CHAPTER 7: CROSS-SECTIONAL DATA ANALYSIS AND REGRESSION

1. Introduction

In all our statistical work to date, we have been dealing with analyses of time-ordered data, or time series: the same variable or variables observed and measured at consecutive points of time. Usually but not necessarily, the points of time are equally spaced. Time-ordered data are very often pertinent for total quality; for example, we need to know whether our processes are in statistical control or whether they are being affected by, say, trends or special causes. We need also to evaluate the effectiveness of interventions aimed at improving our processes and to assure that we are holding the gains from effective interventions from the past.

But not all data are time-ordered. There is also a type of data called **cross-sectional** data, where we are dealing with information about different individuals (or aggregates such as work teams, sales territories, stores, etc.) **at the same point of time or during the same time period**. For example, we might have data on total accidents per worker over the course of the last calendar year for all the workers in a given plant, or we might have questionnaire data on customer satisfaction for a sample of customers last month.

There is also the possibility, to be discussed in Section 6 of this chapter, of a time series of cross sections (or, alternatively, a cross section of time series). For example, we might have monthly sales by each of 37 sales territories for the last 60 months.

We have explained and applied regression tools in the context of time-ordered data. The same tools are directly applicable to cross-sectional data. In one respect the cross-sectional regressions will be simpler: **we do not need to check as to whether the data are in statistical control through time.** We will not need control charts, time-series sequence plots, or runs counts. You can simply skip that part of the analysis, even though by now it has become habitual.[1]

To see what can be learned from cross-sectional data, we now consider the illustration of accidents per worker. Here are some of the things we might be interested in:

- Is there evidence that some workers are more prone to accidents than are others?
- If there are accident-prone workers, who are they and what preventive training may be helpful for them?
- If there are accident-prone workers, are there systematic factors that are associated with higher or lower accident rates?
- If there are systematic factors, can we give them unambiguous causal interpretations?
- Can we do intervention analysis or designed experiments to develop and test accident-prevention policies?

[1]However, the type of question addressed by the checks for statistical control through time has a counterpart for cross-sectional data. In Section 5 we shall discuss briefly how to deal with it.

2. Special Causes and Pareto Analysis

When we have cross-sectional data bearing on a single variable, the time-series analyses are no longer necessary. Rather, our attention focuses on the histogram. The histogram, by its general shape and/or its apparently outlying observations, offers hints as systematic and special causes that may be affecting the data. The analysis of histograms, however, doesn't lend itself quite so easily to a systematic approach to data analysis. Even statisticians may draw more on their knack for detective work than their knowledge of statistical distributions.

The general aim can be illustrated by applications to counting data, in which the Poisson distribution is a first thought for statistical model. If the Poisson distribution is appropriate, the differences between individual measurements are attributable to "chance", and there is neither a "Pareto effect" or any way to single out special causes. This will become clearer if we examine an application to error counts by operators.

Operator Errors

The following study of operator errors gives cross-sectional data on errors in a given month by 10 operators who were keying data into a computer. Even though the data have no time ordering, it is useful, **purely for display,** to look at them with a **c-chart**. The reason is this: if all operators are equally disposed to make errors, the observed cross-sectional histogram of operator errors should be compatible with the Poisson model (see Chapter 4, Section 1). We can get a quick, if rough, check on this assumption by looking for points outside of the control limits on **c-chart**, which are computed on the assumption that the Poisson distribution is applicable.

Here are the notes for the data, contained in the text file OPERROR.sav:

Variables:

operator: ID of operator in data processing department
freq: frequency of data entry errors in December, 1987.

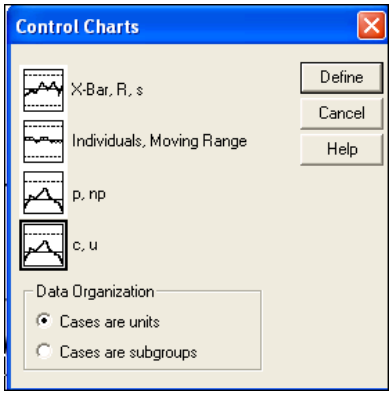
All 10 operators entered about the same amount of data.
Source: Gitlow, Gitlow, Oppenheim, and Oppenheim, "Telling the Quality Story", QUALITY PROGRESS, Sept., 1990, 41-46.

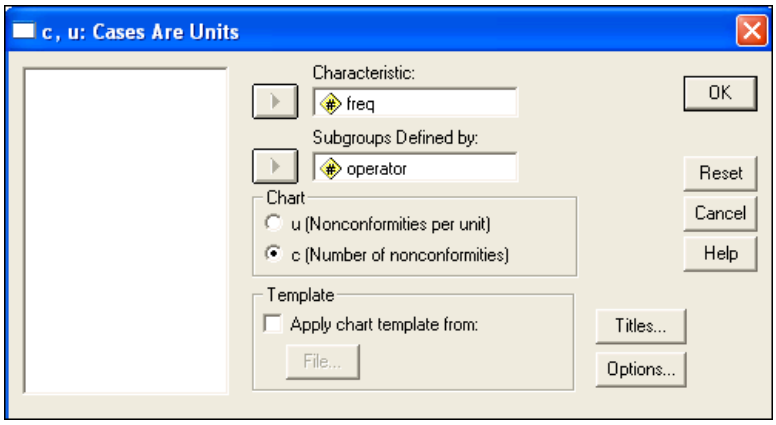
We name the variables **operator** and **freq** as we import the file with ***SPSS*.**

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
freq	10	0	19	5.00	6.928
Valid N (listwise)	10				

Next, we set up the **c-chart** as follows:

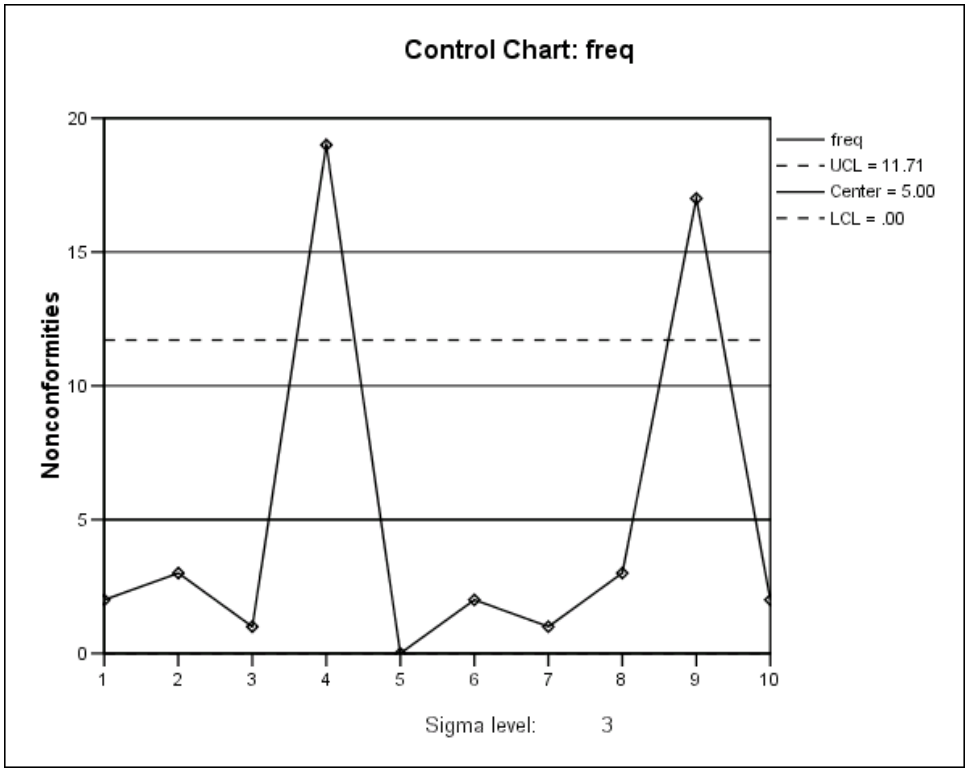




In the chart below, we see that operators 4 and 9 are far above the UCL, suggesting that they were significantly more error prone. In the actual study, this finding was followed up, and it was found that operator 4's problems were correctable by glasses which permitted her to see better the numbers she was working with. (We have no report on operator 9.)

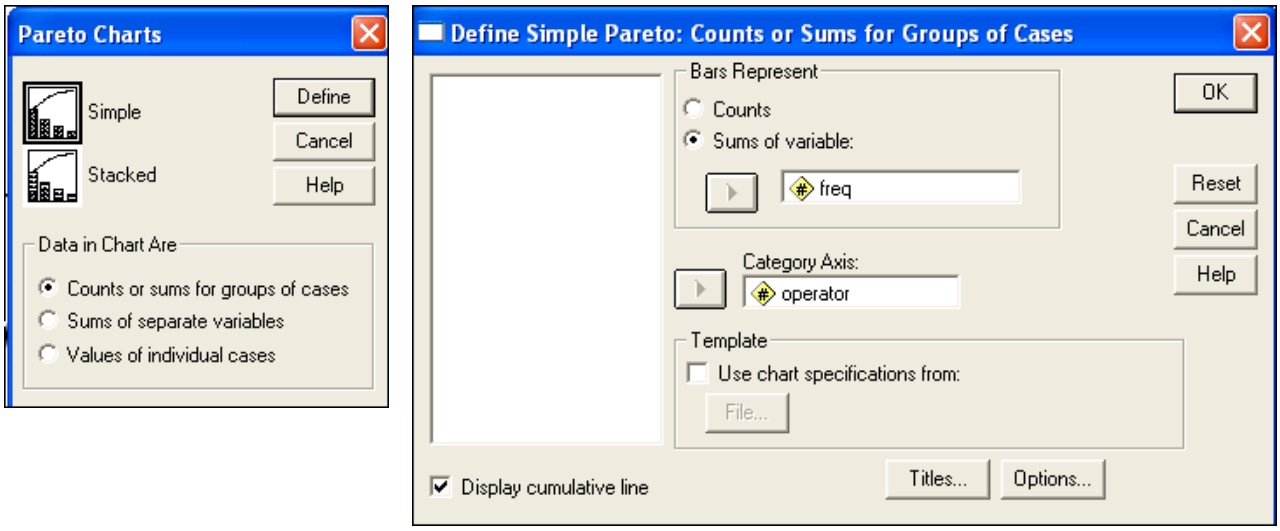
Checking the reasonableness of the Poisson assumption can help in at least two ways:

- Better understanding of the cause system underlying the data.
- Identification of special causes.

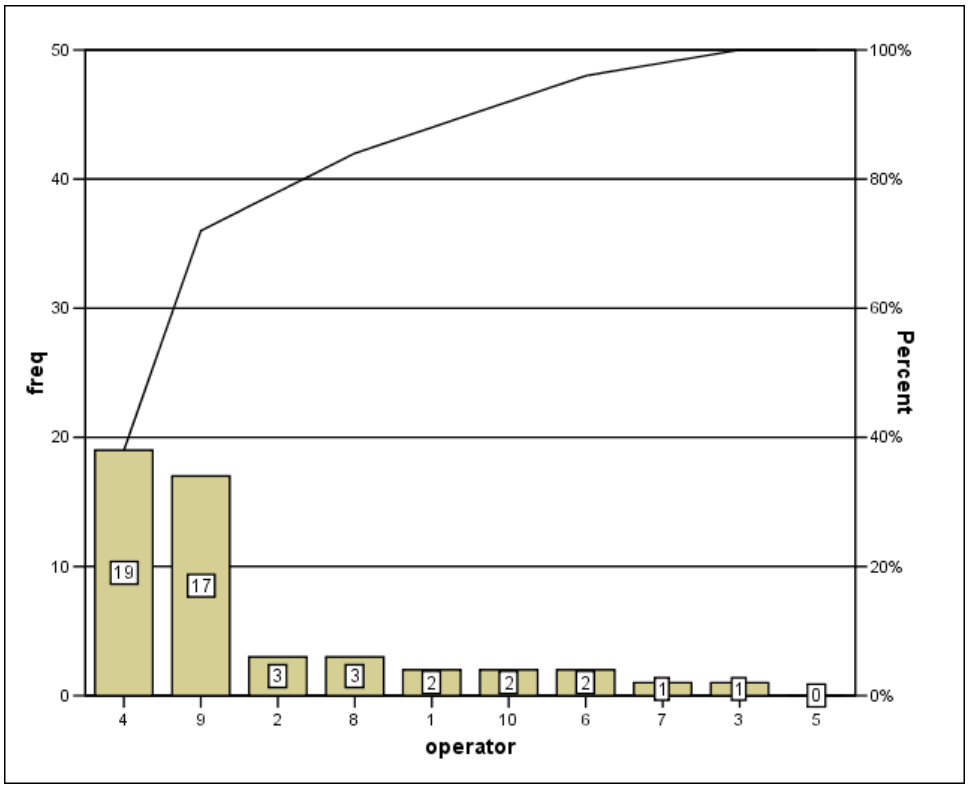


As already mentioned above, the exceptions are for Operators 4 and 9.

Since the Poisson distribution seems inappropriate, there is an apparent "Pareto effect": a few operators account for a major fraction of the accidents. We now use ***SPSS*** to do Pareto analysis. The appropriate procedure is **Graphs/Pareto…**, which brings up the following dialog box:



Notice that **operator** has been entered as the **Category Axis** variable. Be sure, also, to check the box for **Display cumulative line**.



We see that the Pareto Chart is really just a bar chart that has been arranged in a special way. It shows the "defects" from the various sources in descending order of magnitude from left to right. It also shows the cumulative percentage of the each contribution to the total number of defects. Thus we see that Operator 4 was the person who had the most accidents (19 on the left vertical axis) and that her contribution to the total was 38 percent (shown on the right vertical axis). Operator 9 was next with 17 accidents, so that Operators 4 and 9 by themselves accounted for 72 percent of the total. Pareto analysis is one of the most useful of all the elementary statistical tools of quality management. In Juran's

expression, it singles out the "vital few" problems from the "useful many", thus setting priorities for quality improvement.

For example, a manufacturer studied failures of parts and discovered that seven of a very large number of part types accounted for nearly 80 percent of warranty defects, and that three of a large number of branch locations accounted for a large percentage of warranty defects. Improvement efforts could then be concentrated on these parts and branches.

In most applications this "Pareto effect" is so strong that its statistical significance is obvious. However, checking for the assumption of a Poisson distribution, which we have just illustrated by use of **c-chart** in the example of Operator Errors, is useful in cases of doubt. Also, we can compare the mean and the square of the standard deviation (variance), since these two should be roughly equal if the Poisson assumption is valid.

