Yizhan Ao

STAT430

Project2

April. 26, 2022

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)
        1. Our p value is **P-value < 0.001.** and we have set the confident interval to 95%, which means the differences between the variances of the means are stalslcally significant. In other words, this proof the ralonality of using the linear regression.
    b. The r-squared value and what it means. (10points)
        1. Our R-squared value is 0.0681

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

        We have the R^2 to be the value of variance explained by the model divided by the total variance. The r^2 value is low. But we could have guessed that by looking at the scatterplot. It seemed linear for smaller values of hours. Our R squared value is not very good is because of several outliers from CGPA of 2.7-3.3 from the probability plot listed.
    c. Regression results and what it means (p-values on variables, significance, residuals)
        1. Regression equation:
            1. **CGPA = -0.04200* hours spent in library + 3.76723**
                - P value for the independent variable
            2. The p-value for the independent variable(s) (10points)
                - We have the p value to be very small for the intercept so intercept is very significant and the value for each major and other independent variables are shown below so we could see the year is taking a very important place as well.

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	.	.	.

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

- The P-values of independent values Average Hour/ 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. Therefore the only significant variables are architectural and business to my model

4. The residual analysis, look at the normal probability plot and analyze (10points) and the residuals by regression for dependent variable (10points)
    - We are having a roughly symmetric bell shaped curve in our normal distribution. We could infer that the distribution of residuals is followed by a normal distribution
    - Similar result from the regression plot. Most of data points are predicted probably. Therefore, this is model performs pretty well from the residuals perspective. There are very few outliers on the plot
    - From the probability plot for CGPA we have the plot to be basically linear only a few outliers exist in the end points

5. Make sure all assumptions are met. Explain what this does to your results. (10points.)
    - The model fit our data well but the performance of independent variables are not well. Many independent variables are having p values grater than 0.05 which is caused by the data points distributed. The residual distribution is roughly normal, and the input of the R squared value to be very low. We have the data to be well performed by the MSE to be very low which means the error of the model is low. Therefore the accuracy is high

