Program Summary - Homework 7.sas

Execution Environment

Author:	chwang10
File:	/home/chwang10/Homework 7.sas
SAS Platform:	Linux LIN X64 3.10.0-1062.9.1.el7.x86_64
SAS Host:	ODAWS01-USW2.ODA.SAS.COM
SAS Version:	9.04.01M6P11072018
SAS Locale:	en_US
Submission Time:	11/18/2020, 9:51:52 PM
Browser Host:	ASTOUND-66-234-210-119.CA.ASTOUND.NET
User Agent:	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/86.0.4240.198 Safari/537.36
Application Server:	ODAMID01-USW2.ODA.SAS.COM

Code: Homework 7.sas

```
* Programmed by Charles Hwang  *
* Coded in SAS OnDemand        *
* Wednesday, November 18, 2020 *
* Course: STAT 403             *
* Title: Homework 7            *;

/* 1a */ Data CarBrands;
Length Subject$ 2 Car1 - Car4$ 9; * "Chevrolet" appears to be the longest car brand name ;
Infile "/home/chwang10/CarBrands.txt";
Input Subject$ Car1 - Car4$;
Run;

/* 1b */ Data restructure;
Set CarBrands;
Array y[*] Car1 - Car4;
Do i = 1 to dim(y);
Car=y[i];
If Car="XX" then delete; * Deleting 42 placeholder observations ;
Output;
End;
Drop i Car1 - Car4;
Run;

/* 1c(i) */ Proc Freq data=restructure order=freq;
Title "1c(i). List of Car Brands in Descending Order";
Table Car /nocum; * Excluding cumulative statistics ;
Run;
/* 1c(ii)a */ * I will use a Chi-Squared Goodness-of-Fit test for this analysis. Because of the small
stated population proportions for GMC (106 * 4% = 4.24), Hyundai (106 * 4% = 4.24), and
Nissan (106 * 2% = 2.12), the sample size assumption (n > 5) is slightly violated in one-third of
cells (3 brands out of 9). ;
* H0: The sample proportions of car brands are reflective of the stated population proportions.
HA: The sample proportions of car brands are not reflective of the stated population proportions. ;
Proc Freq data=restructure; * Default sort for correct population proportions ;
Title "1c(ii). Chi-Squared Goodness-of-Fit Test on Car Brand Proportions";
Table Car /chisq testp=(12 8 16 4 22 4 9 2 23) nocum; * Because there is only one variable in the dataset,
additional weight counts are not needed. ;
Run; * We fail to reject H0 at the α = .05 level. There is insufficient evidence (χ = 6.4856, p = 0.5930)
that the sample proportions of car brands are not reflective of the stated population proportions. ;
/* 1c(ii)b */ * GMC and Nissan are tied as the most overrepresented brands in the sample compared to
their stated population proportions with approximately |8/(106*4%) - 1| = |47/53| = 88.679245 percent
more cars than expected. Dodge is the most underrepresented brand in the sample compared to its stated
population proportion with |6/(106*8%) - 1| = |-31/106| = 29.245283 percent less cars than expected.
Even though GMC and Nissan are both equally overrepresented in the sample, GMC contributes twice as much
value to the chi-squared test statistic as Nissan, and the most value of any brand in the dataset,
because the numerator is squared when calculating the value for the test
statistic ((8-106*4%)^2/(106*4%) = 3.33433962264 vs. (4-106*2%)^2/(106*2%) = 1.66716981132). ;
```

```
/* 2a */ Data TestScores;
Infile "/home/chwang10/Testscores.txt";
Input Year$;
Input Grade3 - Grade8; * We will consider "Grade" as numeric rather than character for now. ;
Input Score3 - Score8;
Run;
```

```
/* 2b */ Data TestScoreXY;
Set TestScores;
Array a[*] Grade3 - Grade8; * Array placeholder variables must be different from output variables ;
Array b[*] Score3 - Score8;
Do i = 1 to dim(a); * Arrays need to be done simultaneously in order to work ;
Do i = 1 to dim(b);
X=a[i];
Y=b[i];
Output; * Only one "Output" command needed ;
End; * Ending both Do loops ;
End;
Drop Year i Grade3 - Grade8 Score3 - Score8;
Run;
```

```
/* 2c */ Proc Reg data=TestScoreXY;
Title "2c. Linear Regression of Test Score Data";
Model Y=X; * Linear model: Y = 91.23810*X + 103.74603 ;
Run; * Both parameters are significant at the α = .01 level. There is sufficient evidence that both the
intercept (p < 0.0001) and slope (p < 0.0001) are significant to the linear model. ;
* There is no clear nonlinear pattern, but the residuals and studentized residuals appear to be slightly
heteroscedastic. However, according to Cook's D, there is only one slightly high-leverage point, and the
histogram of the data is approximately normal. ;
```

