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=",")

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)? If yes, which gender tends to think that bringing an unruly child is more rude? **Flying**

Contingency table:

# No Somewhat Very

# Female 91 193 158

# Male 56 155 190

# Pearson's Chi-squared test

# data: tbl1

# X-squared = 13.463, df = 2, p-value = 0.001193

##The chi-square test tells us that there is a significant association between gender

## and whether they think it is rude to bring an unruly child on a plane.

##Inspecting the contingency table, we see that males answered more frequently "very" and "somewhat", making it that males are significantly more likely to think unruly children are rude.

##The contingency table tells us that the assumptions for a chi-square test are met. There are no structural zeros and no cell has fewer than 5 responses

1. Is there a significant difference in tuition (tuition) by type of institution (type)? If yes, which type has a higher tuition? **college**

public= subset(college, type=='Public')

private= subset(college, type=='Private nonprofit')

t.test(public[,"tuition"], private[,"tuition"], paired=FALSE)

# Welch Two Sample t-test

# data: public[, "tuition"] and private[, "tuition"]

# t = -22.79, df = 1397.9, p-value < 2.2e-16

# alternative hypothesis: true difference in means is not equal to 0

# 95 percent confidence interval:

# -10258.878 -8632.749

# sample estimates:

# mean of x mean of y

# 18855.88 28301.69

##The Welch’s t-test shows us that the p-value is significant, and the two data subsets are significantly different; the tuition is significantly different depending on the type of institution. Looking at the sample estimates, we can see that the public school tuition is about 18000, and the private tuition is 28000; Private institutions have higher tuition.

shapiro.test(public[,"tuition"])

# Shapiro-Wilk normality test

# data: public[, "tuition"]

# W = 0.96319, p-value = 2.58e-10

shapiro.test(private[,"tuition"])

# Shapiro-Wilk normality test

# data: private[, "tuition"]

# W = 0.98776, p-value = 1.147e-06

## Both shapiro tests tell us that the subset data are not normally distributed, but since there are 500 or 800 samples in each, we can use the data anyway. However, looking at the histogram of each, they seem close enough to a normal distribution.

var.test(public[,"tuition"], private[,"tuition"])

# F test to compare two variances

# data: public[, "tuition"] and private[, "tuition"]

# F = 0.43469, num df = 534, denom df = 871, p-value < 2.2e-16

# alternative hypothesis: true ratio of variances is not equal to 1

# 95 percent confidence interval:

# 0.3737381 0.5070016

# sample estimates:

# ratio of variances

# 0.4346851

##The var.test tells us that the variances are significantly different. To address this, we use the Welch’s t-test.

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

boxplot(Hscore~Region, data=happy)

happy.region=aov(Hscore~Region, data=happy)

summary(happy.region)

# Df Sum Sq Mean Sq F value Pr(>F)

# Region 3 64.37 21.456 23.62 1.28e-12 \*\*\*

# Residuals 153 138.96 0.908

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

##Based on the output, we can see that there is a significant difference in happiness by region.

## Assumptions: Normality is satisfied because there are 157 observations.

## We are assuming each country is an independent observation

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

pairs(happy[, c("Corruption","GDP", "Freedom", "Generosity")])

cor(happy[, c("Corruption","GDP", "Freedom", "Generosity")])

#Corruption and freedom have coefficient of .5; correlated, but ok

modelq4=lm(Corruption~ GDP+Freedom+Generosity, data=happy)

summary(modelq4)

# Call:

# lm(formula = Corruption ~ GDP + Freedom + Generosity, data = happy)

# Residuals:

# Min 1Q Median 3Q Max

# -0.20056 -0.06895 -0.00811 0.05827 0.34540

# Coefficients:

# Estimate Std. Error t value Pr(>|t|)

# (Intercept) -0.03623 0.02536 -1.429 0.1551

# GDP 0.04313 0.01998 2.159 0.0324 \*

# Freedom 0.29137 0.06074 4.797 3.79e-06 \*\*\*

# Generosity 0.14270 0.06161 2.316 0.0219 \*

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 0.09438 on 153 degrees of freedom

# Multiple R-squared: 0.2914, Adjusted R-squared: 0.2775

# F-statistic: 20.98 on 3 and 153 DF, p-value: 1.95e-11

plot(Corruption~ GDP+Freedom+Generosity, data=happy)

plot(modelq4)

#check autocorrelation

library(lmtest)

dwtest(modelq4, alternative=c("two.sided"))

##Check homoscedasticity

plot(residuals(modelq4)~fitted(modelq4))

abline(lm(residuals(modelq4)~fitted(modelq4)), col="red")

bptest(modelq4)

