Erin Barton

23 April 2017

Natural Resource Statistics–Final Take Home Exam

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**
   1. **Hypothesis:** 
      1. H0: There is not a significant association between gender and whether people think it’s rude to bring an unruly child on the plane.
      2. Ha: There is not a significant association between gender and whether people think it’s rude to bring an unruly child on the plane.
   2. **Variables**
      1. Gender: categorical, independent variable
      2. Unruly Child: categorical, dependent variable
   3. **Test: Chi-Squared test of independence**
      1. Assumptions/requirements
         1. Observations are independent
         2. No structural 0s
         3. <20% of cells have expected values <5

table1 <- table(flying$unruly\_child, flying$gender)

Female Male

No 91 56

Somewhat 193 155

Very 158 190

* + 1. transformations? –> no
    2. run the test:
       1. chisq.test(flying$unruly\_child, flying$gender)

Pearson's Chi-squared test

data: flying$unruly\_child and flying$gender

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

* + 1. interpret test results:
       1. The results show that there is a significant association between gender and whether you think it’s rude to bring an unruly child on the plane. Men tend to think it is very rude than women do, although women are tend to think it is somewhat rude more than men do. Women tend to think it is not rude at all more so than men.

1. Is there a significant difference in tuition (tuition) by type of institution (type)? If yes, which type has a higher tuition? **College**
   1. **Hypothesis**
      1. H0: There is no significant difference in tuition based on type of institution
      2. Ha: There is a significant difference in tuition based on type of institution.
   2. **Variables**
      1. Tuition: continuous, dependent variable
      2. Type of Institution: categorical, independent variable
   3. **Test:** Two Sample T-Test
      1. Assumptions/requirements
         1. Data are continuous
         2. Samples randomly selected from population
         3. Observations are independent
         4. Values are nearly normal or N≥30
            1. N>30
         5. Variance is equal between two populations
            1. priv.college <- subset(college, type=="Private nonprofit")

pub.college <- subset(college,type=="Public")

* + - * 1. var.test(priv.college$tuition, pub.college$tuition)

p-value < 2.2e-16 , which means the null hypothesis, that the variance is equal, can be rejected.

* + 1. Transformation: A Welch’s Two Sample T-Test should be run instead of a regular t-test as the variance between groups is not equal.
    2. Run test: t.test(priv.college$tuition, pub.college$tuition, paired=FALSE)
       1. Welch Two Sample t-test

data: priv.college$tuition and pub.college$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:

8632.749 10258.878

sample estimates:

mean of x mean of y

28301.69 18855.88

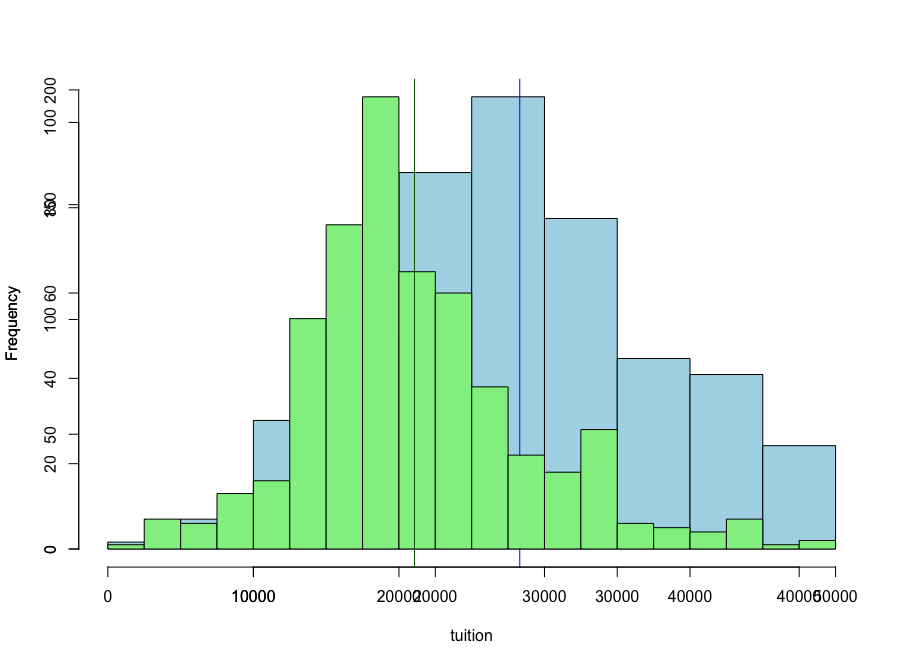
* + - 1. This test shows that Ha is supported: there is a significant difference in tuition by type of institution. This can be seen by how the p-value is less than .05. Using a histogram plot, we can see that the mean tuition of private colleges is higher than the mean tuition of public nonprofit colleges.
         1. hist(priv.college$tuition, breaks=15, col="light blue", xlab="tuition", main="")

