YINGQIAO GOU

STAT430
HW4

PROC PRINT – WORK.HW4FINAL

```
/* Generated Code (IMPORT) */
/* Source File: HW_4_Final_Exam_Work.csv */
/* Source Path: /home/u50368724/my_shared_file_links/schimiak
*/
/* Code generated on: 03/07/2021 11:09 */

%web_drop_table(WORK.HW4FINAL);


FILENAME REFFILE '/home/u50368724/my_shared_file_links/
schimiak/HW_4_Final_Exam_Work.csv';

PROC IMPORT DATAFILE=REFFILE
     DBMS=CSV
     OUT=WORK.HW4FINAL;
     GETNAMES=NO;
RUN;

PROC CONTENTS DATA=WORK.HW4FINAL; RUN;


%web_open_table(WORK.HW4FINAL);
```

Then, renamed the column names from VAR1, VAR2 and VAR3 to FinalGrade FinalExam and ClassWork.

Show Table

```
PROC SQL;
CREATE TABLE WORK.query AS
SELECT FinalGrade , FinalExam , ClassWork FROM WORK.HW4FINAL;
RUN;
QUIT;

PROC DATASETS NOLIST NODETAILS;
CONTENTS DATA=WORK.query OUT=WORK.details;
RUN;

PROC PRINT DATA=WORK.details;
RUN;
```

Obs	FinalGrade	FinalExam	ClassWork
1	61.60777778	51	76.94444444
2	98.59166667	88	100
3	75.22111111	54	98.44444444
4	90.63944444	88	93.27777778
5	78.79611111	65	96.61111111
6	68.81944444	51	79.61111111
7	85.01444444	71	97.94444444
8	90.66444444	77	96.61111111
9	76.245	58	96.33333333
10	86.50777778	87	96.11111111
11	91.13166667	67	100
12	94.21611111	85	92.11111111
13	86.87166667	60	100
14	55.56222222	44	43.05555556
15	73.73666667	64	9.333333333
16	75.60777778	56	41.94444444
17	66.85166667	67	17
18	73.18666667	63	58.33333333
19	59.37388889	22	5.555555556
20	71	29	10.61111111
21	62.08666667	60	0
22	78.01222222	54	83.38888889
23	92.55	77	98.66666667
24	100.0083333	100	100
25	74.36722222	55	87.72222222
26	81.93444444	65	93.94444444
27	82.30888889	77	99.22222222
28	86.21	85	100

1. Do a PROC FORMAT to change the following:

   Determine the letter grade (LETTER_GRADE) for the Final_Grade (10 point scale: 90-100 A, 80-<90 B, 70-<80 C, 60-<70D, and <60 F)

CODE

```
data work.transform;
      set WORK.HW4FINAL;
      if FinalGrade>90 then LETTER_GRADE='A';
      ELSE IF FinalGrade>80 and FinalGrade<=90 then LETTER_GRADE='B';
      ELSE IF FinalGrade>70 and FinalGrade<=80 then LETTER_GRADE='C';
      ELSE IF FinalGrade>60 and FinalGrade<=70 then LETTER_GRADE='D';
      ELSE LETTER_GRADE='F';
run;
```

RESULT

Obs	FinalGrade	FinalExam	ClassWork	LETTER_GRADE
1	61.60777778	51	76.94444444	D
2	98.59166667	88	100	A
3	75.22111111	54	98.44444444	C
4	90.63944444	88	93.27777778	A
5	78.79611111	65	96.61111111	C
6	68.81944444	51	79.61111111	D
7	85.01444444	71	97.94444444	B
8	90.66444444	77	96.61111111	A
9	76.245	58	96.33333333	C
10	86.50777778	87	96.11111111	B
11	91.13166667	67	100	A
12	94.21611111	85	92.11111111	A
13	86.87166667	60	100	B
14	55.56222222	44	43.05555556	E
15	73.73666667	64	9.333333333	C
16	75.60777778	56	41.94444444	C
17	66.85166667	67	17	D
18	73.18666667	63	58.33333333	C
19	59.37388889	22	5.555555556	E
20	71	29	10.61111111	C
21	62.08666667	60	0	D
22	78.01222222	54	83.38888889	C
23	92.55	77	98.66666667	A
24	100.0083333	100	100	A
25	74.36722222	55	87.72222222	C
26	81.93444444	65	93.94444444	B
27	82.30888889	77	99.22222222	B
28	86.21	85	100	B

