Test 2

Econometrics 3112.003, Spring 2019
Belk College of Business, UNCC
Ercan Karadas

[1] A large random sample of 81 filtered cigarettes is obtained, and the tar content of each cigarette is measured. The sample has a mean ($\overline{Y}$) of 23.0 mg and a standard deviation (s_Y) of 6.7 mg. We are interested in testing the claim that the mean tar content of filtered cigarettes is greater than 21.1 mg, which is the mean for unfiltered cigarettes.

a) Formulate the appropriate null and alternative hypotheses.

$$H_0 : \mu_Y \leq 21.1$$
$$H_1 : \mu_Y > 21.1$$

b) What is the distribution of $\overline{Y}$?
Under the null-hypothesis, Central Limit Theorem gives

$$\overline{Y} \sim N\left(21.1, \frac{6.7^2}{\sqrt{81}}\right) \qquad \text{(approximately)}$$

This result also allows us to use the standard normal distribution as an approximation to t distribution.

c) Calculate the appropriate test statistic and conclude the test at $\alpha = 0.05$?

$$Z^a = \frac{23.0 - 21.1}{6.7/9} = 2.55 > Z_{0.025} = 1.96$$

Therefore, we reject the null-hypothesis that $\mu_Y \leq 21.1$ and this conclusion provides some support for the original claim that $\mu_Y > 21.1$.

d) Find the p-value of the test in the previous part?

$$\begin{aligned} p - \text{value} &= P\left(Z > Z^a\right) \\ &= P\left(Z > 2.55\right) \\ &= 1 - P\left(Z \leq 2.55\right) \\ &= 1 - 0.9946 \\ &= 0.0054 \end{aligned}$$

[2] In any year, the weather may cause damages to a home. On a year-to-year basis, the damage is random. Let Y denote the dollar value of damages in any given year. Suppose that during 95% of the year $Y = \$0$, but during the other 5% $Y = 20,000$.

Basically, Y is a binary (or Benouilli) r.v. with the probability distribution

$$P(Y = 0) = 0.95, \quad P(Y = 20,000) = 0.05$$

a) What are the mean and standard deviation of damages caused in a year?

$$\begin{aligned} \mu_Y &= E(Y) = 0 \times 0.95 + 20,000 \times 0.05 = 1000 \\ \sigma_Y^2 &= E\left[(Y - \mu_Y)^2\right] \\ &= (0 - 1000)^2 \times 0.95 + (20,000 - 1000)^2 \times 0.05 = 19 \times 10^6 \end{aligned}$$

and the standard deviation of Y is $\sigma_Y = \sqrt{19 \times 10^6} = 4359$.

For the rest of the problem, consider an "insurance pool" of 100 people whose homes are sufficiently dispersed so that, in any year, the damage to different homes can be viewed as independently distributed random variables. Let $\overline{Y}$ denote the average damage caused to these 100 homes in one year.

In this part, we can think of we have a sample consisting 100 i.i.d draws from the Bernouilli given above. Then the CLT applies and we can write the distribution of $\overline{Y}$ (the average damage caused to these 100 homes) as $\overline{Y} \sim N(\mu_Y, \sigma_Y^2/100)$

b) What is the expected value of the average damage $\overline{Y}$?

$E(\overline{Y}) = \mu_Y = 1000$ and $V(\overline{Y}) = \sigma_Y^2/100 = 19 \times 10^6/100 = 19 \times 10^4$

c) What is the probability that $\overline{Y}$ exceeds \$2,000? (Hint: Use the central limit theorem to compute an approximate answer.)

Since $\overline{Y} \sim N(1000, 19 \times 10^4)$

$$\begin{aligned} P(\overline{Y} > 2,000) &= 1 - P(\overline{Y} \leq 2,000) \\ &= 1 - P\left(Z \leq \frac{2000 - 1000}{\sqrt{19 \times 10^4}}\right) \\ &= 1 - P\left(Z \leq 2.29\right) = 0.0109 \end{aligned}$$

[3] A soda vendor at UNC Charlotte football games observes that more sodas are sold the warmer the temperature at game time is. Based on 32 home games covering five years, the vendor estimates the relationship between soda sales and temperature to be

$$\widehat{Y} = -240 + 8\,X, \qquad R^2 = 0.34, \qquad SER = 12.8$$

where Y is the number of sodas she sells and X is temperature in degrees Fahrenheit.

a) Interpret the estimated slope and intercept. Do the estimates make sense? Why, or why not?

The interpretation of the slope term is straightforward: as the temperature rises by one degrees Fahrenheit the number of sodas sold increases by 8 cans, on average. However, here the intercept term has no economically meaningful interpretation. But note that, at least we can see that it is meaningless to interpret it as the number of sodas sold when the temperature is zero degrees Fahrenheit.

b) Interpret R^2 and SER.

