ST3241 Categorical Data Analysis I
Generalized Linear Models

Introduction and Some Examples

Introduction

- We have discussed methods for analyzing associations in two-way and three-way tables.
- Now we will use models as the basis of such analysis.
- Models can handle more complicated situations than discussed so far.
- We can also estimate the parameters, which describe the effects in a more informative way.

Example: Challenger O-ring

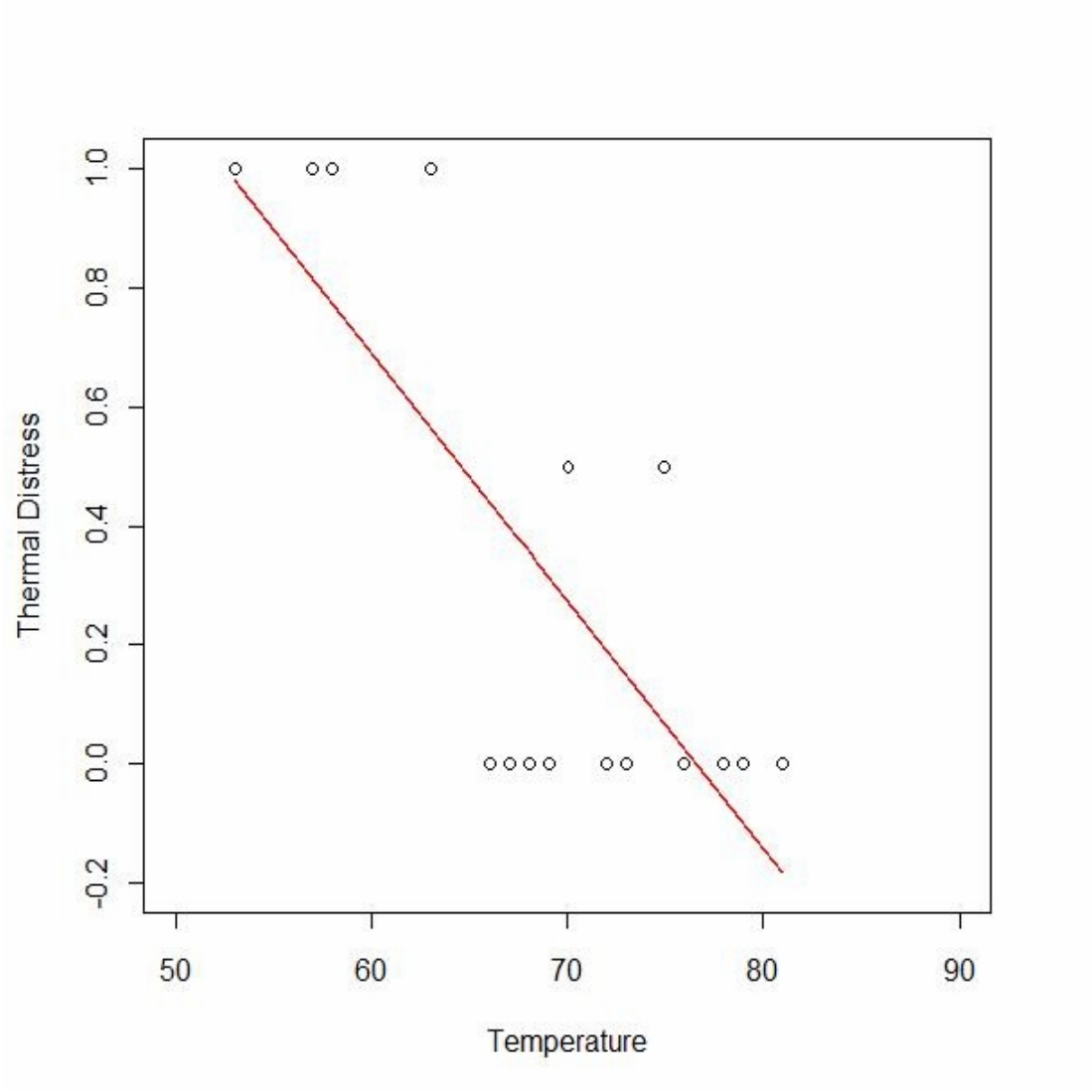
- For the 23 space shuttle flights that occurred before the Challenger mission disaster in 1986, the following table shows the temperature at the time of flight and whether at least one primary O-ring suffered thermal distress.

