

Analysis of Counts and Proportions

Bailey and Gatrell Chapter 8

Lecture 20

November 19, 2009

1

Models of Area Data

So far we have assumed distributions were approximately normal

Occasionally need to consider alternate distributions of the response variable

Data collected over areas often take the form of counts or proportions

Such data are more likely to have binomial or Poisson distributions

The problem with such data is that the variance is related to the mean

2

Visualization and exploration of rates and counts

Rates

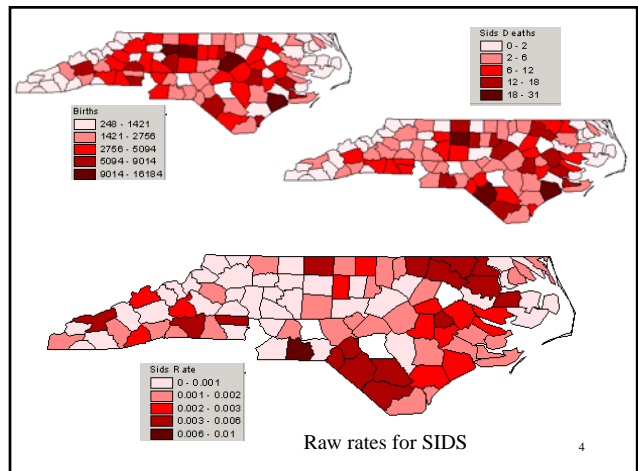
Ratio of number of cases to number of individuals at risk of being a case

Ratio of observed to expected – Standardized mortality ratio (SMR)

Rate maps can be misleading since the variability of the rates is a function of the underlying population

- areas with small number of expected cases are likely to have high or low rates
- areas with large number of expected cases are likely to look average.

3



4

Visualization and exploration of rates and counts

Want to highlight anomalies (higher or lower than expected rates) given the variation in the underlying population in each area

One approach for health related data is to map measures of **relative risk**

y_i is count (of cases) in some area, n_i is the population (at risk) for the area

5

Relative Risk

Assume the overall rate of occurrence is constant for all areas and all areas are independent

Then we can assume counts y_i are observations on an independent Poisson random variable with an expected mean value μ_i

$$\hat{\mu}_i = n_i \left(\frac{\sum y_i}{\sum n_i} \right)$$

The **relative risk** is the observed count divided by the estimated expected value, μ_i multiplied by 100

$$\hat{r}_i = \frac{y_i}{\hat{\mu}_i} * 100$$

6

Relative Risk

$$\text{Relative risk} = \frac{y_i}{\hat{\mu}_i} * 100$$

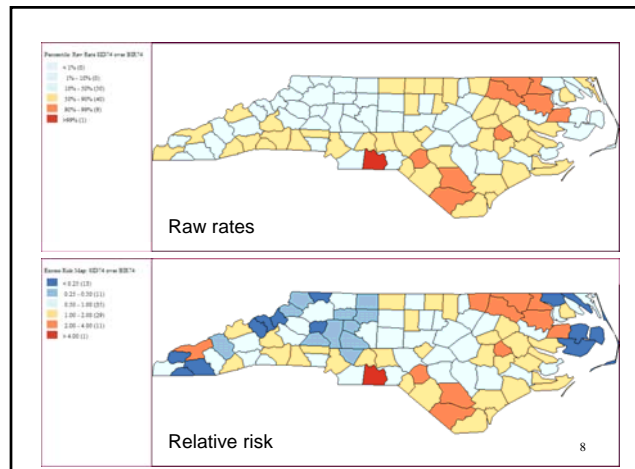
In case of disease, indicates whether an observed rate exceeds an expected disease rate.

Values near 100 are not much different than expected

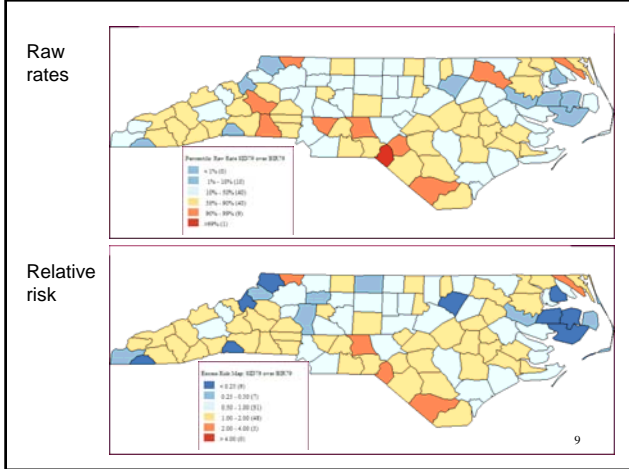
Values less than 100 are smaller than expected

Values larger than 100 are greater than expected

7



8



Relative Risk

The relative risk statistic is sensitive to small values.

