CHAPTER 4: INTRODUCTION TO REGRESSION

1. Fitting Trends by Regression: the Running Example

In Chapter 3, Section 5, we discussed two applications in which out-of-control data sets displayed apparent trends -- systematic level shifts through time -- a running application and the data on mortality in intensive care.

- In the running example, slowing down during the workout showed up on a control chart as a clear tendency for the split times to move upward through time, a tendency that was easily seen in spite of substantial variation from point to point.

- In the intensive care application, a downtrend in mortality rates was barely visible on the control chart since the point-to-point variation was very large. In fact, you may have thought that we were claiming too much when we asserted that a trend was present.

If there is a trend, the process **cannot be** in statistical control. Trends, when they are present, are of great practical importance. We cannot be content simply to monitor the process, looking for special causes. Systematic change is taking place.

This systematic change may be either favorable or unfavorable. If the change is unfavorable, there is need for action to stop or reverse the trend. If the change is favorable, there is need for action to ensure that whatever is causing the favorable change is maintained.

Here are important ideas concerning trends:

- Although trends do not tell **why** systematic change is taking place, they provide a signal to look for root causes of this change.

- They may also provide hints as to where and how to look for root causes.

- Trends are useful in forecasting, at least for short periods into the future. Professor Walter Fackler of the University of Chicago, whose one-year-ahead macroeconomic forecasting accuracy was unmatched, said that awareness of underlying economic trends was essential for his forecasting methodology.

- Trends furnish a basis for performance comparisons. If the general trend of health care costs is upward, a company or hospital that holds these costs constant is doing well.

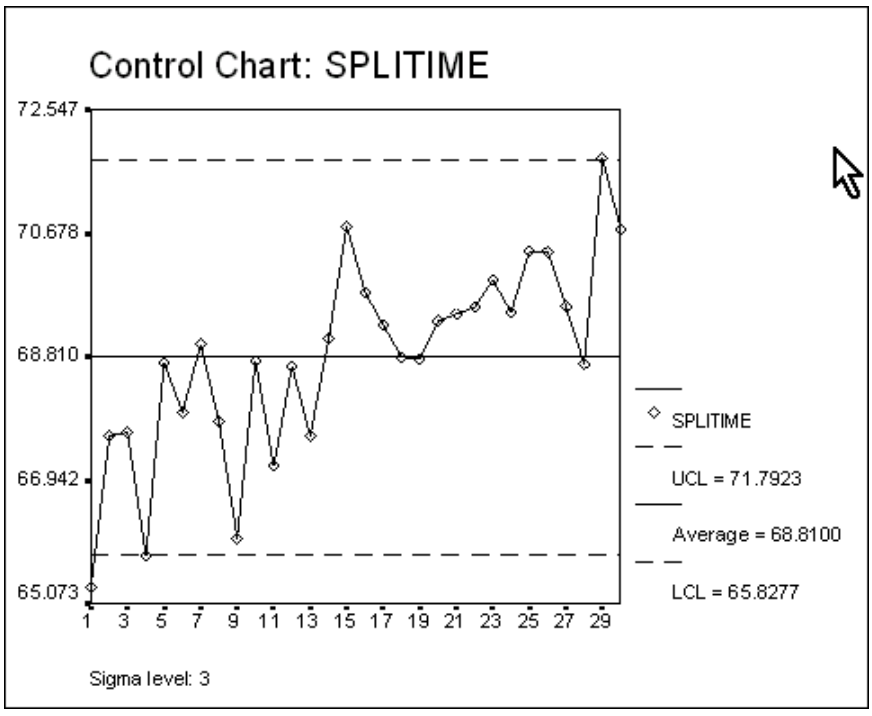
We need statistical tools to detect and confirm the presence of trend, and to fit trends that are confirmed. By fitting a trend, we mean a procedure that estimates how rapidly the systematic change of process level is occurring as time passes. One way to do this is by **regression**, a tool that also has many other statistical applications besides trend fitting.

In this section and the next, we shall introduce regression concepts by showing how they can be applied, using ***SPSS***, to fit trends in the two examples of trend from Chapter 3-- the running data and the intensive care mortality data.

The ***SPSS*** command sequence for **linear regression** will give more computer output than we have immediate need for. Consistent with the policy of explaining only what you need to know when you need to know it, we shall explain only those aspects of the output that are relevant at the moment. (If you are impatient about information in the output that we deliberately pass over, you may wish to read ahead. We will cover almost all of it somewhere in **STM**.)

The running example (see LAPSPLIT.sav) has a trend that is easy to spot visually on a control chart, so we will start with that, beginning with a brief review of background.

> Data from a running workout around a block that is approximately 3/8 miles, divided by markers into approximately three 1/8 mile segments. Timed with a Casio Lap Memory 30 digital watch to obtain timings for each of 30 1/8 mile segments (3 3/4 miles in total). Attempted to run at constant (and easy) perceived effort, without looking at the splits at each 1/8 mile marker. Afternoon of 19 September 93. Data are in seconds.



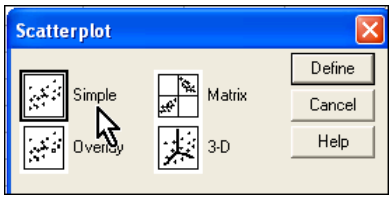
The upward trend is clearly visible.

We can also think of this control chart as a **scatter plot**, that is, a plot that relates a vertical variable **Y** to a horizontal variable **X** in the standard graphical display with coordinate axes.

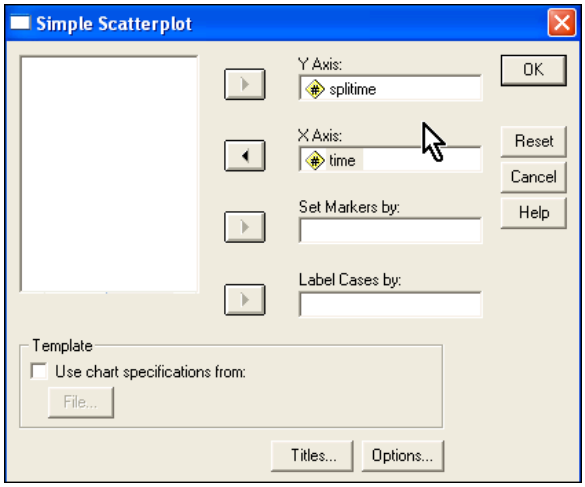
For example, just imagine the label "**TIME**" for the horizontal axis of the time series plot above. "**SPLITIME**" is the label for the vertical axis. Then the plot can be interpreted as a scatter plot of **splitime** against another variable, **time**.

We can make a scatter plot directly, using the ***SPSS*** plotting routine, as shown below. First, we create the variable **time** explicitly. The way to do it is to click on **Transform/ Compute…** and type in **"time = $casenum"**. **$CASENUM** is a built-in ***SPSS*** function that assigns the numbers of each row in the spreadsheet to the new variable **time**. If you look at **Data Editor** after the transformation you will see that there is now a second column in the interior of the spreadsheet containing the case numbers under the heading, **time**.

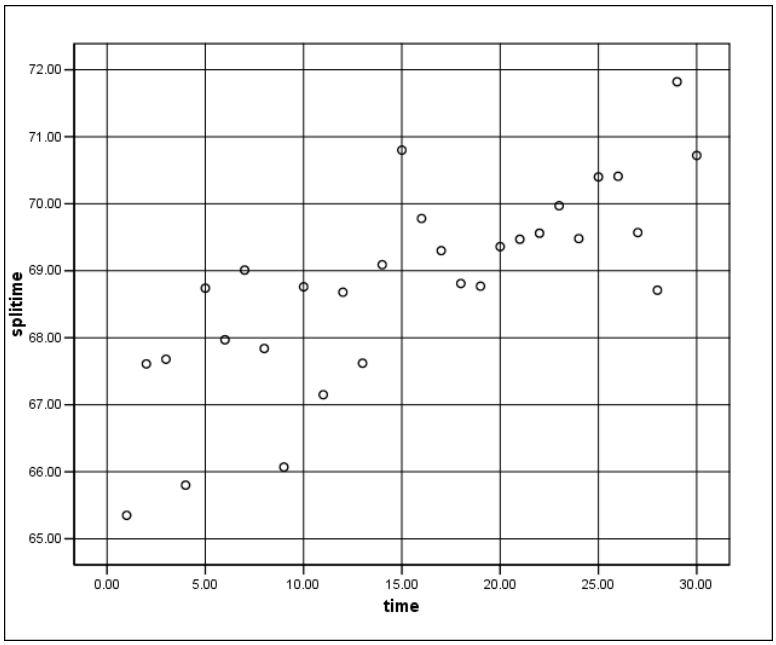
At the next step we execute **Graphs/Scatter**…, which brings up the following dialog box:



We then highlight the icon for **Simple** and click on the **Define** button. We have selected **splitime** (read values on the Y Axis) to be plotted against **time** (on the X Axis).



Here is the scatter plot that appears after we click on **OK**:



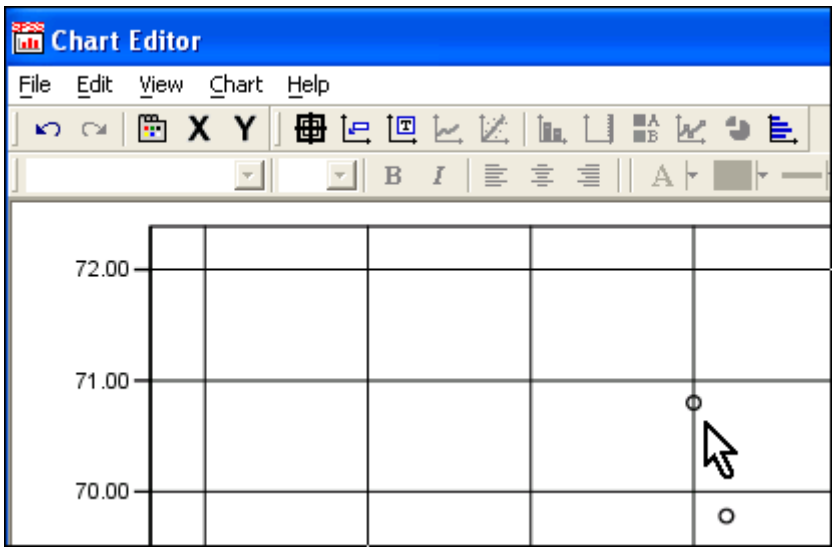
Before moving ahead, compare this scatter plot with the control chart. There is a close relationship between the two graphs. Both show essentially the same information. For example, for any time-ordered data, we can plot the variable of interest on the vertical axis of a scatter plot and the variable **time** on the horizontal axis of the scatter plot to obtain a run chart. We admit that the scatter plot is a bit more difficult to digest because the points are not connected by lines to emphasize their sequential occurrence, but that is a problem of aesthetics that we can fix later.

For now, we shall think of **regression** as a tool that will fit a line to provide a quantitative description of the upward trend that can be seen on the scatter plot as a general tendency of the points to move upward as we move from left to right. For any given **time**, this line gives the trend value -- that is, the **"predicted"** value -- of **splitime.** It is assumed (later we'll learn to check the reasonableness of the assumption) that for the various values of **time** the expected values of **splitime** trace out a straight line.

It is useful to recall that the mathematical equation of a straight line is of the form **Y = a + bX**, where:

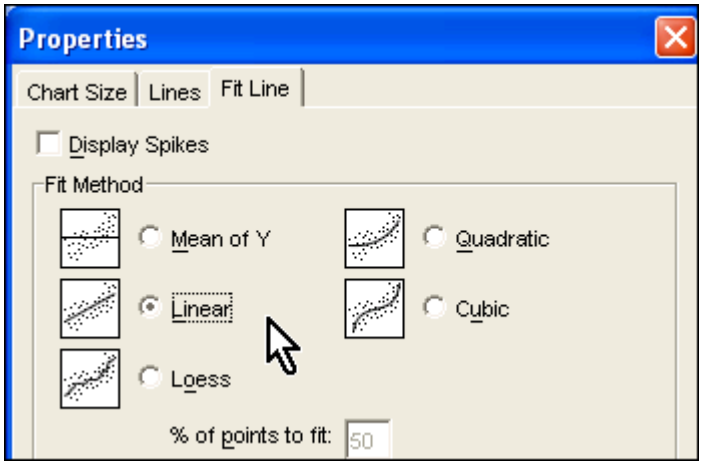
- **Y** is the vertical variable, often called the "dependent variable".
- **X** is the horizontal variable, often called the "independent variable".
- **b** is the slope of the line, that is, the change of Y for a unit increase of X. If **b** is positive, the line slopes upwards; if **b** is negative, the line slopes downwards.
- **a** is the "constant". If the origin of the horizontal and vertical axes is at zero, **a** is also the "intercept", that is, the fitted Y value given by the line when X = 0.

Go back to your scatter plot that we have shown above and, placing your pointer within the plot, double click on the left mouse button. You should immediately get a window entitled **Chart Editor**. The top section looks like this:

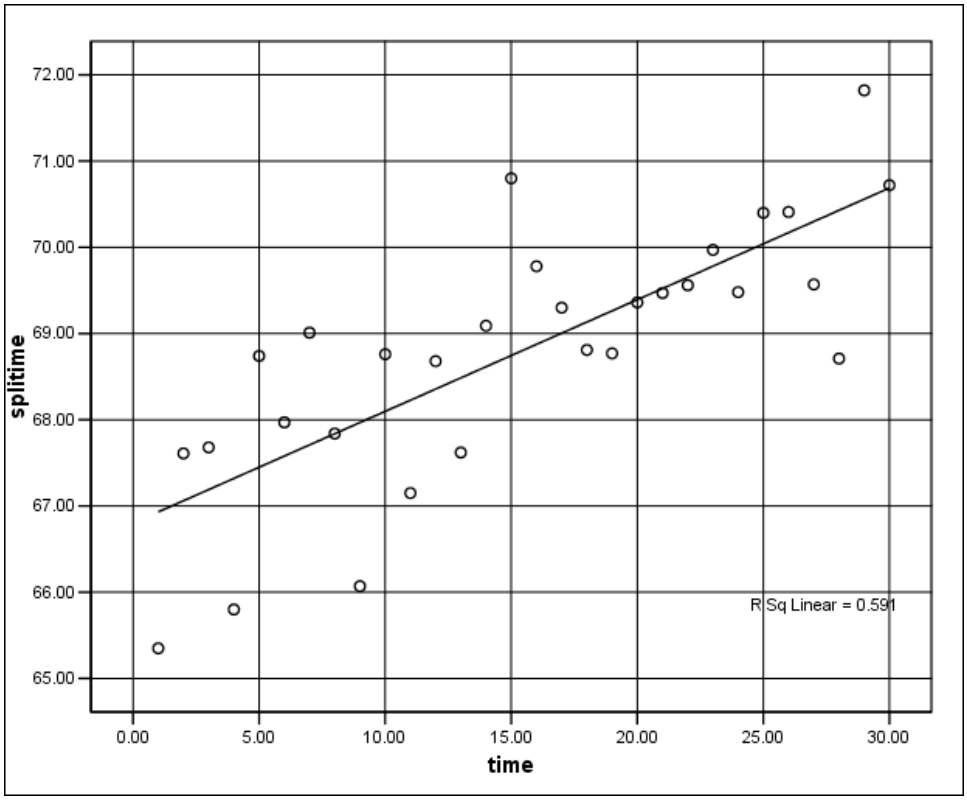


Then, placing your pointer tip on one of the points—any one will do—double click again. You will get another window showing a color palette, but you can quickly close that. The important result is that your action will also cause some additional icons on the toolbar that were previously grayed out to become active. We are especially interested in the one that looks like this:

If you double click on that icon you will get the following Properties window in which you should mark the little circle next to **Linear**:



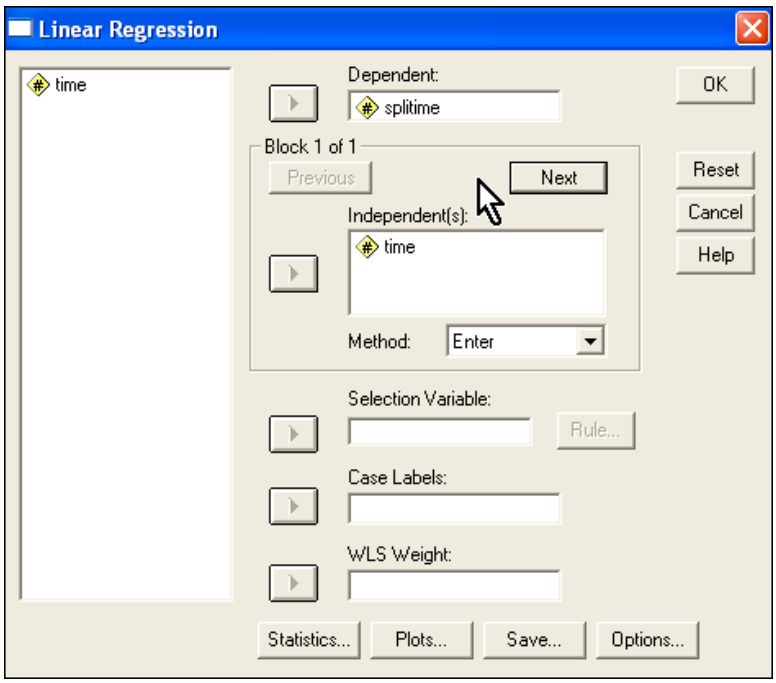
Finally, when you click on the **Apply** button, a straight line appears in the plot. It appears to pass roughly through the center of the mass of points—the precise method of fitting will be discussed later. For now just think of it as the "**line of best fit"** in the plot of **splitime** against **time**. (Note also the annotation in the plot that says "**RSq Linear = 0.591"**. That also will be explained eventually in this chapter.)



We now need to show the way in which the regression line is determined by ***SPSS*** via the sequence **Analyze/Regression/Linear…**, and explain the essentials of the output from the command.

The Basic Regression Command:

The sequence of clicks on **Analyze/Regression/Linear…** brings up the following dialog box:



We have indicated the variable to place on the left-hand-side of the equation Y = a + bX, called the **dependent variable**. It is **splitime**. Also we must tell ***SPSS*** the name(s) of the **independent variable(s)**, i.e., the variable(s) on the right-hand side, in this case **time**. The variable called **time** plays the role of X in the straight line equation. Note that the box for **Method:,** just under the one for **Independent(s),** is set at **Enter**. Finally, click on the button labeled **Statistics…**at the bottom of the window and when a new window appears make sure that only two boxes are checked for now—**Estimates** and **Model Fit**.

Don't worry now about the other boxes or settings. We will discuss some of them as they are required; others need not be discussed for the level of this course. So all that is necessary now is to press **OK**, and we obtain the following displays:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	time[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: splitime

This first exhibit merely summarizes the operations that have been performed. We may not even show it in subsequent discussions unless it is absolutely necessary to make a point.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.769[a]	.591	.576	.96588

a. Predictors: (Constant), time

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37.717	1	37.717	40.428	.000[a]
	Residual	26.122	28	.933		
	Total	63.839	29			

a. Predictors: (Constant), time

b. Dependent Variable: splitime

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	66.802	.362		184.691	.000
	time	.130	.020	.769	6.358	.000

a. Dependent Variable: splitime

Do not be daunted by all these confusing numbers. We will review the key parts of the output, one step at a time.

The Key Result: the Fitted Line

In the last table image shown above the mouse pointer indicates the two numbers under the column heading "**B**". These are the estimated parameters[1], for the regression line determined by **a** = **66.802**, called the "constant" or "intercept", and **b** = **0.130**, the "slope". The official statistical term for these numbers is "regression coefficients." With these two numbers we can write an expression for the straight line

splitime = 66.802 + 0.130 **time**

[1]The parameters of our statistical model, Y=a + bX, are **a** and **b**. Y and X are variables. Thus, when the parameters are given specific values such as 66.802 and 0.130, the line is made concrete and, for example, it may be drawn on the graph. We call the parameters "estimated" because if we took another sample of data they would come out to be similar (we hope), but slightly different.

and you should do this in your notes-- perhaps right under the title in the regression output display above. We prefer to write it this way:

predicted splitime = 66.802 + 0.130 **time**

to emphasize that the line does not represent a deterministic relationship between actual **splitime** and **time**. Rather the line is a description of the **expected relationship** of **splitime** to **time**, abstracting from chance variation, just as the center line on a control chart describes the **expected level** of the data for a process that is in a state of statistical control and varying unpredictably above and below the line.

This relationship says:

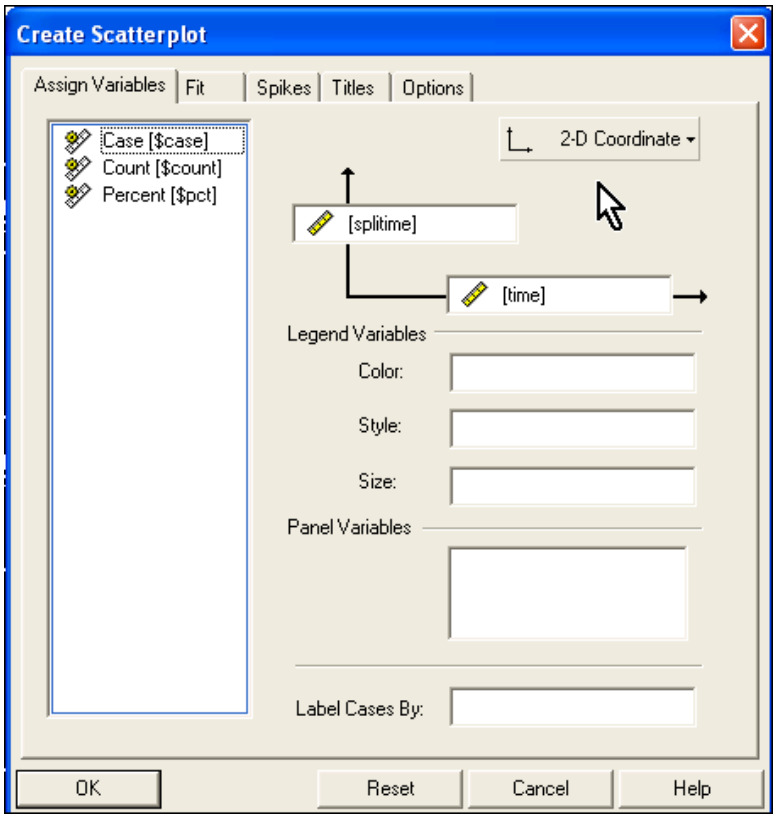
- **On average**, for an increase of one unit in **time**, the regression line rises 0.13 seconds, the value of the **slope coefficient** or **regression coefficient**.