2. Make a frequency chart for Letter_Grade

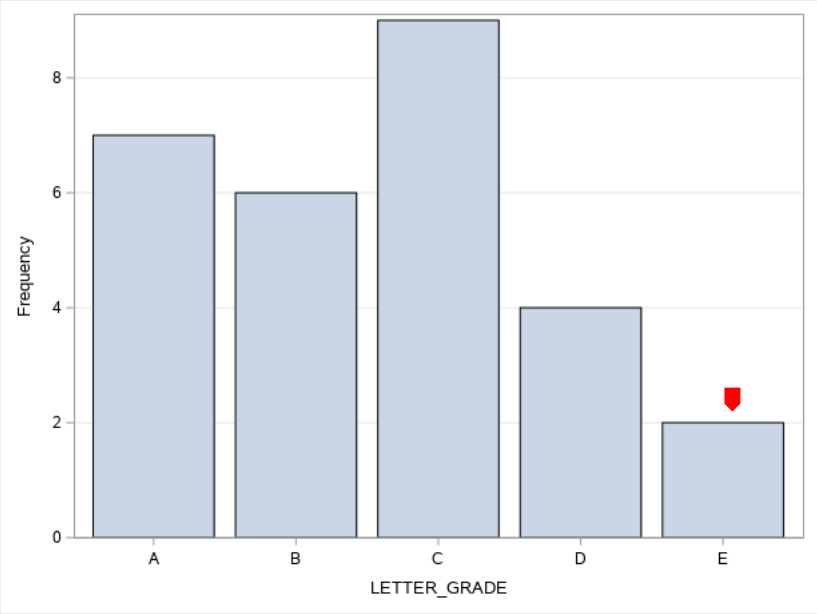
CODE

```
ods graphics / reset width=6.4in height=4.8in
imagemap;

proc sgplot data=WORK.TRANSFORM;
     vbar LETTER_GRADE /;
     yaxis grid;
run;

ods graphics / reset;
```

BAR CHART



3. Create a Scatterplot of the final grade vs final exam grade.

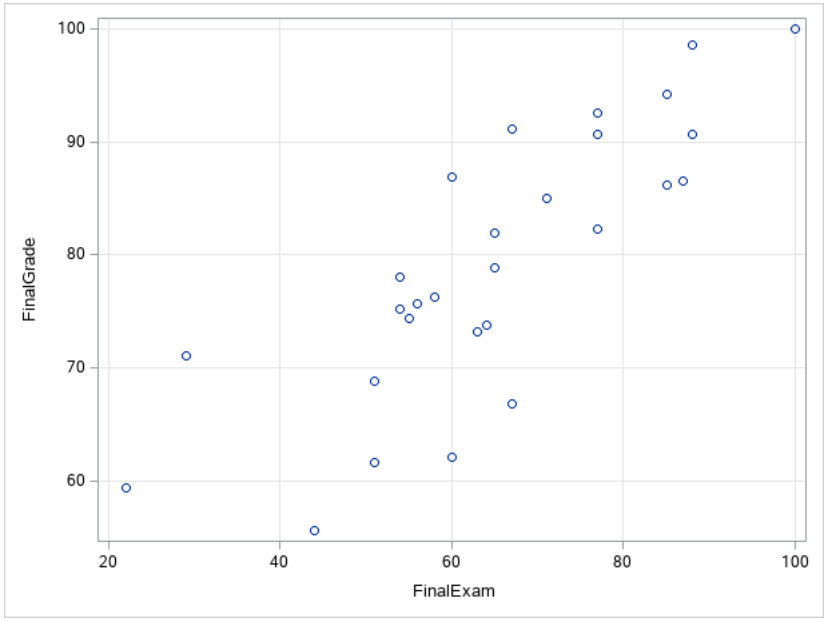
CODE

```
ods graphics / reset width=6.4in height=4.8in
imagemap;

proc sgplot data=WORK.TRANSFORM;
     scatter x=FinalExam y=FinalGrade /;
     xaxis grid;
     yaxis grid;
run;

ods graphics / reset;
```

CHART



4. Create a scatterplot of the final grade vs class work.

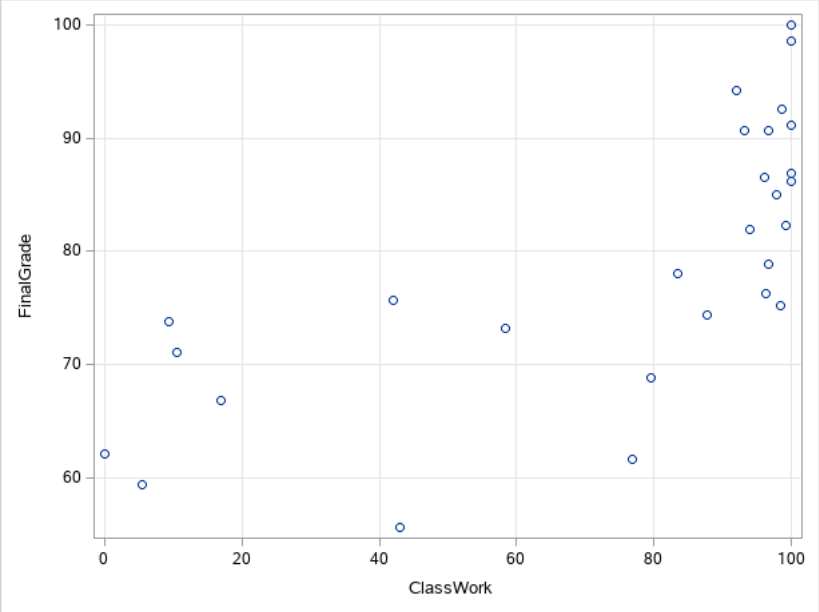
CODE

```
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.TRANSFORM;
     scatter x=ClassWork y=FinalGrade /;
     xaxis grid;
     yaxis grid;
run;

ods graphics / reset;
```

