**Ho Hsieh\_NRE538\_Final\_TakeHome\_Exam**

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**

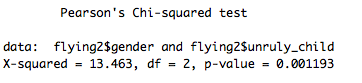
Test: Chi-square

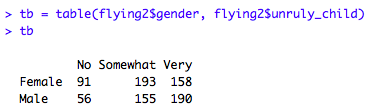
Check for the assumptions:

* 1. Random sampled: unknown.
  2. Independent observation: unknown.
  3. >20% of the groups have enough expected and observed values (>= 5 or 10): Yes. The least number of observations is 56.
  4. No structured 0: Yes.

Run the test:

**chisq.test(flying2$gender, flying2$unruly\_child)**

The result of the chi-square test indicates that gender is significant associated with whether people think it's rude to bring an unruly child on the plane (p-value < 0.05).  
  
Based on the chi-square table, we can tell that males tend to think it's more rude to bring an unruly child on the plane.



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

Test: T-test

Check for the assumptions:

1) Random sampled: unknown.

2) Independent observations: unknown.

3) Normal distribution OR sample size > 30: Yes.

The result of the shapiro test indicates that the distribution of the data is significant different than normal distribution (p-value < 0.05). But in the qq plot, the distribution is not deviated from the line a lot. Most of the values in the middle fit the line. As well, the sample size is larger than 30 (sample size = 1407). Therefore, we can assume the normality of the data.

4) Equal variance: No. **🡪 Run Welch’s T-test**

p-value < 2.2e-16 < 0.05 --> reject null hypothesis that the variance between the groups is equal

---> the variance between the groups is not equal.

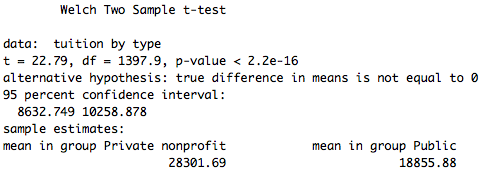
Run the test:

**t.test(tuition~type, var.equal = FALSE, data = college)**

The p-value < 2.2e-16 < 0.05 --> reject H0 that the tuition between the two types of institution are the same.

--> the tuition between the two types of institution are significantly different.

The mean of "Private nonprofit" is 28301.69 and the mean of "Public" is 18855.88 ---> The Private nonprofit has a higher tuition.



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

Test: One-way ANOVA

Check for the assumptions:

1) Random samples: unknown

2) Independent observations: unknown.

3) Normal distribution OR sample size > 30: Yes.

The result of the shapiro test indicates that the distribution of the data is significant different than normal distribution (p-value < 0.05). But in the qq plot, the distribution is not deviated from the line a lot. Most of the values in the middle fit the line. As well, the sample size is larger than 30 (sample size = 157). Therefore, we can assume the normality of the data.

4) equal variance: Yes.

The p value > 0.05 ---> we cannot reject the H0 that the variance among the groups are the same.

---> the variances are equal among the groups.

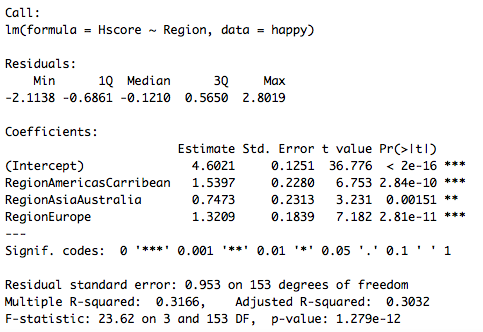
Run the test:

I use the "lm" function to run the ANOVA.

**lm3 = lm(Hscore~Region, data = happy)**

The result suggests that the happiness of RegionAmericasCarribean, RegionAsiaAustralia, and RegionEurope are significantly different than that of AfricaMideast (each of the regions other than AfricaMideast has a significant effect (p value < 0.05) on happiness).

Based on the result, we can say that there is at least one region (AfricaMideast) that is significantly different than other regions in terms of happiness.



