W271 Group Lab 3

Group 3- Allison Godfrey, Frank Yungfong Tang, Ren Tu, December 12 2020

Exercise 1

```
library(car); library(dplyr); library(gplots); library(ggplot2);
library(grid); library(gridExtra); library(plm); library(Hmisc);
load("driving.Rdata")
# Display the label of all variable except the year dummy start with "d"
excludeYearDummy=(substring(desc$variable, 1, 1) != "d")
paste("There are", sum(excludeYearDummy == FALSE), "year dummy variables")
```

```
## [1] "There are 25 year dummy variables"
```

```
desc[excludeYearDummy, c(1,2)]
```

```
##        variable                                         label
## 1          year                             1980 through 2004
## 2         state            48 continental states, alphabetical
## 3          sl55                             speed limit == 55
## 4          sl65                             speed limit == 65
## 5          sl70                             speed limit == 70
## 6          sl75                             speed limit == 75
## 7        slnone                                no speed limit
## 8      seatbelt     =0 if none, =1 if primary, =2 if secondary
## 9        minage                          minimum drinking age
## 10      zerotol                            zero tolerance law
## 11          gdl                 graduated drivers license law
## 12        bac10                        blood alcohol limit .10
## 13        bac08                        blood alcohol limit .08
## 14        perse administrative license revocation (per se law)
## 15       totfat                       total traffic fatalities
## 16      nghtfat                     total nighttime fatalities
## 17      wkndfat                       total weekend fatalities
## 18    totfatpvm        total fatalities per 100 million miles
## 19   nghtfatpvm    nighttime fatalities per 100 million miles
## 20   wkndfatpvm      weekend fatalities per 100 million miles
## 21     statepop                              state population
## 22    totfatrte        total fatalities per 100,000 population
## 23   nghtfatrte    nighttime fatalities per 100,000 population
## 24   wkndfatrte      weekend accidents per 100,000 population
## 25   vehicmiles               vehicle miles traveled, billions
## 26         unem                      unemployment rate, percent
## 27    perc14_24           percent population aged 14 through 24
## 28     sl70plus                         sl70 + sl75 + slnone
## 29       sbprim                       =1 if primary seatbelt law
## 30      sbsecon                     =1 if secondary seatbelt law
## 56 vehicmilespc
```

As we can see from the above descriptions, there are many potential variables to explore in their relationships with the total fatality rate.

```
# The dependent variable totfatrte and the potential explanatory variables
varsToEDA = c("year", "state", "totfatrte", "bac08", "bac10", "perse", "sbprim",
              "sbsecon", "sl70plus", "gdl", "perc14_24", "unem", "vehicmilespc")
```

With a large number of variables, we will focus our EDA on the variables used in our analysis (13 variables out of 56). Due to page limit of this lab, we are not print out the statistics in this report but only provide the code to do so below:

```
describe(data[varsToEDA], condense=TRUE)
```

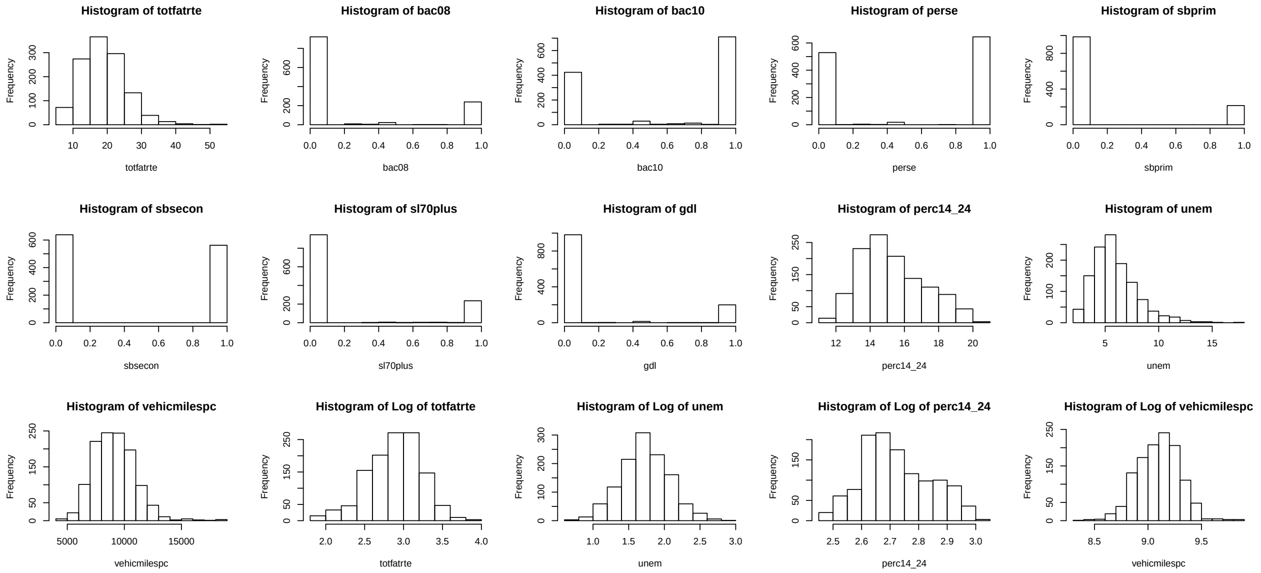
```
## data[varsToEDA]
##
##  13  Variables      1200  Observations
## ---------------------------------------------------------------------------
## year
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1200        0       25    0.998     1992    8.327     1981     1982
##      .25      .50      .75      .90      .95
##     1986     1992     1998     2002     2003
##
## lowest : 1980 1981 1982 1983 1984, highest: 2000 2001 2002 2003 2004
## ---------------------------------------------------------------------------
## state
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1200        0       48        1    27.15     16.6     4.00     6.00
##      .25      .50      .75      .90      .95
##    15.75    27.50    39.25    47.00    49.00
##
## lowest :  1  3  4  5  6, highest: 47 48 49 50 51
## ---------------------------------------------------------------------------
## totfatrte
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1200        0      916        1    18.92    7.032    9.578   11.458
##      .25      .50      .75      .90      .95
##   14.377   18.435   22.773   26.790   29.895
##
## lowest :  6.20  6.47  6.55  6.75  6.76, highest: 42.31 43.22 46.51 52.18 53.32
## ---------------------------------------------------------------------------
## bac08
##        n  missing distinct     Info     Mean      Gmd
##     1200        0        8     0.54   0.2135   0.3358
##
## lowest : 0.000 0.250 0.333 0.417 0.500, highest: 0.417 0.500 0.667 0.750 1.000
##
## Value      0.000 0.250 0.333 0.417 0.500 0.667 0.750 1.000
## Frequency    921     9     5     4    19     1     2   239
## Proportion 0.768 0.008 0.004 0.003 0.016 0.001 0.002 0.199
## ----------------------------------------------------------------------------
## bac10
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1200        0       10    0.748   0.6231   0.4691        0        0
##      .25      .50      .75      .90      .95
##        0        1        1        1        1
##
## lowest : 0.000 0.250 0.333 0.417 0.500, highest: 0.583 0.667 0.750 0.833 1.000
##
## Value      0.000 0.250 0.333 0.417 0.500 0.583 0.667 0.750 0.833 1.000
## Frequency    424     4     4     1    28     4     8    13     3   711
## Proportion 0.353 0.003 0.003 0.001 0.023 0.003 0.007 0.011 0.002 0.592
## ----------------------------------------------------------------------------
## perse
##        n  missing distinct     Info     Mean      Gmd
##     1200        0        9     0.76   0.5471   0.4958
##
## lowest : 0.000 0.083 0.167 0.250 0.333, highest: 0.333 0.417 0.500 0.750 1.000
##
## Value      0.000 0.083 0.167 0.250 0.333 0.417 0.500 0.750 1.000
## Frequency    528     1     1     4     2     2    16     1   645
## Proportion 0.440 0.001 0.001 0.003 0.002 0.002 0.013 0.001 0.538
## ----------------------------------------------------------------------------
## sbprim
##        n  missing distinct     Info      Sum     Mean      Gmd
##     1200        0        2    0.441      215   0.1792   0.2944
##
## ----------------------------------------------------------------------------
## sbsecon
##        n  missing distinct     Info      Sum     Mean      Gmd
##     1200        0        2    0.747      562   0.4683   0.4984
##
## ----------------------------------------------------------------------------
## sl70plus
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1200        0       15    0.515   0.2068   0.3283        0        0
##      .25      .50      .75      .90      .95
##        0        0        0        1        1
##
## lowest : 0.000 0.042 0.083 0.333 0.375, highest: 0.750 0.792 0.833 0.984 1.000
##
## Value      0.000 0.042 0.083 0.333 0.375 0.417 0.500 0.583 0.625 0.667
## Frequency    938     1     5     1     1     3     3     2     1     3
## Proportion 0.782 0.001 0.004 0.001 0.001 0.002 0.002 0.002 0.001 0.002
##
## Value      0.750 0.792 0.833 0.984 1.000
## Frequency      3     2     2     1   234
## Proportion 0.002 0.002 0.002 0.001 0.195
## ---------------------------------------------------------------------------
## gdl
##        n  missing distinct     Info     Mean      Gmd
##     1200        0        8    0.449   0.1741   0.2877
##
## lowest : 0.000 0.167 0.250 0.500 0.670, highest: 0.500 0.670 0.750 0.833 1.000
##
## Value      0.000 0.167 0.250 0.500 0.670 0.750 0.833 1.000
## Frequency    981     1     2    14     1     1     1   199
## Proportion 0.818 0.001 0.002 0.012 0.001 0.001 0.001 0.166
## ---------------------------------------------------------------------------
## perc14_24
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1200        0       87        1    15.33    2.116     12.6     13.2
##      .25      .50      .75      .90      .95
##     13.9     14.9     16.6     18.2     18.9
##
## lowest : 11.7 11.8 11.9 12.0 12.1, highest: 19.9 20.0 20.1 20.2 20.3
## ---------------------------------------------------------------------------
## unem
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1200        0      112        1    5.951    2.235      3.2      3.7
##      .25      .50      .75      .90      .95
##      4.5      5.6      7.0      8.6      9.9
##
## lowest :  2.2  2.3  2.4  2.5  2.6, highest: 14.2 14.4 15.0 15.5 18.0
## ---------------------------------------------------------------------------
## vehicmilespc
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1200        0     1200        1     9129     2014     6573     6968
##      .25      .50      .75      .90      .95
##     7788     9013    10327    11348    12197
##
## lowest :  4372.046  4504.285  4569.239  4735.135  4918.824
## highest: 16373.844 17440.082 18093.619 18276.135 18390.080
## ---------------------------------------------------------------------------
```

