Advanced Computer Processing of Statistical Data

Homework 1

Allen Rahrooh

22 October 2019

Question 1

Part A)

```
DATA states1;
        INPUT state $ avgRain roadLength population;
        datalines;
FL 54.5 122391 19552860
GA 50.7 127492 9992167
AL 58.3 102018 4833722
NY 41.8 114800 19651127
MI 32.8 122284 9895622
TX 28.9 313596 26448193
VT 42.7 14238 626630
OR 27.4 73479 3930065
;
RUN;
```

Part B)

```
PROC PRINT data=states1;
RUN;
```

Obs	state	avgRain	roadLength	population
1	FL	54.5	122391	19552860
2	GA	50.7	127492	9992167
3	AL	58.3	102018	4833722
4	NY	41.8	114800	19651127
5	MI	32.8	122284	9895622
6	TX	28.9	313596	26448193
7	VT	42.7	14238	626630
8	OR	27.4	73479	3930065

Figure 1 Problem 1 Part B

Part C)

```
PROC PRINT data=states1;
TITLE 'Average Rain, Road Length and Population Per State';
RUN;
```

Average Rain, Road Length and Population Per State

Obs	state	avgRain	roadLength	population
1	FL	54.5	122391	19552860
2	GA	50.7	127492	9992167
3	AL	58.3	102018	4833722
4	NY	41.8	114800	19651127
5	MI	32.8	122284	9895622
6	TX	28.9	313596	26448193
7	VT	42.7	14238	626630
8	OR	27.4	73479	3930065

Figure 2 Problem 1 Part C

Part D)

```
PROC PRINT data=states1 NOOBS;
TITLE 'Average Rain, Road Length, and Population Per State';
RUN;
```

Average Rain, Road Length, and Population Per State

state	avgRain	roadLength	population
FL	54.5	122391	19552860
GA	50.7	127492	9992167
AL	58.3	102018	4833722
NY	41.8	114800	19651127
MI	32.8	122284	9895622
TX	28.9	313596	26448193
VT	42.7	14238	626630
OR	27.4	73479	3930065

Figure 3 Problem 1 Part D

Part E)

```
PROC PRINT data=states1 NOOBS;
FORMAT roadLength COMMA10.;
FORMAT population COMMA10.;
TITLE 'Average Rain, Road Length, and Population Per State';
RUN;
```

Average Rain, Road Length, and Population Per State

state	avgRain	roadLength	population
FL	54.5	122,391	19,552,860
GA	50.7	127,492	9,992,167
AL	58.3	102,018	4,833,722
NY	41.8	114,800	19,651,127
MI	32.8	122,284	9,895,622
TX	28.9	313,596	26,448,193
VT	42.7	14,238	626,630
OR	27.4	73,479	3,930,065

Figure 4 Problem 1 Part E

Part F)

```
DATA states2;
        INPUT state $ avgRain roadLength population;
        datalines;
HI 39.1 4439 1404054
KS 28.9 140476 2893957
MD 44.5 31984 5928814
MT 15.3 74983 1015165
OH 39.1 123297 11570808
;
RUN;
```

Part G)

```
DATA states3;
      SET states1 states2;
      FORMAT roadLength COMMA10.;
```

```
      FORMAT population COMMA10.;
RUN;
```

Part H)

```
PROC SORT DATA=states3;
       BY state;
RUN;

DATA states4;
       SET states3;
RUN;
```

Part I)

```
PROC PRINT data=states4 NOOBS;
       TITLE 'DATA SORTED BY STATE';
RUN;
```

DATA SORTED BY STATE

state	avgRain	roadLength	population
AL	58.3	102,018	4,833,722
FL	54.5	122,391	19,552,860
GA	50.7	127,492	9,992,167
HI	39.1	4,439	1,404,054
KS	28.9	140,476	2,893,957
MD	44.5	31,984	5,928,814
MI	32.8	122,284	9,895,622
MT	15.3	74,983	1,015,165
NY	41.8	114,800	19,651,127
OH	39.1	123,297	11,570,808
OR	27.4	73,479	3,930,065
TX	28.9	313,596	26,448,193
VT	42.7	14,238	626,630

Figure 5 Problem 1 Part I

Part J)

```
PROC MEANS DATA = states4;
```

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
avgRain	13	38.7692308	12.0396833	15.3000000	58.3000000
roadLength	13	105036.69	77402.51	4439.00	313596.00
population	13	9057168.00	8261099.96	626630.00	26448193.00

Figure 6 Problem 1 Part J

Part K)

```
PROC PRINT data = states4 NOOBS;
SUM avgRain roadLength population;
TITLE 'Average Rain, Road Length and Population Per State';
RUN;
```

Average Rain, Road Length and Population Per State

state	avgRain	roadLength	population
AL	58.3	102,018	4,833,722
FL	54.5	122,391	19,552,860
GA	50.7	127,492	9,992,167
HI	39.1	4,439	1,404,054
KS	28.9	140,476	2,893,957
MD	44.5	31,984	5,928,814
MI	32.8	122,284	9,895,622
MT	15.3	74,983	1,015,165
NY	41.8	114,800	19,651,127
OH	39.1	123,297	11,570,808
OR	27.4	73,479	3,930,065
TX	28.9	313,596	26,448,193
VT	42.7	14,238	626,630
	504.0	**1,365,477**	**117,743,184**

Figure 7 Problem 1 Part K

Part L)

```
DATA states5;
        SET states4;
        IF population < 1000000 THEN DELETE;
   IF avgRain < 20 THEN DELETE;
```

```
RUN;

PROC PRINT data = states5 NOOBS;
TITLE 'FILTERED DATA';
SUM avgRain roadLength population;
RUN;
```

FILTERED DATA

state	avgRain	roadLength	population
AL	58.3	102,018	4,833,722
FL	54.5	122,391	19,552,860
GA	50.7	127,492	9,992,167
HI	39.1	4,439	1,404,054
KS	28.9	140,476	2,893,957
MD	44.5	31,984	5,928,814
MI	32.8	122,284	9,895,622
NY	41.8	114,800	19,651,127
OH	39.1	123,297	11,570,808
OR	27.4	73,479	3,930,065
TX	28.9	313,596	26,448,193
	446.0	**1,276,256**	**116,101,389**

Figure 8 Problem 1 Part L

The states that got removed are Montana (MT) and Vermont (VT).

Part M)

```
DATA states6;
      SET states5;
      roadLengthKM = roadLength * 1.61;
RUN;

PROC PRINT data = states6 NOOBS;
FORMAT roadLengthKM COMMA14.;
TITLE 'UPDATED DATA WITH KM CONVERSION';
SUM avgRain roadLength population roadLengthKM;
RUN;
```

UPDATED DATA WITH KM CONVERSION

state	avgRain	roadLength	population	roadLengthKM
AL	58.3	102,018	4,833,722	164,249
FL	54.5	122,391	19,552,860	197,050
GA	50.7	127,492	9,992,167	205,262
HI	39.1	4,439	1,404,054	7,147
KS	28.9	140,476	2,893,957	226,166
MD	44.5	31,984	5,928,814	51,494
MI	32.8	122,284	9,895,622	196,877
NY	41.8	114,800	19,651,127	184,828
OH	39.1	123,297	11,570,808	198,508
OR	27.4	73,479	3,930,065	118,301
TX	28.9	313,596	26,448,193	504,890
	446.0	**1,276,256**	**116,101,389**	**2,054,772**

Figure 9 Problem 1 Part M

Part N)

```
PROC CORR DATA = states6 PEARSON;
RUN;
```

The CORR Procedure

4 Variables:	avgRain roadLength population roadLengthKM

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
avgRain	11	40.54545	10.66999	446.00000	27.40000	58.30000
roadLength	11	116023	78370	1276256	4439	313596
population	11	10554672	8115010	116101389	1404054	26448193
roadLengthKM	11	186797	126176	2054772	7147	504890

Pearson Correlation Coefficients, N = 11
Prob > |r| under H0: Rho=0

	avgRain	roadLength	population	roadLengthKM
avgRain	1.00000	-0.27520 0.4128	0.02249 0.9477	-0.27520 0.4128
roadLength	-0.27520 0.4128	1.00000	0.74965 0.0079	1.00000 <.0001
population	0.02249 0.9477	0.74965 0.0079	1.00000	0.74965 0.0079
roadLengthKM	-0.27520 0.4128	1.00000 <.0001	0.74965 0.0079	1.00000

Figure 10 Problem 1 Part N

Why are the diagonal elements value1?

The diagonal elements are value 1 because it represents the correlation between the variable and itself.

Which variables have the largest correlation with the population variable?

The variables roadLength and roadLengthKM have the highest correlation of 0.74965.