PLOT



5. Determine then state the Pearson Correlation Coefficient for the following:
    a. final grade and final exam
    b. final grade and class work

CODE

```
ods noproctitle;
ods graphics / imagemap=on;

proc corr data=WORK.TRANSFORM pearson nosimple
noprob plots=none;
     var FinalGrade;
     with FinalExam ClassWork;
run;
```

CORRELATIONS

2 With Variables:	FinalExam ClassWork
1 Variables:	FinalGrade

Pearson Correlation Coefficients, N = 28	
	FinalGrade
FinalExam FinalExam	0.81452
ClassWork ClassWork	0.71299

6. Do a regression on the following:

a. Predict final grade based on final exam grade

CODE

```
ODS NOPROCTITLE;
ODS GRAPHICS / IMAGEMAP=ON;

PROC REG DATA=WORK.TRANSFORM ALPHA=0.05
PLOTS(ONLY)=(DIAGNOSTICS RESIDUALS
          FITPLOT OBSERVEDBYPREDICTED);
     MODEL FINALGRADE=FINALEXAM /;
     OUTPUT OUT=WORK.REG_STATS0001 P=P_ LCL=LCL_
UCL=UCL_ LCLM=LCLM_ UCLM=UCLM_
          R=R_;
     RUN;
QUIT;
```

ANOVA TABLE

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2547.46062	2547.46062	51.25	<.0001
Error	26	1292.32267	49.70472		
Corrected Total	27	3839.78328			

TABLE 2

Root MSE	7.05016	R-Square	0.6634
Dependent Mean	79.18298	Adj R-Sq	0.6505
Coeff Var	8.90363		

PARAMETER TABLE

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	43.69879	5.13250	8.51	<.0001
FinalExam	FinalExam	1	0.54591	0.07625	7.16	<.0001

i. Speak to meaning and implication of the following:

1. P-value of ANOVA table

   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.

2. P-value of the slope and intercept

   P-value of slope is < 0.0001.

   P-value of intercept is <0.0001.

   This means both the FinalExam and Intercept are significant to this model predicting.

3. R^2 value

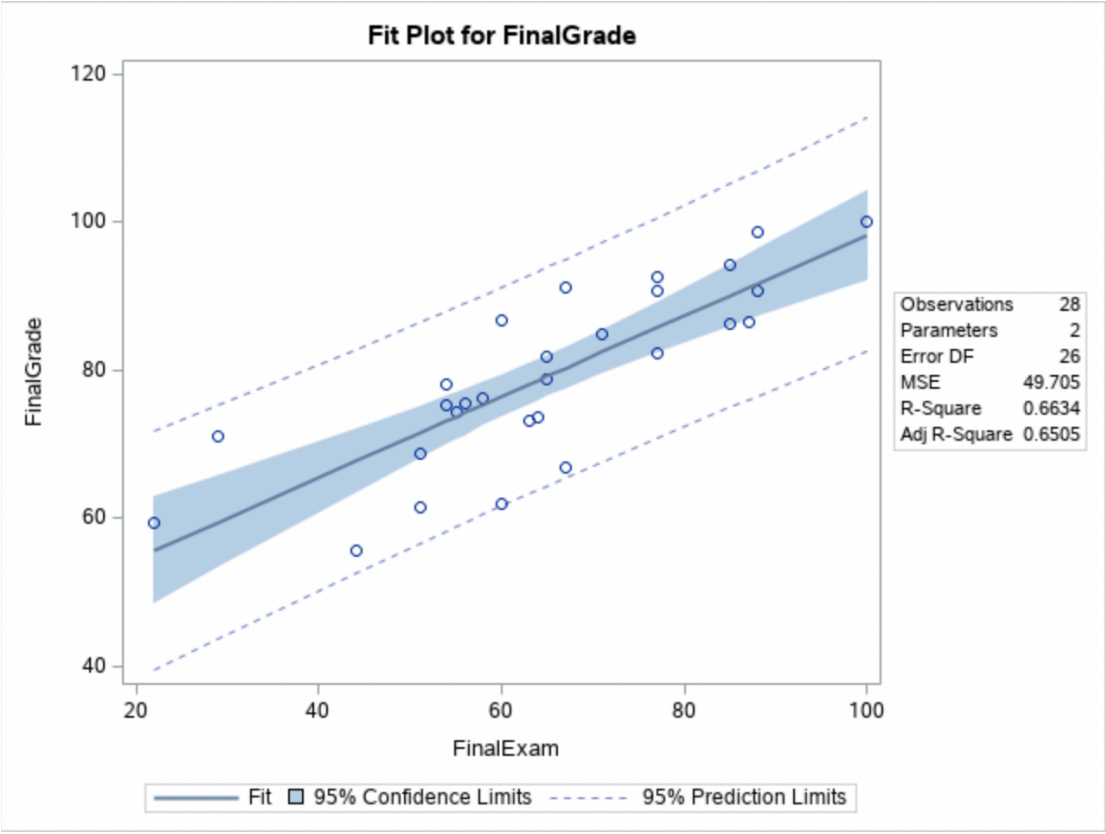
   0.6634.

