

Regression Analysis for Bivariate Data

I. The Regression Equations:

The error variation for a line $y = b_0 + b_1x$ and a set of measured scores $(x_i, y_i), 1 \leq i \leq n$, is given

by $S(b_0, b_1) = \sum_{i=1}^n (y_i - b_0 - b_1x_i)^2$. This expression defines S as a function of the two linear

parameters b_0 and b_1 ; b_0 is the intercept and b_1 is the slope. The so called “best fit” line is the one that minimizes the function $S(b_0, b_1)$. For this reason, regression analysis is sometimes called the method of “least squares”. We are minimizing the sum of squared vertical deviations from the line. More generally, we can model the “response” or output variable, y , in terms of the value of the “control” or input variable, x , by some function $y = F(x; b_0, b_1, \dots, b_k)$ where the

variables b_0, b_1, \dots, b_k are the parameters of the model. According to the method of least squares we determine values for these parameters from the observed (x, y) data points by minimizing the sum of squared y deviations from the values predicted by the model. That is, for a given set of

data, we choose b_0, b_1, \dots, b_k so as to minimize the sum $\sum_{i=1}^n [y_i - F(x_i; b_0, b_1, \dots, b_k)]^2$. If F is a

linear function of b_0, b_1, \dots, b_k the problem results in a linear system of k equations in the k unknown parameters. Even if F does not depend linearly on x , this method is still referred to as a linear regression, since the model is **linear** in the **parameters**. The solution of a non-linear least squares problem is also very important in modeling; however, the resulting equations are more complicated and generally the solutions can not be stated in closed form. The initial case presented above of y as a linear function of x is often called “simple linear regression” to distinguish it from cases where F is linear in b_0, b_1, \dots, b_k , but not linear in x .

The problem of minimizing a differentiable function of several variables is a standard one in calculus. A necessary condition is that the gradient of the function must vanish at the input values

that give a minimum. For $S(b_0, b_1)$ this means that $\frac{\partial S}{\partial b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1x_i) = 0$ and

$\frac{\partial S}{\partial b_1} = -2 \sum_{i=1}^n x_i (y_i - b_0 - b_1x_i) = 0$. These are called the **normal equations** of the regression

analysis. We thus have the following two by two system of equations for b_0 and b_1 .

$$\sum_{i=1}^n (y_i - b_0 - b_1x_i) = 0 \quad (1)$$

$$\sum_{i=1}^n x_i (y_i - b_0 - b_1x_i) = 0 \quad (2)$$

Since $S(b_0, b_1)$ can be made arbitrarily large by letting the values of either b_0 and b_1 get sufficiently large, S does not have an absolute maximum. However, $S(b_0, b_1)$ does have a lower bound of zero. Therefore $S(b_0, b_1)$ must have an absolute minimum. Since the above system has

only one solution, it must correspond to the absolute minimum. We will designate the values of (b_0, b_1) that solve (1) and (2) by the labels $(\hat{\beta}_0, \hat{\beta}_1)$. Then equation (1) implies that

$$\sum_{i=1}^n y_i - \hat{\beta}_0 n - \hat{\beta}_1 \sum_{i=1}^n x_i = 0. \text{ Dividing through this equation by } n \text{ gives the result that}$$

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}. \quad (3)$$

This says that the regression line must pass through the point (\bar{x}, \bar{y}) . Stated differently, on the average the regression model works! Since it takes two points to determine a line this is a necessary but not a sufficient condition to determine the values of $(\hat{\beta}_0, \hat{\beta}_1)$. Solving for $\hat{\beta}_0$, gives the identity

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \quad (4)$$

This result and equation (2) imply that $\sum_{i=1}^n x_i (y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})) = 0$.

Since $\hat{\beta}_1$ is a constant it can be factored out of the sum and isolated.

$$\hat{\beta}_1 \sum_{i=1}^n x_i (x_i - \bar{x}) = \sum_{i=1}^n x_i (y_i - \bar{y}) \quad (5)$$

Now the sum of the deviations about the mean must vanish; $\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n (y_i - \bar{y}) = 0$.

This, in turn, leads to the familiar form of the variation in x ,

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

A similar result holds for the covariation of x and y :

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

Giving names to the variation in x and covariation of x and y , we solve equation (5) for $\hat{\beta}_1$.

$$SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i (x_i - \bar{x}) = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \quad (6)$$

$$SS_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i (y_i - \bar{y}) = \sum_{i=1}^n y_i (x_i - \bar{x}) = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n} \quad (7)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i (y_i - \bar{y})}{\sum_{i=1}^n x_i (x_i - \bar{x})} = \frac{SS_{xy}}{SS_{xx}} \quad (8)$$

Substituting this result into equation (4),

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = \sum_{i=1}^n \frac{y_i}{n} - \bar{x} \frac{SS_{xy}}{SS_{xx}} = \sum_{i=1}^n \frac{y_i}{n} - \frac{\bar{x}}{SS_{xx}} \sum_{i=1}^n y_i (x_i - \bar{x}) \\ &= \sum_{i=1}^n y_i \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SS_{xx}} \right) \quad (9) \\ &= \frac{\sum_{i=1}^n y_i (SS_{xx} - n\bar{x}(x_i - \bar{x}))}{nSS_{xx}} \\ &= \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n y_i \left(\frac{\sum_{i=1}^n x_i \right)^2}{nSS_{xx}} - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i + \sum_{i=1}^n x_i \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}}{nSS_{xx}} \\ &= \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{nSS_{xx}} = \frac{\bar{y} \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i y_i}{SS_{xx}} \\ &= \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \quad (10) \end{aligned}$$

To represent formulas more concisely we use the shorthand notation that $\sum U = \sum_{i=1}^n U_i$, where the index is not stipulated and the lower and upper limits are understood to be 1 and n , respectively.

Also an average can be represented using bar notation, $\bar{U} = \frac{\sum U}{n}$. Hence, we can write the following formulas for the population variance and covariance.

$$\begin{aligned} \frac{SS_{xx}}{n} &= \overline{(x - \bar{x})^2} = \overline{x^2} - (\bar{x})^2 \\ \frac{SS_{yy}}{n} &= \overline{(y - \bar{y})^2} = \overline{y^2} - (\bar{y})^2 \\ \frac{SS_{xy}}{n} &= \overline{(x - \bar{x})(y - \bar{y})} = \overline{x(y - \bar{y})} = \overline{y(x - \bar{x})} = \overline{xy} - (\bar{x})(\bar{y}) \end{aligned} \tag{11}$$

The sample variances and covariances are given by

$$\begin{aligned} var(x) &= s_x^2 = \frac{SS_{xx}}{n-1} \\ var(y) &= s_y^2 = \frac{SS_{yy}}{n-1} \\ cov(x, y) &= \frac{SS_{xy}}{n-1} \end{aligned} \tag{12}$$

The regression equations for determining $\hat{\beta}_0$ and $\hat{\beta}_1$ can be written in a variety of equivalent forms. Different authors and disciplines prefer some forms more than others, but the different equations all give the same results. The table below provides a summary.

