Homework 1

Richard Albright
ISYE6414
Spring 2020

Part A. ANOVA

Additional Material: ANOVA tutorial

https://datascienceplus.com/one-way-anova-in-r/

Jet lag is a common problem for people traveling across multiple time zones, but people can gradually adjust to the new time zone since the exposure of the shifted light schedule to their eyes can resets the internal circadian rhythm in a process called "phase shift". Campbell and Murphy (1998) in a highly controversial study reported that the human circadian clock can also be reset by only exposing the back of the knee to light, with some hailing this as a major discovery and others challenging aspects of the experimental design. The table below is taken from a later experiment by Wright and Czeisler (2002) that re-examined the phenomenon. The new experiment measured circadian rhythm through the daily cycle of melatonin production in 22 subjects randomly assigned to one of three light treatments. Subjects were woken from sleep and for three hours were exposed to bright lights applied to the eyes only, to the knees only or to neither (control group). The effects of treatment to the circadian rhythm were measured two days later by the magnitude of phase shift (measured in hours) in each subject's daily cycle of melatonin production. A negative measurement indicates a delay in melatonin production, a predicted effect of light treatment, while a positive number indicates an advance.

Raw data of phase shift, in hours, for the circadian rhythm experiment

Treatment	Phase Shift (hr)
Control	0.53, 0.36, 0.20, -0.37, -0.60, -0.64, -0.68, -1.27
Knees	0.73, 0.31, 0.03, -0.29, -0.56, -0.96, -1.61
Eyes	-0.78, -0.86, -1.35, -1.48, -1.52, -2.04, -2.83

```
control <- c(0.53, 0.36, 0.20, -0.37, -0.60, -0.64, -0.68, -1.27)
knees <- c(0.73, 0.31, 0.03, -0.29, -0.56, -0.96, -1.61)
eyes <- c(-0.78, -0.86, -1.35, -1.48, -1.52, -2.04, -2.83)
data <- data.frame(
       Y=c(control, knees, eyes),
       Circadian.Rythm = factor(
         rep(
           c("control", "knees", "eyes"),
           times=c(length(control), length(knees), length(eyes)))
       ))
fm1 <- aov(Y~Circadian.Rythm, data=data)

xtable(anova(fm1), type='latex', comment=FALSE)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Circadian.Rythm	2	7.22	3.61	7.29	0.0045
Residuals	19	9.42	0.50		

Question A1 - 7 pts

Consider the following incomplete R output:

Fill in the missing values in the analysis of the variance table.

Source	Df	Sum of Squares	Mean Squares	F-statistics	p-value
Treatments	2	7.225	3.6122	7.2894	0.004
Error	19	9.415	0.4955		
TOTAL	21	16.64			

Question A2 - 3 pts

Use μ_1, μ_2, and μ_3 as notation for the three mean parameters and define these parameters clearly based on the context of the topic above. Find the estimates of these parameters.

```
model.tables(fm1, type = "means")
```

Tables of means Grand mean

-0.7127273

Circadian.Rythm control eyes knees -0.3087 -1.551 -0.3357 rep 8.0000 7.000 7.0000

```
mu_1 <- mean(control)
mu_2 <- mean(knees)
mu_3 <- mean(eyes)
variable <- c("Control", "Knees", "Eyes")
symbol <- c("mu_1", "mu_2", "mu_3")
mean <- c(mu_1, mu_2, mu_3)
xtable(data.frame(variable, symbol, mean))
```

	variable	symbol	mean
1	Control	mu_1	-0.31
2	Knees	mu_2	-0.34
3	Eyes	mu_3	-1.55

