Problem 2.8

Stephen Blatti

October 15, 2017

2.8 Use the Cars2004EU data frame from the PASWR package which contains the numbers of cars per 1000 inhabitants (cars), the total number of known mortal accidents (deaths) and the country population/1000 (population) for the 25 member countries of the European Union for the year 2004.

```
library(PASWR2)

## Warning: package 'PASWR2' was built under R version 3.4.2

## Loading required package: lattice

## Loading required package: ggplot2
```

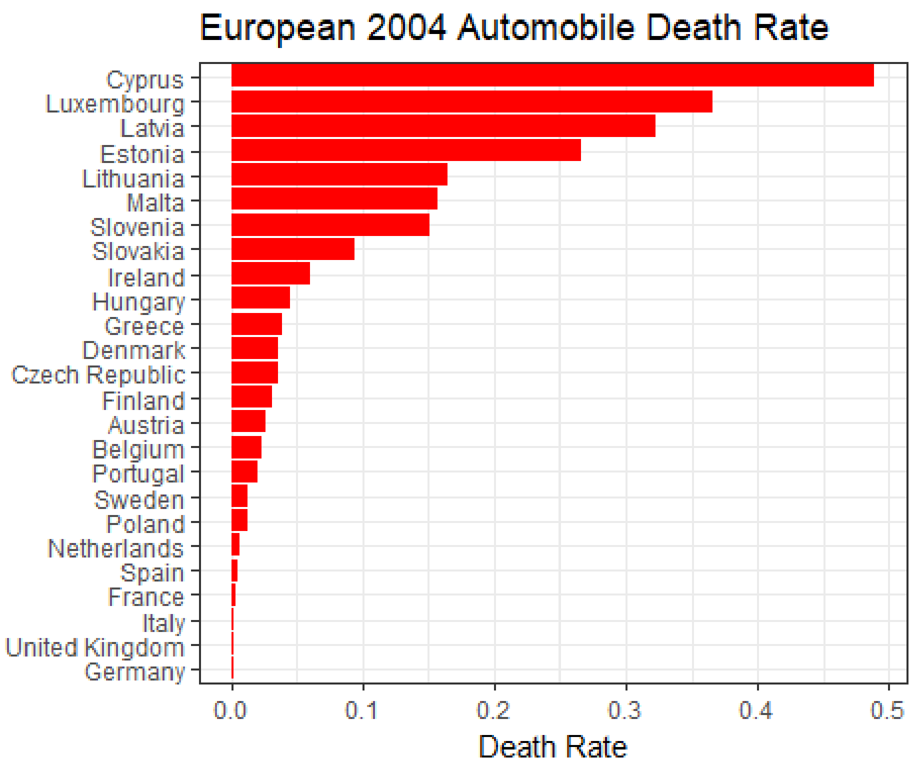
(a) Compute the total number of cars per 1000 inhabitants in each country, and store the result in an object named total.cars. Determine the total number of known automobile fatalities in 2004 divided by the total number of cars for each country and store the result in an object named death.rate.

```
CARS2004 <- within(data = CARS2004, expr = {
  total.cars = cars * population / 1000
  death.rate = deaths / total.cars
  })
head(CARS2004)

##          country cars deaths population death.rate total.cars
## 1        Belgium  467    112      10396 0.02306932   4854.932
## 2 Czech Republic  373    135      10212 0.03544167   3809.076
## 3        Denmark  354     68       5398 0.03558548   1910.892
## 4        Germany  546     71      82532 0.00157559  45062.472
## 5        Estonia  350    126       1351 0.26646928    472.850
## 6         Greece  348    147      11041 0.03825865   3842.268
```

(b) Create a barplot showing the automobile death rate for each of the European Union member countries. Make the bars increase in magnitude so that the countries with the smallest automobile death rates appear first.

```
ggplot(data =CARS2004, aes(x = reorder(country, death.rate), y = death.rate))
+
  geom_bar(stat = "identity", fill = "red") +
  coord_flip() + labs(x = "", y = "Death Rate",
                      title = "European 2004 Automobile Death Rate") +
theme_bw()
```



