Linear Regression

Case Study

Najib Mozahem

May 1, 2019

The Dataset

- gpa: overall GPA of the student (this is the dependent variable)
- english: the average grade on all English courses taken by the student (data is taken from a non-English speaking country where the language of instruction in university is English)
- college: whether the student is in the engineering school or the business school (zero means business, one means engineering)
- credits: the total number of credits completed so far by the student
- gender: whether the student is a male or a female (zero means female, one means male)
- attendance: attendance and participation grade last semester
- siblings: Number of brothers and sisters that the student has
- income: family income per year ($)
- work: records whether the student works full time, part time, or whether the student doesnt work at all.

Continuous Variables

There are two types of independent variables in our dataset, continuous and binary. We start by looking at the continuous variables.

Continuous Variables - Attendance (Scatter plot)

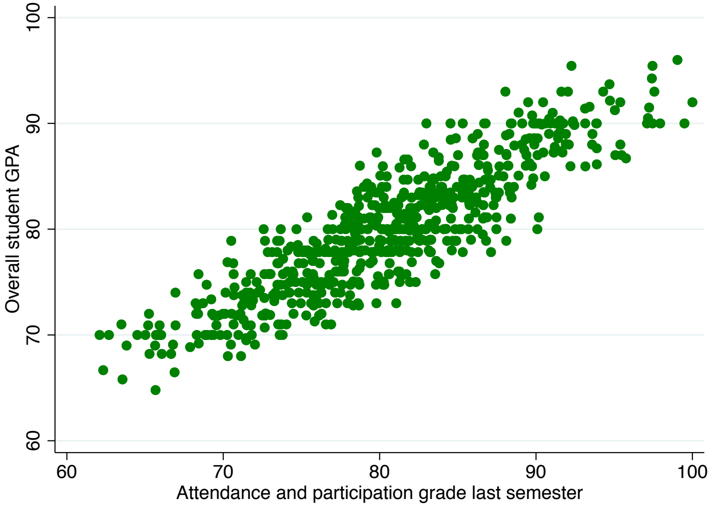

Figure: Scatter plot of GPA and attendance.

Continuous Variables - Attendance (Regression)

Source	SS	df	MS			
Model	18376.0567	1	18376.0567	Number of obs	=	666
Residual	5292.10793	664	7.97004207	F(1, 664)	=	2305.64
Total	23668.1646	665	35.591225	Prob > F	=	0.0000
				R-squared	=	0.7764
				Adj R-squared	=	0.7761
				Root MSE	=	2.8231

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	.7406767	.0154253	48.02	0.000	.7103885	.7709649
_cons	20.27157	1.246345	16.26	0.000	17.82432	22.71882

Continuous Variables - English (Scatter plot)

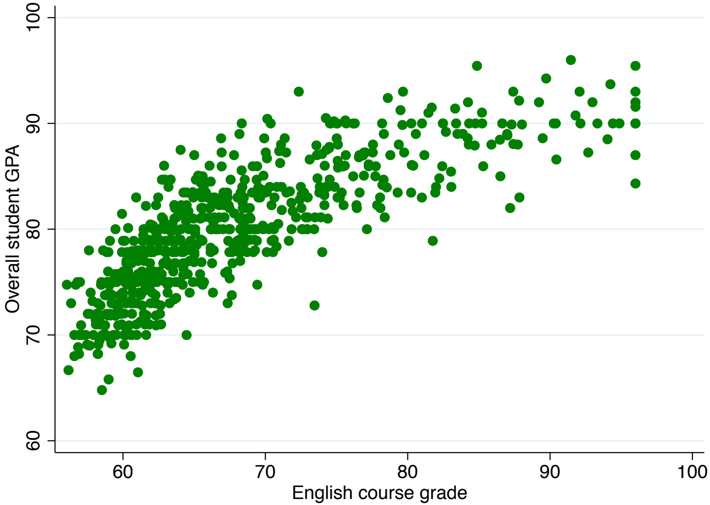

Figure: Scatter plot of GPA and english.

