CHAPTER 6: SPECIFICATION – VARIABLES

Recall, we had the following **six assumptions** required for the Gauss-Markov Theorem:

1. The regression model is linear, correctly specified, and has an additive error term.
2. The error term has a zero population mean.
3. All explanatory variables are uncorrelated with the error term.
4. Observations of the error term are uncorrelated with each other (no serial correlation).
5. The error term has a constant variance (no heteroskedasticity).
6. No explanatory variable is a perfect linear function of any other explanatory variables (no perfect multicollinearity).

We have assumed these assumptions have been satisfied. For the rest of the course, we will deal with violations of these assumptions.

This chapter deals with **Assumption 1**.

Specifying an equation consists of three parts:

1. Choosing the correct independent variables
2. Choosing the correct functional form
3. Choosing the correct form of the random error term

A **specification error** occurs when the model is misspecified in terms of the choice of variables, functional form or error structure.

In choosing explanatory variables, two types of errors are likely:

- Omitting a variable that belongs in the model
- Including an irrelevant variable

We examine the theoretical consequences of each of these specification errors.

Omitted Variables

What happens when an important variable that belongs in the model is omitted?

Suppose the *true* model is: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

But we instead *estimate* the model: $Y_i = \beta_0^* + \beta_1^* X_{1i} + \varepsilon_i^*$, where $\varepsilon_i^* = \varepsilon_i + \beta_2 X_{2i}$

Consequence:

The estimated values of all the other regression coefficients included in the model will be biased **unless**:

- the true regression coefficient excluded from the model is zero or
- the excluded variable is uncorrelated with every included variable

Why?

Example

Let’s think about trying to explain yearly earnings ($). We believe education and ability should be in the regression equation.

Problem: We do not observe ability and therefore do not have data for it.

Correcting for an Omitted Variable

Omitted variable bias is hard to detect:

- invest time in thinking about the equation before you even look at the data
- estimated coefficient has the wrong sign (and significant) or magnitude

Corrections:

- Include the variable
- Report the expected bias

Irrelevant Variables

What happens when a variable that does not belong in the model is included in the equation?

Suppose the *true* model is: $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$

But we instead *estimate* the model: $Y_i = \beta_0^* + \beta_1^* X_{1i} + \beta_2^* X_{2i} + \varepsilon_i^{**}, \quad \varepsilon_i^{**} = \varepsilon_i - \beta_2^* X_{2i}$

Consequence:

The estimated values of all the other regression coefficients included in the model will still be unbiased, their variance however will be higher so we can expect lower $\bar{R}^2$ and larger standard errors for our estimated coefficients.

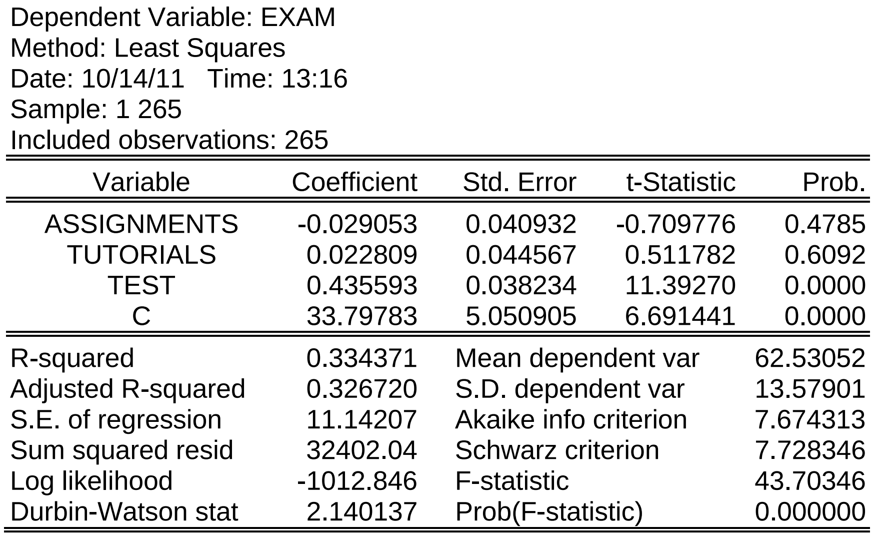
This will happen **unless**:

- the irrelevant variable is uncorrelated with every included variable

Example

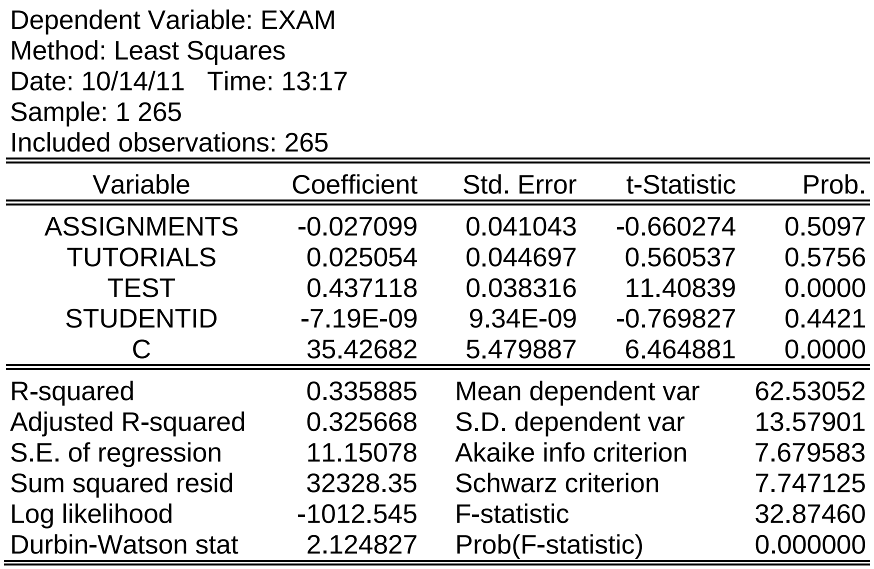
Let’s go back to our example of trying to explain BUEC 333 final exam grades (out of 100) using term test grades (out of 100), assignment grades (out of 100) and tutorial grades (out of 100). Suppose we consider adding **Student ID** as an explanatory variable.

