Correlation and Regression

Minerva University

CS51: Formal Analyses

Prof. Volkan

January 30, 2022.

Introduction

Infants spend on average 38 weeks in gestation to be born as a healthy baby (Jukic et al., 2013). This biological mechanism was studied for our dataset in North Carolina with 150 mothers. This study examines a linear relationship between weeks spent in gestation and a baby’s birth weight by using correlation and regression analyses. Our research seeks inference from the infant mortality rate of 6.8 deaths per 1000 births in North Carolina, which is a higher mortality rate than the average rate of 5.6 deaths per 1000 births in the U.S (Elflein, 2021), (Centers of Disease Control and Prevention, 2021). This will help us infer whether gestation weeks can affect infant mortality as babies with low birth weight have a higher probability of dying in states of the U.S. that have higher than average infant mortality rates. We will further use Pearson’s R, p-value and a linear regression model to analyze the variables using a slope and R^2.

Dataset

North Carolina Births data was taken from 150 couples between the ages of 15 and 46 from an online database (OpenIntro, n.d.). The first 10 rows of the dataset are shown **(Appendix A)**. For this analysis, we will be focusing on two variables for single regression analysis: **weeks** and **weight** in lbs. Our research question is whether a linear relationship exists between gestation weeks and a baby’s birth weight. This will help us find implications of time spent in a mother’s womb as a factor for a healthy baby with an ideal birth weight.

Weeks is the total time spent in gestation (time of pregnancy). This is being used as a quantitative discreet variable in the study as the dataset number of weeks was rounded off to nearest weeks. Usually, gestation periods are weeks, and some days, even hours, can be included to mark when the baby is born, but to make the analysis simpler we are taking only the weeks rounded off to act as a discreet variable. Here we take it as a predictor variable as we predict that it influences how healthy the baby is born. The second variable is **weight** in pounds which is the infant's weight at birth**.** This acts as a response variable as it is a proxy variable to check the baby's health and is dependent on pre-birth factors like gestation period (Vilanova et al., 2019). Weight is being considered as a quantitative continuous variable as the weight of the infant can vary between 2 decimal places, and as it is an uncountable set of values between two set points, it is continuous.[1]

Method

Hypotheses:

Null and Alternative hypotheses are formulated:

$$H_0: \beta 1 \ = \ 0$$

There is no linear relationship between weeks spent in gestation period and birth weight of the infants.

$$H_A: \beta 1 \,!= \ 0$$

There is a linear relationship between weeks spent in the gestation period and the birth weight of the infants.

[1] **#variables:** I have analyzed the variables, classified them into respective categories. I have explained and justifed why weeks will be considered discreet instead of continuous. I have mentioned use of weight as a proxy variable for infant's health.

The dataset was evaluated using Python: panda and stats packages **(Appendix A & B)**. Summary statistics are represented **(Fig. 1)**.

	Weeks	**Weight(pounds)**
Mean	38.5	7.0
Standard Deviation	2.75	1.50
Range	26.0	18.0

Fig. 1 shows summary statistics of High (HAG) and Low (LAG) anxiety levels in form of mean, standard deviation, median, mode, and range.

Assumptions and Conditions:

To analyze the correlation and regression of the above-mentioned variables we need to meet some basic assumptions and conditions based on **LINER** guidelines and plot computation shown in **Appendix E** and (**Fig. 2).**

- **Linearity:** There is a linear relationship between x & y (**Blue** plot).
- **Independent Observation:** The sample size is less than 10% of the population size.
- **Normal Distribution:** The residuals are normally distributed (**Grey** plot).
- **Equal Variance:** There is a strong sign of homoscedasticity (this means that the variance of the dependent variable should be the same for all the data) as points have approximately equal spread and standard deviation with some outliers (**Red** plot).

- **Randomness:** The data comes from a random sample which is why there is a difference in the number of smokers vs non-smokers.[2]

