MSiA 401: Homework 2
Due: October 12, 3pm
Professor Malthouse

Work in your assigned teams. Turn in one copy of your answers. All group members must put their name on the homework.

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

   (a) The `origin` variable is categorical, where 1=US, 2=Europe and 3=Japan. Type the following command to make it a factor variable and assign meaningful labels:

   ```
   auto$origin = factor(auto$origin, 1:3, c("US", "Europe", "Japan"))
   ```

   Submit a `table` of the variable (i.e., frequency distribution).

   (b) Regress `mpg` on `origin`, `weight` and `year`. Examine the diagnostic plots and comment on which assumptions of the linear model, if any, are violated.

   (c) Regress `log(mpg)` on `origin`, `log(weight)`, `year` and `year` squared. Examine the diagnostic plots and the summary. Comment on whether the model assumptions are roughly satisfied.

   (d) Use the results from the previous part to describe the effect of `year` on `log(mpg)`, i.e., is it U-shaped, inverted-U shaped, or linear? If it is nonlinear, where is the minimum or maximum. Submit a graph showing the effect. For you to think about but not turn in: why would `year` have this effect?

   (e) What does the coefficient for log(weight) tell you? How is unlogged mpg related to unlogged weight?

2. Consider the following model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i, \quad i = 1, \ldots, n.$$

   This will be called the *uncentered* model. We discussed in class how, when $x \geq 0$, x and x^2 are often (highly) correlated. One way to reduce the correlation, yet have an equivalent model, is to mean center the x variables prior to estimation, i.e., let $\bar{x}$ be the mean of x. Let $\tilde{x}_i = x_i - \bar{x}$, then regress y on $\tilde{x}$ and $\tilde{x}^2$, i.e.,

$$y_i = \gamma_0 + \gamma_1 \tilde{x}_i + \gamma_2 \tilde{x}_i^2 + e_i = \gamma_0 + \gamma_1 (x_i - \bar{x}) + \gamma_2 (x_i - \bar{x})^2 + e_i,$$

   where γ_j are coefficients for the *centered* model. This problem will show you how the two models are equivalent to each other.

   (a) Write β_0, β_1 and β_2 as a functions of the γ_j's and $\bar{x}$. Hint: start with the second expression above, distribute γ_1 across $(x_i - \bar{x})$ and γ_2 across the expanded square. Collect the terms and set β_2 to the coefficient of x^2, β_1 equal to the coefficient of x, and β_0 equal to anything that does not include an x.

(b) You will now check your work with the auto data set. Regress `mpg` on `year` and `year`^{2}. Note the coefficients.

(c) What is the correlation between `year` and `year`^{2}?

(d) What is the mean of year?

(e) What is the correlation between `year` and `year`^{2} after mean centering?

(f) To understand why the correlation is reduced, plot year squared against year, and separately centered year squared versus centered year.

(g) Regress `mpg` on centered year and centered year squared. Note the coefficients.

(h) Substitute your estimates from the previous part into the expressions you derived in part (a) and show that the equal the estimates from part (b).

3. (8 points) ACT 3.2: Hat matrix for simple linear regression.

4. (15 points) ACT 3.3: Hat matrix for the Anscombe Data Set IV.

5. (6 points) ACT 3.6: Mean and covariance of the GLS estimator.

6. (8 points) Let Y_1 and Y_2 be independent random variables with $\mathbb{E}(Y_i) = \mu$ and $\mathbb{V}(Y_i) = \sigma_i^2$ $(i = 1, 2)$. Consider estimates of μ having the form $\bar{y}_w = w_1Y_1 + w_2Y_2$, where w_1 and w_2 are non-negative constants.

   (a) Under what circumstances will $\bar{y}_w$ be an unbiased estimator of μ?

   (b) What is the variance of $\bar{y}_w$?

   (c) Among all unbiased estimates, show that $\mathbb{V}(\bar{y}_w)$ is minimized when $w_i \propto 1/\sigma_i^2$.

7. (6 points) How does the type and amount of crime around a Divvy bike station affect the **demand** for bikes, as measured by the number of rentals per time period? We also want to assess how other independent variables are related to demand. I will be providing you with a data set giving the demand at $n = 300$ bike stations, but I want you to think about what results you would expect before looking at the data. I do **not** expect you to read other research articles. Instead, I want you to think about what could happen in the model and explanations for why. Your data will have data on how often 31 different crimes occurred in the area around the bike share station during the previous year—I have lagged the crime data to avoid problems with reverse causality. Many of the crimes are rare, so we will focus on the following eight (you could use the other if you want): theft, battery, deceptive practice, assault, burglary, robbery, criminal trespassing, narcotics, and homicide. You might want to google these terms for more precise definitions. For example, my understanding is that assault involves a threat, but not bodily harm, while battery implies harm. Deceptive practice is sometimes called fraud, and a examples include passing bad checks or trying to withdraw money from the bank as someone else. You will also have data on: number

of bus stops in the area, number of train stops in the area, station capacity (number of bikes), number of marked bike routes, number of businesses in the area, population density, park area, percent minority residents, average education level, and average per capita income. For this problem, I want you to develop a theory to explain how different types of crime will affect demand. It may be that some types of crime have no relationship with demand, and your theory should allow for this. **Why** might some types have an association and others not? Another consideration is that you have actual crime statistics from the Chicago Police Department instead of *perceptions* about crime. The two could be different (why?). Your next assignment will be to test your theories against the data.