Correlation and Regression Report

1. Introduction

The intrinsic value of education is to witness progress in students. Progress can be demonstrated not only through grades or teacher and parental evaluation but also through students' own self-efficacy. From an academic viewpoint, we would like to explore the improvements made by secondary students in study results of Maths throughout the second period. In this report, we investigate the correlation between students' second period grades and their final grades. We first process the data by calculating the descriptive statistics and compute the distribution of residuals to verify if the conditions hold to conduct further regression analysis. Then, we calculate and interpret the Pearson's correlation coefficient (r), the coefficient of determination (R-squared), and a regression equation that describes the model. We also perform a test for statistical significance to determine whether students's final grades are higher than their second period grades. In other words, is there any convincing evidence that there is any linear relationship between students' second period grades and their final grades. We really want to see the validity of one upcoming conclusion about students' progress throughout the period.

2. Dataset

The two datasets aim at approaching Portuguese secondary student achievement and are collected through school reports and questionnaires. Data is extracted from the source of Paulo Cortez, University of Minho, Portugal. As I am more interested in Mathematics than Portuguese subjects, in this report I look at the Maths dataset. As I am curious about students' study progress throughout the second period, I would like to decide to filter data to make a comparison between second period grades and final grades in Maths of students. One variable of interest is grades, a quantitative variable. To specify, in this dataset, the variable of grades is discrete because grades are countable and only take limited values ranging from 0 to 20. Moreover, in reality, no one is graded at 9.5424 for example. Therefore, we can pick a specific number for a grade, which goes against a continuous variable. Another variable involved in this analysis is subjects. Mathematics would be a categorical qualitative variable. Moreover, we can consider second period grades and final grades to be the predictor variable and response variable, respectively, because we want to predict students' progress at the end of the period based on the regression model. In a regression analysis, we also have to take extraneous variables into consideration. The study progress of

students may also be attributed to higher quality teachers, more manageable problems in final tests or better health conditions of students at the end of the period.[1]

3. Methods

a. Summary Statistics

The dataset was processed in Python. We first calculated the descriptive statistics for second period grades and final grades, including mean, median, mode, range, and standard deviation. Table 1 provides the summary statistics for both second period grades and final grades (these are computed in Appendix A). The sample distributions for each quantity are displayed in Figures 1 and 2 (these are computed in Appendix B).

Table 1: Summary statistics for the variables of interest

	Second Period Grades	Final Grades
Count	$n = 395$	$n = 395$
Mean	$\bar{x} = 10.71$	$\overline{y} = 10.42$
Median	11	11
Mode	9	10
Standard Deviation	$s_x = 3.76$	$s_y = 4.58$
Range	19	20

[1] #variables: I classified and defined the appropriate type of variables in this model, ranging from quantitative, qualitative to extraneous variables considered. I also explained detailed their distinction as well as the correlation relationship between the predictor and response variables

