Institut de Recerca en Economia Aplicada Regional i Pública
Research Institute of Applied Economics

Document de Treball 2024/11 1/33 pág.
Working Paper 2024/11 1/33 pag.

"Difference-in-Difference models to estimate causal effects on auto insurers behavior"

Catalina Bolancé, Montserrat Guillen, Ana María Pérez-Marín & Anna-Patrícia Orteu

UBIREA Institut de Recerca en Economia Aplicada Regional i Pública

UBIREA

Institut de Recerca en Economia
Aplicada Regional i Pública
UNIVERSITAT DE BARCELONA

WEBSITE: www.ub.edu/irea/ • CONTACT: irea@ub.edu

Abstract

The Difference-in-Difference (DiD) method is useful to test if an event has effects in a given outcome using non-experimental data. Based on DiD method, we propose alternative panel models to estimate the causal effects of the traffic accidents on driving behavior patterns: the total annual driving distance in km, the percent of km circulated above the speed limits, in urban areas and at night. We use a data set provided by an ”insurtech” company that uses car sensors to measure driving data over a period of three years. The estimation results show as the causal effects of accidents are different if we consider frequency of accidents, type of damages and whose fault is the accident. Furthermore, different profiles of policyholders in function of drivers and cars characteristics are associated with specific causal effects.

JEL classification: C21 C55 G22

Keywords: Telematics, average treatment effect, panel data, damages, fault, drivers heterogeneity.

Catalina Bolancé (corresponding author). Department of Econometrics Riskcenter-IREA. University of Barcelona, Barcelona, Spain. E-mail: bolance@ub.edu

Montserrat Guillen. Department of Econometrics Riskcenter-IREA. University of Barcelona, Barcelona, Spain. E-mail: mguillen@ub.edu

Ana María Pérez-Marín. Department of Econometrics Riskcenter-IREA. University of Barcelona, Barcelona, Spain. E-mail: amperez@ub.edu

Anna-Patrícia Orteu. Department of Econometrics Riskcenter-IREA. University of Barcelona, Barcelona, Spain. E-mail: annap.orteu@ub.edu

Acknowledgments: This work was supported by Grant TED2021-130187B-I00 funded by MCIN/AEI/10.13039/501100011033 and by the ”European Union NextGenerationEU/PRTR”.

1 Introduction

We analyze how an accident causes changes in the auto policyholder's way of driving, e.g., if after the accident the driver becomes more careful or reduces the number of kilometers driven.

We carried out a Difference-in-Differences (DiD) analysis with alternative outcomes on driving patterns and driving conditions, a type of information that is available through sensors data regularly collected by "insurtech" firms. With this aim, we selected a sample of policyholders that are observed over three years and they have the following characteristics: the first year (pre- treatment period) no drivers had accidents, the second year (treatment period) a few drivers had one or more accidents and the third year (post-treatment period) no drivers reported accidents either. We assume that having accident(s) in the second year is the "treatment" that could cause changes in the way of driving in the third year. We tested if these changes are statistically significant or not.

In short, we show how insurance companies can use their data in a context similar to an experimental design and identify accidents causal effects, which can help better adjust the price of insurance to the anticipated profile of the insured. Furthermore, in a more general context, analyzing changes in driving patterns also has implications with regard to road safety.

The DiD is a statistical technique used to estimate the causal effect of a treatment on an outcome of interest by using non-experimental data sets, this means that we can not control the assignment of the treatment, this could be assigned by chance. The basic idea behind the DiD analysis is to compare the change in the outcome over time, pre-treatment and post-treatment periods, between a treatment group (those who reported accidents) and a control group (those who did not report accidents). The untreated group should allow us to identify the temporal variation without presence of treatment in the treated group. By comparing the difference in the changes between the two groups it is estimated the causal effect of the treatment.

The DiD based analysis can be applied to a wide range of research questions, such as evaluating the impact of a new policy on health outcomes, estimating the effect of a marketing campaign on sales or assessing the impact of a natural disaster on economic outcomes. Some examples of application are found in Di Tella and Schargrodsky (2004), Galiani et al. (2005) and Jeffrey et al. (2011). Furthermore, for a recent review on innovative methodologies on DiD based analysis see Roth et al. (2023).

An important point is that the DiD method can help to address the issue of selection bias, which arises when treatment and control groups differ in ways that might affect the outcome of interest. By comparing the change in the outcome over time within each group, we can take into account any pre-existing differences between the groups.

The DiD based analysis can be carried out using a linear regression where the treatment effect is estimated with the coefficient associated with the binary variable that identifies the treated group in the post-treatment period, this is the DiD model. In this work, this model allows us to estimate the causal effects on alternative outcomes of the occurrence of one o more accidents. We present an innovative application on auto insurance, where the treatment group includes those policyholders who have one or more accidents along a given treatment period and the control group includes those insureds who do not have accidents during the same period.

We compare the differences one year before (pre-treatment period) and one year after (post- treatment period) the accident(s) between the treatment group and the control group, for a set of variables that measure annual driving patterns (”percentage distances driven above the posted speed limits” and ”total distance in kilometers”) and driving conditions (”percentage kilometers at night” and ”percentage kilometers in urban areas”). The aim is to estimate the average treatment effect on the treated group (ATT), i.e. the causal effect on the group of policyholders reporting one or more accidents.

The causal effect of having accidents could depend on the number or type of accidents. For example, the causal effect of an accident with body injuries (BI) may be different from that with only property damage (PD). Alternatively, the effect of an accident where the driver is at fault may also be different from that of an accident where the driver is not at fault. We note that the treatment and control groups are defined in a simplified way and the DiD model is generalized adding multiple treatment effects.

Some relevant assumptions have to be accepted when estimating the DiD model. The main assumption is the parallel trends, which means that the trend in the outcome would have been the same for the treatment and control groups in the absence of the event. If this assumption is violated the DiD classical estimator based on linear regression is biased and inconsistent, if this assumption is not met we will have to consider other techniques, e.g., some semi-parametric estimator can be used (see Abadie, 2005; Athey and Imbens, 2006). However, there is not statistical inference to verify whether there are parallel trends and their analysis has to be based on the results found in similar studies and/or the past experiences.

Alternatively, parallel trends can be assumed conditional on the values of a set of covariates, i.e. heterogeneity assumption. In this case the model has to be expanded by adding multiplicative effects of covariates with post-treatment period and treatment effects. In our study, this assumption is fundamental, given that causal effects of accidents on driving behavior depend on the characteristics of the drivers and cars.

Another assumption of the DiD model is consistency or non anticipatory effects, that implies that in the pre-treatment period the outcome of the treatment group is not affected by the future event. In our case, this implies that in the period before treatment the driving patterns are not affected by a possible future accident(s), i.e., if the accident did not occur, driving behavior would not change. However, these driving patterns could be different in both groups in the pre-treatment period. We have commented previously in this introduction that DiD method addresses these differences between groups, in other words, it takes into account the selection bias.

The structure of the paper is as follows. In Section 2, we describe the general notation and the DiD model. A new DiD model in insurance is presented in Section 3. In Section 4, the data and the estimation results are described. Finally, the paper ends with conclusions in Section 5.

2 The Difference-in-Difference method

We have a set of N individuals that are observed for at least three times: pre-treatment, treatment and post-treatment periods. Commonly, in analysis based on the DiD methodology, the treatment period is reduced to an instant in time, so all is considered to happen in two periods, before and after the treatment. Here, taking into account the available information, the occurrence of the accident is assumed to possibly occur over a certain period, which in our case is one year. We identify the two compared periods as $t = 1, 2$, that are the pre-treatment and the post-treatment periods, respectively, in these two periods the policyholders did not report accident(s). The individuals form two groups, depending on whether they had reported accident(s) in the treatment period, the control group of those insureds that did not have accidents and the treatment group of those who suffered one or more accidents. From now on, in this section, we will use some similar notation as Roth et al. (2023).

Let D_i, $i = 1, \ldots, N$, be the variable that identifies the control ($D_i = 0$) and the treatment ($D_i = 1$) groups. Let Tr_{it}, $t = 1, 2$, be the variable that identifies the period of occurrence of the treatment in the treatment group, note that $Tr_{it} = D_i \times I(t = 2)$, where $I(\cdot)$ is the indicator function, that is equal 1 if the condition in parentheses is true and equal 0 on the contrary, i.e. Tr_{it} is equal to 1 when the treatment group is doing in the last period.

If we could observe all possible outcomes in both groups and both periods, for each individual i in period t we would have the following information:

- $Y_{it}(0, 0)$ is the outcome in period t if the individual i is untreated in both periods.
- $Y_{it}(0, 1)$ is the outcome in period t if the individual i is untreated in the first period and treated in the second period.
- $Y_{it}(1, 0)$ is the outcome in period t if the individual i is treated in the first period and untreated in the second period.
- $Y_{it}(1, 1)$ is the outcome in period t if the individual i is treated in both periods.

In practice, we can only observe the values of $Y_{it}(0, 0)$ for the individual in the control group and $Y_{it}(0, 1)$ for the individual in the treatment group; the values of $Y_{it}(1, 0)$ and $Y_{it}(1, 1)$ are potential outcomes that we can not observe. Given the observed outcomes, we can simplify the notation in function of the post-treatment period, $Y_{it}(0) = Y_{it}(0, 0)$ and $Y_{it}(1) = Y_{it}(0, 1)$. In general, for $i = 1, \ldots, N$ and $t = 1, 2$, the random variable that is measured can be expressed as:

$$Y_{it} = D_i Y_{it}(1) + (1 - D_i) Y_{it}(0).$$

The aim of the DiD method is to estimate the ATT in period $t = 2$, this is:

$$ATT_2 = E\left[Y_{i2}(1) - Y_{i2}(0) | D_i = 1\right]. \tag{1}$$

The problem with the expression (1) is that $Y_{it}(0)$ for the treatment group can not be observed in $t = 2$ and then we must have a value of $Y_{i2}(0)$ to replace in (1). So, we have to use the control group information to identify it. A way to can calculate the ATT_2 consists of adding two assumptions. The first is the parallel trends, that is expressed as:

$$E\left[Y_{i2}(0) - Y_{i1}(0)|D_i = 1\right] = E\left[Y_{i2}(0) - Y_{i1}(0)|D_i = 0\right],$$

i.e., if the treatment was not given, the difference between pre- and post-treatment periods is the same for the two groups. An example is shown in Figure 1. The parallel trends assumption can not be tested, it is usually analyzed using previous studies.