Continuous Variables - English (Checking for nonlinearity)

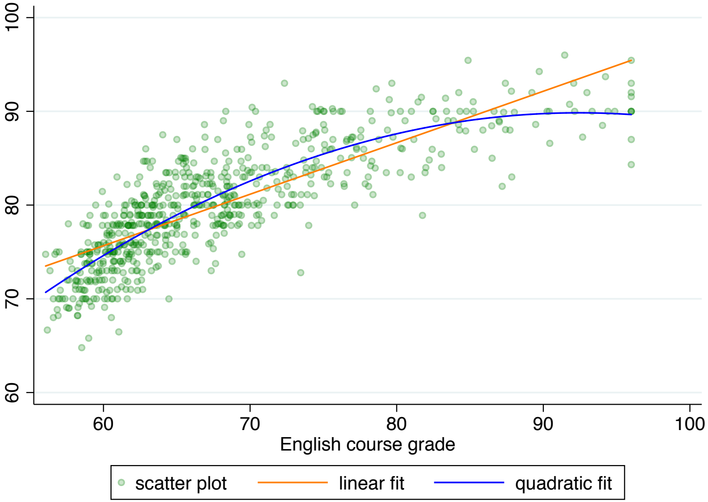

Figure: Scatter plot of GPA and english.

Continuous Variables - English (Regression)

Source	SS	df	MS
Model	17101.4609	2	8550.73045
Residual	7119.02181	674	10.5623469
Total	24220.4827	676	35.8291164

Number of obs	=	677
F(2, 674)	=	809.55
Prob > F	=	0.0000
R-squared	=	0.7061
Adj R-squared	=	0.7052
Root MSE	=	3.25

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
english	2.674998	.1910748	14.00	0.000	2.299824	3.050171
english2	-.0144694	.0012971	-11.16	0.000	-.0170162	-.0119227
_cons	-33.79013	6.925419	-4.88	0.000	-47.38812	-20.19214

Continuous Variables - Income (Scatter plot)

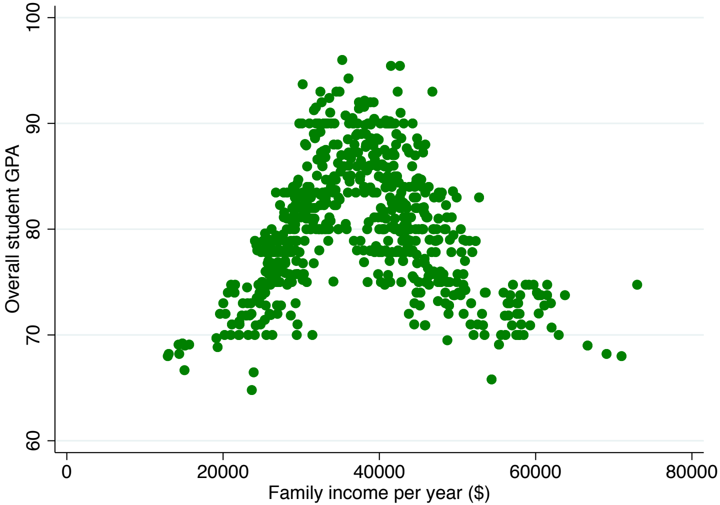

Figure: Scatter plot of GPA and income.

Continuous Variables - Income (Regression)

Source	SS	df	MS
Model	9410.93931	2	4705.46966
Residual	14809.5434	674	21.9726163
Total	24220.4827	676	35.8291164

Number of obs	=	677
F(2, 674)	=	214.15
Prob > F	=	0.0000
R-squared	=	0.3886
Adj R-squared	=	0.3867
Root MSE	=	4.6875

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.0021508	.0001057	20.35	0.000	.0019433	.0023584
income2	-2.74e-08	1.32e-09	-20.69	0.000	-3.00e-08	-2.48e-08
_cons	40.61942	2.032082	19.99	0.000	36.62944	44.60939

Continuous Variables - Credits (Scatter plot)

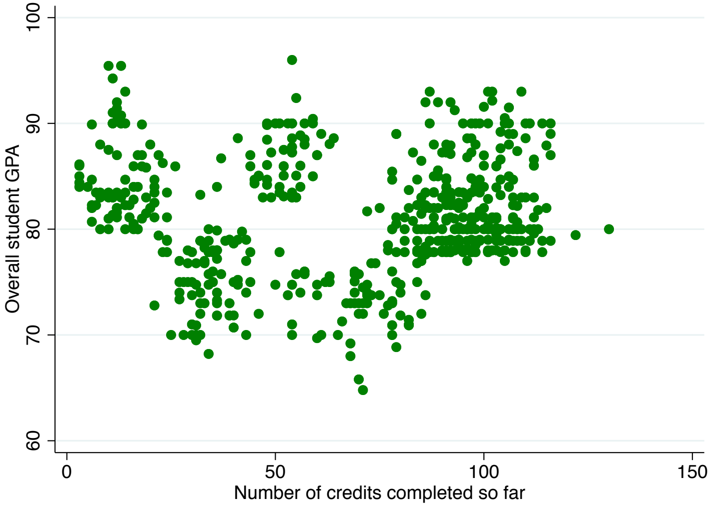

Figure: Scatter plot of GPA and credits.

