# Elementary Statistics

Brian E. Blank

March 4, 2018

FIRST PRESIDENT UNIVERSITY PRESS

# 5. Regression

When bivariate data in a scatter plot appears to have a linear association, then we often seek a linear model about which the swarm of points clusters. Determining such a line is the focus of this chapter.

## 5.1 The Regression Line

In this section, we will begin by plotting some tabulated data. The scatter plot of the data will suggest a linear association, which Pearson's $r$ will corroborate. A very specific linear model for the data in the form of an explicit equation for a line will be proposed. The mathematical strategy for obtaining the equation of the linear model will be explicated. Finally, the appropriateness of the linear model will be evaluated.

### Raw Bits

> Do you qualify for Raw Bits, the natural cereal made from oat hulls and wheat shafts? It's not just a cereal; it's a moral statement!
> —Commercial message from Raw Bits, longtime sponsor of *A Prairie Home Companion*

It is reasonable to believe that the calories contained in a serving of breakfast cereal will have at least a moderate linear association with the quantity of sugar in the serving. To investigate this belief, data from a somewhat random assortment of 26 cereals was collected. We have used X for the explanatory variable (the sugar content) and Y for the response variable (the caloric content). The scatter plot of the 26 points is shown in Figure 5.1.1. It confirms our hypothesis of a moderate positive linear association.

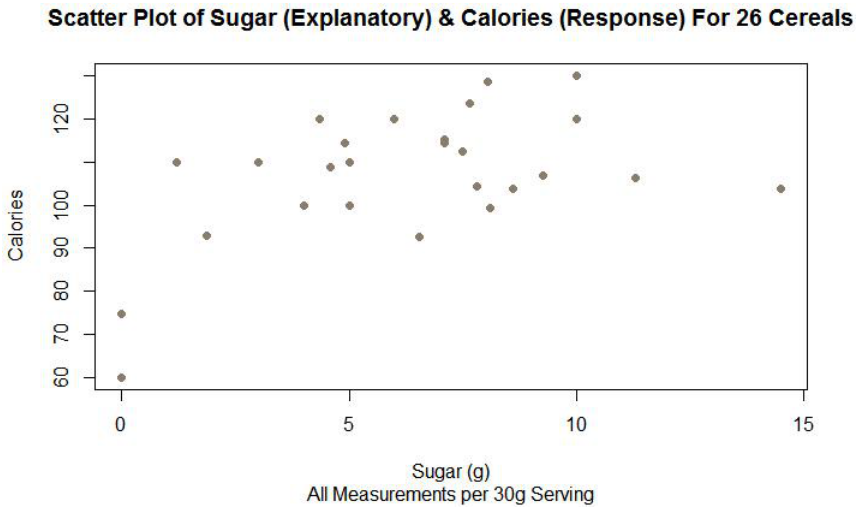



**Figure 5.1.1 Scatter Plot of (Sugar,Calories)**

With a value of about 0.504, Pearson's $r$ quantitatively confirms our suspicion that there is a moderate positive linear association. In addition to $r$ we will need some other summary statistics. They can be found in Table 5.1.1.

| $\bar{x}$ | $s_X$ | $\bar{y}$ | $s_Y$ | $r$ |
|---|---|---|---|---|
| 6.286154 | 3.464582 | 107.0231 | 15.30467 | 0.5039541 |

**Table 5.1.1: Key Statistics of Sugar Content (X) and Caloric Content (Y) for 26 Fine Cereal Products**

For details on the cereals selected for the investigation and their sugar and caloric contents, see Table 5.1.2.

| **Cereal (Producer and Product Name)** | **Sugar (grams)** | **Calories** |
|---|---|---|
| General Mills Cheerios | 1.20 | 110.0 |
| General Mills Fiber One | 0.00 | 60.0 |
| General Mills Kix | 3.00 | 110.0 |
| General Mills Oatmeal Crisp | 8.60 | 103.9 |
| General Mills Total | 5.00 | 100.0 |
| General Mills Trix | 10.0 | 120.0 |
| General Mills Wheaties | 4.00 | 100.0 |
| Kellogg's Corn Flakes | 1.87 | 93.0 |
| Kellogg's Meuslix | 9.27 | 106.9 |
| Kellogg's Raisin Bran Crunch | 11.3 | 106.4 |
| Kellogg's Special K Red Berries | 8.10 | 99.3 |
| Kashi Granola | 4.60 | 108.9 |
| Kashi Heart to Heart Oatmeal | 7.80 | 104.2 |
| Kashi Raisin Vineyard | 6.55 | 92.7 |
| Malt-O-Meal Cinnamon Toasters | 10.0 | 130.0 |
| Malt-O-Meal Oat Blenders | 6.00 | 120.0 |
| Malt-O-Meal Golden Puffs | 14.5 | 103.7 |
| Post Great Grains Digestive Blends Berry | 4.91 | 114.5 |
| Post Great Grains Cranberry, Almond | 7.50 | 112.5 |
| Post Great Grains Honey, Oats, Seeds | 4.36 | 120.0 |
| Post Great Grains Raisins, Dates, Pecans | 7.09 | 114.5 |
| Quaker Oats Oats and Honey | 7.10 | 115.3 |
| Quaker Oats Granola | 7.65 | 123.5 |
| Quaker Oats Puffed Rice | 0.00 | 74.7 |
| Quaker Oats Real Medleys | 8.04 | 128.6 |
| Quaker Oatmeal Squares Brown Sugar | 5.00 | 110.0 |

**Table 5.1.2: Sugar Content (grams) and Calories in 30 gram Servings of 26 Cereals**

The question we now ask is, *If the data is more or less scattered along a line, then what is the equation of that line?* Good question! With $y$ representing the caloric content (the response variable) and $x$ representing the quantity of sugar in grams (the explanatory or predictor variable), the equation of the line we seek has the form $y = m\,x + b$. Our question becomes, *What values of $m$ and $b$ should we use?* In this section we will show that the general answer to this question is

$$m = r \cdot \frac{s_Y}{s_X} \qquad \text{and} \qquad b = \bar{y} - m\,\bar{x}. \tag{5.1.0}$$

With the values of $\bar{x}, s_X, \bar{y}, s_Y,$ and $r$ found in Table 5.1.1, the equations of line (5.1.0) give us $m = 2.2262$ and $b = 93.02884$ for the 26 cereals we have been considering. We will say that the resulting line with equation $y = 2.2262\,x + 93.02884$ is a ***linear model*** for the data. In Figure 5.1.2 we have superimposed this linear model on the scatter plot of cereal data.

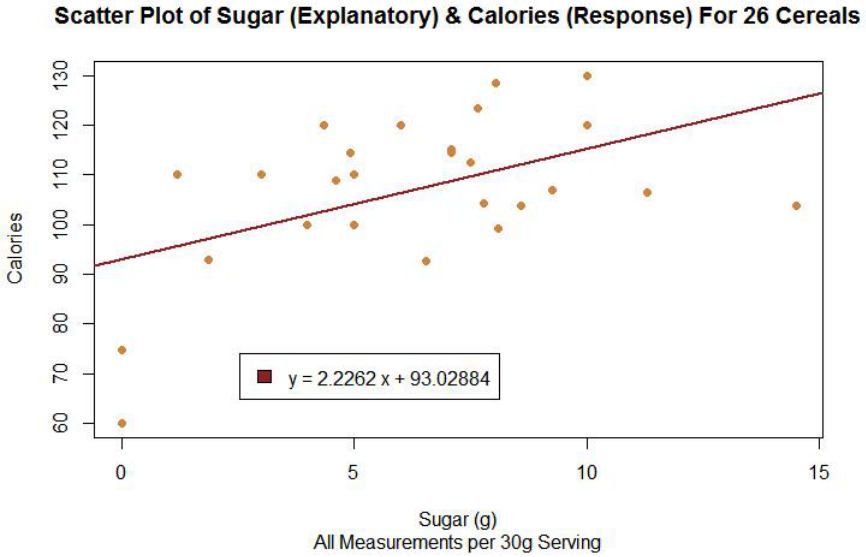



**Figure 5.1.2 Scatter Plot of (Sugar,Calories) with Linear Model**

We can see that the linear model $y = 2.2262\,x + 93.02884$ passes nicely among the points in the scatter plot. But why do we use this line and not another, similar one? Were we to change the slope just a bit from 2.2262 and were we to change the $y$-intercept just a bit from 93.02884, then surely we would still have a line passing among the scattered points. What is so special about the numbers 2.2262 and 93.02884 that we have chosen for the slope and $y$-intercept of the linear model? In this section We will answer the question that we have just posed.

## The $y$-Intercept of the Best Fitting Line

We assume that X and Y are distributions with $N$ observed values $x_1, x_2, \ldots, x_N$ and $y_1, y_2, \ldots, y_N$ respectively, that a scatterplot reveals that the $N$ points $(x_1, y_1)\,,\ (x_2, y_2)\,,\ \ldots\ ,(x_N, y_N)$ display a linear relationship, and that the linear correlation coefficient $r$ confirms our assessment of the scatterplot. As usual, we will measure $x$ along the horizontal axis and $y$ along the vertical axis. It will useful to use the vertical axis to also represent a variable we will denote by $\widehat{y}$. We will write the model for the linear association between X and Y as $\widehat{y} = m\,x + b$. With this notation we do not have to worry that our model does not pass through an observed data point $(x_j, y_j)$. We simply set $x = x_j$ in the equation for the linear model: doing do results in a value $\widehat{y}_j = m\,x_j + b$. The notation $\widehat{y}_j$ does not suggest that the value obtained from the model is precisely the observed value $y_j$. We will refer to $\widehat{y}_j$ as a ***predicted value*** or a ***fitted value***.

We now turn our attention to selecting advantageous values for the slope $m$ and $y$-intercept $b$ of the linear model $\widehat{y} = m\,x + b$. Once we have determined such a line, we can use it to estimate or predict the value $mx + b$ of the response variable $y$ that corresponds to a given value $x$ of the predictor variable. Of course, we can also let $x$ assume the value of an observation $x_j$ and compare the predicted response $m\,x_j + b$ with the observed responce $y_j$. For each $j$ from 1 to $N$, we will let

$$\widehat{y}_j = m\,x_j + b$$

denote the value that the linear model predicts for the data value $x_j$. As Figure 5.1.2 indicates, there might not be a single $j$ for which the predicted value $\widehat{y}_j$ equals the observed value $y_j$. Indeed, if the observation measurements are sufficiently precise, then we would not expect any fitted value $widehaty_j$ to exactly equal the observed value $y_j$. The amount $y_j - \widehat{y}_j$ by which the predicted value misses the observed value is called the $j^{\text{th}}$ ***residual*** and is denoted by $e_j$:

$$e_j = y_j - \widehat{y}_j.$$

By rewriting the equation for the residual as

$$y_j = \widehat{y}_j + e_j,$$

we see the equation,

**Observed datum = linear model prediction + residual.**

In terms of the parameters $m$ and $b$, the equations for the residuals are

$$e_j = y_j - m\,x_j - b \qquad (1 \le j \le N). \tag{5.1.1}$$

If a residual $e_j$ is nonzero, then the predicted value $\widehat{y}_j$ does not equal the observed value $y_j$ and is in error. The letter $e$, in fact, stands for *error*. We would like to find the line that minimizes the totality of these errors, but the same considerations that arise in defining the concept of *variance* apply here: we do not want positive errors and negative errors to cancel, causing the calculated total error to be smaller than it really is. Using the absolute values of the errors fixes that problem, but absolute values are difficult to work with. Therefore, we square $e_j$ to obtain a nonnegative measure of the $j^{\text{th}}$ error. Table 5.1.3 below is an extension of Table 5.1.1 that includes the caloric values predicted by the linear model $y = 2.2262\,x + 93.2884$. Table 5.1.3 also tabulates the errors and square errors associated with these predicted response values.

| Observation $j$ | $x_j$ | $y_j$ | $\widehat{y}_j$ | $e_j$ | $e_j^2$ |
|---|---|---|---|---|---|
| 1 | 1.200 | 110.000 | 95.700 | 14.300 | 204.482 |
| 2 | 0.000 | 60.000 | 93.029 | -33.029 | 1090.904 |
| 3 | 3.000 | 110.000 | 99.707 | 10.293 | 105.937 |
| 4 | 8.600 | 103.900 | 112.174 | -8.274 | 68.462 |
| 5 | 5.000 | 100.000 | 104.160 | -4.160 | 17.304 |
| 6 | 10.000 | 120.000 | 115.291 | 4.709 | 22.176 |
| 7 | 4.000 | 100.000 | 101.934 | -1.934 | 3.739 |
| 8 | 1.870 | 93.000 | 97.192 | -4.192 | 17.571 |
| 9 | 9.270 | 106.900 | 113.666 | -6.766 | 45.775 |
| 10 | 11.300 | 106.400 | 118.185 | -11.785 | 138.884 |
| 11 | 8.100 | 99.300 | 111.061 | -11.761 | 138.323 |
| 12 | 4.600 | 108.900 | 103.269 | 5.631 | 31.704 |
| 13 | 7.800 | 104.200 | 110.393 | -6.193 | 38.356 |
| 14 | 6.550 | 92.700 | 107.610 | -14.910 | 222.322 |
| 15 | 10.000 | 130.000 | 115.291 | 14.709 | 216.359 |
| 16 | 6.000 | 120.000 | 106.386 | 13.614 | 185.340 |
| 17 | 14.500 | 103.700 | 125.309 | -21.609 | 466.938 |
| 18 | 4.910 | 114.500 | 103.959 | 10.541 | 111.103 |
| 19 | 7.500 | 112.500 | 109.725 | 2.775 | 7.699 |
| 20 | 4.360 | 120.000 | 102.735 | 17.265 | 298.078 |
| 21 | 7.090 | 114.500 | 108.813 | 5.687 | 32.347 |
| 22 | 7.100 | 115.300 | 108.835 | 6.465 | 41.798 |
| 23 | 7.650 | 123.500 | 110.059 | 13.441 | 180.653 |
| 24 | 0.000 | 74.700 | 93.029 | -18.329 | 335.946 |
| 25 | 8.040 | 128.600 | 110.927 | 17.673 | 312.318 |
| 26 | 5.000 | 110.000 | 104.160 | 5.840 | 34.107 |
| | | | | | 4368.624 |

**Table 5.1.3: Predicted Caloric Content $\widehat{y}_j$, Error $e_j$, Square Error $e_j^2$**

You will note that no error is particularly small, let alone 0. The fact that no error is 0 tells us that our linear model does not pass through a single point in the scatter plot. Also notice that the marginal row of Table 5.1.3 contains one entry, 4368.624, which is the sum of the square errors tabulated above it in the final column. This marginal entry is the ***sum of the squares of the errors***, which is defined, in general, as

$$\text{SSE}(m,b) = \sum_{j=1}^{N} e_j^2 = \sum_{j=1}^{N} (y_j - \widehat{y}_j)^2 = \sum_{j=1}^{N} (y_j - m\,x_j - b)^2\,. \tag{5.1.2}$$

It is evident that the value of the sum in equation (5.1.2) depends on the values of $m$ and $b$. That is the reason for using the functional notation $\text{SSE}(m,b)$. Our task is to find the values for $m$ and $b$ that result in $\text{SSE}(m,b)$ being a *minimum*. For this reason, the line $y = m\,x + b$ that results from these minimizing values is called the ***least squares line***. It is also called the ***best fitting line*** and, for reasons that we shall soon understand, the ***regression line***. After we have derived the formulas for $m$ and $b$ that result in the regression line, the formulas

that we gave in line (5.1.0), we will not consider any other values for $m$ and $b$, and we will write SSE rather than $\text{SSE}(m, b)$.

As we have mentioned, Table 5.1.3 includes the value 4368.624 of $\text{SSE}(m, b)$ for $m = 2.2262$ and $b = 93.02884$. Every other candidate linear model, no matter how poorly chosen, gives rise to a value of $\text{SSE}(m, b)$. For example, if we simplify the parameters a bit so that the candidate linear model is $y = 2.2\,x + 83$, then the value of $\text{SSE}(m, b)$ is 4369.803. If we try $m = 2.3$ and $b = 93.5$, then $\text{SSE}(m, b)$ is 4392.992. No matter what other values of $m$ and $b$ we try, we will get a value of $\text{SSE}(m, b)$ that is larger than 4368.624. The line $y = 7\,x + 60$, which passes through the second and fifteenth observations $(0, 60)$ and $(10, 130)$ produces a whopping $\text{SSE}(m, b)$ of 11444.400.