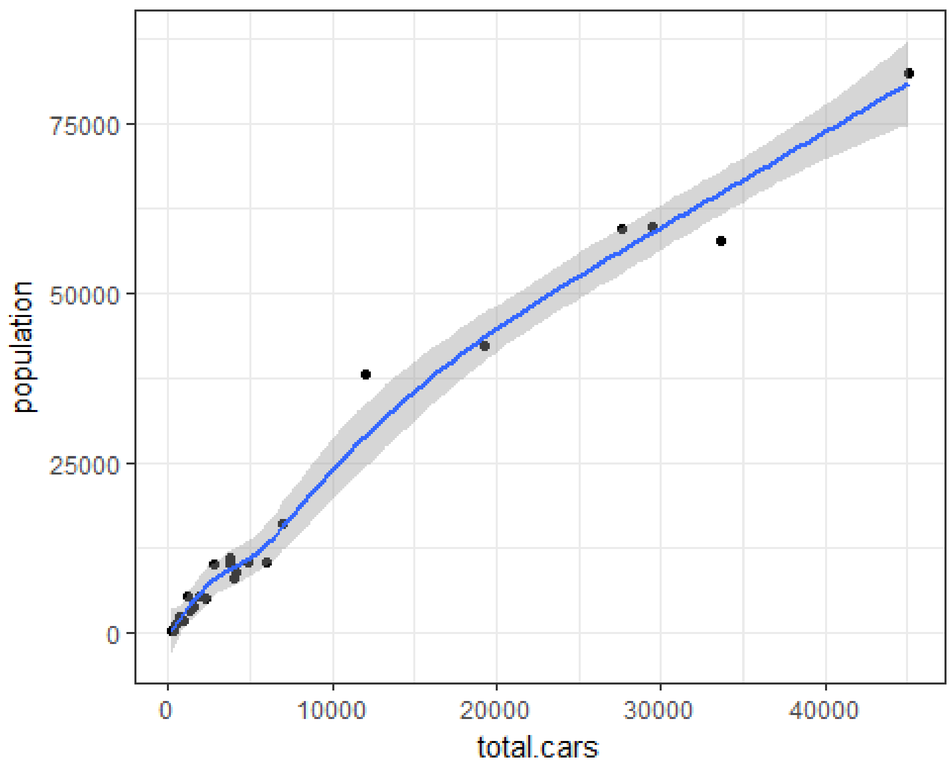
(c) Which country has the lowest automobile death rate? Which country has the highest automobile death rate?

```
# Germany - Lowest
# Cyprus - highest
```

(d) Create a scatterplot of population versus total.cars. How would you characterize the relationship?

```
ggplot(data = CARS2004, aes(x = total.cars, y = population)) + geom_point() +
  geom_smooth() + theme_bw()

## `geom_smooth()` using method = 'loess'
```




```
# Positive curvilinear relationship between total.cars and population.
```

