Chapter 3: Regression Analysis and Forecasting

Rong Fan

South Dakota State University

rong.fan@sdstate.edu

3.1 Intro I

Some stuff in this Chapter should look familiar.

A simple linear regression model involves a single predictor variable and is written as

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{1}$$

where y is the response, x is the predictor, β_0 and β_1 are parameters which are called intercept and slope respectively, and ϵ is an error term. We usually assume that the error term ϵ is normally distributed with mean 0 and variance σ^2, $N(0, \sigma^2)$.

A multiple linear regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon. \tag{2}$$

3.1 Intro II

We will talk about the multiple linear model for two situations.

1. Data are collected in a single time period. This type is called cross-section data. The model can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i, \qquad i = 1, 2, \cdots, n \quad (3)$$

   where i denotes individual observation and n is the total number of observations.

2. Time series data. The regressors and the response are time series. The model can be written as

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_k x_{tk} + \epsilon_t, \qquad t = 1, 2, \cdots, T \quad (4)$$

Notice that the both models are estimated by **Least Square Estimation**. Also notice the total number of regressors is k.

3.2 Least Squares Estimation in Linear Regression Models I

For cross section data, denote the ith observation of variable x_j by x_{ij}, so $1 \le j \le k$.

It would be much easier to present the results if matrix notations are used. Let $\underset{\sim}{y} = [y_1, y_2, ..., y_n]'$, $\underset{\sim}{\beta} = [\beta_0, \beta_1, ..., \beta_k]'$, $\underset{\sim}{\epsilon} = [\epsilon_1, \epsilon_2, ..., \epsilon_n]'$ and

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots & \\ 1 & x_{n1} & x_{n2} & x_{n3} & \dots & x_{nk} \end{bmatrix} \tag{5}$$

Then the matrix form model is

$$\underset{\sim}{y} = \mathbf{X}\underset{\sim}{\beta} + \underset{\sim}{\epsilon}. \tag{6}$$

3.2 Least Squares Estimation in Linear Regression Models II

Then the least square estimate of β which is denoted by $\widehat{\beta}$ is

$$\underset{\sim}{\widehat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underset{\sim}{y}. \tag{7}$$

A special case is that there is only one regressor, or there are two parameters.

$$y = \beta_0 + \beta_1 x \tag{8}$$

Here there is a short cut to find the least square estimates, that is

$$\hat{\beta}_1 = r\frac{s_y}{s_x} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{9}$$

where r is the correlation, s_x, s_y are the standard deviations of the two variables, and $\bar{x}$, $\bar{y}$ are the mean of the two variables. From this one can see that the fitted line has slope=0 if and only if the two variables are uncorrelated ($r = 0$).

3.2 Least Squares Estimation in Linear Regression Models III

Then the fitted values for the response is

$$\hat{\underset{\sim}{y}} = \mathbf{X}\underset{\sim}{\widehat{\beta}}, \tag{10}$$

or

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_k x_{ik}, \qquad i = 1, 2, \cdots, n \tag{11}$$

Residual vector is

$$\underset{\sim}{e} = \underset{\sim}{y} - \hat{\underset{\sim}{y}} = \underset{\sim}{y} - \mathbf{X}\underset{\sim}{\hat{\beta}}. \tag{12}$$

The sum of squares of the residuals is

$$(\underset{\sim}{y} - \mathbf{X}\underset{\sim}{\hat{\beta}})'(\underset{\sim}{y} - \mathbf{X}\underset{\sim}{\hat{\beta}}). \tag{13}$$

3.2 Least Squares Estimation in Linear Regression Models IV

After dividing the degree of freedom $n-p$, where $p=k+1$, we get an estimator of the error variance σ^2:

$$\hat{\sigma}^2 = \text{Mean Square Error} = \frac{\text{SSE}}{n-p}. \tag{14}$$

$\underset{\sim}{\hat{\beta}}$ is in fact an unbiased estimator of $\underset{\sim}{\beta}$ meaning $E(\underset{\sim}{\hat{\beta}}) = \underset{\sim}{\beta}$.

3.2 Least Squares Estimation in Linear Regression Models V

Since $Var(\mathbf{A}\underset{\sim}{y}) = \mathbf{A}Var(\underset{\sim}{y})\mathbf{A}'$. Let $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, therefore $\hat{\underset{\sim}{\beta}} = \mathbf{A}\underset{\sim}{y}$ and the covariance matrix of $\hat{\underset{\sim}{\beta}}$ is

$$
\begin{aligned}
Var(\hat{\underset{\sim}{\beta}}) &= Var(\mathbf{A}\underset{\sim}{y}) && (15)\\
&= \mathbf{A}Var(\underset{\sim}{y})\mathbf{A}' && (16)\\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2 I_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} && (17)\\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} && (18)
\end{aligned}
$$

The main diagonal of the matrix consists of the variance for each $\underset{\sim}{\beta}_j$.

A hospital is attempting to evaluate patient satisfaction.

TABLE 3.2 Patient Satisfaction Survey Data

Observation	Age (x_1)	Severity (x_2)	Satisfaction (y)
1	55	50	68
2	46	24	77
3	30	46	96
4	35	48	80
5	59	58	43
6	61	60	44
7	74	65	26
8	38	42	88
9	27	42	75
10	51	50	57
11	53	38	56
12	41	30	88
13	37	31	88
14	24	34	102
15	42	30	88
16	50	48	70
17	58	61	52
18	60	71	43
19	62	62	46
20	68	38	56
21	70	41	59
22	79	66	26
23	63	31	52
24	39	42	83
25	49	40	75

3.2 An example II

The model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

or in matrix form $\underset{\sim}{y} = \mathbf{X}\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$ where

$$\mathbf{X} = \begin{bmatrix} 1 & 55 & 50 \\ 1 & 46 & 24 \\ 1 & 30 & 46 \\ 1 & 35 & 48 \\ 1 & 59 & 58 \\ 1 & 61 & 60 \\ 1 & 74 & 65 \\ 1 & 38 & 42 \\ 1 & 27 & 42 \\ 1 & 51 & 50 \\ 1 & 53 & 38 \\ 1 & 41 & 30 \\ 1 & 37 & 31 \\ 1 & 24 & 34 \\ 1 & 42 & 30 \\ 1 & 50 & 48 \\ 1 & 58 & 61 \\ 1 & 60 & 71 \\ 1 & 62 & 62 \\ 1 & 68 & 38 \\ 1 & 70 & 41 \\ 1 & 79 & 66 \\ 1 & 63 & 31 \\ 1 & 39 & 42 \\ 1 & 49 & 40 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 68 \\ 77 \\ 96 \\ 80 \\ 43 \\ 44 \\ 26 \\ 88 \\ 75 \\ 57 \\ 56 \\ 88 \\ 88 \\ 102 \\ 88 \\ 70 \\ 52 \\ 43 \\ 46 \\ 56 \\ 59 \\ 26 \\ 52 \\ 83 \\ 75 \end{bmatrix}$$

3.2 An example III

Then we can find the least squares estimates as

$$\underset{\sim}{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underset{\sim}{y} = [143.4720118, -1.031053414, -0.55603781]' \quad (19)$$

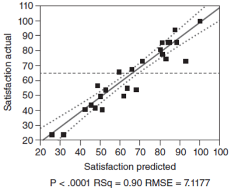
Therefore the regression model is

$$\hat{y} = 143.472 - 1.031x_1 - 0.556x_2$$

3.2 An example IV

TABLE 3.3 JMP Output for the Patient Satisfaction Data in Table 3.2

Actual by Predicted Plot



Summary of Fit

RSquare	0.896593
RSquare Adj	0.887192
Root mean square error	7.117667
Mean of response	65.52
Observations (or Sum Wgts)	25

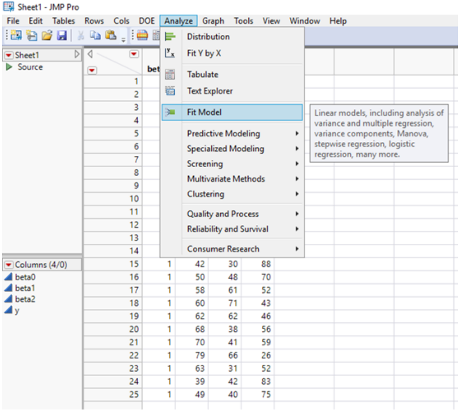
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	9663.694	4831.85	95.3757
Error	22	1114.546	50.66	Prob > F
C. Total	24	10778.240		<.0001*

Parameter Estimates

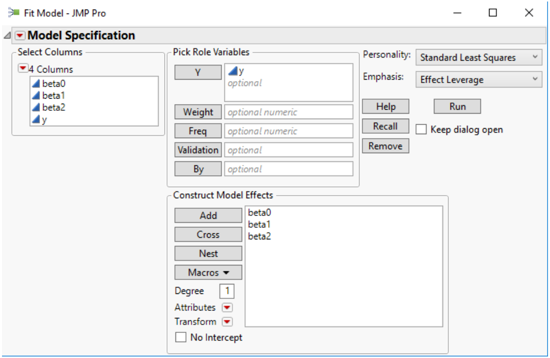
Term	Estimate	Std Error	t Ratio	Prob> $\|t\|$
Intercept	143.47201	5.954838	24.09	<.0001*
Age	−1.031053	0.115611	−8.92	<.0001*
Severity	−0.556038	0.13141	−4.23	0.0003*

1. Importing the data into JMP.
2. **Analyze → Fit Model**



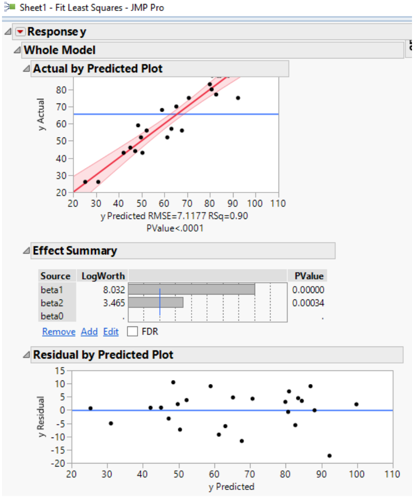
3.2 Supplement: Regression Model in JMP II

3. Select y (dependent variable) to **Y** under **Pick Role Variables**.
4. Add beta0, beta1, and beta2 (regressors or independent variables) into **Construct Model Effects**.
5. Click **Run**.



3.2 Supplement: Regression Model in JMP III

Output of Regression Model I



Output of Regression Model II.

Summary of Fit

RSquare	0.896593
RSquare Adj	0.887192
Root Mean Square Error	7.117667
Mean of Response	65.52
Observations (or Sum Wgts)	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	9663.694	4831.85	95.3757
Error	22	1114.546	50.66	Prob > F
C. Total	24	10778.240		<.0001*

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept	Biased	143.47201	5.954838	24.09	<.0001*
beta0	Zeroed	0	0	.	.
beta1		-1.031053	0.115611	-8.92	<.0001*
beta2		-0.556038	0.13141	-4.23	0.0003*

For time series data, we often fit a regression line if there is a linear trend. Recall the only one regressor is time t, i.e.

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & T \end{bmatrix} \tag{20}$$

So the least squares estimates are

$$\hat{\beta}_0 = \frac{2(2T+1)}{T(T-1)} \sum_{t=1}^{T} y_t - \frac{6}{T(T-1)} \sum_{t=1}^{T} t y_t$$

$$\hat{\beta}_1 = \frac{12}{T(T^2-1)} \sum_{t=1}^{T} t y_t - \frac{6}{T(T-1)} \sum_{t=1}^{T} y_t$$

3.2 Least Squares Estimation in Linear Regression Models II

Notice that these estimates depend on T, the time at which they were computed. So it is convenient to denote them by $\underset{\sim}{\hat{\beta}}_0(T)$ and $\underset{\sim}{\hat{\beta}}_1(T)$.

Then the forecast of the next observation is

$$\underset{\sim}{\hat{y}}_{T+1}(T) = \underset{\sim}{\hat{\beta}}_0(T) + \underset{\sim}{\hat{\beta}}_1(T)(T+1). \tag{21}$$

When a new observation becomes available, the parameter estimates would be updated to $\underset{\sim}{\hat{\beta}}_0(T+1)$ and $\underset{\sim}{\hat{\beta}}_1(T+1)$ to reflect the new information.

3.3.1 Test for significance of linear regression model I

In this section, we assume the errors ϵ_i are normally and independently distributed with mean zero and variance σ^2, i.e. $\text{NID}(0, \sigma^2)$. So the y_is are normally and independently distributed with variance σ^2 as well.

Test for significance of the linear regression model is to determine whether there is a linear relationship between $\underset{\sim}{y}$s and $\mathbf{X}$s.
The hypotheses are

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \quad VS \quad H_1 : \text{at least one } \beta_j \neq 0 \tag{22}$$

Rejection of H_0 implies that at least one predictor contributes to the model.

3.3.1 Test for significance of linear regression model II

The test statistic F_0 is based on the total sum of squares

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2, \tag{23}$$

and the sum of squares of the error

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{24}$$

Then

$$F_0 = \frac{SS_R/k}{SS_E/(n-p)} \tag{25}$$

where $SS_R = SS_T - SS_E$ is called the sum of squares of the regression.

We reject H_0 at significance level α, if F_0 exceeds a critical value $F_{\alpha,k,n-p}$. See Table A.4.

3.3.1 Test for significance of linear regression model III

Equivalently, we could use the p-value. We reject H_0, if the p-value is less than α.

The test is usually summarized in an analysis of variance table (ANOVA).

TABLE 3.4 Analysis of Variance for Testing Significance of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Test Statistic, F_0
Regression	SS_R	k	$\frac{SS_R}{k}$	$F_0 = \frac{SS_R/k}{SS_E/(n-p)}$
Residual (error)	SS_E	$n-p$	$\frac{SS_E}{n-p}$	
Total	SS_T	$n-1$		

3.3.1 Test for significance of linear regression model IV

Recall the output of regression model for Patient Satisfaction:

Summary of Fit

RSquare	0.896593
RSquare Adj	0.887192
Root Mean Square Error	7.117667
Mean of Response	65.52
Observations (or Sum Wgts)	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	9663.694	4831.85	95.3757
Error	22	1114.546	50.66	**Prob > F**
C. Total	24	10778.240		<.0001*

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept	Biased	143.47201	5.954838	24.09	<.0001*
beta0	Zeroed	0	0	.	.
beta1		-1.031053	0.115611	-8.92	<.0001*
beta2		-0.556038	0.13141	-4.23	0.0003*

$$F_0 = \frac{9663.694/2}{1114.546/22} = \frac{4831.85}{50.66} = 95.38 > 3.44 = F_{0.05,2,22} \qquad (26)$$

and the p-value is < 0.0001. So there is evidence that we reject the null hypotheses and conclude that at least one from age and severity contributes significantly to patient satisfaction.

