Correlation and Regression

Minerva University

CS51: Formal Analysis

Prof. A. Terrana

January 27, 2022

Correlation and Regression

Introduction

During 2020, the rate of childhood obesity increased from 19.3% to 22.4% in the US (Helping All Children Grow Up Healthy, 2020). Since physical activity is one of the main ways to decrease body weight, this report aims to discover which factors influence the frequency of physical activity most. The "Youth Risk Behavior Surveillance System" dataset was used as sample data. The potential independent variables were found, conditions for correlation were checked, and the independent variable with the strongest correlation coefficient was chosen for creating a regression model line. The confidence interval was calculated to estimate the true slope of the regression line for the population. The sample analysis results were interpreted as an estimation for the population parameter.

Dataset

The source of the dataset is the Center for Disease Control and Prevention. The data was collected from 1991 to 2019 across 2,100 independent surveys of 9th - 12th-grade students in the US (Center for Disease Control and Prevention, 2019). The research question is " What is a good predictor of children's physical activity frequency? " The research question will be investigated by finding the correlation between the dependent and independent variable. The dependent variable is the number of days per week children reported to be active since it will be tested in the study. It is a quantitative discrete variable since it is measured in specific numeric values (days). The relationships between a dependent variable and potential independent variables using scatterplots and correlation coefficients were checked as a preparation step before the study. The

variable with the highest correlation with the dependent variable was chosen as an independent variable for analysis. The independent variable is the number of days children reported doing strength exercises since the study will focus on analysing the manipulation with this variable to acssess changes in dependent varibale. It is also a quantitative discrete variable measured in days. Quantitative data type enables the use of regressions which would be impossible with categorical data (qualitative nominal variables)[1]. Data cleaning was performed before the analysis to handle missing values (Appendix A).

Data Analysis

The data were analyzed with the Python built-in libraries: Pandas, Matplotlib, Seaborn, Statistics, Statsmodels, SciPy. Table 1 summarises the descriptive stats for the variables(Detailed calculation in Appendix B).

Table 1: Summary statistics for the number of days with strength exercises and with physical activity(walking, running, etc)

	A number of days with strength exercises	**A number of days with physical activity(walking, running, etc)**
Count	$n_x = 100$	$n_y = 100$

[1] #**variables**:I explained the choice of variables illustrating how they will be used to answer the research question. I specified and explained the type of variable (quantitative discrete), units of varibales and explained why it is important to have this variable type for the study. I explained the relationships between the variables by identifying the independent, dependent why they have this roles in the study and how it will be used further. I also mentioned possible extraneous variables and how it may affect dependent variable in conclusion of the paper.

Mean	$\bar{x}_1 = 2.788$	$\bar{x}_2 = 3.75$
Mode	0	7
Median	2.5	4
Standard Deviation	$s_1 = 2.58$	$s_2 = 2.56$

The distributions of the variables are demonstrated in figure 1 and figure 2. Adolescents tend to have 60+ minutes of physical activities more than do strength exercises in general.

