**Advanced Statistics Project**

**From : 01st Nov,2020 to 18th Nov,2020**

**Group Members : Niranjan Dhavan.**

**Problem 1:**

**A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (A & B) in the compound were varied at three levels each. Randomization was used in assigning four volunteers to each of the nine treatments. The data on hours of relief can be found in the following .csv file: Fever.csv.**

* 1. **State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables ‘A’ and ‘B’ individually.**

**Insights:** An Analysis of Variance Test, or ANOVA, can be thought of as a generalization of the t-tests for more than 2 groups. The independent t-test is used to compare the means of a condition between two groups. ANOVA is used when we want to compare the means of a condition between more than two groups.

ANOVA tests if there is a difference in the mean somewhere in the model (testing if there was an overall effect), but it does not tell us where the difference is (if there is one). To find where the difference is between the groups, we have to conduct post-hoc tests.

To perform any tests, we first need to define the Null (H0) and Alternate Hypothesis (𝐻𝑎):

**The Hypothesis for the One-Way ANOVA For Variable “A” are:**

𝐻o : 𝐴𝑙𝑙 𝑡ℎ𝑒 𝑖𝑛𝑔𝑟𝑒𝑑𝑖𝑒𝑛𝑡 “𝐴” 𝑚𝑒𝑎𝑛𝑠 𝑜𝑓𝑎𝑙𝑙 𝑡ℎ𝑒 3 𝑙𝑒𝑣𝑒𝑙𝑠 𝑢𝑛𝑑𝑒𝑟𝑐𝑜𝑛𝑠𝑖𝑑𝑒𝑟𝑎𝑡𝑖𝑜𝑛 𝑎𝑟𝑒 𝑒𝑞𝑢𝑎𝑙.

𝐻a : 𝐴𝑡𝑙𝑒𝑎𝑠𝑡 𝑜𝑛𝑒 𝑜𝑓 𝑡ℎ𝑒 𝑖𝑛𝑔𝑟𝑒𝑑𝑖𝑒𝑛𝑡 “𝐴” 𝑚𝑒𝑎𝑛𝑠 𝑜𝑓 𝑎𝑙𝑙 𝑡ℎ𝑒 3 𝑙𝑒𝑣𝑒𝑙𝑠 𝑐𝑜𝑛𝑠𝑖𝑑𝑒𝑟𝑎𝑡𝑖𝑜𝑛 𝑎𝑟𝑒𝑢𝑛𝑒𝑞𝑢𝑎𝑙.

**The Hypothesis for the One-Way ANOVA For Variable “B” are:**

𝐻o : 𝐴𝑙𝑙 𝑡ℎ𝑒 𝑖𝑛𝑔𝑟𝑒𝑑𝑖𝑒𝑛𝑡 “B” 𝑚𝑒𝑎𝑛𝑠 𝑜𝑓𝑎𝑙𝑙 𝑡ℎ𝑒 3 𝑙𝑒𝑣𝑒𝑙𝑠 𝑢𝑛𝑑𝑒𝑟 𝑐𝑜𝑛𝑠𝑖𝑑𝑒𝑟𝑎𝑡𝑖𝑜𝑛 𝑎𝑟𝑒 𝑒𝑞𝑢𝑎𝑙.

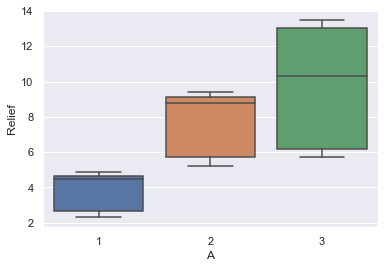
𝐻a : 𝐴𝑡𝑙𝑒𝑎𝑠𝑡 𝑜𝑛𝑒 𝑜𝑓 𝑡ℎ𝑒 𝑖𝑛𝑔𝑟𝑒𝑑𝑖𝑒𝑛𝑡“B” 𝑚𝑒𝑎𝑛𝑠 𝑜𝑓𝑎𝑙𝑙 𝑡ℎ𝑒 3 𝑙𝑒𝑣𝑒𝑙𝑠 𝑐𝑜𝑛𝑠𝑖𝑑𝑒𝑟𝑎𝑡𝑖𝑜𝑛 𝑎𝑟𝑒 𝑢𝑛𝑒𝑞𝑢𝑎𝑙.

* 1. **Perform one-way ANOVA for variable ‘A’ with respect to the variable ‘Relief’. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results**.

**First Step :-** After defining the hypothesis for conducting One Way Anova the first step is to load the data set into pandas and perform EDA using python functions like describe for descriptive statistics ,head- to know the columns, tail- to check for totals if any and info for checking null values and conversion of integer/float data into categorical.

**Second Step: -** The second step in analysing the data is to examine the data visually. This allows the experimenter of Research Laboratory to gauge how the amounts of the active ingredient A in the compound at three different levels has impact on Reliefof severe cases of hay fever. Let’s load the data and view it as a box chart.

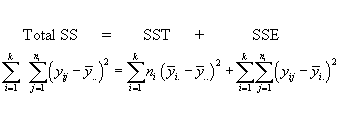
The visual Representation Variable “A” against Y Variable “Relief”.



**Inference:** So, what do we see? Visually, we can see that level 2 and level 3 has more impact on relief as compared to level 1 of active ingredients A. Additionally, it appears that active ingredient A with level 2 and level 3 might contribute to increasing relief around the mean but I can’t say for certain. So, what is someone to do? Run the ANOVA on the data.

# **Third Step: - Sum of Squares**

Now comes the heart of the ANOVA, calculating the sum of squares but first let’s define a few more things.

Where,

SST is the sum of squares between treatments SST

SSE is the sum of squares within treatments (error),

Total SS is the total of SST and SSE

Let’s convert these equations into a python function to quickly compute these equations for us.

One Way ANOVA with the variable 𝑖𝑛𝑔𝑟𝑒𝑑𝑖𝑒𝑛𝑡 ‘𝐴’ with respect to 'Relief'. The output of One Way Anova by running a code in python is:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Particulars** | **df** | **sum\_sq** | **mean\_sq** | **F** | **PR(>F)** |
| **C(A)** | 2 | 220.02 | 110.01 | 23.46539 | 4.58E-07 |
| **Residual** | 33 | 154.71 | 4.688182 | NaN | NaN |

**Inference:** Now, we see that the corresponding p-value is less than alpha (0.05). Thus, we 𝐟𝐚𝐢𝐥 𝐭𝐨 𝐀𝐜𝐜𝐞𝐩𝐭 the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 (𝐻o). This means at least one particular 𝑖𝑛𝑔𝑟𝑒𝑑𝑖𝑒𝑛𝑡 ‘A’ type has different mean of 'Relief' as compared to the other 𝑖𝑛𝑔𝑟𝑒𝑑𝑖𝑒𝑛𝑡 A Type.

As we rejected the Null Hypothesis in the F stat, we need to check which of particular 𝑖𝑛𝑔𝑟𝑒𝑑𝑖𝑒𝑛𝑡 ‘A’ type has a different mean.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Multiple comparison of means |  |  | Tukey HSD |  | FWER=0.05 | |
| group1 | group2 | meandiff | p-adj | lower | upper | reject |
| 1 | 2 | 3.95 | 0.001 | 1.7814 | 6.1186 | TRUE |
| 1 | 3 | 5.95 | 0.001 | 3.7814 | 8.1186 | TRUE |
| 2 | 3 | 2 | 0.0755 | -0.1686 | 4.1686 | FALSE |

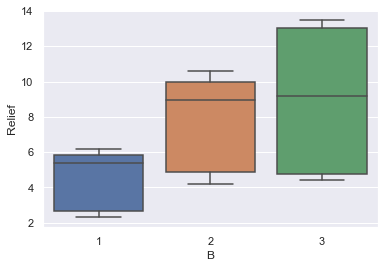
Basis the Tukey HSD it was identified that level 2 & level 3 have same mean relief time for ingredient 'A' with respect to the Relief. Whereas Level 1 & 2, level 1 & 3 have different means relief time for ingredient 'A ‘with respect to the Relief. **Thus, we reject the null hypothesis for 1 & 2 and 1 & 3 pairs but fail to reject for the 2 & 3 pair. This means at Level 2 and Level 3 for ingredient A the relief would be higher as compared with level 1.**

* 1. **Perform one-way ANOVA for variable ‘B’ with respect to the variable ‘Relief’. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.**

**First Step :-** After defining the hypothesis for conducting One Way Anova the first step is to load the data set into pandas and perform EDA using python functions like describe for descriptive statistics ,head- to know the columns, tail- to check for totals if any and info for checking null values and conversion of integer/float data into categorical.