3.3.1 Test for significance of linear regression model V

The output also shows R-square $R^2 = 0.896593$ and adjusted R-square $R^2_{adj} = 0.887192$. Notice that they are computed from sum of squares. The inference of R-square is that about 89.7% of the variability in the data is explained by the model.

Both R-square and adjusted R-square are very closed, usually a good sign that the regression model does not contain unnecessary predictors.

3.3.2 Tests on Individual Regression Coefficients I

Usually we are interested in testing the individual regression coefficients. For example, the model might be more effective with the deletion of one variable that is not significant.

The hypotheses for **testing any individual regression coefficient**, say β_j, are

$$H_0 : \beta_j = 0 \quad VS \quad H_1 : \beta_j \neq 0 \tag{27}$$

If null hypothesis in not rejected, then this indicates that the predictor x_j can be deleted from the model.

3.3.2 Tests on Individual Regression Coefficients II

The test statistic is

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \tag{28}$$

where C_{jj} is the diagonal element of the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to $\hat{\beta}_j$
H_0 is rejected if the absolute value of the test statistics is greater than $t_{\alpha/2,n-p}$ as in Table A.3. Similarly, there is the p-value approach.

The term $\sqrt{\hat{\sigma}^2 C_{jj}}$ is called the standard error of $\hat{\beta}_j$, that is

$$se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}} \tag{29}$$

which is usually provided by softwares.

3.3.2 Tests on Individual Regression Coefficients III

Summary of Fit

RSquare	0.896593
RSquare Adj	0.887192
Root Mean Square Error	7.117667
Mean of Response	65.52
Observations (or Sum Wgts)	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	9663.694	4831.85	95.3757
Error	22	1114.546	50.66	Prob > F
C. Total	24	10778.240		<.0001*

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept	Biased	143.47201	5.954838	24.09	<.0001*
beta0	Zeroed	0	0	.	.
beta1		-1.031053	0.115611	-8.92	<.0001*
beta2		-0.556038	0.13141	-4.23	0.0003*

Suppose we want to test whether the predictor patient age is significant to the model.

$$t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{-1.031053}{0.115611} = -8.92 < t_{0.025,22} = -2.0739 \tag{30}$$

So we reject $H_0 : \beta_1 = 0$ and conclude that patient age is statistically significant to the model.

3.3.2 Tests Groups of Coefficients I

We can also **test a group of predictors**, given that other predictors are included in the model by using extra sum of squares.

The "big idea" is to consider:

1. The model with all predictors(the full model),
2. The model with the other predictors only(the reduced model),
3. Then compare the sum of squares.

3.3.2 Tests Groups of Coefficients II

Consider the model with k predictors

$$\underset{\sim}{y} = \mathbf{X}\underset{\sim}{\beta} + \underset{\sim}{\epsilon} \tag{31}$$

where $\underset{\sim}{y}$ and $\underset{\sim}{\epsilon}$ are $(n \times 1)$, $\mathbf{X}$ is $(n \times p)$, $\underset{\sim}{\beta}$ is $(p \times 1)$ and $p = k + 1$. We want to determine if a group of predictors, say x_1, $x_2 ... x_r$ are significant to the model. Let the vector $\underset{\sim}{\beta}$ be partitioned as follows:

$$\underset{\sim}{\beta} = \begin{bmatrix} \underset{\sim}{\beta}_1 \\ \underset{\sim}{\beta}_2 \end{bmatrix} \tag{32}$$

where $\underset{\sim}{\beta}_1$ is $(r \times 1)$ and is formed by the group of predictors, and $\underset{\sim}{\beta}_2$ is $(p - r) \times 1$ and takes care of the leftover.

3.3.2 Tests Groups of Coefficients III

Rewrite $\underset{\sim}{y}$ as:

$$\underset{\sim}{y} = \begin{bmatrix} \mathbf{X_1} & \mathbf{X_2} \end{bmatrix} \begin{bmatrix} \underset{\sim}{\beta}_1 \\ \underset{\sim}{\beta}_2 \end{bmatrix} + \underset{\sim}{\epsilon} = \mathbf{X_1}\underset{\sim}{\beta}_1 + \mathbf{X_2}\underset{\sim}{\beta}_2 + \underset{\sim}{\epsilon} \tag{33}$$

where $\mathbf{X_1}$ and $\mathbf{X_2}$ represent the columns of $\mathbf{X}$ associated with $\underset{\sim}{\beta}_1$ and $\underset{\sim}{\beta}_2$.

3.3.2 Tests Groups of Coefficients IV

The hypotheses are

$$H_0 : \underset{\sim}{\beta}_1 = 0 \quad VS \quad H_1 : \underset{\sim}{\beta}_1 \neq 0 \tag{34}$$

The full model $\underset{\sim}{y} = \mathbf{X}\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$ has regression sum of squares $SS_R(\underset{\sim}{\beta})$.
The reduced model $\underset{\sim}{y} = \mathbf{X_2}\underset{\sim}{\beta}_2 + \underset{\sim}{\epsilon}$ has regression sum of squares $SS_R(\underset{\sim}{\beta}_2)$.

Define the extra sum of squares due to β_1

$$SS_R(\underset{\sim}{\beta}_1|\underset{\sim}{\beta}_2) = SS_R(\underset{\sim}{\beta}) - SS_R(\underset{\sim}{\beta}_2) = \hat{\underset{\sim}{\beta}}'\mathbf{X}'\underset{\sim}{y} - \hat{\underset{\sim}{\beta}}'_2\mathbf{X}'_2 y \tag{35}$$

which describes the regression sum of squares due to $\underset{\sim}{\beta}_1$, given that $\underset{\sim}{\beta}_2$ is already in the model.

The test statistic

$$F_0 = \frac{SS_R(\underset{\sim}{\beta}_1|\underset{\sim}{\beta}_2)/r}{\text{Full Model } SS_E/(n-p)} \tag{36}$$

$$= \frac{SS_R(\underset{\sim}{\beta}_1|\underset{\sim}{\beta}_2)/r}{(\underset{\sim}{y}'\underset{\sim}{y} - \hat{\underset{\sim}{\beta}}'\mathbf{X}'\underset{\sim}{y})/(n-p)} \tag{37}$$

We reject H_0 if $F_0 > F_{\alpha,r,n-p}$. So there is at least one predictor in the group $\beta_1, \cdots, \beta_r$ that is significant.

Notice that when $r = 1$, this is equivalent to the t-test for a single predictor.

3.3 An Example I

Consider the patient example again. Suppose we fit a more elaborate model by incorporating the second degree terms:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon \tag{38}$$

This model is fitted by letting the **X** matrix be

$$\begin{bmatrix} 1 & x_{11} & x_{12} & x_{11}x_{12} & x_{11}^2 & x_{12}^2 \\ 1 & x_{21} & x_{22} & x_{21}x_{22} & x_{21}^2 & x_{22}^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ 1 & x_{n1} & x_{n2} & x_{n1}x_{n2} & x_{n1}^2 & x_{n2}^2 \end{bmatrix}$$

3.3 An Example II

Suppose we want to test if the second degree terms are significant:

$$H_0 : \beta_{12} = \beta_{11} = \beta_{22} = 0 \quad VS \quad H_1 : \text{at least one} \neq 0 \tag{39}$$

TABLE 3.5 JMP Output for the Second-Order Model for the Patient Satisfaction Data

Summary of Fit

RSquare	0.900772
RSquare Adj	0.874659
Root mean square error	7.502639
Mean of response	65.52
Observations (or Sum Wgts)	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	9708.738	1941.75	34.4957
Error	19	1069.502	56.29	Prob > F
C. Total	24	10,778.240		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> |t|
Intercept	143.74009	6.774622	21.22	<0.0001*
Age	−0.986524	0.135366	−7.29	<0.0001*
Severity	−0.571637	0.158928	−3.60	0.0019*
(Severity-45.92)*(Age-50.84)	0.0064566	0.016546	0.39	0.7007
(Age-50.84)*(Age-50.84)	−0.00283	0.008588	−0.33	0.7453
(Severity-45.92)*(Severity-45.92)	−0.011368	0.013533	−0.84	0.4113

Summary of Fit

RSquare	0.896593
RSquare Adj	0.887192
Root Mean Square Error	7.117667
Mean of Response	65.52
Observations (or Sum Wgts)	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	9663.694	4831.85	95.3757
Error	22	1114.546	50.66	Prob > F
C. Total	24	10778.240		<.0001*

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept	Biased	143.47201	5.954838	24.09	<.0001*
beta0	Zeroed	0	0	.	.
beta1		-1.031053	0.115611	-8.92	<.0001*
beta2		-0.556038	0.13141	-4.23	0.0003*

3.3 An Example III

1. The full model has $SS_R(\underset{\sim}{\beta}) = 9708.738$.
2. The reduced model, $SS_R(\underset{\sim}{\beta}_2) = 9663.694$.
3. The extra sum of squares is $9708.738 - 9663.694 = 45.044$.

Then the test statistic

$$F_0 = \frac{45.044/3}{56.29} = 0.267 < F_{0.05,3,19} = 3.1274 \tag{40}$$

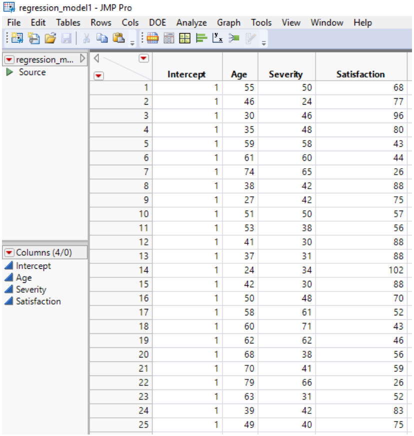
So there is no evidence against the null hypothesis. There is no reason to believe that the model would be improved by adding the second degree terms.

3.3 An Example IV

Notice also the the adjusted R-square for the full model is less than the reduced model. This is also an indication that the second degree terms should not be added.

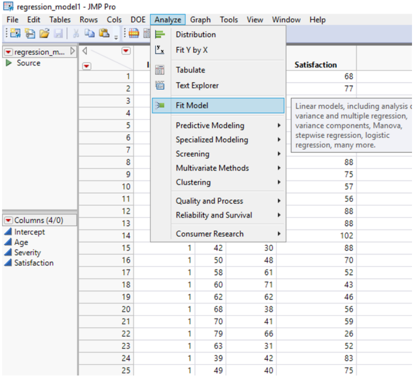
Second Order of Regression Model in JMP I

Import data:



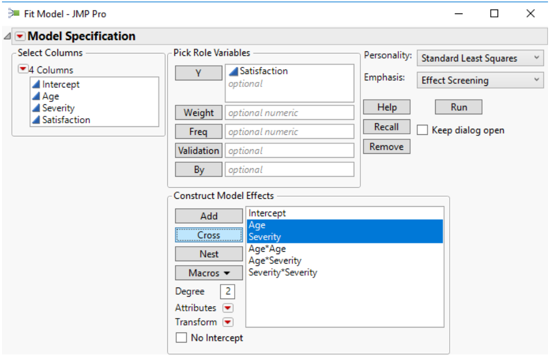
	Intercept	Age	Severity	Satisfaction
1	1	55	50	68
2	1	46	24	77
3	1	30	46	96
4	1	35	48	80
5	1	59	58	43
6	1	61	60	44
7	1	74	65	26
8	1	38	42	88
9	1	27	42	75
10	1	51	50	57
11	1	53	38	56
12	1	41	30	88
13	1	37	31	88
14	1	24	34	102
15	1	42	30	88
16	1	50	48	70
17	1	58	61	52
18	1	60	71	43
19	1	62	62	46
20	1	68	38	56
21	1	70	41	59
22	1	79	66	26
23	1	63	31	52
24	1	39	42	83
25	1	49	40	75

Select Analyze → Fit Model:



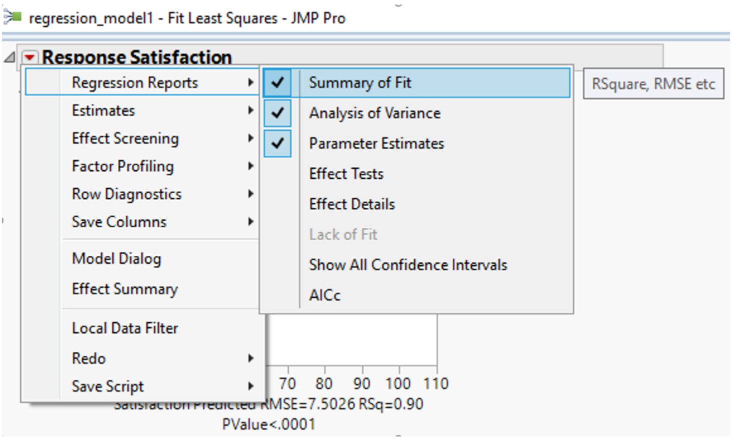
Second Order of Regression Model in JMP III

1. Add Age and Severity to Conduct Model Effects.
2. Select both Age and Severity, then click on "Cross" to add the the mutual effects of the two variables to the model.



Second Order of Regression Model in JMP IV

3. Response Satisfaction (Red Triangle)
4. Regression Reports (Right Triangle).
5. Check Summary of Fit, Analysis of Variance and Parameter Estimates.



Second Order of Regression Model in JMP V

Output of the second order of regression model (Table 3.5).

