Bill's Ice Cream Store

Noah Lomnitz, Abbas Siddiqui, Achyuth Kolluru

April 11, 2024

Contents

1 Importance and Context

Each day millions of Americans place their lives in the hands of other drivers on the roads. Road safety is a paramount concern in society, our research delves into how the age of drivers amongst many other factors plays a crucial role in the severity of car crashes. This report aims to answer the question:

What is the relation between drivers' ages in car crashes and the rates of injuries and fatalities, and what are the primary risk factors contributing to drivers of varying ages?

The dataset spans a wide range of driver ages, from 6 to 99 years, allowing for a thorough examination of age-related patterns in crash fatalities. Additionally, variables such as alcohol use, drug use, gender, and seat belt usage provide contextual information for each crash incident.

Through the FARS dataset, we gain valuable insights into the complexities of car crashes and their outcomes, particularly regarding the relationship between driver age and the percentage of crashes resulting in fatalities. This analysis aims to contribute to road safety discussions by uncovering patterns and potential risk factors associated with different age groups of drivers.

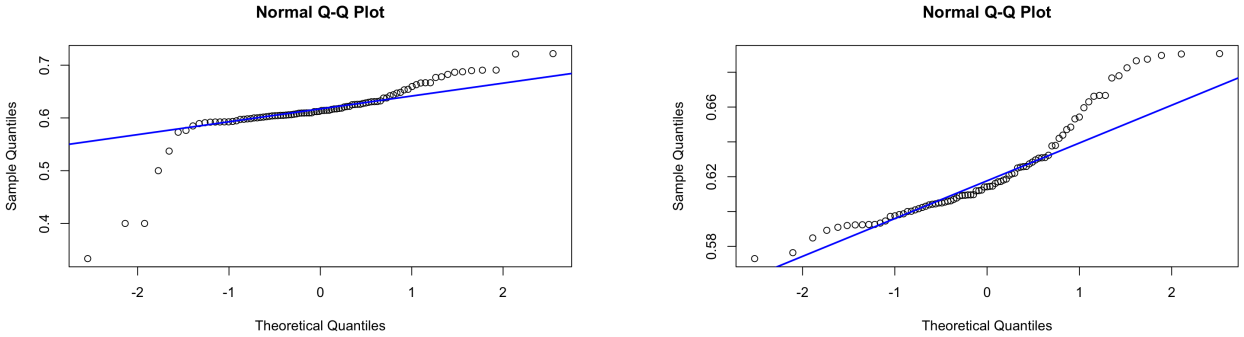
In this analysis, the key concept we are interested in is the percentage of injuries or deaths (PCT_INJ_or_DEATH) resulting from car accidents, which is our dependent variable (Y). The primary independent variable (X) chosen to investigate its relationship with PCT_INJ_or_DEATH is the age of the drivers involved in the accidents (AGE). We hypothesize that there may be a relationship between the age of drivers and the likelihood of injuries or deaths in accidents.

Additionally, several other variables were used to further explore the dataset:

- **Total_Number_DRIVER_ALC:** This variable represents the total number of drivers involved in alcohol-related accidents. It is used to calculate the percentage of drivers involved in alcohol-related accidents (Pct_Driver_Alc).
- **Male_Count and Female_Count:** These variables indicate the number of male and female drivers involved in accidents, respectively. They are used to calculate the percentage of male and female drivers involved in accidents (Pct_Male and Pct_Female).
- **RUR_Count and URB_Count:** These variables represent the number of accidents that occurred in rural and urban areas, respectively. They are used to calculate the percentage of accidents in rural and urban areas (Pct_RUR and Pct_URB)
- **HIGHWAY_Count:** This variable denotes the number of accidents that occurred on highways. It is used to calculate the percentage of accidents on highways (Pct_Highway).
- **Vehicle_Related_Accident, Nature_Related_Accident, Infrastructure_Related_Accident, Living_Entities_Accident, Other_Accident:** These variables represent the number of accidents related to different causes, such as vehicle-related, nature-related, infrastructure-related, accidents involving living entities, and other causes. They are used to calculate the percentage of accidents related to each cause.
- **Passenger_Cars, Trucks_and_Vans, Motorcycles, Off_Road_Vehicles, Others:** These variables indicate the number of accidents involving different types of vehicles, such as passenger cars, trucks and vans, motorcycles, off-road vehicles, and others. They are used to calculate the percentage of accidents involving each type of vehicle.
- **DRINKING_Count:** This variable represents the total count of alcohol-related accidents. It was used in some models alongside other variables to investigate its impact on the percentage of injuries or deaths.
- **NUM_PPL_INVOLVED:** This variable denotes the total number of people involved in an accident. While it was not found to be statistically significant in some model tests, it was considered as a potential explanatory variable.

