THE LINEAR MODEL WITH NORMAL ERROR

7.1 THE SIMPLEST LINEAR MODEL - THE LINE

Researchers attempt to summarize the relationship between *y* and *x* through equations as well as graphs. We have previously, in section 6.8, seen that an equation given by

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

is a straight line. In order to predict *y* from *x*, we substitute into the equation. Suppose a researcher has a graph of *y* versus *x* and wishes to determine the best fitting line to this set of points. The best fitting line is called a ***regression*** line, for historical reasons. The method of ***least squares*** is a method of finding this best line. Basically, deviations of the point from the line in the vertical direction are used as criteria for selecting the line. Some of the deviations will be positive, and others negative; however, all of the squares of their deviations are positive. The line that makes this sum smallest is the "best" line. In other words, the line

$$\hat{y}_i = b_0 + b_1 x_i$$

(read $\hat{y}$ as "*y* hat" -- predicted *y* by the fitted line)

that minimizes the sum of the squared deviations in the vertical direction.

$$\begin{aligned} \textbf{residual}_i^2 &= (\textbf{observed.}y_i - \textbf{predicted.}\hat{y}_i)^2 \\ &= (y_i - b_0 - b_1 x_i)^2 \\ &= e_i^2 \end{aligned}$$

is the best line. These deviations $(e_i = y_i - \hat{y})$ are called ***residuals*** or ***errors***.

For example, see Figure 7.1 below:

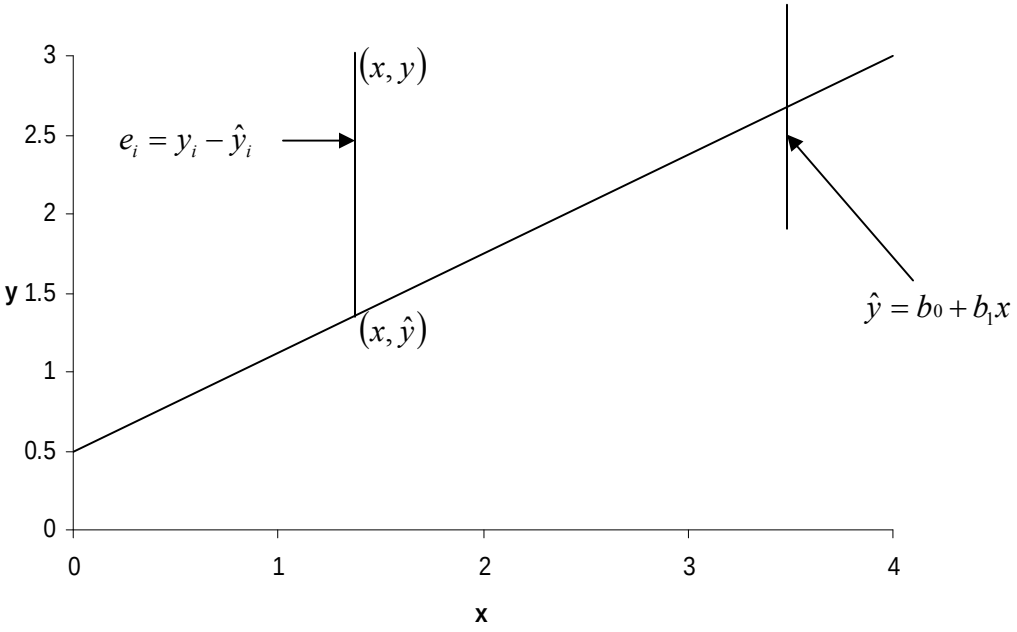


Figure7.1

The least squares regression line, $\hat{\boldsymbol{y}} = \boldsymbol{b_0} + \boldsymbol{b_1 x}$, is given using the following formulas:

$$\boldsymbol{b_1} = \frac{\sum (\boldsymbol{x} - \bar{\boldsymbol{x}})(\boldsymbol{y} - \bar{\boldsymbol{y}})}{\sum (\boldsymbol{x} - \bar{\boldsymbol{x}})^2}$$

$$\boldsymbol{b_0} = \bar{\boldsymbol{y}} - \boldsymbol{b_1}\bar{\boldsymbol{x}}$$

Note that the researcher would hypothesize a line would be an appropriate model for the data. If the line is an appropriate model, then the reasoning in the previous section would be used to select the best line given the data collected on the system.

In other words, the ***model*** specified by the researcher is given by

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

where β_0 and β_1 are population parameters that represent the y-intercept and slope of the line, but are unknown. Once the researcher has a sample or data from the population, s/he can use least squares to calculate the best fitting line.

$$\hat{y} = b_0 + b_1 x$$

Here, b_0 and b_1 are the least squares estimators of β_0 and β_1 derived from the data, assuming the true model to be a line. Remember the data is composed of two parts: (1) that explained by the model and (2) that not accounted for by the model, called ***error***.

$$\boldsymbol{DATA = MODEL + ERROR}$$

$$y = b_0 + b_1 x + e$$

We display the relationship graphically in Figure 7.2 below:

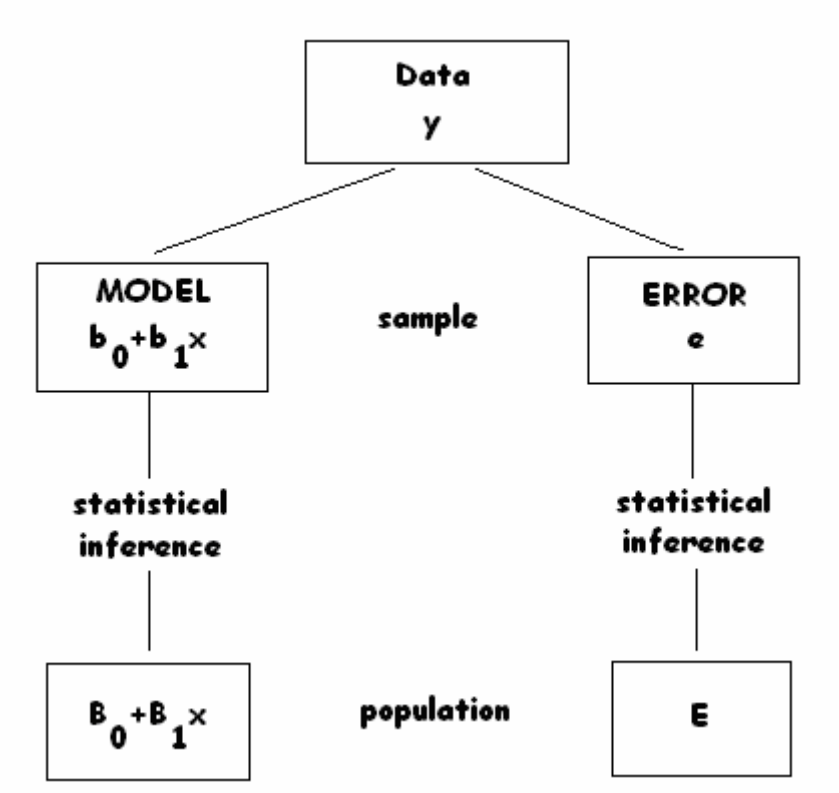


Figure 7.2

If the model predicts the data exactly, then researchers call their type of model ***deterministic***. Note that when the data is accounted for by a model part and an error part, researchers speak of ***probabilistic*** models, since the error part is assumed to be the product of other random, unexplainable effects. The population parameters in the model are unknown and are estimated using the sample data. Clearly, it is assumed that the model, in this case a line, is the correct one for this system. We will examine other models (i.e., polynomials) in Chapter 8.

CLICK here for overview of Polynomials.

