STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 1

The UNIVARIATE Procedure
Variable: VALUE

Moments			
N	880	**Sum Weights**	880
Mean	127.833523	**Sum Observations**	112493.5
Std Deviation	12.3263345	**Variance**	151.938523
Skewness	-0.0503064	**Kurtosis**	-0.056344
Uncorrected SS	14513994.4	**Corrected SS**	133553.962
Coeff Variation	9.64248992	**Std Error Mean**	0.41552065

Basic Statistical Measures			
Location		**Variability**	
Mean	127.8335	**Std Deviation**	12.32633
Median	128.0300	**Variance**	151.93852
Mode	121.0400	**Range**	79.23000
		Interquartile Range	17.03500

Note: The mode displayed is the smallest of 3 modes with a count of 3.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	307.6466	**Pr > |t|**	<.0001
Sign	**M**	440	**Pr >= |M|**	<.0001
Signed Rank	**S**	193820	**Pr >= |S|**	<.0001

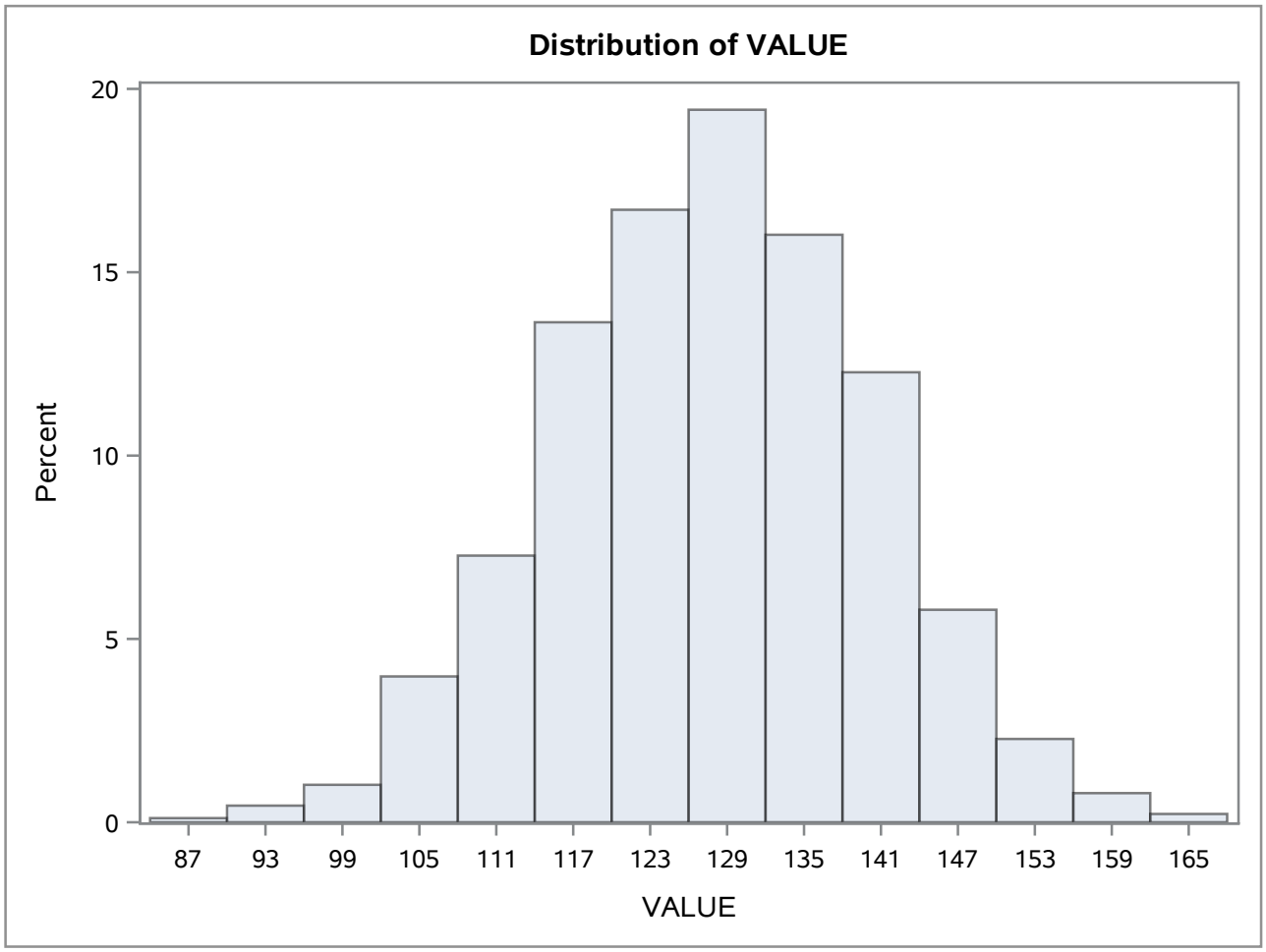
Quantiles (Definition 5)	
Level	**Quantile**
100% Max	165.700
99%	156.200
95%	148.055
90%	143.470
75% Q3	136.390
50% Median	128.030
25% Q1	119.355
10%	112.340
5%	107.345
1%	98.400
0% Min	86.470

