Driving Experience and Road Fatalities in the USA

RESEARCH QUESTION

Does the minimum legal drinking age in a state (X1), the percentage of young drivers in the state (X2), and mandatory community service (X3) have a significant linear impact on night fatalities in the state?

Background

While all traffic fatalities are a source of accidents, night fatalities are more concerning as there is low visibility and a higher chance of drivers being drunk. Research shows that because of lesser visibility, drivers are more likely to collide with pedestrians or cyclists, whereas, in the case of drunk driving, drivers get into more fatal collisions with other vehicles (Owens & Sivak, 1996). Regarding night driving, the driver's expertise plays an essential role which is correlated to the driver's age. Generally, an experienced driver is more likely to be better at avoiding a fatality than an amateur one (Wood, 2020). Moreover, given the significant impact of alcohol consumption on fatalities, it seems plausible for the minimum drinking age in a state to significantly impact the number of night fatalities.

Variables

In order to test the hypothesis, **three independent variables are considered** for this question,

1. The minimum drinking age in the state – continuous variable.
2. The percentage of young drivers in the state – continuous variable.
3. Mandatory community service in the state – categorical variable.

The **dependent variable for the study is night fatalities across states** from the year 1982 to the year 1988. Since the number of night fatalities is obviously going to be high for states with a bigger population, it is important to standardize the quantity. Standardization was done by dividing fatalities by the population of the state in that year.

Hypothesis

The proposed hypothesis is that the legal drinking age should have a linear inverse (negative) effect on night fatalities, as lowering the minimum age can have an adverse impact on fatalities (Asch, & Levy, 1987). Whereas the percentage of young drivers in the state should have a linear positive impact on the night fatalities as younger drivers are less experienced to prevent a collision while driving at night. Further, it appears that mandated minimum jail sentences or community service for driving under the influence (DUI) engrains fear in individuals and they avoid driving under the influence, hence service is also speculated to have a linear impact on night fatalities (Ruhm, 1996).

Data Analysis

In order to test the hypothesis, there is a full model with speculated non-zero beta values whereas the reduced model has all beta-values zero for X1, X2, and X3.

Model	$\boldsymbol{Reduced\ Model}: Y = \beta_0 + \epsilon$	$\boldsymbol{Full\ Model}: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
Hypothesis	$H_0: \beta_1 = \beta_2 = \beta_3 = 0$	$H_a: \beta_1 \neq \beta_2 \neq \beta_3 \neq 0$
DoF	$dfE\ (Reduced) = n - p$ $= 336 - 1 = 335$	$dfE\ (Full) = n - p$ $= 336 - 4 = 332$

For the multivariate linear regression, first, the diagnostic tests were run to ensure there are no assumption violations. The three main tests conducted are:

- ## Test for Constant Variance

Residual plot for residuals against the fitted value was plotted to see if there is any change in variance. Figure 1 shows that there was no major violation for variance in the data. To further ensure the findings, the Breusch-Pagan test was conducted with the null hypothesis that there is no variation in variance. The p-value for the test was 0.082, indicating that the null hypothesis cannot be rejected.

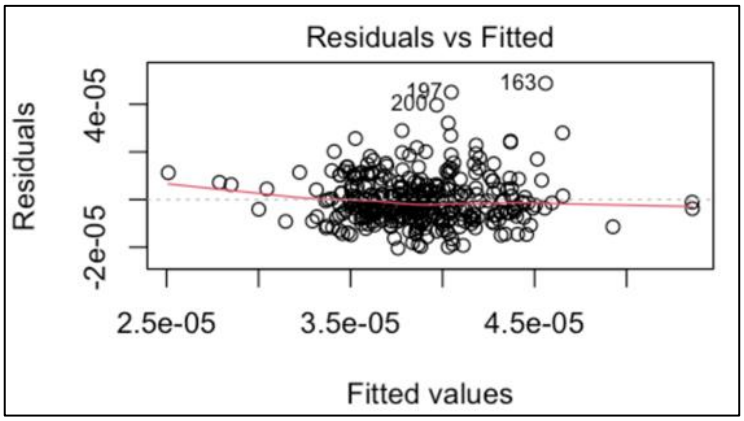

Figure 1: Residual Plot

- ## Test for Multicollinearity

Multicollinearity is a test to confirm if there is any redundancy between the different independent variables.
One way to test it is to conduct a variance inflation factor test. If the values are greater than 10, there is multicollinearity among variables. The results for the independent variables of the study turned out to be,

```
> vif
   youngdrivers        drinkage factor(service)
       1.087091        1.086938        1.004660
```

Since all the values are less than 10, it can be said that there is no collinearity between variables.

