Chapter 17
Poisson Regression

Chapter Table of Contents

Chapter 17
Poisson Regression

In Chapter 16, “Logistic Regression,” you examined logistic regression as an example of a generalized linear model.

In this chapter, you will examine another example of a generalized linear model, Poisson regression. You can choose **Analyze:Fit (Y X)** to carry out a Poisson regression analysis when the response variable represents counts. You can use the fit variables and methods dialogs to specify this generalized linear model.

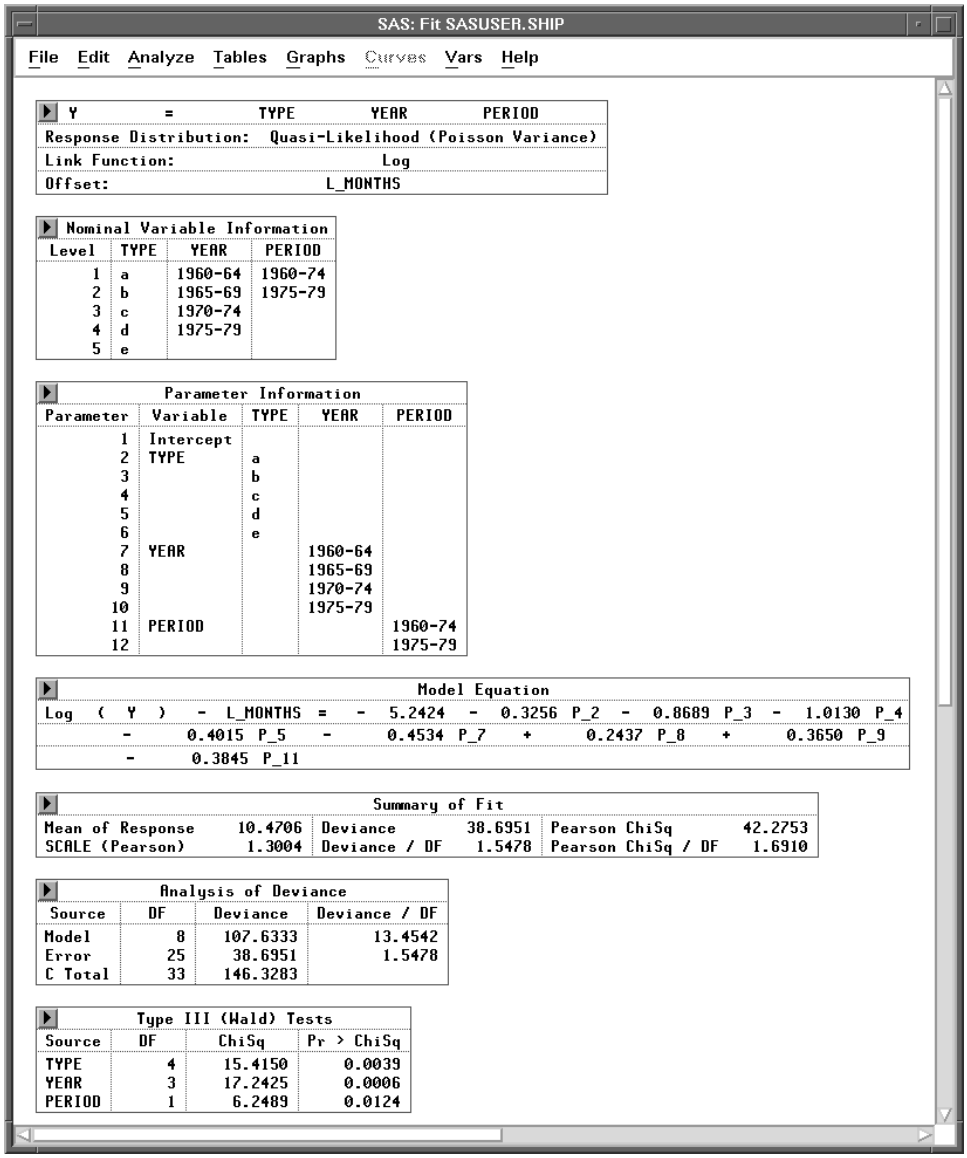
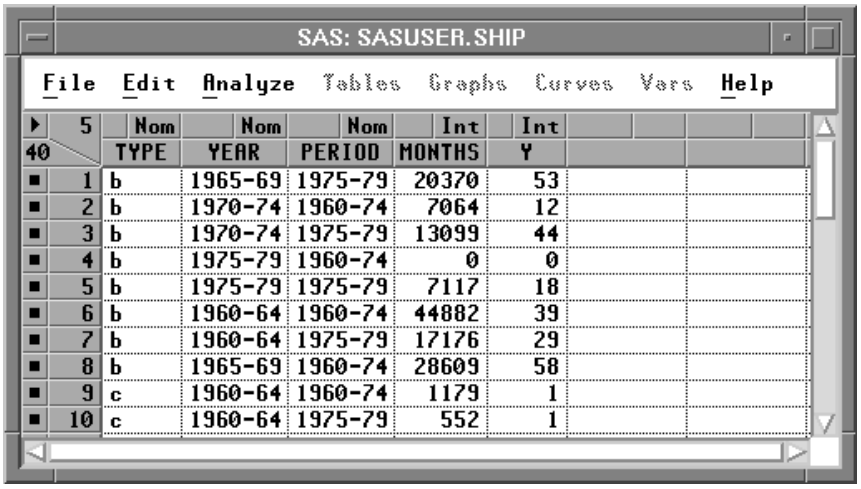


