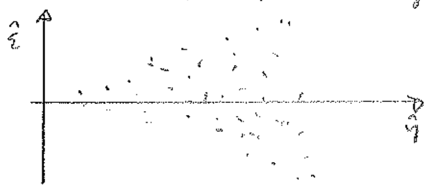
Applied Statistics, Practicals Tue 9-10. 15/10/2009

4) Some Diagnostics Model: $y = X\beta + \varepsilon$, ε iid $N(0, \sigma^2)$

To see whether a larger model is needed, add terms (e.g. x^2) and see if they are significant.

Note: (i) $\hat{\varepsilon}$ and $\hat{y}$ are uncorrelated
(proof follows from $\mathrm{cov}(\hat{\varepsilon}, \hat{y}) = \mathrm{cov}((I-H)Y, HY)$ & Handout 2)
(ii) $\hat{\varepsilon}$ and $\hat{\beta}$ are uncorrelated

- Plot $\hat{\varepsilon}$ against $\hat{y}$ we expect no systematic relationship. If we see a syst. pattern, might make us question the model.



We may see $|\hat{\varepsilon}_i|$ is increasing as $\hat{y}_i$ increases, i.e. "fanning out" suggests that variance is increasing with fitted value.
See Handout 1.1 for boys birthweight data

If we see fanning out, we could try to fix this up by taking a transformation of the y_i's, e.g. $\log(y_i) = \alpha + \beta x_i + \varepsilon_i$

If $Y_i \sim N(\mu, \sigma^2(\mu))$ and $f(\cdot)$ is "well behaved"

$$f(Y) \approx f(\mu) + (Y-\mu) f'(\mu)$$

this implies $f(Y) \overset{\text{approx.}}{\sim} N(f(\mu), \sigma^2(\mu)(f'(\mu))^2)$

Idea: try to have an $f(\cdot)$ so that $\sigma^2(\mu)(f'(\mu))^2$ is approx. constant, i.e. choose f to stabilize the variance.

OR use Box-Cox transformations (see later, see practical)

- Plot $\hat{\varepsilon}_i$'s against covariates both in and not yet in the model to see if there is any dependence.

- Normality: Q-Q plot

Model assumes $P(\varepsilon_i \le x) = \Phi(\frac{x}{\sigma})$ where $\Phi(x) = P(N(0,1) \le x)$

Approximate $P(\varepsilon_i \le x)$ by

$$\hat{F}_n(x) = \frac{\#\{i : \hat{\varepsilon}_i \le x\}}{n}.$$

$\hat{F}_n$ is the empirical distribution function of residuals

If assumptions OK, then $\hat{F}_n(x) \approx \Phi(\frac{x}{\sigma})$

$$\Phi^{-1}(\hat{F}_n(x)) \approx \frac{x}{\sigma}.$$

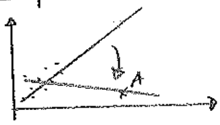
$E(\hat{\varepsilon}) = E(M_X Y) = M_X E(Y) = M_X X\beta = 0.$

A QQ plot plots $\hat{\varepsilon}_i$'s against $\Phi^{-1}(\hat{F}_n(\hat{\varepsilon}_i))$, which should be close to a straight line if model is OK.

$$\sigma^2_{\hat{\varepsilon}} = \overbrace{E(\hat{\varepsilon}\hat{\varepsilon}')}^{E(\hat{\varepsilon}\hat{\varepsilon}') + \underbrace{E(\hat{\varepsilon})^2}_{=0}};\quad \hat{\varepsilon} = \underbrace{(I - P_X)}_{M_X} Y$$
$$= E[M_X Y (M_X Y)'] = M_X \mathrm{var}(Y) M_X' = M_X(\sigma^2 I_n) M_X' = \sigma^2 M_X = \sigma^2(I - P_X).\ M_X \text{ idempotent.}$$