- ## Test for Normality

The last test for diagnosis was the Shapiro test to evaluate if the data is normally distributed or not. The null hypothesis for the test is that the data is normally distributed, which had to be rejected due to the low p-value. The Q-Q plot also showed the normality issue as shown in Figure 2.

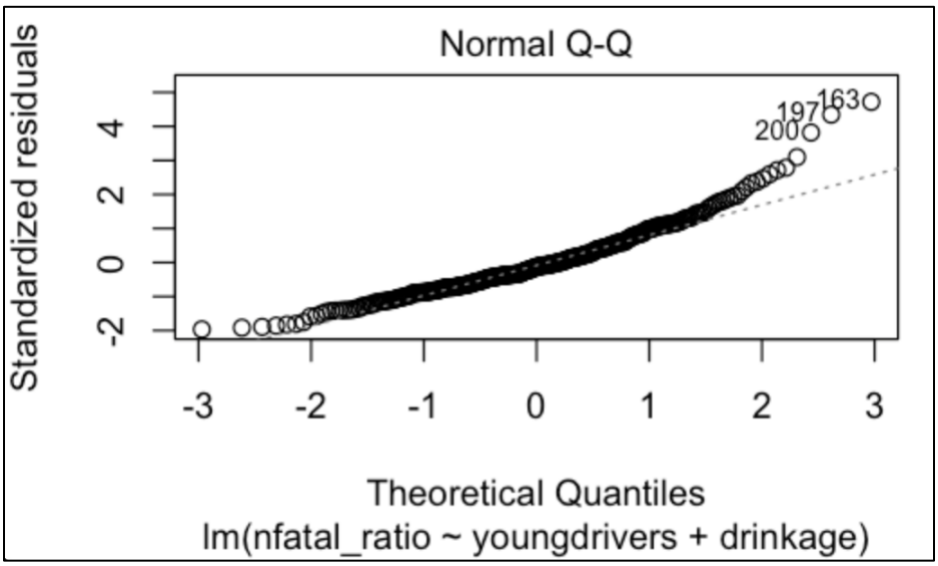

Figure 2: Test for normality.

```
        Shapiro-Wilk normality test

data:  nfatal_ratio
W = 0.93604, p-value = 7.655e-11
```

The issue was fixed by transforming Y using Box-cox transformation. Following is the model after transformation and the respective Q-Q plot.

$$\ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

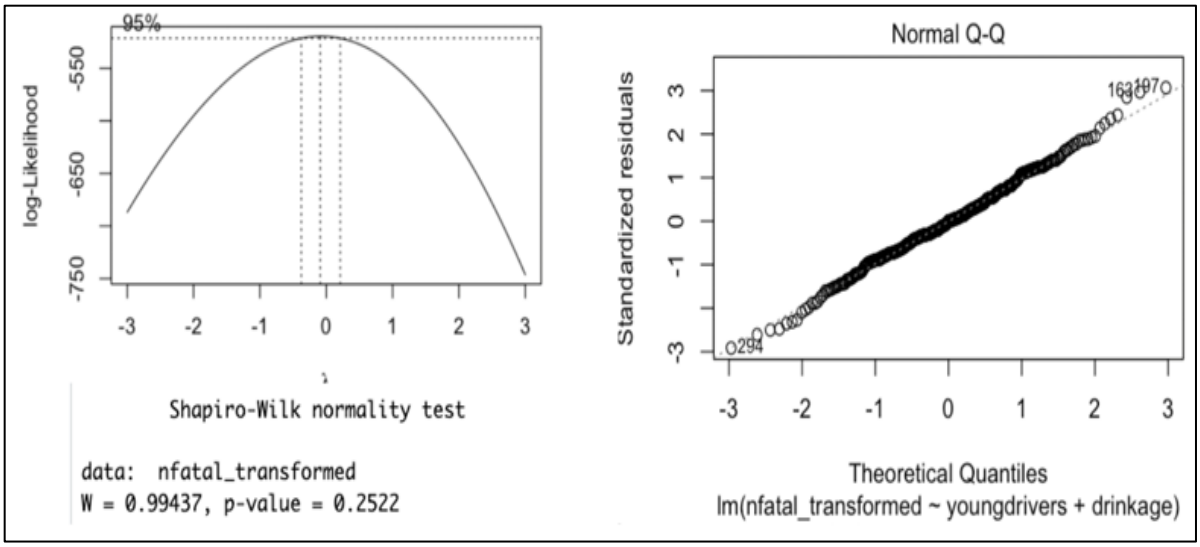

Figure 3: Transformation to fix the normality issue.