We opted to summarize 97 thousand different crashes to a total of 92 by the age of drivers involved in order to study the risks of each age group. Since we are interested in studying age groups in the range of 15 to 97, we dropped a total of 9 observations from the dataset during pre-processing. Additionally, these points appeared to be outliers that significantly differed from the majority of the data. Leaving them in could have a substantial impact on statistical analyses, potentially skewing results and interpretations.

Upon visual examination using a quantile-quantile (QQ) plot, it became apparent that these 9 observations had extreme values for the percentage of injuries or deaths. The QQ plot compares the distribution of a dataset to a normal distribution, and deviations from the straight line indicate potential outliers. In this case, the outliers were distorting the linearity of the QQ plot, suggesting that they were not representative of the majority of accidents in the dataset.



(a) Original Data

(b) Cleaned Data

To maintain the integrity and reliability of the analysis, it was decided to exclude these outliers. By omitting these points the analysis aimed to capture a subset of accidents that were more typical and representative of the majority. This filtering process was essential to ensure that the relationship between age and the percentage of injuries or deaths in car accidents was not disproportionately influenced by these extreme values. This decision allowed for a more accurate and meaningful examination of the relationship between age and the percentage of injuries or deaths in car accidents.

2 Modeling

Moving on to the modeling stage, several key transformations and model specifications were applied to explore the relationship between various variables and the percentage of injuries or deaths. The data was split into exploration and confirmation sets, with 30% of the data allocated for exploration via random stratified sampling to ensure that the exploration and confirmation subsets span the entire range of age groups.

We created a base linear model of just percentage of injuries or deaths (PCT_INJ_or_DEATH) fitted to age (AGE). The resulting residual plot from this model displayed a parabolic behavior.

When deciding which features and transformations to use in advanced modeling, we opted to start by testing various transformations of our variables. Informed by the residual plot of our base model, we fit models using the polynomial transformation of (AGE) over many degrees. We found that the second degree polynomial of (AGE) fit the data much better than in its linear form. Additionally, we found interaction terms such as (AGE)*(DRINKING_Count) and (AGE)*(Pct_Veh_Rel_Acc) to provide important information to our models.

We tested many different variables such as DRUGS_Count, Male_Count, Female_Count, RUR_Count, URB_Count, and HIGHWAY_Count to investigate their influence on the percentage of injuries or deaths in this way; we used the correlation matrix to support modeling decisions. We also created normalized versions of certain variables such as Pct_Passenger_Cars so we could aggregate statistics by age group rather than look at each crash instance, and uniformly standardize their range of values to be between 0 and 1.

One important point to note is that because our target variable (PCT_INJ_or_DEATH) is a ratio, it imposes linearity on our dataset and loses information from the original variables. As a result, we need to acknowledge that variables that may have a nonlinear relationship with PCT_INJ_or_DEATH may have altered relations. For the final set of variables we consider in our modeling, we believe that this influence likely doesn't change variables such as Male_Count, RUR_Count, etc but may have changed the relations between variables such as DRINKING_Count.

