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**

In this question, we want to perform hypothesis testing with a nominal dependent variable (gender) and ordinal independent variable (opinion on unruly children).

Null: Gender of passenger has no effect on they think it is rude to bring unruly children on the plane.

Alternate: Gender of passenger affects whether they think it is rude to bring unruly children on the plane.

The appropriate test to use is the chi-square test of independence. I started by creating a contingency table:

No Somewhat Very

Female 91 193 158

Male 56 155 190

Chi-square test assumptions check:

1. I assume observations are independent.
2. No structural zeros.
3. Values in cells all >5.

Performing the chi-square test, we get a p-value of 0.001193, which is smaller than 0.05. We can therefore reject the null hypothesis.

By examining the contingency table, we can clearly see that females are less likely to find bringing an unruly child on board an airplane rude.

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

In this question, we want to perform hypothesis testing with a categorical independent variable (type of institution) and a continuous dependent variable (tuition fee).

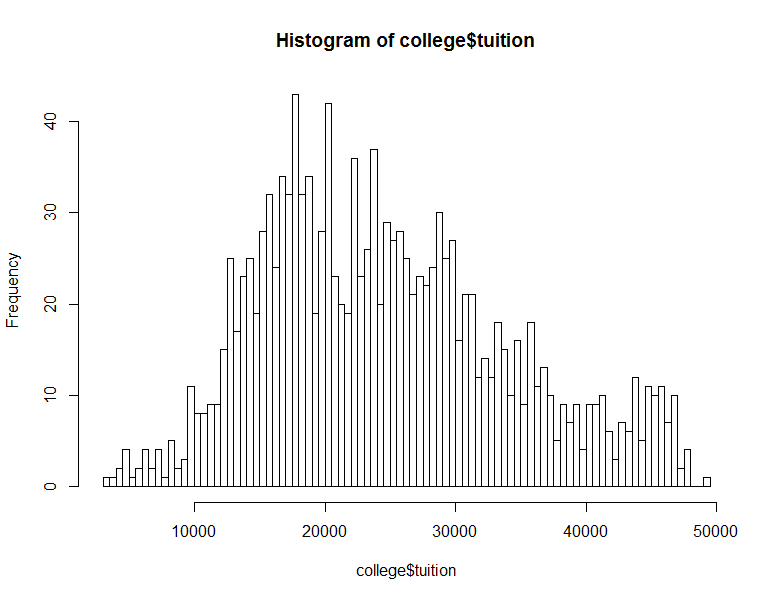
Null: There is no significant difference in tuition between public and private non-profit colleges.

Alternate: There is a significant difference in tuition between public and private non-profit colleges.

An unpaired t-test or Welch’s t-test is most appropriate since there are only two groups.

Assumptions checking:

1. Data is continuous.
2. Assume that samples are randomly selected from population.
3. Assume observations are independent.
4. Dependent variable (tuition) is normally distributed:

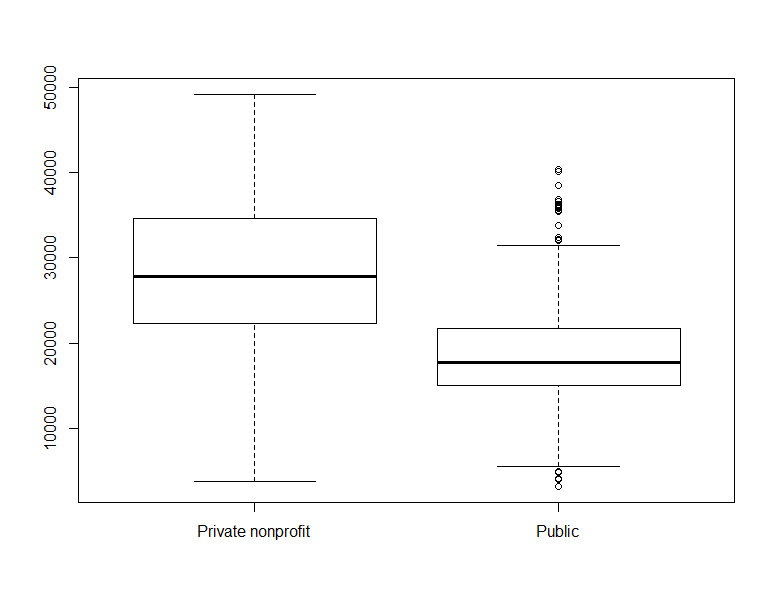


Shapiro test results: p = 3.193e-15

Data is not normally distributed, but since n is large, we can assume normality because of CLT.