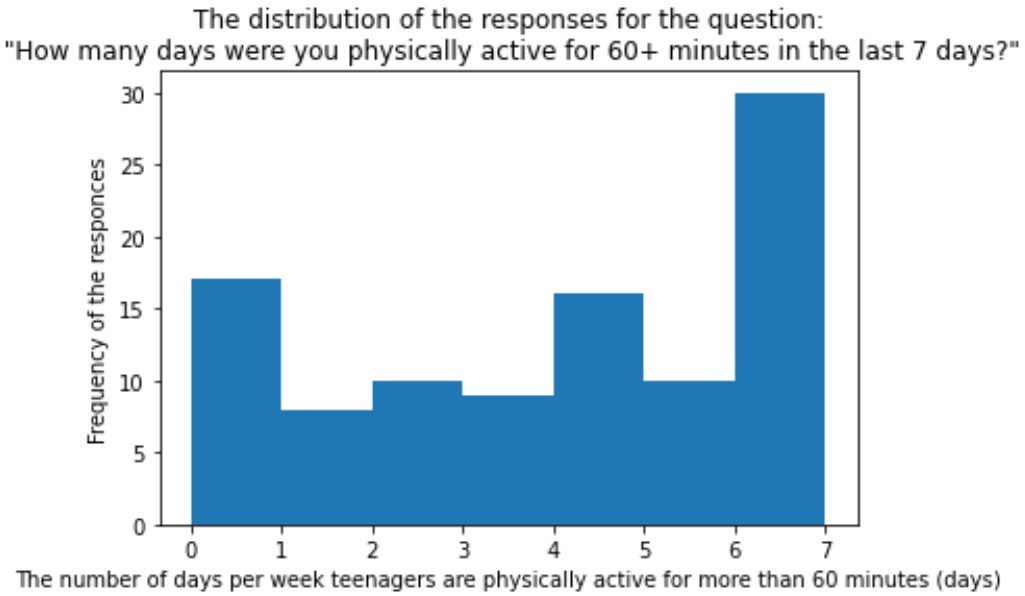


Figure 1: the figure illustrates the distribution of the number of days adolescents reported being physically active for more than 60 minutes. The X-axes represent the number of days values. The Y-axes represent the number of respondents that chose the corresponding number of days. The distribution is skewed left with most responses "7 days".

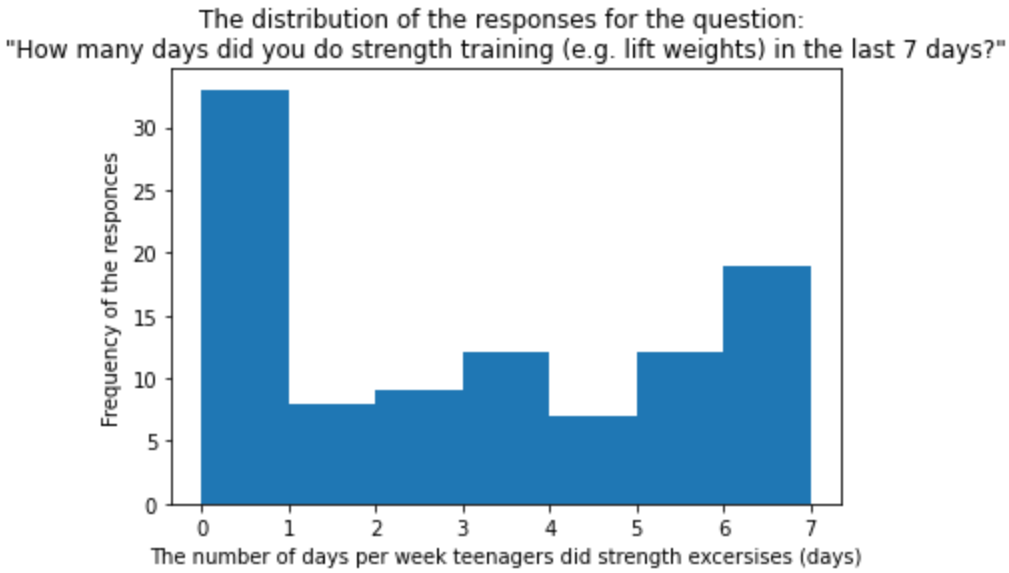


Figure 2: the figure illustrates the distribution of the number of days adolescents reported doing strength exercises. The X-axes represent the number of days values. The Y-axes represent the number of respondents that chose the corresponding number of days. The distribution is skewed right with most of the responses "0 days".

Assumptions check

Correlation coefficient and regression line will produce effective results only when these conditions are met (calculations in Appendix C):

L - linearity. There is a positive linear trend in the data, so as the number of strength training increases, the number of physical activity increases (figure 3). Therefore, the assumption about the linearity of the data can be considered plausible.

