HANDOUT #14: SIMPLE LINEAR REGRESSION

1. Model for Regression
   (a) Modeling Mean of Response Variable Y as function of Explanatory Variable X: $\mu_{Y|X} = \beta_o + \beta_1 X$
   (b) Distribution of Y given $X = x$: $N(\mu_{Y|X}, \sigma^2)$
   (c) Data: $(X_i, Y_i), i = 1, \ldots, n$ independent pairs
2. Least Squares Estimators of β_o, β_1, and σ
3. Properties of LSE
4. Using Residuals: $e_i = Y_i - \hat{\beta}_o - \hat{\beta}_1 X_i$ to evaluated validity of regression model conditions
5. Prediction and Confidence Intervals
6. ANOVA, R^2, Lack of Fit
7. Correlation: Pearson and Spearman
8. Problems in the Use of Regression Analysis

Supplemental Reading

- Chapters 10 & 11 in Tamhane/Dunlop book

Simple Linear Regression

(Straight-Line Regression)

In many studies the researcher is interested in studying the relationship between two or more variables. These variables have a non-deterministic relationship which involves a degree of uncertainty in the functional relationship between model parameters and the random variation of population values about the functional relationship. In order to understand the modeling, we will first consider a deterministic relationship:

Deterministic Relationship: Two variables x and y have a deterministic relationship if knowing the value of x completely specifies the value of y. This relationship may be expressed as

$$y = h(x)$$

Example: Suppose the cost of renting an apartment is a \$1000 deposit with a \$750 monthly payment. The total cost y for renting the apartment for x months is

$$y = 1000 + 750x$$

Non-Deterministic Relationship: A nondeterministic relationship or random relationship involves a degree of uncertainty between the values of x and y. Typically, the model will involve both an underlying deterministic component and a random component which reflects that for each value of x there is a population of y values with the many populations related through $h(x)$:

$$y = h(x) + \epsilon$$

where ϵ has a specified distribution which may involve unknown parameters.

Example: Suppose in our previous example, y is the total cost of renting an apartment but x is the number of months in a lease agreement. We want to relate total cost (y) to months (x) in lease agreement for all apartments in Houston. Certainly, y is related to x in some relationship but for a fixed value of x, say, $x =$ 1-year, there would be a wide variation in the total cost y depending on the size of the apartment and its location in Houston.
Regression analysis involves a systematic approach to determining *reasonable* approximations to both $h(\cdot)$ and the distribution of ϵ.

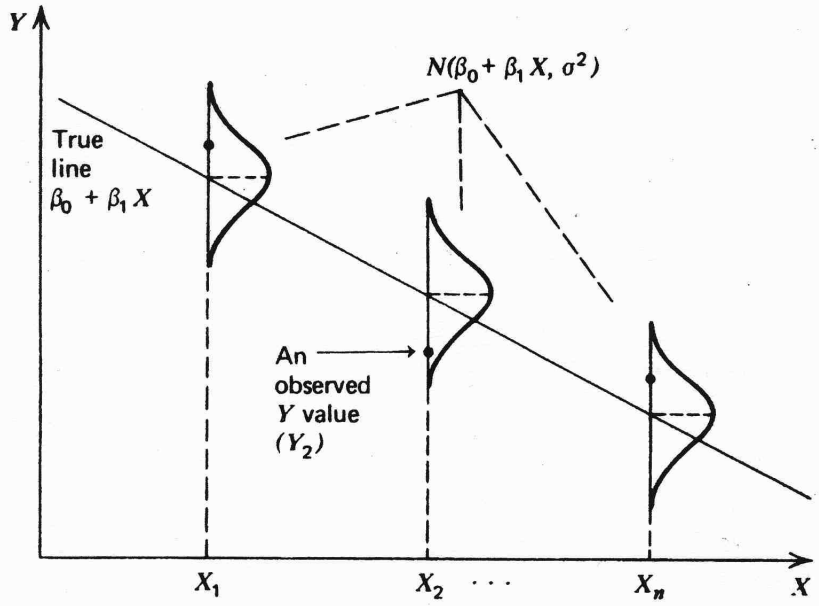
The simplest of these models species a straight-line for the deterministic component, i.e., $h(x) = \beta_o + \beta_1 x$ and a $N(0, \sigma^2)$ distribution for ϵ:

$$\text{Simple Linear Regression Model (SLR):} \quad y = \beta_o + \beta_1 x + \epsilon$$

Implications of this model: For each value of x, there is a population of $y-$values having a $N(\mu_{y|x}, \sigma^2_{y|x})$ distribution where

$$\mu_{y|x} = \beta_o + \beta_1 x : \quad \text{the mean of } y \text{ depends on } x \text{ through the straight-line: } \beta_o + \beta_1 x$$

$$\text{BUT} \quad \sigma^2_{y|x} = \sigma^2 : \quad \text{the variance of } y \text{ is constant for all values of } x$$



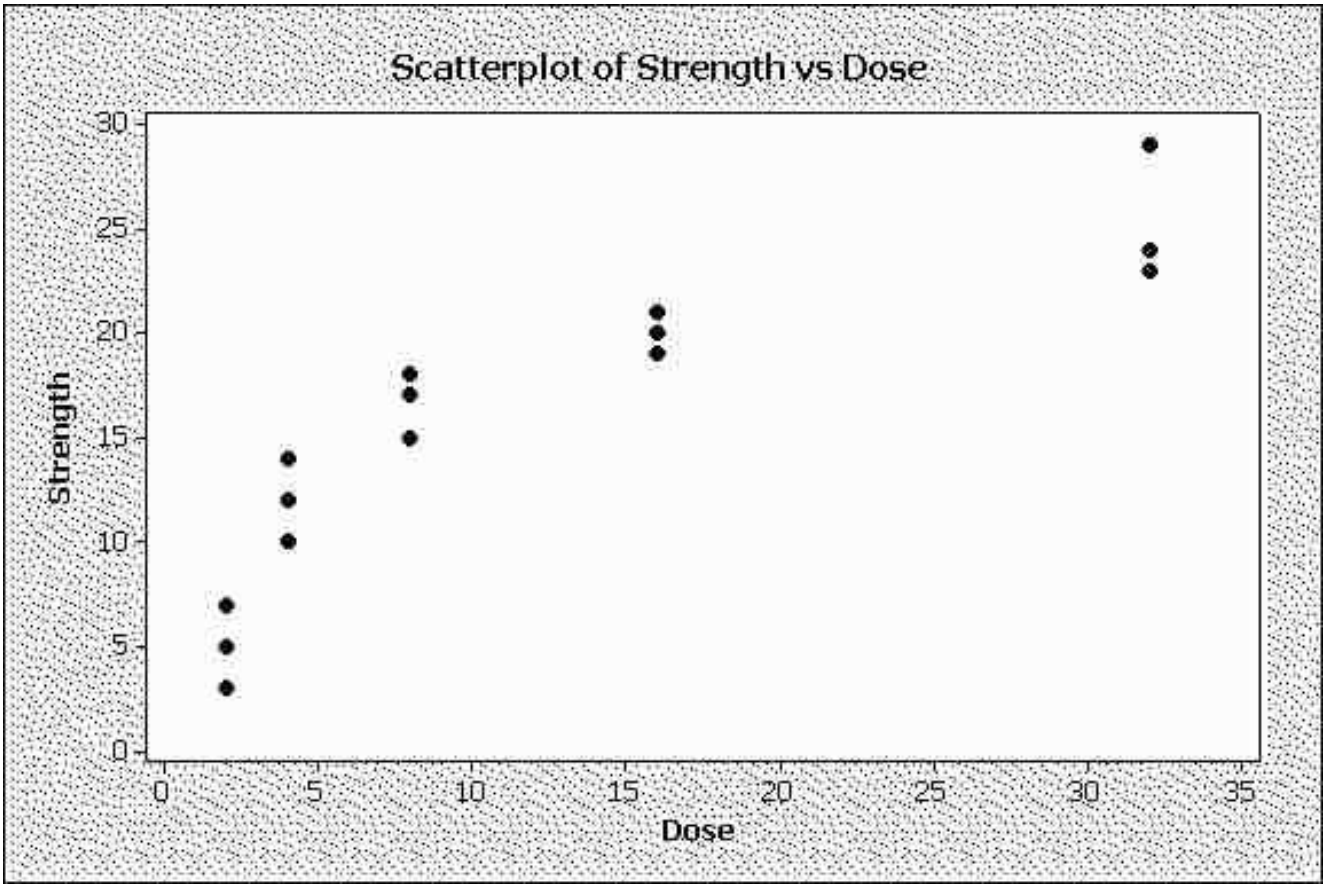
Each response observation is assumed to come from a normal distribution centered vertically at the level implied by the assumed model. The variance of each normal distribution is assumed to be the same, σ^2.

Case Study: What Dose Level is Needed?

A pharmaceutical firm is investigating the relationship betweeen the dose level of a drug and its potency. A study is designed in which 15 test tubes are inoculated with a virus culture and then incubated for 5 days at $30°C$. Three test tubes are randomly assigned to each of the 5 different dose levels to be investigated (0, 4, 8, 16, and 32 milligrams). Each test tube was injected with a single dose level of the drug. The response, y, a measure of the protective strength of the drug against the virus culture was obtained for each of the 15 test tubes. The data is given here.

Tube	Dose	Response	Tube	Dose	Response	Tube	Dose	Response
1	2	5	6	4	14	11	16	21
2	2	7	7	8	15	12	16	19
3	2	3	8	8	17	13	32	23
4	4	10	9	8	18	14	32	24
5	4	12	10	16	20	15	32	29

The following plot of the data reveals a general increasing in the protective strength as the dose level increases.



From the data we can observe that the data does not perfectly fit a straight line, however, it is "reasonably" close. We will next write a model to describe the relationship between the dose level applied to the test tube and the strength of the drug's response. Let y_i, be the ith observation on the **Dependent Variable**, response observed from the ith test tube. Let x_i be the ith value of the **Independent Variable**, the dose level placed in the ith test tube. The SLR model will relate y_i to x_i: For each value of dose x, the strength of drug y is a random variable having

C_1: Normal Distribution

C_2: Standard Deviation, $\sigma_{y|x} = \sigma$, the variation in strength y is the same for all dose levels x (constant variance condition).

C_3: Mean value, $\mu_{y|x} = \beta_o + \beta_1 x$, that is, the mean strength of the drug y is related to the dose level x by a straight-line: $\beta_o + \beta_1 x_i$

C_4: The n pairs of dose-response values, $(x_i, y_i), i = 1, \ldots, n$ are independent.

The above conditions are stating that the distribution of the dependent variable y is a normal distribution which is related to the independent variable x only through the mean value of y and not through its standard deviation.

That is,

$$Y_i = \beta_o + \beta_1 x_i + \epsilon_i, \quad \text{for} \quad i = 1, \ldots, n.$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent normal variables with $\mu_\epsilon = 0$ and σ_ϵ the same value of all i. The variable ϵ_i describes the variation of the dependent variable y about the regression line, $\mu_{Y|x} = \beta_o + \beta_1 x$. For our experiment, the values of ϵ describe the size of the deviation of a particular test tubes strength measurement from the mean strength of all possible test tubes having the **same dose level**.

