Homework 3

STAT448 - Advanced Data Analysis

Due: Thursday, October 12 at 11pm

To complete this assignment, you will need to access the data sets in **Program_HW3_Data.sas** in the Homework 3 folder on the course website. The `psych` data set is for Exercise 1, and the `heart` data set is for Exercises 2, 3, and 4.

1. (4 parts) The psychology department at a hypothetical university has been accused of underpaying female faculty members. The data in the `psych` data set represent salary (in thousands of dollars) for all 22 professors in the department. This data is borrowed from Designing Experiments and Analyzing Data (2004). [1]

   (a) For the **salary** variable, create a cross-tabulation of the mean, standard deviation, and counts by **sex** (F, M) and **rank** (Assistant, Associate). Comment on any interesting features (e.g., any apparent differences between sexes and/or ranks, check if all cells seem to have the same variance, whether it is balanced data or not, etc.).

   (b) Fit a two-way ANOVA model including the interaction term. Comment on significance of the model and variation explained. What do the Type I and Type III sums of squares tell us about significance of effects? Is the interaction between **rank** and **sex** significant?

   (c) Refit the model without the interaction term. Comment on significance of the model and variation explained. Report and interpret the Type I and Type III tests of the main effects. Are the main effects of **rank** and **sex** significant?

   (d) Choose a final model based on your results from parts (b) and (c). For the chosen model, comment on the significance of the model and the individual terms in the model, and discuss the amount of variation explained by the model. State the differences in **salary** across different main effect groups and the interaction(s) between them, if any. Obtain model diagnostics to validate your assumptions.

   Hint: For interpretations of differences for the main effects, state quantitative interpretations of the significantly different groups (e.g. difference estimates and 95%

[1] S. Maxwell and H. Delaney (2004). Designing experiments and analyzing data: a model comparison perspective. New York, NY: Taylor and Francis Group

confidence intervals of the differences, and what the difference tells us about **salary**).

From Exercises 2-4, we would like to investigate the relationships between cholesterol, weight and/or blood pressure. The data set is built upon the `sashelp.heart` data set included in SAS and includes measurement variables for “alive” subjects. See the brief summary in the table below.

Variable Name	**Description**
weight	subject’s weight
systolic	top number in a blood pressure reading, indicating the blood pressure level when the heart contracts
diastolic	bottom number in a blood pressure reading, indicating the blood pressure level when the heart is at rest or between beats
cholesterol	measured cholesterol
bp_status	blood pressure status
weight_status	weight status

2. (2 parts) The medical director at your company wants to know if weight alone can predict cholesterol outcome. Consider modeling cholesterol as a function of weight.

   (a) Fit a linear regression model for **cholesterol** as a function of **weight**. If any points are unduly influential, note those points, then remove them and refit the model. Consider Cook’s distance cut off to be 1.

   (b) Comment on the quality of the final model, significance of the parameters, variation explained by the model, and any remaining issues noted in the diagnostics. What does this model tell us about the relationship between cholesterol and weight? Explain to the medical director whether this is a good model based on variation explained and diagnostics.

3. (2 parts) The medical director wants to know if blood pressures and weight can better predict cholesterol outcome. Consider modeling **cholesterol** as a function of **diastolic, systolic**, and **weight**.

   (a) Fit a linear regression model for **cholesterol** as a function of **diastolic, systolic**, and **weight**. Analyze the diagnostics, describe any issues that need to be remedied, make any necessary adjustments for undue influence, and re-fit the model, if necessary. For Cook’s distances, do not leave any points in the final model that have Cook’s distance greater than 1 in the plot.

(b) Comment on how much variation in **cholesterol** is described by the model. Comment on the relationship between cholesterol and the remaining significant predictor(s). Check multicollinearity issue among predictors. Explain to the medical director whether this is a good model based on variation explained and diagnostics.

4. (2 parts) Now consider forward model selection for the **cholesterol** model. Keep predictors that add at least 1% of additional explained variation to the model.

   (a) Perform forward model selection and address any issues of undue influence like in part (a) of Exercise 2.

   (b) Interpret the final model, comment on the variation in **cholesterol** explained, and usefulness of the model like in part (b) of Exercise 2 and 3. Compare the variations explained by the models from Exercise 2 and 3.