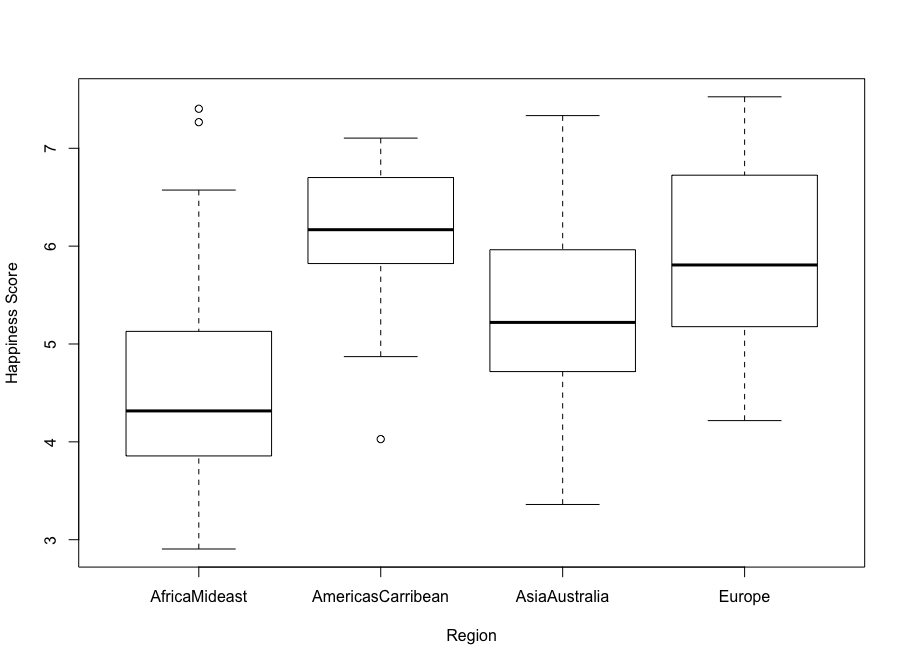
abline(v=mean(priv.college$tuition), col="blue")

par(new=TRUE)

hist(pub.college$tuition, breaks=15, col="light green", xlab="", main="")

abline(v=mean(pub.college$tuition), col="dark green")



1. Is there a significant difference in happiness (Hscore) by region (Region)? **Happy**
   1. **Variables:** 
      1. Happiness: continuous, dependent
      2. Region: categorical (4 levels), independent
   2. **Test: One-Way ANOVA**
      1. Visualize the data:
         1. boxplot(Hscore~Region, data=happy, xlab="Region", ylab="Happiness Score")
         2. 
      2. Assumptions
         1. Population must be normally distributed or N≥30
            1. N>30
         2. Samples must be independent
         3. Each population must have same variance
            1. leveneTest(Hscore~Region, data=happy)

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

Df F value Pr(>F)

group 3 0.7605 0.5179

153

p-value>.05, indicating that the population is of equal variance

* + 1. Run test:
       1. modaov <- aov(Hscore~Region, data=happy)

summary(modaov)

* + - * 1. 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

* + - * 1. The aov test shows that there is a significant difference in happiness by region.
      1. TukeyHSD(modaov
         1. 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