In order to estimate the regression line that **best** fits the data, we use a technique called the **least squared error** fit of a line to the data. Goal: Find the line

$$y = \beta_o + \beta_1 x.$$

which minimizes the deviation between observed y_i and the value obtained by evaluating the line at x_i: $\hat{y}_i = \beta_o + \beta_1 x_i$. Then the error in the estimation, $e_i = y_i - \hat{y}_i$, is called the ith residual. We want to determine the values of β_o and β_1 which make the **sum of squares residuals**, SSE, as small as possible.

$$Q(\beta_o, \beta_1) = SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2 = \sum_{i=1}^{n} [y_i - \hat{\beta}_o - \hat{\beta}_1 x_i]^2$$

By differentiation of SSE with respect to β_o and β_1 we obtain

$$\frac{\partial Q}{\partial \beta_o} = -2 \sum_{i=1}^{n} [y_i - \hat{\beta}_o - \hat{\beta}_1 x_i]$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i [y_i - \hat{\beta}_o - \hat{\beta}_1 x_i]$$

Setting the partial derivates to zero, we obtain the *normal equations:*

$$(n)\hat{\beta}_o + \Big(\sum_{i=1}^{n} x_i \Big) \hat{\beta}_1 = \sum_{i=1}^{n} y_i$$

$$\Big(\sum_{i=1}^{n} x_i \Big) \hat{\beta}_o + \Big(\sum_{i=1}^{n} x_i^2 \Big) \hat{\beta}_1 = \sum_{i=1}^{n} x_i y_i$$

The solutions of the two equations yield the values of $\hat{\beta}_o$, and $\hat{\beta}_1$ which minimize the sum of squares residuals:

Least Squares Estimates(LSE): $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$ $\hat{\beta}_o = \bar{y} - \hat{\beta}_1 \bar{x}$

where

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

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

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

Furthermore, an estimator of σ is obtained by noting that $\mu_{y|x} = \beta_o + \beta_1 x$ which yields $\hat{\mu}_{y|x} = \hat{\beta}_o + \hat{\beta}_1 x$. Therefore, the estimated standard deviation of the y_i's is obtained from

$$\hat{\sigma}_\epsilon = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{\mu}_{y|x})^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}, \qquad \text{where} \qquad \text{SSE} = \sum_{i=1}^n (y_i - \hat{\mu}_{y|x})^2 = \sum_{i=1}^n (y_i - (\hat{\beta}_o + \hat{\beta}_1 x_i))^2$$

Properties of LSE and $\hat{\sigma}$

1. $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators:

$$\begin{aligned}
E[\hat{\beta}_1] &= E\left[\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{SS_{xx}}\right] \\
&= \frac{\sum(x_i - \bar{x})E[y_i - \bar{y}]}{SS_{xx}} \\
&= \frac{\sum(x_i - \bar{x})E[\beta_o + \beta_1 x_i + \epsilon_i - \beta_o - \beta_1 \bar{x} - \bar{\epsilon}]}{SS_{xx}} \\
&= \frac{\sum(x_i - \bar{x})(\beta_o + \beta_1 x_i + E[\epsilon_i]) - \beta_o - \beta_1 \bar{x} - E[\bar{\epsilon}])}{SS_{xx}} \\
&= \frac{\sum(x_i - \bar{x})(\beta_1 x_i - \beta_1 \bar{x})}{SS_{xx}} = \beta_1 \\
E[\hat{\beta}_0] &= E[\bar{y}] - E\left[\hat{\beta}_1 \bar{x}\right] \\
&= E[\beta_o + \beta_1 \bar{x} + \bar{\epsilon}] - E\left[\hat{\beta}_1 \bar{x}\right] \\
&= \beta_o + \beta_1 \bar{x} + E[\bar{\epsilon}] - E\left[\hat{\beta}_1\right] \bar{x} \\
&= \beta_o + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_o
\end{aligned}$$

2. $Var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}}{SS_{xx}}\right]$ and $Var(\hat{\beta}_1) = \frac{\sigma^2}{SS_{xx}}$

   In deriving the variance of $\hat{\beta}_1$, note that

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{SS_{xx}} = \frac{(\sum(x_i - \bar{x})y_i}{SS_{xx}}$$

$$\begin{aligned}
Var\left[\hat{\beta}_1\right] &= \frac{\sum(x_i - \bar{x})^2 Var[y_i]}{(SS_{xx})^2} \\
&= \frac{\sum(x_i - \bar{x})^2 (Var[\epsilon_i]}{(SS_{xx})^2} \\
&= \frac{\sigma^2 \sum(x_i - \bar{x})^2}{(SS_{xx})^2} = \frac{\sigma^2}{SS_{xx}}
\end{aligned}$$

In deriving the variance of $\hat{\beta}_0$, note that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \sum \left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SS_{xx}} \right] y_i$$

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

3. Sampling Distribution of $\hat{\beta}_o$ and $\hat{\beta}_1$

   Using

$$\hat{\beta}_0 = \sum \left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SS_{xx}} \right] y_i = \sum_{i=1}^{n} a_i y_i$$

   and

$$\hat{\beta}_1 = \sum \frac{x_i - \bar{x}}{SS_{xx}} y_i = \sum_{i=1}^{n} b_i y_i$$

   it is seen that $\hat{\beta}_o$ and $\hat{\beta}_1$ are linear combinations of independent normal random variables. Therefore, both $\hat{\beta}_o$ and $\hat{\beta}_1$ are normally distributed with means and variances given on the previous page. However, $\hat{\beta}_o$ and $\hat{\beta}_1$ are not independent because

$$\begin{aligned} Cov(\hat{\beta}_o, \hat{\beta}_1) &= Cov\left(\sum_{i=1}^{n} a_i y_i, \sum_{i=1}^{n} b_i y_i \right) \\ &= \sum_{i=1}^{n} a_i b_i Var(y_i) \\ &= \sigma^2 \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SS_{xx}} \right] \left[\frac{x_i - \bar{x}}{SS_{xx}} \right] \\ &= \sigma^2 \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{nSS_{xx}} \right] - \sum_{i=1}^{n} \left[\frac{\bar{x}(x_i - \bar{x})^2}{(SS_{xx})^2} \right] \\ &= -\sigma^2 \sum_{i=1}^{n} \frac{\bar{x}}{SS_{xx}} \end{aligned}$$

4. Using $\hat{\sigma}^2 = \frac{1}{n-2}SSE$, the estimated standard errors of $\hat{\beta}_o$ and $\hat{\beta}_1$ are given by

$$\hat{SE}(\hat{\beta}_o) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}}$$

$$\hat{SE}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SS_{xx}}}$$

5. It can be shown that $\frac{(n-2)\hat{\sigma}^2}{\sigma^2}$ has a Chi-squared distribution with df=n-2 and is distributed independently of both $\hat{\beta}_o$ and $\hat{\beta}_1$. Therefore, we can use the t-distribution in tests of hypotheses and confidence intervals of $\mu_{y|x}$, β_o and β_1.

6. The fitted residuals $e_i = y_i - \hat{y}_i$ are used to assess whether the required model conditions are satisfied in a given experiment. When the model is correct, the residuals have the following properties:

$$e_i's \quad \text{are normally distributed with} \quad E[e_i] = 0, \quad Var[e_i] = \sigma^2\left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{SS_{xx}}\right] \approx \sigma^2$$

for reasonably large n. Thus, we can evaluate the model conditions by examining whether the e_i's appear to be normally distributed with mean 0 and essentially constant variance.

We will now apply these formulas to the dose-strength of response data.

$$SS_{xx} = 1785.6,$$

$$SS_{yy} = 764.4,$$

$$SS_{xy} = 1027.2$$

Thus

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{1027.2}{1785.6} = .5753$$

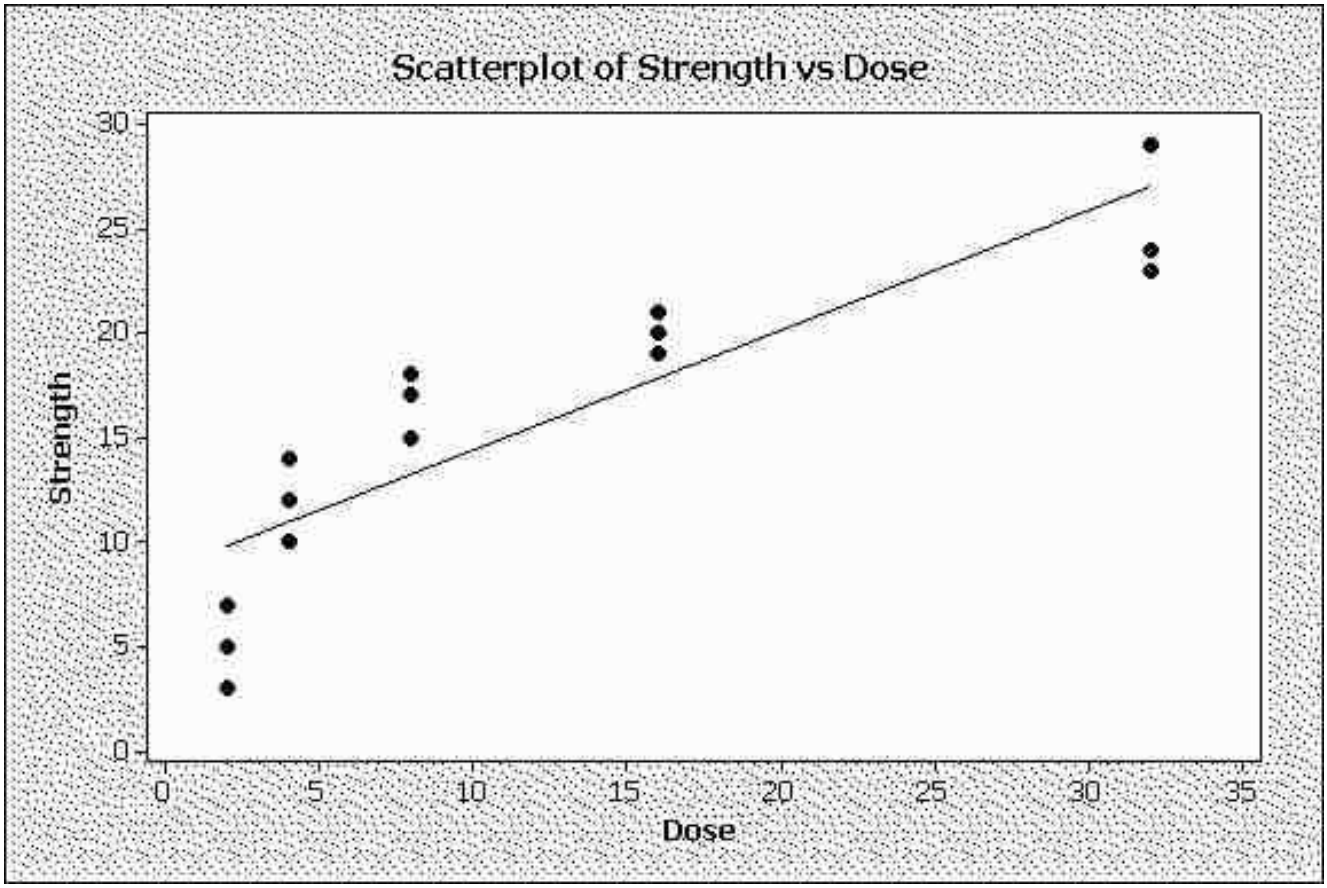
$$\bar{y} = \frac{237}{15} = 15.8 \quad \text{and} \quad \bar{x} = \frac{186}{15} = 12.4 \quad \text{thus}$$

