## CHAPTER 9: LINEAR MODELS WITH SEVERAL VARIABLES

Suppose we have a single dependent variable $y$ and m independent variables. Further, let $y = f(x_1$ $,\ldots,x_m)$ denote the functional relationship between $(x_1 ,\ldots,x_m)$. Note that we are in a different situation than the one discussed in the previous chapter. Previously with only one single independent variable, one could graph the data ($x$ vs. $y$) in order to help determine the type of model or degree of polynomial necessary to model the system. With m independent variables the graph is an m + 1 dimensional surface. The surface may be visualized when m is small (say 2), but in cases of many $x$'s the researcher may not be able to visualize the surface and must depend on the techniques we will describe here to help determine the important independent variables.

### 9.1 QUANTITATIVE VARIABLES (CONTINUOUS)

Suppose m = 1 and $x$ is a quantitative variable. Further suppose the relationship between $y$ and $x$ can be approximated by a polynomial relationship (Chapter 8). $y$ is a function of $x$ as described by

$$\boxed{\begin{aligned} y &= f(x) \\ &= \beta_0 + \beta_1 x^1 + \ldots + \beta_k x^k \\ &= x_{1xk}{}' \beta_{kx1} \end{aligned}}$$

**Example One**

Now suppose we have 2 continuous quantitative variables, $x_1$ and $x_2$. Assume that $f$ can be approximated by a polynomial of degree 1 in $x_1$, for example:

$$E[y|x] = \beta_0 + \beta_1 x_1$$

and a polynomial of degree 2 in $x_2$, for example:

$$E[y|x] = \beta_0 + \beta_2 x_2$$

Then we can write

$$\begin{aligned} y &= f(x_1, x_2) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ &= x'_{1x2}\, \beta_{2x1} \end{aligned}$$

which gives the full model that describes the relationship between $y$ and $x_1$ and $x_2$. Note there is no curvature (i.e. $x^2$, $x^3$) in this model.

The model of degree 1 can be written as

$$\boxed{E\left(y \mid x\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2}$$

whereas before the $\beta$'s represent the unknown population parameters that are estimated by the b's using least squares.

**$\beta_0$ represents the $y$ intercept of this 3-dimensional surface.**

**$\beta_1$ is the change in E($y$) for a one unit increase in $x_1$ when $x_2$ is held constant.**

**$\beta_2$ is the change in E($y$) for a one unit increase in $x_2$ when $x_1$ is held constant.**

The interpretation of $\beta_1$ is given in Figure 9.1 below. If $x_2$ is held constant, then the model is $E[y|x] = \beta_0 + \beta_1 x_1$ , a line for each value of $x_2$. $\beta_1$ is the change in $y$ per unit increase in $x_1$. The graph below gives the relationship for $x_2 = 4$ and $x_2 = 5$.

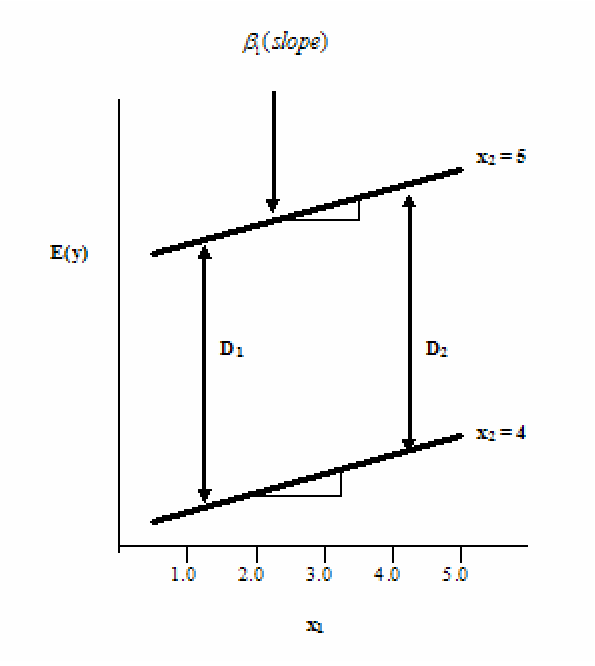


Figure 9.1  No Interaction

Note that the lines will be parallel, or using the notation in Chapter 5, for each value of $x_1$ the same change in $x_2$ produces the same change in $y$. In case of an interaction of the variables $x_1$ and $x_2$ , the change in $y$ will depend on $x_1$ and $x_2$ .

Clearly there is no curvature in the surface (i.e. higher degree of polynomial terms) and the $x$'s effect $y$ independent of one another. In other words, the variables do not interact. Exercise one will show the parallelism of surfaces when the variables are independent. Another way of saying this using the terminology in Chapter 5, is for each value of $x_2$ the same change in $x_1$ produces the same change in $y$, or $D_1 = D_2$.

**Example One (continued)**

Consider a related model where model one is as before

$$E[y|x] = \beta_0 + \beta_1 x_1$$

and model two is now a polynomial of degree 2 in $x_2$,

$$E(y \mid x) = \beta_0 + \beta_2 x_2 + \beta_3 x_2^{\ 2}$$

This permits curvature in the model. We can then write the overall model as

$$\boxed{\boldsymbol{E(y \mid x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^{\ 2}}}$$

The model and interpretation is given in Figure 9.2. If we examine the plot on the next page of $y$ vs. $x_2$ for constant $s_1$ we clearly see the parallelism and curvature.

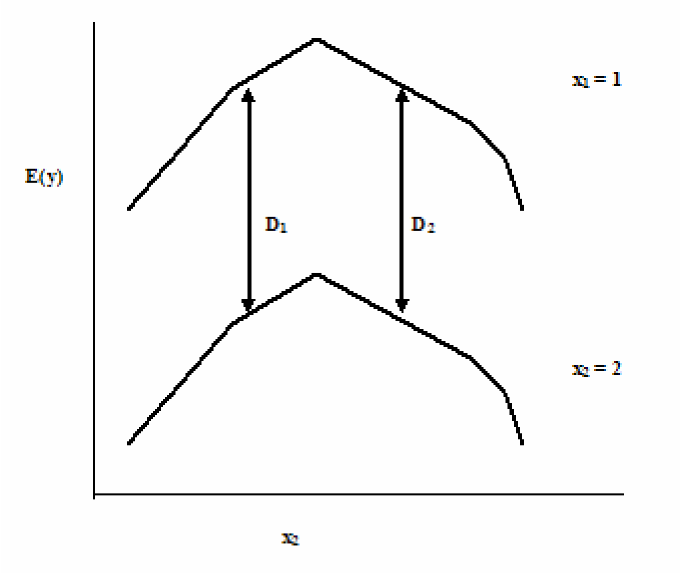


Figure 9.2 No Interaction

The model is $E\left(y\,|\,x\right)=\beta_0+\beta_2x_2+\beta_3x_2^{\;2}$, a quadratic for each vale of $x_1$. Note that once again the lines are parallel, or in the notation of Chapter 5, for each value of $x_1$ the same change in $x_2$ produces the same change in $y$. In other words, $D_1$ , or difference one, is equal to difference two, or $D_2$, for all values of $x_2$ ($D_1 = D_2$).

