Spring 2015 Statistics 151a (Linear Models) : Lecture Nine

Aditya Guntuboyina

17 February 2015

1 Hypothesis Testing and GLRT

Consider the linear model $Y = X\beta + e$ with $e \sim N_n(0, \sigma^2 I_n)$. We denote this model by M and refer to it as the full model.

Suppose that we want to test a linear constraint on β. This constraint can be incorporated in the full model to yield a reduced model m. In the last class, we saw that the following test statistic can be used for this:

$$\frac{(RSS(m) - RSS(M))/(p-q)}{RSS(M)/(n-p-1)}$$

where p is the number of explanatory variables in M and q is the number of explanatory variables in m (we are assuming that both M and m are full rank models i.e., the X-matrices have full column rank).

We saw in the last class that under the null hypothesis (i.e., the model m), the above test statistic has the $F_{p-q,n-p-1}$ distribution. Therefore, the p-value is given by

$$\mathbb{P}\left\{F_{p-q,n-p-1} > \frac{(RSS(m) - RSS(M))/(p-q)}{RSS(M)/(n-p-1)}\right\}.$$

We arrived at this test via an argument based on the fact that it is quite natural to look at $RSS(m) - RSS(M)$ and then adjusting this because σ^2 is unknown. However, one can ask if there are any optimality properties associated with this test. It turns out that this test is the Generalized Likelihood Ratio Test (GLRT) in this setting.

Consider a general testing situation where θ is an unknown vector parameter which takes values in Θ which is a vector space of dimension k. Suppose Θ_0 is a subspace of Θ with dimension l with $l < k$. Then in order to test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \notin \Theta_0$, the GLRT is based on the test statistic:

$$\ell(y) = \frac{\sup_{\theta \in \Theta_0} f_\theta(y)}{\sup_{\theta \in \Theta} f_\theta(y)} \tag{1}$$

where $f_\theta(y)$ denotes the density of the observations when the true parameter is θ. Clearly $\ell(y)$ is in the range $[0, 1]$. The closer it is to zero, the less credible is the null hypothesis. Thus the GLRT rejects H_0 when $\ell(y) \leq c$ where c is determined by the level condition. The famous Wilks's theorem states that subject to some regularity conditions, $-2\log \ell(y)$ has an asymptotic χ^2_{k-l} distribution under H_0 as the sample size goes to infinity. The GLRT often leads one to optimal tests (UMP or UMPU) when these exist. Such testing optimality is beyond the scope of this course.

In our setting of testing $H_0 : m$ against $H_1 : M$ in the linear model, what is the GLRT? Recall that m is obtained from M by putting a constraint on β. The GLRT statistic is

$$\ell = \frac{\max_{\sigma^2, \beta \textbf{ constrained}} (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\|Y - X\beta\|^2\right]}{\max_{\sigma^2, \beta} (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\|Y - X\beta\|^2\right]}$$

In the denominator the maximizers of β and σ^2 are the maximum likelihood estimators: $\hat{\beta}_{ML}$ minimizes $\|Y - X\beta\|^2$ over all β and $\hat{\sigma^2}_{ML}$ equals

$$\hat{\sigma^2}_{ML} := \frac{1}{n} \min_{\beta} \|Y - X\beta\|^2 = \frac{RSS(M)}{n}$$

Thus $\hat{\beta}_{ML}$ is the same as the least squares estimate while $\hat{\sigma^2}_{ML}$ is different from the previous unbiased estimator we used which was $RSS(M)/(n-p-1)$. Asymptotically (as n becomes much larger than p), both estimators give the same result but when n is comparable to p, one gets different results.

Coming back to ℓ, in the numerator, the maximizers of β is the minimizer of $\|Y - X\beta\|^2$ under the constraint and the maximizer of σ^2 is:

$$\frac{\min_{\beta \textbf{ constrained}} \|Y - X\beta\|^2}{n} = \frac{RSS(m)}{n}.$$

Plugging these maximizing values in the definition of ℓ, we obtain

$$\ell = \left(\frac{RSS(m)}{RSS(M)} \right)^{-n/2}.$$

This is clearly a decreasing function of $(RSS(m) - RSS(M))/RSS(M)$. Thus the GLRT is equivalent to rejecting when $(RSS(m) - RSS(M))/RSS(M)$ is large. Taking the null distribution of this statistic into account, it follows that the GLRT is equivalent to the test based on (1).

