Homework Assignment 3

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

1 House prices

Load the file `Ames_house41100.csv` and run MLR of `Price` on all other variables using the commands:

```
house = read.csv("Ames_house41100.csv")
ols = lm(Price ~ ., data=house)
```

(*Note:* Meta-data description for this dataset are available online; e.g., `https://github.com/at-tan/Cracking_Ames_Housing_OLS/blob/master/Data/data%20description.txt`).

(a) Which variables are significant at the 5% level? By looking at the magnitude of the coefficients, is `Quality` or `Condition` more important for the price of a house?

(b) What are the `LocationXX` terms in the `summary` of the `ols` model? What is the baseline value for `Location`? Do they correspond to different intercepts or different slopes?

(c) We want to check whether being adjacent to something "positive" (e.g., a school or park), captured in the `PosA` variable, adds value to the house relative to the baseline. State the relevant null hypothesis and test it.

(d) The results from (a) and (c) are rather surprising, and we want to investigate further. A colleague suggests to interact `Condition` with `Location` through:

```
ols2 = lm(Price ~ . + Location*Condition, data=house)
```

(i) What has changed now with respect to (a) and the coefficients of `Condition` and `Quality`?

(ii) Is `PosA` significant with respect to the baseline? i.e., is your answer from (c) now different?

(iii) Argue why `Price` is more strongly correlated with `Condition` for those houses with `Location = PosA` compared to the baseline. State the relevant hypothesis and test it.

(iv) You remembered to visualize! For that you run:

```
par(mfrow=c(2, 1))
d1 = subset(house, Location=="Artery")
plot(Price ~ Condition, data=d1, pch=20)
d2 = subset(house, Location=="PosA")
plot(Price ~ Condition, data=d2, pch=20)
```

Show visual evidence to confirm your findings at (d)(iii).

2 Infant Nutrition

This question involves data from a study on the "nutrition of infants and preschool children in the north central region of the United States of America".[1] It is available on the course web page as

[1] by E.S. Eppright, H.M. Fox, B.A. Fryer, G.H. Lamkin, V.M. Vivian and E.S. Fuller in *World Review of Nutrition and Dietetics*, **14**, 1972, pp. 269–332.

nutrition.csv, and contains 72 observations of boys' weight/height ratio (woh) for equally spaced values of age in months.

(a) Plot the data ($Y =$ woh), and overlay the least squares line. Comment on the goodness of fit. Also, plot the residuals (on y-axis) and age on x-axis from the above fit and comment on any patterns you see. Is this a good residual plot? Explain why or why not.

(b) Create a new variable as follows:

```
age2 = nut$age^2
```

This simply creates a new vector with all age values squared. Run MLR of woh on both age and age2. Repeat the plots from (a) and show the results. Are these better? Why or why not?

(c) The authors of the study have reason to believe that the observations fall into two groups: (1) the first seven boys and (2) the remaining 65. Introduce an appropriate dummy variable to distinguish these groups. Then, using interactions produce two different lines for the two groups. Repeat the diagnostic plots from (a) and present the results.

(d) Of the three, which model do you prefer? Why?

3 Insurance charges

We follow up on the Insurance example we saw in class, and will investigate the role of the BMI variable on insurance charges.

(a) Create a new variable BmiHigh which is TRUE when BMI exceeds the sample median value, and is FALSE otherwise. Make sure BmiHigh is a factor.

(b) Continuing from the example in the sides, we now run a *three-way interaction* as follows:

```
> reg = lm(charges ~ age * smoker * BmiHigh, data=ins)
> summary(reg)
...
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -1954.804    605.734  -3.227  0.00128 **
age                        262.049     15.051  17.411  < 2e-16 ***
smokeryes                15644.530   1351.784  11.573  < 2e-16 ***
BmiHighTRUE               -234.487    872.045  -0.269  0.78805
age:smokeryes              -31.697     33.914  -0.935  0.35014
age:BmiHighTRUE              9.083     20.875   0.435  0.66353
smokeryes:BmiHighTRUE    17543.889   1910.678   9.182  < 2e-16 ***
age:smokeryes:BmiHighTRUE   33.049     46.544   0.710  0.47779
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Write down the implied regression model as we did in class, i.e.,
E(charges | ..) = β_0 + β_1*age...

Remember to translate a symbol ':' into a multiplication and terms such as BmiTrue into indicators.

How many subpopulations have we created with this model? How many observations are there within each subpopulation?

(c) Using the `lm` output, test if any of the slopes of `age` across subpopulations are significantly different.

(d) Explain what it means that the coefficients for `smokeryes` and `smokeryes:BmiHighTRUE` are significant. Does this help you identify the three groups in the original `charges` ∼ `age` plot? Visualize these groups in the original plot, by coloring or otherwise. (*Note:* GenAI can help!).

(e) Let's do an exercise in policy making. Suppose that a US government program has a fixed budget to be allocated to improving `bmi` or `smoker` prevalence in the general population.

Load the original dataset and do not include `BmiHigh` this time. The following model runs MLR with all possible two-way interactions:

```
> ins = read.csv("Insurance.csv", stringsAsFactors = T)
> reg = lm(charges ~.^2 , ins)
```

Check whether there is significant interaction between `smoker` and `bmi`. Explain briefly why this result is intuitive.

(f) Let's try to estimate the benefit from reducing the smoker population by 1%. The following code creates a "bootstrap" dataset where 1% of smokers are random are switched to being non-smokers, and then reports the predicted cost/patient in the new hypothetical population.

```
bootstrap_US_population = function() {
  ins_boot = ins   # "bootstrap" dataset
  idx.smokers = which(ins$smoker=="yes")
  make_nonsmoker = sample(idx.smokers, size=round(0.01*length(idx.smokers)))
  ins_boot$smoker[make_nonsmoker] <- "no"  # make 1% of smokers into non-smokers

  pred = predict(reg, ins_boot) # make prediction of charges.
  mean(pred) # average
}
```

Run this code 1,000 times and average over all predictions. This gives our estimate of medical expenses *per patient* after the intervention. Estimate the total $benefit from this intervention (*Note:* Assume 50 million adult patients in the US).

Ask ChatGPT (or other AI) for an independent estimate and compare the numbers.

(g) Suppose we considered another intervention that reduced `bmi`. By changing the above code, what % reduction in `bmi` would lead to the same benefits as the benefit we found in (f)?