The Data

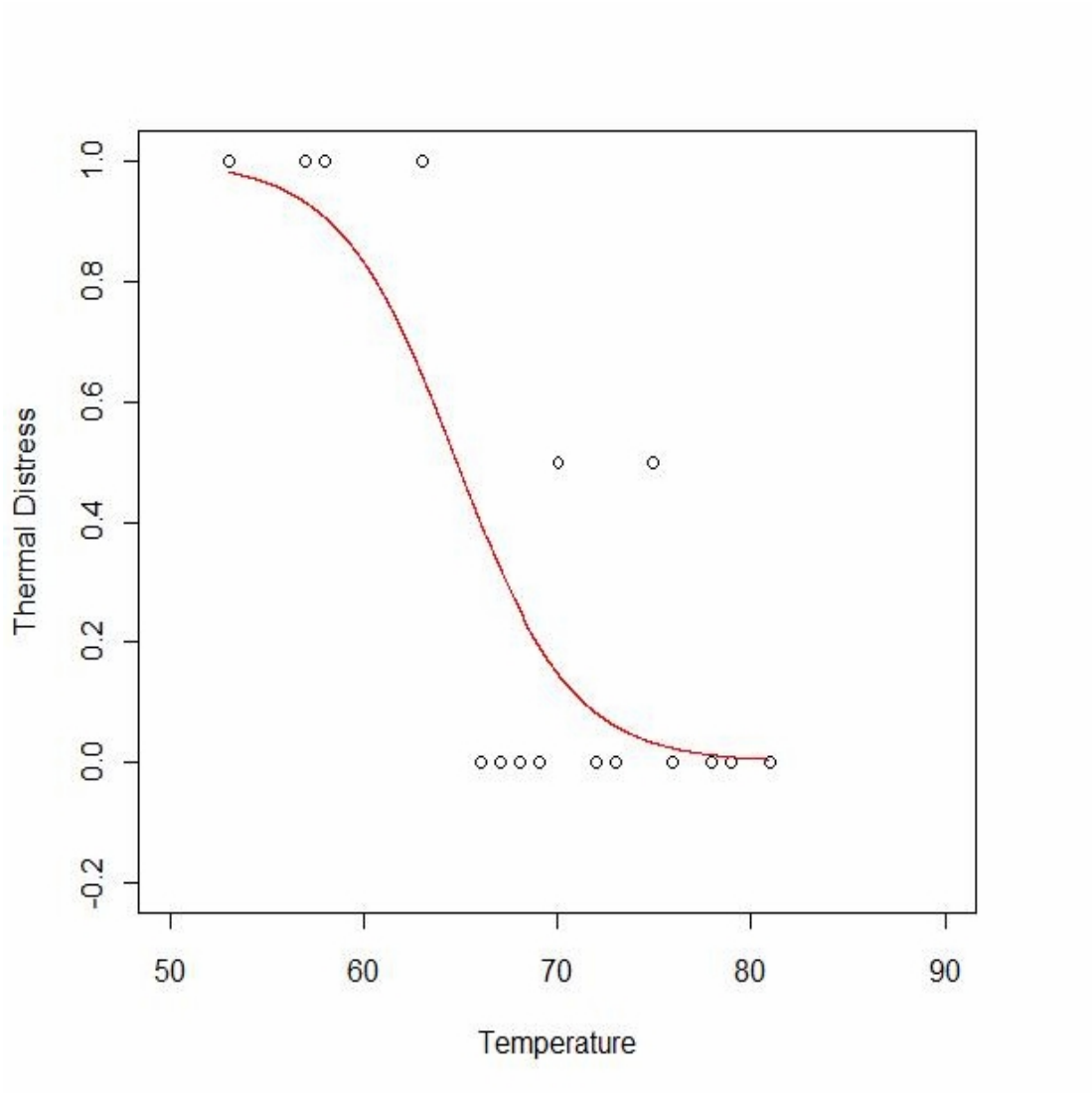
Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD
1	66	0	9	57	1	17	70	0
2	70	1	10	63	1	18	81	0
3	69	0	11	70	1	19	76	0
4	68	0	12	78	0	20	79	0
5	67	0	13	67	0	21	75	1
6	72	0	14	53	1	22	76	0
7	73	0	15	67	0	23	58	1
8	70	0	16	75	0			

- Is there any association between Temperature and thermal distress?