$$\hat{\beta}_o = \bar{y} - \hat{\beta}_1\bar{x} = 15.8 - (.5753)(12.4) = 8.67$$

$$\text{SSE} = SS_{yy} - \beta_1 SS_{xy} = 764.4 - (.5753)(1027.2) = 173.45$$

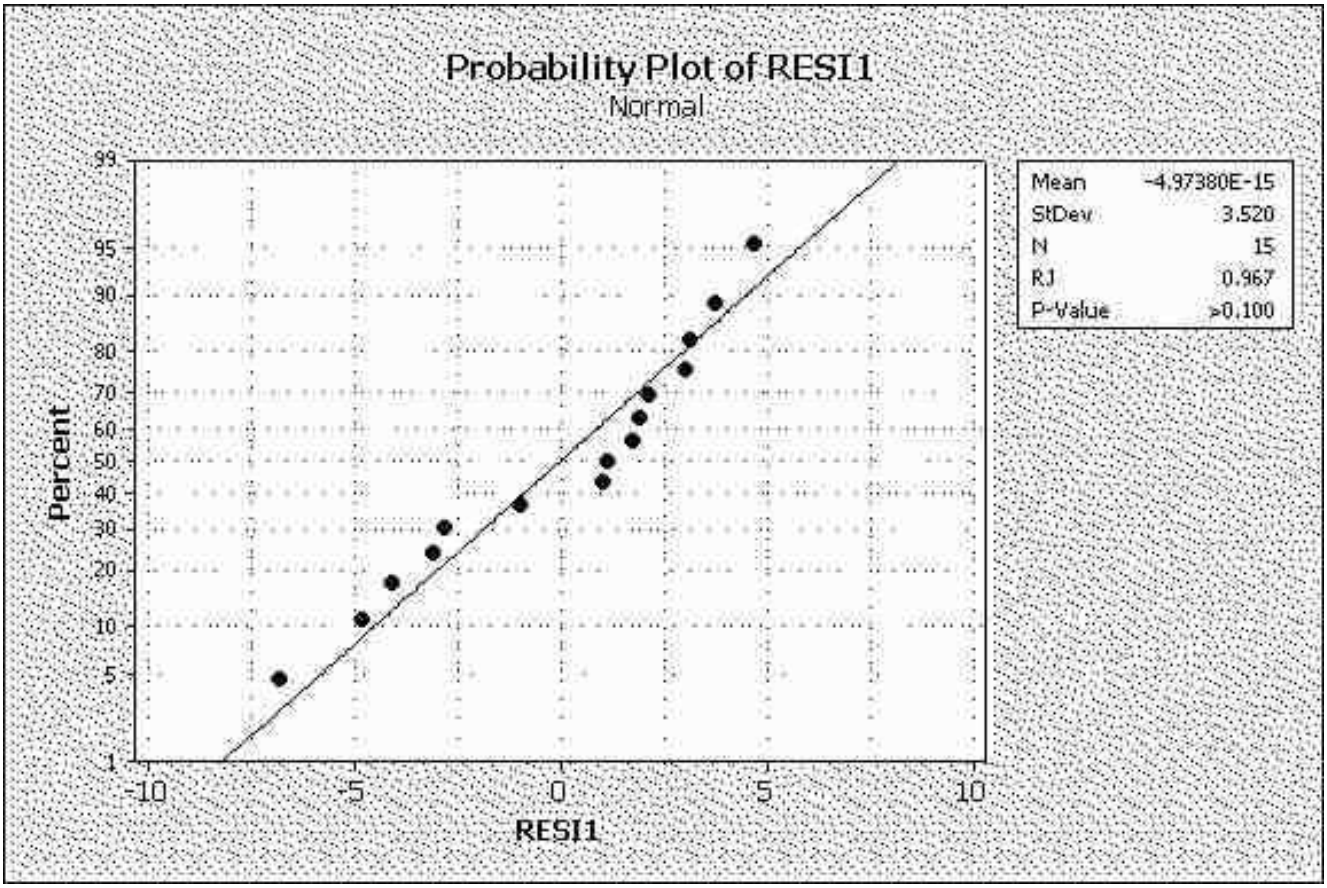
$$\hat{\sigma}_\epsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{173.45}{15-2}} = 3.653$$

We can next plot the **least squares line**, $\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x = 8.67 + .5753x$ on the scatter-plot of the 15 data values (x_i, y_i) and observe how well the least squares line fits the data values.



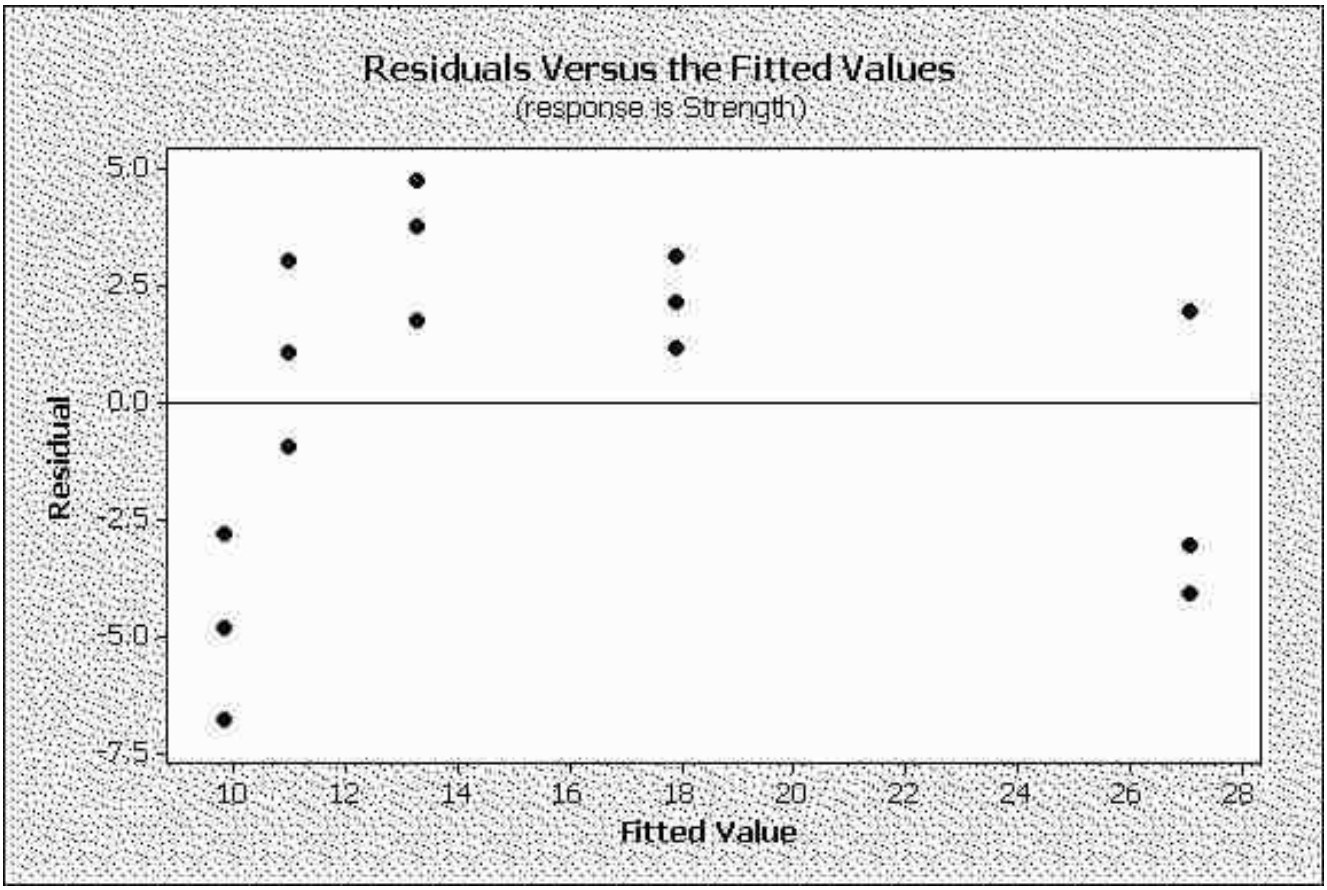
The predicted values of the strength variable y are $\hat{y}_i = \hat{\beta}_o + \hat{\beta}_1 x_i = 8.67 + .5753 x_i$ with computed residuals $e_i = y_i - \hat{y}_i$. We will use these values in assessing whether the conditions imposed by our model are satisfied by the particular experimental conditions and the manner in which the experiment was conducted.

C_1 : Normality of the Population Distributions.



The plotted points are close to the straight-line and the p-value=.169 which would indicate that the null hypothesis of Normality is NOT rejected.

C_2 : Constant Variance across the values of the independent varible x, the dose levels



The spread in the values of the residuals e_i remains fairly constant as the fitted values are increased. This would indicate that the constant variance assumption has been satisfied.

C_3 : Mean value, $\mu_{Y|x} = \beta_o + \beta_1 x$

We can evaluate this condition with the above plot. If the straight-line was an appropriate model, the residuals should be randomly scattered about an horizontal line through 0. The above plot indicates that the straight-line may not be the best possible model for this experiment.

Evaluation of Fit of the Model:

The above plots give us a graphical depiction of how well our model fits the data from the experiments. The following analysis will quantify the degree to which the model fits the data.

ANOVA for Regression

Similarly to the AOV for an experiment with a treatment variable, we can compute the SS's for a regression model.

$$SS_{TOT} = SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

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

$$SSE = SS_{Error} = SS_{Resid} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS_{TOT} - SS_{Reg}$$

For the dose-strength data, we have

$SS_{TOT} = 764.40$ $SS_{Reg} = (1027.2)^2/1785.6 = 590.92$ $SS_{Error} = 764.4-590.92 = 173.48$

We can summarize these calculations in an AOV table:

Source	df	SS	MS	F	p-value
Model	1	590.92	590.92	44.28	.000016
Error	15-2	173.48	13.34		
Model	15-1	764.40			

The F-ratio is the test statistic for testing:

$$H_o : \beta_1 = 0 \qquad \text{versus} \qquad H_a : \beta_1 \neq 0$$

Since the p-value = .000016, we have very significant evidence that β_1 is not zero and thus the linear relationship between the dose level of the drug and the strength of the response is present. This merely informs us that a straight-line relationship is a possible model between dose level and strength. To further investigate the relationship, we want to compute how much of the variation in the strength values are *explained* by the straight-line model. We compute the **coefficient of determination**, R^2 which is defined as follows:

$$R^2 = 1 - \frac{SSE}{SS_{TOT}} = \frac{SS_{Reg}}{SS_{TOT}} = Cov(y_i, \hat{y}_i)$$

Since $SS_{TOT} = SS_{Reg} + SSE$, we can think of R^2 as the proportion of the variation in the dependent variable **explained** by the regression model. In our example, we have

$$R^2 = 1 - \frac{173.48}{764.40} = 1 - .227 = .773 = 77.3\%$$

Thus, we would state that the straight-line model explains 77.3% of the variation in the 15 strength of response readings observed in the experiment. The remaining 22.7% of the variation could be due to such sources as

1. A more complex model relating Strength to Dose Level is needed.

2. There are variables other than Dose Level which affect Strength, such as, the conditions in the laboratory, the amount of virus in the test tubes, variations in the production of the drug from dose to dose, measurement errors in recording the value of Strength during the experiment, etc.

