9 Descriptive Statistics: Regression Analysis

Homer: Hello, Police? Are you sitting down? Good! I wish to report a robbery.
Wiggum: [bored] A robbery, right. Thanks for the report. [hangs up] That's another one, Lou...723 Evergreen Terrace. [Looks at a map with the robbery locations marked on it] Well, there doesn't seem to be any pattern yet, but if I take this one and move it here...and I move these over here...hello! It almost looks like an arrow!

From: *The Simpsons*

9.1 Introduction

In the last chapter we looked at data with an eye to determining if a linear correlation exists. If a scatter plot of the data suggests such a relation and the correlation coefficient confirms it, we might then want to consider how we could quantify the relationship.

For example, the following scatter plot shows the grades of 92 students on the first and second exams in a course. This data set is contained in the file *exam grades.xls*.

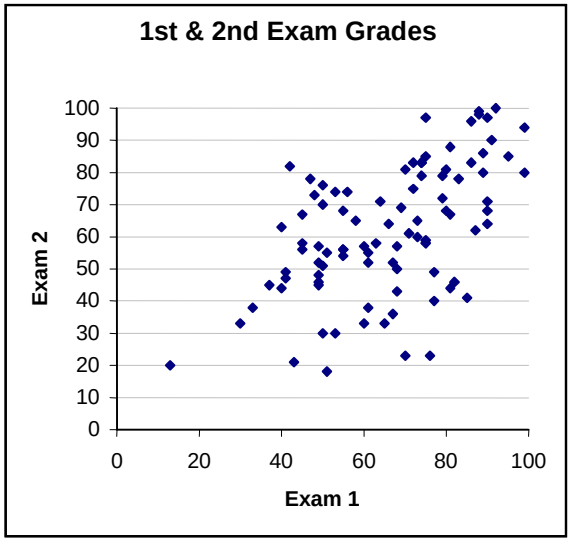


Figure 9.1

The correlation coefficient r equals 0.5532, indicating a modest positive association. Can the data be used to predict a score on exam 2, given the score on exam 1? If so, how accurately will our prediction mirror the actual performance of an individual student? When we attempt to predict the value of one variable knowing the value of the other, we are performing a *regression analysis*. In doing so we specify one variable as the independent variable (grade on exam 1) and the other variable (grade on exam 2) as the dependent variable. In the regression analysis the two variables are treated differently. If we wish to switch the roles of the variables (for instance, suppose a

student's grade on exam 1 was lost and we want to estimate it based on the grade on exam 2) then we will have to do a different regression analysis.

9.1 Method of Least Squares

In doing regression analysis we try to find the equation of a line which best fits the linear trend in the data. In order to obtain a precise and unequivocal answer we must specify what we mean by "best fit". The most commonly used criterion is the so-called least squares fit. Let's explain how this works.

Example 9.1: Find the line that best fits the points (1, 1), (3, 12), and (5, 10) in the sense of least squares.

Solution:

The points are shown in the figure below (•) with a possible line as a candidate for the best fit.

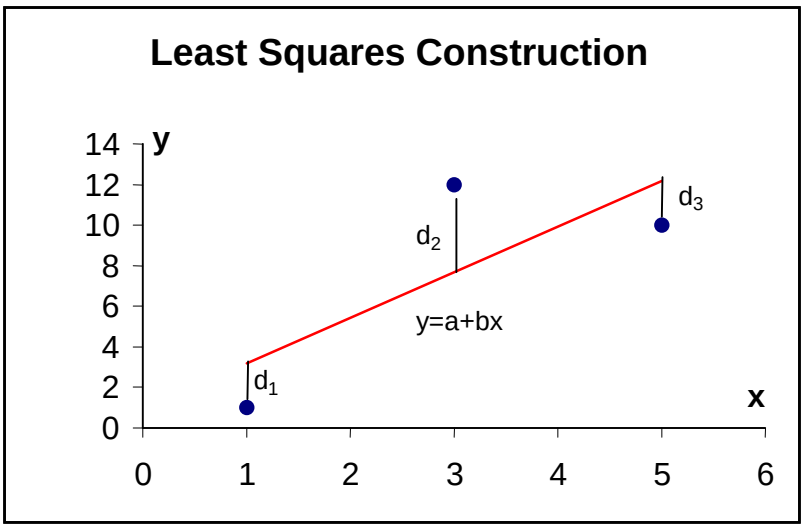


Figure 9.2

The vertical segments from the three points to the line are labeled d_1, d_2, and d_3. These represent the vertical distances from the given points to the candidate line. If the points were exactly on the line these distances would all be zero. That would be a perfect fit. As there is no line that passes through all three points, we might choose to measure the fit by adding the individual distances and selecting the line for which the sum of the vertical distances is as small as possible. This is quite reasonable, but the mathematics is somewhat less tractable than the alternative approach we consider below.

The situation is similar to difficulties we encountered in defining the standard deviation in Chapter 7. A similar remedy can be used. Namely, we define the fitness of the line as the sum of the

squares of the distances d_1, d_2, and d_3. In symbols, using L to stand for the candidate line, we have $fitness(L) = d_1^2 + d_2^2 + d_3^2$. The line L has an equation of the form $y = a + bx$. (The reader may recall the convention in algebra of writing the equation of a line using the symbolic form $y = mx + b$. Unfortunately, we have already used m as the median and statistical practice has enshrined the use of b for the slope, rather than for the intercept.) We can obtain a formula for the fitness in terms of the unknown coefficients a, and b. To do this we first find expressions for d_1^2, d_2^2, and d_3^2.

- d_1: The data point is (1, 1). We need the point on the line that has the same abscissa $x = 1$. This is the point $(1, a+b)$, where the second coordinate is obtained by putting $x = 1$ in the equation of the line $y = a + bx$. The absolute difference of the ordinates gives the vertical distance between the points so $d_1 = |a + b - 1|$ and $d_1^2 = (a + b - 1)^2$.

- d_2: The data point is (3, 12). The point on the line is $(3, a+3b)$. As above, we obtain that $d_2^2 = (a + 3b - 12)^2$.

- d_3: The data point is (5, 10). The point on the line with the same abscissa is $(5, a+5b)$. Therefore, $d_3^2 = (a + 5b - 10)^2$.

We can now write a formula for the fitness function in terms of a and b.

$$fitness(a,b) = (a + b - 1)^2 + (a + 3b - 12)^2 + (a + 5b - 10)^2 \qquad (9.1)$$

Our problem is to find values of a and b that minimize this expression. Looking at Figure 9.2 you might think that selecting the line joining the points (1, 1) and (5, 10) would produce the smallest value of the fit. Indeed, for that line the first and third terms will be zero (why?). The equation of the line in question is $y = -1.25 + 2.25x$, so $a = -1.25$ and $b = 2.25$. Substituting these values for a and b in the fitness formula gives $fitness(-1.25, 2.25) = (6.5)^2 = 42.25$. However, we can do better than this. Raising the line a bit so its equation is $y = 1 + 2.25x$, with $a = 1$ and $b = 2.25$, gives a fitness value of 28.1875. Is this the smallest value of the fitness expression or can we find a line that produces a smaller value?

Such a question should sound familiar. In calculus you learned how to find the minimum value of functions using the derivative. We have a similar problem here except the expression we need to minimize is a function of two independent variables a and b, instead of the single variable problems you encountered in calculus. Nevertheless, those methods can be applied, although instead of setting the ordinary derivative equal to zero, we must compute two partial derivatives and set each to zero to obtain two equations for a and b. We will not go into the details. It can be done and it's not that hard (the equations turn out to be first degree equations in the two unknowns). The result in this example is the line $y = .9167 + 2.25x$. It is the line shown in Figure

9.2. The fitness value for this line is 28.167, slightly smaller than the value for the line $y = 1 + 2.25x$.■

The discussion of Example 9.1 can be summarized as follows. We have n data points $(x_1, y_1), \ldots, (x_n, y_n)$. The line that bests fits these points in the sense of least squares is specified by the following definition.