Library Book Borrowing

The statistical approach just illustrated is widely applicable. The next application concerns book borrowing by the "customers" of a library. (It is also another example of a "lightning data set" -- the data contained in LIBRARY.sav took just a few minutes to collect during a visit to the library.) The descriptive notes are:

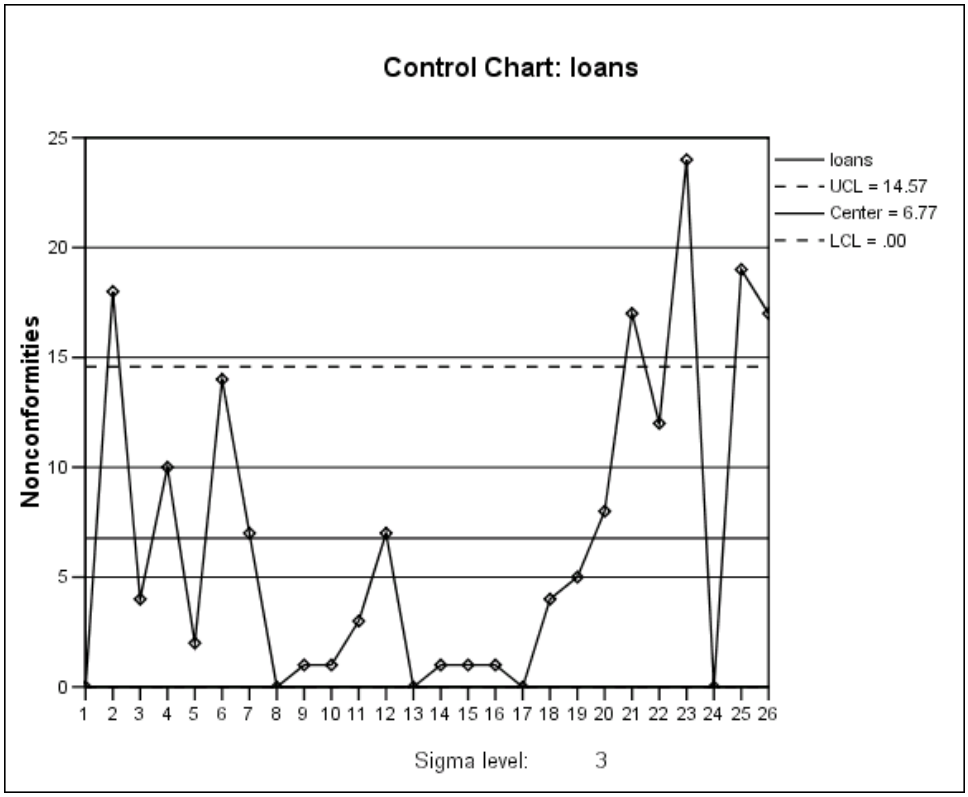
> Number of loans of books in the Morton Arboretum Library, catalog category QK 477, as of 20 November 1993. Data collected for 26 different books. Renewals not counted, but one borrower might account for more than one loan of a given book.

After naming the variable **loans**, for "number of loans", we execute **Descriptive Statistics:**

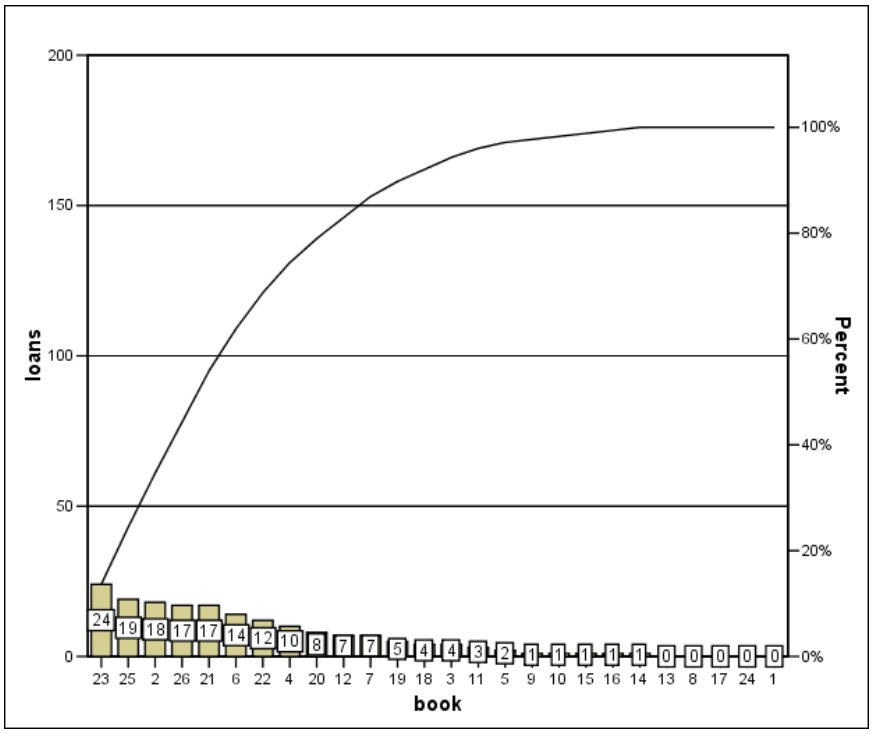
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
loans	26	0	24	6.77	7.279
Valid N (listwise)	26				

Note that 7.28*7.28 = 53.0 is much greater than the mean of 6.77. Hence, as is also shown by a **c-chart** below, the Poisson assumption is not tenable: some books are borrowed significantly more frequently than others.



Here is the Pareto chart:



Five of the 26 books account for over half of the total loans, and half the books account for 92 percent of the total. The Pareto effect is clearly at work here.

Air Delays in October, 1987

We can study histograms to look for systematic and special causes in applications other than those entailing counting data. The application below is based on percentages of flights delayed for a cross section of airlines, and we can use the normal distribution as a rough reference point for outlier search. The data are in the file AIRDELAY.sav:

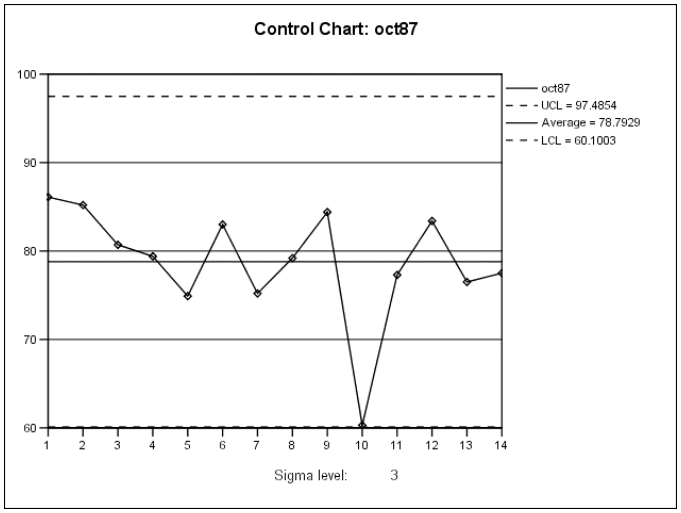
> Percent of airline flights reported arriving on time in October, November, and December of 1987. "On time" means within 15 minutes of schedule. From the *Wall Street Journal*, 9 February 1988, page 7. There are two small revisions for November as reported in the *WSJ* of 7 January 1988.

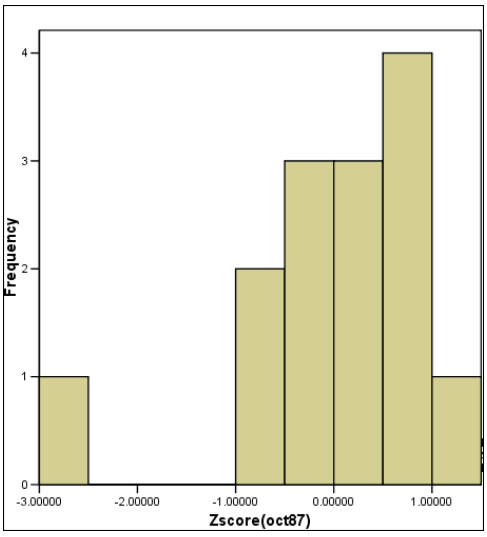
You see that the variables are named **airline, dec87, nov87**, and **oct87** as we open the data in ***SPSS***. Then we apply **Descriptive Statistics** to **oct87.** (This time in the **Descriptives…** setup window check the little box labeled **Save standardized values as variables**.)

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
oct87	14	60.30	86.10	78.7929	6.49372
Valid N (listwise)	14				

We are no longer working with counts so we will use a **control chart** to check on the behavior of the data. Remember that these are not time series data, but even when the sequence is irrelevant, the chart and its control limits can be useful in detecting outliers. We observe that Airline #10, **Pacific Southwest**, is right on the LCL. The histogram for **Zoct87** gives a similar result:





Pacific Southwest is a good candidate for search for an assignable cause since its on-time performance is much worse than that of the other thirteen airlines, as is shown by the simple histogram above. If, to the contrary, all fourteen airlines had been a part of the same "system" (to use an expression of Deming), it is reasonable that we should see a histogram resembling what would be expected from a

normal distribution; or at least that we should not see such a big gap between **Pacific Southwest** and the other thirteen airlines.

Are there significant differences among these other thirteen? In Section 6, we shall show a methodology for approaching this question; as we shall see, it requires more data than the on-time arrivals for one single month.

Performance Differences among Baseball Teams

The ***SPSS*** data file named BASEBALL.sav contains five variables, **won, lost, league, division**, and **city**. The first two variables are the numbers of games **won** and **lost** at the time that activity ceased because of the 1994 players' strike. The indicator variable **league** equals zero for the American League and one for the National League. For **division** the values are 1 for East, 2 for Central, and 3 for West. In the next few steps we shall look at the percentage of games won for each team. If the teams were about equal in their performances we would expect the distribution of those percentages to have the same shape and spread as a binomial distribution where the chance of winning is 0.50[2].

Before continuing, we must use **Transform/Compute…** to define two new variables:

total = won + lost and

pct = 100*won/total

Here is a partial glimpse of the ***SPSS*** worksheet after the transformations above:

	won	lost	league	division	city	total	pct
1	70	43	0	1	New York	113.00	61.95
2	63	49	0	1	Baltimore	112.00	56.25
3	55	60	0	1	Toronto	115.00	47.83
4	54	61	0	1	Boston	115.00	46.96

Next consider this display of **Descriptive Statistics**:

Descriptive Statistics

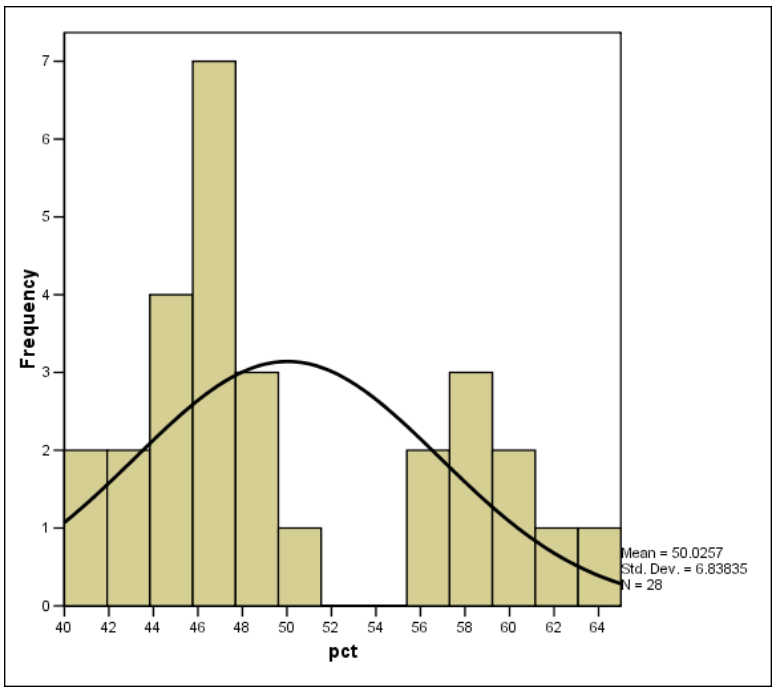
	N	Minimum	Maximum	Mean	Std. Deviation
total	28	112.00	117.00	114.2143	1.22798
pct	28	40.17	64.91	50.0257	6.83835
Valid N (listwise)	28				

We see that the average percentage of wins is close to 50 percent, and the mean number of games played by each team is about 114. If the performances of the teams differ only by chance, then we can approximate the standard deviation of **pct** by the formula

$$SD = \frac{100 * 0.5}{\sqrt{114}} = 4.6829$$

[2] If this is not intuitively obvious, think of each game as a coin toss with the chance of heads equal to the chance of tails (assuming that at each game the teams are evenly matched).

which comes from the mathematical theory of the binomial distribution. The actual standard deviation, however, is 6.84-- substantially larger than 4.68. Hence it appears, as we would expect, that the teams differ by more than chance, and we must reject the hypothesis that they are from the same general "performance process." The histogram shows something even more interesting:



The distribution is bimodal (It has two peaks). **There are too few teams close to 50 percent, the overall average.** (Note: To make your own histogram look that same as that above you will have to go into the Chart Editor and fiddle with the number of intervals and their width.)

INTLEAGUE.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : won 78

	won	lost	division	city
1	78	64	0	Pawtucket
2	71	71	0	Syracuse
3	70	72	0	Ottawa
4	67	74	0	Rochester
5	62	80	0	Scranton_WB
6	80	61	1	Richmond
7	77	65	1	Charlotte
8	74	68	1	Columbus"
9	67	75	1	Norfolk
10	63	79	1	Toledo
11				

For a contrasting example, the data set on the left gives no evidence that anything other than chance factors captured by the binomial distribution are at work. These data are the final 1994 standings for the International League. You may enjoy working with them!

3. Simple Cross-Sectional Regressions

When we have more than a single variable in cross-sectional applications, we can use regression tools in much the same way as for time-series data, but we have to be even more cautious about causal interpretation. In this section, we consider two marketing applications to illustrate both the regression mechanics and the interpretation of results.

Sales Proposals and Sales

The first example, contained in SALEPROP.sav, entails data from a cross-sectional sample of 31 sales representatives. It leads to a simple regression analysis in which the dependent variable is sales for each sales representative and the independent variable is the number of written proposals prepared by the sales representative, each for a single month's time. Note that the data are listed by number of proposals, from low to high. There is no time sequence: there are 31 sales representatives for one single time period.

> Study of the relation of sales volume and selling proposals
> prepared by sales representatives in geographical territories of
> the Chicago branch of an office supplies company during one month.
> "Management believed that there is a direct relationship between
> the number of proposals a sales representative prepares and the
> dollar volume of sales achieved in that month. It is therefore
> company policy to require each person to prepare at least eight
> written proposals per month."
> **sales$**: sales in $1000.
> **numprop**: number of proposals.