Figure 17.1. Poisson Regression Analysis

Displaying the Poisson Regression Analysis

The **SHIP** data shown in Figure 17.2 represent damage caused by waves to the forward section of certain cargo-carrying vessels. The purpose of the investigation was to set standards for future hull construction. In order to do so, the investigators needed to know the risk of damage associated with five ship types (**TYPE**), year of construction (**YEAR**), and period of operation (**PERIOD**). These three variables are the classification variables. **MONTHS** is the aggregate number of months in service and is an explanatory variable. **Y** is the response variable and represents the number of damage incidents (McCullagh and Nelder 1989).



40	TYPE	YEAR	PERIOD	MONTHS	Y
	Nom	Nom	Nom	Int	Int
1	b	1965-69	1975-79	20370	53
2	b	1970-74	1960-74	7064	12
3	b	1970-74	1975-79	13099	44
4	b	1975-79	1960-74	0	0
5	b	1975-79	1975-79	7117	18
6	b	1960-64	1960-74	44882	39
7	b	1960-64	1975-79	17176	29
8	b	1965-69	1960-74	28609	58
9	c	1960-64	1960-74	1179	1
10	c	1960-64	1975-79	552	1

Figure 17.2. **SHIP** Data Set

Recall from Chapter 16 that the generalized linear model has three basic components:

- a linear function of explanatory variables. For this example, the function is

$$\beta_0 + \beta_1 \log(\text{MONTHS}) + \gamma_i + \tau_j + \delta_k + (\gamma\tau)_{ij} + (\gamma\delta)_{ik} + (\tau\delta)_{jk}$$

  where $log(\text{MONTHS})$ is a variable whose coefficient β_1 is believed to be 1. An effect such as this is commonly referred to as an *offset*. γ_i is the effect of the *i*th level of **TYPE**, τ_j is the effect of the *j*th level of **YEAR**, δ_k is the effect of the *k*th level of **PERIOD**, $(\gamma\tau)_{ij}$ is the effect of the *ij*th level of the **TYPE** by **YEAR** interaction, $(\gamma\delta)_{ik}$ is the effect of the *ik*th level of the **TYPE** by **PERIOD** interaction, and $(\tau\delta)_{jk}$ is the effect of the *jk*th level of the **YEAR** by **PERIOD** interaction.