- Test for Influential points

These are the outliers that seem to have a significant impact on the model, hence it is important to identify them. Box plots and Cooks distance were used to identify these points (Figure 4), and robust regression was run to fix them (Figure 5). Later, Bootstrap was also run with robust regression to obtain confidence intervals for linear impacts. The intervals were then back-transformed to obtain reasonable results.

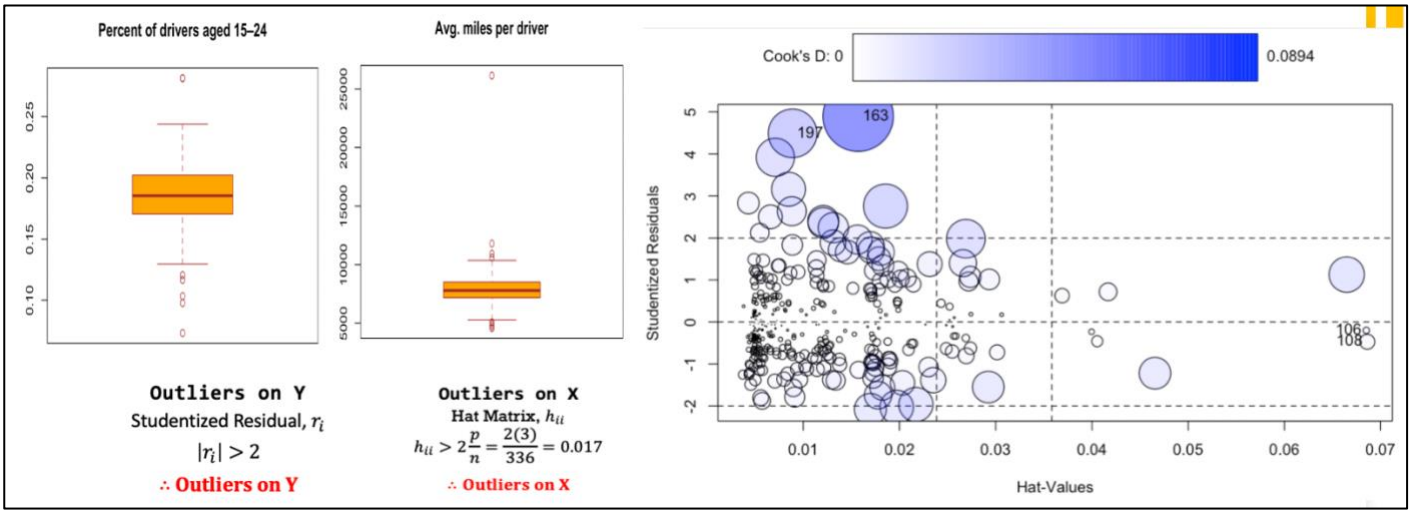

Figure 4: Testing for influential points

Result and Discussion

As for the final results, F critical and F statistics need to be calculated to answer the research question.

$$F_s = \frac{MSR}{MSE} = \frac{(24.1607 - 22.0513)/3}{(22.0513/332)}$$
$$= 10.586$$
$$F_c = (0.95,\ 3,\ 332) = 2.6318$$

Since $F_s > F_c$, we reject the null hypothesis and conclude that all the minimum drinking age, community service, and percentage of young drivers within a state have a significant linear impact on the night fatalities in a state. The results align with the literature, where these factors seem to have had a significant linear impact on fatalities.

```
> rbMod<-rlm(nfatal_transformed ~ youngdrivers+drinkage+factor(service), psi=psi.huber)
> summary(rbMod)

Call: rlm(formula = nfatal_transformed ~ youngdrivers + drinkage +
    factor(service), psi = psi.huber)
Residuals:
       Min         1Q     Median         3Q        Max
-0.7974669 -0.1655179 -0.0005653  0.1623762  0.8059682

Coefficients:
                   Value    Std. Error t value
(Intercept)         -9.8504   0.3754   -26.2367
youngdrivers         2.4047   0.5852     4.1092
drinkage            -0.0393   0.0161    -2.4322
factor(service)yes   0.0586   0.0359     1.6328

Residual standard error: 0.2434 on 331 degrees of freedom
  (1 observation deleted due to missingness)
> anova(rbMod)
Analysis of Variance Table

Response: nfatal_transformed
                Df  Sum Sq Mean Sq F value Pr(>F)
youngdrivers     1  1.5984 1.59844
drinkage         1  0.3461 0.34606
factor(service)  1  0.1649 0.16490
Residuals          22.0513
>
```

Figure 5 Results for robust regression.