7.2 Example

This example was adopted from Gebotys and Roberts *(Canadian Journal of Behavioral Scienc*e, 1987, vol. 19, p. 479-488). Gebotys and Roberts examined the effect of age on how a person viewed crimes' seriousness. A measure of seriousness (0-100) was developed, where a score of 0 indicates a crime that is not serious and a score of 100 indicates a very serious crime. Ten people were asked to rate the seriousness of car theft and, among other things, their age was recorded. The data are given below:

age	seriousness
20	21
25	28
26	27
25	26
30	33
34	36
40	31
40	35
40	41
80	95

The researcher wants to predict seriousness from age, and a literature review suggests a line would be an appropriate model. S/he first plots the data. See figure 7.3 on the next page.

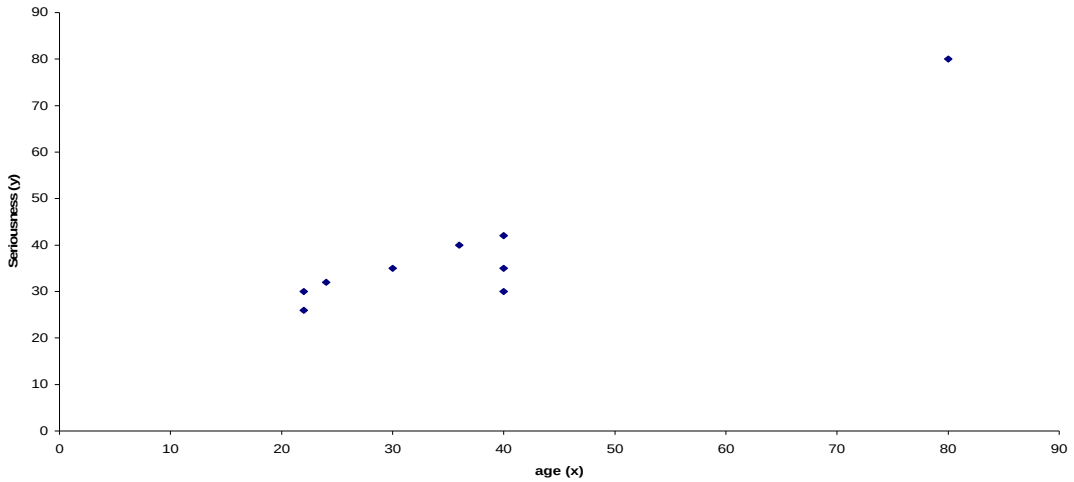


Figure 7.3

The plot confirms the researcher's suspicions that a line would be a reasonable model. Note that most of the ages are clustered between 20 and 40 years, with one elderly person at age 80 years.

Using Table 7.1, and the formula defined previously, we can easily calculate the least squares estimators of β_0 and β_1. In this example, β_0 and β_1 represent the intercept and slope of a line that relates age to crime seriousness for all individuals in the country (the population). The data will enable the researcher to calculate least squares estimates of these values (b_0 and b_1) based on ten people (the sample).

Table 7.1

	x_i	y_i	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	20	21	-16	-16.3	256	265.69	260.8
2	25	28	-11	-9.3	121	86.49	102.3
3	26	27	-10	-10.3	100	106.09	103.0
4	25	26	-11	-11.3	121	127.69	124.3
5	30	33	-6	-4.3	36	18.49	25.8
6	34	36	-2	-1.3	4	1.69	2.6
7	40	31	4	-6.3	16	39.69	-25.2
8	40	35	4	-2.3	16	5.29	-9.2
9	40	41	4	3.7	16	13.69	14.8
10	80	95	44	57.7	1936	3329.29	2538.8
Totals	360	373	0	0	2622	3994.1	3138.0

$\bar{x} = 36$ $\bar{y} = 37.3$

The least squares estimate of β_1, the slope of the line, is

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$= \frac{3138}{2622}$$

$$= 1.197$$

The least squares estimate of β_0, the y-intercept of the line, is

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$= 37.3 - 1.197(36)$$

$$= -5.792$$

The least squares line is given by

$$y = -5.792 + 1.197x$$

In other words, ratings of crime seriousness increase, on average, 1.197 units per year of age. The researcher can now predict seriousness for a given age by substituting into the above equation. Remember the model has been estimated on the basis of only 10 individuals. A larger sample with a larger range of ages (note the gap between $x = 40$ and $x = 80$ years) would improve the researcher's confidence in the model.

7.3 STATISTICAL INFERENCE FOR THE SIMPLE LINEAR MODEL

For the model $\boldsymbol{y = \beta_0 + \beta_1 x}$ the population of values for y is assumed to be ***normally*** distributed about a mean that depends on x. We write $E(y|x)$ to read "the expected or average value of y given x" to represent these means. With this model, the researcher assumes all the means lie on a line when plotted against x. The line

$$\boldsymbol{E(y \mid x) = \beta_0 + \beta_1 x}$$

is given in Figure 7.4 on the next page.

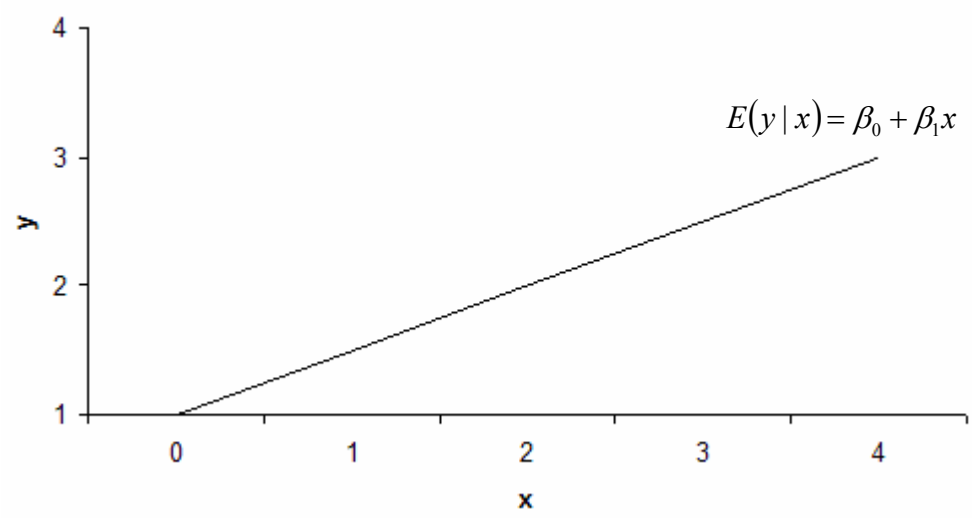


Figure 7.4

The residuals *(e_i)* or errors are assumed to:

1. Be independent;
2. Follow a normal distribution with mean equal to 0; and,
3. Have unknown variance equal to σ^2.

Researchers usually check the normality assumption (Assumption 2) using a ***probability plot*** (see section 3.7). We have previously seen in 3.7 that if the distribution of observations is normal, the probability plot of residuals will reveal a straight line. Deviations from the line indicate non-normality.

CLICK here for overview of Probability Plots.

We estimate σ^2 in the population of residuals using s^2, our estimate from the data.

$$s^2 = \frac{\sum e_i^{\,2}}{n-2}$$

$$= \frac{\sum (y_i - \hat{y}_i)}{n-2}$$

n-2 is called the ***degrees of freedom*** for s^2, an unbiased estimator of σ^2. We lose two degrees of freedom since we estimate two parameters (β_0, β_1) in the model. σ^2 is the spread or scaling of the observations about the line. This spread of the normal distribution is assumed constant across *x* (see Figure 7.4 above).

7.4 Example (cont'd)

For the Gebotys and Roberts (1987) data, calculate s^2, our estimate of σ^2, with the necessary information from Table 7.2.

