MODEL MISSPECIFICATION
## Consequences of Variable Misspecification

<table>
<thead>
<tr>
<th>True Model</th>
<th>Fitted Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y = \alpha + \beta_1 x_1 + u$</td>
<td>$\hat{y} = a + b_1 x_1 + b_2 x_2$</td>
</tr>
<tr>
<td>$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u$</td>
<td>$\hat{y} = a + b_1 x_1$</td>
</tr>
</tbody>
</table>

To keep the analysis simple, we will assume that there are only two possibilities. Either $y$ depends only on $x_1$, or it depends on both $x_1$ and $x_2$. 


### Consequences of Variable Misspecification

<table>
<thead>
<tr>
<th></th>
<th>True Model</th>
<th>Fitted Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[ y = \alpha + \beta_1 x_1 + u ]</td>
<td>[ \hat{y} = a + b_1 x_1 ]</td>
</tr>
<tr>
<td></td>
<td>Correct specification, no problems</td>
<td>Correct specification, no problems</td>
</tr>
<tr>
<td></td>
<td>[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u ]</td>
<td>[ \hat{y} = a + b_1 x_1 + b_2 x_2 ]</td>
</tr>
<tr>
<td></td>
<td>Coefficients are biased (in general). Standard errors are invalid.</td>
<td>Coefficients are unbiased but Inefficient</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct specification, no problems</td>
</tr>
</tbody>
</table>
MODEL MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u \]

\[ \hat{y} = a + b_1 x_1 \]

\[ E(b_1) = \beta_1 + \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} \]

In the present case, the omission of \( x_2 \) causes \( b_1 \) to be biased by an amount \( \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} \). We will demonstrate this first intuitively and then mathematically.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u \]

\[ \hat{y} = a + b_1 x_1 \]

\[ E(b_1) = \beta_1 + \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} \]

The intuitive reason is that, in addition to its direct effect \( \beta_1 \), \( x_1 \) has an apparent indirect effect as a consequence of acting as a proxy for the missing \( x_2 \).
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u \]

\[ \hat{y} = a + b_1 x_1 \]

\[ E(b_1) = \beta_1 + \beta_2 \frac{Cov(x_1, x_2)}{Var(x_1)} \]

The strength of the proxy effect depends on two factors: the strength of the effect of \( x_2 \) on \( y \), which is given by \( \beta_2 \), and the ability of \( x_1 \) to mimic \( x_2 \).
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u \]

\[ \hat{y} = a + b_1 x_1 \]

\[ E(b_1) = \beta_1 + \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} \]

The ability of \( x_1 \) to mimic \( x_2 \) is determined by the slope coefficient obtained when \( x_2 \) is regressed on \( x_1 \), which of course is \( \text{Cov}(x_1, x_2)/\text{Var}(x_1) \).
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u \]

\[ \hat{y} = a + b_1 x_1 \]

\[ b_1 = \frac{\text{Cov}(x_1, y)}{\text{Var}(x_1)} = \frac{\text{Cov}(x_1, [\alpha + \beta_1 x_1 + \beta_2 x_2 + u])}{\text{Var}(x_1)} \]

We will now derive the expression for the bias mathematically. Since we are mistakenly fitting a simple regression model, the slope coefficient is \( \text{Cov}(x_1, y)/\text{Var}(x_1) \). The first step is to substitute for \( y \) from the true model.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u \]

\[ \hat{y} = a + b_1 x_1 \]

\[
\begin{align*}
    b_1 &= \frac{\text{Cov}(x_1, y)}{\text{Var}(x_1)} = \frac{\text{Cov}(x_1, [\alpha + \beta_1 x_1 + \beta_2 x_2 + u])}{\text{Var}(x_1)} \\
    &= \frac{\text{Cov}(x_1, \alpha) + \text{Cov}(x_1, \beta_1 x_1) + \text{Cov}(x_1, \beta_2 x_2) + \text{Cov}(x_1, u)}{\text{Var}(x_1)}
\end{align*}
\]

We use Covariance Rule 1 to decompose the numerator.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u \]

\[ \hat{y} = a + b_1 x_1 \]

\[ b_1 = \frac{\text{Cov}(x_1, y)}{\text{Var}(x_1)} = \frac{\text{Cov}(x_1, [\alpha + \beta_1 x_1 + \beta_2 x_2 + u])}{\text{Var}(x_1)} \]

\[ = \frac{\text{Cov}(x_1, \alpha) + \text{Cov}(x_1, \beta_1 x_1) + \text{Cov}(x_1, \beta_2 x_2) + \text{Cov}(x_1, u)}{\text{Var}(x_1)} \]

\[ = \frac{0 + \beta_1 \text{Cov}(x_1, x_1) + \beta_2 \text{Cov}(x_1, x_2) + \text{Cov}(x_1, u)}{\text{Var}(x_1)} \]

The first term is 0 because \( \alpha \) is a constant. The \( \beta \) coefficients can be taken out of the second and third terms.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u \]

\[ \hat{y} = a + b_1 x_1 \]

\[
\begin{align*}
b_1 &= \frac{\text{Cov}(x_1, y)}{\text{Var}(x_1)} = \frac{\text{Cov}(x_1, [\alpha + \beta_1 x_1 + \beta_2 x_2 + u])}{\text{Var}(x_1)} \\
&= \frac{\text{Cov}(x_1, \alpha) + \text{Cov}(x_1, \beta_1 x_1) + \text{Cov}(x_1, \beta_2 x_2) + \text{Cov}(x_1, u)}{\text{Var}(x_1)} \\
&= \frac{0 + \beta_1 \text{Cov}(x_1, x_1) + \beta_2 \text{Cov}(x_1, x_2) + \text{Cov}(x_1, u)}{\text{Var}(x_1)} \\
&= \beta_1 + \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} + \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)}
\end{align*}
\]

Hence we have demonstrated that \( b_1 \) has three components.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u \]

\[ \hat{y} = a + b_1 x_1 \]

\[ b_1 = \frac{\text{Cov}(x_1, y)}{\text{Var}(x_1)} = \beta_1 + \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} + \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)} \]

\[ E(b_1) = E \left\{ \beta_1 + \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} + \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)} \right\} \]

To investigate biasedness or unbiasedness, we take the expected value of \( b_1 \).
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u \]

\[ \hat{y} = a + b_1 x_1 \]

\[ b_1 = \frac{\text{Cov}(x_1, y)}{\text{Var}(x_1)} = \beta_1 + \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} + \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)} \]

\[ E(b_1) = E\left\{ \beta_1 + \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} + \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)} \right\} \]

\[ = E(\beta_1) + E\left\{ \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} \right\} + E\left\{ \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)} \right\} \]

Using Expected Value Rule 1, we can decompose the expected value as the sum of the expected values of the components.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u \]

\[ \hat{y} = a + b_1 x_1 \]

\[ b_1 = \frac{\text{Cov}(x_1, y)}{\text{Var}(x_1)} = \beta_1 + \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} + \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)} \]

\[ E(b_1) = E \left\{ \beta_1 + \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} + \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)} \right\} \]

\[ = E(\beta_1) + E \left\{ \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} \right\} + E \left\{ \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)} \right\} \]

\[ = \beta_1 + \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} \]

The first two components are fixed since \( \beta_1 \) and \( \beta_2 \) are constants and \( x_1 \) and \( x_2 \) are nonstochastic, by assumption. The expected value of the third term is 0. (For a proof, see the second sequence in Chapter 3.)
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u \]

\[ \hat{y} = a + b_1 x_1 \]

\[ b_1 = \frac{\text{Cov}(x_1, y)}{\text{Var}(x_1)} = \beta_1 + \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} + \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)} \]

\[ E(b_1) = E\left\{ \beta_1 + \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} + \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)} \right\} \]

\[ = E(\beta_1) + E\left\{ \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} \right\} + E\left\{ \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)} \right\} \]

\[ = \beta_1 + \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} \]

Thus \( b_1 \) is biased by an amount \( \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} \). As a consequence of the misspecification, the standard errors, \( t \) tests and \( F \) test are invalid.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

```
reg hgc asvabc hgcm
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 570</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1230.2039</td>
<td>2</td>
<td>615.101949</td>
<td>F( 2, 567) = 156.81</td>
</tr>
<tr>
<td>Residual</td>
<td>2224.04347</td>
<td>567</td>
<td>3.92247526</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>3454.24737</td>
<td>569</td>
<td>6.07073351</td>
<td>R-squared = 0.3561</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.3539</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = 1.9805</td>
</tr>
</tbody>
</table>

| hgc    | Coef.       | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|--------|-------------|-----------|-------|------|---------------------|
| asvabc | 0.1381062   | 0.0097494 | 14.166| 0.000| 0.1189567           |
|         |             |           |       |      | 0.1572556           |
| hgcm   | 0.154783    | 0.0350728 | 4.413| 0.000| 0.0858946           |
|         |             |           |       |      | 0.2236715           |
| _cons  | 4.791277    | 0.5102431 | 9.390| 0.000| 3.789088            |
|         |             |           |       |      | 5.793475            |

We will illustrate the bias using an educational attainment model. To keep the analysis simple, we will assume that HGC depends on ASVABC and HGCM. The output above shows the corresponding regression using EAEF Data Set 21.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

