Econ 103: Homework 3 Solutions

Manu Navjeevan

September 6, 2021

Single Linear Regression Review

1. (Challenge, Linear Regression as Line of Best Fit). Recall that our single linear regression model, defined in terms of the "line of best fit" is an approximation of the true conditional mean rather than the true conditional mean. However, in the case that X is binary, ($X \in \{0, 1\}$), the parameters β_0 and β_1 from the linear model

$$Y = \beta_0 + \beta_1 X + \epsilon, \;\; \mathbb{E}[\epsilon] = \mathbb{E}[\epsilon X] = 0.$$

exactly describe the conditional mean. In this exercise we will show this.

   (a) Use the following equalities, true for a random variable X that takes values $X \in \{0, 1\}$, to get an expression for $\text{Cov}(X, Y)$.

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[Y|X = 0]\Pr(X = 0) + \mathbb{E}[Y|X = 1]\Pr(X = 1) \\
\mathbb{E}[X] &= \Pr(X = 1) \\
\mathbb{E}[XY] &= \mathbb{E}[Y|X = 1]\Pr(X = 1)
\end{aligned}$$

   It may be helpful to let $p = \Pr(X = 1)$ and note that $\Pr(X = 0) = 1 - p$.

   Answer: Let $p = \Pr(X = 1)$. Using the equalities:

$$\begin{aligned}
\mathbb{E}[X] &= p \\
\mathbb{E}[Y] &= \mathbb{E}[Y|X = 1]p + \mathbb{E}[Y|X = 0](1 - p) \\
\mathbb{E}[YX] &= \mathbb{E}[Y|X = 1]p
\end{aligned}$$

   we can write

$$\begin{aligned}
\text{Cov}(X, Y) &= \mathbb{E}[YX] - \mathbb{E}[X]\mathbb{E}[Y] \\
&= \mathbb{E}[Y|X = 1]p - p\left(\mathbb{E}[Y|X = 1]p + \mathbb{E}[Y|X = 0](1 - p)\right) \\
&= \mathbb{E}[Y|X = 1]p - p\left(\mathbb{E}[Y|X = 1]p + \mathbb{E}[Y|X = 0] - \mathbb{E}[Y|X = 0]p\right) \\
&= \underbrace{\mathbb{E}[Y|X = 1]p - p^2\mathbb{E}[Y|X = 1]}_{\text{pull out p}} - \underbrace{p\mathbb{E}[Y|X = 0] + \mathbb{E}[Y|X = 0]p^2}_{\text{pull out p}} \\
&= p\underbrace{\left(\mathbb{E}[Y|X = 1] - p\mathbb{E}[Y|X = 1]\right)}_{\text{pull out } \mathbb{E}[Y|X=1]} - p\underbrace{\left(\mathbb{E}[Y|X = 0] - p\mathbb{E}[Y|X = 0]\right)}_{\text{pull out } \mathbb{E}[Y|X=0]} \\
&= p(1 - p)\mathbb{E}[Y|X = 1] - p(1 - p)\mathbb{E}[Y|X = 0]
\end{aligned}$$

   Any answer after line 2 would be accepted. Just important to set up for part (b)

   (b) Use the following expression, true for a random variable X that takes values $X \in \{0, 1\}$, to get a simplified expression for $\beta_1 = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$:

$$\text{Var}(X) = \Pr(X = 1)\Pr(X = 0).$$

Answer: Using the definition $p = \Pr(X = 1)$ and noting that $\Pr(X = 0) = 1 - \Pr(X - 1) = 1 - p$, we can write $\text{Var}(X) = p(1 - p)$. Using our answer from part (a), we can simplify:

$$\begin{aligned}\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} &= \frac{p(1-p)\mathbb{E}[Y|X=1] - p(1-p)\mathbb{E}[Y|X=0]}{p(1-p)} \\ &= \mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0]\end{aligned}$$

