Advanced Methods in Biostatistics II

Lecture 2

October 26, 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 will continue discussing hypothesis testing in the context of the linear model.

- We will also discuss confidence intervals and prediction intervals.

Testing

- Let us consider the linear contrast $\mathbf{c}'\beta$, where $\mathbf{c}$ is a vector of length p.
- Recall that the uniformly minimum variance unbiased estimator of $\mathbf{c}'\beta$ is given by $\mathbf{c}'\hat{\beta}$.
- We now consider methods for testing:

$$H_0 : \mathbf{c}'\beta = 0.$$

General linear hypothesis test

- One way to approach the problem is to use the general linear hypothesis framework discussed last time.
- In this setting, we may consider testing the hypothesis:

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

where $\mathbf{K} = \mathbf{c}'$.

General linear hypothesis test

- Recall from last lecture the statistic:

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

- If $H_0 : \mathbf{K}\beta = \mathbf{0}$ is true, then $F \sim F_{q,n-p}$.

General linear hypothesis test

- If $\mathbf{K} = \mathbf{c}'$, then

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

- If $H_0 : \mathbf{c}'\beta = \mathbf{0}$ is true, then $F \sim F_{1,n-p}$

General linear hypothesis test

- To test $H_0 : \beta_j = 0$, we use $\mathbf{c} = (0, \ldots, 0, 1, 0, \ldots, 0)$, where the 1 is in the j^{th} position.
- This gives,

$$F = \frac{\hat{\beta}_j^2}{s^2 g_{jj}}$$

where g_{jj} is the j^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.

General linear hypothesis test

- Since the F-statistic has 1 and $n - p$ degrees of freedom, we can equivalently use the t-statistic:

$$t = \frac{\hat{\beta}_j}{s\sqrt{g_{jj}}}$$

- If $H_0 : \mathbf{c}'\beta = \mathbf{0}$ is true, then $t \sim t_{n-p}$.
- We reject $H_0 : \beta_j = 0$ if $|t| \geq t_{n-p,1-\alpha/2}$

T statistics

- As an alternative approach recall that

$$\mathbf{c}'\hat{\beta} \sim N(\mathbf{c}'\beta, \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}\sigma^2).$$

- Therefore,

$$\frac{\mathbf{c}'\hat{\beta} - \mathbf{c}'\beta}{\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}\sigma^2}} \sim N(0,1)$$

- Furthermore, $\mathbf{c}'\hat{\beta}$ and s^2 are independent.

T statistics

- Hence,

$$\frac{\mathbf{c}'\hat{\beta} - \mathbf{c}'\beta}{\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}\sigma^2}} \Big/ \sqrt{\frac{n-p}{\sigma^2} s^2/(n-p)}$$

  is a standard normal divided by the square root of an independent χ^2 over its degrees of freedom.

- Therefore,

$$\frac{\mathbf{c}'\hat{\beta} - \mathbf{c}'\beta}{s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \sim t_{n-p}.$$

Testing

- To test $H_0 : \beta_j = 0$, we use $\mathbf{c} = (0, \ldots, 0, 1, 0, \ldots, 0)$.
- This again gives

$$t = \frac{\hat{\beta}_j}{s\sqrt{g_{jj}}}.$$

- If $H_0 : \mathbf{c}'\beta = \mathbf{0}$ is true, then $t \sim t_{n-p}$.

Confidence and prediction intervals

- Next we use our distributional results to create confidence intervals and prediction intervals for the linear model.
- In particular, we focus on creating a confidence region for $\boldsymbol{\beta}$, confidence intervals for β_j, $\mathbf{c}'\boldsymbol{\beta}$, $E(y)$, and σ^2, as well as prediction intervals for future observations.

