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? **flying**

Because both variables are categorical and you want to see if there is a significant difference between the two a chi-square test for independence is most appropriate.

Hypothesis testing:

H0: there is no significant association between gender and perception of rudeness

H1: there is a significant association between gender and perception of rudeness

**R Output:**

1. Contingency table:

No Somewhat Very

Female 91 193 158

Male 56 155 190

2. Pearson's Chi-squared test

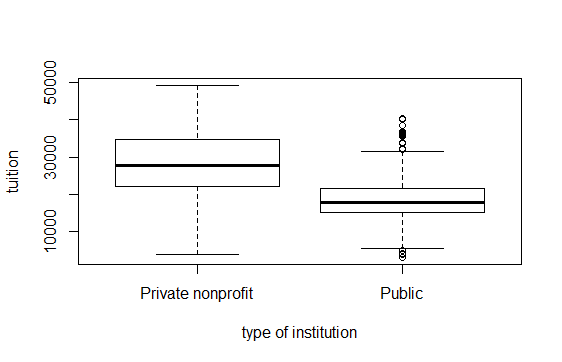
data: table(flying$gender, flying$unruly\_child)

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

**Interpretation:**

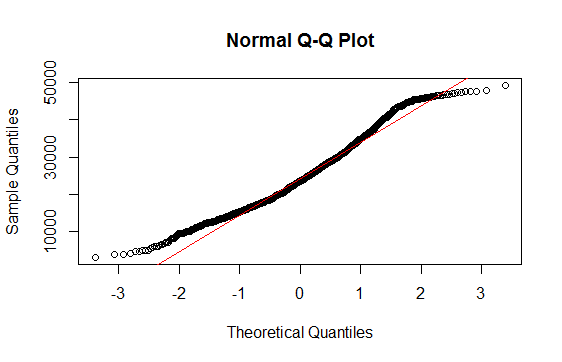
The results of the test indicate that there is a significant difference between gender and their perceptions on whether it’s rude to bring an unruly child on a plane or not. The contigency table indicates that a higher proportion of men than women think that brining an unruly child on a plane is very rude.

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



**Assumptions:**

* Normality: the data are not normal, but the sample size is fairly large, so following the central limit theorem, I will just assume normality.



* Variance: The variance among the data is unequal as shown below:

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

* Independence: I assume that the data points are independent

**Statistical test:** Because the variance is unequal, the non-parametric test, Welch’s t-test is most appropriate.

Hypothesis testing:

H0: There is no significant difference in tuition between public and private colleges.

H1: There is a significant difference in tuition between public and private colleges

**R Output:**

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:

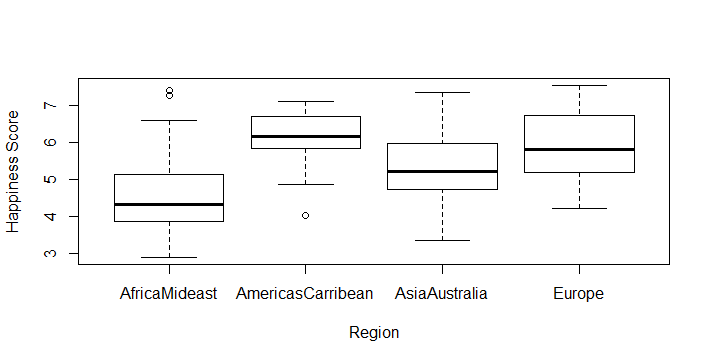
mean of x mean of y

18855.88 28301.69

**Interpretation:**

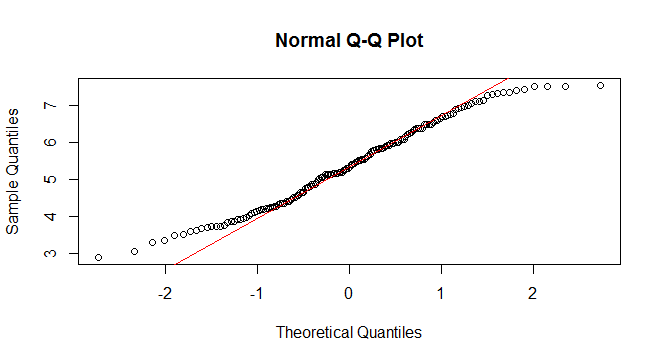
The results of the t-test indicate that there is a significant difference in tuition between public and private colleges (we reject the null hypothesis). The boxplot indicates that the mean tuition of private institutions is higher than the mean tuition of public institutions.

