Econ 103: Midterm Questions

Manu Navjeevan

August 26, 2021

1 Econ 41 Review

1. Suppose I know the joint distribution of X and Y. Using the joint pdf of X and Y I calculate the covariance between Y and X using the following formulas:

$$\begin{aligned} \text{Cov}(X,Y) &= \mathbb{E}[YX] - \mathbb{E}[Y]\mathbb{E}[X] = 30 \\ \text{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = 100 \\ \text{Var}(Y) &= \mathbb{E}[(Y - \mathbb{E}[Y])^2] = 4 \end{aligned}$$

Have I made any errors?

(a) No, all formulas are correct and there are no apparent errors in calculation.

(b) Yes, the formula for $\text{Var}(Y)$ is incorrect.

(c) Yes, the formula for $\text{Cov}(X,Y)$ is incorrect.

(d) Yes, the formula for $\text{Var}(X)$ is incorrect.

(e) Yes, all formulas are correct but I have made an error when computing either the variances or the covariance.

Answer: All formulas are correct. However, recall from Homework that, by the Cauchy-Schwarz inequality the correlation coeffecient must be bounded between -1 and 1. Here the correlation coeffecient is given by

$$\frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{30}{10 \cdot 2} = 1.5.$$

So, I must have made a calculation error.

2. Suppose that I can pay \$0.8 to enter a lottery whose payout (in dollars) is described by a random variable Y with pdf:

$$f_Y(a) = \begin{cases} ca^2 & \text{if } a \in [0,1] \\ 0 & \text{otherwise} \end{cases},$$

for some unknown constant $c > 0$. Should I pay to enter this lottery?

(a) It is impossible to tell because we do not know c.

(b) Yes, because the expected value, $\mathbb{E}[Y]$, is positive.

(c) Yes, because the expected value, $\mathbb{E}[Y]$, is greater than \$0.8.

(d) No, because the expected value, $\mathbb{E}[Y]$, is negative.

(e) No, because the expected value, $\mathbb{E}[Y]$, is less than \$0.8.

Answer: Using the fact that the pdf integrates to one, we can solve for c:

$$\int_0^1 ca^2 \, da = 1 \implies c \cdot \frac{a^3}{3}\bigg|_0^1 = 1 \implies \frac{c}{3} = 1 \implies c = 3.$$

Now, we can solve for the exptected value of Y:

$$\mathbb{E}[Y] = \int_0^1 3f_Y(a) \cdot a \, da = \int_0^1 3a^3 \, da = 3 \cdot \frac{a^4}{4}\bigg|_0^1 = \frac{3}{4}.$$

Because the expected value of this lottery is lower than the cost to enter it, we should not enter the lottery.

3. Suppose we are interested in the distribution of the random variable Y which has pdf given by:

$$f_Y(a) = \frac{1}{10},$$

if $a \in [0, 10]$ and equal to zero otherwise. What is the probability that $Y = 4$, that is what is $\mathbb{P}_Y(\{4\})$?

Fill in the Blank: Answer should be zero as the probability that a continuous random variable takes up a specific value is equal to zero.

4. Suppose we are interested in the relationship (in the population of UCLA students) between the average number of cups of coffee drank per day by UCLA students, Y, and the number of classes being taken currently, X. We find that $\text{Cov}(X, Y) = 0.75$. Which of the following statements is most correct?

(a) A one unit increase in the number of classes taken is associated with a 0.75 unit increase in the average number of cups of coffee drank.

(b) X should not be considered a random variable because the number of classes someone is taking does not vary throughout the quarter, even if this number varies from person to person in the population.

(c) 75% of the variance in Y can be explained by a single linear regression model with X.

(d) The expected value of Y given $X = 0$ is 0.75.

(e) There is a positive relationship between X and Y.

Answer: The covariance is just a measure of linear association. Without knowing other quantites we cannot compute R^2 or $\hat{\beta}_1$.