```
.reg hgc asvabc hgcm
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 570</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1230.2039</td>
<td>2</td>
<td>615.101949</td>
<td>F( 2, 567) = 156.81</td>
</tr>
<tr>
<td>Residual</td>
<td>2224.04347</td>
<td>567</td>
<td>3.92247526</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>3454.24737</td>
<td>569</td>
<td>6.07073351</td>
<td>R-squared = 0.3561</td>
</tr>
</tbody>
</table>

| hgc | Coef.    | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-----|----------|-----------|-------|------|----------------------|
| asvabc | .1381062 | .0097494  | 14.166 | 0.000 | .1189567 .1572556    |
| hgcm  | .154783  | .0350728  | 4.413  | 0.000 | .0858946 .2236715    |
| _cons | 4.791277 | .5102431  | 9.390  | 0.000 | 3.78908  5.793475    |

\[
HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + u
\]

\[
E(b_1) = \beta_1 + \beta_2 \frac{\text{Cov}(ASVABC, HGCM)}{\text{Var}(ASVABC)}
\]

We will run the regression a second time, omitting HGCM. Before we do this, we will try to predict the direction of the bias in the coefficient of ASVABC.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

```
.reg hgc asvabc hgcm
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 570</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1230.2039</td>
<td>2</td>
<td>615.101949</td>
<td>F(  2,  567) = 156.81</td>
</tr>
<tr>
<td>Residual</td>
<td>2224.04347</td>
<td>567</td>
<td>3.92247526</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>3454.24737</td>
<td>569</td>
<td>6.07073351</td>
<td>Adj R-squared = 0.3539</td>
</tr>
</tbody>
</table>

| hgc      | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|----------|--------|-----------|-------|------|-----------------------|
| asvabc   | 0.1381062 | 0.0097494 | 14.166 | 0.000 | 0.1189567-0.1572556   |
| hgcm     | 0.154783 | 0.0350728 | 4.413  | 0.000 | 0.0858946-0.2236715  |
| _cons    | 4.791277 | 0.5102431 | 9.390  | 0.000 | 3.78908-5.793475     |

\[
HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + u
\]

\[
E(b_1) = \beta_1 + \beta_2 \frac{\text{Cov}(ASVABC, HGCM)}{\text{Var}(ASVABC)}
\]

It is reasonable to suppose, as a matter of common sense, that \( \beta_2 \) is positive. This assumption is strongly supported by the fact that its estimate in the multiple regression is positive and highly significant.
**VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE**

```
. reg hgc asvabc hgcm
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1230.2039</td>
<td>2</td>
<td>615.101949</td>
</tr>
<tr>
<td>Residual</td>
<td>2224.04347</td>
<td>567</td>
<td>3.92247526</td>
</tr>
<tr>
<td>Total</td>
<td>3454.24737</td>
<td>569</td>
<td>6.07073351</td>
</tr>
</tbody>
</table>

|         | Coef.      | Std. Err. | t       | P>|t| | [95% Conf. Interval] |
|---------|------------|-----------|---------|-----|---------------------|
| asvabc  | .1381062   | .0097494  | 14.166  | 0.000 | .1189567 .1572556   |
| hgcm    | .154783    | .0350728  | 4.413   | 0.000 | .0858946 .2236715   |
| _cons   | 4.791277   | .5102431  | 9.390   | 0.000 | 3.78908 5.793475   |

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + u \]

\[ E(b_1) = \beta_1 + \beta_2 \frac{\text{Cov}(ASVABC,HGCM)}{\text{Var}(ASVABC)} \]

The correlation between ASVABC and HGCM is positive, so their covariance must be positive. Var(ASVABC) is automatically positive. Hence the bias should be positive.
**VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE**

```
. reg hgc asvabc
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 570</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1153.80864</td>
<td>1</td>
<td>1153.80864</td>
<td>F( 1, 568) = 284.89</td>
</tr>
<tr>
<td>Residual</td>
<td>2300.43873</td>
<td>568</td>
<td>4.05006818</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>3454.24737</td>
<td>569</td>
<td>6.07073351</td>
<td>R-squared = 0.3340</td>
</tr>
</tbody>
</table>

|             | Coef.    | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|-------------|----------|-----------|-------|------|---------------------|
| hgc         |          |           |       |      |                     |
| asvabc      | 0.1545378| 0.0091559 | 16.879| 0.000| 0.1365543 - 0.1725213|
| _cons       | 5.770845 | 0.4668473 | 12.361| 0.000| 4.853888 - 6.687803 |

Here is the regression omitting *HGCM*. 
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

```
. reg hgc asvabc hgcm

------------------------------------------------------------------------------
hgc |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
asvabc |   .1381062   .0097494     14.166   0.000       .1189567    .1572556
hgcm |    .154783   .0350728      4.413   0.000       .0858946    .2236715
    _cons |   4.791277   .5102431      9.390   0.000        3.78908    5.793475
------------------------------------------------------------------------------
```

```
. reg hgc asvabc

------------------------------------------------------------------------------
hgc | Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
asvabc |   .1545378   .0091559     16.879   0.000       .1365543    .1725213
    _cons |   5.770845   .4668473     12.361   0.000       4.853888    6.687803
------------------------------------------------------------------------------
```

As you can see, the coefficient of \textit{ASVABC} is indeed higher when \textit{HGCM} is omitted. Part of the difference may be due to pure chance, but part is attributable to the bias.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

```
. reg hgc hgcm
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 570</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>443.110436</td>
<td>1</td>
<td>443.110436</td>
<td>F( 1, 568) = 83.59</td>
</tr>
<tr>
<td>Residual</td>
<td>3011.13693</td>
<td>568</td>
<td>5.30129742</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>3454.24737</td>
<td>569</td>
<td>6.07073351</td>
<td>Adj R-squared = 0.1267</td>
</tr>
</tbody>
</table>

```
| hgc   | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-------|--------|-----------|-------|-----|----------------------|
| hgcm  | .3445198 | .0376833  | 9.142 | 0.000 | .2705041 .4185354    |
| _cons | 9.506491 | .4495754  | 21.145 | 0.000 | 8.623458 10.38952   |
```

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + u \]

\[ E(b_2) = \beta_2 + \beta_1 \frac{\text{Cov}(ASVABC,HGCM)}{\text{Var}(HGCM)} \]

Here is the regression omitting ASVABC instead of HGCM. We would expect \( b_2 \) to be upwards biased. We anticipate that \( \beta_1 \) is positive and we know that both the covariance and variance terms in the bias expression are positive.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

```
. reg hgc asvabc hgcm

|        | Coef.    | Std. Err. |    t     | P>|t|    |      [95% Conf. Interval]  |
|--------|----------|-----------|----------|--------|----------------------------|
| asvabc | 0.1381062| 0.0097494 | 14.166   | 0.000  | 0.1189567 - 0.1572556      |
| hgcm   | 0.154783 | 0.0350728 | 4.413    | 0.000  | 0.0858946 - 0.2236715      |
| _cons  | 4.791277 | 0.5102431 | 9.390    | 0.000  | 3.78908 - 5.793475         |
```

```
. reg hgc hgcm

|        | Coef.    | Std. Err. |    t     | P>|t|    |      [95% Conf. Interval]  |
|--------|----------|-----------|----------|--------|----------------------------|
| hgcm   | 0.3445198| 0.0376833 | 9.142    | 0.000  | 0.2705041 - 0.4185354      |
| _cons  | 9.506491 | 0.4495754 | 21.145   | 0.000  | 8.623458 - 10.38952        |
```

In this case the bias is quite dramatic. The coefficient of HGCM has more than doubled. (The reason for the bigger effect is that Var(HGCM) is much smaller than Var(AVSABC), while $\beta_1$ and $\beta_2$ are similar in size, judging by their estimates.)
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

\[ . \text{reg hgc asvabc hgc} \]

\[
\begin{array}{l}
\text{Source} | \quad \text{SS} \quad \text{df} \quad \text{MS} \\
\hline
\text{Model} | \quad 1230.2039 \quad 2 \quad 615.101949 \\
\text{Residual} | \quad 2224.04347 \quad 567 \quad 3.92247526 \\
\hline
\text{Total} | \quad 3454.24737 \quad 569 \quad 6.07073351 \\
\end{array}
\]

\[ . \text{reg hgc asvabc} \]

\[
\begin{array}{l}
\text{Source} | \quad \text{SS} \quad \text{df} \quad \text{MS} \\
\hline
\text{Model} | \quad 1153.80864 \quad 1 \quad 1153.80864 \\
\text{Residual} | \quad 2300.43873 \quad 568 \quad 4.05006818 \\
\hline
\text{Total} | \quad 3454.24737 \quad 569 \quad 6.07073351 \\
\end{array}
\]

\[ . \text{reg hgc hgc} \]

\[
\begin{array}{l}
\text{Source} | \quad \text{SS} \quad \text{df} \quad \text{MS} \\
\hline
\text{Model} | \quad 443.110436 \quad 1 \quad 443.110436 \\
\text{Residual} | \quad 3011.13693 \quad 568 \quad 5.30129742 \\
\hline
\text{Total} | \quad 3454.24737 \quad 569 \quad 6.07073351 \\
\end{array}
\]