Fit From Linear Regression



Fit From Logistic Regression

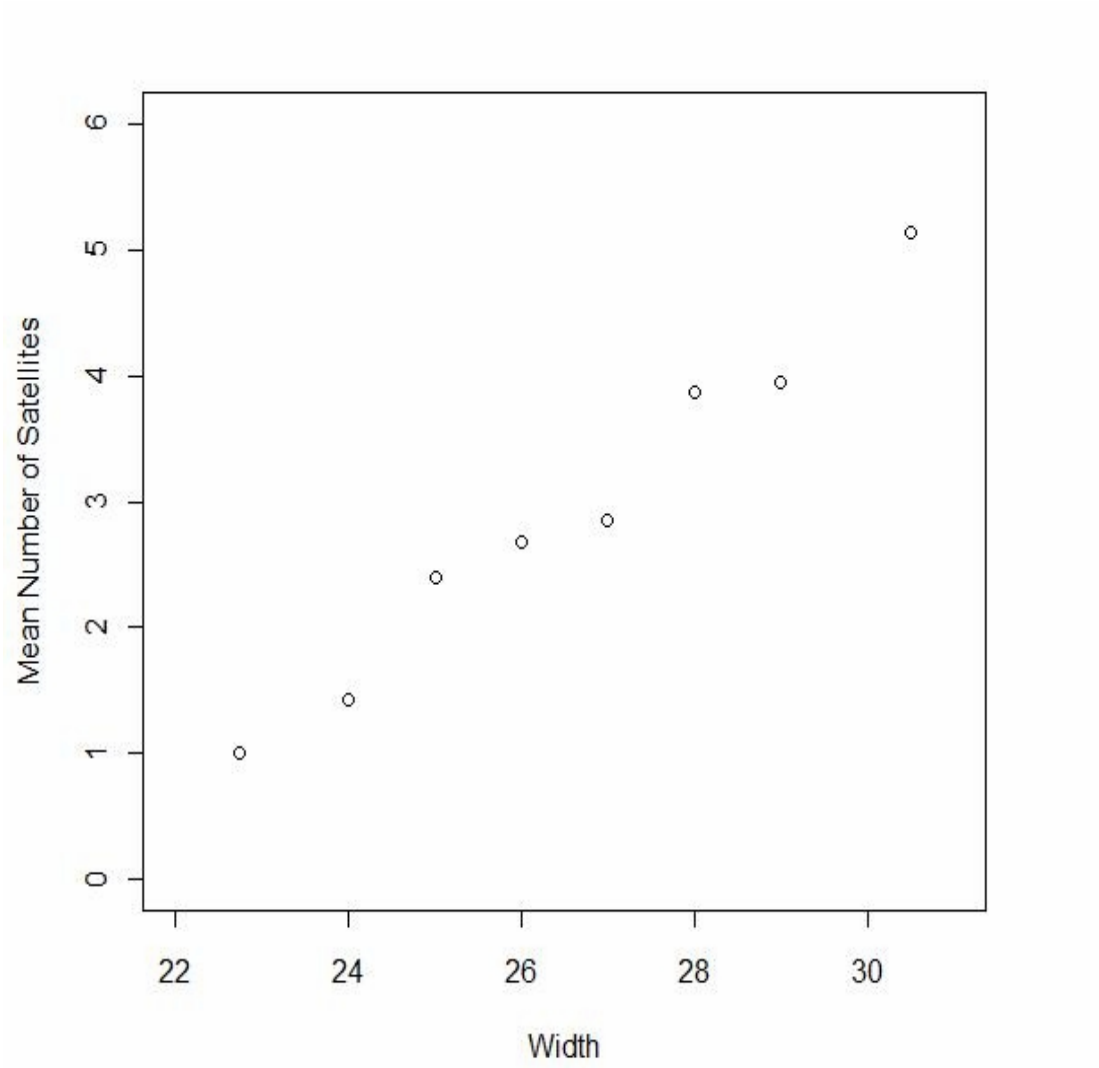


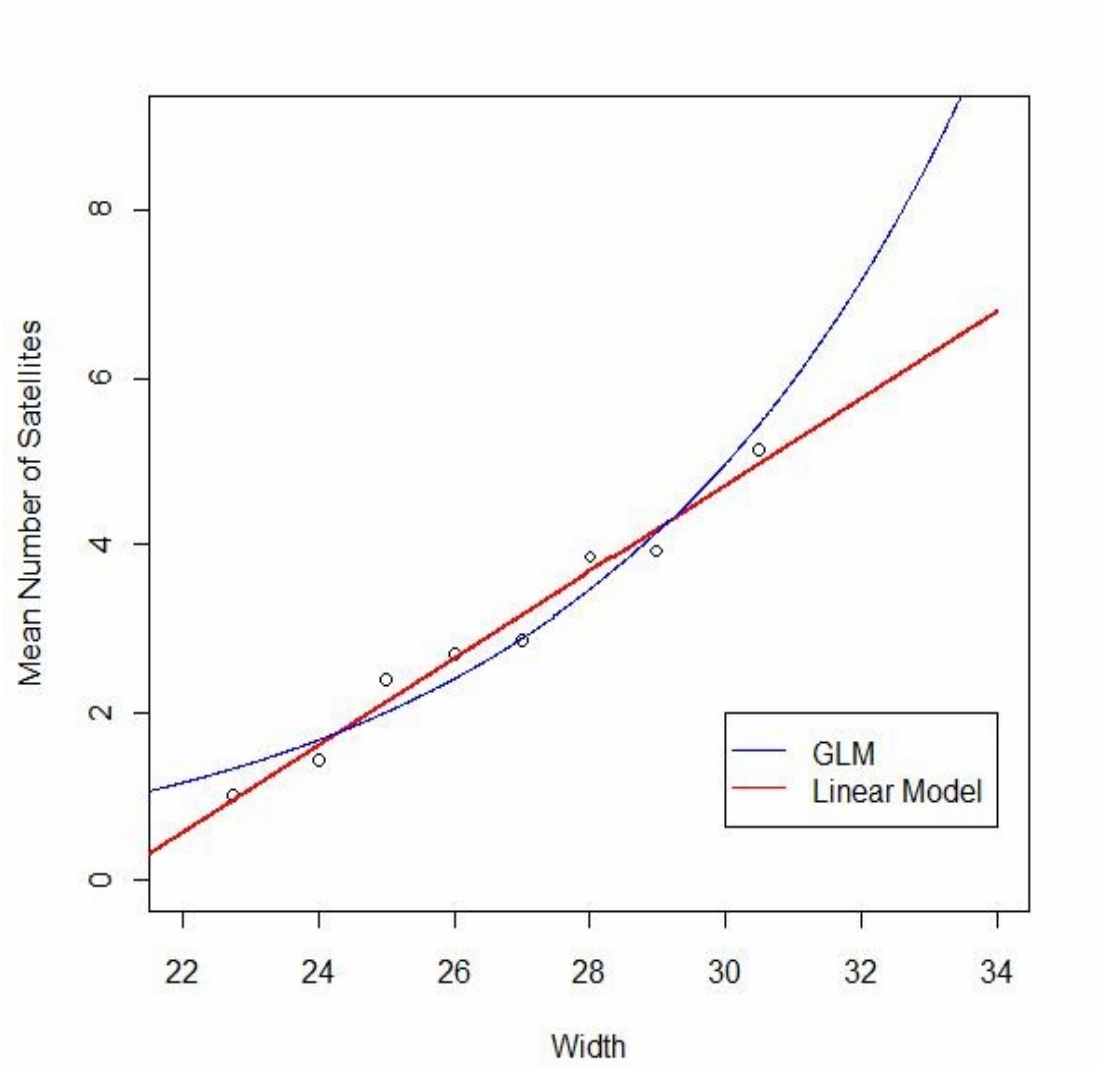
Example: Horseshoe Crabs

- Each female horseshoe crab in the crab in the study had a male crab attached to her in her nest.
- The study investigated factors that affect whether the female crab had any other males, called *satellites*, residing nearby her.
- Explanatory variables included the female crab's color, spine condition, weight, and carapace width.
- The response outcome for each female crab is her number of satellites.

