# Econ 21020 - Problem Set 3 Solutions

Due 11/3 at 11:59 PM. Submit to Canvas. May be completed in groups of up to 6 students. Only one submission is required per group. Note that this assignment will be graded for COMPLETION and not for accuracy.

## Problem 1

We learned that one interpretation of the linear regression equation is that the coefficients $\beta_0$ and $\beta_1$ determine the function of $X$ that is the best linear approximation to $E[Y|X]$. However, "best" need not always imply "good." Consider $X \sim N(0,1)$ and $Y = X^2$. Then, $E[Y|X] = X^2$.

(a) We said that under the (equivalent) best linear predictor of $Y$ given $X$ interpretation, $\beta_0$ and $\beta_1$ will satisfy the following first order conditions:

$$\begin{aligned} E[Y - \beta_0 - \beta_1 X] &= 0 \\ E[X(Y - \beta_0 - \beta_1 X)] &= 0 \end{aligned}$$

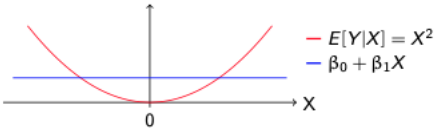
Solve this system of equations for $\beta_0$ and $\beta_1$. It may be useful to note that for $X \sim N(0,1)$, $E[X^2] = 1$ and $E[X^3] = 0$.

SOLUTION:

$$\begin{aligned} 0 &= E[Y - \beta_0 - \beta_1 X] \\ &= E[Y] - \beta_0 - \beta_1 E[X] \\ &= E[X^2] - \beta_0 - \beta_1 E[X] \\ &= 1 - \beta_0 \\ \Rightarrow \beta_0 &= 1 \\ 0 &= E[X(Y - \beta_0 - \beta_1 X)] \\ &= E[XY] - \beta_0 E[X] - \beta_1 E[X^2] \\ &= E[X^3] - E[X] - \beta_1 E[X^2] \\ &= 0 - \beta_1 \\ \Rightarrow \beta_1 &= 0 \end{aligned}$$

(b) Draw pictures of the best linear approximation to $E[Y|X]$, $\beta_0 + \beta_1 X$, and the actual $E[Y|X]$ on the same graph. The graph need not be extremely precise - it just needs to capture the major features of the functions. Interpret the result in light of the best linear approximation interpretation.

SOLUTION:



Among all possible linear functions of $X$, aka straight lines that can be drawn on the above graph, the blue line does the best job at approximating the red line (in the sense of minimizing the mean squared difference between the two). However, a reasonable subjective assessment of the above would say that the blue line and red lines are not similar, so we might consider the best linear approximation to conditional expectation to be "bad" even if it is "best" in this case.

(c) Say that our econometrician has a theory about the functional form of the conditional expectation of $Y$, and runs the regression

$$\sqrt{Y} = \beta_0 + \beta_1 X$$

on the interval $Y \geq 0$ instead. How will the performance of the linear approximation to the conditional expectation for this regression compare to that of the original estimating equation? There is no need to show any math for this part if you do not feel that you need to.

SOLUTION: Defined over the specified region of $Y$, $\sqrt{Y} = X$, and so, $E[\sqrt{Y}|X] = X$. This is of course a linear function of $X$, so the best linear approximation to $E[\sqrt{Y}|X] = X$ will simply be $E[\sqrt{Y}|X] = X$, itself - i.e. the linear approximation will now be perfect. We could notice that, in this case, interpretations 1 and 2 of the linear regression are identical.

## Problem 2

Say we are interested in studying the effect of sentencing on recidivism of juvenile offenders. Consider the all causes model:

$$Y = g(X, U)$$

where $Y = 1$ if the offender committed another crime and $Y = 0$ otherwise. Let $X = 1$ if the offender's sentence included prison time and $X = 0$ if otherwise. $U$ has the typical meaning it would take if we are claiming that $g$ is a causal model. We have observational data on $Y$ and $X$.

(a) Give two examples of unobserved determinants you think would be a part of $U$.

SOLUTION: Under the causal model interpretation, we are defining $U$ as every causal determinant of recidivism apart from the nature of the offender's original sentence. This could include an extremely wide range of things, including: the laws governing the place where the offender lives, the economic conditions in the place where the offender lives, whether or not the offender is part of any kind of criminal organization, the offender's mental condition or personality (ideally defined at the point prior to the their original sentence, so as to avoid the issue of prison sentences affecting these characteristics)...

(b) Under what assumptions would we be able to interpret the results of the regression model

$$Y = \beta_0 + \beta_1 X + U$$

causally? Does that assumption(s) seem plausible in this context? Why or why not?