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 1

The UNIVARIATE Procedure
Variable: VALUE

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
86.47	558	157.15	643
90.32	588	159.28	146
92.63	563	159.70	583
93.96	515	163.71	793
94.86	735	165.70	235

The UNIVARIATE Procedure

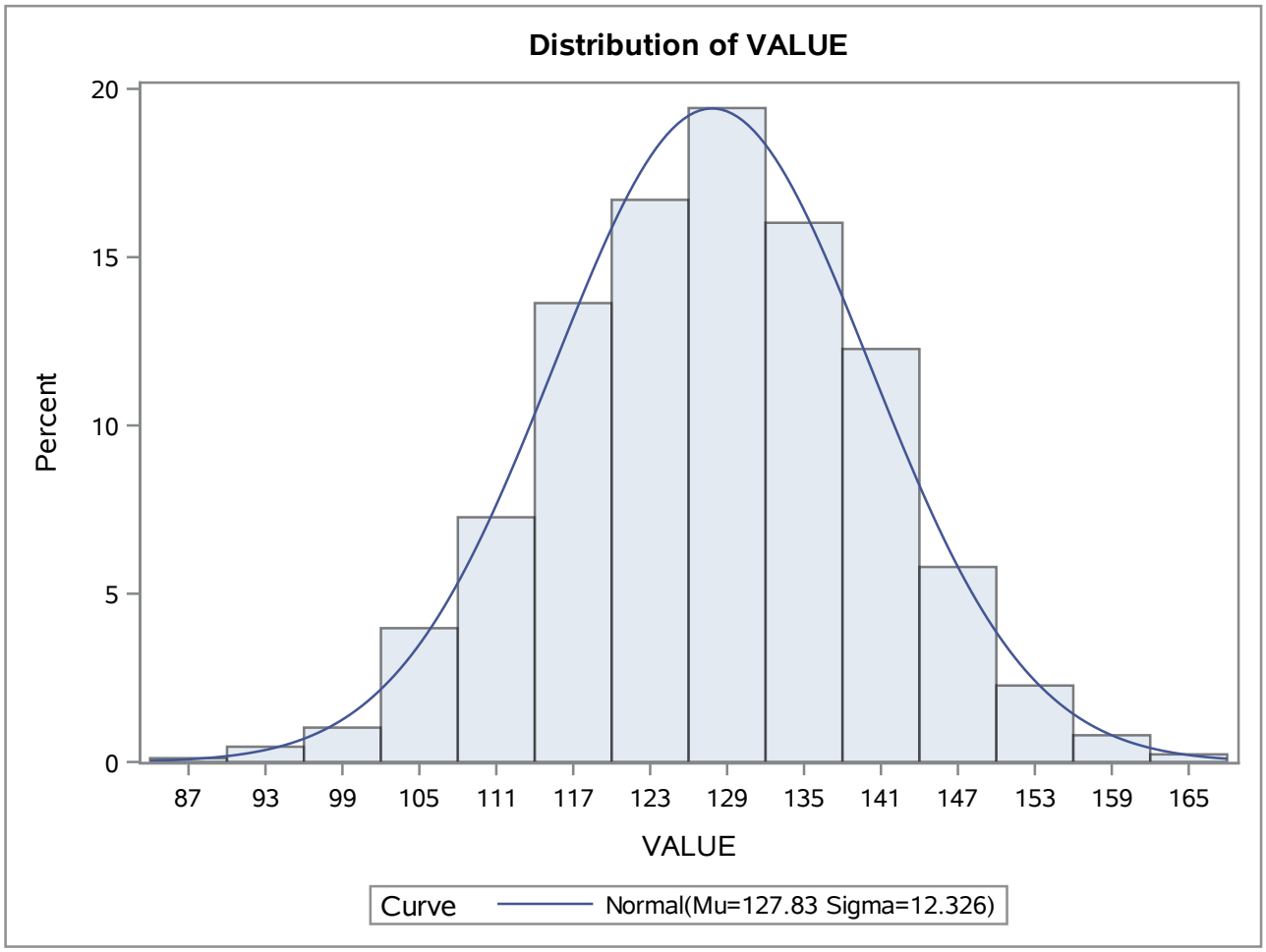


STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 1

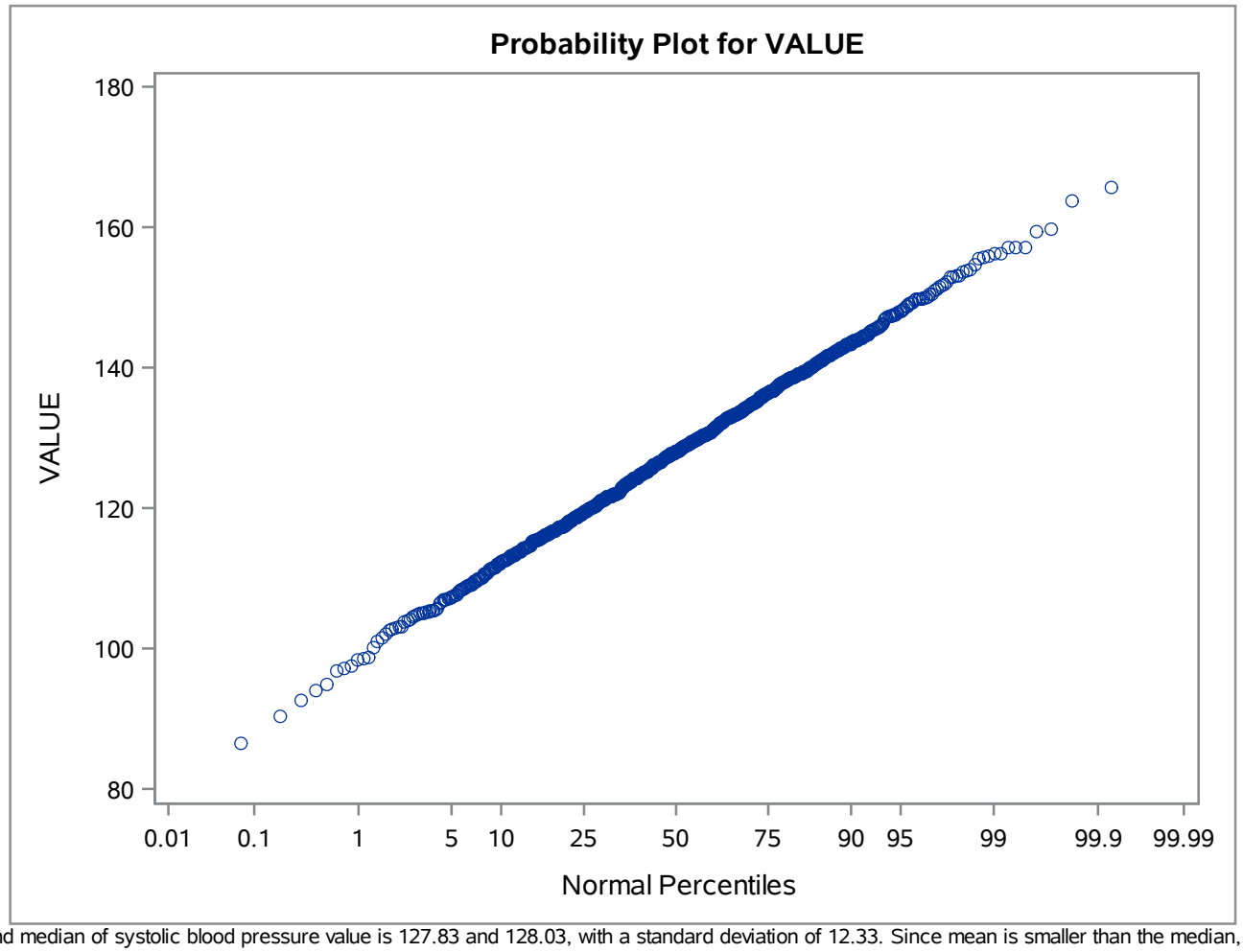
The UNIVARIATE Procedure
Variable: VALUE

