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 (do not include only the R output)
* 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=",")

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?**

check whether the data meet the requirements/assumptions of the test you plan to run

**Chi-square** assumptions:

INDEPENDENT SAMPLES: **CHECK**

NO STRUCTURAL ZEROS: **CHECK**

**No Somewhat Very**

**Female** 91 193 158

**Male** 56 155 190

LESS THAN 20% OF CELLS HAVE EXPECTED VALUES LESS

THAN 5: **CHECK (no expected values in this case)**

run the test

Pearson's Chi-squared test

data: tbl

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

interpret the results (do not include only the R output)

With a p-value less than .o5, **we reject the null hypothesis**.

**There is a statistically significant association between gender and whether people think it is rude to bring an unruly child flying.**

A higher percentage of **men (~47%)** than **women (~36%)** think it is **very rude** to bring an unruly child.

------------------------------------------------------------------------------------------------------------

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

check whether the data meet the requirements/assumptions of the test you plan to run

Running a **two-sample, unpaired t-test (after checking assumptions, I decided to run a Welch’s t-test)**

ASSUMPTIONS:

Continuous data: **CHECK**

Random sample: **CHECK**

Independent observations: **CHECK**  
Normal dist of n>=30: **NO (fails Shaprio Wilks, QQ could be better, transform needed)**

Equal variance: **NO**, **will need to use Welch’s t-test (var.equal = FALSE)**

complete any transforms needed to make the data meet the required assumptions

Square root-transformed to get a better qq-plot. Both the original data and the transformed data fail Shapiro-Wilks. Log transform plot was significantly worse than the original data.



BEFORE SQRT TRANSFORM (left) AFTER SQRT TRANSFORM (right)

Fits to the line a little better, so I proceeded with the transformed data.

run the test

Welch Two Sample t-test (using the transformed data)

data: public$Ttrans and private$Ttrans

t = -21.847, df = 1328.9, **p-value < 2.2e-16**

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

95 percent confidence interval:

-32.97451 -27.54052

sample estimates:

**mean of x mean of y**

**135.4532 165.7107**

interpret the results (do not include only the R output)

based on **p-value**, **we reject the null hypothesis that the mean tuition between public and private colleges are equal.**

There is a statistically significant difference in tuition based on institution type. **Private schools have a higher average tuition.**

I also ran the results on the original data, and they were the same.

------------------------------------------------------------------------------------------------------------

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

check whether the data meet the requirements/assumptions of the test you plan to run

**ANOVA (one-way) assumptions:**

Normal distribution: **CHECK, passes Shapiro-Wilks, qqplot looks good**

Independent samples: **CHECK**

Equal variance: **CHECK, passes Levene’s Test**

Does not require equal same sizes: **CHECK**

run the test

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

interpret the results (do not include only the R output)

Ho: There is no statistical difference in happiness score based on region.

**Based on the p-value (1.28e-12), we can reject the null hypothesis.**

**There is a significant difference in happiness score by region.**

------------------------------------------------------------------------------------------------------------

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

check whether the data meet the requirements/assumptions of the test you plan to run

**Multiple linear regression assumptions:**

Linear relationship: **QUESTIONABLE**, few of the variables have a linear relationship with the dependent variable, either before or after transformation. I chose to ignore this ☺

Homoscedasicity: **CHECK** residuals pass Breusch-Pagan test

Statistical independence of errors: **QUESTIONABLE** residuals fail Shapiro-Wilks, but the QQplot looks relatively OK

Normality of errors: **CHECK** residuals pass Durbin-Watson test

complete any transforms needed to make the data meet the required assumptions

**I log transformed the dependent variable before running the model to approximate a more normal distribution. The log-transformed data passed the Shapiro-Wilks test.**

run the test

Call:

lm(formula = happy$Ctrans1 ~ happy$Life + happy$Generosity + happy$Freedom)

Residuals:

Min 1Q Median 3Q Max

-2.36070 -0.41907 0.06779 0.53881 1.27486

Coefficients:

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

(Intercept) -3.1091 0.1870 -16.625 < 2e-16 \*\*\*

happy$Life -0.0733 0.2578 -0.284 0.777

happy$Generosity 0.5186 0.4458 1.163 0.246

**happy$Freedom 2.2601 0.4347 5.199 6.31e-07 \*\*\***

---

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

Residual standard error: 0.6933 on 153 degrees of freedom

Multiple R-squared: 0.2107, **Adjusted R-squared: 0.1952**

F-statistic: 13.61 on 3 and 153 DF, p-value: 6.445e-08

interpret the results (do not include only the R output)

The model says that **Freedom is the only variable in the model that is significantly effecting our dependent variable** (Corruption). As Freedom increases one unit, Corruption goes up 2.26 log-transformed units. Other variables are not significant.

check model fit in the case of linear regressions and/or glms

Based on the Adjusted R-squared (chosen because there are multiple variables in the regression), this linear model explains 19.5% of the variation in the data.