Table 7.2

	x_i	y_i	$\hat{y}_i = b_0 + b_1 x_i$	$e_i = y_i - \hat{y}$	e_i^2
1	20	21	18.15	2.85	8.12
2	25	28	24.13	3.87	14.98
3	26	27	25.33	1.67	2.79
4	25	26	24.13	1.87	3.50
5	30	33	30.12	2.88	8.29
6	34	36	34.91	1.09	1.19
7	40	31	42.09	-11.09	122.99
8	40	35	42.09	-7.09	50.27
9	40	41	42.09	-1.09	1.19
10	80	95	89.97	5.03	25.30
Totals	360	373	373.01	-0.01	238.62

$$s^2 = \frac{\sum e_i^2}{n-2}$$

$$= \frac{238.62}{10-2}$$

$$= 29.83$$

In other words, the spread, or variance, of the residuals about the line is 29.83. Researchers often report the square root of s^2, since, given the assumption of normality, some quick calculations can be made about the residuals and their distribution. We know that +/- 1 standard deviation for the normal gives approximately 68% of the data.

In our example,

$s = 5.46$

therefore, +/- 5.46 about the predicted mean from the model would give approximately 68% of the data. It is also convenient to have *s* for future calculations.

7.5 SIGNIFICANCE TESTS AND CONFIDENCE INTERVALS

Previously, we reviewed significance tests and confidence intervals for means using the *t* distribution (see section 3.4). Similar procedures are applicable to the slope and intercept parameters of the simple linear model when the sample has been selected randomly, as described in Chapter 2.

All tests of hypothesis, parameter equal to zero, have the following form, with the *T* statistic equal to

$$T = \frac{estimate}{standard.error.of.estimate}$$

For the slope, we have, for example

$$H_o : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

$$T = \frac{b_1}{s(b_1)}$$

where the *T* statistic has a student *t* distribution with *n*-2 degrees of freedom.

Similarly, all of the confidence intervals are of the form

$$estimate \pm t_{\alpha/2} \; standard\ error\ of\ the\ estimate$$

For example, a $1 - \alpha$ confidence interval for the slope β_1 is

$$b_1 \pm t_{\alpha/2} s(b_1)$$

where t is the value from the t distribution with n-2 degrees of freedom.

To estimate $s(b_1)$, we use the formula

$$s(b_1) = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Similarly, we have the standard error $s(b_0)$

$$s(b_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

The computer usually includes tests of $H_o : \beta_0 = 0$; however, this information is usually not of interest to the researcher unless $x = 0$ exists or has practical importance.

7.6 Example (cont'd)

For the Gebotys and Roberts (1987) data, test the hypothesis that the slope is equal to zero. If the slope is zero, then we conclude as age changes, there is no effect on seriousness except for random variation. If we reject H_0, then we conclude as age changes, seriousness changes.

Perform the test at $\alpha = .05$.

$$H_o : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

The standard error is easily calculated below. We know *s* is equal to 5.46 from section 7.4.

$$s(b_1) = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

$$= \frac{5.46}{51.21}$$

$$= .11$$

The *T* statistic is calculated next using the formula below:

$$T = \frac{b_1}{s(b_1)} = \frac{1.197}{.11} = 10.88$$

The degrees of freedom are $df = (n - 2) = 10 - 2 = 8$.

The error critical value, $t_{critical} = 2.306$, is found in tables of the *t* distribution with 8 degrees of freedom and $\alpha/2 = .025$. Since the value of the *T* statistic, 10.88, is greater than 2.306 (the critical value for a two tailed test when $\alpha = .05$), we reject $H_o : \beta_1 = 0$, the slope is equal to zero, and say the result is significant at $\alpha = .05$. Age is important in predicting crime seriousness. Since the slope is positive, the greater the *x* value (age) the greater the *y* value (seriousness).

A 95% confidence interval for the slope β_1 is

$$b_1 \pm t_{\alpha/2} s(b_1)$$

$$= 1.197 \pm 2.306(.11)$$

$$= (.943, 1.451)$$

The slope of the line (β_1) is between .943 and 1.451, with 95% confidence. Researchers often report confidence intervals to quantify their uncertainty with respect to the slope. Remember that the 10 individuals have given us an estimate (b^2) of the population's slope (β_1). Replications of the above study would give different estimates. A confidence interval gives the researcher a degree of certainty where the true value of the slope (β_1) lies with a given probability.

7.7 CONFIDENCE INTERVALS FOR $E(y|x)$

The model

$$E(y \mid x) = \beta_0 + \beta_1 x$$

gives the mean value of y for a given x. In order to estimate the mean from the sample, we use the least squares estimators b_0, b_1 of β_0 and β_1.

$$E(y \mid x) = b_0 + b_1 x$$

To construct a confidence interval for $E(y|x_0)$, for a particular value of x , say x_0, we need the standard error of $E(y|x_0)$.

$$s[E(y \mid x_0)] = s\sqrt{1/n + (x_0 - s)^2 / \sum (x_i - \bar{x})^2}$$

A 1 - α confidence interval for the mean when $x = x_0$ is

$$E(y \mid x_0) \pm t_{\alpha/2} s[E(y \mid x_0)]$$

where $t_{\alpha/2}$ is the α/2 critical value of the t distribution with n-2 degrees of freedom.

Example

Compute a 95% confidence interval for mean seriousness rating when age is 25 years.

First, substitute into the estimated model $x = 25$.

$$\begin{aligned} E(y \mid x) &= b_0 + b_1 x \\ &= -5.792 + 1.197(25) \\ &= 24.13 \end{aligned}$$

When $x = 25$, the mean seriousness rating is 24.13.

Next, calculate the standard error.

(See Table 7.1 for the value of $\sum (x_i - \bar{x})^2$)

$$s[E(y \mid x_0)] = s\sqrt{1/n + (x_0 - s)^2 / \sum (x_i - \bar{x})^2}$$

$$= 5.46\sqrt{1/10 + 121/2622}$$

$$= 2.09$$

For a 95% confidence interval, α = .05, α/2 = .025, the degrees of freedom are 10-2 = 8, and the upper critical value for the *t* distribution with 8 degrees of freedom is 2.306. Substituting into the following formula, we obtain

$$E(y \mid x_0) \pm t_{\alpha/2} s[E(y \mid x_0)]$$

$$= 24.13 \pm 2.306(2.09)$$

$$= (19.31, 28.95)$$

For people with an average age of 25 years, the mean seriousness rating is between 19.31 and 28.95 with 95% confidence. We will soon see how to easily calculate this using SPSS.

7.8 PREDICTION INTERVALS FOR *Y*

The predicted value for a future *y* can also be estimated using

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

Note that the result is the same for predicting the average value $E(y|x_0)$ for a particular *x*, say x_0. The different notation serves as a reminder of the two cases. The error, however, is larger when predicting a single observation than when a mean of a number of observations is considered. This is a result of the fact that the observation *y* will vary about a sample mean, as well as vary about the true model. The two sources of error (about a mean and the model) are reflected in the standard error for the prediction interval.

$$s(\hat{y}) = \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

The 1 – α prediction interval for a future observation y when $x = x_0$ is

$$\hat{y} \pm t_{\alpha/2} s(\hat{y})$$

where $t_{\alpha/2}$ is the upper α/2 critical value for the t distribution with n - 2 degrees of freedom.

Example

Construct a 95% prediction interval for seriousness when the age of the particular person is 25.

$$\hat{y} = b_0 + b_1 x$$

$$= 24.13$$

$$s(\hat{y}) = \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$= \sqrt{1 + \frac{1}{10} + \frac{121}{2622}}$$

$$= 5.85$$

The 95% predicted interval is

$$\hat{y} \pm t_{\alpha/2} s(\hat{y})$$

Substituting $\hat{y}$ and $s(\hat{y})$ above, with $t_{\alpha/2}$ and 8 degrees of freedom, we obtain

$$24.13 \pm 2.306(5.85)$$

$$= (10.64, 37.62)$$

For a person who is 25 years old, we predict a crime seriousness score between 10.62 and 37.62 with 95% confidence. Note that the interval is larger for an individual than the average prediction given in section 7.6, for reasons given at the beginning of this section concerning the variability of an observation versus a mean. We will use SPSS to calculate this interval shortly.

