Molly Watters

NRE 538 Final Exam

Take Home

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***

**Test**: Chi-squared, because we are finding out if there is an association between categorical independent and categorical dependent variables.

Pearson's Chi-squared test

data: flying$gender and flying$unruly\_child

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

**Assumptions**:

*Independent Observations:* I assume that observations were made independently because each observation was taken from a different individual.

*No structural zeros:* I assume that every cell of my table can be filled as every individual surveyed could have an answer to each question asked and N/As do not affect my results.

**Results/Interpretation:** The chi-squared test showed a p-value of 0.001193 which is less than 0.05, so there is a significant association between gender and whether people think it's rude to bring an unruly child on the plane. By adding up the number of males who answered no, somewhat, and very, and dividing the number who answered “very” by that total, we see that the number of males who think it is very rude to bring an unruly child on a plane is 47.38%. We compare this with 35.74% of females who think it is very rude. This suggests that men tend to think it is more rude (with a higher percentage thinking it's very rude) to bring an unruly child on an airplane than do women.

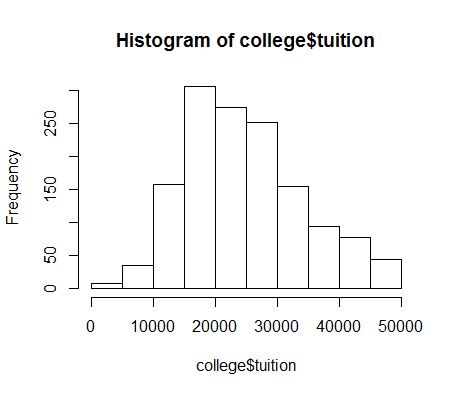
1. *Is there a significant difference in tuition (tuition) by type of institution (type)? If yes, which type has a higher tuition?* ***College***

**Test:** Two sample t-test, because we have a continuous dependent variable (tuition) and categorical independent variable (type). It is a two sample t-test because we are comparing the tuitions of type of institution as two samples, public and private nonprofit.

**Assumptions:***Continuous data:* Tuition is a continuous variable, so this assumption is met.

*Random sampling:* We can assume that the data was randomly sampled from students (I assume that’s the unit of measurement in this dataset) from both types of institutions.

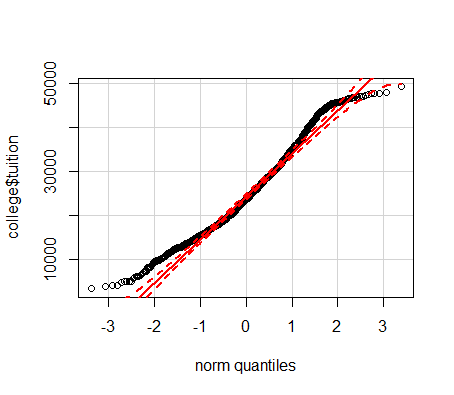
*Normal distribution:* Based on the histogram below the data looks normally distributed. However, the Shapiro test shows a low p-value so that suggests that we must reject our null hypothesis of normality and we cannot assume our data is normally distributed. The qqplot below looks good though so we will consider this assumption met.



Shapiro-Wilk normality test

data: college$tuition

W = 0.97407, p-value = 3.193e-15



*Equal variance:* Based on the variance test below, we cannot meet the assumption of equal variances between the two samples because the p-value is less than 0.05 (<2.2e-16) so we must reject the null that they are equal. We can address this by using the Welch’s t-test.

var.test(type.pub[,"tuition"], type.priv[,"tuition"])

F test to compare two variances

data: type.pub[, "tuition"] and type.priv[, "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

*Independent observations:* We can assume observations were collected independently because each student (still assuming that’s the unit of measurement) is unique.

**Run & interpret the model:**  The t-test results show a p-value of less than 2.2e-16 which is less than 0.05 and means we can reject the null hypothesis that tuition is the same for the two samples. Therefore we can say there is a significant difference between tuition at public and private nonprofit schools.

