CHAPTER 6: AUTOREGRESSION

1. Introduction

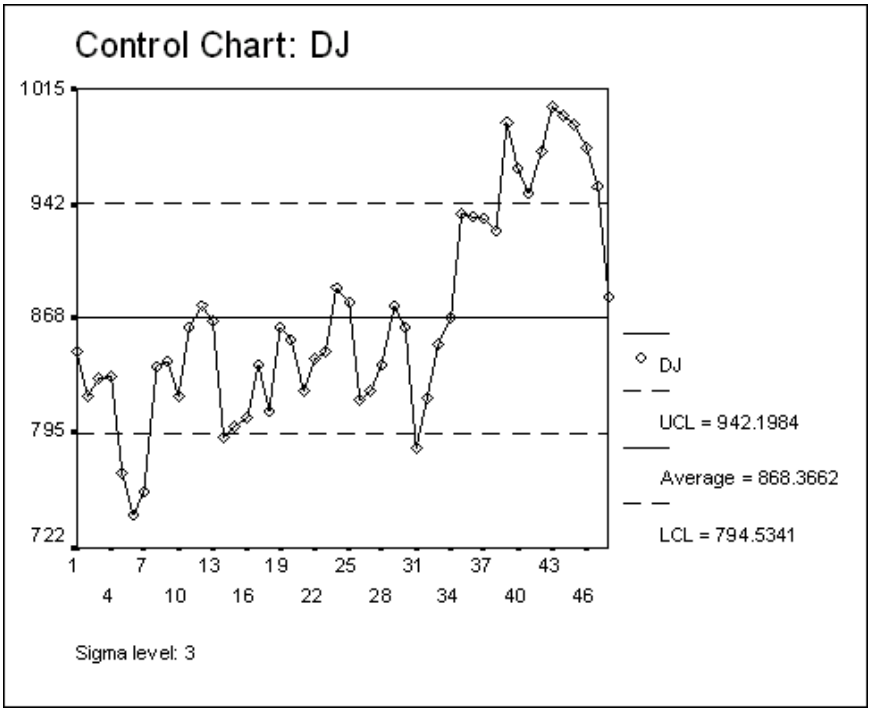
In Chapters 4 and 5, we have introduced regression analysis for time-ordered data. We have learned how to check for the presence of trend effects, periodic effects, special causes, and intervention effects. If we decide that any of these are present, we have learned to estimate their magnitude by including them in a regression model for the process. Understanding of such effects is essential both to quality improvement efforts and to holding the gains from past improvements.

We saw also how to make allowance for the effects of variables other than the series of immediate interest, as in the example of severity of illness in the study of mortality in intensive care in Chapter 4.

When there are several possible independent variables, we learned in Chapter 5 how to use stepwise regression for a rough screening of the data to find variables that seem to be most useful in explaining the variation of the dependent variable of interest.

The tools developed in Chapters 4 and 5 suffice to provide a good understanding of many data sets that you will encounter in practice. They do not, however, deal with **lagged effects**, in which what has happened in the past helps to predict the future.

We encountered one example of lagged effects, the monthly closings of the Dow Jones Industrial Average. A given month's closing tended to be relatively close to that of the previous month. Recall the control chart:



We saw that the process was wildly out of control. Instead of varying unpredictably about a fixed level, the points **meandered** through time, with each month usually closer to the previous month than to the earlier readings in general, and with wide swings of the general level with the passage of time. We can call such a relationship a **lagged effect** because the result of one time period tends to spill over into the next period or periods.

It turned out that the Dow Jones application could be simply analyzed by computing changes or differences from month to month and noting that these appeared to be in a state of statistical control.[1]

There are many applications in which the meandering tendency is much weaker than in the Dow Jones application, but in which lagged effects are present and must be contended with. Company sales data often provide a good example. In these applications, however, the lagged effects are less strong, and differencing is usually not a good strategy for analysis.

Instead, we will use earlier values of the dependent variable -- "**lagged variables**" -- **as independent variables in our regression models**. The term "autoregression" -- "self regression" -- is used for such regression models.

2. A Chemical Reactor Process

For illustration of the idea of autoregression, we shall use an application from chemical engineering in which temperature measurements were made on a chemical reactor process at one-minute time intervals. (In thinking of this process, you might be helped by the analogy of temperature of the water in a shower. If you could measure temperature every few seconds, you would probably find that, in general, the current measurement was closer to the prior reading than to earlier measurements.) The data are contained in the ***SPSS*** file called REACTOR.sav. They are taken from an example in Box and Jenkins, *Time Series Analysis: Forecasting and Control*, revised edition, Holden-Day, 1976.

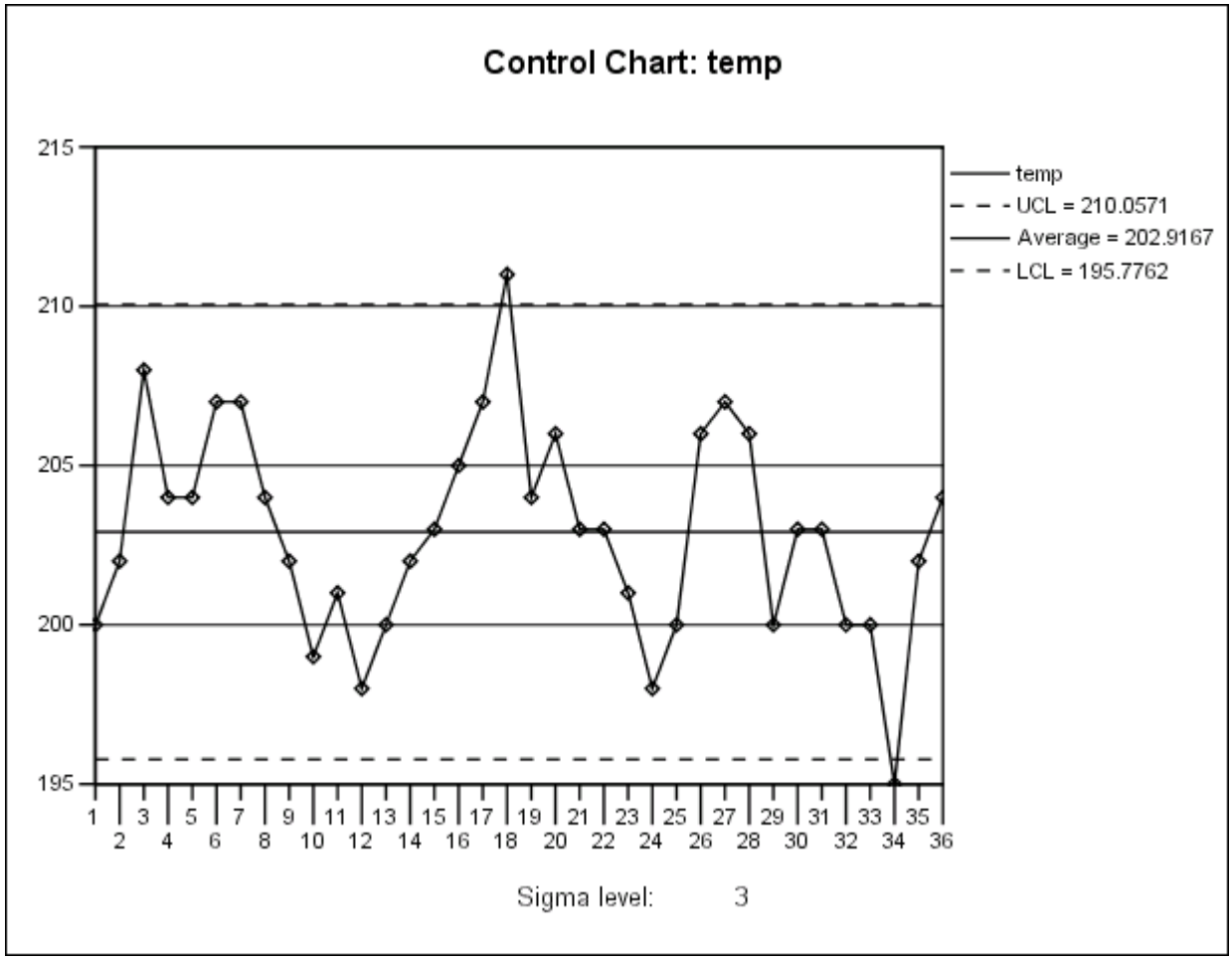
Naming the single variable in the data set **temp**, we open the file and call up **Analyze/Descriptive Statistics/ Descriptives…**

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
temp	36	195.00	211.00	202.9167	3.31555
Valid N (listwise)	36				

[1]The simple trick of differencing works for a wide variety of price series for organized exchanges. It is seldom useful in applications in quality management. If you are tempted to try it, here is a simple check on its potential value: difference the data and compare the standard deviation of the differences with the standard deviation of the original data. Unless the standard deviation of the differences is sharply lower than that of the original data, forget the idea of differencing and work with the original data.

We next examine the control chart for **temp**:



You can easily see the meandering tendency. Our task will be to model this tendency by means of autoregression.

In case you had any doubt about the meandering tendency, look at the runs count below. There are significantly fewer runs than the expected number which you can calculate to be 18.8. The data tend to persist above or below the mean longer than we would expect for in-control, or random, data.

Runs Test

	temp
Test Value[a]	202.9167
Cases < Test Value	16
Cases >= Test Value	20
Total Cases	36
Number of Runs	10
Z	-2.836
Asymp. Sig. (2-tailed)	.005

a. Mean

Our strategy for data analysis begins by **lagging**, as shown next, to create a new variable **temp_1** that is simply the immediately preceding value of **temp**[2]. Using the ***SPSS*** sequence **Transform/Compute…,** we define

temp_1= LAG(temp)

Here are the first few rows of the spreadsheet:

	temp	temp_1
1	200.00	.
2	202.00	200.00
3	208.00	202.00
4	204.00	208.00
5	204.00	204.00
6	207.00	204.00
7	207.00	207.00
8	204.00	207.00
9	202.00	204.00

Notice that the column for **temp_1** is the same as that for **temp** except that it is "pushed down" one row. Obviously, the first case of **temp_1** is missing because we do not know the lagged value of the first observation of **temp**. It is good to remember that **whenever you lag by k time periods the first k rows of the new lagged variable will be missing.**

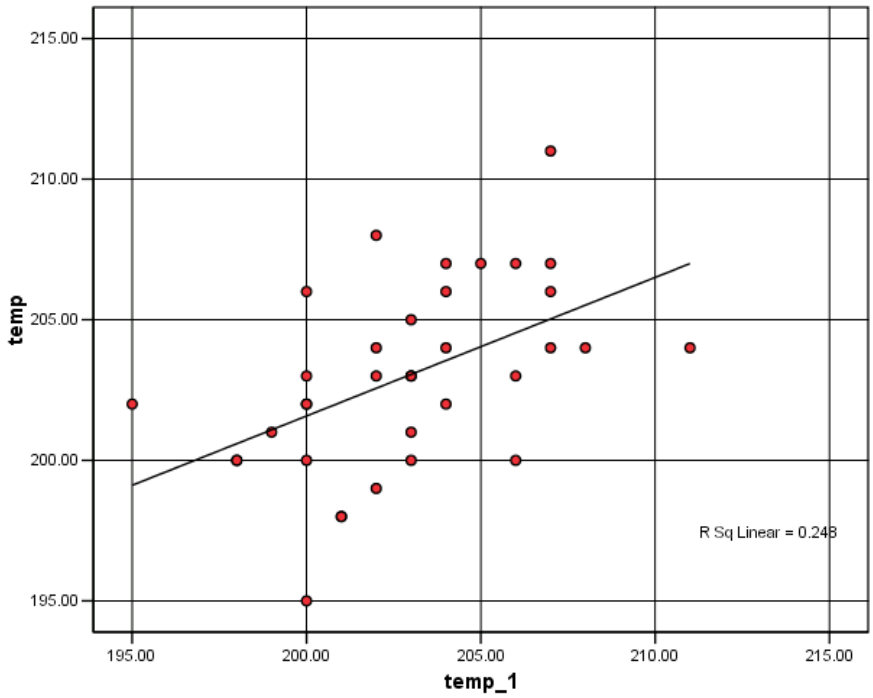
The key step in understanding "meandering" is to understand the scatter plot of **temp** on **temp_1**. "Meandering" means that successive pairs of points more often than not are either both below the mean or both above it. This tendency translates into a scatter plot that displays positive correlation:

[2]We would like to call the lagged variable "**temp-1**", using a minus sign, but ***SPSS*** will not allow the use of mathematical symbols in names. Thus in creating lagged variables (and we will be doing that frequently from now on) we shall indicate the lag with the underscore character "_".

Correlations

		temp	temp_1
temp	Pearson Correlation	1	.498**
	Sig. (2-tailed)	.	.002
	N	36	35
temp_1	Pearson Correlation	.498**	1
	Sig. (2-tailed)	.002	.
	N	35	35

**. Correlation is significant at the 0.01 level



The strategy for data analysis follows from the scatter plot: to help to explain the variation of **temp**, we regress **temp** on **temp_1**, the lagged value of temperature. To get an idea of the nature and significance of the lagged effect, we use the procedure **Analyze/Regression/Linear…**, making sure that we save the residuals for further analysis:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.498[a]	.248	.225	2.92774

a. Predictors: (Constant), temp_1

b. Dependent Variable: temp

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	103.023	30.334		3.396	.002
	temp_1	.493	.149	.498	3.296	.002

a. Dependent Variable: temp

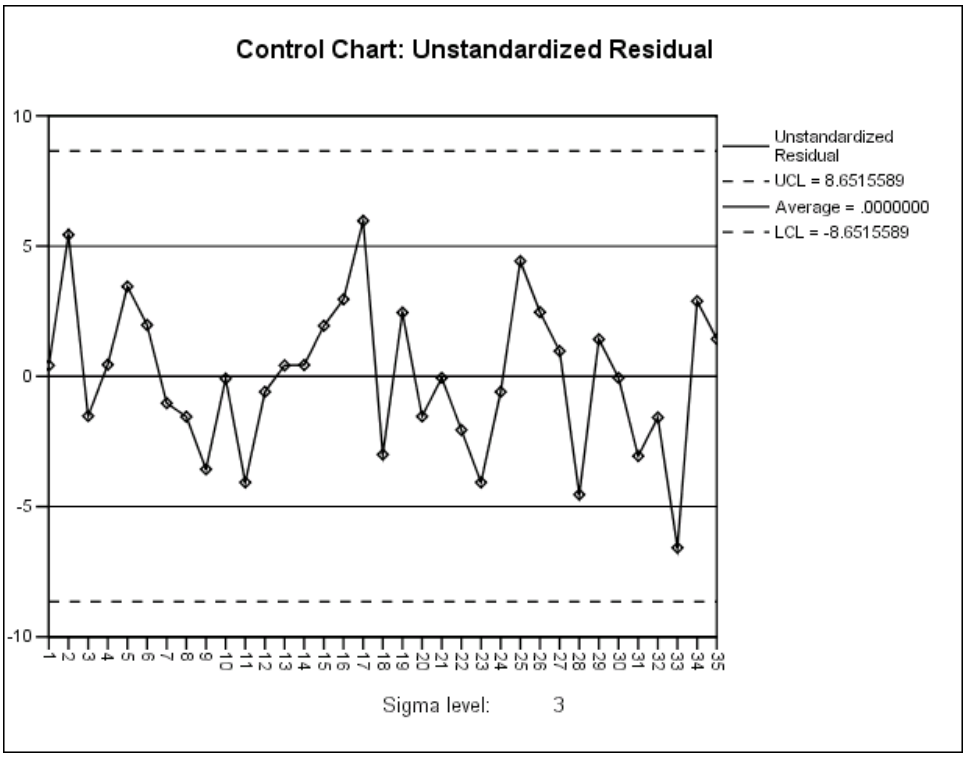
The autoregression is clearly significant. The t-ratio for **temp_1** is 3.30, and the Std. Error of the Estimate (residual standard error) is 2.928, substantially lower than the standard deviation, 3.316, for the original data.

