Correlation and Regression Report

Introduction

Around 140 million babies are born every year in the world (The World Count,2021). Pregnant mothers have to visit the hospital several times before giving birth. As some babies are heavier than others, one interesting question to consider was whether there is any relation between weight of babies and visits? Is it possible that babies who are heavier, their mothers have to visit the doctor more often? Moreover, if there is some correlation, can we predict one variable using another? This assignment will serve as the basis to answer all these questions using a regression model.

Dataset

The dataset used for this assignment was featured in Openintro Statistics Textbook (Diez, 2012). It is a sample of births in North Carolina. The data set has 150 births, includes various variables like mothers age, smoking habits, weight gained, visits to the hospital, and gender.

However, this paper focuses only on two variables, namely “Weight” - the weight of baby, which is a quantitative continuous variable. They are numbers with decimals and have indefinite values between two numbers. We are considering it an independent variable, which means it does not get affected when there is change in other variables. And “Visits,” - the number of hospital visits, a quantitative discrete variable as it can only take a specific finite number of values. This is considered a dependent variable that changes as per several factors. As the assignment focuses on correlation and regression, we have considered both our variables as quantitative, which can be counted.[1]

[1] #variables: I identified and distinguished between the two different type of variables related to my research question.

Methods

Summary Statistics

Before diving into the analysis as a preliminary step, we will import and analyze our data, find the summary statistics using pandas in python. Appendix A shows the relevant coding on the calculation of various statistics, and table 1 below indicates only relevant statistics.

Table 1: Summary statistics

	Weights	**Visits**
Count	n_1 =150	n_2 =150
Mean	$\bar{x}_1$= 7.04	$\bar{x}_2$= 11.50
Standard Deviation	s_1= 1.50	s_2=3.63

The next step is to visualize the given data to better understand the relationship between the variables and construct residual plot. Figure 1 shows a linear regression, the line fit with regard to the variables. In addition, figure 2 shows a scatterplot of the residuals where x-axis represents the independent variable, and y-axis shows the residuals. Residual is the difference between predicted and the observed value of the variable -visits. Lastly, figure 3 shows us a histogram of the residuals accompanied by a QQ plot in figure 4 showing normal distribution. Check Appendix B for relevant coding of the graphs.