(c) Use the expressions for $\mathbb{E}[Y]$ and $\mathbb{E}[X]$ above, as well as the expression for β_1 that you derived in part (b) to get a simplified expression for

$$\beta_0 = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X].$$

Answer: Using $\beta_1 = \mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 0]$, $\mathbb{E}[X] = p$, and $\mathbb{E}[Y] = p\mathbb{E}[Y|X = 1] + (1 - p)\mathbb{E}[Y|X = 0]$:

$$\begin{aligned}\beta_0 = \mathbb{E}[Y] - \beta_1\mathbb{E}[X] &= \underbrace{p\mathbb{E}[Y|X=1] + (1-p)\mathbb{E}[Y|X=0]}_{\mathbb{E}[Y]} - \underbrace{p(\mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0])}_{\beta_1\mathbb{E}[X]} \\ &= (1-p)\mathbb{E}[Y|X=0] + p\mathbb{E}[Y|X=0] \\ &= \mathbb{E}[Y|X=0]\end{aligned}$$

(d) Use the expressions for β_0 and β_1 from above as well as the linear model:

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

What is the predicted value of Y when $X = 0$? What about when $X = 1$?

Answer: The predicted value of Y when $X = 0$ is $\beta_0 = \mathbb{E}[Y|X = 0]$. The predicted value of Y when $X = 1$ is $\beta_0 + \beta_1 = \mathbb{E}[Y|X = 1]$. While normally the line of best fit does not coincide with the true conditional expectation, in this case it does.

Multiple Linear Regression

1. (Single Hypothesis Testing). Consider the linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

We want to test the hypotheses:

$$H_0 : \beta_2 = 0 \quad \text{vs.} \quad H_1 : \beta_2 \neq 0$$

at level $\alpha = 0.05$.

(a) Suppose on a sample of size $n = 100$ we find that $\sigma_\epsilon^2 = 400$, $\sigma_{X_2}^2 = 200$, $\hat{\beta}_2 = 1$, and $\rho_{12}^2 = 0.5$, where we recall that ρ_{12} is the sample correlation coeffecient between X_1 and X_2. Conduct the hypothesis test in the setup of this problem.

Answer: We first use this information to calculate

$$\hat{\sigma}_{\beta_2}^2 = \frac{\sigma_\epsilon^2}{(1 - \rho_{12})^2 \sigma_X^2} = 4 \implies \sigma_{\beta_2}/\sqrt{n} = 2/10 = 0.2.$$

We can then construct our test statistic

$$t^* = \frac{\hat{\beta}_2 - 0}{\sqrt{\beta_2}/\sqrt{n}} = \frac{1}{0.2} = 5.$$

Since this test statistic is larger than $z_{1-\alpha/2} = 1.96$ we reject the null hypothesis and conclude in favor of the alternative that $\beta_2 \neq 0$.

(b) Give an intuitive explanation for why the variance of $\hat{\beta}_1$ is increasing with the correlation between X_1 and X_2.

Answer: Recall that in a multiple regression model, we interpret the coeffecient β_1 as the (approximate) association between X_1 and Y while holding X_2 constant. If,. in our data, X_1 and X_2 are highly correlated, it is difficult for us to parse out this effect since whenever X_1 moves, X_2 tends to move too. This makes it difficult to distinguish the effect of X_1 on Y from the effect of X_2 on Y and reduces our certainty in what the true parameter β_1 is. The variance of our estimator $\hat{\beta}_1$ can be interpreted as a measure of our uncertainty about the true parameter β_1, so a high correlation will increase the variance of $\hat{\beta}_1$. (Variations on this answer will be accepted, the main point is to remark that it is difficult to parse out the effect of X_1 on Y from the effect on X_2 on Y).

2. (Single Hypothesis Testing). Suppose we are interested in exploring the relationship between income, years of education, and experience. To investigate this relationship, we consider the following model:

$$\ln(\text{Income}) = \beta_0 + \beta_1 \text{Edu} + \beta_2 \text{Exper} + \epsilon.$$

After fitting this model with sample size $n = 100$ we find the following variance covariance matrix.

$$\text{Cov}(\hat{\beta}) = \begin{array}{c} \\ \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{array} \begin{array}{c} \begin{array}{ccc} \hat{\beta}_0 & \hat{\beta}_1 & \hat{\beta}_2 \end{array} \\ \begin{pmatrix} 0.05 & 0.25 & 0.16 \\ 0.25 & 0.08 & 0.1 \\ 0.16 & 0.1 & 0.36 \end{pmatrix} \end{array}$$

We want to prove that returns to education are larger than returns to experience.

