University of Toronto

School of Continuing Studies

SCS 3251 Statistics of Data Science

Final Project – Autumn 2018

Factors Affecting GPA

and

GPA Performance Between

Males and Females

Student: Isaac Aktam

Professor: Dr. Sergiy Nokhrin

Due Date: Tuesday, 6:30 PM, December 4, 2018 Section: 016

Table of Contents

Summary/Abstract

The final project is based upon the analysis that was performed in Quantitiatve Methods in Economics course in my undergrad. Initial analysis was performed in Excel and with erroneous conclusions. It was erroneously concluded that there is a statistically significant difference between Male GPA and Female GPA due to an error in T-Statistic calculation.

The purpose of this project was to use to use Python to determine what factors have the most significant effect on student's grade point average (GPA) and subsequently determine if there is a significant difference in academic performance between males and females. To do such analysis, we have initially surveyed 103 University of Toronto Scarborough students from various departments. From our most recent analysis using Python, we gather the following results:

- For all of the 103 students, GPA is affected by *Classes Missed, Studying Hours,* and *Happiness*
- For 50 male students, GPA is affected by *Classes Missed*, *Studying Hours, Employment Hours*, and *Happiness*
- For 53 female students, we achieved inconclusive results with GPA being affected by *Happiness* only

Introduction

The purpose of this project was to recreate initial findings from the undergraduate university cours and therefore to determine what factors have the most significant effect on student's grade point average (GPA) and subsequently determine if there is a significant difference in academic achievement between males and females. Students often question themselves where they should concentrate the most to improve their GPA. This is the primary reason we decided to do this research. Therefore, our research question are:

- What factors affect GPA the most?
- Who performs better academically, males and females?

Factors affecting GPA and GPA performance between males and females

To do such analysis, we have initially surveyed 103 University of Toronto Scarborough students from variaous deparments in order to ensure that our sample is random enough and represents the overall student population. There are many factors affecting students GPA, however we decided to focus on factors like Extracurricular Hours, Amount of Courses Taken, Amount of Classes Missed, Hours Studying, Happiness Level, Hours Spent Working, Hours Spent on Commuting, and Hours Spent on Relationships.

The survey questions cover both personal, academic, and social life of students thus encompassing a great range of activities that we hypothesize to have a great effect on student's GPA. We divide our analysis into two parts:

1. Linear regression analysis of gathered data from 103 surveys for the purpose of determining factors influencing student's GPA. Additionally, we employ the method of Backward Elimination.
2. Hypothesis analysis of GPA of 50 male students (Sample 1) and GPA of 53 female students (Sample 2) for the purposes of determining who performs better academically.

For this study, we hypothesize the following results to be true:

1. Studying hours and happiness level to have high positive correlation with GPA.
2. Work hours, classes missed, and relationship commitments have strong negative correlation with GPA.
3. Amount of courses taken have either weak positive correlation or weak negative correlation with GPA.
4. Commute hours, classes missed, and extracurricular activities have weak negative correlation with GPA.
5. Equal academic performance between male students and female students.

Sample Survey Questions

- What is your gender?
- How many hours per week do you work?
- How many hours per week do you spend on commuting to and from university?
- If you are in a relationship, how many hours per week do you spend on your partner? • How many hours per week do you spend on extracurricular activities?
- How many courses are you taking this semester?
- How many lectures, in total, have you missed during this semester?
- How many hours per week do you study?
- What is your GPA?
- One a scale from 1 to 5, specify your happiness level for this semester: 1 2 3 4 5

Such that:

- 1 – Very Unhappy
- 2 - Unhappy
- 3 - Content
- 4 - Happy
- 5 – Very Happy

Methodology and Analysis

To gather the necessary data, we surveyed 150 University of Toronto Scarborough students. Each of them received a single survey. Our survey included short and concise questions so as to let students finish them quickly and not to discourage students from not answering them. We were not guaranteed to receive a completely answered survey. As a result, after gathering all of the surveys, we had to drop 47 faulty questionnaires (i.e. people would leave some of the questions blank or give a vague answer). As a result we were left with 103 surveys. Next, we divided 103 remaining surveys into a Male group of 50 surveys and a Female group of 53 surveys. Lastly, we gathered all of the necessary results for the further analysis. Since the size of each group/sample is bigger than 30, we are guaranteed to have a normally distributed data.

For the analysis part, we use the following libraries:

```
import pandas as pd
import numpy as np
import scipy.stats as st

# to make this notebook's output stable across runs
np.random.seed(123)

%matplotlib inline
import matplotlib
import matplotlib.pyplot as plt
import plotly.graph_objs as go
import plotly.offline as py
py.init_notebook_mode(connected=True)
from plotly.offline import download_plotlyjs, init_notebook_mode, plot, iplot
init_notebook_mode(connected=True)
plt.rcParams['axes.labelsize'] = 14
plt.rcParams['xtick.labelsize'] = 12
plt.rcParams['ytick.labelsize'] = 12

import plotly.figure_factory as ff
from statsmodels.formula.api import ols
import pylab

import seaborn as sns
import statsmodels.stats.api as sms
import plotly.tools as tls
```

Sample of Male data and its statistics:

In [180]:
```
male_data.head(5)
```

Out[180]:

	GPA	Extracurricular Hours	Number of Courses	Classes missed	Studying Hours	Happiness	Employment Hours	Commute Hours	Relationship Hours
0	3.10	2.0	4	30	25.0	3	0.0	2.00	10
1	3.05	14.0	6	3	30.0	2	4.0	18.00	0
2	3.63	3.0	6	0	10.0	4	0.0	0.50	0
3	2.50	3.0	4	0	26.0	3	0.0	1.00	0
4	4.00	5.0	4	2	120.0	4	0.0	0.25	0

In [305]:
```
male_data.describe()
```

Out[305]:

	GPA	Extracurricular Hours	Number of Courses	Classes missed	Studying Hours	Happiness	Employment Hours	Commute Hours	Relationship Hours
count	50.000000	50.000000	50.000000	50.000000	50.000000	50.000000	50.000000	50.000000	50.000000
mean	3.123800	4.930000	4.380000	7.100000	23.890000	3.000000	5.080000	5.484200	12.140000
std	0.473234	5.460442	0.945235	7.804368	25.359515	1.087968	11.568942	4.678884	34.005408
min	1.980000	0.000000	2.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
25%	2.817500	0.500000	4.000000	2.000000	5.750000	2.000000	0.000000	1.350000	0.000000
50%	3.100000	4.000000	4.000000	4.000000	17.500000	3.000000	0.000000	5.000000	0.000000
75%	3.477500	6.375000	5.000000	9.500000	30.000000	3.750000	6.500000	8.000000	4.750000
max	4.000000	28.000000	6.000000	30.000000	120.000000	5.000000	70.000000	18.000000	184.000000

Sample of Female data:

In [182]:
```
female_data.head(5)
```

Out[182]:

	GPA	Extracurricular Hours	Number of Courses	Classes missed	Studying Hours	Happiness	Employment hours	Commute Hours	Relationship Hours
0	3.22	0.0	4	2	21.0	4	0.0	5.0	40
1	2.78	10.0	5	6	10.0	4	7.5	4.0	10
2	3.40	0.0	5	2	34.0	2	0.0	5.0	0
3	3.50	4.0	5	0	10.0	1	0.0	10.0	0
4	3.00	4.0	5	0	25.0	1	0.0	5.0	0

In [306]:
```
female_data.describe()
```

Out[306]:

	GPA	Extracurricular Hours	Number of Courses	Classes missed	Studying Hours	Happiness	Employment hours	Commute Hours	Relationship Hours
count	53.000000	53.000000	53.000000	53.000000	53.000000	53.000000	53.000000	53.000000	53.000000
mean	3.212823	5.707547	4.773585	5.528302	18.216981	3.226415	4.905660	4.960377	12.396226
std	0.543054	6.509190	0.823723	5.960136	12.626860	1.031084	7.660498	5.180152	33.675681
min	1.620000	0.000000	2.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
25%	2.780000	2.000000	4.000000	1.000000	10.000000	3.000000	0.000000	0.500000	0.000000
50%	3.300000	4.000000	5.000000	4.000000	14.000000	3.000000	0.000000	4.000000	0.000000
75%	3.740000	6.000000	5.000000	8.000000	28.000000	4.000000	10.000000	9.000000	5.000000
max	4.000000	36.000000	6.000000	32.000000	60.000000	5.000000	30.000000	30.000000	184.000000

Sample of Combined data:

In [314]:
```
combined_data.head(5)
```

Out[314]:

	Gender	Extracurricular Hours	Number of Courses	Classes missed	Studying Hours	Happiness	Employment Hours	Commute Hours	Relationship Hours	GPA
0	Male	5.0	4	4	0.0	4	0.0	10.0	0	3.33
1	Male	0.0	2	15	3.0	2	0.0	11.0	0	2.33
2	Female	0.0	4	4	3.0	5	12.0	1.6	15	2.70
3	Female	1.0	4	6	14.0	4	12.5	2.5	3	2.75
4	Female	4.0	4	1	10.0	4	0.0	2.0	0	3.00

In [315]:
```
combined_data.describe()
```

Out[315]:

	Extracurricular Hours	Number of Courses	Classes missed	Studying Hours	Happiness	Employment Hours	Commute Hours	Relationship Hours	GPA
count	103.000000	103.000000	103.000000	103.000000	103.000000	103.000000	103.000000	103.000000	103.000000
mean	5.330097	4.582524	6.291262	20.970874	3.116505	4.990291	5.214660	12.271845	3.169608
std	6.006350	0.902331	6.927681	19.958514	1.059976	9.706724	4.926053	33.670023	0.509832
min	0.000000	2.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	1.620000
25%	2.000000	4.000000	2.000000	10.000000	2.000000	0.000000	1.150000	0.000000	2.800000
50%	4.000000	5.000000	4.000000	15.000000	3.000000	0.000000	4.000000	0.000000	3.200000
75%	6.000000	5.000000	8.500000	30.000000	4.000000	7.250000	8.500000	5.000000	3.580000
max	36.000000	6.000000	32.000000	120.000000	5.000000	70.000000	30.000000	184.000000	4.000000

Next, we peformed graphical analysis of the GPA:

	Male GPA	Female GPA	Combined GPA
Mean	3.123800	3.212823	3.169608
Standard Deviation	0.468478	0.537906	0.507351

Next, we need to prepare our data. To do so, we binarized the Happiness feature as it is categorical feature:

Turn happiness values into text

In [199]:
```
male_data['Happiness'] = male_data['Happiness'].map({5 : 'Very Happy' , 4 : 'Happy',  3 : "Content",
                                                      2 : "Unhappy", 1:  "Very Unhappy"})
```

In [200]:
```
female_data['Happiness'] = female_data['Happiness'].map({5 : 'Very Happy' , 4 : 'Happy',  3 : "Content",
                                                          2 : "Unhappy", 1:  "Very Unhappy"})
```

In [201]:
```
combined_data['Happiness'] = combined_data['Happiness'].map({5 : 'Very Happy' , 4 : 'Happy',  3 : "Content",
                                                              2 : "Unhappy", 1:  "Very Unhappy"})
```

Binarize Happiness

In [202]:
```
male_data = pd.concat([male_data, pd.get_dummies(male_data["Happiness"], prefix = "Happiness")], axis = 1)
male_data = male_data.drop("Happiness", axis = 1)
```

In [203]:
```
female_data = pd.concat([female_data, pd.get_dummies(female_data["Happiness"], prefix = "Happiness")], axis = 1)
female_data = female_data.drop("Happiness", axis = 1)
```

