A RAND NOTE

USING STATISTICAL TOOLS

John E. Rolph

April 1984

N-2121-PSSP

Prepared for

The Private Sector Sponsors Program

Rand
1700 MAIN STREET
P.O. BOX 2138
SANTA MONICA, CA 90406-2138

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
APR 1984		**00-00-1984 to 00-00-1984**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Using Statistical Tools	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Rand Corporation,1776 Main Street,PO Box 2138,Santa Monica,CA,90407-2138**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**83**	

The research described in this report was sponsored by the Private Sector Sponsors Program.

Published by The Rand Corporation

A RAND NOTE

USING STATISTICAL TOOLS

John E. Rolph

April 1984

N-2121-PSSP

Prepared for

The Private Sector Sponsors Program

Rand
1700 MAIN STREET
P.O. BOX 2138
SANTA MONICA, CA 90406-2138

PREFACE

This Note consists of the text and charts of a presentation given to the first Private Sector Sponsors Program (PSSP) Process Industries Workshop, held at The Rand Corporation on September 28 through September 30, 1983. This presentation--on using elementary statistical tools--was given in the opening session of the workshop, which was attended by managers and working-level staff from PSSP sponsoring organizations. The purpose of the session was to give the audience enough background in statistical methods in general, and in regression methods in particular, to allow them to follow the statistical presentations on cost estimation, performance estimation, and schedule slippage estimation given later in the workshop. The statistical background of the audience was assumed to be rudimentary.

This presentation is reproduced here as a Rand Note, both to give the attendees a written record of this portion of the workshop and to provide a reference for similar PSSP workshops in the future.

The material presented here is not original. Examples were chosen from a variety of statistics texts and other sources that were relevant to the interests of the workshop participants. Sources are given below in the Acknowledgments section and in the references.

SUMMARY

The briefing presented here has two goals: to review the logic behind statistical reasoning, with particular emphasis on regression techniques; and to discuss, through a series of examples, how to interpret output from regression programs.

The material is organized into three sections. The first discusses model building. The second describes how regression methods can be used to fit equations to data. The third describes how to make statistical inferences about the parameters that describe a regression line or surface, and discusses potential pitfalls in making such inferences.

ACKNOWLEDGMENTS

I appreciate the support and encouragement of Edward Merrow and Robert Perry and the suggestions of Ronald Hess and Chris Myers. I would like especially to acknowledge the major role that Mary Vaiana played. She took a transcript of the oral presentation and transformed it into a readable draft that I could work from to produce the current note. Gus Haggstrom reviewed an earlier version of the Note and made many valuable suggestions. Will Harriss ably edited the manuscript.

The data examples were drawn from several sources that I will describe here by reference to page numbers. Pages 6, 7, and 8 are based on data in Draper and Smith (1981, page 414). Pages 11, 12, 15, and 19 are based on data in Freedman, Pisani and Purves (1978, Chapters 8 through 11). The Forbes data on pages 22 through 29 were taken from Weisberg (1980, Chapter 1). The wages regression example on pages 35 through 38 was taken from Morris and Rolph (1981, Chapter 4). The tree data on pages 55 through 62 were taken from Ryan, Joiner and Ryan (1976, Appendix A). Finally, the hypothetical graduate school admission data on pages 68 and 69 were taken from Hooke (1982, Chapter 13).

CONTENTS

I. MODEL BUILDING

INTRODUCTION

This Note first reviews elements of statistical reasoning, with particular emphasis on regression methods, as those will be the basis for the parametric cost estimation discussed later in this workshop. In interpreting regression output, I will discuss estimated regression coefficients and how to interpret them; define measures of the estimates' accuracy; and discuss how to construct confidence intervals and make statistical tests, both about regression coefficients and other aspects of the regression equations. Finally, I will examine measures of the strength of the predicted relationships.

The discussion will be divided into three sections. First, the discussion of model building covers problem definition, measuring outcomes, data collection issues, data quality issues, and so forth. Next, is a description of using regression methods to fit equations to data, both equations with a single explanatory variable and those with several explanatory variables. Following is how to interpret the output from these fits. Finally, the issue of making statistical inferences from these fits is addressed, by introducing the idea of a model for the error around the regression line and making statistical inferences about the parameters that describe the regression line or surface. The Note concludes with a discussion of potential pitfalls.

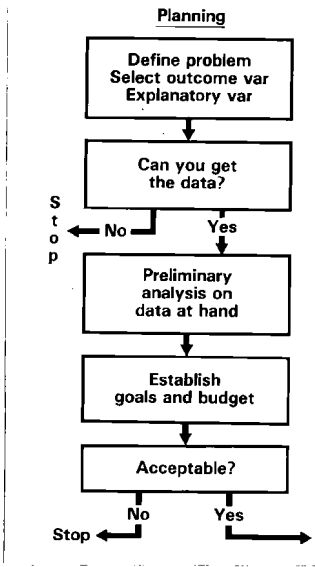
MODELS

- **Types**
  - **Functional — e.g., chemical process**
  - **Control — e.g., designed experiment**
  - **Predictive — e.g., cost estimation**
- **Characteristics**
  - **Deterministic**
  - **Probabilistic**

There are several ways to look at models. The first is by type. For example, in a *functional model*, the phenomenon being modeled is fairly well understood and unambiguous. An example would be a chemical process that is determined by mass balance equations. Functional models may be good for explaining how a process works; but if we want to control the process we need a *control model,* which not only explains the phenomenon under consideration but has some variables in it that can be changed. This is typically the situation in a designed experiment, where we can control some important variables of interest. *Predictive models* are like black boxes. They are often used with messy data; they are frequently not functional models, in the sense that we don't fully understand the phenomenon we are modeling. However, the data can generate enough information about the phenomenon to render these models usable for making predictions. That is the process we will be discussing when we consider cost estimation.

Another way of categorizing models is by characteristics--for example, *deterministic* vs. *probabilistic* models. "Deterministic" doesn't mean that we understand the world perfectly, but rather that we treat the model as a reasonable approximation to reality, without worrying about how far off it is. A good example is an engineering cost estimation model. By contrast, in probabilistic or statistical models, we will pay some attention to the errors in the model--that is, we will want to know how uncertainty about the model will affect the accuracy of the predictions that we will make with the model.

SUMMARY OF MODEL BUILDING: THE PLANNING STAGE

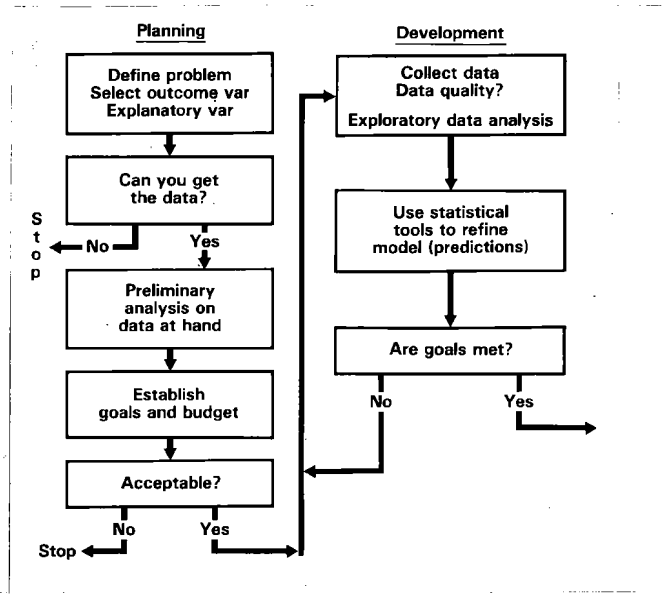


This diagram provides an overview of the first step in the model building process. It is important to note that statistics is only one part of the process. Practical knowledge about the process being modeled is equally important.

The first part of the model building process is really the planning or brainstorming stage. Here we define the problem. We decide what we are trying to measure--that is, the outcome variables we are interested in--and possibly the kinds of variables that are available to explain this outcome. Once we've settled on tentative definitions of the outcome variables and of possible explanatory variables, we must ask whether we can gather information on these variables. If it is impossible to get the required data, the model building process stops.

If we believe we will be able to gather data, we will want to do some planning. This planning can range from educated guesses to preliminary analysis of data that are related to the problem. On the basis of this planning, we can establish some goals for what we hope to be able to predict, how well we expect to do it, and roughly how much it will cost. Of course, if these goals and budget are not acceptable, we are out of business. If they are acceptable, we move on to the process of model development itself.

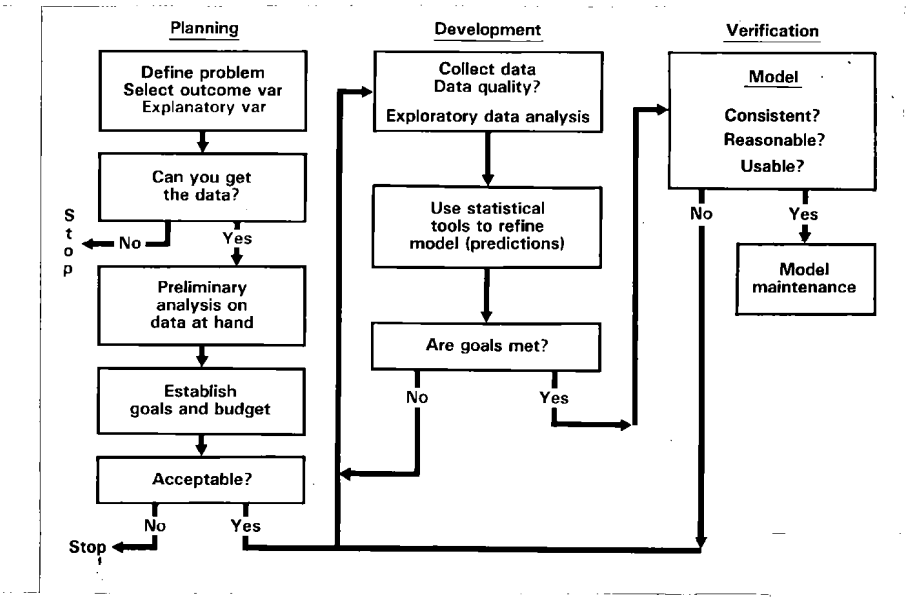
SUMMARY OF MODEL BUILDING: MODEL DEVELOPMENT



Model development starts with data collection. Oftentimes questions about data collection revolve about quality issues: Are the data sufficiently good that we can believe the answers based on them?

Once the data are collected and checked, we begin a two-phase analysis. First is exploratory data analysis. This consists of trying to get a feel for the data--making plots, perhaps doing cross-tabulations, trying to understand what the relationships are between the variables we are measuring. Once we have a general grasp of these relationships, we begin more formal statistical modeling. This process includes refining outcome and explanatory variables, using various diagnostics, and so forth. It is an iterative process that typically takes many steps. At the end of this process, we must ask whether the model we have developed will meet our goals. If it will not, we go back to square one.