For organizational purposes, we have divided the task of finding the regression line into several parts. Remember: Our task is to derive the equations given as a preview in line (5.1.0). We are not assuming the equations of line (5.1.0) and we will not use them until they have been derived. Once we have shown that the equations for $m$ and $b$ stated in line (5.1.0) minimize $\text{SSE}(m, b)$, we will use only these values for $m$ and $b$. From that time onward, $\text{SSE}(m, b)$ will refer to the one value that results from the regression line parameter: it will cease to be a function and it will be written simply as SSE.

A good deal of algebra is about to begin. The amount you read should depend on your interests.
At the end of the algebra, there will be a Summary subsection enumerating the facts you should know.

**Theorem 1.** *Suppose that $(x_1, y_1), \ldots, (x_N, y_N)$ are points in the xy-plane. For any real numbers $m$ and $b$, let SSE$(m, b)$ be the sum of the squares of the errors, as defined by equation (5.1.2). Then, if $m$ and $b$ are the values that minimize SSE$(m, b)$, we have*

$$b = \bar{y} - m\,\bar{x}. \tag{5.1.3}$$

*Proof.* Notice that we are not yet saying what the slope $m$ of the least squares line must be. What we are asserting is that the $y$-intercept of the least squares line can be expressed in terms of that still-unknown slope by equation (5.1.3). Therefore, once we determine $m$, we will have the complete equation of the least squares line.

We first prepare the quantities that are to be squared in $\text{SSE}(m, b)$. Then we expand the squares. Three of the sums that result from the expansion can be simplified. Completing the square is the final step. This is analogous to the algebraic procedure for finding the low point or vertex of a parabola:

$$\begin{aligned}
\text{SSE}(m, b) &= \sum_{j=1}^{N} \left(b - (y_j - m\,x_j)\right)^2 \\
&= \sum_{j=1}^{N} b^2 - 2\,b \sum_{j=1}^{N} (y_j - m\,x_j) + \sum_{j=1}^{N} (y_j - m\,x_j)^2 \\
&= N\,b^2 - 2\,b\,(N\,\bar{y} - m\,N\,\bar{x}) + \sum_{j=1}^{N} (y_j - m\,x_j)^2 \\
&= N\left(\left(b^2 - 2\,b\,(\bar{y} - m\,\bar{x})\right) + \frac{1}{N}\sum_{j=1}^{N} (y_j - m\,x_j)^2\right) \\
&= N\left(\left(b^2 - 2\,b\,(\bar{y} - m\,\bar{x}) + (\bar{y} - m\,\bar{x})^2\right) + \frac{1}{N}\sum_{j=1}^{N} (y_j - m\,x_j)^2 - (\bar{y} - m\,\bar{x})^2\right) \\
&= N\left(\left(b - (\bar{y} - m\,\bar{x})\right)^2 + \frac{1}{N}\sum_{j=1}^{N} (y_j - m\,x_j)^2 - (\bar{y} - m\,\bar{x})^2\right).
\end{aligned}$$

Notice that, of the three summands inside the large parentheses of the last line, only the first involves $b$. Therefore, whatever specification of $b$ is made, the values of the final two sums are not affected. As a result, we can minimize the total expression inside the large parentheses by finding the value of $m$ that minimizes the combination of the final two summands, and then specifying $b$ to minimize the first summand. But that first summand is a square, hence nonnegative. We minimize its value by choosing $b$ so that it is 0. In other words, if $\text{SSE}(m, b)$ is minimized, then $b = \bar{y} - m\,\bar{x}$. □

There are several important implications of equation (5.1.3). The most obvious one is that we have reduced the problem of finding two unknowns, $m$ and $b$, to the simpler problem of finding one unknown, $m$. Once we succeed

at obtaining an expression for $m$ we are done because equation (5.1.3) expresses $b$ in terms of $m$ and the sample means $\bar{x}$ and $\bar{y}$ of the two observed distributions. The next theorem states several less obvious implications of equation (5.1.3).

**Theorem 2.** *Let $\widehat{y} = m\,x + b$ be the regression line for $(x_1, y_1)\,,\ (x_2, y_2)\,,\ \ldots\,,(x_N, y_N)$. Let $\bar{x}$ and $\bar{y}$ be the means of the $X$ and $Y$ observations. Let $\widehat{y}_1 = m\,x_1 + b, \widehat{y}_2 = m\,x_2 + b, \ldots,\ \widehat{y}_N = m\,x_N + b$ be the predicted responses to the observed explanatory values $x_1, x_2, \ldots,\ x_N$. Let $\overline{\widehat{y}}$ be the mean of these predicted values. Let $e_1, e_2, \ldots,\ e_N$ be the residuals defined by equation (5.1.1). Then*

*i) The point $(\bar{x}, \bar{y})$ lies on the regression line.*
*ii) The mean of the predicted responses equals the mean of the observed responses: $\overline{\widehat{y}} = \bar{y}$.*
*iii) The sum of the residuals is zero:*

$$\sum_{j=1}^{N} e_j = 0. \tag{5.1.4}$$

*Proof.* For part (i), we note that for $x = \bar{x}$, the value of $\widehat{y} = m\,x + b$ is $m\,\bar{x} + b$, or $m\,\bar{x} + (\bar{y} - m\,\bar{x})$, or $\bar{y}$. Thus, $(\bar{x}, \bar{y})$ lies on the regression line.

For part (ii), we calculate

$$\begin{aligned}
\overline{\widehat{y}} &= \frac{1}{N}\sum_{j=1}^{N} \widehat{y}_j \\
&= \frac{1}{N}\sum_{j=1}^{N} (m\,x_j + b) \\
&= m\,\frac{1}{N}\sum_{j=1}^{N} x_j + \frac{1}{N}\sum_{j=1}^{N} b \\
&= m\,\bar{x} + \frac{1}{N}\left(N \times b\right) \\
&= m\,\bar{x} + b \\
&= m\,\bar{x} + (\bar{y} - m\,\bar{x}) \\
&= \bar{y}.
\end{aligned}$$

For part (iii), we calculate

$$\sum_{j=1}^{N} e_j = \sum_{j=1}^{N} (y_j - \widehat{y}_j) = N \times \bar{y} - N \times \overline{\widehat{y}} = N\,\left(\bar{y} - \overline{\widehat{y}}\right) = 0 \qquad \text{(by part (ii) of this theorem).}$$

□

Next we will derive the formulas for $m$ and $b$ that we announced in line (5.1.0). The derivation is simplified now that we have reduced the problem to a minimization problem with only one variable. The standard method of minimization of Calculus I suffices.

**Theorem 3.** *Let SSE$(m, b)$ be defined by formula (5.1.2). Then among all possible values of $m$ and $b$, the sum of squares SSE$(m, b)$ is minimized when $m$ and $b$ are given by*

$$m = r_{XY} \cdot \frac{s_Y}{s_X} \qquad \textit{and} \qquad b = \bar{y} - m\,\bar{x}. \tag{5.1.0}$$

*The regression line is therefore*

$$\widehat{y} = r_{XY} \cdot \frac{s_Y}{s_X}\,x + \bar{y} - r_{XY} \cdot \frac{s_Y}{s_X}\,\bar{x}$$

*or, equivalently,*

$$\widehat{y} = r_{XY} \cdot \frac{s_Y}{s_X}\,(x - \bar{x}) + \bar{y}.$$

*Proof.* We begin by rewriting the sum of squares in a more convenient form. Toward that end, we replace $b$ by the right side of formula (5.1.3). You will note that the resulting formula expresses $\mathrm{SSE}(m, b)$ in terms of the $x$ and $y$ deviations:

$$\mathrm{SSE}(m,b) = \sum_{j=1}^{N} (y_j - (m\,x_j + b))^2 = \sum_{j=1}^{N} (y_j - (m\,x_j + \bar{y} - m\,\bar{x}))^2 = \sum_{j=1}^{N} ((y_j - \bar{y}) - m\ (x_j - \bar{x}))^2\,.$$

Using a standard method of minimizing a function of one variable $m$, we differentiate with respect to $m$ and set the derivative equal to 0. We remark that an even more elementary method would be to complete the square.

$$\begin{aligned}
0 &= \frac{d}{dm} \sum_{j=1}^{N} \Big( (y_j - \bar{y}) - m\ (x_j - \bar{x}) \Big)^2 \\
&= -2 \sum_{j=1}^{N} \Big( (y_j - \bar{y}) - m\ (x_j - \bar{x}) \Big)\ (x_j - \bar{x}) \\
&= -2 \sum_{j=1}^{N} \Big( (y_j - \bar{y})\ (x_j - \bar{x}) \Big) + 2\,m \sum_{j=1}^{N} (x_j - \bar{x})^2 \\
&= -2 \sum_{j=1}^{N} \Big( (y_j - \bar{y})\ (x_j - \bar{x}) \Big) + 2\,m\,(N-1) \cdot \frac{1}{N-1} \sum_{j=1}^{N} (x_j - \bar{x})^2 \\
&= -2 \sum_{j=1}^{N} \Big( (y_j - \bar{y})\ (x_j - \bar{x}) \Big) + 2\,m\,(N-1) \cdot s_X^2.
\end{aligned}$$

On solving for $m$ we find

$$m = \frac{1}{(N-1)s_X^2} \left( \sum_{j=1}^{N} (y_j - \bar{y})\ (x_j - \bar{x}) \right) = \frac{1}{N-1} \left( \sum_{j=1}^{N} \left( \frac{(x_j - \bar{x})}{s_X} \right) \left( \frac{(y_j - \bar{y})}{s_Y} \right) \right) \frac{s_Y}{s_X} = r_{XY} \frac{s_Y}{s_X}.$$

□

Let E denote the distribution of residuals $e_1, e_2, \ldots, e_N$ that result from the regression line. It follows from part (iii) of the preceding theorem that the mean $\bar{e}$ of E is 0. Let $r_{XE}$ denote Pearson's linear correlation between the explanatory variable X and the residuals E. Let $\widehat{Y}$ denote the distribution of fitted values.

**Theorem 4.** *The linear correlation $r_{XE}$ between the observations of the explanatory variable X and the distribution E of residuals is 0. Additionally, the correlation $r_{\widehat{Y}E}$ between the predicted values of the response variable Y and the residuals E is also 0:*

$$r_{XE} = 0 \qquad \textit{and} \qquad r_{\widehat{Y}E} = 0.$$

*Proof.* For the values

$$m = r_{\mathrm{XY}}\,\frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}} \qquad \text{and} \qquad b = \bar{y} - r_{\mathrm{XY}}\,\frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}}\,\bar{x}$$

that minimize $\mathrm{SSE}(m, b)$, we have

$$e_j = y_j - \widehat{y}_j = y_j - \left( r_{\mathrm{XY}}\,\frac{s_Y}{s_X}\,x_j + \bar{y} - r_{\mathrm{XY}}\,\frac{s_Y}{s_X}\,\bar{x} \right) = (y_j - \bar{y}) - r_{\mathrm{XY}}\,\frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}}\ (x_j - \bar{x})\,.$$

One more preliminary: Equation (5.1.4) tells us that $\bar{e} = 0$. If we use the formula for $e_j$ from the preceding line and subtract 0 in the form of $\bar{e}$ from the left side, we obtain

$$e_j - \bar{e} = (y_j - \bar{y}) - r_{\mathrm{XY}}\,\frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}}\ (x_j - \bar{x}) \tag{5.1.5}$$

It follows from (5.1.5) that

$$\begin{aligned}
(N-1)\, s_{\mathrm{E}}\, s_{\mathrm{X}}\, r_{\mathrm{E,X}} &= \sum_{j=1}^{N} (e_j - \bar{e})\,(x_j - \bar{x}) \\
&= \sum_{j=1}^{N} \Big( (y_j - \bar{y}) - r_{\mathrm{XY}}\, \frac{s_Y}{s_X}\, (x_j - \bar{x}) \Big)\,(x_j - \bar{x}) \\
&= \sum_{j=1}^{N} \Big( (y_j - \bar{y})\,(x_j - \bar{x}) \Big) - r_{\mathrm{XY}}\, \frac{s_Y}{s_X} \sum_{j=1}^{N} (x_j - \bar{x})^2 \\
&= (N-1)\, s_{\mathrm{X}}\, s_{\mathrm{Y}}\, r_{\mathrm{XY}} - r_{\mathrm{XY}}\, \frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}}\, (N-1)\, s_{\mathrm{X}}^2 \\
&= 0,
\end{aligned}$$

which shows that $r_{\mathrm{E,X}} = 0$.

For the second correlation assertion, we have

$$\begin{aligned}
r_{\hat{\mathrm{Y}}\mathrm{E}} &= \frac{1}{N-1} \sum_{j=1}^{N} \left( \frac{\widehat{y}_j - \bar{\widehat{y}}}{s_{\widehat{Y}}} \right) \left( \frac{e_j - \bar{e}}{s_{\mathrm{E}}} \right) \\
&= \frac{1}{N-1}\, \frac{1}{s_{\widehat{\mathrm{Y}}} s_{\mathrm{E}}} \sum_{j=1}^{N} (m\, x_j + b - \bar{y})\,(e_j - \bar{e}) \\
&= \frac{1}{N-1}\, \frac{1}{s_{\widehat{\mathrm{Y}}} s_{\mathrm{E}}} \sum_{j=1}^{N} (m\, x_j + b - \bar{y})\; e_j \\
&= \frac{1}{N-1}\, \frac{1}{s_{\widehat{\mathrm{Y}}} s_{\mathrm{E}}} \sum_{j=1}^{N} (m\, x_j e_j + (b - \bar{y})\, e_j) \\
&= \frac{1}{N-1}\, \frac{1}{s_{\widehat{\mathrm{Y}}} s_{\mathrm{E}}} \left( m \sum_{j=1}^{N} x_j e_j - (b - \bar{y}) \sum_{j=1}^{N} e_j \right) \\
&= \frac{1}{N-1}\, \frac{1}{s_{\widehat{\mathrm{Y}}} s_E}\, (m \cdot 0 - (b - \bar{y}) \cdot 0) \\
&= 0.
\end{aligned}$$

□

## The Regression Line for z-Scores

Suppose that we want to work with the standardized variables

$$z_{\mathrm{X}} = \frac{x - \bar{x}}{s_{\mathrm{X}}} \qquad \text{and} \qquad z_{\mathrm{Y}} = \frac{y - \bar{y}}{s_{\mathrm{Y}}}.$$

There are two easy ways to find the regression line for these standard scores. The regression line for $\mathrm{Y} \sim \mathrm{X}$,

$$\widehat{y} = r_{\mathrm{XY}} \frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}}\, x + \bar{y} - r_{\mathrm{XY}} \frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}}\, \bar{x}$$

can be written as

$$\widehat{y} - \bar{y} = r_{\mathrm{XY}} \frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}}\, (x - \bar{x})$$

or

$$\frac{\widehat{y} - \bar{y}}{s_{\mathrm{Y}}} = r_{\mathrm{XY}} \frac{x - \bar{x}}{s_{\mathrm{X}}}$$

or

$$\widehat{z}_{\mathrm{Y}} = r_{\mathrm{XY}}\, z_{\mathrm{X}}. \tag{5.1.6}$$

An alternative derivation of equation (5.1.6) is just as easy. We use the equation

$$\widehat{y} = r_{\mathrm{XY}} \cdot \frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}} \, (x - \bar{x}) + \bar{y},$$

replacing the parameters for X and Y with corresponding parameters for $Z_{\mathrm{X}}$ and $Z_{\mathrm{Y}}$. The means of the standardized variables are both 0 (so $\bar{x}$ and $\bar{y}$ are replaced by 0) and the standard deviations are both 1 (so $s_{\mathrm{X}}$ and $s_{\mathrm{Y}}$ are both replaced by 1). The corresponding formula for z-scores is therefore

$$\widehat{z}_{\mathrm{Y}} = r_{z_{\mathrm{X}} z_{\mathrm{Y}}} \, z_{\mathrm{X}}.$$

We obtain equation (5.1.6) from this equation by using the equality of the correlations $r_{z_{\mathrm{X}} z_{\mathrm{Y}}}$ and $r_{\mathrm{XY}}$: $r_{z_{\mathrm{X}} z_{\mathrm{Y}}} = r_{\mathrm{XY}}$.