Confidence ellipsoids

- We begin by exploring joint confidence regions for the elements $\beta_1, \beta_2, \ldots \beta_p$ in $\boldsymbol{\beta}$.
- First, consider the F statistic:

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

- Now, let $\mathbf{K} = \mathbf{I}$ and $\mathbf{t} = \boldsymbol{\beta}$. This implies that $q = p$ and

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{ps^2} \sim F_{p,n-p}.$$

Confidence ellipsoids

- Now it must hold that:

$$P\left(\frac{(\hat{\beta}-\beta)'\mathbf{X}'\mathbf{X}(\hat{\beta}-\beta)}{ps^2} \leq F_{p,n-p,1-\alpha}\right) = 1-\alpha.$$

- Hence, a $100(1-\alpha)\%$ joint confidence region for β is defined to consist of all vectors β that satisfy:

$$(\hat{\beta}-\beta)'\mathbf{X}'\mathbf{X}(\hat{\beta}-\beta) \leq ps^2 F_{p,n-p,1-\alpha}.$$

- For $p = 2$ this is an ellipse. For $p > 2$ this is a hyperellipse.

Confidence ellipsoids

- This multivariate form of a confidence interval is called a confidence ellipse.
- They are most useful when the dimension is such that we can visualize it as an actual ellipse.
- For larger dimensions we instead focus on obtaining confidence intervals for each individual element of β.

Confidence interval for β_j

- Turning our attention to confidence intervals for β_j, we begin by recalling that

$$t = \frac{\hat{\beta}_j - \beta_j}{s\sqrt{g_{jj}}}$$

follows a t_{n-p} distribution.

- Thus,

$$P\left(-t_{n-p,1-\alpha/2} \leq \frac{\hat{\beta}_j - \beta_j}{s\sqrt{g_{jj}}} \leq t_{n-p,1-\alpha/2}\right) = 1 - \alpha$$

Confidence interval for β_j

- Solving the inequality for β_j, gives

$$P(\hat{\beta}_j - t_{n-p,1-\alpha/2} s \sqrt{g_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{n-p,1-\alpha/2} s \sqrt{g_{jj}}) = 1 - \alpha$$

- Thus, a $100(1-\alpha)\%$ confidence interval for β_j is given by:

$$\hat{\beta}_j \pm t_{n-p,1-\alpha/2} s \sqrt{g_{jj}}$$

Confidence interval for $\mathbf{c}'\beta$

- Similarly,

$$\frac{\mathbf{c}'\hat{\beta} - \mathbf{c}'\beta}{s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \sim t_{n-p}.$$

- Thus, following a similar procedure as above, we can show that a $100(1-\alpha)\%$ confidence interval for $\mathbf{c}'\beta$ is given by:

$$\mathbf{c}'\hat{\beta} \pm t_{n-p,1-\alpha/2}\, s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}$$

Confidence interval for $E(y)$

- Let $\mathbf{x}_0$ denote a particular choice of $\mathbf{x}$, and let y_0 be the corresponding observation.
- Then, we can write

$$y_0 = \mathbf{x}_0'\boldsymbol{\beta} + \epsilon$$

and

$$E(y_0) = \mathbf{x}_0'\boldsymbol{\beta}.$$

Confidence interval for $E(y)$

- Suppose we want to find a confidence interval for $E(y_0)$, i.e., the mean of the distribution of y corresponding to $\mathbf{x}_0$.
- The minimum variance unbiased estimator of $E(y_0)$ is given by $\mathbf{x}_0'\hat{\beta}$.
- Since this is of the form $\mathbf{c}'\hat{\beta}$ we can write a $100(1-\alpha)\%$ confidence interval using the previous result as follows:

$$\mathbf{x}_0'\hat{\beta} \pm t_{n-p,1-\alpha/2}s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$$

Prediction intervals

- A confidence interval for a future observation y_0 corresponding to $\mathbf{x}_0$ is called a prediction interval.
- We give it a specific name because y_0 is an individual observation, and thus a random variable rather than a parameter.
- This impacts the width of the subsequent interval.

Prediction intervals

- While $E(y_0)$ is the mean of the distribution of y at $\mathbf{x}_0$, y_0 instead represents the prediction of an individual outcome drawn from the distribution of y at $\mathbf{x}_0$.
- The point estimate will be the same for both, i.e. $\mathbf{x}_0'\hat{\beta}$.
- However, the variance is larger when predicting an individual outcome due to the additional variation of an individual about the mean.

Prediction intervals

- For a prediction interval, we seek to estimate a range of possible values for y at $\mathbf{x}_0$, a different statement than trying to estimate the average value of y at $\mathbf{x}_0$.
- As the number of observations increase, the estimate of the average should improve.
- However, predicting a single new value involves intrinsic variability that remains no matter how much data we use to build our model.