Example:

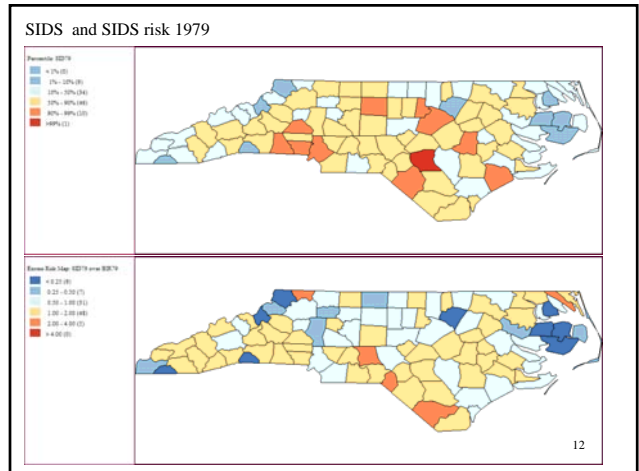
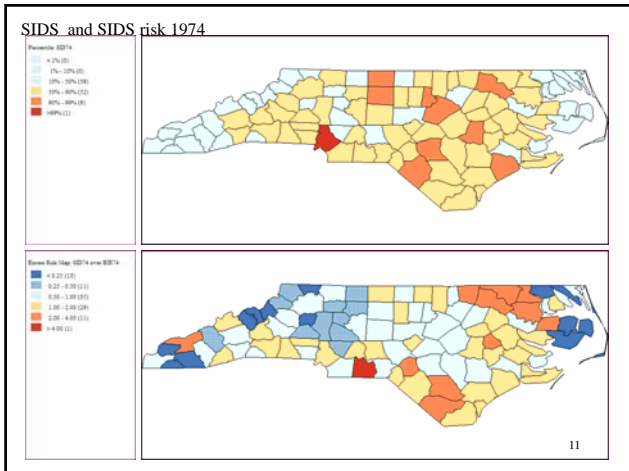
Swain County had 675 births over the period and 3 deaths from SIDS. The standardized risk is 2.20, quite high.

If there were only 2 cases:

The risk is only 1.46.

Four cases raises the risk to 2.93.

10



Probability mapping

An alternative to mapping relative risk is to map the probability of getting a count that is more extreme than the observed count

Counts are assumed to be Poisson distributed with mean value μ_i

$$p_i = \begin{cases} \sum_{x \geq y_i} \frac{\hat{\mu}_i^x e^{-\hat{\mu}_i}}{x!} & y_i \geq \hat{\mu}_i \\ \sum_{x \leq y_i} \frac{\hat{\mu}_i^x e^{-\hat{\mu}_i}}{x!} & y_i < \hat{\mu}_i \end{cases}$$

Small values of p (< 0.05) indicate an unusually high or low rate

13

Probability mapping

Objective is to standardize rates onto a probability scale for proper comparison.

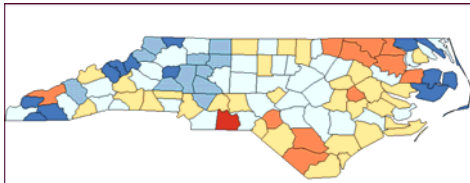
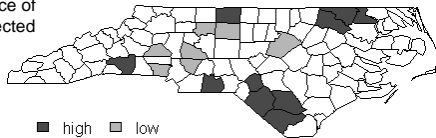
Transforms data into probability of deviations from homogeneity.