- The **regression coefficient** with numerical value 0.13 describes the runner's tendency to slow down in this running workout at constant perceived effort. The pace tends to slow by 0.13 seconds per 1/8 mile, or by 8*0.13 = 1.04 seconds per mile.

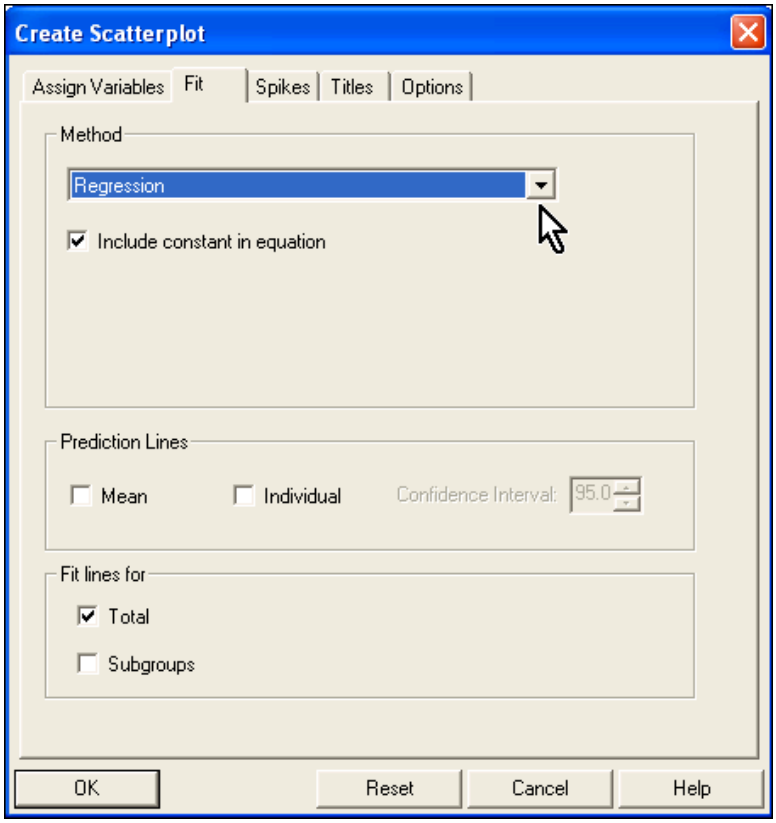
- For **time** = 0 (at the start of the workout), the height of the line is 66.8 seconds. Technically, this is the **constant or intercept**. The constant is not directly relevant in this particular application because the first actual reading is at **time** = 1. At **time** = 1, the height of the regression line is 66.8 + 0.13 = 66.9.

Making the Plot Look Pretty

This is perhaps as good a place as any to discuss how to make our graphic of the regression line representing trend somewhat fancier and more like a control chart. This time instead of executing **Graphs/Scatter…,** we follow a different route via **Graphs/Interactive/Scatterplot…**
This brings up the following dialog window in which we have dragged the variable names **splitime** and **time** into the boxes for the Y and X axes respectively.

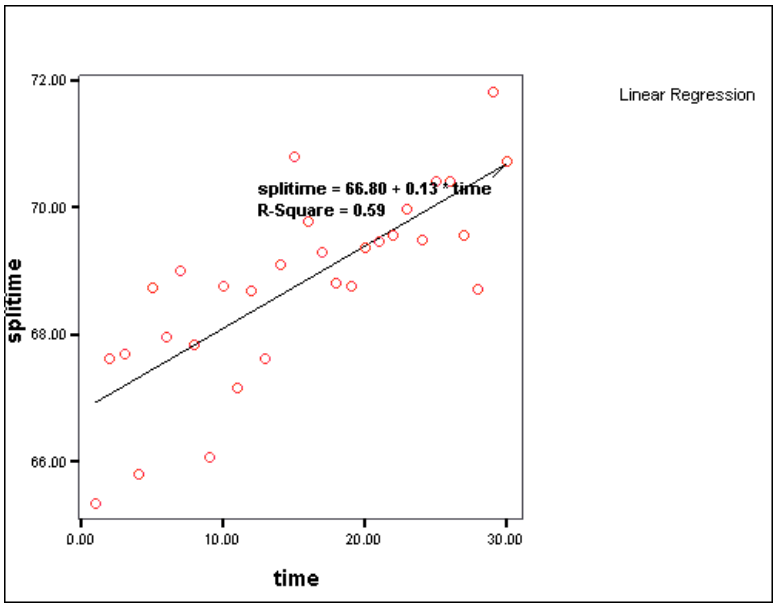


Before clicking on the **OK** button, however, we touch the **Fit** tab. This produces the following image:

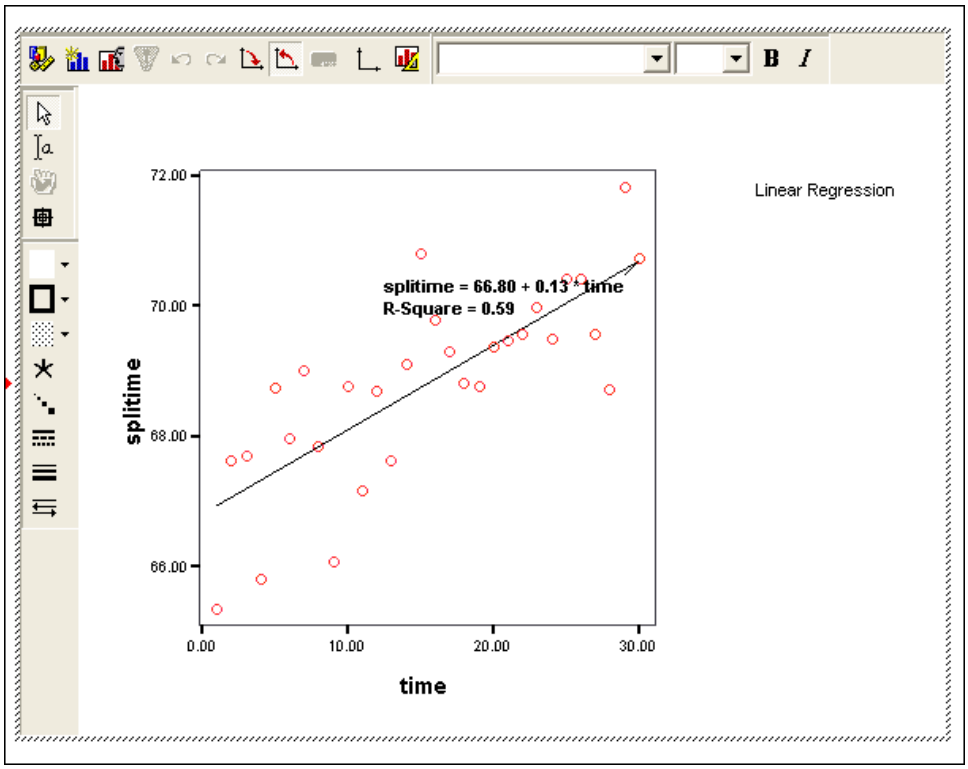


From the pull-down menu under **Method** we choose **Regression**, and we make sure that we have checked the box for **Include constant in equation**. Also the box for **Total** under **Fit lines for** must be checked.

After we click **OK** on the **Fit** page we get this new scatter plot:



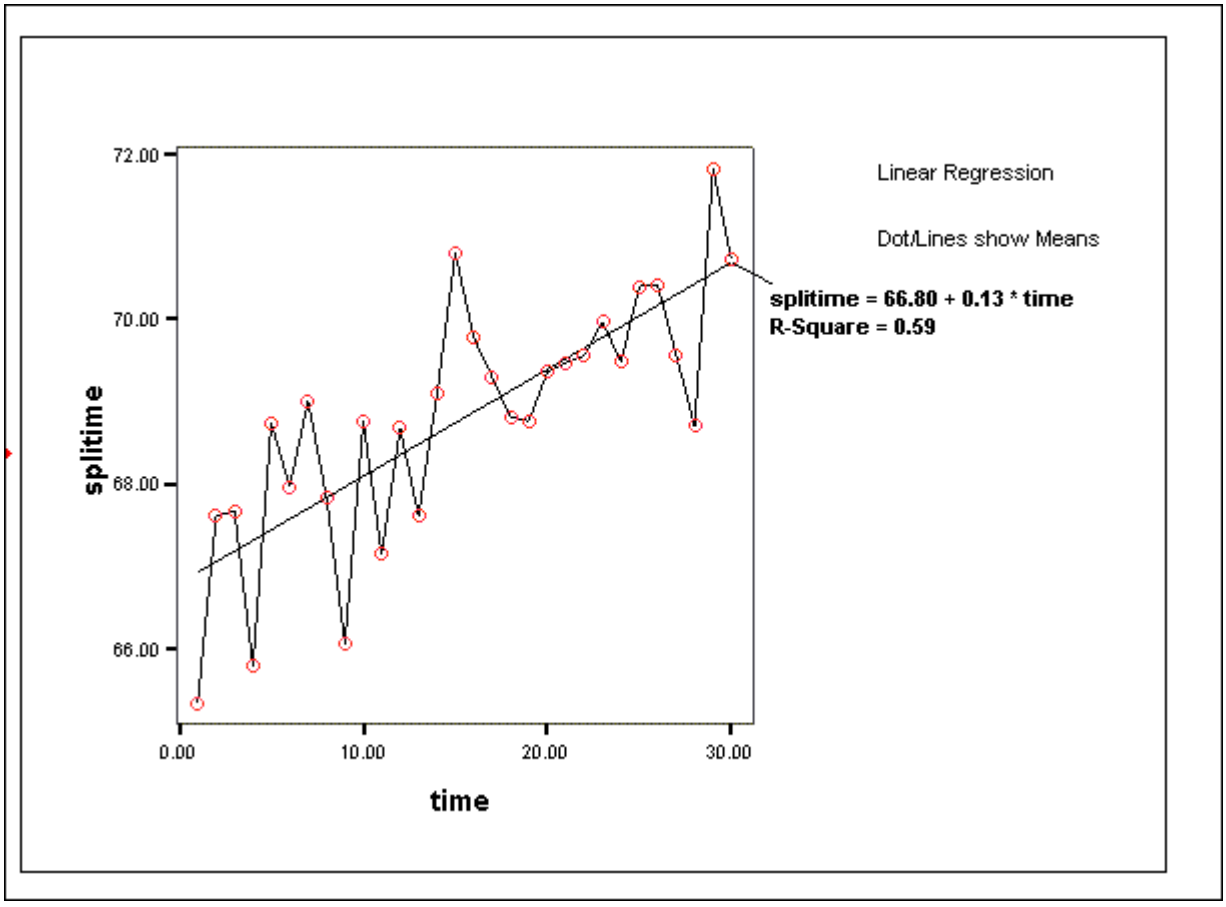
You see that this time the equation for the linear regression line is nicely displayed. We are not done yet, however, in dressing up the display. The next step is to double-click on the image to bring up some new tools displayed along the left and the top borders of the plot:



If we click once with our pointer within the annotation showing the linear equation, a dotted blue outline of the text box will appear. We can then drag the box to the right of the plot to avoid obscuring the data points. After you have the box in the desired position then click on the tool in

the upper left that looks like this  . This is the **Insert Element** tool. The mouse click causes a long menu of various gadgets to drop down. Second in the list is this one: 

. Just a single click on that icon and--Voila!-- our plot now looks like this:

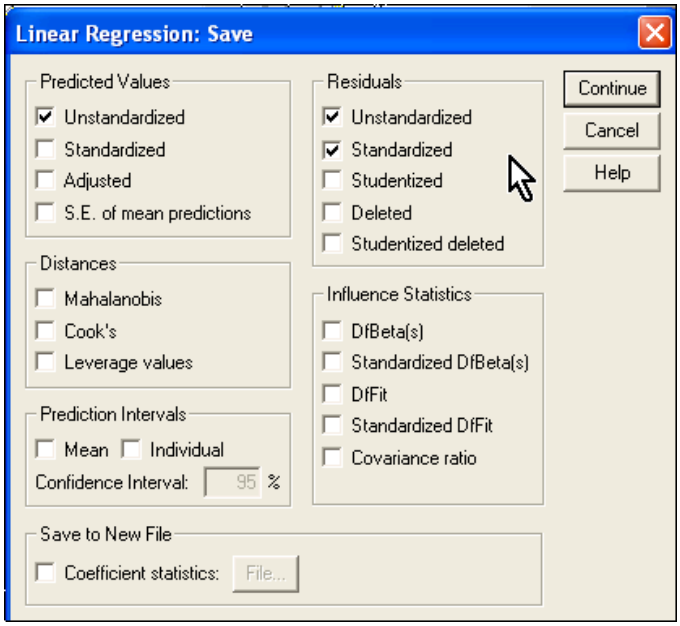


There is just one more final touch that is necessary. Double click on the annotation that says "**Dot/Lines show Means**" and in the **Dots and Lines** window that appears uncheck the **Display Key** box. That causes the mysterious note to disappear and now we have an image that looks just like the control chart, but with the trend line and its equation clearly shown.

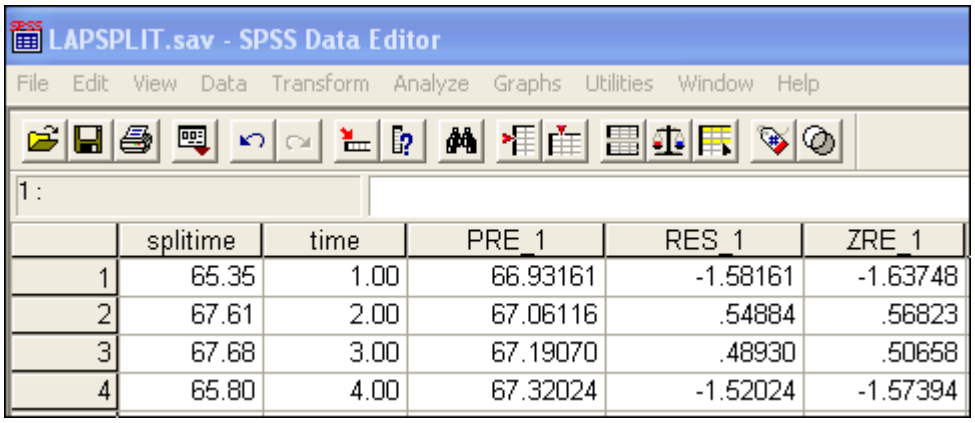
We have shown how powerful the graphics features of ***SPSS*** can be and how, with some effort, we can produce some very elegant displays. We shall show more of these "little niceties" as we continue, and some of you may discover others that are even more fancy and illuminating. We repeat, however, that **you should not take valuable time away from learning the fundamentals** just to make more beautiful plots. For almost all purposes in your homework and in the final project you will not need to go beyond the simplest of graphic displays. To annotate a graph, for example, you can just as easily type a nearby description in your word processing document rather than inserting and properly placing a text box in the plot itself.

The Save Feature in *SPSS* Regression

Flip back to the page above where we first displayed the window for setting up the regression model that we wanted to fit and recall that we discussed the button at the bottom for **Statistics…** Note that there are three other buttons in the same location. If we click on the one labeled **Save…**, we get the following new window:



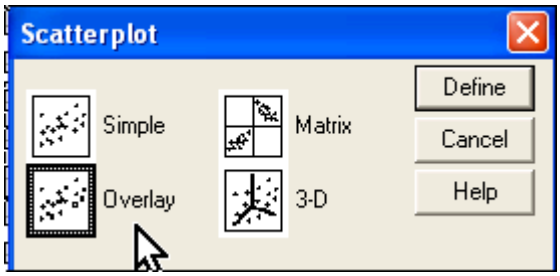
Later we will use some of the other features offered in this menu, but for now we need only check the boxes for **Unstandardized** under **Predicted Values** and those for both **Unstandardized** and **Standardized** under **Residuals**. Look now at the **Data Editor** and you see that three additional variables are created when the regression is executed. The variables are named **PRE_1**, **RES_1,** and **ZRE_1**.[2]



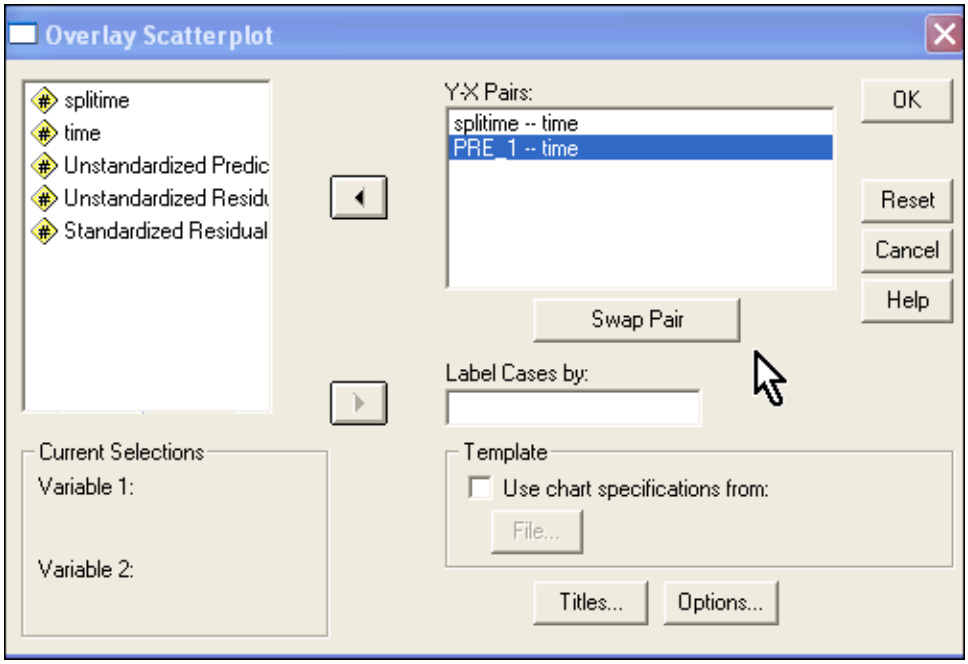
	splitime	time	PRE_1	RES_1	ZRE_1
1	65.35	1.00	66.93161	-1.58161	-1.63748
2	67.61	2.00	67.06116	.54884	.56823
3	67.68	3.00	67.19070	.48930	.50658
4	65.80	4.00	67.32024	-1.52024	-1.57394

[2] The addition to each variable name of "_1" is to indicate that these are the first predicted and residual values to be saved. Later on we may want to fit another model of a different form and save predicted and residual values from that model to compare with those from the first. Those new variables will carry the label "_2". If confusing now, this will all become clearer after we have discussed multiple regression models.

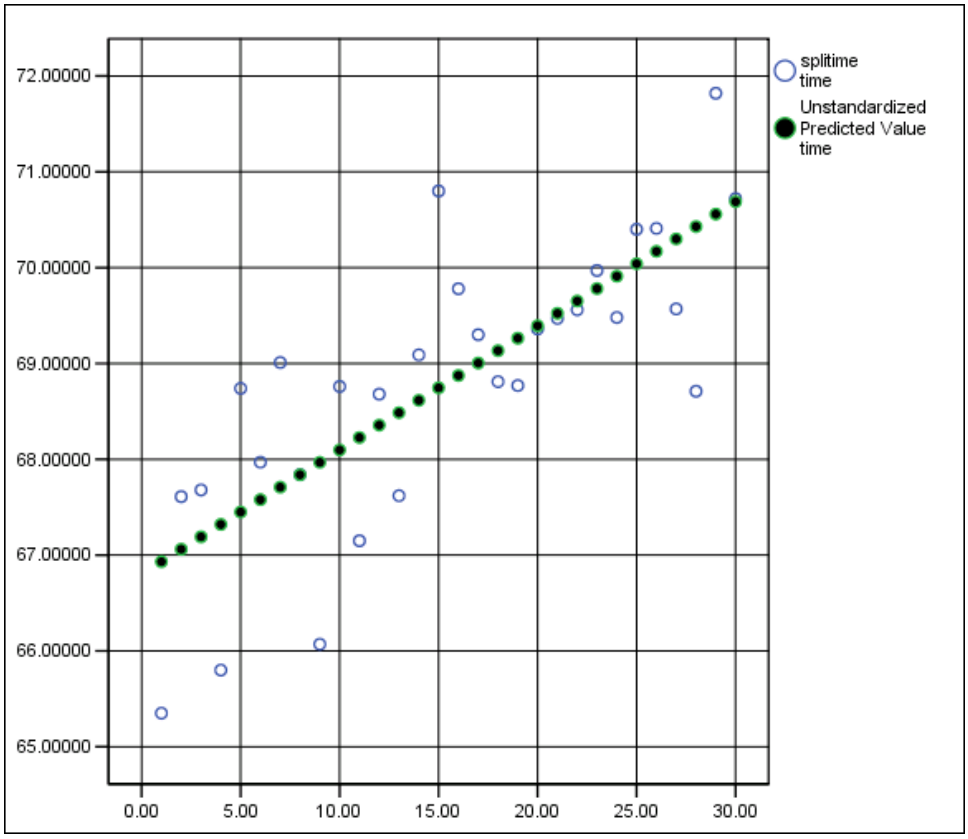
PRE_1 stands for **predicted**, i.e., the height of the regression line for any given value of **time**. To show this graphically, let's return to the sequence **Graphs/Scatter…** This time, however, in the window below, instead of highlighting the icon for **Simple** we highlight **Overlay**.



Next, we set up the dialog window as follows. Note that because of an idiosyncracy of ***SPSS*** you will have to use the button for **Swap Pair** to make the setup just right:



Here is our new scatter plot:



At first, when you create the plot above on your own PC, the little circles showing the predicted line will not be filled, and except for having a different color they may be difficult to distinguish from the regular data points. We have filled them in with the color black by right-clicking on the chart, and then left-clicking on **SPSS Chart Object** followed by **Open**. This action opens the **Chart Editor**, which makes available many options for "dressing up" the chart that we shall illustrate from time-to-time in class.

Be sure to read the Appendix after you have finished this chapter for further important discussion of the predicted values and their meaning.

