Estimating the impact of traffic regulation on accident numbers and gravity

Project Report - Inferential Statistics

Jonathan Janke - M2016053

Vladimirs Racejevs - M2016048

Steffen Vering - M2016033

NOVA IMS
Master in Advanced Analytics
07.07.2017

Abstract

Changing traffic regulation oftentimes requires drivers to adapt their driving behavior or speed. In the 1990s the State of California, US has introduced two regulations for drivers. One required them to wear a seat belt and another one increased the speed limit on interstates and other rural highways from 55 to 65 mph. Using a time-series regression model the impact of each of each of these legislation's was estimated. The created regression model accounts for seasonality, serial correlation as well as the time trend and corrects heteroscedasticity. The introduction of the seat belt law did reduce the percentage of fatal accidents on average by 0.245 percentage points controlling for all other factors. This was can be considered a significant decrease with a 1% significance level. On the other hand the group did not find evidence for a significant impact of the increase of the speed limit on the total number of accidents on those rural roads, which were included in the increase.

Contents

Chapter 1

Introduction

The aim of this project is to apply the methods discussed in the Inferential Statistics class at NOVA IMS on a new dataset. This document is supposed to give the reader an overview over the problem that the group chose to investigate, the methodology chosen to do so (chapter 1), the solutions that were found (chapter 2), as well as a short summary and outlook (chapter 3). The problem that is being investigated revolves around the number of accidents in California, US in relation to the introduction of two separate regulations that might or might not have impacted the number and gravity of these accidents. In this introductory chapter the underlying data will be discussed, research goals will be established and the methodology that was used will be outlined.

1.1 Data Description

The data being used in this project was retrieved from a public repository[1] that supplies supplemental information for the Woolridge textbook[2], that was also used in class.

For this project the group decided to use the *traffic2* dataset, that contains monthly data on the number and type of accidents in California between 1981 and 1989. The dataset contains 108 observations. There is no missing data and according to the textbook the mode of measurement is constant and therefore this dataset is viable to be treated as a time-series.

Table 1.1 shows the summary statistics for all non-transformed variables and non-dummy variables in the data. The table 4.1 also contains a description of all variables and their meaning and can be found in the Appendix. There are 16 variables that could be used as the dependent variable in a regression model. These specify the number of total, fatal, injury and property damage only accidents by road type (i.e. non-interstate, highways, county roads, state roads and rural 65 mph roads). It can be observed that there are on average 42,831 accidents overall, of which 17,861 leave one or more participants injured and 378, which

[1] http://www.cengage.com/

[2] Woolridge, Introductory Econometrics: A Modern Approach, 5th Edition

include at least one fatality. It can also be seen that in about 29.6 percent of the observations the increased speed limit was applicable and for 44.4 percent of the observations the belt requirement regulation was in place. Besides these two laws, the unemployment rate and the number of weekend days per month (these include Fridays) are the other potential independent variables that are present in the dataset.

Table 1.1: Data Summary

Statistic	Mean	St. Dev.	Min	Max
totacc	42,831.260	4,608.328	32,699	52,971
fatacc	377.935	48.547	266	500
injacc	17,861.480	1,963.107	13,268	21,741
pdoacc	24,591.840	2,773.219	19,162	31,425
ntotacc	39,522.970	3,797.032	30,759	47,874
nfatacc	335.306	41.130	237	434
ninjacc	16,578.780	1,695.411	12,492	19,963
npdoacc	22,608.890	2,252.886	18,030	28,338
rtotacc	324.824	82.207	181	555
rfatacc	13.250	5.372	3	31
rinjacc	146.213	41.333	71	261
rpdoacc	165.361	39.450	90	267
ushigh	1,200.546	139.759	914	1,523
cntyrds	6,641.472	505.426	5,334	7,805
strtes	3,640.602	313.766	2,826	4,447
unem	7.201	1.790	4	12
spdlaw	0.296	0.459	0	1
beltlaw	0.444	0.499	0	1
wkends	13.074	1.011	12	15

In order to get a better understanding of the data, the project group decided to calculate the correlations between the variables in the data-set. The corresponding correlation table was not included in this report because of it's questionable value and its large size. Instead the value of the correlation between the independent variables and potentially relevant dependant variables is included. Figure 1.1 shows this correlation plot. It can be observed, that there are relatively high absolute correlation values for most of the potential explanatory variables. Whereas the speedlaw and the beltlaw variables have a positive correlation that is higher than 0.5 for all dependent variables, but the number of fatal accidents, the unemployment rate has a negative correlation of up to -0.8 with the dependent variables. The number of weekend days in contrast, does not show such a strong correlation and even though there is a slight positive correlation, this never

exceeds 0.25. It might make sense to further transform some of these variables

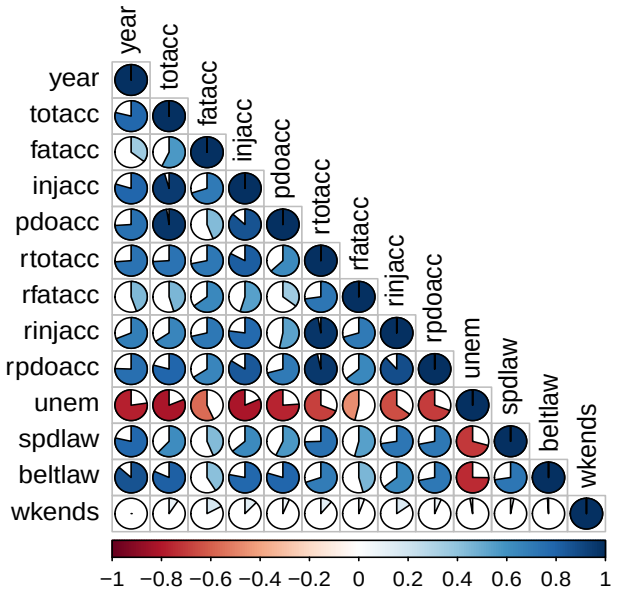

Figure 1.1: Correlation of number of all accidents with dependent variables

for the creation of the model or to include transformed dependent or independent variables into the model, but these steps will be included in chapter 2. For now it should be noted that the dataset that was retrieved from the Woolridge repository includes the percentage values of the type of accidents, the log-transformed numbers of accidents as well as multiplication of the speedlaw/beltlaw dummies with the current time. It also includes the squared time variable of the time trend variable. All variable names can be found in the appendix.

Before defining the research goals for this project, a visual analysis of the impact of the introduction of the belt and the speed regulation were conducted. Therefore the group has plotted the number of accidents by type over time with different colours referring to the introduction of regulations. 0 represents the time before the introduction of the regulations, 1 represents the time, where only the belt requirement was in place and 2 represents the time, when the increased speed limit as well as the belt requirement were in place at the same time. The group noticed that there was no clear visual distinction that could be made between the plots for all accidents, accidents involving injuries and accidents only damaging property. In contrast, the number of fatal accidents had a very different curve, that is shown in 1.2 next to the plot for the number of total accidents. While it seems that there is an increasing number of accidents over time, this trend cannot be observed as clearly for the number of fatal accidents. In fact the number of fatal accidents over times curve visually suggests that this value might even be stationary, but this would need to be tested in the following chapters.