Based on the histogram below and the sample estimates in the Welch’s t-test (mean tuition of 18855.88 for public institutions and 28301.69 for private nonprofit institutions), private nonprofit institutions have a higher tuition than public institutions.

Welch Two Sample t-test

data: type.pub[, "tuition"] and type.priv[, "tuition"]

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

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

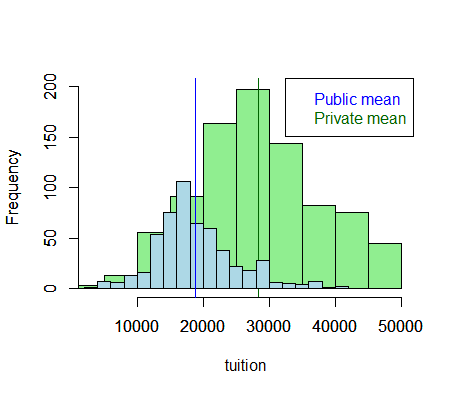
95 percent confidence interval:

-10258.878 -8632.749

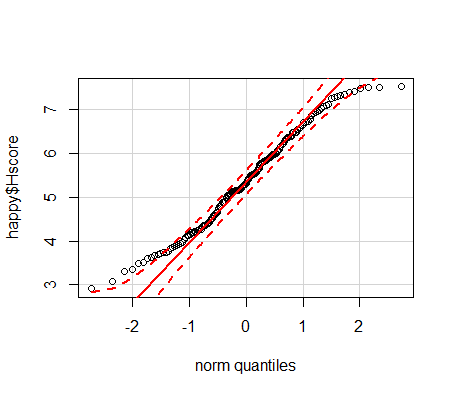
sample estimates:

mean of x mean of y

18855.88 28301.69



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

**Test:** One-way ANOVA, because we have one categorical independent variable (region) with four levels for different regions and one continuous dependent variable of happiness score.

**Assumptions:**

*Normal distribution*: Using a shapiro test we see that the p-value is smaller than 0.05 (0.0123) which means we would reject the null hypothesis of normality. However, the qqplot looks good and the sample size is large, so we can treat this assumption as met.

*Independent samples*: We can assume that happiness scores are independent

because they were calculated based on separate data from each country.

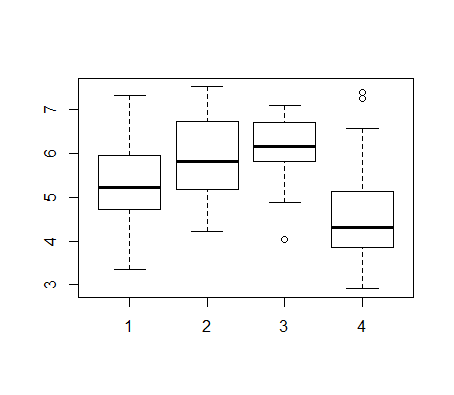
*Equal variance of populations:* The Levene’s test shows a high p-value of 0.52, which means variances are equal.

Levene's Test for Homogeneity of Variance (center = median)

Df F value Pr(>F)

group 3 0.7605 0.5179

153



**Results:**

A boxplot shows that AmericasCarribean has the highest mean happiness score (labeled #3) followed by Europe, AsiaAustralia, and AfricaMideast. The fact that several of the boxes overlap suggests there may not be significant differences between them, however. To test this, we use a one way ANOVA (see below). Since the p-value is smaller than 0.05 we can say that there is a significant difference between in happiness by region.

> summary(happy.aov)

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

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

**Test**: Linear model, because we are trying to find out what continuous independent factors are significantly associated with a continuous dependent variable. First I will test the correlation of the possible factors and then I will choose ones that look suitable for my model. I will choose freedom, generosity, and GDP as the continuous independent variables in my model because they have the highest correlation with Corruption. This is visually apparent as well in the pairs plot. I can also see in the pairs plot and by looking at the correlation coefficients that, with the exception of family and GDP which is higher, these factors are not highly correlated with each other. As I discover when I test my assumptions, I will first need to log transform my dependent variable of corruption so that I can run my linear model.