The regression equation is

predicted temp = 103 + 0.493 temp_1 ,

which is the equation of the line plotted in the scatter diagram above.

Next, we must plot a control chart and do a runs test for the saved residuals, **RES_1**:



Runs Test

	Unstandardized Residual
Test Value[a]	.0000000
Cases < Test Value	18
Cases >= Test Value	17
Total Cases	35
Number of Runs	13
Z	-1.712
Asymp. Sig. (2-tailed)	.087

a. Mean

You should study the above output carefully. The residuals clearly do not meander; they are not obviously out of control. You might, however, be concerned that the number of runs, although not significant at the 0.05 level, is rather close with p = 0.087.

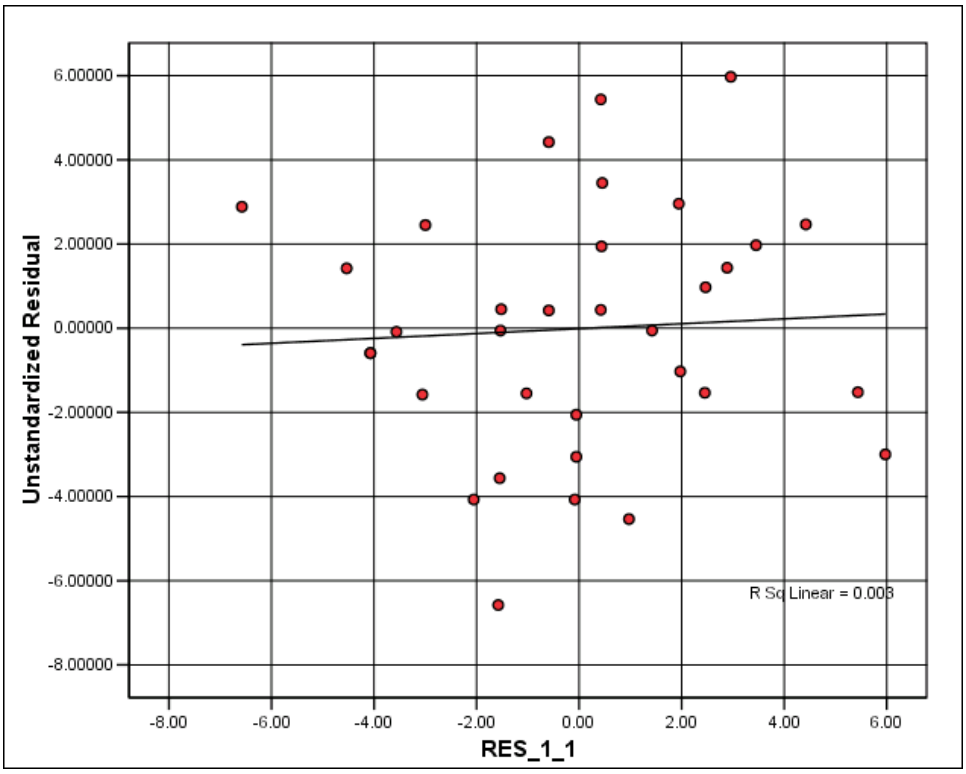
A new diagnostic check will give further understanding of what is going on. We begin by looking at the correlation between residuals separated by one time period, that is, between **RES_1** and **RES_1_1**.[3] This is called the **autocorrelation coefficient** of **RES_1**. For comparison with the result below, recall that the correlation coefficient between **temp** and **temp_1** -- **the autocorrelation coefficient of temp --** was about **0.50.**

First we must perform the transformation **RES_1_1 = LAG(RESIDU)**. Then we examine the correlation between **RES_1** and **RES_1_1** and check the scatter plot.

Correlations

		Unstandardized Residual	RES_1_1
Unstandardized Residual	Pearson Correlation	1	.058
	Sig. (2-tailed)	.	.745
	N	35	34
RES_1_1	Pearson Correlation	.058	1
	Sig. (2-tailed)	.745	.
	N	34	34

[3] This notation is rather unfortunate because ***SPSS*** automatically assigns the name **RES_1** to the **unlagged** residuals. Since, however, we will not be explicitly naming lagged residuals after this illustrative example, we will continue to use the underscore symbol in designating variables as lagged, hoping that the reader can wade through the somewhat confusing notation.



If the residuals from regression are in a state of statistical control and the sample size were very large, the correlation of **RES_1** and **RES_1_1** should be virtually zero. We see that for these 35 residuals, the correlation is 0.058, close to zero. A rough rule-of-thumb for judging significance of the departure from zero of any correlation coefficient is the following. If the correlation coefficient is computed from **n** pairs of observations, the standard error for the sample coefficient under the assumption of zero true correlation is

$$\frac{1}{\sqrt{n}}$$

Since there are 36-2 = 34 pairs of residuals, the standard error is 1 over $\sqrt{34}$ = 0.171. Hence the t-ratio for judging significance is

$$(0.058-0)/0.171 = 0.34$$

which is far less than the conventional requirement of 2 for significance.

The conclusion is that **RES_1** and **RES_1_1** are **not** significantly correlated, which is consistent with the assumption that the residuals from regression are in a state of statistical control. This new evidence reduces any suspicion raised by the significance level of the runs count.

Additional information can be obtained by looking at the correlation between **RES_1** and **RES_1_2**, that is, the correlation between the current **RES_1** and its lagged value **two periods back**. (We cannot infer this new correlation from what we have already done.) If the residuals from regression are in state of statistical control and the sample were large, this correlation, too,

should be virtually zero. We shall explore in a moment whether the correlation of residuals at lag two is significantly different from zero.

Here is the terminology that we shall be using:

- The correlation between **RES_1** and **RES_1_1**, which we have just seen to be 0.058, is called the **first order autocorrelation coefficient** of residuals.

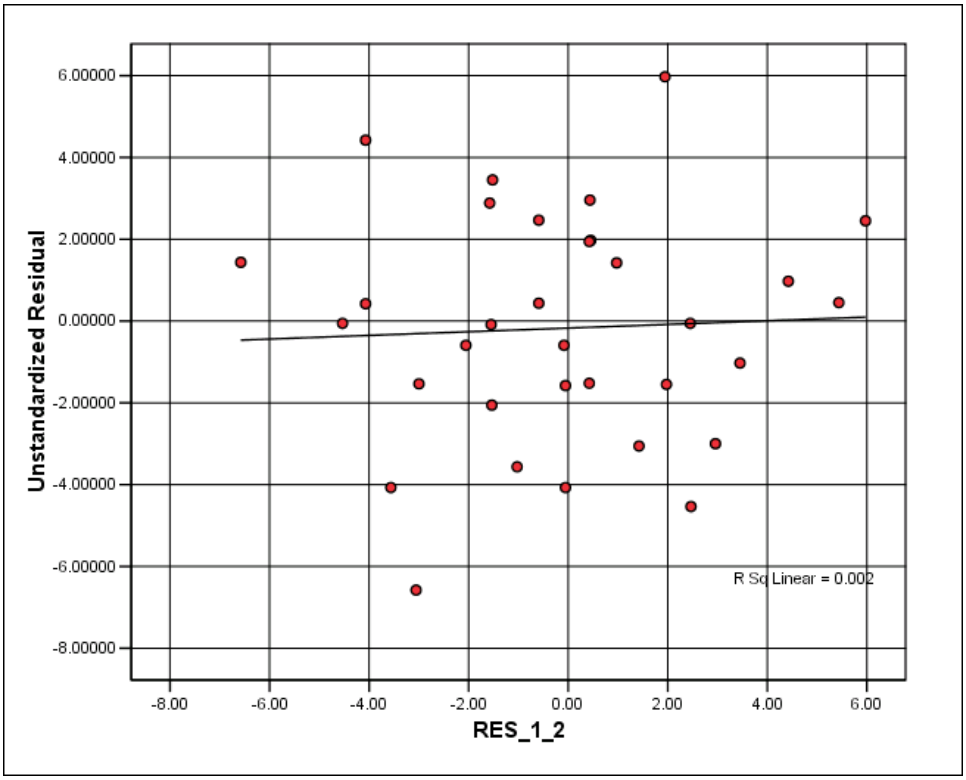
- The correlation between **RES_1** and **RES_1_2**, which we are about to look at, is called the **second order autocorrelation coefficient** of residuals.

After defining **RES_1_2= LAG(RES_1, 2)** in the **Transform** procedure, we obtain the following[4]:

Correlations

		Unstandardized Residual	RES_1_2
Unstandardized Residual	Pearson Correlation	1	.046
	Sig. (2-tailed)	.	.798
	N	35	33
RES_1_2	Pearson Correlation	.046	1
	Sig. (2-tailed)	.798	.
	N	33	33

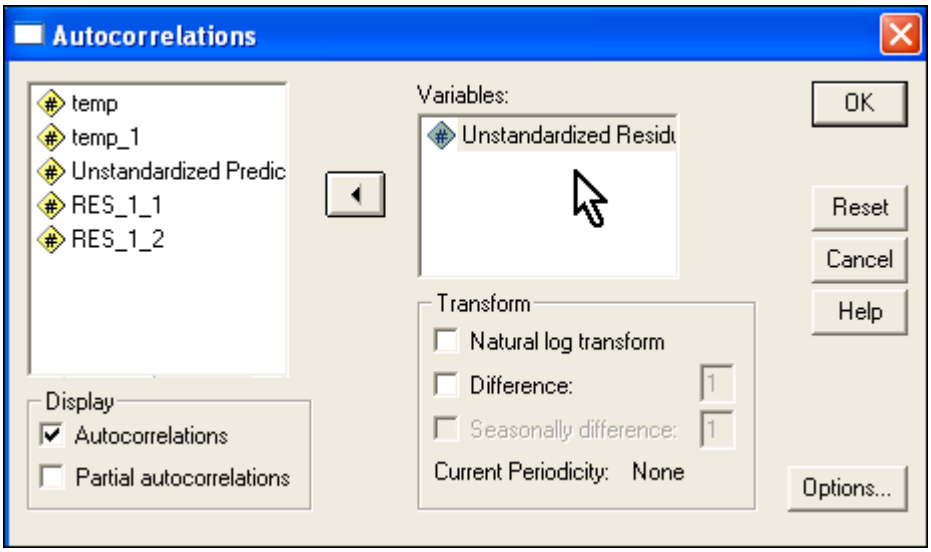
[4]Alternatively, we could have executed **RES_1_2 = LAG(RES_1)**. Do you see why?



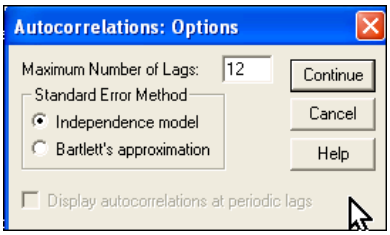
Again the autocorrelation coefficient, this time for lag 2, is close to zero, further supporting the assumption that the residuals are in statistical control. Once you get the idea of what we are doing, you will see that we could continue this line of analysis. We could look, for example, at the third autocorrelation coefficient of residuals, the correlation between **RES_1** and **RES_1_3**.

SPSS, however, provides us a shortcut. We can use the sequence **Graphs/Time Series/Autocorrelations…** to get a whole set of autocorrelation coefficients, one for each lag up to some maximum. This is what we shall normally do from here on. The explicit computations shown above for the first and second autocorrelation coefficients served only to explain the meaning of autocorrelation coefficients at varying lags. **Autocorrelations…** will give the results we need for the first 12 autocorrelation coefficients.[5] Here is the dialog window:

[5] Note: if you compute autocorrelations by directly lagging residuals, as we did above for lags 1 and 2, you will often get slight differences from those printed out in the **Autocorrelations…**graph. This stems from the fact that the correct computational formulas to compute autocorrelations differ slightly from those used in obtaining the Pearson correlation coefficients that we have seen earlier. The differences are not important for our purposes.



Clicking on the **Options…** button opens this new window:



When you open it you will notice that the default value for the maximum number of lags is 16. We have changed the maximum to 12 to reduce the size of the output.

After we click on **Continue** followed by **OK** we get the following display:

```
MODEL:  MOD_1.

Variable:  RES_1     Missing cases:  1     Valid cases:  35
□


Autocorrelations:   RES_1   Unstandardized Residual

     Auto- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1   Box-Ljung  Prob.
                  |----|----|----|----|----|----|----|----|
  1   .058   .162                .     |*    .                    .126   .722
  2   .046   .160                .     |*    .                    .208   .901
  3  -.103   .157                .   **|     .                    .641   .887
  4  -.062   .155                .    *|     .                    .800   .938
  5  -.005   .152                .     *     .                    .801   .977
  6  -.361   .150               *.*****|     .                   6.626   .357
  7  -.294   .147                ******|     .                  10.626   .156
  8  -.070   .144                .    *|     .                  10.861   .210
  9   .104   .142                .     |**   .                  11.401   .249
 10   .248   .139                .     |*****.                  14.594   .148
 11   .083   .136                 .    |**   .                  14.962   .184
 12   .058   .133                 .    |*    .                  15.151   .233

Plot Symbols:      Autocorrelations *     Two Standard Error Limits .

Total cases:  36     Computable first lags:  34
```

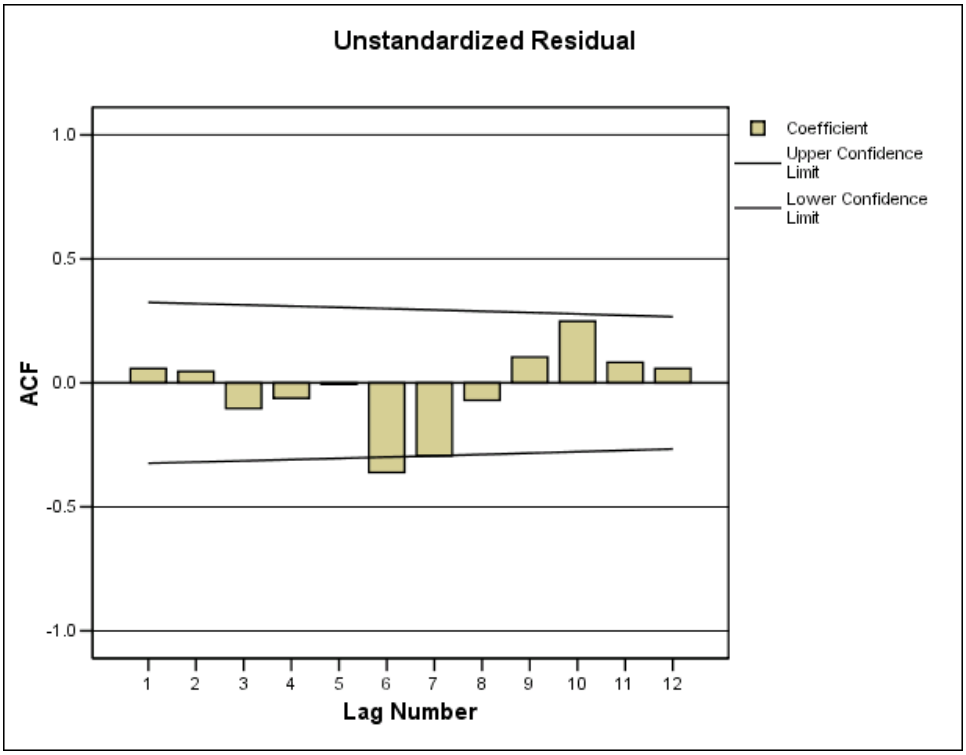
Here is how the output above is interpreted. On the first line, starting with "1", we read the first-order autocorrelation coefficient 0.058, which we have already seen above. We see also its graphical location on the scale running across the top of the plot.