Figure 1. Histogram for second period Mathematics grades of secondary students

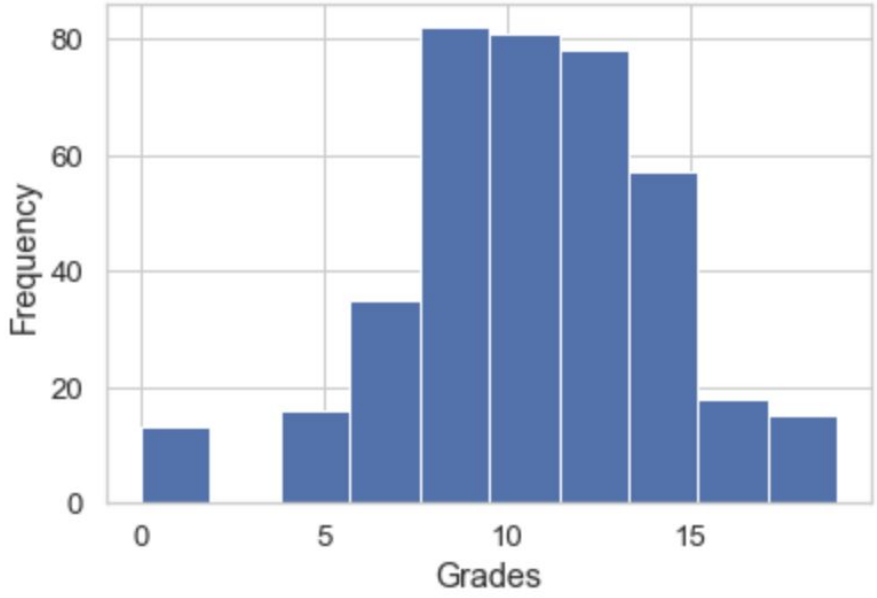
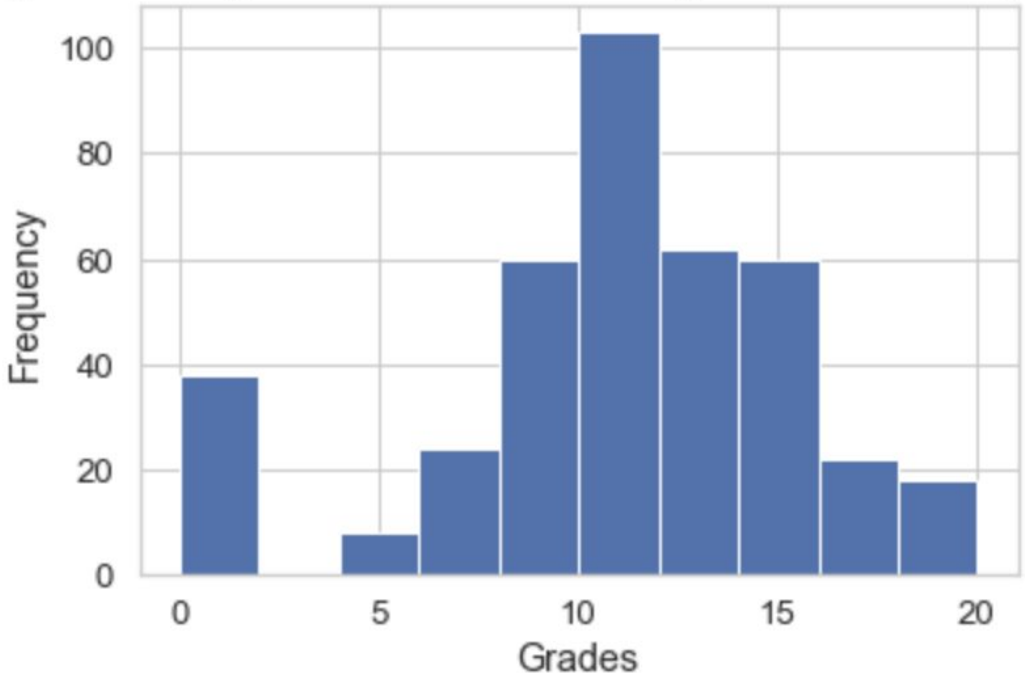


Figure 2. Histogram for final Mathematics grades of secondary students



From the two histograms, we can see that both share nearly normal distributions. However, there are some minor irregularities or outliers in the left that we really need to be cautious about.

Furthermore, we can see that in both samples, the medians are larger than the mean, yet to a slight degree, which suggests no serious concerns.

b. Evaluating Conditions

Since we do not have data of all students' grades, we do not know the population standard deviation. We choose the t-distribution for inference and estimate the standard error using the sample standard deviation. Furthermore, we also use Bessle's correction to the standard deviation to correct the bias in the estimation of the population variance.

The three following figures are computed in Appendix C.

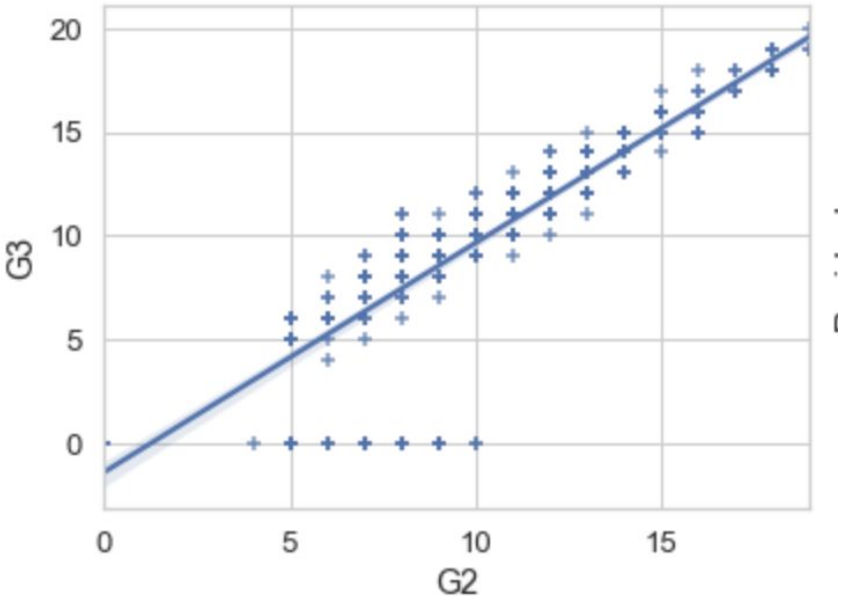


Figure 3. The scatter plot represents the relationship between G2 and G3 (second period grades and final grades). It shows a strong positive linear relationship between the secondary period grades and the final grades of students in Mathematics.