The second assumption is called no anticipatory effects, this is related with the treatment group and it is expressed as $Y_{i1}(0) = Y_{i1}(1)$ with $D_i = 1$. This assumption implies that, for the treated group, the value of the outcome in the first period would have been the same no matter what happens in the future.

Using these two assumptions, parallel trends and non anticipatory effects, we can identify the ATT_2 as follows. First, from parallel trends expression we obtain that:

$$E\left[Y_{i2}(0)|D_i = 1\right] = E\left[Y_{i1}(0)|D_i = 1\right] + E\left[Y_{i2}(0) - Y_{i1}(0)|D_i = 0\right]$$

and, second, from non anticipatory effect we have

$$E\left[Y_{i1}(0)|D_i = 1\right] = E\left[Y_{i1}(1)|D_i = 1\right].$$

So, we can write $Y_{i2}(0)$ in function of the observed value as:

$$\begin{aligned} E\left[Y_{i2}(0)|D_i = 1\right] &= E\left[Y_{i1}(1)|D_i = 1\right] + E\left[Y_{i2}(0) - Y_{i1}(0)|D_i = 0\right] = \\ &= E\left[Y_{i1}|D_i = 1\right] + E\left[Y_{i2} - Y_{i1}|D_i = 0\right]. \end{aligned}$$

Replacing the previous expression in (1) we obtain that, in function of the observed outcomes, the ATT_2 is:

$$\begin{aligned} ATT_2 &= E\left[Y_{i2}(1) - Y_{i2}(0)|D_i = 1\right] = E\left[Y_{i2}(1)|D_i = 1\right] - E\left[Y_{i2}(0)|D_i = 1\right] = \\ &= E\left[Y_{i2}|D_i = 1\right] - E\left[Y_{i1}|D_i = 1\right] - E\left[Y_{i2} - Y_{i1}|D_i = 0\right] = \\ &= E\left[Y_{i2} - Y_{i1}|D_i = 1\right] - E\left[Y_{i2} - Y_{i1}|D_i = 0\right]. \qquad (2) \end{aligned}$$

The result expressed in (2) gives its name to the DiD method and can be estimated from sample means differences.

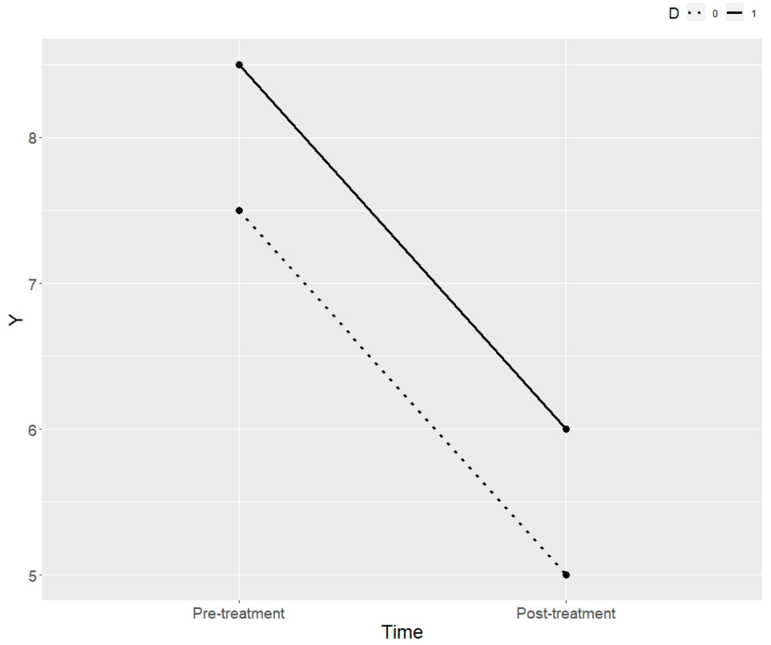

Figure 1: Parallel trends assumption.

2.1 The alternative models

The average treatment effect can also be identified from the following linear model that can be estimated using Ordinary Least Squares (OLS):

$$Y_{it} = \alpha + \theta \cdot D_i + \zeta \cdot I(t = 2) + \tau \cdot Tr_{it} + \epsilon_{it}, \tag{3}$$

where ϵ_{it}, for $i = 1, ..., N$ and $t = 1, 2$, are independent and identically distributed (iid) random errors. We remember that $Tr_{it} = D_i \times I(t = 2)$ takes value 1 for treatment group in post-treatment period. From (3) we obtain $E[Y_{i2} - Y_{i1}|D_i = 1] = (\alpha + \theta + \zeta + \tau) - (\alpha + \theta)$ and $E[Y_{i2} - Y_{i1}|D_i = 0] = (\alpha + \zeta) - \alpha$, then it is direct to obtain that the ATT_2 defined in (2) is equal to the parameter τ, which can be interpreted as the causal effect of the treatment on Y_{it}.

The dependent variable in the model defined in (3) refers to individual i in period t, i.e. a panel data is required to estimate causal effect τ. If individual effects in panel data model specification are significant the ϵ_{it} are not iid and the OLS estimator of τ is not efficient. Alternatively, a panel data model with individual fixed effects α_i and temporal fixed effects δ_t is used, this is called the two-way fixed effects (TWFE) model and it is specified as follows:

$$Y_{it} = \alpha_i + \delta_t + \tau \cdot Tr_{it} + \epsilon_{it}, \tag{4}$$

where to prove that τ is equal to ATT_2 is direct taking into account that $E\left[Y_{i2} - Y_{i1}|D_i = 1\right] = (\alpha_i + \delta_2 + \tau) - (\alpha_i + \delta_1)$ and $E\left[Y_{i2} - Y_{i1}|D_i = 0\right] = (\alpha_i + \delta_2) - (\alpha_i + \delta_1)$. The TWFE model can be estimated with the classical within estimator for individual and temporal fixed effects in panel data models. If individual fixed effects do not exist the OLS estimator of (3) is the best. Both estimators, the OLS and within, are consistent if the parallel trends and non anticipatory effects assumption is true. This last assumption is generally acceptable, but about the parallel trends there could be reasonable doubts, given that we have to assume that for all the policyholders the trends are the same.

An alternative is to assume parallel trends conditional on covariates:

$$E\left[Y_{i2}(0) - Y_{i1}(0)|D_i = 1, X_i\right] = E\left[Y_{i2}(0) - Y_{i1}(0)|D_i = 0, X_i\right],$$

where $X_i = (X_{i1}, ..., X_{ik})'$ is a vector of values of a set of time-invariant covariates that can be observed on the pre-treatment period. Directly, it is easy to deduce that the ATT_2 conditioned on X_i can be expressed as:

$$ATT_2(X_i) = E\left[Y_{i2} - Y_{i1}|D_i = 1, X_i\right] - E\left[Y_{i2} - Y_{i1}|D_i = 0, X_i\right]. \qquad (5)$$

To identify the $ATT_2(X_i)$, for each treated case with values X_i there must be at least some untreated cases with the same values of covariates. The model in (4) adding covariates can be extended as follows:

$$Y_{it} = \alpha_i + \delta_t + \tau \cdot Tr_{it} + \beta_1 \cdot X_{i1} \cdot I(t = 2) + ... + \beta_k \cdot X_{ik} \cdot I(t = 2) + \epsilon_{it}. \qquad (6)$$

With the model specified in (6) we can avoid specification errors by adding explanatory variables that have significant effects on the dependent variable. However, the causal effect of the treatment remains homogeneous for all individuals. To take into account the heterogeneity assumption, i.e. the ATT is different for each vector X_i, the TWFE model must include the interaction effects between treatment variable and covariates, in this case the TWFE model is:

$$Y_{it} = \alpha_i + \delta_t + \tau \cdot Tr_{it} + \beta_1 \cdot X_{i1} \cdot I(t = 2) + ... + \beta_k \cdot X_{ik} \cdot I(t = 2) + \gamma_1 \cdot X_{i1} \cdot Tr_{it} + ... + \gamma_k \cdot X_{ik} \cdot Tr_{it} + \epsilon_{it}, \qquad (7)$$

Using the model specified in (7), the causal effects are identified as:

$$\begin{aligned} ATT_2(X_i) &= E\left[Y_{i2} - Y_{i1}|D_i = 1, X_{ij}\right] - E\left[Y_{i2} - Y_{i1}|D_i = 0\right] \\ &= \tau + \beta_1 \cdot X_{i1} + ... + \beta_k \cdot X_{ik} + \gamma_1 \cdot X_{i1} + ... + \gamma_k \cdot X_{ik} - (\beta_1 \cdot X_{i1} + ... + \beta_k \cdot X_{ik}) \\ &= \tau + \gamma_1 \cdot X_{i1} + ... + \gamma_k \cdot X_{ik}. \end{aligned} \qquad (8)$$

Note that the covariates may be time-variant, in this case $X_{it} = (X_{it1}, ..., X_{itk})'$, $t = 1, 2$, and, in addition to the interaction effects of covariates with $t = 2$ $(X_{itj} \cdot I(t = 2), j = 1, ..., k)$ and Tr_{it} $(X_{itj} \cdot Tr_{it}, j = 1, ..., k)$, the covariates effects without interactions must be added in the model (7). In the insurance application presented in this paper we had time-invariant covariates.

