Econometrics HW6

陈子睿 15220212202842

November 27, 2023

EX 1

Solution From the graph we can see that there's a non-linear(quadratic) relationship between X and Y. If we try attempt to estimate this relationship using a simple linear regression model, we are likely to introduce specification error, which may lead to omitted variable bias, which is given by:

$$Bias = \frac{Cov(X,Z)}{Var(X)} \times \beta_Z$$

where β_Z is the effect of the omitted variable Z (which captures the non-linear component of the relationship) on the dependent variable Y.

In this case, we consider two separate intervals of X:

1. For lower values of X, the slope of the curve is negative, meaning that an increase in X leads to a decrease in Y. If we fit a straight line to the entire data, the negative slope of the lower values will pull the estimated linear effect (slope) downwards.

2. For higher values of X, the slope of the curve is positive, meaning that an increase in X leads to an increase in Y. This positive slope for higher values of X will pull the estimated linear effect upwards.

When we combine these two opposing effects in a simple regression model, they will partially cancel each other out, leading to an estimated slope (β) that is closer to zero than the true relationship at any given point on the curve. This is because the linear model cannot capture the curvature of the relationship and instead averages out the effects across all values of X, leading to a biased estimate of the relationship between X and Y.

EX 2

Solution (a) We're given two equations:

$$L_i = 2H_i + u_i$$

$$H_i = -L_i + \epsilon_i$$

To find the OLS estimates of H_i on L_i, we can simply project L_i on H_i using:

$$\hat{\beta}_{OLS} = \frac{Cov(H_i, L_i)}{Var(H_i)} = Cov(H_i, L_i)$$

(b) But health (H) affects labor supply (L), and labor supply also affects health, which leads to endogeneity.The presence of endogeneity shows that $Cov(H_i, L_i)$ is NOT the correct measure to use since H_i is determined simultaneously with L_i.

If we don't addressing endogeneity, we can substitute H_i to get:

$$\begin{aligned} L_i &= 2H_i + u_i \\ &= 2(-L_i + \epsilon_i) + u_i \\ &= \frac{2}{3}\epsilon_i + \frac{1}{3}u_i \end{aligned}$$

So the computed estimator $\hat{\beta}_{OLS}$ would be based on the covariance between L_i and ϵ_i, which is zero by assumption, but the true effect of health on labor supply is a 2-hour increase in working hours for each 1 unit increase in health.

EX 3

Solution For simplicity, we assume that both ϵ_i and u_i, ϵ_i and X_i are uncorrelated, and ϵ_i is a zero-mean distubance with positive variance, i.e.

$$Cov(\epsilon_i, u_i) = 0, \quad Cov(\epsilon_i, X_i) = 0, \quad \mathbb{E}(\epsilon_i) = 0, \quad Var(\epsilon_i) > 0$$

Mathematically, the biased OLS estimator $\tilde{\beta}_{OLS}$ is given by:

$$\begin{aligned} \tilde{\beta}_{OLS} &\xrightarrow{p} \frac{Cov(Y, \tilde{X})}{Var(\tilde{X})} \\ &= \frac{Cov(Y, X + \epsilon)}{Var(X) + Var(\epsilon)} \\ &= \frac{Cov(Y, X) + Cov(Y, \epsilon)}{Var(X) + Var(\epsilon)} \\ &= \frac{Cov(Y, X)}{Var(X)} \times \frac{Var(X)}{Var(X) + Var(\epsilon)} + \frac{Cov(Y, \epsilon)}{Var(X) + Var(\epsilon)} \\ &= \beta^\circ \frac{Var(X)}{Var(X) + Var(\epsilon)} + \frac{Cov(\beta^\circ X + u, \epsilon)}{Var(X) + Var(\epsilon)} \\ &= \beta^\circ \frac{Var(X)}{Var(X) + Var(\epsilon)} + \beta^\circ \frac{Cov(X, \epsilon)}{Var(X) + Var(\epsilon)} + \frac{Cov(u, \epsilon)}{Var(X) + Var(\epsilon)} \\ &= \beta^\circ \frac{Var(X)}{Var(X) + Var(\epsilon)} \end{aligned}$$

Where β° is the true OLS coefficient. Since $Var(\epsilon) > 0$, OLS estimator $\tilde{\beta}_{OLS}$ is biased.

Therefore, the OLS estimator of the coefficient on $\tilde{X}_i$ is NOT consistent.

Next we consider the variance of the OLS estimator of the coefficient on $\tilde{X}_i$, i.e. $\tilde{\beta}_{OLS}$.