Summary of Fit

RSquare	0.900772
RSquare Adj	0.874659
Root Mean Square Error	7.502639
Mean of Response	65.52
Observations (or Sum Wgts)	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	9708.738	1941.75	34.4957
Error	19	1069.502	56.29	Prob > F
C. Total	24	10778.240		<.0001*

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept	Biased	143.74009	6.774622	21.22	<.0001*
Intercept	Zeroed	0	0	.	.
Age		-0.986524	0.135366	-7.29	<.0001*
Severity		-0.571637	0.158928	-3.60	0.0019*
(Age-50.84)*(Age-50.84)		-0.00283	0.008588	-0.33	0.7453
(Age-50.84)*(Severity-45.92)		0.0064566	0.016546	0.39	0.7007
(Severity-45.92)*(Severity-45.92)		-0.011368	0.013533	-0.84	0.4113

A $100(1-\alpha)$ confidence interval (CI) for the parameter, say β_j is

$$(\hat{\beta}_j - t_{\alpha/2,n-p} se(\hat{\beta}_j), \quad \hat{\beta}_j + t_{\alpha/2,n-p} se(\hat{\beta}_j)) \tag{41}$$

Summary of Fit

RSquare	0.896593
RSquare Adj	0.887192
Root Mean Square Error	7.117667
Mean of Response	65.52
Observations (or Sum Wgts)	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	9663.694	4831.85	95.3757
Error	22	1114.546	50.66	Prob > F
C. Total	24	10778.240		<.0001*

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept	Biased	143.47201	5.954838	24.09	<.0001*
beta0	Zeroed	0	0	.	.
beta1		-1.031053	0.115611	-8.92	<.0001*
beta2		-0.556038	0.13141	-4.23	0.0003*

According to the summary table above, we have $\hat{\beta}_1 = -1.0311$ and $se(\hat{\beta}_j) = 0.1156$, so a 95% CI for β_1 is

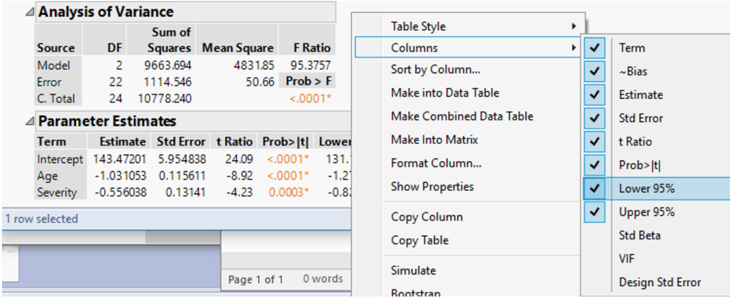
$$
\begin{aligned}
&(-1.0311 - (2.074)(0.1156),\ -1.0311 + (2.074)(0.1156)) && (42)\\
&= (-1.2709,\ -0.7913) && (43)
\end{aligned}
$$

Recall that an $100(1-\alpha)$ CI does not contain 0 if and only if the predictor is significant at α.

3.3 Statistical Inference in Linear Regression III

Steps for obtaining CIs for coefficients $\beta_0, \beta_1, \cdots, \beta_k$:

1. Right click any place within the block of **Parameter Estimates area**.
2. Select **Columns**.
3. Select **Lower 95%** and **Upper 95%** for obtaining 95% confidence intervals for the coefficients $\beta_i, i = 1, 2, \cdots, k$.



4. Output of the CIs. For example, the CIs for coefficient of Age (β_1) is (-1.270816, -0.791291).

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	143.47201	5.954838	24.09	<.0001*	131.12243	155.82159
Age	-1.031053	0.115611	-8.92	<.0001*	-1.270816	-0.791291
Severity	-0.556038	0.13141	-4.23	0.0003*	-0.828566	-0.28351

3.3 Statistical Inference in Linear Regression V

Let us say the model is fitted. Then we want to make inference on the mean response

$$E(y(\underset{\sim}{x}_0)) = \mu_{\underset{\sim}{y}|\underset{\sim}{x}_0} = \underset{\sim}{x}_0' \underset{\sim}{\beta} \tag{44}$$

for any particular values of the predictors, say x_{01}, $x_{02} ... x_{0k}$. Denote the vector $\underset{\sim}{x}_0 = [1, x_{01}, x_{02} ... x_{0k}]'$

A $100(1-\alpha)$ CI on the mean response is

$$\left(\hat{y}(\underset{\sim}{x}_0) - t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 \underset{\sim}{x}_0 (\mathbf{X}'\mathbf{X})^{-1} \underset{\sim}{x}_0},\ \ \hat{y}(\underset{\sim}{x}_0) + t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 \underset{\sim}{x}_0 (\mathbf{X}'\mathbf{X})^{-1} \underset{\sim}{x}_0}, \right. \tag{45}$$

where $\hat{\sigma}^2$ is the estimate of error variance.
Notice that

$$se(\hat{y}(\underset{\sim}{x}_0)) = \sqrt{\hat{\sigma}^2 \underset{\sim}{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \underset{\sim}{x}_0} \tag{46}$$

Reconsider the patient example. Suppose we want to find the a CI on mean patient satisfaction for the point where age=55 and severity=50. The fitted mean response $\hat{y}(\underset{\sim}{x}_0) = 58.96$, the standard error

$$se(\hat{y}(\underset{\sim}{x}_0)) = \sqrt{\hat{\sigma}^2 \underset{\sim}{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \underset{\sim}{x}_0} = 1.51, \qquad (47)$$

and $t_{0.05/2,22} = 2.074$.

So A 95% CI on the mean patient satisfaction for the case where age=55 and severity=50 is

$$(58.96 - 2.074 \cdot 1.51,\ 58.96 + 2.074 \cdot 1.51) = (55.83,\ 62.09). \qquad (48)$$

In the case where the combination of x_1 and x_2 is not in the observations in the sample, the same formula can be used to find CIs as well. JMP has a convenient way of doing this.

TABLE 3.6 JMP Calculations of the Standard Errors of the Fitted Values and Predicted Responses for the Patient Satisfaction Data

Observation	Age	Severity	Satisfaction	Predicted	Residual	SE (Fit)	SE (Predicted)
1	55	50	68	58.96218	9.037816	1.507444	7.275546
2	46	24	77	82.69865	−5.69865	2.988566	7.719631
3	30	46	96	86.96267	9.03733	2.803259	7.6498
4	35	48	80	80.69533	−0.69533	2.446294	7.526323
5	59	58	43	50.38967	−7.38967	1.962621	7.383296
6	61	60	44	47.21548	−3.21548	2.128407	7.429084
7	74	65	26	31.0316	−5.0316	2.89468	7.683772
8	38	42	88	80.93839	7.061606	1.919979	7.372076
9	27	42	75	92.27998	−17.28	2.895873	7.684221
10	51	50	57	63.0864	−6.0864	1.517813	7.277701
11	53	38	56	67.69674	−11.6967	1.856609	7.355826
12	41	30	88	84.51769	3.482312	2.275784	7.472641
13	37	31	88	88.08586	−0.08586	2.260863	7.468111
14	24	34	102	99.82144	2.178556	2.994326	7.721863
15	42	30	88	83.48663	4.513366	2.277152	7.473058
16	50	48	70	65.22953	4.770474	1.462421	7.266351
17	58	61	52	49.75261	2.247393	2.214287	7.454143
18	60	71	43	42.13012	0.869878	3.21204	7.808866
19	62	62	46	45.07236	0.927644	2.296	7.478823
20	68	38	56	52.23094	3.769057	3.038105	7.738945
21	70	41	59	48.50072	10.49928	2.97766	7.715416
22	79	66	26	25.3203	0.679703	3.24021	7.820495
23	63	31	52	61.27847	−9.27847	3.28074	7.837374
24	39	42	83	79.90734	3.092659	1.849178	7.353954
25	49	40	75	70.70888	4.291118	1.58171	7.291295
—	75	60	—	32.78074	—	2.78991	7.644918
—	60	60	—	48.24654	—	2.120899	7.426937

3.4 Prediction of New Observations I

A regression model can be used to predict future observations on the response y corresponding to a particular set of values of the predictor or regressor variables, say $(x_{01}, x_{02}, \cdots, x_{0k})$. Let $\underset{\sim}{x}_0'$ represent this point, then $\mathbf{x_0'}$ contains the coordinates of the point of interest and unity to account for the intercept term, so $\underset{\sim}{x}_0' = [1, x_{01}, x_{02}, \cdots, x_{0k}]$.
A point estimate of the future observation $y(\underset{\sim}{x}_0)$ at the point $(x_{01}, x_{02}, \cdots, x_{0k})$ is computed from

$$\hat{y}(\underset{\sim}{x}_0) = \underset{\sim}{x}_0' \hat{\beta}. \tag{49}$$

The prediction error in using $\hat{y}(x_0)$ to estimate $y(x_0)$ is $y(x_0) - \hat{y}(x_0)$. Because $y(x_0)$ and $\hat{y}(x_0)$ are independent, the variance of this prediction error is

$$\begin{aligned} Var[y(x_0) - \hat{y}(x_0)] &= Var[y(x_0)] + Var[\hat{y}(x_0)] && (50) \\ &= \sigma^2 + \sigma^2 x_0'(X'X)^{-1}x_0 && (51) \\ &= \sigma^2[1 + x_0'(X'X)^{-1}x_0]. && (52) \end{aligned}$$

If we use $\hat{\sigma}^2$ from equation (49) to estimate the error variance σ^2, then we have

$$\frac{y(x_0) - \hat{y}(x_0)}{\sqrt{\hat{\sigma}^2[1 + x_0'(X'X)^{-1}x_0]}} \sim t_{\alpha/2, n-p} \qquad (53)$$

3.4 Prediction of New Observations III

Therefore, a $100(1-\alpha)$ percent prediction interval for the future observation $y(\mathbf{x_0})$ is

$$\hat{y}(\mathbf{x_0}) \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2[1+\mathbf{x_0'}(\mathbf{X'X})^{-1}\mathbf{x_0}]} \tag{54}$$

Compared with CIs for the mean response, the difference is that PIs have standard errors

$$\sqrt{\hat{\sigma}^2[1+\mathbf{x_0'}(\mathbf{X'X})^{-1}\mathbf{x_0}]} \tag{55}$$

which is larger than the standard errors used in CIs. The effect of this is that PIs will be wider than CIs. This makes sense because PIs are made for a single future observation which should be associated with more variability.

3.4 An Example I

Reconsider the patient example. Suppose we want to find the a PI on mean patient satisfaction for the point where age=75 and severity=60. The fitted mean response $\hat{y}(\mathbf{x_0}) = 32.78$, the standard error

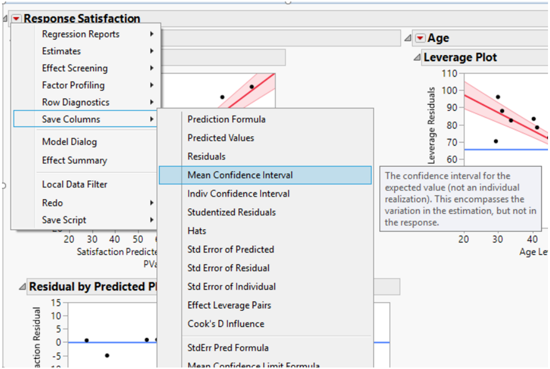
$$se(\hat{y}(\mathbf{x_0})) = \sqrt{\hat{\sigma}^2[1 + \mathbf{x_0}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}]} = 7.65, \tag{56}$$

and $t_{0.05/2,22} = 2.074$.

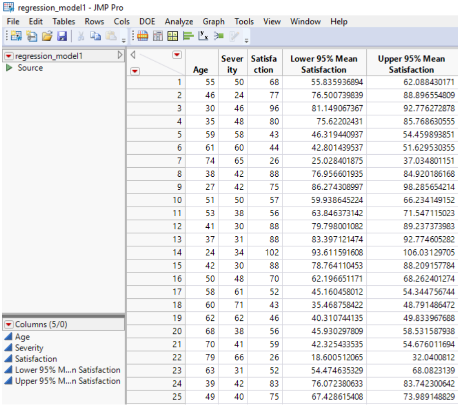
So a 95% PI on the mean patient satisfaction for the casse where age=75 and severity=60 is $(16.93, 48.64)$

JMP for CI and PI I

1. Response Satisfaction (Red Triangle)
2. **Save Columns** (Right Triangle).
3. To get the CI for mean response, select **Mean Confidence Interval**. Otherwise, select other options according to your needs.



4. Columns will be added to the original data table based on your selection in step 3. The output for **Mean Confidence Interval** is presented below.



	Age	Severity	Satisfaction	Lower 95% Mean Satisfaction	Upper 95% Mean Satisfaction
1	55	50	68	55.835936894	62.088430171
2	46	24	77	76.500739839	88.896554809
3	30	46	96	81.149067367	92.776272878
4	35	48	80	75.62202431	85.768630555
5	59	58	43	46.319440937	54.459893851
6	61	60	44	42.801439537	51.629530355
7	74	65	26	25.028401875	37.034801151
8	38	42	88	76.956601935	84.920186168
9	27	42	75	86.274308997	98.285654214
10	51	50	57	59.938645224	66.234149152
11	53	38	56	63.846373142	71.547115023
12	41	30	88	79.798001082	89.237373983
13	37	31	88	83.397121474	92.774605282
14	24	34	102	93.611591608	106.03129705
15	42	30	88	78.764110453	88.209157784
16	50	48	70	62.196651171	68.262401274
17	58	61	52	45.160458012	54.344756744
18	60	71	43	35.468758422	48.791486472
19	62	62	46	40.310744135	49.833967688
20	68	38	56	45.930297809	58.531587938
21	70	41	59	42.325433535	54.676011694
22	79	66	26	18.600512065	32.0400812
23	63	31	52	54.474635329	68.0823139
24	39	42	83	76.072380633	83.742300642
25	49	40	75	67.428615408	73.989148829

3.5 Model Adequacy Checking I

An important part of any data analysis and model-building procedure is checking the adequacy of the model.

Model adequacy checking is particularly important in building regression models for purposes of forecasting. Because forecasting will almost always involve some extrapolation or projection of the model into the future, and unless the model is reasonable the forecasting process is almost certainly doomed to failure.

3.5.1 Residual Plots I

Regression model residuals are very useful in model adequacy checking and to get some sense of how well the regression model assumptions of normally and independently distributed model errors with constant variance are satisfied.

$y_i, i = 1, 2, \cdots, n$. Observed value of the response variable.
$\hat{y}_i, i = 1, 2, \cdots, n$. Fitted value from the model.
$e_i = y_i - \hat{y}_i, i = 1, 2, \cdots, n$. Residuals.

3.5.1 Residual Plots II

Residual plots are the primary approach to model adequacy checking.

1. Normality - Normal probability plot/ QQ plot.
2. Constant Variance - residuals vs fitted values. If the constant variance assumption is satisfied, this plot should exhibit a random scatter of residuals around zero. If this plot shows a pattern, such as an outward-going funnel pattern, indicating that the variance of the observations is increasing as the mean increases. Data **transformations** are useful in stabilizing the variance.

3.5.1 Residual Plots III

3. When the data are a time series, plot the residuals vs time order. The anticipated pattern on this plot is random scatter. Trends, cycles, or other patterns in the plot of residuals vs time indicate model inadequacies, possibly due to missing terms or some other model specification issue. A funnel-shaped pattern that increases in width with time is an indication that the variance of the time series is increasing with time. This happens frequently in economic time series data, and in data that span a long period of time. Log transformations are often useful in stabilizing the variance of these types of time series.