Definition 9.1: The *least squares regression line* of y on x for the data $(x_1, y_1), \ldots, (x_n, y_n)$ is the line $y = a + bx$ for which the sum

$$(a + bx_1 - y_1)^2 + (a + bx_2 - y_2)^2 + \cdots (a + bx_n - y_n)^2$$

has its smallest possible value.■

The computation of the coefficients in the regression equation requires the use of calculus. The result, however, is a purely algebraic formula that we summarize in

Theorem 9.1: The regression line of y on x is determined by the following properties:

a) The slope of the regression line (b in the expression $y = a + bx$) is given by $b = r\dfrac{s_y}{s_x}$, where r is the correlation coefficient and s_x and s_y are the standard deviation of the x and y values respectively.

b) The regression line passes through the point $(\overline{x}, \overline{y})$.■

The slope of the regression line $r\dfrac{s_y}{s_x}$ is usually abbreviated $\hat{b}$ (read "b hat"). Properties a) and b) enable us to write the equation of the regression line using the point-slope method.

Theorem 9.2: The equation of the regression line of y on x is given by any of the following equivalent forms

a) $\dfrac{y - \overline{y}}{x - \overline{x}} = \hat{b}$

b) $y - \overline{y} = \hat{b}(x - \overline{x})$

c) $y = (\overline{y} - \hat{b}\overline{x}) + \hat{b}x$

Proof:

Form a) is the direct statement of the equation of a line using the point-slope method, with slope $\hat{b}$ and point $(\overline{x}, \overline{y})$. Form b) follows immediately from this and c) is the expanded form of b) in

which we solve for y in terms of x. Note that form c) has the shape $y = a + bx$ so that the y intercept of the regression line, $\hat{a}$, is given by $\hat{a} = \overline{y} - \hat{b}\overline{x}$.■

For the record we state the last part of the proof as a formal property.

Corollary 9.1: The y intercept of the regression line of y on x is given by the formula $\hat{a} = \overline{y} - \hat{b}\overline{x}$, where $\overline{x}$ is the mean of the x values, $\overline{y}$ the mean of the y values and $\hat{b}$ is the slope of the regression line.■

Example 9.2: Find the regression line of y on x for the three data points given in Example 9.1.

Solution:

We organize the most important ingredients of the calculation as a table.

x	y	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})(y-\overline{y})$
1	1	-2	-6.67	13.34
3	12	0	4.33	0
5	10	2	2.33	4.66

Using columns 1 and 2 we find that $\overline{x} = 3$ and $\overline{y} = 7.67$ (rounded to three significant figures). We then compute $s_x = \sqrt{\frac{(-2)^2 + 0^2 + (2)^2}{2}} = 2$ and similarly $s_y = 5.86$. The covariance $c_{xy} = \frac{13.34 + 0 + 4.66}{2} = 9$, which gives for the correlation coefficient $r = \frac{c_{xy}}{s_x s_y} = \frac{9}{(2)(5.86)} = .768$.

For the slope of the regression line we obtain $\hat{b} = \frac{rs_y}{s_x} = \frac{(.768)(5.86)}{2} = 2.25$. From Corollary 9.1 the intercept is given by $\hat{a} = \overline{y} - \hat{b}\overline{x} = 7.67 - (2.25)(3) = .92$. Thus the regression line is $y = .92 + 2.25x$. Except for some slight error due to round off, this agrees with the formula given at the end of Example 9.1, which was computed directly using *Excel*.■

9.2 Prediction

A regression line is computed to predict unknown y values from known x values. Making the prediction is trivial; understanding what it means is not.

Example 9.3: Referring to the data plotted in Figure 9.1, if a student scores 75 on the first exam what grade might we predict for the second exam?

Solution:

For this data it turns out that

$\bar{x} = 65.85 \qquad s_x = 17.89 \qquad \bar{y} = 61.83 \qquad s_y = 20.26 \qquad r = .5532.$

The regression line has equation $y = 20.6 + .626x$ (verify this is accurate to three significant figures). If we knew nothing about the student's performance on exam 1 then the best prediction we could make would be the mean $\bar{y}$ or 61.83. But we know the student did above average on the first test and since there is a modest positive correlation, we would expect the student to do better than average on the second exam. The regression line provides our estimate. The predicted grade is $20.6 + (.626)(75) = 67.55$, obtained by substituting $x = 75$ in the regression equation. We denote this predicted value by $\hat{y}$. This notation serves to remind us that this is a predicted value of the grade on exam 2 using the regression line and not the actual value y that the student finally achieves. We will sometimes emphasize this by writing the equation of the regression line in the form $\hat{y} = 20.6 + .626x$. ■

Having made a prediction we would like to know what it means and how good it is. At this point our explanation will be somewhat impressionistic. To make the discussion more rigorous requires tools from probability theory that we will gradually develop in the next several chapters. The reader can then consult any of the more specialized texts on statistics for an elaboration of this topic.

The scatter plot in Figure 9.1 has been redrawn below, with the regression line added.

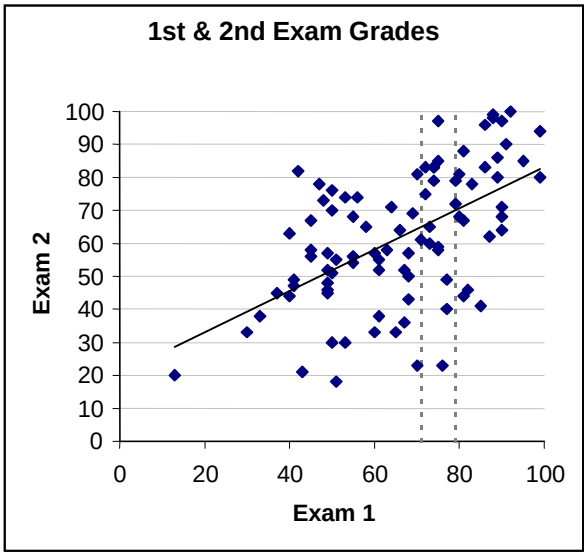


Figure 9.3

The vertical dashed lines enclose all data values for which the score on exam 1 was between 71 and 79. The table below lists all the data points contained between these vertical lines.

Exam 1	Exam 2	Exam 1	Exam 2	Exam 1	Exam 2	Exam 1	Exam 2
71	61	74	79	75	97	79	79
72	75	74	83	75	59	79	72
72	83	74	83	76	23		
73	65	75	58	77	40		
73	60	75	85	77	49		

Table 9.1

The second exam grades range from 23 to 97. The average of these 17 grades is 67.7. Notice that this is quite close to the value $\hat{y} = 67.55$ on the regression line for $x = 75$. This leads to the first regression property.

Regression Property 9.1: The value $\hat{y}$ predicted from the regression equation for a given $\tilde{x}$ approximates the average of the y values that would be attained for a large number of different observations whose x values are near the specified value $\tilde{x}$.■

Thus $\hat{y}$ is the prediction of an average. For students scoring around 75 on the first exam, the $\hat{y}$ value of 67.55 is a prediction of their average performance on the second exam. From our discussion of univariate data, however, we would like to know not only the predicted average value, but some estimate for the spread of the actual values around the prediction.

If we know nothing about the first exam score then we know that the spread of the y values in general is given by $s_y = 20.26$. We would expect that knowing the first exam grade would reduce this spread, because the first exam is some indicator of a student's ability. In fact, as we have seen above, knowing the score x on the first exam, the second exam score will vary around the corresponding value $\hat{y}$ on the regression line. We need then a measure of how far an actual y value might lie from the point $\hat{y}$ on the regression line with the same x value. The following example illustrates the point.