Two derivations of equation (5.1.6) certainly suffice. Nevertheless, a third derivation might also be eduactional. Suppose that we are back at the beginning and we do not know the slope of the regression line for the standardized variables $Z_{\mathrm{X}}$ and $Z_{\mathrm{Y}}$. We *will* use our knowledge that the point $(\bar{z}_{\mathrm{X}}, \bar{z}_{\mathrm{Y}})$, which is to say $(0, 0)$, lies on the regression line. This fact tells us that 0 is the $z_{\mathrm{Y}}$-intercept. The line we seek has the equation $\widehat{z}_{\mathrm{Y}} = m_z \, z_{\mathrm{X}}$. We seek the value of $m_z$ that minimizes the sum of the square errors $\mathrm{SSE}_z$ that result from the linear model $\widehat{z}_{\mathrm{Y}} = m_z \, z_{\mathrm{X}}$. Note: the observations, the regression line, the fitted values, the residuals, and the sum of the squared errors for the standardized variables $Z_{\mathrm{X}}$ and $Z_{\mathrm{Y}}$ all differ from the corresponding values for the raw variables X and Y. It is for that reason that we will write $m_z$ and $\mathrm{SSE}_z$ to designate the regression line slope and the sum of squared errors for the standardized variables.

To find the value of $m_z$ that minimizes $\mathrm{SSE}_z \, (m_z)$, we will rely on three equations involving the $z_{\mathrm{X}}$ and $z_{\mathrm{Y}}$ variables. The first two of these equations apply to any standardized variable, which, by virtue of the standardization, has sample mean 0 and sample standard deviation 1. (We are not making any assumption about normality, nor are we claiming that the z-scores are standard normal.) Because the variance of the z-scores of Z is 1, we have

$$1 = \frac{1}{N-1} \sum \left( z_{\mathrm{X}} - \overline{z_{\mathrm{X}}} \right)^2 = \frac{1}{N-1} \sum \left( z_{\mathrm{X}} - 0 \right)^2 = \frac{1}{N-1} \sum z_{\mathrm{X}}^2.$$

The same equation holds for $z_{\mathrm{Y}}$, which also has variance 1. Thus,

$$\frac{1}{N-1} \sum z_{\mathrm{X}}^2 = 1 \qquad \text{and} \qquad \frac{1}{N-1} \sum z_{\mathrm{Y}}^2 = 1. \tag{5.1.7}$$

The last ingredient we need is the formula for Pearson's linear correlation coefficient $r_{\mathrm{XY}}$ between X and Y (which we will write simply as $r$):

$$\frac{1}{N-1} \sum z_{\mathrm{X}} \, z_{\mathrm{Y}} = r. \tag{5.1.8}$$

We are now ready to minimize the sum of the squares of the residuals in $z_{\mathrm{X}} z_{\mathrm{Y}}$-coordinates. Bare in mind that this quantity, $\mathrm{SSE}_z \, (m_z)$, is composed of the square errors between the z-scores of the response variable and the z-scores predicted by the linear model in the z-score plane. To simplify $\mathrm{SSE}_z \, (m_z)$, we will use the three equations of lines (5.1.7) and (5.1.8). After the simplification, the minimization of $\mathrm{SSE}_z \, (m_z)$ is achieved by completing a square.

**Theorem 5.** *Let $r$ be Pearson's linear correlation coefficient of the $N$ points $(x_1, y_1)$, $(x_2, y_2)$, $\ldots$, $(x_N, y_N)$ in the xy-plane. Let $Q_1, Q_2, \ldots, Q_N$ be the points in the $z_X z_Y$-plane that result from transforming each $x_j$ and $y_j$ into its z-score. Let $SSE_z \, (m_z)$ be the sum of the square errors with respect to the points $Q_1, Q_2, \ldots, Q_N$ and the line $\widehat{z}_Y = m_z \, z_X$ in the $z_X z_Y$-plane. Then the minimum value of $SSE_z \, (m_z)$ occurs when $m_z = r$. Furthermore,*

$$\frac{1}{N-1} SSE_z(r) = 1 - r^2. \tag{5.1.9}$$

*The regression line in the $z_X z_Y$-plane has equation*

$$\widehat{z}_{_Y} = r \, z_{_X}. \tag{5.1.10}$$

*Proof.* In view of the factor $1/(N-1)$ in lines (5.1.7) and (5.1.8), we scale $\mathrm{SSE}_z\,(m_z)$ by this factor as well:

$$\begin{aligned}
\frac{1}{N-1}\mathrm{SSE}_z\,(m_z) &= \frac{1}{N-1}\sum\left(z_{\mathrm{Y}}-m_z z_{\mathrm{X}}\right)^2 \\
&= \frac{1}{N-1}\sum\left(z_{\mathrm{Y}}^2-2\,m_z\,z_{\mathrm{X}}\,z_{\mathrm{Y}}+m_z^2 z_{\mathrm{X}}^2\right) \\
&= \frac{1}{N-1}\sum z_{\mathrm{Y}}^2-2\,m_z\frac{1}{N-1}\sum z_{\mathrm{X}}\,z_{\mathrm{Y}}+m_z^2\frac{1}{N-1}\sum z_{\mathrm{X}}^2 \\
&= 1-2\,r\,m_z+m_z^2 \qquad \text{(using (5.1.7) and (5.1.8))} \\
&= \left(m_z^2-2\,r\,m_z+r^2\right)+1-r^2 \qquad \text{(completing the square)} \\
&= \left(m_z-r\right)^2+\left(1-r^2\right).
\end{aligned}$$

Let us restate what we have just proved, leaving out all the intermediate lines:

$$\frac{1}{N-1}\mathrm{SSE}_z\,(m_z)=\left(m_z-r\right)^2+\left(1-r^2\right). \tag{5.1.11}$$

We can now make several observations. The expression $\left(m_z-r\right)^2$, being a square, cannot be negative. Therefore, the right side of equation (5.1.11) is minimized when $\left(m_z-r\right)^2=0$, which occurs for $m_z=r$ and results in equation (5.1.9). Equation (5.1.10) is an immediate consequence. □

Equation (5.1.9) allows us to deduce an important inequality that we asserted without proof in Chapter 4. Because the left side of equation (5.1.11), a sum of squares, is nonnegative, we see that $1-r^2\geq 0$, or, equivalently, $-1\leq r\leq 1$. This proves the assertion made in Chapter 4 that Pearson's correlation coefficient satisfies the inequality $|\,r\,|\leq 1$. This fact, together with the equation $\widehat{z}_{\mathrm{Y}}=r\,z_{\mathrm{X}}$, gives rise to a remarkable observation. When the data points are not all on a line, in which case $|r|<1$, the regression equation $\widehat{z}_{\mathrm{Y}}=r\,z_{\mathrm{X}}$ tells us that $|\widehat{z}_{\mathrm{Y}}|<|z_{\mathrm{X}}|$: the predicted deviation $z_{\mathrm{Y}}$ of the response variable from the mean is *less* than the deviation $z_{\mathrm{X}}$ of the explanatory variable from the mean.

The least squares line was used in astronomy by Pierre-Simon Laplace and Karl Friedrich Gauss in the late $18^{\text{th}}$ century. It was introduced to statistics in 1886 by Sir Francis Galton, whose paper focused on what he termed the *Law of Regression.* Indeed, the title of his paper was Regression *towards* Mediocrity in Hereditary Stature. Galton's word *regression* has stuck, but we now say *regression towards the mean* instead of *regression towards mediocrity.* Galton discovered the law while studying the relationship between the heights of children and their parents. A plate from his paper appears as Figure 5.1.3. The mid-parental height to which Galton refers is the average of the height of the father and 1.08 times the height of the mother.

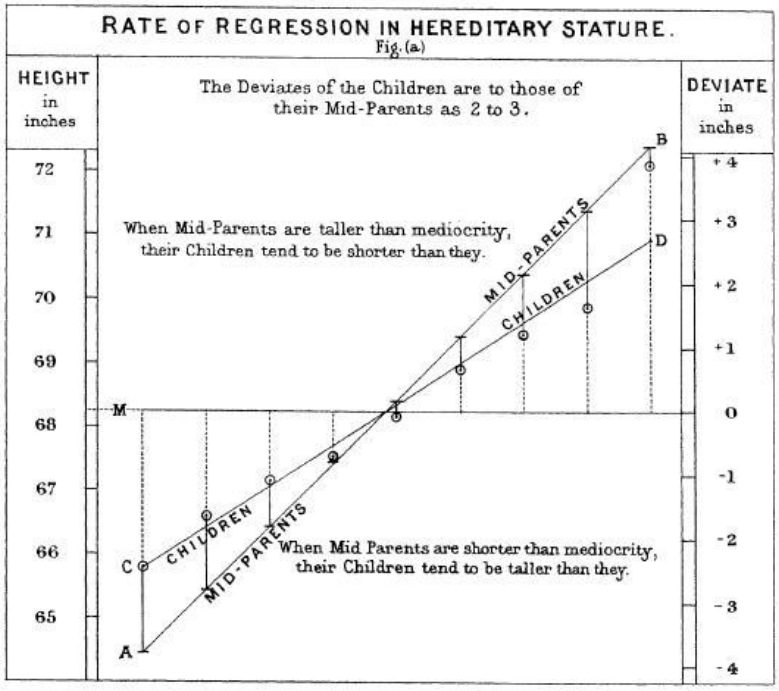



**Figure 5.1.3 Law of Regression as Reported by Galton, 1886**

Before we move on to the next subsection, we will check the theory developed in this section with the data from our cereal study. Table 5.1.4 is a reworking of Table 5.1.3 in which every raw observation has been replaced with its z-score. The errors are the differences $z_y - \widehat{z}_y$, and the square errors are $(\widehat{z}_y)^2$.

| Observation $j$ | $z_x$ | $z_y$ | $\widehat{z}_y$ | $e_j$ | $e_j^2$ |
|---|---|---|---|---|---|
| 1 | -1.468 | .195 | -.740 | .934 | .873 |
| 2 | -1.814 | -3.072 | -.914 | -2.158 | 4.657 |
| 3 | -.948 | .195 | -.478 | .673 | .452 |
| 4 | .668 | -.204 | .337 | -.541 | .292 |
| 5 | -.371 | -.459 | -.187 | -.272 | .074 |
| 6 | 1.072 | .848 | .540 | .308 | .095 |
| 7 | -.660 | -.459 | -.333 | -.126 | .016 |
| 8 | -1.275 | -.916 | -.642 | -.274 | .075 |
| 9 | .861 | -.008 | .434 | -.442 | .195 |
| 10 | 1.447 | -.041 | .729 | -.770 | .593 |
| 11 | .524 | -.505 | .264 | -.768 | .591 |
| 12 | -.487 | .123 | -.245 | .368 | .135 |
| 13 | .437 | -.184 | .220 | -.405 | .164 |
| 14 | .076 | -.936 | .038 | -.974 | .949 |
| 15 | 1.072 | 1.501 | .540 | .961 | .924 |
| 16 | -.083 | .848 | -.042 | .890 | .791 |
| 17 | 2.371 | -.217 | 1.195 | -1.412 | 1.993 |
| 18 | -.397 | .489 | -.200 | .689 | .474 |
| 19 | .350 | .358 | .177 | .181 | .033 |
| 20 | -.556 | .848 | -.280 | 1.128 | 1.273 |
| 21 | .232 | .489 | .117 | .372 | .138 |
| 22 | .235 | .541 | .118 | .422 | .178 |
| 23 | .394 | 1.077 | .198 | .878 | .771 |
| 24 | -1.814 | -2.112 | -.914 | -1.198 | 1.434 |
| 25 | .506 | 1.410 | .255 | 1.155 | 1.333 |
| 26 | -.371 | .195 | -.187 | .382 | .146 |
| | | | | | 18.651 |

**Table 5.1.4: Predicted Caloric Content z-Score $\widehat{z}_y$, Error $e_j$, Square Error $e_j^2$**

The first thing to note is that the value of SSE in the z-score plane, namely 18.651, is very different from the value of SSE in the raw data plane. According to the theory we developed, the z-score SSE should equal $1 - r^2$ when it is divided by $N - 1$. Let us see:

$$\frac{1}{N-1}\text{SSE} = \frac{1}{25}\,18.651 = 0.746 \qquad \text{and} \qquad 1 - r^2 = 1 - (0.50395)^2 = 0.746.$$

## Summary

- Regression line for z-scores: $\widehat{z}_{\text{Y}} = r_{\text{XY}}\, z_{\text{X}}$
- Slope $m$ of Regression Line for Y $\sim$ X: $r_{\text{XY}}\,\dfrac{s_{\text{Y}}}{s_{\text{X}}}$
- Y-intercept $b$ of Regression Line for Y $\sim$ X: $\bar{y} - m\,\bar{x}$, or $\bar{y} - r_{\text{XY}}\,\dfrac{s_{\text{Y}}}{s_{\text{X}}}\,\bar{x}$
- Equation of Regression Line for Y $\sim$ X (Slope-Intercept Form): $\widehat{y} = r_{\text{XY}}\,\dfrac{s_{\text{Y}}}{s_{\text{X}}}\,x + \left(\bar{y} - r_{\text{XY}}\,\dfrac{s_{\text{Y}}}{s_{\text{X}}}\,\bar{x}\right)$
- Equation of Regression Line for Y $\sim$ X (Point-Slope Form): $\widehat{y} = r_{\text{XY}}\,\dfrac{s_{\text{Y}}}{s_{\text{X}}}\,(x - \bar{x}) + \bar{y}$
- Fitted or Predicted Values: $\widehat{y}_j = r_{\text{XY}}\,\dfrac{s_{\text{Y}}}{s_{\text{X}}}\,x_j + \left(\bar{y} - r_{\text{XY}}\,\dfrac{s_{\text{Y}}}{s_{\text{X}}}\,\bar{x}\right)$

- Residuals (Errors): $e_j = y_j - \widehat{y}_j$
- Zero Linear Correlation: X (observations of explanatory variables) and E (errors)
- Zero Linear Correlation: $\widehat{\mathrm{Y}}$ (Fitted Values) and: E (errors)

## Regression Lines in R

Only two new R commands are needed for obtaining and graphing a regression line. If data for the explanatory and response variables `X` and `Y` have been entered in the usual way, namely `X = c(x1, x2, ... , xN)` and `Y = c(y1, y2, ... , yN)`, or, more commonly, from extracting two columns from a table that has been read into the session, then the two parameters of the regression line, slope $m$ and $y$-intercept $b$, are returned by the command `lm( Y ~ X )`. The $y$-intercept of the regression line will appear under `(intercept)` and the slope of the regression line will appear under the name of the explanatory variable. If a scatter plot has been created by the command `plot(X,Y)`, then the command `abline(lm(Y ~ X ))` superimposes a plot of the regression line on the scatter plot. For example, here is a screen capture of the R session used to create Figure 5.1.2.

```
> sugar.calories = read.table("C:/stats/data/cereal26.txt", header = FALSE)
> sugar.calories[1:3 , ]
                        V1  V2  V3
1  General Mills Cheerios 1.2 110
2 General Mills Fiber One 0.0  60
3       General Mills Kix 3.0 110
> sugar = sugar.calories[ , 2]
> calories = sugar.calories[ , 3]
> plot.title = "Scatter Plot of Sugar (Explanatory) & Calories (Response) For 26 Cereals"
> plot.subtitle = "All Measurements per 30g Serving"
> plot(sugar,calories, main = plot.title, xlab = "Sugar (g)", ylab = "Calories",
+ sub = plot.subtitle, pch = 16, col = "tan3")
> sugar.calories.rl = lm( calories ~ sugar )
> sugar.calories.rl

Call:
lm(formula = calories ~ sugar)

Coefficients:
(Intercept)        sugar
     93.029        2.226

> abline( sugar.calories.rl , col = "brown4", lwd = 2)
> legend.txt = "y = 2.2262 x + 93.02884"
> legend("bottomleft", inset = c(0.2,0.1), fill="brown4", legend = legend.txt)
```

In the first line, the sugar and calorie observations have been read in by means of a file stored on the author's hard drive. In the session, the table has been named `sugar.calories`. The command `sugar.calories[1:3 , ]` prints to screen the first three rows. Seeing a few entries of each column can be useful orientation when you have not worked with the table recently. The command `head( sugar.calories )` would have used a built-in function to display six rows. The next two lines extract the second and third columns and name the vectors `sugar` and `calories` respectively. With a view toward requesting a scatterplot, the title and subtitle of the plot are entered and assigned to the user-chosen names `plot.title` and `plot.subtitle`. The input `plot( sugar, calories, ... )` generates a scatterplot. The parameter `pch = 16` specifies that each observation is represented by a plot character that is a filled circle. The code `sugar.calories.rl = lm( calories ~ sugar )` calculates the linear model and assigns the structure to the user-chosen variable `sugar.calories.rl`. When we refer to "structure", we signify that there is a good deal of information stored in `sugar.calories.rl`: much more than the equation calories $= m\,\text{sugar} + b$ of the regression line. Calling on `sugar.calories.rl` results in the screen display the two parameters $m$ and $b$. The input `abline( sugar.calories.rl , col = ''brown4'', lwd = 2)` superimposes the regression line on the scatterplot. Adding a plot legend is the finishing touch.