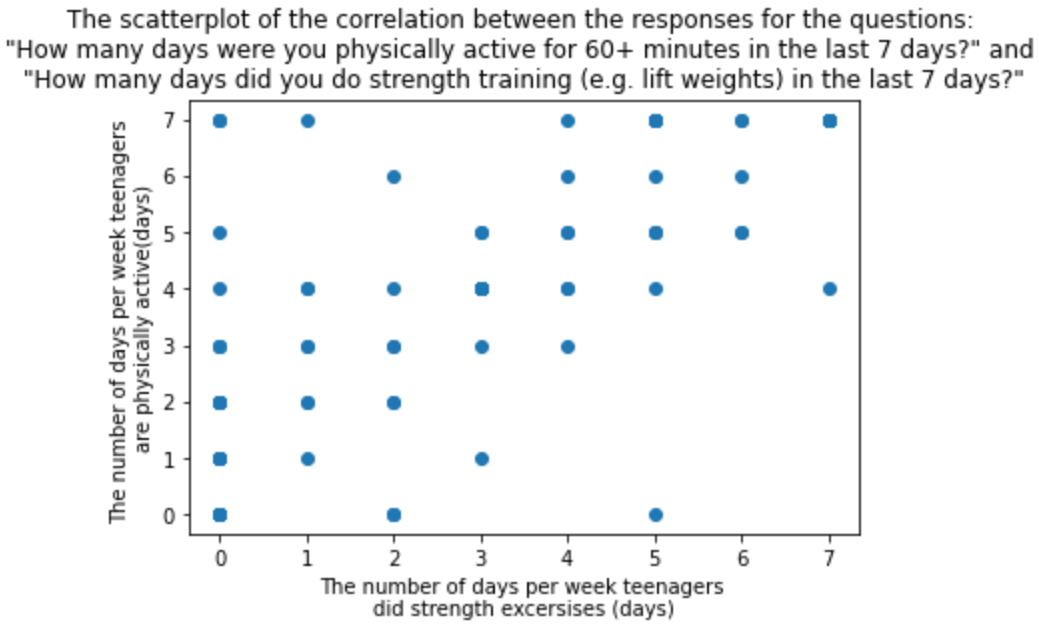


Figure 3: The scatterplot illustrates the relationship between the independent (strength training) on the X-axes and the dependent variable (active days) on the Y-axes. There is a positive linear association between the variables [2].

I- independence. Since there are only 100 respondents, the sample size is smaller than 10 % of the population, and the sampling was without replacement, so the assumption of independence is plausible.

[2] **#dataviz**: I used an appropriate type of datavisualisation to effective present the pattern. I added labels for axis specifying the units of variables represented on the axes. I added title and captions to explain the trend on the graph. I discussed the most interesping features of the graph (linearity) and used it for proving the condition required for further data analysis

N - normal residuals. The data pattern in different parts of the graph is different - more residuals above the approximate line of a relationship than below(figure 4). So, there is a trend in the errors, but it is larger or smaller across the graph (as opposed to the errors influencing each other). The fact influences the sum of squares and consequently affects the regression line, so the regression does not explain all trends in the dataset. Therefore the assumption about the normality of residuals is not plausible

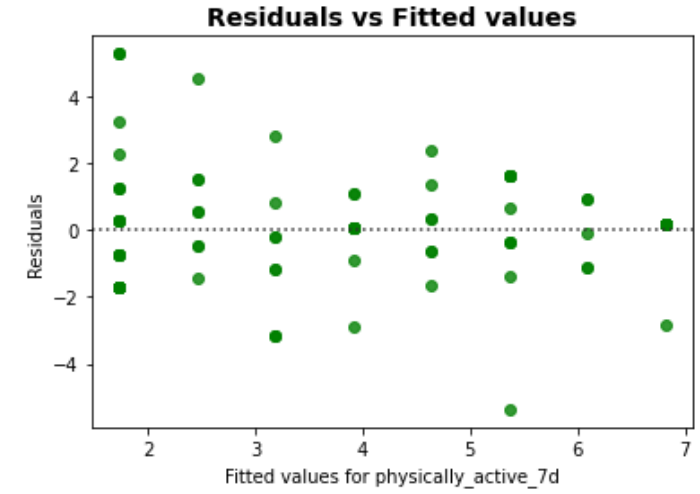


Figure 4: The figure illustrates the distribution of residuals on the regression model line. X-axes represent the fitted values for the dependent variable. Y-axes represent residuals. There are more residuals in the left part of the graph than in the right side.