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



**Assumptions:**

* Normality: the data is not normal, but the sample size is fairly large, so following the central limit theorem, I will just assume normality.



* Variance: The variance among the data is equal
* Independence: I assume that the data points are independent

**Statistical Test:** ANOVA test is most appropriate considering that all of the assumptions are met.

Hypothesis testing:

H0: There is no significant difference in happiness across regions.

H1: There is a significant difference in happiness across regions.

**R Output:**

**ANOVA 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

**Post-Hoc Tukey HSD test:**

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = Hscore ~ Region, data = happy)

$Region

diff lwr upr p adj

AmericasCarribean-AfricaMideast 1.5397193 0.94745533 2.13198329 0.0000000

AsiaAustralia-AfricaMideast 0.7472543 0.14642988 1.34807874 0.0081572

Europe-AfricaMideast 1.3208593 0.84313961 1.79857901 0.0000000

AsiaAustralia-AmericasCarribean -0.7924650 -1.49989372 -0.08503628 0.0214193

Europe-AmericasCarribean -0.2188600 -0.82522748 0.38750748 0.7847082

Europe-AsiaAustralia 0.5736050 -0.04112656 1.18833656 0.0768934

**Interpretation:**

The results of the ANOVA test indicate that there is a significant difference in happiness across regions, and we reject the null hypothesis. The Tukey HSD test results indicate that there is a significant difference (p <= 0.05) in happiness levels between the following regions:

* Americas/Caribbean and Africa/Mideast (happiness levels are greater in the Americas than in Africa)
* Asia/Australia and Africa/Mideast (happiness levels are greater in Asia than in Africa)
* Europe and Africa/Mideast (happiness levels are greater in Europe than in Africa)
* Asia/Australia and Americas/Caribbean (happiness levels in Asia are less than in the Americas)

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

**Correlation coefficients:** The table and plot below indicate the correlation coefficients between all of the continuous variables in the data set. Specifically related to corruption, there appears to be low correlation between corruption and the remaining variables (which might lead to a weak linear model). For the linear model in this exercise I chose the variables that are most correlated with corruption (had the highest correlation coefficients): GDP, Freedom, and Generosity. Additionally, the correlation coefficients between the three explanatory variables are low, indicating that there is little correlation between them, so it is appropriate to proceed with the model.

H0: GDP, freedom, and generosity *are not* significantly associated with corruption.

H1: GDP, freedom, and generosity *are* significantly associated with corruption

GDP Family Life Freedom Corruption Generosity

GDP 1.00000000 0.66953969 0.83706723 0.3622828 0.2941848 -0.02553066

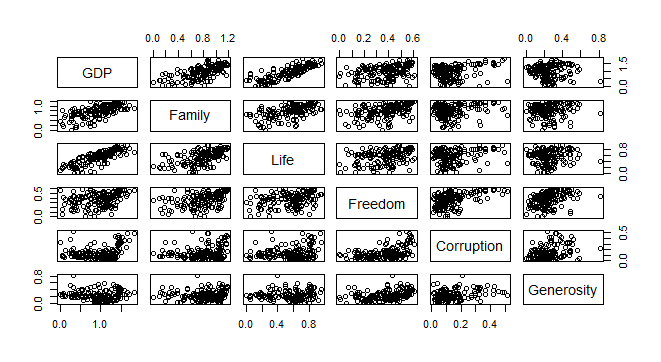
Family 0.66953969 1.00000000 0.58837678 0.4502082 0.2135609 0.08962885

