Lecture notes ~~[illegible]~~ V

Linear regression with multiple regressors.

SW Ch. 6:

(1) $Y_i = \beta_0 + \beta_1 X_{1i} + u_i$

u_i contains all omitted factors - observed and unobserved.

Example:

Y_i = log income

X_{1i} = years of schooling

Then u_i contains effect of e.g.

- parents' level of schooling
- i' years of experience
- type of occupation

Is u_i uncorrelated with X_{1i}?

Assume ~~[illegible]~~ ~~[illegible]~~ true model is

(2) $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \tilde{u}_i$, where

X_{2i} = years of experience, $\tilde{u}_i$ error term such that $E(\tilde{u}_i \mid X_{1i}, X_{2i}) = 0$

$\Rightarrow$ (3) $u_i = \beta_2 X_{2i} + \tilde{u}_i$

Then $Cov(X_{1i}, u_i)$ in ① becomes

$$Cov(X_{1i}, \beta_2 X_{2i} + \tilde{u}_i) = \beta_2 Cov(X_{1i}, X_{2i})$$
$$+ \underbrace{Cov(X_{1i}, \tilde{u}_i)}_{\left(=0 \text{ if } (2) \text{ is } \underline{\text{true}} \text{ model}\right)} = \beta_2 Cov(X_{1i}, X_{2i}) \neq 0$$

Recall

$$\hat{\beta}_1 = \beta_1 + \frac{\sum (X_{1i} - \bar{X}_1) u_i}{\sum (X_{1i} - \bar{X}_1)^2} \xrightarrow{p} \beta_1 + \frac{Cov(X_{1i}, u_i)}{Var(X_{1i})}$$

$$\neq \beta_1$$

$\hat{\beta}_1$ is therefore inconsistent

Sign of the (asymptotic) bias depends on sign of the covariance between regressor and the ~~[illegible]~~ error term

Example 2

Model is:

(1) $$Output_i = \beta_0 + \beta_1 \cdot manhours_i + u_i$$

where i denotes firm i

But true model is:

(2) $$Output_i = \beta_0 + \beta_1 manhours_i + \beta_2 capital_i + \tilde{u}_i$$

where $\tilde{u}_i$ is (correctly specified) error term, i.e. $E(\tilde{u}_i \mid manhours, capital) = 0$

Bias of $\hat{\beta}_1$ in equation (1) is

$$\frac{Cov(u_i, manhours_i)}{Var(manhours_i)} = \frac{Cov(\beta_2 capital_i + \tilde{u}_i, manhours_i)}{Var(manhours_i)}$$

$$= \beta_2 \frac{Cov(capital_i, manhours_i)}{Var(manhours_i)} + \underbrace{\frac{Cov(\tilde{u}_i, manhours_i)}{Var(manhours_i)}}_{=0}$$

$$= \underbrace{\beta_2}_{>0} \frac{Cov(capital_i, manhours)}{\underbrace{Var(manhours_i)}_{>0}}$$

Example 3 : Model says that :

$$grade_i = \beta_0 + \beta_1 \underbrace{\frac{\#\,teachers}{\#\,students}}_{\equiv (T/S)_i} + u_i$$

$$\equiv \beta_0 + \beta_1 (T/S)_i + u_i$$

True model is:

$$grad_i = \beta_0 + \beta_1 (T/S)_i + \beta_2 \, Minority + \tilde{u}_i$$

where $E(\tilde{u}_i \mid (T/S), Minority) = 0$

and $Minority = \begin{cases} 1 & \text{if has minority background} \\ 0 & \text{else} \end{cases}$

$$Bias = \frac{Cov((T/S)_i, \beta_2 Minority_i + \tilde{u}_i)}{Var((T/S_i))}$$

$$= \frac{\beta_2 \, Cov((T/S)_i, Minority_i)}{Var((T/S_i))}$$

In general : Important to include variables that are determinants of Y_i and correlated with the regressor of interest.

Multiple regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i$$

where

$$E(u_i \mid X_{1i}, \dots, X_{ki}) = 0 \Rightarrow$$

$$E(Y_i \mid X_{1i}, \dots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki}$$

— i.e. the regression line

Partial or marginal effects:

$$\beta_k = \frac{\partial E(Y_i \mid X_{1i}, \dots, X_{ki})}{\partial X_{ki}},$$

i.e. holding all other variables constant

One cannot in general obtain $\hat{\beta}_k$ from single-regressor equation:

$$Y_i = \beta_0 + \beta_k X_{ki} + u_i$$

— because of omitted variable bias.

Ordinary least squares (OLS)

The OLS estimator of $\beta_0, \beta_1, \ldots, \beta_k$ is given by

$$\min_{\beta_0, \ldots, \beta_k} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_{1i} - \cdots - \beta_k X_{ki} \right)^2$$

Predicted values:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}$$

Residual:

$$\hat{u}_i = Y_i - \hat{Y}_i \iff Y_i = \hat{Y}_i + \hat{u}_i$$

The 1. order conditions give:

$$\frac{\partial}{\partial \beta_\ell} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \cdots - \hat{\beta}_k X_{ki} \right)^2$$

$$= \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \cdots - \hat{\beta}_k X_{ki} \right) X_{\ell i} = 0$$

$\iff \sum_{i=1}^{n} \hat{u}_i X_{\ell i} = 0$ – the residuals are uncorrelated with all the regressors!

The least squaes assumptions:

A.1 $$E(u_i \mid X_{1i}, \ldots, X_{ki}) = 0$$

This implies that

$$\mathrm{Cov}(u_i, X_{\ell i}) = E(u_i X_{\ell i})$$

$$= E\big(E(u_i X_{\ell i} \mid X_{1i}, \ldots, X_{ki})\big)$$ [double expectation]

$$= E\big(X_{\ell i}\, E(u_i \mid X_{1i}, \ldots, X_{ki})\big)$$

$$= E(X_{\ell i} \cdot 0) = 0$$

Thus the OLS estimator is a moment estimator:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i X_{\ell i} = E(u_i X_{\ell i}) = 0$$

which follows fra 1. order conditions for OLS and A.1.

A.2 :

$(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i)$ $i = 1, \ldots, n$

are i.i.d

A.3 : $E(X_{1i}^4) < \infty, \ldots, E(X_{ki}^4) < \infty,$

$E(Y_i^4) < \infty$

— i.e. large outliers are unlikely

A.4. No perfect multicollinearity: no regressor should be a perfect linear function of other regressors.

More about multicollinearity:

Write $Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + u_i$

where $X_{0i} \equiv 1$ for all i

Multicollinearity means that one of the regressors can be written as a linear function of the other ones:

E.g. $X_{ki} = \sum_{j=0}^{k-1} a_j X_{ji} = a_0 X_{0i} + \dots + a_{k-1} X_{k-1,i}$

for all i and some constants $a_0, \dots, a_{k-1}$

Example:

$$y_i = \beta_0 + \beta_1 Man_i + \beta_2 Woman_i + u_i$$

$Man_i = \begin{cases} 1 & \text{if } i \text{ is man} \\ 0 & \text{else} \end{cases}$, $Woman_i = \begin{cases} 1 & \text{if woman} \\ 0 & \text{else} \end{cases}$

$\Rightarrow Man_i + Woman_i = 1 \iff Man_i = 1 - Woman_i$

$$\Rightarrow Y_i = \beta_0 + \beta_1 (1 - Woman_i) + \beta_2 Woman_i + u_i$$
$$= \beta_0 + \beta_1 + (\beta_2 - \beta_1) Woman_i + u_i$$