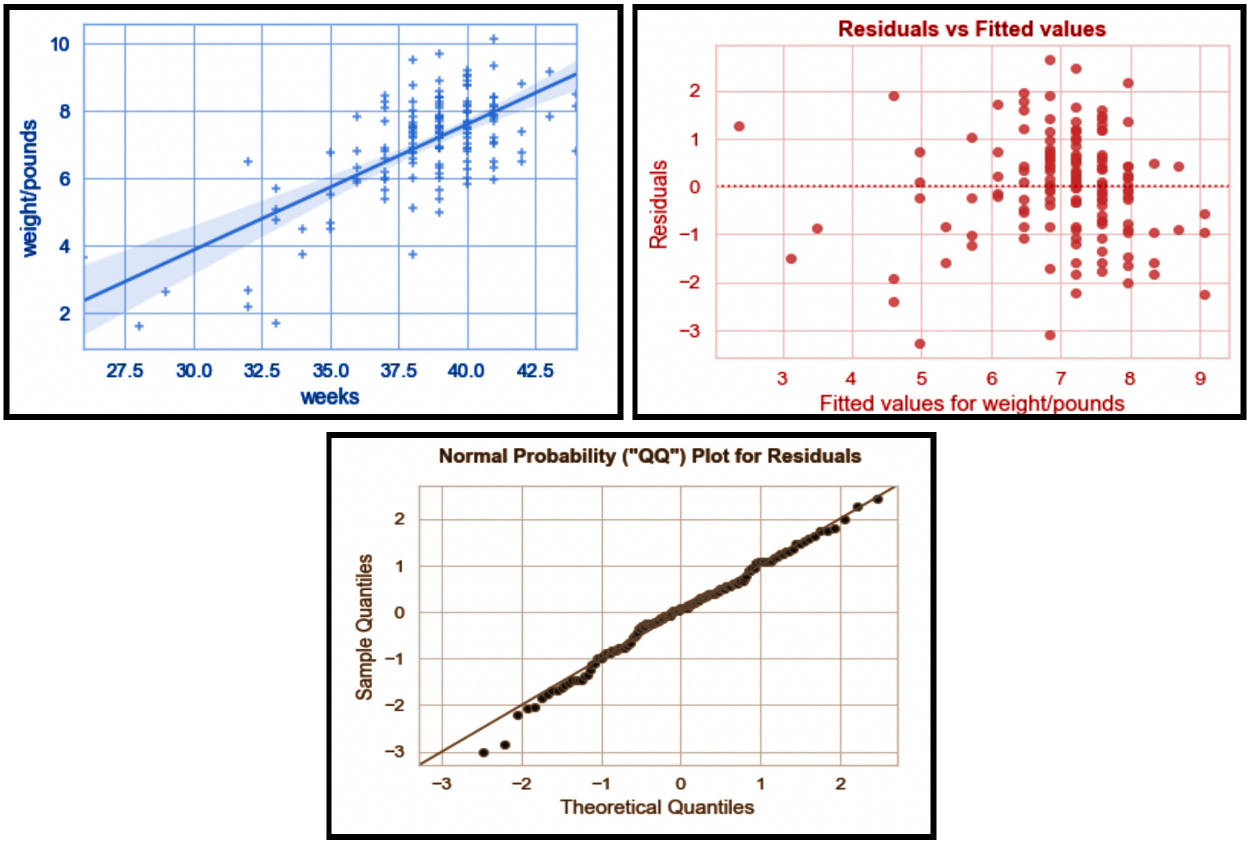


Fig. 2 shows the three plots Scatter plot with best-fit line **(Blue)**, Homoscedasticity of residuals **(Red),** Distribution of residuals **(Grey)**.

There are some outliers that reduce the predictive power of our model but we will assume the conditions can pass the standards for using correlation and regression analyses.

[2] **#dataviz:** I have plotted the data effectively using appropriate means of plotting. All necessary components are included e.g. captions, x and y labels with units. I have elaborated how the plots can be used for justifying that the assumptions/conditions are met for regression and I have used seaborn to make a better visualization.

Correlation:

We have used Pearson's R to determine the strength of bivariate correlations and determine if the two variables are linearly related. Its value can lie between -1 and 1, with the extremes showing a strong correlation, and closer to 0 indicating minimal correlation. Given that the conditions are met we can either use the formulae:

$$r = \frac{\Sigma\,(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2\ \Sigma(y_i - \bar{y})^2}}$$

where n presents the number of data points, x,y the sample means and s_x,s_y sample standard deviations or use a scipy library from pandas to find the value using a tool called pearsonr as shown in **Appendix C.**

The computation of Pearson's R shows a correlation value of ~ 0.68, demonstrating a moderately strong positive correlation present between the baby's weight and their delivery week. The sign of the correlation shows a positive or negative correlation and the magnitude of the value indicates the strength of correlation. Given that R^2 minimizes the residuals and confirms linear relationship using a single regression model showing the prediction of variability, we will use it. However, we must not assume that there is any causation even though there is a strong correlation and that could have been caused by an extraneous variable like m_age as high age can cause other complications during pregnancy (Chung et al., 2020).[3]

[3] **#correlation:** I have computed Pearson's correlation coefficient with detailed explanation of use and need. I have addressed why we cannot assume causation from this and have indicated an extraneous variable that can play a role to prevent causation.

