

What Do r and r^2 Really Measure?

Before looking at what r and r^2 do, let's consider what they do not do. They cannot be used to determine if a data set is linear. For example, the data in Table 1 describe the relationship between the depth of water in an urn and time as the urn is being emptied. The first question we ask is, "Is the relationship between the variables linear?"

Time (secs)	0	30	60	100	140	180	210	265	330	390
Depth (in)	12.5	11.25	10.5	9.25	8.0	7.0	6.0	4.75	3.5	2.5

Table 1: Time and Depth of lemonade in an urn

At some time in their lives, most students have gotten a drink from an urn with a stop-cock drain at the bottom. They recognize that a cup fills quickly if the urn is nearly full and very slowly if the urn is nearly empty. If the relationship is linear, then the change in the depth of the liquid in the urn over any fixed interval of time will be the same (to within measurement accuracy).

If students can judge whether the urn is nearly full or nearly empty by how quickly the cup fills, then the relationship cannot be linear. If the change in the dependent variable allows students to predict the value of the independent variable, then the relationship cannot be linear. As one student told me, "Being linear means that you don't know where you are." This is an important way to think about linear functions.

Data analysis offers support to these experiential arguments (see Figure 1). If we fit a line to the data, the curved shape of the data becomes apparent as the characteristic U-shaped residual plot indicates curvature in the data.

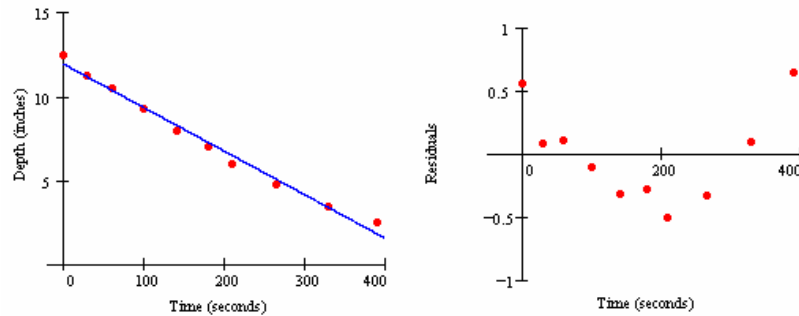


Figure 1: Linear Model of Depth and Time with Residuals

If we compute the correlation between Time and Depth for this data, we find that $r = -0.9935$ and $r^2 = 0.9871$. Notice that r is only 0.0065 from a perfect negative correlation, yet we know from our own personal experience that this phenomena is not linear, and sufficiently non-linear for us to detect from the simple experience of filling cup from an urn. The moral here is that we **cannot** use either r or r^2 as a means to determine if a set of data is best modeled by a line.

Bivariate Fit of child ht By parent ht

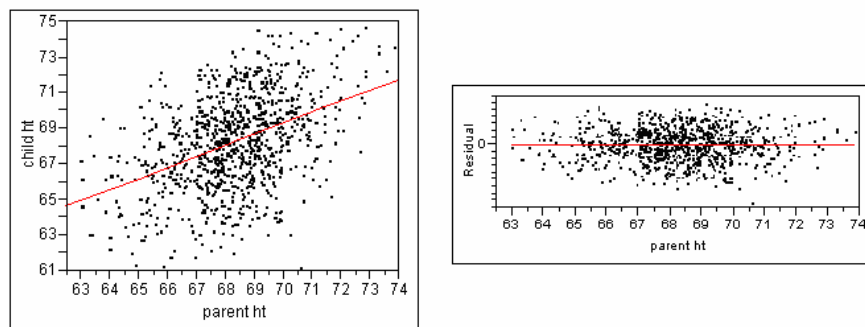


Figure 2: Galton's Father/Son data

To emphasize this point, consider the data shown in Figure 2. This is the famous data gathered by Fances Galton on 952 father/son combinations. Galton measured the height of the father (in inches) and the height of the first-born son (in inches). The relationship between these two variables is best described by a linear function, although the correlation between the two variables is $r = 0.4209$ and the coefficient of determination is $r^2 = 0.1772$.

So, what good is a “measure of linearity” that varies from 0 to 1, with the closer to 1 being “more linear” if a correlation of 0.42 is linear but 0.99 is not? In what way do r and r^2 “measure linearity”?

The Coefficient of Determination (r^2)

To understand what r^2 measures and how to interpret its value, we need to consider how it is computed. For our example problem, we will use some data collected in class. The data in Table 2 represent the heights from which a hard rubber ball was dropped (x in inches) and the height of its first bounce (y in inches). The ball was dropped times from each of five heights. A scatterplot of the 15 measurements is shown in Figure 3.