Continuous Variables - Credits (Regression)

Source	SS	df	MS
Model	.02126165	1	.02126165
Residual	17994.8615	569	31.6254157
Total	17994.8828	570	31.5699698

Number of obs	=	571
F(1, 569)	=	0.00
Prob > F	=	0.9793
R-squared	=	0.0000
Adj R-squared	=	-0.0018
Root MSE	=	5.6236

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
credits	.0001823	.0070315	0.03	0.979	-.0136285	.0139931
_cons	81.13538	.5316944	152.60	0.000	80.09106	82.1797

Continuous Variables - Siblings (Scatter plot)

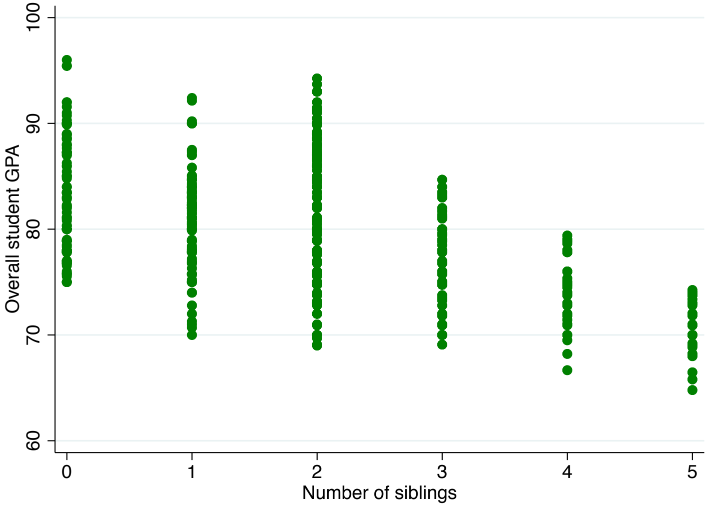

Figure: Scatter plot of GPA and siblings.

Continuous Variables - Siblings (Smoothing the scatter plot)

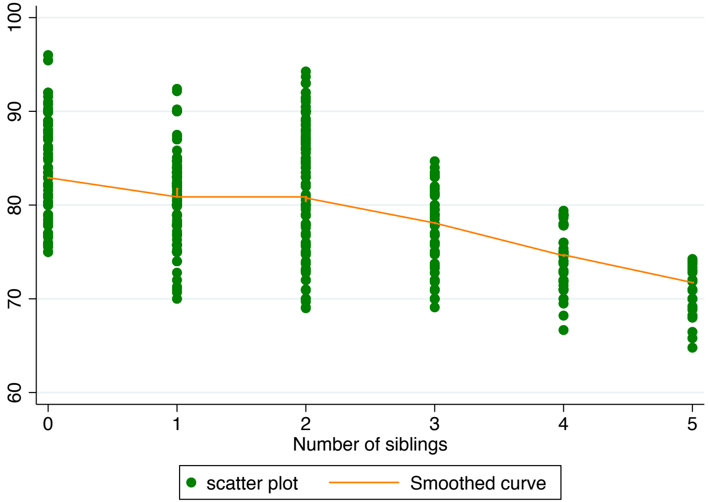

Figure: Scatter plot of GPA and siblings.

Continuous Variables - Siblings (Regression)

Source	SS	df	MS
Model	6072.64126	1	6072.64126
Residual	18147.8415	675	26.885691
Total	24220.4827	676	35.8291164

Number of obs	=	677
F(1, 675)	=	225.87
Prob > F	=	0.0000
R-squared	=	0.2507
Adj R-squared	=	0.2496
Root MSE	=	5.1851

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
siblings	-2.092668	.1392426	-15.03	0.000	-2.366069	-1.819267
_cons	83.96477	.3324845	252.54	0.000	83.31195	84.6176

Binary Variables

We now turn our attention towards the binary variables. Our dataset contains two binary variables, and they are college and gender.