We know that the large sample variance of $\hat{\beta}_{OLS}$ is:

$$Var(\hat{\beta}_{OLS}) = \frac{1}{n} \frac{var[(X_i - \mu_X)u_i]}{[Var(X_i)]^2}$$

Then for $\tilde{\beta}_{OLS}$, we have that:

$$\begin{aligned} Var(\tilde{\beta}_{OLS}) &= \frac{1}{n} \frac{Var[(\tilde{X}_i - \mu_{\tilde{X}})u_i]}{[Var(\tilde{X}_i)]^2} \\ &= \frac{1}{n} \frac{Var[(X_i + \epsilon_i - \mu_X - \mu_\epsilon)u_i]}{[Var(X_i) + Var(\epsilon)]^2} \\ &= \frac{1}{n} \{ \frac{Var[(X_i - \mu_X)u_i]}{[Var(X_i) + Var(\epsilon)]^2} + \frac{Var[(\epsilon_i - \mu_\epsilon)u_i]}{[Var(X_i) + Var(\epsilon)]^2} \} \\ &= Var(\hat{\beta}_{OLS}) \times [\frac{Var(X_i)}{Var(X_i) + Var(\epsilon)}]^2 + \frac{1}{n} \frac{Var[(\epsilon_i - \mu_\epsilon)u_i]}{[Var(X_i) + Var(\epsilon)]^2} \end{aligned}$$

The above discussion is quite complicated, here we provide another approach. Consider the classical OLS estimator:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$$

Therefore, the OLS estimator of the coefficient on $\tilde{X}_i$ is given by:

$$\tilde{\beta}_{OLS} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y$$

For the variance of the OLS estimator, we have:

$$\begin{aligned}
Var(\hat{\beta}) &= \mathbb{E}[Var(\hat{\beta}|X)] \\
&= \mathbb{E}[(\hat{\beta} - \beta^{\circ})(\hat{\beta} - \beta^{\circ})'|X] \\
&= (X'X)^{-1}X'\mathbb{E}[uu'|X]X(X'X)^{-1} \\
&= \sigma^2\mathbb{E}(X'X)^{-1}
\end{aligned}$$

Under conditional homoskedasticity, i.e. $Var(u|X) = \sigma^2$. Actually, through our discussion, we can find that:

$$Var(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$$

We've known that the key for the consistency of the OLS estimator $\tilde{\beta}$ for β° is to check if $\mathbb{E}(X_iu_i) = 0$, we are forced to estimate the following regression model:

$$Y = \beta^{\circ}\tilde{X} + \nu$$

where ν is some unobservable disturbance.

Obviously, the disturbance ν is different from the true disturbance u. Although the linear regression model is correctly specified, we no longer have $\mathbb{E}(\nu|X) = 0$ due too the existence of the measurement errors. From above equation, we can obtain:

$$\begin{aligned}
\nu &= Y - \beta^{\circ}\tilde{X} \\
&= (\beta^{\circ}X + u) - \beta^{\circ}(X + \epsilon) \\
&= u - \beta^{\circ}\epsilon
\end{aligned}$$

The regression error ν contains the true disturbance u and a linear combination of measurement error.

The for the expectation we have:

$$\begin{aligned}
\mathbb{E}(\tilde{X}_i\nu_i) &= \mathbb{E}[(X_i + \epsilon_i)\nu_i] \\
&= \mathbb{E}(X_i\nu_i) + \mathbb{E}(\epsilon_i\nu_i) \\
&= 0 - \beta^{\circ}\mathbb{E}(\epsilon_i^2) \\
&= -\beta^{\circ}Var(\epsilon) \\
&\neq 0
\end{aligned}$$

Therefore, applying WLLN, the OLS estimator

$$\begin{aligned}
\tilde{\beta} - \beta^{\circ} &= Q_{\tilde{x}\tilde{x}}^{-1}n^{-1}\hat{\sum}_{i=1}^{n}\tilde{X}_i\nu_i \\
&\xrightarrow{p} Q_{\tilde{x}\tilde{x}}^{-1}\mathbb{E}(\tilde{X}_i\nu_i) \\
&= -\beta^{\circ}Var(\epsilon)Q_{\tilde{x}\tilde{x}}^{-1} \\
&\neq 0
\end{aligned}$$

In the words, $\tilde{\beta}$ is not consistency for β°.

To discuss the efficiency of $\tilde{\beta}$, we first note that

$$\tilde{\beta} = \beta^\circ[1 - Var(\epsilon)Q_{\tilde{x}\tilde{x}}^{-1}]$$

It's easy to find that

$$Var(\tilde{\beta}) = Var(\beta^\circ[1 - Var(\epsilon)Q_{\tilde{x}\tilde{x}}^{-1}]) = [1 - Var(\epsilon)Q_{\tilde{x}\tilde{x}}^{-1}]^2 Var(\beta^\circ)$$