Height (x)	20	30	40	50	60
Bounce 1 (y)	16.5	24	33	39.5	47.5
Bounce 2 (y)	15.75	24.25	32.75	38.5	47.5
Bounce 3 (y)	16.75	25.5	33	39.75	49.25

Table 2: Ball Drop Data (height of drop, height of 1st bounce)

The mean height of the drop is $\bar{x} = 40$ and the mean height of the first bounce is $\bar{y} = 32.233$. The standard deviation of the heights of the bounces is $s = 11.469$, so the variance of the heights of the bounces is $s^2 = 131.54$. The regression line for the data is $\hat{y} = 0.96667 + 0.78167x$ ($Bounce = 0.967 + 0.782Drop$), and the standard deviation of the residuals is 0.78148. The coefficient of determination is $r^2 = 0.99536$. What is it about this data and this line that is 0.99536?

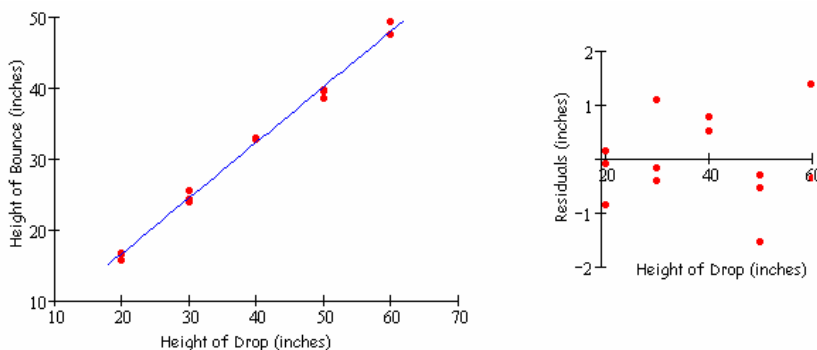


Figure 3: Ball Drop Data with Residual Plot

From Figure 3 we observe that the y -values in the data set are not all the same. They vary. Why? If y depends on x (Bounce Height depends on Height of Drop), and we change x (Height of Drop), then naturally y (Bounce Height) will change. How much of the variation that we see in y can we attribute to changing x and how much from elsewhere? Well, we say that the variation we see in the fitted y -values comes from the linear equation and x , and the rest from elsewhere.

So, we compare the variation in the fitted values (by computing the variance of the fitted y -values) to the variation in the original values (by computing the variance of y). The ratio is the proportion or fraction of the variation in y that is attributed to the linear relationship with x .

The fitted values (values on the line) are shown in Table 2.5 below.

Height (x)	20	30	40	50	60
Bounce 1 (y)	16.6	24.417	32.233	40.05	47.867
Bounce 2 (y)	16.6	24.417	32.233	40.05	47.867
Bounce 3 (y)	16.6	24.417	32.233	40.05	47.867

Table 2.5: Fitted Values for Ball Drop

The variance of the y -values

{16.5, 24, 33, 39.5, 47.5, 15.75, 24.25, 32.75, 38.5, 47.5, 16.75, 25.5, 33, 39.75, 49.25} is $s_y^2 = 131.54$. The variance of the fitted y -values {16.6, 24.417, 32.233, 40.05, 47.867, 16.6, 24.417, 32.233, 40.05, 47.867, 16.6, 24.417, 32.233, 40.05, 47.867} is $s_{\hat{y}} = 11.442477^2 = 130.9303$. Notice that the variance of the original y 's is larger than the variance of the fitted \hat{y} 's. Since $r^2 = 0.99536$, we know that the ratio of the variance of the fitted \hat{y} -values to the variance of the original y -values is $r^2 = \frac{130.93}{131.54} = 0.99536$. Also, notice that the variance of the residuals is $(0.78148)^2 = 0.61071$ and $131.54 = 130.93 + 0.61$, so $s_y^2 = s_{\hat{y}}^2 + s_r^2$, the variance of the y -values is the sum of the variance of the fitted values (\hat{y} 's) and the residuals.

Interpreting r^2

Let's think about the following situation. I'm going to drop this ball one more time. You must guess how high the first bounce will be. What is your guess? If you have been paying any attention at all to what we have been doing, you will want to know the height from which I will drop the ball.

If I don't tell you the height from which I will drop the ball, what is your best guess? How is that different than your guess if I do tell you the height from which I will drop the ball? If you don't know the height from which I will drop the ball, your best guess will be the average of all bounces. If you do know the height from which I will drop the ball, your best guess will be the value of the regression line evaluated at that height.

In a nutshell:

- If you don't know x , your best guess is \bar{y} .
- If you do know x , your best guess is \hat{y}

How much better is \hat{y} as a guess than is \bar{y} ? Another way to phrase this question is, how more do you know about the value of y (the actual result of the drop) if you know x , than if you don't know x ? These are the questions r^2 is attempting to answer.

To see how r^2 answers these questions, we focus on a single data point. Figure 4 illustrates the situation. We represent the location of the data point with y , the location of the position on the regression line with \hat{y} , and the mean value with \bar{y} . For every point y , we can write

$y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y})$. Algebraically, we see that this is an identity. Recall that if we don't know x , we guess \bar{y} and if we do know x we guess \hat{y} . So, $(\hat{y} - \bar{y})$ is, in some sense, the amount of additional information (beyond just \bar{y}) about y that we got by knowing x . We know this much more about the value of y . We don't know everything, since $\hat{y} \neq y$. But $(\hat{y} - \bar{y})$ of the difference between y and \bar{y} can be attributed to the linear relationship with x . This leaves $(y - \hat{y})$ unexplained (notice, this is the residual for this point).