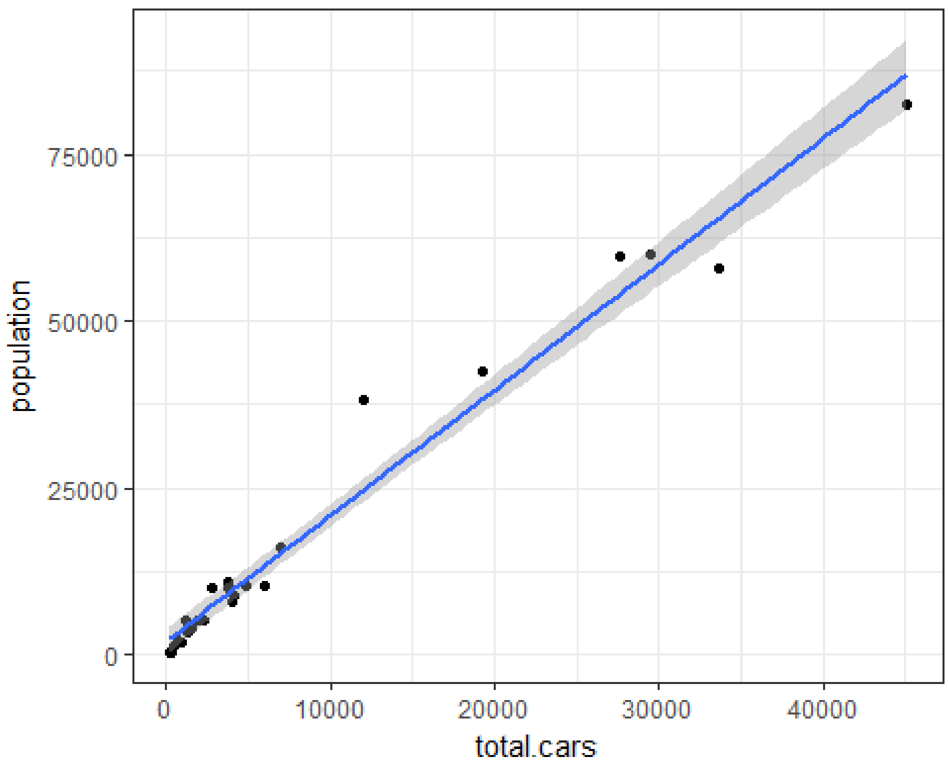
(e) Find the least squares estimates for regressing population on total.cars. Superimpose the least squares line on the scatterplot from (d). What population does the least squares model predict for a country with a total.cars value of 19224.630? Find the difference between the population predicted from the least squares model and the actual population for the country with a total.cars value of 19224.630.

```
fit <- lm(population ~ total.cars, data = CARS2004)
summary(fit)

##
## Call:
## lm(formula = population ~ total.cars, data = CARS2004)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7500  -1840  -1013   1015  13510
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.124e+03  9.731e+02   2.183   0.0395 *
## total.cars  1.881e+00  6.561e-02  28.668   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3967 on 23 degrees of freedom
```

```
## Multiple R-squared:  0.9728, Adjusted R-squared:  0.9716
## F-statistic: 821.8 on 1 and 23 DF,  p-value: < 2.2e-16
```

```
ggplot(data = CARS2004, aes(x = total.cars, y = population)) + geom_point() +
  geom_smooth(method = "lm") + theme_bw()
```




```
population <- predict(fit, newdata = data.frame(total.cars = 19224.630)) *
1000
population
```

```
##        1
## 38285550
```

```
residuals(fit)[7]*1000 # Spain is num 7
```