Example 9.4: Find the standard deviation for the 17 scores on exam 2 given in Table 9.1 and discuss the relationship with the overall spread of the scores on exam 2.

Solution:

Consider again the data in Table 9.1. For the 17 data values in that set we found a mean of 67.7. The standard deviation of these data values from this mean is 18.68. For the entire set of exam 2 grades the standard deviation around the mean of 61.8 is $s_y = 20.3$. Knowing the first exam score reduces the variability in the second exam score. Part of the latter variability is due to the tendency of the second exam to reflect the performance on the first exam. The regression line

captures this portion of the variability, leaving only an unexplained variation around the regression line.■

We call the unexplained variation of the y values around the regression estimate $\hat{y}$, *the standard deviation of y given x*. We denote this by $s_{y|x}$. Fortunately, it is possible to estimate this quantity. Section 9.7 provides some additional mathematical detail.

Regression Property 9.2: The value of $s_{y|x}$ is approximately $\sqrt{1-r^2}s_y$, where r is the correlation coefficient and s_y is the standard deviation of the y values about their mean.■

Regression Property 9.2 says that the variation in the data as measured by s_y is reduced by the regression. The amount of the reduction is determined by the factor $\sqrt{1-r^2}$. Although we omit a derivation of Regression Property 9.2, the extreme cases are plausible and lend some belief to the validity of the estimate. First, if $r \approx \pm 1$, the data points have strong linear correlation so we expect little spread around the $\hat{y}$ values predicted by the regression equation. In this situation we should have $s_{y|x}$ close to zero, as the formula indicates. When r is close to ± 1 knowing the x value greatly reduces the uncertainty in predicting the y value. On the other hand, if $r = 0$ then there is no correlation between the x and y values and knowing the value of x provides no additional useful information in pinning down the value of y. In other words $s_{y|x} \approx s_y$, as stated by the regression property.

Example 9.5: Use Regression Property 9.2 to estimate $s_{y|x}$ and compare the result to the standard deviation computed in Example 9.4**.**

Solution:

From Regression Property 9.2 the value of $s_{y|x}$ is approximately $\sqrt{1-r^2}s_y$, which using $r = .55$ and $s_y = 20.3$, gives $s_{y|x} \approx 17$. For the data values in Table 9.1 the actual standard deviation of 18.68 for the y values is slightly above our prediction, but within the statistical limits of expected variation.■

Regression Property 9.2 shows the importance of the quantity r^2 in regression analysis. The closer r^2 is to one, the smaller will be the value of $s_{y|x}$ and hence the more tightly the points will cluster around the regression line. r^2 is called the *coefficient of determination.* It determines by how much the spread in the dependent variable is reduced by the regression. This quantity should always be reported when presenting a regression analysis.

One final comment regarding prediction. ***The estimates provided by*** Regression Property 9.1 ***and*** Regression Property 9.2 ***should only be used when the value of x is between the smallest and***

largest x ***values in the data***. Using values of x beyond this region is called *extrapolation*. Estimates for $s_{y|x}$ in such cases may be considerably larger than that provided Regression Property 9.2. In general, more exact estimates of $s_{y|x}$ show that it depends somewhat on x, getting larger as x becomes further from the mean $\overline{x}$.

9.3 The Regression Effect

Let's consider again the grade data plotted in Figure 9.1. Table 9.2 shows the performance of the top 10 students on exam 1 and their grades on exam 2.

Exam 1	89	90	90	90	90	91	92	95	99	99
Exam 2	86	68	71	97	64	90	100	85	80	94

Table 9.2

The average grade for these students on exam 1 was 92.5, while on exam 2 the average fell to 83.5, a 9 point drop, although still well above the average for exam 2 of 61.8. Perhaps these students became complacent after doing so well on the first exam. Let's look at what happened on exam 2 to the students who were the lowest scoring on exam 1.

Exam 1	13	30	33	37	40	40	41	41	42	43
Exam 2	20	33	38	45	63	44	49	47	82	21

Table 9.3

The average grade for these students increased from 36 to 44.2, more than 8 points, although still quite below the average grade of 61.8 for exam 2. You might conclude that these students worked harder after doing so poorly on the first exam, or perhaps the instructor made an extra effort to assist them. Finally, let's consider 10 students who scored slightly above the average of 65.8 on the first exam.

Exam 1	70	70	71	72	72	73	73	74	74	74
Exam 2	81	23	61	75	83	65	60	79	83	83

Table 9.4

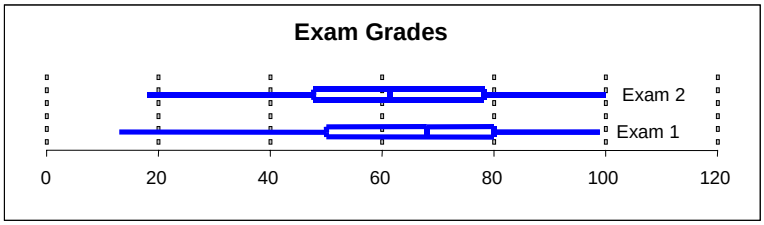
For these students the average went from 72.3 to 69.3, so they stayed slightly above average, although on the whole there was not a lot of change. In summary, it appears that the better performing students on the first exam did worse on the second, average students stayed about average and weaker students picked up a little bit. In short, to use the terminology of Sir Francis Galton who first observed this effect, there was a regression (moving back) to the mean. We speculated above that this behavior was a consequence of the students' or the instructor's efforts. In fact, it is a statistical artifact created precisely because the exam scores are not highly correlated. Let's explain.

Figure 9.3 shows that the data is spread around the regression line, so the first exam score does not have a very high predictive value for the second score. If a student begins with a high score on the first exam the second exam score can either be bigger or smaller. However, if you've scored say 95 on the first exam, random variation from that score is much more likely to produce a lower grade than a higher one. Indeed in Table 9.2, seven out of the ten top performers on exam 1 scored lower on exam 2. Similarly if you did poorly on exam1 it's likely your next grade will be better. Indeed this is the case for nine out of the ten students listed in Table 9.3. For students scoring about average to begin with, their second exam grade is likely to have an equal chance of increasing or decreasing. In Table 9.4, six out of ten students saw an increase in grade.

We summarize the regression effect below. Exercise 15 gives a more detailed mathematical justification than the argument sketched above.

The Regression Effect (Regression to the Mean): If a bivariate data set shows only modest correlation and we consider x values that are large or small compared to the mean, then the corresponding y values will tend to be less extreme.■

It is important to understand that the Regression Effect does not tell us anything about the overall distribution of y values compared to the distribution of x values. In fact, for the exam grades the comparative box plots shown below indicate that the individual distributions are almost identical. Although the top performers on exam 1 on the whole do worse on exam 2, students who moved up in their performance have taken their place. In short, the Regression Effect is a consequence of the variation of scores for the individual, not of a variation in the ensemble.



The Regression Effect can be, and often is, misinterpreted as a real effect due to a treatment. For example, suppose different instructors had taught the class during the period leading up to each of the exams. The data in Table 9.2 might be presented as evidence that the second instructor was ineffective in teaching the better students, since their performance obviously declined during that period. However, unless the exam scores are usually highly correlated, this observation can be explained by the Regression Effect. ***Incorrectly interpreting the result of the Regression Effect as a real difference due to a treatment is sometimes called the Regression Fallacy***. This fallacy is particularly prevalent when the subjects being studied are chosen because in one instance they scored at the high or low end for some variable, for example blood pressure, that exhibits fairly wide variation over time.

