Econ 103: Homework 3

Manu Navjeevan

Due: End of Day, Monday, September 6th

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 $\Pr(X = 0) = 1 - p$.

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

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

   (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 $X = 1$?

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:
$$H_0 : \beta_2 = 0 \;\; \text{vs.} \;\; H_1 : \beta_2 \neq 0$$
   at level $\alpha = 0.05$.

(a) Suppose on a sample of size $n = 100$ we find that $\sigma^2_\epsilon = 400$, $\sigma^2_{X_2} = 200$, $\hat{\beta}_2 = 1$, and $\rho^2_{12} = 0.5$, 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.

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

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).

(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 $X - Y = X + (-Y)$).

(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).

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 $H_0 : \beta_1 = 2\beta_2 = 20\beta_3$).

(b) Interpret this null hypothesis in context.

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

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

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

(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.

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?

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

   (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?

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.

   (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°?

   (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.