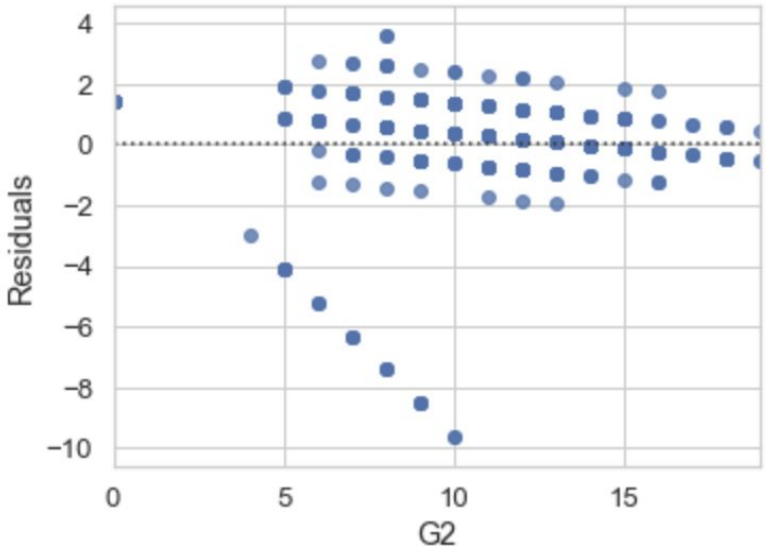


Figure 4. Residuals vs fitted values plot

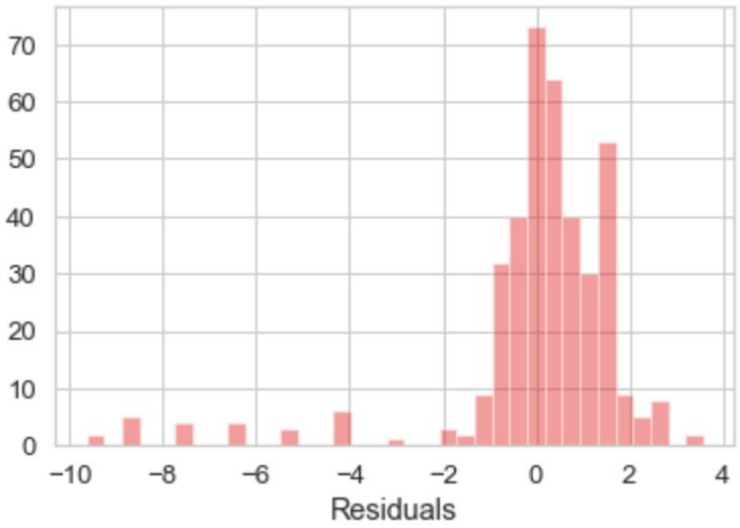


Figure 5. The histogram shows the distribution of residuals. It indicates normal distribution with few outliers.

Before calculating the correlation coefficient or the coefficient of determination, we must check if four LINE conditions are met.

- Linearity: From Figure 3, we can see a positive linear relationship between the second period grades and the final grades. Therefore, this condition holds.
- Independent observations: This condition also holds because one data point represents grades of one student, which is independent data points.

- Nearly normal residuals: From Figure 5, we can see a normal distribution of residuals with few outliers on the left. However, with such a large dataset, this should not be a serious concern. For the sake of our regression analysis, we assume that this condition holds.
- Equal variances: The Figure 4 shows that the graph is quite homoscedastic. However, there are some outliers which can be detrimental to predicting the values in the model. Therefore, to conduct the regression analysis, we will assume that this condition holds but we should be cautious of the results which can be less accurate.

Since all conditions are met, we are justified to conduct our regression analysis. [2]

c. Correlation Coefficient

We proceed to calculate the Pearson's correlation coefficient to evaluate the strength of this linear relationship. We first compute the mean. Then, we subtract the original data from the mean and divide each of them by the standard deviation to get the standard units. We repeat this process for the other variable. The correlation coefficient is the average product of corresponding pairs of the two variables in standard units. In this report, we use Python to compute the correlation coefficient.

