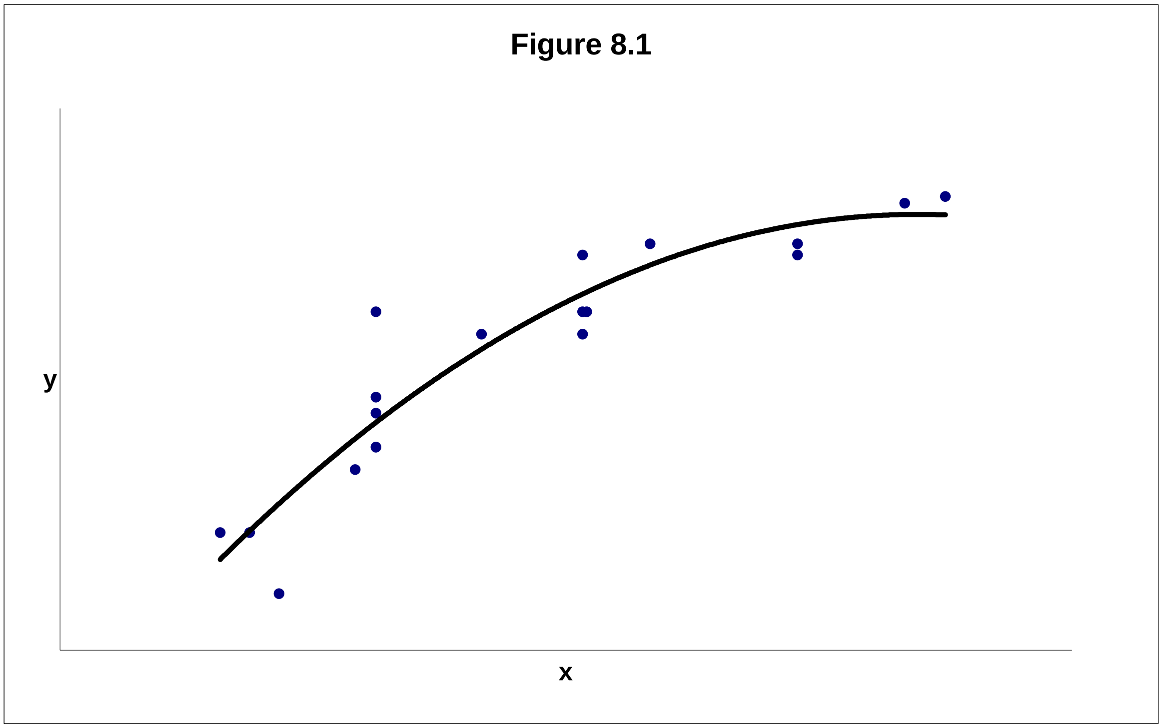
CHAPTER 8: POLYNOMIAL MODELS AND MATRIX NOTATION

8.1 POLYNOMIAL MODELS

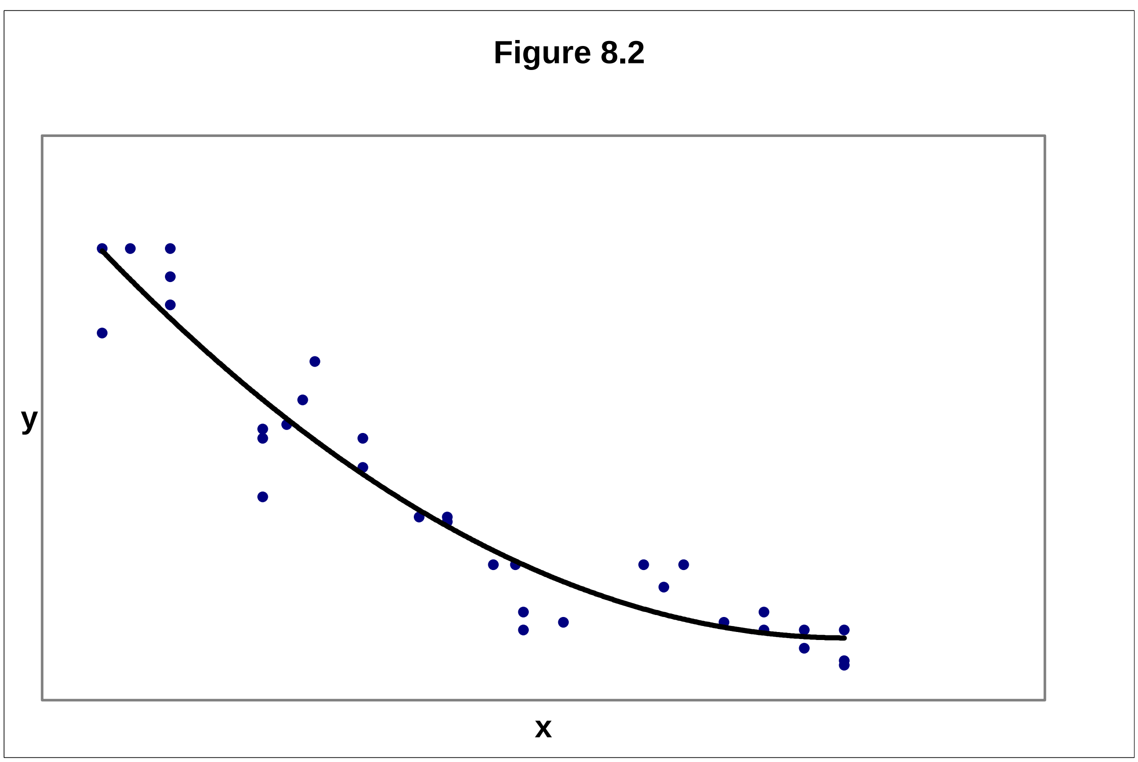
The methods we have used in the previous section can be easily extended to problems where the functions form of the relationship is not described by a line. Polynomial relationships (see section 6.12) are used by psychologists to model systems in a wide variety of areas. For models of this type researchers assume that the means of the y variable depend on x according to the model; which is called a polynomial of degree k-1.