3.5.1 Residual Plots IV

The residual plots for the regression model for the patient satisfaction data.

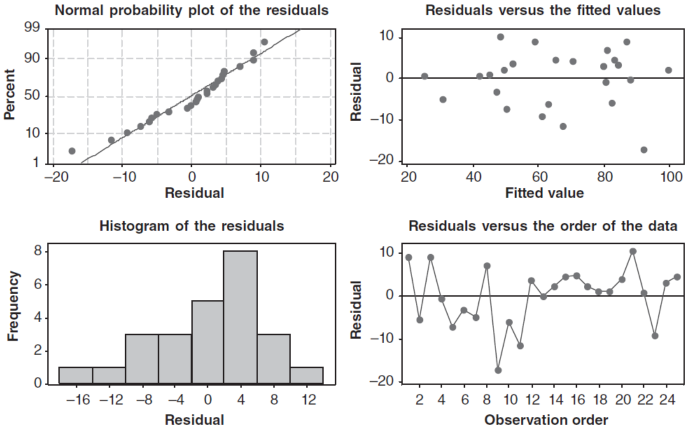


FIGURE 3.1 Plots of residuals for the patient satisfaction model.

1. The residuals lie generally along a straight line, so there so there is no obvious reason to be concerned with the normality assumption. There is a very mild indication that one of the residuals (in the lower tail) may be slightly larger than expected, so this could be an indication of an outlier (a very mild one).
2. Histograms are more useful for large samples of data than small ones, so since there are only 25 residuals, this display is probably not as reliable as the normal probability plot. However, the histogram does not give any serious indication of nonnormality.
3. The Residual vs Fitted value plot indicates essentially random scatter in the residuals, the ideal pattern. If this plot had exhibited a funnel shape, it could indicate problems with the equality of variance assumption.
4. If Residual vs Observation order plot this was the order in which the data were collected, or if the data were a time series, this plot could reveal information about how the data may be changing over time. For example, a funnel shape on this plot might indicate that the variability of the observations was changing with time.

3.5.2 Scaled Residuals and PRESS I

1. **Standardized Residuals**

   Scaled residuals frequently convey more information than do the ordinary residuals. One type of scaled residuals is the Standardized Residual,

$$d_i = \frac{e_i}{\hat{\sigma}}, \quad i = 1, 2, \cdots, n \tag{57}$$

   where $\hat{\sigma}$ is computed by the square root of MSE: $\hat{\sigma} = \sqrt{MSE}$.

   Since the standardized residuals have mean zero and unit variance = 1. Consequently, they are useful in looking for **outliers**. Any observation with a standardized residual outside this interval $-3 \leq d_i \leq 3$ is a potential outlier.

2. **Studentized Residuals**
The standardizing process in equation (57) scales the residuals by dividing them by their approximate average standard deviation. For some data sets, residuals may have different standard deviations. To address this issue, we can estimate the standard deviations by the following method.

$$\underset{\sim}{\hat{y}} = \mathbf{X}\underset{\sim}{\hat{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underset{\sim}{y} = \mathbf{H}\underset{\sim}{y}, \tag{58}$$

where the $n \times n$ matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is usually called the "hat" matrix because it maps the vector of observed values into a vector of fitted values. Then

$$\underset{\sim}{e} = \underset{\sim}{y} - \underset{\sim}{\hat{y}} = \underset{\sim}{y} - \mathbf{H}\underset{\sim}{y} = (\mathbf{I} - \mathbf{H})\underset{\sim}{y} \tag{59}$$

3.5.2 Scaled Residuals and PRESS III

The covariance matrix for $\underset{\sim}{e}$ is

$$
\begin{aligned}
Cov(\underset{\sim}{e}) &= V((\mathbf{I} - \mathbf{H})\underset{\sim}{y}) && (60)\\
&= (\mathbf{I} - \mathbf{H})V(\underset{\sim}{y})(\mathbf{I} - \mathbf{H})' && (61)\\
&= \sigma^2((\mathbf{I} - \mathbf{H}))(\mathbf{I} - \mathbf{H})' && (62)\\
&= \sigma^2(\mathbf{I} - \mathbf{H}) && (63)
\end{aligned}
$$

Denote the diagonal elements of $\mathbf{H}$ by h_{ii}. We get the variance of the ith residual is

$$Var(e_i) = \sigma^2(1 - h_{ii}) \leq \sigma^2 \tag{64}$$

where $0 \leq h_{ii} \leq 1$. Notice that this shows the MSE overestimates the variance of residuals.

Then the **studentized residuals** are

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}} \tag{65}$$

The studentized residuals have unit variance (i.e., $V(r_i) = 1$).

Large values of the studentized residuals are usually an indication of potential unusual values or outliers in the data.

Absolute values of the studentized residuals that are larger than three or four indicate potentially problematic observations.

3.5.2 Scaled Residuals and PRESS V

TABLE 3.6 Residuals and Other Diagnostics for the Regression Model for the Patient Satisfaction Data in Example 3.1

Observation	Residuals	Studentized Residuals	R-Student	h_{ii}	Cook's Distance
1	9.0378	1.29925	1.32107	0.044855	0.026424
2	−5.6986	−0.88216	−0.87754	0.176299	0.055521
3	9.0373	1.38135	1.41222	0.155114	0.116772
4	−0.6953	−0.10403	−0.10166	0.118125	0.000483
5	−7.3897	−1.08009	−1.08440	0.076032	0.031999
6	−3.2155	−0.47342	−0.46491	0.089420	0.007337
7	−5.0316	−0.77380	−0.76651	0.165396	0.039553
8	7.0616	1.03032	1.03183	0.072764	0.027768
9	−17.2800	−2.65767	−3.15124	0.165533	0.467041
10	−6.0864	−0.87524	−0.87041	0.045474	0.012165
11	−11.6967	−1.70227	−1.78483	0.068040	0.070519
12	3.4823	0.51635	0.50757	0.102232	0.010120
13	−0.0859	−0.01272	−0.01243	0.100896	0.000006
14	2.1786	0.33738	0.33048	0.176979	0.008159
15	4.5134	0.66928	0.66066	0.102355	0.017026
16	4.7705	0.68484	0.67634	0.042215	0.006891
17	2.2474	0.33223	0.32541	0.096782	0.003942
18	0.8699	0.13695	0.13386	0.203651	0.001599
19	0.9276	0.13769	0.13458	0.104056	0.000734
20	3.7691	0.58556	0.57661	0.182192	0.025462
21	10.4993	1.62405	1.69133	0.175015	0.186511
22	0.6797	0.10725	0.10481	0.207239	0.001002
23	−9.2785	−1.46893	−1.51118	0.212456	0.194033
24	3.0927	0.44996	0.44165	0.067497	0.004885
25	4.2911	0.61834	0.60945	0.049383	0.006621

3.5.2 Scaled Residuals and PRESS VI

To illustrate the calculations, consider the 1st and 9th observation. The studentized residuals are calculated as follows:

$$r_1 = \frac{e_1}{\sqrt{\hat{\sigma}^2(1-h_{11})}} \tag{66}$$

$$= \frac{e_1}{\hat{\sigma}\sqrt{(1-h_{11})}} \tag{67}$$

$$= \frac{9.0378}{7.11767\sqrt{1-0.044855}} \tag{68}$$

$$= 1.2992 \tag{69}$$

$$r_9 = \frac{e_9}{\hat{\sigma}\sqrt{(1-h_{99})}} \tag{70}$$

$$= \frac{-17.2800}{7.11767\sqrt{1-0.165533}} \tag{71}$$

$$= -2.65767 \tag{72}$$

3.5.2 Scaled Residuals and PRESS VII

Note that none of the studentized residuals in Table 3.7 is this large. The largest studentized residual, 2.65767, is associated with observation 9. This observation does show up on the normal probability plot of residuals in Figure 3.1 as a very mild outlier, but there is no indication of a significant problem with this observation.

3. **PRESS**: The Prediction Error Sum of Squares (PRESS)
To calculate PRESS, we select an observation, for example, i. We fit the regression model to the remaining $n-1$ observations and use this equation to predict the withheld observation y_i. Denoting this predicted value by $\hat{y}_{(i)}$.

$$e_{(i)} = y_i - \hat{y}_{(i)}, \text{Prediction Error for the } i\text{th observation.}$$

$$\text{PRESS} = \sum_{i=1}^{n} e_{(i)}^2 \tag{73}$$

$$R^2_{Prediction} = 1 - \frac{PRESS}{SS_T}. \tag{74}$$

Small PRESS values imply that the model is useful in predicting new observations. $R^2_{Prediction}$ is similar to the ordinary $R^2 = 1 - \frac{SS_E}{SS_T}$. The model would be expected to explain $R^2_{Prediction}$ of the variability in new data.

3.5.2 Scaled Residuals and PRESS X

A useful remark is that for linear regression model, we do not need to fit the model n times to find PRESS. In fact, we can compute $e_{(i)}$ as

$$e_{(i)} = \frac{e_i}{1 - h_{ii}} \tag{75}$$

In the patient example, the value of PRESS is calculated as follow:

$$\begin{aligned}
\text{PRESS} &= \sum_{i=1}^{n} \frac{e_i^2}{(1 - h_{ii})^2} \\
&= \frac{9.037816^2}{(1 - 0.044854)^2} + \frac{(-5.698647)^2}{(1 - 0.176299)^2} + \cdots + \frac{(4.291117)^2}{(1 - 0.049383)^2} \\
&= 1484.9335
\end{aligned}$$

3.5.2 Scaled Residuals and PRESS XI

If the value of the $R^2_{Prediction}$ is not much smaller than the ordinary R^2, this is a good indication about potential model predictive performance.

$$R^2_{Prediction} = 1 - \frac{PRESS}{SS_T} = 1 - 1484.9335/10778.2 = 0.8622. \quad (76)$$

The model would be expected to explain about 86.22% of the variability in new data.

4. **R-student / Externally Studentized Residual**

Review: The (**internally**) Studentized Residuals $r_i, i = 1, \ldots, n$ is often considered an outlier diagnostic. It is defined as

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}} \tag{77}$$

where $\hat{\sigma} = MS_E$. This is referred to as **internal scaling** of the residual because MS_E is an internally generated estimate of σ^2 obtained from fitting the model to all n observations.

3.5.2 Scaled Residuals and PRESS XIII

Another approach would be to use an estimate of σ^2 based on a data set with the ith observation removed. We denote the estimate of σ^2 so obtained by $S_{(i)}^2$.

R-student for ith observation is defined as

$$t_i = \frac{e_i}{\sqrt{S_{(i)}^2(1 - h_{ii})}} \tag{78}$$

where $S_{(i)}^2$ is an estimate of σ^2 with the ith observation removed,

$$S_{(i)}^2 = \frac{(n-p)MS_E - e_i^2/(1-h_{ii})}{n-p-1}. \tag{79}$$

Most of the time, t_i will be closed to r_i. However, for influential observation, they can differ significantly. See Table 3.6.

3.5.2 Scaled Residuals and PRESS XIV

Under the standard assumptions, the R-student residuals $t_i \sim t_{n-p-1}$. Thus we compare t_i with a t critical value $t_{\alpha,n-p-1}$ to determine if the i observation is an influential observation or not.

3.5.3 Measures of Leverage and Influence I

It is possible that a small subset of the observations have a large influence on the fitted model. Then we need to locate them and check the influence.

Recall the hat matrix

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

And

$$\underset{\sim}{\hat{y}} = \mathbf{H}\underset{\sim}{y}$$

Then the fitted value for ith observation is

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j \tag{80}$$

Therefore the element h_{ij} can be interpreted as the amount of **leverage** exerted by the observation y_j on the predicted value $\hat{y}_i$. So h_{ij} is a useful tool in identifying influential observations.

3.5.3 Measures of Leverage and Influence II

Meanwhile it can be shown that

$$\sum_{i=1}^{n} h_{ii} = rank(\mathbf{H}) = rank(\mathbf{X}) = p,$$

So the aveage size of the diagonal elements of the $\mathbf{H}$ matrix is $\frac{p}{n}$. A widely used rough guideline is to compare h_{ii} to $\frac{2p}{n}$ which is twice the average. If any hat diagonal exceeds $\frac{2p}{n}$ to consider that observation as a high-leverage point.

3.5.3 Measures of Leverage and Influence III

In the patient example,

$$\frac{2p}{n} = \frac{2(3)}{25} = 0.24.$$

The largest hat diagonal is $h_{23,23} = 0.212456$. This does not exceed twice the average size of a hat diagonal, so there are no observations that are considered as high-leverage ones.

3.5.3 Measures of Leverage and Influence IV

When considering **influential** observations, it is desirable to look at the response variable as well in addition to $\mathbf{X}$.

Cook's distance for ith observation is defined as

$$D_i = \frac{(\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_{(\mathbf{i})})'\mathbf{X}'\mathbf{X}(\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_{(\mathbf{i})})}{p \cdot MSE} = \frac{r_i^2}{p} \cdot \frac{h_{ii}}{1 - h_{ii}}, \quad i = 1, 2, \cdots, n \qquad (81)$$

where $\underset{\sim}{\beta}_{(i)}$ is the LSE obtained by deleting the ith observation. We usually consider observation for which $D_i > 1$ to be influential.

3.5.3 Measures of Leverage and Influence V

In the patient example, observation 9 has the largest value of D_i which is computed from

$$D_9 = \frac{(-2.65767)^1}{3} \cdot \frac{0.165533}{1 - 0.165533} = 0.467 < 1. \tag{82}$$

So there are no influential observations in the data set.

3.6 Variable Selection Methods in Regression I

Model selection: selecting the best subset from all candidate predictors.

1. All possible regressions: All 2^K possible models are examined to identify the potentially good one(s), where K is the total number of predictors. One can see the when K is large, the number of possible models becomes out of control rapidly.

2. Stepwise selection:
   - Forward Selection: begin with a model with no predictor, then sequentially insert predictors into the model until no remaining predictors are good enough to enter the model.

   - Backward Selection: begin with a full model, then sequentially remove predictors to produce a final model.

Reconsider the patient example with some additional variables.