Finally, we will investigate how $R^2$ behaves when a variable is omitted. In the simple regression of $HGC$ on $ASVABC$, $R^2$ is 0.33, and in the simple regression of $HGC$ on $HGCM$ it is 0.13.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

\[ . \text{reg hgc asvabc hgm} \]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1230.2039</td>
<td>2</td>
<td>615.101949</td>
</tr>
<tr>
<td>Residual</td>
<td>2224.04347</td>
<td>567</td>
<td>3.92247526</td>
</tr>
<tr>
<td>Total</td>
<td>3454.24737</td>
<td>569</td>
<td>6.07073351</td>
</tr>
</tbody>
</table>

\[ . \text{reg hgc asvabc} \]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1153.80864</td>
<td>1</td>
<td>1153.80864</td>
</tr>
<tr>
<td>Residual</td>
<td>2300.43873</td>
<td>568</td>
<td>4.05006818</td>
</tr>
<tr>
<td>Total</td>
<td>3454.24737</td>
<td>569</td>
<td>6.07073351</td>
</tr>
</tbody>
</table>

\[ . \text{reg hgc hgm} \]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>443.110436</td>
<td>1</td>
<td>443.110436</td>
</tr>
<tr>
<td>Residual</td>
<td>3011.13693</td>
<td>568</td>
<td>5.30129742</td>
</tr>
<tr>
<td>Total</td>
<td>3454.24737</td>
<td>569</td>
<td>6.07073351</td>
</tr>
</tbody>
</table>

Number of obs = 570
F( 2, 567) = 156.81
Prob > F = 0.0000
R-squared = 0.3561
Adj R-squared = 0.3539
Root MSE = 1.9805

Number of obs = 570
F( 1, 568) = 284.89
Prob > F = 0.0000
R-squared = 0.3340
Adj R-squared = 0.3329
Root MSE = 2.0125

Number of obs = 570
F( 1, 568) = 83.59
Prob > F = 0.0000
R-squared = 0.1283
Adj R-squared = 0.1267
Root MSE = 2.3025

Does this imply that ASVABC explains 33% of the variance in HGC and HGCM 13%? No, because the multiple regression reveals that their joint explanatory power is 0.36, not 0.46.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

In the second regression, ASVABC is partly acting as a proxy for HGCM, and this inflates its apparent explanatory power. Similarly, in the third regression, HGCM is partly acting as a proxy for ASVABC, again inflating its apparent explanatory power.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

\[ \text{. reg \ lgearn hgc male} \]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>28.951332</td>
<td>2</td>
<td>14.475666</td>
</tr>
<tr>
<td>Residual</td>
<td>124.850561</td>
<td>567</td>
<td>.220194992</td>
</tr>
<tr>
<td>Total</td>
<td>153.801893</td>
<td>569</td>
<td>.270302096</td>
</tr>
</tbody>
</table>

\[ \text{Number of obs = 570} \]

\[ F( 2, 567) = 65.74 \]

\[ \text{Prob > F} = 0.0000 \]

\[ \text{R-squared} = 0.1882 \]

\[ \text{Adj R-squared} = 0.1854 \]

\[ \text{Root MSE} = 0.46925 \]

| lgearn | Coef.     | Std. Err. | t    | P>|t|  | [95% Conf. Interval] |
|--------|-----------|-----------|------|------|----------------------|
| hgc    | 0.0818944 | 0.0079976 | 10.240 | 0.000 | 0.0661858 0.0976036 |
| male   | 0.2285156 | 0.0397695 | 5.746 | 0.000 | 0.1504021 0.3066291 |
| _cons  | 1.19254   | 1134845   | 10.508 | 0.000 | 0.9696386 1.4154415 |

\[ \text{Exp(.2285) = 1.2567} \]

However, it is also possible for omitted variable bias to lead to a reduction in the apparent explanatory power of a variable. This will be demonstrated using a simple earnings function model, supposing the logarithm of earnings to depend on HGC and MALE.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

```
. reg lgearn hgc male
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 570</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>28.951332</td>
<td>2</td>
<td>14.475666</td>
<td>F( 2, 567) = 65.74</td>
</tr>
<tr>
<td>Residual</td>
<td>124.850561</td>
<td>567</td>
<td>.220194992</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>153.801893</td>
<td>569</td>
<td>.270302096</td>
<td>R-squared = 0.1882</td>
</tr>
</tbody>
</table>

| lgearn | Coef.   | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|--------|---------|-----------|-------|------|---------------------|
| hgc    | .0818944 | .0079976  | 10.240| 0.000| .0661858 .097603   |
| male   | .2285156 | .0397695  | 5.746 | 0.000| .1504021 .3066291 |
| _cons  | 1.19254  | .1134845  | 10.508| 0.000| .9696386 1.415441  |

\[
LGREARN = \alpha + \beta_1 HGC + \beta_2 MALE + u
\]

\[
E(b_1) = \beta_1 + \beta_2 \frac{\text{Cov}(HGC, MALE)}{\text{Var}(HGC)}
\]

\[r_{HGC,MALE} = -0.002510\]

If we omit MALE from the regression, the coefficient of HGC should be subject to a downward bias. \(\beta_2\) is likely to be positive, and we know that Var(HGC) is positive, but Cov(HGC, MALE) is negative because HGC and MALE are negatively correlated.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

```
.reg lgearn hgc male
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 570</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>28.951332</td>
<td>2</td>
<td>14.475666</td>
<td>F( 2, 567) = 65.74</td>
</tr>
<tr>
<td>Residual</td>
<td>124.850561</td>
<td>567</td>
<td>.220194992</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>153.801893</td>
<td>569</td>
<td>.270302096</td>
<td>R-squared = 0.1882</td>
</tr>
</tbody>
</table>

| lgearn     | Coef.       | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|------------|-------------|-----------|-------|------|---------------------|
| hgc        | 0.0818944   | 0.0079976 | 10.240| 0.000| 0.0661858 0.097603 |
| male       | 0.2285156   | 0.0397695 | 5.746 | 0.000| 0.1504021 0.3066291|
| _cons      | 1.19254     | 0.1134845 | 10.508| 0.000| 0.9696386 1.415441 |

\[ LGEARN = \alpha + \beta_1 HGC + \beta_2 MALE + u \]

\[ E(b_2) = \beta_2 + \beta_1 \frac{\text{Cov}(HGC, MALE)}{\text{Var}(MALE)} \]

For the same reasons, the coefficient of MALE in a simple regression of LGEARN on MALE should be downwards biased.
VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

<table>
<thead>
<tr>
<th>. reg lgearn hgc male</th>
</tr>
</thead>
</table>

| lgearn | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|--------|--------|-----------|-------|-------|----------------------|
| hgc    | .0818944 | .0079976  | 10.240| 0.000 | .0661858 .097603 |
| male   | .2285156 | .0397695  | 5.746 | 0.000 | .1504021 .3066291 |
| _cons  | 1.19254  | .1134845  | 10.508| 0.000 | .9696386 1.415441 |

| . reg lgearn hgc |

| lgearn | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|--------|--------|-----------|-------|-------|----------------------|
| hgc    | .0792256 | .0082061  | 9.655 | 0.000 | .0631077 .0953435 |
| _cons  | 1.358919 | .1127785  | 12.049| 0.000 | 1.137406 1.580433 |

| . reg lgearn male |

| lgearn | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|--------|--------|-----------|-------|-------|----------------------|
| male   | .2048652 | .0431797  | 4.744 | 0.000 | .1200538 .2896767 |
| _cons  | 2.313324 | .032605   | 70.950| 0.000 | 2.249282 2.377365 |

As can be seen, the coefficients of HGC and MALE are indeed lower in the simple regressions.
A comparison of $R^2$ for the three regressions shows that the sum of $R^2$ in the simple regressions is actually less than $R^2$ in the multiple regression. However, in this case the difference is small because the negative correlation between HGC and MALE is small (-0.05)
MODEL MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE
### Consequences of Variable Misspecification

<table>
<thead>
<tr>
<th>True Model</th>
<th>Fitted Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y = \alpha + \beta_1 x_1 + u$</td>
<td>$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u$</td>
</tr>
<tr>
<td>Correct specification, no problems</td>
<td>Coefficients are biased (in general). Standard errors are invalid.</td>
</tr>
</tbody>
</table>

In this sequence we will investigate the consequences of including an irrelevant variable in a regression model.
## Variable Misspecification II: Inclusion of an Irrelevant Variable

### Consequences of Variable Misspecification

<table>
<thead>
<tr>
<th></th>
<th>True Model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( y = \alpha + \beta_1 x_1 + u )</td>
<td>( y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u )</td>
</tr>
<tr>
<td>Fitted Model</td>
<td>( \hat{y} = a + b_1 x_1 )</td>
<td>Correct specification, no problems</td>
</tr>
<tr>
<td></td>
<td>( \hat{y} = a + b_1 x_1 + b_2 x_2 )</td>
<td>Coefficients are biased (in general). Standard errors are invalid.</td>
</tr>
<tr>
<td></td>
<td>Correct specification, but inefficient. Standard errors are valid (in general)</td>
<td>Correct specification, no problems</td>
</tr>
</tbody>
</table>

The effects are different from those of omitted variable misspecification. In this case the coefficients in general remain unbiased, but they are inefficient. The standard errors remain valid, but are needlessly large.
VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + u \]

\[ \hat{y} = a + b_1 x_1 + b_2 x_2 \]

These results can be demonstrated quickly.
VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + u \]

\[ \hat{y} = a + b_1 x_1 + b_2 x_2 \]

\[ y = \alpha + \beta_1 x_1 + 0 x_2 + u \]

Rewrite the true model adding \( x_2 \) as an explanatory variable, with a coefficient of 0. Now the true model and the fitted model coincide. Hence \( b_1 \) will be an unbiased estimator of \( \beta_1 \) and \( b_2 \) will be an unbiased estimator of 0.
VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + u \]

\[ \hat{y} = a + b_1 x_1 + b_2 x_2 \]

\[ y = \alpha + \beta_1 x_1 + 0x_2 + u \]

\[ \sigma_{b_1}^2 = \frac{\sigma_u^2}{n \text{Var}(x_1)} \times \frac{1}{1 - r_{x_1,x_2}^2} \]