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

- a link function that relates the mean to the linear function of explanatory variables. For this example, the link function is the log

$$\log(\text{expected number of damage incidents})$$

$$= \quad \beta_0 + \beta_1 \log(\text{MONTHS}) + \gamma_i + \tau_j + \delta_k + (\gamma\tau)_{ij} + (\gamma\delta)_{ik} + (\tau\delta)_{jk}$$

$\Longrightarrow$ **Open the SHIP data set.**

Recall from the previous equation that **Y** is assumed to be directly proportional to **MONTHS**. Since $log(Y)$ is being modeled, you need to carry out a log transformation on **MONTHS**. Follow these steps to create a new variable that represents the log of **MONTHS**.

$\Longrightarrow$ **Select MONTHS in the data window.**

$\Longrightarrow$ **Choose Edit:Variables:log(Y).**

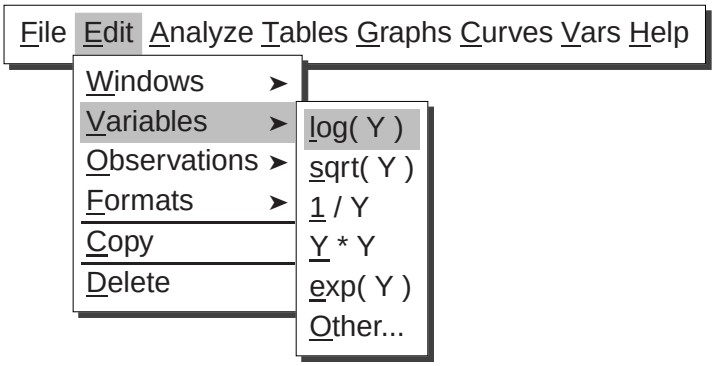
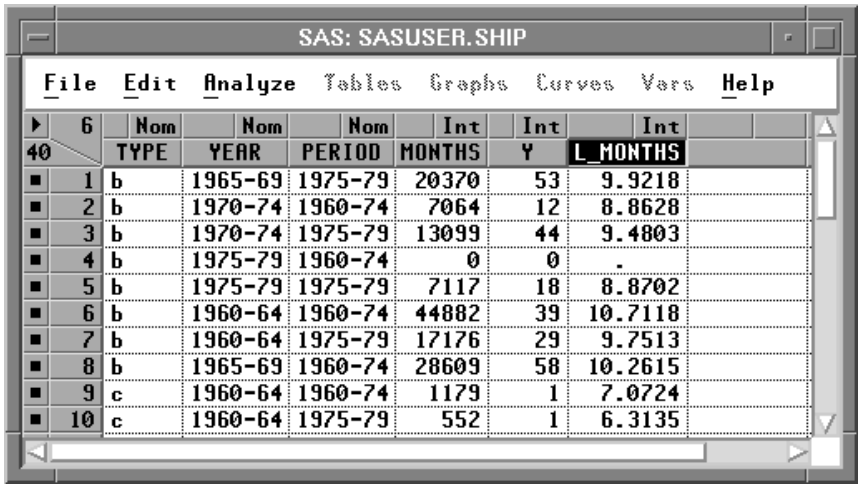


Figure 17.3. Edit:Variables Menu

A new variable, **L_MONTHS**, now appears in the data window.



40	TYPE	YEAR	PERIOD	MONTHS	Y	L_MONTHS
1	b	1965-69	1975-79	20370	53	9.9218
2	b	1970-74	1960-74	7064	12	8.8628
3	b	1970-74	1975-79	13099	44	9.4803
4	b	1975-79	1960-74	0	0	.
5	b	1975-79	1975-79	7117	18	8.8702
6	b	1960-64	1960-74	44882	39	10.7118
7	b	1960-64	1975-79	17176	29	9.7513
8	b	1965-69	1960-74	28609	58	10.2615
9	c	1960-64	1960-74	1179	1	7.0724
10	c	1960-64	1975-79	552	1	6.3135

Figure 17.4. Data Window with **L_MONTHS** Added

⟹ **Deselect L_MONTHS in the data window.** Some values of **MONTHS** are **0**, meaning that this kind of ship has not seen service. You need to restrict these observations from entering into the model fit. The log transformation does this automatically since *log*(**MONTHS**) becomes a missing value for the observations with a value of **0** for **MONTH**. Observations with missing values for the explanatory variables or the response variable are not used in the model fit.

