STAT 540, Fall 2015

Due on **12:00 A.M., Friday, Oct 30th**

Homework 8

Homework format for all STAT 540 homework this term: Please label all problems clearly and turn in an organized homework assignment. You don't need to spend hours producing beautifully typeset homework, but you won't get credit if we can't find or read your answer. Unless noted otherwise, turn in the following (as appropriate for the problem).

- Theoretical derivation (when asked for).
- Numerical results **with an explanation of your solution**, written in complete sentences. If computer code is absolutely necessary to provide context here, then include it–nicely formatted–within the solution (otherwise, see below).
- Appropriate graphics. Use informative labels, including titles and axis labels. Try to put multiple plots on the page by using, for example, the R command `par(mfrow=c(2,2))`.
- **Only as necessary**: Final clean computer code used to answer the problem **attached to the end of your homework**. Only include the rare code excerpts without which we wouldn't be able to figure out what you did. Annotate your code. Number and order the code in order of the problems. When in doubt, leave it out; consider that we will probably never read it.
- Some problems will be relatively open-ended, such as "Here are some data. Analyze them and write a report." I will provide further instructions about reports later. They should be self-contained, with suitable EDA, graphs, numerical results, and **scientific interpretation**. No computer code should be included. The report should be concise: "no longer than necessary".

(1) Write out the following models of elementary/intermediate statistical analysis in the matrix form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$:

(a) A "linear spline" regression model with knots at 3.5 and 7.5 is given as

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 (x_i - 3.5)\mathrm{I}(x_i \geq 3.5) + \alpha_3 (x_i - 7.5)\mathrm{I}(x_i \geq 7.5) x_i + \epsilon_i$$

for $x_i = i$ for $i = 1, 2, \ldots, 10$.

(b) A two way ANCOVA model without interactions but with two quantitative predictors is given by

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_1 x_{1,ijk} + \gamma_2 x_{2,ijk} + \epsilon_{ijk}$$

for $i = 1, 2; j = 1, 2; k = 1, 2$.

(2) Check $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ on page 61 in lecture 8 are both the solution to the normal equation on page 60 in the same lecture. Also, find two different generalized inverse of $(\mathbf{X}'\mathbf{X})$, $\mathbf{A}_1$ and $\mathbf{A}_2$ that $(\mathbf{X}'\mathbf{X})\mathbf{A}_i(\mathbf{X}'\mathbf{X}) = (\mathbf{X}'\mathbf{X})$ for $i = 1, 2$, respectively and they should give you $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$, respectively.

(3) The following problems concerns about the estimability discussed in class.

(a) Consider a factor effect model for a study with a balanced two-way factorial treatment design

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$$

for $i = 1, 2; j = 1, 2, 3$ and $k = 1, 2$. Then, the "LSMEAN" for the two row means are given by $\mu + \alpha_1 + (\beta_1 + \beta_2 + \beta_3)/3 + (\alpha\beta_{11} + \alpha\beta_{12} + \alpha\beta_{13}/3$ and $\mu + \alpha_2 + (\beta_1 + \beta_2 + \beta_3)/3 + (\alpha\beta_{21} + \alpha\beta_{22} + \alpha\beta_{23}/3$. The design matrix is given by

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Are the two LSMEANS estimable? If so, give vectors $\mathbf{A}$ (one for each LSMEANS) that demonstrate your conclusions.

(b) A recent "two-way" factorial study was conducted where factor A was State: Iowa or Missouri, and factor B was a type of stream: Grazed, Buffered, or Grazed and Buffered. As it happened, the investigators could find all three types of streams in Iowa, but only two types in Missouri. The number of observations for each combination of state and type of stream are:

State/ Type of Stream	B	G	B/G
Iowa	1	2	2
Missouri	1	2	0

Consider a factor effects model for this study

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$$

for $i = 1, 2; j = 1, 2, 3$ and $k = 1, 2, \ldots, n_i$. The design matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

For each of the following quantities, is it estimable? If so, what **A** vector demonstrates it is estimable?

(i) The mean for B/G streams in Iowa.
(ii) The mean for B/G streams in Missouri.
(iii) The "LSMEAN" for B streams, i.e. the average over IA and MO.
(iv) The "LSMEAN" for B/G streams, i.e. the average over IA and MO.
(v) The average difference between B and G streams, i.e. averaged over IA and MO.
(vi) The average difference between G and B/G streams, i.e. averaged over IA and MO.

(c) Consider an additive effects model for the stream study:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

for $i = 1, 2; j = 1, 2, 3$ and $k = 1, 2, \ldots, n_i$. The design matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

For each of the following quantities, is it estimable? If so, what **A** vector demonstrates it is estimable?

(i) The mean for B/G streams in Iowa.
(ii) The mean for B/G streams in Missouri.
(iii) The "LSMEAN" for B streams, i.e. the average over IA and MO.
(iv) The "LSMEAN" for B/G streams, i.e. the average over IA and MO.
(v) The average difference between B and G streams, i.e. averaged over IA and MO.
(vi) The average difference between G and B/G streams, i.e. averaged over IA and MO.