TABLE 3.8 Expanded Patient Satisfaction Data

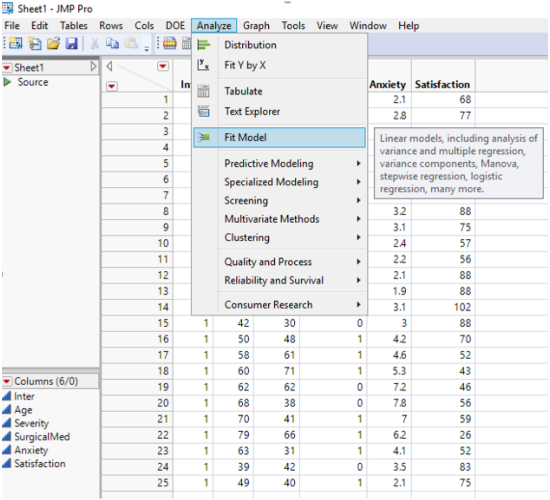
Observation	Age	Severity	Surgical–Medical	Anxiety	Satisfaction
1	55	50	0	2.1	68
2	46	24	1	2.8	77
3	30	46	1	3.3	96
4	35	48	1	4.5	80
5	59	58	0	2.0	43
6	61	60	0	5.1	44
7	74	65	1	5.5	26
8	38	42	1	3.2	88
9	27	42	0	3.1	75
10	51	50	1	2.4	57
11	53	38	1	2.2	56
12	41	30	0	2.1	88
13	37	31	0	1.9	88
14	24	34	0	3.1	102
15	42	30	0	3.0	88
16	50	48	1	4.2	70
17	58	61	1	4.6	52
18	60	71	1	5.3	43
19	62	62	0	7.2	46
20	68	38	0	7.8	56
21	70	41	1	7.0	59
22	79	66	1	6.2	26
23	63	31	1	4.1	52
24	39	42	0	3.5	83
25	49	40	1	2.1	75

Forward selection starts with a model with no predictors.

```
> satisfaction3.fit <- lm(satisfaction~ 1, data=patientex)
> step.for <- step(satisfaction3.fit, direction='forward',scope=~age + severity + surgmed + anxiety )
Start:  AIC=153.66
satisfaction ~ 1

           Df Sum of Sq     RSS    AIC
+ age       1    8756.7  2021.6 113.82
+ severity  1    5634.3  5143.9 137.17
+ anxiety   1    3100.2  7678.0 147.18
<none>                  10778.2 153.66
+ surgmed   1     589.9 10188.4 154.25

Step:  AIC=113.82
satisfaction ~ age

           Df Sum of Sq    RSS    AIC
+ severity  1    907.04 1114.5 100.93
<none>                  2021.6 113.82
+ anxiety   1      9.81 2011.8 115.70
+ surgmed   1      1.79 2019.8 115.80

Step:  AIC=100.93
satisfaction ~ age + severity

          Df Sum of Sq    RSS    AIC
<none>                 1114.5 100.93
+ anxiety  1    74.611 1039.9 101.20
+ surgmed  1     0.163 1114.4 102.93
```

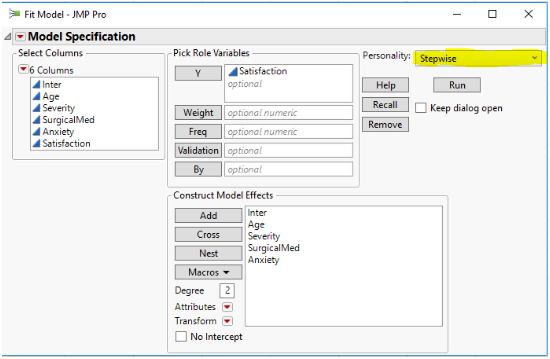
One can see the procedure stops until age and severity are added.

1. Import the patient data and **Fit Model**

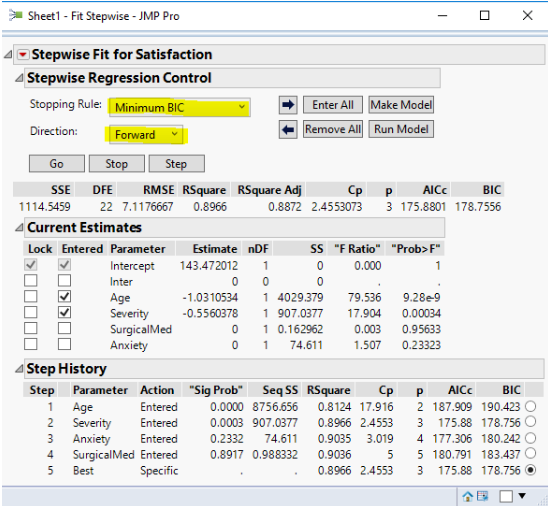


Forward Selection by JMP II

2. Select **Stepwise** among the drop-down options for **Personality**.

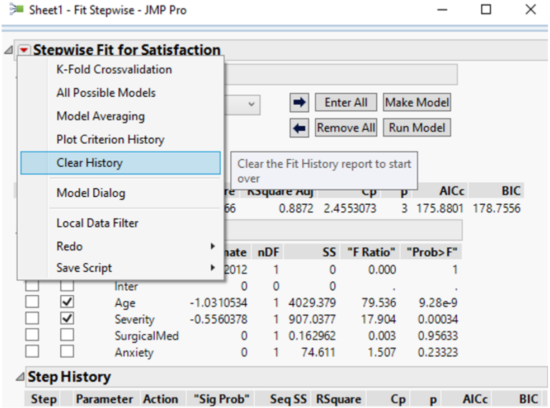


3. The default criterion is **Minumum BIC** and use **Forward** selection. Click **Go** to start the selection.



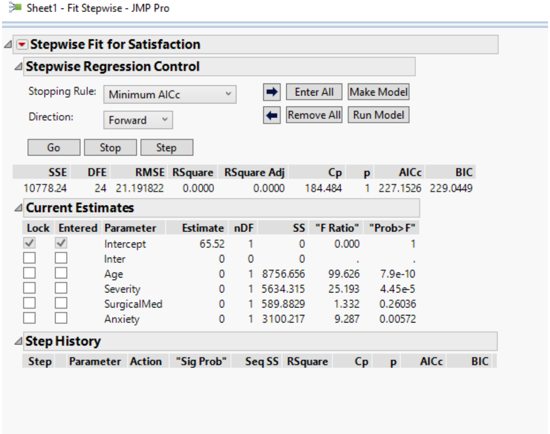
Forward Selection by JMP IV

4. Clear Step history.



Forward Selection by JMP V

5. Select **Remove All** to re-start the selection by **Forward** method. If you like, you can change the selection criterion to **Minimum AICc**. Then click **Go**.



Sheet1 - Fit Stepwise - JMP Pro

Stepwise Fit for Satisfaction

Stepwise Regression Control

Stopping Rule: Minimum AICc

Direction: Forward

Enter All | Make Model | Remove All | Run Model

Go | Stop | Step

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
10778.24	24	21.191822	0.0000	0.0000	184.484	1	227.1526	229.0449

Current Estimates

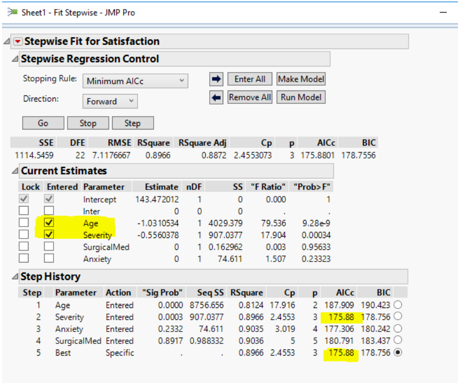
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
✓	✓	Intercept	65.52	1	0	0.000	1
		Inter	0	0	0	.	.
		Age	0	1	8756.656	99.626	7.9e-10
		Severity	0	1	5634.315	25.193	4.45e-5
		SurgicalMed	0	1	589.8829	1.332	0.26036
		Anxiety	0	1	3100.217	9.287	0.00572

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC

Forward Selection by JMP VI

6. Two variables **Age** and **Severity** are selected into the model to get the minimmum AICc 175.88. See highlighted.



Sheet1 - Fit Stepwise - JMP Pro

Stepwise Fit for Satisfaction

Stepwise Regression Control

Stopping Rule: Minimum AICc

Direction: Forward

Go | Stop | Step | Enter All | Make Model | Remove All | Run Model

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
1114.5459	22	7.1176667	0.8966	0.8872	2.4553073	3	175.8801	178.7556

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	143.472012	1	0	0.000	1
☐	☐	Inter	0	0	0	.	.
☐	☑	Age	-1.0310534	1	4029.379	79.536	9.28e-9
☐	☑	Severity	-0.5560378	1	907.0377	17.904	0.00034
☐	☐	SurgicalMed	0	1	0.162962	0.003	0.95633
☐	☐	Anxiety	0	1	74.611	1.507	0.23323

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	Age	Entered	0.0000	8756.656	0.8124	17.916	2	187.909	190.423	○
2	Severity	Entered	0.0003	907.0377	0.8966	2.4553	3	175.88	178.756	○
3	Anxiety	Entered	0.2332	74.611	0.9035	3.019	4	177.306	180.242	○
4	SurgicalMed	Entered	0.8917	0.988332	0.9036	5	5	180.791	183.437	○
5	Best	Specific	.	.	0.8966	2.4553	3	175.88	178.756	◉

7. Table 3.9

TABLE 3.9 JMP Forward Selection for the Patient Satisfaction Data in Table 3.8

Stepwise Fit for Satisfaction

Stepwise Regression Control

Stopping rule: Minimum AICc

Direction: Forward

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
1114.5459	22	7.1176667	0.8966	0.8872	2.4553073	3	175.8801	178.7556

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	143.472012	1	0	0.000	1
☐	☑	Age	−1.0310534	1	4029.379	79.536	9.28e-9
☐	☑	Severity	−0.5560378	1	907.0377	17.904	0.00034
☐	☐	Surg-Med	0	1	0.162962	0.003	0.95633
☐	☐	Anxiety	0	1	74.611	1.507	0.23323

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	Age	Entered	0.0000	8756.656	0.8124	17.916	2	187.909	190.423	○
2	Severity	Entered	0.0003	907.0377	0.8966	2.4553	3	175.88	178.756	○
3	Anxiety	Entered	0.2332	74.611	0.9035	3.019	4	177.306	180.242	○
4	Surg-Med	Entered	0.8917	0.988332	0.9036	5	5	180.791	183.437	○
5	Best	Specific	.	.	0.8966	2.4553	3	175.88	178.756	◉

Backward Selection by R I

Backward selection starts with a full model.

```
> satisfaction4.fit <- lm(satisfaction~ age +severity +surgmed +anxiety, data=patientex)
> step.back <- step(satisfaction4.fit, direction='backward')
Start:  AIC=103.18
satisfaction ~ age + severity + surgmed + anxiety

           Df Sum of Sq    RSS    AIC
- surgmed   1       1.0 1039.9 101.20
- anxiety   1      75.4 1114.4 102.93
<none>                  1038.9 103.18
- severity  1     971.5 2010.4 117.68
- age       1    3387.7 4426.6 137.41

Step:  AIC=101.2
satisfaction ~ age + severity + anxiety

           Df Sum of Sq    RSS    AIC
- anxiety   1      74.6 1114.5 100.93
<none>                  1039.9 101.20
- severity  1     971.8 2011.8 115.70
- age       1    3492.7 4532.6 136.00

Step:  AIC=100.93
satisfaction ~ age + severity

           Df Sum of Sq    RSS    AIC
<none>                  1114.5 100.93
- severity  1     907.0 2021.6 113.82
- age       1    4029.4 5143.9 137.17
```

One can see the procedure stops until surgmed and anxiety are removed.

Backward Selection by R II

R lists the best model of each size.

```
> step.best <- regsubsets(satisfaction ~ age+severity+surgmed+anxiety, data=patientex)
> summary(step.best)
Subset selection object
Call: regsubsets.formula(satisfaction ~ age + severity + surgmed +
    anxiety, data = patientex)
4 Variables  (and intercept)
         Forced in Forced out
age          FALSE      FALSE
severity     FALSE      FALSE
surgmed      FALSE      FALSE
anxiety      FALSE      FALSE
1 subsets of each size up to 4
Selection Algorithm: exhaustive
         age severity surgmed anxiety
1  ( 1 ) "*" " "      " "     " "
2  ( 1 ) "*" "*"      " "     " "
3  ( 1 ) "*" "*"      " "     "*"
4  ( 1 ) "*" "*"      "*"     "*"
```

The best model with 1 predictor is age; the best model with 2 predictors is age+severity; the best model with 3 predictors is age+severity+anxiety.

Backward Selection by JMP I

TABLE 3.10 JMP Backward Elimination for the Patient Satisfaction Data in Table 3.8

Stepwise Fit for Satisfaction

Stepwise Regression Control

Stopping rule: Minimum AICc

Direction: Backward

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
1114.5459	22	7.1176667	0.8966	0.8872	2.4553073	3	175.8801	178.7556

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	143.472012	1	0	0.000	1
☐	☑	Age	−1.0310534	1	4029.379	79.536	9.28e-9
☐	☑	Severity	−0.5560378	1	907.0377	17.904	0.00034
☐	☐	Surg-Med	0	1	0.162962	0.003	0.95633
☐	☐	Anxiety	0	1	74.611	1.507	0.23323

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	Age	Entered	0.0000	8756.656	0.8124	17.916	2	187.909	190.423	○
2	Severity	Entered	0.0003	907.0377	0.8966	2.4553	3	175.88	178.756	○
3	Surg-Med	Entered	0.9563	0.162962	0.8966	4.4522	4	179.034	181.971	○
4	Anxiety	Entered	0.2422	75.43637	0.9036	5	5	180.791	183.437	○
5	Surg-Med	Removed	0.8917	0.988332	0.9035	3.019	4	177.306	180.242	○
6	Anxiety	Removed	0.2332	74.611	0.8966	2.4553	3	175.88	178.756	○
7	Severity	Removed	0.0003	907.0377	0.8124	17.916	2	187.909	190.423	○
8	Best	Specific	.	.	0.8966	2.4553	3	175.88	178.756	◉

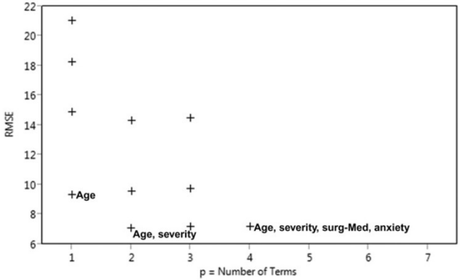
Backward Selection by JMP II

TABLE 3.12 JMP All Possible Models Regression for the Patient Satisfaction Data in Table 3.8

All Possible Models

Ordered up to best 4 models up to 4 terms per model.

