Chapter 16
Logistic Regression

Chapter Table of Contents

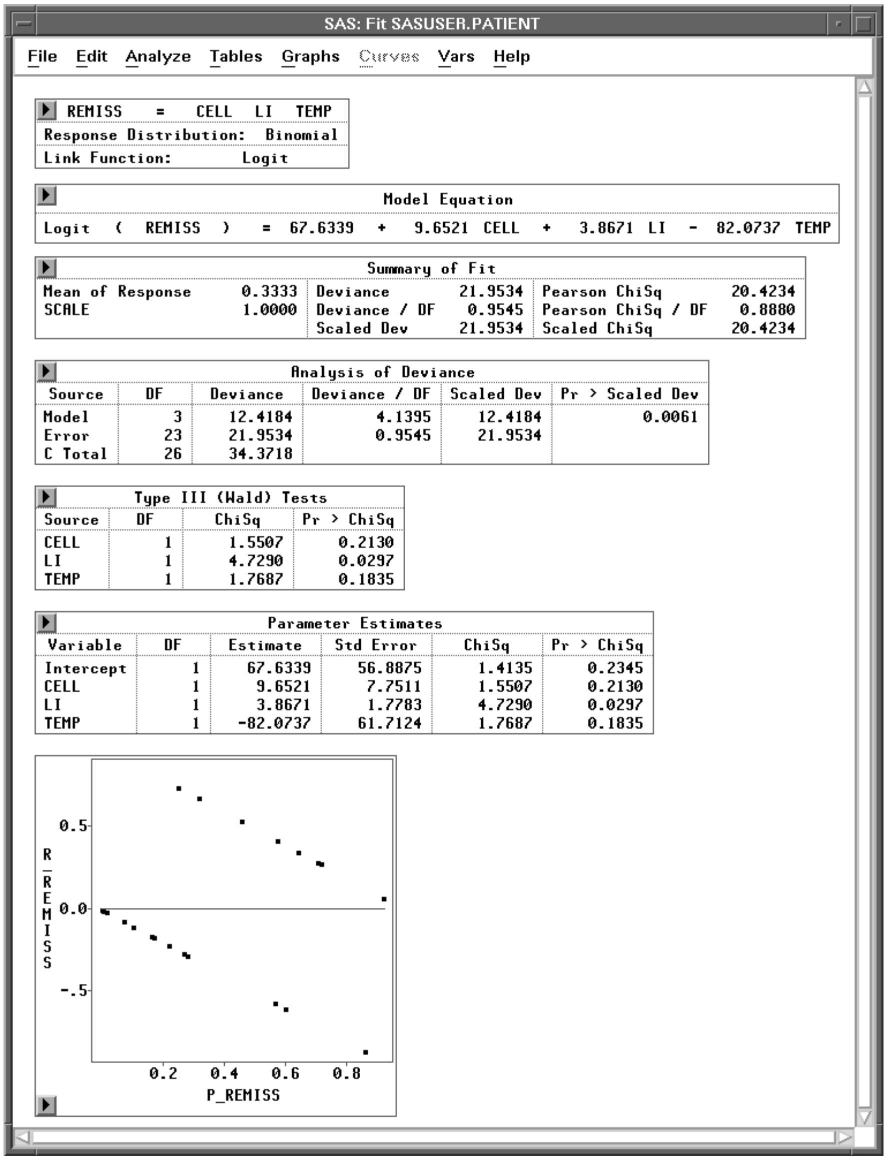
DISPLAYING THE LOGISTIC REGRESSION ANALYSIS 252
Model Equation 255
Summary of Fit 256
Analysis of Deviance 256
Type III (Wald) Tests 256
Parameter Estimates Table 256
Residuals-by-Predicted Plot 256

MODIFYING THE MODEL 257

REFERENCES 260

Chapter 16
Logistic Regression

In the last two chapters, you used least-squares methods to fit linear models. In this chapter, you use maximum-likelihood methods to fit *generalized* linear models. You can choose **Analyze:Fit (Y X)** to carry out a logistic regression analysis. You can use the fit variables and method dialogs to specify generalized linear models and to add and delete variables from the model.



```
REMISS    =    CELL   LI   TEMP
Response Distribution:  Binomial
Link Function:          Logit
```

Model Equation

```
Logit  (  REMISS  )  =  67.6339  +  9.6521 CELL  +  3.8671 LI  -  82.0737 TEMP
```

Summary of Fit

Mean of Response	0.3333	Deviance	21.9534	Pearson ChiSq	20.4234
SCALE	1.0000	Deviance / DF	0.9545	Pearson ChiSq / DF	0.8880
		Scaled Dev	21.9534	Scaled ChiSq	20.4234

Analysis of Deviance

Source	DF	Deviance	Deviance / DF	Scaled Dev	Pr > Scaled Dev
Model	3	12.4184	4.1395	12.4184	0.0061
Error	23	21.9534	0.9545	21.9534	
C Total	26	34.3718			