E - equal variance (homoscedasticity). From figure 5 can be inferred that the variance is approximately normal. Since the points in the QQ-normal plot(figure 4) approximately lie on a straight diagonal line, the data is approximately normally distributed. Therefore the assumption of equal variance is plausible.

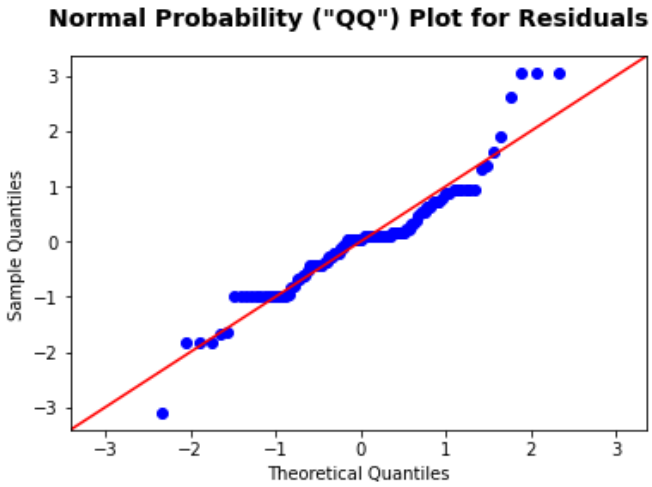


Figure 5: The figure illustrates the distribution of residuals on the regression model line. X-axes represent theoretical quantities(those that are expected if the trend is the same all over the model line). Y-axes represent sample quantities (those that are actually true for this model line). Despite some outliers, the general trend follows the straight line.

R- randomness. Since the data was collected from more than 4.8 million American high school students across 2,100 independent surveys(Center for Disease Control and Prevention, 2019) it proves the plausibility of randomness assumption by satisfying the requirements of the equality of chances for all members to get into the dataset resulting in the unbiased process.

Although not all the assumptions are plausible (residuals are not normally distributed) the

study will consider the underlying consequences of not satisfying all the assumptions and interpret the results accordingly.

Pearson's correlation coefficient

Pearson's r is an indicator of the strength of the linear relationship between variables, also known as correlation coefficient. The correlation coefficient is 0.7324 (without units) (calculated in Appendix D). Since the maximum possible value for Pearson's r is 1, 0.7324 is a number that indicates a strong relationship between the variables. So, points fall close to the straight line. A positive value indicates a positive linear relationship between the variables and implies that the line slopes upward. However, we cannot infer if x causes y or vice versa since correlation does not mean causation since the high correlatio coefficient may be a result of coincidence[3].

Coefficient of determination

The coefficient of determination (R^2) is 0.54 (calculations in appendix D). This means 54% of the variance in the number of days children were physically active can be explained by the number of days children did strength exercises. The equation of the regression line:

$$\hat{y} = 0.7266x + 1.73$$

The hat sign above y means that this is an estimation for y. The coefficient for x is 0.7266 and since both variables are measured in days/days that will cancel out. A positive slope tells that the

[3] **#correlation**: I computed and effectively interpreted correlation coefficient explaining the meaning of it for the data. I justified why the relationship is strong by giving the highes possible bound for the correlation coefficient. I recognized and effectively explained the difference between correlation and causation. In conclusion part I provided example of further actions that are necessary to conclude about possible causation(interventional study)

trend is upward. When X increases for 0.7266 days y increases for 1 day. 1.73 is the intercept with the x line, expressed in days (the same units as y). It represents the value of y when x is 0. In this case, it is possible to set the x value to zero and we have data in that part of the graph, so we can interpret this intercept for real life. So if teenagers do not do strength exercises at all we can expect them to move for 60+ minutes on average for 1.73 days per week[4].