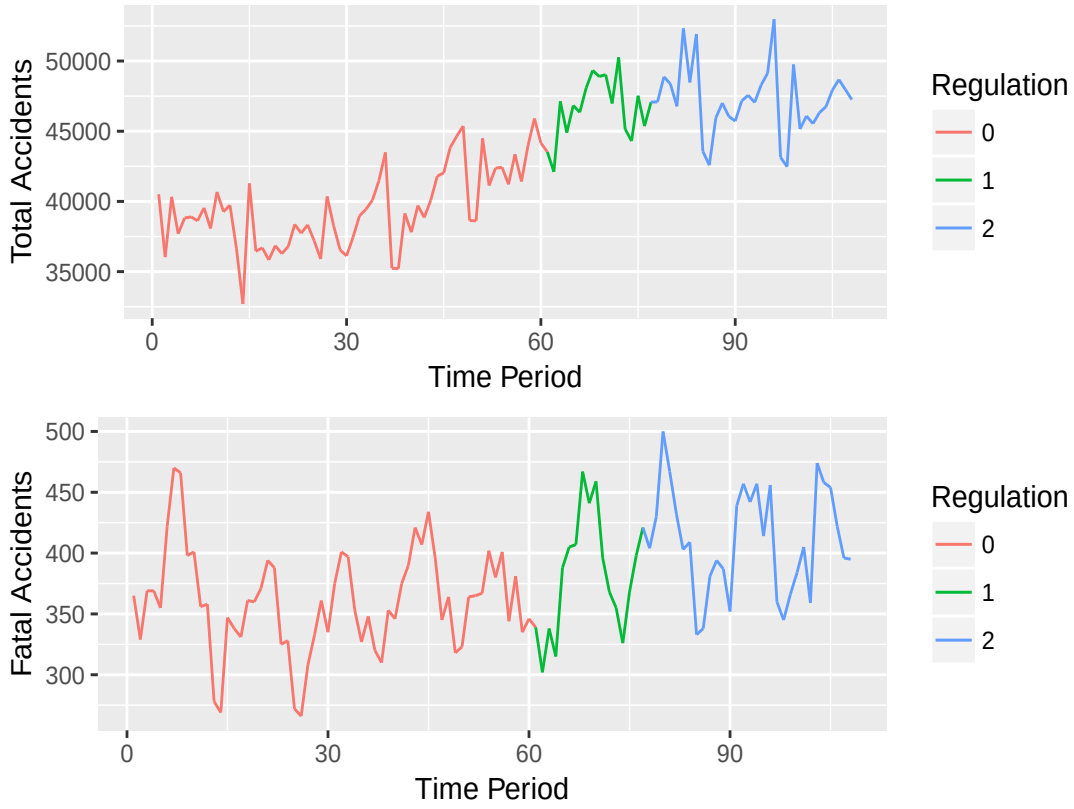

Figure 1.2: Targets over time

Overall the traffic dataset serves as an optimal starting point for this project, as it contains enough variables, no missing data and explanatory variables that are correlated with the target. This allows the progress to a more distant view of the data and investigate, which research questions might be appropriate and valuable to answer.

1.2 Research goals

While it was already determined that the data is appropriate for an econometric analysis, more importantly the real-world representation of the data and the potential inference that is contained needs to be interesting enough for the group to analyze. To come up with an interesting research question the group decided to separately consider independent and dependent variables.

For the independent variables (*t, tsq, unem, spdlaw, bltlaw, wkends*), potential research questions could lead us to ask, whether "The number of accidents(or others) has a significant trend?" or whether "The number of weekend days in a month significantly impacts the number of fatalities in accidents?". The result of this question could then lead to further research questions or could be leveraged in further accident reduction efforts. For this project, the group focused on the analysis of the increase in the speed limit in 1987/1988 and the introduction of a belt requirement in 1986. A recent study on occupant protection in passenger vehicles found, that about half of all passengers, who died in car accidents in the US in 2015, were not secured by a seat-belt.[3] This could lead to the assumption that the introduction of a seat belt requirement has significantly reduced the number of fatal accidents. Also other analyses have found that there is a linkage between the number of crashes and the speed limit[4]. The goal of this project is to asses, whether this impact is present in the Californian data as well and to estimate the effect using a time-series regression model.

Two separate research questions based on the assumptions and previous work that each assess different criteria, are proposed:

- **Did the introduction of the seat belt requirement impact the percentage of fatal accidents in California)**
- **Did the increase of the speed limit from 55 to 65 miles per hour impact the total number of accidents in California)**

If the underlying hypotheses (each research question is answered to be true) are correct, then the final result of this project will be a model, from which one will be able to derive the impact of the introduction of the laws on the dependent variable.

[3] https://www.cdc.gov/motorvehiclesafety/seatbelts/facts.html
[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2724439/

1.3 Methodology

To answer these research questions, the group developed a number of functions in the statistical programming language R. R was chosen, because of it's flexibility and completeness. The group is able to easily manipulate the data, while in the same function being able to perform a statistical test, by including an external library. This section will explain the process and methods that the group has developed to build a reliable time-series regression model.

Overall the work was split in two major categories: The Checks for the Gauss-Markov assumptions and the adaptions and transformations required if one of these assumptions was violated. Since the Gauss-Markov assumptions need to be checked for both models that will be develop to answer the research questions, they were put into a separate R file. This file has the following structure:

- Load Packages (corrplot, het.test, lmtest)
- General function that calls all methods "test_gm_assumptions" (Parameters: dependent data, model, significance level)
- 1. Check H2: No Multicollinearity
    - Plot the correlation matrix
    - Check whether the matrix of independent data has full rank. If this is the case, return TRUE
  2. Check H3: Zero Conditional Mean
    - Plot the correlation matrix
    - Check whether any variables has a correlation of 1 with the residuals. If this is not the case, return TRUE
  3. Check H4: Homoscedasticity
    - Perform the Breusch-Pagan Test
    - Check wether the p-value of the test is higher than the significance. If it is, return TRUE
  4. Check H5: No serial correlation
    - Perform the Durbin-Watson Test. If the p-value is higher than the specified significance, return TRUE
    - Apply Breusch-Godfrey Test iterating through increasing indexes and return the first index, under which the p-value is higher than the specified significance.

5. Check H6: Normal Distribution of the errors
   - Perform the Shapiro-Wilk test on the residuals of the model. If the p-value of the test is higher than the specified significance, return true.

Each of these functions can be either called through the overall assumption-test method described above function or individually. This is important as some of the results might not correct, i.e. if the OLS estimator is biased or inconsistent. The aforementioned tests were all included using R packages. More precisely, the Shapiro-Wilk test was used from the *stats* package[5], the Breusch-Godfrey, Durbin-Watson and Breusch-Pagan test were used from the *lmtest*package[6]. As these tests, the Gauss-Markov theorem and all assumptions were already discussed in class, these will not be explained further.

The other part of the R code that was developed is focused on the creation of a model and its adaption according to each research question. The detailed adaption can be seen in chapter 2. The main tests, corrections and plots that might be used, will be explained in this chapter.

For each of the research questions the first step is to develop a linear model that can be estimated using OLS (applied using the lm function from the *stats* package). All other steps will only be needed, should at least one assumption of the Gauss-Markov theorem be violated.

Throughout the project, the group might use the diff function from the R *base* to differentiate parts of the model to account for serial correlation. The differentiation allows the group to only look at the changes from one observation to the following instead of the overall estimation. The Wald-test might be used to check for the joint significance of some parameters, like the seasonal dummies or others. A waldtest function is also included in the *lmtest* package.

In case the observations will be serially correlated, the group decided to rely on the Prais-Winsten estimation, which corrects the errors of an OLS estimation for serial correlation of AR(1).

Alternatively, and also in the case of Heteroscedasticity, the Newey-West estimation will be used to get robust standard errors for models that are not consistent, when applying the OLS method. This method is included in the *sandwich* R package [7]

In addition to the aforementioned Breusch-Godfrey method to retrieve the order of auto-correlation, the auto-correlation estimates were also plotted using the

[5]https://stat.ethz.ch/R-manual/R-devel/library/stats/html/shapiro.test.html
[6]https://cran.r-project.org/web/packages/lmtest/index.html
[7]https://cran.r-project.org/web/packages/sandwich/sandwich.pdf

auto-correlation function from the *stats* package.

The Dickey-Fuller test might be applied using the adf.test function from the *tseries* package[8]. It can be used to test for Stationarity in time-series and was discussed in the theoretical class.

1.3.1 Introduction of other correction methods

Out of the tests and corrections the group will use, the Prais-Winsten and Newey-West estimations were not yet discussed in the Inferential statistics class and will therefore be introduced in this chapter.