Type III (Wald) Tests

Source	DF	ChiSq	Pr > ChiSq
CELL	1	1.5507	0.2130
LI	1	4.7290	0.0297
TEMP	1	1.7687	0.1835

Parameter Estimates

Variable	DF	Estimate	Std Error	ChiSq	Pr > ChiSq
Intercept	1	67.6339	56.8875	1.4135	0.2345
CELL	1	9.6521	7.7511	1.5507	0.2130
LI	1	3.8671	1.7783	4.7290	0.0297
TEMP	1	-82.0737	61.7124	1.7687	0.1835

Figure 16.1. Logistic Regression Analysis

Displaying the Logistic Regression Analysis

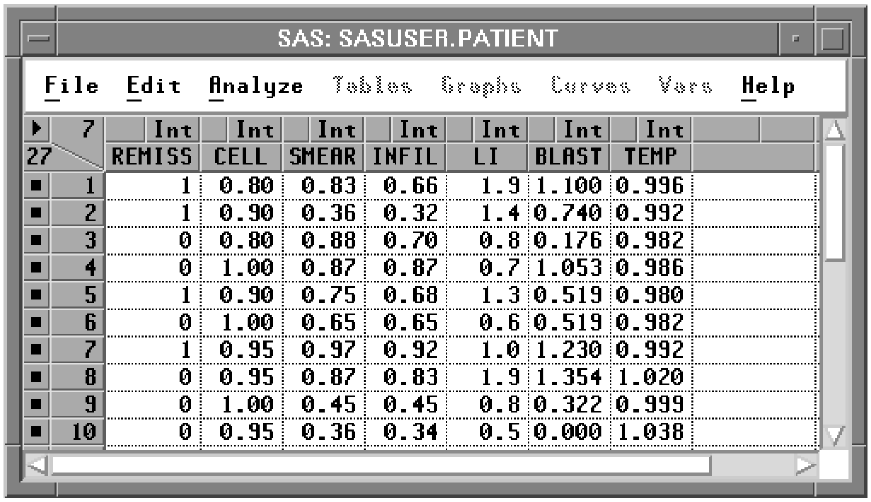
The **PATIENT** data set, described by Lee (1974), contains data collected on 27 cancer patients. The response variable, **REMISS**, is binary and indicates whether cancer remission occurred:

REMISS = 1 indicates success (remission occurred)

REMISS = 0 indicates failure (remission did not occur)

Several other variables containing patient characteristics thought to affect cancer remission were also included in the study. For this example, consider the following three explanatory variables: **CELL**, **LI**, and **TEMP**. (You may want to carry out a more complete analysis on your own.)

⟹ **Open the PATIENT data set.**



7	Int	Int	Int	Int	Int	Int	Int
27	REMISS	CELL	SMEAR	INFIL	LI	BLAST	TEMP
1	1	0.80	0.83	0.66	1.9	1.100	0.996
2	1	0.90	0.36	0.32	1.4	0.740	0.992
3	0	0.80	0.88	0.70	0.8	0.176	0.982
4	0	1.00	0.87	0.87	0.7	1.053	0.986
5	1	0.90	0.75	0.68	1.3	0.519	0.980
6	0	1.00	0.65	0.65	0.6	0.519	0.982
7	1	0.95	0.97	0.92	1.0	1.230	0.992
8	0	0.95	0.87	0.83	1.9	1.354	1.020
9	0	1.00	0.45	0.45	0.8	0.322	0.999
10	0	0.95	0.36	0.34	0.5	0.000	1.038

Figure 16.2. Data Window

The generalized linear model has three components:

- a linear predictor function constructed from explanatory variables. For this example, the function is

  $$\theta_i = \beta_0 + \beta_1 \mathsf{CELL}_i + \beta_2 \mathsf{LI}_i + \beta_3 \mathsf{TEMP}_i$$

  where $\beta_0, \beta_1, \beta_2$ and β_3 are coefficients (parameters) for the linear predictor, and CELL_i, LI_i, and TEMP_i are the values of the explanatory variables.

- a distribution or probability function for the response variable that depends on the mean μ and sometimes other parameters as well. For this example, the probability function is binomial.

- a link function, $g(.)$, that relates the mean to the linear predictor function. For logistic regression, the link function is the logit

  $$g(p_i) = \mathrm{logit}(\mathrm{p_i}) = \log(\frac{\mathrm{p_i}}{1-\mathrm{p_i}}) = \theta_\mathrm{i}$$

  where p_i = Pr(REMISS=1 | x_i) is the response probability to be modeled, and x_i is the set of explanatory variables for the *i*th patient.

