MULTIVARIATE REGRESSION I
Steps in Statistical Analysis, especially model specification and goodness of fit

We will be going over various aspects of regression problems during the next few weeks. These will include:

1. Omitted variables
2. Nonlinear relationships
3. Non-constant error variance
4. Correlation among errors
5. Nonnormal errors
6. Influential cases

First, however, we want to focus on

Steps in statistical analysis

1. Specify the model
   Usually a linear model with normally distributed homoscedastic error terms
   Transform data (dependent and/or independent variables) to meet normality assumptions

2. Estimate the parameters
   Frequently use the least squares criterion - choose as estimates of the parameters those that yield the smallest RSS (residual sum of squares).
   Later in the course, we will discuss another criterion that is frequently used, the maximum likelihood criterion for finding estimates.
   We have also briefly referred to robust estimates, those that are little affected by one or two unusual values.

3. Consider goodness of fit
   One criterion is the $R^2$ value - the proportion of the original variation about the mean (sum of the squared residuals about the mean) that is "explained" by the model.
   Another consideration is the pattern of residuals about the line -
   - are there unusual values (outliers)?
   - is there heteroscedasticity? or are they distributed about the predicted value in the same way, no matter what that predicted value?
   A third question to ask is whether the model fits better, or as well, as another model.

4. Interpret the results
   The b's are partial regression coefficients. If the coefficient of $X_i$ is estimated as $b$, we say that, if $X_i$ increases by 1, the expected of $Y$ increases by $b$.
   If the b's are standardized coefficients, we talk about everything in standardized units. The interpretation then is, if $X_i$ increases by one standard deviation of $X_i$, then the expected value of $Y$ increases by (standardized) $b$ standard deviations of $Y$.
   When we do cross-sectional studies, we can ONLY be sure that the coefficients represent association between the X variable and the dependent variable, $Y$. They do NOT represent causation. We will discuss this point further in later lectures.

We will return to this schema repeatedly.

Today I wanted to focus on points 1 and 3: specifying the model, and goodness of fit.

MODELS

In statistics, we usually talk about two models for our data:

- the model where $Y$ is the sum of a constant and an error term, which I will refer to as Model C (for constant):
- the model where $Y$ is the sum of its linear relationship to one predictor, $X$, and an error term, which I will refer to as Model 2p (2 parameters, $\hat{\alpha}_0$ and $\hat{\alpha}_1$):

$$Y = \hat{\alpha}_0 + \hat{\alpha}_1 X + \epsilon$$

Model C could have been called Model 1p, since it only has one parameter ($\hat{\mu}$).

Each of these models says that $Y$ is the sum of its expected value and an error term. The expected value in Model C is a constant, the mean; the expected value in Model 2p is the point on the line corresponding to the particular value of $X_i$.

In each case, we choose as the estimates of the parameters the values of $\hat{\mu}$ in Model C and $\hat{\alpha}_0$ and $\hat{\alpha}_1$ that minimize the sum of the squared errors. For Model C, the estimate of $\hat{\mu}$ is the sample mean. For model 2p, the equations for $\hat{\alpha}_0$ and $\hat{\alpha}_1$, the estimates of $\hat{\alpha}_0$ and $\hat{\alpha}_1$, were given last semester. I'd like to go over what we mean by these statements in some detail.

To do so, I will use a small dataset on the relationship of number of breeding pairs to area in 22 colonies.

```
use h1-1.dta
list colony area pairs
```