```
xtable(tidy(TukeyHSD(fm1)))
```

	term	comparison	estimate	conf.low	conf.high	adj.p.value
1	Circadian.Rythm	eyes-control	-1.24	-2.17	-0.32	0.01
2	Circadian.Rythm	knees-control	-0.03	-0.95	0.90	1.00
3	Circadian.Rythm	knees-eyes	1.22	0.26	2.17	0.01

“‘ ## Question A3 - 10 pts

Use the ANOVA table in Question A1 to write the:

a. **2 pts** Write the null hypothesis of the ANOVA F-test, H_0

H_0: $\mu_1 = \mu_2 = \mu_3$

b. **2 pts** Write the alternative hypothesis of the ANOVA F-test, H_A

H_A: some means are different

c. **2 pts** Fill in the blanks for the degrees of freedom of the ANOVA F-test statistic: F(_____, _____)

$F(2, 19)$

d. **2 pts** What is the p-value of the ANOVA F-test?

0.004

e. **2 pts** According the the results of the ANOVA F-test, does light treatment affect phase shift? Use an α-value of 0.05.

Yes, $0.004 < 0.05$. Light treatment affects phase shift. We reject the null hypothesis and accept the alternative hypothesis. Knees vs the control group have similar means. Eyes vs control and Eyes vs Knees are different.

Part B. Simple Linear Regression

Additional Material: Simple Linear Regression tutorial (8 modules)

http://www.r-tutor.com/elementary-statistics/simple-linear-regression

It is common knowledge that obeying the traffic signs while driving reduces the number of accidents on the road. Is the previous really true? If it is, the more signs the safer the highway? In this problem we will analyze data from 39 sections of large highways in Minnesota in 1973 to try to give answers to these questions.

The data file includes the following columns:

rate: 1973 accident rate per million vehicle miles.

sigs1: signs per mile of roadway, adjusted to have no zero values.

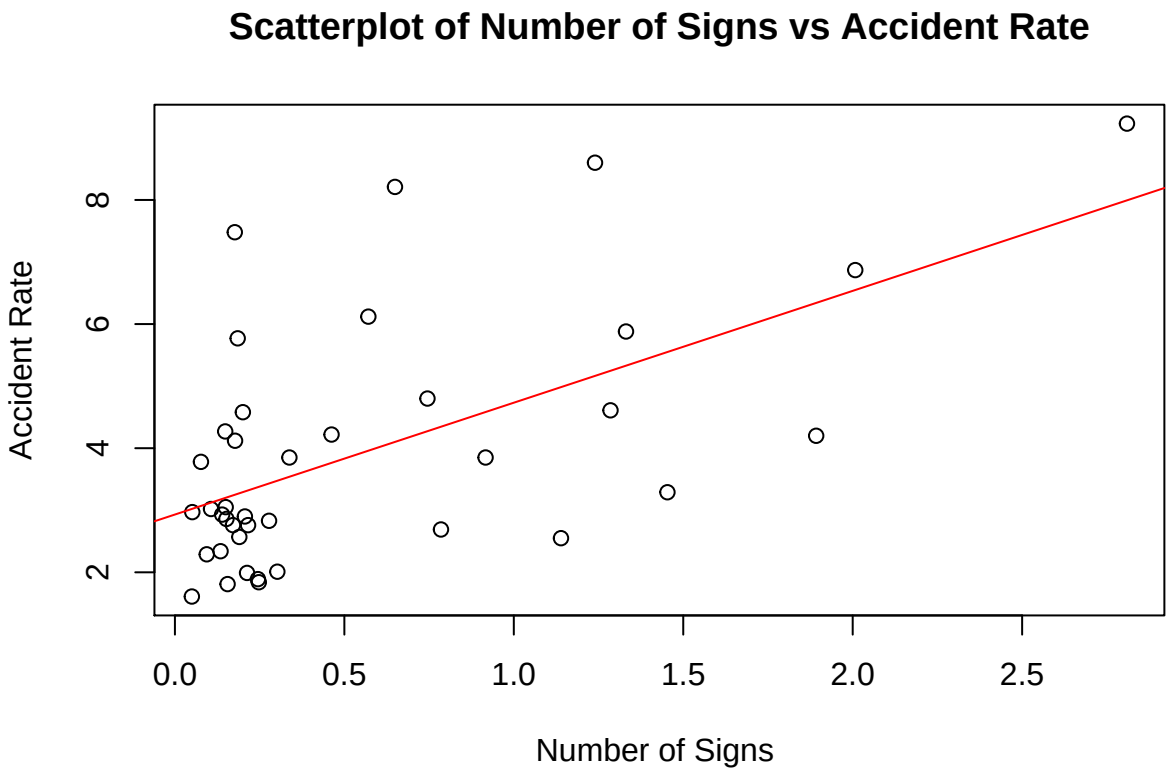
The data is in the file "Highway.csv". To read the data in `R`, save the file in your working directory (make sure you have changed the directory if different from the R working directory) and read the data using the `R` function `read.csv()`, and we will extract the variables of interest into two vectors.

