Handout 12

Poisson Regression

Consider the data set in which we have number of broken ampules in cartons of 1000. Also recorded is the number of times each carton was transferred during the shipment. In Handout 9, the Ampule Breakage data had an outlier which turned out to be a typo. In this handout, the typo has been corrected and there is no outlier.

Denote by Y_i the number of broken ampules in the i^{th} carton which had X_i transfers. Let us assume that Y_i is distributed as Poisson(μ_i). Note that Y_i is indeed Binomial(N,π_i), with $N = 1000$. It is known from probability theory that when $N\pi_i$ is small to moderate, then the binomial distribution can be modeled by Poisson with mean $\mu_i = N\pi_i$. Note that if $Y_i \sim$Poisson(μ_i), then $E(Y_i) = \mu_i$ and $Var(Y_i) = \mu_i$. We will model $\log(\mu_i)$ by a linear function of X_i, i.e., we will write $\log(\mu_i) = \beta_0 + \beta_1 X_i$. So the "link" function here is the natural logarithm. In the accompanying graph, we have plots of Y vs X, $log(Y + 1/2)$ vs X, and $(Y + 3/8)^{1/2}$ vs X. From the plot of Y vs X, it seems that both the mean and variance of Y vary with X. This is not surprising since we see that for the Poisson distribution $E(Y_i) = \mu_i$ and $Var(Y_i) = \mu_i$.

We have plotted $\log(Y + 1/2)$ against X to see if the relation is linear as postulated by the model. We could have plotted $\log(Y)$ vs X, but $E[\log(Y + 1/2)]$ is closer to $log(\mu)$ than $E[\log(Y)]$ is. The plot looks reasonably linear, thus it may be justifiable to model $\log(\mu_i)$ as a linear function of X_i. Finally, when the counts are not small, one may bypass Poisson modeling by taking a square root transformation of Y and then use a standard regression model for $\sqrt{Y}$ against X. In such a case, it is better to use Anscombe transformation $\sqrt{Y + 3/8}$ instead of $\sqrt{Y}$ since the distribution of$\sqrt{Y + 3/8}$ is approximately $N\left(\sqrt{\mu}, 1/4\right)$ if the value of μ is moderate to large.

Here is the output from the R command glm(Y~X,family=poisson):

Call:

glm(formula = Y ~X, family = poisson)

Deviance Residuals

Min	1Q	Median	3Q	Max
−1.08628	−0.40777	−0.00965	0.33945	1.20392

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.21895	0.10150	21.86	$< 2e-16 **$
X	0.35711	0.02729	13.09	$< 2e-16 **$

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 197.5184 on 19 degrees of freedom

Residual deviance: 7.6906 on 18 degrees of freedom

AIC: 109

Number of Fisher Scoring iterations: 4

xxxxxxxxxxx

Clearly, the slope is highly significant, and it indicates that the mean breakage $\mu(X)$ depends on X, the number of transfers.

Note that

$$\frac{\hat{\mu}(X+1)}{\hat{\mu}(X)} = \exp(\hat{\beta}_1) = 1.4292.$$

Thus we can say that, for any X, the mean ampule breakage $\mu(X+1)$ at $X+1$ is estimated to be 1.4 times the mean ampule breakage $\mu(X)$ at X, ie, the mean amuple breakage is estimated to be 40% higher for unit change in the number of transfers. So for the model $\mu(X) = \exp(\beta_0 + \beta_1 X)$, for any X, $\mu(X+1)$ is proportional to $\mu(X)$ and the constant of proportionality is $\exp(\beta_1)$. It should be noted that if $\log(\mu(X))$ is modeled as a quadratic function of X, the ratio $\mu(X+1)/\mu(X)$ would depend on X.

Goodness-of-fit

We wish to test if the model $\log(\mu_i) = \beta_0 + \beta_1 X_i$ is reasonable for the Ampule Breakage data. So we test $H_0 : \log(\mu_i) = \beta_0 + \beta_1 X_i$ for all i, vs $H_1 : \log(\mu_i)$'s do not lie on a straight line. For the **saturated** model, $\mu_1, \ldots, \mu_n$ are allowed to be arbitrary and the estimate of μ_i is Y_i. Thus G^2 is the Residual deviance, ie, $G^2 = -2[\log(L_M)) - \log(L_S)]$, where L_M is the likelihood for the linear model in $\log(\mu_i)$. Since all the counts are 5 or larger, we may use the result that $G^2 \sim \chi^2_{n-p}$ under H_0, where p is the number of beta parameter estimated in the log-linear model. Here $n - p = 20 - 2 = 18$, and the p-value of the test is area to the right of 7.6906 under the χ^2_{18} curve, and the area is about 0.983. Thus we cannot reject the null. Conclusion: the log-linear model $\mu(X) = \exp(\beta_0 + \beta_1 X)$ is reasonable for the Ampule Breakage data.