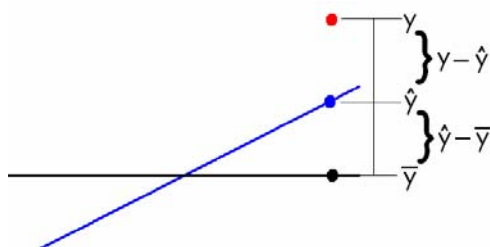


Figure 3: Partitioning $y - \bar{y}$ for a Single Point

Now, divide both sides of $y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y})$ by $y - \bar{y}$. This gives $1 = \frac{(y - \hat{y})}{(y - \bar{y})} + \frac{(\hat{y} - \bar{y})}{(y - \bar{y})}$.

Now, we can interpret $\frac{(\hat{y} - \bar{y})}{(y - \bar{y})}$ as the fraction or proportion of $y - \bar{y}$ attributed to the linear

relationship with x (and knowing x). Also, $\frac{(y - \hat{y})}{(y - \bar{y})}$ is the fraction or proportion of $y - \bar{y}$ unexplained by the linear relationship with x .

So, for a single point, we have this very nice interpretation of $\frac{(\hat{y} - \bar{y})}{(y - \bar{y})}$ as the fraction or

proportion of $y - \bar{y}$ attributed to the linear relationship with x . The value $\frac{(\hat{y} - \bar{y})}{(y - \bar{y})}$ is a measure of

how much information x gives us about y , or how much better \hat{y} is than \bar{y} as a guess for y . Recall that if we don't know x , we guess \bar{y} and if we do know x we guess \hat{y} . If $\hat{y} = y$, from

$1 = \frac{(y - \hat{y})}{(y - \bar{y})} + \frac{(\hat{y} - \bar{y})}{(y - \bar{y})}$ we see that $\frac{(\hat{y} - \bar{y})}{(y - \bar{y})} = 1$, and if $\hat{y} = \bar{y}$, then from $\frac{(\hat{y} - \bar{y})}{(y - \bar{y})} = 0$.

Moving From a Point to the Data Set as a Whole

We want to apply this interpretation at a point to the data set as a whole. We only need to change a few things. Instead of considering $y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y})$, we square both sides and sum over all y -values. So, the expression we have is $\sum_i (y_i - \bar{y})^2 = \sum_i [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2$. Now, squaring the right sides gives

$$\sum_i [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2 \sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}), \text{ so}$$

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

This equation has almost the same structure as $y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y})$. Unfortunately, it has the messy last term. However, this term is always zero. One attribute of least squares regression is that $2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$. It takes quite a lot to prove this, and we will not do it here. (The short

reason is that vector $(y_i - \hat{y}_i)$ is perpendicular to vector $(\hat{y}_i - \bar{y})$ and so the dot product

$$\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0. \text{ See } A \text{ Vector Approach to Regression, Pages 9-12}.$$

But, now we have $\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$. If we divide both sides by $(n-1)$,

we have $\frac{\sum_i (y_i - \bar{y})^2}{n-1} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-1} + \frac{\sum_i (\hat{y}_i - \bar{y})^2}{n-1}$. The left side of this equation, $\frac{\sum_i (y_i - \bar{y})^2}{n-1}$, is the

definition of the variance of y , while $\frac{\sum_i (y_i - \hat{y}_i)^2}{n-1}$ is the variance of the residuals (since the mean of

the residuals is zero). Following the model for a single point above, the term $\frac{\sum_i (\hat{y}_i - \bar{y})^2}{n-1}$ is the *amount of the variance of y we can attribute to the linear relationship with x .*

If we divide by $\frac{\sum_i (y_i - \bar{y})^2}{n-1}$, we have $1 = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} + \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$. So now, $\frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$

is the fraction or *proportion of the variance of y that is attributable to the linear relationship with x .*

This is the coefficient of determination. So, we know that $r^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$.

As before, recall that if we don't know x , we guess \bar{y} and if we do know x , we guess \hat{y} . If

$\hat{y}_i = y_i$ for all i , from $1 = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} + \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$ we see that $r^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1$, and if $\hat{y}_i = \bar{y}_i$

(knowing x tells us nothing about y), then $r^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 0$.

There are many other ways to compute r^2 . This is the way to do it that gives the book interpretation as the proportion of the variation (it is actually *variance*) in y that can be attributed or

explained by the linear relationship with x . Notice also that $r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$. Recall that

$\frac{\sum_i (y_i - \hat{y}_i)^2}{n-1}$ is the variance of the residuals (since the mean residual is zero) and $\frac{\sum_i (y_i - \bar{y})^2}{n-1}$ is the

variance of y . So $r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$ says that r^2 is one minus the ratio of the variance of the

residuals to the variance of y . r^2 is the portion of the variance of y that is *not* in the residuals.