Now you are ready to begin the analysis.

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

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

⟹ **Select TYPE, YEAR, and PERIOD, then click the Expand button.**
TYPE, YEAR, and **PERIOD**, along with all two-way interaction effects, appear in the **X** variables list. Your variables dialog should now appear as shown in Figure 17.5.

Figure 17.5. Fit Variables Dialog with Variable Roles Assigned

The **Expand** button provides a convenient way to specify interactions of any order. The order **2** is the default. You can change the order by entering a different value to replace the **2** or by clicking on the buttons to the right or left of the **2** to increase or decrease the order, respectively.

⟹ **Click the Method button to display the fit method dialog.**
This dialog enables you to specify the probability function or the quasi-likelihood function for the response variable and the link function.

Overdispersion is a phenomenon that occurs occasionally with binomial and Poisson data. For Poisson data, it occurs when the variance of the response Y exceeds the Poisson variance Var(y)=μ. To account for the overdispersion that might occur in the **SHIP** data set, a quasi-likelihood function with variance function Var(μ)=μ (Poisson variance) will be used for the response variable. The variance is given by

$$\mathrm{Var}(\mathrm{y}) = \sigma^2 \mu$$

where σ^2 is the dispersion parameter with value greater than 1 for overdispersion.

⟹ **Select the check box for Quasi-Likelihood.**

⟹ **Click on Poisson under Response Dist.**
This uses the Poisson variance function Var(μ) = μ for the quasi-likelihood function.

⟹ **Click on Pearson under Scale.**
This uses the scale parameter based on the Pearson χ^2 statistic.

⟹ **Select L_MONTHS in the list at the left, then click the Offset button.**
L_MONTHS appears in the **Offset** variables list. Your method dialog should now appear as shown in Figure 17.6.

Figure 17.6. Fit Method Dialog

It is not necessary to specify a **Link Function**. **Canonical** is the default and allows **Fit (Y X)** to choose an appropriate link. For this example, it is equivalent to choosing **Log** as the **Link Function**.