[2] #distributions: This is a strong application of this HC because I correctly identified the type of distribution as well as checked all four conditions fully. I dealt with the conditions sufficiently by explaining some outstanding features and their influence. From this influence, I am aware of my conclusions made to reduce the reliability of the inductive reasoning in this report.

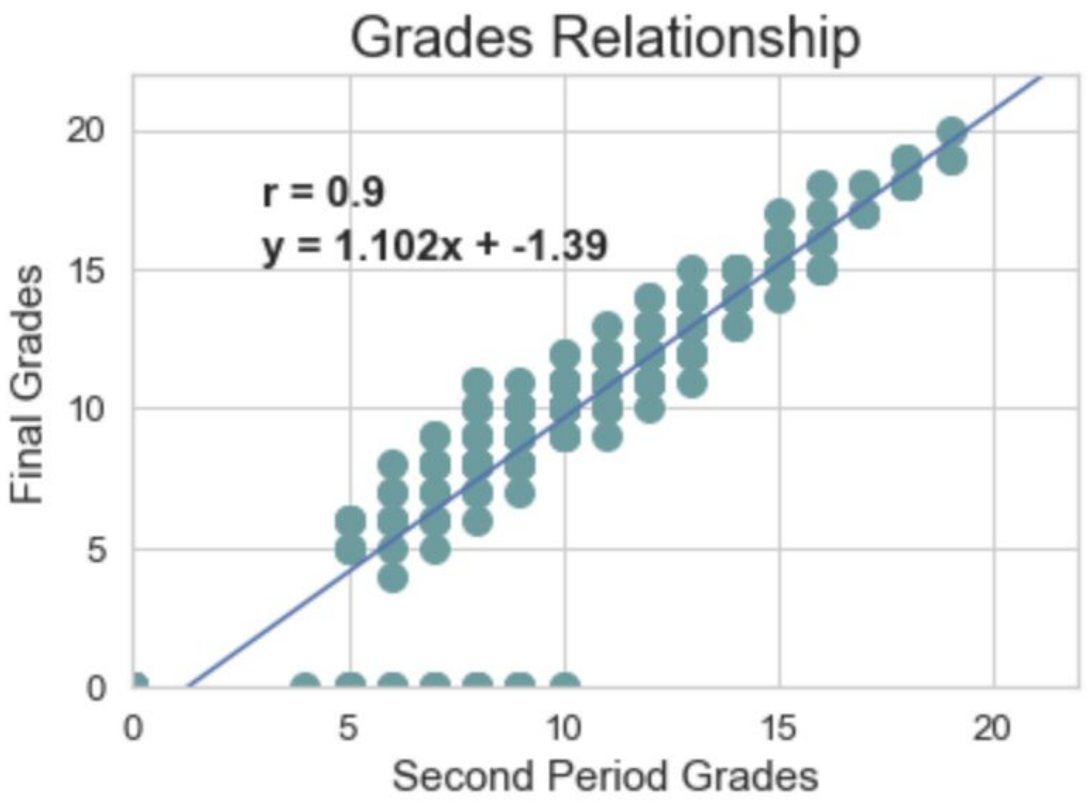


Figure 6. The positive linear relationship between the second period grades and the final grades of students in Mathematics. The graph also includes the trendline which plots the best fit maintaining the closest difference with all data points.

From Figure 6 (computed in Appendix D), we can conclude that the correlation coefficient is 0.9. *This is a comparatively high and positive correlation coefficient which proves a strong positive linear relationship between second period grades and final grades in Mathematics among secondary students*. Indeed, there are a lot of data points which stay close to the trendline. We should also be cautious to infer any causal links between the two variables. Specifically, we cannot infer any causal relationship between second period grades and final grades even though the correlation coefficient is high. Sometimes, there exist some extraneous variables involved in students' progress, namely change in teacher, less challenging study workload or students' health. These extraneous variables prevent us from concluding a causality that second period grades give rise to final grades. [3][4]

d. Coefficient of Determination

[3] #correlation: This is a strong application of this HC because I correctly calculated and interpreted the correlation coefficient. Moreover, I also presented some reasons for why we should not infer causality with appropriated extraneous variables.

[4] #dataviz: In this paper, I have applied #dataviz so effectively. I not only generated appropriate data visualizations but also interpreted detailedly for the use of condition checking or observations.