Question 2

Part A)

```
DATA drinks1;
INPUT name $ drink $;
CARDS;

alex coffee
bob tea
cat coffee
debra tea
eric tea
fred tea
greg tea
heather coffee
irene coffee
jack tea
karen tea
laura coffee
mark tea
;
RUN;
```

Part B)

```
DATA drinksmore;
INPUT name $ sugar $ gender $;
CARDS;

alex yes m
bob yes m
cat no f
debra no f
eric no m
fred no m
```

```
greg yes m
heather no f
irene no f
jack no m
karen yes f
laura no f
mark no m
;
RUN;
```

Part C)

```
DATA drinks2;
SET drinks1;
SET drinksmore;
PROC PRINT DATA = drinks2 NOOBS;
RUN;
```

COMBINED DATA

name	drink	sugar	gender
alex	coffee	yes	m
bob	tea	yes	m
cat	coffee	no	f
debra	tea	no	f
eric	tea	no	m
fred	tea	no	m
greg	tea	yes	m
heather	coffee	no	f
irene	coffee	no	f
jack	tea	no	m
karen	tea	yes	f
laura	coffee	no	f
mark	tea	no	m

Figure 11 Problem 1 Part C

Which table operation is being performed?

The table operation being performed is column concatenation.

Part D)

```
PROC FREQ DATA = drinks2;
TABLES drink;
RUN;
```

The FREQ Procedure

drink	Frequency	Percent	Cumulative Frequency	Cumulative Percent
coffee	5	38.46	5	38.46
tea	8	61.54	13	100.00

Figure 12 Problem 1 Part D

Part E)

```
PROC FREQ DATA = drinks2;
TABLES drink*sugar;
RUN;
```

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of drink by sugar			
	sugar		
drink	no	yes	Total
coffee	4 30.77 80.00 44.44	1 7.69 20.00 25.00	5 38.46
tea	5 38.46 62.50 55.56	3 23.08 37.50 75.00	8 61.54
Total	9 69.23	4 30.77	13 100.00

Figure 13 Problem 1 Part E

Part F)

What comments can you make on the data displayed in the cross tabulation?

From the cross tabulation of the drink and sugar variables we see that most people prefer tea over coffee and that they prefer not to have sugar in their coffee or tea.

Part G)

```
PROC FREQ DATA = drinks2;
TABLES drink*sugar*gender;
RUN;
```

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table 1 of sugar by gender Controlling for drink=coffee		
	gender		
sugar	f	m	Total
no	4 80.00 100.00 100.00	0 0.00 0.00 0.00	4 80.00
yes	0 0.00 0.00 0.00	1 20.00 100.00 100.00	1 20.00
Total	4 80.00	1 20.00	5 100.00

Frequency Percent Row Pct Col Pct	Table 2 of sugar by gender Controlling for drink=tea		
	gender		
sugar	f	m	Total
no	1 12.50 20.00 50.00	4 50.00 80.00 66.67	5 62.50
yes	1 12.50 33.33 50.00	2 25.00 66.67 33.33	3 37.50
Total	2 25.00	6 75.00	8 100.00

Figure 14 Problem 1 Part G

What new information does this variable tell you about the data?

Adding the gender information to the cross tabulation of drink and sugar shows us which gender likes tea or coffee with or without sugar. So we see that females do not prefer sugar in their coffee and males mostly do not have sugar with their tea.

Question 3

Part A)

```
DATA students1;
       INPUT NAME $ ID DEGREE $14.;
       DATALINES;
alex 2210 stats
bob 1121 physics
cat 5412 stats
debra 1981 history
eric 1990 moder-history
fred 2211 Statistics
greg 1221 Physics
heather 7221 education
igor 7112 Phys-ed
;
```

```
RUN;
```

Part B)

```
DATA students1;
        INPUT NAME $ ID DEGREE $14.;
        DATALINES;
alex 2210 Statistics
bob 1121 Physics
cat 5412 Statistics
debra 1981 History
eric 1990 History
fred 2211 Statistics
greg 1221 Physics
heather 7221 Education
igor 7112 Education
;
RUN;

PROC PRINT DATA=students1 NOOBS;
TITLE 'UPDATED DATA BY SUBJECT';
```

UPDATED DATA BY SUBJECT

NAME	ID	DEGREE
alex	2210	Statistics
bob	1121	Physics
cat	5412	Statisitics
debra	1981	History
eric	1990	History
fred	2211	Statistics
greg	1221	Physics
heather	7221	Education
igor	7112	Education

Figure 15 Problem 3 Part B

Part C)

```
DATA grades;
        INPUT ID RESULTS $;
        DATALINES;

2210 B
2210 B
2210 B
```

```
1121 A
1121 B
5412 A
5412 B
5412 C
1990 D
2211 A
2211 A
2211 C
1221 A
1221 B
7221 C
;
```

Part D)

Which method to merge data?

To merge the data sets students1 and grades I used the MERGE keyword once both datasets are sorted by ID then I dropped the names column.

```
PROC SORT DATA = students1;
BY ID;
RUN;

PROC SORT DATA = grades;
BY ID;
RUN;

DATA students3;
        MERGE students1 grades;
        BY ID;
        DROP NAME;
        RUN;

PROC PRINT DATA = students3 NOOBS;
TITLE 'MERGED DATA';
RUN;
```

MERGED DATA

ID	DEGREE	RESULTS
1121	Physics	A
1121	Physics	B
1221	Physics	A
1221	Physics	B
1981	History	
1990	History	D
2210	Statistics	B
2210	Statistics	B
2210	Statistics	B
2211	Statistics	A
2211	Statistics	A
2211	Statistics	C
5412	Statisitics	A
5412	Statisitics	B
5412	Statisitics	C
7112	Education	
7221	Education	C

Figure 16 Problem 3 Part D

Part E)

Which statement will print us the id entries of the students with no grades?
And all other associated info

The WHERE and IF keywords will print the id entries of students with no grades

Part F)

```
DATA students3A;
       SET students3;
       WHERE RESULTS = " ";
       IF RESULTS = " " THEN RESULTS = "F";

RUN;
```

Part G)

```
PROC SORT DATA = students3;
BY ID;
RUN;

PROC SORT DATA = students3A;
BY ID;
RUN;

DATA students4;
UPDATE students3 students3A;
BY ID;
DROP NAME;
RUN;

PROC PRINT DATA = students4 NOOBS;
RUN;
```

MERGED DATA WITH UPDATED MISSING VALUES

ID	DEGREE	RESULTS
1121	Physics	A
1121	Physics	B
1221	Physics	A
1221	Physics	B
1981	History	F
1990	History	D
2210	Statistics	B
2210	Statistics	B
2210	Statistics	B
2211	Statistics	A
2211	Statistics	A
2211	Statistics	C
5412	Statisitics	A
5412	Statisitics	B
5412	Statisitics	C
7112	Education	F
7221	Education	C

Figure 17 Problem 3 Part G

Part H)

```
PROC FREQ DATA = students4;
TABLES DEGREE*RESULTS;
TITLE 'FREQUENCY TABLE OF COURSES AND GRADES'
RUN;
```

FREQUENCY TABLE OF COURSES AND GRADES

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of DEGREE by RESULTS

DEGREE	RESULTS					
	A	B	C	D	F	Total
Education	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 5.88 50.00 33.33	0 0.00 0.00 0.00	1 5.88 50.00 50.00	2 11.76
History	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 5.88 50.00 100.00	1 5.88 50.00 50.00	2 11.76
Physics	2 11.76 50.00 40.00	2 11.76 50.00 33.33	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4 23.53
Statisitics	1 5.88 33.33 20.00	1 5.88 33.33 16.67	1 5.88 33.33 33.33	0 0.00 0.00 0.00	0 0.00 0.00 0.00	3 17.65
Statistics	2 11.76 33.33 40.00	3 17.65 50.00 50.00	1 5.88 16.67 33.33	0 0.00 0.00 0.00	0 0.00 0.00 0.00	6 35.29
Total	5 29.41	6 35.29	3 17.65	1 5.88	2 11.76	17 100.00

Figure 18 Problem 3 Part H

We see from the PROC FREQ cross tabulation that History was the worst performer with only D & F grades. The best performer was Statistics with 7 grades above a C and the second best performing course was physics with 4 grades above a C.