<table>
<thead>
<tr>
<th>colony</th>
<th>area</th>
<th>pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out Skerries</td>
<td>1616</td>
<td>284</td>
</tr>
<tr>
<td>N.E. Unst</td>
<td>1588</td>
<td>495</td>
</tr>
<tr>
<td>Fetlar</td>
<td>931</td>
<td>372</td>
</tr>
<tr>
<td>W. Yell</td>
<td>126</td>
<td>134</td>
</tr>
<tr>
<td>Vaila</td>
<td>302</td>
<td>278</td>
</tr>
<tr>
<td>Fitful Head</td>
<td>596</td>
<td>500</td>
</tr>
<tr>
<td>Uyea</td>
<td>898</td>
<td>731</td>
</tr>
<tr>
<td>W. Unst</td>
<td>208</td>
<td>311</td>
</tr>
<tr>
<td>Buravoe</td>
<td>353</td>
<td>485</td>
</tr>
<tr>
<td>St. Ninians</td>
<td>106</td>
<td>250</td>
</tr>
<tr>
<td>Papa Stour</td>
<td>809</td>
<td>1036</td>
</tr>
<tr>
<td>Gruney</td>
<td>565</td>
<td>1364</td>
</tr>
<tr>
<td>Eshaness</td>
<td>1069</td>
<td>2430</td>
</tr>
<tr>
<td>S. Havra</td>
<td>242</td>
<td>925</td>
</tr>
<tr>
<td>Foula</td>
<td>2927</td>
<td>5570</td>
</tr>
<tr>
<td>Moussa</td>
<td>614</td>
<td>1975</td>
</tr>
<tr>
<td>Sumburage</td>
<td>1273</td>
<td>3243</td>
</tr>
<tr>
<td>Hermaness</td>
<td>1570</td>
<td>3872</td>
</tr>
<tr>
<td>Reawick</td>
<td>111</td>
<td>970</td>
</tr>
<tr>
<td>Dalsetter</td>
<td>60</td>
<td>970</td>
</tr>
<tr>
<td>Fair Isle</td>
<td>3957</td>
<td>17000</td>
</tr>
<tr>
<td>Noss</td>
<td>1317</td>
<td>10767</td>
</tr>
</tbody>
</table>

**Regress**

```
regress Y X
```

```
Source | SS    | df | MS    | Number of obs = 22
--------|-------|----|-------|-----------------------
Instal  |       |    |       | F(  1,    20) = 32.06
```

---

Install Equation Editor and double-click here to view equation.

Install Equation Editor and double-click here to view equation.
Model | 213268475 1 213268475 | Prob > F = 0.0000
Residual | 133050026 20 6652501.31 | R-square = 0.6158
----------+--------------------------| Adj R-square = 0.5966
Total | 346318501 21 16491357.2 | Root MSE = 2579.2

| Y | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|---+------|----------|---|-----|------------------|
| X | 3.302147 | .5832108 | 5.662 | 0.000 | 2.085591 4.518704 |
| _cons | -734.9547 | 786.9989 | -0.934 | 0.362 | -2376.606 906.6963 |

Here I want to note that the Root MSE in the top panel of the table produced by regress is the estimate of the sd of the error about the line, i.e. the Root MSE estimates the sd of \( \epsilon \). It is calculated for you by STATA, but you could calculate it from other information in the table, specifically, the Root MSE.

Predicted value of Y, Y hat

We can find the predicted value of y, yhat2p, in 2 ways: brute force do it yourself or by using the command predict after a regression calculation; the results are exactly the same (but for rounding error):

```
. gen yhat2pm = -734.9547 + 3.302147*X  
  Brute force - my calculation
. predict yhat2p  
  predict used after regress
```

To convince ourselves both procedures give us the same results:

```
. list yhat2p  yhat2pm in 1/5  
  I could also use the command list yhat2p* in 1/5
   yhat2p  yhat2pm
  1. 4601.315 4601.315
  2. 4508.855 4508.854
  3. 2339.344 2339.344
  4. -318.8842 -318.8842
  5. 262.2938 262.2937
```

Generate the predicted value under Model C - all areas are predicted to have the same number of pairs and that number is the mean of pairs. Again we can do it two ways - brute force, or by using the command egen:

```
. gen yhatcm = 2452.818  
  Brute force - my calculation
. egen yhatc = mean(Y)  egen with the function mean
. list yhatc yhatcm in 1/5
```

Plot the observed Y 's, and those predicted under the two models, Model C and Model 2p, against X
Where do the SS given as a result of `regress` come from?