⟹ **Click the OK button to close both dialogs and display the analysis.**

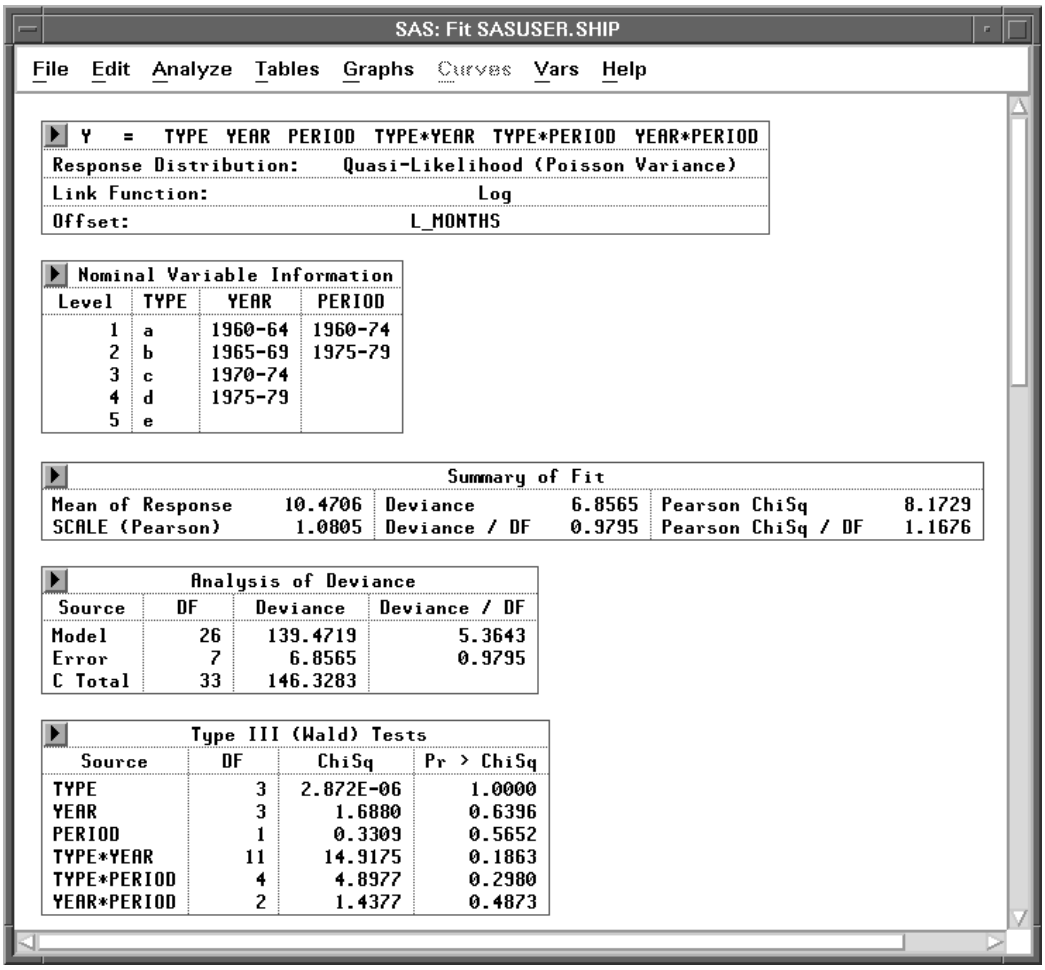


Figure 17.7. Fit Window

By default, the window includes many tables, but only a few are shown in Figure 17.7. These tables are described in the following sections. For more information about the other tables and graphs in the window, see Chapter 39, "Fit Analyses."

† **Note:** A warning message—The negative of the Hessian is not positive definite. The convergence is questionable—appears when the specified model does not converge, as in this example. The output tables, graphs, and variables are based on the results from the last iteration.

Model Information

Begin by examining the table at the top of the fit window that describes the model. The first line gives the effects in the model. The second line gives the response distribution from which the variance function used in the quasi-likelihood function is obtained. The third line gives the link function of **Y**. When an **Offset** variable is also specified in the fit method dialog, the fourth line gives the offset in the model.

The **Nominal Variable Information** table contains the levels of the nominal variables. The **Parameter Information** table, as displayed in Figure 17.1, shows the variable indices for the parameters.

Summary of Fit

The **Summary of Fit** table contains summary statistics including **Mean of Response**, **Deviance**, and **Pearson Chi-Square**. **SCALE (Pearson)** gives the scale parameter estimated from the Pearson χ^2 statistic.

Analysis of Deviance

The **Analysis of Deviance** table summarizes the information related to the sources of variation in the data. **Deviance** represents variation present in the data. **Error** gives the deviance for the current model, and **C Total**, corrected for an overall mean, is the deviance for the model with intercept only. **Model** gives the variation modeled by the explanatory variables, and it is the difference between **C Total** and **Error** deviances.

Type III (Wald) Tests

The **Type III (Wald) Tests** table in this example is a further breakdown of the variation due to **MODEL**. The **DF** for **Model** are broken down into terms corresponding to the main effects for **YEAR**, **TYPE**, and **PERIOD**, and the interaction effects for **TYPE*YEAR**, **YEAR*PERIOD**, and **TYPE*PERIOD**. The composite explanatory power of the set of parameters associated with each effect is measured by the **Chi-Square** statistic. The *p*-value corresponding to each **Chi-Square** statistic is the probability of observing a statistic of equal or greater value, given that the corresponding parameters are all 0.