**Example One (Continued)**

Suppose further that both $x_1$ and $x_2$ interact, for example:

$$\beta_4x_1s_2$$

Then we can write

$$\boxed{\begin{aligned} y &= f(x_1, x_2) \\ &= \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_2^{\,2}+\beta_4x_1x_2 \\ &= x'_{1x4}\beta_{4x1} \end{aligned}}$$

as the full model describing the relationship between the dependent variable $y$ and the independent variables $x_1$ and $x_2$.

Note that the effect of $y$ of a change in one variable will depend on the other. For example, Figure 9.3 below describes an interaction.

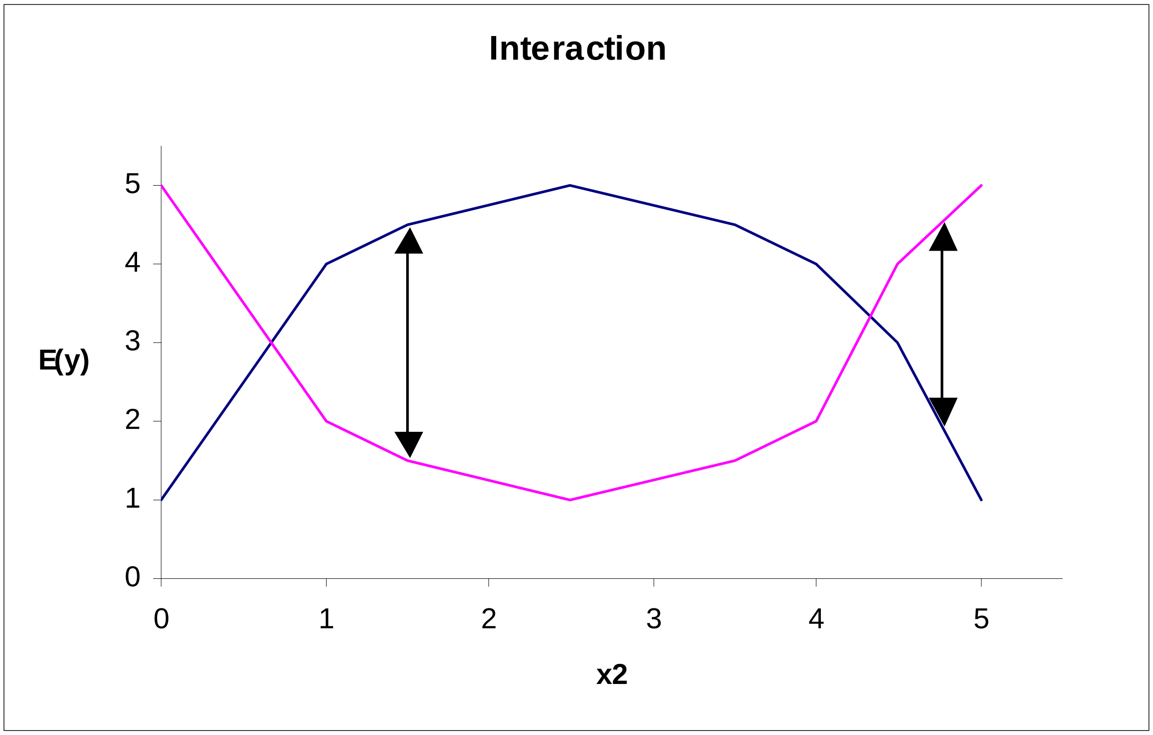


Figure 9.3 Interaction

Note that the lines are not parallel, and changes in E($y$) depend both on $x_1$ and $x_2$. In other words, $D_1 \neq D_2$ in the case of an interaction.

**SUMMARY**

In summary, the introduction of the degree 2 term ($x_2$) allows the surface curvature rather than restricting it to straight planes. The introduction of the interaction term permits the planes to be non parallel; in other words, the change in E($y$) will depend on the values of both $x_1$ and $x_2$. The effect of a unit change in $x_1$ will depend on the level of $x_2$. Exercise one will demonstrate this effect. Using the terminology in Chapter 5, we say that for all $x_1$ the same change in $x_2$ does not produce the same change in $y$.

Researchers combine quantitative variables using polynomial and interaction expressions in order to have maximum flexibility when trying to model a system.

**Example 1**

A researcher wishes to study the public's satisfaction with some community organizations ($y$). Two independent variables are considered: the amount of volunteer time at each organization in hours ($x_1$) and the amount of money spent on each organization in thousands of dollars ($x_2$). The psychologist suspects, from previous research, that satisfaction is related linearly with volunteer time and in a quadratic fashion with dollars spent. There is also the possibility of an interaction between volunteer time and dollars spent. The data are listed below.

| $y$ | $x_1$ | $x_2$ |
|---|---|---|
| 4.2 | 60 | 4.5 |
| 6.5 | 70 | 3.2 |
| 3.1 | 80 | 6.1 |
| . | . | . |
| . | . | . |
| . | . | . |

The linear model is given by

$$y = X\beta$$

where an individual response is described by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^{\,2} + \beta_4 x_1 x_2$$

The following columns are also calculated in order to fit the above model.

| $x_2^2$ | $x_1x_2$ |
|---|---|
| $(4.5)^2 = 20.25$ | 60 x 4.5 = 270.0 |
| $(3.2)^2 = 10.24$ | 70 x 3.2 = 224.0 |

The model is therefore written as

$$y = X\beta$$

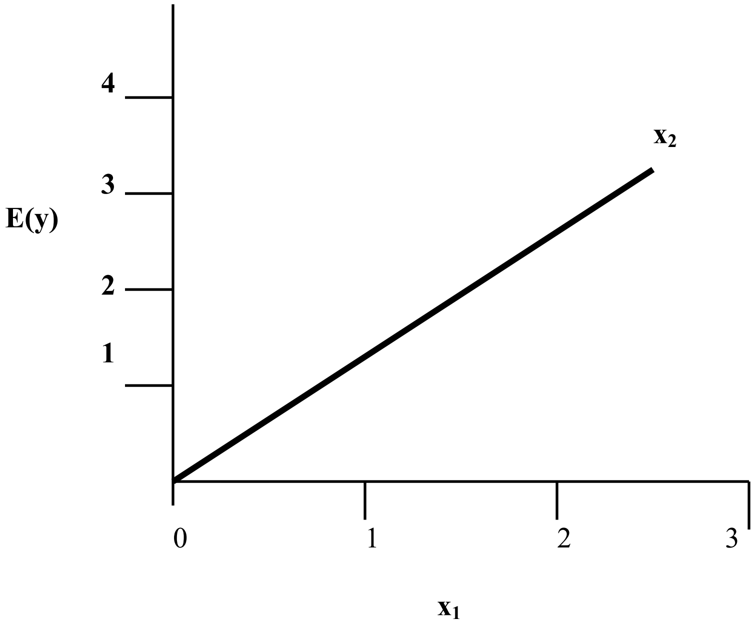
$$\begin{bmatrix} 4.2 \\ 6.5 \\ 3.1 \\ . \\ . \end{bmatrix} = \begin{bmatrix} 1 & 60 & 4.5 & 20.25 & 270.0 \\ 1 & 70 & 3.2 & 10.24 & 224.0 \\ 1 & 80 & 6.1 & 48.61 & 480.8 \\ . & . & . & . & . \\ . & . & . & . & . \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_{11} \\ \beta_{21} \\ \beta_{22} \\ \beta_{12} \end{bmatrix}$$

Clearly we can generalize this to the situation where we have m quantitative variables and obtain an equation of the form

$$\begin{aligned} y &= f(x_1,\ldots,x_m) \\ &= X\beta \end{aligned}$$

**9.2 GEOMETRY OF THE MODEL**

A geometric interpretation of the parameters in the model $E[y|x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ is given below.