From $\hat{\varepsilon} \sim N(0, \sigma^2(I-H))$, we know $\mathrm{var}(\hat{\varepsilon}_i) = (1-h_i)\sigma^2$ where h_i is (i,i) element of H. So could replace $\hat{\varepsilon}_i$ by $\underbrace{\frac{\hat{\varepsilon}_i}{\sqrt{1-h_i}}}_{\text{standardised residuals}}$ or could replace it by $\underbrace{\frac{\hat{\varepsilon}_i}{s\sqrt{1-h_i}}}_{\text{studentised residuals}}$

• Influential points (Cook Statistic)



An influential point is one whose deletion causes a large change in the analysis

Point A has a large effect on estimate of the slope.

Fitting the model without point i, lead to $\hat{\beta}_{-(i)}$

The Cook statistics are $D_i = (\hat{Y} - \hat{Y}_{-(i)})^T (\hat{Y} - \hat{Y}_{(i)}) / s^2 p$, $i = 1, \dots, n$, $p =$ # parameters

Large values of them could indicate an influential observation.

Once an outlier or influential observation is identified, we consider: is it a mistake?

Only remove it permanently from dataset if there is a very good reason.

Otherwise, could report results from 2 analyses, one with the point one without.

See Handout 2

See H'out 2.2 for Box-Cox.

Some proofs: • $\hat{\varepsilon}$ and $\hat{y}$ are uncorrelated:

$$\mathrm{cov}(\hat{\varepsilon}, \hat{y}) = \mathrm{cov}((I-H)Y, HY)$$
$$= E\Big[\big[(I-H)Y - (I-H)E(Y)\big]' \cdot \underbrace{\big[HY - HE(Y)\big]}_{X(X'X)^{-1}X'Y - X(X'X)^{-1}X'X\beta}\Big]$$
$$\hat{Y} - X\beta = 0 \quad \blacksquare$$

• $\hat{\varepsilon}$ and $\hat{\beta}$ are uncorrelated:

$$\mathrm{cov}(\hat{\varepsilon}, \hat{\beta}) = \mathrm{cov}\big[(I-H)y, (X^TX)^{-1}X^Ty\big]$$
$$= E\Big[\big[(I-H)y - (I-H)E(y)\big]^T \cdot \underbrace{\big[(X^TX)^{-1}X^Ty - (X^TX)^{-1}X^T E(y)\big]}_{(X^TX)^{-1}X^Ty - (X^TX)^{-1}X^TXb = b - b = \underline{0}}\Big] \quad \square$$

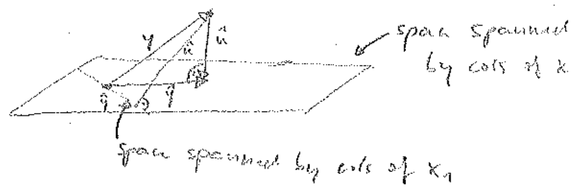
5. Hypothesis testing and analysis of variance

Ω $\quad y = \underset{n\times\tilde{p}}{X}\underset{\tilde{p}\times 1}{\beta} + \varepsilon$

Suppose $\underset{n\times\tilde{p}}{X} = (\underset{n\times p_1}{X_1}\ \underset{n\times p_2}{X_2})$ $\qquad \tilde{p} = p_1 + p_2$

$$\beta = \begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix} \begin{matrix} p_1\times 1 \\ p_2 \times 1\end{matrix}$$

ω_1 $\quad y = X_1\beta_1 + \varepsilon \qquad$ i.e. $\beta_2 = 0$



Then $RSS_\Omega \leq RSS_\omega$

Intuitively, $RSS_{\omega_1} - RSS_\Omega$ large then this provides evidence against $\beta_2 = 0$.

