SIMPLE LINEAR REGRESSION

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

Cover photo: Montauk lighthouse

Revised 4 AUG 2004

Consider a set of 10 data points:

x:	1	2	3	4	4	5	5	6	6	7
Y:	7	8	9	8	9	11	10	13	14	13

Begin by finding $\Sigma x_i = 1 + 2 + 3 + \ldots + 7 = 43$

$\Sigma x_i^2 = 1^2 + 2^2 + 3^2 + \ldots + 7^2 = 217$

$\Sigma y_i = 7 + 8 + 9 + \ldots + 13 = 102$

$\Sigma y_i^2 = 7^2 + 8^2 + 9^2 + \ldots + 13^2 = 1{,}094$

$\Sigma x_i y_i = 1\times7 + 2\times8 + 3\times9 + \ldots + 7\times13 = 476$

Then find $\bar{x} = \dfrac{43}{10} = 4.3$ and $\bar{y} = \dfrac{102}{10} = 10.2$.

Next $S_{xx} = \Sigma x_i^2 - \dfrac{(\Sigma x_i)^2}{n} = 217 - \dfrac{43^2}{10} = 32.1$

$S_{xy} = \Sigma x_i y_i - \dfrac{(\Sigma x_i)(\Sigma y_i)}{n} = 476 - \dfrac{43 \cdot 102}{10} = 37.4$

$S_{yy} = \Sigma y_i^2 - \dfrac{(\Sigma y_i)^2}{n} = 1{,}094 - \dfrac{102^2}{10} = 53.6$

This leads to $b_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{37.4}{32.1} \approx 1.1651$ and then $b_0 = \bar{y} - b_1 \bar{x} = 10.2 - 1.1651 \times 4.3$ ≈ 5.1900.

The regression line can be reported as $\hat{Y} = 5.1900 + 1.1651\, x$. If the spurious precision annoys you, report the line instead as $\hat{Y} = 5.19 + 1.17\, x$.

The quantity S_{yy} was not used here. It has many other uses in regression calculations, so it is worth the trouble to find it early in the work.

Here are the calculations needed to do a simple regression.

> Aside: The word *simple* here refers to the use of just one x to predict y. Problems in which two or more variables are used to predict y are called *multiple regression*.

The input data are (x_1, y_1), (x_2, y_2), …, (x_n, y_n). The outputs in which we are intereseted (so far) are the values of b_1 (estimated regression slope) and b_0 (estimated regression intercept). These will allow us to write the fitted regression line $\hat{Y} = b_0 + b_1 x$.

(1) Find the five sums $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} y_i^2$, $\sum_{i=1}^{n} x_i y_i$.

(2) Find the five expressions $\bar{x}$, $\bar{y}$, $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$,

$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$, $S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$.

(3) Give the slope estimate as $b_1 = \frac{S_{xy}}{S_{xx}}$ and the intercept estimate as $b_0 = \bar{y} - b_1 \bar{x}$.

(4) For later use, record $S_{yy|x} = S_{yy} - \frac{\left(S_{xy}\right)^2}{S_{xx}}$.

Virtually all the calculations for simple regression are based on the five quantities found in step (2). The regression fitting procedure is known as *least squares*. It gets this name because the resulting values of b_0 and b_1 minimize the expression $\sum_{i=1}^{n} \left(y_i - \left[b_0 + b_1 x_i\right]\right)^2$. This is a good criterion to optimize for many reasons, but understanding these reasons will force us to go into the regression model.

As an example, consider a data set with $n = 10$ and with

$$\Sigma\, x_i = 200 \qquad \Sigma\, x_i^2 = 4{,}250 \qquad \Sigma\, y_i = 1{,}000 \qquad \Sigma\, y_i^2 = 106{,}250$$

$$\Sigma\, x_i\, y_i = 20{,}750$$

It follows that

$$\bar{x} = \frac{200}{10} = 20 \qquad\qquad \bar{y} = \frac{1{,}000}{10} = 100$$

$$S_{xx} = 4{,}250 - \frac{200^2}{10} = 250$$

$$S_{xy} = 20{,}750 - \frac{200 \times 1{,}000}{10} = 750$$

It follows next that $b_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{750}{250} = 3$ and $b_0 = \bar{y} - b_1\, \bar{x} = 100 - 3(20) = 40$.

The fitted regression line would be given as $\hat{Y} = 40 + 3\, x$.

We could note also $S_{yy} = 106{,}250 - \dfrac{1{,}000^2}{10} = 6{,}250$. Then $S_{yy|x} = 6{,}250 - \dfrac{750^2}{250} = 4{,}000$.

We use $S_{yy|x}$ to get s_ε, the estimate of the noise standard deviation. The relationship is $s_\varepsilon = \sqrt{\dfrac{S_{yy|x}}{n-2}}$, and here that value is $\sqrt{\dfrac{4{,}000}{10-2}} = \sqrt{500} \approx 22.36$.

In fact, we can use these simple quantities to compute the regression analysis of variance table. The table is built on the identity

$$SS_{total} = SS_{regression} + SS_{residual}$$

The quantity $SS_{residual}$ is often named SS_{error}.

> The subscripts are often abbreviated. Thus, you will see reference to SS_{tot} , SS_{regr} , SS_{resid} , and SS_{err} .

For the simple regression case, these are computed as

$$SS_{tot} = S_{yy}$$

$$SS_{regr} = \frac{\left(S_{xy}\right)^2}{S_{xx}}$$

$$SS_{resid} = S_{yy} - \frac{\left(S_{xy}\right)^2}{S_{xx}}$$

The analysis of variance table for simple regression is set up as follows:

Source of Variation	Degrees of freedom	Sum of Squares	Mean Squares	F
Regression	1	$\frac{\left(S_{xy}\right)^2}{S_{xx}}$	$\frac{\left(S_{xy}\right)^2}{S_{xx}}$	$\frac{MS_{Regression}}{MS_{Resid}}$
Residual	n - 2	$S_{yy} - \frac{\left(S_{xy}\right)^2}{S_{xx}}$	$\frac{S_{yy} - \frac{\left(S_{xy}\right)^2}{S_{xx}}}{n-2}$	
Total	n - 1	S_{yy}		

For the data set used here, the analysis of variance table would be

Source of Variation	Degrees of freedom	Sum of Squares	Mean Squares	F
Regression	1	2,250	2,250	4.50
Residual	8	4,000	500	
Total	9	6,250		

Just for the record, let's note some other computations commonly done for regression. The information given next applies to regressions with K predictors. To see the forms for simple regression, just use $K = 1$ as needed.

The estimate for the noise standard deviation is the square root of the mean square in the residual line. This is $\sqrt{500} \approx 22.36$, as noted previously. The symbol s is frequently used for this, as are $s_{Y|X}$ and s_ε.

The R^2 statistic is the ration $\frac{SS_{regr}}{SS_{tot}}$, which is here $\frac{2,250}{6,250} = 0.36$.

The standard deviation of Y can be given as $\sqrt{\frac{SS_{tot}}{n-1}}$, which is here $\sqrt{\frac{6,250}{9}}$ $\approx \sqrt{694.4444} \approx 26.35$.

It is sometimes interesting to compare s_ε (the estimate for the noise standard deviation) to s_Y (the standard deviation of Y). It can be shown that the ratio of these is

$$\frac{s_\varepsilon}{s_Y} = \sqrt{\frac{n-1}{n-1-K}\left(1-R^2\right)}$$

The quantity $1 - \left(\frac{s_\varepsilon}{s_Y}\right)^2 = 1 - \frac{n-1}{n-1-K}\left(1-R^2\right)$ is called the *adjusted* R^2 statistic, R^2_{adj}.

The following data are found in the file X:\SOR\B011305\M\WOLVES.MTP:

TLength	Weight
10	79
13	72
19	100
19	116
20	85
20	88
23	100
24	80
25	160
27	120

These refer to the tail lengths (in inches) and the weights (in pounds) of 10 wolves. The idea is predict weight from tail lengths. Here are some useful summaries:

Descriptive Statistics

Variable	N	Mean	Median	Tr Mean	StDev	SE Mean
TLength	10	20.00	20.00	20.37	5.27	1.67
Weight	10	100.00	94.00	96.00	26.35	8.33

Variable	Min	Max	Q1	Q3
TLength	10.00	27.00	17.50	24.25
Weight	72.00	160.00	79.75	117.00

Correlations (Pearson)

Correlation of TLength and Weight = 0.600

Here are the results of a regression request:

Regression Analysis: Weight versus TLength

The regression equation is
Weight = 40.0 + 3.00 TLength

Predictor	Coef	SE Coef	T	P
Constant	40.00	29.15	1.37	0.207
TLength	3.000	1.414	2.12	0.067

S = 22.36 R-Sq = 36.0% R-Sq(adj) = 28.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	2250.0	2250.0	4.50	0.067
Residual Error	8	4000.0	500.0		
Total	9	6250.0			

Unusual Observations

Obs	TLength	Weight	Fit	SE Fit	Residual	St Resid
9	25.0	160.00	115.00	10.00	45.00	2.25R

R denotes an observation with a large standardized residual

We must, of course, examine scatterplots.

Formally, the regression activity is using the model $WEIGHT_i = \beta_0 + \beta_1\ TLENGTH_i + \varepsilon_i$, where $i = 1, 2, \ldots, 10$, where β_0 and β_1 are unknown parameters, and where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{10}$ are statistical noise terms. It is assumed that the noise terms are independent with mean 0 and unknown standard deviation σ.

The *fitted* regression equation is that obtained from the computer output. Namely, it's $\hat{WEIGHT} = 40 + 3\ TLENGTH$. Here $b_0 = 40$ is the estimate of β_0, and $b_1 = 3$ is the estimate of β_1. (We sometimes replace the symbols b_0 and b_1 by $\hat{\beta}_0$ and $\hat{\beta}_1$.)

If you wish to check the computational formulas, use x for the Tlength variable and use y for the Weight variable. Then, it happens that

$$\Sigma\, x_i = 200 \qquad \Sigma\, x_i^2 = 4{,}250 \qquad \Sigma\, y_i = 1{,}000 \qquad \Sigma\, y_i^2 = 106{,}250 \qquad \Sigma\, x_i\, y_i = 20{,}750$$

It follows that

$$\bar{x} = \frac{200}{10} = 20 \qquad\qquad \bar{y} = \frac{1{,}000}{10} = 100$$

$$S_{xx} = 4{,}250 - \frac{200^2}{10} = 250$$

$$S_{xy} = 20{,}750 - \frac{200 \cdot 1{,}000}{10} = 750$$

It follows then that $b_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{750}{250} = 3$ and $b_0 = \bar{y} - b_1\, \bar{x} = 100 - 3(20) = 40$.

The data below give the suds height in millimeters as a function of grams of soap used in a standard dishpan.

SOAP	SUDS
3.5	24.4
4.0	32.1
4.5	37.1
5.0	40.4
5.5	43.3
6.0	51.4
6.5	61.9
7.0	66.1
7.5	77.2
8.0	79.2

Let's fit the ordinary regression model and examine the residuals. You can arrange to have the residuals saved by doing

Stat ⇒ **Regression** ⇒ **Regression** ⇒ **Storage** ⇒
[⊙ **Residuals**
OK ⇒]

Here is the regression output:

Regression Analysis

The regression equation is
SUDS = - 20.2 + 12.4 SOAP

Predictor	Coef	StDev	T	P
Constant	-20.234	3.700	-5.47	0.000
SOAP	12.4424	0.6242	19.93	0.000

S = 2.835 R-Sq = 98.0% R-Sq(adj) = 97.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	3193.0	3193.0	397.32	0.000
Error	8	64.3	8.0		
Total	9	3257.3			

The fitted model is $\widehat{SUDS}$ = -20.234 + 12.4424 SOAP. Using this fitted model we can get the residuals as

e_i = $SUDS_i$ - [-20.234 + 12.4424 $SOAP_i$]

= [Actual SUDS value for point i] - [Retro-fit SUDS value for point i]

For instance, for point 1, this value is 24.4 - [-20.234 + 12.4424(3.5)].

Actually, our **Storage** request to Minitab did the arithmetic. The residuals were left in a new column in Minitab's Data window under the name RESI1. (Residuals from subsequent regressions would have different names, and you also have the option of editing the name RESI1.)

