FORMULA DRIVEN POISSON REGRESSION ANALYSIS IN EXCEL

William P . Fox
Department of Defense Analysis
Naval Postgraduate School

Abstract

In one of our elective courses, Dark Networks, the student must use Poisson Regression in their analysis. To the students this is a black box routine for which they do not understand the relationships that exist among the inputs or the outputs. We added a block on regression analysis to our statistical modeling for decision making to help students better understand these relationships across many regression techniques. Furthermore, we did it in Excel which is the software these students will have back in their professions in the real world. We provide two examples with their solutions from the literature to show Poisson regression in Excel. Additionally, we present a pattern recognition method for the Hessian matrix to find the Variance–Covariance matrix in Poisson regression which is used to obtain the coefficient's standard errors.

Disclaimer

"The views expressed in this document are those of the author and do not reflect the official policy or position of the Department of Defense or U.S. Government."

Introduction

We support our interdisciplinary department of Defense Analysis at the Naval Postgraduate School by teaching a three course sequence in mathematical modeling. One of our carry through topics in mathematical modeling is *regression*. Our students initially encountered data where the outcome variable is numeric and normally distributed allowing them to use simple least squares techniques. In the first course in modeling the students had handled linear regression, polynomial regression, and even multivariable regression. In the second modeling course, we teach advanced regression techniques and forecasting as well as nonlinear regression for dealing with oscillating data. Each application is performed in Excel because this will be the software that our students will have after graduation.

In our Common Operational Research Environment (CORE) Lab work our students encounter situations where the outcome variable is not only numeric but also in the form of discrete counts. Often, it is a count of rare events such as the number of new cases of terrorist activities occurring within a population over a certain period of time, the number of certain types of IEDS encountered, or the number of significant acts of violence within a region. The goal 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 explanatory or predictor variables). Although in the CORE lab much of this will be a Black-Box, it is necessary to understand the "output" values to better understand the model itself. It is more important that the students have a tool to reproduce their analysis in a software package that will be available to them in the future, such as Excel.

According to Devore [1] simple linear regression is defined as follows. "There exists parameters B_0, B_1, and σ^2 such that for any fixed input value of x, the dependent variable is related to x through the model equation, $Y = \beta_0 + \beta_1 X_1 + \varepsilon$. The quantity ε in the model equation is a random variable assumed to be normally distributed with mean =0 and variance = σ^2. We expand this definition to when the response variable, y_i, is assumed to have a normal distribution with mean, μ_y, and standard

deviation, σ, we found that the mean could be modeled as a function of our multiple predictor variables $\{X_1,X_2,\ldots,X_n\}$ using the linear function $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 \ldots\ldots + \beta_k X_k$. We have used linear models for bivariate data such as $y = a + bx$ or $y=a+bx+cx^2$ and used models such as $y=a_0+a_1x_1+a_2x_2+\ldots+a_nx_n$ when we have n different independent predictor variables. The key assumptions that we used for least squares are the linearity of the relationship between dependent and independent variables, independence and normality of the errors, and homoscedasticity (constant variance) of the errors. If any of these assumptions is violated then the adequacy of the model is diminished. In our first courses, we show the use of residual plots to give the students information about *the adequacy of the model* depending on the patterns seen or not seen in the residual plot. Our analysis and check for these assumptions generally concern examining the residual plot, a plot of the errors versus the model values, for patterns or no patterns [2, 3].

Normality Assumption Lost

According to Neter [4] and Montgomery [5], in the case of logistic and Poisson regression, the fact that probability lies between 0-1 imposes a constraint. We lose both the normality assumption of multiple linear regression and the assumption of constant variance. Without these assumptions the *F* and *t* tests have no basis for the analysis. When this happens, we must transform the model and the data. The new solution involves using the logistic transformation of the probability *p* or logit **p**, such that

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1X_1 + \beta_2X_2\ldots\ldots + \beta_nX_n.$$

They go on to explain that the β coefficients could now be interpreted as increasing or decreasing the log odds of an event, and exp(β) (the odds multiplier) could be used as the odds ratio for a unit increase or decrease in the explanatory variable [4,5].

