Chapter 17 Poisson Regression

Chapter Table of Contents

DISPLAYING THE P	OIS	SS ()N	ł	RE	G	R	ES	SS	IC)N	A	N	A	L	S	IS					. 2	64
Model Information .																		•				. 2	69
Summary of Fit					•											•		•				. 2	69
Analysis of Deviance					•											•		•				. 2	69
Type III (Wald) Tests					•				•	•	•	•			•	•	•	•			•	. 2	:69
MODIFYING THE M	OD)EI																				. 2	270
Parameter Estimates					•				•	•	•	•			•	•	•	•			•	. 2	:72
REFERENCES																						. 2	272

Part 2. Introduction

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 (YX)** 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.

-			S	AS: Fit SA	ASUSER.SHIP
<u>F</u> ile <u>E</u> dit <u>/</u>	<u>A</u> nalyze <u>T</u> ak	oles <u>G</u>	raphs <u>Cu</u>	irves <u>V</u> a	ars Help
N Y	-	TYPE	YE	AR	PERIOD
Response I	_)istributior	n: Qua	si-Likeli	hood (Poi	isson Variance)
Link Funct	tion:			Log	
Offset:			L_MONTI	HS	
Nominal	Variable I	oformat	ion		
Level T	YPE YEAR	PER			
1 a	1960-64	1960	-74		
2 b 3 c	1965-69) 1975 	-79		
4 d	1975-79	9			
5 e					
	Paramete	r Info	rmation		
Parameter	Variable	TYPE	YEAR	PERIOD	
1	Intercept	_			
3	1116	b			
4		C			
6		e			
7	YEAR		1960-64		
9			1965-65		
10	DEDIOD		1975-79	1000 74	
11	PEKIUD			1975-79	
Þ		HOUTUG		Model	l Equation
	י , – L. ה 401י	_MUNIH: 5 P 5) = -	5.2929 0 4534 P	$-0.3256 P_2 - 0.8683 P_3 - 1.0130 P_4$ P 7 + 0.2437 P 8 + 0.3650 P 9
	0.384	5 P_11			
			S	ummary of	f Fit
SCALE (Pea	esponse arson)	1.3004) Devianc Devianc	e e/DF	38.6351 Pearson Chi5q 42.2753 1.5478 Pearson Chi5q / DF 1.6910
					· · · · · · · · · · · · · · · · · · ·
	Analysis	of Dev	iance	/ 05	
Source Model	0r Dev 8 16	/1ance)7.6333	Deviance	e / Ur 3.4542	
Error	25 3	38.6951	1	1.5478	
C Total	33 14	6.3283			
	Type III (W	ald) T	ests		
Source	DF Ch	i Sq	Pr → Chi	5q	
ТҮРЕ	4 15	5.4150	0.00	39	
PERIOD	з 17 1 б	5.2489	0.00	24	

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).

				SAS: SA	SUSER. SH	IIP		·
E	ile	Edit	Analyza	e Table	s Graph	s Curves	s Vars	Help
	5	Nom	Nom	Nom	Int	Int		
40	\sim	TYPE	YEAR	PERIOD	MONTHS	Y		
	1	Ь	1965-69	1975-79	20370	53		Ì
	2	Ь	1970-74	1960-74	7064	12		
	3	Ь	1970-74	1975-79	13099	44		
	4	Ь	1975-79	1960-74	0	0		
	5	Ь	1975 - 79	1975-79	7117	18		
	6	Ь	1960-64	1960-74	44882	39		
	7	Ь	1960-64	1975-79	17176	29		
	8	Ь	1965 - 69	1960-74	28609	58		
	9	C	1960-64	1960-74	1179	1		
	10	C	1960-64	1975-79	552	1		
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(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(expected number of damage incidents)

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

 \implies 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**.

- \implies Select MONTHS in the data window.
- \implies Choose Edit:Variables:log(Y).