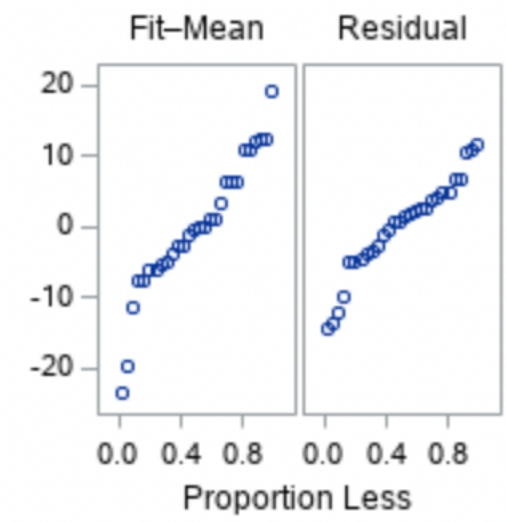
$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

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

4. Residual Analysis

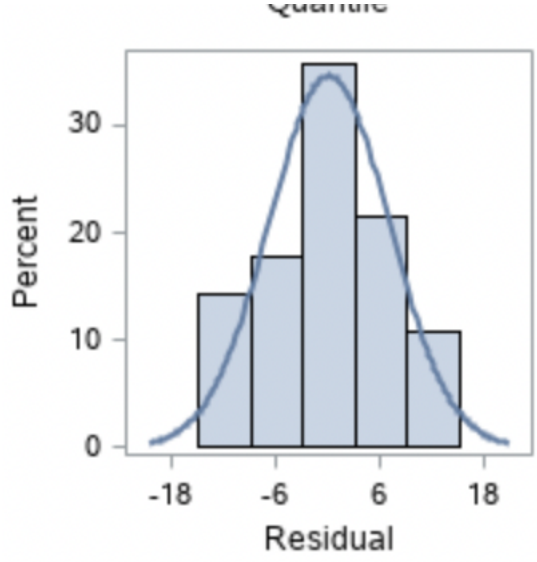
    a. Fit vs Residuals





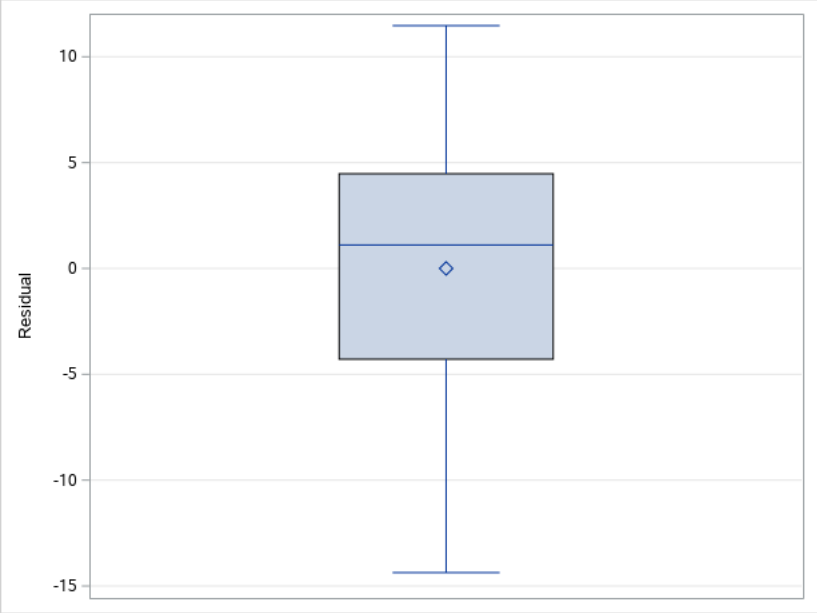
In the first plot, we could see that all points are in the area of 95% CI. Therefore, in the case of 95% CI, the current model performs well.

b. Probability Plot



The residual followed a normal distribution.

c. Boxplot



The distribution of residuals is roughly Symmetrical.

ii. Should you use this regression equation, if so, what is the regression equation?

