Review of concepts & theorems

Part I Omitted variable bias

Let the true model be $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$

If we perform a regression omitting the variable x_2, we estimate a simple linear regression model and we know that the estimated coefficient β_1 is:

$$\widetilde{\beta}_1 = \frac{\widehat{Cov}(y, x_1)}{\widehat{Var}(x_1)}$$

In this formula, we replace the true y (from the complete model) and we obtain the following:

$$\begin{aligned}
\widetilde{\beta}_1 &= \frac{\widehat{Cov}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, x_1)}{\widehat{Var}(x_1)} \\
&= \frac{\widehat{Cov}(\beta_0, x_1)}{\widehat{Var}(x_1)} + \beta_1 \frac{\widehat{Cov}(x_1, x_1)}{\widehat{Var}(x_1)} + \beta_2 \frac{\widehat{Cov}(x_2, x_1)}{\widehat{Var}(x_1)} + \frac{\widehat{Cov}(u, x_1)}{\widehat{Var}(x_1)} \\
&= 0 + \beta_1 \frac{\widehat{Var}(x_1)}{\widehat{Var}(x_1)} + \beta_2 \frac{\widehat{Cov}(x_2, x_1)}{\widehat{Var}(x_1)} + 0
\end{aligned}$$

$$\widetilde{\beta}_1 = \beta_1 + \beta_2 \frac{\widehat{Cov}(x_2, x_1)}{\widehat{Var}(x_1)}$$

Then, the sign of the bias depends on the sign of β_2 and the sign of $Cov_e(x_2, x_1)$ ($Var_e(x_1) \geq 0$)

If the regression equation we estimate excludes one variable which has a causal effect on the variable y, there is a risk that this omission will bias the coefficients for the explanatory variables that are included. This bias exists with certainty when the omitted variable is correlated with one or more of the included regressors. In this case, the variation in the omitted variable which affects y will be correlated with the variation in the included variable. Thus, part of the effect of the omitted variable on y will be attributed to the effect of the included variable on y. (see TD 4 for omitted variable bias derivations)

The incomplete model (when we have not included the effect of experience) under-estimates the effect of education on income: the omission causes downward bias, which suggests a negative relationship between education and experience.

Part II

Frisch-Waugh (partial regressions)

$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\theta_2 + \varepsilon$: we partition $X = [X_1 \quad X_2]$ i.e. we separate the explanatory variables in two groups.

In our example, $X_1 = [educ]$ and $X_2 = \begin{bmatrix} 1 & Exp & Exp^2 \end{bmatrix}$

$\theta_2 = [\beta_0 \quad \beta_2 \quad \beta_3]'$

Theorem: **(a)** The vector $\widehat{\beta}_1$ (OLS estimator of β_1) is equal to the coefficients obtained if Y is regressed on the residuals obtained from regressing each column of X_1 on X_2 $(M_{X_2}X_1)$ (Proof in TD)

Theorem: **(b)** The vector $\widehat{\beta}_1$ (OLS estimator of β_1) is obtained as the coefficients from a regression of the residuals from regressing Y on X_2 $(M_{X_2}Y)$ on the residuals from regressing every column of X_1 on X_2 $(M_{X_2}X_1)$ (Proof in TD).

Theorem: **(c)** The vector $\widehat{\beta}_2$ (OLS estimator of β_2) can be found by regressing the unexplained part $Y - X_1\widehat{\beta}_1$ on X_2 (demonstration in TD).

Part III Outliers & Influential Values

1 concepts

We call outliers all values which are different in a significant way from the overall tendency or trend of the other observations coming from the same data set.

An observation which is influential is an observation whose presence "strongly" affects the result of the model, i.e. small changes in this observation can cause major modifications in the estimated coefficients.

Remark: An observation which is an outlier is not necessarily influential and vica versa.

Remark 2: We observe outliers when certain observations in our sample are influenced by exceptional events, such as wars, climatic anomalies, strikes, etc.

---

2 How to find?

An observation i is called an outlier at a level α if the residuals t_i^* does not belong to the interval:

$$\left[\quad -t\left[(n-1)-(k+1)\right]_{\alpha/2} \quad ; \quad t\left[(n-1)-(k+1)\right]_{\alpha/2} \quad \right]$$

In our example, n=4503 et p=6 and we approximate the Student by N(0,1) and we check if t_i^* belongs to the interval.

$$\left[\quad -N(0,1)_{\alpha/2} \quad ; \quad N(0,1)_{\alpha/2} \right] = [-1,96; 1,96]$$

By definition $\widehat{Y} = PY$ with $P = X\left(X'X\right)^{-1}X'$

we have $\widehat{y}_i = p_{ii}y_i + \sum_{j\neq i} p_{ij}y_j$.

When p_{ii} is large, this indicates that the observation i is important, i.e. large p_{ii} means i is influential.

$\Rightarrow$ We call an observation influential if p_{ii} is "large" enough: generally, the criterion used to judge the influence of an observation i consists in comparing a parameter estimated with and without the presence of observation i.

We will examine two criteria:

- Leverage:
  If $p_{ii} > 2\dfrac{(k+1)}{n} = 2 * (\text{mean of } p_{jj}) \Rightarrow i$ influential.
  (we know that $rg(P) = Tr(P) = k+1 = \sum p_{jj}$)

- Distance of Cook:
  $c_{ii} = \dfrac{p_{ii} t_i^2}{(k+1)(1-p_{ii})}$ with t_i studentized residuals
  If $c_{ii} > F(k+1, n-(k+1))_\alpha \Rightarrow i$ is influential

---

To be able to use the estimated residuals in analysis, we have to standardize them:

- Standardised residuals: $\dfrac{\widehat{\varepsilon}_i}{\widehat{\sigma}}$
  Problem: $\widehat{\sigma}$ is not the standard deviation of $\widehat{\varepsilon}_i$
- Studentized residuals: $t_i = \dfrac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{(1-p_{ii})}}$
  Problem: $\widehat{\sigma}$ depends on all $\widehat{\varepsilon}$ and in particular on $\widehat{\varepsilon}_i$
- Cross-validated residuals:
  $t_i^* = \dfrac{\widehat{\varepsilon}_i}{\widehat{\sigma}_{(i)}\sqrt{(1-p_{ii})}}$
  with $\widehat{\sigma}_{(i)}$ as the estimator of σ in the regression model obtained by excluding the observation $i \Rightarrow \widehat{\sigma}_{(i)}$ is independent from $\widehat{\varepsilon}_i$.
  (an unbiased estimator of the variance of $\widehat{\varepsilon}_i$ does not include the i-th observation.)

  Proposition: If $\varepsilon_i \sim N(0, \sigma^2)$ then $t_i^* \sim t\,[(n-1)-(k+1)]$
  with k = number of explanatory variables (without the constant)

  In practice, if n is large, then $t_i \simeq t_i^* \sim N(0,1)$