Linear Regression:

To show a linear relationship between our bivariate data (two variables are involved), we will use the Single Linear Regression model. It assumes that the relationship between two variables, x and y, can be modeled by a straight line.

$$y = \beta 0 \ + \ \beta 1x$$

Here $\beta 0$ and $\beta 1$ are model parameters but for calculation purposes, we will take their point estimates as b_0 and b_1. The b_0 is the y-intercept of the regression line and b_1 is the straight-line slope. We use the following formula to calculate b_1:

$$b_1 \ = \ \frac{s_y}{s_x} r$$

Where s_y is the standard deviation of y and s_x is the standard deviation for x. R is the coefficient of correlation. The regression equation shows b_1.

For our variable, the regression equation as values shown in **Appendix E** is

$$Weight/lbs \ = \ 0.37\, Weeks \ - \ 7.3$$

The coefficient of weeks(x) tells us that weight(y) is increased by 0.37 lbs when weeks for complete gestation is increased by 1 week. The y-intercept shows that the weight of the baby will be -7.3 if the baby is not born when weeks are 0. Although it is impractical, we calculate it to maintain the mathematical model's integrity.

For our regression, we will use R^2 which describes the amount of variation in the response variable that can be explained by the predictor variable from the least-squares line. It can be calculated either squaring R or using:

$$R^2 = 1 - \frac{SSE}{SSTO}$$

SSE is the "error sum of squares" and quantifies that the data points vary around the estimated regression line and **SSTO** is the "total sum of squares" and quantifies the data points vary around their mean. For ease, we used computation to find R^2 shown in **Appendix E.** In our case R^2 is 0.467 which means that it can predict 46.7% of the variance in weight of the infant by the weeks of gestation. The stronger the R^2, more accurate our prediction and in our case, it is moderately low as it falls between 0.3 and 0.5 (Alhyari, 2016).[4]

Significance test:

In order to assess the acceptance of our null hypothesis, we will do a significance test to find the p-value for our slope (b_1). Given that we used a two-tailed test is examined with α = 0.05 is set as a significance level to assess the null hypothesis taking into account the difference between a sample and population size. By using scientific field's conventions, we can assume a threshold of 5% risk of committing a Type I error (rejecting the null hypothesis when it is true) to conclude that the probability of getting a sample mean more extreme than the observed x̄ (sample mean) is not likely by random chance. As shown in **Appendix D** and rechecked using **Appendix E.** For ur t-value calculation we first need Standard Error as we do not have the population standard deviation:

$$SE = \frac{s_y}{s_x} \times \sqrt{\frac{(1-r)^2}{n-2}}$$

[4] **#regression:** I have computed parameters of the linear regression equation (b1 and b0) and justified a relationship between the dependent and independent variable. I have computed and analyzed R-squared with relevant inferences and need for our study. I have further used a regression model to provide significant insights into two important variables.

Here SE is the stand error, s_y and s_x are the standard deviations of respective variables, r is the coefficient of correlation and n-2 is the degrees of freedom. The t-value is hence calculated using:

$$t = \frac{b_1 - 0}{SE}$$

Where b_1 is the slope, 0 is the null value and SE is the standard error (SE = 0.033).

The t-value is 11. 4 which is quite large and our p-value is approximately 0.0 (computed using a python) which provides strong evidence to reject the null hypothesis. Given that,

$$p - value < 0.05$$

We will reject the null hypothesis and conclude that a significant difference does exist.[5]

Results and Conclusion

The study analyzed variables: infant weight at birth/ lbs and the weeks spent in gestation. Observations made are: There is a high positive correlation r = 0.68 between the two variables; the slope of the linear regression equation is 0.37; the R^2= 0.467 is moderately low; the p-value is less than alpha hence there is strong evidence to reject H_0.

As proved that there is a strong linear relationship even as much as a moderately strong positive correlation which can be predicted with an acceptable R^2 we can make a **strong** inductive argument about the population. The study can provide evidence to carry research to find causation about the high infant mortality rate of 6.8 per 1000 deaths to know if an infant spends less time in gestation will it have less weight at birth and lower chances of survival.

[5] **#signficance:** I effectively calculated statistical significance tests and interpreted the results with relevant interpretations. I have mentioned why we are selecting 0.05 as the alpha value and justified the use as well as explained the reason for rejecting the null hypothesis. Lastly, I have explained the importance of p-value for the significance test.