However, the population variance of \( b_1 \) will be larger than it would have been if the correct simple regression had been run because it includes the factor \( 1/\sqrt{1-r_{x_1,x_2}^2} \).
VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + u \]

\[ \hat{y} = a + b_1 x_1 + b_2 x_2 \]

\[ y = \alpha + \beta_1 x_1 + 0x_2 + u \]

\[ \sigma_{b_1}^2 = \frac{\sigma_u^2}{n\text{Var}(x_1)} \times \frac{1}{1 - r_{x_1,x_2}^2} \]

The estimator of \( b_1 \) using the multiple regression model will therefore be less efficient than the alternative using the simple regression model.
VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + u \]

\[ \hat{y} = a + b_1 x_1 + b_2 x_2 \]

\[ y = \alpha + \beta_1 x_1 + 0 x_2 + u \]

\[ \sigma_{\beta_1}^2 = \frac{\sigma_u^2}{n \text{Var}(x_1)} \times \frac{1}{1 - r_{x_1, x_2}^2} \]

The intuitive reason for this is that the simple regression model exploits the information that \( x_2 \) should not be in the regression, while with the multiple regression model you find this out from the regression results.
VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + u \]

\[ \hat{y} = a + b_1 x_1 + b_2 x_2 \]

\[ y = \alpha + \beta_1 x_1 + 0x_2 + u \]

\[ \sigma_{b_1}^2 = \frac{\sigma_u^2}{n \text{Var}(x_1)} \times \frac{1}{1 - r_{x_1,x_2}^2} \]

The standard errors remain valid, because the model is formally correctly specified, but they will tend to be larger than those obtained in a simple regression, reflecting the loss of efficiency.
VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

\[ y = \alpha + \beta_1 x_1 + u \]

\[ \hat{y} = a + b_1 x_1 + b_2 x_2 \]

\[ y = \alpha + \beta_1 x_1 + 0 x_2 + u \]

\[ \sigma_{b_1}^2 = \frac{\sigma_u^2}{n \text{Var}(x_1)} \times \frac{1}{1 - r_{x_1, x_2}^2} \]

These are the results in general. Note that if \( x_1 \) and \( x_2 \) happen to be uncorrelated, there will be no loss of efficiency after all.
The analysis will be illustrated using a basic semilogarithmic earnings function. The result of regressing $LGEARN$ on $HGC$ and $ASVABC$ is shown above.
Now add the parental education variables, *HGCM* and *HGCF*. These variables are determinants of educational attainment, and hence indirectly affect earnings, but there is no evidence that they have any additional direct effect on earnings.
### VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

\[ \text{reg lgearn hgc asvabc hgcm hgcf} \]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 570</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>26.3617806</td>
<td>4</td>
<td>6.59044515</td>
<td>F( 4, 565) = 29.22</td>
</tr>
<tr>
<td>Residual</td>
<td>127.440112</td>
<td>565</td>
<td>.22555772</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>153.801893</td>
<td>569</td>
<td>.270302096</td>
<td>R-squared = 0.1714</td>
</tr>
</tbody>
</table>

\[ \text{adj R-squared = 0.1655; Root MSE = 0.47493} \]

| lgearn | Coef.   | Std. Err. | t     | P>|t| | 95% Conf. Interval |
|--------|---------|-----------|-------|-----|-------------------|
| hgc    | .0511811| .0101812  | 5.027 | 0.000 | .0311835 - .0711788 |
| asvabc | .010444 | .0027481  | 3.800 | 0.000 | .0050463 - .0158417 |
| hgcm   | .0071835| .0102695  | 0.699 | 0.485 | -.0129876 - .0273547 |
| hgcf   | .004794 | .0076389  | 0.628 | 0.531 | -.0102101 - .0197981 |
| _cons  | 1.073972| .1324621  | 8.108 | 0.000 | .8137933 - 1.33415 |

The fact that the *t* statistics of both variables are low is evidence that they are probably irrelevant.
VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

```
. reg lgearn hgc asvabc

| lgearn | Coef.   | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|--------|---------|-----------|-------|------|---------------------|
| hgc    | 0.0544266 | 0.0099018 | 5.497 | 0.000 | 0.034978 0.0738753   |
| asvabc | 0.0114733 | 0.0026476 | 4.333 | 0.000 | 0.0062729 0.0166736  |
| _cons  | 1.118832  | 0.124107  | 9.015 | 0.000 | 0.8750665 1.362598   |
```

```
. reg lgearn hgc asvabc hgcm hgcf

| lgearn | Coef.   | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|--------|---------|-----------|-------|------|---------------------|
| hgc    | 0.0511811 | 0.0101812 | 5.027 | 0.000 | 0.0311835 0.0711788 |
| asvabc | 0.010444  | 0.0027481 | 3.800 | 0.000 | 0.0050463 0.0158417 |
| hgcm   | 0.0071835 | 0.0102695 | 0.699 | 0.485 | -0.0129876 0.0273547|
| hgcf   | 0.004794  | 0.0076389 | 0.628 | 0.531 | -0.0102101 0.0197981|
| _cons  | 1.073972  | 0.1324621 | 8.108 | 0.000 | 0.8137933 1.33415   |
```

There is no evidence that the inclusion of the parental education variables has caused the other coefficients to be biased. The other coefficients have changed, but the changes are small in relation to the standard errors and appear to be chance movements.
VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

```
. reg lgearn hgc asvabc

         lgearn      |      Coef.     Std. Err.     t     P>|t|     [95% Conf. Interval]
-----------------+-----------------------------------------------
           hgc     |    0.054427   0.0099018     5.497   0.000       0.034978    0.0738753
       asvabc     |    0.011473   0.0026476     4.333   0.000       0.006273    0.0166736
          _cons    |    1.118832   0.124107     9.015   0.000       0.875066    1.362598

. reg lgearn hgc asvabc hgcm hgcf

         lgearn      |      Coef.     Std. Err.     t     P>|t|     [95% Conf. Interval]
-----------------+-----------------------------------------------
           hgc     |    0.051181   0.0101812     5.027   0.000       0.031184    0.0711788
       asvabc     |    0.010444   0.0027481     3.800   0.000       0.005047    0.0158417
          hgcm     |    0.007184   0.0102695     0.699   0.485      -0.012988    0.0273547
           hgcf    |    0.004794   0.0076389     0.628   0.531      -0.010210    0.0197981
          _cons    |    1.073972   0.1324621     8.108   0.000       0.813793    1.334150
```

The standard errors are larger in the misspecified model, reflecting the loss of efficiency.
However, the loss of efficiency is not very great. The parental education variables are correlated with both *HGC* and *ASVABC* but, with a sample as large as the present one, the correlation has to be greater for the loss of efficiency to become a serious problem.
PROXY VARIABLES
PROXY VARIABLES

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \]

Suppose that a variable \( y \) is hypothesized to depend on a set of explanatory variables \( x_1, \ldots, x_k \) as shown above, and suppose that for some reason there are no data on \( x_1 \).
PROXY VARIABLES

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \]

As we have seen, a regression of \( y \) on \( x_2, \ldots, x_k \) would yield biased estimates of the coefficients and invalid standard errors and tests.
Sometimes, however, these problems can be reduced or eliminated by using a proxy variable in the place of $x_1$. A proxy variable is one which is hypothesized to be linearly related to the missing variable. In the present example, $z$ could act as a proxy for $x_1$. 

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$  

$$x_1 = \lambda + \mu z$$
The validity of the proxy relationship must be justified on the basis of theory, common sense, or experience. It cannot be checked directly because there are no data on $x_1$. 

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

$$x_1 = \lambda + \mu z$$
PROXY VARIABLES

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \]

\[ x_1 = \lambda + \mu z \]

\[ y = \alpha + \beta_1 (\lambda + \mu z) + \beta_2 x_2 + \ldots + \beta_k x_k + u \]

If a suitable proxy has been identified, the regression model can be rewritten as shown.
PROXY VARIABLES

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \]

\[ x_1 = \lambda + \mu z \]

\[ y = \alpha + \beta_1 (\lambda + \mu z) + \beta_2 x_2 + \ldots + \beta_k x_k + u \]

\[ = (\alpha + \beta_1 \lambda) + \beta_1 \mu z + \beta_2 x_2 + \ldots + \beta_k x_k + u \]

We thus obtain a model with all variables observable. If the proxy relationship is an exact one, and we fit this relationship, most of the regression results will be rescued.
PROXY VARIABLES

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \]

\[ x_1 = \lambda + \mu z \]

\[ y = \alpha + \beta_1 (\lambda + \mu z) + \beta_2 x_2 + \ldots + \beta_k x_k + u \]

\[ = (\alpha + \beta_1 \lambda) + \beta_1 \mu z + \beta_2 x_2 + \ldots + \beta_k x_k + u \]

1. The estimates of the coefficients of \( x_2, \ldots, x_k \) will be the same as those that would have been obtained if it had been possible to regress \( y \) on \( x_1, \ldots, x_k \).

2. The standard errors and \( t \) statistics of the coefficients of \( x_2, \ldots, x_k \) will be the same as those that would have been obtained if it had been possible to regress \( y \) on \( x_1, \ldots, x_k \).

3. \( R^2 \) will be the same as it would have been if it had been possible to regress \( y \) on \( x_1, \ldots, x_k \).

