Homework 2

Due: Saturday, Sep 30, at 11:59 pm

Data: **heartbpchol.csv** for Exercise1, **bupa.csv** for Exercise 2, **psych.csv** for Exercise3, **cars_new.csv** for Exercise 4
Use the significance level of .05

Exercise 1: Analysis of Variance

The **heartbpchol.csv** data set contains continuous cholesterol (**Cholesterol**) and blood pressure status (**BP_Status**) (category: High/ Normal/ Optimal) for alive patients.

For the **heartbpchol** data set, consider a one-way ANOVA model to identify differences between group cholesterol means. The normality assumption is reasonable, so you can proceed without testing normality.

a) Perform a one-way ANOVA for **Cholesterol** with **BP_Status** as the categorical predictor. Comment on statistical significance of **BP_Status**, the amount of variation described by the model, and whether or not the equal variance assumption can be trusted.

b) Comment on any significantly different cholesterol means as determined by the post-hoc test comparing all pairwise differences. Specifically explain what that tells us about differences in cholesterol levels across blood pressure status groups, like which group has the highest or lowest mean values of **Cholesterol**.

Exercise 2: Analysis of Variance

For this problem use the **bupa.csv** data set. Check UCI Machine Learning Repository for more information (http://archive.ics.uci.edu/ml/datasets/Liver+Disorders). The mean corpuscular volume and alkaline phosphatase are blood tests thought to be sensitive to liver disorder related to excessive alcohol consumption. We assume that normality and independence assumptions are valid.

Variable Name	Description
mcv	mean corpuscular volume
alkphos	alkaline phosphatase
drinkgroup	categorization of the half-pint equivalents of alcoholic beverages drunk per day: **group 1** : less than 1 drink. **group 2** : at least 1 but fewer than 3 drinks. **group 3** : at least 3 but fewer than 6 drinks. **group 4** : at 6 but fewer than 9 drinks. **group 5** : 9 or more drinks.

a) Perform a one-way ANOVA for **mcv** as a function of **drinkgroup**. Comment on significance of the **drinkgroup**, the amount of variation described by the model, and whether or not the equal variance assumption can be trusted.

b) Perform a one-way ANOVA for **alkphos** as a function of **drinkgroup**. Comment on statistical significance of the **drinkgroup**, the amount of variation described by the model, and whether or not the equal variance assumption can be trusted.

c) Perform post-hoc tests for models in a) and b). Comment on any similarities or differences you observe from their results.

Exercise 3:

The psychology department at a hypothetical university has been accused of underpaying female faculty members. The data represent salary (in thousands of dollars) for all 22 professors in the department. This problem is from Maxwell and Delaney (2004).

a) Fit a two-way ANOVA model including **sex** (F, M) and **rank** (Assistant, Associate) the interaction term. What do the Type 1 and Type 3 sums of squares tell us about significance of effects? Is the interaction between **sex** and **rank** significant? Also comment on the variation explained by the model.

b) Refit the model without the interaction term. Comment on the significance of effects and variation explained. Report and interpret the Type 1 and Type 3 tests of the main effects. Are the main effects of **rank** and **sex** significant?

c) Obtain model diagnostics to validate your Normality assumptions.

d) Choose a final model based on your results from parts (a) and (b). Comment on any significant group differences through the post-hoc test. State the differences in **salary** across different main effect groups and interaction (if included) between them.

Hint: For interpretations of differences for the main effects, state quantitative interpretations of the significantly different groups (e.g. estimated differences between groups and what the difference tells us about **salary**). For interaction term, identify significant interactions, but no need to interpret it quantitatively.

Exercise 4:

Use the **cars_new.csv**. See HW1 for detailed information of variables.

a) Start with a three-way main effects ANOVA and choose the best main effects ANOVA model for **mpg_highway** as a function of **cylinders**, **origin**, and **type** for the cars in this set. Comment on which terms should be kept in a model for **mpg_highway** and why based on Type 3 SS. For the model with just predictors you decide to keep, comment on the significant effects in the model and comment on how much variation in highway fuel efficiency the model describes.

b) Starting with main effects chosen in part (a), find your best ANOVA model by adding in any additional interaction terms that will significantly improve the model. For your final model, comment on the significant effects and variation explained by the model.

c) Comment on any significant group differences through the post-hoc test. What does this tell us about fuel efficiency differences across **cylinders**, **origin**, or **type** groups? See Hint in Exercise 3.