- $\beta_0$ is the $y$-intercept (the value of $E(y|x)$ when $x_1 = x_2 = 0$) of this three dimensional surface.
- $\beta_1$ is the change in $E(y|x)$ for 1 unit increase in $x_1$ when $x_2$ is held fixed.
- $\beta_2$ is the change in $E(y|x)$ for 1 unit increase in $x_2$ when $x_1$ is held fixed.

**Exercises**

1. Consider the model $E(y|x) = 2 + x_1 + 3x_2$.

   a. Graph $y$ for $x_2$, $x_1 = 1, 2, 3$. Notice the shape of the surface.

   b. Fix $x_1 = 1$, and sketch a graph of $x_2$ vs. $E(y|x)$ for $x_2 = 1, 2, 3$.
   Fix $x_1 = 2$, and sketch a graph of $x_2$ vs. $E(y|x)$ for $x_2 = 1, 2, 3$.
   Fix $x_1 = 3$, and sketch a graph of $x_2$ vs. $E(y|x)$ for $x_2 = 1, 2, 3$.
   Plot all three on one graph and show that $\beta_{21}$ above is defined correctly.

2. Add the interaction term $2\, x_1\, x_2$ to the model in 1. Redo a and b. Notice that $\beta_3$ of the model $E(y|x) = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_3 x_1\, x_2$ is the interaction term that controls the shape of the surface.

## 9.3 QUALITATIVE VARIABLES

In contrast to the previous discussion of quantitative variables, consider the problem of modeling a qualitative variable. The variable(s) in the linear model that represent qualitative information are called ***dummy*** or indicator variables. Researchers usually use 1 and 0 as the coding system for these variables since this makes parameter interpretation of the model easy. The following method is introduced to help understand the experimental design sections of Part III.

Suppose m = 1 and $x$ is a qualitative variable taking k levels. Let $x_i = 1$ if $x$ takes $i^{th}$ level and $x_i = 0$ otherwise.

The variables, $x_i$ , are called ***dummy variables***.

We can write

$$\boxed{\begin{aligned} y &= f(x) \\ &= f(x_1,\ldots,x_k) \\ &= \beta_1 x_1 + \ldots + \beta_k x_k \\ &= x'_{1xk}\beta_{kx1} \end{aligned}}$$

where

$\beta_i$ = $f(0,\ldots 1, 0,\ldots 0)$,

$x_{kx1} = (x_1,\ldots,x_k)$,

$\beta_{kx1} = (\beta_1,\ldots, \beta_k)$

***then $\beta_i$ is the value of y when x is at its $i^{th}$ level.***

In summary, we have:

$$E(y \mid x) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

where $x_i$ is the dummy variable for the $i^{th}$ level.

$x_i$ = 1 if $x$ takes its $i^{th}$ level

= 0 *otherwise*

then

$E(y) = \beta_1$ for x at level 1( $\mu_1$ )

$E(y) = \beta_2$ for x at level 2( $\mu_2$ )

•

•

•

$E(y) = \beta_k$ for x at level k( $\mu_k$ )

For a variable with **k** levels we have

**k** dummy variables.

This approach is sometimes called the ***cell means*** parameterization.

**Example 1**

A researcher is interested in sex differences in performance on a cognitive task. The independent variable is sex ($x$) with two levels (k = 2). The dependent measure is time in seconds to complete the task ($y$). The data is given below.

| **sex** | ***y*** |
|---|---|
| M | 20 |
| F | 10 |
| M | 16 |
| F | 14 |
| F | 22 |

The linear model which describes the functional relationship between $x$ and $y$ is given by

$$y = f(x) = \beta_1 x_1 + \beta_2 x_2$$

where $x_1$ is a dummy variable having the value 1 if the subject is male and 0 if not male. The variable $x_2$ is 1 if the subject is female and 0 if not female. In the above example,

| $X_1$ | $X_2$ |
|---|---|
| 1 | 0 |
| 0 | 1 |
| 1 | 0 |
| 0 | 1 |
| 0 | 1 |

The design matrix X is written as

$$\mathrm{X} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

The linear model is given by $y = X\beta$

$$\begin{bmatrix} 20 \\ 10 \\ 16 \\ 14 \\ 22 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

where $\beta_1$ is the value of $y$ when the subject is male and $\beta_2$ is the value of $y$ when the subject is female.

Note the clear interpretation of the parameters in the model since $\beta_1$ represents the mean value of $y$ for men and $\beta_2$ represents the mean value of $y$ for women.

In summary, suppose m = 2 and $x_1$ and $x_2$ are qualitative variables taking $k_1$ and $k_2$ levels, respectively. Let $x_{ij}$ = 1 if $x_1$ takes its i^th^ level and $x_2$ takes its j^th^ level and $x_{ij} = 0$ otherwise. Then we can write

$$\boxed{\begin{aligned} y &= f(x_1, x_2) \\ &= f(x_{11},\ldots, x_{k1}x_{k2}) \\ &= \beta_{11}x_{11} + \ldots + \beta_{k1k2}x_{k1k2} \\ &= x'\beta \end{aligned}}$$

where $\beta_{ij}$ is the value of $y$ when $x_1$ is at its $i^{th}$ level and $x_2$ is at its $j^{th}$ level.

**Example 2**

A social psychologist is interested in examining self-esteem in the handicapped. Two independent variables are seen as important: sex ($x_1$) with two levels ($k_1 = 2$) and person's status ($x_2$) of whether they are handicapped or not – a control with two levels ($k_2 = 2$). An observation from the above system can be denoted by $x_{ij}$ , where $i = 1, 2$ denotes the person's sex, and $j = 1, 2$ denotes the person's status. The data are listed on the next page.

| $x_1$ | $x_2$ | $y$ |
|---|---|---|
| F | H | 7 |
| M | H | 9 |
| F | C | 4 |
| F | C | 2 |
| M | C | 6 |

The linear model for an individual's response is given by

$$y = \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_{21}x_{21} + \beta_{22}x_{22}$$

where

- $\beta_{11}$ is the mean value of $y$ when the person is a man and handicapped.
- $\beta_{12}$ is the value of $y$ when the person is a man and not handicapped.
- $\beta_{21}$ is the value of $y$ when the person is a woman and handicapped.
- $\beta_{22}$ is the value of $y$ when the person is a woman and not handicapped.