The main drawback of the model (7) is the assumption that there is a linear relationship between the covariates and the dependent variable. Alternatively, Abadie (2005) proposes a nonparametric approach based on inverse weight estimators (IWE) to directly estimate the ATT_2 conditional on covariates.

3 Using Difference-in-Difference model in insurance

Each policyholder is observed along three consecutive periods (three years), in the first year (pre-treatment period or $t = 1$) the policyholders have not had accidents, in the second year (treatment period) a few policyholders have one or more accidents and, finally, in the third year (post-treatment period or $t = 2$) the insureds have not had accidents either. The aim is to compare pre- and post-treatment periods, before and after the accident(s), and identify the causal effects of accident(s) on a set of outcomes related with the behavior of the drivers.

Let n_i, $i = 1, ..., N$, be the number of reported accidents of policyholder i in the treatment period. First, in the model specified in (7) we define:

- $Tr_{it} = 1$ if $n_i > 0$ and $t = 2$ and
- $Tr_{it} = 0$ on the contrary,
- $X_i = (X_{i1}, ..., X_{ik})'$ is a set of covariates related with characteristics, which are observed in the pre-treatment period and are time-invariant.

Furthermore, we know that the treatment causal effect could differ depending on the number or type of accidents, i.e. we need to consider the intensity of treatment taking into account the number of accidents or their different types, depending on their severity or the driver's responsibility.

An alternative to the model specified in (7) is to replace the binary treatment variable (Tr_{it}) by a treatment variable that is equal to the number of accidents in the treatment period. However, this strategy has two difficulties. First, we must take into account in the model that covariates effects could be different for the different treatment intensities. Second, in general, the number of reported accidents is a variable with high right skewness, i.e., a lot of values are equal to 0 or 1 and there are very few cases with values larger than 2. In practice, we have observed that we do not have sufficient information to identify causal effects in function of the number of accidents.

We propose specify the TWFE model with two treatment variables $Tr_{it}^{(1)}$ and $Tr_{it}^{(2)}$, that represent two different intensities in the treatment:

- in function of the frequency, $Tr_{it}^{(1)} = 1$ if $n_i = 1$ and $Tr_{it}^{(1)} = 0$ otherwise, and $Tr_{it}^{(2)} = 1$ if $n_i > 1$ and $Tr_{it}^{(2)} = 0$ otherwise,

- in function of the damage type, $Tr_{it}^{(1)} = 1$ if there are one or more accidents with BI and $Tr_{it}^{(1)} = 0$ otherwise, and $Tr_{it}^{(2)} = 1$ if all the accidents are only with PD and $Tr_{it}^{(2)} = 0$ otherwise, or

- in function of fault, $Tr_{it}^{(1)} = 1$ if there are one or more accidents where the driver is at fault and $Tr_{it}^{(1)} = 0$ otherwise, and $Tr_{it}^{(2)} = 1$if in all the accidents the driver is not at fault and $Tr_{it}^{(2)} = 0$ otherwise.

Note that in the three cases the treatment variables are disjointed. With two treatment variables the TWFE DiD model especified in (4) and (6) are obtained replacing $\tau \cdot Tr_{it}$ by $\tau_1 \cdot Tr_{it}^{(1)} + \tau_2 \cdot Tr_{it}^{(2)}$. If we want taking into account the heterogeneity assumption the model is:

$$\begin{aligned} Y_{it} = \alpha_i + \delta_t + \tau_1 \cdot Tr_{it}^{(1)} + \tau_2 \cdot Tr_{it}^{(2)} &+ \beta_1 \cdot X_{i1} \cdot I(t = 2) + \ldots + \beta_k \cdot X_{ik} \cdot I(t = 2) \\ &+ \gamma_1^{(1)} \cdot X_{i1} \cdot Tr_{it}^{(1)} + \ldots + \gamma_k^{(1)} \cdot X_{ik} \cdot Tr_{it}^{(1)} \\ &+ \gamma_1^{(2)} \cdot X_{i1} \cdot Tr_{it}^{(2)} + \ldots + \gamma_k^{(2)} \cdot X_{ik} \cdot Tr_{it}^{(2)} + \epsilon_{it}. \end{aligned} \tag{9}$$

From model (9), the heterogeneous causal effect depending on X_i and treatment type l, $l = 1, 2$, is:

$$ATT_2^{(l)}(X_i) = \tau_l + \gamma_1^{(l)} \cdot X_{i1} + \ldots + \gamma_k^{(l)} \cdot X_{ik}. \tag{10}$$

For the particular case of homogeneous causal effect, the average treatment l, $l = 1, 2$, in treatment period is:

$$ATT_2^{(l)} = \tau_l \tag{11}$$

4 Results in insurance data analysis

We have a sample of $N = 3,611$ policyholders having a car insurance in a Spanish company for at least three consecutive years (2009, 2010 and 2011). The first year (t = 2009) is the pre-treatment period, the last year (t = 2011) is the post-treatment period and the middle (t = 2010) is the treatment period when a small part of the drivers have one or more accidents. This database has been extracted from the one used in Pérez-Mar´ın et al. (2019) for an inference analysis on speed changes in young drivers after an accident. The insureds in the sample were selected as follows:

- The policyholders had driven 100 kilometers or more in the pre-treatment period.

- The policyholders had no accident(s) in the pre-treatment and post-treatment periods.

- The policyholders had not had changes in the characteristic of their car insurance policies over the analyzed period.

4.1 Background on telematics driving data

Telematics driving data refers to the collection of information from individual vehicles in motion -Kirushanth and Kabaso (2018) connect telematics data collection with road safety. There has been an explosion of articles that analyzes this type of data jointly with accident information. Gao et al. (2023) analyze insurance claim frequency of commercial trucks using both Poisson regression and several machine learning models, including regression tree, random forest, gradient boosting tree, XGBoost and neural network. They insist on the need to provide interpretation of predictive models in order to calculate transparent insurance premium calculation, as required by regulators.

Telematics information helps to predict claim frequency (Baecke and Bocca, 2017; Ayuso et al., 2019; Huang and Meng, 2019; Winlaw et al., 2019; Meng et al., 2022; Gao et al., 2022). Guillen et al. (2019) insist on the zero-inflation phenomenon. The use of distance driven was the first factor to be identified as a determinant of increase accident risk. Lemaire et al. (2016) find that annual mileage is an extremely powerful predictor of the number of claims where the driver was at fault. Verbelen et al. (2018) find that such variables increase the predictive power and render the use of gender as a rating variable redundant. Ayuso et al. (2016) discuss differences between male and female drivers and study some specific features of young drivers (Ayuso et al., 2014; Pérez-Marín et al., 2019).

Boucher et al. (2017) include policy duration in the risk exposure component in classical models (Lord et al., 2005; Boucher et al., 2007, 2009). Duval et al. (2024) introduce a longitudinal model that accounts for the dependence between contracts from the same insured. Boucher and Turcotte (2020) show that an approximately linear relationship between distance driven and claim frequency can be derived and that can be used to compute the premium surcharge for additional kilometers driven, or as the underlying model to construct Pay-as-you-drive (PAYD) insurance. Cheng et al. (2023) conclude that PAYD insurance is more efficient than fuel tax in reducing mileage due to the concavity relation of premium and driving distance. Litman (2007) pioneered the defense of distance-based pricing. He stated that distance-base pricing is technically and economically feasible, and can provide significant benefits to motorists and society (Ferreira Jr and Minikel, 2012). Denuit et al. (2019) state that the multivariate nature of telematics signals can be incorporated in usage-based insurance. Considering some dependencies is necessary (Bolancé et al., 2020). Henckaerts and Antonio (2022) establish the first steps towards dynamical real-time pricing. A complete review of papers related to insurance can be found in Eling and Kraft (2020) and for those related to accident risk analysis there are a few revisions (Boylan et al., 2024; Chauhan and Yadav, 2024)

Variable selection has played a central role in model specification with telematics variables (Chan et al., 2022; Jeong, 2022). Duval et al. (2022) develop a method to determine how much information about policyholders' driving should be kept by an insurer. Using real data from a North American insurance company, that find that telematics data become redundant after about 3 months or 4,000 km of observation, at least from a claim classification perspective. Wüthrich and Merz (2019) suggest combining classical generalized linear models and is extensions with neural networks. This is also confirmed by Duval et al. (2024) who state that combined models exhibit superior performance compared to log-linear models that rely on manually engineered telematics features. Pesantez-Narvaez et al. (2019) mention the difficulty to improve the predictive performance of parametric models in real data situation. Another strand of research concentrates on risky events rather than actual accidents (Guillen et al., 2021b, 2020; Sun et al., 2021). Telematics data are also combined with external context data like traffic congestion, road condition and weather (Ma et al., 2018; Reig Torra et al., 2023; Masello et al., 2023) and can be informative of extreme behaviours (Guillen et al., 2021c,a; Pitarque and Guillen, 2022).

4.2 Results

We have combined telematic information with accidents reported to the insurance company for our sample of $3,611$ policyholders The mean of the number of reported accidents in treatment period is 0.1025. In Table 1 we show the frequency distribution of the number of accidents. We have 302 insureds that had accident(s) in treatment period, that represents the 8.36% of the drivers in the sample. On the one hand, among these, 145 drivers (4.02%) had accident(s) where they were at fault and 157 (4.35%) had only accidents where they were not at fault. On the other hand, 58 drivers (1.61%) suffered accident(s) with BI and the rest, 244 (6.75%), only have accident(s) with PD.

Table 1: Distribution of the number of accidents in treatment period (2010) and number of policyholders that have reported accidents.