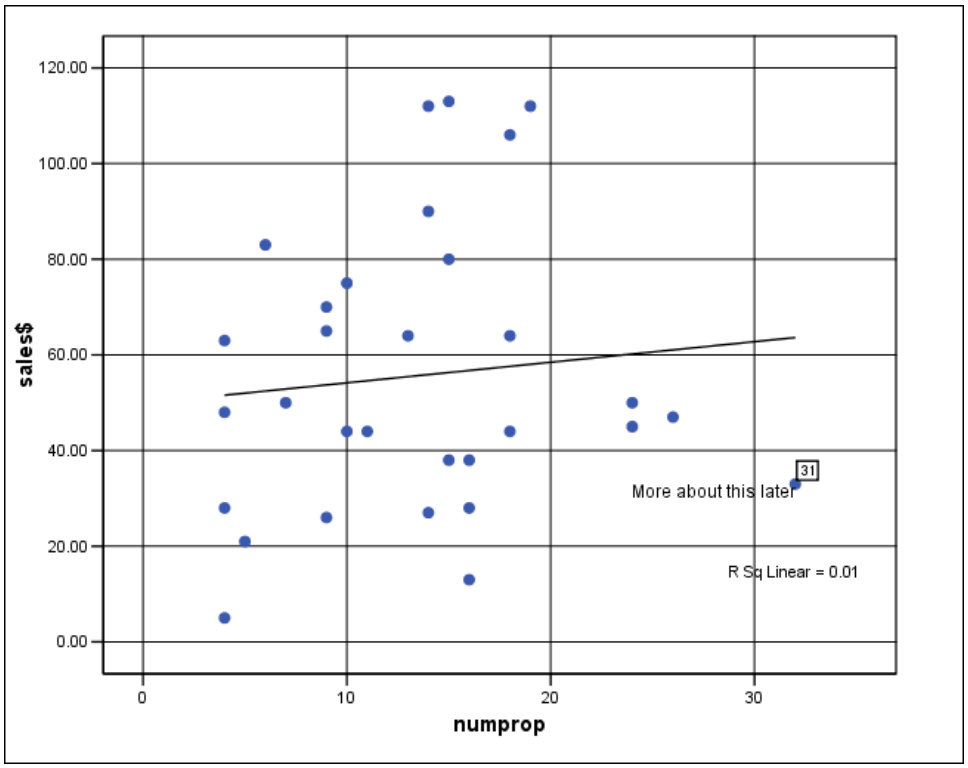
Here are the results of **Descriptive Statistics**:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
sales$	31	5.00	113.00	55.6774	29.56618
numprop	31	4	32	13.52	6.976
Valid N (listwise)	31				

We see that mean sales for the 31 salespersons was $55,677 and the mean number of proposals written was about 13 and a half.

Next, we perform a scatter plot with **sales$** on the vertical axis and **numprop** on the other, followed by a simple linear regression analysis:



The regression line does not appear to have a very large slope nor is R Square large. There doesn't seem to be much of a linear relationship between **sales$** and **numprop**.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.102[a]	.010	-.024	29.91624

a. Predictors: (Constant), numprop

b. Dependent Variable: sales$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	49.861	11.869		4.201	.000	25.586	74.136
	numprop	.430	.783	.102	.550	.587	-1.171	2.032

a. Dependent Variable: sales$

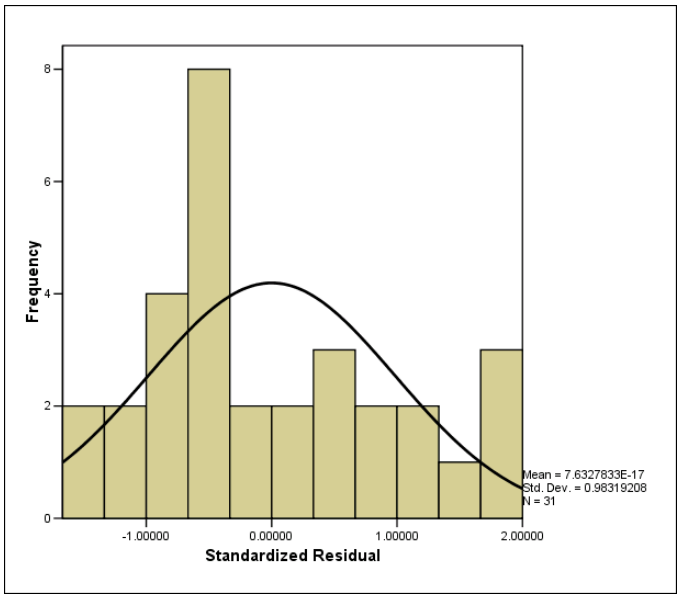
The regression coefficient 0.430 for **numprop** is not even close to having a p-value that would be called significant, although it is positive. (If taken at face value, it would suggest an average increase of $430 in sales for each proposal since **sales$** is expressed in units of $1000.)

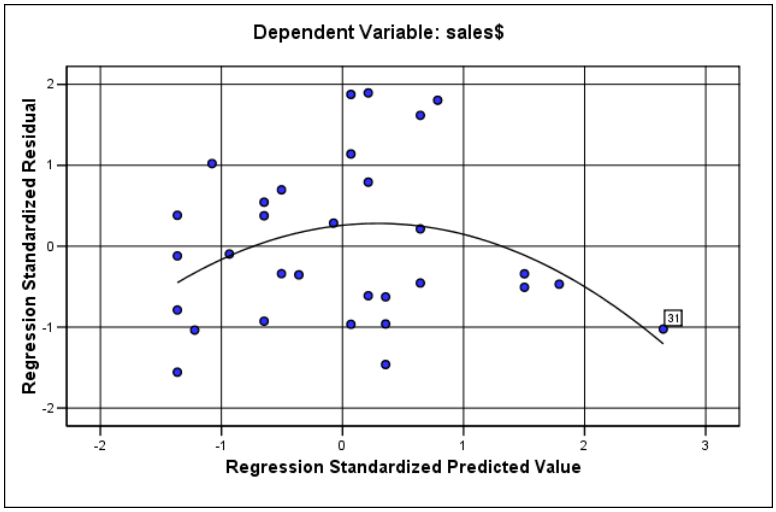
We have not discussed confidence intervals since Section 14 of Chapter 2, but the concept is always useful. In this case we can add and subtract 2*0.78301 from 0.43032 to obtain approximate 95

percent confidence limits for the slope at -1.136 and +1.996. Since the confidence interval contains zero, our conclusion is the same as that from examining the p-value. Note from the display above that we called for confidence intervals while setting up the regression analysis. The more precise limits for the coefficients of **numprop** are -1.171 and 2.032, a little wider than the approximation using plus and minus two as the multiplier of the standard error.

Since the regression does not yield a significant result, we generally would not do diagnostic checking. However, to illustrate the diagnostic checking in cross-sectional regression, we show the two checks that are generally applicable in cross-sectional regression:

- A **histogram** for standardized residuals.
- **Scatter Plot** for the standardized residuals vs. the standardized predicted values.





There is just a hint of nonlinearity in this scatter plot: the small and large fitted values tend to have small residuals and intermediate fitted values tend to have large residuals. Examination of the values of **ZRE_1** would show that the problem is caused by observation 31 at the lower right of the plot.

A point such as 31 is sometimes referred to as an **influence point** because, as you can see in the scatter plots, if it were removed from the plot of **sales$** on **numprop**, the remaining points would be more suggestive of positive correlation. (Imagine moving point 31 even lower and farther to the right-- it would really pull the regression line downwards.)

One way to explore this aspect of the data is to model the possible nonlinearity directly by introducing a new variable

numpropsq = numprop*numprop

and to regress **sales$** on both **numprop** and **numpropsq** as shown below:[3]

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.365[a]	.133	.071	28.48966

a. Predictors: (Constant), numpropsq, numprop

b. Dependent Variable: sales$

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3498.277	2	1749.139	2.155	.135[a]
	Residual	22726.497	28	811.661		
	Total	26224.774	30			

a. Predictors: (Constant), numpropsq, numprop

b. Dependent Variable: sales$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	18.565	19.340		.960	.345	-21.051	58.181
	numprop	5.541	2.669	1.307	2.076	.047	.074	11.008
	numpropsq	-.164	.082	-1.256	-1.994	.056	-.333	.004

a. Dependent Variable: sales$

At best, the regression shows borderline significance. The p-values for the coefficients of **numprop** and **numpropsq** are on either side of 0.05. But **numprop** and **numpropsq** are themselves highly correlated, and this makes the interpretation of the individual coefficients tricky (we'll see more of this in Section 5).

What we need is an assessment of overall significance of the regression model, which now includes two independent variables, **numprop** and **numpropsq**. As we have explained earlier, this assessment is available from the last part of the output showing sources of the sum of squared deviations. The key

[3] As we say in Chapter 6 in connection with fitting a nonlinear time trend in the Diet application, the mathematical basis for this approach is as follows: $Y = a + bX + cX^2$ is the equation a second-degree polynomial (parabola). This equation is often useful in modeling "gently curving" nonlinear relationships.

number is the p-value of **0.135** on the line for **Regression**. This gives a significance assessment for the whole regression model. The number 0.135 does **not** suggest significance, since 0.135 is substantially larger than the rule of thumb, 0.05. So at best there is only a hint of a relation between proposals and sales.

In one sense the study was disappointing: management's beliefs about the effectiveness of written proposals were not borne out. However, even if there had been a strong and positive relationship between sales and the number of written proposals, there would have been ambiguity about the causal interpretation. For example, it could have been possible that the best sales representatives were also well organized and could turn out more written proposals, yet at the same time the proposals themselves might have had little or no effect on sales.

For a cleancut causal interpretation of a regression of Y on X, we would need some assurance that the observed relation in the regression would not be substantially altered if other variables affecting sales -- say X1, X2, and X3 -- had been measured and introduced into the model.

However, although the study is not conclusive, its results cast some doubt on the advisability of management pressure to increase the number of written proposals. There may be other, better, routes to improved sales performance.

One circumstance that provides greater clarity of causal inference is given by **randomized experimentation**. Here, for example, the number of proposals to be made by each sales representative would be determined at random; remember the time-series application of randomized experimentation in Section 3 of Chapter 5.

Absent randomized experimentation, we must simply make the best judgments about causation that we can, given all our knowledge about the application.

This sales study was done many years ago; at that time the following reservations were expressed by the executive MBA student who did it:

- The **quality** of the proposals was not considered, although records existed.
- It might have been worthwhile to try to directly trace sales that could be attributed to specific proposals.
- There may have been a time lag. For example, proposals made this month might have little effect on current sales but substantial effect on sales in succeeding months.

Selling Time and Sales

Another study made nearly two decades later by another executive MBA student illustrates the same points but presents a more challenging statistical challenge. The student commented as follows about the background of his investigation:

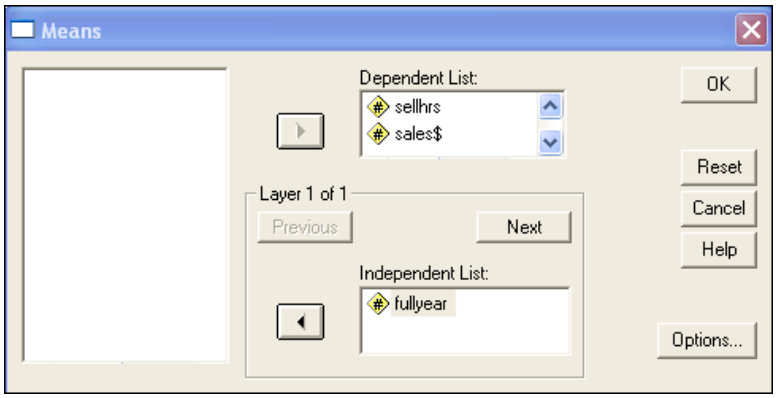
> "One of the great commandments of selling is that the more time spent selling, the more orders are landed. My experience has been with a firm that meticulously keeps track of individual selling

hours and individual sales. These are then reported back to each salesman once a month. Sales managers use their copies of such reports to monitor their operations.

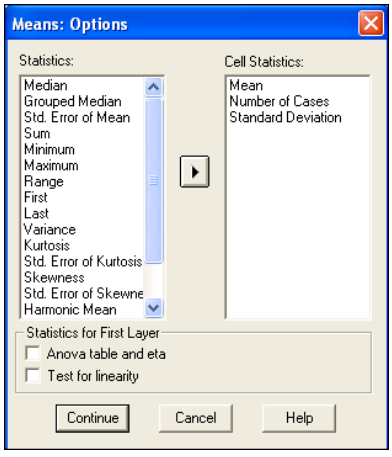
The purpose of my study is to quantify the dependence of sales results on sales efforts. I make simplifying assumptions: accuracy of data, uniformity of sales skills; uniformity of competitive pressures; etc. These assumptions are needed whenever statistics are used for comparative ranking of individual performance."
The data for this example are contained in SELLTIME.sav. The variables are:

sellhrs: hours spent in selling
sales$: sales in units of $1,000
fullyear: an indicator variable equal to 1 if full year spent in sales, 0 otherwise.

To display the descriptives we use the sequence **Analyze/Compare Means/Means…**, with the dialog window looking like this



and selected options as follows:

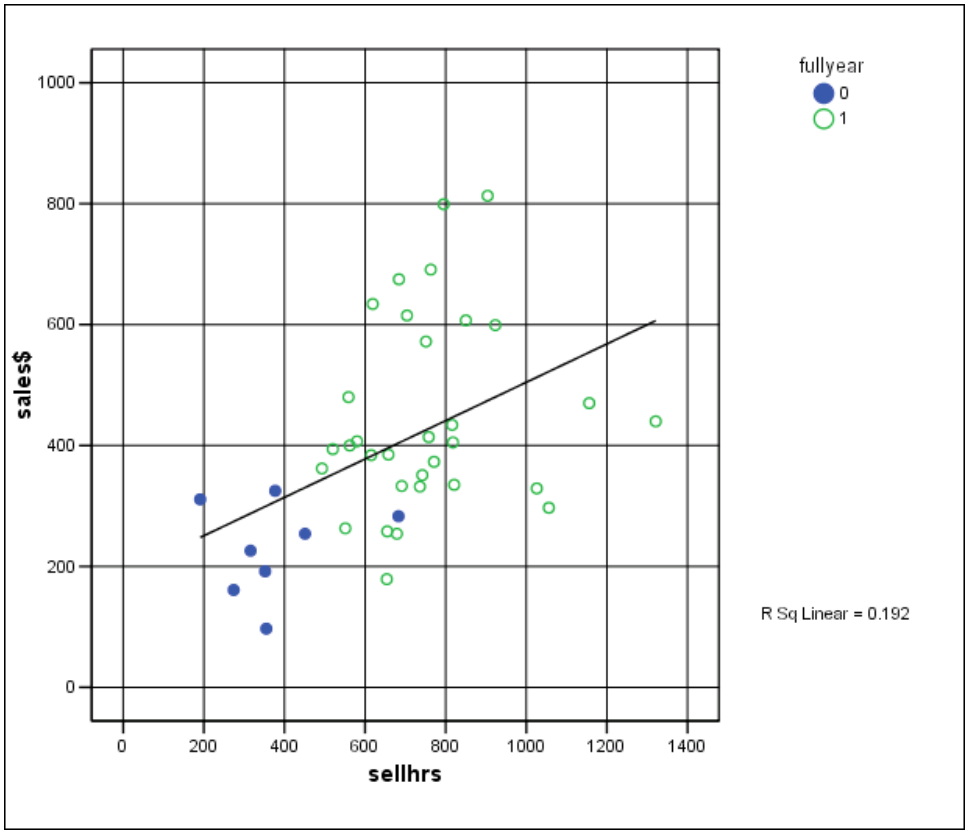


Here is the resulting report:

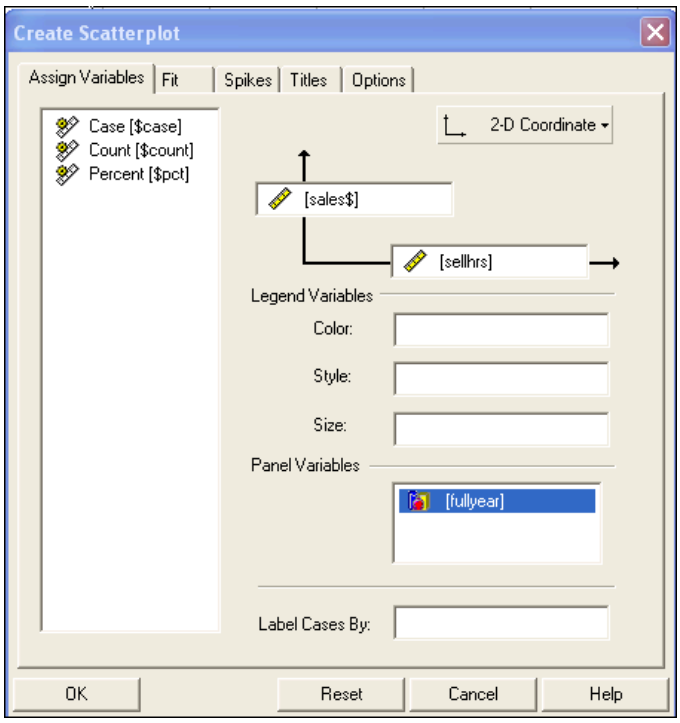
Report

fullyear		sellhrs	sales$
0	Mean	374.88	231.13
	N	8	8
	Std. Deviation	145.951	78.259
1	Mean	757.22	446.38
	N	32	32
	Std. Deviation	186.306	159.815
Total	Mean	680.75	403.33
	N	40	40
	Std. Deviation	235.378	170.306

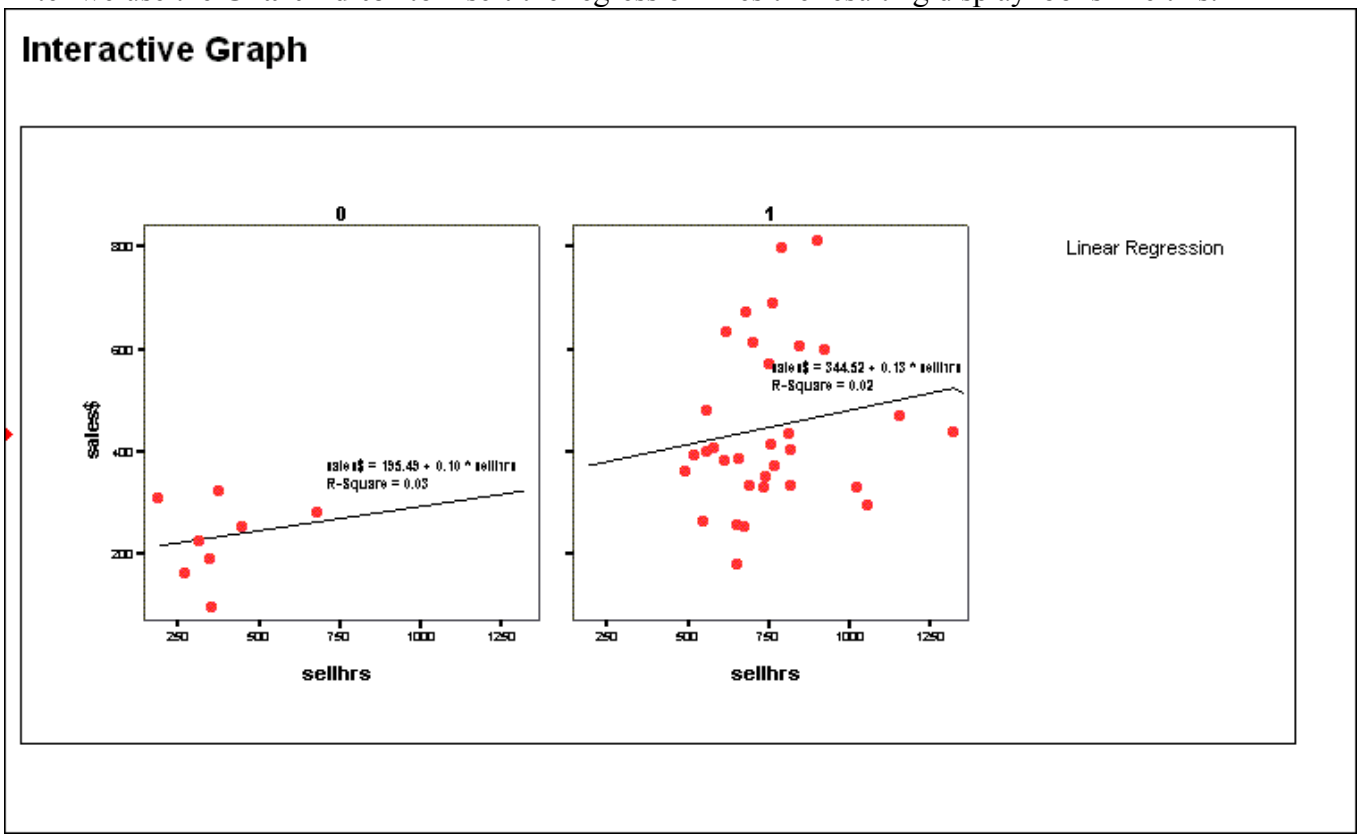
Note that mean **sellhrs** and mean **sales$** are both larger for the 32 sales representatives who worked for the full year. Also make a mental note that the standard deviation of **sales$** is greater for those who worked for a full year; this will have technical consequences that we shall examine later.



This scatter plot suggests a positive relation between **sales$** and **sellhrs**, with possibly higher variance of **sales$** as **sellhrs** increases. We see, however, that if we take **fullyear** into account the relationship will change considerably. We therefore make a new plot, but this time using **Graphs/Interactive/Scatter Plot…** with the following setup:



After we use the **Chart Editor** to insert the regression lines the resulting display looks like this:



The two regression lines appear to be almost parallel, but the values of **R Square** show that the separate linear relationships are much weaker than when the data were combined.

We now pursue the investigation by applying **Stepwise Regression** with **sales$** regressed on both **sellhrs** and **fullyear** to see which, if either, of these variables contributes to the fit.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.512[a]	.262	.243	148.203

a. Predictors: (Constant), fullyear

b. Dependent Variable: sales$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	231.125	52.398		4.411	.000
	fullyear	215.250	58.582	.512	3.674	.001

a. Dependent Variable: sales$

Excluded Variables[b]

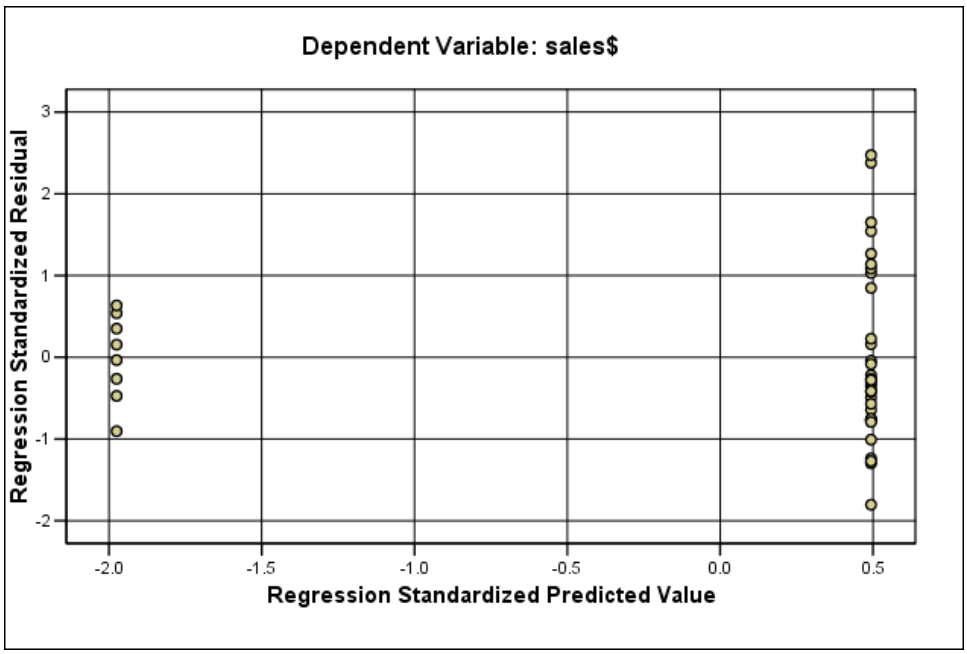
Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	sellhrs	.179[a]	.968	.339	.157	.567

a. Predictors in the Model: (Constant), fullyear

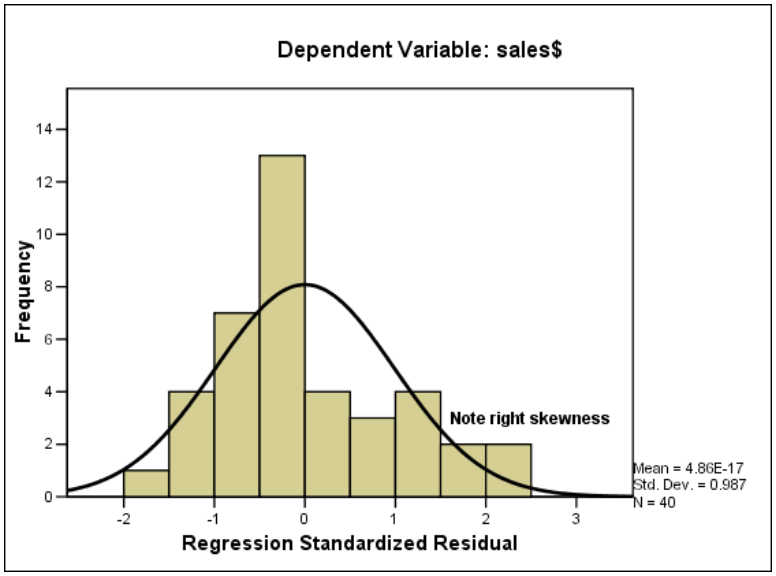
b. Dependent Variable: sales$

Only **fullyear** contributes significantly! We see at the end of the stepwise output that if **sellhrs** were forced into the regression model, it would have a positive coefficient but its t-ratio would only be 0.968-- not statistically significant.[4]

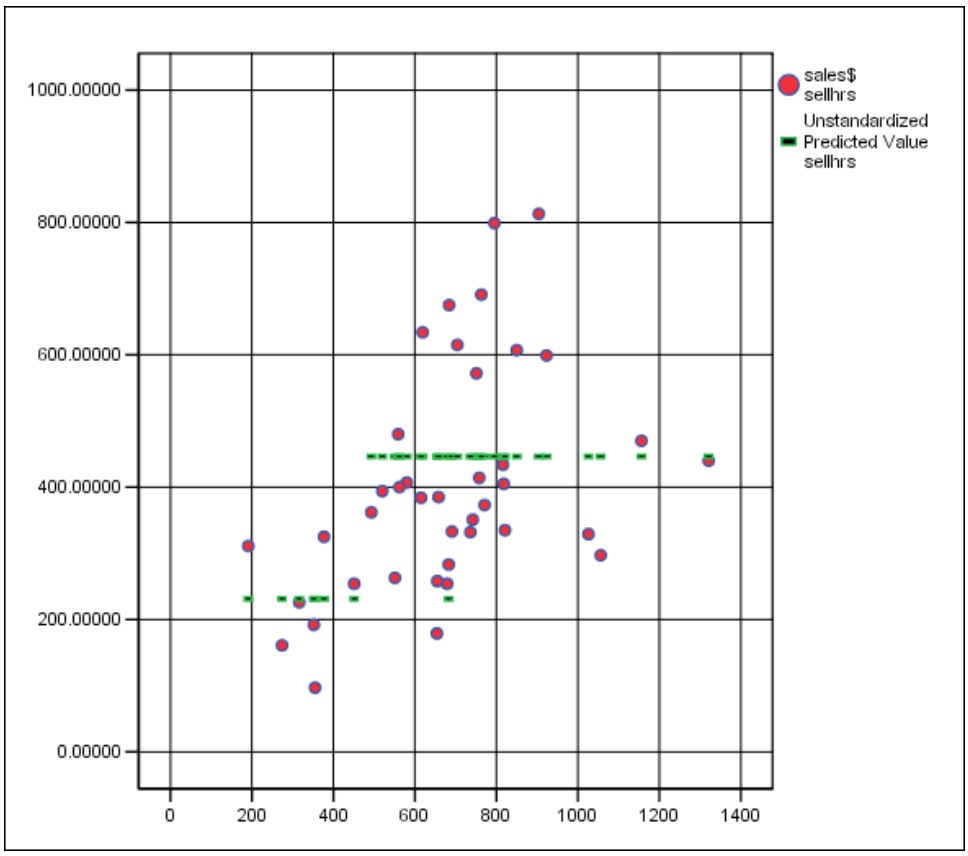
[4]The **partial correlation**, 0.157, for **sellhrs** is a measure of association with **sales$** after the contribution of **fullyear** is already taken into account. Its positive sign indicates that if **sellhrs** were also in the regression model its regression coefficient would be positive.



Notice that the variance of the residuals for the full-time sales representatives is larger, reflecting what we saw earlier with **Descriptive Statistics**.



The moderate skewness in the right tail of the histogram above suggests the consideration of a transformation to natural logarithms. We shall discuss this possibility a little later.



In the scatter plot above we show how the regression model fits the data. There are two horizontal regression lines formed by the solid dashes, the higher one for the full-year sales representatives.

The student's conclusions:

1. "There is a relatively weak association between selling hours and sales results for the same calendar time period.

2. "Use of selling time and sales has limitations as a managerial tool. Hours selling is definitely a measure of effort expended; it does not follow that sales will result from this effort.

3. "The use of sales and selling time to rank and compare individuals is not valid based on these data.

4. "The assumption of equal market potential is the most crucial. My analysis suggests a re-examination of sales force deployment with respect to geographic market potential."

The first three of these conclusions are well on target. The fourth is a bit vague, but seems to be aiming in the right direction. What is meant is that skepticism about the value of proposals would be most nearly justified if the "potential" of each territory were similar; in statistical terms, variables other than **numprop** would have little effect on **sales$**. If this assumption is not tenable, then there is greater reason to believe that allowance for differences in these other variables might put the effect of proposals in a more accurate perspective.