In fact (not proved) indep. $\left\langle \begin{array}{l} (RSS_{\omega_1} - RSS_\Omega)/\sigma^2 \sim \chi^2_{p_2}(\delta) \quad \text{non-central } \chi^2 \quad (\text{see H1.2}) \\ RSS_\Omega/\sigma^2 \sim \chi^2_{n-\tilde{p}} \end{array}\right.$

and non-centrality parameter δ is zero if ω_1 is true (i.e. $\beta_2 = 0$)

So test $\beta_2 = 0$ by referring

$$F = \frac{(RSS_{\omega_1} - RSS_\Omega)/p_2}{RSS_\Omega/(n-\tilde{p})} \quad \text{to } F_{p_2, n-\tilde{p}} \text{ under } H_0: \beta_2 = 0.$$

and reject $\beta_2 = 0$ if $F > F_{p_2, n-\tilde{p}}(\alpha)$ (the appropriate %-age point of F)

or equivalently if p-value, i.e. $P(F_{p_2, n-\tilde{p}} >$ observed value of $F)$ is small.

Note that when $p_2 = 1$, it turns out $F = t^2$.

Simple case: $\quad \Omega: \quad y = \mu 1 + \underset{p\times 1}{X\beta} + \varepsilon \qquad p+1$ parameters $= \tilde{p}$

(vector of 1s.)

$\omega: \quad y = \mu 1 + \varepsilon \qquad$ 1 parameter (NULL MODEL)

Fit ω, MLE is $\hat{\mu} = \bar{y}$ so $RSS_\omega = \sum_{i=1}^n (y_i - \bar{y})^2$

See H3.1 for analysis of variance table.

The coefficient of determination is $R^2 = \frac{S_\beta}{RSS_\omega}$, the proportion of the total variation "explained by" the regression on β. $(0 \leq R^2 \leq 1)$.

Often S_β is further split for testing subsets of parameters.

$\omega_1: \quad y = \mu 1 + X_1\beta_1 + \varepsilon$

See H3.1 for this analysis of variance table.

In general $SS_{\beta_2|\beta_1}$ is NOT same as S_{β_2} so order of fitting matters.

HOWEVER, if $X_1'X_2 = 0$ then we say β_1 and β_2 are orthogonal.

In this case $SS_{\beta_2|\beta_1} = S_{\beta_2}$ and $SS_{\beta_1|\beta_2} = S_{\beta_1}$, i.e. SS for testing $\beta_2 = 0$ is the same whether or not β_1 is in the model.

Further, orthogonality means that the least squares estimate for β_1 in the models $y = \mu 1 + X_1\beta_1 + \varepsilon$ is same as $y = \mu 1 + X\beta + \varepsilon$ $\qquad$ (see practical 3).

6. More structured data

<u>Example 6.1</u> see W1.1 Plant weight example

Aim: determine whether/how 'weight' depends on 'group'

Response variable is "weight"

Explanatory variable "group" is categorical variable (with 3 categories, control, A & B)

This kind of explanatory variable is called a <u>factor</u>

'group' is a factor with 3 <u>levels</u>

Let Y_{ij} be weight for j^{th} plant in group i, $i = \begin{matrix} 1 & \text{control} \\ 2 & A \\ 3 & B \end{matrix}$ and $j = 1, \dots, 10$.

Model: $Y_{ij} = \mu_i + \varepsilon_{ij}$, $\varepsilon_{ij} \sim NID(0, \sigma^2)$

Put this into the linear model framework as follows:

- Let $Y = (Y_{11}, Y_{12}, \dots, Y_{1,10}, Y_{21}, Y_{22}, \dots Y_{2,10}, Y_{31}, \dots, Y_{3,10})^T$
- Similarly $\varepsilon = (\varepsilon_{11} \cdots\ )^T$
- $\beta = (\mu_1, \mu_2, \mu_3)^T$

Then $Y = X\beta + \varepsilon$ where $\underset{30\times 3}{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}\begin{matrix} \} 10\times \\ \\ \} 10\times \\ \\ \} 10\times \\ \\ \end{matrix}$