$n_i =$	0	1	2	3	4	5	Total insureds with accidents
Frequency	3309	241	56	4	0	1	302
Frequency not at fault	3454	120	32	4	0	1	157
Frequency at fault	3466	121	24	0	0	0	145
Frequency PD	3367	184	55	4	0	1	244
Frequency BI	3553	57	1	0	0	0	58

In Table 2 some descriptive statistics of the outcome variables are shown: the mean, the standard deviation (STD), the minimum (Min), the median and the maximum (Max), respectively. These variables are measured annually and they are: total distance in km (Distance), percentage distances driven above the posted speed limits (Speed), percentage of km in urban areas (Urban) and percentage of km at night (Night). The variable Distance is the only one that is measured in absolute values, the rest are percentages. We note that the mean of the variable Distance decreases by more than 1500 km between pre- and post-treatment periods. The average percentage distances driven above the posted speed limits and in urban areas also decrease and the percentage of km at night practically remains constant. We analyze if these changes are different for drivers with or without accident(s) in 2010 (treatment period), i.e. the question is; is there a causal effect of accident(s) on the outcomes? Furthermore, we analyze if the causal effect is homogeneous or depends on a set of covariates. These time-invariant covariates are described in Table 3.

Table 2: Descriptive statistics of outcome variables in pre- and post-treatment periods.

Outcome	t	Mean	STD	Min	Median	Max
Distance	2009	4833.1222	4785.4761	100.4160	3358.6900	34484.3290
	2011	3243.7758	3098.5025	0.0000	2399.5550	34295.6730
Speed	2009	8.6228	9.1769	0.0859	5.2937	68.6549
	2011	6.1667	7.4436	0.0000	3.6083	60.9904
Urban	2009	27.2223	15.4296	0.0000	23.8248	95.9362
	2011	25.6910	15.8737	0.0000	22.1854	100.0000
Night	2009	6.4167	6.6493	0.0000	4.3482	53.8559
	2011	6.8884	8.1766	0.0000	4.1791	100.0000

Table 3: Covariates used in the models (N=3,611).

		Mean	STD
Agei=25	=1 if agei=25, =0 on the contrary	0.6333	0.4820
Age licensei=5	=1 if age of licensei=5, =0 on the contrary	0.6849	0.4646
Parking=yes	=1 if car uses parking, =0 on the contrary	0.6483	0.4776
Woman	=1 if woman, =0 in man	0.4954	0.5000
City	=1 if zone is large metropolitan area, =0 on the contrary	0.1448	0.3520
Car agei=4	=1 if age of cari=4, =0 on the contrary	0.5763	0.4942
Poweri=100	=1 if car poweri=100 horsepower, =0 on the contrary	0.5910	0.4917

In Table 3 we show as the covariates used in the DiD model are defined as binary variables.

We use this strategy to avoid the identification problems, given that for each X_i we need to have observed drivers with and without treatment. The binary variables are defined using criteria that frequently are used by the insurance companies to group their policyholders.

For each outcome variable in Table 2, we estimated four models depending on how the treatment effects are included in the panel model, these models are described in Table 4.

Table 4: Treatment effects in each TWFE panel model.

Model I	Tr_{it}	Accidents	=1 if there were accident(s) in 2010, =0 on the contrary
Model II	Tr^1_{it}	Accidents=1	=1 if there was only one accident in 2010, =0 on the contrary
	Tr^2_{it}	Accidents¿1	=1 if there were 2 or more accidents in 2010, =0 on the contrary
Model III	Tr^1_{it}	At fault	=1 if there were accident(s) where the driver was at fault in 2010, =0 on the contrary
	Tr^2_{it}	Not at fault	=1 if there were accident(s) where the driver was not at fault in 2010, =0 on the contrary
Model IV	Tr^1_{it}	BI	=1 if there were accident(s) with BI in 2010, =0 on the contrary
	Tr^2_{it}	PD	=1 if there were no accidents with BI in 2010, =0 on the contrary

We first estimate the TWFE models assuming homogeneous causal effects of the treatments on outcomes, i.e. unconditional to the covariates vector X_i. In this case the covariates were included additively as in model especified in (6). The covariates included additively in each model for each outcome are selected using the stepwise algorithm of the R function stepAIC(), that is based on selecting the model with the best AIC (Akaike Information Criterion). In Table 5 are shown the estimated ATT_2 associated with each treatment group defined in Table 4, and the p-value of the statistic for testing if the effect is different from zero. We assume that the estimated effects are statistically significant when the p-value is ≤ 0.05. The full results of estimated models are shown in Table A1 in Appendix A, we also add the panel fixed effects test, the determination coefficient (R^2) and the adjusted determination coefficient (Adjusted R^2).

We remember that the estimated parameters in Table 5 are consistent if non anticipatory effects, parallel trends and homogeneity assumptions are true. In our case, the first assumption implies that in 2009 (pre-treatment period) drivers in treatment group have the same behavior as the case of not having any accident in 2010, i.e. the change of the behavior must be due only to the occurrence of the accident. The second assumption involves that the outcome trends for the groups with and without accident(s) would have the same slopes if there had been no accidents. Finally, the homogeneity assumption is related with the fact that the causal effects are the same for all policyholders whatever its characteristics. We have tested that this last homogeneity assumptions is not true, given that we have found that some interaction effects between treatment variable and covariates are statistically significant. So, the estimator use in Table 5 is non consistent. If we compare these results with those associated with different drivers profiles in Tables A2, A3, A4 and A5, respectively, in Appendix B, we can find significant differences.

Table 5: Estimated ATT_2 for each treatment group and each outcome.

		Distance		Speed		Urban		Night	
Model	Treatment	ATT_2	p-value	ATT_2	p-value	ATT_2	p-value	ATT_2	p-value
I	Accidents	765.5157	0.0009	-0.0615	0.8571	-0.1409	0.8083	0.8394	0.0397
II	Accidents=1	885.7809	0.0005	0.1585	0.6755	-0.0989	0.8781	0.8457	0.0616
	Accidents¿1	288.2723	0.5594	-0.9332	0.2034	-0.3074	0.8057	0.8144	0.3529
III	At fault	433.7275	0.1811	-0.5254	0.2752	-0.7209	0.3795	1.4163	0.0139
	Not at fault	1071.4412	0.0006	0.3674	0.4282	0.3941	0.6176	0.3072	0.5792
IV	BI	-43.4732	0.9315	-1.5716	0.0366	0.4774	0.7094	-0.3883	0.6655
	PD	957.3485	0.0002	0.2965	0.4309	-0.2877	0.6537	1.1305	0.0120

To estimate $ATT_2(X_i)$ and $ATT^{(l)}_2(X_i)$ defined in (5) and (10), respectively, that are associate with different profiles of drivers we use the models specified in (7) for one treatment and (9) for two treatments. Similarly to additive model, we use the R function stepAIC() for selecting the covariates and the interaction effects with the treatment variable(s). The causal effects are shown in Table 6, in this case we show the causal effects that are significant at 5% and 10%, although we will focus on those that are significant at 5%. The statistic for testing is associated to the sum of the coefficient following (10). The full results of the estimated panel models are shown in Appendix B.

The estimated values in Table 6 are interpreted as follows:

- The values associated with the treatment variables (Accidents, Accidents=1, Accidents¿1, At fault, Not at fault, BI and PM) are the causal effects of accidents if all the covariates take value 0, i.e. the reference group is made up of older men, with more that 5 years with license, who do not use parking, they do not drive in large metropolitan areas and have older and high powered cars.

- The values associated with the interaction effects of treatments with each covariate are the causal effects when this covariate takes value 1 and the rest of covariates are equal to 0, i.e. in Table 6 we show the results of whether the causal effect changes significantly with respect to the reference group if the covariate value is 1.

Focusing on Distance outcome, we observe that, when the number of accidents is > 1 these cause the decreasing of annual driven distance in the reference group. This same effect is also negative in other groups of drivers that are represented in Table 6. Remaining the rest of the covariates at their reference values, the negative causal effects are found when: the age of the license becomes ≤ 5 or the parking is used or the driver is a woman or the power is ≤ 100. Furthermore, when there are BI the Distance outcome is also reduced in the reference group and this decrease is greater when "City"= 1 and lower when parking is used. In general, these results show that the frequency and the severity of the accidents reduce the risk exposure, measured in annual driven distance in km, in some profiles of drivers. All of these results contradict those shown in Table 5, where the causal effects of accident(s) on the Distance variable were positive or non-significant.

In Table 6 the significant causal effects of accidents on the Speed outcome are mostly negative in the four models. In Model I the effect of accident(s) on percentage distances driven above the posted speed limits is the most negative in the reference group, i.e. older men, with more than 5 years with a license, who do not use parking, they do not drive in large metropolitan areas and have older and high powered cars. This effect is also negative but weaker when the car age becomes ≤ 4. In Model II the most negative significant estimated ATT_2 is associated with Accidents> 1 for policyholders that use parking and the rest of the characteristics are equal to those of the reference group. The negative causal effects in Model III are done when drivers have accident(s) where they are at fault and parking is used, keeping the rest of the covariates at their reference values. When drivers have all accident(s) where they are not at fault the negative effect are significant when the age of license becomes ≤ 5 and, alternatively, the positive effect is significant when the car power becomes ≤ 100.

Focusing in Model IV for Speed outcome, the most negative causal effect is associated to accident(s) with BI for drivers in large metropolitan areas ("City"=1), remaining the rest of the covariates equal to 0. This negative effect also is done for accidents with BI for reference group. When the accident(s) only cause PD there is a significant negative effect when the age of license is ≤ 5 and the rest of the characteristics equal to those of the reference group.

If we compare these results for Speed outcome with those that were shown in Table 5 we can conclude that, if we do not take into account the values of the covariates, we only find negative and significant causal effects when there are accident(s) with BI. Furthermore, we compare our results with those that shown in Pérez-Mar´ın et al. (2019), these authors use a sample of drivers that have an accident and compare the Speed outcome six moth before and after the accident, they conclude that only the men decrease the percent of km circulated above the speed limit given that they tend to have a higher previous percent. Unlike this analysis, the DiD models have allowed us to find different types of drivers who change the value of the outcome Speed in the event of accidents.