$$E(y|x) = \beta_0 + \beta_1 x_1^{1} + \beta_2 x_1^{2} + \ldots + \beta_k x^{k}$$

From Chapter 7 we have seen that in the case of a line y changes by a constant amount with x. There are no curves in this polynomial of degree one. We have seen in 6.12 that polynomials of degree 2 will curve once; polynomial of degree 3 will curve twice, etc. A polynomial of degree k will fit k data points perfectly. When y changes at a constant rate or percentage with x researchers use exponential models (Chapter 13) to model the system under study. When y changes with x at a non-constant rate, then the researcher can consider a polynomial model. See the two examples in Figures 8.1 and 8.2.



y increases with x at a decreasing non-constant rate



y decreases with x at a decreasing non-constant rate

Polynomial models as we have seen previously in section 6.12 give the researcher a great deal of flexibility in modeling the relationship between x and y. No longer is there a restriction that a line models the system. Polynomial models that are quite common are quadratic (degree 2) and cubic (degree 3) relationships. For the sake of parsimony researchers when in the situation that a line does not adequately fit the data, usually try the highest polynomial. A graph of the data is very helpful in determining the degree of polynomial the researcher may wish to consider.

The principle of least squares is used as previously where

$b_0, b_1, \ldots, b_k$

represent the least squares estimators of

$\beta_0, \beta_1, \ldots, \beta_k$

the population parameters.

The predicted value of y is

$$\hat{y} = b_0 + b_1 x^1 + b_2 x^2 + \ldots + b_k x^k$$

where the residual is

$$\boxed{e_i = y_i = \hat{y}_i}$$

once again, the variance of the residuals is given by

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

Since there are (k) parameters estimated the degrees of freedom of s^2 is n - k. Note that there must be at least k+1 data points in order to estimate σ^2, the spread or scaling of the points about the model.

To test the hypothesis

$$H_o : \beta_i = 0$$
$$H_a : \beta i \neq 0$$
use
$$T = \frac{bi}{s(b_i)}$$

where T has a t distribution with n - k degrees of freedom. A common misconception is that the magnitude of the b indicates its importance in predicting y. This is clearly not the case as can be seen in the T statistic. The standard error influences both the test of hypothesis and confidence interval for the parameter.

The ANOVA table has the same structure as before

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

Note that the F statistic now tests the hypothesis

$$H_o : \beta_1 = \ldots \beta_k = 0$$
$$H_a : \beta_1 \neq \ldots \neq \beta_k \neq 0$$

and

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

is distributed with F(k-1, n-k) degrees of freedom under H_0. The Total DF = n-1 and the Model DF = k-1 since β_0, one parameter, is removed from the test of hypothesis. The reasoning for this is much the same as in the previous chapter, researchers wish to examine the effect of x on y and β_0 is usually not of interest.

The proportion of variance in y accounted for by the model using x is called R squared (See Chapter 7).

$$R^2 = \frac{SSM}{SST} = \frac{SSM}{SSM + SSE} = 1 - \frac{SSM}{SSE}$$

The F test can also be written in terms of R^2

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

$$= \frac{R^2 / k-1}{\left(1-R^2\right) / (n-k)}$$

Note that R^2 is defined in terms of sums of squares not mean squares. Clearly R^2 does not reflect the number of parameters or degrees of freedom in the model. As the number of parameters increases, R^2 increases. In fact, a researcher could force R^2 close to one by fitting polynomials of high degree. The question then would be; is the complex polynomial as parsimonious and meaningful a solution for the problem as a lower order polynomial (is a plynomial of degree 2 as reasonable as one of degree 10) and still be significant in the ANOVA test. The ***adjusted*** R^2 or $R_A{}^2$ is used to take into account the number of parameters in the model.

$$R_A{}^2 = 1 - (n-1)\frac{MSE}{SST}$$

In this case $R_A{}^2$ increases only is MSE decreases. It cannot be forced to 1 by adding more parameters to the model.

For example, the researchers that are investigating the relationship between age and seriousness (See Section 7.1) may suspect that a more complicated relationship exists between the two variables. Using the knowledge acquired from a literature review and a plot of x vs. y, these researchers wish to examine a polynomial of degree 2 which they suspect might be a better model for this system.

The model they fit to the data is

$$E(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

All of the above calculations are performed on the computer. For example, in section 8.3 we will see how to fit this model using SPSS.

8.2 INFLUENCIAL OBSERVATIONS

As noted before in Chapter 7, the model can be heavily influenced by extreme observations on the predictor variable. Since these observations shift the model towards them and have small residuals, looking for large residuals is not adequate. A number of techniques exist for identifying influencial observations.

Leverage

$\hat{y}_i$ can be written as a combination of $y_1, y_2, \ldots, y_n$.

$$\hat{y}_i = h_1 y_1 + h_2 y_2 + \ldots + h_n y_n$$

where h_i measures the influence of y_i on $\hat{y}_i$.

