STAT430
Project 2
Yingqiao Gou

1. **Do a write up on the regression of your two quantitative variables of interest.**
    a. Anova results from the PROC REG and what it means (p-value and significance)

    P-value < 0.001.

    P-value is less than 0.001 and we have set the confident interval to 95%, which means the differences between the variances of the means are statistically significant. In other words, this proof the rationality of using the linear regression.

    b. The r-squared value and what it means.

    R-squared is 0.4756.

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

    Usually, the larger the R^2, the better the regression model fits our observations. Therefore, this result seems to be not so good.

    c. Regression results and what it means (p-values on variables, significance, residuals) (You must give the following:

        i. The regression equation

        GPA = 0.01839 * Average Hour/Week Spent in Lib + 2.986878

        ii. The p-value for the independent variable(s) (5 10points)

The P-value of independent value < 0.001.

iii. Explain what the p-values mean (5 10points)

The P-value of independent value less than 0.001 and our CI is 95%, which means our independent value is significant only if p-value < 0.05. Therefore, the variable plays an important part in the linear regression model when predicting the GPA.

iv. The residual analysis, look at the normal probability plot and analyze (5 10points) and the residuals by regression for dependent variable (510points)

If look at the normal probability plot of residual, we could infer that the distribution of residuals is not following a normal distribution since its plot is asymmetric.

Similar result from the regression plot. Most of data points are not predicted probably. Therefore, this is model performs badly from the residuals perspective.

v. Make sure all assumptions are met. Explain what this does to your results. (5points 10points.)

The performance is not so good when this model fitting our dataset even if the p-value of the independent value and intercept is <0.001. This is because

a. The regression line fits model badly, most data points are not distributing along the line.
b. The distribution of residuals is not normal.
c. R-squared value is pretty small.

While the Root MSE is 0.23334, which is not bad.

2. **Do a write up on the regression of all your variables.**

   a. Anova results from the PROC REG and what it means (p-value and significance) (P-value stated, 5 points. Explanation 5 points)

      P-value < 0.0001.

      P-value is less than 0.0001 and we have set the confident interval to 95%, which means the differences between the variances of the means are statistically significant. In other words, this proof the rationality of using the linear regression.

   b. The r-squared value and what it means. (5)

      R-squared is 0.9192.

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

      Usually, the larger the R^2, the better the regression model fits our observations. Therefore, this result seems to great!

   c. Regression results and what it means (p-values on variables, significance, residuals)

(You must give the following:

   i. The regression equation (5)

      GPA = 0.01839 * Average Hour/Week Spent in Lib

      + 0.021758 * Average Hour/Weel Sp

      - 0.036087 * Acdemic Year

      + 0.597687 * What Major ? ART/ARCHITEKT/DESIGN

      + 0.391427 * What Major ? BUSINESS

      - 0.070511 * What Major ?CS

      - 0.023112 * What Major ?ENGINEERING

      + 0.077447 * What Major ?MATH/STAT

- 0.050688 * What Major ?OTHERS

+ 0.318699 * What Major ? PHYSICS/CHEM/BIO

+ 2.930754

ii. The p-value for the independent variable(s) (5)

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.930754	0.083595	35.06	<.0001
Average Hour/Week Sp	1	0.021758	0.002365	9.20	<.0001
Year	1	-0.036087	0.020272	-1.78	0.0883
What Major ? ART/ARCHITEKT/DESIGN	1	0.597687	0.098954	6.04	<.0001
What Major ? BUSINESS	1	0.391427	0.075535	5.18	<.0001
What Major ? CS	1	-0.070511	0.066582	-1.06	0.3006
What Major ? ENGINEERING	1	-0.023112	0.077447	-0.30	0.7681
What Major ? MATH/STAT	1	-0.050688	0.071347	-0.71	0.4846
What Major ? OTHERS	1	0.318699	0.125592	2.54	0.0184
What Major ? PHYSICS/CHEM/BIO	0	0	.	.	.

iii. Explain what the p-values mean (5)

The P-values of independent values **Average Hour/Weel Sp, What Major ? ART/ARCHITEKT/DESIGN, What Major ? BUSINESS and What Major ? OTHERS** are less than 0.05. However, the others are greater than 0.05. Since our CI is 95%, only the values with P-value <0.05 are significant to this regression model.

iv. The residual analysis, look at the normal probability plot and analyze (5) and the residuals by regression for dependent variable (5)

If look at the normal probability plot of residual, we could infer that the distribution of residuals is following a normal distribution since its plot is roughly symmetric like a bell-shaped curve.