The design matrix X is

$$\mathrm{X} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

The linear model for the system is therefore

$$y = X\beta$$

$$\begin{bmatrix} 7 \\ 9 \\ 4 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_{11} \\ \beta_{12} \\ \beta_{21} \\ \beta_{22} \end{bmatrix}$$

## 9.4 SUMMARY

We notice that in the above situations for both quantitative and qualitative variables we can write the relationships between $y$ and the independent variables $x_1,\ldots, x_k$ (which are perhaps functions of the original variables) in the form

$$y = \beta_1 x_1 +\ldots+ \beta_k x_k$$

In the statistical context, we make the assumption that

$$\beta_1 x_1 +\ldots+ \beta_k x_k$$

gives the relationship between the location of the frequency distribution for $y$ and $x_1,\ldots, x_k$, i.e.

$$\boxed{\mathbf{E[y|x_1,\ldots, x_k] = \beta_1 x_1 + \beta_2 x_2 +\ldots+ \beta_k x_k}}$$

Read $E[y|x_1,\ldots, x_k]$ as the average value of $y$ given $x_1,\ldots, x_k$ , and the form of the frequency distribution is otherwise fixed. Thus, changes in $x_1,\ldots, x_k$ change, at most, the ***location*** of the frequency distribution. The form of the frequency distribution is assumed to be normal and the location is given by the mean.

## 9.5 COMPARING A SEQUENCE OF VARIABLES

Suppose we have a single qualitative variable $x_1$ taking $k_1$ levels and a single quantitative variable $x_2$ , such that for each level of $x_1$ the relationship between $y$ and $x_2$ can be well approximated by a polynomial in $x_2$ of degree at most $k_2$. Then we can write

$$\boxed{\begin{array}{l} y = f(x_1, x_2) = \beta_{11} x_{11} +\ldots+ \beta_{k1k2} x_{k1k2} \\ y = x\beta \end{array}}$$

where $x_{ij} = x_i^* x_2^{j-1}$ and $x_i^* = 1$ if $x_1$ is at level i, and is zero otherwise.

**Example:**

$x_1$ = sex

$x_2$ = income

$y$ = satisfaction index

Model: $E(y) = \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_{21}x_{21} + \beta_{22}x_{22}$

**Observations:**

| $x_1$ | $x_2$ | $y$ |
|---|---|---|
| M | 4 | 2 |
| F | 7 | 5 |
| F | 3 | 7 |
| M | 9 | 1 |

| $x_1^*$ | $x_2^*$ | $x_2^0$ | $x_2^1$ |
|---|---|---|---|
| 1 | 0 | 1 | 4 |
| 0 | 1 | 1 | 7 |
| 0 | 1 | 1 | 3 |
| 1 | 0 | 1 | 9 |

| $x_{11}$ | $x_{12}$ | $x_{21}$ | $x_{22}$ |
|---|---|---|---|
| 1 | 4 | 0 | 0 |
| 0 | 0 | 1 | 7 |
| 0 | 0 | 1 | 3 |
| 1 | 9 | 0 | 0 |

$$\text{Therefore, } x = \begin{bmatrix} 1 & 4 & 0 & 0 \\ 0 & 0 & 1 & 7 \\ 0 & 0 & 1 & 3 \\ 1 & 9 & 0 & 0 \end{bmatrix}, y = x\beta$$

where $y = \begin{bmatrix} 4 \\ 7 \\ 3 \\ 9 \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_{11} \\ \beta_{12} \\ \beta_{21} \\ \beta_{22} \end{bmatrix}$

## 9.6 TESTS OF SIGNIFICANCE

To test the hypothesis

$H_0$: $\beta = 0$

$H_a$: $\beta \neq 0$

we use the *t* statistic

$$\boxed{\begin{aligned} T_{n-k} &= (b-\beta)\sqrt{\frac{s^2}{\sum\left(x_1 - \bar{x}\right)^2}} \\ &= \sqrt{\frac{(b-\beta)}{s^2 * x^{ii}}} \end{aligned}}$$

where $s^2 = \dfrac{\sum (y_i - \hat{y})^2}{(n-k)}$

$= MSE$

and $(X'X)^{-1} = (x^{11}\ x^{22} \ldots\ x^{kk})$, the diagonal elements of this matrix, which is distributed student (*n*-k) under $H_0$. Then we compute the observed level of significance (OLS or *p*-value), *P*(|student (*n*-k)| ≥ *t*), and assess accordingly.

To test the hypothesis

$H_0$: $\sigma^2 = \sigma_0^2$

$H_a$: $\sigma^2 \neq \sigma_0^2$

we compute the OLS or *p*-value

$P$ (chi $(n–\mathrm{k}) \geq (n\text{-k})\ s^2 / \sigma_0^2$

since $(n\text{-k})\ s^2 / \sigma_0^2 \sim$ chi (*n*-k) under $H_0$.

## 9.7 DUMMY VARIABLES - ANOTHER PARAMETERIZATION

We have previously discussed, in Example 1, one parameterization of qualitative variables, one that will give a clear interpretation to the parameters of the model (β) when applied to experimental design examples that will be discussed in Part 3. In other words, the $\beta$'s represent the means of the treatment, but more on this later.

Another popular method of modeling qualitative variables is given below. In Example 1 of Section 8.2, we could have examined the model

$$E[y|x] = \beta_0 + \beta_1 x$$

where $x = 1$ if male and $x = 0$ if female. Only one dummy variable is required in this model. The interpretation of the parameters, however, is not as straight forward as before.

For Men, the model is $E[y|x] = \beta_0 + \beta_1$
For Women, the model is $E[y|x] = \beta_0$

$\beta_1$ represents the mean value of $y$ (seconds to complete the task) for women, whereas $\beta_2$ represents the difference in the mean of men and women.

$$\begin{aligned}\beta_2 &= (\beta_0 + \beta_1) - \beta_1 \\ &= E[y|\text{men}] - E[y|\text{women}] \\ &= \mu_m - \mu_w\end{aligned}$$

Notice that the $\beta$'s in the model do not have the straightforward interpretation that was discussed previously.

This type of modeling can be extended easily to three levels ($A_1$, $A_2$, $A_3$), for example:

$$E[y|x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where $x_1 = 1$ if level $A_1$ ; 0 if not.
$x_2 = 1$ if level $A_2$ ; 0 if not.

Then $\beta_1$ is E($y$| $x$) = $\beta_0$ or the mean of $A_3$

$\beta_2$ is E($y$| $x$) = $\beta_0 + \beta_1$ or the mean of $A_1$ minus mean $A_3$

$\beta_3$ is E($y$| $x$) = $\beta_0 + \beta_2$ or the mean $A_2$ minus mean $A_3$

Click here for SPSS program details.

In general, the number of dummy variables required is one less than the number of levels. In summary, we have:

$$E(y \mid x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

where $x_i$ = 1 if E($y$) is mean for level i

= 0 *otherwise*

Note there are k-1 dummy variables or one less than the number of levels.

$E(y) = \beta_0$ when $x$ is at level 1($\mu_1$)

$E(y) = \beta_0 + \beta_1$ when $x$ is at level 2($\mu_2$)

.

.

.

$E(y) = \beta_0 + \beta_k$ when $x$ is at level k($\mu_k$)

The parameters are interpreted as:

$$\beta_2 = \mu_2 - \mu_1$$

$$\beta_3 = \mu_3 - \mu_1$$

.

.

.

$$\beta_k = \mu_k - \mu_1$$

We have discussed two methods of parameterization of the model in the case of qualitative variables. The first will be used extensively in Part III - Experimental Design of the text, whereas the second will be used more frequently in Part II –Linear Models  of the text, since most software uses this type of parameterization for non-design situations.