- **Prais-Winsten estimation**[9]

  "The problem with the GLS estimator is that ρ is rarely known in practice. However, we already know how to get a consistent estimator of ρ: we simply regress the OLS residuals on their lagged counterparts, exactly as in [the GLS estimator]. Next, we use this estimate, $\hat{\rho}$, in place of ρ to obtain the quasi-differenced variables. We then use OLS on the equation

  $$\tilde{y} = \beta_0\tilde{x}_{t0} + \beta_1\tilde{x}_{t1} + ... + \beta_k\tilde{x}_{k1} + error_t \tag{1.1}$$

  where $\tilde{x}_{t0} = (1 - \hat{\rho})$ for $t \leq 2$, and $\tilde{x}_{10} = (1 - \hat{\rho}^2)^{1/2}$. This results in the feasible GLS (FGLS) estimator of the β_j. The error term in 1.1 contains e_t and also the terms involving the estimation error in $\hat{\rho}$. Fortunately, the estimation error in $\hat{\rho}$ does not affect the asymptotic distribution of the FGLS estimators."

  This excerpt from the Woolridge Introcuctory Econometrics book talks about the need to use feasible GLS estimators in case of serial correlation. One of the implementations is the Prais-Winsten estimation, which iteratively applies the GLS regression, uses the ρ estimate to compute the residuals and transform the data and then uses OLS again to estimate the regression. The result of this iterative of this iterative procedure will be a consistent model with appropriate standard errors, even for a serially correlated problem. It should be noted that the Prais-Winsten estimation is only apprropriate for serial correlation of $AR(1)$.

- **Newey-West estimation**[10]

  Similar to the Prais-Winsten estimation, the Newey-West estimation also provides serial-correlation robust standard errors for model with serial correlation. Additionally it also accounts for potential heteroscedasticity in the

[8] https://cran.r-project.org/web/packages/tseries/tseries.pdf
[9] Woolridge, Introductory Econometrics: A Modern Approach, p.425
[10] Woolridge, Introductory Econometrics: A Modern Approach, p.432

model and works with higher-order serial correlations. This means that in case of a serially-correlated and heteroscedastistic model, the group will be using to Newey-West to get robust error estimations.

The estimation itself is not iterative, but instead uses the estimated standard error by a standard OLS model for β and the overall regression. Additionally it uses the residuals $\hat{r}_t$ of the regression x_{t1} on $x_{t1}...x_{tk}$, where x represents the independent variables in the regression model, and the intercept $\hat{u}_t$ to calculate the auxiliary variable $\hat{v}$.

$$\hat{v} = \sum_{t=1}^{n} \alpha_t^2 + 2 \sum_{h=1}^{g} [1 - h/(g+1)] \left(\sum_{t=h+1}^{n} \tilde{\alpha}_t \tilde{\alpha}_{t-h} \right) \tag{1.2}$$

where

$$\hat{\alpha}_t = \hat{r}_t \hat{t}_t \tag{1.3}$$

In the above equation, g is an integer value that represents the order of the serial correlation. The standard errors for one coefficient can then be robustly estimated through the following formula:

$$robust\_se(\hat{\beta}_t) = [se(\hat{\beta}_t)/\hat{\sigma}]^2 \sqrt{\hat{v}} \tag{1.4}$$

This can then be used to calculate confidence intervals and the t-statistic.

Chapter 2

Model selection and hypotheses testing

The following chapter serves to validate our research hypotheses proposed in chapter 1.2. The group tested the two hypotheses separately as described in the following two subchapters. The subchapters outline the group's thought process and their approach to derive at the final conclusion. Furthermore, these chapters describe any data transformations that were conducted during this process.

2.1 H1: The percent of fatal accidents decreased after the introduction of the belt law

The seat belt was introduced to increase passenger safety. It is a safety measure inside the car that aims to reduce the gravity of injuries during accidents. Therefore, the group assumed that the total number of accidents is not affected by the seatbelt law, whereas the fatality of these accidents should be affected by the seatbelt requirement. In order to correct for the general development of accidents due to external factors such as a growing amount of cars, the group chose the percentage of fatal accidents which should behave independently of the number of total accidents. Thus, the group is controlling the model for total accidents. The research question to test for is therefore: Did the introduction of the seat belt law lead to a significant decrease in the percentage of fatal accidents?

The basic assumption of the model can thus be described as following (where percentage of fatal accidents (prcfat) is the dependent variable of a model including the dummy variable beltlaw [0 - law not introduced; 1 - law introduced]):

- $H_0 : \beta_{beltlaw} = 0$
- $H_1 : \beta_{beltlaw} \neq 0$

The above described research question implies the usage of a transformed variable describing the fatal accident relative to the total accidents multiplied by 100: $prcfat = \frac{fatacc}{totacc} \times 100$.

Next, the group plotted the prcfat variable against the time variable to get a visual understanding of the time series development. Furthermore, the group added two

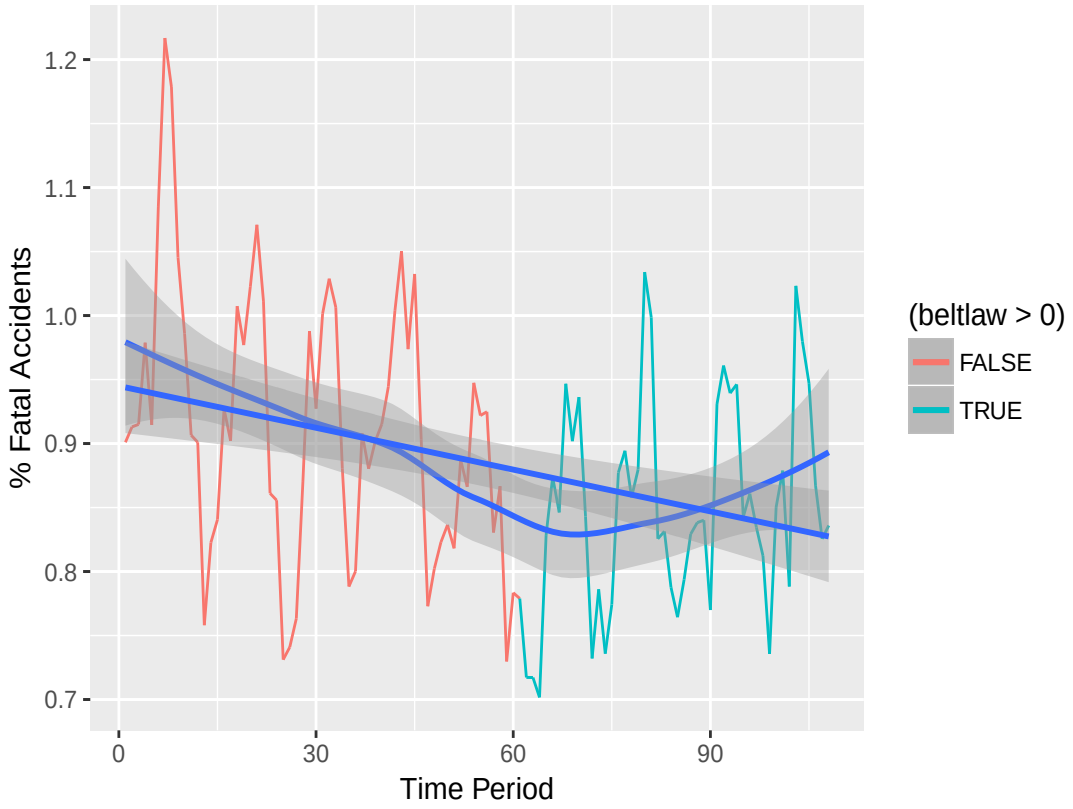

Figure 2.1: Plot of Fatal Accidents over time

possible trend lines in blue to indicate a linear trend and a data-fitted trendline without a particular polynomial shape. As the hypothesis aims to test for the influence of the introduction of the seat belt law (beltlaw = 1), the corresponding data was coloured differently in the line chart (see 2.1). One can clearly identify a trend of some form in the time series. Furthermore, the visual display shows evidence of seasonal differences within a year that should be tested for. The influence of beltlaw is visually difficult to determine as one needs to account for the trend and other factors. The curve of % Fatal Accidents does not indicate the necessity for any transformations (e. g. log) to correct for non-linearity.

