13. Poisson Regression Analysis

We have so far considered situations where the outcome variable is numeric and Normally distributed, or binary. In clinical work one often encounters situations where the outcome variable is numeric, but in the form of counts. Often it is a count of rare events such as the number of new cases of lung cancer occurring in a population over a certain period of time. The aim of regression analysis in such instances is to model the dependent variable Y as the estimate of outcome using some or all of the explanatory variables (in mathematical terminology estimating the outcome as a function of some explanatory variables.

When the response variable had a Normal distribution we found that its mean could be linked to a set of explanatory variables using a linear function like $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_k X_k$.

In the case of binary regression the fact that probability lies between 0-1 imposes a constraint. The normality assumption of multiple linear regression is lost, and so also is the assumption of constant variance. Without these assumptions the *F* and *t* tests have no basis. The solution was to use the logistic transformation of the probability p or logit p, such that $\log_e(p/1- p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots\ldots \beta_n X_n$. The β coefficients could now be interpreted as increasing or decreasing the log odds of an event, and expβ (the odds multiplier) could be used as the odds ratio for a unit increase or decrease in the explanatory variable. In survival analysis we used the natural logarithm of the hazard ratio, that is $\log_e h(t)/h_0 (t) = \beta_0 + \beta_1 X_1 + \ldots\ldots + \beta_n X_n$

When the response variable is in the form of a count we face a yet different constraint. Counts are all positive integers and for rare events the Poisson distribution (rather than the Normal) is more appropriate since the Poisson mean > 0. So the logarithm of the response variable is linked to a linear function of explanatory variables such that $\log_e (Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots$ etc. and so $Y = (e^{\beta 0}) (e^{\beta 1X1}) (e^{\beta 2X2})$ .. etc. In other words, the typical Poisson regression model expresses the log outcome rate as a linear function of a set of predictors.

Assumptions in Poisson Regression

The assumptions include:

1. Logarithm of the disease rate changes linearly with equal increment increases in the exposure variable.
2. Changes in the rate from combined effects of different exposures or risk factors are multiplicative.
3. At each level of the covariates the number of cases has variance equal to the mean.
4. Observations are independent.

Methods to identify violations of assumption (3) i.e. to determine whether variances are too large or too small include plots of residuals versus the mean at different levels of the predictor variable. Recall that in the case of normal linear regression, diagnostics of the model used plots of residuals against fits (fitted values). This means that the same diagnostics can be used in the case of Poisson Regression.

The examples below illustrate the use of Poisson Regression.

Example 1

Births by caesarean section are said to be more frequent in private (fee paying) hospitals as compared to non-fee paying public hospitals. Data about total annual births and the number of caesarean sections carried out were obtained from the records of 4 private hospitals and 16 public hospitals. These are tabulated below:

Births	Hospital type	Caesareans
236	0	8
739	1	16
970	1	15
2371	1	23
309	1	5
679	1	13
26	0	4
1272	1	19
3246	1	33
1904	1	19
357	1	10
1080	1	16
1027	1	22
28	0	2
2507	1	22
138	0	2
502	1	18
1501	1	21
2750	1	24
192	1	9

0 = Private Hospital 1 = Public Hospitals

The result of Poisson regression analysis is described below:

[In Genstat Stats → Regression Analysis → Generalized Linear → General Model in the Analysis field → Distribution → Poisson]

We first regress the response variable ‘Caesarean Sections’ against one explanatory variable viz. ‘number of births’; then add one more explanatory variable ‘number of obstetricians’, and finally add a third variable in the form of indicator variable for ‘public hospital’(i.e. Public hospital =1; Private Hospital =0) in the regression analysis.

```
***** Regression Analysis *****

 Response variate: CAESAREA
     Distribution: Poisson
    Link function: Log
     Fitted terms: Constant, BIRTHS


*** Summary of analysis ***

                                      mean    deviance
              d.f.     deviance     deviance     ratio
Regression       1 63.575488916  63.575488916     63.58
Residual        18 36.414789139   2.023043841
Total           19 99.990278055   5.262646213

Change          -1 -63.575488916 63.575488916     63.58

* MESSAGE: The following units have large standardized residuals:
                  13        2.25
                  14       -2.82
                  16       -2.94
                  17        2.17
* MESSAGE: The following units have high leverage:
                   9        0.41
                  19        0.21


*** Estimates of regression coefficients ***

                  estimate         s.e.      t(*)
Constant             2.132        0.102     20.95
BIRTHS           0.0004405    0.0000540      8.17
* MESSAGE: s.e.s are based on dispersion parameter with value 1
```

