PubH 7405:
REGRESSION ANALYSIS

INTRODUCTION TO

POISSON REGRESSION

PROTOTYPE EXAMPLE #1

We have data for 44 physicians working in emergency medicine at a major hospital system. The concern is about the number of complaints each received the previous year. Why is it different from physician to physician? Could it be explained using other factors? Data available consist of the number of patient visits - and four covariates (the revenue, in dollars per hour; work load at the emergency service, in hours; gender, Female/Male, and residency training in emergency medicine, No/Yes.

Question:

Can we do Regression using the Normal Error Regression Model?

Possible Issue:

The response, Number of Complaints, is not on continuous scale; it's a count – of course, not normally distributed.

PROTOTYPE EXAMPLE #2

Skin Cancer data for different age groups in two metropolitan areas: Minneapolis-St. Paul and Dallas-Fort Worth: (1) Any age effect? If so, is it the same for the two cities? (2) Any weather effect? (difference between two cities) If so, is it the same for all age groups?

Skin Cancer Data				
City:	Minneapolis-St. Paul		Dallas-Ft. Worth	
Age Group	Cases	Population	Cases	Population
15-24	1	172,675	4	181,343
25-34	16	123,063	38	146,207
35-44	30	96,216	119	121,374
45-54	71	92,051	221	111,353
55-64	102	72,159	259	83,004
65-74	130	54,722	310	55,932
75-84	133	32,185	226	29,007
85+	40	8,328	65	7,538

Questions are regression-type questions (about main effects or marginal contributions and about possible effect modifications).

& the same possible issue: The Response or Dependent Variable, the Number of Skin Cancer Cases – a count, is not on the continuous scale and not normally distributed.

In Regression Analysis/Model, we usually impose the condition that the Response Variable Y is on the continuous scale maybe because of the popular "Normal Error Model" - not because Y is always on the continuous scale.

In a previous lecture, the Dependent Variable of interest was represented by an Binary or Indicator Variable Y taking on values 0 and 1. The distribution is "Bernoulli" which is a special case of the Binomial Distribution. And we introduced "Logistic Regression".

We are now introducing a new form of regression, the Poisson Regression, where the Response or Dependent Variable Y represents "count" data – non-negative integers.

RARE EVENTS

It can be shown that the limiting form of the binomial distribution, when n is increasingly large ($n \to \infty$) and π is increasingly small ($\pi \to 0$) while $\theta = n\pi$ (the mean) remains constant, is:

$$\Pr(X = x) = \frac{\theta^x e^{-\theta}}{x!}$$

$$= p(x; \theta)$$

A random variable having this probability function is said to have "Poisson Distribution" $P(\theta)$. For example, with $n = 48$ and $\pi = .05$:

$b(x = 5; n, \pi) = .059$

$p(x = 5; \theta) = .060$

The Poisson Model is often used when the random variable X is supposed to represent the number of occurrences of some random event in an unit interval of time or space, or some volume of matter; numerous applications in health sciences have been documented. For example, the number of virus in a solution, the number of defective teeth per individual, the number of focal lesions in virology, the number of victims of specific diseases, the number of cancer deaths per household, the number of infant deaths in certain locality during a given year, among others.

The mean and variance of the Poisson Distribution are:

$$\mu = \theta$$

$$\sigma^2 = \theta$$

(A very special and “strong” characteristic where the variance is equal to the mean).

REGRESSION NEEDS

The Poisson model is often used when the random variable X is supposed to represent the number of occurrences of some random event in an interval of time or space, or some volume of matter and numerous applications in health sciences have been documented. In some of these applications, one may be interested in to see if the Poisson-distributed dependent variable Y can be predicted from or explained by other variables. The other variables are called predictors, or explanatory or independent variables. For example, we may be interested in the number of defective teeth Y per individual as a function of gender and age of a child, branch of toothpaste, and whether the family has or does not have dental insurance.

& refer to Example 1 (Number of complains) and Example 2 (Number of skin cancer cases).

POISSON REGRESSION MODEL

When the dependent variable Y is assumed to follow a Poisson distribution with mean θ; the Poisson regression model expresses this mean as a function of certain independent variables X_1, X_2, ..., X_k, in addition to the *size* of the observation unit from which one obtained the count of interest. For example, if Y is the number of virus in a solution then the size is the volume of the solution; or if Y is the number of defective teeth then the size is the total number of teeth for that same individual.