Least squares equations $X'X\beta = X'y$ become $\begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix}\begin{pmatrix} \hat\mu_1 \\ \hat\mu_2 \\ \hat\mu_3 \end{pmatrix} = \begin{pmatrix} Y_{1\cdot} \\ Y_{2\cdot} \\ Y_{3\cdot} \end{pmatrix}$ where $Y_{i\cdot} = \sum_{j=1}^{10} Y_{ij}$

so $\hat\mu_i = \bar{Y}_{i\cdot}$ where $\bar{Y}_{i\cdot} = \frac{1}{10}\sum_{j=1}^{10} Y_{ij}$

and $\hat{y}_{ij} = \hat\mu_i = \bar{y}_{i\cdot}$ so $RSS = \sum_{i=1}^{3}\sum_{j=1}^{10}(Y_{ij} - \bar{Y}_{i\cdot})^2$

Alternative parametrisation: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ (4 parameters)

Now $X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}\begin{matrix} \} 10 \\ \\ \} 10 \\ \\ \} 10 \\ \\ \end{matrix}$ $\beta = (\mu, \alpha_1, \alpha_2, \alpha_3)^T$

Least squares equations
$$\begin{aligned} 10\mu + 10\alpha_1 + 10\alpha_2 + 10\alpha_3 &= Y_{++} \\ 10\mu + 10\alpha_1 \qquad\qquad\qquad &= Y_{1+} \\ 10\mu \qquad + 10\alpha_2 \qquad\quad &= Y_{2+} \\ 10\mu \qquad\qquad + 10\alpha_3 &= Y_{3+} \end{aligned}$$
$Y_{++} = \sum_{ij} Y_{ij}$

Indeed parameters are not uniquely determined e.g. $\mu' = \mu + 1$, $\alpha_i' = \alpha_i - 1$

Way out: impose a constraint on the parameters. Several possibilities:

(1) <u>sum-to-zero constraint</u>: $\sum_{i=1}^{3} \alpha_i = 0$ (3 free parameters)

Solve least squares equations & $\sum_{i=1}^{3}\alpha_i = 0$, we find:

$\hat\mu = \bar{Y}_{++}$, $\hat\alpha_i = \bar{Y}_{i+} - \bar{Y}_{++}$ (check) → Dobson page 98.

(where $\bar{Y}_{++} = \frac{1}{30}\sum_{i=1}^{3}\sum_{j=1}^{10} Y_{ij} = \frac{1}{3}\sum_{i=1}^{3}\bar{Y}_{i+}$)

(2) corner point constraint: $\alpha_1 = 0$

Solve least squares equations & $\alpha_1 = 0$

$\hat{\mu} = \bar{Y}_{1+}$ $\quad \hat{\alpha}_i = \bar{Y}_{i+} - \bar{Y}_{1+}$ $\quad i = 2,3$ (check) → Dobson page 99.

For both (1) & (2) $\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i = \bar{Y}_{i+}$ so RSS same.

See H4.1

Note: with FACTORS, first run analysis of variance and F-Test to see whether we need the factor at all (i.e. in this eg. to test $H0: \hat{\alpha}_i = 0 \ \forall i$). If we need the factor THEN look at the summary to see what the effects of the factor levels are.

Example 6.2: Two-way analysis of variance

See H4.2: Observations cross-classified by 2 factors, lab & treatment group

In general, 2 factors A & B A has I levels
B J

Let Y_{ijk} be response for kth observation ($k = 1, \dots, K$) at level i of factor A & level j ~~for~~ of factor B.

$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$ $\quad \varepsilon_{ijk} \sim N(0, \sigma^2)$

Additive model has $\mu_{ij} = \mu + \alpha_i + \beta_j$ $\quad (*)$

model (*) means that the effect (ie the change in $\mathbb{E}$(response)) in going from level i_1 to level i_2 of A is the same for all levels of B.

ie $\mu_{i_2 j} - \mu_{i_1 j} = \alpha_{i_2} - \alpha_{i_1}$ does not depend on j.

