MSiA 401: Homework 2
Due: October 6, 5:00pm
Professor Malthouse

You may work in homework groups. Turn in one copy per group, with all names on it.

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). *Answer: Here's my answer*

   ```
   > table(auto$origin)
      US Europe  Japan
     248     70     79
   ```

   (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.

   ```
   > fit = lm(mpg ~ origin + weight + year, auto)
   > plot(fit, which=1)
   ```

   Answer: The residual plot shows a pattern, with heteroscedasticity and consistently positive values, then negative, then positive again. This indicates that the fit can be improved and we will have to transform both sides. The errors are also non-normal.

   (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. *Answer: Answer: The assumptions are better satisfied. There might be a little lack of fit requiring futher transformations of x variables, but the heteroscedasticity has been addressed. There are some issue with the tails not being completely normal.*

   ```
   > fit = lm(log(mpg) ~ origin + log(weight) + year + I(year^2), data=auto)
   > summary(fit)
                  Estimate Std. Error t value Pr(>|t|)
   (Intercept)  18.4693014  2.6833895   6.883 2.34e-11 ***
   originEurope  0.0668291  0.0176293   3.791 0.000174 ***
   originJapan   0.0319711  0.0179382   1.782 0.075477 .
   log(weight)  -0.8750305  0.0270390 -32.362  < 2e-16 ***
   year         -0.2559684  0.0712094  -3.595 0.000366 ***
   I(year^2)     0.0019051  0.0004687   4.065 5.81e-05 ***
   ```

```
Residual standard error: 0.1136 on 391 degrees of freedom
Multiple R-squared:  0.8898,Adjusted R-squared:  0.8884
F-statistic: 631.7 on 5 and 391 DF,  p-value: < 2.2e-16
```

(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? *Answer: The quadradic coefficient* $0.0019 > 0$ *indicates a U shape, with a min value at* $0.2560/(2 \times 0.001905) = 67.18$*, i.e., around 1967. To make a graph, note that year ranges from 70 to 82 and do something like the code below to see the effect. We see a flatter effect in the early 1970s that becomes steeper over time. Why would this happen? There was an energy crisis in the early 70s that caused to car manufacturers to create more fuel-efficient cars. I suspect the increasing slope is due to the crisis. See effects.*

(e) What does the coefficient for `log(weight)` tell you? How is unlogged mpg related to unlogged weight? *Answer: The coefficient is negative, which indicates that as the weight of the car increases the milage decreases on the average after controlling for year and origin. If we take the exponent of both sides and hold origin and year constant,* $mpg \propto 1/wt^{0.88}$ *and so weight is roughly inversely proportional to mpg.*

2. (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 the form $\bar{y}_w = w_1 Y_1 + w_2 Y_2$, where w_1 and w_2 are non-negative constants.

(a) Under what circumstances will $\bar{y}_w$ be an unbiased estimator of μ? *Answer:* $w_1 + w_2 = 1$*. We can rewrite* $w_1 = w$ *and* $w_2 = 1 - w$*. Then*

$$\mathbb{E}(\bar{y}_w) = \mathbb{E}[wY_1 + (1-w)Y_2] = w\mu + (1-w)\mu = \mu$$

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

$$\mathbb{V}(\bar{y}_w) = \mathbb{V}[wY_1 + (1-w)Y_2] = w^2\sigma_1^2 + (1-w)^2\sigma_2^2$$

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

$$\frac{d\mathbb{V}(\bar{y}_w)}{dw} = 2w\sigma_1^2 - 2(1-w)\sigma_2^2 \quad \text{and} \quad \frac{d^2\mathbb{V}(\bar{y}_w)}{dw^2} = 2\sigma_1^2 + 2\sigma_2^2 > 0$$

If we set the first derivative equal to 0 and solve we get

$$w = w_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \quad \text{and} \quad w_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}.$$

Since the second derivative is positive this is a minimum. To show $w_i \propto 1/\sigma_i^2$, we must find constant k so that $w_i = k/\sigma_i^2$, which is true when $k = \sigma_1^2\sigma_2^2/(\sigma_1^2 + \sigma_2^2)$.

3. 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. *Answer:*

$$y = \underbrace{(\gamma_0 - \gamma_1\bar{x} + \gamma_2\bar{x}^2)}_{\beta_0} + \underbrace{(\gamma_1 - 2\bar{x}\gamma_2)}_{\beta_1} x_i + \underbrace{\gamma_2}_{\beta_2} x_i^2$$

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

```
> fit = lm(mpg ~ year + I(year^2), auto)
> fit$coefficients
(Intercept)        year   I(year^2)
577.2522975 -15.8409008   0.1123014
```

(c) What is the correlation between `year` and `year` squared?

```
> cor(auto$year, auto$year^2)
[1] 0.999759
```

(d) What is the mean of `year`?

```
> (xbar=mean(auto$year))
[1] 75.99496
```

(e) What is the correlation between `year` and `year` squared after mean centering?

```
> auto$yearcent = auto$year - mean(auto$year)
> cor(auto$yearcent, auto$yearcent^2)
[1] 0.014414
```

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

```
> plot(auto$yearcent, auto$yearcent^2)
> plot(auto$weight, auto$weight^2)
```

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

```
> fit2 = lm(mpg ~ yearcent + I(yearcent^2), auto)
> fit2$coefficients
  (Intercept)      yearcent I(yearcent^2)
   21.9906094     1.2277824     0.1123014
```

(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).

```
> g=fit2$coefficients
> g[1]-g[2]*xbar+g[3]*xbar^2 # equals beta_0 from above
(Intercept)
   577.2523

> g[2]-2*xbar*g[3]  # equals beta_1 from above
     yearcent
-15.8409

> g[3]  # equals beta_2 from above
 I(yearcent^2)
    0.1123014
```

4. (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 an example is 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.