The graph above shows two predicted values for each Y -- one from Model C and the other from Model 2p.

and a part of the difference between Y and the mean that we consider explained by Model 2p. This is the difference I have called this the model difference below.

Again, I can use `predict` only after a `regress` command; here, by using the option `residual`, I am asking STATA to produce the residual, or error, terms rather than the predicted values of Y.

```
. graph Y yhatc yhat2p X, connect(.ll) xlabel ylabel
```

```
. gen ec = Y-yhatc
. predict e2p, residual
. gen model = yhat2p-yhatc
```
. list ec e2p model  
  ec        e2p      model       
1.  -2168.818 -4317.315   2148.497  
2.  -1957.818 -4013.855   2056.037  
3.  -2080.818  452.8842  -2771.702  
4.  -2174.818  15.70623  -2190.524  
5.  -1952.818  -733.1251  -1219.693  
6.  -1721.818 -1499.374  -222.4446  
7.  -2141.818  359.1081  -2500.926  
10. -2202.818  634.9271  -2837.745  
11. -1416.818  -900.4824  -516.3357  
12. -1088.818  233.2415  -1322.06  
13. -22.81812 -365.0407   342.2227  
14. -1527.818  860.8351  -2388.653  
15.  3117.182 -3360.43  6477.612  
16. -477.8181  682.4363  -1160.254  
17.  790.1819  -225.6788   1015.861  
18.  1419.182  -577.4165   1996.598  
19. -1482.818  1338.416  -2821.234  
20. -1482.818  1506.826  -2989.644  
21.  14547.18  4668.358   9878.823  
22.  8314.182  7153.027   1161.155

In every case, ec = e2p + model.

The sum of squares of each of these is one of the SS given in the regression output. We can calculate each of these using egen:
. egen ssec = sum(ec*ec)
. egen sse2p = sum(e2p^2)
. egen ssmodel = sum(model^2)

We can also calculate $R^2$ as the Model SS divided by the Total SS
. gen R2 = ssmodel/ssec
. list ssec sse2p ssmodel R2 in 1/5

<table>
<thead>
<tr>
<th></th>
<th>ssec</th>
<th>sse2p</th>
<th>ssmodel</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>3.46e+08</td>
<td>1.33e+08</td>
<td>2.13e+08</td>
<td>.615816</td>
</tr>
<tr>
<td>2.</td>
<td>3.46e+08</td>
<td>1.33e+08</td>
<td>2.13e+08</td>
<td>.615816</td>
</tr>
<tr>
<td>3.</td>
<td>3.46e+08</td>
<td>1.33e+08</td>
<td>2.13e+08</td>
<td>.615816</td>
</tr>
<tr>
<td>4.</td>
<td>3.46e+08</td>
<td>1.33e+08</td>
<td>2.13e+08</td>
<td>.615816</td>
</tr>
<tr>
<td>5.</td>
<td>3.46e+08</td>
<td>1.33e+08</td>
<td>2.13e+08</td>
<td>.615816</td>
</tr>
</tbody>
</table>
Please check to make sure you see that these sums of squares correspond to those in the regression output. Also, here I do the calculation that shows that the error terms can be found either by using the `predict` command in STATA or by generating the values ourselves:

```
. predict e2p, resid
. gen e2pm = Y - yhat2pm
. list e2p e2pm in 1/5
   +-----------------+----------------------+
<table>
<thead>
<tr>
<th>e2p</th>
<th>e2pm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>-4317.315</td>
</tr>
<tr>
<td>2.</td>
<td>-4013.855</td>
</tr>
<tr>
<td>3.</td>
<td>-1967.344</td>
</tr>
<tr>
<td>4.</td>
<td>452.8842</td>
</tr>
<tr>
<td>5.</td>
<td>15.70623</td>
</tr>
</tbody>
</table>
  +-----------------+----------------------+
```

**Goodness of fit**: The important point here: $R^2$ is a measure of how much better Model 2p fits our data than does Model C. The measure is based on the reduction in the sum of the squared errors in prediction of $Y$ when we move from a model with only a constant to one that has 2 parameters and assumes a straight line relationship between the two variables $Y$ and $X$.