SOLUTION: In order to interpret a linear regression of $Y$ on $X$ causally, we'll need to assume that $E[XU] = 0$, where $U$ is defined in the causal model sense as discussed in part (a). As we are using observational data, there is no reason to think this will hold - there is a litany of reasons why someone's sentence might be associated with other determinants of crime. One example: suppose that black offenders are more likely to receive harsher sentences for equivalent crimes. If black offenders are also more likely to live in poorer neighborhoods, this will induce a correlation between sentencing and socio-economic status of the offender's neighborhood, which could reasonably be a component of $U$.

(c) Define and interpret $Y_1$ and $Y_0$ as potential outcomes, in the sense discussed in class.

SOLUTION: We can think of these potential outcomes as being equivalent to $g(X = 1, U)$ and $g(X = 0, U)$, the causal determination function for $Y$ evaluated at the two different types of sentence, holding $U$ fixed. In English, this is whether or not the offender will re-offend if they receive prison time and if they do not receive prison time, respectively, holding all other determinants of recidivism constant across the two scenarios.

(d) Define and interpret the Average Treatment Effect (ATE) between $X = 1$ and $X = 0$. Based on your response to part (b), do you think we can recover this ATE using a linear regression?

SOLUTION: The ATE will be $E[Y_1 - Y_0]$. In this case it might be clearest to interpret if converted to probabilities:

$$\begin{aligned} E[Y_1 - Y_0] &= E[Y_1] - E[Y_0] \\ &= E[\mathbb{1}\{Y_1 = 1\}] - E[\mathbb{1}\{Y_0 = 1\}] \\ &= P\{Y_1 = 1\} - P\{Y_0 = 1\} \end{aligned}$$

aka the difference in probability of recidivism if your sentence includes prison time, averaged across the population of juvenile offenders. We learned in class that $\beta_1$ will equal the ATE of a binary treatment if we can assume random assignment, $X \perp U$. However, as discussed in part (b), there is no particular reason to think this will be the case here, so we probably cannot recover the ATE using a linear regression.

## Problem 3

In class, we made extensive use of the fact that our OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ satisfy:

$$\min_{b_0, b_1} \frac{1}{n} \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2$$

without fully justifying this. Let's walk through a demonstration that this will be the case. Assume that this minimization problem can be solved by taking first-order conditions (the second-order condition is easy to check, if desired). Then, our estimates will satisfy the first-order conditions:

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \tag{1}$$

$$\frac{1}{n} \sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \tag{2}$$

(a) Solve FOC (1) for $\hat{\beta}_0$. This should yield the OLS estimator for $\beta_0$, $\hat{\beta}_0 = \overline{Y}_n - \hat{\beta}_1 \overline{X}_n$

SOLUTION:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) &= 0 \\ \frac{1}{n} \sum_{i=1}^{n} Y_i - \hat{\beta}_0 - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} X_i &= 0 \\ \overline{Y}_n - \hat{\beta}_0 - \hat{\beta}_1 \overline{X}_n &= 0 \\ &\Rightarrow \hat{\beta}_0 = \overline{Y}_n - \hat{\beta}_1 \overline{X}_n \end{aligned}$$

(b) Plug the result from part (a) into FOC (2). Solve for $\hat{\beta}_1$. This should yield the OLS estimator for $\beta_1$, $\hat{\beta}_1 = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}^2_{X,n}}$

SOLUTION:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) &= 0 \\
\frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - (\overline{Y}_n - \hat{\beta}_1 \overline{X}_n) - \hat{\beta}_1 X_i) &= 0 \\
\frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - \overline{Y}_n + \hat{\beta}_1 \overline{X}_n - \hat{\beta}_1 X_i) &= 0 \\
\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{Y}_n \frac{1}{n}\sum_{i=1}^{n} X_i + \hat{\beta}_1 \overline{X}_n \frac{1}{n}\sum_{i=1}^{n} X_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} X_i X_i &= 0 \\
\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{Y}_n \overline{X}_n + \hat{\beta}_1 \overline{X}_n^2 - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} X_i^2 &= 0 \\
\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{Y}_n \overline{X}_n - \hat{\beta}_1 (\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2) &= 0 \\
\hat{\sigma}_{XY} - \hat{\beta}_1 \hat{\sigma}^2_{X,n} &= 0 \\
\Rightarrow \hat{\beta}_1 &= \frac{\hat{\sigma}_{XY}}{\hat{\sigma}^2_{X,n}}
\end{aligned}$$

## Problem 4

Show that the $R^2$ from a regression of $Y$ on $X$ is equal to the $R^2$ from a regression of $X$ on $Y$. Provide some intuition for this result.

