STATISTICS
TEXAS A&M UNIVERSITY

# Handout 13

# Statistical Analysis with the GLIMMIX Procedure

## Applications Using the GLIMMIX Procedure

# Poisson Regression with and without Random Effects

## Objective

- Use PROC GLIMMIX to fit a Poisson regression model with random effects.

- List issues related to overdispersion.

# Poisson Regression without Random Effects

Poisson regression

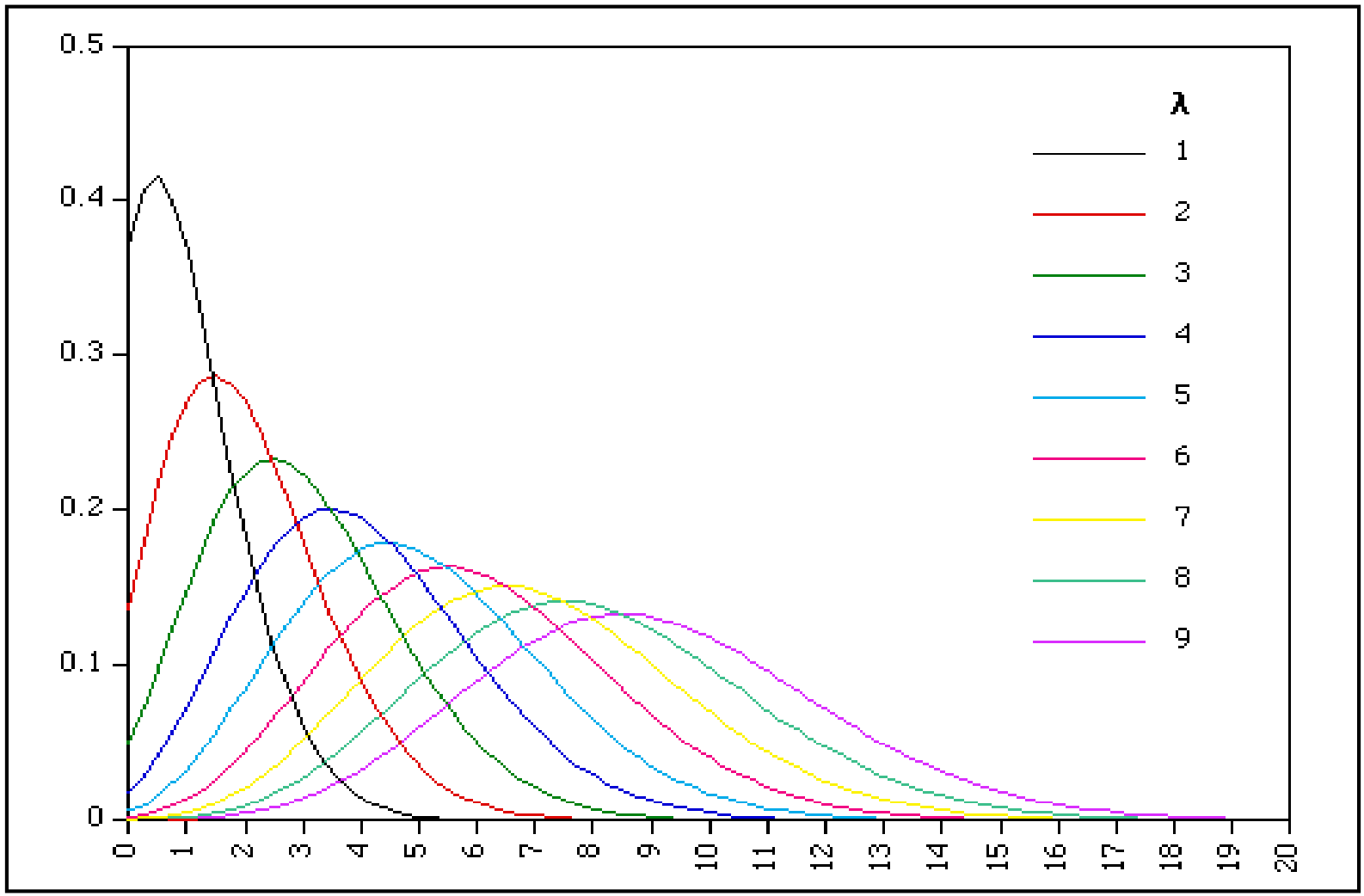
- is one type of generalized linear model
- assumes that the response variable follows a Poisson distribution
- can be used to model the number of occurrences of an event of interest or the rate of occurrence of an event of interest as a function of some predictor variables
- is most appropriate for rare events
- has the log as the canonical link function.

# Poisson Regression Outcome Variables

The following are examples of Poisson regression:

- number of ear infections in infants
- number of equipment failures
- homicide rates
- rate of insurance claims
- number of infected areas per unit volume of a tree
- number of organisms per unit area.

