**Multivariable data analysis and**

**Hypothesis Testing**

**Using Linear and Multiple Regression**

**Ammar Rashed**

**CS 240**

**Prof. Mehmet Baysan**

**Abstract:**

In this project we are analyzing a set of data about (Gender, Age, Time spent on Facebook and GPA) of sample students and reporting the correlation between variables (two or more). Using a simple GUI, the user can choose freely from the data set variables and by a single click he can analyze the correlation, significance of the apparent effect and create a model that he can use in the same GUI to predict certain values accordingly. The user can get Residuals, R-squared, P-Value, F-value, SSE, Mean, Standard Deviation and many other statistics.

**Introduction:**

In part I we are to study the correlation between only two variables, especially the correlation between each variable and the GPA. The first and necessary step to analyze any data is to setup a model for it. There might be other suitable models for our data, but I chose to work with *Least Square* method to create a fit-line for the data, which I find plausible enough as we are dealing with only two variables; the independent variable or the predictor and the dependent variable or the variable we wish to predict according to the predictor. However, there still exists the probability that the result is just due to chance, which requires us to test the significance of the results. This process is called *Hypothesis Testing*where we state couple of hypothesis and test them. The first hypothesis is the *Null Hypothesis* that is the hypothesis that the correlation exists due to chance, or in our case, e.g. the Time spent on Facebook does not significantly affect a student’s GPA. The other hypothesis is the *Alternative Hypothesis* which is simply the opposite of the Null Hypothesis. We are to test the Null Hypothesis and in order to do that we need to determine the proper statistic to use and set our critical level according to which the validity of the hypothesis is to be decided. In part I we are dealing with variables two by two and, therefore, *Difference in Means* can be a suitable statistic to test the significance. Moreover, by calculating the *residuals*, the *R2,* Pearson correlation coefficient and other statistics that might help in analyzing the data, we can simply test the significance using F-statistic which is built on the degrees of freedom and the Mean Square Errors (MSE) of the model. Nonetheless, significance test using different testing statistics might not always give the same result, yet it is always better to have more options and thus understand the data better.

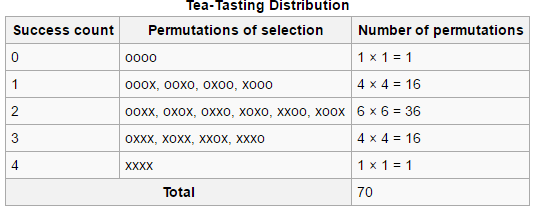
In part II we are studying the effect of multiple variables on a single dependent variable, which is typical problem for Multiple Regression models. After establishing the Multiple Regression Model and calculating the *residuals* and the error parameters of the model we can initiate the Hypothesis Testing process. There is variety of test statistics we can choose from, I chose F-statistic due to its simplicity and, moreover, we wouldn’t calculate more statistics to get the F-value since we had already calculated the Total Sum of Squares TSS, Explained Sum of Squares ESS and Residual Sum of Squares RSS and the degrees of freedom for each of them.

Using the models in Part I and Part II, we can predict a variable given others (e.g. a student’s GPA given his age and gender), assess how accurate our prediction is and lastly decide whether the apparent effect is significant or not.

* **Literature Review:**

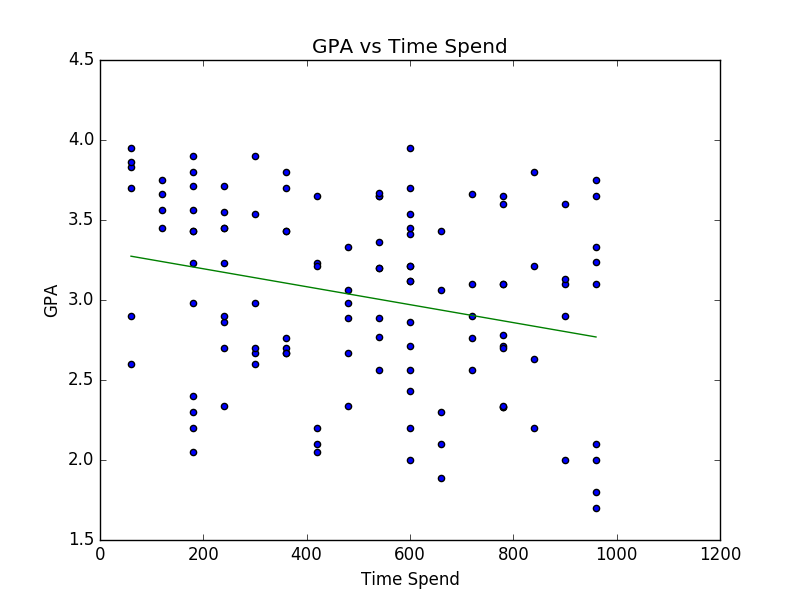
Hypothesis Testing and Regression has been an interesting field of study in statistics, and here are some of those whose names were mentioned coupled to those topics:

