Homework Assignment 2

Max H. Farrell – Chicago Booth
BUS41100 Applied Regression Analysis

Due at the beginning of class of week 3

1 Understanding the Simple Linear Regression Model

Assume the following model: $Y_i = 10.0 + 0.5X_i + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1)$

(a) $\mathbb{E}[Y|X=0] = ?, \quad \mathbb{E}[Y|X=-1] = ?, \quad \text{var}[Y|X] = ?$

(b) What is the probability of $Y > 10$, given $X = 2$?

(c) If X has a mean of zero and variance of 20, what are $\mathbb{E}[Y]$ and $\text{Var}(Y)$?

(d) What is $\text{Cov}(X, Y)$?

2 Simulation from the SLR Model

Use the `rnorm` function in R to generate $n = 100$ samples of $X \sim N(0, \sigma_X^2)$, with $\sigma_X^2 = 2$ (for help use `?rnorm`). For each draw, simulate Y_i from the simple linear regression model $Y_i = 2.5 - 1.0X_i + \epsilon_i$, where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$, with $\sigma_\epsilon^2 = 3$.

(a) Show the scatter plot of Y versus X along with the true regression line.

(b) Split the sample into 2 subsets of size 25 and 75. For each subset, run the regression of Y on X. Add each fitted regression line (use color) to your plot from **(a)**. Why are they not the same?

(c) What is the marginal sample mean for Y? What is the true marginal mean?

(d) Start a fresh scatter plot of Y versus X and add the true regression line and the estimated version (using the full sample).

- **(i)** Add the bounds of the 90% prediction interval to your plot. *Coding hint:*

```
pi <- predict(reg, interval="prediction", level=0.90)
lines(sort(x), pi[order(x), "lwr"], lwd=2, col="red", lty=2)
```

  What is going on with sort(x) and order(x)?

- **(ii)** What percentage of your observations are outside of this interval? *Coding hint: you can test TRUE/FALSE statements like this* `(y > pi[,"upr"])`. *Combine with the same thing for the lower bound.*

- **(iii)** Add the bounds of the *true* 90% prediction interval to your plot. This is the interval that assumes you know the true β_0, β_1, σ_X^2, and σ_ϵ^2 and don't have to use estimates. Thus, estimation of these won't factor into the uncertainty of $\hat{Y}$. (use `?qnorm` for help with quantiles and `lty=2` in `abline()` or `lines()` to get a dashed line).

- **(iv)** What percentage of your observations are outside of this *true* interval?

(e) Repeat part **(d)** for different values of n, σ_X^2, and σ_ϵ^2. What do you learn? What effect do these values have?

3 Maintenance Costs

The cost of the maintenance of a certain type of tractor seems to increase with age. The file `tractor.csv` contains ages (years) and 6-monthly maintenance costs for $n = 17$ such tractors.

(a) Create a plot of tractor maintenance `cost` versus `age`.

(b) Find the least squares fit to the model

$$\texttt{cost}_i = b_0 + b_1\texttt{age}_i + e_i$$

in two ways: first using the `lm` command and second by calculating a correlation and standard deviations [verify that the answers are identical]. Add the fitted line to the scatterplot.

(c) Suppose you were considering buying a tractor that is three years old, what would you expect your six-monthly maintenance costs to be? What is the 95% predictive (cost) interval for the six-monthly maintenance of your tractor? Compare the endpoints of the interval to the observed values of `cost`. What do you conclude about your prediction from this? Why or why not is this conclusion surprising?

4 Broadway Box Office

Let X and Y denote the weekly reports on the box office ticket sales for plays on Broadway in New York for two consecutive weeks, respectively, in October 2004. (You can actually download similar data from `www.playbill.com`.) The regression output for this data set in shown in the table below:

Variable	Coefficient	s.e.	t-value	p-value
Intercept	6805	9929	0.685	0.503
X	0.9821	0.01443	68.071	$< 2 \times 10^{-16}$
	$n = 18$	$R^2 = 0.9966$	$s_\varepsilon = 18007.56$	

Suppose that the model satisfies the usual SLR model assumptions, and that the SST for Y is 1.507773×10^{12}.

(a) What were the degrees of freedom used in calculating s_ε? What are the SSE and SSR?

