EEB 581, Problem Set Eight

Due 23 March 2006

Consider the following data predicting variable y as a function of variables x and sex,

y	x	Sex
236	10	f
266	12	f
301	13	f
306	14	f
230	10	f
97	12	m
84	13	m
120	20	m
82	11	m
118	19	m

Our biological hypothesis is that a line (a linear regression) may best explain the data, but that different lines may be needed for males and females. Let z be the indicator variable for the sex of an observation, with

$$z_i = \begin{cases} 0 & \text{if individual i is male} \\ 1 & \text{if individual i is female} \end{cases}$$

There are four models for us to consider, from basic to the most general.

(i) a simple regression, the same in both sexes

$$y = a + bX + e$$

(ii) a regression with the same slope for both sexes, but different intercepts

$$y = a + bx + cz + e$$

$$(\text{for males} : y = a + bx + e, \qquad \text{for females} : y = (a + c) + bx + e)$$

(iii) a regression with the same intercept for both sexes, but different slopes

$$y = a + bx + d * x * z$$

$$(\text{for males} : y = a + bx + e, \qquad \text{for females} : y = a + (b + d)x + e)$$

(iv) A regression with different slopes and intercepts for the sexes

$$y = a + bx + c * z + d * x * z$$

$$(\text{for males} : y = a + bx + e, \qquad \text{for females} : y = (a + c) + (b + d)x + e)$$

1 : Write each of these models in matrix form.

2 : For each model compute the OLS estimates of the model parameters, the error sum of squares, the estimated residual variance, and the estimated variance-covariance estimate of the parameters.

3 : Given that models (i) – (iii) are subsets of the full model (iv), compute the F statistics, and their resulting p values, that the reduced model has the same fit as the full model.

4 : Using your estimated variances from (2), perform t-tests for all of the parameters for each of the four models (note the df for each test is $N - p$). Note that these are two-sides tests, with the null hypothesis of a zero value, so that p is the probability of getting a value at least that extreme (positive or negative).

$$p = \Pr(t_{N-p} > \frac{|\,\widehat{\beta}_i\,|}{\sqrt{\sigma^2(\widehat{\beta}_i)}}) + \Pr(t_{N-p} < \frac{-|\,\widehat{\beta}_i\,|}{\sqrt{\sigma^2(\widehat{\beta}_i)}}) = 2 \cdot \Pr(t_{N-p} < \frac{-|\,\widehat{\beta}_i\,|}{\sqrt{\sigma^2(\widehat{\beta}_i)}})$$

Be sure and see the Helpful Hints on the next page!

Helpful hints!

(1) Be sure and use **R** and the matrix forms of the GLM results (summarized in the handout "Summary of GLM results" on the course webpage)

(2) You will no doubt notice that only the design matrix $\mathbf{X}$ changes under the models. Further, it does so by the simple addition or subtraction of whole columns (depending on the model). Rather than type in a whole new $\mathbf{X}$ for each model, you can use a cute trick in **R**, based on the fact that it builds matrices one column at a time. Suppose you wish to enter the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 4 \\ 3 & 2 & 3 \\ 5 & 4 & 4 \\ 8 & 9 & 4 \end{pmatrix}$$

in **R**, this is entered as

```
A <- matrix(c(1,3,5,8,2,2,4,9,4,3,4,4), nrow=4)
```

However, consider

$$\mathbf{A} = (\,\mathbf{a} \quad \mathbf{b} \quad \mathbf{c}\,), \quad \text{where} \quad \mathbf{a} = \begin{pmatrix} 1 \\ 3 \\ 5 \\ 8 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 2 \\ 2 \\ 4 \\ 9 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} 4 \\ 3 \\ 4 \\ 4 \end{pmatrix}$$

Entering these three column vectors in **R**

```
a <- matrix(c(1,3,5,8), nrow=4)
b <- matrix(c(2,2,4,9), nrow=4)
c <- matrix(c(4,3,4,4), nrow=4)
```

We can recover $\mathbf{A}$ by combining these column vectors with the **cbind** commannd

```
A <-cbind(a,b,c)
```

Much more interesting, if we wish the matrices

$$\mathbf{D} = \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 5 & 4 \\ 8 & 9 \end{pmatrix}, \quad \mathbf{E} = \begin{pmatrix} 2 & 4 \\ 2 & 3 \\ 4 & 4 \\ 9 & 4 \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} 1 & 2 & 4 \\ 1 & 2 & 3 \\ 1 & 4 & 4 \\ 1 & 9 & 4 \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} 1 \cdot 2 & 2 & 4 \\ 3 \cdot 2 & 2 & 3 \\ 5 \cdot 4 & 4 & 4 \\ 8 \cdot 9 & 9 & 4 \end{pmatrix}$$

we can get these in **R** by simply using

```
D <- cbind(a,b)
E <-cbind(b,c)
F <-cbind(1,b,c)
G <-cbind(a*b,b,c)
```

Rather than trying to retype in everything. Note that using a **1** in **cbind** returns a row of ones! Also note that $a * b$ return a column whose i-th row is just $a_i \cdot b_i$.