Lecture 11. Getting the Model Specification Right

In which you learn what the consequences for bias and efficiency of OLS estimates are if you use the “wrong” set of explanatory variables and are introduced to some specification tests to help guide your model selection
How do we know that the estimated model is the “right” one? (coefficients unbiased and efficient)

What are the consequences of using the wrong model?

Involves investigating

1) Correct functional form
(use logs or levels, squared terms, inverses etc)

2) Whether Gauss-Markov assumptions needed for OLS hold

3) Choice of variables to include in the model
on this consider

a) Omission of (Relevant) Variables

Suppose the true model is given by

\[ y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \]  \hspace{1cm} (1)

But instead you estimate

\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{v} \]  \hspace{1cm} (2)

(ie imposing the restriction that \( \beta_2 = 0 \), \( X_2 \) has no explanatory power, when in fact not true)

We now know that OLS on (2) gives

\[ \hat{\beta}_1 = \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)} \]

when the estimate should be (if used the true model)

\[ \hat{\beta}_1 = \frac{\text{Cov}(X_1, Y)\text{Var}(X_1) - \text{Cov}(X_2, Y)\text{Cov}(X_1, X_2)}{\text{Var}(X_1)\text{Var}(X_2) - [\text{Cov}(X_1, X_2)]^2} \]

So unless \( \text{Cov}(X_1, X_2) = 0 \)

(and if it is then \( X_1 \) and \( X_2 \) are said to be orthogonal ie included & omitted variables uncorrelated)

then the estimates you get from the two models will be different

Does this matter if OLS is supposed to give unbiased efficient estimates?

Estimate is unbiased only if \( E(\hat{\beta}_1) = \beta_1 \)

and since the estimate for the estimated model (2) can always be written as
\[
\hat{\beta}_1^{\text{2var}} = \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)}
\]

Sub. in for TRUE \( y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \)

\[
\hat{\beta}_1^{\text{2var}} = \frac{\text{Cov}(X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u)}{\text{Var}(X_1)}
\]

Using rules on covariances becomes

\[
\hat{\beta}_1^{\text{2var}} = \frac{1}{\text{Var}(X_1)} \left[ \text{Cov}(X_1, \beta_0) + \text{Cov}(X_1, \beta_1 X_1) + \text{Cov}(X_1, \beta_2 X_2) + \text{Cov}(X_1, u) \right]
\]

\[
\hat{\beta}_1^{\text{2var}} = 0 + \beta_1 \frac{\text{Cov}(X_1, X_1)}{\text{Var}(X_1)} + \beta_2 \frac{\text{Cov}(X_1, X_2)}{\text{Var}(X_1)} + \frac{\text{Cov}(X_1, u)}{\text{Var}(X_1)}
\]

and taking expectations (to get bias)

\[
E(\hat{\beta}_1^{\text{2var}}) = \beta_1 + \beta_2 \frac{\text{Cov}(X_1, X_2)}{\text{Var}(X_1)} \neq \beta_1
\]

So OLS is biased when omit relevant variables and the sign of bias depends on

a) the covariance between the variables, \( \text{Cov}(X_1, X_2) \)
b) the sign of the effect \( \beta_2 \) of the extra variable, \( X_2 \), on \( y \)
   (if \( \beta_2 = 0 \) shouldn’t be in model in 1st place)

If \( \beta_2 > 0 \) (and \( \text{Var}(X_2) > 0 \) which it will be since variances are always positive)

then

\[
\hat{\beta}_1^{\text{2var}} > \beta_1 \quad \text{if} \quad \text{Cov}(X_1, X_2) > 0
\]

\[
\hat{\beta}_1^{\text{2var}} < \beta_1 \quad \text{if} \quad \text{Cov}(X_1, X_2) < 0
\]

Hence can tell sign of bias by looking at change in value of estimate in simple and multiple model

Equally important: Not only are estimates biased in presence of omitted variables, but can show that standard errors (hence t, F values etc) are also biased

So it would seem that it is important to include as many variables on the right hand side in order to avoid specification error bias

But…
Example:
Using the data set \textit{ps4data.dta}, a simple regression of the log hourly wage on age (for a particular region and for women only) gives

\begin{verbatim}
.reg lhw age if female==1 & reg==9
\end{verbatim}

\begin{verbatim}
Source |       SS       df       MS              Number of obs = 169
-------------+------------------------------           F(  1,   167) =  2.61
Model |   .76707822     1   .76707822           Prob > F      =  0.1081
Residual |  49.0791051   167  .293886857           R-squared     =  0.0154
-------------+------------------------------           Adj R-squared =  0.0095
Total |  49.8461834   168  .296703472           Root MSE      =  .54211
\end{verbatim}

\begin{verbatim}
|      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lhw |   .0062876   .0038918     1.62   0.108    -.0013959     .013971
       _cons |   1.973012   .1520606    12.98   0.000     1.672803    2.273221
\end{verbatim}

because pay is determined by things other than age there is likely to be omitted variable bias in these estimates (both coefficients and standard errors are likely to be wrong)

