Kyle Mabe

Econometrics

05/03/2025

Project Part 2

**Result Interpretation**

After running the logistic regression model, before looking at the results, we want to look at whether multicollinearity is an issue. If you look at the VIF results (see Appendix Figure 1), it differs from that of a usual VIF model primarily due to the fact that the model changes when it includes any predictor that is represented by more than one column in the design matrix (e.g., multi‑level factors, interactions, or polynomial terms). In that case the model reports the generalized VIF (GVIF) for the whole block of columns, shows the block’s degrees of freedom, and provides the adjusted figure GVIF^(1/(2·Df)) so you can compare it with the usual VIF thresholds of 5.

Based on the VIF model, there appears to be multicollinearity between *Speed.Limit* and *Speed.Squared* as both variables have a GVIF of 13.6. Since *Speed.Squared* is a byproduct of *Speed.Limit,* multicollinearity does not appear to be an issue for our regression output. Based on the adjusted VIF column, GVIF^(1/(2·Df)), that allows you to compare our interaction terms and multi-level factors with the threshold of 5, shows no signs of multicollinearity as their VIF does not exceed 5. Overall, there are not troubling signs of multicollinearity in our model to which we can now interpret the results.

Based on the results, see Appendix figures 2 and 3, the model predicts the probability of someone getting injured or not in a car crash (1=injury, 0=not injured) from 142,549 observations. The overall fit of the model uses a McFadden Pseudo r-squared which is 0.0184. *Speed.Limit* has a positive coefficient meaning every additional mph raises injury odds by 6% (holding all else constant) and is statistically significant with a P-value close to 0. *Speed.Squared* has a negative coefficient which moderates the linear increase, confirming a curved relationship between Speed limit and injury. *Speed.Squared* is also statistically significant as the p-value is close to 0. *Vehicle.Age* is statistically significant having a p-value close to 0, and for every extra year of vehicle age raises injury odds by around 0.8 % holding all else constant. *Driver\_Distracted* is statistically significant with a p-value close to 0 with distracted drivers showing 25 % lower injury odds than someone who is not distracted (holding all else constant). In terms of *Weather*, cloudy and rain appear to be statistically significant with cloudy having a p-value close to 0 and rain having a p-value of 0.03, all the rest of the weather types are not significant. Crashes that occur under cloudy conditions are about 9 % more likely to involve an injury than crashes in clear weather, and crashes in rain are associated with roughly 4 % higher odds of injury compared with clear conditions holding all else constant. For the interaction variable, for non‑distracted drivers (driver\_distracted = 0) each additional year of vehicle age increases 0.8 % odds per year but decreases for distracted drivers, 0.03 % per year in injury odds, as the vehicle gets older.

**Contextualized Results**

The results tell a unique story that supports and contradicts our original hypotheses. Vehicle age behaves as theorized, with each additional year in vehicle age increasing the probability of injury; this relationship echoes Anderson and Searson’s finding that the risk of fatal crashes increase as the car gets older. However, when a crash is marked as involving a distracted driver, the extra risk linked to an older car almost disappears. That could mean two things: newer cars’ safety tech may protect drivers who aren’t paying attention, or police might fail to note distraction in the most serious crashes. In the future, it might be appropriate to include a squared vehicle age variable as once a vehicle reaches a certain age the likelihood of getting injured in a crash plateaus.

Weather has a smaller, mixed impact. Cloudy skies and rain do raise injury risk a little, just as earlier studies found, but snow does not, unlike Becker, Rust, and Ulbrich’s estimate of a big snow effect. A likely reason is that most crashes in our sample happen at slower winter speeds, so slick roads are less dangerous than on the faster rural highways examined in their work. So Hypothesis 3 holds for rain and clouds, but how much weather matters clearly depends on where and how people drive.

