Homework 3

Due: Thursday 2/20/20 by 8:30am

1. Think back to Problem 1 from Homeworks 2 and 3. Suppose Instagram magically knew that every time the number of times user i purchases a product, denoted by Y_i, is related to the number of times the product has been advertised to user i, denoted by X_i, as follows:

$$Y_i = 1 + 2X_i + \epsilon_i$$

where ϵ_i is a *normal* random error term with mean $E\{\epsilon_i\} = 0$ and variance $\sigma^2\{\epsilon_i\} = 0.1$; ϵ_i and ϵ_j are uncorrelated so that their covariance is zero (i.e., $\sigma\{\epsilon_i, \epsilon_j\} = 0$ for all $i \neq j$) for $i = 1, \ldots n$.

   (a) Using `R`, make a plot with three panels. You can make a single plot with three panels by typing `par(mfrow = c(1, 3))` before running any lines of code that create plots. Plot the density of the Y_i for $X_i = 0$, $X_i = 1$, and $X_i = 10$, using a separate panel for each value of X_i. Ensure that the axes are the same across all three plots.

Suppose we were actually able to work with Instagram and conduct an experiment and we randomly selected $n = 10$ of the students in our class. We sent X_i ads for coffee to student i over the course of one day, and recorded Y_i, the number of ounces of coffee student i purchased the following week.

   (b) Suppose that after the experiment concluded, I told you that I fit a simple linear regression model to the data, modeling Y_i as a linear function of X_i. Imagine that I told you that when I tested the null hypothesis $H_0 : \beta_1 \geq 0$ versus the alternative $H_a : \beta_1 < 0$, I failed to reject H_0. Would you conclude that there is no linear association between X and Y?

Suppose that I decided to share the results of the regression with you all.

```
summary(lm(Y~X))
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.940  -9.299   2.344   7.347  35.104
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.489     10.729   1.257    0.244
## X             -1.001      1.788  -0.560    0.591
##
## Residual standard error: 18.75 on 8 degrees of freedom
## Multiple R-squared:  0.03769,    Adjusted R-squared:  -0.0826
## F-statistic: 0.3133 on 1 and 8 DF,  p-value: 0.591
```

   (c) Imagine that an Instagram executive came to class, saw the regression results, and stated "The message I got here is that the more coffee advertisements we send, the less coffee people drink!" Would you agree or disagree? Explain.

2. Refer back to the Toluca Company example we have discussed in class. The `toluca` data has been posted on the Homework page.

```
load("~/Dropbox/Teaching/STAT525-2020/stat525/content/homework/toluca.RData")
X <- data$X
Y <- data$Y
linmod <- lm(Y~X)
summary(linmod)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.876 -34.088  -5.982  38.826 103.528
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.366     26.177   2.382   0.0259 *
## X              3.570      0.347  10.290 4.45e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.82 on 23 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8138
## F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

(a) Label b_0, $s\{b_0\}$, $\frac{b_0}{s\{b_0\}}$, b_1, $s\{b_1\}$, $\frac{b_1}{s\{b_1\}}$ and s on the `R` output given above.

Simulations can be very helpful for building an intuition to interpreting intervals and test results. Using the values of b_0 and s from (a) and `rnorm`, let's simulate $k = 1, \ldots, 1,000$ synthetic datasets $Y_1^{(k)}, \ldots, Y_{25}^{(k)}$ according to:

$$Y_i^{(k)} = b_0 + \epsilon_i^{(k)}, \quad \epsilon_i \overset{i.i.d.}{\sim} \text{normal}\left(0, s^2\right) \quad \text{for } i = 1, \ldots, 25$$

We can think of this as the null model when we are considering the null hypothesis $\beta_1 = 0$, plugging in our estimates of the remaining parameters β_0 and σ^2 which are unknown. The simulated values $Y_1^{(k)}, \ldots, Y_{25}^{(k)}$ represent alternative realizations of the response that we might have observed if $\beta_1 = 0$, e.g. responses we might have observed if the Toluca data were collected in another universe.

For each of the $k = 1, \ldots, 1,000$ synthetic datasets, we will perform a level-$\alpha = 0.05$ test of the null hypothesis that $\beta_1 = 0$, and record whether or not we reject the null hypothesis.

