STAT 608: Homework 5
Total : 100 pts

1. (14pts - 2pts each part)

There are three drugs A, B, and C with unknown efficacies β_1, β_2 and β_3. You are asked to estimate the β_i's, using a clinical trial with the following protocol and observations:

Use drug A and observe y_1 the time to recovery,

Use drug B and observe y_2 the time to recovery,

Use drug C and observe y_3 the time to recovery,

Use a cocktail (mix) of drugs A and B and observe y_4 the time to recovery,

Use a cocktail (mix) of drugs A and C and observe y_5 the time to recovery,

Use a cocktail (mix) of drugs B and C and observe y_6 the time to recovery,

Use a cocktail (mix) of drugs A, B and C and observe y_7 the time to recovery.

It is believed that the measured time to recovery (response) is a combination of the true efficacy plus a random error, and the random errors are independent and identically distributed (i.i.d.).

(a) Write a linear model in matrix form for this experiment, and clearly write the design matrix X and the response vector y.

(b) Compute $X^T X$, $X^T y$ and write the normal equations.

(c) Find the constant c so that

$$A = c \begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix}$$

is the inverse of $X^T X$ (Hint: First compute $A(X^T X)$. Is this matrix diagonal?)

(d) Find the least-squares estimates of the efficacies using the usual formula $(X^T X)^{-1} X^T y$

Now suppose that the variance of errors is proportional to the number of drugs used in each measurement.

(e) Write the appropriate matrix of weights W in this case.

(f) Write the normal equations for the weighted least squares and solve them. (Here you can use R or a computer to invert the matrices you need.)

(g) Do these estimates make sense? Explain first why the estimates are **unbiased**, and second why the individual measurements are now more heavily weighted in the parameter estimates than they were in the ordinary LSE.

Question 2 [3+3+3=9 **points**]: (Old Qualifying Exam Question) A randomized trial was con-ducted to investigate the relationship between a continuous response y and four treatments A, B, C, and D. The sample size was $n = 200$, with 50 observations in each of the four treatment groups. Let y be the 200×1 vector of response values, ordered so that the first 50 entries are for treatment group A, the next 50 for B, then C, and finally D. The regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ was fit, where $\mathbf{X}$ is the 200×4 design matrix given by

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

and where each entry is a column vector of length 50 . The estimated regression coefficients were $\hat{\boldsymbol{\beta}}' = [37.5, -11.5, 1.0, -27.7]$, with standard errors $2.75, 3.89, 3.89, 3.89$, and residual standard deviation $\hat{\sigma} = 19.45$. Also:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.02 & -0.02 & -0.02 & -0.02 \\ -0.02 & 0.04 & 0.02 & 0.02 \\ -0.02 & 0.02 & 0.04 & 0.02 \\ -0.02 & 0.02 & 0.02 & 0.04 \end{bmatrix}$$

(a) Interpret each of the four regression parameters.

(b) What is an approximate 95% confidence interval for the mean difference in response between treatment groups B and A (so, the difference $\mu_B - \mu_A$) ?

(c) What is an approximate 95% confidence interval for the mean response in treatment group B?

3. Question 3: [5+5 =10 pts] . Consider a dataset with 6 observations with 2 covariates. The y-values are {3,2,4,6,7,1}. The residuals are {0.5, 0.25, -0.5, 0.5, -1, 0.25}.
   a. Construct the ANOVA table for the simple linear regression.
   b. Find the value of R^2 and adjusted R^2

Question 4 [3+3= 6 points]: (From Weisberg, 2005) We are interested in the linear model $Y=\beta_0+\beta_1 x_1+\beta_2 x_2+e$.

(a) Suppose we fit the model above to data for which $x_1=2.2x_2$ with no error (that is, all residuals $=0$). For example, X_1 could be a weight in pounds, and X_2 the weight of the same object in kg. Describe the appearance of the added-variable plot for x_2 after x_1 had been added to the above model. Explain why. Assume that Y has a correlation with the predictors that is neither 0 nor 1 . Hint: Think about what goes on the x-axis and the y-axis of the added variable plot. You should notice something interesting about one of those residual vectors.

(b) Again referring to the model above, this time suppose that Y and X_1 are perfectly correlated, so $Y=3X_1$, without any error. Describe the appearance of the addedvariable plot for x_2 after x_1 had been added to the model. Explain. Assume this time that the correlation between the predictors is between 0 and 1 .

Question 5 [2+2+2+2+2+2 = 12 points]: Solve question 3, Chapter 6 (page 216)