On the second line we see "2 0.046", which indicates the autocorrelation of order two and the corresponding computation above. As we go down the plot, we see the 3rd, 4th, etc., autocorrelations, up to the 12th autocorrelation, which is 0.058.

The dot marks that enclose each bar in the graph above indicate approximate two-standard-error limits if the the true autocorrelation coefficient were zero. Thus they lie a distance above and below zero that is about equal to $2/\sqrt{n}$, where n is the number of paired observations available for calculating the particular autocorrelation coefficient.

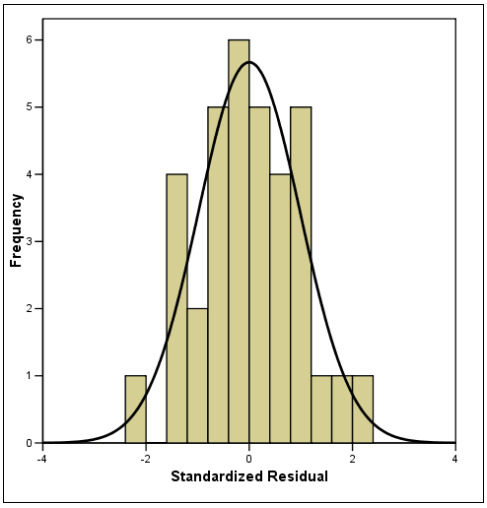
We see that one autocorrelation coefficient, -0.361 at lag 6, is outside the two-standard-error warning limits on the plot above. But recall that roughly 1 in 20 would be expected by chance alone. Hence the evidence provided by **Autocorrelations…** is reasonably consistent with the assumption that the residuals from the autoregression of **temp** on **temp_1** are in a state of statistical control.

Besides this numerical computation, there is a visual check that can be used with **ACF**[6]. It stems from the fact that, if the process is in control, the **ACF plot** -- turned on its side -- should itself look like a control chart for an in-control process. The second display in the output, shown below, enables us to perform this visual check:



[6] Henceforth we shall from time to time refer to the display as the **ACF plot**, where **ACF** stands for **autocorrelation function**.

We now look briefly at the remaining diagnostic checks, which are satisfactory:

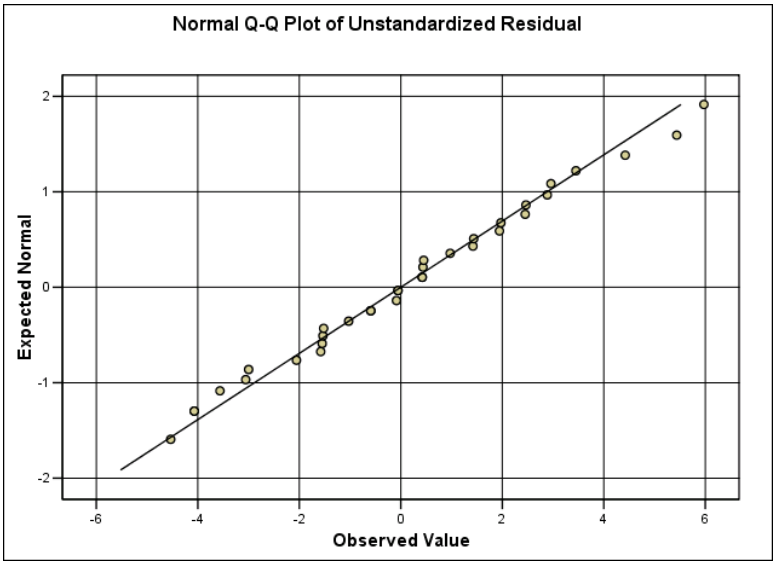


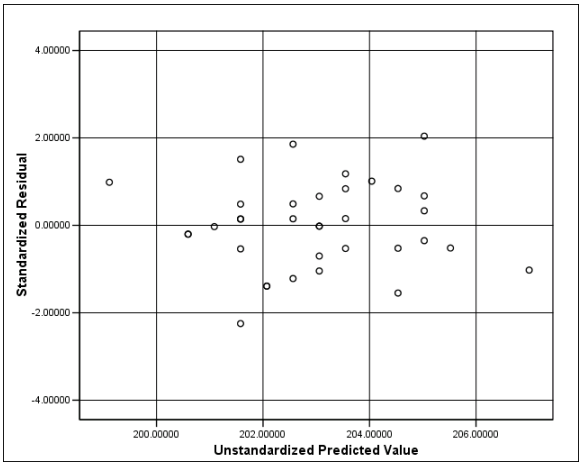
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	.066	35	.200*	.991	35	.992

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction





It would therefore appear that a simple autoregressive scheme, in which the fitted value of each observation is obtained by multiplying the preceding observation by about 0.5 and adding a constant 103 (see the output above), explains the autocorrelated behavior of **temp**. This means that any observations in the control chart for **temp** that appear to be out of control can be interpreted **not as the result of special causes but as the result of the simple autoregression relationship that is operating throughout the data set**. By capturing this relationship in our autoregression model, we see that the

reactor process is **stable**, by which is meant that the residuals from a regression model that does not include trend terms appear to be in a state of statistical control.[7]

- The autoregression model also provides a tool for predicting and monitoring continued observations on the process. For example:

- The autoregression equation can be used to predict the next temperature reading, given the one we have just observed:

predicted = 103 + 0.493(current temp).

The standard deviation of residuals, 2.928, provides a rough guide for deciding whether prediction errors (**actual - predicted**) reflect special causes. For example,

predicted ± 3*2.928 = predicted ± 8.8

gives approximate upper and lower control limits for the next actual observation.[8]

Have We Overlooked Anything?

Although we have done a pretty thorough job of diagnostic testing and found no serious blemishes on the simple autoregressive model, we may still be uneasy about the possibility that some important variables could have been left out. For example:

- Could lags of **temp** higher than 1 improve the fit?

- Could we be overlooking a time trend in? (If you stare hard at the original control chart for **temp**, you may see a faint hint of a downward trend.)

Rather than remain uneasy, we can explore these possibilities by using a regression that tries out, say, not only **temp_1** but also **temp_2** and **time** as independent variables. This is shown below. The added variables prove to be **insignificant**, thus increasing our confidence in the analysis above.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	126.217	38.135		3.310	.002
	temp_1	.511	.183	.511	2.796	.009
	temp_2	-.127	.178	-.129	-.715	.480
	time	-.058	.055	-.171	-1.050	.302

a. Dependent Variable: temp

[7] Another term for "**stable**" that we may use from time to time is "**stationary**", which conveys the idea of the mean and standard deviation "holding still and not shifting".

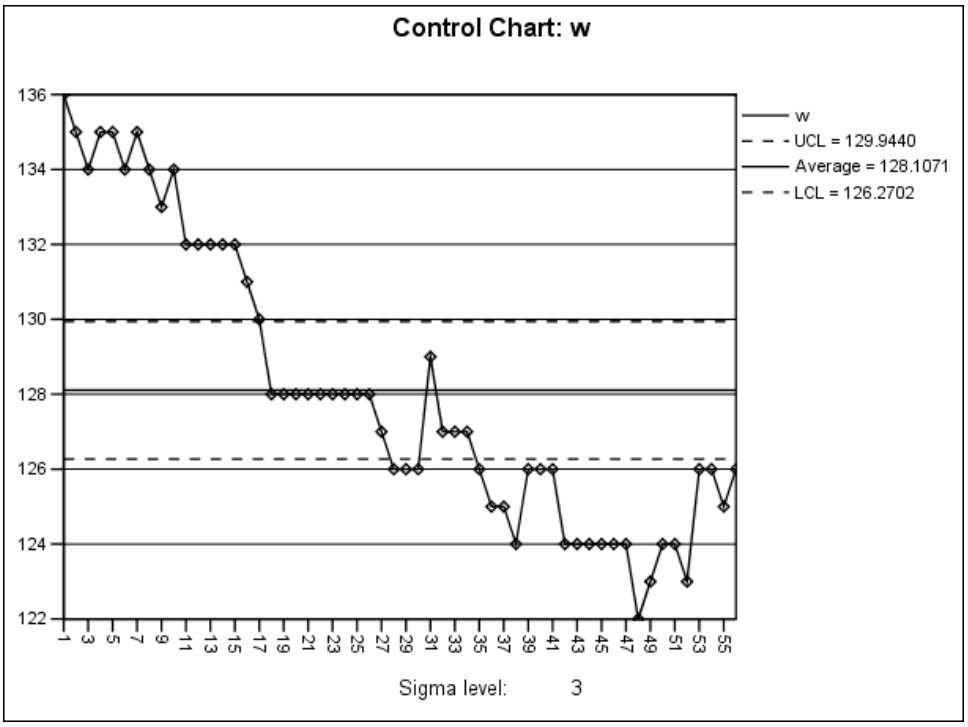
[8] If you obtain the exact confidence limits for predicting a new individual observation, you will see that this calculation underestimates the interval width slightly, but not to a serious extent.

3. A More Complicated Application of Autoregression: History of a Diet

We next use a simple personal quality improvement project, a diet designed by an MBA student to lose weight during the fall quarter at Chicago during the mid-1980s. The role of statistics here is to check on what kind of progress has been made on the key output variable, weight (**w**). Information was also kept on calories consumed (**c**) in order to gain a better understanding of how the diet was going. The data to be studied are contained in the file DIET.sav. The values for weight and calories consumed are daily observations, starting on Thursday, September 28th and ending on the day before Thanksgiving—yielding a set of 56 cases in all.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
w	56	122.00	136.00	128.1071	3.86913
c	56	750.00	6000.00	1573.4821	886.62225
Valid N (listwise)	56				



There is no doubt that the time series is terribly out of control. Note the apparent downward trend, which is possibly nonlinear. There is also a strong possibility of autocorrelation.

Here we interrupt the analysis to mention a mathematical point that is applicable to our next step. The equation of a second-degree polynomial, or parabola, is

$$y = a + bx + cx^2$$

Gently curving trends can often be approximated by parabolas, where **x** is interpreted as the variable we have called **time**. Hence we regress **y** on **two** independent variables, **time** and the **square of time**, which we name **tsq**:[9]

We prepare for the next stage of the analysis by using **Transform/Compute** to define

time = $CASENUM and
tsq = time*time

Then we apply **Stepwise Regression** to **w** as the dependent variable, with **time** and **tsq** as the candidates for the right-hand-side of the equation.

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.920[a]	.847	.844	1.52781
2	.966[b]	.932	.930	1.02597

a. Predictors: (Constant), time

b. Predictors: (Constant), time, tsq

c. Dependent Variable: w

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	134.329	.414		324.579	.000
	time	-.218	.013	-.920	-17.284	.000
2	(Constant)	136.972	.426		321.195	.000
	time	-.492	.035	-2.073	-14.244	.000
	tsq	.005	.001	1.189	8.170	.000

a. Dependent Variable: w

We see that there is a significant, nonlinear downward trend with highly significant t-ratios for both **time** and **tsq**. The original standard deviation of **w** was 3.869. The standard deviation of the residuals is only 1.026 – a considerable reduction in uncertainty.

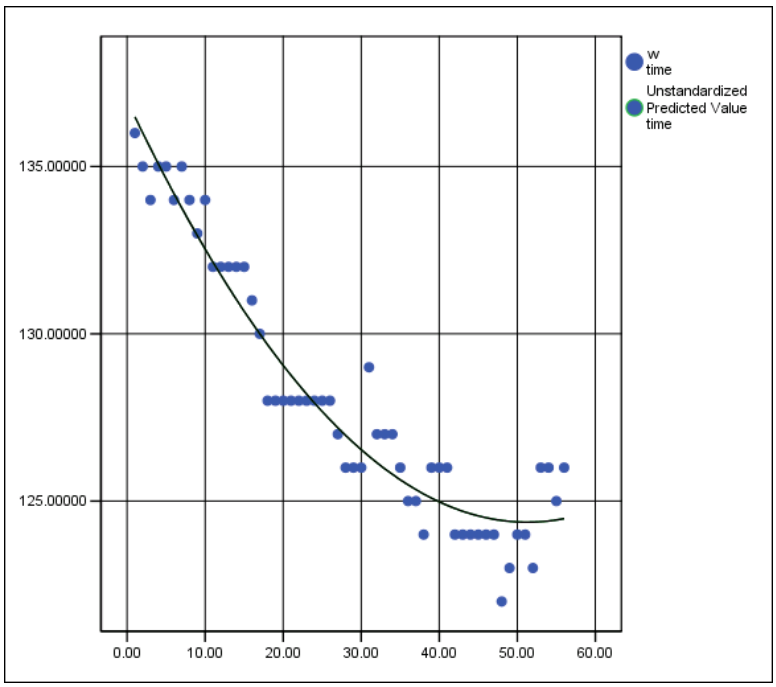
The regression model is

predicted w = 136.972 - 0.492 time + 0.005 tsq

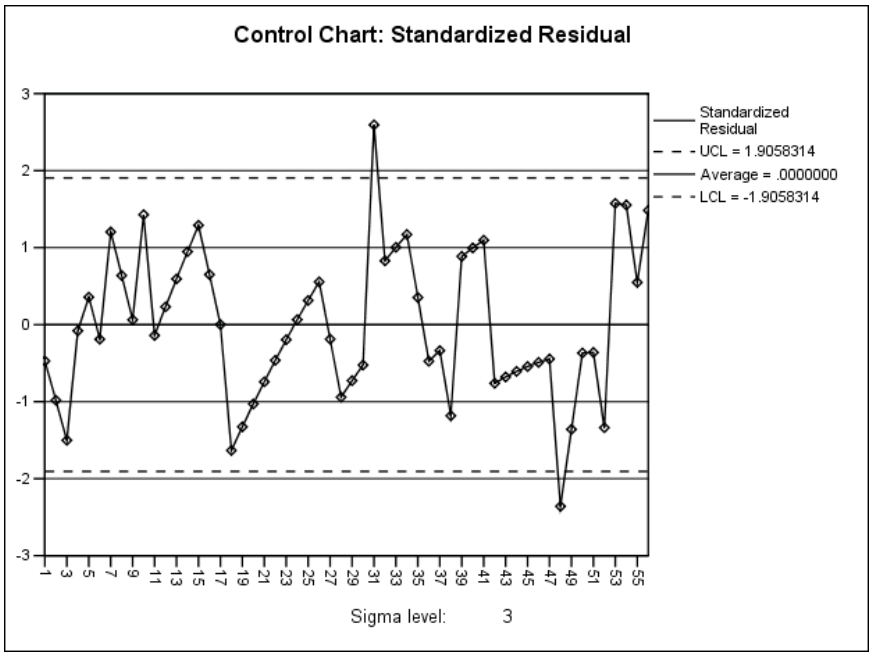
[9] An alternative mathematical representation that can be useful when the curving trend flattens out to become horizontal involves **1/x**, the reciprocal of **x**:

y = a' + b'x + c'(1/x)

We have saved the residuals and the predicted values, but before we look at the residuals, we examine the nature of the trend by means of a scatter plot (overlay). In the image below the predicted values are shown by the curved line:



It is apparent that weight loss was rapid at first, but then it slowed and actually seemed to stop at the end. The trend is clearly nonlinear. This information is important, but the diagnostic checks will show that there is more to be learned from the data than the existence of a nonlinear trend alone.



