

STAT 442: Statistical Methods for Categorical Data

Winter 2008: Term 1

Instructor: Dr. Rollin Brant

e-mail: rollin@stat.ubc.ca

Lectures: Tuesday, Thursday, 8:00-9:30am

Pre-requisite: STAT 306 or equivalent

Text: *Analyzing Categorical Data*, Jeffrey S. Simonoff

(available in the bookstore or on-line through library)

Computing: Exercises will require the use of appropriate software. In lectures examples will be based on the use of \mathcal{R} , which is freely available (see www.R-project.org).

Evaluation:

Assignments	25%
Two Midterms	20% each
Final Exam	35%

WHAT IS CATEGORICAL DATA?

Statistics = interpreting scientific observations

- Measurements
- Classifications
- Recorded events

CATEGORICAL DATA:

- binary
- multi-category
 - nominal
 - ordinal
- events

Finding relationships in data (Stat 306 - Pre-req!)

y = response, outcome, [dependent]

$X_1, X_2, X_3, \dots, X_p$ = explanatory, predictor

[independent]

Stat 306: continuous and binary response

Stat 442: counts, binary response,

multi-category response

Quick Outline = the whole book

Review: Chapters 2 & 3

Chapter 4: Basic building blocks

Binomial, Multinomial, Poisson Distributions

Testing goodness of fit

Chapter 5: Regression models for counts

- Poisson regression, Generalized linear models

Chapter 6: Basic methods for cross-classified data

- two-way contingency tables

Chapter 7: Methods for "special" tables

- ordered categories, square tables

Chapter 8: Multidimensional contingency tables

- allowing for confounders
- log-linear models

Chapter 9: Logistic regression

Chapter 10: Regression for multi-category responses

Review Material:

Chapter 2 - read (on-line!!)

the method of maximum likelihood (more later!)

The multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \quad (2.7)$$

A framework for relating y to the x 's

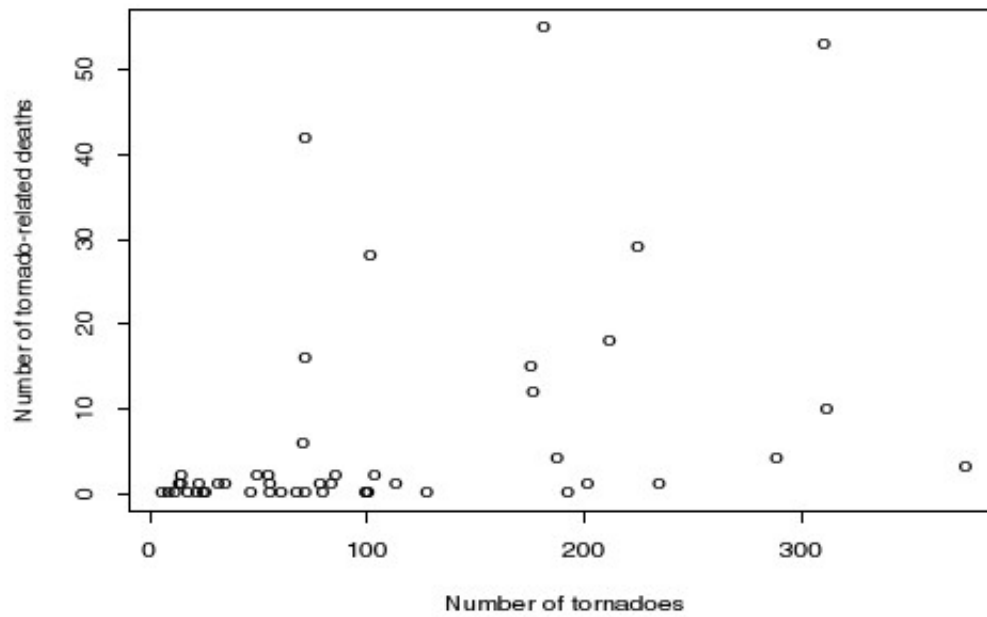
e.g. (text) y = monthly # of tornado-related deaths

in US, 1996-1999

x_1 = # tornadoes for the month

x_2 = # killer tornadoes for the month

Number of deaths versus number of tornadoes



Holding the x 's fixed: Regression as stratification

Stratify by x_1 and x_2 : Each stratum is a set of

months which have the same values for x_1 and x_2

e.g. months with 50 tornadoes, 10 killer tornadoes

Model assumptions describe the behaviour of y within

each stratum determined by the x 's:

Relating the mean of y to the x 's:

$$\mu = E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Relating the standard deviation of y to the x 's:

$$\sigma = \sqrt{\text{Var}(y)} = \text{is constant (i.e. same for all strata)}.$$

The distribution of y : y is normally distributed within strata.

One variable picture (with 3 strata indicated)

2.2 Linear Regression and Least Squares 13

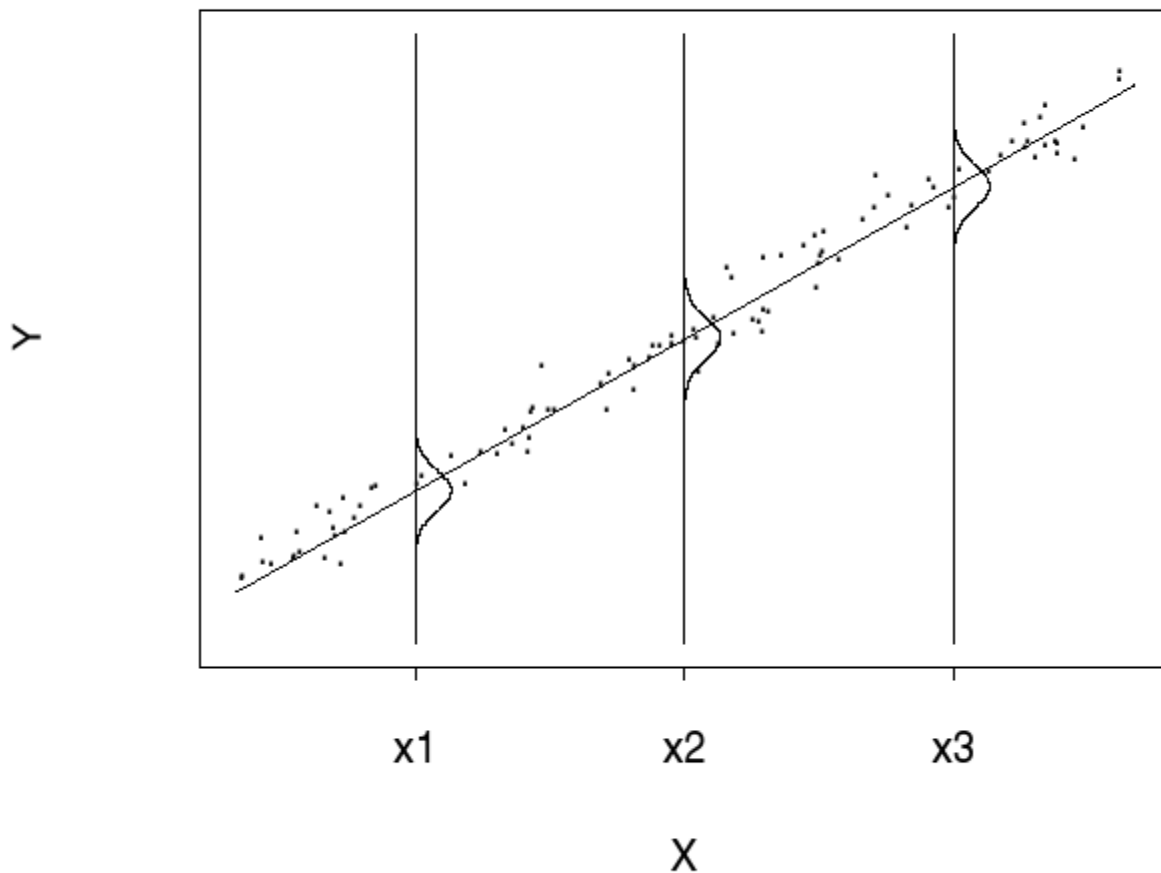


FIGURE 2.2. The simple regression model.