**Is $R^2$ significantly different from 0?** One test is that the coefficient $b_1$ is significantly different from 0, as we saw from the regression output. This is equivalent to a test that the reduction in the SS is significant --- IF there is only one predictor variable. Otherwise we use the F test, which will be discussed in handout 2.

$R^2$, in this case, is .62, so that the variation of $Y$ about the line is 62% less than the variation of $Y$ about its sample mean.

But there are other aspects of goodness of fit. We can look at the residuals, the $e_{2p}$, themselves. One assumption of regression analysis is that the errors are homoscedastic — the distribution should be the same no matter what the values of $X$ - or of the predicted $Y$. That is, if the model fits well, then the estimated errors, the $e_i$'s, should not show a pattern. We, therefore, frequently look at the so-called residual plot, that plots $e$'s against the predicted $Y$, i.e. the points ($e_i$, $Y_i$). Usually the line $e=0$ is added to this type of plot to help see whether there is a pattern - which, ideally, should not be present.

```
. graph e2p yhat2p, yline(0)
```
Another plot that is helpful is the qnorm plot of the residuals, which helps us judge whether or not they are normally distributed.

```
. qnorm e2p
```

![Qnorm plot](image)

In this case, there seem to be both larger than expected residuals when $Y$ hat is large, and some evidence of non-normality.

**"FIXING" NON-NORMALITY OF THE RESIDUALS**

We use the scatterplots of $Y$ against $X$ and residual plots to help us determine, in the same way as for a single variable, whether a transformation will make our errors more normal.

Find transformations that normalize each variable -- here I go back to the original variable names and show part of the output of ladder.

```
. ladder area
Transformation   formula    Chi-sq(2)   P(Chi-sq)
-------------------------
    log        log(area)    1.23       0.540

. ladder pairs
Transformation   formula    Chi-sq(2)   P(Chi-sq)
-------------------------
    log        log(pairs)    1.85       0.397
```
. gen logpair = log(pairs)
. gen logarea = log(area)
. regress logpair logarea

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 22</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>12.4906569</td>
<td>1</td>
<td>12.4906569</td>
<td>F( 1, 20) = 11.45</td>
</tr>
<tr>
<td>Residual</td>
<td>21.8082911</td>
<td>20</td>
<td>1.09041455</td>
<td>Prob &gt; F = 0.0029</td>
</tr>
<tr>
<td>Total</td>
<td>34.2989479</td>
<td>21</td>
<td>1.63328323</td>
<td>R-square = 0.3642</td>
</tr>
</tbody>
</table>

| logpair | Coef.  | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|---------|--------|-----------|-------|------|-----------------------|
| logarea | .6783232 | .2004194 | 3.385 | 0.003 | .2602555    1.096391  |
| _cons  | 2.627175 | 1.293557 | 2.031 | 0.056 | -.0711384  5.325488  |

. * find predicted logpairs, error, standard error for predicting an area's
logpair and standard error for predicting the mean for a particular area
. predict logphat
. predict elogp, resid

Look at the same diagnostic graphs as before:

. graph logpair logphat logarea, connect(.l)
. graph elogp logphat
. qnorm elogp

The transformations, although they reduce $R^2$, seem to have reduced the deviations from normality with which we were concerned.
CONFIDENCE INTERVALS FOR THE PREDICTED VALUE

We can consider prediction of

1. the mean value of $Y$ for all cases with a specific value of $X$ and
2. the value of $Y$ for an individual with a specific value of $X$.

The model says the mean lies on the regression line, while an individual has an expected value on the line and an error that represents that case's deviation from the expected value.