4. The coefficient of \( z \) will be an estimate of \( \beta_1 \mu \), and so it will not be possible to obtain an estimate of \( \beta_1 \), unless you are able to guess the value of \( \mu \). However the \( t \) statistic for \( z \) will be the same as that which would have been obtained for \( x_1 \) if it had been possible to regress \( y \) on \( x_1, \ldots, x_k \), and so you are able to assess the significance of \( x_1 \), even if you are not able to estimate its coefficient.
PROXY VARIABLES

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \]

\[ x_1 = \lambda + \mu z \]

\[ y = \alpha + \beta_1 (\lambda + \mu z) + \beta_2 x_2 + \ldots + \beta_k x_k + u \]

\[ = (\alpha + \beta_1 \lambda) + \beta_1 \mu z + \beta_2 x_2 + \ldots + \beta_k x_k + u \]

It will not be possible to obtain an estimate of \( \alpha \) since the intercept in the revised model is \((\alpha + \beta_1 \lambda)\), but usually \( \alpha \) is of relatively little interest, anyway.
PROXY VARIABLES

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \]

\[ x_1 = \lambda + \mu z \]

\[ y = \alpha + \beta_1 (\lambda + \mu z) + \beta_2 x_2 + \ldots + \beta_k x_k + u \]

\[ = (\alpha + \beta_1 \lambda) + \beta_1 \mu z + \beta_2 x_2 + \ldots + \beta_k x_k + u \]

It is generally more realistic to hypothesize that the relationship between \( x_1 \) and \( z \) is approximate, rather than exact. In that case the four results listed above will hold approximately.

However, if \( z \) is a poor proxy for \( x_1 \), then it is possible that some of the other \( x \) variables will try to act as proxies for it, and there will still be a problem of omitted variable bias.
The use of a proxy variable will be illustrated with an educational attainment model. We will suppose that educational attainment depends jointly on cognitive ability and family background.

As usual, ASVABC will be used as the measure of cognitive ability. However, there is no "family background" variable in the data set. Indeed, it is difficult to conceive how such a variable might be defined.
Instead, we will try to find a proxy. One obvious variable is the mother's educational attainment, \( HGCM \). However, father's educational attainment may also be relevant. So we will hypothesize that the family background index depends on both.
PROXY VARIABLES

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 INDEX + u \]

\[ INDEX = \lambda + \mu_1 HGCM + \mu_2 HGCF \]

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 (\lambda + \mu_1 HGCM + \mu_2 HGCF) + u \]
\[ = (\alpha + \beta_2 \lambda) + \beta_1 ASVABC + \beta_2 \mu_1 HGCM + \beta_2 \mu_2 HGCF + u \]

Thus we obtain a relationship expressing \( HGC \) as a function of \( ASVABC, HGCM, \) and \( HGCF \).
Here is the corresponding regression using *EAEF* Data Set 21.
Here is the regression of HGC on ASVABC alone.
A comparison of the regressions indicates that the coefficient of ASVABC is biased upwards if we make no attempt to control for family background.
This is what we should expect. Both *HGCM* and *HGCF* are likely to have positive effects on educational attainment, and they are both positively correlated with *ASVABC*. 
PROXY VARIABLES

. reg hgc asvabc hgcm hgcf library siblings

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 570</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1285.58208</td>
<td>5</td>
<td>257.116416</td>
<td>F( 5, 564) = 66.87</td>
</tr>
<tr>
<td>Residual</td>
<td>2168.66529</td>
<td>564</td>
<td>3.84515122</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>3454.24737</td>
<td>569</td>
<td>6.07073351</td>
<td>R-squared = 0.3722</td>
</tr>
</tbody>
</table>

| hgc     | Coef.   | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|---------|---------|-----------|-------|-------|----------------------|
| asvabc  | 0.1277852 | 0.010054 | 12.710 | 0.000 | 0.1080373  0.147533 |
| hgcm    | 0.0619975 | 0.0427558 | 1.450 | 0.148 | -0.0219826  0.1459775 |
| hgcf    | 0.1045035 | 0.0314928 | 3.318 | 0.001 | 0.042646  0.166361 |
| library | 0.1151269 | 0.1969844 | 0.584 | 0.559 | -0.2717856  0.5020394 |
| siblings| -0.0509486 | 0.039956 | -1.275 | 0.203 | -0.1294293  0.027532 |
| _cons   | 5.236995  | 0.5665539 | 9.244 | 0.000 | 4.124181  6.349808 |

**LIBRARY** (a dummy variable equal to 1 if anyone in the family owned a library card when the respondent was 14) and **SIBLINGS** (number of brothers and sisters of the respondent) are two other variables in the data set which might act as proxies for family background.
The **LIBRARY** variable was one of three variables included in the National Longitudinal Survey of Youth to help pick up the influence of family background on education. It has the anticipated positive coefficient, but it is not significant.
There is a tendency for parents who are ambitious for their children to limit their number, so
*SIBLINGS* should be expected to have a negative coefficient. It does, but it is also not
significant.
There are further background variables which may be relevant for educational attainment: faith, ethnicity, and region of residence. These variables are supplied in the data set, but it will be left to you to experiment with them.
TESTING A LINEAR RESTRICTION
TESTING A LINEAR RESTRICTION

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + \beta_3 HGCF + u \]

In the last sequence it was argued that educational attainment might be related to cognitive ability and family background, with mother's and father's educational attainment proxying for the latter.
TESTING A LINEAR RESTRICTION

```
.reg hgc asvabc hgcm hgcf
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 570</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1278.24153</td>
<td>3</td>
<td>426.080508</td>
<td>F( 3, 566) = 110.83</td>
</tr>
<tr>
<td>Residual</td>
<td>2176.00584</td>
<td>566</td>
<td>3.84453329</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>3454.24737</td>
<td>569</td>
<td>6.07073351</td>
<td>R-squared = 0.3700</td>
</tr>
</tbody>
</table>

| hgc            | Coef.    | Std. Err. | t     | P>|t|    | [95% Conf. Interval] |
|----------------|----------|-----------|-------|--------|---------------------|
| asvabc         | .1295006 | .0099544  | 13.009| 0.000  | .1099486             | .1490527            |
| hgcm           | .069403  | .0422974  | 1.641 | 0.101  | -.013676             | .152482             |
| hgcf           | .1102684 | .0311948  | 3.535 | 0.000  | .0489967             | .1715401            |
| _cons          | 4.914654 | .5063527  | 9.706 | 0.000  | 3.920094             | 5.909214            |

However, when we run the regression using Data Set 21, we find the coefficient of mother's education is not significant.
TESTING A LINEAR RESTRICTION

```
. reg hgc asvabc hgcm hgcf
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 570</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1278.24153</td>
<td>3</td>
<td>426.080508</td>
<td>F( 3, 566) = 110.83</td>
</tr>
<tr>
<td>Residual</td>
<td>2176.00584</td>
<td>566</td>
<td>3.84453329</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>3454.24737</td>
<td>569</td>
<td>6.07073351</td>
<td>R-squared = 0.3700</td>
</tr>
</tbody>
</table>

```
| hgc | Coef.  | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|-----|--------|-----------|-------|-----|-------------------|
| asvabc | .1295006 | .0099544  | 13.009| 0.000| .1099486 - .1490527 |
| hgcm  | .069403  | .0422974  | 1.641 | 0.101| -.013676 - .152482  |
| hgcf  | .1102684 | .0311948  | 3.535 | 0.000| .0489967 - .1715401 |
| _cons | 4.914654 | .5063527  | 9.706 | 0.000| 3.920094 - 5.909214 |
```

```
. cor hgcm hgcf
(obs=570)

<table>
<thead>
<tr>
<th></th>
<th>hgcm</th>
<th>hgcf</th>
</tr>
</thead>
<tbody>
<tr>
<td>hgcm</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>hgcf</td>
<td>0.6391</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
```

As was noted in one of the sequences for Chapter 4, this might be due to multicollinearity, because mother's education and father's education are highly correlated.
Testing a Linear Restriction

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + \beta_3 HGCF + u \]

In the discussion of multicollinearity, several measures for alleviating the problem were suggested, among them the use of an appropriate theoretical restriction.
TESTING A LINEAR RESTRICTION

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + \beta_3 HGCF + u \]

\[ \beta_3 = \beta_2 \]

In particular, in the case of the present model, it was suggested that the impact of parental education might be the same for both parents, that is, that \( \beta_2 \) and \( \beta_3 \) might be equal.
TESTING A LINEAR RESTRICTION

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + \beta_3 HGCF + u \]

\[ \beta_3 = \beta_2 \]

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 (HGCM + HGCF) + u \]

\[ = \alpha + \beta_1 ASVABC + \beta_2 HGCP + u \]

\[ HGCP = HGCM + HGCF \]

If this is the case, the model may be rewritten as shown. We now have a total parental education variable, HGCP, instead of separate variables for mother's and father's education, and the multicollinearity caused by the correlation between the latter has been eliminated.
TESTING A LINEAR RESTRICTION

```stata
. g hgcp = hgcm + hgcf
. reg hgc asvabc hgcp
```

| Source | SS       | df    | MS      | Number of obs = 570 |
|--------+---------|-------|---------|---------------------|
| Model  | 1276.73764 | 2     | 638.368819 | F( 2, 567) = 166.22 |
| Residual | 2177.50973 | 567   | 3.84040517 | Prob > F = 0.0000 |
| Total  | 3454.24737 | 569   | 6.07073351 | R-squared = 0.3696 |

| hgc     | Coef.  | Std. Err. | t      | P>|t|  | [95% Conf. Interval] |
|---------|--------|-----------|--------|-----|-----------------------|
| asvabc  | 0.1295653 | 0.0099485 | 13.024 | 0.000 | 0.1100249          | 0.1491057 |
| hgcp    | 0.093741  | 0.0165688 | 5.658  | 0.000 | 0.0611973          | 0.1262847 |
| _cons   | 4.823123  | 0.4844829 | 9.955  | 0.000 | 3.871523           | 5.774724  |