Modifying the Model

For this model and this set of data, there does not appear to be sufficient explanatory power in the **YEAR*PERIOD** effect to include it in the model.

⟹ **Click on YEAR*PERIOD in the fit window.**

⟹ **Choose Edit:Delete from the menu.**

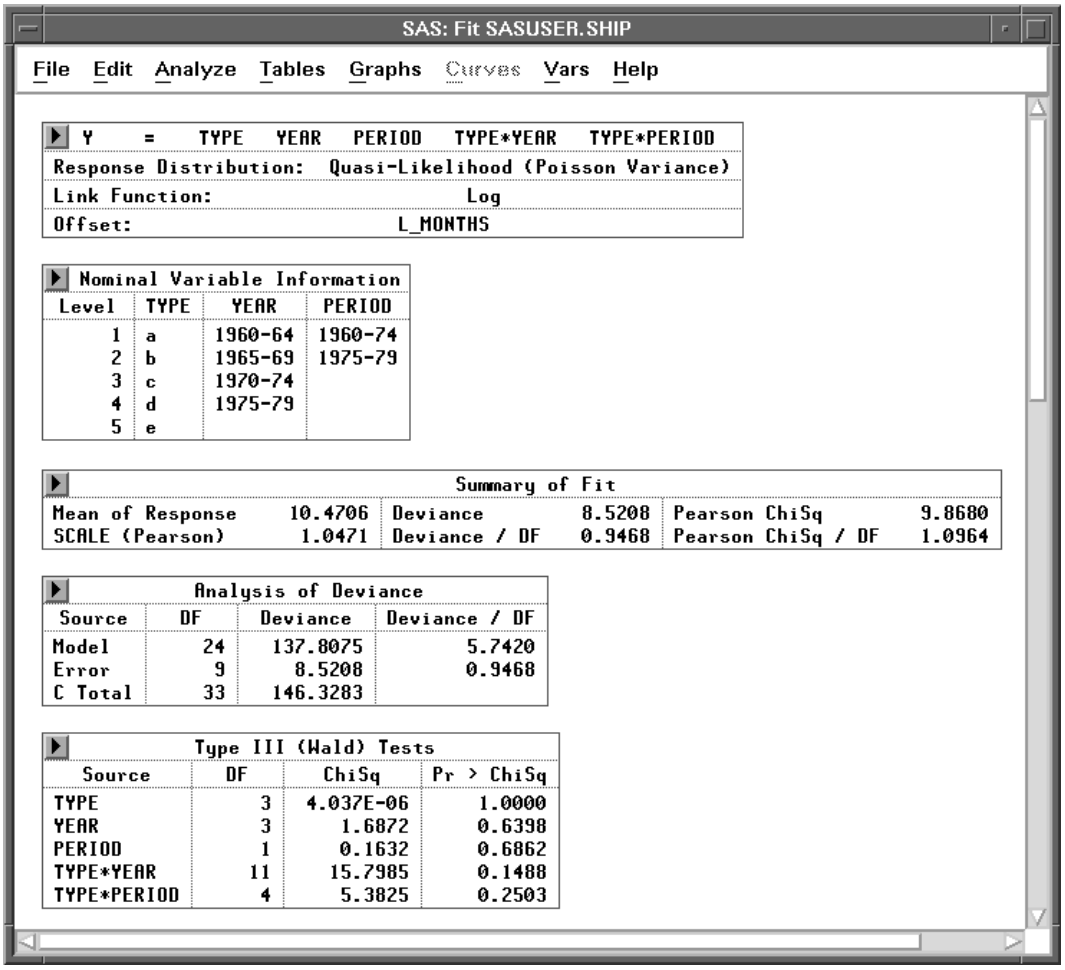


Figure 17.8. Modified Fit Model

Follow the previous steps to remove the other two interaction terms from the model. The resulting main effects model is shown in Figure 17.9.

