

MULTIPLE REGRESSION BASICS

Documents prepared for use in course B01.1305,
New York University, Stern School of Business

- Introductory thoughts about multiple regression page 3
Why do we do a multiple regression? What do we expect to learn from it?
What is the multiple regression model? How can we sort out all the
notation?
- Scaling and transforming variables page 9
Some variables cannot be used in their original forms. The most common
strategy is taking logarithms, but sometimes ratios are used. The “gross
size” concept is noted.
- Data cleaning page 11
Here are some strategies for checking a data set for coding errors.
- Interpretation of coefficients in multiple regression page 13
The interpretations are more complicated than in a simple regression.
Also, we need to think about interpretations after logarithms have been
used.
- Pathologies in interpreting regression coefficients page 15
Just when you thought you knew what regression coefficients meant . . .

Regression analysis of variance table page 18
Here is the layout of the analysis of variance table associated with regression. There is some simple structure to this table. Several of the important quantities associated with the regression are obtained directly from the analysis of variance table.

Indicator variables page 20
Special techniques are needed in dealing with non-ordinal categorical independent variables with three or more values. A few comments relate to model selection, the topic of another document.

Noise in a regression page 32
Random noise obscures the exact relationship between the dependent and independent variables. Here are pictures showing the consequences of increasing noise standard deviation. There is a technical discussion of the consequences of measurement noise in an independent variable. This entire discussion is done for simple regression, but the ideas carry over in a complicated way to multiple regression.

Cover photo: Praying mantis, 2003

© Gary Simon, 2003

INPUT TO A REGRESSION PROBLEM

Simple regression: $(x_1, Y_1), (x_1, Y_2), \dots, (x_n, Y_n)$

Multiple regression: $((x1)_1, (x2)_1, (x3)_1, \dots, (xK)_1, Y_1),$
 $((x1)_2, (x2)_2, (x3)_2, \dots, (xK)_2, Y_2),$
 $((x1)_3, (x2)_3, (x3)_3, \dots, (xK)_3, Y_3),$
 $\dots,$
 $((x1)_n, (x2)_n, (x3)_n, \dots, (xK)_n, Y_n),$

The variable Y is designated as the “dependent variable.” The only distinction between the two situations above is whether there is just one x predictor or many. The predictors are called “independent variables.”

There is a certain awkwardness about giving generic names for the independent variables in the multiple regression case. In this notation, $x1$ is the name of the first independent variable, and its values are $(x1)_1, (x1)_2, (x1)_3, \dots, (x1)_n$. In any application, this awkwardness disappears, as the independent variables will have application-based names such as *SALES*, *STAFF*, *RESERVE*, *BACKLOG*, and so on. Then *SALES* would be the first independent variable, and its values would be $SALES_1, SALES_2, SALES_3, \dots, SALES_n$.

The listing for the multiple regression case suggests that the data are found in a spreadsheet. In application programs like Minitab, the variables can appear in any of the spreadsheet columns. The dependent variable and the independent variables may appear in any columns in any order. Microsoft’s EXCEL requires that you identify the independent variables by blocking off a section of the spreadsheet; this means that the independent variables must appear in consecutive columns.

MINDLESS COMPUTATIONAL POINT OF VIEW

The output from a regression exercise is a “fitted regression model.”

Simple regression: $\hat{Y} = b_0 + b_1 x$

Multiple regression: $\hat{Y} = b_0 + b_1(x1) + b_2(x2) + b_3(x3) + \dots + b_K(xK)$

Many statistical summaries are also produced. These are R^2 , standard error of estimate, t statistics for the b ’s, an F statistic for the whole regression, leverage values, path coefficients, and on and on and on and This work is generally done by a computer program, and we’ll give a separate document listing and explaining the output.

WHY DO PEOPLE DO REGRESSIONS?

A cheap answer is that they want to explore the relationships among the variables.

A slightly better answer is that we would like to use the framework of the methodology to get a yes-or-no answer to this question: Is there a significant relationship between variable Y and one or more of the predictors? Be aware that the word *significant* has a very special jargon meaning.

An simple but honest answer pleads curiosity.

The most valuable (and correct) use of regression is in making predictions; see the next point. Only a small minority of regression exercises end up by making a prediction, however.

HOW DO WE USE REGRESSIONS TO MAKE PREDICTIONS?

The prediction situation is one in which we have new predictor variables but do not yet have the corresponding Y .

Simple regression: We have a new x value, call it x_{new} , and the predicted (or fitted) value for the corresponding Y value is

$$\hat{Y}_{new} = b_0 + b_1 x_{new} .$$

Multiple regression: We have new predictors, call them $(x1)_{new}$, $(x2)_{new}$, $(x3)_{new}$, ..., $(xK)_{new}$. The predicted (or fitted) value for the corresponding Y value is

$$\hat{Y}_{new} = b_0 + b_1(x1)_{new} + b_2(x2)_{new} + b_3(x3)_{new} + \dots + b_K(xK)_{new}$$

CAN I PERFORM REGRESSIONS WITHOUT ANY UNDERSTANDING OF THE UNDERLYING MODEL AND WHAT THE OUTPUT MEANS?

Yes, many people do. In fact, we'll be able to come up with rote directions that will work in the great majority of cases. Of course, these rote directions will sometimes mislead you. And wisdom still works better than ignorance.

WHAT'S THE REGRESSION MODEL?

The model says that Y is a linear function of the predictors, plus statistical noise.

$$\text{Simple regression: } Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\text{Multiple regression: } Y_i = \beta_0 + \beta_1 (x1)_i + \beta_2 (x2)_i + \beta_3 (x3)_i + \dots + \beta_K (xK)_i + \varepsilon_i$$

The coefficients (the β 's) are nonrandom but unknown quantities. The noise terms $\varepsilon_1, \varepsilon_2, \varepsilon_3, \dots, \varepsilon_n$ are random and unobserved. Moreover, we assume that these ε 's are statistically independent, each with mean 0 and (unknown) standard deviation σ .

The model is simple, except for the details about the ε 's. We're just saying that each data point is obscured by noise of unknown magnitude. We assume that the noise terms are not out to deceive us by lining up in perverse ways, and this is accomplished by making the noise terms independent.

Sometimes we also assume that the noise terms are taken from normal populations, but this assumption is rarely crucial.

WHO GIVES ANYONE THE RIGHT TO MAKE A REGRESSION MODEL? DOES THIS MEAN THAT WE CAN JUST SAY SOMETHING AND IT AUTOMATICALLY IS CONSIDERED AS TRUE?

Good questions. Merely claiming that a model is correct does not make it correct. A model is a mathematical abstraction of reality. Models are selected on the basis of simplicity and credibility. The regression model used here has proved very effective. A careful user of regression will make a number of checks to determine if the regression model is believable. If the model is not believable, remedial action must be taken.