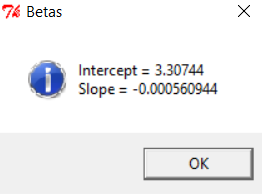
* **Ronald Fisher: Lady tasting tea 1913**
* The experiment provided the Lady with 8 randomly ordered cups of tea—4 prepared by first adding milk, 4 prepared by first adding the tea. She was to select the 4 cups prepared by one method.
* This offered the Lady the advantage of judging cups by comparison.
* The Lady was fully informed of the experimental method.
* The null hypothesis was that the Lady had no such ability.
* Note that in Fisher's approach, there is no alternative hypothesis;[2] this is instead a feature of the Neyman–Pearson approach.
* The test statistic was a simple count of the number of successes in selecting the 4 cups.
* The null hypothesis distribution was computed by the number of permutations. The number of selected permutations equalled the number of unselected permutations.

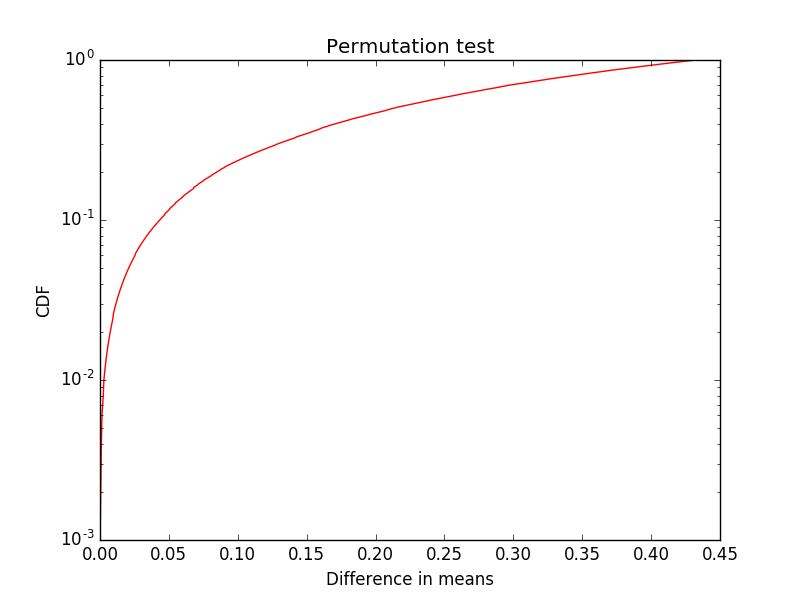
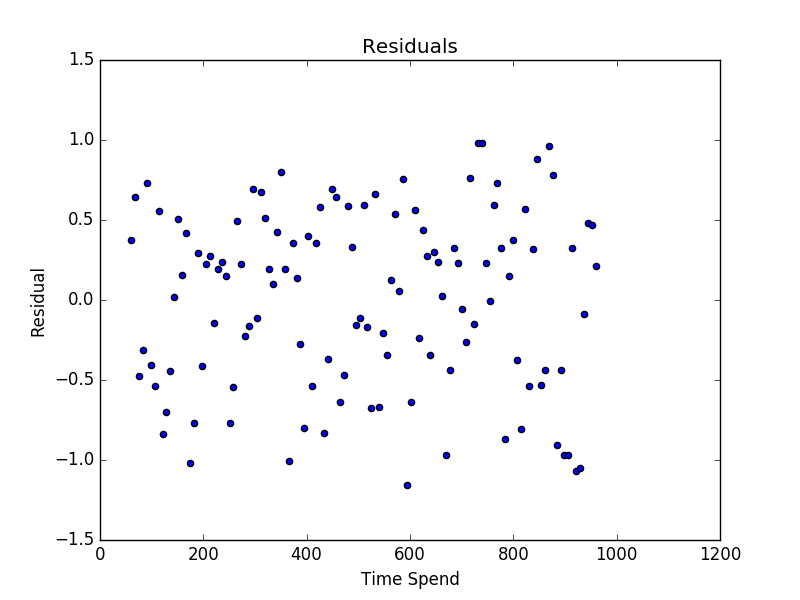