The great surprise is distraction. Contrary to Hypothesis 1 and to Moreira, Sachsida, and Loureiro, as well as Abouk & Adams, who document higher crash risk under phone use or smoking, our model associates recorded distraction with lower odds of injury. One possible reasoning is measurement bias. Police may record distraction only when crashes are minor and drivers are willing to admit fault, leaving serious incidents unreported. Another explanation involves the cars themselves. Newer vehicles are packed with safety features like lane‑keeping, automatic braking, and collision alerts that can rescue a driver who looks away. That idea fits our data that injury risk for drivers who are distracted nearly disappears in newer cars, suggesting these systems soften the consequences of not paying attention.

When we added the interaction term, something that wasn’t looked at in the literature review, we found something new. Newer cars’ safety features help most when the driver isn’t paying attention, showing that technology and careful driving need to work together, not one instead of the other. That said, the model still explains only about 2 % of injury differences, hinting that key factors such as driver age, seat‑belt use, alcohol, and true travel speed are missing and must be captured in future work.

Speed stands out as the clearest factor for reducing injuries. Our model shows that each one mile per hour increase in the posted limit raises injury odds by about six percent, and the same upward slope before plateauing at very high limits. This is similar to what Ashenfelter and Greenstone found when rural interstate limits jumped from 55 to 65 mph, resulting in a 35 percent increase in fatalities. Greenawalt’s state‑level analysis reaches a similar conclusion, showing that every minor increase in speed‑limit exposure increases fatalities. Together, these studies and our results reinforce a simple message that keeping speeds in check remains one of the most effective, evidence‑backed ways to reduce the risk of injuries.

Looking ahead, the next step is to pair crash reports with richer data like black‑box speed logs, phone‑usage records, and even dash‑cam footage so we can measure how fast cars were truly going and whether drivers were actually distracted at the moment of impact. With those details we could test whether lowering limits by five miles per hour on major urban roads really cuts injuries as much as the model suggests and see if advanced safety features in newer cars reduce risky behavior at high speeds. We could also break injuries into levels (no harm, minor hurt, serious hurt, fatal) so policy makers can target interventions where they matter most. Finally, repeating the analysis in rural areas and in states that recently changed their speed laws would show whether the patterns we found here hold up in different driving environments and after real‑world policy shifts.

In conclusion, our model confirms that higher speed limits and older cars make crash injuries more likely, while clouds and rain add only modest extra risk. The surprise is that crashes recorded as distracted seem less harmful, most likely because newer cars’ safety tech softens the blow or because serious distraction often goes unreported. With only a small piece of the variables explaining the risk of injuries in a crash, better data on actual driving speed, seat‑belt use, and in‑car phone activity are potential next steps. Until then, the simplest way to reduce injuries remains the most obvious, reducing the speed at which you are traveling, especially on busy urban roads, and keep upgrading the safety features built into the cars we drive.

**Appendix:**

*Figure 1*

**A close-up of a number

Description automatically generated**

*Figure 2*

***A table of numbers and symbols

Description automatically generated with medium confidence***

*Figure 3*

**

*Code for regression*

#Clear Global Environment

rm(list = ls())

# ------------------------------------------------------

# 1. Import Dataset and Clean Dataset

# ------------------------------------------------------

#load necessary libraries

library(dplyr)

library(psych) # for describe()

library(ggplot2)

library(readxl) # For reading Excel files

library(tidyverse)

library(stats)

library(Hmisc)

library(kableExtra)

library(car)

# Read the CSV file

data <- read.csv('/Users/student/Documents/Spring Babson 2025/Econometrics/Project/Crash\_Reporting\_-\_Drivers\_Data.csv', stringsAsFactors = FALSE, na.strings = c("", "N/A"))

# View the first few rows of the data

head(data)

##CLEANING DATASET

sum(is.na(data))

# Keep only the relevant columns

data\_subset <- data %>%

select(Injury.Severity, Weather, Speed.Limit, Driver.Distracted.By, Vehicle.Year)