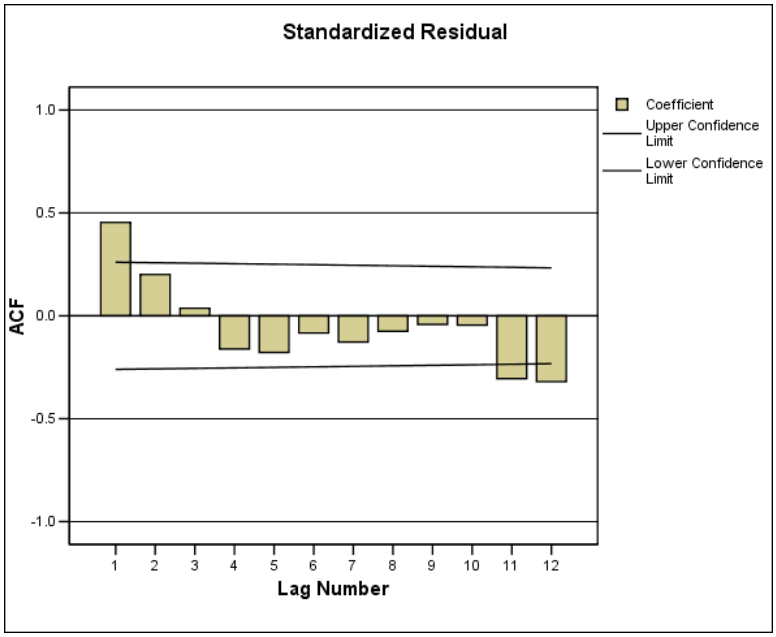
There are clear indications in the above control chart that the residuals are not in control. Let's take a look with the autocorrelation plot:

Autocorrelations: ZRE_1 Standardized Residual

Lag	Auto-Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	.454	.130	. ****.****	12.146	.000
2	.200	.129	. ****.	14.560	.001
3	.037	.128	. * .	14.643	.002
4	-.161	.127	. *** .	16.272	.003
5	-.178	.125	.**** .	18.292	.003
6	-.084	.124	. ** .	18.745	.005
7	-.128	.123	. *** .	19.824	.006
8	-.075	.122	. ** .	20.205	.010
9	-.042	.120	. * .	20.327	.016
10	-.045	.119	. * .	20.471	.025
11	-.305	.118	*.**** .	27.205	.004
12	-.320	.116	*.**** .	34.763	.001

Plot Symbols: Autocorrelations * Two Standard Error Limits .

Total cases: 56 Computable first lags: 55



It is apparent from the autocorrelations that the residuals are far from a state of statistical control:

- 3 of 12 coefficients are outside the two-standard-error limits; the first autocorrelation coefficient 0.454 is 3.5 times its standard error, 0.130.

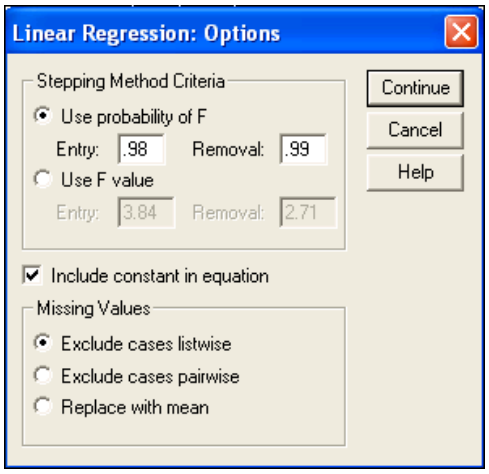
- If you turn the plot on its side and interpret it as a time series, you will see only two runs of +'s and -'s and an overall appearance of nonlinear trend.

There is obvious room for improvement of the trend model. We consider lags of **w** as additional independent variables. Since the first autocorrelation coefficient of residuals is 0.454, you would naturally think of trying the first lagged variable, **w_1**. This choice will actually work here, but in general,

it is hard for a nonspecialist in statistical time series analysis to figure out the useful **lags of w** solely from the **autocorrelation function of residuals** from an earlier model.[10]

We suggest an alternative strategy: start by introducing into the previous regression model, the lowest or first lag, **w_1**. If **w_1** makes a significant contribution to the model, we'll check residuals. **If there is evidence of remaining autocorrelation of residuals, we can then try higher lags, such as w_2 and w_3.** It turns out in the current application that **w_1** does make a significant contribution, and that higher lags are not needed:

After creating the new variable **w_1 = LAG(w)**, we apply **Stepwise Regression**, clicking on **Options…** to set the entry probability of F at 0.98 and the removal setting at 0.99. (See page 5-15 if you have forgotten why we do this.)



Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.963[a]	.927	.925	1.02655
2	.964[b]	.930	.927	1.01440
3	.971[c]	.943	.940	.92253

a. Predictors: (Constant), w_1

b. Predictors: (Constant), w_1, time

c. Predictors: (Constant), w_1, time, tsq

d. Dependent Variable: w

[10]There is a potential trap in looking at the autocorrelation function alone, noting that the autocorrelations of residuals are significant at lags 1, 11, and 12, and then trying **w_1**, **w_11**, and **w_12**. Here this would cause loss of data (from lagging up to 12), and would lead to an unnecessarily complicated model. The route actually followed in the text leads to a much simpler model, as we shall see.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	9.137	4.599		1.987	.052
	w_1	.927	.036	.963	25.848	.000
2	(Constant)	27.595	13.048		2.115	.039
	w_1	.791	.097	.821	8.171	.000
	time	-.036	.024	-.152	-1.509	.137
3	(Constant)	71.448	17.401		4.106	.000
	w_1	.478	.127	.496	3.772	.000
	time	-.265	.070	-1.132	-3.788	.000
	tsq	.003	.001	.698	3.446	.001

a. Dependent Variable: w

We see that **w_1** was the most powerful "explainer of the variance" of **w** and thus entered first. It is interesting that at the second step, of the two remaining variables, **time** and **tsq**, neither by itself would have a |t-ratio| greater than 2. Thus, if we had not relaxed the entrance requirement, the stepwise regression would have come to a halt. After **time**, however, enters the model, the situation is changed, and in the final equation, all three of the independent variables have t-ratios greater than 2 in absolute value.[11]

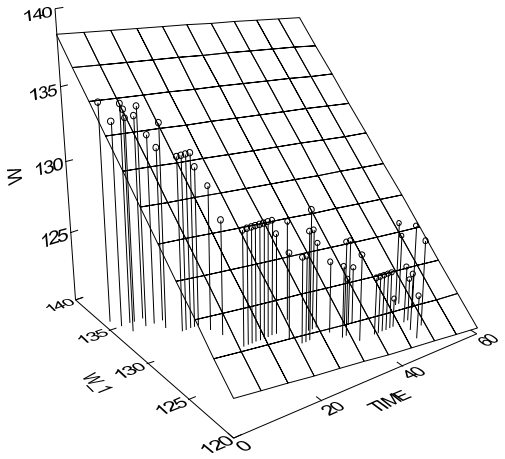
The regression equation is now

predicted w = 71.448 -0.265 time + 0.003 tsq +0.478 w_1

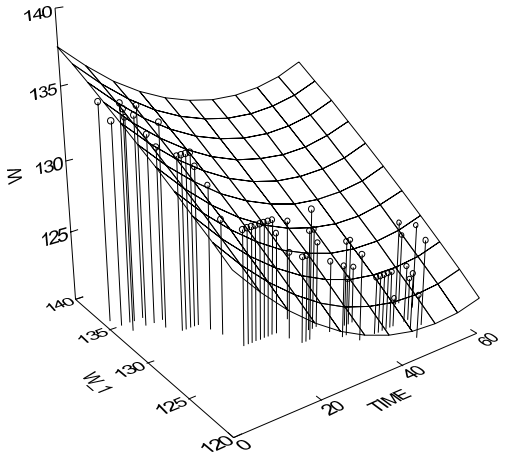
Note also that the standard deviation of residuals is now down to 0.92253.

Of course, we cannot present a four-dimensional graph so that you can see the regression plane (technically called a "hyperplane"), but it is instructive to look at a 3-D plot of the data when **tsq** is omitted. In both of the following plots **time** is on the X-axis, running along the front edge, and **w_1** is on the Y-axis. We have used the spikes to bring out the obvious curvilinear relationship between the two independent variables.

[11]This example shows clearly that it is not always easy to anticipate the best fitting model from the steps that have gone before. Another risk with stepwise regression is that because it is a "hierarchical method", to some degree committing itself to a chosen path, it can occasionally miss a best fitting model entirely.

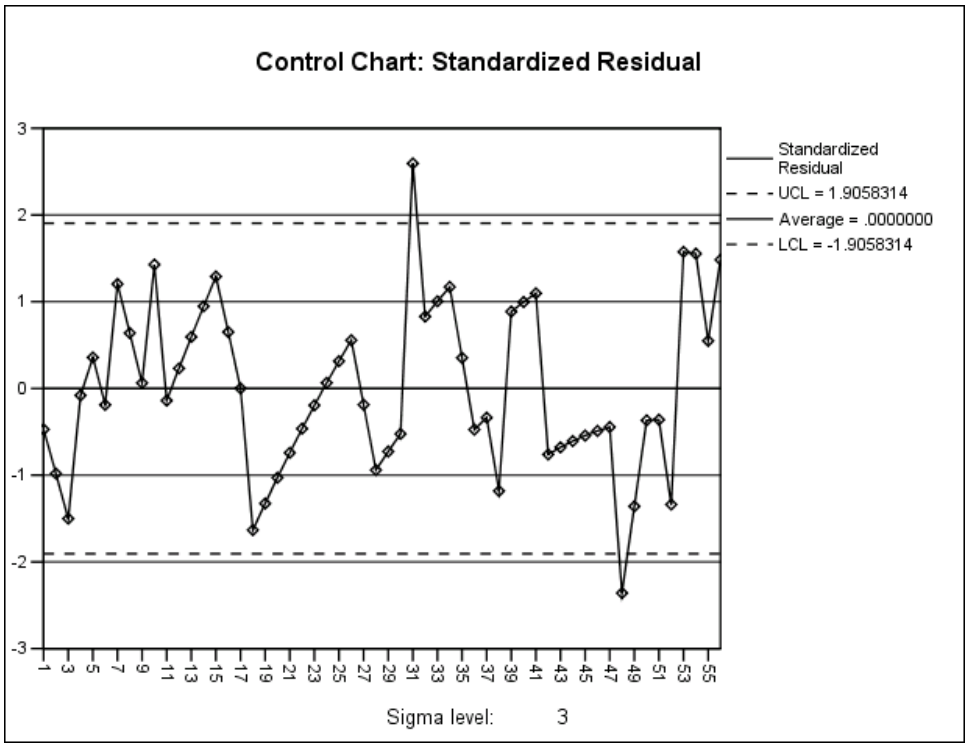


The most important point to note is that when a linear regression is fitted without **tsq** (see the figure above), it looks like **time** has very little effect, corresponding to its low t-ratio at Step 2 in the stepwise regression output above. When we smooth the points (tops of the spikes), however, with a quadratic fit, there is a distinct bend to the surface, indicating that **tsq** is an important variable.[12]



We shall now look at the diagnostics for the residuals from the regression:

[12]Our example illustrates a point that is often a bit mysterious to the uninitiated. "How can a parabolic curve (quadratic function) , which is nonlinear, be fitted with linear regression?" The answer is that **it is nonlinear** when depicted in 3-D, but in a higher dimension, in our example, 4-D, it becomes linear because we treat **time** and its square, **tsq**, as distinct independent variables. By the way, the two graphs shown here were beyond the capability of the student version of ***SPSS.*** As on page 4-52, the plotting routine used was from ***SYSTAT for Windows*®, Version 6.**



The large positive residual at observation 31 suggests the possibility of a special cause. Our subsequent analysis, which will bring **c** (calories consumed) into the picture, will explain what this special cause is. The remaining diagnostics, shown below, suggest only minor problems with this three-variable regression model. We hope that even these will be alleviated when we bring **c** into the picture.