Not useful for very widely varying populations (eg. number of live births)

14

Probability mapping

Statistical significance of departure from expected



15

Probability mapping

Breast Cancer Incidence, Relative Risks
Not Age-Adjusted

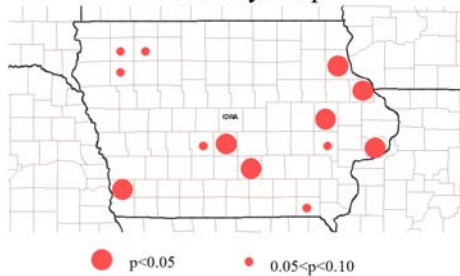


16

Probability mapping

Statistical significance of departure from expected

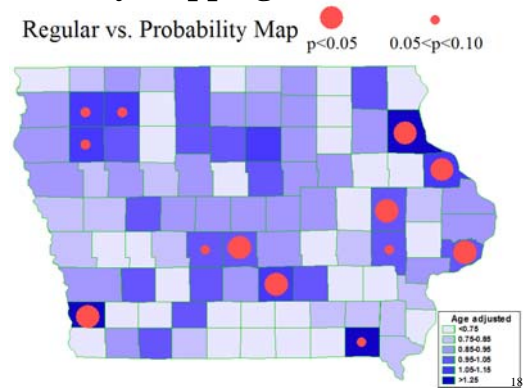
Probability Map



17

Probability mapping

Regular vs. Probability Map



18

Probability mapping

The model underlying the probability map contrasts spatially independent areas with heterogeneous mean against spatially independent areas with homogeneous means

There is no consideration of spatial dependence between areas

Extreme values could be the result of lack of fit to spatially independent model rather than heterogeneous rates

How to improve?

19

Empirical Bayes Estimation

Bayesian statistics - statistical estimation where prior knowledge or beliefs about parameters of interest are taken into account as well as observed data when estimating their values

An unconditional prior probability distribution for values of a parameter of interest is converted to a posterior distribution for the values of that parameter using data actually observed

Bayes theorem derives a posterior distribution by combining the likelihood for the data with the prior distribution

20

Bayes Formula

Let A_1, A_2, \dots, A_k be mutually exclusive events with probabilities of occurrence $P(A_j)$. These are called the prior probabilities.

Let there be m possible observable events B_1, B_2, \dots, B_m . The probability of observing B_j is conditional upon the underlying state of nature A_i . The conditional probabilities of $P(B_j/A_i)$ are known and are called the likelihoods.

We want to calculate posterior probabilities $P(A_i/B)$

$$P(A_i | B) = \frac{P(B | A_i) * P(A_i)}{P(B)}$$

Intuitively, says one's beliefs about observing 'A' are updated by having observed 'B'

23

Empirical Bayes Estimation

Empirical Bayesian estimation refers to the case when the prior distribution is based on certain global aspects of the data

In this case information on the overall rate across all areas

Assume the true but unknown rate for each area is θ_i

$r_i = \frac{y_i}{n_i}$ is the observed rate

The non-Bayesian best estimate of θ_i is just r_i

22

Empirical Bayes Estimation

Consider the North Carolina SIDS data.

We wish to identify counties with anomalously high rates of SIDS.

The observed count data are not sufficient to determine whether a county has a high risk for SIDS. The proportion of deaths to births is not sufficient, especially when the raw number of deaths or births is small

We would like to account for the variability of small changes in the raw numbers when assessing the relative risk for a county.

23

Empirical Bayes Estimation

Suppose we have a prior probability distribution for each θ_i with mean γ_i and variance ϕ_i

The best Bayes estimate of θ_i is then based on combining these prior distributions with the observed rates

$$\hat{\theta}_i = w_i r_i + (1 - w_i) \gamma_i \quad \text{Known as shrinkage estimate}$$

where

$$w_i = \frac{(\phi_i)}{(\phi_i + \gamma_i / n_i)} \quad \text{A function of the population at risk and the variance of the prior distribution}$$

24

Empirical Bayes Estimation

$$\hat{\theta}_i = w_i r_i + (1 - w_i) \gamma_i \quad w_i = \frac{(\phi_i)}{(\phi_i + \gamma_i / n_i)}$$

As w_i approaches 1, weight goes to the observed rate - as it approaches 0, weight goes to the prior mean

If the population is large, we do not shrink the estimate of the area toward the prior so much - there is more confidence in the observed rate

If the population is small there is less confidence in the observed rate so the estimate moves more toward the prior distribution.

25

Empirical Bayes Estimation

Where does the prior distribution information come from?