In [204]:
```
combined_data = pd.concat([combined_data, pd.get_dummies(combined_data["Happiness"], prefix = "Happiness")],
                          axis = 1)
combined_data = combined_data.drop("Happiness", axis = 1)
```

Therefore, we get 5 new binary features - Happiness_Content, Happiness_Happy, Happiness_Unhappy, Happiness_Very Happy, Happiness_Very Unhappy – for Maled data, Female data, and Combined data:

In [457]:
```
male_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 13 columns):
GPA                       50 non-null float64
Extracurricular Hours     50 non-null float64
Number of Courses         50 non-null int64
Classes missed            50 non-null int64
Studying Hours            50 non-null float64
Employment Hours          50 non-null float64
Commute Hours             50 non-null float64
Relationship Hours        50 non-null int64
Happiness_Content         50 non-null uint8
Happiness_Happy           50 non-null uint8
Happiness_Unhappy         50 non-null uint8
Happiness_Very Happy      50 non-null uint8
Happiness_Very Unhappy    50 non-null uint8
dtypes: float64(5), int64(3), uint8(5)
memory usage: 3.4 KB
```

In [458]:
```
female_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53 entries, 0 to 52
Data columns (total 13 columns):
GPA                       53 non-null float64
Extracurricular Hours     53 non-null float64
Number of Courses         53 non-null int64
Classes missed            53 non-null int64
Studying Hours            53 non-null float64
Employment hours          53 non-null float64
Commute Hours             53 non-null float64
Relationship Hours        53 non-null int64
Happiness_Content         53 non-null uint8
Happiness_Happy           53 non-null uint8
Happiness_Unhappy         53 non-null uint8
Happiness_Very Happy      53 non-null uint8
Happiness_Very Unhappy    53 non-null uint8
dtypes: float64(5), int64(3), uint8(5)
memory usage: 3.6 KB
```

In [459]:
```
combined_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103 entries, 0 to 102
Data columns (total 13 columns):
GPA                       103 non-null float64
Extracurricular Hours     103 non-null float64
Number of Courses         103 non-null int64
Classes missed            103 non-null int64
Studying Hours            103 non-null float64
Employment Hours          103 non-null float64
Commute Hours             103 non-null float64
Relationship Hours        103 non-null int64
Happiness_Content         103 non-null uint8
Happiness_Happy           103 non-null uint8
Happiness_Unhappy         103 non-null uint8
Happiness_Very Happy      103 non-null uint8
Happiness_Very Unhappy    103 non-null uint8
dtypes: float64(5), int64(3), uint8(5)
memory usage: 7.0 KB
```

Part 1. Regression Analysis.

In general, population GPA can be described by the following linear regression model

Factors affecting GPA and GPA performance between males and females

$$Y=\beta_0+\beta_1*X_1+\beta_2*X_2+\beta_3*X_3+\beta_4*X_4+\beta_5*X_5+\beta_6*X_6+\beta_7*X_7+\beta_8*X_8+\beta_9*X_9+\beta_{10}*X_{10}+\beta_{11}*X_{11}+\beta_{12}*X_{12}+\varepsilon$$

But, since we are dealing with samples instead of populations, we can only estimate the GPA results.

Therefore, we use the regression estimation line

$$\hat{Y}=b_0+b_1*X_1+b_2*X_2+b_3*X_3+b_4*X_4+b_5*X_5+b_6*X_6+b_7*X_7+b_8*X_8+b_9*X_9+b_{10}*X_{10}+b_{11}*X_{11}+b_{12}*X_{12}$$

э:

$$X_1=Extracurricular\ Hours$$
$$X_2=Number\ of\ Courses$$
$$X_3=Classes\ missed$$
$$X_4=Studying\ Hours$$
$$X_5=Employment\ hours$$
$$X_6=Commute\ Hours$$
$$X_7=Relationship\ Hours$$
$$X_8=Happiness\ Content*I_{[0,1]}$$
$$X_9=Happiness\ Happy*I_{[0,1]}$$
$$X_{10}=Happiness\ Unhappy*I_{[0,1]}$$
$$X_{11}=Happiness\ Very\ Happy*I_{[0,1]}$$
$$X_{12}=Happiness\ Very\ Unhappy*I_{[0,1]}$$

To build the estimated regression equation, we employ the method of Backward Elimination.

Part 1.1: Regression Analysis– Backward Elimination, Grouped Data

First of all, let's us analyze the relationship between each independent variable and dependent variable GPA. The reason we are doing this is to get an idea what independent variables have the most effect on GPA.

Therefore, we have the following results:

- Extracurricular Hours, Commute Hours, Relationship Hours, and Happiness_Very Unhappy have very weak positive correlation with GPA
- Number of Courses, Studying Hours, Happiness_Happy, and Happiness_Very Happy have weak positive correlation with GPA.

	GPA
GPA	1
Extracurricular Hours	0.00164348
Number of Courses	0.21531
Classes missed	-0.320596
Studying Hours	0.274095
Employment Hours	-0.139038
Commute Hours	0.0168107
Relationship Hours	0.0115503
Happiness_Content	-0.196911
Happiness_Happy	0.139615
Happiness_Unhappy	-0.0879658
Happiness_Very Happy	0.143224
Happiness_Very Unhappy	0.0967447

Isaac Aktam

- Classis missed, Employment Hours, and Happiness_Content have weak negative correlation with GPA
- And, Happiness_Unhappy has very weak negative correlation with GPA

We build our estimated regression equation using the following algorithm:

1. For the male sample of 103 students, we start with 8 independent variables.
2. We compute p-value for each independent variable in the model.
3. Look for the independent variable with the maximum p-value > α = 0.05
4. Independent variable with the largest p-value is removed from the model.
5. Stop if p-value for all the remaining independent variables is less than or equal to α = 0.05. Else, return to step 3 and continue on with the process.