Part I)

```
DATA students5(KEEP = DEGREE RESULTSNUM);
SET students4;
IF RESULTS = "A" THEN RESULTSNUM = 4;
ELSE IF RESULTS = "B" THEN RESULTSNUM  = 3;
```

```
ELSE IF RESULTS = "C" THEN RESULTSNUM = 2;
ELSE IF RESULTS = "D" THEN RESULTSNUM = 1;
ELSE IF RESULTS = "F" THEN RESULTSNUM = 0;
RUN;


PROC PRINT DATA = students5 NOOBS;
TITLE 'NUMERICAL GRADES DATA BY ID';
RUN;
```

NUMERICAL GRADES DATA BY ID

DEGREE	RESULTSNUM
Physics	4
Physics	3
Physics	4
Physics	3
History	0
History	1
Statistics	3
Statistics	3
Statistics	3
Statistics	4
Statistics	4
Statistics	2
Statisitics	4
Statisitics	3
Statisitics	2
Education	0
Education	2

Figure 19 Problem 3 Part I

Part J)

```
PROC MEANS DATA = students5;
CLASS DEGREE;
RUN;
```

The MEANS Procedure

Analysis Variable : RESULTSNUM						
DEGREE	N Obs	N	Mean	Std Dev	Minimum	Maximum
Education	2	2	1.0000000	1.4142136	0	2.0000000
History	2	2	0.5000000	0.7071068	0	1.0000000
Physics	4	4	3.5000000	0.5773503	3.0000000	4.0000000
Statisitics	3	3	3.0000000	1.0000000	2.0000000	4.0000000
Statistics	6	6	3.1666667	0.7527727	2.0000000	4.0000000

Figure 20 Problem 3 Part J

What conclusions can you make of this data?

I can conclude that physics and statistics are the best performing courses in terms of grade because the means are above a 3 (B) and the worst performer is history with mean of 0.5 which is below a 1.0 (D).

How does it differ from the previous use of proc freq?

The PROC FREQ tables outputs the relative frequency of each course and each grade. Whereas the PROC MEANS gives the mean, standard deviation, minimum, and maximum of each degree in terms of numeric grade.

Question 4

Part A)

```
DATA mystery;
INPUT x1  x2  x3  x4  y1  y2  y3  y4;
DATALINES;
10 10 10 8 8.04 9.14 7.46 6.58
8 8 8 8 6.95 8.14 6.77 5.76
13 13 13 8 7.58 8.74 12.74 7.71
9 9 9 8 8.81 8.77 7.11 8.84
11 11 11 8 8.33 9.26 7.81 8.47
14 14 14 8 9.96 8.1 8.84 7.04
6 6 6 8 7.24 6.13 6.08 5.25
4 4 4 19 4.26 3.1 5.39 12.5
12 12 12 8 10.84 9.13 8.15 5.56
7 7 7 8 4.82 7.26 6.42 7.91
5 5 5 8 5.68 4.74 5.73 6.89
;
RUN;
```

```
PROC MEANS DATA = mystery;
        VAR x1 x2 x3 x4 y1 y2 y3 y4;
RUN;
```

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
x1	11	9.0000000	3.3166248	4.0000000	14.0000000
x2	11	9.0000000	3.3166248	4.0000000	14.0000000
x3	11	9.0000000	3.3166248	4.0000000	14.0000000
x4	11	9.0000000	3.3166248	8.0000000	19.0000000
y1	11	7.5009091	2.0315681	4.2600000	10.8400000
y2	11	7.5009091	2.0316567	3.1000000	9.2600000
y3	11	7.5000000	2.0304236	5.3900000	12.7400000
y4	11	7.5009091	2.0305785	5.2500000	12.5000000

Figure 21 Problem 4 Part A

From the PROC MEANS output of the mystery data we see that the means of x1,x2,x3,x4 are the same value at 9.0 and the means of y1,y2,y3, and y4 are close to the same value of 7.5. This is interesting because x4 has a higher minimum and maximum value then x2,x3, and x4. The same for the y variables they all have different minimum and maximum values but similar means.

```
PROC SGPLOT DATA = mystery;
  SCATTER x=x1 y=y1;
  TITLE 'SCATTER PLOT OF X1,Y1';
RUN;
```

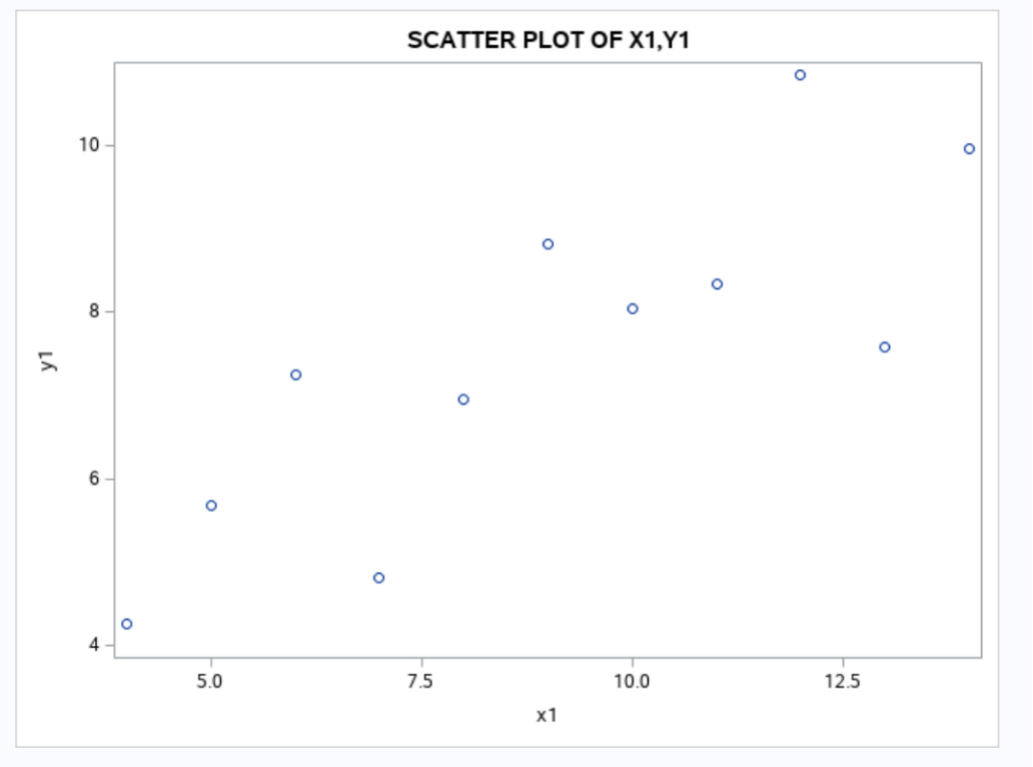

Figure 22 Scatter Plot of X1,Y1

Looking at the scatter plot for the regression pair (x1,y1) it looks like the R-Squared will be close to one.

```
PROC SGPLOT DATA = mystery;
  SCATTER x=x2 y=y2;
  TITLE 'SCATTER PLOT OF X2,Y2';
RUN;
```