1) The mean $Y$ for all cases with a specific value of $X$

Recall that if $y$ and $b$ are variables that are not correlated. Statisticians can prove that the estimates of the mean of $Y$ and of $b$ are not correlated. Therefore

We know what those variances are, so that
The confidence interval for the mean $Y$ corresponding to a specific value of $X$, $X_i$, uses, as the SE, the square root of the variance given above.

2. Confidence interval for the individual - person $i$: For a specific person, recall that

For the individual, we have to add the error term to the variance above. {Unfortunately, nearly all statistics texts use the same notation for both of these variances.}

Thus, our predictions, whether for the mean or for individuals, are "better" the closer the $X$ value is to the mean of the $X$'s. We can also plot the confidence intervals.

STATA will calculate for you the standard errors (sqrt of the variance) for both the predicted mean $Y$ and the predicted individual $Y$. To get confidence intervals, you have to multiply by the value needed to obtain the level of confidence you choose (e.g. 1.96 for a 95% confidence interval when the sample size is large).

```
predict newvar, stdf          standard error of predicted individual Y
predict newvar, stdp          standard error of predicted mean Y
```
I have created these variables for the relationship of logpair to logarea and graphed the predicted line with the confidence intervals for the predicted values for an individual

\[ \text{. predict selogi, stdf} \]
\[ \text{. predict selogm, stdp} \]
\[ \text{. list selog* in 1/5} \]

<table>
<thead>
<tr>
<th></th>
<th>selogi</th>
<th>selogm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.087462</td>
<td>0.3035774</td>
</tr>
<tr>
<td>2</td>
<td>1.086803</td>
<td>0.3012068</td>
</tr>
<tr>
<td>3</td>
<td>1.116194</td>
<td>0.3943033</td>
</tr>
<tr>
<td>4</td>
<td>1.071993</td>
<td>0.2423927</td>
</tr>
<tr>
<td>5</td>
<td>1.07137</td>
<td>0.2396242</td>
</tr>
</tbody>
</table>

Note that the se for the individual is always much larger than for the mean.

\[ \star \text{ t value for 95\% confidence interval when df=20 is 2.086 (from book)} \]
\[ \text{. gen loghi = logphat + 2.086* selogi} \]
\[ \text{. gen loglo = logphat - 2.086* selogi} \]
\[ \text{. graph loghi loglo logpair logphat logarea, connect(l.l) symbol(iiT) xlabel ylabel} \]
MUltivariate Regression

We will start by looking at a simple example, and then return to notation and theory.

The expected value of Y is now a linear function of more than one predictor variable, labeled as $X_1 \ldots X_{k-1}$. Y itself is equal to this linear function plus an error term

The model:  

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \ldots + \beta_{k-1} X_{i,k-1} + \epsilon_i$$

As with bivariate regression, we estimate the coefficients using the least squares criterion - finding the set of $\beta$'s that produce the smallest RSS.