Example Continued

- We consider the width alone as a predictor.
- To obtain a clearer picture, we grouped the female crabs into a set of width categories
- $\leq$ 23.25, 23.25-24.25, 24.25-25.25, 25.25-26.25, 26.25-27.25, 27.25-28.25,28.25-29.25, >29.25
- Calculated sample mean number of satellites for female crabs in each category.


Mean Number of Satellites
0
1
2
3
4
5
6
22
24
26
28
30
Width


Mean Number of Satellites
0
2
4
6
8
22
24
26
28
30
32
34
Width
GLM
Linear Model

Components of A GLM

- Random component
  - Identifies the response variable Y and assumes a probability distribution for it
- Systematic component
  - Specifies the explanatory variables used as predictors in the model
- Link
  - Describes the functional relation between the systematic component and expected value of the random component

Random Component

- Let $Y_1, \cdots, Y_N$ denote the N observations on the response variables Y.
- The random component specifies a probability distribution for $Y_1, \cdots, Y_N$.
- If the potential outcome for each observation Y_i are binary such as "*success*" or "*failure*"; or, more generally, each Y_i might be number of "successes" out of a certain fixed number of trials, we can assume a binomial distribution for the random component.
- If each response observation is a non-negative count, such as cell count in a contingency table, then we may assume a Poisson distribution for the random component.

Systematic Component

- The systematic component specifies the explanatory variables.
- It specifies the variables that play the roles of x_j in the formula $\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$.
- This linear combination of explanatory variables is called the linear predictors.
- Some x_j may be based on others in the model; for instance, perhaps $x_3 = x_1 x_2$, to allow interaction between x_1 and x_2 in their effects on Y, or perhaps $x_3 = x_1^2$ to allow a curvilinear effect of x_1.

Link

- It specifies how $\mu = E(Y)$ relates to explanatory variables in the linear predictor.
- The model formula states that

$$g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

The function $g(.)$ is called the *link function*.

Some Popular Link Functions

- Identity Link

$$g(\mu) = \mu = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

- Log link

$$g(\mu) = \log(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

- Logit

$$g(\mu) = \log[\frac{\mu}{1-\mu}] = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

More On Link Functions $\cdots$

- Each potential probability distribution has one special function of the mean that is called its natural parameter.
- For the normal distribution, it is mean itself.
- For the Poisson, the natural parameter is the log of the mean.
- For the Binomial, the natural parameter is the logit of the success probability.
- The link function that uses the natural parameter as $g(\mu)$ in the GLM is called the *canonical link*.
- Though other links are possible, in practice the canonical links are most common

GLM For Binary Data: Random Component

- The distribution of a binary response is specified by probabilities $P(Y = 1) = \pi$ of success and $P(Y = 0) = 1 - \pi$ of failure.
- For n independent observations on a binary response with parameter π, the number of successes has the binomial distribution specified by parameters n and π.

Linear Probability Model

- To model the effect of X, use ordinary linear regression, by which the expected value of Y is a linear function of X.
- The model

$$\pi(x) = \alpha + \beta x$$

  is called a linear probability model.
- Probabilities fall between 0 and 1 but for large of small values of x, the model may predict $\pi(x) < 0$ or $\pi(x) > 1$.
- This model is valid only for a finite range of x values

Example: Snoring

	Heart Disease			
Snoring	Yes	No	Proportion Yes	Linear Fit
Never	24	1355	0.017	0.017
Occasional	35	603	0.055	0.057
Nearly Every Night	21	192	0.099	0.096
Every Night	30	224	0.118	0.116

Ordinal Data

Example:

- We use (0,2,4,5) for their snoring categories, treating the last two levels closer.
- Linear Fit using maximizing likelihood:

$$\pi(x) = 0.0172 + 0.0198x$$

- The least squares fit is slightly different.

Logistic Regression Model

- Relationship between $\pi(x)$ and x are usually nonlinear rather than linear. The most important function describing this nonlinear relationship has the form

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

- That is,

$$\pi(x) = F_0(\alpha + \beta x), \text{ where } F_0(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}},$$

  where $F_0(x)$ is the cdf of the logistic distribution. Its pdf is $F_0(x)(1 - F_0(x))$.

- The associated GLM is called the *logistic regression function*.

- Logistic regression models are often referred as *logit* models as the link in this GLM is the *logit* link.

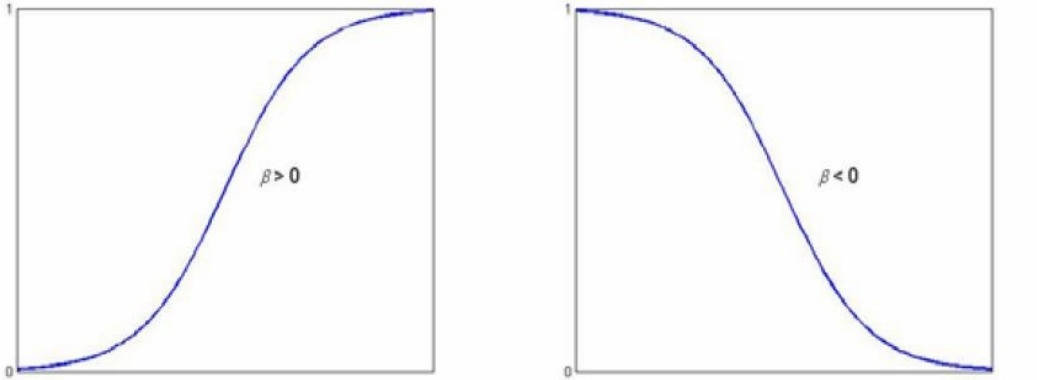
Parameters

- The parameter β determines the rate of increase or decrease of the curve.
- When $\beta > 0$, $\pi(x)$ increases with x.
- When $\beta < 0$, $\pi(x)$ decreases as x increases.
- The magnitude of β determines how fast the curve increases or decreases.
- As $|\beta|$ increases, the curve has a steeper rate of change.