R^2: The variation in temperature explains 34% of the variation in the number of sodas sold. And $SER = 12.8$ is the magnitude of average OLS residual.

c) On a day when the temperature at game time is forecast to be 80 degrees Fahrenheit, predict how many sodas the vendor expects to sell?

On a day when the temperature at game time is forecast to be 80 degrees Fahrenheit, the predicted sodas sales: $\widehat{Y} = -240 + 8 \times 80 = 400$

d) Below what temperature are the predicted sales zero?

Solving $\widehat{Y} = -240 + 8\,X \leq 0$ gives $X \leq 30$, i.e. when the temperature is below 30 degrees Fahrenheit the vendor doesn't sell anything.

e) Sketch a graph of the estimated regression line.

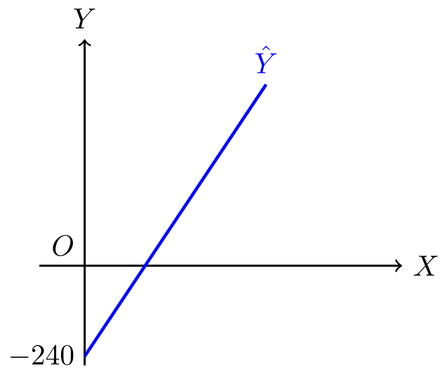

Figure 1: The Sample Regression Line: $\widehat{Y} = -240 + 8\,X$

f) Suppose that another researcher uses the same dataset but instead of measuring the temperature in Fahrenheit, she measures in Celsius. In that case, which of the following would change: $\widehat{\beta}_0$, $\widehat{\beta}_1$, R^2, SER.

Both $\widehat{\beta}_0$ and $\widehat{\beta}_1$ would change, and R^2 and SER would remain the same.

[4] Suppose that a researcher, using data on class size (CS) and average test scores from 50 third-grade classes, estimates the OLS regression:

$$\widehat{\text{Test Score}} = 640.3 - 4.93 \times \text{CS}, \qquad R^2 = 0.11, \qquad SER = 8.7$$

a) A classroom has 28 students. What is the regressions prediction for that classrooms average test score?

The predicted average test score for a classroom of size 28 is

$$\widehat{\text{Test Score}} = 640.3 - 4.93 \times 28 = 502$$

b) Last year a classroom had 25 students, and this year it has 21 students. What is the regression's prediction for the change in the classroom average test score?

The change in the classroom average test score:

$$\Delta\widehat{\text{Test Score}} = (-4.93 \times 21) - (-4.93 \times 25) = 19.72$$

c) The sample average class size across the 50 classrooms is 22.8. What is the sample average of the test scores across the 50 classrooms? (Hint: Review the formulas for the OLS estimators.)

Recall that

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{X} \quad \text{or} \quad \overline{Y} = \widehat{\beta}_0 + \widehat{\beta}_1\overline{X}$$

In this example, this can be written as,

$$\overline{\text{Test Score}} = 640.3 - 4.93 \times \overline{\text{CS}} = 640.3 - 4.93 \times 22.8 = 527.896$$

d) What is the sample standard deviation of test scores across the 50 classrooms? (Hint: Review the formulas for the R^2 and SER.)

Recall that

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\widehat{u}_i^2} = \sqrt{\frac{1}{n-2}SSR}$$

Solving this for SSR gives

$$SSR = (n-2)SER^2 = (50-2)8.7^2 = 3633.12$$

On the other hand, using the formula for R^2

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - SSR}{TSS} = 1 - \frac{SSR}{TSS}$$

where we used the fact that $TSS = ESS + SSR$. Now solving this for TSS gives

$$TSS = \frac{SSR}{1-R^2} = \frac{3633.12}{1-0.11} = 4082.16$$

Finally,

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \frac{1}{n-1}TSS = \frac{1}{50-1}4082.16 = 83.31$$

and the standard deviation of test scores is $\sqrt{83.31} = 9.13$.

[5] R Questions.

a) Suppose I have a vector `x <- 1:5` and `y <- 3:4`. What is produced by the expression `x + y`.

```
x + y = 4 6 6 8
```

b) Suppose I have a vector `x <- c(5, 3, 8, 12, 2, -1)` and I want to set all elements of this vector that are greater than 7 to be equal to zero. Write an R code that achieves this.

```
x[x > 7] <- 0
```

c) Consider the following expression

```
x <- 5
y <- if (x<3) {
NA
}else{
10
}
```

What is the value of `y` after evaluating this expression.

```
y=10
```