STAT430

Homework 4

Yizhan Ao

(Suggestion: Do a PROC PRINT to see the data, this will be very helpful to you)

CODE:

```
Data Homework4;

Infile
'/home/u58594663/my_shared_file_links/schimiak/HW_4_Final_Exam_W
ork.csv' Delimiter=',' DSD;

INPUT

Final_Grade

Final_Exam
```

Obs	Final_Grade	Final_Exam	Class_Work
1	61.608	51	76.944
2	98.592	88	100.000
3	75.221	54	98.444
4	90.639	88	93.278
5	78.796	65	96.611
6	68.819	51	79.611
7	85.014	71	97.944
8	90.664	77	96.611
9	76.245	58	96.333

Obs	Final_Grade	Final_Exam	Class_Work
10	86.508	87	96.111
11	91.132	67	100.000
12	94.216	85	92.111
13	86.872	60	100.000
14	55.562	44	43.056
15	73.737	64	9.333
16	75.608	56	41.944
17	66.852	67	17.000
18	73.187	63	58.333
19	59.374	22	5.556
20	71.000	29	10.611
21	62.087	60	0.000
22	78.012	54	83.389
23	92.550	77	98.667
24	100.008	100	100.000
25	74.367	55	87.722
26	81.934	65	93.944
27	82.309	77	99.222
28	86.210	85	100.000

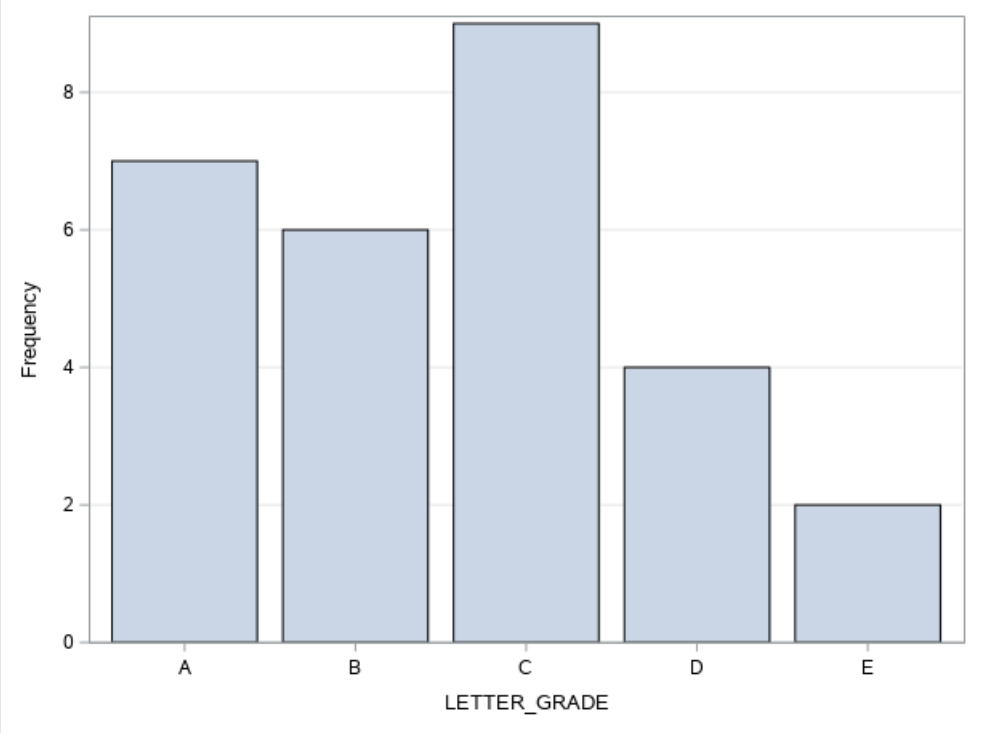
1. Do a PROC FORMAT to change the following: Determine the letter grade (LETTER_GRADE) for the Final_Grade (10 point scale: 90-100 A, 80-

```
IF Final_Grade >= 90 then LETTER_GRADE = 'A';
ELSE IF Final_Grade >= 80 and Final_Grade < 90 then LETTER_GRADE = 'B';
ELSE IF Final_Grade >= 70 and Final_Grade < 80 then LETTER_GRADE = 'C';
ELSE IF Final_Grade >= 60 and Final_Grade < 70 then LETTER_GRADE = 'D';
ELSE LETTER_GRADE = 'F';
```

