Misspecification

We will consider 2 types of misspecification (1) omission of relevant variables, and (2) inclusion of irrelevant variables.

1. Omission of relevant variables

1.1 Properties of estimators

Consider the simple 2 variable case, where we estimate the models:

$$y_i = \beta_0 + \beta_1 x_i + v_i \qquad \text{(F)}$$

when the true model is of the form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i \qquad \text{(T)}$$

and the error process ε_i satisfies all of the CLRM assumptions. Now in estimating (F) (which clearly has z_i omitted) by OLS yields

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_i (x_i - \bar{x})(v_i - \bar{v})}{\sum_i (x_i - \bar{x})^2}, \qquad (2)$$

Now from (F) $v_i - \bar{v} = \beta_2 (z_i - \bar{z}) + \varepsilon_i$, and therefore from equation (2) we get:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_i (x_i - \bar{x})[\beta_2 (z_i - \bar{z}) + \varepsilon_i]}{\sum_i (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \beta_1 + \beta_2 \frac{\sum_i (x_i - \bar{x})(z_i - \bar{z})}{\sum_i (x_i - \bar{x})^2} + \frac{\sum_i (x_i - \bar{x})\varepsilon_i}{\sum_i (x_i - \bar{x})^2}$$

as the final term must be zero as ε_i is random and unrelated to everything, we get:

$$\hat{\beta}_1 = \beta_1 + \beta_2 \frac{\text{cov}(x_i, z_i)}{\text{var}(x_i)}.$$

and hence

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{\text{cov}(x_i, z_i)}{\text{var}(x_i)} \neq \beta_1 \qquad (3)$$

OLS is therefore biased, unless

(i) $\beta_2 = 0$ in which case z_i is not omitted from the (F) equation.

(ii) $\text{cov}(x_i, z_i) = 0$, that is, the omitted variable z_i is unrelated to the variable x_i.

In general therefore, OLS is BIASED, with Bias $= \beta_2 \frac{\text{cov}(x_i, z_i)}{\text{var}(x_i)}$ (from (3)). If the coefficients are biased then the standard errors (and hence t-ratios) are wrong and all hypothesis testing is wrong. This idea can be generalised to many explanatory variables. In general omitting relevant variables will yield biased coefficients, unless the omitted variable(s) are unrelated to ALL of the included explanatory variables.

2. Inclusion of irrelevant variables

2.1 Properties of estimators

The true model is written as

$$y_i = \beta_0 + \beta_1 x_i + u_i \qquad \text{(T)}$$

while the estimated model is of the form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + v_i \qquad \text{(F)}$$

where $\beta_2 = 0$. In this case $E(\hat{\beta}_1) = \beta_1$, that is, the OLS estimate of the estimated model is unbiased; however, the standard errors are generally too large compared with those obtained from the true model (T), this implies that the t-ratios will be too small, implying that some coefficients might be found to be insignificant, which are truly significant. This can be seen using the idea of partitioned regression because in (T) $V(b_1) = \frac{\sigma^2}{\sum_{i=1}(x_{1i} - \bar{x}_1)^2}$, whereas in (F) $V(b_1) = \frac{\sigma^2}{\sum_{i=1}(\tilde{x}_{1i} - \bar{\tilde{x}}_1)^2}$ and $\sum_{i=1}(x_{1i} - \bar{x}_1)^2$ is the total amount of variation in x_1 and $\sum_{i=1}(\tilde{x}_{1i} - \bar{\tilde{x}}_1)^2$ is the variation in x_1 over and above z, in particular $(1 - R^2_{x_1})\sum_{i=1}(x_{1i} - \bar{x}_1)^2 = \sum_{i=1}(\tilde{x}_{1i} - \bar{\tilde{x}}_1)^2$, where $R^2_{x_1}$ is the usual R-squared from a regression of x_1 on z and so $\sum_{i=1}(x_{1i} - \bar{x}_1)^2 \geq \sum_{i=1}(\tilde{x}_{1i} - \bar{\tilde{x}}_1)^2$. If $R^2_{x_1} = 0$ then $\sum_{i=1}(x_{1i} - \bar{x}_1)^2 = \sum_{i=1}(\tilde{x}_{1i} - \bar{\tilde{x}}_1)^2$ and there is no cost to estimating the (F) model.