Then, we calculate the coefficient of determination (R-squared) to explain the strength of a linear fit. We use the formula:

$$R^2 = 1 - \frac{variability\ in\ residuals}{variablity\ in\ the\ outcome} = 1 - \frac{SSE}{SSTO}$$

In this model, SSE refers to the sum of the squares of residuals and SSTO represents the sum squared total error. However, we use Python to compute the R-squared as well as many other results, listed in the Table 2 (computed in Appendix E)

Table 2: The model summary of our regression analysis:

Dep. Variable:	G3	**R-squared:**	0.819
Model:	OLS	**Adj. R-squared:**	0.818
Method:	Least Squares	**F-statistic:**	1776.
Date:	Fri, 29 Jan 2021	**Prob (F-statistic):**	7.63e-148
Time:	16:22:10	**Log-Likelihood:**	-823.83
No. Observations:	395	**AIC:**	1652.
Df Residuals:	393	**BIC:**	1660.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.3928	0.297	-4.690	0.000	-1.977	-0.809
G2	1.1021	0.026	42.139	0.000	1.051	1.154

Omnibus:	246.646	**Durbin-Watson:**	1.866
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1776.620
Skew:	-2.699	**Prob(JB):**	0.00
Kurtosis:	11.877	**Cond. No.**	34.5

Therefore, the correlation of determination is 0.819. For an interpretation, an R-squared value of 0.819 justifies *81.9% of the total variation in the final grades (the response variable) is explained in the second period grades (the predictor variable)*. This high percentage proves a

really close difference between data points and the best fit line. It is not 100% because the approximately remaining 20% of the grades does not depend on the students, but on something else. For example, the workload at the end of the period can be lighter to some extent or teachers imposed less control on students' communication during the final test.

e. Regression Equation

To construct a regression equation, we need to first figure out the slope as well as the y-intercept. We use a formula to calculate the slope of the regression model:

$$\beta_1 = \frac{s_y}{s_x} * r$$

s_y , s_x refers to the standard deviation of final grades and second period grades (computed in Table 1), respectively. We can compute the slope with the given formula. Moreover, because the regression line always passes through the point ($\bar{x}$, $\bar{y}$), we can calculate the y-intercept thanks to the computed values of means in Table 1.

In this report, we use Python to compute the regression equation in Figure 6:

$$\widehat{finalgrades} = -1.39 + 1.102 \times \widehat{secondperiodgrades}$$

The slope here is 1.102 which can be interpreted that *for each additional score in second period grades, there will be a 1.102 increase in the final grades of students in Mathematics*. The slope is positive, which also justifies a positive linear relationship between second period grades and final grades.

Moreover, the y-intercept of -1.39 also shows that *students who scored a 0 in the second period are expected to earn a -1.39 final grade*. It does not make sense to have a final grade of -1.39 because grades are never negative. Grades in this dataset only range from 0 to 20. Here, the y-intercept serves only to adjust the height of the line and is meaningless by itself.[5]

f. Confidence Intervals

To ensure our confidence interval works well, we must check if three conditions are met:

- Random: Students' school profiles and questionnaires are randomly collected from a Portuguese school. Therefore, we check this condition
- Normal: n > 30. This condition holds because both sample groups of grades have sample size larger than 30 (395 specifically)

[5] #regression: This is an awesome application of this HC. I not only calculated and interpreted detailedly the R-squared, the slope and the intercept but also showed a deep understanding of their underlying meaning.

- Independent: We check this condition because n < 10% of population size.

Since these conditions are met, we are justified to use the t-distribution. As a result, we can apply the Central Limit Theorem for all samples to be normally distributed when the number of samples gets larger.

We choose to construct a 95% confidence interval for the slope. We choose t-score instead of z-score because population standard deviation is unknown and we need greater exactitude. We need to use degrees of freedom instead of sample size to make our results more accurate, using the formula: $df = n - 1$. With confidence level and degrees of freedom, we can find out the t-score thanks to the calculator. We then compute the standard error using the usual formula: $SE = \frac{SD}{\sqrt{df}}$. The sample standard deviation is known, thus facilitating us to do the calculation. The margin of error would be the product of t-score and the standard error.