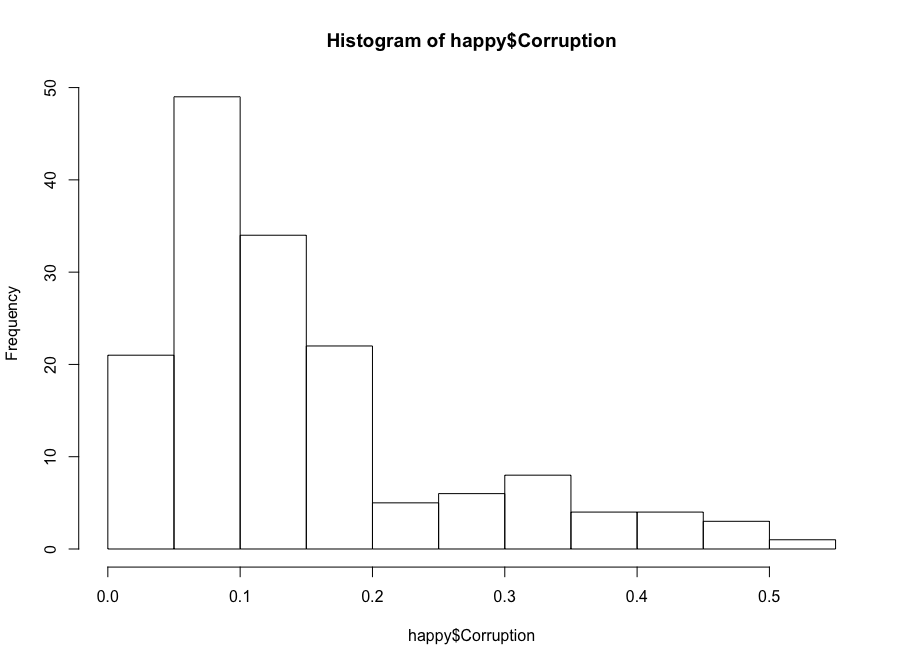
* + - * 1. The Tukey’s HSD test suggests that the relationship between happiness and region is not significant for Europe-AmericasCarribean and Europe–AsiaAustralia, but it is significant between all other groups.