## 9.8 MULTIPLE REGRESSION EXAMPLE USING SPSS

Gebotys and Roberts (1989) were interested in examining the effects of two more variables on the seriousness rating of the crime. The following table provides the new information concerning the amount of TV news watched in hours per week and whether the person had been a previous victim of crime.

| $y$<br>**serious** | $x_1$<br>**age** | $x_2$<br>**amount of TV news watched (hrs/wk)** | $x_3$<br>**previous victim of crime (1=yes, 0=no)** |
|---|---|---|---|
| 21 | 20 | 4 | 1 |
| 28 | 25 | 5 | 1 |
| 27 | 26 | 5 | 1 |
| 26 | 25 | 4.5 | 1 |
| 33 | 30 | 6 | 0 |
| 36 | 34 | 7 | 0 |
| 31 | 40 | 5.5 | 1 |
| 35 | 40 | 6 | 0 |
| 41 | 40 | 7 | 0 |
| 95 | 80 | 9 | 0 |

The calculations are again performed on the computer. It is hypothesized that there may be a quadratic relationship between seriousness and age. The researchers' hypothesize that the following model is a reasonable one:

$$E(y \mid x) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3$$

The model includes age($x_1$, $x_1^2$), TV news ($x_2$), and victimization($x_3$).

In this case, we asked SPSS to fit the model

$$E(y \mid x) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^{\ 2} + \beta_3 x_2 + \beta_4 x_3$$

The SPSS output and interpretation is given below.

Click here for SPSS program details.

**Model Summary[b]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Durbin-Watson |
|---|---|---|---|---|---|
| 1 | .998[a] | .997 | .994 | 1.63 | 1.712 |

a. Predictors: (Constant), Age Squared (years squared), Previous Victim of Crime, Amount of TV News Watched (hrs/wk), Age (years)

b. Dependent Variable: Crime Seriousness

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 3980.808 | 4 | 995.202 | 374.364 | .000[a] |
| | Residual | 13.292 | 5 | 2.658 | | |
| | Total | 3994.100 | 9 | | | |

a. Predictors: (Constant), Age Squared (years squared), Previous Victim of Crime, Amount of TV News Watched (hrs/wk), Age (years)

b. Dependent Variable: Crime Seriousness

In order to determine whether the model is adequate we examine the ANOVA table. Note the degrees of freedom and *F*-statistic values.

$F = 374.364$

which has an *F* distribution with 4 (number of parameters - intercept = 5-1 = 4) and 5 (number of observations - number of parameters = 10-5 = 5) degrees of freedom (*df*). We reject the null hypothesis

$H_0$: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

$H_a$: $\beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \neq 0$

with a *p*-value equal to .001, 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 $\sigma^2$.

$$s^2 = MSE = 2.658$$

$$s = \sqrt{MSE} = 1.63$$

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

$$\begin{aligned} R^2 &= \frac{SSM}{SST} \\ &= \frac{3980.808}{3994.100} \\ &= .99667 \end{aligned}$$

In other words, 99.667% of the variance in seriousness is accounted for by the model (age, agesq, tvnews, victim).

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | 15.923 | 7.923 | | 2.010 | .101 | -4.443 | 36.289 |
| | Age (years) | -.939 | .234 | -.760 | -4.003 | .010 | -1.541 | -.336 |
| | Amount of TV News Watched (hrs/wk) | 4.834 | 1.397 | .337 | 3.460 | .018 | 1.243 | 8.425 |
| | Previous Victim of Crime | -.567 | 2.158 | -.014 | -.263 | .803 | -6.115 | 4.981 |
| | Age Squared (years squared) | 1.728E-02 | .002 | 1.444 | 8.709 | .000 | .012 | .022 |

a. Dependent Variable: Crime Seriousness

In the Variables in the equation section, the column variable lists the variables victim, agesq, tvnews, age, and constant, which refer to the variables associated with the parameters $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$, and $\beta_4$ in the model. The column labeled **B** gives the least squares ($b_0 = 15.923$, $b_1 = -.939$, $b_2 = .01728$, $b_3 = 4.834$, $b_4 = -.567$) estimator for $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$, and $\beta_4$. The equation is therefore

$$E(y|x) = 15.923 - .939x_1 + .01728x_1^2 + 4.834x_2 - .567x_3$$

The **Std. Error** column is the standard error for each of the parameters, for example:

$s(b_0) = 7.923$

$s(b_1) = .234$

$s(b_2) = .002$

$s(b_3) = 1.397$

$s(b_4) = 2.158$

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

$H_0$: $\beta_1 = 0$

$H_a$: $\beta_1 \neq 0$

$$T = \frac{b_2}{s(b_2)} = \frac{-.939}{.234} = -.4003 \; T = b_2 \, / \, s(b_2) = -.938544 \, / \, .234448 = -4.003$$