Example: Snoring

Snoring	Heart Disease		Proportion	Linear	Logit
	Yes	No	Yes	Fit	Fit
Never	24	1355	0.017	0.017	0.021
Occasional	35	603	0.055	0.057	0.044
Nearly Every Night	21	192	0.099	0.096	0.093
Every Night	30	224	0.118	0.116	0.132

Effect of Parameters



Alternative Binary Links

- For logistic regression curves, the probability of a success increases or decreases continuously as x increases.
- Let X denote a random variable, the cumulative distribution function (cdf) $F(x)$ is defined as

$$F(x) = P(X \leq x), \ -\infty < x < \infty$$

- Such a function, plotted as a function of x, has appearance like that of the logistic function in the previous figures.
- It suggests a class of models for binary responses of the form

$$\pi(x) = F(\alpha + \beta x)$$

where F is a cdf for some distribution.

Alternative Binary Links

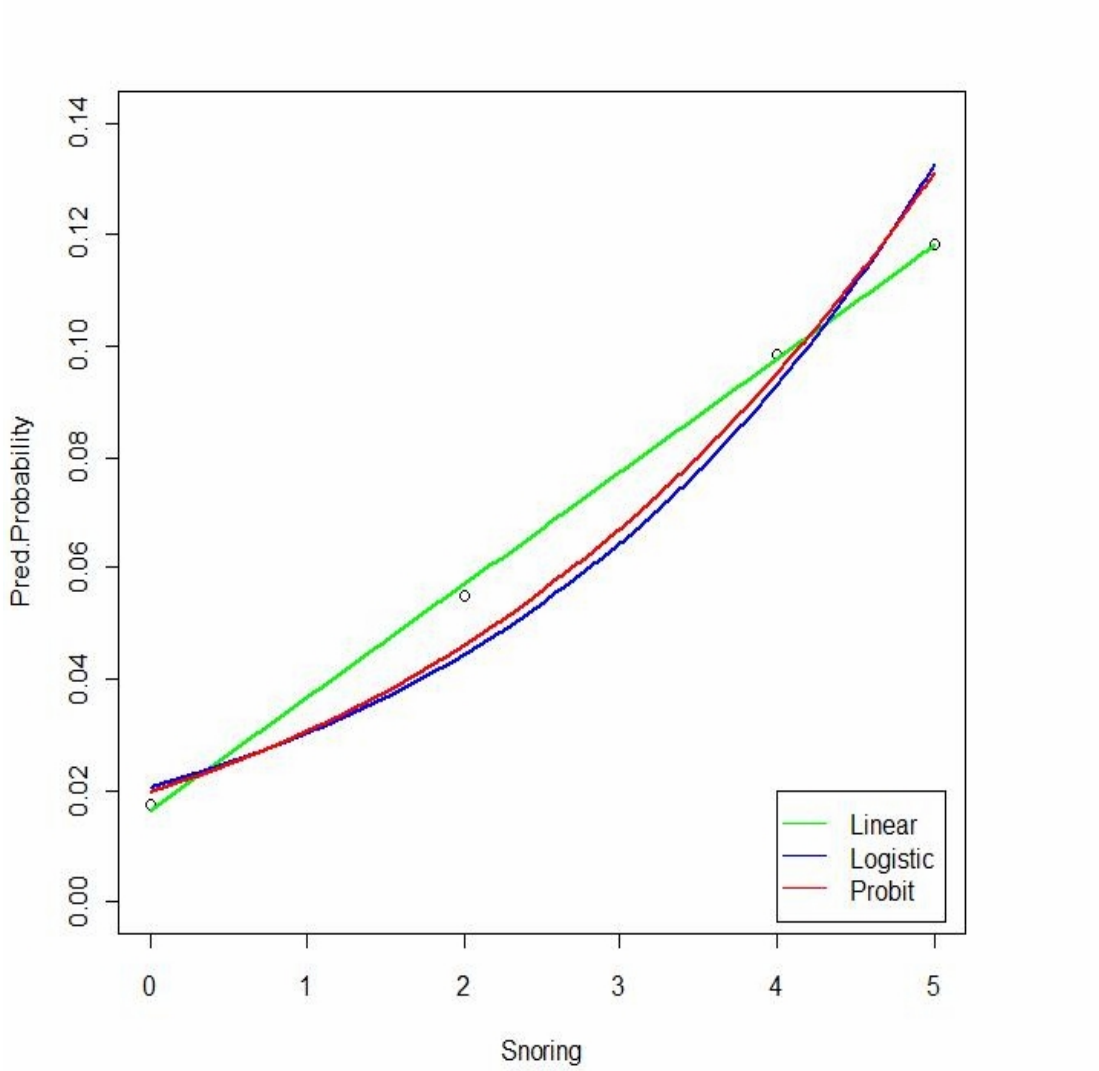
- The logistic regression curve has this form.
- When $\beta > 0$, $\pi(x) = F(\alpha + \beta x)$ has the shape of the cdf of the two-parameter logistic distribution.
- When $\beta < 0$, $1 - \pi(x) = 1 - F(\alpha + \beta x)$ has the shape of the cdf of the two-parameter logistic distribution.
- Each choice of α and $\beta > 0$ corresponds to a different logistic distribution.
- The logistic cdf $F_0(x)$ corresponds to a probability distribution $F_0(x)(1 - F_0(x))$ with a symmetric, bell shape and very similar looking to a normal distribution.

