STA 6443: Algorithms I

Homework 1

Ashley Cortez

October 2, 2021

**Exercise 1: Analysis of Variance**

**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.**

H0: mu(Normal) = mu(High) = mu(Optimal)

Ha: At least ONE of the group means is significantly different from the others in the population

The data summary has a small p-value of 0.00137, which indicates a significant effect, there is at least one mean that differs. The LeveneTest shows a large p-value of 0.5201, meaning the data has equal variance. 2.42% of the variation of cholesterol can by explained by the BP\_Status, since the variation is low, I do not think the equal variance assumption can be trusted.

Chart, diagram

Description automatically generated

The plots match the normaility assumption, specifically the Normal QQ plot.

**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.**

Two groups have a different mean of y

Normal-High:

H0: mu(Normal) = mu(High)

Ha: mu(Normal) ≠ mu (High)

With a small p-value of 0.0159 we reject the null, concluding that mu(Normal) < mu(High)

Optimal-High:

H0: mu(Optimal) = mu(High)

Ha: mu(Optimal) ≠ mu(High)

With a small p-value of 0.0089, we reject the null, concluding that mu(Optimal) < mu(High)

Optimal-Normal:

H0: mu(Optimal) = mu(Normal)

Ha: mu(Optimal) ≠ mu(Normal)

With a large p-value 0.4958 we fail to reject the null, concluding that optimal and normal have the same mean.

High BP status has a different effect on cholesterol levels (mu(High) > mu(Normal) & mu(Optimal)).

**Exercise 2: Analysis of Variance**

**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.**

When we run the LeveneTest we get a large p-value of 0.8744, indicating that we fail to reject the null and we keep the ANOVA model. The summary gives us a small p-value of nearly 0 which indicates there is a significant effect and at least one of the means differs. 11% of the variation of mcv can be explained by the drink group, since the variation is low, I do not think the variation can be trusted.

Graphical user interface, diagram

Description automatically generated

The plots match the normaility assumption, specifically the Normal QQ plot.

**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.**

When we run the LeveneTest we get a large p-value of 0.5201, indicating that we fail to reject the null and we keep the ANOVA model. The summary gives us a small p-value of 0.00495, which indicates there is a significant effect and at least one of the means differs. 4% of the variation of mcv can be explained by the drink group, and since the variation is low, I do not think the variation can be trusted.

Graphical user interface

Description automatically generated with low confidence

The plots match the normaility assumption, specifically the Normal QQ plot.

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

Mcv:

2-1 pvalue 0.5410 > 0.05 we fail to reject the null and conclude that mu(2) = mu(1).

3-1 pvalue 0.6495 > 0.05 we fail to reject the null and conclude that mu(3) = mu(1).

4-1 pvalue 1.9e-06 < 0.05 we reject the null and conclude mu(4) > mu(1)

5-1 pvalue 0.0081 < 0.05 we reject the null and conclude mu(5) > mu(1)

3-2 pvalue 0.9966 > 0.05 we fail to reject the null and conclude that mu(3) = mu(2).

4-2 pvalue 0.0380 < 0.05 we reject the null and conclude mu(4) > mu(2)

5-2 pvalue 0.2646 > 0.05 we fail to reject the null and conclude that mu(5) = mu(2).

4-3 pvalue 0.0025 < 0.05 we reject the null and conclude mu(4) > mu(3)

5-3 pvalue 0.1151 > 0.05 we fail to reject the null and conclude that mu(5) = mu(3).

5-4 pvalue 1.0000 > 0.05 we fail to reject the null and conclude that mu(5) = mu(4)

Group 4 has a significant difference on MCV from group 1, 2, and 3, specifically

Group 4 > Group 1

Group 4 > Group 2

Group 4 > Group 3

Group 5 has a significant difference effect on MCV from Group 1, specifically Group 5 > Group 1

Alkphos:

2-1 pvalue 0.9419 > 0.05 we fail to reject the null and conclude that mu(2) = mu(1).

3-1 pvalue 0.6389 > 0.05 we fail to reject the null and conclude that mu(3) = mu(1).

4-1 pvalue 0.9965 > 0.05 we fail to reject the null and conclude that mu(4) = mu(1).

5-1 pvalue 0.0734 > 0.05 we fail to reject the null and conclude that mu(5) = mu(1).

3-2 pvalue 0.9953 > 0.05 we fail to reject the null and conclude that mu(3) = mu(2).

4-2 pvalue 0.9952 > 0.05 we fail to reject the null and conclude that mu(4) = mu(2).

5-2 pvalue 0.0329 < 0.05 we reject the null and conclude mu(5) > mu(2)

4-3 pvalue 0.9117 > 0.05 we fail to reject the null and conclude that mu(4) = mu(3).

