Linear Regression Models
P8111

Lecture 10

Jeff Goldsmith
February 23, 2016

THE DEPARTMENT OF BIOSTATISTICS
Columbia University
MAILMAN SCHOOL
OF PUBLIC HEALTH

Today's Lecture

- Review of tests
- Two new tests
- Confidence intervals
- Foreshadowing

Some review notes

- Do we have a test for the null $H_0 : \beta_3 = -14$
- Do we have a test for the null $H_0 : \beta_2 + \beta_3 = \pi$
- Do we have a test for the null $H_0 : \beta_2 \beta_3 = \pi$

Individual coefficients

For individual coefficients

- We can use the test statistic

$$T = \frac{\hat{\beta}_j - \beta_j}{\widehat{se}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}_{jj}}} \sim t_{n-p-1}$$

- For a two-sided test of size α, we reject if

$$|T| > t_{1-\alpha/2, n-p-1}$$

- The p-value gives $P(|t_{n-p-1}| > |T_{obs}| | H_0)$

Note that t is a symmetric distribution that converges to a Normal as $n - p - 1$ increases.

Inference for linear combinations

Sometimes we are interested in making claims about $c^T\boldsymbol{\beta}$ for some c.

- Define $H_0 : c^T\boldsymbol{\beta} = c^T\boldsymbol{\beta}_0$ or $H_0 : c^T\boldsymbol{\beta} = 0$
- We can use the test statistic

$$T = \frac{c^T\hat{\boldsymbol{\beta}} - c^T\boldsymbol{\beta}}{\widehat{se}(c^T\hat{\boldsymbol{\beta}})} = \frac{c^T\hat{\boldsymbol{\beta}} - c^T\boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2 c^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}c}}$$

- This test statistic is asymptotically Normally distributed
- For a two-sided test of size α, we reject if

$$|T| > z_{1-\alpha/2}$$

Global F tests

- Compute the test statistic

$$F_{obs} = \frac{(RSS_S - RSS_L)/(df_S - df_L)}{RSS_L/df_L}$$

- If H_0 (the null model) is true, then $F_{obs} \sim F_{df_S - df_L, df_L}$
- Note $df_s = n - p_S - 1$ and $df_L = n - p_L - 1$
- We reject the null hypothesis if the p-value is above α, where

$$\text{p-value} = P(F_{df_S - df_L, df_L} > F_{obs})$$

Alternative global tests: the Wald test

For a vector of coefficients, we can test $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$:

- Use the test statistic

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T [Var(\hat{\boldsymbol{\beta}})]^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

- Under the null, this test statistic has an asymptotic χ^2_p distribution
- In practice, we replace $Var(\hat{\boldsymbol{\beta}})$ with $\widehat{Var}(\hat{\boldsymbol{\beta}})$ and use an F distribution

Alternative global tests: the Wald test

The previous test is special case of $H_0 : R\boldsymbol{\beta} = R\boldsymbol{\beta}_0$ for a $d \times p$ matrix R, using $R = I_{p \times p}$:

- Use the test statistic

$$W = (R\hat{\boldsymbol{\beta}} - R\boldsymbol{\beta}_0)^T [Var(R\hat{\boldsymbol{\beta}})]^{-1} (R\hat{\boldsymbol{\beta}} - R\boldsymbol{\beta}_0)$$

- Under the null, this test statistic has an asymptotic χ^2_d distribution, where

$$d = \text{rank}(Var(R\hat{\boldsymbol{\beta}}))$$

- This formulation is useful for testing subsets (e.g. $H_0 : \beta_1 = \beta_2 = 0$)

Alternative global tests: the likelihood ratio test

If we are using maximum likelihood estimation (we'll cover this soon – turns out to be least squares in MLR), we can use a LRT:

- Use the test statistics

$$\Delta = -2 \log \frac{L_0}{L_1} = -2(l_0 - l_1)$$

- This test statistic has an asymptotic χ^2_d distribution where d is the difference in the number of parameters between the two models.
- Must compare nest models

Example: LRT

```
> linmod = lm(runs ~ at_bats + hits + homeruns + stolen_bases, data = mlb11)
> linmod.null1 = lm(runs ~ hits + homeruns, data = mlb11)
> anova(linmod.null1, linmod)
Model 1: runs ~ hits + homeruns
Model 2: runs ~ at_bats + hits + homeruns + stolen_bases
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1     27 27128
2     25 18020  2    9107.8 6.3178 0.006015 **
>
> ## LRT
> Delta = -2*(logLik(linmod.null1) - logLik(linmod))
> 1-pchisq(Delta, 2)
'log Lik.' 0.002163305 (df=4)
```

Confidence intervals: individual parameters

- A confidence interval with coverage $(1-\alpha)$ is given by

$$\hat{\beta}_j \pm t_{1-\alpha/2,n-p-1}\widehat{se}(\hat{\beta}_j)$$

- Assuming all the standard assumptions hold,

$$(1-\alpha) \text{ "=" } P(LB < \beta_j < UB)$$

Note there is a one-to-one correspondence between this confidence interval and the hypothesis test.