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

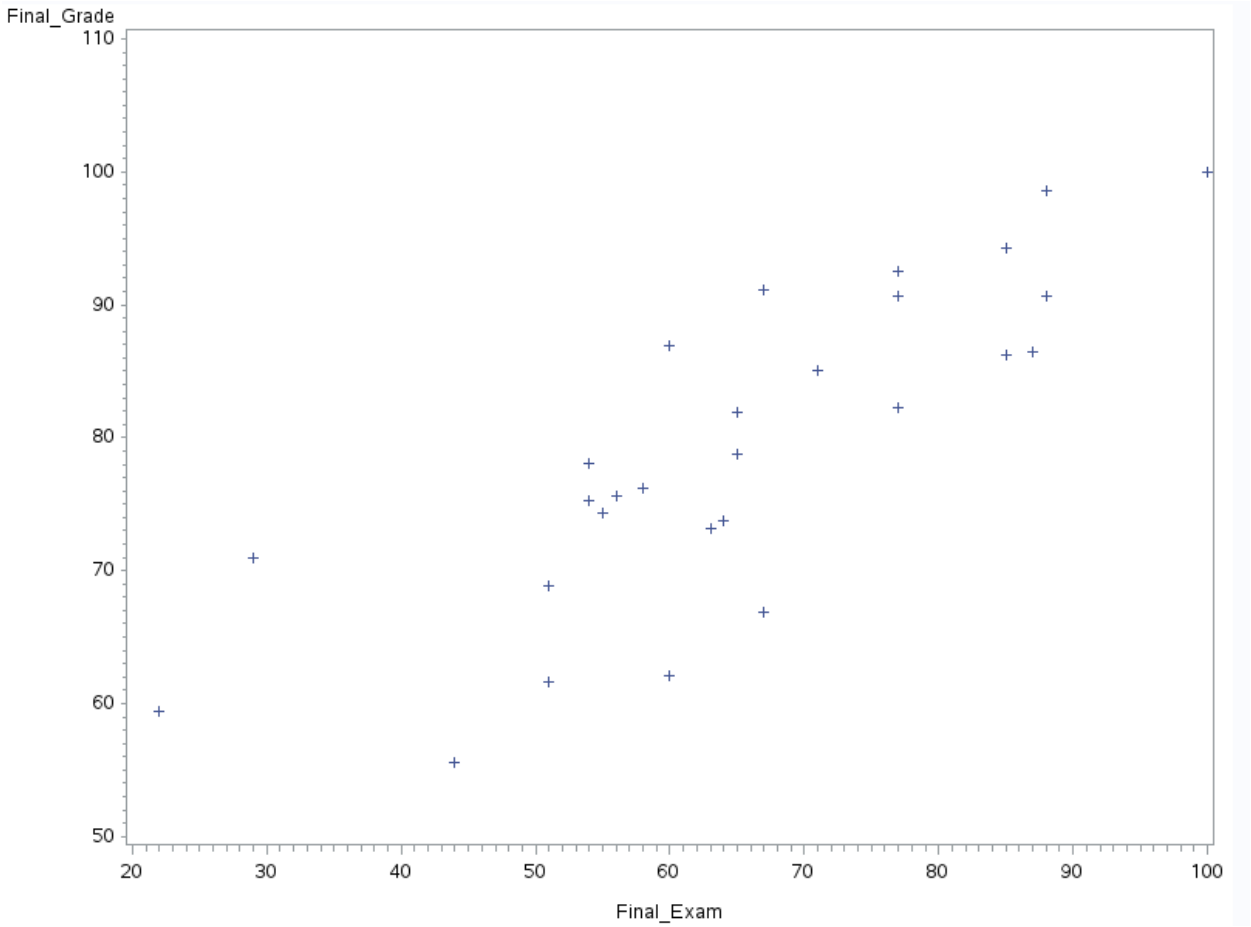
2. Make a frequency chart for Letter_Grade