Life 0.83706723 0.58837678 1.00000000 0.3411993 0.2495833 0.07598731

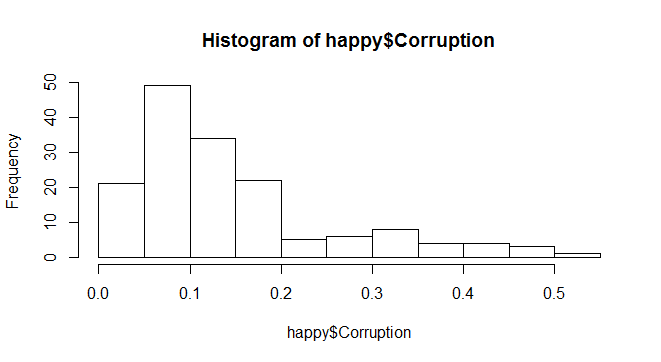
Freedom 0.36228285 0.45020820 0.34119929 1.0000000 0.5020540 0.36175133

Corr. 0.29418478 0.21356094 0.24958329 0.5020540 1.0000000 0.30592986

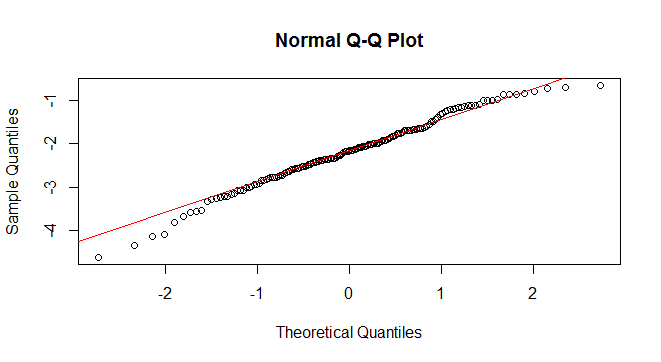
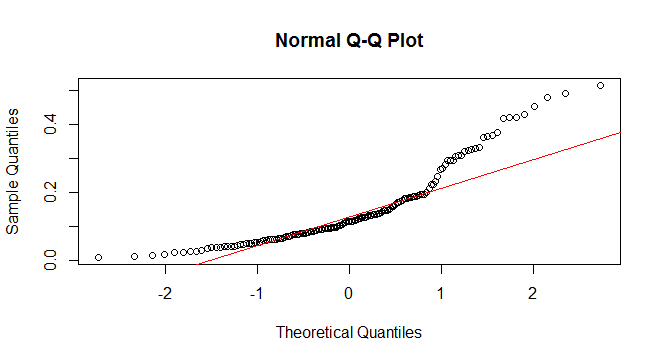
Gen. -0.02553066 0.08962885 0.07598731 0.3617513 0.3059299 1.00000000



* Normality check of the dependent variable (Corruption): the distribution of Corruption is not normal (verified by the Shapiro test and qq-plot), so I decided to transform the variable with a log transformation which resulted in a normal distribution of y.



Before log transformation: After log transformation



**Linear Model**

* Output:

lm(formula = log.cor ~ GDP + Freedom + Generosity, data = happy)

Residuals:

Min 1Q Median 3Q Max

-2.36387 -0.42476 0.04111 0.52922 1.27487

Coefficients:

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

(Intercept) -3.16388 0.18628 -16.985 < 2e-16 \*\*\*

GDP 0.04341 0.14673 0.296 0.768

Freedom 2.16578 0.44615 4.854 2.95e-06 \*\*\*

Generosity 0.54962 0.45252 1.215 0.226

---

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

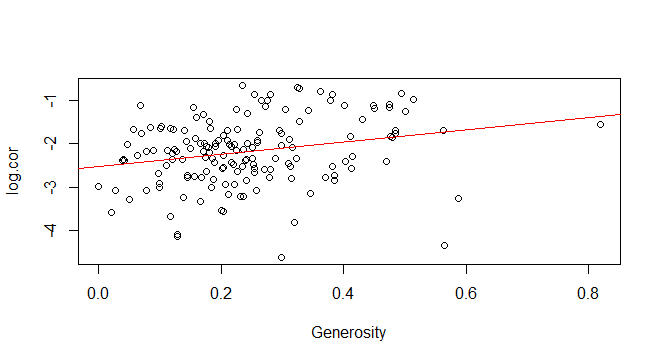
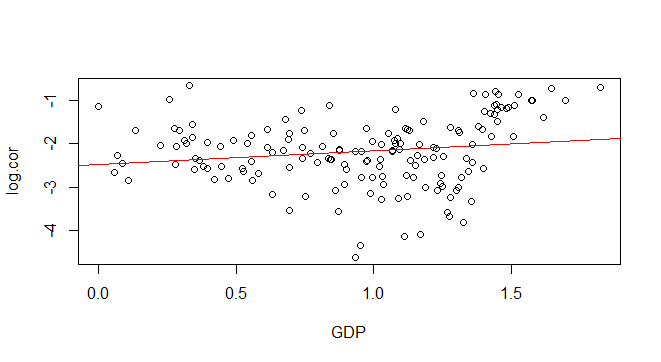
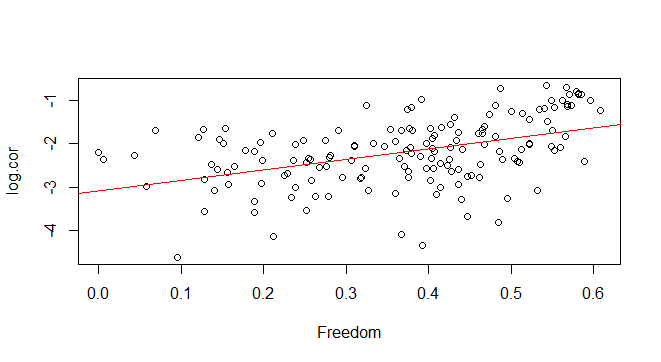
Residual standard error: 0.6933 on 153 degrees of freedom

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

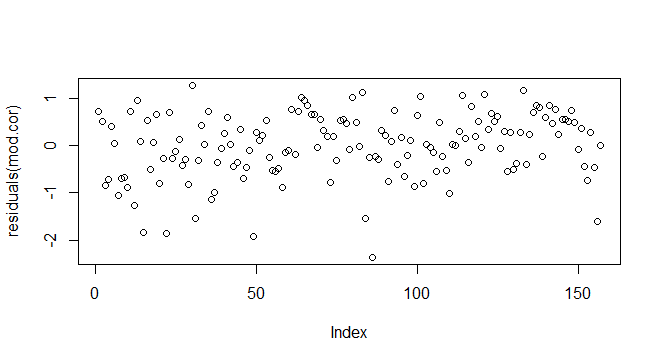
F-statistic: 13.61 on 3 and 153 DF, p-value: 6.424e-08

**Model Check (assumptions)**

* Linearity (coefficient of determination): The plots below show the relationships between corruption (after the log transformation) and each of the independent variables (GDP, Freedom, and Generosity). The plots indicate that there is a linear relationship of varying strength between the variables. For example, the relationship between log.cor and freedom is more pronounced than the relationship between log.cor and GDP which is flatter.



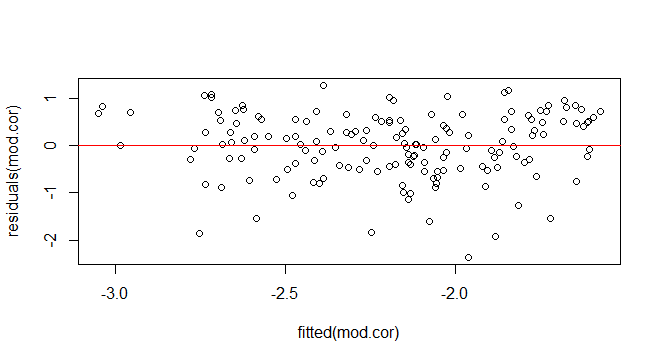
* Multicollinearity: The calculated VIF values (shown below) are all low and less that five, indicating that collinearity between independent variables in the model is low. Additionally the plot indicates that there is no pattern in the distribution of the residuals.