Therefore, we will have K ≥ 0 runs so as to reach our final estimated regression equation. Additionally, once we get the final regression equation, we will be able to run a hypothesis test on each single independent variable to determine if it has any effect on GPA. Variables in red text have the greatest p-value and thus dropped in the next run. Therefore, we have the following results:

Run	Equation
1	Yhat = 2.2468-0.0091*Extracurricular Hours+0.1097*Number of Courses-0.0238*Classes missed+0.071*Studying Hours-0.0085*Employment Hours+0.0135*Commute Hours+0.019*Relationship Hours+0.3565*Happiness_Content+0.5390*Happiness_Happy+0.2450*Happiness_Unhappy+0.5109*Happiness_Very Happy+0.5955*Happiness_Very Unhappy
2	Yhat = 2.25 +0.1017*Number of Courses-0.0227*Classes missed+0.007*Studying Hours-0.0082*Employment Hours+0.0102*Commute Hours+0.0018*Relationship Hours+0.3482*Happiness_Content+0.5230*Happiness_Happy+0.2551*Happiness_Unhappy+0.5066*Happiness_Very Happy+0.6171*Happiness_Very Unhappy
3	Yhat = 2.3039 +0.0985*Number of Courses-0.023*Classes missed+0.007*Studying Hours-0.0066*Employment Hours+0.0016*Relationship Hours+0.3532*Happiness_Content+0.5249*Happiness_Happy+0.2828*Happiness_Unhappy+0.4856*Happiness_Very Happy+0.6574*Happiness_Very Unhappy
4	Yhat = 2.3593 +0.0872*Number of Courses-0.0222*Classes missed+0.0071*Studying Hours-0.0072*Employment Hours+0.3824*Happiness_Content+0.5305*Happiness_Happy+0.2819*Happiness_Unhappy+0.5084*Happiness_Very Happy+0.6561*Happiness_Very Unhappy

5	Yhat = 2.2892 +0.0976*Number of Courses-0.0223*Classes missed+0.0073*Studying Hours+0.3683*Happiness_Content+0.4921*Happiness_Happy+0.2897*Happiness_Unhappy+0.4892*Happiness_Very Happy+0.6499*Happiness_Very Unhappy
6	Yhat = 2.6674-0.229*Classes missed+0.0074*Studying Hours+0.4249*Happiness_Content+0.5686*Happiness_Happy+0.37*Happiness_Unhappy+0.5647*Happiness_Very Happy+0.7392*Happiness_Very Unhappy R-Squared = 0.228

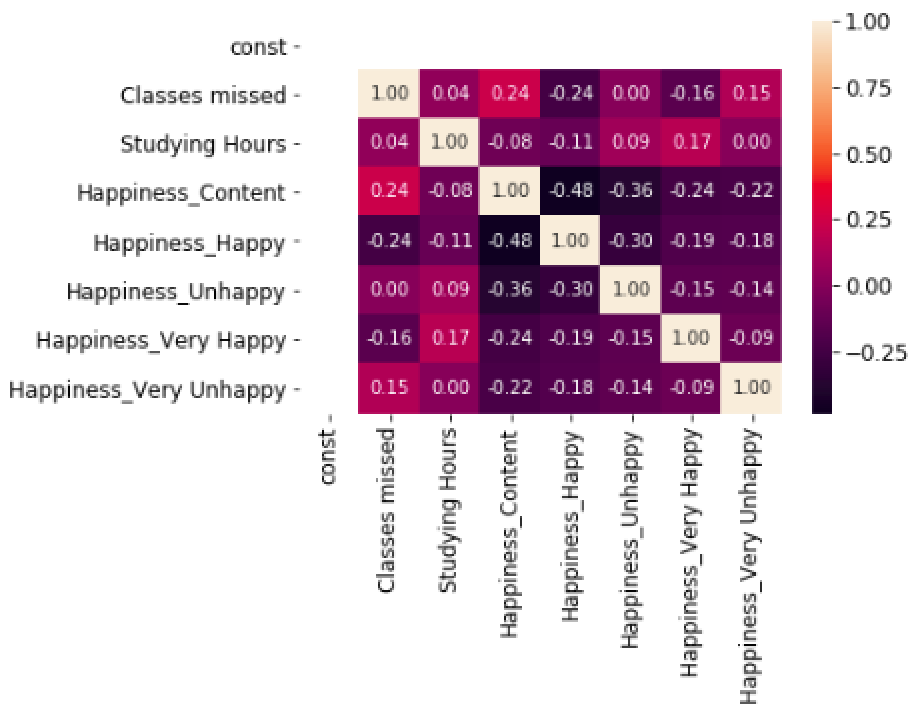
Therefore, we run a hypothesis test to determine if the final set of variables as a whole have effect on GPA.

$$H_0: \beta_{Classes\ Missed} = \beta_{Studying\ Hours} = \beta_{Happiness\ Content} = \beta_{Happiness\ Happy} = \beta_{Happiness\ Unhappy} = \beta_{Happiness\ Very\ Happy} = \beta_{Happiness\ Very\ Unhappy} = 0$$

$$H_a: At\ least\ one\ \beta_i \neq 0$$

From the model summy output, the p-value = 0.000298 < α = 0.05. Therefore, we reject the null hypothesis and conclude that the above variables have effect on GPA.