Residuals

Next, let us consider the deviations or **residuals,** labeled **RES_1**, of actual **splitime** from the values of **PRE_1** given by the regression line. The basic relationship is:

RESIDUAL = ACTUAL SPLITIME – PREDICTED SPLITIME ,

RES_1 = splitime – PRE_1

or, in general:

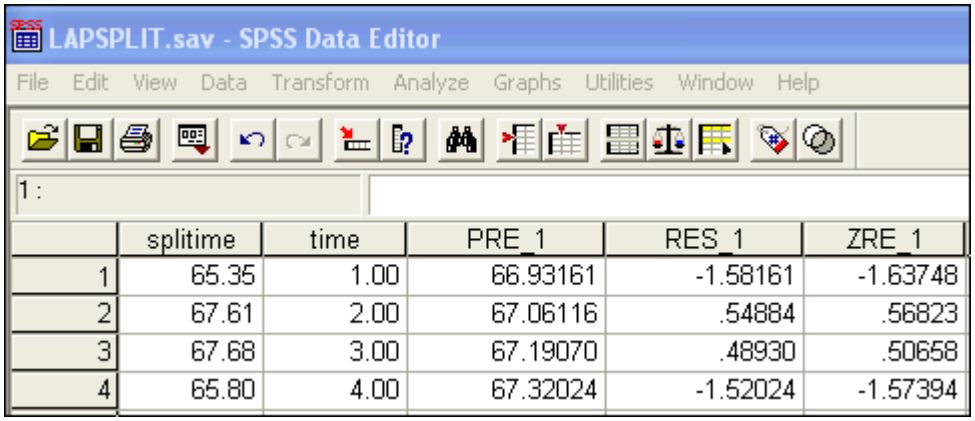
RESIDUAL=ACTUAL - PREDICTED

It follows from the above definition of **RESIDUAL** that we can write

ACTUAL = PREDICTED + RESIDUAL

This statement that the actual value of a variable of interest is made up of two parts-- the first **fitted through statistical modelling**, and the second part **a residual that is left over** and, so to speak, "unexplained" by the model fitting-- can well serve as a slogan for this course. Indeed, almost all that we shall be doing in the rest of ***STM*** can be thought of as involving the analysis of a variable into a **fitted** and a **residual** part, with the aim of reducing our uncertainty about "the way that the process works."

To clarify this point we repeat the abbreviated image of **Data Editor** from above:



	splitime	time	PRE_1	RES_1	ZRE_1
1	65.35	1.00	66.93161	-1.58161	-1.63748
2	67.61	2.00	67.06116	.54884	.56823
3	67.68	3.00	67.19070	.48930	.50658
4	65.80	4.00	67.32024	-1.52024	-1.57394

You should verify with your hand calculator that for **time** = 1, **RES_1** is given by

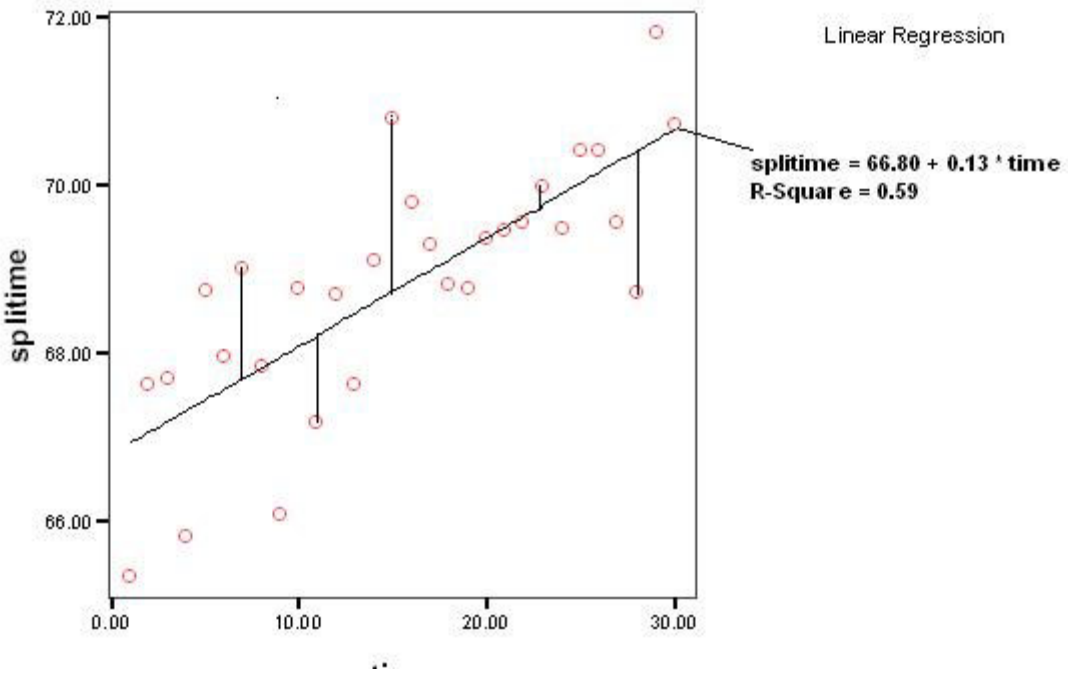
65.35 – 66.93 = -1.58 .

The first 1/8 mile segment was timed in 65.35 seconds, which was 1.58 seconds faster than the fitted value, 66.93. Hence this residual is negative, reflecting the fact that the actual time was below the regression line.

Similarly, the third value of **RES_1** is **67.38 – 67.19 = 0.51**, a positive value because **splitime** is above the regression line. By the way, an interesting fact resulting from the method of fitting the line is that the sum of all of the residuals is **zero**.

Geometric Interpretation of Residuals

Let us now make a scatter plot with both **splitime** and **PRE_1** plotted against **time** to show the geometrical meaning of these residuals. **Each residual can be represented by the signed length of a vertical line segment connecting the values of splitime and of PRE_1**. We have drawn a few of these line segments in the plot below:



The Principle of Fitting the Line

Now we turn to the principle of fitting the regression line. The parameters of **our particular line** are computed so that **the sum of squared residuals over all the data points is minimized.** For this reason, the standard regression fitting approach used by ***SPSS*** is called **the method of least squares**. There are many different straight lines that could be passed through the points in the plot above, but there is **only one** that minimizes the sum of squared residuals, and the calculation of its parameters, **a** and **b**, i.e., the intercept and slope, is a straightforward calculus problem.

The method of least squares is not the only method that could have been used to obtain the fitted line. It is, however, the most widely used method, and it works well in a wide variety of applications. (Even if you know more sophisticated methods of fitting, least squares is a good place to start in almost any practical application.) If the residuals behave like in-control observations that are also normally distributed, the method of least squares is especially attractive.

Diagnostic Checking

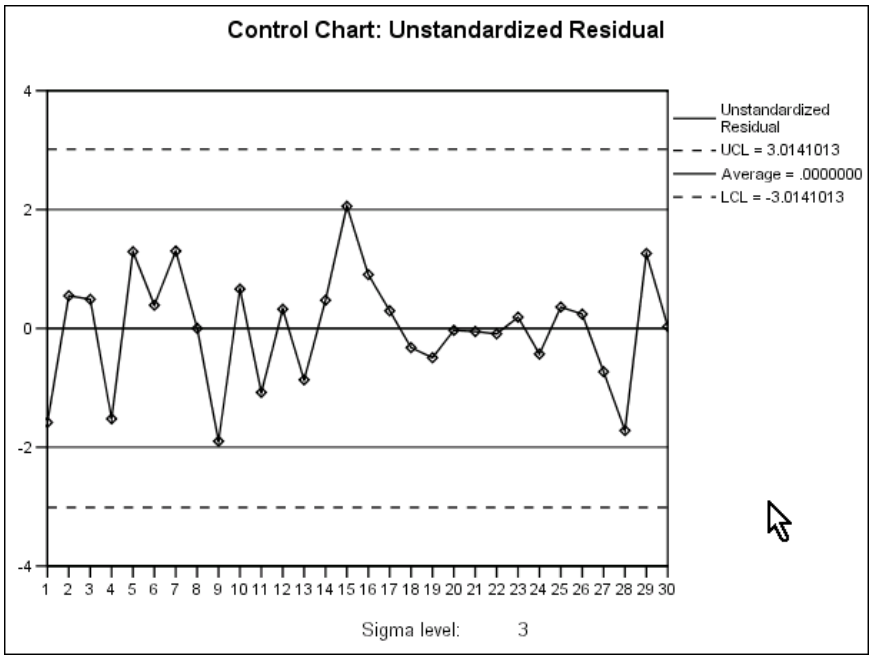
If the regression assumptions are satisfied, then the residuals should behave like any other random, normally distributed, data. Hence we can look at the residuals from our regression from the same perspectives used in looking at the original observations in the applications in Chapter 2 and 3:

We shall refer to this residual analysis as "**diagnostic checking**", because data analysis of residuals is designed to check whether the actual data conform to the regression assumptions:

(1) The regression relationship is a straight line.

(2) The residuals about the line should behave as a process in statistical control with an approximately normal histogram[3].

We'll begin with a control chart for individuals, using the default version in which ***SPSS*** bases control limits on the moving range method of estimating the standard deviation.



Observe that the center line is located at zero. **The mean value of the residuals is always zero**. The method of least squares estimation makes it turn out that way. Note also that when we specify **RES_1** to be plotted, ***SPSS*** knows that it is the residuals from regression and it therefore adds horizontal lines at plus and minus two standard deviations—a kind of alarm feature.

Runs Test

	Unstandardized Residual
Test Value[a]	.0000000
Cases < Test Value	13
Cases >= Test Value	17
Total Cases	30
Number of Runs	16
Z	.000
Asymp. Sig. (2-tailed)	1.000

a. Mean

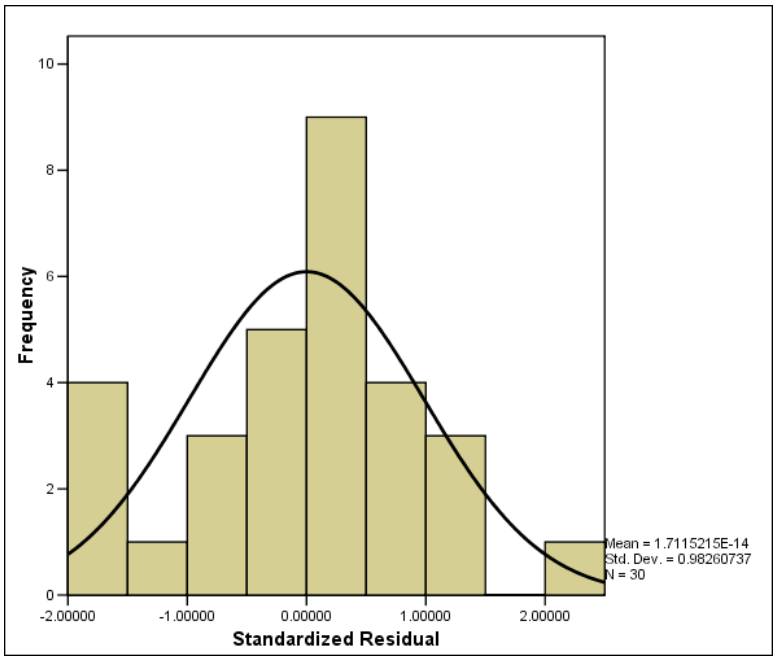
The results of the runs test also look very good.

[3]Students may legitimately ask at this point, "**Why** do these regression assumptions have to be satisfied?" One answer is that the residuals represent that part of the process that we are still uncertain about. A common way of talking about statistical modeling is that the fitted values "explain" the dependent variable, and the residuals are "unexplained." If the residuals were always zero there would be no uncertainty about **splitime**, for example. We would be able to predict its value perfectly by just plugging **time** into the regression equation. Since the residuals stand for the unexplained part of the process, and hence are fraught with uncertainty, we are interested in getting a statistical handle on that uncertainty by means of probability statements. We might, for example, need to answer questions such as, "When **time** = 35, what is the chance that the **splitime** will be greater than 75 seconds?" This is really a question about how big the residual is likely to be. Having residuals that are in statistical control, that is, random with constant mean and standard deviation, and approximately normally distributed, makes it easier to answer such questions.

Standardized Residuals for Checking Normality

Now we must explain one other aspect of the regression output. When we called for the computation and storage of **PRE_1** and **RES_1** after the regression parameter estimates were calculated, we also asked for the standardized residuals, which were named **ZRE_1**.

We can think of **ZRE_1** (approximately, but not exactly) as the kind of standardized residuals that would be produced if we had asked for z-scores to be saved in the dialog window for **Analyze/Descriptive Statistics/ Descriptive…** Hence we can proceed with our usual graphical check for approximate normality **of residuals**, using the histogram command in the usual way:



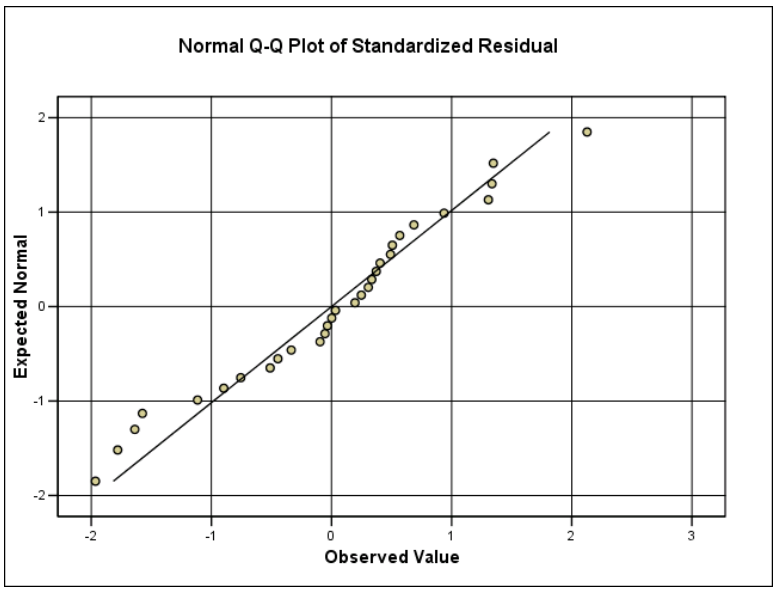
Although 1 of 30 values of **ZRE_1** is outside plus or minus two standard deviations, there is little reason to suspect the occurrence of special causes since 1/30 is only about 3 percent. Here are the Q-Q plot and the Shapiro-Wilk test results:

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.128	30	.200*	.970	30	.527

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

a. Lilliefors Significance Correction



We have now seen that residual behavior is roughly consistent with an in-control process and a normal distribution. There is therefore no evidence of special causes at work.

In summary, we have found one **systematic cause** -- the fatigue effect associated with **time** -- but no **special causes**. The analysis illustrates our general regression strategy:

- To explore for systematic causes that affect all observations. These systematic causes are represented by independent variables in the regression.

- Then to look for special causes, affecting only individual observations or groups of observations. For this purpose, we study the behavior of residuals from the regression.

Interpretation of Trend

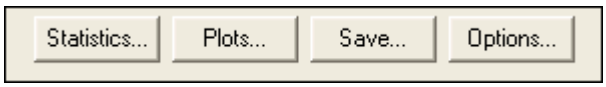
The systematic cause isolated in this application is called a "trend". Here the trend is so obvious that we haven't discussed its statistical significance; the procedure for doing that is given in Section 2.

The presence of trend does not tell **why** the trend is there. To decide about "why", we have to draw on other knowledge and, in many applications, further investigation. In the running example, it is reasonable to ascribe the trend effect to "fatigue": constant effort seems to result in gradually slowing performance as fatigue accumulates.

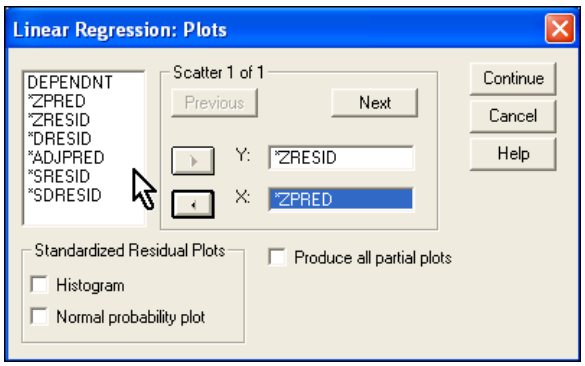
In running races, there is evidence to suggest that an even pace is the best strategy. For example, most world records in distance running were run at nearly an even pace. Our study suggests, therefore, that to run an even pace, one must start easily and gradually increase effort.

Diagnostic Checking for Linearity

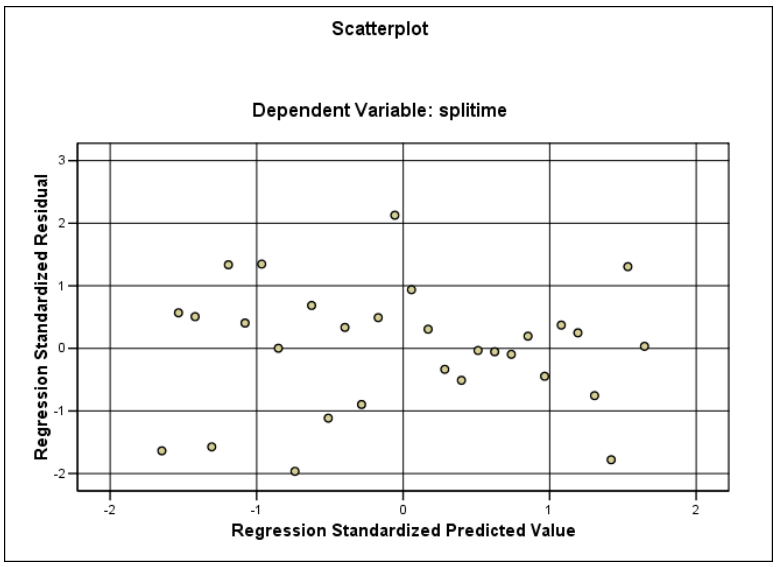
For diagnostic checking of the regression assumptions, there is another useful residual analysis based on a plot-- the scatter plot of **ZRE_1** versus the standardized values of **PRE_1**. This plot is obtained through the original setup of the regression model. Recall that in the dialog window for linear regression there are four buttons at the bottom:



Clicking on **Plots…**opens this window:



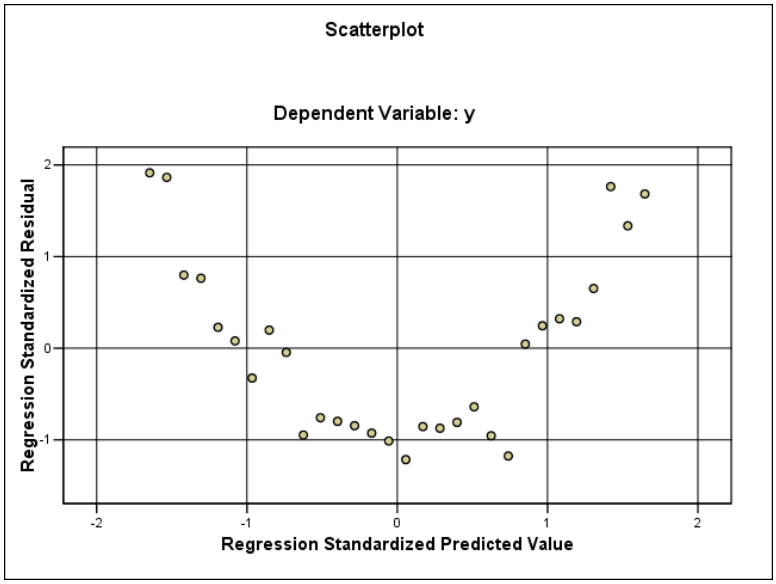
We have moved two variables from the list on the left-hand side to plot **Y = *ZRESID** against **X = *ZPRED**. (Do not be disturbed by the slightly different naming used by ***SPSS*** in this feature.)



This plot serves as a check on the assumption that the trend is linear. If that assumption is tenable, as it is here, the visual impression should be what you would see on a control chart of an in-control process with normally distributed deviations from the center line. Here the horizontal line at height zero, the mean of all residuals, is analogous to the center line. If **RES_1** is regressed on **PRE_1** in any regression model, no matter how complex, the fitted line will always be horizontal (i.e., the slope is zero), which is another way of showing that the method of least squares does not allow a linear relationship between the residuals and the predicted values. A perfectly horizontal regression line, however, does not prevent there being some residuals that appear to be seriously out of control. It is evidence of such out-of-control residuals that we are looking for here.

Back in Chapter 2 when we were discussing histograms and plots to check on normality, we said that to decide what is acceptable or not it is good to keep a picture in your mind of plots where the assumptions are clearly violated. Here, for example, is a case of some data created by

us through simulation that are much better described as a quadratic (curvilinear) function of **time** rather than simple linear. Suppose that we try the simple linear model nevertheless, and plot ***ZRESID** vs. ***ZPREDID**:



This plot ought to convince you that the plot of residuals against fitted values can be useful in regression diagnostics[4]

Regression and Prediction

Before leaving this application, there is another aspect of regression that can be developed. Here we were primarily interested in the trend for its value in better understanding the process. But the trend is also a tool for predicting the dependent variable given any particular value of the independent variable.

To see how this works, think back again to simple control chart analysis. We can always interpret the center line as a "best guess" as to what the next observation will be. The uncertainty about that best guess can be expressed in terms of upper and lower control limits, each three standard deviations removed from the center line.

If, contrary to fact, **splitime** had been in a state of statistical control, we would have predicted the next **splitime** with the **overall mean** as a "best guess" and the control limits around the best guess based on the standard deviation of all observations. The numbers we need come from the **Analyze/ Descriptive Statistics/Descriptive…** sequence executed in Section 5 of Chapter 3:

[4]Experienced statisticians (including you at the end of this course) would comment that the simple linear model never should have made it to the post-regression diagnostic stage, because a simple scatter plot of the dependent variable vs. time would have shown the quadratic nature of the relationship. **Moral: Always look at your data before you leap into an analysis.**

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
SPLITIME	30	65.35	71.82	68.8100	1.48369
Valid N (listwise)	30				

Thus, a prediction of the next point, for segment 31, (assuming continuation of the workout) would have been 68.81 seconds plus or minus 3*1.484 = 4.45 seconds. This would not have been a good prediction because the process was not in a state of statistical control. By the end of the workout, the runner was running slower than his average for the whole workout. Regression offers a prediction that is better in two respects:

- The regression "best guess" takes the trend into account: substitute **time** = 31 into the regression equation 66.8 + 0.130 **time** to get 66.8 + 4.0 = 70.8 seconds. Given the runner's tendency to slow down, this is a more realistic best guess than the overall mean of 68.8.