9.4 Transforming Data

In this and the last chapter we have considered linear or straight-line relationships between variables. The correlation coefficient measures the strength of such linear association and the regression line provides the best possible linear fit. We have seen though, for instance Chapter 8, Example 4b, that a scatter plot of a data set may reveal a non-linear pattern that looks fairly regular, but cannot be captured by the tools we have discussed. In this section we show how data can sometimes be transformed so that the methods of regression analysis, in particular the least squares method for fitting a line, can be applied to find an approximate fit for non-linear data. We consider an important example from astronomy.

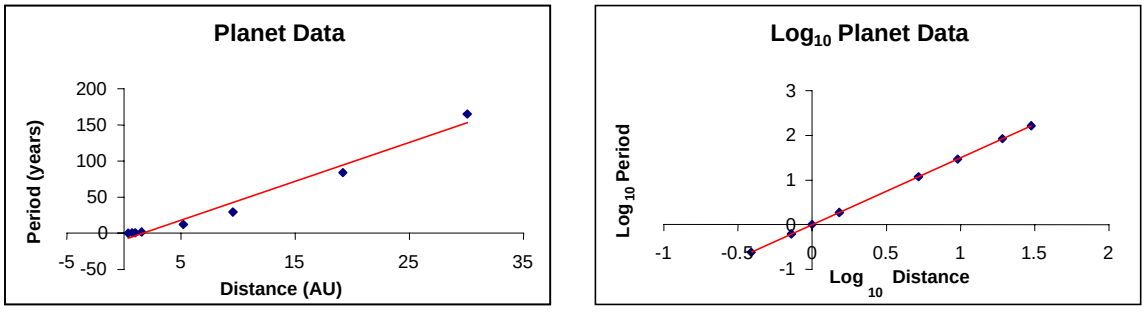
Example 9.6: Use a logarithmic transformation of variables and least squares regression to derive Kepler's law relating the period of a planet's orbit to its distance from the sun.

Solution:

The following table gives the period of revolution (P in years) of each planet around the sun and its distance (D) from the sun, more precisely the length of the semi-major axis of its orbit. The units of the latter are so-called astronomical units (AU), where the semi-major axis of the Earth's orbit equals 1 AU.

Planet	Mercury	Venus	Earth	Mars	Jupiter	Saturn	Uranus	Neptune
D (in AU)	.387	.723	1.000	1.524	5.203	9.539	19.18	30.06
P (in years)	.241	.615	1.000	1.880	11.86	29.46	84.01	164.8

Panel (A) of Figure 9.4 below shows a scatter plot of the data together with the regression line. The correlation coefficient is .988. However, the excellent fit is rather an illusion. The equation of the regression line is approximately $P = 5.4D - 8.8$. For any of the planets from Mercury to Mars this equation produces negative estimates for the period. As you can see from the scatter plot, the first four data points appear almost as one point compared to the much larger data values determined by the outer planets. When there are only a few points whose coordinates are significantly larger than the others, the least squares procedure will be controlled by these "high leverage" points and may yield a line whose excellent statistical fit is of little practical value.



(A) (B)

Figure 9.4

In this example we notice that the numerical values for the coordinates of the points span two and three orders of magnitudes, i.e. the largest value of D is about 100 = 10^2 times as large as the smallest, and the largest value of P is about 1000 times as large as the smallest. In such cases a linear relationship is not really compatible with the variation in order of magnitude. In a linear relation, scaling the independent variable by a factor of 100 should scale the dependent variable by roughly the same multiple.

We can often gain insight into such data by computing the common logarithm ($\log_{10}$) of each data value. This will place all the data values on a comparable numerical scale.

Planet	Mercury	Venus	Earth	Mars	Jupiter	Saturn	Uranus	Neptune
$\log_{10}(D)$	-0.412	-0.141	0.000	0.183	0.716	0.980	1.283	1.478
$\log_{10}(P)$	-0.618	-0.211	0.000	0.274	1.074	1.469	1.924	2.217

A scatter plot (Figure 9.4(B)) shows the strong linear relationship between the logarithmic values. The regression equation in the new coordinates is (rounding to two decimal places) $\log(P) = 1.50\ \log(D)$, with $r = 1.00$. From the logarithmic relation, we obtain $P = D^{1.5}$ or $P^2 = D^3$, a relationship first discovered empirically by Kepler (often referred to as his 3rd law) and later derived by Newton from his laws of motion and the inverse square law of gravitation.■

In the last example we used the technique of transforming the data to discover an exact physical law. The particular transformation involved comparing the common logarithms of the dependent and independent variables. Note that this requires that both variables take on only positive values. The resulting plot is called a *log-log* plot. By exponentiating we find that a linear relation in the log-log plot corresponds to a power relation in the original variables.

Theorem 9.3: If x and y are variables for which a log-log plot shows a linear relationship, then a scatter plot of the quantities x and y will follow a curve of the form $y = cx^b$. Conversely, data that follows a power law $y = cx^b$ will yield a log-log plot that is a straight line of slope b.

Proof:

If the log-log plot shows a straight line relationship then the variables $\log(y)$ and $\log(x)$ must be related through an equation $\log(y) = a + b\log(x)$. Since these are base 10 logarithms, we deduce that

$$y = 10^{\log(y)} = 10^{a+b\log(x)} = 10^a 10^{b\log(x)} = cx^b,$$

as stated in the theorem. To prove the converse, apply the logarithm function to both sides of the equation $y = cx^b$, using that $\log(cx^b) = \log(c) + b\log(x)$.■

When x and y are biological quantities, the power relation $y = cx^b$ is often called an *allometric* relation. Such relationships have been found to hold empirically for many biological variables, for example

- species abundance vs. island size
- average body weight vs. average metabolism rate across species
- average brain weight vs. average body weight across species (see exercise 21)
- size of an organ vs. the overall size of body across the growth period of an organism

If the exponent b in an allometric relation is different from one, then the quantities x and y will grow at different rates, hence the term "allometric". For example, if $b = 0.3$ then doubling x will cause y to increase by a factor of $2^{0.3} \approx 1.23$.

The primary goal in data transformation is to find a representation in which the transformed data exhibits a linear pattern, since we have many tools for analyzing such situations, not least of which is our visual system. Finding an appropriate transformation is a matter of exploration and often ingenuity. In fact, there may be no reasonable way to produce a linear relation from the data. The following theorem summarizes the results obtained using some of the most common transformations besides the log-log plot already discussed. The proof is left to the reader.

Theorem 9.4: Suppose y and x are numerical variables for which:

a) *(semi-log plot)* a plot of $\log(y)$ vs. x shows a linear relation, then $y = c10^{bx}$ and conversely.

b) (*reciprocal plot*) a plot of y vs. $1/x$ shows a linear relation, then $y = a + b/x$ and conversely.■

In this section we have used linear regression as a tool to perform non-linear curve fitting. The resulting expressions can be used for predictive purposes as we did in section 9.2, but providing a correct interpretation of the answers is considerably more difficult in this case.

9.5 Tech Notes

Adding a regression line to a scatter plot is quite simple in *Excel*. Follow the steps described below.

Example 9.7: Adding a regression line to a scatter plot.

Solution:

1. Click the left mouse button with the arrow touching one of your data points. This selects the data.

2. Click the right mouse button and select the item, *Trendline*…. from the pop-up menu. *Excel* refers to a regression line as a "trendline" (The right mouse button brings up a menu of options that is relevant to the selected item).

3. Select Linear from the type choices. Click on the Options tab and check the two boxes marked *Display equation on chart* and *Display R-squared value on chart*.

4. After the equation and trendline have been drawn you can, if necessary, select the written equation and value of r^2 and drag them to a part of the plot area in which they will be more legible.