References:

Asch, P., & Levy, D. T. (1987). Does the minimum drinking age affect traffic fatalities?. *Journal of Policy Analysis and Management*, *6*(2), 180-192.

Owens, D. A., & Sivak, M. (1996). Differentiation of visibility and alcohol as contributors to twilight road fatalities. *Human Factors*, *38*(4), 680-689.

Ruhm, C. J. (1996). Alcohol policies and highway vehicle fatalities. *Journal of health economics*, *15*(4), 435-454.

Wood, J. M. (2020). Nighttime driving: visual, lighting and visibility challenges. *Ophthalmic and physiological optics*, *40*(2), 187-20.

Appendix:
R-Code

```
#Loading Data From Stats Folder
Final_Data<-read.csv("/Users/fatimaperwaizkhan/Documents/STAT 512/projdata.csv", header = TRUE, sep=",")
Full_data<-read.csv("/Users/fatimaperwaizkhan/Documents/STAT 512/Fatalities.csv", header = TRUE, sep=",")

#Define Continuous Variables
fatal<-Final_Data$fatal
fatal_age<-Final_Data$fatal2124
nfatal<-Full_data$nfatal # Y VARIABLE
nfatal_ratio<-nfatal/pop
nfatal_fatal<-nfatal/fatal
drinkage<-Final_Data$drinkage
pop<-Full_data$pop
drycount<-Final_Data$dry #SIGNIFICANT
beertax<-Final_Data$beertax
unemploy<-Final_Data$unemp #SIGNIFICANT
emppopus<-Final_Data$emppopus
afatal<-Full_data$afatal
baptist<-Full_data$baptist #SIGNIFICANT
mormon<-Full_data$mormon #SIGNIFICANT
miles<-Full_data$miles #SIGNIFICANT
gsp<-Full_data$gsp
spirit<-Full_data$spirits
youngdrivers<-Full_data$youngdrivers
breath<-Full_data$breath
service<-Full_data$service

fatalities1<-lm(nfatal_fatal ~ youngdrivers, Full_data ) #SIGNIFICANT
summary(fatalities1)
anova(fatalities1)

fatalities2<-lm(nfatal_ratio ~ miles, Full_data ) #SIGNIFICANT
summary(fatalities2)
```

```
anova(fatalities2)

#Define Categorical Variables
breath<-Final_Data$breath
service<-Final_Data$service

#Box Plots for Cont. Variable
boxplot( miles, main = "Avg. miles per driver", col = "orange",border = "brown")
boxplot(youngdrivers, main = "Percent of drivers aged 15–24", col = "orange",border = "brown")
boxplot(beertax, main = "Beer Tax", col = "orange",border = "brown")
boxplot(unemploy, main = "Unemployment Rate", col = "orange",border = "brown")

#SLR for individual variables
mod_drinkage <- lm(nfatal_ratio ~ drinkage) #SIGNIFICANT
summary(mod_drinkage)
anova(mod_drinkage)
plot(residuals(mod_drinkage))

mod_yd<-lm(nfatal_ratio ~ youngdrivers) #SIGNIFICANT
summary(mod_yd)
anova(mod_unemploy)
plot(residuals(mod_yd))

#Scatter Plots for Cont. Variables
plot(Final_Data, pch=20, cex=1.5, col='steelblue')
pairs(~drinkage+youngdrivers , col='brown')

#MLR Model
mod_fullmodel<-lm(nfatal_ratio ~ youngdrivers+drinkage+factor(service))
summary(mod_fullmodel)
anova(mod_fullmodel)

plot(mod_fullmodel)
plot(residuals(mod_fullmodel)) #major outliers
plot(density(residuals(mod_fullmodel)))
plot(fitted(mod_fullmodel), residuals(mod_fullmodel))
residualPlots(mod_fullmodel)


#BOXCOX TRANSFORMATION TO FIX NORMALITY
library(MASS)
bcmle<-boxcox(lm(nfatal_ratio ~ youngdrivers+drinkage+factor(service)),lambda=seq(-3,3, by=0.1))
bcmle
lambda<-bcmle$x[which.max(bcmle$y)]
lambda #lambda value
nfatal_transformed<-log(nfatal_ratio)#Transformed Y
mod_fullmodel2<-lm(nfatal_transformed ~ youngdrivers+drinkage+factor(service))
plot(mod_fullmodel2) #New QQ Plot

#Anova After Fixing Normality
```

```
summary(mod_fullmodel2)
anova(mod_fullmodel2)