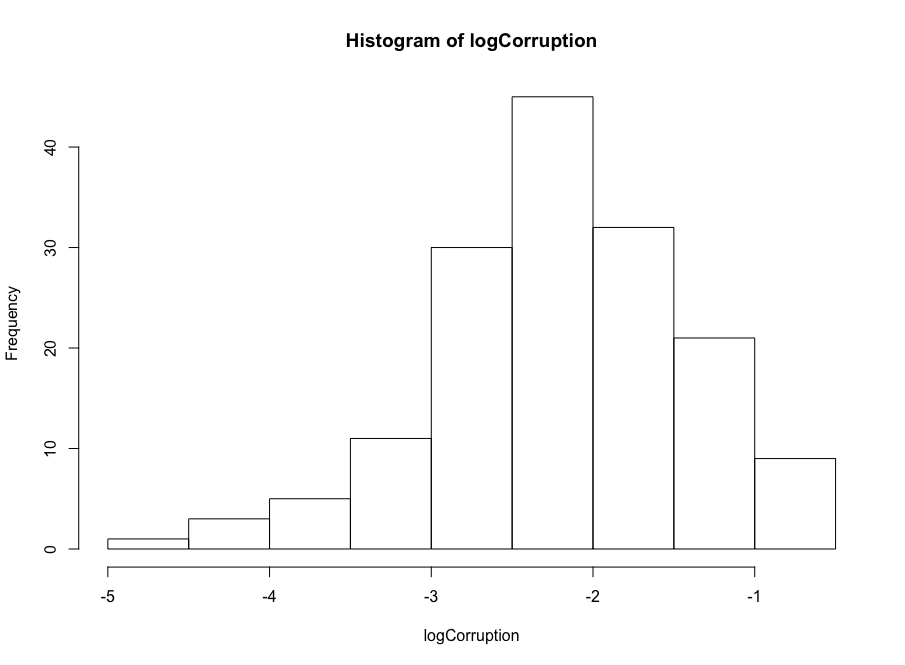
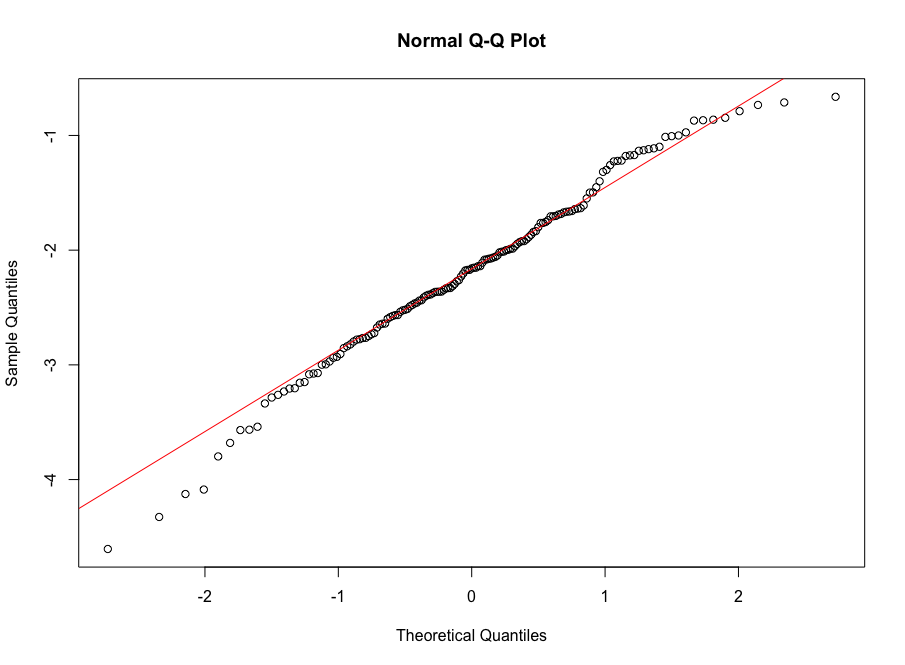
1. What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model. **Happy**
   1. **Hypothesis:** 
      1. H0: Corruption is not significantly associated with Hscore, Generosity or GDP.
      2. Ha: Corruption is significantly associated with Hscore, Generosity or GDP.
   2. **Variables:**
      1. Corruption levels: continuous, dependent variable
      2. Hscore, Generosity, GDP: continuous, independent variables
   3. **Test: Multiple Linear Regression**
      1. Check collinearity
         1. Library(car)
         2. vif(lm(Corruption~Hscore+Generosity+GDP, data=happy))

Hscore Generosity GDP

2.906700 1.091861 2.837041

* + - 1. The VIF shows that none of the selected variables are highly correlated as vif<5 for all of them.
    1. Histogram:
       1. hist(happy$Corruption)



* + - 1. Histogram indicates that the data is not normally distributed and should be transformed
    1. Tranformations:
       1. Log transformation: logCorruption<–log(happy$Corruption)
          1. 
          2. 
          3. The histogram and qqplot suggest that the log transformation made the data more normal and good enough to run through the regression analysis.
    2. Run test:
       1. mod1<–lm(logCorruption~Hscore+Generosity+GDP, data=happy)
          1. Call:

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

Residuals:

Min 1Q Median 3Q Max

-2.45521 -0.43889 0.07901 0.48115 1.86303

Coefficients:

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

(Intercept) -3.6062 0.3110 -11.595 < 2e-16 \*\*\*

Hscore 0.2614 0.0865 3.023 0.00294 \*\*

Generosity 1.0290 0.4525 2.274 0.02435 \*

GDP -0.2477 0.2365 -1.047 0.29660

---

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

Residual standard error: 0.7234 on 153 degrees of freedom

Multiple R-squared: 0.1404, Adjusted R-squared: 0.1236

F-statistic: 8.333 on 3 and 153 DF, p-value: 3.616e-05

* + - 1. The model indicates that there is a significant positive relationship between the log of corruption and happiness, and the log of corruption and generosity. So When happiness score goes up one unit, corruption increases by the log of 0.2614, and the same is true of Generosity. There is not significant relationship between the log of corruption and GDP.
    1. Assumptions
       1. Linear relationship
          1. The relationship are linear

plot(df)