Transformation for Nonconstant Variance

The model just fitted suffices for practical purposes, but two of the diagnostic checks are ragged: the results are somewhat right skewed and the variance of residuals is higher for the full-year sales representatives. Although the practical conclusions will not be affected, we can use this application to illustrate transformation to express data on a scale on which the key regression assumptions -- normality and constant variance -- may be more nearly satisfied. The discussion that follows is based on the assumption that negative values are impossible. (For the logarithmic transformation, developed below, zero values are ruled out.)

First, it is helpful to provide a simple check for whether a transformation even needs to be considered. We apply **Descriptive Statistics** to **sales$**, this time without grouping:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
sales$	40	97	813	403.33	170.306
Valid N (listwise)	40				

There is a simple rule based on the ratio between the standard deviation and the mean, here 170.306/403.33 = 0.42. If this ratio is greater than 0.10 (10 percent), as it is here, then a transformation of SALES **may be** useful for statistical analysis. **If not, then you need not even consider a transformation.**

In earlier chapters, we have already made use of two transformations: the square root (for counting data that may conform approximately to the Poisson distribution) and the cube root (for data that may conform approximately to the exponential distribution). A third transformation, the logarithmic, is also useful -- as we saw in modeling Calories in the diet example of Chapter 6. We can think of a sequence of possible transformations:

square root ➲ cube root ➲ logarithmic[5]

[5]Still another step is the reciprocal, 1/Y, which can be helpful in certain applications. If, for example, the dependent variable in an automobile study is "miles per gallon", the reciprocal transformation would convert to "gallons per mile".

The successive transformations are increasingly drastic in altering the pattern of variation. Most applications presenting nonconstant variance, including the present application to sales, show a higher variance of the dependent variable at higher levels. All three transformations will make the variance of the **fullyear** sales reps more nearly equal to that of the other reps, but we must look more closely to see which works best in the current application.

The log transformation works very well to stabilize variance when the standard deviation of the variable of interest increases proportionally to variable itself. For example, this means that if the level of sales doubles, the standard deviation of sales will also double. Often this is approximately true for business and economic data, such as company sales, stock prices, or the gross domestic product. If you look back to the original **Descriptive Statistics** output in the current application, you will see that the **fullyear** sales reps have roughly twice the mean **sales$** and twice the standard deviation of **sales$** as the less-then-full year reps. (When the log transformation does stabilize variance, we can interpret this to mean that the **percentage changes** of the original series show constant variance through time.)

It turns out that both the cube root and log transformations are good approximations in the current application, and that the cube root is slightly preferable. You will find it instructive to work through the details below, and to see that the diagnostic checks are somewhat better than for the analysis developed above.

The first step is to use **Transform/Compute…** to create

curtsales$ = sales$(1/3)**

Report

curtsales$

fullyear	Mean	N	Std. Deviation
0	6.0555	8	.76747
1	7.5407	32	.89173
Total	7.2436	40	1.04869

Note that the **standard deviations** for the two rep groups are much closer now.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.574[a]	.329	.311	.87017

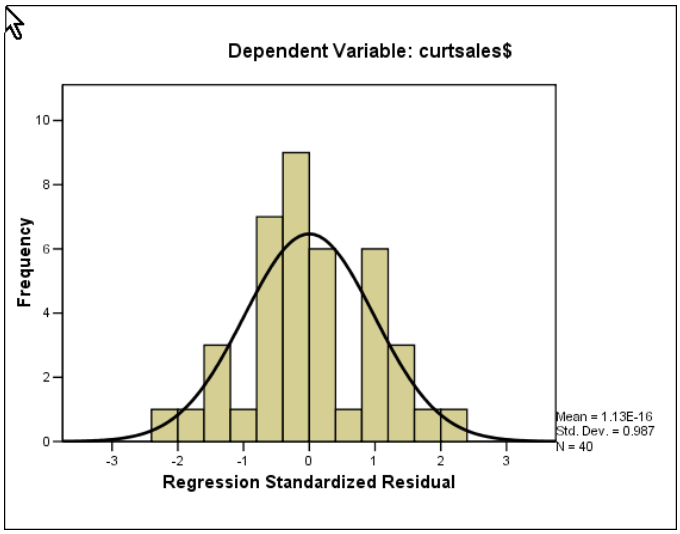
a. Predictors: (Constant), fullyear

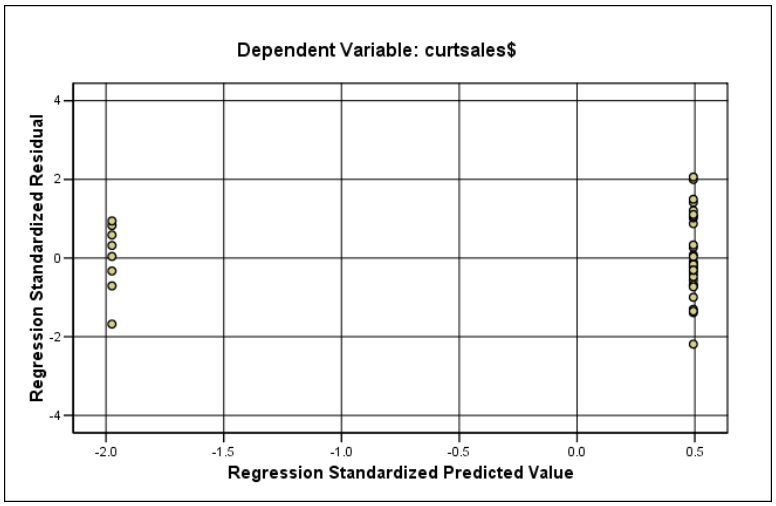
b. Dependent Variable: curtsales$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	6.055	.308		19.683	.000
	fullyear	1.485	.344	.574	4.318	.000

a. Dependent Variable: curtsales$





Note the improved symmetry of the histogram of standardized residuals and that the spreads of the residuals for **fullyear** and **less-than-fullyear** are more nearly the same.

Lightning Data Set: Predictability of Professional Football

Another example of cross-sectional regression is provided by the data for the opening week of the 1994 NFL season contained in BETLINE.sav:

BETLINE.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : betline 3

	betline	vpts	hpts	townsnd	betltot	game
1	3.00	30.0	17.0	1.0	39.00	KansasCity at New Orleans
2	-3.00	23.0	28.0	-1.0	37.00	Philadephia at New York Giants
3	-3.50	10.0	16.0	-1.0	38.00	Minnesota at Green Bay
4	-5.00	28.0	31.0	-1.0	45.00	Atlanta at Detroit
5	4.00	28.0	20.0	1.0	36.00	Cleveland at Cincinnati
6	3.00	21.0	45.0	1.0	38.00	Houston at Indianopolis
7	-1.50	28.0	7.0	-1.0	36.00	Seattle at Washington
8	-6.50	9.0	21.0	1.0	35.00	Tampa Bay at Chicago
9	6.00	12.0	14.0	-1.0	36.00	Arizona at LA Rams
10	5.50	26.0	9.0	-1.0	39.00	Dallas at Pittsburgh
11	-5.50	35.0	39.0	1.0	38.50	New England at Miami
12	-7.00	23.0	3.0	-1.0	41.00	NY Jets at Buffalo
13	-7.00	37.0	34.0	-1.0	45.00	San Diego at Denver
14	-7.00	14.0	44.0	-1.0	45.50	LA Raiders at San Francisco
15	.	.	.	.	.	

The outcome of each game is expressed as the visitor's points, **vpts**, minus the home team's points, **hpts**. One prediction of outcome is provided by **betline**, the betting line spread. A second is **townsend,** the prediction of a sports writer, Murray Townsend, equal to **1** if he predicts the visitors to win and **-1** if he predicts the home team to win. The last variable, **betltot,** is the betting line prediction of the total points for the game. It turns out that these variables do not come close to statistical significance, as the summary analysis below shows. Of course, the sample consists of only 14 games. Other studies, using more data, have shown that the betting line spreads do have some predictive power, but it is surprisingly low. There is a lot of uncertainty in NFL games!

Before we can begin the analysis we must transform as follows:

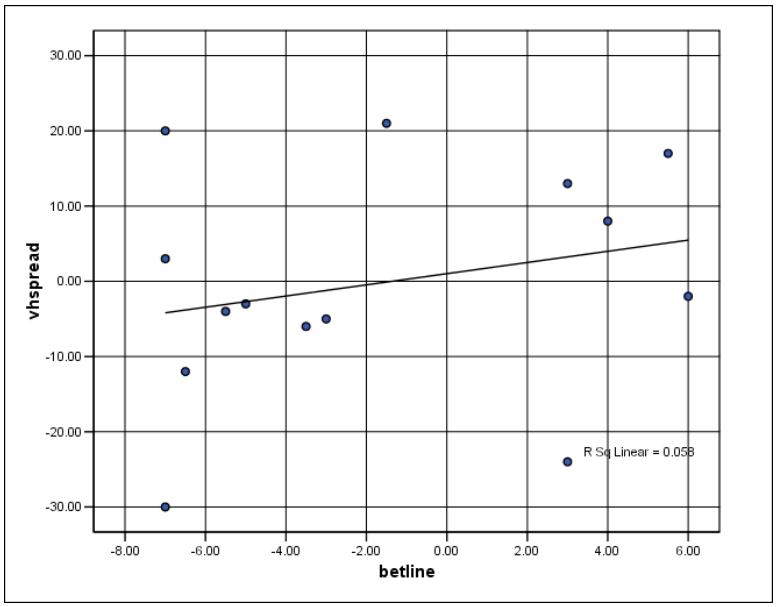
vhspread = vpts - hpts

totpts = vpts + hpts

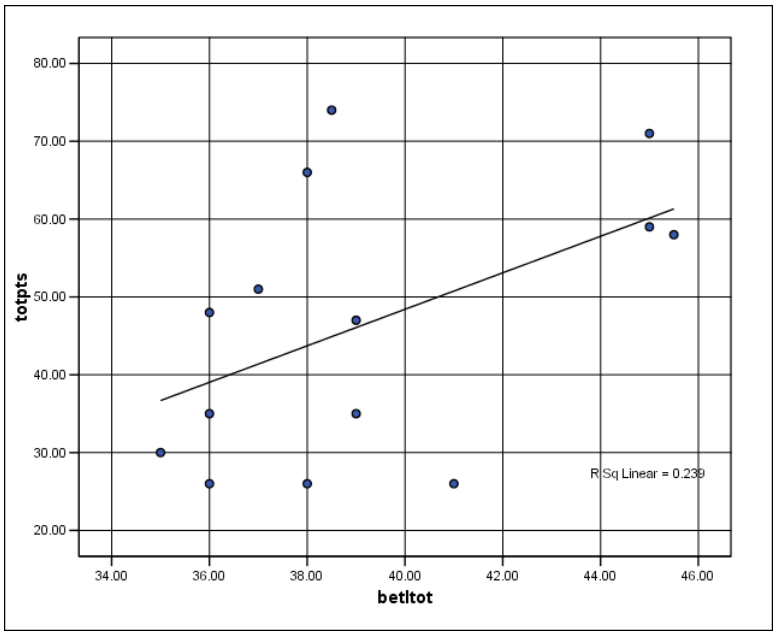
Here is what the **Data Editor** looks like now:

	betline	vpts	hpts	townsnd	betltot	game	vhspread	totpts
1	3.00	30.0	17.0	1.0	39.00	KansasCity at New Orleans	13.00	47.00
2	-3.00	23.0	28.0	-1.0	37.00	Philadephia at New York Giants	-5.00	51.00
3	-3.50	10.0	16.0	-1.0	38.00	Minnesota at Green Bay	-6.00	26.00
4	-5.00	28.0	31.0	-1.0	45.00	Atlanta at Detroit	-3.00	59.00

The actual regressions are not shown, but you can see from this plot that there is no clear relationship between **betline** and **vhspread**.



There is also a betting line on total points -- **betltot** -- and here we finding borderline significance with actual total points by the two teams in each game:



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.488[a]	.239	.175	15.62010

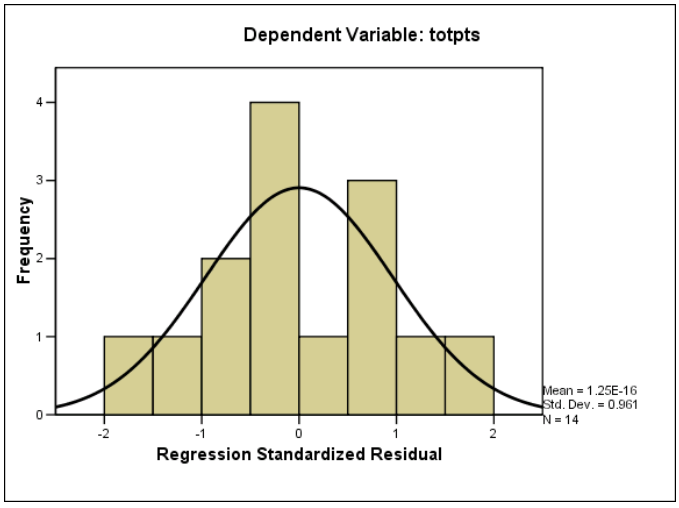
a. Predictors: (Constant), betltot

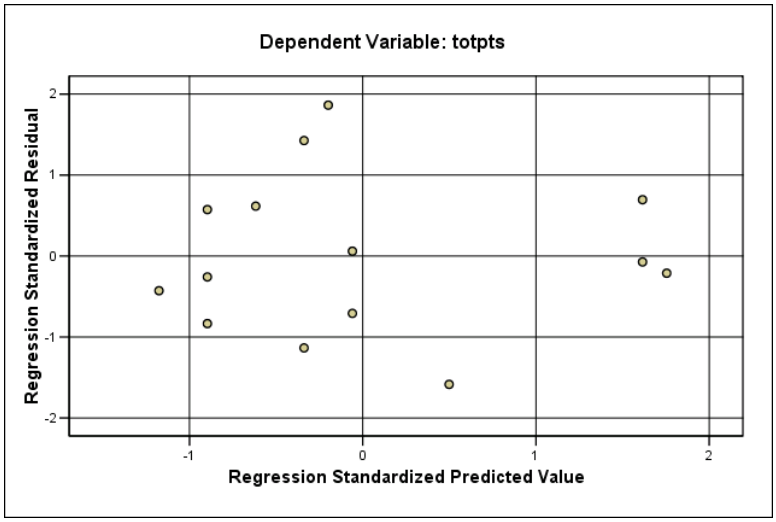
b. Dependent Variable: totpts

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-45.387	47.603		-.953	.359
	betltot	2.345	1.209	.488	1.939	.076

a. Dependent Variable: totpts



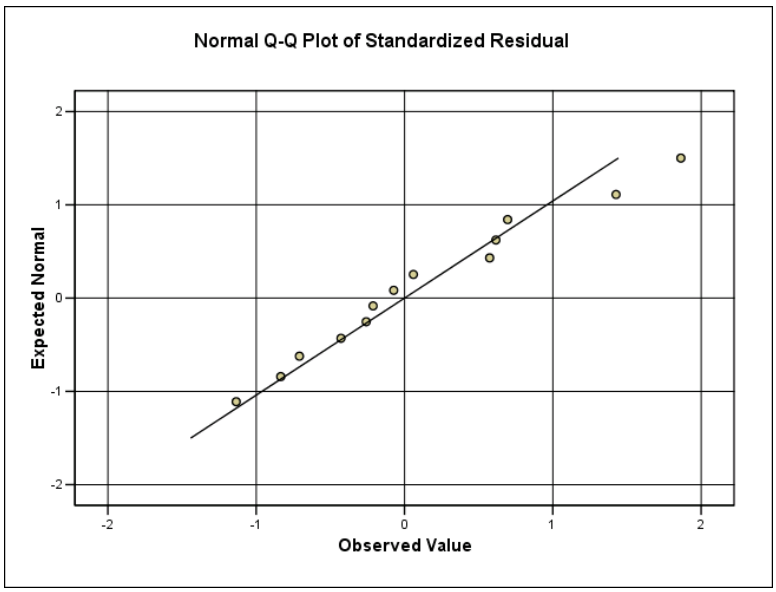


Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.118	14	.200*	.978	14	.963

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

a. Lilliefors Significance Correction

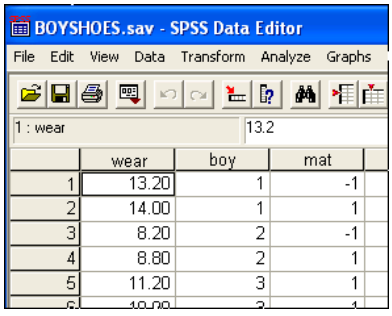


4. Paired-sample Experiment for Cross-Sectional Data: BOYSHOES

The next application illustrates how randomized experiments can be performed with cross-sectional data. This is a cross-sectional analog to the paired sample design in Section 4 of Chapter 5, which was used in the study of the possible effect of hyperventilation on blood pressure readings. You may wish to review that section quickly before reading ahead.