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.99942	**Pr < W**	0.9971
Kolmogorov-Smirnov	**D**	0.015657	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.01714	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.105707	**Pr > A-Sq**	>0.2500

The UNIVARIATE Procedure



The UNIVARIATE Procedure



The mean and median of systolic blood pressure value is 127.83 and 128.03, with a standard deviation of 12.33. Since mean is smaller than the median, the distribution is likely to have a tail to the left. The negative skewness of -0.05 supports it. The range of systolic blood pressure value is 79.23. Systolic blood pressure value seems to follow normal distribution. The quantitative tests agree. All the four tests indicate a p-value of greater than 0.05, showing that we fail to reject H0, which is the data is normally distributed.

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 1

The UNIVARIATE Procedure
Variable: VALUE

RTRTN=1

Moments			
N	220	**Sum Weights**	220
Mean	131.468818	**Sum Observations**	28923.14
Std Deviation	12.2674483	**Variance**	150.490288
Skewness	-0.0010869	**Kurtosis**	-0.2313581
Uncorrected SS	3835448.41	**Corrected SS**	32957.3731
Coeff Variation	9.33107065	**Std Error Mean**	0.8270712

Basic Statistical Measures			
Location		**Variability**	
Mean	131.4688	**Std Deviation**	12.26745
Median	131.2050	**Variance**	150.49029
Mode	121.0400	**Range**	66.95000
		Interquartile Range	16.53500

Note: The mode displayed is the smallest of 2 modes with a count of 2.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	158.9571	**Pr > |t|**	<.0001
Sign	**M**	110	**Pr >= |M|**	<.0001
Signed Rank	**S**	12155	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	163.710
99%	159.280
95%	151.930
90%	147.560
75% Q3	139.925
50% Median	131.205
25% Q1	123.390
10%	114.635
5%	111.375