SUMMARY OF MODEL BUILDING: VERIFICATION

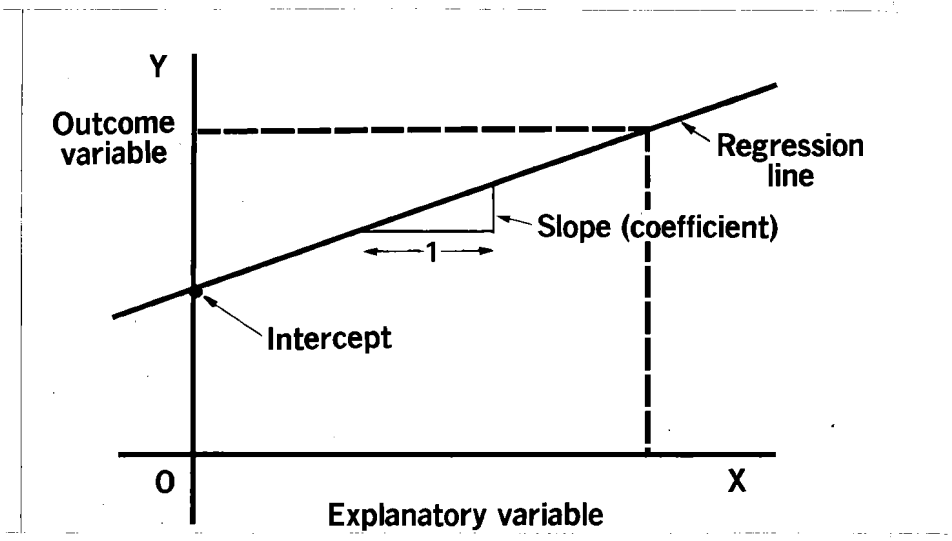


If the output of the model meets our goals, we enter what I have labeled the model verification process. This process is by and large nonstatistical, and it is extremely important. To verify a model, we ask three questions: (1) Is the model consistent? That is, are the estimated relationships stable over time? For example, does innovation appear to have the same effect on cost growth for early plants in our sample as it does for later plants? Is there any obvious, systematic lack of fit of the model to the data? Both commonsense tests and statistical tests are needed. (2) Is the model reasonable? Do experts agree that the important explanatory variables have been included? Are we using the model in the relevant range of these variables? Does the model agree with common knowledge about the process under investigation? (3) Is the model usable? That is, can we use it to explain or predict the outcome variables that we are interested in?

Finally, if we are going to be using the model more than once, it must be maintained. We must ensure that if the process we are modeling changes, we can update our model to reflect those changes.

II. USING REGRESSION METHODS TO FIT EQUATIONS TO DATA

SIMPLE LINEAR REGRESSION

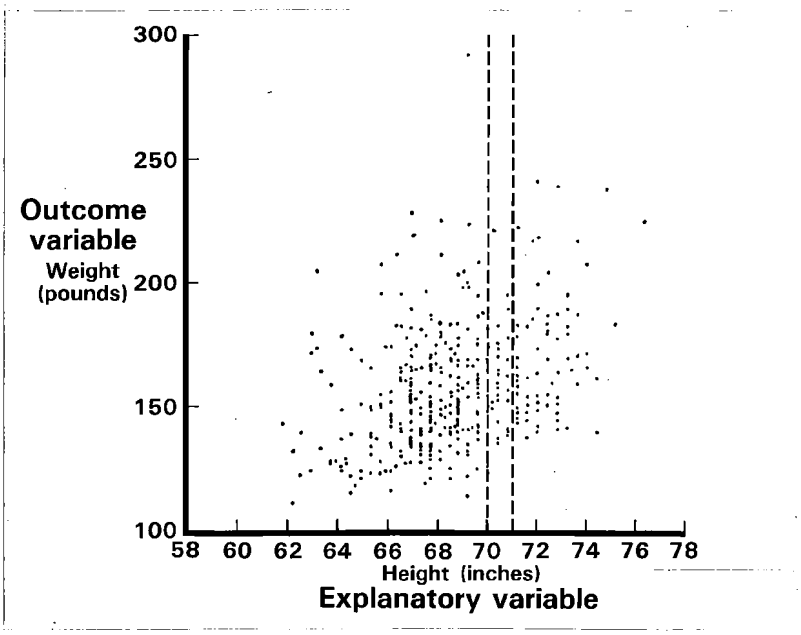


In this era of home computers and hand calculators, most of you have had the capability of doing regressions for quite some time. We will not discuss what goes on inside the calculator or the computer; instead, we will address issues of application. When is it appropriate to run a regression? How do we understand, interpret and use the output?

The simplest possible situation is simple linear regression. In it we have a single explanatory variable, labeled "X" on the horizontal axis, in the diagram above and an outcome variable, labeled "Y", on the vertical axis. By a regression line we mean is simply a line or a graph that tells us, for a given value of X, what the predicted value is for Y.

Regression lines have two characteristics that define them. First, the *intercept* is where the line crosses the axis at 0. The intercept gives the value of Y when X is equal to 0. Second, the *slope* of the line is the amount that the predicted value of Y increases for every one-unit increase in X.

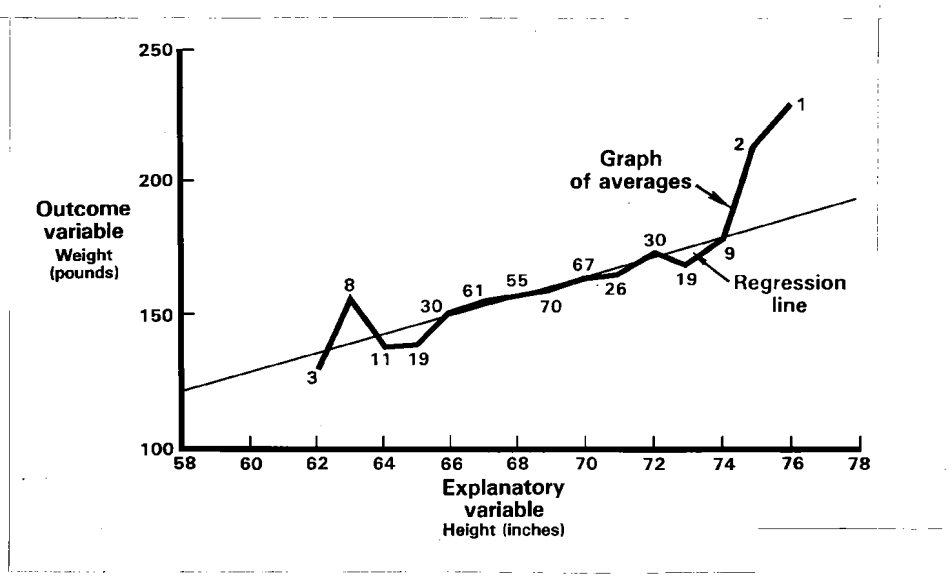
SCATTER PLOT



Let me work slowly through the process of using data to develop a regression line. We will use the following example. Think of predicting a person's weight from his height. Imagine the explanatory variable as being the person's height in inches and the outcome variable as being the person's weight in pounds. Depicted here is a "scatter plot." In this case, it is a plot of the heights and weights of 411 men aged 18 to 24. Each point corresponds to one man. For example, the uppermost point corresponds to a man who is about 5 feet 9 inches tall and weighs about 280 pounds.

One way to predict weight from height is to divide the graph into strips, and for some interval of height (for example, one inch) take the average of all of the weights in that strip and use that as our prediction.

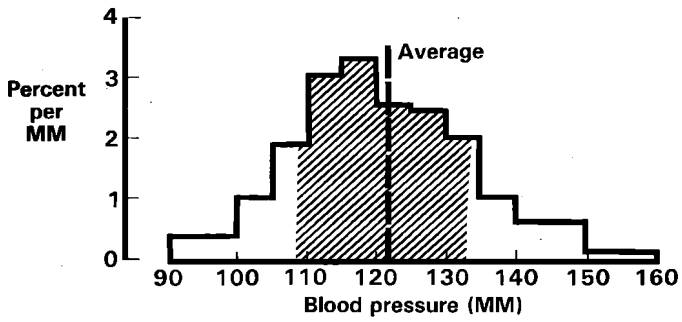
GRAPH OF AVERAGES AND REGRESSION LINE



If we to divided the entire graph in that way, we would get something called the *graph of averages*. A number is associated with each point on this graph, indicating the number of men whose heights fall in a given strip. For example, the number "30" means that there were 30 men between 71 and 72 inches tall, and their average weight was about 165 pounds.

A graph of averages is not a bad way to assess the relationship between height and weight. However, it has considerable variation in it. Of course, there is nothing wrong *a priori* with a graph being variable, but we might wonder whether the unevenness of the graph is due to there being only a few people associated with some points, particularly at the extremes. Perhaps the true relationship between height and weight is a little smoother. Regression is a way of smoothing this line out. More precisely, the regression line is that line which is closest to those 411 points in the sense of minimizing the sum of all the squared distances between the points and the line.

STANDARD DEVIATION



Average = 121 MM

Standard deviation σ = 12.5 MM

$$\sigma = \sqrt{\text{Average of (deviations from average)}^2}$$

68% of women are within $\pm\sigma$ of 121

In order to understand how regression lines are computed and how we should interpret them, we briefly discuss two concepts. The first concept is *standard deviation*. An example is the above graph of the diastolic blood pressures in millimeters of mercury of 1747 women aged 25 to 34. The blood pressures range from about 90 to 160; the average is 121.

In addition to knowing the average value, we would like to know how much the values are spread out. The measure statisticians commonly use for this is the standard deviation, usually denoted by the Greek letter σ. In this particular graph, the standard deviation is 12.5.

The standard deviation is a measure of the distance that the points are away from the average. The formula shows how to compute the standard deviation: Take the average value of all the points, then determine the deviations from the average and square them, take the average of that total, and then to get back into the original units, take the square root of it. In this particular picture--and this will usually be the case--about 68 percent of the blood pressures are within one standard deviation of the average; that is, 68 percent are between 118.5 and 133.5 millimeters of mercury.

CALCULATING A STANDARD DEVIATION

$$\sigma = \sqrt{\text{Average of (deviations from average)}^2}$$

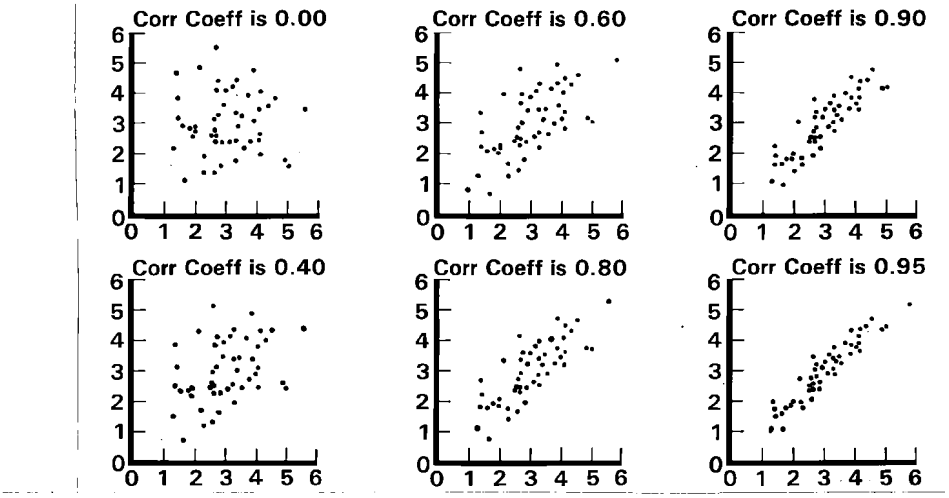
Example: 20, 15, 10, 15

$$\text{Average} = \frac{20 + 15 + 10 + 15}{4} = 15$$

$$\sigma = \sqrt{\frac{5^2 + 0^2 + (-5)^2 + 0^2}{4}} = \sqrt{\frac{25 + 25}{4}} = \sqrt{12.5} = 3.5$$

Let's calculate a quick example to fix the idea. What is the standard deviation of 20, 15, 10, and 15? First we take the average: 20 + 15 + 10 + 15 divided by 4 is 15. Then we add up the squared deviations from the average: 20 is 5 units away from 15--that's 5 squared; 15 is 0 squared; 10 is -5 units away--that's -5 squared; 15 is again 0 squared. The average is (25 + 25)/4 = 12.5. The square root of 12.5 is 3.5. The standard deviation of these four numbers is 3.5.

