Stat 440/540:

Instructions:

- For problems related to a data set, you can do either by hand (if it is a small data set) or using R. But you will only hand in one version of the homework, either by hand or by using R.
- If using R,

  (1) You also need to hand in R code

  (2) Output from R is not a solution. You shall provide comments/interpretation for the output. In most cases, you should reorganize the output in a table or use your own words to state/explain the output.

  (3) Your appendix should only include R code. You should write a complete solution for each problem. For example, if you need to talk about a graph, paste the graph right before or after your statement. Do not refer the grader to somewhere else.

Assignment 1: Due Thursday, Sep 2 Friday in class

Chapter 1, 1.1, 1.7, 1.12, 1.19, 1.35, 1.37

Assignment 2: Due Sep 16 Friday in class

Chapter 2, 2.1, 2.3, 2.4, 2.13, 2.23, 2.34, 2.50, 2.51

Assignment 3: Due Oct 25 Sunday, but not collected. I will discuss the homework on Oct 26 Monday's class

1. Refer to Grade point average problem 1.19, answer the following questions:

(a) Fit the simple linear regression model with GPA as the response and entrance test score as the predictor. Evaluate the linear relationship assumption. Which diagnostic plot(s) is appropriate for this? Create this plot(s) and comment.

(b) Create a diagnostic plot(s) to evaluate the normality assumption. Comment.

(c) Create a plot(s) to assess the equal variance assumption. Comment.

(d) Are there any outliers in this data set? If so which observation(s)? Suppose you learn that the student corresponding to the 9th observation had received the answers on the entrance exam from his friend. Refit the regression model with this observation removed. How does this affect the fit? Comment on the equal variance and normality assumptions with this observation removed.

(e) Chapter 3, 3.3 (f)

2. Chapter 3, 3.15, 3.16, 3.19, 3.20

Midterm: Oct 3rd, Monday in class

Assignment 4: Due Oct 21 Friday in class

- Chapter 4, 4.5
- Chapter 5, 5.5 (1), (2), (3) and $(4)(\mathbf{X}'\mathbf{X})^{-1}, (5)\mathbf{b}, (6)\mathbf{e}, (7)\mathbf{H}, (8)SSE, (9)\mathrm{Var}(b_0), \mathrm{Var}(b_1)$, (10) $s^2(\hat{y}_h)$ when $x_h = 2$. You can either calculate by hand or by R, but you need to write the steps using matrix approach.
- Chapter 6, 6.22
- Grade Point Average Problem

  Refer to the Grade Point Average problem.

  (a)The Dean wishes to fit a linear regression model with GPA as the response and IQ, HS-Avg, and test-score as the predictors. What assumptions are necessary for this model?

  (b) Fit the model and evaluate the linear relationship assumption. Which diagnostic plots are appropriate for this? Create these plots and comment.

  (c) Create a diagnostic plot(s) to evaluate the normality assumption. Comment.

  (d) Create a plot(s) to assess the constant variance assumption. Comment.

  (e) Are there any outlying residuals in this data set? Are there any outliers in any of the predictor variables? If so which observation(s) and which predictor? Explain how this observation can be a fairly extreme outlier in terms of its x value but still not produce an outlying e_i value.

(f) Conduct statistical tests to determine the importance of each predictor variable. State hypotheses, p-values, and decisions. What do you conclude?

(g) Does your answer from part (f) imply that entrance test score is unrelated to GPA?

(h) What is the multiple coefficient of determination? Comment.

(i) Obtain 95% CI for Expected freshman GPA for a student with test score= 32, IQ = 82, HS-Avg= 72. Interpret.

(j) Recall that Frank Buffay obtained a score of 32 on the entrance test. He also has IQ = 82, HS-Avg= 72. Predict his freshman GPA using a 95% prediction interval. Interpret your prediction interval.

(k) Consider the linear regression model with GPA as the response and IQ (X_1), test score (X_2), and HSAvg (X_3) as the predictors. Comment on the correlation structure between the predictors.

(l) Fit the model with X_2 as the only predictor. Then fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$. Comment on the b_2 estimate in each of these models. How do you explain the difference?