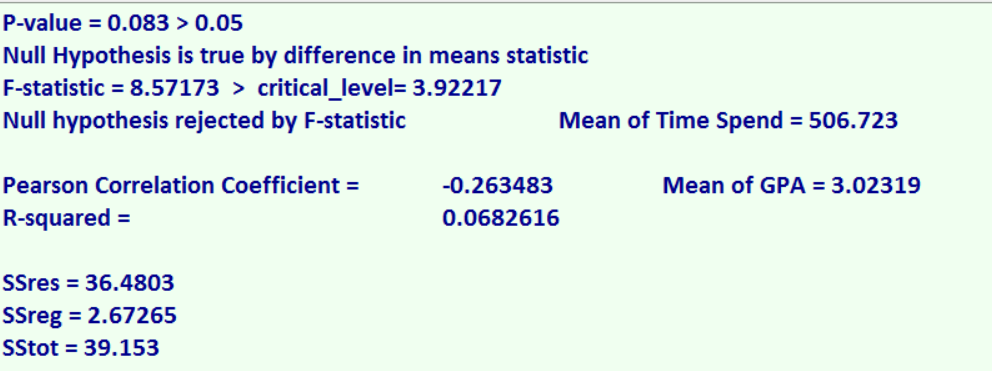
* **Philosopher's Beans:**
* produced by a philosopher describing scientific methods generations before hypothesis testing was formalized and popularized.
* “Few beans of this handful are white.  
  Most beans in this bag are white.  
  Therefore: Probably, these beans were taken from another bag.  
  This is an hypothetical inference.”
* The beans in the bag are the population. The handful are the sample. The null hypothesis is that the sample originated from the population. The criterion for rejecting the null-hypothesis is the "obvious" difference in appearance (an informal difference in the mean). The interesting result is that consideration of a real population and a real sample produced an imaginary bag. The philosopher was considering logic rather than probability. To be a real statistical hypothesis test, this example requires the formalities of a probability calculation and a comparison of that probability to a standard.
* A simple generalization of the example considers a mixed bag of beans and a handful that contain either very few or very many white beans. The generalization considers both extremes. It requires more calculations and more comparisons to arrive at a formal answer, but the core philosophy is unchanged; If the composition of the handful is greatly different from that of the bag, then the sample probably originated from another bag. The original example is termed a one-sided or a one-tailed test while the generalization is termed a two-sided or two-tailed test.
* The statement also relies on the inference that the sampling was random. If someone had been picking through the bag to find white beans, then it would explain why the handful had so many white beans, and also explain why the number of white beans in the bag was depleted (although the bag is probably intended to be assumed much larger than one's hand).
* ***Honorable mentions***
* **Allen B. Downey, *ThinkStats2, chapters 10,11.***
* **Prof. Serna.**
* **Results:**

