Margaret Lindman

Final Take Home Exam

NRE 538

04/22/2017

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
* 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 If you download the datasets manually and load them, please use the following notation. Alternatively Oscar will host these datasets on GitHub soon so you can read them in directly like you did for the last exam.

flying = read.csv('/Users/mehajain/Desktop/flying.csv')

college = read.csv('/Users/mehajain/Desktop/college\_final.csv')

happy = read.csv('/Users/mehajain/Desktop/happy.csv')

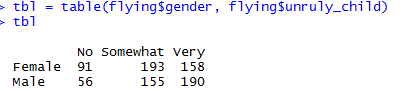
cancer = read.csv('/Users/mehajain/Desktop/cancer.csv')

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***

**-check whether the data meet the requirements/assumptions of the test you plan to run**

Categorical (ordinal) dependent variable (unruly\_child), and Categorical independent variable (gender)🡪 chi-square

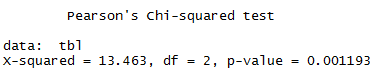


ASSUMPTIONS:

1. Independence of observations 🡪 assume, PASS
2. No ‘structural’ zeros (i.e. every single cell within the chi-square table has a probability > 0 that it can be ﬁlled – though it is possible in your sample it is 0). 🡪 PASS (contingency table)

3) Likely not appropriate when more than 20% of the cells in a table have expected values < 5 🡪 PASS (contingency table)

**-run the test**



**-interpret the results**

At a significance level of 0.05, there is a significant association between gender and whether people think it’s rude to bring an unruly child on the plane. Males tend to think that bringing an unruly child on a plane is more rude.

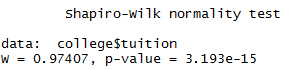
1. ***Is there a significant*** ***difference in tuition (tuition) by type of institution (type)? If yes, which type has a higher tuition? college***

**-check whether the data meet the requirements/assumptions of the test you plan to run**

Continuous dependent variable (Tuition) and categorical independent variable (college type🡪 with two groups)🡪Two sample t-test:

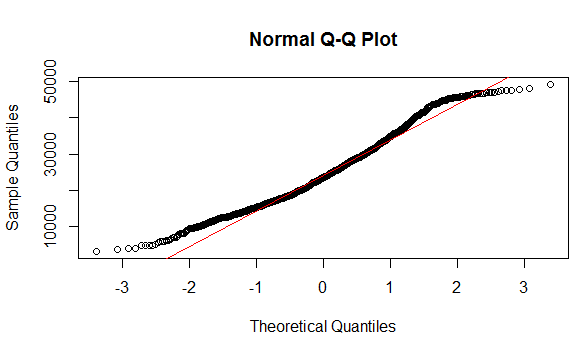
ASSUMPTIONS:

1. Data are continuous🡪 assume PASS
2. Sample is randomly selected from the population🡪 assume PASS
3. Our observations are independent🡪 assume PASS
4. Values are nearly normal OR the sample size is large enough (CLT)🡪 shapiro.test(college$tuition)🡪 FAIL (reject the null hypothesis of normality)



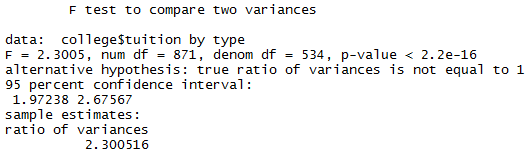
qqplot🡪 does not look normal

**BUT, we can assume normality based on the Central Limit Theorem because the sample size is >30.**

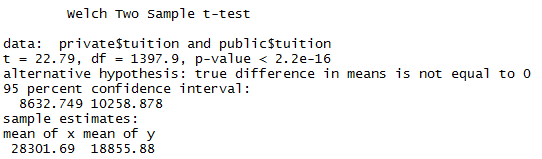


1. Equal variance between 2 populations (for two-sample t-test)🡪

var.test(college$tuition~type,data=college)🡪 FAIL (reject the null hypothesis of equal variance)🡪 need Welch’s T-test