**Second Step: -** The second step in analysing the data is to examine the data visually. This allows the experimenter of Research Laboratory to gauge how the amounts of the active ingredient B in the compound at three different levels has impact on Reliefof severe cases of hay fever. Let’s load the data and view it as a box chart.

The visual Representation Variable “B” against Y Variable “Relief”.



**Inference:** So, what do we see? Visually, we can see that level 2 and level 3 has more impact on relief as compared to level 1 of active ingredients B. Additionally, it appears that active ingredient B with level 2 and level 3 might contribute to increasing relief around the mean but I can’t say for certain. So, what is someone to do? Run the ANOVA on the data.

# **Third Step: - Sum of Squares**

Let us now perform One Way ANOVA with the variable 𝑖𝑛𝑔𝑟𝑒𝑑𝑖𝑒𝑛𝑡 ‘B’ with respect to ' Relief '.

The output of One Way Anova by running a code in python is:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Particulars** | **df** | **sum\_sq** | **mean\_sq** | **F** | **PR(>F)** |
| **C(B)** | 2 | 123.66 | 61.83 | 8.126777 | 0.00135 |
| **Residual** | 33 | 251.07 | 7.608182 | NaN | NaN |

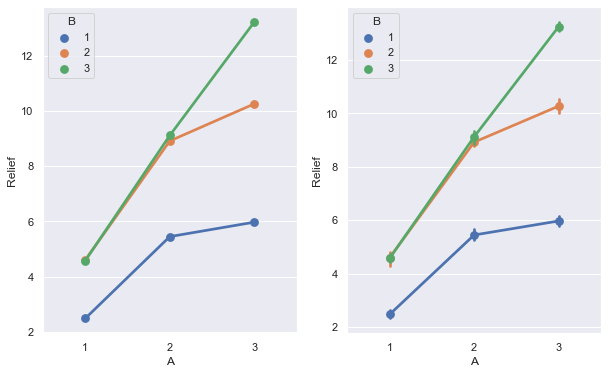
**Inference:** Now, we see that the corresponding p-value is less than alpha (0.05). Thus, we 𝐟𝐚𝐢𝐥 𝐭𝐨 𝐀𝐜𝐜𝐞𝐩𝐭 the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 (𝐻o). This means at least one particular𝑖𝑛𝑔𝑟𝑒𝑑𝑖𝑒𝑛𝑡 ‘B’ type has a different average Relief value as compared to the other 𝑖𝑛𝑔𝑟𝑒𝑑𝑖𝑒𝑛𝑡 ‘B’ Type.

As we rejected the Null Hypothesis in the F stat, we need to check which particular 𝑖𝑛𝑔𝑟𝑒𝑑𝑖𝑒𝑛𝑡 ‘B’ type has a different mean.

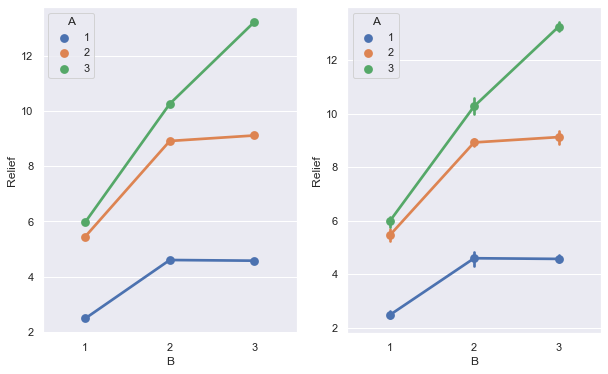
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Multiple comparison of means** |  |  | **Tukey HSD** |  | **FWER=0.05** | |
| **group1** | **group2** | **meandiff** | **p-adj** | **lower** | **upper** | **reject** |
| 1 | 2 | 3.3 | 0.0164 | 0.5374 | 6.0626 | TRUE |
| 1 | 3 | 4.35 | 0.0014 | 1.5874 | 7.1126 | TRUE |
| 2 | 3 | 1.05 | 0.6164 | -1.7126 | 3.8126 | FALSE |
|  |  |  |  |  |  |  |

Basis the Tukey HSD it was identified that level 2 & level 3 have same mean relief time for ingredient 'B' with respect to the Relief. Whereas Level 1 & 2, level 1 & 3 have different means relief time for ingredient 'B' with respect to the Relief. **Thus, we reject the null hypothesis for the 1 & 2 and 1 & 3 pairs but fail to reject for the 2 & 3 pair. This means at Level 2 and Level 3 for ingredient B the relief would be higher as compared with level 1.**

* 1. **Analyse the effects of one variable on another with the help of an interaction plot.**
* Visual interpretation of Interaction of ingredient 'A' with Y variable ‘Relief’ Considering Hue as Variable ‘B’ with confidence intervals ci=99 & ci=None.



* Visual interpretation of Interaction of ingredient 'B' with Y variable ‘Relief’ Considering Hue as Variable ‘A’ with confidence intervals ci=99 & ci=None.



**Inference:** Basis the point plot we identify that there is slight interaction between variables.

* 1. **Perform a two-way ANOVA based on the different ingredients (variable ‘A’ & ‘B’) with the variable 'Relief' and state your results.**

**Insights:** To perform any tests, we first need to define the Null (H0) and Alternate Hypothesis (𝐻𝑎):

**The Hypothesis for the Two-Way ANOVA For Variables “A” and “B” are:**

𝐻o: The mean values for Ingredient “A” and Ingredient “B” 𝑎𝑟𝑒 𝑒𝑞𝑢𝑎𝑙.

𝐻a: 𝐴𝑡𝑙𝑒𝑎𝑠𝑡 𝑜𝑛𝑒 group from Ingredient “A” and Ingredient “B” is significantly different from another.

Let us now perform the Two-Way ANOVA. We will now analyse the effect of both the ingredient ‘A’& ingredient ‘B’ on the Relief variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **df** | **sum\_sq** | **mean\_sq** | **F** | **PR(>F)** |
| **C(A)** | 2 | 220.02 | 110.01 | 109.8329 | 8.51E-15 |
| **C(B)** | 2 | 123.66 | 61.83 | 61.73044 | 1.55E-11 |
| **Residual** | 31 | 31.05 | 1.001613 | NaN | NaN |

**Inference:** The p-value in the both the treatments is less than 𝛼(0.05) and thus we will reject the Null Hypothesis.

Let us check whether there is any interaction effect between the ingredient A & ingredient B.

As we can see that there is some sort of interaction between the two variables from above Point plot (Question 1.4), we shall introduce a new term while performing the Two-Way ANOVA.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **df** | **sum\_sq** | **mean\_sq** | **F** | **PR(>F)** |
| **C(A)** | 2 | 220.02 | 110.01 | 1827.858 | 1.51E-29 |
| **C(B)** | 2 | 123.66 | 61.83 | 1027.329 | 3.35E-26 |
| **C(A):C(B)** | 4 | 29.425 | 7.35625 | 122.2269 | 6.97E-17 |
| **Residual** | 27 | 1.625 | 0.060185 | NaN | NaN |

Due to the inclusion of the interaction effect term, we can see a change in the p-value of the first two treatments as compared to the Two-Way ANOVA without the interaction effect terms. As we see p-value of the interaction effect term of 'ingredient A' and 'ingredient B' is smaller than Alpha - suggesting that the fail to accept Null Hypothesis. Hence there is interaction between the ingredient A & ingredient B.