Similar result from the regression plot. Most of data points are predicted probably. Therefore, this is model performs pretty well from the residuals perspective.

v. Make sure all assumptions are met. Explain what this does to your results.(5)

This model fit our data very well. The performance of this model on our dataset is much better than the simple linear regression even though the p-value of the some independent values are greater than 0,05. This is because

a. The regression line fits model very well, most data points are distributing along the line.

b. The distribution of residuals is roughly normal.

c. R-squared value is greater than 0.9 which is definitely a strong evidence to prove that it is a good model.

d. Root MSE is 0.10459, which means the error of this model is low, in another words, its accuracy is quite high.

Supporting Documentation

1. SAS code

```
%web_drop_table(WORK.pro1);
FILENAME REFFILE '/home/u50368724/GPA vs. Hours of Study in
Lib.csv.csv';
PROC IMPORT DATAFILE=REFFILE
     DBMS=CSV
     OUT=WORK.pro1;
     GETNAMES=YES;
RUN;
PROC CONTENTS DATA=WORK.pro1; RUN;
%web_open_table(WORK.pro1);
ods noproctitle;
ods graphics / imagemap=on;
proc reg data=WORK.PRO1 alpha=0.05 plots(only)=(diagnostics
residuals fitplot
          observedbypredicted);
     model 'Your GPA ?'n='Average Hour/Week Spent in Lib'n /;
     run;
quit;
data WORK.PRO1;
     set WORK.PRO1;

     select ('What Year Are You ?'n);
          when ('Freshman') Year=1;
          when ('Sophomore') Year=2;
          when ('Junior') Year=3;
          when ('Senior') Year=4;
          when ('Graduate') Year=5;
          otherwise Year='What Year Are You ?'n;
     end;
run;
ods noproctitle;
ods graphics / imagemap=on;
```

```
proc glmselect data=WORK.PRO1
outdesign(addinputvars)=Work.reg_design;
     class 'What Major ?'n / param=glm;
     model 'Your GPA ?'n='Average Hour/Week Spent in Lib'n
Year 'What Major ?'n /
          showpvalues selection=none;
run;
proc reg data=Work.reg_design alpha=0.05
plots(only)=(diagnostics residuals
          observedbypredicted);
     where 'What Major ?'n is not missing;
     ods select DiagnosticsPanel ResidualPlot
ObservedByPredicted;
     model 'Your GPA ?'n=&_GLSMOD /;
     run;
quit;
proc delete data=Work.reg_design;
run;
```