(d) A very messed up experiment has a $\boldsymbol{\beta}$ vector and $\mathbf{X}$ matrix given by:

$$\mathbf{X} = \begin{bmatrix} \mu & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

(i) Is μ estimable? Is α_1 estimable? Is α_3 estimable?

(ii) If $\sigma^2 = 2$, compute $\text{Var}(\hat{\mu})$, $\text{Var}(\hat{\alpha}_1)$, and $\text{Var}(\hat{\alpha}_4 - (\hat{\alpha}_2 + \hat{\alpha}_5 + \hat{\alpha}_6))$. Notice $\alpha_4 - (\alpha_2 + \alpha_5 + \alpha_6)$ is estimable.

(iii) How many degree of freedoms are associated with MSE?

(4) Consider a completely randomized experiment in which a total of 10 rats were randomly assigned to 5 treatment groups with 2 rats in each treatment group. Suppose the different treatments correspond to different doses of a drug in milliliters per gram of body weight as indicated in the following table.

Treatment	1	2	3	4	5
Dose of Drug (mL/g)	0	2	4	8	16

Suppose for $i = 1, 2, 3, 4, 5$ and $j = 1, 2$, y_{ij} denotes the weight at the end of the study of the jth rat from the ith treatment group. Furthermore, suppose

$$y_{ij} = \mu_i + \epsilon_{ij}$$

where $\mu_1, \ldots, \mu_5$ are unknown parameters and the ϵ_{ij} are i.i.d. $N(0, \sigma^2)$ for some unknown $\sigma^2 > 0$. Use the `R` code and partial output provided below to answer the following questions.

```
> d=rep(c(0,2,4,8,16),each=2)
> #y is the data vector with entries ordered to appropriately
> #match the vector d.
> dose=factor(d)
> o1=lm(y~ose)
> summary(o1)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 351.000 6.576 53.372 4.37e-08 ***
dose2 -10.000 9.301 -1.075 0.331406
```

```
dose4 -6.000 9.301 -0.645 0.547277
dose8 -17.000 9.301 -1.828 0.127119
dose16 -70.500 9.301 -7.580 0.000634 ***

> anova(o1)
Analysis of Variance Table
Response: y
Df Sum Sq Mean Sq F value Pr(>F)
dose 6505.6
Residuals 432.5
> is.numeric(d)
[1] TRUE
> o2=lm(y~d)
> anova(o2)
Analysis of Variance Table
Response: y
Df Sum Sq Mean Sq F value Pr(>F)
d 5899.6
Residuals 1038.5
> o3=lm(y~d+dose)
> anova(o3)
Analysis of Variance Table
Response: y
Df Sum Sq Mean Sq F value Pr(>F)
d 0.0004245 ***
dose 0.1907591
Residuals
```

(a) Provide the BLUE of μ_1

(b) Provide the BLUE of μ_2

(c) Determine the standard error of the BLUE of μ_2

(d) Conduct a test of $H_0 : \mu_1 = \mu_2$. Provide a test statistic, the distribution of that test statistic (be very precise), a p-value, and a conclusion.

(e) Provide an F-statistic for testing $H_0 : \mu_3 = \mu_4$

(f) Does a simple linear regression model with body weight as a response and dose as a quantitative explanatory variable fit these data adequately? Provide a test statistic, its degrees of freedom, a p-value, and a conclusion.

(g) Provide a matrix $\mathbf{C}$ and a vector $\mathbf{d}$ so that the null hypothesis of the test in part (f) may be written as $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$, where $\boldsymbol{\beta} = (\mu_1, \ldots, \mu_5)'$.

(h) Fill in the missing entries in the ANOVA table produced by the `R` command `anova(o3)`. (This is the last `R` command in the provided code.)

(5) A soft drink bottler is analyzing the vending machine service routes in his distribution system. He is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. It is thought that the two most important variables affecting delivery time (Y) are the number of cases of product stocked (X_1) and the distance walked by the route driver (X_2). An industrial engineer collected 25 measurements on these three variables. The data are found in Montgomery, Peck, and Vining (2001) and are reproduced in the file `softdrink.dat` on the course webpage.

(a) Prepare a scatterplot matrix for these three variables. Comment on your observations.

(b) Fit the linear regression model Y on the two predictor variables. Report the summary table from the estimated model.

(c) Test for an overall regression relation between Y and the two predictor variables.

(d) Plot the internally studentized and externally studentized residuals versus the fitted values from the model in part (b). Comment on your graphs.

(e) For the model in part (b) compute and investigate various regression diagnostics including h_{ii} values, DFFITS, Cook's distance, and DFBETAS Write a short paragraph summarizing the results.

(f) Based on your results in part (e), select a final model. Provide the summary table and discuss the differences between this model and your model in part (b).