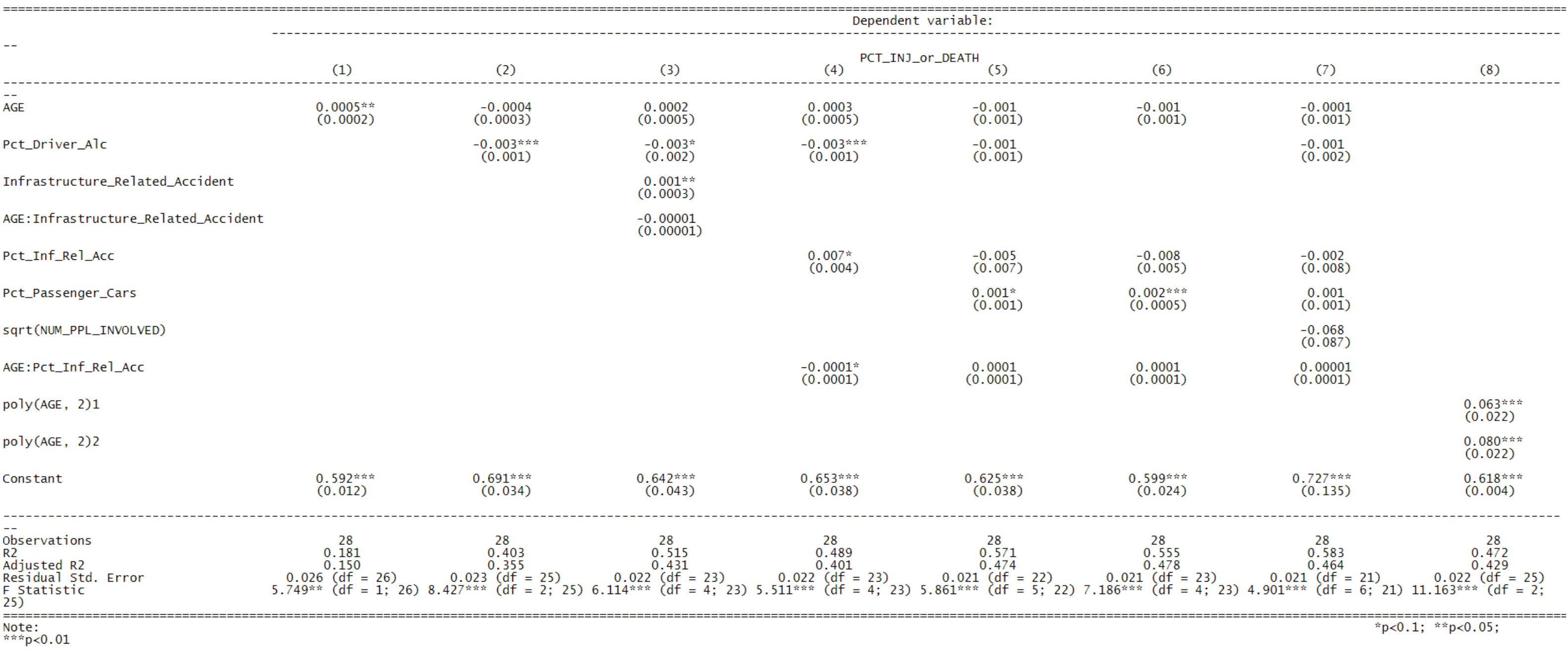
Having narrowed down the breadth of transformations and combinations of features to consider, we used this information to inform our ANOVA testing process for developing our final set of linear models. We utilized forward step-wise selection when deciding which variables to keep in each iteration of our ANOVA testing. Additionally, we studied each model's residual plot to ensure normally distributed residuals since without this the t-statistics and p-value were not guaranteed to be meaningful.

In the end we arrived at this model:

$$lm(PCT\_INJ\_or\_DEATH \sim AGE + Pct\_Driver\_Alc + AGE{*}Pct\_Inf\_Rel\_Acc + Pct\_Passenger\_Cars + \sqrt{NUM\_PPL\_INVOLVED})$$

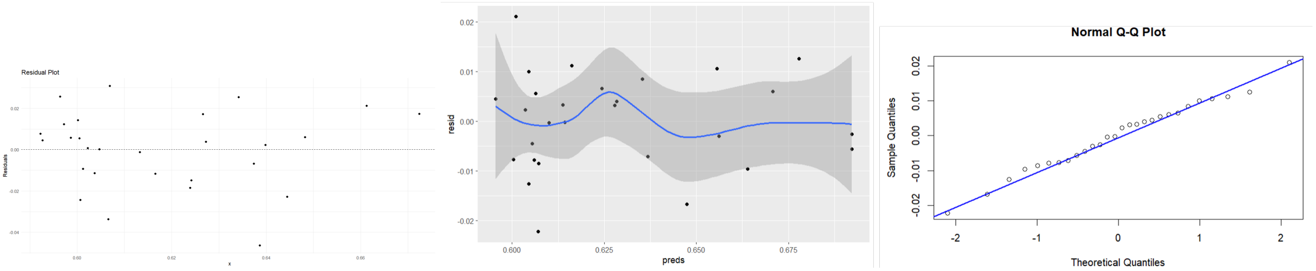
Compared to the base_model and the quadratic AGE model, we can see that our ANOVA model does a better job of fitting to the subtle changes along our data's distribution. However we opted to not using the quadratic model as it altered the distribution of the residuals needed to satisfy the CLM condition.