Note: errors $(\epsilon_i = y_i - \mu_i)$ are $N(0, \sigma^2)$

- mathematical construct for notational convenience

Model Fitting:

Least squares (maximum likelihood) estimation

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi})^2. \quad (2.10)$$

Let $\hat{\beta} = \{\hat{\beta}_0, \dots, \hat{\beta}_p\}$ be the maximum likelihood estimates from a given set of data. Substituting these estimates into (2.7) gives the *fitted values*,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi}.$$

The set of differences between the observed and fitted target values,

$$r_i = y_i - \hat{y}_i,$$

Matrix notation for algebraic solutions:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{pmatrix} \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

The regression model can then be written succinctly as

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

The normal equations (which determine the least squares estimates of $\boldsymbol{\beta}$) can be shown (using multivariable calculus) to be

$$(X'X)\boldsymbol{\beta} = X'\mathbf{y},$$

Least squares estimates, fitted values, residuals

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}.$$

The fitted values are then

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = X(X'X)^{-1}X'\mathbf{y} \equiv H\mathbf{y},$$

where $H = X(X'X)^{-1}X'$ is the so-called “hat” matrix. The residuals $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$ thus satisfy

$$\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X(X'X)^{-1}X'\mathbf{y} = (I - X(X'X)^{-1}X')\mathbf{y},$$

or

$$\mathbf{r} = (I - H)\mathbf{y}.$$

Checking the model assumptions

George Box:

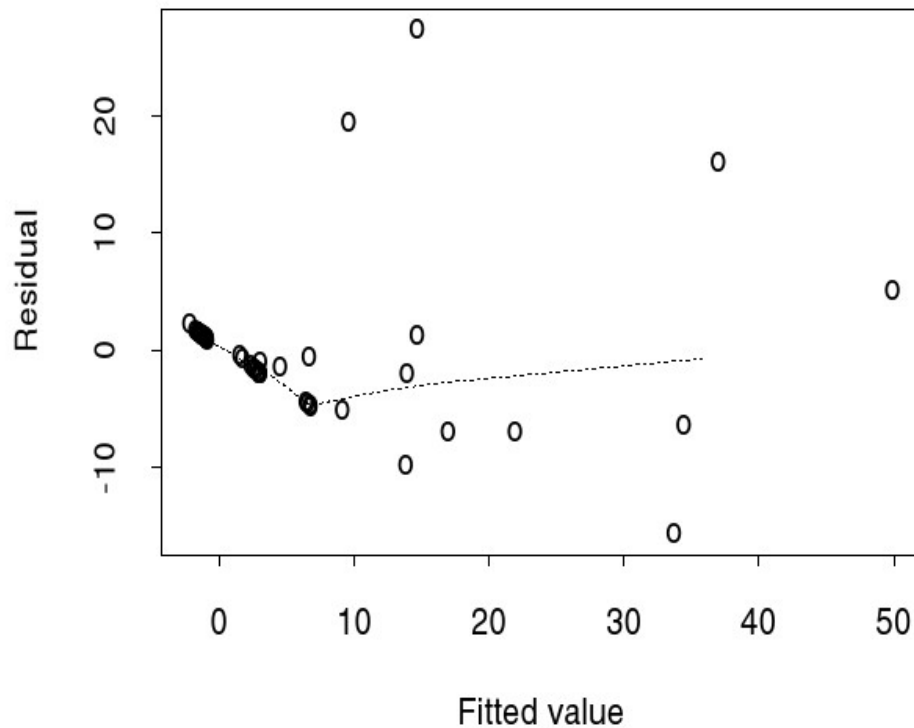
Essentially, all models are wrong, but some are useful

Model checks - are the assumptions "close enough"?