- There are 1200 records in the dataset, each has 56 variables:
- The year variable ranges from 1980-2004- 25 years across 48 states.
- The state variable range from 1 to 51, with a total 48 distinct states (some numbers in the 1-51 range are not used)
- There are a lot of variables that carry redundant information:
    - There are 25 dummy variables (d80..d99, d00..d04) for years, which is 1 if the data is for that year, which carry the same information as variable year.
    - The variable sl70plus is a transformation of the set of sl55, sl60, sl65, sl70, sl75 and slnone which is sl70+sl75+slnone.

- sbprim and sbsecon jointly carry the same information as seatbelt.
- $totfatpvm = \dfrac{totfat}{10 \times vehicmiles}$;
- $nghtfatpvm = \dfrac{nghtfat}{10 \times vehicmiles}$;
- $wkndfatpvm = \dfrac{wkndfat}{10 \times vehicmiles}$;
- $totfatrte = \dfrac{100,000 \times totfat}{statepop}$;
- $nghtfatrte = \dfrac{100,000 \times nghtfat}{statepop}$;
- $wkndfatrte = \dfrac{100,000 \times wkndfat}{statepop}$;

The histograms of the variables of interest (excludes year and state dummy variables) shows their distributions as compared to their logged value distributions.

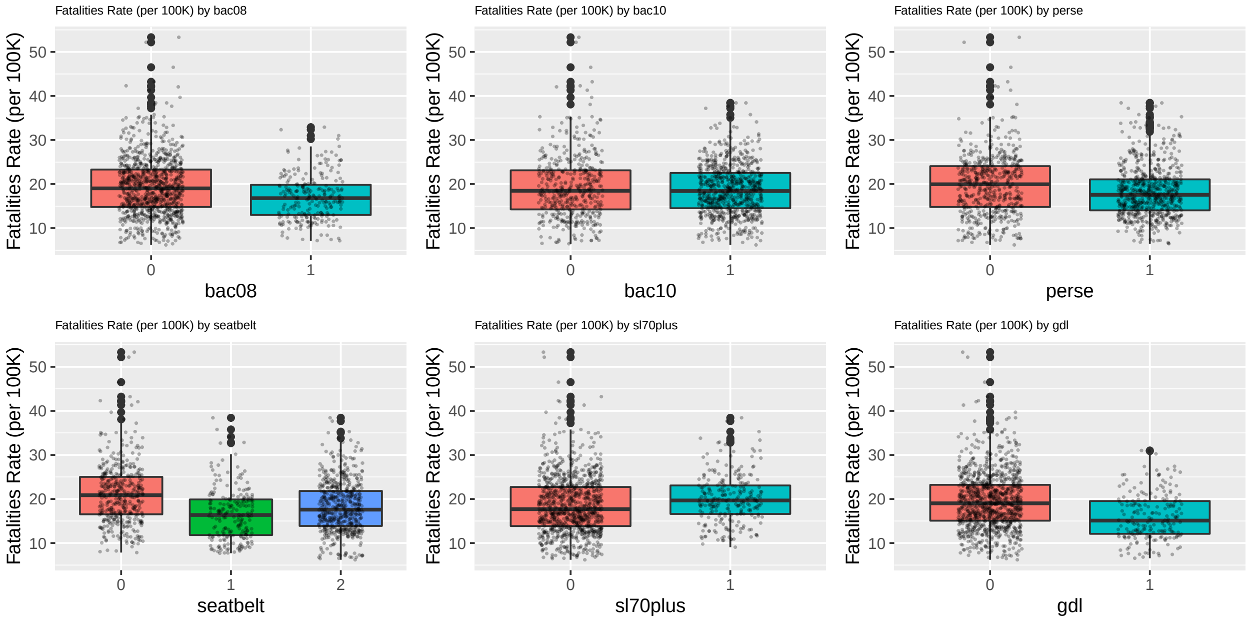
```r
par(mfrow=c(3, 5))
for (v in names(data[varsToEDA])) {
  if (!(v %in% c("year", "state"))) {
    hist(data[[v]], main=paste("Histogram of", v), xlab=v)
  }
}
for (v in c("totfatrte", "unem", "perc14_24", "vehicmilespc")) {
   hist(log(data[[v]]), main=paste("Histogram of Log of", v), xlab=v)
}
```



Many of these variables indicate the proportion of the year the law was enacted in that state during that year, and should have most of the value on either 0 or 1. These variables will not turn into normal distribution even if we apply logorithm transformation. We consider transformations of other variables that will be affected by a log transformation. For example, we see that unem, totfatrte, and perc_14-24 are pretty heavily right skewed and would benefit from a log transformation. The variable vehicmilespc is generally normal, but also slightly right skewed. We can see the histograms after their log transformations and we observe a much more normal distribution across the variables, while also noting that log(perc14_24) still does not look completely normally distributed.

```
tflab<-"Fatalities Rate (per 100K)"
cplot = function(data, plotvar, condvar, xlab, alpha=0.3) {
  g <- ggplot(data, aes(as.factor(condvar), plotvar), alpha=alpha)
  g + geom_boxplot(aes(fill=as.factor(condvar))) + xlab(xlab) + ylab(tflab) +
    geom_jitter(width = 0.2, size=0.3, alpha=alpha) +
    theme(legend.position = "none", plot.title = element_text(size=6.5)) +
    ggtitle(paste(tflab,"by",xlab))
}
cplot(data,data$totfatrte,round(data$bac08,0), xlab="bac08")->p1
cplot(data,data$totfatrte,round(data$bac10,0), xlab="bac10")->p2
cplot(data,data$totfatrte,round(data$perse,0), xlab="perse")->p3
cplot(data,data$totfatrte,data$seatbelt,  xlab="seatbelt")->p4
cplot(data,data$totfatrte,round(data$sl70plus,0),xlab="sl70plus")->p5
cplot(data,data$totfatrte,round(data$gdl,0),xlab="gdl")->p6

grid.arrange(p1, p2, p3, p4, p5, p6, ncol =3, nrow=2)
```



From the above plots, there seems to be a significant difference in the fatality rate with the blood alcohol limit being .08 (bac08). In addition, the fatality rate is lower when there is at least a primary seatbelt fastened and when the speed limit is below 70. Lastly, the total fatality rate is slightly lower for years where the graduated drivers license law is enacted than for years when it is not enacted.