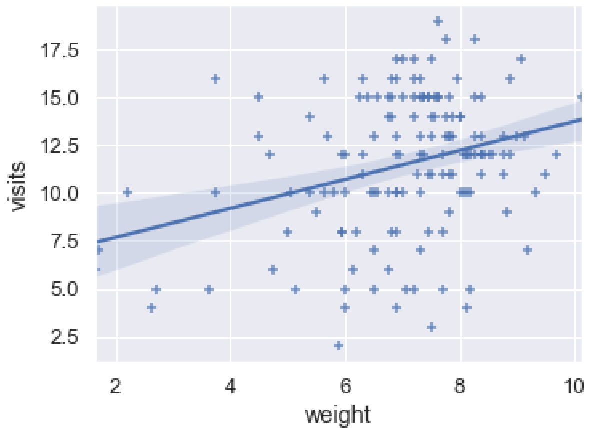


Fig 1. Regression line showing the fit of the variables

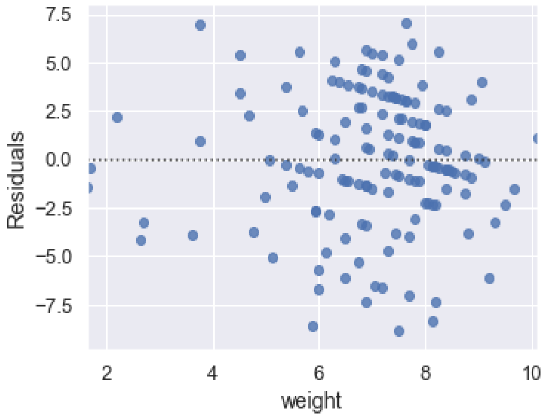


Fig 2. Residual Plot

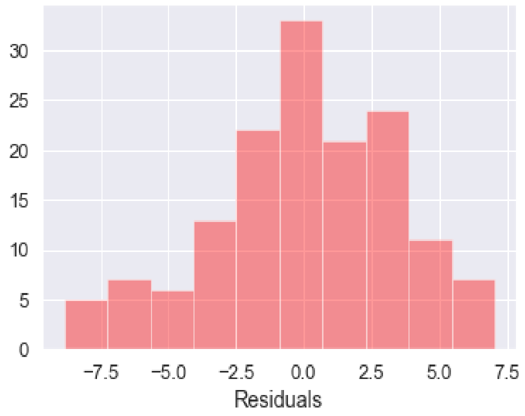


Fig 3. Histogram showing the frequency of residuals

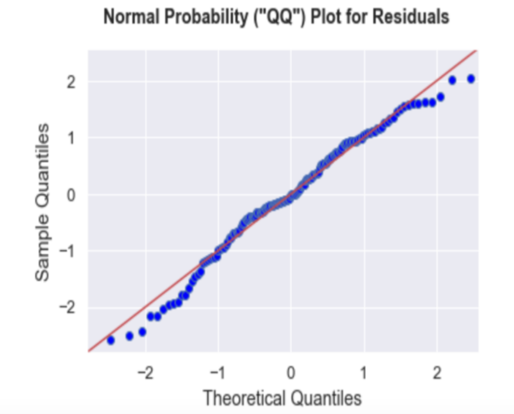


Fig 4. QQ plot showing normality of the residuals[2]

[2] #dataviz: I generated various graphs using python to discuss about the various conditions for inferences

Conditions for inference

Some conditions need to be satisfied before making inferences from our regression model.

1. Linearity

From figure 1, we can see a linear relationship between the two variables, but to consider the linear association in more depth, let us do hypothesis testing

Null hypothesis (H0): There is no linear relationship between the weight of babies and visits to the hospital; $\beta 1 = 0$

Alternate hypothesis (HA): There is some linear relationship between the weight of babies and visits to the hospital; $\beta 1 \neq 0$

To check if the hypothesis is true or false, we will find p-value, which gives the probability that a random chance generated in the data is rare or not. The p-value is calculated using python; check AppendixD for calculations. P-value = 0.001, which is smaller than 0.005. As we know, if pvalue<0.005, then it is statistically significant. This indicates strong evidence against the null hypothesis. There is less than a 5% probability the null is correct. We reject the null hypothesis, which means there is some linear relationship between the given variables.[3]

2. Independence

As we do not have enough information about our data set, we assume the individual observations are independent of each other.

3. Normal

Further, looking at the histogram in figure 3 supported by the QQ plot in figure 4, we clearly see a normal distribution.

[3] #significance: I rejected the null hypothesis on the basis of the p value which was <0.05 and justified one of the conditions of inference.

4. Equal Variability

In figure 2, we see that the scatter plots of the residuals are all around and are not equally distributed. This shows that it is heteroscedastic, and we fail to satisfy the condition of equal variance, i.e., the standard deviation of visits is not the same for all weights. This condition may cause the analysis to be invalid. However, this can be fixed, but as it is out of the scope of the assignment, we will not consider it.

5. Randomness

The last condition is randomness; as the dataset mentions, it is a random sample of 150 births in North Carolina; we also satisfy this condition.

As almost all the conditions are satisfied, we can make inferences from our regression model.

Correlation Coefficient

The first thing to consider is Pearson's correlation coefficient “r”. It is the specific measure that quantifies the strength of the linear relationship between two variables. The formula is $R = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{s_x}\frac{y_i - \bar{y}}{s_y}$. In Appendix D, we directly calculated Pearsons r which is 0.31. As the value is positive, this indicates there is a positive association; with every increase in independent variable weights, there is an increase in dependent variable visits. However, the coefficient is more close to 0 which means that there is not much closeness of association of the points in a scatter plot to a linear regression line, Also in figure 1 the points are not very close to the regression line, which means there is a very weak association between the variables. Correlation tests are useful for a relationship between variables. Seeing two variables moving together does not necessarily mean we know whether one variable causes the other. This is why we would not conclude anything about the causation on the basis of the correlation and avoid the fallacy of Post hoc ergo propter hoc.[4]

[4] #correlation: I discussed about the peasrsons r and how weak the association is. Also I discussed how correlation does not imply causation.

R squared and the regression equation

Further let us take into account R squared, which is the statistical measure of how close the data are to the fitted regression line. Consider the formula of R2 for better interpretation

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

here SSR is the regression sum of squares; SSE is the sum of errors and SST = SSR + SSE. R^2 is the percentage of the actual variation that is due to regression. Refer figure 5 for more clarity.