We can track the progression of models by observing the stargazer table of our ANOVA models.



```
==================================================================================================================================================================================================================
                                                                                                    Dependent variable:
                                     -----------------------------------------------------------------------------------------------------------------------------------------------------------------------------
--
                                                                                                      PCT_INJ_or_DEATH
                                              (1)                   (2)                   (3)                   (4)                   (5)                   (6)                   (7)                   (8)
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
--
AGE                                        0.0005**              -0.0004                0.0002                0.0003                -0.001                -0.001               -0.0001
                                          (0.0002)              (0.0003)              (0.0005)              (0.0005)              (0.001)               (0.001)               (0.001)

Pct_Driver_Alc                                                   -0.003***             -0.003*               -0.003***             -0.001                                      -0.001
                                                                 (0.001)               (0.002)               (0.001)               (0.001)                                     (0.002)

Infrastructure_Related_Accident                                                         0.001**
                                                                                      (0.0003)

AGE:Infrastructure_Related_Accident                                                   -0.00001
                                                                                      (0.00001)

Pct_Inf_Rel_Acc                                                                                              0.007*               -0.005                -0.008                -0.002
                                                                                                             (0.004)               (0.007)               (0.005)               (0.008)

Pct_Passenger_Cars                                                                                                                  0.001*               0.002***               0.001
                                                                                                                                   (0.001)               (0.0005)              (0.001)

sqrt(NUM_PPL_INVOLVED)                                                                                                                                                        -0.068
                                                                                                                                                                              (0.087)

AGE:Pct_Inf_Rel_Acc                                                                                         -0.0001*               0.0001                0.0001               0.00001
                                                                                                            (0.0001)              (0.0001)              (0.0001)              (0.0001)

poly(AGE, 2)1                                                                                                                                                                                       0.063***
                                                                                                                                                                                                    (0.022)

poly(AGE, 2)2                                                                                                                                                                                       0.080***
                                                                                                                                                                                                    (0.022)

Constant                                    0.592***              0.691***              0.642***              0.653***              0.625***              0.599***              0.727***              0.618***
                                            (0.012)               (0.034)               (0.043)               (0.038)               (0.038)               (0.024)               (0.135)               (0.004)

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
--
Observations                                   28                    28                    28                    28                    28                    28                    28                    28
R2                                           0.181                 0.403                 0.515                 0.489                 0.571                 0.555                 0.583                 0.472
Adjusted R2                                  0.150                 0.355                 0.431                 0.401                 0.474                 0.478                 0.464                 0.429
Residual Std. Error                     0.026 (df = 26)       0.023 (df = 25)       0.022 (df = 23)       0.022 (df = 23)       0.021 (df = 22)       0.021 (df = 23)       0.021 (df = 21)       0.022 (df = 25)
F Statistic                          5.749** (df = 1; 26) 8.427*** (df = 2; 25) 6.114*** (df = 4; 23) 5.511*** (df = 4; 23) 5.861*** (df = 5; 22) 7.186*** (df = 4; 23) 4.901*** (df = 6; 21) 11.163*** (df = 2;
25)
==================================================================================================================================================================================================================
Note:                                                                                                                                                                                  *p<0.1; **p<0.05;
***p<0.01
```

Presented in our Stargazer table, AGE consistently showed statistical significance. The square root transformation of NUM_PPL_INVOLVED not only improved our model's adaptability but assissted in meeting CLM conditions. Every factor in our final model had a non-zero coefficient, ensuring no perfect collinearity.



Linear Condition Expectation

Homoscedasticity

Normally Distributed Residuals

Figure 2: Other CLM conditions being met

To ensure independence for the CLM, we checked each data point, represented by its unique ST_CASE value, as distinct crashes. This approach ensured that each crash was treated independently within our analysis.

The above graphs show the other conditions of the CLM conditions being met. The linear condition expectation for our final model is met as points are both above and below the reference line indicating a linear trend. The second plot shows an even variance across the reference line indicating for no evidence for heteroscedasticity. Since the residuals follow the reference line closely, the model satisfies the normally distributed residuals condition.

3 Conclusion

Our analysis revealed a quadratic relationship between Age and PCT_INJ_or_DEATH, indicating that very young or very old drivers are associated with more severe crashes compared to middle-aged drivers. Among different crash types, Infrastructure Related Accidents stood out with the lowest number of survivors.

In conclusion, our findings emphasize the vulnerability of specific age groups, such as very young and elderly drivers, to more severe accidents. These insights are crucial for informing policies and interventions aimed at improving road safety and reducing the severity of car accidents.