HOW CAN WE TELL IF A REGRESSION MODEL IS BELIEVABLE? AND WHAT'S THIS REMEDIAL ACTION STUFF?

Patience, please. It helps to examine some successful regression exercises before moving on to these questions.

THERE SEEMS TO BE SOME PARALLEL STRUCTURE INVOLVING THE MODEL AND THE FITTED MODEL.

It helps to see these things side-by-side.

Simple regression:

The model is $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

The fitted model is $\hat{Y} = b_0 + b_1 x$

Multiple regression:

The model is $Y_i = \beta_0 + \beta_1 (x1)_i + \beta_2 (x2)_i + \beta_3 (x3)_i + \dots + \beta_K (xK)_i + \varepsilon_i$

The fitted model is $\hat{Y} = b_0 + b_1 (x1) + b_2 (x2) + b_3 (x3) + \dots + b_K (xK)$

The Roman letters (the b 's) are estimates of the corresponding Greek letters (the β 's).

WHAT ARE THE FITTED VALUES?

In any regression, we can “predict” or retro-fit the Y values that we’ve already observed, in the spirit of the PREDICTIONS section above.

Simple regression:

The model is
$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

The fitted model is
$$\hat{Y} = a + bx$$

The fitted value for point i is

$$\hat{Y}_i = a + bx_i$$

Multiple regression:

The model is
$$Y_i = \beta_0 + \beta_1 (x1)_i + \beta_2 (x2)_i + \beta_3 (x3)_i + \dots + \beta_K (xK)_i + \varepsilon_i$$

The fitted model is
$$\hat{Y} = b_0 + b_1 (x1) + b_2 (x2) + b_3 (x3) + \dots + b_K (xK)$$

The fitted value for point i is

$$\hat{Y}_i = b_0 + b_1 (x1)_i + b_2 (x2)_i + b_3 (x3)_i + \dots + b_K (xK)_i$$

Indeed, one way to assess the success of the regression is the closeness of these fitted Y values, namely $\hat{Y}_1, \hat{Y}_2, \hat{Y}_3, \dots, \hat{Y}_n$ to the actual observed Y values $Y_1, Y_2, Y_3, \dots, Y_n$.

THIS IS LOOKING COMPUTATIONALLY HOPELESS.

Indeed it is. These calculations should only be done by computer. Even a careful, well-intentioned person is going to make arithmetic errors if attempting this by a non-computer method. You should also be aware that computer programs seem to compete in using the latest innovations. Many of these innovations are passing fads, so don’t feel too bad about not being up-to-the-minute on the latest changes.

The notation used here in the models is not universal. Here are some other possibilities.

Notation here	Other notation
Y_i	y_i
x_i	X_i
$\beta_0 + \beta_1 x_i$	$\alpha + \beta x_i$
ε_i	e_i or r_i
$(x1)_i, (x2)_i, (x3)_i, \dots, (xK)_i$	$x_{i1}, x_{i2}, x_{i3}, \dots, x_{iK}$
b_j	$\hat{\beta}_j$

In many regression problems, the data points differ dramatically in gross size.

EXAMPLE 1: In studying corporate accounting, the data base might involve firms ranging in size from 120 employees to 15,000 employees.

EXAMPLE 2: In studying international quality of life indices, the data base might involve countries ranging in population from 0.8 million to 1,000 millions.

In Example 1, some of the variables might be highly dependent on the firm sizes. For example, the firm with 120 employees probably has low values for gross sales, assets, profits, and corporate debt.

In Example 2, some of the variables might be highly dependent on country sizes. For example, the county with population 0.8 million would have low values for GNP, imports, exports, savings, telephones, newspaper circulation, and doctors.

Regressions performed with such gross size variables tend to have very large R^2 values, but prove nothing. In Example 1, one would simply show that big firms have big profits. In Example 2, one would show that big countries have big GNPs. The explanation is excellent, but rather uninformative.

There are two common ways for dealing with the gross size issue: ratios and logarithms.

The ratio idea just puts the variables on a “per dollar” or “per person” basis.

For Example 1, suppose that you wanted to explain profits in terms of number of employees, sales, assets, corporate debt, and (numerically coded) bond rating. A regression of profits on the other variables would have a high R^2 but still be quite uninformative. A more interesting regression would create the dependent variable profits/assets and use as the independent variables employees/assets, sales/assets, debt/assets. The regression model is

$$\frac{PROFIT_i}{ASSETS_i} = \beta_0 + \beta_1 \frac{EMPLOYEES_i}{ASSETS_i} + \beta_2 \frac{SALES_i}{ASSETS_i} + \beta_3 \frac{DEBT_i}{ASSETS_i} + \beta_4 BOND_i + \varepsilon_i$$

(Model 1)

Observe that BOND, the bond rating, is not a “gross size” variable; there is no need to scale it by dividing by ASSETS.

In Example 1, the scaling might be described in terms of quantities per \$1,000,000 of ASSETS. It might also be reasonable to use SALES as the scaling variable, rather than ASSETS.

For Example 2, suppose that you wanted to explain number of doctors in terms of imports, exports, savings, telephones, newspaper circulation, and inflation rate. The populations give you the best scaling variable. The regression model is

$$\begin{aligned} \frac{DOCTORS_i}{POPNI_i} = & \beta_0 + \beta_1 \frac{IMPORTS_i}{POPNI_i} + \beta_2 \frac{EXPORTS_i}{POPNI_i} + \beta_3 \frac{SAVINGS_i}{POPNI_i} \\ & + \beta_4 \frac{PHONES_i}{POPNI_i} + \beta_5 \frac{PAPERS_i}{POPNI_i} + \beta_6 INFLATE_i + \varepsilon_i \end{aligned} \quad (\text{Model 2})$$

All the ratios used here could be described as “per capita” quantities. The inflation rate is not a “gross size” variable and need not be put on a per capita basis.

An alternate strategy is to take logarithms of all gross size variables. In Example 1, one might use the model

$$\begin{aligned} \log(PROFIT_i) = & \gamma_0 + \gamma_1 \log(ASSETS_i) + \gamma_2 \log(EMPLOYEES_i) + \gamma_3 \log(SALES_i) \\ & + \gamma_4 \log(DEBT_i) + \gamma_5 BOND_i + \varepsilon_i \end{aligned}$$

Of course, the coefficients γ_0 through γ_5 are not simply related to β_0 through β_4 in the original form of the model. Unless the distribution of values of BOND is very unusual, one would not replace it with its logarithm.

Similarly, the logarithm version of model 2 is

$$\begin{aligned} \log(DOCTORS_i) = & \gamma_0 + \gamma_1 \log(POPNI_i) + \gamma_2 \log(IMPORTS_i) + \gamma_3 \log(EXPORTS_i) \\ & + \gamma_4 \log(SAVINGS_i) + \gamma_5 \log(PHONES_i) + \gamma_6 \log(PAPERS_i) + \gamma_7 INFLATE_i + \varepsilon_i \end{aligned}$$

Since INFLATE is not a “gross size” variable, we are not immediately led to taking its logarithm. If this variable has other distributional defects, such as being highly skewed, then we might indeed want its logarithm.