**1.6 Mention the business implications of performing ANOVA for this particular case study.**

ANOVA is test of difference in group means. It is used when more than two group means are compared. For two group means, we can do t-test. If p-value is less than significance level, it can be interpreted that there is no evidence in support of hypothesis that there is no difference in group means. Or, there is significant difference among group means. If there is no difference than the group may not be useful for model building and can be excluded and vise versa.

ANOVA is widely used across businesses and industries for a variety of purposes and is a technique that enables companies to identify problems, trends, risks, and opportunities that impact both short and long-term viability. Below are just a few of its many applications within business scenarios:

**Quality and cost comparison:** In the automotive industry, a car manufacturing company can use ANOVA when purchasing materials to compare the quality of the material to the costs so that they know which supplier to buy from or who should build their products. For example, if they need to purchase steel for door frames, they can compare the strength of the steel to how much it will cost to purchase.

**Product safety tests:** In the beauty industry, cosmetics companies can use ANOVA to test the safety and effectiveness of certain makeups or sunscreen products, for example. They can evaluate these products across different groups of people and then choose to use the ingredients that provide the desired outcomes while minimizing health risks.

**Optimize production:** In entertainment and media, companies can use ANOVA to test what different locations for filming an upcoming movie, and determine which site would be best based on the time of the year, or how much materials would cost for building a set, enabling them to choose the most cost-effective production company.

**Industry-Wide Approach**: ANOVA is effective for a wide variety of uses across different industries, including financial services, ecommerce, industrial, R&D, and more.

**Product Development**: Organizations can better pinpoint and understand what product features to improve or adapt for the best results.

**Inference:**

In the given case a research laboratory is developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (A & B) in the compound were varied at three levels each. Randomization was used in assigning four volunteers to each of the nine treatments. We can see that the research laboratory’s is engaged in product development and performing R&D.

Basis this model we can say that Relief time can be increased based on combination of both the ingredient “A” & ingredient “B”. Except the combination of A1--B2 A1--B3, A2--B1 A3--B1, A2--B2 A2--B3. Rest all of the combination has the interaction amongst both the ingredient A & ingredient B. So therefore, we can conclude that on combination of all ingredient A & ingredient B except combination of A1--B2 A1--B3, A2--B1 A3--B1, A2--B2 A2--B3 can increase the relief time of severe cases of hay fever.

**Problem 2:**

**A company performed a survey to understand the income of households in various neighborhoods of a country. The data dictionary is also present. You can access the data dictionary from the following file**[Income\_Data Dictionary](https://olympus.greatlearning.in/courses/10649/files/1371547/download?verifier=KYMhnNbpD0JG9w1QLWv8erqJs2ZLB8dDfZT5j0df&wrap=1" \t "_blank" \o "Income_Data Dictionary-1.pdf)**. Please refer to the following data set to solve the problem**[Income.csv](https://olympus.greatlearning.in/courses/10649/files/1371546/download?verifier=wphtKCgO4MQkdLHwKfhu2DhIChHtVVl3f4QcoFqW&wrap=1)**. ['FamilyIncome' is the target variable].**

**Income Data**

Number of observations: 753

Observation: Individuals

Continent: Europe

A data frame containing: Description

1. Working Hours Wife - Hours that a wife is working

2. Wife Age - Age of the wife

3. Education Wife - Educational attained by wife in years

4. Wife Hour Earnings - The average hourly earnings of the wife in euros

5. Wife Wage - The wage of the wife reported at the time

6. Working Hours Husband - Hours that the husband is working

7. Husband Age - Age of the husband

8. Education Husband - Education attained by husband in years

9. Husband Wage - The wage of the husband reported at the time

10. Education Wife Mother - Education attained by the wife’s mother in years

11. Education Wife Father - Education attained by the wife’s father in years

12. Unemployment Rate - unemployment rate in county of residence, in percentage points

13. Wife Experience - Actual years of wife's previous labor market experience

14. Family Income-Income of the family in euros

**2.1) Perform exploratory data analysis on the dataset. Showcase some charts, graphs.**

1. Loading the data set- We will be loading the” Income.csv” file using pandas library in python. For this, we will be using read\_csv file.

2. The head function will tell you the top head records in the data set. By default, python shows you only the top 5 records, but we will check top 10 records.

3. The tail function will tell you the top tail records in the data set. By default, python shows you only the top 5 records, but we will check top 10 records for the totals/subtotals if any. The Income dataset doesn’t contain any total/subtotals.

4. The shape attribute tells us a number of observations and variables we have in the data set. It is used to check the dimension of data. The Income data set has 753 observations and 14 variables in the data set.

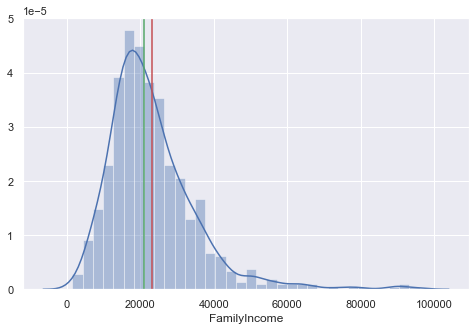
5. info() is used to check the Information about the data and the datatypes of each respective attribute.

Looking at the data in the head function and in info, we come to know that the all the variables comprise of float and integer data types which doesn’t requires conversion. Further there are no null values in dataset and the count is 753 for all the variables.

6. The described method will help to see how data has been spread for numerical values. We can clearly see the minimum value, mean values, different percentile values, and maximum values for the Income data set.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Particulars** | **WorkingHoursWife** | **WifeAge** | **EducationWife** | **WifeHourEarnings** | **WifeWage** | **WorkingHoursHusband** | **HusbandAge** | **EducationHusband** | **HusbandWage** | **EducationWifeMother** | **EducationWifeFather** | **UnemploymentRate** | **WifeExperience** | **FamilyIncome** |
| **count** | 753.0 | 753.0 | 753.0 | 753.0 | 753.0 | 753.0 | 753.0 | 753.0 | 753.0 | 753.0 | 753.0 | 753.0 | 753.0 | 753.0 |
| **mean** | 740.6 | 42.5 | 12.3 | 2.4 | 1.8 | 2267.3 | 45.1 | 12.5 | 7.5 | 9.3 | 8.8 | 8.6 | 10.6 | 23080.6 |
| **std** | 871.3 | 8.1 | 2.3 | 3.2 | 2.4 | 595.6 | 8.1 | 3.0 | 4.2 | 3.4 | 3.6 | 3.1 | 8.1 | 12190.2 |
| **min** | 0.0 | 30.0 | 5.0 | 0.0 | 0.0 | 175.0 | 30.0 | 3.0 | 0.4 | 0.0 | 0.0 | 3.0 | 0.0 | 1500.0 |
| **25%** | 0.0 | 36.0 | 12.0 | 0.0 | 0.0 | 1928.0 | 38.0 | 11.0 | 4.8 | 7.0 | 7.0 | 7.5 | 4.0 | 15428.0 |
| **50%** | 288.0 | 43.0 | 12.0 | 1.6 | 0.0 | 2164.0 | 46.0 | 12.0 | 7.0 | 10.0 | 7.0 | 7.5 | 9.0 | 20880.0 |
| **75%** | 1516.0 | 49.0 | 13.0 | 3.8 | 3.6 | 2553.0 | 52.0 | 15.0 | 9.2 | 12.0 | 12.0 | 11.0 | 15.0 | 28200.0 |
| **max** | 4950.0 | 60.0 | 17.0 | 25.0 | 10.0 | 5010.0 | 60.0 | 17.0 | 40.5 | 17.0 | 17.0 | 14.0 | 45.0 | 96000.0 |

7. Check whether the dependent variable is normally distributed or not (symmetric or not)

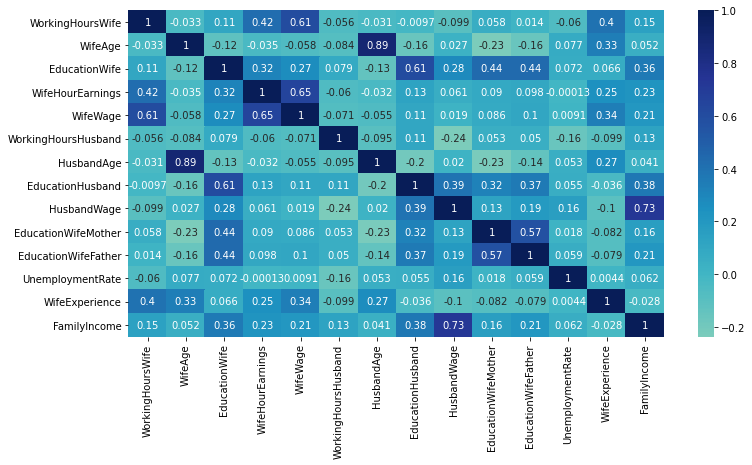


The distribution doesn’t seem to be perfectly symmetric as the mean and median values are different for dependent variable. The distribution tends to be right skewed.

