The OLS model assumes there is no perfect linear relationship between any of the independent variables in the model. Sometimes one may create such relationships by accident (see the example of “perfect collinearity” below). When a perfect linear relationship exists, Stata must drop one of the variables. (If more than one linear relationship exists, Stata must drop all but one of the linearly related variables.) Linearly related variables have a correlation of plus or minus 1.0.

The OLS model is less able to determine the relationship between independent and dependent variables as the amount of correlation between two or more independents grows. When two or more independents are highly correlated, the model is said to suffer from “multicollinearity.” Stata will report findings, but the standard errors on the coefficients will be quite large. If you think about it, this makes sense. If two variables are highly correlated we can’t be sure which variable is at work – is it one’s age or one’s experience that matters? They move together. So Stata reports high standard errors – it can’t be sure which coefficient matters.

In this first example, we will create perfect collinearity by generating a new variable, agecalc, that is explicitly a linear combination of two other variables – experience (exper) and years of education (yrseduc).

```
. use "H:\Rockefeller Courses\PAD705\Problem Set Data\cps83.dta", clear
. gen agecalc=6+exper+ yrseduc
. reg  wklywage agecalc exper yrseduc
```

**Perfect Collinearity**

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs</th>
<th>F(  2,   997)</th>
<th>Prob &gt; F</th>
<th>R-squared</th>
<th>Adj R-squared</th>
<th>Root MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>7746871.24</td>
<td>2</td>
<td>3873435.62</td>
<td>1000</td>
<td>109.60</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>35236171.4</td>
<td>997</td>
<td>35342.198</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>42983042.6</td>
<td>999</td>
<td>43026.0687</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| wklywage   | Coef.    | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|------------|----------|-----------|-------|------|----------------------|
| agecalc    | 31.45905 | 2.226584  | 14.13 | 0.000 | 27.08972 – 35.82837  |
| exper      | -27.89783| 2.145028  | -13.01| 0.000 | -32.10712 – -23.68854|
| yrseduc    | (dropped)|          |       |      |                      |
| _cons      | -339.9829| 45.20905  | -7.52 | 0.000 | -428.6987 – -251.2671|

Revised: February 6, 2003
Correlation between variables

```
correlate agecalc age
(obs=1000)

<table>
<thead>
<tr>
<th></th>
<th>agecalc</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>agecalc</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
```

As expected, the correlation between the variables is perfect.

Multicollinearity

This example demonstrates less than perfect correlation between the variables. Here, age, experience (exper) and years of education (yrseduc) are all highly correlated with one another for obvious reasons.

```
reg wklywage age exper yrseduc

Source | SS        | df | MS        | Number of obs = 1000
--------+-----------+----+-----------+-----------------------
Model   | 7819653.38| 3  | 2606551.13| F( 3, 996) = 73.83
Residual| 35163389.2| 996| 35304.6077| Prob > F = 0.0000
--------+-----------+----+-----------+-----------------------
Total   | 42983042.6| 999| 43026.0687| Adj R-squared = 0.1795
--------+-----------+----+-----------+-----------------------

wklywage | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval]
---------|-------|-----------|-------|------|-------------------
age      | 191.2913 | 133.229 | 1.44  | 0.151 | -70.15031 to 452.733
exper    | -187.7669 | 133.2553 | -1.41 | 0.159 | -449.2603 to 73.7264
yrseduc  | -159.877 | 133.2787 | -1.20 | 0.231 | -421.4162 to 101.6622
_cons    | -1297.343 | 798.8881 | -1.62 | 0.105 | -2865.04 to 270.3536
---------|-------|-----------|-------|------|-------------------
```

