Homework One Answer Key

Michael D. Siciliano

January 26th, 2021

PART ONE

QUESTION 1 (12 pts)

The following will require you to use the tools and verbs we learned during week 1 to wrangle data. The results of these tasks will produce a tibble. You only need to copy and paste the tibble itself (what R reports) and not all of the variables or observations (i.e., don't print out the whole dataset).

```
fatality2 = fatality %>%
  select(fatal, state, year, spirits,unemp, income, dry, pop, miles)
fatality2
```

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. (2pts)

```
## # A tibble: 336 x 9
##    fatal state  year spirits unemp income   dry      pop miles
##    <dbl> <chr> <dbl>   <dbl> <dbl>  <dbl> <dbl>    <dbl> <dbl>
##  1   839 al     1982    1.37 14.4   10544.  25.0 3942002. 7234.
##  2   930 al     1983    1.36 13.7   10733.  23.0 3960008  7836.
##  3   932 al     1984    1.32 11.1   11109.  24.0 3988992. 8263.
##  4   882 al     1985    1.28  8.90 11333.  23.6 4021008. 8727.
##  5  1081 al     1986    1.23  9.80 11662.  23.5 4049994. 8953.
##  6  1110 al     1987    1.18  7.80 11944   23.8 4082999  9166.
##  7  1023 al     1988    1.17  7.20 12369.  23.8 4101992. 9674.
##  8   724 az     1982    1.97  9.90 12309.   0   2896996. 6810.
##  9   675 az     1983    1.90  9.10 12694.   0   2977004. 6587.
## 10   869 az     1984    2.14  5    13266.   0   3071996. 6710.
## # ... with 326 more rows
```

```
fatality2 %>%
  group_by(year) %>%
  summarize(total.fatalities = sum(fatal))
```

b. For each year available in the dataset (i.e., 1982 – 1988), how many total fatalities were there in each of those years? (2pts)

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 7 x 2
##    year total.fatalities
```

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

```
fatality2 %>%
  filter(year == 1982) %>%
  arrange(desc(fatal))
```

c. Which state had the largest number of fatalities in 1982? (2pts)

```
## # A tibble: 48 x 9
##    fatal state  year spirits unemp income    dry       pop miles
##    <dbl> <chr> <dbl>   <dbl> <dbl>  <dbl>  <dbl>     <dbl> <dbl>
##  1  4615 ca     1982    2.21  9.90 15797.  0     24785976  6859.
##  2  4213 tx     1982    1.56  6.90 13943. 11.3   15374004  8145.
##  3  2653 fl     1982    2.51  8.20 13502.  0     10478007  7587.
##  4  2162 ny     1982    2.16  8.60 15159.  0.208 17586958  4576.
##  5  1819 pa     1982    1.36 10.9  13652. 10.8   11879029  6003.
##  6  1651 il     1982    2.04 11.3  14743.  6.40  11478031  5697.
##  7  1607 oh     1982    1.27 12.5  13039. 11.6   10774027  6660.
##  8  1392 mi     1982    1.88 15.5  13247.  0      9116988  6713.
##  9  1303 nc     1982    1.64  9    11079. 27.6    6016002. 7164.
## 10  1229 ga     1982    1.94  7.80 11774.  0.499  5650990. 8623.
## # ... with 38 more rows
```

```
fatality2 %>%
  filter(fatal >1000, dry > 20)
```

d. Which states in which years had more than 1,000 fatalities and more than 20% of its population residing in dry counties. (2pts)

```
## # A tibble: 10 x 9
##    fatal state  year spirits unemp income   dry      pop miles
##    <dbl> <chr> <dbl>   <dbl> <dbl>  <dbl> <dbl>    <dbl> <dbl>
##  1  1081 al     1986    1.23  9.80 11662.  23.5 4049994. 8953.
##  2  1110 al     1987    1.18  7.80 11944   23.8 4082999  9166.
##  3  1023 al     1988    1.17  7.20 12369.  23.8 4101992. 9674.
##  4  1303 nc     1982    1.64  9    11079.  27.6 6016002. 7164.
##  5  1234 nc     1983    1.56  8.90 11455.  26.7 6076992. 7411.
##  6  1450 nc     1984    1.53  6.70 12089.  26.1 6165988. 7814.
##  7  1482 nc     1985    1.5   5.40 12354.  25.6 6255012  7981.
##  8  1647 nc     1986    1.45  5.30 12839.  26.0 6331012. 8255.
##  9  1584 nc     1987    1.40  4.5  13325   25.7 6413007  8514.
## 10  1573 nc     1988    1.34  3.60 13767.  25.7 6489006. 8929.
```