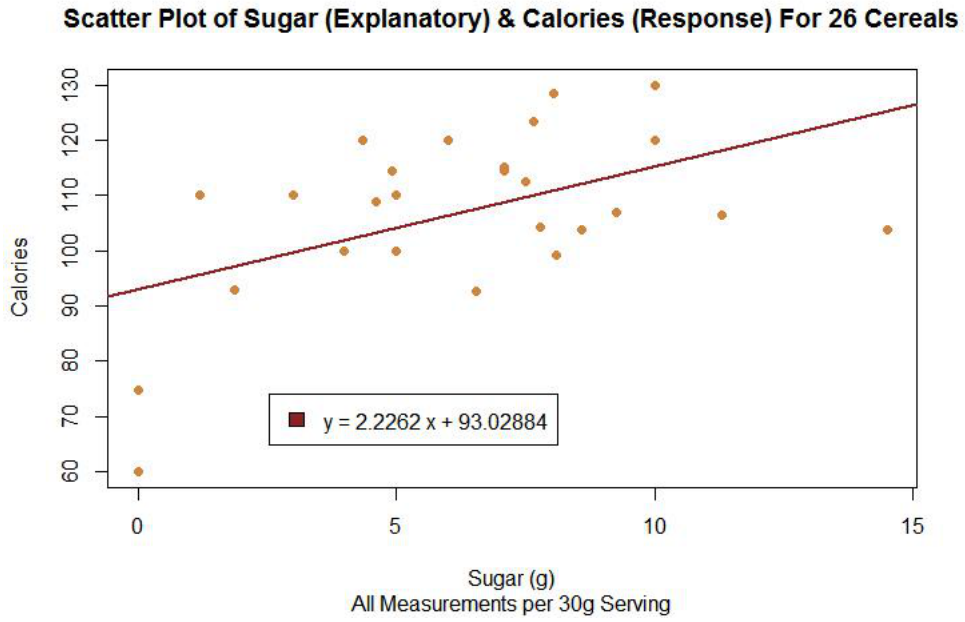



**Figure 5.1.4 Scatter Plot of (Sugar,Calories) with Linear Model**

## 5.2 SST, SSR, and The Coefficient of Determination

The work we have done in the preceding section allows us to apportion contributions to the variance $s_Y^2$ of $Y$. We begin by partitioning 1 into two nonnegative parts,

$$1 = \left(1 - r^2\right) + r^2. \tag{5.2.1}$$

The two summands on the right side of equation (5.2.1) are fractions that combine to make a whole. According to equation (5.1.9), the first of these two summands, namely $1 - r^2$, is equal to $\frac{1}{N-1}\mathrm{SSE}_z$. This is the part of the variability that can be attributed to the *error* of the linear model of the standardized variables. Stated another way, it is the part of the variability that the linear model $z_{\mathrm{Y}} \sim z_{\mathrm{X}}$ does *not* explain. The other summand, $r^2$, must be the part of the variability that the linear model $z_{\mathrm{Y}} \sim z_{\mathrm{X}}$ *does* explain. This fraction, $r^2$, is called the ***coefficient of determination*** and is generally reported when a regression line is used. Sometimes it is reported as the percentage $100\,r^2\%$. We remark that there are other calculations of statistics for which a "coefficient of determination" is defined. The generic symbol for a coefficient of determination is $R^2$, but, unlike the present context, a coefficient of determination is not always equal to $r^2$.

### A Sum of Some Sums: SST = SSR + SSE

We continue to use the notation introduced in Section 1. We will shift our attention back to the raw variables X and Y. For a collection of points $(x_1, y_1)\,,\ (x_2, y_2)\,,\ \ldots\,,\ (x_N, y_N)$ that is modelled by the regression line

$$\widehat{y} = \left(r \cdot \frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}}\right)(x - \bar{x}) + \bar{y},$$

we set

$$\widehat{y}_j = \left(r \cdot \frac{s_Y}{s_X}\right)(x_j - \bar{x}) + \bar{y} \qquad \text{and} \qquad e_j = y_j - \widehat{y}_j.$$

In this subsection, we will state an equation involving the following three sums of squares:

$$\mathrm{SST} = \sum_{j=1}^{N} (y_j - \bar{y})^2 \tag{5.2.2}$$

$$\mathrm{SSR} = \sum_{j=1}^{N} (\widehat{y}_j - \bar{y})^2 \tag{5.2.3}$$

$$\text{SSE} \;= \sum_{j=1}^{N} (y_j - \widehat{y}_j)^2 \,. \tag{5.2.4}$$

Of these three sums, SSE, the Sum of the Squares of the Errors, has already been defined in Section 1. In that section, we actually referred to two different SSEs: one involving the regression line in the $xy$-plane, namely SSE, and one involving the regression line in the $z_X z_Y$-plane, namelt $\text{SSE}_z$. Notice that equation (5.2.4) clearly refers to the sum of the square errors with respect to the regression line in the $xy$-plane. For the other two sums SST and SSR defined by formulas (5.2.3) and (5.2.4), SST stands for the ***Sum of Squares in Total*** and SSR stands for the ***Sum of Squares explained by the Regression line***: from equation (5.2.3), we see that the squares $(\widehat{y}_j - \bar{y})^2$ that contribute to SSR involve the deviations from the mean of the values predicted by the regression line—remember that $\overline{\widehat{y}} = \bar{y}$ by part (ii) of theorem 2 of Section 1. The reason for calling SST a "total" sum of squares can be understood by observing that SST is proportional to the variance of Y:

$$\frac{1}{N-1}\text{SST} \;= \frac{1}{N-1} \sum_{j=1}^{N} (y_j - \bar{y})^2 = s_Y^2. \tag{5.2.5}$$

As a result, every square that is needed to account for the variability of $Y$ can be found in SST. The next subsection is devoted to proving the following important identity, which states that SST is the sum of SSR and SSE:

$$\text{SST} \;= \text{SSR} \;+ \text{SSE} \tag{5.2.6}$$

## Derivation of the Equation SST = SSR + SSE (Optional)

Equation (5.2.6), SST = SSR + SSE, is not obvious. That being the case, a derivation is in order. The key to our derivation will be equation (5.2.7), which will be stated below. The verification of equation (5.2.7) that we are about to provide is not especially enlightening and may be skipped over with no loss of understanding. We begin as follows:

$$\begin{aligned}
\text{SST} &= \sum_{j=1}^{N} (y_j - \bar{y})^2 \\
&= \sum_{j=1}^{N} \Big( (y_j - \widehat{y}_j) + (\widehat{y}_j - \bar{y}) \Big)^2 \\
&= \sum_{j=1}^{N} (y_j - \widehat{y}_j)^2 + 2 \sum_{j=1}^{N} (y_j - \widehat{y}_j)\,(\widehat{y}_j - \bar{y}) + \sum_{j=1}^{N} (\widehat{y}_j - \bar{y})^2 \\
&= \text{SSE} \;+ 2 \sum_{j=1}^{N} (y_j - \widehat{y}_j)\,(\widehat{y}_j - \bar{y}) + \text{SSR}\,.
\end{aligned}$$

We must establish that

$$\sum_{j=1}^{N} (y_j - \widehat{y}_j)\,(\widehat{y}_j - \bar{y}) = 0. \tag{5.2.7}$$

The left side of equation (5.2.7) is seen to be 0 by writing the factor $y_j - \widehat{y}_j$ as $e_j$ and the factor $\widehat{y}_j - \bar{y}$ as $m\,x_j + b - \bar{y}$, or $m\,x_j + \bar{y} - m\,\bar{x} - \bar{y}$, or $m\,x_j - m\,\bar{x}$. Equations (5.1.4) and (5.1.5) finish the calculation:

$$\begin{aligned}
\sum_{j=1}^{N} (y_j - \widehat{y}_j)\,(\widehat{y}_j - \bar{y}) &= \sum_{j=1}^{N} e_j\;(m\,x_j - m\,\bar{x}) \\
&= m \sum_{j=1}^{N} x_j e_j - m\,\bar{x} \sum_{j=1}^{N} e_j \\
&= m \cdot 0 - m\,\bar{x} \cdot 0 \\
&= 0.
\end{aligned}$$

—Student: I got lost in that calculation, somewhere. Can you summarize it?
—Statistician: Certainly! Here is the sum of some sums summary: SST = SSR + SSE

## The Coefficient of Determination ($R^2$)—A Second Look

In the preceding subsection, we went to some trouble to establish that SST = SSR + SSE. We will now use this identity to rigorously characterize the coefficient of determination $r^2$. We again remind the reader that SSE, the sum of squared errors that pertain to the regression line in the $xy$-plane, and $\text{SSE}_z$, the sum of squared errors that pertain to the regression line in the $z_X z_Y$-plane, are not the same sums. Our first step will be to establish the relationship between these two sums, SSE and $\text{SSE}_z$. Doing so allows us to derive a formula for SSE from the known formula for $\text{SSE}_z$.

**Theorem 1.** *The sum SSE of the square errors in the xy-plane of the raw observations, and the sum $SSE_z$ of the square errors in the $z_X z_Y$-plane of the z-scores are related by the equation*

$$SSE = SSE_z \, s_Y^2. \tag{5.2.8}$$

*The sum SSE of the square errors in the xy-plane of the raw observations is given by*

$$\frac{1}{N-1} SSE = \left(1 - r^2\right) \, s_Y^2. \tag{5.2.9}$$

*Proof.* We have

$$\begin{aligned}
\text{SSE} &= \sum_{j=1}^{N} (y_j - \widehat{y}_j)^2 \\
&= \sum_{j=1}^{N} \Big( (y_j - \bar{y}) - (\widehat{y}_j - \bar{y}) \Big)^2 \\
&= \sum_{j=1}^{N} \left( \frac{(y_j - \bar{y})}{s_\text{Y}} - \frac{(\widehat{y}_j - \bar{y})}{s_\text{Y}} \right)^2 s_\text{Y}^2 \\
&= \sum_{j=1}^{N} (z_y - \widehat{z}_y)^2 \; s_\text{Y}^2 \\
&= \text{SSE}_z \, s_\text{Y}^2.
\end{aligned}$$

Now we combine equation (5.2.8) with equation (5.1.9) to obtain formula (5.2.9) for SSE:

$$\frac{1}{N-1} \text{SSE} = \frac{1}{N-1} \text{SSE}_z \, s_\text{Y}^2 = \left(1 - r^2\right) \, s_Y^2.$$

□

**Theorem 2.** *The sum SSR of the squares explained by the regression line is given by*

$$\frac{1}{N-1} SSR = r^2 \, s_Y^2. \tag{5.2.10}$$

*Proof.* The formula for SSR is obtained by combining equations (5.2.6), (5.2.5), and (5.2.10) as follows

$$\begin{aligned}
\frac{1}{N-1}\text{SSR} &= \frac{1}{N-1} (\text{SST} - \text{SSE}) \\
&= \frac{1}{N-1}\text{SST} - \frac{1}{N-1}\text{SSE} \\
&= s_\text{Y}^2 - \left(1 - r^2\right) \, s_\text{Y}^2 \\
&= s_\text{Y}^2 \Big( 1 - \left(1 - r^2\right) \Big) \\
&= r^2 \, s_\text{Y}^2.
\end{aligned}$$

□

Equation (5.2.10) states precisely that the fraction of the variance $s_Y^2$ that the linear model accounts for is $r^2$. Equation (5.2.10) can be useful in other ways. For example, equation (5.2.10) provides another proof of the inequalities $-1 \le r \le 1$:

$$\begin{aligned} r^2\, s_Y^2 &= \frac{1}{N-1}\,\mathrm{SSR} \quad \text{(by equation (5.2.10))} \\ &\le \frac{1}{N-1}\,\mathrm{SSR} + \frac{1}{N-1}\,\mathrm{SSE} \\ &= \frac{1}{N-1}\,\mathrm{SST} \quad \text{(by equation (5.2.6))} \\ &= s_Y^2 \quad \text{(by equation (5.2.5)).} \end{aligned}$$

Dividing each side of this inequality by $s_Y^2$ shows that $r^2 \le 1$ and therefore $-1 \le r \le 1$.

Equation (5.2.10) is also instructive when applied to the two extreme cases, $r^2 = 0$ and $r^2 = 1$. If $r^2 = 0$, then SSR = 0 (from equation (5.2.10)) and therefore SSE = SST (from equation (5.2.6)). This equality between sums of squares tells us that if a linear model were to be (unwisely) imposed on bivariate data with $r = 0$, then the error SSE resulting from the use of the linear model would absorb *all* of the variance of Y. We would therefore not expect the linear model to help us predict a response value $\widehat{y}$ of $y$ corresponding to a value of $x$. Indeed, the regression line when $r = 0$ simplifies to $y = \bar{y}$, an equation of a horizontal line composed of points with the same ordinate $\bar{y}$ for every abscissa $x$. Put another way, the value of $x$ does not help at all in predicting the value of $y$: the linear model makes the same prediction $\widehat{y} = \bar{y}$ for every value of $x$. In the other extreme case, when $r^2 = 1$, the right side of equation (5.2.10) is $s_Y^2$ and so $\mathrm{SSR}/(N-1) = s_Y^2 = \mathrm{SST}/(N-1)$ by equation (5.2.5). It follows that SST = SSR when $r^2 = 1$, and, as a result, SSE = 0. This equation tells us that there are no prediction errors: $\widehat{y}_j = y_j$ for every $j$. Stated differently, *every* observed datum pair lies on the regression line. We have proved another assertion made in Chapter 4: the equation $r = \pm 1$ corresponds to perfect linearity.

Our next theorem has no new information: it is a summary of the relationships between the three sums of squares, SST, SSR, and SSE, and the variance $s_Y^2$.