When the response variable is in the form of a ***count*** we face a yet different constraint. Counts are all positive integers and stand for rare events. Thus, the Poisson distribution (rather than the Normal distribution) is more appropriate since the Poisson has a mean greater than 0 and our counts are all positive counting numbers. So the logarithm of the response variable is linked to a linear function of explanatory variables such that

$$ln\,(Y) = \beta_0 + \beta_1x_1 + \beta_2x_2 \ldots + \beta_nx_n$$

and thus,

$$Y = (e^{B_0})(e^{B_1x_1})(e^{B_2x_2}) \ldots (e^{B_nx_n}) .$$

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

There are several key assumptions in Poisson regression that are different than the assumptions in the simple linear regression model. These include that the logarithm of the dependent variable changes linearly with equal incremental increases in the exposure variable. For example, if we measure risk in exposure per unit time and one group is counts per month and another is count per years we can convert all exposures to strictly counts. We find that changes in the rate from combined effects of different exposures or risk factors are multiplicative. We find for each level of the covariates, the number of cases has variance equal to the mean which makes it follow a Poisson distribution. Further, we assume the observations are independent.

We use diagnostic methods to identify violations of the assumption to determine whether variances are too large or too small including 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 implies that

some of the same diagnostics can be used in the case of Poisson Regression. We will use the residual or deviation plot, deviations versus the model to look for patterns as our main diagnostic method.

In Poisson regression we start with the basic model shown in equation *(1)*,

$$Y_i = E[Y_i] + \varepsilon_i \text{ for } i = 1, 2, \ldots, n. \quad (1)$$

The $_i^{th}$ case mean response is denoted by u_i, where u_i can be one of many defined functions[4] but we elect to use only the form shown in equation *(2)*,

$$u_i = u(\boldsymbol{X_i, B}) = exp(\boldsymbol{X'_i B}) \text{ where } u_i \geq 0. \quad (2)$$

We assume that the variable, Y_i, are independent Poisson random variables with expected value u_i.

In order to apply regression techniques, we will use the likelihood function [4, 5]. The likelihood function, *L*, is given in equation *(3)*.

$$L = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \frac{(u(X_i,B)^{Y_i} \exp[-u(X_i,B)]}{Y_i!} \quad (3)$$

Most texts explain that maximizing this function is quite difficult, so they use the logarithm of the likelihood function shown in equation (4):

$$\ln(L) = \sum_{i=1}^{n} Y_i \ln(u_i) - \sum_{i=1}^{n} u_i - \sum_{i=1}^{n} \ln(Y_i!) \quad (4)$$

where u_i is the fitted model.

We maximize this function to obtain the best estimates for the coefficients of the model. Numerical search techniques are used to obtain these estimates. We mention here that "good" starting points are required to possibly obtain convergence [6].

Within the model development we are concerned about the deviations or residuals as we previously mentioned. In Poisson regression, the deviance is modeled as shown in equation *(5)*:

$$Dev = 2\left[\sum_{i=1}^{n} Y_i ln\left(\frac{Y_i}{u_i}\right) - \sum_{i=1}^{n}(Y_i - u_i)\right] \quad (5)$$

where u_i is the fitted model. We note that because of term $\ln(Y_i/u_i)$ that if $Y_i=0$ we must set the $\ln(Y_i/u_i)=0$.

Inferences for the coefficients are carried out in the same fashion as with logistics regression. To estimate the variance-covariance matrix we require the use of the Hessian matrix. We define the Hessian, $H(\boldsymbol{X})$, as the matrix of second partial derivatives of the $ln(L)$ function. The variance-covariance matrix, $VC(\boldsymbol{X},\boldsymbol{B})$, is minus the inverse of this Hessian matrix evaluated with the final estimates of the coefficients, $\boldsymbol{B}$.

$$VC(\boldsymbol{X},\boldsymbol{B}) = - H(\boldsymbol{X})^{-1}$$

The main diagonal of the matrix are the estimates for the variance. Since we need the estimated standard deviations, se_B, we take the square root of each main diagonal entry to obtain this estimate. We may then perform hypothesis tests of the coefficients using the *t*-test.

We use the logarithm of the likelihood function, equation *(4)*. The Hessian is defined as the matrix of second partial derivatives. We will illustrate two Hessian modeling examples and then we will make a useful observation.