```
# Read in the data
data = read.csv("Highway.csv", head = TRUE, sep = ",")
# Extract the predictor and response variables
rate = as.numeric(data[,2])
signs = as.numeric(data[,6])
```

Question B1: Exploratory Data Analysis - 8 pts

a. **2 pts** Use a scatter plot to describe the relationship between the rate of accidents and the number of signs. Describe the general trend (direction and form). Include plots and R-code used.

```r
plot(rate ~ signs, main="Scatterplot of Number of Signs vs Accident Rate",
   xlab="Number of Signs ", ylab="Accident Rate")
abline(lm(rate ~ signs), col="red")
```



The data is clustered in the lower values and there is increased variablility as the number of signs increases. The relationship is non-linear.

b. **2 pts** What is the value of the correlation coefficient? (Use the cor() function in R with the two input variables (signs,rate)). Please interpret. Interpret the strength of the correlation based on the correlation coefficient.

```r
corr <- cor(signs, rate)
```

58.2907219% of the variability of the accident rate can be explained by the variablility in the number of signs.

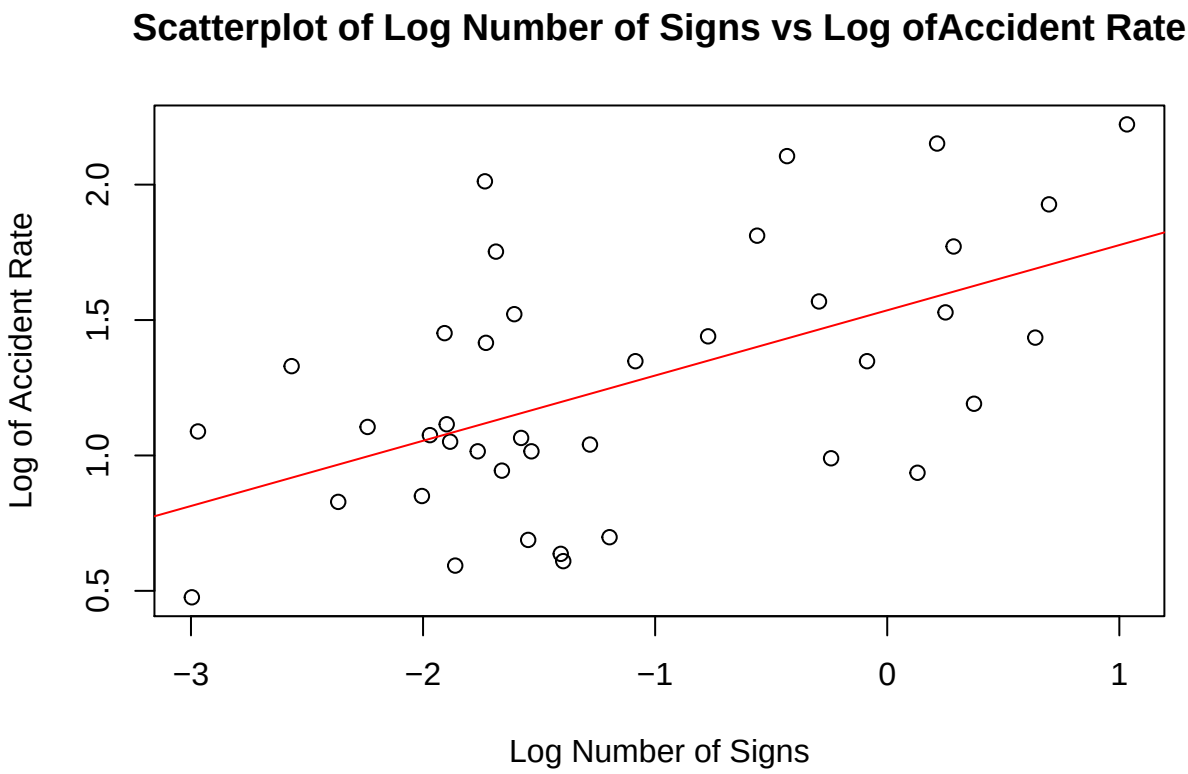
c. **2 pts** Based on this exploratory analysis, is it reasonable to assume a simple linear regression model for the relationship between rate of accidents and the number of signs?

No. I would not assume a simple linear regression model without attempting to transform the data first. The data is non-linear.

d. **2 pts** Based on the analysis above, would you pursue a transformation of the data?

Yes. I would log transform the both the rate and signs variable. The resulting scatter plot would look ike the plot below.