vif(mod.cor)

GDP Freedom Generosity

1.189712 1.367952 1.189185

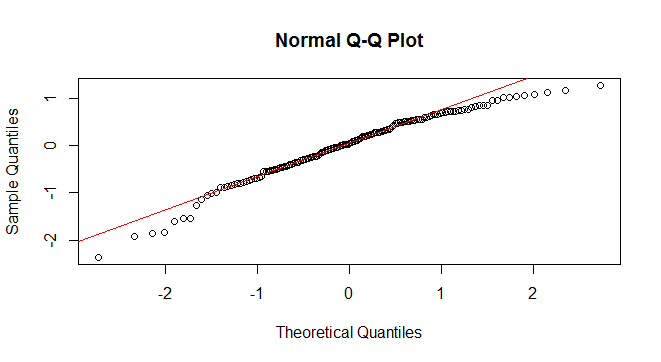
* Homoscedasticity: the plot indicates that the residuals are evenly distributed above and below the plot line. The Breusch-Pagan test confirms that the residuals are homoscedastic.

studentized Breusch-Pagan test

data: mod.cor

BP = 7.1001, df = 3, p-value = 0.06878

* Normality of Residuals: The residuals failed the Shapiro test, but the qq-plot indicates that with the exception of some outliers the majority of the residual points lie on the plot line.

Shapiro-Wilk normality test

data: residuals(mod.cor)

W = 0.96589, p-value = 0.0006382

**Interpretation:** The fit of the model is low (as indicated by the low Adjusted R-squared value) – only 19.52% of the variation of corruption (after the log transformation) is accounted for by the explanatory variables. The output of the linear model indicates that the intercept of the log transformation of corruption is significant and crosses the y-axis at -3.16. The regression coefficient for “GDP” estimates the effect of GDP (or wealth) on the log transformed values of corruption while controlling for the effect of freedom to make individual choices and generosity on the log transformation of corruption. The same is true for the regression coefficients of “Freedom” and “Generosity”. The model indicates that there is a significant positive relationship between the log transformation of corruption and freedom to make individual choices, where for every increase in freedom, there is an increase of 2.17 units of the log transformation of corruption. Additionally, the model indicates that there is not a significant relationship between GDP and Generosity on the log transformation of corruption. In conclusion, the overall fit of the model to the data is low – the model tells us that the effect of the explanatory variables on the dependent variable after transformation is statistically significant, but it accounts for less than 20% of the variation in the data.

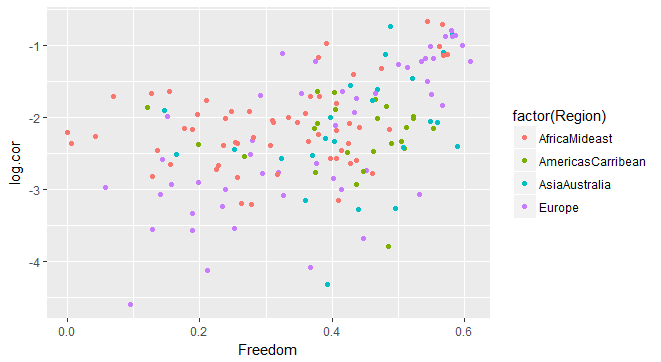
1. Choose one of the continuous independent variables that was significant in the model for Question 4 and interact it with region (Region) to predict corruption (Corruption). This model should only include one continuous independent variable and its interaction with region. Does the influence of your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

For this ANCOVA model, I will be using the variable “Freedom” in order to determine the effect of freedom on the log transformation of corruption after controlling for region.