There is little multicollinearity between the three variables I chose (Life, Generosity, and Freedom).

------------------------------------------------------------------------------------------------------------

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**

ANCOVA assumptions (run with lm):

check whether the data meet the requirements/assumptions of the test you plan to run

Linear relationship: **CHECK** Freedom and Corruption have a somewhat linear relationship and a correlation coefficient of .45.

Homoscedasicity: **CHECK** residuals pass Breusch-Pagan test

Statistical independence of errors: **QUESTIONABLE** residuals fail Shapiro-Wilks, but the QQplot and histogram look OK

Normality of errors: **CHECK** residuals pass Durbin-Watson test

complete any transforms needed to make the data meet the required assumptions

As in the last question, I log-transformed the dependent variable (Corruption)

run the test

#Call:

# lm(formula = happy$Ctrans1 ~ happy$Freedom \* happy$Region)

#Residuals:

# Min 1Q Median 3Q Max

#-2.08487 -0.31735 0.07803 0.40588 1.38756

#Coefficients:

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

**#(Intercept) (Africa Mideast) -2.6096 0.2102 -12.416 < 2e-16 \*\*\***

**# Freedom 1.5929 0.6050 2.633 0.009360 \*\***

#RegionAmericasCarribean 0.3985 0.5898 0.676 0.500360

# RegionAsiaAustralia -0.5177 0.5410 -0.957 0.340110

**# RegionEurope -1.3755 0.3155 -4.360 2.42e-05 \*\*\***

# Freedom: RegionAmericasCarribean -1.6566 1.3882 -1.193 0.234643

# Freedom: RegionAsiaAustralia 0.6657 1.2700 0.524 0.600949

**# Freedom: RegionEurope 2.9826 0.8373 3.562 0.000494 \*\*\***

# ---

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

#Residual standard error: 0.6441 on 149 degrees of freedom

#Multiple R-squared: 0.3365, **Adjusted R-squared: 0.3053**

#F-statistic: 10.79 on 7 and 149 DF, p-value: 5.739e-11

interpret the results (do not include only the R output)

The intercept in the output is AfricaMidEast. There is a significant interaction between Freedom and countries in the Europe region to influence Corruption.

There is a significant relationship between countries in Europe and Corruption. Based on the estimate of -1.3755, Corruption in the Europe region is significantly lower than the intercept, which is AfricaMidEast region. Other regions are not significantly different from the intercept. Moreover, the interaction between Freedom and Region is only significant in the Europe region.

check model fit in the case of linear regressions and/or glms

Based on the adjusted R-squared the model explains about 30% of the variation in the data.

------------------------------------------------------------------------------------------------------------

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**

check whether the data meet the requirements/assumptions of the test you plan to run

GLM Assumptions (binomial):

Independent observations: **CHECK**

Broader exponential family: **CHECK**, binomial

Linear relationship between transformed dependent and independent variable: **CHECK**

complete any transforms needed to make the data meet the required assumptions

I used a GLM with a binomial distribution

run the test

Call:

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

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

Deviance Residuals:

Min 1Q Median 3Q Max

-2.16715 -0.17908 -0.03790 0.03779 3.00637

Coefficients:

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

(Intercept) -40.40324 4.36615 -9.254 **< 2e-16 \*\*\***

perimeter\_mean 0.21008 0.02336 8.993 **< 2e-16 \*\*\***

texture\_mean 0.37371 0.05779 6.466 **1.01e-10 \*\*\***

smoothness\_mean 134.02497 18.73540 7.154 **8.46e-13 \*\*\***

---

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: 181.68 on 565 degrees of freedom

**AIC: 189.68**

Number of Fisher Scoring iterations: 8

interpret the results (do not include only the R output)

The R output above indicates that all three of the independent variables significantly impact whether a tumor is malignant or not. The significant intercept means that even if all of the independent variables were 0, there is still a possibility that the tumor could be malignant.

check model fit in the case of linear regressions and/or glms

I originally used radius, perimeter, and area\_mean in the model, but they were too highly correlated. The AIC value for that model was 248.94. Once I noticed the multicollinearity, I changed the independent variables I was testing. The new model has a much lower AIC value (189.68).

------------------------------------------------------------------------------------------------------------

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.**

malignant

malignant 1.0000000

**radius\_mean 0.7300285**

texture\_mean 0.4151853

**perimeter\_mean 0.7426355**

**area\_mean 0.7089838**

smoothness\_mean 0.3585600

Based on the correlation coefficients (generated using the code: *cor(cancer)*), radius\_mean, perimeter\_mean, and smoothness\_mean are the most highly correlated with whether or not a tumor is malignant. This is why I initially chose them for the last model (I later changed by variables because they were highly correlated).