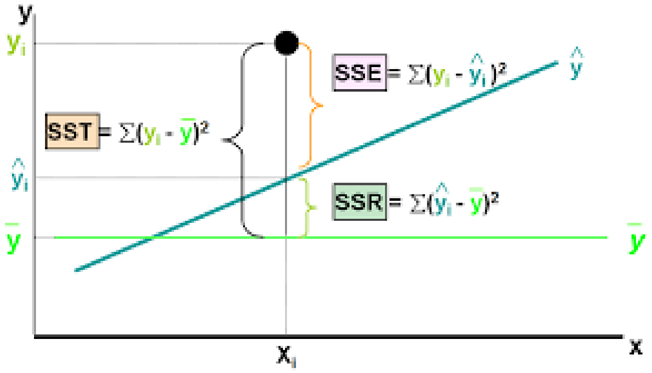

y
y_i
$\hat{y}_i$
$\bar{y}$
SST = $\Sigma(y_i - \bar{y})^2$
SSE = $\Sigma(y_i - \hat{y}_i)^2$
SSR = $\Sigma(\hat{y}_i - \bar{y})^2$
$\hat{y}$
$\bar{y}$
x_i
x

Fig 5. Graphical interpretation of R squared formula

The coefficient R2 is calculated using python in Appendix E which is 0.098, the actual square of Pearson r 0.32 too, it means 9% of the variation in the visits to the hospital is explained by the weight of the babies.

The regression equation is calculated using a code in appendix F.

visits = 0.757 * weight + 6.171 where y-intercept 6.171 tells us how many visits we had expected when the weight is zero, the slope tells us how much our visits determined by the weight. In simple words, on average increase in 1 pound weight is associated with 0.757 visits useful for predicting visits when given weights for the population.[5]

[5] #regression: I discussed the formula of R2 and what does its value convey. Also I interpreted what the regression equation meant.

Results and Conclusion

Regression models are helpful in finding relationships between quantitative variables and predicting the dependent variable like visits in terms of independent variable weight of babies. As discussed above our model has weak correlation (0.31) hence the R2 is also low, 9% variation can only be explained which shows that we cannot rely much on the derived regression equation. However, this model can be useful to compute the relation between various other independent variables like weight gained by mothers or mothers age, adding these variables step by step to see if they have an effect on the dependent variable this is called forward selection. This helps us infer which variable is the most associated.

Our inference was inductive because we took sample sizes and then discussed the population. So we went from specific to more general and the conclusion was not contained within the scope of the premises. This is a type of induction called generalization As we justified all the 5 conditions for inference and used significance test calculated p-value makes our induction strong. The effect size was also large which also convinces that the inference was reliable.[6]

Reflection

Statistics has always been fun, it is making sense out of the data available. As last semester we conducted a difference of means test to get various insights from data using p-value supported by confidence interval. This semester we focused on finding associations between variables and just before talking about regression we used p-value to satisfy the condition for inference that the variables have a linear relationship through hypothesis testing.

Word Count: 1362

[6] #induction: I identified why statistical inference are inductive in nature and discussed about the strength and reliability of the inference.

References

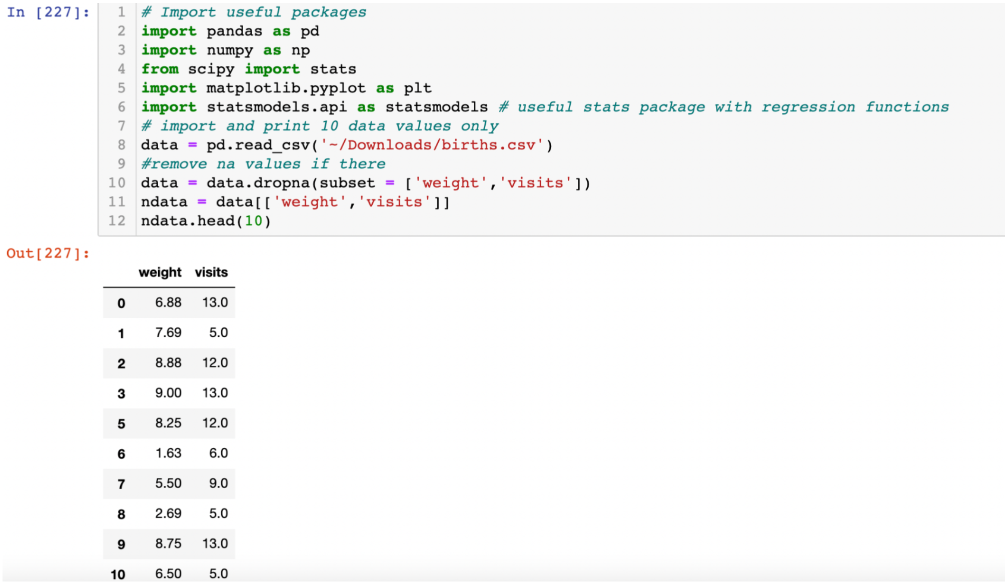
The World Counts. (2021). *Number of births per year.* https://www.theworldcounts.com/populations/world/births