2.1.1 Base Model

The initial model regressed prcfat as the dependent variable against the independent variables t, unem , spdlaw, beltlaw, wkends, feb, mar, apr, may, jun, jul, aug, sep, oct, nov, dec. Thus the model accounts for the trend (t), seasonality (monthly dummy variables), the unemployment rate (unem) and the dummy variables spdlaw and beltlaw.

Table 2.1: Base Model OLS results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.030	0.103	10.003	0
t	−0.002	0.0004	−5.312	0.00000
unem	−0.015	0.006	−2.782	0.007
spdlaw	0.067	0.021	3.262	0.002
beltlaw	−0.030	0.023	−1.270	0.207
wkends	0.001	0.006	0.102	0.919
feb	0.001	0.029	0.030	0.976
mar	0.0001	0.027	0.003	0.997
apr	0.058	0.028	2.093	0.039
may	0.072	0.028	2.592	0.011
jun	0.101	0.028	3.604	0.001
jul	0.177	0.027	6.479	0
aug	0.193	0.027	7.018	0
sep	0.160	0.028	5.674	0.00000
oct	0.101	0.028	3.651	0.0004
nov	0.014	0.028	0.496	0.621
dec	0.009	0.028	0.330	0.742

Observations: 108
R^2: 0.717
Adjusted R^2: 0.668
Residual Std. Error: 0.058 (df = 91)
F Statistic: 14.439^{***} (df = 16; 91)

Assuming that all the GM assumptions hold, we have some intuitive evidence of seasonality as some month, especially jun, jul aug, sep and oct show individual significance at the 0.1% level. Furthermore, the trend in the model is significant at the 0.1% level, indicating that one additional month in average reduces the percentage of fatal accidents by -2.235e-03 percentage points holding other factors fixed. The introduction of the speed law (spdlaw) seems to be significant on the percentage of fatal accidents at a 1% level leading to an average reduction of 6.709e-02 percentage points after its introduction holding other factors fixed whereas the belt law (beltlaw) is not even significant at the 10 % threshold. Interestingly, this model detects an influence of the unemployment on the percentage of fatal accidents at the 1% level of -1.543e-02 percentage points of fatal accidents per unit increase in state unemployment rate c. p. Testing for the GM assumptions yields the following results (see section 1.3):

1. H1: Linear in Parameters: True
2. H2: Zero Conditional Mean: True
3. H3: No Perfect Multicollinearity: True
4. H4: Homoscedasticity: True
5. H5: No serial correlation: False
6. H6: Normal distribution of the errors: True

Following the above described violation of the fifth GM assumption, the described coefficient estimates are still unbiased but not necessarily BLUE. Furthermore, the estimated standard errors are not correct which makes the statistical inference described above invalid.

Considering the serial correlation, the ACF plot suggests a seasonal behaviour of the target variable where multiple lagged instances pass the threshold of 0.2. Nonetheless, the ACF's results can be further used as a visual understanding of the presence of serial correlation, possibly up to order 3. When testing for serial correlation, the Durbin-Watson test rejects the Null-Hypothesis of no serial correlation at the 1% level and thus shows strong evidence of serial correlation within the time series by supporting the alternative hypothesis that true autocorrelation is greater than 0. The Breusch-Godfrey test comes to similar results that align well with the discoveries in figure 2.2 by suggesting serial correlation of order 3. This means that the test was only able to not reject the H0 of no serial correlation at the 5% level when testing for serial correlation of order 4 (p-value = 0.07444).

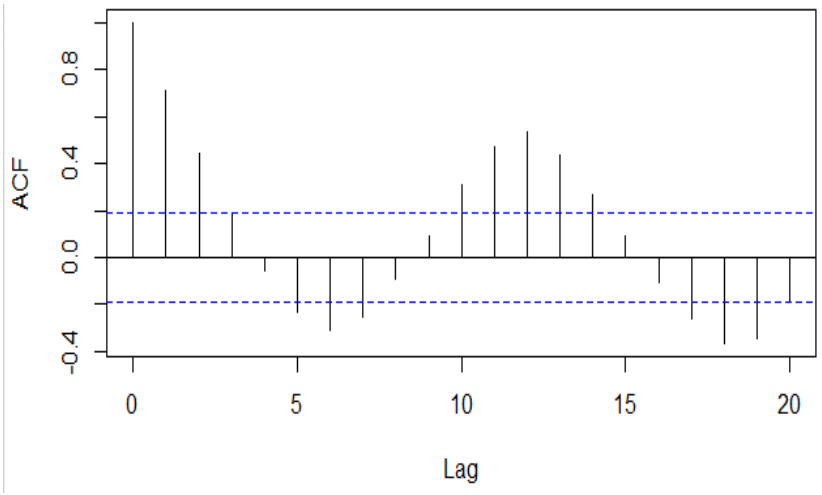

Figure 2.2: ACF plot of rfatacc

2.1.2 Model Corrected for serial correlation

The group suggested two possible methods to correct for the serial correlation in the data

- Prais-Winsten Method
- Differentiating

The first approach corrects for the errors using the Prais-Winsten method. The idea was to correct for serial correlation of order 1 and then retest the model. The test is explained in 1.3. After correcting for the serial correlation of order 1, all GM assumptions held. Therefore, the model has unbiased estimators that are BLUE and allow statistical inference.

Interpreting this model (shown in Table 2.2), we can see that there is strong evidence (0,1% level) of a trend, indicating that the percentage of fatal accidents decreases by 0.002 percentage points per month over time in average c.p. Unemployment (unem) is not significant at the 5% level anymore. The spdlaw still has a significant impact at the 5% level on the percentage of fatal accidents by increasing the percentage by 0.064 percentage points after its introduction in average holding other factors fixed, whereas the beltlaw is not even significant at the 10% level now.

The second approach to correcting for serial correlation is differentiating the time-series: $\Delta Y = Y_t - Y_{t-1}$.

Table 2.2: Serial Correlation Corrected OLS results

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	1.009	0.102	9.928	0.00
t	−0.002	0.001	−3.922	0.000
unem	−0.013	0.007	−1.855	0.067
spdlaw	0.064	0.027	2.394	0.019
beltlaw	−0.025	0.030	−0.824	0.412
wkends	0.001	0.005	0.123	0.902
feb	−0.001	0.024	−0.037	0.970
mar	−0.001	0.026	−0.044	0.965
apr	0.058	0.028	2.086	0.040
may	0.072	0.028	2.578	0.012
jun	0.101	0.028	3.590	0.001
jul	0.175	0.027	6.404	0.000
aug	0.192	0.028	6.957	0.000
sep	0.160	0.028	5.647	0.000
oct	0.101	0.028	3.638	0.001
nov	0.013	0.027	0.487	0.627
dec	0.009	0.025	0.348	0.729

Observations: 108
R^2: 0.9951
Adjusted R^2: 0.9942
F Statistic: 1024^{***} (df = 18; 90)
Rho: 0.2159793

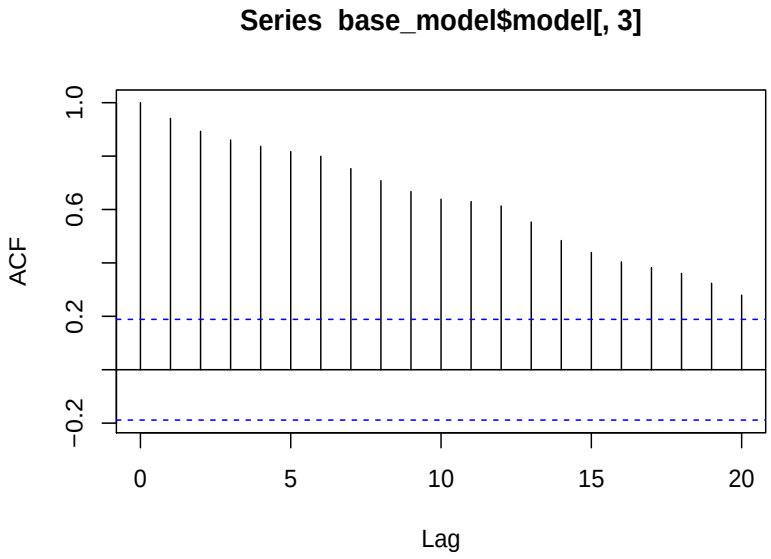

Figure 2.3: ACF of unemployment (unem)