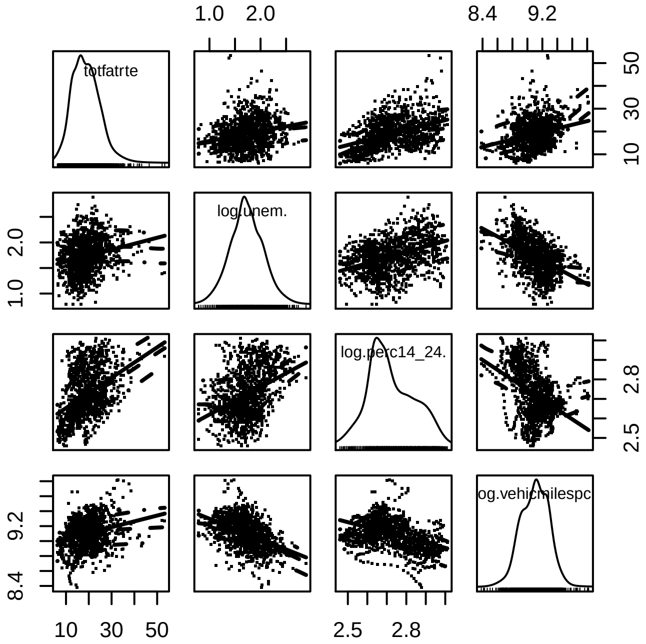
The following scatter plot matrix and the corresponding correlations table shows possible correlations between variables:

```
cor(cbind(totfatrte=data$totfatrte, year=data$year, logunem=log(data$unem),
      logperc14_24=log(data$perc14_24),logvehicmilespc=log(data$vehicmilespc)))
```

```
##                  totfatrte       year    logunem logperc14_24
## totfatrte        1.0000000 -0.3131999  0.2235160    0.4123963
## year            -0.3131999  1.0000000 -0.5181999   -0.7249677
## logunem          0.2235160 -0.5181999  1.0000000    0.3921816
## logperc14_24     0.4123963 -0.7249677  0.3921816    1.0000000
## logvehicmilespc  0.2499636  0.6739773 -0.4609198   -0.4172102
##                 logvehicmilespc
## totfatrte             0.2499636
## year                  0.6739773
## logunem              -0.4609198
## logperc14_24         -0.4172102
## logvehicmilespc       1.0000000
```

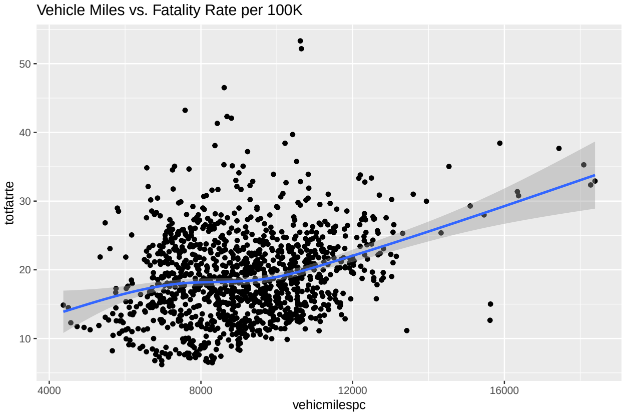
There is a relatively strong negative relationship between the percent of the population between 14 and 24 and the year, indicating that this subset of the population is decreasing over time. In addition, unemployment as well as total fatalities per 100,000 people are decreasing over time, while the vehicle miles are increasing over time. There is a relatively strong negative relationship between the vehicle miles and the unemployment rate. This makes intuitive sense because as unemployment rises, car ownership may generally decline or people can no longer afford gas and therefore drive fewer miles.

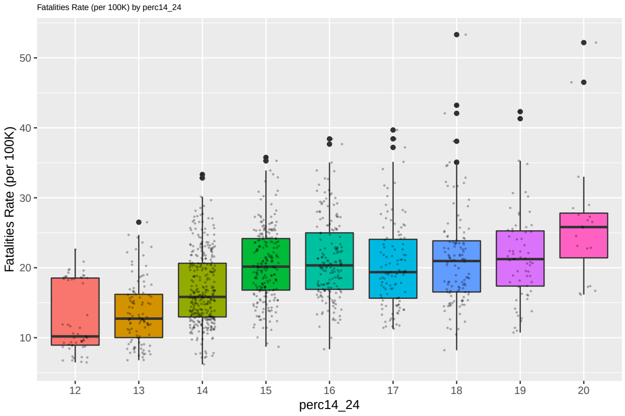
```
scatterplotMatrix(~ totfatrte +  log(unem) + log(perc14_24) + log(vehicmilespc),
                  data=data, col="black", pch=".")
```



Let's look specifically at the relationship between vehicle miles per capita and total fatality rate as well as the relationship between perc_14_24 and total fatality rate:

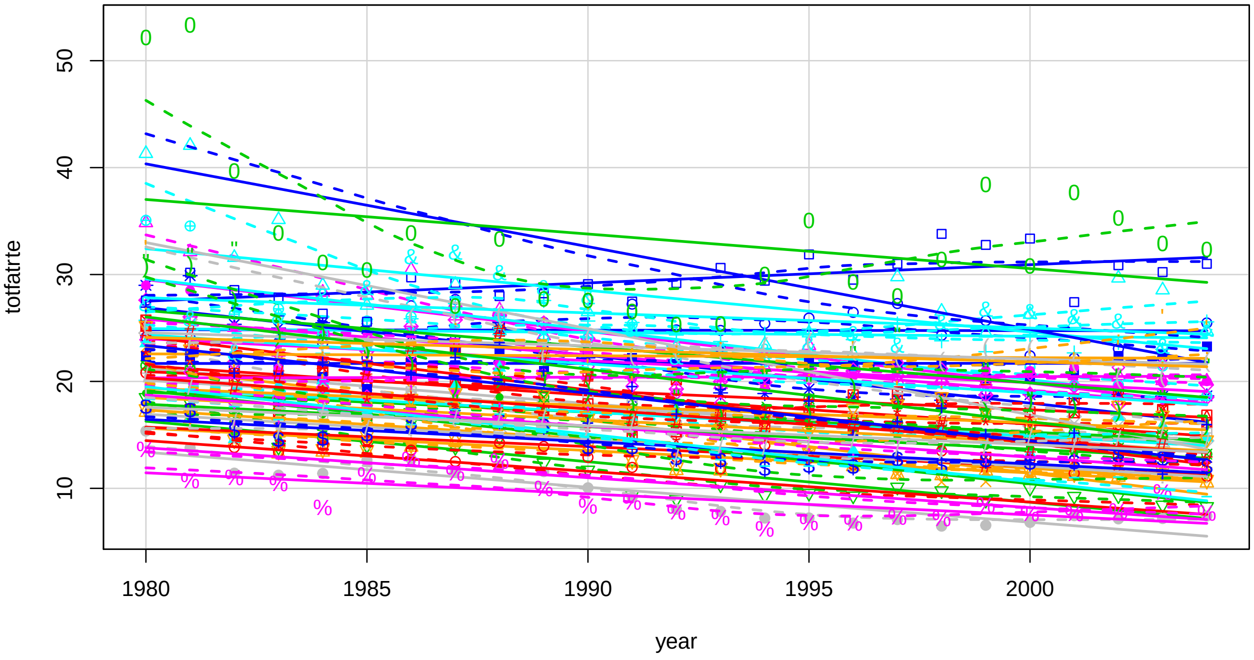
```
ggplot(data, aes(x=vehicmilespc, y=totfatrte)) + geom_point() +
  geom_smooth()+ggtitle("Vehicle Miles vs. Fatality Rate per 100K") -> p1
cplot(data,data$totfatrte,round(data$perc14_24,0), xlab="perc14_24") ->p2
grid.arrange(p1, p2, ncol =2, nrow=1)
```