Finally, it should be noted that one does not generally combine these methods. After all, since

$$\log\left(\frac{A}{B}\right) = \log(A) - \log(B)$$

the logarithm makes the ratio a moot issue.

Dividing logarithms, as in $\log(DOCTORS_i)/\log(POPNI_i)$ is not likely to be useful.

One always has the option of doing a “weighted” regression. One can use one of the variables as a weight in doing the regression. The company assets might be used for Example 1 and the populations used for Example 2. The problem with this approach is that the solution will depend overwhelmingly on the large firms (or large countries).

Data cleaning steps

We will describe the operations in terms of the computer program Minitab.

We will assume here that we are working with a spreadsheet. The columns of this spreadsheet will represent variables; each number in a column must be in the same units. The rows of the spreadsheet will represent data points.

As a preliminary step, check each column for basic integrity. Minitab distinguishes columns of two major varieties, ordinary data and text. (There are also minor varieties, including dates.) If a column is labeled C5-T, then Minitab has interpreted this column as text information.

It sometimes happens that a column which is supposed to be numeric ends up as text. What should you do in such a case?

Scan the column to check for odd characters, such as N/A, DK, ?, unk; some people use markers like this to indicate missing or uncertain values. The Minitab missing numeric data code is the asterisk *, and this should be used to replace things like the above. The expression 2 1/2 was intended to represent 2.5 but Minitab can only interpret it as text; this repair is obvious.

If you edit a text column so that all information is interpretable as numeric, Minitab will not instantly recognize the change. Use **Manipulate** \Rightarrow **Change Data Type** \Rightarrow **Text to Numeric**. If you do this to a column that still has text information, the corresponding entries will end up as *, the numeric missing data code.

It sometimes happens that a column given as numeric really represents a nominal categorical variable and you would prefer to use the names. For example, a column might have used 1, 2, 3, 4 to represent single, married, widowed, and divorced. You would prefer the names. Use **Manipulate** \Rightarrow **Code** \Rightarrow **Numeric to Text**. You will be presented with a conversion table which allows you to do this.

The command **Stat** \Rightarrow **Basic Statistics** \Rightarrow **Display Descriptive Statistics** will give you the minimum and maximum of each column. The minimum and maximum values should make sense; unbelievable numbers for the minimum or the maximum could well be data coding errors. This same command will give you the number of missing values, noted as N^* . The count on missing values should make sense.

For many analyses you would prefer to deal with reasonably symmetric values. One of the cures for right-skewness is the taking of logarithms. Here are some general comments about this process:

Base e logarithms are usually preferred because of certain advantages in interpretation. It is still correct, however, to use base 10 logarithms.

Some variables are of the “gross size” variety. The minimum to maximum span runs over several orders of magnitudes. For example, in a data set on countries of the world, the variable POPULATION will run from 10^5 to 10^9 with many countries at the low end of the scale. This variable should be replaced by its logarithm. In a data set on the Fortune 500 companies, the variable REVENUES will run over several orders of magnitude with most companies toward the low end of the scale. This variable should be replaced by its logarithm.

The command **Stat** \Rightarrow **Basic Statistics** \Rightarrow **Display Descriptive Statistics** will allow you to compare the mean and the standard deviation. If a variable which is always (or nearly always) positive has a standard deviation about as large as the mean, or even larger, is certainly positively skewed.

What should you do with data that are skewed but not necessarily of the “gross size” variety? This is a matter of judgment. Generally you prefer to keep variables in their original units. If most of the other variables are to be transformed by logarithms, then maybe you want to transform this one as well.

If the skewed variable is going to be the dependent variable in a regression, then you will almost certainly want to take its logarithm. (If you don’t take the logarithms immediately, you may find expanding residuals on the residual versus fitted plot. Then you’ll have take logarithms anyhow.)

If the variable to be transformed by logarithms has zero or negative values, then taking logarithms in Minitab will result in missing values! This is not usually what you want to do. Pick a value c so that all values of $X + c$ are positive. Then consider $\log(X + c)$.

Logarithms will *not* cure left-skewed data. If X is such a variable and if M is a number larger than the biggest X , then you can consider $\log(M - X)$, provided you can make a sensible interpretation for this.

Logarithms should *not* be applied to binary variables. If a variable has only two values, then the logarithms will also have only two values.

Suppose that we regress Y on other variables, including J . The fitted model will be

$$\hat{Y} = b_0 + \dots + b_J J + \dots$$

The interpretation of b_J is this:

As J increases by 1, there is an associated increase in Y of b_J , while holding all other predictors fixed.

There's an important WARNING.

WARNING: This interpretation should note that b_J is the “effect” of J on Y after adjusting for the presence of all other variables. (In particular, regressing Y on J without any other predictors could produce a very different value of b_J .) Also, this interpretation carries the disclaimer “while holding all other predictors fixed.” Realistically, it may not be possible to change the value of J while leaving the other predictors unchanged.

Now... suppose that Y is really the base- e logarithm of Z , meaning $Y = \log Z$. What's the link between J and Z ? The fitted model is

$$\log \hat{Z} = b_0 + \dots + b_J J + \dots$$

Here the interpretation of b_J is this:

As J increases by 1, there is an associated increase in $\log Z$ of b_J . This means that $\log Z$ changes to $\log Z + b_J$. By exponentiating, we find that $e^{\log Z} = Z$ changes to $e^{\log Z + b_J} = e^{\log Z} e^{b_J} = Z e^{b_J}$. Using the approximation that $e^t \approx 1 + t$ when t is near zero, we find that Z changes (approximately) to $Z(1+b_J)$. This is interpretable as a percent increase. We summarize thus: as J increases by 1, there is an associated proportional increase of b_J in Z .

If, for example, $b_J = 0.03$, then as J increases by 1, the associated increase in Z is 3%.

This next case is encountered only rarely.

Next suppose that Y is *not* the result of a transformation, but that $J = \log R$ is the base- e logarithm of variable R . What's the link between R and Y ? Let's talk about increasing J by 0.01. (The reason why we consider an increase of 0.01 rather than an increase of 1 will be mentioned below.) Certainly we can say this:

The fitted model is $\hat{Y} = b_0 + \dots + b_J \log R + \dots$

As $J = \log R$ increases by 0.01, there is an associated increase in Y of $0.01 b_J$. Saying that J increases by 0.01 is also saying that $\log R$ increases to $\log R + 0.01$. By exponentiating, we find that $e^{\log R} = R$ changes to $e^{\log R + 0.01} = e^{\log R} e^{0.01} = R e^{0.01} \approx R(1+0.01)$, which is a 1% increase in R .

