Today and tomorrow

Goodness of fit statistics

- Score statistic (Tuesday)
- Wald statistic (Tuesday)
- Deviance (Today)

Hypothesis testing (Today)

- Nested models

Chp 6 Normal Linear Models (Today and tomorrow)

- Focus on GLM formulation, outlier detection and colinearity.

Chapter 5, Inference

- Goodness of fit statistics:
  - Score statistic
$$U^T \Im^{-1} U \sim \chi^2(p)$$
  - Wald statistic, b MLE
$$(b-\beta)^T \Im^{-1} (b-\beta) \sim \chi^2(p)$$
  - Log-likelihood ratio statistic $\Rightarrow$ Deviance
- Hypothesis tests

Shooting balloons

- N trail subjects, $i = 1, 2, \ldots, N$
- Each shot n_i times, trying to hit balloons.
- Count hits y_i.
- Explanatory variables:
  - Experienced / non-experienced gunman
  - Wind speed

Data:

Trail person	1	2	3	...
Experienced	1	0	0	...
Wind speed	2.13	0.59	1.03	...
n_i	6	3	5	...
y_i	2	1	1	...

Shooting balloons, model

- $Y_i \sim bin(n_i, \pi_i)$, $i = 1, 2, \ldots, N$
- $\eta_i = logit(\pi_i)$
-
  1. $\eta_1 = \beta_0 \Rightarrow Y_i \sim bin(n_i, \pi)$
  2. $\eta_i = \beta_0 + \beta_1 x_1 \Rightarrow Y_i \sim bin(n_i, \pi_i)$
  3. $\eta_i = \beta_0 + \beta_2 x_2 \Rightarrow Y_i \sim bin(n_i, \pi_i)$
  4. $\eta_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \Rightarrow Y_i \sim bin(n_i, \pi_i)$

Where $x_1 = 1$ for experienced gunman, otherwise $x_1 = 0$ and x_2 is wind speed.

Saturated model

The richest possible model. Each combination of (all possible known) explanatory variables have their own θ_i. $b = b_{max}$

Example Balloons

$N = 10$ persons trying.

$Y_i \sim bin(n_i, p_i)$, p_i unique for each y_i

- Model with one factor, and this factor has N levels; one for each observation/person.

$m = length(b_{max}) = 10$

Example: Chronically medical conditions

- Women in rural area see GP less then women in urban area.
- Why? Less sick or less accessible?

Saturated model:

Example: Chronically medical conditions

- Women in rural area see GP less then women in urban area.
- Why? Less sick or less accessible?

Data

Group 1: No. of chronically conditions for 26 town women with ≤ 3 GP visits.

Group 2: No. of chronically conditions for 23 country women with ≤ 3 GP visits.

Do women in the two groups with the same number of visits have the same need?

Saturated model:

Example: Chronically medical conditions

- Women in rural area see GP less then women in urban area.
- Why? Less sick or less accessible?

Data

Group 1: No. of chronically conditions for 26 town women with ≤ 3 GP visits.

Group 2: No. of chronically conditions for 23 country women with ≤ 3 GP visits.

Do women in the two groups with the same number of visits have the same need?

Saturated model:

- One explanatory variable (town/country), 26 towm replicates and 23 country replicates.

5.3 Taylor series approximations for log-likelihood

Taylor approximations for $l(\beta)$ near estimate b:

$$l(\beta) = l(b) + (\beta - b)U(b) + \frac{1}{2}(\beta - b)^2 U'(b)$$

Approximate $U'(b)$ with $E(U') = -\Im(b)$:

$$l(\beta) = l(b) + (\beta - b)U(b) - \frac{1}{2}(\beta - b)^2 \Im$$

For a vector b

$$l(\beta) = l(b) + (\beta - b)U(b) - \frac{1}{2}(\beta - b)^T \Im(\beta - b)$$

$\chi^2()$ results ch 1.4 and 1.5

Definition χ^2

If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2(1)$.
If $Z_1, Z_2, \ldots Z_n$ are independent identical distributed $Z_i \sim N(0,1)$, the $\sum_{i=1}^n Z_i^2 \sim \chi^2(n)$

Non iid

If $Y \sim MVN(\mu, \Sigma)$, then $(Y-\mu)^T \Sigma^{-1} (Y-\mu) \sim \chi^2(n)$

Definition non-central χ^2

If $Z_1, Z_2, \ldots Z_n$ are independent identical distributed $Z_i \sim N(0,1)$, the $\sum_{i=1}^n (Z_i - \mu_i)^2 \sim \chi^2(n, \nu)$ with $\nu = \sum \mu_i^2$.

Subtraction

If $X_1^2 \sim \chi^2(m)$ and $X_2^2 \sim \chi^2(k)$, $m > k$, and X_1^2 and X_2^2 are independent, we have: $X^2 = X_1^2 - X_2^2 \sim \chi^2(m-k)$

Chapter 6, Linear Normal Models

Properties:

- As GLM
- Maximum Likelihood Estimate (MLE)
- Least Square Estimate
- Deviance
- Hypothesis testing

Models:

- Multiple linear regression
  - Outlier detection / influential observation
  - Collinearity / multicollinearity
- Analysis of variance (ANOVA)
  - One factor ANOVA
  - Two factor ANOVA
- Analysis of covariance
- General linear model

Deviance

Let β_{max} be the parameter vector for the *saturated* modeled, and β for the model of our interest. Let $l(\beta; y)$ be the log-likelihood function. The *deviance* of the model is

$$D = 2(l(b_{max}; y) - l(b; y)$$

where b and b_{max} are (ML) estimates.

Gaussian pdf

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-0.5\frac{(y-\mu)^2}{\sigma^2})$$