$H_0$: $\beta_2 = 0$

$H_a$: $\beta_2 \neq 0$

$T = 8.709$

$H_0$: $\beta_3 = 0$

$H_a$: $\beta_3 \neq 0$

$T = 3.460$

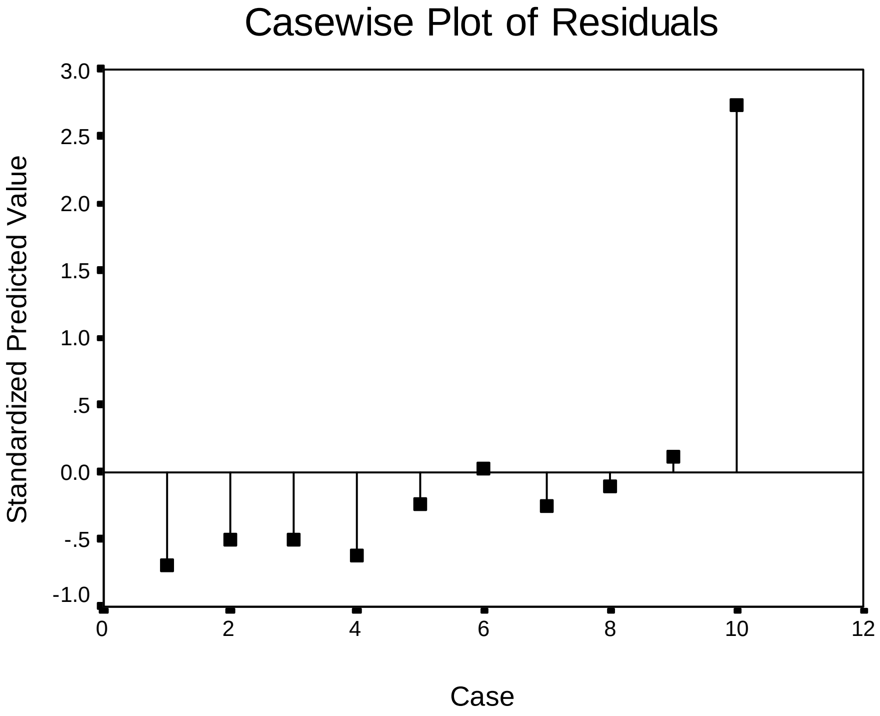
$H_0$: $\beta_4 = 0$

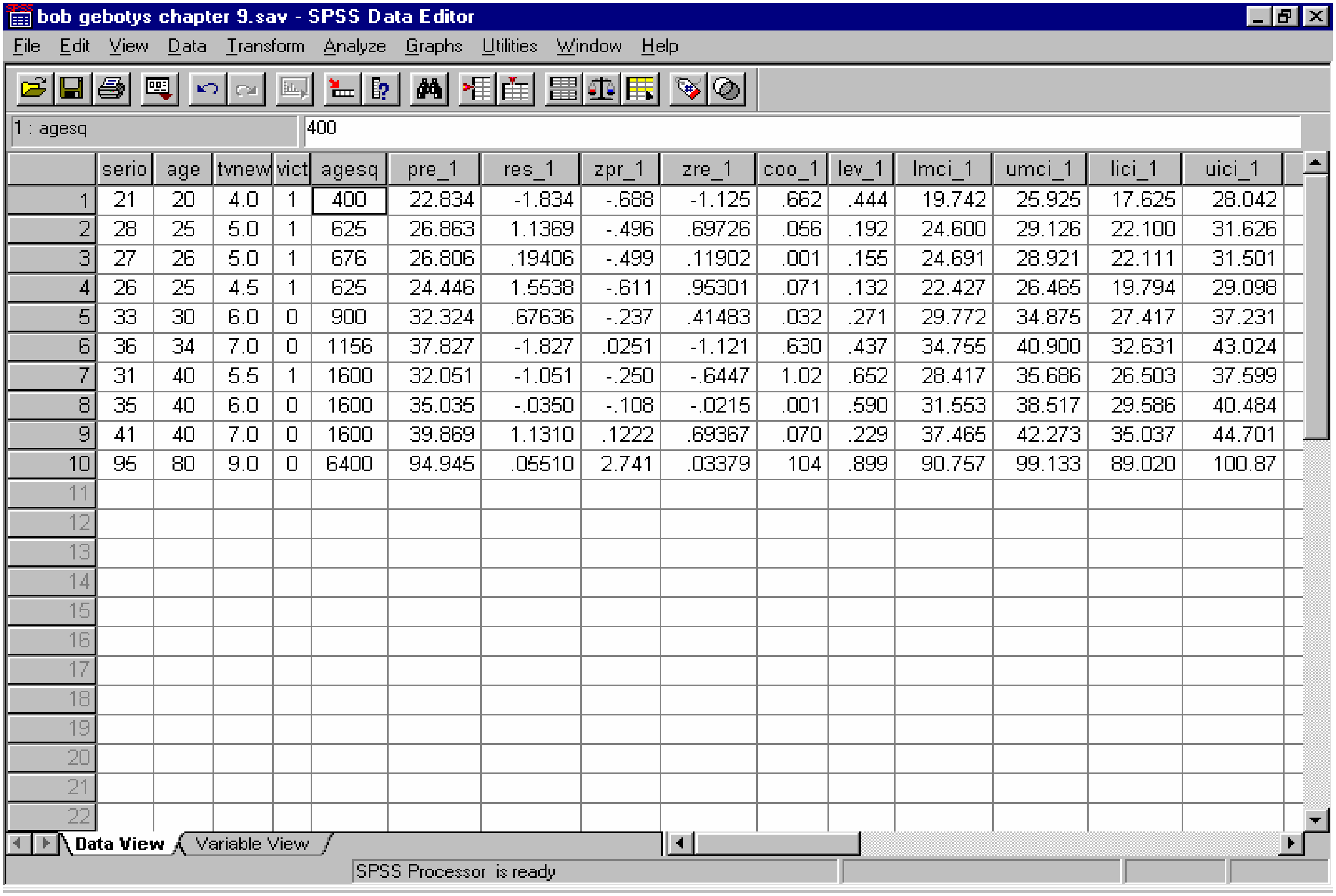
$H_a$: $\beta_4 \neq 0$

$T = -.263$

The column **SIG** gives the OLS or *p*-values for the tests above. In this case, we have $p = .010$ for $\beta_1$ (significant, therefore we reject $H_0$), $p = .001$ for $\beta_2$ (significant, therefore we reject $H_0$), $p = .018$ for $\beta_3$ (significant, therefore we reject $H_0$), and $p = .803$ for $\beta_4$ (not significant, therefore we cannot reject $H_0$). All are with 5 *df*. These statistics indicate age, agesq, and tvnews are all important in predicting seriousness of the crime, but victim (whether or not the participant had been a victim of crime) is not important.

The casewise plot of residuals looks reasonable in that it displays a band pattern over the range of values. The leverage (**LEVER**) and Cook's distance (**COOK D**) values for the 10th observation are large (leverage = .8985, Cook D = 104.14), indicating this is an influencial observation.





1 : agesq 400

| | serio | age | tvnew | vict | agesq | pre_1 | res_1 | zpr_1 | zre_1 | coo_1 | lev_1 | lmci_1 | umci_1 | lici_1 | uici_1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 21 | 20 | 4.0 | 1 | 400 | 22.834 | -1.834 | -.688 | -1.125 | .662 | .444 | 19.742 | 25.925 | 17.625 | 28.042 |
| 2 | 28 | 25 | 5.0 | 1 | 625 | 26.863 | 1.1369 | -.496 | .69726 | .056 | .192 | 24.600 | 29.126 | 22.100 | 31.626 |
| 3 | 27 | 26 | 5.0 | 1 | 676 | 26.806 | .19406 | -.499 | .11902 | .001 | .155 | 24.691 | 28.921 | 22.111 | 31.501 |
| 4 | 26 | 25 | 4.5 | 1 | 625 | 24.446 | 1.5538 | -.611 | .95301 | .071 | .132 | 22.427 | 26.465 | 19.794 | 29.098 |
| 5 | 33 | 30 | 6.0 | 0 | 900 | 32.324 | .67636 | -.237 | .41483 | .032 | .271 | 29.772 | 34.875 | 27.417 | 37.231 |
| 6 | 36 | 34 | 7.0 | 0 | 1156 | 37.827 | -1.827 | .0251 | -1.121 | .630 | .437 | 34.755 | 40.900 | 32.631 | 43.024 |
| 7 | 31 | 40 | 5.5 | 1 | 1600 | 32.051 | -1.051 | -.250 | -.6447 | 1.02 | .652 | 28.417 | 35.686 | 26.503 | 37.599 |
| 8 | 35 | 40 | 6.0 | 0 | 1600 | 35.035 | -.0350 | -.108 | -.0215 | .001 | .590 | 31.553 | 38.517 | 29.586 | 40.484 |
| 9 | 41 | 40 | 7.0 | 0 | 1600 | 39.869 | 1.1310 | .1222 | .69367 | .070 | .229 | 37.465 | 42.273 | 35.037 | 44.701 |
| 10 | 95 | 80 | 9.0 | 0 | 6400 | 94.945 | .05510 | 2.741 | .03379 | 104 | .899 | 90.757 | 99.133 | 89.020 | 100.87 |
| 11 | | | | | | | | | | | | | | | |
| 12 | | | | | | | | | | | | | | | |
| 13 | | | | | | | | | | | | | | | |
| 14 | | | | | | | | | | | | | | | |
| 15 | | | | | | | | | | | | | | | |
| 16 | | | | | | | | | | | | | | | |
| 17 | | | | | | | | | | | | | | | |
| 18 | | | | | | | | | | | | | | | |
| 19 | | | | | | | | | | | | | | | |
| 20 | | | | | | | | | | | | | | | |
| 21 | | | | | | | | | | | | | | | |
| 22 | | | | | | | | | | | | | | | |