If we in fact were to fit a more complex model, the value of R^2 would generally increase somewhat, but the factors mentioned above would prevent the value of R^2 from becoming 1.

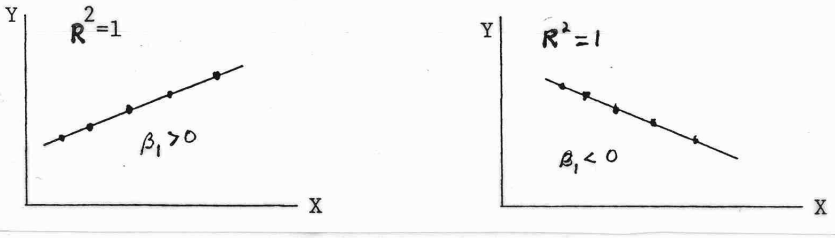
Note that $0 \leq R^2 \leq 1$ with its limits characterized by

1. $R^2 = 0$ corresponding to $SSE = SS_{TOT}$. This would indicate that the model does not explain any of the variation in the dependent variable, i.e., $\beta_1 = 0$. As we change the value of the independent variable, the value of the dependent variable varies in a purely random fashion.

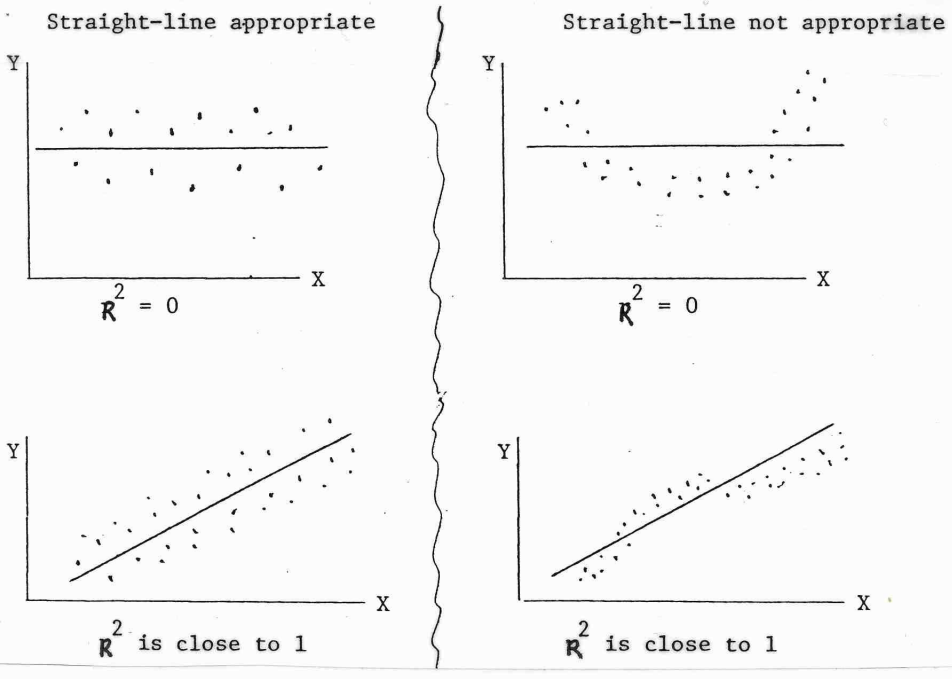
2. $R^2 = 1$ corresponding to $SSE = 0$. This would indicate that all the residuals, $e_i = 0$. Thus, all n data values fall perfectly on the straight-line.

The following plots will illustrate what R^2 does not measure:

1. R^2 does not measure the magnitude of the slope in the regression line:



2. R^2 does not measure whether a straight-line model is the most appropriate model:



After verifying that the required conditions for the model have been satisfied, that is, the normality and equal variance of the residuals, and straight-line relationship in the means, we can develop the following confidence intervals.

1. A $100(1-\alpha)\%$ confidence interval for the slope β_1 is

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\frac{\hat{\sigma}_\epsilon}{\sqrt{SS_{xx}}} \qquad \text{where} \qquad \hat{\sigma}_\epsilon = \sqrt{MSE}$$

   For our example, we have a 95% c.i. for β_1 is

$$.575 \pm 2.160\frac{\sqrt{13.34}}{\sqrt{1785.6}} \quad \text{which yields} \quad .575 \pm .127 \qquad \text{which yields} \quad (.45, .70)$$

2. A $100(1-\alpha)\%$ confidence interval for the intercept β_o is

$$\hat{\beta}_o \pm t_{\alpha/2,n-2}\hat{\sigma}_\epsilon\sqrt{\frac{\sum x_i^2}{nSS_{xx}}}$$

   For our example, we have a 95% c.i. for β_o is

$$8.67 \pm 2.160\sqrt{13.34}\sqrt{\frac{4092}{(15)(1785.6)}} \qquad \text{which yields} \qquad 8.67 \pm 3.08 \qquad \text{which yields} \quad (5.53, 11.69)$$

3. To test the hypotheses: $H_o : \beta_1 \leq C \quad \text{vs} \quad H_a : \beta_1 > C$, we can use the test statistic:

$$T.S. = \frac{\hat{\beta}_1 - C}{\hat{\sigma}_\epsilon/\sqrt{SS_{xx}}} \quad \text{reject} \quad H_o \quad \text{if} \quad T.S. \geq t_{\alpha,n-2}$$

   For our example, suppose we want to test $H_o : \beta_1 \leq .5 \quad \text{vs} H_a : \beta_1 > .5$,

$$T.S. = \frac{.575 - .5}{\sqrt{13.34}/\sqrt{1785.6}} \quad \text{reject} \quad H_o \quad \text{if} \quad T.S. \geq t_{.05,13} = 1.771$$

   Since T.S. = .87, which is not greater than 1.771 we fail to reject H_o, and conclude there is not sufficient evidence to support the research hypothesis that the slope is greater than .5.

   A test of the hypotheses: $H_o : \beta_1 = 0 \quad \text{vs} \quad H_a : \beta_1 \neq 0$, using the test statistic given above with C=0, i.e., reject H_o if $|T.S.| \geq t_{\alpha/2,n-2}$, is equivalent to using the F-test from the AOV table. This holds since when C=0, $(T.S)^2 = F$ and $(t_{\alpha/2,n-2})^2 = F_{\alpha,1,n-2}$.

4. A $100(1-\alpha)\%$ confidence interval for the mean value of y when $x = x_o$ is computed as follows: the point estimator of $\mu_{y|x_o}$ is $\hat{\mu}_{y|x_o} = \hat{\beta}_o + \hat{\beta}_1 x_o$ and the c.i. is given by

$$\hat{\mu}_{y|x_o} \pm t_{\alpha/2,n-2}\hat{\sigma}_\epsilon\sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_{xx}}}$$

For our example, a 95% c.i. for the mean strength of the drug when the dose level is $x_o = 5$ is given here. First, the point estimator of the mean is $\hat{\mu}_{y|5} = \hat{\beta}_o + \hat{\beta}_1 x_o = 8.67 + (.575)(5) = 11.545$. Next we compute the 95% confidence interval on $\mu_{y|5}$

$$11.545 \pm 2.160\sqrt{13.34}\sqrt{\frac{1}{15} + \frac{(5-12.4)^2}{1785.6}}$$

$$\text{which yields} \quad 11.545 \pm 2.461 \qquad \text{which yields} \quad (9.08, 14.01)$$

Thus, we are 95% confident that the mean strength of the drug is between 9.08 and 14.01 when a dose level of 5 milligrams is used.

5. A $100(1-\alpha)\%$ prediction interval for the value of y when $x = x_o$ is computed as follows: This interval is estimating the value that will result for y if one more experiment is run in which the value of x is x_o : the point estimator of y when $x = x_o$ is $\hat{y}_{x_o} = \hat{\beta}_o + \hat{\beta}_1 x_o$ and the prediction interval is given by

$$\hat{y}_{x_o} \pm t_{\alpha/2,n-2}\hat{\sigma}_\epsilon\sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_{xx}}}$$

For our example, a 95% prediction interval for the strength of the drug in a single test tube when the dose level is $x_o = 5$ is given here. First, the predicted strength when using a dose of 5 milligrams is $\hat{y}_5 = \hat{\beta}_o + \hat{\beta}_1 x_o = 8.67 + (.575)(5) = 11.545$. Next we compute the 95% prediction interval on $\hat{\mu}_{y|5}$