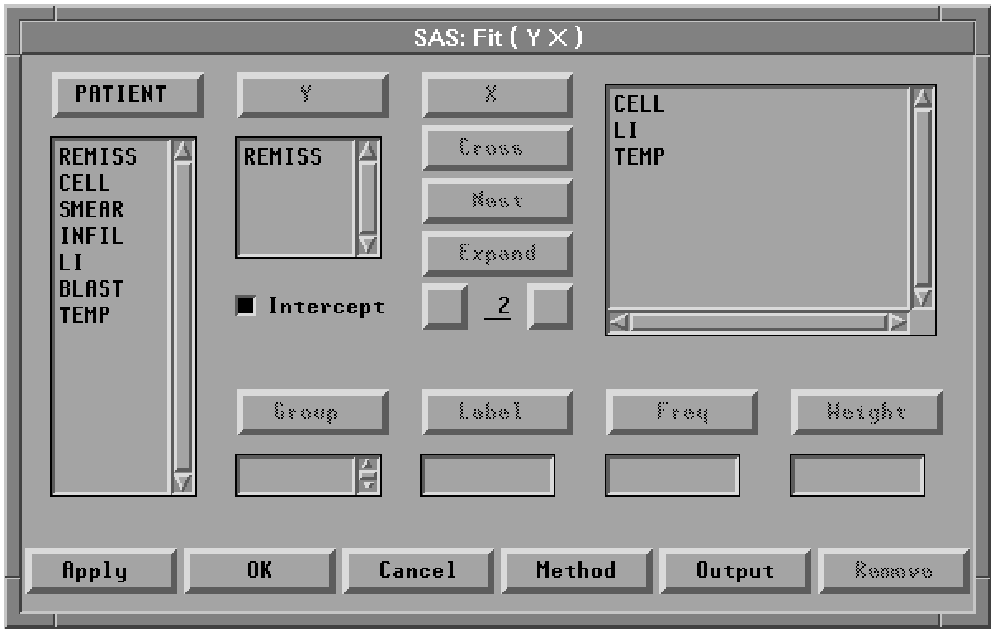
You can specify these three components to fit a generalized linear model by following these steps.

⟹ **Choose Analyze:Fit (Y X) to display the fit variables dialog.**

⟹ **Select REMISS in the list at the left, then click the Y button.**

⟹ **Select CELL, LI, and TEMP in the variables list, then click the X button.**

Your variables dialog should now appear, as shown in Figure 16.3.



Y: REMISS

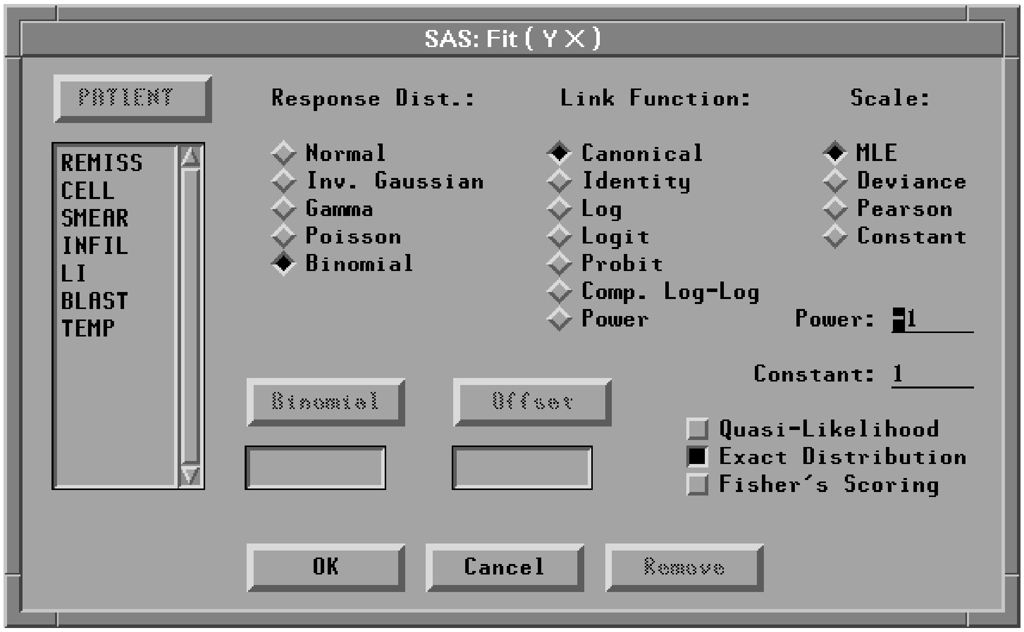
■ Intercept

X: CELL LI TEMP

Figure 16.3. Fit Variables Dialog with Variable Roles Assigned

To specify the probability distribution for the response variable and the link function, follow these steps.