In this report, we use Python to compute the confidence interval which is (1.051, 1.154) (Table 2). This confidence interval is significant because it does not capture 0, thus reinforcing the possibility that students will make progress in the final exams. For an interpretation, *we are 95% confident that students' Maths final grades are higher than their second period grades from 1.051 scores to 1.154 scores*. It also means that *if we conduct taking samples with the same sample size multiple times, 95% of confidence intervals we construct from these samples will contain the population coefficient.*

4. Results and Conclusion

Throughout the report of examining the data about students' Maths grades in the second period and final, we successfully computed the correlation coefficient of 0.9 which justifies a strong positive linear relationship between the second period grades and final grades. These two variables are strongly and positively correlated. Moreover, with a coefficient of determination of 0.819, we can conclude that 81.9% of the total variation of final grades are explained by the model. We also successfully constructed a 95% confidence interval for the slope of a regression equation, (1.051, 1.154) (outputted in Table 2). From that, we can learn that with the same sample size, if we take out several samples and construct one confidence interval. Correspondingly, 95% of the confidence intervals will contain the population mean. The results should be carefully considered to be applied to real life because this inductive reasoning, despite with large samples, still contains some limitations concerning the residual distribution as well as few outliers.[6]

[6] #confidenceintervals: This is a strong application of this HC because I calculated several components and successfully constructed an appropriate confidence interval with clearly stated steps. Also, I interpreted exactly the meaning of the confidence interval and applied this into real-life understanding.

These conclusions are supposed to be inductive because we only examine the data about samples, and process the data to gain insights into statistics of students in a Portuguese school. Based on sample examination, we reached the conclusion about population groups. Therefore, the conclusion is not contained within the scope of premises and we are using the method of generalization. We still need more samples with larger sample sizes for the sake of greater accuracy. Furthermore, the strength of this inductive reasoning is also weakened with some assumptions and outliers in the Evaluating Conditions part.

In addition, we should be cautious to not infer any causal relationships between variables here because although the correlation coefficient is high, there still exist many extraneous variables we have not addressed. To illustrate, the progress in study performance can result from higher quality in teachers, better health conditions in students or less difficult exercises in the final exams. As a result, we would say that the reliability of this inductive reasoning is limited to some extent. [7]

5. Reflection

The Statistics unit really inspired the way I can derive conclusions and any relationships between things in life. Data is always confusing to me. I was helped a lot to process and manipulate data to highlight its patterns for my conclusions. While correlation and regression specifies on justifying the relationship between two variables, confidence intervals bring the conclusion closer to the reality with more uncertainty and wider ranges to embrace the data.

6. References

The full Jupyter notebook file can be accessed here [http://localhost:8888/notebooks/Downloads/Untitled1.ipynb?kernel_name=python3].
The data can be accessed here.

7. Appendix

Appendix A: Import Data

[7] #induction: This is an awesome application of this HC because I accurately and effectively identified inductive reasoning and justified it with inductive characteristics. I also specified the type of induction and explained it carefully. Moreover, I justified the reason why this inductive reasoning can be limited (due to some imperfections in data)

In [16]:

```
import pandas
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
from scipy.stats import linregress# very nice plotting package

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

# import data
data = pandas.read_csv("csv-student-mat.csv")
#print data
data
```

Out[16]:

	Unnamed: 0	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	0	GP	F	18	U	GT3	A	4	4	at_home	...	4	3	4	1	1	3	6	5	6	6
1	1	GP	F	17	U	GT3	T	1	1	at_home	...	5	3	3	1	1	3	4	5	5	6
2	2	GP	F	15	U	LE3	T	1	1	at_home	...	4	3	2	2	3	3	10	7	8	10
3	3	GP	F	15	U	GT3	T	4	2	health	...	3	2	2	1	1	5	2	15	14	15
4	4	GP	F	16	U	GT3	T	3	3	other	...	4	3	2	1	2	5	4	6	10	10
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
390	390	MS	M	20	U	LE3	A	2	2	services	...	5	5	4	4	5	4	11	9	9	9
391	391	MS	M	17	U	LE3	T	3	1	services	...	2	4	5	3	4	2	3	14	16	16
392	392	MS	M	21	R	GT3	T	1	1	other	...	5	5	3	3	3	3	3	10	8	7
393	393	MS	M	18	R	LE3	T	3	2	services	...	4	4	1	3	4	5	0	11	12	10
394	394	MS	M	19	U	LE3	T	1	1	other	...	3	2	3	3	3	5	5	8	9	9

