CHAPTER 9

Simple Linear Regression and Correlation

Regression – used to predict or estimate the value of one variable corresponding to a given value of another variable.

$X$ = independent variable.

$Y$ = dependent variable.

Assumptions for Simple Linear Regression of $Y$ on $X$:

1. Values of $X$ are fixed (preselected).
2. $X$ is measured with negligible error.
(3) For each $X$, there is a subpopulation of $Y$ values that is normal. However, estimates of coefficients and their standard errors are robust to nonnormal distributions. 

(4) Variances of subpopulations of each $Y$ are all equal. 

(5) **Assumption of Linearity** – the means of the subpopulations of $Y$ lie on a line,

$$
\mu_{y|x} = \beta_0 + \beta_1 x,
$$

where $\mu_{y|x}$ is the mean of the subpopulation of $Y$ for a given value $x$ of $X$, and $\beta_0$ and $\beta_1$ are the *population regression coefficients*. 

(6) $Y$ values are statistically independent. 

One can remember **LINE** for the primary assumptions.

**Regression Model:**

$$
y = \beta_0 + \beta_1 x + \epsilon
$$

$$
\implies \epsilon = y - (\beta_0 + \beta_1 x)
$$

$$
\epsilon = y - \mu_{y|x}
$$

Thus $\epsilon$ is the amount by which $y$ differs from the mean of the given subpopulation.

**Regression Analysis** – Four step sample regression equation process:

**Problem** (9.3.2).

(1) Are the assumptions met?

(2) Obtain sample regression equation on the TI:
   (a) Enter the data (in lists $x, y$)
(b) Create a scatter diagram: Set a window \([-0.25 \times 0.25]\); Turn on a plot: \((\text{Diamond}>Y=>\text{Plot 1}>\text{Enter})\); Then fill in the table as in the diagram on the left below; Make sure all the graphs are cleared;

(c) Obtain the least squares regression line (where the squares of the errors are minimized). Go back to the **Stats/List Editor**;

Press (**F4**:Calc>3:Regressions> 1:LinReg \((a+bx)\)) and fill in the table that opens as in the diagram on the right below.

![Diagram of scatter plot and regression line]

Our regression equation is

\[
\hat{y} = 1.2112 + 1.0823x.
\]

\(\hat{y}\) means computed from the regression equation, not observed. View the line with the points (\text{Diamond}>\text{Graph}).

![Graph showing regression line and points]
(3) Evaluate the strength of the relationship between $x$ and $y$ and the usefulness of the regression equation for predicting and estimating.

$$r = .9119 \implies r^2 = .8316.$$  

$r^2$ is called the coefficient of determination and gives the fraction of variation in the values of $y$ that is explained by regression on $x$. Thus we have a good relationship here.

(4) Use the equation to predict and estimate. We have (18, 18) and (18, 23) as data pairs. What is the approximation to the mean of $Y$ for $X = 18$? With the graph still showing, press $\text{F5:Math}>1:Value$. Then put in 18 for $x_c$ and press $\textbf{Enter}$ to get $\hat{y} = 20.6925$.

$\text{NOTE.}$ We have a $\underline{\text{direct}}$ as opposed to an $\underline{\text{inverse}}$ relationship here.

**EXAMPLE (9.3.1).** This is fully described on pages 23-26 of the SPSS manual. The data file is *example 9.3.1.sav*.

**Evaluating the Regression Equation**

$H_0 : \beta_1 = 0$ not rejected means there is no evidence of a linear relationship – see the scatterplots on page 428.

$H_0 : \beta_1 = 0$ rejected means there is evidence of a linear relationship – see the scatterplots on page 429.
Testing $H_0: \beta_1 = 0$ with the $F$ Statistic

EXAMPLE (9.3.1) – continued

For each $i$ and corresponding $x_i$,

$$\left( y_i - \bar{y} \right) = \left( \hat{y} - \bar{y} \right) + \left( y_i - \hat{y} \right),$$

total deviation explained deviation unexplained deviation

$$y_i = \hat{y} + (y_i - \hat{y}),$$
data fit residual

and

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y_i - \hat{y})^2.$$

This is
\[ \text{SST} = \text{total sum of squares} \]

\[ \text{SSR} + \text{SSE} = \text{sum of squares due to regression} \quad \text{residual sum of squares} \]

\[ r^2 = \frac{\text{SSR}}{\text{SST}} = \text{fraction of variation in } y \text{ explained by regression on } x. \]

Testing

\[ H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0 \]

with \( F \):

\[
F = V.R. = \frac{\text{MSR}}{\text{MSE}}.
\]