7.9 THE ANALYSIS OF VARIANCE (ANOVA) FOR SIMPLE REGRESSION

The researcher summarizes the total amount of variation in the data into two components. The two components are:

1. The amount of variation in the data accounted for by the model.
2. The amount of variation in the data not accounted for by the model.

In other words,

Data = 1 + 2

Data = Model + Error

or graphically, in Figure 7.5, the variation in the data is partitioned or analyzed as Model (a line in this case) and Error (that amount not accounted for by the line).

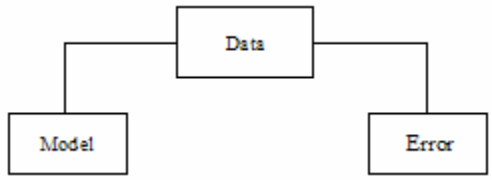


Figure 7.5

The total variation in the data is given by $(y_i - \bar{y})$. Notice that if all these deviations were zero, all observations would be equal and there would be zero variance. The y_i do not equal $\bar{y}$ since subpopulation means differ (i.e., $(\hat{y}_i - \bar{y})$) and are estimated by $\hat{y}_i$. Also, within populations, individual observations will vary about their mean in a normal fashion.

Mathematically, we have

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

If we square the terms

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

We can also write this as sums of squares (*SS*)

$$\boldsymbol{SST = SSM + SSE}$$

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

where we denote

SST = Total sums of squares
SSM = Model sums of squares
SSE = Error sum of squares

The degrees of freedom associated with each *SS* are:

- 1 for *SSM*, since we are only interested in testing H_0: $\beta_1 = 0$ (one parameter);
- n-2 for *SSE*, since there are two parameters in the model (β_0, β_1);
- n-1 for *SST*, since we are not interested in testing H_0: $\beta_0 = 0$, and therefore remove the effect of β_0 (one parameter).

Each component (1 = Model, 2 = Error) has a mean square. The mean square (*MS*) is

$$\boldsymbol{MS = \frac{sums.of.squares}{degrees.of.freedom}}$$

For example, note that

$$MSE = \frac{SSE}{DFE} = \frac{\sum (y_i - \hat{y}_i)}{n-2} = s^2$$

In other words, $MSE = s^2$, an estimate of σ^2, the spread of the residuals about the model.

To test the hypothesis that

$$H_o : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

use the mean square of the Model (*MSM*) and the *MSE* to construct an *F* statistic.

$$\boldsymbol{F = \frac{MSM}{MSE}}$$

where the *p*-value is calculated using a $F(1, n\text{-}2)$ distribution. Note that in this simple case of a line, $F(1, n-2) = t^2_{n-2}$ and both the ANOVA and *t* test techniques give equivalent results in terms of hypotheses tested and *p*-value. In future chapters, when more complex models are considered, it will be shown that one cannot conduct a series of *t* test comparisons without compromising α . The *F* test in the ANOVA table is the proper global test of hypothesis for these models. The line is a special case where both techniques give the same result. This is one reason why there is not a "*t*-table". A more in depth discussion of the ANOVA technique will be given in Chapter 11.

Researchers display the results of the above calculations in an ANOVA table. The table is given below:

Source	Sums of squares (*SS*)	Degrees of freedom (*DF*)	Mean square (*MS*)	*F*
Model	$\sum(\hat{y}_i - \bar{y})^2$	1	$\frac{SSM}{DFM}$	$\frac{MSM}{MSE}$
Error	$\sum(y_i - \hat{y}_i)^2$	$n-2$	$\frac{SSE}{DFE}$	
Total	$\sum(y_i - \bar{y})^2$	$n-1$		

In the case of the simple line, the square of the correlation coefficient, R^2 is the proportion of variation in *y* accounted for by *x*. An $R^2 = 0$ indicates the model accounts for 0 variance, whereas an $R^2 = 1$ indicates a perfect fit for the model. It is interesting to note that R^2 is

a ratio of *SS*. In other words, the degrees of freedom, or number of parameters in the model, are not incorporated directly into the calculation.

$$\boldsymbol{R^2 = \frac{SSM}{SST}}$$

Values can be readily obtained from the ANOVA table. R^2 is a popular summary statistic reported by researchers since it gives an indication of how well the model has accounted for the data. R^2 should be used in conjunction with the ANOVA table since one can have high R^2 and a non-significant model. This can arise when, for example, the sample size for a study is small.

7.10 Example (cont'd)

For the Gebotys and Roberts (1987) data, construct an ANOVA table. Indicate how well the model accounts for the data.

Refer to table 7.1 for *SST*.

$$SST = \sum (y - \bar{y})^2 = 3994.1$$

See table 7.2 for *SSE* and *SSM*.

$SSE = \Sigma e_i^2 = 238.62$

$SSM = SST - SSE$

$= 3994.1 - 238.62 = 3755.48$

Since $n = 10$

DFM = 1

DFE = $n - 2 = 8$

DFT = $n - 1 = 9$

$$MSM = \frac{SSM}{DFM} = \frac{3775.48}{1} = 3755.48$$

$$MSE = \frac{SSE}{DFE} = \frac{238.62}{8} = 29.83$$

$$F = \frac{MSM}{MSE} = \frac{3755.48}{29.83} = 125.90$$

The results are summarized in an ANOVA table.

Source	*SS*	*DF*	*MS*	*F*
Model	3755.48	1	3755.48	125.90
Error	238.62	8	29.83	
Total	3994.1	9		

Note that the *F* statistic is equal to 125.90. Since the OLS or *p*-value is < .0001 with 1, 8 degrees of freedom, we reject H_0: $\beta_1 = 0$ and conclude there is a significant linear relationship between age and seriousness. The ANOVA table is a standard method of communicating results across a wide variety of disciplines. The variance in *y* is partitioned into a model and an error part that are clearly displayed in the table. An estimate of σ^2, the spread about the line, can also be easily calculated from the table since $s^2 = MSE$.

$$R^2 = \frac{SSM}{SST} = \frac{3755.48}{3994.1} = .94$$

In other words, 94% of the variance in crime seriousness (*y*) is accounted for by the model (i.e., age). The model is adequate from an ANOVA and percent variance accounted for (R^2) point of view. An examination of the residuals is next.

7.11 EXAMINATION OF RESIDUALS

For the model

$$\boldsymbol{E(y \mid x) = \beta_0 + \beta_1 x}$$

we have assumed the residuals are normally distributed about a mean that depends on *x*. Three assumptions were made about the normally distributed residuals, e_i. They are as follows:

1. The mean or sum of the residuals is zero;
2. The variance of the residuals σ^2 was the same for all values of *x*. In other words, the spread of the residuals about the model was homogeneous; and,
3. The residuals are independent.

The primary tool we will use for assessing normality is the ***probability plot***, which was discussed in section 3.7. However, a number of other methods will be discussed here to assist the researcher in checking the properties of residuals.

Homogeneous Variance

In order to assess the validity of the assumption that σ^2 of the residuals is constant across *x*, we examine plots of the predicted value, $\hat{y}_i$ versus the residual, e_i.