\*Further graphs are mentioned in below solution.

**2.2) Is there evidence of multicollinearity? Showcase your analysis.**

**Insights: -** Multicollinearity is problem that you can run into when you’re fitting a regression model. Simply put, multicollinearity is when two or more independent variables in a regression are highly related to one another, such that they do not provide unique or independent information to the regression. We can check multicollinearity using a correlation plot or pair plot to see which parameters have multicollinearity issue.



**Inference: -** From the above heatmap we see that there is some degree of correlation amongst the variables, but no perfect correlation exists between independent variables,

1.There are no variables with Perfectly Positive or Negative correlation.

2. Fairly Strong Positive Relationship

- Wife’s age and Husbands age.

3.Moderate Positive Relationship

-Between Wife wage and Working hours wife.

- Between Education Husband and Education Wife.

- Between Wife wage and Wife Hour Earning.

- Between Family Income and Husband Wage.

- Between Education wife father and Education wife mother.

4.No correlation

-Wife Experience and unemployment rate

-Wife wage and unemployment rate

-Wife Houe earnings and Unemployment Rate

-Education husband and Working Hour Wife

**2.3) Perform Multiple Linear Regression (using the 'statsmodels' library) and comment on the model thus built.**

**Insights:** Multiple Linear Regression attempts to model the relationship between two or more features and a response by fitting a linear equation to observed data. The steps to perform multiple linear Regression are almost similar to that of simple linear Regression. The Difference Lies in the evaluation. We can use it to find out which factor has the highest impact on the predicted output and now different variable relate to each other.

Here : **Y = b0 + b1 \* x1 + b2 \* x2 + b3 \* x3 + …… bn \* xn**

# Y- Dependent Variable (FamilyIncome) and

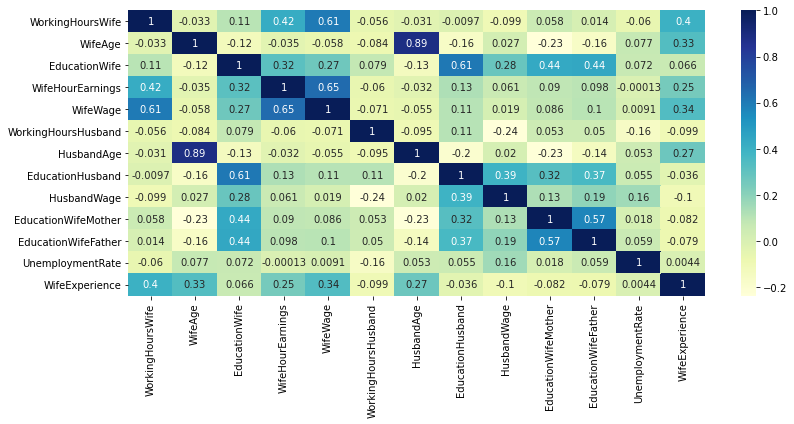
# X- Independent variables (All other variables)

'WorkingHoursWife', 'WifeAge', 'EducationWife', 'WifeHourEarnings', 'WifeWage', 'WorkingHoursHusband', 'HusbandAge', 'EducationHusband', 'HusbandWage', 'EducationWifeMother', 'EducationWifeFather', 'UnemploymentRate', 'WifeExperience'

Before we get to the actual model building exercise, it is important to separate the Independent variables (X) from the Dependent Variables (Y). The goal here is to predict/estimate the “Family Income” based on variables: “working hours, age and earnings of husband and wife.

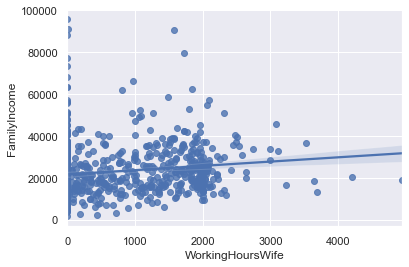
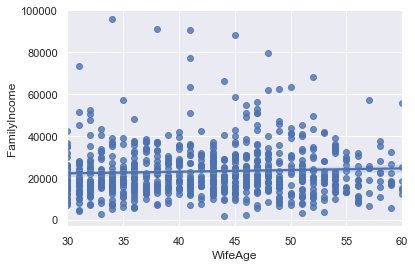
### ****Assumptions of Linear Regression****

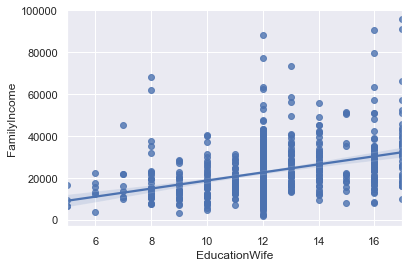
**Assumption 1: Independent Variables should not exhibit correlations- This can be checked by plotting correlation matrix for the independent variables or more visually intuitive way of checking correlations is to plot a heat map from the seaborn library.**

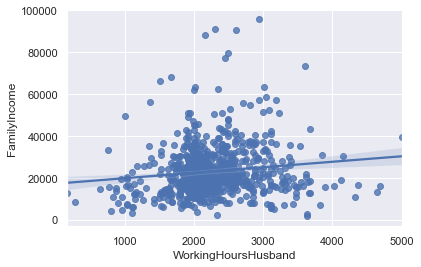
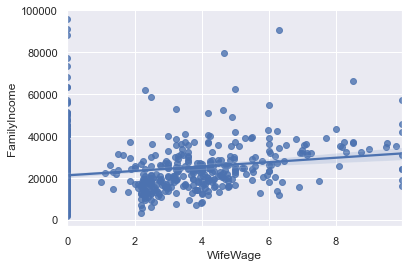


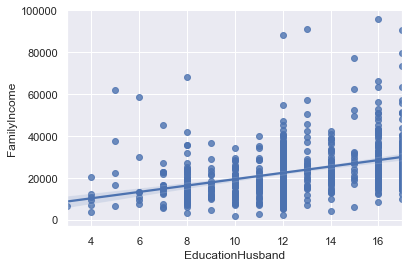
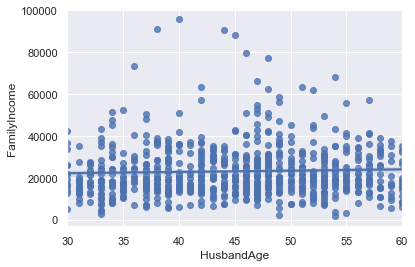
**Inferences*: Here we can see that there is a strong relationship between variable WifeAge and HusbandAge but we assume that the primary objective is to make predictions and we don't want to understand the role of each independent variables hence severity of multicollinearity is ignored.So Assumption 1 holds true.***

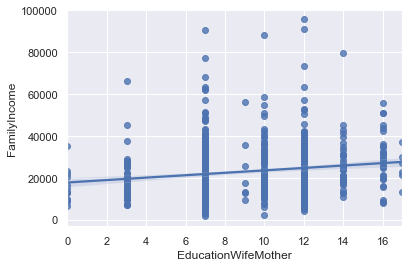
**Assumption 2: Linear Regression means that the dependent variable should be linearly related with the coefficients. - We can Check Linearity with all Variables using Regplot (Seaborn Library Python)**