For the Urban outcome the significant effects in Table 6 are less numerous than for the rest of outcomes. In Model I, a significant negative effect of the treatment is found when the age of car is ≤ 4 and the same effect is positive for women. In Model II the negative treatment effect is found when frequency of accidents is 1 for drivers with age ≤ 25. On the contrary, the positive treatment effect is done when the frequency is greater than 1 for women. In Model III we have also found a negative and a positive treatment effect, the former is found when the drivers have accident(s) at fault for Car age ≤ 4 and the latter is done when the drivers are not at fault for women. Having accident(s) with BI have a negative effects when the car power is ≤ 100 or the age of driver is ≤ 25. For PD, a negative causal effect is found when the age of car is ≤ 4. Having an accidents with BI increases significantly the Urban outcome for women comparing with men. In all cases keeping the rest of the covariates at their reference values equal zero.

In general, after accident(s) the percent of distance at night does not reduce but in many cases increases. We obtain four negative significant effects after having accidents during the treatment period that are associated with the younger drivers or with less experience. Again, keeping the rest of the covariates equal to zero.

In Figure 2, 3 and 4, respectively, we plot the trends between post- and pre-treatment period, for control and treatment groups for three driver profiles. These profiles are selected taking into account that we have drivers with these profiles in control and alternative treatment groups that are plotted (No accidents, Accidents, Accidents¿1, At fault and BI). In all plots the trend associated with control group is the solid line and those associated with the treatment groups are represented with different dashed lines.

The trends plotted in Figure 2 are associated with men that are 25 or less years old, with a driven license with 5 or less years, that use parking at night, they do not drive in large metropolitan areas and have cars that are 4 or less years old, with power greater or equal to 100 horsepower. The Speed (top right) is the outcome where the treatment groups have a trend more different of control group, i.e. to have accident(s) cause a greater decrease in the percentage distances driven above the posted speed limits than if there had been no accident. On the contrary, for this same profile, we observe that accident(s) with BI cause a lower decrease in the anual distance driven in km (top left) that the rest of groups. Causal effects on Urban at Night for this profile are weaker, i.e. trends appear more parallel.

In Figures 3 and 4 we plots the trends for two women profiles. The former corresponds to women 25 years old or younger with a driven license with 5 or less years, that use parking at night, they drive in large metropolitan areas and have cars that are ≤ 4 years old, with power ≤ 100. The latter are similar but with “City”=0, i.e. the same women who do not drive in large metropolitan areas. We highlight some interesting results in both figures. For example, focusing on the treatment group with BI we observe that, when “City”=1 (Figure 3), this type of accident causes a decrease in the annual distance and the percentage distance driven above the posted speed limits (top left and right, respectively) compared with other groups, that can even have positive trends between pre- and post-treatment periods. However, when City=0 the treatment group with BI accidents has positive trends for the Distance and Speed outcomes.

Table 6: Average treatment effects conditional on a given group of drivers that are statistically significant at 5% and 10%.

	Distance		Speed		Urban		Night	
	*ATT*2	p-value	*ATT*2	p-value	*ATT*2	p-value	*ATT*2	p-value
Treatment	Model I							
Accidents	1800.9167	0.0001	-2.1943	0.0014			3.1725	0.0002
Age≤ 25×Accidents	774.6821	0.0134						
Woman×Accidents	1113.4460	0.0126			2.4324	0.0161		
Car age≤ 4×Accidents			-1.1773	0.0313	-2.0776	0.0118	6.9745	0.0000
	Model II							
Accidents= 1	2598.8106	0.0000	-1.3648	0.0169			2.0676	0.0437
Accidents> 1	-6212.5020	0.0002	-3.4448	0.0690			4.8410	0.0480
Age≤ 25× Accidents= 1					-3.8241	0.0029	-1.7776	0.0545
Age license≤ 5× Accidents= 1	1110.0370	0.0022						
Woman× Accidents= 1	1143.9520	0.0170					4.2103	0.0001
Car age≤ 4× Accidents= 1							4.7105	0.0001
Power≤ 100× Accidents= 1			1.3439	0.0038	2.2053	0.0587		
Age≤ 25×Accidents> 1							-2.4141	0.0435
Age license≤ 5×Accidents> 1	-2449.3760	0.0071						
Parking=yes×Accidents> 1	-4026.1460	0.0022	-6.1464	0.0000			7.5545	0.0000
Woman×Accidents> 1	-4418.8390	0.0070			9.9955	0.0027		
Power≤ 100×Accidents> 1	-3904.5680	0.0100						
	Model III							
At fault			-1.8128	0.0722				
Not at fault	2328.9677	0.0012			-1.8450	0.0960	4.1360	0.0003
Age≤ 25×At fault							-5.0394	0.0011
Age license≤ 5×At fault							5.6167	0.0000
Parking=yes×At fault			-4.1358	0.0000				
Woman×At fault							4.5788	0.0010
City×At fault	2547.3550	0.0069						
Car age≤ 4×At fault					-2.1360	0.0177		
Age≤ 25×Not at fault	857.4130	0.0555						
Age license≤ 5×Not at fault			-3.9393	0.0020			-3.5596	0.0216
Parking=yes×Not at fault							1.7443	0.0613
Woman×Not at fault					2.5710	0.0097		
Car age≤ 4×Not at fault			2.3166	0.0959			9.2730	0.0000
Power≤ 100×Not at fault	3335.6350	0.0000	2.9702	0.0008				
	Model IV							
BI	-5376.1413	0.0001	-5.7546	0.0004				
PD	2745.2758	0.0000					2.1741	0.0203
Age≤ 25×BI			-1.5722	0.0803				
Age license ≤ 5×BI							-5.7154	0.0200
Parking=yes×BI	-1861.3750	0.0420						
Woman×BI					12.2256	0.0000		
City×BI	-10110.2700	0.0000	-11.0628	0.0002				
Car age≤ 4×BI							8.2204	0.0115
Power≤ 100×BI					-7.7852	0.0005		
Age≤ 25×PD	991.1634	0.0055			-4.2491	0.0133	-2.4996	0.0058
Age license≤ 5×PD			-2.9793	0.0048				
Woman×PD	1516.8160	0.0021					3.5200	0.0002
Car age≤ 4×PD			1.8464	0.0998	-5.7907	0.0147	4.7800	0.0000
Power≤ 100×PD			1.1580	0.0658				

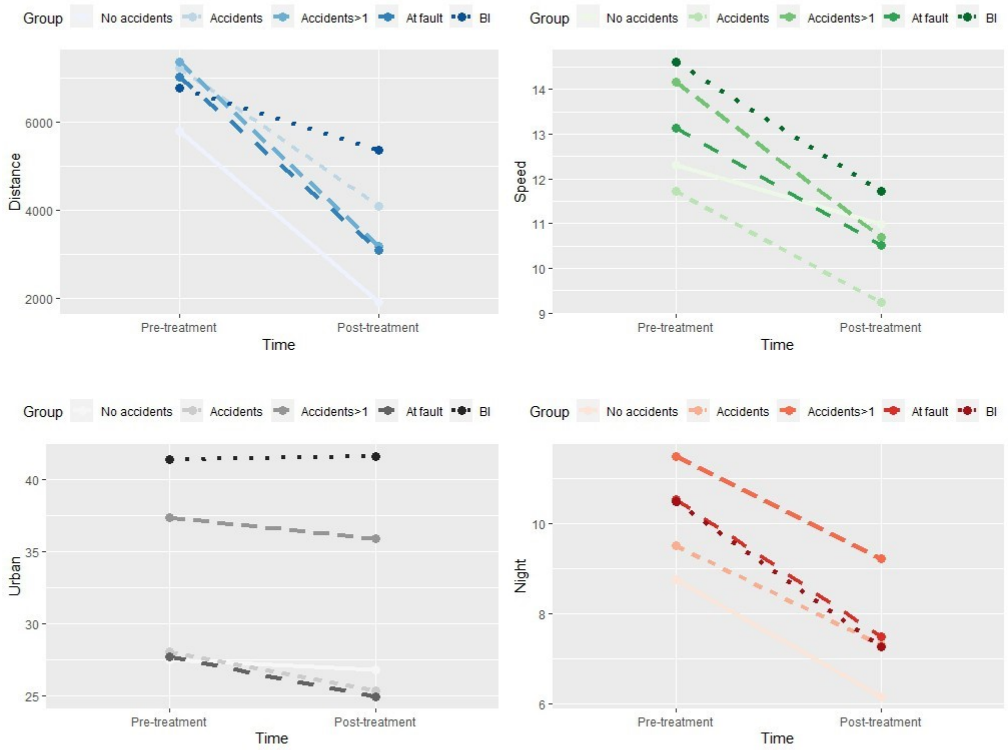

Figure 2: Estimated trends without accidents (control group) and with accidents (treatment groups) for a driver with the following values of covariates: “Age≤25”=1, “Age license≤5”=1, “Parking=yes”=1, “Woman”=0, “City”=0, “Car age≤4”=1 and “Power≤100”=0.