Here are the actual values for this data set:

SOAP	SUDS	RESI1
3.5	24.4	1.08545
4.0	32.1	2.56424
4.5	37.1	1.34303
5.0	40.4	-1.57818
5.5	43.3	-4.89939
6.0	51.4	-3.02061
6.5	61.9	1.25818
7.0	66.1	-0.76303
7.5	77.2	4.11576
8.0	79.2	-0.10545

The data for a Y-on-X regression problem come in the form (x_1, Y_1), (x_2, Y_2), …., (x_n, Y_n). These may be conveniently laid out in a matrix or spreadsheet:

Case	x	Y
1	x_1	Y_1
2	x_2	Y_2
.	.	.
.	.	.
n	x_n	Y_n

The word "case" might be replaced by "point" or "data point" or "sequence number" or might even be completely absent. The labels x and Y could be other names, such as "year" or "sales." In a data file in Minitab, the values for the x's and y's will be actual numbers, rather than algebra symbols. In an Excel spreadsheet, these could be either numbers or implicit values.

If a computer program is asked for the regression of Y on x, then numeric calculations will be done. These calculations have something to say about the regression *model*, which we discuss now.

The most common linear regression model is this.

> The values $x_1, x_2, \ldots, x_n$ are known non-random quantities which are measured without error. If in fact the x values really are random, then we assume that they are fixed once we have observed them. This is a verbal sleight of hand; technically we say we are doing the analysis "conditional on the x's."
>
> The Y-values are independent of each other, and they are related to the x's through the *model equation*
>
> $$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for } i = 1, 2, 3, \ldots, n$$
>
> The symbols β_0 and β_1 in the model equation are nonrandom unknown *parameters.*
>
> The symbols $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are called "statistical noise" or "errors." The ε-values prevent us from seeing the exact linear relationship between x and Y. These ε-values are unobserved random quantities. They are assumed to be statistically independent of each other, and they are assumed to have expected value zero. It is also assumed that (using SD for standard deviation) $SD(\varepsilon_1) = SD(\varepsilon_2) = \ldots = SE(\varepsilon_n) = \sigma_\varepsilon$. The symbol σ_ε is another nonrandom unknown parameter.

The calculations that we will do for a regression will make statements about the model. For example, the estimated regression slope $b_1 = \frac{S_{xy}}{S_{xx}}$ is an *estimate* of the parameter β_1.

Here is a summary of a few regression calculations, along with the statements that they make about the model.

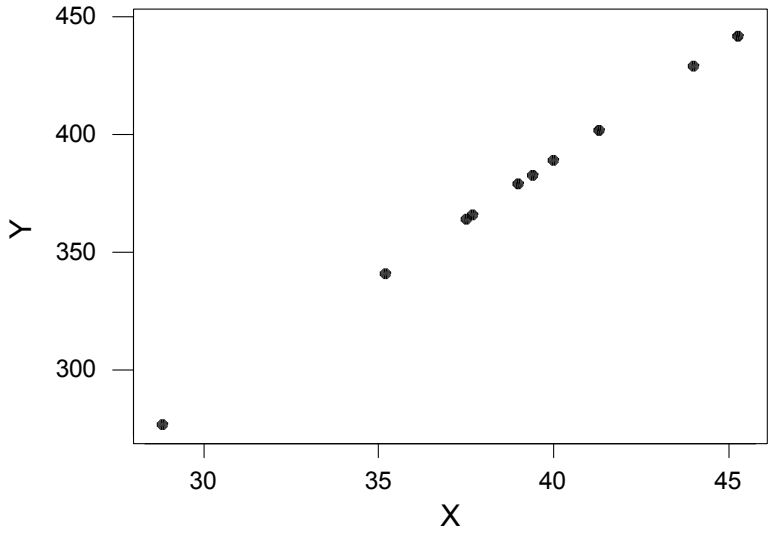
Calculation	What it means
$b_1 = \frac{S_{xy}}{S_{xx}}$ ($\hat{\beta}_1$ used also)	Estimate of regression slope β_1
$b_0 = \bar{y} - b_1 \bar{x}$ ($\hat{\beta}_0$ used also)	Estimate of regression intercept β_0
Residual mean square	Estimate of σ^2
Root mean square residual (standard error of regression)	Estimate of σ
Standard error of an estimated coefficient	Estimate of the standard deviation of that coefficient
t (of an estimated coefficient)	Estimated coefficient, divided by its standard error

The linear regression model with one predictor says that

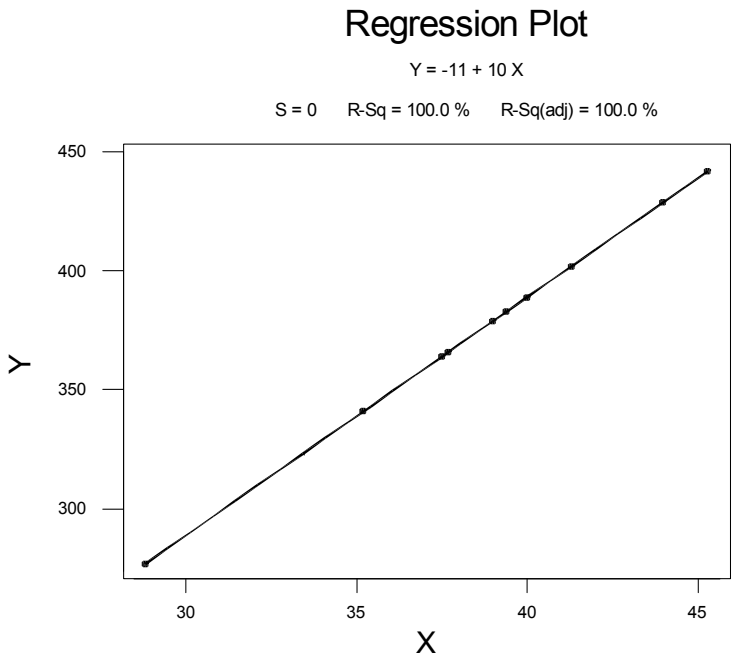
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{for } i = 1, 2, \ldots, n$$

The ε's represent *noise* terms. These are assumed to be drawn from a population with mean 0 and with standard deviation σ.

Let's make the initial observation that if σ = 0, then all the ε's are zero and we should see the line exactly. Here is such a situation:

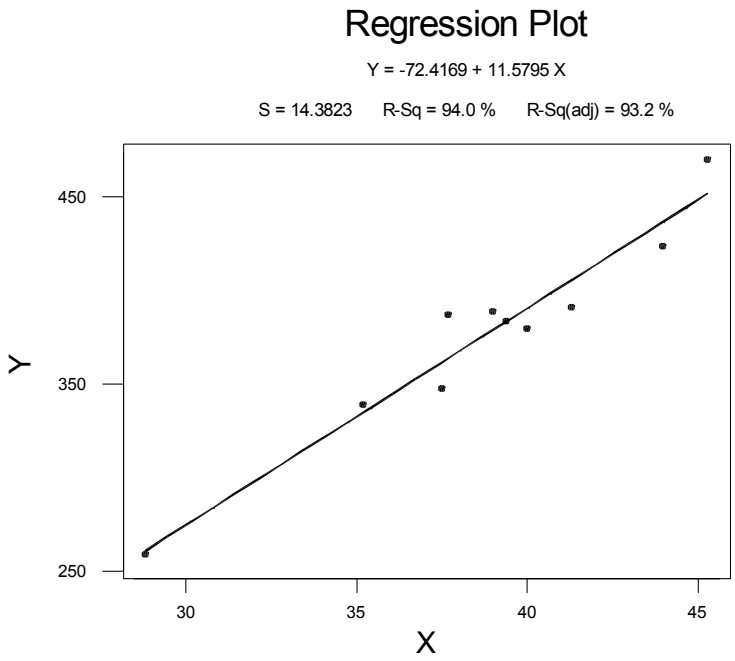


Indeed, if you do the regression computations, you'll get to see the true line exactly.



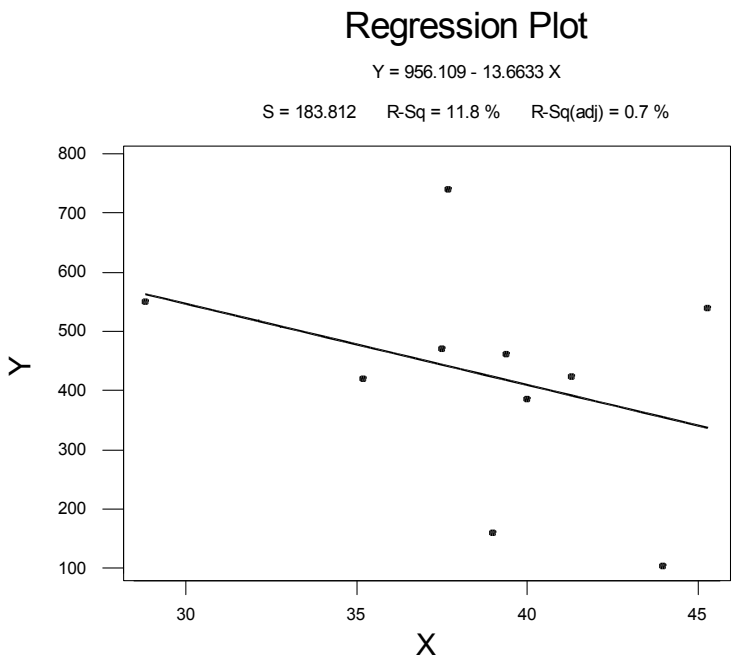
The equation is revealed as $Y = -11 + 10\,x$.

Now, what if there really were some noise? Suppose that $\sigma = 20$. The picture below shows what might happen.



The points stray from the true line. As a result, the fitted line we get, here $\hat{Y} = -72.4 + 11.58\,x$, is somewhat different from the true line.

What would happen if we had large, disturbing noise? Suppose that $\sigma = 150$. The picture below shows this problem:

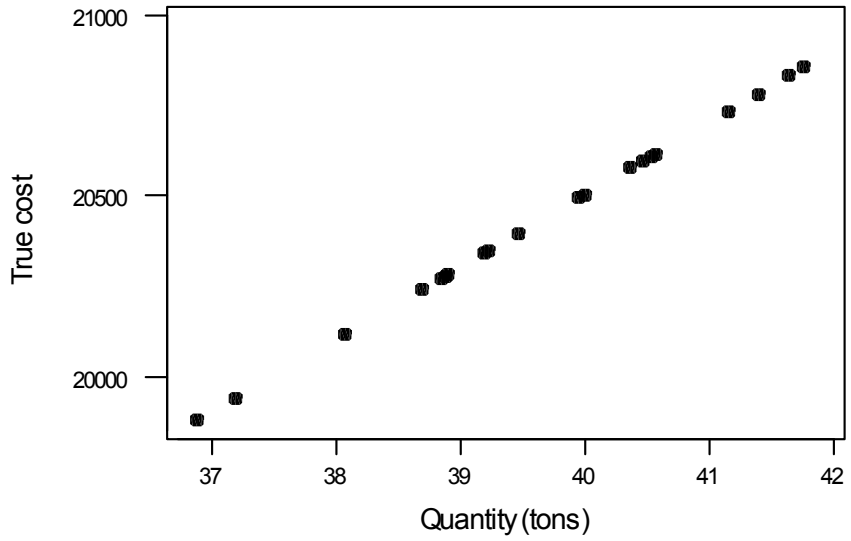


You might notice the change in the vertical scale! We didn't do a very good job of finding the correct line.