SCATTER PLOTS SHOWING CORRELATION COEFFICIENT



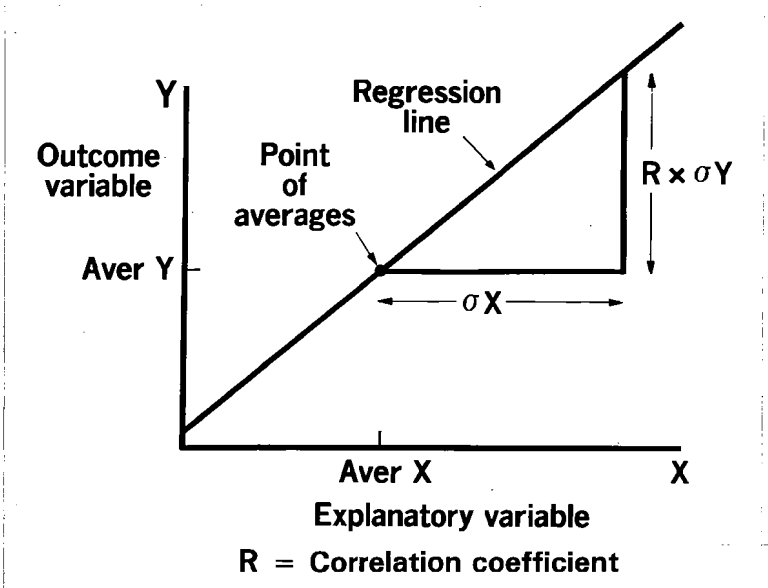
The second concept I want to discuss is *correlation*, or the *correlation coefficient*. This quantity is a measure of association between two quantities that remains the same no matter what units you measure the quantities in. For example, if in our earlier example we had measured the heights in meters and the weights in kilograms, we would like to be sure that the relationship between those two quantities is the same as it would be if we were measuring in inches and pounds. The correlation coefficient allows us to do this. It is a scale-free measure of (linear) association. It has a minimum value of -1 and a maximum value of +1. When it is 0, there is no association; when it is 1 or -1 it means there is a perfect association.

Ordinarily, we use computers to calculate correlation coefficients, but we can get a feel for correlation by examining these scatterplots. Each plot has the same scale, the same averages (3) for both variables, and the same standard deviations (1). The plots differ only in the value of their correlation coefficients. When the correlation is 0, the points on the plot seem to be more or less scattered at random. When

the correlation is largest--.95--the high and low values on the X axis correspond to the high and low values on the Y axis. A correlation of 1 would be a situation in which the data lie exactly on a straight 45-degree line. As the correlation decreases from .95, this pattern becomes less pronounced.

These graphs are all for positive correlations; simply turning the graphs on their sides yields analogous plots for negative correlations.

COMPUTING A REGRESSION



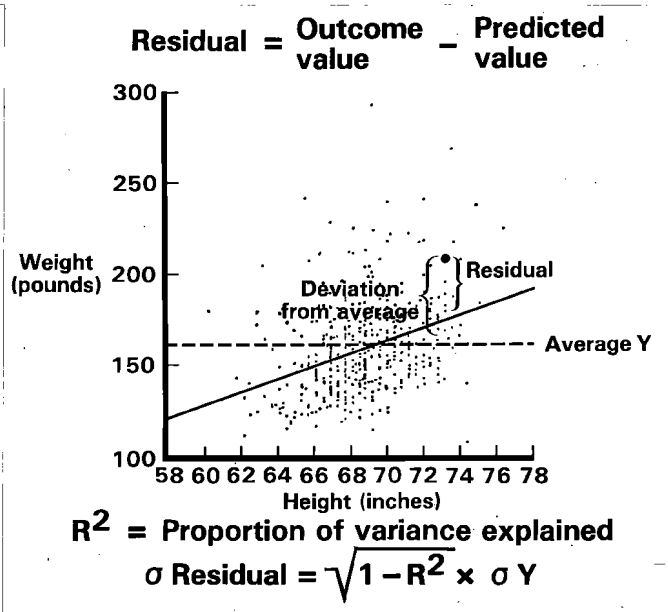
How do we use standard deviations and correlations to understand regression lines? Let's return to our scatter diagram, take the average of all the heights and the average of all the weights, and locate that point as the point of averages. The regression line always will go through that point. However, in order to draw our line, we need more than one point. We also need to know the slope of the line.

The slope of a regression line is very simply defined. If you increase the explanatory variable by one standard deviation (in this example, one standard deviation of the height), the regression line will increase by R times one standard deviation of weight, where R is the correlation coefficient.

As an example, let's look at two special cases. First, suppose that the correlation coefficient is 0. This is the case where there is no association. If the correlation coefficient is 0, then the height of the triangle would be zero, so the regression line would be horizontal. This means that changes in the value of the explanatory variable X will not affect the regression prediction of the outcome variable, Y.

Conversely, if R is equal to 1, it means that for a one standard deviation increase in the explanatory variable, there is a one standard deviation increase in the outcome variable. In this case, the points will lie all exactly on the regression line.

RESIDUALS



Another important concept in regression is how much the predicted value misses the actual value. The difference between the two is called the *residual*. To calculate it, you subtract the predicted value--that is, the value given by the regression line--from the actual value--in this case, each actual weight.

Consider the above scatter diagram of heights and weights. The marked point represents a person who is 73 inches tall and weighs about 220 pounds. His residual is the difference between his actual weight, 220 pounds, and his predicted weight, about 170 pounds.

The size of the residuals gives us information about the usefulness of our explanatory variable. If we didn't know heights at all and we had to predict someone's weight, we should use the simple average of all the weights. We can compare the residual with the difference between the actual value of Y and the average of the Y values. The difference between the point on the graph representing our 220-pound person and the corresponding point on the average line is the amount by which our estimate would miss. If the explanatory variable is providing useful

information, then we expect the residuals to be smaller than the deviations from the average value.

We can make this notion more precise by computing "R^2". It is the proportion of variability explained by the predictor variable. The formula for R^2 gives us the relationship between how large the residuals are (their standard deviation) and how large the deviations from the average are (the standard deviation of Y). If R^2 is 0, then none of the variability is explained, and the standard deviation of the residual is the same as the standard deviation of the outcome variable--that is, the regression line is simply the average line. On the other hand, if R^2 is 1, that means that all the variability is explained and the standard deviation of the residual is 0. If something has a standard deviation of 0, it must be constant; therefore, we would find that all of the points lie exactly on the regression line.

BUILDING A REGRESSION MODEL

1. **Choose variables**
2. **Fit line**
3. **Plot residuals to**
   - **A. Diagnostic—find outliers**
   - **B. Improve predictions**
4. **Iterate until satisfactory**

Now we will put together the concepts we have just discussed and use them for a simple linear regression. I will begin by talking about how we choose the explanatory variable. Then we will fit a regression line to the data using this explanatory variable. We will plot the residuals to find outliers as well as to improve our regression. Finally, we will repeat this process until we are satisfied with the results.

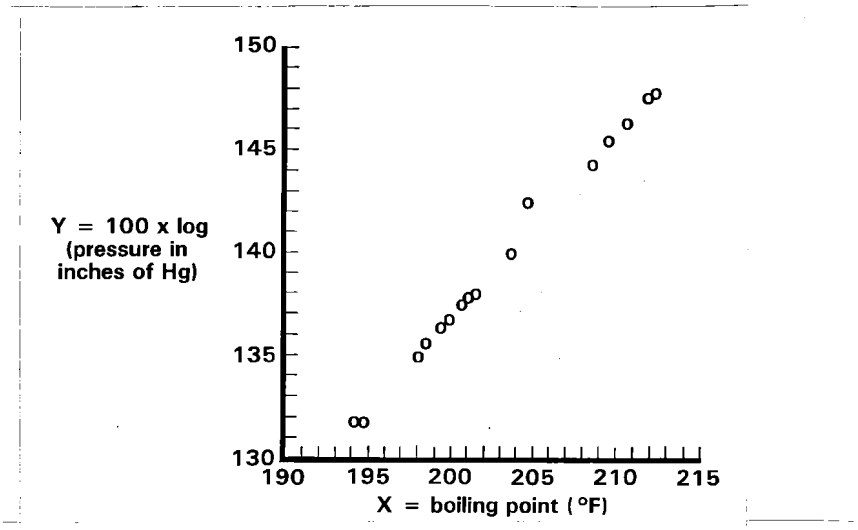
FORBES DATA EXAMPLE

Case Number	Boiling Point (°F)	Pressure (in. Hg)	Log(Pressure)	100 × Log(Pressure)
1	194.5	20.79	1.3179	131.79
2	194.3	20.79	1.3179	131.79
3	197.9	22.40	1.3502	135.02
4	198.4	22.67	1.3555	135.55
5	199.4	23.15	1.3646	136.46
6	199.9	23.35	1.3683	136.83
7	200.9	23.89	1.3782	137.82
8	201.1	23.99	1.3800	138.00
9	201.4	24.02	1.3806	138.06
10	201.3	24.01	1.3805	138.05
11	203.6	25.14	1.4004	140.04
12	204.6	26.57	1.4244	142.44
13	209.5	28.49	1.4547	145.47
14	208.6	27.76	1.4434	144.34
15	210.7	29.04	1.4630	146.30
16	211.9	29.88	1.4754	147.54
17	212.2	30.06	1.4780	147.80

Our example data were collected in the 1840s and 1850s by a Scot named James D. Forbes. Mr. Forbes was interested in estimating altitude by measuring the boiling point of water. His interest was motivated by the difficulty in transporting the fragile barometers of that time. His idea was that if the relationship between the boiling point of water and altitude was predictable, then travelers who could measure the boiling point of water could estimate what altitude they were at. Altitude can be measured in feet or it can be characterized by pressure. Forbes investigated the relationship between boiling point of water and pressure. He wandered around the Alps and Scotland, making observations on the relationship between these two quantities in 17 different places. In the paper that he published in 1857 he addressed the problem of predicting pressure from boiling point. How well are they related; is the relationship strong or weak? Forbes had a theory based on the physics of the situation. His theory predicted that over the range of observed values in this data, the graph of boiling point of water against the logarithm of pressure should be a straight line. Following

Forbes, we will take this data, take the logarithm of the pressure and see if the relationship between the logarithm of the pressure and the boiling point can be approximated by a straight line.

SCATTER PLOT FOR FORBES DATA



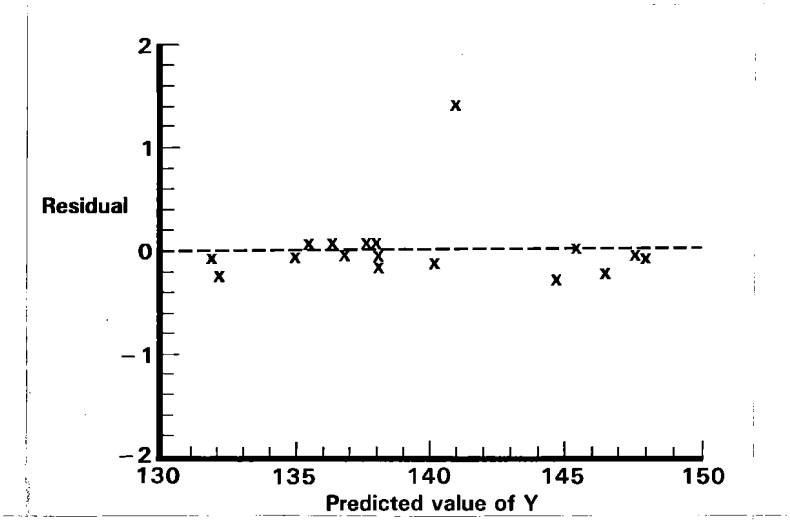
We can begin to investigate Forbes' hypothesis by making a scatterplot of his data. (The pressure is multiplied by 100 so the numbers are not quite so close together; this has no effect conceptually on the models we are using.) The pattern of the data seems to approximate a straight line.