Must be estimated from the data with some simplifying assumptions

Assume all prior means and variances are the same for all areas and assume a particular form of distribution

Gamma distribution as a prior distribution

The gamma distribution has two parameters ν (*scale*) and α (*shape*)

Mean is ν/α Variance is ν/α^2

$$\text{so } \gamma = \nu / \alpha \quad \phi = \nu / \alpha^2$$

26

Empirical Bayes Estimation

$$\hat{w}_i = \frac{\hat{\phi}}{(\hat{\phi} + \hat{\gamma} / n_i)} \quad \hat{\theta}_i = \hat{w}_i r_i + (1 - \hat{w}_i) \hat{\gamma}_i$$

$$\hat{w}_i = \frac{\hat{\nu} / \hat{\alpha}^2}{(\hat{\nu} / \hat{\alpha}^2 + \hat{\nu} / n_i \hat{\alpha})} \quad \hat{\theta}_i = \hat{w}_i r_i + \frac{(1 - \hat{w}_i) \nu_i}{\hat{\alpha}}$$

$$\hat{\theta}_i = \frac{y_i + \hat{\nu}_i}{n_i + \hat{\alpha}}$$

$$\hat{w}_i = \frac{n_i}{(n_i + \hat{\alpha})}$$

If n_i is large w_i is close to 1 so weight goes to r_i

If n_i is small weight goes to $\hat{\nu} / \hat{\alpha}$

27

Empirical Bayes Estimation

SIDS Example:

Mecklenburg County had the highest number of SIDS cases (44) in the state over the study period. However, it also had a large number of births (21,588).

Suppose $\gamma_i = 0.00202$ and $\phi_i = 7.692931e-007$ for the county.

$$\hat{w}_i = \frac{\hat{\phi}}{(\hat{\phi} + \hat{\gamma} / n_i)} \quad \hat{w}_i = \frac{7.69231e-007}{(7.69231e-007 + 0.00202 / 21588)}$$

$w = 0.89$

$$\hat{\theta}_i = \hat{w}_i r_i + (1 - \hat{w}_i) \hat{\gamma}_i$$

$$\hat{\theta}_i = .89 r_i + (1 - .89) 0.00202$$

means that 90% of the weight would be given to the local, observed rate.

28

Empirical Bayes Estimation

SIDS Example:

Swain County had 3 cases of SIDS out of only 675 births. Given the above prior mean and variance, $w = 0.20$. Much less weight is given to the observed rate, due to the small number of births.

$$\hat{w}_i = \frac{7.69231e-007}{(7.69231e-007 + 0.00202/675)}$$

$$w = 0.20$$

$$\hat{\theta}_i = \hat{w}_i r_i + (1 - \hat{w}_i) \hat{\gamma}_i$$

$$\hat{\theta}_i = .20r_i + (1 - .20)0.00202$$

only 20% of the weight given to the local, observed rate.

29

Method of Moments Estimation

Estimate γ by the global mean of observed rates:

$$\hat{\gamma} = \frac{\sum y_i}{\sum n_i}$$

Estimate ϕ by the weighted sample variance of the observed rates around the mean:

$$\hat{\phi} = \frac{\sum n_i \cdot (r_i - \hat{\gamma})^2}{\sum n_i} - \frac{\hat{\gamma}}{\bar{n}}$$

30

Method of Moments Estimation

Then the shrinkage factor is estimated as:

$$\hat{w}_i = \frac{\hat{\phi}}{(\hat{\phi} + \hat{\gamma}/n_i)}$$

$$\hat{\gamma} = \frac{\sum y_i}{\sum n_i}$$

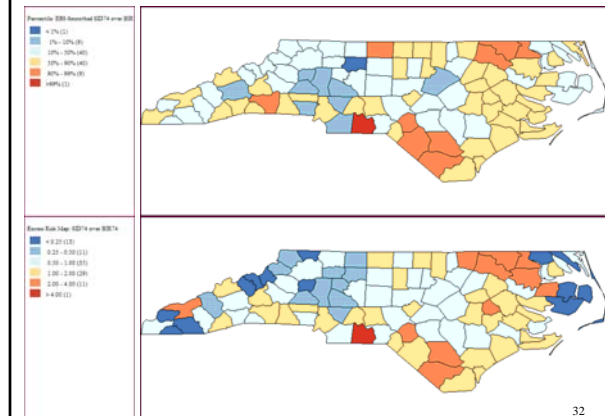
$$\hat{\phi} = \frac{\sum n_i \cdot (r_i - \hat{\gamma})^2}{\sum n_i} - \frac{\hat{\gamma}}{\bar{n}}$$

And the Bayes estimates of the rates are

$$\hat{\theta}_i = \hat{\gamma} + \frac{\hat{\phi}(r_i - \hat{\gamma})}{(\hat{\phi} + \hat{\gamma}/n_i)}$$