The value h_i is called the ***leverage*** of the i^{th} observation with respect to the values of the x's since $h_i = f(x)$.

The larger the leverage the more influence this observation has on the model. h_i is compared **to average leverage** $\bar{h} = {(k-1)}/{n}$ which can be calculated from the number of parameters in the model.

If $h_i > 2\bar{h}$ then this observation should be carefully examined within the context of the study. This will be illustrated with an example shortly.

Cook's Distance

A measure of overall influence an observation has on the estimation of β was proposed by Cook (1979). Cook's distance is defined as

$$D_i = \frac{e_i}{kMSE}\left[\frac{h_i}{(1-h_i)^2}\right]$$

A large value of D_i indicates a strong influence. Values of D can be compared to the F distribution with $df_1 = k$, $df_2 = n - k$ degrees of freedom. Observations that fall above the 50^{th} percentile are considered influencial.

Studentized Deleted Residual

The researcher deletes observations one at a time refitting the model with n-1 observations. The deleted residual

$$d_i = y_i - \hat{y}_{(i)}$$

where $\hat{y}_{(i)}$ indicates the predicted value of y with the i^{th} observation deleted. Large d_i indicate high influence. The ***Studentized deleted residual*** (***d_i****) is

$$\boxed{d_i^* = \frac{d_i}{s_{d_i}}}$$

where s_{d_i} is the standard error of the deleted residual d_i . We compare d_i* with a t distribution with n-k degrees of freedom.

SUMMARY

There are no set solutions for handling influential observations. Researchers might elect to collect more data to dampen the influence of these observations or remove these points and treat them as special cases. The context and nature of the system under study will assist the researcher in selecting a reasonable approach.

8.3 COMPUTER IMPLEMENTATION USING SPSS

The following output fits the model $E(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2$ to the Gebotys and Roberts data in section 7.1. A polynomial of degree 2 (quadratic) is used to describe the relationship between age and seriousness. Click here for the SPSS program details.

In this case we have asked SPSS to fit the model $E(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2$

Descriptive Statistics

	Mean	Std. Deviation	N
Crime Seriousness	37.30	21.07	10
Age (years)	36.00	17.07	10
Age Squared (years squared)	1558.20	1760.83	10

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Age Squared (years squared), Age (years)[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Crime Seriousness

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3896.297	2	1948.149	139.434	.000[a]
	Residual	97.803	7	13.972		
	Total	3994.100	9			

a. Predictors: (Constant), Age Squared (years squared), Age (years)

b. Dependent Variable: Crime Seriousness

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.988[a]	.976	.969	3.74	2.081

a. Predictors: (Constant), Age Squared (years squared), Age (years)

b. Dependent Variable: Crime Seriousness

The output is interpreted as follows.

Model

In order to determine if the model is adequate we examine the ANOVA table. Note the degrees of freedom and F-statistic values. F = 139.434 which has an F distribution with 2 (number of parameters [3] - intercept β_0 [1] = 3 - 1 = 2) and 7 (number of observations - number of parameters = 10 -3 = 7) degrees of freedom. We reject

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

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

with p-value less than .0001, 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.

$$s^2 = MSE = 13.972$$

$$s = \sqrt{MSE} = 3.74$$

note s is also printed in the **STD ERROR** column. In the same area we also have $\boldsymbol{R^2}$, R SQUARE printed where

$$R^2 = \frac{SSM}{SST} = \frac{3896.297}{3994.1} = .97551$$

In other words 97.551% of the variance in seriousness is accounted for by the model (age, agesq).

Variables

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	20.683	8.823		2.344	.052	-.180	41.545
	Age (years)	-8.49E-02	.410	-.069	-.207	.842	-1.055	.885
	Age Squared (years squared)	1.262E-02	.004	1.055	3.174	.016	.003	.022

a. Dependent Variable: Crime Seriousness

In the Variables in the equation section the column variable lists the variable age, agesq and constant these refer to the variables associated with the parameters β_0, β_1, and β_2 in the model. The column labeled **B** given the least squares ($b_0 = 20.683$, $b_1 = -.0849$, $b_2 = .01262$) estimator for β_0, β_1 and β_2. The equation is therefore $E(y|x) = 20.683 - .0849x + .01262x^2$. The **Std Error** column is the standard error for each of the parameters for example