(m) Fit the model with X_1 as the only predictor. Then fit the model with X_1 and X_2 as the predictors. Finally fit the model with X_1, X_2, and X_3 as the predictors. Calculate $SSR(X_1)$, $SSR(X_2|X_1)$, and $SSR(X_3|X_1, X_2)$. Also calculate $R^2(X_2, X_1)$ and $R^2_{Y2|1}$ and interpret. How much does it help to add test score to a model that already has IQ?

Assignment 5: Due Nov 4th Friday in class

Chapter 8: 8.4, 8.40, 8.41, prove 8.12(a),8.12(b),8.12(c) on page 299

Assignment 6: Due Nov 18 Friday in class

Chapter 9: 9.25 (for (a), let's prepare boxplot for each predictor variable instead of dot plot), 9.27

Chapter 10: 10.3, 10.27

Assignment 7: Due Dec 2rd Friday in class

Do problem 1. Methyl mercury of August 2016 take home qual.

UNM Statistics Qualifying Exam Aug 2016 Name: ______________________
Due: 3 P.M., Mon Aug 15, 2016 CODENAME

Qual Take Home (100 points) Complete both problems in this exam. Your report is to be typed, double-spaced, no smaller than ten-point font with one-inch margins, and should be identified by your CODENAME (do not include your name or UNM ID number). Each problem is to be no longer than four pages, and an additional four-page appendix is allowed for each problem but will be examined only at the discretion of the graders; the better constructed your appendix with cross-references from the text, the more likely it is to get examined.

Write your answers completely, but concisely. Insert tables and figures to support your points. Tables and figures should be well-labelled and cross-referenced from text, such as, "in Table 1 . . . ", or if in the appendix, "in Table A1 . . . ". Figures should include appropriate symbols suitable for black-and-white reproduction (that is, avoid use of color if possible; consider symbols, line types, and distinct shades of gray to distinguish categories or values). Computer output without explanation will not be reviewed. As necessary:

1. Plot and describe the data (that is, plot all the individual observations, in addition to summaries of data you might present with the results, such as the mean and confidence intervals).
2. Clearly define population parameters and sample statistics.
3. Clearly specify hypotheses tested and explicitly state the associated model at least once (i.e., write the model equation).
4. Define and assess method assumptions.
5. Write a coherent evidence-based conclusion that a layperson can understand.

You may **not** consult any other person when working on this exam or discuss your exam with anyone else regardless of whether or not the person is taking the exam. You may use your course notes as well as any available books or web resources for the exam. If including computer text tables where alignment is important, then please use a fixed-width font, such as `Courier`, for that text. Any points of clarification can be directed to Prof. Erik Erhardt, `erike@stat.unm.edu`.

Due: Email Ana Parra Lombard `<aparra@math.unm.edu>` with solutions by 3 P.M., Mon Aug 15, 2016, Department of Mathematics and Statistics, University of New Mexico. Please do not turn in a physical copy of your solutions.

(50$^{\text{pts}}$) **1. Methyl mercury**

In a study on methyl mercury in the hair of fishermen in Kuwait, the following variables are collected: fisherman indicator (fisherman): 0=not fisherman, 1=fisherman; age in years (age); residence time in years (restime); height in cm (height); weight in kg (weight); fish meals per week (fishmlwk); parts of fish consumed (fishpart): 0=none, 1=muscle tissue only, 2=muscle tissue and sometimes whole fish, 3=whole fish; methyl mercury in mg/g (MeHg). The primary objective is to assess how the factors affect the methyl mercury levels among fishermen and a control group of non-fishermen.