But, does there exist interrelationship between the independent variables and GPA? To check this, we need to take a look at the correlation matrix:



As we can

see from the above, the correlation between independent variables is low. We care about correlation because t-test tends to give faulty answers when correlation between independent variables is high. But, in our case it is low. Let us take a look at the p-values for each of the indepenet variables whether to test the following hypothesis:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    GPA   R-squared:                       0.228
Model:                            OLS   Adj. R-squared:                  0.180
Method:                 Least Squares   F-statistic:                     4.719
Date:                Sun, 02 Dec 2018   Prob (F-statistic):           0.000298
Time:                        18:17:57   Log-Likelihood:                -62.948
No. Observations:                 103   AIC:                             139.9
Df Residuals:                      96   BIC:                             158.3
Df Model:                           6
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                      2.6674      0.072     36.954      0.000       2.524       2.811
Classes missed            -0.0229      0.007     -3.256      0.002      -0.037      -0.009
Studying Hours             0.0074      0.002      3.140      0.002       0.003       0.012
Happiness_Content          0.4249      0.080      5.334      0.000       0.267       0.583
Happiness_Happy            0.5686      0.084      6.749      0.000       0.401       0.736
Happiness_Unhappy          0.3700      0.099      3.739      0.000       0.174       0.566
Happiness_Very Happy       0.5647      0.138      4.083      0.000       0.290       0.839
Happiness_Very Unhappy     0.7392      0.145      5.113      0.000       0.452       1.026
==============================================================================
Omnibus:                        1.427   Durbin-Watson:                   2.181
Prob(Omnibus):                  0.490   Jarque-Bera (JB):                1.358
Skew:                          -0.163   Prob(JB):                        0.507
Kurtosis:                       2.542   Cond. No.                     4.66e+17
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 4.07e-31. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

As we can see from the above table, the p-value for the all of the above independent variables is less than 0.05. Therefore, we reject the null hypothesis and conclude that none of the coefficients is 0. Therefore, given above t-test hypothesis results reinforce our F-test hypothesis results.

Part 1.2: Regression Analysis– Backward Elimination, Male Data

Factors affecting GPA and GPA performance between males and females

First, let's take a look at the correlation table for the Male Data.

	GPA
GPA	1
Extracurricular Hours	0.0358026
Number of Courses	0.272272
Classes missed	-0.303026
Studying Hours	0.387521
Employment Hours	-0.226996
Commute Hours	-0.111029
Relationship Hours	0.0180885
Happiness_Content	-0.123296
Happiness_Happy	0.0484019
Happiness_Unhappy	-0.153742
Happiness_Very Happy	0.257125
Happiness_Very Unhappy	0.0904521

- There is a weak positive relationship between Number of Courses, Studying Hours, and Happiness_Very Happy and GPA
- There is a very weak positive relationship between Extracurricular Hours, Relationship Hours, Happiness_Happy, and Happiness_Very Unhappy and GPA
- There is a weak negative relationship between Classes Missed, Employment Hours, Commute Hours, Happiness_Content, and Happiness_Unhappy and GPA.

We have the following regression models from the backward elimination:

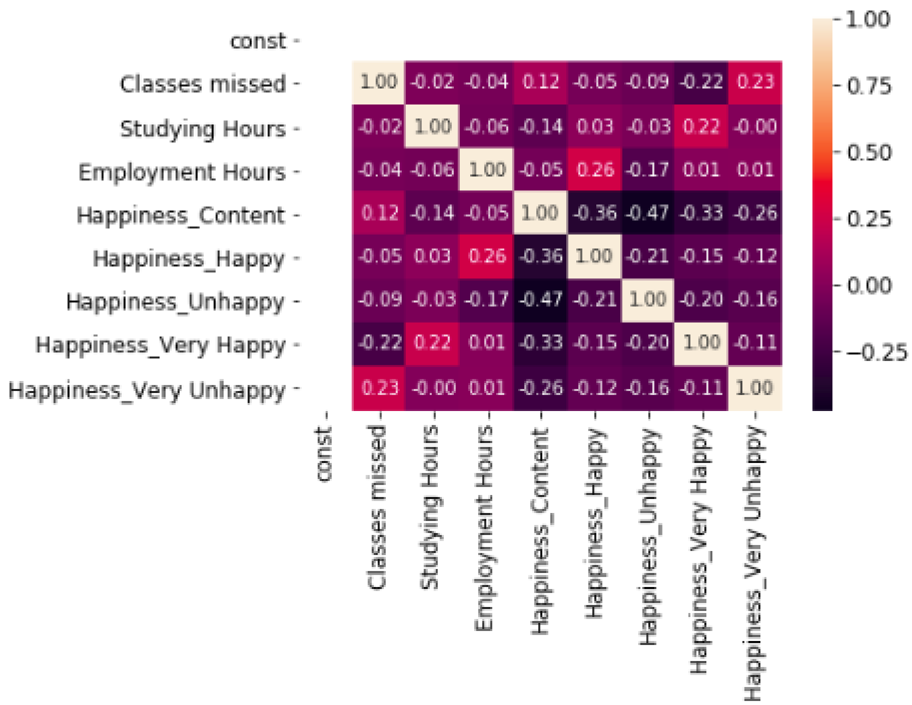
Run	Model
1	Yhat = 2.2724-0.0043*Extracurricular Hours+0.1110*Number of Courses-0.0205*Classes missed+0.0064*Studying Hours-0.0108*Employment Hours+0.0041*Commute Hours+0.0013*Relationship Hours+0.3658*Happiness_Content+0.5308*Happiness_Happy+0.2454*Happiness_Unhappy+0.5110*Happiness_Very Happy+0.6494*Happiness_Very Unhappy
2	Yhat = 2.2806-0.0034*Extracurricular Hours+0.1127*Number of Courses-0.0203*Classes missed+0.0063*Studying Hours-0.0105*Employment Hours+0.0012*Relationship Hours+0.3667*Happiness_Content+0.5250*Happiness_Happy+0.2261*Happiness_Unhappy+0.5042*Happiness_Very Happy+0.6586*Happiness_Very Unhappy
3	Yhat = 2.2859+0.1076*Number of Courses-0.02*Classes missed+0.0063*Studying Hours-0.0108*Employment Hours+0.0012*Relationship Hours+0.3651*Happiness_Content+0.5263*Happiness_Happy+0.2257*Happiness_Unhappy+0.5031*Happiness_Very Happy+0.6656*Happiness_Very Unhappy
4	Yhat = 2.2955+0.1066*Number of Courses-0.0188*Classes missed+0.0063*Studying Hours-0.0109*Employment Hours+0.3771*Happiness_Content+0.5194*Happiness_Happy+0.2231*Happiness_Unhappy+0.5183*Happiness_Very Happy+0.6576*Happiness_Very Unhappy
5	Yhat = 2.6977-0.0201*Classes missed+0.0062*Studying Hours-0.0113*Employment Hours+0.4519*Happiness_Content+0.5869*Happiness_Happy+0.2843*Happiness_Unhappy+0.6201*Happiness_Very Happy+0.7544*Happiness_Very Unhappy R-Squared = 0.366

