Topic 32 - Poisson Regression and Categorical Data Analysis

STAT 525 - Fall 2003

Outline

- Poisson Regression
  - Background
  - Model
  - Inference
- Categorical Data Analysis
  - Goodness-of-fit test
  - Test of Independence
  - Log-linear models

Poisson Regression

- Like logistic regression, this is a nonlinear regression model for discrete outcomes
- Used when response variable is a count
  - Number of acute asthma attacks in a week
  - Number of visits to mall during December
- Useful when large count is a rare event. Otherwise, standard linear model with (nonconstant) normal errors reasonable
- Poisson counts $\rightarrow$ use $\sqrt{\ }$ transformation to stabilize variance

Regression Model

- Standard linear/nonlinear regression form:

$$Y_i = \mathrm{E(Y_i)} + \varepsilon_\mathrm{i}$$

- Generalized linear model: Exponential family distribution for Y and link between the mean and covariates
- Common links $\mu(\mathbf{X_i}, \boldsymbol{\beta})$ for Poisson
  - $\mu_i = \mathbf{X_i'}\boldsymbol{\beta}$
  - $\mu_i = \exp(\mathbf{X_i'}\boldsymbol{\beta})$
  - $\mu_i = \log(\mathbf{X_i'}\boldsymbol{\beta})$
  - All must result in μ_i nonnegative
- Model is such that the Y_i's are independent Poisson random variables with expected values $\mu_i = \mu(\mathbf{X_i}, \boldsymbol{\beta})$

Estimation

- Given the distribution of Y_i, we can formulate the likelihood function

$$\log(L) \quad = \quad \sum Y_i \log(\mu(\mathbf{X_i}, \boldsymbol{\beta})) - \sum \mu(\mathbf{X_i}, \boldsymbol{\beta}) + C$$

- MLEs do not have closed forms
- Iterative reweighted least squares or other numerical search procedures used to solve for β

Hypothesis Testing

- Hypothesis testing and formulation of test statistics is done in similar manner to logistic regression.
- The deviance of a fitted model is the difference between the log-likelihood of the fitted model and a model that has a parameter (μ_i) for each observation Y_i *(i.e., use all of the degrees of freedom so that the residuals will be zero)*.
- For Poisson regression

$$\mathrm{DEV(X_1, X_2, ..., X_{p-1})} \quad = \quad -2\left[\sum Y_i \log\left(\frac{\hat{\mu}_i}{Y_i}\right) + \sum (Y_i - \hat{\mu}_i)\right]$$

Hypothesis Testing

- Can use deviance to compare models
- Models must be hierarchical (Full/Reduced)

$$\begin{aligned}\mathrm{DEV(X_q, ..., X_{p-1} | X_0, ..., X_{q-1})} \quad &= \quad \mathrm{DEV(X_0, ..., X_{q-1})} \\ &- \quad \mathrm{DEV(X_0, ..., X_{p-1})}\end{aligned}$$

- Partial deviance approx χ^2 with $p - q$ df
- Must compute by hand or with GENMOD
- Prediction
  - Mean response for predictors $\mathbf{X}$
  - Probability of specific count (i.e., $Y = 0$)

Example Page 621

- Lumber company interested in the relationship between the number of customers from a census tract and
  - X_1 : Number of housing units
  - X_2 : Average income
  - X_3 : Average housing unit age
  - X_4 : Distance to nearest competitor
  - X_5 : Distance to store
- Will use exp link $\rightarrow \log(\mu_i) = \mathbf{X}'\boldsymbol{\beta}$

SAS Commands

```
data poi;
 infile 'U:\.www\datasets525\CH14TA14.TXT';
 input ncust x1 x2 x3 x4 x5;

proc genmod plots=all;
 model ncust= x1 x2 x3 x4 x5 / link=log dist=poi
                               type3 obstats;
proc print;
run;
```


Output

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	104	114.9854	1.1056
Scaled Deviance	104	114.9854	1.1056
Pearson Chi-Square	104	101.8808	0.9796
Scaled Pearson X2	104	101.8808	0.9796
Log Likelihood		1898.0224	

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	P>ChiSq
Intercept	1	2.9424	0.2072	2.5362	3.3486	201.57	<.0001
x1	1	0.0006	0.0001	0.0003	0.0009	18.17	<.0001
x2	1	-0.0000	0.0000	-0.0000	-0.0000	30.63	<.0001
x3	1	-0.0037	0.0018	-0.0072	-0.0002	4.37	0.0365
x4	1	0.1684	0.0258	0.1179	0.2189	42.70	<.0001
x5	1	-0.1288	0.0162	-0.1605	-0.0970	63.17	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

