Biostatistics 140.653
Third Term, 2022
Problem Set 2

Instructions: Feel free to work with other students on interpreting the questions posed in the problem set and analysis strategy and implementation (i.e. coding and model fitting). However, each student must write-up their own solutions. Write as if for a scientific journal. Be brief and accurate. Be numerate and avoid non- essential statistical jargon. Submit your text and code in an .Rmd file and the compiled report as a .pdf file.

Due in CoursePlus drop box: Friday, February 25 by 12:00pm EST

I. Matrix Representation of Multiple Linear Regression

Upon successful completion of this problem, a student should be able to:

- Write and explain the classical multiple linear regression model in scalar and vector notations.
- Conduct a particular MLR by "hand" using matrix manipulations.
- Demonstrate the link between SLR and two-sample t-test
- Explore the impact of violation of the common variance assumption
- (Optional) Derive the least squares (Gaussian maximum likelihood) estimates of the regression coefficients, predicted values and residuals using matrix notation.

1. Use the following 5 observations and write the simple linear regression model in matrix terms. Then using the least squares calculations in matrix notation, compute estimates for the simple linear regression intercept and slope.

Y	X
-0.1	1
2.9	3
6.2	5
7.3	7
10.7	9

2. Write an R function that takes the vector Y and matrix X as input then calculates and returns each the following components:
   a. the least squares estimates of the regression coefficients
   b. the variance-covariance matrix of the least squares estimates
   c. the correlation between the two regression coefficients
   d. the vector of predicted values $X(X'X)^{-1}X'Y = HY$
   e. the vector of residuals $(I- X(X'X)^{-1}X')Y = (I-H)Y$.

3. Using the R function from Question 2, verify your estimates of the simple linear regression intercept and slope computed in Question 1. Using the standard error estimate for the simple linear regression model slope, construct a 95% confidence interval for the true slope.

4. Suppose you have conducted a randomized controlled trial of an intervention (TRT = 1) vs. placebo (TRT = 0), where n_1 and n_0 patients received the intervention and placebo, respectively. For each patient, you have measured a continuous outcome Y with the goal of comparing E(Y|TRT=1) to E(Y|TRT=0). I ask that you fit the following linear regression model:

$$Y_i = B_0 + B_1 X_i + \varepsilon_i, \varepsilon_i \; iid \; N(0, \sigma^2), X_i = 1 \; if \; TRT = 1, 0 \; if \; TRT = 0$$

   a. Write out the model above using matrix notation and then using matrix calculations solve for the least squares estimates of B_0 and B_1 and $Var(\hat{B}_1)$. HINT: You will show that the model above is the same as conducting a two-sample t-test, assuming the same variance in the intervention and placebo groups. The estimate of the intercept should be the sample mean in the placebo arm, the estimate of the slope should be the difference in the sample means comparing the intervention and control groups and the $Var(\hat{B}_1) = \sigma^2/n_0 + \sigma^2/n_1$.

   b. Now suppose that the true model is:

$$Y_i = B_0 + B_1 X_i + \varepsilon_i$$

   where $X_i = 1 \; if \; TRT = 1, 0 \; if \; TRT = 0$ and $\varepsilon_i \sim N(0, \sigma^2(X_i))$, where $\sigma^2(X_i) = \sigma^2$ if $X_i = 1$ and $\sigma^2(X_i) = 2\sigma^2$ if $X_i = 0$ and $Cov(\varepsilon_i, \varepsilon_j)$ for all i and j. Under the true model, $Var(\hat{B}_1) = \sigma^2/n_1 + 2\sigma^2/n_0$.

   Make a figure to compare the width of the 95% confidence interval for B_1 based on the model I asked you to fit (part a) and the true model. Set values for n_1 and n_0 and allow σ^2 to range from 1 to 100. Describe the impact of fitting the model I asked you to fit (i.e. a model that assumes the same variance in each group) vs. a more flexible model that would allow the variance of the residuals to depend on the assigned treatment group.