From the above plot on the left, we can see there is a relatively strong positive correlation between the vehicle miles per capita and the total fatality rate. In addition, looking at the above plot on the right, we can see that greater populations of people of age 14-24 is associated with an increase in total fatality rate. Let's look at the total fatality rate of each state over time:

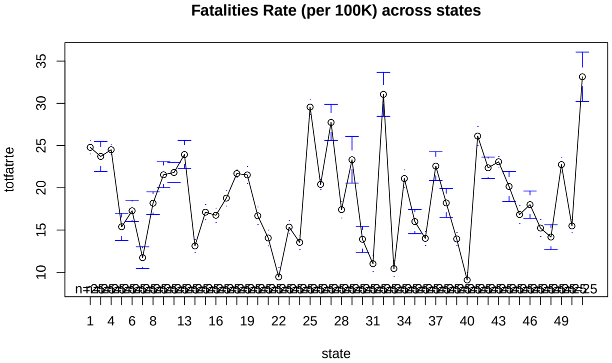
```r
scatterplot(totfatrte~year|state, boxplots=FALSE, smooth=T, data=data,legend=F)
```

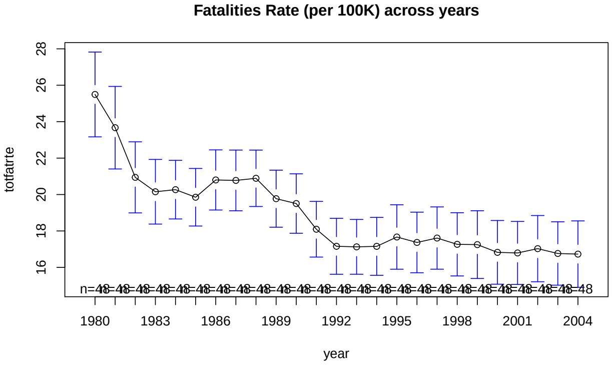


In general, the total fatalities in most states trend downwards over time, but we can see a few anomalous states exhibiting an upward trend over time.

Let's now look at the total fatality rate across states and across years.

```r
par(mfrow=c(1, 2))
# Heterogeineity across state
plotmeans(totfatrte ~ state, main=paste(tflab,"across states"),data=data)
plotmeans(totfatrte ~ year, main=paste(tflab, "across years"),data=data)
```



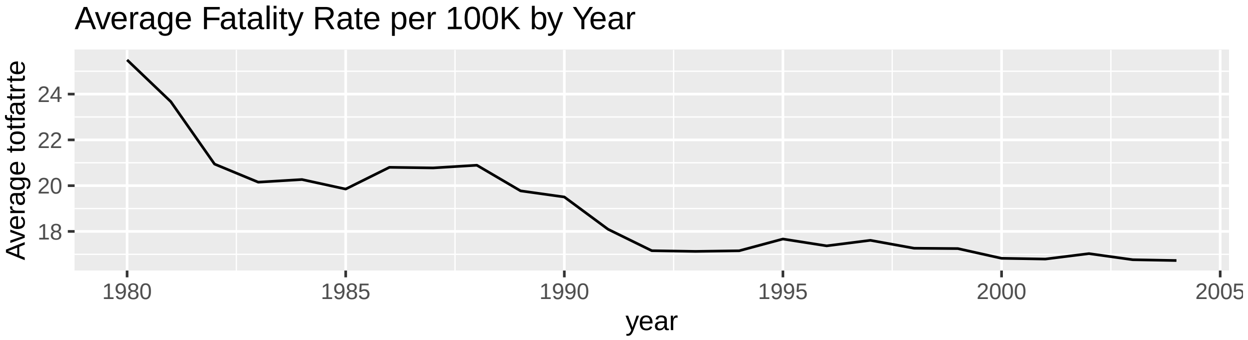


As we can see in the graph above, total fatalities (per 100,000 people) is steadily declining from 1980-2004, but at a decreasing rate in absolute value. The total fatality rates have a large variance when measured across different states.

Exercise 2 Linear Model With Year Dummies

We note that using plm with year and model="pooling" is the same as using these dummy variables. Our dependent variable of interest $totfatrte$ is total fatalities per 100,000 population in a state in a given year. This is equivalent to: $totfatrte_{s,y} = \frac{10,000 \times totfat_{s,y}}{statepop_{s,y}}$ We can calculate the average of $totfatrte$ for each year in the dataset and look at it in a tabular and graphical form.

```
### Average Fatality Plot
mean_fatality<-data %>% group_by(year)%>%summarise(m_totfatrte=mean(totfatrte))
ggplot(mean_fatality, aes(x=year, y=m_totfatrte)) + ylab("Average totfatrte") +
  geom_line() + ggtitle("Average Fatality Rate per 100K by Year")
```



```
avg_totfatrte <- data.frame(data) %>% group_by(year) %>%
  summarise(totfatrte=round(mean(totfatrte), 2), .groups = 'drop')
at <- avg_totfatrte
knitr::kable(list(at[1:5,], at[6:10,], at[11:15,],at[16:20,], at[21:25,]),
  caption="Average Fatality Rate per 100K by Year", booktabs = TRUE)
```

Table 1 and the graph above show that average fatalities are generally declining year over year from 1980 to 2004.

Below, we estimate a linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004. We note that using plm with year and model="pooling" is the same as using

Table 1: Average Fatality Rate per 100K by Year

year	totfatrte	year	totfatrte	year	totfatrte	year	totfatrte	year	totfatrte
1980	25.49	1985	19.85	1990	19.51	1995	17.67	2000	16.83
1981	23.67	1986	20.80	1991	18.09	1996	17.37	2001	16.79
1982	20.94	1987	20.77	1992	17.16	1997	17.61	2002	17.03
1983	20.15	1988	20.89	1993	17.13	1998	17.27	2003	16.76
1984	20.27	1989	19.77	1994	17.16	1999	17.25	2004	16.73