```
set.seed(1)
reject <- rep(NA, 1000)
for (i in 1:length(reject)) {
  Y.sim <- rnorm(n = 25, mean = 62.366, sd = 48.82)
  linmod <- lm(Y.sim~X)
  b1 <- summary(linmod)$coef["X", "Estimate"]
  sb1 <- summary(linmod)$coef["X", "Std. Error"]
  reject[i] <- !((b1 + qt(0.05/2, df = 34)*sb1) <= 0 &
                   (b1 + qt(1 - 0.05/2, df = 34)*sb1) >= 0)
}
```

If we look at the proportion of times we reject the null hypothesis that $\beta_1 = 0$, we get 0.049. This just about matches α, which makes sense, because α conveys how often we would expect to reject the null that $\beta_1 = 0$

across different realizations of data generated according to the null model with $\beta_1 = 0$

(b) Simulate $k = 1, \ldots, 1,000$ synthetic datasets $Y_1^{(k)}, \ldots, Y_{25}^{(k)}$ according to:

$$Y_i^{(k)} = b_0 + 3X_i + \epsilon_i^{(k)}, \quad \epsilon_i \overset{i.i.d.}{\sim} \text{normal}\left(0, s^2\right) \quad \text{for } i = 1, \ldots, 25$$

For each of the $k = 1, \ldots, 1,000$ synthetic datasets, perform a level-$\alpha = 0.05$ test of the null hypothesis that $\beta_1 = 3$, and record whether or not we reject the null hypothesis. In what proportion/percent of simulations do you reject the null? We call this the *level* of the test, it tells us how often a level-$\alpha = 0.05$ test would lead us to reject the null hypothesis that $\beta_1 = 3$ if the true value of β_1 were in fact 3.

(c) Simulate $k = 1, \ldots, 1,000$ synthetic datasets $Y_1^{(k)}, \ldots, Y_{25}^{(k)}$ according to:

$$Y_i^{(k)} = b_0 + 3.5X_i + \epsilon_i^{(k)}, \quad \epsilon_i \overset{i.i.d.}{\sim} \text{normal}\left(0, s^2\right) \quad \text{for } i = 1, \ldots, 25$$

For each of the $k = 1, \ldots, 1,000$ synthetic datasets, perform a level-$\alpha = 0.05$ test of the null hypothesis that $\beta_1 = 3$, and record whether or not we reject the null hypothesis. In what proportion/percent of simulations do you reject the null? We call proportion this the *power* of the test for $\beta_1 = 3.5$, it tells us how often a level-$\alpha = 0.05$ test would lead us to reject the null hypothesis that $\beta_1 = 3$ if the true value of β_1 were in fact 3.5.

(d) Simulate $k = 1, \ldots, 1,000$ synthetic datasets $Y_1^{(k)}, \ldots, Y_{25}^{(k)}$ according to:

$$Y_i^{(k)} = b_0 + 6X_i + \epsilon_i^{(k)}, \quad \epsilon_i \overset{i.i.d.}{\sim} \text{normal}\left(0, s^2\right) \quad \text{for } i = 1, \ldots, 25$$

For each of the $k = 1, \ldots, 1,000$ synthetic datasets, perform a level-$\alpha = 0.05$ test of the null hypothesis that $\beta_1 = 3$, and record whether or not we reject the null hypothesis. In what proportion/percent of simulations do you reject the null? We call proportion this a *power* of a test for $\beta_1 = 6$, it tells us how often a level-$\alpha = 0.05$ test would lead us to reject the null hypothesis that $\beta_1 = 3$ if the true value of β_1 were in fact 6.

(e) Explain how the results of (a) , (c) and (d) differ, and comment on how the power of the test of the null hypothesis that $\beta_1 = 3$ depends on the true value of β_1.

3. Problem 2.27 from the `.pdf` version of the textbook. Requires use of the `muscle` data that has been posted on the Homework page.

4. Problem 2.30 from the `.pdf` version of the textbook. Requires use of the `crime` data that has been posted on the Homework page.