```
plot(log(rate) ~ log(signs), main="Scatterplot of Log Number of Signs vs Log ofAccident Rate",
   xlab="Log Number of Signs", ylab="Log of Accident Rate")
abline(lm(log(rate) ~ log(signs)), col="red")
```



##

Question B2: Fitting the Simple Linear Regression Model - 12 pts

Fit a linear regression model to evaluate the relationship between the rate of accidents and the number of signs. Do not transform the data. The function you should use in R is:

```
# Create the model
model = lm(rate ~ signs)
```

a. **3 pts** What are the model parameters and what are their estimates?

```
xtable(summary(model))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9310	0.3483	8.41	0.0000
signs	1.8021	0.4130	4.36	0.0001

The predictor variable is the number of signs and the response variable is the accident rate.

b. **3 pts** Write down the equation for the simple linear regression model.

Accident rate per million miles = 2.931 + 1.8021 signs per mile

c. **3 pts** Interpret the estimated value of the β_1 parameter in the context of the problem. Include its standard error in your interpretation.

For every additional 1.802 signs per mile there is an increase of one accident per million miles. The sign predictor variables are on average 0.4130 points away the mean.

d. **3 pts** Find a 95% confidence interval for the β_1 parameter. Is β_1 statistically significant at this level?

```
xtable(confint(model))
```

	2.5 %	97.5 %
(Intercept)	2.23	3.64
signs	0.97	2.64

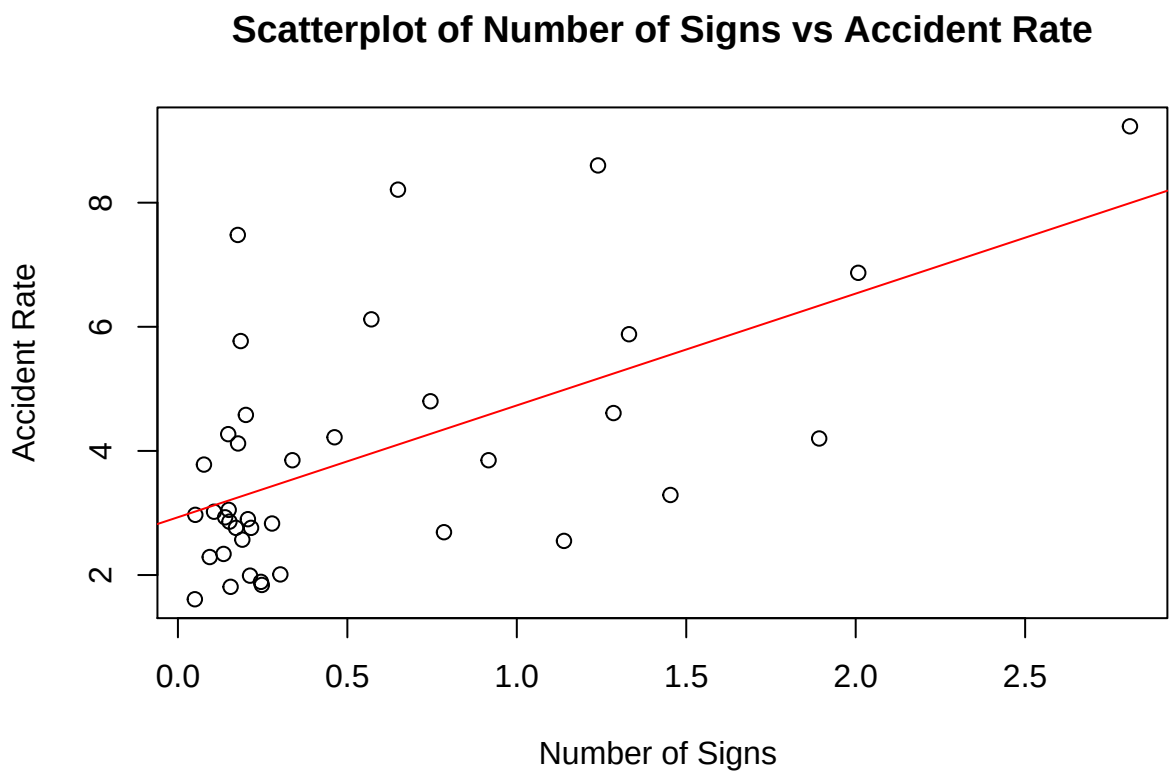
At a confidence interval of 95%, we can assume that the actual slope of the estimated regression line of 1.8021 for the number of signs vs the accident rate is between 0.97 and 2.64. The p-value for signs variable being statistically significant is 0.0001 which is < 0.05. The number is statistically significant.