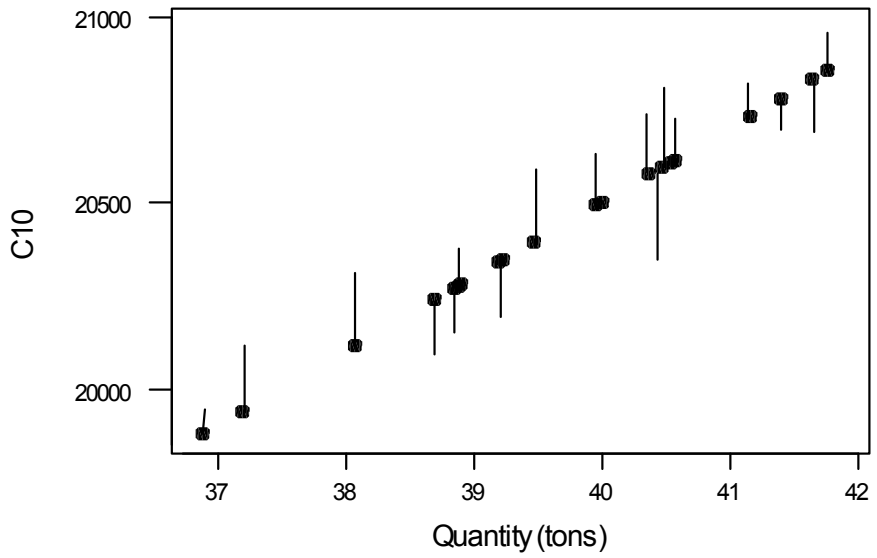
The points in this picture are so scattered that it's not even clear that we have any relationship at all between Y and x.

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 Cost = \$12,500 + 200 $\frac{\$}{\text{ton}}$ × 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:

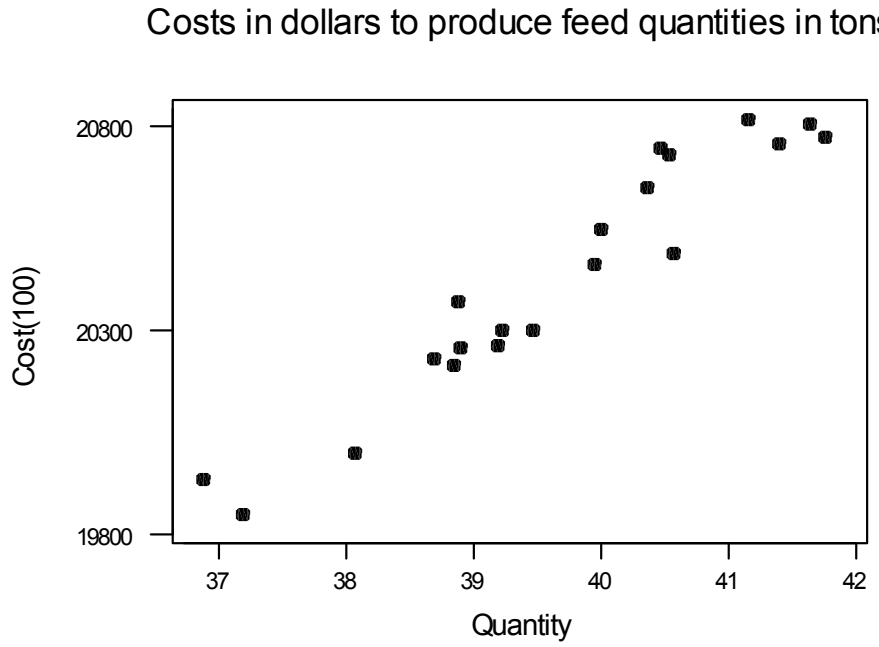


Here then are the data which we actually see:

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:



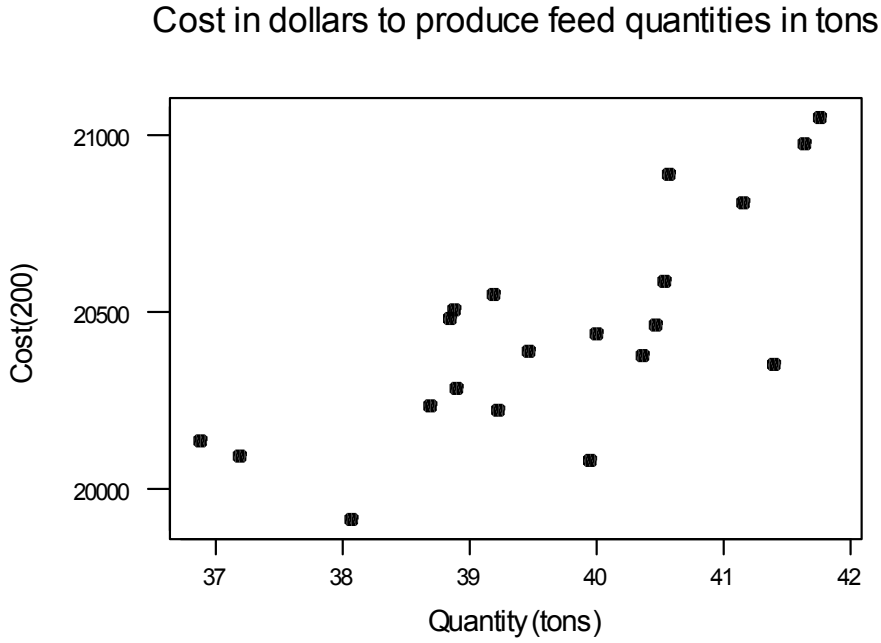
(There is a noise standard deviation of $100 in computing costs.)

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 Côst = \$12,088 + 210 $\frac{\$}{\text{ton}}\times$ 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_ε = \$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/ton
Noise standard deviation	\$100	s_ε = \$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:



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

For this scatterplot, the fitted regression equation is Côst = \$13,910 + 165 $\frac{\$}{\text{ton}}$ × 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 $\frac{\$}{\text{ton}}$ 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_ε = \$210.10.

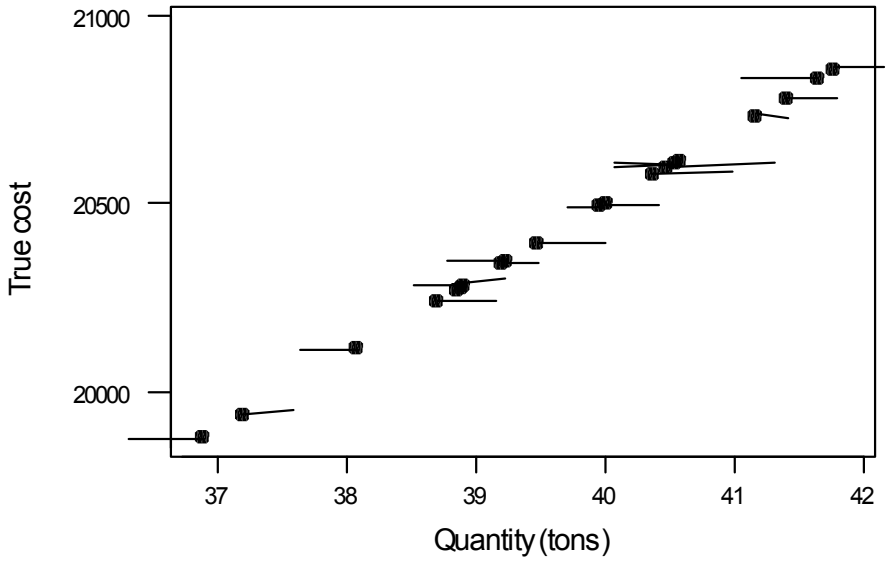
Quantity	True value	Value estimated from data
Fixed cost	\$12,500	b_0 = \$13,910
Variable cost	\$200/ton	b_1 = \$165/ton
Noise standard deviation	\$200	s_ε = \$210.10

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

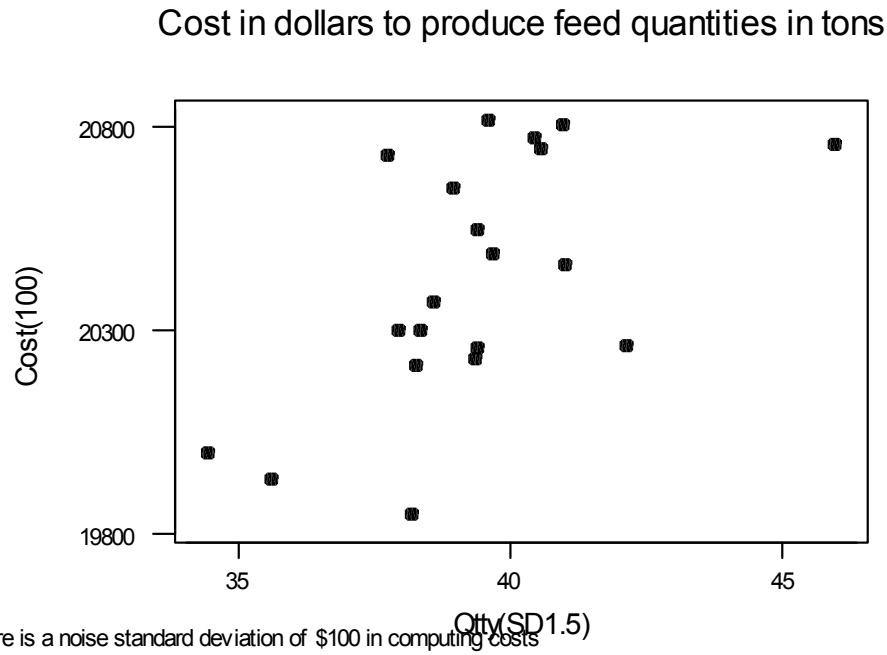
It is important to note that noise in assessing cost, the vertical variable, still gives us a statistically valid procedure. The uncertainty can be overcome with a larger sample size.

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.



(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 Côst = \$17,511 + 74.2 $\frac{\$}{\text{ton}}$ × Quantity. Also, this has s_ε = \$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_ε = \$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

True cost = $\beta_0 + \beta_1 \times$ True quantity

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

Suppose that we observe

$Y =$ True cost + ε

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

and

x = True quantity + ζ

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 $\beta_1 \dfrac{\sigma_x^2}{\sigma_x^2 + \sigma_\zeta^2}$.

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

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

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.

Quantity ⇔ X ⇔ independent variable Cost ⇔ Y ⇔ dependent variable	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.

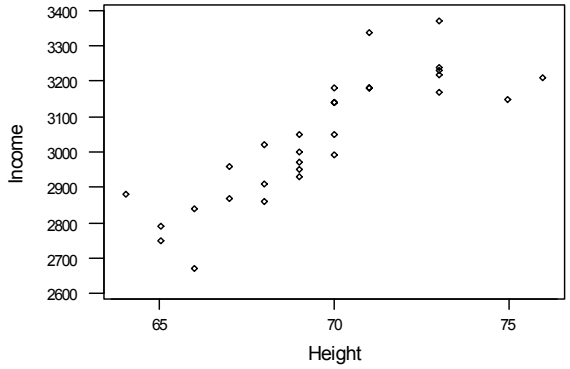
If you know or suspect that σ_ζ will be large relative to σ_x, there are two possible actions:

> 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.

The following data file shows information on 30 male MBA candidates at the University of Pittsburgh. The first column gives height in inches, and the second column gives the monthly income of the initial post-MBA job. (These appeared in the Wall Street Journal, 30 DEC 86.)

```
70 2990      68 2910      75 3150
67 2870      66 2840      68 2860
69 2950      71 3180      69 2930
70 3140      68 3020      76 3210
65 2790      73 3220      71 3180
73 3230      73 3370      66 2670
64 2880      70 3180      69 3050
70 3140      71 3340      65 2750
69 3000      69 2970      67 2960
73 3170      73 3240      70 3050
```

Here is a scatterplot:



This certainly suggests some form of relationship!

The results of the regression are these:

Regression Analysis

```
The regression equation is
Income = - 451 + 50.2 Height

Predictor       Coef       StDev          T        P
Constant      -451.1       418.5      -1.08    0.290
Height        50.179       6.008       8.35    0.000

S = 96.35       R-Sq = 71.4%     R-Sq(adj) = 70.3%

Analysis of Variance
Source       DF          SS          MS         F        P
Regression    1      647608      647608     69.76    0.000
Error        28      259938        9284
Total        29      907547


Unusual Observations
```

```
Obs    Height     Income        Fit  StDev Fit   Residual    St Resid
 18      71.0     3340.0     3111.6       19.5      228.4       2.42R
 26      66.0     2670.0     2860.7       27.9     -190.7      -2.07R

R denotes an observation with a large standardized residual
```

The section below gives fitted values corresponding to $\text{HEIGHT}_{\text{new}} = 66.5$. Note that Minitab lists the 66.5 value in its output; this is to remind you that you asked for this particular prediction.

```
Predicted Values for New Observations

New Obs     Fit     SE Fit          95.0% CI              95.0% PI
1        2885.8       25.6   (  2833.3,  2938.3)   (  2681.6,  3090.0)

Values of Predictors for New Observations

New Obs    Height
1            66.5
```

We see that the fitted equation is INCÔME = -451 + 50.2 HEIGHT. The obvious interpretation is that each additional inch of height is worth $50.20 per month. Can we believe any cause-and-effect here?

