Generalized Linear Models

Categorical and Non-normal Data

Generalized Linear Models

- Binomial variable– Responses with only two possible outcomes, e.g., defective items, sick patients. We try to see how a treatment (variable) affects the probability of an outcome. These probabilities are often estimated with sample proportions. These proportions often follow the binomial distribution.

Count Data

- Poisson Distribution– used to estimate count data
- Gamma Distribution– Skewed data
  - Standard deviation increases as the mean increases

Components of theGeneralized Linear Model

- **Random component**– identifies a probability distribution for the response variable. (normal, gamma, poisson, binomial)
- **Systematic component–** specifies the explanatory variables used as predictors in the model.
- **Link function**– specifies the function of the expected value of Y that is linearly predicted by the explanatory variables. (log, identity, logit)

Ordinary regression

- Random component (distribution of Y)
  - Normal
- Systematic component
  - model statement

$$\beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

- Link function
  g(u)=u (identity)

An Example Birth Rate and GNP

- Table 14.5 on Page 546

SAS Setup

```
proc plot; plot birth*gnp;run;
proc univariate plot; var birth; run;
proc reg data=brate; model birth=gnp; run;
proc genmod data=brate; model birth=gnp / dist=nor
link=identity;
proc genmod data=brate; model birth=gnp / dist=nor
link=log;run;
proc genmod data=brate; model birth=gnp / dist= gam
link=log; run;
```

Results from Proc Reg

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1450.26031	1450.26031	27.52	<.0001
Error	23	1212.02529	52.69675		
Corrected Total	24	2662.28560			

Results from Proc Reg

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	29.62270	2.03742	14.54	<.0001
gnp	1	-0.81331	0.15503	-5.25	<.0001

Results from Genmod

Model Information

Data Set	WORK.BRATE
Distribution	Normal
Link Function	Identity
Dependent Variable	birth
Observations Used	25

Results from Genmod

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	23	1212.0253	52.6968
Scaled Deviance	23	25.0000	1.0870
Pearson Chi-Square	23	1212.0253	52.6968
Scaled Pearson X2	23	25.0000	1.0870
Log Likelihood		-83.9881	

Results from Proc Reg

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1450.26031	1450.26031	27.52	<.0001
Error	23	1212.02529	52.69675		
Corrected Total	24	2662.28560			

Results from Proc Genmod

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	29.6227	1.9542	25.792	33.452	229.77	<.0001
gnp	1	-0.8133	0.1487	-1.1048	-0.5219	29.91	<.0001
Scale	1	6.9628	0.9847	5.2773	9.1868		

Results from Proc Reg

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	29.62270	2.03742	14.54	<.0001
gnp	1	-0.81331	0.15503	-5.25	<.0001

Results from Genmod

Model Information

Data Set	WORK.BRATE
Distribution	Normal
Link Function	Log
Dependent Variable	birth
Observations Used	25

Results from Genmod

Criteria For Assessing Goodness Of Fit

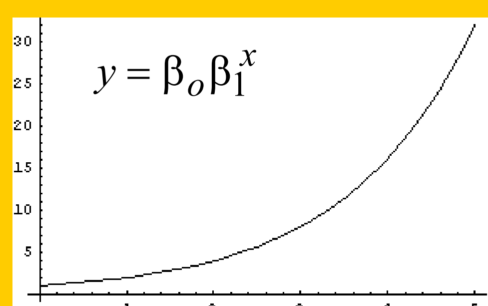
Criterion	DF	Value	Value/DF
Deviance	23	1083.8625	47.1245
Scaled Deviance	23	25.0000	1.0870
Pearson Chi-Square	23	1083.8625	47.1245
Scaled Pearson X2	23	25.0000	1.0870
Log Likelihood		-82.5911	

Results from Genmod: log link

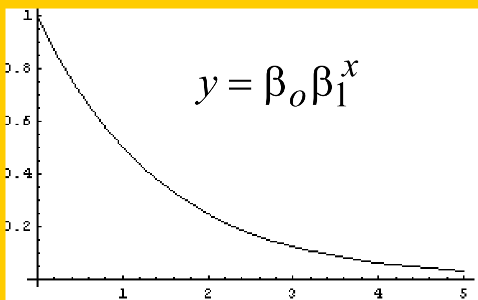
Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	3.4404	0.0676	3.3079	3.5730	2588.16	<.0001
gnp	1	-0.0476	0.0105	-0.068	-0.026	20.40	<.0001
Scale	1	6.5844	0.9312	4.9904	8.6875		

Exponential with B>1



Exponential with B<1



Exponential Growth Model

$$\mu = \beta_o \beta_1^x$$
$$\log \mu = \log(\beta_o \beta_1^x)$$
$$\log \mu = \log \beta_o + \log \beta_1^x$$
$$\log \mu = \alpha + \beta^* x$$

Exponential Growth Model

$$e^{\alpha} = \beta_o$$
$$e^{\beta^*} = \beta_1$$

Obtaining the Regression Coefficients in a log link model

$$\hat{y} = (e^{3.44})(e^{-.0476})^x$$
$$\hat{\beta}_o = (e^{3.44}) = 31.2$$
$$\hat{\beta}_1 = (e^{-.0476}) = .954$$

Interpretation

- Intercept: mean of y when x=0
- Slope: the exponential increase factor
  - For every unit change of gnp, the (predicted) birth rate is multiplied by .954
  - A decrease of 4.6% in birth rate for every unit increase in gnp.

US Population Growth

```
• data pop; input dec pop; cards;
• 0 62.95
• 1 75.99
• 2 91.97
• 3 105.71
• 4 122.78
• 5 131.67
• 7 179.32
• 8 203.30
• 9 226.54
• 10 248.71
• ;
• proc genmod data=pop; model pop=dec / dist=nor
  link=log; run;
```

Population Growth Example

Model Information	
Data Set	WORK.POP
Distribution	Normal
Link Function	Log
Dependent Variable	pop
Observations Used	10

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	8	223.387	27.923
Scaled Deviance	8	10.000	1.250
Pearson Chi-Square	8	223.387	27.923
Scaled Pearson X2	8	10.000	1.250
Log Likelihood		-29.721	

Analysis Of Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
Intercept	1	4.2567	0.0272	4.203	4.310	24564.2	<.0001
dec	1	0.1289	0.0034	0.122	0.135	1406.08	<.0001
Scale	1	4.7264	1.0569	3.049	7.326		

Intercept=70.4, ($e^{4.25}$, at 0 decade the population was 70 millions)

Slope= 1.14 ($e^{.1289}$, every decade the population grows by a factor of 1.14 (114%).