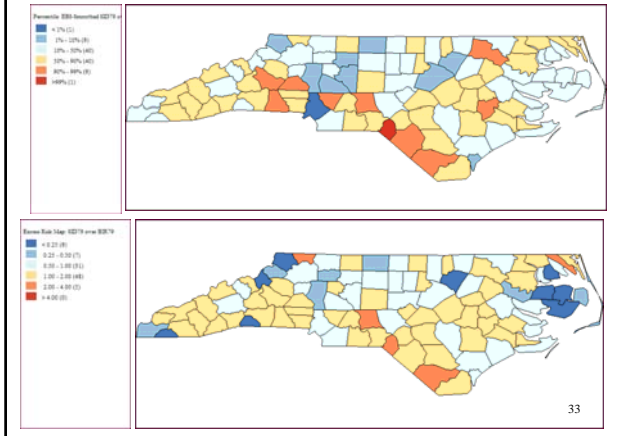
31

Empirical Bayes 1974



32

Empirical Bayes 1979



Generalized Linear Model

Review of regression modeling

A response variable y_i is assumed to be the outcome of a random variable Y_i , with mean value μ_i .

The μ_i is related to a set of explanatory variables x_j by a set of parameters. These parameters are unknown and must be estimated.

Values are distributed around the mean; the manner in which they are distributed is the error distribution.

The form of this distribution is typically assumed to follow specified forms.

In general, to derive optimal parameters and their associated standard errors for μ , assumptions about the distribution of errors must be made.

Generalized Linear Model

The OLS case is highly restrictive: it models only linear relationships, we assume errors are independent, normal, and have constant variance.

Violations of these assumptions don't always affect the estimation of the coefficients, but estimates of parameter significance are invalid.

Weighted least squares relaxes the constant variance assumption.

Non-linear relationships can be handled with transformations

Such transformations do not ensure that the error distribution remains normal with constant variance.

35

Generalized Linear Model

Some compromise between the linearity of the relationship and stabilization of the variance must often be made.

Generalized linear models are less restrictive than OLS, because the distributional assumptions of OLS and the strict linearity assumptions are not necessary. These models consist of two components

1. A link function supporting non-linear relationships between the response and explanatory variables
2. A set of supported error distributions beyond the normal

36

Generalized Linear Model

Cases that require a generalized linear model:

when y_i and μ_i are bounded. For example, if y represents the amount of some physical substance then we may have $y > 0$ and $\mu > 0$.

if y is binary. $y = 1$ if an animal survives and $y = 0$ if it does not, then $0 < \mu < 1$.

The linear model is inadequate in these cases because complicated and unnatural constraints on β would be required to make sure that μ stays in the possible range.

Generalized linear models instead assume a *link linear* relationship

37

Generalized Linear Model

Suppose that we have a sample of n observations y_1, y_2, \dots, y_n which can be treated as realizations of independent Poisson random variables, with $Y_i \sim P(\mu_i)$, and suppose that we want to let the mean μ_i (and therefore the variance!) depend on a vector of explanatory variables \mathbf{x}_i .

We could have a simple linear model of the form

$$\mu_i = \mathbf{x}_i \beta$$

The linear predictor on the right hand side can assume any real value.

The Poisson mean on the left hand side, an expected count, has to be non-negative

38

Generalized Linear Model

To fix this we can model the *logarithm* of the mean using a linear model.

We take logs calculating $h_i = \log(\mu_i)$ and assume that the transformed mean follows a linear model $h_i = \mathbf{x}_i \beta$.

Thus, we consider a generalized linear model with log link. Combining these two steps we can write the log-linear model as

$$\log(\mu_i) = \mathbf{x}_i \beta$$

In this model the regression coefficient β_j represents the expected change in the *log* of the mean per unit change in the predictor x_j . In other words increasing x_j by one unit is associated with an increase of β_j in the log of the mean.

39

Generalized Linear Model

GLM is a generalization of the linear regression model such that

- nonlinear, as well as linear, effects can be tested
- supports categorical, ordinal, count, response variables, as well as continuous variables
- supports dependent variables whose distribution follows several special members of the exponential family of distributions as well as normally-distributed dependent variables

40

Generalized Linear Model

These models can be used for dependent (response) variables that assume any distribution from the exponential family

- Normal
- Binomial
- Poisson
- Gamma

Non-constant variance (as with the Poisson and Gamma distributions) are handled by the model.

