Homework 9 with Solutions

Directions. This homework pertains to materials on Poisson regression in Lesson 9. The assignment should be typed, with your name on the document, and with properly labeled computer output. I suggest you attach your SAS/R input code at the end of your file clearly indicating the problem it corresponds to. If you choose to collaborate, the write-up should be your own. Please show your work! Upload the file to the HW9 Dropbox on ANGEL.

1. *40 pts* An experiment analyzes imperfection rates for two processes used to fabricate silicon wafers for computer chips. For treatment A applied to 10 wafers, the number of imperfections are 8, 7, 6, 6, 3, 4, 7, 2, 3, 4. Treatment B applied to 10 wafers has 9, 9, 8, 14, 8, 13, 11, 5, 7, 7 imperfections. Treat the counts as independent Poisson variates having means μ_A and μ_B.

   (a) Fit the Poisson regression model $\log(\mu) = \alpha + \beta x$, where $x = 1$ for treatment A and $x = 0$ for treatment B. Show that $\beta = log\mu_A - \log \mu_B$. Interpret the estimate.

   Solutions:

   For treatment A, X=1, so $\log \mu_A = \alpha + \beta$ and for treatment B, X=0, so $\log \mu_B = \alpha + \beta(0) = \alpha$. Therefore, $\log \mu_A - \log \mu_B = \alpha + \beta - \alpha = \beta$.

   From SAS/R, we get the following estimates:

	DF	Estimate	P-value
Intercept	1	2.2083	<-.0001
X	1	-0.5988	0.0007

   Since $\exp(\beta) = \exp(-0.5988) = 0.549$, so compared to treatment B, treatment A decreases the imperfections by a rate of about 55%.

   (b) Test $H_0 : \mu_A = \mu_B$, using either a Wald test or a likelihood ratio test from you SAS or R output from the Poisson regression model you fitted. Interpret.

   Solutions: $H_0 : \mu_A - \mu_B = 0$

   The Wald chi-square statistic for β is 11.57 with DF=1, so p-value=0.0007. Therefore, we reject the null hypothesis and conclude that the mean imperfections are different for two treatment groups.

R code/Output

```
> data_wafer = read.table("wafer.txt")
> fit_wafer = glm(data_wafer, family=poisson(link=log))
> summary(fit_wafer)

Call:
glm(formula = data_wafer, family = poisson(link = log))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5280  -0.7259  -0.2028   0.6680   1.5040

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.2083     0.1048  21.066  < 2e-16 ***
V2           -0.5988     0.1760  -3.402  0.00067 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 27.700  on 19  degrees of freedom
Residual deviance: 15.604  on 18  degrees of freedom
AIC: 93.835

Number of Fisher Scoring iterations: 4

> fit_wafer$fitted
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
17  18  19  20
5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 9.1 9.1 9.1 9.1 9.1
 9.1 9.1 9.1 9.1 9.1
>
```

SAS code/Output:

```
data wafers;
input Imperf Treatment $;
cards;
8  1
7  1
```

```
6  1
6  1
3  1
4  1
7  1
2  1
3  1
4  1
9  0
9  0
8  0
14 0
8  0
13 0
11 0
5  0
7  0
7  0
;
run
;

Proc Genmod data = wafers;
class Treatment;
Model Imperf = Treatment / dist=poi link=log obstats;
output out=data2;
proc print;
run;
```

The SAS System

The GENMOD Procedure

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	2.2083	0.1048	2.0028	2.4137	443.76	<.0001
wafer	A	1	-0.5988	0.1760	-0.9439	-0.2538	11.57	0.0007
wafer	B	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Coefficients for Contrast test

Label	Row	Prm1	Prm2	Prm3
test	1	0	1	-1

Contrast Results

Contrast	DF	Chi-Square	Pr > ChiSq	Type
test	1	12.10	0.0005	LR

2. *60 pts* The data below were reported by Laird and Olivier (1981) on the survival of patients after heart-valve replacement surgery. Varying numbers of patients fell into the two categories of age (Under 55, and 55+), two types of types of heart valve (Aortic and Mitral), and they were followed for different lengths of time in terms of days (values under label exposure), and the the last column is the total number of deaths for the combination of the three predictors. See heart.sas, and/or heart.R and poisson.R. (Note: For R users first run poisson.R which was written for this class then heart.R or you can use glm() and write your own code like we do in the online notes).

Age	*Type*	Exposure	Deaths
Under 55	Aortic	1,259	4
	Mitral	2,082	1
55+	Aortic	1,417	7
	Mitral	1,647	9

a. Under a saturated model, we can estimate the mean death rates directly. Let λ_1 be the mean death rate for individuals Under55-Aortic combination. Just looking at the table (even without running the SAS or R code), do you know the estimate of this value? When you take the natural log of this estimate, which parameter estimate in your model do you expect to get?