2 Single Linear Regression

1. Suppose I run a regression of Y on X and obtain the following estimated regression line

$$\widehat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

What would happen to these parameters if I were to add 20 to each value of Y_i before estimating my regression line?

(a) The estimated slope parameter and the estimated intercept parameter would both increase by 20.

(b) The estimated slope parameter would increase by 20 while the estimated intercept parameter would remain unchanged.

(c) The estimated slope parameter would remain unchanged while the estimated intercept parameter would decrease by 20.

(d) None of the other options are correct.

(e) The estimated slope parameter would remain unchanged while the estimated intercept parameter would increase by 20

Answer: To see this, let's let $\tilde{Y} = Y + 20$ and consider estimating the regression line

$$\tilde{Y} = \beta_0^\circ + \beta_1^\circ + \epsilon.$$

Using our formulas we find that

$$\begin{aligned}
\hat{\beta}_1^\circ &= \frac{\sum_{i=1}^n (\tilde{Y} - \bar{\tilde{Y}})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \frac{\sum_{i=1}^n (Y_i + 20 - (\bar{Y} + 20))(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \hat{\beta}_1
\end{aligned}$$

so the slope parameter remains the same. The estimated intercept parameter is given:

$$\begin{aligned}
\hat{\beta}_0^\circ &= \bar{\tilde{Y}} - \hat{\beta}_1^\circ \bar{X} \\
&= \bar{Y} + 20 - \hat{\beta}_1 \bar{X} \\
&= \hat{\beta}_0 + 20.
\end{aligned}$$

so the intercept parameter increases by 20.

2. Suppose that we estimate the model

$$Y = \hat{\beta}_0 + \hat{\beta}_1 \ln(X) + \epsilon.$$

After computing standard errors and computing our parameter estimates $\hat{\beta}_1$ and $\hat{\beta}_0$, we find that the 95% confidence interval for β_1 is given by $[1, 5]$. What is $\hat{\beta}_1$?

(a) We cannot say using this information because X has been transformed.

(b) We cannot use this information to solve for $\hat{\beta}_1$ unless we know $\hat{\beta}_0$.

(c) $\hat{\beta}_1 = 2$

(d) $\hat{\beta}_1 = 3$

(e) $\hat{\beta}_1 = 4$

Answer: The estimated parameter is always in the middle of our confidence interval. So $\hat{\beta}_1 = 3$.

3. Suppose that we estimate the model

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \epsilon.$$

After computing standard errors and computing our paramter estimates $\hat{\beta}_1$ and $\hat{\beta}_0$, we find that the 95% confidence interval for β_1 is given by $[3, 5]$. Which of the following statements is definetly true about the following test?

$$H_0 : \beta_1 \leq 2.99 \quad \text{vs.} \quad H_1 : \beta_1 > 2.99.$$

(a) We cannot use the confidence interval to determine whether we can reject this null hypothesis at level $\alpha = 0.05$ because confidence intervals can only be used to run a two sided test.

(b) We would reject this null hypothesis in favor of the one sided alternative at level $\alpha = 0.01$.

(c) We would fail to reject this null hypothesis in favor of the one sided alternative at level $\alpha = 0.05$.

(d) None of the other answers are correct.

(e) We would reject this null hypothesis in favor of the one sided alternative at level $\alpha = 0.025$.

Answer: Because 2.99 is not in our 95% confidence interval, we know that we would reject the null hypothesis

$$H_0 : \beta_1 = 2.99 \;\text{ vs. }\; H_1 : \beta_1 \neq 2.99.$$

at level $\alpha = 0.05$. As we went over in homework, because $\hat{\beta}_1 > 2.99$, this means that we would reject the null hypothesis

$$H_0 : \beta_1 \leq 2.99 \;\text{ vs. }\; H_1 : \beta_1 > 2.99.$$

at level $\alpha/2 = 0.025$.

4. Which statment best describes how we arrived at the result that $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ is approximately normal for large n?

(a) For large n, the quantity $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ is approximately equal to the scaled sample variance of X over the sample mean of $\epsilon_i(X_i - \mu_X)$. By law of large numbers this sample mean is close to the true mean and the scaled sample variance of X is close to the true variance of X.