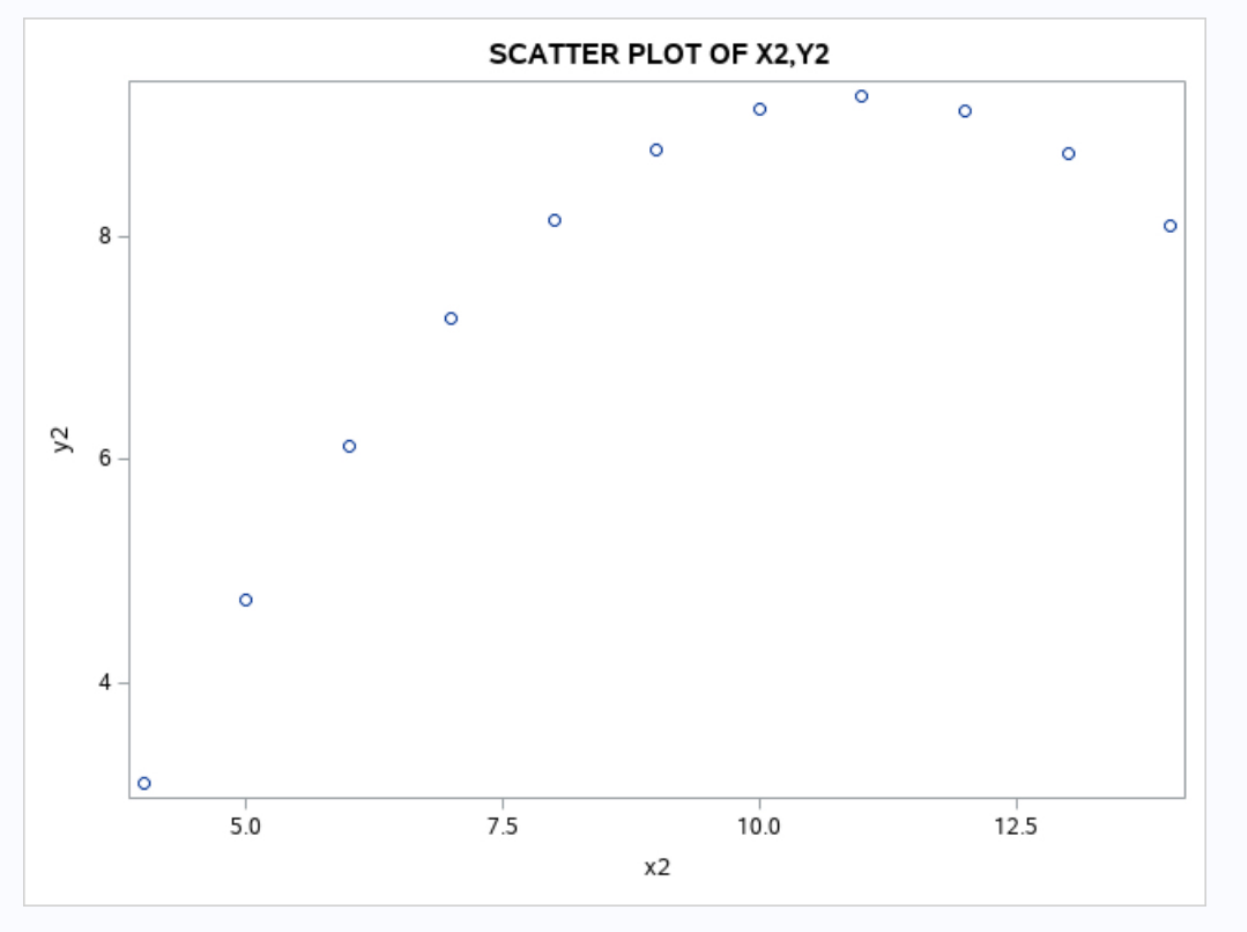

Figure 23 Scatter Plot of X2,Y2

Looking at the scatter plot for the regression pair (x2,y2) it looks like the R-Squared will be close to 0.5 because most of the data is in the top diagonal of the scatter plot.

```
PROC SGPLOT DATA = mystery;
  SCATTER x=x3 y=y3;
  TITLE 'SCATTER PLOT OF X3,Y3';
RUN;
```

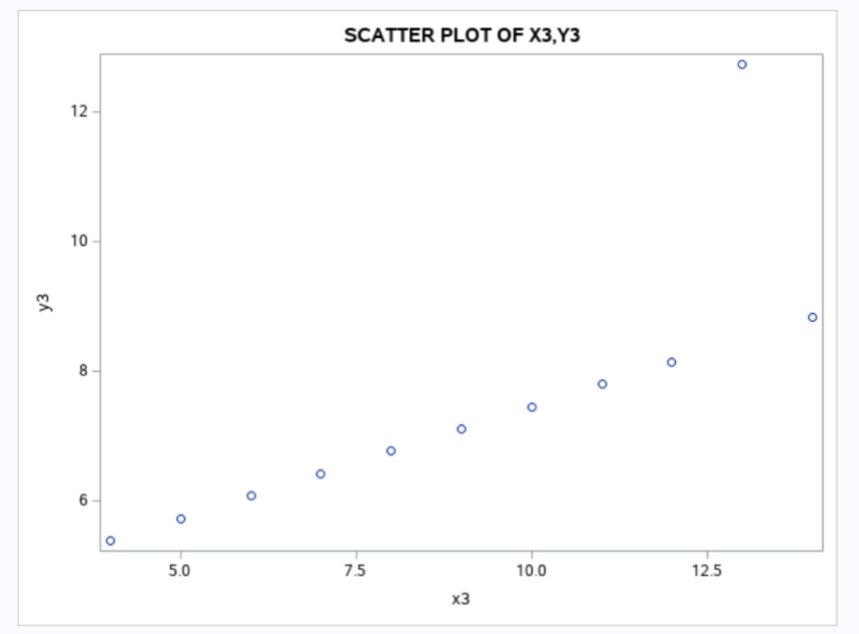

Figure 24 Scatter Plot of X3,Y3

Looking at the scatter plot for the regression pair (x3,y3) it looks like the R-Squared will be greater than 0.95 because the data looks very linear compared to the regression pairs (x1,y1) and (x2,y2) scatter plots.

```
PROC SGPLOT DATA = mystery;
  SCATTER x=x4 y=y4;
  TITLE 'SCATTER PLOT OF X4,Y4';
RUN;
```

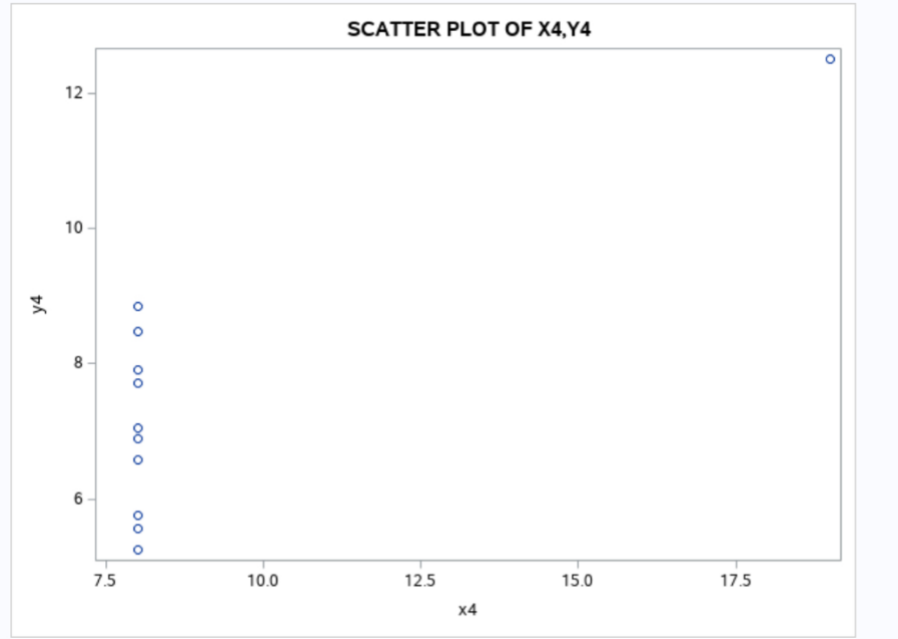

Figure 25 Scatter Plot of X4,Y4

Looking at the scatter plot for the regression pair (x4,y4) it looks like the R-Squared will be close to zero since all the data is horizontally stacked besides one outlier at about x4 = 19.

Part B)

```
PROC REG DATA = mystery;
        MODEL y1 = x1;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: y1

Number of Observations Read	11
Number of Observations Used	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27.51000	27.51000	17.99	0.0022
Error	9	13.76269	1.52919		
Corrected Total	10	41.27269			

Root MSE	1.23660	R-Square	0.6665
Dependent Mean	7.50091	Adj R-Sq	0.6295
Coeff Var	16.48605		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.00009	1.12475	2.67	0.0257
x1	1	0.50009	0.11791	4.24	0.0022

Figure 26 Regression Statistics for Pair X1,Y1

From the regression statistics we see that the associated p value for the intercept is 0.0257 with t Value 2.67 and for x1 it is 0.0022 with t Value 4.24.

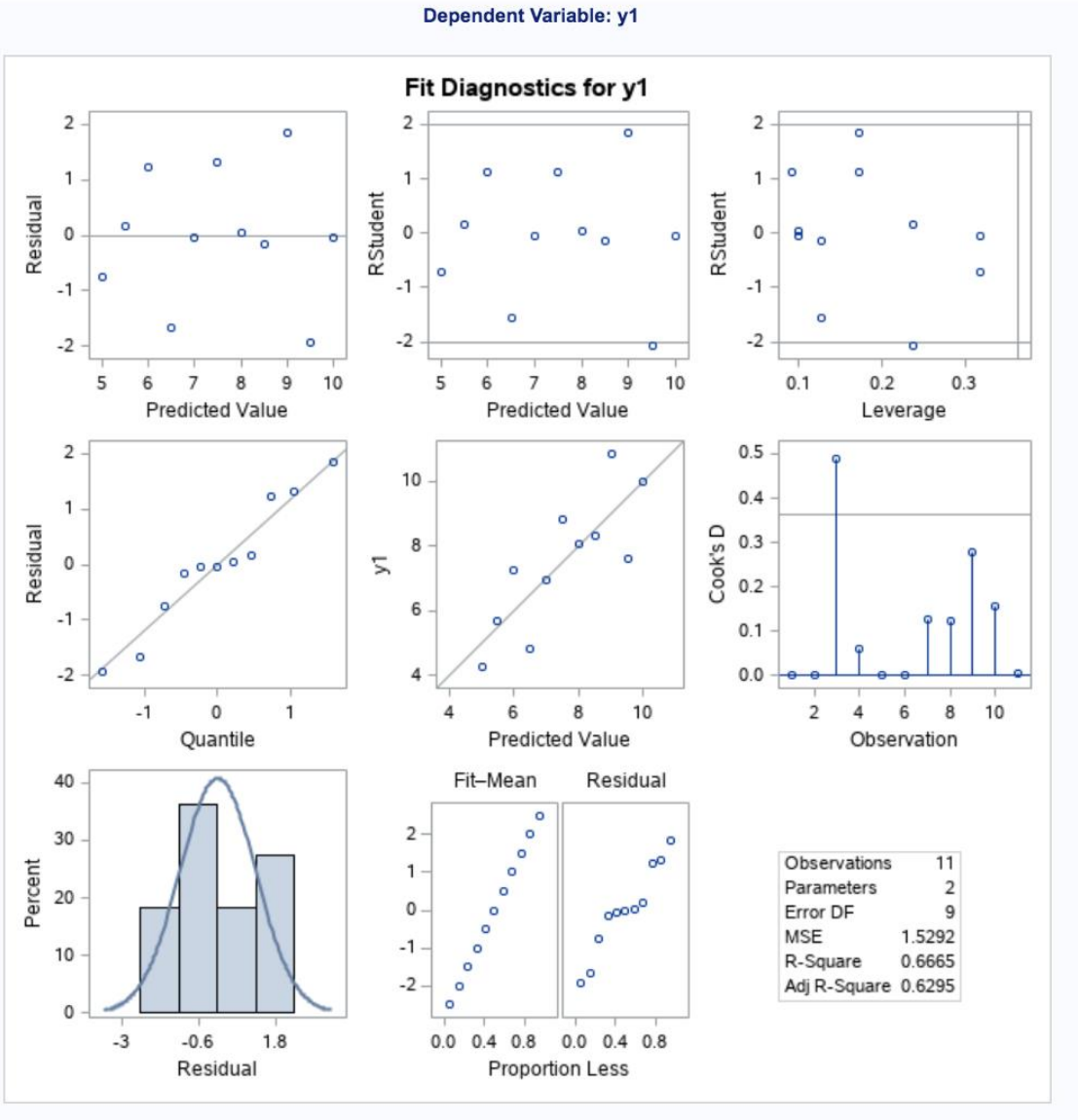

Figure 27 Regression Models for Pair X1,Y1

Looking at the fit diagnostics for (x1,y1) the R-Squared value is 0.6665.