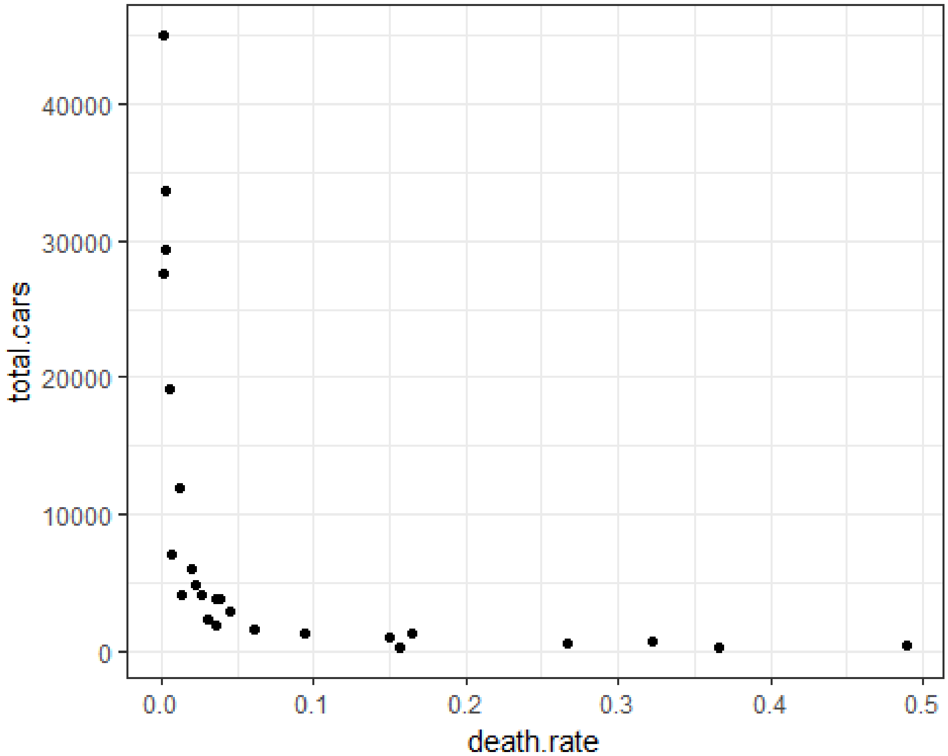
```
##       7
## 4059450
```

```
# The difference between Spain's actual population and the value predicted
with least squares (the seventh residual) 42,345,000 - 38,285,550 =
4,059,450.
CARS2004$population[CARS2004$country=="Spain"] * 1000 - population
```

```
##       1
## 4059450
```

(f) create a scatterplot of total.cars versus death.rate. How would you characterize the relationship between the two variables?

```r
ggplot(data = CARS2004, aes(x = death.rate, y = total.cars)) + geom_point() +
  theme_bw()
```




```r
# Decreasing monotonic relationship between total.cars and death.rate.
```

(g) Compute Spearman's rank correlation of total.cars and death.rate. What is this coefficient measuring?

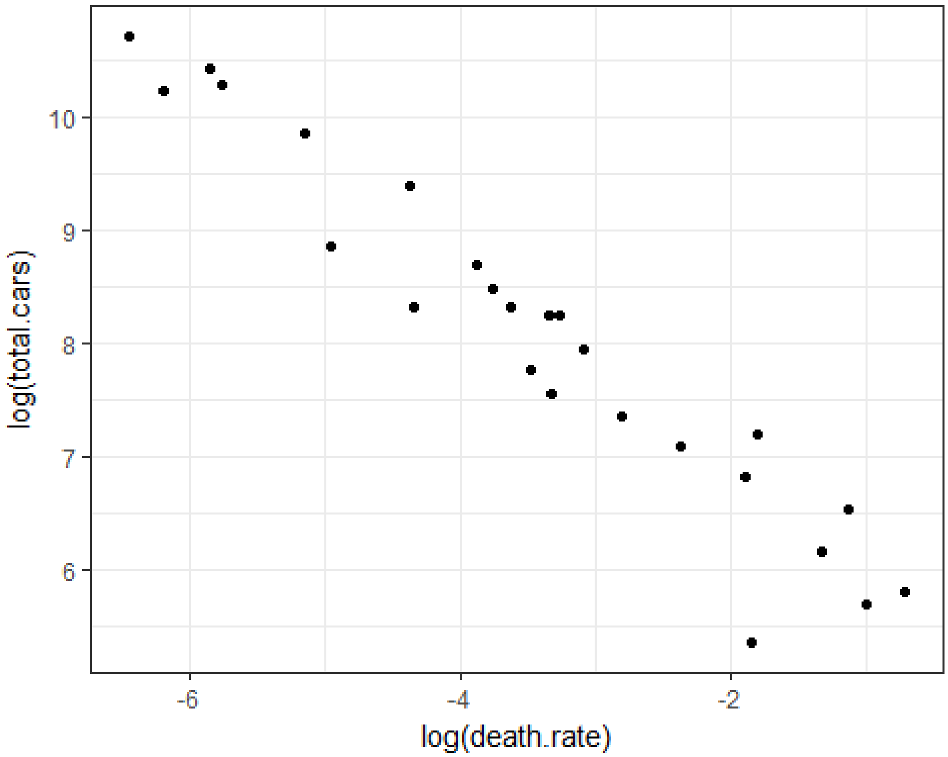
```r
with(data = CARS2004, cor(total.cars, death.rate, method = "spearman"))

## [1] -0.9676923

# Spearman’s rank correlation coefficient measures the monotonic relationship
between two variables.
```

