Name: Solution

AMS 256 Exam 3, Thursday, May 26th, 2016

You MUST show all your work and justify all steps! Solutions are due 5 pm Friday, May 27th as a hard coy (you can drop in my office) or as a pdf file via email (juheelee@soe.ucsc.edu). Do *not* discuss with anyone.

1 (40 pts) Christensen (in Exercise 7.5) presents mathematics ineptitude scores (Score y_{jik}) for a group of $N = 35$ students [1] categorized by

- Major i (1 = Economics, 2 = Anthroplology, and 3 = Sociology);
- High school background ("BG") j (1 = Rural and 2 = Urban).

The output from fitting a 2-way ANOVA model with interaction is on the last page. The model is

$$y_{ijk} = \mu + \alpha_i + \eta_j + \gamma_{ij} + e_{ijk}.$$

Also, you do not need to read the §7.2 ("2-way ANOVA with interaction"), but it might help just getting familiar with the model.

1a. Which group of students has the lowest score? (What is it?) Which group of students has the highest score? (What is it?)

I find the largest and lowest fitted ineptitude score for:

$\hat{y}$	value	Major	BG
$\mu + Major2$	$0.89 + 1.99 = 2.88$	2 (Ant)	1 (rural);
μ	$= 0.89$	1 (Econ)	1 (rural).

Note that R reports LS fits of the parameters with $Major1 = BG1 = 0$.

1b. In the `summary(.)` output there is an F-statistic, $F = 2.553$ with 5 and 29 degrees of freedom.

(i) What are the null and alternative hypotheses being tested?

The F-test tests the reduced model $y_{ijk} = \mu$ versus the full model.

(ii) What conclusion would you make? (Please state in general terms that relate to the groups rather than parameters).

There is (mild) evidence that there are some differences across majors or backgrounds.

[1] I have fudged the data a bit – not the same as in the book.

1c. In the `anova(.)` output the p-value on the line corresponding to BG is large, yet in the summary from lm the p-value for BG2 is small. Do the p-values from the two summaries contradict each other? Explain what is being tested and what it means in this context. Is the students background relevant for predicting the score?

The R function `anova`:
uses type-I SS's That is, the p-value for BG is for testing `Score` $\sim$ `Major + BG` against `Score` $\sim$ `Major`.

The R function `summary(lm)`:
reports t-tests for one coefficient being zero. That is, it tests `Score` $\sim$ `Major*BG` (full model with main effects and interaction effects) against `Score` $\sim$ `Major + Major:BG` (reduced model with main effects of "major" and interaction effects only).

2. (50 pts) Consider the model $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$, with $e_i \sim N(0, \sigma^2)$, i.i.d. Use the following data;

obs	1	2	3	4	5	6	7	8
y_i	82	79	74	83	80	81	84	81
x_{i1}	10	9	9	11	11	10	10	12
x_{i2}	15	14	13	15	14	14	16	13

Please show your work. Do *not* use a regression or linear models computer program. Using R for simple algebra is okay.

2a. Estimate β_1, β_2 and σ^2.

$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = [2.6, 3.7]^T$, $\hat{\sigma}^2 = MSE = 4.7$ (see below)

2b. Give 95% confidence intervals for β_1 and $\beta_1 + \beta_2$

Use

$$t = \frac{\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}} - \boldsymbol{\lambda}^T\boldsymbol{\beta}}{\sqrt{MSE\,\boldsymbol{\lambda}^T H \boldsymbol{\lambda}}} \sim t_{n-2} \Rightarrow p(\boldsymbol{\lambda}^T\boldsymbol{\beta} \in [\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}} \pm q\sqrt{MSE\,\boldsymbol{\lambda}^T H \boldsymbol{\lambda}}]) = 1 - \alpha$$

where q is the 2.5% upper tail cutoff of the central t_{n-2} distribution; $H = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\boldsymbol{\lambda}^T = [1, 0]$ (for β_1) and $[1, 1]$ (for $\beta_1 + \beta_2$). We find:
95% C.I. for β_1 is $(1.1, 4.2)$.
95% C.I. for $\beta_1 + \beta_2$ is $(5.9, 6.8)$.

2c. Perform a $\alpha = 0.01$ test for $H_0:\ \beta_2 = 3$

Again use the same test statistic t, as in **2b**, now for $\boldsymbol{\lambda}^T = [0, 1]$ and $\boldsymbol{\lambda}^T\boldsymbol{\beta} = 3$, to find $t = 1.6$ and (2-sided) p-value of $p = 0.15$. We fail to reject.

2d. Find the p-value for the test of $H_0: \ \beta_1 - \beta_2 = 0$.