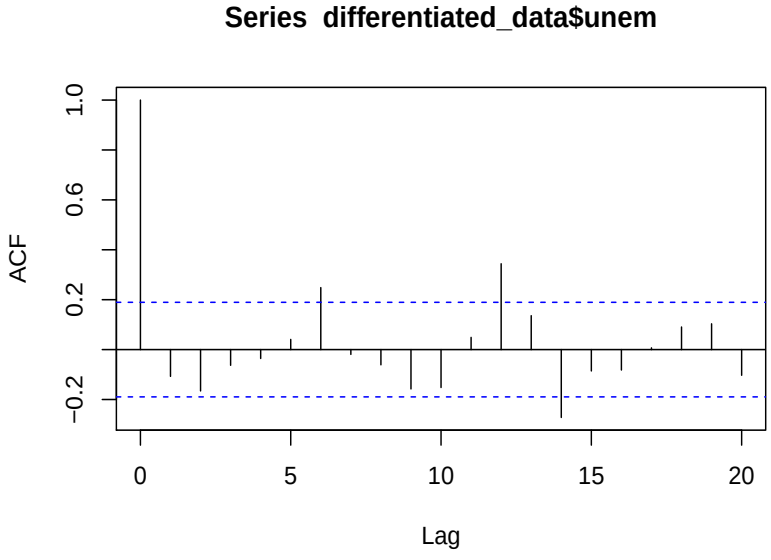

Figure 2.4: ACF of differentiated unemployment (Δunem)

The Adjusted-Dickey-Fuller test for prcfat was able to reject the H0 hypothesis of non-stationarity and thus suggests strong evidence of stationarity at 1% significance level. Contrarily, the unemployment is not able to reject the same Null-Hypothesis at the 5% significance level and thus suggests non-stationarity which can be visually validated using the ACF plot in figure 2.3. Therefore, the prcfat variable is differentiated in order to account for serial correlation and the unem variable is differentiated to make it stationary. After differentiating, unem suggests strong evidence of stationarity at 1% significance level which is also indicated in the ACF plot of the differentiated variable in figure 2.4.

The OLS model using the differentiated values for unem and prcfat does not violate any of the GM assumptions. Thus, the model is unbiased, BLUE and statistical inference is valid.

Table 2.3: Differentiated Model OLS results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−0.127	0.105	−1.210	0.229
t	0.0001	0.0005	0.296	0.768
unem	0.013	0.016	0.778	0.439
spdlaw	−0.007	0.024	−0.302	0.763
beltlaw	0.001	0.027	0.031	0.975
wkends	0.007	0.007	0.942	0.349
feb	0.035	0.037	0.935	0.352
mar	0.042	0.039	1.077	0.284
apr	0.099	0.038	2.574	0.012
may	0.057	0.037	1.517	0.133
jun	0.054	0.035	1.554	0.124
jul	0.088	0.033	2.653	0.009
aug	0.059	0.040	1.485	0.141
sep	0.007	0.038	0.172	0.864
oct	−0.032	0.035	−0.920	0.360
nov	−0.059	0.035	−1.669	0.099
dec	0.027	0.036	0.751	0.455
Observations: 107				
R^2: 0.344				
Adjusted R^2: 0.227				
Residual Std. Error: 0.067 (df = 90)				
F Statistic: 2.945^{***} (df = 16; 90)				

The coefficient estimates described in 2.3 the estimated effects on the differentiated prcfat. Accounting for autocorrelation, stationarity, time trend and season-

ality and other factors, there is no evidence that beltlaw has any effect on the change in percentage of fatal accidents (prcfat) from the sample data. Nonetheless, most independent variables do not show statistical signficance at a 5% level, such as unemployment (unem), speed law (spdlaw), belt law (beltlaw), weekend days (wkends). Although this model does not violate any GM assumptions, it is not very helpful in explaining prcfat. Nonetheless, both the Prais-Winsten and the differentiated model do not show evidence of the beltlaw having a significant impact on the fatality of the accidents (prcfat), which is the original research question of this chapter.

The group decided to continue with the Prais-Winsten corrected model of the undifferentiated prcfat variable for further tests because it was able to provide more significant model parameters and does not require to reduce the number of observations.

2.1.3 Explaining the dependent variable with a squared time trend

This OLS model was developed using the Prais-Winsten method. The resulting model does not violate any GM assumption, thus making statistical inference valid. The $\hat{t2}$ (tsq) parameter estimate is 2.156e-05, meaning a one unit increase in tsq leads to a 2.156e-05 unit increase in the percentage of fatal accidents. The std. error of tsq is 1.127e-05, leading to a t-value of 1.913. This does not indicate significance at the 5% level meaning the model does not show particularly strong evidence for having a squared time trend. Therefore, the squared timetrend was not included in the model.

2.1.4 Allowing for a lagged influence of the belt law on the percentage of fatal accidents

The idea of this model adjustment is that there is a certain adaption rate in using a seat belt after the law was introduced. This means that people do not start abiding the law from day one, especially since they were not fined at the beginning. As official statistics suggest, there was an increasing rate of seatbelt usage after the introduction of the law[1]. If the seat belt law has an influence on the percentage of fatal accidents, then this effect might not be seen from the start of the introduction of the law but there might be an adoption rate. This convergence is accounted for by introducing the variable txbeltlaw = t*beltlaw.

[1]Standard enforcement saves lives: the case for strong seat belt laws; US National Highway Traffic Safety Administration

This variable takes the value 0 until the introduction of the seatbelt law (01/1986) and then steadily increases with t from 61 to 108.

Inclusion of txbeltlaw in the base model described in section 2.1.1 proposes that a one unit change in txbeltlaw has a 0.0034469 unit change in the percentage of fatal accidents with a 1% signifiance level. Nonetheless, the statistical inference is not valid since H5 of the GM assumptions is also violated under this model.

Therefore, the group applied the Prais-Winsten method to obtain estimates for serially uncorrelated coefficients for the OLS model including the txbeltlaw variable. The corrected model does not violate any GM assumption.

Table 2.4: Serially Corrected Model including t*beltlaw as indep. var.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	1.026	0.097	10.595	0.000
t	−0.003	0.001	−5.130	0.000
unem	−0.014	0.006	−2.152	0.034
spdlaw	−0.004	0.035	−0.123	0.902
beltlaw	−0.245	0.086	−2.844	0.006
wkends	0.001	0.005	0.172	0.864
feb	−0.0004	0.024	−0.015	0.988
mar	−0.002	0.026	−0.071	0.944
apr	0.056	0.026	2.114	0.037
may	0.077	0.027	2.878	0.005
jun	0.105	0.027	3.910	0.000
jul	0.178	0.026	6.839	0.000
aug	0.194	0.026	7.379	0.000
sep	0.161	0.027	5.960	0.000
oct	0.100	0.026	3.798	0.000
nov	0.012	0.026	0.462	0.645
dec	0.006	0.024	0.247	0.805
txbeltlaw	0.003	0.001	2.670	0.009

Observations: 108
R^2: 0.9937
Adjusted R^2: 0.9925
F Statistic: 842.7^{***} (df = 17; 90)
Rho: 0.2887052