$s(b_0) = 8.823$ from the Constant row

$s(b_1) = .410$ from the Age row

$s(b_2) = .004$ from the Agesq row

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

$$H_o : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

$$T = \frac{b_1}{s(b_1)} = -.207$$

$$H_o : \beta_2 = 0$$

$$H_a : \beta_2 \neq 0$$

$$T = \frac{b_2}{s(b_2)} = 3.174$$

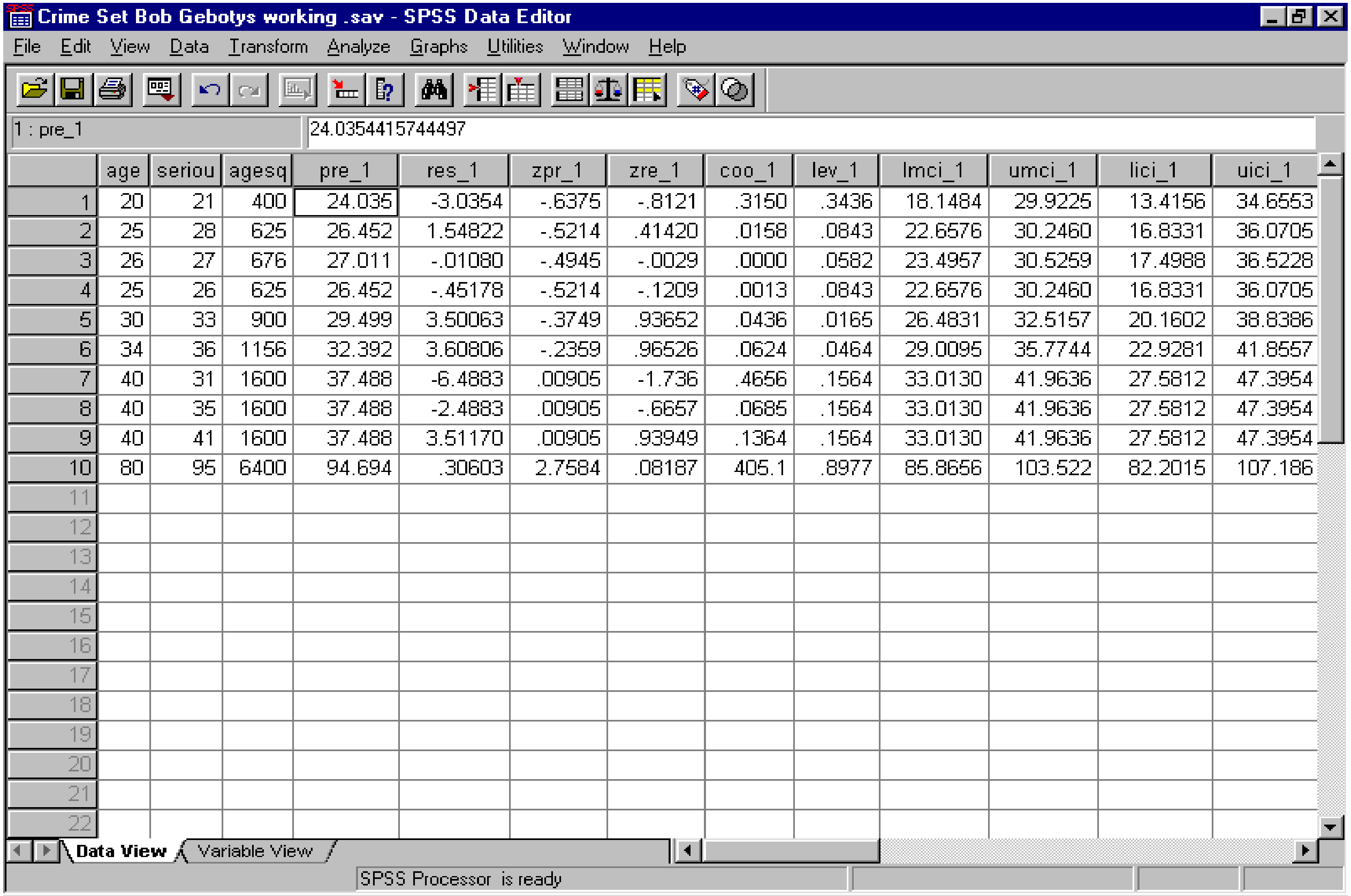
For β_1 the t statistic has the value -.207, and for β_2 the t statistic has the value 3.174. The column **SIG** gives the OLS or p-value for the t test above. In this case we have p = .842 for β_1 (not significant, therefore we cannot reject H_0), and p = .016 for β_2 (significant, therefore we strong evidence against H_0). Both are with 8 degrees of freedom.

Residuals

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	24.04	94.69	37.30	20.81	10
Std. Predicted Value	-.638	2.758	.000	1.000	10
Standard Error of Predicted Value	1.28	3.73	1.93	.72	10
Adjusted Predicted Value	-35.46	39.73	24.58	21.72	10
Residual	-6.49	3.61	-7.11E-16	3.30	10
Std. Residual	-1.736	.965	.000	.882	10
Stud. Residual	-2.013	1.690	.127	1.156	10
Deleted Residual	-8.73	130.46	12.72	41.60	10
Stud. Deleted Residual	-2.872	2.034	.077	1.400	10
Mahal. Distance	.148	8.079	1.800	2.357	10
Cook's Distance	.000	405.070	40.618	128.055	10
Centered Leverage Value	.016	.898	.200	.262	10