these dummy variables.

```r
summary(lm <- plm(totfatrte ~ year, data=data, model="pooling") )
```

```
## Pooling Model
##
## Call:
## plm(formula = totfatrte ~ year, data = data, model = "pooling")
##
## Balanced Panel: n = 25, T = 48, N = 1200
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -12.93021  -4.34682  -0.73052   3.74875  29.64979
##
## Coefficients:
##             Estimate Std. Error t-value  Pr(>|t|)
## (Intercept) 25.49458    0.86712 29.4015 < 2.2e-16 ***
## year1981    -1.82438    1.22629 -1.4877 0.1370936
## year1982    -4.55208    1.22629 -3.7121 0.0002152 ***
## year1983    -5.34167    1.22629 -4.3560 1.440e-05 ***
## year1984    -5.22708    1.22629 -4.2625 2.183e-05 ***
## year1985    -5.64313    1.22629 -4.6018 4.644e-06 ***
## year1986    -4.69417    1.22629 -3.8279 0.0001360 ***
## year1987    -4.71979    1.22629 -3.8488 0.0001251 ***
## year1988    -4.60292    1.22629 -3.7535 0.0001829 ***
## year1989    -5.72229    1.22629 -4.6663 3.418e-06 ***
## year1990    -5.98938    1.22629 -4.8841 1.182e-06 ***
## year1991    -7.39979    1.22629 -6.0343 2.137e-09 ***
## year1992    -8.33667    1.22629 -6.7983 1.681e-11 ***
## year1993    -8.36688    1.22629 -6.8229 1.425e-11 ***
## year1994    -8.33938    1.22629 -6.8005 1.656e-11 ***
## year1995    -7.82604    1.22629 -6.3819 2.512e-10 ***
## year1996    -8.12521    1.22629 -6.6258 5.246e-11 ***
## year1997    -7.88396    1.22629 -6.4291 1.863e-10 ***
## year1998    -8.22917    1.22629 -6.7106 3.007e-11 ***
## year1999    -8.24417    1.22629 -6.7228 2.774e-11 ***
## year2000    -8.66896    1.22629 -7.0692 2.666e-12 ***
## year2001    -8.70188    1.22629 -7.0961 2.214e-12 ***
## year2002    -8.46500    1.22629 -6.9029 8.316e-12 ***
## year2003    -8.73104    1.22629 -7.1199 1.877e-12 ***
## year2004    -8.76563    1.22629 -7.1481 1.542e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    48612
## Residual Sum of Squares: 42407
## R-Squared:      0.12765
## Adj. R-Squared: 0.10983
## F-statistic: 7.16387 on 24 and 1175 DF, p-value: < 2.22e-16
```

The above model can be represented as:

$$
\begin{aligned}
totfatrte =& \beta_0 + \beta_{81} d_{81} + \beta_{82} d_{82} + ... + \beta_{04} d_{04} \\
=& 25.49 - 1.82 d_{81} - 4.55 d_{82} - 5.34 d_{83} - 5.23 d_{84} - 5.64 d_{85} - 4.69 d_{86} - 4.72 d_{87} - 4.60 d_{88} \\
& - 5.72 d_{89} - 5.99 d_{90} - 7.40 d_{91} - 8.34 d_{92} - 8.37 d_{93} - 8.34 d_{94} - 7.83 d_{95} - 8.13 d_{96} \\
& - 7.88 d_{97} - 8.23 d_{98} - 8.24 d_{99} - 8.67 d_{00} - 8.70 d_{01} - 8.47 d_{02} - 8.73 d_{03} - 8.77 d_{04}
\end{aligned}
$$

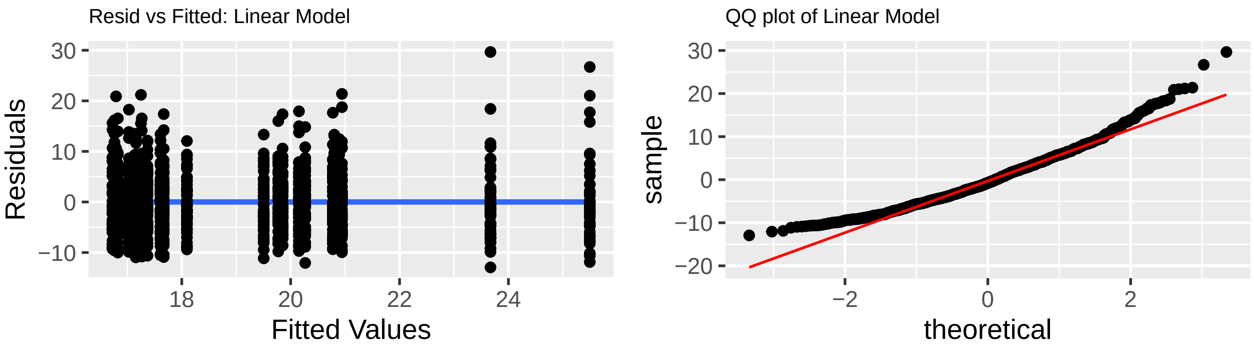
In the model above, the year 1980 (d80) is our base level to which all other years' values are compared. The intercept value indicates that in 1980, the total fatality rate is 25.49 per 100,000 population. The coefficients on each year's dummy variable shows us the difference in fatality rates in that year as compared to 1980. For example, the estimated coefficient of d04 is -8.7656, which means the estimated total fatality rate in 2004 is 8.77 person less than the fatality rate of 1980 per 100,000 population. The p-values for every dummy variable, except for that of 1981, indicate statistical significance of the coefficient. For 1981, we do not have strong evidence to indicate there are any differences from 1980, which is partly expected since 1981 is the closest year to 1980 and we would thus expect a smaller change in fatality rates.

We note that there seems to be certain years of "shocks" to the total fatality rate, specifically in 1982 and 1991. We see the coefficients on those particular years jump a lot from their previous year and then stay pretty stable until the next "shock." If we use the total fatality rate as a measure of driving safety, this shows that driving between 1982-1990 is indeed safer than before 1982, and that driving after 1991 is safer than before 1991.

Let's now plot the residuals versus fitted values and the QQ plot to test the normality and homoskedasticity of our residuals in the linear model.

```
plotplm <- function(plm, name) {
  plmdata <-data.frame(cbind(year=plm$model$year, fitted=fitted(plm),
                             residuals=residuals(plm)))
  ggplot(aes(x=(fitted), y=residuals), data=plmdata)+geom_smooth()+geom_point()+
  ggtitle(paste("Resid vs Fitted:",name))+labs(x="Fitted Values",y="Residuals")+
  theme(plot.title=element_text(size=8)) -> p1
  ggplot(plmdata, aes(sample=residuals))+stat_qq() + geom_qq_line(color="red")+
  theme(plot.title=element_text(size=8))+ggtitle(paste("QQ plot of", name))-> p2
  grid.arrange(p1, p2, ncol = 2, nrow=1)
```

```r
}
plotplm(lm, "Linear Model")
```



The QQ plot shows evidence that the residuals do not assume a purely normal distribution. The residuals do, however, have a mean of 0 and are generally homoskedastic, as seen in the left plot above.

Exercise 3 Extended Model

From our EDA, we chose to log transform perc14_24, unem, and vehicmilespc due to the skewness. We know that applying log transformations on them will make their distribution more normal. Other variables are basically indicator variables and taking log transformation will not make them distribute normally. Applying a log transformation to totfatrte will also make it distribute closer to normal, but it will also make the model much harder to interpret. Therefore, for interpretability reasons, we chose to not transform this variable.

```r
summary(expanded_OLS <-plm(totfatrte~bac08+bac10+perse+sbprim+sbsecon+sl70plus+
        gdl+log(perc14_24)+log(unem)+log(vehicmilespc), data=data,
        index=c("state", "year"), model="pooling"))
```

```
## Pooling Model
##
## Call:
## plm(formula = totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon +
##     sl70plus + gdl + log(perc14_24) + log(unem) + log(vehicmilespc),
##     data = data, model = "pooling", index = c("state", "year"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -12.39918  -2.80339  -0.41472   2.45618  24.10095
##
## Coefficients:
##                      Estimate Std. Error  t-value  Pr(>|t|)
## (Intercept)        -250.87046    9.42696 -26.6120 < 2.2e-16 ***
## bac08                -4.12030    0.54312  -7.5864 6.611e-14 ***
## bac10              -2.02414    0.40443  -5.0049 6.432e-07 ***
## perse              -0.85378    0.31789  -2.6858 0.0073372 **
## sbprim             -2.07684    0.47102  -4.4092 1.131e-05 ***
## sbsecon            -2.00195    0.39520  -5.0657 4.715e-07 ***
## sl70plus            1.41422    0.40394   3.5011 0.0004806 ***
## gdl                -2.74916    0.41198  -6.6731 3.831e-11 ***
## log(perc14_24)     19.47303    1.49309  13.0421 < 2.2e-16 ***
## log(unem)           5.02817    0.44598  11.2745 < 2.2e-16 ***
## log(vehicmilespc)  23.31882    0.88461  26.3607 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    48612
## Residual Sum of Squares: 22835
## R-Squared:      0.53027
## Adj. R-Squared: 0.52632
## F-statistic: 134.224 on 10 and 1189 DF, p-value: < 2.22e-16
```