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: Multiple linear regression

Check for collinearity:

Based on the results of the correlation table, I choose three continuous variables that are not highly correlated to each other (correlation cefficient < 0.5). The continuous independent variables I choose are "Generosity", "Freedom", and "GDP".

Check for the assumptions:

1) Independent observations: unknown.

2) Normal distribution of dependent variable: No. --> Transform the data.

The shapiro test indicates that the data is not normally distributed (p-value < 0.05 --> reject H0 that the data is normally distributed). In the qq plot, the data does not fit the line well. Therefore, I would try to transform the data.

---> log transform the data

The shapiro test indicates that the data is normally distributed after log transform (p-value > 0.05 --> cannot reject H0 that the data is normally distributed). In the qq plot, the data fit the line better.

3) Linear relationship between the dependent variable and the independent variables: assumed.

By plotting the X-Y graph, there might be a positive linear relationship between log(Corruption) and the three independent variables.

Run the test:

**lm4 = lm(Corruption.log~Generosity+Freedom+GDP, data = happy)**

(Assume we set the level of significance as 0.05)

**Intercept:** The intercept is significant (p-value < 2e-16) which means that controlled for the effect of Generosity, Freedom, and GDP, the average log(Corruption) level is -3.16388 and is significant different from 0. (That is, the average corruption level is e^-3.16388 = 0.04226145)

**Generosity**: For every unit increase in Generosity, the log(Corruption) would increase by 0.54962 unit. But this effect is not significant (p-value = 0.226 > 0.05).

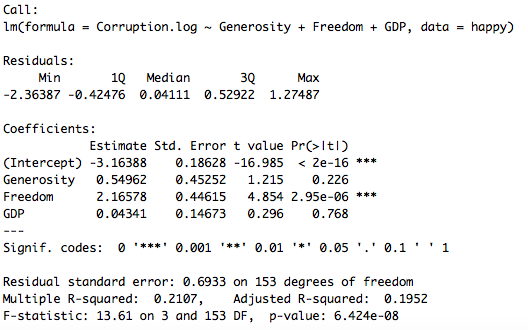
**Freedom**: For every unit increase in Freedom, the log(Corruption) would increase by 2.16578 unit. That is, for every unit increase in Freedom, the Corruption would be e^2.16578 (= 8.721402) times of the original value. This effect is significant (p value = 2.95e-06).

**GDP:** For every unit increase in GDP, the log(Corruption) would increase by 0.04341 unit. But this effect is not significant (p value = 0.786).

**The multiple R-squared** is 0.2107, which means the model can explain 21.07% of the variance in the data.

**The adjusted R-squared** is 0.1952 which suggests that the model fit the data okay.

The **p value of the F test** is 6.424e-08 < 0.05, which indicate that the model is significant different than null model or randomness.



Check model fit:

I use “plot(model)” function as well as shapiro test, dwtest, and bptest.

1. Normality of residuals

The result of the shapiro test shows that the p-value is smaller than 0.05 indicating that we reject the H0 that the residual distribution is not different from normal distribution. --> the residuals are not normally distributed.

However, from the qq plot, the residuals fit the normal line fairly and are not hugely deviated from normal distribution. Therefore, I would think the residuals are normally distributed.