Figure 17.3. Edit: Variables Menu

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

-	4			SAS: SAS	SUSER.S	HIP		• □
	<u>F</u> ile	Edit	<u>A</u> nalyze	e Table	s Grapt	os Con	rves Vars	Help
	6	Nom	Nom	Nom	Int	Int	Int	
40		TYPE	YEAR	PERIOD	MONTHS	Y	L_MONTHS	
	1	Ь	1965-69	1975-79	20370	53	9.9218	
	2	Ь	1970-74	1960-74	7064	12	8.8628	
	3	Ь	1970-74	1975-79	13099	44	9.4803	
	4	Ь	1975-79	1960-74	0	0	•	
	5	Ь	1975-79	1975-79	7117	18	8.8702	
	6	Ь	1960-64	1960-74	44882	39	10.7118	
	7	Ь	1960-64	1975-79	17176	29	9.7513	
	8	Ь	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	
E	i							

Figure 17.4. Data Window with L_MONTHS Added

 \implies 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.

- \implies Choose Analyze:Fit (Y X) to display the fit variables dialog.
- $\implies Select Y in the list at the left, then click the Y button.$ Y appears in the Y variables list.
- \implies 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.

\implies 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 $Var(y)=\mu$. To account for the overdispersion that might occur in the **SHIP** data set, a quasi-likelihood function with variance function $Var(\mu)=\mu$ (Poisson variance) will be used for the response variable. The variance is given by

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

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

- \implies Select the check box for Quasi-Likelihood.
- \implies Click on Poisson under Response Dist.

This uses the Poisson variance function $Var(\mu) = \mu$ for the quasi-likelihood function.

\implies Click on Pearson under Scale.

This uses the scale parameter based on the Pearson χ^2 statistic.

\implies 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.

	SAS: Fi	t(Y×)	
SHIP TYPE YEAR PERIOD MONTHS Y L_MONTHS	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: <u>1</u>
V	Binomiol 0 L_MC	Co Ifset NTHS Cel Remove	nstant: <u>1</u> i-Likelihood t Distribution er's Scoring



It is not necessary to specify a Link Function. Canonical is the default and allows

Fit (YX) to choose an appropriate link. For this example, it is equivalent to choosing Log as the Link Function.

 \Longrightarrow Click the OK button to close both dialogs and display the analysis.

-		S,	AS: Fit SASUSE	R.SHIP		
File Edit Ana	lyze Tabl	es Graphs	Curves V	ars Help		
▶ Y = TY	PE YEAR P	ERIOD TYPE	*YEAR TYPE*	PERIOD Y	EAR*PERIOD	
Response Dis	tribution:	Quasi-	Likelihood (°oisson V	ariance)	
Link Function	n :		Log			
Offset:			L_MONTHS			
Nominal Va	riable Inf	formation				
Level IYPE	YEHK	PERIOD				
1 a 2 b	1965-69	1960-74				
3 c	1970-74	1313 13				
4 d	1975-79					
5 e						
				е г		
			Summary o	f fit		
Mean of Kespo	onse I m)	0.4706 Dev 1.0805 Dev	/iance /iance / DF	6.8565 0 9795	Pearson ChiSq Pearson ChiSq / DF	8.1729
	5117	1.0003 000	Viance / Di	0.0/00		
	Analysis (of Deviance	•			
Source D	F Devi	iance Dev	iance / DF			
Model	26 139	9.4719	5.3643			
Error	7 6	6.8565	0.9795			
L Total	33 146	. 3283				
	Tune III ((Wald) Test	۹.			
Source	DF	ChiSa	Pr > ChiSa			
ТҮРЕ	3	2.872E-06	1.0000			
YEAR	3	1.6880	0.6396			
	1	0.3309	0.5652			
TYPE*YEHR	11	14.9175	0.1863			
YEAR*PERIOD	2	1.4377	0.4873			
	- 1					
N						

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 \mathbf{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.

- \implies Click on YEAR*PERIOD in the fit window.
- \implies Choose Edit:Delete from the menu.

