un nivell de significació $\alpha = 0.05$:

$$|\mathrm{Z}| = 1{,}791957448 < N_{\alpha/2} = 1{,}959963985$$

Com a conseqüència, la sèrie temporal que formen els valors de la proporció de noies que pertany al CFIS del grau de Matemàtiques durant els anys 2001-2019 no té tendència. Per tant, a causa d'aquest resultat no s'aplicarà el mètode de la tendència lineal.

[46] A l'annex es troben els càlculs realitzats amb Excel: Test de Daniel per a la variable *proporció d'alumnes noies que pertany al CFIS del grau de Matemàtiques.*

7. Conclusions

Un cop realitzat l'estudi s'han arribat a diverses conclusions. Per una banda, s'ha vist que sí existeixen diferències entre sexes al grau de Matemàtiques de la Universitat Politècnica de Catalunya, però s'ha comprovat que aquestes no afecten a tots els àmbits d'aquests estudis. Per començar, ja es troben diferències entre el nombre d'alumnes de cada sexe que accedeix a la carrera, tot i així, aquesta no sempre ha estat la mateixa, sinó que a mesura que passen els anys augmenta el percentatge que representen els estudiants del sexe masculí, i per tant, disminueix el de les del sexe femení. A més a més, pel que fa als alumnes que pertanyen al CFIS també s'observen discrepàncies entre les proporcions dels dos sexes, tot i que aquestes s'acostumen a mantenir constants al llarg del temps, de forma més concreta, el nombre de nois supera de forma rellevant el de noies que hi pertanyen.

A continuació, pel que fa a la fase inicial del grau, que és una de les barreres d'entrada que es troben aquests estudiants i la qual han de superar per a poder seguir-los cursant, no es detecten diferències entre sexes, és a dir, aquest no té una influència en el fet d'aprovar-la o no. L'únic factor que sí té afecte és la nota amb la qual els alumnes accedeixen al grau, és a dir, aquells que accedeixen amb notes elevades (per tant, amb un nivell més alt) tenen més possibilitats de superar-la.

Pel que fa referència a la nota final, el sexe sí té una influència. Concretament, les noies acostumen a tenir notes més baixes que els del nois. Tot i així, la diferència causada per aquest factor no és gaire destacable (0,12 punts), sinó que són altres variables les que determinen en major part el valor d'aquesta nota. És a dir, el fet de ser home o dona no provoca tenir una nota més elevada o més baixa, en canvi, sí ho fa el rendiment de cada estudiant a les diverses assignatures, i cal destacar que aquest no depèn del sexe. Finalment, pel que fa al nombre d'alumnes graduats, sí hi ha discrepàncies entre els dos sexes. De forma més concreta, la proporció de graduats homes supera normalment a la de dones, i a més a més, aquesta diferència s'acostuma a mantenir a mesura que passen els anys.

Pel que respecta a l'evolució dels alumnes al grau, s'ha d'esmentar que aquests estudis tenen una gran rellevància, ja que la majoria d'anys aconsegueixen omplir totes les seves places. Per altra banda, en referència al nombre d'estudiants que supera la fase inicial s'ha observat que el fet d'haver passat de ser una llicenciatura a un grau ha provocat que la proporció dels que la superen abandoni la seva tendència a disminuir amb el temps i passi a augmentar, per tant, es pot considerar que aquest canvi ha estat beneficiós pels alumnes.

Per acabar, es pot concloure que el fet de ser home o dona no té un impacte rellevant en l'èxit en aquesta carrera. Tot i així, si hi ha diverses diferències entre aquests dos. Per altra banda, actualment, el nombre de nois sí supera al de dones, tot i que aquesta discrepància ja ve provocada des de l'inici del grau, ja que en el últims anys el nombre de persones del sexe femení que accedeix a cursar aquests estudis és menor al del sexe masculí, és a dir, les diferències entre sexes no apareixen a mesura que passen els anys de la carrera, sinó que ja es troben des del principi. Per tant, a causa d'aquests resultats per a disminuir les diferències existents caldria aplicar mesures que incentivessin a les noies a triar de forma més freqüent el grau de Matemàtiques.

8. Annexos

Dades dels apartats 4. Anàlisi de la nota final dels estudiants del grau de Matemàtiques i 5. Anàlisi de la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres.

id	any_entrada	CFIS	nota_acces	nota_acces_min_curs	provaNivell	sexe	nombre_quad_FI	nota_Fonaments	nota_AlgLineal	...	nota_ModelsMatematicsTecnologia	nombre_assigs_aprovades_1	nota_final
1	2009	NO	9,620	6,830	8,25	DONA	2	9,00	6,10	...	9,00	25,00	8,003
2	2009	NO	8,610	6,830	5,27	DONA	4	5,10	6,60	...	7,00	17,00	6,831
3	2009	NO	7,760	6,830		HOME	2	7,00	6,20	...	7,10	24,00	6,594
4	2009	SI	8,930	6,830		HOME	3	10,00	8,00	...	8,50	25,00	8,57
5	2009	NO	9,570	6,830	7,36	DONA	2	8,00	7,00	...	7,00	25,00	8,379
6	2009	NO	9,600	6,830		HOME	4	8,50	6,70	...	10,00	20,00	7,57
7	2009	SI	7,830	6,830		HOME	4	9,50	5,00	...	6,80	25,00	7,169
8	2009	SI	8,470	6,830		HOME	3	9,00	5,00	...	7,00	20,00	7,303
9	2009	SI	9,750	6,830		HOME	2	10,00	9,50	...	8,10	25,00	9,235
10	2009	SI	8,290	6,830	7,09	HOME	2	6,50	7,30	...	6,90	25,00	7,632
11	2009	NO	9,230	6,830	8,42	DONA	2	8,50	7,30	...	7,00	25,00	8,201
12	2009	NO	7,740	6,830	6,22	HOME	4	5,00	5,00	...	9,00	19,00	6,083
13	2009	NO	7,100	6,830	5,6	HOME	2	5,50	5,00	...	8,10	24,00	6,497
14	2009	NO	8,170	6,830	5,43	HOME	4	5,50	8,50	...	7,00	21,00	7,5
15	2009	SI	7,580	6,830	5,18	HOME	2	5,10	5,40	...	7,00	25,00	7,427
16	2009	NO	7,130	6,830	6,53	HOME	3	5,90	6,80	...	7,00	15,00	6,213
17	2009	NO	8,950	6,830	8,68	HOME	2	7,40	5,00	...	8,00	25,00	7,058
18	2009	NO	7,730	6,830	8,85	HOME	3	8,00	8,40	...	7,00	14,00	7,612
19	2009	NO	9,210	6,830	7,68	HOME	2	7,00	6,60	...	9,00	25,00	7,688
20	2009	NO	8,620	6,830	5,77	DONA	4	5,70	6,80	...	7,50	15,00	6,691
...	...	...	...	...	...	...	...	...	...	...	...	...	...
263	2015	NO	12,684	11,894	6,9	HOME	2	9,00	8,00	...	8,20	25,00	7,678
264	2015	NO	13,725	11,894	7,24	HOME	2	7,00	6,30	...	8,10	25,00	7,209
265	2015	NO	12,868	11,894	7,07	HOME	2	7,00	7,50	...	10,00	25,00	7,913
266	2015	NO	12,878	11,894	6,38	DONA	2	6,20	9,00	...	7,40	25,00	7,474
267	2015	NO	12,914	11,894	8,1	HOME	3	6,30	5,30	...	6,00	23,00	6,399
268	2015	NO	12,796	11,894	5,52	HOME	2	8,00	7,20	...	9,50	25,00	7,393
269	2015	NO	12,950	11,894	6,55	HOME	2	6,40	5,30	...	8,40	25,00	7,094
270	2015	SI	12,500	11,894		HOME	3	9,50	8,20	...	8,00	25,00	7,869
271	2015	SI	13,505	11,894		HOME	4	5,00	7,10	...	7,50	22,00	7,418
272	2015	NO	12,584	11,894	7,76	HOME	2	6,40	6,50	...	6,00	25,00	6,794