Additionally, we can make inferences about U.S. states with higher infant mortality rates, like Mississippi, to use gestation weeks as a prenatal test and do modeling based on the baby's size to estimate the weight. This has been done using ultrasound which proves that all the premises: baby size and gestational age, can be used (Azagidi et al., 2019). Our argument is strong as it follows from the statistical tests and is reliable because their assumptions about regression are true. The model is limited as it has a small sample size from a single location and can only predict 46.7% of the variance.[6]

WORD COUNT: 1396 words excluding tables, figure captions and equations.[7][8]

Reflection

I believe I learned quite a lot in terms of regression, correlation and significance. My goal is to be a data scientist and learning correlation and regression analyses with hypothesis testing provided a practical understanding of how research is dependent on statistical inference. Especially how correlation/ linear regression were complemented by significance testing where the hypothesis rejection needs to be confirmed via p-value.

Word Count: 63 words

[6] **#induction:** I have stated an inductive argument using the observations made with type identified and justified. I have further discussed the strength and reliability of the argument and the methods to improve them by showing the premises are true in real world as ultrasound has been used to test the variables, thus showing a strong inductive argument. I have also mentioned the limitations for the strength and reliability.

[7] **#professionalism:** I have maintained the world count rule, avoided grammatical and spelling errors, plotted the graphs using seaborn to have a better visualization. I have used citations to cite my work and took precautions to avoid plagiarism and biases by applying “considering the opposite” method. I have used all the guidelines from APA format, tables and equations to better communicate the information.

[8] **#organization:** I have made subsections of Methods to improve clarity and used introduction to summarize the report to improve understanding and relevance. I have referenced the appendix and created subsections with relevant code. I could have made it better by creating further subsections for Linear Regression and Dateset headings but I contemplated that it might scatter the content so avoided that.

References

Alhyari, S. (2016, January 18). What is the acceptable R-squared value? - researchgate. Research Gate. Retrieved January 30, 2022, from https://www.researchgate.net/post/what_is_the_acceptable_r-squared_value

Azagidi, A. S., Ibitoye, B. O., Makinde, O. N., Idowu, B. M., &amp; Aderibigbe, A. S. (2019, October 9). Fetal gestational age determination using ultrasound placental thickness. Journal of medical ultrasound. Retrieved January 30, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7194423/

Centers of Disease Control and Prevention. (2021, September 8). Infant mortality. Centers for Disease Control and Prevention. Retrieved January 30, 2022, from https://www.cdc.gov/reproductivehealth/maternalinfanthealth/infantmortality.htm#:~:text=About%20Infant%20Mortality,-Infant%20mortality%20is&amp;text=In%202019%2C%20the%20infant%20mortality,the%20United%20States%2C%202019).

Chung, C., Wang, S., & Yang, L. (2020, November 30). Changing trends of birth weight with maternal age: a cross-sectional study in Xi'an city of Northwestern China - BMC Pregnancy and Childbirth. BMC Pregnancy and Childbirth. Retrieved January 31, 2022, from https://bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/s12884-020-03445-2

Elflein, J. (2021, April 29). Infant mortality rates in U.S. by State. Statista. Retrieved January 30, 2022, from https://www.statista.com/statistics/252064/us-infant-mortality-rate-by-ethnicity-2011/

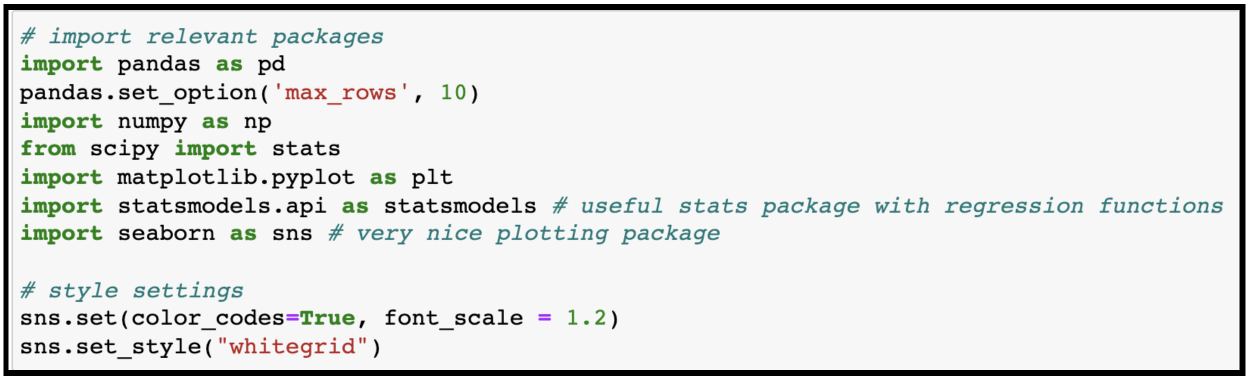
Jukic, A. M., Baird, D. D., Weinberg, C. R., McConnaughey, D. R., &amp; Wilcox, A. J. (2013, October). Length of human pregnancy and contributors to its natural variation. Human reproduction (Oxford, England). Retrieved January 30, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3777570/