(b) For large n, the quantity $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ is approximately equal to the scaled sample variance of X over the sample mean of $\epsilon_i(X_i - \mu_X)$. By central limit theorem, the scaled sample variance of X is approximately normally distributed and by law of large numbers this sample mean is close to the true mean.

(c) For large n, the quantity $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ is approximately equal to the scaled sample mean of $\epsilon_i(X_i - \mu_X)$ over the sample variance of X. By law of large numbers this scaled sample mean is close to the mean and the sample variance of X is close to the true variance of X.

(d) For large n, the quantity $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ is approximately equal to the scaled sample mean of $\epsilon_i(X_i - \mu_X)$ over the sample variance of X. By the central limit theorem this sample mean approximately normally distributed and by the law of large numbers the sample variance of X is close to the true variance of X.

Answer: This is covered in the first set of single linear regression slides.

5. Suppose that β_0 and β_1 are the true line of best fit parameters. After collecting a sample of data $\{Y_i, X_i\}_{i=1}^n$ we obtain for estimates for these parameters $\hat{\beta}_0$ and $\hat{\beta}_1$. Recall that $\hat{\epsilon}_i = Y_i - \hat{\beta}_i - \hat{\beta}_1 X_i$ and $\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$. Which of the following statements is necessarily true for all possible samples?

(a) $\frac{1}{n}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 \leq \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$.

(b) $\mathbb{E}[(Y - \hat{\beta}_0 - \hat{\beta}_1 X)^2] \leq \mathbb{E}[(Y - \beta_0 - \beta_1 X)^2]$.

(c) $\mathbb{E}[(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X)X] = 0$

(d) $\frac{1}{n}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i) = 0$.

(e) $\sum_{i=1}^n \hat{\epsilon}_i^2 \leq \sum_{i=1}^n \epsilon_i^2$.

Answer: Recall that $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize

$$\hat{\beta}_0, \hat{\beta}_1 = \arg\min_{b_0, b_1} \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2.$$

This means that for any $(b_0, b_1) \neq (\hat{\beta}_0, \hat{\beta}_1)$,

$$\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \leq \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2.$$

In particular, for the true parameters β_0, β_1 :

$$\sum_{i=1}^n (\underbrace{Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i}_{\hat{\epsilon}_i})^2 \leq \sum_{i=1}^n (\underbrace{Y_i - \beta_0 - \beta_1 X_i}_{\epsilon_i})^2 \implies \sum_{i=1}^n \hat{\epsilon}_i^2 \leq \sum_{i=1}^n \epsilon_i^2.$$

The other answers are incorrect for the various reasons that they hold for the true parameters but not the estimated parameters or vice-versa.

6. Suppose we are interested in the relationship between our outcome variable Y, a measure of anxiety levels, and our explanatory X, the number of energy drinks a person drinks. Given a sample of observational data from UCLA student, $\{Y_i, X_i\}_{i=1}^n$ we estimate the line

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

and find that $\hat{\beta}_0 = 5$ and $\hat{\beta}_1 = 0.5$. Which interpretation of these estimated parameters is most correct?

(a) We estimate that, on average, drinking an additional energy drink causes a 0.5 unit increase in anxiety levels.

(b) We estimate that the average UCLA student has an anxiety level measure of 5.

(c) We estimate that, on average, a one unit increase in anxiety levels is associated with a 0.5 unit increase in energy drink consumption.

(d) We estimate that, on average, a 0.5 unit increase in energy drink consumptions is associated with a 1 unit increase in anxiety levels.

(e) We estimate that, on average, a 1 unit increase in energy drink consumptions is associated with a 0.5 unit increase in anxiety levels.

Answer: The correct answer correctly interprets $\hat{\beta}_1$ without any causal implication.

7. Which of the following assumptions are necessary for $\hat{\beta}_1$ from the single linear regression model to be approximately normally distributed with mean β_1 when n is large?