# Remove rows with any missing (NA) values

cleaned\_data <- na.omit(data\_subset)

#Create new varaible for age of car

cleaned\_data <- cleaned\_data %>%

mutate(Vehicle.Year = as.numeric(Vehicle.Year)) %>%

filter(!is.na(Vehicle.Year), Vehicle.Year >= 1950, Vehicle.Year <= 2025) %>% # Keep only plausible years

mutate(vehicle\_age = 2025 - Vehicle.Year)

cleaned\_data$Vehicle.Year <- NULL

#Group injury.severity into 1 = injury, 0 = no injury

cleaned\_data <- cleaned\_data %>%

mutate(injury = case\_when(

Injury.Severity %in% c("No Apparent Injury", "NO APPARENT INJURY") ~ 0,

Injury.Severity %in% c("Possible Injury", "POSSIBLE INJURY",

"Suspected Minor Injury", "SUSPECTED MINOR INJURY",

"Suspected Serious Injury", "SUSPECTED SERIOUS INJURY",

"Fatal Injury", "FATAL INJURY") ~ 1,

TRUE ~ NA\_real\_

))

table(cleaned\_data$injury)

sum(is.na(cleaned\_data$injury))

cleaned\_data$Injury.Severity = NULL

#clean up weather variable to group into similar groups

cleaned\_data <- cleaned\_data %>%

filter(!Weather %in% c("Unknown", "UNKNOWN")) %>%

mutate(weather = case\_when(

Weather %in% c("Clear", "CLEAR") ~ "Clear",

Weather %in% c("Cloudy", "CLOUDY") ~ "Cloudy",

Weather %in% c("Rain", "RAINING", "Freezing Rain Or Freezing Drizzle") ~ "Rain",

Weather %in% c("Snow", "SNOW", "Blowing Snow", "BLOWING SNOW",

"SLEET", "Sleet Or Hail", "WINTRY MIX") ~ "Snow",

Weather %in% c("Fog, Smog, Smoke", "FOGGY") ~ "Fog/Smog",

Weather %in% c("Severe Crosswinds", "SEVERE WINDS") ~ "Wind",

Weather %in% c("BLOWING SAND, SOIL, DIRT") ~ "Sand/Dirt",

Weather %in% c("OTHER") ~ "Other",

TRUE ~ NA\_character\_

))

table(cleaned\_data$weather)

sum(is.na(cleaned\_data$weather))

cleaned\_data$Weather= NULL

#clean up driver.distracted.by variable to group into similar groups

cleaned\_data <- cleaned\_data %>%

filter(!Driver.Distracted.By %in% c("Unknown", "UNKNOWN")) %>%

mutate(driver\_distracted = case\_when(

Driver.Distracted.By %in% c("Not Distracted", "NOT DISTRACTED") ~ 0,

TRUE ~ 1

))

table(cleaned\_data$driver\_distracted)

sum(is.na(cleaned\_data$driver\_distracted))

cleaned\_data$Driver.Distracted.By= NULL

#convert variables into correct types

cleaned\_data$driver\_distracted = as.factor(cleaned\_data$driver\_distracted)

cleaned\_data$injury = as.factor(cleaned\_data$injury)

cleaned\_data$weather = as.factor(cleaned\_data$weather)

cleaned\_data$Speed.Limit = as.numeric(cleaned\_data$Speed.Limit)

#remove unecessary datasets

data = NULL

data\_subset = NULL

# ------------------------------------------------------

# 2. Appendix 1: Mean, STDV, n, histograms, scatter plots

# ------------------------------------------------------

#Mean/Mode, STDV, and n for each observation

# Get mode for each factor variable

injury\_mode <- cleaned\_data %>%

count(injury) %>%

slice\_max(n, n = 1) %>%

pull(injury)

weather\_mode <- cleaned\_data %>%

count(weather) %>%

slice\_max(n, n = 1) %>%

pull(weather)

driver\_mode <- cleaned\_data %>%

count(driver\_distracted) %>%

slice\_max(n, n = 1) %>%

pull(driver\_distracted)