* + - 1. Independent errors
         1. dwtest(mod1, alternative=c(“two.sided”))

Durbin-Watson test

data: mod1

DW = 1.8485, p-value = 0.286

alternative hypothesis: true autocorrelation is not 0

* + - * 1. The Durbin Watson test returns a p-value>.05, indicating that the errors are independent.
      1. Homoscedasticity
         1. bptest(mod1)

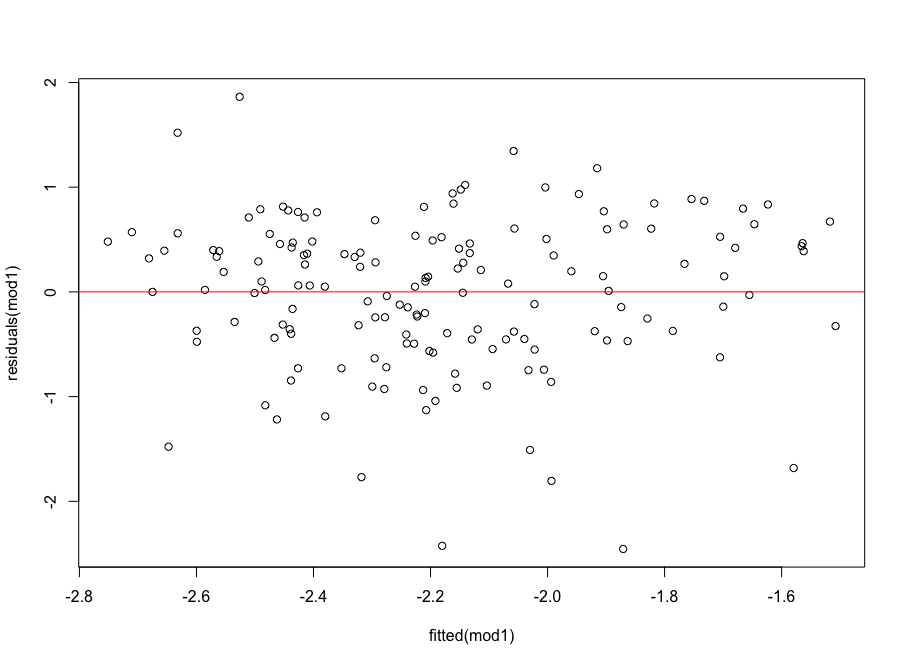
studentized Breusch-Pagan test

data: mod1

BP = 7.6033, df = 3, p-value = 0.05496

* + - * 1. The Bresch-Pagan test shows that there is a borderline possibility of heteroscedasticity as the p-value = .055. A plot of the residuals, though, shows that there does not seem to be a problem with heteroscedasticity in the errors.