#Shapiro Normality Test
shapiro.test(nfatal_transformed)
shapiro.test(nfatal_ratio)

#Multicollinearity Test
library(car)
vif<-vif(lm(nfatal_transformed ~ youngdrivers+drinkage+factor(service)))
vif

library(lmtest)
library(onewaytests)
library(MASS)
library(rmarkdown)

#BP TEST for constant Variance
bptest(mod_fullmodel, data=Full_data) #with old data with normality issue
bf.test(mod_fullmodel, data=Full_data)

bptest(mod_fullmodel2, data=Full_data) #with new data without normality issue
#Variance is Constant

#WLS
wts1<-1/fitted(lm(abs(residuals(mod_fullmodel))~youngdrivers+drinkage+factor(service)))^2
full.mod2<-lm(nfatal_ratio ~ youngdrivers++drinkage+factor(service), weight=wts1)
summary(full.mod2)
anova(full.mod2)
residualPlots(full.mod2)
plot(full.mod2)

#Robust for outliers
rbMod<-rlm(nfatal_transformed ~ youngdrivers+drinkage+factor(service), psi=psi.huber)
summary(rbMod)
anova(rbMod)
confint(rbMod, 'temp', level=0.95)
confint(rbMod, 'hum', level=0.95)
confint(rbMod, 'ws', level=0.95)
residRb<-resid(rbMod)
residualPlots(rbMod)
library(car)

library(boot)
#Bootstrapping
boot.rfit <- function(data, indices, maxit=100) {
  data <- data[indices,]
  mod <- rlm(nfatal_transformed ~ youngdrivers+drinkage+factor(service), data=data,psi=psi.huber)
  return(coef(mod))
}
```

```
bodyfat_model_bootfit <- boot(data=data.frame(Full_data), statistic = boot.rfit, R=1000, maxit=100)

# view results
bodyfat_model_bootfit
# see plots
plot(bodyfat_model_bootfit, index=3)
# 95% confidence intervals
boot.ci(bodyfat_model_bootfit, type="perc", index=4)

#Cooks Distance
cooks_dist<-cooks.distance(mod_fullmodel)
ifpoint<- which(cooks_dist > 0.91)
ifpoint

#influence Point
influencePlot(lm(nfatal_ratio ~ youngdrivers+drinkage+factor(service)))

#HatMatrix Influential points
lm.influence(lm(Full_datanew$nfatal ~ Full_datanew$youngdrivers+Full_datanew$miles))$hat

#Removing Outliers using Boxplot
for (x in c('miles','youngdrivers'))
{
  value = Full_data[,x][Full_data[,x] %in% boxplot.stats(Full_data[,x])$out]
  Full_data[,x][Full_data[,x] %in% value] = NA
}

#Checking whether the outliers in the above defined columns are replaced by NULL or not
sum(is.na(Full_data$miles))
sum(is.na(Full_data$youngdrivers))
as.data.frame(colSums(is.na(Full_data)))

#Removing the null values
library(tidyr)
Full_datanew = drop_na(Full_data)
as.data.frame(colSums(is.na(Full_data)))

mod_fullmodel <- lm(nfatal ~ youngdrivers+miles, Full_datanew)
summary(mod_fullmodel)
anova(mod_fullmodel)

#Scatter Plots for Cont. Variables
plot(nfatal_transformed ~ youngdrivers)
pairs(~youngdrivers+miles , col='brown')

#Box Plots for Cont. Variable
boxplot( Full_datanew$miles, main = "Avg. miles per driver", col = "orange",border = "brown")
boxplot(Full_datanew$youngdrivers, main = "Percent of drivers aged 15–24", col = "orange",border =
"brown")
```

```
#Adding categorical Var
mod_fullmodel3<-lm(nfatal_transformed ~
youngdrivers+drinkage+factor(breath)+youngdrivers:factor(breath)+drinkage:factor(breath))
summary(mod_fullmodel3)
anova(mod_fullmodel3)

mod_fullmodel3<-lm(nfatal_transformed ~ youngdrivers+drinkage+factor(service))
mod_fullmodel3<-lm(nfatal_transformed ~
youngdrivers+drinkage+factor(service)+youngdrivers:factor(service)+drinkage:factor(service))
summary(mod_fullmodel3)
anova(mod_fullmodel3)
```