2. SAS Output Data

Obs	Timestamp	What Year Are You ?	What Major ?	Your GPA ?	Average Hour/Week Spent in Lib	Year
1	2021/06/23 6:04:35 PM GMT+8	Freshman	CS	3	10	1
2	2021/06/23 6:05:01 PM GMT+8	Freshman	BUSINESS	3.5	12	1
3	2021/06/23 6:05:26 PM GMT+8	Sophomore	PHYSICS/CHEM/BIO	3.3	20	2
4	2021/06/23 6:05:56 PM GMT+8	Sophomore	CS	3.4	26	2
5	2021/06/23 6:06:27 PM GMT+8	Freshman	ART/ARCHITEKT/DESIGN	3.6	5	1
6	2021/06/23 6:07:02 PM GMT+8	Freshman	MATH/STAT	3.4	15	1
7	2021/06/23 6:07:41 PM GMT+8	Junior	CS	3.6	35	3
8	2021/06/23 6:08:03 PM GMT+8	Junior	PHYSICS/CHEM/BIO	3.7	40	3
9	2021/06/23 6:08:38 PM GMT+8	Sophomore	MATH/STAT	2.8	7	2
10	2021/06/23 6:08:51 PM GMT+8	Sophomore	CS	2.9	9	2
11	2021/06/23 6:10:25 PM GMT+8	Senior	BUSINESS	3.8	28	4
12	2021/06/23 6:10:42 PM GMT+8	Junior	ART/ARCHITEKT/DESIGN	3.9	22	3
13	2021/06/23 6:11:51 PM GMT+8	Graduate	BUSINESS	4	42	5
14	2021/06/23 6:12:21 PM GMT+8	Senior	CS	3.2	28	4
15	2021/06/23 6:12:36 PM GMT+8	Senior	CS	3.1	18	4
16	2021/06/23 6:12:51 PM GMT+8	Senior	ENGINEERING	3.2	24	4
17	2021/06/23 6:26:15 PM GMT+8	Senior	PHYSICS/CHEM/BIO	3.6	34	4
18	2021/06/23 6:26:30 PM GMT+8	Senior	MATH/STAT	3.5	38	4
19	2021/06/23 6:26:41 PM GMT+8	Senior	CS	3.6	40	4
20	2021/06/23 6:26:53 PM GMT+8	Graduate	CS	3.6	46	5
21	2021/06/23 6:27:41 PM GMT+8	Junior	BUSINESS	4	32	3
22	2021/06/23 6:27:59 PM GMT+8	Freshman	ENGINEERING	2.9	6	1
23	2021/06/23 6:28:33 PM GMT+8	Graduate	OTHERS	3.7	29	5
24	2021/06/23 6:28:56 PM GMT+8	Senior	CS	3.3	18	4
25	2021/06/23 6:29:08 PM GMT+8	Senior	MATH/STAT	3.2	16	4
26	2021/06/23 6:29:30 PM GMT+8	Freshman	MATH/STAT	3.1	16	1
27	2021/06/23 6:30:16 PM GMT+8	Sophomore	MATH/STAT	3.3	25	2
28	2021/06/23 6:30:28 PM GMT+8	Sophomore	ENGINEERING	3.6	32	2
29	2021/06/23 6:30:47 PM GMT+8	Senior	PHYSICS/CHEM/BIO	3.7	46	4
30	2021/06/23 6:30:58 PM GMT+8	Senior	ENGINEERING	3.1	16	4
31	2021/06/23 6:31:23 PM GMT+8	Junior	ENGINEERING	3.3	17	3
32	2021/06/23 6:31:47 PM GMT+8	Freshman	MATH/STAT	3.3	20	1

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
5	Average Hour/Week Spent in Lib	Num	8	BEST12.	BEST32.
1	Timestamp	Char	29	$29.	$29.
3	What Major ?	Char	22	$22.	$22.
2	What Year Are You ?	Char	11	$11.	$11.
4	Your GPA ?	Num	8	BEST12.	BEST32.

Simple Linear Regression Model

Model: MODEL1
Dependent Variable: Your GPA ?

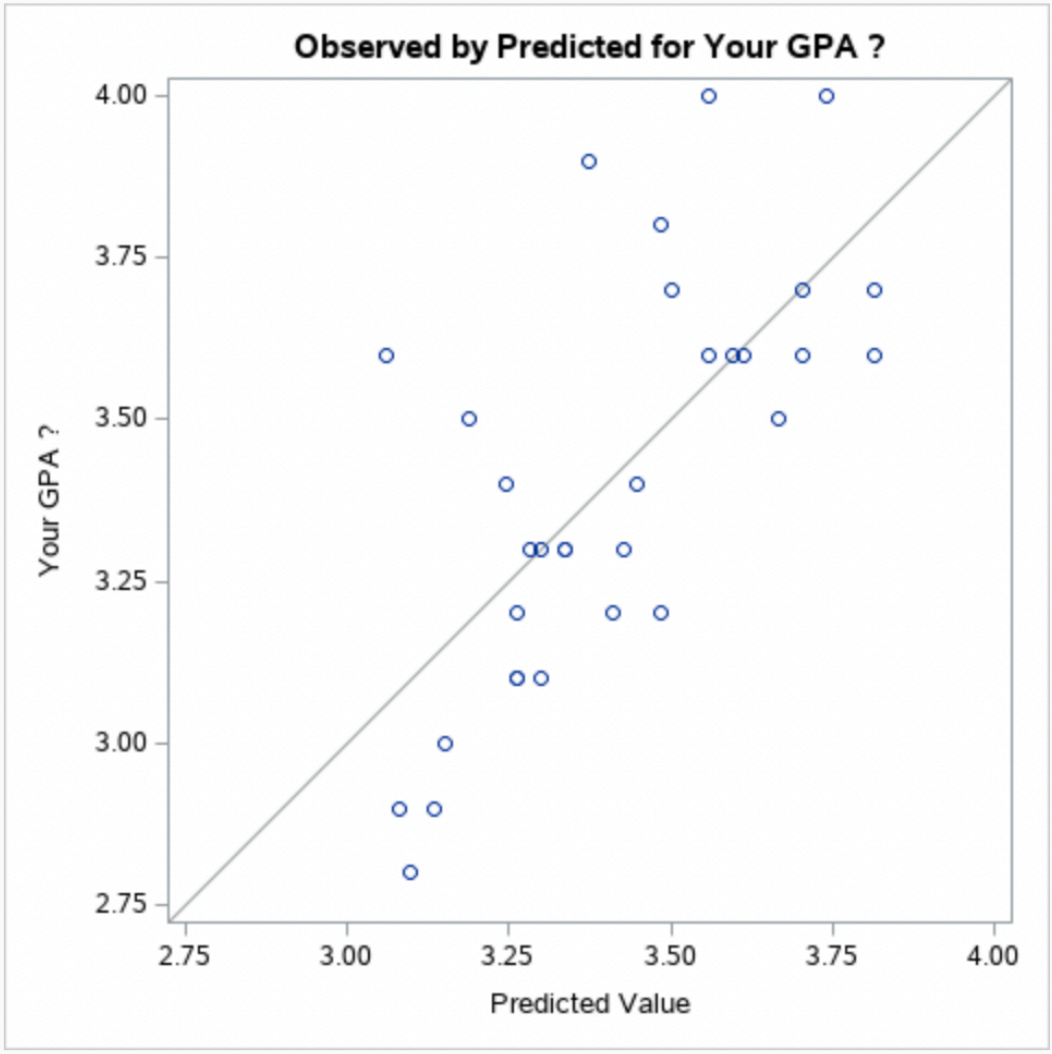
Number of Observations Read	32
Number of Observations Used	32