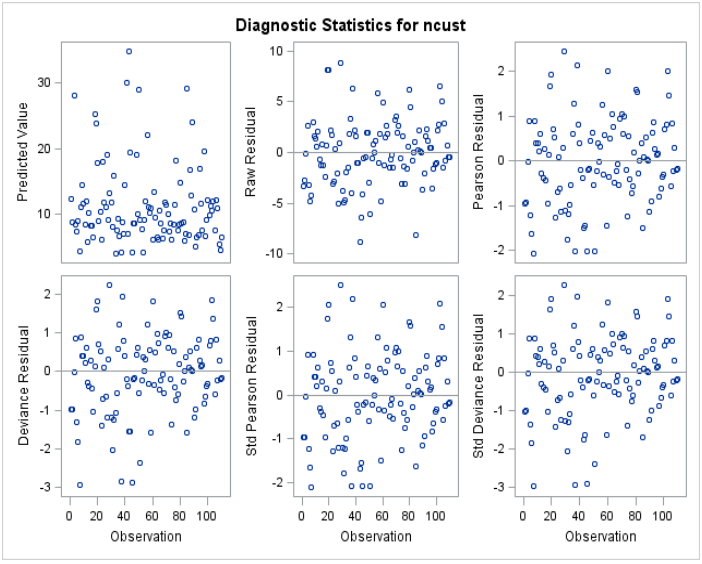
Diagnostics

- Goodness of fit
  - Deviance goodness of fit : Compare overall fit with χ^2_{n-p}
- "Residual" analysis
  - Distribution of residuals under correct model is unknown and thus common residual plot uninformative.
  - Deviance residual is the signed square root of the contribution to the model deviance

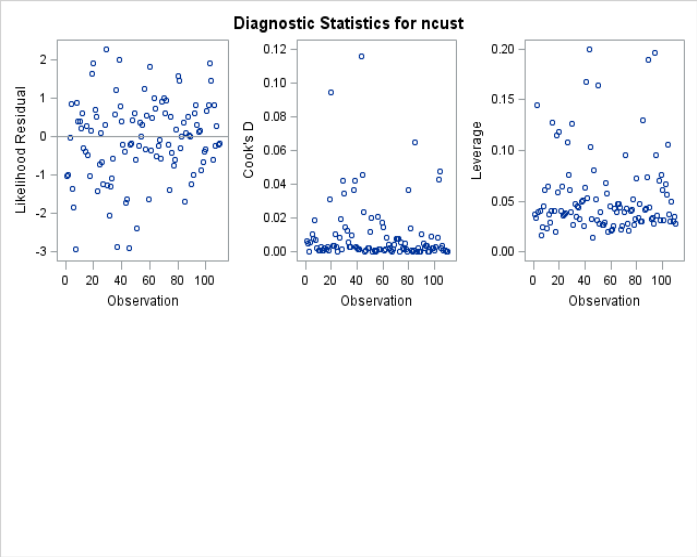
$$\sum (dev_i)^2 = \mathrm{DEV}(\mathrm{X}_0, ..., \mathrm{X}_{\mathrm{p}-1})$$

  - Sign depends on $\hat{\mu}_i > Y_i$
  - Can plot dev_i by observation (index plot)
  - Can generate half probability plot with envelope