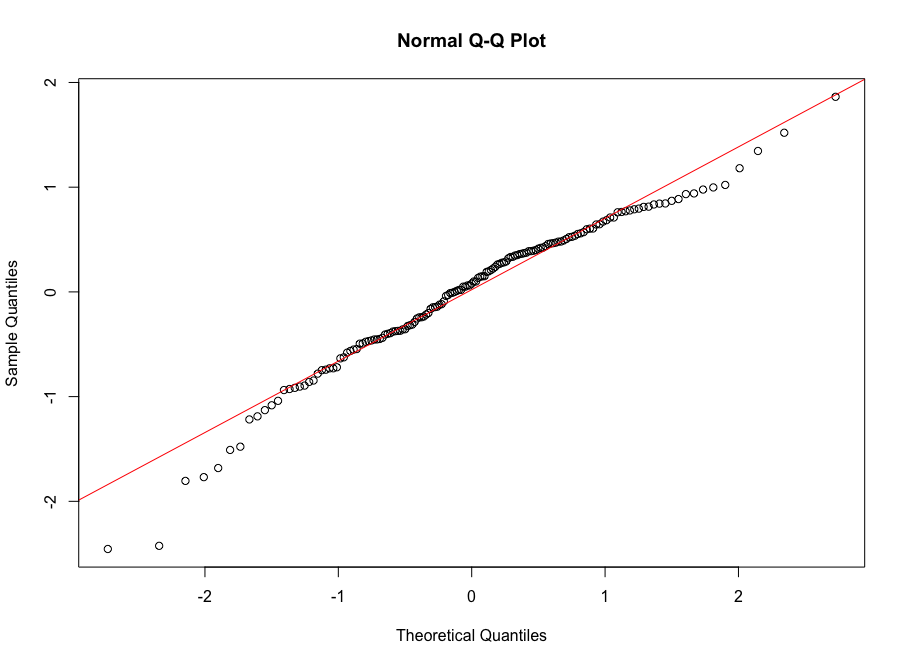
plot(residuals(mod1)~fitted(mod1))

abline(lm(residuals(mod1)~fitted(mod1,col="red")

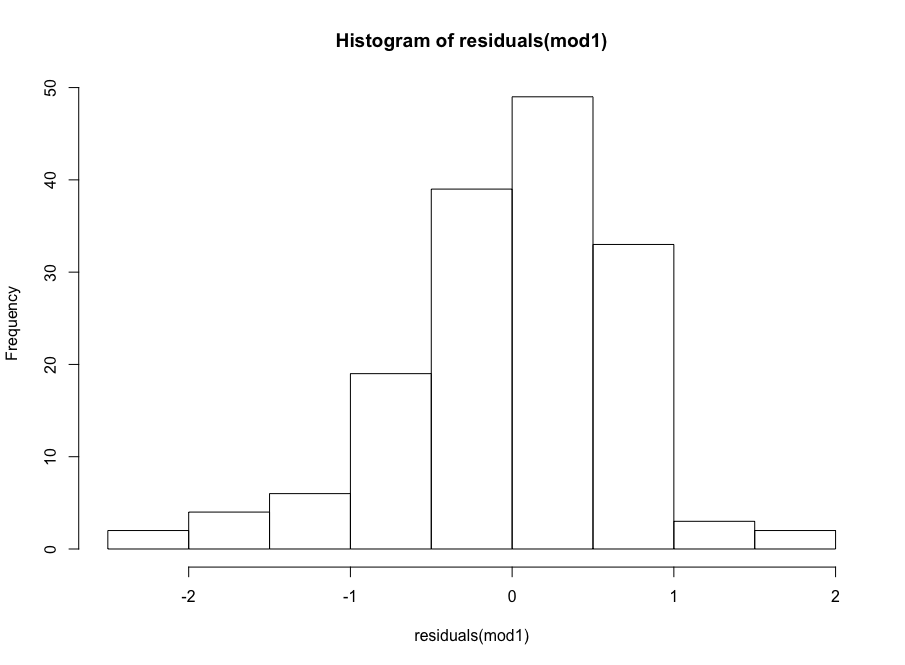
* + - 1. Normal errors
         1. A qqplot and a histogram both show that the errors appear nearly normal, even though the shapiro.wilkes test suggest that the errors are not normal (p-value<.05)

qqnorm(residuals(mod1))

qqline(residuals(mod1), col="red")



hist(residuals(mod1))



shapiro.test(residuals(mod1))

data: residuals(mod1)

W = 0.96973, p-value = 0.001588

* 1. **Check model fit**
  2. Based on the check of the assumptions, the model is a decent fit for the data. It does not grossly violate any assumptions even if it does not fit within the bounds of the more conservative tests, such as the Shapiro-Wilkes test. The adjusted R-Squared value, however, is not very high. It suggests that the model only explains 12% of the variation. To check and see if this can be higher, the insignificant GDP can be taken out. This results in an R-Squared of .123, which is not any better than the other model. The AIC test also shows no real difference in the two models.