Confidence intervals

Hull hypothesis: B = 0; the number of days teenagers do strength exercises is not in linear relationship with the number of days teenagers are active for at least 60 minutes (the slope of the correlation line is 0)

Alternative hypothesis: B ~= 0; the number of days teenagers do strength exercises is in linear relationship with the number of days teenagers are active for at least 60 minutes (slope is >0)

The alpha is set to 0.05 as standard. The 95% confidence interval was calculated to illustrate the range of the slope of the regression line illustrating the relationship between the variables. The first two conditions for the confidence interval calculations independence and randomness were proved above. The third condition is the normality of the distribution of residuals. From figure 4 it is clear that distribution is not normally shaped, so the intervals may have some degree of error.

[4] **#regression**: I accurately constructed and interpreted regression model justifying the relationship between the independent and devepndent varibale. I explained the significance of coefficient of determination in a given context. I computed the equation of the model and explained all its components. I addressed the possibility of getting the value for slope in real life conditions. In the conclusion part I used the regression model to predict the behaviour of the dependent variable and explained the result in the context of children's obesity.

95% confidence intervals for slope of regression line = [0.5999715842568739, 0.8772048863313614] (calculations in Appendix E)

This interval tells that in 95% of the cases, the true slope of the regression line of the population data will be within the interval[5]. This also serves as strong evidence to reject the null hypothesis since the null value was not in the interval. However, to fully reject the null hypothesis statistical significance test should have been performed.

Results and Conclusions

Data showed a strong correlation between the number of days teenagers do strength exercises and the number of days teenagers are active for at least 60 minutes, and 54% of the variance in the dependent variable can be explained with the independent variable. This information can be used for further research, specifically interventional study with treatment and control groups, to investigate the possible causal relationship between variables since correlation is not causation. If causality is confirmed, the model provides helpful insight for changing the dynamics of the dependent variable. For example, strength training for children may be encouraged by installing appropriate equipment in schools, and consequently, physical activity will be expected to increase. The study might have extraneous variables, for example, a preferred type of transportation(qualitative nominal variable) that influence the daily movement, that was

[5] **#confidenceintervals**: I checked the assumptions for confidence intervals, found the condition that is not satisfied and explained how it affects the accuracy of the calculations.I calculated and effectively interpreted the meaning of confidence intervals with the inferences for population. I explained confidence intervals as an estimator of population parament (slope). I also used confidence intervals in context of statistical significance by explaining the potential use of confidence intervals as an evidence against the null hypothesis.

not considered.

The conclusion is strong since if the the premises that are diverse, data is from reliable sources and are it is based on statistical tools. So, the conclusion is likely follows from premises. However, the conclusion is not reliable since one of the premises - normality of residuals condition was not satisfied. The conclusion is supported by generalization based on the sample was used since the study was performed on the sample but the conclusions were made about the population (all children). Induction was used is the prediction of the dependent variable using the regression line. Althoug the regression line is the best possible model based on the minimal sum of squares, it is only an estimate of the actual relationship.[6]

Word Count: 1433 **words**

Reflection

Statistical significance combined with correlation analysis will be valuable in strengthening the results produced by research. We can use confidence intervals to find the range of values for a slope that would work 95(90-99)% of the time. Also, the evaluation of null and alternative hypotheses will help evaluate the probability of the existence of the high relationship between the dependent and independent variables.

[6] **#induction**: I explained how inductive reasoning was applied to make conclusion. I commented on the strength of the argument explaining why I consider it to be strong. I explained how the implausible conditions of normality of residuals distributions affects the reliability of the bias.