(b) Compute the sample variance for Y (s_Y^2) and sample correlation between X and Y (r_{XY}).

(c) Suppose that the ticket sales in the first week for a particular play was $822,000. What is the expected sales for the same play in the following week?

(d) Suppose further that $\bar{X} = 822186.6$ and $s_X = 302724.5$. Construct the 95% forecast interval for the estimate in **(c)**.

(e) Construct the 95% confidence interval for the slope of the true regression line β_1

(f) Some Broadway plays use the rule of thumb that next week's gross box office results will be the same as this week's. Is this reasonable? (Justify/Refute using an appropriate hypothesis test.)

(g) If Y and X were reversed in the above regression, what would you expect R^2 to be? (*Read ahead a little! R^2 is discussed in the week 4 slides.*)

5 Topic: Measurement Error

This question illustrates conceptual material, and thus it has extra exposition.

In class so far (and most of the rest of class too!), we have taken for granted that whatever that the X and Y variables in our data set were measured perfectly. That is, they really contained the information they were supposed to. But this is not always the case! Think of our favorite example: house prices. It is reasonable to assume that the `price` variable really was the sale price, but do you think `square feet` is perfectly measured for any house? Of course not. In this problem, we will explore this topic of *measurement error* (sometimes called *errors-in-variables*). We will only look at the simplest, easiest type of measurement error, so the problem is usually worse than this in real life!

We have an outcome Y and a predictor X, related through the equation (nothing different yet):

$$Y = \beta_0 + \beta_1 X + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

The parameters and predictions that we actually care about for decision-making are *always* those from this model.

But what if X and/or Y are *not* perfectly measured? Instead of the true X and Y, we will imagine we have data on one or both of $\tilde{X}$ and $\tilde{Y}$, which are just X and Y corrupted by Normally distributed errors:

$$\tilde{X} = X + \epsilon_x, \qquad \tilde{Y} = Y + \epsilon_y, \qquad \text{where} \quad \epsilon_x \sim \mathcal{N}(0, \delta_x^2) \quad \text{and} \quad \epsilon_y \sim \mathcal{N}(0, \delta_y^2). \tag{2}$$

In words, instead of the true variable(s) we instead have the true variable(s) *plus* an idiosyncratic error term that corrupts the truth up or down. As usual, we will assume that the errors are independent, and in this case, we assume they are independent of all other variables. (Is this realistic? For example, do you think the error in measuring square footage is the same for all sizes of house?) This is called *classical* measurement error, and it's far from the only type, but it's all we'll deal with in this class.

(a) Suppose that there is no measurement error at all, so we have n i.i.d. observations (Y_i, X_i), like usual. Use the code file `homework2-MeasurementErrorMonteCarlo.R`, set $\delta_x^2 = \delta_y^2 = 0$, and verify that b_1 estimated with this data has all the nice properties found in class.

(b) Suppose that only Y is measured with error, so we have n i.i.d. observations $(\tilde{Y}_i, X_i)$. We will based our estimation on the model:

$$\tilde{Y} = \tilde{\beta}_0 + \tilde{\beta}_1 X + \tilde{\varepsilon}, \qquad \tilde{\varepsilon} \sim \mathcal{N}(0, \tilde{\sigma}^2), \tag{3}$$

instead of (1). Using this model and the data on $(\tilde{Y}_i, X_i)$ we form the estimator $\tilde{b}_1$ of $\tilde{\beta}_1$. Remember, β_1 from (1) is what we care about at the end of the day, not $\tilde{\beta}_1$. That is, we assume throughout that (1) is true.

(i) Recall from slide 47 of week 1 that $\beta_1 = \text{cov}(X, Y) \ / \ \text{var}(X)$. Following the same steps, show that $\tilde{\beta}_1 = \text{cov}(X, \tilde{Y}) \ / \ \text{var}(X)$. Then compute $\text{cov}(X, \tilde{Y})$ by plugging in $\tilde{Y} = Y + \epsilon_y$ to find a relationship between $\tilde{\beta}_1$ and β_1. What does this tell you about the value of $\tilde{b}_1$ as an estimator of β_1?