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**
   1. mod3<–lm(logCorruption~Hscore\*Region, data=happy)
      1. Call:

lm(formula = logCorruption ~ Hscore \* Region, data = happy)

Residuals:

Min 1Q Median 3Q Max

-2.16163 -0.37780 0.08123 0.39173 2.29089

Coefficients:

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

(Intercept) -3.08241 0.41246 -7.473 6.17e-12 \*\*\*

Hscore 0.21281 0.08753 2.431 0.016236 \*

RegionAmericasCarribean 0.20804 1.22562 0.170 0.865446

RegionAsiaAustralia -0.52791 0.86694 -0.609 0.543494

RegionEurope -3.20486 0.74055 -4.328 2.75e-05 \*\*\*

Hscore:RegionAmericasCarribean -0.10927 0.20613 -0.530 0.596823

Hscore:RegionAsiaAustralia 0.05929 0.16528 0.359 0.720319

Hscore:RegionEurope 0.46548 0.13485 3.452 0.000725 \*\*\*

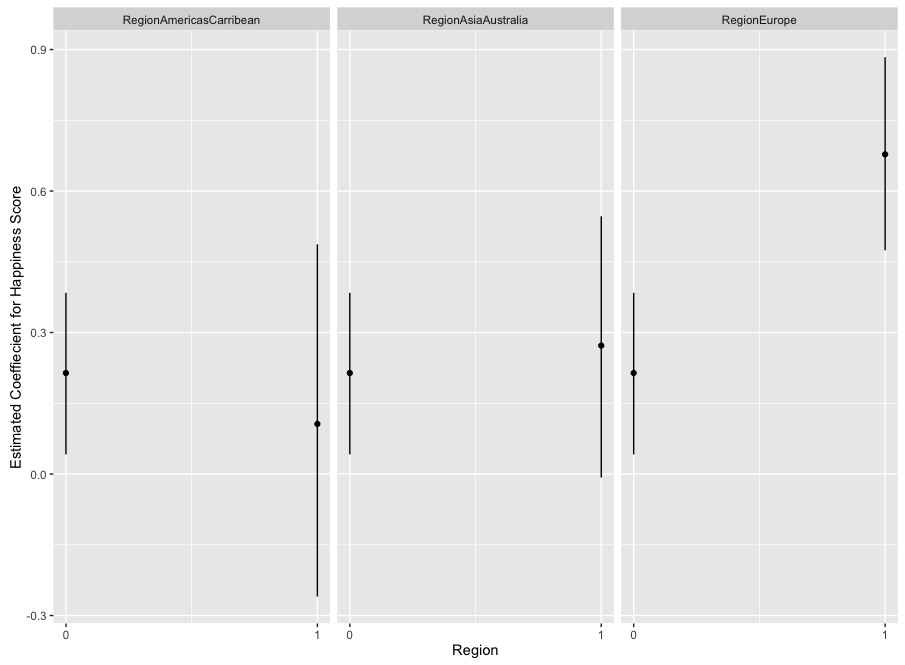
---

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

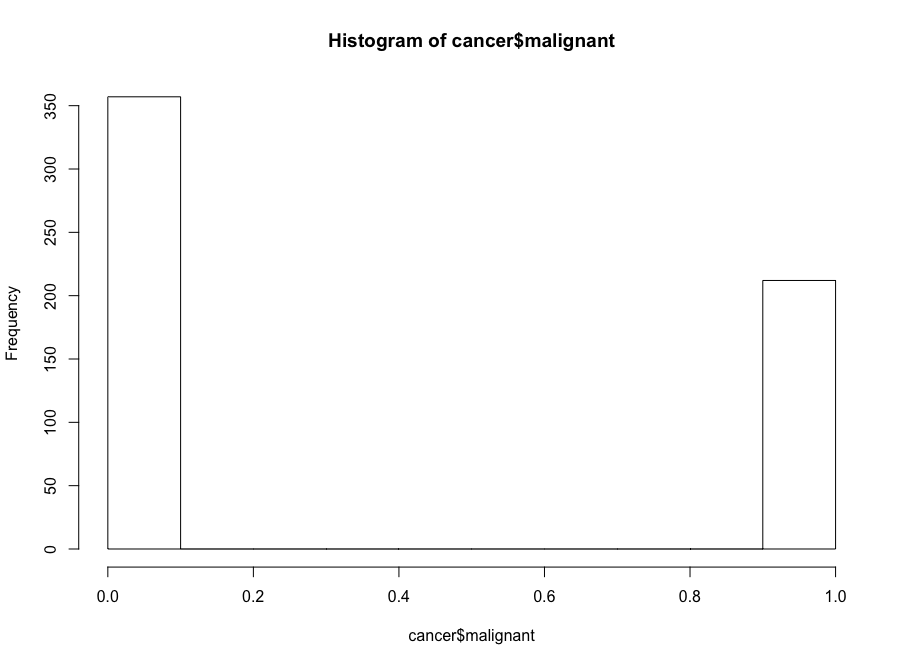
Residual standard error: 0.6749 on 149 degrees of freedom

Multiple R-squared: 0.2716, Adjusted R-squared: 0.2374

F-statistic: 7.936 on 7 and 149 DF, p-value: 3.588e-08

* 1. The model shows that the influence of happiness score does vary by region. The effect of Hscore on corruption increases in Europe compared to Africa and the Middle East by the log of 0.46548.
  2. An interplot further shows that the effect of Hscore on the log of Corruption is only significant in Europe
     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**
   1. **Hypothesis:** 
      1. H0: Mean Radius, Mean Texture, Mean Smoothness are not significantly associated with whether a breast cancer tumor is malignant or not.
      2. Ha: Mean Radius, Mean Texture, Mean Smoothness are significantly associated with whether a breast cancer tumor is malignant or not.
   2. **Variables:**
      1. malignant: binomial categorical, dependent variable
      2. Mean Radius, Mean Texture, Mean Smoothness: continuous, independent variables
   3. **Test: Generalized Linear Model–Binomial**
      1. Histogram:
         1. hist(cancer$malignant



* + - 1. The histogram suggests that the y variable is binomial, supporting the use of a binomial GLM.
    1. Check collinearity
       1. Library(car)
       2. vif(glm(malignant~radius\_mean+texture\_mean+smoothness\_mean, data=cancer, family='binomial'))

radius\_mean texture\_mean smoothness\_mean

1.740809 1.633201 2.066766

* + - 1. The VIF shows that none of the selected variables are highly correlated as vif<5 for all of them. The cor(cancer) function also reveals no collinearity.
  1. **Run test**
     1. glm\_mod = glm(malignant~radius\_mean+texture\_mean+smoothness\_mean, data=cancer, family=binomial(link="logit"))

summary(glm\_mod)

* + - 1. 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

* + - 1. This model shows that radius, texture, and smoothness are all significant in determining the probability that a tumor will be malignant. So the coefficient of the mean radius is the effect of mean radius on the probability of tumor malignancy, and the same for texture and smoothness. By increasing on unit of mean radius, the log odds of malignancy increase by 1.397.
  1. **Check model fit**
     1. mod0 = glm(malignant~1, data=cancer, family=binomial(link=logit))
     2. AIC(mod0,glm\_mod)
        1. df AIC

mod0 1 753.4400

glm\_mod 4 195.2902

* + - 1. The AIC model shows that the glm model has a lower AIC score than the null model (mod0). This indicates it is a better fit for the data.

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.**
   1. You can scale the variables to figure out which ones are the most important for explaining malignancy.
      1. 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) -1.0020 0.2035 -4.924 8.46e-07 \*\*\*

radius\_mean 4.9231 0.5428 9.069 < 2e-16 \*\*\*

texture\_mean 1.6368 0.2456 6.663 2.68e-11 \*\*\*

smoothness\_mean 2.0347 0.2679 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

* + 1. This scaled model shows that mean radius affects the log odds of whether or not a tumor is malignant the most.