References

Center for Disease Control and Prevention. (2019). *YRBSS Data & Documentation | DASH*. CDC. Retrieved December 9, 2021, from https://www.cdc.gov/healthyyouth/data/yrbs/data.htm

Helping All Children Grow Up Healthy. (2020). *Obesity Rates & Trend Data*. The State of Childhood Obesity. Retrieved December 9, 2021, from https://stateofchildhoodobesity.org/data/

Appendix

Appendix A: Importing dataset and performing data cleaning

```
#import libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
import statistics
import statsmodels.api as statsmodels


df = pd.read_csv (r'C:\Users\tetia\Downloads\yrbss_samp.csv') #read file with data
df = df.dropna(subset = ['physically_active_7d', 'strength_training_7d']) #clean rows that will be used in case there
                                                                          #are missing values
df.head()  #show the top 5 rows of the datast
```

]:

	age	gender	grade	hispanic	race	height	weight	helmet_12m	text_while_driving_30d	physically_active_7d	hours_tv_per_school_day	strength_trai
0	16.0	female	11.0	not	Black or African American	1.50	52.62	never	1-2	0	4	
1	17.0	male	11.0	not	White	1.78	74.84	rarely	0	7	1	
2	17.0	male	11.0	not	White	1.75	106.60	never	0	7	2	
3	15.0	male	10.0	hispanic	NaN	1.68	66.68	never	did not drive	3	2	
4	18.0	male	12.0	not	Black or African American	1.70	80.29	never	did not drive	0	2	

The libraries that were used for the data analysis are Pandas(for work with data frame), Matplotlib and Seaborn(for building graphs, data visualization), Scipy(for calculations), Statistics(for calculations), Statsmodels(for mathematical modeling the regression line)

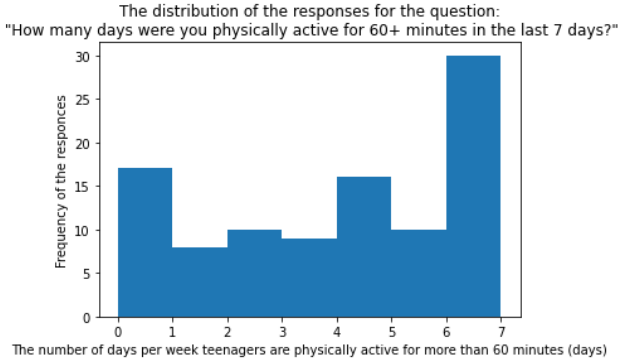
‘Dropna’ function was used to clean the data from the rows with missing values. This was done to ease the calculations and avoid errors. Python calculated only data that has the same type. If there is a missing value, it is understood as a different data type, resulting in an error.

Appendix B: Calculating descriptive statistics and plotting histograms

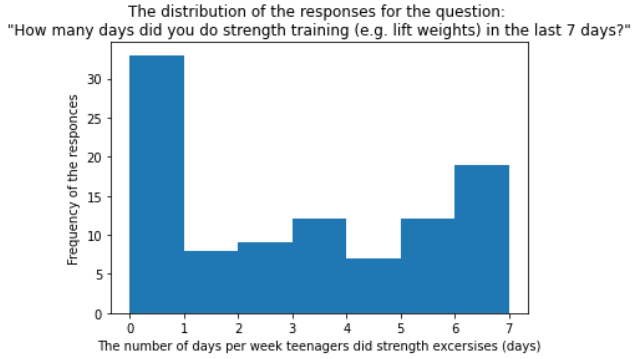
```
ny = len(df['physically_active_7d'])  #size of the sample (number of elements which is the same as length of the list)
meany = sum(df['physically_active_7d'])/ny #sum of the elements in the sample devided by the number of elements to get an ave
modey = statistics.mode(df['physically_active_7d']) #the most popular value
mediany = df['physically_active_7d'].median()
stdy = statistics.stdev(df['physically_active_7d'])#standard deviation of the sample(this function calculates it with n-1 in


print('stats for dependent y: n = {}, mean = {}, mode = {}, median = {}, std = {}'.format(ny,meany,modey,mediany,round(stdy,2

nx = len(df['strength_training_7d'])  #size of the sample (number of elements which is the same as length of the list)
meanx = sum(df['strength_training_7d'])/ny #sum of the elements in the sample devided by the number of elements to get an ave
modex = statistics.mode(df['strength_training_7d']) #the most popular value
medianx = df['strength_training_7d'].median()
stdx = statistics.stdev(df['strength_training_7d'])#standard deviation of the sample(this function calculates it with n-1 in


print('stats for independent x: n = {}, mean = {}, mode = {}, median = {}, std = {}'.format(nx,meanx,modex,medianx,round(stdx
```

```
stats for dependent y: n = 100, mean = 3.75, mode = 7, median = 4.0, std = 2.56
stats for independent x: n = 100, mean = 2.78, mode = 0, median = 2.5, std = 2.58
```

```
plt.hist(df['physically_active_7d'],bins = 7)
plt.xlabel('The number of days per week teenagers are physically active for more than 60 minutes (days)')
plt.ylabel('Frequency of the responces')
plt.title('The distribution of the responses for the question:\n "How many days were you physically active for 60+ minutes in
plt.show()
```