Model	Number	RSquare	RMSE	AICc	BIC
Age	1	0.8124	9.3752	187.909	190.423
Severity	1	0.5227	14.9549	211.257	213.771
Anxiety	1	0.2876	18.2709	221.271	223.785
Surg-Med	1	0.0547	21.0469	228.343	230.857
Age, severity	2	0.8966	7.1177	175.880	178.756
Age, anxiety	2	0.8133	9.5626	190.644	193.520
Age, surg-Med	2	0.8126	9.5817	190.744	193.619
Severity, anxiety	2	0.5795	14.3537	210.951	213.827
Age, severity, anxiety	3	0.9035	7.0371	177.306	180.242
Age, severity, surg-Med	3	0.8966	7.2846	179.034	181.971
Age, surg-Med, anxiety	3	0.8135	9.7844	193.785	196.722
Severity, surg-Med, anxiety	3	0.5893	14.5186	213.518	216.454
Age, severity, surg-Med, anxiety	4	0.9036	7.2074	180.791	183.437



3.7 Generalized and Weighted Least Squares I

Nonconstant error variance issue

A previous approach is to apply a transformation of the response variable. Another thing we can do is to use the method of weighted least squares.

For example, for the simple linear regression, the original Least Square estimates minimize

$$L = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{2} (y_i - \beta_0 - \beta_1 x_i)^2$$

If we introduce a sequence of weights $w_i = \frac{1}{\sigma_i^2}$ which is the reciprocal of the variance of y_i, then the weighted least squares estimates minimize

$$L = \sum_{i=1}^{n} w_i e_i^2 = \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_i)^2$$

3.7.1 Generalized Least Squares I

Sometimes, it is unrealistic to assume the error terms are uncorrelated, i.e. $Var(\epsilon) = \sigma^2 \mathbf{I}$. We now assume $Var(\epsilon) = \sigma^2 \mathbf{V}$, where $\mathbf{V}$ is also a $n \times n$ matrix. $\mathbf{V}$ must be a nonsingular and positive definite matrix so it can be decomposited as $\mathbf{V} = \mathbf{KK}$, where $\mathbf{K}$ is symmetric, i.e. $\mathbf{K} = \mathbf{K}'$.

We use some algebra tricks to solve the model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \quad E(\epsilon) = 0 \quad \text{and} \quad Var(\epsilon) = \sigma^2 \mathbf{V}$$

Multiplying $\mathbf{K}^{-\mathbf{1}}$ on the left to get

$$\mathbf{K}^{-\mathbf{1}}\mathbf{y} = \mathbf{K}^{-\mathbf{1}}\mathbf{X}\beta + \mathbf{K}^{-\mathbf{1}}\epsilon$$

It can be shown that the "new" error term $\mathbf{K}^{-\mathbf{1}}\epsilon$ satisfy mean=0 and variance= $\sigma^2 \mathbf{I}$, which bring us back to the assumption of the ordinary least square case.

3.7.1 Generalized Least Squares II

So the estimate of β, $\hat{\beta}_{GLS}$ is equal to

$$\hat{\beta}_{GLS} = (\mathbf{X}'\mathbf{K}^{-\mathbf{1}'}\mathbf{K}^{-\mathbf{1}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}^{-\mathbf{1}'}\mathbf{K}^{-\mathbf{1}}\mathbf{y} = (\mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{y}$$

Notice here we used the facts $\mathbf{K}^{-\mathbf{1}'} = \mathbf{K}^{-\mathbf{1}}$, and $\mathbf{K}^{-\mathbf{1}}\mathbf{K}^{-\mathbf{1}} = \mathbf{V}^{-\mathbf{1}}$. It can be shown the GLS estimator $\hat{\beta}_{GLS}$ is an unbiased estimator.

$$E(\hat{\beta}_{GLS}) = (\mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-\mathbf{1}}E(\mathbf{y}) \tag{83}$$

$$= (\mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-\mathbf{1}}X\beta \tag{84}$$

$$= \beta \tag{85}$$

The covariance matrix is

$$Var(\hat{\beta}_{GLS}) = \sigma^2(\mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{X})^{-1} \tag{86}$$

The $\hat{\beta}_{GLS}$ is a best linear unbiased estimator of the model parameters β, where "best" means the minimum variance.

3.7.2 Weighted Least Squares I

Weighted least square is a special case of GLS where the matrix $\mathbf{V}$ is a diagonal matrix with diagonal elements $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$.

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Let $\mathbf{W} = \mathbf{V}^{-1}$.

$$\mathbf{W} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix}$$

3.7.2 Weighted Least Squares II

Therefore the WLS estimator is

$$\hat{\beta}_{WLS} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}.$$

The WLS estimator is an unbiased estimator

$$E(\hat{\beta}_{WLS}) = \beta. \tag{87}$$

The covariance matrix of $\hat{\beta}_{WLS}$ is

$$Var(\hat{\beta}_{WLS}) = \sigma^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}.$$

The ideal case is that the weights $w_i = \frac{1}{\sigma_i^2}$ is known from prior knowledge or experience or research.
However, most of the time, they are unknown and need to be estimated.

3.7.2 Weighted Least Squares III

Case 1: Prior knowledge or experience or information from an underlying theoretical model can be used to determine the weights.

For example, suppose that a significant sources of error is **measurement error** and different observations are measured by different instruments of unequal but know or well-estimated accuracy. Then the weights could be chosen **inversely proportional** to the variance of measurement error.

Case 2: Inequality of variance.

1. Method 1: Estimation of a Variance Equation.

   Suppose: The plot of the OLS residuals e_i vs $\hat{y}_i$ exhibit an outward-opening funnel shape. Plots of the OLS e_i vs x indicate that the variance of the observations is a function of one of the predictors x.

$$\sigma_i^2 = f(x), \text{or}, \sigma_i^2 = g(y),$$
$$\hat{\sigma}_i^2 = e_i^2$$
$$\hat{\sigma}_i = |e_i|$$

3.7.2 Weighted Least Squares V

Then, we can find a **variance equation** by the following process:

- Fit the model by using OLS and find the OLS residuals.
- Use residual analysis to determine potential relationship between σ_i^2 and either the mean of y or some of the predictors.
- Regress the squared OLS residuals or the absolute values of OLS residuals to obtain an equation for predicting the variance of each observation, say $\hat{s}_i^2 = f(x)$ or $g(y)$.
- Use fitted values to obtain estimates of the weights, $w_i = 1/\hat{s}_i^2$
- Form matrix **W** and obtain the WLS estimates.

3.7.2 Weighted Least Squares VI

2. Method 2: Using Replicates or Nearest Neighbors.

   If the data set contains replicate observations meaning the observations with exactly the same values of the predictors, we can use the reciprocal of sample variance to estimate weights w_i. This is usually not practical for time series data.

3.7.2 An example I

The data set contains the strength (y) of a connector and the age in weeks (x) of the glue used to bond the components of the connector together.

TABLE 3.12 Connector Strength Data

Observation	Weeks	Strength	Residual	Absolute Residual	Weights
1	20	34	0.5454	0.5454	73.9274
2	21	35	1.1695	1.1695	5.8114
3	23	33	−1.5824	1.5824	0.9767
4	24	36	1.0417	1.0417	0.5824
5	25	35	−0.3342	0.3342	0.3863
6	28	34	−2.4620	2.4620	0.1594
7	29	37	0.1621	0.1621	0.1273
8	30	34	−3.2139	3.2139	0.1040
9	32	42	4.0343	4.0343	0.0731
10	33	35	−3.3416	3.3416	0.0626
11	35	33	−6.0935	6.0935	0.0474
12	37	46	6.1546	6.1546	0.0371
13	38	43	2.7787	2.7787	0.0332
14	40	32	−8.9731	8.9731	0.0270
15	41	37	−4.3491	4.3491	0.0245
16	43	50	7.8991	7.8991	0.0205
17	44	34	−8.4769	8.4769	0.0189
18	45	54	11.1472	11.1472	0.0174
19	46	49	5.7713	5.7713	0.0161
20	48	55	11.0194	11.0194	0.0139
21	50	40	−4.7324	4.7324	0.0122
22	51	33	−12.1084	12.1084	0.0114
23	52	56	10.5157	10.5157	0.0107
24	55	58	11.3879	11.3879	0.0090
25	56	45	−1.9880	1.9880	0.0085
26	57	33	−14.3639	14.3639	0.0080
27	59	60	11.8842	11.8842	0.0072
28	60	35	−13.4917	13.4917	0.0069

3.7.2 An example II

The OLS regression model is

$$\hat{y} = 25.936 + 0.3759x.$$

The OLS residuals are shown as well. We plot both residual VS Weeks (x) and absolute residual VS weeks (x).

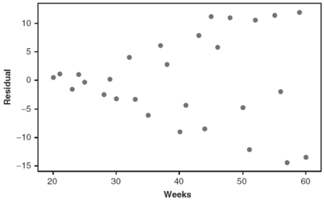


FIGURE 3.3 Plot of residuals versus weeks.

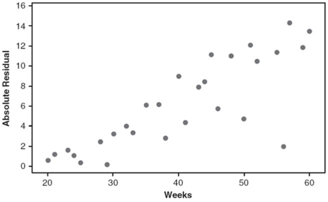


FIGURE 3.4 Scatter plot of absolute residuals versus weeks.

3.7.2 An example III

There is a potential linear relationship between the absolute residuals $|e_i|s$ and weeks.
Regressing the absolute value residuals to the weeks yields

$$\hat{s} = -5.854 + 0.29852x,$$

or

$$\hat{s}_i = -5.854 + 0.29852x_i, i = 1, 2, \cdots, n$$

which is used to fit the s value for each observation. Then the weights are just equal to the reciprocal of the squared s values.

$$w_i = \frac{1}{\hat{s}_i^2} = \frac{1}{(-5.854 + 0.29852x_i)^2}, i = 1, 2, \cdots, n$$

3.7.2 An example IV

Plug w_is into the WLS estimator to obtain the $\hat{\beta}_{WLS}$,

$$\hat{\beta}_{WLS} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}.$$

Finally, the model fitted by WLS method is

$$\hat{y} = 27.545 + 0.32383x.$$

3.7.2 An example V

If the WLS estimates differ significantly from their OLS counterparts,

- use a new WLS residuals
- reestimate the variance equation to produce a new set of weights and a revised set of WLS estimates using these new weights.

This procedure is called **iteratively reweighted least squares**.

In this example, the difference between the two models is acceptable, which is an indication that no iteration is needed.
OLS regression model: $\hat{y} = 25.963 + 0.3759x$.
WLS regression model: $\hat{y} = 27.545 + 0.32383x$.

Discounted Least Squares DLS I

The Weighted Least Squares estimates can be used to a time series linear model as well.
For examples:

$$y_t = \beta_0 + \beta_1 t + \epsilon$$

$$y_t = \beta_0 + \beta_1 \sin \frac{2\pi}{d} t + \beta_2 \cos \frac{2\pi}{d} t + \epsilon$$

The general form of this type of models is

$$y_t = \beta_1 x_1(t) + ... + \beta_p x_p(t) + \epsilon_t = \mathbf{x}(t)'\beta + \epsilon$$

where $\mathbf{x}(t)' = (x_1(t), \ldots, x_p(t))$ and $\beta' = (\beta_1, \ldots, \beta_p)$

Discounted Least Squares DLS II

The matrix form is

$$\mathbf{y} = \mathbf{X}(T)\beta + \epsilon$$

where

$$\mathbf{X}(T) = \begin{bmatrix} x_1(1) & x_2(1) & x_3(1) & \dots & x_p(1) \\ x_1(2) & x_2(2) & x_3(2) & \dots & x_p(2) \\ \dots & \dots & \dots & \dots & \dots \\ x_1(T) & x_2(T) & x_3(T) & \dots & x_p(T) \end{bmatrix}$$

Now consider the weights. It seems reasonable to assume the recent observations are weighted more heavily than older ones. Specifically, let the weight for observation y_{T-j} be θ^j, for some $0 < \theta < 1$. So the weight matrix

$$\mathbf{W} = \begin{bmatrix} \theta^{T-1} & 0 & 0 & \dots & 0 \\ 0 & \theta^{T-2} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \theta^0 \end{bmatrix}$$

Discounted Least Squares DLS III

From discussion for Weighted Least Square estimates, we have

$$\hat{\beta}(T) = (\mathbf{X}(T)'\mathbf{W}\mathbf{X}(T))^{-1}\mathbf{X}(T)'\mathbf{W}\mathbf{y}$$

Let $\mathbf{G}(T) = \mathbf{X}(T)'\mathbf{W}\mathbf{X}(T)$ and $\mathbf{g}(T) = \mathbf{X}(T)'\mathbf{W}\mathbf{y}$, then we have

$$\hat{\beta}(T) = \mathbf{G}(T)^{-1}\mathbf{g}(T)$$

$\hat{\beta}(T)$ is called the **discounted least squares estimator** of β.
We can rewrite

$$\mathbf{G}(T) = \sum_{j=0}^{T-1} \theta^j \mathbf{x}(-j)\mathbf{x}(-j)' \tag{88}$$

$$= \mathbf{G}(T-1) + \theta^{T-1}\mathbf{x}(-(T-1))\mathbf{x}(-(T-1))' \tag{89}$$

Discounted Least Squares DLS IV

In many important applications, the predictor variables $x_i(t)$ are functions linear combinations of their values at the previous time period. That is

$$x_i(t+1) = L_{i1}x_1(t) + \cdots + L_{ip}x_p(t), i = 1, \ldots, p.$$

In matrix form

$$\mathbf{x}(t+1) = \mathbf{L}\mathbf{x}(t),$$

where $\mathbf{x}(t) = [x_1(t), x_2(t) ... x_p(t)]'$, $\mathbf{L}$ is the $p \times p$ matrix of constants L_{ij}. Further,

$$\mathbf{x}(t+2) = \mathbf{L}\mathbf{x}(t+1) = \mathbf{L}\mathbf{L}\mathbf{x}(t) = \mathbf{L}^2\mathbf{x}(t).$$

$$\mathbf{x}(t+k) = \mathbf{L}^k\mathbf{x}(t), k = 1, 2, \cdots .$$

Equivalently,

$$\mathbf{x}(t) = \mathbf{L}^t\mathbf{x}(0), t = 1, 2, \cdots , T.$$

Discounted Least Squares DLS V

For example, $\mathbf{y}_t = \beta_0 + \beta_1 t$,
we have $p = 2$, $x_1(t) = 1$ and $x_2(t) = t$, i.e. $\mathbf{y}_t = \mathbf{x}(t)'\beta = \beta_0 + \beta_1 t$.
where

$$\mathbf{x}(t) = \begin{bmatrix} 1 \\ t \end{bmatrix}.$$

Then

$$\mathbf{L} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

We have

$$\mathbf{x}(t+1) = \mathbf{L}\mathbf{x}(t) = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \mathbf{x}(t) = \begin{bmatrix} 1 \\ t+1 \end{bmatrix}.$$