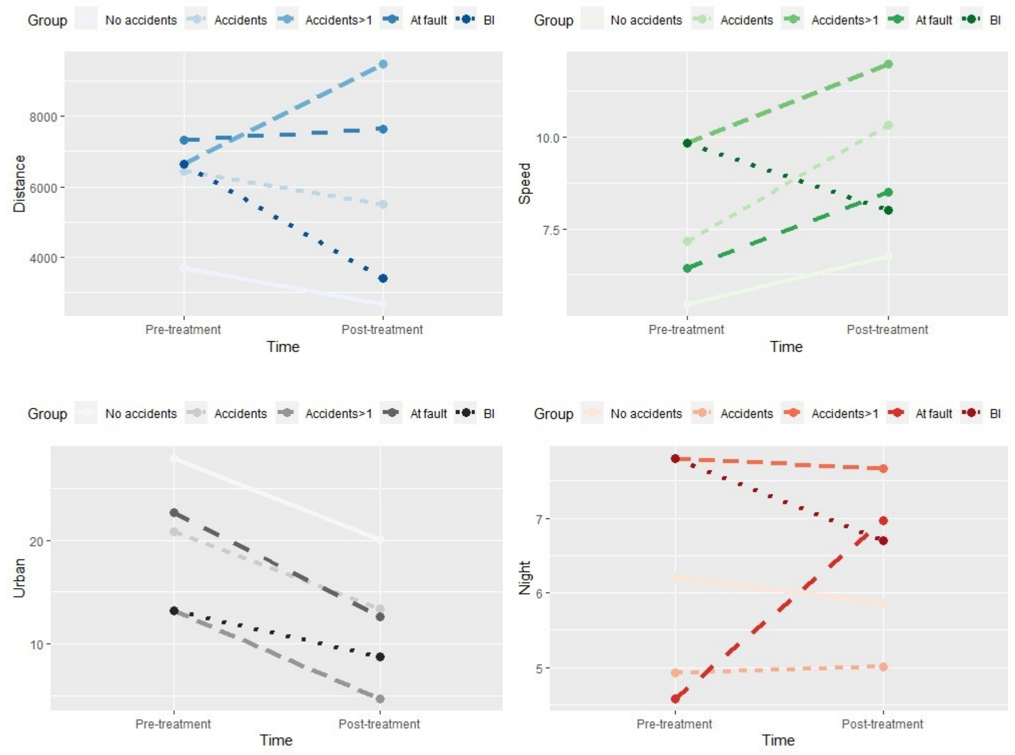

Figure 3: Estimated trends without accidents (control group) and with accidents (treatment groups) for a driver with the following values of covariates: “Age ≤25”=1, “Age license≤5”=1, “Parking=yes”=1, “Woman”=1, “City”=1, “Car age≤4”=1 and “Power≤100”=1.

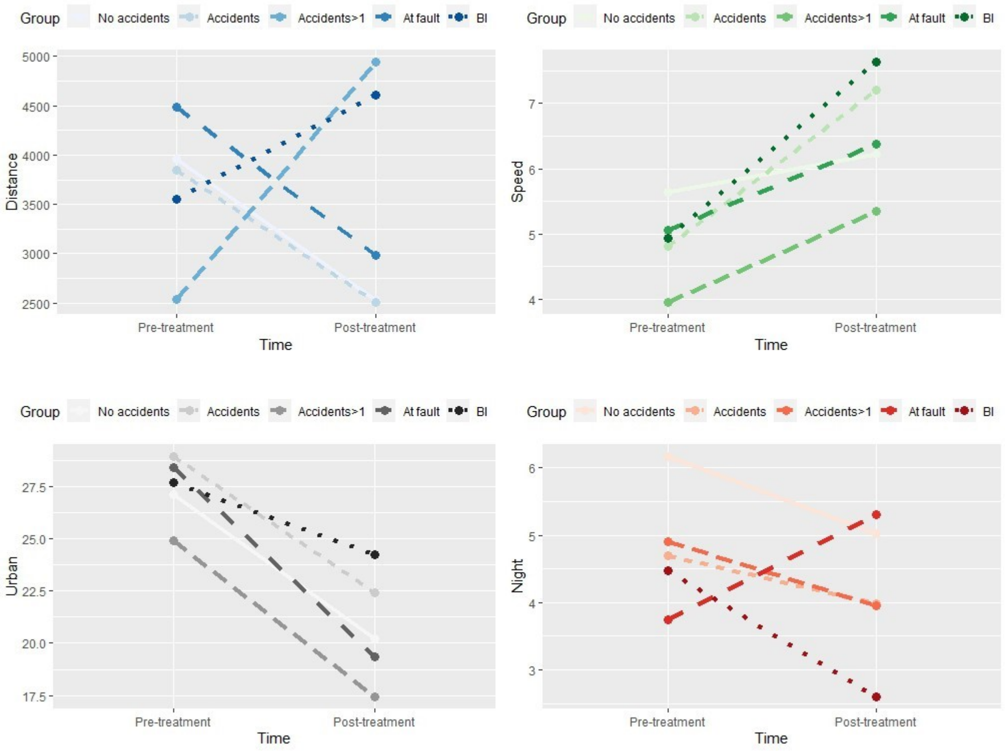

Figure 4: Estimated trends without accidents (control group) and with accidents (treatment groups) for a driver with the following values of covariates: "Age ≤25"=1, "Age license≤5"=1, "Parking=yes"=1, "Woman"=1, "City"=0, "Car age≤4"=1 and "Power≤100"=1.

5 Conclusions

We have used the DiD method to estimate the causal effects of accident(s) on a set of four outcomes: the annual distance in km, the percentage distances driven above the posted speed limits, the percentage of km in urban areas and the percentage of km at night. These variables measure the behavior and driving habits of the drivers. This data was collected by sensors that were used by an ”insurtech” company. We have found that these causal effects are not homogeneous, i.e., they change in function of the profiles of policyholders.

The basic DiD model has been generalized in three ways, to take into account the frequency (if it is equal 1 or is greater than 1), the severity (if there was any bodily injury or if there was only property damage) and if the driver was at fault or not of the accident(s), i.e. different types of treatments. The estimated results for the different expanded models show how the causal effects are different depending on the type of treatment.

We have defined disjointed pairs of treatment variables but, in fact, we can include different treatment variables that are not disjointed. The difficulty is that we need to have sufficient information to identify the different treatment effects given by the different treatments under control (for example, more than 1 accident, BI and at fault) conditional on the values of covariates.

We have identified different driver profiles with different causal effects of treatment. These results could allow insurance companies to better adjust premiums to insured profiles.

We have estimated positive and negative significant treatment effects. Comparing the four outcomes, that with the least significant effects is the percentage of km in urban areas. This outcome is the least sensitive to the occurrence of accident(s).

The larger negative causal effects are obtained when the percentage distances driven above the posted speed limits is analyzed and, furthermore, the most negative effect is associated to accident(s) with bodily injury.

About the total annual km, I have observed that on average this outcome decreases in more than 1500 km between pre- and post-treatment periods, this could be justified by the economic crisis that began in Spain in 2008, however, we found some positive causal effects of accidents that are basically associated to frequency = 1, accidents where the driver was not at fault (or at fault in city) and/or with only property damage. These results could be associated with drivers whose need to take the car increased.

An issue associated with the data set is that we have to work with calendar years instead of with the annuity associated with each policy that takes into account the renewal date. Furthermore, we do not know the exact date on which the accident occurs. The causal effects in the post-treatment period could be reduced as the accidents occur closer to the beginning of the treatment period, due to a forgetting effect. This could cause some estimated ATT_2 to not be statistically significant.

Finally, we note that using additive models can lead to negative predictions, as in our case it occurs for some profiles of drivers in the annual distance outcome. In this case, estimating the multiplicative model using the log-transformation of the dependent variable is a easy solution.

A Homogeneous models

Table A1: Results of the withing estimations of the TWFE models with additive covariates.

Model I	Distance		Speed		Urban		Night	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
Accidents	765.5157	0.0009	-0.0615	0.8571	-0.1409	0.8083	0.8394	0.0397
Age≤<= 25*c*¸	-462.2032	0.0033	0.5306	0.0173			-2.2966	0.0000
Age license≤ 5	-398.4944	0.0891			-0.8191	0.0183	1.1080	0.0077
Parking=yes	-823.3772	0.0000					-0.3996	0.0964
Woman	203.0456	0.1135					0.6968	0.0030
City	396.0378	0.0292	0.7574	0.0050	-0.8614	0.0600	0.8133	0.0117
Car age≤ 4	318.1814	0.1335	0.9102	0.0000			-0.5728	0.1291
Power≤ 100			1.0895	0.0000			-0.5006	0.0375
Panel fixed effects test	1.2144	0.0000	3.0648	0.0000	4.1653	0.0000	1.3353	0.0000
R^2	0.7810		0.8725		0.9506		0.7707	
Adjusted R^2	0.5608		0.7444		0.9011		0.5401	
Model II	Distance		Speed		Urban		Night	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
Accidents=1	885.7809	0.0005	0.1585	0.6755	-0.0989	0.8781	0.8457	0.0616
Accidents¿1	288.2723	0.5594	-0.9332	0.2034	-0.3074	0.8057	0.8144	0.3529
Age≤ 25	-461.5125	0.0033	0.5327	0.0169			-2.2966	0.0000
Age license≤ 5	-396.4131	0.0907			-0.8173	0.0186	1.1082	0.0077
Parking=yes	-823.6921	0.0000					-0.3996	0.0964
Woman	200.5855	0.1180					0.6967	0.0030
City	395.7338	0.0293	0.7564	0.0050	-0.8615	0.0600	0.8132	0.0118
Car age≤ 4	321.8557	0.1291	0.9190	0.0000			-0.5726	0.1293
Power≤ 100			1.0880	0.0000			-0.5006	0.0375
Panel fixed effects test	1.2138	0.0000	3.0646	0.0000	4.1638	0.0000	1.3347	0.0000
R^2	0.7810		0.8725		0.9506		0.7707	
Adjusted R^2	0.5608		0.7444		0.9010		0.5399	
Model III	Distance		Speed		Urban		Night	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
At fault	433.7275	0.1811	-0.5254	0.2752	-0.7209	0.3795	1.4163	0.0139
Not at fault	1071.4412	0.0006	0.3674	0.4282	0.3941	0.6176	0.3072	0.5792
Age≤ 25	-467.4646	0.0029	0.5243	0.0187			-2.2875	0.0000
Age license≤ 5	-395.1734	0.0917			-0.8122	0.0193	1.1024	0.0081
Parking=yes	-824.2571	0.0000					-0.3985	0.0973
Woman	196.7375	0.1254					0.7087	0.0026
City	397.0321	0.0288	0.7582	0.0049	-0.8598	0.0605	0.8113	0.0120
Car age≤ 4	323.4534	0.1272	0.9192	0.0000			-0.5815	0.1235
Power≤ 100			1.0905	0.0000			-0.5045	0.0361
Panel fixed effects test	1.2144	0.0000	3.0648	0.0000	4.1653	0.0000	1.3353	0.0000
R^2	0.7810		0.8725		0.9507		0.7708	
Adjusted R^2	0.5608		0.7444		0.9011		0.5401	
Model IV	Distance		Speed		Urban		Night	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
BI	-43.4732	0.9315	-1.5716	0.0366	0.4774	0.7094	-0.3883	0.6655
PD	957.3485	0.0002	0.2965	0.4309	-0.2877	0.6537	1.1305	0.0120
Age≤ 25	-460.9415	0.0034	0.5342	0.0166			-2.2947	0.0000
Age license≤ 5	-396.1436	0.0909			-0.8242	0.0176	1.1116	0.0075
Parking=yes	-819.0927	0.0000					-0.3930	0.1020
Woman	197.7035	0.1234					0.6886	0.0034
City	388.1634	0.0326	0.7413	0.0060	-0.8551	0.0620	0.8013	0.0131
Car age≤ 4	322.7042	0.1280	0.9216	0.0000			-0.5660	0.1338
Power≤ 100			1.0865	0.0000			-0.5003	0.0376
Panel fixed effects test	1.2141	0.0000	3.0650	0.0000	4.1636	0.0000	1.3353	0.0000
R^2	0.7811		0.8725		0.9506		0.7708	
Adjusted R^2	0.5609		0.7445		0.9011		0.5401	