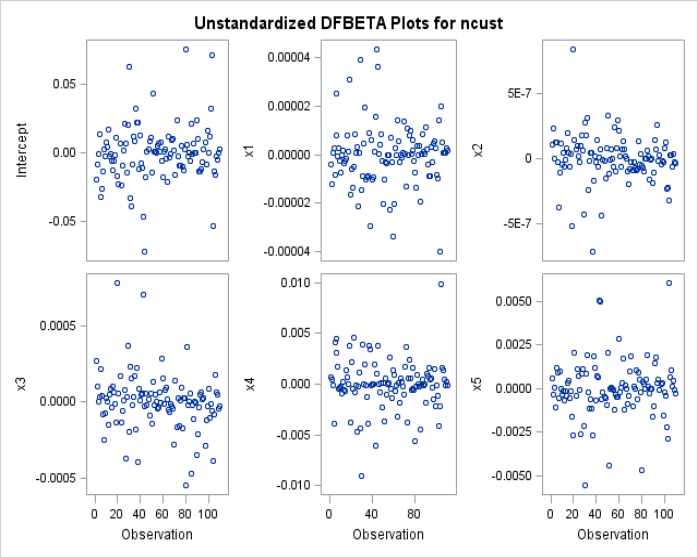
Influence/Residual plots



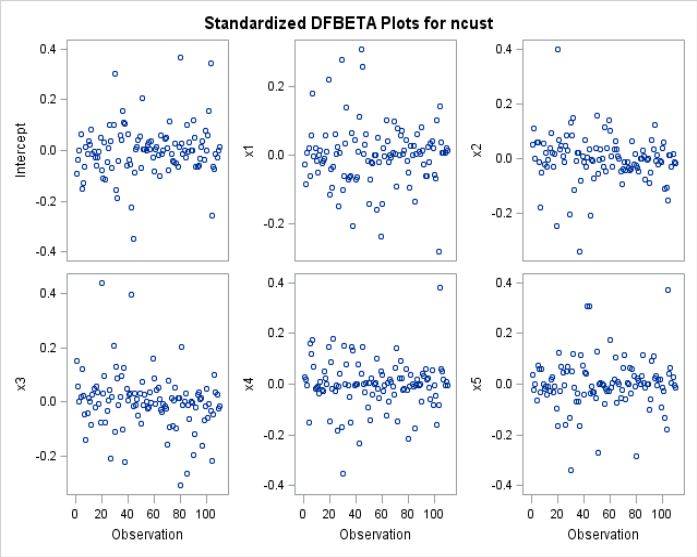
Influence/Residual plots



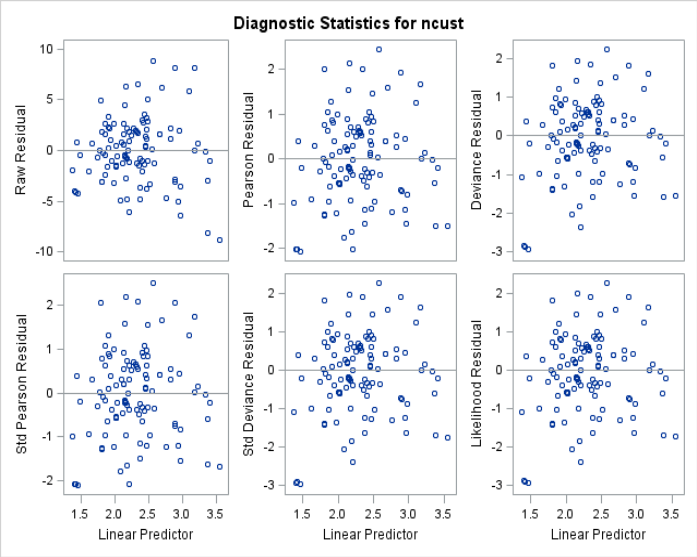
Influence/Residual plots



Influence/Residual plots



Influence/Residual plots



Poisson Model Variations

- Often more variation than expected under Poisson model.
  - Could consider using scale parameter to adjust variability
  - Move to a negative binomial distribution setting
- Situations where there is an excessive number of zeros
  - Zero-inflated Poisson (ZIP) often utilized
  - Available in COUNTREG and FMM procedures

Categorical Analysis

- Logistic regression used to assess the relationship between categorical response explanatory variables
- Will now present a few alternative approaches for categorical variables
- Example:
  - Each student in a sample of 100 is classified according to enrollment (full-time or part-time) and status (Fr, So, Jr, or Sr)
- First, consider univariate analysis
  - Is the proportion of full-time students at least 90%?
  - Is there equal distribution in terms of status?

Notation and Hypotheses

- Have k possible values (levels or categories) of explanatory variable
- Let π_i be the true proportion for category i with $\sum_i \pi_i = 1$
- Observe n_i in category i with $\sum_i n_i = n$
- Null Hypotheses:
  - $H_0 : \pi_1 \geq 0.90$
  - $H_0 : \pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25$

Goodness of Fit Test

- Conceptually, we want to compare the observed cell counts to the expected cell counts under the null hypothesis
- The expected cell counts come from substituting the corresponding hypothesized proportions for each π
- Thus, the expected counts are $n\pi_i$
- A quantitative measure of the extent to which the observed counts differ from the expected counts is

$$\chi^2 = \sum \frac{(\text{observed cell count - expected cell count})^2}{\text{expected cell count}}$$

Goodness of Fit Test

- The magnitude of this statistic reflects the magnitude of the discrepancies between the observed and expected counts
- As long as none of the expected cell counts are too small, when H_0 is true, this statistic is approximately chi-square distributed with $k-1$ degrees of freedom.
- General rule... expected cell count should be 5 or larger

Extensions

- Consider the question "Is there an association between status and enrollment?"
  - Is the proportion of full-time enrollment the same regardless of status?
  - Does knowledge of status provide information about the value of enrollment?
  - Are the two variables independent?
- Can address hypothesis using χ^2 test
- Requires bivariate analysis

Example

- Is there a difference in the drinking behavior of male and female students?
- Form 2×2 table

Drinking level	Men	Women
None	140	186
Low	478	661
Moderate	300	173
High	63	16

Test of Independence

- H_0: The two variables are independent
- Thus, the expected cell counts are $n\pi_i\pi_j$
- With π_i and π_j unknown, the expected value is obtained by multiplying the respective row and column marginals and dividing by the overall total
- The df is now (rows-1)(columns-1)

Example

Data are taken from a study that investigated the effect of oral contraceptive use on the likelihood of heart attacks. The 58 subjects in the first column represent married women under the age of 45 years of age treated for myocardial infarction in two hospital regions in England and Wales during 1968-1972. Each case was matched with three control patients in the same hospitals who were not being treated for myocardial infarction. All subjects were then asked whether they had ever used oral contraceptives.