1. Sample size between two groups not equal, and variance is also not equal (p<2.2e-16), suggesting we need to use a Welch’s t-test.

Performing the t-test, we see that there is a significant difference between the tuition of public schools and private schools, with a very small p-value of <2.2e-16. From this box plot, we can see that private non-profit colleges have more expensive tuitions:



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

In this question, we want to perform hypothesis testing with a continuous dependent variable (HScore) and categorical independent variable (Region).

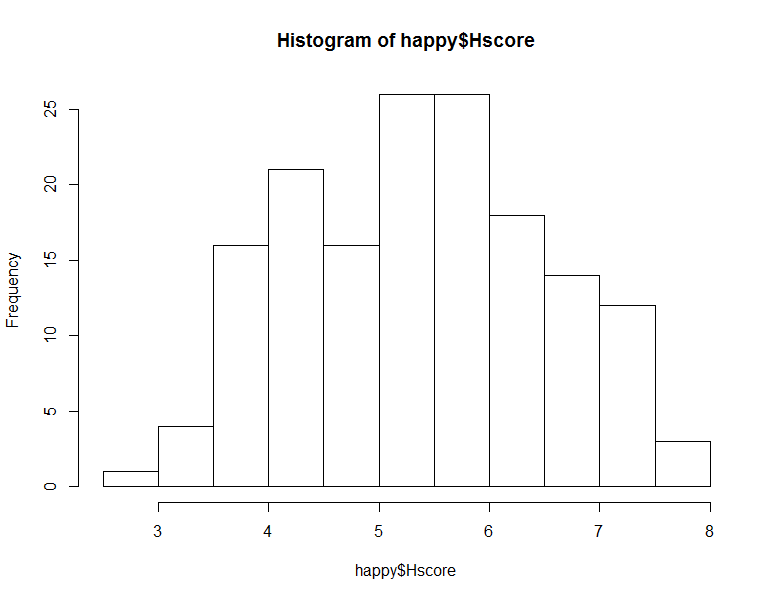
Null: Region has no significant effect on HScores.

Alternate: Region has a significant effect on HScores.

Since there are more than two regions, a one-way ANOVA test should be used for the hypothesis.

Assumptions checking:

1. Assume independence of observations, randomness of selection etc.
2. Normality of dependent variable:



Shapiro test: 0.01248

Not normal but since n >30, we can assume normality based on CLT.

1. Equal variance: Homogeniety of variances can be assumed after checking with Bartlett test (p=0.3428) and levene test (p=0.5179).

We can therefore run a one-way ANOVA with no need to transform. Performing an aov in R, we see that there is a significant difference between the HScores of different regions (p=1.28e-12). Null hypothesis rejected. With a post-hoc test, we can see the magnitude and significance of the differences between different regions:

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

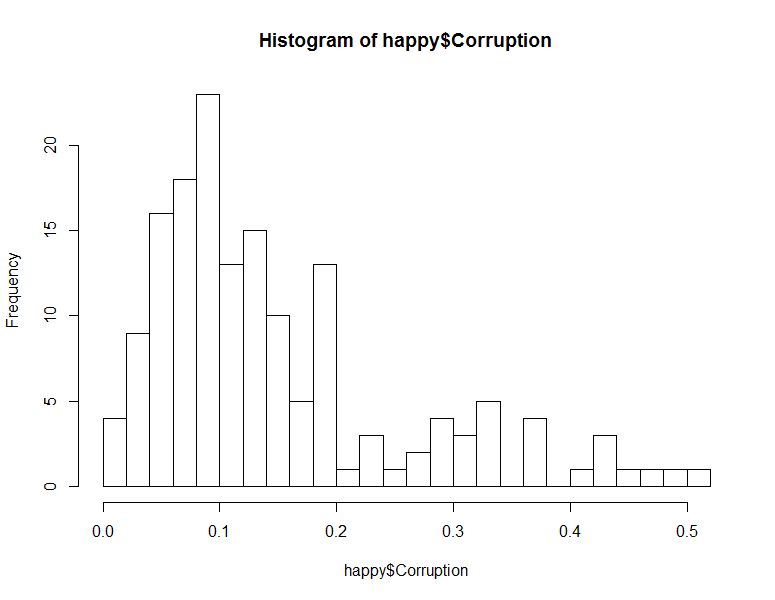
Europe-AsiaAustralia 0.5736050 -0.04112656 1.18833656 0.0768934