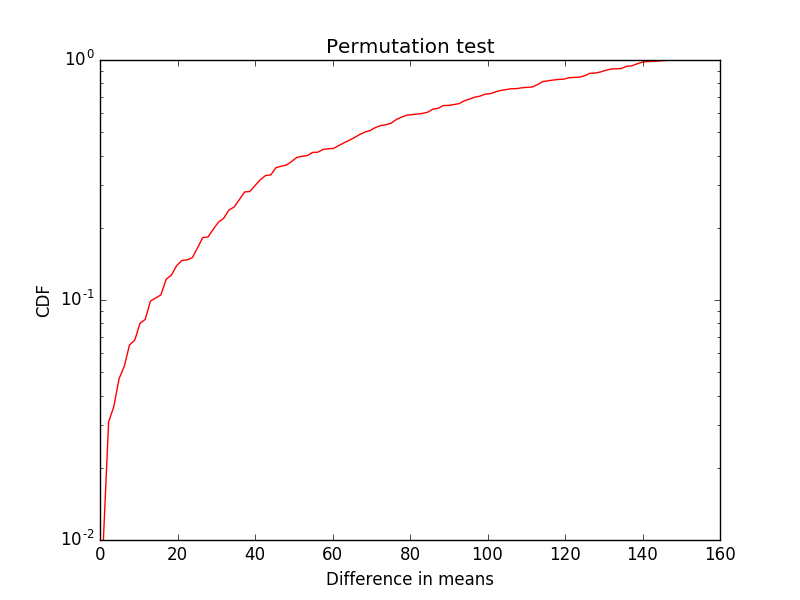
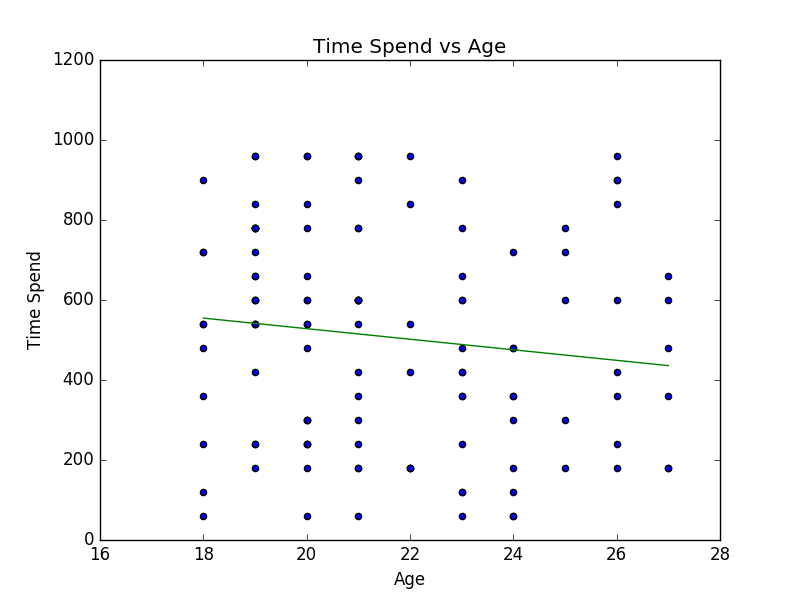
**Time spent on FB vs GPA:**

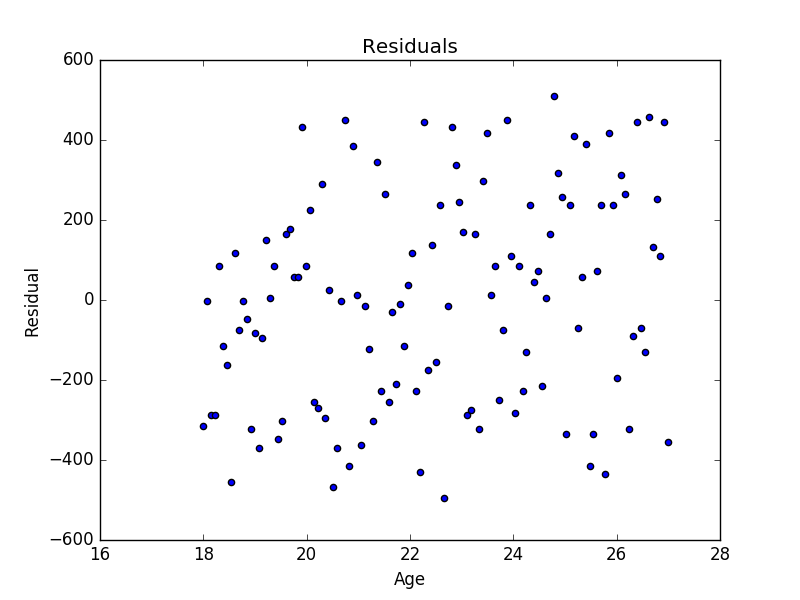
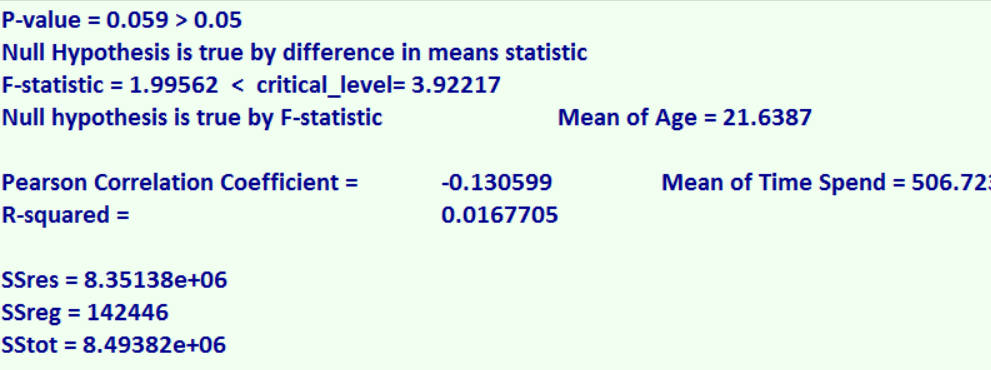
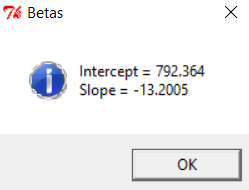
****