Log: Homework 7.sas

Notes (18)

```
 1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
 70
 71         * Programmed by Charles Hwang  *
 72         * Coded in SAS OnDemand        *
 73         * Wednesday, November 18, 2020 *
 74         * Course: STAT 403             *
 75         * Title: Homework 7            *;
 76
 77         /* 1a */
 77       !          Data CarBrands;
 78         Length Subject$ 2 Car1 - Car4$ 9; * "Chevrolet" appears to be the longest car brand name ;
 79         Infile "/home/chwang10/CarBrands.txt";
 80         Input Subject$ Car1 - Car4$;
 81         Run;

 NOTE: The infile "/home/chwang10/CarBrands.txt" is:
       Filename=/home/chwang10/CarBrands.txt,
       Owner Name=chwang10,Group Name=oda,
       Access Permission=-rw-r--r--,
       Last Modified=12Nov2020:01:42:55,
       File Size (bytes)=944

 NOTE: 37 records were read from the infile "/home/chwang10/CarBrands.txt".
       The minimum record length was 18.
       The maximum record length was 33.
 NOTE: The data set WORK.CARBRANDS has 37 observations and 5 variables.
 NOTE: DATA statement used (Total process time):
       real time           0.00 seconds
       user cpu time       0.00 seconds
       system cpu time     0.00 seconds
       memory              876.03k
       OS Memory           38312.00k
       Timestamp           11/19/2020 05:51:51 AM
       Step Count                        715  Switch Count  2
       Page Faults                       0
       Page Reclaims                     92
       Page Swaps                        0
       Voluntary Context Switches        14
       Involuntary Context Switches      0
```

```
      Block Input Operations            0
      Block Output Operations           264


82
83         /* 1b */
83       !          Data restructure;
84         Set CarBrands;
85         Array y[*] Car1 - Car4;
86         Do i = 1 to dim(y);
87         Car=y[i];
88         If Car="XX" then delete; * Deleting 42 placeholder observations ;
89         Output;
90         End;
91         Drop i Car1 - Car4;
92         Run;

NOTE: There were 37 observations read from the data set WORK.CARBRANDS.
NOTE: The data set WORK.RESTRUCTURE has 106 observations and 2 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1170.75k
      OS Memory           38572.00k
      Timestamp           11/19/2020 05:51:51 AM
      Step Count                        716  Switch Count  2
      Page Faults                       0
      Page Reclaims                     126
      Page Swaps                        0
      Voluntary Context Switches        11
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


93
94         /* 1c(i) */
94       !             Proc Freq data=restructure order=freq;
95         Title "1c(i). List of Car Brands in Descending Order";
96         Table Car /nocum; * Excluding cumulative statistics ;
97         Run;

NOTE: There were 106 observations read from the data set WORK.RESTRUCTURE.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.02 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              2570.00k
      OS Memory           38572.00k
      Timestamp           11/19/2020 05:51:51 AM
      Step Count                        717  Switch Count  3
      Page Faults                       0
      Page Reclaims                     129
      Page Swaps                        0
      Voluntary Context Switches        19
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           280


98         /* 1c(ii)a */ * I will use a Chi-Squared Goodness-of-Fit test for this analysis. Because of the small
99         stated population proportions for GMC (106 * 4% = 4.24), Hyundai (106 * 4% = 4.24), and
100        Nissan (106 * 2% = 2.12), the sample size assumption (n > 5) is slightly violated in one-third of
101        cells (3 brands out of 9). ;
102        * H0: The sample proportions of car brands are reflective of the stated population proportions.
103        HA: The sample proportions of car brands are not reflective of the stated population proportions. ;
104        Proc Freq data=restructure; * Default sort for correct population proportions ;
105        Title "1c(ii). Chi-Squared Goodness-of-Fit Test on Car Brand Proportions";
106        Table Car /chisq testp=(12 8 16 4 22 4 9 2 23) nocum; * Because there is only one variable in the dataset,
107        additional weight counts are not needed. ;
108        Run;

NOTE: There were 106 observations read from the data set WORK.RESTRUCTURE.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.18 seconds
      user cpu time       0.10 seconds
      system cpu time     0.00 seconds
      memory              14144.75k
      OS Memory           46252.00k
      Timestamp           11/19/2020 05:51:51 AM