Runs Test

	Unstandardized Residual
Test Value[a]	.0000000
Cases < Test Value	31
Cases >= Test Value	24
Total Cases	55
Number of Runs	22
Z	-1.676
Asymp. Sig. (2-tailed)	.094

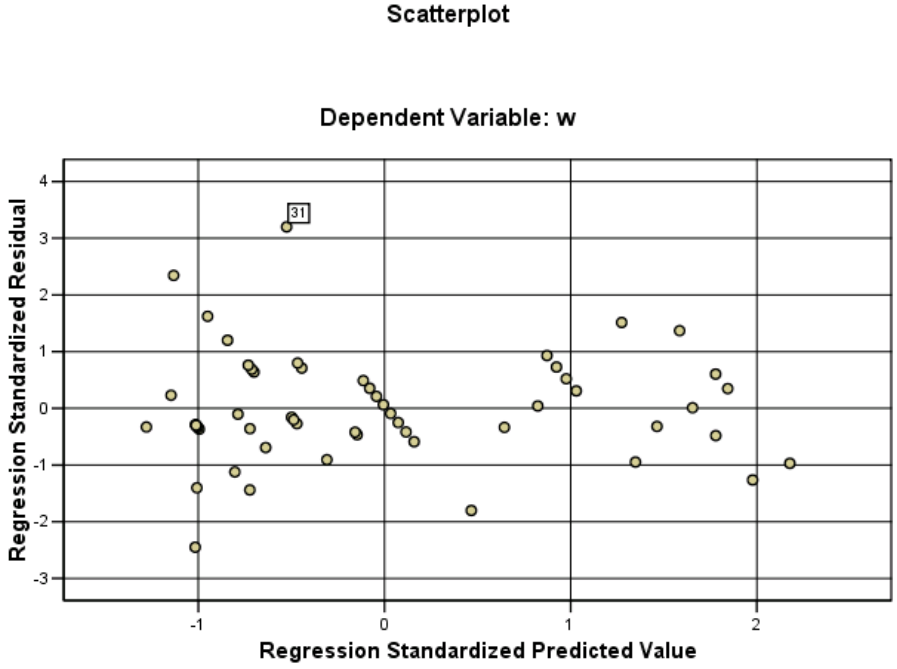
a. Mean

Autocorrelations: RES_2 Unstandardized Residual

Lag	Auto-Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	-.011	.131	. * .	.007	.933
2	-.006	.130	. * .	.010	.995
3	.053	.129	. * .	.178	.981
4	-.118	.128	. ** .	1.040	.904
5	-.153	.126	. *** .	2.503	.776
6	.073	.125	. * .	2.844	.828
7	-.091	.124	. ** .	3.386	.847
8	-.015	.122	. * .	3.401	.907
9	-.010	.121	. * .	3.408	.946
10	.129	.120	. *** .	4.563	.918
11	-.258	.118	***** .	9.324	.592
12	-.130	.117	. *** .	10.557	.567

Plot Symbols: Autocorrelations * Two Standard Error Limits .

Total cases: 56 Computable first lags: 54

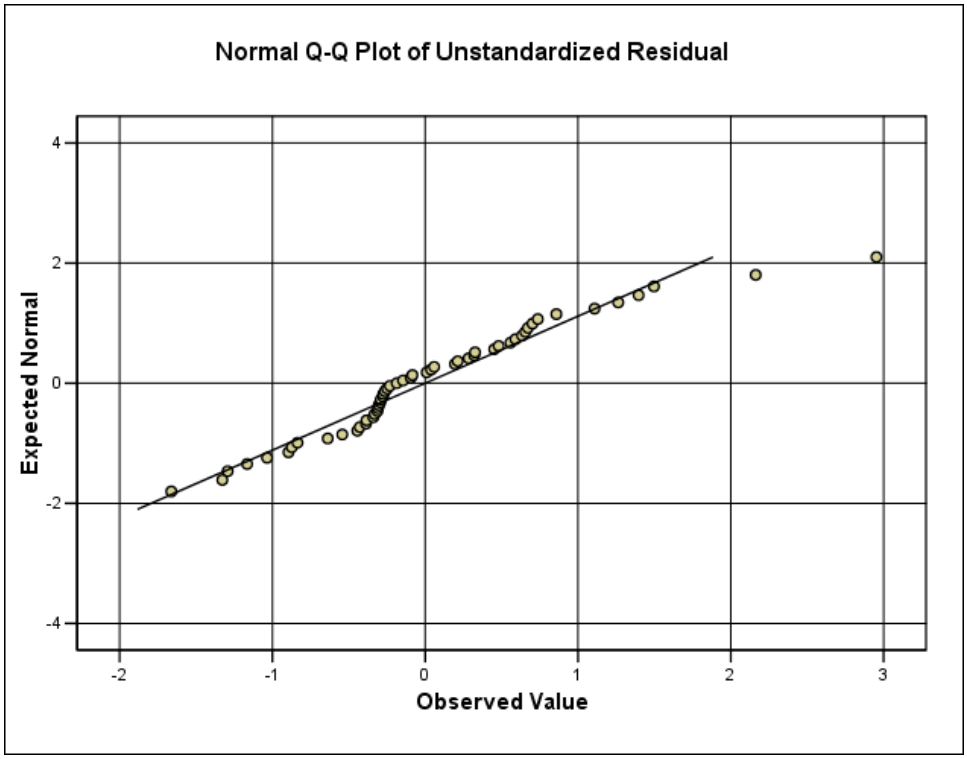


← Ask your instructor how he or she was able to label the outlier as Case 31.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	.110	55	.094	.960	55	.062

a. Lilliefors Significance Correction



Aside from the outlying residual at observation 31, the diagnostic checks are satisfactory.

There is, however, a loose end. We considered only the first lag of **w**, **w_1**. Do we need also to look at higher lags? The autocorrelation analysis as applied to residuals, however, suggests that this would not be necessary, the residuals show no sign of significant autocorrelations. Had the autocorrelation function provided evidence of significant autocorrelations of residuals, we would have needed to go back to consider additional lags of **w**, say **w_2** or maybe **w_2** and **w_3** in addition to **w_1**. (We suggest always working from lower to higher back consecutively rather than trying to guess at isolated "hot lags" where there are large autocorrelation coefficients of residuals.)

An alternative strategy would have been to try both **w_1** and **w_2**, along with trend variables, in the first place. In fact, this would have shown that **w_2** did not contribute significantly to the fit, and we would have been back to the model actually reached above.

This illustrates a general strategy choice:

- Include **selected promising variables**, fit the regression, and then see if further promising variables will improve the fit. This leads to several steps of model building.

or

- Start with **all promising variables**, and simplify by dropping out variables that do not contribute to the fit.

Ordinarily, the two strategies will lead to the same result. The second strategy is faster if you know all the promising variables at the start. The first strategy is more intuitive and sometimes leads you to think of non-obvious possibilities.

In the above analysis, mainly for pedagogical reasons, we followed the first strategy. We started with only trend variables. They explained a good deal of the fit, but the diagnostic checks showed autocorrelated residuals, a problem that we remedied by introducing the first lag, **w_1**.

In the next section, we extend this strategy by introducing **c, caloric intake,** and lags of **c**. This splitting up of the model development was also mainly for pedagogical reasons: we started in this section by seeing how well we can do when we ignore the information provided by **c**. Next we see what improvement is attainable by use of **c**.[13]

What Lies Behind the Regression Model?

We have established the existence of a nonlinear trend and also an autoregressive relationship. What can we say about root causes of these relationships?

- The downward trend appears to be related to caloric reduction achieved by the diet. The reason for the nonlinearity of the trend, however, is not yet apparent. Further light will be shed on this question in Sections 4 and 5 below.

- The autoregressive relationship in addition to the trend effect has a possible explanation in terms of lagged effects in body fluid balance. Thus the effects of increased or decreased fluid intake (or outgo in exercising) can have carryover effects that affect weight measurements for some time into the future. That is why people who weigh themselves frequently are surprised at the extent of variation in measured weights. Thus they speak of "gaining five pounds as a result of large meal", or "losing six pounds in a strenuous workout". Such gains and losses are essentially short run fluctuations. One support for this position will be seen in the next section, when **c** is introduced into the picture. However, **w_1** still contributes to the fit, even when calories are introduced into the fit.

4. Introduction of Information on Calories (c) to Improve the Diet Regression Model

In studying the progress of the diet, we have thus far considered time trend and lags of **w**. The time trend turned out to be nonlinear (**time** and **tsq**) and the lagged effect of **w** was captured by **w_1**. The model thus reached at the end of Section 3 represented the best fit we knew how to obtain with the available information. This model represented a great advance in fit: the standard deviation of **w** was 3.87 pounds, while the standard deviation of residuals was only 0.93 pounds.

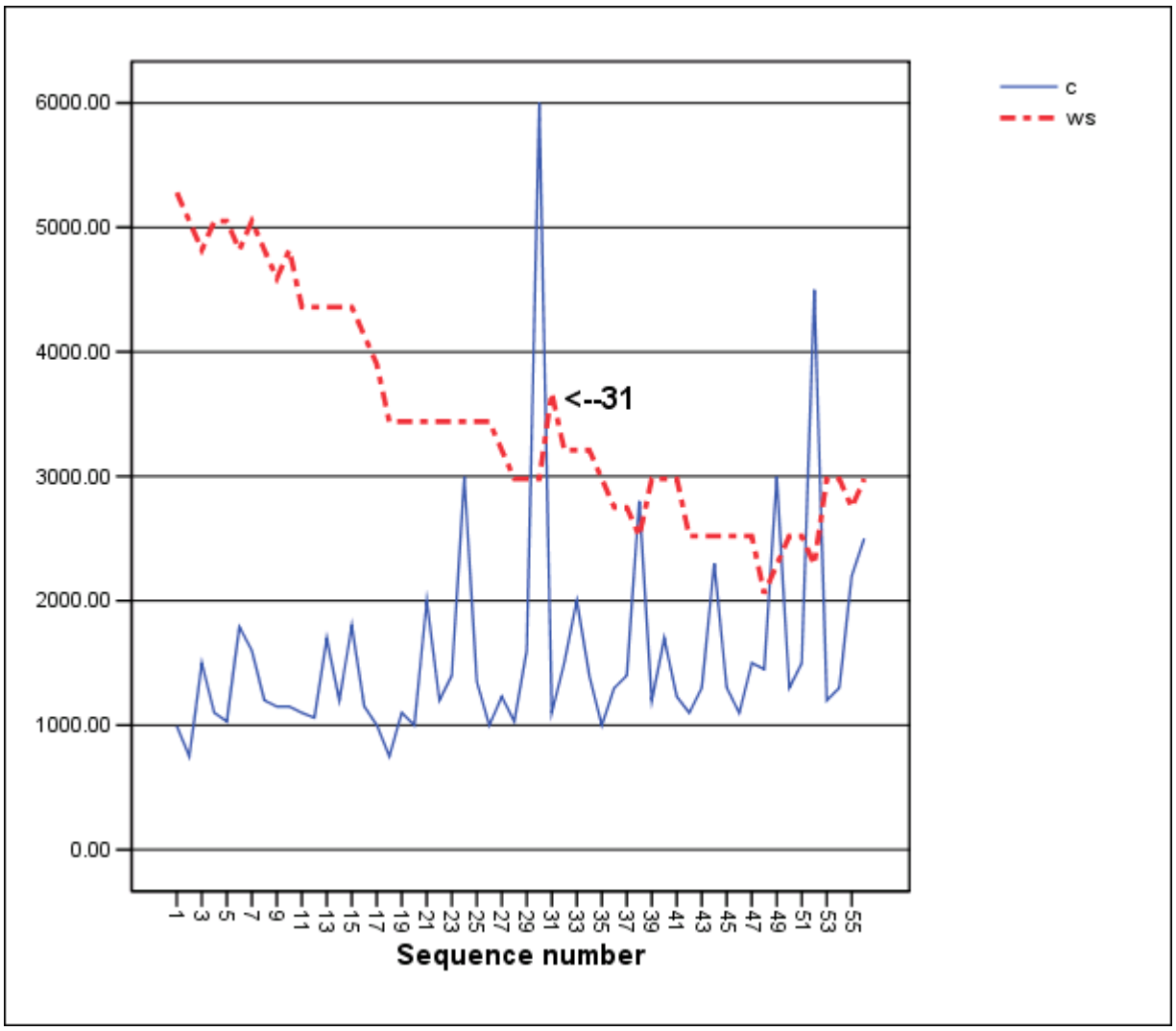
With additional information, however, it is possible that we could achieve a better fit and also add insight into what was happening during the diet. So now we bring calories consumed (**c**) into the picture. The strategy is just an extension of what we have been doing, and minimal commentary will be provided for the data analysis shown next.

Let's begin with a multiple sequence plot showing **w** and **c** on the same graph. To make the comparison meaningful we have temporarily rescaled **w** according to

ws = 230*w - 26000 .

[13]Once we see the completed analysis so that we understand the strategy, we can consolidate all the model building of the two sections into one step in which **all** promising variables are introduced into one large stepwise regression.

All that this does is to make the standard deviation of **w** about the same as **c** and to make the means of the two variables close enough so that we can see their comovement on the plot. (We will continue to work with the unscaled **w** after we are finished discussing the multiple time series plot.)



Careful examination of the time series plot shows a not unexpected relationship between **c** and **w** (rescaled): Most of the major upward jumps in weight appear to be **preceded** by an increase in caloric intake. We have labeled in particular the 31st observation of weight where there was a dramatic increase, but there are other examples as well of a jump in calories preceding an increase in weight.

Recall that observation 31 was an outlier in the control chart for residuals in our last regression.

We shall now include **c** and its one-period lagged value, **c_1**, along with the other independent variables in a new stepwise regression analysis:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	w_1	.	Stepwise (Criteria: Probability-of-F-to-enter <= .980, Probability-of-F-to-remove >= .990).
2	c_1	.	Stepwise (Criteria: Probability-of-F-to-enter <= .980, Probability-of-F-to-remove >= .990).
3	time	.	Stepwise (Criteria: Probability-of-F-to-enter <= .980, Probability-of-F-to-remove >= .990).
4	tsq	.	Stepwise (Criteria: Probability-of-F-to-enter <= .980, Probability-of-F-to-remove >= .990).
5	c	.	Stepwise (Criteria: Probability-of-F-to-enter <= .980, Probability-of-F-to-remove >= .990).