B Heterogeneous models

Table A2: Results for Model I of the withing estimations of the TWFE models with additive and multiplicative covariates.

	Distance		Speed		Urban		Night	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
Accidents	1800.9167	0.0001	-2.1943	0.0014	-0.0016	0.9988	3.1725	0.0002
Age≤ 25	-380.5376	0.0186	0.5383	0.0158			-1.9703	0.0000
Age license≤ 5	-396.9077	0.0904			-1.3965	0.0136	1.3678	0.0016
Parking=yes	-824.1886	0.0000					-0.4073	0.0897
Woman	261.0746	0.0513					0.7013	0.0028
City	399.3320	0.0278	0.7445	0.0057	-0.9408	0.0404	0.8089	0.0121
Car age≤ 4	305.1440	0.1502	0.8338	0.0003	0.9216	0.0908	-0.9575	0.0153
Power≤ 100			0.9449	0.0000	-0.5016	0.1305	-0.4992	0.0378
Age≤ 25×Accidents	-1026.2346	0.0412					-3.8708	0.0002
Age license≤ 5× Accidents							-2.6787	0.0834
Woman×Accidents	-687.4710	0.1359	1.6542	0.0149	2.4340	0.0301		
City×Accidents								
Car age≤ 4×Accidents			1.0170	0.1466	-2.0760	0.0819	3.8020	0.0040
Power≤ 100×Accidents			1.3526	0.0568				
Panel fixed effects test	1.1971	0.0000	3.0706	0.0000	4.0814	0.0000	1.3403	0.0000
R^2	0.7812		0.8727		0.9507		0.7715	
Adjusted R^2	0.5610		0.7447		0.9012		0.5413	

Table A3: Results for Model II of the withing estimations of the TWFE models with additive and multiplicative covariates.

	Distance		Speed		Urban		Night	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
Accidents=1	2598.8106	0.0000	-1.3648	0.0169	-0.3899	0.7980	2.0676	0.0437
Accidents¿1	-6212.5020	0.0002	-3.4448	0.0690	3.6726	0.2276	4.8410	0.0480
Age≤ 25	-442.0150	0.0049	0.5419	0.0151			-1.9379	0.0000
Age license≤ 5	-345.6365	0.1455			-1.3102	0.0211	1.1985	0.0039
Parking=yes	-836.1211	0.0000					-0.3569	0.1554
Woman	255.0043	0.0567					0.5900	0.0148
City	415.4509	0.0220	0.7403	0.0060	-1.0073	0.0283	0.8244	0.0105
Car age≤ 4	322.7443	0.1274	0.8669	0.0001	0.8100	0.1312	-0.8397	0.0290
Power≤ 100			0.9166	0.0000	-0.5349	0.1217	-0.5360	0.0258
Age≤ 25×Accidents=1					-3.4342	0.0120	-3.8453	0.0004
Age license≤ 5×Accidents=1	-1488.7732	0.0071						
Parking=yes×Accidents=1					1.9992	0.1192	-1.9017	0.0416
Woman×Accidents=1	-1454.8582	0.0043					2.1427	0.0182
City×Accidents=1								
Car age≤ 4×Accidents=1							2.6428	0.0096
Power≤ 100×Accidents=1			2.7087	0.0004	2.5952	0.0469		
Age≤ 25×Accidents¿1							-7.2551	0.0010
Age license≤ 5× Accidents¿1	3763.1262	0.0066						
Parking=yes×Accidents¿1	2186.3561	0.0405	-2.7016	0.0733			2.7135	0.1427
Woman×Accidents¿1	1793.6631	0.0858	3.0004	0.0444	6.3229	0.0142		
City×Accidents¿1								
Car age≤ 4×Accidents¿1			3.9602	0.0178	-4.4166	0.1243		
Power≤ 100×Accidents¿1	2307.9340	0.0250			-6.0810	0.0196		
Panel fixed effects test	1.2040	0.0000	3.0764	0.0000	4.0828	0.0000	1.3424	0.0000
R^2	0.7821		0.8730		0.9508		0.7720	
Adjusted R^2	0.5622		0.7451		0.9012		0.5417	

Table A4: Results for Model III of the withing estimations of the TWFE models with additive and multiplicative covariates.

	Distance		Speed		Urban		Night	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
At fault	1075.4656	0.1048	-1.8128	0.0722	1.8502	0.1829	1.3552	0.2903
Not at fault	2328.9677	0.0012	-0.1667	0.8625	-1.8450	0.0960	4.1360	0.0003
Age≤ 25	-415.2950	0.0091	0.5709	0.0106			-2.0458	0.0000
Age license≤ 5	-350.4346	0.1369			-1.3916	0.0139	1.3349	0.0018
Parking=yes	-821.2686	0.0000						
Woman	237.2905	0.0706					0.5981	0.0122
City	334.8529	0.0708	0.7451	0.0056	-0.9338	0.0418	0.8114	0.0116
Car age≤ 4	314.0550	0.1385	0.8195	0.0003	0.9102	0.0914	-0.9227	0.0166
Power≤ 100			0.9372	0.0000	-0.5207	0.1162	-0.4567	0.0552
Age≤ 25× At fault							-6.3946	0.0000
Age license≤ 5×At fault	-1142.5612	0.1299					4.2615	0.0059
Parking=yes×At fault			-2.3230	0.0171				
Woman×At fault			2.0463	0.0346			3.2236	0.0063
City×At fault	1471.8894	0.1033						
Car age≤ 4×At fault			2.8213	0.0061	-3.9863	0.0205		
Power≤ 100×At fault								
Age≤ 25×Not at fault	-1471.5547	0.0335						
Age license≤ 5×Not at fault			-3.7726	0.0144			-7.6956	0.0000
Parking=yes×Not at fault							-2.3918	0.0307
Woman×Not at fault	-1501.5149	0.0224			4.4160	0.0043		
City×Not at fault								
Car age≤ 4×Not at fault			2.4833	0.0879			5.1370	0.0033
Power≤ 100×Not at fault	1006.6672	0.1196	3.1369	0.0008				
Panel fixed effects test	1.1968	0.0000	3.0799	0.0000	4.0840	0.0000	1.3437	0.0000
R^2	0.7815		0.8731		0.9508		0.7722	
Adjusted R^2	0.5611		0.7452		0.9012		0.5423	

Table A5: Results for Model IV of the withing estimations of the TWFE models with additive and multiplicative covariates.

	Distance		Speed		Urban		Night	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
BI	-5376.1413	0.0001	-5.7546	0.0004	0.9331	0.6344	3.1062	0.1299
PD	2745.2758	0.0000	-0.7618	0.3393	-1.3551	0.3967	2.1741	0.0203
Age≤ 25	-397.0468	0.0139	0.5227	0.0204			-2.0089	0.0000
Age license≤ 5	-366.3320	0.1175			-1.5369	0.0087	1.3090	0.0018
Parking=yes	-853.6193	0.0000					-0.4085	0.0889
Woman	263.8871	0.0466					0.5969	0.0138
City	422.8072	0.0203	0.7875	0.0036	-0.9380	0.0409	0.7676	0.0173
Car age≤ 4	310.7016	0.1421	0.8631	0.0001	1.0127	0.0676	-0.8793	0.0235
Power≤ 100			0.9218	0.0000	-0.5375	0.1198	-0.4963	0.0389
Age≤ 25×BI	4351.6117	0.0005	4.1824	0.0234				
Age license≤ 5×BI							-8.8215	0.0107
Parking=yes×BI	3514.7660	0.0014						
Woman×BI			3.6934	0.0176	11.2925	0.0001		
City×BI	-4734.1244	0.0096	-5.3082	0.0493				
Car age≤ 4×BI							5.1142	0.0861
Power≤ 100× BI					-8.7184	0.0020		
Age≤ 25×PD	-1754.1125	0.0013			-2.8940	0.0741	-4.6737	0.0000
Age license≤ 5×PD			-2.2176	0.0805	3.8610	0.1125		
Parking=yes×PD					2.4921	0.0505		
Woman×PD	-1228.4595	0.0152					1.3459	0.1350
City×PD								
Car age≤ 4×PD			2.6082	0.0289	-4.4356	0.0324	2.6059	0.0099
Power≤ 100×PD			1.9198	0.0114	2.6914	0.0390		
Panel fixed effects test	1.2048	0.0000	3.0733	0.0000	4.0892	0.0000	1.3397	0.0000
R^2	0.7823		0.8730		0.9509		0.7717	
Adjusted R^2	0.5626		0.7450		0.9014		0.5413	

References

Abadie, A., 2005. Semiparametric difference-in-differences estimators. The Review of Economic Studies 72, 1–19. URL: http://www.jstor.org/stable/3700681.

