PA 541: Homework 1

Alexis Kwan

Part 1

Question 1

1 a

First, let's select a handful of variables to focus on and remove the others. Create a new dataset, call it fatality2, that contains only the following variables: fatal, state, year, spirits, unemp, income, dry, pop, and miles. Use this dataset for all steps below.

```
fatality2 <- fatalities %>%
             select(fatal, state, year, spirits, unemp, income, dry, pop, miles)
```

1 b

For each year available in the dataset (i.e., 1982 – 1988), how many total fatalities were there in each of those years?

```
fatality2 %>% select(fatal, year) %>%
              group_by(year) %>%
              summarise( total = sum(fatal) )
```

```
## # A tibble: 7 x 2
##    year total
##   <dbl> <dbl>
## 1  1982 43642
## 2  1983 42232
## 3  1984 43921
## 4  1985 43512
## 5  1986 45818
## 6  1987 46118
## 7  1988 46788
```

There were fatalities in the 40,000's for each of the years, increasing at about a thousand for each year.

1 c

Which state had the largest number of fatalities in 1982?

```
fatality2[fatality2$fatal == max(fatality2$fatal), c("state","fatal")]
```

```
## # A tibble: 1 x 2
##   state fatal
##   <chr> <dbl>
## 1 ca     5504
```

California had the most fatalities in 1982 at 5504 fatalities.

1 d

Which states in which years had more than 1,000 fatalities and more than 20% of its population residing in dry counties?

```
fatality2[fatality2$fatal > 1000 & fatality2$dry > 20, c("state","year","dry")]
```

```
## # A tibble: 10 x 3
##    state  year   dry
##    <chr> <dbl> <dbl>
##  1 al     1986  23.5
##  2 al     1987  23.8
##  3 al     1988  23.8
##  4 nc     1982  27.6
##  5 nc     1983  26.7
##  6 nc     1984  26.1
##  7 nc     1985  25.6
##  8 nc     1986  26.0
##  9 nc     1987  25.7
## 10 nc     1988  25.7
```

Alabama and North Carolina had more than 1,000 fatalities with at least 20% of its populations residing in dry counties. It spans the years of 1982 to 1988.

1 e

What is the average number of fatalities in each state?

```
fatality2 %>% select(fatal, state) %>%
              group_by(state) %>%
              summarise( avg = mean(fatal) )
```

```
## # A tibble: 48 x 2
##    state   avg
##    <chr> <dbl>
##  1 al     971
##  2 ar     574
##  3 az     864.
##  4 ca    5045
##  5 co     599.
##  6 ct     465.
##  7 de     130.
##  8 fl    2819.
##  9 ga    1440.
## 10 ia     482.
## # ... with 38 more rows
```

QUESTION 2

Create a new variable, 'fatal.cat' that breaks the continuous variable fatal down into three categories:

i. $0 - 300$
ii. $> 300 - 1000$
iii. > 1000

Please label the categories "low", "mid", "high". Set this new variable to be a factor. What is the mean of miles in each of the fatal categories?

```
fatality2$fatal.cat <-
  case_when(
    (fatality2$fatal > 0 & fatality2$fatal <= 300) ~ "low",
    (fatality2$fatal > 300 & fatality2$fatal <= 1000) ~ "mid",
    (fatality2$fatal > 1000) ~ "high"
  )
fatality2 %>% select(fatal.cat, miles) %>%
              group_by(fatal.cat) %>%
              summarise( avg_by_cat = mean(miles) )
```

```
## # A tibble: 3 x 2
##   fatal.cat avg_by_cat
##   <chr>          <dbl>
## 1 high           7645.
## 2 low            8509.
## 3 mid            7689.
```

The mean number of miles for the high fatality category is 7645.021, while it is 7689.332 and 8509.254 miles average for the middle and low number of fatality categories respectively.

PART TWO

Regression. For part 2, let's limit the fatality2 data from above to only the year 1987. So, to begin part 2, create this new dataset and call it fatality3.

```
fatality3 <- fatality2[fatality2$year == 1987,]
fatality3
```

```
## # A tibble: 48 x 10
##    fatal state  year spirits unemp income     dry       pop miles fatal.cat
##    <dbl> <chr> <dbl>   <dbl> <dbl>  <dbl>   <dbl>     <dbl> <dbl> <chr>
##  1  1110 al     1987    1.18  7.80  11944 23.8     4082999  9166. high
##  2   937 az     1987    1.72  6.20  14241  0       3385996. 9371. mid
##  3   639 ar     1987    1.01  8.10  11537 39.3     2388000. 7666. mid
##  4  5504 ca     1987    1.78  5.80  17846  0      27663018  8181. high
##  5   591 co     1987    1.78  7.70  15605  0.0581  3296004. 8182. mid
##  6   449 ct     1987    2.25  3.30  21192  0.0810  3210996  8339. mid
##  7   146 de     1987    2.37  3.20  16407  0        644000. 9450. low
##  8  2839 fl     1987    2.17  5.30  15584  0      12022987  7788. high
##  9  1599 ga     1987    1.75  5.5   14306  0.207   6222008. 9690. high
## 10   262 id     1987    1.06  8     11859  0        998000. 8135. low
## # ... with 38 more rows
```