SPSS Processor is ready

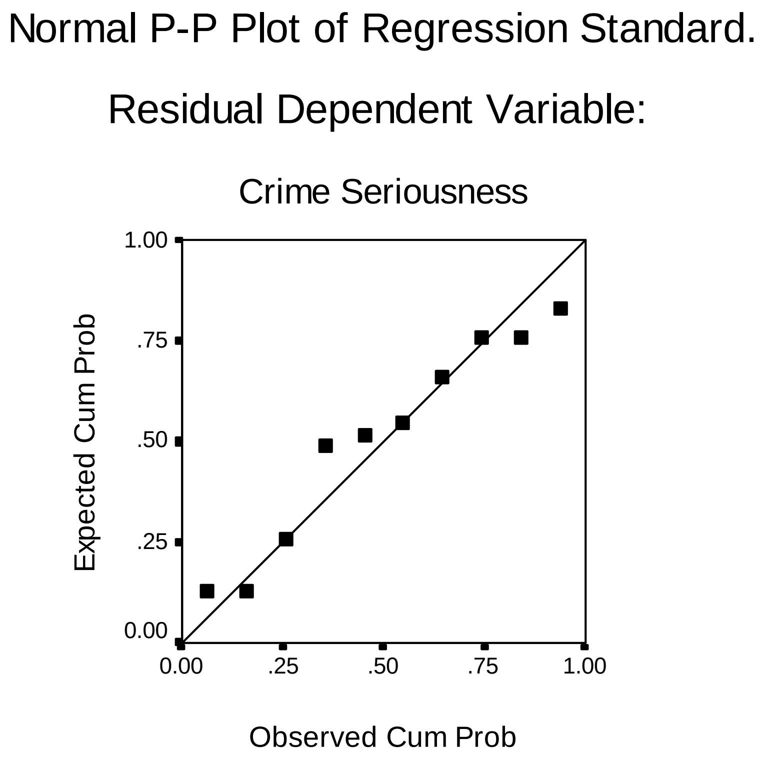
The summary statistics for a number of measures are given below. The average leverage is .4, which indicates the 10th observation is clearly influential. The Dubin Watson test indicates no serial correlation.

**Residuals Statistics**[a]

| | Minimum | Maximum | Mean | Std. Deviation | N |
|---|---|---|---|---|---|
| Predicted Value | 22.83 | 94.94 | 37.30 | 21.03 | 10 |
| Std. Predicted Value | -.688 | 2.741 | .000 | 1.000 | 10 |
| Standard Error of Predicted Value | .79 | 1.63 | 1.12 | .28 | 10 |
| Adjusted Predicted Value | 23.98 | 57.77 | 34.14 | 10.15 | 10 |
| Residual | -1.83 | 1.55 | 2.49E-15 | 1.22 | 10 |
| Std. Residual | -1.125 | .953 | .000 | .745 | 10 |
| Stud. Residual | -1.666 | 1.088 | -.034 | 1.096 | 10 |
| Deleted Residual | -4.24 | 37.23 | 3.16 | 12.24 | 10 |
| Stud. Deleted Residual | -2.233 | 1.113 | -.168 | 1.291 | 10 |
| Mahal. Distance | 1.189 | 8.087 | 3.600 | 2.271 | 10 |
| Cook's Distance | .001 | 104.149 | 10.669 | 32.848 | 10 |
| Centered Leverage Value | .132 | .899 | .400 | .252 | 10 |

a. Dependent Variable: Crime Seriousness

The residual plots look reasonable. The normal probability plot approximates a line.

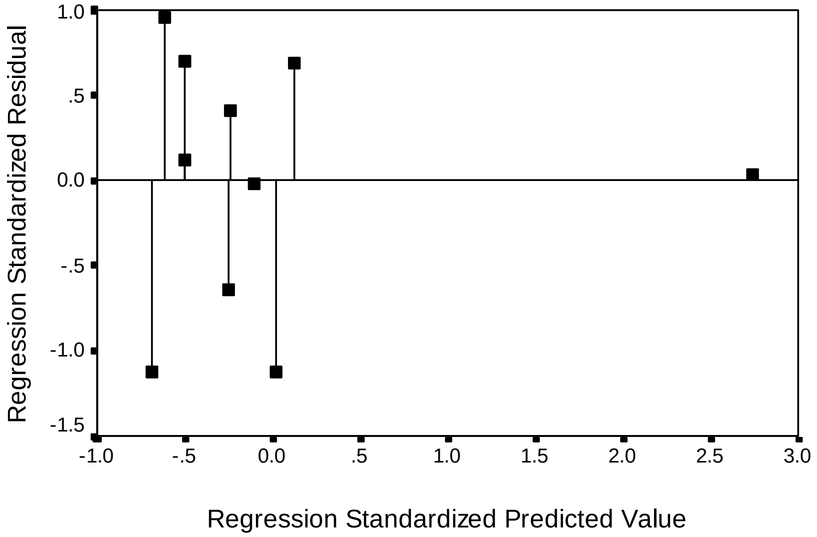


Normal P-P Plot of Regression Standard.

Residual Dependent Variable:

Crime Seriousness

The Standardized Scatterplot of residuals displays a reasonable band shaped pattern.


Scatterplot Dependent Variable:

Crime Seriousness

In summary, from our analysis we conclude that age ($x_1$, $x_1^2$) influences crime seriousness ($y$) in a quadratic manner; the more tvnews ($x_2$) a person is exposed to the more serious the crime rating ($y$), and victimization ($x_3$) has no impact on seriousness ratings ($y$). There is one influential observation, which is an 80 year old person that deserves further study. The residual plots indicate the normality assumption is reasonable.

**9.10 COMPUTER IMPLEMENTATION USING SAS**

```
DATA JUSTICE;
INPUT ID AGE SERIOUS TVNEWS VICTIM;
AGESQ = AGE**2;
CARDS;
1 20 21 4.0 1
2 25 28 5.0 1
….
```

```
…..
…..
10 80 95 9.0 0
PROC REG;
MODEL SERIOUS = AGE AGESQ TVNEWS VICTIM/R INFLUENCE DW;
OUTPUT OUT = RESIDS P = YHAT R = RESID;
PROC PLOT;
PLOT RESID*(YHAT,AGE);
PROC UNIVARIATE PLOT NORMAL;
VAR RESID;
```

**9.11 Exercises**

1. Fit the model $E[y|x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ to the above data.

   a. What is the equation for the model?
   b. Test the hypothesis $\beta_1 = \beta_2 = \beta_3 = 0$ in an ANOVA table.
   c. Test the hypothesis $\beta_i = 0$. Give a 95% confidence interval for $\beta_1$.
   d. What is $R^2$?
   e. Are the residual plots reasonable?
   f. State your conclusions concerning the model clearly.

