BST 140.752
Problem Set 2
Due: November 16, 2017

1 Inference

1. There are three frequently occurring test statistics, the likelihood ratio test statistic, the Wald test, and the score test. If $\mathbf{Y}$ has the probability density function $f(\mathbf{y};\boldsymbol{\beta})$ at $\mathbf{Y}=\mathbf{y}$, where $\boldsymbol{\beta}$ is $p\times 1$, then hypothesis of interest are often of the form $H_0 : \mathbf{L}'\boldsymbol{\beta}=\xi$ versus $H_1 : \mathbf{L}'\boldsymbol{\beta}\neq\xi$, where $\mathbf{L}'$ is $s\times p$ of rank $s\leq p$. Let

   - $\hat{\boldsymbol{\beta}}$ denote the maximum likelihood estimate of $\boldsymbol{\beta}$ under the full model,
   - $\tilde{\boldsymbol{\beta}}$ denote the maximum likelihood estimate of $\boldsymbol{\beta}$ under the model assuming the null hypothesis is true,
   - $\ell(\boldsymbol{\beta})=\log[f(\mathbf{y};\boldsymbol{\beta})]$ denote the log likelihood function,
   - $\mathbf{s}(\boldsymbol{\beta})$ be the vector of scores with j^{th} component

   $$s_j(\boldsymbol{\beta})=\frac{\delta\ell(\boldsymbol{\beta})}{\delta\beta_j},$$

   - $\Im(\boldsymbol{\beta})$ be Fisher's information matrix which has j,k element equal to

   $$-E\left[\frac{\delta^2\ell(\boldsymbol{\beta})}{\delta\beta_j\delta\beta_k}\right].$$

   The three test statistics in this case are the likelihood ratio test statistic, given by

   $$-2[\ell(\tilde{\boldsymbol{\beta}})-\ell(\hat{\boldsymbol{\beta}})],$$

   the Wald test statistic, given by

   $$(\mathbf{L}'\hat{\boldsymbol{\beta}}-\mathbf{xi})'[\mathbf{L}'\Im(\hat{\boldsymbol{\beta}})^{-1}\mathbf{L}]^{-1}(\mathbf{L}'\hat{\boldsymbol{\beta}}-\mathbf{xi}),$$

   and the score test, given by

   $$\mathbf{s}'(\tilde{\boldsymbol{\beta}})\Im(\tilde{\boldsymbol{\beta}})^{-1}\mathbf{s}(\tilde{\boldsymbol{\beta}}).$$

   For the linear model $\mathbf{Y}\sim \mathsf{N}_p(\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2,\sigma^2\mathbf{I})$, where $\mathbf{X}_1$ is $n\times q$ of rank q, $\mathbf{X}_2$ is $n\times(p-q)$ of rank $p-q$, $\mathbf{X}=[\mathbf{X}_1,\mathbf{X}_2]$ is $n\times p$ of rank p, and σ^2 is known, derive the three test statistics for testing $H_0 : \boldsymbol{\beta}_2=\mathbf{0}$ versus $H_1 : \boldsymbol{\beta}_2\neq\mathbf{0}$. Comment.

2 Residuals

1. Consider a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \delta\Delta + \boldsymbol{\varepsilon}$ where δ is a vector with a 1 at position i_0 and 0 elsewhere. Argue the following:

   A. The residual corresponding to i_0 is 0 for this model.

   B. The fitted value for β using this model and all of the data is equivalent to the least-squares estimate using only the data with the i_0 observation deleted.

   C. Argue that the standardized PRESS residuals are a test statistic for $\Delta = 0$.

2. Show that the internally and externally studentized residuals (denoted r_i and t_i, respectively) are monotonically related as follows:

$$t_i = r_i \sqrt{\frac{n-p-1}{n-p-r_i^2}}.$$

3 Inference under incorrectly specified models

1. Suppose the true model is given by $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and we fit the model $y_i = \tilde{\beta}_1 x_i + \tilde{\epsilon}_i$. Compute the bias of $\tilde{\beta}_1$.

4 Multiple Comparisons

Recall the family-wise error rate (FWER) and false detection rate (FDR).

1. Show that if all null hypothesis are true, the FDR is equivalent to the FWER.

2. Show that any procedure that controls the FWER also controls the FDR.

5 Coding and data analysis exercises

1. Extend the `R` function `mylm()` you created in a previous homework to return the nternally and externally studentized residuals, the PRESS residuals, and the Cook's distance. Find a dataset to try out your function (you can simulate one if you like).