We can see this clearly in the Ball Bounce data. The variance of the heights of the bounces is $s_y^2 = 131.54$. The variance of the residuals is $s_R^2 = 0.61071$. The coefficient of determination is $r^2 = 0.99536$. We asked earlier, “what is it about this data and this line that is 0.99536?” We now

know the answer. $1 - \frac{s_R^2}{s_y^2} = 1 - \frac{0.61071}{131.54} = 0.99536$.

So, how do we interpret r^2 ? In a scatterplot, not all the y -values are the same. Why do the y -values vary? Well, if y is linearly related to x , and we change x , then y also changes. Part of the variation we see in y can be explained by the fact that y is related to x and we changed x .

If we look at the ordered pairs (1,2), (3,6), (4,8), and (5, 10), we see that $y = 2x$. There is some variation in the y -values {2, 6, 8, 10}. We can measure this variation and say that the variance of y is 11.666. If we didn't change the x 's, what would have been the variance of y ? If $y = 2x$ and there is no variation in the x 's, there will be no variation in the y 's, so all of the variation in y is attributed to the variation in x . Thus, $r^2 = 1$.

Now, in the real world we don't really expect that all of the y -values will fall exactly on our regression line. As we have seen, the r^2 value is comparing the variation of the y -values of the data and the variation in the \hat{y} -values on the regression line. The variation of the \hat{y} -values can be attributed to the linear relationship between x and y and the variation in x . The variance of the y -values is larger than the variance in the \hat{y} -values. The extra variation comes from the residuals. The value of r^2 is the ratio of the variance of the fitted \hat{y} -values to the variance of the y -values. The variance of the \hat{y} -values is attributed to the linear relationship between x and y . We changed the x -values so the y -values had to also change.

What does all this have to do with linearity? Unfortunately, very little. The coefficient of determination only has this interpretation if the data is modeled by a line. It can't help us determine if a line is appropriate. The value of r^2 gives us information about how close to a line the data fall, by comparing how the variance of y compares to the variance of \hat{y} , but that's all it does.

What About the Correlation Coefficient, r ?

Clearly, $r = \pm\sqrt{r^2}$, but is there more to it than that? Consider the Galton Father/Son data again. For most data sets, we can think of the data as fitting inside a circumscribing ellipse. The ellipse for the Galton data is quite fat, while for the Ball Drop Data would be quite thin. One way to interpret r , is that it is describing how fat the circumscribing ellipse will be.

In Figure 5, we have drawn in the major axis of the ellipse for both sets of data.

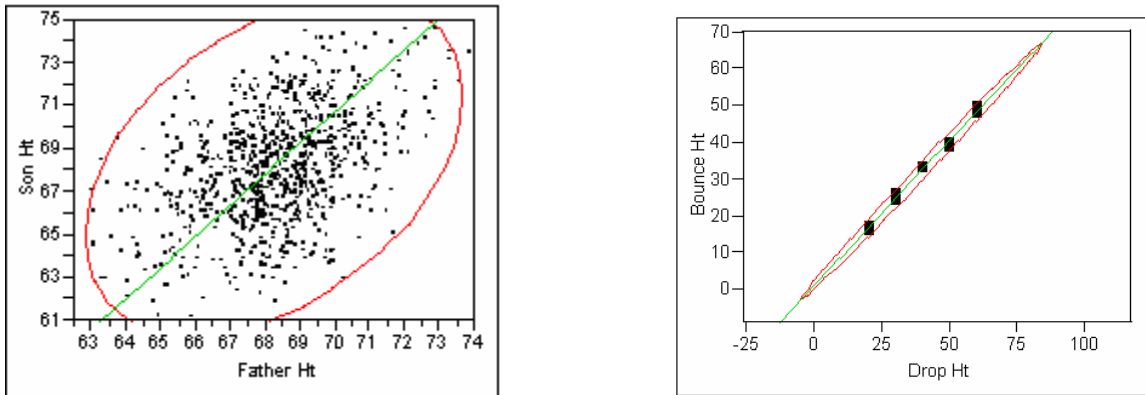


Figure 5: Galton Data ($r = 0.421$) and Ball Drop Data ($r = 0.997$) with Circumscribing Ellipses

For the Galton Data, the slope of the major axis is $\frac{s_y}{s_x} = \frac{s_{Son}}{s_{Father}} = \frac{2.5969696}{1.7876488} = 1.4527$. The line

passing through the center of the ellipse, $(\bar{x}, \bar{y}) = (68.267, 68.202)$, with a slope of $\frac{s_y}{s_x} = 1.4527$ is

called, the standard deviation line by Freedman, Pisani, and Purves.

The equation of the standard deviation line for this data is $Son = -30.8 + 1.45Father$. This is not the least squares line. The job of the least squares line is to estimate the average value of y for a given value of x .