**Theorem 3.** *The three sums of squares SST, SSR, and SSE are related to the variance $s_Y^2$ by the following equations*

$$\frac{1}{N-1}\,SST = s_Y^2 \qquad \frac{1}{N-1}\,SSR = r^2\, s_Y^2 \qquad \frac{1}{N-1}\,SSE = \left(1 - r^2\right)\, s_Y^2.$$

## Alternative Formulas for the Coefficient of Determination

When we rearrange equation (5.2.10) by solving for $r^2$, we obtain

$$r^2 = \frac{\mathrm{SSR}}{(N-1)\, s_Y^2}.$$

Equation (5.2.5) allows us to identify the denominator of this fraction:

$$r^2 = \frac{\mathrm{SSR}}{\mathrm{SST}}. \tag{5.2.11}$$

In view of this equation, some textbook authors define the coefficient of determination by

$$\text{Coefficient of Determination} = R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}}.$$

Alternatively, because SSR = SST - SSE, we have

$$\text{Coefficient of Determination} = R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}.$$

The last of these equations can be used to highlight a common misunderstanding in which it is presumed that the greater $R^2$ is, the better the regression line fits the data. *Not necessarily!* Consider the bivariate data (2,3), (4,3), (10,11). For this data set, the least squares line is $y = (14\,x - 1)/13$, the residuals are 12/13, -16/13, and 4/13, SST = 42.6667, SSE = 2.4615, and $R^2 = 0.9423$. Now consider the bivariate data (2,101), (4,199), (10,501). For this data set, the least squares line is $y = (651\,x - 1)/13$, the residuals are 12/13, -16/13, and 4/13, SST = 86936.0, SSE = 2.4615, and $R^2 = 0.99997$. Note the identical residuals and, consequently, the same SSE. The fits of the regression lines to the observed data are identical. Yet one coefficient of determination exceeds the other.

### The Standard Error

The standard deviation $s_Y$ of Y is defined by

$$s_Y = \sqrt{\frac{1}{N-1}\,\mathrm{SST}}.$$

The ***standard error*** SE of the linear model is defined analogously:

$$\mathrm{SE} = \sqrt{\frac{1}{N-2}\,\mathrm{SSE}}. \tag{5.2.12}$$

Why is the denominator $N-2$ and neither $N$ nor $N-1$? It is because *two* degrees of freedom among the $N$ errors are lost. Together, the equations (5.1.4) and (5.1.5) imply that two of the residuals can be determined from the other $N-2$ residuals. For example, if $N=3$, then $e_1$ can have any value but the equations $e_1+e_2+e_3=0$ and $x_1\,e_1+x_2\,e_2+x_3\,e_3=0$ determine the values of $e_2$ and $e_3$ to be

$$e_2 = e_1\,\frac{(x_3-x_1)}{(x_2-x_3)} \qquad \text{and} \qquad e_3 = e_1\,\frac{(x_1-x_2)}{(x_2-x_3)}.$$

The standard error is sometimes expressed in terms of the standard deviation $s_Y$:

$$\mathrm{SE} = \sqrt{\frac{N-1}{N-2}\,(1-r^2)}\;s_Y. \tag{5.2.13}$$

We obtain equation (5.2.13) from (5.2.12) by using formula (5.2.9) for SSE:

$$\mathrm{SE} = \sqrt{\frac{1}{N-2}\,\mathrm{SSE}} = \sqrt{\frac{1}{N-2}\,(N-1)\,(1-r^2)\,s_Y^2} = \sqrt{\frac{N-1}{N-2}\,(1-r^2)}\;s_Y.$$

## 5.3 And the Crooked Shall be Made Straight: Transformations

With some familiarity with certain planer curves and some skill in algebra, it is possible to transform nonlinear data to linear data, find a linear model, and to transform back. We will illustrate the techniques involved with some examples.

### An Exponential Model

Table 5.3.1 shows the median Miami home price in July for each of the first seven years of the twenty-first century. The table is followed by a scatterplot of the data.

| x = Year | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|---|---|
| y = Price | 207757 | 229282 | 257648 | 290270 | 339682 | 433140 | 469748 |

**Table 5.3.1: Median Miami Home Price**

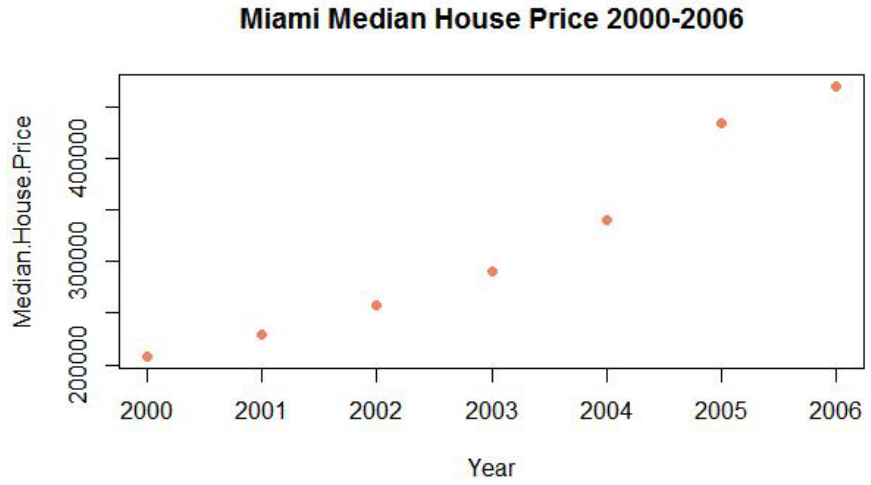



**Figure 5.3.1 Scatter Plot of Year (Predictor) and Median Miami Home Price (Response), 2000-2006**

The linear association is very strong: $r = 0.975$. Based on the scatterplot as well as a Pearson correlation coefficient near 1, it is reasonable to fit the data with a linear model. In Figure 5.3.2, the regression line, $y = 45562\,x - 90941538$, is superimposed on the scatterplot.

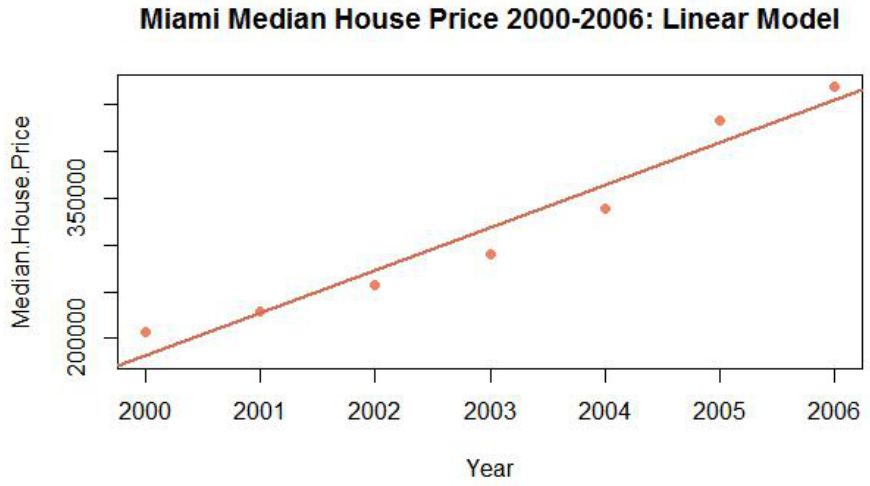



**Figure 5.3.2 Regression Line and Scatter Plot of Year and Median Miami Home Price, 2000-2006**

Here is the R session from which this plot was obtained:

```
> Year = c(2000, 2001, 2002, 2003, 2004, 2005, 2006)
> Median.House.Price = c(207757, 229282, 257648, 290270, 339682, 433140, 469748)
> cor(Year, Median.House.Price)
[1] 0.9746227
> # High linear correlation between Year and Median.House.Price
> plot.title = "Miami Median House Price 2000-2006: Linear Model"
> plot(Year, Median.House.Price, xlim = c(2000,2006), ylim = c(180000,470000),
+ main = plot.title, pch = 19, col = "salmon2")
> Miami.lm = lm( Median.House.Price ~ Year )
> abline( Miami.lm, col = "salmon3", lwd = 2)
```

The regression line fit of the observed points is not terrible. And yet, it is undeniable that the points of the scatterplot in Figure 5.3.1 do not seem to be arranged along a straight line. When such a thing happens, it helps to have a good library of curves in one's memory: the shapes of the curves and their equations as well. In the case of the median Miami house price data, the scatterplot looks more like the graph of an exponential function than a straight line. We therefore will consider a model of the form

$$y = C\,\exp(k\,x) \tag{5.3.1}$$

where $k$ and $C$ are positive constants, and where $x$ and $y$ represent Year and Price respectively. Our theory of regression lines allows us to fit scatterplots with least squares lines—it does not tell us anything about other models such as exponential models. We therefore seek a way to transform our goal model for the original data into a linear of transformed data. In this case applying the natural logarithm to each side of equation (5.3.1) does the job. Using key properties of the logarithm function, we obtain

$$\begin{aligned} \ln(y) &= \ln\left(C\,\exp(k\,x)\right) \\ &= \ln(C) + \ln\left(\exp(k\,x)\right) \\ &= \ln(C) + k\,x \end{aligned}$$

or

$$Y = b + m\,X \tag{5.3.2}$$

where $Y = \ln(y)$, $b = \ln(C)$, $m = k$, and $X = x$. We are in business! Equation (5.3.2) is linear. Once we find $m$ and $b$, we can immediately reverse the transformation and obtain equation (5.3.1). To do so, we use the data in Table 5.3.2:

| X = Year | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|---|---|
| Y = ln(Price) | 12.24412 | 12.34271 | 12.45935 | 12.57857 | 12.73577 | 12.97882 | 13.05995 |

**Table 5.3.2: Year and ln(Price)**

The linear correlation between Year and ln(Price) is extremely close to 1: $r_{XY} = 0.9908465$. The linear model is $Y = -273.2363978 + 0.1427184\,X$. It follows that $C = \exp(-273.2363978)$ and

$$\text{Price} = \exp(-273.2363978)\,\exp(0.1427184\,\text{Year}).$$

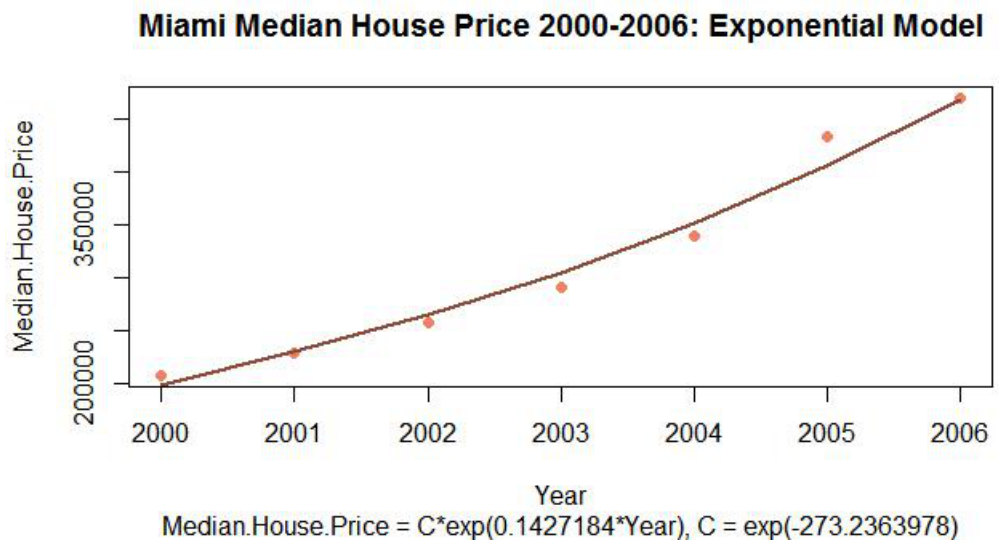



**Figure 5.3.3 Exponential Model and Scatter Plot of Year and Median Miami Home Price, 2000-2006**

The following is a continuation of the R session that produced the linear model.

```
> # Not too bad ... But curved scatter plot suggests the model y = C*exp(k*x), not y = m*x + b
> cor( log(Year), log(Median.House.Price))
[1] 0.990808
> # Higher correlation! Proceed with exponential model
> # y = C*exp(k*x) implies log(y) = log(C) + k*x,
> #    or Y = b + m*X where Y = log(y), b = log(C), m = k, and X = x.
> Miami.House.Model = lm( log( Median.House.Price ) ~ Year )
> Miami.House.Model$coefficients
 (Intercept)         Year
-273.2363978    0.1427184
> C = exp(-273.2363978)
> k = 0.1427184
> exp.model = function(yr){ C*exp(k*yr) }
> plot(Year, Median.House.Price, pch = 19, col = "salmon2",
+ main = "Miami Median House Price 2000-2006: Exponential Model")
> lines( Year, exp.model(Year), col = "salmon4", lwd = 2, type = "l")
> title(sub = "Median.House.Price = C*exp(0.1427184*Year), C = exp(-273.2363978)")
```

## A Higher Parabolic Model

Table 5.3.3 shows the orbital periods (measured in Earth years) and orbital semimajor axes[1] (measured in millions of miles) of the nine planets (including demoted dwarf planet Pluto). Figure 5.3.4, which follows the table, shows a scatter plot of the data.

[1]Planets travel around the sun in ellipses. The longer central axis, which passes through the foci of the ellipse, is called the major axis. Half its length is called the semimajor axis.

| Planet | Mercury | Venus | Earth | Mars | Jupiter | Saturn | Uranus | Neptune | Pluto |
|---|---|---|---|---|---|---|---|---|---|
| Semimajor Axis | 36 | 67 | 93 | 142 | 484 | 887 | 1784 | 2796 | 3666 |
| Period | 0.24 | 0.61 | 1 | 1.88 | 11.86 | 29.46 | 84.07 | 164.82 | 247.68 |

**Table 5.3.3: Orbital Periods and Semimajors Axes of the Planets**

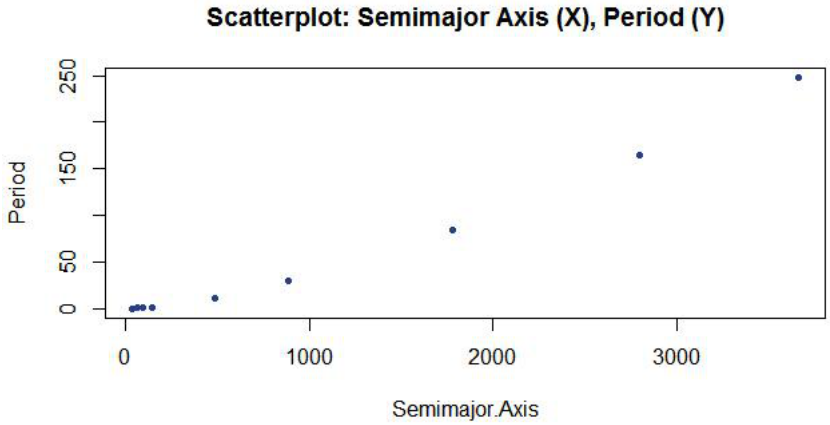



**Figure 5.3.4 Scatter Plot of Semimajor Axis (Predictor) and Orbital Period (Response)**

The scatter plot does not rule out a linear model, but an examination of the data does. For example, the slopes of the line segments joining the point for Mercury to the point for Earth and the point for Saturn to the point for Neptune are 0.013 and 0.071, respectively. A better fit would seem to be a model of the form $y = C\,x^p$ for some unknown values of $C$ and $p$. By applying the logarithm to each side, we can transform the equation into a linear equation. Letting $X = \ln(x)$, $Y = \ln(y)$, $m = p$, and $b = \ln(C)$, we have

$$\begin{aligned} Y &= \ln(y) \\ &= \ln\left(C\,x^p\right) \\ &= \ln\left(C\right) + \ln\left(x^p\right) \\ &= \ln\left(C\right) + p\,\ln(x) \\ &= m\,X + b. \end{aligned}$$

In Table 5.3.4 we have tabulated the natural logarithms of the entries of the preceding table.

| Planet | Mercury | Venus | Earth | Mars | Jupiter | Saturn | Uranus | Neptune | Pluto |
|---|---|---|---|---|---|---|---|---|---|
| ln(Semimajor Axis) | 3.584 | 4.205 | 4.533 | 4.956 | 6.182 | 6.788 | 7.487 | 7.936 | 8.2076 |
| ln(Period) | -1.427 | -0.4943 | 0.0 | 0.6313 | 2.473 | 3.383 | 4.432 | 5.105 | 5.512 |

**Table 5.3.4: ln(Orbital Periods) and ln(Semimajors Axes) of the Planets**

Pearson's $r$ for the transformed data is 0.9999999289. The implication of finding a correlation so close to 1 is that we have found a genuine line—not merely a linear model. The regression line is $Y = 1.501\,X - 6.805$, and when we transform back to the original variables by exponentiating, we obtain

$$\begin{aligned} \text{Orbital Period} &= y \\ &= \exp\big(\ln(y)\big) \\ &= \exp(Y) \\ &= \exp(1.501\,X - 6.805) \\ &= \exp(-6.805)\exp(1.501\,X) \\ &= 0.001108\,\exp(1.501\,\ln(x)) \\ &= 0.001108\,\exp\left(\ln\left(x^{1.501}\right)\right) \\ &= 0.001108\,x^{1.501} \\ &= 0.001108\,\text{Semimajor Axis}^{(3/2)}. \end{aligned}$$

The quation we have just found is Kepler's Third Law. In Figure 5.3.5 we have superimposed the graph of the equation Orbital Period $= 0.001108\,\text{Semimajor Axis}^{(3/2)}$, a so-called *higher parabola*, on the scatterplot.