OpenIntro. (n.d.). North Carolina Births, 100 cases. Data Sets. Retrieved January 30, 2022, from https://www.openintro.org/data/index.php?data=births

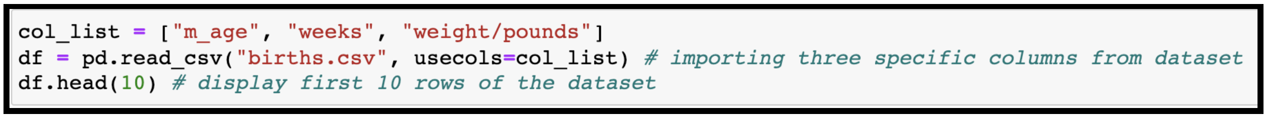
Vilanova, C. S., Hirakata, V. N., de Souza Buriol, V. C., Nunes, M., Goldani, M. Z., &amp; da Silva, C. H. (2019, November 27). The relationship between the different low birth weight strata of newborns with infant mortality and the influence of the main health determinants in the extreme south of brazil - population health metrics. BioMed Central. Retrieved January 30, 2022, from https://pophealthmetrics.biomedcentral.com/articles/10.1186/s12963-019-0195-7#:~:text=Regarding%20the%20risk%20for%20infant,the%20first%20year%20of%20life.

Appendix

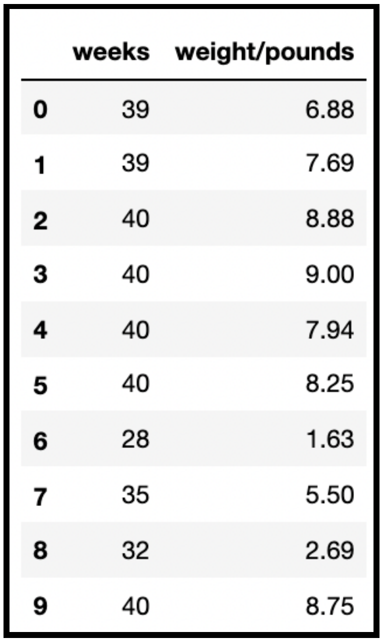
Appendix A: Import, Analyze, and Visualize Data



```
# import relevant packages
import pandas as pd
pandas.set_option('max_rows', 10)
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 seaborn as sns # very nice plotting package

# style settings
sns.set(color_codes=True, font_scale = 1.2)
sns.set_style("whitegrid")
```

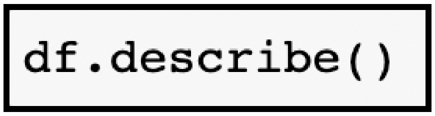


```
col_list = ["m_age", "weeks", "weight/pounds"]
df = pd.read_csv("births.csv", usecols=col_list) # importing three specific columns from dataset
df.head(10) # display first 10 rows of the dataset
```

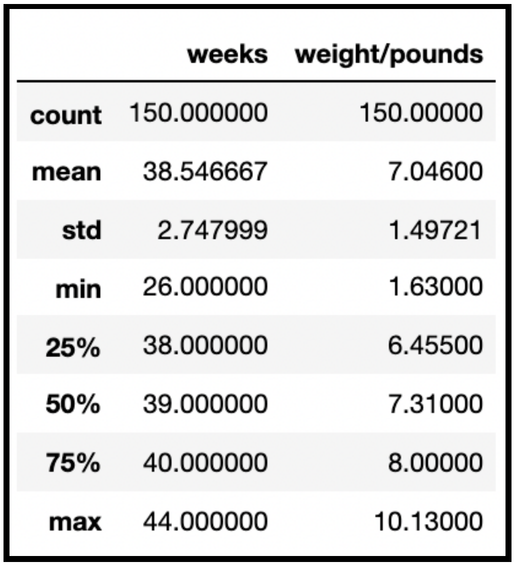


	weeks	weight/pounds
0	39	6.88
1	39	7.69
2	40	8.88
3	40	9.00
4	40	7.94
5	40	8.25
6	28	1.63
7	35	5.50
8	32	2.69
9	40	8.75