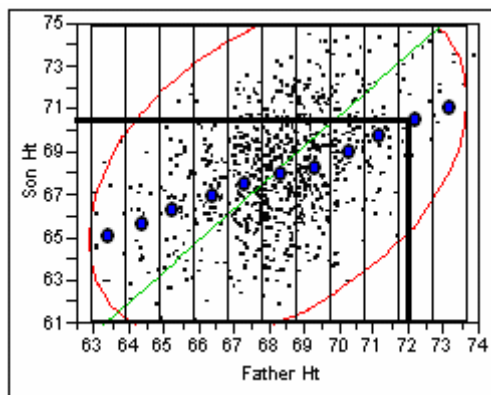


Figure 6: Galton Data with Vertical Strips

In this context, the job of the regression line is to estimate the average Son's height for a given Father's Height. To visualize this, divide the data into vertical strips associated with different values of the independent variable. For the vertical strip with a Father's Height of 72, it looks like the average Son's Height is around 70. The least squares regression line should attempt to pass through or estimate the

centers of all these vertical strips. In Figure 6, we estimate the center of each vertical strip with a blue circle. The least square regression line should approximate those circles.

Notice that the slope the line through the center of the vertical strips is smaller than the slope of the major axis (the standard deviation line). How much smaller? The line through the centers is always regressed towards the mean line $y = \bar{y}$. We need a number between 0 and 1 that we can

multiply $\frac{s_y}{s_x}$, the slope of the major axis, to describe how different the regression line is from the

standard deviation line. If the ellipse is quite thin, the two slopes will be very close, so the multiplier will be close to one. If the ellipse is fat, then the two lines will diverge, and the multiplier will be close to zero. Do we have a measure between 0 and 1 that might fit the bill? Yes, the correlation coefficient is just the multiplier we need.

Bivariate Fit of child ht By parent ht

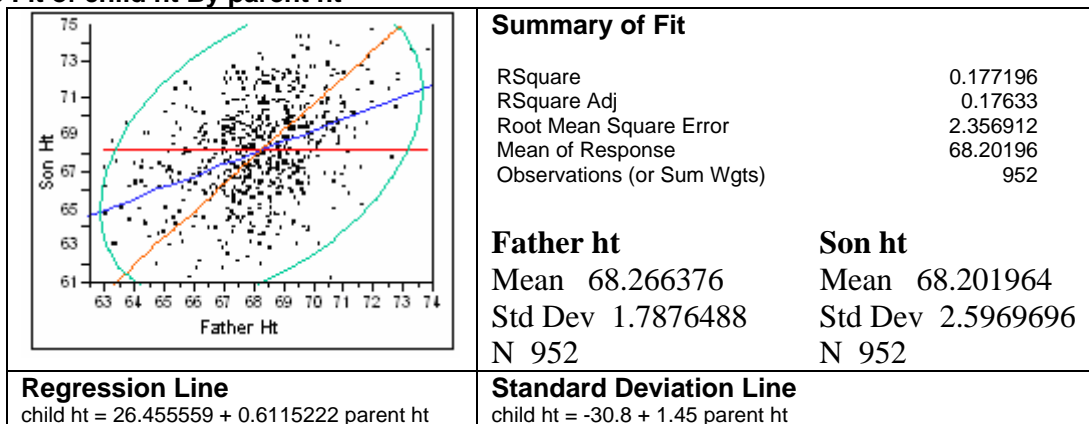


Figure 7: Standard Deviation Line, Regression Line, and Mean Line

The slope of the regression line is exactly $r \frac{s_y}{s_x}$. This means that for a 1 standard deviation increase in x , we expect an increase in y of r standard deviations. For the Galton data, the slope of the regression line is $m_{SF} = r \cdot \frac{s_{Son}}{s_{Father}} = (0.42095) \frac{2.5969696}{1.7876488} = 0.6115$. Since the regression line also passes through $(\bar{x}, \bar{y}) = (68.267, 68.202)$, the equation of the regression line is $Son = 26.46 + 0.6115 Father$. Figure 7 illustrates the relationship between the standard deviation line, the regression line, and the line of the mean.

Regressing y on x and x on y

This diagram also helps explain why we cannot use the regression line y on x to predict values of x from y . If we want to predict Father's Height from the Son's height, then we want a line that passes through the horizontal strips shown in Figure 8. In Figure 7, we saw that the average Son's height for Father's who were 72 inches tall was 70 inches tall. But the average Father's height for son's that are 70 inches tall is not 72 inches. Judging from Figure 8, we see that the average Father's height for son's that are 70 inches tall is about 68.5 inches.