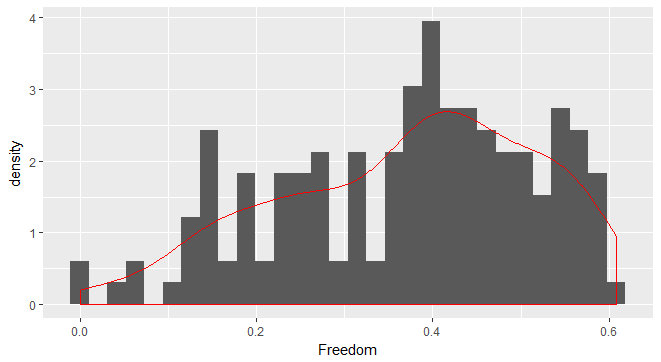
H0: The effect of freedom of choice on corruption does not vary significantly across regions

H1: The effect of freedom of choice on corruption does vary significantly across regions

The following plot shows the distribution of freedom in relation to corruption (after the transformation) according to region. As indicated by the plot, there appears to be a positive trend across all four regions between corruption and freedom; in other words, as freedom of choice increases, the values of corruption after the log transformation increases.



The following histogram shows the distribution of the x-variable (freedom):



From the histogram, the distribution of the data for freedom appears not to be normal, and the Shapiro test confirms this assumption. However, because this is a relatively large sample size, I will assume normality under the Central Limit Theorem.

**Linear model (with interaction term) output:**

Call:

lm(formula = log.cor ~ Freedom \* Region, data = happy)

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) -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

summary(ANCOVA)

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

Freedom 1 18.92 18.916 45.595 3.04e-10 \*\*\*

Region 3 4.60 1.533 3.696 0.013276 \*

Freedom:Region 3 7.83 2.610 6.290 0.000478 \*\*\*

Residuals 149 61.82 0.415

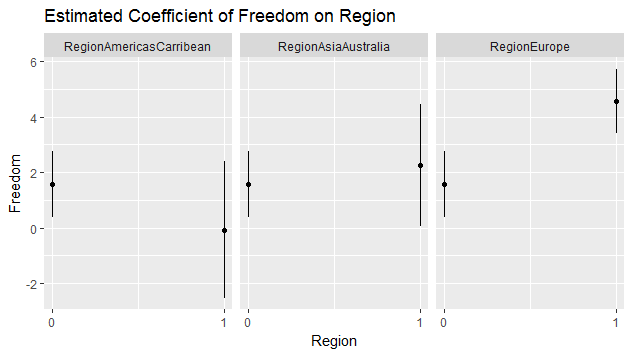
---

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

**Interpretation:** The output of the linear model indicates that the intercept of the log transformation of corruption in the Region “Africa/MidEast” is significant and crosses the y-axis at -2.61. Additionally, the output indicates that there is a significant positive association between freedom of choice and corruption in “Africa/Mideast” (the log value of corruption increases 1.59 units for every unit increase in freedom). The estimates of the remaining regions indicate the average difference in the log value of corruption between “Africa/MidEast” and “Americas/Caribbean”, “Asia/Australia”, and “Europe. The only significant difference in corruption across regions was that between Africa and Europe.