Linearity of y w.r.t. x_j : plot residuals versus x_j

Homogeneous dispersion: plot residuals versus \hat{y} 's

Normality: Q-Q plot of residuals



The residual standard deviation : $\sqrt{\hat{\sigma}^2}$ where

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p - 1}.$$

Least squares decomposition:

$$\sum_{i=1}^n (y_i - \bar{Y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{Y})^2.$$

The coefficient of determination: R^2

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{Y})^2}{\sum_i (y_i - \bar{Y})^2} \equiv \frac{\text{Regression SS}}{\text{Corrected total SS}} = 1 - \frac{\text{Residual SS}}{\text{Corrected total SS}}.$$

$$R^2 = \text{corr}(y_i, \hat{y}_i)^2, \quad (2.11)$$

where

$$\text{corr}(y_i, \hat{y}_i) = \frac{\sum_i (y_i - \bar{Y})(\hat{y}_i - \bar{\hat{Y}})}{\sqrt{\sum_i (y_i - \bar{Y})^2 \sum_i (\hat{y}_i - \bar{\hat{Y}})^2}}$$

Inference: If the model is reasonable $\hat{\beta}$'s and $\hat{\sigma}^2$ are unbiased estimates.

e.g. (vector notation) $E(\hat{\beta}) = \beta$ the true value

$$E(\hat{\sigma}^2) = \sigma^2$$

Sampling distribution of $\hat{\beta}$

- multivariate normal, $\text{Var}(\hat{\beta}) = \sigma^2 (X^t X)^{-1}$

Diagonal elements of $\hat{\sigma}^2 (X^t X)^{-1}$ yield standard errors

(notation: $\text{s.e.}(\hat{\beta}_j)$)

Confidence interval for β_j : $\hat{\beta}_j \pm t_{\alpha/2}^{n-p-1} \text{s.e.}(\hat{\beta}_j)$

(note text uses additional hat!!)

Testing the overall significance of the regression:

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$

versus

$$H_a : \text{some } \beta_j \neq 0, \quad j = 1, \dots, p.$$

The test of these hypotheses is the *F-test*,

$$F = \frac{\text{Regression MS}}{\text{Residual MS}} \equiv \frac{\text{Regression SS}/p}{\text{Residual SS}/(n - p - 1)}. \quad (2.12)$$

Test for a single coefficient, β_j

$$t_j = \frac{\hat{\beta}_j}{\widehat{\text{s.e.}}(\hat{\beta}_j)}, \quad (2.13)$$

which is compared to a *t*-distribution on $n - p - 1$ degrees of freedom. Other values of β_j can be specified in the null hypothesis (say β_{j0}), with the *t*-statistic becoming

$$t_j = \frac{\hat{\beta}_j - \beta_{j0}}{\widehat{\text{s.e.}}(\hat{\beta}_j)}.$$

Analysis of Variance

Source	df	SS	MS	F	p
Regression	2	6227.2	3113.6	67.61	0.000
Residual	45	2072.4	46.1		
Total	47	8299.7			

Predictor	Coef	s.e.	t	p
Constant	-0.846	1.510	-0.56	0.578
Tornadoes	-0.007	0.013	-0.56	0.577
Killer tornadoes	4.005	0.390	10.28	0.000

Making predictions:

$$\widehat{\text{s.e.}}(\hat{y}_0^P) = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum (x_i - \bar{X})^2}}$$

(we might term this more precisely the *estimated* standard error of \hat{y}_0^P , since an estimate of σ is used). More generally,

$$\hat{V}(\hat{y}_0^P) = [1 + \mathbf{x}'_0 (X'X)^{-1} \mathbf{x}_0] \hat{\sigma}^2$$