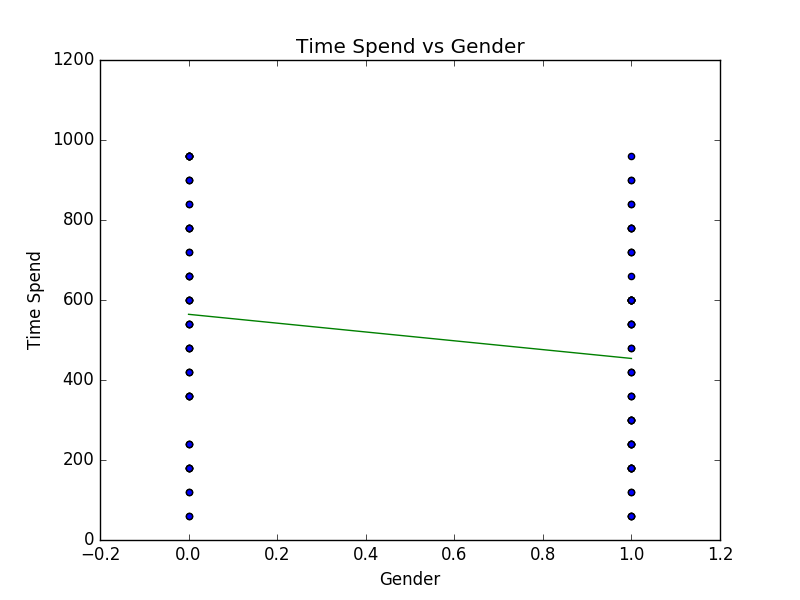
****

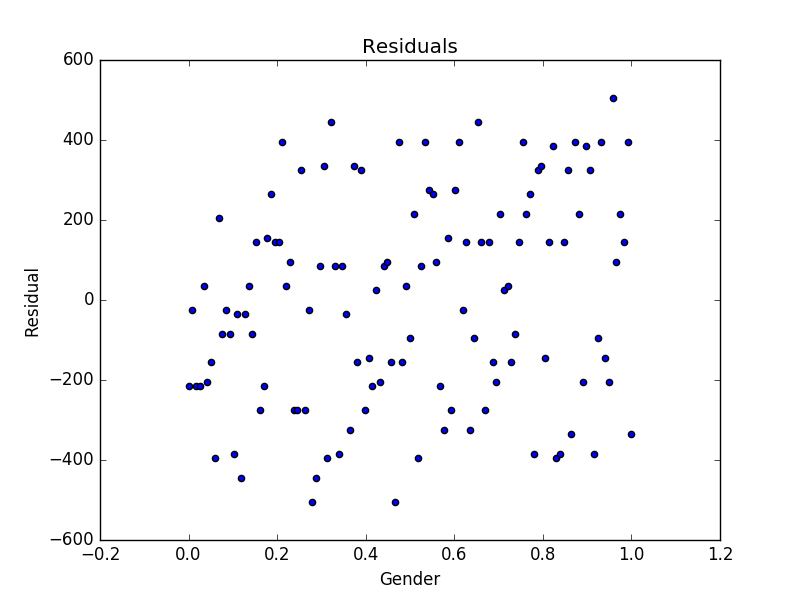
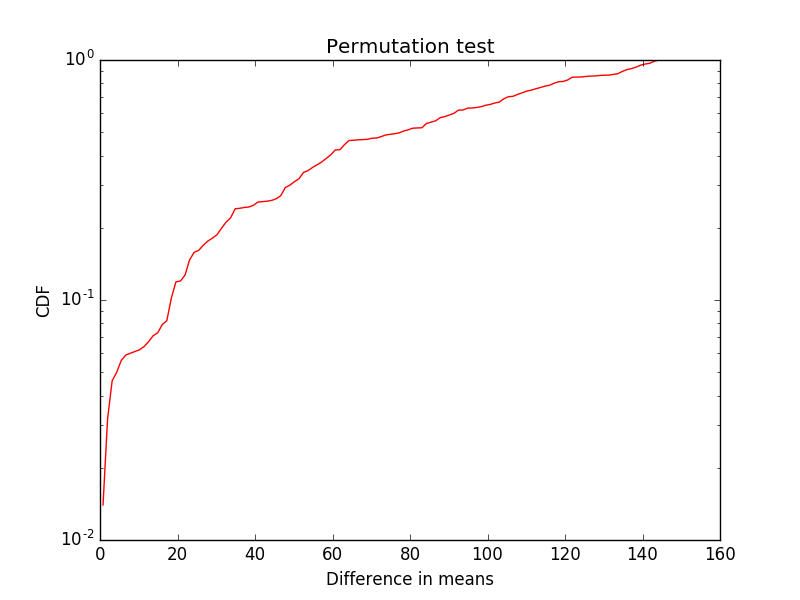
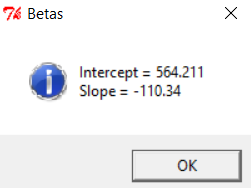
****