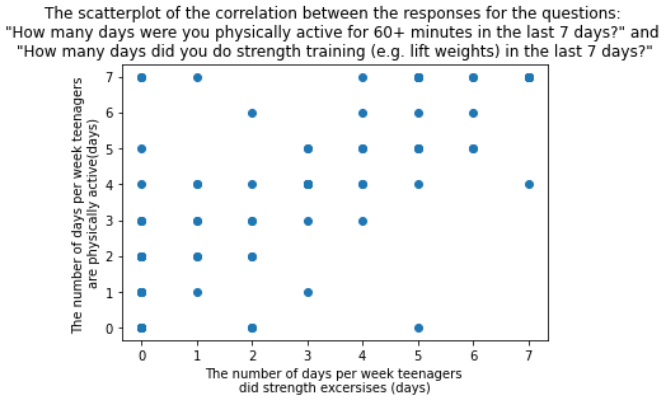


```
plt.hist(df['strength_training_7d'],bins = 7)
plt.xlabel('The number of days per week teenagers did strength excersises (days)')
plt.ylabel('Frequency of the responces')
plt.title('The distribution of the responses for the question: \n"How many days did you do strength training (e.g. lift weigh
plt.show()
```

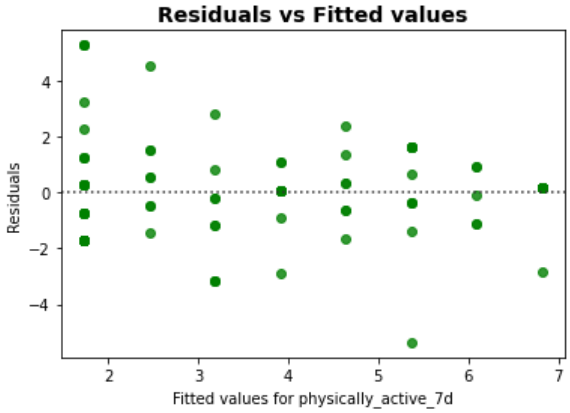


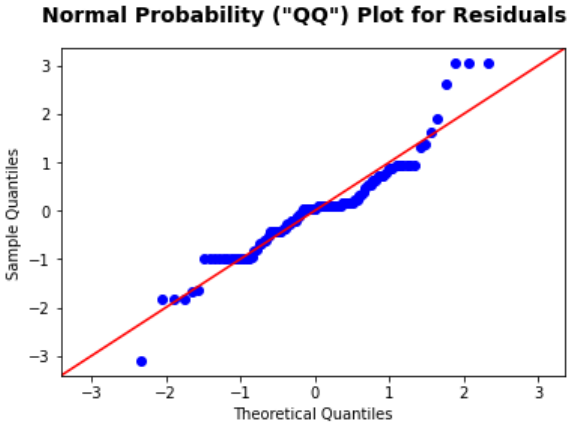
Appendix C: Checking conditions for regression

```
plt.scatter( df['strength_training_7d'], df['physically_active_7d'])
plt.ylabel('The number of days per week teenagers \nare physically active(days)')
plt.xlabel('The number of days per week teenagers \ndid strength excersises (days)')
plt.title('The scatterplot of the correlation between the responses for the questions: \n"How many days were you physically a
plt.show()
```




```
# residual plot:
plt.figure()
residualplot = sns.residplot(x=regressionmodel.predict(), y=regressionmodel.resid, color='green')
residualplot.set(xlabel='Fitted values for '+column_y, ylabel='Residuals')
residualplot.set_title('Residuals vs Fitted values',fontweight='bold',fontsize=14)

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


Residuals vs Fitted values
Residuals
Fitted values for physically_active_7d


Normal Probability ("QQ") Plot for Residuals
Sample Quantiles
Theoretical Quantiles

Appendix D: Calculating correlation coefficient and the regression line

```
active_list = df['physically_active_7d'].tolist() #convert dataframe to to list
strong_list = df['strength_training_7d'].tolist()


stats.pearsonr(strong_list ,active_list) #use function to salculate correlation coefficient
```

```
]: (0.7324358559228007, 4.730879238529133e-18)
```

```
def mult_regression(column_x, column_y): #define the function for regression line calculation
    if len(column_x)==1:  # If there is only one predictor variable, plot the regression line
        plt.figure() #plot figure
        sns.regplot(x=column_x[0], y=column_y, data=data, marker="+",fit_reg=True,color='orange')