Residuals.

Note that for Poisson regression:

Pearson residuals: $e_i = (Y_i - \hat{\mu}_i)/\sqrt{\hat{\mu}_i}$,

Deviance residuals: $dev_i = sign(Y_i - \hat{\mu}_i)[-2Y_i \log(\hat{\mu}_i/Y_i) - 2(Y_i - \hat{\mu}_i)]^{1/2}$,

where $\hat{\mu}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)$, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the maximum likelihood estimates of β_0 and β_1.

If the null hypothesis were true, then both e_i and dev_i are approximately normally distributed with mean 0, but their variances are a bit smaller than 1. Standardized Pearson or Deviance residuals can be obtained from R, and these stanardized residuals are approximately distributed as $N(0,1)$ if the null were true. We have plotted the deviance residuals $\{e_i\}$ against the transfers $\{X_i\}$. All the deviance residuals are between -1.5 and 1.5. This indicated that the model $\mu(X) = \exp(\beta_0 + \beta_1 X)$ is a reasonable model for the mean transfers.

Estimation of parameters:

The likelihood function here is

$$\begin{aligned} L &= \prod_{i=1}^{n} \left[e^{-\mu_i} \mu_i^{Y_i} / Y_i! \right], \\ \log L &= \sum_{i=1}^{n} [-\mu_i + Y_i \log \mu_i - \log(Y_i!)] \end{aligned}$$

If $\mu_i = \exp(\beta_0 + \beta_1 X_i)$, then one maximizes L with respect to β_0 and β_1 or equivalently maximize $\log(L)$ with respect to β_0 and β_1. If $\log(L)$ is differentiated with respect to β_0 and β_1, then setting the derivatives

to zero lead to the following likelihood equations

$$\sum \mu_i = \sum Y_i \text{ and } \sum X_i \mu_i = X_i Y_i,$$

or in vector matrix notations, the likelihood equations are

$$\boldsymbol{X}^T \boldsymbol{\mu} = \boldsymbol{X}^T \boldsymbol{Y},$$

where $\boldsymbol{X}$ is $n \times 2$ a matrix whose first column consists of 1's and the second column has $X_1, \ldots, X_n$, $\boldsymbol{Y}$ is the vector of $Y_1, \ldots, Y_n$, and $\boldsymbol{\mu}$ is the vector of $\exp(\beta_0 + \beta_1 X_1), \ldots, \exp(\beta_0 + \beta_1 X_n)$. There are no closed form solutions and the ML estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained by solving the likelihood equations using iterative methods.

If $\hat{\mu}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the ML estimates of β_0 and β_1, then the loglikelihood is

$$\log L_M = \sum_{i=1}^{n} [-\hat{\mu}_i + Y_i \log \hat{\mu}_i - \log(Y_i!)]$$

Under the saturated model, estimate of μ_i is Y_i, and hence

$$\log L_S = \sum_{i=1}^{n} [-Y_i + Y_i \log Y_i - \log(Y_i!)].$$

Thus the residual deviance is

$$\begin{aligned} & -2[\log(L_M) - \log(L_S)] \\ = \; & -2 \sum [(Y_i - \hat{\mu}_i) + Y_i \log(\hat{\mu}_i / Y_i)] = \sum [-2 Y_i \log(\hat{\mu}_i / Y_i) - 2(Y_i - \hat{\mu}_i)]. \end{aligned}$$

Ampule Breakage data:

X = number of transfers during shipment

Y = number of broken ampules in a carton 1000.

Y	11	9	17	55	12	22	13	40	8	25	8	47
X	1	0	2	5	0	3	1	4	0	3	0	5
$\hat{\mu}$	13.15	9.20	18.79	54.85	9.20	26.85	13.15	38.38	9.20	26.85	9.20	54.85

Y	19	16	34	64	19	11	28	44
X	2	1	4	5	2	0	3	4
$\hat{\mu}$	18.79	13.15	38.38	54.85	18.79	9.20	26.85	38.38

Non-melanoma skin cancer among women.

The following are the number of cases of nonmelanoma skin cancer among women reported in 1974 in two cities Minneapolis-St.Paul (Minn) and Dallas-Ft.Worth (Dallas). Also reported are the age groups and the number of individuals in each group. For the purpose of analysis we will take the midpoint of each age group as the representative age of that group and we will take 90 to be the representative age of the 85+ group (not the best thing to do and there are other methods for handling this).