In our frame work, the dependent variable Y is assumed to follow a Poisson distribution; its values y_i's are available from n "observation unit" which is also characterized by an independent variable X. For the observation unit "i" ($i \le n$), let s_i be the size and x_i be the covariate value. The Poisson regression model assumes that the relationship between the mean of Y and the covariate X is described by:

$$E(Y_i) = s_i \lambda(x_i)$$
$$= s_i \exp(\beta_0 + \beta_1 x_i)$$

where $\lambda(x_i)$ is called the "risk" of/for observation unit i ($1 \le i \le n$).

The basic rationale for using the term "risk" is the approximation of the Binomial distribution by the Poisson distribution. Recall that, when n goes to infinity, π tends 0 while $\theta = n\pi$ remains constant, the binomial distribution $B(n,\pi)$ can be approximated by the Poisson distribution $P(\theta)$. The number n is the size of the observation unit; so the ratio between the mean and the size represents π (or $\lambda(x)$ in the new model); that's the "probability" or "risk" (and, the ratio of risks is called the "risks ratio" or "relative risk").

Model with Several Covariates

Suppose we want to consider k covariates, X_1, X_2, …, X_k, simultaneously. The simple Poisson regression model of previous section can be easily generalized and expressed as:

$$E(Y_i) = s_i \lambda(x_{1i}, x_{2i}, ..., x_{ki})$$
$$= s_i \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_1 x_{ki})$$
$$= s_i \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji})$$

where $\lambda(x_{ji}$'s) is called the "risk" of/for observation unit i ($1 \le i \le n$), x_{ji} is the value of the covariate X_j measured from subject i.

EXAMPLE

The purpose of this study was to examine the data for 44 physicians working for an emergency at a major hospital so as to determine which of the following four variables are related to the number of complaints received during the previous year. In addition to the number of complaints, served as the dependent variable, data available consist of the number of visits - which serves as the *size* for the observation unit, the physician - and four covariates. Table 6.2 presents the complete data set. For each of the 44 physician there are two continuous independent variables, the revenue (dollars per hour) and work load at the emergency service (hours) and two binary variables, gender (Female/Male) and residency training in emergency services (No/Yes).

No. of Visits	Complaint	Gender	Residency	Revenue	Hours
2014	2	Y	F	263.02	1287.25
3091	3	N	M	334.94	1588.00
879	1	Y	M	206.42	705.25
1780	1	N	M	236.32	1005.50
3646	11	N	M	288.91	1667.25
2690	1	N	M	275.94	1517.75
1864	2	Y	M	295.71	967.00
2782	6	N	M	224.91	1609.25
3071	9	N	F	249.32	1747.75
1502	3	Y	M	269.00	906.25
2438	2	N	F	225.61	1787.75
2278	2	N	M	212.43	1480.50
2458	5	N	M	211.05	1733.50
2269	2	N	F	213.23	1847.25
2431	7	N	M	257.30	1433.00
3010	2	Y	M	326.49	1520.00
2234	5	Y	M	290.53	1404.75
2906	4	N	M	268.73	1608.50
2043	2	Y	M	231.61	1220.00
3022	7	N	M	241.04	1917.25
2123	5	N	F	238.65	1506.25
1029	1	Y	F	287.76	589.00
3003	3	Y	F	280.52	1552.75
2178	2	N	M	237.31	1518.00
2504	1	Y	F	218.70	1793.75
2211	1	N	F	250.01	1548.00
2338	6	Y	M	251.54	1446.00
3060	2	Y	M	270.52	1858.25
2302	1	N	M	247.31	1486.25
1486	1	Y	F	277.78	933.95
1863	1	Y	M	259.68	1168.25
1661	0	N	M	260.92	877.25
2008	2	N	M	240.22	1387.25
2138	2	N	M	217.49	1312.00
2556	5	N	M	250.31	1551.50
1451	3	Y	F	229.43	9.73.75
3328	3	Y	M	313.48	1638.25
2928	8	N	M	293.47	1668.25
2701	8	N	M	275.40	16.52.75
2046	1	Y	M	289.56	1029.75
2548	2	Y	M	305.67	1127.00
2592	1	N	M	252.35	1547.25
2741	1	Y	F	276.86	1499.25
3763	10	Y	M	308.84	1747.50

The interpretation or meaning of the "Regression Coefficients" could be seen as follows – which is similar to the case of the "normal Error Regression Model".