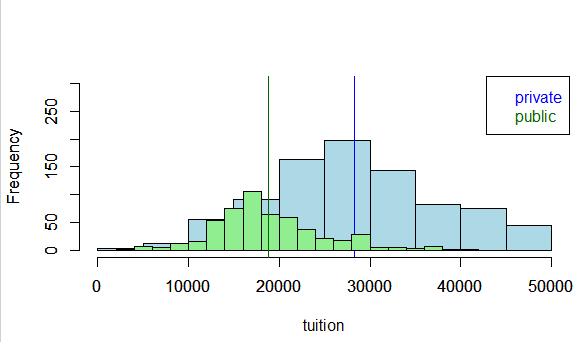


**-run the test**



**-interpret the results**

P-value<0.05, there is a significant difference in tuition (tuition) by type of institution (type). Private colleges have higher tuition.



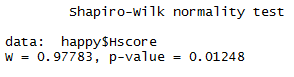
1. ***Is there a significant difference in happiness (Hscore) by region (Region)? happy***

**-check whether the data meet the requirements/assumptions of the test you plan to run**

Continuous dependent variable (happiness), Categorical independent variable with multiple groups (Region)🡪 One-way ANOVA

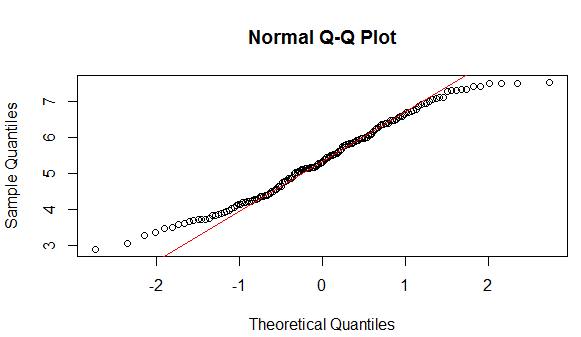
ASSUMPTIONS:

1)The populations of interest must be normally distributed🡪 shapiro.test(happy$Hscore) 🡪 FAIL (p-value less than 0.05, therefore we reject the null hypothesis of normality)



qqplot: Does not appear to be normal

**BUT, we can assume normality because sample size is > 30 with at least 15 per group.**



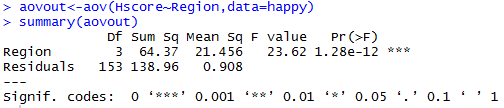
2) The samples must be independent of one another🡪 Assume, PASS

3) Each population must have the same variance (Levene’s test)🡪 PASS (fail to reject the null hypothesis that variances are equal)



4) (Does **not** require for equal sample sizes across groups)

**-run the test**



**-interpret the results**

P-value < 0.05, thus there a significant difference in happiness (Hscore) by region (Region). Time for the **Tukey-Kramer procedure** to decide which region(s) is different (cannot use Tukey HSD because the groups/Regions have different sample sizes).

ASSUMPTIONS of Tukey Kramer:

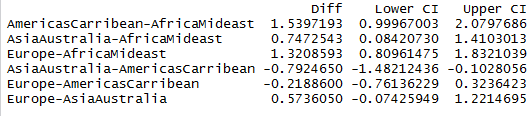
1)Observations are independent within and among groups🡪 assume PASS

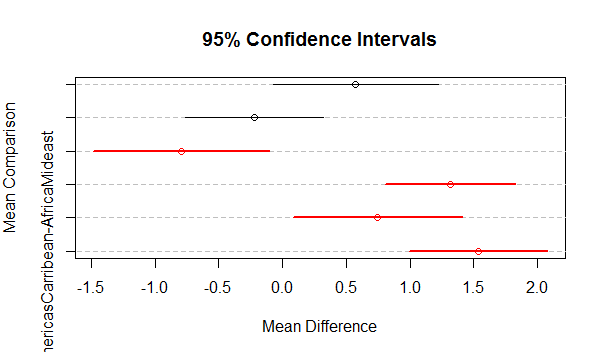
2)Groups are normally distributed🡪 (see above) shapiro.test and qqplot non-normal, **BUT can assume normality because sample size >30 and each region has at least 15 observations.**