a. Dependent Variable: w

Model Summary[f]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.963[a]	.927	.925	1.02655
2	.979[b]	.958	.956	.78340
3	.980[c]	.960	.958	.76770
4	.985[d]	.970	.967	.67733
5	.985[e]	.970	.967	.68419

a. Predictors: (Constant), w_1

b. Predictors: (Constant), w_1, c_1

c. Predictors: (Constant), w_1, c_1, time

d. Predictors: (Constant), w_1, c_1, time, tsq

e. Predictors: (Constant), w_1, c_1, time, tsq, c

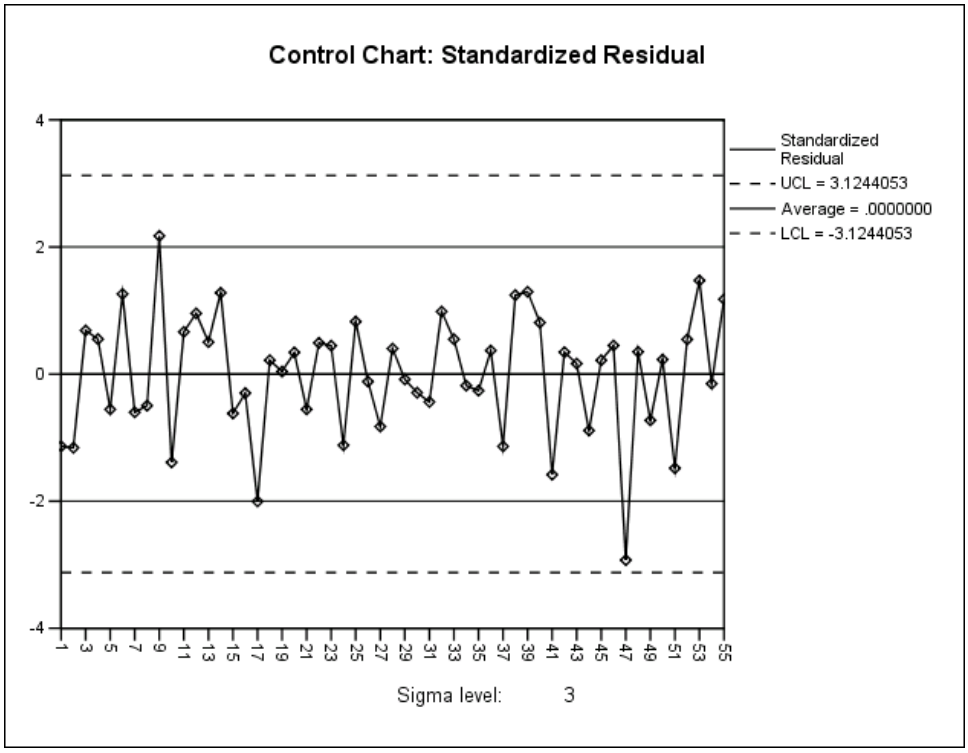
f. Dependent Variable: w

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.137	4.599		1.987	.052
	w_1	.927	.036	.963	25.848	.000
2	(Constant)	1.016	3.743		.271	.787
	w_1	.981	.029	1.018	34.184	.000
	c_1	.001	.000	.186	6.245	.000
3	(Constant)	17.529	10.003		1.752	.086
	w_1	.859	.074	.892	11.602	.000
	c_1	.001	.000	.184	6.308	.000
	time	-.032	.018	-.135	-1.774	.082
4	(Constant)	55.164	13.006		4.241	.000
	w_1	.591	.095	.613	6.249	.000
	c_1	.001	.000	.173	6.679	.000
	time	-.226	.052	-.964	-4.364	.000
	tsq	.002	.001	.589	3.939	.000
5	(Constant)	55.257	13.335		4.144	.000
	w_1	.590	.097	.612	6.099	.000
	c_1	.001	.000	.173	6.532	.000
	time	-.226	.053	-.965	-4.300	.000
	tsq	.002	.001	.590	3.878	.000
	c	.000	.000	-.001	-.041	.968

a. Dependent Variable: w

Note that at Step 5, although forced into the regression by the specification of "F to enter", **c**, the contemporaneous caloric intake, contributes nothing to the model. Thus we will rerun the regression analysis without using the stepwise procedure, specifying only the independent variables that are included in Step 4 above. That way we can obtain the correct residuals on which to perform our diagnostics, shown as follows:



The behavior of the residuals is much improved over the last regression. There are no reports of possible special causes although there is one very negative residual at time 48 that is close to the LCL. The outlier at time 31, however, is no longer present due to the relationship between **c_1** and **w** shown in the time series plot above. All that we can say about point 48 in looking at the time series is that there was a moderate drop in **w** from 47 to 48, but a rise in **c** from 46 to 47. This accounts for the negative residual but it does not tell us why it happened. The problem will remain unresolved here, but it would invite investigation for a special cause if this study were being done in real time.

Runs Test

	Standardized Residual
Test Value[a]	.0000000
Cases < Test Value	25
Cases >= Test Value	30
Total Cases	55
Number of Runs	34
Z	1.572
Asymp. Sig. (2-tailed)	.116

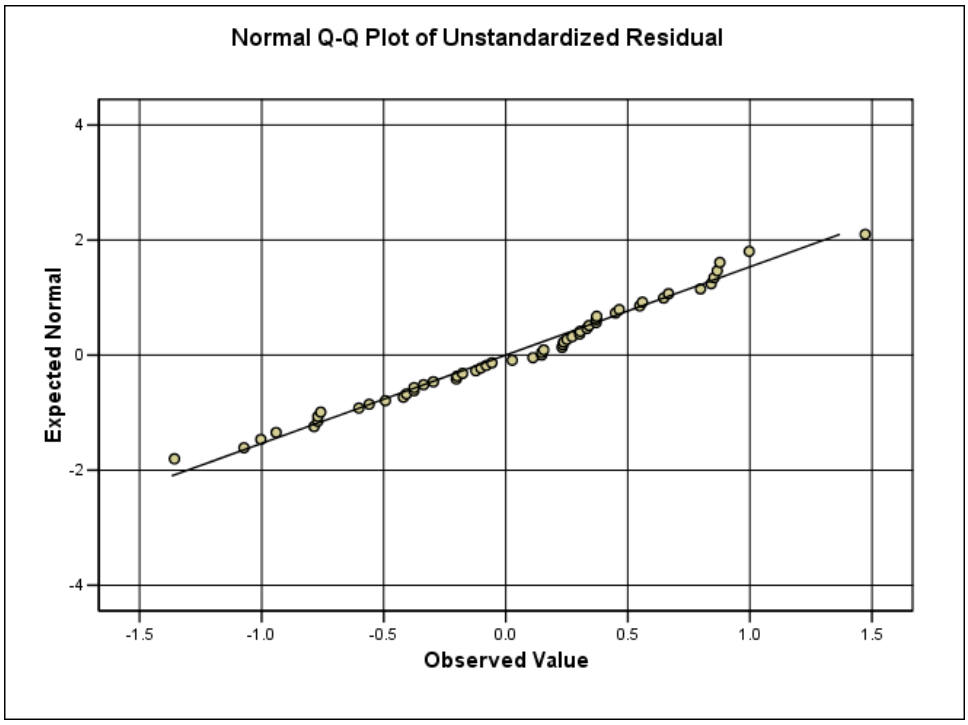
a. Mean

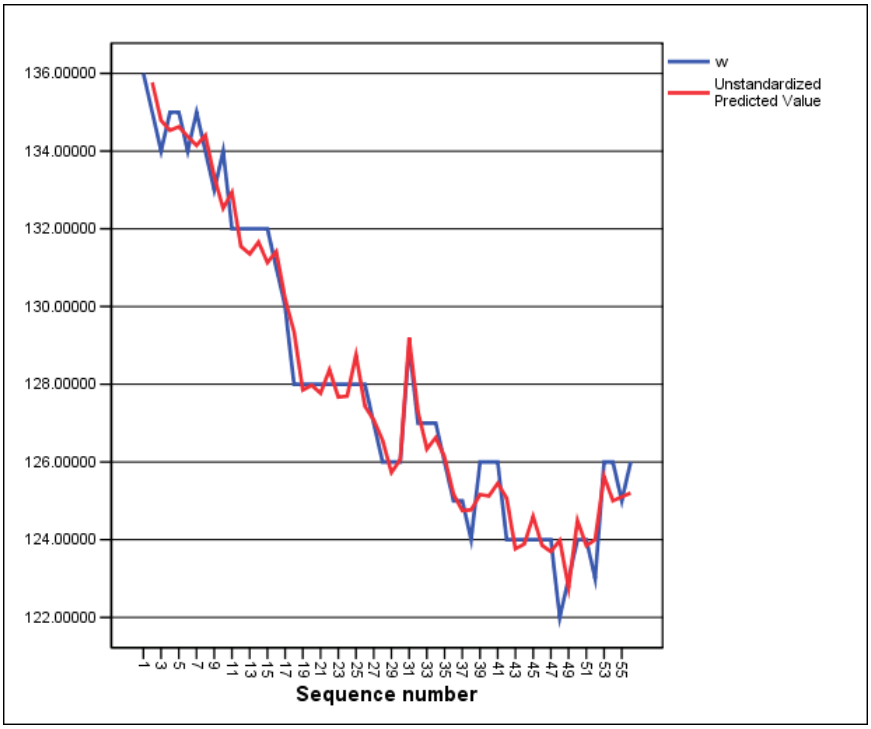
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	.098	55	.200*	.980	55	.475

*. This is a lower bound of the true significance.

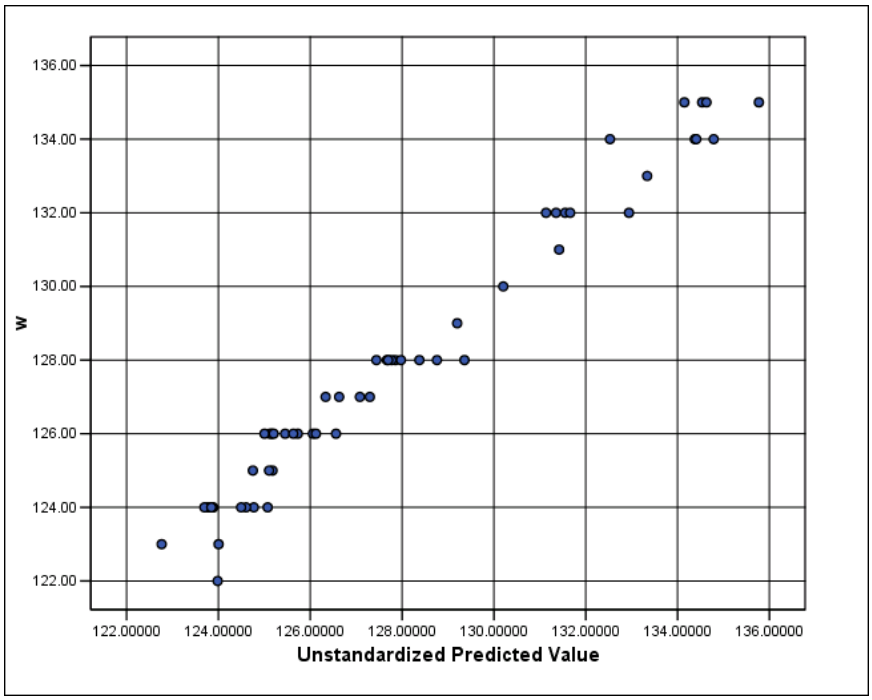
a. Lilliefors Significance Correction





If you look carefully at the plot above, you will see that the predicted values fall rapidly at first, then slowly, then flatten out, and finally appear to turn up. It appears that backsliding in the diet has

begun to occur. We will say more about this in the next section. This plot is one way to look at overall fit. The following plot is another:



Correlations

		w	Unstandardized Predicted Value
w	Pearson Correlation	1	.985**
	Sig. (2-tailed)	.	.000
	N	56	55
Unstandardized Predicted Value	Pearson Correlation	.985**	1
	Sig. (2-tailed)	.000	.
	N	55	55

**. Correlation is significant at the 0.01 level (2-tailed).

Recall that the correlation between **w** and **predicted**, 0.9848, is the multiple correlation coefficient, and note that 0.9848*0.9848 = 0.9698, which is shown in the regression output as **R Square**.

R SQUARE

Up to now we have talked only about R, not RSQ. RSQ (R SQUARE), is very commonly used as an overall summary of goodness of fit of a regression model. It is often said to represent the "percentage (or proportion) of variance of the dependent variable explained by the regression model". Here 97 percent is "explained". More precisely, RSQ is (to a pretty close approximation) the ratio of the variance (squared standard deviation) of the fitted values of the regression to the variance (squared standard deviation) of the dependent variable. The idea is that the larger the variance of fitted values, the more the regression "explains".

However, RSQ is less easy to visualize than is R itself. Recall that R is simply the slope of the regression line on the standardized plot of dependent variable versus fitted.

Finally, there is a widespread and serious misunderstanding about RSQ that you should be aware of: it is the erroneous belief that a "low" RSQ means a poor statistical analysis and a "high" RSQ means a good analysis. In fact, given competent statistical analysis, the value of RSQ depends on the data set being analyzed. You can't get many calories from a turnip, and you can't avoid getting a lot of calories from a pizza.

- A low but significant RSQ may be associated with a model that is highly useful for prediction and understanding. Its low value tells you that there is much room for improvement if you can find additional useful independent variables; it does not invalidate the good work you have done with the independent variables now available.

- A high and nominally significant RSQ can come from a "rash regression", a disastrous blunder that we shall explain at the end of Section 6 below.

- Beware the "RSQ putdown": "Your RSQ is only 30 percent? Your results are useless". People who say that often do not really know what they are talking about.

- Finally, for a measure of goodness of fit, rely primarily on the standard deviation of residuals.

ADJUSTED R SQUARE versus UNADJUSTED R SQUARE

In our current application, ***SPSS*** also prints out "Adjusted R Square" with the value 0.967, slightly lower than "R Square", which equals 0.970. Our shorthand for "Adjusted R Square" is RSQ(ADJ). The adjusted value, RSQ(ADJ), makes allowance for a curious shortcoming of RSQ: As you add variables to a regression model, RSQ can never decline and typically increases, even if the variables do not contribute significantly to the fit[14]. Hence the more variables in the model, the lower is RSQ(ADJ) compared to RSQ.

We can illustrate this phenomenon with the following very simple example. The file below, named SILLY.sav, consists of only three rows of data on three variables: **w**, for weight, **c_1**, for calories lagged one day, and **tempyaku**, the daily temperature in degrees Fahrenheit in Yakutsk, Siberia.

[14] As we show in a moment, we could say "...even if the new variables make no sense whatsoever."

SILLY.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Ut

1 : w 134

	w	c_1	tempyaku
1	134.00	750.00	-35.00
2	136.00	1230.00	-29.00
3	135.00	1150.00	-24.00
4			
5			

As ridiculous as this example may seem, we can carry out the mechanics of fitting a linear regression model. Here are the results for the simple linear regression of **w** on **c_1**:

Model Summary

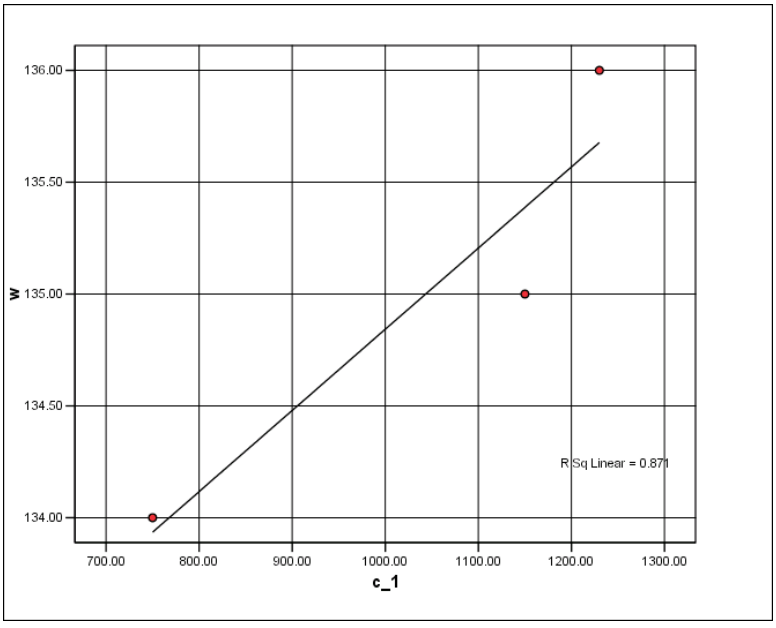
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.933[a]	.871	.742	.50800

a. Predictors: (Constant), c_1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	131.214	1.487		88.266	.007
	c_1	.004	.001	.933	2.598	.234

a. Dependent Variable: w



We see that R Square is 0.871 and Std. Error of the Estimate, 0.508. You can see that with only three data points, by moving any one of them around we can make RSQ as large or as small as we desire. Now look what happens in three dimensions when we introduce the third variable, **tempyaku**, the temperature in Yakutsk, Siberia:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	1.000[a]	1.000	1.000	.