Codi d'R per a l'anàlisi descriptiu de la base de dades dels apartats 4. Anàlisi de la nota final dels estudiants del grau de Matemàtiques i 5. Anàlisi de la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres.

```
library(readxl)
dades_graduats <- read_excel("dades_graduats.xlsx", col_types = c("numeric", "numeric",
"text", "numeric", "numeric", "numeric", "text", "numeric", "numeric", "numeric", "numeric",
"numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric",
"numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric",
"numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric"))
dd <- as.data.frame(dades_graduats)
dd$CFIS <- factor(dd$CFIS)
dd$sexe <- factor(dd$sexe, levels = levels(factor(dd$sexe))[c(2,1)])

#Any entrada
table(dd$any_entrada)
barplot(table(dd$any_entrada), main = "Alumnes segons l'any d'entrada", xlab = 'Any', ylab =
'Nbre. alumnes', col = colors()[44])
barplot(table(dd$sexe, dd$any_entrada), main = "Alumnes segons l'any d'entrada", xlab =
'Any', ylab = 'Nbre. alumnes', col = colors()[c(28,44)])
legend ("topright", c("Home", "Dona"), col = colors()[c(28,44)], pch = 16)
prop.table(table(dd$sexe, dd$any_entrada))

#CFIS
table(dd$CFIS)
pie(table(dd$CFIS), col = colors()[c(28,44)], main = "Alumnes CFIS")
barplot(table(dd$sexe, dd$CFIS), col = colors()[c(28,44)], ylab = "Nbre. alumnes", main =
"Estudiants que pertanyen al CFIS")
legend ("topright", c("Home", "Dona"), col = colors()[c(28,44)], pch = 16)
prop.table(table(dd$sexe, dd$CFIS))

#Nota_accés
dd$dif_acces_min <- dd$nota_acces - dd$nota_acces_min_curs
summary(dd$dif_acces_min)
hist(dd$dif_acces_min)
with(dd[dd$sexe == "HOME",], hist(dif_acces_min, col = colors()[44], ylab = "Freqüència",
xlab = "Diferència notes", main = "Diferència entre nota i nota mínima d'accés pels nois"))
with(dd[dd$sexe == "DONA",], hist(dif_acces_min, col = colors()[44], ylab = "Freqüència",
xlab = "Diferència notes", main = "Diferència entre nota i nota mínima d'accés per les noies"))
with(dd[dd$sexe == "HOME",], summary(dif_acces_min))
with(dd[dd$sexe == "DONA",], summary(dif_acces_min))

#Prova Nivell
summary(dd$provaNivell)
hist(dd$provaNivell)
```

```
with(dd[dd$sexe == "HOME",], hist(provaNivell, col = colors()[44], ylab = "Freqüència", xlab = "Nota", main = "Notes de la prova de nivell pels nois"))
with(dd[dd$sexe == "DONA",], hist(provaNivell, col = colors()[44], ylab = "Freqüència", xlab = "Notes", main = "Notes de la prova de nivell per les noies"))
with(dd[dd$sexe == "HOME",], summary(provaNivell))
with(dd[dd$sexe == "DONA",], summary(provaNivell))
t.test(dd$provaNivell ~ dd$sexe, alternative = "greater")

#sexe
table(dd$sexe)
pie(table(dd$sexe), col = colors()[c(28,44)], main = "Sexe dels alumnes")
prop.table(table(dd$sexe))

#Nombre quadrimestres Fase Inicial
table(dd$nombre_quad_FI)
prop.table(table(dd$nombre_quad_FI))
summary(dd$nombre_quad_FI)
barplot(table(dd$sexe, dd$nombre_quad_FI), col = colors()[c(28,44)], xlab = "Semestres", ylab = "Nbre. alumnes", main = "Nombre de semestres per aprovar la Fase Inicial")
legend ("topright", c("Home", "Dona"), col = colors()[c(28,44)], pch = 16)
with(dd[dd$sexe == "HOME",], prop.table(table(nombre_quad_FI)))
with(dd[dd$sexe == "DONA",], prop.table(table(nombre_quad_FI)))
with(dd[dd$sexe == "HOME",], summary(nombre_quad_FI))
with(dd[dd$sexe == "DONA",], summary(nombre_quad_FI))

#Funció per analitzar les variables: Notes de les assignatures obligatòries
#Fa un histograma i resum numèric de la variable i també, separant per sexe.
descriptiu <- function(x, y){
  TOTAL <- summary(x)
  hist(x, col = colors()[44], ylab ="Freqüència", xlab = "Nota", main = paste("Nota assignatura:", y))
  hist(x[dd$sexe == "HOME"], main = paste("Nota assignatura:", y,        "pels nois"), col = colors()[44], ylab = "Freqüència", xlab = "Nota")
  hist(x[dd$sexe == "DONA"], main = paste("Nota assignatura:", y,
"per les noies"), col = colors()[44], ylab = "Freqüència", xlab = "Nota")
  HOME <- summary(x[dd$sexe == "HOME"])
  DONA <- summary(x[dd$sexe == "DONA"])
  return(rbind(HOME, DONA, TOTAL))
}

descriptiu(dd$nota_Fonaments, "Fonaments de la Matemàtica")
descriptiu(dd$nota_AlgLineal, "Àlgebra Lineal")
descriptiu(dd$nota_CalculUnaVariable, "Càlcul en una Variable")
descriptiu(dd$nota_Informatica, "Informàtica")
descriptiu(dd$nota_CalculDiferencial, "Càlcul Diferencial")
```

```
descriptiu(dd$nota_GeometriaAfiEuclidiana, "Geometria Afí i Euclidiana")
descriptiu(dd$nota_AlgebraLinealNumerica, "Àlgebra Lineal Numèrica")
descriptiu(dd$nota_MatematicaDiscreta, "Matemàtica Discreta")

descriptiu(dd$nota_CalculIntegral, "Càlcul Integral")
descriptiu(dd$nota_AlgebraMultilinealIGeometria, "Àlgebra Multilineal i Geometria")
descriptiu(dd$nota_Algorismia, "Algorísmia")
descriptiu(dd$nota_ProgramacioMatematica, "Programació Matemàtica")
descriptiu(dd$nota_FuncionsVariableComplexa, "Funcions de Variable Complexa")
descriptiu(dd$nota_Topologia, "Topologia")
descriptiu(dd$nota_Fisica, "Física")
descriptiu(dd$nota_AnalisiReal, "Anàlisi Real")

descriptiu(dd$nota_EquacionsDiferencialsOrd, "Equacions Diferencials Ordinàries")
descriptiu(dd$nota_EstructuresAlgebraiques, "Estructures Algebraiques")
descriptiu(dd$nota_CalculNumeric, "Càlcul Numèric")
descriptiu(dd$nota_TeoriaProbabilitat, "Teoria de la Probabilitat")
descriptiu(dd$nota_EquacionsDerivadesParcials, "Equacions en Derivades Parcials")
descriptiu(dd$nota_GeometriaDiferencial, "Geometria Diferencial")
descriptiu(dd$nota_ModelsMatematicsFisica, "Models Matemàtics de la Física")
descriptiu(dd$nota_Estadistica, "Estadistica")

descriptiu(dd$nota_ModelsMatematicsTecnologia, "Models Matemàtics de la Tecnologia")

#Nombre assignatures aprovades a la primera matrícula
summary(dd$nombre_assigs_aprovades_1)
barplot(table(dd$sexe, dd$nombre_assigs_aprovades_1), col = colors()[c(28,44)], xlab = "Nbre. assignatures", ylab = "Nbre. alumnes", main = "Nombre d'assignatures aprovades en la primera matrícula")
legend (1,140, c("Home", "Dona"), col = colors()[c(28,44)], pch = 16)
with(dd[dd$sexe == "HOME",], prop.table(table(nombre_quad_FI)))
with(dd[dd$sexe == "DONA",], prop.table(table(nombre_quad_FI)))
with(dd[dd$sexe == "HOME",], summary(nombre_quad_FI))
with(dd[dd$sexe == "DONA",], summary(nombre_quad_FI)

#Nota final de la carrera
summary(dd$nota_final)
hist(dd$nota_final)
with(dd[dd$sexe == "HOME",], hist(nota_final, col = colors()[44], ylab = "Freqüència", xlab = "Nota", main = "Notes finals del grau pels nois"))
with(dd[dd$sexe == "DONA",], hist(nota_final, col = colors()[44], ylab = "Freqüència", xlab = "Nota", main = "Notes finals del grau per les noies"))
with(dd[dd$sexe == "HOME",], summary(nota_final))
```

