MLDS 401 Homework 3
Due: Monday, October 16, 3:00
Professor Malthouse

Work in your assigned teams. Submit a single assignment with all names.

1. Use the auto data set from JWHT problem 3.9 on page 122.

   (a) Now regress `mpg` on `cylinders`, `displacement`, `weight`, and `year`. Comment on the signs of the estimated coefficients and note which are significantly different from 0. What is value of R^2?

   (b) Compute the variance inflation factors. What do they tell you?

   (c) Drop `weight` from the model. What happens to the parameter estimates and R^2?

   (d) Drop `weight` and `displacement` from the model. What happens to the parameter estimates and R^2?

   (e) For class discussion: Draw a DAG for weight, horsepower, displacement, cylinders and mpg.

2. In class I mentioned that you should control for a variable if it is a fork, but not if it is a pipe or collider. The next three problems will ask you to generate some data to see why. Let $n = 500$ be the number of observations. Generate $w \sim \mathcal{U}[0, 5]$. Then let $x = w + \delta$ where $\delta \sim \mathcal{N}(0, 1)$. Then let $y = 4 + 2x - 3w + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. **For all three problems, our interest is in understanding the effect of x on y, and w is another variable that you are considering including.**

   (a) Is this a pipe, fork or collider?

   (b) Generate a correlation matrix and basic descriptive statistics (min, max, mean, sd).

   (c) Regress y on x. Is the coefficient of x significant at the .05 level? Does a 95% CI cover the true slope for x, namely 2?

   (d) Now regress y on both x and w. Is the coefficient of x significant at the .05 level? Does a 95% CI cover the true slope for x, namely 2?

   (e) What are the values of VIF from this second regression?

3. Let $n = 500$ be the number of observations. Generate $x \sim \mathcal{U}[0, 5]$. Then let $y = x + \delta$ where $\delta \sim \mathcal{N}(0, 1)$. Then let $w = 4 + 2x + 3y + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$.

   (a) Is this a pipe, fork or collider?

   (b) Generate a correlation matrix and basic descriptive statistics (min, max, mean, sd).

   (c) Regress y on x. Is the coefficient of x significant at the .05 level? Does a 95% CI cover the true slope for x, namely 1?

(d) Now regress y on both x and w. Is the coefficient of x significant at the .05 level? Does a 95% CI cover the true slope for x, namely 1?

(e) What are the values of VIF from this second regression?

(f) Compare the values of R^2 (or S_e) between the two models. Which model is better according or R^2 (or S_e). Is this the right model?

4. Let $n = 500$ be the number of observations. Generate $x \sim \mathcal{U}[0, 5]$. Then let $w = x + \delta$ where $\delta \sim \mathcal{N}(0, 1)$. Then let $y = 2w + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$.

   (a) Is this a pipe, fork or collider?

   (b) Generate a correlation matrix and basic descriptive statistics (min, max, mean, sd).

   (c) Regress y on x. Is the coefficient of x significant at the .05 level?

   (d) Now regress y on both x and w. Is the coefficient of x significant at the .05 level?

   (e) Which model has a better R^2? (You could also compare adjusted R^2, although we have not defined it yet.)

5. (12 points) JWHT problem 3.14a–f on page 125.

   (a) Generate the data as shown below, but for those working in Python, I will translate the data generation process. Generate $n = 100$ points where $x_1 \sim \mathcal{U}[0, 1]$ (`runif(100)`) and `rnorm(100)` produces 100 $\mathcal{N}(0, 1)$ variables.

   ```
   > set.seed(1)
   > x1 = runif(100) # part a
   > x2 = 0.5*x1 + rnorm(100)/10
   > y = 2 + 2*x1 + .3*x2 + rnorm(100)
   ```

   Write out the form of the linear model. What are the true regression coefficents and standard deviation of errors.

   (b) What is the correlation between x_1 and x_2?

   (c) Using this data, regress y on x_1 and x_2. Describe the results. What are the parameter estimates and how do they relate to the true parameters? Which coefficients are significantly different from 0? Are the true parameters “covered” by 95% CIs?

   (d) Regress y on x_1 alone. Is β_1 significantly different from 0? Is the true β_1 covered by by a 95% CI?

   (e) Regress y on x_2 alone. Is β_2 significantly different from 0? Is the true β_2 covered by by a 95% CI? Are the true parameters “covered” by the 95% confidence intervals?

```
> fit = lm(y~x2)   # Part e
> summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
x2            2.8996     0.6330    4.58 1.37e-05 ***
> confint(fit)
               2.5 %   97.5 %
(Intercept) 2.003116 2.776783
x2          1.643324 4.155846
```

(f) Do the results in (c)–(e) contradict each other? Explain.

6. (10 points) ACT 2.4: Omitted variables. Suppose that there are two predictor variables, x_1 and x_2, but we fit the straight line model $y = \beta_0 + \beta_1 x_1 + \epsilon$ omitting x_2. If, in fact, the true model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, show that

$$\mathbb{E}(\hat{\beta}_1) = \beta_1 + \beta_2 \sum_{i=1}^{n} c_i x_{i2} = \beta_1 + \frac{\beta_2}{S_{11}} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = \beta_1 + \beta_2 r \frac{s_2}{s_1},$$

where $c_i = (x_{i1} - \bar{x}_1)/S_{11}$, $S_{11} = \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2$, r is the sample correlation coefficient between x_1 and x_2, and $s_1 > 0$, $s_2 > 0$ are the sample SD's of x_1, x_2, respectively. Thus $\hat{\beta}_1$ is biased with the bias given by $\beta_2 r s_2/s_1$. Under what condition is this bias zero? Discuss how this result applies to JWHT 3.14 that you worked in problem 5.

7. In class I mentioned that one of the effects of multicollinearity is to increase the variance of the slope estimates. This problem will show you why this is the case. Suppose that $y = \beta_1 z_1 + \beta_2 z_2 + e$, where all variables have been standardized prior to estimation making the intercept unnecessary. You have observed (z_{i1}, z_{i2}, y_i), for $i = 1, \ldots, n$, where the sample means are 0 and the sample variances are 1, i.e.,

$$\bar{z}_j = \frac{1}{n} \sum_{i=1}^{n} z_{ij} = 0, \quad S_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (z_{ij} - 0)^2 = 1, \quad \text{and} \quad r = \frac{1}{n-1} \sum_{i=1}^{n} z_{i1} z_{i2},$$

where r is sample correlation between z_1 and z_2. How does $\text{VIF} = 1/(1 - r^2)$ affect the variance of the slope estimates? Hint: recall that $V(\mathbf{b}) = \sigma^2 (\mathbf{Z}^\mathsf{T} \mathbf{Z})^{-1}$ and

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

8. Start analyzing the bike data. There is nothing to turn in.