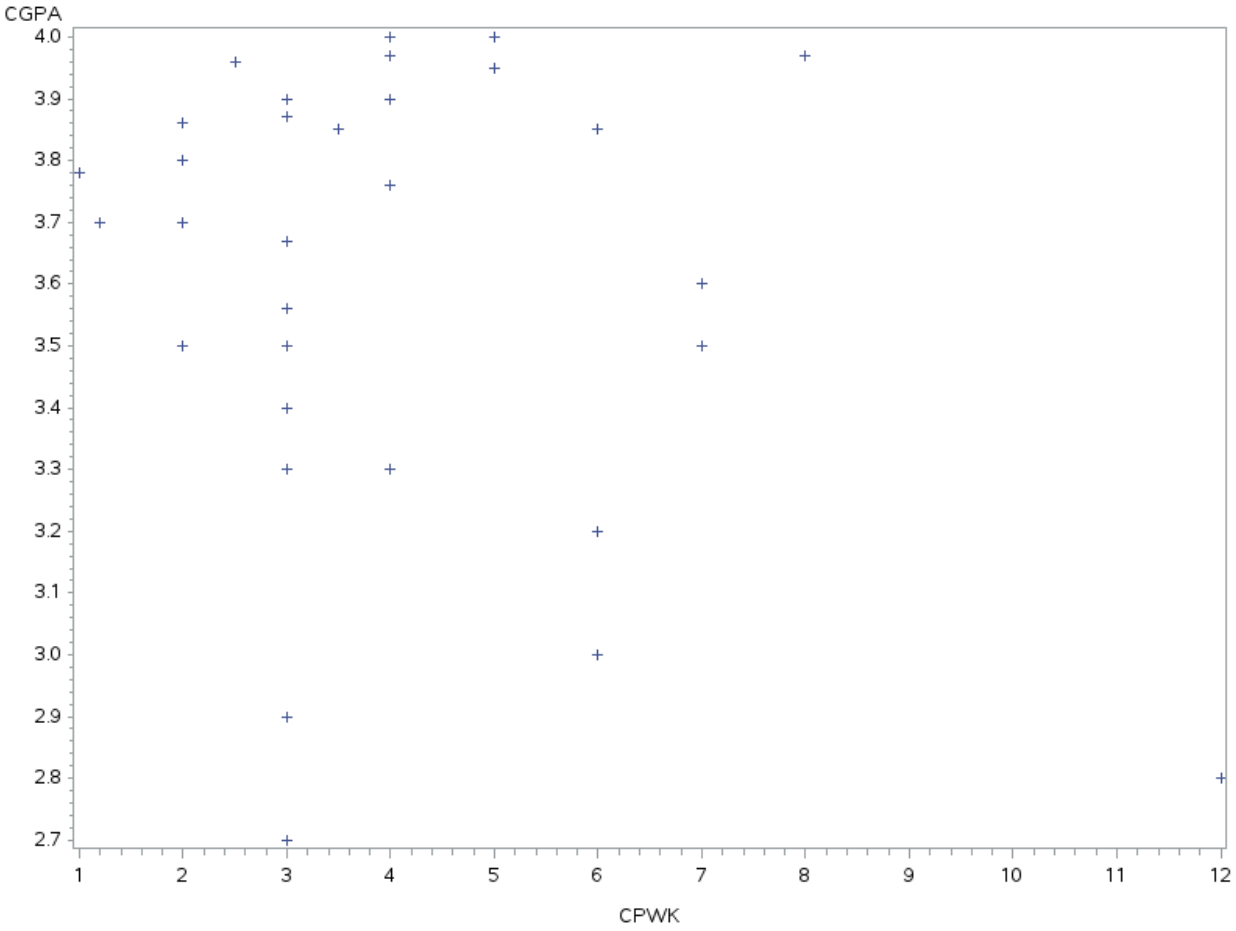
CODE:

```
DATA temp ;
INFILE '/home/u58594663/project1.csv' delimiter=',' dsd;
INPUT
    TIMESTAMP $
    YEAR $
    MAJOR $
    CGPA
    CPWK
    SMOKE $;
RUN;

PROC PRINT DATA = temp;
RUN;
ODS GRAPHICS ON;
PROC GPLOT DATA = temp;
PLOT CGPA*CPWK;
RUN;
data WORK.PRO1;
set WORK.PRO1;
select ('Which year of study are you? 'n);
when ('Freshman') Year=1;
when ('Sophomore') Year=2;
when ('Junior') Year=3;
when ('Senior') Year=4;
otherwise Year='Which year of study are you? 'n;
end;
run;
ods noproctitle;
ods graphics / imagemap=on;
proc glmselect data=WORK.PRO1
outdesign(addinputvars)=Work.reg_design;
class 'What is your major'n / param=glm;
model 'CPWK'n /
showpvalues selection=none;
run;
PROC CORR DATA = temp;
VAR CGPA CPWK;
RUN;
PROC REG DATA = temp PLOTS=DIAGNOSTICS(STATS=NONE);
MODEL CGPA = CPWK;
RUN;
```

Obs	TIMESTAMP	YEAR	MAJOR	CGPA	CPWK	SMOKE
1	2/16/202	Sophomor	CS	2.70	3.0	Yes
2	2/16/202	Sophomor	MATH	2.90	3.0	Yes
3	2/16/202	Senior	CS	3.86	2.0	No
4	2/16/202	Senior	MATH	3.90	4.0	No
5	2/16/202	Sophomor	STAT	3.40	3.0	Yes
6	2/16/202	Senior	ECON	3.56	3.0	Yes
7	2/16/202	Senior	CS	3.97	8.0	Yes
8	2/16/202	Senior	ECON	3.78	1.0	Yes
9	2/16/202	Senior	Media de	3.70	2.0	No
10	2/16/202	Senior	BIO	2.80	12.0	Yes
11	2/16/202	Freshman	FINA	3.50	7.0	No
12	2/16/202	Senior	INFOSCI	3.95	5.0	Yes
13	2/16/202	Freshman	ECON	4.00	4.0	No
14	2/16/202	Freshman	CS	3.67	3.0	No
15	2/16/202	Senior	PHS	3.70	2.0	Yes
16	2/16/202	Junior	CS	3.87	3.0	No
17	2/16/202	Freshman	FINA	3.96	2.5	No
18	2/16/202	Senior	CS/MATH	3.30	3.0	No
19	2/16/202	Junior	CS	3.20	6.0	Yes
20	2/16/202	Junior	ECON	3.00	6.0	No