If you want the equation of the regression line without bothering to plot the data, (not recommended) you can use the command *=slope(y range, x range)* to find the slope of the regression line. The intercept can be found using the command *=intercept(y range, x range).* Note the order of the *y* and *x* variable is important here (unlike the correlation coefficient). The *y* range appears first if we are doing a regression of *y* on *x*.■

Data transformations are readily carried out in *Excel* and the transformed data can be plotted in the usual way. Figure 9.4(B) was obtained in that way. There is an alternative approach that is particularly convenient when exploring the suitability of logarithmic transformations. If you have already constructed a scatter plot of variables *y* vs. *x* you can replot the graph using logarithmic coordinates on the *x* and/or *y* axes. Follow these steps.

Example 9.8: Transforming a scatter plot to a log-log or semi-log plot.

Solution:

1. Click on the x (or y) axis to select it.

2. Click the right mouse button to bring up the pop-up menu. Select *Format Axis.*

3. Click the tab labeled "*Scale*" and then check the box marked "*Logarithmic Scale*".

Note that it is necessary that the axis you wish to transform via logarithms have only positive values. Repeating this for each axis will produce the log-log graph below for the planet data described in Example 9.6.

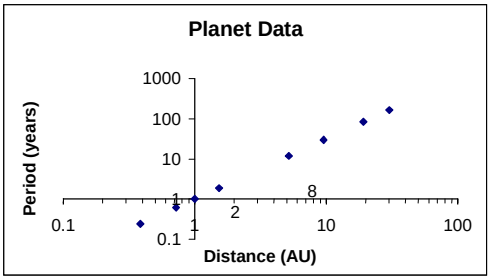


Figure 9.5

Figure 9.5 is similar to Figure 9.4(B), except for the peculiar scale on the axes. This scale is called a logarithmic scale. In this scale a unit of length represents a power of 10 so there is equal spacing between successive powers. Intermediary values are located according to the value of their logarithm. For example, the tickmark for 2 is placed at the position $\log_{10}(2) \approx 0.3$ units along the horizontal scale. The tickmark for $8 = 2^3$ is at 0.9, three times as far from the origin.

Before the widespread availability of computer graphing programs, it was very common for researchers who needed to use logarithms in their data analysis to use log-log or semi-log graphing paper, whose grid lines were prepared in this way. With such paper, the user does not have to compute the logarithms of the data values to create log plots. The scale on the paper does this automatically. For the same reason, the axes in Figure 9.5 are not described as $\log_{10}(\text{Distance})$, as was done in Figure 9.4(B). In Figure 9.5 we read the actual numerical values of the distance along the horizontal axis, not their logarithms.■

The method described above for doing log plots is convenient for a quick check on the usefulness of a logarithmic transformation, but it has one serious drawback. *Excel* cannot find the regression line relating the transformed variables using a logarithmic plot constructed as in Figure 9.5. You must numerically transform the data and then find the regression line using either of the methods described at the beginning of this section.

9.6 Summary

If a scatter plot of a set of bivariate data exhibits a linear trend we may want to use one of the variables, say *x*, to predict the value of the other, *y*. The ***regression line*** is the tool for doing this. This line passes through the point $(\overline{x}, \overline{y})$ and has slope rs_y / s_x.

The value $\hat{y}$ obtained from the regression equation for a specific *x* is a prediction. The actual value of *y* observed when the independent variable is close to *x* will vary, but will on average be close to this prediction (Regression Property 9.1). The value of r^2 (the ***coefficient of determination***) controls the variability of the actual value about the prediction $\hat{y}$ (Regression Property 9.2). When r^2 is close to one the actual *y* values are likely to show only modest spread about the regression line.

Transformations can often be applied to data so that the transformed data shows a linear trend. Using regression and working backwards, we can then find approximate non-linear relations in our original data. Such relations, while not as easy to interpret statistically, provide useful empirical tools for building mathematical models.

The ***regression effect*** arises when, having measured a population with respect to some variable, we categorize the data into groups, for example, taking the top 10%. When we observe some other characteristic, the individuals in our group will usually exhibit less extreme variation. For example, the top 10% according to the first measurement may only be in the top 20% with respect to the second measure. This leads to the Regression Fallacy when some treatment intervenes

between the two measurements and the observed change in variability is attributed to the treatment.

9.7 Mathematical Excursions

Regression Property 9.1 rests on an important mathematical identity. The decomposition provided by this identity plays an important role in many advanced statistical techniques used in multivariate data analysis. The reader may find an elementary introduction to these ideas useful in later studies.

Lemma 9.1: Suppose $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ satisfy

$$a_1 b_1 + a_2 b_2 + \cdots + a_n b_n = 0, \tag{9.2}$$

then

$$(a_1^2 + \cdots + a_n^2) + (b_1^2 + \cdots + b_n^2) = (a_1 + b_1)^2 + \cdots + (a_n + b_n)^2. \tag{9.3}$$

Proof:

Simply expand the right side of (9.3) and rearrange the terms. Using (9.2) the terms $a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$ drop out and we are left with precisely the left side of (9.3). ■

Lemma 9.1 is an algebraic generalization of the Pythagorean Theorem, which is the geometric interpretation of the case $n = 2$.

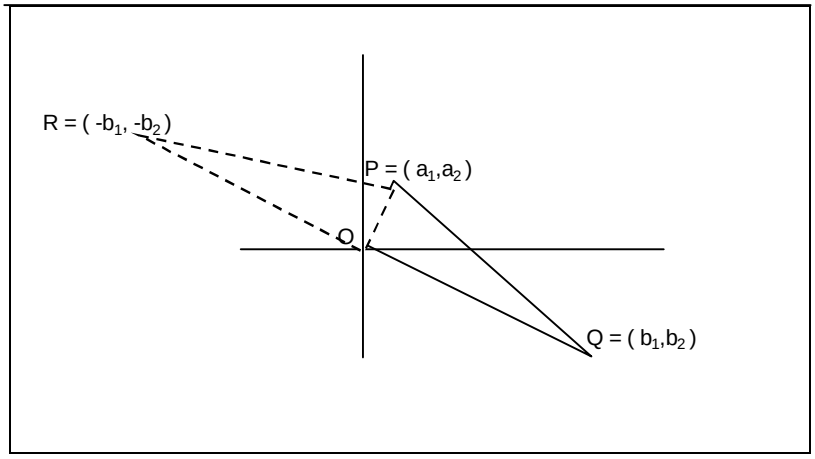


Figure 9.6

The reader should verify that in the diagram above the squared lengths of sides OP and OR are $a_1^2 + a_2^2$ and $b_1^2 + b_2^2$, respectively. The segments are perpendicular when the slopes are negative reciprocals. This translates into the condition $a_1b_1 + a_2b_2 = 0$. Since $|\mathrm{PR}|^2 = (a_1 + b_1)^2 + (a_2 + b_2)^2$, Lemma 9.1 is precisely the Pythagorean Theorem applied to the right triangle POR. In this sense the Lemma is the algebraic generalization of a familiar geometric property and leads us to introduce the following geometric language.

Definition 9.2: Two sequences $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ are said to be *perpendicular* or *orthogonal* if $a_1b_1 + a_2b_2 + \cdots + a_nb_n = 0$.■

What does all this have to do with Regression Property 9.2? The latter property tries to compare the variability s_y of the entire data set with the variability in the *y* values when *x* is fixed. For s_y we have the equation

$$s_y^2 = \frac{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n-1} = \frac{\mathrm{SS_T}}{n-1}. \tag{9.4}$$

The numerator in this expression is called the *total sum of squares*, abbreviated $\mathrm{SS_T}$. Lemma 9.1 allows us to decompose this numerator into two separate sums of squares, one measuring the variability around the regression line and the other measuring the variability due to the regression. To be precise we have the following important theorem.

