Advanced Methods in Biostatistics II

Lecture 1

October 24, 2017

Linear model

- Consider the linear model

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

where $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

- Today we begin discussing hypothesis testing in the context of the linear model.
- Hypothesis testing provides a formal tool for choosing between a reduced model and an associated full model.

Test of overall regression

- Let us write $\mathbf{X} = [\mathbf{J}_n \;\; \mathbf{X}_1]$ and $\boldsymbol{\beta} = (\beta_0 \;\; \boldsymbol{\beta}_1')'$.
- Here $\mathbf{J}_n$ is an n-dimensional vector of ones and $\mathbf{X}_1$ is an $n \times (p-1)$ matrix.
- Suppose now we want to test $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$.
- This is sometimes referred to as the overall regression hypothesis as it tests the significance of all explanatory variables except the intercept term.

Sums of square

- Recall, that we can partition the sums of square as follows:

$$\mathbf{y}'(\mathbf{I}-\mathbf{H_J})\mathbf{y} = \mathbf{y}'(\mathbf{I}-\mathbf{H_X})\mathbf{y} + \mathbf{y}'(\mathbf{H_X}-\mathbf{H_J})\mathbf{y}$$

- Alternatively, we can express this as follows:

$$||\mathbf{y} - \bar{y}\mathbf{J}_n||^2 = ||\mathbf{y} - \hat{\mathbf{y}}||^2 + ||\hat{\mathbf{y}} - \bar{y}\mathbf{J}_n||^2.$$

Sums of square

- We refer to $||\mathbf{y} - \bar{y}\mathbf{J}_n||^2$ as the sum of square total (SST), $||\mathbf{y} - \hat{\mathbf{y}}||^2$ as the sum of square error (SSE), and $||\hat{\mathbf{y}} - \bar{y}\mathbf{J}_n||^2$ as the sum of square regression (SSR).
- Hence, we can write $SST = SSR + SSE$.

Mean squares

- Note, that

$$SSR/\sigma^2 \sim \chi^2_{p-1}(\lambda)$$

where $\lambda = \beta'\mathbf{X}'(\mathbf{H_X} - \mathbf{H_J})\mathbf{X}\beta/2\sigma^2$.

- In addition,

$$SSE/\sigma^2 \sim \chi^2_{n-p}.$$

- Also, note that *SSE* and *SSR* are independent.

F-statistic

- Hence,

$$F = \frac{SSR/(p-1)}{SSE/(n-p)} \sim F_{p-1,n-p}(\lambda)$$

with $\lambda = \beta' \mathbf{X}' (\mathbf{H_X} - \mathbf{H_J}) \mathbf{X} \beta / 2\sigma^2$.

F-statistic

- Importantly, under H_0 the term $\lambda = 0$ and $F \sim F_{p-1,n-p}$.
- However, if H_0 is false we need to use the non-central F-distribution.
- This is an important result in the context of power analysis.

Testing

- To test $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$, we reject H_0 if $F \geq F_{p-1,n-p,1-\alpha}$.
- Here the term $F_{p-1,n-p,1-\alpha}$ is the upper α percentile of the central F-distribution.

Mean squares

- The ratio of the sums of square to the degrees of freedom gives the mean squares.
- For example,

$$MSR = \frac{SSR}{p-1}$$

and

$$MSE = \frac{SSE}{n-p}.$$

Expected mean square

- As the F-statistic is the ratio of MSR and MSE we can motivate it by studying their expected values.
- This allows us to gain a better understanding of the behavior of the test.