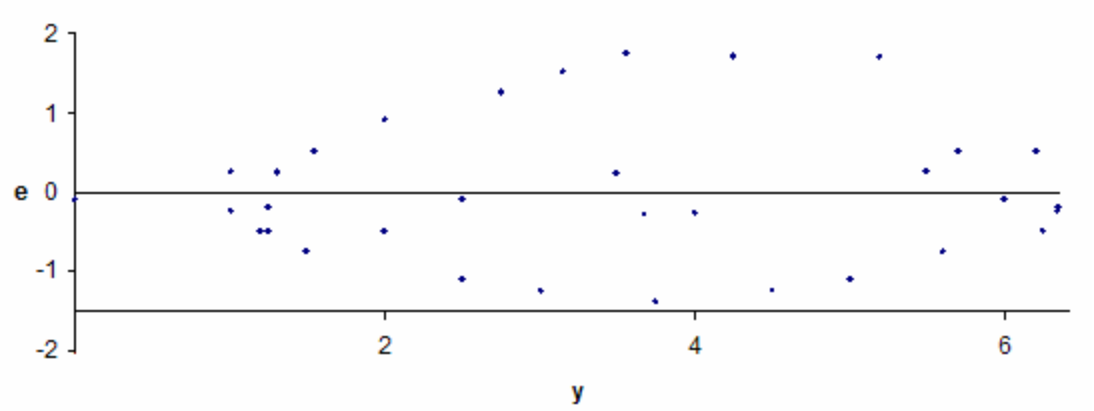


Figure 1

Figure 1 describes the pattern of residuals for ***binomial*** data. Notice the range 0 through 1 and the football shape. Residual patterns such as this are obtained when researchers analyze proportions or percentages.

If we perform an arcsine transformation of *y*

$$y^t = \sin^{-1}\sqrt{y}$$

where the superscript *t* refers to the transformed data, the residuals will more accurately reflect the assumption of homogeneity. See figure 2 on next page.

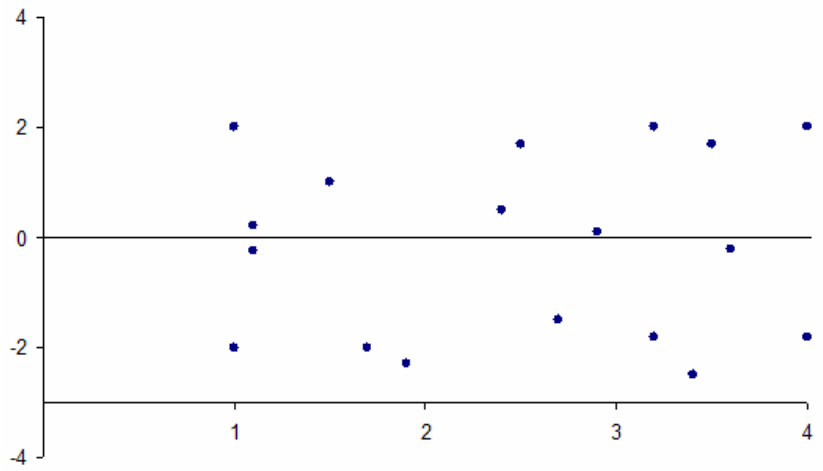


Figure 2

Clearly, the residuals form a symmetric band about 0, with a generally uniform scatter throughout the band. For standardized residuals, $N(0,1)$, most observations fall within ±2 standard deviations, and almost all within ±3 standard deviations.

If the residuals display a fan shape as in figure 3, this indicates the variance is increasing proportionally to the square of the mean.

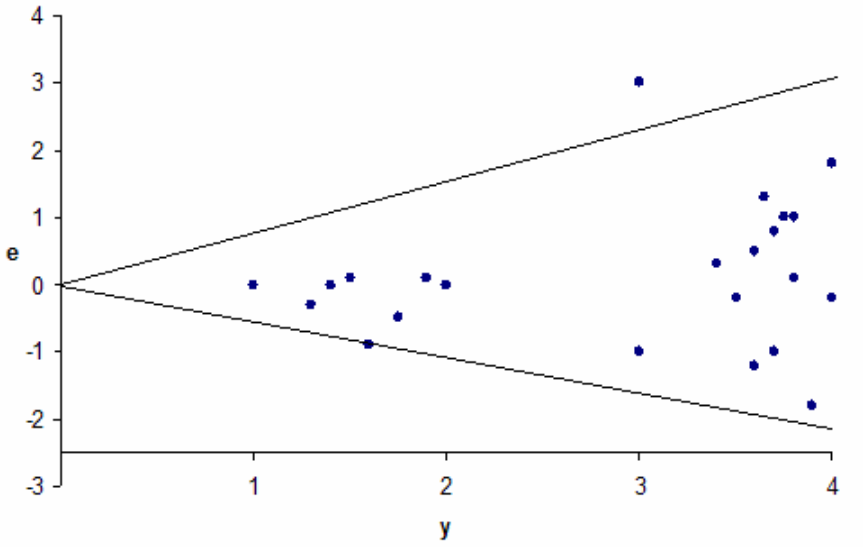


Figure 3

The transformation

$$y^t = logy$$

will stabilize this variance.

When researchers study counts per unit time or area, the variance is typically proportional to E(*y*). This follows a ***Poisson*** distribution which has a cone shaped pattern, as displayed in Figure 4.

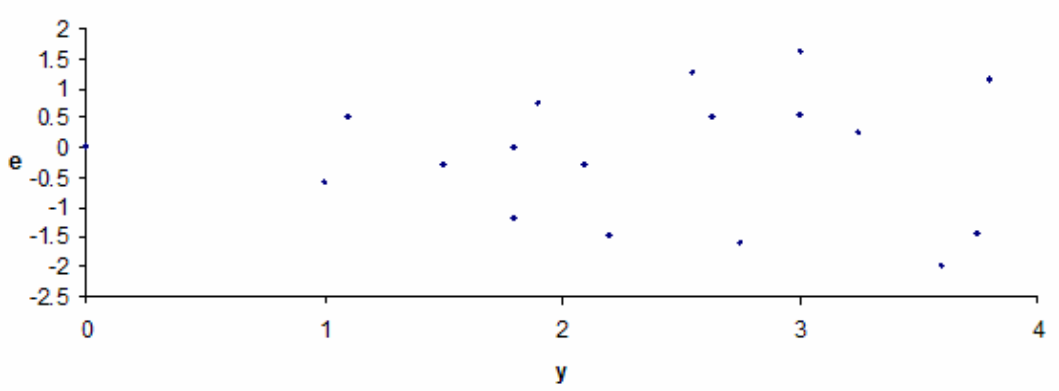


Figure 4

If one takes the square root of *y*

$$y^t = \sqrt{y}$$

the variance is stabilized.

Note that the choice of transformation used to stabilize variance has been selected on the basis of graphical displays. These displays are idealized patterns that with research experience will assist the researcher to properly analyze the data. The validity of the analysis is not compromised by using transformations since the relationship between the raw data and the transformed data is always reported.

A simple test for Homogeneity of Variance

A simple method to compare two variances is to use the *F* statistic, as described in section 3.5. The researcher divides the sample into two sub-samples on the basis of some criteria. For example, say in the Gebotys and Roberts (1987) data, we wanted to compare the seriousness

score variance of the over 40 year old group with the 40 and under group, since there is a suggestion older individuals will have more consistent seriousness ratings. Two groups are clearly defined on the basis of age and a test of equality of seriousness score variance can easily be conducted using

$$H_o : \sigma_1^{\,2} = \sigma_2^{\,2}$$

$$H_a : \sigma_1^{\,2} \neq \sigma_2^{\,2}$$

with the test statistic

$$F = \frac{s_1^{\,2}}{s_2^{\,2}}$$

See section 3.5 for an example of this calculation. H_0 describes the homogenous case, whereas H_a describes a violation of the equal variance assumption. Checking the equality of variance assumption for several variances will be discussed in Chapter 14. Note that the previous section on ***transformations*** of the sample data may help stabilize the variances.

Independence

Another assumption of the linear model is independence of residuals e_i or independence of y. If the values of e_i at time t are correlated with e_{i+1} then this assumption is violated. The test of hypothesis that researchers typically wish to test is that there is zero correlation. The **Durbin-Watson *d*-statistic** is commonly used for this test.

$$H_o : p(residuals) = 0$$

$$H_a : p \neq 0$$

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^{\,2}}$$

Where if the correlation is zero $\rho = 0$, *then* $d \approx 2$

if the residuals are positively correlated ρ^+, *then* $d < 2$

if the residuals are negatively correlated ρ^-, *then* $d > 2$

The *p*-value and *d*-value are printed on the computer output. Notice that the test assumes the residuals are normally distributed. Chapter 24 and Chapter 10 discuss some Time series models that incorporate serial correlation. We will discuss some techniques that handle correlated errors at that time. For the moment, knowledge that the independence assumption may have been violated will be valuable information for the researcher.