The model can also be interpreted as:

$$\begin{aligned} totfatrte =&\beta_0 + \beta_{d81}d_{81} + \beta_{d82}d_{82} + .. + \beta_{d04}d_{04} \\ &+ \beta_{bac08}bac08 + \beta_{bac10}bac10 + \beta_{perse}perse + \beta_{sbprim}sbprim + \beta_{sbsecon}sbsecon \\ &+ \beta_{sl70plus}sl70plus + \beta_{gdl}gdl + +\beta_{perc14_24}log(perc14_24) \\ &+ \beta_{unem}log(unem) + \beta_{vehicmilespc}log(vehicmilespc) \end{aligned}$$

$$\begin{aligned} totfatrte =& -242.32 - 2.14d_{81} - 6.50d_{82} - 7.4d_{83} - 6.26d_{84} - 7.03d_{85} - 6.40d_{86} \\ &- 7.02d_{87} - 7.21d_{88} - 8.81d_{89} - 9.84d_{90} - 12.02d_{91} - 13.84d_{92} - 13.67d_{93} \\ &- 13.15d_{94} - 12.58d_{95} - 14.47d_{96} - 14.60d_{97} - 15.16d_{98} - 14.97d_{99} \\ &- 15.12d_{00} - 16.05d_{01} - 16.78d_{02} - 17.14d_{03} - 16.60d_{04} \\ &- 2.55bac08 - 1.31bac10 - 0.66perse - 0.072sbprim + 0.026sbsecon \\ &+ 3.32sl70plus - 0.89gdl + 2.99log(perc14_24) + 5.35log(unem) + 28.14log(vehicmilespc) \end{aligned}$$

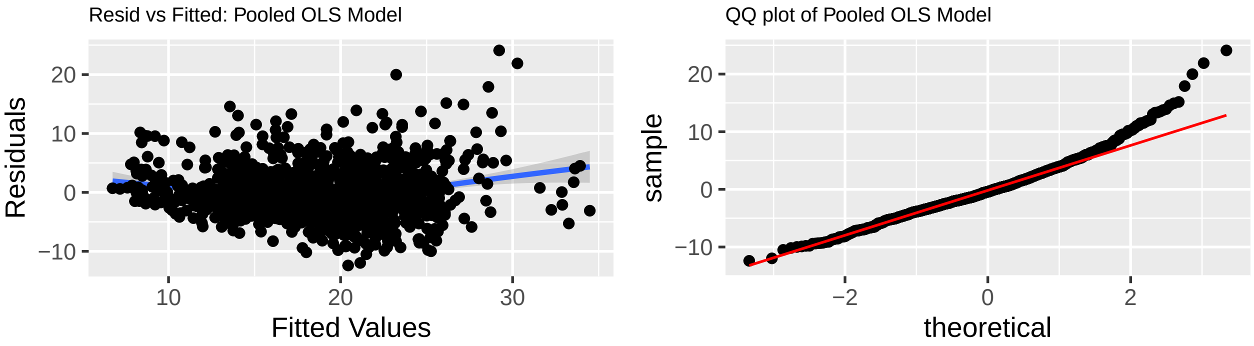
From the above results, we can see that the coefficients on bac08, bac10, perse, sl70plus, log(unem), and log(vehicmilespc) are all statistically significant (p value < 0.05), while the coefficients on sbprim, sbsecon, gdl, and log(perc14_24) are not statistically significant in this model (p value > 0.05).

BAC08 and BAC10: BAC stands for Blood Alcohol Content. BAC10 indicates the proportion of the year that the state law considers the BAC limit for a DUI as 0.10 or higher. BAC08 indicates the proportion of the year that the state law considers the BAC limit for a DUI as 0.08 or higher. In our model, the estimated coefficient for bac08 is -2.55, meaning that the state law considering the BAC limit as 0.08 for the whole year is associated with the decreased fatality rate of 2.55 persons per 100,000 compared to when no BAC law is enacted, keeping all other variables fixed in the model. The estimated coefficient for bac10 is -1.31, meaning that the state law considering the BAC limit as 0.10 for the whole year is associated with the decreased fatality rate of 1.31 persons persons per 100,000 compared to when no BAC law is enacted, keeping all other variables fixed. Intuitively, it makes sense that enacting a harsher law of making the BAC limit equal to 0.08 would be associated with fewer fatalities because, theoretically, not as many people with high BAC levels would be driving legally, and therefore discourage them to drive. Both coefficients are statistically significant at the 95% level.

Per Se Laws: The coefficient on perse is statistically significant at the 95% confidence level, but is relatively low in absolute value. The model indicates that the enactment of per se laws (versus no per se laws) in a given year is associated with a decreased fatality rate of 0.66 persons per 100,000, keeping all other variables fixed.

Primary Seat Belt Law: The coefficients on sbprim and sbecon are both not statistically significant (p values > 0.05), indicating that we do not have enough evidence to claim an effect difference than 0.

```r
plotplm(expanded_OLS, "Pooled OLS Model")
```



As we can see in the plots above, the residuals of this pooled OLS model assume a much more normal distribution and generally centers around a mean of 0. However, this model shows slight evidence of heteroskedasticity of the residuals, potentially violating the OLS assumption of homoskedasticity of the residuals.

Exercise 4 Fixed Effect Model

```r
summary(fe_model <- plm(totfatrte ~ bac08+bac10+perse+sbprim+sbsecon+sl70plus+
  gdl+log(perc14_24)+log(unem)+log(vehicmilespc), data=data,
  index=c("state", "year"), model="within"))
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon +
##     sl70plus + gdl + log(perc14_24) + log(unem) + log(vehicmilespc),
##     data = data, model = "within", index = c("state", "year"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -6.67258 -1.23075 -0.04362  1.13121 14.40568
```

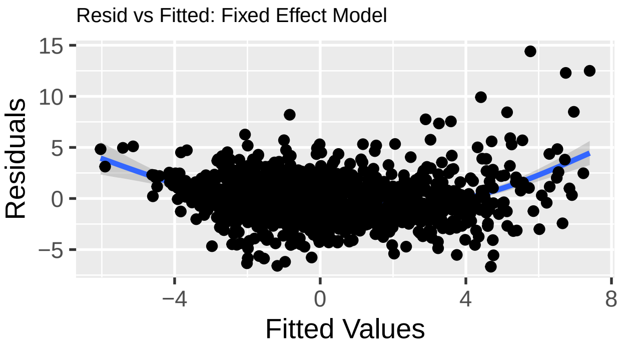
```
##
## Coefficients:
##                     Estimate Std. Error t-value  Pr(>|t|)
## bac08               -1.87013    0.38488 -4.8590 1.344e-06 ***
## bac10               -1.49165    0.26791 -5.5678 3.214e-08 ***
## perse               -1.66381    0.24817 -6.7044 3.172e-11 ***
## sbprim              -1.78404    0.34832 -5.1219 3.549e-07 ***
## sbsecon             -0.82948    0.24989 -3.3194 0.0009306 ***
## sl70plus            -1.15208    0.24250 -4.7509 2.283e-06 ***
## gdl                 -0.65260    0.23146 -2.8195 0.0048923 **
## log(perc14_24)      15.26578    1.14528 13.3293 < 2.2e-16 ***
## log(unem)           -3.18184    0.32670 -9.7392 < 2.2e-16 ***
## log(vehicmilespc)    3.89973    1.02549  3.8028 0.0001506 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12134
## Residual Sum of Squares: 5604.5
## R-Squared:      0.53812
## Adj. R-Squared: 0.51507
## F-statistic: 133.05 on 10 and 1142 DF, p-value: < 2.22e-16
```