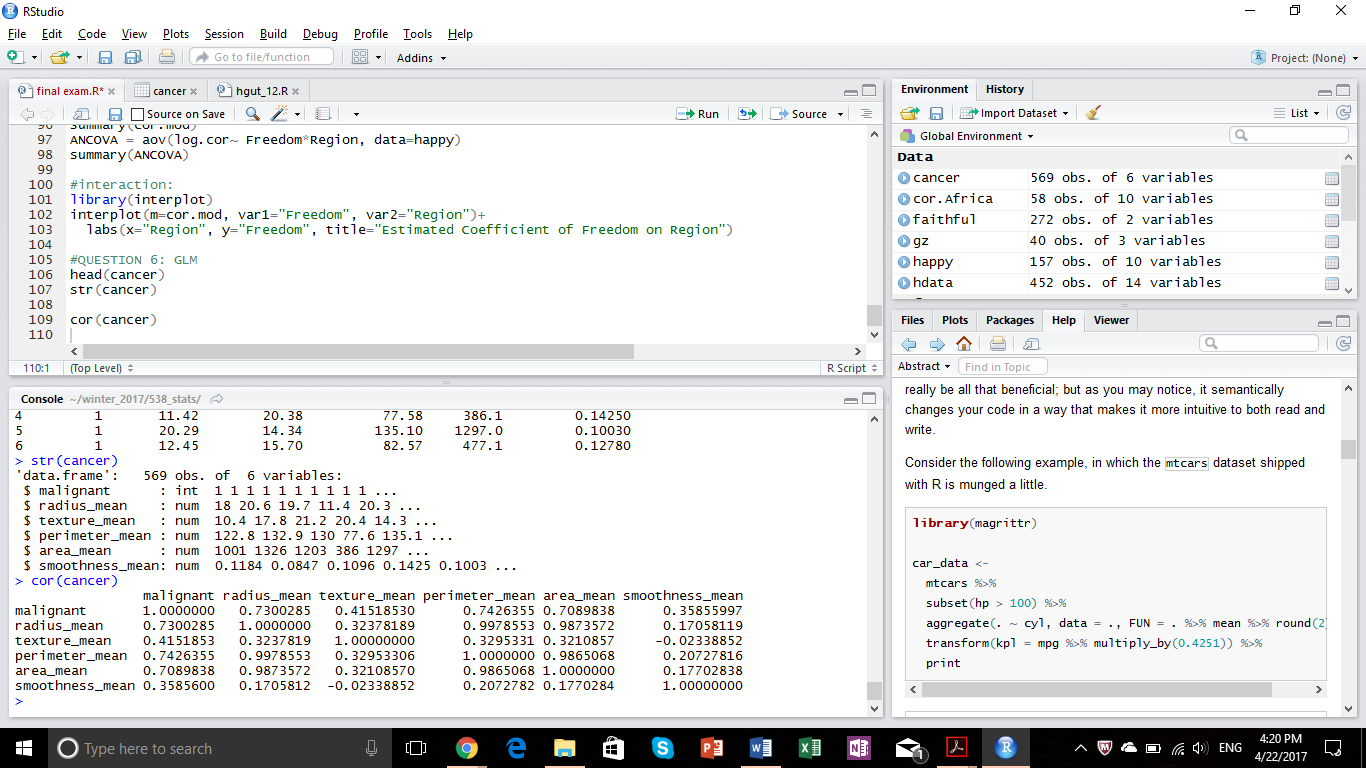
Additionally, the only significant interaction term between region and freedom was with Europe, which means that the relationship between freedom of choice and corruption is only significantly different in Africa and Europe. This interaction (demonstrated in the plot below) indicates that the effect (regression coefficient) of freedom on corruption is significantly lower in Africa/MidEast than it is in Europe.