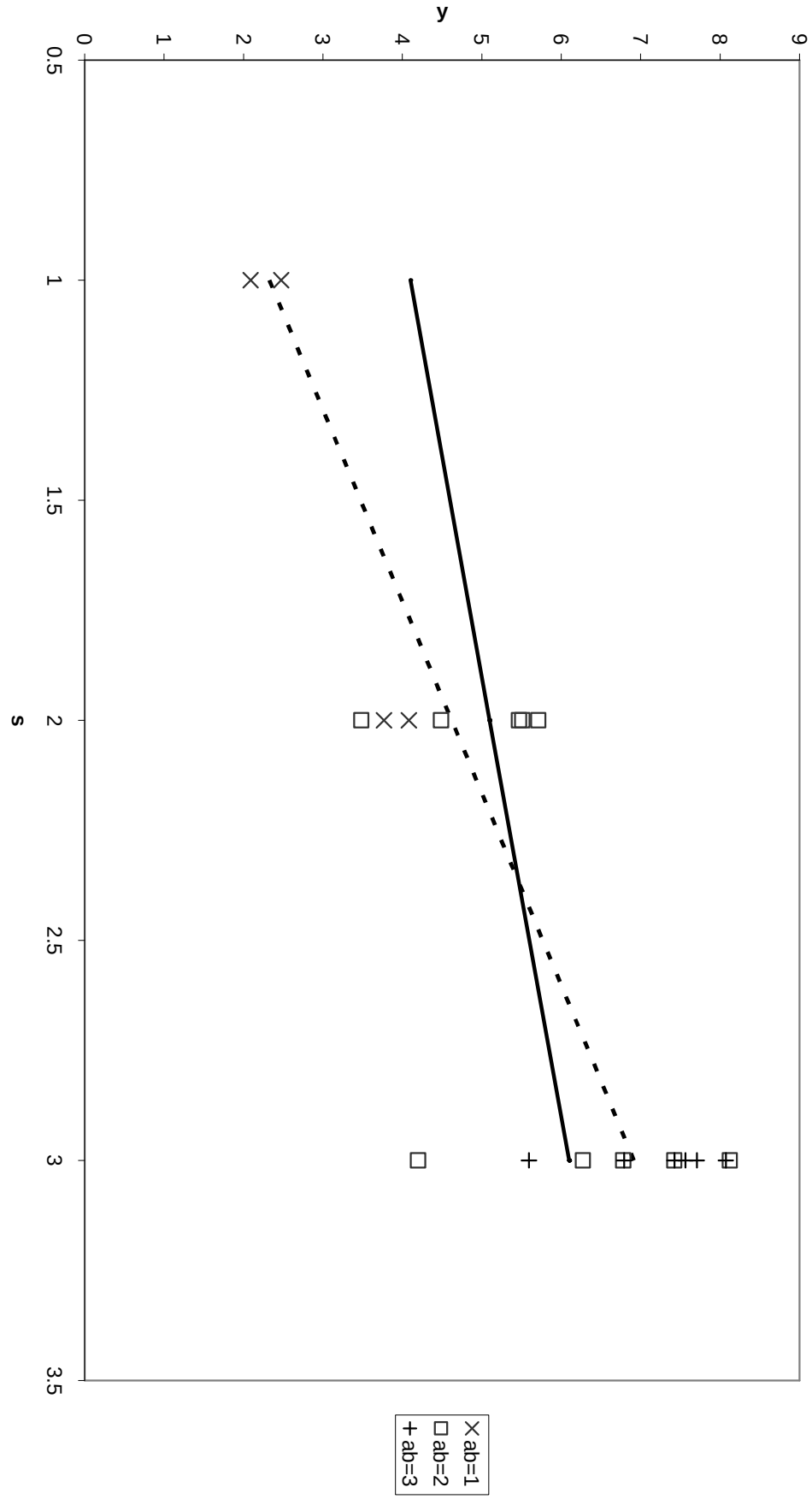

Figure 1: Scatter plot of earnings versus schooling