Assume that our model is, $y_i=exp(b_0+b_1*x_i)$. Putting this model into equation *(4)* we have,

$$\ln(L) = \sum_{i=1}^{n} Y_i \ln(\exp(b_0 + b_1 x_i)) - \sum_{i=1}^{n} \exp(b_0 + b_1 x_i) - \sum_{i=1}^{n} Y_i!$$

We define the second partial derivatives as follows in equation *(6)*:

$$g_{ij} = \frac{\partial^2(\ln(L))}{\partial b_i \partial b_{ij}} \text{ for all } i \text{ and } j. \quad (6)$$

The estimates for the variance-covariance matrix are defined and are displayed in equation *(7):*

$$s^2(\boldsymbol{b}) = [(-g_{ij})_{\boldsymbol{B}=\boldsymbol{b}}]^{-1} \qquad (7)$$

We take these partial derivatives and set up the Hessian matrix, g_{ij} as shown in the matrix below*:*

$$g_{ij} = \begin{bmatrix} -\left(\sum_{i=1}^{n} e^{b_0+b_{1x_i}}\right) & -\left(\sum_{i=1}^{n} x_i e^{b_0+b_{1x_i}}\right) \\ -\left(\sum_{i=1}^{n} x_i e^{b_0+b_{1x_i}}\right) & -\left(\sum_{i=1}^{n} x_i^2 e^{b_0+b_{1x_i}}\right) \end{bmatrix}$$

When our model slightly differs, such as $y_i=exp(b_0+b_1x_{1i}+b_2x_{2i})$, then we find the Hessian matrix, g_{ij}. We note the similarities between the last two Hessian matrices.

$$g_{ij} = \begin{bmatrix} -\left(\sum_{i=1}^{n} e^{b_0+b_{1x_{1i}}+b_2x_{2i}}\right) & -\left(\sum_{i=1}^{n} x_{1i} e^{b_0+b_{1x_{1i}}+b_2x_{2i}}\right) & -\left(\sum_{i=1}^{n} x_{2i} e^{b_0+b_{1x_{1i}}+b_2x_{2i}}\right) \\ -\left(\sum_{i=1}^{n} x_{1i} e^{b_0+b_{1x_{1i}}+b_2x_{2i}}\right) & -\left(\sum_{i=1}^{n} x_{1i}^2 e^{b_0+b_{1x_{1i}}+b_2x_{2i}}\right) & -\left(\sum_{i=1}^{n} x_{1i}x_{2i} e^{b_0+b_{1x_{1i}}+b_2x_{2i}}\right) \\ -\left(\sum_{i=1}^{n} x_{2i} e^{b_0+b_{1x_{1i}}+b_2x_{2i}}\right) & -\left(\sum_{i=1}^{n} x_{1i}x_{2i} e^{b_0+b_{1x_{1i}}+b_2x_{2i}}\right) & -\left(\sum_{i=1}^{n} x_{2i}^2 e^{b_0+b_{1x_{1i}}+b_2x_{2i}}\right) \end{bmatrix}$$

We see the pattern in the matrix of partial derivatives and we can extend the pattern to easily obtain the Hessian for a model when we have *n* independent variables, $y_i=exp(b_0+b_1x_{1i}+b_2x_{2i}+\ldots+b_nx_{ni})$ and we identify the common term in the matrix as the summation, $\Sigma\ exp(b_0+b_1x_{1i}+b_2x_{2i}+\ldots+b_nx_{ni})$. We call this summation *P*. This gives us a generic Hessian matrix for Poisson regression to use with our choice of the model from $y_i=exp(b_0+b_1x_{1i}+b_2x_{2i}+\ldots+b_nx_{ni})$, depending on the number of independent variables.

$$g_{ij} = -\begin{bmatrix} P & \sum x_{1i}P & \sum x_{2i}P & \sum x_{3i}P & \dots & \sum x_{ni}P \\ \sum x_{1i}P & \sum x_{1i}^2P & \sum x_{1i}x_{2i}P & \sum x_{1i}x_{3i}P & \dots & \sum x_{1i}x_{ni}P \\ \sum x_{2i}P & \sum x_{1i}x_{2i}P & \sum x_{2i}^2P & \sum x_{2i}x_{3i}P & \dots & \sum x_{2i}x_{ni}P \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \sum x_{ni}P & \sum x_{1i}x_{ni}P & \sum x_{2i}x_{ni}P & \sum x_{3i}x_{ni}P & \dots & \sum x_{ni}^2P \end{bmatrix}$$

This is the generic Hessian matrix, so we need to replace the formulas with numerical values and compute the inverse of the negative of this matrix. Once we replace the variables with their respective values we should have a non-singular square matrix that we can take the inverse of. The main diagonal entries of this matrix inverse are the estimates for the variances of the coefficients to the estimates of ***b***. The square root of the entries of the main diagonal are the estimates of the *se* of the coefficients of ***b*** to be used in the hypothesis testing for each coefficient, ***b*** *as*

$$t^*=bi/se(b_i)$$

where *se* is the standard error associated with b_i.