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

Corruption GDP Family Life

Corruption 1.0000000 0.29418478 0.21356094 0.24958329

GDP 0.2941848 1.00000000 0.66953969 0.83706723

Family 0.2135609 0.66953969 1.00000000 0.58837678

Life 0.2495833 0.83706723 0.58837678 1.00000000

Freedom 0.5020540 0.36228285 0.45020820 0.34119929

Generosity 0.3059299 -0.02553066 0.08962885 0.07598731

Freedom Generosity

Corruption 0.5020540 0.30592986

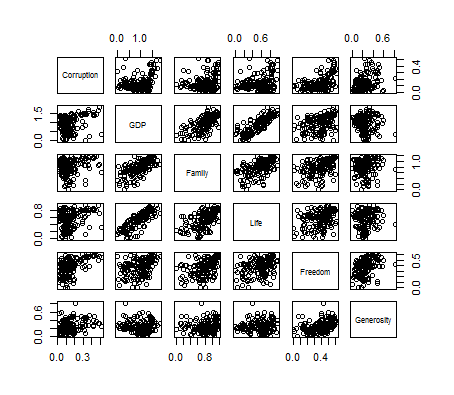
GDP 0.3622828 -0.02553066

Family 0.4502082 0.08962885

Life 0.3411993 0.07598731

Freedom 1.0000000 0.36175133

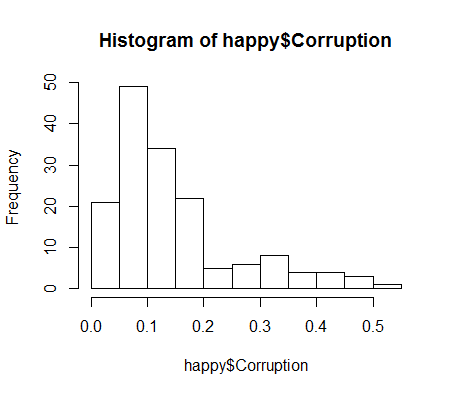
Generosity 0.3617513 1.00000000



**Assumptions:**

*Normal errors:* First we can look at a histogram of the dependent variable as a proxy for what the errors will look like after I run my model. The data does not look very normally distributed. I then use a shapiro test to further determine if the dependent variable is normally distributed. Since the p-value is less than 0.05, we must say that the data is not normally distributed because we reject the null of normality. This means that I will need to log transform my data. Once I do the transformation, both the histogram and shapiro test show that my data is normal (shapiro test p-value is 0.085, so I do not reject the null of normality). See below for plots and shapiro tests before and after transforming my data.

Before transforming:

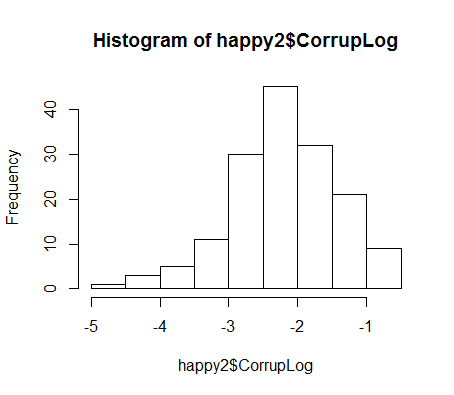


Shapiro-Wilk normality test

data: happy$Corruption

W = 0.85796, p-value = 5.128e-11

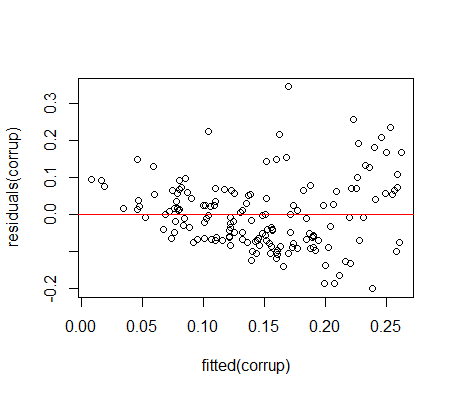
After Transforming:



Shapiro-Wilk normality test

data: happy2$CorrupLog

W = 0.9849, p-value = 0.08535

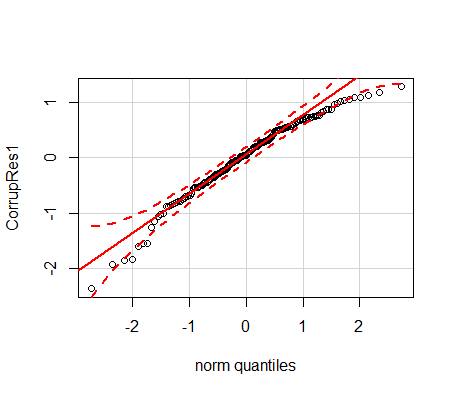


Once I test the residuals for normality with their own shapiro test, I see that errors are not normal because the p-value is small (0.00053) which means I must still reject the null of normal distribution. However, the qqplot looks good so I will consider this assumption met.

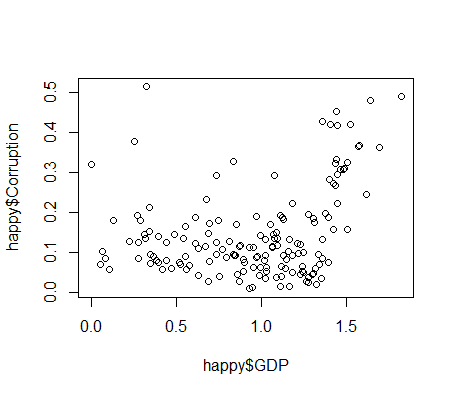
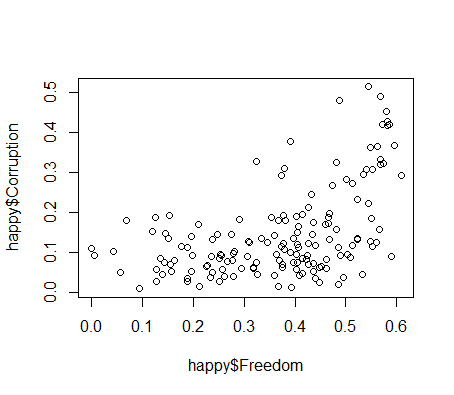
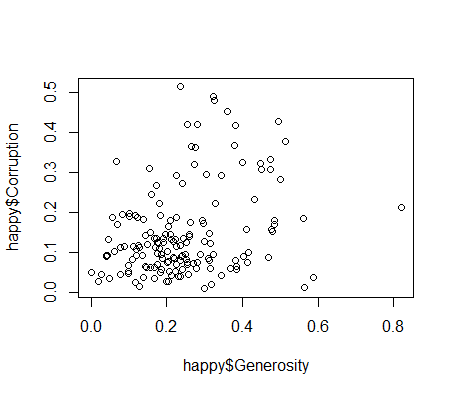
Shapiro-Wilk normality test

data: CorrupRes1

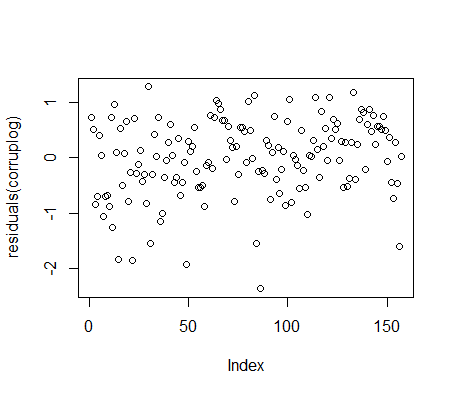
W = 0.96589, p-value = 0.0006382



*Linear relationship:* Based on plots below, the relationships appear to be linear.