We see, for example, that the differences in HScores between Europe and Americas is not significant, and the largest significant mean difference in HScores is between Americas and AfricaMideast, with Americas averaging ~1.54pts higher.

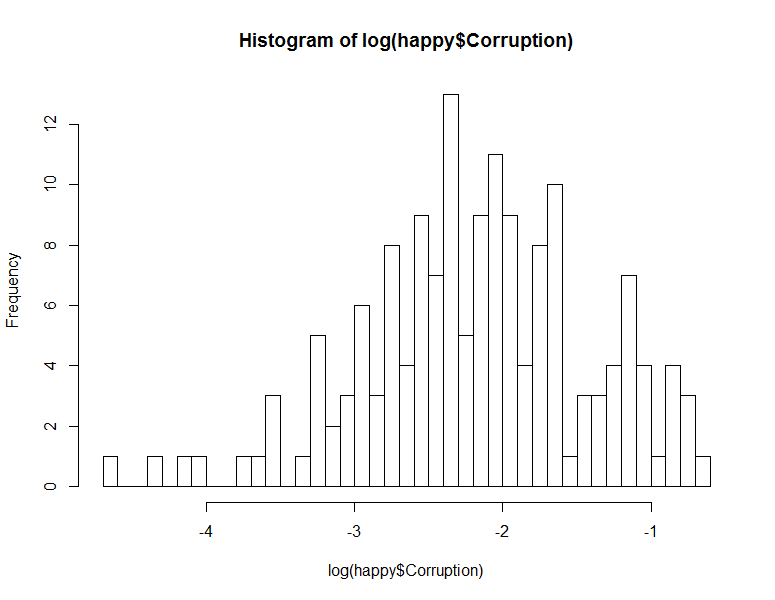
1. What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model. **happy**

This question is asking to create a multivariate regression model that examines factors associated with corruption levels in countries.

Before running the model, check if dependent variable is normally distributed to give a sense of what errors will look like:

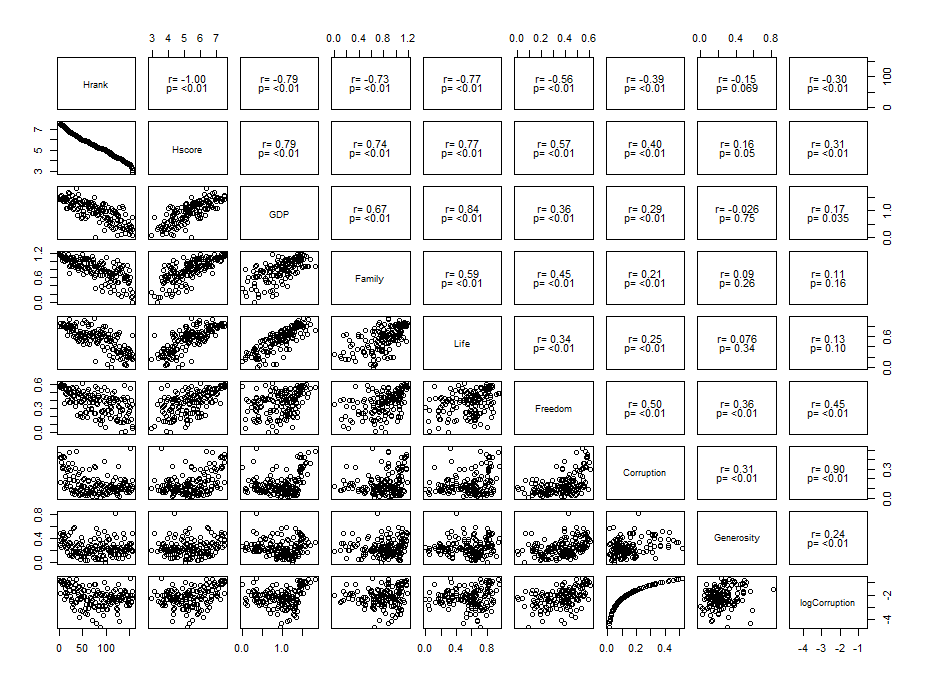


Shapiro test p-value = 5.128e-11. Log transformation seems like it will help to make the data more normal:



New shapiro test p-value = 0.08535. The transformed data can now be assumed normal and we can expect normal errors.