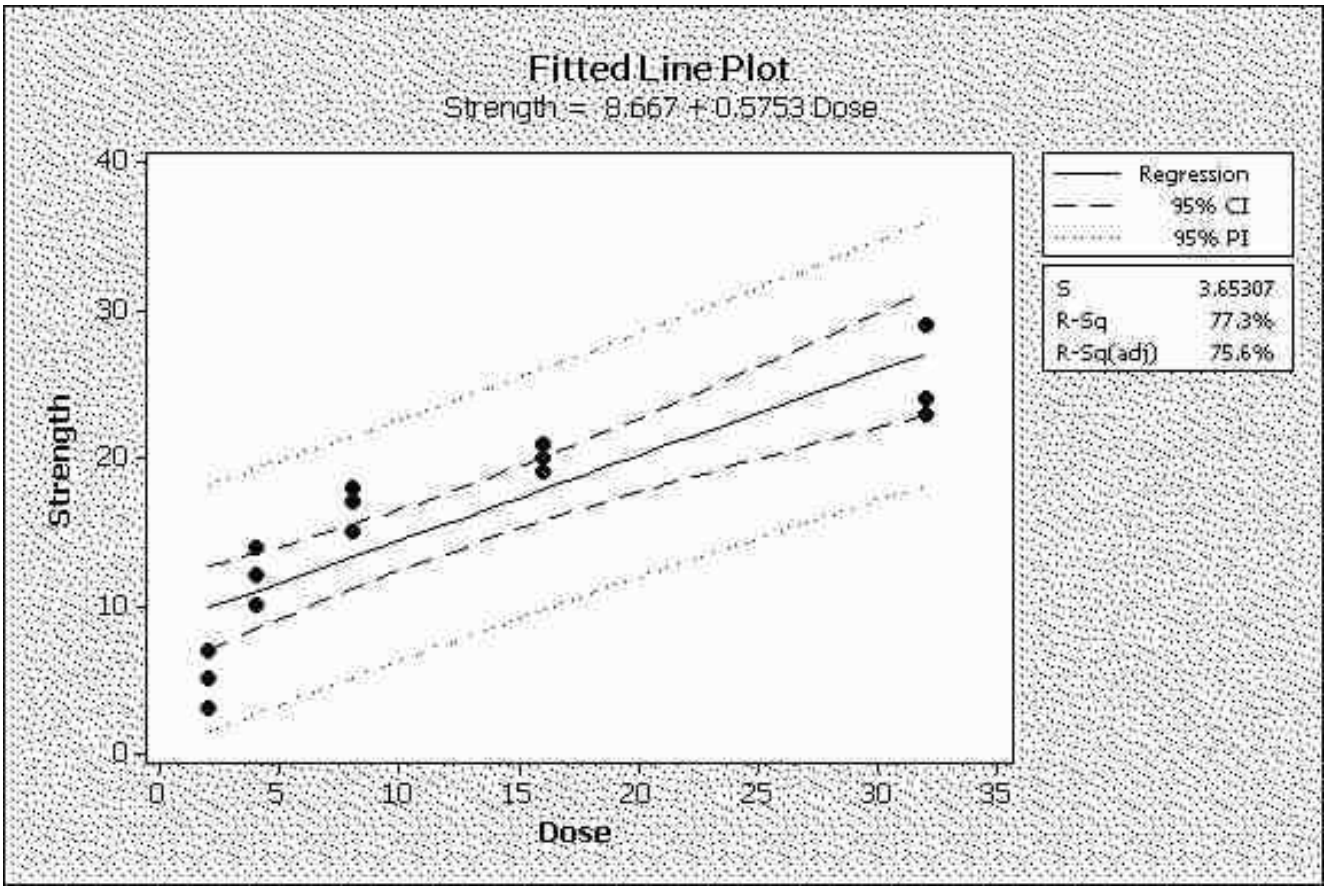
$$11.545 \pm 2.160\sqrt{13.34}\sqrt{1 + \frac{1}{15} + \frac{(5-12.4)^2}{1785.6}} \quad \text{which yields}$$

$$11.545 \pm 8.264 \qquad \text{which yields} \quad (3.28, 19.81)$$

Thus, we are 95% confident that if a dose level of 5 milligrams is applied to a test tube the resulting strength will be between 3.28 and 19.81.

Note that the prediction interval is considerably wider than the confidence interval since we are attempting to predict a single value from the population of strength values rather the mean of the population. Both the prediction interval and confidence interval are narrowest when $x_o = \bar{x}$ as can be seen in the following plot. This occurs since the mean is the center of the x-values and we have "equal" knowledge about the relationship between y and x on both sides of $\bar{x}$.

Prediction and Confidence Intervals Plot



6. We can also test hypotheses concerning the mean value of y when $x = x_o$. Suppose we want to test whether the mean strength of the drug is greater than 15 when x=20. We can use the following test statistic to test the hypotheses $H_o : \mu_{y|x_o} \leq C$ versus $H_a : \mu_{y|x_o} > C$

$$T.S. = \frac{\hat{\mu}_{y|x_o} - C}{\hat{\sigma}_\epsilon \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_{xx}}}} \qquad \text{reject} \quad H_o \quad \text{if} \qquad T.S. \geq t_{\alpha,n-2}$$

For example, to test $H_o : \mu_{y|20} \leq 15$ versus $H_a : \mu_{y|20} > 15$, we would compute $\hat{\mu}_{y|20} = 8.67 + (.575)(20) = 20.17$. Thus,

$$T.S. = \frac{20.17 - 15}{\sqrt{13.34}\sqrt{\frac{1}{15} + \frac{(20-12.4)^2}{1785.6}}} = 4.498 > t_{.05,13} = 1.771.$$

Thus, we reject H_o and conclude there is significant evidence (p-value=.0003) that the mean strength of the drug is greater than 15 when the dose level is 20 milligrams.

Lack-of-Fit-Test

If repeated observatios of y occur a several values of x_i, then it is possible to test the hypotheses:

$$H_o : \mu_{y|x} = \beta_o + \beta_1 x \quad \text{vs}$$

$$H_o : \mu_{y|x} \neq \beta_o + \beta_1 x$$

Suppose we have m distinct values of x and we observe n_i values of y at the ith distinct value of x, that is,

$$y_{1,1}, y_{1,2}, \cdots, y_{1,n_1} \quad \text{observations at} \quad x_1$$

$$y_{2,1}, y_{2,2}, \cdots, y_{2,n_2} \quad \text{observations at} \quad x_2$$

$$\cdots$$

$$y_{m,1}, y_{m,2}, \cdots, y_{m,n_m} \quad \text{observations at} \quad x_m$$

In the dose-strength case study, $m = 5$ and $n_i = 3$.
Define

$$SS_{PE} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \quad \text{and}$$

$$SS_{LOF} = SSE - SS_{PE}$$

Under H_o, the test statistics,

$$F = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)}$$

has an F-distribution with df=m-2 and n-m. We would thus reject H_o if $F \geq F_{\alpha,m-2,n-m}$.

In the drug example, $SSE = 173.45, \quad SS_{PE} = 43.3 \quad SS_{LOF} = 173.45 - 43.3 = 130.15$, this yields

$$F = \frac{130.15/(5-2)}{43.3/(15-5)} = 10.02 \Rightarrow P - value = 0.0023$$

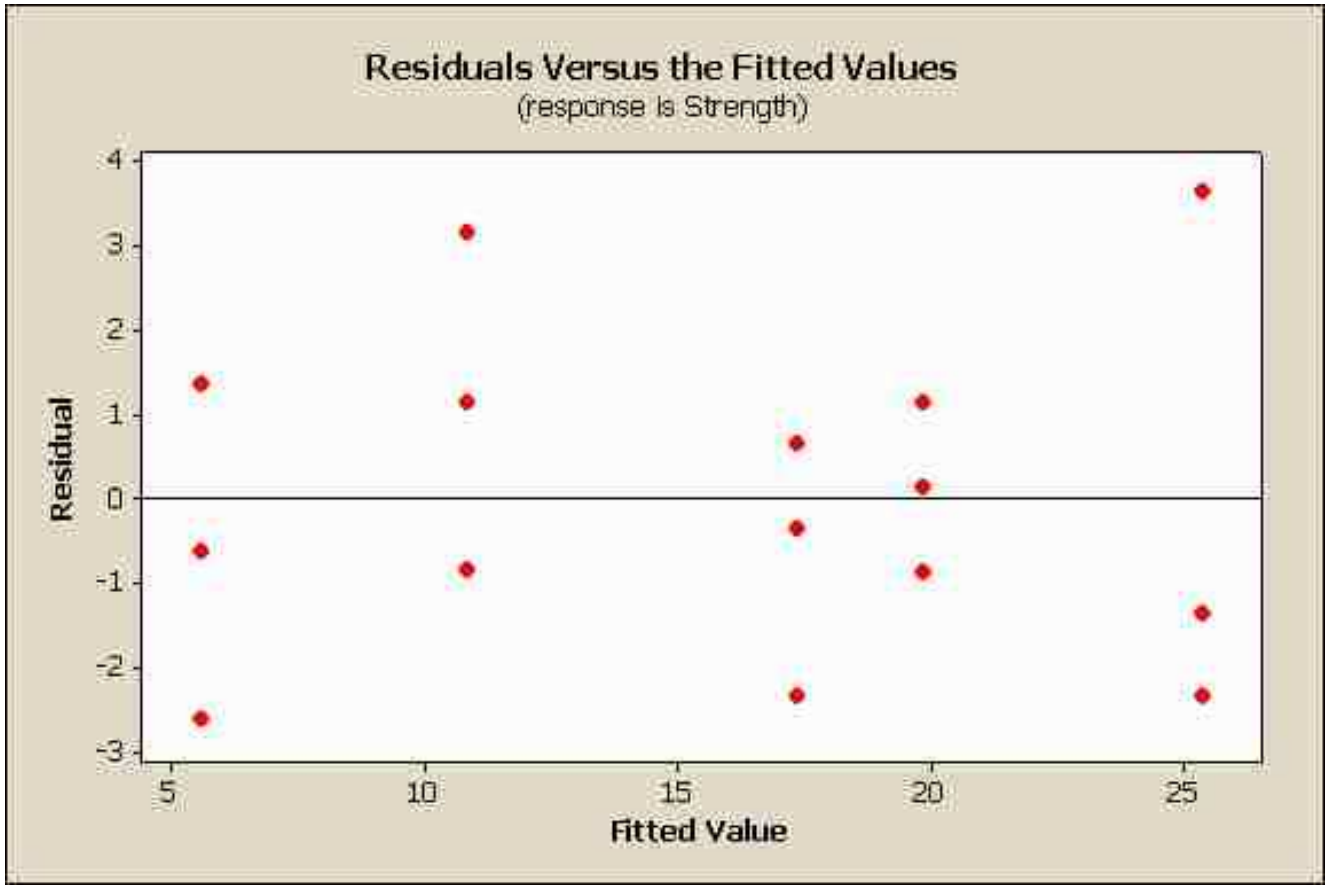
Thus, there is a strong rejection of H_o which would indicate that the straight-line model is not appropriate. In these situations we can often achieve a more precise fit between y and x by fitting a higher order polynomial in x. In our example, I refit the model using $y = \beta_o + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$ and obtained the following results.