Age	City	Pop	Cases
$[15, 25)$	Minn	172,675	1
$[25, 35)$	Minn	123,065	16
$[35, 45)$	Minn	96,216	30
$[45, 55)$	Minn	92,051	71
$[55, 65)$	Minn	72,159	102
$[65, 75)$	Minn	54,722	130
$[75, 85)$	Minn	32,185	133
85+	Minn	8,328	40
$[15, 25)$	Dallas	181,343	4
$[25, 35)$	Dallas	146,207	38
$[35, 45)$	Dallas	121,374	119
$[45, 55)$	Dallas	111,353	221
$[55, 65)$	Dallas	83,004	259
$[65, 75)$	Dallas	55,932	310
$[75, 85)$	Dallas	29,007	226
85+	Dallas	7,538	65

Note that the number of cases in any age group should be proportional to the population size in the group. **If μ_i is the expected number of cases for the i^{th} case, then we should be modeling the rate, i.e., the number of cases per 1000 women.** Let w_i be the population size of the i^{th} case divided by 1000, ie, $w_1 = 172.675, w_2 = 123.065$ etc. **In the computer packages, w_i is called the "offset"**. We may model $\log(\mu_i/w_i)$ as a linear function of city and age. Let X_{i1} be the age and X_{i2} be the city (Dallas=0 and Minnesota=1). So the simplest model is of the form

$$\log(\mu_i/w_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}, \text{ i.e., } \log(\mu_i) = \log(w_i) + \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}.$$

For each city, the plot of $\log(Y/w)$ against age shows nonlinearity. In order to account for the nonlinearity, we ran a bit more more complex model

$$\log(\mu_i/w_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + \beta_4 X_{i1} X_{i2} \tag{1}$$

Here is the R output.

Call:

glm(formula = y ~age + age2 + city + age * city + offset(log(w)),

family = poisson)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.3798	-0.7818	0.2523	0.6865	2.3927

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	$-5.478e+00$	$2.982e-01$	-18.373	$< 2e-16 ***$
age	$1.678e-01$	$9.940e-03$	16.883	$< 2e-16 ***$
age2	$-9.262e-04$	$8.165e-05$	-11.343	$< 2e-16 ***$
city	$-1.351e+00$	$2.380e-01$	-5.677	$1.37e-08 ***$
age:city	$8.395e-03$	$3.551e-03$	2.364	$0.0181*$

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2790.34 on 15 degrees of freedom

Residual deviance: 31.99 on 11 degrees of freedom

AIC: 136.24

Number of Fisher Scoring iterations: 4

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Note that the all the parameters are highly significant. It should be pointed out that residual deviance is 31.99 with 11 df with a p-value of 0.0008 (area to the right of 31.99 under the χ^2_{11} curve). This indicates that the model given in (1) may not be adequate. We obtained the standradized residuals using the R command: rstudent(PoissonReg), where 'PoissonReg' is the R object when fitting model (1). The plot of the studentized residuals against age show that there are few residuals which our outside $(-2.5, 2.5)$. This seems to validate the formal conclusion that model (1) is not quite appropriate here.

Better modeling may be possible in this case, and one may try a cubic model. Though not shown here, even a cubic model turns out to be inadequate. We can try another modeling scheme by transforming age as described below.

Note that the relation between $\log(y/w)$ is an increasing nonlinear function of age and the rate of increase seems to decrease for ages higher than 50. A plot of $\log(y/w)$ against $\log(age)$ is given here. We try a cubic polynomial model for $\log(y/w)$ in $\log(age)$. We fitted the model

$$\begin{aligned} \log(\mu_i/w_i) &= \beta_0 + \beta_1 \log(X_{i1}) + \beta_2 \log(X_{i1})^2 + \beta_3 \log(X_{i1})^3 + \beta_4 X_{i2} + \beta_5 \log(X_{i1}) X_{i2} \\ &\quad + \beta_6 \log(X_{i1})^2 X_{i2} + \beta_7 \log(X_{i1})^3 X_{i2}. \end{aligned} \tag{2}$$

Note that fitting this model is equivalent to fitting two separate cubic polynomials. The R output for is not shown here for model (2). Backward stepwise regression (using the 'step' function in R) kept the following variables: $\log(X_{i1}), \log(X_{i1})^2, X_{i2}$ and $\log(X_{i1})X_{i2}$, and the residual deviance for the final model is 12.961 with df=11, and a p-value=0.296. This clearly suggests that the final model is a reasonable description for the data.

The following is the R output when we fit the final model ('lage', 'lage2' and 'lage3' refer to $\log(age)$, $\log(age)^2$ and $\log(age)^3$).