Appendix B: Descriptive Statistics

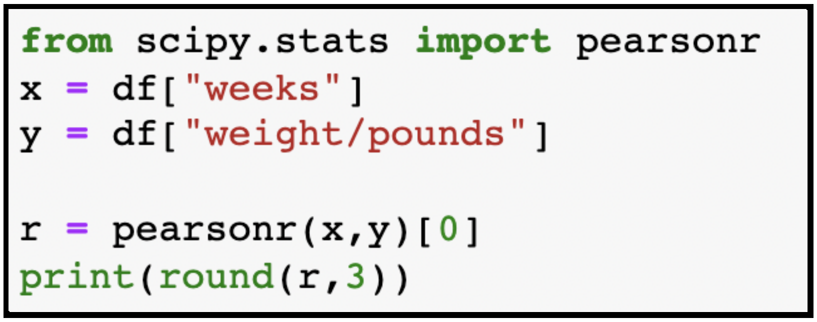


```
df.describe()
```

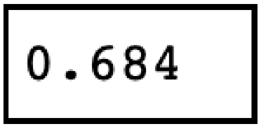


	weeks	weight/pounds
count	150.000000	150.00000
mean	38.546667	7.04600
std	2.747999	1.49721
min	26.000000	1.63000
25%	38.000000	6.45500
50%	39.000000	7.31000
75%	40.000000	8.00000
max	44.000000	10.13000

Appendix C: Pearson's Correlation (R)

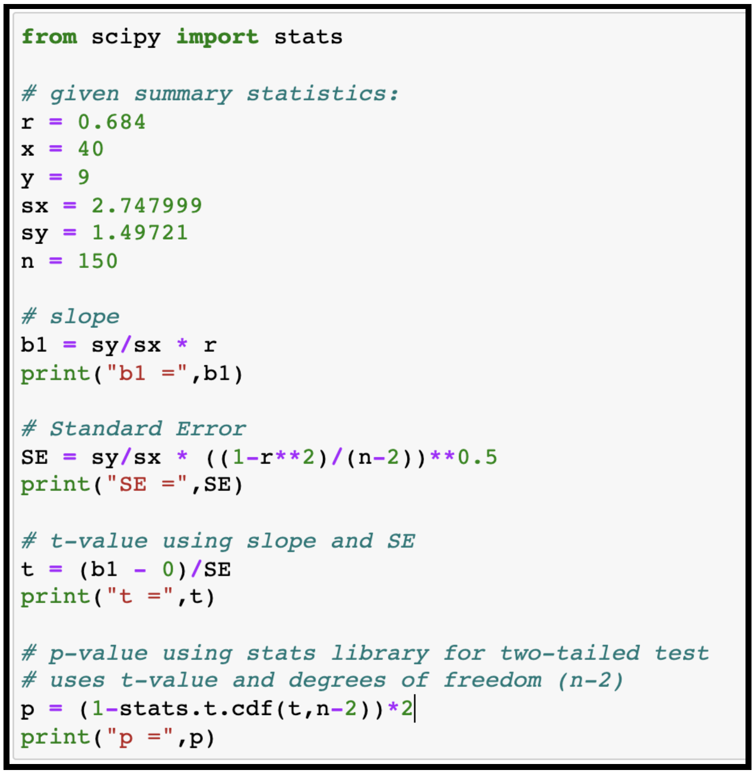


```
from scipy.stats import pearsonr
x = df["weeks"]
y = df["weight/pounds"]

r = pearsonr(x,y)[0]
print(round(r,3))
```

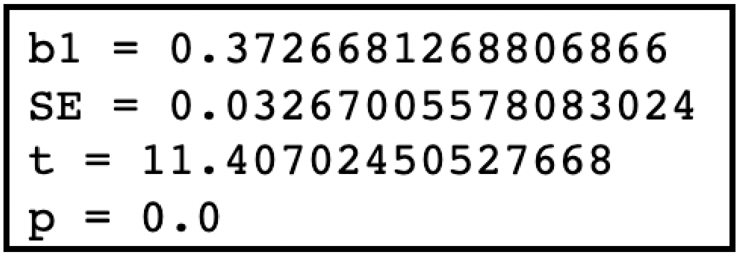


```
0.684
```

Appendix D: Significance Test (p-value)