Binary Variables - College (Regression)

Source	SS	df	MS
Model	361.527806	1	361.527806
Residual	23858.9549	675	35.3465999
Total	24220.4827	676	35.8291164

Number of obs	=	677
F(1, 675)	=	10.23
Prob > F	=	0.0014
R-squared	=	0.0149
Adj R-squared	=	0.0135
Root MSE	=	5.9453

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
college						
Engineering	1.470097	.4596731	3.20	0.001	.5675363	2.372658
_cons	79.1506	.3421136	231.36	0.000	78.47886	79.82233

Binary Variables - GPA (Regression)

Source	SS	df	MS
Model	1261.55583	1	1261.55583
Residual	22406.6088	664	33.7448927
Total	23668.1646	665	35.591225

Number of obs	=	666
F(1, 664)	=	37.39
Prob > F	=	0.0000
R-squared	=	0.0533
Adj R-squared	=	0.0519
Root MSE	=	5.809

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
gender						
Female	2.810086	.4595897	6.11	0.000	1.907661	3.71251
_cons	78.76412	.2904518	271.18	0.000	78.19381	79.33444

Categorical Variables (more than two groups)

The dataset that we are using also contains the variable work. Unlike binary variables, this variable divides the observations into three groups: those that have a full time job, those that have a part time job, and those that have no job at all.

Categorical Variables - Work (Regression)

Source	SS	df	MS		
Model	5284.63957	2	2642.31979	Number of obs =	677
Residual	18935.8431	674	28.0947228	F(2, 674) =	94.05
				Prob > F =	0.0000
Total	24220.4827	676	35.8291164	R-squared =	0.2182
				Adj R-squared =	0.2159
				Root MSE =	5.3004

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
work						
Part time	4.98294	.4317714	11.54	0.000	4.135161	5.830718
Full time	-2.616398	.6943671	-3.77	0.000	-3.979781	-1.253016
_cons	78.17105	.2940158	265.87	0.000	77.59375	78.74834

Multiple Regression

From the previous section, it seems that we need a model that includes the variables attendance, english, the square of english, income, the square of income, siblings, college, gender, and work. We can now fit a multiple regression model that includes all of these variables.

Multiple Regression

Source	SS	df	MS
Model	20923.9788	10	2092.39788
Residual	2744.18585	655	4.18959671
Total	23668.1646	665	35.591225

Number of obs	=	666
F(10, 655)	=	499.43
Prob > F	=	0.0000
R-squared	=	0.8841
Adj R-squared	=	0.8823
Root MSE	=	2.0469

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	.4184037	.018439	22.69	0.000	.382197	.4546103
english	.7920346	.1375871	5.76	0.000	.5218697	1.0622
english2	-.0037094	.0009115	-4.07	0.000	-.0054992	-.0019196
income	.0003589	.0000607	5.92	0.000	.0002398	.000478
income2	-4.63e-09	7.61e-10	-6.09	0.000	-6.13e-09	-3.14e-09
siblings	-.2505416	.0662263	-3.78	0.000	-.3805832	-.1205001
college						
Engineering	.5750044	.1653606	3.48	0.001	.2503035	.8997053
gender						
Female	-.2993557	.1764917	-1.70	0.090	-.6459134	.047202
work						
Part time	.9500899	.1860619	5.11	0.000	.58474	1.31544
Full time	-.5795809	.2849956	-2.03	0.042	-1.139196	-.0199657
_cons	3.349104	4.754391	0.70	0.481	-5.986581	12.68479

Multiple Regression

- If you look at the output, you will notice something interesting, and that is that the variable gender is no longer significant.
- In our dataset, the average GPA for males is 78.76 and the average GPA for females is 81.57.
- The average attendance grade for males is 78.62 and that the average attendance grade for females is 83.29.
- Therefore, it seems that the difference in GPAs between males and females is due to females attending more.
- Given the above, we can go ahead and fit a model that does not include gender.