Therefore, we run a hypothesis test to determine if the final set of variables as a whole have effect on GPA.

$$H_0: \beta_{Classes\ Missed} = \beta_{Studying\ Hours} = \beta_{Employment\ Hours} = \beta_{Happiness\ Content} =$$
$$\beta_{Happiness\ Happy} = \beta_{Happiness\ Unhappy} = \beta_{Happiness\ Very\ Happy} = \beta_{Happiness\ Very\ Unhappy} = 0$$
$$H_a: At\ least\ one\ \beta_i \neq 0$$

From the model summy output, the p-value = 0.00518 < α = 0.05. Therefore, we reject the null hypothesis and conclude that the above variables have effect on GPA.

But, does there exist interrelationship between the independent variables and GPA? To check this, we need to take a look at the correlation matrix:



As we can see from the above, the correlation between independent variables is low. We care about correlation because t-test tends to give faulty answers when correlation between independent variables is high. But, in our case it is low. Let us take a look at the p-values for each of the indepenet variables whether to test the following hypothesis:

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    GPA   R-squared:                       0.366
Model:                            OLS   Adj. R-squared:                  0.260
Method:                 Least Squares   F-statistic:                     3.458
Date:                Sun, 02 Dec 2018   Prob (F-statistic):            0.00518
Time:                        18:16:57   Log-Likelihood:                -21.657
No. Observations:                  50   AIC:                             59.31
Df Residuals:                      42   BIC:                             74.61
Df Model:                           7
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                      2.6977      0.094     28.569      0.000       2.507       2.888
Classes missed            -0.0201      0.008     -2.544      0.015      -0.036      -0.004
Studying Hours             0.0062      0.002      2.614      0.012       0.001       0.011
Employment Hours          -0.0113      0.005     -2.144      0.038      -0.022      -0.001
Happiness_Content          0.4519      0.093      4.884      0.000       0.265       0.639
Happiness_Happy            0.5869      0.144      4.069      0.000       0.296       0.878
Happiness_Unhappy          0.2843      0.117      2.439      0.019       0.049       0.520
Happiness_Very Happy       0.6201      0.155      3.988      0.000       0.306       0.934
Happiness_Very Unhappy     0.7544      0.185      4.081      0.000       0.381       1.127
==============================================================================
Omnibus:                        5.364   Durbin-Watson:                   2.431
Prob(Omnibus):                  0.068   Jarque-Bera (JB):                4.535
Skew:                          -0.726   Prob(JB):                        0.104
Kurtosis:                       3.260   Cond. No.                     7.55e+17
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.09e-31. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

As we can see from the above table, the p-value for the all of the above independent variables is less than 0.05. Therefore, we reject the null hypothesis and conclude that none of the coefficients is 0. Therefore, given above t-test hypothesis results reinforce our F-test hypothesis results.

Part 1.3: Regression Analysis– Backward Elimination, Female Data

First, let’s take a look at the correlation table for the Female Data.

	GPA
GPA	1
Extracurricular Hours	-0.0315586
Number of Courses	0.139026
Classes missed	-0.339015
Studying Hours	0.195183
Employment hours	-0.0410907
Commute Hours	0.11999
Relationship Hours	0.00565062
Happiness_Content	-0.24566
Happiness_Happy	0.160733
Happiness_Unhappy	-0.0130296
Happiness_Very Happy	0.04577
Happiness_Very Unhappy	0.104733

- There is a weak positive relationship between Number of Courses, Studying Hours, Commut Hours, and Happiness_Happy and GPA
- There is a very weak positive relationship between Relationship Hours, Happiness_Very Happy, and Happiness_Very Unhappy and GPA
- There is a weak negative relationshop between Classes Missed and Happiness_Content and GPA
- Thtere is a very a weak negative relationship between Extracurricular Hours, Employment Hours, and Happiness Unhappy and GPA.

We have the following regression models from the backward elimination:

Run	Model
1	Yhat = 2.1768-0.0097*Extracurricular Hours+0.1210*Number of Courses-0.0289*Classes missed+0.0083*Studying Hours-0.0022*Employment Hours+0.0166*Commute Hours+0.0024*Relationship Hours+0.3262*Happiness_Content+0.5023*Happiness_Happy+0.2968*Happiness_Unhappy+0.5044*Happiness_Very Happy+0.5471*Happiness_Very Unhappy
2	Yhat = 2.1562-0.0091*Extracurricular Hours+0.1248*Number of Courses-0.0290*Classes missed+0.0084*Studying Hours+0.0150*Commute Hours+0.0025*Relationship Hours+0.3158*Happiness_Content+0.4922*Happiness_Happy+0.2977*Happiness_Unhappy+0.4944*Happiness_Very Happy+0.5561*Happiness_Very Unhappy
3	Yhat = 2.1532+0.1180*Number of Courses-0.0279*Classes missed+0.0087*Studying Hours+0.0121*Commute Hours+0.0023*Relationship Hours+0.3022*Happiness_Content+0.4710*Happiness_Happy+0.3027*Happiness_Unhappy+0.5049*Happiness_Very Happy+0.5724*Happiness_Very Unhappy
4	Yhat = 2.3175+0.0864*Number of Courses-0.0282*Classes missed+0.0098*Studying Hours++0.0018*Relationship Hours+0.3251*Happiness_Content+0.51*Happiness_Happy+0.3446*Happiness_Unhappy+0.4840*Happiness_Very Happy+0.6538*Happiness_Very Unhappy
5	Yhat = 2.4349+0.0589*Number of Courses-0.0287*Classes missed+0.0105*Studying Hours++0.3776*Happiness_Content+0.5327*Happiness_Happy+0.3565*Happiness_Unhappy+0.5138*Happiness_Very Happy+0.6633*Happiness_Very Unhappy
6	Yhat = 2.6599-0.0282*Classes missed+0.0111*Studying Hours++0.4022*Happiness_Content+0.5690*Happiness_Happy+0.4197*Happiness_Unhappy+0.551*Happiness_Very Happy+0.7180*Happiness_Very Unhappy
7	Yhat = 2.8351-0.0271*Classes missed++0.4141*Happiness_Content+0.5568*Happiness_Happy+0.5510*Happiness_Unhappy+0.5687*Happiness_Very Happy+0.7444*Happiness_Very Unhappy