- The standard deviation 1.484 of **splitime** is too large to serve as the basis for realistic control limits, because it reflects the variation attributable to the trend effect. The appropriate computation for constructing control limits around the prediction of 70.8 is not 3 times 1.484, or 4.45, but 3 times the **standard deviation of residuals (**called by ***SPSS*** the **" Std. Error of the Estimate")**. We will refer to it and to its equivalent in more complex regression models as the **residual standard error**. This very important number is found in the second table that is produced by the linear regression output:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.769[a]	.591	.576	.96588

a. Predictors: (Constant), time

b. Dependent Variable: splitime

The key figure shown above is 0.96588. Hence for a better prediction, we have 70.8 plus or minus 3 times 0.9659, or 70.8 plus or minus 2.90.

The fact that the **residual standard error** is lower than the original standard deviation of the dependent variable is often expressed by saying that the regression equation has **explained** some of the variation originally present in the data. Another way to say it is that the **residual standard error** is a measure of the uncertainty in predicting **splitime** after the trend is removed.

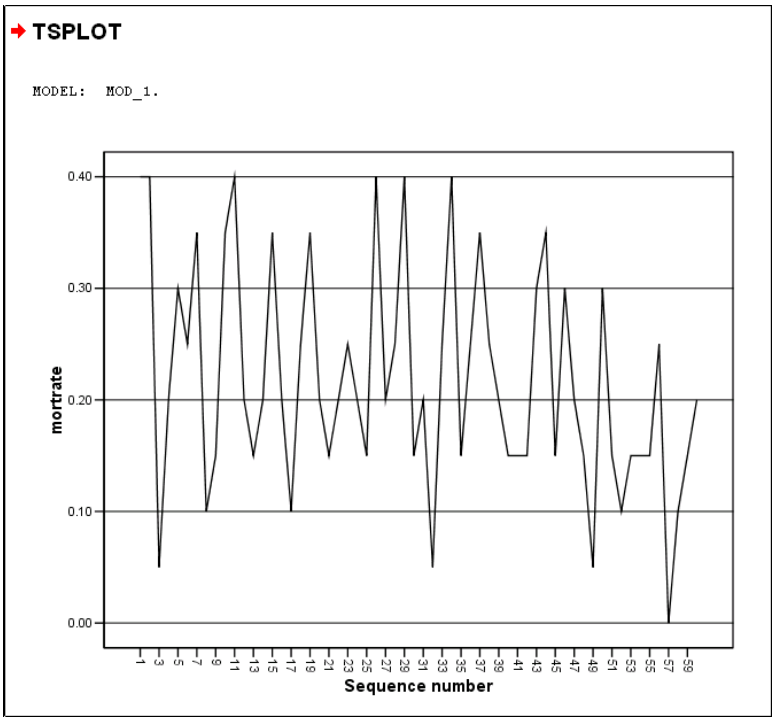
- We defer until later an explanation of **R Square**. **R** is the **correlation coefficient**, and **R-Square** is its square. At the end of this chapter, we shall explain the correlation coefficient **R**.

2. Trend Fitting by Regression: the Intensive Care Example

We turn now to the data on mortality in intensive care in INTCARE.sav, where a trend appears to be present, but is much less obvious visually. We didn't explain this back then, but when we were done with the session we saved the spreadsheet containing three variables: **mort**, the number of deaths; **n**, the sample size for each group; and **mortrate**, the mortality rate formed by dividing **mort** by **n**. The data are found in the file called INTCAREa.sav**.** You will find the ability to save the contents of **Data Editor** especially convenient if, in the course of your own project, you create many new variables through data transformations. If you do not save your work before you quit a session, when you return to the project you will have to start all over again repeating all of the necessary transformations.

We will tentatively assume that approximate normality is applicable here (even though the binomial distribution might have been slightly more accurate), and then check the normality assumption later in the analysis. The steps that follow are mainly a review of the approach shown in Section 1, but we shall make a few special comments as needed. Otherwise, you should follow the output closely to be sure that you understand it.

Here is the graph obtained from the ***SPSS*** procedure **Graphs/Sequence** applied to **mortrate**. The title, TSPLOT, stands for Time Series Plot, one of several useful tools for visualizing sequential data. If you picture the trend line, running from the beginning to the end of the plot, you can see that it is downward:



Now we turn to the regression fitting of trend. We first must create the new variable **time=$CASENUM**. We then apply the **Analyze/Regression/Linear…**sequence, using **mortrate** as the dependent variable and **time** as the independent variable. As shown in the previous example, we also save **RES_1, PRE_1,** and **ZRE_1** for diagnostic tests.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.325[a]	.106	.090	.09652

a. Predictors: (Constant), time

b. Dependent Variable: mortrate

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.064	1	.064	6.845	.011[a]
	Residual	.540	58	.009		
	Total	.604	59			

a. Predictors: (Constant), time

b. Dependent Variable: mortrate

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.275	.025		10.893	.000
	time	-.002	.001	-.325	-2.616	.011

a. Dependent Variable: mortrate

The first thing that we note is the equation for the fitted line:

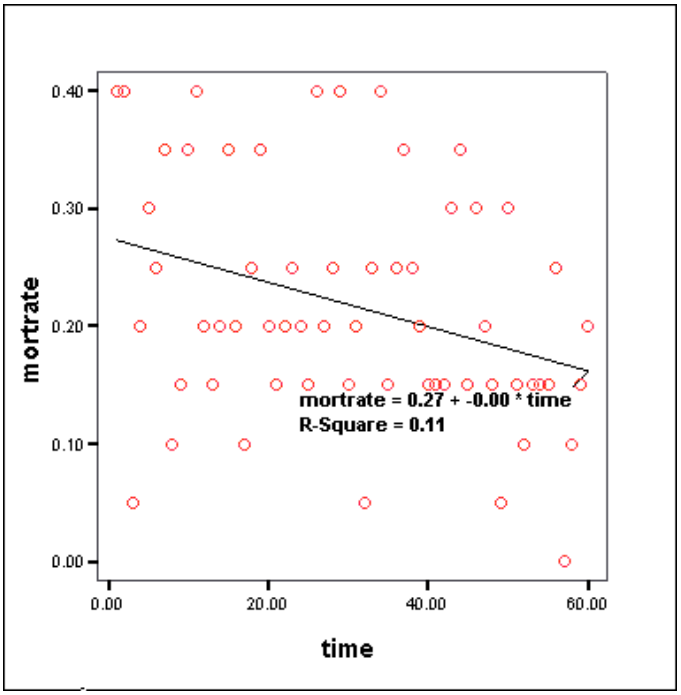
predicted mortrate = 0.275 - 0.002 **time**

This says that starting from a height of 0.275 (27.5 percent) at **time** = 0 (just before the first observation), the line trended down by 0.002 units for each observation, or about 0.12 units (12 percentage points) over the course of the 60 observations, that is: -0.002 * 60 = -0.12.

At face value, this translates to two extra survivors per each successive group of 20 patients. This would be obviously gratifying, but there are two important questions that we will investigate in a moment:

- Is the downtrend **statistically significant**, or could it reasonably be ascribed to the effects of chance variation in a sample of 60?
- If the trend is significant, what root causes lie behind it?

Before considering these questions, we go through the main outlines of the statistical analysis with only minimal commentary. We shall not explain which ***SPSS*** sequences were necessary to obtain these results. As a self-quiz, see if you can duplicate them.

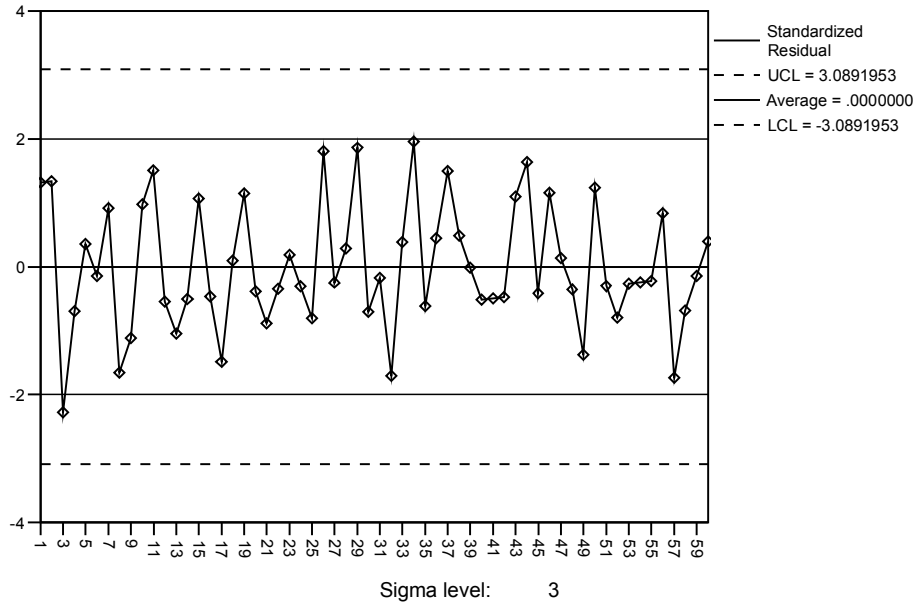


Runs Test

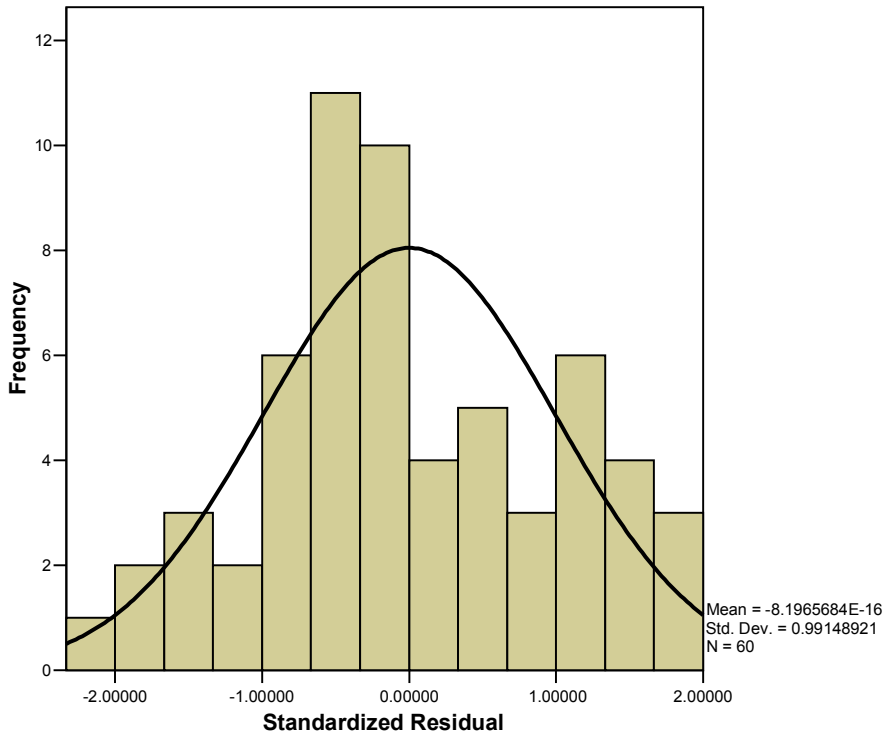
	Unstandardized Residual
Test Value[a]	.0000000
Cases < Test Value	35
Cases >= Test Value	25
Total Cases	60
Number of Runs	31
Z	.223
Asymp. Sig. (2-tailed)	.823

a. Mean

Control Chart: Standardized Residual



Sigma level: 3

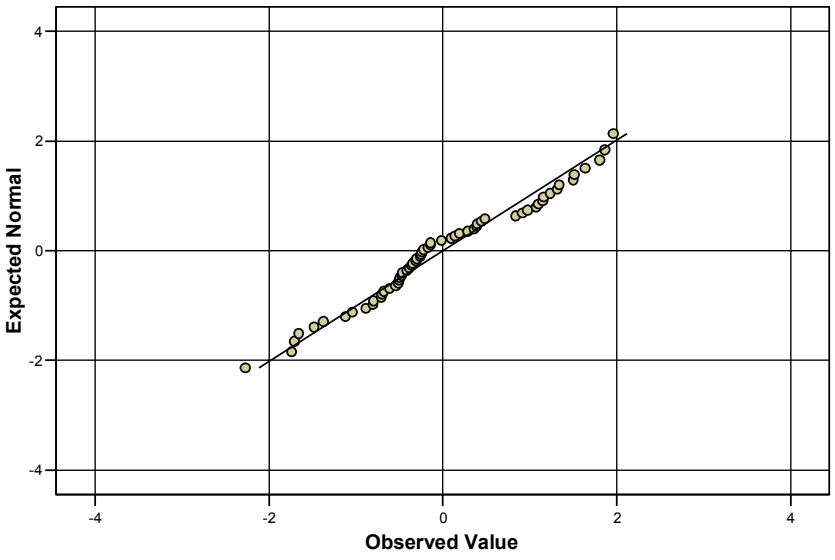


Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.123	60	.024	.972	60	.178

a Lilliefors Significance Correction

Normal Q-Q Plot of Standardized Residual



Overall, the fit looks good, and no special causes appear to be at work.

Statistical Significance of the Apparent Trend

We now return to the question of statistical significance of the downward trend. For that purpose, we display again some of the output from the simple linear regression:

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.275	.025		10.893	.000
	time	-.002	.001	-.325	-2.616	.011

a Dependent Variable: mortrate

The bottom row of the table above gives important information about the independent variable, **time**:

- The value -0.002 is the regression coefficient itself.

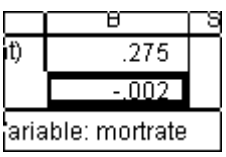
- The "Std. Error" -- 0.001 -- is a measure of uncertainty about the slope coefficient attributable to the fact that we are working with a sample of only 60 observations, and there is considerable variability even in the residuals from regression.

  This measure is a special kind of standard deviation. All of the statistical measures that we shall calculate from data analyses are estimates of model parameters and therefore have theoretical sampling distributions with a mean and a standard deviation. That is, we realize that from sample to sample the particular value of the estimate is likely to change. It is common to use the term "standard error" when referring to the standard deviation of the sampling distribution of any statistic of interest. Be careful, however, not to confuse this particular **standard error of a coefficient** with the **residual standard error** of the regression that we discussed above. They are related but not the same.

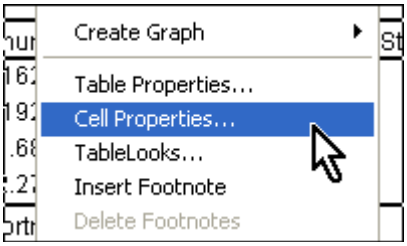
The numbers that are displayed in the coefficient table above are expressed with three decimal places. Thus the values -.002 and .001 for the slope and its standard error are rounded up from more precise figures. In order to clearly explain another important number in the output above, we must take advantage of one of the more sophisticated features of ***SPSS*** – the ability to reformat the table:

- Placing the mouse pointer within the table above, we double click and a border appears to indicate that we are in editing mode.

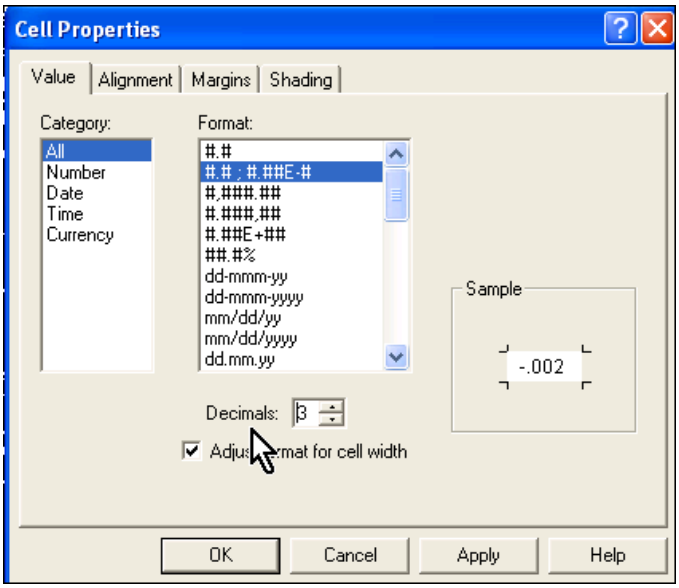
- We then click once with the left button on the slope coefficient, -.002, causing that cell to be encased in a heavy black border.



- Continuing, we place the pointer within the highlighted cell and right-click one time. This action produces the following menu. (We only show the center section here.)



- When we left-click on the highlighted selection, Cell Properties…, the window below appears:



- Finally, we change the number of decimals from 3 to 7 -- note the arrow in the image above.

After we have repeated the same reformatting operation for the **Std. Error** the coefficient table now looks like this:

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.275	.025		10.893	.000
	time	-.0018825	.0007195	-.325	-2.616	.011

a Dependent Variable: mortrate

We see that the slope and its standard error are more precisely expressed as -.0018825 and .0007195.

Now we can show with greater clarity the reasoning behind the concept of **statistical significance:**

- Suppose that there were in fact no trend at all: the "true slope" is assumed to be zero.
- Then the slope coefficient, -0.0018825 for this sample, is less than zero by a distance of 2.616 standard errors.

That is,

(-0.0018825 - 0)/0.0007195 = -2.616

The number -2.616 is shown in the table above under the column heading "**t".** We will often refer to this figure as the "t-ratio".[5]

The most common rule-of-thumb guide is that the slope coefficient is judged "statistically significant" if the absolute value of the t-ratio exceeds 2.

This requirement is more than satisfied by our coefficient (rounded) of -0.002. Hence by this rule-of-thumb, the negative slope of mortality rates would be judged statistically significant; the regression coefficient would be said to be "significantly less than zero".

"Significance" or "p-value"

A similar rule of thumb can be based on the final number, 0.011, which is in the column labeled "**Sig.**". (The term "p-value" means exactly the same thing as "significance". We have already used it in our discussion of the runs count, explained in Chapter 2.)

In this application, Sig. = 0.011 means:

- **If the true slope were 0 and we were looking only at sampling variation**, we would get a sample slope coefficient **at least as far from zero** as is 0.002 only about 11 times in a thousand.
- 11 times in a thousand would be considered "relatively unusual" **if the true slope were zero (in this application, no real trend).**

[5] There is no need to permanently change the number of decimals shown in your table settings. We have only done it here to better illustrate the calculation of the t-ratio. You see that the value of **t** is correct even though the other displayed figures are rounded.

- Hence the significance level, or p-value, can be thought of as a measure of unusualness: the smaller the p-value, the more unusual the result.

How unusual is the result for it to be judged statistically significant? Again, there is a conventional rule of thumb:

If Sig. is less than 0.05, the observed coefficient is deemed sufficiently unusual to be judged a significant departure from zero.

This rule of thumb comes to much the same thing as the conventional rule of thumb that requires that the absolute value of **t** exceed 2 for statistical significance.

Both rules of thumb are just that-- rules of thumb. Neither an absolute t-ratio greater than 2 nor a p-value less than 0.05 is so sacred an occurrence that the rule must apply without exception. As you gain experience with data analysis, you may wish to modify these rules of thumb with judgment. The rule of thumb gives rough guidance, however, and we recommend its use, especially by those who are just learning basic statistical techniques.

Significance of Trend in the Running Data

In the running data of Section 1, the t-ratio was 6.358, and the p-value as rounded to three decimal places, was 0.000, which means less than 5 in 10,000. Hence significance was overwhelming, so much so that we did not discuss it explicitly at the time, saying only that the visual evidence was strong.

Practical Interpretation of "Significance"

Three important ideas about the word "significance" merit discussion:

- In the context of our current application, "statistical significance" means **only** that we judge the difference between -0.002 and 0 as too large to be written off as a chance fluke. **We conclude that the decline of mortality rates in intensive care is real.**

- A **statistically significant** trend is not necessarily a **practically important** trend. Practical importance must be judged in the context of the application. In the intensive-care application, the downtrend of mortality rates, if real, would surely be of practical importance because two additional survivors for each group of 20 patients is clearly important (see the discussion above immediately following the first regression printout).

  In the running application of Section 1, the uptrend of running times in a constant-effort workout is surely worth knowing about, and is in that sense important.

- **A statistically significant trend does not prove a particular causal mechanism**. Mortality rates in the intensive care unit were improving, and that is an important fact. What lies behind the improvement, however, is not clear from the finding that the trend is significant. It could be better medical care, but it could be some factor that has nothing to do with medical care. For example, as time passed during the period of this study, perhaps the patients who were coming to the intensive care unit were, on average, less seriously ill.

  This last possibility can be investigated by further statistical analysis since there is additional information about the average severity of illness for each group of 20 patients. In Section 5 we present an analysis that takes this information into account. That analysis will rule out the possibility that the improvement is an artifact of changes in the seriousness of illness of the patients coming to the intensive care unit.

  But the root cause of the improvement is not known, and further study of the intensive-care process remains an important task for future research.

3. Fitting Periodic Effects by Regression: the Ishikawa data

Recall the data set contained in ISHIKAWA.sav, discussed in Section 6 of Chapter 3:

First 50 observations from Ishikawa, *Guide to Quality Control*, Table 7.2, page 66. Each successive five observations represent measures taken at successive times on the same data: 6:00, 10:00, 14:00, 18:00, and 22:00. Thus the 50 observations represent 10 days, which we shall assume to be consecutive working days. The original data set has 75 further observations for another 25 working days; to save space we will not look at these at this time.