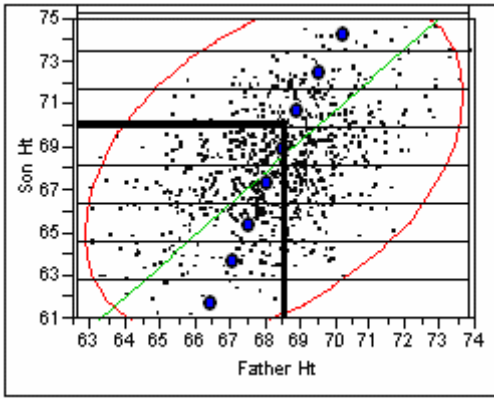


Figure 8: Horizontal Strips

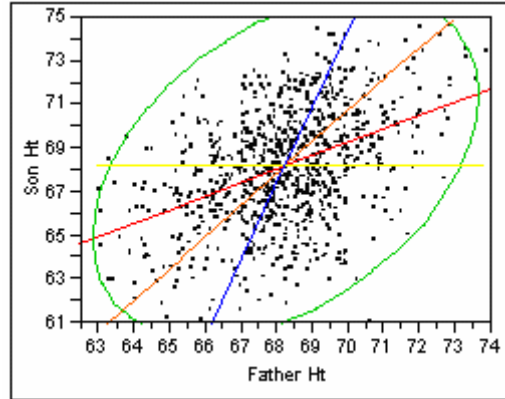


Figure 9: Regressing y on x and x on y

Finally, we also see that regression line for Fathers on Sons will have a slope of $m_{FS} = r \cdot \frac{s_{Father}}{s_{Son}} = (0.42095) \frac{1.7876488}{2.5969696} = 0.2898$. The line is as far away from the major axis of the ellipse as was the other regression line. So, $(m_{FS})(m_{SF}) = \left(r \cdot \frac{s_{Father}}{s_{Son}} \right) \left(r \cdot \frac{s_{Son}}{s_{Father}} \right) = r^2$ and $r = \pm \sqrt{(m_{FS})(m_{SF})}$. The correlation coefficient is, among other things, the geometric mean of the two slopes. Naturally, if the two slopes are reciprocals of each other, then $r = 1$.

A Vector Approach to Linear Regression

The question most often asked when students begin their study of linear regression and curve fitting is, “why do we minimize the sum of the **squares** of the errors?” Squaring the errors seems like an artificial measure of the total error of the fit. Typically, we fumble around with answers like “we want to make all the errors positive, so positive and negative errors won't negate each other”. Then we are faced with explaining why working with squares is simpler than working with absolute values, which accomplish the same task without altering the size of the errors. To understand why the sum of the squares of the errors is the “natural” measure of the fit rather than the artificial measure it appears to be, we need to think about the problem geometrically.

Simple Linear Regression

Given a set of n data points $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), \dots, (x_n, y_n)$, we can think of the data as defining two vectors \vec{x} and \vec{y} , with

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \text{ and } \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}.$$

With this interpretation, we re-consider the linear equation $\vec{y} = a + b\vec{x}$. This vector equation now makes no sense, since a and b are scalars and \vec{x} and \vec{y} are $n \times 1$ vectors. Implied by the equation is an $n \times 1$ vector of 1's, which we will call $\vec{1}$. Now the equation

$$\vec{y} = a\vec{1} + b\vec{x}$$

is well defined.

If the equation $\vec{y} = a\vec{1} + b\vec{x}$ is satisfied, then all of the data lie precisely on a line, as shown in Figure 10. More importantly, we interpret the vector equation $\vec{y} = a\vec{1} + b\vec{x}$ as saying that vector \vec{y} lives in the plane defined by \vec{x} and $\vec{1}$.

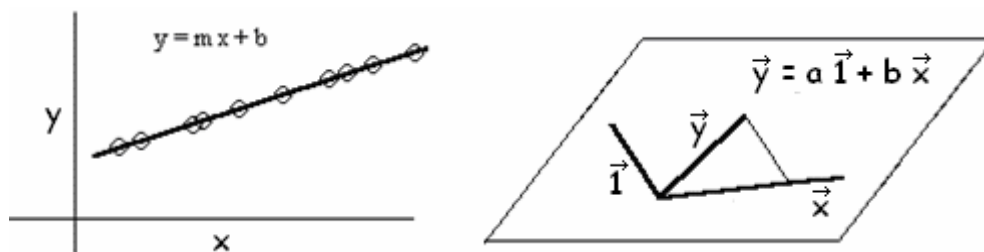


Figure 10: Data lie exactly on the line $y = a + bx$

The ordered pairs (1, 5), (2, 8), (4, 14), (5, 17), and (8, 26) illustrate this idea. We have the

vector equation, $\begin{bmatrix} 5 \\ 8 \\ 14 \\ 17 \\ 26 \end{bmatrix} = a \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + b \cdot \begin{bmatrix} 1 \\ 2 \\ 4 \\ 5 \\ 8 \end{bmatrix}$. If $a = 2$ and $b = 3$, we have an identity $\begin{bmatrix} 5 \\ 8 \\ 14 \\ 17 \\ 26 \end{bmatrix} = 2 \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + 3 \cdot \begin{bmatrix} 1 \\ 2 \\ 4 \\ 5 \\ 8 \end{bmatrix}$.

The vector $\begin{bmatrix} 5 \\ 8 \\ 14 \\ 17 \\ 26 \end{bmatrix}$ is indeed the sum of vectors $2 \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$ and $3 \cdot \begin{bmatrix} 1 \\ 2 \\ 4 \\ 5 \\ 8 \end{bmatrix}$. This means that these three vectors all live in the same space (a plane in 5-space).