The R^2 value is reasonably large, so that this is certainly a useful regression.

Suppose that you wanted a 95% confidence interval for the true slope β. This would be given as $b \pm t_{\alpha/2;n-2}\,\text{SE}(b)$, which is $50.179 \pm 2.0484 \times 6.008$, or 50.179 ± 12.307. You should be able to locate the values 50.179 and 6.008 in the listing above.

Suppose that you'd like to make a prediction for a person with height 66.5. Minitab will give this to you, and reminds you by repeating the value 66.5 in its Session window. You can see from the above that the fit (or point prediction) is 2,885.8. You could of course have obtained this as $-451.1 + 50.179 \times 66.5 \approx 2{,}885.8$. Minitab provides several other facts near this 2,885.8 figure. The only thing likely to be useful to you is identified as the `95.0% PI`, meaning 95% prediction interval. This is (2,681.5, 3,090.1), meaning that you're 95% sure that the INCOME for a person 66.5 inches tall would be between $2,681.5 and $3,090.1.

The residual-versus-fitted plot must of course be examined. It's not shown here, just to save space, but it should be noted that this plot showed no difficulties.

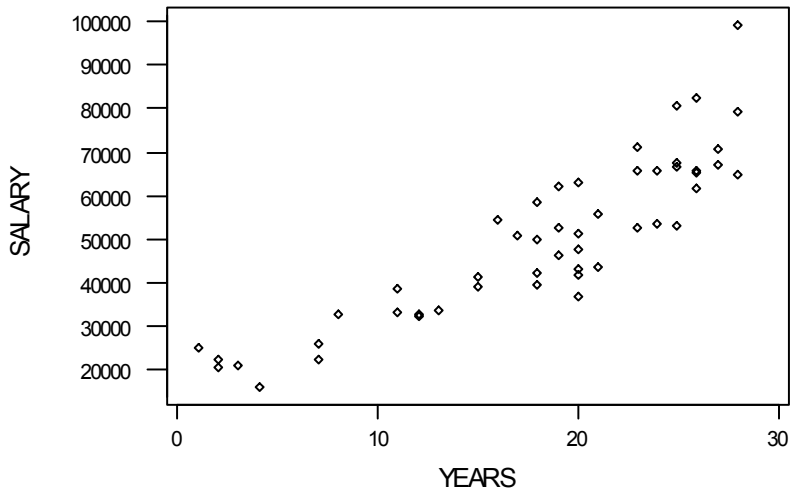
The data listed below give two numbers for each of 50 middle-level managers at a particular company. The first number is annual salary, and the second number is years of experience.

We are going to examine the relationship between salary and years of experience. Then we'll use the residuals to identify individuals whose salary is out of line with their experience.

26075 7	43076 20	63022 20	24833 1	54288 16
79370 28	56000 21	47780 20	65929 26	20844 3
65726 23	58667 18	38853 15	41721 20	32586 12
41983 18	22210 7	66537 25	82641 26	71235 23
62309 19	20521 2	67447 25	99139 28	36530 20
41154 15	49717 18	64785 28	52624 23	52745 19
53610 24	33233 11	61581 26	50594 17	67282 27
33697 13	43628 21	70678 27	53272 25	80931 25
22444 2	16105 4	51301 20	65343 26	32303 12
32562 8	65644 24	39346 18	46216 19	38371 11

Note that the dependent variable is SALARY, and the independent variable is YEARS, meaning years of experience.

Here is a scatterplot showing these data:



Certainly there seems to be a relationship!

We can get the regression work by doing this:

Stat ⇒ **Regression** ⇒ **Regression** ⇒
[**Response:** SALARY
Predictors: YEARS
Storage ⇒ [⊙ **Residuals**
OK ⇒]
OK ⇒]

There are other interesting features of these values, and this exercise will not exhaust the work that might be done. Here are the regression results:

Regression Analysis

The regression equation is
SALARY = 11369 + 2141 YEARS

Predictor	Coef	SE Coef	T	P
Constant	11369	3160	3.60	0.001
YEARS	2141.3	160.8	13.31	0.000

S = 8642 R-Sq = 78.7% R-Sq(adj) = 78.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	13238780430	13238780430	177.25	0.000
Error	48	3585076845	74689101		
Total	49	16823857275			

Unusual Observations

Obs	YEARS	SALARY	Fit	SE Fit	Residual	St Resid
31	1.0	24833	13511	3013	11322	1.40 X
35	28.0	99139	71326	2005	27813	3.31R
45	20.0	36530	54195	1259	-17665	-2.07R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.

There's more information we need right now, but you can at least read the fitted regression line as

$$\hat{\text{SALARY}} = 11{,}369 + 2{,}141\ \text{YEARS}$$

with the interpretation that a year of experience is worth $2,141.

Let's give a 95% confidence interval for the regression slope. This particular topic will pursued vigorously later. The slope is estimated as 2,141.3, and its standard error is given as 160.8. A standard error, or SE, is simply a data-based estimate of a standard deviation of a statistic. Thus, the 95% confidence interval is $2{,}141.3 \pm t_{0.025;48}\ 160.8$, or $2{,}141.3 \pm 2.0106 \times 160.8$, or $2{,}141.3 \pm 323.30448$. This precision is clearly not called for, so we can give the interval as simply 2,141 ± 323. You might observe that this is very close to (estimate) ± 2 SE.

The value of S, meaning s_ε, is 8,642. The interpretation is that the regression explains salary to within a noise with standard deviation $8,642. The standard deviation of

SALARY (for which the computation is not shown) is \$18,530; the regression would thus have to be appraised as quite successful.

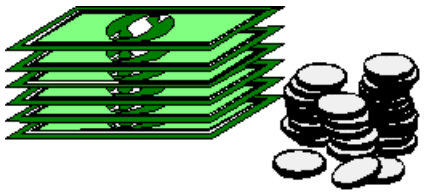
The fitted value for the i^{th} person in the list is $\text{SALÂRY}_i = 11{,}369 + 2{,}141\ \text{YEARS}_i$ and the residual for the i^{th} person is $\text{SALARY}_i - \text{SALÂRY}_i$, which we will call e_i. Clearly the value of e_i indicates how far above or below the regression line this person is.

The Minitab request **Storage** ⇒ [⊙ **Residuals]** caused the residuals to be calculated and saved. These will appear in a new column in the spreadsheet, RESI1. (Subsequent uses would create RESI2, RESI3, and so on.) Here are the values, displayed next to the original data:

SALARY	YEARS	RESI1	SALARY	YEARS	RESI1	SALARY	YEARS	RESI1
26075	7	-283.4	43628	21	-12708.7	99139	28	27813.1
79370	28	8044.1	16105	4	-3829.5	52624	23	-7995.4
65726	23	5106.6	65644	24	2883.3	50594	17	2822.5
41983	18	-7929.8	63022	20	8826.6	53272	25	-11630.0
62309	19	10254.9	47780	20	-6415.4	65343	26	-1700.3
41154	15	-2334.9	38853	15	-4635.9	46216	19	-5838.1
53610	24	-9150.7	66537	25	1635.0	54288	16	8657.8
33697	13	-5509.3	67447	25	2545.0	20844	3	3050.8
22444	2	6792.2	64785	28	-6540.9	32586	12	-4478.9
32562	8	4062.3	61581	26	-5462.3	71235	23	10615.6
43076	20	-11119.4	70678	27	1493.4	36530	20	-17665.4
56000	21	-336.7	51301	20	-2894.4	52745	19	690.9
58667	18	8754.2	39346	18	-10566.8	67282	27	-1902.6
22210	7	-4148.4	24833	1	11322.5	80931	25	16029.0
20521	2	4869.2	65929	26	-1114.3	32303	12	-4761.9
49717	18	-195.8	41721	20	-12474.4	38371	11	3447.4
33233	11	-1690.6	82641	26	15597.7			

It's easy to pick out the most extreme residuals. These are -\$17,665.4 (point 45, a 20-year person with salary \$36,530) and \$27,813.1 (point 35, a 28-year person with \$99,139). These two points are reported after the regression as being unusual observations, and they are noted to have large residuals.

Point 31 is a high influence point, but that issue is not the subject of this document.



Let's think of role of b_1 in fitted model $\hat{Y} = b_0 + b_1 x$. This says that as x goes up by one unit, Y (tends to) go up by b_1. In fact, it's precisely what we call $\frac{dY}{dx}$ in calculus. This measures the sensitivity of Y to x. Suppose that x is the price at which something is sold and Y is the quantity that clears the market. It helps to rewrite this as $Q = b_0 + b_1 P$. We'd expect b_1 to be negative, since the quantities consumed will usually decrease as the price rises. If currently the price is P_0 and currently the quantity is Q_0, then $Q_0 = b_0 + b_1 P_0$. Now a change of price from P_0 to $P_0 + \theta$ leads to a change in quantity from Q_0 to $b_0 + b_1 (P_0 + \theta) = b_0 + b_1 P_0 + b_1\theta = Q_0 + b_1 \theta$. Thus, the change in quantity is $b_1 \theta$. The proportional change ratio - $\frac{\frac{\text{change in } Q}{Q_0}}{\frac{\text{change in } P}{P_0}}$ is called an elasticity. (We use minus sign since they tend to move in opposite direction, and we like positive elasticities.) Of course, we can simplify this a bit, writing it finally as - $b_1 \frac{P_0}{Q_0}$.

Suppose that a demand "curve" has equation $Q = 400{,}000 - 2{,}000\, P$. We put "curve" in quotes as this equation is actually a straight line. Indeed the illustrations in most microeconomics texts use straight lines. Suppose that the current price is $P_0 = \$60$. The quantity corresponding to this is $Q_0 = 400{,}000 - 2{,}000 \times 60 = 280{,}000$. Now suppose that the price rises by 1%, to \$60.60. The quantity is now $400{,}000 - 2{,}000 \times 60.60 = 278{,}800$. This is a decrease of 1,200 in quantity consumed, which is a decrease of $\frac{1{,}200}{280{,}000} \approx 0.0043 = 0.43\%$. Thus a 1% increase in price led to a decrease in quantity of 0.43%, so we would give the elasticity as 0.43. This is less than 1, so that the demand is *inelastic*.

We could of course obtain this as - $b_1 \frac{P_0}{Q_0} = 2{,}000 \times \frac{60}{280{,}000} \approx 0.43$.

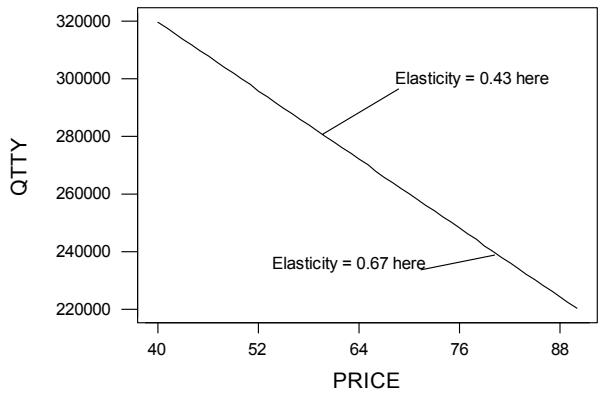
> One approach follows through on the "1% change in price" idea. The other approach simply uses - $b_1 \frac{P_0}{Q_0}$. These will not give exactly the same numerical result, but the values will be very close.

