***Jordon Tourville***

This exam is open book and open internet but you are NOT allowed to work with anyone else or ask anyone other than Meha or Oscar any questions about the exam. It is due at noon on Sunday, April 23.

Please answer the following questions by analyzing the associated datasets. For all tests, please:

* check whether the data meet the requirements/assumptions of the test you plan to run
* complete any transforms needed to make the data meet the required assumptions
* run the test
* interpret the results
* check model fit in the case of linear regressions and/or glms
* if you have the option between running a linear model with a transformed y variable or a glm, choose the linear model with a transformed y variable. only run a glm when you have to.

Provide all answers in R or R markdown (similar to the take home quiz 4). Use the following scripts to load the datasets. The dataset to be used for each question is provided in bold at the end of the question.

Dataset Please use the following scripts to load in the data from GitHub

flying = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/flying.csv",header=TRUE, sep=",")

college = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/college.csv",header=TRUE, sep=",")

happy = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/happy.csv",header=TRUE, sep=",")

cancer = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/cancer.csv",header=TRUE, sep=",")

Red=code

1. Is there a significant association between gender (gender) and whether people think it’s rude to bring an unruly child on the plane (unruly\_child)? **Flying**

***Use chi-square (both variables are categorical)***

***Checking assumptions-assume independent observations and no structural 0's.***

***chisq.test(flying$gender, flying$unruly\_child)#p=0.001193***

***ANSWER: There is a significant association between gender and whether people think it's rude to bring an unruly child onto a plane.***

1. Is there a significant difference in tuition (tuition) by type of institution (type)? **College**

***Use 2 sample t-test (only 2 types of school)***

***checking assumptions:***

***data are continuous, assume random sampling and independent observations***

***hist(college$tuition)***

***shapiro.test(college$tuition)#p<<0.0001-data not normally distributed***

***qqnorm(college$tuition)***

***qqline(college$tuition, col = "red")#histogram, shapiro.test, and qqplot show data is not normally distributed, but sample size is large (>500) so we assume normality under CLT.***

***public1 = subset(college, type=="Public")***

***private1 = subset(college, type=="Private nonprofit")#subsetting data***

***var.test(public1$tuition, private1$tuition)#variance test shows unequal variances so we use Welch's 2 sample t-test.***

***t.test(public1$tuition, private1$tuition)#p<<<0.001 and mean of public school tuition is far lower than private.***

***ANSWER: There is a significant difference in tuition by type of institution, and tuition is significantly lower for public schools than private nonprofit schools.***

1. Is there a significant difference in happiness (Hscore) by region (Region)? **Happy**

***Use one-way anova***

***Checking assumptions:***

***data assumed to be independent, sample sizes do not need to be equal.***

***hist(happy$Hscore)***

***shapiro.test(happy$Hscore)#p=0.0125-not normally distributed***

***qqnorm(happy$Hscore)***

***qqline(happy$Hscore, col = "red")#histogram, shapiro.test, and qqplot show data is not normally distributed but is close in qqplot, but sample size is large (157) so we assume normality under CLT.***

***install.packages("lawstat")***

***library(lawstat)***

***levene.test(happy$Hscore, happy$Region, location = "mean")#p=0.342, variances are equal.***

***aov1 = aov(happy$Hscore ~ happy$Region, data = happy)#p<<0.001 shows that there is a difference in Hscore by Region.***

***summary(aov1)***

***aov2 = lm(happy$Hscore ~ happy$Region, data = happy)#comparisons of region to Africa.***

***summary(aov2)***

***aov3 = lm(happy$Hscore-1 ~ happy$Region, data = happy)#comparisons of region to region mean***

***summary(aov3)***

***install.packages("DTK")***

***library(DTK)***

***Happy = DTK.test(happy$Hscore, happy$Region)#Tukey-kramer Post-hoc test with unequal sample sizes.***

***DTK.plot(Happy)***

***Happy***

***ANSWER: There is a significant difference in happiness by region. The comparisons that are significantly different from each other are Europe to Africa, Americas to Asia/Australia, Americas to Africa, and Asia to Africa.***