When first examined, this data set initially appeared to be in a state of statistical control and approximately normally distributed. Then, after more detailed analysis, we discovered that the first period of each day tended to be higher than the observations for the subsequent four periods:

"Indicator" Independent Variables for Periodic Effects

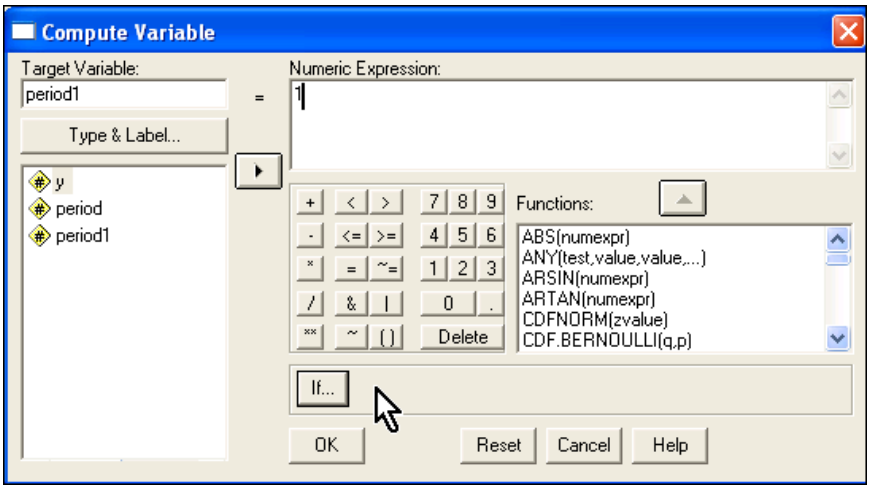
Now we come to the critical step. In our modeling of possible time trends, we have created an independent variable **time** -- 1, 2, 3, ... -- that permits us to use regression to estimate the trend and test its significance. We need to do a parallel step to create an independent variable for a possible first-period effect. We begin by typing in by hand the values of a new variable, **period**-- 1,2,3,4,5,1,2,3,4,5,1,…,etc.-- just as we did back in Chapter 3:

	y	period
1	14.00	1
2	12.60	2
3	13.20	3
4	13.10	4
5	12.10	5
6	13.20	1
7	13.30	2

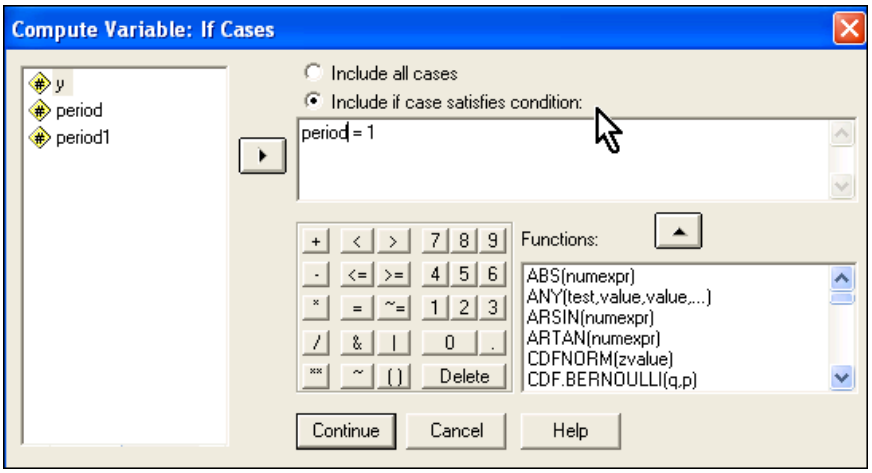
The next step is to set up an **indicator variable** named **period1** that takes just two possible values, 0 and 1. For an observation taken at the first period of the day, **period1** = 1; for observations taken at any of the four other periods, **period1** = 0.

We accomplish this via **Transform/Compute…:**

- At the first step define **period1 = 0** for all cases.
- Then repeat **Transform/Compute…** as shown below:

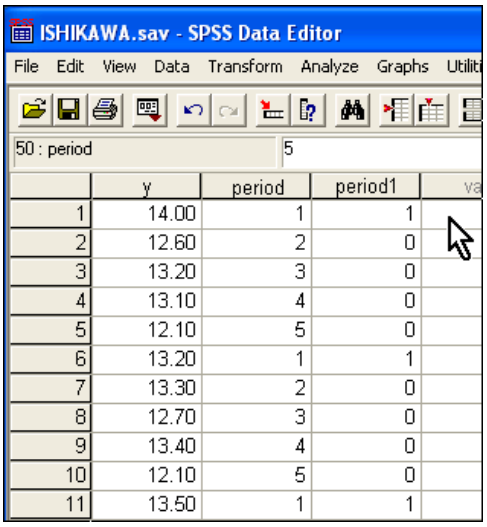


- Note that we are resetting **period1** to the value 1. Before we click on **OK**, however, we must instead click on the **If…** button (see arrow). That action opens this window:



- Note carefully that we have done two things here. First, we marked the little circle for **Include if case satisfies condition:.** Then we typed in the condition, **period=1**. Recall that the new variable **period1** is already set to zero for all cases. With the two windows above set up as displayed, **period1** will be changed to the value 1, but only if **period=1**.

After we hit **Continue** and the subsequent **OK** button the **Data Editor** looks like this:



	y	period	period1	va
1	14.00	1	1	
2	12.60	2	0	
3	13.20	3	0	
4	13.10	4	0	
5	12.10	5	0	
6	13.20	1	1	
7	13.30	2	0	
8	12.70	3	0	
9	13.40	4	0	
10	12.10	5	0	
11	13.50	1	1	

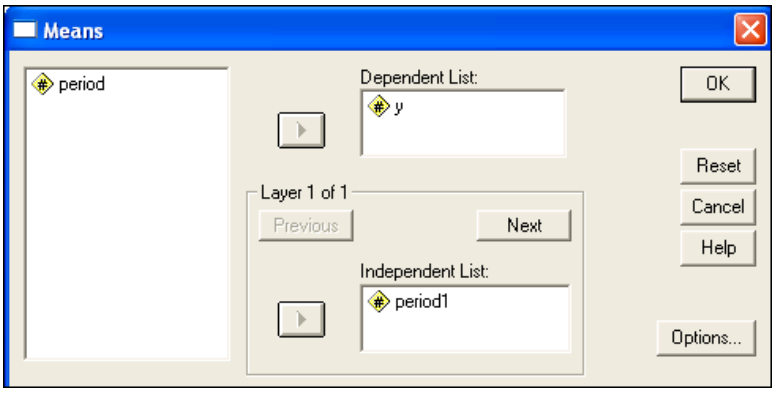
Note that we have hit the **Variable View** tab and changed the number of decimal places for **period** and **period1** to 0.

Here are the descriptive statistics for **y** that we ran back in Chapter 3:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
y	50	12.00	14.40	12.9120	.57237
Valid N (listwise)	50				

Next, we execute the ***SPSS*** sequence **Analyze/Compare Means/Means…**which opens this window:



Specifying **y** as the dependent and **period1** as the independent variable, we obtain this display:

y

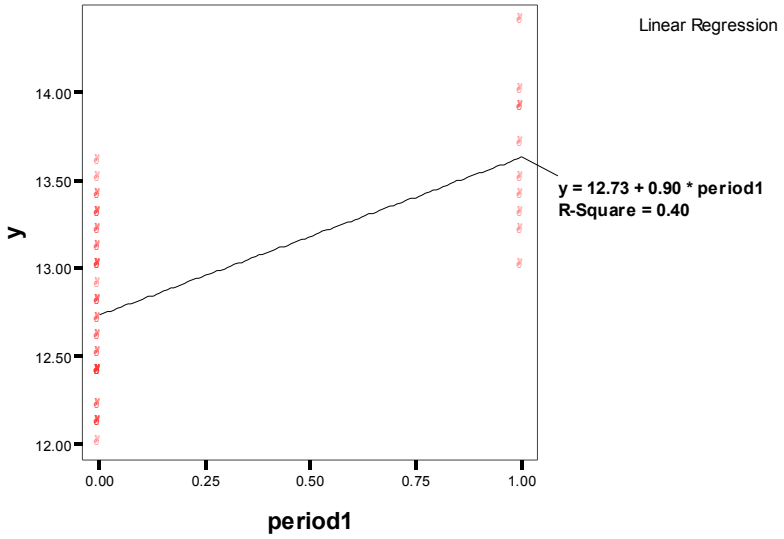
period1	Mean	N	Std. Deviation
0	12.7325	40	.45200
1	13.6300	10	.42701
Total	12.9120	50	.57237

- When **period1** = 1, the mean of **y** is 13.63

- When **period1** = 0, the mean of **y** is 12.73.

Thus the mean readings are higher for **period1** by 13.6300 - 12.7325 = 0.8975 which can be rounded up to 0.90.

We can get the same information, and more, from regression analysis. We begin with an interactive scatter plot of **y** against the indicator variable **period1** in which we also ask for the regression line to be plotted:



There are just two vertical columns of observations.

- At the left are the 40 observations at periods 2, 3, 4, 5 (**period1** = 0), which extend from about 12.0 to 13.60

- At the right are the 10 observations at period 1 (**period1** = 1), which extend from 13 to 14.4.

There is only a partial overlap between the **y** values corresponding to the two columns. Intuitively, that suggests that the difference between the first period average and the average of the other four periods is going to be significant.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.634[a]	.401	.389	.44742

a. Predictors: (Constant), period1

b. Dependent Variable: y

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.444	1	6.444	32.191	.000[a]
	Residual	9.609	48	.200		
	Total	16.053	49			

a. Predictors: (Constant), period1

b. Dependent Variable: y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	12.732	.071		179.983	.000
	period1	.898	.158	.634	5.674	.000

a. Dependent Variable: y

Here is the key result from the display above:

The regression equation is

predicted y = 12.732 + 0.898 period1 ,

which agrees with the note in the scatter plot above.

The salient conclusions are:

- In the regression equation, the constant 12.732 is just the mean of **y** when **period1** = 0.
- The regression coefficient, 0.898, is the difference of means between **period1** = 0 and **period1** = 1.
- The t-ratio 5.674 is much greater than 2.
- The regression coefficient is clearly significant by the rule of thumb for **t** introduced above.
- "Sig. = 0.000" means that the actual p-value rounds to 0.000 when only three decimal places can be shown; that is, the actual value must be less than 0.0005. If there were no true **period1** effect, there would be fewer than five chances in 10,000 of a difference between the actual regression coefficient and zero that is at least as large as 0.898.

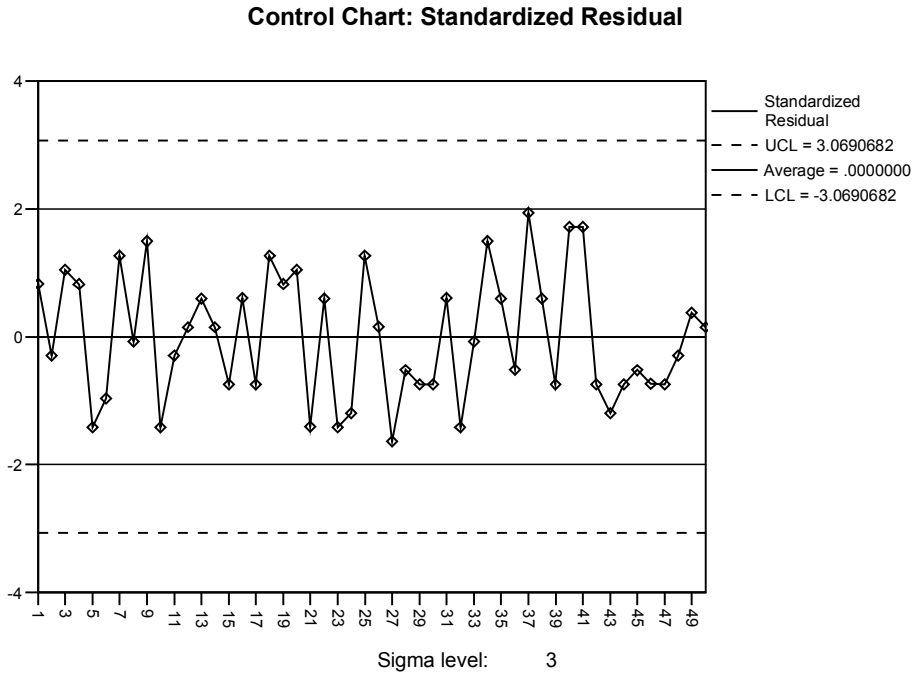
- 0.0005 is much less than 0.05, so that by this rule of thumb as well, the first period difference is **statistically significant.**

We can't tell whether the difference is practically important, since we have no descriptive background on the example from our source, but it is may well be so.

Notice in the plot on the previous page that the regression line passes through the means for each group. Since the distance between the two values of **period1** is one unit, the slope of the line is equal to the difference between the means.

Diagnostic Checking

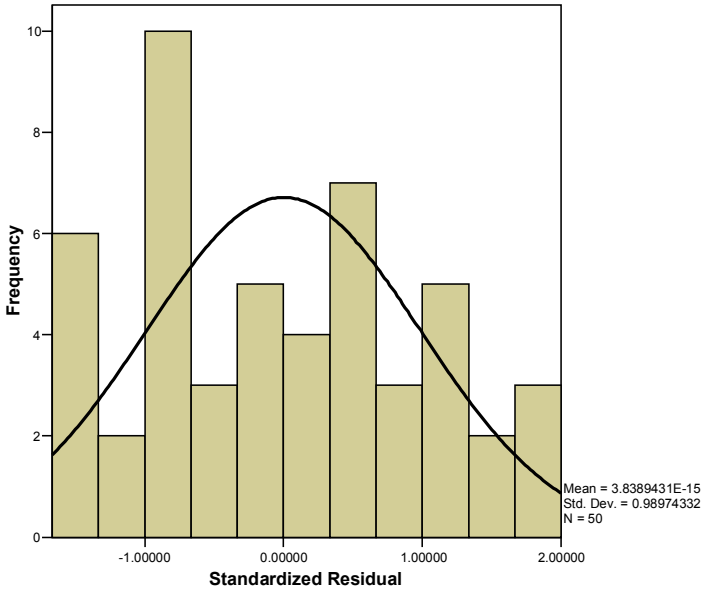
Next we look at the behavior of regression residuals as a means of checking diagnostically whether the assumptions of the regression model appear to hold in this application. (We will always try to follow more or less the same sequence in doing these checks, so that you will become familiar with them. When more checks are added later, we'll fit those in.)



Runs Test

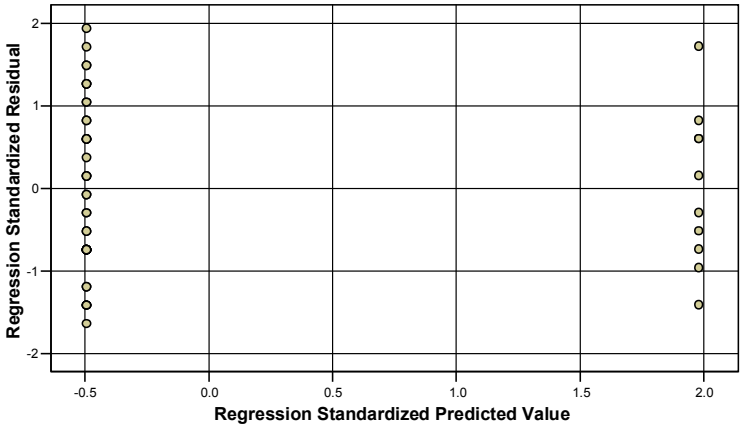
	Standardized Residual
Test Value(a)	.0000000
Cases < Test Value	26
Cases >= Test Value	24
Total Cases	50
Number of Runs	27
Z	.298
Asymp. Sig. (2-tailed)	.766

a Mean



Scatterplot

Dependent Variable: y

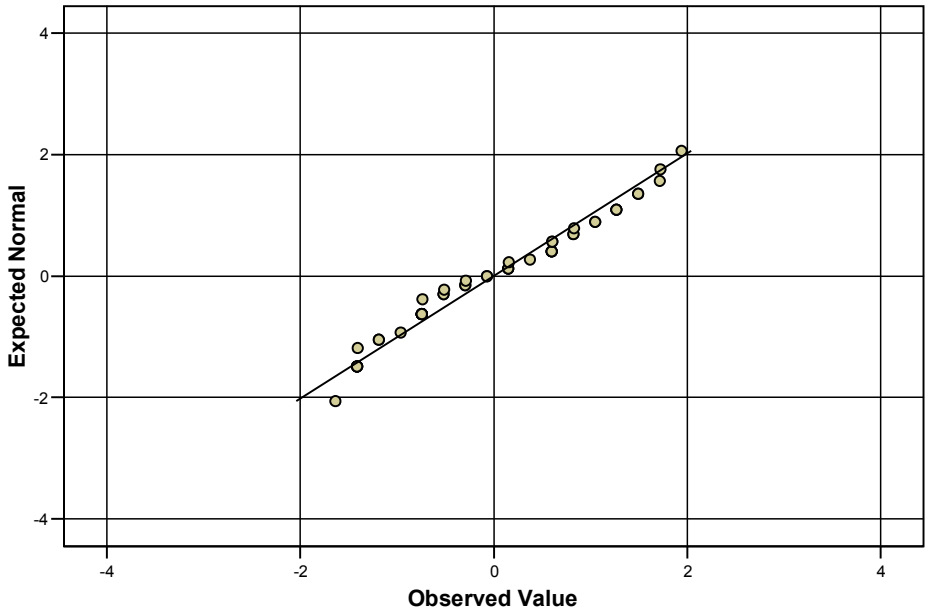


Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.132	50	.030	.953	50	**.045**

a Lilliefors Significance Correction

Normal Q-Q Plot of Standardized Residual



Although the residuals appear to be random, they fail the Shapiro-Wilk test for normality. (See highlighted value of Sig.) We shall consider this to be a minor problem, however, and continue on to the next example.

4. Fitting Intervention Effects by Regression: the Putting Data

Recall the example from Section 7 of Chapter 3:

> Indoor putting experiment. Putts sunk per 100 trials from fixed distance. At the end of the first 10 groups of 100, it was noticed that 136 of 443 misses were left misses and 307 were right misses. It was reasoned that the position of the ball relative to the putting stance was a problem. "I concluded that the ball was too far "back" (too much in the middle) of my putting stance. I moved the ball several inches forward in my

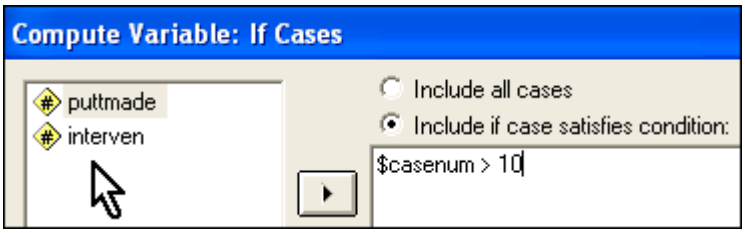
stance, keeping just inside my left toe." The final 10 observations were made with the modified stance.

The variable of interest is **puttmade**, the percentage of successful shots out of each group of 100. We now need to create an indicator variable that shows if the putts were shot before or after the change in stance. We will call the new variable **interven**, for "intervention". We want the variable to have the value "0" for the first ten groups of putts, and "1" for the last ten groups.

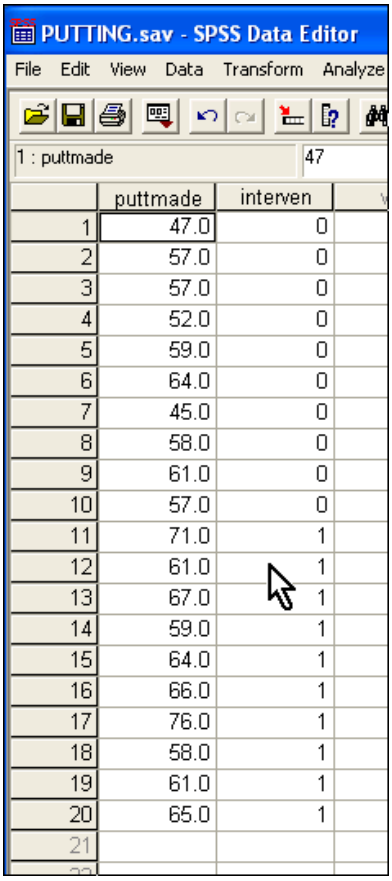
As in the last example, we begin by setting the new variable equal to zero for all cases:

interven = 0

Then we repeat the sequence **Transform/Compute…,** but this time setting **interven = 1** subject to the **If…**condition shown below:



After we execute the transformation and change the number of decimal places for **interven** to zero the **Data Editor** looks like this:



	puttmade	interven
1	47.0	0
2	57.0	0
3	57.0	0
4	52.0	0
5	59.0	0
6	64.0	0
7	45.0	0
8	58.0	0
9	61.0	0
10	57.0	0
11	71.0	1
12	61.0	1
13	67.0	1
14	59.0	1
15	64.0	1
16	66.0	1
17	76.0	1
18	58.0	1
19	61.0	1
20	65.0	1
21		

Here are the means and standard deviations for each set of putts-- **before vs. after** the intervention:

Report

puttmade

interven	Mean	N	Std. Deviation
0	55.700	10	5.9824
1	64.800	10	5.5737
Total	60.250	20	7.3117

The following displays, in the same spirit as our analysis of the ISHIKAWA.sav data, are presented without comment until the end:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.638[a]	.408	.375	5.7817

a. Predictors: (Constant), interven

b. Dependent Variable: puttmade

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	414.050	1	414.050	12.386	.002[a]
	Residual	601.700	18	33.428		
	Total	1015.750	19			