The regression equation may now be written as:
$\text{Log}_e(Y) = \beta_0 + \beta_1X_1$

On substituting the values of Y and X, the equation can be written as:
$\text{Log}_e(\text{Caesarean}) = 2.132 + 0.000441$ Births
Which leads to caesarean = $(e)^{2.132}$ x $(e)^{0.0004}$ Births

Recall that we have used poisson regression to model the data and thereby obtained estimates of caesarean sections based on just one explanatory variable viz. number of births. We next add a second term to the model, - type of hospital and obtain the following.

```
***** Regression Analysis *****

 Response variate: CAESAREA
     Distribution: Poisson
    Link function: Log
     Fitted terms: Constant, BIRTHS, HOSP_TYP

 *** Summary of analysis ***
```

```
                                        mean  deviance
              d.f.     deviance     deviance     ratio
Regression       2 81.951077606 40.975538803     40.98
Residual        17 18.039200449  1.061129438
Total           19 99.990278055  5.262646213

Change          -2 -81.951077606 40.975538803     40.98
* MESSAGE: ratios are based on dispersion parameter with value 1


* MESSAGE: The following units have large standardized residuals:
                   5       -2.48
* MESSAGE: The following units have high leverage:
                   9        0.41


*** Estimates of regression coefficients ***

                  estimate         s.e.      t(*)
Constant             1.351        0.249      5.43
BIRTHS           0.0003261    0.0000603      5.41
HOSP_TYP             1.045        0.272      3.84
   •  MESSAGE: s.e.s are based on dispersion parameter with value 1
```

This is the full model for which the regression equation may be written as:
$\text{Log}\ (Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

On substituting the values we have $\text{Log}_e\ (Y) = 1.351 + 0.00033$ Births + 1.045 Hosptype for Public Hospitals (For private hospitals X_2 is 0 and $\beta_2 X_2 = 0$).

This leads to caesarean sections = $(e)^{1.351}$ x $(e)^{0.00033}$births x $(e)^{1.045}$ Public Hospitals
Caesarean sections in Public Hospitals = 3.86 x 1 x 2.84 = 10.97 =approximately 11.
Caesarean sections in Private Hospitals = 3.86 x 1 = 3.86 = Approximately 4.

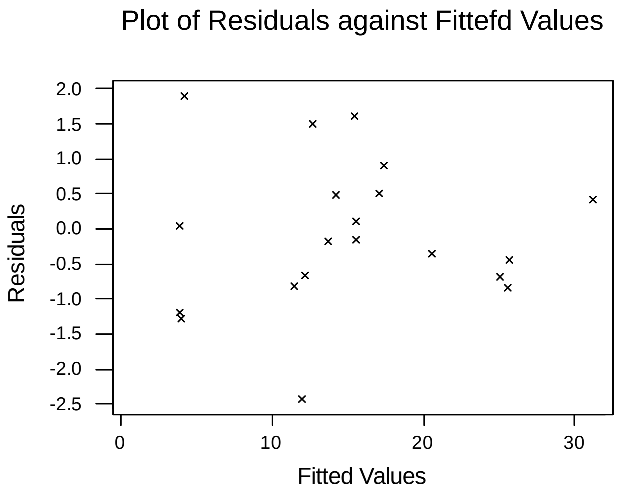
Conclusion

According to evidence presented all things being equal it is the other way round. Caesarean sections are about twice as common in Public Hospitals than in Private ones.

Of the two models presented which one gives best estimates? For the answer we look at the tables which give the values for deviance. Deviance serves the same purpose as sum of squares in multiple linear regression. An important use of deviance, and difference between deviances, is in the comparison of fits of two models when additional explanatory variable(s) get added to the initial simple model.