395 rows × 34 columns

Appendix B: Analyze and Visualize Data

In [17]:
```
#create a list to give out values
secondperiod_list = data['G2'].tolist()
final_list = data['G3'].tolist()

#calculate descriptive stats
mode_secondperiod = stats.mode(secondperiod_list)[0][0]
mean_secondperiod = data['G2'].mean()
median_secondperiod = data['G2'].median()
sd_secondperiod = data['G2'].std()

print('- Mean =', mean_secondperiod)
print('- Median =', median_secondperiod)
print('- Mode =', mode_secondperiod)
print('- Standard deviation =', sd_secondperiod)
print('- Range =', max(secondperiod_list) - min(secondperiod_list))
```

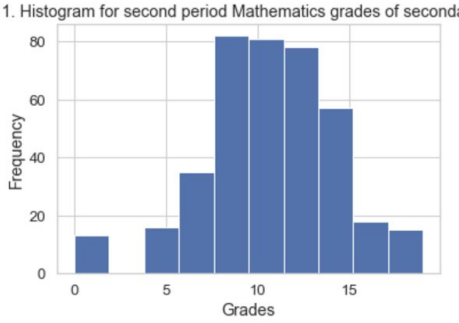
```
- Mean = 10.713924050632912
- Median = 11.0
- Mode = 9
- Standard deviation = 3.761504659556034
- Range = 19
```

In [18]:
```
mode_final = stats.mode(final_list)[0][0]
mean_final = data['G3'].mean()
median_final = data['G3'].median()
sd_final = data['G3'].std()

print('- Mean =', mean_final)
print('- Median =', median_final)
print('- Mode =', mode_final)
print('- Standard deviation =', sd_final)
print('- Range =', max(final_list) - min(final_list))


```

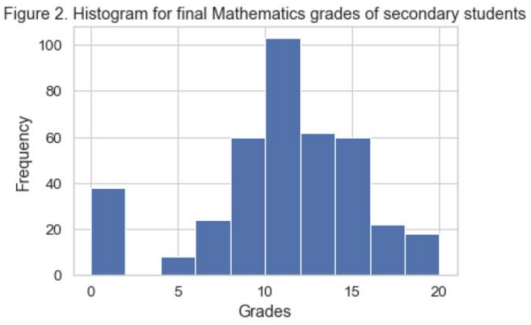
```
- Mean = 10.415189873417722
- Median = 11.0
- Mode = 10
- Standard deviation = 4.5814426109978434
- Range = 20
```

In [26]:
```
secondperiod = list(data['G2'].values)
final = list(data['G3'].values)
#plot histograms
plt.hist(secondperiod, bins = 10)
plt.title('Figure 1. Histogram for second period Mathematics grades of secondary students')
plt.xlabel('Grades')
plt.ylabel('Frequency')
plt.show()

```

Figure 1. Histogram for second period Mathematics grades of secondary students

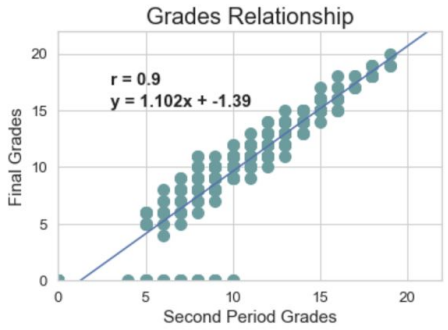


In [28]:
```
plt.hist(final, bins = 10)
plt.title('Figure 2. Histogram for final Mathematics grades of secondary students')
plt.xlabel('Grades')
plt.ylabel('Frequency')
plt.show()
```



Appendix C: Calculate the Correlation Coefficient and the Regression Equation

In [29]:
```
def plotscatter(x, y, title, color, xmax):
    plt.figure()
    plt.scatter(x, y, s=100, c=color) #choose the type of data visualization - here is scatter plot
    plt.title(title, fontsize=20) #name a title
    plt.xlabel('Second Period Grades', fontsize=15)
    plt.ylabel('Final Grades', fontsize=15)
    plt.xlim(0,xmax) #put a limit to the horizontal and vertical axes
    plt.ylim(0,22)

    # calculates slope, intercept, r_value for dataset
    slope, intercept, r_value, p_value, std_err = linregress(x, y)
    plt.plot([0, xmax], [intercept, slope * xmax + intercept])

    # adds legend to figure with regression line equation
    equation = 'y = ' + str(round(slope,3)) + 'x' + ' + ' + str(round(intercept,2))
    rvalue = 'r = ' + str(round(r_value,2))
    #locate where to put the correlation coefficient and the regression equation
    plt.text(3, 15.3, equation,fontsize=15,fontweight='bold')
    plt.text(3, 17.3, rvalue,fontsize=15,fontweight='bold')

#give out the plot with results
plotscatter(secondperiod, final, 'Grades Relationship', 'cadetblue', 22)
```