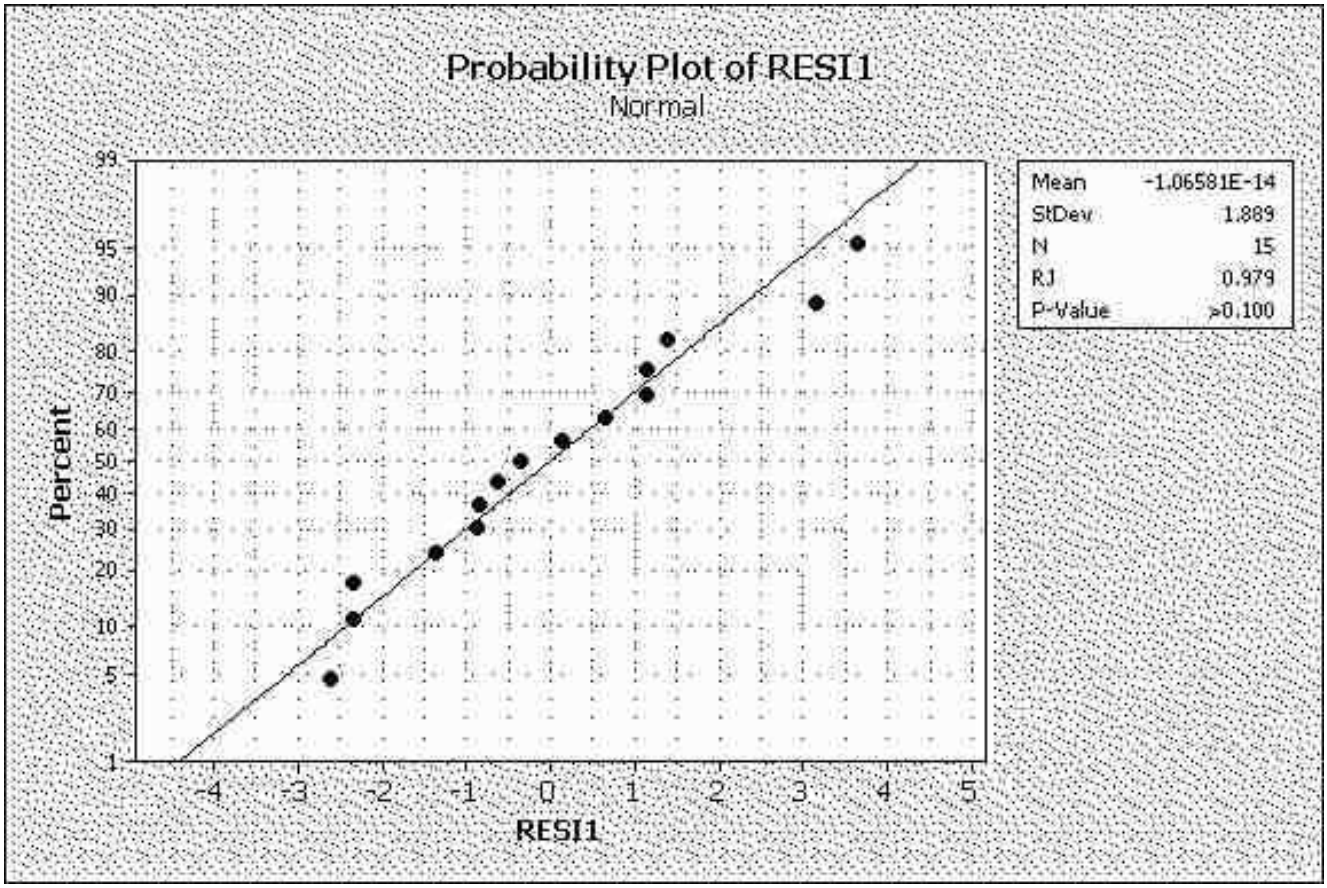
1. $\hat{y} = -1.175 + 3.820x - .220x^2 + .00395x^3$
2. p-value = .0006 for testing $H_o : \beta_1 = 0$

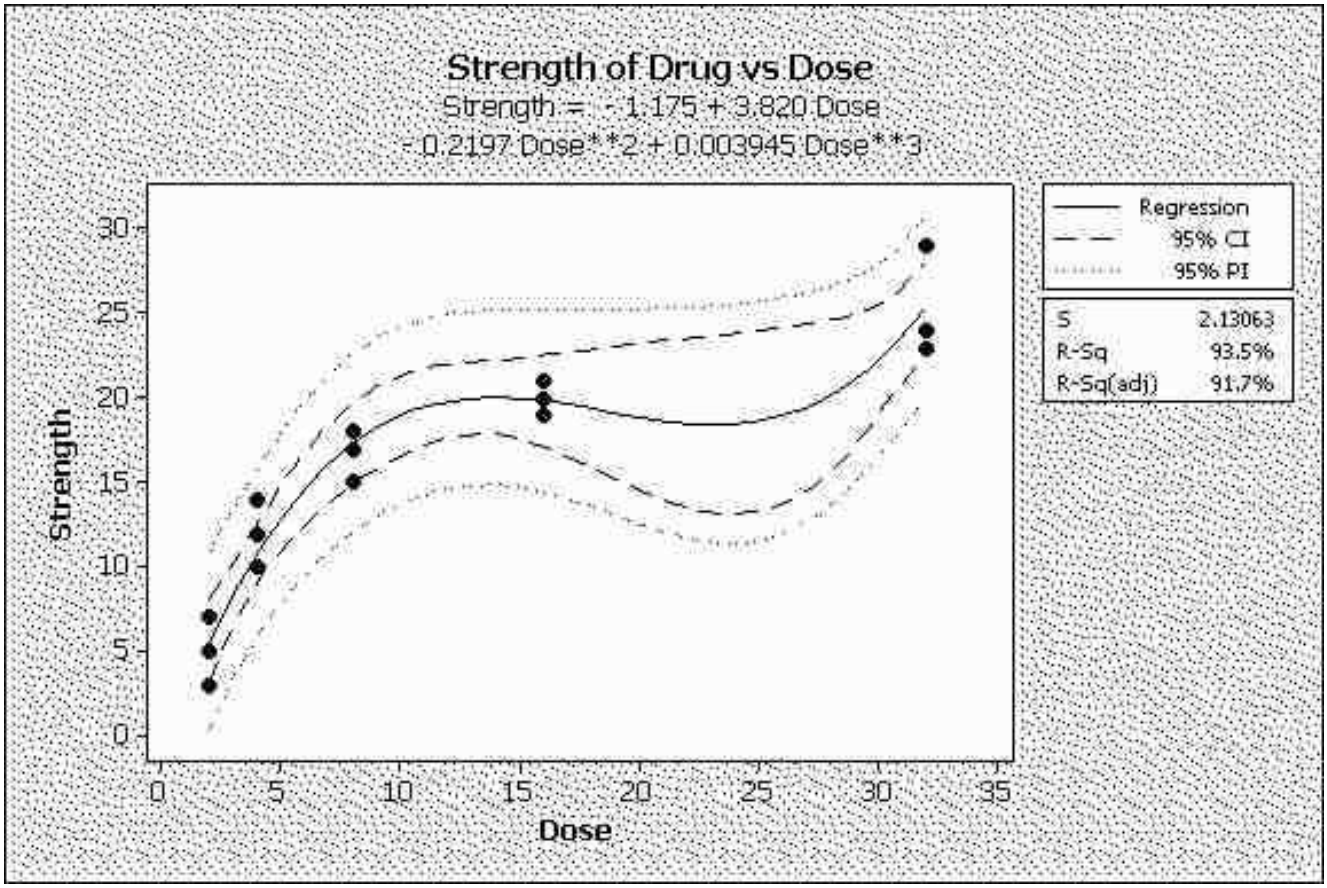
   p-value = .006 for testing $H_o : \beta_2 = 0$

   p-value = .012 for testing $H_o : \beta_3 = 0$
3. $R^2 = .935$ compared to $R^2 = .773$ in our previous model.
4. The plot of the Strength vs Dose shows a cubic relationship

5. The residual plots have the points closer to a horizonal line through 0




Probability Plot of RESI1
Normal
Percent
RESI1
Mean -1.06581E-14
StDev 1.889
N 15
RJ 0.979
P-Value >0.100


Strength of Drug vs Dose
Strength = - 1.175 + 3.820 Dose
- 0.2197 Dose**2 + 0.003945 Dose**3
Regression
95% CI
95% PI
S 2.13063
R-Sq 93.5%
R-Sq(adj) 91.7%
Strength
Dose
0
5
10
15
20
25
30
35

Correlation Coefficients

The covariance between two random variables, X and Y, measures the simultaneous dispersion of the variables about their means. The covariance is defined as

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)].$$

Positive values of covariance implies that values of X greater than μ_X occur when the values of Y tend to be larger than μ_Y, similarly, values of X less than μ_X occur when the values of Y tend to be less than μ_Y Negative values of covariance implies that values of X greater than μ_X occur when the values of Y tend to be less than μ_Y and values of X less than μ_X occur when the values of Y tend to be greater than μ_Y. The values of Cov(X,Y) have units which depend on the units of X and Y. Therefore, the values of Cov(X,Y) vary greatly not because of the strength of the association between X and Y but simply based on the selection of the units of measurements for X and Y. To alleviate this problem, a standardized form of covariance is defined by the correlation coefficient:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

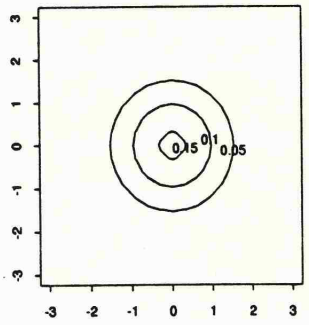
The correlation coefficient is a unit-free measure of association and hence yields the same value no matter what units are selected for X and Y. In fact, $-1 \leq \rho_{XY} \leq 1$, with $\rho_{XY} = \pm 1$ implying that $Y = \beta_o + \beta_1 X$. That is, there is a perfect linear relationship between X and Y.
If the population distribution is bivariate normal, then the marginal distributions of X and Y is normal with the joint pdf given by

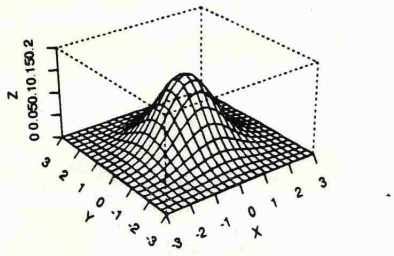
$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} exp\Big\{-\frac{1}{2(1-\rho^2)}\Big[\Big(\frac{x-\mu_X}{\sigma_X}\Big)^2 + \Big(\frac{y-\mu_Y}{\sigma_Y}\Big)^2 - 2\rho\Big(\frac{x-\mu_X}{\sigma_X}\Big)\Big(\frac{y-\mu_Y}{\sigma_Y}\Big)\Big]\Big\}$$

A plot of this pdf for several values of ρ is given on the next page.

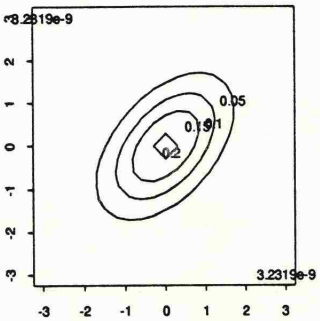
Bivariate Normal Density

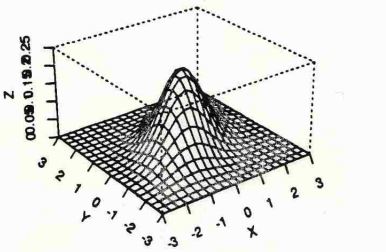
Bivariate Normal Distribution with Rho=0.0



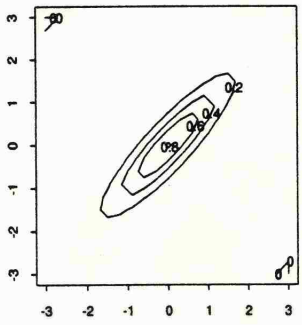


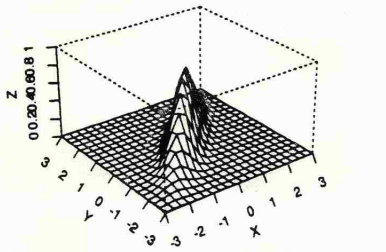
Bivariate Normal Distribution with Rho=0.50





Bivariate Normal Distribution with Rho=0.90





When ρ is unknown for a given population, we want to make inferences about ρ based on a random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ from a population which is represented by a pair of random variables (X,Y). For example, before-after data in a learning experiment:

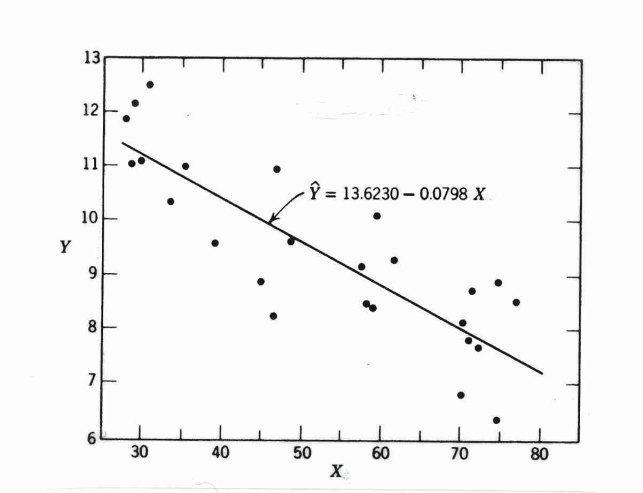
X=reading comprehension before training and Y=reading comprehension after training or two measures of the quality of a product:

X=tensile strength of a new allow and Y=surface corrosion resistance of alloy

In assessing the strength of the association between X and Y, the first step is to plot the n data values: $(X_1, Y_1), \cdots, (X_n, Y_n)$ in a scatterplot to display the relationship between X and Y.