      Step Count                        718  Switch Count  3
      Page Faults                       0
      Page Reclaims                     2488
      Page Swaps                        0
      Voluntary Context Switches        249
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           1080


 108      !      * We fail to reject H0 at the α = .05 level. There is insufficient evidence (χ = 6.4856, p = 0.5930)
 109        that the sample proportions of car brands are not reflective of the stated population proportions. ;
 110        /* 1c(ii)b */ * GMC and Nissan are tied as the most overrepresented brands in the sample compared to
 111        their stated population proportions with approximately |8/(106*4%) – 1| = |47/53| = 88.679245 percent
 112        more cars than expected. Dodge is the most underrepresented brand in the sample compared to its stated
 113        population proportion with |6/(106*8%) – 1| = |-31/106| = 29.245283 percent less cars than expected.
 114        Even though GMC and Nissan are both equally overrepresented in the sample, GMC contributes twice as much
 115        value to the chi-squared test statistic as Nissan, and the most value of any brand in the dataset,
 116        because the numerator is squared when calculating the value for the test
 117        statistic ((8-106*4%)^2/(106*4%) = 3.33433962264 vs. (4-106*2%)^2/(106*2%) = 1.66716981132). ;
 118
 119        /* 2a */
 119      !          Data TestScores;
 120        Infile "/home/chwang10/Testscores.txt";
 121        Input Year$;
 122        Input Grade3 – Grade8; * We will consider "Grade" as numeric rather than character for now. ;
 123        Input Score3 – Score8;
 124        Run;

NOTE: The infile "/home/chwang10/Testscores.txt" is:
      Filename=/home/chwang10/Testscores.txt,
      Owner Name=chwang10,Group Name=oda,
      Access Permission=-rw-r--r--,
      Last Modified=12Nov2020:01:42:55,
      File Size (bytes)=263

NOTE: 18 records were read from the infile "/home/chwang10/Testscores.txt".
      The minimum record length was 4.
      The maximum record length was 23.
NOTE: The data set WORK.TESTSCORES has 6 observations and 13 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              779.84k
      OS Memory           45992.00k
      Timestamp           11/19/2020 05:51:51 AM
      Step Count                        719  Switch Count  2
      Page Faults                       0
      Page Reclaims                     105
      Page Swaps                        0
      Voluntary Context Switches        16
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272


 125
 126        /* 2b */
 126      !          Data TestScoreXY;
 127        Set TestScores;
 128        Array a[*] Grade3 – Grade8; * Array placeholder variables must be different from output variables ;
 129        Array b[*] Score3 – Score8;
 130        Do i = 1 to dim(a); * Arrays need to be done simultaneously in order to work ;
 131        Do i = 1 to dim(b);
 132        X=a[i];
 133        Y=b[i];
 134        Output; * Only one "Output" command needed ;
 135        End; * Ending both Do loops ;
 136        End;
 137        Drop Year i Grade3 – Grade8 Score3 – Score8;
 138        Run;

NOTE: There were 6 observations read from the data set WORK.TESTSCORES.
NOTE: The data set WORK.TESTSCOREXY has 36 observations and 2 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.01 seconds
      memory              959.50k
      OS Memory           46252.00k
      Timestamp           11/19/2020 05:51:51 AM
      Step Count                        720  Switch Count  2
      Page Faults                         0
      Page Reclaims                       128
      Page Swaps                          0
      Voluntary Context Switches          17
      Involuntary Context Switches        0
      Block Input Operations              0
      Block Output Operations             264


139
140          /* 2c */
140        !          Proc Reg data=TestScoreXY;
141          Title "2c. Linear Regression of Test Score Data";
142          Model Y=X; * Linear model: Y = 91.23810*X + 103.74603 ;
143          Run;

143        !      * Both parameters are significant at the α = .01 level. There is sufficient evidence that both the
144          intercept (p < 0.0001) and slope (p < 0.0001) are significant to the linear model. ;
145          * There is no clear nonlinear pattern, but the residuals and studentized residuals appear to be slightly
146          heteroscedastic. However, according to Cook's D, there is only one slightly high-leverage point, and the
147          histogram of the data is approximately normal. ;
148
149          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
160
```