Codi d'R per a l'anàlisi descriptiu bivariant de les variables de la base de dades dels apartats 4. Anàlisi de la nota final dels estudiants del grau de Matemàtiques i 5. Anàlisi de la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres amb la variable nota final.

```
#Any d'entrada
plot(dd$nota_final[dd$any_entrada == "2009" & dd$sexe == "HOME"], 1:26,  col = 2, pch = 1, ylab = 'Id. estudiant', ylim = c(0,40), xlim = c(5,10), main = "Nota final vs Any d'entrada", xlab = "Nota final")
points(dd$nota_final[dd$any_entrada == "2009" & dd$sexe == "DONA"],1:12,col = 2, pch = 2)
points(dd$nota_final[dd$any_entrada == "2010" & dd$sexe == "HOME"],1:30,col = 3, pch = 1)
points(dd$nota_final[dd$any_entrada == "2010" & dd$sexe == "DONA"],1:13,col = 3, pch = 2)
points(dd$nota_final[dd$any_entrada == "2011" & dd$sexe == "HOME"],1:35,col = 4, pch = 1)
points(dd$nota_final[dd$any_entrada == "2011" & dd$sexe == "DONA"],1:12,col = 4, pch = 2)
points(dd$nota_final[dd$any_entrada == "2012" & dd$sexe == "HOME"],1:37,col = 5, pch = 1)
points(dd$nota_final[dd$any_entrada == "2012" & dd$sexe == "DONA"],1:12,col = 5, pch = 2)
points(dd$nota_final[dd$any_entrada == "2013" & dd$sexe == "HOME"],1:28,col = 6, pch = 1)
points(dd$nota_final[dd$any_entrada == "2013" & dd$sexe == "DONA"],1:15,col = 6, pch = 2)
points(dd$nota_final[dd$any_entrada == "2014" & dd$sexe == "HOME"],1:36,col = 7, pch = 1)
points(dd$nota_final[dd$any_entrada == "2014" & dd$sexe == "DONA"],1:5,col = 7, pch = 2)
points(dd$nota_final[dd$any_entrada == "2015" & dd$sexe == "HOME"],1:10,col = 8, pch = 1)
points(dd$nota_final[dd$any_entrada == "2015" & dd$sexe == "DONA"],1:1,col = 8, pch = 2)
legend ("topleft", c("2009", "2010", "2011", "2012", "2013", "2014", "2015", "Home", "Dona"), col = c(2:8,1,1), pch = c(rep(1,8),2))

#CFIS
plot(dd$nota_final[dd$CFIS == "SI" & dd$sexe == "HOME"],1:99,ylim = c(1,110), xlim = c(5,10),col = 3, pch = 1, ylab = "Id. estudiant", main = "Nota final vs CFIS", xlab = "Nota final")
points(dd$nota_final[dd$CFIS == "SI" & dd$sexe == "DONA"], 1:14, col = 3, pch = 2)
points(dd$nota_final[dd$CFIS == "NO" & dd$sexe == "HOME"], 1:103, col = 2, pch = 1)
points(dd$nota_final[dd$CFIS == "NO" & dd$sexe == "DONA"], 1:56, col = 2, pch = 2)
legend ("topleft", c("NO", "SI", "Home", "Dona"), col = c(2:3,1,1), pch = c(rep(1,3),2))
with(dd[dd$CFIS == "SI" & dd$sexe == "HOME",],summary(nota_final))
with(dd[dd$CFIS == "SI" & dd$sexe == "DONA",],summary(nota_final))
with(dd[dd$CFIS == "NO" & dd$sexe == "HOME",],summary(nota_final))
with(dd[dd$CFIS == "NO" & dd$sexe == "DONA",],summary(nota_final))

#Nota accés - nota min
with(dd[dd$sexe == "HOME",], plot(nota_final, dif_acces_min, col = 2, main = "Nota final vs Nota d'accés", pch = 16, xlab = "Nota final", ylab = "Diferència nota d'accés i nota mínima"))
with(dd[dd$sexe == "DONA",], points(nota_final, dif_acces_min, col = 3, pch = 16))
legend ("topleft", c("Home", "Dona"), col = c(2:3), pch = 16)
```

```
#Prova nivell
with(dd[dd$sexe == "HOME",], plot(nota_final, provaNivell, col = 2, main = "Nota final vs Nota prova de nivell", pch = 16, xlab = "Nota final", ylab = "Nota prova de nivell", xlim = c(5,10)))
with(dd[dd$sexe == "DONA",], points(nota_final, provaNivell, col = 3, pch = 16))
legend ("topleft", c("Home", "Dona"), col = c(2:3), pch = 16)

#Nombre de quadrimestres FI
with(dd[dd$sexe == "HOME" & dd$nombre_quad_FI == '1',], plot(nota_final, 1:5, col = 2, pch = 1, main = "Nota final vs Nombre de quadrimestres FI", ylim = c(0,105), xlim = c(5,10), ylab = 'Id. estudiant', xlab = "Nota final"))
with(dd[dd$sexe == "HOME" & dd$nombre_quad_FI == '2',], points(nota_final, 1:104, col = 3, pch = 1))
with(dd[dd$sexe == "DONA" & dd$nombre_quad_FI == '2',], points(nota_final, 1:37, col = 3, pch = 2))
with(dd[dd$sexe == "HOME" & dd$nombre_quad_FI == '3',], points(nota_final, 1:64, col = 4, pch = 1))
with(dd[dd$sexe == "DONA" & dd$nombre_quad_FI == '3',], points(nota_final, 1:13, col = 4, pch = 2))
with(dd[dd$sexe == "HOME" & dd$nombre_quad_FI == '4',], points(nota_final, 1:28,  col = 6, pch = 1))
with(dd[dd$sexe == "DONA" & dd$nombre_quad_FI == '4',], points(nota_final, 1:18, col = 6, pch = 2))
with(dd[dd$sexe == "HOME" & dd$nombre_quad_FI == '5',], points(nota_final, 1, col = 7, pch = 1))
with(dd[dd$sexe == "DONA" & dd$nombre_quad_FI == '5',], points(nota_final, 1:2, col = 7, pch = 2))
legend ("topleft", c("1", "2", "3", "4", "5", "Home", "Dona"), col = c(2:4,6:7,1,1), pch = c(rep(1,6),2))

#Nota assignatures obligatòries
#Funció que crea un diagrama bivariant i calcula la correlació total i en funció del sexe
descriptiubi <- function(x, y){
  plot(dd$nota_final[dd$sexe == "HOME"], x[dd$sexe == "HOME"], main = paste("Nota final vs Nota assignatura:", y), col = 2, ylab = paste("Nota",y), xlab = "Nota final", pch = 16)
  points(dd$nota_final[dd$sexe == "DONA"], x[dd$sexe == "DONA"], col = 3, pch = 16)
  legend("bottomright", c("Home", "Dona"), col = 2:3, pch = 16)
  TOTAL <- cor(dd$nota_final, x)
  HOME <- cor(dd$nota_final[dd$sexe == "HOME"], x[dd$sexe == "HOME"])
  DONA <- cor(dd$nota_final[dd$sexe == "DONA"], x[dd$sexe == "DONA"])
  return (rbind(TOTAL, HOME, DONA))
}
```

```
descriptiubi(dd$nota_Fonaments, "Fonaments de la Matemàtica")
descriptiubi(dd$nota_AlgLineal, "Àlgebra Lineal")
descriptiubi(dd$nota_CalculUnaVariable, "Càlcul en una Variable")
descriptiubi(dd$nota_Informatica, "Informàtica")
descriptiubi(dd$nota_CalculDiferencial, "Càlcul Diferencial")
descriptiubi(dd$nota_GeometriaAfiEuclidiana, "Geometria Afí i Euclidiana")
descriptiubi(dd$nota_AlgebraLinealNumerica, "Àlgebra Lineal Numèrica")
descriptiubi(dd$nota_MatematicaDiscreta, "Matemàtica Discreta")

descriptiubi(dd$nota_CalculIntegral, "Càlcul Integral")
descriptiubi(dd$nota_AlgebraMultilinealIGeometria, "Àlgebra Multilineal i Geometria")
descriptiubi(dd$nota_Algorismia, "Algorísmia")
descriptiubi(dd$nota_ProgramacioMatematica, "Programació Matemàtica")
descriptiubi(dd$nota_FuncionsVariableComplexa, "Funcions de Variable Complexa")
descriptiubi(dd$nota_Topologia, "Topologia")
descriptiubi(dd$nota_Fisica, "Física")
descriptiubi(dd$nota_AnalisiReal, "Anàlisi Real")

descriptiubi(dd$nota_EquacionsDiferencialsOrd, "Equacions Diferencials Ordinàries")
descriptiubi(dd$nota_EstructuresAlgebraiques,"Estructures Algebraiques")
descriptiubi(dd$nota_CalculNumeric, "Càlcul Numèric")
descriptiubi(dd$nota_TeoriaProbabilitat, "Teoria de la Probabilitat")
descriptiubi(dd$nota_EquacionsDerivadesParcials,"Equacions en Derivades Parcials")
descriptiubi(dd$nota_GeometriaDiferencial, "Geometria Diferencial")
descriptiubi(dd$nota_ModelsMatematicsFisica, "Models Matemàtics de la Física")
descriptiubi(dd$nota_Estadistica, "Estadistica")

descriptiubi(dd$nota_ModelsMatematicsTecnologia, "Models Matemàtics de la Tecnologia")

#Nombre assignatures aprovades a la primera matrícula
barplot(with(dd[dd$nota_final < 7,], table(sexe, nombre_assigs_aprovades_1)), main = "Nota final menor a 7 vs Nombre d'assignatures aprovades en la primera matrícula", xlab = "Nbre. assignatures", ylab = "Freqüència", col = colors()[c(28,44)])
legend ("topleft", c("Home", "Dona"), col = colors()[c(28,44)], pch = 16)
barplot(with(dd[dd$nota_final <= 8 & dd$nota_final >= 7,], table(sexe, nombre_assigs_aprovades_1)), main = "Nota final entre 7 i 8 vs Nombre d'assignatures aprovades en la primera matrícula", xlab = "Nbre. assignatures", ylab = "Freqüència", col = colors()[c(28,44)])
legend ("topleft", c("Home", "Dona"), col = colors()[c(28,44)], pch = 16)
with(dd[dd$nota_final > 8,], table(nombre_assigs_aprovades_1, sexe))
```