	S	SAS: Fit SASUSE	R. SHIP		· []
File Edit Analyze	e Tables Graph	s Curves Va	rs Help		1
					Δ
🕨 Y 😑 TYPI	E YEAR PERIO	D TYPE*YEAR	TYPE*PE	ERIOD	
Response Distrib	ution: Quasi-L	ikelihood (Poi	sson Var	iance)	
Link Function:		Log			
Offset:	L	_MONTHS			
Nominal Vaniah	la Information	1			
level TYPF	YFAR PFRINN				
1 a 19	160-64 1960-74				
2 Ь 19	165-69 1975-79				
3 c 19	170-74				
5 e	67-63				
i i	i	I			
		Summary of	Fit		
Mean of Response	e 10.4706 De	eviance	8.5208	Pearson ChiSq	9.8680
SCALE (Pearson)	1.0471 De	eviance / DF	0.9468	Pearson ChiSq / DF	1.0964
D Anal	lusis of Deviand	e			
Source DF	Deviance De	viance / DF			
Model 24	137.8075	5.7420			
Error 9	8.5208	0.9468			
C Total 33	146.3283				
D Tune	a III (Wald) Tes	ts			
Source D	IF ChiSa	Pr > ChiSa			
ТҮРЕ	3 4.037E-06	1.0000			
YEAR	3 1.6872	0.6398			
PERIOD	1 0.1632	0.6862			
	4 5.3825	0.2503			
					V

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.

-			8	SAS: Fit SAS	USER.SH	IP			• 🗆
File Edit	Analy	ze Table	s Graph	s Curves	Vars I	lelp			
	_								
► Y	-		ТЧРЕ	YEAR	PEF	RIOD]		
Response	e Disti	ribution:	Quasi-Li	ikelihood (Poisson	Variance)			- 11
Link Fu	nction:			Log					- 11
Offset:			L_	MONTHS					- 11
Nomin	al Var	iable Inf	ormation						
Level	ITPE	TEHK	PERIUU						- 14
2	a h	1965-69	1960-74						- 14
3	c	1970-74							- 11
4	d	1975-79							- 14
5	e								- 18
N				S.uma and	. of Fit				ıЦ
Mean of	Resnor	10 10	4706 De	Juliance	38 6	951 Pearse	n ChiSa	42 2753	
SCALE (Pearson	n) 1	.3004 De	viance / D	F 1.5	478 Pearso	on ChiSq / DF	1.6910	
					_				
	A	nalysis o	f Devianc	e					
Source	DF	Devia	ance Dev	viance / DF					
Model		8 107. 25 20	.6333	13.4542	2				
C Total		25 30. 33 146.	. 3283	1.3470	,				
	:								
	Type	e III (Wal	d) Tests						
Source	DF	ChiS	q Pr 3	≻ ChiSq					
TYPE	4	15.4	4150	0.0039					
YEAR PEPTOD		3 17.7	2425	0.0005					
		. 0.,	LTUJ	0.0127					∇
<₽									

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.

le <u>E</u> dit <u>A</u>	nalyze	Tables	Graphs	Curves	<u>V</u> ars <u>H</u> elp			
				Parame	ter Estimate	5		
Variable	TYPE	YEAR	PERIOD	DF	Estimate	Std Error	ChiSq	Pr → ChiSq
Intercept				1	-5.2424	0.3216	265.7331	<.0001
TYPE	а			1	-0.3256	0.3067	1.1266	0.2885
	Ь			1	-0.8689	0.2580	11.3417	0.0008
	с			1	-1.0130	0.4414	5.2662	0.0217
	d			1	-0.4015	0.3994	1.0109	0.3147
	е			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

Copyright © 1999 by SAS Institute Inc., Cary, NC, USA.

ISBN 1-58025-490-X

All rights reserved. Printed in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

U.S. Government Restricted Rights Notice. Use, duplication, or disclosure of the software by the government is subject to restrictions as set forth in FAR 52.227–19 Commercial Computer Software-Restricted Rights (June 1987).

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

1st printing, October 1999

 SAS^{\circledast} and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. $^{\circledast}$ indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

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