(a) Homoskedasticity: The variance of ϵ does not change with various values of X, this is formally presented as either $\mathbb{E}[\epsilon^2|X = x]$ or $\text{Var}(\epsilon|X = x)$ is constant for all values of x.

(b) Rank Condition: There are at least two values of X in the population.

(c) Random Sampling: The samples $\{Y_i, X_i\}_{i=1}^n$ are independently and identically drawn from the underlying population of interest (Y, X).

(d) Linearity: The true conditional mean is linear. That is $\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x$ for all x.

(e) Normality. The linear regression error terms ϵ are normally distributed.

Answer: Mostly from the first set of regression slides. Recall that homoskedasticity is not required for asymptotic normality, it just helps us simplify the form of the asymptotic variance.

8. We are interested in the relationship between smoking, X, and developing lung disease Y. To investigate this relationship, we collect a random sample of data $\{Y_i, X_i\}_{i=1}^n$ and estimate the following linear regression equation

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Which statement best describes why we consider $\hat{\beta}_0$ and $\hat{\beta}_1$ random variables?

(a) The relationship between X and Y is not deterministic, some people who smoke never develop lung disease while others who do not smoke do develop lung disease.

(b) By the central limit theorem $\hat{\beta}_0$ and $\hat{\beta}_1$ are approximately normally distributed when n is large.

(c) We do not consider $\hat{\beta}_0$ and $\hat{\beta}_1$ random variables, they are fixed quantities that we compute from our sample.

(d) We do not consider $\hat{\beta}_0$ and $\hat{\beta}_1$ random variables as they are fixed quantities that come from the joint population distribution of (Y, X).

(e) $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables because they are functions of the sample, which is randomly collected from the population.

Answer: From the first set of regression slides.

9. We are interested in the relationship between smoking, X, and developing lung disease Y. To investigate this relationship, we collect a random sample of data $\{Y_i, X_i\}_{i=1}^n$ and estimate the following linear regression equation

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

In our specific sample, we find that $\hat{\beta}_1 = 4$. Under the assumption that $\beta_1 = 0$, we estimate the standard error of $\hat{\beta}_1$ and use the normal approximation for the distribution of $\hat{\beta}_1$ to find that (approximately) $\Pr(|\hat{\beta}_1| > 4) = 0.04$. Which of the following claims can we make about the result of the hypothesis test:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0?$$

(a) We would reject this null hypothesis at level $\alpha = 0.03$ because the p-value is equal to $0.04/2 = 0.02 < \alpha$.

(b) We would fail to reject this null hypothesis at level $\alpha = 0.05$ because the p-value is equal to $0.04 \cdot 2 = 0.08 > \alpha$.

(c) We would fail to reject this null hypothesis at level $\alpha = 0.1$ because the test statistic $t^* = 0.04 \cdot 2 = 0.08$ is less than $z_{0,9} \approx 1.28$.

(d) We would fail reject this null hypothesis at level $\alpha = 0.1$ because the test statistic $t^* = 4$ is greater than $z_{0.9} \approx 1.28$.

(e) We would reject this null hypothesis at level $\alpha = 0.05$ because the p-value is equal to $0.04 < \alpha$.

Answer: The value $\Pr(|\hat{\beta}_1| > 4)$ under the assumption that $\beta_1 = 0$ is the motivation for the p-value. It is the probability that we would obtain our result, or something even further from the null hypothesis, if the null hypothesis was true. Should refer to the first example when motivating the asymptotic variance.

10. Suppose I collect a random sample $\{Y_i, X_i\}_{i=1}^n$ and estimate the linear model:

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

After doing so I find my estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and use these to construct my estimated residuals $\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$. I find that the sample correlation coeffecient between $\hat{\epsilon}_i^2$ and X is large. This is evidence against which assumption?

(a) Rank Condition: There are at least two values of X in the population.

(b) Random Sampling: The samples $\{Y_i, X_i\}_{i=1}^n$ are independently and identically drawn from the underlying population of interest (Y, X).