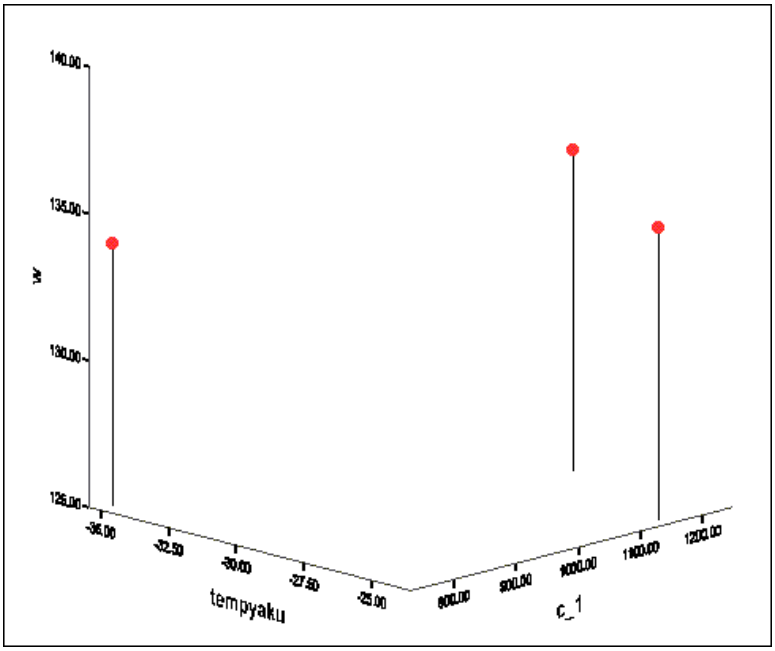
a. Predictors: (Constant), tempyaku, c_1

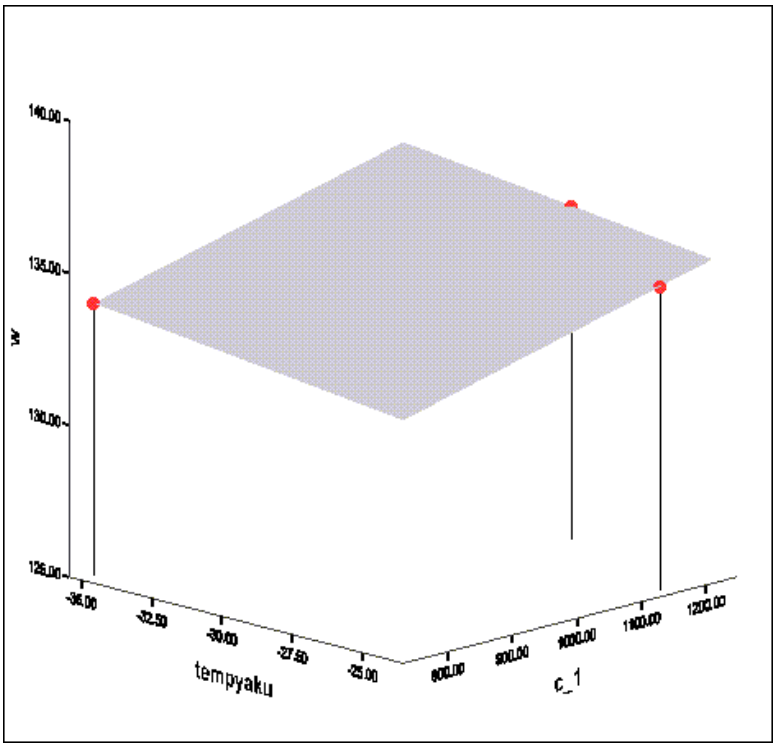
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	125.944	.000		.	.
	c_1	.006	.000	1.429	.	.
	tempyaku	-.111	.000	-.612	.	.

a. Dependent Variable: w

R Square is now equal to one, as high a value as is possible. The reason for this is simple. With only three data points, the fitting of a regression plane in three dimensions is **without any residual error**, as shown in the figures below:





Notice that in the **Coefficients** table displayed above the standard errors for each of the coefficients are zero and the t-ratios and significance levels cannot even be reported (t is infinite, and the significance level is zero). **Note also that there is one critical error in the Model Summary table**: Adjusted R Square is reported to equal one, the same as R Square. When you see the formulas in the appendix at the end of this chapter you can show that this reported value for RSQ(ADJ) cannot be correct. The Adjusted R Square column in the **Model Summary** table above should be blank, as are the columns for t and Sig. in the other table. (Perhaps ***SPSS*** will correct this error in future versions.)

Finally, here is a useful fact that you should remember:

As an additional variable is added to a regression model, Adjusted R Square will increase if and only if the Std. Error of the Estimate (Residual Standard Error) decreases.

This point is illustrated in the **Model Summary** table displayed on page 6-27 above. We repeat the table here, but with the interior cell format changed to show more decimal places for some of the reported values:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.963[a]	.92650	.92512	1.02655
2	.979[b]	.95801	.95639	.78340
3	.980[c]	.96045	.95812	.76770
4	.985[d]	.96981	.96740	.67733
5	.985[e]	.96982	.96674	.68419

a. Predictors: (Constant), w_1

b. Predictors: (Constant), w_1, c_1

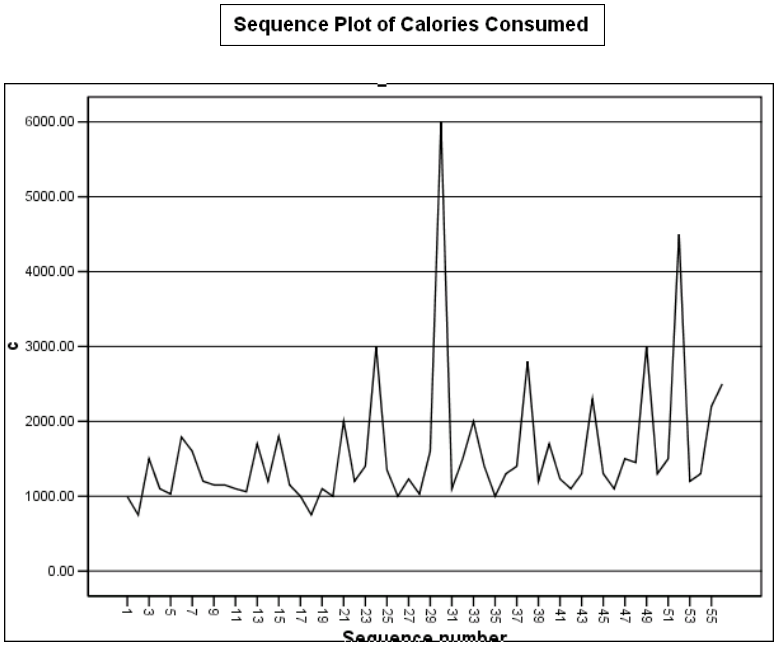
c. Predictors: (Constant), w_1, c_1, time

d. Predictors: (Constant), w_1, c_1, time, tsq

e. Predictors: (Constant), w_1, c_1, time, tsq, c

Observe that R Square increases steadily. If, for a given number of cases, we had enough independent variables, no matter how nonsensical they might be, we could drive R Square to its maximum value, one, as in our example using SILLY.sav. Note also that Adjusted R Square increases steadily, but only through Step 4, while, as stated in the shaded box above, the residual standard error steadily decreases. At Step 5, however, Adjusted R Square decreases, consistent with the increase in the standard error. This halt in the increase in RSQ(ADJ) is also consistent with the variable **c** not being significant as it is forced into the model at Step 5.

5. Study of Calories Consumed (c); Lessons Learned



The variable **c** has some interest in its own right, so we will apply our standard approach to analyzing it. In the analysis it will turn out that another data transformation will provide a useful refinement: we will work with the logarithm of **c** rather than **c** itself. The reasoning is

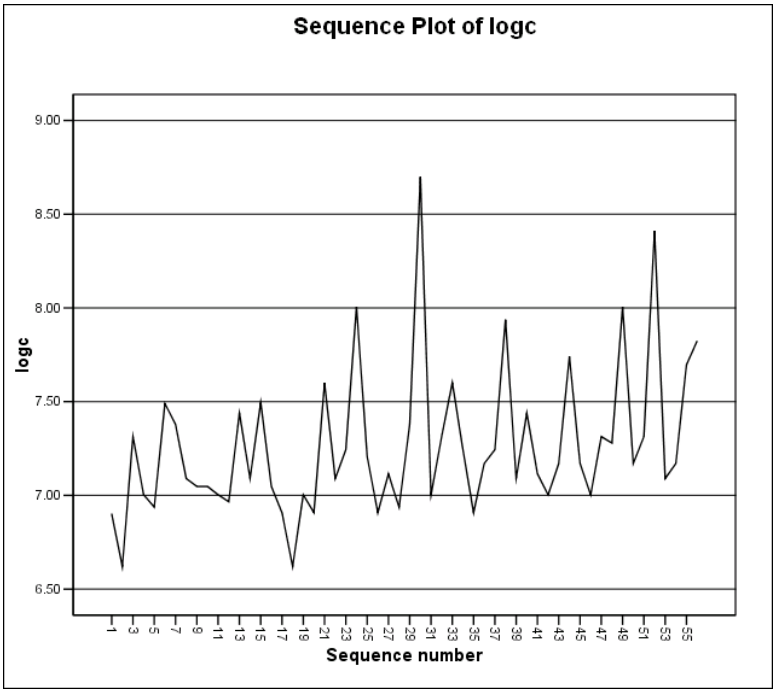
parallel to the reasoning applied in Section 2 of Chapter 5 in the Medical Audit example, where a square root transformation of **tonsils** -- number of tonsillectomies per month -- made the variance more nearly constant. In the present example, the logarithm of **c** will make the variance more nearly constant through time. You may wish to review the first part of Medical Audit before going ahead.[15] The time series for calories, **c**, is trending upwards with increasing variance (vertical dispersion) around the trend. To try to stabilize the variance, we execute the transformation

logc = LN(c) .

In ***SPSS***, the expression **LN** stands for "natural logarithm", as opposed to "common logarithm". It really does not matter which of the log systems you use in transforming, but we will work with natural logs in ***STM***. Before we go any further, you should read the following box and try to remember the warning in future analyses:

> The logarithmic transformation is used to stabilize the variance of a process with nonconstant variance. It also can sometimes make data look more normal than they do before the transformation. But the transformation can **never make nonrandom data more random**.

Here is the time series after the log transformation has been applied:



[15] In Section 2 of Chapter 3, in the example of time intervals between successive cars passing a fixed point, we used a cube root transformation. In this instance it was the normality assumption rather than constant variance that made the transformation advantageous.

If you overlook momentarily the large spike in the middle of the series, the increase in relative variability in the early part of the series does make the overall variance about the trend line appear more nearly constant.

We shall now fit a time trend to the logarithmic series using **Linear Regression**:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.365[a]	.133	.117	.37265

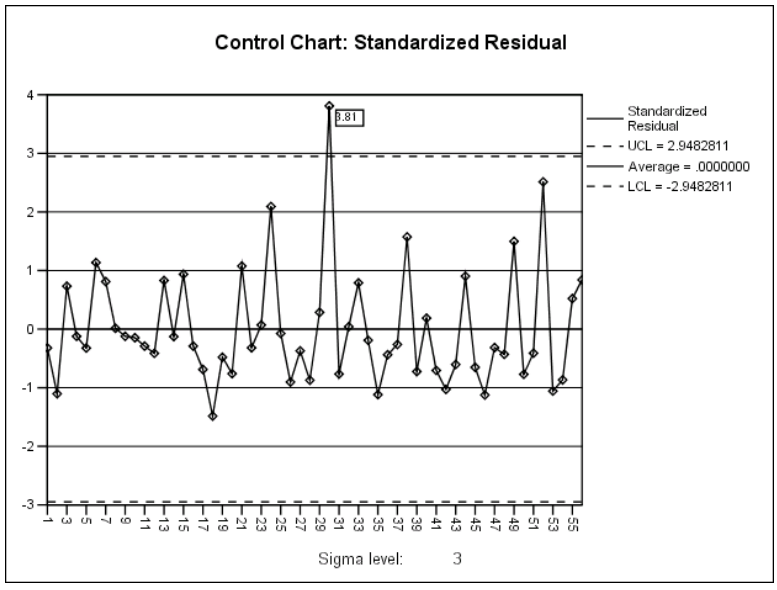
a. Predictors: (Constant), time

b. Dependent Variable: logc

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	7.013	.101		69.478	.000
	time	.009	.003	.365	2.882	.006

a. Dependent Variable: logc



Note the large positive residual at **time**=30. We saw the same phenomenon back in the multiple time series plot for both **w** and **c**, and we have already related it to the large increase in W on the following day. It must have been a real food splurge! Other diagnostics are shown below and they are generally satisfactory:

Runs Test

	Unstandardized Residual
Test Value[a]	.0000000
Cases < Test Value	36
Cases >= Test Value	20
Total Cases	56
Number of Runs	28
Z	.378
Asymp. Sig. (2-tailed)	.705

a. Mean

Autocorrelations: RES_1 Unstandardized Residual

Lag	Auto-Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	-.089	.130	. ** .	.472	.492
2	-.212	.129	.**** .	3.174	.205
3	.078	.128	. ** .	3.546	.315
4	-.138	.127	. *** .	4.735	.316
5	-.146	.125	. *** .	6.093	.297
6	.091	.124	. ** .	6.631	.356
7	-.063	.123	. * .	6.891	.440
8	.225	.122	. ****.	10.314	.244
9	.112	.120	. ** .	11.179	.264
10	-.139	.119	. *** .	12.541	.250
11	-.039	.118	. * .	12.653	.317
12	-.210	.116	.**** .	15.894	.196

Plot Symbols: Autocorrelations * Two Standard Error Limits .

Total cases: 56 Computable first lags: 55

6. Time-Series Regression Strategy and Hazards

At the end of Section 2, Chapter 5, we offered "Systematic Strategies for Regression Analysis". The aim was to provide quick, direct routes to the selection and fitting of regression models, given a list of promising independent variables. Instead of developing the regression model step by step, as we have often done for pedagogical purposes, you can simply use **Stepwise Regression** to screen for the variables that appear to contribute significantly. Then you apply diagnostic checks to the model that emerges.

Sometimes several models will fit the data about equally well. If this happens, we give weight to **simplicity or parsimony: it is advantageous to use a model with fewer independent variables if there is modest or zero loss in goodness of fit as measured by the standard deviation of residuals.**

However, if the fit is roughly the same for several regression models, the choice is unlikely to be critical. All models are "in the right ballpark".