Regression after dropping age

```
.reg wklywage exper yrseduc

Source | SS        | df | MS        | Number of obs = 1000
--------+-----------+----+-----------+-----------------------
Model   | 7746871.24| 2  | 3873435.62| F(  2, 997) = 109.60
Residual| 35236171.4| 997| 35342.198 | Prob > F = 0.0000
--------+-----------+----+-----------+-----------------------
Total   | 42983042.6| 999| 43026.0687| Adj R-squared = 0.1786
--------+-----------+----+-----------+-----------------------

wklywage | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval]
---------|-------|-----------|-------|------|-------------------
exper    | 3.561217 | .4326322 | 8.23  | 0.000 | 2.712243 to 4.410191
yrseduc  | 31.45905 | 2.226584 | 14.13 | 0.000 | 27.08972 to 35.82837
_cons    | -151.2286 | 32.28669 | -4.68 | 0.000 | -214.5863 to -87.87096
---------|-------|-----------|-------|------|-------------------
```
Correlation between variables

```
correlate exper yrseduc age
(obs=1000)

   |    exper  yrseduc      age
|---------------------------
exper |   1.0000
yrseduc | -0.2822   1.0000
age |   0.9811  -0.0912   1.0000
```

Once again, we see that there was a greater deal of correlation – though not perfect correlation – between two of the included explanatory variables – age and exper.

Another example: Using the VIF measurement in Stata.

Stata includes a “post-regression” command called “vif” (variance inflation factor) that can be used to detect multicollinearity. VIF is a “rule of thumb” technique – it creates a number but does not have a known theoretical distribution to tell us whether a VIF measurement of a certain size is a cause for concern to a mathematical certainty. The usual rule of thumb is that any variable with a VIF greater than 10 is probably a concern and if the average VIF is “substantially” greater than 1.0, there may be one or more collinear explanatory. In this example, weight violates the “no greater than 10” rule and length almost does.

```
use "C:\Stata\auto.dta"
(1978 Automobile Data)
regress price mpg weight length
```

```
| Source | SS     | df | MS     | Number of obs =  74
|--------|--------|----|--------|------------------
| Model  | 226957412 |  3 | 75652470.6 | F(  3,  70) = 12.98
| Residual | 408107984 | 70 | 5830114.06 | Prob > F  =  0.0000
| Total   | 635065396 | 73 | 8699525.97 | R-squared =  0.3574

| price | Coef. | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|-------|-------|-----------|-------|------|-----------------|
| mpg   | -86.78928 | 83.94335  | -1.03 | 0.305 | -254.209 80.63046 |
| weight | 4.364798  | 1.167455  | 3.74  | 0.000 | 2.036383 6.693213 |
| length | -104.8682 | 39.72154  | -2.64 | 0.010 | -184.0903 -25.64607 |
| _cons | 14542.43  | 5890.632  | 2.47  | 0.016 | 2793.94 26290.93 |
```

```
vif

<table>
<thead>
<tr>
<th>Variable</th>
<th>VIF</th>
<th>1/VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>weight</td>
<td>10.31</td>
<td>0.097010</td>
</tr>
<tr>
<td>length</td>
<td>9.79</td>
<td>0.102095</td>
</tr>
<tr>
<td>mpg</td>
<td>2.95</td>
<td>0.338610</td>
</tr>
</tbody>
</table>

Mean VIF | 7.69
```
Correlation between variables

```
corr price weight length mpg  
(obs=74)

|    price   weight   length      mpg  
|-----------------------------
|price |   1.0000
|weight|   0.5386   1.0000
|length|   0.4318   0.9460   1.0000
|mpg  |  -0.4686  -0.8072  -0.7958   1.0000
```

As the correlations show, length, weight, and mpg are all highly correlated. One solution is to drop weight from the regression:

```
regress price  mpg length
```

```
Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  2,    71) =   10.55
Model |   145463467     2  72731733.7           Prob > F      =  0.0001
Residual |   489601929    71  6895801.81           R-squared     =  0.2291
-------------+------------------------------           Adj R-squared =  0.2073
Total |   635065396    73  8699525.97           Root MSE      =  2626.0

------------------------------------------------------------------------------
price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
mpg |  -173.7022   87.72302    -1.98   0.052    -348.6169    1.212556
length |   21.28605   22.79323     0.93   0.354    -24.16237    66.73446
   _cons |   5864.305   5888.103     1.00   0.323    -5876.238    17604.85
------------------------------------------------------------------------------

.vif

|Variable|      VIF| 1/VIF |
-------------|---------|-------|
length |      2.73| 0.366735 |
mpg |      2.73| 0.366735 |

Mean VIF | 2.73
```

Now the first rule is met – no VIF over 10, but the mean is somewhat greater than 1.0 – whether a mean of 2.73 is too high is hard to say with certainty.