Observe that the elasticity calculation depends here on where we started. Suppose that we had started at $P_0 = \$80$. The quantity consumed at this price of \$80 is $Q_0 = 400{,}000 - 2{,}000 \times 80 = 240{,}000$. An increase in price by 1%, meaning to \$80.80, leads to a new quantity of $400{,}000 - 2{,}000 \times 80.80 = 238{,}400$. This is a decrease of 1,600 in

quantity consumed. In percentage terms, this decrease is $\frac{1{,}600}{240{,}000} \approx 0.0067 = 0.67\%$.
We would give the elasticity as 0.67.

This is also found as $- b_1 \frac{P_0}{Q_0} = 2{,}000 \times \frac{80}{240{,}000} \approx 0.67$.

We can see that the elasticity for this demand curve $Q = 400{,}000 - 2{,}000\,P$ depends on where we start.



You might check that at starting price $P_0 = \$120$, the elasticity would be computed as 1.50, and we would now claim that the demand is (highly) *elastic*.

> Details: Find $Q_0 = 400{,}000 - 2{,}000 \times 120 = 160{,}000$. The 1% increase in price would lead to a consumption decrease of $1.2 \times 2000 = 2{,}400$. This is a percentage decrease of $\frac{2{,}400}{160{,}000} = 0.015 = 1.5\%$.

What is the shape of a demand curve on which the elasticity is the same at all prices?
This is equivalent to asking about the curve for which $\frac{\frac{dQ}{Q}}{\frac{dP}{P}} = -c$, where c is constant. (The minus sign is used simply as a convenience; of course c will be positive.) This condition can be expressed as

$$\frac{dQ}{Q} = -c\,\frac{dP}{P}$$

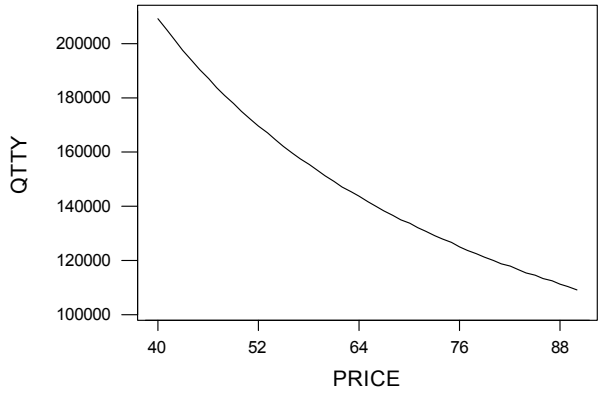
for which the solution is $\log Q = -c \log P + d$. This uses the calculus fact that $\frac{d}{dt}\log f(t) = \frac{f'(t)}{f(t)}$. The result can be reexpressed as

$$e^{\log Q} = e^{-c\log P + d}$$

or

$$Q = mP^{-c}$$

where $m = e^d$. The picture below shows the graph of $Q = 4{,}000{,}000\ P^{-0.8}$. This curve has elasticity 0.80 at every price.



The equation $\log Q = -c \log P + d$ is a simple linear relationship between $\log Q$ and $\log P$. Thus, if we base our work on a regression of log(quantity) on log(price), then we can claim that the resulting elasticity is the same at every price.

If you began by taking logs of BOTH variables and fitted the regression (log-on-log), you'd get the elasticity directly from the slope (with no need to worry about P_0 or Q_0). That is, in a log-on-log regression, the elasticity is exactly $-b_1$.

Now, where would we get the information to actually perform one of these regressions? It would be wonderful to have prices and quantities in a number of different markets which are believed to be otherwise homogeneous. For example, we could consider per capita cigarette sales in states with different tax rates. Unfortunately, it is rather rare that we can get cleanly organized price and quantity on a geographic basis. It is more common by far to have price and quantity information over a large area at many points in time. In such a case, we would also have to worry about the possibility that the market changes over time.

Measures of quality for a regression:

R^2 want big (as close as possible to 100%)

s_ε want small

s_ε/s_Y want small (as close as possible to 0%)

F want big (up to ∞)

t statistic for b_1 want far away from 0

Finally, here's the summary of what to do with the regression "game" with one independent variable.

Begin by making a scatterplot of (X, Y). You might see the need to transform.
Indications: Excessive curvature or extreme clustering of points in one region of the plot.

Note SD(Y).

Perform regression of Y on X. Note R^2, t for b_1, and s_ε.

Another useful summary is the correlation between x and Y. Formally, this is computed as $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$. It happens that b_1 and r have the same sign.

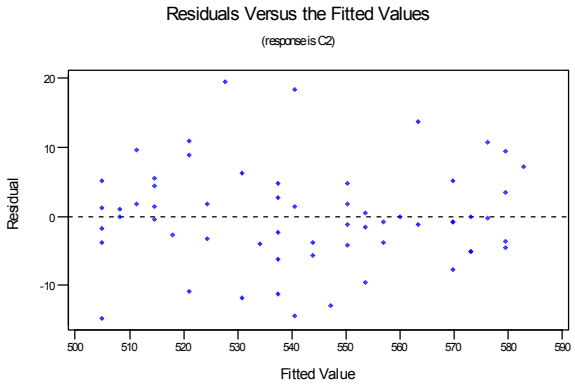
Plot residual versus fitted. If you have pathologies, correct them and start over.

Use the regression to make your relevant inference. We'll check up on this later.

It is now standard practice to examine the plot of the residuals against the fitted values to check for appropriateness of the regression model. Patterns in this plot are used to detect violations of assumptions. The plot is obtained easily in most computer packages.

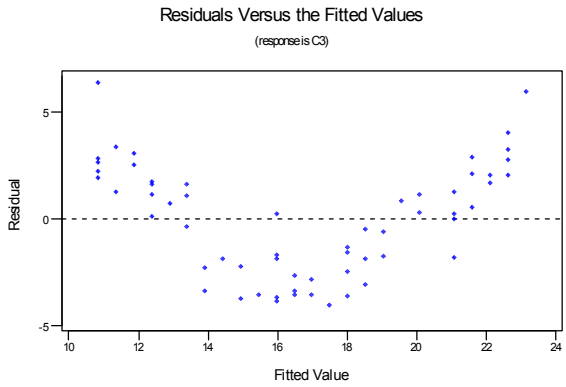
> If you are working in Minitab, click on **Stat** ⇒ **Regression** ⇒ **Regression**, indicate the dependent and independent variable(s), then click on **Graphs**, and check off **Residuals versus fits**. In the resulting plot, you can identify individual points by using **Editor** ⇒ **Brush**.

You HOPE that plot looks like this:



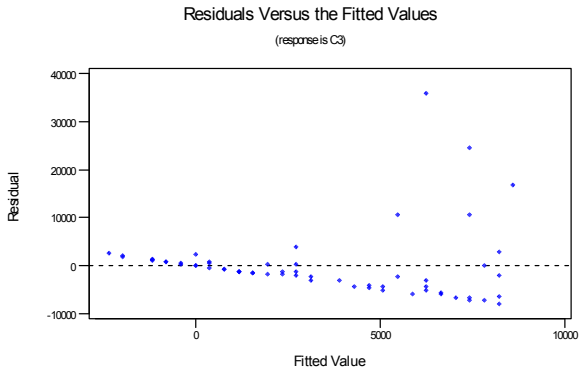
This picture would be described as "patternless."

There are a number of common pathologies that you will encounter. The next picture shows curvature:



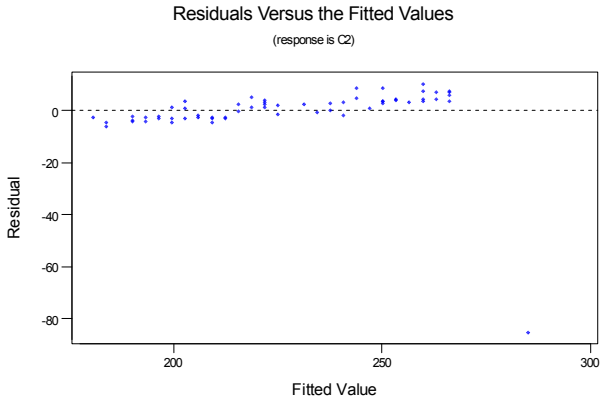
The cure for curvature consists of revising the model to be curved in one or more of the predictor variables. Generally this is done by using X^2 (in addition to X), but it can also be done by replacing X by $\log X$, by $\sqrt{X}$, or some other nonlinear function.

A very frequent problem is that of residuals that expand rightward on the residual versus fitted plot.



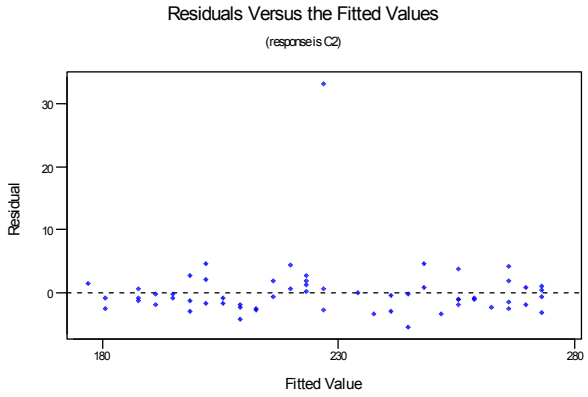
On this picture there is considerably more scatter on the right side. It appears that large values of $\hat{Y}$ also have widely scattered residuals. Since the residuals are supposed to be independent of all other parts of the problem, this picture suggests that there is a violation of assumptions. This problem can be cured by replacing Y by its logarithm.

The picture below shows a destructive outlier.



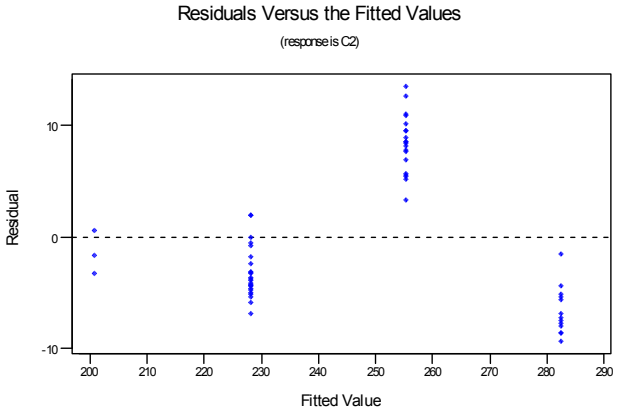
The point at the lower right has a very unusual value in the X-space, and it also has an inappropriate Y. This point must be examined. If it is incorrect, it should be corrected. If the correct value cannot be determined, the point must certainly be removed. If the point is in fact correct as coded, then it must be set aside. You can observe that this point is masking a straight-line relationship which would otherwise be incorporated in the regression coefficient estimates.

In the picture below, there is a very unusual point, but it is not destructive.



The unusual point in this plot must be checked, of course, but it is not particularly harmful to the regression. At worst, it slightly elevates the value of b_0, the estimated intercept.

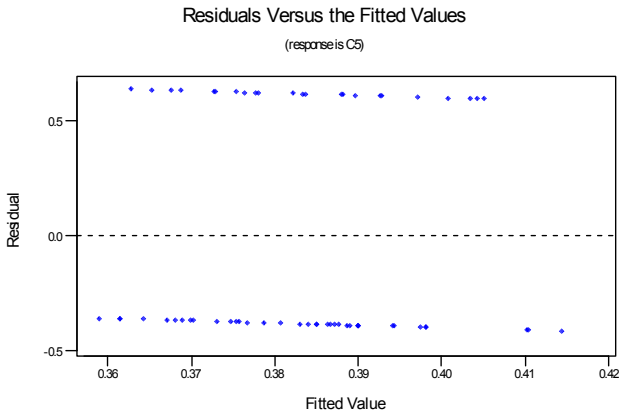
On the picture below, there are several vertical stripes. This indicates that the X-values consisted of a small finite number (here 4) of different patterns.



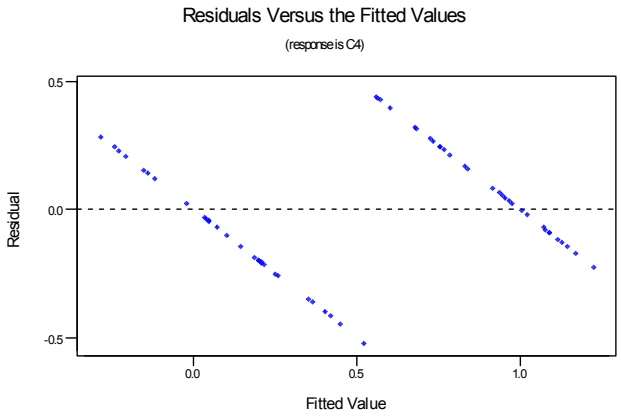
The required action depends on the discreteness in the X-values that caused the stripes.