Here is the regression with HGCP replacing HGCM and HGCF.
A comparison of the regressions reveals that the standard error of the coefficient of HGCP is much smaller than those of HGCM and HGCF, and consequently its t statistic is higher. Its coefficient is a compromise between those of HGCM and HGCF, as might be expected.
TESTING A LINEAR RESTRICTION

However, the use of a restriction will lead to a gain in efficiency only if the restriction is valid. If it is not valid, its use will lead to biased coefficients and invalid standard errors and tests.
TESTING A LINEAR RESTRICTION

```
. reg hgc asvabc hgcm hgcf

------------------------------------------------------------------------------
hgc | Coef. Std. Err. t P>|t| [95% Conf. Interval]
---------+--------------------------------------------------------------------
 asvabc |  .1295006  .0099544  13.009 0.000  .1099486  .1490527
 hgcm |  .069403  .0422974  1.641 0.101  -.013676  .152482
 hgcf |  .1102684  .0311948  3.535 0.000  .0489967  .1715401
 _cons |  4.914654  .5063527  9.706 0.000  3.920094  5.909214
------------------------------------------------------------------------------
```

```
. reg hgc asvabc hgcp

------------------------------------------------------------------------------
hgc | Coef. Std. Err. t P>|t| [95% Conf. Interval]
---------+--------------------------------------------------------------------
 asvabc |  .1295653  .0099485  13.024 0.000  .1100249  .1491057
 hgcp |  .093741  .0165688  5.658 0.000  .0611973  .1262847
 _cons |  4.823123  .4844829  9.955 0.000  3.871523  5.774724
------------------------------------------------------------------------------
```

Do the coefficients of HGCM and HGCF in the unrestricted regression look as if they satisfy the restriction? Not really, in this case. The coefficient of HGCM is much smaller than that of HGCF, but then it should be noted that the standard errors are quite large.
We will now perform a proper test. The imposition of a restriction makes it more difficult for the regression model to fit the data because there is one fewer parameter to adjust. There will therefore be an increase in $RSS$ (and a decrease in $R^2$) when it is imposed.
## TESTING A LINEAR RESTRICTION

```
. reg hgc asvabc hgcm hgcf
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 570</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1278.24153</td>
<td>3</td>
<td>426.080508</td>
<td>F( 3, 566) = 110.83</td>
</tr>
<tr>
<td>Residual</td>
<td>2176.00584</td>
<td>566</td>
<td>3.84453329</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>3454.24737</td>
<td>569</td>
<td>6.07073351</td>
<td>R-squared = 0.3700</td>
</tr>
</tbody>
</table>

```
. reg hgc asvabc hgcp
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 570</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1276.73764</td>
<td>2</td>
<td>638.368819</td>
<td>F( 2, 567) = 166.22</td>
</tr>
<tr>
<td>Residual</td>
<td>2177.50973</td>
<td>567</td>
<td>3.84040517</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>3454.24737</td>
<td>569</td>
<td>6.07073351</td>
<td>R-squared = 0.3696</td>
</tr>
</tbody>
</table>

If the restriction is valid, the deterioration in the fit should be a small, random amount. However, if the restriction is invalid, the distortion caused by its imposition will lead to a significant deterioration in the fit.

12
### TESTING A LINEAR RESTRICTION

```
. reg hgc asvabc hgcm hgcf

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1278.24153</td>
<td>3</td>
<td>426.080508</td>
</tr>
<tr>
<td>Residual</td>
<td>2176.00584</td>
<td>566</td>
<td>3.84453329</td>
</tr>
<tr>
<td>Total</td>
<td>3454.24737</td>
<td>569</td>
<td>6.07073351</td>
</tr>
</tbody>
</table>
```

Number of obs = 570
F( 3, 566) = 110.83
Prob > F = 0.0000
R-squared = 0.3700
Adj R-squared = 0.3667
Root MSE = 1.9607

```
. reg hgc asvabc hgcp

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1276.73764</td>
<td>2</td>
<td>638.368819</td>
</tr>
<tr>
<td>Residual</td>
<td>2177.50973</td>
<td>567</td>
<td>3.84040517</td>
</tr>
<tr>
<td>Total</td>
<td>3454.24737</td>
<td>569</td>
<td>6.07073351</td>
</tr>
</tbody>
</table>
```

Number of obs = 570
F( 2, 567) = 166.22
Prob > F = 0.0000
R-squared = 0.3696
Adj R-squared = 0.3674
Root MSE = 1.9597

In the present case, we can see that the increase in RSS is very small, and hence we are unlikely to reject the restriction.
TESTING A LINEAR RESTRICTION

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + \beta_3 HGCF + u \]

\[ \beta_3 = \beta_2 \]

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 (HGCM + HGCF) + u \]

\[ = \alpha + \beta_1 ASVABC + \beta_2 HGCP + u \]

\[ HGCP = HGCM + HGCF \]

\[ H_0 : \beta_3 = \beta_2, \quad H_1 : \beta_3 \neq \beta_2 \]

The null hypothesis is that the restriction is valid, and the alternative one is that it is invalid.
TESTING A LINEAR RESTRICTION

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + \beta_3 HGCF + u \]

\[ \beta_3 = \beta_2 \]

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 (HGCM + HGCF) + u \]
\[ = \alpha + \beta_1 ASVABC + \beta_2 HGCP + u \]

\[ HGCP = HGCM + HGCF \]

\[ H_0 : \beta_3 = \beta_2, \quad H_1 : \beta_3 \neq \beta_2 \]

\[ F = \frac{(RSS_R - RSS_U) / 1}{RSS_U / (n - k - 1)} = \frac{2177.51 - 2176.01}{2176.01 / 566} = 0.39 \]

The test statistic is a member of the family of \( F \) tests where the numerator is the improvement in the fit on relaxing the restriction, divided by the cost of relaxing it (one degree of freedom, because one additional parameter has to be estimated).
TESTING A LINEAR RESTRICTION

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + \beta_3 HGCF + u \]

\[ \beta_3 = \beta_2 \]

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 (HGCM + HGCF) + u \]
\[ = \alpha + \beta_1 ASVABC + \beta_2 HGCP + u \]

\[ HGCP = HGCM + HGCF \]

\[ H_0 : \beta_3 = \beta_2, \quad H_1 : \beta_3 \neq \beta_2 \]

\[ F = \frac{(RSS_R - RSS_U)/1}{RSS_U/(n - k - 1)} = \frac{2177.51 - 2176.01}{2176.01/566} = 0.39 \]

The denominator of the test statistic is RSS after making the improvement (that is, RSS for the unrestricted model), divided by the number of degrees of freedom remaining.
TESTING A LINEAR RESTRICTION

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + \beta_3 HGCF + u \]

\[ \beta_3 = \beta_2 \]

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 (HGCM + HGCF) + u \]

\[ = \alpha + \beta_1 ASVABC + \beta_2 HGCP + u \]

\[ HGCP = HGCM + HGCF \]

\[ H_0 : \beta_3 = \beta_2, \quad H_1 : \beta_3 \neq \beta_2 \]

\[ F = \frac{(RSS_R - RSS_U)}{RSS_U / (n - k - 1)} = \frac{2177.51 - 2176.01}{2176.01 / 566} = 0.39 \]

The \( F \) statistic is 0.39. An \( F \) statistic below 1 is never significant (look at the \( F \) table), so we do not reject \( H_0 \). The restriction appears to be valid. At least, it is not rejected by the data.
TESTING A LINEAR RESTRICTION

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + \beta_3 HGCF + u \]

\[ \beta_3 = \beta_2 \]

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 (HGCM + HGCF) + u \]

\[ = \alpha + \beta_1 ASVABC + \beta_2 HGCP + u \]

Linear restrictions can also be tested using a \( t \) test. This involves writing down the model for the restricted version and adding the term which would convert it back to the unrestricted version. The test evaluates whether this additional term is needed.
TESTING A LINEAR RESTRICTION

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + \beta_3 HGCF + u \]

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCP + u \]

\[ 0 = \beta_2 HGCM + \beta_3 HGCF - \beta_2 HGCP \]

To find the conversion term, we write the restricted version of the model under the unrestricted version and subtract.
TESTING A LINEAR RESTRICTION

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + \beta_3 HGCF + u \]

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCP + u \]

\[ 0 = \beta_2 HGCM + \beta_3 HGCF - \beta_2 HGCP \]
\[ = \beta_2 HGCM + \beta_3 HGCF - \beta_2 (HGCM + HGCF) \]
\[ = (\beta_3 - \beta_2) HGCF \]

We see that the term which converts the restricted model back to the unrestricted one is \((\beta_3 - \beta_2) HGCF\).
TESTING A LINEAR RESTRICTION

\[ \begin{align*}
&HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + \beta_3 HGCF + u \\
&HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCP + u \\
&0 = \beta_2 HGCM + \beta_3 HGCF - \beta_2 HGCP \\
&= \beta_2 HGCM + \beta_3 HGCF - \beta_2 (HGCM + HGCF) \\
&= (\beta_3 - \beta_2) HGCF \\
&HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCP + (\beta_3 - \beta_2) HGCF + u
\end{align*} \]

We add this term to the restricted model and investigate whether it is needed.
TESTING A LINEAR RESTRICTION

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + \beta_3 HGCF + u \]