We now have all the equations that we need to build the tables of outputs for Poisson regression that are similar to Excel's prepackaged regression outputs.

Estimates of Regression Coefficients

We use one for the constant plus one for every predictor variable in the model being examined for the number of coefficients. Estimates are the final values (that converged) for the numerical search method to maximize the *ln(L)* equation. The values of *se* are the square roots of the main diagonal of the inverse of *(-)* the Hessian matrix. The values of *t*= (final coefficient estimate)/se* and the *p*-value are displayed, where the *p*-value is the probability associated with the $|t^*|$ from $P(T>|t^*|)$. In our summary of Poisson regression analysis, let *m* = number of variables in the model, let *k* = number of data elements of the dependent variable, *Y*. We present the statistical formulas.

	Degrees of freedom (df)	Deviance	Mean deviance, *MDev*	Ratio
Regression	*m*	$D_{reg}=D_t-D_{res}$	$MDev(reg)=D_{reg}/m$	$\|MDev(reg)\|$
Residual	*k-1-m*	D_{res}=Result from equation *(5)* using the full model with *m* predictors.	*MDEV(res)= Dres/(k-1-m)*	
Total	*k-1*	D_t=Result from equation *(5)* using only $y=exp(b_o)$ as the best model	$MDev(t)= D_t/(k-1)$	

Illustrative Examples

The first example will be explained in more detail than the second example, for illustrative purposes, to show how we used the equations and Excel to perform Poisson Regression. We note that a prerequisite for using Poisson regression is that data for the dependent variable, *Y*, must be discrete counts data with large numbers a rare event. We have chosen two data sets [7] that have published solutions in the literature to be our examples.

Example 1: Caesarian Births

The data is defined as follows:

Csec= number of C-sections performed

Hosp= type of hospital public or private, coded as (*0*-public or *1*-private)

Birth= number of births at the hospital

	Csec	Hosp	Birth		Csec	Hosp	Birth
1	8	0	236	11	10	1	357
2	16	1	739	12	16	1	1080
3	15	1	970	13	22	1	1027
4	23	1	2371	14	2	0	28
5	5	1	309	15	22	1	2507
6	13	1	679	16	2	0	138
7	4	0	26	17	18	1	502
8	19	1	1272	18	21	1	1501
9	33	1	3246	19	24	1	2750
10	19	1	1904	20	9	1	192

This data was obtained through the record at 4 private hospitals and 16 public hospitals. We desired to build a model to predict the number of *c-section* births as a function of the type of hospital and number of births. Since the *y* variable represents discrete counts of *C-section* births, *Csec*, with large numbers of *Csec* births being rare we should use Poisson regression.

We list the steps required to obtain a regression model using Excel and Poisson regression.

Step 1. Calculate the baseline constant model, *y=exp(constant),* using the Solver to obtain the value of the constant that minimizes the deviations, equation *(5)*. We minimize equation *(5)* by varying the value cell of the constant.

Step 2. Repeat Step 1 for the full model, $y=exp(b_0+b_1x_1+b_2x_2+….+b_nx_n)$, using the Solver to obtain the values of the parameters $\{b_0, b_1, b_2, …., b_n)$ that minimize the deviations using equation *(5)*.

Step 3. Using the pattern recognition format as explained for the Hessian concerning equation (7), compute the standard errors, *se*, for the parameter estimates.

Step 4. Compute the individual deviations and plot the indexed deviations.

Step 5. Compute the *p*-values as appropriate.

Step 6. Put all values into tables.

Step 7. Plot the deviation.

We illustrate these steps using Excel.

Step 1:We have entered the data so first we calculate the information for the constant model *y=exp(constant).*

For the constant model, $y = e^{b_0}$, we initially set the decision variable for the solver, b_0, to 0. The final result from the solver is that b_0=2.711377991 and the constant model is $y = e^{2.711377991}$. Our objective function value when b_0=2.711377991 is 99.99028.

Step 2. We repeat using the solver for the full model, $y=exp(b_0+b_1*hosp+b_2*birth)$. We initially set the coefficients, decision variables b_0,b_1,b_2 to *0*. Our final coefficients are $b_0=1.350998944$, $b_1 = 1.0045138972$, and $b_2=0.00032607$. Our final objectives function value is 18.0392.