Multiple Regression

Source	SS	df	MS			
Model	20911.9257	9	2323.5473			
Residual	2756.23895	656	4.20158377			
Total	23668.1646	665	35.591225			

Number of obs	=	666
F(9, 656)	=	553.02
Prob > F	=	0.0000
R-squared	=	0.8835
Adj R-squared	=	0.8819
Root MSE	=	2.0498

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	.4100926	.0178014	23.04	0.000	.375138	.4450473
english	.8025509	.1376438	5.83	0.000	.5322754	1.072826
english2	-.0037703	.0009121	-4.13	0.000	-.0055613	-.0019794
income	.0003616	.0000607	5.95	0.000	.0002423	.0004808
income2	-4.66e-09	7.62e-10	-6.12	0.000	-6.16e-09	-3.17e-09
siblings	-.2452277	.0662468	-3.70	0.000	-.375309	-.1151464
college						
Engineering	.6364283	.1615772	3.94	0.000	.3191574	.9536991
work						
Part time	.9510869	.186327	5.10	0.000	.5852176	1.316956
Full time	-.5791741	.2854029	-2.03	0.043	-1.139587	-.0187607
_cons	3.370133	4.761171	0.71	0.479	-5.978839	12.71911

R-squared

We see that the value of R-squared is 0.88, which is high. This means that the model is explaining around 88% of the observed variability in the dependent variable.

Plotting predicted values against observed values

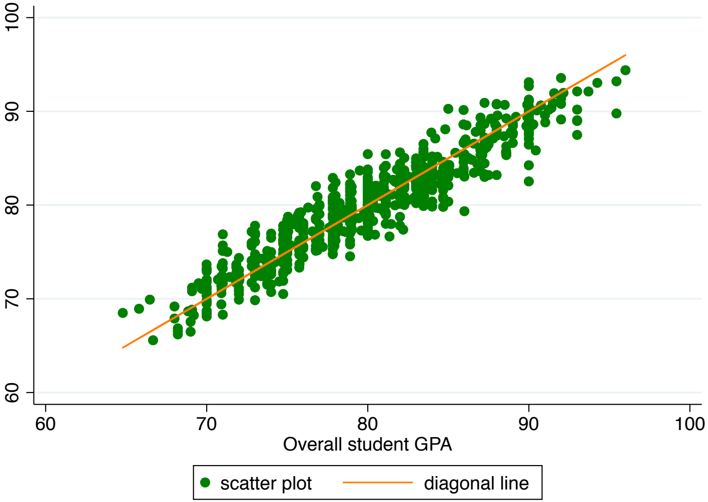

Figure: Comparing predicted values to observed values.

Normality of the Residuals - Histogram

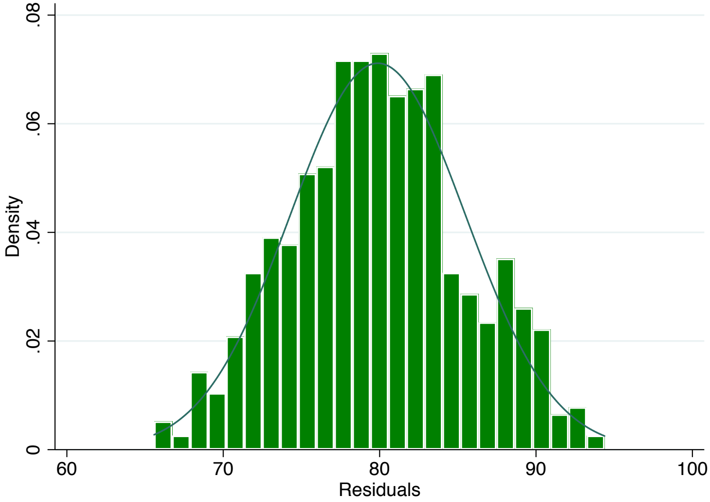

Figure: Comparing predicted values to observed values.

