Problem Set #5: The MLR Model

Economics 435: Quantitative Methods

Fall 2011

1 Consequences of leaving out a quadratic term

Suppose that:

$$E(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

and you estimate the (misspecified) regression:

$$y = \gamma_0 + \gamma_1 x + u$$

where you also assume that SLR1-SLR4 hold.

a) Let $\gamma_1 \equiv \text{plim } \hat{\gamma}_1$. Find γ_1 in terms of β_1, β_2, $E(x)$, $E(x^2)$, and $E(x^3)$. You will find it helpful to make use of the formula:

$$\hat{\gamma}_1 = \frac{c\hat{o}v(x,y)}{v\hat{a}r(x)}$$

b) Let $\delta(X) = \frac{\partial E(y|x=X)}{\partial X}$ be the marginal predictive effect of x on y. Find $\delta(X)$ as a function of β_1 and β_2.

c) If the true CEF is linear, i.e., if $\beta_2 = 0$, then $\hat{\gamma}_1$ will be a consistent estimate of the marginal effect $\delta(X)$. If it is not, then the linear regression model can be considered as a linear approximation to the true (quadratic) CEF. The quality of this approximation will depend on the value of X. Specifically, there will be one value X^* for which $\delta(X^*) = \gamma_1$ exactly. Near that value, the two will be similar but not identical. Find X^*.

2 Best linear predictors

The MLR model imposes the strong assumption that $E(y|x)$ is linear. However, we often find ourselves in situations where this assumption is probably false - for example, where the dependent variable is discrete. These situations are sometimes handled with nonparametric regression, and sometimes with more complex nonlinear models (like the probit, logit, and tobit models we will learn in a few weeks). However, we might wonder what will happen if we just use OLS.

Let (x, y) be a pair of random variables. The *best linear predictor* (BLP) of y given x is defined as the linear function

$$blp(x) = b_0 + b_1 x$$

that minimizes:

$$E\left((y - b_0 - b_1 x)^2\right)$$

Note that this definition makes no assumptions about the statistical relationship between y and x. In particular, we are not assuming that $E(y|x)$ is linear in x.

a) Show that:

$$\begin{aligned} b_1 &= cov(x,y)/var(x) \\ b_0 &= E(y) - b_1 E(x) \end{aligned}$$

Hint: to find the minimum of a convex function, take the derivative and set it equal to zero.

b) Suppose that assumptions MLR2 and MLR4 hold, i.e., we have a random sample of size n on (y, x), and that the sample exhibits variation in x. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the OLS regression coefficients from this sample. Are $\hat{\beta}_0$ and $\hat{\beta}_1$ consistent estimators of b_0 and b_1? Prove it.

c) Suppose that we are interested in finding the *best linear approximation* to the true CEF, i.e.:

$$BLA(x) = a_0 + a_1 x$$

where a_0 and a_1 minimize:

$$E\left((E(y|x) - a_0 - a_1 x)^2\right)$$

Show that the BLA is the same as the BLP, i.e., $a_0 = b_0$ and $a_1 = b_1$.

3 Fitted values and residuals

Suppose we are estimating an SLR[1] model. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the usual OLS estimators of β_0 and β_1. Let:

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ \hat{u}_i &= y_i - \hat{y}_i \end{aligned}$$

a) Is the fitted value $\hat{y}_i$ a consistent estimator of y_i? That is, is it the case[2] that:

$$\text{plim } \hat{y}_i = y_i$$

Provide an argument (ideally, a proof) that your answer is correct.

b) Is the residual $\hat{u}_i$ a consistent estimator of u_i? That is, is it the case that:

$$\text{plim } \hat{u}_i = u_i$$

Provide an argument (ideally, a proof) that your answer is correct.

c) Suppose that you are concerned that the data have heteroskedasticity. You are willing to assume that any heteroskedasticity takes a linear form:

$$E(u^2|x) = \alpha_0 + \alpha_1 x$$

After estimating the original regression (of y on x) and calculating the fitted values, you use OLS to estimate a regression of the squared residual $\hat{u}_i^2$ on x_i. Let the resulting regression coefficients be $\hat{\alpha}_0$ and $\hat{\alpha}_1$. Is it the case that:

$$\text{plim } \hat{\alpha}_1 = \alpha_1$$

Provide an argument (a proof is not necessary here) that your answer is correct.

[1] In case you didn't write down the definition of the SLR model: Let (y, x, u) be a triplet of random variables such that (SLR1) $y = \beta_0 + \beta_1 x + u$ (SLR2) We have a random sample (y_i, x_i) of size n on the random variables (y, x) (SLR3) $E(u|x) = 0$ (SLR4) There is variation in x_i in the sample.

[2] If you're uncomfortable with the idea of a random variable having another random variable as its probability limit, then just think of an equivalent statement: plim $(\hat{y}_i - y_i) = 0$.