The critical value for \( F \) is \( F_{1, n-2}^{1-\alpha} \). For our example it is

\[
F_{1.107}^{.95} = 3.93
\]

We have \( F = 217.279 \), which is greater than the critical value of 3.93, and \( p = \text{Sig.} = .000 \), which for SPSS means \( p < .001 \), we reject \( H_0 \).
Testing $H_0 : \beta_1 = 0$ with the t Statistic

We assume $\hat{\beta}_1$ is an unbiased point estimator for $\beta_1$ and $(\beta_1)_0$ is the hypothesized value for $\beta_1$. If $\sigma_{y|x}^2$ is known, we can use

$$z = \frac{\hat{\beta}_1 - (\beta_1)_0}{\sigma_{\hat{\beta}_1}}.$$ 

Most often $(\beta_1)_0 = 0$. If $\sigma_{y|x}^2$ is unknown, the usual case, the test statistic is

$$t = \frac{\hat{\beta}_1 - (\beta_1)_0}{s_{\hat{\beta}_1}}$$

where $s_{\hat{\beta}_1}$ is an estimate of $\sigma_{\hat{\beta}_1}$. The critical value of $t$ for this example is

$$t^{n-2}_{1-\alpha/2} = t^{107}_{.975} = 1.982$$

<table>
<thead>
<tr>
<th>Coefficients³</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Constant)</td>
</tr>
<tr>
<td>Unstandardized Coefficients</td>
<td>B</td>
<td>-215.981</td>
</tr>
<tr>
<td>Std. Error</td>
<td></td>
<td>21.796</td>
</tr>
<tr>
<td>Standardized Coefficients</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>t</td>
<td></td>
<td>-9.909</td>
</tr>
<tr>
<td>Sig.</td>
<td></td>
<td>.000</td>
</tr>
<tr>
<td>95% Confidence Interval for B</td>
<td>Lower Bound</td>
<td>-259.190</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upper Bound</td>
</tr>
</tbody>
</table>

a. Dependent Variable: y

Since $t = 14.70$ with $p = \text{Sig.} = .000$, we conclude $p < .001$. Thus we reject $H_0$. Recalling that the point estimate for $\beta_1$ is 3.459, we also see that a 95% CI for $\beta_1$ is (2.994, 3.924). Note that 0 is not in the CI, which is again sufficient evidence to reject $H_0$. 
Estimating the Population Coefficient of Determination

In general, \( r^2 > \rho^2 \), the population coefficient of determination. Thus, \( r^2 \) is a biased estimator of \( \rho^2 \).

An unbiased estimator of \( \rho^2 \) is

\[
\tilde{r}^2 = 1 - \frac{SSE}{SST} \cdot \frac{n - 1}{n - 2},
\]

called the adjusted \( r^2 \).

\[
\tilde{r}^2 \to r^2 \text{ as } n \to \infty
\]

since

\[
\frac{n - 1}{n - 2} \to 1 \text{ and } 1 - \frac{SSE}{SST} = \frac{SST - SSE}{SST} = \frac{SSR}{SST} = r^2.
\]

Problem (9.3.2).

\[
r = .9119 \implies r^2 = .8316 \implies r^2 = 1 - .1684.
\]

Since \( n = 10 \),

\[
\tilde{r}^2 = 1 - .1684 \left( \frac{9}{8} \right) = .8105.
\]

Using the Regression Equation

We assume \( \alpha = .05 \).

Estimating the Mean of \( Y \) for a Given value of \( X \).

For each \( x \) we have a point estimate

\[
\hat{y} = \beta_0 + \beta_1 x.
\]

In the graph that follows at the top of the next page, the horizontal line shows the mean of the \( y \)-values, 101.894. We see that the scatter about the regression line is much less than the scatter about the mean line, which is as it should be when the null hypothesis \( \beta_1 = 0 \) has been rejected. The bands about the regression line give the 95\% confidence interval for the mean values \( \mu_{y|x} \) for each \( x \), or from another point of view, the probability is .95 that the population regression line \( \mu_{y|x} = \beta_0 + \beta_1 x \) lies within these bands.
In general, the $100(1 - \alpha)\%$ CI for $\mu_y|x$ when $\sigma^2_{y|x}$ is unknown is

$$\hat{y} \pm t^{n-2}_{(1-\alpha/2)} s_{y|x} \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

where $x_p$ is the particular value of $x$ at which we wish to obtain a prediction interval for $Y$.