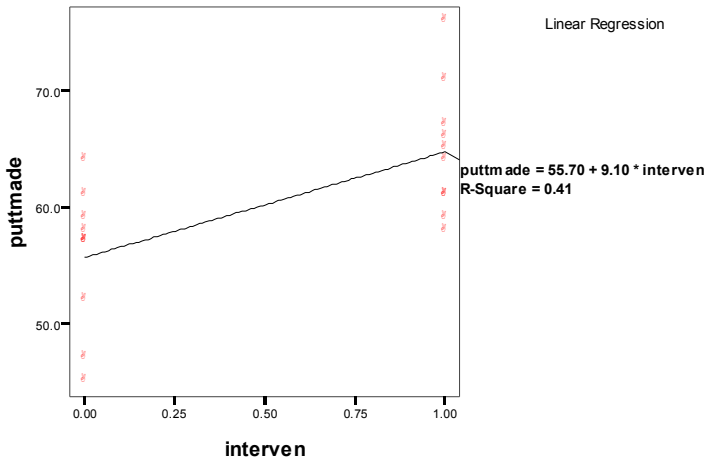
a. Predictors: (Constant), interven

b. Dependent Variable: puttmade

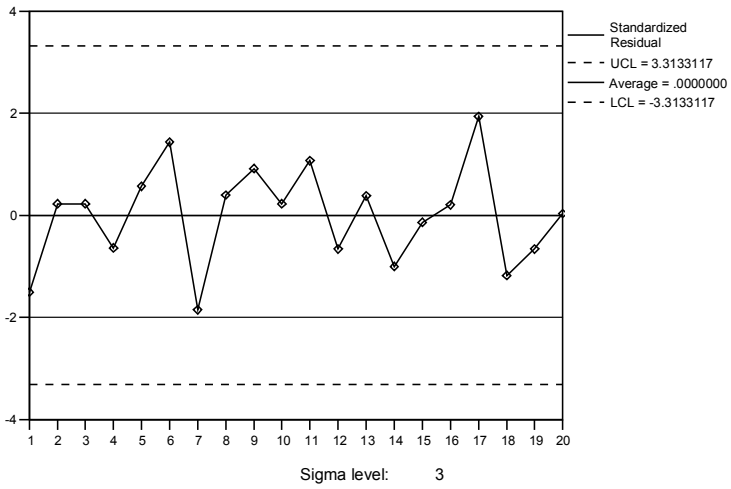
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	55.700	1.828		30.465	.000
	interven	9.100	2.586	.638	3.519	.002

a. Dependent Variable: puttmade



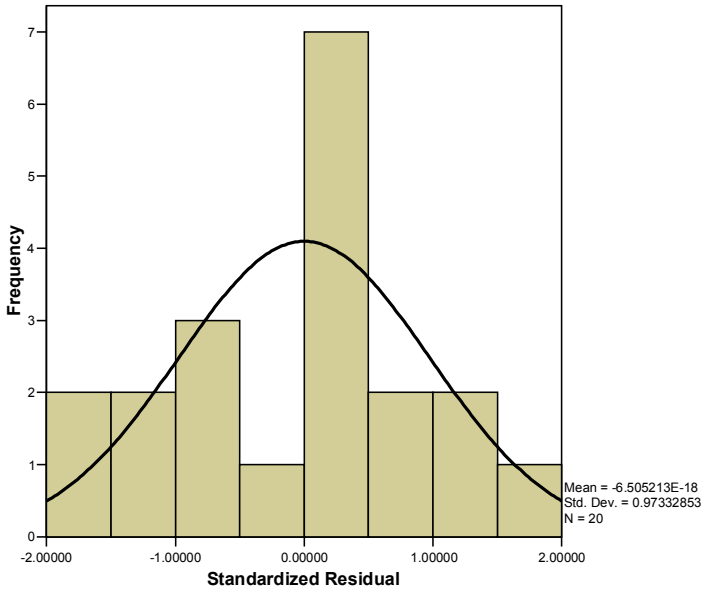
Control Chart: Standardized Residual



Runs Test

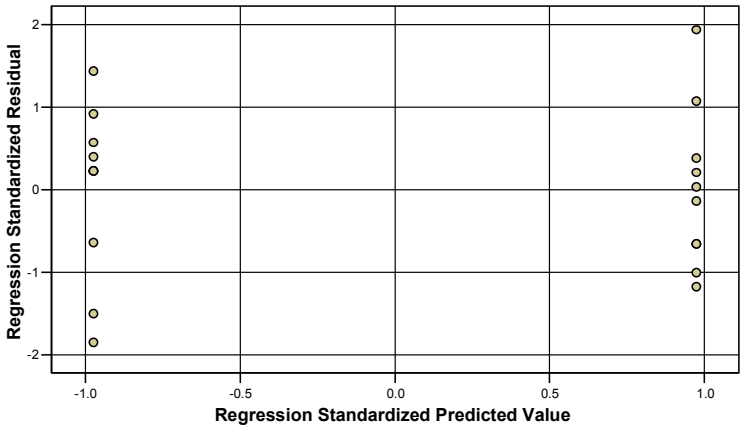
	Standardized Residual
Test Value(a)	.0000000
Cases < Test Value	8
Cases >= Test Value	12
Total Cases	20
Number of Runs	12
Z	.432
Asymp. Sig. (2-tailed)	**.666**

a Mean



Scatterplot

Dependent Variable: puttmade

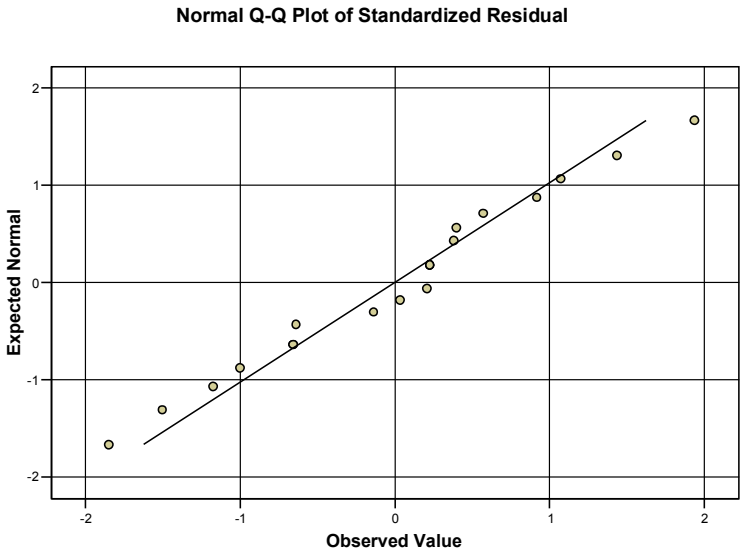


Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.134	20	.200(*)	.980	20	**.932**

* This is a lower bound of the true significance.

a Lilliefors Significance Correction



The increase of 9.1 points in the percentage of putts sunk after the intervention is clearly statistically significant, and nowhere in the above displays do we see anything to make us doubt the validity of the regression assumptions.

We have noted the similarity between this analysis and the previous, but one essential difference is that in the Ishikawa data, the effect that was studied was a periodic phenomenon, that is, it occurred in every first observation out of successive groups of five. We really do not have any good "causal" explanation for the phenomenon-- it is merely there. In the present putting example we have estimated an **intervention effect** that is a before and after phenomenon with respect to the imposition on the process of a possible cause, **the change in stance**. This raises the question of whether or not a regression analysis of the putting data would show evidence of a time trend, which would support the alternative hypothesis that the improvement is due to practice (i.e., a learning effect):

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.591(a)	.349	.313	**6.0613**

a Predictors: (Constant), time
b Dependent Variable: puttmade

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	52.584	2.816		18.676	.000
	time	.730	.235	.591	3.106	.006

a Dependent Variable: puttmade

We leave it for you to carry out the diagnostics for this regression to confirm that the necessary assumptions appear to be met. It looks like **time** is also a significant "explainer" of the behavior of the dependent variable **puttmade**. How can we decide which model is better?

We can compare the goodness of fit of the two regression models by comparing the **residual standard errors**, where smaller is better:

Std. Error of the Estimate	`5.7817`	(Intervention model) ✓
Std. Error of the Estimate	`6.0613`	(Time trend model)

Hence the data support the idea that the improvement can be attributed to the change of stance, not to a steady improvement with practice. (It may occur to you that we might do better using both **interven** and **time** in a multiple regression, an approach that will be introduced in the next section of this chapter. The answer, however, is that multiple regression does not significantly improve the fit.)

Where do we stand, then, on causal interpretation?

- The successful-intervention interpretation fits somewhat better than the improvement-with-practice interpretation.

- The background data, including especially the line of reasoning that led to the intervention in the first place, supports the intervention argument.

These two facts are helpful but not decisive. In the actual study, the student went ahead with further experimentation to **confirm** the causal interpretation: further data showed continued performance at the higher level attained during the final 10 trials above. (A confirmatory study like this is always desirable when statistical analysis of past data suggests that a quality improvement has been achieved.)

5. Introduction to Multiple Regression: Allowance for Severity in the Intensive Care Application

As we pointed out at the end of Section 4, regression need not be confined to use of a single independent variable. The intensive care application of Section 2 can be used to illustrate how we can proceed when there is more than one independent variable.

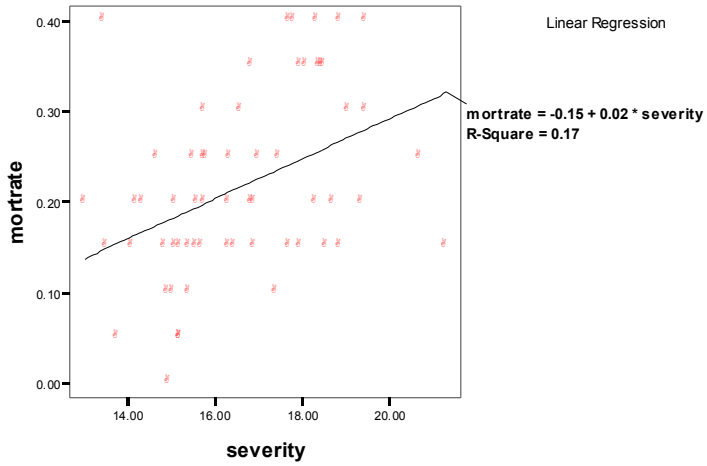
Recall that there was a significant downtrend in mortality, that is, the regression of **mortrate** on **time** produced a significantly negative regression coefficient for **time**. In reaching a causal inference, however, we would like to know about the severity of illness of the patients. If, through time, the intensive care patients were on average less sick, we could get a downtrend in mortality without any change in the treatment process.

For the same data previously analyzed, information was actually available on severity. We have it incorporated in the file that is named INTCAREb.sav.

The measures of severity for each case are average values of an index of mortality risk called the APACHE II score. When a patient is admitted to the intensive care unit, an assessment is made by the examining staff and an APACHE II score based on a set of basic indicators is

assigned. The values for the variable named **severity** in the data set are averages over the patients in each group of 20.

The scatter plot of **mortrate** on **severity** is sloping upward from the lower-left to the upper-right corner, suggesting a positive association between the two variables, which is confirmed by the regression output below:



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.409(a)	.168	.153	**.09311**

a Predictors: (Constant), severity

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.101	1	.101	11.681	.001(a)
	Residual	.503	58	.009		
	Total	.604	59			

a Predictors: (Constant), severity
b Dependent Variable: mortrate

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.150	.108		-1.384	.172
	severity	.022	.006	.409	3.418	.001

a Dependent Variable: mortrate

Keep a mental note of the **residual standard error**, 0.09311, in boldface above. We are going to use it for comparison purposes. Recall also that the regression on **time** showed a **residual standard error** equal to 0.09652, so the two simple linear regression models are about equally successful in reducing the uncertainty surrounding **mortrate** (with the regression on **severity** just a hair better).

Multiple Regression of Mortrate on both Time and Severity

The multiple regression is carried out via the same ***SPSS*** sequence, **Analyze/Regression/Linear…**, as before. The only change is that, where asked to specify the independent variables in the dialog box, we type in the names of both variables, **time** and **severity**. Here are the results:

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.511(a)	.261	.235	**.08849**

a Predictors: (Constant), severity, time
b Dependent Variable: mortrate

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.158	2	.079	10.073	.000(a)
	Residual	.446	57	.008		
	Total	.604	59			

a Predictors: (Constant), severity, time
b Dependent Variable: mortrate

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	**-.082**	.106		-.779	.439
	time	**-.002**	.001	-.306	**-2.686**	**.009**
	severity	**.021**	.006	.395	**3.465**	**.001**

a Dependent Variable: mortrate

The fitted regression equation is:

predicted mortrate = - 0.082 - 0.002 time + 0.021 severity

The regression or slope coefficient of **time** is -0.002, which is the same when rounded as the result of -0.002 in the simple regression of **mortrate** on **time**. The interpretation of -0.002 is this:

- for a unit increase in **time**, fitted **mortrate** decreases by 0.002 unit, **for any level of severity.**

This, in turn, suggests that the trend found in the original simple regression was not just a reflection of changing severity of illness.

Similarly for **severity**, where the regression coefficient is 0.021:

- for a unit increase on the **severity** scale (Apache II scale), fitted **mortrate** increases 0.021 units, **for any value of time.**

This, in turn, suggests that there is a **severity** effect as well as a trend effect.

Assessment of statistical significance of these two coefficients is done as before: for both coefficients we see by the highlighting in the regression output above that the absolute t-ratios were substantially greater than 2, and the p-values (or significance levels) were much less than 0.05:

Note also that the **residual standard error**, equal to **0.08849**, is less than it was for either simple regression (that is, for **mortrate** regressed on **severity** or for **mortrate** on **time**). Thus the multiple regression model promises better predictability.

Analysis of Variance

So far we have not commented on the ***SPSS*** regression output in the second table of the display, i.e., the ANOVA display. The term "ANOVA" stands for "Analysis of Variance". It is called the "Analysis of Variance" because it analyzes, i.e., "breaks down" a measure of variability of the dependent variable into two parts-- (1) the part "explained" by the fitted model, and (2) the residual part. We shall not try to give a detailed explanation here because that would take us beyond the theoretical scope of ***STM***. However, there is one key number, the significance level, shown in bold below:

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.158	2	.079	10.073	**.000(a)**
	Residual	.446	57	.008		
	Total	.604	59			

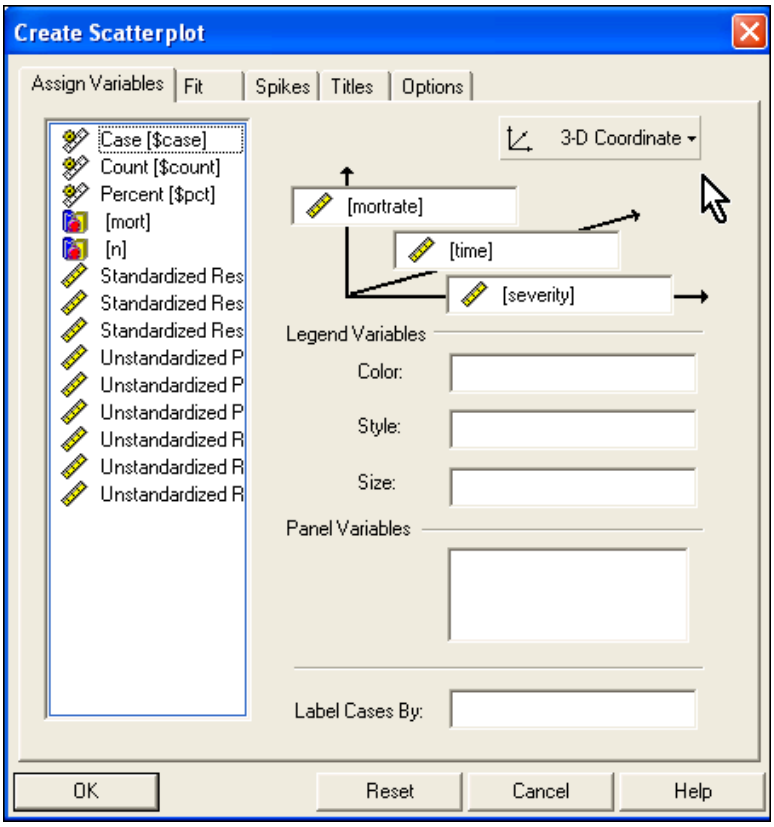
a Predictors: (Constant), severity, time
b Dependent Variable: mortrate

This p-value is less than 0.0005, which can be interpreted as a p-value or level of significance for the model as a whole, including both independent variables, **time** and **severity**. That is, if **neither** independent variable really contributed to the fit except by chance, there would be fewer than 5 chances in 10,000 of getting coefficients at least this far from 0.

By contrast, the p-values were 0.011 for **time** and 0.001 for **severity** in the regressions on a single independent variable, so the significance is clearer for the model as a whole than for either of the component variables. This will not always be so, and you would do well always to check the overall significance level for the analysis of variance as well as the individual p-values (and/or t-ratios) for the coefficients of the individual variables of the multiple regression.

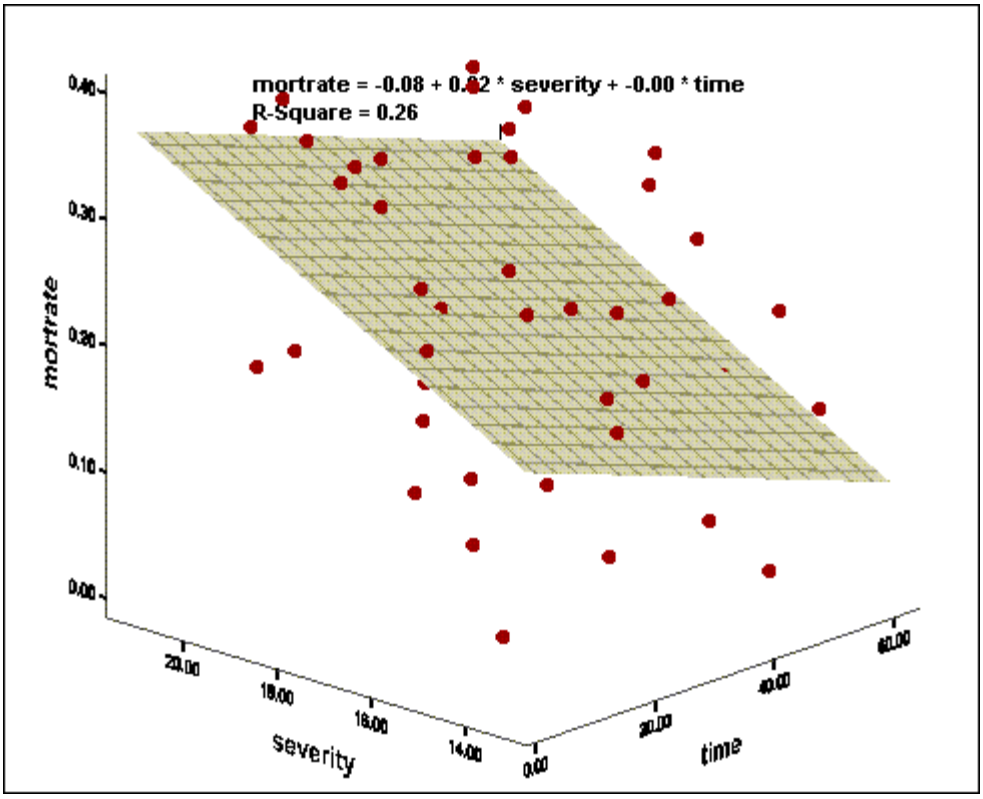
Graphical Display of Multiple Regression

After running the multiple regression above, we click on **Graphs/Interactive/ Scatterplot…** again. This time, however, we choose the 3-D Coordinate form and set up the dialog window as follows:



As you might expect, the fitted linear model for the regression of a dependent variable on **two** independent variables is a plane. The point of origin for all three of the axes is the corner that appears to be closest to you when looking straight at the plot. If you follow the orientation of the edges of the plane shown below, you can see that as **time** increases, for a fixed value of **severity**, the fitted value of **mortrate** decreases-- corresponding to the negative coefficient of **time**. Similarly, you can see by following the edge of the plane along the **severity** axis that **mortrate** increases as **severity** increases, holding the value of **time** fixed.[6]

[6]Actually, the fitted regression plane should extend infinitely in all directions. The "edges" here are really the lines of intersection of the plane with the front, back, and side walls of the cube defined by the three coordinate axes of the plot. When the linear regression involves more than one independent variable, the regression coefficients for the variables are called "partial slopes". (Technically, they are the first-order partial derivatives of the linear function.)



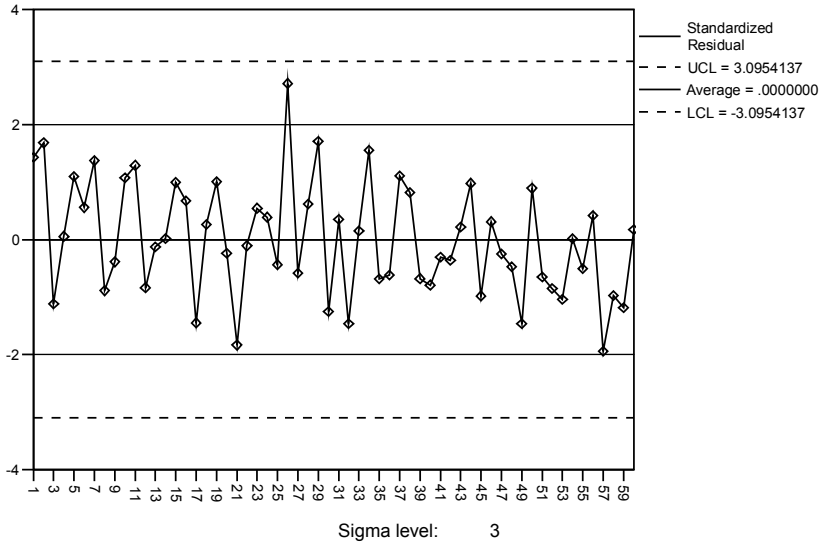
There is another nice feature of ***SPSS*** with respect to 3-D plots that may provide some amusement as well as education when you examine the image above, or a similar one, on your own PC. If you double-click and enter the chart editor you will see the window below:

This display shows a feature that enables you to rotate the 3-D plot shown above about any of the three axes. Try it out just for fun. Just place your mouse pointer on one of the wheels and then, pressing the left button, move it. In all seriousness, this feature sometimes enables an analyst to detect indications of special causes that might not be evident in the first image is produced. Sometimes the detection just requires looking at the picture from a slightly different perspective. We mention this point here for future reference.

Diagnostic Checking of Multiple Regression

Now, just as in simple regression, we check the multiple-regression assumptions by "diagnostic checks" of residual behavior. These parallel the checks used in simple regression. All are satisfactory, as you should confirm.