Of course, in a real data situation, we cannot expect the equation $\vec{y} = a\vec{1} + b\vec{x}$ to hold. If there is any error at all, then $\vec{y} \neq a\vec{1} + b\vec{x}$, that is, the data do not all fall on the line. This means the vector \vec{y} is not in the plane of \vec{x} and $\vec{1}$. Nevertheless, we do want to write a linear model to represent the data set. That is, we want to know the y -values, \hat{y} , for which $\vec{y} = a\vec{1} + b\vec{x}$ is the best linear model for the data.

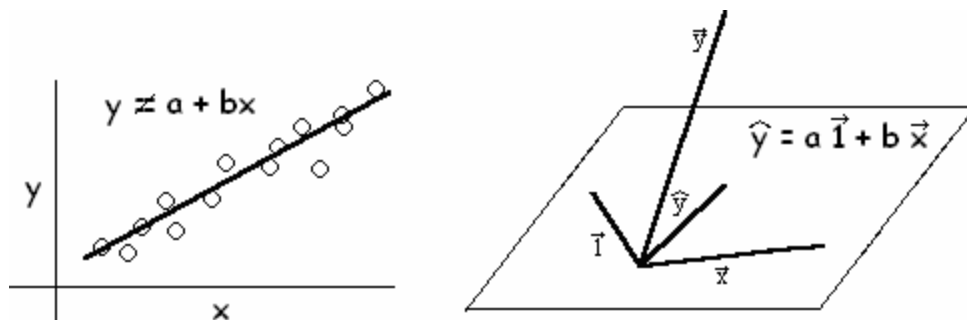


Figure 11: Data do not lie exactly on the line $y = a + bx$

The regression question can be stated geometrically as, "of all vectors in the plane of \vec{x} and $\vec{1}$, which is the closest to \vec{y} ?" We will use this vector as our model \hat{y} . Figure 12 illustrates the geometry of the regression problem.

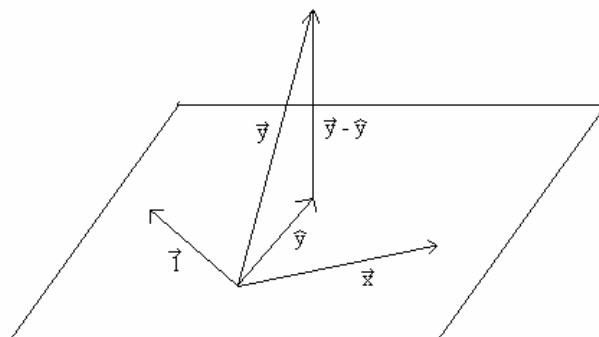


Figure 12: The Geometry of Linear Regression

The vector $\vec{r} = \vec{y} - \hat{y} = \vec{y} - (a\vec{1} + b\vec{x})$, known as the residual vector, is the difference in the actual vector \vec{y} and the model vector \hat{y} . The residual vector represents the error in the approximation of \vec{y} by \hat{y} . We want \hat{y} to be as close (literally) to \vec{y} as possible. To achieve the best fit, we want to minimize the length of \vec{r} . Remember, the length of a vector is found by the generalized Pythagorean Theorem, $\|\vec{r}\| = \sqrt{r_1^2 + r_2^2 + r_3^2 + \dots + r_n^2}$. To minimize the length of this vector, you must minimize the sum of the **squares** of its components. What is at first viewed as an artificial device to make terms positive is actually the “natural” (in the sense of Pythagorean) measure of the length of a vector!

Figure 12 contains all of the information required to compute the least squares components as well. Minimizing the length of \vec{r} requires that \vec{r} be perpendicular to the plane of \vec{x} and $\vec{1}$. Since \vec{r} is perpendicular to both \vec{x} and $\vec{1}$, we know that the dot products $\vec{r} \cdot \vec{x}$ and $\vec{r} \cdot \vec{1}$ are zero. Since $\vec{r} \cdot \vec{1} = 0$ we know immediately that the sum of the residuals is zero. Further, if

$$\vec{r} \cdot \vec{1} = 0,$$

then

$$(\vec{y} - \hat{y}) \cdot \vec{1} = (\vec{y} - a\vec{1} - b\vec{x}) \cdot \vec{1} = 0.$$

This gives the linear equation

$$\vec{y} \cdot \vec{1} = a(\vec{1} \cdot \vec{1}) + b(\vec{x} \cdot \vec{1}).$$

Similarly, if

$$(\vec{y} - \hat{y}) \cdot \vec{x} = 0,$$

then

$$(\vec{y} - a\vec{1} - b\vec{x}) \cdot \vec{x} = 0.$$

This generates the linear equation

$$\vec{y} \cdot \vec{x} = a(\vec{1} \cdot \vec{x}) + b(\vec{x} \cdot \vec{x}).$$

To determine the values of a and b that minimize the size of the residual vector, we only have to solve a system of two linear equations in two unknowns,