*Independent errors/no autocorrelation:* Errors are independent because they are not correlated with each other by something like time or location. The Durbin-Watson test shows a p-value of above 0.05 (0.37) which means we cannot reject the null that errors are independent; so we can say that they are independent and there is no autocorrelation.



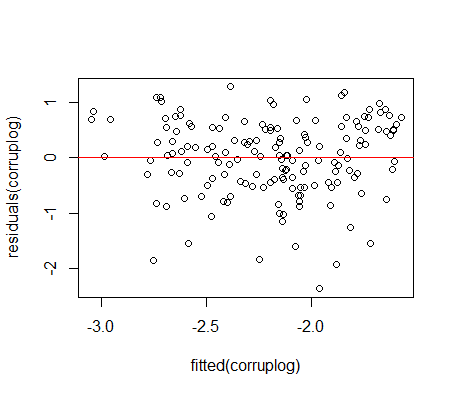
Durbin-Watson test

data: corruplog

DW = 1.8727, p-value = 0.3701

alternative hypothesis: true autocorrelation is not 0

*Homoscedasticity:* The residuals look evenly dispersed based on the plot below, and the bp test gives a p-value above 0.05 (0.068) which means we cannot reject the null hypothesis of homoscedasticity. We can say the assumption of homoscedasticity is met.

**

studentized Breusch-Pagan test

data: corruplog

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

**Run & Interpret the model:** Now I actually run my linear model with Freedom, Generosity, and GDP as independent variables and corruption as dependent variable. I see freedom is the only factor that significantly influences my model (freedom has a p-value of 2.95e-06); generosity and GDP are not significant with p-values of 0.226 and 0.768 respectively. With every unit increase in freedom, the log of corruption increases by 2.166.

Call:

lm(formula = CorrupLog ~ Freedom + Generosity + GDP, data = happy2)

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 \*\*\*

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

Generosity 0.54962 0.45252 1.215 0.226

GDP 0.04341 0.14673 0.296 0.768

---

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 Checking:** Based on the AIC test below, my model is not the best fit model. This makes sense as only one of my variables was significant: freedom. It follows that the model only containing freedom had the lowest AIC and is therefore the best predictor of log corruption. The models containing freedom and generosity or GDP in fact were also better predictors than the model containing all three (corruplog) used in this exam question.

> AIC(corruplog,corruplog1,corruplog2,corruplog3,corruplog4,corruplog5,corruplog6)