Codi d'R per a la creació del model lineal de l'apartat 4. Anàlisi de la nota final dels estudiants del grau de Matemàtiques.

```
#STEP només notes assignatures
mod1 <- lm(nota_final ~ nota_Fonaments + nota_AlgLineal + nota_CalculUnaVariable + nota_Informatica + nota_CalculDiferencial + nota_GeometriaAfiEuclidiana + nota_AlgebraLinealNumerica + nota_MatematicaDiscreta + nota_CalculIntegral+ nota_AlgebraMultilinealIGeometria + nota_Algorismia + nota_ProgramacioMatematica + nota_FuncionsVariableComplexa + nota_Topologia + nota_Fisica + nota_AnalisiReal + nota_EquacionsDiferencialsOrd + nota_EstructuresAlgebraiques + nota_CalculNumeric + nota_TeoriaProbabilitat + nota_EquacionsDerivadesParcials + nota_GeometriaDiferencial + nota_ModelsMatematicsFisica + nota_Estadistica + nota_ModelsMatematicsTecnologia, data = dd)

mod11 <- step(lm(nota_final ~ 1, data = dd), formula(mod1))
summary(mod11)

#Sense nota assignatures
library(car)
mod21 <- lm(nota_final ~ dif_acces_min*sexe + dif_acces_min*CFIS + nombre_quad_FI*sexe + nombre_quad_FI*CFIS + nombre_assigs_aprovades_1 *sexe + nombre_assigs_aprovades_1*CFIS, data = dd)
summary(mod21)
Anova(mod21) #Interacció dif_acces_min*CFIS no significativa

mod22 <- lm(nota_final ~ dif_acces_min*sexe + nombre_quad_FI*sexe + nombre_quad_FI*CFIS + nombre_assigs_aprovades_1*sexe + nombre_assigs_aprovades_1*CFIS, data = dd)
summary(mod22)
Anova(mod22) #Interacció nombre_assigs_aprovades_1*sexe no significativa

mod23 <- lm(nota_final ~ dif_acces_min*sexe + nombre_quad_FI*sexe + nombre_quad_FI*CFIS + nombre_assigs_aprovades_1*CFIS, data = dd)
summary(mod23)
Anova(mod23) #Interacció nombre_quad_FI*sexe no significativa

mod24 <- lm(nota_final ~ dif_acces_min*sexe + nombre_quad_FI*CFIS + nombre_assigs_aprovades_1*CFIS, data = dd)
summary(mod24)
Anova(mod24) #Interacció nombre_quad_FI*CFIS no significativa

mod25 <- lm(nota_final ~ dif_acces_min*sexe + nombre_quad_FI + nombre_assigs_aprovades_1*CFIS, data = dd)
summary(mod25)
Anova(mod25) #Tot significatiu
```

```
#Comparació
anova(mod22, mod21) #Millor mod22
anova(mod23, mod22) #Millor mod23
anova(mod24, mod23) #Millor mod24
anova(mod25, mod24) #Millor mod25

vif(mod25)

plot(residuals(mod25) ~ fitted.values(mod25), xlab = "Valors ajustats", ylab = "Residus", main = "Residus VS Valors ajustats")
abline(h = 0, lty = 2)
plot(mod25, which = 2)
shapiro.test(residuals(mod25)) #Residus no normals

mod26 <- lm((1/nota_final) ~ dif_acces_min*sexe + CFIS + nombre_quad_FI + nombre_assigs_aprovades_1, data = dd)

plot(residuals(mod26) ~ fitted.values(mod26), xlab = "Valors ajustats", ylab = "Residus", main = "Residus VS Valors ajustats")
abline(h = 0, lty = 2)
plot(mod26, which = 2)
shapiro.test(residuals(mod26)) #Residus sí normals

#ACP només amb assingatures
library(FactoMineR)
ACP <- PCA(dd[,c(4,8,10:34)], quali.sup = c(1,2))

#Model amb 5 assignatures més rellevants (seleccionades amb STEP)
mod31 <-  lm(nota_final ~ dif_acces_min*sexe + dif_acces_min*CFIS + nombre_quad_FI*sexe + nombre_quad_FI*CFIS + nombre_assigs_aprovades_1 *sexe + nombre_assigs_aprovades_1*CFIS + nota_CalculNumeric*sexe + nota_CalculNumeric*CFIS + nota_TeoriaProbabilitat*sexe + nota_TeoriaProbabilitat*CFIS + nota_ProgramacioMatematica*sexe + nota_ProgramacioMatematica*CFIS + nota_EstructuresAlgebraiques*sexe + nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*sexe + nota_Estadistica*CFIS, data = dd)
summary(mod31)
Anova(mod31) #Interacció nota_EstructuresAlgebraiques*sexe no significativa

mod32 <-  lm(nota_final ~ dif_acces_min*sexe + dif_acces_min*CFIS + nombre_quad_FI*sexe + nombre_quad_FI*CFIS + nombre_assigs_aprovades_1 *sexe + nombre_assigs_aprovades_1*CFIS + nota_CalculNumeric*sexe + nota_CalculNumeric*CFIS + nota_TeoriaProbabilitat*sexe + nota_TeoriaProbabilitat*CFIS + nota_ProgramacioMatematica*sexe + nota_ProgramacioMatematica*CFIS + nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*sexe + nota_Estadistica*CFIS, data = dd)
```

```
summary(mod32)
Anova(mod32) #Interacció nota_ProgramacioMatematica*CFIS no significativa

mod33 <-  lm(nota_final ~ dif_acces_min*sexe + dif_acces_min*CFIS + nombre_quad_FI*sexe + nombre_quad_FI*CFIS + nombre_assigs_aprovades_1 *sexe + nombre_assigs_aprovades_1*CFIS + nota_CalculNumeric*sexe + nota_CalculNumeric*CFIS + nota_TeoriaProbabilitat*sexe + nota_TeoriaProbabilitat*CFIS + nota_ProgramacioMatematica*sexe + nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*sexe + nota_Estadistica*CFIS, data = dd)
summary(mod33)
Anova(mod33) #Interacció nota_CalculNumeric*CFIS no significativa

mod34 <-  lm(nota_final ~ dif_acces_min*sexe + dif_acces_min*CFIS + nombre_quad_FI*sexe + nombre_quad_FI*CFIS + nombre_assigs_aprovades_1 *sexe + nombre_assigs_aprovades_1*CFIS + nota_CalculNumeric*sexe + nota_TeoriaProbabilitat*sexe + nota_TeoriaProbabilitat*CFIS + nota_ProgramacioMatematica*sexe + nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*sexe + nota_Estadistica*CFIS,data=dd)
summary(mod34)
Anova(mod34) #Interacció nota_ProgramacioMatematica*sexe no significativa

mod35 <-  lm(nota_final ~ dif_acces_min*sexe + dif_acces_min*CFIS + nombre_quad_FI*sexe + nombre_quad_FI*CFIS + nombre_assigs_aprovades_1 *sexe + nombre_assigs_aprovades_1*CFIS + nota_CalculNumeric*sexe + nota_TeoriaProbabilitat*sexe + nota_TeoriaProbabilitat*CFIS + nota_ProgramacioMatematica + nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*sexe + nota_Estadistica*CFIS,data=dd)
summary(mod35)
Anova(mod35) #Interacció nombre_quad_FI*CFIS no significativa

mod36 <-  lm(nota_final ~ dif_acces_min*sexe + dif_acces_min*CFIS + nombre_quad_FI*sexe + nombre_assigs_aprovades_1*sexe + nombre_assigs_aprovades_1*CFIS + nota_CalculNumeric*sexe + nota_TeoriaProbabilitat*sexe + nota_TeoriaProbabilitat*CFIS + nota_ProgramacioMatematica + nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*sexe + nota_Estadistica*CFIS, data = dd)
summary(mod36)
Anova(mod36) #Interacció nota_Estadistica*sexe no significativa

mod37 <-  lm(nota_final ~ dif_acces_min*sexe + dif_acces_min*CFIS + nombre_quad_FI*sexe + nombre_assigs_aprovades_1*sexe + nombre_assigs_aprovades_1*CFIS + nota_CalculNumeric*sexe + nota_TeoriaProbabilitat*sexe + nota_TeoriaProbabilitat*CFIS + nota_ProgramacioMatematica + nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*CFIS, data = dd)
summary(mod37)
Anova(mod37) #Interacció nota_TeoriaProbabilitat*sexe no significativa
```

```
mod38 <-  lm(nota_final ~ dif_acces_min*sexe + dif_acces_min*CFIS + nombre_quad_FI*sexe
+ nombre_assigs_aprovades_1*sexe + nombre_assigs_aprovades_1*CFIS +
nota_CalculNumeric*sexe + nota_TeoriaProbabilitat*CFIS + nota_ProgramacioMatematica +
nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*sexe + nota_Estadistica*CFIS,data=dd)
summary(mod38)
Anova(mod38) #Interacció nota_Estadistica*sexe no significativa

mod39 <-  lm(nota_final ~ dif_acces_min*sexe + dif_acces_min*CFIS + nombre_quad_FI*sexe
+ nombre_assigs_aprovades_1*sexe + nombre_assigs_aprovades_1*CFIS +
nota_CalculNumeric*sexe + nota_TeoriaProbabilitat*CFIS + nota_ProgramacioMatematica +
nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*CFIS, data = dd)
summary(mod39)
Anova(mod39) #Interacció nota_CalculNumeric*sexe no significativa

mod310 <- lm(nota_final ~ dif_acces_min*sexe + dif_acces_min*CFIS +
nombre_quad_FI*sexe + nombre_assigs_aprovades_1*sexe +
nombre_assigs_aprovades_1*CFIS + nota_CalculNumeric + nota_TeoriaProbabilitat*CFIS +
nota_ProgramacioMatematica + nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*CFIS,
data = dd)
summary(mod310)
Anova(mod310) #Interacció nombre_assigs_aprovades_1*sexe no significativa

mod311 <- lm(nota_final ~ dif_acces_min*sexe + dif_acces_min*CFIS +
nombre_quad_FI*sexe + nombre_assigs_aprovades_1*CFIS + nota_CalculNumeric +
nota_TeoriaProbabilitat*CFIS + nota_ProgramacioMatematica +
nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*CFIS, data = dd)
summary(mod311)
Anova(mod311) #Interacció nombre_assigs_aprovades_1*CFIS no significativa

mod312 <- lm(nota_final ~ dif_acces_min*sexe + dif_acces_min*CFIS +
nombre_quad_FI*sexe + nombre_assigs_aprovades_1 + nota_CalculNumeric +
nota_TeoriaProbabilitat*CFIS + nota_ProgramacioMatematica +
nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*CFIS, data = dd)
summary(mod312)
Anova(mod312) #Interacció dif_acces_min*CFIS no significativa

mod313 <- lm(nota_final ~ dif_acces_min*sexe + nombre_quad_FI*sexe +
nombre_assigs_aprovades_1 + nota_CalculNumeric + nota_TeoriaProbabilitat*CFIS +
nota_ProgramacioMatematica + nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*CFIS,
data = dd)
summary(mod313)
Anova(mod313) #Interacció nota_TeoriaProbabilitat*CFIS no significativa

mod314 <- lm(nota_final ~ dif_acces_min*sexe + nombre_quad_FI*sexe +
nombre_assigs_aprovades_1 + nota_CalculNumeric + nota_TeoriaProbabilitat +
```

```
nota_ProgramacioMatematica + nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*CFIS,
data = dd)
summary(mod314)
Anova(mod314) #Interacció dif_acces_min*sexe no significativa

mod315 <- lm(nota_final ~ dif_acces_min + nombre_quad_FI*sexe +
nombre_assigs_aprovades_1 + nota_CalculNumeric + nota_TeoriaProbabilitat +
nota_ProgramacioMatematica + nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*CFIS,
data = dd)
summary(mod315)
Anova(mod315) #Nombre_assigs_aprovades_1 no significativa

mod316 <- lm(nota_final ~ dif_acces_min + nombre_quad_FI*sexe + nota_CalculNumeric +
nota_TeoriaProbabilitat + nota_ProgramacioMatematica +
nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*CFIS, data = dd)
summary(mod316)
Anova(mod316) #Interacció nombre_quad_FI*sexe no significativa

mod317 <- lm(nota_final ~ dif_acces_min + nombre_quad_FI + sexe + nota_CalculNumeric
+ nota_TeoriaProbabilitat + nota_ProgramacioMatematica +
nota_EstructuresAlgebraiques*CFIS + nota_Estadistica*CFIS, data = dd)
summary(mod317)
Anova(mod317) #Interacció nota_Estadistica*CFIS no significativa

mod318 <- lm(nota_final ~ dif_acces_min + nombre_quad_FI + sexe + nota_CalculNumeric +
nota_TeoriaProbabilitat + nota_ProgramacioMatematica +
nota_EstructuresAlgebraiques*CFIS + nota_Estadistica, data = dd)
summary(mod318)
Anova(mod318) #Interacció nombre_quad_FI no significativa

mod319 <- lm(nota_final ~ dif_acces_min + sexe + nota_CalculNumeric +
nota_TeoriaProbabilitat + nota_ProgramacioMatematica +
nota_EstructuresAlgebraiques*CFIS + nota_Estadistica, data = dd)
summary(mod319)
Anova(mod319) #Tot significatiu

#Comparació
anova(mod32, mod31) #Millor mod32
anova(mod33, mod32) #Millor mod33
anova(mod34, mod33) #Millor mod34
anova(mod35, mod34) #Millor mod35
anova(mod36, mod35) #Millor mod36
anova(mod37, mod36) #Millor mod37
anova(mod38, mod36) #Millor mod38
anova(mod39, mod38) #Millor mod39
```

```
anova(mod310, mod39) #Millor mod310
anova(mod311, mod310) #Millor mod311
anova(mod312, mod311) #Millor mod312
anova(mod313, mod312) #Millor mod313
anova(mod314, mod313) #Millor mod314
anova(mod315, mod314) #Millor mod315
anova(mod316, mod315) #Millor mod316
anova(mod317, mod316) #Millor mod317
anova(mod318, mod317) #Millor mod318
anova(mod319, mod318) #Millor mod319

vif(mod319)

plot(residuals(mod319) ~ fitted.values(mod319), xlab = "Valors ajustats", ylab = "Residus",
main =  "Residus VS Valors ajustats")
abline(h = 0, lty = 2)
plot(mod319, which = 2)
shapiro.test(residuals(mod319)) #Residus sí normals

c(summary(mod319)$adj.r.squared)

#Model final: MOD319

summary(mod319)
confint(mod319)
```