$$\vec{y} \cdot \vec{1} = a(\vec{1} \cdot \vec{1}) + b(\vec{x} \cdot \vec{1})$$

$$\vec{y} \cdot \vec{x} = a(\vec{1} \cdot \vec{x}) + b(\vec{x} \cdot \vec{x})$$

Solving this system, we find that

$$b = \frac{(\vec{y} \cdot \vec{x})(\vec{1} \cdot \vec{1}) - (\vec{y} \cdot \vec{1})(\vec{1} \cdot \vec{x})}{(\vec{x} \cdot \vec{x})(\vec{1} \cdot \vec{1}) - (\vec{x} \cdot \vec{1})(\vec{1} \cdot \vec{x})} \quad \text{and} \quad a = \frac{(\vec{y} \cdot \vec{1})(\vec{x} \cdot \vec{x}) - (\vec{y} \cdot \vec{x})(\vec{x} \cdot \vec{1})}{(\vec{x} \cdot \vec{x})(\vec{1} \cdot \vec{1}) - (\vec{x} \cdot \vec{1})(\vec{1} \cdot \vec{x})}.$$

Evaluating the dot products generates the more conventional forms

$$b = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \quad \text{and} \quad a = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}.$$

For the data set (2, 7), (3, 10), (4, 10), (5, 14) and (6, 15) we have $\vec{x} = \begin{bmatrix} 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix}$ and $\vec{y} = \begin{bmatrix} 7 \\ 10 \\ 10 \\ 14 \\ 15 \end{bmatrix}$, with

$$\sum x_i = 20, \quad \sum y_i = 56, \quad \sum x_i y_i = 244, \quad \sum x_i^2 = 90, \quad \left(\sum x_i \right)^2 = 400 \quad \text{and} \quad n = 5.$$

Substituting, we have $b = \frac{5(244) - (20)(56)}{5(90) - 400} = 2$ and $a = \frac{(56)(90) - (244)(20)}{5(90) - 400} = 3.2$, so the least squares

linear fit is $y = 3.2 + 2x$. From the vector form of the equation, we find that $\hat{y} = 3.2 \cdot \vec{1} + 2 \cdot \vec{x} = \begin{bmatrix} 7.2 \\ 9.2 \\ 11.1 \\ 13.2 \\ 15.2 \end{bmatrix}$ and

the residual vector is $\vec{y} - \hat{y} = \begin{bmatrix} 7 \\ 10 \\ 10 \\ 14 \\ 15 \end{bmatrix} - \begin{bmatrix} 7.2 \\ 9.2 \\ 11.1 \\ 13.2 \\ 15.2 \end{bmatrix} = \begin{bmatrix} -0.2 \\ 0.8 \\ -1.2 \\ 0.8 \\ -0.2 \end{bmatrix} = \vec{r}.$

Correlation Coefficient

The correlation coefficient, r , can also be related to the vectors in Figure 12. One form of the

computational formula for r^2 is $r^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$, where \bar{y} is the vector whose elements are all the average

of the y -values. The vector \bar{y} can be found by projecting the vector \hat{y} onto the $\vec{1}$ vector. All that is necessary, though, is to recognize that the vector \bar{y} is in the direction of $\vec{1}$.

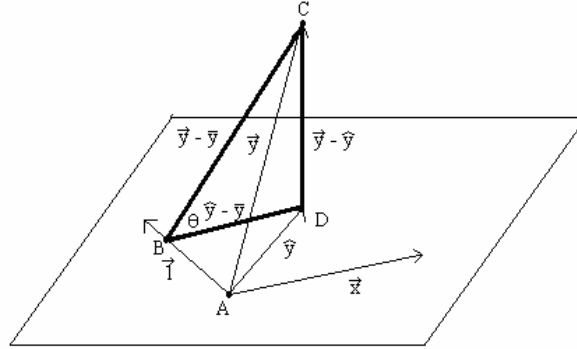


Figure 13: The correlation triangle

In Figure 13, the vector from A to B is vector \bar{y} . So, the vector from B to C, then, is vector $\hat{y} - \bar{y}$ and the vector from B to D is $\hat{y} - \bar{y}$. The length of $\bar{y} - \hat{y}$ is $|\bar{y} - \hat{y}| = \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}$ and of $\hat{y} - \bar{y}$ is

$|\hat{y} - \bar{y}| = \sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}$. These two vectors form the side and hypotenuse of a right triangle, so the ratio of

these two lengths is $\cos(\theta) = \frac{\sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$. But this is just $\sqrt{r^2} = |r|$. The correlation coefficient is the

cosine of the angle between $\bar{y} - \hat{y}$ and $\hat{y} - \bar{y}$. The value of the cosine varies from -1 to 1 , as expected for r . In our example,

$$\cos(\theta) = \frac{(\hat{y} - \bar{y}) \cdot (\bar{y} - \hat{y})}{|\hat{y} - \bar{y}| |\bar{y} - \hat{y}|} = \frac{40}{\sqrt{40} \sqrt{42.8}} \approx 0.9667.$$

References:

- Box, George, William Hunter, and J. Stuart Hunter, *Statistics for Experimenters*, John Wiley and Sons, 1978.
 Freedman, Pisani, Purves, *Statistics*, Norton, 1978.
 Saville, David J. and Wood, Graham R., *Statistical Methods, The Geometric Approach*, Springer-Verlag, 1991.