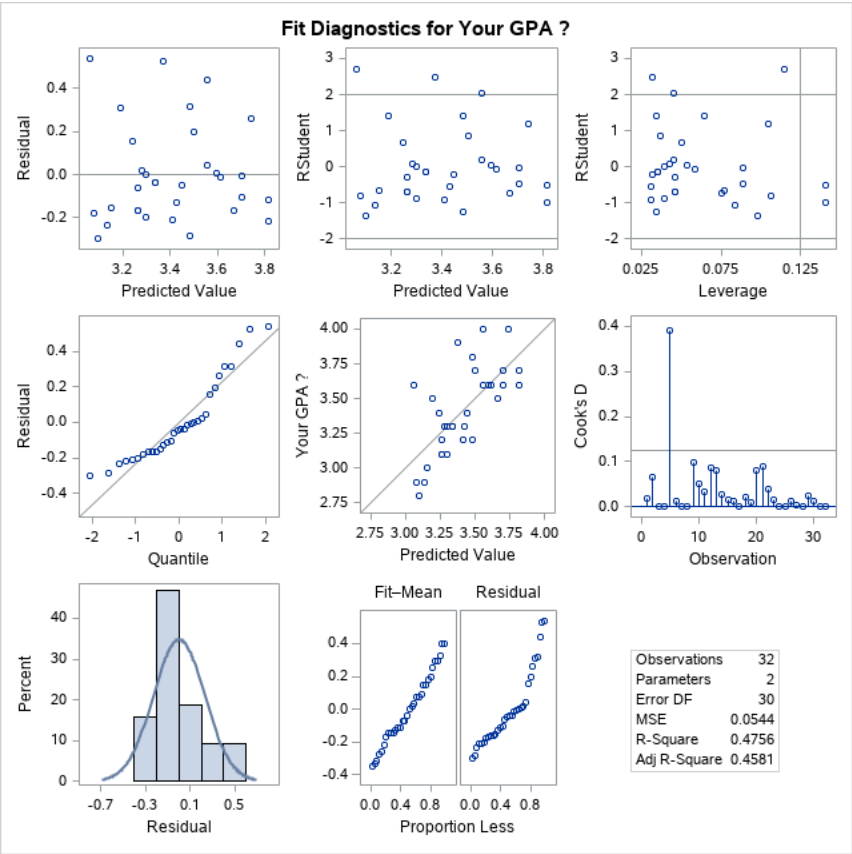
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.48152	1.48152	27.21	<.0001
Error	30	1.63348	0.05445		
Corrected Total	31	3.11500			