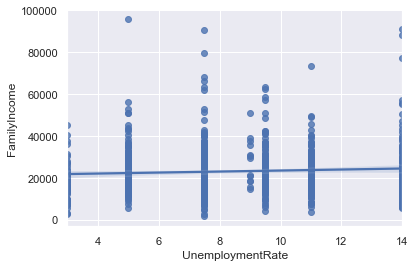
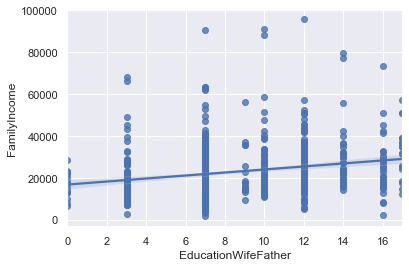
**** ****

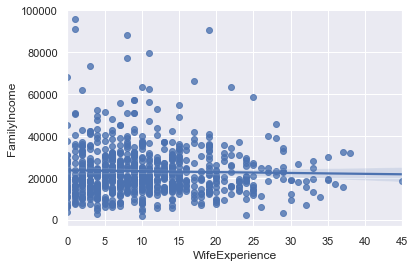
****

****

****

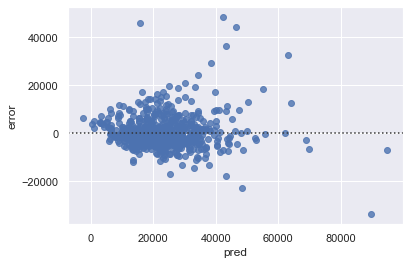
**** ****

****

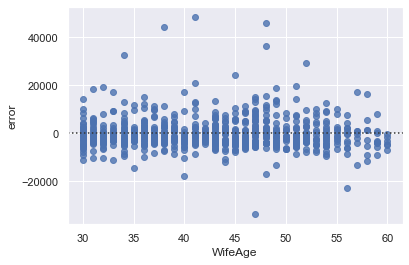


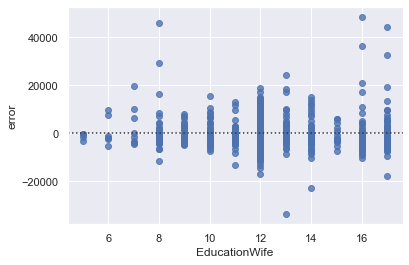
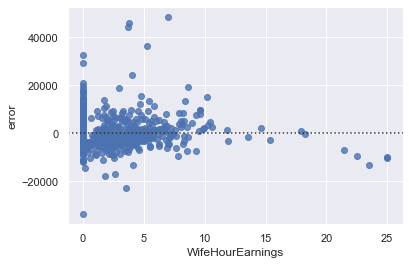
**Inference:** Here we can see that the independent variables has a liner relation with the dependent variable Family Income**.**

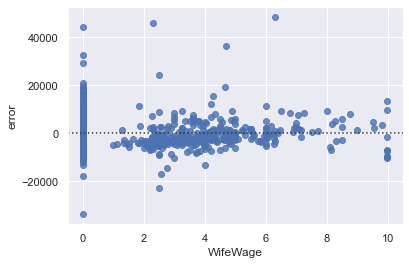
**Assumption 3: The error terms has a constant variance i.e. they are homoscedastic in nature. - This assumption can be checked by obtaining predictions and errors. We will first plot the Residuals (or errors) against the Predicted value (Y-hat).**

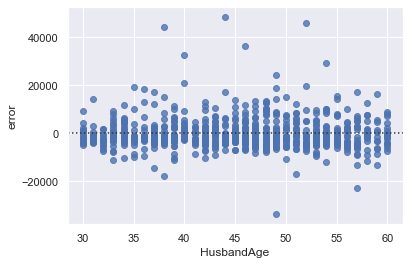
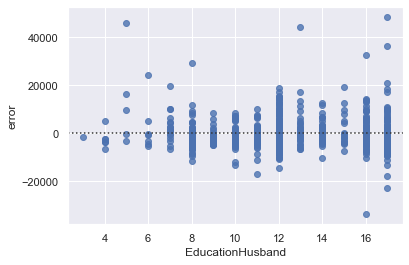
****

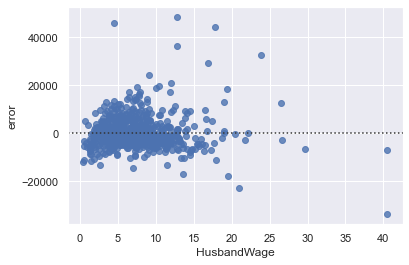
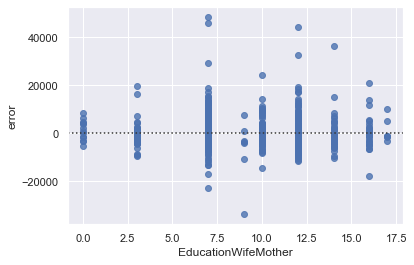
**Inference :** In the given problem we can see that there is some sort of pattern between error and pred. Typically, if there is any hint of Heteroscedasticity between the Residuals and Predicted scores, as a second level of analysis, we should be looking at the residuals against each X variable to isolate the problem variables(s) causing this problem and come up with a course of action. Visually (sample images) the residuals doesn’t seems to be homoscedastic.

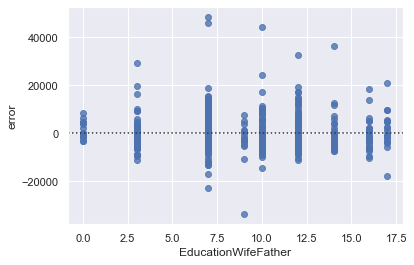
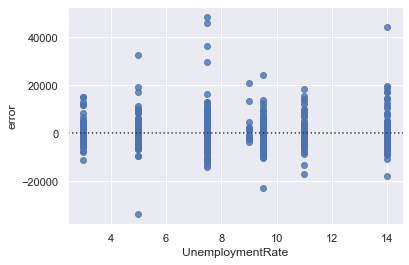
 

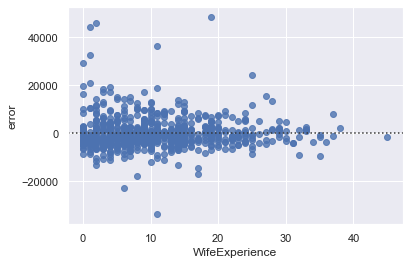
** **

** **

** **

** **

** **

****

**Assumption 4: There should not be any auto-correlation between the error terms. -** The variation in the independent variable which is explained by the dependent variables is 70.50 %

**Inference: Here, we see that the Durbin-Watson test statistic is 2.07 and thus we can say that this particular assumption of Linear Regression is also verified.**

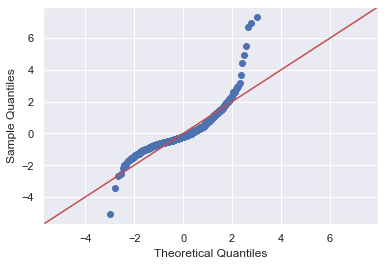
**Assumption 5: The errors are assumed to be normally distributed - Let us run the Shapiro test of normality to check whether the errors are normally distributed.**

Hypothesis for Shapiro Test

Ho: The distribution is normal

Ha: The distribution is not normal

**Inference:** Since the p-value (9.731716305268925e-28) is less than α (0.05), we can say that the errors are not normally distributed, and this particular assumption does not hold true. But by using the QQ plot we can say that the Residuals (or errors) are normally distributed.