. use h1-2
. list y x1 x2

<table>
<thead>
<tr>
<th></th>
<th>y</th>
<th>x1</th>
<th>x2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>190</td>
<td>73</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>254</td>
<td>57</td>
<td>23</td>
</tr>
<tr>
<td>3</td>
<td>181</td>
<td>67</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>209</td>
<td>27</td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>302</td>
<td>69</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>288</td>
<td>63</td>
<td>28</td>
</tr>
<tr>
<td>7</td>
<td>263</td>
<td>70</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>244</td>
<td>63</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>290</td>
<td>89</td>
<td>31</td>
</tr>
<tr>
<td>10</td>
<td>220</td>
<td>82</td>
<td>34</td>
</tr>
<tr>
<td>11</td>
<td>220</td>
<td>60</td>
<td>34</td>
</tr>
<tr>
<td>12</td>
<td>385</td>
<td>72</td>
<td>36</td>
</tr>
<tr>
<td>13</td>
<td>274</td>
<td>85</td>
<td>37</td>
</tr>
<tr>
<td>14</td>
<td>303</td>
<td>55</td>
<td>40</td>
</tr>
<tr>
<td>15</td>
<td>365</td>
<td>75</td>
<td>44</td>
</tr>
<tr>
<td>16</td>
<td>311</td>
<td>59</td>
<td>46</td>
</tr>
<tr>
<td>17</td>
<td>354</td>
<td>84</td>
<td>46</td>
</tr>
<tr>
<td>18</td>
<td>434</td>
<td>69</td>
<td>48</td>
</tr>
<tr>
<td>19</td>
<td>308</td>
<td>75</td>
<td>50</td>
</tr>
<tr>
<td>20</td>
<td>374</td>
<td>79</td>
<td>51</td>
</tr>
<tr>
<td>21</td>
<td>405</td>
<td>65</td>
<td>52</td>
</tr>
<tr>
<td>22</td>
<td>346</td>
<td>65</td>
<td>52</td>
</tr>
<tr>
<td>23</td>
<td>402</td>
<td>79</td>
<td>57</td>
</tr>
<tr>
<td>24</td>
<td>451</td>
<td>76</td>
<td>57</td>
</tr>
<tr>
<td>25</td>
<td>395</td>
<td>59</td>
<td>60</td>
</tr>
</tbody>
</table>

. summ y x1 x2

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>25</td>
<td>310.72</td>
<td>77.82915</td>
<td>181</td>
<td>451</td>
</tr>
<tr>
<td>x1</td>
<td>25</td>
<td>68.68</td>
<td>12.72701</td>
<td>27</td>
<td>89</td>
</tr>
<tr>
<td>x2</td>
<td>25</td>
<td>39.12</td>
<td>12.24854</td>
<td>20</td>
<td>60</td>
</tr>
</tbody>
</table>
. regress y x1 x2

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>102570.815</td>
<td>2</td>
<td>51285.4073</td>
<td>F( 2, 22) = 26.36</td>
</tr>
<tr>
<td>Residual</td>
<td>42806.2253</td>
<td>22</td>
<td>1945.73752</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>145377.04</td>
<td>24</td>
<td>6057.37667</td>
<td>R-square = 0.7056</td>
</tr>
</tbody>
</table>

| y          | Coef.     | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|------------|-----------|-----------|-------|-------|---------------------|
| x1         | 0.4173621 | 0.7287761 | 0.573 | 0.573 | -1.094027 to 1.928751 |
| x2         | 5.216591  | 0.7572445 | 6.889 | 0.000 | 3.646162 to 6.78702  |
| _cons      | 77.98254  | 52.42964  | 1.487 | 0.151 | -30.74988 to 186.715  |

The interpretation of the coefficients above is that y increases by .42 for each increase of 1 in x1 and that y increases by 5.2 for each increase of 1 in x2. The proportion of the variation in y that is explained by the association to both x1 and x2 is given by R-square - and is, in this case, .71.

The coefficients in multiple regression are also referred to as the partial coefficients. What is meant by this is that the interpretation I gave above is incomplete. It really is that, holding x2 constant, y increases by .42 for each increase in x1. Further, it says that this gives the relationship between y and x1 when the effect of x2 is eliminated from both y and x1. We will go on to this next time.
APPENDIX: LEAST SQUARES ESTIMATES

Whatever model we chose to use, the least squares approach leads to errors that average 0.

```
. sum e2p

         Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
         e2p   |      22    .0001162   4060.955  -2318.818   14547.18
---------+-----------------------------------------------------
```

In addition, the residual sum of squares is smaller than with any other line. You can test this by generating a different line with any intercept and slope you choose, finding the predicted y, eg.

```
. gen myyhat = -600 + 4.302147*area
```

and finding

```
Install Equation Editor and double-click here to view equation.
```

This sum will be larger than the Residual SS from the least squares line. Similarly, were you to calculate

```
Install Equation Editor and double-click here to view equation.
```

with a constant that does not equal the sample mean, you’d find it larger than the TSS.