```
PROC REG DATA = mystery;
        MODEL y2 = x2;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: y2

Number of Observations Read	11
Number of Observations Used	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27.50000	27.50000	17.97	0.0022
Error	9	13.77629	1.53070		
Corrected Total	10	41.27629			

Root MSE	1.23721	R-Square	0.6662
Dependent Mean	7.50091	Adj R-Sq	0.6292
Coeff Var	16.49419		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.00091	1.12530	2.67	0.0258
x2	1	0.50000	0.11796	4.24	0.0022

Figure 28 Regression Statistics for Pair X2,Y2

From the regression statistics we see that the associated p value for the intercept is 0.0258 with t Value 2.67 and for x2 it is 0.0022 with t Value 4.24.

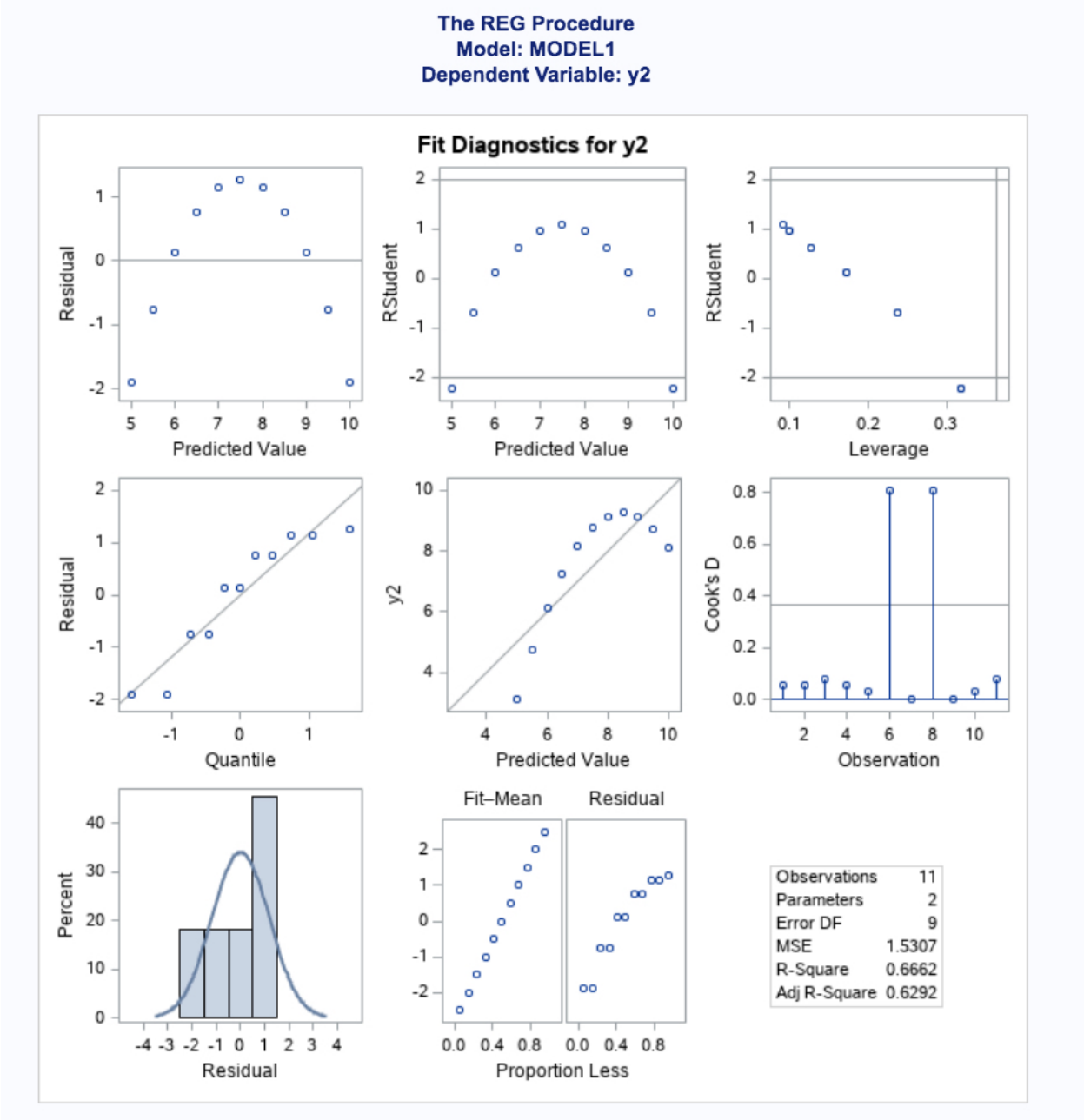

Figure 29 Regression Models for Pair X2,Y2

Looking at the fit diagnostics for (x2,y2) the R-Squared value is 0.6662 which is relatively close to the regression model of (x1,y1).

```
PROC REG DATA = mystery;
        MODEL y3 = x3;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: y3

Number of Observations Read	11
Number of Observations Used	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27.47001	27.47001	17.97	0.0022
Error	9	13.75619	1.52847		
Corrected Total	10	41.22620			

Root MSE	1.23631	R-Square	0.6663
Dependent Mean	7.50000	Adj R-Sq	0.6292
Coeff Var	16.48415		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.00245	1.12448	2.67	0.0256
x3	1	0.49973	0.11788	4.24	0.0022

Figure 30 Regression Statistics for Pair X3,Y3

From the regression statistics we see that the associated p value for the intercept is 0.0256 with t Value 2.67 and for x3 it is 0.0022 with t Value 4.24.

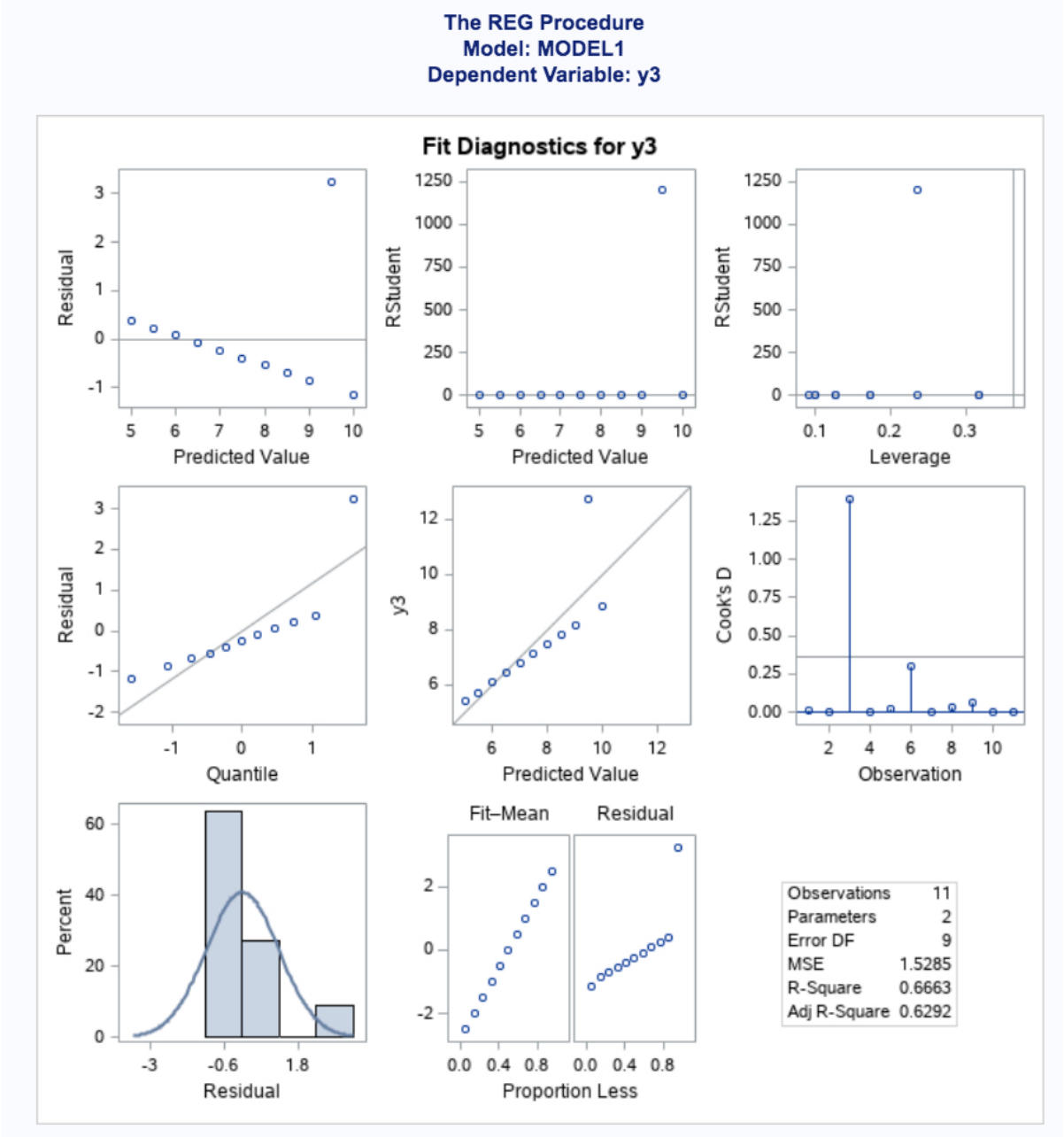

Figure 31 Regression Models for Pair X3,Y3

Looking at the fit diagnostics for (x3,y3) the R-Squared value is 0.6663 which is relatively close to the regression model of (x1,y1) and (x2,y2).

```
PROC REG DATA = mystery;
        MODEL y4 = x4;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: y4

Number of Observations Read	11