Example revisited

```
> confint(linmod)
                   2.5 %       97.5 %
(Intercept)  -502.9429878 1.665365e+03
at_bats        -0.4440385 3.938287e-02
hits            0.4643923 9.304364e-01
homeruns        0.9253836 1.581629e+00
stolen_bases    0.1756711 8.702772e-01
```

Confidence intervals: multiple parameters

We might want a confidence region for multiple coefficients simultaneously

- Invert Wald test for multiple coefficients – find region containing all values $\boldsymbol{\beta}_0$ for which p-value from global Wald test is $> \alpha$
- Then

$$(1 - \alpha) \text{"="} P[\boldsymbol{\beta} \in \text{region}]$$

- This region is an ellipsoid in higher dimensions; we can visualize in 2D most easily and 3D pretty well.

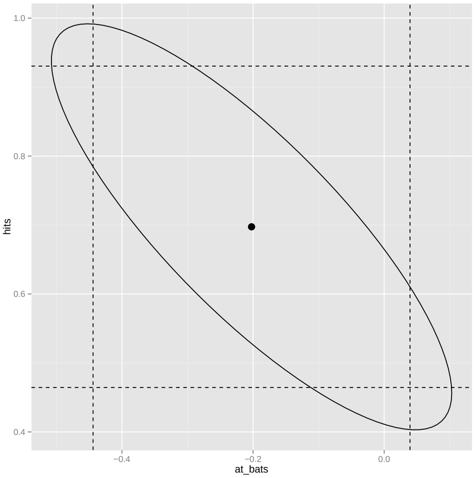
Confidence intervals: multiple parameters

```
library(ellipse)

CI.ellipse = as.data.frame(ellipse(linmod,c(2,3)))
est = as.data.frame(t(as.matrix(coef(linmod)[2:3])))

## plot the joint confidence region
ggplot(CI.ellipse, aes(x = at_bats, y = hits)) + geom_path() +
  geom_hline(yintercept = confint(linmod)[3,], linetype = 2) +
  geom_vline(xintercept = confint(linmod)[2,], linetype = 2) +
  geom_point(data = est, size = 4)
```

Confidence intervals: multiple parameters



Expected value and it's variance

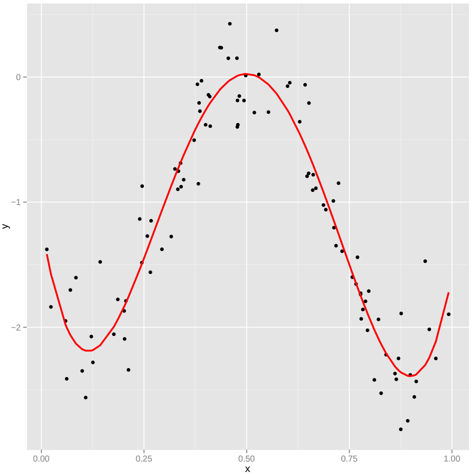
- What is $\hat{E}(y|\boldsymbol{x} = \boldsymbol{x}_0)$?
- How can we estimate the variance of $\hat{E}(y|\boldsymbol{x} = \boldsymbol{x}_0)$?

In particular, a confidence interval for $E(y|\boldsymbol{x} = \boldsymbol{x}_0)$ is given by

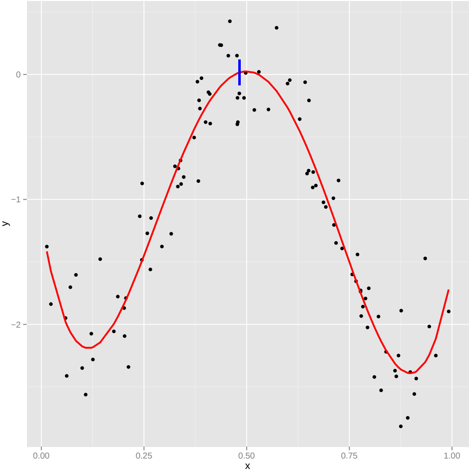
$$(\hat{y}|\boldsymbol{x} = \boldsymbol{x}_0) \pm t_{1-\alpha/2,n-p-1}\widehat{se}_{fit}(\hat{y}|\boldsymbol{x}_0)$$

This can be estimated for any $\boldsymbol{x}_0$.