Codi d'R per a l'anàlisi descriptiu bivariant de les variables de la base de dades dels apartats 4. Anàlisi de la nota final dels estudiants del grau de Matemàtiques i 5. Anàlisi de la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres amb la variable dicotòmica que indica si s'ha aprovat la fase inicial del grau en dos quadrimestres o no.

```
dd$dos_quadFI <- ifelse(dd$nombre_quad_FI == '2', "SI", "NO")

#Sexe
table(dd$sexe, dd$dos_quadFI)
barplot(table(dd$sexe, dd$dos_quadFI), ylim = c(0,170), col = colors()[c(28,44)], ylab =
"Freqüència", main = "Alumnes que aproven la FI en 2 quadrimestres en funció del sexe")
legend("topleft", c("Home", "Dona"), pch = 16, col = colors()[c(28,44)])

#Any d'entrada
with(dd[dd$any_entrada == "2009",],table(sexe, dos_quadFI))
with(dd[dd$any_entrada == "2010",],table(sexe, dos_quadFI))
with(dd[dd$any_entrada == "2011",],table(sexe, dos_quadFI))
with(dd[dd$any_entrada == "2012",],table(sexe, dos_quadFI))
with(dd[dd$any_entrada == "2013",],table(sexe, dos_quadFI))
with(dd[dd$any_entrada == "2014",],table(sexe, dos_quadFI))
with(dd[dd$any_entrada == "2015",],table(sexe, dos_quadFI))

#CFIS
barplot(with(dd[dd$CFIS == "SI",],table(sexe, dos_quadFI)), col = colors()[c(28,44)], ylab =
"Freqüència", main = "Alumnes que pertanyen al CFIS vs 2 quadrimestres en aprovar FI", ylim
= c(0, 70))
legend("topleft", c("Home", "Dona"), pch = 16, col = colors()[c(28,44)])
barplot(with(dd[dd$CFIS == "NO",],table(sexe, dos_quadFI)), col = colors()[c(28,44)], ylab =
"Freqüència", main = "Alumnes que no pertanyen al CFIS vs 2 quadrimestres en aprovar FI",
ylim = c(0, 105))
legend("topleft", c("Home", "Dona"), pch = 16, col = colors()[c(28,44)])
with(dd[dd$CFIS == "SI",],table(sexe, dos_quadFI))
with(dd[dd$CFIS == "NO",],table(sexe, dos_quadFI))

#Nota accés - nota min
with(dd[dd$sexe == "HOME" & dd$dos_quadFI == "NO",], plot(dif_acces_min, 1:98, pch = 1,
col = 2, xlab = "Diferència entre nota d'accés i nota mínima", ylab = "Id. estudiant", main =
"Diferència entre la nota d'accés i la nota mínima vs 2 quadrimestres en aprovar FI", ylim =
c(-1,106), xlim = c(0,7.5)))
with(dd[dd$sexe == "HOME" & dd$dos_quadFI == "SI",], points(dif_acces_min, 1:104, pch =
1, col = 3))
with(dd[dd$sexe == "DONA" & dd$dos_quadFI == "NO",], points(dif_acces_min, 1:33, pch = 2,
col = 2))
```

```
with(dd[dd$sexe == "DONA" & dd$dos_quadFI == "SI",], points(dif_acces_min, 1:37, pch = 2, col = 3))
legend("topright", c( "NO", "SÍ", "Home", "Dona"), col = c(2,3,1,1), pch = c(1,1,1,2), cex = 0.8)

#Prova nivell
with(dd[dd$sexe == "HOME" & dd$dos_quadFI == "NO",], plot(provaNivell, 1:98, pch = 1, col = 2, xlab = "Nota prova de nivell", ylab = "Id. estudiant", main = "Nota prova de nivell vs 2 quadrimestres en aprovar FI", ylim = c(-1,106), xlim = c(3.2, 10)))
with(dd[dd$sexe == "HOME" & dd$dos_quadFI == "SI",], points(provaNivell, 1:104, pch = 1, col = 3))
with(dd[dd$sexe == "DONA" & dd$dos_quadFI == "NO",], points(provaNivell, 1:33, pch = 2, col = 2))
with(dd[dd$sexe == "DONA" & dd$dos_quadFI == "SI",], points(provaNivell, 1:37, pch = 2, col = 3))
legend("topleft", c("No", "Sí","Home", "Dona"), col = c(2,3,1,1), pch = c(1,1,1,2))
```