Yes, we should use this regression. The equation is

FinalGrade = 0.54591 * FinalExam +43.69879.

b. Predict final grade based on class work

CODE

```
ODS NOPROCTITLE;
ODS GRAPHICS / IMAGEMAP=ON;

PROC REG DATA=WORK.TRANSFORM ALPHA=0.05
PLOTS(ONLY)=(DIAGNOSTICS RESIDUALS
           FITPLOT OBSERVEDBYPREDICTED);
     MODEL FINALGRADE=CLASSWORK /;
     OUTPUT OUT=WORK.REG_STATS0001 DFFITS=DFFITS_
P=P_ LCL=LCL_ UCL=UCL_ LCLM=LCLM_
           UCLM=UCLM_ PRESS=PRESS_ R=R_
STUDENT=STUDENT_ RSTUDENT=RSTUDENT_;
     RUN;
QUIT;
```

ANOVA TABLE

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1951.96344	1951.96344	26.88	<.0001
Error	26	1887.81984	72.60846		
Corrected Total	27	3839.78328			

TABLE 2

Root MSE	8.52106	R-Square	0.5084
Dependent Mean	79.18298	Adj R-Sq	0.4894
Coeff Var	10.76123		

PARAMETER TABLE

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	61.19508	3.82479	16.00	<.0001
ClassWork	ClassWork	1	0.24299	0.04686	5.18	<.0001

i. Speak to meaning and implication of the following:

1. P-value of ANOVA table

   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.

2. P-value of the slope and intercept

   P-value of slope is < 0.0001.

   P-value of intercept is <0.0001.

   This means both the ClassWork and Intercept are significant to this model predicting.

3. R^2 value

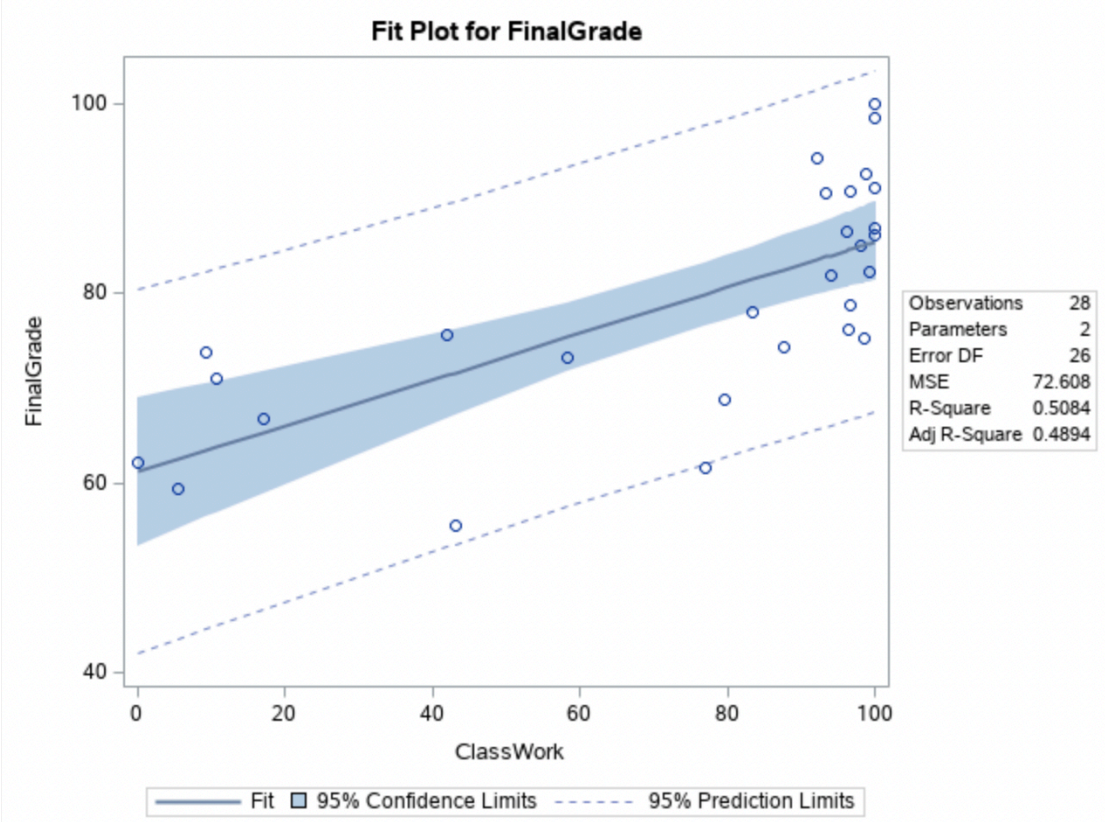
   0.5084.