Diez, D. (2012). *OpenIntro Statistics*. https://www.openintro.org/data/index.php?data=births

Appendix

The full Jupyter notebook file and the data can be accessed in the zipped folder submitted as a secondary file.

Appendix A : Importing data and finding summary statistics

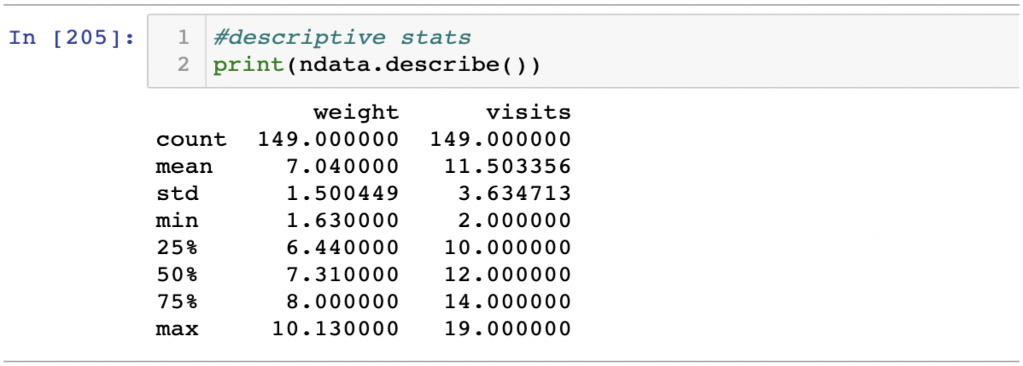


In [227]:

```python
# Import useful packages
import pandas as pd
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt
import statsmodels.api as statsmodels # useful stats package with regression functions
# import and print 10 data values only
data = pd.read_csv('~/Downloads/births.csv')
#remove na values if there
data = data.dropna(subset = ['weight','visits'])
ndata = data[['weight','visits']]
ndata.head(10)
```

Out[227]:

	weight	visits
0	6.88	13.0
1	7.69	5.0
2	8.88	12.0
3	9.00	13.0
5	8.25	12.0
6	1.63	6.0
7	5.50	9.0
8	2.69	5.0
9	8.75	13.0
10	6.50	5.0