Table 1

$\hat{\beta}_0$	$\hat{\beta}_1$
$\frac{\sum y \sum x^2 - \sum xy \sum x}{n \sum x^2 - (\sum x)^2}$	$\frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$
$\frac{\bar{y} \sum x^2 - \bar{x} \sum xy}{\sum x^2 - \frac{(\sum x)^2}{n}}$	$\frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$
$\bar{y} - \hat{\beta}_1 \bar{x}$	$\frac{SS_{xy}}{SS_{xx}}$
$\overline{(y - \hat{\beta}_1 x)}$	$\frac{\overline{(x - \bar{x})(y - \bar{y})}}{\overline{(x - \bar{x})^2}} = \frac{\overline{xy} - (\bar{x})(\bar{y})}{\overline{x^2} - (\bar{x})^2}$
$\frac{cov(xy, x) - cov(x^2, y)}{var(x)}$	$\frac{cov(x, y)}{var(x)}$
$\frac{\overline{(x^2)\bar{y}} - \bar{x}\overline{(xy)}}{\overline{(x(x - \bar{x}))}} = \frac{\overline{(xy(x - \bar{x}))} - \overline{(x^2(y - \bar{y}))}}{\overline{(x(x - \bar{x}))}}$	$\frac{\overline{(y(x - \bar{x}))}}{\overline{(x(x - \bar{x}))}}$

Note: Johnson in his text uses α for the regression intercept instead of $\hat{\beta}_0$ and β for the regression slope instead of $\hat{\beta}_1$. This could easily cause confusion with Type I and II error probabilities. However, in section 11.3 he does adopt to the more standard $\hat{\beta}_j$ notation consistent with these notes.

As the formulas in Table 1 illustrate there are two distinct approaches in doing a regression analysis. One approach calculates the following five sums and then formulates all remaining calculations in terms of these sums and the number of data points, n .

$$\sum x \quad \sum y \quad \sum xy \quad \sum x^2 \quad \sum y^2$$

The approach of these notes is to first calculate the above five sums and then use these sums to calculate the five additional quantities shown below. Any additional calculations are then formulated in terms of these quantities and n .

$$\bar{x} \quad \bar{y} \quad SS_{xy} \quad SS_{xx} \quad SS_{yy}$$

As you may have noticed none of the regression equations encountered so far use $\sum y^2$ (or equivalently, SS_{yy}). That will change in the next two sections

II. Inference and Regression

In order to perform statistical inference on a regression model we need to consider more carefully the nature of the relationship between the x and y variables. Implicit in the least squares approach of fitting two variables is what is called a probabilistic model. This means that one variable is the response (or output, or dependent) variable and the other is the control (or input, or independent) variable. We will let y designate the response variable and x designate the control variable. Since x is under experimental control, we assume that its values were set and the error or uncertainty present in determining these values is essentially zero. The measured values of y depend on x , but they are also subject to an inherent error. The probabilistic model asserts that

$$y = F(x; \beta_0, \beta_1, \dots, \beta_k) + \varepsilon_x.$$

F is a function of x and a set of fixed parameters $\beta_0, \beta_1, \dots, \beta_k$. Epsilon, ε_x , represents the **random** error present in measuring y at the given value of x . If no error were present the model would be called deterministic. That would mean that the value of the input completely determines the value of the output, the essential meaning of a function. However, life is usually not so simple and random errors are often part of our data. We assume that for each value of x , $\langle \varepsilon_x \rangle = \mu_{\varepsilon_x} = 0$. On the average the error vanishes. If the expected value of the error is not zero, this is considered deterministic and it means we need to modify the function F to explain it. Ultimately random error is random. It can't be predicted in advance of the measurement. Any changes in the y values that can be explained by changes in the x value need to be incorporated into the model function.

In fitting a model to measured data it made perfect sense to minimize just the vertical deviations of the data from the model predictions. The horizontal values, since we control x , are “right where they ought to be”. In a least squares fit there is a fundamental asymmetry between x and y . In

algebra, if $y = mx + b$, then $x = \frac{1}{m}y - \frac{b}{m}$ is an equivalent equation. But in a linear regression, this

is not the case. Suppose we perform a simple linear regression on y and x , with x as the control variable. Now exchange the roles of x and y and do a second simple linear regression that fits x in terms of y . The two lines we get are not equivalent! Their slopes are not reciprocals nor are their intercepts related as expected. The lines are not equivalent since the two variables are treated differently in the regression analysis. All of the uncertainty is assumed to reside in the response variable and none in the control variable.

In practice, there is usually error or random uncertainty in both of our variables and in a sense a least squares fit of y in terms of x is not appropriate. Nevertheless, it is still often performed. For example, the error in x may be quite small compared to that of y . Furthermore, a linear regression in the age of computers and calculators is very easy to perform and enables us to make needed predictions on what response to expect for a given input. Correlation analysis (section 11.6 of the R. Johnson text) does not assume that x has zero uncertainty and so treats both variables equivalently. The correlation coefficient of y versus x is the same as the correlation coefficient of x versus y . There is also a procedure, known as “generalized least squares”. For a linear model this method minimizes the sum of the squared distances of the data points from the fitted line, i.e.,

$$\sum_{i=1}^n \frac{(y_i - b_0 - b_1 x_i)^2}{b_1^2 + 1}$$
, rather than just the sum of the vertical distances squared. The problem has a straightforward solution, but the equation for the slope is quadratic rather than linear.

To summarize the above discussion, in our inferences about the regression model we will assume that x is the control variable and there is no random error in determining x values. However, to be precise in our confidence intervals and hypothesis tests about the population parameters $\beta_0, \beta_1, \dots, \beta_k$ that characterize the model, we need to make further assumptions about the distribution of y values. In particular, we assume that the distribution of the random error, ϵ_x , is normal with a mean of zero and a standard deviation, σ , that does not depend on x . Thus, the random errors in y are independent. If we were to look at the probability density function of y , $f(x, y)$, it will be a normal distribution about $y = F(x; \beta_0, \beta_1, \dots, \beta_k)$, i.e., the cross sections of the surface, $z = f(x, y)$ at fixed x will be Gaussians with a mean of $F(x; \beta_0, \beta_1, \dots, \beta_k)$. If the standard deviation, σ , is a constant, then all of the cross sections have the same spread. This is illustrated below in Figure 1 for the linear model. It is still possible to do statistical inference if the variance of the random error in y is a function of x . However, this leads to a different set of normal equations using a procedure known as “weighted least squares”. This type of analysis requires either prior knowledge of the variance as a function of x , or, as is more commonly the case, repeated measurements of y at fixed x in order to estimate this variance. An illustration of a variable spread in y for the linear model is shown in Figure 2.