(c) None of the assumptions listed appear to be violated.

(d) This can be interpreted as evidence against all of these listed assumptions.

(e) Homoskedasticity: The variance of ϵ does not change with various values of X.

Answer: A correlation between $\hat{\epsilon}_i^2$ and X indicates that the **variance** of the residuals is changing with X.

11. Suppose I collect a random sample $\{Y_i, X_i\}_{i=1}^n$ and estimate the linear model:

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

After doing so I find my estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and use these to construct my estimated residuals $\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$. I find that the sample correlation coeffecient between $\hat{\epsilon}_i$ and X is large (positive). Which of the following is most correct:

(a) Homoskedasticity looks to be violated.

(b) There is a large positive relationship between Y and X.

(c) The estimator $\hat{\beta}_1$ is not approximately normally distributed, even for large n.

(d) I should transform either Y or X to get a better fit.

(e) I have made a calculation error when computing the sample correlation coeffecient.

Answer: Recall from the first order conditions of the estimating equations that

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i X_i = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i = 0.$$

The sample covariance between $\hat{\epsilon}_i$ and X is then

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i(X_i - \bar{X}) = 0.$$

This means that the correlation coeffecient between X and $\hat{\epsilon}$ should also be zero.

12. After collecting a random sample of size $n = 100$, $\{Y_i, X_i\}_{i=1}^{100}$ I estimate the linear model

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

I find that $\hat{\beta}_0 = \hat{\beta}_1 = 1.5$ and that $\sigma^2_{\beta_1} = 100$, $\sigma^2_{\beta_0} = 200$, and $\sigma_{\beta_{01}} = 50$. I also know that $z_{0.95} \approx 1.64$ and that $z_{0.9} \approx 1.28$. What is the result of testing the hypothesis:

$$H_0 : \lambda = \beta_0 + \beta_1 = 0 \text{ vs. } H_1 : \lambda = \beta_0 + \beta_1 \neq 0.$$

Hint: Recall that $\sigma^2_{\beta_0}$ is such that, for large n, $\sqrt{n}(\hat{\beta}_0 - \beta_0) \sim N(0, \sigma^2_{\beta_0})$, $\sigma^2_{\beta_1}$ is such that, for large n, $\sqrt{n}(\hat{\beta}_1 - \beta_1) \sim N(0, \sigma^2_{\beta_1})$, and $\sigma_{\beta_{01}} = \text{Cov}(\sqrt{n}(\hat{\beta}_0 - \beta_0), \sqrt{n}(\hat{\beta}_1 - \beta_1))$.

(a) Reject the null hypothesis in favor of the alternative hypothesis at level $\alpha = 0.1$.

(b) Not enough information to answer this question.

(c) Fail to reject the null hypothesis in favor of the alternative hypothesis at level $\alpha = 0.1$

Answer: This problem should be similar to what we saw on Homework 2. First, we calculate $\text{Var}(\hat{\lambda})$:

$$\begin{aligned}\text{Var}(\hat{\lambda}) &= \text{Var}(\hat{\beta}_0) + \text{Var}(\hat{\beta}_1) + 2\,\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)\\ &= \frac{\sigma^2_{\beta_0}}{n} + \frac{\sigma^2_{\beta_1}}{n} + 2\frac{\sigma_{\beta_{01}}}{n}\\ &= \frac{100}{100} + \frac{200}{100} + 2\frac{50}{100}\\ &= 4\end{aligned}$$

Now, we calculate our test statistic t^*, noting that $\hat{\lambda} = \hat{\beta}_0 + \hat{\beta}_1 = 3$:

$$t^* = \frac{\hat{\lambda}_0 - 0}{\sqrt{\text{Var}(\hat{\lambda})}} = \frac{3}{\sqrt{4}} = 1.5.$$

Since this is a two sided test at level $\alpha = 0.1$, we compare our test statistic to $z_{1-\alpha/2} = z_{0.95} \approx 1.64$. Since $t^* < 1.64$ we fail to reject this null hypothesis.