Codi d'R per a la creació del model lineal generalitzat de l'apartat 5. Anàlisi de la probabilitat d'aprovar la fase inicial del grau en dos quadrimestres.

```
#Model Lineal Generalitzat
library(car)
dd$dos_quadFI <- ifelse(dd$dos_quadFI == "SI", 1, 0)

mod1 <- glm (dos_quadFI ~ CFIS + dif_acces_min + dif_acces_min*CFIS + dif_acces_min*sexe
+ sexe, family = binomial, data = dd)
summary(mod1)
Anova(mod1) #Interacció sexe*dif_acces_min no significativa

mod2 <- glm (dos_quadFI ~ CFIS + dif_acces_min + dif_acces_min*CFIS + sexe, family =
binomial, data = dd)
summary(mod2)
Anova(mod2) #Sexe no signficativa

mod3 <- glm (dos_quadFI ~ CFIS + dif_acces_min + dif_acces_min*CFIS, family = binomial,
data = dd)
summary(mod3)
Anova(mod3) #Interacció CFIS*dif_acces_min no significativa

mod4 <- glm (dos_quadFI ~ CFIS + dif_acces_min, family = binomial, data = dd)
summary(mod4)
Anova(mod4) #CFIS no significativa

mod5 <- glm (dos_quadFI ~ dif_acces_min, family = binomial, data = dd)
summary(mod5)
Anova(mod5) #Tot significatiu

#Elecció del millor model
anova(mod2, mod1, test = 'Chisq') #Millor mod2
anova(mod3, mod2, test = 'Chisq') #Millor mod3
anova(mod4, mod3, test = 'Chisq') #Millor mod4
anova(mod5, mod4, test = 'Chisq') #Millor mod5

cbind(AIC(mod1), AIC(mod2), AIC(mod3), AIC(mod4), AIC(mod5))
#Segons AIC: Millor mod5
cbind(BIC(mod1), BIC(mod2), BIC(mod3), BIC(mod4), BIC(mod5))
#Segons BIC: Millor mod5
cbind((1-mod1$dev/mod1$null.dev),          (1-mod2$dev/mod2$null.dev),          (1-
mod3$dev/mod3$null.dev),
    (1-mod4$dev/mod4$null.dev), (1-mod5$dev/mod5$null.dev))
#Segons Pseudo-R^2: Millor mod1
#Millor model: mod5
```

```
#AUC: Capacitat predictiva
#install.packages('AUC')
#install.packages('ROCR')
library(AUC)
library(ROCR)
y <- dd$dos_quadFI
pr <- predict(mod5, type="response")
boxplot(pr ~ y, main = "Valors predits vs Valors reals", xlab = '2 quadrimestres en aprovar FI',
ylab = "Probabilitat")
dadesroc <- prediction(pr,y)
par(mfrow=c(1,2))
roc.perf <- performance(dadesroc,"auc",fpr.stop=0.05)
plot(performance(dadesroc,"err"))
plot(performance(dadesroc,"tpr"))
par(mfrow = c(1,1))
plot(performance(dadesroc,"tpr","fpr"), ylab = 'Sensibilitat', xlab = 'Taxa de falsos positius',
main = 'Corba ROC') #Corba ROC
abline(0,1, col = 'red')
auc <- auc(roc(pr,factor(y)))

# Mesures de capacitat predictiva
y_esp <- ifelse(pr<0.5, "0", "1")
taula <- table(y_esp, y)
sum(diag(taula))/sum(taula) # proporció d'encert
taula[2,2]/sum(taula[,2])   # sensibilitat
taula[1,1]/sum(taula[,1])   # Especificitat
taula[2,2]/sum(taula[2,])   # Valor Predictiu Positiu
taula[1,1]/sum(taula[1,])   # Valor Predictiu negatiu

# Validació del model: Gràfic
residualPlots(mod5)

# Model final
summary(mod5)
confint(mod5)
exp(coef(mod5)[2])
#L'increment en 1 punt en la diferència de notes provoca l'increment del LOGODDS en
0.2048358.
#A més diferència entre la nota d'accés i la nota mínima més  probabilitats d'aprovar la fase
inicial en 2 quadrimestres.
#L'ODDS d'aprovar la FI en 2 quadrimestres és 1.227323 vegades més gran en augmentar la
diferència en 1 punt. L'increment d'1 punt per a la diferència de notes provoca l'increment
del ODDS d'aprovar la FI en 2 quad. en 1.227323.
exp(confint(mod5))
```

Codis d'R per a la comparació de proporcions de nois i noies al grau de Matemàtiques i test d'homogeneïtat.

```
#Comparació de proporcions
dd$sexe <- factor(dd$sexe, levels = levels(factor(dd$sexe))[c(2,1)])
ddaux <- matrix(table(dd$any, dd$sexe), nrow = 7, ncol = 2)
ddaux <- ddaux[1:6,]
rownames(ddaux) <- 2009:2014

ddaux <- cbind(ddaux, ddaux[,1] + ddaux[,2])
prop_nois <- ddaux[,1]/ddaux[,3]
prop_noies <- ddaux[,2]/ddaux[,3]
ddaux <- cbind(ddaux, prop_nois, prop_noies)
colnames(ddaux) <- c("HOME", "DONA", "TOTAL", "PROP_NOIS", "PROP_NOIES")

Z <- (ddaux[,4] - 0.5)/sqrt(((ddaux[,4])*(1 - ddaux[,4]))/ddaux[,3])
ddaux <- cbind(ddaux, Z)
Z_alpha <- qnorm(0.05, lower.tail = FALSE)
Z_alpha2 <- qnorm(0.05/2, lower.tail = FALSE)

Lim_inf_nois_bi <- ifelse((ddaux[,4] - Z_alpha2*sqrt(((ddaux[,4])*(1 - ddaux[,4]))/ddaux[,3])) < 0, 0, ddaux[,4] - Z_alpha2*sqrt(((ddaux[,4])*(1 - ddaux[,4]))/ddaux[,3]))
Lim_sup_nois_bi <- ifelse((ddaux[,4] + Z_alpha2*sqrt(((ddaux[,4])*(1 - ddaux[,4]))/ddaux[,3])) > 1, 1, ddaux[,4] + Z_alpha2*sqrt(((ddaux[,4])*(1 - ddaux[,4]))/ddaux[,3]))
Lim_inf_nois_uni <- ifelse((ddaux[,4] - Z_alpha*sqrt(((ddaux[,4])*(1 - ddaux[,4]))/ddaux[,3])) > 1, 1, ddaux[,4] - Z_alpha*sqrt(((ddaux[,4])*(1 - ddaux[,4]))/ddaux[,3]))
Lim_sup_nois_uni <- rep(1,length = nrow(ddaux))

Lim_inf_noies_bi <- ifelse((ddaux[,5] - Z_alpha2*sqrt(((ddaux[,5])*(1 - ddaux[,5]))/ddaux[,3])) < 0, 0, ddaux[,5] - Z_alpha2*sqrt(((ddaux[,5])*(1 - ddaux[,5]))/ddaux[,3]))
Lim_sup_noies_bi <- ifelse((ddaux[,5] + Z_alpha2*sqrt(((ddaux[,5])*(1 - ddaux[,5]))/ddaux[,3])) > 1, 1, ddaux[,5] + Z_alpha2*sqrt(((ddaux[,5])*(1 - ddaux[,5]))/ddaux[,3]))
Lim_inf_noies_uni <- rep(0,length = nrow(ddaux))
Lim_sup_noies_uni <- ifelse((ddaux[,5] + Z_alpha*sqrt(((ddaux[,5])*(1 - ddaux[,5]))/ddaux[,3])) > 1, 1, ddaux[,5] + Z_alpha * sqrt(((ddaux[,5])*(1 - ddaux[,5]))/ddaux[,3]))

nois <- cbind(ddaux[,4], Lim_inf_nois_bi, Lim_sup_nois_bi, Lim_inf_nois_uni, Lim_sup_nois_uni)
```

```
noies <- cbind(ddaux[,5], Lim_inf_noies_bi, Lim_sup_noies_bi, Lim_inf_noies_uni, Lim_sup_noies_uni)

#Test d'homogeneïtat
test <- chisq.test(ddaux[,c(1,2)])

test$observed
test$expected

qchisq(0.05, df=5, lower.tail = FALSE)
```

Dades apartat 6. Evolució dels nois i noies al grau de Matemàtiques.

Total:

any	tipus_pla	superaFI	supera15	nois	noies	total
1997	LLICENCIATURA	44	45	20	26	46
1998	LLICENCIATURA	39	43	19	26	45
1999	LLICENCIATURA	32	38	21	25	46
2000	LLICENCIATURA	32	37	20	19	39
2001	LLICENCIATURA	36	40	26	18	44
2002	LLICENCIATURA	34	34	22	23	45
2003	LLICENCIATURA	27	32	19	17	36
2004	LLICENCIATURA	29	33	23	21	44
2005	LLICENCIATURA	33	38	30	14	44
2006	LLICENCIATURA	31	36	27	22	49
2007	LLICENCIATURA	27	32	31	17	48
2008	LLICENCIATURA	21	29	29	21	50
2009	GRAU	38	45	34	25	59
2010	GRAU	37	44	34	20	54
2011	GRAU	37	41	38	16	54
2012	GRAU	44	48	42	14	56
2013	GRAU	40	44	35	22	57
2014	GRAU	39	46	43	11	54
2015	GRAU	35	38	26	20	46
2016	GRAU	39	44	35	14	49
2017	GRAU	33	43	35	16	51
2018	GRAU	19	51	35	20	55
2019	GRAU	0	40	38	15	53

CFIS:

any	tipus_pla	nois	noies	total
1997	LLICENCIATURA			
1998	LLICENCIATURA			
1999	LLICENCIATURA			
2000	LLICENCIATURA			
2001	LLICENCIATURA	1	0	1
2002	LLICENCIATURA	2	0	2
2003	LLICENCIATURA	2	0	2
2004	LLICENCIATURA	0	1	1
2005	LLICENCIATURA	3	0	3
2006	LLICENCIATURA	5	1	6
2007	LLICENCIATURA	5	0	5
2008	LLICENCIATURA	4	1	5
2009	GRAU	6	0	6
2010	GRAU	9	4	13
2011	GRAU	4	1	5
2012	GRAU	9	1	10
2013	GRAU	4	2	6
2014	GRAU	14	0	14
2015	GRAU	9	2	11
2016	GRAU	15	1	16
2017	GRAU	14	2	16
2018	GRAU	18	5	23
2019	GRAU	13	4	17

Test de Daniel per a la variable nombre total d'alumnes de la llicenciatura de Matemàtiques.