(a) Formally state, in terms of parameters of the model, the null and alternative hypotheses associated with this test (Hint: Recall the null is that returns to education are smaller than returns to experience, our goal will be to provide evidence against this null hypothesis).

Answer: The null hypothesis can be expressed as $H_0 : \beta_1 \leq \beta_2 \iff \beta_1 - \beta_2 \leq 0$. The alternative hypothesis is then $H_1 : \beta_1 > \beta_2 \iff \beta_1 - \beta_2 > 0$.

(b) Suppose we find that $\hat{\beta}_1 = 1.1$ and $\hat{\beta}_2 = 0.7$. What is the result of running the hypothesis test specified in part (a) at level $\alpha = 0.05$? (Hint: It may be useful to recall that we can write $X - Y = X + (-Y)$).

Answer: We will consider the linear combination of parameters $\lambda = \beta_1 - \beta_2$ and test the null hypothesis $H_0 : \lambda \leq 0$ against the one sided alternative $H_1 : \lambda > 0$. Using the covariance matrix above, we find that

$$\text{Var}(\hat{\lambda}) = \text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) - 2\,\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = 0.08 + 0.36 - 2 \cdot 0.1 = 0.24.$$

Then we can construct our test statistic

$$t^* = \frac{\hat{\lambda} - 0}{\sqrt{\text{Var}(\hat{\lambda})}} = \frac{0.4}{0.489} = 0.8164.$$

Given this, since $t^* \leq z_{1-\alpha} = 1.64$, we fail to reject this null hypothesis. We cannot reject the hypothesis that returns to experience are higher than returns to education.

(c) Keeping all other values the same, what is the largest value of $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$ for which we would reject this null hypothesis? (This may be larger or smaller than the existing covariance).

Answer: So, the correct answer to this is to notice that the variance of $\hat{\lambda}$ is decreasing as the Covariance is increasing. For small variances we reject the null hypothesis. Our only restriction

is that the variance needs to positive. So, we can take $2\,\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) < \mathrm{Var}(\hat{\beta}_0) + \mathrm{Var}(\hat{\beta}_1) \implies \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) < 0.22$.

However, in office hours, I was confused and mentioned that you can find an upper bound by inverting the test statistic (using $t^* > 1.64$). This gives a lower bound on rejecting the null hypothesis $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) > 0.19$. This answer will also be accepted as well as an answer that claims that the upper bound is unattainable.

3. (Multiple Hypotheses Testing). Suppose a hamburger restaurant is investigating the relationship between the number of burgers it sells in a month, the price of a burger in dollars, and the money it spends on advertising in tens of thousands of dollars, and whether or not it is open on Saturdays.

   Consider the following unrestricted model:

$$\text{Sales} = \beta_0 + \beta_1 \text{Price} + \beta_2 \text{Advert} + \beta_3 \text{Saturdays} + \epsilon.$$

   And the restricted model

$$\text{Sales} = \beta_0 + \beta_1 (\text{Advert} + \text{Saturdays} - \text{Price}) + \epsilon.$$

   (a) In terms of the unrestricted model parameters, state the null hypothesis being imposed by the restricted model (something like $H_0 : \beta_1 = 2\beta_2 = 20\beta_3$).

   Answer: The restricted model is imposing that a ten thousand dollar increase in advertising is associated with the same change in sales as being open on Saturday or a one dollar decrease in price. Formally, we can write this as:

$$H_0 : -\beta_1 = \beta_2 = \beta_3.$$

   (b) Interpret this null hypothesis in context.

   Answer: See above. A correct answer should include units. It is ok to use language like "the effect of. . . "

   (c) Suppose $n = 104$ and, after estimating both the restricted and unrestricted models, we find that $\mathrm{SSE_R} = 1000$, $\mathrm{SSE_U} = 800$. Use this information the compute the F-statistic.