> If the X-values are quantitative, then the regression is appropriate. The user should check for curvature, of course.
>
> If the X-values are numerically-valued steps of an ordered categorical variable, then the regression is appropriate. The user should check for curvature.
>
> If the X-values are numerically-valued levels of a qualitative non-ordinal categorical variable, then the problem should be recast as an analysis of variance or as an analysis of covariance.

Finally, we have a pattern in which imperfect stripes go across the residual-versus-fitted pattern.



These stripes can be generally horizontal, as above, or they can be oblique:



These pictures suggest that there are only two values of *Y*. In such a situation, the appropriate tool is logistic regression.

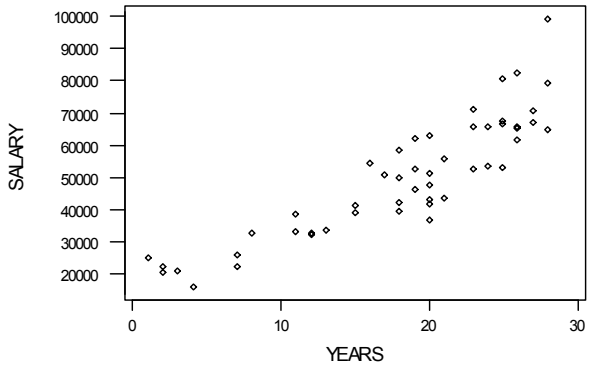
If there are three or more such stripes, indicating the same number of possible *Y*-values, then the recommended technique is multiple logistic regression. Multiple logistic regression can be either "ordinal" or "nominal" according to whether *Y* is nominal or ordinal.

The file X:\SOR\B011305\M\SALARY.MTP gives SALARY and YEARS of experience for a number of middle-level executives.

Let's first find the regression of SALARY on YEARS. The data set consists of 50 points and looks like this:

CASE	SALARY	YEARS	CASE	SALARY	YEARS
1	26075	7	.	.	.
2	79370	28	47	67282	27
3	65726	23	48	80931	25
4	41983	18	49	32303	12
.	.	.	50	38371	11

Here is a scatterplot:



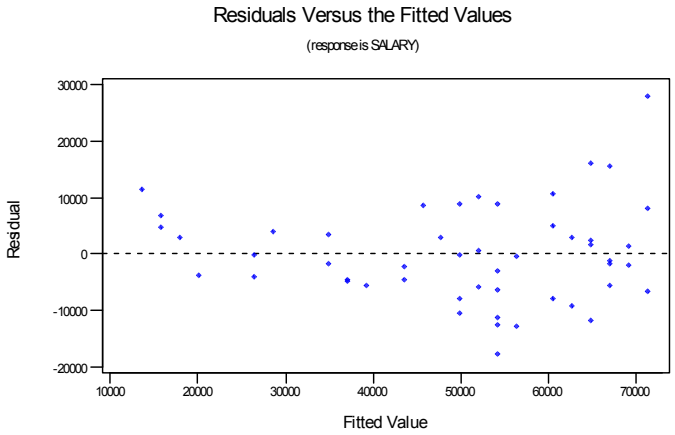
The regression results from Minitab are these:

```
The regression equation is
SALARY = 11369 + 2141 YEARS

Predictor        Coef     SE Coef          T        P
Constant        11369        3160       3.60    0.001
YEARS          2141.3       160.8      13.31    0.000

S = 8642        R-Sq = 78.7%     R-Sq(adj) = 78.2%

Analysis of Variance
Source            DF            SS            MS         F       P
Regression         1   13238780430   13238780430    177.25   0.000
Residual Error    48    3585076845      74689101
Total             49   16823857275

Unusual Observations
Obs    YEARS     SALARY         Fit      SE Fit    Residual    St Resid
 31      1.0      24833       13511        3013       11322        1.40 X
 35     28.0      99139       71326        2005       27813        3.31R
 45     20.0      36530       54195        1259      -17665       -2.07R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

The fitted regression equation is certainly $\widehat{SALARY} = 11{,}369 + 2{,}141\ YEARS$.

What interpretation can we give for the estimated regression slope? The slope 2,141 suggests that each year of experience translates into $2,141 of salary.

Here is the residual versus fitted plot. It shows a common pathological pattern.



The residuals spread out to a greater degree at the right end of the graph. This is very common.

> This same observation might have been made in the original plot of Salary vs Years. In the case of *simple* regression (just one independent variable) the residual-versus-fitted plot is a rescaled and tilted version of the original plot.

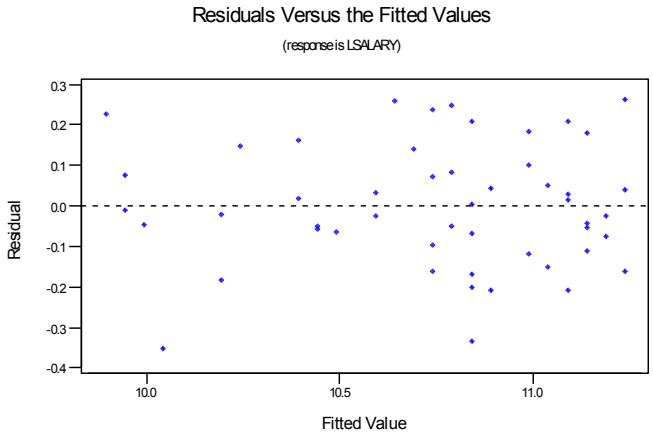
We'll handle this violation of model assumptions by making a transformation on SALARY. The usual solution is to use the logarithm of salary. Here, we'll let LSALARY be the log (base *e*) of SALARY. The output from Minitab is this:

Regression Analysis

```
The regression equation is
LSALARY = 9.84 + 0.0500 YEARS

Predictor       Coef     SE Coef          T        P
Constant     9.84132     0.05635     174.63    0.000
YEARS       0.049978    0.002868      17.43    0.000

S = 0.1541      R-Sq = 86.4%     R-Sq(adj) = 86.1%

Analysis of Variance
Source       DF          SS          MS         F        P
Regression    1      7.2118      7.2118    303.65    0.000
Error        48      1.1400      0.0238
Total        49      8.3519
```

```
Unusual Observations
Obs     YEARS    LSALARY        Fit  StDev Fit   Residual    St Resid
 19       4.0     9.6869    10.0412     0.0460    -0.3543      -2.41R
 31       1.0    10.1199     9.8913     0.0537     0.2286       1.58 X
 45      20.0    10.5059    10.8409     0.0225    -0.3350      -2.20R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

The residual-versus-fitted plot for this revised regression is shown next:



This plot shows that the problem of expanding residuals has been cured.

Let's also note that in the original regression, $R^2 = 78.7\%$, whereas in the regression using LSALARY, $R^2 = 86.4\%$.

The fitted regression is now $\widehat{\text{LSALARY}} = 9.84132 + 0.049978\ \text{YEARS}$, which can be rounded to $\widehat{\text{LSALARY}} = 9.84 + 0.05\ \text{YEARS}$.

This suggests the interpretation that each year of experience is worth 0.05 in the *logarithm* of salary. You can exponentiate the expression above to get

$$e^{\widehat{\text{LSALARY}}} = e^{9.84+0.05\text{YEARS}} = \widehat{\text{SALARY}} = e^{9.84} \times (e^{0.05})^{\text{YEARS}}$$

In this form, increasing YEARS by 1 causes the fitted salary to be multiplied by $e^{0.05}$. You can use a calculator to get a number out of this, but we have a simple approximation $e^{0.05} \approx 1.05$. Thus, accumulating a year of experience suggests that salary is to be multiplied by 1.05; this is a 5% raise.

> You might be helped by this useful approximation. If t is near zero, then $e^t \approx 1 + t$. Thus, getting a coefficient of 0.05 in this regression leads directly to the 5% raise interpretation.

Let's reconsider the consequences of our work in regressing SALARY on YEARS. The regression model for this problem (in original units) was

$$\text{SALARY}_i = \beta_0 + \beta_1 \text{YEARS}_i + \varepsilon_i \qquad [1]$$

The noise terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ were assumed to be a sample from a population with mean zero and with standard deviation σ.

Our fitted regression was

$$\hat{\text{SALARY}} = 11{,}369 + 2{,}141.3 \text{ YEARS}$$

The estimate of σ was computed as 8,642.29. This was labeled S by Minitab, but you'll also see the symbols s or $s_{Y|x}$ or s_ε.

Predicted salaries can be obtained by direct plug-in. For someone with 5 years of experience, the prediction would be $11{,}369 + (2{,}141.3 \times 5) = 22{,}075.5$. Here is a short table showing some predictions; this refers to the column "Predicted SALARY with basic model [1]."

YEARS	Predicted SALARY with basic model [1]	Predicted SALARY with logarithm model [2]	95% prediction interval for SALARY, using basic model [1]	95% prediction interval for SALARY, using logarithm model [2]
5	22,076	24,130	4,020.8 to 40,130.7	17,487.6 to 33,296.2
10	32,782	30,980	15,037.6 to 50,527.1	22,577.3 to 42,510.2
15	43,488	39,775	25,910.5 to 61,067.2	29,071.1 to 54,420.7
20	54,195	51,066	36,635.5 to 71,755.3	37,335.5 to 69,842.6
25	64,902	65,563	47,212.1 to 82,591.8	47,824.8 to 89,877.2

It happens that model [1] will lead to a residual versus fitted plot with the very common pattern of expanding residuals. The cure comes in replacing SALARY with LSALARY and considering the model

$$\text{LSALARY}_i = \beta_0 + \beta_1 \text{YEARS}_i + \varepsilon_i \qquad [2]$$

For this model, the residual versus fitted plot shows a benign pattern, and we are able to believe the assumption that the ε_i's are a sample from a population with mean 0 and standard deviation σ. The LOG here is base e.

> Please note that we have "recycled" notation. The parameters β_0, β_1, and σ, along with the random variables ε_1 through ε_n, do not have the same meanings in [1] and [2].

The fitted model corresponding to [2] is

$$\widehat{\text{LSALARY}} = 9.84132 + 0.049978\ \text{YEARS}$$

For this model, the predicted log-salary for someone with 5 years of experience would be $9.84132 + (5 \times 0.049978) = 10.09121$. In original units, this would be $e^{10.09121} \approx 24{,}130$. This is the first entry in the column "Predicted SALARY with logarithm model [2]."

So what's the big deal about making a distinction between models [1] and [2]? The fitted values are different, with model [2] giving larger values at the end and smaller values in the middle. Here are the major answers:

(*a*) The difference of about $4,000 in the model is not trivial. The two models are certainly not giving very similar answers.

(*b*) Model [1] has assumed equal noise standard deviations throughout the entire range of YEARS values. The plot of (SALARY, YEARS) suggests that this is not realistic. The residual versus fitted plot for model [1] makes this painfully obvious. The consequences will be seen in predictions. Examine the column "95% prediction interval for SALARY, using basic model [1]." Each of these prediction intervals is about $36,000 wide. This may be realistic for those people with high seniority (high values of YEARS), but it's clearly off the mark for those with low seniority.

(*c*) Model [2] has assumed equal noise standard deviations, in terms of LSALARY, throughout the entire range of YEARS values. This is more believable. In the column "95% prediction interval for SALARY, using logarithm model [2]" the lengths vary, with the longer intervals associated with longer seniority.

Consider the linear regression problem with data (x_1, Y_1), (x_2, Y_2), …, (x_n, Y_n). We form the model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

The assumptions which accompany this model include these statements about the ε_i's :

> The noise terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent of each other and of all the other symbols in the problem.
>
> The noise terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are drawn from a population with mean zero.
>
> The noise terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are drawn from a population with standard deviation σ.