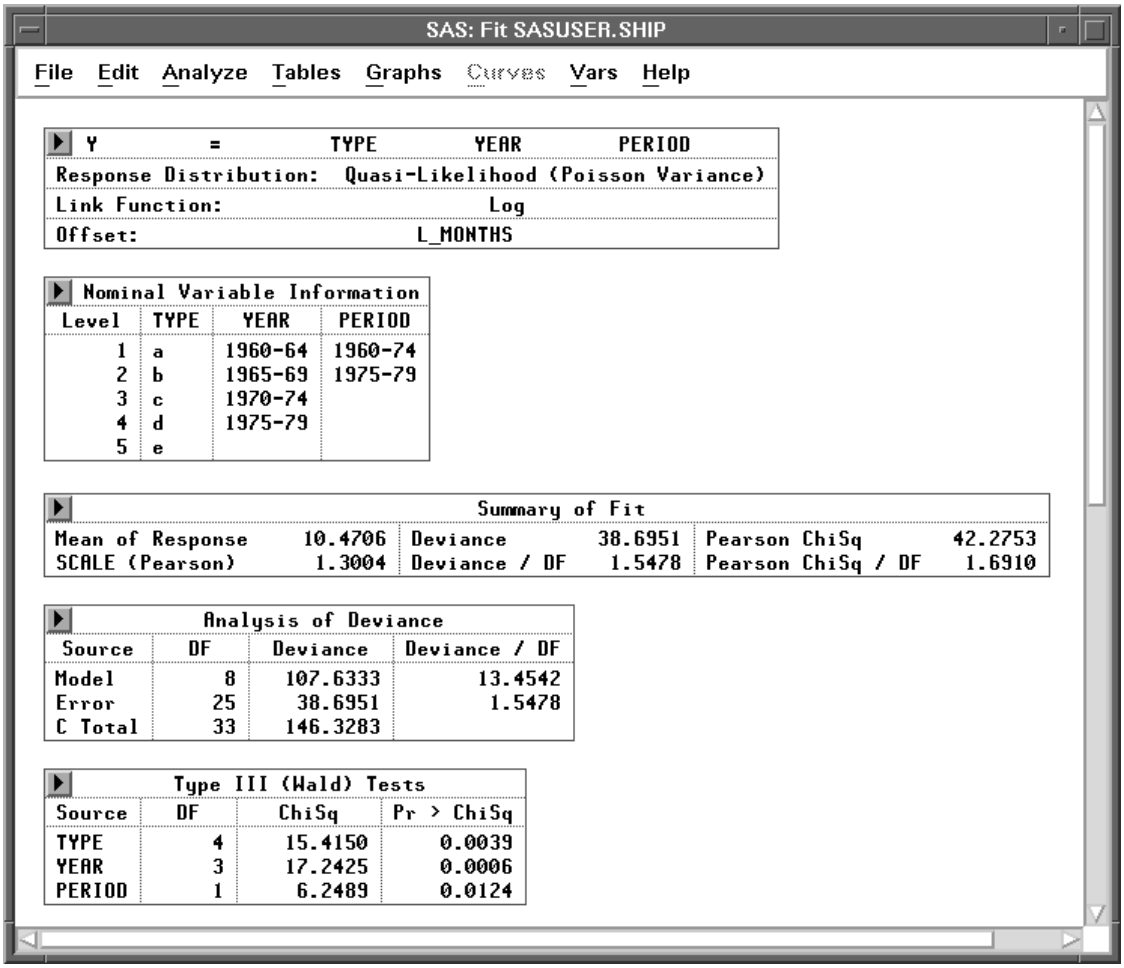
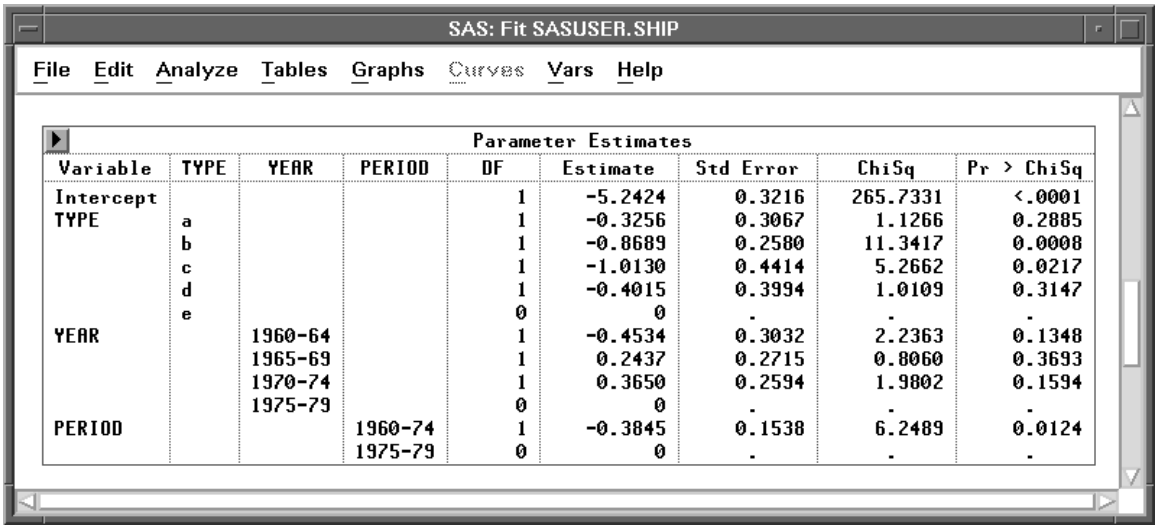