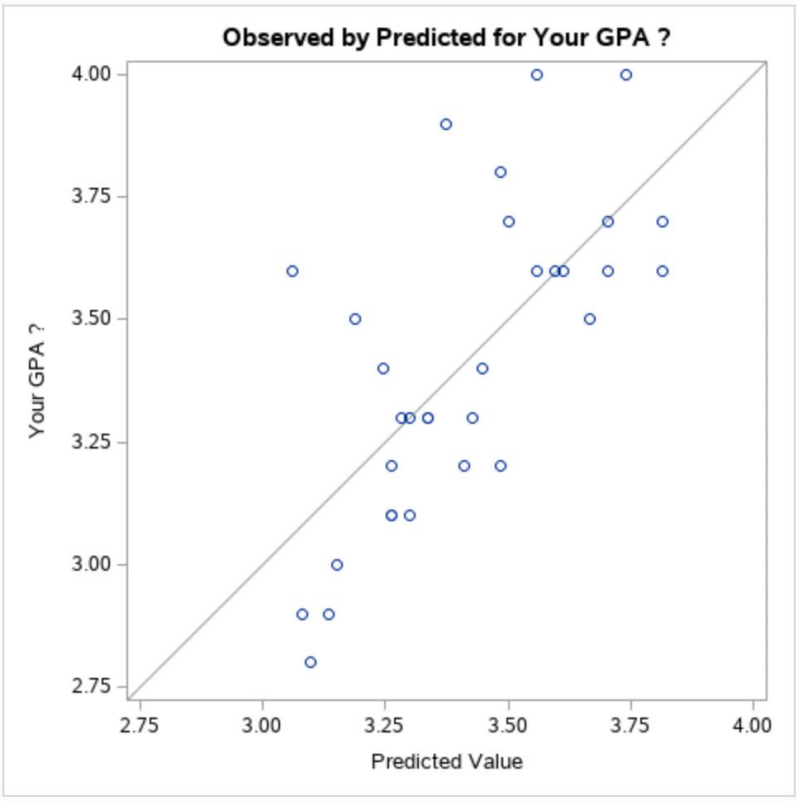
Obs	TIMESTAMP	YEAR	MAJOR	CGPA	CPWK	SMOKE
21	2/16/202	Senior	CS	3.30	4.0	No
22	2/16/202	Junior	CS	3.85	3.5	No
23	2/16/202	Senior	BIO	3.70	1.2	Yes
24	2/16/202	Junior	IS & OMB	3.60	7.0	No
25	2/16/202	Junior	ECON	3.97	4.0	No
26	2/16/202	Senior	SUPPLYC	3.80	2.0	Yes
27	2/16/202	Junior	ARCH	3.50	2.0	No
28	2/16/202	Junior	CS	3.85	6.0	No
29	2/27/202	Senior	ECON	3.50	3.0	No
30	2/28/202	Sophomor	CS	3.40	3.0	Yes
31	2/28/202	Senior	ECON	3.76	4.0	No
32	2/28/202	Sophomor	CS	4.00	5.0	No
33	2/28/202	Senior	CS	3.90	3.0	No


CGPA
4.0
3.9
3.8
3.7
3.6
3.5
3.4
3.3
3.2
3.1
3.0
2.9
2.8
2.7
1
2
3
4
5
6
7
8
9
10
11
12
CPWK

```
DATA temp ;
INFILE '/home/u58594663/project1.csv' delimiter=',' dsd;
INPUT
    TIMESTAMP $
    YEAR $
    MAJOR $
    CGPA
    CPWK
    SMOKE $;
RUN;

PROC PRINT DATA = temp;
RUN;
ODS GRAPHICS ON;
PROC GPLOT DATA = temp;
PLOT CGPA*CPWK;
RUN;
data WORK.PRO1;
set WORK.PRO1;
select ('Which year of study are you? 'n);
when ('Freshman') Year=1;
when ('Sophomore') Year=2;
when ('Junior') Year=3;
when ('Senior') Year=4;
otherwise Year='Which year of study are you? 'n;
end;
run;
ods noproctitle;
ods graphics / imagemap=on;
proc glmselect data=WORK.PRO1
outdesign(addinputvars)=Work.reg_design;
class 'What is your major'n / param=glm;
model 'CPWK'n /
showpvalues selection=none;
run;
PROC CORR DATA = temp;
VAR CGPA CPWK;
RUN;
PROC REG DATA = temp PLOTS=DIAGNOSTICS(STATS=NONE);
MODEL CGPA = CPWK;
RUN;
```

