Going the Distance

Author: Michael Govaerts

In [352]:

```
import pandas as pd
```

In [353]:

```
# Read csv file, noting that the data is separated with semicolons rather than the defaul

df = pd.read_csv("student(demo+grades)\student-mat.csv", sep=';')
```

Attribute Information:

Source:
http://web.archive.org/web/20210209231731/https://archive.ics.uci.edu/ml/datasets/Student+Perform

# #

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

2 sex - student's sex (binary: 'F' - female or 'M' - male)

3 age - student's age (numeric: from 15 to 22)

4 address - student's home address type (binary: 'U' - urban or 'R' - rural)

5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)

8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)

9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')

13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)

16 schoolsup - extra educational support (binary: yes or no)

17 famsup - family educational support (binary: yes or no)

18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

19 activities - extra-curricular activities (binary: yes or no)

20 nursery - attended nursery school (binary: yes or no)

21 higher - wants to take higher education (binary: yes or no)

22 internet - Internet access at home (binary: yes or no)

23 romantic - with a romantic relationship (binary: yes or no)

24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)

26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)

27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

29 health - current health status (numeric: from 1 - very bad to 5 - very good)

30 absences - number of school absences (numeric: from 0 to 93)

# #

these grades are related with the course subject, Math or Portuguese:

31 G1 - first period grade (numeric: from 0 to 20)

31 G2 - second period grade (numeric: from 0 to 20)

32 G3 - final grade (numeric: from 0 to 20, output target)

In [355]:

```
#Review data set in tabular form to confirm it has been processed correctly

df.head()
```

Out[355]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	travelt
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	
1	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	
2	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	
3	GP	F	15	U	GT3	T	4	2	health	services	home	mother	
4	GP	F	16	U	GT3	T	3	3	other	other	home	father	

In [356]:

```
# Limit data set to values that seem relevant to travel time and performance

df2 = df[['reason', 'traveltime', 'studytime', 'activities', 'G1', 'G2', 'G3', 'absences'
```

In [357]:

```
df2.head()
```

Out[357]:

	reason	traveltime	studytime	activities	G1	G2	G3	absences
0	course	2	2	no	5	6	6	6
1	course	1	2	no	5	5	6	4
2	other	1	2	no	7	8	10	10
3	home	1	3	yes	15	14	15	2

	reason	traveltime	studytime	activities	G1	G2	G3	absences
4	home	1	2	no	6	10	10	4

In [358]:

```
# Review descriptive statistics of values to begin to analyze data

df2.describe()
```

Out[358]:

	traveltime	studytime	G1	G2	G3	absences
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
mean	1.448101	2.035443	10.908861	10.713924	10.415190	5.708861
std	0.697505	0.839240	3.319195	3.761505	4.581443	8.003096
min	1.000000	1.000000	3.000000	0.000000	0.000000	0.000000
25%	1.000000	1.000000	8.000000	9.000000	8.000000	0.000000
50%	1.000000	2.000000	11.000000	11.000000	11.000000	4.000000
75%	2.000000	2.000000	13.000000	13.000000	14.000000	8.000000
max	4.000000	4.000000	19.000000	19.000000	20.000000	75.000000

Unfortunately, we are able to see from this data set that few students traveled very far to get to the school they attended.

In [359]:

```
# Confirm trend away from long travel times to school.  As we can see from the below outp

df2.mode()
```

Out[359]:

	reason	traveltime	studytime	activities	G1	G2	G3	absences
0	course	1	2	yes	10	9	10	0

In [360]:

```
# Determine if there are any clear correlations between travel time, study time, absences

df2.corr()
```

Out[360]:

	traveltime	studytime	G1	G2	G3	absences
traveltime	1.000000	-0.100909	-0.093040	-0.153198	-0.117142	-0.012944
studytime	-0.100909	1.000000	0.160612	0.135880	0.097820	-0.062700
G1	-0.093040	0.160612	1.000000	0.852118	0.801468	-0.031003
G2	-0.153198	0.135880	0.852118	1.000000	0.904868	-0.031777
G3	-0.117142	0.097820	0.801468	0.904868	1.000000	0.034247
absences	-0.012944	-0.062700	-0.031003	-0.031777	0.034247	1.000000