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 1

The UNIVARIATE Procedure
Variable: VALUE

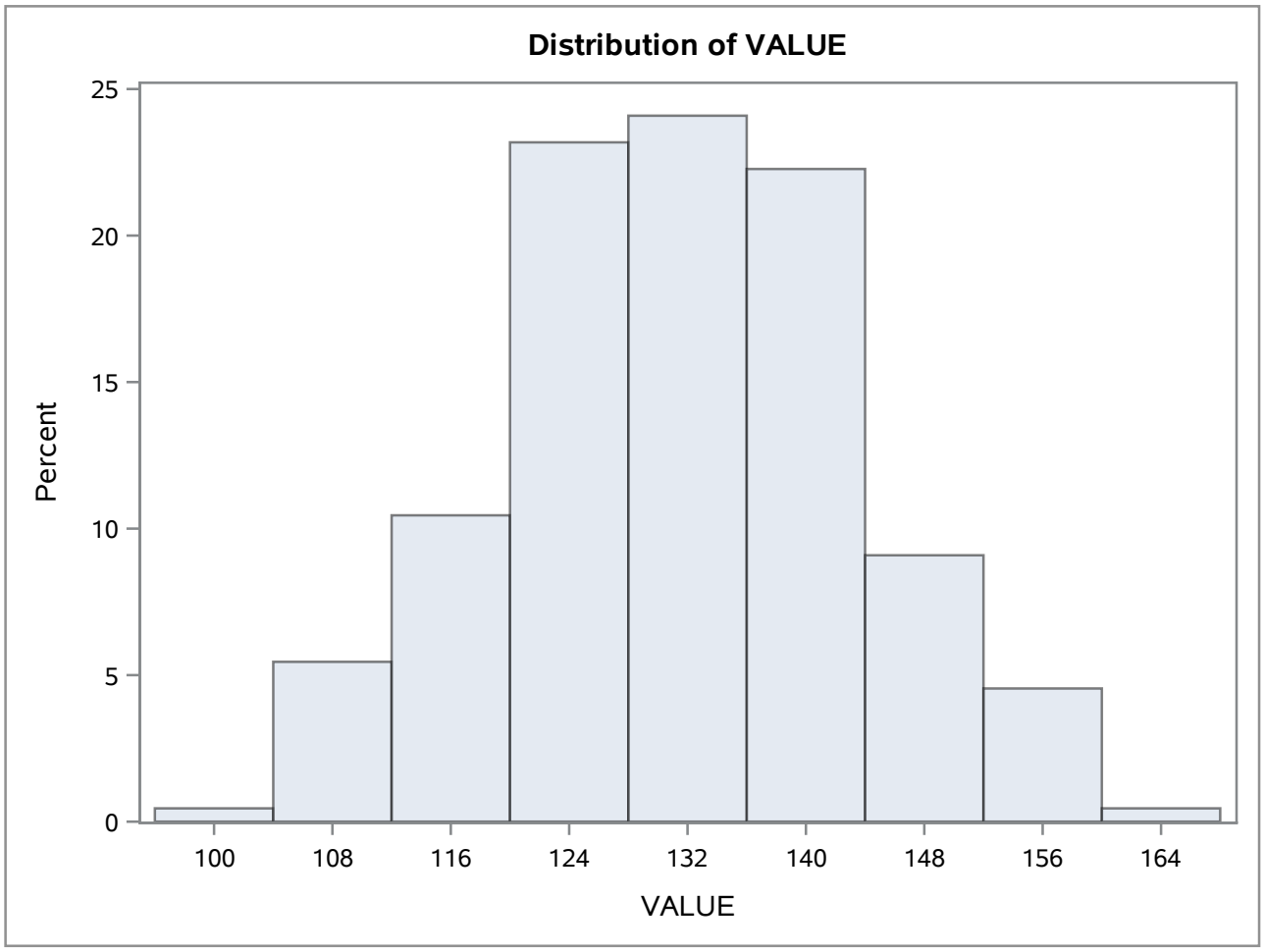
RTRTN=1

Quantiles (Definition 5)	
Level	**Quantile**
1%	105.490
0% Min	96.760

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
96.76	202	156.22	8
104.64	126	157.04	76
105.49	66	159.28	37
106.88	167	159.70	146
107.19	176	163.71	199

The UNIVARIATE Procedure

RTRTN=1



STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 1

The UNIVARIATE Procedure
Variable: VALUE

RTRTN=2

Moments			
N	220	**Sum Weights**	220
Mean	127.699136	**Sum Observations**	28093.81
Std Deviation	12.6911749	**Variance**	161.065919
Skewness	0.07224663	**Kurtosis**	-0.2093769
Uncorrected SS	3622828.71	**Corrected SS**	35273.4363
Coeff Variation	9.9383404	**Std Error Mean**	0.85563883

Basic Statistical Measures			
Location		**Variability**	
Mean	127.6991	**Std Deviation**	12.69117
Median	127.4250	**Variance**	161.06592
Mode	112.4600	**Range**	71.74000
		Interquartile Range	17.60000

Note: The mode displayed is the smallest of 3 modes with a count of 2.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	149.2442	**Pr > |t|**	<.0001
Sign	**M**	110	**Pr >= |M|**	<.0001
Signed Rank	**S**	12155	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	165.700
99%	157.030
95%	148.395
90%	143.970
75% Q3	136.665
50% Median	127.425
25% Q1	119.065
10%	110.635
5%	107.295