3)Equal variance within and among groups 🡪(see above) Levene’s Test🡪 PASS

4) Does **not** require equal sample size across groups

From the Tukey Kramer test, we know that Europe-AmericasCarribean and Europe-AsiaAustralia are the only two regions that are NOT significantly different in terms of happiness score (contain “0” in the confidence interval mean difference)





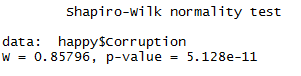
1. ***What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model. Happy***

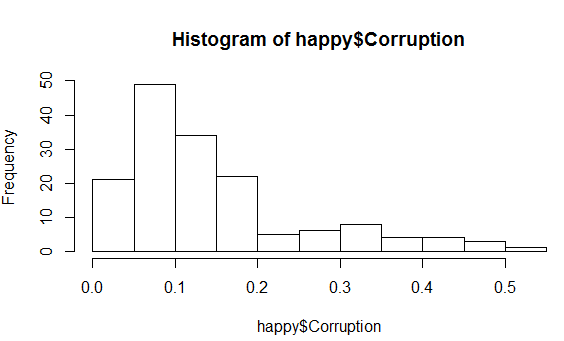
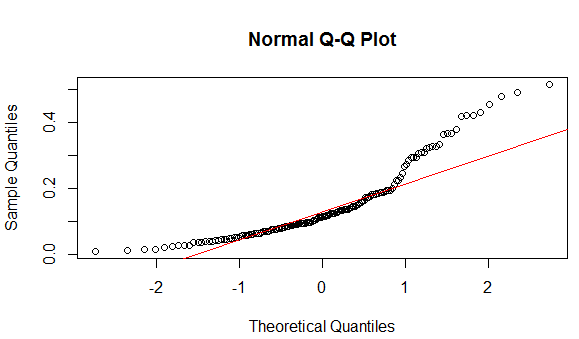
**-check whether the data meet the requirements/assumptions of the test you plan to run**

Multilinear regression (continuous dependent, and multiple continuous independent variables)

**ASSUMPTIONS:**

1. Independent samples🡪 assume PASS
2. Linear relationship between dependent and independent variables🡪 assume PASS
3. Normally distributed dependent variable🡪 FAIL (reject the null hypothesis for normality), and plots do not appear normal

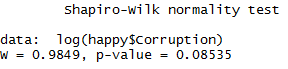


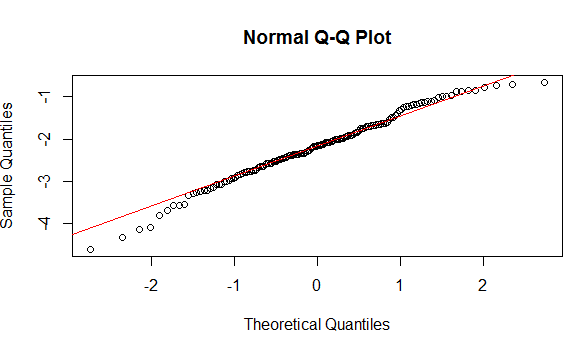
**-complete any transforms needed to make the data meet the required assumptions**

Log transform of dependent variable (Corruption) to address failed assumption of normality:

PASSES Shapiro-Wilk test after log transformation (cannot reject the null hypothesis for normality):

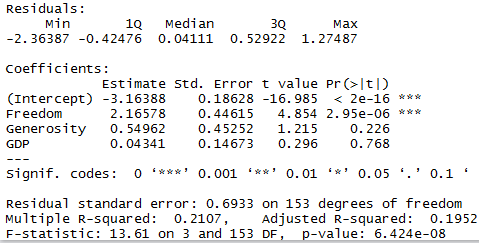


Better (more normal) qqplot of dependent variable (Corruption) after log transformation:



**-run the test**





**-interpret the results**

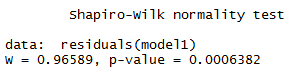
From the output, we see that Freedom is the only independent variable that has a significant effect on Corruption. When Freedom increases by 1 unit, Corruption increases by 2.17 units.

