Multiple Linear Regression (cont.)

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

February 13, 2007

Hypothesis Testing

- The summary of a linear model and the ANOVA table of a linear model summarize different hypothesis tests.
- Regression coefficients must be understood in the context of which other variables are included in the model.
- The R command `summary` shows results of t-tests that test if a parameter is zero in a model that includes all other coefficients.
- The R command `anova` shows the results of F-tests that test if a parameter is zero in a model including only parameters listed earlier in the table.
- However, tests in the ANOVA table use the full model to estimate error for all tests.

R Commands to Fit Models

```
> toxic = read.table("toxic.txt", header = T)
> str(toxic)
> attach(toxic)
> toxic0.lm = lm(effect ~ 1)
> toxic1.lm = lm(effect ~ dose)
> toxic2.lm = lm(effect ~ weight)
> toxic12.lm = lm(effect ~ dose + weight)
> toxic21.lm = lm(effect ~ weight + dose)
```


Differences in Tests

- We need to distinguish between

$$H_0 : [\beta_1 = 0 \,|\, \beta_0]$$

(i.e., $\beta_1 = 0$ given that β_0 is in the model), and

$$H_0 : [\beta_1 = 0 \,|\, \beta_0, \beta_2]$$

(i.e., $\beta_1 = 0$ given that β_0, β_2 are in the model).

Pesticide Example (cont.)

```
> summary(toxic12.lm)

Call:
lm(formula = effect ~ dose + weight)

Residuals:
      Min        1Q    Median        3Q       Max
-0.081512 -0.023945 -0.003421  0.022278  0.094310

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.22281    0.08364   2.664  0.01698 *
dose         0.65139    0.17305   3.764  0.00170 **
weight      -1.13321    0.18044  -6.280 1.10e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0466 on 16 degrees of freedom
Multiple R-Squared: 0.7796, Adjusted R-squared: 0.752
F-statistic:  28.3 on 2 and 16 DF,  p-value: 5.57e-06

> anova(toxic12.lm)

Analysis of Variance Table

Response: effect
          Df   Sum Sq  Mean Sq F value    Pr(>F)
dose       1 0.037239 0.037239  17.152 0.0007669 ***
weight     1 0.085629 0.085629  39.440 1.097e-05 ***
Residuals 16 0.034738 0.002171
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null	p-value
$H_0: [\beta_0 = 0 \mid \beta_1, \beta_2]$	0.01698
$H_0: [\beta_1 = 0 \mid \beta_0, \beta_2]$	0.001697
$H_0: [\beta_2 = 0 \mid \beta_1, \beta_2]$	1.097e-05
$H_0: [\beta_1 = \beta_2 = 0 \mid \beta_0]$	5.57×10^{-6}
$H_0: [\beta_1 = 0 \mid \beta_0]$	0.0007669
$H_0: [\beta_2 = 0 \mid \beta_0, \beta_1]$	1.097e-05

Coefficient Summary

Model	Intercept	Dose	Weight
`toxic0.lm`	0.3714		
`toxic1.lm`	0.6049	-0.3206	
`toxic2.lm`	0.5226		-0.5258
`toxic12.lm`	0.2228	0.6514	-1.133
`toxic21.lm`	0.2228	0.6514	-1.133

- Coefficient interpretation depends on other variables present in the model!

Regression Through the Origin

- The usual model is:
  $y_i = \beta_0 + \beta_1 x_i + e_i$, where $e_i \sim \text{iid N}(0, \sigma^2)$
- Suppose we accept $H_0\colon \beta_0 = 0$. Then the model reduces to:
  $y_i = \beta_1 x_i + e_i$, where $e_i \sim \text{iid N}(0, \sigma^2)$.
- Model parameters are β_1, σ^2.
- The equations for least squares regression change.
- The least squares criterion becomes to minimize $\sum_{i=1}^n (y_i - (b_1 x_i))^2$.
- The solution to this problem is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

- The estimated variance is $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{\beta}_1 x_i)^2}{n-1}$

- Inference using regression through the origin has one more degree of freedom for error as one fewer parameter is needed for the mean.

Source	df	SS	MS
Regression	1	$\frac{(\sum_{i=1}^n x_i y_i)^2}{\sum_{i=1}^n x_i^2}$	$\frac{(\sum_{i=1}^n x_i y_i)^2}{\sum_{i=1}^n x_i^2}$
Error	$n-1$	By subtraction	SSError$/(n-1)$
Total	n	$\sum_{i=1}^n y_i^2$	–

Example

```
> x = c(40, 43, 49, 51, 52, 52, 54, 55, 60)
> y = c(44, 34, 46, 53, 55, 56, 51, 55, 60)
> origin = data.frame(x = x, y = y)
> origin.lm = lm(y ~ x, data = origin)
> summary(origin.lm)$coefficients

             Estimate Std. Error   t value    Pr(>|t|)
(Intercept) -5.585586 13.2655664 -0.421059 0.686338805
x            1.105856  0.2601592  4.250690 0.003789901

> anova(origin.lm)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
x          1 361.98  361.98  18.068 0.00379 **
Residuals  7 140.24   20.03
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Fit Without an Intercept

```
> origin2.lm = lm(y ~ x - 1, data = origin)
> summary(origin2.lm)$coefficients

   Estimate Std. Error  t value     Pr(>|t|)
x 0.9970085 0.02771484 35.97382 3.905618e-10

> anova(origin2.lm)

Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 23260.2 23260.2  1294.1 3.906e-10 ***
Residuals  8   143.8    18.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> par(mar = c(4, 4, 0, 0), las = 1, pch = 16)
> plot(x, y, xlim = c(0, 60))
> abline(origin.lm)
> abline(origin2.lm, col = "red", lty = 2)
```