Some Notes on Linear Models in R

Dan Nettleton

Iowa State University

April 2, 2009

Single Factor ANOVA Model: $y_{ij} = \mu + \tau_i + e_{ij}$

Treatment	i	j	y_{ij}	Mean
1	1	1	9	$\mu + \tau_1$
1	1	2	7	$\mu + \tau_1$
2	2	1	0	$\mu + \tau_2$
2	2	2	4	$\mu + \tau_2$
3	3	1	4	$\mu + \tau_3$
3	3	2	6	$\mu + \tau_3$
4	4	1	0	$\mu + \tau_4$
4	4	2	2	$\mu + \tau_4$

Least-Squares Estimation of Mean Parameters

y_{ij}	Mean	Error^2
9	$\mu + \tau_1$	$\{9 - (\hat{\mu} + \hat{\tau}_1)\}^2$
7	$\mu + \tau_1$	$\{7 - (\hat{\mu} + \hat{\tau}_1)\}^2$
0	$\mu + \tau_2$	$\{0 - (\hat{\mu} + \hat{\tau}_2)\}^2$
4	$\mu + \tau_2$	$\{4 - (\hat{\mu} + \hat{\tau}_2)\}^2$
4	$\mu + \tau_3$	$\{4 - (\hat{\mu} + \hat{\tau}_3)\}^2$
6	$\mu + \tau_3$	$\{6 - (\hat{\mu} + \hat{\tau}_3)\}^2$
0	$\mu + \tau_4$	$\{0 - (\hat{\mu} + \hat{\tau}_4)\}^2$
2	$\mu + \tau_4$	$\{2 - (\hat{\mu} + \hat{\tau}_4)\}^2$

Solutions to the least squares equations are values $\hat{\mu}$, $\hat{\tau}_1$, $\hat{\tau}_2$, $\hat{\tau}_3$, and $\hat{\tau}_4$ that make the sum of the squared errors as small as possible.

Least-Squares Estimation of Mean Parameters

y_{ij}	Mean	Error^2
9	$\mu + \tau_1$	$\{9 - (\hat{\mu} + \hat{\tau}_1)\}^2$
7	$\mu + \tau_1$	$\{7 - (\hat{\mu} + \hat{\tau}_1)\}^2$
0	$\mu + \tau_2$	$\{0 - (\hat{\mu} + \hat{\tau}_2)\}^2$
4	$\mu + \tau_2$	$\{4 - (\hat{\mu} + \hat{\tau}_2)\}^2$
4	$\mu + \tau_3$	$\{4 - (\hat{\mu} + \hat{\tau}_3)\}^2$
6	$\mu + \tau_3$	$\{6 - (\hat{\mu} + \hat{\tau}_3)\}^2$
0	$\mu + \tau_4$	$\{0 - (\hat{\mu} + \hat{\tau}_4)\}^2$
2	$\mu + \tau_4$	$\{2 - (\hat{\mu} + \hat{\tau}_4)\}^2$

Note that the solutions are not unique because the sum of the squared errors depends on $\hat{\mu}$, $\hat{\tau}_1$, $\hat{\tau}_2$, $\hat{\tau}_3$, and $\hat{\tau}_4$ only through $\hat{\mu} + \hat{\tau}_1$, $\hat{\mu} + \hat{\tau}_2$, $\hat{\mu} + \hat{\tau}_3$, and $\hat{\mu} + \hat{\tau}_4$.

Least-Squares Estimation of Mean Parameters

y_{ij}	Mean	Error^2
9	$\mu + \tau_1$	$\{9 - (\hat{\mu} + \hat{\tau}_1)\}^2$
7	$\mu + \tau_1$	$\{7 - (\hat{\mu} + \hat{\tau}_1)\}^2$
0	$\mu + \tau_2$	$\{0 - (\hat{\mu} + \hat{\tau}_2)\}^2$
4	$\mu + \tau_2$	$\{4 - (\hat{\mu} + \hat{\tau}_2)\}^2$
4	$\mu + \tau_3$	$\{4 - (\hat{\mu} + \hat{\tau}_3)\}^2$
6	$\mu + \tau_3$	$\{6 - (\hat{\mu} + \hat{\tau}_3)\}^2$
0	$\mu + \tau_4$	$\{0 - (\hat{\mu} + \hat{\tau}_4)\}^2$
2	$\mu + \tau_4$	$\{2 - (\hat{\mu} + \hat{\tau}_4)\}^2$

In this case, any values satisfying $\hat{\mu} + \hat{\tau}_1 = \frac{9+7}{2} = 8$ $\hat{\mu} + \hat{\tau}_2 = \frac{0+4}{2} = 2$ $\hat{\mu} + \hat{\tau}_3 = \frac{4+6}{2} = 5$ and $\hat{\mu} + \hat{\tau}_4 = \frac{0+2}{2} = 1$ will minimize the sum of squared errors.

