Advanced Methods in Biostatistics I

Lecture 15

October 17, 2017

Linear model

- Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1,p-1} \\ x_{20} & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Least-squares estimate

- The least-squares estimate is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

- The vector of fitted values is given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}.$$

- The vector of residuals is given by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}.$$

Distributional results

- Let us assume the that errors are uncorrelated with mean zero and common variance, i.e. $E[\varepsilon] = \mathbf{0}$ and $\mathrm{var}(\varepsilon) = \sigma^2\mathbf{I}$.
- These assumptions imply that

$$E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$$

and

$$\mathrm{var}(\mathbf{y}) = \sigma^2\mathbf{I}.$$

Least squares estimate

- The least squares estimate is unbiased:

$$E[\hat{\beta}] = \beta.$$

- The covariance matrix of the least squares estimate is

$$\mathrm{var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

- The least squares estimator is the best linear unbiased estimator (BLUE).

Distributional results

- If $\mathbf{X}$ has rank p, we can define

$$\begin{aligned} s^2 &= (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})/(n-p) \\ &= RSS/(n-p). \end{aligned}$$

- s^2 is an unbiased estimate of σ^2.

Distributional results

- Now let us now assume that ε also follows a multivariate normal distribution, i.e. $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$.
- This implies that $\mathbf{y} \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I})$.

Properties

- It is relatively straightforward to show that
$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$
- Normality holds since $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is a linear function of $\mathbf{y}$ of the form $\hat{\boldsymbol{\beta}} = \mathbf{A}\mathbf{y}$, where $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a constant matrix.

Quadratic forms

Theorem

If $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \Sigma)$ then

$$\mathbf{y}'\mathbf{A}\mathbf{y} \sim \chi^2(r, \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}/2))$$

if and only if $\mathbf{A}\Sigma$ is idempotent of rank r

Properties

- The estimate of the variance is

$$s^2 = \frac{1}{n-p}(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}).$$

- We previously showed that this estimate was unbiased, i.e.

$$E[s^2] = \sigma^2.$$

- We can alternatively express s^2 as follows:

$$\frac{n-p}{\sigma^2}s^2 = \frac{1}{\sigma^2}\mathbf{y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}.$$

Properties

- Note this can be expressed as

$$\mathbf{y}'\mathbf{A}\mathbf{y}$$

  where

$$\mathbf{A} = \sigma^{-2}(\mathbf{I} - \mathbf{H}).$$

- Furthermore, note that

$$\mathbf{A}\Sigma = (\mathbf{I} - \mathbf{H})$$

  is idempotent with rank $n - p$.

Properties

- Also note that

$$\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} = \frac{1}{2}(\mathbf{X}\beta)'(\mathbf{I} - \mathbf{H})\mathbf{X}\beta = 0.$$

- Thus,

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

- The special case of this where **X** has only an intercept yields the usual empirical variance estimate.

Confidence interval for the variance

- We can use this result to develop a confidence interval for the variance.
- Let $\chi^2_{n-p,\alpha}$ be the α quantile from the chi squared distribution with $n-p$ degrees of freedom.
- Therefore

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

Confidence interval for the variance

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

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

Independence

Theorem

Let $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \Sigma)$, $\mathbf{A}$ be symmetric idempotent matrix, and $\mathbf{B}$ a matrix of constants, and suppose $\mathbf{B}\Sigma\mathbf{A} = \mathbf{0}$. Then $\mathbf{y}'\mathbf{A}\mathbf{y}$ and $\mathbf{B}\mathbf{y}$ are independent.

Independence

- Recall that

$$\frac{n-p}{\sigma^2}s^2 = \mathbf{y}'\mathbf{A}\mathbf{y}$$

where $\mathbf{A} = \sigma^{-2}(\mathbf{I} - \mathbf{H})$ and

$$\hat{\beta} = \mathbf{B}\mathbf{y}$$

where $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Independence

- Note

$$\begin{aligned}\mathbf{B}\Sigma\mathbf{A} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\sigma^{-2}(\mathbf{I}-\mathbf{H}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{H}) \\ &= 0\end{aligned}$$

- Thus, $\hat{\beta}$ and $(n-p)s^2/\sigma^2$ are independent, which implies that $\hat{\beta}$ and s^2 are independent.

Properties

- Recall that we showed that under the normality assumption, $\hat{\beta}$ and s^2 are sufficient statistics for β and σ^2.
- In addition, $\hat{\beta}$ and s^2 are complete statistics.

Properties

Theorem

Let $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, where $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Let $t(\beta, \sigma^2)$ be any function of the parameters β and σ^2 for which an unbiased estimator exists. Then there exists a function of the sufficient statistics $\hat{\beta}$ and s^2, say $q(\hat{\beta}, s^2)$, that is also an unbiased estimator of $t(\beta, \sigma^2)$. In addition, $q(\hat{\beta}, s^2)$ is the uniformly minimum variance unbiased (UMVU) estimator for $t(\beta, \sigma^2)$.