Situations in which $SD(\varepsilon_i)$ appears to systematically vary would violate these assumptions. The residual versus fitted plot can detect these situations; this plot shows the points $(e_1, \hat{Y}_1)$, $(e_2, \hat{Y}_2)$, …, $(e_n, \hat{Y}_n)$. We will sometimes see that the residuals have greater variability when the fitted values are large. The recommended cure is that Y_i be replaced by $\log(Y_i)$.

> If some of the Y_i's are zero or negative, we'd use $\log(Y_i + c)$ for some c big enough to make all values of $Y_i + c$ positive.

Why does this work? It's a bit of insanely tricky math, but it's fun.

Suppose that the residual versus fitted plot suggests that $SD(Y_i)$ is big when $\beta_0 + \beta_1 x_i$ is big. We'll use the symbol $\mu_i = \beta_0 + \beta_1 x_i$ for the expected value of Y_i. The observation is that $SD(Y_i)$ is big when μ_i is big.

The are many ways of operationalizing the statement "$SD(Y_i)$ is big when μ_i is big." The description that will work for us is

$$SD(Y_i) = \alpha \mu_i$$

That is, the standard deviation grows proportional to the mean. The symbol α is just a proportionality constant, and it's quite irrelevant to the ultimate solution.

Let's seek a functional transformation $Y \mapsto g(Y)$ that will solve our problem.

> The symbol g represents a function to be found; perhaps we'll decide $g(t) = t^3$ or $g(t) = \cos(2\pi t)$ or $g(t) = \dfrac{t}{t^2 + 1}$ or something else.

It's convenient to drop the symbol *i* for now. We'll be talking about the general phenomenon rather that about our specific *n* data points.

By "solve our problem" we are suggesting these two notions:

If $\mu_k \neq \mu_\ell$ then $E(\, g(Y_k)\,) \neq E(\, g(Y_\ell)\,)$; that is, the function *g* preserves differentness of means (expected values).

If $\mu_k \neq \mu_\ell$ then $SD(\, g(Y_k)\,) \approx SD(\, g(Y_\ell)\,)$; that is, the function *g* allows Y_k and Y_ℓ to have unequal means but approximately equal standard deviations.

We're going to use two mathematical facts and one statistical fact.

MATH FACT 1: If *g* is any well-behaved function, then

$$g(y) \approx g(\mu) + g'(\mu)\,(y - \mu)$$

This is just Taylor's theorem. It can be thought of as an approximate version of the mean value theorem in calculus. The symbol μ can be any convenient value. When we use this with a random variable *Y*, we'll take μ as the mean of the random variable.

MATH FACT 2: If the function *g* has a derivative for which

$$g'(t) = \frac{k}{t}$$

then the function $g(t) = \log t$ is one possible solution. (There are many possible solutions, but this one is simplest. Also, we'll use base-*e* logs. The most detailed solution would be $g(t) = A + k \log t$.)

STATISTICAL FACT: If Y is a random variable with mean μ, and if g is a well-behaved function, then

$$E(\, g(Y)\,) \;\approx\; g(\mu)$$

$$\mathrm{Var}(\, g(Y)\,) \;\approx\; \left[\, g'(\mu)\right]^2 \mathrm{Var}(Y)$$

This is described in terms of the variance, which is the square of the standard deviation. In terms of the standard deviation, we could state this as

$$\mathrm{SD}(\, g(Y)\,) \;\approx\; \left|\, g'(\mu)\right| \mathrm{SD}(Y)$$

If we end up with an *in*creasing function for g, then $g'(\mu) > 0$, and we can write this as simply

$$\mathrm{SD}(\, g(Y)\,) \;\approx\; g'(\mu)\;\mathrm{SD}(Y)$$

Now we're ready to accomplish the objective.

We have

$$\begin{cases} \mathrm{E}(Y) &= \mu \\ \mathrm{SD}(Y) &= \alpha\sqrt{\mu} \end{cases}$$

We seek a function g so that

$$\begin{cases} \mathrm{E}\left(g(Y)\right) & = \text{expression that varies as } \mu \text{ varies} \\ \mathrm{SD}\left(g(Y)\right) & = \text{expression that is approximately constant as } \mu \text{ varies} \end{cases}$$

We'll use MATH FACT 1 to write

$$g(Y) \approx g(\mu) + g'(\mu)\,(Y - \mu)$$

It's important that the symbol μ here is $E(Y)$, the mean of Y.

We'll use the STATISTICAL FACT to write

$$\text{SD}(\, g(Y)\,) \;\approx\; g'(\mu)\;\text{SD}(Y)$$

We wrote this without the absolute value sign. The ultimate g will turn out to be an increasing function, so we have not created any difficulties.

Now we'd like this to be *un*varying as μ varies. We'll set up this condition:

$$g'(\mu)\;\text{SD}(Y) \overset{\text{want}}{=} c$$

The symbol c is just some generic constant, a quantity that does not change. Indeed, we don't even care about the value of c; all we want is that it doesn't move around.

Now insert our assumed relationship $\text{SD}(Y) = \alpha\,\mu$. We have this:

$$g'(\mu)\,\alpha\,\mu \overset{\text{want}}{=} c$$

We can solve this as

$$g'(\mu) \overset{\text{want}}{=} \frac{c}{\alpha\,\mu} = \frac{k}{\mu}$$

At the last step, we wrote $k = \frac{c}{\alpha}$; after all, if c is a constant that we don't care about and α is a mere proportionality constant, we can certainly smash them together into the same symbol.

Now we exploit MATH FACT 2 to claim that $g(\mu) = \log \mu$ is a function that solves this. Since this last form just uses μ as a place-holder, we could write it as $g(t) = \log t$ or as $g(Y) = \log(Y)$.

It's this last thing that solves our problem! We now see that

$$\begin{cases} \text{E}(\log Y) & \approx \log(\mu) & \text{which changes as } \mu \text{ changes} \\ \text{SD}(\log Y) & \approx c & \text{which does not change as } \mu \text{ changes} \end{cases}$$

For these reasons, a residual versus fitted plot which shows right-expanding residuals leads to the cure of replacing the Y_i's with their logarithms.

The sample correlation coefficient between data columns (variables) x and y is calculated as

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}}$$

where

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y} = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

Here n is the number of data points. Equivalently, n is the number of entries in each of the data columns.

Clearly, we can write

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}} = \frac{\frac{1}{n-1} S_{xy}}{\sqrt{\frac{1}{n-1} S_{xx}\, \frac{1}{n-1} S_{yy}}} = \frac{\frac{1}{n-1} S_{xy}}{s_x\, s_y}$$

Correlations are always between -1 and +1.

If r = -1 then the n data points lie exactly on a line of negative slope. (On a conventional graph, this line would go from northwest to southeast.)

If r = +1 then the n data points lie exactly on a line of positive slope. (On a conventional graph, this line would go from southwest to northeast.)

If r is near zero, then we would say that the linear relationship between x and y is weak. This does not preclude a strong non-linear relationship, but such situations are rare.

The numerator of r, namely $\frac{1}{n-1} S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$, is called the *sample covariance* of x and y, written Cov(x, y).

Certainly Cov(x, y) = $r\, s_x\, s_y$.

The sample covariance Cov(x, y) estimates a population covariance. It's hard to make a good notation that distinguishes these, but here we'll use Cov(X, Y) for the population covariance.

The sample correlation coefficient r estimates the population correlation coefficient ρ. The formal definition of ρ requires integrals.

For the population quantities, Cov(X, Y) = $\rho\, \sigma_x\, \sigma_y$.

If X and Y are random variables (as distinguished from data) with standard deviations σ_x and σ_y , then the standard deviation of $aX + bY$, where a and b are numbers, is

$$\text{SD}(aX + bY) = \sqrt{a^2\sigma_x^2 + 2ab\,\text{Cov}(X, Y) + b^2\sigma_y^2} = \sqrt{a^2\sigma_x^2 + 2ab\,\rho\sigma_x\sigma_y + b^2\sigma_y^2}\,.$$

This form is interesting for risk minimization. Suppose that you have investments $\mathbb{X}$ and $\mathbb{Y}$, and that random variables X and Y represent the returns, for which the standard deviations are σ_x and σ_y . Suppose that you decide to apportion your investments with proportion a going to $\mathbb{X}$ and proportion $b = 1 - a$ going to $\mathbb{Y}$. Then SD($aX + bY$) is the standard deviation of your return. The standard deviation of your return is minimized for $a = \dfrac{\sigma_y^2 - \rho\sigma_x\sigma_y}{\sigma_x^2 - 2\rho\sigma_x\sigma_y + \sigma_y^2} = \dfrac{\sigma_y^2 - \rho\sigma_x\sigma_y}{\left(\sigma_x^2 - \rho\sigma_x\sigma_y\right) + \left(\sigma_y^2 - \rho\sigma_x\sigma_y\right)}$. If a turns out to be 0.46, then 46% of the investment goes to $\mathbb{X}$ and 54% goes to $\mathbb{Y}$.

> This argument assumes that the expected value of $aX + bY$ is fixed. The entire discussion then centers around minimizing the variability, so we want the value of a which makes SD($aX + bY$) as small as possible.

This diversification always works convincingly if $\rho < 0$, in which case it is an obvious strategy. That is, you should always diversify in dealing the negatively correlated investments.

If the correlation ρ is zero or positive, there is still some benefit to diversification. If ρ = 0, then the investment strategy will be based on $a = \dfrac{\sigma_y^2}{\sigma_x^2 + \sigma_y^2}$.

For positive ρ, there are some interesting ways to break up special cases.

Case 1: Suppose that $\sigma_y < \sigma_x$. Then it will happen that $a = 0$ if $\rho = \dfrac{\sigma_y}{\sigma_x}$. Since $\sigma_y < \sigma_x$ this value of ρ is less than 1 and thus a real possibility. At this correlation $\rho = \dfrac{\sigma_y}{\sigma_x}$, 100% is going to investment 𝕐 (the less volatile investment).

> If $\rho > \dfrac{\sigma_y}{\sigma_x}$, the formula seems to ask you to make a *negative* investment in 𝕏, using the extra funds to buy more of 𝕐. Other issues now come into the problem, and we'll set it aside.

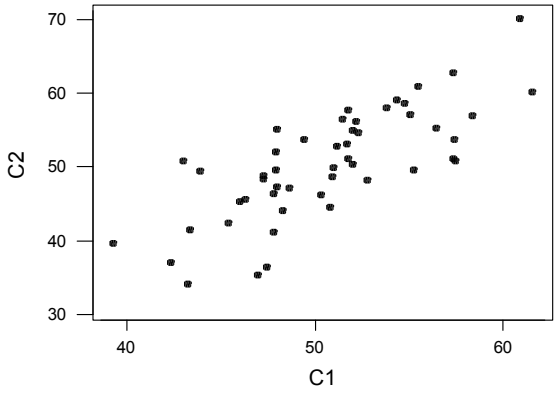
Case 2: Suppose that $\sigma_y = \sigma_x$. Then the function being optimized is $\mathrm{SD}(aX + bY) = \sqrt{a^2\sigma_x^2 + 2ab\rho\sigma_x\sigma_y + b^2\sigma_y^2} = \sqrt{a^2\sigma_x^2 + 2ab\rho\sigma_x^2 + b^2\sigma_x^2}$ $= \sigma_x\sqrt{a^2 + 2ab\rho + b^2}$. The optimum is $a = 0.50$, so that the investment is split 50-50 between 𝕏 and 𝕐. This is the optimum strategy no matter what the value of ρ. The optimized standard deviation is $\sigma_x\sqrt{\dfrac{1+\rho}{2}}$, so that you do better as ρ gets smaller.

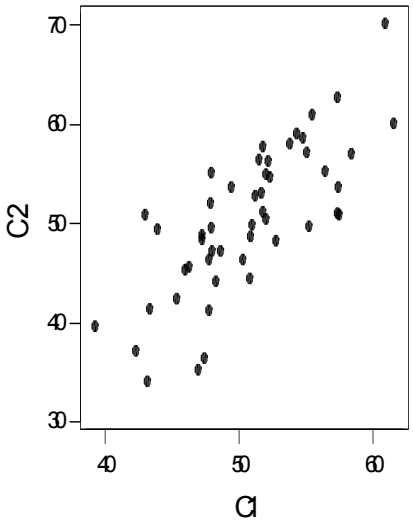
Case 3: This is the mirrored version of Case 1. Suppose that $\sigma_x < \sigma_y$. Then it will happen that $b = 0$ if $\rho = \dfrac{\sigma_x}{\sigma_y}$. Since $\sigma_x < \sigma_y$ this value of ρ is less than 1 and thus a real possibility. At this correlation $\rho = \dfrac{\sigma_x}{\sigma_y}$, 100% is going to investment 𝕏 (the less volatile investment).