df AIC

corruplog 5 336.4578

corruplog1 4 334.5476

corruplog2 4 335.9643

corruplog3 4 356.9480

corruplog4 3 333.9705

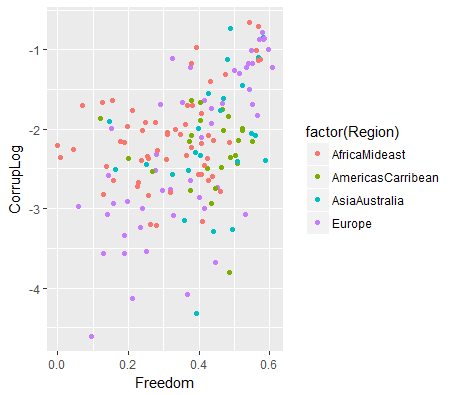
corruplog5 3 365.0821

corruplog6 3 360.1237

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***

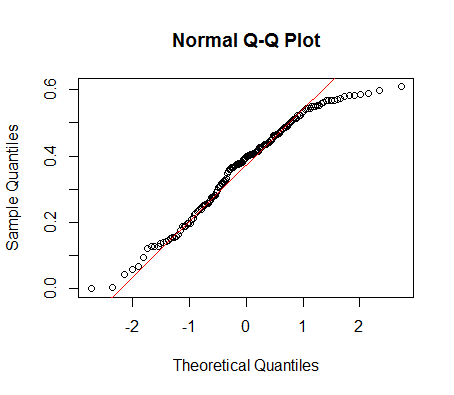
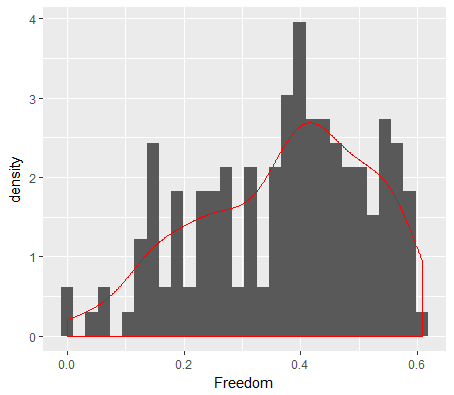
**Test:** I will use freedom as the independent variable because it was the factor that was significantly associated with corruption. I will use an ANCOVA because my model has a continuous dependent variable but both continuous and categorical independent variables. I will also include interaction with region in my model. I will use the log transformed version of my corruption data because I know it will work well based on the previous question.

**Assumptions:**

*Linear relationship:* Based on the plot below, it looks like there is a linear relationship between corruption and freedom across regions. 

*Normal distribution:* Freedom does not officially appear to be normally distributed based on the p-value of 0.0003 shown by the Shapiro test, but the histogram and qqplot looks good. Corruption, which has already been log transformed, is normally distributed; the qqplot, histogram, and p-value of 0.08 in the Shapiro test all confirm this assumption is met.

Freedom:

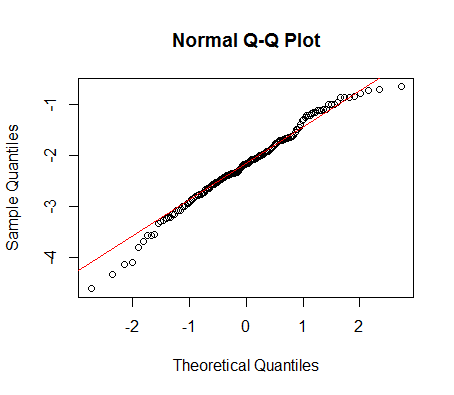
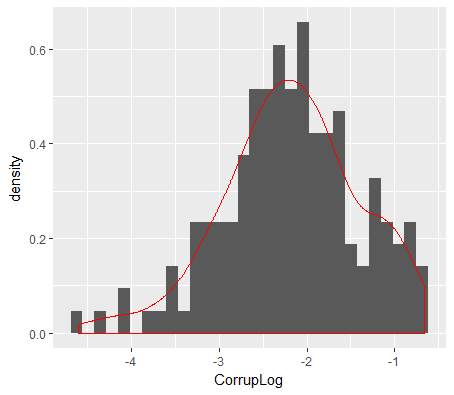


Shapiro-Wilk normality test

data: happy2[, "Freedom"]

W = 0.96245, p-value = 0.0002907

Corruption:



Shapiro-Wilk normality test

data: happy2[, "CorrupLog"]

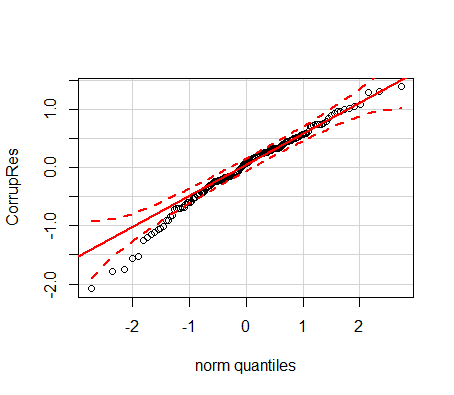
W = 0.9849, p-value = 0.08535

We can also test the residuals themselves for corruption to see if they are normally distributed; based on the Shapiro test’s p-value of 0.015 (less than 0.05) we must reject the null of normal distribution. However, the qqplot below looks good so we will deem this assumption met.