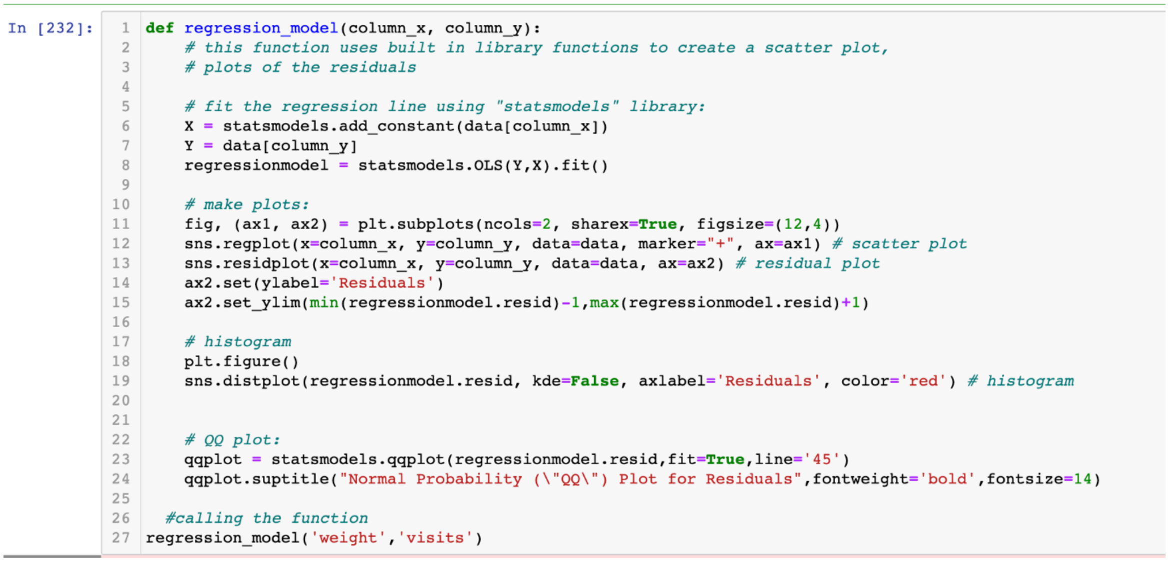


In [205]:

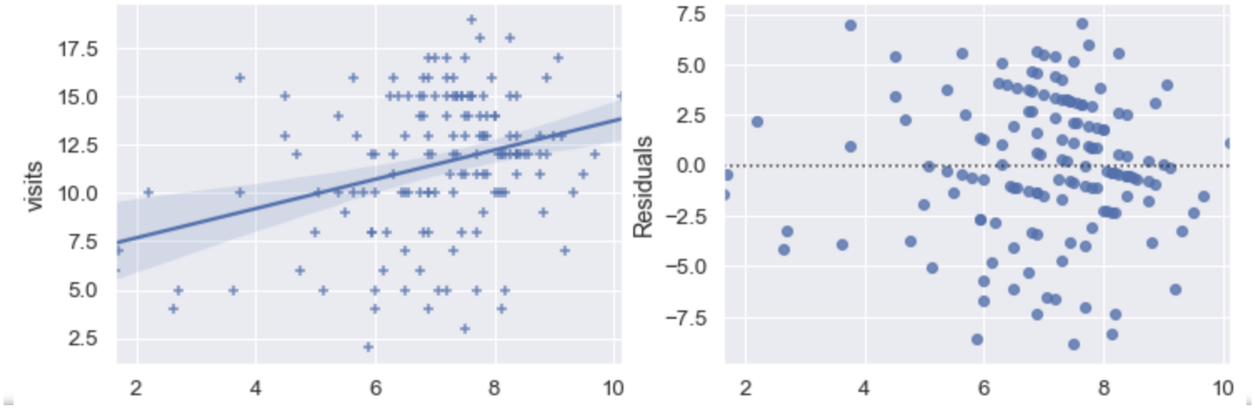
```python
#descriptive stats
print(ndata.describe())
```

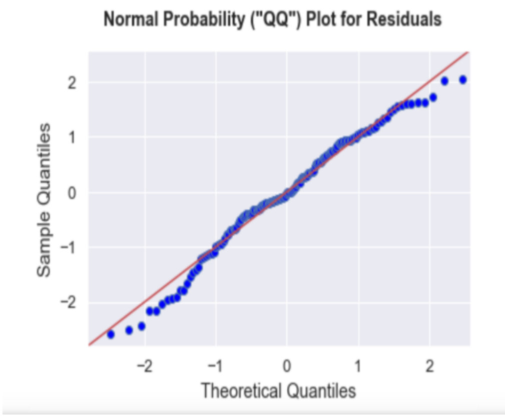
```
           weight      visits
count  149.000000  149.000000
mean     7.040000   11.503356
std      1.500449    3.634713
min      1.630000    2.000000
25%      6.440000   10.000000
50%      7.310000   12.000000
75%      8.000000   14.000000
max     10.130000   19.000000
```