Probit Models

- The probability of success, $\pi(x)$, has the form $\Phi(\alpha + \beta x)$ where Φ is the cdf of a standard normal distribution $N(0, 1)$.
- The link function is known as probit link: $g(\pi) = \Phi^{-1}(\pi)$.
- The *probit transform* maps $\pi(x)$ so that the regression curve for $\pi(x)$ (or $1 - \pi(x)$, when $\beta < 0$) has the appearance of the normal cdf with mean $\mu = -\alpha/\beta$ and standard deviation $\sigma = 1/|\beta|$.

Example: Snoring

Snoring	Heart Disease		Proportion	Linear	Logit	Probit
	Yes	No	Yes	Fit	Fit	Fit
Never	24	1355	0.017	0.017	0.021	0.020
Occasional	35	603	0.055	0.057	0.044	0.046
Nearly Every Night	21	192	0.099	0.096	0.093	0.095
Every Night	30	224	0.118	0.116	0.132	0.131


Pred.Probability
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0
1
2
3
4
5
Snoring
Linear
Logistic
Probit

GLM for Count Data

- Many discrete response variables have counts as possible outcomes.
  - For a sample of cities worldwide, each observation might be the number of automobile thefts in 2003.
  - For a sample of silicon wafers used in computer chips, each observation might be the number of imperfections on a wafer.
- We have earlier seen the Poisson distribution as a sampling model for counts.

Poisson Regression

- Assume a Poisson distribution for the random component.
- One can model the Poisson mean using the *identity link*.
- But more common to model the *log* of the mean.
- A Poisson *loglinear model* is a GLM that assumes a Poisson distribution for Y and uses the $log - link$.