MEASURE OF ASSOCIATION

Consider the case of a binary covariate X, say, representing an exposure (1 = exposed, 0 = not exposed). We have the following:

$$\ln\lambda_i = \begin{cases} \beta_0 + \beta_1 + \ln s_i \text{ if exposed (or } x = 1) \\ \beta_0 + \ln s_i \text{ if unexposed (or } x = 0) \end{cases}$$

$$\frac{\lambda_i(\text{exposed})}{\lambda_i(\text{unexposed})} = \exp(\beta_1)$$

This quantity, represented by exp(β1), is the relative risk associated with the exposure.

Similarly, we have for a continuous covariate X and consider any value x of X,

$$\ln\lambda_i = \begin{cases} \beta_0 + \beta_1 x + \ln s_i \text{ if } X = x \\ \beta_0 + \beta_1 (x+1) + \ln s_i \text{ if } X = x+1 \end{cases}$$

$$\frac{\lambda_i (X = x+1)}{\lambda_i (X = x)} = \exp(\beta_1)$$

This quantity, represented by exp(β1), is the relative risk associated with one unit increase in the value of X, from x to (x+1).

ESTIMATION OF PARAMETERS

Under the assumption that Yi is distributed as Poisson with the above mean, the Likelihood Function is given by:

$$L(y;\beta) = \prod_{i=1}^{n} \left\{ \frac{[s_i \lambda(x_i)]^{y_i} \exp[-s_i \lambda(x_i)]}{y_i!} \right\}$$

$$\ln L(y;\beta) = \sum_i \left\{ y_i \ln s_i - \ln y_i! + y_i[\beta_0 + \beta_1 x_i] - s_i \exp[\beta_0 + \beta_1 x_i] \right\}$$

from which estimates of the two regression coefficients β_0 and β_1 can be obtained by the Maximum Likelihood procedure.

EXAMPLE #1

Refer to the Emergency Service data and suppose we want to investigate the relationship between the number of complaints Y (adjusted for number of visits) and residency training X. It may be perceived that by having training in the specialty a physician would perform better and, therefore, less likely to provoke complaints. An application of the simple Poisson regression analysis yields:

Variable	Coefficient	St Error	z-Statistic	p-Value
Intercept	-6.7566	0.1387	-48.714	<0.0001
No Residency	0.3041	0.1725	1.763	0.0779

The result indicates that the common perception is almost true, that the relationship between the number of complaints and no residency training in emergency service is marginally significant (p = 0.0779); the relative risk associated with no residency training is:

exp(.3041) = 1.36

Those without previous training is 36 percent more likely to receive the same number of complaints as those who were trained in the specialty.

SAS SAMPLE

a SAS program would include these instructions:

```
DATA EMERGENCY;
INPUT VISITS CASES RESIDENCY;
LN = LOG(VISITS);
CARDS;
(Data);
PROC GENMOD DATA EMERGENCY;
CLASS RESIDENCY;
MODEL CASES = RESIDENCY/ DIST = POISSON LINK = LOG OFFSET = LN;
```

where EMERGENCY is the name assigned to the data set, VISITS is the number of visits, CASES is the number of complaints (Y), and RESIDENCY (X) is the binary covariate indicating whether the physician received residency training in the specialty. The option CLASS is used to declare that the covariate is categorical.

Model with Several Covariates

Suppose we want to consider k covariates, X_1, X_2, …, X_k, simultaneously. The model is:

$$\begin{aligned} E(Y_i) &= s_i\lambda(x_{1i}, x_{2i}, ..., x_{ki}) \\ &= s_i \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_1 x_{ki}) \\ &= s_i \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}) \end{aligned}$$

where $\lambda(x_{ji}$'s) is called the "risk" of/for observation unit i ($1 \le i \le n$), x_{ji} is the value of the covariate X_j measured from subject i.

$$L(y;\beta) = \prod_{i=1}^{n} \left\{ \frac{[s_i\lambda(x_{ji})]^{y_i} \exp[-s_i\lambda(x_{ji})]}{y_i!} \right\}$$

$$\ln L(y;\beta) = \sum_{i} \left\{ y_i \ln s_i - \ln y_i! + y_i[\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}] - s_i \exp[\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}] \right\}$$

Also similar to the simple regression case, exp(β_i) represents:

(i) The Relative Risk associated with, say, an exposure if X_i is binary (exposed or X_i=1 versus unexposed or X_i=0),

or:

(ii) The Relative Risk due to one unit increase in the value of X_i if X_i is continuous (X_i=x+1 versus X_i=x).