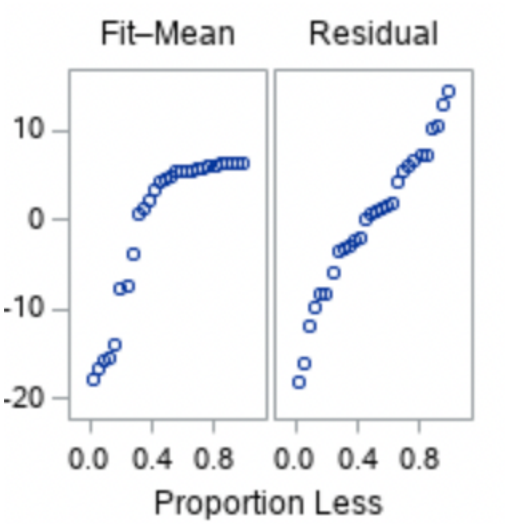
$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

Usually, the larger the R^2, the better the regression model fits your observations. Therefore, although the value of this R-squared less than the FinalExam ones, this result still seems to be not bad.

4. Residual Analysis

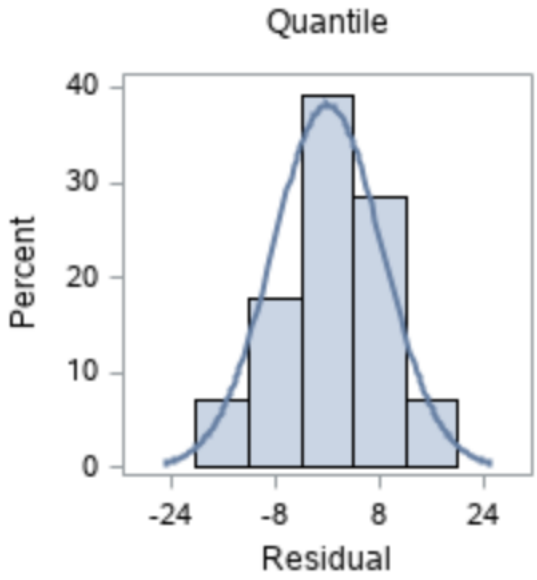
    a. Fit vs Residuals





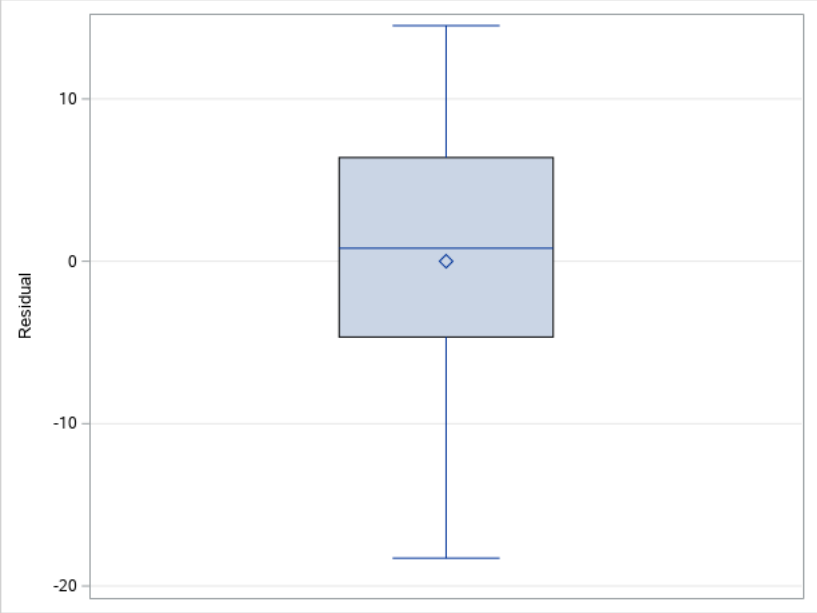
In the first plot, we could see that all points are in the area of 95% CI. Therefore, in the case of 95% CI, the current model performs well.

b. Probability Plot



The residual followed a normal distribution.

c. Boxplot



The distribution of residuals is roughly Symmetrical.

ii. Should you use this regression equation, if so, what is the regression equation?

Yes, we should use this regression. The equation is

FinalGrade = 0.24299 * ClassWork +61.19508.

REG_Test1_2

It 2 variables: Test_1 and Test_2
(Suggestion: Do a PROC PRINT to see the data, this will be very helpful to you.)

PROC PRINT – WORK.REG

```
%web_drop_table(WORK.REG);


FILENAME REFFILE '/home/u50368724/my_shared_file_links/
schimiak/REG_Test_1_2.csv';

PROC IMPORT DATAFILE=REFFILE
     DBMS=CSV
     OUT=WORK.REG;
     GETNAMES=NO;
RUN;

PROC CONTENTS DATA=WORK.REG; RUN;


%web_open_table(WORK.REG);
```

Then, renamed the column names from VAR1 and VAR2 to Test_1 and Test_2.