1. Independence of residuals

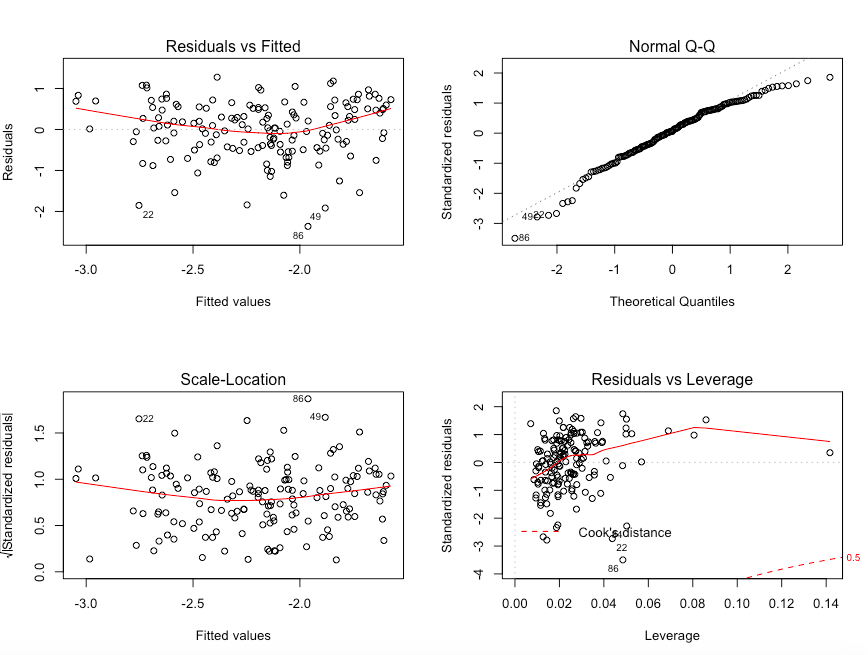
The dwtest shows that the p-value = 0.3701 > 0.05. We cannot reject the H0 that the residuals are independent. ---> The residuals are independent.

1. Homoscedasticity

The bptest shows that the p-value = 0.06878 > 0.05 indicating that we cannot reject the H0 that the residual distribution is homoscedasticity.

---> The residuals are homoscedasticity.

Based on the results of model fit checks, I would say the model fit the data pretty well.



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: ANCOVA

Check for assumptions:

Based on the assumption checks done in the last question, I would log transform the dependent variable (Corruption).

Run the test:

**lm5 = lm(Corruption.log~Region\*Freedom, data = happy)**

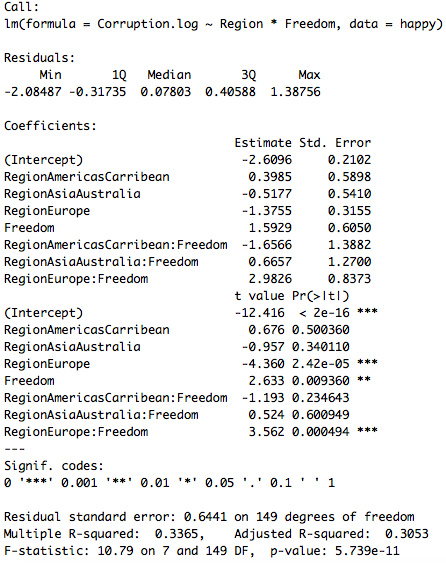
(Assume that we set 0.05 as the level of significance)

The result of the linear model show that:

* **Intercept**: The estimate of the intercept is -2.6096 which means that controlled for other independent variables, the average log(Corruption) in AfricaMideast is -2.6096. The intercept is significant (p-value < 2e-16).
* **RegionAmericasCarribean:** The estimate of the RegionAmericasCarribean is 0.3985 which means that controlled for other independent variables, the average log(Corruption) in AmericasCarribean is -2.6096 + 0.3985 = -2.2111. But this effect of RegionAmericasCarribean is not significant (p-value = 0.500360 > 0.05).
* **RegionAsiaAustralia**: The estimate of the RegionAsiaAustralia is -0.5177 which means that controlled for other independent variables, the average log(Corruption) in AsiaAustralia is -2.6096 - 0.5177 = -3.1273. But this effect of AsiaAustralia is not significant (p-value = 0.340110 > 0.05).
* **RegionEurope**: The estimate of the RegionEurope is 0.3985 which means that controlled for other independent variables, the average log(Corruption) in Europe is -2.6096 - 1.3755 = -3.9851. This effect of Europe is significant (p-value = 2.42e-05 < 0.05).