Model: MODEL1
Dependent Variable: Your GPA ?



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	.	.	.

---

2 Variables:	CGPA CPWK

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
CGPA	33	3.60152	0.36124	118.85000	2.70000	4.00000
CPWK	33	3.94545	2.24473	130.20000	1.00000	12.00000

Pearson Correlation Coefficients, N = 33
Prob > |r| under H0: Rho=0

	CGPA	CPWK
CGPA	1.00000	-0.26099 0.1424
CPWK	-0.26099 0.1424	1.00000

Model: MODEL1
Dependent Variable: CGPA

Number of Observations Read	33
Number of Observations Used	33

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.28444	0.28444	2.27	<.0001
Error	31	3.89138	0.12553		
Corrected Total	32	4.17582			

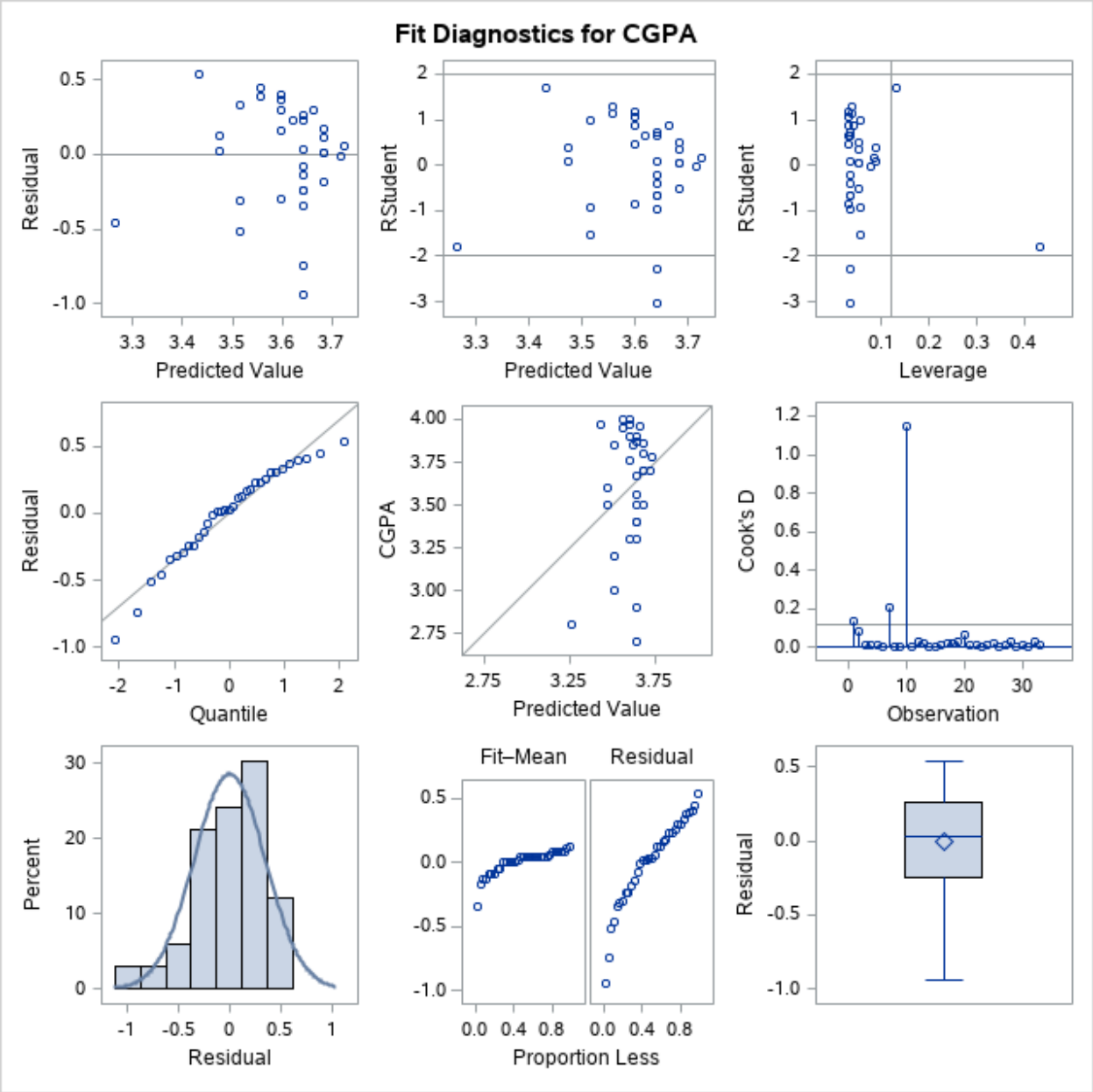
Root MSE	0.35430	R-Square	0.0681
Dependent Mean	3.60152	Adj R-Sq	0.0381