Number of Observations Used	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27.49000	27.49000	18.00	0.0022
Error	9	13.74249	1.52694		
Corrected Total	10	41.23249			

Root MSE	1.23570	R-Square	0.6667
Dependent Mean	7.50091	Adj R-Sq	0.6297
Coeff Var	16.47394		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.00173	1.12392	2.67	0.0256
x4	1	0.49991	0.11782	4.24	0.0022

Figure 32 Regression Statistics for Pair X4,Y4

From the regression statistics we see that the associated p value for the intercept is 0.0256 with t Value 2.67 and for x4 it is 0.0022 with t Value 4.24.

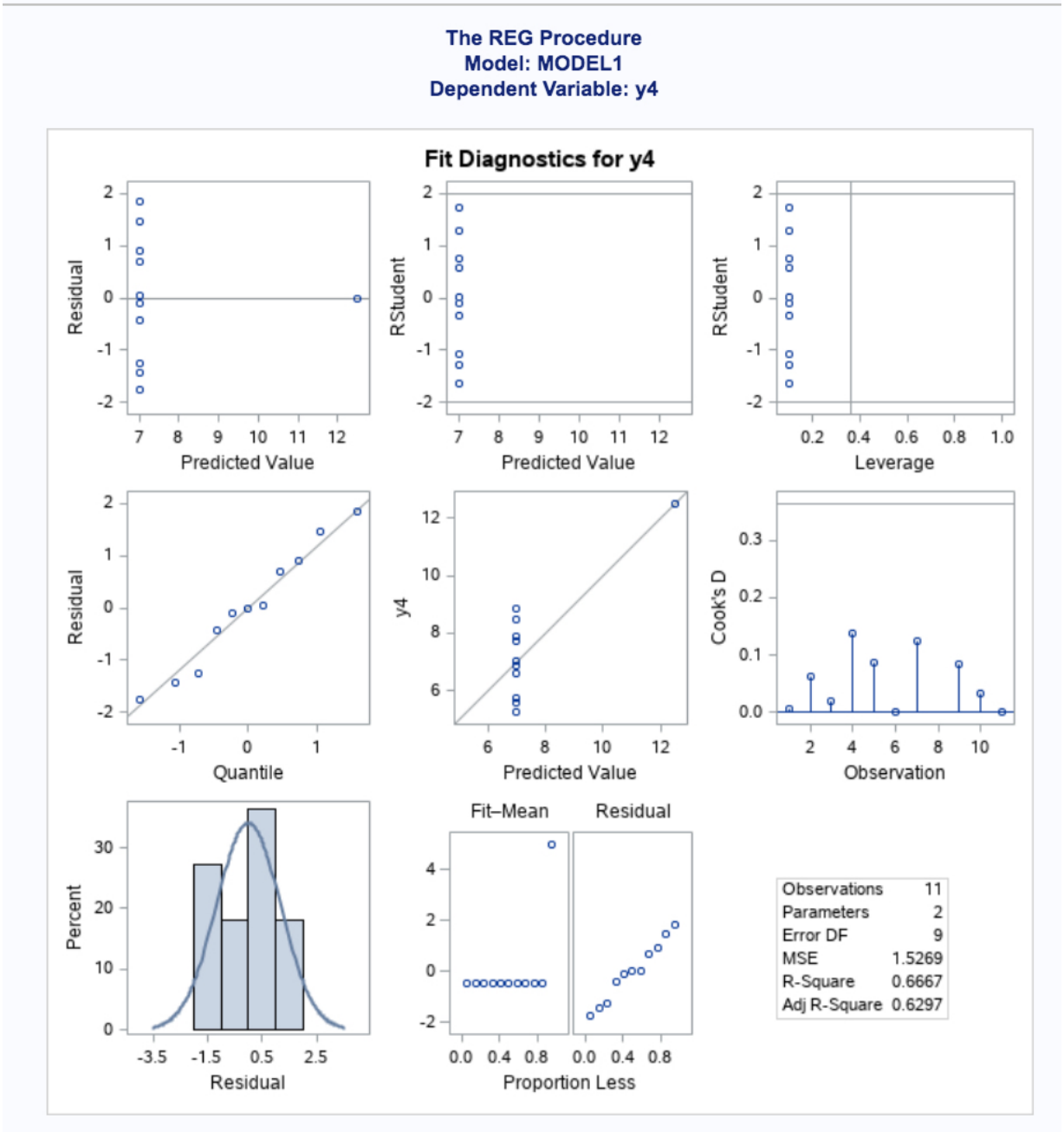

Figure 33 Regression Models for Pair X4,Y4

Looking at the fit diagnostics for (x4,y4) the R-Squared value is 0.667 which is the same as the regression models of (x1,y1),(x2,y2), and (x3,y3). It is interesting to see that given the different scatter plots of the regression pairs that they all have relatively the same R-Square value.

Part C)

What conclusion can you arrive from this exploration in terms of the suitability of descriptive statistics and regression in terms of data exploration?

I can conclude that the descriptive statistics give a baseline of analysis without having to do a regression analysis you can quickly observe the descriptive statistics and describe the data. If the descriptive statistics is not providing a clear outcome like the mystery dataset then doing the scatter plots and regression analysis can help better understand the data.

What is the recommendation that you would provide further data explorations to necessarily include as a results?

The recommendation that I would provide is to combine the variables x1,x2,x3,x4 as a single variable (X) and the variables y1,y2,y3,y4 as a single variable (Y) and then do a scatter plot, descriptive statistics, and the regression model to see if a better fit is possible since the R-Squared value for the regression pairs (x1,y1),(x2,y2),(x3,y3) and (x4,y4) all had the same R-Squared value around 0.6667.

Part D)

Can you include a simple yet effective solution to differentiate these associations included in the dataset?