(h) Plot the logarithm of total.cars vs the logarithm of death.rate. How would you characterize the relatinship.

```r
ggplot(data = CARS2004, aes(x = log(death.rate), y = log(total.cars))) +
geom_point() +
  theme_bw()
```



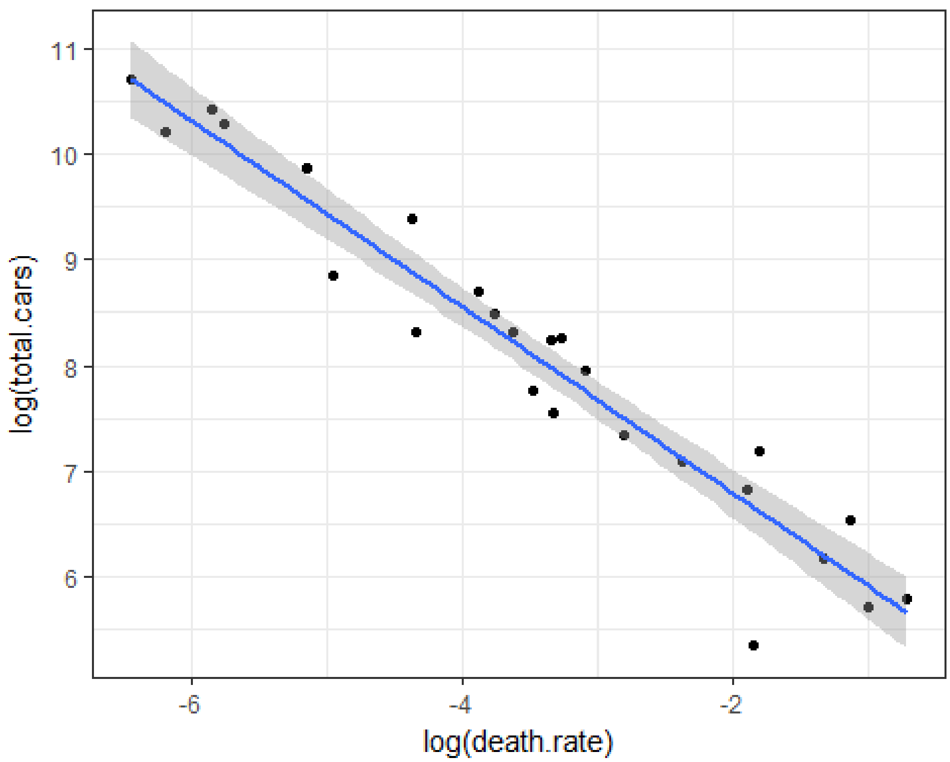

```
# The relationship is strong, linear, and negative between the logarithm of
total.cars
# and the logarithm of death.rate.
```

(i) What are the least squares estimates for the regression of log(death.rate) on log(total.cars). Superimpose the least squares line on the scatterplot from (h).

```
log.fit <- lm(log(total.cars) ~ log(death.rate), data = CARS2004)
coef(summary(log.fit))

##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)      5.0206666 0.19568324  25.65711 1.994256e-18
## log(death.rate) -0.8833401 0.05142204 -17.17824 1.293676e-14

ggplot(data = CARS2004, aes(x = log(death.rate), y = log(total.cars))) +
geom_point() +
  theme_bw() + geom_smooth(method = "lm")
```



11
10
9
8
7
6
log(total.cars)
-6
-4
-2
log(death.rate)