The only modification needed in the light of autoregression is that we have to consider lags of potential independent variables as well as trends, seasonals, etc. In the analysis in Section 5, for example, we ended up with **w_1** and **c_1** plus trend variables. With this choice of model,

the diagnostic checks of residuals -- particularly **Autocorrelation** and **Runs test** -- were satisfactory, and there was no need to consider higher lags -- **w_2, w_3, ... , c_2, c_3, ...**

In developing this model we arbitrarily considered only the first lags of **c** and **w**. In general, we need to decide how many lags to consider. Fortunately, this choice is not critical:

- If we try too many lags, the insignificant lags will drop out of the stepwise regression.
- If we try too few lags, the diagnostic checks of the model will suggest that we need to go back to try more lags.

"Rash Regression"

One disastrous error that is often committed by those who are not "statistically savvy" is to be unaware of the need to lag when autoregressive effects are present. For example, someone might reason that **w** obviously depends on **c**, so the simple regression of **w** on **c** should tell the story. Here's where that reasoning leads:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.307[a]	.094	.077	3.71644

a. Predictors: (Constant), c

b. Dependent Variable: w

Coefficients[a]

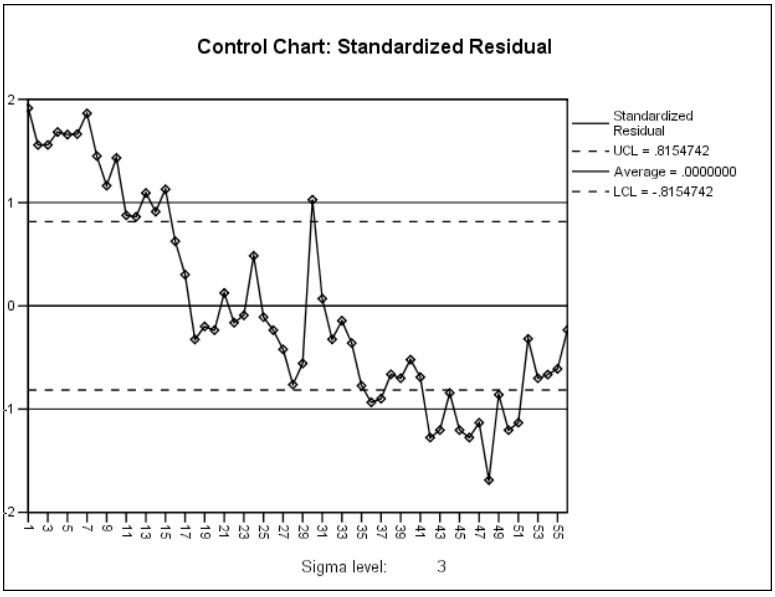
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	130.214	1.019		127.835	.000
	c	-.001	.001	-.307	-2.369	.021

a. Dependent Variable: w

There is indeed a significant effect for calories, but it is in the wrong direction! The negative sign for the regression coefficient of **c** suggests that the more one eats, the less one weighs.

Moreover, the **Std. Error of the Estimate** is 3.716-- several times larger than the 0.677 that we got in Section 4.

Finally, the diagnostic checks for this regression model are terrible, as shown by the examples below:



Runs Test

	Unstandardized Residual
Test Value[a]	.0000000
Cases < Test Value	35
Cases >= Test Value	21
Total Cases	56
Number of Runs	8
Z	-5.545
Asymp. Sig. (2-tailed)	.000

a. Mean

Autocorrelations: RES_2 Unstandardized Residual

Lag	Auto-Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	.877	.130	****.**********	45.463	.000
2	.810	.129	****.**********	84.920	.000
3	.774	.128	****.**********	121.613	.000
4	.699	.127	****.*********	152.109	.000
5	.653	.125	****.********	179.261	.000
6	.617	.124	****.*******	203.962	.000
7	.546	.123	****.******	223.710	.000
8	.495	.122	****.*****	240.287	.000
9	.451	.120	****.****	254.318	.000
10	.364	.119	****.**	263.656	.000
11	.299	.118	****.*	270.091	.000
12	.219	.116	****.	273.646	.000

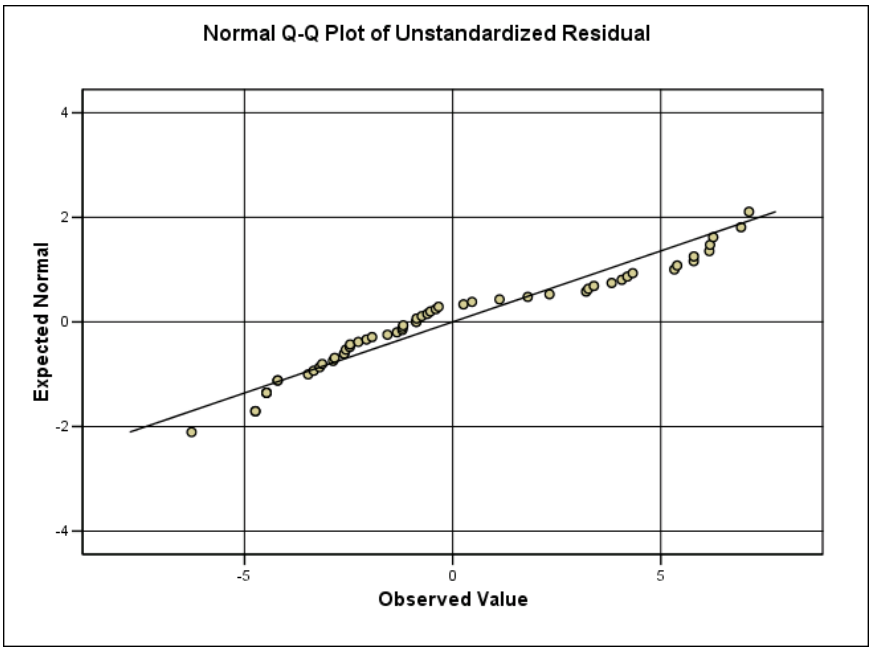
Plot Symbols: Autocorrelations * Two Standard Error Limits .

Total cases: 56 Computable first lags: 55

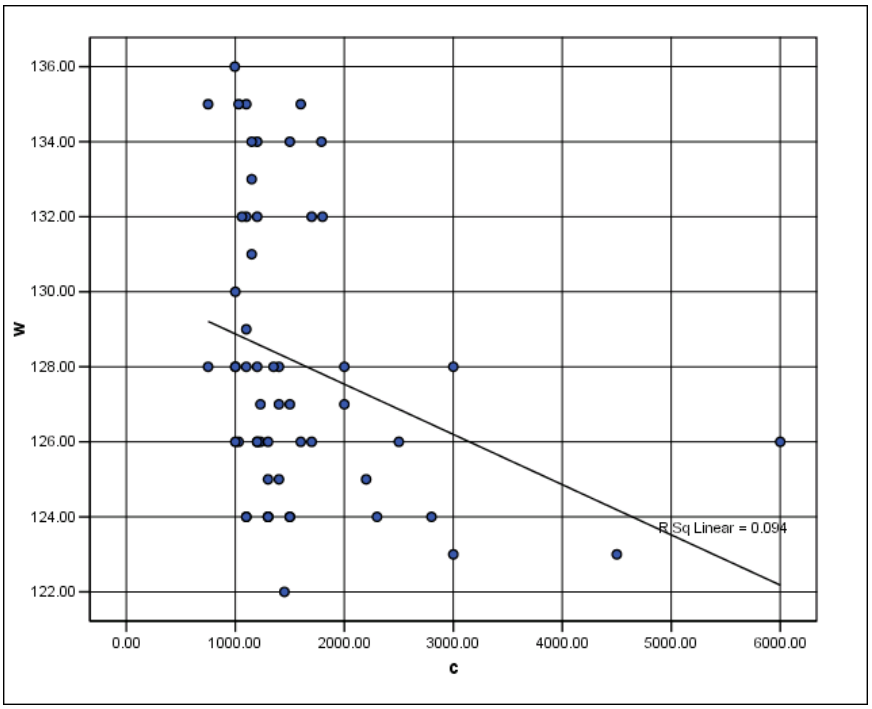
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	.162	56	.001	.924	56	.002

a. Lilliefors Significance Correction



The Shapiro-Wilk value of 0.924 for the wiggly line in the normal probability plot above is significant at a p-value below 0.05. Hence we cannot accept the hypothesis that the residuals are normally distributed.



The plot of **w** vs. **c** shows that the negative slope is caused by a few large values of **c**. The time series showed that this tended to occur later in the diet, when a good deal of weight had been lost.

The kind of statistical blunder illustrated by this example is very common. Its essence is ignoring the time-series aspects of regression when the variables are ordered in time. If there are trends, seasonals, and autoregressive effects in **Y** and/or **X**, the simple regression of **Y** on **X** is almost sure to give nonsensical results.[16] We call this type of ill-advised analysis **"rash regression."**

[16]G. Udny Yule, considered by many to be the father of time series analysis, published a paper entitled "Why Do We Sometimes Get Nonsense-Correlations between Time-Series?" in 1926. His example showed a high correlation between the mortality rate for England and Wales and the proportion of first marriages that are performed in the Church of England.

Appendix: Further Remarks on RSQ and RSQ(ADJ)

We have already discussed RSQ and defined it as the sum of squares due to regression divided by the total sum of squares. This statement requires further amplification. In our discussion of the output from a linear regression analysis we have thus far ignored the table entitled **ANOVA,** which stands for "Analysis of Variance". For illustrative purposes we repeat the regression output shown in the main body of this chapter, but this time we include the **ANOVA** table.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.985[a]	.970	.967	.67733

a. Predictors: (Constant), c_1, tsq, w_1, time

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	736.989	4	184.247	401.612	.000[a]
	Residual	22.938	50	.459		
	Total	759.927	54			

a. Predictors: (Constant), c_1, tsq, w_1, time

b. Dependent Variable: w

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	55.164	13.006		4.241	.000
	time	-.226	.052	-.964	-4.364	.000
	tsq	.002	.001	.589	3.939	.000
	w_1	.591	.095	.613	6.249	.000
	c_1	.001	.000	.173	6.679	.000

a. Dependent Variable: w

The ANOVA table shows an analysis of the sum of squares, the numerator of the sample variance of **w**. (You may remember from another class in which formulas were emphasized that the standard deviation is the square root of the variance and that the variance is the sum of squared deviations about the mean, divided by the sample size minus one. If you do not remember this or never heard it before, no matter.) Recall that we said earlier that the method of least squares fits a regression surface in a way that minimizes the residual sum of squares. At any rate, the **ANOVA** table above shows a total sum of squares equal to 759.927. If you divide this figure by 54, the sample size minus one, you get 14.07. Take the square root of that and you get 3.75, close to the original standard deviation of **w,** 3.87, that we obtained back on page 6-15.[17] What the **ANOVA** table does is to break the total sum of squares into two parts: The first part that is "explained" by the regression, and second or "residual" part that is left over and still "unexplained". Thus, we can calculate from the table above that R Square, the sum of squares due to regression divided by the total sum of squares, is 736.989/759.927 = 0.970, exactly as shown in the **Model Summary** table.

In the following discussion, for typographical reasons, we are going to write "RSQ" and "RSQ(ADJ)" as "$\boldsymbol{R}^2$" and "$\boldsymbol{R}^2_{adj}$". It follows from the paragraph above that $1-\boldsymbol{R}^2$ is the ratio of the residual sum of squares to the total-- i.e., 22.9385/759.927 = .0302 in the example. We might say that 3.02 percent of the sum of squares remains "unexplained".

[17] The reason for a slight numerical discrepancy is that we are using only 55 cases because of lagging as opposed to the full 56 cases used on page 6-15.

Now, the purpose of the silly example involving the temperature in Yakutsk, Siberia was to show graphically that if you have enough independent variables in your regression model, no matter how ridiculous they may seem, you can have a perfect fit-- that is, you can force $1-\boldsymbol{R}^2$ all the way to zero. In fitting a regression model with an intercept, the "perfect fit" will happen if the number of independent variables is one less than the number of observations. Since it does not seem fair to give a modeler credit for a good fit when the model contains meaningless variables, the idea of $\boldsymbol{R}^2_{adj}$ was developed. You will see that with a fixed number of observations, $\boldsymbol{R}^2_{adj}$ gives a handicap to the modeler who rashly includes too many independent variables in the regression equation.

Before we proceed, some more notational definitions:

SST = the total sum of squares
SSR = the sum of squares due to regression
SSE = the residual (error) sum of squares
n = the number of observations (sample size)
k = the number of independent variables plus one (for the intercept)

Thus we can write

$$1-\boldsymbol{R}^2=\frac{\boldsymbol{SSE}}{\boldsymbol{SST}}$$

and mention again that as ***k*** increases, $1-\boldsymbol{R}^2$ **cannot increase**-- the more independent variables, the closer the fit.

Adjusted R Square, $\boldsymbol{R}^2_{adj}$, is defined by the following relationship:

$$1-\boldsymbol{R}^2_{adj}=\frac{(\boldsymbol{n}-1)}{(\boldsymbol{n}-\boldsymbol{k})}(1-\boldsymbol{R}^2)=\frac{\boldsymbol{SSE}/(\boldsymbol{n}-\boldsymbol{k})}{\boldsymbol{SST}/(\boldsymbol{n}-1)}$$

Note that the numerator in the expression above, ***SSE/(n-k)***, is the formula for the **Mean Square for Residuals**, the square of the **Std. Error of the Estimate (Residual Standard Error)**; and the denominator, ***SST/(n-1)***, is the formula for the sample variance (the square of the standard deviation of the dependent variable, as explained in the first part of the appendix). Thus, whereas we can refer to $1-\boldsymbol{R}^2$ as "the unexplained proportion of the total sum of squares," $1-\boldsymbol{R}^2_{adj}$ is "the unexplained proportion of the original sample variance."

If the sample size, ***n***, is held fixed, then as ***k***, the number of independent variables plus one, increases, the ***SSE*** (sum of squared residuals) cannot increase-- assume that it decreases. The factor ***(n - k)***, however, decreases. Thus we cannot be sure whether the numerator in the

expression above decreases or increases. If the decrease in ***(n - k***) is less than the decrease in the ***SSE***, then $1-R^2_{adj}$ will decrease. But if the decrease in ***(n - k***) exceeds the decrease in ***SSE***, then $1-R^2_{adj}$ will actually become larger. As you continue to introduce new variables into your model, at some point an increase in $1-R^2_{adj}$ will indicate that you should stop doing that. Note also that whenever $1-R^2_{adj}$ decreases you will also see a decrease in the M**ean Square for Residuals**, as equivalently stated in the shaded box on page 6-35 of the textbook, thus another stopping rule is to quit adding independent variables when you detect an increase in the standard deviation of the residuals.

Unfortunately, this discussion makes the whole procedure of model building seem too mechanical. As we discuss these matters further we will emphasize two points:

(1) It is not a good idea to keep adding independent variables to the regression equation in a mindless way, even if the procedure seems to lead to decreases in the standard deviation of the residuals. You should always be guided by theory, if possible, and at least by common sense.

(2) If the introduction of an additional independent variable leads to only a slight improvement in the fit, back up and retain the simpler model. This is the "principle of parsimony", which, according to our preachments, will lead to better results in the long run.