In the current application (taken from Box, Hunter, and Hunter, *Statistics for Experimenters*, Wiley, 1978), the objective was to find out whether sole material B, a less expensive substitute for the standard sole material A, would result in greater wear for boys shoes. Since there is apt to be enormous variation between boys in the amount of wear, but little variation for a given boy between the right and left shoe, material A was used for one shoe and material B for the other. The "pair" in this randomized pair experiment was therefore the two shoes of a given boy. The assignment of material A or B to right or left foot was done randomly.

It is necessary to use an indicator variable for the type of material assigned. There is a slight advantage to expressing this variable not as 0 and 1, but -1 and +1. As we have been doing, we include all relevant variables in a stepwise regression to identify the significant ones, then examine the selected model in detail. We begin by looking at the contents of the file BOYSHOES.sav:

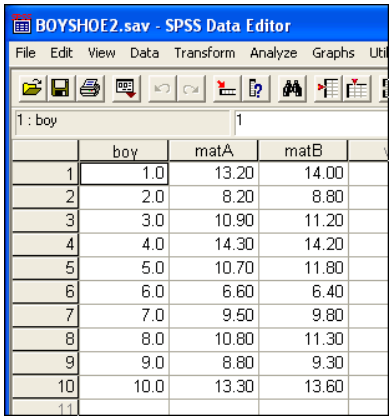


The variable **wear** is of principal interest—it is the amount of wear on the shoe identified by **mat** (-1 for Material A, 1 for Material B).The next step is use **Transform/Compute…** to create indicator variables for each boy, naming the 0-1 variables

boy1 boy2 boy3 boy4 boy5 boy6 boy7 boy8 boy9 boy10

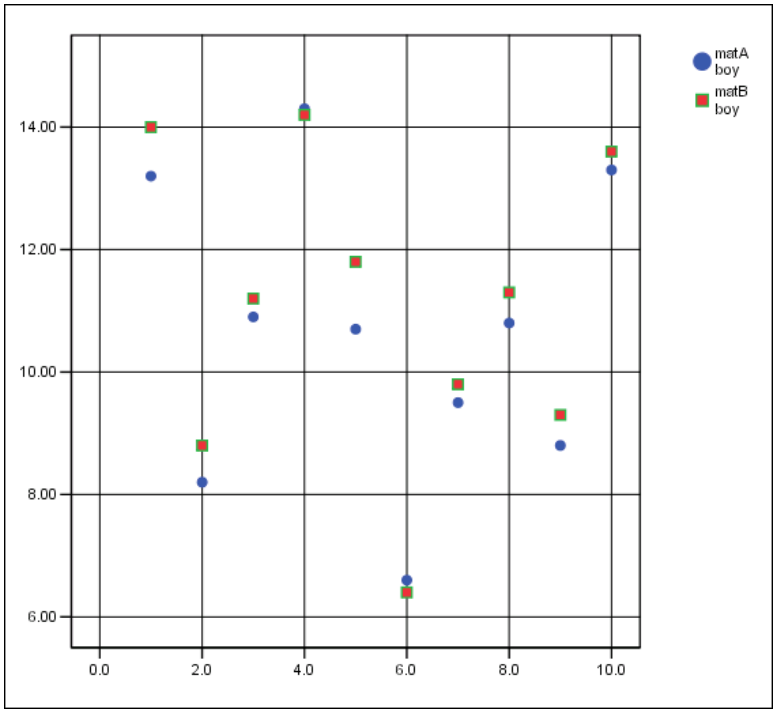
similar to our setup of the hyperventilation example on page 5-48.

If you have the time and the inclination it will be very instructive for you to continue the direct linear regression analysis, following the setup and the approach that we used back in Chapter 5. We shall not, however, illustrate that analysis here You will recall that **Paired-Samples T Test** was demonstrated in Appendix 2 of Chapter 5. In the blood pressure example we changed the data set into the form that is required for the **Paired-Samples T Test** by means of some rather complicated sorting. To avoid that complication with the present example, we have created the file BOYSHOE2.sav:



	boy	matA	matB
1	1.0	13.20	14.00
2	2.0	8.20	8.80
3	3.0	10.90	11.20
4	4.0	14.30	14.20
5	5.0	10.70	11.80
6	6.0	6.60	6.40
7	7.0	9.50	9.80
8	8.0	10.80	11.30
9	9.0	8.80	9.30
10	10.0	13.30	13.60
11			

Each row of BOYSHOE2.sav consists of two readings, the first for Material A and the second for Material B. This layout enables us to make the following scatter plot. Notice how close to each other are the readings for the two shoes worn by each individual boy. This confirms the realism of the assumption that the differences between boys would be more substantial than the differences between the two shoes of individual boys.



We can also see that for most of the pairs Material B shows more wear than Material A. We shall have to wait until our analysis is complete, however, to see whether the difference between B and A is due to more than chance variation. Here is the full output from that analysis after applying the ***SPSS*** procedure **Analyze/Compare Means/Paired-Samples T Test…:**

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	matA	10.6300	10	2.45133	.77518
	matB	11.0400	10	2.51847	.79641

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	matA & matB	10	.988	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	matA - matB	-.41000	.38715	.12243	-.68695	-.13305	-3.349	9	.009

The results show that the estimated difference between the mean **wear** for Material A and Material B is -0.41. In other words, on average, Material A shows -0.41 units less **wear** than B. The finding is highly statistically significant as shown by a t-ratio of -3.349 and a p-value equal to 0.009.

Note also (in the second table above) the very high correlation, 0.988, between the measurements for each foot on the ten boys. This confirms the superiority of the paired-sample design, i.e., because there is much less variability within a boy than among the various boys. Suppose that instead of pairing the experimenters had assigned Material A to one sample of ten boys and Material B to a separate and independent sample of ten other boys, and that they had applied **Analyze/CompareMeans/Independent-Samples T Test…**. If, by chance, one sample of boys were less active than the boys in the other sample, the difference in materials could have been obscured. We might say, then, that pairing has controlled for the unseen variable, "**activity level**".

5. Car Gas Mileage, Studied by Cross-Sectional Regression

Our next application, from the file MPGCARS.sav, is a rather typical example of a study designed to tease causal inference from cross-sectional data. The data set below is an extract made from a much larger data set collected by a University of Chicago MBA student, Edmund Muth, during the fall of 1980. From a larger number of cars studied, the extract includes only non-diesel sedans listing at less than $13,000 in 1980, and it includes only selected variables from a larger list. Muth obtained most of the data from *Car and Driver* magazine, a well-respected publication that had established standardized testing procedures and methods for each of the variables used in the study. Do not be put off by the fact that these data were obtained almost twenty-five years ago, before some of you were born. The basic statistical analysis shown below follows familiar lines and is still relevant. The variables are defined as follows:

auto: Name and model of automobile
mpg: EPA estimates of mile per gallon
V_engine: Indicator, =1 for V6 or V8 engine, 0 otherwise
hp: Horsepower

weight: Weight in pounds
numcyl: Number of cylinders
japan: Indicator, =1 for Japan, 0 for U.S. or Europe
age: Age of design in years

Here are the basic descriptive statistics:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
mpg	35	14.0	36.0	22.800	6.2299
V_engine	35	0	1	.26	.443
hp	35	55.00	205.00	101.7429	36.77223
weight	35	1800.00	3740.00	2679.0000	610.28658
numcyl	35	4.00	8.00	5.0857	1.63368
japan	35	0	1	.29	.458
age	35	1.0	15.0	3.400	3.3184
Valid N (listwise)	35				

In studies such as this one, it often a good preliminary step to obtain the correlations among the variables:

Correlations

		mpg	V_engine	hp	weight	numcyl	japan	age
mpg	Pearson Correlation	1	-.577**	-.726**	-.884**	-.608**	.597**	-.266
	Sig. (2-tailed)	.	.000	.000	.000	.000	.000	.122
	N	35	35	35	35	35	35	35
V_engine	Pearson Correlation	-.577**	1	.747**	.668**	.821**	-.372*	.168
	Sig. (2-tailed)	.000	.	.000	.000	.000	.028	.335
	N	35	35	35	35	35	35	35
hp	Pearson Correlation	-.726**	.747**	1	.795**	.683**	-.434**	.333
	Sig. (2-tailed)	.000	.000	.	.000	.000	.009	.051
	N	35	35	35	35	35	35	35
weight	Pearson Correlation	-.884**	.668**	.795**	1	.690**	-.452**	.284
	Sig. (2-tailed)	.000	.000	.000	.	.000	.006	.098
	N	35	35	35	35	35	35	35
numcyl	Pearson Correlation	-.608**	.821**	.683**	.690**	1	-.348*	.254
	Sig. (2-tailed)	.000	.000	.000	.000	.	.041	.141
	N	35	35	35	35	35	35	35
japan	Pearson Correlation	.597**	-.372*	-.434**	-.452**	-.348*	1	-.193
	Sig. (2-tailed)	.000	.028	.009	.006	.041	.	.266
	N	35	35	35	35	35	35	35
age	Pearson Correlation	-.266	.168	.333	.284	.254	-.193	1
	Sig. (2-tailed)	.122	.335	.051	.098	.141	.266	.
	N	35	35	35	35	35	35	35

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

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

In the first row we see a significant positive correlation of the dependent variable, **mpg**, with **japan**. There are significant negative correlations with **V_engine, hp, weight**, and **numcyl.**
From the remainder of the display we see that there are significant correlations between most pairs of

independent variables. We have already seen examples of the way in which this phenomenon, sometimes called "**multicollinearity**" can obscure important relationships among variables. We will suggest a new way of dealing with it in a moment.

We now perform **Stepwise Regression** with **mpg** as the dependent variable and all of the others (except **auto**) as candidates for inclusion on the right-hand-side of the equation:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	weight	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	japan	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: mpg

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.884[a]	.781	.774	2.9596
2	.911[b]	.830	.820	2.6465

a. Predictors: (Constant), weight

b. Predictors: (Constant), weight, japan

c. Dependent Variable: mpg

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	46.968	2.284		20.568	.000
	weight	-.009	.001	-.884	-10.847	.000
2	(Constant)	42.927	2.435		17.627	.000
	weight	-.008	.001	-.771	-9.443	.000
	japan	3.380	1.110	.249	3.045	.005

a. Dependent Variable: mpg

Excluded Variables[c]

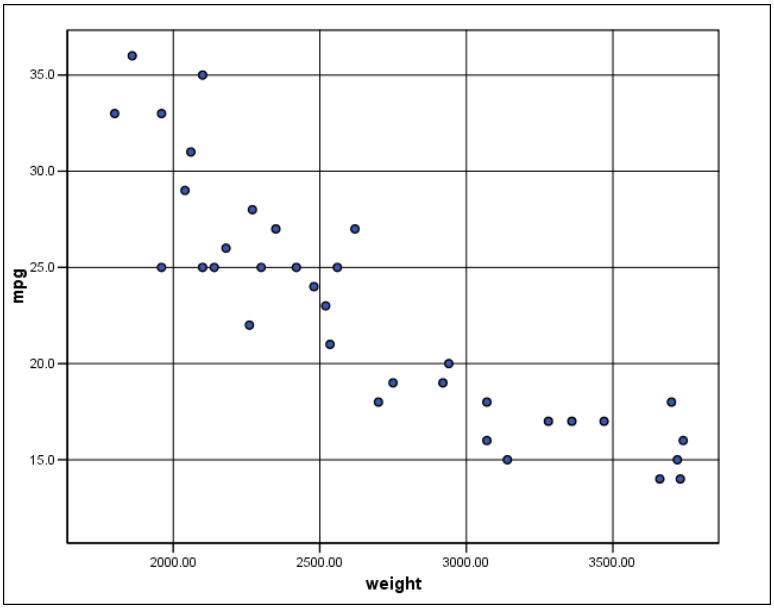
Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	V_engine	.024[a]	.212	.833	.037	.554
	hp	-.063[a]	-.460	.648	-.081	.368
	numcyl	.003[a]	.028	.978	.005	.524
	japan	.249[a]	3.045	.005	.474	.796
	age	-.016[a]	-.190	.851	-.034	.919
2	V_engine	.056[b]	.560	.580	.100	.548
	hp	-.013[b]	-.104	.918	-.019	.361
	numcyl	.020[b]	.199	.844	.036	.522
	age	.001[b]	.015	.988	.003	.914

a. Predictors in the Model: (Constant), weight

b. Predictors in the Model: (Constant), weight, japan

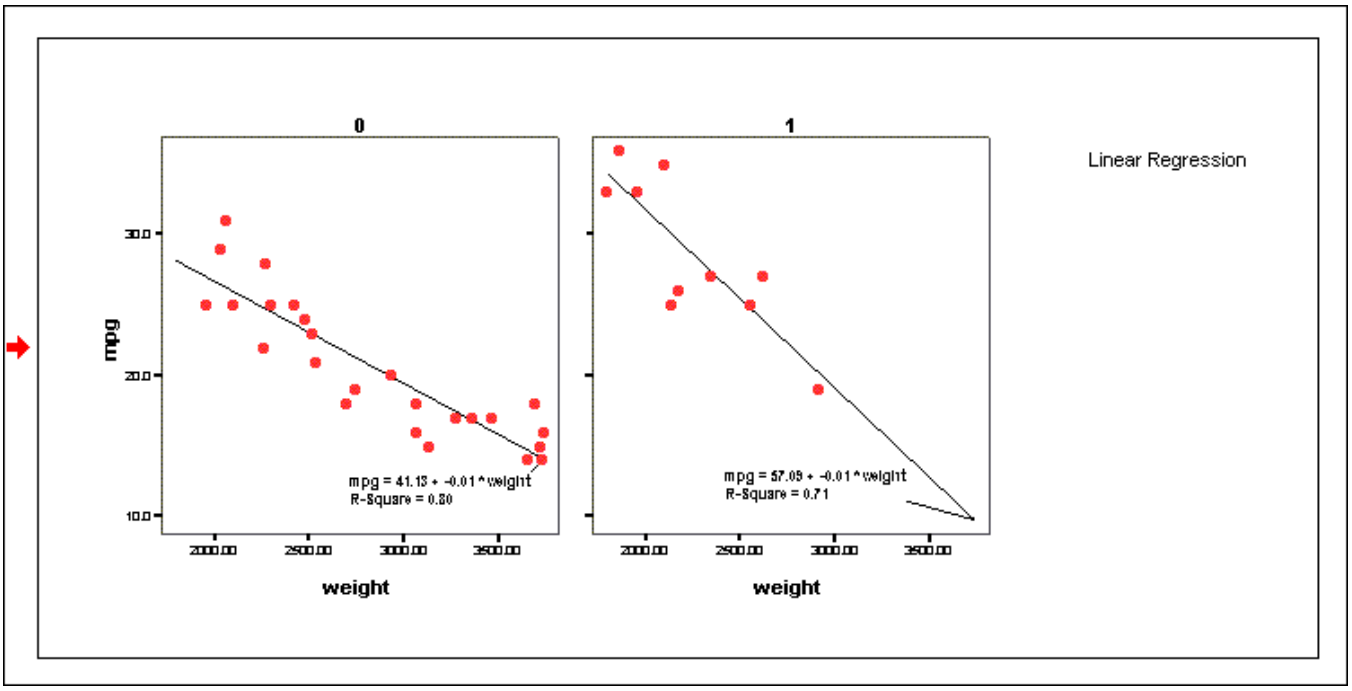
c. Dependent Variable: mpg

Only two of the candidate variables, **weight** and **japan**, contribute significantly to the multiple regression. Let's look at the plot of **mpg** against **weight**:

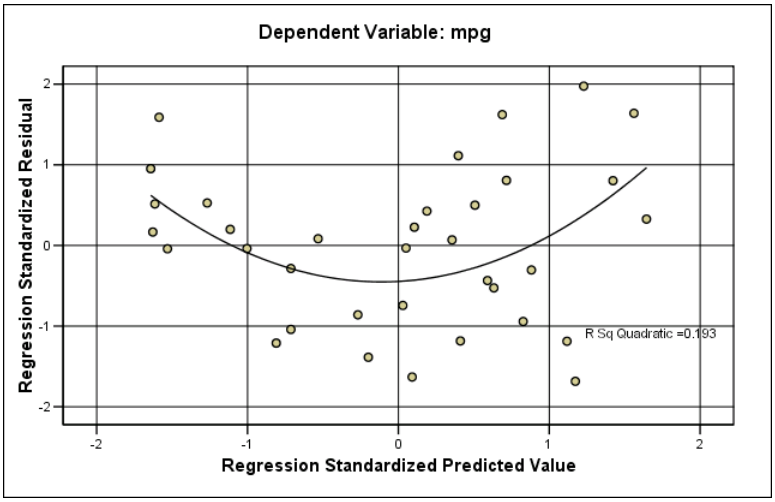


The relationship appears to be nonlinear!

We repeat the approach involving **Graphs/Intervactive/Scatterplot…**that we used earlier in this chapter to show which of the points above are Japanese automobiles and which are not. In the setup we make **japan** a panel variable which results in this dual plot:



The separate regression lines for **japan** equal to **0** or **1** are quite different. Note that the hint of curvilinearity remains in the scatter for non-Japanese cars.



The plot of **standardized residuals** vs. **standardized predicted** looks very bad indeed! We have to do something about the nonlinearity. With **Transform/Compute…** we create:

weightsq = weight*weight

Then we run the multiple regression again, this time including **weightsq**.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.934[a]	.873	.860	2.3271

a. Predictors: (Constant), weightsq, japan, weight

b. Dependent Variable: mpg

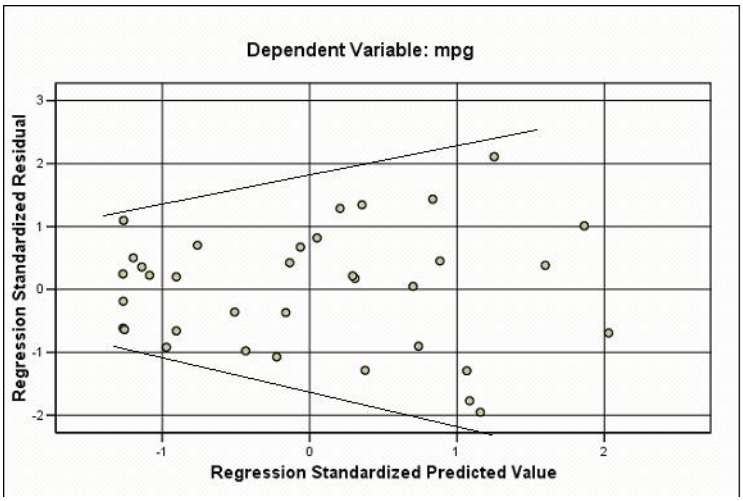
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	74.242	9.950		7.461	.000
	japan	2.988	.984	.220	3.038	.005
	weight	-.031	.007	-3.042	-4.295	.000
	weightsq	.000	.000	2.267	3.223	.003

a. Dependent Variable: mpg

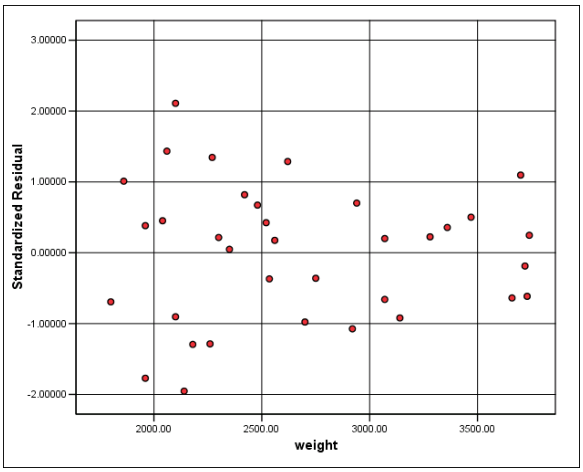
All three of the independent variables, **weight, weightsq**, and **japan** have regression coefficients that are significantly different from zero.

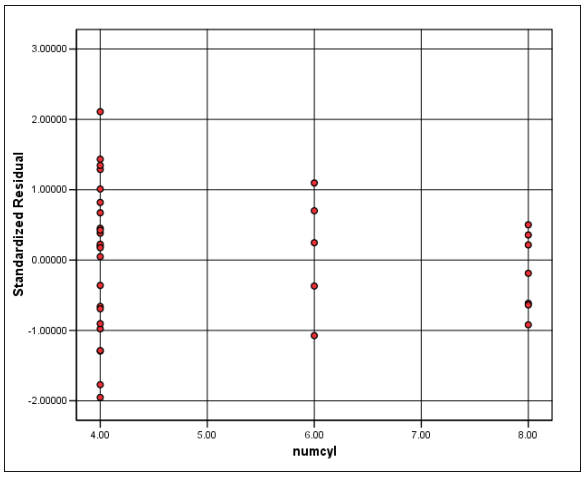
Note also that the **standard deviation** is 2.33, considerably smaller than 6.23, the **std. dev.** of **mpg** before the regression. This tells us that if we want to predict the **mpg** of another automobile with a weight that is not too far from those in the sample we are likely to be within plus or minus 5 miles per gallon of the correct figure.



We still have a problem, however. The plot above shows that the variance of the residuals is not constant. **We must keep in mind the possibility of a logarithmic transformation**. The histogram and normal probability plot for the **residuals** are satisfactory, however, and we will put off transforming for the time being.

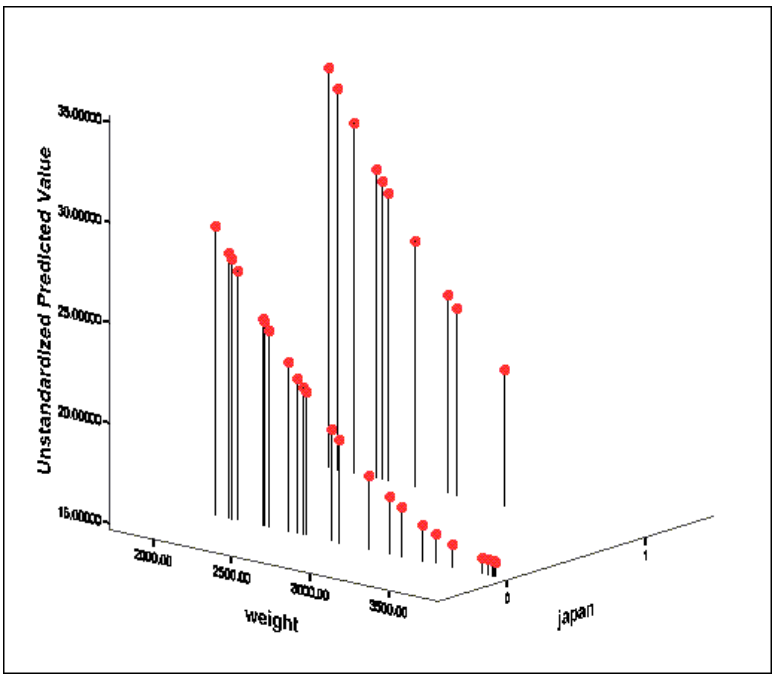
Before reporting that the residuals look pretty good (with the exception of nonconstant variance) it is always a wise procedure to check for possible further nonlinear relationships with the variables that are already in the model, as well as relationships (both linear and nonlinear) with those that have been left out. We will only show two examples here, but the **residuals** should be plotted against all of the potential independent variables.





There is no evidence of higher order nonlinearity with respect to **weight**, nor does there appear to be any interesting relationship with **numcyl**. Both plots, however, confirm the previous indication of nonconstant variance.

Finally, here is an interactive 3-D plot of **predicted** vs. **weight**, controlling for **japan**:



We have added spikes running vertically from the points to the floor of the plot to emphasize that the predicted values of **mpg** for the Japanese autos are uniformly higher than those for the non-Japanese vehicles. The Japan effect is quite dramatic. (We leave it to readers to speculate over whether we would see the same phenomenon with 2004 vehicles.) Also we see that the pattern of points for the non-Japanese cars exhibits a greater degree of curvilinearity.

Let's return to the question of correlation among the potential independent variables. Statistically, our main objective is to find a simple or "parsimonious" model (few independent variables) that fits the data well. We saw before we began the regression modeling that some pairs of potential independent variables had high correlation coefficients. We did not create **weightsq** until after that check on correlation, but now we can see that **weightsq** and **weight** are also very highly correlated.

Correlations

		weightsq	weight
weightsq	Pearson Correlation	1	.996**
	Sig. (2-tailed)	.	.000
	N	35	35
weight	Pearson Correlation	.996**	1
	Sig. (2-tailed)	.000	.
	N	35	35

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

When such a high degree of multicollinearity is present it often happens that the estimated coefficients for both variables (in this case **weight** and **weightsq**) are rendered insignificant, a strong argument for going to a more parsimonious model. In this instance, however, as shown by the regression analysis above, retaining both variables yields a better linear fit. The problem, if any, comes in trying to

interpret the two coefficients separately. Because of the correlation, the coefficient of **weight** changes drastically according to whether or not **weightsq** is included; and vice versa.

We saw above that there was some evidence of nonconstant variance of residuals. Now we explore that issue further, both to make a small refinement in the model and to follow up on the discussion of transformations in Section 3. With **Transform/Compute…** we create

logmpg = LN(mpg)

and repeat the regression analysis using this new dependent variable.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.939[a]	.882	.871	.09809

a. Predictors: (Constant), weightsq, japan, weight

b. Dependent Variable: logmpg

Coefficients[a]

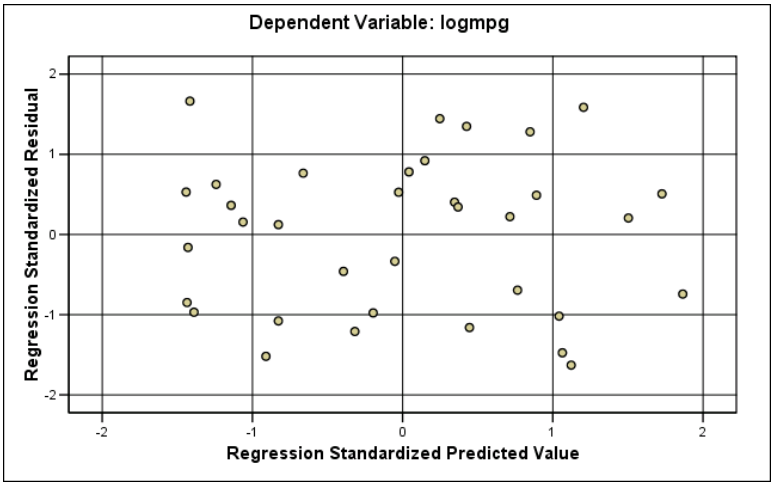
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.9028188	.41942343		11.689	.000
	japan	.11277380	.04146687	.189	2.720	.011
	weight	-.00100759	.00030476	-2.251	-3.306	.002
	weightsq	.00000011	.00000005	1.431	2.116	.043

a. Dependent Variable: logmpg

Note that in order to display significant digits for the coefficients and std. errors for **weightsq** we had to increase the number of decimal places for those columns from three to eight. The regression equation is thus

predicted logmpg = 4.9028 + 0.11277380 japan – 0.00100759 weight + 0.00000011 weightsq .

Let's see if the residuals are improved:



The answer is "Yes", the variance now appears to be constant.

Looking at the regression output above, you might wonder if the variable **weightsq** is really needed. Although its regression coefficient is statistically significant, it is of such small magnitude that perhaps it is not of practical importance. Let's run the regression without it:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.930[a]	.865	.857	.10328

a. Predictors: (Constant), weight, japan

b. Dependent Variable: logmpg

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	4.03627051	.09504242		42.468	.000
	japan	.12361665	.04332614	.207	2.853	.008
	weight	-.00036616	.00003254	-.818	-11.253	.000

a. Dependent Variable: logmpg

With a few numerical examples you can verify through the two regression equations that the effect of an increase in automobile weight of 1,000 pounds has about the same effect on **logmpg** under both models.[6] If, however, we were to plot the **residuals** vs. **weightsq** for this last model with **weightsq** omitted, we would still see the curvilinearity that led us to include it in the first place. **Hence it is best to work with the less parsimonious model.**

Two more comments are in order with respect to comparing the **logmpg** model with that using **mpg**:

[6]Are you really up to verifying this statement? Try it.