Here's the conclusion: as R increases by 1%, there is an associated increase in Y of $0.01 b_J$.

If, for example, $b_J = 25,400$, then a 1% increase in R is associated with an approximate increase in Y of 254.

We used an increase of 0.01 (rather than 1) to exploit the approximation $e^{0.01} \approx 1.01$.

Finally, suppose that both Y and J are obtained by taking logs. That is $Y = \log Z$ and $J = \log R$. What is the link between R and Z ? Suppose we consider J increasing by 0.01; as in the previous note, this is approximately a 1% change in R .

As J increases by 0.01, there is an associated change from Y to $Y + 0.01 b_J$. As $Y = \log Z$, we see that Z changes (approximately) to $Z(1+0.01 b_J)$. Thus: as R increases by 1%, we find that there is an associated change in Z of $0.01 b_J$, interpreted as a percent.

If, for example, $b_J = 1.26$, then a 1% increase in R is associated with an approximate increase of 1.26% in Z .

This document points out an interesting misunderstanding about multiple regression. There can be serious disagreement between

the regression coefficient b_H in the regression $\hat{Y} = b_0 + b_G X_G + b_H H$
 and
 the regression coefficient b_H in the regression $\hat{Y} = b_0 + b_H H$

While most people would not expect the values of b_H to be identical in these two regressions, it is somewhat shocking as to how far apart they can be.

Consider this very simple set of data with $n = 20$:

<i>G</i>	<i>H</i>	<i>Y</i>	<i>G</i>	<i>H</i>	<i>Y</i>
73	7.3	3096	80	0.8	3326
87	-6.0	3519	82	-2.4	3365
83	-3.7	3383	77	2.9	3215
78	2.5	3261	81	-1.5	3306
82	-2.2	3360	79	1.1	3266
80	0.7	3334	78	1.9	3229
83	-2.9	3388	76	3.5	3193
86	-6.2	3481	80	0.5	3315
75	5.1	3120	80	-0.3	3280
82	-1.3	3378	81	-0.6	3335

Here is the regression of Y on (G, H) :

The regression equation is
 $Y = -751 + 50.6 G + 20.5 H$

Predictor	Coef	StDev	T	P
Constant	-751.2	515.9	-1.46	0.164
G	50.649	6.439	7.87	0.000
H	20.505	6.449	3.18	0.005

S = 13.63 R-Sq = 98.5% R-Sq(adj) = 98.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	209106	104553	562.64	0.000
Error	17	3159	186		
Total	19	212265			

This shows a highly significant regression. The F statistic is enormous, and the individual t statistics are positive and significant.

Now, suppose that you regressed Y on H only. You'd get the following:

The regression equation is
 $Y = 3306 - 29.7 H$

Predictor	Coef	StDev	T	P
Constant	3306.31	6.38	518.17	0.000
H	-29.708	1.907	-15.58	0.000

S = 28.53 R-Sq = 93.1% R-Sq(adj) = 92.7%

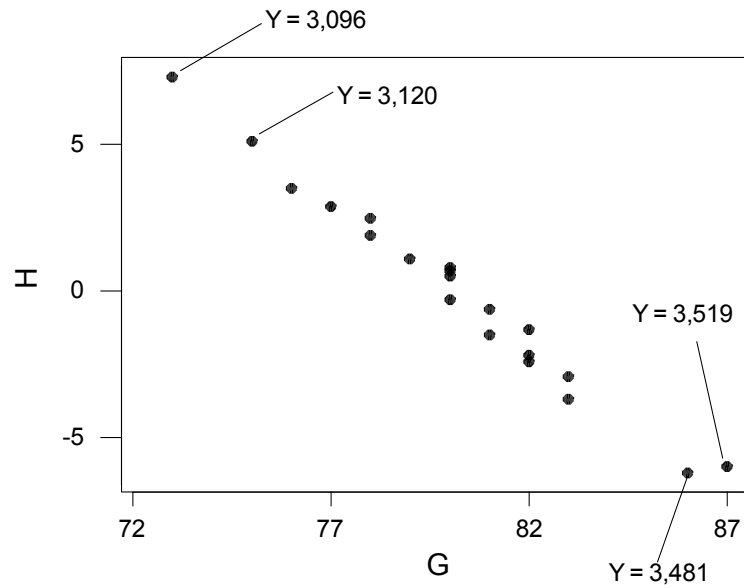
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	197610	197610	242.71	0.000
Error	18	14655	814		
Total	19	212265			

This regression is also highly significant. However, it now happens that the relationship with H is significantly *negative*.

How could this possibly happen? It turns out that these data were strung out in the (G, H) plane with a negative relationship. The coefficient of Y on G was somewhat larger than the coefficient on H , so that when we look at Y and H alone we see a negative relationship.

The picture below shows the locations of the points in the (G, H) plane. The values of Y are shown at some extreme points, suggesting why the apparent relationship between Y and H appears to be negative.



The quantity $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$ measures variation in Y . Indeed we get s_y from this as $s_y = \sqrt{\frac{S_{yy}}{n-1}}$. We use the symbol \hat{y}_i to denote the fitted value for point i .

One can show that $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$. These sums have the names SS_{total} , $SS_{\text{regression}}$, and SS_{error} . They have other names or abbreviations. For instance

SS_{total} may be written as SS_{tot} .

$SS_{\text{regression}}$ may be written as SS_{reg} , SS_{fit} , or SS_{model} .

SS_{error} may be written as SS_{err} , SS_{residual} , SS_{resid} , or SS_{res} .

The degrees of freedom accounting is this:

SS_{total} has $n - 1$ degrees of freedom

$SS_{\text{regression}}$ has K degrees of freedom (K is the number of independent variables)

SS_{error} has $n - 1 - K$ degrees of freedom

Here is how the quantities would be laid out in an analysis of variance table:

Source of Variation	Degrees of freedom	Sum of Squares	Mean Squares	F
Regression	K	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{K}$	$\frac{MS_{\text{Regression}}}{MS_{\text{Error}}}$
Error	$n - 1 - K$	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 1 - K}$	
Total	$n - 1$	$\sum_{i=1}^n (y_i - \bar{y})^2$		

A measure of quality of the regression is the F statistic. Formally, this F statistic tests

$$H_0 : \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \dots, \beta_K = 0 \quad [\text{Note that } \beta_0 \text{ does not appear.}]$$

versus

$$H_1 : \text{at least one of } \beta_1, \beta_2, \beta_3, \dots, \beta_K \text{ is not zero}$$

Note that β_0 is not involved in this test.

Also, note that $s_\epsilon = \sqrt{MS_{\text{Error}}}$ is the estimate of σ_ϵ . This has many names:

- standard error of estimate
- standard error of regression
- estimated noise standard deviation
- root mean square error (RMS error)
- root mean square residual (RMS residual)

The measure called R^2 is computed as $\frac{SS_{\text{Regression}}}{SS_{\text{Total}}}$. This is often described as the “fraction of the variation in Y explained by the regression.”