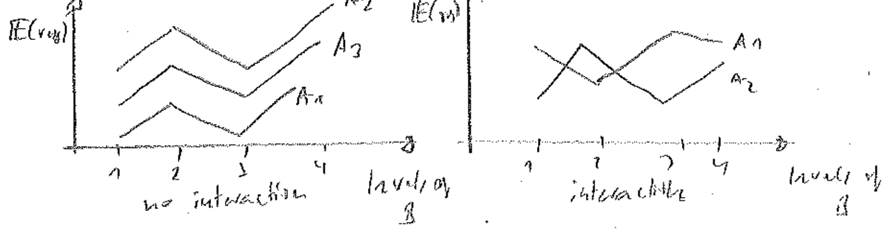
We say there is no interaction between factors A & B.

If instead of (*) we have

$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ $\quad (+)$

then $\mu_{i_2 j} - \mu_{i_1 j} = \alpha_{i_2} - \alpha_{i_1} + \gamma_{i_2 j} - \gamma_{i_1 j}$ → does depend on j.

In (+) there is interaction between factors A & B.



Treat } factors factor
Lab

weight ~ Treat + Lab fits (*)

weight ~ Treat * Lab fits (+)

(Treat + Lab + Treat · Lab)

In (+) corner point constraints

are $\alpha_1 = 0$ $\quad \beta_1 = 0$

$\gamma_{1j} = 0 \ \forall j$ $\quad \gamma_{i1} = 0 \ \forall i$

Can extend to

(i) 3-way or higher tables (so could have eg 3 factor interactions)

(ii) factors and non-factors (i.e. regression variables) together in the same model

Eg birthweight example response is weight - y

explanatory variables sex (factor)

age (regression variable)

$y \sim age$ $\quad Y_{ij} = \mu + \beta x_{ij} + \varepsilon_{ij}$ $\quad i = 1,2 \quad j = 1,..,12$

$y \sim age + sex$ $\quad Y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij}$ $\quad \alpha_1 = 0$

$y \sim age * sex$ $\quad Y_{ij} = \mu + \alpha_i + (\beta + \gamma_i) x_{ij} + \varepsilon_{ij}$ $\quad \alpha_1 = \gamma_1 = 0$

II. Generalized Linear Models (glms)

7. Introduction to glms

Linear Model $y = X\beta + \varepsilon$ $\quad \varepsilon \sim N(0, \sigma^2 I)$ $\quad \varepsilon_i \sim NID(0, \sigma^2)$

so $y_i = \beta' x_i + \varepsilon_i$

Alternatively $\underbrace{y_i \sim N(\mu_i, \sigma^2)}_{\text{(i) distribution}}$ independent & $\underbrace{\mu_i = \beta' x_i}_{\text{(ii) relationship between } E(Y_i) = \mu_i \text{ and the linear predictor } \beta' x_i \text{ (linear in parameters)}}$

Glms generalize both (i) and (ii)

See H5.1

H5.2

Examples (i) $Y \sim N(\mu, \sigma^2)$

Density may be written

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} = \exp\left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right\}$$

$$= \exp\left\{\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}\right\}$$

which is of the form $f(y; \theta, \phi)$ on H5.1 with $\theta = \mu$ $\quad b(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2}$

$\phi = \sigma^2$ $\quad c(y, \phi) = -\frac{1}{2}\log(2\pi\phi) - \frac{y^2}{2\phi}$ $\qquad b'(\theta) = \theta$ so we do get $E(Y) = b'(\theta)$

$b''(\theta) = 1$ $\quad$ -- " -- $\quad Var(Y) = \sigma^2 = \phi b''(\theta)$

The canonical link function for normal is $g(\cdot)$ such that

$g(\mu) = \theta$ i.e. $g(\mu) = \mu$ [identity]

So the model $Y_i = \beta' x_i + \varepsilon_i$ has $Y_i \sim N(\mu_i, \sigma^2)$ $\quad \mu_i = \beta' x_i$

is a glm with canonical link function.