Now I need to analyze which independent continuous variables should be included in the model. I begin by checking for collinearity among the variables:



From this table, I see that freedom (r=0.45), HScore (r=0.31) and Hrank (r=0.30) are most correlated with the transformed corruption. However, as Hscore and Hrank are highly correlated (r=1.0), only one of them should be included. Checking again, we see that Hscore and Freedom are also correlated at r=0.57, suggesting that there could be variance inflation problems that I should check after running the regression. Other potential independent variables to use are Generosity (r=0.24) and GDP (r=0.17). I decided to replace Hrank with generosity.

I run the model logCorruption~Freedom+Hscore+Generosity and obtain the following results:

Call:

lm(formula = logCorruption ~ Freedom + Hscore + Generosity, data = happy)

Residuals:

Min 1Q Median 3Q Max

-2.35652 -0.41685 0.06602 0.51886 1.32674

Coefficients:

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

(Intercept) -3.35861 0.27504 -12.211 < 2e-16 \*\*\*

Freedom 1.94799 0.49004 3.975 0.000108 \*\*\*

Hscore 0.05872 0.05896 0.996 0.320816

Generosity 0.55330 0.44468 1.244 0.215312

---

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

Residual standard error: 0.6912 on 153 degrees of freedom

Multiple R-squared: 0.2153, Adjusted R-squared: 0.1999

F-statistic: 14 on 3 and 153 DF, p-value: 4.136e-08

From these results, we see that only Freedom has a significant relationship with logCorruption. The R-squared value is low at ~0.2, suggesting that our model only explains about 20% of the variation in the data. Based on this model, we can expect, while controlling for the other two variables, for every unit increase in Freedom there is an 1.95 unit increase of logCorruption, or a exp(1.95)=7.03 increase in Corruption score.

Checking model:

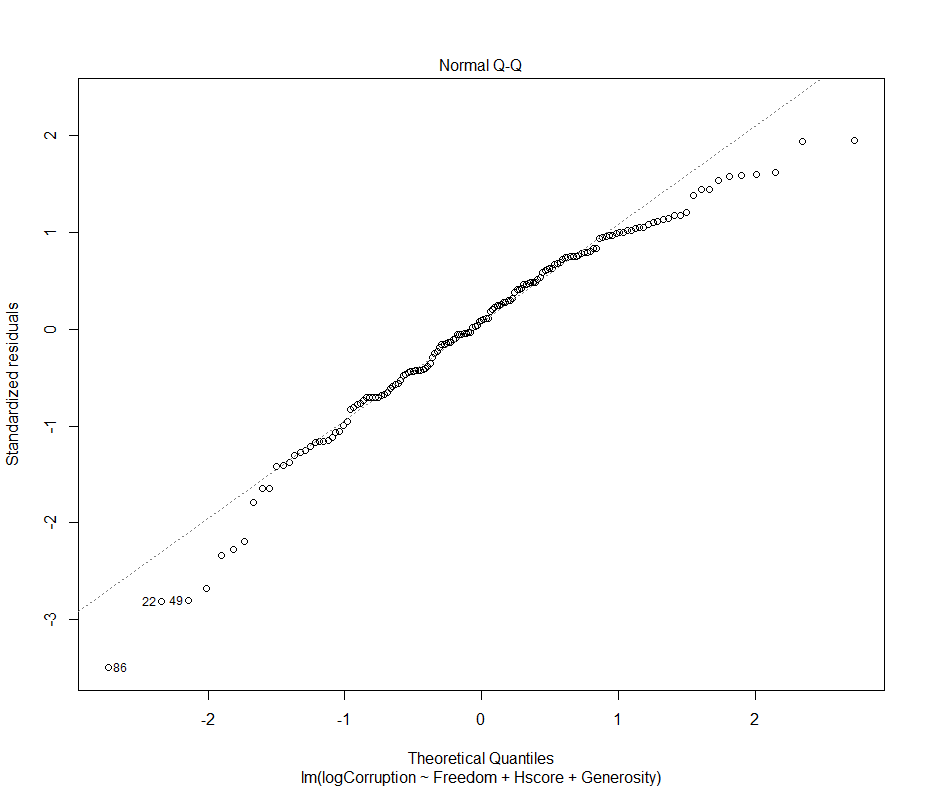
1. Check VIF: previously we saw that there was some correlation between Hscore and Freedom:

Freedom Hscore Generosity

1.660068 1.479215 1.155118

Values look good. No issues here.

1. Check normality of residuals with Shapiro test, p-value = 0.00067, residuals cannot be assumed normal. QQplot:



Issue seems to be with outliers at either end, most values seem alright.

1. Check homoscedasticity of residuals with Breusch-Pagan test, p-value = 0.2571, residuals are homoscedastic.
2. Check independence of residuals with Durbin Watson test, p-value = 0.1783, residuals are independent.
3. 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**

After including the interaction term for Freedom, the new model will be:

logCorruption ~ Freedom\*Region

I obtain the following output:

Call:

lm(formula = logCorruption ~ 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

Interpreting the output:

1. (Intercept) indicates where the fitted line crosses the Y-axis for AfricasMideast. Where Freedom score is zero we expect logCorruption to be -2.6096 (or corruption to be exp(-2.6096)) for AfricasMideast region.
2. The estimate for ‘Freedom’ indicates the slope of the fitted line for the AfricasMideast region.
3. The next three items in the Estimate column indicates the change in intercept for the different fitted lines for each region as compared to AfricasMideast. For example, the intercept for AmericasCarribean is -2.6096 + 0.3985, and -2.6096 – 0.5177 for AsiaAustralia. However, the only significant value is for Europe. This means that only the intercept for the European region is significantly different from the AfricaMideast region. This y-intercept is -2.6096 – 1.3755 =

-3.9851.

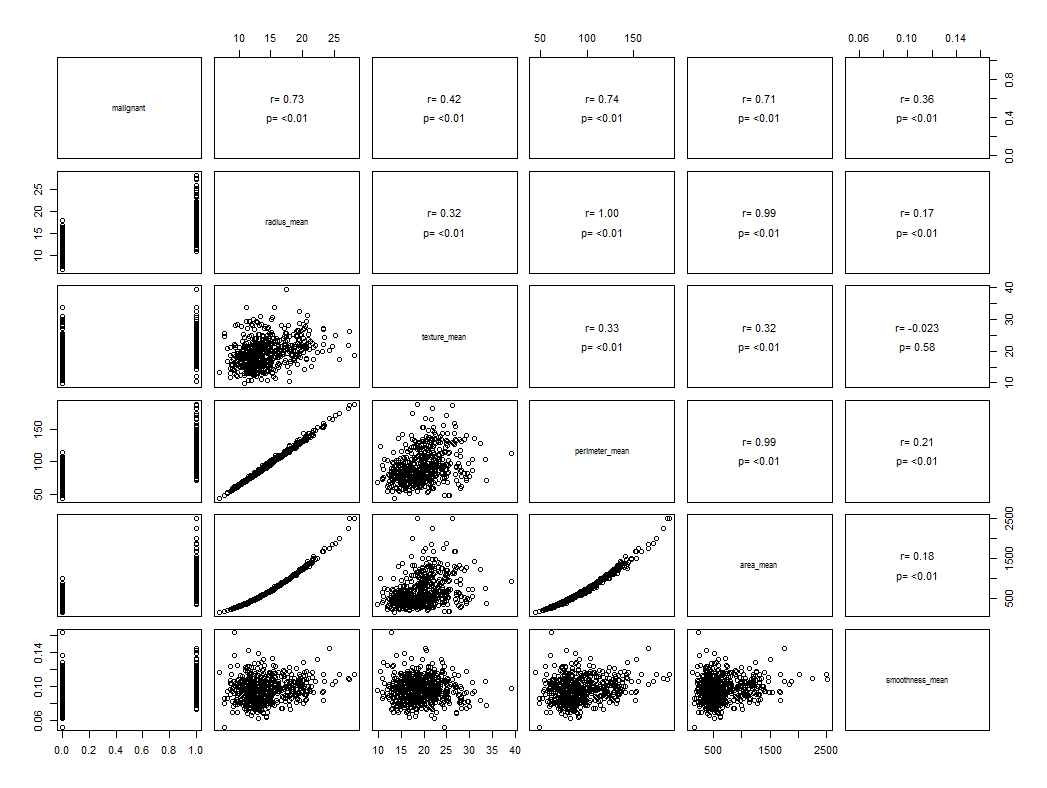
1. The last three items in the Estimate column indicates the change in slope for the fitted lines of the different regions compared to AfricaMideast. Of these, only the Europe region is significant. It has a slope of 1.5929 + 2.9826 = 4.5755. For the European region, logCorruption increases at 4.5755 per unit increase in Freedom as opposed to 1.5929 in the AfricaMideast region.
2. Adjusted R-squared is 0.3053, indicating that about 30.5% of the variance in the data can be explained by this model.