8	Yhat = 2.7078+0.3041*Happiness_Content+0.6076*Happiness_Happy+0.4884*Happiness_Unhappy +0.6055*Happiness_Very Happy+0.7022*Happiness_Very Unhappy R-squared = 0.07

Therefore, we run a hypothesis test to determine if the final set of variables as a whole have effect on GPA.

$$H_0 : \beta_{Happiness Content} = \beta_{Happiness\ Happy} = \beta_{Happiness\ Unhappy} = \beta_{Happiness Very\ Happy} = \beta_{Happiness Very\ Unhappy} = 0$$
$$H_a : At\ least\ one\ \beta_i \neq 0$$

From the model summy output, the p-value = 0.474 !< α = 0.05. Therefore, we do reject the null hypothesis and conclude that the above variables do not have effect on GPA. Thus, we need to gather more data and perform further analysis. In short, regression model for Female Data is inconclusive. Interestingly, F-test result conflicts with the t-test (please see below) result for each of the variables as can be seen below.

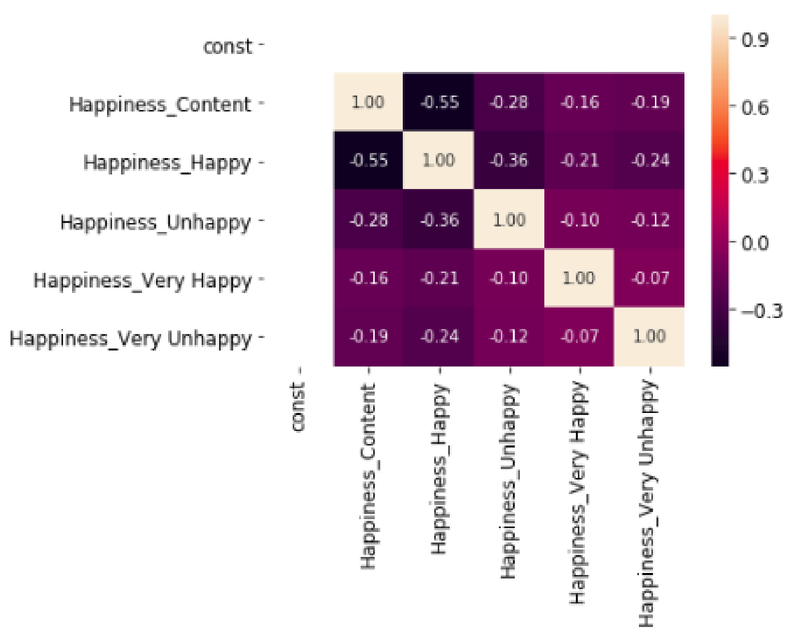
$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

for i being each of the independent variables from the final regresion model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    GPA   R-squared:                       0.070
Model:                            OLS   Adj. R-squared:                 -0.008
Method:                 Least Squares   F-statistic:                    0.8963
Date:                Sun, 02 Dec 2018   Prob (F-statistic):              0.474
Time:                        18:17:30   Log-Likelihood:                -40.431
No. Observations:                  53   AIC:                             90.86
Df Residuals:                      48   BIC:                             100.7
Df Model:                           4
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                      2.7078      0.082     32.981      0.000       2.543       2.873
Happiness_Content          0.3041      0.138      2.199      0.033       0.026       0.582
Happiness_Happy            0.6076      0.125      4.842      0.000       0.355       0.860
Happiness_Unhappy          0.4884      0.178      2.751      0.008       0.131       0.845
Happiness_Very Happy       0.6055      0.270      2.244      0.029       0.063       1.148
Happiness_Very Unhappy     0.7022      0.237      2.960      0.005       0.225       1.179
==============================================================================
Omnibus:                        2.096   Durbin-Watson:                   1.607
Prob(Omnibus):                  0.351   Jarque-Bera (JB):                2.012
Skew:                          -0.448   Prob(JB):                        0.366
Kurtosis:                       2.670   Cond. No.                     7.03e+15
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.41e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Lastly, independent variables are uncorrlated:



Part 2. Hypothesis Testing

Check if there exists a statistically significant difference between Male GPA and Female GPA

$$H_0: \mu_{Male\,GPA} - \mu_{Female\,GPA} = 0$$
$$H_a: \mu_{Male\,GPA} - \mu_{Female\,GPA} \neq 0$$

Since we are dealing with samples instead of actual populations, we have to work with t-test:

```
In [277]: st.ttest_ind(y_male, y_female, equal_var=False)
Out[277]: Ttest_indResult(statistic=-0.88830501314215926, pvalue=0.37650029348579206)
```

As we can see from above p-value = 0.3765 > α = 0.05. Therefore, do no reject the Null Hypothesis and conclude that there does not exist a statistically significant difference between Mean Male GPA and Mean Female GPA.