To differentiate the associations I would do different regression pairs like (x1,y4) or (x2,y3) this will provide more regression models that may have a better fit then (xi,yi) models.

Question 5

```
DATA carprice;
        INPUT TYPE $ PRICE MAX_PRICE RANGE_PRICE ROUGH_RANGE GPM100 MPG_CITY
                MPG_HIGHWAY;
        DATALINES;
Midsize         14.2 15.7 17.3 3.1 3.09 3.8 22 31
Large 19.9 20.8 21.7 1.8 1.79 4.2 19 28
Large 22.6 23.7 24.9 2.3 2.31 4.9 16 25
Midsize 26.3 26.3 26.3 0 -0.01 4.3 19 27
Large 33 34.7 36.3 3.3 3.3 4.9 16 25
Midsize 37.5 40.1 42.7 5.2 5.18 4.9 16 25
Compact 8.5 13.4 18.3 9.8 9.8 3.3 25 36
Compact 11.4 11.4 11.4 0 -0.01 3.4 25 34
Sporty 13.4 15.1 16.8 3.4 3.38 4.2 19 28
Midsize 13.4 15.9 18.4 5 5.01 4 21 29
Van 14.7 16.3 18 3.3 3.31 4.9 18 23
Van 14.7 16.6 18.6 3.9 3.9 5.7 15 20
Large 18 18.8 19.6 1.6 1.6 4.7 17 26
Sporty 34.6 38 41.5 6.9 6.88 4.8 17 25
Large 18.4 18.4 18.4 0 -0.01 4.2 20 28
```

```
Compact 14.5 15.8 17.1 2.6 2.59 3.9 23 28
Large 29.5 29.5 29.5 0 0.02 4.3 20 26
Small 7.9 9.2 10.6 2.7 2.68 3.2 29 33
Small 8.4 11.3 14.2 5.8 5.8 3.8 23 29
Compact 11.9 13.3 14.7 2.8 2.81 4.1 22 27
Van 13.6 19 24.4 10.8 10.77 5.3 17 21
Midsize 14.8 15.6 16.4 1.6 1.6 4.2 21 27
Sporty 18.5 25.8 33.1 14.6 14.6 4.8 18 24
Small 7.9 12.2 16.5 8.6 8.6 3.2 29 33
Large 17.5 19.3 21.2 3.7 3.69 4.2 20 28
Small 6.9 7.4 7.9 1 1 3.1 31 33
Small 8.4 10.1 11.9 3.5 3.49 3.8 23 30
Compact 10.4 11.3 12.2 1.8 1.82 4.1 22 27
Sporty 10.8 15.9 21 10.2 10.21 3.9 22 29
Sporty 12.8 14 15.2 2.4 2.4 3.7 24 30
Van 14.5 19.9 25.3 10.8 10.82 5.7 15 20
Midsize 15.6 20.2 24.8 9.2 9.21 3.9 21 30
Large 20.1 20.9 21.7 1.6 1.59 4.5 18 26
Midsize 33.3 34.3 35.3 2 1.99 4.7 17 26
Large 34.4 36.1 37.8 3.4 3.42 4.5 18 26
Sporty 13.3 14.1 15 1.7 1.71 4.1 23 26
Midsize 14.9 14.9 14.9 0 -0.02 4.4 19 26
Compact 13 13.5 14 1 0.99 3.6 24 31
Midsize 14.2 16.3 18.4 4.2 4.19 3.7 23 31
Van 19.5 19.5 19.5 0 0 4.9 18 23
Large 19.5 20.7 21.9 2.4 2.41 4.2 19 28
Sporty 11.4 14.4 17.4 6 6.01 3.8 23 30
Small 8.2 9 9.9 1.7 1.69 2.8 31 41
Compact 9.4 11.1 12.8 3.4 3.39 3.7 23 31
Sporty 14 17.7 21.4 7.4 7.4 4.2 19 28
Midsize 15.4 18.5 21.6 6.2 6.19 4.3 19 27
Large 19.4 24.4 29.4 10 10 4.2 19 28
Small 9.2 11.1 12.9 3.7 3.7 3 28 38
;
```

Part A)

```
PROC TTEST DATA = carprice;
WHERE TYPE IN ("Large", "Sporty");
CLASS TYPE;
VAR PRICE;
RUN;
```

The TTEST Procedure

Variable: PRICE

TYPE	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Large		11	22.9364	6.2607	1.8877	17.5000	34.4000
Sporty		8	16.1000	7.8251	2.7666	10.8000	34.6000
Diff (1-2)	Pooled		6.8364	6.9476	3.2283		
Diff (1-2)	Satterthwaite		6.8364		3.3492		

TYPE	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Large		22.9364	18.7304	27.1424	6.2607	4.3745	10.9871
Sporty		16.1000	9.5581	22.6419	7.8251	5.1737	15.9261
Diff (1-2)	Pooled	6.8364	0.0253	13.6475	6.9476	5.2134	10.4155
Diff (1-2)	Satterthwaite	6.8364	-0.3961	14.0688			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	17	2.12	0.0492
Satterthwaite	Unequal	13.054	2.04	0.0620

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	7	10	1.56	0.5038

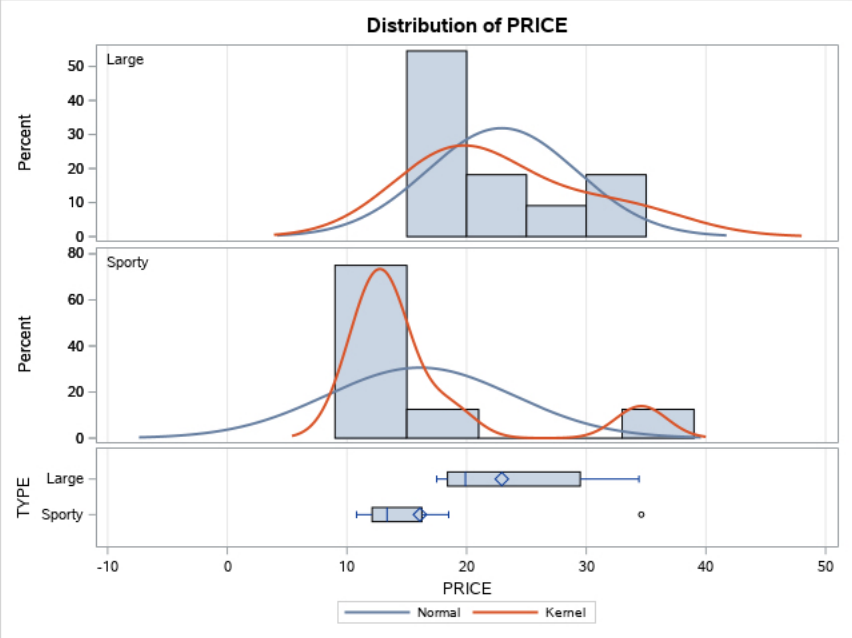

Figure 34 Problem 5 Part A T-Test Large/Sporty

For the T-test between the large and sporty cars the means are 22.93 for large and 16.10 for sporty vehicles this means we do not accept the null hypothesis of equal means.

```
PROC TTEST DATA = carprice;
WHERE TYPE IN ("Van", "Large");
```

```
CLASS TYPE;
VAR PRICE;
RUN;
```

The TTEST Procedure

Variable: PRICE

TYPE	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Large		11	22.9364	6.2607	1.8877	17.5000	34.4000
Van		5	15.4000	2.3367	1.0450	13.6000	19.5000
Diff (1-2)	Pooled		7.5364	5.4367	2.9323		
Diff (1-2)	Satterthwaite		7.5364		2.1576		

TYPE	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Large		22.9364	18.7304	27.1424	6.2607	4.3745	10.9871
Van		15.4000	12.4986	18.3014	2.3367	1.4000	6.7145
Diff (1-2)	Pooled	7.5364	1.2471	13.8256	5.4367	3.9803	8.5742
Diff (1-2)	Satterthwaite	7.5364	2.9032	12.1696			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	14	2.57	0.0222
Satterthwaite	Unequal	13.823	3.49	0.0036