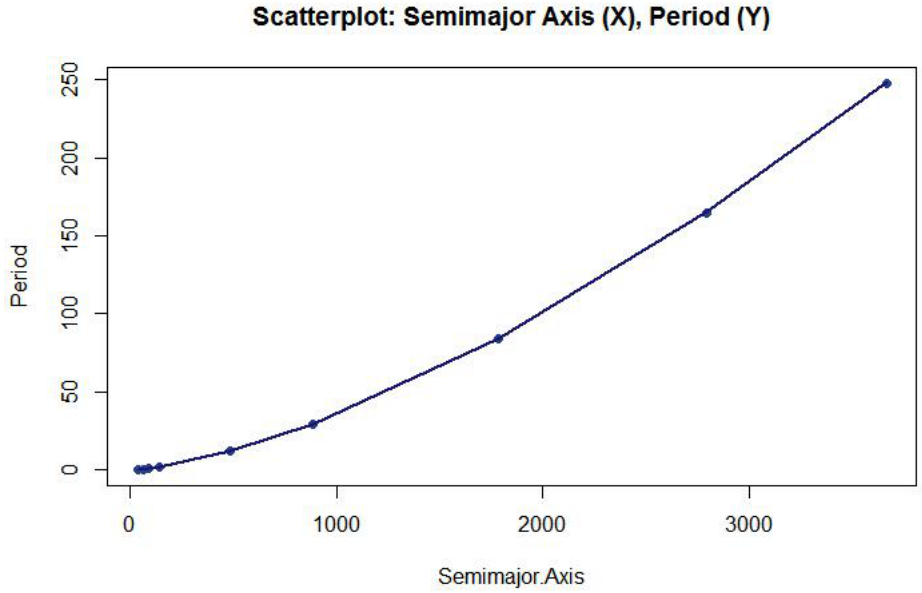



**Figure 5.3.5 Higher Parabolic Model of Orbital Period as Response to Semimajor Axis**

Here is the R code that produced Figure 5.3.5

```
> Semimajor.Axis = c(36, 67, 93, 142, 484, 887, 1784, 2796, 3666)
> Period = c(0.24, 0.61, 1, 1.88, 11.86, 29.46, 84.07, 164.82, 247.68)
> plot(Semimajor.Axis,Period, pch = 16, col = "royalblue4",
+ main = "Scatterplot: Semimajor Axis (X), Period (Y)")
> lm(log(Period) ~ log(Semimajor.Axis))

Call:
lm(formula = log(Period) ~ log(Semimajor.Axis))

Coefficients:
        (Intercept)  log(Semimajor.Axis)
             -6.805                1.501

> lines(Semimajor.Axis, exp(- 6.805)*Semimajor.Axis^1.501, type = "l", lwd=2, col = "midnightblue")
```

Let's try the equation out on Ceres, the largest asteroid (aka minor planet aka dwarf planet) in the asteroid belt between Mars and Jupiter. Its semimajor axis is 257.2 million miles. Plug this value into the formula we found and the result is that the orbital period is 4.57 Earth years. In fact, the orbital period is 1681 Earth days, or 4.61 Earth years.

## 5.4. Regression Considerations

Bivariate data $(x_1, y_1)\,,\,(x_2, y_2)\,,\,\ldots\,,(x_N, y_N)$ is collected, a scatter plot is produced, a linear association is perceived, Pearson's $r$ corroborates the relationship, and a regression line is calculated. End of story? Not so fast!

### The Plot of Residuals

After a regression line has been calculated, it is good practice to generate a scatter plot of the residuals. Recall that the $j^{\text{th}}$ residual $e_j$ is defined by $y_j - \widehat{y}_j$ where $\widehat{y}_j$ is the ordinate of the point on the regression line that has abscissa $x_j$. There are two common ways to produce a plot of residuals. Either the points $(x_1, e_1)\,,\,(x_2, e_2)\,,\,\ldots\,,(x_N, e_N)$ are plotted, or the points $(\widehat{y}_1, e_1)\,,\,(\widehat{y}_2, e_2)\,,\,\ldots\,,(\widehat{y}_N, e_N)$ are plotted. It does not matter which of these two types

of residual plots is investigated. For the purpose of definiteness, we will adopt the former: for us, the abscissas of the plotted points will be the X observations.

What are we looking for in a plot of residuals? Absolutely nothing! If a linear model is a good one, then there should be no visible patterns in the plot of residuals: the points will be uniformly distributed throughout the top and bottom halves of the viewing rectangle. The plot will not thicken[2] anywhere, nor will there be any voids. Any feature of the plot that catches our eye should be explained, if possible.[3]

Figure 5.4.1 shows a plot of the residuals of the 26 cereals considered in Section 5.1. The abscissas of the plotted points are the observations of the explanatory variable. A grey rectangle has been superimposed to highlight the region that contains the main swarm of points. Two points, marked in scarlet, stand out for being distant from the pack. A large void, delineated by a scarlet rectangle, is also very noticeable.

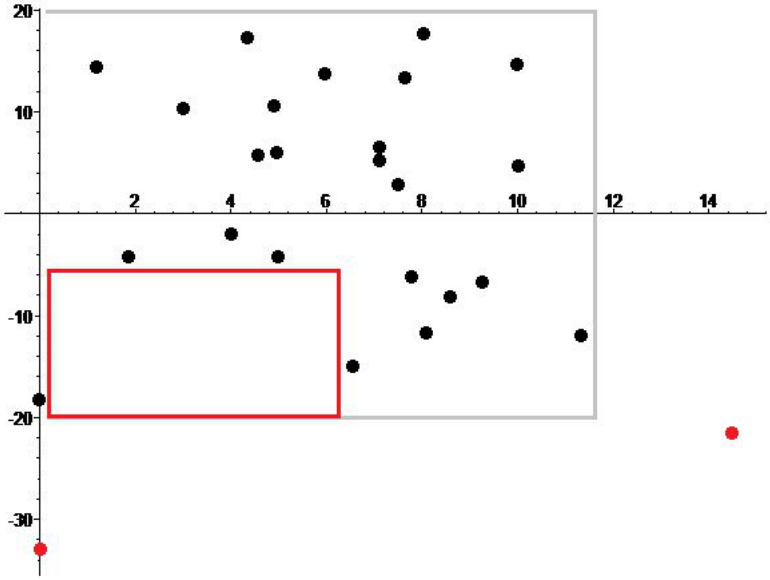



**Figure 5.4.1 Scatter Plot of (Sugar,Residuals)**

The two visual features that stand out suggest a flaw in our model. The two points painted red both have large negative residuals. That means that the predicted caloric content for each of those cereals is substantially greater than the actual content. Why are those actual contents so much lower than the predictions produced by our model?

Recall that in Section 5.1 we used the cereal data to obtain the regression line $y = 2.226\,x + 93.03$. Think about the $y$-intercept, 93.03, which provides a clue. The $y$-intercept represents the predicted calories of a cereal with no sugar. Based on the data, the linear model has a built-in "start-up" caloric cost, which is augmented by the addition of sugar. Perhaps the two highlighted cereals, Quaker Oats Puffed Rice and Malt-O-Meal Golden Puffs, have very little of the other ingredients that contribute to the start-up caloric content of cereal. Indeed, fat is a high calorie component of food, and both of our stand-out cereals, Puffed Rice and Golden Puffs, have 0 fat content.

The considerations just introduced can also account for the empty rectangle that encloses most of the region at the lower left end of the residual plot. This is the region that corresponds to cereals with negative residuals (less than the fat content built into the model) and lower than average sugar content. Why are there no cereals represented by points in the rectangle? Well, such cereals might exist, but, by chance, were not among the 26 cereals selected for our sample. However, our sample does include ten cereals that lie above the rectangle. These cereals would have fallen inside the rectangle had they had lower fat content. Why don't they? The evidence is that cereal producers are using the reduction of sugar content to allow the inclusion of fats that are considered heart-healthy. For example, Post's promotion of *Post Great Grains Honey, Oats, Seeds*, a cereal that would have fallen inside the rectangle had it had lower fat content, states "While some of the competition add artificial sweeteners, we add nutritious fruit, nuts and/or seeds.[4]" The list of ingredients includes almonds, pumpkin seeds,

[2]The author does not disapprove of either clichés or the puns that are based on them.

[3]According to Theorem 3 of Section 1, both the distribution of explanatory observations and the distribution of regression line predictions have Pearson correlation 0 with the distribution of residuals. It is this 0 correlation that tells us that there should be no discernible pattern in either of the two possible plots of the residuals.

[4]`http://www.postfoods.com/our-brands/great-grains/protein-blend-honey,-oats-seeds/`

sunflower seeds, and flaxseed, all of which are considered to be healthy foods despite their relatively high fat content.

Even an imperfect linear model can be enlightening, if only to prompt us to think more deeply about what we are investigating.

## Regression Outliers

A point in a scatter plot that is distant from the regression line is called a ***regression outlier***. It is often the case that a regression outlier will repay investigation with an interesting tale. The twenty-first century was only eleven months old when the regression outlier of the century occurred.

On November 7, 2000, Americans elected Republican candidate George W. Bush the new president. Although the Democratic candidate, Al Gore, won the popular vote, the presidency was decided by the electoral vote, and that hinged on swing state Florida. When the dust settled, the official Florida count was 2,912,790 votes for Bush and 2,912,253 votes for Gore. The official margin of official victory for Bush was 537 votes.

Several issues concerning the Florida vote can be debated, but one issue is really not in doubt: Al Gore lost a few thousand votes in Palm Beach County due to voter confusion. The background of this story is that there were several other candidates for the presidency. Consumer advocate Ralph Nader placed third, but our story concerns Reform Party candidate, Pat Buchanan, who finished fourth. In the scatter plot shown in Figure 5.4.2, the $x$-axis represents the vote count for Bush for each of the 67 counties in Florida. The $y$-axis represents the vote count for Buchanan.

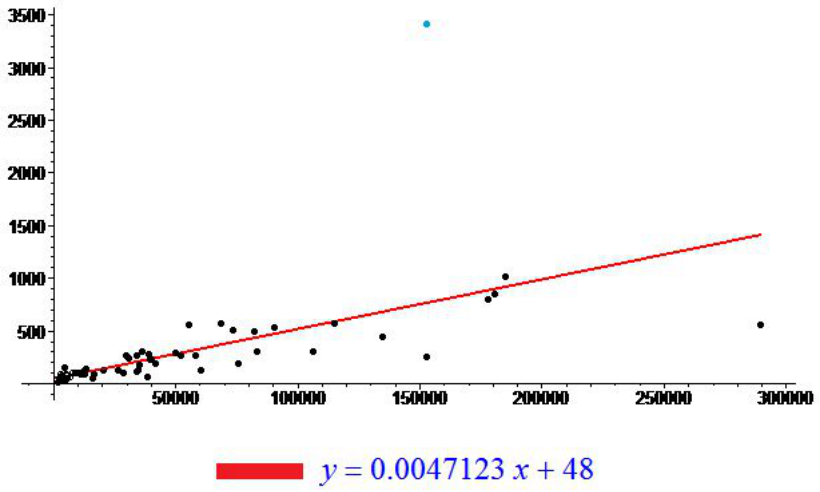



**Figure 5.4.2 Scatter Plot of (Votes for Bush,Votes for Buchanan)**

Pearson's $r$ for the data is 0.6. The regression line, $y = 0.0047123\,x + 48$, is superimposed. The quickest of glances will bring to mind the Sesame Street song, *One of These Points is Not Like the Others*. It is the point representing the voters of Palm Beach County. In that county, Bush received 152,964 votes. The regression line predicts 769 votes for Buchanan. A political analysis of that county predicted even fewer: it was a county with liberal leanings and Buchanan was an arch conservative; it was a county with a large Jewish population and Buchanan was widely regarded as antisemitic. The Reform Party's own polling predicted 400–500 votes. And yet, Buchanan received 3407 votes.

How can one explain the discrepancy? The explanation is the now-infamous "butterfly ballot." See Figure 5.4.3.

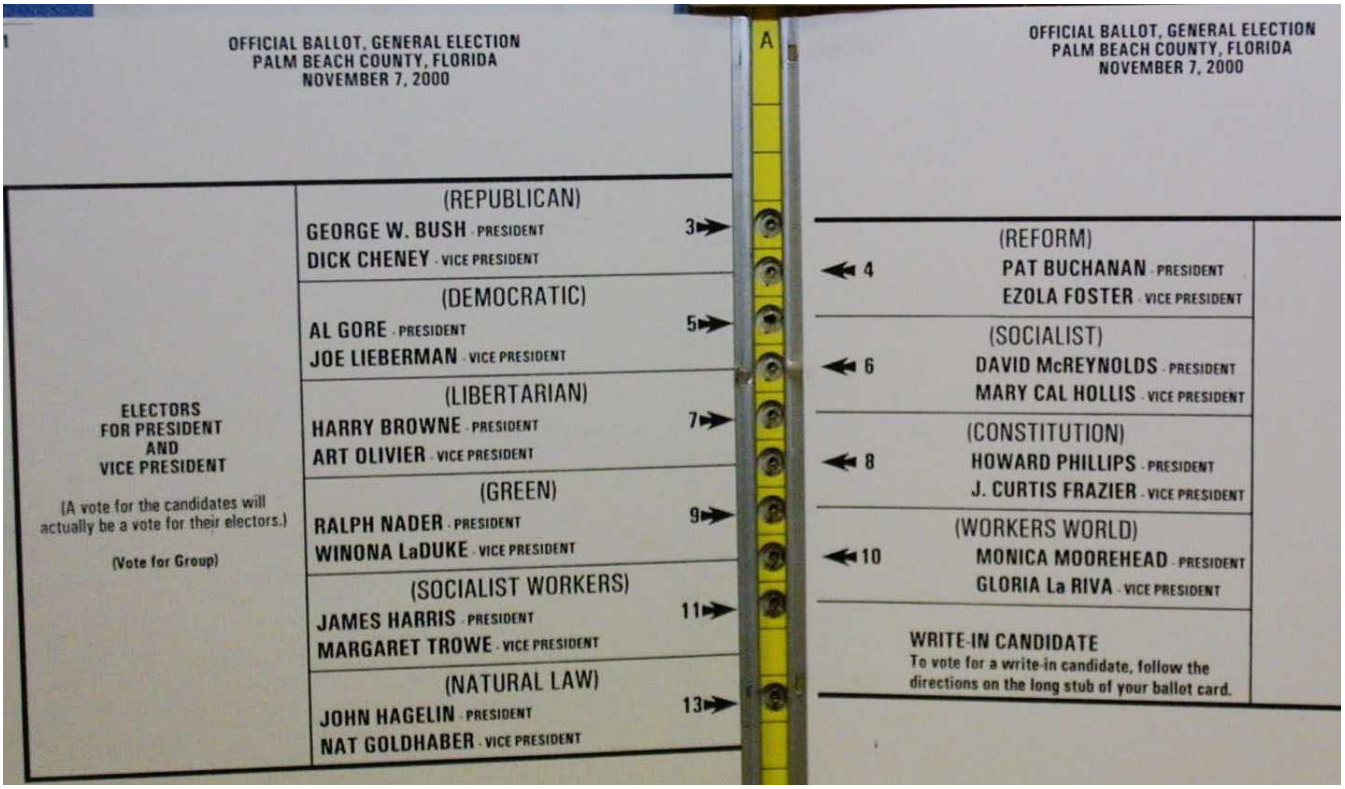



**Figure 5.4.3 The Palm Beach County Butterfly Ballot**

Of the 1,131,184 persons living in Palm beach County in 2000, over 260,000 were 65 years old or older. It is presumed that many of the elderly were either confused by the ballot or had eyesight issues. On the left side of the ballot, Gore was the second presidential candidate listed from the top. The presumption is that some three thousand voters punched the second hole from the top instead of the third, inadvertantly voting for Buchanan. As Pat Buchanan himself said in an interview, "What cost Al Gore Florida in 2000, and the presidency, was the 'butterfly ballot.'"