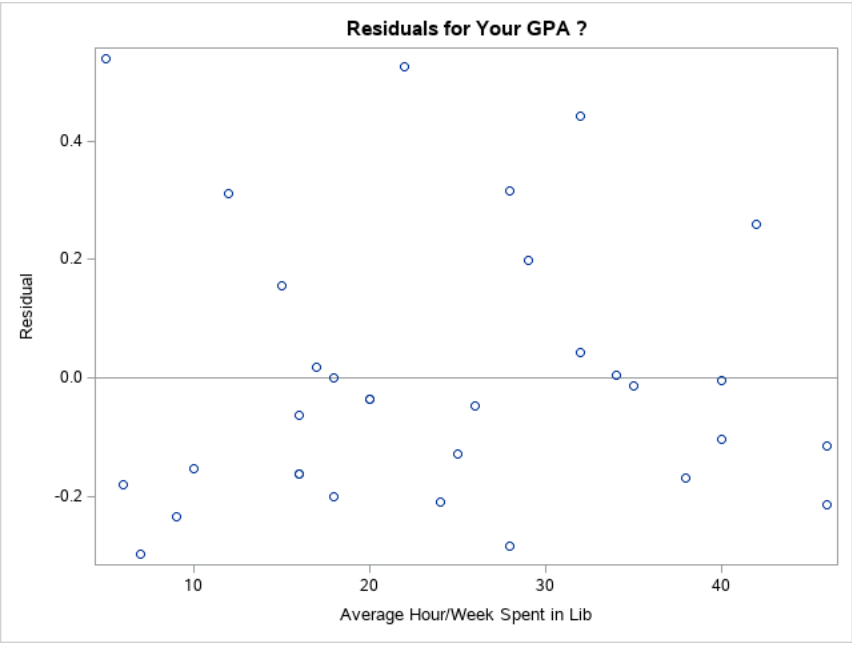
Root MSE	0.23334	R-Square	0.4756
Dependent Mean	3.41250	Adj R-Sq	0.4581
Coeff Var	6.83792		

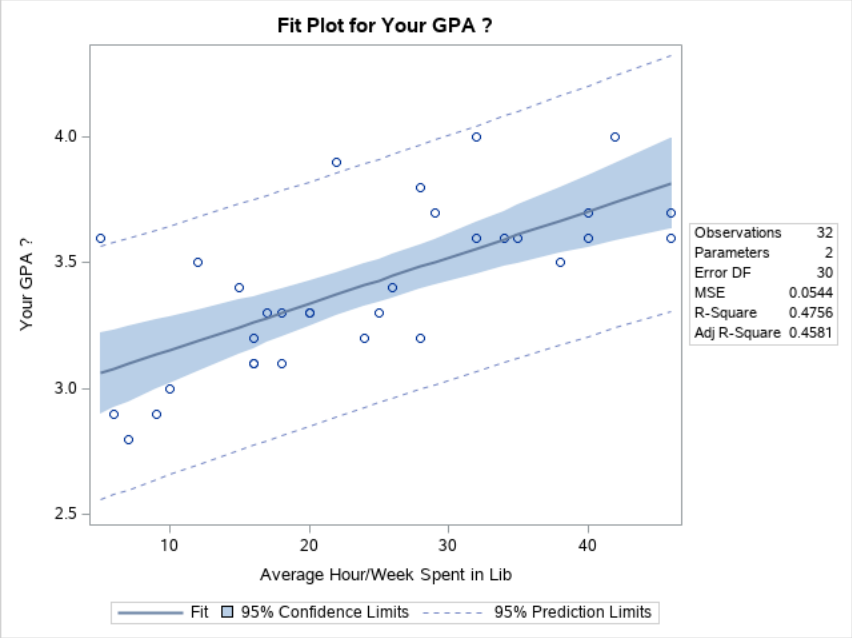
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.96878	0.09454	31.40	<.0001
Average Hour/Week Spent in Lib	1	0.01839	0.00353	5.22	<.0001

Model: MODEL1
Dependent Variable: Your GPA ?