Figure 1
Normal Distribution of y about $y = \beta_0 + \beta_1 x$
Sigma a constant

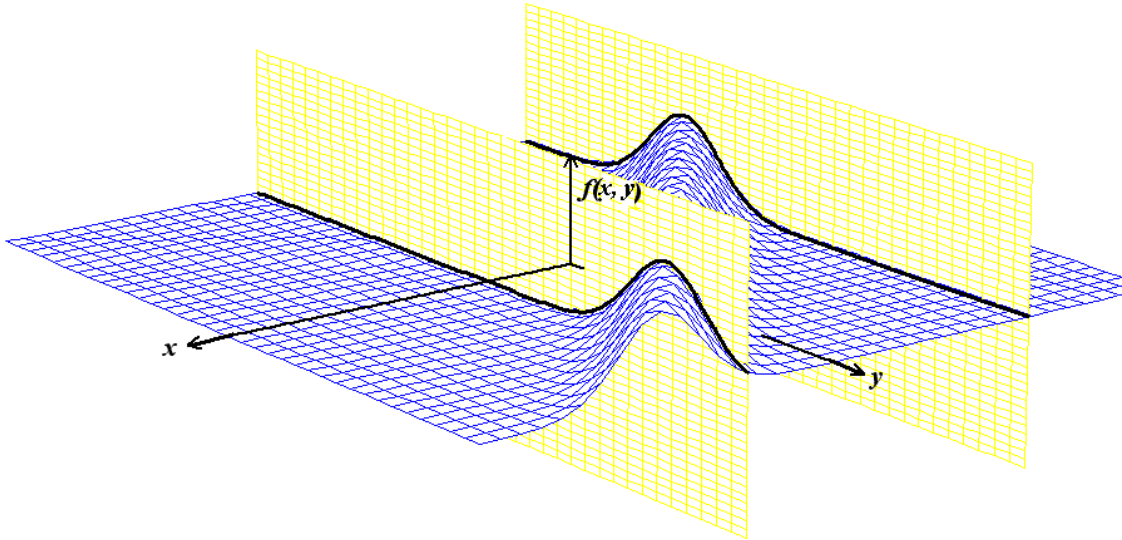
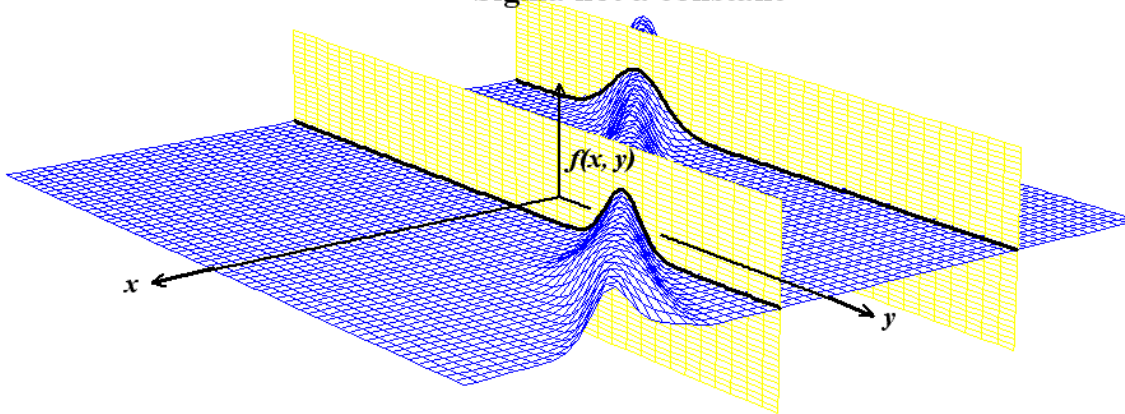


Figure 2
Normal Distribution of y about $y = \beta_0 + \beta_1 x$
Sigma not a constant



From this point on we will assume that $y = \beta_0 + \beta_1 x + \epsilon$ and ϵ is in $N(0, \sigma)$. The parameter β_1 is the population slope and the parameter β_0 is the population intercept. The regression equations for $\hat{\beta}_1$ and $\hat{\beta}_0$ shown in Table 1 are linear in the measured y values. By our assumptions, there is no random error in the x values and $\langle y_i \rangle = \beta_0 + \beta_1 x_i$. The properties of the expected value and variance of a linear combination of independent random variables lead to the following results.

$$\begin{aligned} \langle \hat{\beta}_1 \rangle &= \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{SS_{xx}} \right) \langle y_i \rangle = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{SS_{xx}} \right) (\beta_0 + \beta_1 x_i) = \frac{\beta_0}{SS_{xx}} \sum_{i=1}^n (x_i - \bar{x}) + \frac{\beta_1}{SS_{xx}} \sum_{i=1}^n x_i (x_i - \bar{x}) \\ &= \frac{\beta_0}{SS_{xx}} (0) + \frac{\beta_1}{SS_{xx}} \cancel{SS_{xx}} \end{aligned}$$

$$\langle \hat{\beta}_1 \rangle = \beta_1 \quad (13)$$

$$\text{Var}(\hat{\beta}_1) = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{SS_{xx}} \right)^2 \text{Var}(y_i) = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{SS_{xx}} \right)^2 \text{Var}(\beta_0 + \beta_1 x_i + \varepsilon_i) = \frac{\sigma^2}{SS_{xx}^2} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{\sigma^2}{SS_{xx}}$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SS_{xx}} \quad (14)$$

For the regression slope we make use of equation (9).

$$\begin{aligned} \langle \hat{\beta}_0 \rangle &= \sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SS_{xx}} \right) \langle y_i \rangle = \sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SS_{xx}} \right) (\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SS_{xx}} \right) + \beta_1 \sum_{i=1}^n \left(\frac{x_i}{n} - \frac{\bar{x}}{SS_{xx}} x_i (x_i - \bar{x}) \right) \\ &= \beta_0 \left(\sum_{i=1}^n \frac{1}{n} - \frac{\bar{x}}{SS_{xx}} \sum_{i=1}^n (x_i - \bar{x}) \right) + \beta_1 \left(\sum_{i=1}^n \frac{x_i}{n} - \frac{\bar{x}}{SS_{xx}} \sum_{i=1}^n x_i (x_i - \bar{x}) \right) \\ &= \beta_0 \left(\frac{n}{n} - \frac{\bar{x}}{SS_{xx}} (0) \right) + \beta_1 \left(\bar{x} - \frac{\bar{x}}{SS_{xx}} \cancel{SS_{xx}} \right) = \beta_0 (1) + \beta_1 (0) \end{aligned}$$

$$\langle \hat{\beta}_0 \rangle = \beta_0 \quad (15)$$

$$\begin{aligned} \text{Var}(\hat{\beta}_0) &= \sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SS_{xx}} \right)^2 \text{Var}(y_i) = \sum_{i=1}^n \left(\frac{1}{n^2} - \frac{2\bar{x}(x_i - \bar{x})}{nSS_{xx}} + \left(\frac{\bar{x}}{SS_{xx}} \right)^2 (x_i - \bar{x})^2 \right) \sigma^2 \\ &= \sigma^2 \left(\frac{n}{n^2} - \frac{2\bar{x}}{nSS_{xx}} \sum_{i=1}^n (x_i - \bar{x}) + \left(\frac{\bar{x}}{SS_{xx}} \right)^2 \sum_{i=1}^n (x_i - \bar{x})^2 \right) = \sigma^2 \left(\frac{1}{n} - 0 + \left(\frac{\bar{x}}{SS_{xx}} \right)^2 SS_{xx} \right) \end{aligned}$$

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x})^2}{SS_{xx}} \right) \quad (16)$$

Thus, $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased sample estimates of the population slope and population intercept, respectively.