## High leverage and Influential Points

> Give me a lever and a place to stand and I will move the earth.
> —Archimedes (in one wording of the thought)

A point with an abscissa ($x$-coordinate) that is much greater or smaller than $\bar{x}$ is said to be a ***high leverage*** point. Think of a lever extending from the high leverage point to a fulcrum at $(\bar{x}, \bar{y})$. A high leverage point has the potential to greatly influence the parameters of a regression line. It is therefore advisable to investigate high leverage points.

An ***influential*** point is one that has a demonstrably larger impact on the slope and/or $y$-intercept of the regression line than have most of the other points. A regression outlier might be an influential point, as might be a high leverage point. That said, a high leverage point that is close to the regression line might not be influential. The trouble is, the visual evidence of a scatter point with the regression line superimposed can be misleading: a high leverage point might be so influential that it drags the regression line down to its vicinity (and therefore does not appear to be far away from the regression line). The surest way to tell if a high leverage point is influential is to calculate the regression line that would have occurred without the point and compare it to the one calculated with the point.

By way of an example, consider the cereal data that was presented earlier in this chapter. For the $x$-variable (sugar content in grams), the mean $\bar{x}$ is 6.29. The $x$-value for Malt-O-Meal Golden Puffs is 14.5, making the point $(14.5, 103.7)$ a high leverage point. This is the point colored red at the far right of the scatter plot shown in Figure 5.4.4.

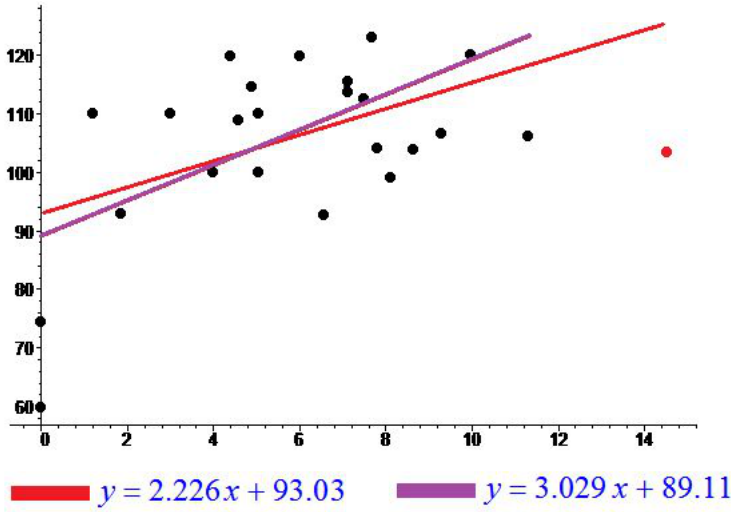



**Figure 5.4.4 High Leverage Point—Malt-O-Meal Golden Puffs**

The regression line based on all 26 cereals, colored red in Figure 5.4.4, is $y = 2.226\,x + 93.03$. The regression line that is colored purple in Figure 5.4.4 is based on the 25 cereals that remain when Golden Puffs is removed from the calculation. Its equation is $y = 3.029\,x + 89.11$. A little arithmetic reveals that the single point representing Golden Puffs changes the slope of the regression point by 36%. That's influence!

A regression outlier also has the potential to be influential. Consider the (blue) point representing Palm Beach County in Figure 5.4.2. It is a regression outlier of the most spectacular kind, but it is not a high leverage point. With the inclusion of the Palm beach County vote, the regression line has equation $y = 0.00471\,x + 48$. Without the point, the regression line has equation $y = 0.00327\,x + 68$. A little arithmetic shows that the single point representing Palm Beach County changes the slope of the regression line by 44% and the $y$-intercept by 29%. That's influence!

## Causality

In a 1926 address and accompanying publication, George Udny Yule (What a GUY!) called attention to two sequences of data collected every year from 1866 to 1911—X: Standard Mortality per 1000 population in England and Wales, and Y: Church of England Marriages per 1000 marriages. In the scatterplot below, we have scaled the latter to obtain percentages.

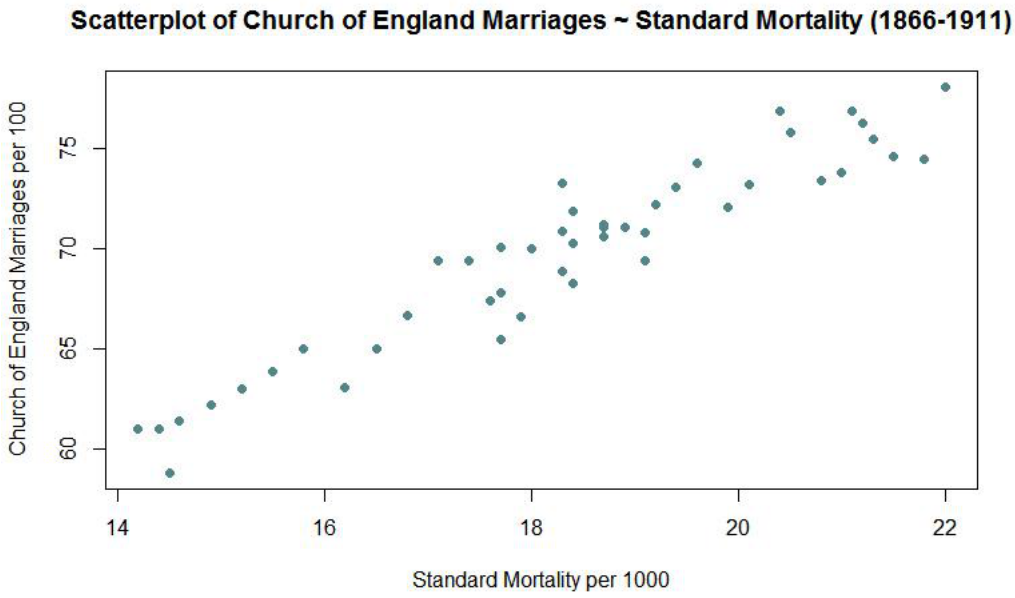



**Figure 5.4.5 Spurious Correlation**

To any eye it is very obvious that there is a strong linear correlation. In fact, Pearson's correlation coefficient is 0.9515. Do Church of England marriages cause mortality? Clearly such an idea is farfetched. Here is what Yule had to say:

> Now I suppose it is possible, given a little ingenuity and goodwill, to rationalize very nearly anything. And I can imagine some enthusiast arguing that the fall in the proportion of Church of England marriages is simply due to the Spread of Scientific Thinking since 1866, and the fall in mortality is also clearly to be ascribed to the Progress of Science: hence both variables are largely or mainly influenced by a common factor and consequently ought to be highly correlated. But most people would, I think, agree with me that the correlation is simply sheer nonsense. ...It is absurd to suppose that the two variables in question are in any sort of way, however indirect, causally related to one another.

The correlation Yule drew attention to was utterly spurious. Even when correlation is not spurious, there may be no causation. University admissions officers may use standardized test scores applicants have taken as a predictor of university success as measured by Freshperson GPA, but there is clearly not even an iota of causation. Whatever your SAT or ACT score was, it will not in any way cause your next statistics exam score. For that matter, neither will your last statistics exam score.

In the case of Yule's mortality and marriage correlation, the statistician considered a possible factor, the progression of the variable Year, and dismissed it. In such a situation we refer to the variable *Year* as a "confounding variable". We will discuss *confounding variables* in greater detail later in this section. In the case of standardized test scores and freshperson GPA, the correlation, which is moderate, may be attributed in part to confounding factors of student ability and knowledge.

- *Correlation may be spurious.*
- *Correlation may be the result of a confounding or lurking variable.*
- *Correlation, even if it very strong, does not imply a causal relationship.*[5]

If X is an explanatory variable and Y is a response variable, then we call a third variable V a ***confounding variable*** if there is a correlation between V and X and between V and Y. To what degree is any association between Y and X actually attributable to the common influence of V?

Here are some examples of confounding variables. Suppose that Y is the likelihood of a driver being a burglar looking for a burglary opportunity. Suppose that X is the value of the vehicle being driven. It is the author's impression that, in the last several years of life of a junker he drove from 1982 to 2004, police officers, lying in wait to nab speeders, would, upon spotting the author's car, tail his car as he drove it through the affluent Clayton community. The author presumes that police officers believe there is a strong negative correlation between the value of a car and the likelihood of the driver's intent to burglarize. Perhaps this belief is justified. But clearly Poverty is a confounding variable. Millionaires may commit many crimes, but burglary would be low on the list. Poverty obviously influences both the value of the car a person drives and his or her motivation to risk burglary.

Down Syndrome is a genetic abnormality that occurs by chance. Let Y be the likelihood of a Down Syndrome birth. Let X be the order of the birth: 1 = the first child to which the mother gave birth, 2 = the second child to which the mother gave birth, and so on. Evidence suggests that Y $\sim$ X. But Maternal Age is clearly a confounding variable: Y $\sim$ Maternal Age and X $\sim$ Maternal Age.

Let Y be the incidence of coronary heart disease. Let X be Weight. Both X and Y are associated with Age. Both X and Y are associated with aspects of Diet, such as consumption of saturated fat. There are several confounding variables.

Let Y be the number of jellyfish stings in the water off a beach. Let X be the weight of litter picked up off the beach. There is a spurious correlation Y $\sim$ X. The association is not due to jellyfish become enraged by human litterers. A confounding variable is the number of visitors to the beach.

[5]Where do babies come from? There is no need for the author to answer this question. Everybody knows: Storks deliver them. In ancient times, particularly in northern Europe, the days surrounding the summer solstice in June were the occasion for pagan fertility rituals. Nine months later, in March, there would be an observable increase in births. At the same time, storks, flying north on their migratory route, were seen in increasing numbers. The association of birds and births was noted. Clearly, birds had to be the explanation, and births the response. Human knowledge can advance from even the most off-hand observations.

Let Y be the incidence of dementia. It is known that Y $\sim$ X, where X is the amount of trauma to the brain. A confounding variable is Age.

A new anesthetic called *halothane* became available in American hospitals in 1958. Within a few years, there was a suspicion that halothane was asoociated with a greater number of postoperative deaths than other anesthetic agents. A large statistical study was undertaken to ascertain the safety of halothane. A database of 850,000 surgeries between 1962 and 1964 was examined. The number of deaths among the patients was 17,000 during a six week postoperative period: in other words, the death rate was about 2%

Four specific anesthetic agents, Halothane, Pentothal, Cyclopropane, and Ether, were singled out, and the was a fifth group of all other anesthetics used.

| Halothane | Pentothal | Cyclopropane | Ether | Others |
|---|---|---|---|---|
| 1.7% | 1.7% | 3.4% | 1.9% | 3.0% |

**Table 5.4.1: Death Rates Associated with Various Anesthetics**

A first glance at the data suggests that fears of the danger of halothane were misplaced. However, there were confounding variables. For example, surgeons were well-aware that Cyclopropane was used much more frequently than other agents in risky, difficult operations. Type of Operation was a confounding variable. Statisticians also made adjustments for other confounding variables: Age, Gender, and Physical Status Prior to Operation. Here is the adjusted table:

| Halothane | Pentothal | Cyclopropane | Ether | Others |
|---|---|---|---|---|
| 2.1% | 2.0% | 2.6% | 2.0% | 2.5% |

**Table 5.4.2: Adjusted Death Rates Associated with Various Anesthetics**

The study showed that Halothane was about as safe as any other anesthetic and safer than Cyclopropane. As it happened, both Halothane and Cyclopropane were replaced with better anesthetics within 20 years of the study. Our interest is the effect that the confounding variables had in generating the initial, raw data.

There is another sort of variable to which frequent reference is made. If we are investigating Y $\sim$ X, then we say that V is a ***lurking variable*** if Y $\sim$ V and X $\sim$ V and we do not consider V at all. The difference between confounding and lurking variables is that are aware of a confounding variable and we try to control for it in the design of our study or explain its role after the study. A lurking variable is a confounding variable could-have-been had we not remained blissfully unaware of its potential influence.

## Extrapolation

Let us return to Florida! Not Palm Beach, but Miami. Why Miami? Put aside the heat. Put aside the humidity. Put aside the insects. Miami is an engine of opportunity. Miami is the point of departure on the road to riches. Table 5.4.3 shows the median Miami home price in July for each of the first seven years of the twenty-first century.

| Year | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|---|---|
| Price | 207757 | 229282 | 257648 | 290270 | 339682 | 433140 | 469748 |

**Table 5.4.3: Median Miami Home Price**

The linear association is very strong: $r = 0.975$. The equation of the regression line is $y = 45562\,x - 90941538$. Figure 5.3.3 in the preceding section displays the fit of the linear model to the median Miami house price data. The slope, 45562, of the regression line means that the median home price increases by \$45,562 each year. That comes to a \$227,810 profit if we buy a median price house in 2006 and sell it five years later in 2011. The linear model predicts that the house will be worth $\$\big((45562)(2011) - 90941538\big)$, or \$ 683644.

But wait! July 2011 has come and gone. We don't have to rely on a regression line prediction for the sale price. We have historical records that tell us what the actual median home price was in Miami in July 2011. Let us project our magic carpet to prosperity. The following R code has been used to generate Figure 5.4.6, which extends our previous linear model through July 2014. The regression line has only been plotted through 2011 for obvious reasons. (The viewing window is $[2000, 2014] \times [180000, 683644]$.)

```
> Year = 2000:2014
> Median.House.Price = c(207757, 229282, 257648, 290270, 339682, 433140,
+ 469748, 422918, 287600, 238229, 232877, 214285, 225280, 250861, 251000)
```

```
> f = function(x){45562*x-90941538}
> plot.title = "Miami Median House Price 2000-2014: Linear Model"
> plot(Year, Median.House.Price, xlim = c(2000,2014), ylim = c(180000,683644),
+ main = plot.title, pch = 19, col = "salmon2")
> lines( Year[1:12], f(Year[1:12]), xlim = c(2000,2014), ylim = c(180000,683644),
+ type = "l", lwd = 2, col = "salmon4")
```

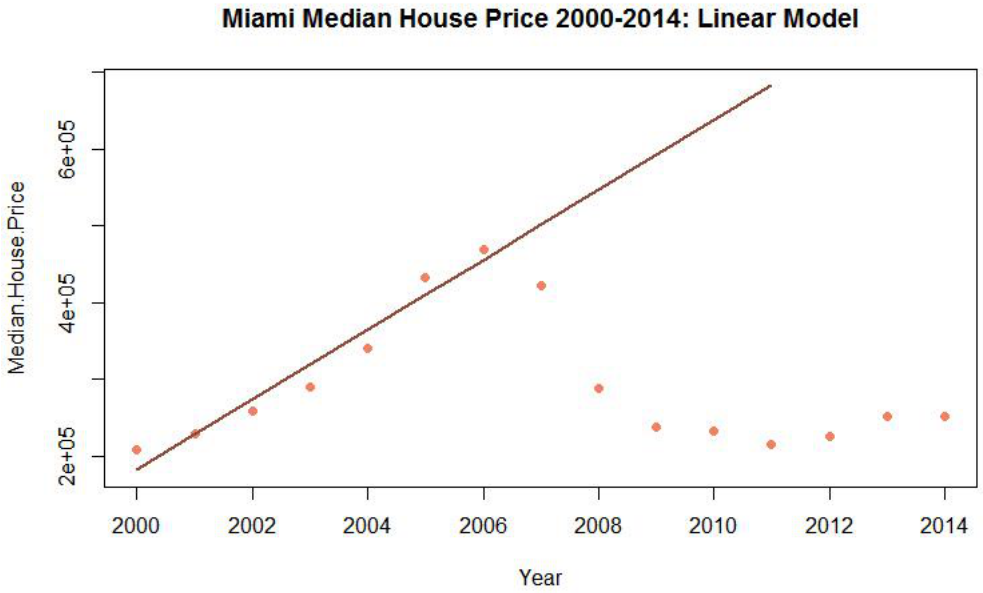



**Figure 5.4.6 Scatter Plot of Year (Predictor) and Median Miami Home Price (Response), 2000-2014**

OMG!!! What happened to our magic carpet ride? Who pulled the rug out from under our feet? Who bent our sequence of linear observations?