Model 1	y*ln(exp(M1)	ln(Fact(Y))		Deviation equation		
15.05	21.69102393	10.6046029		-5.05549	-7.05	
15.05	43.38204786	30.6718601		0.979372	0.950000002	
15.05	40.67066987	27.8992714		-0.04992	-0.05	
15.05	62.36169379	51.6066756		9.754673	7.950000002	
15.05	13.55688996	4.78749174		-5.5097	-10.05	
15.05	35.24791388	22.5521639		-1.90357	-2.05	
15.05	10.84551196	3.17805383		-5.30033	-11.05	
15.05	51.51618183	39.3398842		4.428159	3.950000002	
15.05	89.47547371	85.054467		25.90928	17.95	
15.05	51.51618183	39.3398842		4.428159	3.950000002	
15.05	27.11377991	15.1044126		-4.08793	-5.05	
15.05	43.38204786	30.6718601		0.979372	0.950000002	
15.05	59.6503158	48.4711814		8.352618	6.950000002	
15.05	5.422755982	0.69314718		-4.03646	-13.05	
15.05	59.6503158	48.4711814		8.352618	6.950000002	
15.05	5.422755982	0.69314718		-4.03646	-13.05	
15.05	48.80480384	36.3954452		3.221888	2.950000002	
15.05	56.93893781	45.3801389		6.996033	5.950000002	
15.05	65.07307179	54.7847294		11.20022	8.950000002	
15.05	24.40240192	12.8018275		-4.62738	-6.05	
301	816.1247753	608.501426	Sums	49.99514	3.70058E-08	
OBJ Func	-93.37665016					
				99.99028	Total Deivations	

Model 2	y*ln(exp(m2))	ln(fact(Y))		Dev Eq	y-u
4.17015	11.42361596	10.6046		5.211916	3.82985
13.9727	42.19368496	30.67186		2.167735	2.027302
15.06581	40.68642	27.89927		-0.06567	-0.06581
23.78976	72.89287012	51.60668		-0.7765	-0.78976
12.14472	12.48447121	4.787492		-4.43728	-7.14472
13.70199	34.02803267	22.55216		-0.68369	-0.70199
3.894155	5.437907292	3.178054		0.10727	0.105845
16.6249	53.40713457	39.33988		2.537206	2.375096
31.64463	114.0007586	85.05447		1.383991	1.355371
20.42949	57.32261015	39.33988		-1.37827	-1.42949
12.33629	25.1254571	15.10441		-2.09961	-2.33629
15.616	43.97273516	30.67186		0.388684	0.384001
15.34844	60.0823106	48.47118		7.920623	6.651556
3.896696	2.720257935	0.693147		-1.33396	-1.8967
24.86848	70.69922307	48.47118		-2.69629	-2.86848
4.038999	2.79199383	0.693147		-1.4057	-2.039
12.93357	46.07687136	36.39545		5.94982	5.066431
17.91382	60.59701966	45.38014		3.337952	3.086182
26.91911	79.02807853	54.78473		-2.75479	-2.91911
11.69012	22.1286941	12.80183		-2.35367	-2.69012
300.9998	857.1001469	608.5014		9.019768	0.000167
OBJ Func	-52.40111134			Dev due to residuals	
				18.0392	

Step 3. We built the Hessian matrix from knowing the pattern and then we take the inverse of (-) the Hessian matrix. After we take the inverse we take the square root of the values along the main diagonal as the standard errors.

(-)Hessian		
300.9998327	284.9998324	467947.3576
284.9998324	284.9998324	466195.4648
467947.3576	466195.4648	1037620324
Inverse		
0.06254362	-0.06189194	-3.98392E-07
-0.06189194	0.074484805	-5.55339E-06
-3.98392E-07	-5.55339E-06	3.63851E-09

We take the square root of the values along the main diagonal to obtain the standard errors of the estimates,

$$se(b_0) = \sqrt{0.06254362} = 0.250087225$$
$$se(b_1) = \sqrt{0.074484805} = 0.272919045$$
$$se(b_2) = \sqrt{3.63851E-09} = 6.03291E-05$$

Step 4. We compute the deviations and obtain a plot, as shown in Figure 1 below. We interpret the plot as we examine the plot for patterns or no patterns.

Step 5 & 6. We place the information into the tables noting that there are 20 data elements in *Y*, the full model has two variables plus a constant. Where appropriate we computed the *p*-values.