After estimates b's of regression coefficients β's and their standard errors have been obtained, a 95 percent confidence interval for the log of the Relative Risk associated with the ith factor is given by: $b_i \pm SE(b_i)$. These results are necessary in the effort to identify important risk factors for the Poisson outcome, the "count".

Note:

We form confidence intervals of the regression coefficient before exponentiating (the endpoints) to obtain relative risks.

Before such analyses are done, the problem and the data have to be examined carefully.

If some of the variables are highly correlated, then one or fewer of the correlated factors are likely to be as good predictors as all of them; information from other similar studies also has to be incorporated so as to drop some of these correlated explanatory variables.

The uses of products, such as $x_1 * x_2$, and higher power terms, such as x_1^2, may be necessary and can improve the goodness-of-fit.

We are assuming a log-linear regression model in which, for example, the Relative Risk due to one unit increase in the value of a continuous X_i ($X_i = x+1$ versus $X_i = x$) is independent of x. Therefore, if this "linearity" seems to be violated, the incorporation of powers of X_i should be seriously considered. The use of products will help in the investigation of possible effect modifications.

There are no simple diagnostic tool, such as Scatter Diagram for Normal Error Regression Model, for detecting lack of linearity. One might consider to include a quadratic term and see if it's significant.

And, finally, the messy problem of missing data; most packaged programs would delete a subject if one or more covariate values are missing.

TESTING HYPOTHESES

Once we have fit a multiple Poisson regression model and obtained estimates for the various parameters of interest, we want to answer questions about the contributions of various factors to the prediction of the Poisson-distributed response variable. There are three types of such questions:

(i) An overall test: taken collectively, does the entire set of explanatory or independent variables contribute significantly to the prediction of response?

(ii) Test for the value of a single factor: does the addition of one particular variable of interest add significantly to the prediction of response over and above that achieved by other independent variables?

(iii) Test for contribution of a group of variables: does the addition of a group of variables add significantly to the prediction of response over and above that achieved by other independent variables?

OVERALLL REGRESSION EFFECTS

We first consider the first question stated above concerning an overall test for a model containing k factors. The null hypothesis for this test may be stated as: "all k independent variables considered together do not explain the variation in the response any more than the size alone". In other words,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_o : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

Since:

$$E(Y_i) = s_i \lambda(x_{1i}, x_{2i}, \ldots, x_{ki})$$
$$= s_i \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki})$$
$$= s_i \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji})$$

If H_0 is true, the average/mean response has nothing to do with the predictors.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

This can be tested using the Likelihood Ratio Chi-square test at k degrees of freedom:

$$X^2 = 2\{\ln L_k - \ln L_0\}$$

where $\ln L_k$ is the log likelihood value for the model containing all k covariates and lnL0 is the log likelihood value for the model containing only the intercept. Computer packaged program, such as SAS PROC GENMOD, provides these log likelihood values but in separate runs.

If this overall test is “significant”, it only means “one or some of the regression coefficients (the “slopes”) not equal to zero; it does not tell us which coefficients are not zero or which factor or factors are related to the response. This is similar to performing the one-way ANOVA F-test before checking for pairwise differences.

TESTING FOR SINGLE FACTOR

- The question is: “Does the addition of one particular factor of interest add significantly to the prediction of Dependent Variable over and above that achieved by other factors?”.
- The Null Hypothesis for this test may stated as: "Factor X_i does not have any value added to the explain the variation in Y-values over and above that achieved by other factors ". In other words,

$$H_0 : \beta_i = 0$$

- The Null Hypothesis is $H_0 : \beta_i = 0$
- Regardless of the number of variables in the model, one simple approach is using $z_i = \frac{b_i}{SE(b_i)}$
- where b_i is the corresponding estimated regression coefficient for factor X_i and $SE(b_i)$ is the estimate of the standard error of bi, both of which are printed by standard computer packaged programs such as SAS. In performing this test, we refer the value of the z score to percentiles of the standard normal distribution; for example, we compare the absolute value of z to 1.96 for a two-sided test at the 5 percent level. Note: No use of t-distribution here.