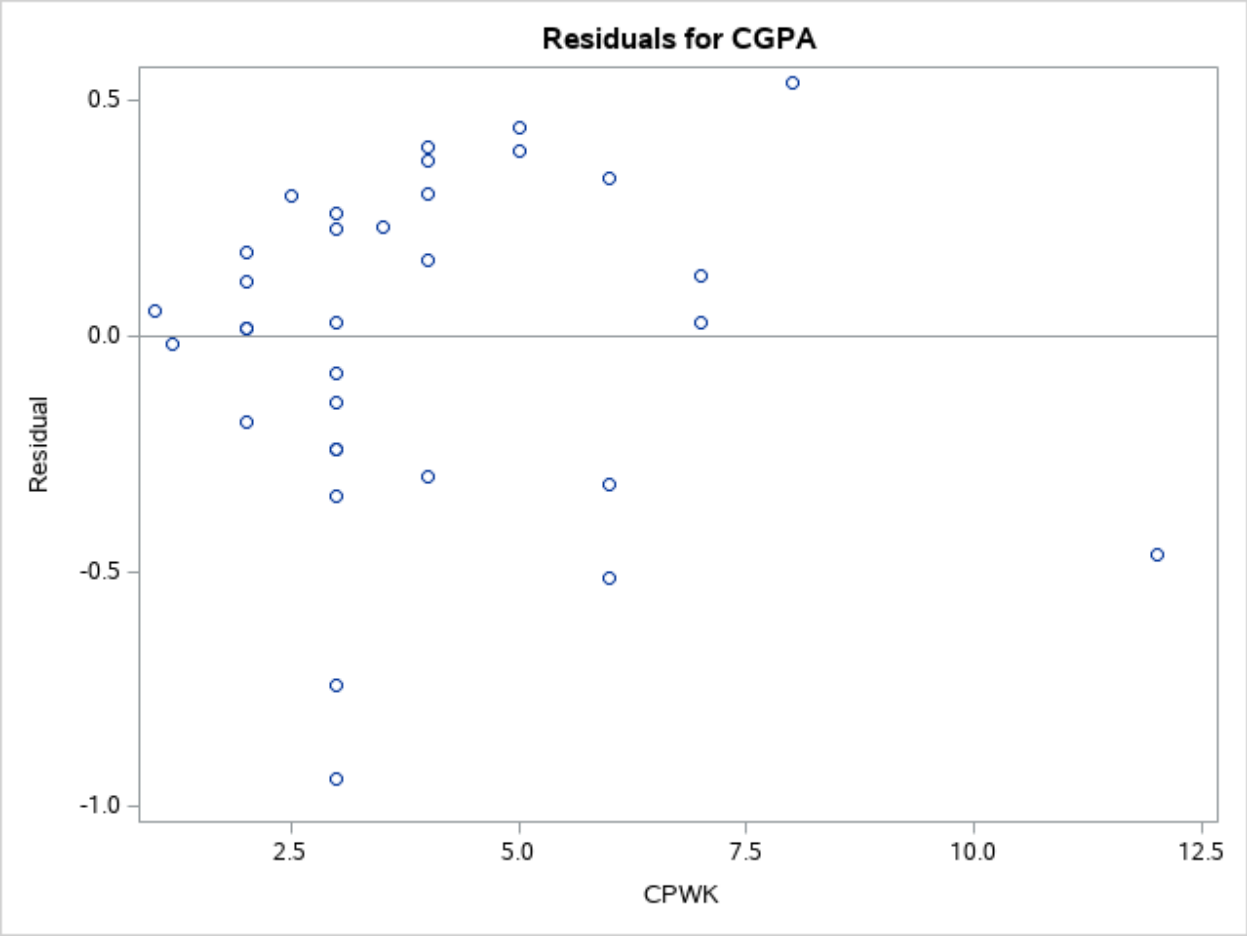
Coeff Var	9.83753			

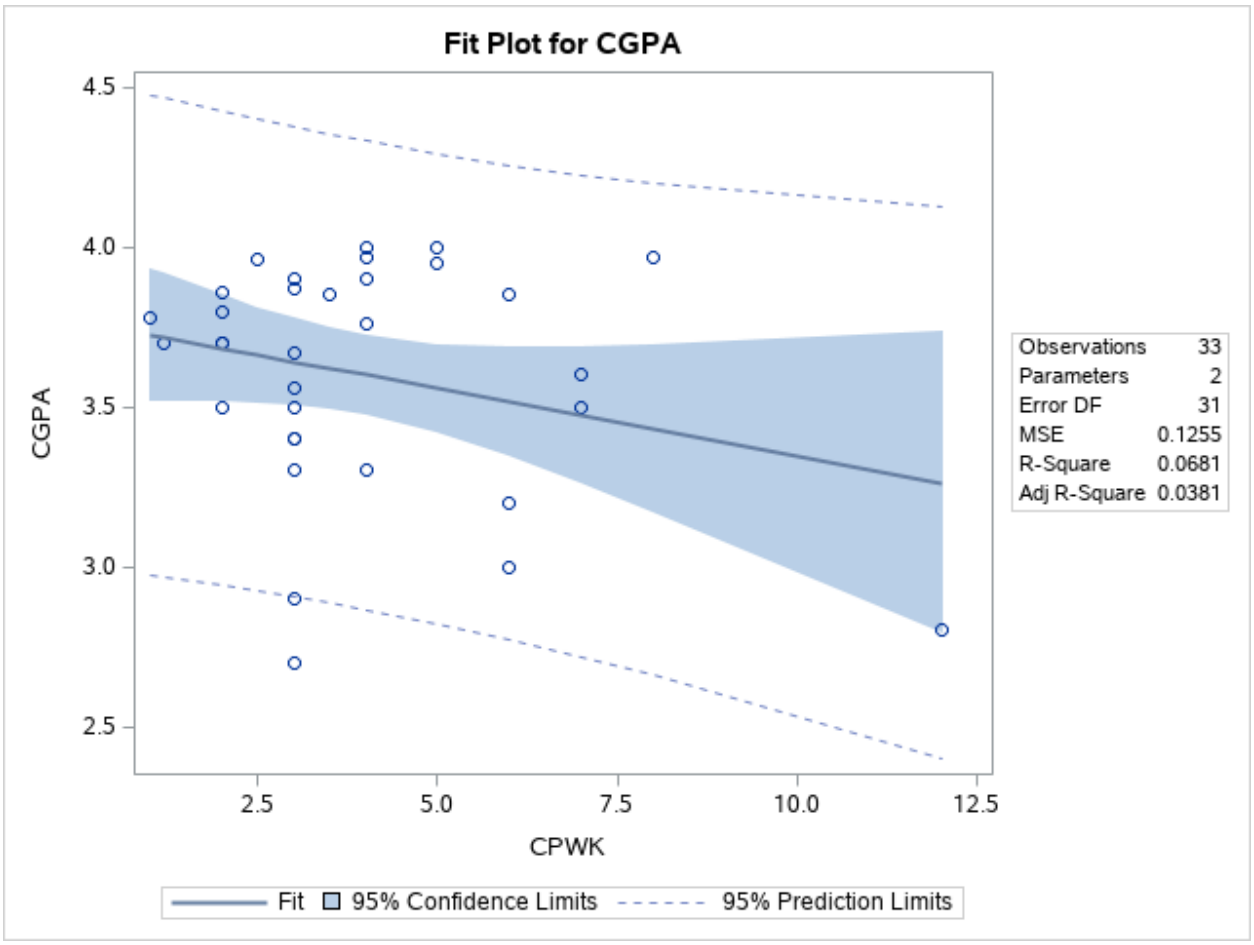
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.76723	0.12619	29.85	<.0001
CPWK	1	-0.04200	0.02790	-1.51	<.0001