FORBES DATA DIAGNOSTICS--OUTLIERS

Case Number	X	Y	Predicted	Residual
1	194.50	131.79	132.04	− 0.25
2	194.30	131.79	131.86	− 0.07
3	197.90	135.02	135.08	− 0.06
4	198.40	135.55	135.53	0.02
5	199.40	136.46	136.42	0.04
6	199.90	136.83	136.87	− 0.04
7	200.90	137.82	137.77	0.05
8	201.10	138.00	137.95	0.05
9	201.40	138.06	138.22	− 0.16
10	201.30	138.05	138.13	− 0.08
11	203.60	140.04	140.19	− 0.15
12	204.60	142.44	141.08	1.36
13	209.50	145.47	145.47	0.00
14	208.60	144.34	144.66	− 0.32
15	210.70	146.30	146.54	− 0.24
16	211.90	147.54	147.62	− 0.08
17	212.20	147.80	147.89	− 0.09

Forbes fit a regression line to his data. For each of the 17 points listed here, we show the values of the line. That is, for each X, there is a predicted value of Y. For each of these predictions, there is a difference--the residual--between Y and the predicted value of Y. We see that one of these residuals--the one for Case 12--is much larger than the others.

RESIDUAL PLOT FOR FORBES DATA

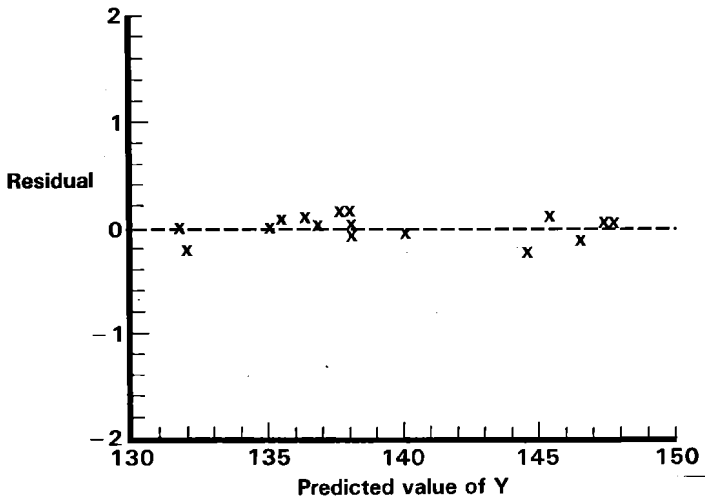


Residuals are easier to interpret--and to see--if we plot them. The 1.36 residual for Case 12 is an outlier. All the other residuals are within about 0.35 of 0. Forbes said of Case 12 that "It evidently was an error."

What do we do after we have used the residual plot to identify points that are unusual? The first thing is to decide whether the unusual points are valid. Was something miscopied or misread? (Of course, it's too late to verify Forbes' data.) Alternatively, an outlying point could be the result of another process; for example, the relationship between boiling point and altitude might be stable throughout a limited range of boiling points, but change at others. Or the outlying point could be a valid point that needs to be investigated further. Indeed, outliers are sometimes the most important cases for a study. The fact that outliers are by definition unlike the other points sometimes signals something unusual going on that is worthy of further investigation. For example, in a study of effectiveness of schools a few years ago at Rand, the researchers used regression to predict test

score outcomes from the usual explanatory characteristics of schools. They then studied those schools who were outliers on the high side for clues to particularly effective teaching techniques.

RESIDUAL PLOT FOR FORBES DATA WITH CASE 12 DELETED



Let's suppose we thought the point might have been miscopied. We can refit Forbes' data with case 12 deleted, and see that all the residuals are now fairly close to 0.

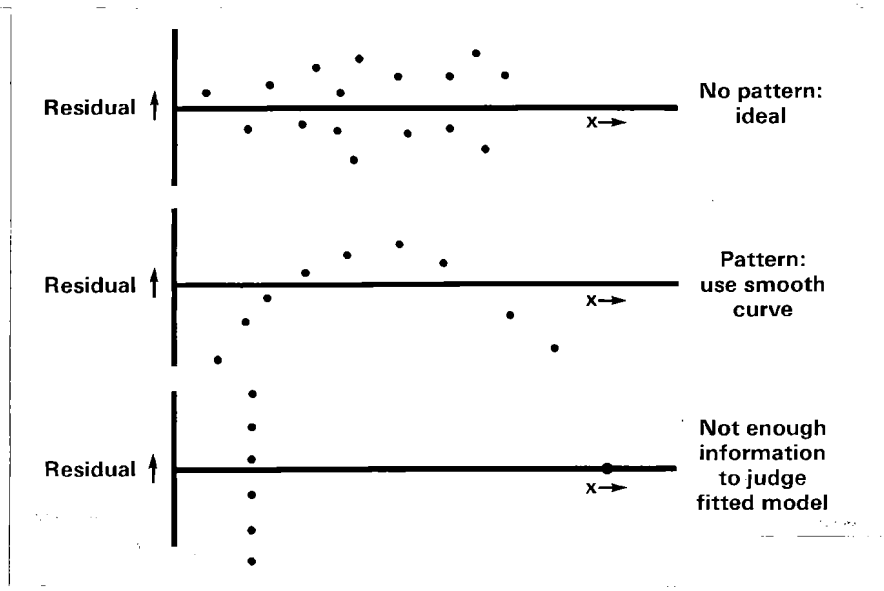
SUMMARY FOR FORBES DATA

Quantity	Value using all data	Value without case 12
Intercept	-42.131	-41.203
Slope	0.895	0.891
σ residuals (SEE)	0.379	0.113
R^2	0.995	0.999+

The real question about Case 12 is whether it makes a large difference in the estimated relationship between altitude and boiling point. In this instance, both the intercept of the regression line and its slope change only slightly when we eliminate Case 12, so including it makes little difference in the prediction. But it does make a difference in our measure of how good the prediction is, namely the standard deviation of the residuals, which is often referred to as the *standard error of estimate* (SEE). When Case 12 is included, the standard deviation of the residuals is about three times as large as it is without Case 12. Since we are not in a position to investigate the validity of Case 12, we would probably want to report both of these equations.

I have included the value for R^2 to make the point that in the physical sciences you frequently get relationships that are extremely good--over 99 percent of the variability explained. As we will see, when we are estimating phenomenon such as cost and performance, we won't be able to do nearly that well.

POSSIBLE RESIDUAL PLOTS



We can also use residual plots to improve our predictions. Here are some other possible residual plots that we might have seen. In each case, the residual is plotted against the explanatory variable. The first residual plot reveals no pattern.

If there is a pattern, we want to exploit it to improve our predictions. The second residual plot is an example of a pattern. It looks like a smooth curve. In this case, we would be inclined to choose a different explanatory variable to see if we can get a better prediction of Y.

The last residual plot is a little strange. There are only two possible values for the explanatory variable, and there are a number of different values for the outcome variable for the same value of X. This residual plot should be a warning sign. The slope of the line is determined by only one point. Thus we lack sufficient information to judge how well this model will fit other data sets. Some of the measures of accuracy that we will discuss apply to situations like this.

MULTIPLE REGRESSION

Definition

- **Several explanatory variables**
- **Best fit of surface to data points**

Predicted outcome $= a$ (Intercept) $+ b \times X_1$ (Explanatory variable 1) $+ c \times X_2$ (Explanatory variable 2)

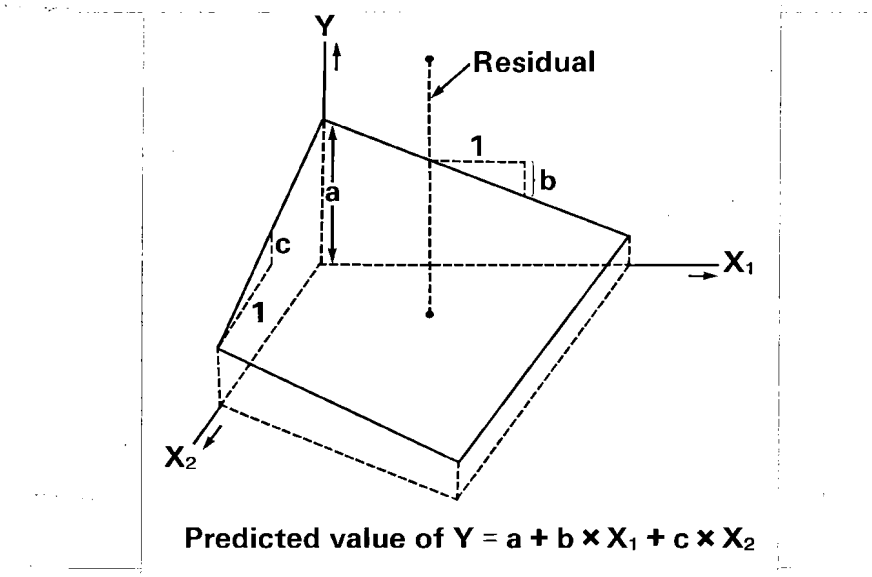
Residual = Outcome value - Predicted value

Residual $\sigma = \sqrt{1 - R^2} \times \sigma Y$

We noted that a pattern in the residual plot might lead us to look for another explanatory variable. But we might not want to throw away the first one. That brings us to the subject of multiple regression. As the name implies, multiple regression is the same as simple linear regression except that we have several explanatory variables. Instead of fitting a line, we fit a surface. We fit the surface the same way we fit the line: minimizing the sum of the squared deviations.

For example, in a multiple regression with two explanatory variables, a regression surface would be described by the intercept term, the coefficient of the first explanatory variable, and the coefficient of the second explanatory variable. Just as before, the residual is the outcome value minus the predicted value. Just as before, the residual standard deviation will be related to the proportion of variance explained and the standard deviation of the outcome variable.

MULTIPLE REGRESSION WITH TWO VARIABLES



The figure above illustrates what multiple regression looks like in the case of two explanatory variables. This figure represents a regression surface. We have two explanatory variables, X_1 and X_2. The regression surface for predicting the outcome variable Y is described by three quantities. First is the intercept value a (where it intercepts the Y axis). Second is the slope relative to X_1; if we look where X_2 is 0, it is depicted by the line in the Y-X_1 plane and has the value b. In this case the slope is negative. Third is the slope in the X_2 direction. That is, we hold X_1 constant and look in the Y-X_2 plane. The distance labeled c is the slope here. The residual of a point is shown as the distance between the point and the plane as the vertical dotted line depicts.

Finding the predicted value of Y from a multiple regression is analogous to the simple linear regression case. It is the intercept a plus b times X_1, plus c times X_2.

MULTIPLE REGRESSION: NEW FEATURES

- **Relationship between explanatory variables**

  Example

  Innovation { % new / Number of new steps

- **Creation of new explanatory variables — interactions**

  Example

  Level of project definition
  and
  Process development stage

We need to note two new features of multiple regression. The first is the relationship *between* the various explanatory variables. You will recall an example of this situation in the workshop discussion of the cost growth model. There are two possible measures of innovation in that model: the percent of the plant's capital cost that is associated with new technology, and the number of new steps in the process. Because these two variables are very closely related--that is, one predicts the other quite well--having both in a regression equation does not add much new information. In cases like this, we want to be sure that we are not double counting and fooling ourselves by thinking we are getting more information than there is.

Second, having more than one explanatory variable gives us a new dimension: *interactions*. Rather than simply adding up the estimated effects of the two new explanatory variables, we may create a third variable that is a combination of the two. In this case there may be some synergism between the two variables, that is to say, the whole is not simply the sum of its parts. Two possible explanatory variables in

the cost growth model are the level of project definition and the stage of process development. The model predicts cost growth more accurately if we allow the level of project definition to have a different effect depending on the stage of process development.

