STAT 423/504 - Homework 2

Due date: Wednesday, February 12. Please submit your homework to the 423/504 mailbox at Padelford B-313 on Wednesday by 5:00 PM on the designated day. Please submit the output and plots, but not your R code unless the quesiton specifically asks for it. Total possible points: 28.5 + 4 Bonus points.

1. (16 points) (ALR, 3.3) The Berkeley Guidance Study enrolled children born in Berkeley, California, between January 1928 and June 1929, and then measured them periodically until age eighteen (Tuddenham and Snyder, 1954). The data we use includes height in centimeters at age 2,9 and 18, (`HT2`, `HT9`, `HT18`), weight in kilograms (`WT2`, `WT9`, `WT18`), and strength in kilograms (`ST2`, `ST9`, `ST18`). As a response we will use the variable `BMI18` which is computed as

$$\texttt{BMI18} = \frac{\texttt{WT18}}{(\texttt{HT18}/100)^2}.$$

   Use data `BGSgirls` in R package `alr4` which is a subset of data set `BGSall` in the same R package containing data only on girls. You can obtain a description of the data set by loading package `alr4` and executing `?BGSgirls`.

   (a) (1 point) Draw the scatterplot matrix of `HT2`, `HT9`, `WT2`, `WT9`, `ST9` and `BMI18`. Compare the scatterplot matrix with the matrix of sample correlations for these variables. What do you observe? Compare the relationship of the predictors with the response and the pairwise predictor relationships.

   R Hint: To obtain the sample correlations matrix use

   ```
   print(cor(BGSgirls[,c(1,2,3,4,6,11)]),3)
   ```

   (b) (2 points) Fit two linear models:

      1. E[`BMI18 | WT9, ST9`], and
      2. E[`BMI18 | HT2, WT2, HT9, WT9, ST9`]

      Print their summaries and comment on the output. Which of the estimates that are in model 2 and not model 1 are significant at the 5% level ($\alpha = 0.05$)?

   (c) (4 points) Conduct an anova comparison of the above 2 models.Test the hypothesis $H_0 : (\beta_{HT2}, \beta_{WT2}, \beta_{HT9}) = (0, 0, 0)$ at the 5% level.

      i. (1 point) What is the test statistic for this test?
      ii. (1 point) What distribution does this test statistic follow under the null hypothesis (specify the degreres of freedom)?
      iii. (1 point) Do you reject the null hypothesis?
      iv. (1 point) Based on the result of the hypothesis test, which model would you choose?

   (d) (2 points) Plot the histogram of the residuals and the TA plot from the model 2 above. Do the normality and the constant variance assumptions appear to hold?

   (e) (2 points) What are the $\hat{\sigma}$ and the in-sample MSE of model 2 above?

   (f) (5 points) Consider model 2 summary output and testing null hypotheses:

      i. $H_0^1 : \beta_{HT2} = 0$
      ii. $H_0^2 : \beta_{WT2} = 0$

iii. $H_0^3 : \beta_{HT9} = 0$

iv. $H_0^4 : \beta_{WT9} = 0$

v. $H_0^5 : \beta_{ST9} = 0$

Which of these null hypothesis would be rejected if:

i. (1 point) You control the FWER at the 0.1 level using the Bonferroni correction?

ii. (2 points) You control the FWER at the 0.1 level using the Holm correction?

iii. (2 points) You control the FDR at the 0.1 level using the Benjamini-Hochberg procedure?

Make sure to explain each of the correction procedures you are applying above!

2. (12.5 + 4 Bonus points) (ALR 5.14, 5.15) Refer to the Berkeley Guidance study described in the previous problem. We will now use the data set `BGSall` in R package `alr4`.

(a) (1 point) Consider the regression of `HT18` (response) on `HT9` and `Sex` (predictors). Draw the scatterplot of `HT18` vs `HT9` using a different symbol or color for men and women (with a legend). Comment on the appropriate mean function for the data. Looking at the scatterplot, do you think including the factor `Sex` in the linear model is justified? Explain.

R Hint: Make sure that `Sex` is encoded as a factor!

(b) (6 points) Fit the linear model (call it fit.height) with `HT18` as a response and `HT9` and `Sex` as predictors.

i. (1 point) Consider the fitted regression line for women. What is the intercept of this fitted line?

ii. (1 point) What is the predicted height at 18 years of age for a 135cm tall 9-year-old girl (heights given in the data set are in centimeters - cm)?

iii. (2 points) What would be the predicted average change in height at age 18, for the same girl if in fact her height at age 9 was 137cm, but was measured wrongly as 135cm?

iv. (1 point) Based on this model, what is the 95% confidence interval for the differnece in height between men and women?

v. (1 point) Was your suspicion from part (a) correct? Test the hypothesis $H_0 : \beta_{\texttt{Sex}} = 0$ at the 5% level. Do you reject the null hypothesis? Explain.

(c) (5.5 + 4 Bonus points) Consider the following three models in addition to the above model in fit.height. (These are written in Wilkinson-Rogers notation, 1 indicates that the intercept is present in the model.)

1. `HT18 | 1 + HT2 + HT9 + Sex`
2. `HT18 | 1 + HT2 + HT9 + Sex + Sex:HT2 + Sex:HT9`
3. `HT18 | 1 + HT2 + HT9 + HT2:HT9 + Sex + Sex:HT2 + Sex:HT9 + Sex:HT2:HT9`

Call these fitted models fit.height2, fit.height3, and fit.height4 respectively.

i. (2 points) How many parameters are estimated in each of the proposed models (fit.height, fit.height2, fit.height3, fit.height4)?

ii. (1.5 points) What is the predicted average height of a girl who is 135cm tall at age 9 and 90cm tall at age 2 accoriding to fit.height2, fit.height3, and fit.height4 respectively?

iii. (Bonus 2 points) Consider a sequential procedure where you start from the smallest model (fit.height) and compare it with model fit.height2 using the anova test at the 10% level. If model fit.height is not rejected the procedure ends and you opt for model fit.height as the preferred model.
However, if model fit.height is rejected, you proceed to compare models fit.height2 and fit.height3 with an anova test at the 10% level. This continues until one of the null hypotheses is **not rejected** (and you opt for the smaller model and stop) or you reject all null hypotheses associated with the anova tests for comparing:

- fit.height and fit.height2,
- fit.height2 and fit.height3,
- fit.height3 and fit.height4,

in which case you opt for the largest model.
Which model would you choose based on this sequential procedure?

iv. (Bonus 2 points) Consider a converse sequential procedure from above. You start with largest model (from fit.height4) and compare it with model fit.height3 using the anova test at the 10% level. If the p-value is smaller than 10% you opt for the bigger model.
Otherwise, you compare model fit.height2 with model fit.height3 at the 10% level. This continues until you **reject the null hypothesis** (p-value smaller than 10%) at some point (in which case you opt for the larger model and stop), or until you do not manage to reject the null hypothesis for any of the comparisons
   - fit.height3 and fit.height4,
   - fit.height2 and fit.height3,
   - fit.height and fit.height2,

   in which case you opt for the smallest model.
   Which model would you choose based on this sequential procedure?

v. (2 points) What is the $\hat{\sigma}$ associated with models fit.height, fit.height2, fit.height3 and fit.height4? Based on the $\hat{\sigma}$, what is your preferred model?