Normality of the Residuals - Quantile Normal Plots

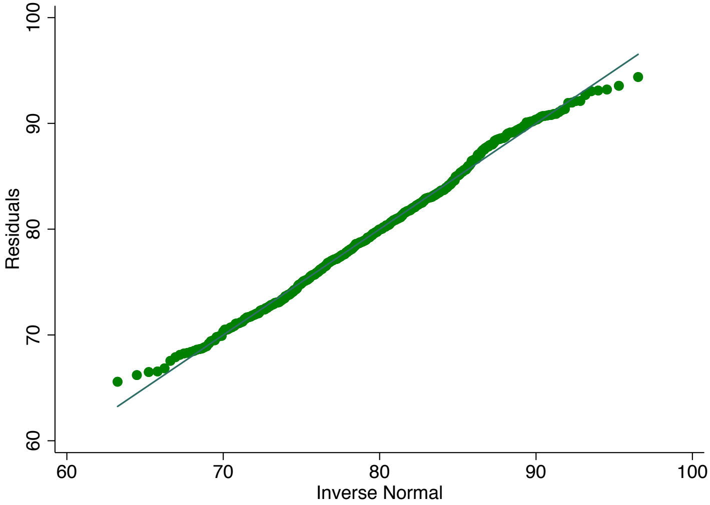

Figure: Comparing predicted values to observed values.

Normality of the Residuals - Skewness/Kurtosis Test

Skewness/Kurtosis tests for Normality

Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	joint adj chi2(2)	joint Prob>chi2
residuals	666	0.4072	0.0162	6.46	0.0396

Normality of the Residuals - Shapiro/Wilk Test

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
residuals	666	0.99441	2.438	2.170	0.01501

Homoscedasticity - Breusch/Pagan Test

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of gpa

         chi2(1)      =     5.75
         Prob > chi2  =   0.0165
```

Addressing Assumption Violations

- Given that we have rejected the assumptions of normality and homoscedasticity, does this mean that we disregard our regression results?
- Fortunately no. As mentioned in the theory part, what we can do in this case is to fit the model while telling the statistical software to use robust standard errors.
- This way, the assumptions are relaxed and we can have more faith in the resulting model.

Addressing Assumption Violations - Robust Standard Errors

```
Linear regression                               Number of obs     =        666
                                                F(8, 656)         =          .
                                                Prob > F          =          .
                                                R-squared         =     0.8835
                                                Root MSE          =     2.0498