⟹ **Click the Method button in the variables dialog to display the method dialog.**



REMISS
CELL
SMEAR
INFIL
LI
BLAST
TEMP

Response Dist.:
○ Normal
○ Inv. Gaussian
○ Gamma
○ Poisson
● Binomial

Link Function:
● Canonical
○ Identity
○ Log
○ Logit
○ Probit
○ Comp. Log-Log
○ Power

Scale:
● MLE
○ Deviance
○ Pearson
○ Constant

Power: 1

Constant: 1

□ Quasi-Likelihood
■ Exact Distribution
□ Fisher´s Scoring

Figure 16.4. Fit Method Dialog

⟹ **Click on Binomial under Response Dist to specify the probability distribution.**

You do not need to specify a **Link Function** for this example. **Canonical**, the default, allows **Fit (Y X)** to choose a link dependent on the probability distribution. For the binomial distribution, as in this example, it is equivalent to choosing **Logit**, which yields a logistic regression.

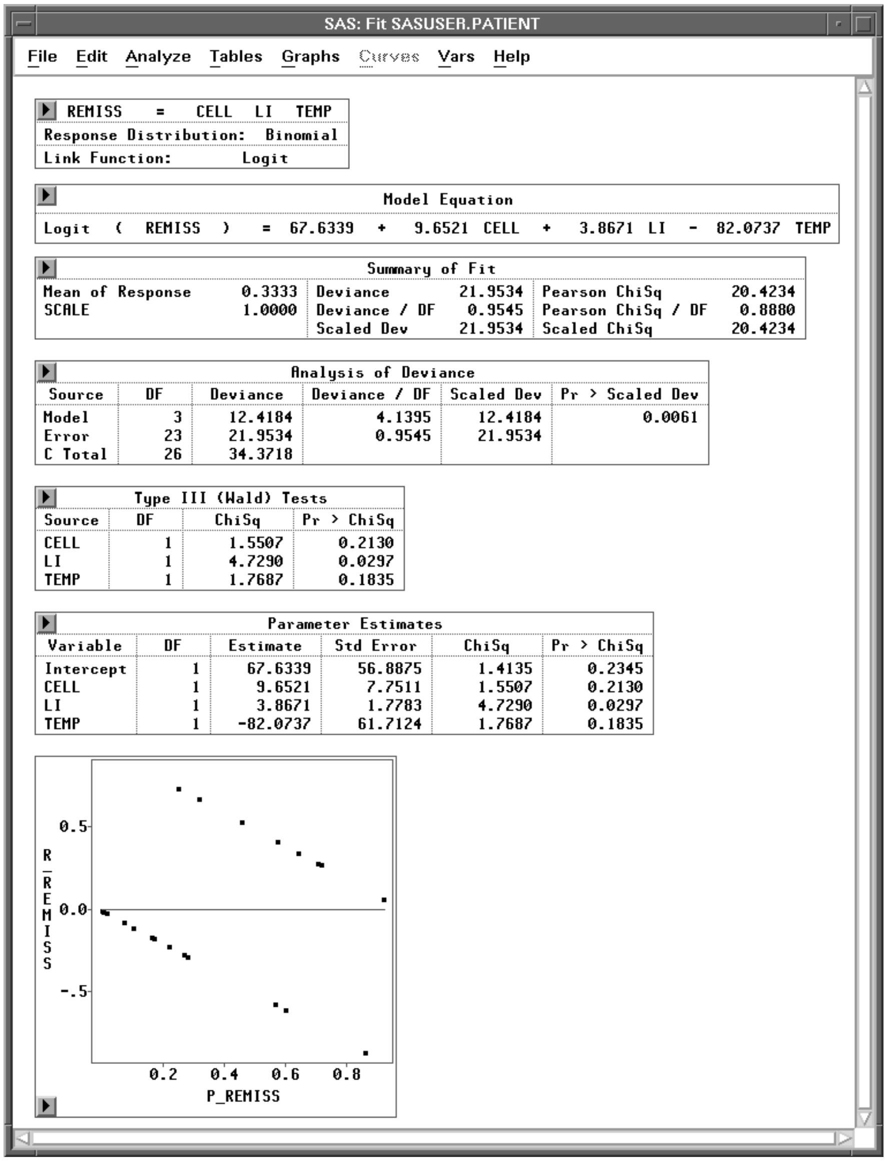
⟹ **Click the OK button to close the method dialog.**

⟹ **Click the Apply button in the variables dialog.**

This creates the analysis shown in Figure 16.5. Recall that the **Apply** button causes the variables dialog to stay on the screen after the fit window appears. This is convenient for adding and deleting variables from the model.

By default, the fit window displays tables for model information, **Model Equation**, **Summary of Fit**, **Analysis of Deviance**, **Type III (Wald) Tests**, and **Parameter Estimates**, and a residual-by-predicted plot. You can control the tables and graphs displayed by clicking on the **Output** button in the fit variables dialog or by choosing from the **Tables** and **Graphs** menus.