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



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

```
PROC GPLOT DATA = HW4;
PLOT Final_Grade*Final_Exam;
RUN;
```

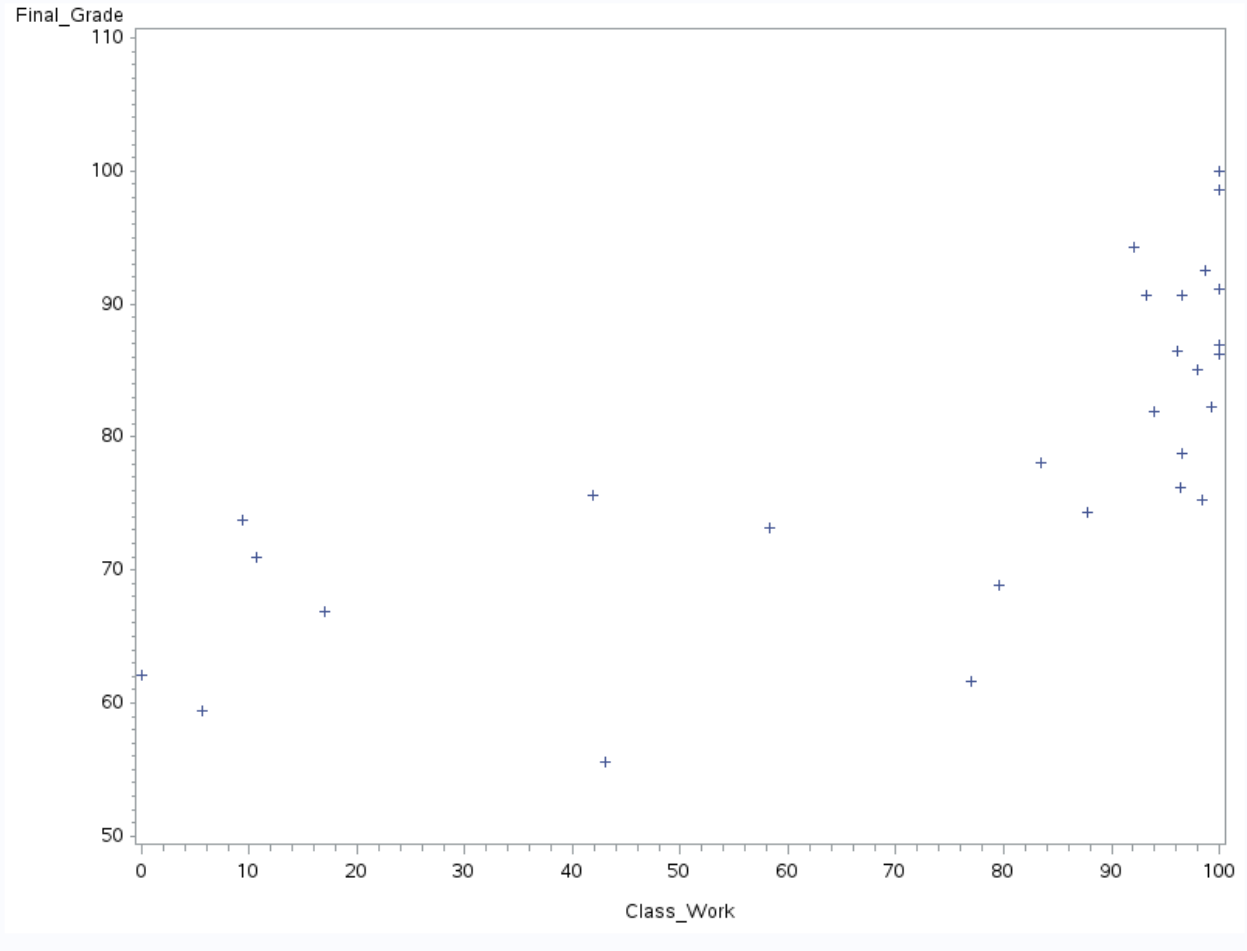


4. Create a scatterplot of the final grade vs class work.

```
PROC GPLOT DATA = Homework4;

PLOT Final_Grade*Class_Work;

RUN;
```



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

```
/*  */

PROC CORR DATA = Homework4;

VAR Final_Grade Final_Exam;

RUN;
```

The CORR Procedure

2 Variables:	Final_Grade Final_Exam

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Final_Grade	28	79.18298	11.92536	2217	55.56222	100.00833
Final_Exam	28	65.00000	17.79305	1820	22.00000	100.00000

Pearson Correlation Coefficients, N = 28 Prob > |r| under H0: Rho=0		
	Final_Grade	Final_Exam
Final_Grade	1.00000	0.81452 <.0001
Final_Exam	0.81452 <.0001	1.00000

Based on Pearson Correlation coefficient of the final grade and final exam is 0.81452

b. final grade and class work

```
PROC CORR DATA = Homework4;
VAR Final_Grade Class_Work;
RUN;
```

The CORR Procedure

2 Variables:	Final_Grade Class_Work

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Final_Grade	28	79.18298	11.92536	2217	55.56222	100.00833
Class_Work	28	74.02778	34.99197	2073	0	100.00000

Pearson Correlation Coefficients, N = 28
Prob > |r| under H0: Rho=0

	Final_Grade	Class_Work
Final_Grade	1.00000	0.71299 <.0001
Class_Work	0.71299 <.0001	1.00000

Based on Pearson Correlation coefficient of the final grade and final exam is 0.71299.