Interpreting the parameters, we can see that both txbeltlaw and beltlaw become significant at the 0.01 significance level, whereas spdlaw becomes insignificant even at the 0.5 level. Despite correcting for the trend and seasonality, unemployment is still significant at a 5% level. The introduction of the beltlaw in average

lead to a reduction of the percentage of fatal accidents of 0.245 percentage points c.p. On the other hand, as time increases from the introduction of the law, each additional month increases the percentage of fatal accidents by 0.003 according to the variable txbeltlaw holding other factors fixed. The model shows strong evidence of a linear trend in the time series with each month in average reducing the probability of fatal accidents by 0.003 percentage points. Interestingly, looking at the time trend t and the timetrend after the introduction of the seat belt law *txbeltlaw*, they equalize each other once the seat belt law was introduced (beltlaw=1), meaning that in average the time trend should be 0 c.p. after the introduction of the beltlaw.The group reformulated the model to test the following hypothesis:

$prcfat = \beta_1 + \theta \times t + b_{txbeltlaw} \times (txbeltlaw - t) + ...\ where\ \theta = \beta_t + \beta_{txbeltlaw}$

1. $H_0 : \beta_t = -\beta_{txbeltlaw} \implies \theta = 0$
   $H_1 : \beta_t \neq -\beta_{txbeltlaw} \implies \theta \neq 0$
2. Under the Null, $t = \frac{\theta}{se(\theta)} \sim t_{n-k}$
3. $W_{5\%} = \{t : t_{abs} > 1.96\}$
4. $t = \frac{\theta}{se(\theta)} = \frac{0.0007765}{0.0012099} \approx 0.642$
5. $t_{abs} \notin W_{5\%} \implies Cannot\ reject\ H_0$
6. We can only reject H_0 at a 50% (p = 0.522640) significance level meaning that there is strong evidence in the model that $\beta_t = \beta_{txbeltlaw}$, meaning that the model does not show a significant trend after the introduction of the seat belt law.

Looking at the parameters individually, there is strong evidence of some month having a significant influence on the percentage of fatal accidents, such as june, july or august. To validate the presence of seasonality, the group conducted a test:

1. $H_0 : \beta_{month} = 0\ where\ month = \{feb,\ mar,\ ...\ ,\ nov,\ dec\}$
   $H_1 : \exists\ \beta_{month} \neq 0\ where\ month = \{feb,\ mar,\ ...\ ,\ nov,\ dec\}$
2. Under the Null,
$$F = \frac{\frac{R^2_{UR} - R^2_R}{q}}{\frac{1 - R^2_{UR}}{n-k}} \sim F_{q,\ n-k}$$
3. $W_{5\%} = \{F : F_{obs} > F_{11,\ 108-17}\} \implies W_{5\%} = \{F : F_{obs} > 3.95\}$
4. F observation
$$F_{obs} = \frac{\frac{0.9951 - 0.9621}{2}}{\frac{1 - 0.9951}{91}} \approx 306.43$$

5. $F_{obs} \in W_{5\%} \implies$ *Reject* H_0 *at* 5% *significance level.*
6. The model shows strong evidence of seasonality in the dataset.

Interestingly, the warmer months (e. g. jun, jul, aug, sep) seem to increase the percentage of fatal accidents compared to the colder months (e. g. nov, dec, jan, feb, mar) in average holding other factors fixed. The smallest percentage of fatal accidents occurs during february with an average decrease of 0.0004 per cent compared to january. The highest percentage of fatal accidents occurs in average durring august with an increase of 0.194 percentage points of fatal accidents.

To sum up, the percent of fatal accidents decreased after the introduction of the belt law. Using the serially corrected model from table 2.4 which does not violate any GM assumption, one can identify a statistically significant decrease of prcfat by 0.245 percentage points in average holding other factors fixed and controlling for trend and seasonality. There is a generally decreasing trend in the percentage of fatal accidents by 0.003 percentage points per month in the timeseries. This trend can be observed until the introduction of the seatbelt law. Afterwards, no significant trend is identifiable, thus the timeseries does not move into any particular direction. It is difficult to find a logical explanation for this behaviour and it can be due to external, unexplained factors. Furthermore, there is strong evidence of monthly seasonality in the timeseries with a higher percentage of fatal accidents during the warmer month compared to the colder month.

2.2 H2: Testing Speed Law Effect on Number of Fatalities in Traffic Accidents on Rural Roads

The group's initial hypothesis is that after the introduction of the law that raises the speed limit on rural roads to 65 mph the number of fatalities on those roads would also increase controlling for other factors.

2.2.1 Base Model

Our initial model specification proposes the relationship of the dependent variable rtotacc with others to be of form

$$rtotacc = \beta_1\ t + \beta_2\ unem + \beta_3\ spdlaw + \beta_4\ beltlaw + \beta_5\ wkends + \beta_6\ feb + \beta_7\ mar + \beta_8\ apr + \beta_9\ may + \beta_{10}\ jun + \beta_{11}\ jul + \beta_{12}\ aug + \beta_{13}\ sep + \beta_{14}\ oct + \beta_{15}\ nov + \beta_{16}\ dec$$

and thus theoretically accounting for the seasonality, time trend, belt law, speed law, unemployment rate and number of weekends in a month. The initial model

shows the coefficient on spdlaw to be positive and highly significant with p-value close to zero.

The basic assumption of the model will be formulated as follows (where spdlaw has value 1 for when it has taken effect and 0 otherwise):

- $H_0 : \beta_{spdlaw} = 0$
- $H_1 : \beta_{spdlaw} \neq 0$

Table 2.5: Rural Roads Accidents Base Model OLS results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	179.794	52.584	3.419	0.001
t	0.853	0.215	3.968	0.000
unem	−2.986	2.832	−1.055	0.294
spdlaw	50.009	10.506	4.760	0.000
beltlaw	27.635	11.865	2.329	0.022
wkends	2.377	3.148	0.755	0.452
feb	−6.437	14.811	−0.435	0.665
mar	48.486	13.998	3.464	0.001
apr	42.767	14.209	3.010	0.003
may	68.507	14.119	4.852	0.000
jun	84.110	14.349	5.862	0.000
jul	122.479	13.923	8.797	0.000
aug	134.379	14.018	9.586	0.000
sep	79.692	14.405	5.532	0.000
oct	38.321	14.133	2.711	0.008
nov	58.351	14.375	4.059	0.000
dec	71.933	14.229	5.055	0.000

Observations : 108
R^2 : 0.891
Adjusted R^2 : 0.872
Residual Std. Error : 29.376 (df = 91)
F Statistic : 46.683^{***} (df = 16; 91)

However, to have more conclusive results we must first test if the Gauss-Markov assumptions hold and consequently statistical tests for significance of coefficients are valid. Following the tests described in 1.3 the following results were obtained for the six G-M assumptions:

1. H1: Linear in Parameters: True
2. H2: Zero Conditional Mean: True

3. H3: No Perfect Multicollinearity: True
4. H4: Homoskedasticity: False
5. H5: No serial correlation: False
6. H6: Normal distribution of the errors: True

We have some evidence of heteskedasticity in the residuals as the Breusch-Pagan test statistic was >38 with a respective p-value of 0.001 which made us reject the null hypothesis that there is no evidence of heteroskedasticity in the residuals. This can also be seen visually on the residuals plot Figure 2.5.

