Homework 1

Math 430, Winter 2017

Due Date: Wednesday, January 18 by 4:30 p.m.

Problem 1

The file `snow.csv` on the class web site contains data from a snow gauge calibration study. A snow gauge is an instrument that measures the wetness of snow, which is crucial in western states for predicting water availability. Wet snow is denser than dry snow. Density is time consuming to measure directly; the snow gauge instrument measures a quantity called gain that depends on the density. The data at hand were collected to calibrate the instrument, that is describe how gain is related to density. When the snow gauge is used, the gain is measured and used to predict the snow density.

Polyethylene blocks were used as substitute for snow. These can be manufactured in different densities. The density is set by the process used to manufacture the blocks. The data set (in `snow.csv`) includes 9 densities. Ten blocks of each density were measured.

(a) The investigators plan to use regression to describe the relationship between gain and density. Which variable (gain or density) should be used as the X variable? Which is the Y variable? If it doesn't matter, say so. Briefly explain.

SOLUTION: The investigators set the density values, so they should be the X variables. The gain values are measured with error, so they should be the Y variables.

(b) Rightly or wrongly, the investigators decide to use density as the X variable and gain as the Y variable. For all this and subsequent parts of this question, please assume that the usual simple linear regression model is appropriate. Estimate the slope and intercept of the regression of gain on density.

```
snow <- read.csv("https://github.com/math430-lu/data/raw/master/snow.csv")
snow.lm <- lm(gain ~ density, data = snow)
snow.lm$coefficients
```

```
## (Intercept)     density
##    348.4060   -579.9309
```

(c) If you assume a linear relationship, is density related to gain? In other words, test whether the slope = 0. Report your p-value and a short conclusion.

```
summary(snow.lm)
```

```
##
## Call:
## lm(formula = gain ~ density, data = snow)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -151.18  -58.48  -12.20   43.23  198.89
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   348.41      13.41   25.98   <2e-16 ***
## density      -579.93      33.50  -17.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 71.5 on 88 degrees of freedom
## Multiple R-squared:  0.7731, Adjusted R-squared:  0.7705
## F-statistic: 299.8 on 1 and 88 DF,  p-value: < 2.2e-16
```

Based on a test statistic of $T = -17.31$ and associated p-value < 0.0001, there is very strong evidence to indicate that the slope is not zero.

(d) Predict the average gain when the density = 0.2

```
predict(snow.lm, newdata = data.frame(density = 0.2))
```

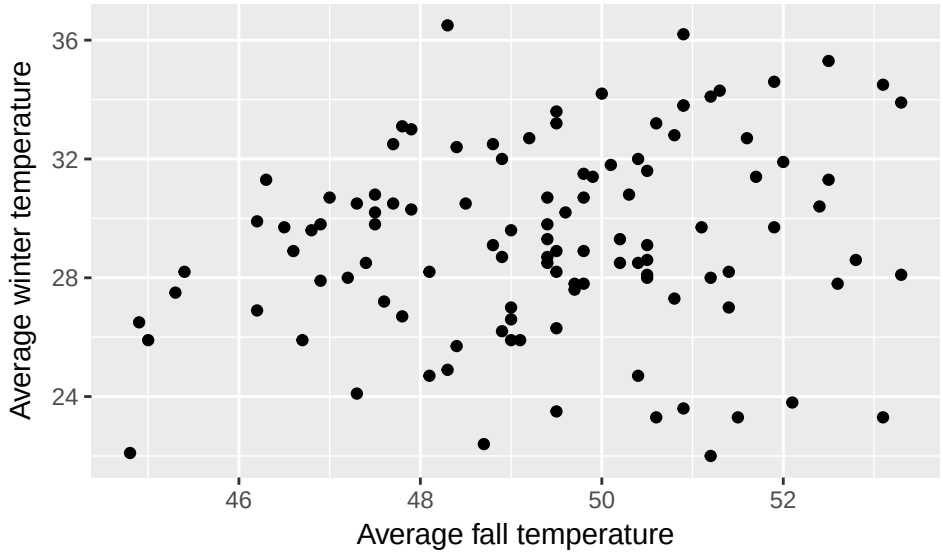
```
##        1
## 232.4198
```

Problem 2

The data file `ftcollinstemp.csv` gives the mean temperature in the `fall` of each year, defined as September 1 to November 30, and the mean temperature in the following `winter`, defined as December 1 to the end of February in the following calendar year, in degrees Farenheit, for Fort Collins, CO (Colorado Climate Center, 2012). These data cover the time period from 1900 to 2010. The question of interest is: Does the average `fall` temperature predict the average `winter` temperature?