- The **logmpg** model is no longer linear in the **mpg** scale. It says that

$$predicted\ mpg = e^{4.90 - 0.001 weight + 0.113 japan - 0.0000001134 weightsq}$$

Recall that e^x is the exponential function. The mathematical form of the equation doesn't matter if our aim is primarily to predict, but if we are trying to understand the causal mechanism behind fuel efficiency it may be important.

- The model in **logmpg** has an **Adjusted R Square** equal to 0.871, slightly higher than that for the **mpg** model, 0.860. Does that mean that the log model gives a closer fit than the original model? Not necessarily, because we must remember that the scales of measurement for the two models are different. To answer the question, it is necessary to transform the fitted values for one model back into the scale of the other model, and compare the residual sums of squares in the common metric. **In the present case, regardless of which model is slightly superior in fit, the better behavior of the residuals recommends the log model.**

Time-Series Checks Are Not Applicable

Finally, a word on the absence of time-series checks in our analyses of cross-sectional data. The time-series checks, by definition, are based on the time-ordering of data. With time-series data, we can check easily for trends, autocorrelation, and other patterns that cause deviation from the assumption of statistical control for the variable or variables of interest.

Technically, the assumption of statistical control can be expressed by saying that it is assumed that the data -- or residuals from regression models -- arise **randomly**.[7]

There is no easy way to check on this assumption in cross-sectional data, where the sequence of listing is often arbitrary, as in the present application.[8] Hence in most cross-sectional applications, you can forget about the time-series checks, even though, by now, you are in the habit of doing them routinely.

To get at the randomness assumption in cross-sectional studies, we have to take a different tack, which we now illustrate. Cars 63 and 64 in the data set just analyzed were both Chrysler "K" cars. Car 64 had a slightly larger engine and slightly poorer mileage. Hence we do not really have two independent pieces of information. Fortunately, that is the only such problem in the data set. However, if we were dealing with a data set with just a few distinctly different car models and many slight variations on the models included, the effective sample size would be much smaller than it appears to be on the surface.

[7]Still more technically, the expression is **"independent and identically distributed"**.

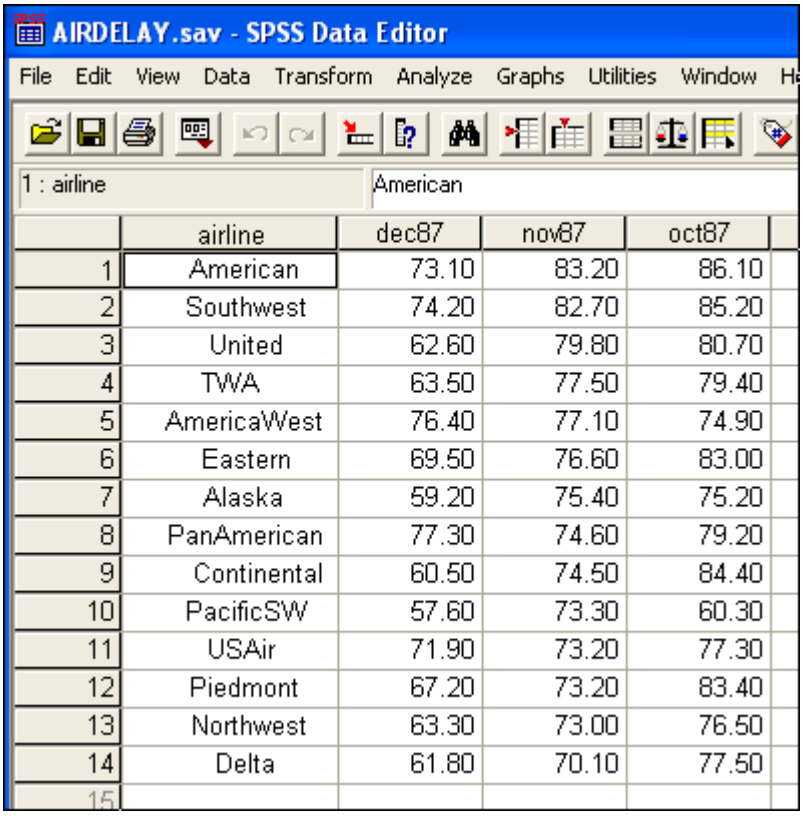
[8]In some cross-sectional applications, the sequence of listing of data may not be arbitrary. For example, in a study of accidents for employees in a plant, the data file might list employees from low to high seniority. Then a "time-series check" becomes a "seniority check", and it would be worth doing.

Similarly, in the sales examples of Section 3, there is the possibility that efforts of a sales representative in one territory may spill over into another. This is the cross-sectional analogue of autocorrelation in time series. Unfortunately, it is usually very hard to detect such problems in cross-sectional data. The simple statistical tools of time-series analysis do not do the job. The best we can do is to try to get to know our cross-sectional data better: to note, for example, when two cars are slight variations of each other or two adjacent territories might have spillover sales effects.

6. Successive Cross Sections and Performance Assessment

When we have cross-sectional performance measures for two successive cross sections (or in general, for more than two), we stand to learn much more than can be learned from a single cross section. At the end of Section 2, we studied Air Delays in October, 1987, for 14 airlines. One airline, Pacific Southwest, had a much poorer on-time record than the others. In a histogram of percent on-time for all airlines, Pacific Southwest stood out nearly three standard deviations below the mean, suggesting the possibility of a special cause, and the desirability of further investigation of its precise nature.

But how about the other 13 airlines? Are there real differences between them, or are we just seeing chance variations? There is little that we can do to answer this question when we have data from only a single cross section. But if we have data on successive cross sections, we can make important progress. To illustrate, we shall use data on the same 14 airlines for the next two months, November and December of 1987, as shown below in the contents of AIRDELAY.sav:


AIRDELAY.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window H

1 : airline | American

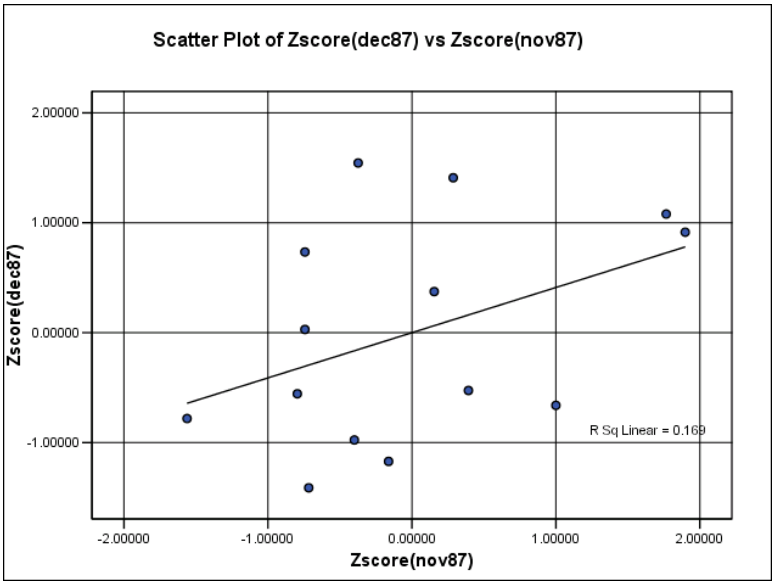
	airline	dec87	nov87	oct87
1	American	73.10	83.20	86.10
2	Southwest	74.20	82.70	85.20
3	United	62.60	79.80	80.70
4	TWA	63.50	77.50	79.40
5	AmericaWest	76.40	77.10	74.90
6	Eastern	69.50	76.60	83.00
7	Alaska	59.20	75.40	75.20
8	PanAmerican	77.30	74.60	79.20
9	Continental	60.50	74.50	84.40
10	PacificSW	57.60	73.30	60.30
11	USAir	71.90	73.20	77.30
12	Piedmont	67.20	73.20	83.40
13	Northwest	63.30	73.00	76.50
14	Delta	61.80	70.10	77.50
15				

Recalling that the variable under study is the percentage of flights that were on time, we apply **Descriptive Statistics** to **dec87** and **nov87**. Be sure to check the little box for **Save standardized values as variables:**

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
dec87	14	57.60	77.30	67.0071	6.66558
nov87	14	70.10	83.20	76.0143	3.78598
Valid N (listwise)	14				

Note that the mean delay is lower in December, but the standard deviation is greater. Can that be due to poorer weather in December? For our present purposes, however, we are not interested in the average on-time performance of all 14 airlines, but in how the airlines perform relative to each other. We pursue that question now, looking at the **standardized** versions of **dec87** and **nov87**:



Correlations

		dec87	nov87
dec87	Pearson Correlation	1	.411
	Sig. (2-tailed)	.	.144
	N	14	14
nov87	Pearson Correlation	.411	1
	Sig. (2-tailed)	.144	.
	N	14	14

(For the purpose of this application, we shall treat 0.411 as if it were deemed statistically significant, but its significance is in fact borderline: the standard error is $1/\sqrt{14} = 0.267$.)

The correlation 0.411 between the successive cross sections is the key to our analysis of, say, the December on-time percentages. To see intuitively why this is so, consider two extremes:

- If the correlation were zero, it would appear that the performance measures for both months are the result of independent chance factors that have nothing to do with the "long-run" on-time capability. Then the best estimate of individual airline performance in December would put all 14 airlines at the December mean of 67.007 percent.

- If the correlation were +1, it would appear that each month's performance measures reflect nothing but consistent "long-run" performance capability. The December readings would then be accepted at face value: all differences between airlines would be taken as "real".

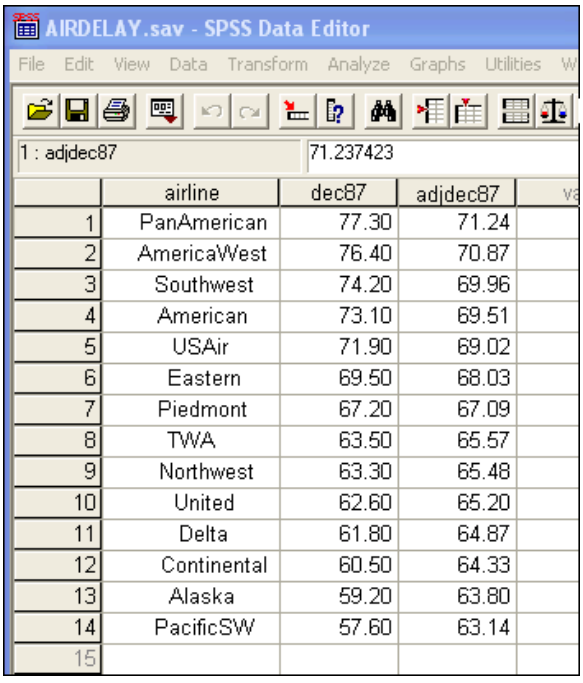
Of course, neither extreme assumption is correct. However, under reasonable statistical assumptions we can adjust the observed measures by the following procedure for each airlines:

- Calculate the airline's deviation from the December mean of 67.007.
- Multiply this deviation by the correlation coefficient, 0.411.
- Add the result to the overall mean of 67.007

Now we are ready to apply **Transform/Compute…** and make the desired adjustment to the data:

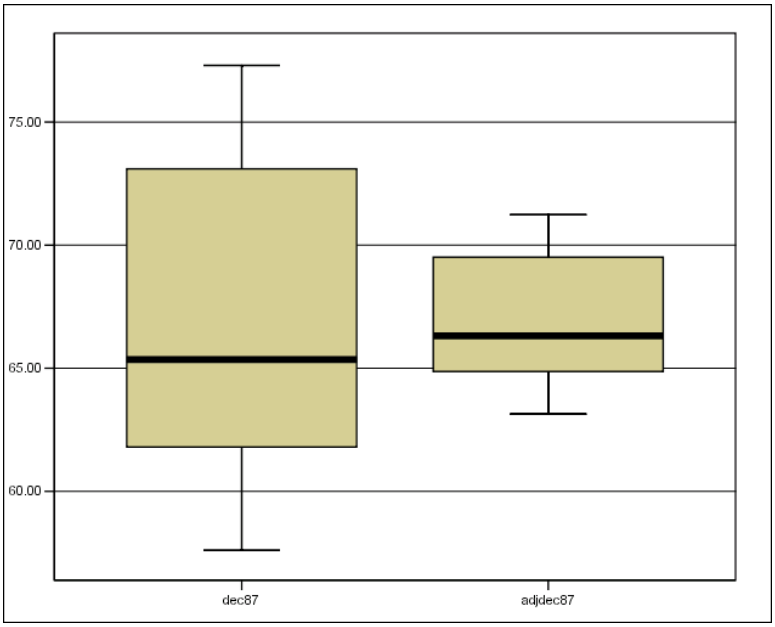
adjdec87 = 67.007 + 0.411*(dec87 - 67.007)

After clicking on **OK** to carry out the adjustment, we then use **Data/Sort Cases…** to rank the data from highest to lowest on the **dec87** performance measure, making sure that we carry **adjdec87** along in the sort. With some judicious reordering of variables in the spreadsheet, we can display the following:



	airline	dec87	adjdec87
1	PanAmerican	77.30	71.24
2	AmericaWest	76.40	70.87
3	Southwest	74.20	69.96
4	American	73.10	69.51
5	USAir	71.90	69.02
6	Eastern	69.50	68.03
7	Piedmont	67.20	67.09
8	TWA	63.50	65.57
9	Northwest	63.30	65.48
10	United	62.60	65.20
11	Delta	61.80	64.87
12	Continental	60.50	64.33
13	Alaska	59.20	63.80
14	PacificSW	57.60	63.14
15			

For example, Pan American (listed #1 on the sorted array above) had 77.3 percent on time, the best of the 14 airlines in December. The above computation "shrinks" that number toward the mean, yielding 71.2 percent as the adjusted performance. Pan American is still rated "best" but the margin of difference is much less. Similarly, PSA (listed #14 on the sorted array) had 57.6 percent on time, the worst of the 14 airlines in December. This is shrunk upward towards the mean, to 63.1 percent. We can visualize the results of this shrinkage by the following **Box Plot** and **Descriptive Statistics:**



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
dec87	14	57.60	77.30	67.0071	6.66558
adjdec87	14	63.14	71.24	67.0071	2.73955
Valid N (listwise)	14				

There is an important general lesson here. The example is not atypical. Observed performance measures reflect both "long-run" performance capability and chance, in varying mixtures. The natural tendency is to assume away the role of chance, to regard each difference in performance as a reflection of a special cause that distinguishes one individual from another. The airline example is relatively typical in that there is moderate (albeit at best only borderline-significant) correlation between one month and another. When data on two time periods are available, we can use the simple "shrinkage" approach above to make allowance for the role of chance factors.

Sports teams are an interesting example when we compare standings, based on winning percentages, from one season to another. In professional baseball the correlations from season to season have been so low in recent years that shrinkage brings all teams close to 0.500. In professional football the correlation from season to season is much higher -- on the order of 0.600 -- but shrinkage shows that the disparities between teams are still substantial. Luck plays an important role!