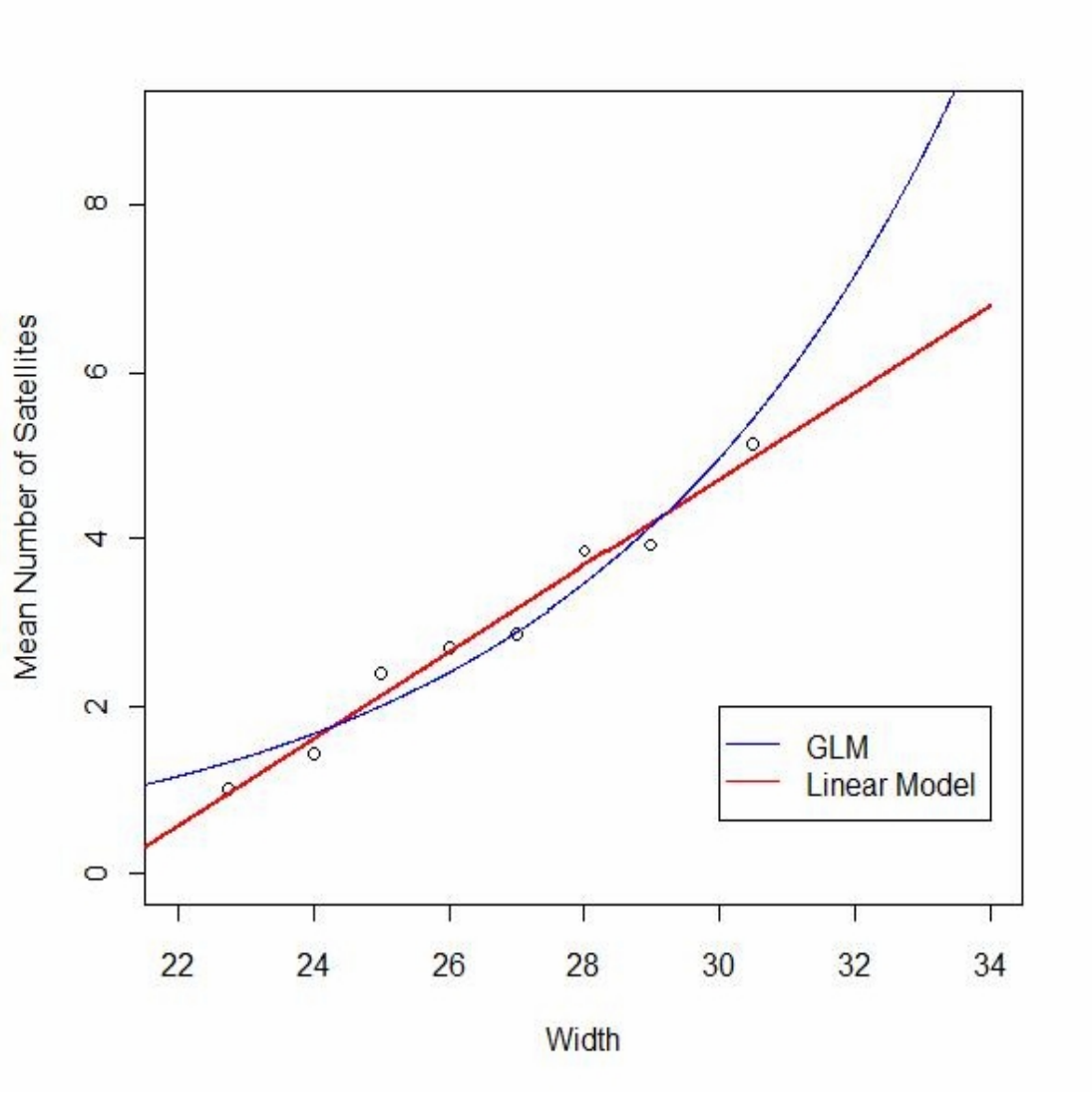
Poisson Regression - Continued

- Let μ denote the expected value of Y and let X denote an explanatory variable.
- Then the Poisson log-linear model has the form

$$\log(\mu) = \alpha + \beta x$$

- For this model: $\mu = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x}$

Example: Horseshoe Data



Poisson Regression For Rate Data

- It is often relevant to certain types of events which occur over time, space, or some other index of size, to model the *rate* at which events occur.
- In modeling numbers of auto thefts in 2003 for a sample of cities, we could form a rate for each city by dividing the number of thefts by the city's population size.
- The model describes how the *rate* depends on some other explanatory variables.

Poisson Regression For Rate Data

- When a response count Y_i has index (such as population size) equal to t_i, the sample rate of outcomes is Y_i/t_i.
- The expected value of the rate is μ_i/t_i.
- A log-linear model for the expected rate has form:

$$\log(\mu_i/t_i) = \alpha + \beta x_i$$

- This has an equivalent representation:

$$\log \mu_i - \log t_i = \alpha + \beta x_i$$

- The adjustment term, $-\log t_i$, to the above equation is called an $offset$.

Exponential Family

- The random variable Y has a distribution in the exponential family, if its *p.d.f* (or *p.m.f*.) can be written as

$$f(y;\theta,\phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y,\phi)\}$$

  for some specific function $a(\phi)$, $b(\theta)$ and $c(y,\phi)$.

- The parameter θ is called the *natural parameter* and ϕ is called the *dispersion* (or *scale*) *parameter*.

Examples: Normal Distribution

- The $p.d.f.$ of $N(\mu, \sigma^2)$

$$f(y;\theta,\phi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(y-\mu)^2/(2\sigma^2)\}$$
$$= \exp\{(y\mu - \mu^2/2)/\sigma^2 - (y^2/\sigma^2 + \log(2\pi\sigma^2))/2\}$$

- Here, $\theta = \mu$, $\phi = \sigma^2$, and $a(\phi) = \phi$, $b(\theta) = \theta^2/2$ and $c(y,\phi) = -\{y^2/\sigma^2 + \log(2\pi\sigma^2)\}/2$
- The canonical link: $g(\mu) = \mu$.

Examples: Binomial Distribution

- The $p.d.f.$ of Bernoulli(π):

$$f(y;\theta,\phi) = \pi^y(1-\pi)^{1-y} = (1-\pi)(\frac{\pi}{1-\pi})^y$$
$$= \exp\{y\log(\frac{\pi}{1-\pi}) - \log(\frac{1}{1-\pi})\}$$

- Here $\theta = \log(\frac{\pi}{1-\pi})$,$\phi = 1$,$b(\theta) = \log(1+e^\theta)$
  $a(\phi) = 1$, $c(y,\phi) = 0$
- The canonical link: $g(\pi) = \log(\frac{\pi}{1-\pi})$.

Examples: Poisson Distribution

- The *p.d.f.* of Poisson(λ):

$$f(y;\theta,\phi) = e^{-\lambda}\frac{\lambda^y}{y!}$$
$$= \exp\{(y\log\lambda - \lambda) - \log y!\}$$

- Here $\theta = \log\lambda$,$\phi = 1$, $b(\theta) = e^{\theta}$,
  $a(\phi) = 1$, $c(y,\phi) = -\log y!$
- The canonical link: $g(\lambda) = \log(\lambda)$.

Log-Likelihood Functions

- The log-likelihood function

$$l(\theta, \phi; y) = \log f(y; \theta, \phi) = (y\theta - b(\theta))/a(\phi) + c(y, \phi)$$

- We use general likelihood results, applicable to exponential families $E(\frac{\partial l}{\partial \theta}) = 0$ and $E(\frac{\partial l}{\partial \theta})^2 = -E(\frac{\partial^2 l}{\partial \theta^2})$
- Here
  $\frac{\partial l}{\partial \theta} = (y - b'(\theta))/a(\phi)$ and $\frac{\partial^2 l}{\partial \theta^2} = -b''(\theta)/a(\phi)$

Mean and Variances

- Now, we have $0 = E(\frac{\partial l}{\partial \theta}) = \{E(y) - b'(\theta)\}/a(\phi)$
- So that, $E(Y) = b'(\theta)$.
- Similarly,

$$\frac{var(Y)}{a^2(\phi)} = \frac{b''(\theta)}{a(\phi)}$$

- So that, $var(Y) = b''(\theta)a(\phi)$.

Examples

- Normal Distribution: $b(\theta) = \theta^2/2$
  - $E(Y) = b'(\theta) = \theta = \mu$
  - $var(Y) = b''(\theta)a(\phi) = \phi = \sigma^2$.
- Bernoulli Distribution: $b(\theta) = \log(1 + e^\theta)$
  - $E(Y) = b'(\theta) = e^\theta/(1 + e^\theta) = \pi$.
  - $var(Y) = b''(\theta) = e^\theta/(1 + e^\theta)^2 = \pi(1 - \pi)$
- Poisson Distribution: $b(\theta) = \exp(\theta)$
  - $E(Y) = b'(\theta) = \exp(\theta) = \lambda$.
  - $var(Y) = b''(\theta) = \exp(\theta) = \lambda$.