Theorem 9.5:

$$\underbrace{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}_{\mathrm{SS_T}} = \underbrace{[(y_1 - \hat{y}_1)^2 + \cdots + (y_n - \hat{y}_n)^2]}_{\mathrm{SS_E}} + \underbrace{[(\hat{y}_1 - \bar{y})^2 + \cdots + (\hat{y}_n - \bar{y})^2]}_{\mathrm{SS_R}}.$$

Proof:

We apply Lemma 9.1 with $a_1 = y_1 - \hat{y}_1, \ldots, a_n = y_n - \hat{y}_n$, and $b_1 = \hat{y}_1 - \bar{y}, \ldots, b_n = \hat{y}_n - \bar{y}$. To do this we must verify that the orthogonality condition (9.2) is satisfied. This requires a bit more of algebraic labor than we expect most readers would care to experience, so we omit that detail. If the reader will grant (9.2) then clearly $a_1 + b_1 = y_1 - \bar{y}$, etc. so Lemma 9.1 yields the result stated in the theorem.■

We have labeled the sums of squares on the right side of Theorem 9.5 with their usual designations. The first, $\mathrm{SS_E}$, or *sum of squares due to "error"*, measures the variability of the data points around the regression line. In effect, this is the variation in the data that is not accounted for ("explained") by the regression. It is also referred to as the *residual sum of squares*. The second, $\mathrm{SS_R}$, or *sum of squares due to regression* measures the variability of the regression values $\hat{y}_i$ about $\bar{y}$, which happens to also be the mean of the regression values. Theorem 9.5 expresses the important result that

$$SS_T = SS_E + SS_R \tag{9.5}$$

or in words

total sum of squares = sum of squares due to errors + sum of squares due to regression.

Orthogonal decompositions of this sort play a very important role in multivariate statistics. They enable one to separate data variability into different components, thereby getting a handle on the contribution each makes to the total variability. Although all three quantities in (9.5) can be computed directly from the data, it is fairly easy to find simple expressions for two of them, namely SS_T and SS_R.

Theorem 9.6: $SS_T = (n-1)s_y^2$ and $SS_R = (n-1)r^2 s_y^2$

Proof:

The first statement is simply a rewriting of (9.4). To prove the second, note that from Theorem 9.2b) we have for each index *i*, $\hat{y}_i - \bar{y} = \hat{b}(x_i - \bar{x})$, where $\hat{b} = \frac{rs_y}{s_x}$ is the slope of the regression line. This yields

$$SS_R = (\hat{y}_1 - \bar{y})^2 + \cdots + (\hat{y}_n - \bar{y})^2 = \frac{r^2 s_y^2}{s_x^2}((x_1^2 - \bar{x})^2 + \cdots + (x_n^2 - \bar{x})^2) = (n-1)r^2 s_y^2,$$

since $(x_1^2 - \bar{x})^2 + \cdots + (x_n^2 - \bar{x})^2 = (n-1)s_x^2$. ■

From Theorem 9.7 and (9.5) we obtain the main result of this section.

Theorem 9.7:

a) $SS_E = (n-1)(1-r^2)s_y^2$

b) $\frac{SS_R}{SS_T} = r^2$

Proof:

a) From (9.5) and Theorem 9.6 we have $SS_E = SS_T - SS_R = (n-1)s_y^2(1-r^2)$.

b) This follows immediately from the previous theorem.■

Theorem 9.7b) asserts that r^2 gives the fraction of the total variation in the data that is due to the regression. Thus, for example, if $r = 0.5$ then $r^2 = 0.25$ so the regression accounts for about 25% of the overall variation in the data, as measured by the sum of squares criterion. On the other hand, 75% of the total variation is not accounted for by the regression. This viewpoint is an alternative, and perhaps more direct interpretation of the coefficient of determination, r^2, than our earlier discussion centering on Regression Property 9.2.

Regression Property 9.2 resembles the statement of Theorem 9.7a). Indeed, if we assert that $s_{y|x}^2$ (the variance of *y* for fixed *x*) satisfies

$$s_{y|x}^2 \approx \frac{\mathrm{SS_E}}{n-1},$$

then Regression Property 9.2 follows from Theorem 9.7a). We are allowed to make this approximation because a working assumption in regression analysis is that the *y* values associated with any fixed *x* show the same variation around the predicted value $\hat{y}$. This common variance can then be estimated using the average total residual variation given by $\frac{\mathrm{SS_E}}{n-1}$.

9.8 Exercises

1. a) Consider the three points (1,1), (4,0) and (0,8). If $y = a + bx$ is any (non-vertical) line, set up an expression for how well this line fits the three points in the sense of (vertical) least squares (see (9.1)).

   a) Using the expression you found in a) determine which of the following lines has the best fit to the three given points in the sense of least squares. Draw a sketch showing the three points and each line.

      i) The line through (4,0) and (0,8)

      ii) $y = 5.7 - 1.6x$

2. a) Find the quantities s_x, s_y and r for the three data points in exercise 1 and use these to find the regression line.

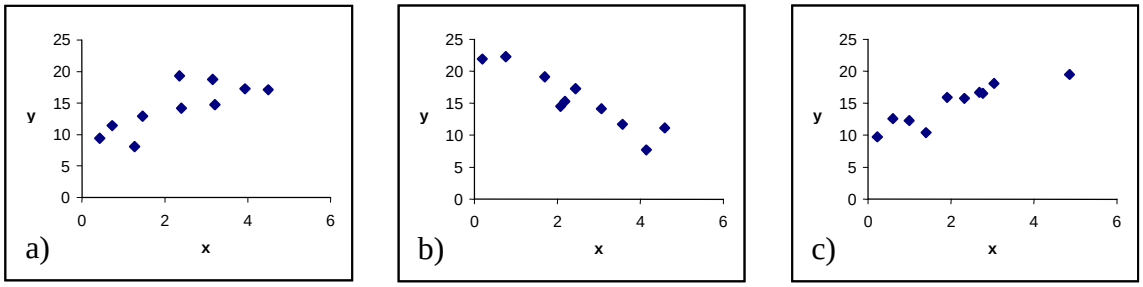
   b) Using the fitness expression you found in 1a) show that the regression line has a better fit than either of the lines considered in 1a).

3. The three points (1, 3), (2.5, 6) and (7, 15) lie on a straight line. Find the equation of this line and then verify that the formulas of Theorem 9.1 and Theorem 9.2 actually produce this line as the least squares regression line of *y* on *x*. Is this result always to be expected when the data points are collinear? Explain.

4. a) Suppose a set of bivariate data satisfies $\bar{x} = 3.5$, $s_x = 1.5$, $\bar{y} = 5.5$, $s_y = 0.75$ and $r = -.8$. Compute the regression line of y on x.

   b) What estimate would you give for the value of y when $x = 6$?

5. The graphs below show scatter plots of three data sets labeled a, b, and c:



   The regression lines of y on x for these are (not necessarily in the order shown)

   i) $\hat{y} = 10.1 + 2.2x, \quad r^2 = .83$

   ii) $\hat{y} = 9.2 + 2.2x, \quad r^2 = .60$

   iii) $\hat{y} = 23 - 3.1x, \quad r^2 = .87$

   Determine which regression line (i, ii, or iii) corresponds to which scatter plot. Indicate briefly how you arrived at your answer.