Fit Diagnostics for Your GPA ?
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
Your GPA ?
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 32
Parameters 2
Error DF 30
MSE 0.0544
R-Square 0.4756
Adj R-Square 0.4581


Residuals for Your GPA ?
Residual
Average Hour/Week Spent in Lib



Multilinear Regression

Data Set	WORK.PRO1
Dependent Variable	Your GPA ?
Selection Method	None

Number of Observations Read	32
Number of Observations Used	32

Class Level Information		
Class	**Levels**	**Values**
What Major ?	7	ART/ARCHITEKT/DESIGN BUSINESS CS ENGINEERING MATH/STAT OTHERS PHYSICS/CHEM/BIO

Dimensions	
Number of Effects	4
Number of Parameters	10

Least Squares Summary

Step	Effect Entered	Number Effects In	Number Parms In	SBC
0	**Intercept**	1	1	-71.0785
1	**Average Hour/Week Sp**	2	2	-88.2692
2	**Year**	3	3	-85.1411
3	**What Major ?**	4	9	-123.8710*

*** Optimal Value of Criterion**

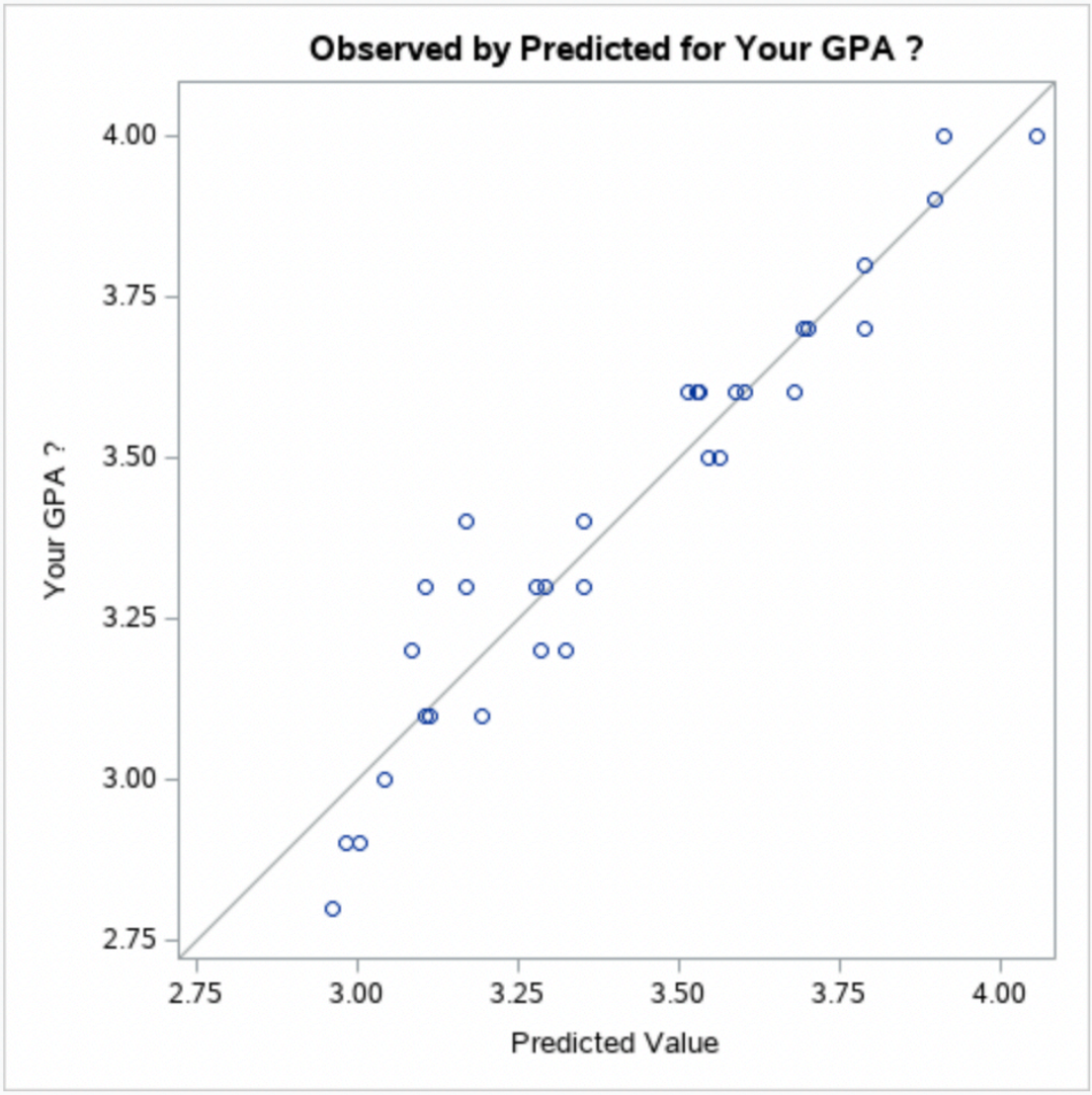
Least Squares Model (No Selection)

