Homework Assignment 2

Panagiotis (Panos) Toulis – Chicago Booth
BUS41100 Applied Regression Analysis

1 Basics of Simple Linear Regression Model

Consider the SLR model: $Y = 10.0 + 0.5X + \varepsilon,\ \varepsilon \sim N(0, 1)$. and answer the following questions.

(a) What is β_0, β_1 and σ^2?

(b) What is the expected value of Y if we fix $X = 0$?

(c) What is the probability of $Y > 10$ if we fix $X = 2$? (*Hint:* You could simulate it.)

2 Salary Data

We would like to understand the relationship between years of experience in a job and salary. The file `Salary_dataset.csv` contains a random sample of 30 workers for the City of Chicago. We want to do inference using the SLR model:

$$\texttt{Salary} = \beta_0 + \beta_1 \texttt{YearsExperience} + \varepsilon.$$

(a) Load the data and plot `Salary` against `YearsExperience` as a scatterplot. Do you think the linear model is plausible for these data?

(b) Give an intuitive interpretation of the β_1 parameter in the SLR model.

(c) Fit the SLR model using `lm()` and test whether β_1 is statistically significant. Also, produce a 95% confidence interval for this parameter. Do you conclude that the years of experience have a strong correlation with salary?

(d) Suppose that a friend of yours has been working for 5 years with annual pay around \$60,000. However, your friend is thinking of quitting. Try to convince your friend of staying for another 5 years by supplying the appropriate 95% prediction interval for the estimated salary.

3 Sampling Distributions

Start with the following `R` code:

```
beta0 = 2
beta1 = 5
sigma2 = 4
X = rnorm(100)
```

As in class, we will pretend that we don't know the true parameter values.

(a) Simulate 100 values for Y from the linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Calculate the least squares estimates b_0 and b_1 of the true parameters β_0, β_1, respectively. Also get s^2, the estimate of σ^2 from `R`'s output. Are the estimates close to the true values? Why are the estimates not exactly the same as the true values?

(b) With fixed X vector, generate the sampling distribution of b_1 and plot its histogram. Verify that the center of this histogram is around 5 and briefly explain why this happens.

(c) From (b) plot the histogram of $(b_1 - \beta_1)/s_1$ where s_1 is the standard error of b_1 from `R`'s output. Show that this is very similar to the standard normal distribution.

4 The Case of Zillow Offers

Here, we consider the case of Zillow Offers that used complex predictive analytics to flip houses for profit, but unfortunately ended in a debacle.[1]

(a) Load the full Ames House dataset and plot `Price` against `Quality`. Then, consider the SLR model

$$\texttt{Price} = \beta_0 + \beta_1 \texttt{Quality} + \varepsilon.$$

Test whether $\beta_1 = 0$. What is a plausible range for the monetary value of a "one-unit of house quality" ?

(b) Now, let's focus on a subpopulation of "fixer-uppers", houses of low quality we would like to renovate and sell for profit. Define this subpopulation as follows.

```
fixer_uppers = house %>% filter(Quality <= 5) %>% select(Quality, Price)
```

How many houses are in this subpopulation? What is their average quality?

(c) Suppose we renovated these houses and increased their `Quality` by 3 units:

```
Xf = fixer_uppers %>% select(Quality) %>% mutate(Quality=Quality+3)  # quality+ 3
P = fixer_uppers$Price                                               # original prices
P_new = predict(..., newdata=Xf, interval="prediction")             # new prices
```

Fill in the '...' above to produce a 95% prediction interval for the prices of houses in `Xf`.

(d) In (c) , `P` is the cost of buying and `P_new` contains estimates of the sell prices after renovation.

- Use `sum(P_new[,1] - P)` to estimate the total profit from selling after renovation.
- Also, use `sum(P_new[,2] - P)` for the same calculation. Why is this a *conservative* estimate of the total profit from selling after renovation?

(e) An economist in your team suggests that the above model is too simplistic to predict profit because it does not take labor and selling costs into account. The economist proposes the following model:

- After we buy the house, we need to wait `T_wait` (months) to find a crew to complete the renovation.
- A crew finishes the work in `T_labor` months with a cost of `$labor_cost` per month.
- After renovation, it takes `T_sell` months for the house to be sold.
- The maintenance cost of a house is `$maint_cost` per month (incl. repairs, taxes, etc). The property "sits idle" for `T_wait + T_labor + T_sell` months from the time of purchase.