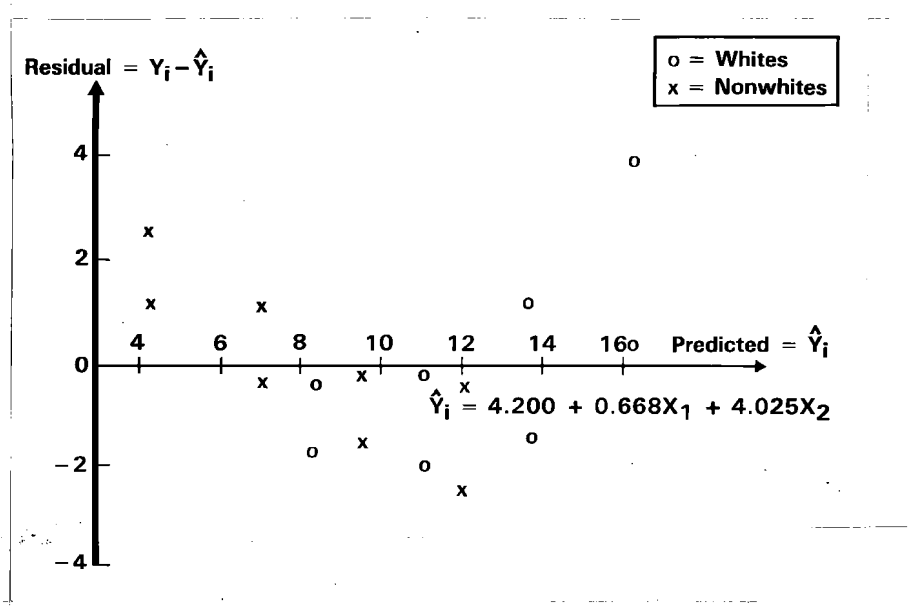
MULTIPLE REGRESSION: EXAMPLE

Outcome variable: Annual wages
Explanatory variables: Years in labor force and race

Worker	Annual wages in thousands of dollars Y	Years in labor force X_1	Race X_2
1	5.455	0	0
2	6.600	0	0
3	6.630	4	0
4	8.022	4	0
5	8.059	8	0
6	9.751	8	0
7	9.751	12	0
8	11.853	12	0
9	6.545	0	1
10	7.920	0	1
11	8.955	4	1
12	10.835	4	1
13	12.250	8	1
14	14.823	8	1
15	16.759	12	1
16	20.279	12	1

Let's look at an example of how to use plots in developing a regression equation. In this hypothetical example, the outcome variable is annual wages and the two explanatory variables are years in the labor force and race. We coded race as 0 for white and 1 for nonwhite. Years in the labor force range from 0 to 12 years.

PLOT OF RESIDUAL WAGES ON PREDICTED WAGES

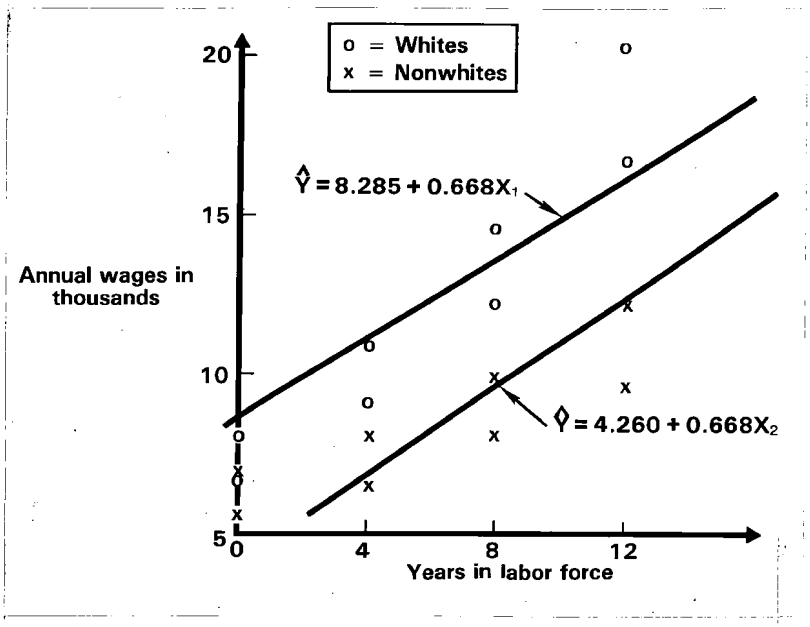


We can fit a regression of annual wages on years in the labor force and race. The fitted regression line is 4.2 plus about .7 times the number of years in the labor force plus 4 times the race. That is, on average, all other things being equal, whites earn about $4,000 more than nonwhites. Also on average, all other things being equal, each additional year of experience is worth about $600 to $700 in wages.

We can also plot the residuals. Here I have plotted them against the predicted value of Y. Note that we have an outlying point. Also note that instead of plotting the residuals against race, I have used a different symbol for the two values, 0 and 1, of the race variable. This is called a *dummy variable*, and a number of models that you will see later in the workshop use them.

When we use these dummy variables, we see that the pattern of the residuals is different for points corresponding to nonwhites as compared with whites.

WAGE PREDICTIONS WITHOUT INTERACTIONS



The difference in the residual patterns might suggest that whites and nonwhites, on average, earn different amounts of money. Here is another way to see the different patterns. In this diagram, we have shown the separate intercept for the whites and the nonwhites. Here the predicted wage for nonwhites is about $4,000+, again with about $700 additional for each year of experience. For whites, of course, it's about $8000. You can see a definite pattern suggesting that nonwhites and whites may not achieve the same gain in wages for each year of experience.

WAGE PREDICTIONS USING INTERACTION

We want to see whether the relationship between wages and years of experience is the same for whites and nonwhites. To do that, we will introduce an interaction term. Using an interaction means that we are allowing, on average, the possibility of a different relationship between wages and experience for whites than for nonwhites. What the data show us very clearly is that, on average, each additional year of experience is worth on average about an additional $400 a year for nonwhites, whereas it is worth an additional $2000 a year for whites. Interestingly enough, nonwhites and whites without experience start at fairly similar points.

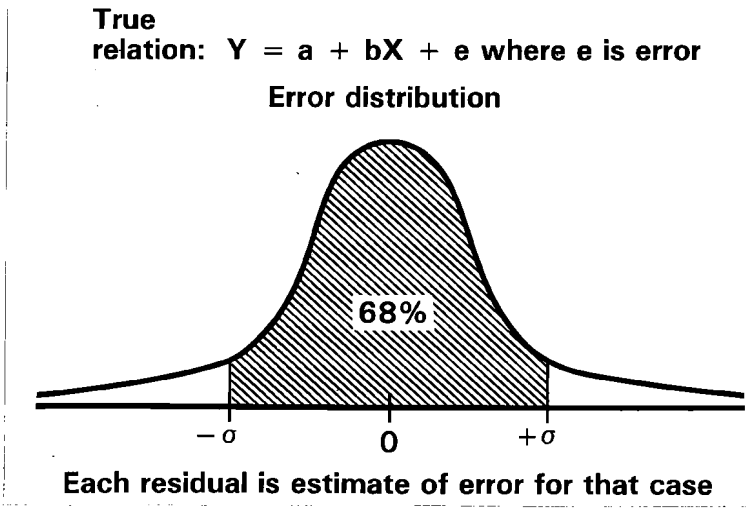
III. MAKING STATISTICAL INFERENCES FROM REGRESSIONS

MAKING STATISTICAL INFERENCES FROM REGRESSIONS

- **How sure are we that true relationship is "close" to fitted equation?**
- **How good a predictor is fitted regression for new point?**

There are two main reasons why we are interested in this topic. First, after we fit a line or a surface to a set of points, we want to know how sure we can be that the true relationship between the outcome variable and the explanatory variables is close to what our fit says it is. In particular, we want to know how confident we can be that the relationship would hold in a new situation. For example, if we had chosen a different set of plants to estimate cost growth, would we have gotten substantially the same results--that is, much the same line? Second, if we want to use our model to make a prediction for a new point--for example, predict cost growth for a new plant--how good would that prediction be?

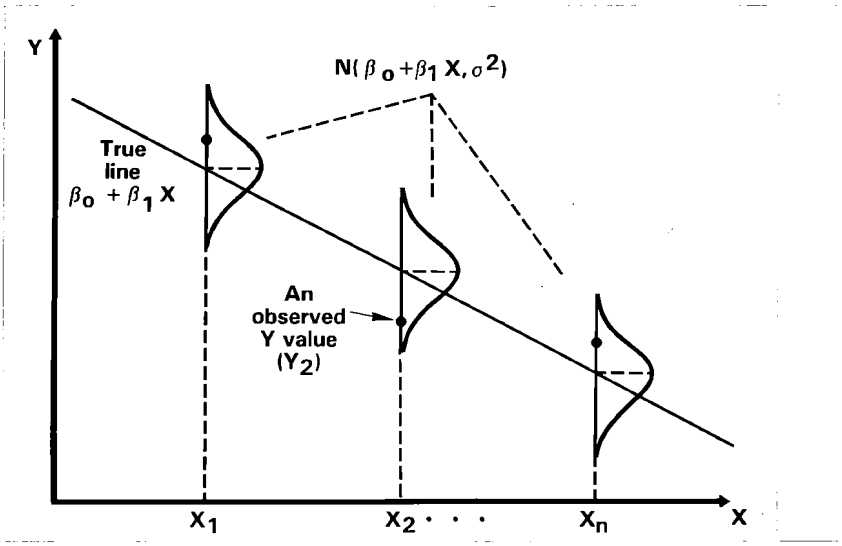
MODEL OF ERRORS IN REGRESSION



To address these issues, let us look at a model of errors for a regression equation. Consider the situation of a single explanatory variable; conceptually, the idea is identical for multiple regression.

Think of the true relationship between the outcome variable Y and the explanatory variable X as being the intercept value, plus the slope times the explanatory variable, plus some error--something left over. This could be measurement error, or it could be that we would need to know many things to predict Y perfectly and we only know a few. We are going to have to make some assumptions about the error. The standard assumption is that the error distribution tends to come from a bell-shaped curve, as shown here. This curve is called a *normal curve* or *Gaussian curve*. Because it is for errors, we assume it is centered at 0. Like any other quantity, it has a standard deviation. The interval from $-\sigma$ to $+\sigma$ will contain about 68 percent of the errors; plus or minus 2 standard deviations will contain 95 percent of the errors. You should think of the residuals we talked about earlier as estimates of the error for each individual case.

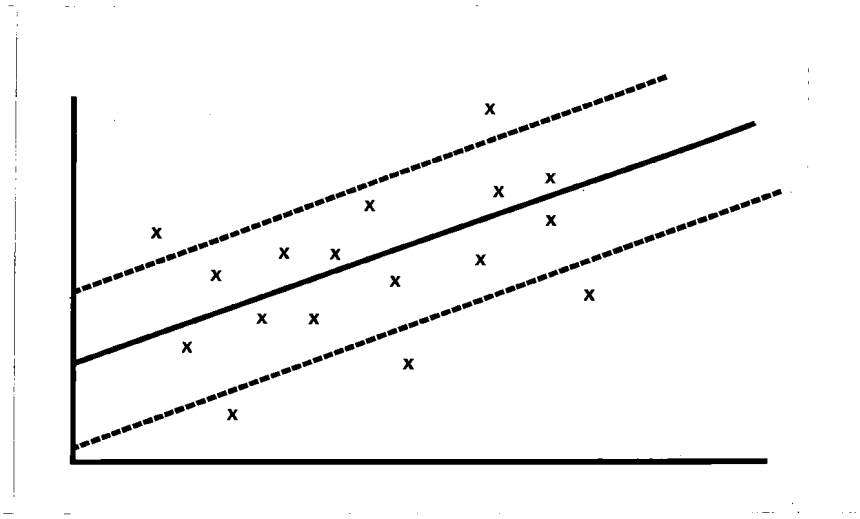
SCHEMATIC OF ERRORS IN REGRESSION



This diagram depicts errors around the regression line. The explanatory variable is on the X-axis, the outcome variable on the Y-axis. We can think of the line as indicating the "true" relationship between these variables. I have indicated this line with Greek letters-- β_0 plus β_1 times X.