Question B3: Checking the Assumptions of the Model - 16 pts

Interpret the following graphs with respect to the assumptions of the linear regression model. In other words, comment on whether there are any apparent departures from the assumptions of the linear regression model. Make sure that you state the model assumptions and assess each one. Each graph may be used to assess one or more model assumptions.

a. Scatterplot of the data with signs on the x-axis and rate on the y-axis

```
plot(rate ~ signs, main="Scatterplot of Number of Signs vs Accident Rate",
   xlab="Number of Signs ", ylab="Accident Rate")
abline(lm(rate ~ signs), col="red")
```



Model Assumption(s) it checks:

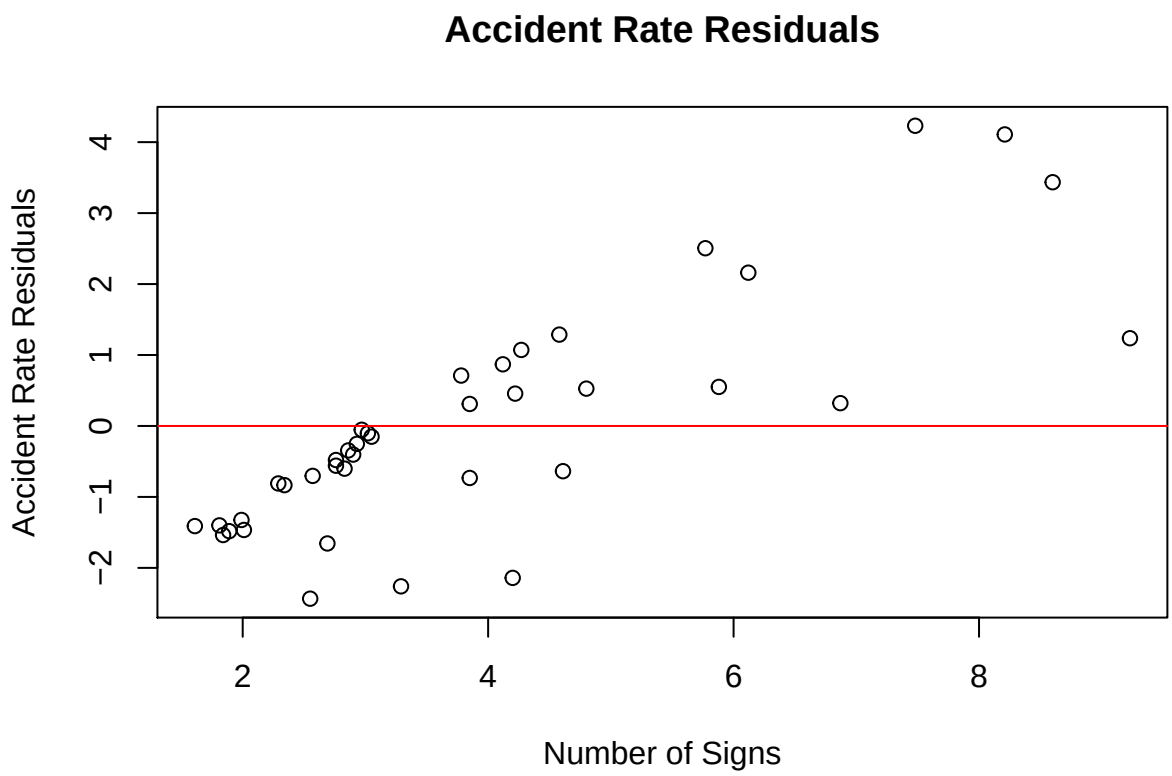
It checks for linearity of the model.