Solutions: The mean death for individuals under 55 and Aortic is 4/1259=0.0032. The model is

$$log(\lambda) = \beta_0 + \beta_1 X_{age} + \beta_2 X_{type} + \beta_3 X_{age} X_{type}$$

where λ is the death rate, $X_{age} = I(age \leq 55)$ is the indicator for age, $X_{type} = I_{Aortic}$ is the indicator for type. In the above model $log(0.0032) = -5.75$ is the estimate for $\beta_0 + \beta_1 + \beta_2 + \beta_3$.

b. Examine the Wald statistics (that is z-values) of the saturated model output. Which predictors are significant? Interpret the parameters of this model.

Solutions:

From the SAS/R output summarized below, we can say that the intercept and age are significant.

	estimate	Chi-square value	p-value
β_0	-5.2095	244.25	<0.0001
β_1	-2.4316	5.32	0.0211
β_2	-0.1009	0.04	0.8413
β_3	1.9902	2.63	0.1046

Interpretation:

β_0: For individuals with Age $>$ 55 and Mitral, the mean death rate is exp(β_0) $= 0.0546$.

β_1: For individuals with Mitral, the death rate of those with age < 55 decreases by multiplicative factor exp(β_1) $= 0.08789$ compared to those with age > 55.

β_2: For individuals of Age $>$ 55, the mean death rate of those with Aortic decreases by a multiplicative factor exp(β_2) $= 0.904$ compared to those with Mitral.

$\beta_3 \neq 0$ means that when comparing individuals of Age $<$ 55 and Age $>$ 55, the change in death rate is different for individuals with Mitral or Aortic.

c. Why did we use the *offset* in this model?

Solutions:

We used an offset term (exposure) because it is not reasonable to look at death count itself. In the study four different groups of patients were followed for different time spans, and groups with longer exposure will more likely to have more death counts. Therefore, the death rates rather than the death counts are more comparable, and that is why we used the exposure as an offset term.

In our model the response (death rate) with log link is

$$log(\lambda) = log(\mu/n) = log(\mu) - log(n) \tag{1}$$

where λ is the death rate, μ is the death counts and n is the exposure counts. The offset term is $log(n)$.

d. What would be your criticism of this model, if any? Do you think that main effects model would be better?

Solutions:

If we look at the Wald statistics, only the main effect of age is significant. However, we should not conclude that type is not an important variable. We suggest a lack of fit test to see whether a model with only age fits the data well.

R code/Output

```
> ### Problem 2  ###
> Y = c(4,1,7,9)
> expo = c(1259,2082,1417,1647)
> age = c(1,1,0,0)
> type = c(1,0,1,0)
> X = cbind(intercept=1, age=age, type=type, age_type=age*type)
> fitmodel = poisson.regression(X,Y,offset=log(exposure))
1...2...3...4...5...
> poisson.print(fitmodel)
The Newton-Raphson algorithm converged in 5 iterations.

                coef        SE coef/SE  pval
intercept -5.2094862 0.3333333  -15.63 0.000
age       -2.4315981 1.0540926   -2.31 0.021
type      -0.1009009 0.5039526   -0.20 0.841
age_type   1.9902065 1.2263638    1.62 0.105

Loglikelihood = 17.9415696838927
Pearson's X^2 = 2.69968494366806e-23
 Deviance G^2 = -1.11022305162167e-15
           df = 0
50% of observations have expected counts below 5.0
The minimum expected cell count is 1
>
```

SAS code/Output

```
/*heart-valve example on Poisson regression */
/* re: Lesson 9*/
options nocenter nodate nonumber linesize=72;
```

```
data heart;
input age $ type $ exposure y;
o=log(exposure);
cards;
Under55 Aortic 1259 4
Under55 Mitral 2082 1
55+ Aortic 1417 7
55+ Mitral 1647 9
;

/*saturated poisson regression with offset*/
proc genmod data=heart order=data;
class age type;
model y = age type age*type / dist=poisson link=log offset=o;
run;
```

The SAS System

The GENMOD Procedure

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept			1	-5.2095	0.3333	-5.8628	-4.5562	244.25	<.0001
AGE	<55		1	-2.4316	1.0541	-4.4976	-0.3656	5.32	0.0211
AGE	55+		0	0.0000	0.0000	0.0000	0.0000	.	.
TYPE	Aortic		1	-0.1009	0.5040	-1.0886	0.8868	0.04	0.8413
TYPE	Mitral		0	0.0000	0.0000	0.0000	0.0000	.	.
AGE*TYPE	<55	Aortic	1	1.9902	1.2264	-0.4134	4.3938	2.63	0.1046
AGE*TYPE	<55	Mitral	0	0.0000	0.0000	0.0000	0.0000	.	.
AGE*TYPE	55+	Aortic	0	0.0000	0.0000	0.0000	0.0000	.	.
AGE*TYPE	55+	Mitral	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale			0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.