More regression modelling

Response variable y

$\underline{k}$ explanatory variables $x = (x_1, \dots, x_k)^T$

Observe y and x for n subjects, hence we have obs. (y_i, x_i) for $i = 1, \dots, n$, where $x_i = (x_{i1}, x_{i2}, \dots, x_{ik})^T$

Useful to write

$$\underset{n\times 1}{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \underset{n\times k}{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & & x_{2k} \\ x_{31} & \vdots & & \vdots \\ \vdots & & & \\ x_{n1} & x_{n2} & & x_{nk} \end{bmatrix} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$$

Generalized linear models (includes linear models as a special case)

- Linear predictor $\eta = \underset{n\times k}{X}\underset{k\times 1}{\beta}$, where $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$ are parameters

  Often $x_{i1} = 1, \forall i$, hence β_1 is an intercept

- Link function: Links the linear predictor to the mean of the response variable

  $g(\mu) = \eta = X\beta$ — Linear regression: $g(\mu) = \mu$, $\mu = X\beta$

  $\mu = g^{-1}(\eta) = g^{-1}(X\beta)$ — Poisson data with log-link-function: $g(\mu) = \log(\mu) = X\beta$, $\mu = \exp\{X\beta\}$

nov 17-12:15

- Distribution of the response variable y with mean $E[y|X] = \mu$, perhaps with a over-dispersion parameter ϕ

  Ordinary linear regression: $y_i \sim N(\mu_i, \sigma^2)$, $i = 1, \dots, n$

  Poisson regr. $y_i \sim Pois(\mu_i)$, $i = 1, \dots, n$

  Ex. over-dispersed Poisson regression with a log-link

  $$y_i \sim Pois(\mu_i)$$
  $$g(\mu_i) = \log(\mu_i)$$
  $$\log \mu_i \sim N(X\beta, \sigma^2)$$

Offsets Often it is appropriate to assume that one of the explanatory variables has a known coefficient, hence is an offset.

Example: Poisson regression, where μ_i is a rate of occurrence per time unit. If y_i is observed for time length t_i, then $E[y_i] = t_i \mu_i$

Now $y_i \sim Pois(E[y_i]) = Pois(t_i \cdot \mu_i)$

$$\log(\mu_i) = \log\left(\frac{E[y_i]}{t_i}\right) = X\beta, \text{ or } \log E[y_i] = \log t_i + X\beta$$

nov 17-12:31

Interpreting the β_i's

Linear regression: β_i is the effect of changing x_i by one unit

GLM with $g(\mu) \neq \mu$:

Consider a "standard" case with x_0 and y_0, then changing x_0 by Δx changes the mean response from $g^{-1}(x_0\beta)$ to

$g^{-1}(g(y_0) \pm (\Delta x)\cdot\beta)$.

Some guidelines for the explanatory variable matrix X

- Identifiability and collinearity – β cannot be uniquely identified from data alone
  - Rank of X less than k
    - $\underline{k > n}$: There are methods to handle this (Lasso, ...)
- Transform the explanatory variables appropriately, or for example: Both x_j, x_j^2, $\log x_j$
- Categorical variables: One indicator variable x_j for each category -1 (when you have an intercept)

nov 17-12:42

The ordinary linear regression with non-inf prior

$$\underset{n\times 1}{y} \mid \beta, \sigma, X \sim N(\underset{n\times k}{X}\underset{k\times 1}{\beta}, \sigma^2 \underset{n\times n}{I})$$

Here: X is fixed, known
Can generalize to having a distr. on X

$p(\beta, \sigma^2) \propto \sigma^{-2}$ ← This gives a proper posterior distribution for (1) $n > k$
(2) Rank of X is k

Posterior distribution given by $p(\beta, \sigma^2|y) = p(\beta|\sigma^2, y)\cdot p(\sigma^2|y)$

$$\beta|\sigma, y \sim N(\hat{\beta}, V_\beta \sigma^2), \quad \hat{\beta} = (X^TX)^{-1}X^Ty, \quad V_\beta = (X^TX)^{-1}$$

$$\sigma^2|y \sim \text{Inv-}\chi^2(n-k, s^2), \quad s^2 = \frac{1}{n-k}(y - X\hat{\beta})^T(y - X\hat{\beta})$$

For k large, $(X^TX)^{-1}$ can be computationally hard. Methods such as QR decomposition can be used

nov 17-12:54