Residuals for Model Building

The mathematical model of a line and the assumptions concerning the residuals clearly specify the form and shape the residuals should take in a variety of plots. Residuals are used not only to test assumptions about the distribution of e_i, but also to test how appropriate the mathematical model is for the psychological system under study. For example, if the residual plot of x versus e_i displays a curved pattern as shown in Figure 5 below,

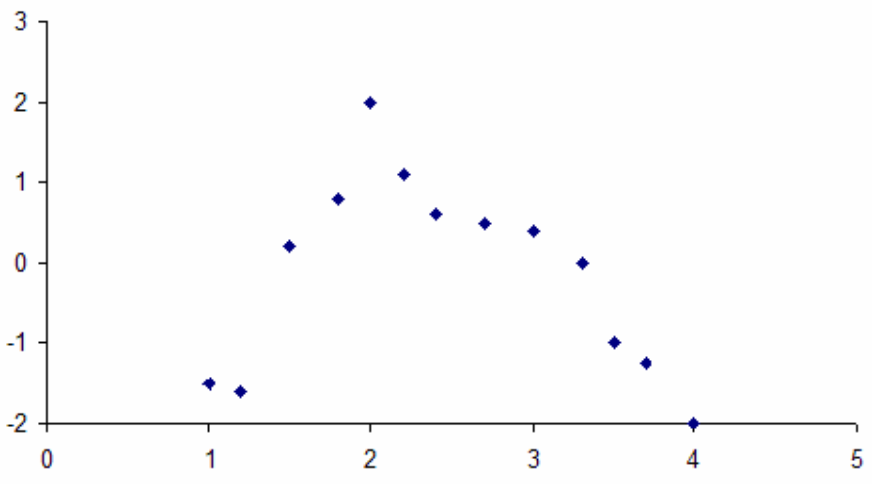


Figure 5

this may indicate a non linear dependence of y on x. For example, it may indicate a quadratic model, as discussed in section 6.10.

7.12 LINEAR MODEL USING SPSS

The **REGRESSION** command fits linear models by least squares. The **VARIABLES** sub-command tells SPSS what variables are in the model. The **DEPENDENT** sub-command tells SPSS which is the y variable. The **ENTER** command defines the x variable. In this case, we have asked SPSS to fit the model, $E(y \mid x) = \beta_0 + \beta_1 x$.

CLICK here for SPSS details.

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	AGE[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: SERIOUS

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.970[a]	.940	.933	5.46	1.040

a. Predictors: (Constant), AGE

b. Dependent Variable: SERIOUS

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3755.547	1	3755.547	125.944	.000[a]
	Residual	238.553	8	29.819		
	Total	3994.100	9			

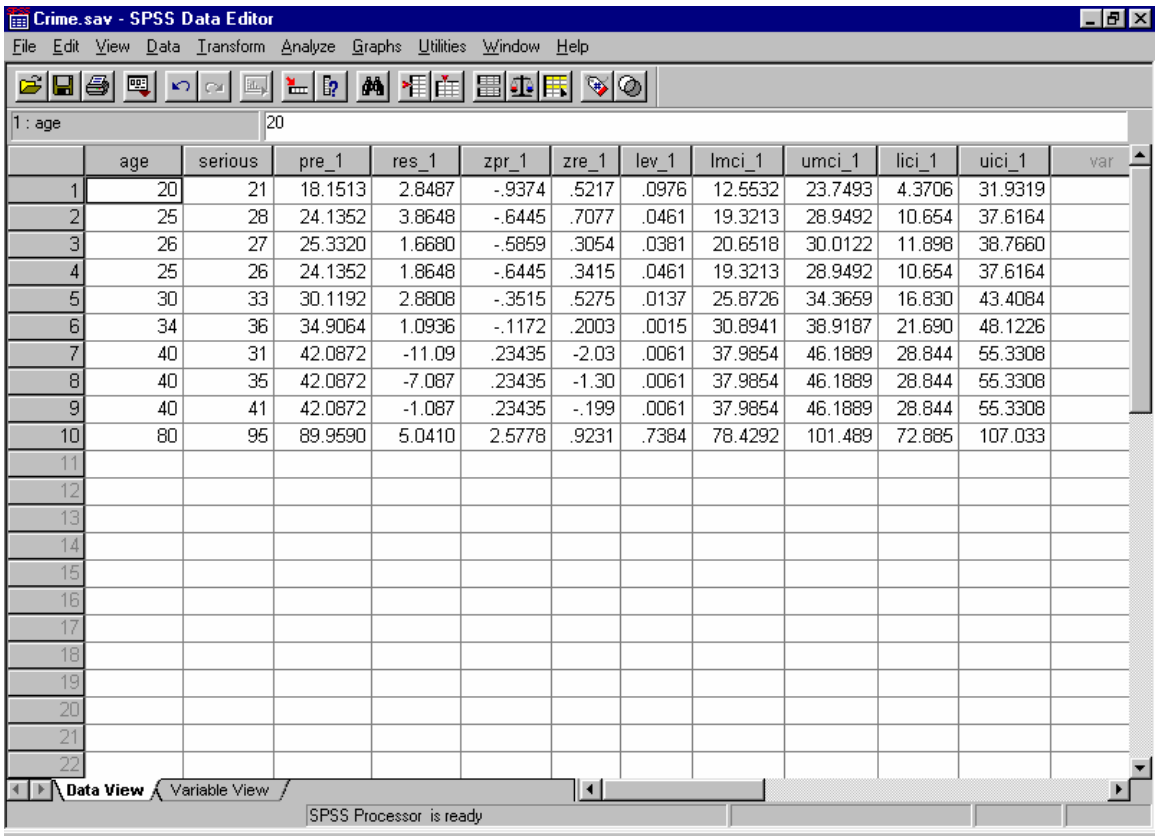
a. Predictors: (Constant), AGE

b. Dependent Variable: SERIOUS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-5.785	4.210		-1.374	.207	-15.492	3.923
	AGE	1.197	.107	.970	11.222	.000	.951	1.443

a. Dependent Variable: SERIOUS



	age	serious	pre_1	res_1	zpr_1	zre_1	lev_1	lmci_1	umci_1	lici_1	uici_1	var
1	20	21	18.1513	2.8487	-.9374	.5217	.0976	12.5532	23.7493	4.3706	31.9319	
2	25	28	24.1352	3.8648	-.6445	.7077	.0461	19.3213	28.9492	10.654	37.6164	
3	26	27	25.3320	1.6680	-.5859	.3054	.0381	20.6518	30.0122	11.898	38.7660	
4	25	26	24.1352	1.8648	-.6445	.3415	.0461	19.3213	28.9492	10.654	37.6164	
5	30	33	30.1192	2.8808	-.3515	.5275	.0137	25.8726	34.3659	16.830	43.4084	
6	34	36	34.9064	1.0936	-.1172	.2003	.0015	30.8941	38.9187	21.690	48.1226	
7	40	31	42.0872	-11.09	.23435	-2.03	.0061	37.9854	46.1889	28.844	55.3308	
8	40	35	42.0872	-7.087	.23435	-1.30	.0061	37.9854	46.1889	28.844	55.3308	
9	40	41	42.0872	-1.087	.23435	-.199	.0061	37.9854	46.1889	28.844	55.3308	
10	80	95	89.9590	5.0410	2.5778	.9231	.7384	78.4292	101.489	72.885	107.033	
11												
12												
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												

The analysis of variance table above is important as a summary table. Note the degrees of freedom and F-statistic values.

$$F = 125.944$$

which has an F distribution with 1 and 8 degrees of freedom. We reject

$$H_o : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

that the slope is equal to zero with p-value less than .0000 (.0000 is to 4 figures only), the **SIG** F value on the output. The **REGRESSION** row refers to the ***model*** and the **RESIDUAL** row refers to the ***error*** component. The mean square of the residual is equal to s^2, our estimate of σ^2.

$$s^2 = \text{MSE} = 29.819$$

$$s = \sqrt{MSE} = 5.461$$

Note *s* is also printed in the **STANDARD ERROR** column. In the same area, we also have ***R*^{2}, *R* SQUARE,** printed where

$$R^2 = \frac{SSM}{SST}$$
$$= \frac{3755.547}{3994.100}$$
$$= .94027$$

In other words, 94.027% of the variance in seriousness is accounted for by age.