Ex. (ii) $Z \sim Bin(n,p)$ Let $Y = \frac{Z}{n}$ be proportion of successes

Note: $\mathbb{E}(Y) = \frac{1}{n}\mathbb{E}(Z) = p \; (=\mu)$, $var(y) = \frac{p(1-p)}{n}$ (check)

$P(Y=y) = P(Z=ny) = \binom{n}{ny} p^{ny}(1-p)^{n-np}$

$= \exp\left\{ \frac{y\log(\frac{p}{1-p}) + \log(1-p)}{1/n} + \log\binom{n}{ny} \right\}$ (check)

This is of form $f(y,\theta,\delta)$ with $\theta = \log(\frac{p}{1-p})$ so $p = \frac{e^\theta}{1+e^\theta}$ (check)

$b(\theta) = -\log(1-p) = \log(1+e^\theta)$ (check)

$\delta = \frac{1}{n}$

$b'(\theta) = \frac{e^\theta}{1+e^\theta} = p \; (=\mu)$ ✓

$b''(\theta) = \frac{e^\theta}{(1+e^\theta)^2}$ & this is $p(1-p)$

So we do get $var(Y) = \frac{p(1-p)}{n} = \delta b''(\theta)$ ✓

The variance function is $V(\mu) = \mu(1-\mu)$

Canonical link is $g(\cdot)$ so that $g(\mu) = \log(\frac{\mu}{1-\mu})$ $\left[= \log(\frac{p}{1-p})\right]$

→ logistic link

$\boxed{\delta = \phi}$ above.

H5.2 again (2nd page)

Example: Normal deviance see H6.0

8. Binomial data $Y_1, \dots, Y_n$ independent $Y_i \sim Bin(n_i p)$

$g(p_i) = \beta^T x_i$

where $g(\cdot)$ is the canonical link: $\log\left(\frac{p_i}{1-p_i}\right) = \beta^T x_i$

logit (p_i); log odds

which means that $p_i = p_i(\beta) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$. $1-p_i = \frac{1}{1+e^{\beta^T x_i}}$

likelihood $= \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i}$

So loglike $= \ell(\beta) = \sum_{i=1}^{n}\left\{ y_i \log(p_i) + (n_i - y_i)\log(1-p_i) + \log\binom{n_i}{y_i}\right\}$ (8.1)

$= \sum_{i=1}^{n}\left\{ y_i \log\left(\frac{p_i}{1-p_i}\right) + n_i \log(1-p_i) + \log\binom{n_i}{y_i}\right\}$

$= \sum_{i=1}^{n}\left\{ y_i \beta^T x_i - n_i \log(1+e^{\beta^T x_i}) + \log\binom{n_i}{y_i}\right\}$

$= \beta^T\left(\sum_{i=1}^{n} y_i x_i\right) - \sum_{i=1}^{n} n_i \log(1+e^{\beta^T x_i}) + \sum_{i=1}^{n}\log\binom{n_i}{y_i}$

$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} n_i x_i \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$ (check)

So mle $\hat{\beta}$ solves $\sum y_i x_i = \sum n_i x_i \underbrace{\frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}}_{p_i}$ Solve via iteration (see HS.1)

$$\frac{\partial^2 \ell}{\partial\beta\,\partial\beta^T} = -\sum_i n_i x_i x_i^T \frac{e^{\beta^T x_i}}{(1+e^{\beta^T x_i})^2} \qquad \text{(check)}.$$

Does not depend on Y_i's.