Figure 17.9. Main Effects Model

The estimate of the dispersion parameter $\phi = \sigma^2 = 1.6910$ suggests that overdispersion exists in the model. **Type III (Wald) Tests** table shows that all of the main effects are significant.

Parameter Estimates

Analyses where some effects are classification variables yield different parameter estimates from those observed in a regression setting. They represent a different additive contribution for each level value (or combination of level values for interaction effects), and thus the individual elements in the table are not as easily interpretable as they are in multiple regression.



Parameter Estimates

Variable	TYPE	YEAR	PERIOD	DF	Estimate	Std Error	ChiSq	Pr > ChiSq
Intercept				1	-5.2424	0.3216	265.7331	<.0001
TYPE	a			1	-0.3256	0.3067	1.1266	0.2885
	b			1	-0.8689	0.2580	11.3417	0.0008
	c			1	-1.0130	0.4414	5.2662	0.0217
	d			1	-0.4015	0.3994	1.0109	0.3147
	e			0	0	.	.	.
YEAR		1960-64		1	-0.4534	0.3032	2.2363	0.1348
		1965-69		1	0.2437	0.2715	0.8060	0.3693
		1970-74		1	0.3650	0.2594	1.9802	0.1594
		1975-79		0	0	.	.	.
PERIOD			1960-74	1	-0.3845	0.1538	6.2489	0.0124
			1975-79	0	0	.	.	.

Figure 17.10. Parameter Estimates Table

Because the overall level is set by the **INTERCEPT** parameter, the set of parameters associated with an effect is redundant. This shows up in the **Parameter Estimates** table as parameters with degrees of freedom (**DF**) that are **0** and estimates that are **0**. An example of this is the parameter for the **e** level of the **TYPE** variable.

From the **Parameter Estimates** table, ships of types **b** and **c** have the lowest risk, and ships of type **e** the highest. The oldest ships (built between 1960 and 1964) have the lowest risk and ships built between 1965 and 1974 have the highest risk. Ships operated between 1960 to 1974 have a lower risk than ships operated between 1975 to 1979.

The analysis provides a table for the complete fitted model, but you should not use these parameter estimates and their associated statistics individually to determine which parameters have an effect. For further information on parameter estimates and other features of the Fit window, see Chapter 39, “Fit Analyses.”

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

References

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