Thus

$$\mathbf{y}_{t+1} = \mathbf{x}(t+1)'\beta = \beta_0 + \beta_1(t+1).$$

Discounted Least Squares DLS VI

Similar, we have

$$\mathbf{x}(t+2) = \mathbf{L}\mathbf{x}(t+1) = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ t+1 \end{bmatrix} = \begin{bmatrix} 1 \\ t+2 \end{bmatrix}.$$

Or,

$$\mathbf{x}(t+2) = \mathbf{L}^2\mathbf{x}(t) = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ t \end{bmatrix} = \begin{bmatrix} 1 \\ t+2 \end{bmatrix}.$$

$$\mathbf{y}_{t+2} = \mathbf{x}(t+2)'\beta = \beta_0 + \beta_1(t+2).$$

It can be shown the following is true:

Let $\mathbf{G}$ be a steady-state limiting value of $\mathbf{G}(T)$, that is $\mathbf{G} = \lim_{T\to\infty} \mathbf{G}(T)$. Then the DLS estimator satisfies

$$\hat{\beta}(T) = \mathbf{G}(T)^{-1}\mathbf{X}(T)'\mathbf{W}\mathbf{y} = \cdots = \mathbf{L}'\hat{\beta}(T-1) + \mathbf{G}^{-1}\mathbf{x}(0)e_T(1) \quad (90)$$

This is a recurrence relationship between the DLS estimator at time period $T-1$ and at time period T. Here $e_T(1)$ is the one-step-ahead forecast error for time period T. The origin of time is shifted to the end of the current time period, forecasting is easy with discounted least squares. The forecast of the observation at a future time period $T+\tau$, made at the end of the time period T is

$$\hat{y}_{T+\tau}(T) = \mathbf{x}(\tau)'\hat{\beta}(T) = \sum_{j=1}^{p} \mathbf{x}_j(\tau)\hat{\beta}_j(T) \quad (91)$$

An Example DLS I

Consider the simple linear trend model

$$y_t = \beta_0 + \beta_1 t + \epsilon_t, \quad t = 1, 2, ..., T$$

So as we discussed before, the **L** matrix is

$$L = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

$$\mathbf{X}(T) = \begin{bmatrix} 1 & -(T-1) \\ 1 & -(T-2) \\ \vdots & \vdots \\ 1 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{x}(-(T-1))' \\ \mathbf{x}(-(T-2))' \\ \vdots \\ \mathbf{x}(-1)' \\ \mathbf{x}(0)' \end{bmatrix},$$

An Example DLS II

The discounted weight matrix is

$$\mathbf{W} = \begin{bmatrix} \theta^{T-1} & 0 & 0 & \dots & 0 \\ 0 & \theta^{T-2} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \theta^0 \end{bmatrix}$$

After matrix algebra calculation, we get

$$\mathbf{G}(T) = \mathbf{X}(T)'\mathbf{W}\mathbf{X}(T) = \begin{bmatrix} \frac{1-\theta^T}{1-\theta} & -\frac{\theta(1-\theta^T)}{1-\theta} \\ -\frac{\theta(1-\theta^T)}{1-\theta} & \frac{\theta(1+\theta)(1-\theta^T)}{(1-\theta)^3} \end{bmatrix}.$$

Let $T \to \infty$, we have

$$\mathbf{G} = \lim_{T\to\infty} \mathbf{G}(T) = \begin{bmatrix} \frac{1}{1-\theta} & -\frac{\theta}{1-\theta} \\ -\frac{\theta}{1-\theta} & \frac{\theta(1+\theta)}{(1-\theta)^3} \end{bmatrix} \quad \text{and} \quad \mathbf{G}^{-1} = \begin{bmatrix} 1-\theta^2 & (1-\theta)^2 \\ (1-\theta)^2 & \frac{(1-\theta)^2}{\theta} \end{bmatrix}.$$

An Example DLS III

So, we are ready to use equation (90)

$$\hat{\beta}(T) = \mathbf{L}'\hat{\beta}(T-1) + \mathbf{G}^{-1}\mathbf{x}(0)e_T(1)$$

and obtain

$$\hat{\beta}(T) = \begin{bmatrix} \hat{\beta}_0(T) \\ \hat{\beta}_1(T) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0(T-1) \\ \hat{\beta}_1(T-1) \end{bmatrix} + \begin{bmatrix} 1-\theta^2 & (1-\theta)^2 \\ (1-\theta)^2 & \frac{(1-\theta)^2}{\theta} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} e_T(1),$$

which is equivalent to

$$\hat{\beta}_0(T) = \hat{\beta}_0(T-1) + \hat{\beta}_1(T-1) + (1-\theta^2)e_T(1) \tag{92}$$

$$\hat{\beta}_1(T) = \hat{\beta}_1(T-1) + (1-\theta)^2 e_T(1) \tag{93}$$

The equation (92) updates intercept at time T as the old intercept plus the old slope, plus a fraction of the current forecast error. The equation (93) revises the slope by adding a fraction of the current forecast error. It

An Example DLS IV

states that in discounted least squares the vector of parameters estimates computed at the end of the time period T can be computed as a simple linear combination of the estimates made at the end of the previous time period $T - 1$ and the one-step-ahead forecast error for the observation in period T.

An Example DLS V

To illustrate the process, suppose we are forecasting a time series with a linear trend and we have initial estimates at time t=0 of the slope $\hat{\beta}_1(0) = 1.5$ and the intercept $\hat{\beta}_0(0) = 50$.

We use $\theta = 0.9$ and assume the actual observation in time period 1 and 2 are $y_1 = 52$, and $y_2 = 55$ respectively. The process goes as follows:

1. The forecast for time period t=1, made at the end of the time period t=0 (the lead time $\tau = 1$) is computed below:
   $\hat{y}_1(0) = \hat{\beta}(0)'\mathbf{x}(1) = \hat{\beta}_0(0) + \hat{\beta}_1(0)(t) = 50 + 1.5(1) = 51.5$

2. The forecast for time period at time t=2, made at the end of the time period t=1, (the lead time $\tau = 1$) is computed below:
   $e_1(1) = y_1 - \hat{y}_1(0) = 0.5$,
   $\hat{\beta}_0(1) = \hat{\beta}_0(0) + \hat{\beta}_1(0) + (1 - \theta^2)e_1(1) = 51.60$,
   $\hat{\beta}_1(1) = \hat{\beta}_1(0) + (1 - \theta)^2 e_1(1) = 1.55$ and
   $\hat{y}_2(1) = \hat{\beta}_0(1) + \hat{\beta}_1(1) = 53.15$.

An Example DLS VI

3. The forecast for time period at time t=3, made at the end of the time period t=2, (the lead time $\tau = 1$) is computed below:
At time t=2, $e_2(1) = y_2 - \hat{y}_2(1) = 1.85$,
$\hat{\beta}_0(2) = \hat{\beta}_0(1) + \hat{\beta}_1(1) + (1 - \theta^2)e_2(1) = 53.50$,
$\hat{\beta}_1(2) = \hat{\beta}_1(1) + (1 - \theta)^2 e_2(1) = 1.57$ and
$\hat{y}_3(2) = \hat{\beta}_0(2) + \hat{\beta}_1(2) = 55.07$

Suppose a forecast at the time period t=12, (the lead time $\tau = 10$) is required at time t=2, the desired forecast is

$$\hat{y}_{12}(2) = \hat{\beta}_0(2) + \hat{\beta}_1(2) \cdot 10 = 69.2$$

Example:

Response variable:

- Annual sales of a product in a particular region of the country.

Predictor variables:

- Annual advertising expenditures
- Population in the region over the period of time

If the population size is not included in the model, this may cause the errors in the model to be **positively autocorrelated**, because if the per capita demand for the product is either constant or increasing with time, population size is positively correlated with product sales.

Autocorrelated error problems

Problem of autocorrelated errors:

1. The OLS regression coefficients are still unbiased, but they are no longer minimum-variance estimates. (GLS)
2. When the errors are positively autocorrelated, the *MSE* may seriously underestimate the error variance σ^2. Consequently, the standard errors of the $\beta_i's$ may be small. As a result, CIs and PIs are shorter than they really should be, and tests of hypotheses on $\beta_i's$ may be misleading, in that they may indicate that one or more predictor variables contribute significantly to the model when they really do not.
3. The CIs, PIs, and tests of hypotheses based on the t and F distributions are no longer exact procedures.

Approaches to deal with autocorrelated error

Approaches to deal with the problem of autocorrelated errors:

1. If possible, identify and include missing predictors in the model.
2. Use WLS or GLS if the autocorrelation structure is known.
3. If not, the analyst must turn to a model that specifically incorporates the autocorrelation structure.

Detecting Autocorrelation I

We introduce a commonly used test developed by Durbin and Watson. This test is based on the assumption that the errors are generated by a first-order autoregressive process, that is

$$\epsilon_t = \phi\epsilon_{t-1} + a_t$$

where ϵ_t is the error term at time t, a_t is an NID$(0, \sigma_a^2)$ random variable. We will require that $|\phi| < 1$. Thus the model with first-order autoregressive errors would be

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t, \quad \epsilon_t = \phi\epsilon_{t-1} + a_t$$

Successively substituting for ϵ_t, ϵ_{t-1},... yields

$$\epsilon_t = \sum_{i=0}^{\infty} \phi^i a_{t-i}$$

Detecting Autocorrelation II

that is, the error term for period t is a linear combination of all of the current and previous a_t. We can also show that

$$E(\epsilon_t) = 0$$

$$Var(\epsilon_t) = \sigma_a^2\left(\frac{1}{1-\phi^2}\right)$$

$$Cov(\epsilon_t, \epsilon_{t\pm k}) = \phi^k \sigma_a^2\left(\frac{1}{1-\phi^2}\right)$$

The lag one autocorrelation is

$$\rho_1 = \frac{Cov(\epsilon_t, \epsilon_{t+1})}{\sqrt{Var(\epsilon_t)}\sqrt{Var(\epsilon_t)}} = \phi$$

The autocorrelation between two errors that are k periods apart is

$$\rho_k = \phi^k$$

Durbin-Watson Test I

The test is a statistical test for the presence of *positive* autocorrelation in regression model errors.
The hypotheses are

$$H_0 : \phi = 0 \quad VS \quad H_1 : \phi > 0$$

and the Durbin-Watson test statistic is

$$\begin{aligned} d &= \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \\ &= \frac{\sum_{t=2}^{T} e_t^2 + \sum_{t=2}^{T} e_{t-1}^2 - 2\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_t^2} \\ &\approx 2(1 - r_1) \end{aligned}$$

where r_1 is the lag one error autocorrelation. So if errors are uncorrelated, the statistic should be closed to 2.

Durbin-Watson Test II

The decision procedure is as follows:

$$\text{If} \quad d < d_L \quad \text{reject } H_0$$

$$\text{If} \quad d > d_U \quad \text{do not reject } H_0$$

$$\text{If} \quad d_L \leq d \leq d_U \quad \text{inconclusive}$$

where d_L and d_U can be found in Appendix A.

If a test for negative autocorrelation is desired,

$$H_0 : \phi = 0 \quad VS \quad H_1 : \phi < 0$$

one can use statistic

$$4 - d.$$

Then the decision rules for the test hypotheses are the same as those used in testing for $\phi > 0$.

Durbin-Watson Test III

The data set below presents the **annual advertising expenses** and **annual concentrate sales** for a soft drink company for 20 years.

TABLE 3.13 Soft Drink Concentrate Sales Data

Year	Sales (Units)	Expenditures (10^3 dollars)	Residuals
1	3083	75	−32.3298
2	3149	78	−26.6027
3	3218	80	2.2154
4	3239	82	−16.9665
5	3295	84	−1.1484
6	3374	88	−2.5123
7	3475	93	−1.9671
8	3569	97	11.6691
9	3597	99	−0.5128
10	3725	104	27.0324
11	3794	109	−4.4224
12	3959	115	40.0318
13	4043	120	23.5770
14	4194	127	33.9403
15	4318	135	−2.7874
16	4493	144	−8.6060
17	4683	153	0.5753
18	4850	161	6.8476
19	5005	170	−18.9710
20	5236	182	−29.0625

A linear model is fitted

$$\hat{Sales} = 1609 + 20.1 Expenditures.$$

The residuals are plotted as Figure 3.5.The residual plot has a upward-downward pattern indicative of potential autocorrelation.

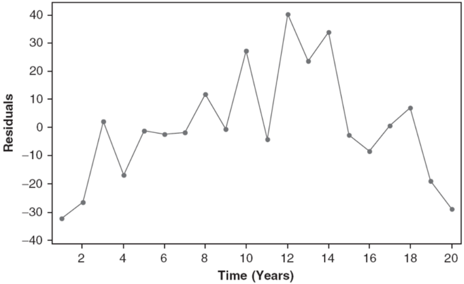


FIGURE 3.5 Plot of residuals versus time for the soft drink concentrate sales model.

Durbin-Watson Test V

Durbin-Watson test:

$$H_0 : \phi = 0 \quad VS \quad H_1 : \phi > 0,$$

the Durbin-Watson test statistic is

$$d = \frac{\sum_{t=2}^{20}(e_t - e_{t-1})^2}{\sum_{t=1}^{20} e_t^2} = 1.08$$

The critical values for significant level 0.05 are $d_L = 1.20$ and $d_U = 1.41$.

Sample Size	Probability in Lower Tail (Significance Level = α)	k = Number of Regressors (Excluding the Intercept)									
		1		2		3		4		5	
		d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
	0.01	0.81	1.07	0.70	1.25	0.59	1.46	0.49	1.70	0.39	1.96
15	0.025	0.95	1.23	0.83	1.40	0.71	1.61	0.59	1.84	0.48	2.09
	0.05	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
	0.01	0.95	1.15	0.86	1.27	0.77	1.41	0.63	1.57	0.60	1.74
20	0.025	1.08	1.28	0.99	1.41	0.89	1.55	0.79	1.70	0.70	1.87
	0.05	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99

Durbin-Watson Test VI

Since $d = 1.08 < d_L$, we reject the null hypothesis and conclude that the errors are positively autocorrelated.