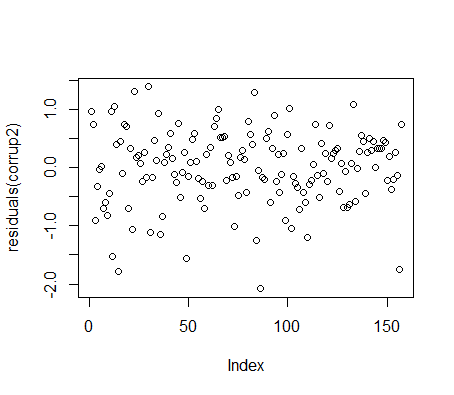
Shapiro-Wilk normality test

data: CorrupRes

W = 0.97848, p-value = 0.01481



*Independent Errors/No Autocorrelation:* There does not appear to be any autocorrelation in the data based on the residuals plot below where residuals are evenly dispersed with no apparent patterns, and the Durbin-Watson test that shows a p-value of 0.56 which means we cannot reject the null of autocorrelation being 0 and deem this assumption met.



Durbin-Watson test

data: corrup2

DW = 2.1408, p-value = 0.5627

alternative hypothesis: true autocorrelation is not 0

*Homoscedasticity:* The data appears to be homoscedastic, because the residuals appear to be evenly dispersed from the line. Additionally, the BP test shows a high p-value of 0.25 which means we cannot reject the null hypothesis of homoscedasticity; so we can say that the assumption of homoscedasticity is met.

studentized Breusch-Pagan test

data: corrup2

BP = 9.0359, df = 7, p-value = 0.2501

**Run and interpret the model:**

In the output, due to alphabetical ordering, the intercept is corruption in the AfricaMideast region when freedom is zero. This intercept of AfricaMideast is significant with a p-value of less than 2e-16, and the estimate shows that with every unit increase in AfricaMideast, the corruption variable decreases by 2.6096 units. The region of Europe was significant alone as well, with a p-value of 2.42e-05, and the estimate shows that with every unit increase in Europe, corruption decreases by 1.376. The interaction of freedom and the AfricaMideast region on corruption is also significant, with a p-value of 0.0094; the estimate suggests that with every unit increase in freedom in AfricaMideast, corruption increases by 1.593 units. Also significant is the interaction of freedom and the region of Europe, with a p-value of 0.00049; the estimate suggests that with every unit increase in freedom in Europe, corruption increases by 2.98 units. These significant interaction outputs mean that the effect of freedom depends on region for AfricaMideast and Europe. However, AmericasCarribean and AsiaAustralia did not significantly predict corruption alone or interacting with freedom.

Model: corrup2 = lm(CorrupLog~Freedom\*Region, data=happy2)

lm(formula = CorrupLog ~ Freedom \* Region, data = happy2)

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 \*\*

Region[T.AmericasCarribean] 0.3985 0.5898 0.676 0.500360

Region[T.AsiaAustralia] -0.5177 0.5410 -0.957 0.340110

Region[T.Europe] -1.3755 0.3155 -4.360 2.42e-05 \*\*\*

Freedom:Region[T.AmericasCarribean] -1.6566 1.3882 -1.193 0.234643

Freedom:Region[T.AsiaAustralia] 0.6657 1.2700 0.524 0.600949

Freedom:Region[T.Europe] 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

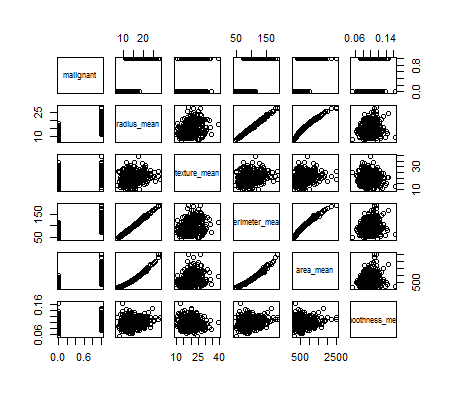
Multiple R-squared: 0.3365, Adjusted R-squared: 0.3053

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

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***

**Test:** I will need to use a logistic model with a GLM function because I have a categorical dependent variable and continuous independent variables.