> If $\rho > \dfrac{\sigma_x}{\sigma_y}$, the formula seems to ask you to make a *negative* investment in 𝕐, using the extra funds to buy more of 𝕏. Other issues are now involved, so we'll set this aside.

Here are some unusual things about correlation coefficients that seem to be guarded secrets.

* You cannot appraise correlation coefficients visually. The pictures below have a correlation coefficient of 0.763. In fact, the pictures are identical, except for the aspect ratio. But changing the aspect ratio alters the visual impression of clustering.





* Correlations are related to the rate of regression to mediocrity. In fact the estimated regression coefficient b_1 can be expressed as $b_1 = r \frac{s_y}{s_x}$. Recall also that the intercept can be written as $b_0 = \bar{y} - b_1 \bar{x}$. Write the fitted regression line as

$$\hat{Y} = b_0 + b_1 x = (\bar{y} - b_1 \bar{x}) + b_1 x = \bar{y} + b_1 (x - \bar{x})$$

$$= \bar{y} + r \frac{s_y}{s_x} (x - \bar{x})$$

One more simple manipulation gives

$$\frac{\hat{Y} - \bar{y}}{s_y} = r \frac{x - \bar{x}}{s_x}$$

Now suppose that you obtained a new x-point, call it x_{new}. You'd predict the corresponding Y, call it $\hat{Y}_{\text{new}}$, as $\hat{Y}_{\text{new}} = b_0 + b_1 x_{\text{new}}$. By exactly the logic above, this can be written as

$$\frac{\hat{Y}_{\text{new}} - \bar{y}}{s_y} = r \frac{x_{\text{new}} - \bar{x}}{s_x}$$

Now suppose that it happened that x_{new} was exactly one standard deviation above average, relative to the x's. That is, $\frac{x_{\text{new}} - \bar{x}}{s_x} = 1$. The obvious prediction for $\hat{Y}_{\text{new}}$ would be that it would then be one standard deviation above average, relative to the y's, meaning $\frac{\hat{Y}_{\text{new}} - \bar{y}}{s_y} = 1$. However, the regression prediction is not this! The regression prediction is $\frac{\hat{Y}_{\text{new}} - \bar{y}}{s_y} = r$. As $|r| < 1$, we are setting $\hat{Y}_{\text{new}}$ closer to the y-mean than x_{new} was to the x-mean. This phenomenon is called "regression to the mean." You will also see the phrase "regression to mediocrity."

Suppose that $x_1, x_2, \ldots, x_n$ is a list of values with mean $\bar{x}$. The (sample) variance of the x's is defined as

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The calculation of s_x^2 by hand or by computer is usually done through the formula

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

The square root is s_x, the sample standard deviation of the x's.

If we had matching random variables $y_1, y_2, \ldots, y_n$ then we could make parallel calculations. By matching, we mean that x_1 and y_1 are collected from the same data point, x_2 and y_2 are collected from the same data point, and so on. Thus, the sample variance of the y's is defined as

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

In parallel with the above, the calculation of s_y^2 by hand or by computer is usually done through the formula

$$s_y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]$$

The square root is s_y, the sample standard deviation of the y's.

We can compute one additional quantity which tells how the x's and y's tend to behave relative to each other. This quantity is the sample covariance, defined as

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

The calculation of s_{xy} is usually done through the formula

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i\, y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}\right]$$

Covariances are not that easy to interpret. They are calculated in product units; for example, if the x's are in dollars and the y's in tons, it happens that s_{xy} is in units of dollar-tons. Accordingly, statisticians routinely present this in the form of correlations. Specifically, we define the sample correlation between the x's and y's to be

$$r_{xy} = \frac{s_{xy}}{s_x\, s_y}$$

It is always true that $-1 \le r_{xy} \le +1$. Here are some quick interpretations for the correlation:

If $r_{xy} = +1$, then for some numbers a and b (with $b > 0$), it happens that $y_i = a + b\, x_i$. That is, a plot of the data would show that all points lie on a straight line of positive slope.

If $r_{xy} = -1$, then for some numbers a and b (with $b < 0$), it happens that $y_i = a + b\, x_i$. That is, a plot of the data would show that all points lie on a straight line of negative slope.

If $r_{xy} = 0$, then there is a complete absence of a straight-line relationship between the x's and the y's. A plot of the data would show aimless scatter, though it occasionally happens that a non-linear (curved) relationship corresponds to a correlation of zero. In practice, sample correlations are rarely exactly zero, though they can be very close to zero.

In-between values of r_{xy} are taken as measures of the strength of the relationship. Thus a value $r_{xy} = 0.31$ would indicate a weak positive relationship between the x's and the y's, while a value $r_{xy} = -0.82$ would indicate a fairly strong negative relationship.

There is almost nothing in statistics that is quite as mysterious as the regression effect. Suppose that your business operates a chain of franchise stores which are very similar is size, appearance, and functionality. Suppose that the numbers below represent the gross sales for year 1999 at a sample of 10 of these stores:

10.4 8.2 11.6 10.8 10.2 7.9 9.4 8.8 11.0 10.1

These figures are in millions of dollars. The mean value is 9.84, and the standard deviation is 1.2295. You believe that the gross sales for year 2000 should grow at a 10% rate. Predict the gross sales for each of the ten stores in the sample.

The naive solution here is simply to say that each store's sales will grow by 10%. That is

Predict that the store with 10.4 will grow to 10.4 + 1.04 = 11.44
Predict that the store with 8.2 will grow to 8.2 + 0.82 = 9.02
Predict that the store with 11.6 will grow to 11.6 + 1.16 = 12.76
… and so on

This ignores the possibility that stores with good numbers in 1999 were benefited not only by favorable business conditions (such as location) but were also the recipients of lucky statistical randomness. It would seem reasonable to predict that such stores would not do so amazingly well in year 2000.

Let X represent the sales of a randomly-selected store in 1999, and let Y be the corresponding sales in 2000. Obviously the objective is to use X to predict Y, and we must consider the regression relationship to do this. The population version of the regression relationship, meaning the *expected* value of Y in terms of X, is given as

$$\hat{Y} = \beta_0 + \beta_1 X$$

where $\beta_1$is the true regression slope and $\beta_0$is the true regression intercept. (The "hat" symbol on Y notes that this is an expected Y rather than an actual value.) The true relationship is not known, but with data we will get the estimated relationship as

$$\hat{Y} = b_0 + b_1 X$$

It's much more instructive to write the population version of the regression relationship as

$$\frac{\hat{Y} - \mu_Y}{\sigma_Y} = \rho \frac{X - \mu_X}{\sigma_X}$$

where μ_Y are σ_Y are the mean and standard deviation of Y, and where μ_X and σ_X are the mean and standard deviation of X. The symbol ρ is the population correlation between X and Y.

Yes, there is relationship between β_0 and β_1 and these symbols. Specifically, $\beta_1 = \rho \frac{\sigma_Y}{\sigma_X}$ and $\beta_0 = \mu_Y - \beta_1 \mu_X$.

As usual, true relationships are estimated from data. The data-based fitted regression estimate is

$$\frac{\hat{Y} - \bar{y}}{s_Y} = r \frac{X - \bar{x}}{s_X}$$

If we have a new X, call it X_{new}, then the corresponding prediction for Y, call it $\hat{Y}_{\text{new}}$, obeys the relationship

$$\frac{\hat{Y}_{\text{new}} - \bar{y}}{s_Y} = r \frac{X_{\text{new}} - \bar{x}}{s_X}$$

(We use the symbol X_{new} to suggest that the prediction could be made at an X value which does not appear in our original data set. Predictions can, of course, also be made at the X values we've already got – as long as we have not yet seen the Y values.)

Suppose that the value of X_{new} is (relative to $\bar{x}$ and s_X) 1.4 standard deviation above average. That is, suppose that

$$\frac{X_{\text{new}} - \bar{x}}{s_X} = 1.4$$

It would seem natural to predict a Y that was also 1.4 standard deviation above average. That is, one would think that

$$\frac{\hat{Y}_{\text{new}} - \bar{y}}{s_Y} = 1.4$$

However, the fitted relationship has a correlation r in its formula. Thus our real prediction is that

$$\frac{\hat{Y}_{\text{new}} - \bar{y}}{s_Y} = 1.4\, r$$

Observe that X_{new} is somewhat high, and we also predict the new Y to be high, *but not quite as high as the new X*. If, for example, $r = 0.4$, then our prediction is

$$\frac{\hat{Y}_{\text{new}} - \overline{y}}{s_Y} = 0.56$$

Suppose that a man is 6′ 9″ tall, and suppose that he has a son. What height would you predict that this son would grow to? Most people would predict that he would be tall, but it would be quite unusual for him to be as tall as his father. Perhaps a reasonable prediction is 6′ 6″. We are expecting the son to "regress" back to mediocrity, meaning that we expect him to revert to average values.

This is of course a prediction on average. The son could well be even taller than his father, but this is certainly less than a 50% chance.

This regression effect applies also at the other end of the height scale. If the father is 5′ 1″, then his son is likely to be short also, but *the son will probably be taller than his father*.

The regression effect is strongest when far away from the center. For fathers of average height, the probability is about 50% that the sons will be taller and about 50% that the sons will be shorter.

This logic can be modified even if improved nutrition tends to increase the heights of all the sons. In the example that started this document, improving business conditions is doing similar things for the stores!

The regression effect is everywhere.

> People found to have high blood pressure at a mass screening (say at an employer's health fair) will have, on average, lower blood pressures the next time a reading is taken.
>
> Mutual fund managers who have exceptionally good performances in year T will be found to have not-so-great performances, on average, in year $T + 1$.
>
> Professional athletes who have great performances in year T will have less great performances, on average, in year $T + 1$.

The example that started this discussion seems to lack the structure to actually make the numeric predictions. We can, however, get something useful.

If we let the current information be denoted by X, then we've obtained

$\bar{x} = 9.84$ an estimate of μ_X

$s_X = 1.2295$ an estimate of σ_X

If Y denotes the values for the following year, then it would seem reasonable to do this:

$1.10\ \bar{x} = 10.824$ an estimate of μ_Y

$1.10\ s_X = 1.35245$ an estimate of σ_Y

There are some people who would use s_X itself as an estimate of σ_Y, but it seems more reasonable that the standard deviation should grow as the mean grows.

This lets us almost complete a fitted regression relationship:

$$\frac{\hat{Y} - 10.824}{1.35245} = r\,\frac{X - 9.84}{1.2295}$$

This lacks a value for r, the sample correlation between X and Y. Since we have never observed even a single Y, we have no value for r. However, we might be able to make a reasonable guess. Perhaps a good value for r would be 0.60. This ought to be a plausible correlation for data of this type. If we use this value, then the fitted relationship is

$$\frac{\hat{Y} - 10.824}{1.35245} = 0.60\,\frac{X - 9.84}{1.2295}$$

What prediction would correspond to the X value of 11.6, the best store in the sample? If you place $X = 11.6$ into this equation, you get $\hat{Y} = 11.9856$. Your prediction calls for growth of 0.3856 at this store, from 10.824 to 11.9856. This is an increase all right, but it is about 3.32% ell below 10%.

Consider the store with 7.9. If you place $X = 7.9$ into this equation, you get $\hat{Y} = 9.5436$. This predicts a growth of 1.6436, about 20.81%, which is considerably more than 10%.

How about the store with 10.1, very slightly above the X average? If you place $X = 10.1$ into this equation, you get $\hat{Y} = 10.9956$. This represents an increase of 0.8956, or $\frac{0.8956}{10.1} \approx 0.0887 = 8.87\%$. That is, this store started just slightly above average and its predicted growth ended up just below the target 10%.