For a set of n data points (x_i, y_i) , we might try to estimate σ by the sample standard

deviation $\sqrt{\frac{\sum_{i=1}^n (\varepsilon_i)^2}{n-1}}$. However, since we don't know β_1 or β_0 , we can't compute the actual

values of ε_i . Instead, we need to estimate ε_i by using the sample based estimates $\hat{\beta}_1$ and $\hat{\beta}_0$ and the fitted regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad (17)$$

This estimate of random error, $\hat{\varepsilon}_i$, is called the residual of the linear fit at $x = x_i$. It's the vertical deviation of the data point (x_i, y_i) from the sample regression line. From equation (1), the sum of all of the residuals vanishes:

$$\sum_{i=1}^n \hat{\varepsilon}_i = \sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

The sum of squared residuals, called the Error Variation or the Residual Sum of Squares, is given by

$$SSE = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \quad (18)$$

The sample estimates $\hat{\beta}_1$ and $\hat{\beta}_0$ were chosen to make SSE as small as possible. To get a more convenient computational form for SSE , we use equations (4) and (11)

$$\begin{aligned} SSE &= \sum_{i=1}^n (y_i - \bar{y} - \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^n ((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}))^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2\hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= SS_{yy} - 2\hat{\beta}_1 SS_{xy} + \hat{\beta}_1^2 SS_{xx} = SS_{yy} - 2 \frac{SS_{xy}^2}{SS_{xx}} + \left(\frac{SS_{xy}}{SS_{xx}} \right)^2 SS_{xx} \end{aligned}$$

So,

$$SSE = SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}. \quad (19)$$

The variance of the random error can be estimated by

$$S_e^2 = \frac{SSE}{n-2} = \frac{SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}}{n-2}. \quad (20)$$

This is an unbiased estimate of σ^2 and under the assumption of normality, $\frac{(n-2)S_e^2}{\sigma^2}$ is distributed as chi-squared with $n - 2$ degrees of freedom. The number of degrees of freedom is $n - 2$ because the sample mean and sample regression slope each “eat up” a degree of freedom.

The standard deviation, σ , can then be estimated by

$$S_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}}{n-2}} \quad (21)$$

S_e is called the **standard error of estimate** and is used in all of our regression inferences.

Using equations (14) and (16) the standard errors of the regression slope and intercept are estimated by the formulas given below.

$$S_{\hat{\beta}_1} = \frac{S_e}{\sqrt{SS_{xx}}} \quad (22)$$

$$S_{\hat{\beta}_0} = S_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}} \quad (23)$$

The following two scores are both distributed as a t distribution with $n - 2$ degrees of freedom.

$$\frac{\beta_1 - \hat{\beta}_1}{S_{\hat{\beta}_1}} \quad \frac{\beta_0 - \hat{\beta}_0}{S_{\hat{\beta}_0}}$$

Thus we can calculate confidence intervals and perform hypothesis tests on the population slope and intercept. In particular, $100(1 - \alpha)\%$ confidence intervals are computed by the formulas.

$$\beta_1 : \hat{\beta}_1 \pm t_{\alpha/2} S_{\hat{\beta}_1} \quad (24) \quad \beta_0 : \hat{\beta}_0 \pm t_{\alpha/2} S_{\hat{\beta}_0} \quad (25)$$

One of the primary uses of a regression line is prediction. Suppose we want to estimate y at $x = x_0$. The regression estimate is

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{y} + \hat{\beta}_1 (x_0 - \bar{x}) = \bar{y} + \frac{SS_{xy}}{SS_{xx}} (x_0 - \bar{x}) = \sum_{i=1}^n y_i \left(\frac{1}{n} + \frac{x_0 - \bar{x}}{SS_{xx}} (x_i - \bar{x}) \right).$$

Again the properties of a linear combination of independent random variables allow us to derive the following results.

$$\begin{aligned} \langle \hat{y}_0 \rangle &= \sum_{i=1}^n \langle y_i \rangle \left(\frac{1}{n} + \frac{x_0 - \bar{x}}{SS_{xx}} (x_i - \bar{x}) \right) = \sum_{i=1}^n (\beta_0 + \beta_1 x_i) \left(\frac{1}{n} + \frac{x_0 - \bar{x}}{SS_{xx}} (x_i - \bar{x}) \right) \\ &= \beta_0 \sum_{i=1}^n \left(\frac{1}{n} + \frac{x_0 - \bar{x}}{SS_{xx}} (x_i - \bar{x}) \right) + \beta_1 \sum_{i=1}^n \left(\frac{x_i}{n} + \frac{(x_0 - \bar{x})}{SS_{xx}} x_i (x_i - \bar{x}) \right) \\ &= \beta_0 \left(\frac{n}{n} + \frac{x_0 - \bar{x}}{SS_{xx}} (0) \right) + \beta_1 \left(\bar{x} + \frac{(x_0 - \bar{x})}{SS_{xx}} \cancel{SS_{xx}} \right) = \beta_0 + \beta_1 x_0 \end{aligned}$$

$$\begin{aligned}
 \text{Var}(\hat{y}_0) &= \sum_{i=1}^n \sigma^2 \left(\frac{1}{n} + \frac{x_0 - \bar{x}}{SS_{xx}} (x_i - \bar{x}) \right)^2 = \sigma^2 \sum_{i=1}^n \left(\frac{1}{n^2} + \frac{2(x_0 - \bar{x})}{nSS_{xx}} (x_i - \bar{x}) + \left(\frac{x_0 - \bar{x}}{SS_{xx}} \right)^2 (x_i - \bar{x})^2 \right) \\
 &= \sigma^2 \left(\frac{n}{n^2} + \frac{2(x_0 - \bar{x})}{nSS_{xx}} (0) + \left(\frac{x_0 - \bar{x}}{SS_{xx}} \right)^2 SS_{xx} \right) \\
 &= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}} \right)
 \end{aligned}$$

So the expected value of the **estimated** value of y at $x = x_0$ is the expected value of y at $x = x_0$. Stated differently, what $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ actually estimates is $\mu_{y|x_0}$, the mean or average value of many repeated y measurements at $x = x_0$. From the result for the variance of \hat{y}_0 , the standard error of $\mu_{y|x_0}$ is given by

$$S_{\mu_{y|x_0}} = S_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}. \quad (26)$$

Note: That the uncertainty of a prediction increases as we move farther from the center of the data. This result reminds us of the danger in extrapolating a line beyond any observed data points. In fact, if the underlying relationship between y and x is not linear and we make predictions outside of our data, this formula can drastically underestimate the uncertainty.