The first table displays the model information. The first line gives the model specification. The second and third lines give the error distribution and the link function you specified in the Method dialog.



```
REMISS    =    CELL   LI   TEMP
Response Distribution:  Binomial
Link Function:        Logit
```

```
                         Model Equation
Logit  (  REMISS  )  =  67.6339  +  9.6521  CELL  +  3.8671  LI  -  82.0737  TEMP
```

```
                                  Summary of Fit
Mean of Response      0.3333  Deviance          21.9534  Pearson ChiSq        20.4234
SCALE                 1.0000  Deviance / DF      0.9545  Pearson ChiSq / DF    0.8880
                              Scaled Dev        21.9534  Scaled ChiSq         20.4234
```

```
                         Analysis of Deviance
 Source    DF    Deviance   Deviance / DF  Scaled Dev  Pr > Scaled Dev
Model       3     12.4184          4.1395     12.4184           0.0061
Error      23     21.9534          0.9545     21.9534
C Total    26     34.3718
```

```
       Type III (Wald) Tests
Source    DF     ChiSq    Pr > ChiSq
CELL       1    1.5507       0.2130
LI         1    4.7290       0.0297
TEMP       1    1.7687       0.1835
```

```
                       Parameter Estimates
 Variable    DF    Estimate   Std Error     ChiSq    Pr > ChiSq
Intercept     1     67.6339     56.8875    1.4135       0.2345
CELL          1      9.6521      7.7511    1.5507       0.2130
LI            1      3.8671      1.7783    4.7290       0.0297
TEMP          1    -82.0737     61.7124    1.7687       0.1835
```

Figure 16.5. Fit Window

Model Equation

The **Model Equation** table writes out the fitted model using the estimated regression coefficients:

$$\begin{aligned} &\text{logit}(\text{Prob}(\textsf{REMISS} = 1)) \\ = \;& \textsf{67.6399} + \textsf{9.6521*CELL} + \textsf{3.8671*LI} - \textsf{82.0737*TEMP} \end{aligned}$$

Summary of Fit

The **Summary of Fit** table contains summary statistics for the fit of the model including values for **Deviance** and **Pearson's Chi-Squared** statistics. These values contrast the fit of your model to that of a saturated model that allows a different fit for each observation. If the data are sparse in the sense that most observations have a different set of explanatory variables, as in this set of data, then the quality of these measures is likely to be poor. Inferences drawn from these measures should be treated cautiously.

Analysis of Deviance

The **Analysis of Deviance** table summarizes information about the variation in the response for the set of data. Some of the variation can be explained by the **Model**. The **Error** is the remainder that is not systematically explained. **C Total** (the total corrected or adjusted for the mean) is the sum of **Model** and **Error**. The probability values give a measure of whether the amount of variation is consistent with chance alone or whether there is evidence of additional variation. In this case the **Deviance** associated with the **Model** shows a significant effect for the model, (p = 0.0061).

Type III (Wald) Tests

Wald tests are Chi-square statistics that test the null hypothesis that a parameter is 0; in other words, that the corresponding variable has no effect given that the other variables are in the model. These are approximate tests that are more accurate with larger sample sizes. In this example, only the coefficient for **LI** is statistically significant (p = 0.0297).

Parameter Estimates Table

The **Parameter Estimates** table shows the estimate, standard error, Chi-square statistic and associated degrees of freedom, and p-value for each of the parameters estimated.

Residuals-by-Predicted Plot

In the diagnostic plot of residuals versus predicted values, you can examine residuals for the model. You can point and click to identify individual observations. Because the observed response must either be 0 or 1, the plot of the residuals versus predicted values must lie along two straight lines. Plots of residuals versus the independent variables and other possible explanatory variables may be more useful. You can create scatter plots by selecting the response and explanatory variables in the data window and choosing **Analyze:Scatter Plot (Y X)**.

Modifying the Model

Plots of the residuals against other variables may suggest extensions of the model. Alternatively you may be able to remove some variables and thus simplify the model without losing explanatory power. The **Type III (Wald) Tests** table or the possibly more accurate **Type III (LR) Tests** table contains statistics that can help you decide whether to remove an effect. If the *p*-value associated with the test is large, then there is little evidence for any explanatory value of the corresponding variable.