Athey, S., Imbens, G.W., 2006. Identification and inference in nonlinear difference-in-differences models. Econometrica 74, 431–497. URL: http://www.jstor.org/stable/3598807.

Ayuso, M., Guillen, M., Nielsen, J.P., 2019. Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data. Transportation 46, 735–752.

Ayuso, M., Guillén, M., Pérez-Marín, A.M., 2014. Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. Accident Analysis & Prevention 73, 125–131.

Ayuso, M., Guillen, M., Pérez-Marín, A.M., 2016. Telematics and gender discrimination: Some usage-based evidence on whether men’s risk of accidents differs from women’s. Risks 4, 10.

Baecke, P., Bocca, L., 2017. The value of vehicle telematics data in insurance risk selection processes. Decision Support Systems 98, 69–79.

Bolancé, C., Guillen, M., Pitarque, A., 2020. A sarmanov distribution with beta marginals: An application to motor insurance pricing. Mathematics 8, 2020.

Boucher, J.P., Côté, S., Guillen, M., 2017. Exposure as duration and distance in telematics motor insurance using generalized additive models. Risks 5, 54.

Boucher, J.P., Denuit, M., Guillén, M., 2007. Risk classification for claim counts: a comparative analysis of various zero-inflated mixed poisson and hurdle models. North American Actuarial Journal 11, 110–131.

Boucher, J.P., Denuit, M., Guillen, M., 2009. Number of accidents or number of claims? an approach with zero-inflated poisson models for panel data. Journal of Risk and Insurance 76, 821–846.

Boucher, J.P., Turcotte, R., 2020. A longitudinal analysis of the impact of distance driven on the probability of car accidents. Risks 8, 91.

Boylan, J., Meyer, D., Chen, W.S., 2024. A systematic review of the use of in-vehicle telematics in monitoring driving behaviours. Accident Analysis & Prevention 199, 107519.

Chan, J.S., Choy, S.B., Makov, U., Shamir, A., Shapovalov, V., 2022. Variable Selection Algorithm for a Mixture of Poisson Regression for Handling Overdispersion in Claims Frequency Modeling Using Telematics Car Driving Data. Risks 10, 83.

Chauhan, V., Yadav, J., 2024. Bibliometric review of telematics-based automobile insurance: Mapping the landscape of research and knowledge. Accident Analysis & Prevention 196, 107428.

Cheng, J., Feng, F.Y., Zeng, X., 2023. Pay-as-you-drive insurance: modeling and implications. North American Actuarial Journal 27, 303–321.

Denuit, M., Guillen, M., Trufin, J., 2019. Multivariate credibility modelling for usage-based motor insurance pricing with behavioural data. Annals of Actuarial Science 13, 378–399.

Di Tella, R., Schargrodsky, E., 2004. Do police reduce crime? estimates using the allocation of police forces after a terrorist attack. American Economic Review 94, 115–133. URL: https://www.aeaweb.org/articles?id=10.1257/000282804322970733, doi:10.1257/000282804322970733.

Duval, F., Boucher, J.P., Pigeon, M., 2022. How Much Telematics Information Do Insurers Need for Claim Classification? North American Actuarial Journal 26, 570–590.

Duval, F., Boucher, J.P., Pigeon, M., 2024. Telematics combined actuarial neural networks for cross-sectional and longitudinal claim count data. ASTIN Bulletin: The Journal of the IAA, 1–24.

Eling, M., Kraft, M., 2020. The impact of telematics on the insurability of risks. The Journal of Risk Finance 21, 77–109.

Ferreira Jr, J., Minikel, E., 2012. Measuring per mile risk for pay-as-you-drive automobile insurance. Transportation research record 2297, 97–103.

Galiani, S., Gertler, P., Schargrodsky, E., 2005. Water for life: The impact of the privatization of water services on child mortality. Journal of Political Economy 113, 83–120. URL: https://doi.org/10.1086/426041, doi:10.1086/426041.

Gao, G., Wang, H., Wüthrich, M.V., 2022. Boosting Poisson regression models with telematics car driving data. Machine Learning , 1–30.

Gao, Y., Huang, Y., Meng, S., 2023. Evaluation and interpretation of driving risks: Automobile claim frequency modeling with telematics data. Statistical Analysis and Data Mining: The ASA Data Science Journal 16, 97–119.

Guillen, M., Bermúdez, L., Pitarque, A., 2021a. Joint generalized quantile and conditional tail expectation regression for insurance risk analysis. Insurance: Mathematics and Economics 99, 1–8.

Guillen, M., Nielsen, J.P., Ayuso, M., Pérez-Marín, A.M., 2019. The use of telematics devices to improve automobile insurance rates. Risk analysis 39, 662–672.

Guillen, M., Nielsen, J.P., Pérez-Marín, A.M., 2021b. Near-miss telematics in motor insurance. Journal of Risk and Insurance 88, 569–589.

Guillen, M., Nielsen, J.P., Pérez-Marín, A.M., Elpidorou, V., 2020. Can automobile insurance telematics predict the risk of near-miss events? North American Actuarial Journal 24, 141–152.

Guillen, M., Pérez-Marín, A.M., Alcañiz, M., 2021c. Percentile charts for speeding based on telematics information. Accident Analysis & Prevention 150, 105865.

Henckaerts, R., Antonio, K., 2022. The added value of dynamically updating motor insurance prices with telematics collected driving behavior data. Insurance: Mathematics and Economics 105, 79–95.

Huang, Y., Meng, S., 2019. Automobile insurance classification ratemaking based on telematics driving data. Decision Support Systems 127, 113156.

Jeffrey, S.H., Christy, H.L., Mona, A.A., Allyson, G.H., Robert, P.D., 2011. Changes in per member per month expenditures after implementation of florida's medicaid reform demonstration. Health Services Research 46, 787–804. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3097402/, doi:10.1111/j.1475-6773.2010.01226.x.

Jeong, H., 2022. Dimension reduction techniques for summarized telematics data. The Journal of Risk Management, Forthcoming .

Kirushanth, S., Kabaso, B., 2018. Telematics and road safety, in: 2018 2nd International Conference on Telematics and Future Generation Networks (TAFGEN), IEEE. pp. 103–108.

Lemaire, J., Park, S.C., Wang, K.C., 2016. The use of annual mileage as a rating variable. ASTIN Bulletin: The Journal of the IAA 46, 39–69.

Litman, T., 2007. Distance-based vehicle insurance feasibility, costs and benefits. Victoria 11. URL: https://vtpi.org/dbvi_com.pdf.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis & Prevention 37, 35–46.

Ma, Y.L., Zhu, X., Hu, X., Chiu, Y.C., 2018. The use of context-sensitive insurance telematics data in auto insurance rate making. Transportation Research Part A: Policy and Practice 113, 243–258.

Masello, L., Castignani, G., Sheehan, B., Guillen, M., Murphy, F., 2023. Using contextual data to predict risky driving events: A novel methodology from explainable artificial intelligence. Accident Analysis & Prevention 184, 106997.

Meng, S., Wang, H., Shi, Y., Gao, G., 2022. Improving automobile insurance claims frequency prediction with telematics car driving data. ASTIN Bulletin: The Journal of the IAA 52, 363–391.

Pérez-Marín, A.M., Ayuso, M., Guillen, M., 2019. Do young insured drivers slow down after suffering an accident? Transportation research part F: traffic psychology and behaviour 62, 690–699. doi:https://doi.org/10.1016/j.trf.2019.02.021.

Pesantez-Narvaez, J., Guillen, M., Alcañiz, M., 2019. Predicting motor insurance claims using telematics data—xgboost versus logistic regression. Risks 7, 70.

Pitarque, A., Guillen, M., 2022. Interpolation of quantile regression to estimate driver's risk of traffic accident based on excess speed. Risks 10, 19.

Reig Torra, J., Guillen, M., Pérez-Marín, A.M., Rey Gámez, L., Aguer, G., 2023. Weather conditions and telematics panel data in monthly motor insurance claim frequency models. Risks 11, 57.

Roth, J., Sant'Anna, P.H., Bilinski, A., Poe, J., 2023. What's trending in difference-in-differences? a synthesis of the recent econometrics literature. Journal of Econometrics 235, 2218–2244. URL: https://www.sciencedirect.com/science/article/pii/S0304407623001318, doi:https://doi.org/10.1016/j.jeconom.2023.03.008.

Sun, S., Bi, J., Guillen, M., Pérez-Mar´ın, A.M., 2021. Driving risk assessment using near-miss events based on panel poisson regression and panel negative binomial regression. Entropy 23, 829.

Verbelen, R., Antonio, K., Claeskens, G., 2018. Unravelling the predictive power of telematics data in car insurance pricing. Journal of the Royal Statistical Society Series C: Applied Statistics 67, 1275–1304.

Winlaw, M., Steiner, S.H., MacKay, R.J., Hilal, A.R., 2019. Using telematics data to find risky driver behaviour. Accident Analysis & Prevention 131, 131–136.

Wüthrich, M.V., Merz, M., 2019. Yes, we can! ASTIN Bulletin: The Journal of the IAA 49, 1–3.

UBIREA
Institut de Recerca en Economia Aplicada Regional i Pública
Research Institute of Applied Economics
Universitat de Barcelona
Av. Diagonal, 690 • 08034 Barcelona

WEBSITE: www.ub.edu/irea/ • **CONTACT:** irea@ub.edu