------------------------------------------------------------------------------
             |               Robust
         gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  attendance |   .4100926   .0183261    22.38   0.000     .3741078    .4460775
     english |   .8025509   .1302303     6.16   0.000     .5468324    1.058269
    english2 |  -.0037703   .0008659    -4.35   0.000    -.0054707    -.00207
      income |   .0003616   .0000638     5.67   0.000     .0002363    .0004868
     income2 |  -4.66e-09   8.15e-10    -5.72   0.000    -6.27e-09   -3.06e-09
    siblings |  -.2452277   .0652367    -3.76   0.000    -.3733256   -.1171299
             |
     college |
 Engineering |   .6364283   .1615658     3.94   0.000     .3191797    .9536768
             |
        work |
   Part time |   .9510869   .1933017     4.92   0.000     .5715221    1.330652
   Full time |  -.5791741   .2641098    -2.19   0.029    -1.097777   -.0605715
             |
       _cons |   3.370133   4.477199     0.75   0.452    -5.421235     12.1615
------------------------------------------------------------------------------
```

Multicollinearity - VIF

Variable	VIF	1/VIF
attendance	2.53	0.395829
english	228.58	0.004375
english2	217.20	0.004604
income	59.98	0.016674
income2	60.35	0.016570
siblings	1.43	0.699116
1.college	1.02	0.978012
work		
1	1.34	0.748793
2	1.21	0.823429
Mean VIF	63.74	

Diagnostics

The next step is to investigate whether there are outliers and influential observations in the dataset.

Outliers

- In order to identify whether there are outliers, we can plot a scatter plot of two variables.
- The problem is that this method works when we just have one independent variable.
- However, in our model, there are several independent variables.
- Fortunately, there is a tool that allows us to work around this problem, and this tool is the added-variable plot.
- What these plots do is that they produce a scatter plot of the dependent variable against each independent variable while accounting for the presence of the other independent variables.

Outliers - Added Variable Plots

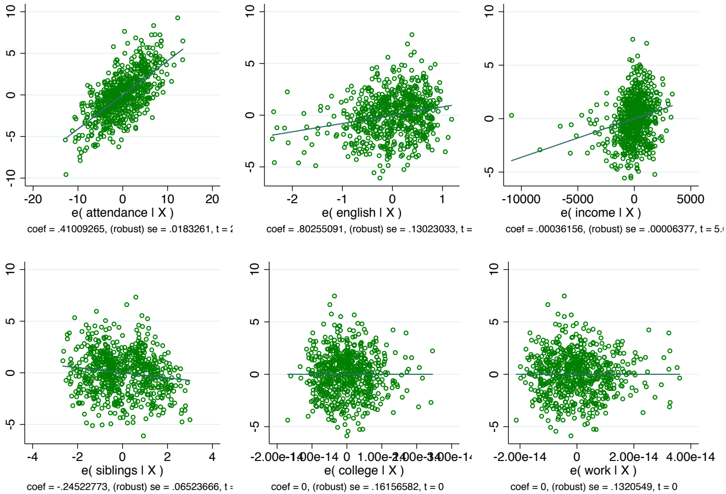

Figure: Added variable plots for each independent variable.

Influential Observations

- We next investigate whether there are any particularly influential observations in our dataset.
- We can do this by calculating the DFBETAS, DFFITS, and Cook's D statistic.