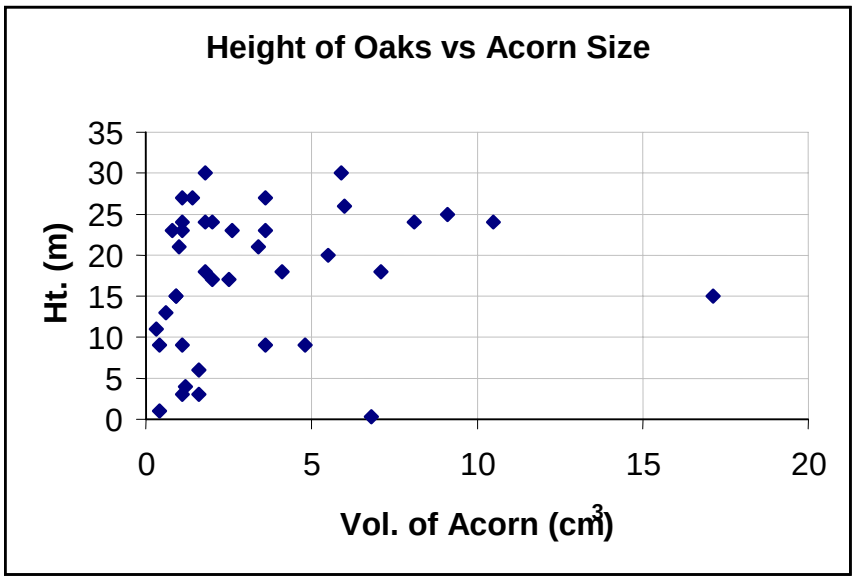
6. In section 8.4 (see Figure 8.4) we displayed a scatter plot of airfares vs. miles for selected flights from New York. The regression line of fares on distance is $\hat{y} = 162 + .092x$.

   a) What fare would you predict for a destination whose distance from New York is 1500 miles?

   b) For the tabulated flights the standard deviation of distances from New York is 1002 miles and the standard deviation of the respective fares is \$125. Find r.

   c) For a destination that is 1500 miles from New York, would you be surprised if the actual fare was \$400? What if the fare was \$500? Explain your reasoning.

7. The scatter plot below shows the heights of the oak species from file *acorn.xls* plotted against the size (in cubic centimeters) of the corresponding acorn for that species.



a) Based on the scatter plot make a guess as to the value of r. Briefly explain your thinking.

b) The regression equation of height (y) on acorn size (x) is given by $\hat{y} = 16 + .42x$. What is the predicted height of a mature oak from a species whose acorns are about 5 cm^3? Explain the meaning of this prediction.

c) The standard deviations s_x and s_y have values $s_x = 3.45$ cm^3 and $s_y = 8.50$ m. Using this information, determine the value of r^2 and then use this to explain why you either agree or disagree with the statement: "Acorn size explains very little of the variation in height of mature oaks."

8. Show that formula for the slope of the regression line $\hat{b} = rs_y / s_x$ can be rewritten as $\hat{b} = \frac{c_{xy}}{s_x^2}$.

9. Tom and Jerry are doing a physics lab together. They hang different weights (x) from a spring and measure the length (y) which the spring is stretched by each weight. Tom records the data with x in kilograms and y in centimeters, while Jerry records the same data with x in kilograms and y in meters.

   a) Whose data will show the larger standard deviation for the y values? Explain

   b) Will there be a difference in the correlation coefficients? Explain

   c) Whose data will show a steeper regression line? Explain

10. A study was done of the increase in the weight (y) of 20 laboratory mice vs. their caloric consumption (x) over a certain time period.

a) If a lab assistant accidentally recorded an additional 10 grams to the weight gain of each mouse, what effect will this have on the regression equation of y vs. x, assuming y is measured in grams?

b) If the lab assistant accidentally recorded the weight gain of each mouse as 10% above its actual value, what effect will this have on the regression equation of y vs. x?

11. a) Consider the same three points as in 1a). If $x = a + by$ is any (non-horizontal) line, set up an expression for how well this line fits the three points in the sense of (horizontal) least squares, i.e. using the horizontal distances from the line to each point.

b) Using the expression you found in 11a) determine which of the following lines has the best fit to the three given points, in the sense of horizontal least squares. Draw a sketch showing the three points and each line.

i) The line through (4,0) and (0,8)

ii) $y = 5.7 - 1.6x$

12. a) In constructing a regression equation of x on y we want to predict the x value from knowledge of the y value. In this case we need an equation of the form $x = \hat{a} + \hat{b}y$ that bests fits the given points in the sense of horizontal least squares. Explain why rotating your head 90° turns this problem into the usual (vertical) least squares problem. Then use Theorem 9.1 and Theorem 9.2 to find expressions for the coefficients $\hat{a}$ and $\hat{b}$ in the equation of the regression line $x = \hat{a} + \hat{b}y$.

b) Show that the equation of the regression line of x on y that you found in 12a) may also be written as $y - \overline{y} = \dfrac{s_y}{rs_x}(x - \overline{x})$.

13. For the three point data in Example 9.1, use exercise 12 (and Example 9.2) to find the regression line of x on y. Draw a graph showing the three points and the two regression lines (y on x and x on y).

14. Suppose x and y are bivariate data for which $r > 0$. Which regression line has a steeper slope (measured as $\Delta y / \Delta x$), the regression line of y on x or the regression line of x on y? Explain.

15. a) A study of weights (y) vs. heights (x) for men yielded for the regression line of weight on height the equation $\hat{y} = -120 + 4.2x$, where x is measured in inches and y in pounds. Is a 64-inch man who weighs 165 pounds above or below average in weight <u>for his height</u>? Explain.

b) Suppose a man is considered obese if his weight is more than 2 standard deviations above the mean weight for men with the same height. If the weight-height data yielded $s_x = 2.3$ and $s_y = 15$ would the man in 15a) be considered obese? Explain.

c) Can you find the equation of the regression line of height on weight from the information in parts a) and b)? If not, can you find the slope of this line? Explain.

d) Why is it invalid to use the regression equation in 15a) to predict the weights of children?

16. The file *sat.xls* presents for each state in the U.S. the average SAT scores y achieved by high-school seniors in that state and the percentage x of seniors who took the test. Some summary statistics for this data are presented in the following table:

SAT scores, y	$\bar{y} = 948$	$s_y = 66$
% seniors taking test, x	$\bar{x} = 35.7\ \%$	$s_x = 27\ \%$

The equation of the regression line is $y = -2.2x + 1024$.

a) As the percentage of students taking the test increases, what happens to the average SAT scores?

b) What average SAT score would you predict if 50% of the H.S. seniors took the exam? (Note the units of x are in percents, not their decimal equivalents.)

c) What is the value of the correlation coefficient?

d) Suppose a state had a 50% participation rate and the average SAT scores for seniors in that state were 1000. Give a statistical explanation as to why this should be considered an extremely good result.

17. The file *OldFaith.xls* contains records of the duration of an eruption of Old Faithful and the time to next eruption.

a) Prepare a scatter plot of time to next eruption vs. duration of previous eruption. What aspect of the distribution of each variable is reflected in the appearance of this scatter plot? (See the discussion in Chapter 7 and exercise 9 of that chapter.)

b) Find the regression line for the scatter plot. Does the data seem adequately described by the regression line?

c) How should Park Rangers use the regression line to post for park visitors the waiting time to the next eruption?

18. The file *bivar.xls* contains 400 computer generated bivariate data points labeled x and y.

a) Construct a scatter plot of the data, including the regression line and r^2 value.

b) Fill in the second row of the following table using the regression equation. Carry out the computations in *Excel* using suitable formulas.

x	1	2	3	4	5	6	7	8	9
$\hat{y}$									
empirical estimate for $\hat{y}$									

c) Fill in the third row by computing average y values using only data points whose x value is near the number in row 1. For example, for $x=1$ you might estimate the regression value $\hat{y}$ from the data using the average of all y values for which the corresponding x is between 0.5 and 1.5. (Note: You may want to copy the data set onto a blank worksheet before manipulating it.)

d) Should the values in row 3 be close to the values in row 2? Explain. Would you consider your results in line with these expectations?