Results: Homework 7.sas

1c(i). List of Car Brands in Descending Order

The FREQ Procedure

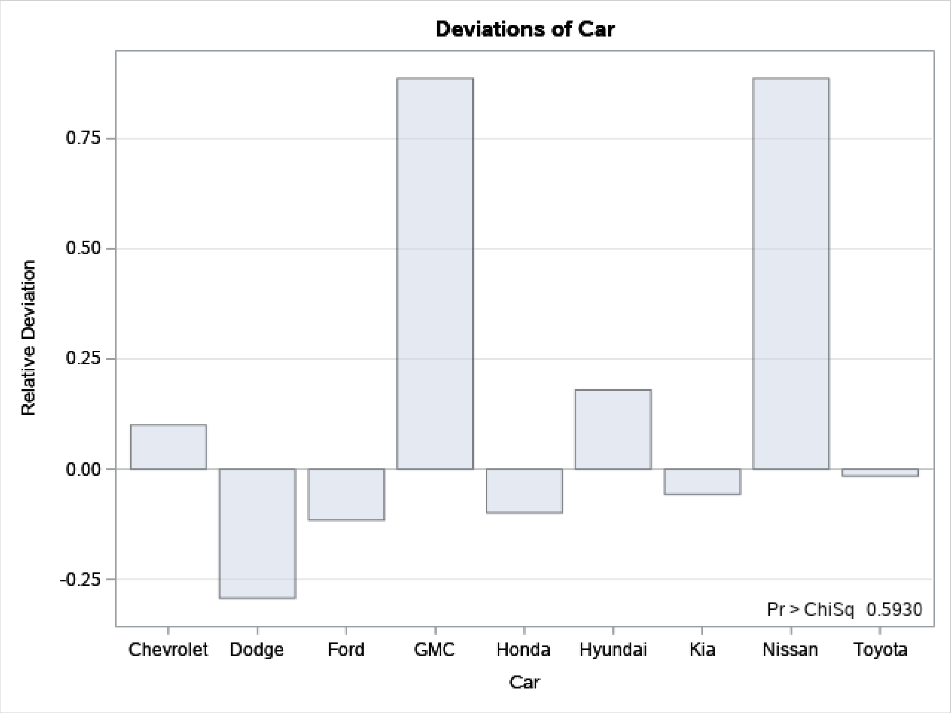
Car	Frequency	Percent
Toyota	24	22.64
Honda	21	19.81
Ford	15	14.15
Chevrolet	14	13.21
Kia	9	8.49
GMC	8	7.55
Dodge	6	5.66
Hyundai	5	4.72
Nissan	4	3.77

1c(ii). Chi-Squared Goodness-of-Fit Test on Car Brand Proportions

The FREQ Procedure

Car	Frequency	Percent	Test Percent
Chevrolet	14	13.21	12.00
Dodge	6	5.66	8.00
Ford	15	14.15	16.00
GMC	8	7.55	4.00
Honda	21	19.81	22.00
Hyundai	5	4.72	4.00
Kia	9	8.49	9.00
Nissan	4	3.77	2.00
Toyota	24	22.64	23.00

Chi-Square Test for Specified Proportions	
Chi-Square	6.4856
DF	8
Pr > ChiSq	0.5930
WARNING: 33% of the cells have expected counts less than 5. Chi-Square may not be a valid test.	