- A useful exercise would be to plot the DFFITS and Cook's D on the same plot.

Influential Observations - Plotting DFFITS against Cook's D

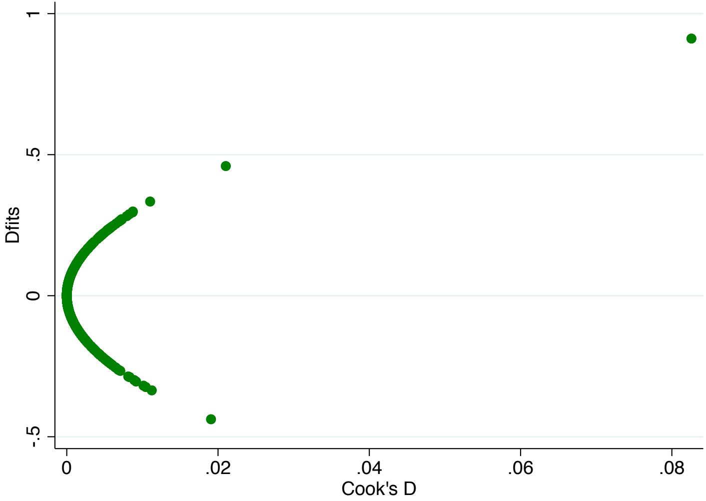

Figure: Plotting DFFITS and Cook's D.

Influential Observations

- Looking at the figure, we see that there is a single point that seems to be problematic since it has a higher than average values of both statistics.
- We note that this data point is the only one that has a Cook's D that is greater than 0.08.
- When discussing the outliers, we noted that there seems to be some outliers with respect to the independent variable income.
- It would be interesting to look at this graph again, but this time while we are paying attention to the value of Cook's D.

Influential Observations - Combining Findings

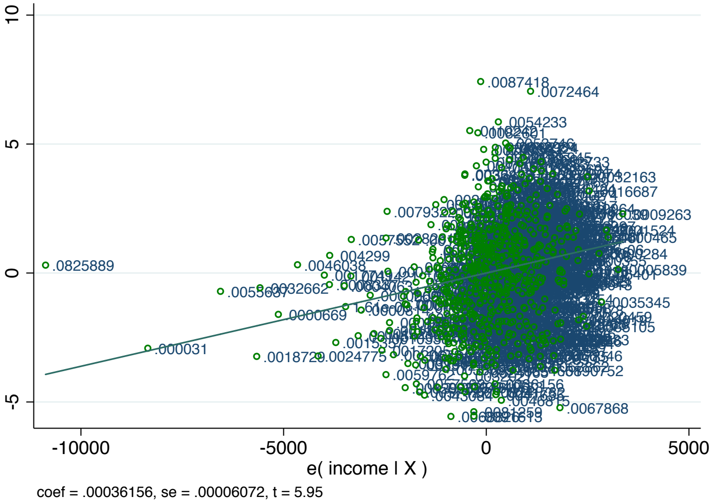

Figure: Added variable plot with Cook's D used as labels.

Influential Observations

- We now see that the outlier on the left hand side is actually the point that also has a Cook's D that is greater than 0.08.
- This means that this point is not only an outlier, but it is also influential.
- What do we do with it?
- The best thing to do about these points is to fit two models, one that includes all observations, and one that excludes these problematic observations.
- We can then compare the results.

Influential Observations - Comparing the Models

Table: Comparing estimates of both models

	(1)	(2)
Attendance and participation grade last semester	0.410***	0.406***
English course grade	0.803***	0.784***
english2	-0.00377***	-0.00364***
Family income per year - U.S. dollars	0.000362***	0.000409***
income2	-4.66e-09***	-5.29e-09***
Number of siblings	-0.245***	-0.239***
Business	0	0
Engineering	0.636***	0.617***
No	0	0
Part time	0.951***	0.938***
Full time	-0.579*	-0.571*
Constant	3.370	3.607
Observations	666	665

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Visualizing the Result

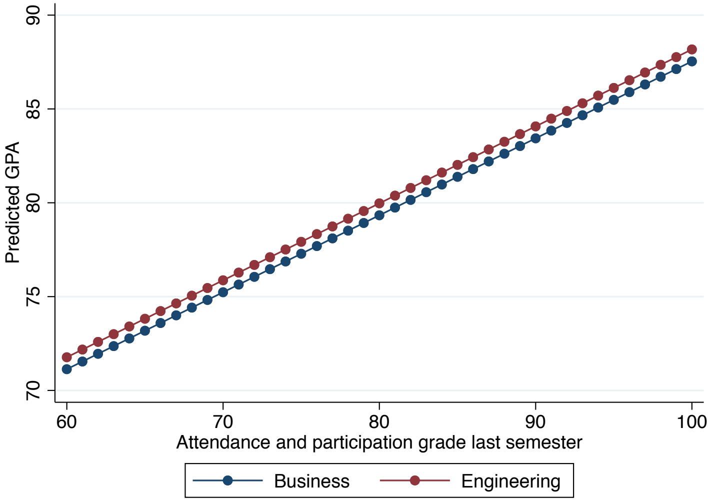

Figure: Visualizing how GPA varies with varying levels of the variables attendance and college.

Visualizing the Result

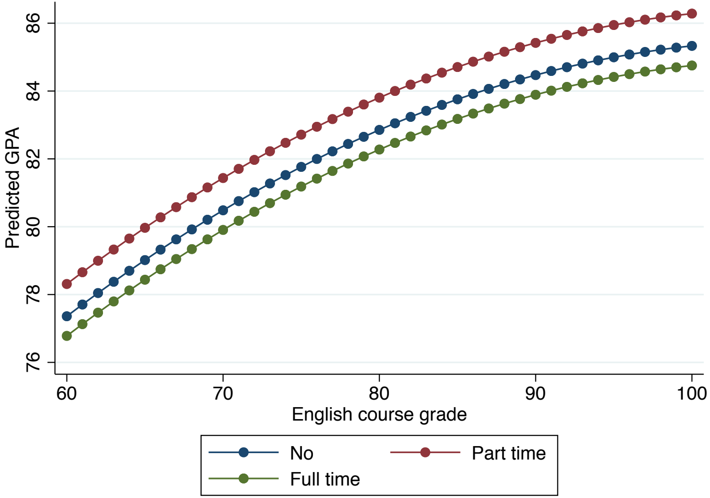

Figure: Visualizing how GPA varies with varying levels of the variables english and work.

Visualizing the Result

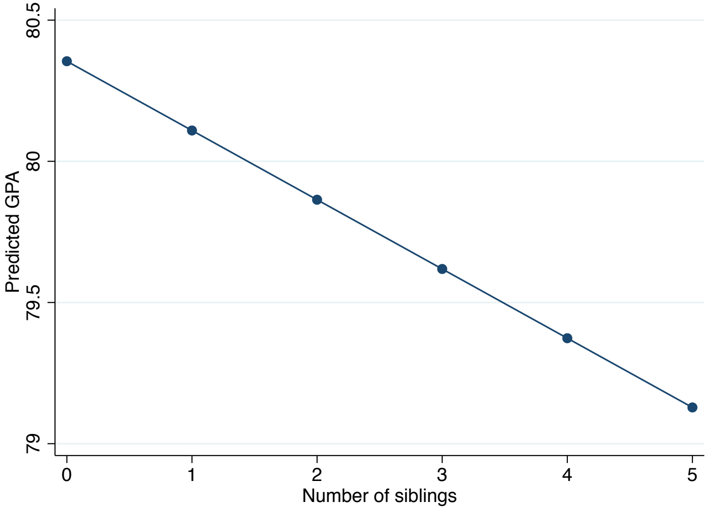

Figure: Visualizing how GPA varies with varying numbers of siblings.