**J. Jedediah Smith**

**BIFX 503: Statistics for Bioinformatics**

**Homework Set #2**

**Due: October 1, 2021**

*Instructions: Use R to complete this assignment. Assignment is to be submitted via Blackboard.*

The dataset **cabbages** (from the R MASS package) contains data from a cabbage field trial. This experiment had a factorial design, evaluating the effects of cultivar and planting date on the vitamin C content of cabbage.

1. Generate an interaction plot to visualize how vitamin C content varies by cultivar and planting date. Do the graphs suggest that interaction might be present? Why or why not?

The graph suggests that there is no interaction, as the lines on it are almost parallel. Clears signs of interaction, such as the lines crossing, are notably absent.

Chart, line chart

Description automatically generated

1. Run an ANOVA model that consists of cultivar, planting date, and their interaction. Is the interaction term significant? If so, what does this mean? If the interaction is significant, perform simple main effects analysis to understand the nature of the interaction.

The P-Value is 0.218627. It is not significant because it is higher than 0.05. This indicates that the Cultivar-Vitamin C and Date-Vitamin C interactions are independent. As a result, we do not perform simple main effects analysis.

Text

Description automatically generated

1. The weight of the cabbage is also included in the dataset. It is possible that weight influences vitamin C content. Assess the relationship between these two variables using a scatterplot and correlation. What are your conclusions?

Cabbage weight and Vitamin C content appear to be correlated. The scatterplot shows that they have a negative linear association. And the correlation value of -0.659892 confirms it.

Chart, scatter chart

Description automatically generated



1. Re-run the ANOVA, this time adding weight as an additional covariate. (Note: you are not interested in the interaction between weight and any of the other variables.) Do your conclusions change after taking weight into account? Why or why not?

This time the P-Value was 0.664. Now the P-Value is larger than before, so it is still above the 0.05 significance threshold. Because of this, our conclusions remain unchanged.

Text

Description automatically generated

1. Explain how multiple comparisons inflate alpha. Choose a multiple comparisons procedure (*e.g.,* Bonferroni or FDR) and explain how this procedure controls type I error.

The more tests that are performed, the more chances you have to get a significant result purely by chance. As a result, multiple comparison procedures like Bonferroni have Family-Wise Error Rate (FWER) built in to control alpha reduce the chance of Type I errors. With Bonferroni, the significance level is: alpha / number of tests.

The dataset **Cars93** (from the R MASS package) contains data on 93 new cars for the model year 1993. The variables include price, fuel efficiency, vehicle size, and various features. You are going to use this data and linear regression to determine the correlates of car price. There are a lot of variables in this dataset so we will narrow our focus to 3 continuous and 3 categorical variables: MPG, Horsepower, Passengers, AirBags, DriveTrain, and Origin.

1. Calculate a new variable, MPG, which is the average of MPG.city and MPG.highway. Make three scatter plots and three box plots to see how each of the six independent variables (MPG, Horsepower, Passengers, AirBags, DriveTrain, and Origin) relate to the dependent variable (Price). Based on the graphs, which variables seem to be associated with Price?

Graphs with linear X-Y trends are indicative of association. There appears to be a negative linear association between MPG and Price, as well as AirBags and Price. There appears to be a positive one between Horsepower and Price, as well DriveTrain and Price. There does not appear to be a strong linear association between Passengers and Price or Origin and Price.



Diagram

Description automatically generated

1. Fit a series of bivariate linear regression models having Price as the dependent variable and each of these independent variables: MPG, Horsepower, Passengers, AirBags, DriveTrain, and Origin. Note these are simple linear regression models: each model has 1 independent variable. Report the results in terms of regression coefficients and 95% confidence intervals. Which variables are significant predictors of Price?

MPG: The regression coefficient is -1.0491. The 95% confidence interval is -1.35093 to

-0.747304. Since the p-value was 6.61e-10, MPG is a significant predictor of Price.

Horsepower: The regression coefficient is 0.1454. The 95% confidence interval is 0.1217377 to 0.1690048. Since the p-value was 2e-16 Horsepower is a significant predictor of Price.

Passengers: The regression coefficient is 0.5379. The 95% confidence interval is -1.394742 to 2.470598. Since the p-value was 0.58170, Passengers is not a significant predictor of Price.

AirBags: Used “Driver & Passenger” as the reference. For “Driver Only,” the regression coefficient was -7.145 and the 95% confidence interval was -11.83075 to -2.46023. For “None,” the coefficient was -15.195 and the interval was -20.04573 to -10.344712. Since the p-values were 0.00319 and 1.51e-08 respectively, AirBags is a significant predictor of Price.

DriveTrain: Used “4WD” as the reference. For “Front,” the regression coefficient was -0.09418 and the 95% confidence interval was -5.974898 to 5.78654. For “Rear,” the coefficient was 11.32000 and the interval was 4.327232 to 18.31277. Since the p-values were 0.97469 and 0.00181 respectively, DriveTrain is a significant predictor of Price as long as you are comparing “Rear” to “4WD” and “Front.”

Origin: Used “USA” as the reference. For “Non-USA,” the regression coefficient was 1.936 and the 95% confidence interval was -2.04682 to 5.918765. Since the p-value was 0.337, Origin is not a significant predictor of Price.

Text, letter

Description automatically generated

A picture containing text

Description automatically generated

Text

Description automatically generated

(Included the output just for MPG since there would otherwise be like 18 screenshots)

1. Fit a multiple linear regression model of predictors of Price:
   1. In which order should IVs enter the model?

They should enter in order of significance with the most significant first. In our case, this would be: Horsepower, MPH, AirBags, DriveTrain. Origin and Passengers should not be included because they are not significant.

* 1. Are any IVs removed? If so, why?

Remove MPH and DriveTrain. Partial F tests were run to see which variables improved model fit when used together with Horsepower. Only AirBags had a significant p-value.

* 1. What is the final model? Report the results in terms of regression coefficients and 95% confidence intervals.

The final model uses Horsepower and AirBags. For Horsepower, the regression coefficient was 0.1238 and the 95% confidence interval was 0.09830454 to 0.1492003. For AirBags, using “Driver & Passenger” as reference, the coefficients were -3.1518 for “Driver Only” and -6.6267 for “None” while the intervals were -6.54459146 to 0.2410308 for “Driver Only” and -10.46329350 to -2.7901664 for “None.”



* 1. Interpret the meaning of each regression coefficient in your final model. What does the model tell you about predictors of automobile price?

For each 1 hp increase in Horsepower, we except to see Price rise by $123.80. Cars with “Driver Only” AirBags are about $3,151.80 cheaper than those with “Driver & Passenger” AirBags. Cars with “None” AirBags are about $6,626.70 cheaper. Together Horsepower and AirBags explain 66% of the variability in Price.

Note: Building multiple regression models is a very subjective process! There is no single “correct” answer. You will get full credit for these problems if you follow the correct procedure and explain what you did at each step and your rationale.