The output is interpreted as follows:

In the ***Variables in the equation*** section, the column variable lists the variable ***age*** and ***constant***. These refer to the ***variables*** associated with the parameters β_0 and β_1 in the model. The column labeled **B** gives the least squares ($b_0 = -5.785$, $b_1 = 1.197$) estimator for β_0 and β_1. The equation is therefore $E(y|x) = -5.785 + 1.197x$. The **Std. Error** column is the standard error for each of the parameters. For example

$$s(b_0) = 4.210$$
$$s(b_1) = .107$$

The ***t*** column gives the corresponding *t* statistic for testing the hypothesis

$$H_o : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

For β_1, the *t* statistic has the value 11.222. The column SIG gives the OLS or *p*-value for the test above. In this case, we have at least $p < .001$ (note .000 value is given to 3 figures only) with 8 degrees of freedom. In other words, we reject H_0 and conclude there is a relationship between seriousness and age, since the slope is not equal to zero. We conclude that a line adequately models the relationship between seriousness and age. The relationship is positive (since b_1 is positive) with seriousness increasing 1.197 units per year of age.

We examine a number of plots to assess the normality assumption.

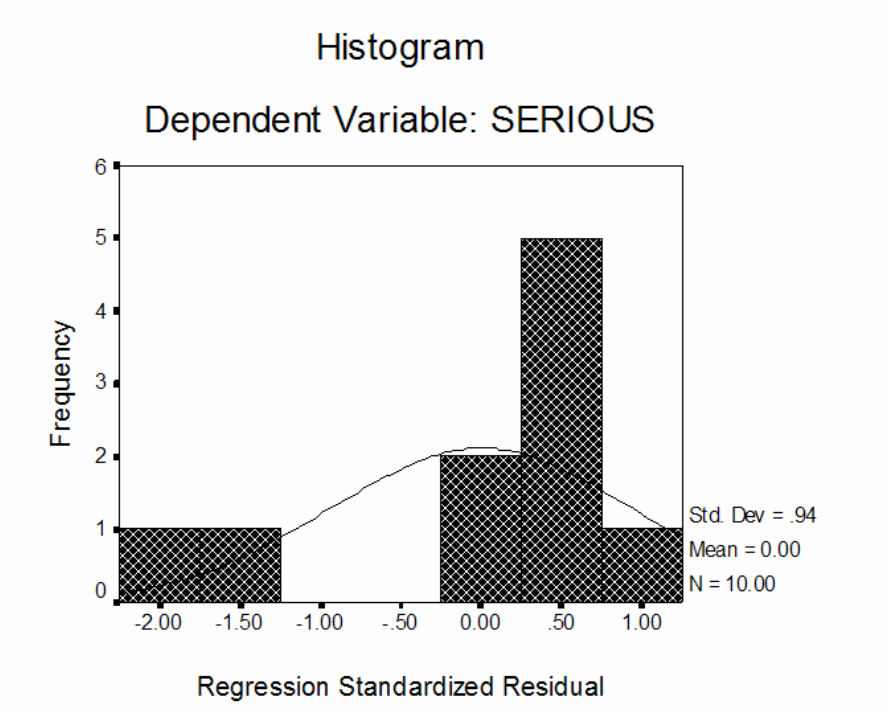
The **CASEWISE** command gives the following output. The actual y (seriousness), predicted y ($\hat{y}$), and residual $e_i = y - \hat{y}$. There is also a plot of the standardized residuals (mean = 0, σ = 1) that is produced by taking the residual and dividing it by the standard deviation of the sample. Given the normality assumption, these residuals should be within 3 standard deviations of their mean (0) and have a band pattern (see Figure 2 for an example), as is the case below. An observation that has a large residual is called an ***outlier***. The casewise plot indicates there are no outliers since the standardized residuals fall within ±3. The Durbin Watson statistic indicates there may be some serial correlation since d = 1.03981; however, given there are only 10 observations, this remains a suspicion.

Casewise Diagnostics[a]

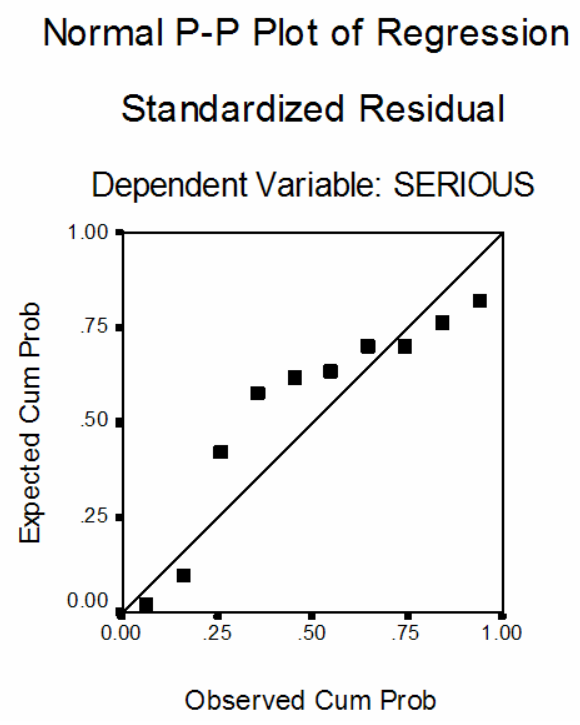
Case Number	Std. Residual	SERIOUS	Predicted Value	Residual
1	.522	21	18.15	2.85
2	.708	28	24.14	3.86
3	.305	27	25.33	1.67
4	.341	26	24.14	1.86
5	.528	33	30.12	2.88
6	.200	36	34.91	1.09
7	-2.030	31	42.09	-11.09
8	-1.298	35	42.09	-7.09
9	-.199	41	42.09	-1.09
10	.923	95	89.96	5.04

a. Dependent Variable: SERIOUS

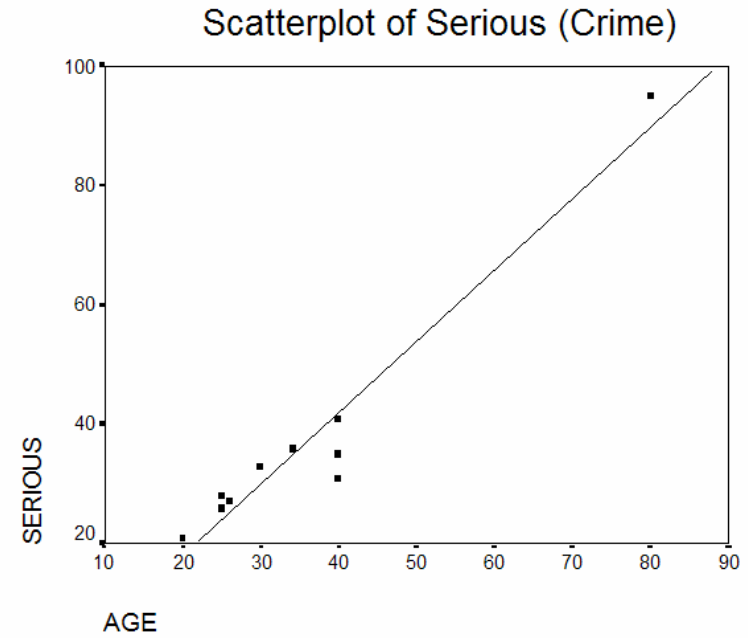
In order to check the normality assumption, we could also look at a histogram of the standardized residuals, which should display a normal pattern. Characteristics of the normal distribution include a bell shaped distribution, symmetry about zero, and 68% of the data within ±1 standard deviation, 95% of the data within ±2 standard deviations and 99% of the data within ±3 standard deviations. Although difficult to judge with 10 observations, the graph on the following page looks reasonable.