Likelihood Equations in GLMs

- Let $(y_1, \cdots, y_N)$ denote responses for N independent observations.
- Let $(x_{i1}, \cdots, x_{ip})$ denote values of p explanatory variables for observation i.
- The systematic component for the i-th observation

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$$

- If $E(y_i) = \mu_i$, then the link for the i-th observation is:

$$\eta_i = g(\mu_i)$$

Likelihood Equations in GLMs

- For N independent observations, the log-likelihood function is:
$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N} \log f(y_i; \boldsymbol{\theta}, \phi) = \sum_{i=1}^{N} \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{N} c(y_i, \phi)$$
- The notation $L(\boldsymbol{\beta})$ reflects the dependence of $\boldsymbol{\theta}$ on the model parameters $\boldsymbol{\beta}$.

Likelihood Equations in GLMs

- The likelihood equations are:

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial L_i}{\partial \beta_j} = 0,$$

  for $j = 1, \cdots, p$.

- Simplifying, we have

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i) x_{ij}}{var(y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0$$

  for $j = 1, \cdots, p$. Notice that $\frac{\partial \mu_i}{\partial \eta_i} = 1/g'(\mu_i)$.

Examples

- Logit Model:

$$\sum_{i=1}^{N}(y_i - \pi_i)x_{ij} = 0, \text{ where } \pi_i = \frac{exp(\sum_{j=1}^{p}\beta_j x_{ij})}{1+exp(\sum_{j=1}^{p}\beta_j x_{ij})}$$

- Probit Model:

$$\sum_{i=1}^{N}\frac{(y_i-\pi_i)x_{ij}}{\pi_i(1-\pi_i)}\phi(\sum_{k=1}^{p}\beta_k x_{ik}) = 0, \text{ where } \pi_i = \Phi(\sum_{j=1}^{p}\beta_j x_{ij})$$

- Log-Linear Model:

$$\sum_{i=1}^{N}(y_i - exp(\sum_{k=1}^{p}\beta_k x_{ik}))x_{ij} = 0$$

Maximum Likelihood Estimates

- ML estimates of β_j's are obtained by solving the likelihood equations using numerical methods.
- The ML estimates $\hat{\beta}_j$'s are approximately normally distributed.
- Thus, a confidence interval for a model parameter β_j equals

$$\hat{\beta}_j \pm z_{\alpha/2} ASE$$

where ASE is the asymptotic standard error of $\hat{\beta}_j$.

Testing For Significance

- To test: $H_0 : \beta_j = 0$.
- The test statistic, $Z = \hat{\beta}_j / ASE$ has an approximate standard normal distribution, when H_0 is true.
- Equivalently, Z^2 has a chi-squared distribution with $d.f. = 1$, which can be used for two-sided alternatives
- This type of test is known as Wald's test.

Likelihood Ratio Test

- The likelihood-ratio test statistic equals

$$-2\log(L_0/L_1) = -2[\log L_0 - \log L_1] = -2[l_0 - l_1]$$

  where L_0 and L_1 are the maximized likelihood functions under the null hypothesis and under the full model, respectively.

- Under H_0, this test statistic also has a large sample chi-squared distribution with $d.f. = 1$.

  Uses more information, higher power than score test

Score Test

- The score statistic or efficient score statistic uses the size of the derivative of the log-likelihood function evaluated at $\beta_j = 0$.
- The score statistic is the square of the ratio of this derivative to its ASE.
- It also has an approximate chi-squared distribution.

Example

- In simple logistic regression with one explanatory variable, the log-likelihood function is:

$$l(\alpha, \beta) = \sum_{i=1}^{N} \{y_i(\alpha + \beta x_i) - \log(1 + \exp(\alpha + \beta x_i))\}$$

Test if ß is correlated with y

- Therefore for $H_0 : \beta = 0$ and $H_1 : \beta \neq 0$, we have

$$\begin{aligned} l_0 &= \sum_{i=1}^{N} \{y_i \log \tfrac{\bar{y}}{1-\bar{y}} - \log \tfrac{1}{1-\bar{y}}\}, \\ l_1 &= \sum_{i=1}^{N} \{y_i(\hat{\alpha} + \hat{\beta} x_i) - \log(1 + \exp(\hat{\alpha} + \hat{\beta} x_i))\} \end{aligned}$$

Model Residuals

- For i-th observation, the raw residual is:

$$r_i = y_i - \hat{\mu}_i = \text{Observed} - \text{fitted},$$

  where y_i is the observed response and $\hat{\mu}_i$ is the fitted value from the model.

- The Pearson residual is defined as
  $\text{Pearson residual} = \frac{\text{Oberved-fitted}}{\sqrt{\hat{var}(\text{observed})}} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{var}(y_i)}}$.

- For Poisson GLMs, it simplifies to
  $e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$

Adjusted Residuals

- The Pearson residuals divided by its estimated standard error are called *adjusted residuals*.
- Adjusted residuals have an approximate standard normal distribution.
- For Poisson GLMs, the general form of the adjusted residual is:

$$\frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\mu}_i(1-h_i)}} = \frac{e_i}{\sqrt{1-h_i}}$$

where h_i is called the leverage of observation i.