# Summary table using this improved mode method

summary\_table <- data.frame(

Variable = c("Speed.Limit", "vehicle\_age", "injury", "weather", "driver\_distracted"),

Mean\_or\_Mode = c(

round(mean(cleaned\_data$Speed.Limit, na.rm = TRUE), 2),

round(mean(cleaned\_data$vehicle\_age, na.rm = TRUE), 2),

as.character(injury\_mode),

as.character(weather\_mode),

as.character(driver\_mode)

),

SD = c(

round(sd(cleaned\_data$Speed.Limit, na.rm = TRUE), 2),

round(sd(cleaned\_data$vehicle\_age, na.rm = TRUE), 2),

NA, NA, NA

),

N = c(

sum(!is.na(cleaned\_data$Speed.Limit)),

sum(!is.na(cleaned\_data$vehicle\_age)),

sum(!is.na(cleaned\_data$injury)),

sum(!is.na(cleaned\_data$weather)),

sum(!is.na(cleaned\_data$driver\_distracted))

)

)

print(summary\_table)

#HISTOGRAMS

# Histogram for Speed.Limit

ggplot(cleaned\_data, aes(x = Speed.Limit)) +

geom\_histogram(binwidth = 5, fill = "#2c7fb8", color = "white") +

labs(title = "Distribution of Speed Limit", x = "Speed Limit (mph)", y = "Frequency") +

theme\_minimal()

# Histogram for Vehicle Age

ggplot(cleaned\_data, aes(x = vehicle\_age)) +

geom\_histogram(binwidth = 1, fill = "#f03b20", color = "white") +

labs(title = "Distribution of Vehicle Age", x = "Vehicle Age (years)", y = "Frequency") +

theme\_minimal()

# Bar plot for Injury

ggplot(cleaned\_data, aes(x = injury)) +

geom\_bar(fill = "#1b9e77") +

labs(title = "Injury Count", x = "Injury (1 = injury, 0 = no injury)", y = "Count") +

theme\_minimal()

# Bar plot for Weather

ggplot(cleaned\_data, aes(x = weather)) +

geom\_bar(fill = "#d95f02") +

labs(title = "Weather Condition Count", x = "Weather", y = "Count") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

# Bar plot for Driver Distraction

ggplot(cleaned\_data, aes(x = driver\_distracted)) +

geom\_bar(fill = "#7570b3") +

labs(title = "Driver Distraction Count", x = "Driver Distracted (1 = Yes, 0 = No)", y = "Count") +

theme\_minimal()

# ------------------------------------------------------

# 3. Initial Regression

# ------------------------------------------------------

# Run logistic regression

logistic\_model <- glm(injury ~ Speed.Limit + vehicle\_age + weather + driver\_distracted,

data = cleaned\_data,

family = binomial(link = "logit"))

# View the summary

summary(logistic\_model)

# Display odds ratios with confidence intervals

exp(cbind(OR = coef(logistic\_model), confint(logistic\_model)))

# Compute McFadden's Pseudo R^2

log\_likelihood\_model <- logLik(logistic\_model)

log\_likelihood\_null <- logLik(glm(injury ~ 1, data = cleaned\_data, family = binomial(link = "logit")))

pseudo\_r2 <- 1 - (log\_likelihood\_model / log\_likelihood\_null)

#not like r squared in linear regression but ranges from 0 to around 1. compares it to null hypothesis or no preditors

#if there is an actual number the the model is better than if it had no predictors

cat("McFadden's Pseudo R^2:", round(as.numeric(pseudo\_r2), 4), "\n")