To compute the p-values, we use *t.dist.2t(x,df)*. We find that all the coefficients are significant at a 0.05 level. We accept the full model to use for predictive analysis. For our example, our model for the number of *C-section births* is

$$Y = e^{(1.350993605 + 1.045142687*hosp + 0.000326073*births)}$$

Step 7. We examine the deviation plot, Figure 1, from the deviations versus the model. We look for one of these patterns: linear, curved, fanning (in or out) or for randomness of the plot [2, 3]. We see no pattern and accept the full model as adequate.

Analysis of Deviance

	Degrees of Freedom (*df*)	Deviance	Mean deviance	Ratio
Regression	2 (two variables in model-hosp, births)	(99.9902-18.0392)=81.951077	81.951077/2=40.9755	|40.9755|= 40.98
Residual	17 (Y-1-2)	18.0392	18.0392/17=1.06112	
Total	19 (Y-1)	99.9902	99.9902/19=5/26264	

Analysis of Regression Coefficients

	Df	Coefficient	*se*	t*	*p*-values	Significant
b_0	2	1.351	0.2505367	5.3924	0.01	Yes
b_1	17	1.0451	0.2617086	3.9934	0.00047	Yes
b_2	19	0.000326	0.00000604	5.397	0.0000165	Yes

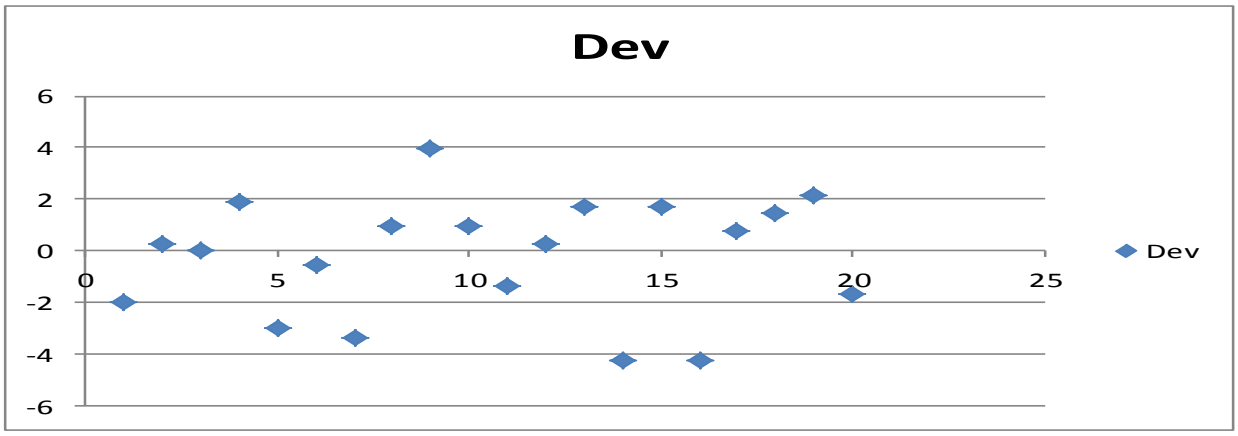

Figure 1. Deviation plot of deviations versus model from *C-sec* example.

Since the model is adequate, we use the model for predictions and interpolations. If we know we have a private hospital with 363 births then our estimate for the number of caesarians births is

$Y=e^{(1.350993605+1.045142687*+0.000326073*363)}=12.36$

or approximately *12*.

Checking our Analysis Decisions

We can always check our analysis to see if a smaller model would be more adequate. We could also build an intermediate model:

$Y=e^{(bo+hosp*b1)}$

We would use all the same equations as before and we could have built the following two tables:

Analysis of Deviance

	df	Deviance	Mean Deviance	Ratio
Regression	1	63.575	63.575	63.575
Residual	18	36.414	2.023	
Total	19	99.9902	5.263	

Analysis of Regression Coefficients

	Coefficient	Estimated SE	t*	Significant at 0.05
b_0	2.132	0.102	20.95	Yes
b_1	0.0004405	.0000540	8.17	Yes

How do we know which equation to use:

$Y=e^{(b0+b1*hosp)}$ or

$Y=e^{(b0+b1*hosp+b2*births)}$

Deviance serves the purpose of comparing models. Model I, the one variable predictor model, has a regression deviance of 63.575 (or about 64% is explained by the model) where Model II, the full model, has a regression deviance of 81.95 (or about 82% is explained by the model). This is a difference of 18.375 with a change of *df* of 1. We can use a χ^2 at 1 degree of

freedom to find it is significant at beyond 0.005. Thus, the model with more *df* and smaller residual deviance is better.