a. Dependent Variable: Crime Seriousness

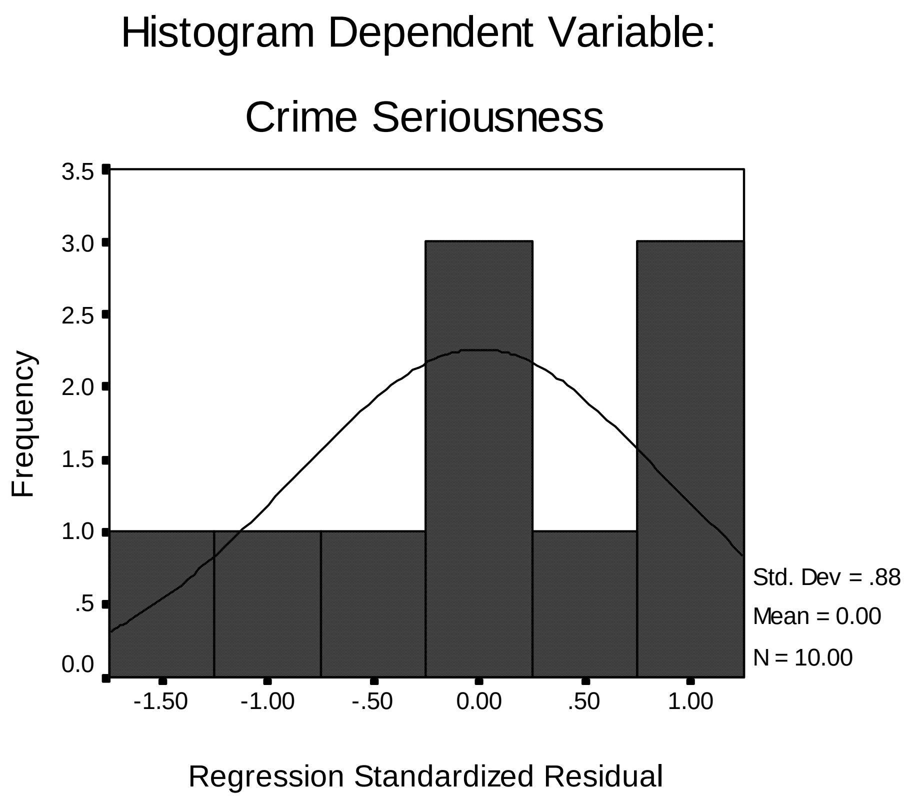


	age	seriou	agesq	pre_1	res_1	zpr_1	zre_1	coo_1	lev_1	lmci_1	umci_1	lici_1	uici_1
1	20	21	400	24.035	-3.0354	-.6375	-.8121	.3150	.3436	18.1484	29.9225	13.4156	34.6553
2	25	28	625	26.452	1.54822	-.5214	.41420	.0158	.0843	22.6576	30.2460	16.8331	36.0705
3	26	27	676	27.011	-.01080	-.4945	-.0029	.0000	.0582	23.4957	30.5259	17.4988	36.5228
4	25	26	625	26.452	-.45178	-.5214	-.1209	.0013	.0843	22.6576	30.2460	16.8331	36.0705
5	30	33	900	29.499	3.50063	-.3749	.93652	.0436	.0165	26.4831	32.5157	20.1602	38.8386
6	34	36	1156	32.392	3.60806	-.2359	.96526	.0624	.0464	29.0095	35.7744	22.9281	41.8557
7	40	31	1600	37.488	-6.4883	.00905	-1.736	.4656	.1564	33.0130	41.9636	27.5812	47.3954
8	40	35	1600	37.488	-2.4883	.00905	-.6657	.0685	.1564	33.0130	41.9636	27.5812	47.3954
9	40	41	1600	37.488	3.51170	.00905	.93949	.1364	.1564	33.0130	41.9636	27.5812	47.3954
10	80	95	6400	94.694	.30603	2.7584	.08187	405.1	.8977	85.8656	103.522	82.2015	107.186
11													
12													
13													
14													
15													
16													
17													
18													
19													
20													
21													
22													

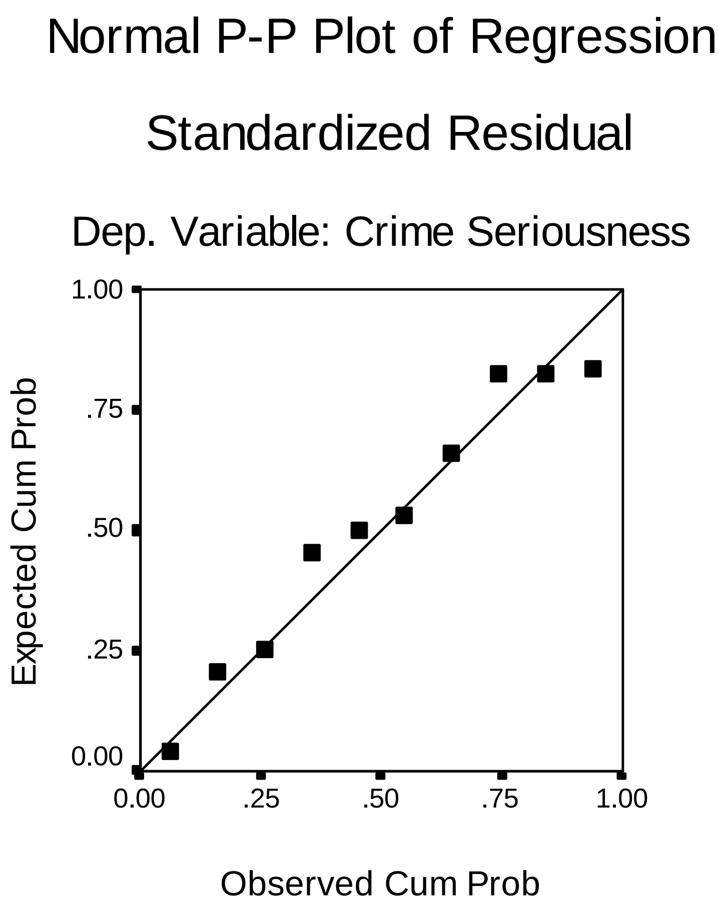
We examined the residuals of the model using a variety of plots and statistics. The Casewise plot gives a band between ±3 standard deviations that is reasonable. The Durbin-Watson Statistic is about 2, which indicates zero correlation. The leverage (LEVER) and Cook's distance (COOK D) values for the 10th observation are relatively large (h_{10} = .8977, D = 405.069) indicating this is an influencial observation. If we compare h_{10} to $2\bar{h}$ where $\bar{h} = .2$ (found in the summary statistics section or use our formula for leverage, (3-1)/10=.2) our suspicion that the 10th observation (a person 80 year old with a high seriousness rating) is influencial is confirmed.

