Chapter Six

Ryan Honea

10/3/2017

Exercise Four

Question

It is well known that the concentration of cholesterol in blood serum increase with age, but it is less clear whether cholesterol level is also associated with body weight. The table below shows for thirty women serum cholesterol, (millimoles per liter), age (years) and body mass index (weight divided by height squared, where weight was measured in kilograms and height in meters). Use multiple regression to test whether serum cholesterol is associated with body mass index when age is already included in the model.

CHOL	Age	BMI	CHOL	Age	BMI
5.94	52	20.7	6.48	65	26.3
4.71	46	21.3	8.83	76	22.7
5.86	51	25.4	5.1	47	21.5
6.52	44	22.7	5.81	43	20.7
6.8	70	23.9	4.65	30	18.9
5.23	33	24.3	6.82	58	23.9
4.97	21	22.2	6.28	78	24.3
8.78	63	26.2	5.15	49	23.8
5.13	56	23.3	2.92	36	19.6
6.74	54	29.2	9.27	67	24.3
5.95	44	22.7	5.57	42	22
5.83	71	21.9	4.92	29	22.5
5.74	39	22.4	6.72	33	24.1
4.92	58	20.2	5.57	42	22.7
6.69	58	24.4	6.25	66	27.3

Solution

We begin by entering our data into collections.

```r
chol <- c(5.94, 4.71, 5.86, 6.52, 6.80, 5.23, 4.97, 8.78, 5.13, 6.74,
          5.95, 5.83, 5.74, 4.92, 6.69, 6.48, 8.83, 5.10, 5.81, 4.65,
          6.82, 6.28, 5.15, 2.92, 9.27, 5.57, 4.92, 6.72, 5.57, 6.25)
bmi  <- c(20.7, 21.3, 25.4, 22.7, 23.9, 24.3, 22.2, 26.2, 23.3, 29.2,
          22.7, 21.9, 22.4, 20.2, 24.4, 26.3, 22.7, 21.5, 20.7, 18.9,
          23.9, 24.3, 23.8, 19.6, 24.3, 22.0, 22.5, 24.1, 22.7, 27.3)
age  <- c(52, 46, 51, 44, 70, 33, 21, 63, 56, 54, 44, 71, 39, 58, 58,
          65, 76, 47, 43, 30, 58, 78, 49, 36, 67, 42, 29, 33, 42, 66)
```

This is followed by ensuring that these are the matrix class for sake of transpose and inversions.

```r
y <- as.matrix(chol)
X_0 <- cbind(rep(1, length(chol)),
           as.matrix(age))
X_1 <- cbind(rep(1, length(chol)),
           as.matrix(age),
           as.matrix(bmi))
```

Now, we can use the `lm` command to find the coefficients.

```
lmod_0 <- lm(chol ~ age)
lmod_1 <- lm(chol ~ age + bmi)
```

With this information, we can now calculate Deviances and therefore the F-statistic.

```
D_0 = t(y)%*%y - t(lmod_0$coefficients)%*%t(X_0)%*%y
D_1 = t(y)%*%y - t(lmod_1$coefficients)%*%t(X_1)%*%y
p = 3
q = 2
N = length(y)
F.stat = ((D_0 - D_1)/(p-q)) / (D_1/(N-p))
cat("F-Statistic is",F.stat,"\n")
```

```
## F-Statistic is 5.147385
```

We also want to know the F-statistic at 95% in order to have a rejection region for our F-statistic.

```
cat("Rejection Region is F > ", qf(.95,p-q,N-p), "\n")
```

```
## Rejection Region is F >  4.210008
```

As our F-statistic is within our rejection region, we reject the null model and conclude that BMI's inclusion makes a stronger model.

Exercise Five

Question

The table below shows plasma inorganic phospate levels (mg/dl) one hour after a standard glucose tolerance test for obese subjects, with or without hyperinsulinemia, and controls (data from Jones 1987).

Hyperinsulinemic	Non-hyperinsulinemic	Controls
obese	obese	
2.3	3	3
4.1	4.1	2.6
4.2	3.9	3.1
4	3.1	2.2
4.6	3.3	2.1
4.6	2.9	2.4
3.8	3.3	2.8
5.2	3.9	3.4
3.1		2.9
3.7		2.6
3.8		3.1
		3.2

Solution

(a): Perform a one-factor analysis of variance to test the hypothesis that there are no mean differences among the three groups. What conclusions can you draw?