⟹ **Choose Tables:Type III (LR) Tests.**

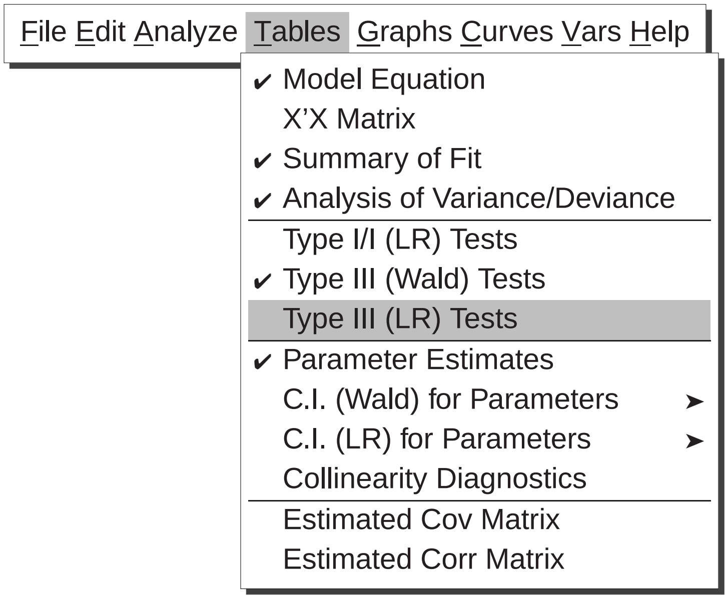
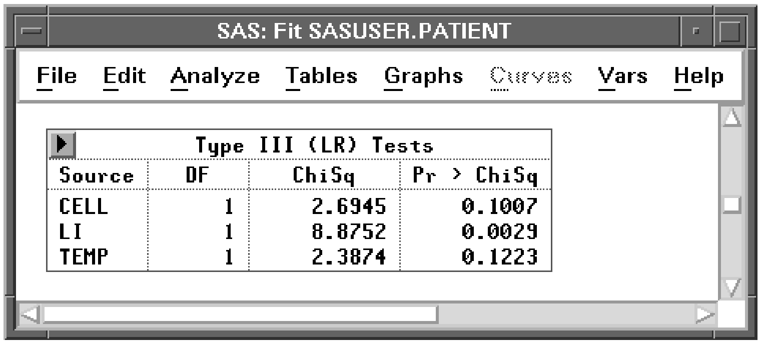


Figure 16.6. Tables Menu

This displays the table shown in Figure 16.7.



Type III (LR) Tests

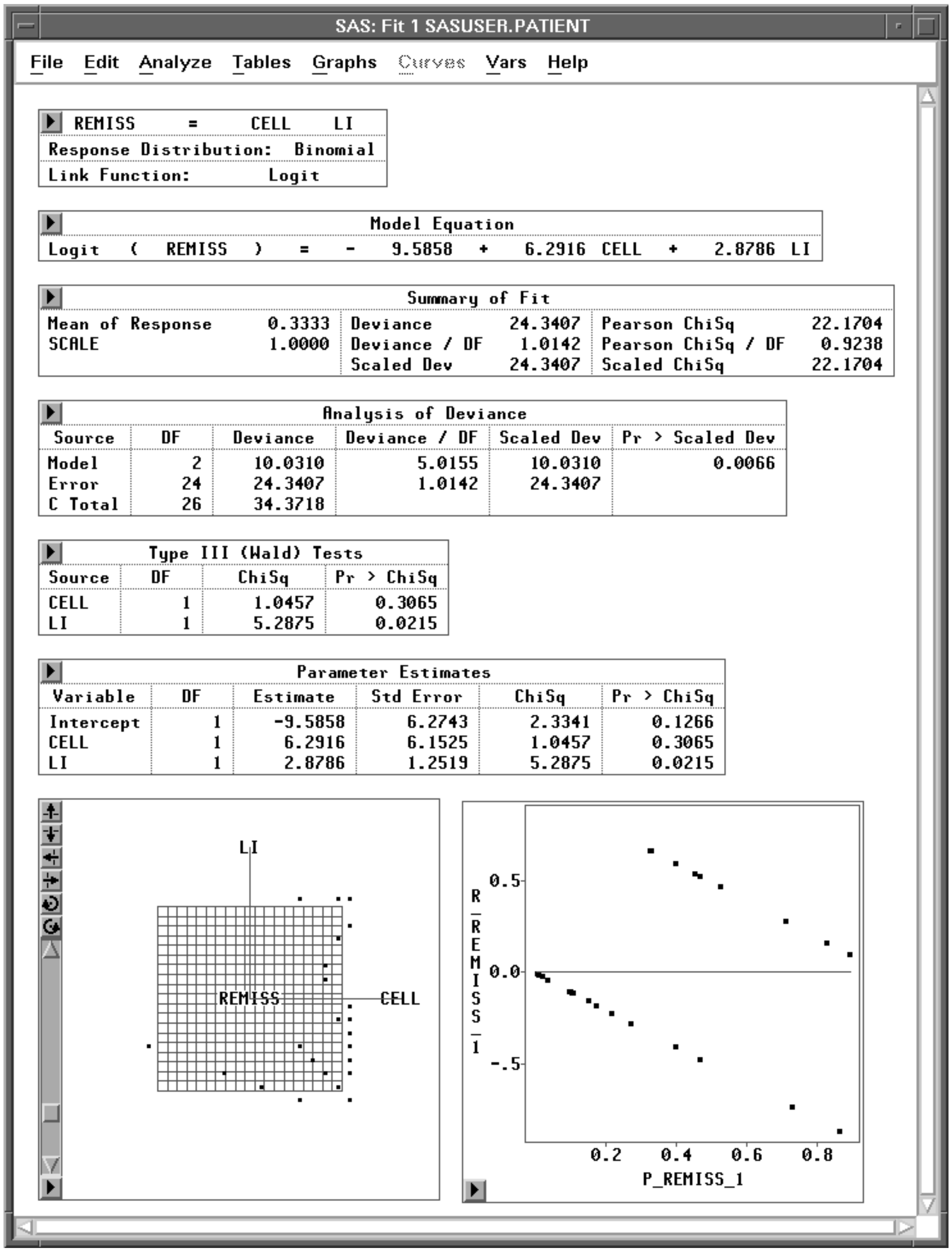
Source	DF	ChiSq	Pr > ChiSq
CELL	1	2.6945	0.1007
LI	1	8.8752	0.0029
TEMP	1	2.3874	0.1223