##Check normality

shapiro.test(residuals(modelq4))

##Based on the linear regression, all variables I chose are significantly

##associated with corruption. Shapiro test is significant, but QQ plot looks normal. Residuals seem to not have

## a pattern, suggesting homoscedasticity, even though bptest is significant. dwtest is not significant.

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? If yes, how do you interpret the interaction? **Happy**

modelq5=lm(Corruption~ Freedom \*Region , data=happy)

summary(modelq5)

# Call:

# lm(formula = Corruption ~ Freedom \* Region, data = happy)

# Residuals:

# Min 1Q Median 3Q Max

# -0.20801 -0.05591 -0.01298 0.04804 0.29995

# Coefficients:

# Estimate Std. Error t value Pr(>|t|)

# (Intercept) 0.03558 0.02973 1.197 0.2333

# Freedom 0.35091 0.08559 4.100 6.77e-05 \*\*\*

# RegionAmericasCarribean 0.07653 0.08344 0.917 0.3605

# RegionAsiaAustralia -0.05444 0.07654 -0.711 0.4780

# RegionEurope -0.10550 0.04464 -2.363 0.0194 \*

# Freedom:RegionAmericasCarribean -0.33785 0.19640 -1.720 0.0875 .

# Freedom:RegionAsiaAustralia 0.05666 0.17967 0.315 0.7529

# Freedom:RegionEurope 0.25902 0.11845 2.187 0.0303 \*

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#

# Residual standard error: 0.09112 on 149 degrees of freedom

# Multiple R-squared: 0.3567, Adjusted R-squared: 0.3265

# F-statistic: 11.8 on 7 and 149 DF, p-value: 6.547e-12

#check autocorrelation

dwtest(modelq5, alternative=c("two.sided"))

##Check homoscedasticity

plot(residuals(modelq5)~fitted(modelq5))

abline(lm(residuals(modelq5)~fitted(modelq5)), col="red")

bptest(modelq5)

##Check normality

shapiro.test(residuals(modelq5))

##The output from the ancova tells us that freedom is significant on the slope

##of relationship between the first region (africa) and corruption. We also see that the

##intercept for europe is significantly different than for africa and is less than africa.

## The interaction terms tell us if the slopes for each region are significantly different

##than the slope for the first region, listed as the coefficient for freedom. The output shows

## that Europe is signifcantly different than the slope for Africa, and is above Africa by 0.25.

##Also, the output shows AmericasCarribean region has a slope -0.33 less than Africa, if significance is at 0.1.

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**

head(cancer)

cor(cancer)

pairs(cancer)

modelq6 = glm(malignant~texture\_mean + smoothness\_mean + area\_mean, data=cancer, family=binomial(link="logit"))

summary(modelq6)

# Call:

# glm(formula = malignant ~ texture\_mean + smoothness\_mean + area\_mean,

# family = binomial(link = "logit"), data = cancer)

# Deviance Residuals:

# Min 1Q Median 3Q Max

# -2.14463 -0.20032 -0.04419 0.01878 2.85800

# Coefficients:

# Estimate Std. Error z value Pr(>|z|)

# (Intercept) -32.619261 3.608250 -9.040 < 2e-16 \*\*\*

# texture\_mean 0.381076 0.057677 6.607 3.92e-11 \*\*\*

# smoothness\_mean 146.766602 19.221507 7.636 2.25e-14 \*\*\*

# area\_mean 0.016260 0.001827 8.901 < 2e-16 \*\*\*

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# (Dispersion parameter for binomial family taken to be 1)

# Null deviance: 751.44 on 568 degrees of freedom

# Residual deviance: 182.53 on 565 degrees of freedom

# AIC: 190.53

# Number of Fisher Scoring iterations: 8

plot(malignant~texture\_mean + smoothness\_mean + area\_mean, data=cancer)

##Based on the output from this GLM model, all three factors I chose,

## texture, area, and smoothness, are significant in determining if the

## tumor is malignant. I avoided selecting perimeter and radius in the same

## model with area, since there is such a high correlation between those variables.

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.**

modelq7a = glm(malignant~texture\_mean, data=cancer, family=binomial(link="logit"))

modelq7b = glm(malignant~smoothness\_mean, data=cancer, family=binomial(link="logit"))

modelq7c = glm(malignant~area\_mean, data=cancer, family=binomial(link="logit"))

AIC(modelq7a, modelq7b, modelq7c)

# df AIC

# modelq7a 2 650.5191

# modelq7b 2 677.9485

# modelq7c 2 329.6565

##Based on picking the lowest AIC value, the area of the tumor is the most

##important factor for determining if it is malignant or not.