Example: The energy consumption of a large corporation is under study. The data from 25 consecutive months are obtained from a steam plant at a large corporation. The corporation is interested in the association between X=Average monthly atmospheric temperature (°F) and Y=Amount of energy used monthly. The twenty-five values are plotted below.

Plot of Energy Use versus Atmospheric Temperature



Using the 25 data values we want to estimate the correlation between X and Y. The sample correlation coefficient (called **Pearson's product-moment correlation**) is given by

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}$$

We can also obtain r very easily from fitting a SLR equation to the n data values and noting that

$$r = \hat{\beta}_1 \frac{S_X}{S_Y}$$

Alternatively,

$$r = +\sqrt{R^2} \text{ if } \hat{\beta}_1 > 0 \text{ and } r = -\sqrt{R^2} \text{ if } \hat{\beta}_1 < 0$$

For the 25 data values, $\hat{\beta}_1 = -0.0798, R^2 = 0.7144$. Therefore,

$$r = -\sqrt{0.7144} = -0.845$$

This indicates a strong negative linear relationship between amount of energy used and the atmospheric temperature. Great care must be taken in the interpretation of the value of r. This is best illustrated by examining a number of scatterplots for various values of r given on the next page. Note there may be a very weak association between X and Y even when r appears to be quite large. Also, recall the plots from the SLR handout which illustrated situations where there was a strong association between Y and X but the association was not linear. Therefore, r may be very close to zero when in fact a strong non-linear relationship exists between X and Y.
A formal test of the hypotheses $H_o : \rho = 0$ vs $H_a : \rho \neq 0$ exists in the situation when the joint distribution of (X,Y) is bivariate normal. The test statistic is

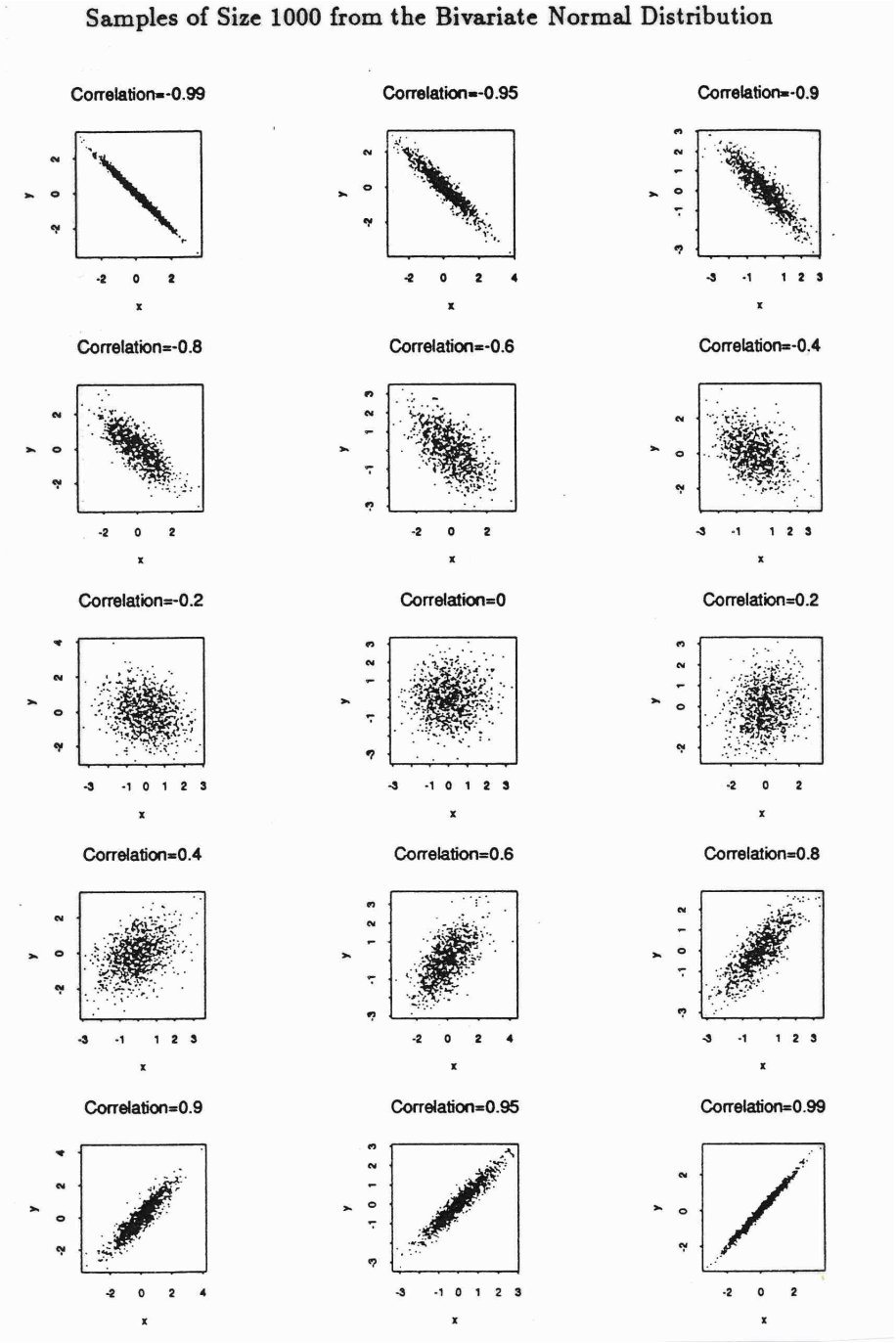
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \quad \text{which has a t-distribution with df=n-2 when } \rho = 0.$$

Often researchers will misinterpret the results from this test. This situation results from a confusion between statistical significance and practical significance. Suppose a psychologist has devised a skills test for production-line workers. It is hoped that the skills test will be an assessment of the worker's potential productivity. The psychologist evaluates the test on a random sample of 40,000 workers. From each worker he obtains their score on the test, X and also a measure of their true productivity, Y. The sample correlation coefficient is computed to be r=0.02. The test of $H_o : \rho = 0$ vs $H_a : \rho \neq 0$ yields

$$t = \frac{.02\sqrt{39998}}{\sqrt{1-(.02)^2}} = 4.0, \quad \text{with df=39998} \Rightarrow \text{P-value} = 0.00006$$

Thus, the null hypothesis would be rejected at any reasonable value for α and the conclusion was stated as **there was a significant correlation between the skills test and the productivity of the worker**. However, r=.02 would indicate nearly no linear relationship between the two variables. There is no practical meaningful relationship between these test scores and the productivity of the workers. The problem is that with very large sample sizes the test has nearly power of size 1.0 to detect very small correlations. The correct statement is that the there is statistically significant evidence that the correlation is not zero but may be very near zero.

A second problem that arises both with correlation and regression models is that of the difference between association and causation. High correlation (or R^2 in a regression model) is often interpreted incorrectly as a **cause and effect** relationship. Such a conclusion should not be reached in observational studies in which many variables are uncontrolled or unobserved. Recall the Simpson paradox in contingency tables. Any correlation between X and Y may have occurred because both X and Y are highly affected by a third variable which is either unknown or unobservable. A high fat diet was found to be associated with a high risk for breast and colon cancer in several studies. However, high fat diets are found in richer, more developed countries which also have other dietary, lifestyle, and environmental differences from less developed countries which have low fat diets. It may be these other factors which are causing the rise in the cancer rate. There are classic examples where it is obvious there is no relationship between two variables but a statistical correlation may exist between their numbers. For example, there is a strong correlation between the winning percentage of the New York Yankees baseball team over a series of y ears and annual coal production in England in the same years.


Samples of Size 1000 from the Bivariate Normal Distribution
Correlation=-0.99
Correlation=-0.95
Correlation=-0.9
Correlation=-0.8
Correlation=-0.6
Correlation=-0.4
Correlation=-0.2
Correlation=0
Correlation=0.2
Correlation=0.4
Correlation=0.6
Correlation=0.8
Correlation=0.9
Correlation=0.95
Correlation=0.99
x
y

Spearman's Rank Correlation Coefficient

Pearson's correlation, r, is a measure of the strength of the *linear association* between X and Y. Furthermore, any inferences using r require the assumption of bivariate normality. An alternative coefficient, Spearman's Rank Correlation Coefficient, r_{SP}, requires only the scale of X and Y be ordinal. It measures the monotonicity of the relationship not just the linearity. It is calculated identically to r after first performing a rank transformation of the X's and Y's separately:

1. Replace the X_i's with their ranks R_i's amongst the X_is
2. Replace the Y_i's with their ranks S_i's amongst the Y_is
3. Compute r on the n (R_i, S_i) pairs yielding r_{SP}

$$Z = r_{SP}\sqrt{n-1} \text{ which is approximately standard normal under } H_o$$

can be used to test H_o : X and Y are independent vs H_a : X and Y are associated.

Using the data from the example discussed on the next page: r_{SP}

Case	Age	Score	Rank-Age	Rank-Score
1	15	95	14.5	11.0
2	26	71	20.0	2.0
3	10	83	6.5	3.5
4	9	91	3.5	8.0
5	15	102	14.5	16.5
6	20	87	18.5	7.0
7	18	93	17.0	9.0
8	11	100	10.5	14.0
9	8	104	2.0	18.0
10	20	94	18.5	10.0
11	7	113	1.0	20.0
12	9	96	3.5	12.0
13	10	83	6.5	3.5
14	11	84	10.5	5.0
15	11	102	10.5	16.5
16	10	100	6.5	14.0
17	12	105	13.0	19.0
18	42	57	21.0	1.0
19	17	121	16.0	21.0
20	11	86	10.5	6.0
21	10	100	6.5	14.0

From the 21 original data values we have $r = -.640$ with p-value=.002 for testing $H_1 : \rho < 0$.