If now add a dummy variable for graduate status

\begin{verbatim}
.reg lhw age grad if female==1 & reg==9
\end{verbatim}

\begin{verbatim}
Source |       SS       df       MS              Number of obs = 169
-------------+------------------------------           F(  2,   166) =  7.79
Model |   4.2781892     2   2.1390946           Prob > F      =  0.0006
Residual |  45.5679942   166  .274505989           R-squared     =  0.0858
-------------+------------------------------           Adj R-squared =  0.0748
Total |  49.8461834   168  .296703472           Root MSE      =  .52393
\end{verbatim}

\begin{verbatim}
|      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    age |   .0083863   .0038068     2.20   0.029     .0008703    .0159023
       grad |   .3723513   .1041134     3.58   0.000     .1667942    .5779083
       _cons |   1.823649   .1527802    11.94   0.000     1.522006    2.125292
\end{verbatim}

Can see, size of age estimate increases by 50% and becomes statistically significant

Also since sign on graduate is positive, reasoning above tells us that age and graduate status are \textit{negatively} correlated. Can see this by looking at simple correlation coefficient

\begin{verbatim}
corr age grad if e(sample) (obs=169)
\end{verbatim}

\begin{verbatim}
|      age     grad
-------------+------------------
    age |  1.0000
    grad | -0.1542   1.0000
\end{verbatim}
Inclusion of Irrelevant Variables

Suppose instead that include more variables than are needed

\[
\text{True: } y = \beta_0 + \beta_1X_1 + u \tag{1}
\]
\[
\text{Estimate: } y = \beta_0 + \beta_1X_1 + \beta_2X_2 + v \tag{2}
\]

(Irrelevant means that the variable \(X_2\) has no explanatory power so model (2) is failing to impose the restriction that \(\beta_2=0\),

OLS on (2) gives

\[
\hat{\beta}_1 = \frac{\text{Cov}(X_1, Y)\text{Var}(X_2) - \text{Cov}(X_2, Y)\text{Cov}(X_1, X_2)}{\text{Var}(X_1)\text{Var}(X_2) - [\text{Cov}(X_1, X_2)]^2} \neq \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)}
\]

However in this case can show OLS estimate of \(\beta_1\) will not be biased

Why?

Just give an intuitive proof (see Dougherty for formal proof)

\[
y = \beta_0 + \beta_1X_1 + \beta_2X_2 + v \tag{2}
\]

Since true effect of \(\beta_2\) is zero and we know OLS gives unbiased estimates of true values then would expect, on average, the OLS estimate of \(\beta_2\) in (2) should also be zero

If it does not then it is only the result of chance. Its presence in the model does not affect the bias of the other variables)

but will be inefficient, since in 3 variable model

\[
\text{Var}(\hat{\beta}_1) = \frac{s^2}{N*\text{Var}(X)} * \frac{1}{1-r_{X_1X_2}^2} \neq \frac{s^2}{N*\text{Var}(X)}
\]

so including extra irrelevant variables has a cost in terms of larger standard errors (smaller \(t\), \(F\) values) than otherwise (and type II error)
Example

Using the data set smokes.dta, suppose we are interested in the association between smoking and pay. A regression of the log of hourly wages on a dummy variable to indicate whether or not someone smokes (smokes) and a continuous variable to indicate the number of cigarettes smoked each week (quant) and controls for age and gender gives the following

```
. reg lhw age female smokes quant

Source |       SS       df       MS              Number of obs =   10061
-------------+------------------------------           F(  4, 10056) =  170.33
Model |  235.697543     4  58.9243857           Prob > F      =  0.0000
Residual |  3478.78205 10056  .345940936           R-squared     =  0.0635
-------------+------------------------------           Adj R-squared =  0.0631
Total |  3714.47959 10060  .369232564           Root MSE      =  .58817

------------------------------------------------------------------------------
    lhw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    age |   .0051492   .0004697    10.96   0.000     .0042286    .0060698
  female |  -.2326228    .011761   -19.78   0.000    -.2556766   -.2095689
     smokes |  -.1190897   .0236027    -5.05   0.000    -.1653558   -.0728236
       quant |  -.000528    .000208    -2.54   0.011    -.0009356   -.0001203
     _cons |   6.828719    .021466   318.12   0.000     6.786641    6.870796
------------------------------------------------------------------------------
```

So it seems that smoking and the amount smoked are associated with lower pay (by how much?)