As we can see from the table above, travel time is negatively correlated with study time and course grades at all periods.
Interestingly, travel time appears to have a stronger negative correlation with final course grade than the positive correlation that study time has with final course grade. This outcome may suggest that

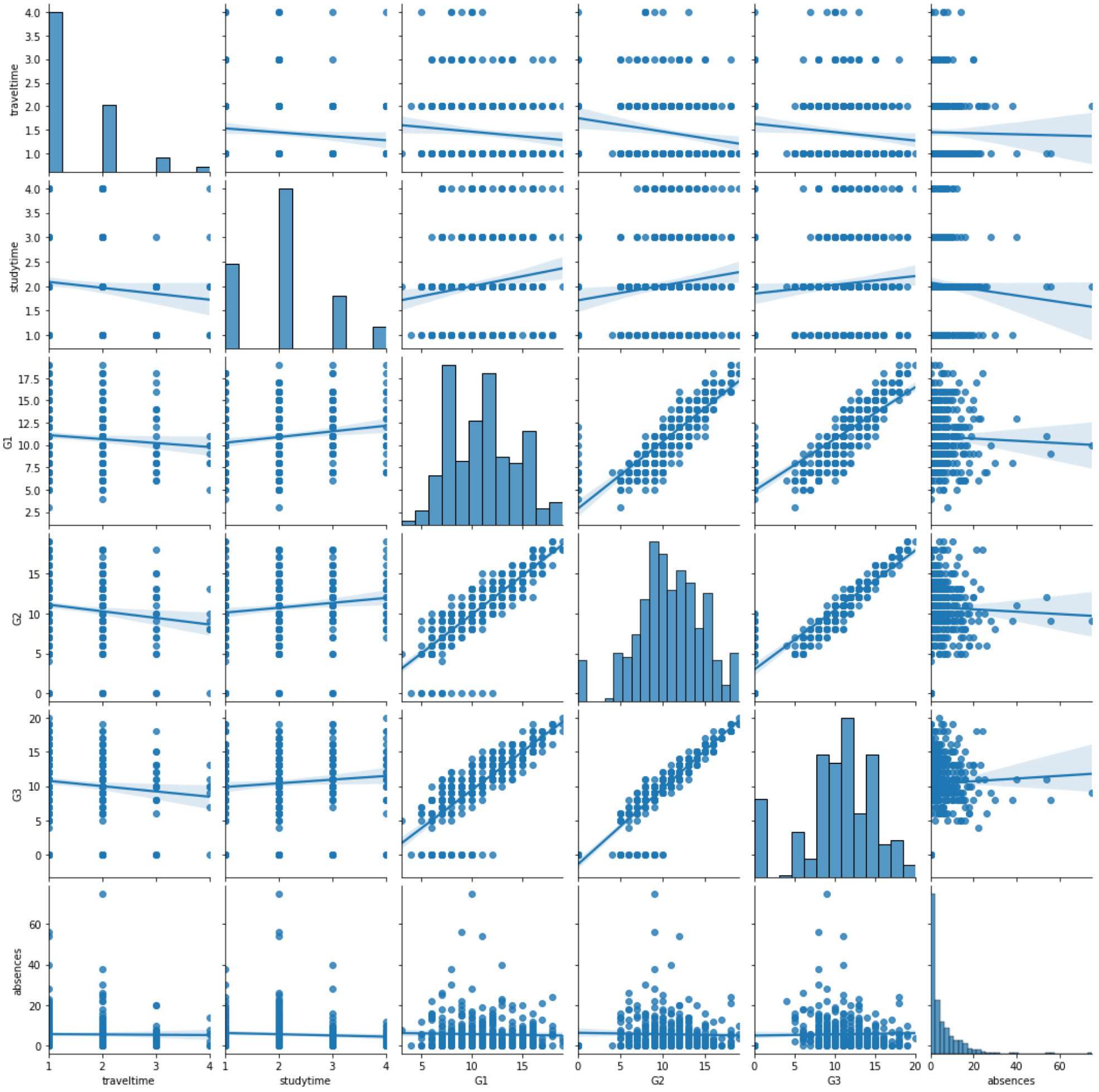
travel time could be a larger factor in student performance than time spent studying. The only time this appears to not have been the case is for the first grade report. Absences are also not very strongly correlated with any of the other performance characteristics, which is surprising.