Any	t	Total	Rang	d	d^2
1997	1	46	8	7	49
1998	2	45	6	4	16
1999	3	46	8	5	25
2000	4	39	2	-2	4
2001	5	44	3	-2	4
2002	6	45	6	0	0
2003	7	36	1	-6	36
2004	8	44	3	-5	25
2005	9	44	3	-6	36
2006	10	49	11	1	1
2007	11	48	10	-1	1
2008	12	50	12	0	0

Contrast de Daniel	
SUMA	197
ESTADÍSTIC Ts	0,31118881
ESTADÍSTIC Z	1,03209653
ABS(Z)	1,03209653

Test de Daniel per a la variable nombre total d'alumnes del grau de Matemàtiques.

Any	t	Total	Rang	d	d^2
2009	1	59	11	10	100
2010	2	54	5	3	9
2011	3	54	5	2	4
2012	4	56	9	5	25
2013	5	57	10	5	25
2014	6	54	5	-1	1
2015	7	46	1	-6	36
2016	8	49	2	-6	36
2017	9	51	3	-6	36
2018	10	55	8	-2	4
2019	11	53	4	-7	49

Contrast de Daniel	
SUMA	325
ESTADÍSTIC Ts	-0,47727273
ESTADÍSTIC Z	-1,50926888
ABS(Z)	1,50926888

Test de Daniel per a la variable proporció d’alumnes noies al grau de Matemàtiques.

Any	t	Total	Noies	Noies/Total	Rang	d	d^2
1997	1	46	26	0,56521739	22	21	441
1998	2	45	26	0,57777778	23	21	441
1999	3	46	25	0,54347826	21	18	324
2000	4	39	19	0,48717949	19	15	225
2001	5	44	18	0,40909091	12	7	49
2002	6	45	23	0,51111111	20	14	196
2003	7	36	17	0,47222222	17	10	100
2004	8	44	21	0,47727273	18	10	100
2005	9	44	14	0,31818182	7	-2	4
2006	10	49	22	0,44897959	16	6	36
2007	11	48	17	0,35416667	8	-3	9
2008	12	50	21	0,42	13	1	1
2009	13	59	25	0,42372881	14	1	1
2010	14	54	20	0,37037037	10	-4	16
2011	15	54	16	0,2962963	5	-10	100
2012	16	56	14	0,25	2	-14	196
2013	17	57	22	0,38596491	11	-6	36
2014	18	54	11	0,2037037	1	-17	289
2015	19	46	20	0,43478261	15	-4	16
2016	20	49	14	0,28571429	4	-16	256
2017	21	51	16	0,31372549	6	-15	225
2018	22	55	20	0,36363636	9	-13	169
2019	23	53	15	0,28301887	3	-20	400

Contrast de Daniel	
SUMA	3630
ESTADÍSTIC Ts	-0,79347826
ESTADÍSTIC Z	-3,72174294
ABS(Z)	3,72174294

Resultat d'aplicar la funció regressió sobre les dades de la proporció de noies al grau de Matemàtiques dels anys 1997-2015.

Resumen								
Estadísticas de la regresión								
Coeficiente de correlación múltiple	0,77186111							
Coeficiente de determinación R^2	0,59576958							
R^2 ajustado	0,57199132							
Error típico	0,06730383							
Observaciones	19							
ANÁLISIS DE VARIANZA								
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*			
Regresión	1	0,11349528	0,113495277	25,05522188	0,00010832			
Residuos	17	0,07700669	0,004529805					
Total	18	0,19050197						
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
Intercepción	0,55950398	0,03214208	17,40720981	2,85645E-12	0,49169011	0,62731785	0,49169011	0,62731785
Variable X 1	-0,01411079	0,00281905	-5,005519142	0,00010832	-0,02005846	-0,00816312	-0,02005846	-0,00816312

Prediccions resultants d'aplicar el mètode de la tendència lineal considerant les dades des del 1997-2015 com a període mostral i les del 2016-2019 com a extra-mostral i càlcul de l'EPAM per a la proporció de noies.

Any	t	Noies/Total	Predicció	Error	ABS(Error)	Error^2	ABS(Error/Yt)
1997	1	0,56521739	0,54539319	0,01982421	0,01982421	0,000393	0,035073595
1998	2	0,57777778	0,53128239	0,04649539	0,04649539	0,00216182	0,080472785
1999	3	0,54347826	0,5171716	0,02630666	0,02630666	0,00069204	0,048404262
2000	4	0,48717949	0,5030608	-0,01588131	0,01588131	0,00025222	0,032598489
2001	5	0,40909091	0,48895001	-0,0798591	0,0798591	0,00637748	0,19521113
2002	6	0,51111111	0,47483921	0,0362719	0,0362719	0,00131565	0,070966757
2003	7	0,47222222	0,46072842	0,0114938	0,0114938	0,00013211	0,024339819
2004	8	0,47727273	0,44661762	0,0306551	0,0306551	0,00093974	0,06422974
2005	9	0,31818182	0,43250683	-0,11432501	0,11432501	0,01307021	0,359307179
2006	10	0,44897959	0,41839604	0,03058356	0,03058356	0,00093535	0,068117922
2007	11	0,35416667	0,40428524	-0,05011857	0,05011857	0,00251187	0,141511268
2008	12	0,42	0,39017445	0,02982555	0,02982555	0,00088956	0,071013223
2009	13	0,42372881	0,37606365	0,04766516	0,04766516	0,00227197	0,112489782
2010	14	0,37037037	0,36195286	0,00841751	0,00841751	7,0855E-05	0,022727285
2011	15	0,2962963	0,34784206	-0,05154577	0,05154577	0,00265697	0,173966962
2012	16	0,25	0,33373127	-0,08373127	0,08373127	0,00701093	0,334925073
2013	17	0,38596491	0,31962047	0,06634444	0,06634444	0,00440158	0,171892409
2014	18	0,2037037	0,30550968	-0,10180598	0,10180598	0,01036446	0,499774789
2015	19	0,43478261	0,29139888	0,14338372	0,14338372	0,02055889	0,329782565
2016	20	0,28571429	0,27728809	0,0084262	0,0084262	7,1001E-05	0,029491684
2017	21	0,31372549	0,2631773	0,05054819	0,05054819	0,00255512	0,16112237
2018	22	0,36363636	0,2490665	0,11456986	0,11456986	0,01312625	0,315067121
2019	23	0,28301887	0,23495571	0,04806316	0,04806316	0,00231007	0,169823169

Beta0	0,55950398
Beta1	-0,01411079

EPAM	16,8876086

Resultat d'aplicar la funció regressió sobre les dades de la proporció de noies al grau de Matemàtiques dels anys 1997-2019.

Resumen

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,78784494
Coeficiente de determinación R^2	0,62069965
R^2 ajustado	0,60263773
Error típico	0,06479473
Observaciones	23

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	0,14427693	0,144276926	34,36509501	8,0841E-06
Residuos	21	0,08816549	0,004198357		
Total	22	0,23244242			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
Intercepción	0,54309082	0,02792726	19,44662013	6,56882E-15	0,48501291	0,601169	0,4850129	0,6011687
Variable X 1	-0,01194011	0,0020368	-5,862174939	8,08413E-06	-0,01617587	-0,007704	-0,016176	-0,007704

Prediccions pels anys 1997-2025 resultants d'aplicar el mètode de la tendència lineal usant les dades des del 1997-2019 de la proporció de noies.

Any	t	Noies/Total	Predicció
1997	1	0,56521739	0,53115072
1998	2	0,57777778	0,51921061
1999	3	0,54347826	0,50727051
2000	4	0,48717949	0,4953304
2001	5	0,40909091	0,48339029
2002	6	0,51111111	0,47145019
2003	7	0,47222222	0,45951008
2004	8	0,47727273	0,44756998
2005	9	0,31818182	0,43562987
2006	10	0,44897959	0,42368976
2007	11	0,35416667	0,41174966
2008	12	0,42	0,39980955
2009	13	0,42372881	0,38786945
2010	14	0,37037037	0,37592934
2011	15	0,2962963	0,36398923
2012	16	0,25	0,35204913
2013	17	0,38596491	0,34010902
2014	18	0,2037037	0,32816891
2015	19	0,43478261	0,31622881
2016	20	0,28571429	0,3042887
2017	21	0,31372549	0,2923486
2018	22	0,36363636	0,28040849
2019	23	0,28301887	0,26846838
2020	24		0,25652828
2021	25		0,24458817
2022	26		0,23264807
2023	27		0,22070796
2024	28		0,20876785
2025	29		0,19682775

Beta0	0,54309082
Beta1	-0,01194011

Test de Daniel per a la variable proporció d'alumnes que superen la fase inicial de la llicenciatura de Matemàtiques dels anys 1997-2008.