T statistics

- We are now in the position to develop inference for the β parameters.
- Consider the linear contrast $\mathbf{q}'\beta$.
- The uniformly minimum variance unbiased estimator of $\mathbf{q}'\beta$ is given by $\mathbf{q}'\hat{\beta}$.

T statistics

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

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

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

T statistics

- Therefore,

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

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

- Thus, we can write

$$\frac{\mathbf{q}'\hat{\beta} - \mathbf{q}'\beta}{\sqrt{\mathbf{q}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{q}s^2}} \sim t_{n-p}.$$

T statistics

- This result forms the basis of inference on the least-squares estimates of β.
- For example, choosing $q = (0, \dots, 0, 1, 0, \dots 0)$ allows us to perform inference on the i^{th} element of β.
- As another example, we can compare the first two elements of β using $q = (1, -1, 0, \dots 0)$.

F tests

- Now consider testing the hypothesis that

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

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

- Note that $\mathbf{K}\hat{\beta} \sim N(\mathbf{K}\beta, \mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}'\sigma^2)$ and thus

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

F tests

- Furthermore, $\mathbf{K}\hat{\beta}$ is independent of s^2.
- Thus,

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

forms the ratio of two independent χ^2 random variables over their degrees of freedom, which is an F distribution.

F tests

- Hence,

$$\frac{(\mathbf{K}\hat{\beta} - \mathbf{K}\beta)'\{\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}'\sigma^2\}^{-1}(\mathbf{K}\hat{\beta} - \mathbf{K}\beta)}{ps^2} \sim F_{p,n-p}.$$

- For example, we can use this result to test whether all elements of β are equal to 0, or alternatively whether both $\beta_i = \beta_j$ and $\beta_k = \beta_l$.

Coding example

Consider the `swiss` fertility dataset. Let's first make sure that we can replicate the coefficient table obtained by R.

```
> fit = lm(Fertility ~ ., data = swiss)
> round(summary(fit)$coef, 3)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        66.915     10.706   6.250    0.000
Agriculture        -0.172      0.070  -2.448    0.019
Examination        -0.258      0.254  -1.016    0.315
Education          -0.871      0.183  -4.758    0.000
Catholic            0.104      0.035   2.953    0.005
Infant.Mortality    1.077      0.382   2.822    0.007
```

Coding example

```
> # Now let's do it more manually
> x = cbind(1, as.matrix(swiss[,-1]))
> y = swiss$Fertility
> beta = solve(t(x) %*% x, t(x) %*% y)
> e = y - x %*% beta
> n = nrow(x); p = ncol(x)
> s = sqrt(sum(e^2) / (n - p))
> #Compare with lm
> c(s, summary(fit)$sigma)
[1] 7.165369 7.165369
```

Coding example

```
> ## Show that standard errors agree with lm
> betaVar = solve(t(x) %*% x) * s ^ 2
> cbind(summary(fit)$coef[,2], sqrt(diag(betaVar)))
                        [,1]        [,2]
(Intercept)      10.70603759 10.70603759
Agriculture       0.07030392  0.07030392
Examination       0.25387820  0.25387820
Education         0.18302860  0.18302860
Catholic          0.03525785  0.03525785
Infant.Mortality  0.38171965  0.38171965
```

Coding example

```
> # Show that the tstats agree
> tstat = beta / sqrt(diag(betaVar))
> cbind(summary(fit)$coef[,3], tstat)
                     [,1]      [,2]
(Intercept)       6.250229  6.250229
Agriculture      -2.448142 -2.448142
Examination      -1.016268 -1.016268
Education        -4.758492 -4.758492
Catholic          2.952969  2.952969
Infant.Mortality  2.821568  2.821568
```

Coding example

```
> # Show that the P-values agree
> cbind(summary(fit)$coef[,4],  2 *pt(-abs(tstat), n-p)
                              [,1]         [,2]
(Intercept)      1.906051e-07 1.906051e-07
Agriculture      1.872715e-02 1.872715e-02
Examination      3.154617e-01 3.154617e-01
Education        2.430605e-05 2.430605e-05
Catholic         5.190079e-03 5.190079e-03
Infant.Mortality 7.335715e-03 7.335715e-03
```

Coding example

```
> # Get the F statistic
> # Set K to grab everything except the intercept
> k = cbind(0, diag(rep(1, p - 1)))
> k
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    1    0    0    0    0
[2,]    0    0    1    0    0    0
[3,]    0    0    0    1    0    0
[4,]    0    0    0    0    1    0
[5,]    0    0    0    0    0    1
```

Coding example

```
> kvar = k %*% solve(t(x) %*% x) %*% t(k)
> fstat = t(k %*% beta) %*% solve(kvar) %*% (k %*% beta)
> #Show that it's equal to what lm is giving
> cbind(summary(fit)$fstat, fstat)
> #Calculate the p-value
> pf(fstat, p - 1, n - p, lower.tail = FALSE)
              [,1]
[1,] 5.593799e-10
> summary(fit)
## ... only showing the one relevant line ...
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```