Now add an indicator for how many cigarettes are smoked at the weekend – Hard to believe that should have true effect on wages but the effect is

```
. reg lhw age female smokes quant quantw

Source |       SS       df       MS              Number of obs =   10061
-------------+------------------------------           F(  5, 10055) =  136.29
Model |  235.758451     5  47.1516902           Prob > F      =  0.0000
Residual |  3478.72114 10055  .345969283           R-squared     =  0.0635
-------------+------------------------------           Adj R-squared =  0.0630
Total |  3714.47959 10060  .369232564           Root MSE      =  .58819

------------------------------------------------------------------------------
    lhw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    age |   .0051537   .0004698    10.97   0.000     .0042328    .0060746
  female |  -.2325969   .0117616   -19.78   0.000    -.2556521   -.2095689
     smokes |  -.1214821   .0242826    -5.00   0.000    -.169081   -.0738833
       quant |  -.0006898   .0004382    -1.57   0.115     -.0015489   .0001692
     quantw |   .0011567   .0027568     0.42   0.675    -.0042472    .0065606
     _cons |    6.82852   .0214721   318.02   0.000     6.786431    6.870609
------------------------------------------------------------------------------
```

```
corr quant quantw
(obs=30049)
   quant | quant  quantw
-------------+------------------
  quant |  1.0000
 quantw |  0.9674   1.0000
```

Can see (compared with original specification) that standard error on quant variable is much larger (t value now insignificant)
Moral: Irrelevant variables – particularly if they are strongly correlated with other rhs variables can inflate standard errors, (introduce unnecessary multicollinearity)

N.B. It may seem confusing that omitting relevant variables can cause bias when we use OLS and including irrelevant variables will not. Isn’t OLS always supposed to give unbiased estimates?

Remember OLS will only give unbiased estimates if all 4 of the Gauss-Markov assumptions hold (see earlier notes).

The case of omitted variables violates the assumption Cov(X,u) = 0 since the omitted variable(s) X2 now form part of the residual and if they are correlated with the included variables X1 then Cov(X1, X2) = Cov(X,u) ≠ 0

In the case of inclusion of irrelevant variables the X2 are not part of the residual and if all the other Gauss-Markov assumptions hold the OLS will be unbiased (though in this case it will be inefficient because the X2 variables should not be in the model).
Testing for Functional Form
To test whether should have included extra variables (strictly higher order terms of the included variables) then do the Ramsey Regression Specification Error Test (RESET)

Given chosen model

1) Estimate: \( y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \)

2) save predicted (fitted) values : \( \hat{y} = \beta_0 + \beta_1 \hat{X}_1 + \beta_2 \hat{X}_2 \)

(predicted value is a weighted average of all the right hand side variables with weights given by size of coefficients)

3) Add higher order powers of this predicted variable to the original equation

\[
y = \beta_0 + \beta_1 \hat{X}_1 + \beta_2 \hat{X}_2 + \hat{y} + \hat{y}^2 + \hat{y}^3 + \ldots + \hat{y}^k
\]

(higher orders of predicted value are weighted averages of higher orders of all the right hand side variables (no. of extra terms is arbitrary – should check robustness of result to variation in no.)

4) F test for inclusion of these extra variables

5) Reject null of no functional form mis-specification if estimated F > F_{critical}

Example:
The data set food_2.dta, contains information on food expenditure and the age in a sample of British adults. You decide to examine the impact of age on food expenditure

```
. reg food age
 Source |       SS       df       MS              Number of obs =     340
-------------+------------------------------           F(  1,   338) =   11.21
Model |  19877.8218     1  19877.8218           Prob > F      =  0.0009
Residual |  599501.744   338   1773.6738           R-squared     =  0.0321
-------------+------------------------------           Adj R-squared =  0.0292
Total |  619379.565   339  1827.07836           Root MSE      =  42.115
-------------+------------------------------
food |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  age |  -.4486191   .1340079    -3.35   0.001    -.7122136   -.1850247
  _cons |   86.93161   7.253069    11.99   0.000     72.66477    101.1984
-------------+----------------------------------------------------------------
```

Simple regression of food expenditure on age suggests a negative association between age and food expenditure

However a scatter plot suggests that the relationship between age and food expenditure is non-linear – so it may be that this simple specification suffers from omitted variable bias.
To test formally for this use the RESET test

```
predict fhat  /* this is how you get predicted values in Stata */
(option xb assumed; fitted values)

. g fhat2=fhat^2   /* now square and cube these fitted values */
. g fhat3=fhat^3

reg food age fhat2
```

```
Source |       SS       df       MS              Number of obs =     340
-----------------+------------------------------           F(  2,   337) =   27.89
Model |  87958.6589     2  43979.3294           Prob > F      =  0.0000
Residual |  531420.906   337  1576.91664           R-squared     =  0.1420
-----------------+------------------------------           Adj R-squared =  0.1369
Total |  619379.565   339  1827.07836           Root MSE      =   39.71

------------------------------------------------------------------------------
food |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age |  -15.20347   2.249123    -6.76   0.000    -19.62756   -10.77938
fhat2 |  -0.259274   .0394594    -6.57   0.000    -.3368919   -.1816562
_cons |   1918.255   278.7967     6.88   0.000     1369.854    2466.655
------------------------------------------------------------------------------
```

and test for joint significance of these extra variables (fhat2 fhat3) using the F test

\[
F = \frac{RSS_{restrict} - RSS_{unrestrict}}{J} \sim F(J, N-K_{unrestrict})
\]

\[
= \frac{599502 - 531421}{531421 / 340 - 3} \sim F(1, 340 - 3)
\]

\[
= 43.2
\]