1. What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model. **Happy**

***Use multiple linear regression***

***checking model assumptions:***

***hist(happy$Corruption)#not normally distributed***

***shapiro.test(happy$Corruption)#p<<0.001, not normally distributed***

***log\_corrupt = log(happy$Corruption)#log transform y***

***hist(log\_corrupt)***

***qqnorm(log\_corrupt)***

***qqline(log\_corrupt, col = "red")***

***shapiro.test(log\_corrupt)#transform results in normality, p=0.0835***

***pairs(happy[, c("GDP", "Freedom", "Generosity")])#no evidence of multicollinearity.***

***mod1 = lm(log\_corrupt ~ GDP + Freedom + Generosity, data = happy)#run model***

***library(lmtest)***

***bptest(mod1)#errors are homoscedastic, p=0.06878***

***dwtest(mod1)#errors are independent, p=0.185***

***summary(mod1)#summary of model, model fit: Adjusted R2=0.1952, explains almost 20% of variance in Corruption (y).***

***ANSWER: Out of GDP, Freedom, and Generosity, only Freedom was significantly associated with Corruption and Corruption increased as Freedom increased (interestingly enough).***

1. Choose one of the continuous independent variables that was significant in the model for Question 4 and interact it with region (Region) to predict corruption (Corruption). This model should only include one continuous independent variable and its interaction with region. Does the influence of your continuous variable on corruption vary by region? **Happy**

***Use ANCOVA lm framework***

***Checking assumptions:***

***We already tested for normality in y-use log transform***

***mod2 = lm(log\_corrupt ~ Freedom \* Region, data = happy)#run ANCOVA model***

***bptest(mod2)#p=0.2501, homoscedaststic errors***

***dwtest(mod2)#p=0.7187, independence of errors***

***summary(mod2)#There is a difference of the influence of freedom on Corruption by region.***

***Adj R2=0.3053***

***ANSWER: The influence of freedom on corruption does vary by region. The only difference is in Europe where corruption increases as freedom increases compared to Africa/Middle East, p=0.000494.***

1. Which factors are significantly associated with whether a breast cancer tumor is malignant or not? Choose three continuous independent variables to include in your model. **Cancer**

***Use a binomial logistic regression glm***

***Checking model assumptions-assume linear relationship, that y falls in binomial distribution (0, 1 data), does not need to assume homoscedasticity.***

***pairs(cancer[, c("texture\_mean", "area\_mean", "smoothness\_mean", "radius\_mean", "perimeter\_mean")])#since area is directly related to radius and perimeter, we can only use area, smoothness, and texture as variables to avoid mulitcollinearity.***

***mod3 = glm(malignant ~ texture\_mean + area\_mean + smoothness\_mean, data = cancer, family = binomial(link="logit"))***

***summary(mod3)#All independent variables are significant.***

***mod4 = glm(malignant ~ texture\_mean + smoothness\_mean, data = cancer, family = binomial(link="logit"))***

***summary(mod4)***

***mod5 = glm(malignant ~ smoothness\_mean, data = cancer, family = binomial(link="logit"))***

***summary(mod5)***

***AIC(mod3, mod4, mod5)#checking main model with other models revealed main model has better fit (far lower AIC).***

***ANSWER: mean texture, mean smoothness, and mean area are all significantly associated with whether or not a tumor is malignant.***

1. BONUS/EXTRA CREDIT: Which independent variables are the most important in explaining whether a breast cancer tumor is malignant or not? Use the same 3 continuous independent variables you chose for question 6. **cancer.**

***Use question 6 model-which variable explains most variation in y?***

***install.packages("caret")***

***library(caret)***

***varImp(mod3, scale = FALSE)#returns value of importance for each variable.***

***ANSWER: area\_mean has highest value of importance as returned by the VarImp function (8.9), so we can conclude that this has a greater effect on malignancy.***