In the case of the first analysis using just one explanatory variable, the deviance explained by the regression is 63.575. This changes to 81.95 when a second variable, Hospital Type is added to the model. The difference is 18.375 (81.95 – 63.575 = 18.375) and the difference in degree of freedom is 1. Looking this up in the table of χ^2 at 1 degree of freedom, we find that the result is significant ($P < .001$). We can opt for this model.

The next step is to check for the fit of the model by carrying out diagnostic plots of deviance against fits



Second Example to illustrate Poisson Regression Analysis

A cohort of subjects, some non-smokers and others smokers, was observed for several years. The number of cases of cancer of the lung diagnosed among the different categories was recorded. Data regarding the number of years of smoking were also obtained from each individual. For each category the person-years of observation were calculated. The investigators wish to address the question of the relative risks of smoking.

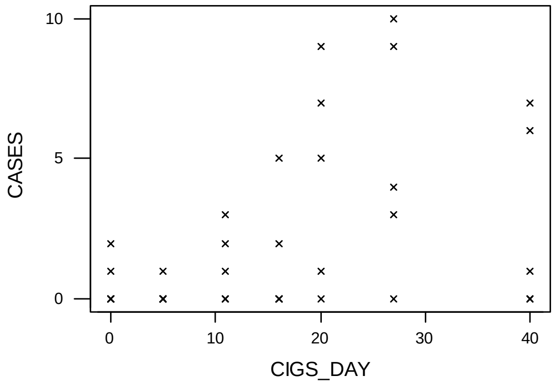
The following records were kept.

CIGS	No.	Person	
Per	years	years	
Day	smoking		CASES
0	15	10366	1
0	25	5969	0
0	35	3512	0
0	45	1421	0
0	55	826	2
5	15	3121	0
5	25	2288	0
5	35	1648	1
5	45	927	0
5	55	606	0
11	15	3577	0
11	25	2546	1
11	35	1826	0
11	45	988	2
11	55	449	3
16	15	4317	0
16	25	3185	0
16	35	1893	0
16	45	849	2
16	55	280	5

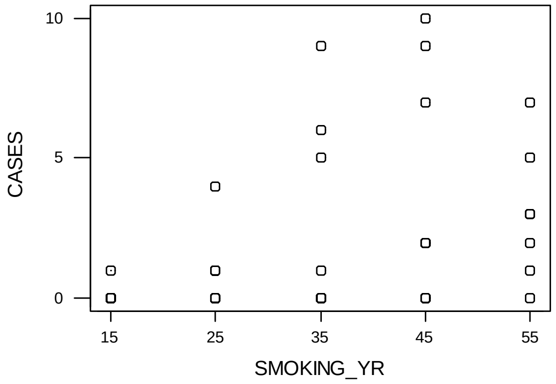
```
20      15      5683    0
20      25      5483    1
20      35      3646    5
20      45      1567    9
20      55      416     7
27      15      3042    0
27      25      4290    4
27      35      3529    9
27      45      1409    10
27      55      284     3
40      15      670     0
40      25      1482    0
40      35      1336    6
40      45      556     7
40      55      104     1
```

In the above data set the average number of cigarettes smoked per day represents the daily dose, and the years of smoking together with the average number of cigarettes smoked daily represents the total dose inhaled over time. Both appear to be related to the outcome as illustrated in the charts below.