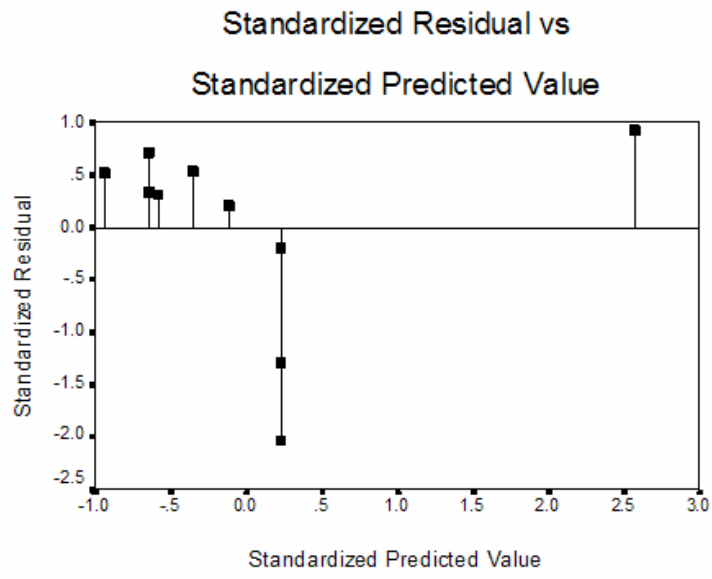
The probability plot for normal data should approximate a line (see section 3.7). The boxes in the figure below represent the data. The boxes fall a bit off of the line. This might suggest using a different model for the data; for example, a polynomial.



A scatterplot of the data suggests the line is a reasonable model, although there is one point isolated from the rest (x = 80, an elderly person). Note this plot is identical to the SCATTERGRAM output, except for the scale.



A plot of standardized $\hat{y}$ versus e_i indicates the model is fitting reasonably since there is a band (see Figure 2 for an example) to the data, as is the situation in the casewise plot.



In summary, a line is a reasonable model for this problem; however, the probability plot suggests a different model might better satisfy the normality assumption. In other words, the researcher's assumption that a line describes the relationship between age and seriousness in the population may be false. There are no outliers; however, the graph of age vs. seriousness reveals a clustering of points with one isolated point. Problem 1 of the exercises, which introduces the concept of an influential observation, should help in understanding the importance of this point.

Since the data were collected without controlling the inputs, as previously discussed in Chapter 4, the researcher can make statements only concerning the association of age and seriousness. No cause-effect statements can be made concerning the variables.

7.13 COMPUTER IMPLEMENTATION USING SAS

Note that SAS commands end in a semicolon (;). The commands for this example are listed below. The output of SAS is similar to the SPSS package and will not be reproduced here; however, a brief explanation of the program commands is listed below.

```
DATA JUSTICE;
INPUT ID AGE SERIOUS;
CARDS;
1 20 21
2 25 28
…
…
…
10 80 95
PROC REG;
MODEL SERIOUS = AGE / R DW;
OUTPUT OUT=RESIDS P=YHAT R=RESID;
PROC PLOT;
PLOT RESID*(YHAT AGE);
PROC UNIVARIATE PLOT NORMAL;
VAR RESID;
```

The first line in the **DATA** statement that tells SAS to create a data set called **justice**. The **PROC REG** command tells the computer to perform a least squares fitting of the model specified in the **MODEL** statement. In this case, the model is $E(y|x) = \beta_0 + \beta_1 x$, where y = seriousness and x = age. The **R** and **DW** subcommands give the residuals and Durbin-Watson statistic. The **OUTPUT** line stores the residuals. Residual plots are produced by the **PLOT** procedure.

7.14 Exercises

For the questions listed below, fit the model $\boldsymbol{E(y|x) = \beta_0 + \beta_1 x}$. From the computer output:

a. Give the least squares estimates of β_0 and β_1.
b. Test the H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$ and state your conclusions clearly.
c. State R^2. What is s^2?
d. Give a 95% confidence interval for β_1.
e. Comment on how the residuals either satisfy or do not satisfy their assumptions.

1. Omit the 10th observation (80, 95) from the above data set and redo the analysis. Compare the two different analyses. Note that the model, in this case a line, is heavily influenced by observations extreme in x. Since these observations shift the line towards them, and therefore have small residuals, looking for large residuals (outliers) is not adequate. These observations are called ***influential*** observations and researchers usually use the scatterplot of x vs. y and knowledge of the psychological system to guide their analysis. Influential observations will be discussed in section 8.3.

2. J.B. Stroud *(Amer. J. Psychol.,* 1932, vol. 44, p. 721-731) had 18 subjects learn 12 pairs of monosyllabic nouns in serial order. Level of learning was measured (by the method of direct recall) after each presentation of the 12 pairs. P = successive tenths of the learning period; R = average number of pairs recalled. Fit $E(R|P) = \beta_0 + \beta_1 P$.

P	R
1	0.7
2	1.9
3	3.2
4	4.5
5	6.4
6	7.6
7	8.6
8	9.9
9	10.9
10	12.0

3. R.M. Bellows *(J. exp. Psychol.,* 1936, vol. 19, p. 716-731) studied the relationship between critical fusion frequency for intermittent puffs of air, applied to the lower lip, and length of time the stimulus was applied. The values are averages for three observers. Fit $E(F|T) = \beta_0 + \beta_1 T$.

time (min.)	critical fusion frequ.	time (min.)	critical fusion frequ.
1	156.2	11	104.9
2	152.7	12	97.5
3	149.3	13	89.7
4	147.8	14	87.7
5	142.8	15	84.2
6	137.9	16	80.8
7	129.1	17	76.4
8	120.2	18	73.9
9	114.8	19	70.4
10	110.3	20	64.0

4. J.A. Gilbert *(Yale Psychol. Stud.,* 1894, vol. 2, p. 40-100) measured the heights of large groups of children (both boys and girls) ranging in age from 6 to 17 years. A = age; H = average height in inches. Fit $E(H|A) = \beta_0 + \beta_1 A$.

A	H
6	45.4
7	47.3
8	49.3
9	51.2
10	53.0
11	55.9
12	57.3
13	59.9
14	60.9
15	62.7
16	64.9
17	65.6

5. J.A. Gilbert *(Yale Psychol. Stud.,* 1894, vol. 2, p. 40-100) measured the weights of large groups of girls ranging in age from 6 to 17 years. A = age; W = average weight in pounds. Fit $E(W|A) = \beta_0 + \beta_1 A$.

A	W
6	44.3
7	50.4
8	53.0
9	58.8
10	62.7
11	70.0
12	84.5
13	92.0
14	98.0
15	104.0
16	113.0
17	113.7

6. H. Ebbinghaus *(Memory*. New York: Teachers College, Columbia University, 1913, p. 56) ran a study of retention as a function of the number of repetitions. He used 70 double lists composed of 6 series of 16 nonsense syllables each, some of which he read 8 times, some 16 times, some 24 times, etc. Retention was determined by the method of relearning, 24 hours after original reading. Fit $E(R|S) = \beta_0 + \beta_1 S$. The data are as follows:

R(# of repetitions)	S(% saved)
8	8
16	15
24	23
32	32
42	45
53	54
64	64