

Stat 511 HW#4 Spring 2008 (corrected)

1. The data set in the following table is due originally to Chau and Kelley (*Journal of Coatings Technology* 1993) and has been used by many authors since. Given are values of

y = coating opacity

x_1 = (weight) fraction of pigment 1

x_2 = (weight) fraction of pigment 2

x_3 = (weight) fraction of a polymeric binder

for $n = 14$ specimens of a coating used on some identification labels.

| x_1 | x_2 | x_3 | y |
|-------|-------|-------|-----------|
| .13 | .67 | .20 | .710,.680 |
| .45 | .35 | .20 | .802,.822 |
| .45 | .21 | .34 | .823,.798 |
| .13 | .53 | .34 | .698,.711 |
| .29 | .51 | .20 | .772 |
| .45 | .28 | .27 | .818 |
| .29 | .37 | .34 | .772 |
| .13 | .60 | .27 | .700 |
| .29 | .44 | .27 | .861,.818 |

Note that $x_1 + x_2 + x_3 = 1$.

a) Argue that two equivalent analyses of this data set can be made using the regression models

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \quad (1A)$$

and

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \varepsilon_i \quad (1B)$$

Then argue that two equivalent analyses of this data set can be made using the regression models

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{1i} x_{2i} + \beta_5 x_{1i} x_{3i} + \beta_6 x_{2i} x_{3i} + \varepsilon_i \quad (2A)$$

and

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{1i}^2 + \alpha_4 x_{2i}^2 + \alpha_5 x_{1i} x_{2i} + \varepsilon_i \quad (2B)$$

(This is a so-called "mixture study." For such a study, the first of each of these two pairs in some sense treats the linearly dependent predictors symmetrically, while the second is probably easier to think about and work with using standard software.)

Use R to do the following.

b) Fit model (1B) using `lm()`. Based on this fit, compute the least squares estimate of the parameter vector $\boldsymbol{\beta}$ in model (1A). Use the estimated covariance matrix for $\boldsymbol{\alpha}_{OLS}$ to find an

estimated covariance matrix for β_{OLS} . Then fit model (1A) directly using `lm()`. (A "-1" in the model specification will fit a no-intercept regression.)

c) Fit model (2B) to these data and normal plot standardized residuals.

d) In the model (2B) test $H_0: \alpha_3 = \alpha_4 = \alpha_5 = 0$. Report a p -value. Does quadratic curvature in response (as a function of the x 's) appear to be statistically detectable?

e) Use some multivariate calculus on the fitted quadratic equation and find the location (x_1, x_2) of an absolute maximum. Use R matrix calculations to find 90% two-sided confidence limits for the mean response here. Then find 90% two-sided prediction limits for a new response from this set of conditions.

2. (Testing "Lack of Fit" ... See Section 6.6 of Christensen) Suppose that in the usual linear model

$$Y = X\beta + \varepsilon$$

X is of full rank (k). Suppose further that there are $m < n$ distinct rows in X and that $m > k$. One can then make up a "cell means" model for Y (where observations having the same corresponding row in X are given the same mean response) say

$$Y = X^*\mu + \varepsilon$$

This model puts no restrictions on the means of the observations except that those with identical corresponding rows of X are equal. It is the case that $C(X) \subset C(X^*)$ and it thus makes sense to test the hypothesis $H_0: EY \in C(X)$ in the cell means model. This can be done using

$$F = \frac{Y'(P_{X^*} - P_X)Y / (m - k)}{Y'(I - P_{X^*})Y / (n - m)}$$

and this is usually known as testing for "lack of fit."

Use R and matrix calculations to find a p -value for testing lack of fit to the quadratic regression model (2B) in Problem 1.

3. Below is a small table of fake 2-way factorial data. Enter them into R in three vectors of length $n = 12$. Call these vectors "y", "A", and "B".

| | Level 1 of B | Level 2 of B | Level 3 of B |
|--------------|--------------|--------------|--------------|
| Level 1 of A | 12 | 13,14,15 | 20 |
| Level 2 of A | 8 | 10 | 6,7 |
| Level 3 of A | 10 | 13 | 7 |

a) Create and print out an R data frame using the commands

```
> d<-data.frame(y,A,B)
> d
```

b) Turn the numerical variables A and B into variables that R will recognize as levels of qualitative factors by issuing the commands

```
> d$A<-as.factor(d$A)
> d$B<-as.factor(d$B)
```

Then compute and print out the cell means by typing

```
> means<-tapply(d$y,list(d$A,d$B),mean)
> means
```

You may find out more about the function `tapply` by typing

```
> ?tapply
```

c) Make a crude interaction plot by doing the following. First type

```
> x.axis<-unique(d$B)
```

to set up horizontal plotting positions for the sample means. Then make a "matrix plot" with lines connecting points by issuing the commands

```
> matplot(c(1,3),c(5,25),type="n",xlab="B",ylab="Mean
Response",main="y")
> matlines(x.axis,means,type="b")
```

The first of these commands sets up the axes and makes a dummy plot with invisible points "plotted" at (1,5) and (3,25). The second puts the lines and identifying A levels (as plotting symbols) on the plot.

d) Set the default for the restriction used to create a full rank model matrix, run the linear models routine and find both sets of "Type I" sums of squares by issuing the following commands

```
> options(contrasts=c("contr.sum","contr.sum"))
> lm.out1<-lm(y~A*B,data=d)
> summary.aov(lm.out1,ssType=1)
> lm.out2<-lm(y~B*A,data=d)
> summary.aov(lm.out2,ssType=1)
```

See if anything changes if you ask R to compute "Type III" sums of squares by issuing the command

```
> summary.aov(lm.out1,ssType=3)
```

(In the past R has failed to respond to the request for Type III sums of squares without warning you that it is going to fail to do so.)

e) Start over with this problem, doing the calculations "from scratch" using your basic linear models knowledge and matrix calculations in R. Compute all of Type I, Type II and Type III sums of squares here, using the sum restriction in the first two cases (and the order of factors A,B). Then compute Type I and Type II sums of squares using the SAS baseline restriction.

f) Now suppose that by some misfortune, the observation from the (1,3) cell of this complete 3×3 factorial somehow gets lost and one has only $n = 11$ observations from $k = 8$ cells (and thus "incomplete factorial" data). Test the hypothesis that at least for the cells where one has data, there are no interactions, i.e. $E\mathbf{Y} \in C\left(\left(\mathbf{1} \mid \mathbf{X}_{\alpha^*} \mid \mathbf{X}_{\beta^*}\right)\right)$. (Note that this matrix $\left(\mathbf{1} \mid \mathbf{X}_{\alpha^*} \mid \mathbf{X}_{\beta^*}\right)$ should be of full rank.)

g) In the incomplete factorial context of part f), the function $\mu + \alpha_1^* + \beta_3^*$ is estimable. What is the OLS estimate for it? (Note that this is the mean response for the missing cell only if the same no-interaction model used to describe the 8 cells extends to the 9th. This is the kind of assumption one makes in regression analysis when using a fitted prediction equation to estimate a mean response at a set of conditions not in one's original data set. It might well be argued, however, that the link between observed and unobserved conditions is intuitively stronger with quantitative factors than with qualitative factors.)