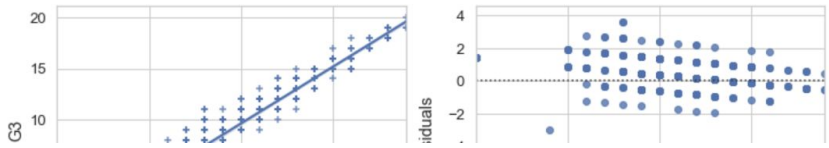


Appendix D: Check Conditions and Calculate the Coefficient of Determination

In [2]:
```
def regression_model(column_x, column_y):
    # this function uses built in library functions to create a scatter plot,
    # plots of the residuals, 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]
    regressionmodel = statsmodels.OLS(Y,X).fit() #OLS stands for "ordinary least squares"

    # extract regression parameters from model, rounded to 3 decimal places:
    Rsquared = round(regressionmodel.rsquared,3)
    slope = round(regressionmodel.params[1],3)
    intercept = round(regressionmodel.params[0],3)

    # 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') #name the values to be Residuals
    ax2.set_ylim(min(regressionmodel.resid)-1,max(regressionmodel.resid)+1)
    plt.figure() # histogram
    sns.distplot(regressionmodel.resid, kde=False, axlabel='Residuals', color='red') # histogram

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

In [6]:
```
regression_model('G2','G3')
```

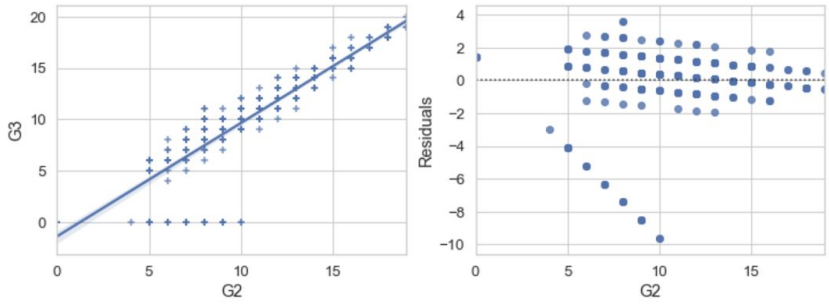
```
R-squared =  0.819
Regression equation: G3 =  1.102 * G2 +  -1.393
```

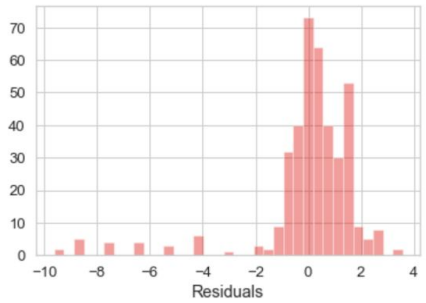



```
    # print the results:
    print("R-squared = ",Rsquared)
    print("Regression equation: "+column_y+" = ",slope,"* "+column_x+" + ",intercept)
```

In [6]:
```
regression_model('G2','G3')
```

```
R-squared =  0.819
Regression equation: G3 =  1.102 * G2 +  -1.393
```



Appendix E: Construct Confidence Intervals

In [8]:
```
#call a variable and assign a column in the dataset to it
predictor_vars = ['G2']

X = data[predictor_vars]
X = statsmodels.add_constant(X) # if excluded, the intercept would default to 0
y = data['G3']
model = statsmodels.OLS(y, X).fit()
model.summary() #print out the summary of all statistics
```

Out[8]:

OLS Regression Results

Dep. Variable:	G3	**R-squared:**	0.819
Model:	OLS	**Adj. R-squared:**	0.818
Method:	Least Squares	**F-statistic:**	1776.
Date:	Fri, 29 Jan 2021	**Prob (F-statistic):**	7.63e-148
Time:	16:22:10	**Log-Likelihood:**	-823.83
No. Observations:	395	**AIC:**	1652.
Df Residuals:	393	**BIC:**	1660.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.3928	0.297	-4.690	0.000	-1.977	-0.809
G2	1.1021	0.026	42.139	0.000	1.051	1.154

Omnibus:	246.646	**Durbin-Watson:**	1.866
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1776.620
Skew:	-2.699	**Prob(JB):**	0.00
Kurtosis:	11.877	**Cond. No.**	34.5

Word count: 1358