Show Table

```
PROC SQL;
CREATE TABLE WORK.query AS
SELECT FinalGrade , FinalExam , ClassWork FROM WORK.HW4FINAL;
RUN;
QUIT;

PROC DATASETS NOLIST NODETAILS;
CONTENTS DATA=WORK.query OUT=WORK.details;
RUN;

PROC PRINT DATA=WORK.details;
RUN;
```

Part of Result

Obs	Test_1	Test_2
1	64	61
2	84	89
3	100	89
4	96	61
5	88	89
6	80	85
7	100	77
8	68	65
9	76	57
10	80	81
11	88	81
12	76	69
13	88	81
14	100	93
15	100	93
16	92	80

7. Do a Regression analysis predicting Test_2 based on Test_1.

a. Speak to meaning and implication of the following:

CODE

```
ods noproctitle;
ods graphics / imagemap=on;

proc reg data=WORK.TRANSFORM alpha=0.05
plots(only)=(diagnostics residuals
          fitplot observedbypredicted);
     model FinalGrade=FinalExam /;
     output out=work.Reg_stats0001 p=p_ lcl=lcl_
ucl=ucl_ lclm=lclm_ uclm=uclm_
          r=r_;
     run;
quit;
```

ANOVA TABLE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	15189	15189	94.88	<.0001
Error	219	35060	160.08912		
Corrected Total	220	50248			

TABLE 2

Root MSE	12.65263	R-Square	0.3023
Dependent Mean	74.64706	Adj R-Sq	0.2991
Coeff Var	16.94994		

PARAMETER TABLE

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.14895	7.69546	0.02	0.9846
Test_1	Test_1	1	0.85912	0.08820	9.74	<.0001

i. P-value of ANOVA table

<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.

ii. P-value of the slope and intercept

P-value of slope is < 0.0001.

P-value of intercept is <0.9846.

This means both the Test_1 is significant to this model predicting, but intercept is not.