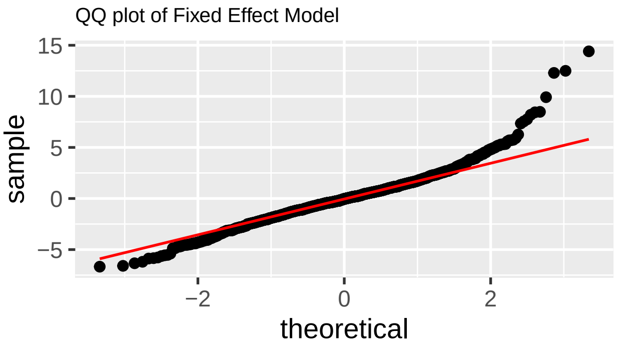
The model can also be interpreted as:

$$\begin{aligned}
(totfatrte_{it} - \overline{totfatrte_i}) =& \beta_{bac08}(bac08_{it} - \overline{bac08_i}) + \beta_{bac10}(bac10_{it} - \overline{bac10_i}) + \\
& \beta_{perse}(perse_{it} - \overline{perse_i}) + \beta_{sbprim}(sbprim_{it} - \overline{sbprim_i}) + \\
& \beta_{sbsecon}(sbsecon_{it} - \overline{sbsecon_i}) + \beta_{sl70plus}(sl70plus_{it} - \overline{sl70plus_i}) + \\
& \beta_{gdl}(gdl_{it} - \overline{gdl_i}) + \\
& \beta_{perc14\_24}(log(perc14\_24_{it}) - log(\overline{perc14\_24_i})) + \\
& \beta_{unem}(log(unem_{it}) - log(\overline{unem_i})) + \\
& \beta_{vehicmilespc}(log(vehicmilespc_{it}) - log(\overline{vehicmilespc_i})) \\
(totfatrte_{it} - \overline{totfatrte_i}) =& -1.87(bac08_{it} - \overline{bac08_i}) - 1.49(bac10_{it} - \overline{bac10_i}) \\
& -1.66(perse_{it} - \overline{perse_i}) - 1.78(sbprim_{it} - \overline{sbprim_i}) \\
& -0.83(sbsecon_{it} - \overline{sbsecon_i}) - 1.15(sl70plus_{it} - \overline{sl70plus_i}) \\
& -0.65(gdl_{it} - \overline{gdl_i}) \\
& +15.27(log(perc14\_24_{it}) - log(\overline{perc14\_24_i})) \\
& -3.18(log(unem_{it}) - log(\overline{unem_i})) \\
& +3.90(log(vehicmilespc_{it}) - log(\overline{vehicmilespc_i}))
\end{aligned}$$

Let's again examine the residuals of the Fixed Effect model:

```r
plotplm(fe_model, "Fixed Effect Model")
```





In the above plots, we can see that the residuals in the fixed effect model are not very normally distributed and are not always centered around a 0 mean. However, they do still show evidence of homoskedasticity as their variance generally stays constant across fitted values.

- The estimated coefficient for bac08 is -1.87, meaning that the state law considering the BAC limit as 0.08 for the whole year is associated with the decreased fatality rate of 1.87 persons per 100,000 compared to when no BAC law is enacted, keeping all other variables fixed in the model.
- The estimated coefficient for bac10 is -1.49, meaning that the state law considering the BAC limit as 0.10 for the whole year is associated with the decreased fatality rate of 1.49 persons persons per 100,000 compared to when no BAC law is enacted, keeping all other variables fixed.
- The model indicates that the enactment of per se laws (versus no per se laws) in a given year is associated with a decreased fatality rate of 1.66 persons per 100,000, keeping all other variables fixed.
- The model indicates that the enactment of primary seat belt law (versus no primary seat belt law) in a given year is associated with a decreased fatality rate of 1.78 persons per 100,000, keeping all other variables fixed.

Let's look at the following table to examine the differences in the coefficients in the Pooled OLS model versus the Fixed Effect model.

```
fl = c("bac08", "bac10", "perse", "sbprim")
comp <- data.frame(PooledOLS=expanded_OLS$coefficients[fl],
          POLS.p_value=(summary(expanded_OLS)$coefficients[fl,4]),
          FixedEffect=fe_model$coefficients[fl],
          FE.p_value=(summary(fe_model)$coefficients[fl,4]),
          Delta=(fe_model$coefficients[fl]-expanded_OLS$coefficients[fl]))
knitr::kable(comp)
```

	PooledOLS	POLS.p_value	FixedEffect	FE.p_value	Delta
bac08	-4.1203000	0.0000000	-1.870129	1.3e-06	2.2501711
bac10	-2.0241366	0.0000006	-1.491653	0.0e+00	0.5324837
perse	-0.8537769	0.0073372	-1.663805	0.0e+00	-0.8100282
sbprim	-2.0768387	0.0000113	-1.784036	4.0e-07	0.2928027

When using a fixed effect model, every one of the coefficients becomes statistically significant at the 95% confidence level. The coefficient on bac08 gets **very slightly** smaller in absolute value in the fixed effect model versus the pooled OLS model. By contrast, the coefficients on the other three variables (bac10, perse, and sbprim) all get substantially larger in absolute value. The p-values for all 4 estimates shrink in the fixed effect model when compared to the pooled OLS model, well past the 0.05 threshold for statistical significance, making our estimates more reliable.

Assumptions: For the fixed effects model, the following assumptions are needed:

FE1 For each i, the model is $y_{it} = \beta_1 x_{it1} + ... + \beta_k x_{itk} + a_i + u_{it}, t = 1, ..., T$, where the β_j are the parameters to estimate and ai is the unobserved effect. **FE2** We have a random sample from the cross section. **FE3** Each explanatory variable changes over time (for at least some i), and no perfect linear relationships exist among the explanatory variables. **FE4** For each t, the expected value of the idiosyncratic error given the explanatory variables in all time periods and the unobserved effect is zero: $E(u_{it}|X_i, a_i) = 0$. Under these first four assumptions—which are identical to the assumptions for the first-differencing estimator—the fixed effects estimator is unbiased. Again, the key is the strict exogeneity assumption, FE4. Under these same assumptions, the FE estimator is consistent with a fixed T as $N\infty$. **FE5** $Var(u_{it}|X_i, a_i) = Var(u_{it}) = \sigma_u^2$, for all $t = 1, ..., T$. **FE6** For all $t \neq s$, the idiosyncratic errors are uncorrelated (conditional on all explanatory variables and a_i): $Cov(u_{it}, u_{is}|X_i, a_i) = 0$). **FE7** Conditional on X_i and a_i, the u_{it} are independent and identically distributed as $Normal(0, \sigma_u^2)$.

These assumptions are reasonable in the dataset context given the multiple observations across time for each state and the likely presence of attributes that stays the same over time within each state. Perhaps most importantly, the Fixed Effect model allows time invariant aspects of each individual state to be correlated with the explanatory variables (since they are canceled out of the model anyways).

By contrast, in a pooled OLS regression, each independent variable x_{it} must be uncorrelated with the time invariant attributes of the individual states. Written more precisely, a pooled OLS model requires that $cov(x_{itj}, a_i) = 0$. In this case, this assumption is unlikely to hold since there is likely unobserved time invariant heterogeneity. For example, the individual state's political leaning could be affecting their ability to pass certain laws or the percent of people between 14 and 24 years old. Due to the many aspects of individual states that may be correlated with the explanatory variables, the Pooled OLS assumption of no time invariant heterogeneity is unlikely to hold, thus making the FE model more reliable in this situation.

