8. Model Specification and Data Problems

8.1 Functional Form Misspecification

A functional form misspecification generally means that the model does not account for some important nonlinearities.

Recall that omitting important variable is also model misspecification.

Generally functional form misspecification causes bias in the remaining parameter estimators

Example 8.1: Suppose that the correct specification of the wage equation is

(1)

$$\log(\mathtt{wage}) = \beta_0 + \beta_1\mathtt{educ} + \beta_2\mathtt{exper} + \beta_3(\mathtt{exper})^2 + u.$$

Then the return for an extra year of experience is

$$\frac{\partial \log(\mathtt{wage})}{\partial\, \mathtt{exper}} = \beta_2 + 2\beta_3\mathtt{exper}. \tag{2}$$

If the second order term is dropped from (1), use of the resulting biased estimate of β_2 can be misleading.

RESET test

Ramsey (1969)* proposed a general functional form misspecification test, Regression Specification Error Test (RESET), which has proven to be useful.

Estimate first

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u, \tag{3}$$

get $\widehat{y}$ and test in the augmented model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \delta_1 \widehat{y}^2 + \delta_2 \widehat{y}^3 + e \tag{4}$$

the null hypothesis

$$H_0 : \delta_1 = \delta_2 = 0. \tag{5}$$

The test is the F-test with numerator $df_1 = 2$ and denominator $df_2 = n - k - 3$.

*Ramsey, J.B. (1969). Tests for specification errors in classical linear least-squares analysis, *Journal of the Royal Statistical Society*, Series B, 71, 350–371.

Example 8.2: Consider the the house price data (Exercise 3.1) and estimate

(6) $\text{price} = \beta_0 + \beta_1 \text{lotsize} + \beta_2 \text{sqrft} + \beta_3 \text{bdrms} + u.$

Estimation results are:

```
Dependent Variable: PRICE
Method: Least Squares
Sample: 1 88
Included observations: 88
==========================================================
Variable     Coefficient   Std. Error   t-Statistic   Prob.
----------------------------------------------------------
C              -21.77031     29.47504     -0.738601  0.4622
LOTSIZE         0.002068     0.000642      3.220096  0.0018
SQRFT           0.122778     0.013237      9.275093  0.0000
BDRMS           13.85252     9.010145      1.537436  0.1279
==========================================================

============================================================
R-squared           0.672362  Mean dependent var    293.5460
Adjusted R-squared  0.660661  S.D. dependent var    102.7134
S.E. of regression  59.83348  Akaike info criterion 11.06540
Sum squared resid   300723.8  Schwarz criterion     11.17800
Log likelihood     -482.8775  F-statistic           57.46023
Durbin-Watson stat  2.109796  Prob(F-statistic)     0.000000
============================================================
```

Estimate next (6) augmented with $(\widehat{\text{price}})^2$ and $(\widehat{\text{price}})^3$ as in (4). The F-statistic for the null hypothesis (5) becomes $F = 4.67$ with 2 and 82 degrees of freedom. The p-value is 0.012, such that we reject the null hypothesis at the 5% level. Thus, there is some evidence of non-linearity.

Estimate next

$$\log(\text{price}) = \beta_0 + \beta_1 \log(\text{lotsize}) + \beta_2 \log(\text{sqrft}) + \beta_3 \text{bdrms} + u. \tag{7}$$

Estimation results:

```
Dependent Variable: LOG(PRICE)
Method: Least Squares
Date: 10/19/06   Time: 00:01
Sample: 1 88
Included observations: 88
============================================================
Variable          Coefficient   Std. Error  t-Statistic   Prob.
============================================================
C                   -1.297042     0.651284    -1.991517  0.0497
LOG(LOTSIZE)         0.167967     0.038281     4.387714  0.0000
LOG(SQRFT)           0.700232     0.092865     7.540306  0.0000
BDRMS                0.036958     0.027531     1.342415  0.1831
============================================================

==============================================================
R-squared             0.642965  Mean dependent var      5.633180
Adjusted R-squared    0.630214  S.D. dependent var      0.303573
S.E. of regression    0.184603  Akaike info criterion  -0.496833
Sum squared resid     2.862563  Schwarz criterion      -0.384227
Log likelihood        25.86066  F-statistic             50.42374
Durbin-Watson stat    2.088996  Prob(F-statistic)       0.000000
==============================================================
```

The F-statistic for the the null hypothesis (5) is now $F = 2.56$ with p-value 0.084. Thus (5) is not rejected at the 5% level. Thus overall, on the basis of the RESET test the log-log model (7) is preferred.

Non-nested alternatives

For example if the model choices are

(8) $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

and

(9) $$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u.$$

Because the models are non-nested the usual F-test does not apply.

A common approach is to estimate a combined model

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 \log(x_1) + \gamma_4 \log(x_2) + u.$$

(10)

$H_0 : \gamma_3 = \gamma_4 = 0$ is a hypothesis for (8) and $H_0 : \gamma_1 = \gamma_2 = 0$ is a hypothesis for (9). The usual F-test applies again here.

Davidson and MacKinnon (1981)* procedure:

For example to test (8), estimate first

(11) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_1 \widehat{\widehat{y}} + v$,

where $\widehat{\widehat{y}}$ is the fitted value of (9). A significant t value of the θ_1-estimate is a rejection of (8).

Similarly, if $\widehat{y}$ denotes the fitted values of (8), the test of (9) is the t-staistic of the θ_1-estimate from

(12)

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + \theta_1 \widehat{y} + v,$$

*Davidson, R. and J.G. MacKinnon (1981). Several tests for model specification in the presence of alternative hypotheses, *Econometrica* 49, 781–793.

Remark 8.1: A clear winner need not emerge. Both models may be rejected or neither may be rejected. In the latter case adjusted R-square can be used to select the better fitting one. If both models are rejected, more work is needed. *

*For more complicated cases, see Wooldridge, J.M. (1994). A simple specification test for the predictive ability of transformation models, *Review of Economics and Statistics* **76**, 59–65.

8.2 Outliers

Particularly in small data sets OLS estimates are influenced by one or several observations. Generally such observations are called *outliers* or *influential observations*.

Loosely, an observation is an outlier if dropping it changes estimation results materially.

In detection of outliers a usual practice is to investigate standardized (or "studentized") residuals.

If an outlier is an obvious mistake in recording the data, it can be corrected. Usual practice also is to eliminate such observations.

Data transformations, like taking logarithms often narrow the range of data and hence may alleviate outlier problems, too.