EXAMPLE #2

Refer to the data set on emergency service; Using all four covariates, we found that only the effect of work load (Hours) is significant at the 5 percent level.

Variable	Coefficient	St Error	z-Statistic	p-Value
Intercept	-8.1338	0.9220	-8.822	<0.0001
No Residency	0.2090	0.2012	1.039	0.2988
Female	-0.1954	0.2182	-0.896	0.3703
Revenue	0.0016	0.0028	0.571	0.5775
Hours	0.0007	0.0004	1.750	0.0452

Given a continuous variable of interest, one can fit a polynomial model and use this type of test to check for linearity. It can also be used to check for a single product representing an effect modification. For example, to focus on the quadratic term of:

$$E(Y_i) = s_i \lambda(x_i)$$

$$= s_i \exp(\beta_0 + \beta_1 Hour_i + \beta_2 Hour_i^2)$$

TESTING FOR A GROUP OF VARABLES

- **The question is: "Does the addition of a group of factors add significantly to the prediction of Y over and above that achieved by other factors?**
- **The Null Hypothesis for this test may stated as: "Factors $\{X_{i+1}, X_{i+2}, \ldots, X_{i+m}\}$, considered together as a group, do not have any value added to the prediction of the Mean of Y that other factors are already included in the model". In other words,**

$$H_0 : \beta_{i+1} = \beta_{i+2} = \ldots = \beta_{i+m} = 0$$

- **This "multiple contribution" test is often used to test whether a similar group of variables, such as demographic characteristics, is important for the prediction of the mean of Y; these variables have some trait in common.**
- **Another application: collection of powers and/or product terms. It is of interest to assess powers & interaction effects collectively before considering individual interaction terms in a model. It reduces the total number of tests & helps to provide better control of overall Type I error rates which may be inflated due to multiple testing.**

The process can also be used to test for the contribution of a categorical covariate which is represented by several “dummy variables”.

EXAMPLE #3

In this example, the dependent variable is the number of cases of skin cancer. Data involve only two covariates; age and location, both are categorical. We use seven dummy variables to represent the eight age groups (with "85+" age group being the baseline) and one for location (with Minneapolis-St. Paul as the baseline)

Skin Cancer Data				
City:	Minneapolis-St. Paul		Dallas-Ft. Worth	
Age Group	Cases	Population	Cases	Population
15-24	1	172,675	4	181,343
25-34	16	123,063	38	146,207
35-44	30	96,216	119	121,374
45-54	71	92,051	221	111,353
55-64	102	72,159	259	83,004
65-74	130	54,722	310	55,932
75-84	133	32,185	226	29,007
85+	40	8,328	65	7,538

SEQUENTIAL ADJUSTMENT

In the type 3 analysis, we test the significance of the effect of each factor added to the model containing all other factors – like in most common multiple regression analyses; that is to investigate the additional contribution of the factor to the explanation of the dependent variable. Sometimes, however, we may be interested in a hierarchical or sequential adjustment. For example, we focus on the quadratic term (adjusted) in addition to the regular term (unadjusted). This can be achieved using PROC GENMOD by requesting the type 1 analysis option

OVERDISPERSION

The Poisson is a very special distribution; its mean μ and its variance σ^2 are equal. If we use the variance-mean ratio as a dispersion parameter then it is 1 in a standard Poisson model, less than 1 in an under-dispersed model, and greater than 1 in an over-dispersed model. Over-dispersion is a common phenomenon in practice and it causes concerns because the implication is serious; the analysis which assumes the Poisson model often under-estimates standard errors and, thus, wrongly inflates the level of significance.

MEASURING OVERDISPERSION

After a Poisson regression model is fitted, dispersion is measured by the scaled deviance or scaled Pearson chi-square; it is the deviance or Pearson chi-square divided by the degrees of freedom (number of observations minus number of parameters). The deviance is defined as twice the difference between the maximum achievable log likelihood and the log likelihood at the maximum likelihood estimates of the regression parameters.

EXAMPLE

Refer to the data set on emergency service with all four covariates, we can see that both indices are greater than 1 indicating an over-dispersion. In this example, we have a sample size of 44 but five degrees of freedom lost due to the estimation of the five regression parameters, including the intercept.