Expected mean square

- Note, we have:

$$
\begin{aligned}
E(MSR) &= E(SSR/(p-1)) \\
&= \frac{\sigma^2}{p-1} E(SSR/\sigma^2) \\
&= \frac{\sigma^2}{p-1} ((p-1) + 2\lambda) \\
&= \sigma^2 + \frac{1}{p-1} \beta' \mathbf{X}' (\mathbf{H_X} - \mathbf{H_J}) \mathbf{X} \beta
\end{aligned}
$$

Expected mean square

- Similarly,

$$
\begin{aligned}
E(MSE) &= E\left(\frac{SSE}{n-p}\right) \\
&= \frac{\sigma^2}{n-p} E(SSE/\sigma^2) \\
&= \sigma^2
\end{aligned}
$$

F-statistic

- If H_0 holds, then both expected values will equal σ^2 and therefore $F \approx 1$.
- If $\beta_2 \neq 0$, then $E(SSR/p) > \sigma^2$ and $F > 1$.
- We therefore reject H_0 for large values of F.

Example

- Data was collected on 100 houses recently sold in a city.
- It consisted of the sales price (in \$), house size (in square feet), the number of bedrooms, the number of bathrooms, the lot size (in square feet), and the annual real estate tax (in \$).
- Want to fit a model for sales price as a function of the five other variables.

Example

```
> Housing = read.table("housing.txt",header=TRUE)
> head(Housing)
  Taxes Bedrooms Baths  Price Size   Lot
1  1360        3   2.0 145000 1240 18000
2  1050        1   1.0  68000  370 25000
3  1010        3   1.5 115000 1130 25000
4   830        3   2.0  69000 1120 17000
5  2150        3   2.0 163000 1710 14000
6  1230        3   2.0  69900 1010  8000
```

Example

```
> results = lm(Price ~ Size + Lot + Taxes + Bedrooms + Baths, data=Housing)
> summary(results)

Call:
lm(formula = Price ~ Size + Lot + Taxes + Bedrooms + Baths, data = Housing)

Residuals:
   Min     1Q Median     3Q    Max
-89978 -16931  -1407  19077  73705

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6633.7997 15834.6177   0.419 0.676214
Size           33.5714     8.8904   3.776 0.000279 ***
Lot             1.6162     0.4948   3.266 0.001522 **
Taxes          20.6436     5.2558   3.928 0.000163 ***
Bedrooms    -6469.6862  5313.1550  -1.218 0.226396
Baths       11824.4881  7320.9445   1.615 0.109628
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 27980 on 94 degrees of freedom
Multiple R-squared:  0.7659,Adjusted R-squared:  0.7535
F-statistic: 61.52 on 5 and 94 DF,  p-value: < 2.2e-16
```

Example

- Test: $H_0 : \beta_1 = \ldots = \beta_5 = 0$.
- The output shows that F = 61.52 ($p < 2.2e - 16$), indicating that we should clearly reject the null hypothesis that the explanatory variables collectively have no effect on Price.

Test of a subset of variables

- Let $\mathbf{X} = [\mathbf{X}_1 \;\; \mathbf{X}_2]$ and $\beta = (\beta_1' \;\; \beta_2')'$.
- Here we assume that $\mathbf{X}_1$ is $n \times p_1$ and $\mathbf{X}_2$ is $n \times p_2$.
- Then,

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \varepsilon \\ &= \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon. \end{aligned}$$

Testing a subset

- Consider testing $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$
- This becomes a problem of comparing the full model with a reduced model $\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1^* + \epsilon^*$.
- Note we typically incorporate the intercept term into $\mathbf{X}_1$.

Example

- Consider the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon.$$

- Suppose we seek to test the hypothesis

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0.$$

- If H_0 is rejected, we would choose the full second-order model over the reduced first-order model.