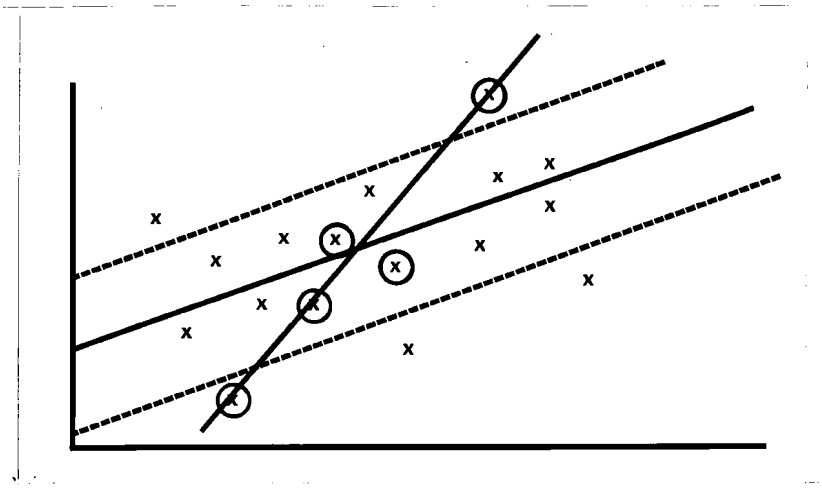
We think about the points as being generated as follows. Take a particular value X and locate its predicted Y value. Then sample a point, as shown by the bell-shaped curves. The diagram shows three sampled points. This is the conceptual model underlying regression.

REPEATED ERROR DRAWS

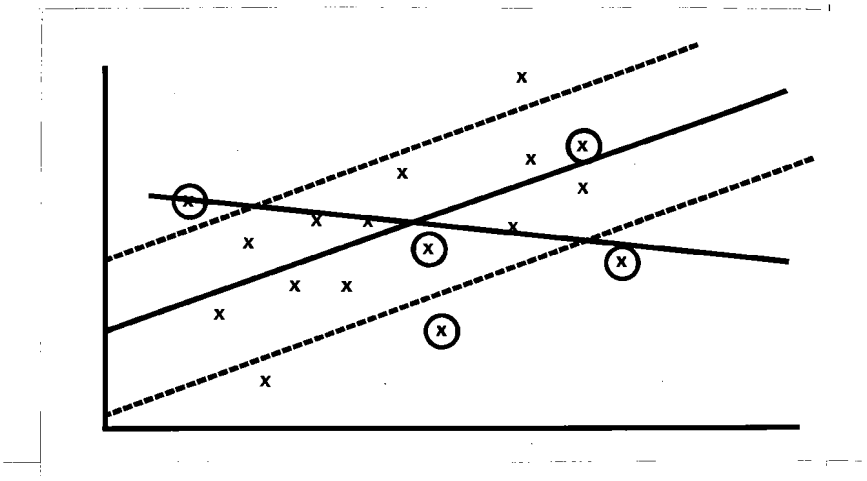


It is clear that you get different estimated lines depending on which points you happened to sample. Let's think of a situation where there is a true line, and we use the bell-shaped curves that we've just finished looking at and draw the dotted parallel lines above and below the true line at a vertical distance of one standard deviation. The X values here represent points that we might sample. Suppose we sample, say, five points and we fit a regression line to them. The fitted line may look quite different from the true regression line.

REPEATED ERROR DRAWS

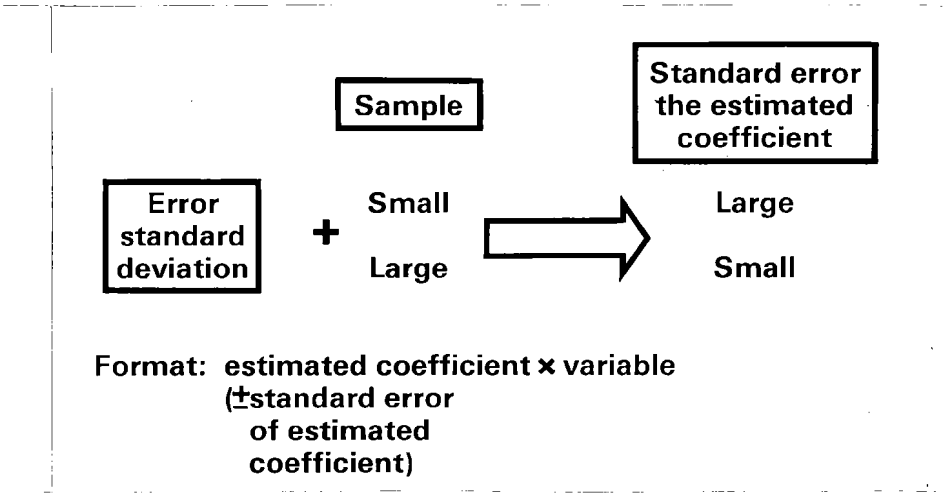


Here is an example of five points we could sample. Note how much steeper the estimated line is than the true line.



On the other hand, if we sampled five other points, we would have gotten a much less steep regression line. These are two possible realizations of the same error distribution.

INFERENCE FOR REGRESSION COEFFICIENTS



What then can we say about the accuracy of the estimated fit of a line to points? The accuracy of the line will depend on how big the errors are--that is, the standard deviation of the errors--and on how many points you sample. The smaller the standard deviation of the errors, the smaller the standard error of the estimated regression coefficient. If you have a small sample, you will tend to have a large standard error of the estimated regression coefficient, that is, the estimated slope. If you have a large sample, you will have a small standard error of the estimated coefficient. (A computer program will compute what these estimated standard errors are.) Thus the standard error of the estimated coefficient is a measure of the variability of the estimate of the slope. When we use the standard error of the estimated coefficient in the rest of this workshop, we will display it in the form shown on the bottom of this chart.

USE OF ERROR MODEL

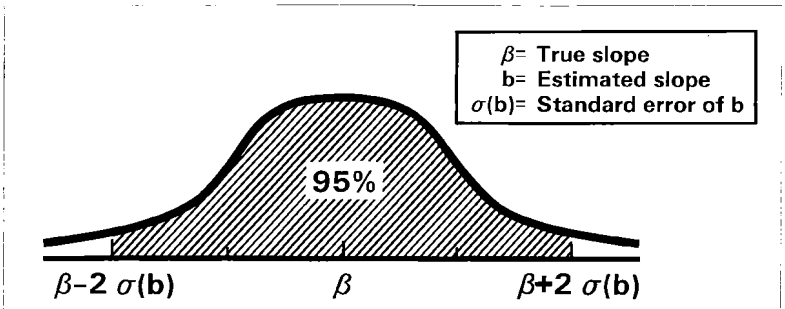
Make statistical inference about:

- **Regression coefficients**
- **Predicted line (surface)**
  e.g. true relationship
- **A new case**

We will use the error model we have just discussed for three purposes. First, we will use it to make inferences about the true relationship for a particular explanatory variable. (In a multiple regression we may be talking about more than one coefficient--more than one slope.) Second, we will use it to make inferences about how close the predicted line, or the predicted surface, is to the true surface. Finally, we will use it to infer how good a prediction we can make about a new case--for example, a new chemical plant.

We will consider each of these purposes in turn.

CONFIDENCE INTERVALS



Distribution of b

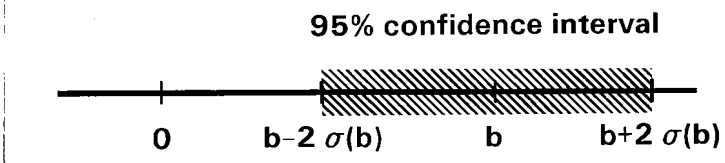
$$\beta - 2\,\sigma(b) \leq b \leq \beta + 2\,\sigma(b)$$

95% confidence interval means

$$b - 2\,\sigma(b) \leq \beta \leq b + 2\,\sigma(b)$$

We need to introduce the concept of a confidence interval. Let's consider the case of confidence intervals for coefficients. In this diagram, the true slope is labeled β, to distinguish it from b, which is the slope we've estimated from our particular data. Our regression program has printed out the standard error of b. What does that mean? That means that b is an estimate of β, and that if you take b plus 2 standard errors and b minus 2 standard errors, you will be 95% confident that the value of β falls within that interval.

STATISTICAL TESTS



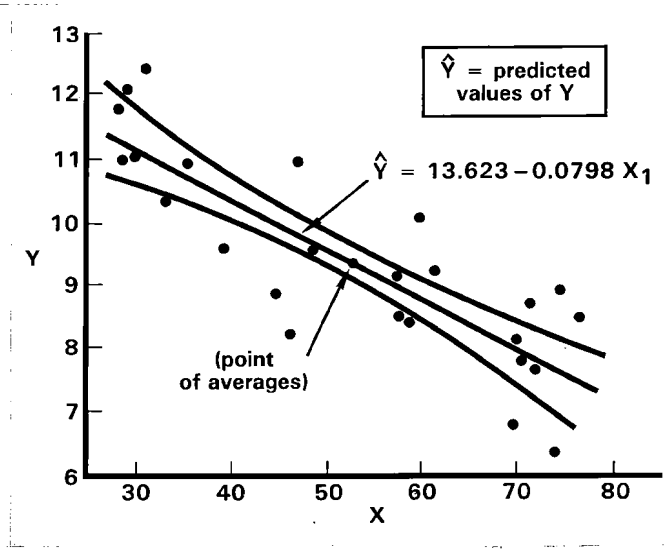
"Reject hypothesis that slope (correlation) is zero at 5% significance level"

That is:

The difference of the estimated coefficient from zero is statistically significant

The 95% confidence interval concept is useful in the definition of statistical significance. That is, is there really a relationship between the outcome variable and the explanatory valuable? Consider the following situation. Our regression program gives us an estimated regression coefficient and the value of 2 standard errors. We observe whether the 95% confidence interval includes 0 or not. If it does not include 0, we can reject the hypothesis that the true slope or correlation is 0 at the 5% significance level (5% coming from 100% - 95%). That is, the difference between the estimated coefficient and zero is statistically significant. If the 95% confidence level does include zero, there is a possibility that there really is no relationship between the explanatory variable and the outcome variable. In this workshop, when we say that a relationship is statistically significant, we mean that there are--at most--only 5 chances in 100 that a value as large as the observed one could occur where in fact the true value is zero. The statistical significance is a measure of how consistent our data is with the hypothesis of no relationship between X and Y. In many of the equations you will see in the workshop, the chances are more like 3 in 1000.

CONFIDENCE INTERVAL FOR TRUE RELATIONSHIP



Let's use the idea of a confidence interval to answer the question, "How sure are we that the regression line we have fitted is close to the true regression line?" In fitting a line the notion of leverage is important. A change in an end point will move the line up or down a lot, while a change in a middle point will not change the slope nearly as much. The width of a 95% confidence interval depends both on the standard errors of the estimated coefficients and the particular values of the explanatory variables that the interval is being computed for. Thus the answer to the above question is: "For a particular value of X, we are 95% confident that the true expected value of Y lies within the interval above X between the bands." Note that, as we would expect from the leverage notion, these bands narrow in the middle and widen at the ends.