Criterion	df	Value	Scaled Value
Deviance	39	54.52	1.3980
Pearson Chi-Square	39	54.42	1.3700

FITTING OVERDISPERSED MODEL

PROC GENMOD allows the specification of a scale parameter to fit over-dispersed Poisson regression models. Instead of a variance equal to the mean,

$$\text{Var}(Y) = \mu$$

it allows the variance function to have a multiplicative "over-dispersion factor" φ (specified by users):

$$\text{Var}(Y) = \varphi\mu$$

EXAMPLE #4

Refer to the data set on emergency service; Using all four covariates, we have the following results by fitting the "regular" Poisson Model.

Variable	Coefficient	St Error	z-Statistic	p-Value
Intercept	-8.1338	0.9220	-8.822	<0.0001
No Residency	0.2090	0.2012	1.039	0.2988
Female	-0.1954	0.2182	-0.896	0.3703
Revenue	0.0016	0.0028	0.571	0.5775
Hours	0.0007	0.0004	1.750	0.0452

Criterion	df	Value	Scaled Value
Deviance	39	54.52	1.3980
Pearson Chi-Square	39	54.42	1.3700

The model could be fitted in the usual way, and the point estimates of regression coefficient are not affected. The covariance matrix, however, is multiplied by φ.

There are two options available for fitting over-dispersed models; the users can control either the scaled deviance (by specifying DSCALE in the model statement) or the scaled Pearson chi-square (by specifying PSCALE in the model statement). The value of the controlled index becomes 1; the value of the other is close to but may not be equal to 1.

Note: a SAS program would include this instruction:

```
MODEL CASES = GENDER RESIDENCY REVENUE HOURS/
              DIST = POISSON LINK = LOG OFFSET = LN(DSCALE);
```

NEW RESULTS

Variable	Coefficient	St Error	z-Statistic	p-Value
Intercept	-8.1338	1.0901	-7.462	<0.0001
No Residency	0.2090	0.2378	0.879	0.3795
Female	-0.1954	0.2679	-0.758	0.4486
Revenue	0.0016	0.0033	0.485	0.6375
Hours	0.0007	0.0004	1.694	0.0903

Criterion	df	Value	Scaled Value
Deviance	39	39.00	1.000
Pearson Chi-Square	39	38.22	0.980

As compared to the results of the regular model, the point estimates remain the same but the standard errors are larger; the effect of work load (Hours) is no longer significant at the 5 percent level.

STEPWISE REGRESSION

In many applications, our major interest is to identify important risk factors. In other words, we wish to identify from many available factors a small subset of factors that relate significantly to the outcome, e.g. the disease under investigation. In that identification process, of course, we wish to avoid a large Type I (false positive) error. In a regression analysis, a Type I error corresponds to including a predictor that has no real relationship to the outcome; such an inclusion can greatly confuse the interpretation of the regression results. One popular procedure is “Stepwise Regression”

With Poisson Regression, it is still desirable and possible to perform stepwise regression. Unfortunately, PROC GENMOD does not have an automatic option; one has to run it many times and, at each step, choose to add or remove a variable “manually”.

With "Poisson Regression", we "lost" these tools that we use with NERM:

R^2 and all coefficients of partial determination (there are some substitutes but not as good)

All graphs (Scatter diagram, all residual plots, and Variable-added Plot)

Least Squares method (but MLE is better)

All other methods/tools are unchanged (test for single factors, stepwise, etc…)

However, if Y is distributed as Poisson, try $Y = \sqrt{Y}$; it both stabilizes the variance and improves normality. Then you could get approximate results using PROC REG and the Normal Error Regression Model (If Y is binary, you would have no choice but Logistic Regression).

Readings & Exercises

- **Readings: A thorough reading of the text's section 14.3 (pp.618-622) is highly recommended.**
- **Exercises: None**

Due As Homework

22.1 Refer to dataset "Emergency Service", let Y = Number of Complaints and four independent variables, X_1 = Gender, X_2 = Residency, X_3 = Revenue, and X_4 = Hours; choose coding 0/1 for X1 and X2.

a) Fit the Poisson Regression Model using PROC GENMOD and confirm the results of Example #2.

b) Using the same data set, take the square root of Y and use as the new Dependent Variable and fit the Normal Error Regression Model using PROC REG; draw your conclusions on effects of covariates.

c) Comment on the similarities or differences between the two sets of results obtained in (a) and (b).