The value for $r_{SP} = -.317$ with a p-value=.162.

If we recompute r and r_{SP} excluding the 18th data value we obtain:

$$r = -.335, \quad p - value = .149 \qquad\qquad r_{SP} = -.208, \quad p - value = .380$$

Problem in Using Regression Analysis

1. **Cause and Effect**

   There is a tendency to think of the explanatory variable X *Causative Agent* and the response varible Y as an *Effect*. With randomized, controlled experiments, *cause-and-effect* terminology is appropriate.

   Example Explanatory Variable: Dose level of Drug

   Response Variable: Decrease in Blood Pressure of Subject

   However, in observational studies, this is NOT appropriate. We should use the term *Association* to describe any relationship between the explanatory and response variables.

   Example Suppose we collect data from 50 countries on

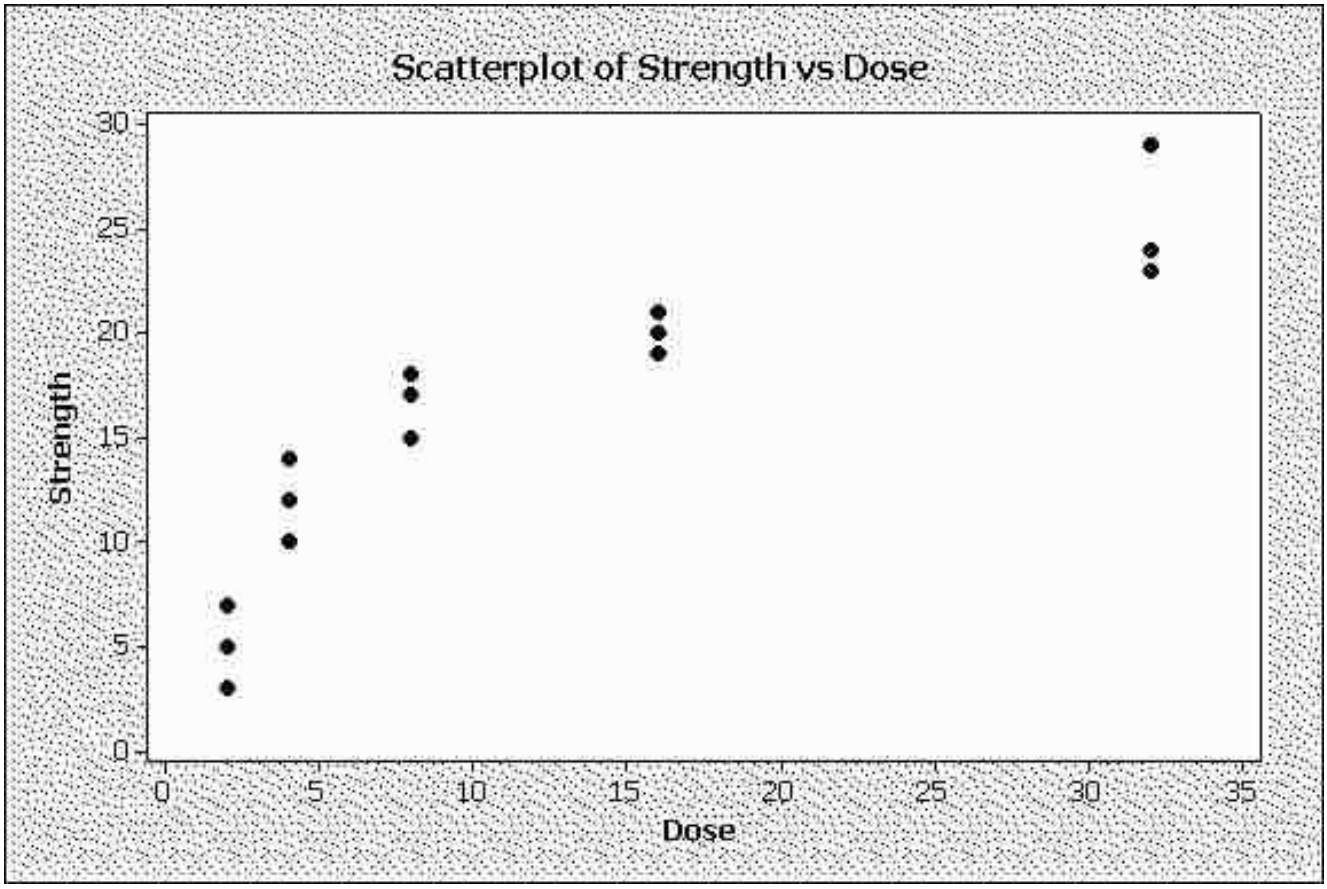
   Explanatory Variable: Wine Consumption (liters per person per year)

   Response Variable: Heart Disease Mortality Rates (deaths per 1000)

   Example Consider the data on the 21 children relating Age at which child first speaks and Score on apptitude test.

2. **Extrapolation and Interpolations**

Interpolation is when estimates are made of $\mu_{Y|X}$ or $\hat{Y}_x$ for values of X which are within the range of the observed values of $X : X_1, \ldots, X_n$ but may not necessarily be one of these n values.



In the Dose-Strength Example, we would want to estimate the strength of the drug for dose levels other than 2, 4, 8, 16, and 32. If we estimates the average strength of the drug at dose levels of 10, 20, and 30 using of least squares equation: $\mu_{Y|x} = 8.7 + .575x$ for $x = 10, 20, 30$, we would be making interpolations from the observed data. However, if we attempted to estimate $\mu_{Y|x}$ for $x = .5$ or $x = 50$ we may be making serious errors because the true relationship between $\mu_{Y|x}$ and X may be very non-linear outside of the range of $x = 2$ and $x = 32$.

When we attempt to estimate $\mu_{Y|x}$ for x beyond the range of the observed data, we are making **Extrapolations**. Only in the case of time series data, when very specific models are used and very specific conditions hold are extrapolations, (called forecasts in time series) a sensible inference.

3. **Outliers**

An *Outlier* is an observation (X_i, Y_i) having a value for its residual $e_i = Y_i - \hat{\beta}_o - \hat{\beta}_1 X_i$ that is large in magnitude relative to what would be expected from a $N(0,1)$ sample. The *Studentized Residual* is defined to be

$$T_i = \frac{e_i}{SE(e_i)}$$

where $SE(e_i) = \hat{\sigma}\sqrt{1-h_i}$, $\hat{\sigma} = \sqrt{MSE}$, and $h_i = \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \frac{1}{n}$.

If the conditions for the model are satisfied, then T_i has a $t-$distribution with $df = n-2$. Thus, observation (X_i, Y_i) is said to be an outlier if $|T_i| > 3$.

If a data point is identified as an outlier, then the data point is not typical of the rest of the data. A careful examination of this data point should be conducted. However, an automatic deletion of the data value if $|T_i| > 3$ should not be done. Outliers should only be deleted when it arises from atypical or unusual experimental conditions. The unusual circumstances may be of vital interest and require further investigation rather than deletion.

General Rule: If $|T_i| > 3$ AND the data value (X_i, Y_i) was obtained under experimental conditions greatly different from the remaining data or if recording errors have occured or any other possible extenuating circumstances have occurred, then remove the data value and refit the regression line. Also, document that this data value was in the original data and explain why it was deleted prior to fitting the line.

4. **Influential Data Values**

A data value (X_i, Y_i) is said to have *High Leverage* for the explanatory variable if the value of X_i is distant in the $X-$ direction from the remaining data values. This will often cause a distortion in the value of both $\hat{\beta}_o$ and $\hat{\beta}_1$.

The *Leverage* of (X_i, Y_i) is defined to be

$$h_i = \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \frac{1}{n}$$

A large value for h_i would indicate that X_i is far from $\bar{X}$. In particular, $\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_i = \frac{2}{n}$. Thus, if h_i is much larger than $\frac{2}{n}$ we would consider (X_i, Y_i) to have *High Leverage*. In particular, a value of $h_i > \frac{4}{n}$ is taken as an indicator of high leverage.

The following example from *Applied Regression Analysis*, by Draper and Smith, illustrates the effect of **influential** data points on the regression line.

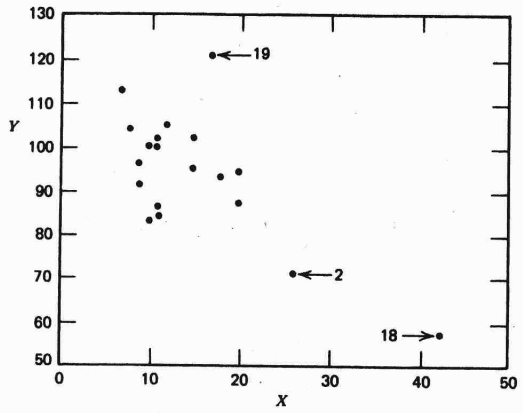
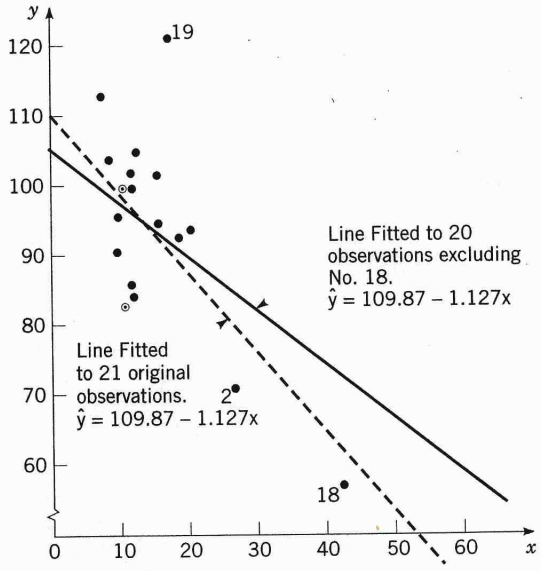


Figure 8.1. A regression with an observation (19) that may not be influential and one (18) that may well be. X represents the age of a child at first word (in months) and Y represents the child's score on an aptitude test. Reproduced by permission from Andrews and Pregibon (1978). The original data were recorded by Dr. L. M. Linde of UCLA and were given by Mickey, Dunn, and Clark (1967). See Table 8.1 for the data.

TABLE 8.1. Age at First Word (X) and Gesell Adaptive Score (Y)

Case	X	Y
1	15	95
2	26	71
3	10	83
4	9	91
5	15	102
6	20	87
7	18	93
8	11	100
9	8	104
10	20	94
11	7	113
12	9	96
13	10	83
14	11	84
15	11	102
16	10	100
17	12	105
18	42	57
19	17	121
20	11	86
21	10	100

Source: Data from Mickey, Dunn, and Clark (1967) but recorded by L. M. Linde of UCLA.



There are a number of other measures of influential data values such as Cook's Distance which measures influence with respect to the influence of (X_i, Y_i) on the prediction of Y_i. Other measures of influence use as a metric the influence on the estimation of the βs.

In the case of a single explanatory variable, measures of influence are not crucial because we can often visual detect extreme data values in either the $Y-$ or $X-$ direction. When there are many explanatory variables the graphical or visual determination is much more complex.