Econ 108 FAll 2021
Problem Set 2

This problem set is due at 7pm, Friday, October 14, 2022.

1. Suppose our data set of $Y_i, i = 1, \dots, 11$ is $(1, 2, 5, 11, 17, 19, 30, 57, 98, 101, 500)$.

   (a) Find the number c that minimizes the *mean square errors*: $\min \sum_{i=1}^{n} (Y_i - c)^2$

   (b) Find the number c that minimizes the *mean absolute errors*: $\min \sum_{i=1}^{n} |Y_i - c|$.

2. In this question we will use the log normal distribution simulated on page 8 of 01uncertainty.pdf to illustrate the bias of the sample standard deviation and the use of bootstrap.

   (a) Use the code on page 8 of 01uncertain.pdf to simulate a large population from a log normal distribution. Compute the *population* standard deviation σ. (Note: we are asking about the standard deviation σ, and not the variance σ^2)

   (b) Modify the code on page 9 of 01uncertain.pdf so that it applies to the *sample standard deviation* instead of the *sample mean*.

   (c) What is the *averaged sample standard deviation* across the simulation draws? This gives you a *simulation based* estimate of the expectation of the *sample standard deviation*. Does it appear to be bias, and in what direction?

   (d) Using the simulation draws, can you calculate the frequentist sampling standard deviation of the sample standard deviation?

   (e) Now draw a *single* sample from the population that has been simulated on page 8 of 01uncertain.pdf. Use the bootstrap method to estimate the standard deviation of the sample standard deviation.

   (f) Given that we are concerned that the sample standard deviation might be biased for the population standard deviation, describe how the bootstrap method can be used to come up with a bias-corrected sample standard deviation.

   (g) What are the three methods of confidence interval construction using the bootstrap? Apply each of these methods to come up with a confidence interval *for the population standard deviation* (σ) based on the sample standard deviation ($\hat{\sigma}$).

3. In this exercise we will use the oj.csv data and the oj.R code from the textbook.

(a) Understand and run the oj.R code. You would need to make a few changes. For example, oj$logmove needs to be replaced by log(oj$sales).

(b) We can use the following linear regression

```
glm(formula = log(sales) ~ log(price) + brand, data = oj)
```

to study how brand names affect the levels of sales without an interactive effect on price elasticity. Based on the output of this regression, how much does sales increase (in multiplicative percentage terms) when the orange brand is changed from dominicks to tropicana? Is this change statistically significant?

(c) In the above regression, how much does sales increase (in multiplicative percentage terms) when the orange brand is changed from minute.maid to tropicana? Is this change statistically significant? (Hint: To answer the second question of statistical significance, you might consider changing the reference level of the brand factor before running the regression again.)

(d) As we saw in class, the interactive regression

```
glm(formula = log(sales) ~ log(price) * brand, data = oj)
```

generates the following output

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  10.95468    0.02070 529.136   <2e-16 ***
log(price)                   -3.37753    0.03619 -93.322   <2e-16 ***
brandtropicana                0.96239    0.04645  20.719   <2e-16 ***
brandminute.maid              0.88825    0.04155  21.376   <2e-16 ***
log(price):brandtropicana     0.66576    0.05352  12.439   <2e-16 ***
log(price):brandminute.maid   0.05679    0.05729   0.991    0.322
```

1. Explain why this table implies the following price elasticities:

   `dominicks: -3.4, minute maid: -3.3, tropicana: -2.7`
2. Is the difference between the dominicks price elasticity and minute.maid price elasticity statistically different from zero?
3. Is the difference between the dominicks price elasticity and tropicana price elasticity statistically different from zero?
4. Is the difference between the minute.maid price elasticity and tropicana price elasticity statistically different from zero? (Hint: to answer this question you might need to changing the reference level of the brand factor before running the regression again.)

(e) In class we also ran the 3-way interactive regression to study the brand-specific effect of advertisement on price elasticity:

```
glm(formula = log(sales) ~ log(price) * brand, data = oj)
```

1. Explain how the output of this regression implies the elasticity table:

```
                Dominicks  Minute.Maid Tropicana
Not Featured     -2.8        -2.0            -2.0
Featured         -3.2         -3.6        -3.5
```

2. During the advertisement period, are the three brand-specific elasticities statistically different from each other?
3. During the non-advertisement period, are the three brand-specific elasticities statistically different from each other?
4. For each of the three brands, does advertisement affect price elasticities statistically significantly?
5. In this question (e), minute.maid appears similar to tropicana. But in the previous question, minute.maid appears similar to dominicks. Do you have an explanation for this difference?
6. To account for the possibility that the price elasticity of sales might not be a constant and might depend on the price level, we decide to add a quadratic term to the linear regression model of

$$\log(\text{sales}) = \beta_0 + \beta_1 \log(\text{price}) + \beta_2 \left(\log(\text{price})\right)^2 + \beta_b' \text{brand} + \epsilon$$

We can implement this in R using the command

```
reg = glm(log(sales) ~ log(price) + I(log(price)^2)+brand, data=oj)
```

At the average price level, what is the price elasticity of sales?