SOLUTION: We'll work from the $R^2 = \frac{ESS}{TSS}$ definition, and consider the numerator and denominator separately, to begin:

$$
\begin{aligned}
ESS &= \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y}_n)^2 \\
&= \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 X_i - \overline{Y}_n)^2 \\
&= \sum_{i=1}^{n}(\overline{Y}_n - \hat{\beta}_1\overline{X}_n + \hat{\beta}_1 X_i - \overline{Y}_n)^2 \\
&= \sum_{i=1}^{n}(\hat{\beta}_1(X_i - \overline{X}_n))^2 \\
&= \hat{\beta}_1^2\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \\
&= (\frac{\hat{\sigma}_{XY}}{\hat{\sigma}_{X,n}^2})^2 n\hat{\sigma}_{X,n}^2 \\
&= n\frac{(\hat{\sigma}_{XY})^2}{\hat{\sigma}_{X,n}^2} \\
TSS &= \sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2 \\
&= n\hat{\sigma}_{Y,n}^2 \\
\Rightarrow R^2 &= \frac{n\frac{(\hat{\sigma}_{XY})^2}{\hat{\sigma}_{X,n}^2}}{n\hat{\sigma}_{Y,n}^2} \\
&= \frac{(\hat{\sigma}_{XY})^2}{\hat{\sigma}_{Y,n}^2\hat{\sigma}_{X,n}^2}
\end{aligned}
$$

If we repeated this all for a regression of $X$ on $Y$, we'd end up with an analogous $ESS = n\frac{(\hat{\sigma}_{XY})^2}{\hat{\sigma}_{Y,n}^2}$ and $TSS = n\hat{\sigma}_{X,n}^2$, and so end up with the same $R^2$.

To get some intuition, consider a graphical argument. We have some data, collected on a graph with a $Y$-axis and $X$-axis, and a line of best fit. The TSS is the total amount of variation in the data. The ESS is the amount of variation attributable to the line of best fit. Taking the ratio gives us the $R^2$. Say now we simply flip the graph so that the former $X$-axis is now the $Y$-axis and vice versa. The data points and the line of best fit remain the same, apart from a rotation. Thus, the total variation in the data and the variation in the line of best fit remain the same, implying the same $R^2$, even if we relabel the axes.

# Problem 5

Say we are comfortable assuming homoskedasticity of $U$. Let's call

$$\hat{\sigma}_{NR}^2 = \frac{\frac{1}{n}\sum_{i=1}^n \hat{U}_i^2}{\hat{\sigma}_X^2}$$

the non-heteroskedasticity robust estimator of $\sigma_1^2$ (the variance of the limiting distribution of $\sqrt{n}(\hat{\beta}_1 - \beta_1)$). Demonstrate that this estimator is consistent for $\sigma_1$, assuming homoskedasticity, our normal maintained assumptions, and $E[Y^4], E[X^4] < \infty$. (Hint 1: I would deal with the numerator on its own first and bring the denominator back in later.) (Hint 2: Remember that $\hat{\beta}_0$, $\hat{\beta}_1$, and sample variances and covariances are consistent under the assumptions that have been made.)

SOLUTION: Let's start by taking a look at the numerator. There's a few different ways we can proceed here - basically we want to rearrange the numerator into a continuous function of objects that we know will converge in probability so that we can apply the CMT and get the entire numerator to converge to something we know. Let's look at one way to do this:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \hat{U}_i^2 &= \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (Y_i - (\overline{Y}_n - \hat{\beta}_1 \overline{X}_n) - \hat{\beta}_1 X_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^n ((Y_i - \overline{Y}_n) - \hat{\beta}_1 (X_i - \overline{X}_n))^2 \\
&= \frac{1}{n}\sum_{i=1}^n ((Y_i - \overline{Y}_n)^2 + \hat{\beta}_1^2 (X_i - \overline{X}_n)^2 - 2\hat{\beta}_1 (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)) \\
&= \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2 + \hat{\beta}_1^2 \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 - 2\hat{\beta}_1 \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n) \\
&= \hat{\sigma}_{Y,n}^2 + \hat{\beta}_1^2 \hat{\sigma}_{X,n}^2 - 2\hat{\beta}_1 \hat{\sigma}_{X,Y} \\
\Rightarrow \frac{1}{n}\sum_{i=1}^n \hat{U}_i^2 &\xrightarrow{p} \sigma_Y^2 + \beta_1^2 \sigma_X^2 - 2\beta_1 \sigma_{X,Y} \qquad \text{(CMT)}
\end{aligned}$$

where the last line follows because the penultimate line is a continuous function of four things, $\hat{\sigma}_{Y,n}^2$, $\hat{\sigma}_{X,n}^2$, $\hat{\sigma}_{X,Y}$, and $\hat{\beta}_1$, that we all know converge in probability to the objects they are respectively estimating. Of course the result that the numerator is converging in probability is only useful if it is converging to the actual numerator of $\sigma_1$. Under the assumption of homoskedasticity the numerator is $Var(U)$, as we showed in class. Lets then show that $Var(U)$ is equal to the final line above:

$$\begin{aligned}
Var(U) &= Var(Y - \beta_0 - \beta_1 X) \\
&= Var(Y) + \beta_1^2 Var(X) - 2\beta_1 Cov(X, Y) \\
\Rightarrow \frac{1}{n}\sum_{i=1}^{n} \hat{U}_i^2 &\xrightarrow{p} Var(U)
\end{aligned}$$

the second line above follows from the nature of variances of sums of random variables. Finally, we know that $\hat{\sigma}_X^2 \xrightarrow{p} Var(X)$. Using this and the assumption that $Var(X) > 0$, we apply the CMT again to finish:

$$\hat{\sigma}_{NR}^2 = \frac{\frac{1}{n}\sum_{i=1}^{n} \hat{U}_i^2}{\hat{\sigma}_X^2} \xrightarrow{p} \frac{Var(U)}{Var(X)} = \sigma_1^2$$

where the last equal sign holds under the assumption of homoskedasticity, as shown in class.

# Problem 6

This exercises uses observational (non-experimental) data on the years of schooling and the (log) weekly wage of 329,509 observations of American men born between 1930-1939, as was used in the paper Angrist and Krueger (199). This data is available on Canvas, under Modules PSet Data, as “ak91.csv”. We will consider the variables “education”, which represents the years of education completed by the man and “log_weekly_wage” which represents the log weekly wage. Consider these variables to be an iid sample $(Y_i, X_i) \sim (Y, X)$ where $Y$ is years of education and $X$ is log wage of American men born between those years. Assume that $E[X^4], E[Y^4] < \infty$.

(a) Which of the three interpretations of linear regression from class do you think would be most appropriate for the regression equation:

$$Y = \beta_0 + \beta_1 X + U$$

where $Y$ and $X$ are defined as above. Why?

SOLUTION: The linear conditional expectation interpretation and causal model interpretations involve meaningful assumptions about the data and world, so lets evaluate if either of those are appropriate first. The linear conditional expectation interpretation necessitates the assumption that the conditional expectation of (log) wage is linear in years of education. There’s no particular reason to think this will be true, as there are many conceivable non-linearities in this conditional expectation. I would, for one thing, imagine that the conditional expectation would be somewhat “jumpy.” Maybe there’s not a huge difference in average pay for people who dropped out of high school in 10th vs 11th grade. However, there might be a jump up upon completion of high school. There might also generally be increasingly or decreasingly large returns to additional years of school (i.e. the difference between a grad degree and a bachelor's is, on average, larger/smaller than the difference between a bachelor's and a HS degree). In any case, its hard to motivate a completely linear conditional expectation.

A causal model interpretation would require that $E[XU] = 0$, where $U$ is defined in the causal model sense of every other determinant of wage other than years of schooling. As discussed in class, this is unlikely to hold. It is easy to imagine a correlation between years of education and family background, for instance, where family background might be part of $U$. Thus, the causal model interpretation seems like a poor fit.

Then, by default, the best linear predictor interpretation seems like the best fit, as it is the most general and requires no assumptions that would be unreasonable for this context.

(b) Perform the indicated regression using the data, without heteroskedasticity-robust standard errors. Interpret the coefficients you get in light of your response to (a).

SOLUTION: Given our answer to (a), I interpret the coefficients as describing the best linear predictor. That would mean, I would predict that someone with "0" years of education would earn a log weekly wage of about 5, and each additional year of education would lead me to predict an increase in the log weekly wage of 0.07 log dollars. This would represent the estimates of the best prediction of wage we can make given a linear function of wage (while it is nice to be "best" in this class of prediction functions, as touched on in part (a), its in the realm of possibility that a linear function may simply be a poorly performing predictor).

(c) Re-run the regression with heteroskedasticity-robust standard errors. Did the coefficients and/or standard errors change? Why?

SOLUTION: We can note that our coefficients did not change at all. This is unsurprising, as the assumption of homoskedasticity has no relevance to how we estimate the parameters of a linear regression: we use the OLS estimators either way. However, assuming homoskedasticity or not *does* change how we estimate standard errors. The standard error is the ratio of the estimated variance of the limiting distribution to the root of the sample size. The way we estimate the variance of the limiting distribution changes depending on whether or not we assume heteroskedasticity, so it makes sense that our errors changed.

(d) In our setting, which type of standard errors would you consider appropriate? Why?

SOLUTION: I would consider the robust standard errors to be more appropriate. There is no particular reason to believe that the variance in the $U$'s will be constant across all amounts of education. As discussed in class, for instance, it may be the case that the distribution of wages increases with additional years of schooling, as more educated people are more likely to be "super-rich," widening the gap between the top percentiles and the average earner at higher levels of education (even if the average earner themselves is earning more).