13. After collecting random sample of size $\{Y_i, X_i\}_{i=1}^{20}$ we find that $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 200$, $\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) = 400$, $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 400$, and $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = 300$. We do not know n, though we assume that our distributional results still hold. Which of the following is most correct?

(a) We can say that, in our sample, $50\% = \sqrt{0.75\%}$ of the variance in Y can be explained by the linear model with X.

(b) We can use the data above to find that $\hat{\beta}_0 = 3$, which means that we estimate the average value of Y when $X = 0$ to be 3.

(c) If we increase X by one, Y will increase by 1.

(d) None of the other options are correct.

(e) We can use the information above to reject the null hypothesis $H_0 : \beta_1 = 0$ against a two sided alternative $H_1 : \beta_1 \neq 0$ at level $\alpha = 0.05$ ($z_{0.975} = 1.95$).

Answer: The first three possible responses are incorrect so we will try to see if we can show the last response is true. The idea is to notice that using the information, we can calculate $\hat{\sigma}^2_{\beta_1}$ as well as $\hat{\beta}_1$. Then, what we have to do to for the hypothesis test is lower bound the test statistic t^*. Using $\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) = 400$ and $\sum_{i=1}^n (X_i - \bar{X})^2 = 200$, we find that $\hat{\beta}_1 = 1$. Now, to compute $\hat{\sigma}^2_{\beta_1}$ recall that

$$\hat{\sigma}^2_{\beta_1} = \frac{\sigma^2_\epsilon}{\sigma^2_X} = \frac{\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\text{SSE}}{200}.$$

Using $\text{SST} = \text{SSR} + \text{SSE}$ we find that $\text{SSE} = 100$. This means that $\hat{\sigma}^2_{\beta_1} = \frac{1}{2}$. Putting this together, we know that

$$t^* = \frac{1}{\frac{1}{\sqrt{2}}\frac{1}{\sqrt{n}}} = \sqrt{2}\sqrt{n}.$$

So long as $n \geq 2$ we will get that $t^* \geq 2$ and so we can reject the null hypothesis since $2 > z_{0.975}$. We know that $n \geq 2$ because $\sum_{i=1}^n (Y_i - \bar{Y})^2 \neq 0$ (if $n = 1$ then $Y_1 = \bar{Y}$).

14. Suppose we are interested in the relationship between price of a toy P and quantity of that toy sold Q. We record prices and quantities sold at various stores around the area, $\{Q_i, P_i\}_{i=1}^n$ estimate the following model

$$\ln(Q) = \beta_0 + \beta_1 \ln(P) + \epsilon.$$

What is the best interpretation of our parameter $\hat{\beta}_1$?

(a) On average, we estimate that a 1% increase in price is associated with a $\hat{\beta}_1/100$ unit change in quantity sold.

(b) On average, we estimate that a one unit increase in price is associated with a $\beta_1\%$ unit change in price.

(c) On average, we estimate that a 1% increase in price is associated with a $100 \cdot \hat{\beta}_1\%$ change in quantity sold.

(d) In this particular case we can interpret the result as a causal effect because price clearly has a causal impact on quantity sold.

(e) None of the other answers are correct.

Answer: From the second set of single linear regression slides. Recall that a one unit change in $\ln(W)$ is associated with an approximately 100% change in W. So a one percent change in P is associated with a 0.01 unit change in $\ln(P)$. This in turn is associated with a $\beta_1 \cdot 0.01$ unit change in $\ln(Q)$, which corresponds to a $\beta_1\%$ change in Q. In total, the correct interpretation would be that a 1% change in P is associated with a $\beta_1\%$ change in Q. This is not a listed solution.

We also cannot interpret this as causal because stores will take into consideration demand when setting prices. A higher prices may correspond to a toy being sold in a store that faces a higher demand, which means that there could be a positive association between price and quantity sold in the data. This would not indicate that there is a positive causal relationship between price and quantity sold.