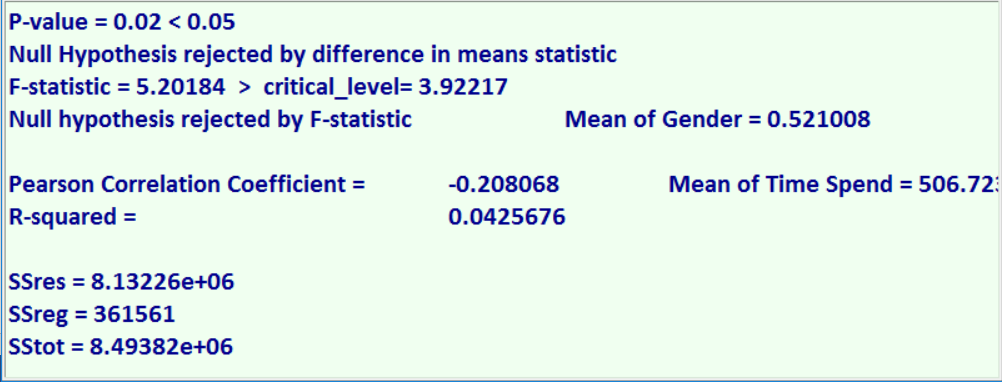
****

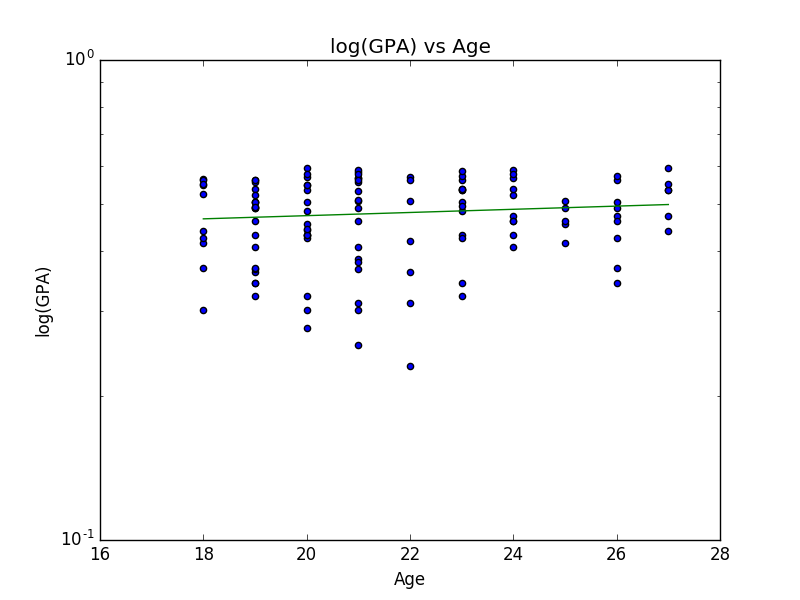
**Time Spent on FB and Age:**