* **Freedom**: The estimate of Freedom is 1.5929 which means for every unit increase in Freedom, the log(Corruption) in AfricaMideast would increase by a rate of 1.5929. This effect is significant (p-value = 0.009360 < 0.05).
* **RegionAmericasCarribean:Freedom**: The estimate of RegionAmericasCarribean:Freedom is -1.6566 which means that for every unit increase in Freedom in AmericasCarribean, the log(Corruption) would increase by a rate of 1.5929 - 1.6566 = -0.0637. But this effect is not significant (p-value = 0.234643 > 0.05).
* **RegionAsiaAustralia:Freedom**: The estimate of RegionAsiaAustralia:Freedom is 0.6657 which means that for every unit increase in Freedom in AsiaAustralia, the log(Corruption) would increase by a rate of 1.5929 + 0.6657 = 2.2586. But this effect is not significant (p-value = 0.600949 > 0.05).
* **RegionEurope:Freedom**: The estimate of RegionEurope:Freedom is 2.9826 which means that for every unit increase in Freedom in Europe, the log(Corruption) would increase by a rate of 1.5929 + 2.9826 = 4.5755. This effect is significant (p-value = 0.000494 < 0.05).
* **The multiple R-squared** is 0.3365, which means the model can explain 33.65% of the variance in the data.
* **The adjusted R-squared** is 0.3053, which suggests that the model fit the data okay.
* **The p value of the F test** is 5.739e-11 < 0.05, which indicate that the model is significant different than null model or randomness.

Yes, the influence of the continuous variable (Freedom) is different among regions. Compared to in AfricaMideast, the influence of Freedom on log(Corruption) in Region Europe is higher (1.5929 + 2.9826 = 4.5755). For AmericasCarribean and AsiaAustralia, the influence of Freedom on log(Corruption) is the same as that of AfricaMideast.



Check model fit:

I use “plot(model)” function as well as shapiro test, dwtest, and bptest.

1. Normality of residuals

The result of the shapiro test shows that the p-value is smaller than 0.05 indicating that we reject the H0 that the residual distribution is not different from normal distribution. --> the residuals are not normally distributed.

However, from the qq plot, the residuals fit the normal line fairly and are not hugely deviated from normal distribution. Therefore, I would think the residuals are normally distributed.

1. Independence of residuals

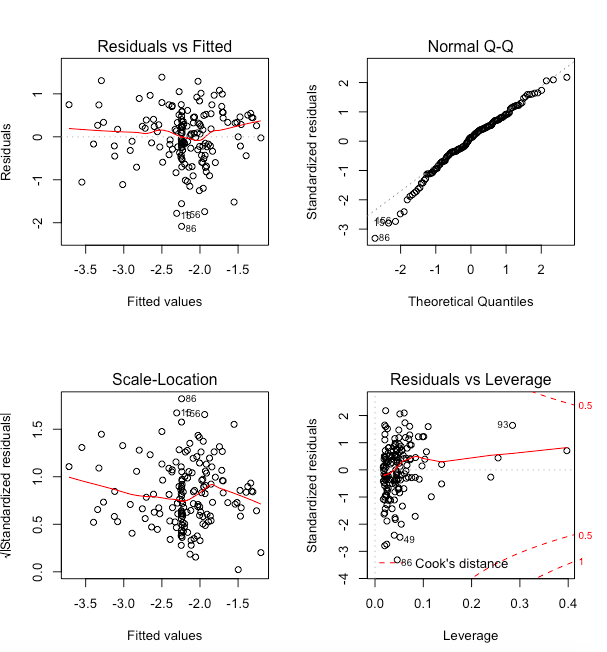
The dwtest shows that the p-value = 0.5627 > 0.05. We cannot reject the H0 that the residuals are independent.---> The residuals are independent.

1. Homoscedasticity

The bptest shows that the p-value = 0.2501 > 0.05 indicating that we cannot reject the H0 that the residual distribution is homoscedasticity.

---> The residuals are homoscedasticity.

Based on the results of model fit checks, I would say the model fit the data pretty well.



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: Generalized linear regression

Check for collinearity:

Based on the results of the correlation table, I choose three continuous variables that are not highly correlated to each other (correlation cefficient < 0.5).

The continuous independent variables I choose are "smoothness\_mean", "texture\_mean", and "radius\_mean".

Check for the assumptions:

1) Independent observations: unknown.