**Inference:** Using the QQ plot we can say that the Residuals (or errors) are normally distributed.

By using the syntax and adding a constant to perform the linear regression in Python using statsmodels we get:

|  |  |  |  |
| --- | --- | --- | --- |
| **OLS Regression Results** |  |  |  |
| **Dep. Variable:** | FamilyIncome | R-squared: | 0.705 |
| **Model:** | OLS | Adj. R-squared: | 0.7 |
| **Method:** | Least Squares | F-statistic: | 135.9 |
| **Date:** | Sun, 15 Nov 2020 | Prob (F-statistic): | 8.01E-186 |
| **Time:** | 21:17:12 | Log-Likelihood: | -7692.7 |
| **No. Observations:** | 753 | AIC: | 1.54E+04 |
| **Df Residuals:** | 739 | BIC: | 1.55E+04 |
| **Df Model:** | 13 |  |  |
| **Covariance Type:** | nonrobust |  |  |

\*The P-value for the analysis of variance F-test (P = 8.01e-186) suggests that the at least one of the beta coefficients is non zero in predicting the dependent variable.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **const** | -2.22E+04 | 2455.663 | -9.057 | 0 | -2.71E+04 | -1.74E+04 |
| **WorkingHoursWife** | 2.7718 | 0.372 | 7.456 | 0 | 2.042 | 3.502 |
| **WifeAge** | 134.1663 | 68.226 | 1.967 | 0.05 | 0.227 | 268.106 |
| **EducationWife** | 375.2334 | 153.202 | 2.449 | 0.015 | 74.471 | 675.995 |
| **WifeHourEarnings** | 310.9829 | 101.621 | 3.06 | 0.002 | 111.482 | 510.484 |
| **WifeWage** | 277.4616 | 154.003 | 1.802 | 0.072 | -24.875 | 579.798 |
| **WorkingHoursHusband** | 6.779 | 0.441 | 15.356 | 0 | 5.912 | 7.646 |
| **HusbandAge** | 27.4612 | 66.582 | 0.412 | 0.68 | -103.251 | 158.174 |
| **EducationHusband** | -75.3354 | 110.423 | -0.682 | 0.495 | -292.115 | 141.444 |
| **HusbandWage** | 2286.1687 | 67.84 | 33.699 | 0 | 2152.986 | 2419.351 |
| **EducationWifeMother** | 31.6954 | 92.865 | 0.341 | 0.733 | -150.616 | 214.007 |
| **EducationWifeFather** | 24.6254 | 87.378 | 0.282 | 0.778 | -146.913 | 196.164 |
| **UnemploymentRate** | -50.9882 | 80.405 | -0.634 | 0.526 | -208.838 | 106.862 |
| **WifeExperience** | -107.5475 | 36.789 | -2.923 | 0.004 | -179.771 | -35.323 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 404.335 | **Durbin-Watson:** | 2.073 |
| **Prob(Omnibus):** | 0 | **Jarque-Bera (JB):** | 5410.657 |
| **Skew:** | 2.107 | **Prob(JB):** | 0 |
| **Kurtosis:** | 15.438 | **Cond. No.** | 2.49E+04 |

## Interpreting the Regression Results

I highlighted several important components within the results:

1. **Adjusted. R-squared** reflects the fit of the model. R-squared values range from 0 to 1, where a higher value generally indicates a better fit, assuming certain conditions are met.
2. **const coefficient** is your Y-intercept. It means that if all the independent variables coefficient are zero, then the expected output (i.e., the Y) would be equal to the const coefficient.
3. **Working Hours Wife coefficient** represents the change in the output Y due to a change of one unit in the working hours of wife (everything else held constant)
4. **std err**reflects the level of accuracy of the coefficients. The lower it is, the higher is the level of accuracy
5. **P >|t|** is your p-value. A p-value of less than 0.05 is considered to be statistically significant
6. **Confidence Interval** represents the range in which our coefficients are likely to fall (with a likelihood of 95%)

**2.4) Perform Principal Component Analysis (on the predictor variables) and extract the Principal Components. Comment on the reason behind choosing the number of Principal Components.**

**Insights:** Since, we will be doing Principal Component Analysis on the data to reduce the dimensions let us go ahead and drop the target variable 'FamilyIncomme'. Further now that the dataset has been loaded, it must be prepared for dimensionality reduction. The majority of machine learning and optimization algorithms perform better when all the features are along the same scale. In order to do this a standardization approach can be implemented.

**Step 1:** Before we go ahead and perform the Principal Component Analysis, let us build the covariance matrix.

Covariance measures how two features vary with each other. A positive covariance indicates that features increase and decrease together. Whereas, a negative covariance indicates that the two features vary in the opposite directions.

Covariance Matrix %s

[[ 1.00132979e+00 -3.31582115e-02 1.06101321e-01 4.23506899e-01 6.07723445e-01 -5.64225183e-02 -3.11300884e-02 -9.66326289e-03 -9.87300540e-02 5.79406915e-02 1.36890778e-02 -6.03700052e-02 4.05497226e-01]

[-3.31582115e-02 1.00132979e+00 -1.20382861e-01 -3.46051004e-02 -5.83924775e-02 -8.44837698e-02 8.89319009e-01 -1.63266264e-01 2.70507312e-02 -2.34953587e-01 -1.60804404e-01 7.71793469e-02 3.34460049e-01]

[ 1.06101321e-01 -1.20382861e-01 1.00132979e+00 3.18801449e-01 2.67930359e-01 7.90208661e-02 -1.33699059e-01 6.12767546e-01 2.85315019e-01 4.35915402e-01 4.43046609e-01 7.22359587e-02 6.63436673e-02]

[ 4.23506899e-01 -3.46051004e-02 3.18801449e-01 1.00132979e+00 6.52507648e-01 -5.99985405e-02 -3.18782667e-02 1.26391585e-01 6.13711932e-02 9.04253051e-02 9.86077958e-02 -1.27896664e-04 2.50913296e-01]

[ 6.07723445e-01 -5.83924775e-02 2.67930359e-01 6.52507648e-01 1.00132979e+00 -7.08913431e-02 -5.54725311e-02 1.07108799e-01 1.93018022e-02 8.57115891e-02 1.02909091e-01 9.13632427e-03 3.42011117e-01]

[-5.64225183e-02 -8.44837698e-02 7.90208661e-02 -5.99985405e-02 -7.08913431e-02 1.00132979e+00 -9.55138763e-02 1.07988079e-01 -2.36334670e-01 5.34247037e-02 5.04123675e-02 -1.55426002e-01 -9.94983848e-02]

[-3.11300884e-02 8.89319009e-01 -1.33699059e-01 -3.18782667e-02 -5.54725311e-02 -9.55138763e-02 1.00132979e+00 -1.95582291e-01 1.97070726e-02 -2.27759226e-01 -1.35179749e-01 5.31644460e-02 2.72272038e-01]

[-9.66326289e-03 -1.63266264e-01 6.12767546e-01 1.26391585e-01 1.07108799e-01 1.07988079e-01 -1.95582291e-01 1.00132979e+00 3.95189608e-01 3.24906192e-01 3.67187193e-01 5.50903396e-02 -3.63490142e-02]

[-9.87300540e-02 2.70507312e-02 2.85315019e-01 6.13711932e-02 1.93018022e-02 -2.36334670e-01 1.97070726e-02 3.95189608e-01 1.00132979e+00 1.26904382e-01 1.93485974e-01 1.58130174e-01 -1.03443569e-01]

[ 5.79406915e-02 -2.34953587e-01 4.35915402e-01 9.04253051e-02 8.57115891e-02 5.34247037e-02 -2.27759226e-01 3.24906192e-01 1.26904382e-01 1.00132979e+00 5.73833806e-01 1.84266371e-02 -8.22881828e-02]

[ 1.36890778e-02 -1.60804404e-01 4.43046609e-01 9.86077958e-02 1.02909091e-01 5.04123675e-02 -1.35179749e-01 3.67187193e-01 1.93485974e-01 5.73833806e-01 1.00132979e+00 5.86096606e-02 -7.89069628e-02]

[-6.03700052e-02 7.71793469e-02 7.22359587e-02 -1.27896664e-04 9.13632427e-03 -1.55426002e-01 5.31644460e-02 5.50903396e-02 1.58130174e-01 1.84266371e-02 5.86096606e-02 1.00132979e+00 4.44127241e-03]

[ 4.05497226e-01 3.34460049e-01 6.63436673e-02 2.50913296e-01 3.42011117e-01 -9.94983848e-02 2.72272038e-01 -3.63490142e-02 -1.03443569e-01 -8.22881828e-02 -7.89069628e-02 4.44127241e-03 1.00132979e+00]]

**Step 2:** Let us now get the eigen values and the eigen vectors.

The eigenvectors represent the principal components (the directions of maximum variance) of the covariance matrix. The eigenvalues are their corresponding magnitude.