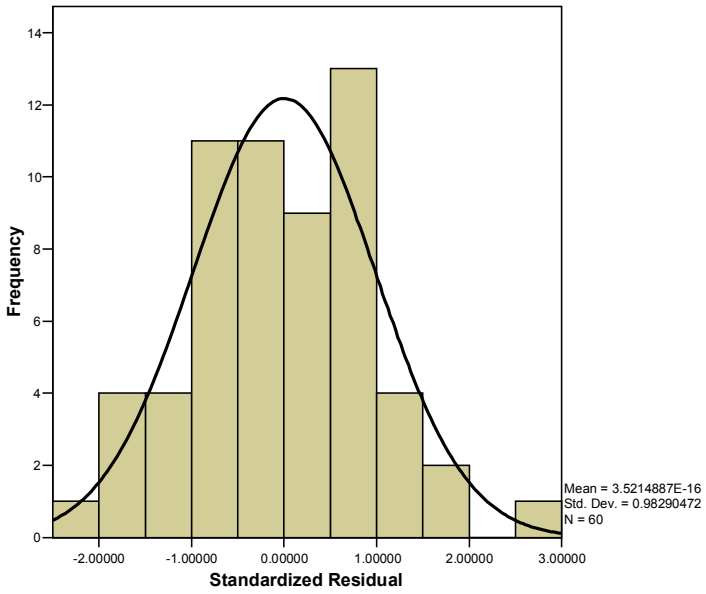
Control Chart: Standardized Residual



Runs Test

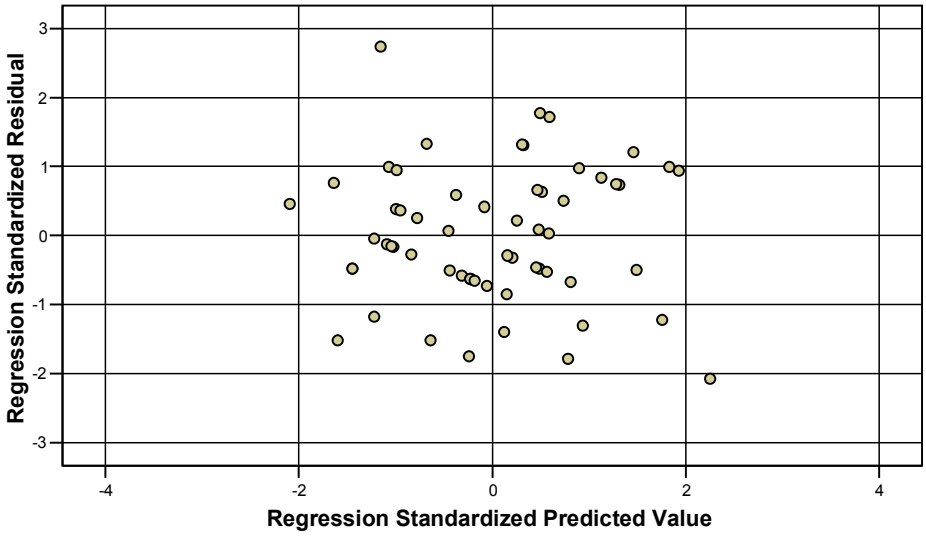
	Standardized Residual
Test Value[a]	.0000000
Cases < Test Value	31
Cases >= Test Value	29
Total Cases	60
Number of Runs	33
Z	.530
Asymp. Sig. (2-tailed)	.596

a. Mean



Scatterplot

Dependent Variable: mortrate

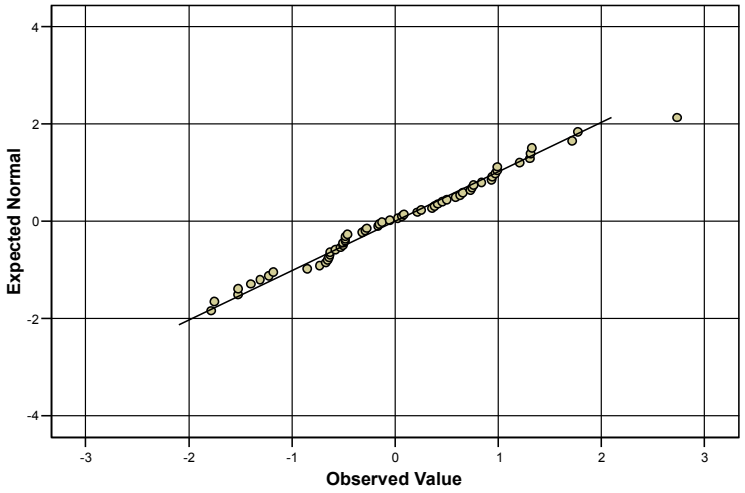


Tests of Normality

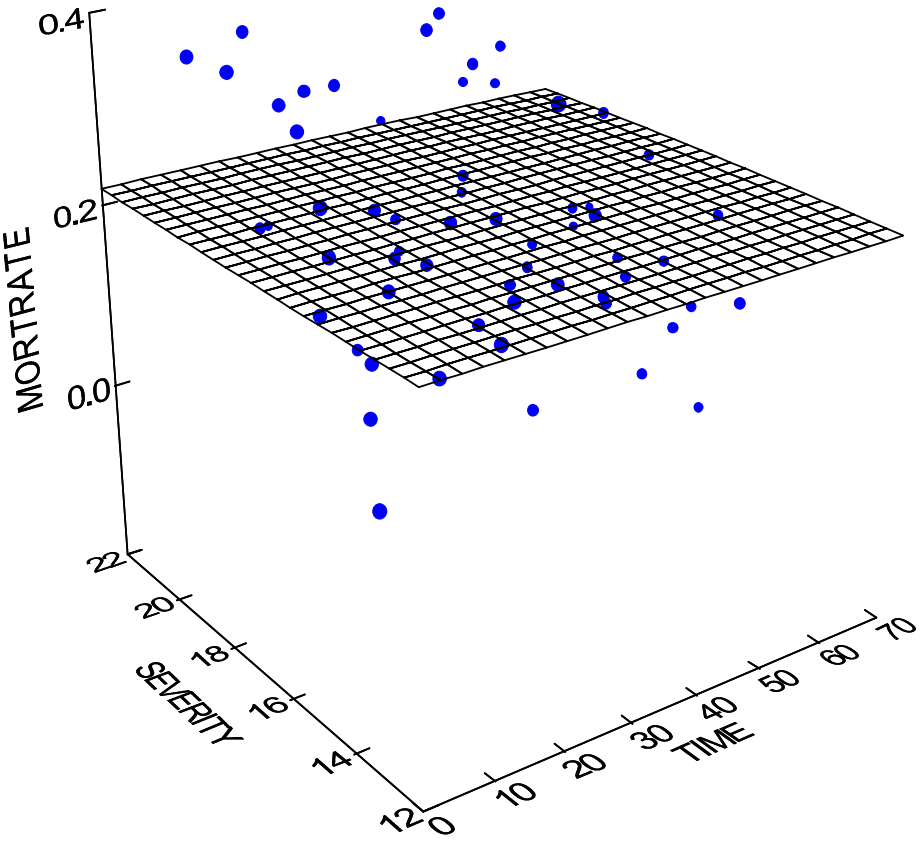
	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.081	60	.200(*)	.987	60	**.763**

* This is a lower bound of the true significance.
a Lilliefors Significance Correction

Normal Q-Q Plot of Standardized Residual



Before continuing our main discussion let's take a few moments to visualize in three dimensions the three levels of linear regression that we have been considering. The first of the three, although it may not seem at first to be a case of regression, is where we used no independent variables at all. We have said that if **mortrate** were in control (which it is not), and without considering its possible relationship with other variables, our best estimate for the next, as yet unobserved, value is just the sample mean. Think about this for a moment. That is the same thing as fitting a linear regression model with just the constant, no other parameters.
Here is the representation of this very, very simple linear model in 3-D[7]:



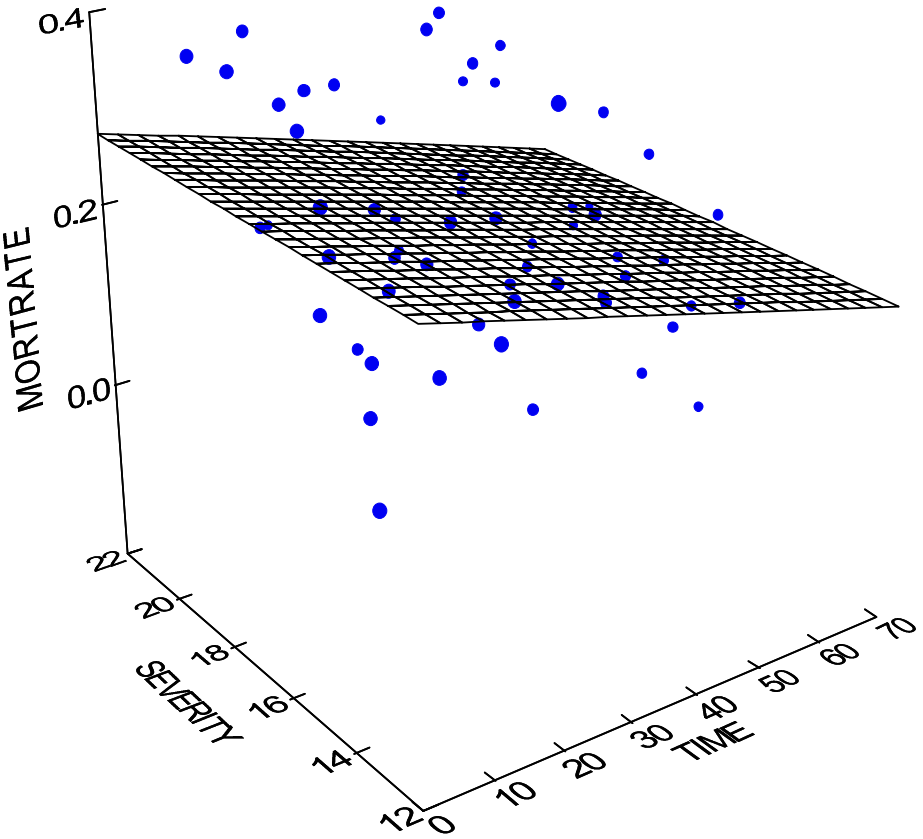
predicted mortrate = 0.2175

Note that the fitted plane in this example is completely horizontal, i.e., the partial slopes for **time** and **severity** are set at zero, and the plane intercepts the vertical axis at the mean value. It is a fact

[7] Most of the plots shown here and on the following pages were made using another statistical software package, ***SYSTAT for Windows*®, Version 6**, a product of SYSTAT Products, SPSS, Inc., of Evanston, IL. We do not expect you to reproduce them.

that if you could measure the vertical distances of the various points from the plane, square the distances, and sum up the squares, you would just get the same sum of squared deviations that is used in the calculation of the standard deviation of **mortrate**.

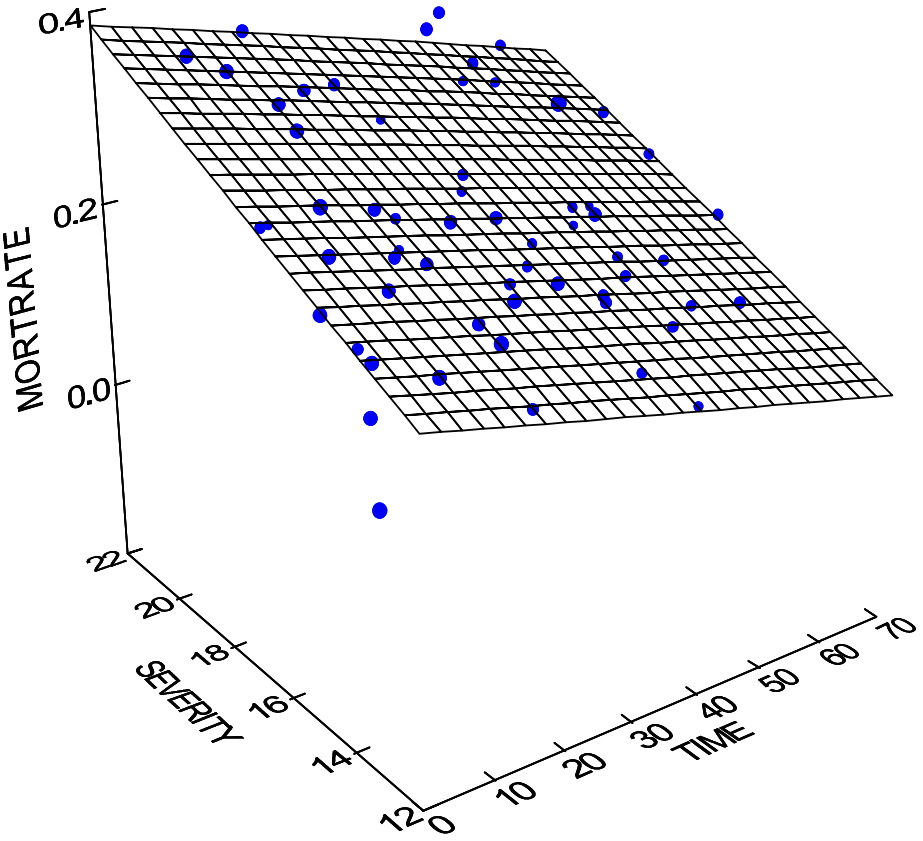
Next we show the 3-D plot for the simple (one independent variable) of **mortrate** on **time**:



predicted mortrate = 0.275 - 0.002 time

Now the plane is still horizontal with respect to the omitted variable **severity.** Along the **time** axis**,** however, it is lifted up a bit, intercepting the vertical axis at 0.275, and it is tilted downward, displaying the negative partial slope for **time**, -0.002. With the ability to move the plane from a totally horizontal position by regressing **mortrate** on **time**, we have made the residual standard error smaller than the original value, i.e., the plane is closer to the individual data points.

Finally we display the 3-D picture of the full model with both **time** and **severity** on the right-hand side of the regression equation:



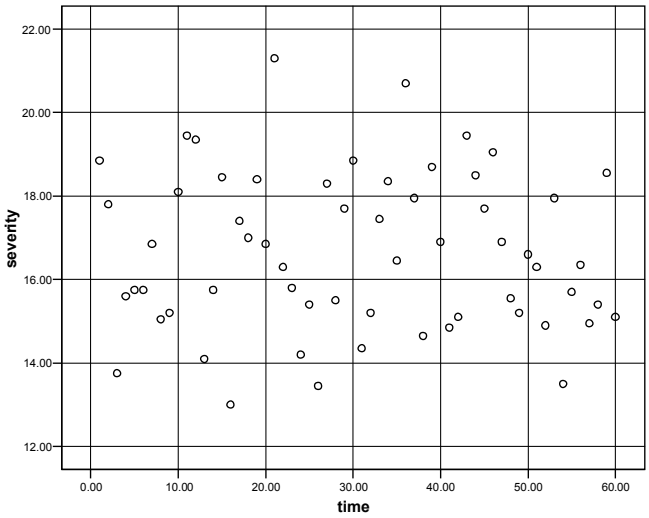
predicted mortrate = -0.082 -0.002 time + 0.021 severity

Now the plane is tilted upward along the **severity** axis and the intercept is decreased in value. When rounded to three decimal places, the partial slope for **time** is the same as before. You see that with freedom to adjust the orientation of the plane more than before by regressing on both variables instead of just one we have succeeded in placing the linear surface closer than ever to the swarm of points. Although we could not show a picture, you can imagine that if we could identify additional independent variables that are related in some way to **mortrate,** we could fit a

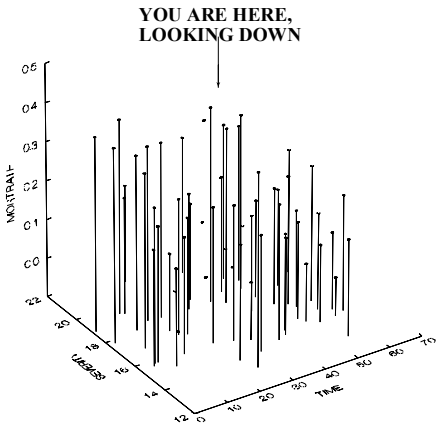
plane (called a **hyperplane** when greater than three dimensions) that gets closer and closer to the points with each increase in dimension, approaching a perfect fit[8].

Relationship between Independent Variables

One other aspect of the data is of interest, the relationship between the two independent variables **time** and **severity**. We see from the scatter plot below that the relationship appears to be weak or nonexistent.

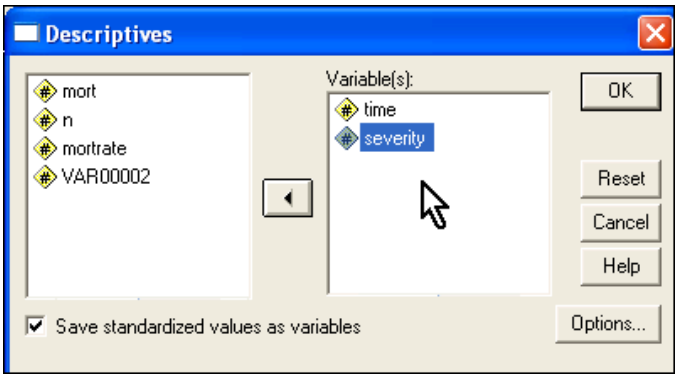


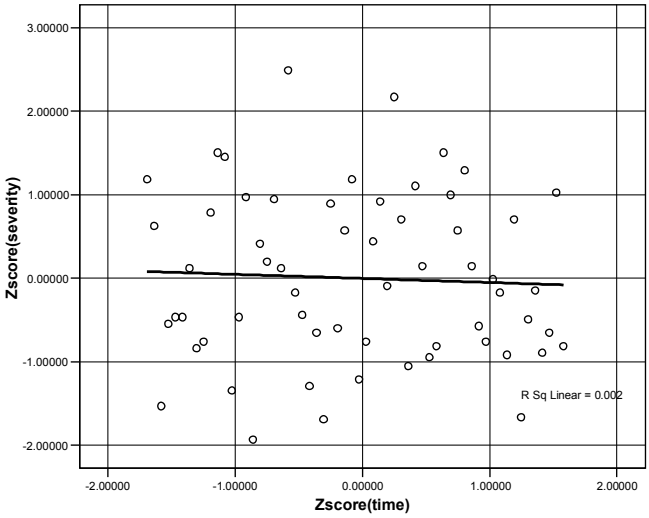
We repeat the 3-D scatter plot, but this time with "spikes" connecting the data points with the floor, that is, the plane with **time** and **severity** for axes. The scatter plot above is what you would see if you were in a helicopter hovering directly above and looking straight down at the floor.



[8] Later we shall discuss the danger and the folly of "overfitting" a regression model.

At the next step we invoke the ***SPSS*** sequence **Analyze/Descriptive Statistics/ Descriptive…** for the sole purpose of producing standardized values of **time** and **severity**. (See the checked box.)





The plot of standardized **severity** against standardized **time** (above and to the right) is very similar to the previous one, but now the overall horizontal and vertical dispersions are the same. (The standard deviation of any standardized variable is 1.) Note that this time we called for the regression line to be drawn. It is almost horizontal (with only a slight negative slope), indicating virtually no linear association between the variables. Here is the output for **zseverity** regressed on **ztime**:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.047[a]	.002	-.015	1.00744629

a. Predictors: (Constant), Zscore(time)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-.003	.130		-.021	.983
	Zscore(time)	-.049	.135	-.047	-.362	.719

a. Dependent Variable: Zscore(severity)

The results confirm that the regression coefficient is close to zero-- the low t-ratio and high p-value show that we **cannot reject** the hypothesis that there is no linear relationship between the two variables.

Next, we run the other regression-- **ztime** regressed on **zseverity**:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.047[a]	.002	-.015	1.00744629

a. Predictors: (Constant), Zscore(severity)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.000	.130		.000	1.000
	Zscore(severity)	-.047	.131	-.047	-.362	.719

a. Dependent Variable: Zscore(time)

Perhaps surprisingly to you, the displays for **zseverity** regressed on **ztime** and **ztime** regressed on **zseverity** are exactly the same except for the labeling of the dependent and the independent variables. Note especially that the regression coefficient is **-0.047** in both cases. In general when we compare the regression of Y on X with the regression of X on Y, the two regression coefficients will not, as they are here, be equal. We are working here, however, with **standardized variables ZX and ZY**, and with standardized variables it does not matter which is the dependent or the independent variable--**the coefficients for the two alternative regression lines must be equal.**

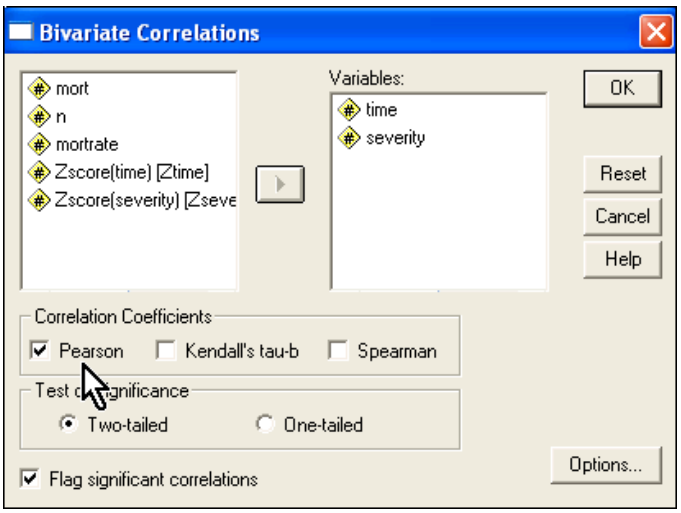
The regression coefficient of ZY on ZX, or of ZX on ZY, can be interpreted as the **correlation coefficient** between X and Y.[9]

In discussing the scatter plot for **zseverity** vs. **ztime** above we observed that the regression line was nearly horizontal--- hence its slope is near zero, meaning nearly zero correlation. If, on the other hand all the points lie exactly on an upward sloping line it will have a slope of 1 (with standardized ZX and ZY) and the correlation coefficient is +1. If all the points lie on a downward sloping line with slope -1, then the correlation is -1, indicating perfectly negative linear association. The correlation coefficient can never be less than -1 or greater than +1. In the real world most data sets yield correlations that lie between these extremes.

The correlation coefficient, **R**, is routinely displayed in the output from linear regression along with its squared value, **R Square.** (See the tables above entitled **Model Summary**.) The square of -0.047 is 0.002 (or 0.2%) when rounded to three decimal places. We shall later give another interpretation of **R Square** in terms of the "percentage of variance explained by the regression".

[9] This is but one of several definitions of the correlation coefficient--- that is, the slope of the regression line when one standardized variable is regressed on another standardized variable. We may mention some of the other interpretations later.

If you would like ***SPSS*** to compute the correlation coefficient for you without running a regression involving standardized variables use the sequence **Analyze/Correlate/Bivariate…:**



The correlation is called "bivariate" because it is a measure of association involving two variables. Note that in the window at left we have checked the box for **Pearson**. Karl Pearson (b.1857-d.1936) was one of the founders of modern statistical science. There are many measures of association that use the name "correlation". Pearson's is just one of them, but arguably the most widely used.