Least-Squares Estimation of Mean Parameters

- By default, R sets the value corresponding to the first level of each factor to be 0 and reports only values for $\hat{\mu}$ and levels other than the first.
- In this example, there is one factor (treatment).
- The levels of the factor treatment correspond to τ_1,τ_2,τ_3, and τ_4.
- Thus, $\hat{\tau}_1$ will be set to 0 by R.
- The values of $\hat{\mu}$, $\hat{\tau}_2$, $\hat{\tau}_3$, and $\hat{\tau}_4$ will be chosen so that $\hat{\mu} = 8$, $\hat{\mu} + \hat{\tau}_2 = 2$, $\hat{\mu} + \hat{\tau}_3 = 5$, and $\hat{\mu} + \hat{\tau}_4 = 1$.

R Code and Output

```
> trt=gl(4,2)

> trt
[1] 1 1 2 2 3 3 4 4
Levels: 1 2 3 4

> y=c(9,7,0,4,4,6,0,2)

> o=lm(y~trt)

> coef(o)
(Intercept)        trt2        trt3        trt4
          8          -6          -3          -7
```

R Code and Output (continued)

```
> anova(o)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
trt        3   60.0    20.0  5.7143 0.06272 .
Residuals  4   14.0     3.5
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1
```

The `trt` line the the ANOVA table corresponds to the F test of

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4.$$

Other Tests

- Let θ denote the vector of mean parameters whose estimates are not set equal to 0. In our case,

$$\theta = [\mu, \tau_2, \tau_3, \tau_4]'.$$

- Let $\hat{\theta}$ denote the vector of mean parameter estimates that are not set equal to 0. In our case,

$$\begin{aligned} \hat{\theta} &= [\hat{\mu}, \hat{\tau}_2, \hat{\tau}_3, \hat{\tau}_4]'. \\ &= [8, -6, -3, -7]'. \end{aligned}$$

- Let V denote the estimated variance matrix for $\hat{\theta}$. The actual numerical values for this matrix can be obtained in R as `vcov(o)`.

Other Tests (continued)

- Suppose we wish to test the hypothesis $H_0 : m'\theta = 0$.
- The test statistic is $t = \frac{m'\hat{\theta}}{\sqrt{m'Vm}}$.
- Under H_0, the distribution of this statistic is t with degrees of freedom equal to the residual degrees of freedom from the ANOVA table.
- Suppose we wish to test for a difference between the means of treatments 2 and 3. What should our m vector be?

R Code and Output (continued)

```
> m=c(0,1,-1,0)
> th=coef(o)
> V=vcov(o)

> tt=t(m)%*%th/sqrt(t(m)%*%V%*%m)
> tt
           [,1]
[1,] -1.603567

> rdf=anova(o)[2,1]
> rdf
[1] 4

> pval=2*(1-pt(abs(tt),rdf))
> pval
           [,1]
[1,] 0.1840740
```

R Code and Output (continued)

```
> get.pvals=function(y)
{
  o=lm(y~trt)
  th=coef(o)
  V=vcov(o)
  a=anova(o)
  rdf=anova(o)[2,1]
  p1=a[1,5]
  m=c(0,1,-1,0)
  tt=t(m)%*%th/sqrt(t(m)%*%V%*%m)
  p2=2*(1-pt(abs(tt),rdf))
  m=c(0,1,0,0)
  tt=t(m)%*%th/sqrt(t(m)%*%V%*%m)
  p3=2*(1-pt(abs(tt),rdf))
  p=c(p1,p2,p3)
  p
}
```

R Code and Output (continued)

- Suppose `d` is a data frame with one row for each gene.
- Suppose the first column of `d` contains the gene ID.
- Suppose that there are 8 other columns containing the data in the same order as the `y` vector.
- Then the code below will compute three p-values for each gene and store the result in a matrix with one row for each gene and one column for each test.

```
results=t(apply(d[,-1],1,get.pvals))
```