# Poisson Distributions with Different Means



# Poisson Regression Model without Random Effects

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

$$\mu = e^{(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}$$

$$= e^{\beta_0} \cdot e^{\beta_1 X_1} \cdots e^{\beta_k X_k}$$

# Poisson Regression Model with Random Effects

$$\log(\mu \mid \gamma) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \gamma_1 Z_1 + \ldots + \gamma_q Z_q$$

$$\mu \mid \gamma = e^{(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \gamma_1 Z_1 + \ldots + \gamma_q Z_q)}$$

$$= e^{\beta_0} \cdot e^{\beta_1 X_1} \cdots e^{\beta_k X_k} \cdot e^{\gamma_1 Z_1} \cdots e^{\gamma_q Z_q}$$

# Poisson Regression Parameter Estimates

$e^{\hat{\beta}} =$ multiplicative effect on conditional mean $\hat{\mu}$ for a one-unit change in X.

For example:

$e^{\hat{\beta}} =$ 1.20, then a one-unit increase in X yields a 20% increase in the estimated conditional mean.

$e^{\hat{\beta}} =$ 0.80, then a one-unit increase in X yields a 20% decrease in the estimated conditional mean.

# Female Horseshoe Crab Example

# Female Horseshoe Crab Example

The data comes from a study on the mating habits of female horseshoe crabs. Each female horseshoe crab had a male crab resident in her nest.

The study investigated factors affecting whether the female horseshoe crab had any other males, named *satellites,* residing nearby.

The response variable for each female horseshoe crab is her number of satellites.

**Satellites:** the number of satellites or male horseshoe crabs residing nearby

**Color:** female horseshoe crab's color (light medium, dark medium, dark)

**Spine:** spine condition (both good, one work/broken, both worn/broken)

**Width:** carapace width in centimeters

**Weight:** weight in kilograms

**Site:** randomly chosen sites where the observations were obtained.

# The Data

| width | weight | color | spine | satellites | site |
|---|---|---|---|---|---|
| 28.3 | 3.05 | 2 | 3 | 8 | 1 |
| 22.5 | 1.55 | 3 | 3 | 0 | 1 |
| 26.0 | 2.30 | 1 | 1 | 9 | 1 |
| 24.8 | 2.10 | 3 | 3 | 0 | 1 |
| 26.0 | 2.60 | 3 | 3 | 4 | 1 |
| 23.8 | 2.10 | 2 | 3 | 0 | 1 |
| 26.5 | 2.35 | 1 | 1 | 0 | 1 |
| 24.7 | 1.90 | 3 | 2 | 0 | 1 |
| 23.7 | 1.95 | 2 | 1 | 0 | 1 |
| 25.6 | 2.15 | 3 | 3 | 0 | 1 |
| 24.3 | 2.15 | 3 | 3 | 0 | 1 |
| 25.8 | 2.65 | 2 | 3 | 0 | 1 |
| 28.2 | 3.05 | 2 | 3 | 11 | 2 |
| 21.0 | 1.85 | 4 | 2 | 0 | 2 |
| 26.0 | 2.30 | 2 | 1 | 14 | 2 |
| 27.1 | 2.95 | 1 | 1 | 8 | 2 |
| ... | | | | | |

**color**: 1=light medium 2=medium 3=dark medium 4=dark

**spine**: 1=both good 2=one worn or broken 3=both worn broken

# Fitting a Poisson Regression Model with and without Random Effects

This demonstration illustrates the concepts discussed previously.

For Poisson regression with random effects using the default estimation method, the value for Gener. Chi-Square / DF should be close to 1 to indicate a good model fit.

- ❍ True
- ❍ False

**crabexample**

# crabexample results

| No random Effect | Random Effect | | Random Effect |
|---|---|---|---|
| **ML** | **Pseudo Likelihood** | **Method** | **ML** |
| 3.24 | N/A | Model Fit | 2.65 |
| N/A | 2.84 | Var(Residual) | N/A |
| 0.0292* | 0.2217 | Pvalue for testing color | 0.2174 |
| 0.4239 | 0.6447 | Pvalue for testing spine | 0.6404 |
| 0.0033* | 0.0025* | Pvalue for testing weight | 0.0025* |
| 0.7325 | 0.8949 | Pvalue for testing width | 0.8918 |
| | 2.6617945 | Var(Pearson residual) | |
| 921.76 | | AICc | 894.39 |
| 946.11 | | BIC | 898.37 |

* Significant test

# Overdispersion

- Poisson regression models assume that, given random effects, the variance is equal to the mean.
- The variability might exceed the mean for count data.
- Overdispersion leads to underestimates of the standard errors of parameter estimates.
- Overdispersion results in overestimates of the test statistic and liberal *p*-values.

# Causes of Overdispersion

- Subject heterogeneity due to an under-specified model in terms of fixed and random effects
- Outliers in the data
- A positive correlation between the responses and a failure to model the correlation

Is overdispersion a problem in ordinary least squares regression?

- ❍ Yes
- ❍ No

# Correction for Overdispersion

- Be certain that you do not have erroneous data.
- Recheck your model to include all important variables.
- Investigate more complex random effect structures if the overdispersion arises from the correlations among the observations.
- Use the RANDOM _RESIDUAL_ statement in PROC GLIMMIX to model the extra scale parameter.
- Use the negative binomial distribution to model the overdispersion (DIST=NEGBIN option in the MODEL statement in PROC GLIMMIX).