From F tables, critical value at 5% level \(F(1, 337) = F(1, \infty) = 3.92\)

So estimated \(F > F_{critical}\)

Still **reject** null that model is correctly specified – so need to look for more variables or change functional form.

N.B. Stata does this test automatically but adds a different number of powers of the predicted values to the regression – how many? Test can be sensitive to number of powers used.

```
. ovtest
```

Ramsey RESET test using powers of the fitted values of food

Ho: model has no omitted variables

\[
F(3, 335) = 15.58
\]

\[
Prob > F = 0.0000
\]
So the Ramsey RESET test suggests that there may be omitted higher order powers of age

Adding the square of age

```
reg food age age2
```


<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs =   340</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F( 2, 337) = 27.89</td>
</tr>
<tr>
<td>Model</td>
<td>87958.728</td>
<td>2</td>
<td>43979.364</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>531420.837</td>
<td>337</td>
<td>1576.91643</td>
<td>R-squared = 0.1420</td>
</tr>
<tr>
<td></td>
<td>619379.565</td>
<td>339</td>
<td>1827.07836</td>
<td>Adj R-squared = 0.1369</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = 39.71</td>
</tr>
</tbody>
</table>

```
-------------+------------------------------           F(  2,   337) =   27.89
Model |   87958.728     2   43979.364           Prob > F      =  0.0000
Residual |  531420.837   337  1576.91643           R-squared     =  0.1420
        |  619379.565   339  1827.07836           Adj R-squared =  0.1369
        |          |    |             | Root MSE      =  39.71
-------------+------------------------------           F(  2,   337) =   27.89
```

```
| food | Coef. | Std. Err. | t    | P>|t|   [95% Conf. Interval] |
|------|-------|-----------|------|-------|-------------------------|
| age  | 5.019477 | .8417381 | 5.96 | 0.000 | 3.363754 | 6.6752 |
| age2 | -.0521813 | .0079416 | -6.57 | 0.000 | -.0678025 | -.03656 |
| _cons | -41.10639 | 20.65161 | -1.99 | 0.047 | -81.7287 | -.4840842 |
```

```
ovtest
Ramsey RESET test using powers of the fitted values of food
Ho: model has no omitted variables
F(3, 334) = 1.37
Prob > F = 0.2526
```

which suggests that no more higher order powers are needed (but that age^2 and possibly age^3 are needed in the model). N.B. this test should not really be used to test whether other variables (like gender) should be included in the model.

```
. reg food age age2 female
```

```
<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs =   340</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F( 3, 336) = 22.98</td>
</tr>
<tr>
<td>Model</td>
<td>105466.689</td>
<td>3</td>
<td>35155.5629</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>513912.877</td>
<td>336</td>
<td>1529.50261</td>
<td>R-squared = 0.1629</td>
</tr>
<tr>
<td></td>
<td>619379.565</td>
<td>339</td>
<td>1827.07836</td>
<td>Adj R-squared = 0.1629</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = 39.109</td>
</tr>
</tbody>
</table>

```
-------------+------------------------------           F(  3,   336) =   22.98
Model |  105466.689     3  35155.5629           Prob > F      =  0.0000
Residual |  513912.877   336  1529.50261           R-squared     =  0.1629
        |  619379.565   339  1827.07836           Adj R-squared =  0.1629
        |          |    |             | Root MSE      =  39.109
-------------+------------------------------           F(  3,   336) =   22.98
```

```
| food | Coef. | Std. Err. | t    | P>|t|   [95% Conf. Interval] |
|------|-------|-----------|------|-------|-------------------------|
| age  | 4.664694 | .8355931 | 5.58 | 0.000 | 3.021041 | 6.308347 |
| age2 | -.0487368 | .0078872 | -6.18 | 0.000 | -.0642514 | -.0332222 |
| female | -14.75513 | 4.361143 | -3.38 | 0.001 | -23.33371 | -6.176544 |
| _cons | -27.02573 | 20.76021 | -1.30 | 0.194 | -67.86208 | 13.81062 |
```

```
ovtest
Ramsey RESET test using powers of the fitted values of food
Ho: model has no omitted variables
F(3, 333) = 1.22
Prob > F = 0.3019
```