Interpretation:

The data is clustered under 0.5 signs per mile. The model may be non-linear.

b. Residual plot - a plot of the residuals, ϵ_i, versus, $\hat{y}_i$

```
model.res = resid(model)
plot(rate, model.res,
     ylab="Accident Rate Residuals", xlab="Number of Signs",
     main="Accident Rate Residuals")
abline(0,0, col='red')
```



Model Assumption(s) it checks:

It checks if the mean of the residuals is 0 and have constant variance.

Interpretation:

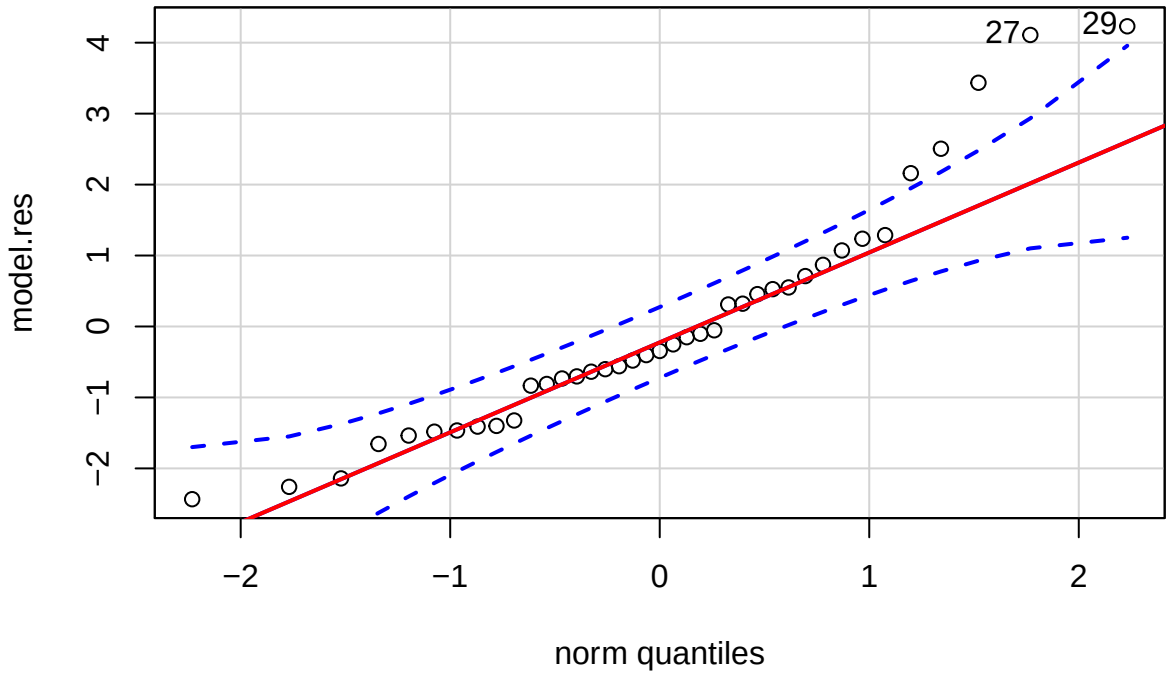
The residuals are not evenly dispersed around the mean of 0. The values of signs approximately < 4 are mostly < 0 while the values $>= 4$ are above zero. The variance spreads out as the number of signs increase. This indicates that there is a non-linear pattern to the data.

c. q-q plot

```
qqPlot(model.res, pch=1)
```

[1] 29 27

```
qqline(model.res, col = "red", lwd = 2)
```



Model Assumption(s) it checks:

It determines if the residuals are normally distributed.

Interpretation:

The data is curved and does not follow the AB line, indicating the data may be exponential.

Question B4: Prediction - 4 pts

Suppose we are interested in predicting future accident rates when `signs = 1.25`. Please make a prediction and provide the 95% prediction interval. What observations can you make about the result?

```
predict = predict(model, data.frame(signs = c(1.25)), interval="predict", level=0.95)
xtable(predict)
```

	fit	lwr	upr
1	5.18	1.78	8.59

The expected accident rate per million miles is 5.18 with a 95% likelyhood the actual rate is somewhere between 1.78 accidents per million miles and 8.59 accidents per million miles.