Solution: We begin by entering the data.

```r
hi.obese  <- c(2.3, 4.1, 4.2, 4.0, 4.6, 4.6, 3.8, 5.2, 3.1, 3.7, 3.8)
nhi.obese <- c(3.0, 4.1, 3.9, 3.1, 3.3, 2.9, 3.3, 3.9)
controls  <- c(3.0, 2.6, 3.1, 2.2, 2.1, 2.4, 2.8, 3.4, 2.9, 2.6, 3.1, 3.2)
```

Now, we find the squared-sums and the fits.

```r
hi.obese.ssq <- sum(hi.obese^2)
hi.obese.fit <- rep(mean(hi.obese),length(hi.obese))
nhi.obese.ssq <- sum(nhi.obese^2)
nhi.obese.fit <- rep(mean(nhi.obese),length(nhi.obese))
controls.ssq <- sum(controls^2)
controls.fit <- rep(mean(controls),length(controls))
```

With this information, we can begin to perform the ANOVA. We will need to find the Sum of Squares Within Groups, Between Groups, and in the Residuals. Truly, we only need to find two of these because the third can be found by subtracting two SS values from the total.

```r
grand.mean = mean(c(hi.obese, nhi.obese, controls))
hi.obese.sqres <- (hi.obese - hi.obese.fit)^2
nhi.obese.sqres <- (nhi.obese - nhi.obese.fit)^2
controls.sqres <- (controls - controls.fit)^2
TOT <- hi.obese.ssq + nhi.obese.ssq + controls.ssq
SSR <- sum(c(hi.obese.sqres,nhi.obese.sqres,controls.sqres))
SSB <-  length(hi.obese)*(mean(hi.obese) - grand.mean)^2 +
        length(nhi.obese)*(mean(nhi.obese) - grand.mean)^2 +
        length(controls)*(mean(controls) - grand.mean)^2
SSW <- TOT - SSR - SSB
cat("Total Sum of Squares:", TOT, "\n",
    "Sum Squares Within:", SSW, "\n",
    "Sum Squares Between:",SSB, "\n",
    "Sum Squares Residuals:", SSR, "\n")
```

```
## Total Sum of Squares: 368.11
##  Sum Squares Within: 350.919
##  Sum Squares Between: 7.808278
##  Sum Squares Residuals: 9.382689
```

We can now construct an ANOVA table to find our F-value.

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Mean (SSW)	1	350.919		
Between treatment (SSB)	2	7.808	3.904	11.653
Residual (SSR)	28	9.382	0.335	
Total	31	368.11		

Now, we check the rejection region.

```r
cat("Rejection Region: F > ",qf(.95,2,28))
```

```
## Rejection Region: F >  3.340386
```

11.653 is within the rejection region, and so we reject the null hypothesis that the means are equal and conclude that there is evidence to suggest they are different.

(b): Obtain a 95% confidence interval for the difference in means between the two obese groups.

Solution: We can do this using t-statistics and pooled standard deviations.

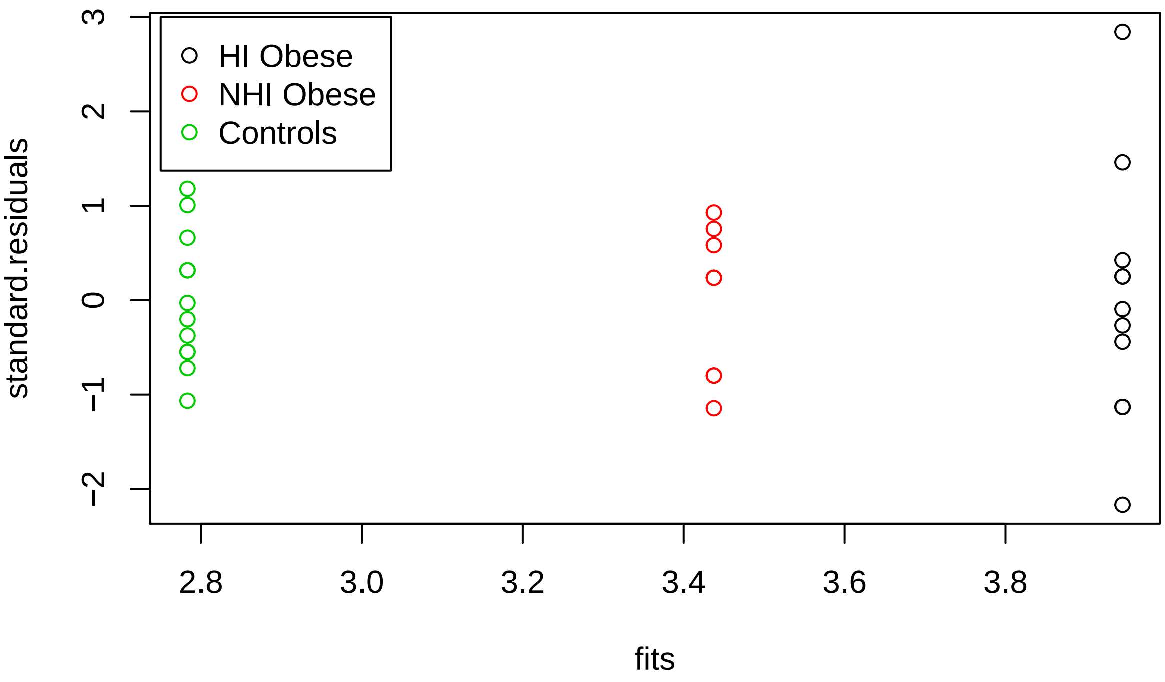
```
st.dev.pool <- sqrt( (10*sd(hi.obese)^2 + 7*sd(nhi.obese)^2 + 11*sd(controls)^2)/ (28) )
mean.dif <- mean(hi.obese) - mean(nhi.obese)
conf.left <- mean.dif - qt(.975,28)*st.dev.pool*sqrt((1/11) + (1/8))
conf.right <- mean.dif + qt(.975,28)*st.dev.pool*sqrt((1/11) + (1/8))
cat("Confidence Interval for Different in Means: (", conf.left,
      ",", conf.right, ")")
```

```
## Confidence Interval for Different in Means: ( -0.04302617 , 1.058935 )
```