You saw confidence intervals around predicted values from regressions in the workshop in a more practical context--namely, when you start extrapolating beyond the range of your data. Here you should be careful because you cannot be sure whether the same relationships hold.

Note that because the band shown gives confidence intervals for *expected values* of Y, the data points will frequently fall outside the band because the points are modeled as the true line plus additional error.

STATISTICAL INFERENCE ABOUT A FUTURE VALUE

Sources of error:

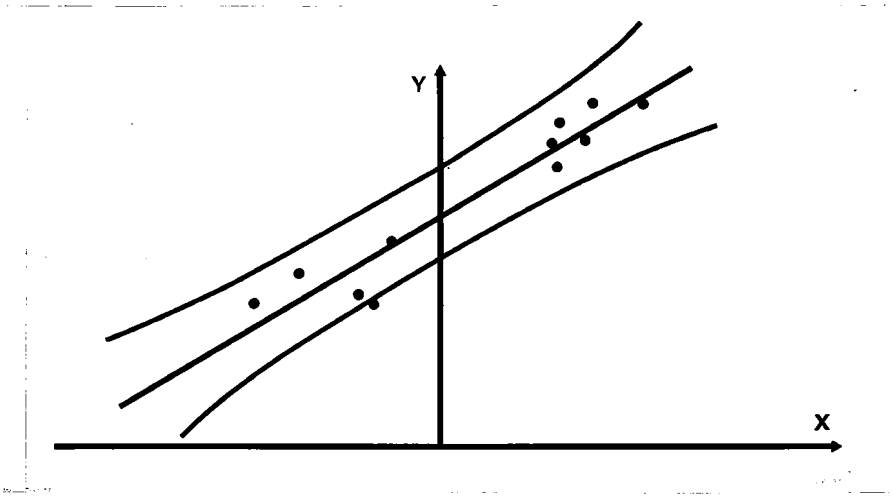
- **Uncertainty about line (surface)**
- **Additional error due to the new observation**

Result : more uncertainty

Finally, what can we say about the future-value problem?

In making a prediction about a case--e.g., a new process plant--there are two sources of error. First, there is the source of error that we've just been talking about: How sure are you that the fitted line is close to the true line? Second, even if you know the line perfectly, some additional error will arise from the new observation. Thus, in predicting a future value, there is more uncertainty than if you were simply assessing how far away the estimated line is from the true line.

CONFIDENCE INTERVAL FOR PREDICTING A NEW VALUE



Let's consider the following case. Here is an example of an explanatory variable, X, and an outcome variable, Y. The middle line is the true line surrounded by the points that were used to estimate it. These slightly flared lines provide 95% confidence intervals for future values. That is to say, if a new Y value came in for a particular value of X, we are 95% confident that it would be in the confidence interval determined by these curves.

Notice two things about these confidence intervals. First, they are wider for X values at the extremes than for ones in the middle. Second, the flaring of the band is not as pronounced as the band shown in the earlier figure. The difference is that there are two sources of error here instead of only one. First is the error in the estimated line itself, and second is the error around the line. The figure shown reflects adding the two together.

CHOOSING A REGRESSION MODEL

- **A priori considerations**
  - **Outcome variables**
  - **Explanatory variables**
- **Goal: choose model to explain data parsimoniously**
- **Use statistical tools to achieve goal**

An important part of the planning process consists of applying our knowledge to specify an outcome variable for the regression model and to identify a set of candidate explanatory variables. How detailed these *a priori* specifications will be depends on how detailed our knowledge is of the phenomenon being modeled. For example, we often know enough to make an intelligent guess about the functional form of the regression model.

The goal of this process is to choose a regression model that will explain the data as parsimoniously as possible. By parsimoniously, I mean using no more coefficients or explanatory variables than are absolutely necessary. I've described some of the statistical tools that you can use to achieve this goal. In particular, statistical tools play a large role in choosing the best functional form for the regression model.

The process of modeling is as much an art as it is a science. Effective modeling requires combining good knowledge of the process being considered with appropriate techniques to exploit the information

contained in the data. The statistical analyst's toolbox contains a variety of graphic methods as well as formal statistical tests. We've seen that we can use residual plots and statistical tests to determine a good scale to use in defining both outcome and explanatory variables. They also help us to decide how to combine our raw explanatory variables into new variables for a better fit. Since changing the scale of the outcome variable will, in general, change how well a given set of explanatory variables fit the data, the process is iterative.

To illustrate how our analysts have gone about building regression models to estimate cost growth, startup costs, and other outcomes of interests in process plants, let us work through a simple example of building and testing a regression model.

APPLICATION--TREE DATA

D = Diameter	*H = Height*	*Vol = Volume*
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
10.5	72	16.4
10.7	81	18.8
10.8	83	19.7
11.0	66	15.6
11.0	75	18.2
11.1	80	22.6
11.2	75	19.9
11.3	79	24.2
11.4	76	21.0
11.4	76	21.4
11.7	69	21.3
12.0	75	19.1
12.9	74	22.2
12.9	85	33.8
13.3	86	27.4
13.7	71	25.7
13.8	64	24.9
14.0	78	34.5
14.2	80	31.7
14.5	74	36.3
16.0	72	38.3
16.3	77	42.6
17.3	81	55.4
17.5	82	55.7
17.9	80	58.3
18.0	80	51.5
18.0	80	51.0
20.6	87	77.0

The example we will work with is the problem of estimating the volume of a tree in cubic feet as a function of the tree's height and its diameter. The purpose of this model is to provide foresters with an easy way of estimating the amount of timber in a given area of forest. To do this, they need to determine the volume of any given tree. Measuring a tree's volume directly is difficult, but measuring height and diameter is relatively easy. Thus, a forester would like to develop a table, or an equation, that enables him to estimate the volume of a tree from its diameter and/or its height.

The data on the chart come from a sample of 31 black cherry trees from the Allegheny National Forest in Pennsylvania. The trees were cut, and their diameter, height, and volume recorded. The numbers in the diameter column represent the diameter of the tree in inches at approximately 4-1/2 feet above the ground. The height is in feet, and the volume is in cubic feet. You can see that a fairly wide range of tree sizes is represented in this sample.

MODELS OF VOLUME

- **Naive model**

  $V = a + bH + cD + error$

- **Other possibilities**
  - **Other explanatory variables e.g. D^2**
  - **Relation of volume, height, diameter of cone**

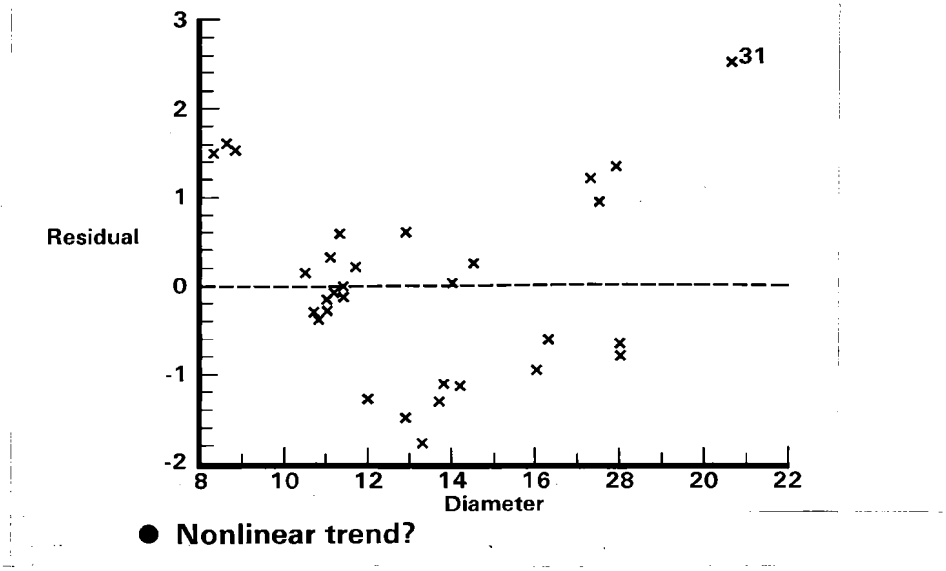
    $V = a \times D^2H$ or $\log V = a + b \log D + c \log H$
  - **Other more complex relationships**

How should we go about modeling the volume of a tree? Since we have three columns of data, our first, naive approach might be to fit a regression with the volume as the outcome variable on two explanatory variables: height and diameter. There are, of course, a number of other possibilities. We could add in other explanatory variables such as the square of the diameter or an interaction term between the height and the diameter.

We could also use what we know about trees and solid geometry to postulate a model. For example, suppose we thought of a tree as a cone. We know that for cones, the volume is proportional to the height times the square of the diameter. That would suggest regressing the logarithm of the volume as the outcome variable on the logarithm of the diameter and the logarithm of the height. Because trees are not perfect cones, we might follow this line of reasoning and develop other, more complex, postulated relationships.

Below, I sketch the steps that a statistical analyst might take to develop a model for this situation. In working through this example, I am less interested in the details of each step than I am in imparting a feel for the analytic *process*.

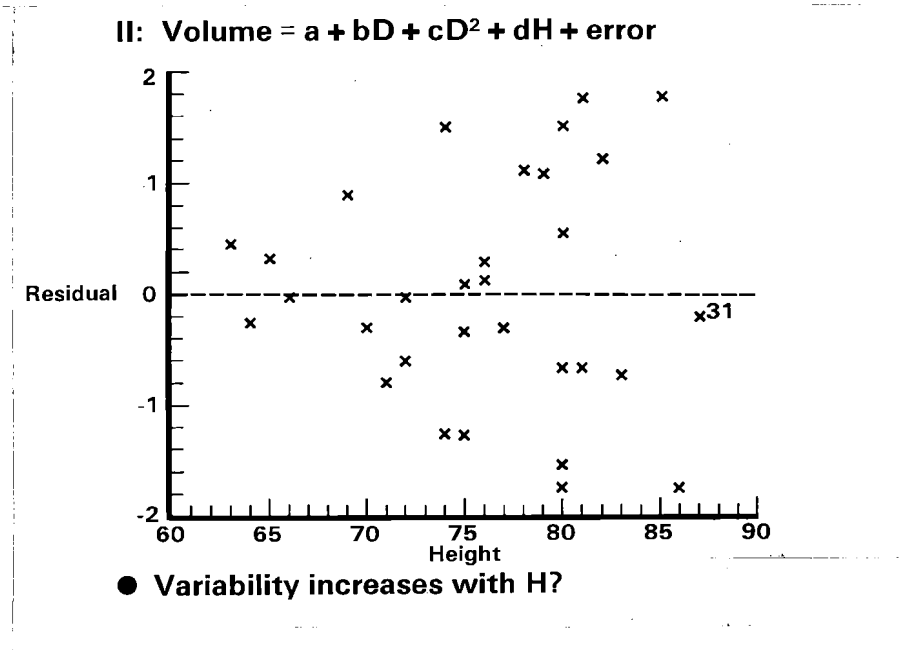
RESIDUAL PLOT FOR NAIVE MODEL OF TREE DATA



Suppose we fit the naive model to the data--that is, we regress volume on height and diameter for the 31 trees. This figure shows a plot of the residuals from this naive model (on the vertical axis) against the diameter (on the horizontal axis). There are several other plots we could make, but we look at just this one.

It is apparent from this plot that there may be some nonlinear dependence of volume on diameter. That is, the plot is somewhat U-shaped. Further, Tree 31 (which you may recall was the largest tree) appears to be something of an outlier. Given the U-shape form of this plot, as a first step we add a squared term in the diameter. That is, we think of this pattern as being a parabola.