Any	t	SuperaFI	Total	SuperaFI/Total	Rang	d	d^2
1997	1	44	46	0,956521739	12	11	121
1998	2	39	45	0,866666667	11	9	81
1999	3	32	46	0,695652174	5	2	4
2000	4	32	39	0,820512821	10	6	36
2001	5	36	44	0,818181818	9	4	16
2002	6	34	45	0,755555556	8	2	4
2003	7	27	36	0,75	6	-1	1
2004	8	29	44	0,659090909	4	-4	16
2005	9	33	44	0,75	6	-3	9
2006	10	31	49	0,632653061	3	-7	49
2007	11	27	48	0,5625	2	-9	81
2008	12	21	50	0,42	1	-11	121

Contrast de Daniel	
SUMA	539
ESTADÍSTIC Ts	-0,884615385
ESTADÍSTIC Z	-2,933937315
ABS(Z)	2,933937315

Test de Daniel per a la variable proporció d'alumnes que superen la fase inicial del grau de Matemàtiques dels anys 2009-2016.

Any	t	SuperaFI	Total	SuperaFI/Tota	Rang	d	d^2
2009	1	38	59	0,6440678	1	0	0
2010	2	37	54	0,68518519	2	0	0
2011	3	37	54	0,68518519	2	-1	1
2012	4	44	56	0,78571429	7	3	9
2013	5	40	57	0,70175439	4	-1	1
2014	6	39	54	0,72222222	5	-1	1
2015	7	35	46	0,76086957	6	-1	1
2016	8	39	49	0,79591837	8	0	0

Contrast de Daniel	
SUMA	13
ESTADÍSTIC Ts	0,845238095
ESTADÍSTIC Z	2,236289799
ABS(Z)	2,236289799

Resultat d'aplicar la funció regressió sobre les dades de la proporció d'alumnes que superen la fase inicial de la llicenciatura de Matemàtiques dels anys 1997-2008.

Resumen								
Estadísticas de la regresión								
Coeficiente de correlación múltiple	0,88809758							
Coeficiente de determinación R^2	0,78871731							
R^2 ajustado	0,76758904							
Error típico	0,07902143							
Observaciones	12							
ANÁLISIS DE VARIANZA								
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*			
Regresión	1	0,23310269	0,233102687	37,32995436	0,0001142			
Residuos	10	0,06244387	0,006244387					
Total	11	0,29554656						
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
Intercepción	0,96778632	0,04863433	19,8992419	2,25468E-09	0,85942228	1,0761504	0,859422	1,0761504
Variable X 1	-0,04037436	0,0066081	-6,109824413	0,000114204	-0,05509813	-0,025651	-0,055098	-0,025651

Resultat d’aplicar la funció regressió sobre les dades de la proporció d’alumnes que superen la fase inicial del grau de Matemàtiques dels anys 2009-2016.

Resumen								
Estadísticas de la regresión								
Coeficiente de correlación múltiple	0,68910685							
Coeficiente de determinación R^2	0,47486824							
R^2 ajustado	0,38734629							
Error típico	0,03850895							
Observaciones	8							
ANÁLISIS DE VARIANZA								
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*			
Regresión	1	0,00804599	0,008045988	5,425704008	0,0586958			
Residuos	6	0,00889763	0,001482939					
Total	7	0,01694362						
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
Intercepción	0,60463773	0,03000593	20,1506083	9,70173E-07	0,53121586	0,67806	0,531216	0,6780596
Variable X 1	0,01384092	0,00594206	2,329314064	0,058695803	-0,00069877	0,028381	-0,0007	0,0283806

Prediccions pels anys 1997-2016 resultants d'aplicar el mètode de la tendència lineal usant les dades des del 2009-2016 de la proporció d'alumnes que superen la fase inicial.

Any	t	superaFI/Total	Predicció
1997	1	0,956521739	0,92741196
1998	2	0,866666667	0,88703761
1999	3	0,695652174	0,84666325
2000	4	0,820512821	0,80628889
2001	5	0,813953488	0,76591453
2002	6	0,744186047	0,72554018
2003	7	0,735294118	0,68516582
2004	8	0,651162791	0,64479146
2005	9	0,731707317	0,6044171
2006	10	0,581395349	0,56404275
2007	11	0,511627907	0,52366839
2008	12	0,355555556	0,48329403
2009	1	0,603773585	0,61847865
2010	2	0,609756098	0,63231957
2011	3	0,653061224	0,64616049
2012	4	0,739130435	0,66000142
2013	5	0,666666667	0,67384234
2014	6	0,65	0,68768326
2015	7	0,685714286	0,70152418
2016	8	0,727272727	0,71536511

Beta0	0,60463773
Beta1	0,01384092

Test de Daniel per a la variable proporció d'alumnes que pertany al CFIS del grau de Matemàtiques.

Any	t	Total CFIS	Total	CFIS/Total	Rang	d	d^2
2001	1	1	44	0,02272727	1,5	0,5	0,25
2002	2	2	45	0,04444444	3	1	1
2003	3	2	36	0,05555556	4	1	1
2004	4	1	44	0,02272727	1,5	-2,5	6,25
2005	5	3	44	0,06818182	5	0	0
2006	6	6	49	0,12244898	11	5	25
2007	7	5	48	0,10416667	9	2	4
2008	8	5	50	0,1	7	-1	1
2009	9	6	59	0,10169492	8	-1	1
2010	10	13	54	0,24074074	14	4	16
2011	11	5	54	0,09259259	6	-5	25
2012	12	10	56	0,17857143	12	0	0
2013	13	6	57	0,10526316	10	-3	9
2014	14	14	54	0,25925926	15	1	1
2015	15	11	46	0,23913043	13	-2	4
2016	16	16	49	0,32653061	18	2	4
2017	17	16	51	0,31372549	16	-1	1
2018	18	23	55	0,41818182	19	1	1
2019	19	17	53	0,32075472	17	-2	4

Contrast de Daniel	
SUMA	104,5
ESTADÍSTIC Ts	0,90833333
ESTADÍSTIC Z	3,85373196
ABS(Z)	3,85373196

Resultat d'aplicar la funció regressió sobre les dades de la proporció d'alumnes que pertany al CFIS del grau de Matemàtiques dels anys 2001-2019.

Resumen								
Estadísticas de la regresión								
Coeficiente de correlación múltiple	0,89900873							
Coeficiente de determinación R^2	0,8082167							
R^2 ajustado	0,79693533							
Error típico	0,05371904							
Observaciones	19							
ANÁLISIS DE VARIANZA								
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*			
Regresión	1	0,20673899	0,20673899	71,64171181	1,6795E-07			
Residuos	17	0,04905749	0,002885735					
Total	18	0,25579648						
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
Intercepción	-0,02535751	0,02565444	-0,988426073	0,336804591	-0,07948364	0,028769	-0,07948	0,028769
Variable X 1	0,01904468	0,00225004	8,46414271	1,6795E-07	0,01429751	0,023792	0,014298	0,023792

Prediccions pels anys 2001-2019 resultants d'aplicar el mètode de la tendència lineal usant les dades des del 2001-2019 de la proporció d'alumnes que pertany al CFIS.

Any	t	CFIS/Total	Predicció
2001	1	0,02272727	0
2002	2	0,04444444	0,01273185
2003	3	0,05555556	0,03177654
2004	4	0,02272727	0,05082122
2005	5	0,06818182	0,06986591
2006	6	0,12244898	0,08891059
2007	7	0,10416667	0,10795527
2008	8	0,1	0,12699996
2009	9	0,10169492	0,14604464
2010	10	0,24074074	0,16508933
2011	11	0,09259259	0,18413401
2012	12	0,17857143	0,20317869
2013	13	0,10526316	0,22222338
2014	14	0,25925926	0,24126806
2015	15	0,23913043	0,26031274
2016	16	0,32653061	0,27935743
2017	17	0,31372549	0,29840211
2018	18	0,41818182	0,3174468
2019	19	0,32075472	0,33649148

Beta0	-0,02535751
Beta1	0,01904468

Test de Daniel per a la variable proporció d'alumnes noies que pertany al CFIS del grau de Matemàtiques.

Any	t	Total	Noies	Noies/Total	Rang	d	d^2
2001	1	1	0	0	4	3	9
2002	2	2	0	0	4	2	4
2003	3	2	0	0	4	1	1
2004	4	1	1	1	19	15	225
2005	5	3	0	0	4	-1	1
2006	6	6	1	0,16666667	11	5	25
2007	7	5	0	0	4	-3	9
2008	8	5	1	0,2	13,5	5,5	30,25
2009	9	6	0	0	4	-5	25
2010	10	13	4	0,30769231	17	7	49
2011	11	5	1	0,2	13,5	2,5	6,25
2012	12	10	1	0,1	9	-3	9
2013	13	6	2	0,33333333	18	5	25
2014	14	14	0	0	4	-10	100
2015	15	11	2	0,18181818	12	-3	9
2016	16	16	1	0,0625	8	-8	64
2017	17	16	2	0,125	10	-7	49
2018	18	23	5	0,2173913	15	-3	9
2019	19	17	4	0,23529412	16	-3	9

Contrast de Daniel	
SUMA	658,5
ESTADÍSTIC Ts	0,42236842
ESTADÍSTIC Z	1,79195745
ABS(Z)	1,79195745