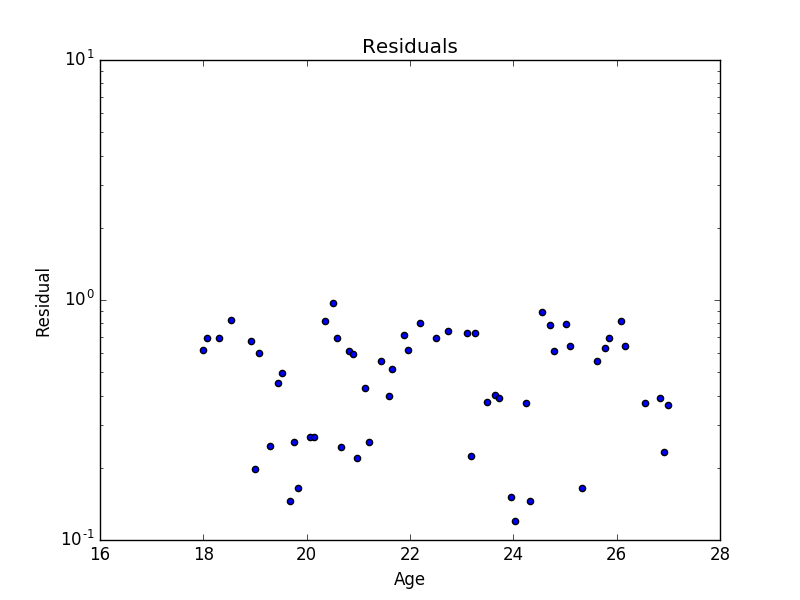
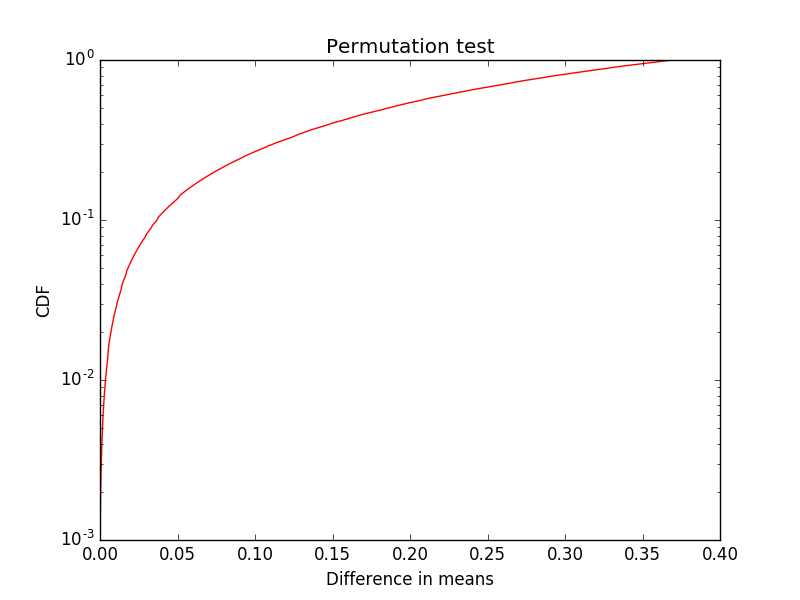
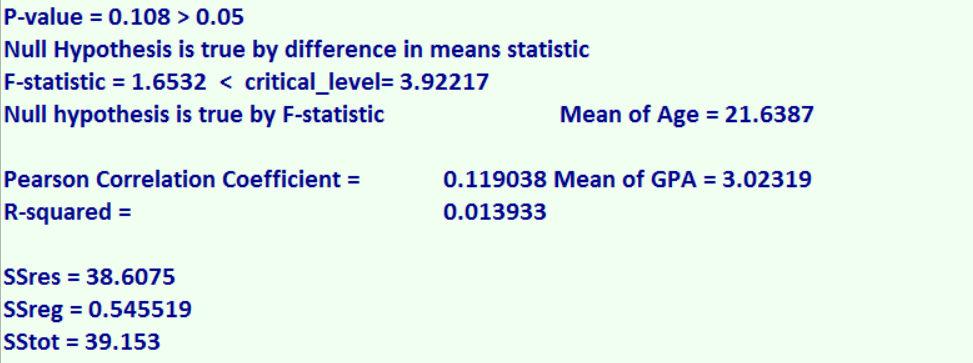
****