[1] https://www.wsj.com/articles/zillow-sells-2-000-homes-in-dismantling-its-house-flipping-business-11636545601

After market research, we come up with the following numbers:

```
T_wait = 0.1 # wait time to find labor
T_labor = 2 # 2 months of work
labor_cost = 10000 # cost of labor/mon
maint_cost = 300 # maintenance cost/mon
T_sell = 6 # 6 months to sell
```

Argue why the following code gives revised profit estimates according to the economist's model:

```
P_new[,1] - labor_cost*T_labor - maint_cost*(T_wait + T_sell + T_labor) - P
```

What is the total revised profit? How does it compare to the estimate from (d)? Also, calculate the conservative total profit estimate from:

```
P_new[,2] - labor_cost*T_labor - maint_cost*(T_wait + T_sell + T_labor) - P
```

(f) The model in (e) does not consider frictions in the labor market. Let's assume that `L` $= 5$ construction crews in the region and we assign our properties sequentially to the available `L` crews. Then, the first `L` properties would have to wait for `T_wait` to be assigned to labor, the next `L` would wait for `2T_wait + T_labor`, the next `L` for `3T_wait + 2 T_labor` , and so on.. Consider the following code:

```
L = 5    # Total crews.
T_wait_new = c()
cnt = 1
while(length(T_wait_new) < length(P) ) {
  T_wait_new = c(T_wait_new, cnt*rep(T_wait, L) + (cnt-1)*rep(T_labor, L) )
  cnt = cnt + 1
}
T_wait_new = head(T_wait_new, length(P))
```

Confirm that `T_wait_new` has the correct waiting times as described above. Then, calculate the revised profits from (e) but using the new waiting times as follows:

```
P_new[,1] - labor_cost*T_labor - maint_cost*(T_wait_new + T_sell + T_labor) - P
```

Is the total profit positive? Consider also the conservative profit estimates:

```
P_new[,2] - labor_cost*T_labor - maint_cost*(T_wait_new + T_sell + T_labor) - P
```

What is the conservative estimate of the total profit? Do you see something concerning?

(g) Argue why `T_sell` may also not be the same across all properties, and argue that this could lead to further reduction of estimated profits.

(h) Based on this exercise, discuss briefly what went wrong for Zillow Offers. Try to relate to anything else you can find about this case from online sources.

5 Hubble's Law and Bootstrap

According to Wikipedia,[2] Hubble's law is the observation in physical cosmology that galaxies are moving away from Earth at speeds proportional to their distance. In other words, the farther they are, the faster they are moving away from Earth. Hubble's law is considered the first observational

[2] `https://en.wikipedia.org/wiki/Hubble%27s_law`

basis for the expansion of the universe, and serves as one of the pieces of evidence in support of the Big Bang model.

In this example, we will use a real dataset containing the `velocity` of various galaxies and their `distance` to Earth. Answer the following questions.

(a) Load the file `hubble.csv`. Fit an appropriate SLR model and provide statistical evidence in support of Hubble's law.

(b) Produce the bootstrap 95% CI for the slope of `velocity` with respect to `distance` by completing the dots '...' below.

```
astro = read.csv("hubble.csv")
sampl.boot = replicate(5000, {
  I = sample(1:nrow(astro), size=nrow(astro), replace=T)
  astro_new = astro[I,] # bootstrapped dataset
  fit = ...
  coef(fit)[...]
})

se = sd(...) # bootstrap standard error
76 + 2*c(-1,1)*se # Bootstrap CI.
```

Is this similar to the 95% CI obtained from (a) using `confint`?

6 Knowledge Transfer

The point of these exercises is to train you in transferring your 41100 knowledge to domains that we haven't seen in class. Below is a regression output from a simple linear regression model in `Stata`, another popular statistical software.

cholesterol	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
time_tv	.0440691	.0105434	4.18	0.000	.0231461	.0649921
_cons	-2.134777	1.813099	-1.18	0.242	-5.732812	1.463259

(a) What do you think is the regression model here, and what is the scientific question behind it?

(b) What is the p-value for the intercept? Is it significant? Is it one-sided or two-sided?

(c) Is `time_tv` a significant variable? What does this mean in relation to (a)?