5-3 pvalue 0.0069 < 0.05 we reject the null and conclude mu(5) > mu(3)

5-4 pvalue 0.0578 > 0.05 we fail to reject the null and conclude that mu(5) = mu(4).

Group 5 has a significant difference on alkaline phosphatase from Group 2 and 3, specifically

Group 5 > Group 2

Group 5 > Group 3

Similarties: In both models Group 5 had a significant difference,i.e larger means then the group they were being comare to.

Differences: Group 4 had a significant difference, i.e larger mean then some of the groups it was being compared to with MCV but not alkaline phosphatase. Group 1 had a significant difference,i.e smaller mean then group 5 with MCV but not alkaline phosphatase. Group 2 and 3 has a significant difference,i.e smaller means then group 4 in the MCV model whereas it was smaller than group 5 in the alkaline phosphatase model.

**Exercise 3:**

**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.**

**Type 1:**

H0: Sex has no effect on salary

Ha: Sex has an effect on salary

With a pvalue 0.000637 < 0.05 we reject the null, so we can conclude that sex does have an effect on salary.

H0: Rank has no effect on salary

Ha: Rank has an effect on salary

With a pvalue 0.000417 < 0.05 we reject the null, so we can conclude that rank does have an effect on salary.

H0: no interaction between sex and rank

Ha: exist an interaction between sex and rank

With a pvalue 0.795101 > 0.05 we fail to reject the null, so we can conclude that there is no interaction between sex and rank.

**Type 3:**

H0: Sex has no effect on salary

Ha: Sex has an effect on salary

With a pvalue 0.09671 > 0.05 we fail to reject the null, so we can conclude that sex does not have an effect on salary.

H0: Rank has no effect on salary

Ha: Rank has an effect on salary

With a pvalue 0.01240 < 0.05 we reject the null, so we can conclude that rank does have an effect on salary.

H0: no interaction between sex and rank

Ha: exist an interaction between sex and rank

With a pvalue 0.795101 > 0.05 we fail to reject the null, so we can conclude that there is no interaction between sex and rank.

66% of the variation of salaries can be explained by the model.

When we ran the LeveneTest we had a large p-value, therefore, we fail to reject the null concluding that the groups have the same variance.

**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?**

**Type 1**

H0: Sex has no effect on salary

Ha: Sex has an effect on salary

With a pvalue 0.000454 < 0.05 we reject the null, so we can conclude that sex does have an effect on salary.

H0: Rank has no effect on salary

Ha: Rank has an effect on salary

With a pvalue 0.000291 < 0.05 we reject the null, so we can conclude that rank does have an effect on salary.

**Type 3:**

Sex has no effect on salary

Ha: Sex has an effect on salary

With a pvalue 0.0092618 < 0.05 we reject the null, so we can conclude that sex does have an effect on salary.

H0: Rank has no effect on salary

Ha: Rank has an effect on salary

With a pvalue 0.0002912 < 0.05 we reject the null, so we can conclude that rank does have an effect on salary.

Both sex and rank have an effect on salaries, with males and associates having the larger mean than females and assistants.

66% of variation in salaries can be explained by the model.

**c) Obtain model diagnostics to validate your Normality assumptions.**

Diagram, schematic

Description automatically generated

The plots match the normaility assumption, specifically the Normal QQ plot.

**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**

The final model I chose based on my results is from part (a).

Sex effect: Ho: mu(M) = mu(F) Ho: mu(M) ≠ mu(F) With a pvalue of 0.0065 < 0.05 we reject the null, sex has an effect on salaries, specifically mu(M) > mu(F)

Rank effect: Ho: mu(Assoc) = mu(Assist) Ho: mu(Assoc) ≠ mu(Assist) With a pvalue of 0.0060 < 0.05 we reject the null, rank has an effect on salaries, specifically mu(Assoc) > mu(Assist).

Interaction effect:

M:Assist-F:Assist with a pvalue of 0.4055 > 0.05 we fail to reject the null mu(M:Assist) = mu(F:Assist)

F:Assoc-F:Assist with a pvalue of 0.0861 > 0.05 we fail to reject the null mu(F:Assoc) = mu(F:Assist)

M:Assoc-F:Assist with a pvalue of 0.0002 < 0.05 we reject the null mu(M:Assoc) > mu(F:Assist)

F:Assoc-M:Assist with a pvalue of 0.8304 > 0.05 we fail to reject the null mu(F:Assoc) = mu(M:Assist)

M:Assoc-M:Assist with a pvalue of 0.0323 < 0.05 we reject the null mu(M:Assoc) > mu(M:Assist)

M:Assoc-F:Assoc with a pvalue of 0.2115 > 0.05 we fail to reject the null mu(M:Assoc) = mu(F:Assoc)

Sex and Rank both have an effect on salary, with males having a higher mean of salaries then females and associates having a higher mean of salaries then assistants. Male Associates have a significant effect on salaries, specifically Male Associates > Male Assistants and Male Associates > Female Assistants. With an r-squared value of 0.6647566, 66.48% of variation in salaries can be explained by the model.