Notice that the residual for this observation is small and therefore not an outlier. The researchers might elect to collect more data in the 40 through 80 year old range to increase their confidence in the model and dampen the influence of this observation. Another option might be to remove the point from the analysis and treat this as a special case that needs further investigation. There are no set solutions for how to handle influential observations, however, the knowledge that they exist is valuable and a logical approach in dealing with them is highly recommended.

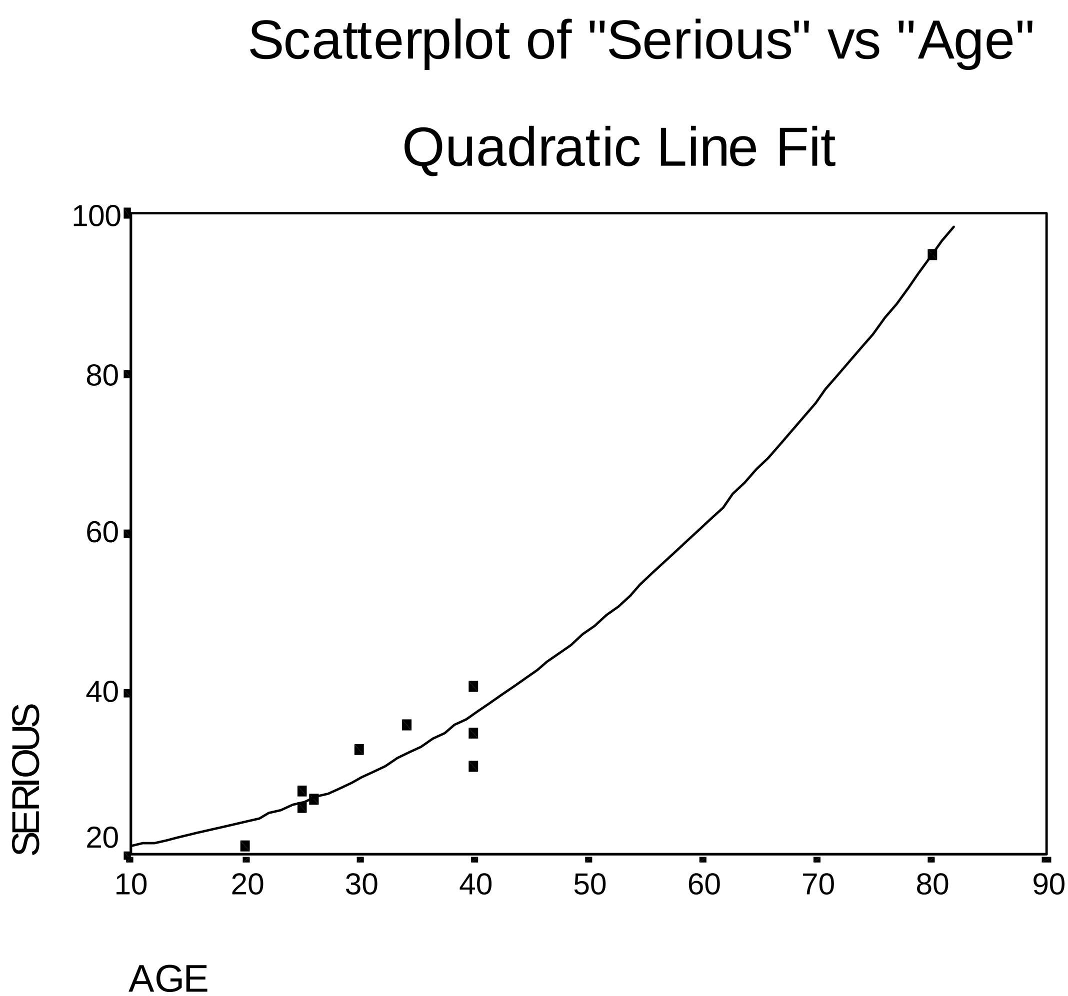
The histogram of residuals looks reasonable, although with 10 observations, this is difficult to judge.