# Adding a Scale Parameter

- For generalized linear models, adding a scale parameter (**`random _residual_;`**) does not change the parameter estimates. The addition of a scale parameter changes the standard errors of the parameter estimates.

- For generalized linear mixed models, adding a scale parameter (**`random _residual_;`**) changes the model. Therefore, the addition changes
  - parameter estimates and the standard errors
  - variances of random effects and residuals.

# Modeling Overdispersion by Adding a Scale Parameter

This demonstration illustrates the concepts discussed previously.

**crabexample**

# crabexample results after adding a scale parameter

| No random Effect | Random Effect | G-side |
|---|---|---|
| Random_residual_ | Random_residual_ | R-side |
| ML | Pseudo Likelihood | Method |
| 3.24 | N/A | Model Fit |
| N/A | 3.0 | Var(Residual) |
| 0.4173 | 0.5318 | Pvalue for testing color |
| 0.7663 | 0.8091 | Pvalue for testing spine |
| 0.0988 | 0.0797 | Pvalue for testing weight |
| 0.8918 | 0.8912 | Pvalue for testing width |
| | 0.9294659 | Var(Pearson residual) |
| 921.76 | | AICc |
| 946.11 | | BIC |

# Question

Which of the following is **false**?

a. In PROC GLIMMIX you use RANDOM _RESIDUAL_ to add the scale parameter for Poisson regression.

b. For Poisson regression without random effects, adding the scale parameter does ad hoc adjustment to the standard errors of fixed effects.

c. For Poisson regression with random effects, adding the scale parameter only does ad hoc adjustment to the standard errors of fixed effects.

d. Overdispersion might arise from correlations in the data. Therefore, you should try to model it rather than adjust it ad hoc.

STATISTICS
TEXAS A&M UNIVERSITY

# Negative Binomial Distribution

The negative binomial distribution

- is the distribution for count data that permits the variance to exceed the mean, given random effects

- enables the model to have greater flexibility in modeling the relationship between the mean and the variance of the response variable than the Poisson model, given random effects.

| **Response Variable** | **Distribution** | **Link Function** | **Variance Function** |
|---|---|---|---|
| count | negative binomial | natural log | $\mu + k\mu^2$ |

# Dispersion Parameter *k*

- The dispersion parameter *k* is not allowed to vary over observations.
- The limit case when the parameter *k* is equal to 0 corresponds to a Poisson regression model.
- When the parameter is greater than 0, overdispersion is evident and the standard errors will increase. The fitted values are similar, but the larger standard errors reflect the overdispersion uncaptured with the Poisson model.
- The pmf for the Negative Binomial Dispersion

$$f(y) = \frac{\Gamma(y+1/k)}{\Gamma(y+1)\Gamma(1/k)} * \frac{(ku)^y}{(1+ku)^{y+1/k}} \quad \text{for } y = 0, 1, 2, 3\ldots$$

# Dispersion Parameter *k*

- The dispersion parameter *k* is not allowed to vary over observations.
- The limit case when the parameter *k* is equal to 0 corresponds to a Poisson regression model.
- When the parameter is greater than 0, overdispersion is evident and the standard errors will increase. The fitted values are similar, but the larger standard errors reflect the overdispersion uncaptured with the Poisson model.

# Modeling Overdispersion by Using the Negative Binomial Distribution

This demonstration illustrates concepts discussed previously.

**crabexample**

# crabexample results after using Negative Binomial

| No random Effect | Random Effect | G-side |
|---|---|---|
| ML | Pseudo Likelihood | Method |
| 0.93 | N/A | Model Fit |
| N/A | 1.03 | Var(Residual) |
| 0.4285 | 0.4406 | Pvalue for testing color |
| 0.7408 | 0.7314 | Pvalue for testing spine |
| 0.0845 | 0.0341* | Pvalue for testing weight |
| 0.9825 | 0.9013 | Pvalue for testing width |
| | 0.9574525 | Var(Pearson residual) |
| 764.42 | | AICc |
| 791.70 | | BIC |

# Residual Analysis

For models dealing with discrete outcomes, be aware of the following circumstances:

- The residual plot usually shows some pattern due to the discreteness of the outcome.
- You can look for potential outliers in residual plots.
- If there are no random effects or correlated errors,
  - there are no distributional assumptions about residuals
  - model diagnostic statistics might be available in certain procedures.
- If there are random effects or correlated errors, the GLIMMIX procedure assumes that the residuals for the linearized pseudo-data follow a normal distribution.

# Questions

- In PROC GLIMMIX, when you specify a distribution, it implies a default link function unless you use the LINK= option to change it.
  - ❍ True
  - ❍ False

- In PROC GLIMMIX, when you specify a link function, it implies a default distribution. For example, specifying LINK=LOG implies a Poisson distribution.
  - ❍ True
  - ❍ False