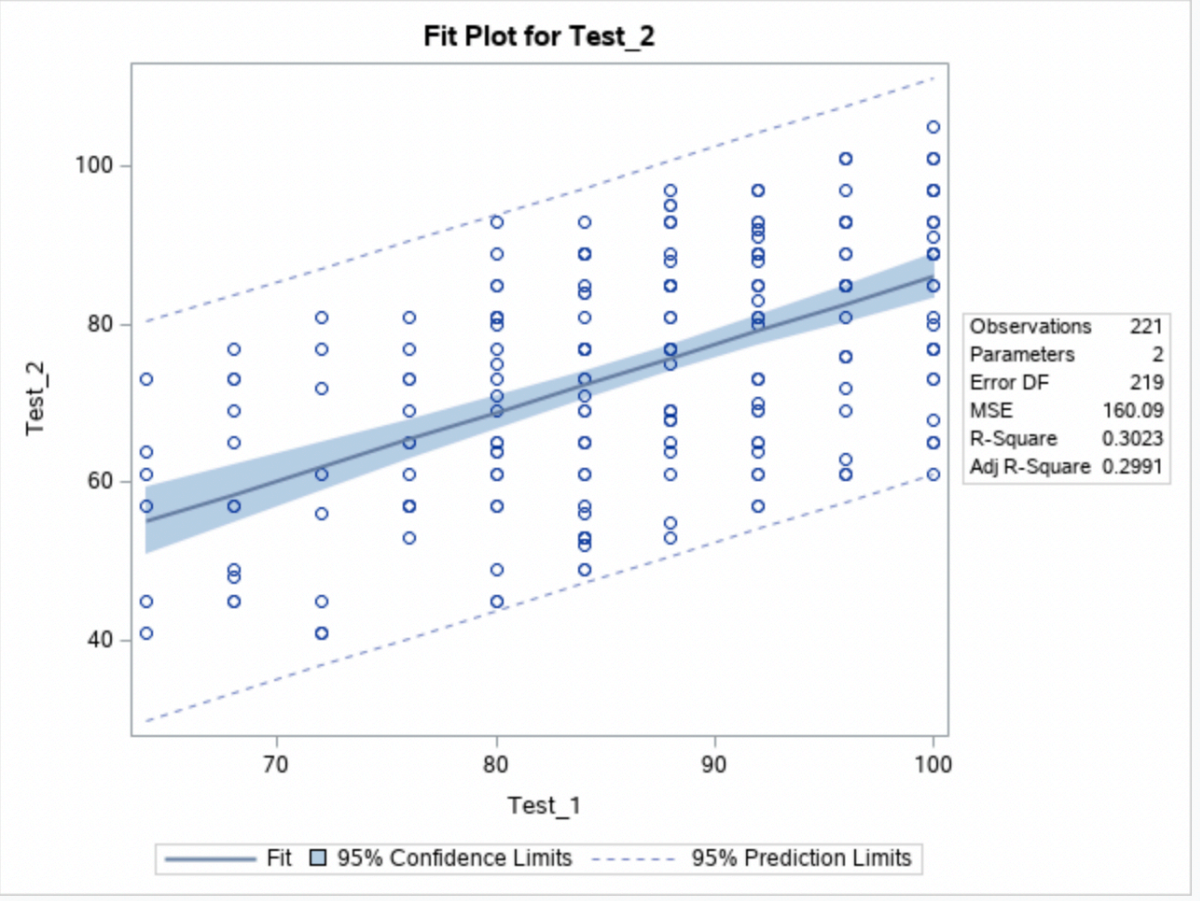
iii. R^2 value

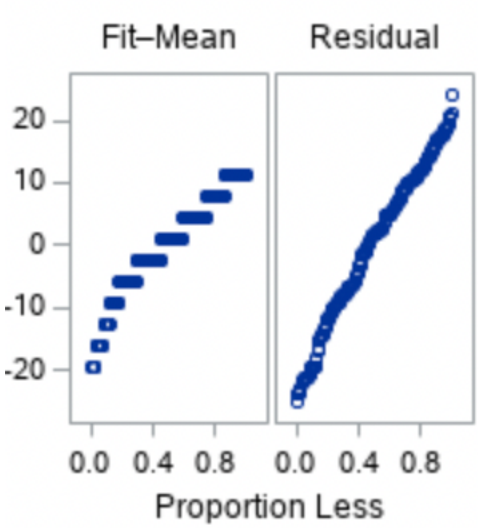
0.3023

Usually, the larger the R^2, the better the regression model fits your observations. Also, the value interval of R-squared is [0, 1], thus o.3 is not a good one.

iv. Residual Analysis

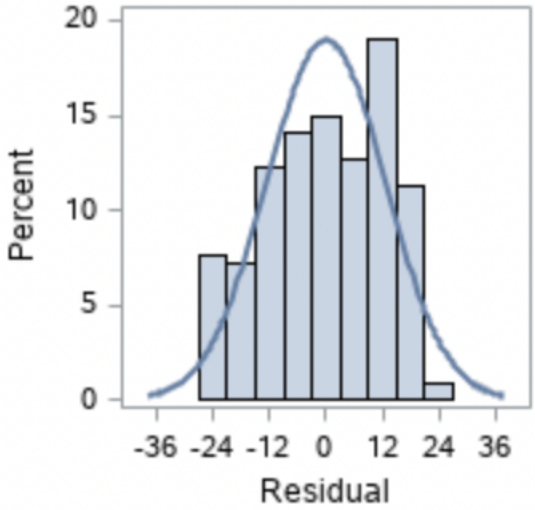
1. Fit vs Residuals





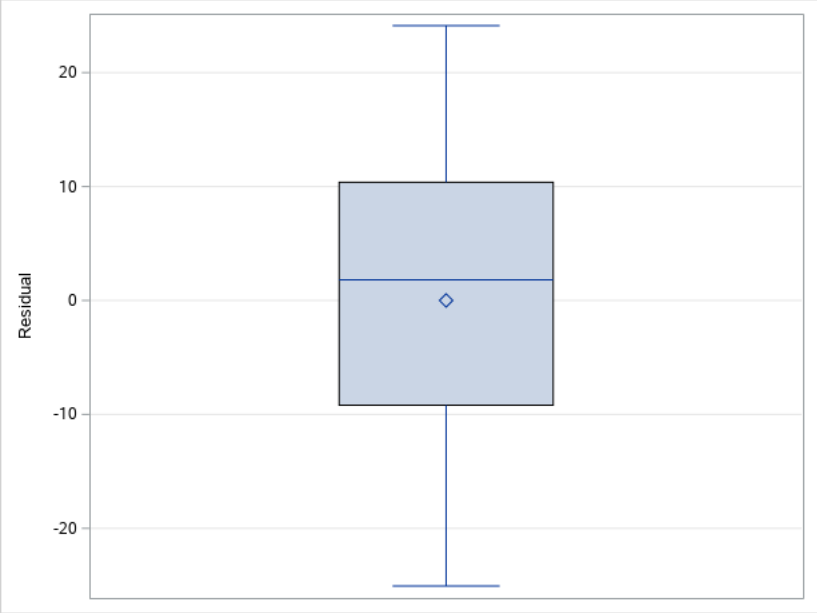
The first plot shows that most of points are outside the 95% Confidence Limits.

2. Probability Plot



This plot shows that the residual of this model does not follow a normal distribution.

3. Boxplot



The distribution of residuals is roughly Symmetrical and more spreading out.

b. Should you use this regression equation, if so, what is the regression equation?

We should not use this regression. This is because that

(1) R-squared (roughly o.3) is not good.
(2) Model fitting the data is bad.
(3) P-value of intercept of this model is too large, is about o.9846.