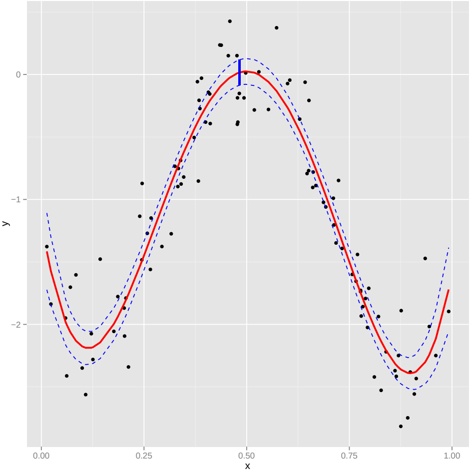
Estimated mean



Estimated mean and variance



Estimated mean and variance



Pointwise and simultaneous CIs

- Pointwise confidence intervals construct CI's at each point independently of all other points
- Implicit multiple comparisons problem
- Simultaneous intervals can be constructed so that

$$(1-\alpha) \text{ "=" } P(f(x) \in SCI)$$

Which is wider?

Predictions and prediction intervals

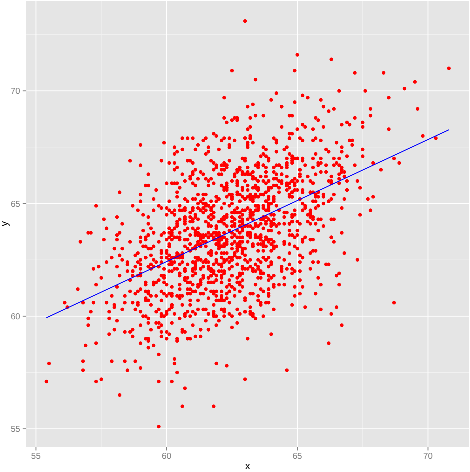
- What is the prediction value y for a given $\boldsymbol{x}_0$
- What range would you give for the value of a new outcome?
- Two sources of variance to consider: variance in estimates and variance in outcome

A prediction interval is given by

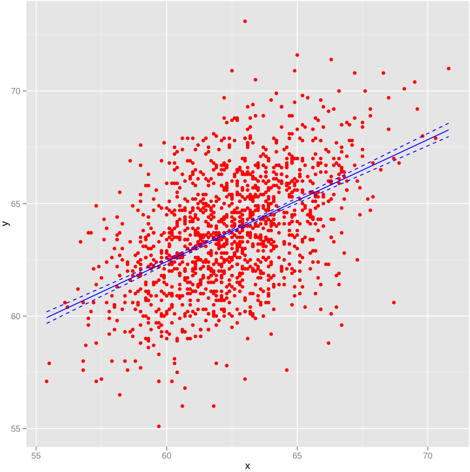
$$(\hat{y}|\boldsymbol{x} = \boldsymbol{x}_0) \pm t_{1-\alpha/2, n-p-1} \widehat{se}_{pred}(\hat{y}|\boldsymbol{x}_0)$$

This can be estimated for any $\boldsymbol{x}_0$.

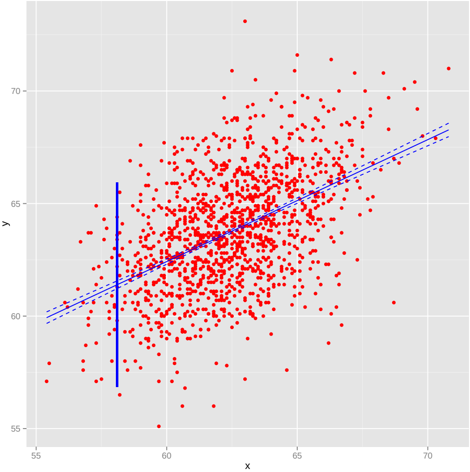
Predictions - mother/daughter height



Variance of fitted values



Prediction interval



Some things to think on

- Why are we building models?
- How should we assess models?
- What kinds of predictors should we included, and how should we decide to include them?

Some things to think on

Three general goals are

- Prediction
- Estimation of association
- Testing of associations

These goals will often not lead to the same final model.

Today's big ideas

- Inference for MLRs: global tests; confidence intervals for coefficients, predictions, functions

---

- Suggested reading: Faraway Ch 3.6 - 3.9