So $\mathbb{E}\left[-\frac{\partial^2\ell}{\partial\beta\,\partial\beta^T}\right] = \sum n_i x_i x_i^T p_i(1-p_i) = V(\beta)^{-1}$ say

Then, by general theory for mle's, we have

$$\hat{\beta} \text{ approx } \underset{\substack{\uparrow \\ \text{multivariate} \\ \text{normal}}}{N}(\beta, V(\beta))$$

Recall $\phi_i = \frac{1}{n_i}$ i.e. of form $c_i\phi$ with c_i's known, $a_i = \frac{1}{n_i}$ and $\phi = 1$.

<u>Deviance</u>

Saturated model ω_S: $Y_i \sim \text{Bin}(n_i, p_i)$ independent $0 \le p_i \le 1$.

For ω_S: $\frac{\partial \ell}{\partial p_i} = 0$ in (8.1) gives us $\tilde{p}_i = \frac{y_i}{n_i}$

So $\ell_{max}^{(S)} = \sum\left\{y_i \log\left(\frac{y_i}{n_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i}\right) + \log\binom{n_i}{y_i}\right\}$

ω_g $\log\left(\frac{p_i}{1-p_i}\right) = \beta^T x_i$ Let mle's be $\hat{p}_i = p_i(\hat{\beta})$ From (8.1)

$$\ell_{max}^{(g)} = \sum\left\{y_i \log(p_i(\hat{\beta})) + (n_i - y_i)\log(1 - p_i(\hat{\beta})) + \log\binom{n_i}{y_i}\right\}$$

So scaled deviance

$$S(\omega_g, \omega_S) = 2\left\{\ell_{max}^{(S)} - \ell_{max}^{(g)}\right\}$$

$$= \sum\left\{ y_i \log \underbrace{\frac{y_i}{n_i p_i(\hat{\beta})}}_{=\frac{y_i}{e_i}} + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - e_i}\right)\right\}$$

where $e_i = n_i p_i(\hat{\beta}) = n_i \hat{p}_i$

are the estimated expected values under the model ω_g.

Here $\phi = 1$, so deviance is $D(\omega_g, \omega_S) = \phi\, S(\omega_g, \omega_S) = S(\omega_g, \omega_S)$.

It can be shown (expansion for log) that (8.2) is approximately

$$\sum_i \left\{\frac{(y_i - e_i)^2}{e_i} + \frac{[(n_i - y_i) - (n_i - e_i)]^2}{n_i - e_i}\right\}$$

this is Pearson's χ^2.

Both this and deviance are approximately χ^2_{n-p} if ω_g is true.

So if ω_g is a bad fit then deviance will be large compared to χ^2_{n-p}.

Logistic link is the most commonly used.

Other possibilities

$$g(p_i) = \Phi^{-1}(p_i) \quad \text{where } \Phi \text{ is } N(0,1) \text{ distribution function}$$

↑ probit link

$$g(p_i) = \log(-\log(1-p_i)) \quad \text{complementary log-log}$$

Example: logistic regression, see H6.1

<u>9. Poisson data</u> $Y \sim \text{Poisson}(\mu)$

$$P(Y=y) = \frac{e^{-\mu}\mu^y}{y!} = \exp\{y\log(\mu) - \mu - \log(y!)\}$$

$\theta = \log(\mu) \quad b(\theta) = \mu = e^\theta \quad \phi = 1$ (check)

Check that $b'(\theta) = \mu$ and $V(\mu) = \mu$

Canonical link $g(\mu) = \log(\mu)$

$Y_1, \dots, Y_n$ independent $Y_i \sim \text{Poisson}(\mu_i)$

Assume canonical link

$\log(\mu_i) = \beta^T x_i \quad$ so $\mu_i = e^{\beta^T x_i} = \mu_i(\beta)$ say