Check if there exists a difference between Variance of Male GPA and Variance of Female GPA.

$$H_0: \sigma^2_{Female\,GPA} = \sigma^2_{Male\,GPA}$$
$$H_a: \sigma^2_{Female\,GPA} \neq \sigma_{Male\,GPA}$$

```
In [322]:
male_var = np.std(y_male)**2

print(male_var)

female_var = np.std(y_female)**2

print(female_var)

f =  max(male_var, female_var) / min(male_var, female_var)

print(f)

0.21947156
0.28934335646849424
1.3183637846675633

In [279]: st.f.ppf(1 - 0.05/2, len(y_female) - 1, len(y_male) - 1)
Out[279]: 1.7515103907336533

In [280]: st.f.ppf(0.05/2, len(y_female) - 1, len(y_male) - 1)
Out[280]: 0.57361090501494394

In [281]:
p_value = 2*(1 - st.f.cdf(f, len(y_female) - 1, len(y_male) - 1))
print(p_value)

0.331486555998
```

Since p-value = 0.3314 > α = 0.05. Therefore, we do no reject the Null Hypothesis and conclude that there does not exist a statistically significant difference between Male GPA Variance and Female GPA Variance.

95% Confidence Intervals

GPA Mean Difference

In [282]:
```
cm = sms.CompareMeans(sms.DescrStatsW(y_male), sms.DescrStatsW(y_female))
print(cm.tconfint_diff(usevar='unequal'))
```
```
(-0.28783963413812308, 0.10979435111925603)
```

As we can see $\mu_{Male\,GPA} - \mu_{Female\,GPA}$ is between negative number and a postive number. Therefore, we can't conclude with certainty what gender performance better academically wise.

Ratio of GPA Variance

In [283]:
```
alpha = 0.05

female_var = np.std(y_female)**2
male_var = np.std(y_male)**2
ratio = female_var / male_var
f_stat = st.f.ppf(1 - alpha/2, len(y_female) - 1, len(y_male))
left = ratio / f_stat
right = ratio * f_stat
print((left, right))
```
```
(0.75553561901606636, 2.3004647629803161)
```

The ratio $\frac{\sigma^2_{Female\,GPA}}{\sigma^2_{Female\,GPA}}$ lies between values that are less than one and greater than one. Therefore, we cannot state with certainty whose GPA is spread more around the mean.

Mean

Male Mean

In [328]:
```
conf_inf_mean(y_male)
```
```
(2.990660048905057, 3.2569399510949415)
```

Female Mean

In [329]:
```
conf_inf_mean(y_female)
```
```
(3.0645572597092681, 3.3610880233096028)
```

Combined Mean

In [330]:
```
conf_inf_mean(y_combined)
```
```
(3.0704513799492696, 3.2687641540313113)
```

Variance

```
In [332]: conf_inf_var(y_male)
```

```
(0.15057514771229896, 0.33235422209142651)
```

Female Variance

```
In [333]: conf_inf_var(y_female)
```

```
(0.20060640721916939, 0.43264643321608598)
```

Group Variance

```
In [335]: conf_inf_var(y_combined)
```

```
(0.19744561482526019, 0.3418271691135511)
```

Conclusion

Our main results are as follows:

- General regression equation:
    - Yhat = 2.6674-0.229*Classes missed+0.0074*Studying Hours+0.4249*Happiness_Content+0.5686*Happiness_Happy * $I_{[0,1]}$ +0.37*Happiness_Unhappy * $I_{[0,1]}$ +0.5647*Happiness_Very Happy * $I_{[0,1]}$ +0.7392*Happiness_Very Unhappy * $I_{[0,1]}$
    - R-Squared = 0.228
- Male regression equation:
    - Yhat = 2.6977-0.0201*Classes missed+0.0062*Studying Hours-0.0113*Employment Hours+0.4519*Happiness_Content * $I_{[0,1]}$ +0.5869*Happiness_Happy * $I_{[0,1]}$ +0.2843*Happiness_Unhappy * $I_{[0,1]}$ +0.6201*Happiness_Very Happy * $I_{[0,1]}$ +0.7544*Happiness_Very Unhappy * $I_{[0,1]}$
    - R-Squared = 0.366
- Female regression equation (inconclusive):

- Yhat = 2.7078+0.3041*Happiness_Content * $I_{[0,1]}$ +0.6076*Happiness_Happy * $I_{[0,1]}$ +0.4884*Happiness_Unhappy * $I_{[0,1]}$ +0.6055*Happiness_Very Happy * $I_{[0,1]}$ +0.7022*Happiness_Very Unhappy * $I_{[0,1]}$
- R-squared = 0.07

- There is a not statistically significant difference between Average Male GPA and Average Female GPA
- There is not a statistically significant difference between Female GPA Variance and Male GPA Variance
- For all of the above regression equations, coefficient of determination is quiet low. This means our regression equations explain/predict less than 40% of the actual data. Does this mean that the actual regression model

$$Y=\beta_0+\beta_1*X_1+\beta_2*X_2+\beta_3*X_3+\beta_4*X_4+\beta_5*X_5+\beta_6*X_6+\beta_7*X_7+ \beta_8*X_8+\beta_9*X_9+\beta_{10}*X_{10}+\beta_{11}*X_{11}+\beta_{12}*X_{12}+\varepsilon$$

can explain the real life GPA data with less than 40% accuracy? No, it does not mean that. Our regression equations are built only for our sample data and sample data results should not be generalized to explain population data. What can be done to improve our results? First of all, we can increase sample data to more than 103 data points. Second, we can use other methods of regression equation building such as Stepwise Regression, Forward Selection, and Best-Subset Regression. Third, to make our hypothesis analysis and confidence intervals more precise, we can use α = 0.025, 0.01, 0.001. Fourth, we can use different sampling methods such as online surveying or phone surveys. Lastly, we can increase the number of independent variables (i.e. add such variables as Leisure Hours, Socializing Hours, and Sleeping Hours) to make our regression equations more precise.