First, I have to decide which three factors are most important to include in my model so I use the correlation function to see which factors are most closely correlated with whether a breast cancer tumor is malignant or not. It looks like perimeter mean, radius mean, and area mean are the most highly correlated, but they are also very highly correlated with each other. This is also apparent in the pairs plot below. So I will not choose all three, and will choose **perimeter** mean since it is the most highly correlated with malignancy. I will then select the next highest or remaining factors of **texture** mean and **smoothness** mean to use in my model.



> cor(cancer)

malignant radius\_mean texture\_mean

malignant 1.0000000 0.7300285 0.41518530

radius\_mean 0.7300285 1.0000000 0.32378189

texture\_mean 0.4151853 0.3237819 1.00000000

perimeter\_mean 0.7426355 0.9978553 0.32953306

area\_mean 0.7089838 0.9873572 0.32108570

smoothness\_mean 0.3585600 0.1705812 -0.02338852

perimeter\_mean area\_mean smoothness\_mean

malignant 0.7426355 0.7089838 0.35855997

radius\_mean 0.9978553 0.9873572 0.17058119

texture\_mean 0.3295331 0.3210857 -0.02338852

perimeter\_mean 1.0000000 0.9865068 0.20727816

area\_mean 0.9865068 1.0000000 0.17702838

smoothness\_mean 0.2072782 0.1770284 1.00000000

**Assumptions:**

*Linear relationship for transformed dependent variable:*

**Run and Interpret the Model:**

All three factors are significantly associated with whether a breast cancer tumor is malignant or not. The intercept and perimeter had a p-value of less than 2e-16, texture had a p-value of 1.01e-10, and smoothness had a p-value of 8.48e-13. For every unit increase in perimeter mean the log odds of malignancy increases by 0.21. For every unit increase in texture, the log odds of malignancy increase by 0.3737. For every unit increase in smoothness, the log odds of malignancy increase by 134.025.

|  |
| --- |
| 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 |
|  |
| |  | | --- | |  | |

**Model checking:** Based on the anova test below, our model with all three factors (mod1) is the best predictor because it has the lowest residuals at 181.68. The AIC test supports this result because our model with all three factors (mod1), has the lowest AIC of 189.68.

> anova(mod0,mod1,mod2,mod3,mod4,mod5,mod6,mod7)

Analysis of Deviance Table

Model 1: malignant ~ 1

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

Model 3: malignant ~ perimeter\_mean

Model 4: malignant ~ texture\_mean

Model 5: malignant ~ smoothness\_mean

Model 6: malignant ~ perimeter\_mean + smoothness\_mean

Model 7: malignant ~ texture\_mean + smoothness\_mean

Model 8: malignant ~ texture\_mean + perimeter\_mean

Resid. Df Resid. Dev Df Deviance

1 568 751.44

2 565 181.68 3 569.76

3 567 304.48 -2 -122.81

4 567 646.52 0 -342.03

5 567 673.95 0 -27.43

6 566 240.67 1 433.28

7 566 546.40 0 -305.74

8 566 268.08 0 278.33

> AIC(mod0,mod1,mod2,mod3,mod4,mod5,mod6,mod7)

df AIC

mod0 1 753.4400

mod1 4 189.6754

mod2 2 308.4844

mod3 2 650.5191

mod4 2 677.9485

mod5 3 246.6674

mod6 3 552.4047

mod7 3 274.0782

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.**

Based on the correlation test from the beginning of question 6, perimeter, area, and radius have the highest correlations with whether a tumor is malignant or not. However, they are also highly correlated and this question is asking to stick to the same 3 continuous independent variables that I chose in question 6. Perimeter, texture, and smoothness are all significantly associated with whether a tumor is malignant or not and contributed to a well-fitting model as compared with different permutations of the three variables. So I would say they are all important, but smoothness appears to be the most important because it has the highest potential to have an impact on malignancy. The estimate for smoothness is 134.025 as compared with 0.21 for perimeter and 0.37 for texture. This means that the log odds of malignancy increase by much, much more for smoothness than for perimeter and texture.