Loglikelihood is

$$\ell(\beta) = \sum_{i=1}^n \{-\mu_i + y_i\log(\mu_i) - \log(y_i!)\} \qquad (9.1)$$

$$= -\sum_{i=1}^n e^{\beta^T x_i} + \beta^T \sum_{i=1}^n y_i x_i - \sum_{i=1}^n \log(y_i!)$$

$$\frac{\partial \ell}{\partial \beta} = -\sum_{i=1}^n x_i e^{\beta^T x_i} + \sum_{i=1}^n y_i x_i \qquad \text{(check)}$$

mle $\hat\beta$ satisfies $\sum_{i=1}^n y_i x_i = \sum_{i=1}^n e^{\hat\beta^T x_i} x_i \qquad (9.2)$

Also $-\frac{\partial^2 \ell}{\partial\beta\partial\beta^T} = +\sum_{i=1}^n x_i x_i^T e^{\beta^T x_i} = V(\beta)^{-1}$ say

So mle $\hat\beta \overset{\text{approx}}{\sim} N(\beta, V(\beta))$ we may replace $V(\beta)$ by $V(\hat\beta)$.

Using Poisson for rates

Suppose $Y_1, \ldots, Y_n$ are counts for different "exposures" $m_1, \ldots, m_n$

$Y_i \sim \text{Poisson}(\mu_i)$ where $\mu_i = m_i \theta_i$ θ_i is rate

Interest lies in modelling the rate, ie in modelling how

θ_i depends on covariates.

Model $\log(\mu_i) = \log(m_i \theta_i) = \log(m_i) + \underbrace{\log(\theta_i)}_{\log(\theta_i) = \beta^T x_i}$

In this type of situations $\log(m_i)$ is an offset.

It's coefficient is forced to be 1.

Contingency Tables

See H8.1 Tables of counts cross-classified by 2 or more categorical variables

Example $r \times c$ table, n individuals

let $p_{ij} = P(\text{in cell } i,j)$ $\sum_{i,j} = 1$

Let Y_{ij} be # in cell i,j

Let $Y = (Y_{11}, Y_{12}, \cdots, Y_{rc})$

Then $P(Y=y) = \dfrac{n!}{\prod_{i,j} y_{ij}} \prod p_{ij}^{y_{ij}}$ if $\sum_{i,j} y_{ij} = n$; 0 otherwise

i.e. $Y \sim \text{Multinomial}(n, p)$ where $p = (p_{11}, p_{12}, \cdots, p_{rc})$

Would like to model $\log(p_{ij}) = \beta^T x_i$

glm - don't have multinomial

Way out can use Poisson

It can be shown that if $Y_{ij} \sim \text{Poisson}(\mu_{ij})$

then $Y \mid Y_{++} = n \sim \text{Multinomial}\left(n, \left(\frac{\mu_{11}}{\mu_{++}}, \frac{\mu_{12}}{\mu_{++}}, \cdots, \frac{\mu_{rc}}{\mu_{++}}\right)\right)$

so we do Poisson glm modelling, with Y_{++} forced to be n.

III, Non parametric Statistics

Here we do not assume that data come from a parametric family.

See H8.2

$X_1, \ldots, X_m$ $Y_1, \ldots, Y_n$ ← $m+n$ numbers (rank them)

e_1 e_m $e_{m+1}+\Delta$ $e_{m+n}+\Delta$ $Z_1, \ldots, Z_{m+n}$

smallest

if this is Y_1 then $S_1 = 1$

Under H_0 ($m+n$ iid random variables), each of the $\binom{m+n}{n}$ possible assignments of ranks to the Y's, are equally likely.

i.e. each assignment has probability $\frac{1}{\binom{m+n}{n}}$.

So can calculate $P(T \geq t \mid H_0)$ (for small m and n).

Otherwise, use asymptotic normality of T.

Matched pairs (Sec 18.2)

There are 2^n ways of assigning +'s and −'s to the ranks of $|z_1|, \dots, |z_n|$.

Under H_0, each of these is equally likely with prob. $\frac{1}{2^n}$

so can calculate null probs (or use N approx).