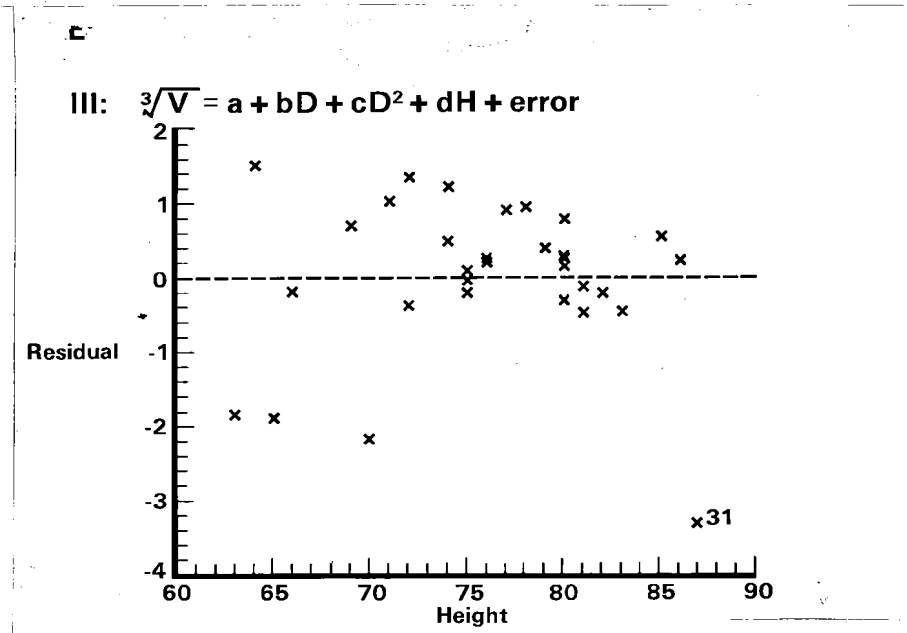
RESIDUAL PLOT TREE DATA: MODEL II



This plot shows a plot based on Model II--the naive model augmented with an additional explanatory variable, the square of diameter. When plotted against diameter, the residuals do not appear to have any pattern. A plot of the residuals from this model (on the vertical axis) is displayed against height (on the horizontal axis). Notice that the point for Tree 31 has been brought in towards the regression surface, and there is no obvious nonlinear shape to the plot. However, it does appear that the variability, i.e., the standard deviation of the error, may increase with the height, thus violating one of the assumptions of our model of errors. That is, each error is assumed to be drawn from the same bell-shaped curve and hence has the same standard deviation.

There are a number of ways to improve the model when we detect nonconstant variability. One very common cause of this problem is that the scale in which the outcome variable is measured is not appropriate. In this case, if we rescale the outcome variable, our regression model will then have constant variability. Various methods can be used to suggest what rescaling might work. For these data, examination of a number of residual plots suggested taking the cube root of volume.

RESIDUAL PLOT WITH NEW OUTCOME VARIABLE
FOR TREE DATA: MODEL III



- **Variability constant but perhaps outlier?**

This chart shows a residual plot for Model III, the cube root of volume regressed against the Model II explanatory variables. The plot indicates that the standard deviation structure for the errors has been improved and, with the possible exception of Tree 31, there is no detectable pattern of the residuals. The fit of the model is quite sensitive to the presence of Tree 31--not a very satisfactory state of affairs.

Are there any other models that either intuition or the data suggest that we should try? Recall the earlier suggestion that perhaps trees tend to be shaped like cones, which suggested taking logarithms of both the outcome variable and the explanatory variables. This and other functions of both the outcome variable and the explanatory variables were tried; I will not report the details here. Let me emphasize that developing a regression equation is an iterative process. In particular, every time you change the scale of the outcome variable, you go back to "square one" in choosing the function form for the explanatory variables.

I will conclude this example by going back to our original outcome variable, Volume, and presenting a regression model that does a satisfactory job.

REGRESSION SUMMARIES FOR TREE DATA: MODELS I AND IV

Outcome variable: V

Explanatory variables	Model I	Model IV
Intercept	−58.0 (±8.6)	−65.6 (±124.7)
D	4.7 (±.3)	−21.5 (±5.1)
H	.3 (±.1)	−1.8 (±9.4)
D log (D)		7.2 (±1.4)
H log (H)		.4 (±1.8)

Recall that the naive model consisted of regressing the volume of the tree as the outcome variable on the diameter and height of the tree as the two explanatory variables. I have listed some of the computer output from two regressions on the chart. I give the estimated regression coefficient and their standard errors under Model I. You can interpret these numbers as saying that for each additional inch of diameter, the tree will on average be about 4-1/2 cubic feet larger in volume. Similarly, for every additional foot of height, the tree gains approximately one third of a cubic foot in lumber. The weakness of the relationship between volume and height reflects the fact that height adds very little strength to the prediction, given that diameter is already being used.

After further analysis, we see that Model IV produces a satisfactory fit to the data with no pattern in the residuals. Model IV has the volume in cubic feet as its outcome variable. There is no rescaling. We have added two additional explanatory variables that are functions of diameter and height. The first is the diameter times the

logarithm of the diameter and the second is the height times the logarithm of height. The estimated regression coefficient on height times log of height suggests that this second new variable really need not be included. In fact, a 95% confidence interval on its estimated coefficient does include zero, so a statistical test would suggest not including it. This test also was confirmed by doing a plot of the residuals against height times log of height.

DISPLAYING RESULTS--BLOCK DIAGRAM

Finally, I want to show you the format in which regression models are presented in the rest of this workshop. First, as an overview, the model will be displayed in the form of a block diagram showing the explanatory and the outcome variables.

DISPLAYING RESULTS: EQUATION

Predicted value = a Intercept

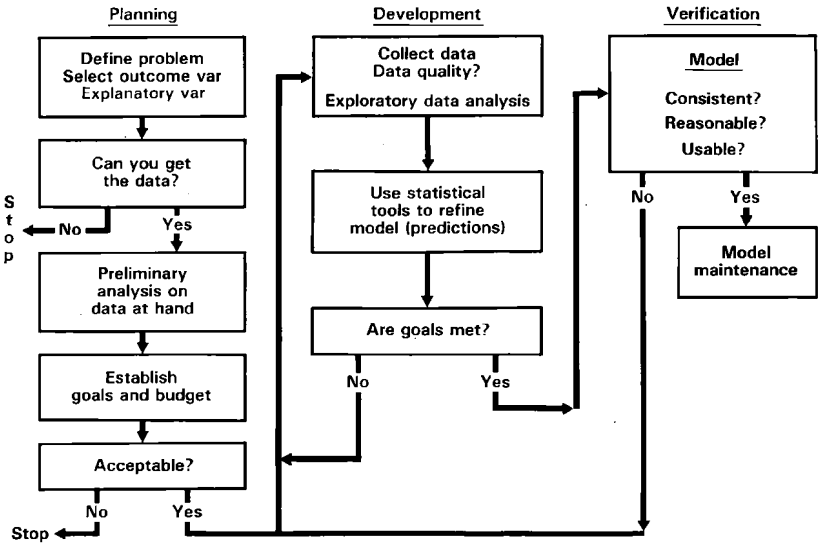
$+ b_1 \times X_1$
$+ b_2 \times X_2$
$+ b_3 \times X_3$
} **Explanatory variables**

•
•
•

R^2
Standard Error of Estimate

After discussion of the model, the equation will be shown. The predicted outcome variable will result from an estimated intercept, plus a series of explanatory variables, each of which has an estimated slope or coefficient. Underneath that coefficient there may be a plus or minus number, which will be the standard error of the estimated coefficient. If the model fits, you can be 95% confident that the true value of the coefficient lies between plus or minus two standard errors of the estimated coefficient. The displayed value of R^2 is the proportion of variability explained by the regression model. We also will display the standard error of the estimate (SEE)--the standard deviation of the residuals--which is a measure of how well the equation fits the data. The size of the SEE is an important factor in assessing the adequacy of the model for predictive purposes.

SUMMARY OF MODEL BUILDING



We conclude by spending a few moments on possible pitfalls in the modeling process. When we talked about model verification in the beginning, we said that the model should be consistent, reasonable, and usable. I want to leave you with some thoughts about the reasonableness of a model.

We are using these models to understand how changes in the explanatory variable will affect the outcome variable. So at least implicitly we are thinking about a *causal* relationship between the explanatory variable and the outcome variable. But obviously, unless we have a well-understood behavioral mechanism, all that statistics and data-fitting are going to tell us is that outcome variables and explanatory variables tend to change together. That is, the two variables are *associated* with one another--and here is where intelligently using additional information to supplement statistical reasoning from the data is helpful. One common problem in this domain goes under the label, for lack of a better one, of "lurking" variables.

A celebrated instance of a lurking variable at work is the example of the storks and babies. Suppose you fit a regression of the relationship between the number of storks in several geographical areas and the number of babies being born in those areas and discover that the number of storks is a very good (inverse) predictor of number of babies. Basic biology suggests that this is a spurious relationship, and further investigation detects a lurking variable on the scene. Storks apparently build nests in chimneys (or perhaps rooftops). Urban areas have more chimneys, hence more babies. Thus a third variable--in this case chimneys and rooftops--is more causally related to the outcome variable than the explanatory variable of storks that was chosen.

GRADUATE SCHOOL ADMISSION--HYPOTHETICAL

	No. applied	No. admitted	% admitted
Total			
Women	**150**	**55**	**37%**
Men	**150**	**65**	**43%**

Lurking variables and more generally omitted explanatory variables can lead to statistical paradoxes. Here's an example. Several years ago, the University of California, Berkeley, decided to study its graduate school admissions because it was concerned about equity. (For ease of exposition, I have simplified some of the numbers from their study; however, the patterns are unchanged.) The study found that 37% of the women and 43% of the men who applied to graduate school were admitted. It appeared that the university might be discriminating against women.

Somebody noted that individual departments, not the university, admitted people to graduate school. Thus one should to look at the individual department records to understand the admissions process.

GRADUATE SCHOOL ADMISSION--HYPOTHETICAL

	No. applied	No. admitted	% admitted
Total			
Women	**150**	**55**	**37%**
Men	**150**	**65**	**43%**
Engineering Dept.			
Women	**50**	**25**	**50%**
Men	**100**	**50**	**50%**
English Dept.			
Women	**100**	**30**	**30%**
Men	**50**	**15**	**30%**

If we look at departments, we see quite a different pattern. In the Engineering department, 50 women applied and 25 of them--50%--were admitted. Of the 100 men who applied, 50 were admitted--also a 50% admissions rate. In the English department, 30 of the 100 female applicants were admitted and 15 of the male applicants--30% in each case. Thus, although the university-wide data suggested that women were being discriminated against, department-level data provided no evidence of discrimination.

Statisticians call this situation Simpson's Paradox. It means that overlooking an important variable can produce misleading results. Let me assure you that in the rest of the presentations in this workshop, we have made every effort to avoid getting into this situation.

REFERENCES

Draper, N. R., and H. Smith, *Applied Regression Analysis*, John Wiley and Sons, Inc., New York, 1981.

Freedman, David, Robert Pisani, and Roger Purves, *Statistics*, W. W. Norton and Company, Inc., New York, 1978.

Hooke, Robert, *How To Tell the Liars from the Statisticians*, Marcel Dekker, Inc., New York, 1983.

Morris, Carl N., and John E. Rolph, *Introduction to Data Analysis and Statistical Inference*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1981.

Ryan, Thomas A., Brian L. Joiner, and Barbara F. Ryan, *Minitab Student Handbook*, Duxbury Press, North Scituate, Massachusetts, 1976.

Weisberg, Sanford, *Applied Linear Regression*, John Wiley and Sons, Inc., New York, 1980.