Data: `www.stat.unm.edu/~erike/exams/UNM_Stat_Exam_Qual_takehome_201608_pr1-DATA_Fisher.txt`

(a) Build a regression model for predicting MeHg: use model selection technique for choosing variables from the dataset; consider all the two-way interactions that include weight as one of the variables; discuss which variables should be retained, which should be dropped. Assess deviations from model assumptions; if the assumptions are violated, try to address those concerns, rebuild the model and reassess the assumptions. State and interpret your final model. (35 pts)

(b) Using the F-test statistic, test whether MeHg concentration is related to fisherman indicator, age, duration of residence, weight, height, fish meals per week, and parts of fish consumed based the final model in part (a). **Only** consider the tests for the variables that are in your final model. Use $\alpha = 0.05$. (5 pts)

(c) Compute the partial determination scores of fisherman indicator, age, duration of residence, fish meals per week, and parts of fish consumed given that weight is in the model for predicting MeHg concentration. Summarize your findings. (5 pts)

(d) Consider the model using only weight and parts of fish consumed as covariate. Examine whether the effect on MeHg concentrations from eating no fish differs from that for each of the other parts consumed by performing three pairwise tests. Use the Bonferroni procedure with $\alpha = 0.05$ and a 95 percent family-wise confidence coefficient. Summarize your findings. (5 pts)

(50$^{\text{pts}}$) **2. Alkaloid concentrations in tea**

As part of a recent consulting project, Erik came across this scenario. The process for producing tea bags with the specified weight and alkaloid concentration from a specific herb ("Herb A") is as follows. A blend of two herbs is made in South Africa (a box weighing roughly 45 kg), then shipped to California to be bagged (producing roughly 29,000 tea bags). The relative contributions in the blend is 4% (60 mg) of Herb A and 96% (1440 mg) of Herb B. Of interest is the total concentration of a specific alkaloid for Herb A as measured by infusion extraction (steeping the tea bag). The process of herb drying, mixing, and bagging was found to be highly reliable (accurate and precise), but the method of growing and harvesting Herb A was found to affect alkaloid concentration. The focus of this problem is to analyze the experimental growing data for Herb A to assess the resulting alkaloid concentration measured in controlled lab conditions.

Three varieties of Herb A (A1, A2, and A3), were grown at research greenhouses with either 0, 23, 45, or 68 g/m^2 of fertilizer (primarily nitrogen supplied as soy bean meal). Eighteen (18) temporary fields were created by placing standard top soil in previously unplanted areas, with small experimental green houses placed over each one. Each of the 3 varieties were randomly assigned to 6 of the 18 green houses, so that each green house had only one variety (to protect from confusion at harvest). The four levels of fertilizer were randomly assigned to equal-sized quarters of each greenhouse. At harvest time, 3 leaves were collected from each of several center plants in each quarter and combined as the sample. Each sample was uniformly processed (dried, crushed, and blended), then portions shipped to three labs for alkaloid concentration measurements. We will analyze the data from one of the labs. Of interest are differences in variety and fertilizer levels.

Data: `www.stat.unm.edu/~erike/exams/UNM_Stat_Exam_Qual_takehome_201608_pr2-DATA_alkaloid.csv`

Analyze the data provided by this experiment. In addition to analyses and comments arising from your own curiousity, please address the following as part of your write-up. It is recommended that you structure your write-up similar to the order of the questions below.

(a) What statistical design is being used, and why? Could a better design have been used, and why or why not?

(b) Is there blocking? If so, what is/are the block(s)?

(c) What is/are the nuisance factor(s) to be "averaged out" in the design?

(d) What is/are the treatment(s)?

(e) What is/are the outcome(s)/response(s)?

(f) Plot the data (not only summaries of the data) in a way that helps you understand what the effects are.

(g) Write out the best full statistical model (in notation, defining the notation you use) and state the model assumptions.

(h) Fit the model written in the previous part (that is, effects in fitted model should be as in the model specification above), and assess and address deviations from model assumptions. This may be an iterative process. Summarize each model fit and the evidence for the decisions made to arrive at your final model, consider moving intermediate model fit details to the appendix. (Note: If model assumptions are not met, try to address that. If you can not address unsatisfied model assumptions, mention this and continue as though the model assumptions are met.)

(i) State and conduct statistical tests for the parameters, and interpret the test results.

(j) Perform pairwise comparisons based on your final model and summarize which pairs of treatment combinations are different.

(k) Discuss anything else of interest, and address the original goal of the experiment.