```
fatality2 %>%
  group_by(state) %>%
```

```
  summarize(mean.fatality = mean(fatal))
```

e. What is the average number of fatalities in each state? (2pts)

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 48 x 2
##    state mean.fatality
##    <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 (8 pts)

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. (4pts)

```
fatality2 = fatality2 %>%
  mutate(fatal.cat = case_when( fatal <= 300 ~ 'low',
                               fatal > 300  & fatal <= 1000 ~ 'mid',
                               fatal > 1000 ~ 'high'))

fatality2$fatal.cat = factor(fatality2$fatal.cat, levels = c("low", "mid", "high") )

fatality2 %>%
  group_by(fatal.cat) %>%
  summarize(mean.miles = mean(miles))
```

What is the mean of miles in each of the fatal categories? (4pts)

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

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

The mean number of miles for low is 8509, for mid 7689, and for high 7645.

PART TWO

```r
fatality3 = fatality2 %>%
  filter(year ==1987)
fatality3
```

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.

```
## # 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> <fct>
##  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 (6 pts)

Using the newly created fatality3 dataset, test the correlation between miles and fatal. (2pts) What are your findings (i.e., what is the size of the correlation and is it significant)? (4pts)

```r
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
```

The correlation indicates that average miles driven per driver and number of fatal crashes are negatively correlated, at -.20, but the value is not significant (p-value = 0.16)

QUESTION 4 (12 pts)

```r
fatality3$pop_100k = fatality3$pop/100000

mod1 = lm(fatal ~ pop_100k, data = fatality3)
summary(mod1)
```

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. (4pts) (a) Interpret the estimates of the slope and intercept coefficients in the context of the problem.

(4pts) (b) What is the percentage of variation in fatal explained by pop_100k? (2pts) (c) Predict the number of fatalities in a state if the population was 8 million. (2pts)

```
##
## 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
```

a. Intercept: For states with a population of zero, the predicted number of fatal crashes is 66.8. This value makes no sense given that states can't have a population of zero. It is simply the point at which the line crosses the y axis.

   Slope: For each additional 100k people in the state, the predicted number of fatalities increases by 17.8. This effect is significant at the .001 level.

b. The R-squared is .93, meaning that 93% of the variation in fatalities is explained by population.

C. The number of fatalities can be predicted as follows:

```
66.85 + 17.79 *(80)
```

```
## [1] 1490.05
```

QUESTION 5 (8 pts)

```
fatality3$resid = resid(mod1)
fatality3$pred = predict(mod1)

fatality3 %>%
  arrange(resid) %>%
  select(state, fatal, pop_100k, pred, resid)
```

Which state has the largest negative residual in our model from question 5? (2pts) Which state has the largest positive residual? (2pts) Tell me what these large positive and large negative residuals mean within the context of our data and model. (4pts)

```
## # A tibble: 48 x 5
##    state fatal pop_100k  pred resid
##    <chr> <dbl>    <dbl> <dbl> <dbl>
##  1 ny     2333    178.   3238. -905.
##  2 il     1660    116.   2128. -468.
##  3 ma      689     58.6  1109. -420.
##  4 nj     1023     76.7  1432. -409.
```

```
##   5 mn      530    42.5    822. -292.
##   6 oh     1772   108.    1986. -214.
##   7 pa     1987   119.    2191. -204.
##   8 ct      449    32.1    638. -189.
##   9 ri      113     9.86   242. -129.
## 10 wi      797    48.1    922. -125.
## # ... with 38 more rows
```

```r
fatality3 %>%
  arrange(desc(resid)) %>%
  select(state, fatal, pop_100k, pred, resid)
```

```
## # A tibble: 48 x 5
##    state fatal pop_100k  pred resid
##    <chr> <dbl>    <dbl> <dbl> <dbl>
##   1 fl      2839     120.   2206.  633.
##   2 ca      5504     277.   4989.  515.
##   3 ga      1599      62.2 1174.  425.
##   4 sc      1086      34.3  676.  410.
##   5 nc      1584      64.1 1208.  376.
##   6 tn      1248      48.5  931.  317.
##   7 al      1110      40.8  793.  317.
##   8 az       937      33.9  669.  268.
##   9 nm       568      15.0  334.  234.
## 10 ms       756      26.2  534.  222.
## # ... with 38 more rows
```