   Answer: First, note that $n-p-1 = 100$ and $J = 2$ (count the equality signs in the null hypothesis). Using this, we compute F^*:

$$F^* = \frac{(\mathrm{SSE_R} - \mathrm{SSE_U})/J}{\mathrm{SSE_U}/(n-p-1)} = \frac{(1000-800)/2}{800/100} = \frac{100}{8} = 12.5.$$

   (d) Using the command $\mathrm{pf}(F^*, J, n-p-1)$ in R, compute the p-value. Recall that:

$$\Pr\left(F(J, n-p-1) \leq c\right) = \mathrm{pf}(c, J, n-p-1).$$

   Answer: We can calculate the p-value via $p = 1 - \mathrm{pf}(F^*, J, n-p-1) = 1 - \mathrm{pf}(12.5, 2, 100) \approx 0$.

   (e) Using this p-value report the result of the test at level $\alpha = 0.05$. Interpret the test result in the context of the problem.

   Answer: Since the p-value is less than $\alpha = 0.05$ we reject this null hypothesis. We conclude in favor of the alternative hypothesis, which is to say that either a one dollar decrease in price is associated with a different change in sales than a ten thousand dollar increase in advertising or a ten thousand dollar increase in advertising is associated with a different change in sales than being open Saturday. (There are many ways to state this, just important that you note that at least one of two restrictions is violated and units are used).

4. (Polynomial Modeling). When estimating wage equations, we expect that young, inexperienced workers will have relatively low wages and that with additional experience their wages will rise, but then begin to decline after middle age, as the worker nears retirement. This life cycle pattern of wages can be captured by introducing experience and experience squared to explain the level of wages. If we also include years of education, we have the equation

$$\text{Wages} = \beta_0 + \beta_1 \text{Educ} + \beta_2 \text{Exper} + \beta_3 \text{Exper}^2 + \epsilon.$$

   (a) In terms of the parameters of this model, what is the expected marginal effect of experience on wages?

   Answer: Taking derivatives gives

$$\frac{\partial \widehat{Y}}{\partial \text{Exper}} = \beta_2 + 2\beta_3 \text{Exper}.$$

   (b) Given the explanation above, what signs do we expect on the coeffecients β_2 and β_3?

   Answer: Since we expect returns on experience to be positive to begin with, but then diminish as experience increases, we would expect $\beta_2 > 0$ and $\beta_3 < 0$.

   (c) Suppose we estimate that $\hat{\beta}_2 = 20$ and $\hat{\beta}_3 = -0.6$. After how many years of experience do we esimate that wages will start to decline?

   Answer: Setting the marginal effect equal to zero and solving for experience gives

$$20 - 2 \cdot 0.6 \text{Exper} = 0 \implies \text{Exper} = \frac{20}{1.2} = 16.6666.$$

   An answer of 16.6666, 16, or 17, would be accepted.

5. (Omitted Variables Bias). Consider the two models:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$
$$Y = \beta_0^\circ + \beta_1^\circ X_1 + \beta_2^\circ X_2 + \epsilon^\circ.$$

   Recall that the omitted variables bias is the difference between β_1 and β_1°, $\text{OVB} = \beta_1 - \beta_1^\circ$.

   (a) From lecture, give the formula for the omitted variables bias.

   Answer: From lecture:

$$\text{OVB} = \beta_1 - \beta_1^\circ = \beta_2^\circ \frac{\text{Cov}(X_1, X_2)}{\text{Var}(X_1)}.$$

   (b) Suppose that X_2 has a negative relationship with the outcome and X_1 and X_2 are negatively related. What is the sign of the omitted variables bias? Which should be larger, β_1 or β_1°?

   Answer: Using the formula above and the fact that a negative times a negative is positve, we find that $\text{OVB} > 0$. This means that $\beta_1 > \beta_1^\circ$.

   (c) (Challenge). Give an example that illustrates this. That is, come up with an example in which X_1 and X_2 are negatively related and X_2 is negatively associated with the outcome. Then, within the context of the example, give an explanation for why excluding X_2 from your model would make the coeffecient on X_1 either larger or smaller. This explanation should not just use the omitted variables formula and rather provide reasoning within the context of the example.

   Answer: Many possible answers. A good answer 1) clearly explains why the covariance should be negative, 2) clearly explains why the association between X_2 and Y should be negative, 3) explains, within the context of the example and without just using the formula, why this would lead to a larger slope parameter on the model that just includes X_1 and excludes X_2.