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCP + u \]

\[ 0 = \beta_2 HGCM + \beta_3 HGCF - \beta_2 HGCP \]
\[ = \beta_2 HGCM + \beta_3 HGCF - \beta_2 (HGCM + HGCF) \]
\[ = (\beta_3 - \beta_2) HGCF \]

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCP + (\beta_3 - \beta_2) HGCF + u \]

\[ H_0 : \beta_3 - \beta_2 = 0, \quad H_1 : \beta_3 - \beta_2 \neq 0 \]

The null hypothesis is that the coefficient of the conversion term is 0, and the alternative hypothesis is that it is different from 0.
TESTING A LINEAR RESTRICTION

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCM + \beta_3 HGCF + u \]

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCP + u \]

\[ 0 = \beta_2 HGCM + \beta_3 HGCF - \beta_2 HGCP \]
\[ = \beta_2 HGCM + \beta_3 HGCF - \beta_2 (HGCM + HGCF) \]
\[ = (\beta_3 - \beta_2) HGCF \]

\[ HGC = \alpha + \beta_1 ASVABC + \beta_2 HGCP + (\beta_3 - \beta_2) HGCF + u \]

\[ H_0 : \beta_3 - \beta_2 = 0, \quad H_1 : \beta_3 - \beta_2 \neq 0 \]

Of course the null hypothesis is that the restriction is valid. If it is valid, the conversion term is not needed, and the restricted version is an adequate representation of the data.
TESTING A LINEAR RESTRICTION

. reg hgc asvabc hgcp hgcf

Source |       SS    df    MS
---------+------------------ F(  3,   566) = 110.83
Model    | 1278.24153     3  426.080508  Prob > F = 0.0000
Residual | 2176.00584   566  3.84453329  R-squared = 0.3700
---------+------------------------------ Adj R-squared = 0.3667
Total    | 3454.24737   569  6.07073351  Root MSE = 1.9607

hgc | Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+---------------------------------------------------------------------
asvabc |   .1295006   .0099544     13.009   0.000       .1099486    .1490527
hgcp  |   .069403   .0422974      1.641   0.101       -.013676     .152482
hgcf  |   .0408654   .0653386      0.625   0.532      -.0874704    .1692012
_cons  |   4.914654   .5063527      9.706   0.000       3.920094    5.909214

Here is the corresponding regression. We see that the coefficient of HGCF is not significantly different from zero, indicating that the term is not needed and that the restricted version is an adequate representation of the data.
It can be shown mathematically that the \textit{F} test and the \textit{t} test are equivalent. The \textit{F} statistic is the square of the \textit{t} statistic and the critical value of \textit{F} is the square of the critical value of \textit{t}.
MULTIPLE RESTRICTIONS
MULTIPLE RESTRICTIONS

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + u \]

\[ \beta_3 = \beta_2, \quad \beta_4 + \beta_5 = 0 \]

Multiple restrictions can be tested by multiple reparameterizations. Each one will result in one of the original parameters being dropped and replaced by a test statistic for the restriction.

For example, suppose that we have the model and hypothetical restrictions shown.
MULTIPLE RESTRICTIONS

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + u \]

\[ \beta_3 = \beta_2, \quad \beta_4 + \beta_5 = 0 \]

\[ \theta = \beta_3 - \beta_2, \quad \phi = \beta_4 + \beta_5 \]

We define the test statistics \( \theta \) and \( \phi \).

The restrictions can be written \( \theta = \phi = 0 \).
MULTIPLE RESTRICTIONS

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + u \]

\[ \beta_3 = \beta_2, \quad \beta_4 + \beta_5 = 0 \]

\[ \theta = \beta_3 - \beta_2, \quad \phi = \beta_4 + \beta_5 \]

\[ \beta_3 = \beta_2 + \theta, \quad \beta_5 = \phi - \beta_4 \]

We use these definitions to express one \( \beta \) parameter in terms of the other \( \beta \) parameter and \( \theta \) or \( \phi \).
MULTIPLE RESTRICTIONS

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + u \]

\[ \beta_3 = \beta_2, \quad \beta_4 + \beta_5 = 0 \]

\[ \theta = \beta_3 - \beta_2, \quad \phi = \beta_4 + \beta_5 \]

\[ \beta_3 = \beta_2 + \theta, \quad \beta_5 = \phi - \beta_4 \]

\[ Y = \beta_1 + \beta_2 X_2 + (\beta_2 + \theta)X_3 + \beta_4 X_4 + (\phi - \beta_4)X_5 + u \]

We substitute into the model.
MULTIPLE RESTRICTIONS

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + u \]

\[ \beta_3 = \beta_2, \quad \beta_4 + \beta_5 = 0 \]

\[ \theta = \beta_3 - \beta_2, \quad \phi = \beta_4 + \beta_5 \]

\[ \beta_3 = \beta_2 + \theta, \quad \beta_5 = \phi - \beta_4 \]

\[ Y = \beta_1 + \beta_2 X_2 + (\beta_2 + \theta)X_3 + \beta_4 X_4 + (\phi - \beta_4) X_5 + u \]

\[ = \beta_1 + \beta_2 (X_2 + X_3) + \beta_4 (X_4 - X_5) + \theta X_3 + \phi X_5 + u \]

We bring the \( \beta_2 \) components together. We do the same for the \( \beta_4 \) components.
MULTIPLE RESTRICTIONS

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + u \]

\[ \beta_3 = \beta_2, \quad \beta_4 + \beta_5 = 0 \]

\[ \theta = \beta_3 - \beta_2, \quad \phi = \beta_4 + \beta_5 \]

\[ \beta_3 = \beta_2 + \theta, \quad \beta_5 = \phi - \beta_4 \]

\[ Y = \beta_1 + \beta_2 X_2 + (\beta_2 + \theta)X_3 + \beta_4 X_4 + (\phi - \beta_4)X_5 + u \]

\[ = \beta_1 + \beta_2 (X_2 + X_3) + \beta_4 (X_4 - X_5) + \theta X_3 + \phi X_5 + u \]

\[ = \beta_1 + \beta_2 Z + \beta_4 W + \theta X_3 + \phi X_5 + u \]

\[ Z = X_2 + X_3 \]
\[ W = X_4 - X_5 \]

Defining \( Z = X_2 + X_3 \) and \( W = X_4 - X_5 \), we can rewrite the model as shown.
MULTIPLE RESTRICTIONS

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + u \]

\[ \beta_3 = \beta_2, \quad \beta_4 + \beta_5 = 0 \]

\[ \theta = \beta_3 - \beta_2, \quad \phi = \beta_4 + \beta_5 \]

\[ \beta_3 = \beta_2 + \theta, \quad \beta_5 = \phi - \beta_4 \]

\[ Y = \beta_1 + \beta_2 X_2 + (\beta_2 + \theta) X_3 + \beta_4 X_4 + (\phi - \beta_4) X_5 + u \]

\[ = \beta_1 + \beta_2 (X_2 + X_3) + \beta_4 (X_4 - X_5) + \theta X_3 + \phi X_5 + u \]

\[ = \beta_1 + \beta_2 Z + \beta_4 W + \theta X_3 + \phi X_5 + u \]

\[ Z = X_2 + X_3 \]

\[ W = X_4 - X_5 \]

We can now test the restrictions by regressing \( Y \) on \( Z, W, X_3, \) and \( X_5 \) and performing \( t \) tests on the coefficients of \( X_3 \) and \( X_5 \).
MULTIPLE RESTRICTIONS

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + u \]

\[ \beta_3 = \beta_2, \quad \beta_4 + \beta_5 = 0 \]

\[ \theta = \beta_3 - \beta_2, \quad \phi = \beta_4 + \beta_5 \]

\[ \beta_3 = \beta_2 + \theta, \quad \beta_5 = \phi - \beta_4 \]

\[ Y = \beta_1 + \beta_2 X_2 + (\beta_2 + \theta)X_3 + \beta_4 X_4 + (\phi - \beta_4)X_5 + u \]

\[ = \beta_1 + \beta_2 (X_2 + X_3) + \beta_4 (X_4 - X_5) + \theta X_3 + \phi X_5 + u \]

\[ = \beta_1 + \beta_2 Z + \beta_4 W + \theta X_3 + \phi X_5 + u \]

\[ Y = \beta_1 + \beta_2 Z + \beta_4 W + u \]

We could also perform a joint test of the restrictions, hypothesizing \( H_0: \theta = \phi = 0 \). This would involve comparing the residual sum of squares with that when fitting the fully restricted model where \( Y \) depends only on \( Z \) and \( W \).
MULTIPLE RESTRICTIONS

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + u \]

\[ \beta_3 = \beta_2, \quad \beta_4 + \beta_5 = 0 \]

\[ \theta = \beta_3 - \beta_2, \quad \phi = \beta_4 + \beta_5 \]

\[ \beta_3 = \beta_2 + \theta \]

\[ Y = \beta_1 + \beta_2 X_2 + (\beta_2 + \theta) X_3 + \beta_4 (X_4 - X_5) + \phi X_5 + u \]

\[ Y = \beta_1 + \beta_2 Z + \beta_4 W + \theta X_3 + \phi X_5 + u \]

\[ F(2, n - k) = \frac{(RSS_R - RSS_U)}{RSS_U / (n - k)} \]

The test statistic would be as shown, where \( RSS_U \) is the residual sum of squares in the unrestricted model, \( RSS_R \) is the residual sum of squares in the model with both restrictions, and \( k \) is the number of parameters in the original, unrestricted version.
MULTIPLE RESTRICTIONS

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + u \]

\[ \beta_3 = \beta_2, \quad \beta_4 + \beta_5 = 0 \]

\[ \theta = \beta_3 - \beta_2, \quad \phi = \beta_4 + \beta_5 \]

\[ \beta_3 = \beta_2 + 0 \]

\[ Y = \beta_1 + \beta_2 X_2 + (\beta_2 + 0) X_3 + \beta_4 (X_4 - X_5) + \theta X_3 + \phi X_5 + u \]

\[ Y = \beta_1 + \beta_2 Z + \beta_4 W + \theta X_3 + \phi X_5 + u \]

In general, if there were \( p \) restrictions being tested simultaneously, the test statistic would be as shown.
You will often encounter references to zero restrictions. This just means that a particular parameter is hypothesized to be equal to zero, for example, $\beta_5$ in the model above. Taken in isolation, the appropriate test is of course the $t$ test. 