Partitioning the data

- Consider the following partitioning:

$$\mathbf{y}'(\mathbf{I}-\mathbf{H_J})\mathbf{y} = \mathbf{y}'(\mathbf{I}-\mathbf{H_X})\mathbf{y} + \mathbf{y}'(\mathbf{H_X}-\mathbf{H_{X_1}})\mathbf{y} + \mathbf{y}'(\mathbf{H_{X_1}}-\mathbf{H_J})\mathbf{y}$$

- We can alternatively write this as:

$$SST = SSE + SS(\beta_2|\beta_1) + SSR(reduced).$$

- Here

$$SS(\beta_2|\beta_1) = \mathbf{y}'(\mathbf{H_X}-\mathbf{H_{X_1}})\mathbf{y}$$

and

$$SSR(reduced) = \mathbf{y}'(\mathbf{H_{X_1}}-\mathbf{H_J})\mathbf{y}.$$

Extra sum of squares

- Note: $SS(\beta_2|\beta_1) = SSR(\mathit{full}) - SSR(\mathit{reduced})$.
- This term, denoted the extra sum of squares, can be viewed as the marginal increase in the regression sum of squares when including additional parameters to the model.

Extra sum of squares

- If $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ is true, we would expect $SS(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)$ to be small.
- It is important to note that this hypothesis tests whether $\boldsymbol{\beta}_2$ contributes in addition to $\boldsymbol{\beta}_1$.

Sums of square

- Recall from a previous lecture that

$$
\begin{aligned}
\sigma^{-2}\mathbf{y}'(\mathbf{H}_{\mathbf{X}} - \mathbf{H}_{\mathbf{X}_1})\mathbf{y} &\sim \chi^2_{p_2}(\beta'\mathbf{X}'(\mathbf{H}_{\mathbf{X}} - \mathbf{H}_{\mathbf{X}_1})\mathbf{X}\beta/2\sigma^2) \\
\sigma^{-2}\mathbf{y}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}})\mathbf{y} &\sim \chi^2_{n-p}
\end{aligned}
$$

- In addition, $\mathbf{y}'(\mathbf{H}_{\mathbf{X}} - \mathbf{H}_{\mathbf{X}_1})\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}})\mathbf{y}$ are independent.

F-statistic

- Thus,

$$F = \frac{\mathbf{y}'(\mathbf{H_X} - \mathbf{H_{X_1}})\mathbf{y}/p_2}{\mathbf{y}'(\mathbf{I} - \mathbf{H_X})\mathbf{y}/(n-p)} \sim F_{p_2, n-p}(\lambda)$$

  where $\lambda = \beta'\mathbf{X}'(\mathbf{H_X} - \mathbf{H_{X_1}})\mathbf{X}\beta/2\sigma^2$.

Expected mean square

- Note, we have:

$$\begin{aligned} E(MSR(\beta_2|\beta_1)) &= E(SSR(\beta_2|\beta_1)/p_2) \\ &= \frac{\sigma^2}{p_2} E(SSR(\beta_2|\beta_1)/\sigma^2) \\ &= \frac{\sigma^2}{p_2}(p_2 + 2\lambda) \\ &= \sigma^2 + \frac{1}{p_2}\beta'\mathbf{X}'(\mathbf{H}_{\mathbf{X}} - \mathbf{H}_{\mathbf{X}_1})\mathbf{X}\beta \end{aligned}$$

Expected mean square

- Similarly,

$$
\begin{aligned}
E(MSE) &= E\left(\frac{SSE}{n-p}\right) \\
&= \frac{\sigma^2}{n-p} E(SSE/\sigma^2) \\
&= \sigma^2
\end{aligned}
$$

F-statistic

- If H_0 holds, then both expected values will equal σ^2 and therefore $F \approx 1$.
- If $\beta_2 \neq 0$, then $E(SSR(\beta_2|\beta_1)/p_2) > \sigma^2$ and $F > 1$.
- We therefore reject H_0 for large values of F.

Testing

- Note if $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ is true, then $\lambda = 0$ and $F \sim F_{p_2, n-p}$
- To test $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$, we reject H_0 if $F \geq F_{p_2, n-p, 1-\alpha}$.
- Here $F_{p_2, n-p, 1-\alpha}$ is the upper α percentile of the central F-distribution.