**Exercise 4:**

**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.**

Cylinder effect: H0: Cylinder has no effect on mpg\_highway

Ha: Cylinder has an effect on mpg\_highway

With a pvalue of 2e-16 < 0.05 we reject the null, so we can conclude that cylinders does have an affect on mpg\_highway.

Origin affect: H0: Origin has no effect on mpg\_highway Ha: Origin has an effect on mpg\_highway With a pvalue of 0.77948 > 0.05 we fail to reject the null, so we can conclude that origin does not have an affect on mpg\_highway.

Type effect: H0: Type has no effect on mpg\_highway Ha: Type has an effect on mpg\_highway With a pvalue of 0.00175 < 0.05 we reject the null, so we can conclude that type does have an affect on mpg\_highway and should be removed.

46% of variation in highway fuel efficiency can be explained by the model.

**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.**

Cylinder effect:

H0: Cylinder has no effect on mpg\_highway

Ha: Cylinder has an effect on mpg\_highway

With a pvalue of 2e-16 < 0.05 we reject the null, so we can conclude that cylinders does have an affect on highway fuel efficiency.

Type effect:

H0: Type has no effect on mpg\_highway

Ha: Type has an effect on mpg\_highway

With a pvalue of 0.0006601 < 0.05 we reject the null, so we can conclude that type does have an affect on highway fuel efficiency.

Interaction effect:

H0: no interaction between cylinder and type

Ha: exist an interaction between cylinder and type

With a pvalue 0.0046958 < 0.05 we reject the null, so we can conclude that there is interaction between cylinder and type.

The pvalues are the same for Type 1 test and Type 3 so the conclusion would be identical

48% of variation in highway fuel efficiency can be explained by the model.

H0: all groups have the same variances

Ha: at least one group has different variance

With a pvalue 0.002465 < 0.05, we reject the null concluding that at least one groups has a different variance, which implies heteroscedasticity. From notes: For multi-way ANOVA especially with interaction terms, it is hard to pass equal variance assumption as we do not have large enough samples in each cell. Thus, we will not seriously consider Levene’s test result and mainly focus on Normality check.

Graphical user interface, diagram, schematic

Description automatically generated

The plots match the normality assumption, specifically the Normal QQ plot.

**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**

Cylinder effect:

H0: mu(6) = mu(4)

Ha: mu(6) ≠ mu(4)

With a pvalue of <2e-16 < 0.05 we reject the null, and conclude that 4 cylinders and 6 cylinders do not have equal highway fuel efficiency.

Type effect:

H0: mu(Sports) = mu(Sedan)

Ha: mu(Sports) ≠ mu(Sedan)

With a pvalue of 0.0117 < 0.05 we reject the null, and conclude that sports and sedans do not have equal highway fuel efficiency.

Interaction effect:

6:Sedan-4:Sedan with a pvalue of 2e-16 < 0.05 we reject the null, mu(6:Sedan) < mu(4:Sedan) 4:Sports-4:Sedan with a pvalue of 0.00036 < 0.05 we reject the null, mu(4:Sports) < mu(4:Sedan)

6:Sports-4:Sedan with a pvalue of 2.7e-06 < 0.05 we reject the null, mu(6:Sports) < mu(4:Sedan) 4:Sports-6:Sedan with a pvalue of 0.88732 > 0.05 we fail to reject the null, mu(4:Sports) = mu(6:Sedan)

6:Sports-6:Sedan with a pvalue of 0.98763 > 0.05 we fail to reject the null, mu(6:Sports) = mu(6:Sedan)

6:Sports-4:Sports with a pvalue of 0.86373 > 0.05 we fail to reject the null, mu(6:Sports) = mu(4:Sedan)

Both cylinders and type of vehicle have a significant effect on a cars highway fuel efficiency, specifically -A 4 cylinder sedan has a higher fuel efficiency then a 6 cylinder sedan. -A 4 cylinder sedan has a higher fuel efficiency then a 4 cylinder sports car. -A 4-cylinder sedan has a higher fuel efficiency then a 6 cylinder sports car.

A 4 cylinder sedan has the best fuel efficiency then the other cars in the dataset.