(ii) Use (1) to compute the variance of Y given X. Use (1) and (2) to compute the variance of $\tilde{Y}$ given X (equivalently, compute $\tilde{\sigma}^2$ in terms of other parameters). Compare the two. What does this tell you about the standard error of $\tilde{b}_1$ as an estimator of β_1?

(iii) Use the code `homework2-MeasurementErrorMonteCarlo.R` with $\delta_x^2 = 0$ and describe what happens to the sampling distribution of $\tilde{b}_1$ as δ_y^2 increases, and discuss how this relates to what you found in **(b)(i)** and **(b)(ii)**.

(c) Suppose that only X is measured with error, so we have n i.i.d. observations $(Y_i, \tilde{X}_i)$. We will based our estimation on the model:

$$Y = \tilde{\tilde{\beta}}_0 + \tilde{\tilde{\beta}}_1 \tilde{X} + \tilde{\tilde{\varepsilon}}, \qquad \tilde{\tilde{\varepsilon}} \sim \mathcal{N}(0, \tilde{\tilde{\sigma}}^2). \tag{4}$$

instead of (1). Using this model and the data on $(Y_i, \tilde{X}_i)$ we form the estimator $\tilde{\tilde{b}}_1$ of $\tilde{\tilde{\beta}}_1$. Remember, β_1 from (1) is what we care about at the end of the day, not $\tilde{\tilde{\beta}}_1$.

- **(i)** Recall from slide 47 of week 1 that $\beta_1 = \text{cov}(X, Y) \;/\; \text{var}(X)$. Following the same steps, show that $\tilde{\tilde{\beta}}_1 = \text{cov}(\tilde{X}, Y) \;/\; \text{var}(\tilde{X})$. Then compute $\text{cov}(\tilde{X}, Y)$ and $\text{var}(\tilde{X})$ by plugging in $\tilde{X} = X + \epsilon_x$ to find a relationship between $\tilde{\tilde{\beta}}_1$ and β_1. What does this tell you about the value of $\tilde{\tilde{b}}_1$ as an estimator of β_1?
- **(ii)** Use (1) to compute the variance of Y given X. Use (1) and (2) to compute the variance of Y given $\tilde{X}$ (equivalently, compute $\tilde{\tilde{\sigma}}^2$ in terms of other parameters). Compare the two. What does this tell you about the standard error of $\tilde{\tilde{b}}_1$ as an estimator of β_1?
- **(iii)** Use the code `homework2-MeasurementErrorMonteCarlo.R` with $\delta_y^2 = 0$ and describe what happens to the sampling distribution of $\tilde{\tilde{b}}_1$ as δ_x^2 increases, and discuss how this relates to what you found in **(c)(i)** and **(c)(ii)**. Pick a fixed, positive value of δ_x^2, (e.g. $\delta_x^2 = 1/2$), and describe what happens as n increases.

(d) Edit the code `homework2-MeasurementErrorMonteCarlo.R` to study the properties of prediction intervals in the presence of measurement error in Y (as in **(b)**) and in X (as in **(c)**).

(e) *[Nothing to turn here.]* Experiment with the code `homework2-MeasurementErrorMonteCarlo.R` to get a feel for what happens with different combinations of δ_x^2, δ_y^2, n, σ^2, β_1, and β_2.

(f) *[Nothing to turn in here either, just concluding the educational part!]* Everything we have done above applies to predicted values, prediction intervals, and anything else that comes from a simple linear regression.

But this changes when we move to multiple linear regression, introduced in week 3. When we have more than one X variable, and some are measured with error, the direction of the bias is unknown. Who knows what happens in situations like logistic regression.

(g) Now let's see what happens in real data. Researchers at General Motors collected data on 60 U.S. Standard Metropolitan Statistical Areas (SMSA's) in a study of whether or not air pollution contributes to mortality; see `smsa.csv`. The dependent variable for analysis is age adjusted `Mortality` rate and the explanatory variable is `HCPot`, the HydroCarbons pollution potential index.

- **(i)** Is it reasonable that `Mortality` and/or `HCPot` are measured with error?
- **(ii)** Interpret the results of a regression of `Mortality` on `HCPot` in light of your answer.
- **(iii)** Do you think the measurement error in this case is classical (independent of everything) or not? State your reasoning.