Example

- Suppose we include the variables bedroom, bath, size and lot in our model and are interested in testing whether the number of bedrooms and bathrooms are significant after taking size and lot into consideration.

R code

```
> reduced = lm(Price ~ Size + Lot, data=Housing)
> full = lm(Price ~ Size + Lot + Bedrooms + Baths, data=Housing)
> anova(reduced, full)
Analysis of Variance Table

Model 1: Price ~ Size + Lot
Model 2: Price ~ Size + Lot + Bedrooms + Baths
  Res.Df        RSS Df  Sum of Sq      F  Pr(>F)
1     97 9.0756e+10
2     95 8.5672e+10  2 5083798629 2.8186 0.06469 .
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

Example

- Since $F = 2.82$ ($p = 0.0647$) we cannot reject the null hypothesis at the 5% significance level.
- It appears that the variables Bedrooms and Baths do not contribute significant information to Price once the variables Size and Lot have been taken into consideration.

General linear hypothesis test

- Now consider testing the hypothesis that

$$H_0 : \mathbf{K}\beta = \mathbf{0}$$

  for $\mathbf{K}$ of rank q.

- This is known as the general linear hypothesis, and the two previous cases discussed today are special cases.

General linear hypothesis test

- Note that $\mathbf{K}\hat{\beta} \sim N_q(\mathbf{K}\beta, \mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}'\sigma^2)$.
- Therefore,

$$(\mathbf{K}\hat{\beta})'\{\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}'\sigma^2\}^{-1}\mathbf{K}\hat{\beta} \sim \chi^2_q(\lambda)$$

  where $\lambda = (\mathbf{K}\beta)'\{\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}'\}^{-1}\mathbf{K}\beta/2\sigma^2$.
- Furthermore, this quadratic form is independent of s^2.

General linear hypothesis test

- Let,

$$F = \frac{(\mathbf{K}\hat{\beta})'\{\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}'\}^{-1}\mathbf{K}\hat{\beta}}{qs^2}.$$

- If $H_0 : \mathbf{K}\beta = \mathbf{0}$ is false, then $F \sim F_{q,n-p}(\lambda)$
- If $H_0 : \mathbf{K}\beta = \mathbf{0}$ is true, then $\lambda = 0$ and $F \sim F_{q,n-p}$
- To test $H_0 : \mathbf{K}\beta = \mathbf{0}$, we reject H_0 if $F \geq F_{q,n-p,1-\alpha}$.

General linear hypothesis test

- To test $H_0 : \mathbf{K}\beta = \mathbf{t}$, we instead use the fact that

$$\mathbf{K}\hat{\beta} - \mathbf{t} \sim N_q(\mathbf{K}\beta - \mathbf{t}, \mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}'\sigma^2).$$

- Therefore,

$$(\mathbf{K}\hat{\beta} - \mathbf{t})'\{\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}'\sigma^2\}^{-1}(\mathbf{K}\hat{\beta} - \mathbf{t}) \sim \chi^2_q(\lambda)$$

where $\lambda = (\mathbf{K}\beta - \mathbf{t})'\{\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}'\}^{-1}(\mathbf{K}\beta - \mathbf{t})/2\sigma^2$.

- Furthermore, this quadratic form is independent of s^2.

General linear hypothesis test

- Let,

$$F = \frac{(\mathbf{K}\hat{\beta} - \mathbf{t})'\{\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}'\}^{-1}(\mathbf{K}\hat{\beta} - \mathbf{t})}{qs^2}.$$

- If $H_0 : \mathbf{K}\beta = \mathbf{t}$ is false, then $F \sim F_{q,n-p}(\lambda)$
- If $H_0 : \mathbf{K}\beta = \mathbf{t}$ is true, then $\lambda = 0$ and $F \sim F_{q,n-p}$
- To test $H_0 : \mathbf{K}\beta = \mathbf{t}$, we reject H_0 if $F \geq F_{q,n-p,1-\alpha}$.