In [361]:

```
# Matplotlib, a plotting library for Python, was imported for visualization purposes

import matplotlib.pyplot as plt

# Seaborn, a matplotlib enhancement, was imported to aid in visualization

import seaborn as sns
```

In [362]:

```
# A pairplot was drawn using seaborn scatterplots
# The kind="reg" parameter was used to add linear regression models to the scatter plots,
# the visualization of any correlations
g = sns.pairplot(df2, kind="reg")
```

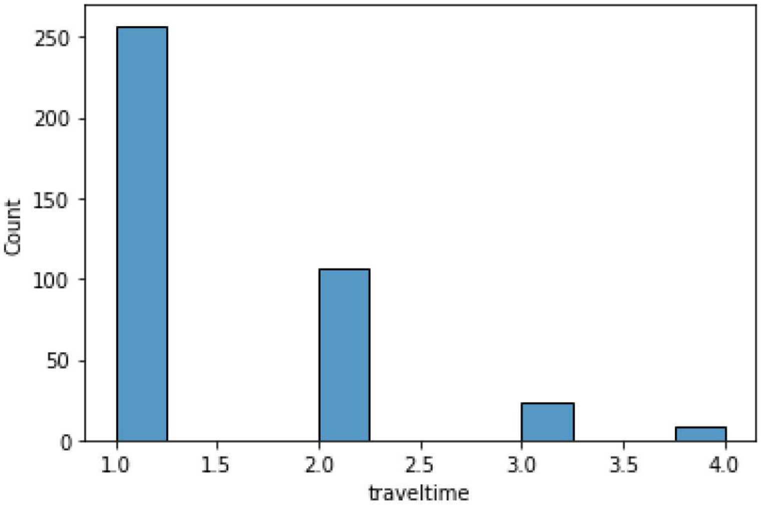


In [363]:

```
# Save pairplot for use in our project report
```

```
g.savefig("math_output.png")
```

In [364]:
```
# The axes labels of a seaborn pairplot, shown above, correspond to the relational scatte
# not the histograms that occur on the diagonal of the pairplot grid.  However, as we can
# plot below, the heights of the histogram bars correspond to the relative frequency of e
# above.

h = sns.histplot(df2['traveltime']);
```



Unfortunately, both the pairplot and the individual traveltime histogram show very few students have significant travel time, so it would be hard to draw firm conclusions about the effects of long travel times using the existing data set.

Now let's see if the same trend exists for the Portuguese grades for the students.

In [365]:
```
# Read csv file, noting that the data is separated with semicolons rather than the defaul

df1 = pd.read_csv("student(demo+grades)\student-por.csv", sep=';')
```

Attribute Information:

Source:
http://web.archive.org/web/20210209231731/https://archive.ics.uci.edu/ml/datasets/Student+Perforn

# #

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

2 sex - student's sex (binary: 'F' - female or 'M' - male)

3 age - student's age (numeric: from 15 to 22)

4 address - student's home address type (binary: 'U' - urban or 'R' - rural)

5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)

8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)

9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')

13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)

16 schoolsup - extra educational support (binary: yes or no)

17 famsup - family educational support (binary: yes or no)

18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

19 activities - extra-curricular activities (binary: yes or no)

20 nursery - attended nursery school (binary: yes or no)

21 higher - wants to take higher education (binary: yes or no)

22 internet - Internet access at home (binary: yes or no)

23 romantic - with a romantic relationship (binary: yes or no)

24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)

26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)

27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

29 health - current health status (numeric: from 1 - very bad to 5 - very good)

30 absences - number of school absences (numeric: from 0 to 93)

# #

these grades are related with the course subject, Math or Portuguese:

31 G1 - first period grade (numeric: from 0 to 20)

31 G2 - second period grade (numeric: from 0 to 20)

32 G3 - final grade (numeric: from 0 to 20, output target)

In [367]:

```
#Review data set in tabular form to confirm it has been processed correctly

df1.head()
```

Out[367]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	travelt
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	travelt
1	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	
2	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	
3	GP	F	15	U	GT3	T	4	2	health	services	home	mother	
4	GP	F	16	U	GT3	T	3	3	other	other	home	father	

In [368]:

```
# Limit data set to values that seem relevant to travel time and performance

df1a = df1[['reason', 'traveltime', 'studytime', 'activities', 'G1', 'G2', 'G3', 'absence
```

In [369]:

```
df1a.head()
```

Out[369]:

	reason	traveltime	studytime	activities	G1	G2	G3	absences
0	course	2	2	no	0	11	11	4
1	course	1	2	no	9	11	11	2
2	other	1	2	no	12	13	12	6
3	home	1	3	yes	14	14	14	0
4	home	1	2	no	11	13	13	0

In [370]:

```
# Review descriptive statistics of values to begin to analyze data

df1a.describe()
```

Out[370]:

	traveltime	studytime	G1	G2	G3	absences
count	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000
mean	1.568567	1.930663	11.399076	11.570108	11.906009	3.659476
std	0.748660	0.829510	2.745265	2.913639	3.230656	4.640759
min	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	1.000000	10.000000	10.000000	10.000000	0.000000
50%	1.000000	2.000000	11.000000	11.000000	12.000000	2.000000
75%	2.000000	2.000000	13.000000	13.000000	14.000000	6.000000
max	4.000000	4.000000	19.000000	19.000000	19.000000	32.000000

Unfortunately, again, we are able to see from this data set that few students traveled very far to get to the school they attended.

We can also see that the 50%-ile student scores for Portuguese and Math were fairly similar across the data sets.

Interestingly, we have 649 respondents for the Portuguese grade data set, whereas we only had 395 for the Math grade data set.

In [371]:

```
# Confirm trend away from long travel times to school.  As we can see from the below outp
# minimal travel time.

df1a.mode()
```

Out[371]:

	reason	traveltime	studytime	activities	G1	G2	G3	absences
0	course	1	2	no	10	11	11	0

In [372]:

```
# Determine if there are any clear correlations between travel time, study time, absences

df1a.corr()
```