A prediction of a single y measurement at $x = x_0$ is called estimating a future observation. We will use the notation $\hat{y} | x_0$ for this estimate. By our assumptions the random error in this future y measurement at $x = x_0$ is independent of the random errors in the y measurements used to calculate the regression slope and intercept. Therefore, the variance of the estimated future observation is calculated as

$$\begin{aligned}
 \text{Var}(\hat{y} | x_0) &= \text{Var}(y - \mu_{y|x_0}) = \text{Var}(y) + (-1)^2 \text{Var}(\mu_{y|x_0}) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}} \right) \\
 &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}} \right)
 \end{aligned}$$

The standard error of this estimate is given by

$$S_{\hat{y}|x_0} = S_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}. \quad (27)$$

Because of the 1 under the radical, we see that a single measurement, as we would expect, is more uncertain than an average. The assumption of normality allows us to calculate $100(1 - \alpha)\%$ confidence intervals for the mean value of y and future observed value of y at $x = x_0$. The upper

and lower confidence limits for $\hat{y} | x_0$ are referred to as the limits of prediction for a future y value.

$$\mu_{y|x_0}: \hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2} S_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} \quad (28) \quad \hat{y} | x_0: \hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2} S_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} \quad (29)$$

III. Goodness of Fit: Checking the Adequacy of the Model

As in the analysis of variance we decompose each y measurement into its contributing terms.

$$y_i = \bar{y} + (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \quad (30)$$

The first term approximates the i 'th score by the sample mean. The second term corrects the sample mean by the regression model estimate to take into account y 's dependence on x . The last term is the i 'th residual and represents y 's supposedly random departure from the regression model. If there is no relationship between x and y , the best estimate for any y value would be the sample mean. Under this circumstance the model estimate \hat{y}_i should be close to \bar{y} for every value of x . For a simple linear regression this means that the regression slope should be approximately zero. Slope represents the rate of change of y with respect to x . If y does not depend on x , a change in x would leave y essentially unchanged. Hence, the slope should be very close to zero. Stated in contra positive form, a non-zero regression slope is evidence of a relationship between x and y .

Of course there is a very easy way to literally see if there's a relationship between x and y . Plot the data! A graph of the data points (x_i, y_i) , called a scatter diagram, should actually precede any calculations. The eye is a great tool for spotting patterns and relationships. It can happen that a sophisticated formula based on faulty assumptions can miss a relationship that is quite obvious from a graph. This is especially important in considering transformations on the data that could result in a linear fit.

Rewriting equation (30) to solve for the variation in y gives the following.

$$\begin{aligned} SS_{yy} &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n [(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)]^2 \\ &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \end{aligned}$$

Now,

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = \sum_{i=1}^n \hat{y}_i (y_i - \hat{y}_i) - \bar{y} \sum_{i=1}^n (y_i - \hat{y}_i) = \hat{\beta}_1 \sum_{i=1}^n x_i (y_i - \hat{y}_i) + (\hat{\beta}_0 - \bar{y}) \sum_{i=1}^n (y_i - \hat{y}_i).$$

From equations (2) and (1) both of the sums in the last expression are zero. Thus, we have

$$SS_{yy} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2. \quad (31)$$

The variation of y about its mean, SS_{yy} , is called the **Total Variation**. The second sum on the right side of equation (31) is the Residual Sum of Squares also called the **Error Variation** or **SSE** which we studied in last section on statistical inference. The first sum on the right side of equation (31) is also a sum of squares. It is called the Sum of Squares due to Regression or the **Explained Variation**. It measures the amount of variation in y which can be explained by

changes in x and y 's **linear dependence** on x . From equation (19), $SS_{yy} = \frac{SS_{xy}^2}{SS_{xx}} + SSE$. Thus, we conclude that the Explained Variation, SS_{ex} , can be easily computed by the following formula.

$$SS_{ex} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \frac{SS_{xy}^2}{SS_{xx}} \quad (32)$$

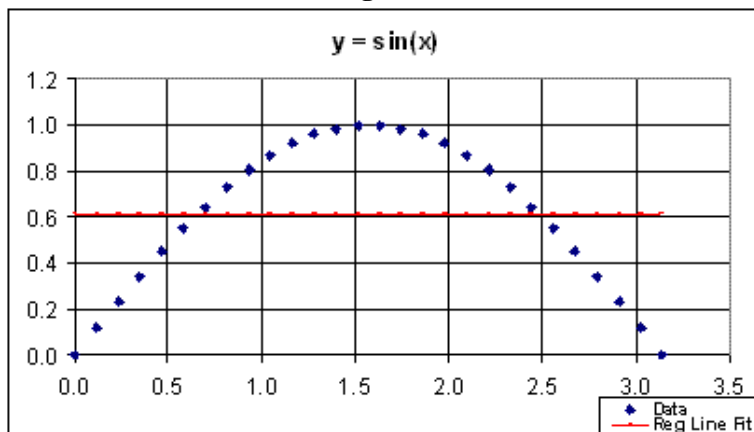
The ratio of the Explained Variation to the Total Variation is always between 0 and 1. A value of 0 would mean that none of y 's variability can be attributed to changes in x . This is the case of a regression slope of 0. If the value of this ratio is 1, then the Error Variation is 0. The linear model explains all of the variability in y . We have a "perfect fit". The ratio is called the coefficient of determination and is designated by r^2 .

$$r^2 = \frac{SS_{ex}}{SS_{yy}} = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}} \quad (33)$$

From equation (8), we can also write $r^2 = \frac{SS_{xx}\hat{\beta}_1^2}{SS_{yy}}$, which again demonstrates that a regression slope of zero is equivalent to a coefficient of determination equal to zero.

Note: We have argued that when there is no relationship between x and y , the regression slope will be close to zero. Thus, establishing at a given level of significance the alternative hypothesis that the regression slope is nonzero, establishes a relationship between x and y . The converse is **not true**. A zero regression slope or coefficient of determination does not establish the lack of a relationship between x and y . Consider the following: Suppose the underlying relationship between x and y is $y = \sin(x)$ and we sample 28 data points with the x values uniformly distributed on the interval $[0, \pi]$. This is shown in Figure 3.