Model: MODEL1
Dependent Variable: CGPA


Fit Diagnostics for CGPA
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
CGPA
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Residual

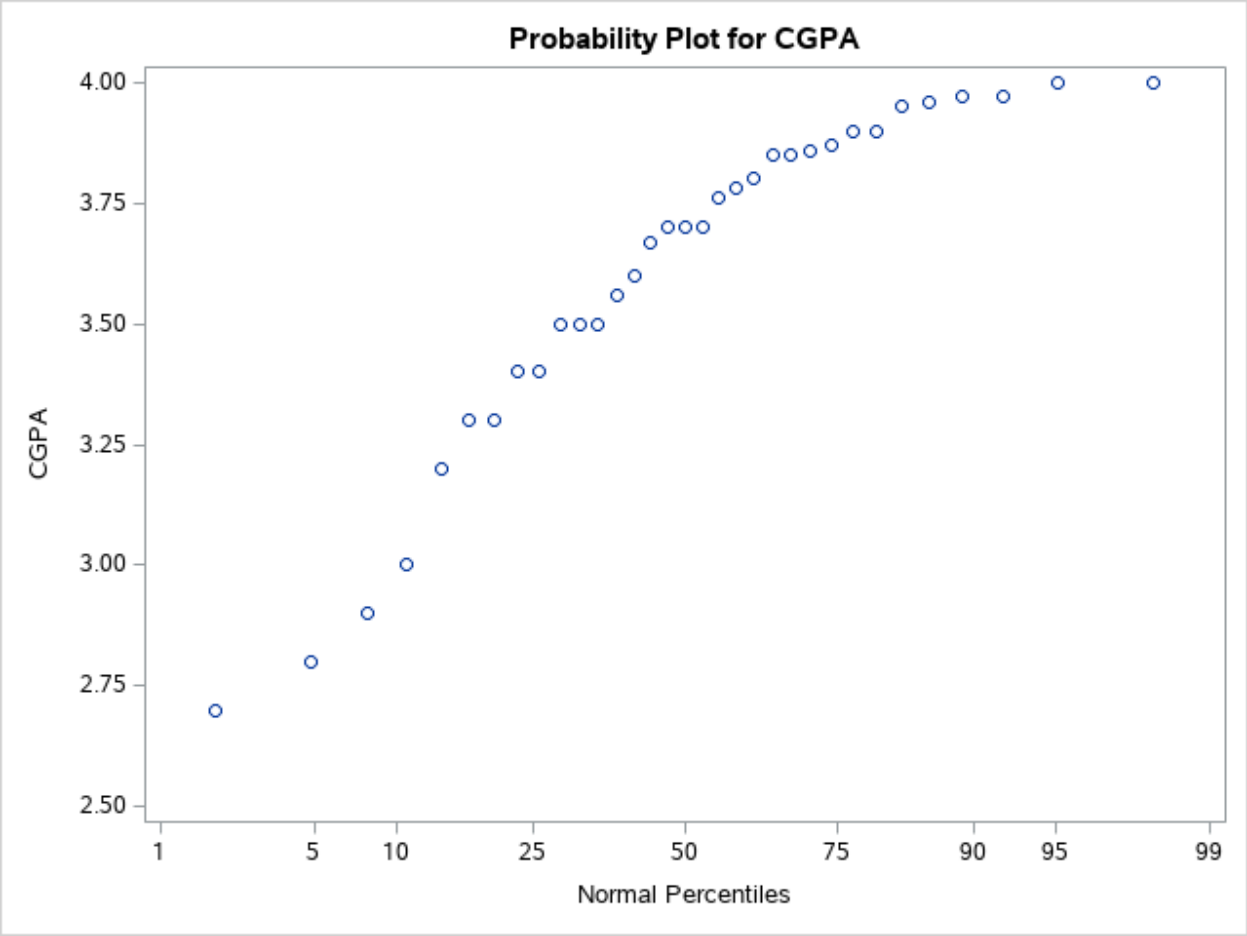

Residuals for CGPA
0.5
0.0
-0.5
-1.0
Residual
2.5
5.0
7.5
10.0
12.5
CPWK


Fit Plot for CGPA
CGPA
4.5
4.0
3.5
3.0
2.5
2.5
5.0
7.5
10.0
12.5
CPWK
Observations 33
Parameters 2
Error DF 31
MSE 0.1255
R-Square 0.0681
Adj R-Square 0.0381
Fit
95% Confidence Limits
95% Prediction Limits

Obs	TIMESTAMP	YEAR	MAJOR	CGPA	CPWK	SMOKE
1	2/16/202	Sophomor	CS	2.70	3.0	Yes
2	2/16/202	Sophomor	MATH	2.90	3.0	Yes
3	2/16/202	Senior	CS	3.86	2.0	No
4	2/16/202	Senior	MATH	3.90	4.0	No
5	2/16/202	Sophomor	STAT	3.40	3.0	Yes
6	2/16/202	Senior	ECON	3.56	3.0	Yes
7	2/16/202	Senior	CS	3.97	8.0	Yes
8	2/16/202	Senior	ECON	3.78	1.0	Yes
9	2/16/202	Senior	Media de	3.70	2.0	No
10	2/16/202	Senior	BIO	2.80	12.0	Yes
11	2/16/202	Freshman	FINA	3.50	7.0	No
12	2/16/202	Senior	INFOSCI	3.95	5.0	Yes
13	2/16/202	Freshman	ECON	4.00	4.0	No
14	2/16/202	Freshman	CS	3.67	3.0	No
15	2/16/202	Senior	PHS	3.70	2.0	Yes
16	2/16/202	Junior	CS	3.87	3.0	No
17	2/16/202	Freshman	FINA	3.96	2.5	No