**-check model fit in the case of linear regressions and/or glms**

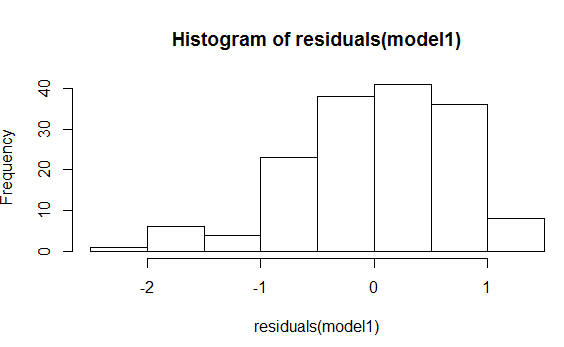
MODEL FIT CHECKS:

1. Normality of errors:

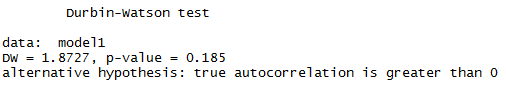
shapiro.test🡪 FAILED (reject null hypothesis of normality), but the Shapiro-Wilk test is very stringent; let’s check the histogram:



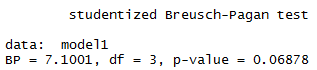
The histogram does not look perfect, but good enough to proceed.



1. Independent errors: dwtest🡪 PASSED (cannot reject the null hypothesis of independent errors)



1. Homoscedascity: bptest🡪 PASSED (cannot reject the null hypothesis of homoscedastic errors)



4) Non-normal errors: Shapiro Wilk:

Check if independent variables are highly correlated using VIF:



Generally, we want the VIF to be smaller than 5, therefore, there is no problematic multicollinearity among my independent variables.

The Adjusted R2 for this model is about 0.20, which means it only explains about 20% of the variation in Corruption. However this model may still be useful in that we now know freedom has a significant effect on corruption.

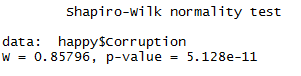
1. ***Choose one of the continuous independent variables that was significant in the model for Question 4 and interact it with region (Region) to predict corruption (Corruption). This model should only include one continuous independent variable and its interaction with region. Does the influence of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

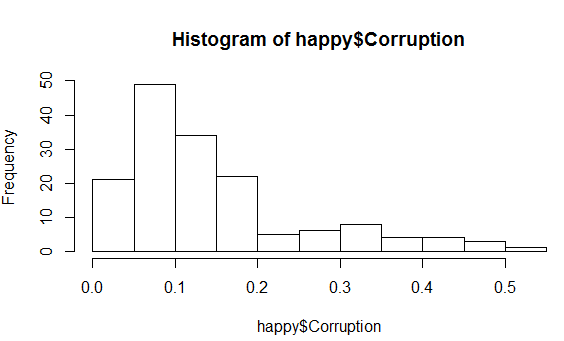
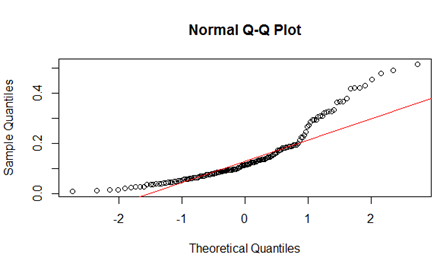
**-check whether the data meet the requirements/assumptions of the test you plan to run**

Continuous dependent variable (Corruption) and one continuous (Freedom) independent variable and interaction term (Freedom x Region) 🡪 ANCOVA

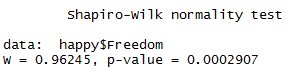
ASSUMPTIONS:

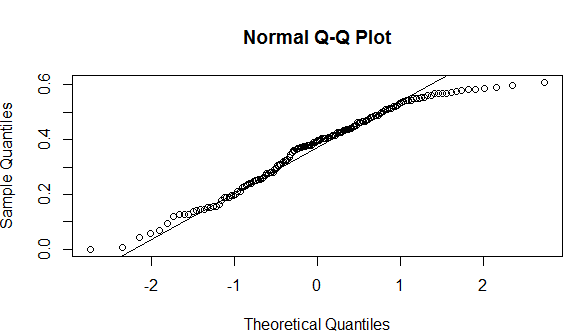
1. Normally of data: Corruption: shapiro.test🡪 FAIL (reject the null hypothesis of normality), and plots do not appear normal