You can show, by the way, that

$$\frac{s_\epsilon}{s_y} = \sqrt{\frac{n-1}{n-1-K}(1-R^2)}$$

The quantity $R_{adj}^2 = 1 - \frac{n-1}{n-1-K}(1-R^2)$ is called the *adjusted R-squared*. This is supposed to adjust the value of R^2 to account for both the sample size and the number of predictors. With a little simple arithmetic,

$$R_{adj}^2 = 1 - \left(\frac{s_\epsilon}{s_y}\right)^2$$



1 1 1 1 1 1 1 1 1 1 1 1 1 1 INDICATOR VARIABLES 1 1 1 1 1 1 1 1 1 1 1 1

In each row of the data sheet, SL + RANCH + COLONIAL + TUDOR will be exactly 1. This just notes that each house is one, and only one, of the four styles.

The command **Calc** ⇒ **Make Indicator Variables** can be applied to a column of alphabetic information. If STYLE contained alphabetic values, say

Split-Level Ranch Colonial Tudor

then Minitab would assign

Colonial to the first-named variable in the **Store results in:** list

Ranch to the second-named variable in the **Store results in:** list

Split-Level to the third-named variable in the **Store results in:** list

Tudor to the fourth-named variable in the **Store results in:** list

This is done based on alphabetical ordering. Again, you need to be careful. You still have the option of listing a set of columns as C11-C14. After you see the results, you can assign names to these columns.

It seems natural now to run the regression of PRICE on (SL, RANCH, COLONIAL, TUDOR, SIZE, BEDROOM). Note that STYLE is not included.

If you do that, you'll get this message at the top of the Minitab run:

* TUDOR is highly correlated with other X variables
* TUDOR has been removed from the equation

This message happens because $SL + RANCH + COLONIAL + TUDOR = 1$ for every line of the data set. This creates total collinearity with the regression intercept, and the regression arithmetic is impossible. Minitab deals with this by removing the last-named variable involved. In this instance, TUDOR was named last and was eliminated.

Minitab then goes on to produce a useful regression run:

The regression equation is
PRICE = 114696 + 21.8 SIZE + 2682 BEDROOM - 21054 SL - 12504 RANCH
- 12639 COLONIAL

Predictor	Coef	SE Coef	T	P
Constant	114696	4160	27.57	0.000
SIZE	21.832	1.993	10.96	0.000
BEDROOM	2682	1006	2.66	0.008
SL	-21054	1871	-11.26	0.000
RANCH	-12504	1821	-6.86	0.000
COLONIAL	-12639	1705	-7.41	0.000

S = 9882 R-Sq = 56.5% R-Sq(adj) = 55.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	46184185424	9236837085	94.58	0.000
Residual Error	364	35546964282	97656495		
Total	369	81731149706			

Parts of the output have been omitted.

The question now is the interpretation of the coefficients. For a split-level home, the indicators have values $SL = 1$, $RANCH = 0$, $COLONIAL = 0$. (Note that TUDOR has been omitted by Minitab). The fitted equation for a split-level home is then

$$PRICE = 114696 + 21.8 \text{ SIZE} + 2682 \text{ BEDROOM} - 21054 \quad \text{Split-Level}$$

A ranch home has indicators $SL = 0$, $RANCH = 1$, $COLONIAL = 0$. This gives the fitted equation

$$PRICE = 114696 + 21.8 \text{ SIZE} + 2682 \text{ BEDROOM} - 12504 \quad \text{Ranch}$$

Similarly, the fitted equation for colonial homes is

$$PRICE = 114696 + 21.8 \text{ SIZE} + 2682 \text{ BEDROOM} - 12639 \quad \text{Colonial}$$

What about the Tudor homes? These have $SL = 0$, $RANCH = 0$, $COLONIAL = 0$, so that the fitted equation for these is

$$PRICE = 114696 + 21.8 \text{ SIZE} + 2682 \text{ BEDROOM} \quad \text{Tudor}$$

The omitted indicator, here TUDOR, gives the base for interpreting the other estimated coefficients.

The suggestion is that a split-level home sells for 21,054 less than a Tudor home, holding all other variables fixed. A ranch sells for 12,504 less than a Tudor home, holding all other variables fixed. It follows that a ranch sells for $21,054 - 12,504 = 8,550$ more than a split-level, holding all other variables fixed.

If we had asked Minitab for the regression of PRICE on (SL, RANCH, TUDOR, SIZE, BEDROOM), we would have produced the following fitted equation:

$$PRICE = 102057 + 21.8 \text{ SIZE} + 2682 \text{ BEDROOM} - 8415 \text{ SL} + 135 \text{ RANCH} + 12639 \text{ TUDOR}$$

This time the indicator for colonial was used as the baseline, and we see that the Tudor homes sell for 12,639 more than the colonial homes, holding all else fixed. Perfectly consistent.

The following display indicates exactly what happens as we change the baseline.

Indicators used in the regression	Estimated coefficients			
	SL	RANCH	COLONIAL	TUDOR
SL, RANCH, COLONIAL	-21,054	-12,504	-12,639	
SL, RANCH, TUDOR	-8,415	135		12,639
SL, COLONIAL, TUDOR	-8,550		-135	12,504
RANCH, COLONIAL, TUDOR		8,550	8,415	21,054

In all parts of this table, the other variables (SIZE, BEDROOM) were used as well.

All four lines of this table represent equivalent fits. All produce the same R^2 , the same F statistic, and the same s_e (S in Minitab). Moreover, the estimated coefficients on SIZE and BEDROOM will be the same in all four lines, as will the corresponding t statistics.

If you are using a set of indicator variables, and if you go through a variable-selection process to remove variables, you must keep the indicator set intact. In the context of this problem, that means that any fitted model must use either
 three out of the four indicators

or

none of the indicators

The indicators only make solid good sense when used together.

The regression of PRICE on (SIZE, BEDROOM, SL, RANCH, COLONIAL) which we saw above had significant t statistics on all independent variables. We would not be tempted to remove any of them. Moreover, a stepwise regression would select all the predictors.

The regression of PRICE on (SIZE, BEDROOM, SL, RANCH, TUDOR) produces this:

The regression equation is
 PRICE = 102057 + 21.8 SIZE + 2682 BEDROOM - 8415 SL + 135 RANCH
 + 12639 TUDOR

Predictor	Coef	SE Coef	T	P
Constant	102057	3674	27.78	0.000
SIZE	21.832	1.993	10.96	0.000
BEDROOM	2682	1006	2.66	0.008
SL	-8415	1365	-6.16	0.000
RANCH	135	1309	0.10	0.918
TUDOR	12639	1705	7.41	0.000

This suggests that we might remove the indicator for RANCH. Indeed, stepwise regression selects all the variables except RANCH.