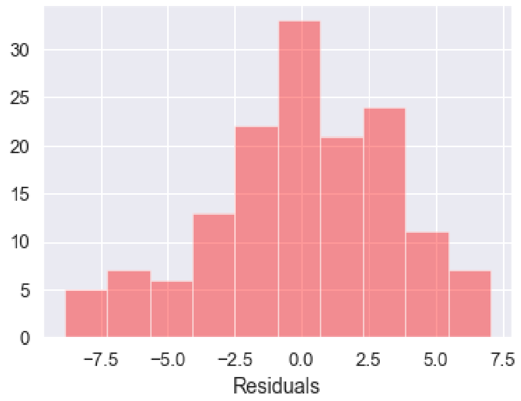
Appendix B: Plotting Graphs



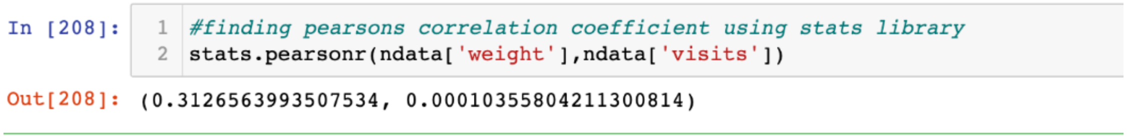
```
In [232]:
def regression_model(column_x, column_y):
    # this function uses built in library functions to create a scatter plot,
    # plots of the residuals

    # fit the regression line using "statsmodels" library:
    X = statsmodels.add_constant(data[column_x])
    Y = data[column_y]
    regressionmodel = statsmodels.OLS(Y,X).fit()

    # make plots:
    fig, (ax1, ax2) = plt.subplots(ncols=2, sharex=True, figsize=(12,4))
    sns.regplot(x=column_x, y=column_y, data=data, marker="+", ax=ax1) # scatter plot
    sns.residplot(x=column_x, y=column_y, data=data, ax=ax2) # residual plot
    ax2.set(ylabel='Residuals')
    ax2.set_ylim(min(regressionmodel.resid)-1,max(regressionmodel.resid)+1)

    # histogram
    plt.figure()
    sns.distplot(regressionmodel.resid, kde=False, axlabel='Residuals', color='red') # histogram


    # QQ plot:
    qqplot = statsmodels.qqplot(regressionmodel.resid,fit=True,line='45')
    qqplot.suptitle("Normal Probability (\"QQ\") Plot for Residuals",fontweight='bold',fontsize=14)

  #calling the function
regression_model('weight','visits')
```