QUESTION 3

Using the newly created fatality3 dataset, test the correlation between miles and fatal. What are your findings (i.e., what is the size of the correlation and is it significant)?

```
cor.test(fatality3$miles, fatality3$fatal)
```

```
##
##  Pearson's product-moment correlation
##
## data:  fatality3$miles and fatality3$fatal
## t = -1.3971, df = 46, p-value = 0.1691
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4595422  0.0873941
## sample estimates:
##        cor
## -0.2017504
```

There is a negative 0.2 correlation which means there is a weak correlation and when the number of fatalities increases the miles traveled would decrease. But since the p-value of 0.1691 $>$ 0.05 the correlation is not statistically significant. That also means that if the null hypothesis is that there is no correlation, we can not reject the null hypothesis under a significance level of 0.05.

QUESTION 4

Create a new population variable, that is population in 100,000s. Call the new variable pop_100k. Run a simple linear regression predicting fatal from pop.100k.

a. Interpret the estimates of the slope and intercept coefficients in the context of the problem.
b. What is the percentage of variation in fatal explained by pop_100k?
c. Predict the number of fatalities in a state if the population was 8 million.

```
fatality3$pop_100k <- fatality3$pop / 100000
mod1 <- lm(fatal ~ pop_100k, data = fatality3)
summary(mod1)
```

```
##
## Call:
## lm(formula = fatal ~ pop_100k, data = fatality3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -905.30  -94.77  -40.39  122.35  632.99
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  66.8469    53.8995    1.24    0.221
## pop_100k     17.7922     0.7444   23.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 268.9 on 46 degrees of freedom
## Multiple R-squared:  0.9255, Adjusted R-squared:  0.9239
## F-statistic: 571.2 on 1 and 46 DF,  p-value: < 2.2e-16
```

```
predict(mod1, list("pop_100k" = 8000))
```

```
##        1
## 142404.7
```

For every 100,000 increase in population, the model predicts an increase of 17.79 deaths.According to the R^2 given in the summary, approximately 92% of the variation is explained by the linear regression model on pop_100k. For a population of 8 million, this model predicts about 142,404.7 fatalities.

QUESTION 5

Which state has the largest negative residual in our model from question 4?
Which state has the largest positive residual?

Tell me what these large positive and large negative residuals mean within the context of our data and model.

```
neg_res <-
  sapply(
    residuals(mod1),
    FUN = function(x) {ifelse(test = (sign(x) == -1),
                              yes = x,
                              no = NA
                              )
    }
  )
c(fatality3$state[which.max(-neg_res)], neg_res[which.max(-neg_res)])
```

```
##                                     30
##                "ny" "-905.302844573234"
```

New York state has the largest residual of approximately -905.30.

```
pos_res <-
  sapply(
    residuals(mod1),
    FUN = function(x) {ifelse(test = (sign(x) == 1),
                              yes = x,
                              no = NA
                              )
    }
  )
c(fatality3$state[which.max(pos_res)], pos_res[which.max(pos_res)])
```

```
##                                     8
##               "fl" "632.994974271745"
```

Florida has the largest positive residual at 632.99.

The size of these residuals means that for the states of Florida and New York, the linear model was the least accurate in the prediction of the number of fatalities for these states.

QUESTION 6

Fit another regression model with fatal as the dependent variable and pop_100k, miles, and dry as the predictors.

a. What percentage of the variation in the dependent variable is explained by the predictors?
b. Ignoring whether the predictor is significant or not, interpret the coefficient estimates for each predictor. Be specific when discussing the relationship.
c. How do we interpret the p-value for dry?
d. By how much did our R-squared increase from our initial model that only included pop_100k as a predictor?

```
mod2 <- lm(fatal ~ pop_100k + miles + dry, data = fatality3)
summary(mod2)
```

```
##
## Call:
## lm(formula = fatal ~ pop_100k + miles + dry, data = fatality3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -595.23 -134.50    1.17  109.70  666.58
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.226e+03  2.966e+02  -4.133 0.000158 ***
## pop_100k     1.878e+01  6.656e-01  28.219  < 2e-16 ***
## miles        1.464e-01  3.373e-02   4.339 8.24e-05 ***
## dry          6.990e+00  3.340e+00   2.093 0.042127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 225.5 on 44 degrees of freedom
## Multiple R-squared:  0.9499, Adjusted R-squared:  0.9464
## F-statistic: 277.8 on 3 and 44 DF,  p-value: < 2.2e-16
```