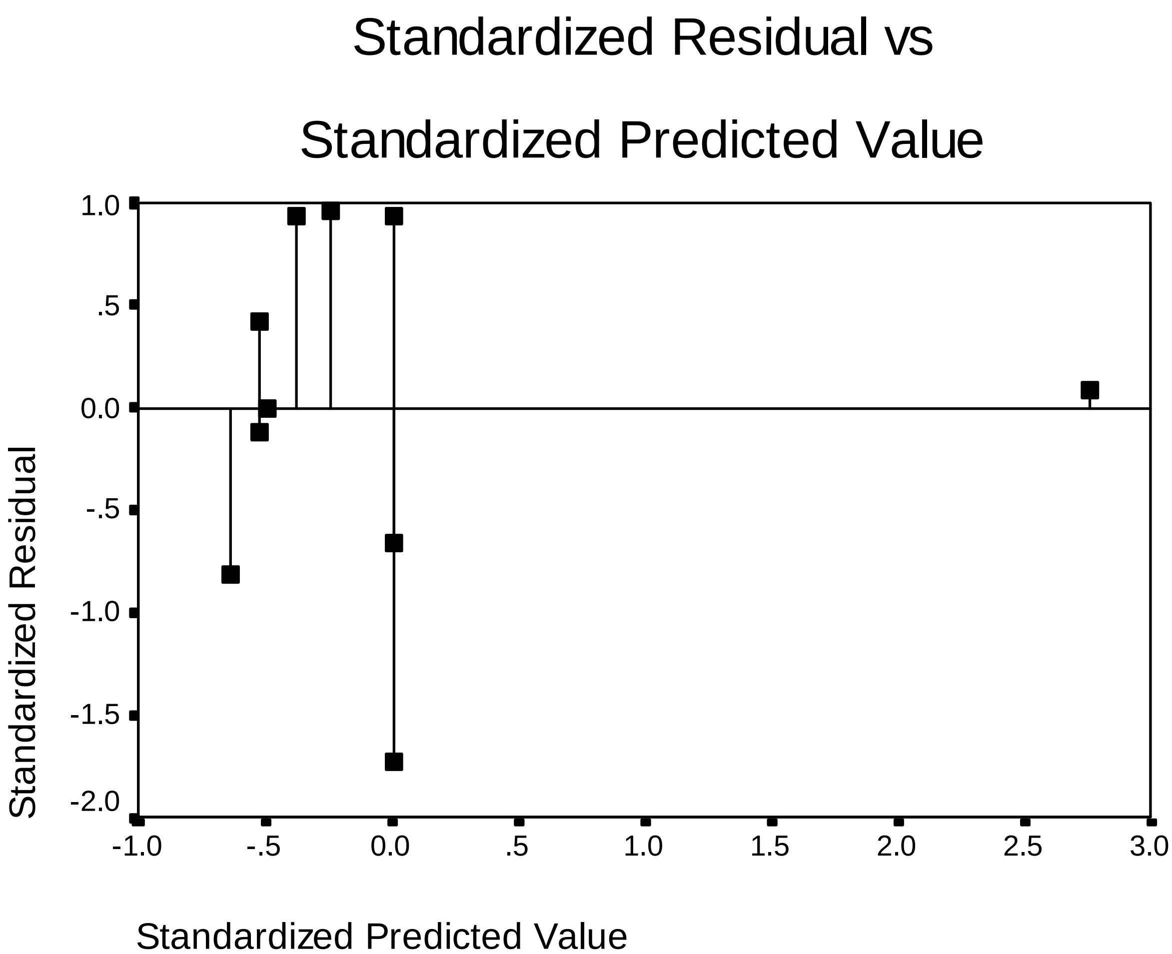
The probability plot (see section 3.7) has improved from the previous problem of fitting a line discussed in Chapter 7 in the sense that the residuals more closely approximate a normal distribution. The large bulge present in the normal probability plot of residuals in Chapter 7 is no longer present in the polynomial of degree 2 model.



The plot of the standardized data indicates a quadratic model would be reasonable. Note that a polynomial model of degree two is being suggested on the basis of one point that is far from the major cluster of points.



The plot of standardized $\hat{y}$ vs. e_i displays a reasonable band shape as well.



In Chapter 11 we will learn how to compare the polynomial model presented here with the line model in Chapter 7 using the ANOVA technique. Although the polynomial model has a higher R^2 than the line we do not know whether this improvement is statistically significant. In Chapter 11 we will learn how to compare these types of nested models.

8.4 COMPUTER IMPLEMENTATION USING SAS

The computer output of SAS is similar to the SPSS package and will not be reproduced here, however, the commands are listed below to fit the model previously discussed.

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

Note that the lines are identical to the program discussed in Chapter 7 however the AGESQ variable has been added in the MODEL statement and at the beginning of the program as a transformation. The INFLUENCE option has also been added to the MODEL statement to give influence statistics.

8.5 Exercises

1. Neifeld, M. and Poffenberger, A., A Mathematical Analysis of Work Curves, J. General Psychol. 1928, 1, 448-458 examined how fatigue was related to work. People were required to lift a 24lb. weight unit exhausted on a number of occasions. The results are given below where x stands for segment of work and y stands for number of arbitrary units of work.

X	Y
1	25.1
2	25.2
3	24.8
4	24.3
5	23.8
6	23.3
7	22.2
8	20.9
9	18.5
10	15.8
11	12.1
12	9.7
13	5.7
14	3.4
15	2.1

Find the best polynomial model for this data. Give reasons for your choice. Consider models up to degree 3.

2. Ioteyko, J., La Fatigue, Paris: Flammarion, 1920 examined fatigue in basically the same manner as above. The data have been converted to 15 work segments (x) so that a comparison can be made with Neifeld and Poffenberger (1928), Problem 1. Y again denotes units of work. The data are listed on the next page.