Example

- Consider the following two tasks: (i) guessing the sales price of a diamond given its weight; and (ii) guessing the average sales price of diamonds given a particular weight.
- With enough data, we should be able to estimate the average sale price very precisely.
- However, we still won't know the exact sales price of an individual diamond of that weight.

Prediction intervals

- Consider estimating y_0 at $\mathbf{x} = \mathbf{x}_0$.
- Note that the random variables y_0 and $\hat{y}_0 = \mathbf{x}_0'\hat{\beta}$ are independent because y_0 is a future observation obtained independently of the n observations used to compute $\hat{y}_0$.
- Hence,

$$Var(y_0 - \mathbf{x}_0'\hat{\beta}) = (1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)\sigma^2$$

Prediction intervals

- Since $E(y_0 - \mathbf{x}_0'\hat{\beta}) = 0$ and s^2 is independent of both y_0 and $\hat{y}_0$, the statistic

$$t = \frac{y_0' - \mathbf{x}_0'\hat{\beta}}{s\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}}$$

follows a t_{n-p} distribution.

Prediction intervals

- Therefore, a $100(1-\alpha)\%$ prediction interval for $\hat{y}_0$ can be written:

$$\mathbf{x}_0'\hat{\beta} \pm t_{n-p,1-\alpha/2}\, s\sqrt{1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$$

- Note that term $\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$ tends to be smaller than 1.
- Hence, prediction intervals for y_0 tend to be much wider than confidence intervals for $E(y_0)$.

R code

- In R the function predict() can be used to make both confidence and prediction intervals.
- To make confidence intervals for the mean response use the option interval="confidence".
- To make a prediction interval instead use the option interval="prediction".

Example

- The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier.
- Data was collected from 45 recent calls to perform routine preventive maintenance service; for each call, x is the number of copiers serviced and y is the total number of minutes spent by the service person.

Example

Estimate the model parameters.

```
> results = lm(Time ~ Copiers)

> results

Call:
lm(formula = Time ~ Copiers)

Coefficients:
(Intercept)      Copiers
    -0.5802      15.0352
```

Example

Test whether there is a linear association between variables.

```
> summary(results)

Call:
lm(formula = Time ~ Copiers)

Residuals:
     Min       1Q   Median       3Q      Max
-22.7723  -3.7371   0.3334   6.3334  15.4039

Coefficients:
                  Estimate   Std. Error  t value    Pr(>|t|)
(Intercept)  -0.5802     2.8039      -0.207    0.837
Copiers      15.0352    0.4831      31.123   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 8.914 on 43 degrees of freedom
Multiple R-Squared: 0.9575,     Adjusted R-squared: 0.9565
F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

Example

Construct individual 95% confidence intervals for the parameters β.

```
> confint(results)
                        2.5 %       97.5 %
(Intercept)   -6.234843    5.074529
Copiers       14.061010   16.009486
```

Example

Construct a 95% confidence interval for the mean service time on calls in which five copiers are serviced.

```
> predict(results,data.frame(Copiers =5),interval="confidence")
           fit          lwr          upr
[1,] 74.59608     71.91422     77.27794
```

Example

Construct a 95% prediction interval for the service time on the next call in which five copiers are serviced.

```
> predict(results,data.frame(Copiers =5),interval="prediction")
          fit         lwr              upr
[1,] 74.59608    56.42133     92.77084
```

Confidence interval for the variance

- Next we seek to develop a confidence interval for the variance.
- Recall that,

$$\frac{n-p}{\sigma^2}s^2 \sim \chi^2_{n-p}.$$

Confidence interval for the variance

- Therefore

$$P\left(\chi^2_{n-p,\alpha/2} \leq \frac{(n-p)s^2}{\sigma^2} \leq \chi^2_{n-p,1-\alpha/2}\right) = 1-\alpha$$

- Solving for σ^2 yields the $100(1-\alpha)\%$ confidence interval:

$$\frac{(n-p)s^2}{\chi^2_{n-p,1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-p)s^2}{\chi^2_{n-p,\alpha/2}}$$