Figure 16.7. Likelihood Ratio Type III Tests

The *p*-values for **TEMP** and **CELL** are relatively large, suggesting these effects could be removed. Although the numbers are different, the same conclusions would be reached from the corresponding **Wald** tests. In the Fit Variables dialog, follow these steps to request a new model with **TEMP** removed.

⟹ **Select TEMP in the effects list, then click the Remove button.**
TEMP disappears from the effects list.

⟹ **Click on Apply, and a new fit window appears, as shown in Figure 16.8.**



```
REMISS      =      CELL     LI
Response Distribution:  Binomial
Link Function:        Logit
```

```
                         Model Equation
Logit   (   REMISS   )   =   -   9.5858   +   6.2916  CELL   +   2.8786  LI
```

```
                              Summary of Fit
Mean of Response     0.3333  Deviance        24.3407  Pearson ChiSq        22.1704
SCALE                1.0000  Deviance / DF    1.0142  Pearson ChiSq / DF    0.9238
                             Scaled Dev      24.3407  Scaled ChiSq         22.1704
```

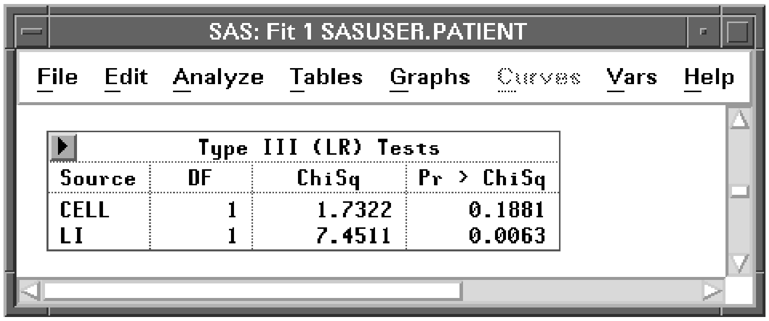
```
                          Analysis of Deviance
 Source     DF   Deviance   Deviance / DF   Scaled Dev   Pr > Scaled Dev
Model        2   10.0310          5.0155     10.0310            0.0066
Error       24   24.3407          1.0142     24.3407
C Total     26   34.3718
```

```
         Type III (Wald) Tests
Source     DF      ChiSq    Pr > ChiSq
CELL        1     1.0457      0.3065
LI          1     5.2875      0.0215
```

```
                         Parameter Estimates
 Variable    DF    Estimate   Std Error     ChiSq    Pr > ChiSq
 Intercept    1     -9.5858      6.2743    2.3341        0.1266
 CELL         1      6.2916      6.1525    1.0457        0.3065
 LI           1      2.8786      1.2519    5.2875        0.0215
```

Figure 16.8. Fit Window with CELL and LI as Explanatory Variables

⟹ **Choose Tables:Type III (LR) Tests in the new fit window.**
This displays a **Type III (LR) Tests** table in the window.