Example 2. Issues when some of the Y values are zero.

This example was chosen because it illustrates what is required when any *Y* values are equal to 0.

A cohort of subjects, some non-smokers and others smokers, was observed for several years. The number of cases of lung cancer diagnosed among the different categories was recorded. Data regarding the number of years of smoking were obtained from each individual. For each category the person-years of observation were calculated. We desire to build a mathematical model that predicts the cases of lung cancer as a function of these other variables.

We define the variables as follows:

Day= average number of cigarettes smoked per day.

YS=number of years smoking

Person Yr= number of total person years observed

Cases=number of cases of lung cancer observed.

The following data records were taken directly from the literature.

#	Day	YS	Person Year	Cases	#	Day	YS	Person Year	Cases
1	0	15	10366	1	19	16	45	1893	2
2	0	25	5969	0	20	16	55	280	5
3	0	35	3512	0	21	20	15	5683	0
4	0	45	1421	0	22	20	25	5483	1
5	0	55	826	2	23	20	35	3646	5
6	5	15	3121	0	24	20	45	1567	9
7	5	25	2288	0	25	20	55	416	7
8	5	35	1648	1	26	27	15	3042	0
9	5	45	927	0	27	27	25	4290	4
10	5	55	606	0	28	27	35	3529	9
11	11	15	3577	0	29	27	45	1409	10
12	11	25	2546	1	30	27	55	284	3
13	11	35	1826	0	31	40	15	670	0
14	11	45	988	2	32	40	25	1482	0
15	11	55	449	3	33	40	35	1336	6
16	16	15	4317	0	34	40	45	556	7
17	16	25	3185	0	35	40	55	104	1
18	16	35	849	0					

We build the simple model, *y=exp(constant)* to obtain the total deviance. The total deviance is found to be 137.291. Next, we will build the full model.

$$Cases = e^{(b0+b1*Day+b2*YS+b3*Person\ Year))}$$

Using Excel we can build the models and obtain the following outputs:

Analysis of Deviance

	df	deviance	Mean deviance	Ratio
Regression	3	63.169	21.056	21.06
Residual	31	74.122	2.391	
Total	34	137.291	4.037	

Analysis of Regression Coefficients

	Coefficients	Estimated SE	*t**	Significant at 0.05
Constant	-4.669	0.988	-4.72	Yes
Day	0.0559	0.01	5.58	Yes
YS	0.0888	0.0166	5.34	Yes
Person Yr	0.000410	0.000104	3.94	Yes

Our best one term model was found to be $y=e^{(b0+b1*Person\ Year)}$ and we display the following tables for later comparison:

Analysis of Deviance

	df	Deviance	Mean deviance	Ratio
Regression	1	8.744	8.744	8.74
Residual	33	128.546	3.895	
Total	34	137.291	4.037	

Analysis of Regression Coefficients

	Coefficients	Estimated *se*	t*	Significant at 0.05
Constant	1.208	0.169	7.16	Yes
Person Yr	-0.0001921	0.0000711	-2.70	No

We compare the models to see the difference in regression deviance is 63.168-8.744=54.415 while *df* changed from 3-1=2. We find the χ^2 value of 54.415 at 2 *df* and it is highly significant. We obtain the deviation plot, Figure 2, for the deviances versus the fitted model. We do not see any pattern. Thus, we accept our model as adequate.