Finally, we need an objective method to test whether an indicator variable set should be used at all. Let's consider the context of our model, namely

$$\begin{aligned} \text{PRICE}_i = & \beta_0 + \beta_{\text{SIZE}} \text{SIZE}_i + \beta_{\text{BEDROOM}} \text{BEDROOM}_i \\ & + \beta_{\text{SL}} \text{SL}_i + \beta_{\text{RANCH}} \text{RANCH}_i + \beta_{\text{COLONIAL}} \text{COLONIAL}_i + \varepsilon_i \end{aligned}$$

The decision about whether or not to use the style indicators is really a test of the null hypothesis $H_0: \beta_{\text{SL}} = 0, \beta_{\text{RANCH}} = 0, \beta_{\text{COLONIAL}} = 0$.

There is a method for testing whether a *set* of coefficients is all zero. This method works for situations beyond what we are testing here. This requires the computation of this F statistic:

$$\frac{\left\{ \begin{array}{l} \text{Regression sum of squares} \\ \text{using SIZE, BEDROOM,} \\ \text{SL, RANCH, COLONIAL} \end{array} \right\} - \left\{ \begin{array}{l} \text{Regression Sum of Squares} \\ \text{using SIZE, BEDROOM} \end{array} \right\}}{\left\{ \begin{array}{l} \text{Residual Mean Square} \\ \text{using SIZE, BEDROOM,} \\ \text{SL, RANCH, COLONIAL} \end{array} \right\}} \div \left\{ \begin{array}{l} \text{Number of coefficients} \\ \text{being investigated} \end{array} \right\}$$

This is to be interpreted as an F statistic. We need to identify the two degrees of freedom numbers associated with F .

The numerator degrees of freedom is “Number of coefficients being investigated” in the calculation above.

The denominator degrees of freedom is the DF for residual in the regression on (SIZE, BEDROOM, SL, RANCH, COLONIAL).

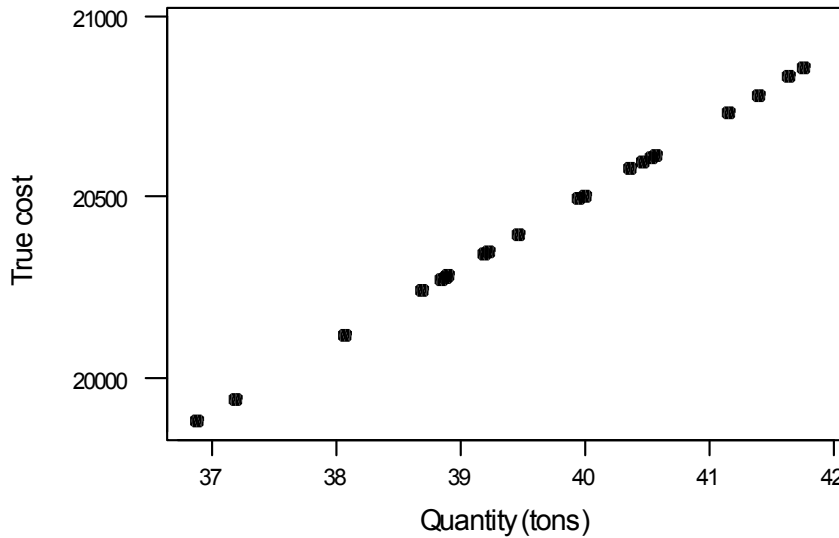
The regression on (SIZE, BEDROOM, SL, RANCH, COLONIAL) had this analysis of variance table:

Analysis of Variance				
Source	DF	SS	MS	F
P				
Regression	5	46184185424	9236837085	94.58
0.000				
Residual Error	364	35546964282	97656495	
Total	369	81731149706		

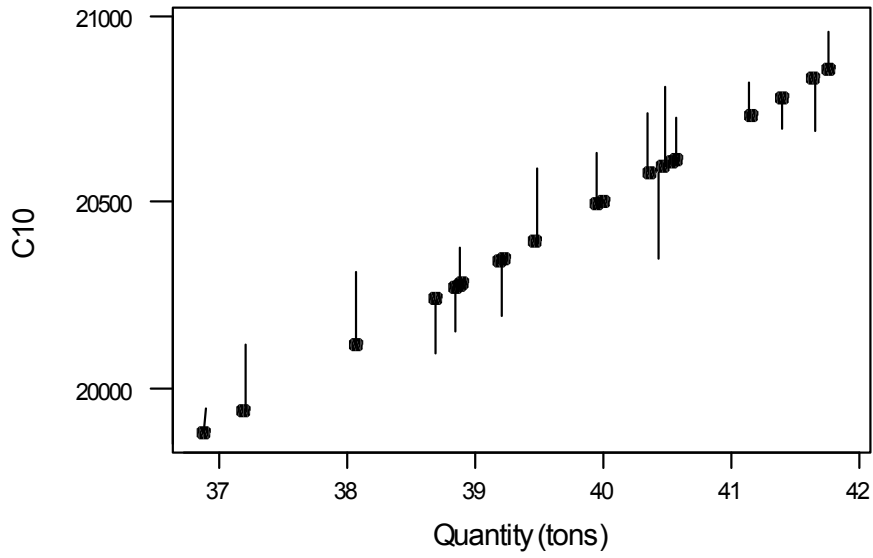
The regression sum of squares is 46,184,185,424. The residual mean square is 97,656,495. We note also that the degrees of freedom in the residual line is 364.

There are many contexts in which regression analysis is used to estimate fixed and variable costs for complicated processes. The following data set involves the quantities produced and the costs for the production of a livestock food mix for each of 20 days. The quantities produced were measured in the obvious way, and the costs were calculated directly as labor costs + raw material costs + lighting + heating + equipment costs. The equipment costs were computed by amortizing purchase costs over the useful lifetimes, and the other costs are reasonably straightforward.

In fact, the actual fixed cost (per day) was \$12,500, and the variable cost was \$200/ton. Thus the exact relationship we see should be $\text{Cost} = \$12,500 + 200 \frac{\$}{\text{ton}} \times \text{Quantity}$. Here is a picture of this exact relationship:



It happens, however, that there is statistical noise in assessing cost, and this noise has a standard deviation of \$100. Schematically, we can think of our original picture as being spread out with vertical noise:



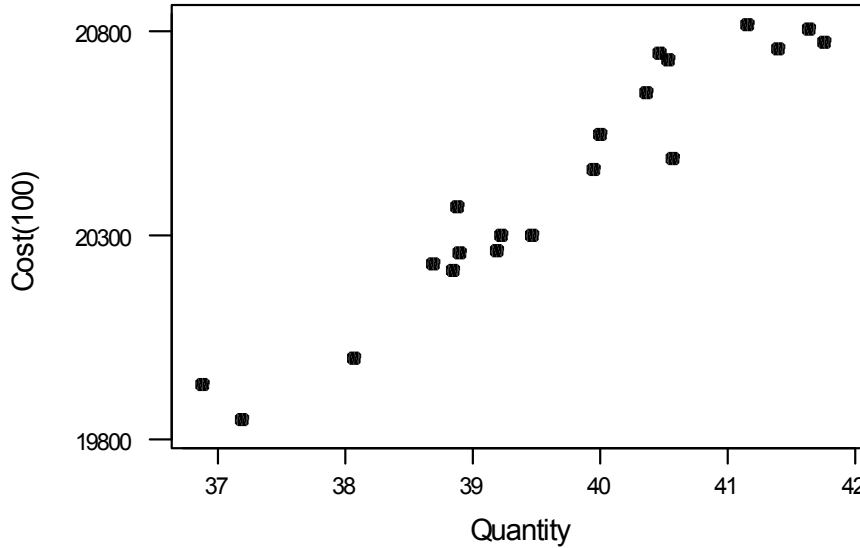
Here then are the data which we actually see:

Quantity	Cost	Quantity	Cost
41.66	20812.70	39.22	20302.30
40.54	20734.90	41.78	20776.70
38.90	20258.70	38.88	20373.00
38.69	20232.40	38.84	20213.70
40.58	20493.40	37.18	19848.70
40.48	20750.30	41.16	20818.90
36.88	19932.80	39.19	20265.10
39.47	20303.70	40.38	20654.50
41.41	20760.30	40.01	20553.00
38.07	20002.20	39.96	20463.10

The quantities are in tons, and the costs are in dollars.

Here is a scatterplot for the actual data:

Costs in dollars to produce feed quantities in tons



(There is a noise standard deviation of \$100 in computing costs.)

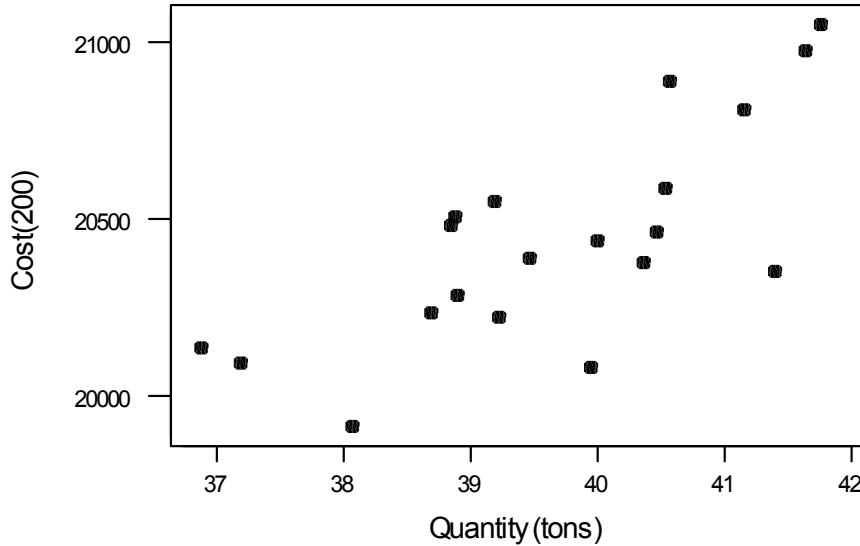
The footnote shows that in the process of assessing costs, there is noise with a standard deviation of \$100. In spite of this noise, the picture is fairly clean. The fitted regression line is $\widehat{Cost} = \$12,088 + 210 \frac{\$}{\text{ton}} \times \text{Quantity}$. The value of R^2 is 92.7%, so we know that this is a good regression. We would assess the daily fixed cost at \$12,088, and we would assess the variable cost at \$210/ton. Please bear in mind that this discussion hinges on knowing the exact fixed and variable costs and knowing about the \$100 noise standard deviation; in other words, this is a simulation in which we really know the facts. An analyst who sees only these data would not know the exact answer. Of course, the analyst would compute $s_e = \$83.74$, so that

Quantity	True value	Value estimated from data
Fixed cost	\$12,500	$b_0 = \$12,088$
Variable cost	\$200/ton	$b_1 = \$210/\text{ton}$
Noise standard deviation	\$100	$s_e = \$83.74$

All in all, this is not bad.

As an extension of this hypothetical exercise, we might ask how the data would behave with a \$200 standard deviation associated with assessing costs. Here is that scatterplot:

Cost in dollars to produce feed quantities in tons



(There is a noise standard deviation of \$200 in computing costs.)

For this scatterplot, the fitted regression equation is $\hat{C}ost = \$13,910 + 165 \frac{\$}{ton} \times Quantity$. Also for this regression we have $R^2 = 55.4\%$. Our estimates of fixed and variable costs are still statistically unbiased, but they are infected with more noise. Thus, our fixed cost estimate of \$13,910 and our variable cost estimate of $165 \frac{\$}{ton}$ are not all that good. Of course, one can overcome the larger standard deviation in computing the cost by taking more data. For this problem, the analyst would see $s_e = \$210.10$.

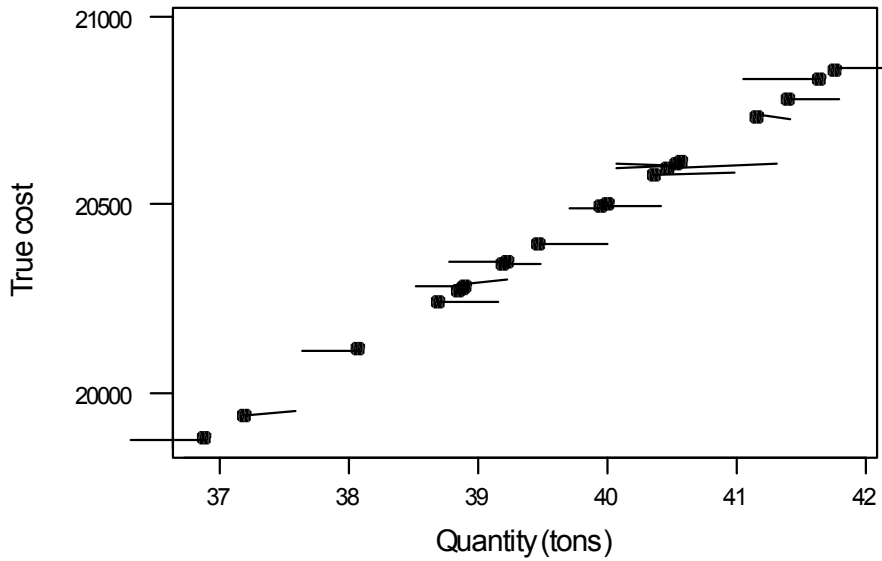
Quantity	True value	Value estimated from data
Fixed cost	\$12,500	$b_0 = \$13,910$
Variable cost	\$200/ton	$b_1 = \$165/ton$
Noise standard deviation	\$200	$s_e = \$210.10$

This is not nearly as good as the above, but this may be more typical.