Call:
glm(formula = cases ~lage + lage2 + city+ lage*city + offset(log(w)),
family = "poisson")

Deviance	Residuals			
Min	1Q	Median	3Q	Max
−1.6708	−0.6539	−0.1969	0.7152	1.5041

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−30.6648	3.5855	−8.553	$< 2e-16 **$
lage	11.5005	1.8035	6.377	$1.81e-10 **$
lage2	−1.0430	0.2264	−4.607	$4.09e-06 **$
city	−2.6223	0.8602	−3.049	$0.0023 **$
lage:city	0.4391	0.2069	2.123	$0.0338*$

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 2790.340 on 15 degrees of freedom
Residual deviance: 12.961 on 11 degrees of freedom
AIC: 117.21
Number of Fisher Scoring iterations: 4

Figure 1: Poisson Regression: Data on broken ampules

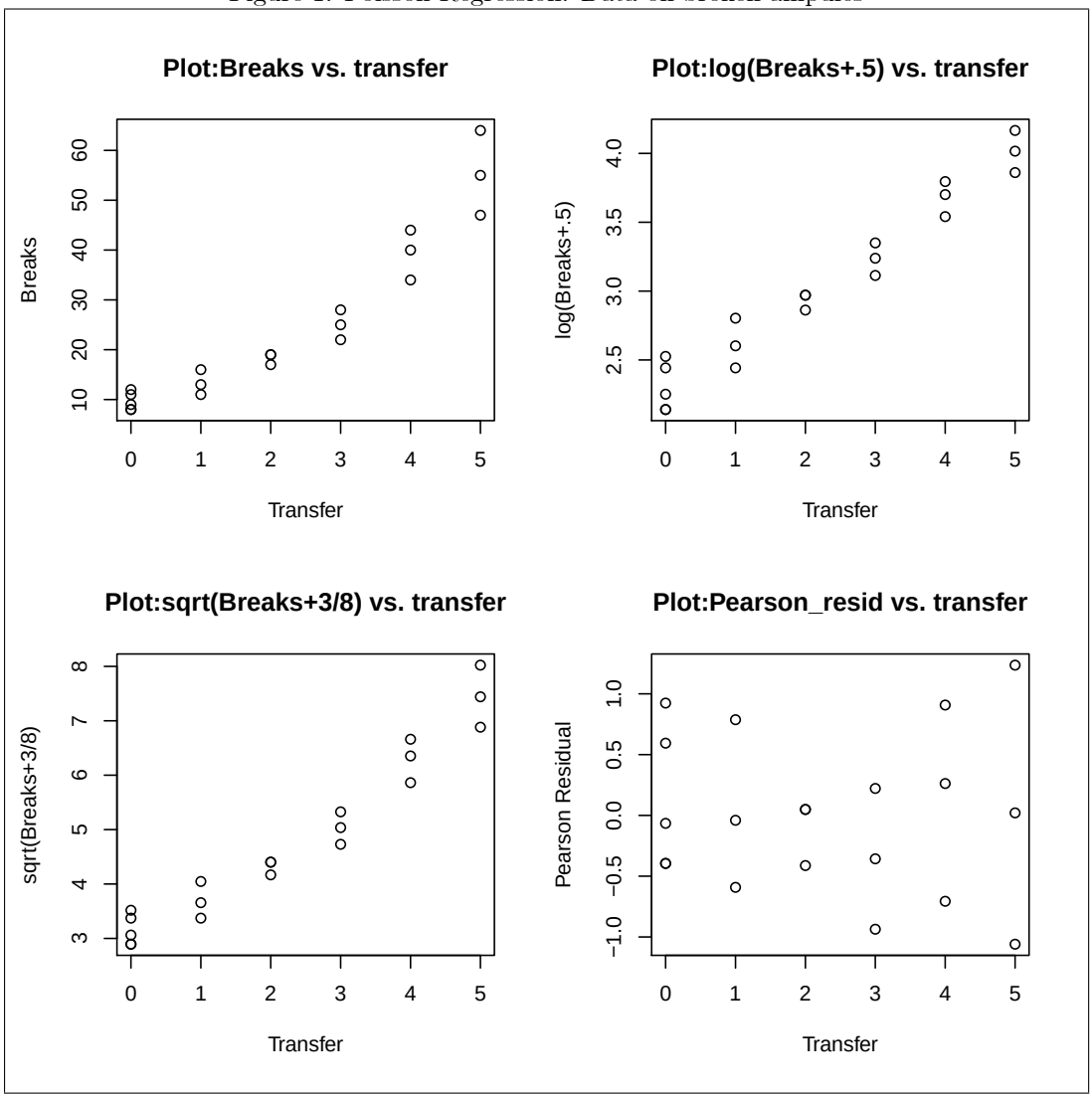

Figure 2: Poisson Regression: Skin Cancer Data