(c): Using an appropriate model, examine the standardized residuals for all the observations to look for any systematic effects and to check the Normality assumption.

Solution: We can use the fits from before and the pooled standard deviation found previously to find standardized residuals for each of our factors.

```
fits <- c(hi.obese.fit,nhi.obese.fit,controls.fit)
fits_labels <- factor(c(rep(0,11), rep(1,8), rep(2,12)))
observed <- c(hi.obese, nhi.obese, controls)
standard.residuals <- (fits-observed)/st.dev.pool
plot(fits,standard.residuals, col = fits_labels)
legend(2.75,3,c("HI Obese","NHI Obese", "Controls"),col = 1:3, pch = 1)
```



The NHI Obese appears as if it might not be normal, but with each factor having so little observations, that is difficult to assess. That being said, the Hyperinsulinemic Obese appear to have much more variation compared to the non-hyperinsulinemic obese and controls. Indeed, the latter two standard residuals are primarily within 1 standard deviation, while the hyperinsulinemic obese goes out to up to three standard deviations.

Exercise Seven

Question

For the balanced data in the table below (Table 6.10), the analyses in Section 6.4.2 showed that the hypothesis tests were independent. An alternative specification of the design matrix for the saturated model (6.9) with the corner point constraints $\alpha_1 = \beta_1 = (\alpha\beta)_{12} = (\alpha\beta)_{21} = (\alpha\beta)_{31} = 0$. so that

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \\ \beta_2 \\ (\alpha\beta)_{22} \\ (\alpha\beta)_{32} \end{bmatrix} \quad \text{and} \quad \boldsymbol{X} = \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & -1 & 0 & -1 \\ 1 & 0 & 1 & -1 & 0 & -1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \end{bmatrix},$$

where the columns of $\boldsymbol{X}$ corresponding to the terms $(\alpha\beta)_{jk}$ are the products of columns corresponding to terms α_j and β_k.

Levels of Factor A	Levels of Factor B		
	B_1	B_2	Total
A_1	6.8, 6.6	5.3, 6.1	24.8
A_2	7.5, 7.4	7.2, 6.5	28.6
A_3	7.8, 9.1	8.8, 9.1	34.8
Total	45.2	43	88.2

Source of variation	Degrees of freedom	Sum of Squares	Mean square	F
Mean	1	648.2700		
Levels of A	2	12.7400	6.3700	25.82
Levels of B	1	0.4033	0.4033	1.63
Interactions	2	1.2067	0.6033	2.45
Residual	6	1.4800	0.2467	
Total	12	664.1000		

Solution

(a): Show that $\boldsymbol{X}^T\boldsymbol{X}$ has the block diagonal form described in Section 6.2.5. Fit the model (6.9) and also the models (6.10) and (6.12) and verify that the results in the second table above (Table 6.12) are the same for this specification of $\boldsymbol{X}$.

Solution: First, we put the matrix in R so we don't have to do this matrix multiplication by hand.

```r
X <- matrix(c( 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
              -1,-1,-1,-1, 1, 1, 1, 1, 0, 0, 0, 0,
              -1,-1,-1,-1, 0, 0, 0, 0, 1, 1, 1, 1,
              -1,-1, 1, 1,-1,-1, 1, 1,-1,-1, 1, 1,
               1, 1,-1,-1,-1,-1, 1, 1, 0, 0, 0, 0,
               1, 1,-1,-1, 0, 0, 0, 0,-1,-1, 1, 1),
            ncol = 6, byrow = FALSE)
t(X)%*%X
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]   12    0    0    0    0    0
## [2,]    0    8    4    0    0    0
## [3,]    0    4    8    0    0    0
## [4,]    0    0    0   12    0    0
## [5,]    0    0    0    0    8    4
## [6,]    0    0    0    0    4    8
```

This is indeed a block diagonal matrix.

(b): Show that the estimates for the means of the subgroup with the treatments A_3 and B_2 for two different models are the same as the values given at the end of Section 6.4.2.