Our original estimation produced:



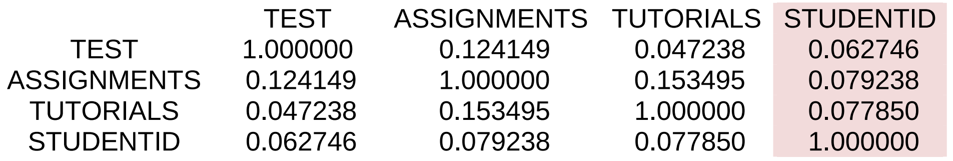
```
Dependent Variable: EXAM
Method: Least Squares
Date: 10/14/11   Time: 13:16
Sample: 1 265
Included observations: 265
======================================================================
        Variable        Coefficient   Std. Error   t-Statistic   Prob.
======================================================================
      ASSIGNMENTS       -0.029053     0.040932    -0.709776    0.4785
       TUTORIALS         0.022809     0.044567     0.511782    0.6092
         TEST            0.435593     0.038234     11.39270    0.0000
           C             33.79783     5.050905     6.691441    0.0000
======================================================================
R-squared                0.334371    Mean dependent var        62.53052
Adjusted R-squared       0.326720    S.D. dependent var        13.57901
S.E. of regression       11.14207    Akaike info criterion     7.674313
Sum squared resid        32402.04    Schwarz criterion         7.728346
Log likelihood          -1012.846    F-statistic               43.70346
Durbin-Watson stat       2.140137    Prob(F-statistic)         0.000000
======================================================================
```

With the inclusion of student ID, we obtain the following regression output:



```
Dependent Variable: EXAM
Method: Least Squares
Date: 10/14/11   Time: 13:17
Sample: 1 265
Included observations: 265
=====================================================================
      Variable        Coefficient    Std. Error    t-Statistic    Prob.
=====================================================================
   ASSIGNMENTS        -0.027099      0.041043     -0.660274     0.5097
     TUTORIALS         0.025054      0.044697      0.560537     0.5756
        TEST           0.437118      0.038316      11.40839     0.0000
     STUDENTID        -7.19E-09      9.34E-09     -0.769827     0.4421
         C             35.42682      5.479887      6.464881     0.0000
=====================================================================
R-squared              0.335885    Mean dependent var       62.53052
Adjusted R-squared     0.325668    S.D. dependent var       13.57901
S.E. of regression     11.15078    Akaike info criterion    7.679583
Sum squared resid      32328.35    Schwarz criterion        7.747125
Log likelihood        -1012.545    F-statistic              32.87460
Durbin-Watson stat     2.124827    Prob(F-statistic)        0.000000
=====================================================================
```

Correlations:



	TEST	ASSIGNMENTS	TUTORIALS	STUDENTID
TEST	1.000000	0.124149	0.047238	0.062746
ASSIGNMENTS	0.124149	1.000000	0.153495	0.079238
TUTORIALS	0.047238	0.153495	1.000000	0.077850
STUDENTID	0.062746	0.079238	0.077850	1.000000

Consequences of including the irrelevant student ID variable into the model:

Choosing which variables belong in the model and which do not is difficult. Let economic theory and careful judgment guide your choice of variables. Thinking about the problem is the hard part and must be done before you estimate the model.

There are some formal specification criteria we will look at.

What you should **not** do:

- Do not test various combination of variables until you find something that you like [Data mining – estimating a lot of specifications before “the” equation has been chosen]
- Sequential specification search – sequentially dropping variables

Formal specification criteria

We consider three such criteria. Specification tests however cannot prove a particular specification is true.

Ramsey's Regression Specification Error Test (RESET)

Most-used test next to $\bar{R}^2$. The RESET test is a general test that measures whether the overall fit of a regression equation can be significantly improved by adding polynomials in $\hat{Y}$. If so, then there is specification error.

RESET test involves three steps:

1. Estimate equation to be tested using OLS and save the fitted (predicted) values:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$$

2. Take these predicted values and create $\hat{Y}_i^2, \hat{Y}_i^3, \hat{Y}_i^4$ terms. Estimate the new regression equation:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki} + \beta_{k+1}\hat{Y}_i^2 + \beta_{k+2}\hat{Y}_i^3 + \beta_{k+3}\hat{Y}_i^4 + \varepsilon_i$$

3. Compare the fits of the two equations using the *F*-test:

$$F = \frac{(RSS_M - RSS)/M}{RSS/(n-k-1)} \sim F_{M,n-k-1}$$

If the two equations are significantly different in fit then *F* will be large enough to reject a null hypothesis of H_0: $\beta_{k+1} = \beta_{k+2} = \beta_{k+3} = 0$.

If you do reject H_0, the test does not tell you what the specification error is!

Akaike's Information Criterion and the Schwarz Criterion

These criteria allow you to compare alternative specifications by adjusting the residual sum of squares (RSS) for the sample size and the number of explanatory variables:

$$\text{AIC} = \log(RSS/n) + 2(k+1)/n$$

$$\text{SC} = \log(RSS/n) + \log(n)(k+1)/n$$

Estimate two specifications, calculate AIC and SC (EViews automatically outputs these), choose model with lower AIC or SC.

In contrast to the RESET test, AIC and SC require you to have alternative specifications to test between.