18	2/16/202	Senior	CS/MATH	3.30	3.0	No
19	2/16/202	Junior	CS	3.20	6.0	Yes
20	2/16/202	Junior	ECON	3.00	6.0	No
21	2/16/202	Senior	CS	3.30	4.0	No
22	2/16/202	Junior	CS	3.85	3.5	No
23	2/16/202	Senior	BIO	3.70	1.2	Yes
24	2/16/202	Junior	IS & OMB	3.60	7.0	No
25	2/16/202	Junior	ECON	3.97	4.0	No
26	2/16/202	Senior	SUPPLYC	3.80	2.0	Yes
27	2/16/202	Junior	ARCH	3.50	2.0	No
28	2/16/202	Junior	CS	3.85	6.0	No
29	2/27/202	Senior	ECON	3.50	3.0	No
30	2/28/202	Sophomor	CS	3.40	3.0	Yes
31	2/28/202	Senior	ECON	3.76	4.0	No
32	2/28/202	Sophomor	CS	4.00	5.0	No
33	2/28/202	Senior	CS	3.90	3.0	No

The MEANS Procedure

Variable	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Mean
CGPA	2.70	3.40	3.70	3.87	4.00	3.60
CPWK	1.00	3.00	3.00	5.00	12.00	3.95


Probability Plot for CGPA
CGPA
4.00
3.75
3.50
3.25
3.00
2.75
2.50
1
5
10
25
50
75
90
95
99
Normal Percentiles

PROC CORR

1 With Variables:	CPWK
1 Variables:	CGPA

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
CPWK	33	3.94545	2.24473	130.20000	1.00000	12.00000
CGPA	33	3.60152	0.36124	118.85000	2.70000	4.00000

Pearson Correlation Coefficients, N = 33
Prob > |r| under H0: Rho=0

	CGPA
CPWK	-0.26099 0.1424