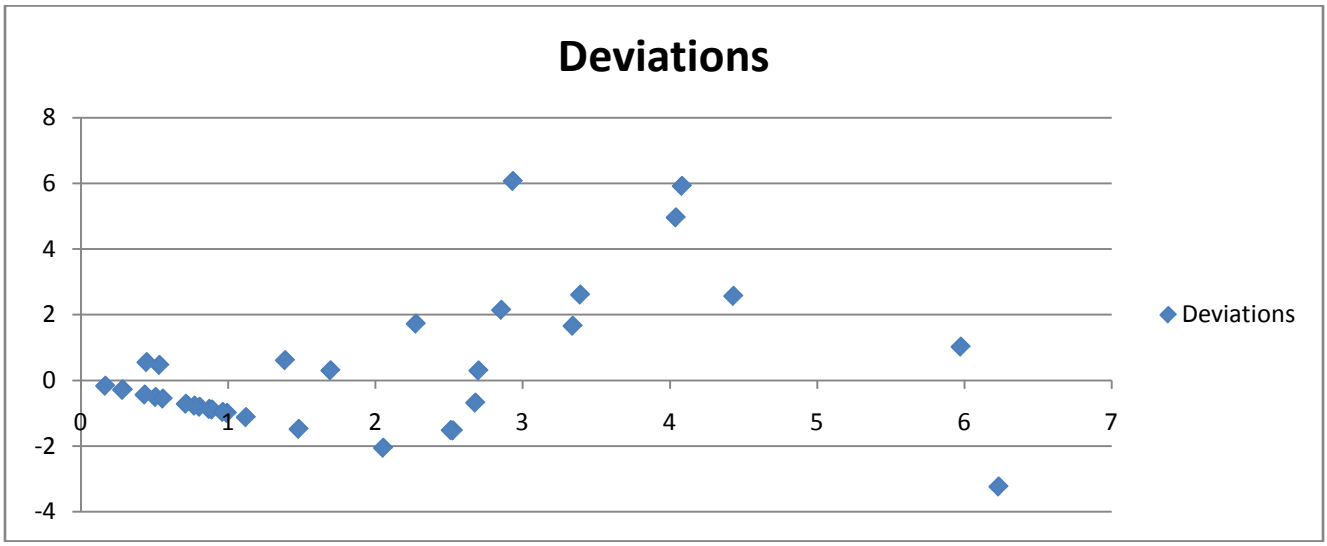

Figure 2. Deviation plot of deviations versus model in the full model from example 2.

Conclusions

Once adequate models are obtained, we may use these models to predict or interpolate information. A measure of the goodness of fit is obtained by using the deviance statistic of a baseline line against the fuller model. We were surprised by the number of statistical packages that do not have Poisson regression as an option. Most textbooks that mention Poisson regression only mention the basic model and then show output from a statistical program such as SAS or R. Our students want to know "what, why, and how". The black box does not provide all these answers initially. Having gone through building the Excel model to see where all the pieces fit the student and the instructor both have a better understanding of the results from the black box.

Furthermore, if a special statistical package is not available, a complete Poisson regression can be built in Excel by the use of the formulas given.

Student reaction has been varied. Although a small number of students still like only the black box approach, the majority responded that developing the models in Excel provided them with insights that they would not get in the black box approach. Some of the insights pointed out by the students include:

(1) Understanding the importance of *Y* being a discrete count where some values are "rare events".

(2) Understanding that independent variables may be both categorical and quantitative inputs which do not affect the model development.

(3) The building of the regression models allowed better understanding of the "concept" of model building as well as model "selection" from real data sets.

(4) Understanding the importance of including or excluding independent variables from the modeling process. The results were no longer just numbers.

References

1. Devore, J. (1995). Probability and Statistics for Engineering and the Sciences, 4th Ed. Wadsworth Publishing, Belmont: CA, pp. 474-509.

2. Afifi, A. & S. Azen. (1979). Statistical Analysis: A Computer Oriented Approach, 2nd Ed. Academic Press, New York: NY, pp. 143-144.

3. Giordano, F., W. Fox, & S. Horton (2013). A First Course in Mathematical Modeling, 5th Ed. Cengage Publishers, Boston: MA, pp. 235-240.

4. Neter, J., M. Kutner, C. Nachtsheim, W. Wasserman. (1996). Applied Linear Statistical Models, 4th Ed. Irwin Press, Boston: MA, pp. 609-614.

5. Montgomery, D. E. Peck, G. Vining. (2006). Introduction to Linear Regression, 4th Ed. John Wiley and Sons, New York: NY, pp. 427-453.

6. Fox, W.P. (2012). Issues and Importance of "Good" Starting Points for Nonlinear Regression for Mathematical Modeling with Maple: Basic Model Fitting to Make Predictions with Oscillating Data", JCMST, **31**(1), pp. 1-16.

7. http://www.oxfordjournals.org/our_journals/tropej/online/ma_chap13.pdf (accessed April 10, 2012)

Biographical Information

Dr. William P. Fox is a professor in the Department of Defense Analysis at the Naval Postgraduate School. He received his BS degree from the United States Military Academy at West Point, New York, his MS from the Naval Postgraduate School, and his Ph.D. from Clemson University. Previously he has taught at the United States Military Academy and Francis Marion University, where he was the chair of the mathematics department for eight years. He has many publications including books, chapters, journal articles, conference presentations, and workshops. He directs several math modeling contests through COMAP. His interests include applied mathematics, optimization (linear and nonlinear), mathematical modeling, statistical models for medical research, and computer simulations. He is currently the President of the Military Application Society in INFORMS.