Predicting $Y$ for a given $X$.

$$\hat{y} = \beta_0 + \beta_1 x$$

is again the point estimate. The outer bands on the graph at the top of the next page give the 95% confidence interval for $y$ for each value of $x$. 
In general, the $100(1 - \alpha)\%$ CI for $y$ when $\sigma_{y|x}^2$ is unknown is

$$\hat{y} \pm t^{n-2}_{(1-\alpha/2)} s_{y|x} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}}.$$  

The confidence bands in the scatter plots relate to the four new columns in our data window, a portion of which is shown at the top of the next page. We interpret the first row of data. For $x=74.5$, the 95% confidence interval for the mean value $\mu_{y|74.5}$ is $(32.41572, 52.72078)$, corresponding to the limits of the inner bands at $x=74.5$ in the scatter plot, and the 95% confidence interval for the individual value $y(74.5)$ is $(-23.7607, 108.8972)$, corresponding to the limits of the outer bands at $x = 74.5$. The first pair of acronyms lmci and umci stand for lower mean confidence interval and upper mean confidence interval, respectively, with the i in the second pair standing for individual.
9. SIMPLE LINEAR REGRESSION AND CORRELATION

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>y</th>
<th>LMCI_1</th>
<th>UMCI_1</th>
<th>LICI_1</th>
<th>UICI_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>74.75</td>
<td>25.72</td>
<td>32.41572</td>
<td>52.72078</td>
<td>-23.7607</td>
<td>108.8972</td>
</tr>
<tr>
<td>2</td>
<td>72.60</td>
<td>25.89</td>
<td>24.17575</td>
<td>46.08766</td>
<td>-31.3250</td>
<td>101.5884</td>
</tr>
<tr>
<td>3</td>
<td>81.80</td>
<td>42.60</td>
<td>59.11112</td>
<td>74.79530</td>
<td>.93840</td>
<td>132.9680</td>
</tr>
<tr>
<td>4</td>
<td>83.95</td>
<td>42.80</td>
<td>67.10283</td>
<td>81.67669</td>
<td>8.43859</td>
<td>140.3409</td>
</tr>
</tbody>
</table>

The Correlation Model

Both $X$ and $Y$ are now random variables, and both are measured from random “units of association” (the element from which the two measurements are taken).

**Example.** Choose 15 CBU students at random and measure their height $X$ and weight $Y$.

Each variable is on equal footing here, and we measure the strength of the relationship. We can also do regression of $Y$ on $X$ or regression of $X$ on $Y$.

Correlation Assumptions:

1. For each value of $X$, there is a normally distributed population of $Y$ values.
2. For each value of $Y$, there is a normally distributed population of $X$ values.
3. The joint distribution of $X$ and $Y$ is a normal distribution called the *bi-variate normal distribution*.
4. The subpopulations of $Y$ values all have the same variance.
5. The subpopulations of $X$ values all have the same variance.

The Correlation Coefficient

$\rho$ (for population) measures the strength of the linear relationship between $X$ and $Y$.

$\rho = \pm \sqrt{\rho^2}$, the previously discussed coefficient of determination.

$-1 \leq \rho \leq 1$
\[ \rho = 1 \] means perfect *direct correlation* \((\beta_1 > 0)\).

\[ \rho = -1 \] means perfect *inverse correlation* \((\beta_1 < 0)\).

\[ \rho = 0 \] means that the variables are not linearly correlated \((\beta_1 = 0 \text{ for regression})\).

We approximate \( \rho \) with \( r \), the sample correlation coefficient.

\[ r = \pm \sqrt{r^2} \], the sample coefficient of determination (the sign of \( r \) is the same as the sign of \( \beta_1 \)).

Our interest is that \( \rho \neq 0 \), i.e., that

\[ H_0 : \rho = 0 \quad H_A : \rho \neq 0 \]

has \( H_0 \) rejected.

SPSS indicates significance at the \( \alpha = .05 \) and \( \alpha = .01 \) levels of significance. Choose one-sided if the direction of relationship is known (direct or inverse), two-sided otherwise.

**Problem (9.7.3).** – using SPSS.

We view the scatterplot.
Since it is clear we have an inverse relationship here, we do a one-sided test of significance at the $\alpha = .05$ level.

$$H_0 : \rho \geq 0 \quad H_A : \rho < 0$$

![Correlations Table]

*Correlation is significant at the 0.05 level (1-tailed).

We have $\rho \approx r = -.812$ and we have $p = .013$. Thus the correlation is marked as significant at the $\alpha = .05$ level, but not at the $\alpha = .01$ level. We reject $H_0$.

**Precautions** – read through these clearly on pages 459-60 of the text.