Sample Size = 106

2c. Linear Regression of Test Score Data

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Number of Observations Read	36
Number of Observations Used	36

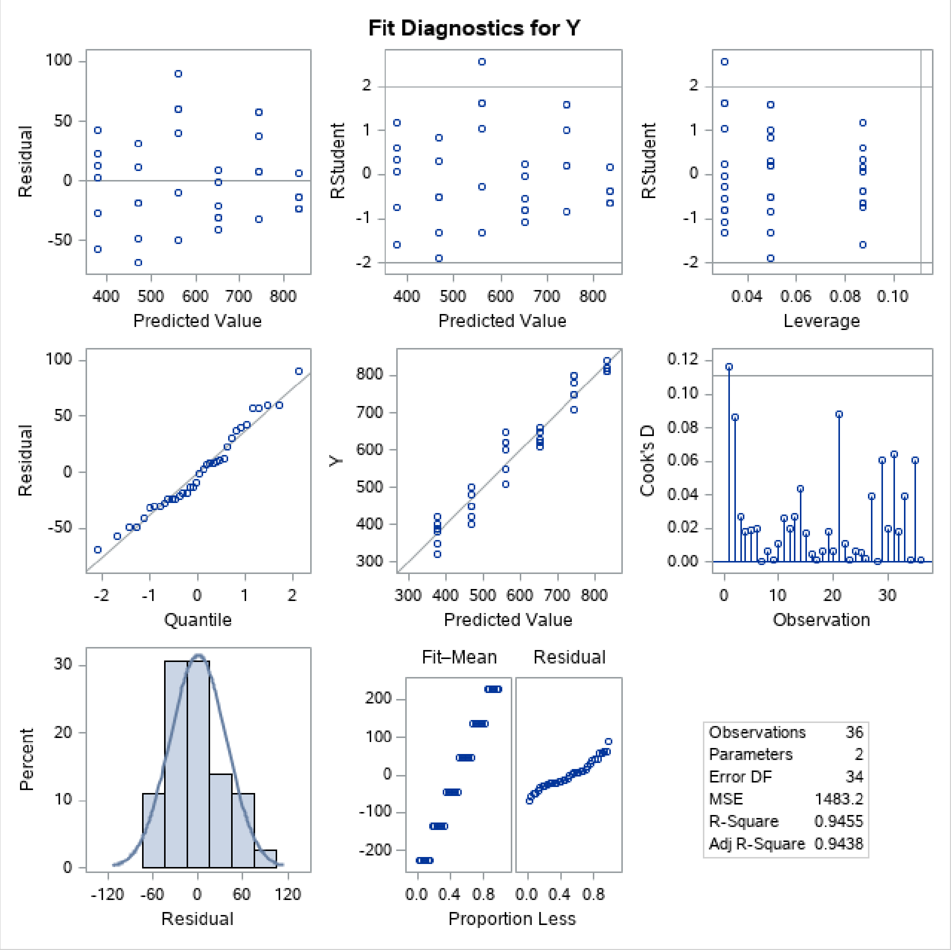
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	874061	874061	589.32	<.0001
Error	34	50428	1483.17460		
Corrected Total	35	924489			

Root MSE	38.51201	**R-Square**	0.9455
Dependent Mean	605.55556	**Adj R-Sq**	0.9438
Coeff Var	6.35978		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	103.74603	21.64474	4.79	<.0001
X	1	91.23810	3.75839	24.28	<.0001