If the Durbin-Watson test indicates a potential problem with autocorrelated model errors, and if these missing predictors can be identified and incorporated into the model, the autocorrelation problem may be eliminated. Reconsider the soft drink example:

TABLE 3.16 Expanded Soft Drink Concentrate Sales Data for Example 3.13

Year	Sales (Units)	Expenditures (10^3 Dollars)	Population	Residuals
1	3083	75	825,000	−4.8290
2	3149	78	830,445	−3.2721
3	3218	80	838,750	14.9179
4	3239	82	842,940	−7.9842
5	3295	84	846,315	5.4817
6	3374	88	852,240	0.7986
7	3475	93	860,760	−4.6749
8	3569	97	865,925	6.9178
9	3597	99	871,640	−11.5443
10	3725	104	877,745	14.0362
11	3794	109	886,520	−23.8654
12	3959	115	894,500	17.1334
13	4043	120	900,400	−0.9420
14	4194	127	904,005	14.9669
15	4318	135	908,525	−16.0945
16	4493	144	912,160	−13.1044
17	4683	153	917,630	1.8053
18	4850	161	922,220	13.6264
19	5005	170	925,910	−3.4759
20	5236	182	929,610	0.1025

The new fitted model is

$$\hat{Sales} = 320 + 18.4 Expenditures + 0.00168 Population.$$

We find the Durbin-Watson statistic for the new model is

$$d = 3.05932 > d_U = 1.54,$$

and conclude that there is no evidence to reject $\phi = 0$.

3.8 Regression Models for General Time Series Data III

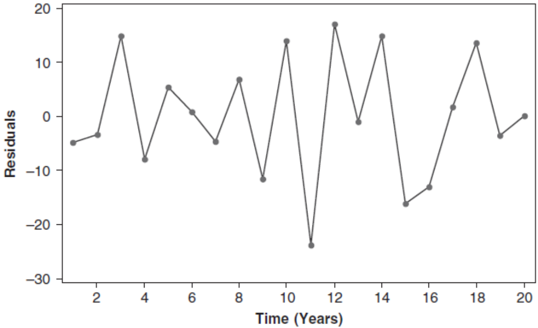


FIGURE 3.6 Plot of residuals versus time for the soft drink concentrate sales model in Example 3.13.

A plot of residual in time order shows improvement compared to the residuals of the old model. Therefore, we say that adding the new predictor population size has eliminated a problem with autocorrelation in the errors.

The Cochrane-Orcutt Method I

Often times, the autocorrelation can not be removed by adding more new predictors, then it is necessary to take explicit account of the autocorrelative structure.

The idea of Cochrane-Orcutt method for the simple linear trend case is the following: Contruct a new response variable $y_t' = y_t - \phi y_{t-1}$, and notice

$$
\begin{aligned}
y_t' =& y_t - \phi y_{t-1} && (94)\\
=& \beta_0 + \beta_1 x_t + \epsilon_t - \phi(\beta_0 + \beta_1 x_{t-1} + \epsilon_{t-1}) && (95)\\
=& \beta_0(1-\phi) + \beta_1(x_t - \phi x_{t-1}) + \epsilon_t - \phi\epsilon_{t-1} && (96)\\
=& \beta_0(1-\phi) + \beta_1(x_t - \phi x_{t-1}) + a_t && (97)\\
=& \beta_0' + \beta_1 x_t' + a_t && (98)
\end{aligned}
$$

Notice here we construct a new predictor $x_t' = x_t - \phi x_{t-1}$ and the error term $a_t = \epsilon_t - \phi\epsilon_{t-1}$ are independent.

The Cochrane-Orcutt Method II

An equivalent form of equation (94) is

$$y_t = \phi y_{t-1} + y_t' \tag{99}$$
$$= \phi y_{t-1} + \beta_0(1-\phi) + \beta_1(x_t - \phi x_{t-1}) + a_t \tag{100}$$

Equation (100) is obtained by equation (97).
The parameter ϕ is unknown and needs to be estimated. Recall $\epsilon_t = \phi\epsilon_{t-1} + a_t$, therefore $\hat{\phi}$ can be estimated by an OLS regression of e_t on e_{t-1}, that is

$$\hat{\phi} = \frac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_t^2}.$$

Then after $\hat{\phi}$ is obtained, we can calculate the new response and predictor

$$y_t' = y_t - \hat{\phi} y_{t-1}, \qquad x_t' = x_t - \hat{\phi} x_{t-1}$$

The Cochrane-Orcutt Method III

Then OLS to y'_t on x'_t yields the slope $\hat{\beta}'_0$, the intercept $\hat{\beta}_1$, and new residuals. If the new residuals are still autocorrelated, repeat the process until the residuals are uncorrelated. Usually one or two iterations are sufficient.

The Cochrane-Orcutt Method IV

The following data set contains the market share of a toothpaste and its price for 30 time periods.

TABLE 3.18 Toothpaste Market Share Data

Time	Market Share	Price	Residuals	y'_t	x'_t	Residuals
1	3.63	0.97	0.281193			
2	4.20	0.95	0.365398	2.715	0.553	−0.189435
3	3.33	0.99	0.466989	1.612	0.601	0.392201
4	4.54	0.91	−0.266193	3.178	0.505	−0.420108
5	2.89	0.98	−0.215909	1.033	0.608	−0.013381
6	4.87	0.90	−0.179091	3.688	0.499	−0.058753
7	4.90	0.89	−0.391989	2.908	0.522	−0.268949
8	5.29	0.86	−0.730682	3.286	0.496	−0.535075
9	6.18	0.85	−0.083580	4.016	0.498	0.244473
10	7.20	0.82	0.207727	4.672	0.472	0.256348
11	7.25	0.79	−0.470966	4.305	0.455	−0.531811
12	6.09	0.83	−0.659375	3.125	0.507	−0.423560
13	6.80	0.81	−0.435170	4.309	0.471	−0.131426
14	8.65	0.77	0.443239	5.869	0.439	0.635804
15	8.43	0.76	−0.019659	4.892	0.445	−0.192552
16	8.29	0.80	0.811932	4.842	0.489	0.847507
17	7.18	0.83	0.430625	3.789	0.503	0.141344
18	7.90	0.79	0.179034	4.963	0.451	0.027093
19	8.45	0.76	0.000341	5.219	0.437	−0.063744
20	8.23	0.78	0.266136	4.774	0.469	0.284026

A simple linear regression model is fitted and the Durbin-Watson statistic for the residuals is $d = 1.13582 < d_L = 1.20$, which indicates evidence to autocorrelated errors.

TABLE 3.19 Minitab Regression Results for the Toothpaste Market Share Data

Regression Analysis: Market Share Versus Price

```
The regression equation is
Market Share = 26.9 - 24.3 Price

Predictor      Coef  SE Coef      T      P
Constant     26.910    1.110  24.25  0.000
Price       -24.290    1.298 -18.72  0.000

S = 0.428710   R-Sq = 95.1%   R-Sq(adj) = 94.8%

Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  64.380  64.380  350.29  0.000
Residual Error  18   3.308   0.184
Total           19  67.688

Durbin-Watson statistic = 1.13582
```

The Cochrane-Orcutt Method VI

So we use Cochrane-Orcutt method to estimate $\hat{\phi} = 0.409$. The transformed variables are computed as:

$$y'_t = y_t - 0.409y_{t-1}$$

$$x'_t = x_t - 0.409x_{t-1}$$

which are listed in the table as well.

A simple regression of y'_t on x'_t is fitted and compared to the previous model.

The Cochrane-Orcutt Method VII

TABLE 3.20 Minitab Regression Results for Fitting the Transformed Model to the Toothpaste Sales Data

Regression Analysis: y′ Versus x′

```
The regression equation is
y-prime = 16.1 - 24.8 x-prime

Predictor      Coef   SE Coef       T      P
Constant    16.1090    0.9610   16.76  0.000
x-prime     -24.774     1.934  -12.81  0.000

S = 0.390963   R-Sq = 90.6%   R-Sq(adj) = 90.1%

Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  25.080  25.080  164.08  0.000
Residual Error  17   2.598   0.153
Total           18  27.679

Unusual Observations

Obs  x-prime  y-prime     Fit  SE Fit  Residual  St Resid
  2    0.601   1.6120  1.2198  0.2242    0.3922      1.22 X
  4    0.608   1.0330  1.0464  0.2367   -0.0134     -0.04 X
 15    0.489   4.8420  3.9945  0.0904    0.8475      2.23R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

Durbin-Watson statistic = 2.15671
```

3.8.2 Forecasting and Prediction Intervals I

It is tempting to ignore the autocorrelation to forecast the response at time, say T+1 by using

$$\hat{y}_{T+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{T+1} \tag{101}$$

assuming the predictor at time T+1 is known.

However, if we do consider the correlation, we would use

$$y_{T+1} = \phi y_T + (1-\phi)\beta_0 + \beta_1(x_{T+1} - \phi x_T) + a_{T+1}$$

which was derived in previous discuss to forecast $\hat{y}_{T+1}(T)$ as

$$\hat{y}_{T+1} = \hat{\phi} y_T + (1-\hat{\phi})\hat{\beta}_0 + \hat{\beta}_1(x_{T+1} - \hat{\phi} x_T) \tag{102}$$

(102) is likely to be very different from (101).

3.8.2 Forecasting and Prediction Intervals II

To find a prediction interval on the forecast, we notice that since the variance of the one-step-ahead forecast error is $Var(a_{T+1}) = \sigma_a^2$, a $100(1-a)$ percent prediction interval for the lead-one forecast is

$$\hat{y}_{T+1}(T) \pm z_{\alpha/2}\sigma_a,$$

which is estimated by

$$\hat{y}_{T+1}(T) \pm z_{\alpha/2}\hat{\sigma}_a.$$

Similarly, for a two-step-ahead forecast at time period T, we have

$$\begin{aligned}
y_{T+2} =& \phi y_{T+1} + (1-\phi)\beta_0 + \beta_1(x_{T+2} - \phi x_{T+1}) + a_{T+2} \\
=& \phi[\phi y_T + (1-\phi)\beta_0 + \beta_1(x_{T+1} - \phi x_T) + a_{T+1}] \\
&+ (1-\phi)\beta_0 + \beta_1(x_{T+2} - \phi x_{T+1}) + a_{T+2}
\end{aligned}$$

3.8.2 Forecasting and Prediction Intervals III

The two-step-ahead forecast error is

$$a_{T+2} + \phi a_{T+1}$$

whose variance is $Var(a_{T+2} + \phi a_{T+1}) = (1 + \phi^2)\sigma_a^2$.

So a $100(1 - \alpha)$ percent PI is

$$\hat{y}_{T+2}(T) \pm z_{\alpha/2}(1 + \phi^2)^{1/2}\hat{\sigma}_a.$$

In general, the τ-step-ahead forecast error is

$$y_{T+\tau} - \hat{y}_{T+\tau}(T) = a_{T+\tau} + \phi a_{T+\tau-1} + ... + \phi^{\tau-1} a_{T+1}$$

and a $100(1 - \alpha)$ percent PI for the lead-τ forecast is

$$\hat{y}_{T+\tau}(T) \pm z_{\alpha/2} \left(\frac{1 - \hat{\phi}^{2\tau}}{1 - \hat{\phi}^2} \right)^{1/2} \hat{\sigma}_a.$$

3.8.2 Forecasting and Prediction Intervals IV

Previously we assumed that the predictor(s), say x_{T+1} were known at the time period $T+1$, which is sometimes unrealistic. If not, we would have to use the forecast for the predictor(s), $\hat{x}_{T+1}$.

Now the forecast for y_{T+1} is

$$\hat{y}_{T+1}(T) = \hat{\phi} y_T + (1-\hat{\phi})\hat{\beta}_0 + \hat{\beta}_1[\hat{x}_{T+1}(T) - \hat{\phi} x_T]$$

so the forecast error is

$$y_{T+1} - \hat{y}_{T+1}(T) = a_{T+1} + \beta_1[x_{T+1} - \hat{x}_{T+1}(T)].$$

The variance of the error is

$$\sigma_a^2 + \beta_1^2 \sigma_x^2(1),$$

so a $100(1-\alpha)$ percent PI for the lead-one-forecast is

$$\hat{y}_{T+1}(T) \pm z_{\alpha/2}[\hat{\sigma}_a^2 + \hat{\beta}_1^2\hat{\sigma}_x^2(1)]^{1/2}$$

Recall the regression model with autocorrelated errors:

$$\begin{aligned} y_t =& \phi y_{t-1} + \beta_0(1-\phi) + \beta_1(x_t - \phi x_{t-1}) + a_t \\ =& \phi y_{t-1} + \beta_0(1-\phi) + \beta_1 x_t - \phi\beta_1 x_{t-1} + a_t \end{aligned}$$

which has a general format

$$y_t = \phi y_{t-1} + \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + a_t \tag{103}$$

Another alternative model is to drop the lagged value of predictor variable x_{t-1} to get

$$y_t = \phi y_{t-1} + \beta_0 + \beta_1 x_t + a_t \tag{104}$$

An Example I

Reconsider the toothpaste example, first we fit the model (3).

TABLE 3.23 Minitab Results for Fitting Model (3.119) to the Toothpaste Market Share Data

Regression Analysis: y Versus y_{t-1}, x, x_{t-1}

```
The regression equation is
y = 16.1 + 0.425 y(t-1) - 22.2 x + 7.56 x(t-1)


Predictor      Coef  SE Coef      T      P
Constant     16.100    6.095   2.64  0.019
y(t-1)       0.4253   0.2239   1.90  0.077
x           -22.250    2.488  -8.94  0.000
x(t-1)        7.562    5.872   1.29  0.217

S = 0.402205   R-Sq = 96.0%   R-Sq(adj) = 95.2%

Analysis of Variance

Source          DF      SS      MS       F      P
Regression       3  58.225  19.408  119.97  0.000
Residual Error  15   2.427   0.162
Total           18  60.651

Source  DF  Seq SS
y(t-1)   1  44.768
x        1  13.188
x(t-1)   1   0.268

Durbin-Watson statistic = 2.04203
```

An Example II

Notice the P-value for x_{t-1} is 0.217 indicating this variable is not significant in the model. So we remove it and refit the model (4)

TABLE 3.24 Minitab Results for Fitting Model (3.120) to the Toothpaste Market Share Data

Regression Analysis: y Versus y_{t-1}, x

```
The regression equation is
y = 23.3 + 0.162 y(t-1) - 21.2 x

Predictor      Coef  SE Coef      T      P
Constant     23.279    2.515   9.26  0.000
y(t-1)      0.16172  0.09238   1.75  0.099
x           -21.181    2.394  -8.85  0.000

S = 0.410394   R-Sq = 95.6%   R-Sq(adj) = 95.0%

Analysis of Variance

Source          DF      SS      MS       F      P
Regression       2  57.956  28.978  172.06  0.000
Residual Error  16   2.695   0.168
Total           18  60.651

Source  DF  Seq SS
y(t-1)   1  44.768
x        1  13.188

Durbin-Watson statistic = 1.61416
```