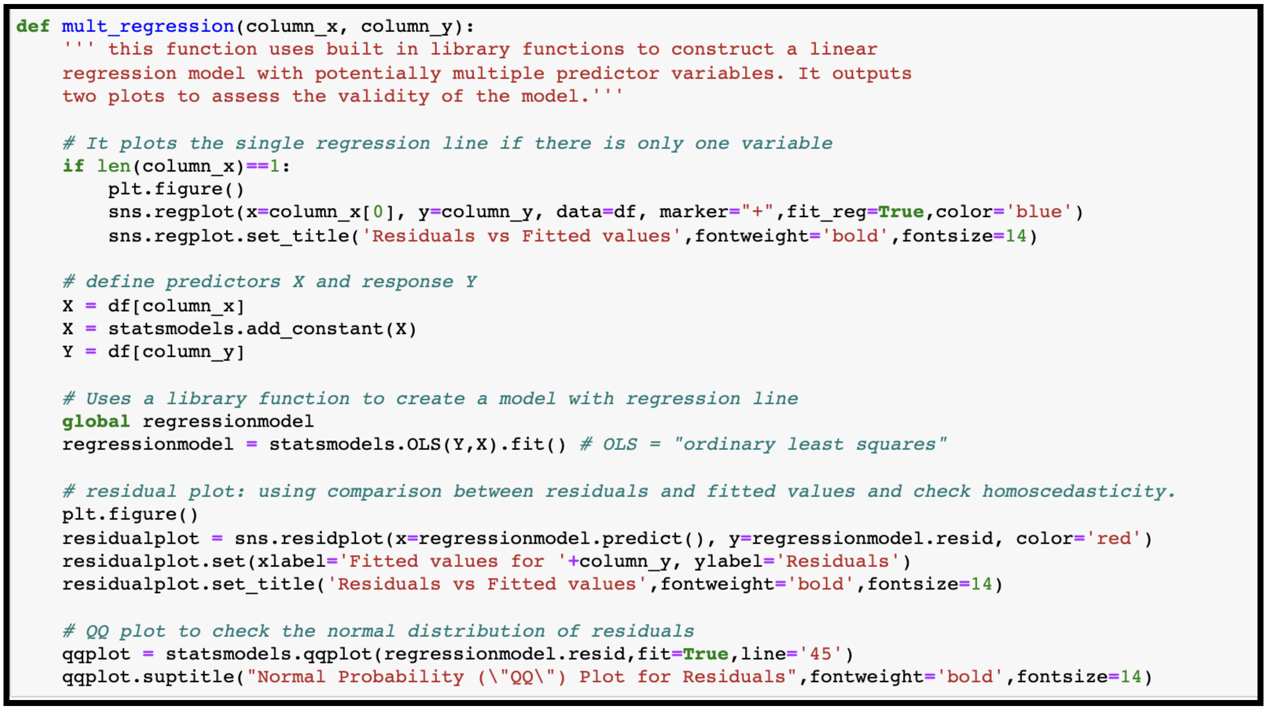


```
from scipy import stats

# given summary statistics:
r = 0.684
x = 40
y = 9
sx = 2.747999
sy = 1.49721
n = 150

# slope
b1 = sy/sx * r
print("b1 =",b1)

# Standard Error
SE = sy/sx * ((1-r**2)/(n-2))**0.5
print("SE =",SE)

# t-value using slope and SE
t = (b1 - 0)/SE
print("t =",t)

# p-value using stats library for two-tailed test
# uses t-value and degrees of freedom (n-2)
p = (1-stats.t.cdf(t,n-2))*2
print("p =",p)
```