	Myocardial Infarction	
Oral Contraceptive Practice	Yes	No
Used	23	34
Never used	35	132
Total	58	166

Taken from Agresti, "Categorical Data Analysis", Wiley.

SAS Code

```
options nocenter ls=75;
data a1;
oc='yes'; myo='yes'; count=23; output;
oc='yes'; myo='no'; count=34; output;
oc='no'; myo='yes'; count=35; output;
oc='no'; myo='no'; count=132; output;

proc freq data=a1;
weight count;
table oc*myo/all measures exact;
run;
```


Fisher's Exact Test

- Proposed by Fisher (1935)
- Based upon the thought that an observed sample could be used to reject the hypothesis if the total probability (under that hypothesis) of that sample is small.
- Test yields the probability of observing a table that gives at least as much evidence of association as the one actually observed, given that the null hypothesis is true.
- Row and columns marginals are assumed fixed. The hypergeometric probability p of every possible table is computed, and the p-value is computed

Output

```
oc         myo
Frequency
Percent
Row Pct
Col Pct       no     yes   Total
---------------------------------------------------
no           132      35     167
           58.93   15.63   74.55
           79.04   20.96
           79.52   60.34
---------------------------------------------------
yes           34      23      57
           15.18   10.27   25.45
           59.65   40.35
           20.48   39.66
---------------------------------------------------
Total         166      58     224
            74.11   25.89   100.00

Statistic                     DF       Value      Prob
Chi-Square                     1      8.3288    0.0039
Likelihood Ratio Chi-Square    1      7.8676    0.0050
Continuity Adj. Chi-Square     1      7.3488    0.0067
Mantel-Haenszel Chi-Square     1      8.2916    0.0040
Phi Coefficient                       0.1928
Contingency Coefficient               0.1893
```


Output

```
       Fisher's Exact Test
Cell (1,1) Frequency (F)       132
Left-sided Pr <= F          0.9986
Right-sided Pr >= F         0.0040
Table Probability (P)       0.0026
Two-sided Pr <= P           0.0052

          Estimates of the Relative Risk (Row1/Row2)

Type of Study               Value  95% Confidence Limits
Case-Control (Odds Ratio)  2.5513   1.3356         4.8734
Cohort (Col1 Risk)         1.3251   1.0556         1.6634
Cohort (Col2 Risk)         0.5194   0.3373         0.7998
Sample Size = 224
```


Proc CATMOD

- A procedure for categorical data modeling, where data are represented by contingency tables
- The **rows** of the table correspond to populations (or samples)
- The **columns** of the table correspond to the responses
- Responses formed on the basis of one or more dependent variables
- n_{ij} is the number of individuals in the i^{th} population with the j^{th} response.

Proc CATMOD

- Different types of categorical analyses
- Consider loglinear model, which is a specialized case of generalized linear models for Poisson-distributed data
- Conditional relationship between two or more discrete, categorical variables is analyzed
- Take log of the cell frequencies within a contingency table
- No distinction made between independent and dependent variables (i.e., models only demonstrate association)

Proc CATMOD

- Consider the oral contraceptive study
- Model

$$\log(F_{ij}) = \mu + R_i + C_j + (RC)_{ij}$$

- To see if there is association, you'd look at the interaction term.
- No interaction $\rightarrow$ independence model

SAS Code

```
data a1;
oc='yes'; myo='yes'; count=23; output;
oc='yes'; myo='no'; count=34; output;
oc='no'; myo='yes'; count=35; output;
oc='no'; myo='no'; count=132; output;

proc catmod data=a1;
weight count;
model oc*myo = _response_ /noparm noresponse
                           pred=freq;
loglin oc|myo;
run;
```


Output

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
oc	1	28.94	<.0001
myo	1	27.08	<.0001
oc*myo	1	8.05	0.0046
Likelihood Ratio	0	.	.

Maximum Likelihood Predicted Values for Frequencies

		Observed	Predicted		
oc	myo	Frequency	Frequency	Std Error	Residual
no	no	132	132	7.362918	3.049E-7
no	yes	35	35	5.433863	-4.07E-7
yes	no	34	34	5.369994	-1.23E-8
yes	yes	23	23	4.542968	1.142E-7

Background Reading

- KNNL Chapter 14
- knnl621.sas