2) Linear relationship between the transformed dependent variable and the independent variables: assumed.

The data of whether a breast cancer tumor is malignant or not is a binomial distribution. Therefore, I would apply a logit transformation on the data.

logit transformation:   
p = the probability that the breast cancer tumor is malignant; odds = (p/(1-p)); y = log(odds)

Run the test:

**glm6 = glm(malignant~smoothness\_mean+texture\_mean+radius\_mean, data = cancer, family = binomial(link="logit"))**

* **Intercept**: When controlled for independent variables, the average log(odds) would be -42.01941.

That is, when controlled for other independent variables, the average probability that the breast cancer tumor would be malignant is exp(-42.01941)/(1+exp(-42.01941)) = 5.639e-19

The intercept is significant different than 0 (p value < 2e-16).

* **smoothness\_mean**: When controlled for other independent variables, for every unit increase in smoothness\_mean, the average log(odds) would increase by a rate of 144.67423.

That is, when controlled for other independent variables, every unit increase in smootheness\_mean would increase the odds by exp(144.67423) = 6.779845e+62 times.

This effect is significant (p value = 3.06e-14).

* **texture\_mean**: When controlled for other independent variables, for every unit increase in texture\_mean, the average log(odds) would increase by a rate of 0.38056.

That is, when controlled for other independent variables, every unit increase in texture\_mean would increase the odds by exp(0.38056) = 1.463104 times.

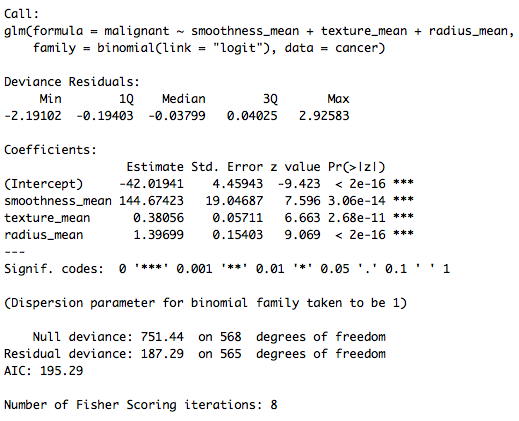
This effect is significant (p value = 2.68e-11).

* **radius\_mean**: When controlled for other independent variables, for every unit increase in radius\_mean, the average log(odds) would increase by a rate of 1.39699.

That is, when controlled for other independent variables, every unit increase in radius\_mean would increase the odds by exp(1.39699) = 4.043012 times.

This effect is significant (p value < 2e-16).

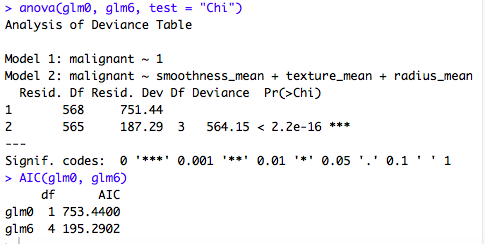
In conclusion, the smoothness\_mean, texture\_mean, and radius\_mean all are significantly associated with whether a breast cancer tumor is malignant or not.



Model fit check:

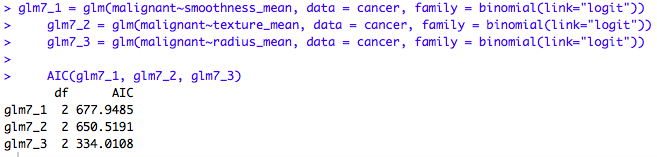
glm0 = glm(malignant~1, data = cancer, family = binomial(link = logit))

summary(glm0)



The results of the two tests (anova and AIC) indicate that the smoothness\_mean, texture\_mean, and radius\_mean as independent variables (glm6) could increase the likelihood of the model and the increase is significant (p-value < 0.05).

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 results of the AIC, we can see that radius\_mean is the most important variable among the three variables in explaining whether a breast cancer tumor is malignant or not.

(glm7\_3 has the lowest AIC, which implies that the model glm7\_3 would lost least amount of information)