(a) Draw (in R) a scatterplot of the response versus the predictor, and describe any pattern you might see in the plot.

```
ftcollinstemp <- read.csv("https://github.com/math430-lu/data/raw/master/ftcollinstemp.csv")
library(ggplot2)
ggplot(data = ftcollinstemp, aes(x = fall, y = winter)) +
  geom_point() +
  labs(x = "Average fall temperature", y = "Average winter temperature")
```



There is a very weak, positive association between the average fall and winter temperatures.

(b) Use R to fit the regression of the response on the predictor.

```
temp_mod <- lm(winter ~ fall, data = ftcollinstemp)
summary(temp_mod)
```

```
##
## Call:
## lm(formula = winter ~ fall, data = ftcollinstemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8186 -1.7837 -0.0873  2.1300  7.5896
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.7843     7.5549   1.825   0.0708 .
## fall          0.3132     0.1528   2.049   0.0428 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.179 on 109 degrees of freedom
## Multiple R-squared:  0.0371, Adjusted R-squared:  0.02826
## F-statistic:   4.2 on 1 and 109 DF,  p-value: 0.04284
```

(c) Interpret the estimated y-intercept in the context of the problem.

$\widehat{\beta}_0 \approx 13.8$. If the average fall temperature is 0°F, we expect the average winter temperature to be 13.8°F. Notice that this is an extrapolation since the average fall temperature has never been 0°F, so we should not trust this prediction.

(d) Interpret the estimated slope in the context of the problem.

For each 1°F increase in the average fall temperature, we expect the average winter temperature to increase by 0.3°F.

(e) Construct a 95% confidence interval for the slope and interpret it in the context of the problem. Does the slope appear to be significantly different from 0?

```
confint(temp_mod)
```

```
##                   2.5 %    97.5 %
## (Intercept) -1.18920049 28.757891
## fall         0.01028623  0.616052
```

We are 95% confident that for a 1°F increase in the average fall temperature, the average winter temperature will, on average, increase between 0.01°F and 0.62°F.

(f) The follwing code snippet breaks the data set into two time periods: an early period from 1900 to 1989 and a late period from 1990 to 2010.

```
early <- subset(ftcollinstemp, year < 1990)
late <- subset(ftcollinstemp, year >= 1990)
```

Fit the regression model to each of the time periods. Are the results different in the two time periods? Justify your answer statistically.

First, we must fit SLR models to each data set.

```
early_lm <- lm(winter ~ fall, data = early)
late_lm <- lm(winter ~ fall, data = late)

library(broom)
tidy(early_lm)
```

```
##          term   estimate std.error statistic     p.value
```

```
## 1 (Intercept) 22.7079122 8.2600078 2.7491393 0.007250303
## 2        fall  0.1208925 0.1681116 0.7191207 0.473971879
```

```
tidy(late_lm)
```

```
##          term   estimate  std.error statistic   p.value
## 1 (Intercept) 24.8259649 17.7972619 1.3949317 0.1791311
## 2        fall  0.1390029  0.3509374 0.3960904 0.6964512
```

Next, calculate CIs for each models coefficients (here I am calculating 95% confidence intervals)

```
confint(early_lm)
```

```
##                 2.5 %     97.5 %
## (Intercept)  6.292882 39.1229421
## fall        -0.213194  0.4549791
```

```
confint(late_lm)
```

```
##                   2.5 %     97.5 %
## (Intercept) -12.4241322 62.0760621
## fall         -0.5955175  0.8735234
```

From the above 95% confidence intervals we find no difference in slopes of the two lines (the intervals clearly overlap). Note that the early model has an intercept that is significantly above 0; however, the confidence intervals overlap, so there is not clear difference between the models.

Problem 3

(a) Show that the estimated slope can be written as $\widehat{\beta}_1 = r \cdot \frac{s_y}{s_x}$, where s_x and s_y are the sample standard deviations of x and y, respectively.

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{SXY}{SXX} \\
&= \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \\
&= \frac{1}{n-1} \sum \left(\frac{x_i - \overline{x}}{s_x} \cdot \frac{y_i - \overline{y}}{s_y} \right) \cdot \frac{s_y}{s_x} \\
&= r \cdot \frac{s_y}{s_x}
\end{aligned}$$

(b) What does this indicate about the relationship between signs of the correlation and the slope?

The fact that $\widehat{\beta}_1 = r \cdot \frac{s_y}{s_x}$ indicates that the slope and the correlation will always have the same sign, as sample standard deviation is not negative.