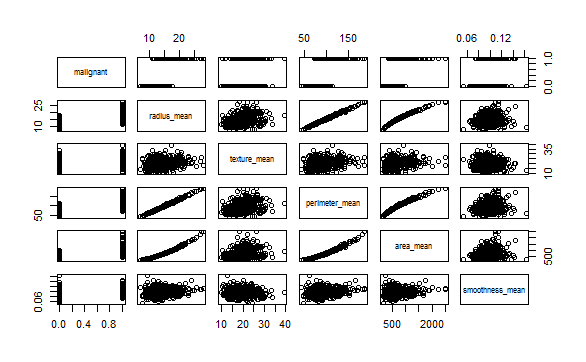
**Interaction plot:**



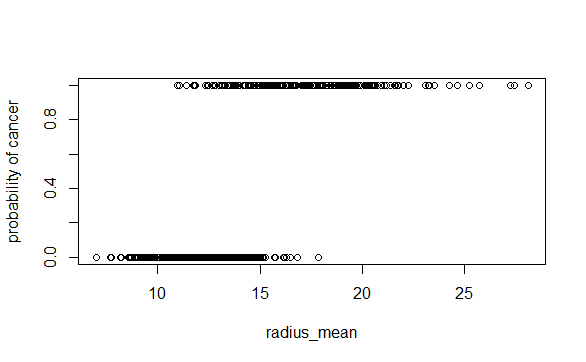
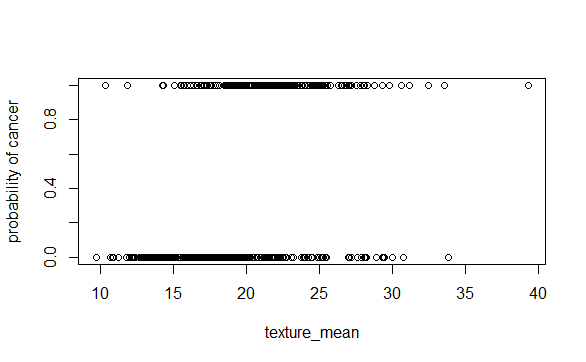
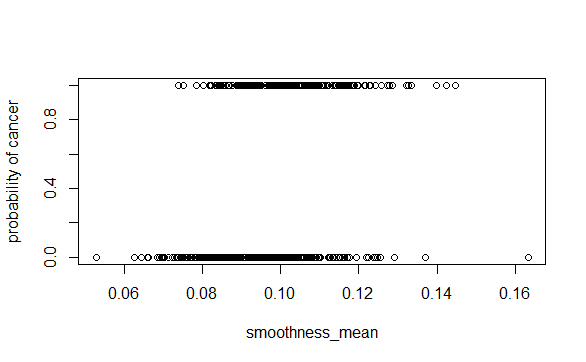
1. Which factors are significantly associated with whether a breast cancer tumor is malignant or not? Choose three continuous independent variables to include in your model. **Cancer**

**Correlation coefficients:** The table and plots below indicate the correlation coefficients between all of the variables in the data set. For this exercise, I chose the following independent variables: “radius\_mean”, “texture\_mean”, and “smoothness\_mean”; these variables are appropriate because there is little correlation between them. I did not include “perimeter\_mean” or “area\_mean” because they are both derived with the radius and are thus highly correlated with one another. The dependent variable is binary (malignant or not), and therefore follows the binomial distribution, thus the model with will be created with the “logit” function.





The plots below indicate the distribution of each explanatory variable for malignant and non-malignant tumors.



Hypothesis testing:

H0: The variables radius, texture, and smoothness do not have a significant effect on the probability of having a malignant tumor.

H1: The variables radius, texture, and smoothness do have a significant effect on the probability of having a malignant tumor.

**Model output:**

Call:

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

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

Deviance Residuals:

Min 1Q Median 3Q Max

-2.19102 -0.19403 -0.03799 0.04025 2.92583

Coefficients:

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

(Intercept) -42.01941 4.45943 -9.423 < 2e-16 \*\*\*

radius\_mean 1.39699 0.15403 9.069 < 2e-16 \*\*\*

texture\_mean 0.38056 0.05711 6.663 2.68e-11 \*\*\*

smoothness\_mean 144.67423 19.04687 7.596 3.06e-14 \*\*\*

---

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: 187.29 on 565 degrees of freedom

AIC: 195.29

Number of Fisher Scoring iterations: 8

**Interpretation:** the model indicates that the probability for a tumor to be malignant (to get cancer) is significantly influenced by all three explanatory variables (radius, texture, and smoothness). The intercept of the probability of a tumor being malignant (log odds) when radius, texture, and smoothness are at zero is significant, meaning that there is still a probability of the tumor being malignant without those factors. The model indicates that for every increase in radius mean, the log odds of the tumor being malignant increases by a factor of 1.39; for every increase in texture mean, the log odds of the tumor being malignant increases by a factor of 0.38; and lastly, for every increase in smoothness mean, the log odds of the tumor being malignant increases by a factor of 144.67.

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.**

The variable that appears to be most important in detecting tumor malignancy is smoothness, because, as indicated by the model, for every increase in the mean smoothness of the tumor, the log odds of the tumor being malignant increases by a factor of 144.67.