About 95% of the variation is explained by this model. The pop_100k coefficient tells us that for every additional 100,000 people, there would be about 18 additional deaths on average, holding the number of miles traveled, and the "dryness" of the county constant. Similarly, the miles and "dry" coefficients tells us that for 1 mile increase in travel or 1% increase in "dryness" there would be on average 0.14 more or 7 more fatalities respectively, holding the other factors constant. The intercept tells us if there was no contribution from the other factors, i.e. at a population of 0, 0 miles traveled, and 0% dry, then there would still be an average of -1226 deaths. Clearly the other variables make up for this negative base value in their contributions to the number of fatalities. The p-value for our dry variable is at 0.0421, which tells us that there is a 4.21% chance that this result is random. At a significance level of 0.05, the coefficient is statistically significant and we can reject the null hypothesis that the coefficient is 0 or that there is no effect from dryness. As compared to our previous model with just the pop_100k as a predictor our R^2 squared increased by 0.02, which means the addition of the miles and dry variables only explained 2% more of the variation.

QUESTION 7

Run the following two models and compare the difference in the size and direction of the coefficient on miles.

$Y_i = b_0 + b_1 miles_i + e_i$ $Y_i = b_0 + b_1 miles_i + b_2 pop100\_k_i + e_i$

What is happening here? Can we trust the estimate of the effect of miles in the first model?

```
mod3 <- lm(fatal ~ miles, data = fatality3)
mod4 <- lm(fatal ~ miles + pop_100k, data = fatality3)
summary(mod3)
```

```
##
## Call:
## lm(formula = fatal ~ miles, data = fatality3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1261.6  -585.9  -240.8   264.2  4522.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2518.5778  1123.7013   2.241   0.0299 *
## miles         -0.1879     0.1345  -1.397   0.1691
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 964.7 on 46 degrees of freedom
```

```
## Multiple R-squared:  0.0407, Adjusted R-squared:  0.01985
## F-statistic: 1.952 on 1 and 46 DF,  p-value: 0.1691
```

We cannot trust the first model for a couple reasons. Firstly, the coefficient for miles fails to be statistically significant even for a level of 0.05 since the p-value is 0.17. Secondly, only 4% of the variation is explained by the model.

```
summary(mod4)
```

```
##
## Call:
## lm(formula = fatal ~ miles + pop_100k, data = fatality3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -640.77 -137.12   -6.78  153.22  636.00
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.127e+03  3.036e+02  -3.711 0.000567 ***
## miles        1.382e-01  3.474e-02   3.978 0.000249 ***
## pop_100k     1.874e+01  6.899e-01  27.166  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 233.8 on 45 degrees of freedom
## Multiple R-squared:  0.9449, Adjusted R-squared:  0.9424
## F-statistic: 385.6 on 2 and 45 DF,  p-value: < 2.2e-16
```

We can see with the second model, not only are both of the coefficients statistically significant with p-values much smaller than 0.05, but much more of variance is explained when the population variable is added to the model. Clearly the second variable, pop_100k, has a positive effect on the independent variable, the number of fatalities. We see that contribution from miles to the fatalities in the model has increased evident from the changes of signs of the coefficient to positive from negative compared to the previous model.

```
cor.test(fatality3$miles, fatality3$pop_100k)
```

```
##
##  Pearson's product-moment correlation
##
## data:  fatality3$miles and fatality3$pop_100k
## t = -2.4975, df = 46, p-value = 0.01615
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5733918 -0.0681075
## sample estimates:
##       cor
## -0.3455551
```

From Pearson's correlation test we see that there is a statistically significant negative correlation, at a significance level of 0.05, between miles and pop_100k. From this we can see there was omitted variable bias, since there is a correlation between our independent variables and the new variable has an effect on the independent variable. Since the correlation is negative and the coefficient of pop_100k is negative, the omitted variable bias is negative. Intuitively, the intercept of the first model seemed to be overcompensating for the smaller coefficient and doing most of the heavy lifting, while in the second model miles traveled is compensating for the new variable in the opposite direction.