Figure 3



A regression fit of this data gives $\hat{y} = \bar{y} + 0x$, i.e., both the regression slope and the coefficient of determination are zero. A plot of the horizontal regression line is also shown in Figure 3. There is certainly a deterministic relationship between y and x . However, the equation that describes this relationship is not well approximated by a straight line. The horizontal regression line explains none of the variation of y about its mean. But that does not imply that no explanation is possible. We just need to use a more sophisticated model. This example, while artificial, does illustrate the importance of first generating the scatter plot. Had we done this, we would not even have attempted the simple linear regression. The calculation of r^2 can be extended to regressions that model y with a nonlinear dependence on x .

$$r^2 = \frac{SS_{ex}}{SS_{yy}} = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{\sum_{i=1}^n [y_i - F(x; \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)]^2}{SS_{yy}}$$

For the data in Figure 3, the model $F(x; \beta_0, \beta_1) = \beta_0 + \beta_1 \sin(x)$ would yield a value of r^2 very close to 1.

The notation r^2 would seem to imply that there is also an r . There is and it is called the sample correlation coefficient. From equation (33) it is given by

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} \tag{34}$$

Since SS_{xy} can be negative r can take on values between -1 and 1 . Either extreme corresponds to $r^2 = 1$ and thus a perfect fit.

Note: The formula for r , unlike the regression slope, is symmetric in x and y . Like the regression slope there are a multitude of equivalent formulas for r . Some are presented in Table 2.

Table 2

$\hat{\beta}_1 \sqrt{\frac{SS_{xx}}{SS_{yy}}} = \hat{\beta}_1 \sqrt{\frac{SS_{xx}/(n-1)}{SS_{yy}/(n-1)}} = \hat{\beta}_1 \frac{s_x}{s_y}$	$\frac{n \sum xy - \sum x \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$
$\frac{cov(x, y)}{\sqrt{var(x)var(y)}} = \frac{cov(x, y)}{s_x s_y}$	$\frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{[\sum x^2 - \frac{(\sum x)^2}{n}][\sum y^2 - \frac{(\sum y)^2}{n}]}}$
$\frac{\sum_{i=1}^n \frac{y_i (x_i - \bar{x})}{n-1}}{s_x s_y} = \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{(n-1) s_x s_y}$	$\frac{\overline{(y(x - \bar{x}))}}{\sqrt{(x(x - \bar{x})) \cdot (y(y - \bar{y}))}}$
$\frac{1}{(n-1)} \sum_{i=1}^n \frac{y_i (x_i - \bar{x})}{s_y s_x}$	$\frac{1}{(n-1)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \frac{1}{(n-1)} \sum z_x z_y$

A negative correlation simply means that x and y move in “opposite” directions. As one variable increases, the other decreases. Of course, this is just what we expect when the regression slope is negative.

In the physical and engineering sciences where the identification and use of control and response variables is quite common, r^2 is usually emphasized and reported more often than is r . In the social sciences, where most variables have random distributions, r is often considered the more important statistic.

From Table 2 we can express the regression slope as $\hat{\beta}_1 = r \frac{s_y}{s_x}$. The linear regression model can

than be expressed as $\hat{y} = r \frac{s_y}{s_x} x + \hat{\beta}_0 = r \frac{s_y}{s_x} x + \bar{y} - r \frac{s_y}{s_x} \bar{x} = \bar{y} + s_y r \left(\frac{x - \bar{x}}{s_x} \right)$. Rearranging this

result, $\frac{\hat{y} - \bar{y}}{s_y} = r \left(\frac{x - \bar{x}}{s_x} \right)$. If we define the model or predicted z score as $\hat{z}_y = \frac{\hat{y} - \bar{y}}{s_y}$, the linear

regression model states that $\hat{z}_y = r \cdot z_x$. This equation explains the origin of the name “regression analysis”. Large z_x scores correspond to x scores which are significantly different than the mean. For these scores the regression model predicts y scores that are also removed from the mean; however, since $-1 < r < 1$, the predicted y scores tend to be closer to the mean of y than the corresponding x scores were to the mean of x . This movement of y to values closer to \bar{y} was called a “regression to the mean” and so the whole analysis was given this name.

We started this section with equation (30) and noting its similarity to the representation of y in an analysis of variance. The Explained Variation associated with the model is like the Treatment Sum of Squares. We compare the variance in y associated with the model to the residual or error variance in y . The linear model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ has two statistics, $\hat{\beta}_0$ and $\hat{\beta}_1$. However, the model has only one degree of freedom, since $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. The second degree of freedom is really due to the mean, \bar{y} , and is not associated with the linear model. Stated differently, two points determine a line, but the regression line must pass through the point (\bar{x}, \bar{y}) . The regression line has only 1 additional “free” point to choose in fitting the data. More generally a model with k parameters, $\hat{y} = F(x; \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$, would have $k - 1$ degrees of freedom. For the linear model we summarize the results in an ANOVA table as shown below.

Source	Sum of Squares	Degrees of Freedom	Mean Square
Linear Model	$SS_{ex} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \frac{SS_{xy}^2}{SS_{xx}}$	1	$\frac{SS_{ex}}{1}$
Error	$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}$	$n - 2$	$S_e^2 = \frac{SSE}{n - 2}$
Total	SS_{yy}	$n - 1$	$s_y^2 = \frac{SS_{yy}}{n - 1}$

The observed F score is $F = \frac{SS_{\text{ex}}}{S_e^2}$, and for a stated level of significance, α , we compare its value against $F_\alpha(v_1=1, v_2=n-2)$. If $F > F_\alpha(1, n-2)$, we conclude that the model can explain more of the variation in y than can be attributed to randomness. Another way to interpret the F ratio in the context of a regression fit is as follows. The P -value, $P = \Pr(f > F)$ in an F distribution with $v_1 = 1$ and $v_2 = n - 2$, is the probability that the observed scatter diagram of y versus x is due to “pure” chance and is not the result of a linear relationship between x and y .

From equation (33), we can relate the observed F score to the coefficient of

determination, $F = \frac{SS_{\text{ex}}}{S_e^2} = \frac{SS_{\text{ex}}}{(SSE)/(n-2)} = \frac{(n-2)SS_{\text{ex}}}{SS_{\text{yy}} - SS_{\text{ex}}} = \frac{(n-2)SS_{\text{ex}}/SS_{\text{yy}}}{1 - SS_{\text{ex}}/SS_{\text{yy}}}$, so

$$F = \frac{(n-2)r^2}{1-r^2}. \quad (35)$$

The fundamental notion of a probabilistic model is that the model should explain everything in the data that is capable of explanation. What’s left over, the residuals, is beyond explanation or prediction and is to be attributed to randomness. In our treatment of inference on regression parameters we assumed that this randomness was normally distributed with a variance that did not vary from point to point. Is there a way to experimentally examine these assumptions? Yes, **look at the residuals!** They are supposed to be random. If we see a discernable pattern, that’s not random. Any pattern seen means an aspect of the data, which is capable of explanation and should, at least in principle, be incorporated into the model.