Again use the same test statistic t, as in **2b**, now for $\boldsymbol{\lambda} = [1, -1]^T$ and $\boldsymbol{\lambda}^T\boldsymbol{\beta} = 0$, to find $t = -1$ and (2-sided) p-value of $p = 0.35$. We fail to reject.

```
y <- c(82, 79, 74, 83, 80, 81, 84, 81)
X <-  cbind(c( 10, 9, 9, 11, 11, 10, 10, 12),
            c( 15, 14, 13, 15, 14, 14, 16, 13))
n <- length(y)

## 2a. estimate b1, b2, sig2
H <- solve(t(X) %*% X)  # (X'X)^-1
b <- H %*% t(X) %*% y   # b-hat:                      2.6, 3.7
e <- y-X%*%b
MSE <- sum(e*e)/(n-2)   # MSE = estimate of sig2:   4.7

## 2b. CI for b1
q <- qt(0.975, n-2)    # 2.5% tail cutoff for t(n-1) distribution
b[1] + c(-1,1)*q*sqrt(MSE*H[1,1])  # C.I. for b1:  1.1 to 4.2

## 2b. CI for b1+b2
lam <- c(1,1)           # lam = (1,1)
sum(b) + c(-1,1)*q*sqrt(MSE*t(lam)%*%H%*%lam) # CI fo (b1+b2)
                                              ## 5.9 to 6.8
## 2c Test b2=3
s <- sqrt(MSE*H[2,2])
tt <- (b[2] - 3)/s      # test statistic t=(b2hat - 3)/sqrt(MSE*..)
                        # tt= 1.6
2*pt(tt,n-2,lower.tail=F) # p-value for hypothesis test = 0.15

## 2d. Test b1-b2=0
lam <- c(1,-1)
s <- sqrt(MSE*t(lam)%*%H%*%lam)
tt <- (lam %*% b - 0)/s   # tt = -1
2*pt(tt,n-2,lower.tail=T) # p-value for hypothesis test = 0.35
```

3. (15 pts) Show that for a linear model with an intercept, R^2 is simply the square of the correlation between the data y_i and the predicted values $\hat{y}_i$, where $\hat{\boldsymbol{y}} = [\hat{y}_1, \ldots, \hat{y}_n]^T = \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

Let $\bar{y} = \frac{1}{n}\sum y_i$ and $\bar{\hat{y}} = \frac{1}{n}\sum \hat{y}_i$ denote the (empirical) mean of y_i and $\hat{y}_i$. Let $S_y = \frac{1}{n}\sum(y_i - \bar{y})^2 = \frac{1}{n}\boldsymbol{y}^T(I - \boldsymbol{P}_1)\boldsymbol{y}$ denote the (empirical) variance of $\boldsymbol{y}$ where $\boldsymbol{P}_1$ is the projection operator onto $C(\mathbf{1})$. Similarly $S_{\hat{y}} = \frac{1}{n}\sum(\hat{y}_i - \bar{\hat{y}})^2 = \frac{1}{n}\hat{\boldsymbol{y}}^T(I - \boldsymbol{P}_1)\hat{\boldsymbol{y}}$ and $S_{y,\hat{y}} = \frac{1}{n}\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})$.

Thus,

$$\begin{aligned}
[\mathrm{corr}(\boldsymbol{y}, \hat{\boldsymbol{y}})]^2 &= \frac{S^2_{y,\hat{y}}}{S_y\, S_{\hat{y}}} = \frac{(\boldsymbol{y}^T(I - \boldsymbol{P}_1)\hat{\boldsymbol{y}})^2}{\boldsymbol{y}^T(I - \boldsymbol{P}_1)\boldsymbol{y}\, \hat{\boldsymbol{y}}^T(I - \boldsymbol{P}_1)\hat{\boldsymbol{y}}} = \frac{(\boldsymbol{y}^T(I - \boldsymbol{P}_1)\boldsymbol{P}\boldsymbol{y})^2}{\boldsymbol{y}^T(I - \boldsymbol{P}_1)\boldsymbol{y}\, \boldsymbol{y}^T\boldsymbol{P}(I - \boldsymbol{P}_1)\boldsymbol{P}\boldsymbol{y}} \\
&= \frac{\left[\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{P}_1\boldsymbol{y}\right]^2}{\boldsymbol{y}^T(I - \boldsymbol{P}_1)\boldsymbol{y}\ (\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{P}_1\boldsymbol{y})} = \frac{\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{P}_1\boldsymbol{y}}{\boldsymbol{y}^T(I - \boldsymbol{P}_1)\boldsymbol{y}} \\
&= \frac{SSReg}{SSTotal(correctedformean)} = R^2
\end{aligned}$$

```
> ## Fit a 2-way ANOVA model: ##################################
> summary(lm(Score1 ~ as.factor(Major)*as.factor(BG), data=dat))

Call:
lm(formula = Score1 ~ as.factor(Major) * as.factor(BG), data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.60236 -0.66773 -0.02406  0.52986  2.17744 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)   
(Intercept)                        0.8893     0.4033   2.205  0.03554 * 
as.factor(Major)2                  1.9860     0.6377   3.114  0.00413 **
as.factor(Major)3                  1.1889     0.6377   1.864  0.07244 . 
as.factor(BG)2                     1.2564     0.5207   2.413  0.02237 * 
as.factor(Major)2:as.factor(BG)2  -1.6631     0.8233  -2.020  0.05270 . 
as.factor(Major)3:as.factor(BG)2  -1.6130     0.8233  -1.959  0.05977 . 
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 0.988 on 29 degrees of freedom
Multiple R-squared:  0.3057,Adjusted R-squared:  0.1859 
F-statistic: 2.553 on 5 and 29 DF,  p-value: 0.04945 

> 
> ## Print the ANOVA Table ####################################
> anova(lm(Score1 ~ as.factor(Major)*as.factor(BG), data=dat))
Analysis of Variance Table

Response: Score1
                               Df  Sum Sq Mean Sq F value  Pr(>F)  
as.factor(Major)                2  6.0755 3.03776  3.1123 0.05964 .
as.factor(BG)                   1  0.8623 0.86233  0.8835 0.35502  
as.factor(Major):as.factor(BG)  2  5.5228 2.76142  2.8291 0.07543 .
Residuals                      29 28.3058 0.97606                  
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 
> 
```