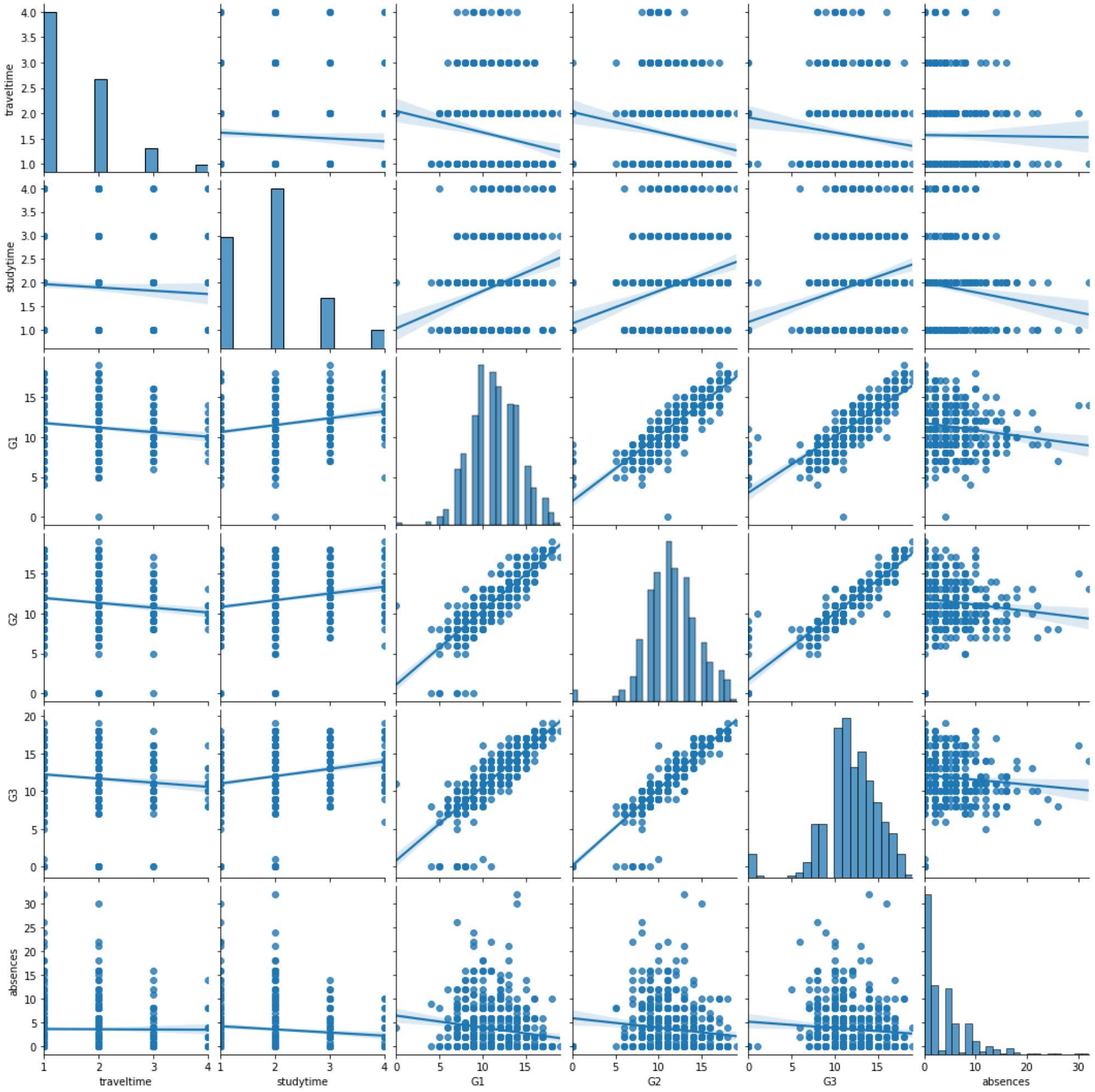
Out[372]:

	traveltime	studytime	G1	G2	G3	absences
traveltime	1.000000	-0.063154	-0.154120	-0.154489	-0.127173	-0.008149
studytime	-0.063154	1.000000	0.260875	0.240498	0.249789	-0.118389
G1	-0.154120	0.260875	1.000000	0.864982	0.826387	-0.147149
G2	-0.154489	0.240498	0.864982	1.000000	0.918548	-0.124745
G3	-0.127173	0.249789	0.826387	0.918548	1.000000	-0.091379
absences	-0.008149	-0.118389	-0.147149	-0.124745	-0.091379	1.000000

Again, we see a negative correlation between travel time and student grades at all three measured performance checkpoints. However, when compared to Math performance, we see a notably stronger positive correlation between study time and Portuguese grades. In this case, the magnitudes of the study time and grades correlations are larger than the magnitudes of the corresponding travel time and grade correlations. Here, absences are more strongly negatively correlated with Portuguese course grades, unlike absences and math grades in the previous analysis. However, absences and commute time are not significantly correlated, aside from a very slight negative correlation, which is consistent across both data sets.

In [373]:

```
# A pairplot was drawn using seaborn scatterplots
# The kind="reg" parameter was used to add linear regression models to the scatter plots,
# the visualization of any correlations
h = sns.pairplot(df1a, kind="reg")
```



In [374]:

```
# Save pairplot for use in our project report

h.savefig("por_output.png")
```

As we can see from the table above, travel time is negatively correlated with study time and course grades at all grading checkpoints.

However, unlike in the Math grade data, travel time does not have a stronger negative correlation with final two course grades than the positive correlation that study time has with these final course grades. For these data sets, it appears that time spent studying for Portuguese had a stronger positive correlation with course grades than time spent studying for Math grades.

We will discuss our findings, and the limitations of this project, further in the associated presentation.

References:

Data set source (linked above in the notebook and which includes the attribute information): https://archive.ics.uci.edu/ml/datasets/Student+Performance