6. Do a regression on the following:
    a. Predict final grade based on final exam grade

```
PROC REG DATA = Homework4 PLOTS = DIAGNOSTICS(STATS = NONE);

MODEL Final_Grade = Final_Exam;

RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: Final_Grade

Number of Observations Read	28
Number of Observations Used	28

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			

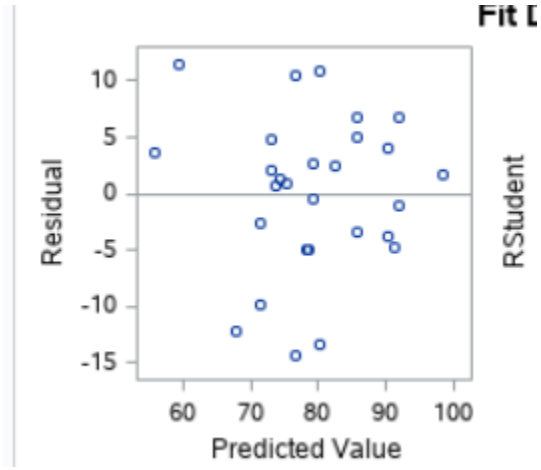
Root MSE	7.05016	R-Square	0.6634
Dependent Mean	79.18298	Adj R-Sq	0.6505
Coeff Var	8.90363		

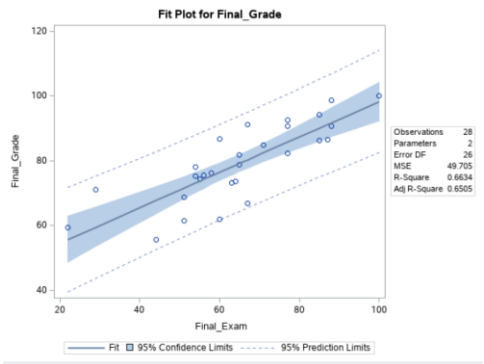
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	43.69879	5.13250	8.51	<.0001
Final_Exam	1	0.54591	0.07625	7.16	<.0001

I Speak to meaning and implication of the following:

1. P-value of ANOVA table
    a. P value is <0.0001 given 95% confidence level, the p value is less than 0.001 and the difference is small meaning the data is statistically important. So the data model is good here
2. P-value of the slope and intercept
    a. The P-value of slope is < 0.0001,
    b. given 95% confidence level, the p value slope is less than 0.001 , and the slope of intercept are <0.0001 which means the intercept and the slope are statistically important
3. R 2 value
    a. The R^2 value is 0.6634
    b. The coefficient of determination (R2) expresses how much variance in one variable can be explained by the variance in another. The R2 value is 0.6634, which suggests that x accounts for 66.34 percent of the variation in the y variable (final grade).
4. Residual Analysis
    a. Fit vs Residuals

i.

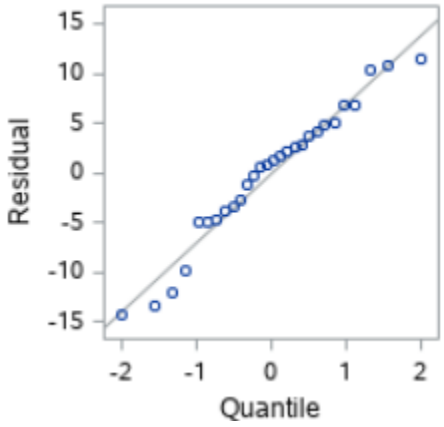




ii. The plot shows that there is no telescoping and that the residual distribution is symmetrically centered on zero. This indicates that the standard deviation of the residuals is constant. Also, because there is no discernible patterning, we can infer the residual's independence. The model is then acceptable.

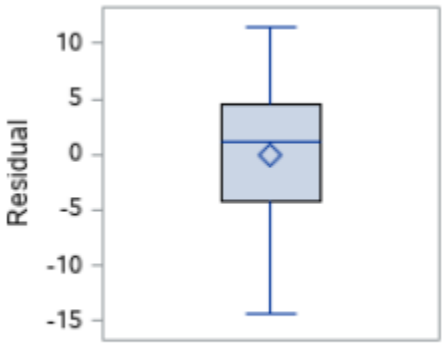
b. Probability Plot

i.



ii. This is basically linear. Given some outliers on the end and start

c. Boxplot



i.

ii. Given Box plot no outlier catching us eyes

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

The model is mainly reliable and can be utilized because the residual is IID, but it is not very reliable because the R2 is not large enough. The following is the regression equation:

Final_Grade = 0.54591 * Final_Exam + 43.69879

b. Predict final grade based on class work

```
PROC REG DATA = HW4 PLOTS = DIAGNOSTICS(STATS = NONE);

MODEL Final_Grade = Final_Exam;

RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: Final_Grade

Number of Observations Read	28
Number of Observations Used	28

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			

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

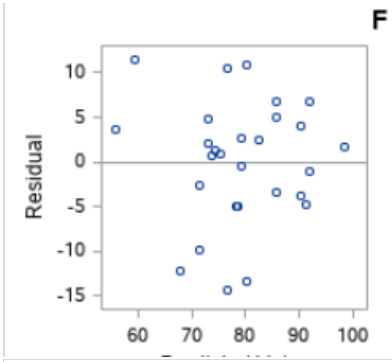
Parameter Estimates

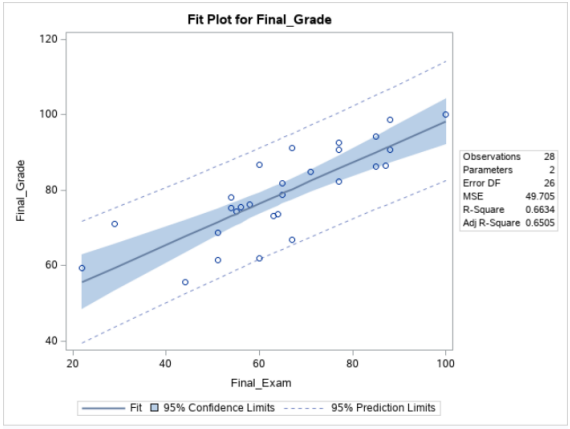
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	43.69879	5.13250	8.51	<.0001
Final_Exam	1	0.54591	0.07625	7.16	<.0001

i.

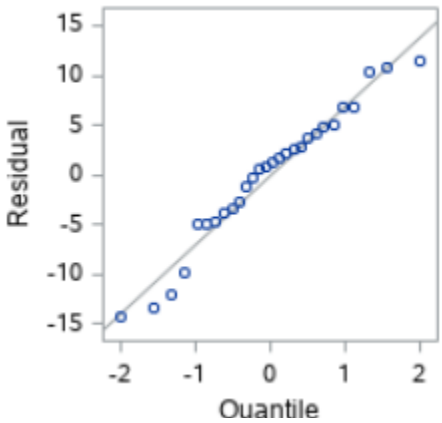
1. P-value of ANOVA table
    a. The P-value of ANOVA table is < 0.0001.
2. Under the condition of the 95% confident interval, P-value < 0.0001 means the differences between the variances of the means are statistically significant. Model looks good
3. P-value of the slope and intercept
    a. The P-value of slope is < 0.0001,
    b. The P-value of intercept is <0.0001.
4. Under the condition of the 95% confident interval, which means our independent variable and the intercept is statistically significant, we reject the null that the slope is 0. Also, we reject the null that the intercept is 0. Therefore, we can say the class work variable plays an important role in the linear regression model when predicting the final grade.
5. 3. R2 value
    a. R^2 is 0.5084
    b. The coefficient of determination (R2) expresses how much variance in one variable can be explained by the variance in another. The R2 value is 0.5084, indicating that the x variable

explains 50.84 percent of the variation in the y variable (final grade) (class work). This R2 value indicates the model's utility; it fits the requirement, but not very well. This poses a 50 percent of not fitting

6. Residual Analysis
    a. Fit vs Residuals
    b. 



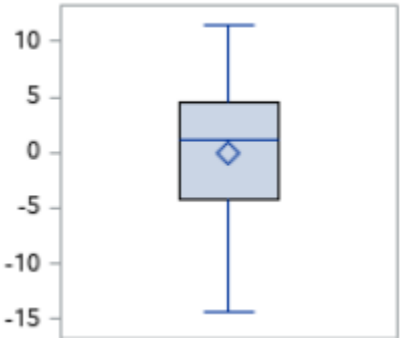
        i. The plot shows that there is no telescoping and that the residual distribution is symmetrically centered on zero. This indicates that the standard deviation of the residuals is constant. Also, because there is no discernible patterning, we can infer the residual's independence. The model is then acceptable.
    c. Probability Plot



i.

ii. The plot of residuals looks basically linear

d. Boxplot



i.

ii. we find that there is no outlier.

e.

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

iii. The model is mainly reliable and can be utilized, however it is not particularly dependable because the R2 is not large enough. The following is the regression equation:

1. 0.24299 * Class Work + 61.19508 = Final Grade

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

```
PROC REG DATA = Homework4_2 PLOTS = DIAGNOSTICS(STATS = NONE);
MODEL Test_2 = Test_1;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: Test_2

Number of Observations Read	221
Number of Observations Used	221

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			

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

Parameter Estimates

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

a. P value
    a. The P-value of ANOVA table is < 0.0001.
    b. Under the condition of the 95% confident interval, P-value is less than 0.0001 means the differences between the variances of the means are statistically significant. That is to say, using the linear regression is reasonable and our model looks good here.

B. P-value of the slope and intercept

a. The P-value of slope is < 0.0001,

b.The P-value of intercept is 0.9846.

c. Therefore, we can say the test_1 variable plays an important role in the linear regression model when predicting the test_2.

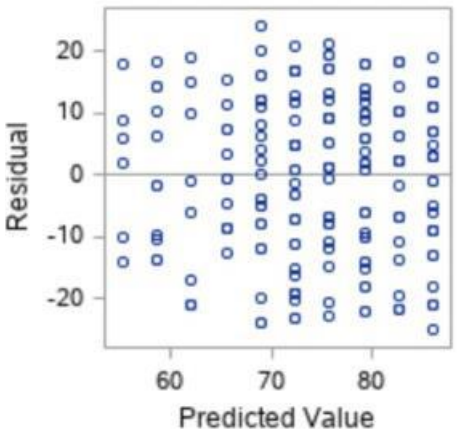
c. R^2 value

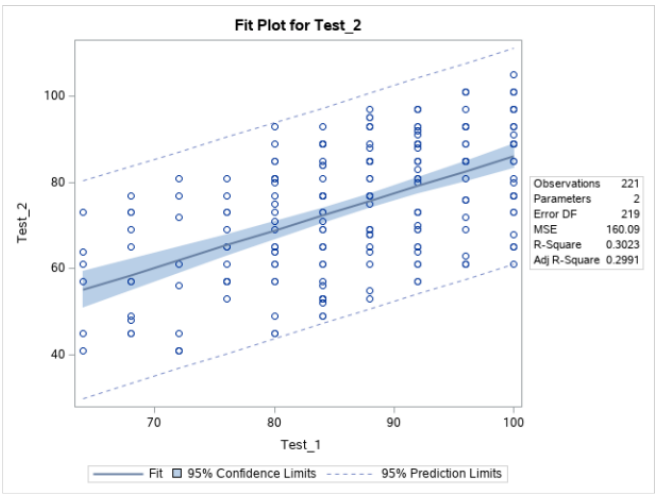
a. The R2 value is 0.3023.

b. This means only 30.23% of variation of y variable (test_2) is explained by x variable (test_1). This R2 value pointing that the model is not good.

d. Residual Analysis

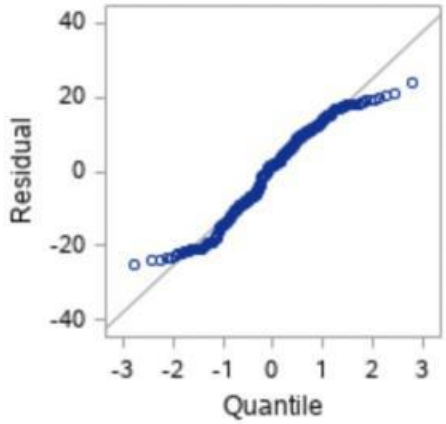
1. Fit and Residual





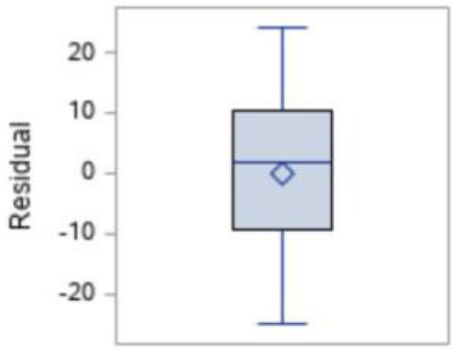
The plot shows that there is no telescoping and that the residual distribution is symmetrically centered on zero. This indicates that the standard deviation of the residuals is constant. Also, because there is no discernible patterning, we can infer the residual's independence. The model is then acceptable.

2. Probability Plot



The plot of residuals looks basically linear

3. Box plot



No outlier

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

We should not apply this model since, despite meeting the assumptions, the low R2 indicates that it is not dependable or relevant.