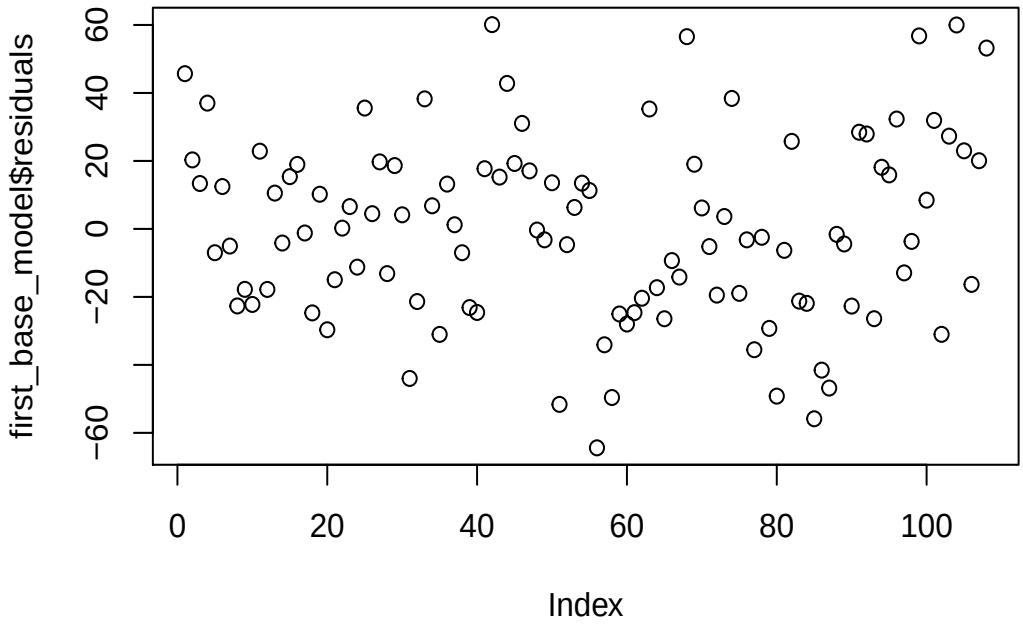

Figure 2.5: Residuals of the first model

Moreover there is also evidence of some serial correlation of the residuals based on the Durbin-Watson statistic of 1.44 and a respective p-value of 0.001, which made us reject the null hypothesis that there is no first-order serial correlation. As a result our estimated coefficients can no longer be BLUE and statistical tests are unreliable since the current standard errors are probably underestimated. In addition, the Breusch-Godfrey test showed signs of correlation for up to 4 lags reporting a LM statistic of 9.8 and a p-value of 0.044 for 4 lags as opposed to 10.49 and 0.063 for 5 lags. Under 0.05 significance we assume 4 lags.

2.2.2 Correcting the First Model for Heteroskedasticity and Serial Correlation

Since, in this model we have both evidence of heteroskedasticity and serial correlation, we have decided to use the Newey-West hetereskedasticity and autocorrelation robust standard errors (or HAC standard errors) mainly because the Newey-West correction can be applied for any number of lags, which we suspect are more than 1 in the model based on the tests of the GM assumptions.

Table 2.6: Rural Roads Accidents Newey-West Corrected coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	179.794	42.541	4.226	0.000
t	0.853	0.262	3.255	0.002
unem	−2.986	2.236	−1.336	0.185
spdlaw	50.009	13.140	3.801	0.000
beltlaw	27.635	14.725	1.877	0.064
wkends	2.377	2.630	0.904	0.368
feb	−6.437	7.083	−0.909	0.366
mar	48.486	12.689	3.821	0.000
apr	42.767	10.050	4.255	0.000
may	68.507	10.591	6.469	0.000
jun	84.110	12.761	6.591	0.000
jul	122.479	11.381	10.671	0.000
aug	134.379	15.675	8.573	0.000
sep	79.560	11.427	6.963	0.000
oct	37.843	12.824	2.951	0.004
nov	58.351	8.638	6.675	0.000
dec	70.626	10.456	6.754	0.000

Based on the HAC robust standard errors, in Table 2.6 we observe that the coefficient on the dummy spdlaw is still statistically significant under 0.05 confidence level. However, we still have some theoretical ground to adjust this model further which will be discussed in the next section 2.2.3.

2.2.3 Adjusting the Model for non-linear Time Trend and Testing Potential Lagged Spead law effect

As in the case with beltlaw hypothesis we suspect there might be factors that influence the effect of the spdlaw variable. Firstly, we felt there might be a

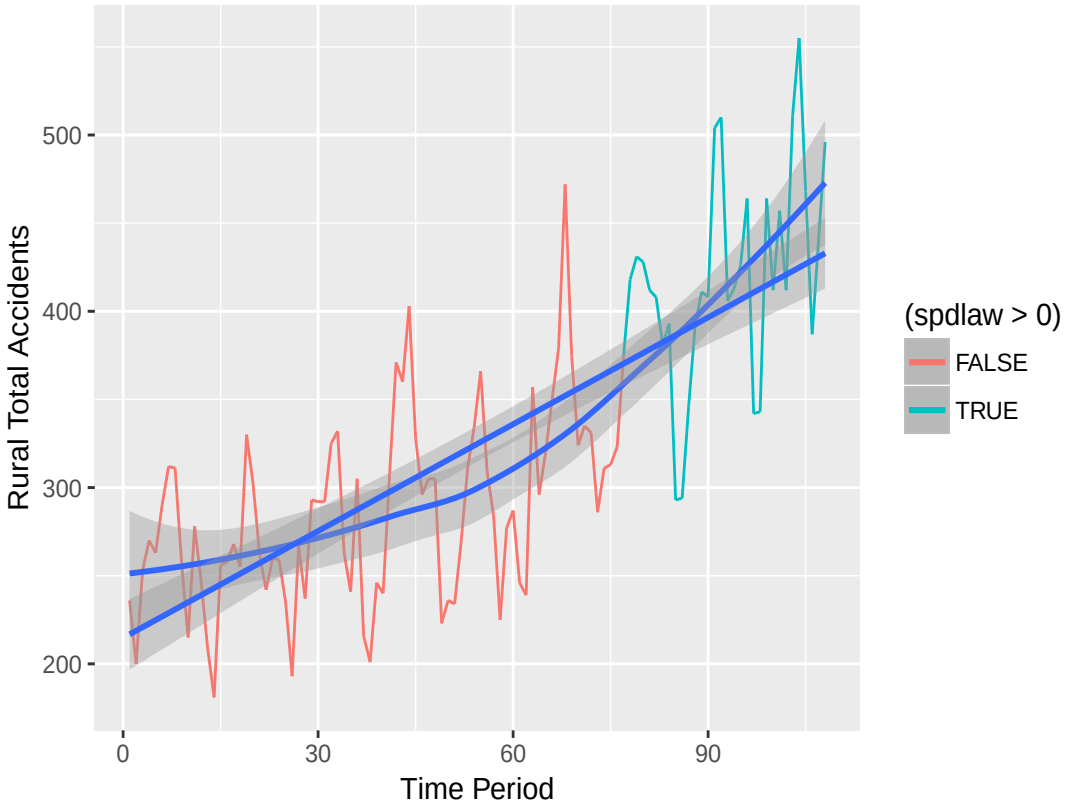

Figure 2.6: Plot of Rural Total Accidents over time

non-linear time trend associated with the series, especially since the dependent variable is nominal for the second hypothesis, which means it is very likely indirectly associated with a polynomial trend in population growth for example. To assess that we firstly plot number of accidents on rural roads against time in Figure 2.6 and added two trend lines linear and quadratic. Visually, the quadratic trend line fits the data better, so we will include a polynomial term t of order 2 (tsq).

Besides the correction for possible quadratic trend, we suspect that there might be an interaction between time and the speed law, since laws usually take some time to get used to. However, we suspect there will not be as pronounced of an effect in case of belt law if any, since it makes more sense that people would more readily drive at the new (higher) speed limit than start wearing a belt. The adjusted model seen in Table 2.7 shows some interesting things.

Even though the new variables tsq and t-spdlaw have insignificant estimated coefficients, the p-values for coefficients on spdlaw and t have increased and stopped being significant even under 0.1 significance level. However we keep in mind that this model suffers the same problems as the base one (heteroskedasticity and

Table 2.7: Adjusted Base Model OLS results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	193.432	50.199	3.853	0.000
t	0.807	0.829	0.972	0.333
unem	−3.723	3.314	−1.123	0.264
spdlaw	−194.060	132.867	−1.461	0.148
beltlaw	42.899	17.378	2.469	0.015
wkends	2.310	2.921	0.791	0.431
feb	−6.654	13.731	−0.485	0.629
mar	47.856	12.986	3.685	0.000
apr	41.714	13.210	3.158	0.002
may	70.159	13.235	5.301	0.000
jun	85.473	13.375	6.390	0.000
jul	123.863	12.912	9.593	0.000
aug	134.920	13.035	10.351	0.000
sep	79.560	13.436	5.921	0.000
oct	37.843	13.152	2.877	0.005
nov	57.503	13.351	4.307	0.000
dec	70.626	13.215	5.344	0.000
t_spdlaw	2.817	1.760	1.601	0.113
tsq	−0.004	0.014	−0.310	0.757

Observations : 108
R^2 : 0.909
Adjusted R^2 : 0.890
Residual Std. Error : 27.23 (df = 89)
F Statistic : 49.22^{***} (df = 18; 89)

serial correlation). To receive our final model we have applied the Newey-West correction to the standard errors of the model 2.7 and we have reeived results seen in table 2.8.