X	Y
1	32.6
2	31.4
3	30.2
4	29.2
5	27.6
6	26.5
7	23.8
8	24.8
9	23.5
10	21.4
11	20.2
12	17.2
13	14.8
14	9.7
15	2.0

Find the best polynomial model for this data. State your reasons clearly. Consider models up to degree 3.

a. Compare and contrast your results to problems 1 and 2.

b. $Y = \beta_0 + \beta_1 X + \beta_2 X^2$

3. Chapanis (Psychometrika, 1953, 18, 327-336) utilized the following set of (X,Y) values to illustrate the shortcomings of the reduction process for fitting a parabolic arc to empirical data:

X	Y
.1	25
.5	39
1.5	42
2.0	43
3.0	54
3.8	53
4.6	58
4.9	51
6.7	51
7.8	50
8.2	40
9.0	34
10.0	33
11.0	16
11.2	24

8.6 MATRIX NOTATION FOR THE LINEAR MODEL

We briefly introduce matrix notation here to assist us in writing the linear model. The use of matrix notation in the text is strictly to assist the student in writing linear models and conceptualizing the fitting and evaluation of models. The computer (SPSS, SAS) will perform all calculations. No matrix algebra will be performed by hand. Matrix notation however is an invaluable tool in writing and conceptualizing model and aids significantly in understanding the computer output. We will use it as a convenient shorthand notation and let the computer perform the calculations. Matrix algebra is not necessary for an understanding of the material in this text.

As we discussed in the previous section we assume hereafter that the relationship between the dependent variable y and the independent variables $x_1,\ldots,x_k$ (perhaps re-expressions of the original independent variables) is of the form, ignoring for the moment the possibility of variation.

$$\begin{aligned} y &= f(x_1,\ldots,x_k) \\ &= \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \\ &= X'_{1xk}\, \beta_{kx1} \\ &= X'\beta \end{aligned}$$

In the statistical context i.e. when a particular value for $(x_1,\ldots,x_k)'$ specifies a frequency distribution for y we ***assume*** that

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

and that changes in $(x_1,\ldots,x_k)$ affect at most the ***means*** of the frequency distributions. Read $E[y|x_1,\ldots,x_k]$ as the average value of y given $x_1,\ldots,x_k$. If we put $e = y - \beta_1 x_1 - \ldots - \beta_k x_k$ then the frequency distribution of e is constant as $(x_1,\ldots,x_k)$ changes.

Thus we can write our model as

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

and e is referred to as the error term.

If f is the frequency of e, then for a particular value of $(x_1,\ldots,x_k)$ the frequency function of y is given by $f(e - \beta_1 x_1 - \ldots - \beta_k x_k)$. We will ***assume*** hereafter that f can be taken to be a density function and that the variance of the frequency distribution for e exists and is equal to σ^2.

In a psychological investigation our primary purpose will be to make inferences about the true value of the coefficients $\beta_1, \beta_2, \ldots, \beta_k$.

To do this we will be required to make a number of observations at different values of $(x_1,\ldots,x_k)'$.

Let y_i denote the observation taken at

$$X'_{(i)} = (x_{i1},\ldots,x_{ik})$$

and let e_i denote the error. Then for n observations we have

$$y_{nx1} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_1 x_{11} + & \ldots & + \beta_k x_{1k} + & e_1 \\ \beta_1 x_{21} + & \ldots & + \beta_k x_{2k} + & e_2 \\ . & & . & . \\ . & & . & . \\ . & & . & . \\ \beta_1 x_{n1} + & \ldots & + \beta_k x_{nk} + & e_n \end{bmatrix} = \begin{bmatrix} x_{11} & \ldots & x_{1k} \\ x_{21} & \ldots & x_{2k} \\ . & & . \\ . & & . \\ . & & . \\ x_{n1} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

$$= X_{nxk}\,\beta_{kx1} + e_{nx1}$$

$$\mathbf{= X\beta + e}$$

where **X** is called the ***design matrix***.

We assume that the form of the frequency distribution for e is normal: i.e.

$$f(e) = (s\pi\sigma^2)^{-\frac{1}{2}} e^{\left(-\frac{1}{2\sigma^2}e^2\right)}$$

or e ~ $N(0, \sigma^2)$

The statistical model (See Section 2.4 for information on statistical models) we have constructed here is

$$(R, (2\pi\sigma^2)^{-\frac{1}{2}} e^{\left(-\frac{1}{2\sigma^2}(y_i - \beta_1 x_1 - ... - \beta_k x_,)^2\right)} \beta_k \in R, \sigma \in R^+)$$

called the ***linear model with normal error***.

For the normal linear model the least squares estimator of β is given by

$$\mathbf{b = (X'X)^{-1}X'y}$$

The vector

$$\mathbf{Y - Xb} = \boldsymbol{e}$$

is called the ***residual vector***. We can write

$$||y||^2 = ||Xb||^2 + ||y - Xb||^2$$

where

$$||Xb||^2 = y'Xb$$

$$||y - Xb||^2 = y'y - y'Xb$$

This can be displayed in an ANOVA table

Source	Degrees of freedom	Sum of squares
Model	k	y'Xb
Residual	n-k	y'y - y'Xb
Total	n	y'y

Computer packages usually remove the constant (β_1) automatically giving a corrected total i.e. (n-1) and Model (k-1) degrees of freedom, sums of squares etc.

The ANOVA table removing the constant (β_1) looks like the following

Source	Degrees of freedom	Sum of squares
Model	k-1	$y'Xb - \frac{\left(\sum y_i\right)^2}{n}$
Residual	n-k	y'y - y'Xb
Total	n-1	$y'y - \frac{\left(\sum y_i\right)^2}{n}$

More information on the ANOVA procedure and computer implementation will be given in Chapter 11. SPSS will perform all matrix calculations however matrix notation will be very useful in writing and conceptualizing the linear model in succeeding chapters.