Cases of Cancer of the Lung by average number of cigarettes smoked



Cases of Cancer of the Lung by number of years of smoking



Since over a number of years of observation some cases of cancer of the lung can be expected to arise from causes not related to smoking, we use this as our base line (uninformative) model and perform the first analysis as below

***** Regression Analysis *****

```
 Response variate: CASES
     Distribution: Poisson
    Link function: Log
     Fitted terms: Constant, PERSONYR


*** Summary of analysis ***

                                        mean  deviance
              d.f.     deviance     deviance     ratio
Regression       1  8.743853032  8.743853032      8.74
Residual        33 128.546991211  3.895363370
Total           34 137.290844242  4.037966007

Change          -1 -8.743853032  8.743853032      8.74
* MESSAGE: ratios are based on dispersion parameter with value 1




*** Estimates of regression coefficients ***

                  estimate         s.e.      t(*)
Constant             1.208        0.169      7.16
PERSONYR        -0.0001921    0.0000711     -2.70
* MESSAGE: s.e.s are based on dispersion parameter with value 1
```

We next perform the second analysis using the full model as below:

```
***** Regression Analysis *****

 Response variate: CASES
     Distribution: Poisson
    Link function: Log
     Fitted terms: Constant, PERSONYR, CIGS_DAY, SMOKING_


*** Summary of analysis ***

                                        mean  deviance
              d.f.     deviance     deviance     ratio
Regression       3 63.168816931 21.056272310     21.06
Residual        31 74.122027311  2.391033139
Total           34 137.290844242  4.037966007

Change          -3 -63.168816931 21.056272310     21.06
* MESSAGE: ratios are based on dispersion parameter with value 1



*** Estimates of regression coefficients ***

                  estimate         s.e.      t(*)
Constant            -4.669        0.988     -4.72
PERSONYR          0.000410     0.000104      3.94
CIGS_DAY            0.0559       0.0100      5.58
SMOKING_            0.0888       0.0166      5.34
•   MESSAGE: s.e.s are based on dispersion parameter with value 1
```

The difference in the deviance is 63.168 – 8.743853 = 54.425 at 3 – 1 = 2 degrees of freedom. Entering the table for values of χ^2 at 2 degrees of freedom, we get a highly significant *P* value for 54.425. This indicates a good fit for the model.

We accept the model and obtain a regression equation which could be written as

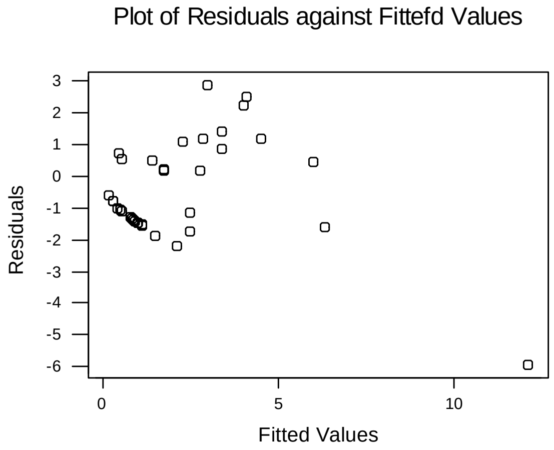
$$\text{Log cases} = \alpha + \beta_1\text{Personyears} + \beta_2\text{Cigarettes/day} + \beta_3\text{Years of Smoking}$$
$$= -4.669 + 0.00041\text{Personyears} + 0.0559\text{Cigarettes/day} + 0.0888\text{Years of Smoking}$$

$$\text{Cases} = (e)^{-4.669} \times (e)^{\mathbf{0.00041personyears}} \times (e)^{\mathbf{0.0559Cigarettes/day}} \times (e)^{\mathbf{0.0888years\ of\ smoking}}$$

The equation is useful for estimating the relative risk of developing lung cancer by the number of cigarettes smoked (i.e. strength of the dose), or by number of years of smoking (total dose).

For example, all things being equal the relative risk of smoking 25 cigarettes/day as compared to 15 can be estimated at $(e)^{0.0559x25} \div (e)^{0.0559x15} = (e)^{0.0559x10}$, and so on. Similar estimates could be used for calculating the relative risk by different years of smoking.

A plot of residuals against fitted values is shown below:



Summary

The typical Poisson regression model expresses the natural logarithm of the event or outcome of interest as a linear function of a set of predictors. The dependent variable is a count of the occurrences of interest e.g. the number of cases of a disease that occur over a period of follow-up. Typically, one can estimate a rate ratio associated with a given predictor or exposure.

A measure of the goodness of fit of the Poisson regression model is obtained by using the deviance statistic of a base-line model against a fuller model.