It is important to note that noise in assessing cost, the vertical variable, still gives us a statistically valid procedure. The uncertainty can be overcome with a larger sample size.

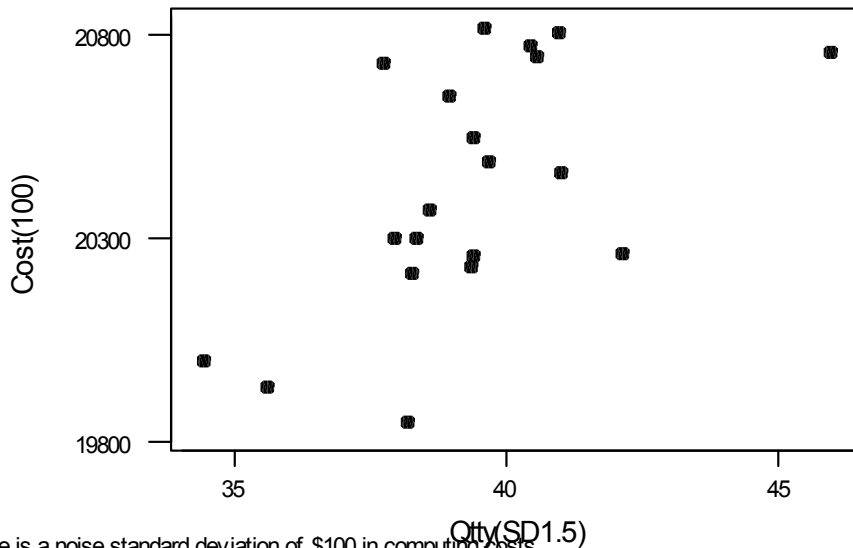
We will now make a distinction between noise in the vertical direction (noise in computing cost) and noise in the horizontal direction (noise in measuring quantity).

A more serious problem occurs when the horizontal variable, here quantity produced, is not measured exactly. It is certainly plausible that one might make such measuring errors when dealing with merchandise such as livestock feed. For these data, the set of 20 quantities has a standard deviation of 1.39 tons. This schematic illustrates the notion that our quantities, the horizontal variable, might not be measured precisely:



Here is a picture showing the hypothetical situation in which costs experienced a standard deviation of measurement of \$200 while the feed quantities had a standard deviation of measurement of 1.5 tons.

Cost in dollars to produce feed quantities in tons



(There is a noise standard deviation of \$100 in computing costs and quantities have been measured with a SD of 1.5 tons.)

For this picture the relationship is much less convincing. In fact, the fitted regression equation is $\hat{C}ost = \$17,511 + 74.2 \frac{\$}{ton} \times Quantity$. Also, this has $s_\epsilon = \$252.60$. This has not helped:

Quantity	True value	Value estimated from data
Fixed cost	\$12,500	$b_0 = \$17,511$
Variable cost	\$200/ton	$b_1 = \$74.20/ton$
Noise standard deviation	\$200	$s_\epsilon = \$252.60$

The value of R^2 here is 34.0%, which suggests that the fit is not good.

Clearly, we would like both cost and quantity to be assessed perfectly. However,

noise in measuring costs leaves our procedure valid (unbiased) but with imprecision that can be overcome with large sample sizes

noise in measuring quantities makes our procedure biased

The data do not generally provide clues as to the situation.

Here then is a summary of our situation.

Suppose that the relationship is

$$\text{True cost} = \beta_0 + \beta_1 \times \text{True quantity}$$

where β_0 is the fixed cost and β_1 is the variable cost

Suppose that we observe

$$Y = \text{True cost} + \varepsilon$$

where ε represents the noise in measuring or assessing the cost, with standard deviation σ_ε

and

$$x = \text{True quantity} + \zeta$$

where ζ represents the noise in measuring or assessing the quantity, with standard deviation σ_ζ

Let us also suppose that the True quantities themselves are drawn from a population with mean μ_x and standard deviation σ_x .

You will do least squares to find the fitted line $\hat{Y} = b_0 + b_1 x$.

It happens that b_1 , the sample version of the variable cost, estimates $\beta_1 \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\zeta^2}$.

Of course, if $\sigma_\zeta = 0$ (no measuring error in the quantities), then b_1 estimates β_1 . It is important to observe that if $\sigma_\zeta > 0$, then b_1 is biased closer to zero.

It happens that b_0 , the sample version of the fixed cost, estimates

$$\beta_0 + \beta_1 \mu_x \frac{\sigma_\zeta^2}{\sigma_x^2 + \sigma_\zeta^2}.$$

If $\sigma_\zeta = 0$, then b_0 correctly estimates the fixed cost β_0 .

The impact in accounting problems is that we will tend to *underestimate* the variable cost and *overestimate* the fixed cost.

You can see that the critical ratio here is $\frac{\sigma_\zeta^2}{\sigma_x^2}$, the ratio of the variance of the noise in x relative to the variance of the population from which the x 's are drawn.

In the real situation, you've got one set of data, you have no idea about the values of β_0 , β_1 , σ_x , σ_ζ , or σ_ε . If you have a large value of R^2 , say over 90%, then you can be pretty sure that b_1 and b_0 are useful as estimates of β_1 and β_0 . If the value of R^2 is not large, you simply do not know whether to attribute this to a large σ_ε , to a large σ_ζ , or to both.

	Small σ_ζ / σ_x (quantity measured precisely relative to its background variation)	Large σ_ζ / σ_x (quantity measured imprecisely relative to its background variation)
Small σ_ε (cost measured precisely)	b_0 and b_1 nearly unbiased with their own standard deviations low; R^2 will be large	b_1 seriously biased downward and b_0 seriously biased upward; R^2 will not be large
Large σ_ε (cost measured imprecisely)	b_0 and b_1 nearly unbiased but their own standard deviations may be large; R^2 will not be large	b_1 seriously biased downward and b_0 seriously biased upward; R^2 will not be large

Do you have any recourse here?

If you know or suspect that σ_ε will be large, meaning poor precision is assessing costs, you can simply recommend a larger sample size.

If you know or suspect that σ_ζ will be large relative to σ_x , there are two possible actions:

By obtaining multiple readings of x for a single true quantity, it may be possible to estimate σ_ζ and thus undo the bias. You will need to obtain the services of a serious statistical expert, and he or she should certainly be well paid.

You can spread out the x -values so as to enlarge σ_x (presumably without altering the value of σ_ζ). In the situation of our animal feed example, it may be procedurally impossible to do this.

CROSS REF LIST (NOT FOR DISTRIBUTION)

Original documents (not part of the formal handout)

introthoughts.doc

grossSize.doc

DataCleaning.doc

coefintr.doc

regpath.doc

anova.doc

indicator.doc (has a few things on variable selection)

errinx.doc