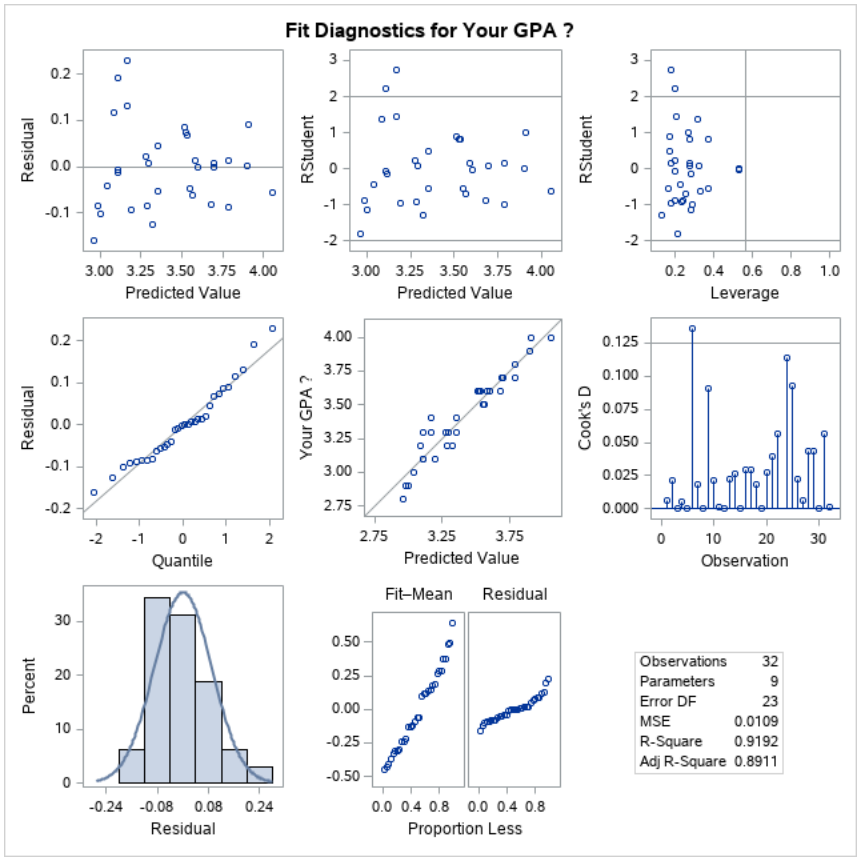
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	2.86341	0.35793	32.72	<.0001
Error	23	0.25159	0.01094		
Corrected Total	31	3.11500			

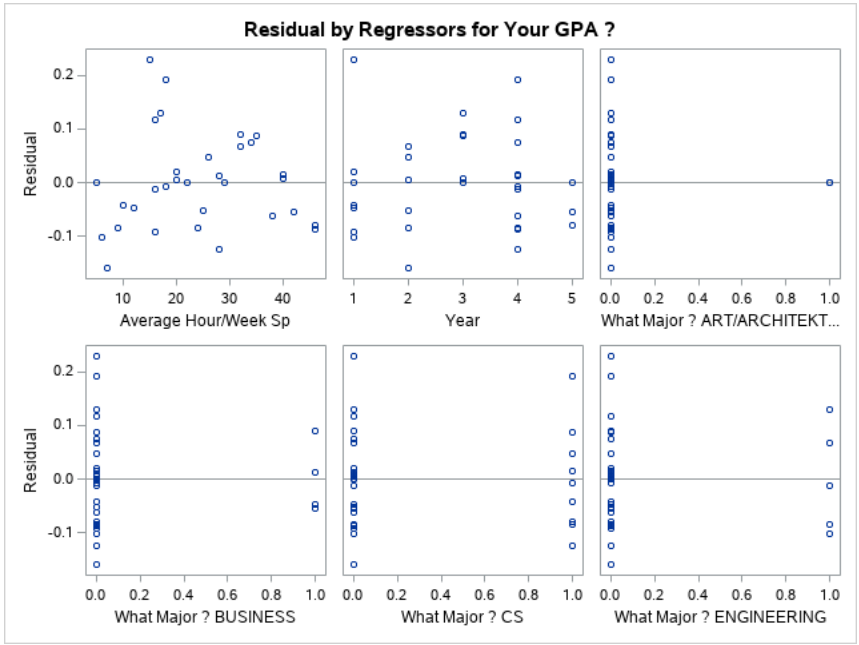
Root MSE	0.10459
Dependent Mean	3.41250
R-Square	0.9192
Adj R-Sq	0.8911
AIC	-103.06259
AICC	-92.58640
SBC	-123.87096