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	10	4	7.18	0.0725

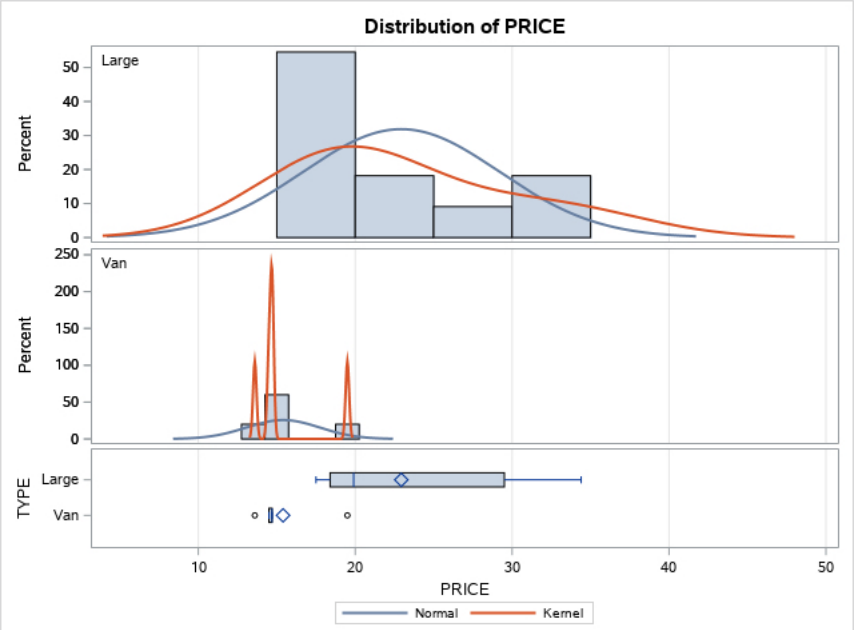

Figure 35 Problem 5 Part A T-Test Van/Large

For the T-test between the large and van cars the means are 22.93 for large and 15.40 for van vehicles this means we do not accept the null hypothesis of equal means.

```
PROC TTEST DATA = carprice;
WHERE TYPE IN ("Small", "Midsize");
CLASS TYPE;
VAR PRICE;
RUN;
```

The TTEST Procedure

Variable: PRICE

TYPE	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Midsize		9	20.6000	9.2895	3.0965	13.4000	37.5000
Small		7	8.1286	0.6969	0.2634	6.9000	9.2000
Diff (1-2)	Pooled		12.4714	7.0370	3.5463		
Diff (1-2)	Satterthwaite		12.4714		3.1077		

TYPE	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Midsize		20.6000	13.4595	27.7405	9.2895	6.2747	17.7966
Small		8.1286	7.4840	8.7731	0.6969	0.4491	1.5347
Diff (1-2)	Pooled	12.4714	4.8653	20.0775	7.0370	5.1520	11.0981
Diff (1-2)	Satterthwaite	12.4714	5.3228	19.6200			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	14	3.52	0.0034
Satterthwaite	Unequal	8.1156	4.01	0.0038

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	8	6	177.67	<.0001

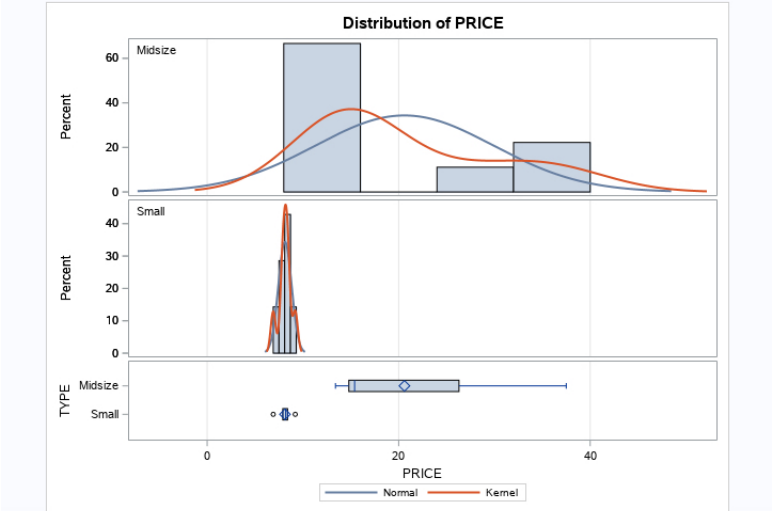

Figure 36 Problem 5 Part A T-Test Small/MIdsize

For the T-test between the midsize and small cars the means are 20.6 for midsize and 8.12 for sporty vehicles this means we do not accept the null hypothesis of equal means.

Part B)

```
PROC REG DATA = carprice alpha = 0.05;
        MODEL GPM100 = PRICE;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: GPM100

Number of Observations Read	48
Number of Observations Used	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6.39722	6.39722	0.53	0.4723
Error	46	560.20675	12.17841		
Corrected Total	47	566.60397			

Root MSE	3.48976	R-Square	0.0113
Dependent Mean	4.10417	Adj R-Sq	-0.0102
Coeff Var	85.02963		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.88822	1.19332	4.10	0.0002
PRICE	1	-0.04733	0.06530	-0.72	0.4723

Figure 37 Problem 5 Part B Linear Regression

From the PROC REG of the linear regression model GPM100 = PRICE we get a p value of 0.0002 with t Value 4.10 for the intercept and 0.4723 with t Value -0.72 for the PRICE variable.

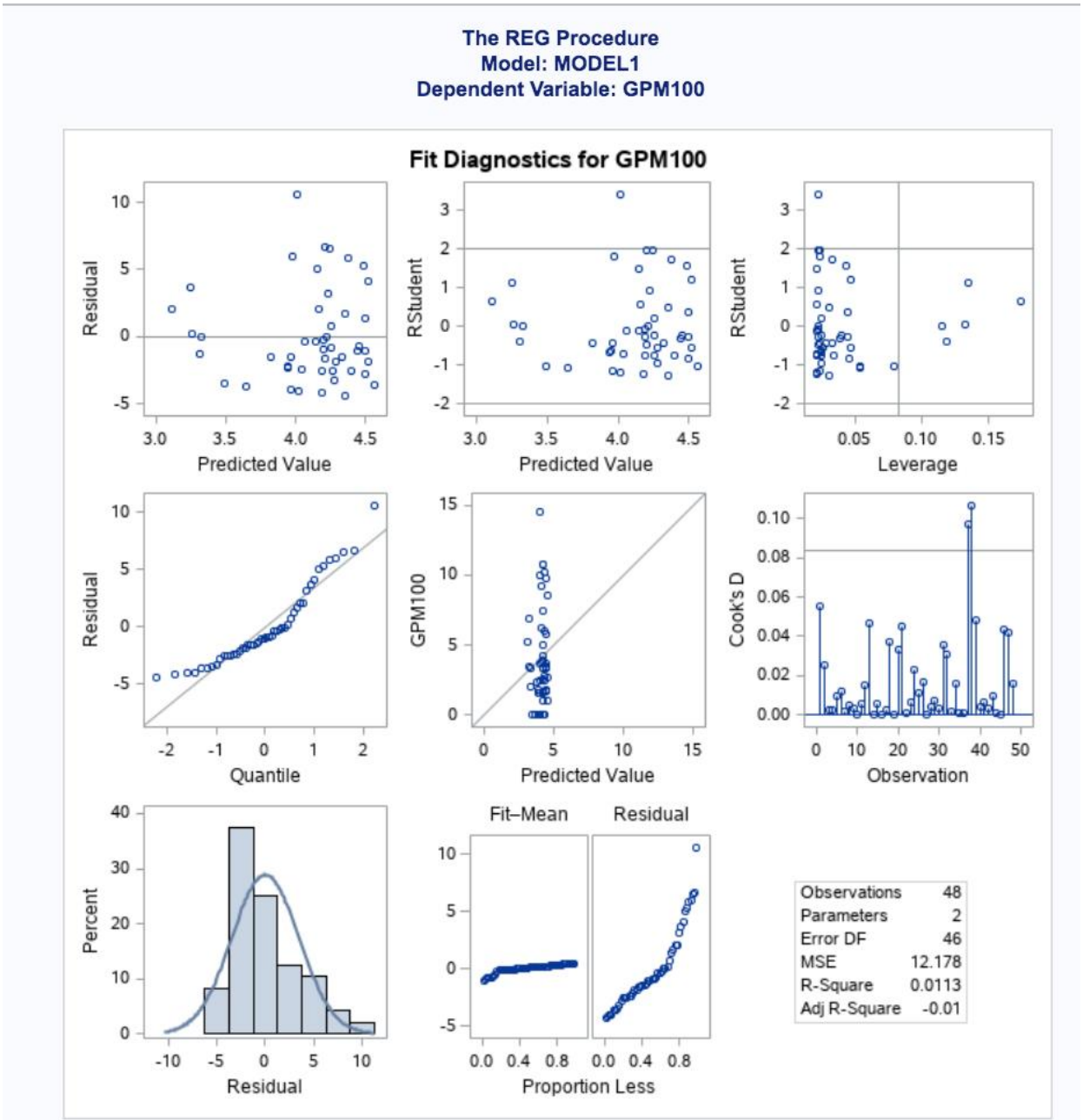

Figure 38 Problem 5 Part B Linear Regression Statistics

From the linear regression statistics of the model gpm100 (Y) = price (X) the associated R-Squared value is 0.0113 which is a very bad fitted model. So we conclude that there is no linear relationship between the variables gpm100 and price.