Table 2.8: Rural Roads Accidents Newey-West Corrected coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	193.430	44.816	4.316	0.000
t	0.806	0.812	0.993	0.323
unem	−3.723	3.392	−1.097	0.275
spdlaw	−194.060	127.290	−1.525	0.131
beltlaw	42.899	17.330	2.475	0.015
wkends	2.308	2.455	0.941	0.349
feb	−6.653	7.132	−0.934	0.353
mar	47.856	12.405	3.858	0.000
apr	41.714	9.536	4.374	0.000
may	70.159	8.193	8.564	0.000
jun	85.473	12.487	6.845	0.000
jul	123.860	10.667	11.612	0.000
aug	134.920	14.099	9.570	0.000
sep	79.560	11.427	6.963	0.000
oct	37.843	12.824	2.951	0.004
nov	57.503	8.702	6.608	0.000
dec	70.626	10.456	6.754	0.000
t_spdlaw	2.817	1.689	1.668	0.099
tsq	−0.004	0.013	−0.318	0.751

After controlling for seasonality, potentially non-linear time trend, unemplyment and number of weekend days, and using heteroskedasticity and serial correlation consistent standard errors we see that the coefficient of the speed law variable has a p-value of 0.131 which is not significant even under 0.1 significance level. We can state a similar conclusion for the interaction variable between time and spdlaw (t_spdlaw), the p-value for which is 0.099 which is exactly at the theshold of significance under the most lax 0.1 significance level. Unfortunately, we were unable to test the joint significance of these two variable since the output for the Newey-West corrected model includes only the coefficients but not $R\hat{2}$.

In conlusion, we did not detect evidence to suggest that the total number of accidents on rural roads have increased due to the change in the speed law (increase of maximum allowed by 10mph) when controlling for other factors.

Chapter 3

Conclusion and outlook

This chapter will summarize the results of the analysis by answering the research questions that were formulated in section 1.3. Additionally this chapter features a short outlook on other potential research questions and enhancements to the developed models.

3.1 Conclusion

The group analysed the *traffic2* dataset from the Woolridge textbook containing monthly timeseries data of number and type of traffic accidents in California between 1981 and 1986. Among other variables, the dataset contains the total number of accidents (totacc), the total number of fatal accidents (fatacc), the unemployment rate in California (unem), the number of weekend days per month including Fridays (wkends) and two binary dummy variables indicating regulatory changes taking effect in the state of California:

- Obligation of wearing a seat belt in the car (1986): beltlaw [1]
- Raise of the speed limit from 55 mp/h to 65 mp/h on freeways: spdlaw [2]

The group conducted an event study to determine whether these events had a significant impact on the traffic accidents in the state of California. To do so, the group used logical reasoning and visual analysis of the timeseries to derive two possible research hypotheses. Firstly, a seat belt is a safety feature within a car and is therefore not expected to have any significant influence on the total number of accidents. Nonetheless, the group wants to test for the significance of the seat belt law on the fatality of accidents. In order to correct for the general development of total accidents, the first hypothesis uses the percentage of fatal accidents (prcfat) as dependent variable. The second research question was directed at the speed limit law. As the speed limit directly influences the driving behaviour, the group wants to test whether the sample data provides evidence of

[1] http://articles.latimes.com/1986-01-11/local/me-26814_1_seat-belt-law
[2] http://articles.latimes.com/1986-02-12/news/mn-27520_1_speed-limit

a significant influence of the speed limit law on total number of accidents. The group formulated the two hypotheses:

1. The percent of fatal accidents decreased after the introduction of the belt law
2. Testing Speed Law Effect on Number of Fatalities in Traffic Accidents on Rural Roads

After correcting for serial correlation, the first model provided a significant impact of the seat belt law taking effect by reducing the percentage of fatal accidents 0.245 percentage points in average holding other factors fixed and controlling for trend and seasonality. This impact is significant at a 1% significance level. Furthermore, the timeseries shows significant seasonal behaviour and a significant linear trend of -0.003 percentage points per month c. p. until the seat belt law took effect. After the seat belt law taking effect, the trend is nullified and no longer significantly observable.

The group's second objective was to test if the speed law would have an impact on the total number of accidents on rural roads, theorising that higher speed limits would lead to more accidents. However, the group found no evidence of speed law having an effect on the total number of accidents on the rural roads after controlling for heteroskedasticity, serial corellation and other factors. The main tranformation that heavily impacted the significance of spdlaw variable was introducing a non-linear time trend and interaction between time trend and spdlaw.

3.2 Outlook

When looking back at the analysis being performed during this project, there is still some room for improvement, that could addressed in future research.

While the model did address the research questions that were relevant during this project, the data set did contain other potentially interesting dependent variables, that could have been analyzed. The group for example did not analyze the impact of the seat belt requirement on overall injury-involving accidents. This might be interesting, since a seat belt does not only protect against fatal injuries, but also (at lower speeds) against non-fatal injuries. Additionally the number of accidents by road type or a log-transformed number of accidents was also not analyzed in this project.

The group also included the independent variables wkends and unem in the model. In the end it turned out that those variables are not always significant and excluding them from the base model could have reduced the time and complexity

to achieve a consistent final model.

While the group did already account for seasonality in the model, the coefficients for the seasonal dummies showed, that both the number of total accidents as well as the percent of fatal accidents were higher in the summer month. This behaviour did not align with the assumption of the group members, who would have expected higher accident numbers in the winter.

Whereas the group did use data from the Woolridge textbook, which was specific to the US state california and the regulation changes in a timeframe in the 1980's it might be more relevant to include more recent data and more recent regulation changes. An example for this might be the introduction of the new traffic fine regulation in May 2014 in Germany. An interesting research area here, might be the influence of the regulation change on the number of accidents or the number of fines. Additionally a more recent data set could include the airbag requirement for passenger cars and their impact on fatal accidents, which was introduced in 1998. It might also have been interesting to include the car brand or price in the anylsis, but this data was not available to the group at the time of the project.

Chapter 4

Appendix

Table 4.1: Data Description

	variable	label
1	year	1981 to 1989
2	totacc	statewide total accidents
3	fatacc	statewide fatal accidents
4	injacc	statewide injury accidents
5	pdoacc	property damage only accidents
6	ntotacc	noninterstate total acc.
7	nfatacc	noninterstate fatal acc.
8	ninjacc	noninterstate injur acc.
9	npdoacc	noninterstate property acc.
10	rtotacc	tot. acc. on rural 65 mph roads
11	rfatacc	fat. acc. on rural 65 mph roads
12	rinjacc	inj. acc. on rural 65 mph roads
13	rpdoacc	prp. acc. on rural 65 mph roads
14	ushigh	acc. on U.S. highways
15	cntyrds	acc. on county roads
16	strtes	acc. on state routes
17	t	time trend
18	tsq	tˆ2
19	unem	state unemployment rate
20	spdlaw	=1 after 65 mph in effect
21	beltlaw	=1 after seatbelt law
22	wkends	# weekends in month
36	prcfat	100∗(fatacc/totacc)
37	prcrfat	100∗(rfatacc/rtotacc)
42	prcnfat	100∗(nfatacc/ntotacc)
46	spdt	spdlaw∗t
47	beltt	beltlaw∗t
48	prcfat_1	prcfat[_n-1]