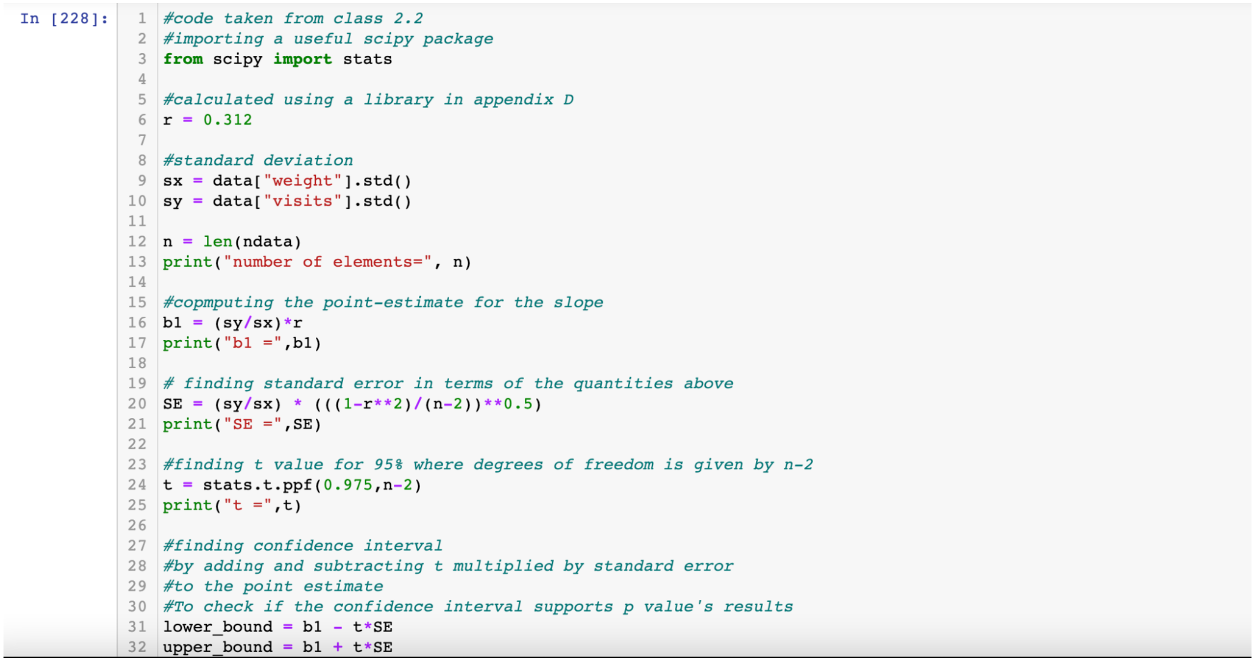


Appendix C: Finding Pearson's R

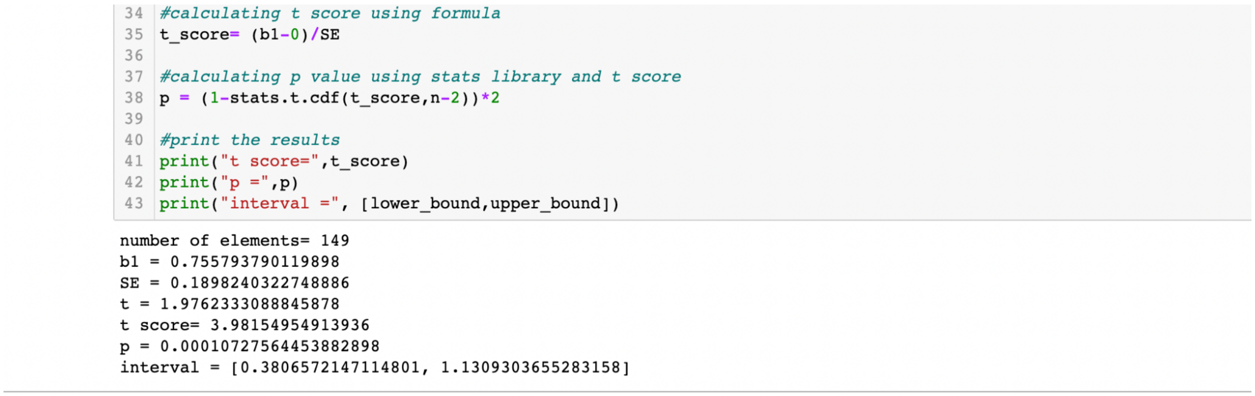


```
In [208]:  #finding pearsons correlation coefficient using stats library
           stats.pearsonr(ndata['weight'],ndata['visits'])

Out[208]: (0.3126563993507534, 0.00010355804211300814)
```

Appendix D: Finding P values and confidence interval



```
In [228]:  #code taken from class 2.2
           #importing a useful scipy package
           from scipy import stats

           #calculated using a library in appendix D
           r = 0.312

           #standard deviation
           sx = data["weight"].std()
           sy = data["visits"].std()

           n = len(ndata)
           print("number of elements=", n)

           #copmputing the point-estimate for the slope
           b1 = (sy/sx)*r
           print("b1 =",b1)

           # finding standard error in terms of the quantities above
           SE = (sy/sx) * (((1-r**2)/(n-2))**0.5)
           print("SE =",SE)

           #finding t value for 95% where degrees of freedom is given by n-2
           t = stats.t.ppf(0.975,n-2)
           print("t =",t)

           #finding confidence interval
           #by adding and subtracting t multiplied by standard error
           #to the point estimate
           #To check if the confidence interval supports p value's results
           lower_bound = b1 - t*SE
           upper_bound = b1 + t*SE
```



```
#calculating t score using formula
t_score= (b1-0)/SE

#calculating p value using stats library and t score
p = (1-stats.t.cdf(t_score,n-2))*2

#print the results
print("t score=",t_score)
print("p =",p)
print("interval =", [lower_bound,upper_bound])
```

```
number of elements= 149
b1 = 0.755793790119898
SE = 0.1898240322748886
t = 1.9762333088845878
t score= 3.98154954913936
p = 0.00010727564453882898
interval = [0.3806572147114801, 1.1309303655283158]
```

Appendix E: R squared and Regression equation

In [233]:

```
#code taken from class 2.2
#defining the function
def regression_model(column_x, column_y):

    # compute R-squared, and display the regression eqn


    # fit the regression line using "statsmodels" library:
    X = statsmodels.add_constant(data[column_x])
    Y = data[column_y]

    # extract regression parameters from model, rounded to 3 decimal places:
    regressionmodel = statsmodels.OLS(Y,X).fit() #using Ordinary least squares
    Rsquared = round(regressionmodel.rsquared,3)
    slope = round(regressionmodel.params[1],3)
    intercept = round(regressionmodel.params[0],3)

    # printing the results:
    print("R-squared = ",Rsquared)
    print("Regression equation: "+column_y+" = ",slope,"* "+column_x+" + ",intercept)

    #calling the function
regression_model('weight','visits')
```

```
R-squared =  0.098
Regression equation: visits =  0.757 * weight +  6.171
```