Eigen Vectors [[-0.17742383 0.42818966 0.27289732 -0.01800772 0.02140832 0.10248336 0.04906251 0.16921007 0.5533919 -0.38195272 -0.14860013 -0.43375229 -0.07761001]

[ 0.2686461 0.32073721 -0.46948303 0.24081122 0.71604282 0.00652029 0.01636776 -0.12039671 0.01878517 0.00937483 0.06594523 -0.1174905 0.02087811]

[-0.44361296 0.02791376 -0.21775337 0.14759737 0.00704313 -0.16389846 -0.10670115 0.07993251 -0.37441051 0.06785842 -0.68811238 -0.26211575 0.03935585]

[-0.2760925 0.38312234 0.13260116 -0.08785453 0.00248872 -0.15014214 -0.05824887 -0.50545136 -0.34121163 -0.48969245 0.16968773 0.24980465 0.15114727]

[-0.27877776 0.43697551 0.20093792 -0.09615522 0.01433441 -0.04139521 -0.05283352 -0.2293794 0.0947791 0.76645265 0.12449233 -0.00541794 -0.1114797 ]

[-0.03355309 -0.13593613 0.16562636 0.65079552 -0.01506437 -0.23533967 -0.49972219 -0.13060046 0.35680902 0.00489237 -0.05394212 0.23643394 0.14453663]

[ 0.27112606 0.30979566 -0.45967565 0.24016386 -0.69104393 0.03175631 0.06216562 -0.21903496 0.04356164 0.0110717 0.02394737 -0.16613812 -0.03082722]

[-0.39030187 -0.10135941 -0.25204265 0.07982126 -0.0355829 -0.40281185 -0.11302597 0.30144107 -0.05714899 -0.0718493 0.60851914 -0.27391906 -0.21580811]

[-0.19662968 -0.05685545 -0.41803185 -0.37587903 -0.0065251 -0.37844575 0.20102704 -0.04765237 0.50427169 0.00380223 -0.1618307 0.34392121 0.23947192]

[-0.37574411 -0.13293597 -0.09920381 0.20490276 -0.00918022 0.49186323 0.20538294 -0.0202883 0.04309676 0.07715042 0.23041677 -0.126309 0.65410627]

[-0.37110939 -0.1133324 -0.19958985 0.20443877 0.03604699 0.43144101 0.16254322 -0.13980461 0.12374186 -0.09663615 -0.04565706 0.34370628 -0.62412078]

[-0.03269247 0.01480671 -0.26834078 -0.41712699 -0.02104951 0.37611081 -0.77723504 0.01153374 0.05903811 -0.03934607 0.02264449 -0.0212219 0.02416107]

[-0.00222212 0.46537947 -0.03229387 0.127637 -0.07522618 0.08326945 -0.01688772 0.68126879 -0.14279958 0.00299443 0.01100643 0.50553325 0.11398277]]

Eigen Values [2.98225435 2.41033838 1.78562682 1.2353852 0.10674915 0.92972654 0.85994247 0.69081582 0.54007441 0.28991404 0.32473922 0.44163312 0.42008771]

**Step 3:** Let us now calculate the variance explained by eigen values and the cumulative variance by the eigen values (TOTAL VARIATION).

The eigenvector that has the largest corresponding eigenvalue represents the direction of maximum variance.

The variance explained by each of eigen values in order is

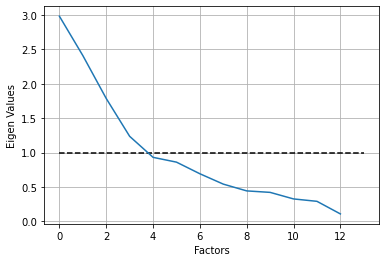
[22.909952727776258, 18.516441546049972, 13.717349758794843, 9.490342942168782, 7.1422449686941984, 6.606157297491247, 5.306910765021288, 4.148901360032483,

3.392666313185718, 3.2271525130873675, 2.494676606741444, 2.227146383598584, 0.8200568173578443]

Cumulative Variance Explained

[22.90995273 41.42639427 55.14374403 64.63408697 71.77633194 78.38248924 83.68940001 87.83830137 91.23096768 94.45812019 96.9527968 99.17994318 100.

**Step 4:** Let us now plot the variance explained by each eigen value with the eigen value.



**Inference:** From the above plot, we can see that the number of components that we can probably take is 5. We also see that if we take 5 components the total amount of variance explained is 71.77%

**Step 5:** Let us now calculate the Principal Components.

From statsmodels.import multivariate.pca import PCA and run the code so as to obtain reduced dimensions which are as follows,

Final Dimensions (head only):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Component No.** | **comp\_0** | **comp\_1** | **comp\_2** | **comp\_3** | **comp\_4** |
| **Renamed As** | **Wife Age & Education** | **Wife Experience and Wife Mother Education** | **Working Hours and Age Wife** | **Unemployment and working Hrs Husband** | **EducationWifeMother and EducationHusband** |
| 0 | -0.01911 | -0.00174 | -0.06764 | -0.02092 | 0.000299 |
| 1 | -0.00518 | -0.01593 | -0.05804 | 0.054033 | -0.00511 |
| 2 | -0.01985 | 0.019512 | -0.06504 | -0.04149 | 0.004189 |
| 3 | 0.022086 | -3E-06 | -0.02862 | -0.00075 | 0.006209 |
| 4 | -0.05516 | -0.00921 | -0.02832 | 0.033167 | -0.03116 |

Now let us check correlation among the components



**Inference:** Now, we see that the correlation amongst the variables have been decreased.

**2.5) Perform Multiple Linear Regression with 'FamilyIncome' as the dependent variable and the Principal Components extracted as the independent variables.**

Multiple Linear Regression attempts to model the relationship between two or more features and a response by fitting a linear equation to observed data. The steps to perform multiple linear Regression are almost similar to that of simple linear Regression. The Difference Lies in the evaluation. We can use it to find out which factor has the highest impact on the predicted output and now different variable relate to each other.

Here : **Y = b0 + b1 \* x1 + b2 \* x2 + b3 \* x3 + …… bn \* xn**

# Y- Dependant Variable(FamilyIncome) and

# X- Independant variables (Final Dimensions)

'Wife Age & Education','Wife Experience and Wife Mother Education','Working Hours and Age Wife','Unemployment and working Hrs Husband','EducationWifeMother and EducationHusband'

Before we get to the actual model building exercise, it is important to seperate the Independent variables (X) from the Dependent Variables (Y). The goal here is to predict/estimate the “Family Income” based on independent variables identified as PCA

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | familyincome | R-squared: | 0.437 |
| **Model:** | OLS | Adj. R-squared: | 0.433 |
| **Method:** | Least Squares | F-statistic: | 115.8 |
| **Date:** | Mon, 16 Nov 2020 | Prob (F-statistic): | 1.34E-90 |
| **Time:** | 14:49:10 | Log-Likelihood: | -7936.4 |
| **No. Observations:** | 753 | AIC: | 1.59E+04 |
| **Df Residuals:** | 747 | BIC: | 1.59E+04 |
| **Df Model:** | 5 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **Std err** | **t** | **p>t** | **[0.025** | **0.975]** |
| **const** | 2.31E+04 | 334.555 | 68.989 | 0 | 2.24E+04 | 2.37E+04 |
| **Wife Age & Education** | -1.39E+05 | 9180.474 | -15.138 | 0 | -1.57E+05 | -1.21E+05 |
| **Wife Experience and Wife Mother Education** | 2.76E+04 | 9180.474 | 3.001 | 0.003 | 9528.514 | 4.56E+04 |
| **Working Hours and Age Wife** | 1.15E+05 | 9180.474 | 12.539 | 0 | 9.71E+04 | 1.33E+05 |
| **Unemployment and working Hrs Husband** | 2.40E+04 | 9180.474 | 2.612 | 0.009 | 5952.926 | 4.20E+04 |
| **EducationWifeMother and EducationHusband** | 1.22E+05 | 9180.474 | 13.292 | 0 | 1.04E+05 | 1.40E+05 |