Freedom:





1. Each population must have the same variance (Levene’s test)

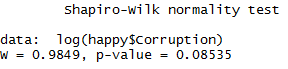


1. Independent samples: assume🡪 PASS

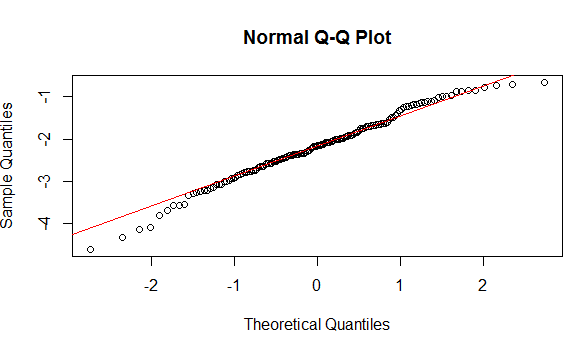
**-complete any transforms needed to make the data meet the required assumptions**

Log transform of dependent variable (Corruption) to address failed assumption of normality:

PASSES Shapiro-Wilk test after log transformation (cannot reject the null hypothesis for normality) :

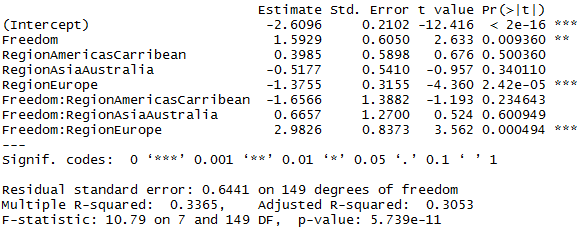


Better qqplot of dependent variable (Corruption) after log transformation.



**-run the test**





**-interpret the results**

Here the intercept is AfricaMideast (because of alphabetical order) when Freedom is 0. The estimate of Freedom (1.59, p-value < 0.05) shows a significant effect of Freedom on Corruption in AfricaMideast. Because the interaction term is significant for Europe, the effect (slope) of the continuous variable (Freedom) on Corruption differs in the case of Europe, when compared to AfricaMideast.

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***

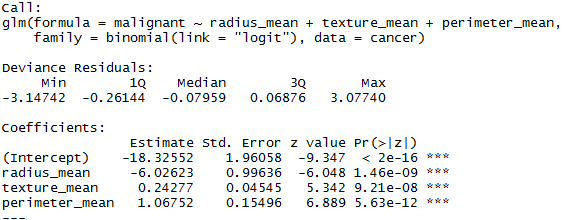
**-check whether the data meet the requirements/assumptions of the test you plan to run**

GLM🡪 binomial (if breast cancer tumor is malignant or not). “Malignant” is a binomial (0 or 1) data. To decide which continuous independent variables to include in my model, I used the cor () function. I chose three independent variables with correlation values <0.5.

ASSUMPTIONS/MODEL CHECKS:

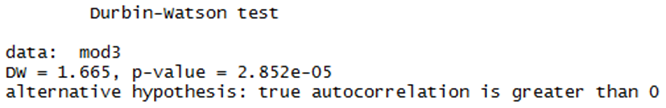
1. Linear relationship🡪 (after ‘transformation’)
2. Independent sample🡪 assume PASS

**-run the test**



MODEL CHECKS:

Independent errors🡪 FAILED dwtest (could try clustering SEs)



**-interpret the results**

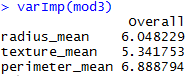
Logit = log odds

Radius\_mean significantly affects the chances that a breast cancer tumor is malignant or not by logit(-6.03).

Texture\_mean significantly affects the chances that a breast cancer tumor is malignant or not by logit(0.24).

Perimeter\_mean significantly affects the chances that a breast cancer tumor is malignant or not by logit(1.07).

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.***



**-interpret the results**

Perimeter\_mean is the most important in explaining whether a breast cancer tumor is malignant or not, radius\_mean is the second most important, and texture\_mean is the third most/least important.