The use of a regression line to predict a value of the response variable corresponding to a predictor value that is between the values of two observations is called ***interpolation***. The use of a regression line to predict a value of the response variable corresponding to a predictor value that is outside the range of observations is called ***extrapolation***. "Extrapolation" is the technical buzzword for predicting the future based on the past. In our adventure in property speculation, the years 2000–2006 were the range of the predictor variable that was the basis of the linear model. The value of the predictor variable used for our prediction was 2011—well outside the linear model basis range. In fancy jargon, we were extrapolating. Two take-away morals of our real estate learning experience are:

- *The farther the predictor variable $x$ is from the sample mean $\bar{x}$, the more dubious is the prediction $\widehat{y}$ based on $x$.*
- *Extrapolation based on a regression line is risky.*

# Chapter 5 Exercises

1. For the following data

| Explanatory X | 1 | 2 | 6 |
|---|---|---|---|
| Response Y | 2 | 8 | 20 |

(i) Find $\bar{x}, \bar{y}, s_X$, and $s_Y$. (ii) Find Pearson's $r$. (iii) Find the regression line. (iv) Find the absolute values of the residuals. (v) Find SST, SSR, and SSE. (vi) Find the coefficient of determination.

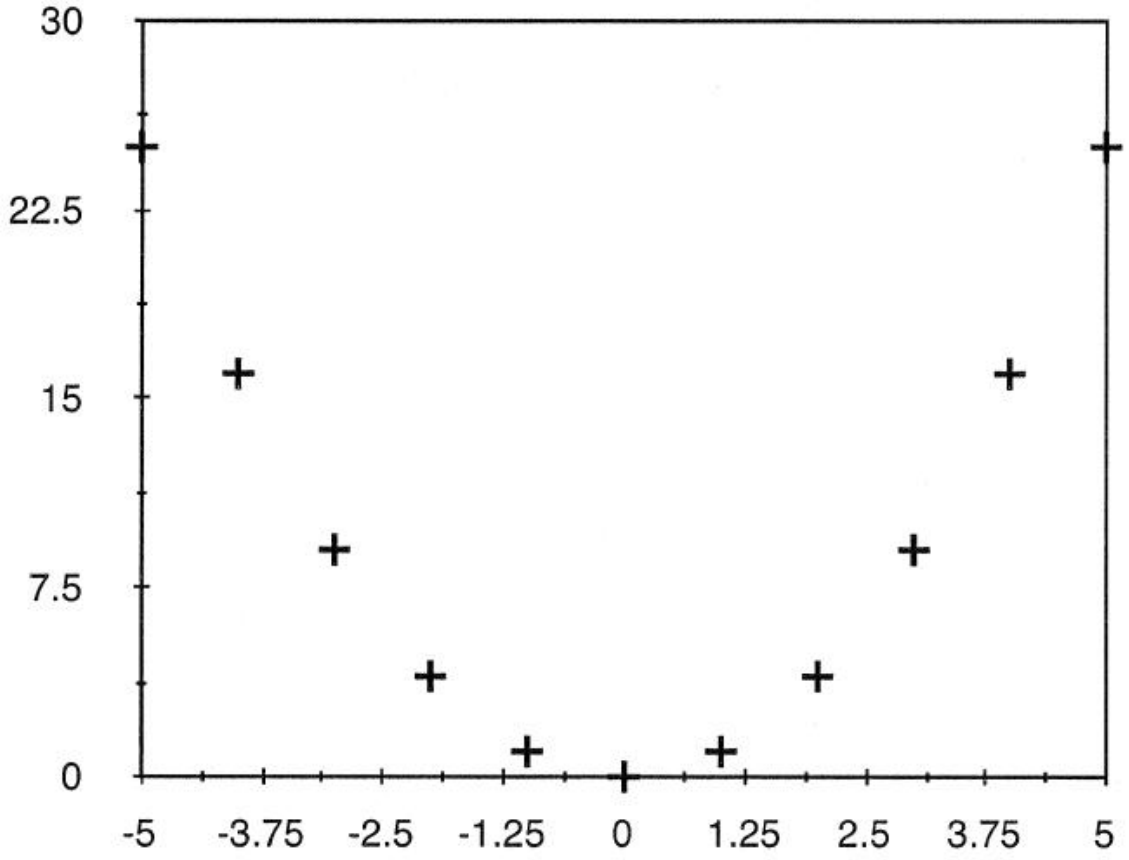
2. (Washington University exam, Fall 2010) For a given bivariate data set (X,Y): $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, the regression line for the z-scores passes through the point (1,0.42). What fraction of the variation in Y is accounted for by the linear model?

A) 0.176 B) 0.352 C) 0.420 D) 0.580 E) 0.828
F) 0.840 G) 0.880 H) 0.940 I) 1.000 J) Other

3. (Washington University exam, Fall 2010. Warning: This problem is faulty due to an inconsistent choice of slopes.) For a given bivariate data set (X,Y): $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, the equation for predicting a $y$-value corresponding to an $x$-value is $y = 2x + 1$, and the equation for predicting an $x$-value corresponding to a $y$-value is $x = 0.6y - 0.4$. What is the coefficient of correlation between $X$ and $Y$.

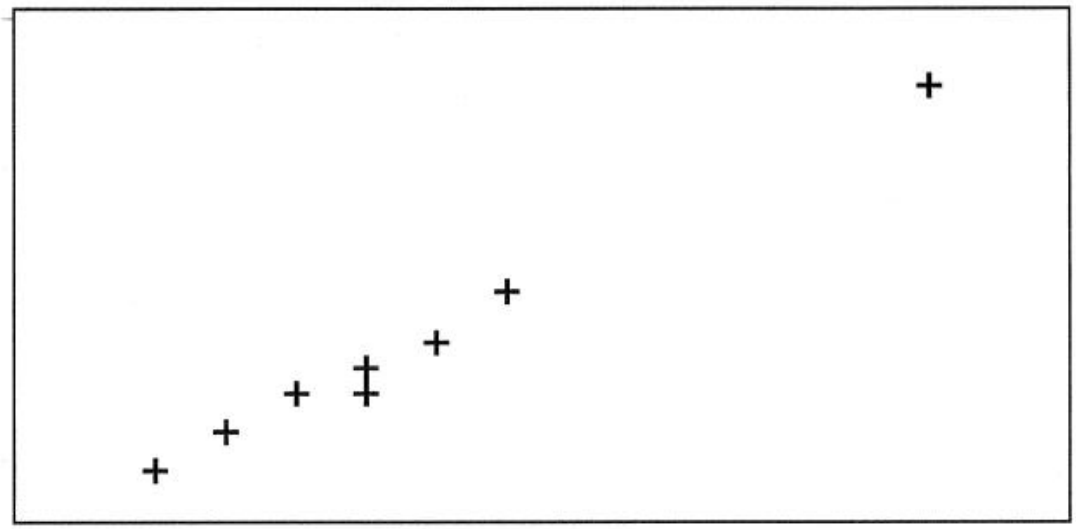
A) 2 B) 0.6 C) 1.2 D) -0.4 E) -0.8
F) 1.063 G) -1.063 H) 1.095 I) -1.095 J) Other

4. (Washington University exam, Fall 2010) Which re-expression would be appropriate to get a linear scatter to the following plot:




A) Re-express Y by square of Y
B) Re-express Y by square-root of Y
C) Re-express Y by log(Y)
D) Re-express Y by $10^{Y}$
E) Re-express Y by 1/Y
F) Re-express Y by 1/(square-root of Y)
G) None of the above

5. (Washington university exam, Fall 2010) Describe the outlier in the scatter plot given below.



A) Influential with high leverage and high residual
B) Influential with high leverage and low residual
C) Influential with low leverage and high residual
D) Influential with low leverage and low residual
E) Not influential with high leverage and high residual
F) Not influential with high leverage and low residual
G) Not influential with low leverage and high residual
H) Not influential with low leverage and low residual

6. In the following early epidemiological study of the effects of smoking on lung health, cigarette consumption in 1930 and lung cancer mortality in 1950 were tabulated. (Source: Freedman, Pisani, and Purves, Op. cit.)

| **Country** | **Cigarettes per Capita** | **Deaths per Million** |
|---|---|---|
| Australia | 480 | 180 |
| Canada | 500 | 150 |
| Denmark | 380 | 170 |
| Finland | 1100 | 350 |
| Great Britain | 1100 | 460 |
| Iceland | 230 | 60 |
| Netherlands | 490 | 240 |
| Norway | 250 | 90 |
| Sweden | 300 | 110 |
| Switzerland | 510 | 250 |
| United States | 1300 | 200 |

**Eleven Country Lung Cancer Mortality, 1950**

(i) Calculate the regression line.
(ii) There are three high leverage points at the right. One at a time, leave out each in the calculation of the regression line. To what degree is each influential?
(iii) Two of the high leverage points are regression line outliers. Which ones? What are the residuals?

7. Maternal age is the most common risk factor for Down syndrome: as a woman gets older, her risk that a pregnancy will involve a chromosome abnormality increases. (There are three types of Down syndrome: nondisjunction aka trisomy 21, translocation, and mosaicism, accounting for 95%, 4%, and 1% of all cases, respectively. Maternal age is not believed to be a risk factor for translocation, which is the only heritable form of Down syndrome. There are no known behavioural or environmental factors that cause any of the three types of Down syndrome.) The following table[6], divided into three parts for typesetting reasons, relates risk (the response variable) to maternal age (the explanatory variable).

| **Maternal Age** ( $x$ ) | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Down Synd. Risk** ( $y$ ) | $\frac{1}{1667}$ | $\frac{1}{1667}$ | $\frac{1}{1429}$ | $\frac{1}{1429}$ | $\frac{1}{1250}$ | $\frac{1}{1250}$ | $\frac{1}{1176}$ | $\frac{1}{1111}$ | $\frac{1}{1053}$ | $\frac{1}{1000}$ |

**Down Syndrome Risk by Maternal Age, 20–29 (yr)**

| **Maternal Age** ( $x$ ) | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Down Syndrome Risk** ( $y$ ) | $\frac{1}{952}$ | $\frac{1}{909}$ | $\frac{1}{769}$ | $\frac{1}{625}$ | $\frac{1}{500}$ | $\frac{1}{385}$ | $\frac{1}{294}$ | $\frac{1}{227}$ | $\frac{1}{175}$ | $\frac{1}{137}$ |

**Down Syndrome Risk by Maternal Age, 30–39 (yr)**

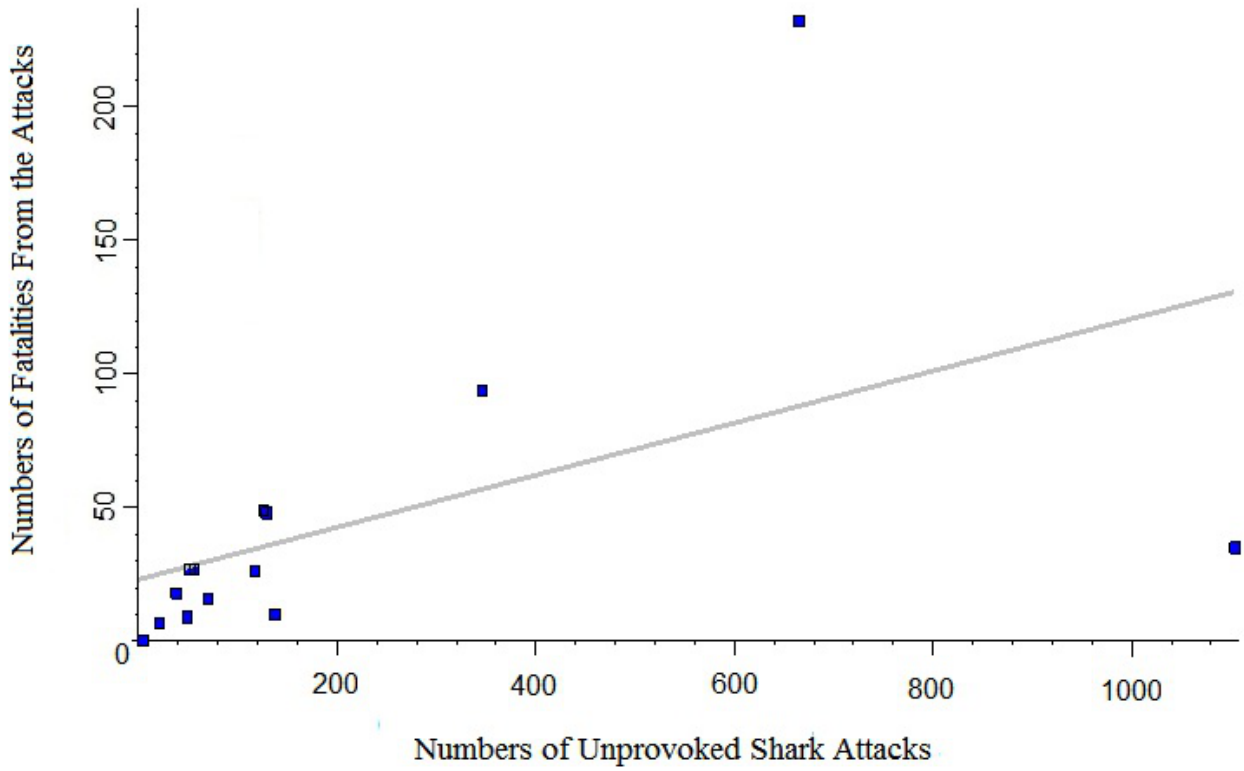
| **Maternal Age** ( $x$ ) | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Down Syndrome Risk** ( $y$ ) | 1/106 | 1/82 | 1/64 | 1/50 | 1/38 | 1/30 | 1/23 | 1/18 | 1/14 | 1/11 |

**Down Syndrome Risk by Maternal Age, 40–49 (yr)**

[6]Source: Hook EB, Cross PK, Schreinemachers DM, Chromosomal abnormality rates at amniocentesis and in live-born infants, JAMA **249(15)** (1983), 2034-38. `http://www.ncbi.nlm.nih.gov/pubmed/6220164` The *National Down Syndrome Society* and the *Mayo Clinic* give slightly different values, but their figures appear to have been rounded and no source for them is cited.

a) Calculate Pearson $r$ for the given 30 data points.
b) The linear association appears to be fairly strong, but is it? Do now what should have been done first (i.e., before the calculation of $r$).
c) Calculate $\ln(y)$ for each value $y$ of risk. Calculate Pearson $r$ for $x$ and the transformed data $\ln(y)$.
d) The linear association between $x$ and $\ln(y)$ appears to be *very* strong, but is it? Do now what should have been done before the latest correlation calculation.
e) Calculate Pearson $r$ for $x$ and the transformed data $\ln(y)$ for $y \geq 32$.
f) The linear association between $x$ and $\ln(y)$ for $y \geq 32$ appears to be *extraordinarily* strong, Confirm this judgment by plotting the relevent observations and superimposing the regression line in the same window.
g) Inverse transform the regression line obtained in part (f) and plot it for $20 \leq x \leq 49$ in a viewing window containing a scatterplot of all the original data.
h) A woman who is unwilling to accept a risk greater than 1/250 should not give birth beyond what age?

8. The figure below shows a scatter plot of the numbers of unprovoked shark attacks ($x$) and resulting fatalities ($y$) worldwide in the years 1958–2014. There are 14 points (although two pairs of points partially overlap, so, depending on resolution, there may appear to be only 12 plotted points). The regression line based on all observations has been superimposed.




Identify the explanatory variable outliers, the response variable outliers, the regression line outliers, the high leverage points, and the influential points. A word to the wise: You're gonna need a bigger boat.