These large positive and negative residuals indicate states that are not well fit given our current model. Because the residual is the observed minus the predicted values, very large positive residuals suggest that the predicted value fell far below what was actually observed. Thus Florida, with a population of 12 million, had 2,839 fatal crashes, but were predicted to have only 2,206. So they have a large positive residual of 632.9. In other words, they had many more fatal crashes than expected by our model. A large negative residual suggests a state that had many fewer fatal crashes then expected. In other words, the model predicted they would have more crashes than they actually did. Here we see New York had a residual of -905.3. They model predicted they would have 3,238 crashes, but only had 2,333.

QUESTION 6 (12 pts)

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

Fit another regression model with fatal as the dependent variable and pop_100k, miles, and dry as the predictors. (2pts) (a) What percentage of the variation in the dependent variable is explained by the predictors? (2pts) (b) Ignoring whether the predictor is significant or not, interpret the coefficient estimates for each predictor. Be specific when discussing the relationship. (4pts) (c) How do we interpret the p-value for dry? (2pts) (d) By how much did our R-squared increase from our initial model that only included pop_100k as a predictor? (2pts)

```
##
## 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
```

a. In our updated model with the three predictors, we explain 95% of the variation.

b. *Pop_100k* - for each additional 100k people in a state the number of fatalities is expected ot increase by 18.8

   miles - for each additional mile driven on average per driver, the number of fatalities increases by .146.

   dry - For each percentage increase in the percentage of residents residing in a dry county the predicted number of fatalities increases by 7. This value seems counter intuitive. Something to think about more regarding what may be going on here.

c. The p-value for dry is .042. This indicates that probability of finding an effect of this size or larger in repeated samples if the true effect is indeed 0. We would say that dry is significant at the .05 level.

d. The R-squared when up by about 2.4%

QUESTION 7 (12pts)

Run the following two models and compare the difference in the size and direction of the coefficient on *miles*. (6pts) What is happening here? Can we trust the estimate of the effect of miles in the first model? (6pts)

$$Y_i = \beta_0 + \beta_1 miles_i + e_i$$

$$Y_i = \beta_0 + \beta_1 miles_i + \beta_2 pop\_100k_i + e_i$$

```
mod3 = lm(fatal ~ miles , data = fatality3)

mod4 = lm(fatal ~ miles + pop_100k, data = fatality3)

screenreg(list(mod3, mod4))
```

```
##
## ======================================
##              Model 1     Model 2
## --------------------------------------
## (Intercept)   2518.58 *  -1126.64 ***
##             (1123.70)    (303.63)
## miles           -0.19        0.14 ***
##                 (0.13)      (0.03)
## pop_100k                   18.74 ***
```

```
##                              (0.69)
## -------------------------------------
## R^2               0.04        0.94
## Adj. R^2          0.02        0.94
## Num. obs.        48          48
## =====================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

So the coefficient on *miles* goes from a negative .19 to a positive .14. This is an example of omitted variable bias. The variable pop_100k is a variable that belongs in the regression and it is correlated with miles. We can see that correlation here:

```r
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
```

Notice that miles and pop_100k are negatively correlated. Thus as miles increases, population decreases. This makes sense as people who live in more rural states likely need to drive more on a day to day basis (things are simply further away) and those in urban areas may rely more on public transit. Given this negative correlation along with a positive effect of pop_100k on fatalities, the coefficient for miles has a negative bias. [Think about it this way, as miles goes up, population goes down, and miles is now picking up the effect of both of those variables.]