Correlations

		time	severity
time	Pearson Correlation	1	-.047
	Sig. (2-tailed)	.	.719
	N	60	60
severity	Pearson Correlation	-.047	1
	Sig. (2-tailed)	.719	.
	N	60	60

Here is the table that results when we click **OK** in the window above:

You see the number -.047, the same as the slope of the two regression lines involving the standardized variables.

Just to give you something to think about we have specified five variables to have correlations calculated. That implies that there will be 10 distinct correlations relating each pair of variables. See if you can figure out how to read the output in the "correlation matrix" displayed below:

Correlations

		mortrate	time	severity	Unstandardized Predicted Value	Unstandardized Residual
mortrate	Pearson Correlation	1	-.325*	.409**	.511**	.860**
	Sig. (2-tailed)	.	.011	.001	.000	.000
	N	60	60	60	60	60
time	Pearson Correlation	-.325*	1	-.047	-.636**	.000
	Sig. (2-tailed)	.011	.	.719	.000	1.000
	N	60	60	60	60	60
severity	Pearson Correlation	.409**	-.047	1	.801**	.000
	Sig. (2-tailed)	.001	.719	.	.000	1.000
	N	60	60	60	60	60
Unstandardized Predicted Value	Pearson Correlation	.511**	-.636**	.801**	1	.000
	Sig. (2-tailed)	.000	.000	.000	.	1.000
	N	60	60	60	60	60
Unstandardized Residual	Pearson Correlation	.860**	.000	.000	.000	1
	Sig. (2-tailed)	.000	1.000	1.000	1.000	.
	N	60	60	60	60	60

*. Correlation is significant at the 0.05 level (2-tailed).

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

Note first the correlation between **time** and **severity** that we have been discussing. It is lightly shaded.

Next, observe that the correlation between **mortrate** and **predicted** is 0.511. The **predicted** variable here is from the multiple regression of **mortrate** on both **time** and **severity**. If you thumb back a few pages to the display of that regression you see that **R Square** was 0.261. You can verify that it is the square of 0.511. This value is the highest correlation that you can achieve with **mortrate** by choosing coefficients for a linear function of **time** and **severity**.

Thus **predicted (PRE_1)** can be called that linear function of **time** and **severity** that has the maximum correlation with the dependent variable, **mortrate**. The method of least squares leads to that achievement.

Notice also that the correlations between **residual (RES_1)** and each of the variables **PRE_1**, **time**, and **severity** are all zero. This is another feature of the least squares method:

> The residuals **must be** uncorrelated with the predicted (fitted) values in linear regression. Similarly, they are uncorrelated with any of the independent variables that were used in the model.

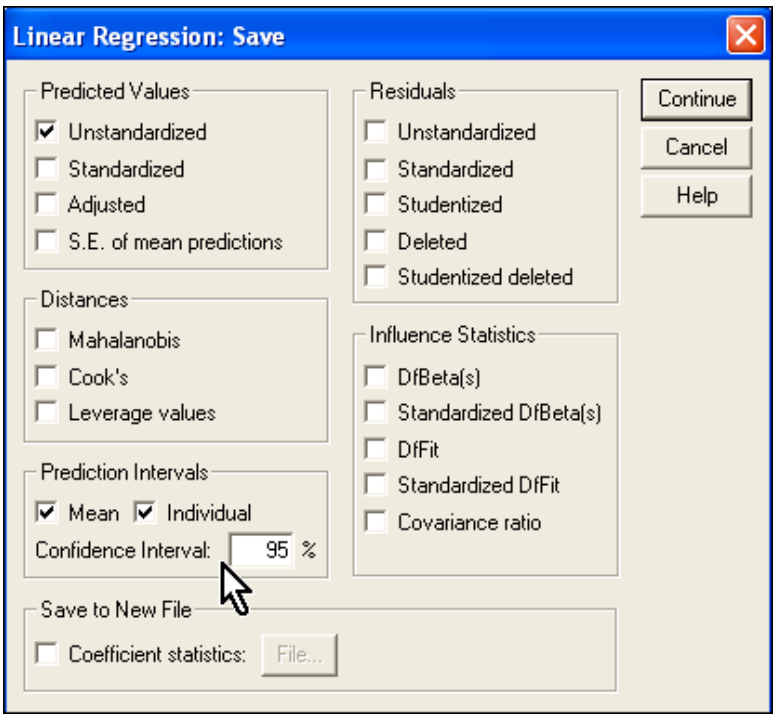
Effects of Correlation between Independent Variables

In this first application of multiple regression, then, the correlation between independent variables **severity** and **time** turns out to be very low. An important consequence of this fact is that the regression coefficients in the multiple regression are close to those in either of the simple regressions, **mortrate** on **severity** or **mortrate** on **time**.

If, on the other hand, we regress Y on X1 and X2, and X1 and X2 are highly correlated, the coefficients in the multiple regression can differ greatly, even in sign, from the coefficients in the simple regressions of Y on X1 and Y on X2. If the correlation between X1 and X2 is extremely high the coefficients for each of the variables may appear to be close to zero, whereas if one of the independent variables were omitted, that remaining would have a significant association with the dependent variable. We will say more later about the complications entailed by this phenomenon, which is sometimes given the ominous name "**multicollinearity**".

APPENDIX: MORE ABOUT PREDICTED VALUES

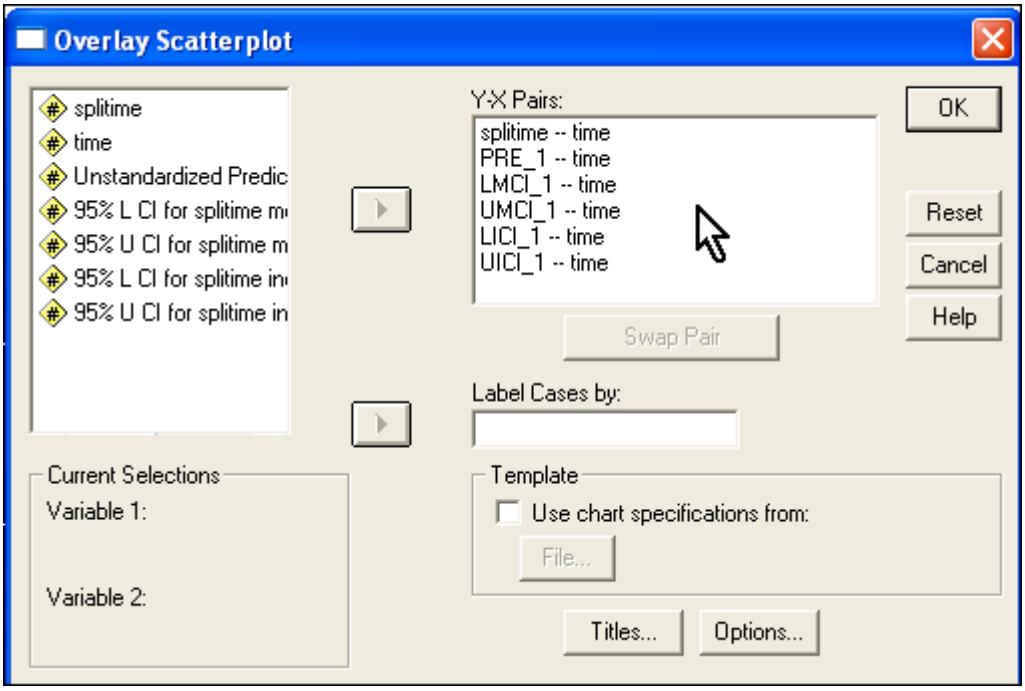
Let's consider again the data file LAPSPLIT.sav. We shall rerun the regression of **splitime** on **time**, but this time, in addition to saving the unstandardized predicted values, **PRE_1**, we also save the **prediction intervals** (both **mean** and **individual**, as shown by the arrow below).



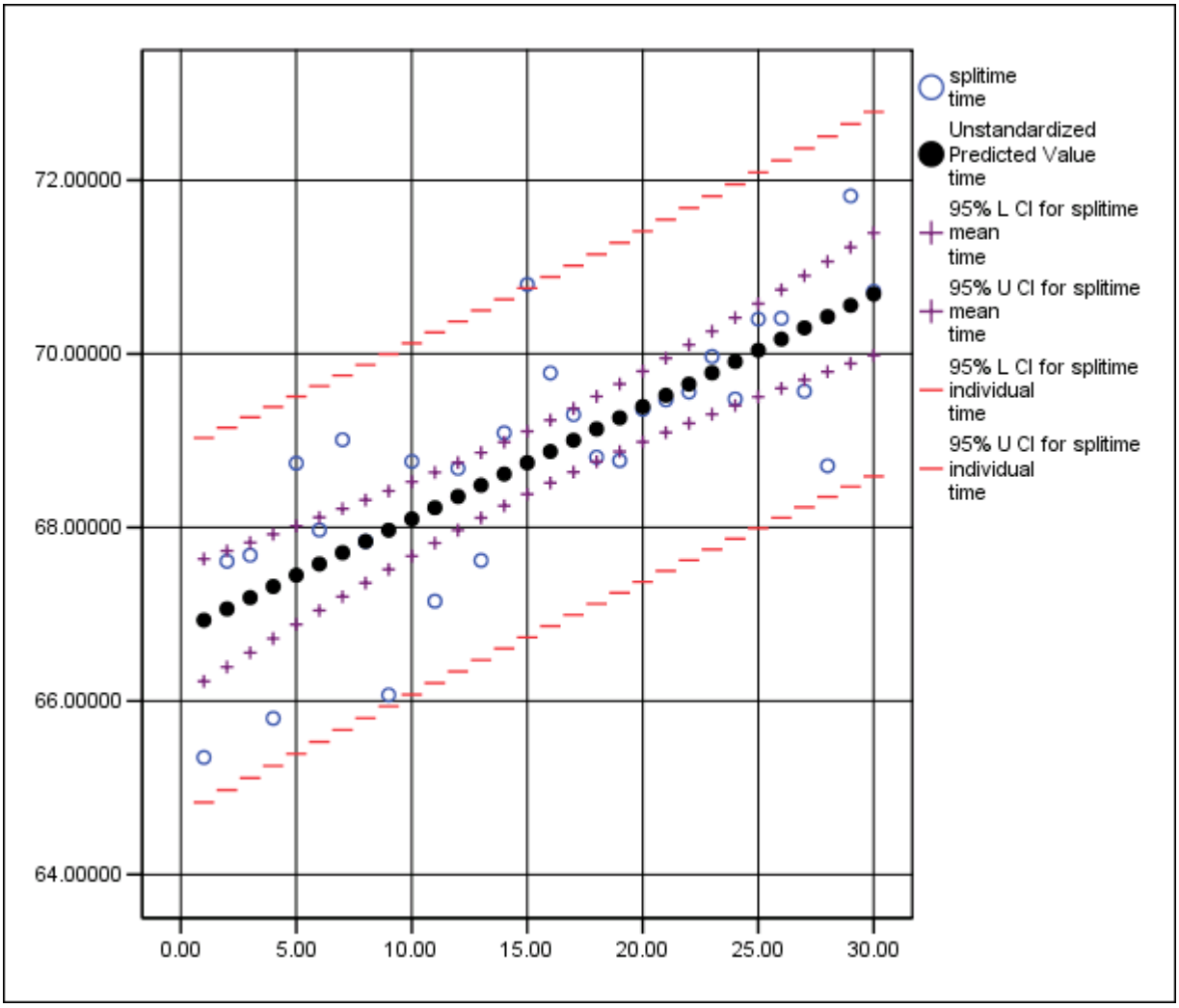
The first few rows of the **Data Editor** now look like this:

	splitime	time	PRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1	va
1	65.35	1.00	66.93161	66.22685	67.63637	64.83132	69.03191	
2	67.61	2.00	67.06116	66.39189	67.73042	64.97250	69.14981	
3	67.68	3.00	67.19070	66.55617	67.82523	65.11292	69.26848	
4	65.80	4.00	67.32024	66.71955	67.92094	65.25254	69.38794	
5	68.74	5.00	67.44979	66.88189	68.01769	65.39138	69.50820	
6	67.97	6.00	67.57933	67.04298	68.11569	65.52940	69.62927	
7	69.01	7.00	67.70888	67.20259	68.21516	65.66660	69.75115	
8	67.84	8.00	67.83842	67.36045	68.31639	65.80298	69.87386	

Next, we set up a scatter plot with overlay as follows:



The result (after editing) is the graph below:



In addition to showing the regression line (the line of predicted values), there are two sets of lines enclosing the fitted line that must be explained. Consider first the pair of new lines indicated by plus signs. These lines are fairly close to the regression line and are bowed in shape—that is, the interval between them becomes wider as the value of time moves farther from its mean, 15.5. The lines are the upper and lower confidence limits for the predicted value itself. They reflect the fact that the regression line is only an estimate based on the present sample, and that if another sample had been drawn, **under the same process conditions**, we would have seen a different result, although one would hope that under conditions of control there would not be much variation from the line shown here. To put it in other terms, if you believe that there is a "true" linear relationship between **splitime** and **time**, before we even selected the sample there was a 95 percent chance that the resulting limits shown by pluses would enclose the correct line. That is why they are called 95% confidence limits.

Although the confidence bounds for the regression line shown by pluses deserve mention, in this course we are more interested in the outer 95% upper and lower confidence limits shown by dashes. Focus, for example, on the vertical grid line at **time = 20**. If you go back to the **Data Editor** you will see the following for that case:

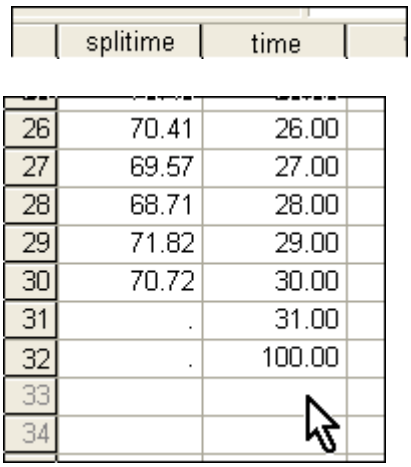
	splitime	time	PRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
20	69.36	20.00	69.39295	68.98582	69.80008	67.37297	71.41293

Now consider this problem: Suppose that we want to repeat the process under exactly the same conditions as those that were in effect when we made our observations of **splitime**, and we ask this question: "What will our new observation of **splitime** be when **time** equals 20?" The answer is that we expect it to be the same as our predicted value (the height of the regression line) which is **69.39**. (See the image above for **time=20**.) We also must ask, however, about our uncertainty concerning that prediction for an individual new observation. The lower and upper confidence limits, **LICI_1** and **UICI_1** tell us that with 95% confidence, although we **expect** the new observation to fall right on the line, it may be as low as **67.37** or as high as **71.41**. Another way to look at the problem is to remember that in asking for a prediction of a new observation of **splitime** we are not only uncertain about the regression line, but we are also uncertain about the **residual** from the line, i.e., the amount by which the actual **splitime** will fall above or below the line. Our uncertainty about the residual is expressed by the **residual standard error** which, you may recall, was reported as **0.96588**. (See the original regression analysis early in this chapter.) You can verify that the intervals between **LICI_1** and **UICI_1** for all rows in the **Data Editor** extend approximately from minus two to plus two times the **residual standard error**.[10]

As a final point of discussion in this appendix, recall that around page 4-22 in this chapter we mentioned the advantage of having the regression model to forecast the value of **splitime** when **time = 31**, one period beyond the data that were actually observed. We explained that instead of simply using the mean value of past data as an estimate, it is better to take the time trend into account by using the formula for the predicted value, a procedure that gave us a forecast value equal to **70.8** seconds. We shall now show how to use ***SPSS*** to accomplish this while also producing confidence limits for the forecast.

[10] The multiplier is actually closer to 2.08, but we shall not quibble in this course about minor technical details. The important point is that the residuals are assumed to be approximately normally distributed.

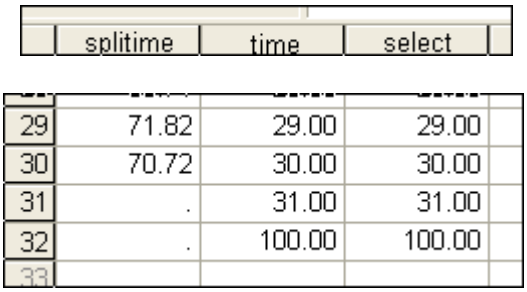
The first step is to clear out all of the columns involving predicted values and confidence limits from the **Data Editor** and then to type in two new values of **time** in rows 31 and 32:



	splitime	time
26	70.41	26.00
27	69.57	27.00
28	68.71	28.00
29	71.82	29.00
30	70.72	30.00
31	.	31.00
32	.	100.00
33		
34		

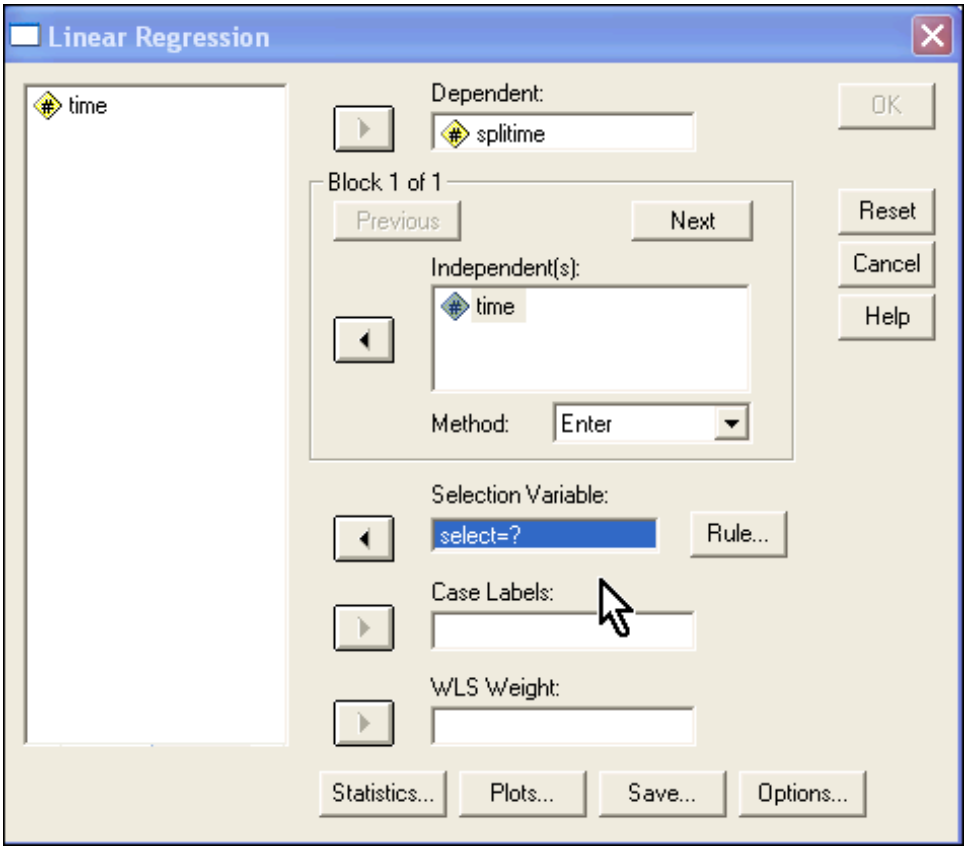
For now, ignore the last value, 100—we shall discuss that later. Concentrate on the new value 31.

At the next step we must create a new variable—we will call it **select**. Through **Transform/ Compute…** we make **select** exactly equal to **time**. The last rows of the Data Editor now look like this:

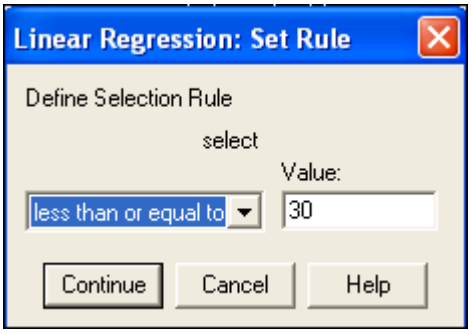


	splitime	time	select
29	71.82	29.00	29.00
30	70.72	30.00	30.00
31	.	31.00	31.00
32	.	100.00	100.00
33			

Then we set up the regression analysis again, but this time we have moved the new variable **select** into the box for **Selection Variable:**



Next, we must click on the **Rule…** button and set up the new dialog window as follows:



Finally we continue back to the regression dialog window and do the analysis, saving the unstandardized predictive values and the confidence limits for individuals. Our results look like this:

	splitime	time	select	PRE_1	LICI_1	UICI_1
29	71.82	29.00	29.00	70.55884	68.47019	72.64750
30	70.72	30.00	30.00	70.68839	68.58809	72.78868
31	.	31.00	31.00	70.81793	68.70523	72.93063
32	.	100.00	100.00	79.75646	75.69672	83.81620
33						

We see that by using the **Selection Variable** feature of linear regression, we can obtain predicted values and confidence limits for values of the independent variable that are not in the data set that was used to fit the model. For **time = 31** we get the same value that we calculated by hand back on page 4-22, **70.8 seconds**, but now we can also report that our 95% confidence interval is **68.7 seconds to 72.9 seconds**, unfortunately fairly wide-ranging.

Finally we explain why we have also included **100** as a value for **time** for which we want to forecast **splitime**. Note in the image above that we predict that for **time = 100** our running time will have degenerated to **79.8** seconds with an interval of uncertainty ranging from **75.7 to 83.8** seconds. You should ask, however, "Why are we trying to predict an observation that is so far removed from our actual data? Is it even sensible to consider that the runner would have the strength and endurance to continue for 100 lap splits? Do we really believe that the straight line would continue with the same slope as during the first 30 splits?" Recognizing that the answer to all of the questions is probably a resounding "No!", we also note that we are at least penalized for such an outlandish prediction by a confidence interval that is very wide. When **time = 100**, the resulting interval is **8.1** seconds wide as opposed to only **4.2** seconds wide for **time = 31**. In our graph above, if the time axis had been extended far enough, the outer limits (the lines of dashes) would be bowed outward, similar to the inner confidence bands (shown by pluses). Thus there is an important lesson contained in this last example:

We caution that in real world applications of linear regression it is extremely dangerous to forecast results using values of the independent variables that are far outside the range observed in the original data set. How can we be sure that the conditions underlying the process that we have observed will not have shifted in some unpredictable way?