One procedure is to plot the residuals versus the independent variable, x . If a pattern is seen, then strictly speaking the model being used is inadequate and needs to be improved in order to explain the pattern. Of course this is easier said than done. Here is where a fundamental knowledge of the system being studied can be crucial. Can a “mechanistic” model with parameters intrinsic to the system be developed? Statistical analysis is a very powerful tool, but it is a poor substitute for a fundamental understanding of the relationship between variables. In situations where no good mechanistic understanding exists or where a mechanistic approach is too complicated, we resort to empirical model building. Based on the pattern of residuals, we try different mathematical modifications of the model. Often this involves adding additional power law terms to the previous model. A full discussion of this problem is far beyond the scope of these notes and provides more than enough material for a separate course.

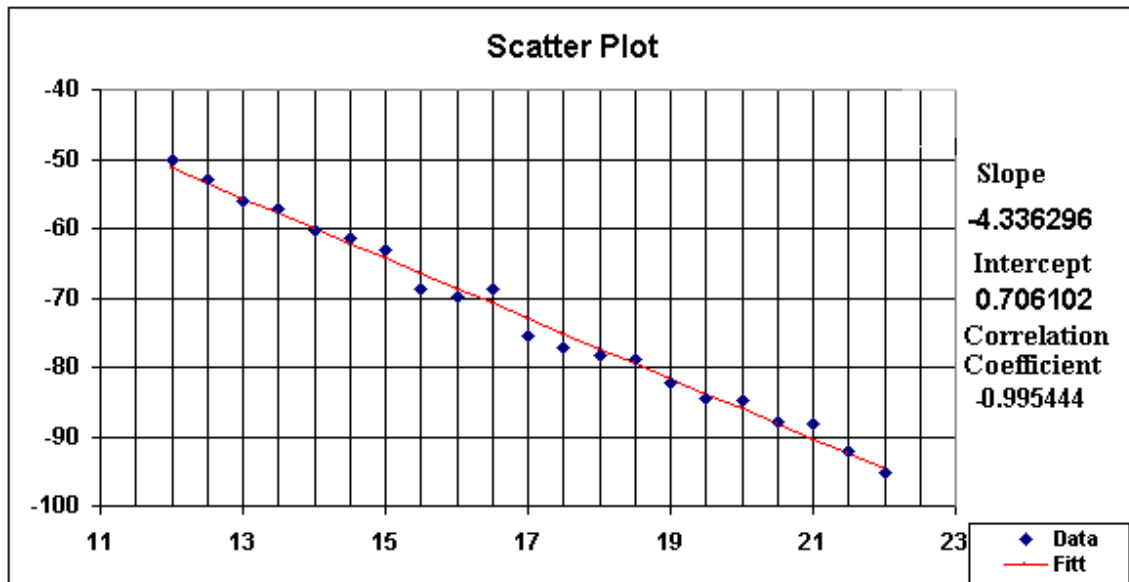
Even when the residuals appear randomly distributed we should examine whether the spread about zero changes as x changes. The assumption of a constant variance was used in deriving the methods for inference on the regression parameters. If there is evidence that the variance is not constant, we should consider doing a weighted least squares analysis.

A normal scores plot of the residuals provides a relatively easy visual check for departures from a normal distribution.

A second procedure in examining residuals is to plot them versus the order in which the data was taken. If any pattern appears this may be indicative of a deterministic error in the experimental set up.

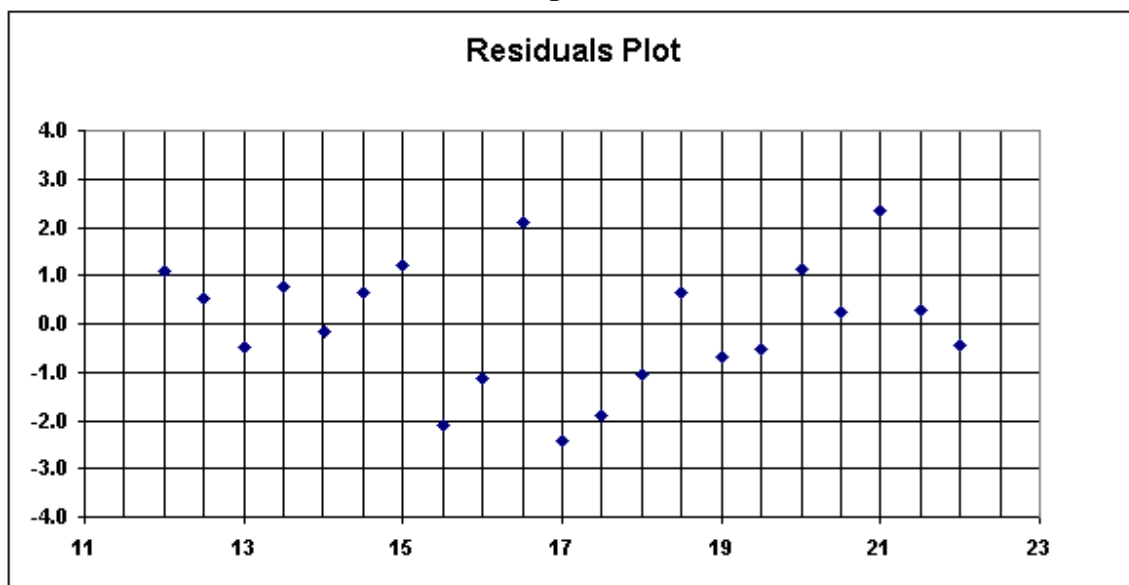
To illustrate the idea of examining the plot of residuals versus the independent variable consider the scatter diagram shown in Figure 4. This was generated in Excel using a linear relationship between x and y with population parameters, $\beta_1 = -4.5$ and $\beta_0 = 3.6$. Random error was simulated by adding to each y value, calculated from $y = \beta_0 + \beta_1 x$, a number, computed by the formula, $\text{sigma} * \text{NORMSINV}(\text{RAND}())$. The value of sigma was 1.4, so that the simulation had ϵ in $N(0, 1.4)$. The regression slope, intercept and sample correlation coefficient are also displayed.

Figure 4



The plot of residuals versus x for this same set of data is displayed in Figure 5.