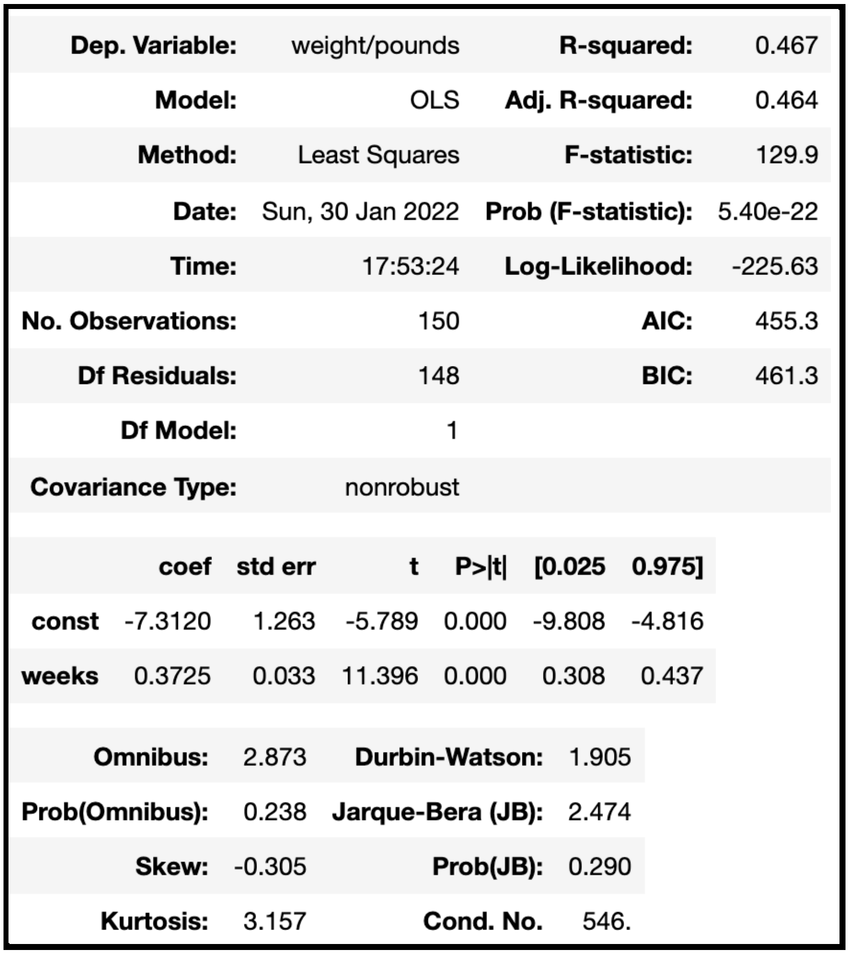
```
b1 = 0.3726681268806866
SE = 0.03267005578083024
t = 11.40702450527668
p = 0.0
```

Appendix E: Linear Regression (R^2)



```
def mult_regression(column_x, column_y):
    ''' this function uses built in library functions to construct a linear
    regression model with potentially multiple predictor variables. It outputs
    two plots to assess the validity of the model.'''

    # It plots the single regression line if there is only one variable
    if len(column_x)==1:
        plt.figure()
        sns.regplot(x=column_x[0], y=column_y, data=df, marker="+",fit_reg=True,color='blue')
        sns.regplot.set_title('Residuals vs Fitted values',fontweight='bold',fontsize=14)

    # define predictors X and response Y
    X = df[column_x]
    X = statsmodels.add_constant(X)
    Y = df[column_y]

    # Uses a library function to create a model with regression line
    global regressionmodel
    regressionmodel = statsmodels.OLS(Y,X).fit() # OLS = "ordinary least squares"

    # residual plot: using comparison between residuals and fitted values and check homoscedasticity.
    plt.figure()
    residualplot = sns.residplot(x=regressionmodel.predict(), y=regressionmodel.resid, color='red')
    residualplot.set(xlabel='Fitted values for '+column_y, ylabel='Residuals')
    residualplot.set_title('Residuals vs Fitted values',fontweight='bold',fontsize=14)

    # QQ plot to check the normal distribution of residuals
    qqplot = statsmodels.qqplot(regressionmodel.resid,fit=True,line='45')
    qqplot.suptitle("Normal Probability (\"QQ\") Plot for Residuals",fontweight='bold',fontsize=14)
```

```
mult_regression(['weeks'],'weight/pounds')
regressionmodel.summary()
```



Dep. Variable:	weight/pounds	**R-squared:**	0.467
Model:	OLS	**Adj. R-squared:**	0.464
Method:	Least Squares	**F-statistic:**	129.9
Date:	Sun, 30 Jan 2022	**Prob (F-statistic):**	5.40e-22
Time:	17:53:24	**Log-Likelihood:**	-225.63
No. Observations:	150	**AIC:**	455.3
Df Residuals:	148	**BIC:**	461.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-7.3120	1.263	-5.789	0.000	-9.808	-4.816
weeks	0.3725	0.033	11.396	0.000	0.308	0.437

Omnibus:	2.873	**Durbin-Watson:**	1.905
Prob(Omnibus):	0.238	**Jarque-Bera (JB):**	2.474
Skew:	-0.305	**Prob(JB):**	0.290
Kurtosis:	3.157	**Cond. No.**	546.