Type III (LR) Tests

Source	DF	ChiSq	Pr > ChiSq
CELL	1	1.7322	0.1881
LI	1	7.4511	0.0063

Figure 16.9. Likelihood Ratio Type III Tests

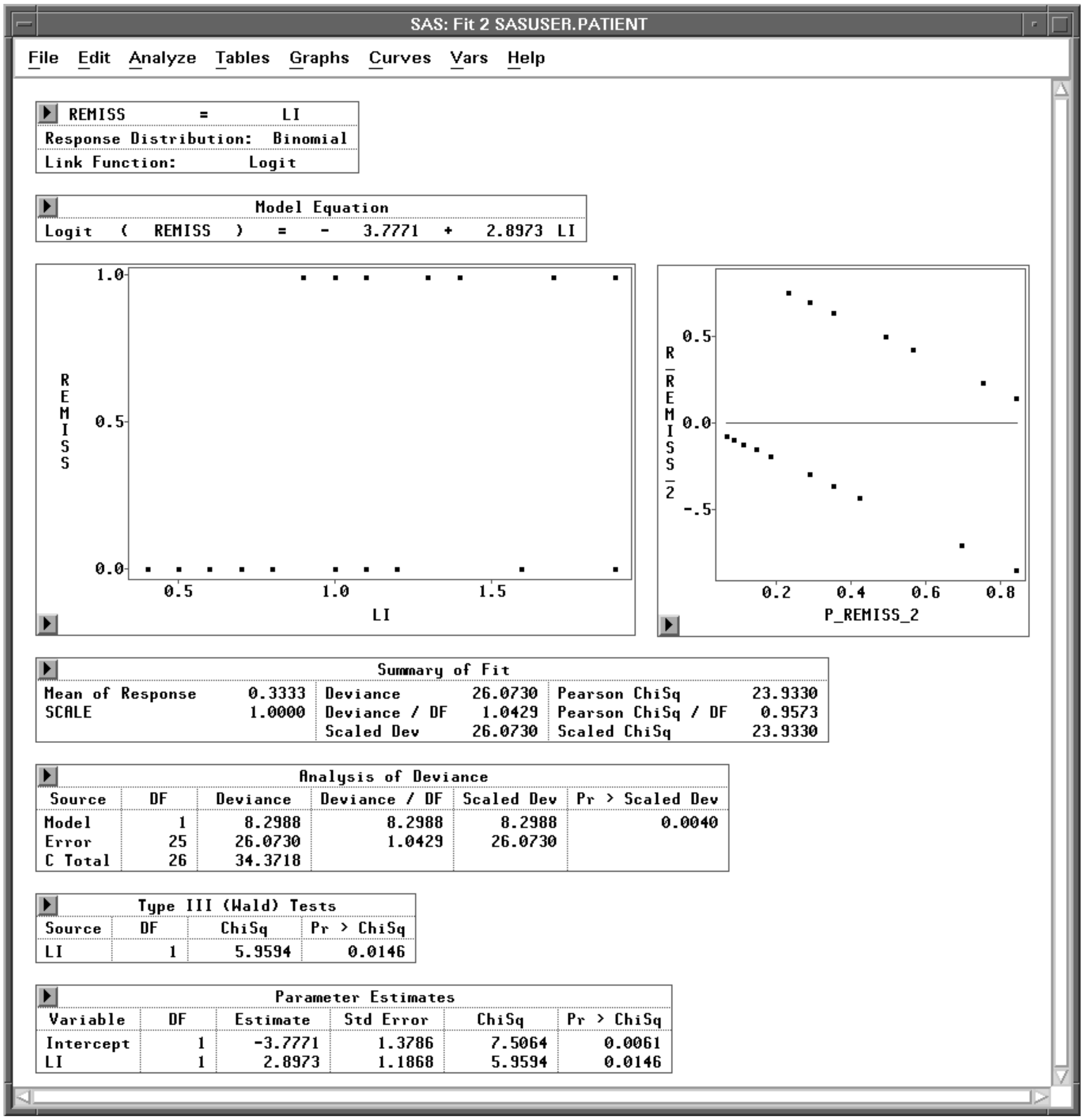
The *p*-value for **CELL** in the LR test suggests that this effect could also be removed.

⟹ **Click on the variable CELL in the effects list in the Fit Dialog.**
Then click on **Remove**. **CELL** disappears from the effects list.

⟹ **Click on Apply, and a new Fit window appears, as shown in Figure 16.10.**
Since the new model contains only one **X** variable, the fit window displays a plot of **REMISS** versus **CELL**.

Using the **Apply** button, you have quickly created three logistic regression models. Logistic regression is only one special case of the generalized linear model. Another case, Poisson regression, is described in the next chapter.

⊕ **Related Reading:** Generalized Linear Models, Chapter 39.



REMISS	=	LI
Response Distribution:		Binomial
Link Function:		Logit

Model Equation

Logit (REMISS) = - 3.7771 + 2.8973 LI

Summary of Fit

Mean of Response	0.3333	Deviance	26.0730	Pearson ChiSq	23.9330
SCALE	1.0000	Deviance / DF	1.0429	Pearson ChiSq / DF	0.9573
		Scaled Dev	26.0730	Scaled ChiSq	23.9330

Analysis of Deviance

Source	DF	Deviance	Deviance / DF	Scaled Dev	Pr > Scaled Dev
Model	1	8.2988	8.2988	8.2988	0.0040
Error	25	26.0730	1.0429	26.0730	
C Total	26	34.3718			

Type III (Wald) Tests

Source	DF	ChiSq	Pr > ChiSq
LI	1	5.9594	0.0146

Parameter Estimates

Variable	DF	Estimate	Std Error	ChiSq	Pr > ChiSq
Intercept	1	-3.7771	1.3786	7.5064	0.0061
LI	1	2.8973	1.1868	5.9594	0.0146

Figure 16.10. Fit Window with LI as the Only Explanatory Variable

References

Lee, E.T. (1974), "A Computer Program for Linear Logistic Regression Analysis," *Computer Programs in Biomedicine*, 80–92.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, Second Edition, London: Chapman and Hall.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/INSIGHT User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999. 752 pp.

SAS/INSIGHT User's Guide, Version 8

ISBN 1–58025–490–X

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999

The Institute is a private company devoted to the support and further development of its software and related services.