# Compute LR Chi-Squared and p-value

lr\_chi2 <- -2 \* (as.numeric(log\_likelihood\_null) - as.numeric(log\_likelihood\_model))

p\_value <- pchisq(lr\_chi2, df = length(coef(logistic\_model)) - 1, lower.tail = FALSE)

#tells you that if theres something in this model that is worth looking at

cat("LR Chi^2:", round(lr\_chi2, 4), "\n")

cat("Prob > chi^2:", round(p\_value, 4), "\n")

# ------------------------------------------------------

# 4. Logistic Regression with interaction and transformed

# ------------------------------------------------------

#Square speed limit variable

cleaned\_data$Speed.Squared <- cleaned\_data$Speed.Limit^2

# Run logistic regression

logistic\_model2 <- glm(injury ~ Speed.Limit + Speed.Squared + vehicle\_age + weather + driver\_distracted + driver\_distracted:vehicle\_age ,

data = cleaned\_data,

family = binomial(link = "logit"))

# View the summary

summary(logistic\_model2)

# Display odds ratios with confidence intervals

exp(cbind(OR = coef(logistic\_model2), confint(logistic\_model2)))

# Compute McFadden's Pseudo R^2

log\_likelihood\_model2 <- logLik(logistic\_model2)

log\_likelihood\_null2 <- logLik(glm(injury ~ 1, data = cleaned\_data, family = binomial(link = "logit")))

pseudo\_r2 <- 1 - (log\_likelihood\_model2 / log\_likelihood\_null2)

#not like r squared in linear regression but ranges from 0 to around 1. compares it to null hypothesis or no preditors

#if there is an actual number the the model is better than if it had no predictors

cat("McFadden's Pseudo R^2:", round(as.numeric(pseudo\_r2), 4), "\n")

# Compute LR Chi-Squared and p-value

lr\_chi2 <- -2 \* (as.numeric(log\_likelihood\_null2) - as.numeric(log\_likelihood\_model2))

p\_value <- pchisq(lr\_chi2, df = length(coef(logistic\_model2)) - 1, lower.tail = FALSE)

#tells you that if theres something in this model that is worth looking at

cat("LR Chi^2:", round(lr\_chi2, 4), "\n")

cat("Prob > chi^2:", round(p\_value, 4), "\n")

# Check for multicollinearity using VIF

vif\_values <- vif(logistic\_model2)

# Display VIF results

print(vif\_values)

**Bibliography:** APA Format

Abouk, R., & Adams, S. (2013). Texting bans and fatal accidents on roadways: Do they work? Or do drivers just react to announcements of bans? *American Economic Journal: Applied Economics, 5*(2), 179–199. <https://doi.org/10.1257/app.5.2.179>

Anderson, R. W. G., & Searson, D. J. (2015). Use of age–period–cohort models to estimate effects of vehicle age, year of crash and year of vehicle manufacture on driver injury and fatality rates. *Accident Analysis & Prevention, 75*, 202–210. <https://doi.org/10.1016/j.aap.2014.11.013>

Ashenfelter, O., & Greenstone, M. (2002). *Using mandated speed limits to measure the value of a statistical life* (IZA Discussion Paper No. 571). Institute for the Study of Labor (IZA). https://ftp.iza.org/dp571.pdf

Becker, N., Rust, H. W., & Ulbrich, U. (2022). Weather impacts on various types of road crashes: A quantitative analysis using generalized additive models. *European Transport Research Review, 14*(37). <https://doi.org/10.1186/s12544-022-00561-2>

Greenawalt, K. R. (2006). *The effect of macroeconomic conditions on traffic fatality rates across the United States* (Master’s thesis). Wright State University. <https://corescholar.libraries.wright.edu/econ_student/15>

Moreira, T., Sachsida, A., & Loureiro, P. (2004). Traffic accidents: An econometric investigation. *Economics Bulletin, 9*(3), 1–7. <https://www.researchgate.net/publication/4804067_Traffic_Accidents_An_Econometric_Investigation>