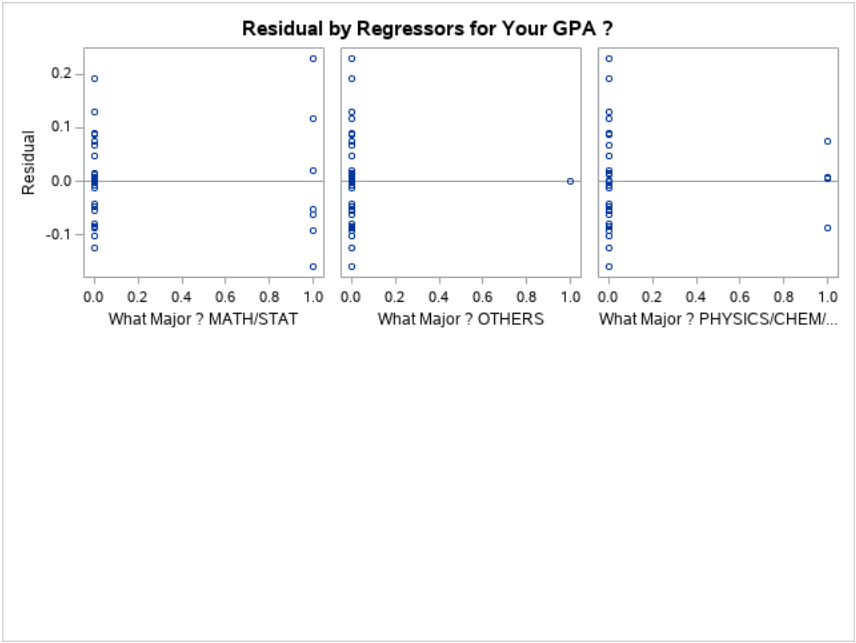
Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.930754	0.083595	35.06	<.0001
Average Hour/Week Sp	1	0.021758	0.002365	9.20	<.0001
Year	1	-0.036087	0.020272	-1.78	0.0883
What Major ? ART/ARCHITEKT/DESIGN	1	0.597687	0.098954	6.04	<.0001
What Major ? BUSINESS	1	0.391427	0.075535	5.18	<.0001
What Major ? CS	1	-0.070511	0.066582	-1.06	0.3006
What Major ? ENGINEERING	1	-0.023112	0.077447	-0.30	0.7681
What Major ? MATH/STAT	1	-0.050688	0.071347	-0.71	0.4846
What Major ? OTHERS	1	0.318699	0.125592	2.54	0.0184
What Major ? PHYSICS/CHEM/BIO	0	0	.	.	.

Model: MODEL1
Dependent Variable: Your GPA ?




Fit Diagnostics for Your GPA ?
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
Your GPA ?
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 32
Parameters 9
Error DF 23
MSE 0.0109
R-Square 0.9192
Adj R-Square 0.8911


Residual by Regressors for Your GPA ?
Residual
Average Hour/Week Sp
Year
What Major ? ART/ARCHITEKT...
Residual
What Major ? BUSINESS
What Major ? CS
What Major ? ENGINEERING


Residual by Regressors for Your GPA ?
Residual
0.2
0.1
0.0
-0.1
0.0 0.2 0.4 0.6 0.8 1.0
What Major ? MATH/STAT
0.0 0.2 0.4 0.6 0.8 1.0
What Major ? OTHERS
0.0 0.2 0.4 0.6 0.8 1.0
What Major ? PHYSICS/CHEM/...