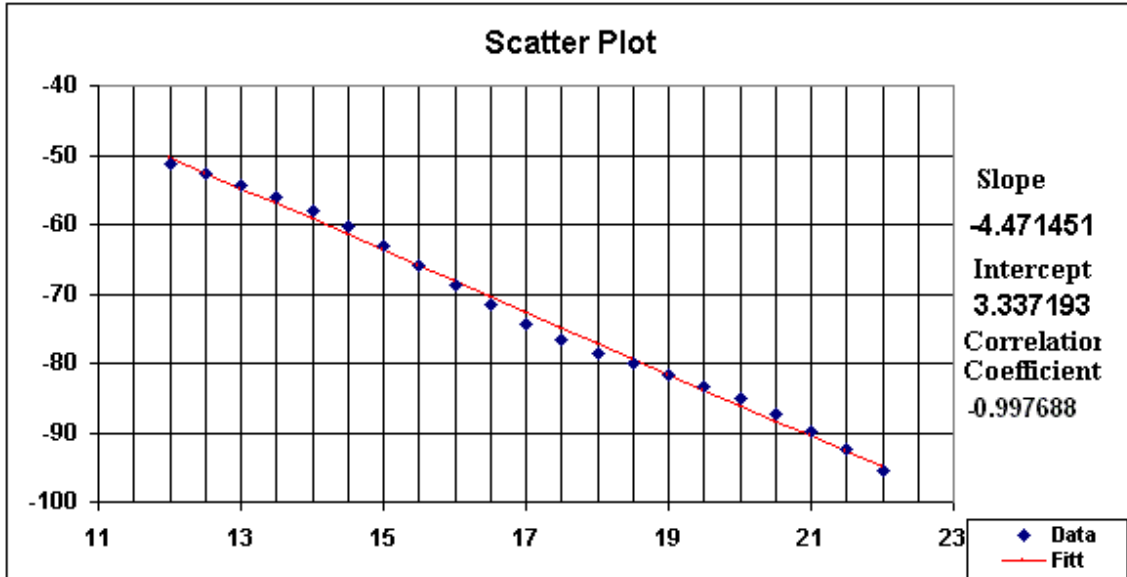
Figure 5



The pattern certainly appears random. Of course, this is to be expected given how the deviations from the linear model were generated.

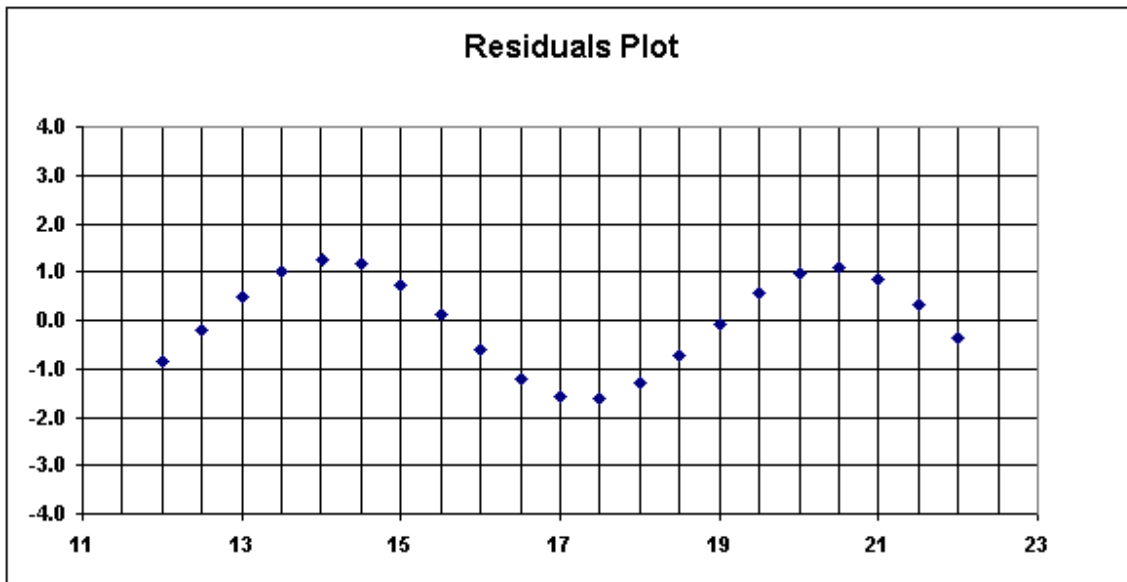
The data displayed in figure 6 was generated in Excel using the same linear relationship, but instead of random error, $1.4 * \text{SIN}(x)$ was added to each y value. This represents a deterministic rather than a random departure from the linear model.

Figure 6



The plot of residuals versus x for this same set of data is displayed in Figure 7.

Figure 7



Clearly the pattern of residuals is not random! The linear model in Figure 6 has a larger r^2 and gives better estimates of the population slope and intercept than the fit in Figure 4. Despite this, the linear model is not really adequate! There are additional variations in y that a truly adequate model should explain.

IV. Summary of Formulas

First from the set of data points $(x_i, y_i), 1 \leq i \leq n$, calculate the following five sums.

$$\sum_{i=1}^n x_i \quad \sum_{i=1}^n y_i \quad \sum_{i=1}^n x_i y_i \quad \sum_{i=1}^n x_i^2 \quad \sum_{i=1}^n y_i^2$$

Table 3

Sample Mean of x	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
Sample Mean of y	$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$
Variation in x	$SS_{xx} = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}$
Variation in y	$SS_{yy} = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}$
The Covariation in xy	$SS_{xy} = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}$
Sample Linear Regression Slope	$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$
Sample Linear Regression Intercept	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
Error Variation	$SSE = SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}$
Standard Error of Estimate	$S_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}}{n-2}}$
Explained Variation	$SS_{ex} = \frac{SS_{xy}^2}{SS_{xx}}$
Standard Error of the Slope	$S_{\hat{\beta}_1} = \frac{S_e}{\sqrt{SS_{xx}}}$

<p>Standard Error of the Intercept</p>	$S_{\hat{\beta}_0} = S_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}}$
<p>Standard Error of a Predicted y Mean</p>	$S_{\mu_{y x_0}} = S_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$
<p>Standard Error of a Predicted Future Observation</p>	$S_{\hat{y} x_0} = S_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$
<p>All 100(1-α)% Confidence Intervals for a Regression Parameter are Computed the Same Way</p>	<p><i>U</i> is the Parameter, \hat{U} is the Sample Estimate, $S_{\hat{U}}$ is the Standard Error of the Sample Estimate, the Critical two-tail t score has <i>n</i> – 2 Degrees of Freedom</p> $U = \hat{U} \pm t_{\alpha/2} S_{\hat{U}}$
<p>All Two Tail Hypothesis Tests are Performed the Same Way</p>	$H_0 : U = U_0 ; H_1 : U \neq U_0 ; t = \frac{ U_0 - \hat{U} }{S_{\hat{U}}}$ <p>If the observed $t > t_{\alpha/2}$ with <i>n</i> – 2 Degrees of Freedom, the Null Hypothesis is Rejected.</p>
<p>All Right Tail Hypothesis Tests are Performed the Same Way</p>	$H_0 : U > U_0 ; H_1 : U \leq U_0 ; t = \frac{\hat{U} - U_0}{S_{\hat{U}}}$ <p>If the observed $t > t_{\alpha}$ with <i>n</i> – 2 Degrees of Freedom, the Null Hypothesis is Rejected.</p>
<p>All Left Tail Hypothesis Tests are Performed the Same Way</p>	$H_0 : U < U_0 ; H_1 : U \geq U_0 ; t = \frac{U_0 - \hat{U}}{S_{\hat{U}}}$ <p>If the observed $t > t_{\alpha}$ with <i>n</i> – 2 Degrees of Freedom, the Null Hypothesis is Rejected.</p>
<p>Coefficient of Determination</p>	$r^2 = \frac{SS_{xy}^2}{SS_{xx} SS_{yy}}$
<p>Correlation Coefficient</p>	$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$