**Time Spent on FB vs Gender:**

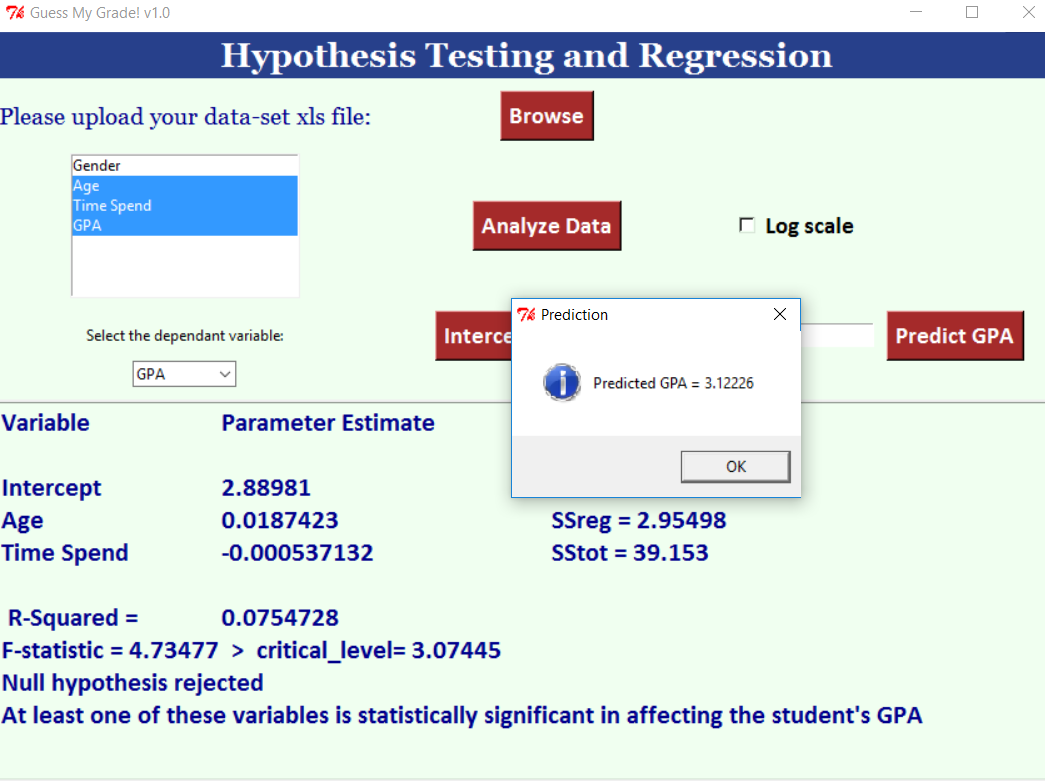
****

****

**Log(GPA) vs Age:**

****

**A student who is 21 years old and spends 300 hours on FB:**

****

**Discussion:**

In part one we can see that none of the apparent effects have enough low p-value to reject the null hypothesis. However, when we use F-statistic instead of difference in means statistic we actually reject the null hypothesis. In order to get the p-value using the differnece in means statistic, I divided the sample into two groups, one with higher values of independent variable (e.g. time spent on facebook) and the other one is with the lower values. By shuffling the two groups together and calculating the difference in means we are assuming the null hupothesis to be true. Repeating the shuffle (running the model) for n times (1000 times used in code) and then calculating the p-value by summing all the difference between the difference in mean values of the model (shuffled groups) and the actual difference in means and then dividing by the number of repitition. The F-Statisic on the other hand is to be calculated by dividing ESS by its degrees of freedom: k= number of independent variables or predictors, (ESS(MS)) or MSE and dividing the RSS by its degrees of freedom: n- (k+1), (RSS(MS)) or MSR and then F-value = MSE/MSR. To get the critical value of F, we can use the F-distribution table with alpha = 0.05 and the degrees of freedom we have. Using F-statistic makes more sense than difference in means as Difference in Means can work with Gender variable, since we have two groups available, however for time spent on facebook for example dividing the sample into two groups might introduce some sampling errors that might affect the pvalue, especially when some p-values are 0.059 which is almost 0.05.

**Weaknesses of the data and the model:**

The data provided might not make sense in the sense that some students have higher GPA while they spend more time on facebook, which doesn’t fit to the reality in a point. That might raise the critics against the sampling process itself.The model doesn’t work efficiently when we go to unrealistic values, e.g. if a student’s age is 200 years he is predicted to spend -1847 hours on FB! That is why models are to be used in specific range of values and errors must always be taken into consideration.