2 Confidence Intervals for β_j

Because $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$, we have $\hat{\beta}_j \sim N(\beta_j, \sigma^2 v_j)$ where v_j is the corresponding diagonal entry of $(X^TX)^{-1}$. A $100(1-\alpha)$ % C.I for β_j is therefore given by

$$\hat{\beta}_j \pm z_{\alpha/2}\sigma\sqrt{v_j}.$$

But σ is not known, so we use the fact that

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{v_j}} \sim t_{n-p-1}$$

to construct the following $100(1-\alpha)$ % C.I for β_j:

$$\hat{\beta}_j \pm t^{\alpha/2}_{n-p-1}\hat{\sigma}\sqrt{v_j}.$$

Because $\hat{\sigma}\sqrt{v_j}$ is the standard error for $\hat{\beta}_j$, we can write this C.I as

$$\hat{\beta}_j \pm t^{\alpha/2}_{n-p-1} s.e(\hat{\beta}_j).$$

If this interval contains the value 0, it means that the hypothesis $H_0 : \beta_j = 0$ will not be rejected at the α level.

3 Prediction Intervals

Suppose we get a new subject whose explanatory variables are $x_{01}, \ldots, x_{0p}$. What would be our prediction for its response? Our linear model says that the response for this new subject will be $y_0 = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + e_0$. Because β is estimated by $\hat{\beta}$ and e_0 is a zero mean error, our prediction for its response is simply $\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \ldots \hat{\beta}_p x_{0p}$.

What is the uncertainty in this prediction? This is captured by providing a prediction interval. There are usually two kinds of prediction intevals:

1. Interval for the mean response
2. Interval for the response

3.1 Interval for the mean response

The mean response is just $\beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p}$. So this interval is just a confidence interval for this parameter. Write $x_0 := (1, x_{01}, \ldots, x_{0p})^T$ so that

$$\beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} = x_0^T \beta.$$

How to find a $100(1-\alpha)\%$ C.I for $x_0^T\beta$? Observe that

$$\frac{x_0^T\hat{\beta} - x_0^T\beta}{\hat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0}} \sim t_{n-p-1}.$$

Therefore

$$x_0^T\hat{\beta} \pm t_{n-p-1}^{(\alpha/2)}\hat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0} \tag{2}$$

is a $100(1-\alpha)\%$ C.I for $x_0^T\beta$.

3.2 Interval for the response

The response for the new subject with these explanatory variables is given by $y_0 = x_0^T\beta + e_0$. It is therefore more sensible to obtain an interval for y_0 itself instead of $x_0^T\beta$. Because e_0 has mean zero, it is natural to center an interval for y_0 around $\hat{y_0} = x_0^T\hat{\beta}$. We want to find a such that

$$\mathbb{P}\{y_0 \in [\hat{y_0} - a, \hat{y_0} + a]\} = \mathbb{P}\{y_0 - \hat{y_0} \in [-a, a]\} = 1 - \alpha.$$

For finding a, we need to look at the distribution of $y_0 - \hat{y_0} = y_0 - x_0^T\hat{\beta}$. It is easy to see that $y_0 - \hat{y_0}$ has a normal distribution with mean zero and variance $\sigma^2\left(1 + x_0^T(X^TX)^{-1}x_0\right)$. Therefore

$$\frac{y_0 - \hat{y_0}}{\hat{\sigma}\sqrt{1 + x_0^T(X^TX)^{-1}x_0}} \sim t_{n-p-1}$$

It is obvious above that $\hat{\sigma}$ and $y_0 - \hat{y_0}$ are independent?

Based on the above t-distribution, the interval

$$x_0^T\hat{\beta} \pm t_{n-p-1}^{(\alpha/2)}\hat{\sigma}\sqrt{1 + x_0^T(X^TX)^{-1}x_0} \tag{3}$$

presents a $100(1-\alpha)\%$ interval for the future response y_0. This is called the prediction interval for the response corresponding to the explanatory variable values $x_{01}, \ldots, x_{0p}$. Note the difference between the intervals in (2) and (3). The interval in (3) also takes into account the randomness present in the error e_0 and is thus wider than the interval in (2).

The additional width in the prediction interval compared to the confidence interval accounts for the error in observing $x_0^T\beta$. The difference between the widths of the two intervals can be quite substantial when $x_0^T(X^TX)^{-1}x_0$ is small which typically happens when the sample size n is large. The prediction error of $x_0^T\hat{\beta}$ equals the sum of the estimation error of $x_0^T\beta$ (which is $x_0^T\hat{\beta} - x_0^T\beta$) and the deviation of the observation y_0 from its mean (which is $y_0 - x_0^T\beta$). We can hope to reduce the estimation error by using a lot of data, but we still have to allow for the variablility in the observations while constructing the prediction interval. The latter component does not depend on n.