41

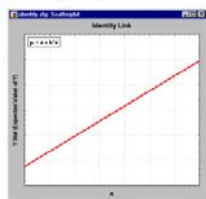
Generalized Linear Model

Common link functions and their associated probability distributions

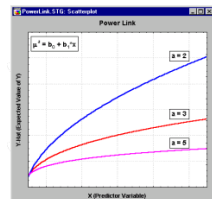
Distribution	Name	Link Function	Mean Function
Normal	Identity	$X\beta = \mu$	$\mu = X\beta$
Exponential Gamma	Inverse	$X\beta = \mu^{-1}$	$\mu = (X\beta)^{-1}$
Inverse Gaussian	Inverse squared	$X\beta = \mu^{-2}$	$\mu = (X\beta)^{-1/2}$
Poisson	Log	$X\beta = \ln(\mu)$	$\mu = \exp(X\beta)$
Binomial Multinomial	Logit	$X\beta = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(X\beta)}{1 + \exp(X\beta)} = \frac{1}{1 + \exp(-X\beta)}$

42

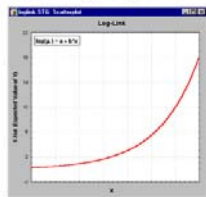
Identity link: $f(x) = x$



Power link: $f(x) = x^a$, for a given a



Log link: $f(x) = \log(x)$



Normal, Gamma, Inverse normal, and Poisson distributions

Binomial, and Ordinal Multinomial distributions:

Logit link: $f(z) = \log(z/(1-z))$

Probit link: $f(z) = \text{invnorm}(z)$

where *invnorm* is the inverse of the standard normal cumulative distribution function

43

Generalized Linear Model

The parameters are solved within the maximum likelihood framework. Instead of employing OLS to estimate parameters for the independent variables, GLM employs iterative re-weighted least squares (IRLS).

1. Reasonable values for β are guessed
2. Using these β a weighted regression is developed, producing revised estimates for β .
3. The revised β are employed in a subsequent weighted regression.
4. Steps 2 and 3 are repeated until the estimates converge.

44

Generalized Linear Model

(Scaled) Deviance is the goodness-of-fit statistic for generalized linear models. It is like the R^2 in OLS, but is a log likelihood statistic.

The deviance of a fitted model compares the log-likelihood of the fitted model to the log-likelihood of a model with n parameters that fits the n observations perfectly (saturated model).

$$\text{Deviance} = -2 * (L_m - L_s)$$

where L_m is the maximized log-likelihood value for the model of interest, and L_s is the log-likelihood for the saturated model

The smaller the deviance, the closer the fitted value is to the saturated model. The larger the deviance, the poorer the fit.

45

Example: Launch temperatures (in degrees Fahrenheit) and an indicator of O-ring failure for 24 space shuttle launches prior to the space shuttle Challenger disaster in 1986.

Let $\pi(x) = \text{Prob}(\text{success}|x)$ and $1 - \pi(x) = \text{Prob}(\text{failure}|x)$. We want a 'model' for $\pi(x)$.

Why not a linear regression model $\pi(x) = \beta_0 + \beta_1 x_1$

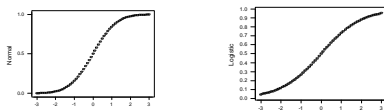
The variance of the outcome of a Bernoulli trial is $[\pi(x)(1-\pi(x))]$ = $[\beta_0 + \beta_1 x_1][1 - (\beta_0 + \beta_1 x_1)]$. The variance of an observation depends on x , meaning the assumption of constant variance is not satisfied.

The errors would be either $0 - [\beta_0 + \beta_1 x_1] = -\beta_0 - \beta_1 x_1$ or $1 - (\beta_0 + \beta_1 x_1)$ --just two possible values for a given x --violating assumption of normality.

46

Since $\pi(x)$ is a probability, its values should be between 0 and 1.

For the O-ring problem, we would expect $\pi(x)$ to increase from values near 0 to values near 1: as temperatures increase the chances of a failure should decrease or the chances of a 'success' --no O-ring failure--should increase.



The expression for the logistic curve: $F(x) = e^x / (1 + e^x)$. The corresponding regression model is

$$\pi(x) = F(\beta_0 + \beta_1 x_1) = \exp(\beta_0 + \beta_1 x_1) / [1 + \exp(\beta_0 + \beta_1 x_1)].$$

47