2. Problem Clinical Psychology

Success of a counseling session ($Y$) is recorded along with the amount of paraphrasing ($X_1$) and amount of empathy ($X_2$) for 14 subjects in the following clinical study. Data are presented on the next page.

| **Subject** | ***Y*** | ***X*₁** | ***X*₂** |
|---|---|---|---|
| 1 | 14.7 | 8.9 | 31.5 |
| 2 | 48.0 | 36.6 | 27.0 |
| 3 | 25.6 | 26.8 | 25.9 |
| 4 | 10.0 | 6.1 | 39.1 |
| 5 | 16.0 | 6.9 | 39.2 |
| 6 | 16.8 | 6.9 | 38.3 |
| 7 | 20.7 | 7.3 | 33.9 |
| 8 | 38.8 | 28.4 | 33.8 |
| 9 | 16.9 | 6.5 | 27.9 |
| 10 | 27.0 | 18.0 | 33.1 |
| 11 | 16.0 | 4.5 | 26.3 |
| 12 | 24.9 | 19.9 | 37.8 |
| 13 | 7.3 | 2.9 | 34.6 |
| 14 | 12.8 | 2.0 | 36.4 |

Use SPSS to fit the following model:

$$E[y|x] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2$$

You will need several COMPUTE statements to compute values of $X_1^2$, $X_2^2$, and $X_1X_2$. For example, we compute $X_1^2$:

```
COMPUTE X1X2 = X1 * X1
```

which creates the new variable for which I have chosen (arbitrarily) the variable name X1X2.

a. What is the estimated regression equation? Do the residual plots suggest that the full model should be modified? Explain.

b. Test at $\alpha = .05$ the null hypothesis that there is no relationship between the dependent variable and the model in an ANOVA table. Test the five hypotheses that $\beta_i$ where I = 1, 2, 3, 4, 5 is equal to zero. What is $R^2$?

c. Assume the simpler model is adequate. It has been observed that the first 7 subjects were advised by clinician A, and the second set of 7 subjects was advised by clinician B. This qualitative variable may be entered into the model by including a variable $X_3$, which is set equal to 0 for the first 7 observations and equal to 1 for the next 7. Fit the model

$$E[y|x] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

and test whether the clinician makes a difference, i.e. does $\beta_3 = 0$? (Qualitative variables with more than 2 levels can also be entered into the regression models.)

3. Tucher (1987) and Darlington (1990) examined homelessness in the United States for 50 cities. The data for 30 cities is reported below. The dependent measure ($Y$) is the homelessness rate.

| **City** | $Y$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|---|---|---|---|
| *Miami* | 15.9 | 24.5 | 7.5 | 29.8 | 372 | 7.0 | 0 | 67 |
| *St.Louis* | 11.6 | 21.8 | 8.4 | 14.0 | 429 | 8.5 | 0 | 29 |
| *San Francisco* | 11.5 | 13.7 | 6.0 | 10.2 | 712 | 1.6 | 1 | 49 |
| *Worcester, Mass* | 10.6 | 14.4 | 3.7 | 14.1 | 160 | 3.0 | 0 | 25 |
| *Los Angeles* | 10.5 | 16.4 | 7.9 | 2.8 | 3097 | 2.2 | 1 | 57 |
| *Santa Monica* | 10.2 | 9.9 | 7.0 | 0.8 | 88 | 1.8 | 1 | 57 |
| *Newark, N.J.* | 9.5 | 32.8 | 5.9 | 41.7 | 314 | 2.3 | 1 | 31 |
| *Hartford* | 8.8 | 25.2 | 7.1 | 20.0 | 136 | 2.6 | 1 | 25 |
| *Washington, D.C.* | 7.5 | 18.6 | 8.4 | 19.8 | 623 | 2.0 | 1 | 31 |
| *Detroit* | 6.8 | 21.9 | 9.1 | 9.7 | 1088 | 5.4 | 0 | 23 |
| *Yonkers* | 6.8 | 9.8 | 4.9 | 10.7 | 191 | 2.1 | 1 | 32 |
| *Chicago* | 6.6 | 20.3 | 8.3 | 13.0 | 2992 | 6.0 | 0 | 21 |
| *Seattle* | 6.5 | 11.2 | 6.6 | 14.6 | 488 | 5.5 | 0 | 39 |

| City | $Y$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|---|---|---|---|
| *Las Vegas* | 6.0 | 10.5 | 8.9 | 14.2 | 183 | 9.0 | 0 | 44 |
| *Boston* | 5.6 | 20.2 | 4.6 | 25.3 | 571 | 2.6 | 1 | 30 |
| *Richmond* | 5.3 | 19.3 | 5.3 | 20.5 | 219 | 5.5 | 0 | 37 |
| *New York* | 5.0 | 20.0 | 7.4 | 21.5 | 7165 | 2.2 | 1 | 32 |
| *Dallas-Fort Worth* | 5.0 | 14.1 | 4.7 | 5.9 | 1388 | 6.0 | 0 | 44 |
| *Denver* | 4.9 | 13.7 | 5.0 | 9.0 | 504 | 4.0 | 0 | 30 |
| *Charleston, W. Va.* | 4.7 | 12.6 | 10.7 | 22.9 | 63 | 5.9 | 0 | 29 |
| *Atlanta* | 4.6 | 27.5 | 5.0 | 35.5 | 426 | 9.0 | 0 | 42 |
| *Fort Wayne* | 4.3 | 11.0 | 6.3 | 5.0 | 165 | 9.2 | 0 | 21 |
| *Portland* | 4.2 | 13.0 | 7.4 | 5.0 | 366 | 5.5 | 0 | 39 |
| *Houston* | 3.7 | 12.7 | 8.4 | 1.9 | 1706 | 7.0 | 0 | 51 |
| *San Diego* | 3.1 | 12.4 | 5.3 | 1.1 | 960 | 5.3 | 0 | 57 |
| *Salt Lake City* | 3.1 | 14.2 | 6.3 | 6.5 | 165 | 4.5 | 0 | 29 |
| *Little Rock* | 2.9 | 14.1 | 5.8 | 16.8 | 170 | 6.5 | 0 | 40 |
| *New Orleans* | 2.8 | 26.4 | 11.0 | 25.2 | 559 | 8.0 | 0 | 52 |
| *Charleston, S.C.* | 2.8 | 14.1 | 4.4 | 30.6 | 69 | 9.0 | 0 | 49 |
| *Albuquerque* | 2.8 | 12.4 | 6.3 | 3.1 | 351 | 9.7 | 0 | 35 |

$Y$ = homelessness rate; $X_1$ = poverty rate; $X_2$ = unemployment rate; $X_3$ = public housing rate; $X_4$ = population (thousands); $X_5$ = vacancy rate; $X_6$ = rent control (1 = rent control exists); $X_7$ = winter temperature.

a. Fit the model

$$E(y|x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_7 X_7$$

b. Is the model adequate?
Test $H_0$: $\beta_1 = \beta_2 = \ldots = \beta_7 = 0$ in an ANOVA table.

c. What variables are important in the model?
Test $H_0$: $\beta_i = 0$, i = 2,…, 7.

d. Are the residuals reasonable?

e. Add 3 two-way interaction terms to the model.
Explain the reasoning behind inclusion of each interaction term and test for their importance.