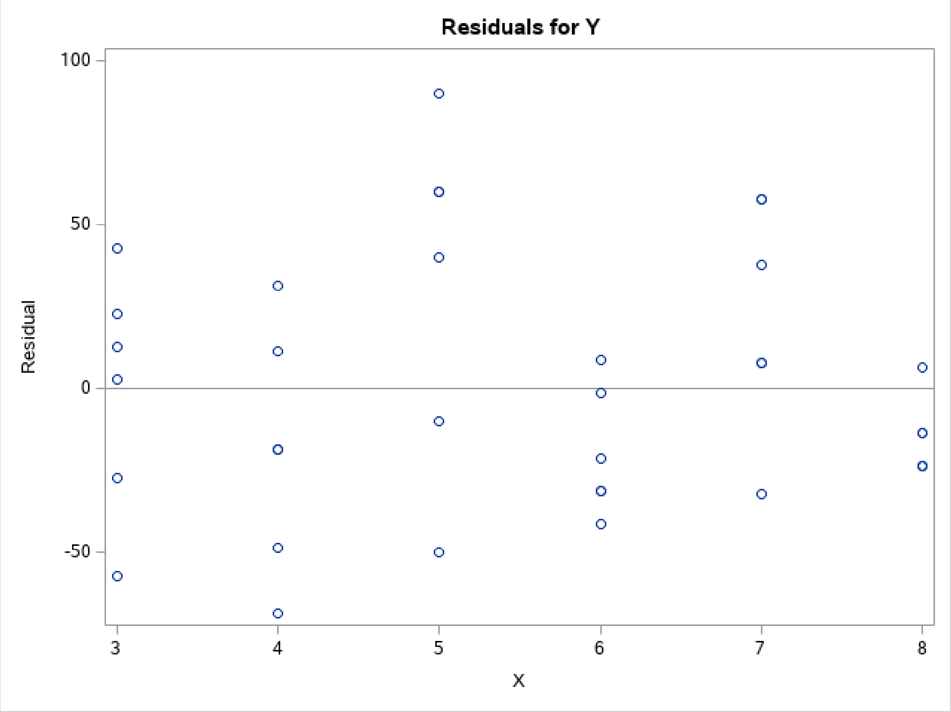
2c. Linear Regression of Test Score Data

The REG Procedure
Model: MODEL1
Dependent Variable: Y

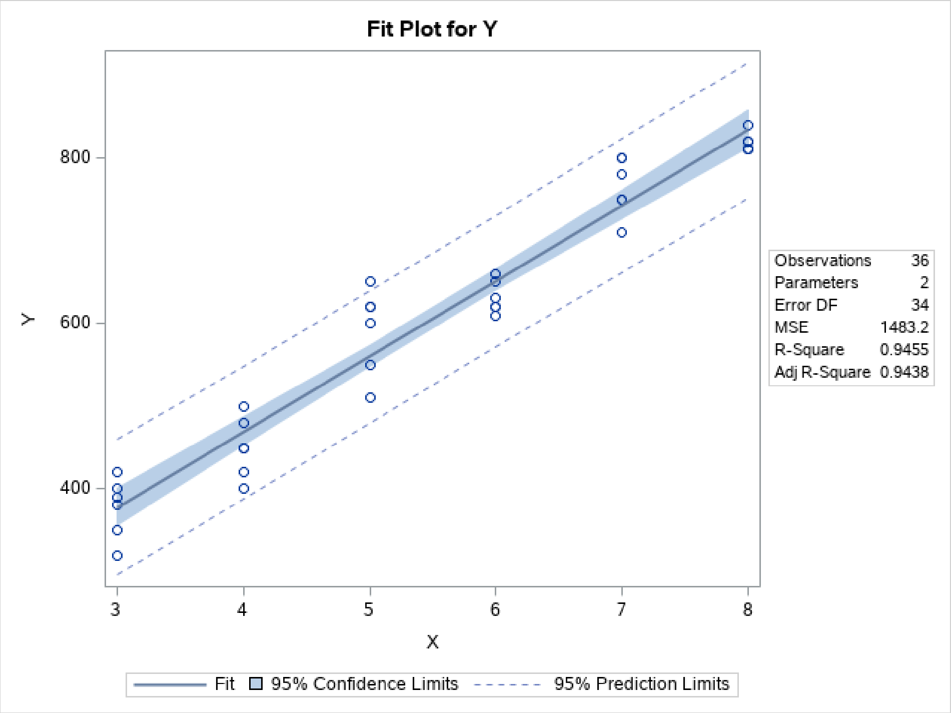


Fit Diagnostics for Y

Observations	36
Parameters	2
Error DF	34
MSE	1483.2
R-Square	0.9455
Adj R-Square	0.9438



Residuals for Y



Fit Plot for Y

Observations	36
Parameters	2
Error DF	34
MSE	1483.2
R-Square	0.9455
Adj R-Square	0.9438