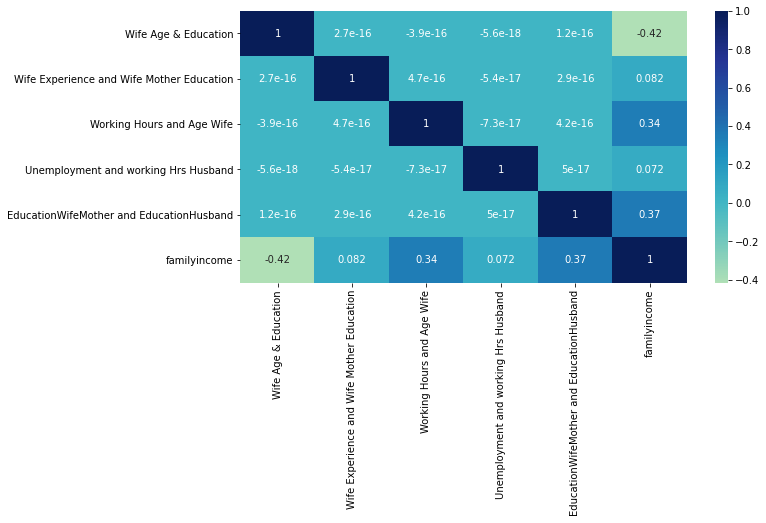
## Interpreting the Regression Results

I highlighted several important components within the results:

1. **Adjusted. R-squared** reflects the fit of the model. R-squared values range from 0 to 1, where a higher value generally indicates a better fit, assuming certain conditions are met.
2. **const coefficient** is your Y-intercept. It means that if all the independent variables coefficient are zero, then the expected output (i.e., the Y) would be equal to the const coefficient.
3. **Wife Age and Education coefficient** represents the change in the output Y due to a change of one unit in the working hours of wife (everything else held constant)
4. **std err**reflects the level of accuracy of the coefficients. The lower it is, the higher is the level of accuracy
5. **P >|t|** is your p-value. A p-value of less than 0.05 is considered to be statistically significant
6. **Confidence Interval** represents the range in which our coefficients are likely to fall (with a likelihood of 95%)

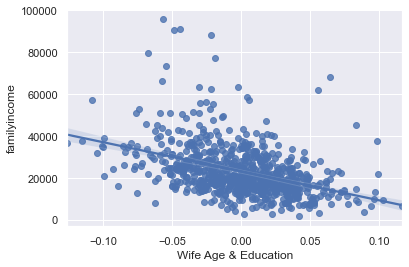
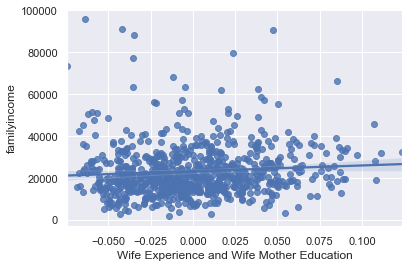
### **Assumptions of Linear Regression**

**Assumption 1: Independent Variables should not exhibit correlations- This can be checked by plotting correlation matrix for the independent variables or more visually intuitive way of checking correlations is to plot a heat map from the seaborn library.**



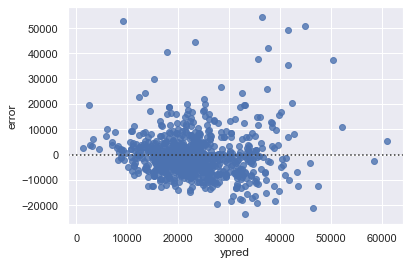
**Inference:** The independent variables don't exhibit very high correlations amongst themselves. So, this assumption holds true.

**Assumption 2: Linear Regression means that the dependent variable should be linearly related with the coefficients.- We can Check Linearity with all Variables using Regplot (Seaborn Library Python)**

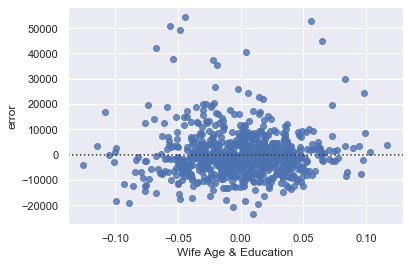
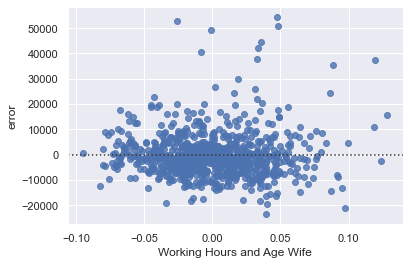
****

**Inference:** It can be seen from the above sample plots that linear relation exists between dependent and independent variables.

**Assumption 3:** The error terms has a constant variance i.e. they are homoscedastic in nature. - This assumption can be checked by obtaining predictions and errors. We will first plot the Residuals (or errors) against the Predicted value (Y-hat).



**Inference :** In the given problem we can see that there is some sort of pattern between error and pred. Typically, if there is any hint of Heteroscedasticity between the Residuals and Predicted scores, as a second level of analysis, we should be looking at the residuals against each X variable to isolate the problem variables(s) causing this problem and come up with a course of action.

Visually (sample images) the residuals seem to be homoscedastic.

**Assumption 4:** There should not be any auto-correlation between the error terms. (One value of the error term should not predict the next value of the error term)

The Durbin-Watson test statistic ranges from a value of 0 to 4. wherever the statistic is exactly equal to 2, we can go ahead and say that there is zero autocorrelation. However, as a thumb rule if this statistic lies between 1.5-2.5, the model may still be deployed.

**Inference: Here, we see that the Durbin-Watson test statistic is 1.95 and thus we can say that this particular assumption of Linear Regression is also verified**

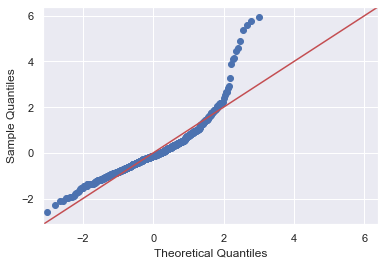
**Assumption 5**: Errors are normally distributed.

Hypothesis for Shapiro Test

\*Ho: The distribution is normal

\* Ha: The distribution is not

Since the p-value (1.398379559175213e-33) is less than α (0.05), we can say that the errors are not normally distributed, and this particular assumption does not hold true. But by using the QQ plot we can say that the Residuals (or errors) are normally distributed.



**Inference:** Using the QQ plot we can say that the Residuals (or errors) are normally distributed.

**2.6) Comment on the Model thus built using the Principal Components and with 'FamilyIncome'.**

Interpretation of the principal components is based on finding which variables are most strongly correlated with each component, i.e., which of these numbers are large in magnitude, the farthest from zero in either direction. Which numbers we consider to be large or small is of course is a subjective decision. You need to determine at what level the correlation is of importance. Here a correlation (*Ref Assumption 1 above*) above 0.5 is deemed important. By doing PCA we have reduced dimensions to 5 from 14 with help of machine learning tool.

**2.7) Mention the business implication and interpretation of the models.**

This model can be used in sanctioning loan /credit card by Banks.

Commercial banks receive a lot of applications for loan/ credit cards. Many of them get rejected for many reasons, like high loan balances, low income levels, or too many inquiries on an individual's credit report, for example. Manually analyzing these applications is mundane, error-prone, and time-consuming (and time is money!). Luckily, this task can be automated with the power of machine learning and pretty much every commercial bank does so nowadays. In this project, we have built an credit card approval predictor using machine learning techniques, just like the real banks do. The dataset used in this project is the [Income].