It can be considered to be a special case of the $t$ test of a restriction discussed above where there is no need for reparameterization. The test statistic is just the $t$ statistic for the parameter in question.
MULTIPLE RESTRICTIONS

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u \quad RSS_U \]
\[ Y = \beta_1 + \beta_2 X_2 + u \quad RSS_R \]

\[ H_0 : \beta_3 = \beta_4 = 0 \]
\[ H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \text{ or both } \beta_3 \text{ and } \beta_4 \neq 0 \]

\[ F(2, n-k) = \frac{(RSS_R - RSS_U)}{RSS_U / (n-k)} / 2 \]

Likewise the testing of multiple zero restrictions can be thought of as a special case of the testing of multiple restrictions, again with no need for reparameterization. The example shown is for a model where there are two zero restrictions.
MULTIPLE RESTRICTIONS

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u \quad RSS_U \]
\[ Y = \beta_1 + \beta_2 X_2 + u \quad RSS_R \]

\[ H_0 : \beta_3 = \beta_4 = 0 \]
\[ H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \text{ or both } \beta_3 \text{ and } \beta_4 \neq 0 \]

\[ F(2, n - k) = \frac{(RSS_R - RSS_U) / 2}{RSS_U / (n - k)} \]

\[ F(p, n - k) = \frac{(RSS_R - RSS_U) / p}{RSS_U / (n - k)} \]

The \( F \) test of the joint explanatory power of a group of explanatory variables discussed in Section 3.5 in the text can be thought of in this way.
MULTIPLE RESTRICTIONS

Unrestricted model: \[ Y = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + u \quad RSS_U \]

Restricted model: \[ Y = \beta_1 + u \quad RSS_R \]

Restrictions: \[ \beta_2 = \beta_3 = \ldots = \beta_k = 0 \]

Hypotheses:
\[ H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0 \]
\[ H_1 : \text{at least one of the slope coefficients} \neq 0 \]

Even the $F$ statistic for the equation as a whole can be treated as a special case. Here the unrestricted and restricted models are as shown.
MULTIPLE RESTRICTIONS

Unrestricted model: \( Y = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + u \) \( RSS_U \)

Restricted model: \( Y = \beta_1 + u \) \( RSS_R \)

Restrictions: \( \beta_2 = \beta_3 = \ldots = \beta_k = 0 \)

Hypotheses: \( H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0 \)
\( H_1 : \) at least one of the slope coefficients \( \neq 0 \)

Fitting the restricted model: \( b_1 = \bar{Y} \)

When we fit the restricted model, we find that the OLS estimator of \( \beta_1 \) is the sample mean of \( Y \) (see Exercise 1.11).
MULTIPLE RESTRICTIONS

Unrestricted model: \[ Y = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + u \quad RSS_U \]

Restricted model: \[ Y = \beta_1 + u \quad RSS_R \]

Restrictions: \[ \beta_2 = \beta_3 = \ldots = \beta_k = 0 \]

Hypotheses: \[ H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0 \]
\[ H_1 : \text{at least one of the slope coefficients } \neq 0 \]

Fitting the restricted model: \[ b_1 = \bar{Y} \]
\[ \hat{Y}_i = b_1 = \bar{Y} \quad \text{for all } i \]

Hence the fitted value of \( Y \) in all observations is equal to the sample mean of \( Y \).
MULTIPLE RESTRICTIONS

Unrestricted model: \[ Y = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + u \quad RSS_U \]

Restricted model: \[ Y = \beta_1 + u \quad RSS_R \]

Restrictions: \[ \beta_2 = \beta_3 = ... = \beta_k = 0 \]

Hypotheses: \[ H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0 \]
\[ H_1 : \text{at least one of the slope coefficients } \neq 0 \]

Fitting the restricted model: \[ b_1 = \bar{Y} \]
\[ \hat{Y}_i = b_1 = \bar{Y} \quad \text{for all } i \]
\[ TSS = ESS + RSS \]
\[ \sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 \]

Now we know that for any OLS regression, \[ TSS = ESS + RSS. \]
MULTIPLE RESTRICTIONS

Unrestricted model: \[ Y = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + u \quad \text{RSS}_U \]

Restricted model: \[ Y = \beta_1 + u \quad \text{RSS}_R \]

Restrictions: \[ \beta_2 = \beta_3 = \ldots = \beta_k = 0 \]

Hypotheses: \[ H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0 \]
\[ H_1 : \text{at least one of the slope coefficients} \neq 0 \]

Fitting the restricted model:
\[ b_1 = \bar{Y} \]
\[ \hat{Y}_i = b_1 = \bar{Y} \quad \text{for all} \ i \]

\[ TSS = ESS + RSS \]
\[ \sum (Y_i - \bar{Y})^2 = 0 + \sum e_i^2 \]

Hence \( TSS = RSS \) for the restricted regression.
MULTIPLE RESTRICTIONS

Unrestricted model: \[ Y = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + u \] \[ RSS_U \]

Restricted model: \[ Y = \beta_1 + u \] \[ RSS_R \]

Restrictions: \[ \beta_2 = \beta_3 = \ldots = \beta_k = 0 \]

Hypotheses: \[ H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0 \]
\[ H_1 : \text{at least one of the slope coefficients} \neq 0 \]

Fitting the restricted model:
\[ b_1 = \bar{Y} \]
\[ \hat{Y}_i = b_1 = \bar{Y} \quad \text{for all } i \]
\[ TSS = ESS + RSS \]
\[ RSS_R = TSS \]

Obviously, if there are no explanatory variables, none of the variation in Y is explained by the model and so RSS is equal to TSS.
MULTIPLE RESTRICTIONS

Unrestricted model: \( Y = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + u \) \( RSS_U \)

Restricted model: \( Y = \beta_1 + u \) \( RSS_R \)

Restrictions: \( \beta_2 = \beta_3 = \ldots = \beta_k = 0 \)

Hypotheses: \( H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0 \)
\( H_1 : \) at least one of the slope coefficients \( \neq 0 \)

\[
F(k-1, n-k) = \frac{(RSS_R - RSS_U) / (k-1)}{RSS_U / (n-k)}
\]

Here is the \( F \) statistic for the comparison of the unrestricted model with all of the \( X \) variables and the restricted model with only the intercept.
MULTIPLE RESTRICTIONS

Unrestricted model: \[ Y = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + u \quad RSS_U \]

Restricted model: \[ Y = \beta_1 + u \quad RSS_R \]

Restrictions: \[ \beta_2 = \beta_3 = \ldots = \beta_k = 0 \]

Hypotheses: \[ H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0 \]
\[ H_1 : \text{at least one of the slope coefficients } \neq 0 \]

\[ F(k-1, n-k) = \frac{(RSS_R - RSS_U) / (k-1)}{RSS_U / (n-k)} \]
\[ = \frac{(TSS - RSS_U) / (k-1)}{RSS_U / (n-k)} \]

We have just seen that RSS from the restricted version is equal to TSS.
MULTIPLE RESTRICTIONS

Unrestricted model: \[ Y = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + u \quad RSS_U \]

Restricted model: \[ Y = \beta_1 + u \quad RSS_R \]

Restrictions: \[ \beta_2 = \beta_3 = \ldots = \beta_k = 0 \]

Hypotheses:
- \( H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0 \)
- \( H_1 : \) at least one of the slope coefficients \( \neq 0 \)

\[ F(k-1, n-k) = \frac{(RSS_R - RSS_U)/(k-1)}{RSS_U/(n-k)} \]

\[ TSS = ESS_U + RSS_U \]

Now we refer to the decomposition of \( TSS \) in the case of the unrestricted regression. This is similar to the decomposition for the restricted model, with the difference that \( RSS_U \) will in be smaller than \( RSS_R \) and \( ESS_U \) will be positive, instead of zero.
MULTIPLE RESTRICTIONS

Unrestricted model:  \[ Y = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + u \quad RSS_U \]

Restricted model: \[ Y = \beta_1 + u \quad RSS_R \]

Restrictions: \[ \beta_2 = \beta_3 = \ldots = \beta_k = 0 \]

Hypotheses: \[ H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0 \]
\[ H_1 : \text{at least one of the slope coefficients } \neq 0 \]

\[ F(k-1, n-k) = \frac{(RSS_R - RSS_U)/(k-1)}{RSS_U / (n-k)} \]

\[ = \frac{(TSS - RSS_U)/(k-1)}{RSS_U / (n-k)} = \frac{ESS_U / (k-1)}{RSS_U / (n-k)} \]

\[ TSS = ESS_U + RSS_U \]

Given the decomposition for the unrestricted version, we can rewrite the \( F \) statistic as shown. This is the expression for the \( F \) statistic for the equation as a whole that was given in Section 3.5.