Overall, this means that the influence of freedom on logCorruption does show some regional variation. After running the model several times using a different region as the basis of comparison (see R script), I can conclude that this regional variation is only significantly different in Europe, where the effect of freedom is greater.

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**

The dependent variable in this case is binary, therefore a logistic regression should be used. I will use a GLM in this case.

I begin by doing exploratory analysis of the data to assess correlation between my variables:



From this plot, we can see that mean radius (r=0.73), mean perimeter (r=0.74) and mean area (0.71) has the highest correlation with whether the tumor is malignant or not. However, all of these variables are, expectedly, highly correlated with one another (r=0.99 – r=1). As such, only one of them should be used. Texture (r=0.42) and smoothness (r=0.36) should be the other two variables that are used. **All factors are significantly correlated with malignance.**

Malignant~perimeter+texture+smoothness

Check assumptions:

1. We have to assume that observations are independent.
2. We have to assume that the link transformed variable has a linear relationship with the independent variables.

Running the model, I obtain the following output:

Call:

glm(formula = malignant ~ texture\_mean + perimeter\_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 \*\*\*

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

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

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

Interpreting the output: From this table, we see that the probability for a tumour to be malignant is significantly influenced by all three of our independent variables. When all three independent variables are at 0, the log odds of the tumour being malignant is -40.4. With one unit increase in mean texture, the log odds increase by 0.373. Log odds increase by 0.210 and 134.025 respectively for perimeter and smoothness. The very high log odds for increase in smoothness value is because smoothness values in the data are all very small (0.1 or less). In order to obtain the probability of the cancer being malignant, we can inverse transform the log odds with the expit function at the values we are interested in. For example, the probability that a tumour is malignant where all three independent variables are zero is 2.839e-18 – very close to zero.

Check model:

To check if the model captures the patterns in the data well, I compared it to a null model with an ANOVA test. I obtained the following output:

> anova(cancer.mod0, cancer.mod1, test="Chi")

Analysis of Deviance Table

Model 1: malignant ~ 1

Model 2: malignant ~ texture\_mean + perimeter\_mean + smoothness\_mean

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 568 751.44

2 565 181.68 3 569.76 < 2.2e-16 \*\*\*

---

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

Based on the output, I can tell that the model is significantly different from the null model (p<2.2e-16). The probability for the null model to replicate the same deviance as the previous model is very low at <2.2e-16.

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.**

Using an R package that includes a function to assess relative importance of variables for logistic regressions, I obtained the following output:

> varImp(cancer.mod1)

Overall

texture\_mean 6.466150

perimeter\_mean 8.993087

smoothness\_mean 7.153568

This package measures importance using the absolute t-statistic for each model parameter. Based on this output, we see that perimeter\_mean is most important, followed by smoothness and then texture.

To check, I tried with another package, which has a function to use standardized beta coefficients to assess relative importance:

> lm.beta(cancer.mod1)

texture\_mean perimeter\_mean smoothness\_mean

3.321510 10.548990 3.895173

Again, we get the same results that perimeter\_mean is most important, followed by smoothness and then texture. We see here, for example, that a change in 1 standard deviation of perimeter\_mean has approximately three times the impact that a change in 1 standard deviation of texture\_mean.