15. Suppose that after collecting a sample of size $n = 50$ we estimate the following regression line

$$\hat{Y} = 1 + 2 \cdot X.$$

We also find that $R^2 = 0.75$, $\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) = 400$ and $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 400$. Using R we find that $\Pr(|Z| > 1) \approx 0.3173$ where $Z \sim N(0, 1)$.

(a) We are 84.134% confident that the true value of β_1 is in the interval $[1.9, 2.1]$.

(b) We are 68.26% confident that the true value of β_1 is in the interval $[1.5, 2.5]$.

(c) We are 84.134% confident that the true value of β_1 is in the interval $[1.5, 2.5]$.

(d) None of the other answers is correct/We do not have enough information to say

(e) We are 68.26% confident that the true value of β_1 is in the interval $[1.9, 2.1]$.

Answer: In order to calculate a $100(1-\alpha)\%$ confidence interval, we have to know $z_{1-\alpha/2}$ and the standard error of $\hat{\beta}_1$. Recall that $z_{1-\alpha/2} > 0$ is such that

$$\Pr(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2 \implies \Pr(Z > z_{1-\alpha/2}) = \frac{\alpha}{2} \implies \Pr(|Z| > z_{1-\alpha/2}) = \alpha.$$

Since we are given that $\Pr(|Z| > 1) \approx 0.3173$ we can say that $z_{1-\alpha/2} = 1$ for $\alpha \approx 0.3173$. We can use this to compute a $100(1-\alpha)\% = 68.26\%$ confidence interval. Now, we have to compute $\hat{\sigma}^2_{\beta_1}$. To do this, note that

$$\hat{\sigma}^2_{\beta_1} = \frac{\hat{\sigma}^2_{\epsilon}}{\hat{\sigma}^2_X} = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\text{SSE}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Using $\hat{\beta}_1 = 2$ and $\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) = 400$ we find that $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 200$. Using $R^2 = 0.75$ and SST $= 400$ we find that SSE $= 100$. So $\hat{\sigma}^2_{\beta_1} = \frac{100}{200} = \frac{1}{2}$. The standard error of $\hat{\beta}_1$ is given by $\sigma_{\beta_1}/\sqrt{n} = \frac{1}{\sqrt{2}}\frac{1}{\sqrt{50}} = \frac{1}{\sqrt{100}} = \frac{1}{10}$. Combining this together with $z_{1-\alpha/2} = 1$ gives the result.

16. Suppose that, before running our regression, we transform our Y variables by scaling (multiplying) them by $\frac{1}{10}$. We do not transform X. Our R^2 from the initial regression is equal to 0.6, whereas the R^2 from the regression is equal to 0.06. Which of the following is most correct:

(a) We can explain 0.6% of the variance in Y using our linear model with X.

(b) We can explain 6% of the variance in Y using our linear model with X.

(c) We can explain 36% of the variance in Y using our linear model with X.

(d) None of the other answers are correct.

(e) I have made a computational error somewhere when computing R^2.

Answer: Recall from the second set of single linear regression slides that when we scale Y by a constant $c \neq 0$, both $\hat{\beta}_0$ and $\hat{\beta}_1$ get scaled by c. That is, for the linear regression of $\tilde{Y} = cY$ against X:

$$\widehat{\tilde{Y}} = c\hat{\beta}_0 + c\hat{\beta}_1 X = c\widehat{Y}.$$

Calculating the R^2 from this regression of $\tilde{Y}$ against X using the formula $R^2 = \frac{\text{SSR}}{\text{SST}}$ gives:

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n}(\widehat{\tilde{Y}}_i - \bar{\tilde{Y}})^2}{\sum_{i=1}^{n}(\tilde{Y}_i - \bar{\tilde{Y}})^2} \\ &= \frac{\sum_{i=1}^{n}(c\widehat{Y}_i - c\bar{Y})^2}{\sum_{i=1}^{n}(cY_i - c\bar{Y})^2} \\ &= \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \end{aligned}$$

This is the same as the R^2 from the initial regression. So, we must have made a computational error somewhere when computing R^2.