

HW 2 for Stat 7249 - Spring 2009

Due February 3

Reading in text for this assignment

- Chapters 3-5

Datasets

- none for this assignment

1. DB Problem 5.1
2. DB Problem 7.1
3. Anscombe residuals are a type of residuals constructed to more closely follow a normal distribution than Pearson's residuals. The idea is to transform Y , using a function $A(Y)$ so that $A(Y)$ is as close to normal as possible. For likelihoods based on glm's, the function $A(\cdot)$ is given by

$$A(\cdot) = \int \frac{d\mu}{V^{1/3}(\mu)} \quad (1)$$

Then, residuals are based on $A(y) - A(\mu)$. The variance of $A(Y)$ can be (first order) approximated as $A'(\mu)\sqrt{V(\mu)}$. Derive the Anscombe residual for the gamma distribution. Using R, take a sample of size 500 from a gamma distribution and examine the distribution of the Pearson and Anscombe residuals. Comment. Do this for several different values for the parameters of the Gamma distribution.

4. Suppose a population of individuals is partitioned into two sub-populations or groups, G_1 and G_2 . Assume measurements Z from group j , $j = 1, 2$ follow a normal distribution with mean μ_j and covariance matrix Σ . Let z^* be an observation made on an individual drawn at random from the combined population. Show that

$$\text{logit}(P(Y = 1 | Z^*)) = \alpha + \beta^T Z^* \quad (2)$$

where $\alpha = \log(\pi/(1/\pi)) + \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2 - \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1$ and $\beta = \Sigma^{-1}(\mu_1 - \mu_2)$ and $\pi_1 = P(Y = 1)$.

5. Let R_i be the unobserved true binary response for unit i with $\pi_i^* = P(R_i = 1)$ satisfying the logistic model

$$\text{logit}(\pi_i^*) = \beta^T x_i. \quad (3)$$

Suppose that the observed binary response, Y_i is subject to misclassification as follows,

$$\begin{aligned}P(Y_i = 1 \mid R_i = 0) &= \delta_i \\P(Y_i = 0 \mid R_i = 1) &= \epsilon_i.\end{aligned}$$

Suppose the misclassification errors satisfy the following condition

$$\frac{\delta_i}{\epsilon_i} = \frac{\pi_i^*}{1 - \pi_i^*}. \quad (4)$$

Under (4), can we consistently estimate β using the Y_i instead of R_i ? [Hint: Just determine the relationship between $\text{logit}(\pi_i)$ and $\beta^T x_i$, where $\pi_i = P(Y_i = 1)$]. Does (4) seem like a plausible condition? Comment. Finally, would you expect the variance of $\hat{\beta}$ to compare if estimating it using Y_i versus R_i ? Explain.