5. OPTIONAL: Under the Gaussian multiple linear regression framework, write the log likelihood function for the regression coefficients and residual variance in matrix terms and derive the mle's for the regression coefficients. Derive their joint distribution, as well as the distribution of the predicted values and residuals.

II. Advanced Inferences for Linear Regression

Upon successful completion of this problem, a student will be able to:

- Display data and fitted values from a multiple linear regression
- Calculate a confidence interval for a linear function of the regression parameters
- Calculate a confidence interval for a non-linear function of the regression parameters
- Test a null hypothesis involving more than one regression coefficient by using a likelihood ratio test (F-test in the linear model)
- Write a coherent, jargon-free summary for a public health journal of a multiple linear regression analysis.

Use the NMES data set on persons 65 years of age and above to address the question of whether older men and women of the same age use roughly the same quantity of medical services. That is, estimate the difference in average medical expenditures between men and women as a function of age.

See the Datasets folder in the on-line library to gain access to the dataset (provided as an R workspace with dataframe "nmes", see NMESRworkspace.zip) and the codebook describing the available variables in the data.

1. Define:
    - agem65 = age-65
    - age_sp1 = $(\text{age}- 75)^{+}$
    - age_sp2=$(\text{age}-85)^{+}$
    - female = 1 for females and 0 for males.

    Fit a MLR of expenditures on age and gender as:

    expenditure ~ (agem65 + age_sp1 + age_sp2) + female +
    female*(agem65 + age_sp1 + age_sp2)

    Write a short, scientific interpretation of each coefficient in the model; use the estimated coefficient with corresponding confidence interval.

2. Create a figure that displays the data and the predicted values from the fit of the MLR model from Question1.

3. Using the model fit in Step 1 above, make a plot of the expected difference between women and men in expenditures as a function of age. Add a horizontal line at 0. Note that this difference is a simple function of the estimated coefficients from the model. (Hint: Start by writing out the regression model for females and males, both will be a function of age and the regression coefficients. Then take the difference and plug in the estimated regression coefficients and allow age to range from 65 to 94.)

4. Use the appropriate linear combination of regression coefficients to calculate the estimated difference between females and males in average expenditures and its standard error at ages 65, 75 and 85 years. Complete the table below.

Age	Estimated difference in expenditures Female vs. male	Linear Model Std Error	Linear Model 95% CI	Bootstrap Std Error	Boostrap 95% CI
65					
75					
85					

5. Now estimate the ratio of the average expenditures comparing women to men at age 65. This is a non-linear function of the regression coefficients from step 1. Use the delta method to estimate the standard error of this statistic and make a 95% confidence interval for the true value given the model.

6. The data used in this regression are highly skewed and heteroscedastic (unequal variances across observations). Hence, the assumptions of the linear regression are not consistent with patterns in the data. As you will learn shortly, the estimates are still unbiased, but the standard errors and confidence intervals are likely biased. Hence, your inferences (tests and CIs) that depend on both the mean and variance estimates may be incorrect.

   To check, use the bootstrap procedure to estimate the standard errors and confidence intervals for the differences in the table in Question 4 and for the ratio in Question 5. Compare the results obtained directly from the linear regression with those obtained using bootstrapping.

7. Test the null hypothesis that on average, males and females use the same quantity of medical services; i.e. are the mean expenditures at any age the same for males and females? Use a likelihood ratio test performed by fitting a null and extended model and comparing the change in –2*log likelihood to the appropriate X^2 statistic. In addition, perform an F-test for the same null hypothesis. Write a sentence or two that summarizes what you learned about the medical expenditures and age from this test and the similarity/difference of the two tests.

8. Using the results of Questions 1-7, write a brief report with sections: Background and objective, data, methods, results, discussion as if for a health services journal. NOTE: The data section should briefly (in a sentence or two) describe the data source (e.g. 1987 NMES). Recall the question: *Do older males and females of the same age use roughly the same quantity of medical services?*