    # define independent X and dependent Y:
    X = df['strength_training_7d']
    X = statsmodels.add_constant(X)
    Y = df['physically_active_7d']

    # construct model:
    global regressionmodel
    regressionmodel = statsmodels.OLS(Y,X).fit() # OLS = "ordinary least squares"

    # residual plot:
    plt.figure()
    residualplot = sns.residplot(x=regressionmodel.predict(), y=regressionmodel.resid, color='green')
    residualplot.set(xlabel='Fitted values for '+column_y, ylabel='Residuals')
    residualplot.set_title('Residuals vs Fitted values',fontweight='bold',fontsize=14)

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

mult_regression(df['strength_training_7d'],'physically_active_7d')
regressionmodel.summary()
```

OLS Regression Results

Dep. Variable:	physically_active_7d	**R-squared:**	0.536
Model:	OLS	**Adj. R-squared:**	0.532
Method:	Least Squares	**F-statistic:**	113.4
Date:	Mon, 24 Jan 2022	**Prob (F-statistic):**	4.73e-18
Time:	10:29:09	**Log-Likelihood:**	-196.79
No. Observations:	100	**AIC:**	397.6
Df Residuals:	98	**BIC:**	402.8
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.7300	0.258	6.705	0.000	1.218	2.242
strength_training_7d	0.7266	0.068	10.650	0.000	0.591	0.862

Omnibus:	12.769	**Durbin-Watson:**	1.923
Prob(Omnibus):	0.002	**Jarque-Bera (JB):**	20.764
Skew:	0.535	**Prob(JB):**	3.10e-05
Kurtosis:	4.959	**Cond. No.**	5.80

Appendix E: Calculating confidence intervals

20

```
r = 0.73 #correlation coefficient

b1 = r*stdy/stdx # the formula for the point-estimate for the slope
SE = stdy/stdx*((1-r**2)/(nx-2))**0.5# formula with the standard error
t = stats.t.ppf(0.975,nx-2) # calculate conficence intervals

lower_bound = b1 - t*SE #define the lower bound
upper_bound = b1 + t*SE #define the upper bound

print("b1 =",b1, "\nSE =",SE, "\nt =",t, "\ninterval =", [lower_bound,upper_bound])
```

```
b1 = 0.7241949938579935
SE = 0.06848958804854584
t = 1.984467454426692
interval = [0.5882796354085629, 0.860110352307424]
```