19. In discussing the Regression Effect we need to be more precise as to how deviation from the mean is measured. The appropriate units are so-called *z*-scores, which we will consider in more detail in Chapter 14. The z-score of a measurement x is the number of standard deviations of x from its mean, i.e. the z-score is $z_x = \dfrac{x-\overline{x}}{s}$. This concept was implicit in the Bell Curve Rule, discussed in Chapter 7.

a) For the grade data in Figure 9.1 we have $\overline{x}=65.9$, $s_x=17.9$, $\overline{y}=61.8$, $s_y=20.3$. Using Table 9.2 and Table 9.3 find the z-score for the highest and lowest exam grades.

b) If $\hat{y}$ is the predicted value of y on the regression line corresponding to x, show that

$$\frac{\hat{y}-\overline{y}}{s_y} = r\left(\frac{x-\overline{x}}{s_x}\right) = rz_x \tag{9.6}$$

where z_x is the z-score of x.

c) Conclude from b) that whenever x and y are not perfectly correlated, the predicted value $\hat{y}$ will have a smaller absolute z-score than the z-score of x. Interpreting $\hat{y}$ as an average, how does this explain the Regression Effect?

20. a) Find the top ten x values for the data in *bivar.xls* and make a chart with the z-scores of each value. What is the average z-score for these ten values?

b) For each of the x values examined in a) find the z-scores for the associated y values. What is the average z-score for these ten y values?

c) For each of the x values examined in a), use (9.6) to determine the z-score for the predicted y value. What is the average z-score of these 10 predicted values?

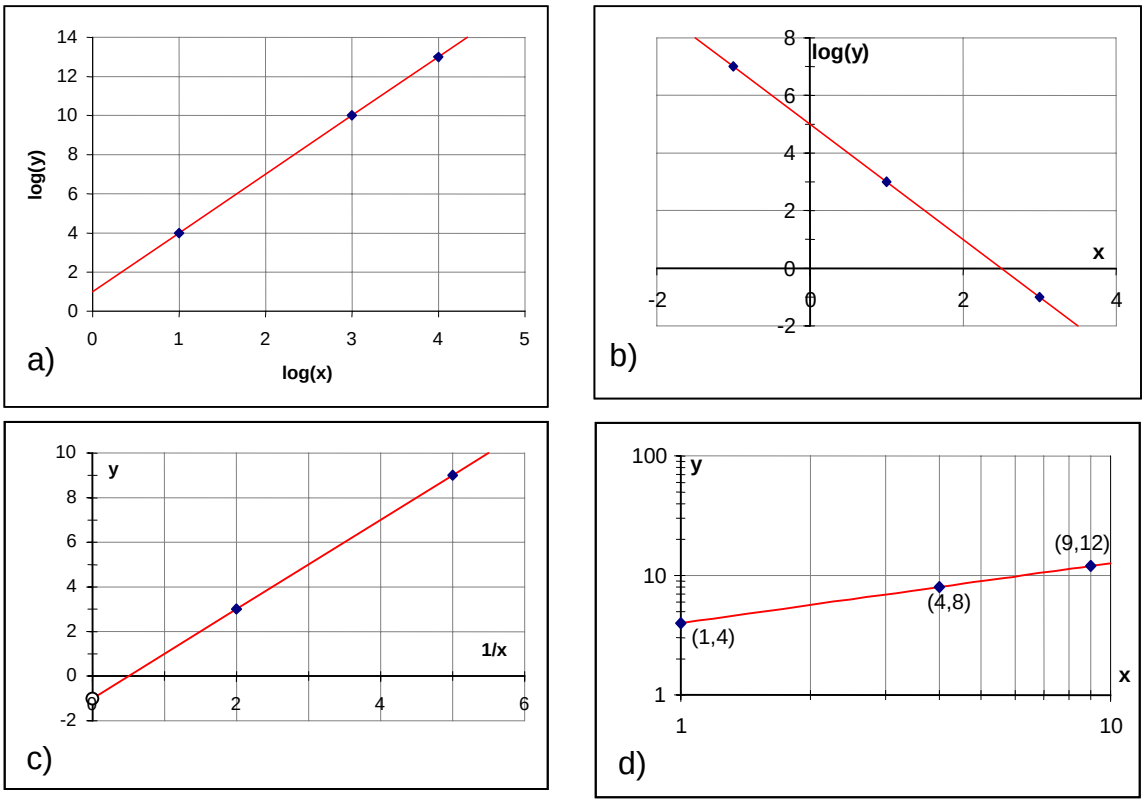
d) Explain why the computations in this exercise illustrate the Regression Effect.

21. a) If $y = 3x^2$ draw a sketch of the log-log plot of $\log(y)$ versus $\log(x)$ for $x \geq 1$.

    b) If $y = 3e^{2x}$ draw a sketch of the semi-log plot of $\log(y)$ versus x.

    c) If $y = c10^{bx}$ show that y can also be written as $y = ce^{kx}$ for some k. Find an expression for k.

22. Using each of the plots a) - d) below find an equation relating y and x.



23. Complete the proof of Theorem 9.4.

24. a) Use the data in *US_pop.xls* for the years 1790 - 1860 to find an exponential fit of the form $y = ce^{rt}$ for the population during this period. (Let $t = 0$ correspond to 1790.)

    b) What prediction does your model give for the population in 1900? Compare to the actual value.

25. Using the data in the file *mammals.xls* find a simple allometric relationship between body weight and brain size for the species listed.

26. Propranolol is a beta blocker that is sometimes prescribed to reduce the frequency of migraine headaches. A patient suffering from migraines agreed to a study in which a certain dosage of propranolol (known to the physician but not the patient) was administered for a six week

period and the number of migraines recorded. The table below records the results for 8 consecutive such periods. (from *Biostatistics in Clinical Medicine*, J.A. Ingelfinger et.al, 1994)

Period	1	2	3	4	5	6	7	8
Dosage (mg)	Placebo	10	40	80	120	160	Placebo	40
Migraine Rate (# per 30 days)	9.5	7.5	5.0	4.5	4.5	3.5	13.1	4.7

a) Construct a plot of migraine rate vs. propranolol dosage and find the regression line. Do you feel a linear relation accurately describes the data?

b) Pharmacological principles suggest that the response should be related to the logarithm of the dosage. Plot the migraine rate vs. the log of the dosage. To avoid the fact that log(0) is not defined use a small dosage, say 1-mg, for the placebo.

c) Fit a regression line to the plot obtained in b). How accurately does the linear relation capture the data?

d) What migraine rate would you predict for a dosage of 100 mg?

27. a) Open the file *auto.xls*. (See exercise 11 in Chapter 8). After preparing a scatter plot find the regression line of mileage on weight together with the coefficient of determination (r^2). What estimate would you give for the mileage for a 4000 lb. car.

b) For cars weighing close to 4000 lbs, approximately how wide is the interval one standard deviation about the mean mileage.

In each of problems 28 - 31 prepare a scatter plot of the data, including a regression line. The plot should have a title and the axes should be appropriately labeled. You will have to decide which variable to select as the independent variable and which as the dependent.

You should give a brief discussion, with reasons, as to whether the data shows a linear relationship and if so how strong the relationship is. Take note of any exceptional data points or outliers, one or both of whose coordinates are far removed from the center of their respective distributions. By deleting these points (remember, you can undelete!) you can investigate the influence these points have on the regression line. Comment on the results.

28. Use the file *cities.xls* (columns A, B and C only). (See exercise 12 of Chapter 8)

29. Use the data file *hibernat.xls*.

30. Work with the data file *sat.xls*.

31. Work with the data file *ozone.xls*.