Exercise 5 Random Effects vs Fixed Effect

If the assumption of no time invariant heterogeneity can be upheld, then we would rather use a Random Effect model because it would be most efficient in that case. However, if this assumption does not hold, we must use a fixed effect model. First, let's set up our Random Effects model:

```
re_model <- plm(totfatrte ~ bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+
                  log(perc14_24)+log(unem)+log(vehicmilespc), data=data,
                index=c("state", "year"), model="random")
```

Let's use the Hausman test to determine whether or not we can reject the null hypothesis (that we can use a random effect model/ our heterogeneity assumption is upheld).

```
phtest(fe_model, re_model)
```

```
##
```

```
##  Hausman Test
##
## data:  totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus +  ...
## chisq = 184.77, df = 10, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

Based on the above results, we can reject the null hypothesis that the random effect model gives us reliable estimates in favor of the alternative to use a fixed effect model (p value < 0.05). This indicates that there are in fact time invariant unobserved attributes of the individual states in our model. More concisely, this indicates that $cov(x_{itj}, a_i) \neq 0$ indicating that we would prefer to use a Fixed Effect model.

Exercise 6 Impact of Increase 1,000 on vehicmilespc

Since our model uses the logarithm transformation of vehicmilespc, the estimated effect of a 1,000 mile change in vehicmilespc depends on the original value of vehicmilespc. Therefore, we create Table 3 to show the impact of the change in the observed range of vehicmilespc:

Table 3: Change in totfatrte for 1000 mile increase in vehicmilespc

Miles	Change		Miles	Change		Miles	Change		Miles	Change
4000	0.87	5	8000	0.46	9	12000	0.31	13	16000	0.24
5000	0.71	6	9000	0.41	10	13000	0.29	14	17000	0.22
6000	0.60	7	10000	0.37	11	14000	0.27	15	18000	0.21
7000	0.52	8	11000	0.34	12	15000	0.25	16	19000	0.20

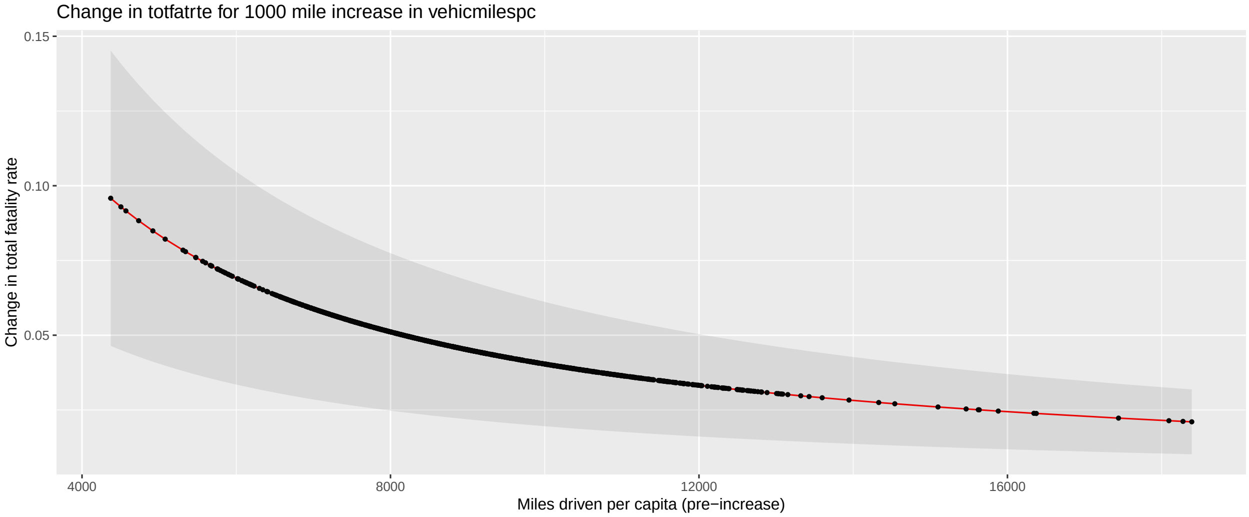
```r
start<-floor(min(data$vehicmilespc)/1000)*1000
end<-ceiling(max(data$vehicmilespc)/1000)*1000
v <- seq(start, end, 1000)
d <- rep(0, length(v))
beta <- fe_model$coefficients[["log(vehicmilespc)"]]
for (i in 1:length(v)) {
  d[i] <- round(beta*(log(v[i]+1000)-log(v[i])),2)
}
t <- data.frame(miles=v,   totfatrte=d)
knitr::kable(list(t[1:4,], t[5:8,], t[9:12,], t[13:16,]),
             col.names=c("Miles", "Change"), booktabs = TRUE,
  caption="Change in totfatrte for 1000 mile increase in vehicmilespc" )
```

Table 3 shows the changes of totfatrte if the vehicmilespc from the value next to it in the table increase 1,000. For example, increasing the value of vehicmilespc 1,000 from 5,000 will increase the value of totfatrte by 0.71. This means, in a state, if the number of miles driven per capita increases 1,000 miles from 5,000 miles to 6,000 miles, we estimated the fatality increase by 0.71 persons for every 100,000 population, or increase 7.1 persons per million population. Increasing the value of vehicmilespc by 1,000 from 13,000 will increase the value of totfatrte by 0.29 only. This means, in a state, if the number of miles driven per capita increases 1,000 from 13,000 miles to 14,000 miles, we estimate the fatality to increase by 0.29 persons for every 100,000 population, or increase 2.9 persons per million population. From Table 3, we conclude that totfatrte increases in the range between 0.20 and 0.87 for increases of 1,000 in the observed range of vehicmilespc (4,000-19,000 miles driven per capita).

Let's also look at this effect on different values of vehicmilespc in graphical form, with the upper and lower confidence intervals shown as well:

```r
# pull out beta from model
beta <- fe_model$coefficients[["log(vehicmilespc)"]]
# calculate change from base value
#change <- beta*(log(data$vehicmilespc+1000)-log(data$vehicmilespc))
change <- fe_model$coefficients[["log(vehicmilespc)"]]/100 *
  (log(data$vehicmilespc + 1000)/log(data$vehicmilespc) - 1)*100
# confidence interval
se <- sqrt(diag(vcov(fe_model)))[["log(vehicmilespc)"]]
#lower <- (beta-1.96*se) * (log(data$vehicmilespc+1000) - log(data$vehicmilespc))
#upper <- (beta+1.96*se) * (log(data$vehicmilespc+1000) - log(data$vehicmilespc))
lower <- (beta-1.96*se) * (log(data$vehicmilespc + 1000)/log(data$vehicmilespc) - 1)
```

```r
upper <- (beta+1.96*se) * (log(data$vehicmilespc + 1000)/log(data$vehicmilespc) - 1)
#create dataframe to plot
df <- data.frame(x = data$vehicmilespc, y = change, l = lower, u = upper)
#plot with confidence interval
ggplot(df, aes(x=x, y=y)) + geom_line(size=0.5, color="red")+geom_point(size=1)+
  geom_ribbon(aes(ymin=l, ymax=u), linetype=2, alpha=0.1) +
  labs(title = "Change in totfatrte for 1000 mile increase in vehicmilespc") +
  xlab("Miles driven per capita (pre-increase)") +
  ylab("Change in total fatality rate")
```



The plot above shows the effect on total fatality rate resulting from a 1000 mile increase in vehicmilespc. Since we log transformed vehicmilespc, the effect of a 1000 mile increase will be different depending on the starting value of vehicmilespc. As we can observe in the plot above, the effect of a 1000 mile increase in vehicmilespc on the total fatality rate is always positive, but becomes less positive the higher the starting value for vehicmilespc. This makes intuitive sense because if the vehicle miles driven is already very high, a 1000 mile increase will have less of an effect on the total fatality rate since 1000 miles is a very small percentage increase compared to when there is a very low level of miles driven per capita.

Exercise 7 Serial Correlation or Heteroskedasticity in the Idiosyncratic Errors

If there is heteroskedasticity in the idiosyncratic errors of our model, then the estimated variances in our model are biased and the standard errors associated with explanatory variables are likely to be understated. If there is serial correlation in the idiosyncratic errors, then the model likely has omitted variables that lead to biased estimators. The combination of biased estimators and biased standard errors could lead us to incorrectly conclude whether explanatory variables in the model are statistically significant.