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 1

The UNIVARIATE Procedure
Variable: VALUE

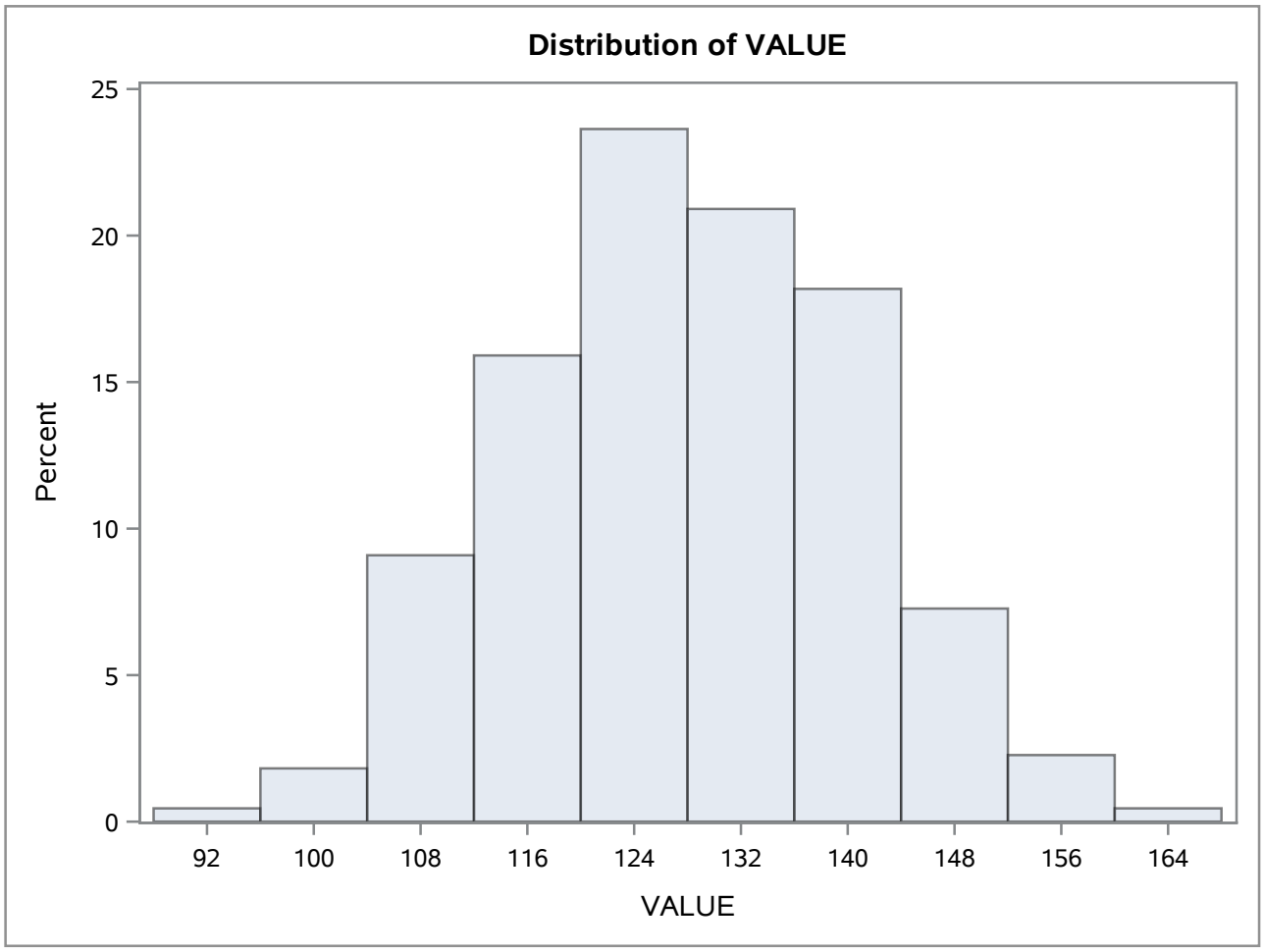
RTRTN=2

Quantiles (Definition 5)	
Level	**Quantile**
1%	101.050
0% Min	93.960

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
93.96	349	153.78	406
98.40	236	155.61	284
101.05	222	157.03	272
102.60	241	157.15	381
103.10	327	165.70	279

The UNIVARIATE Procedure

RTRTN=2



STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 1

The UNIVARIATE Procedure
Variable: VALUE

RTRTN=3

Moments			
N	220	**Sum Weights**	220
Mean	126.213227	**Sum Observations**	27766.91
Std Deviation	11.60215	**Variance**	134.609884
Skewness	-0.2289691	**Kurtosis**	0.15971785
Uncorrected SS	3534030.89	**Corrected SS**	29479.5646
Coeff Variation	9.19249925	**Std Error Mean**	0.78221679

Basic Statistical Measures			
Location		**Variability**	
Mean	126.2132	**Std Deviation**	11.60215
Median	126.5500	**Variance**	134.60988
Mode	116.2300	**Range**	69.32000
		Interquartile Range	16.38500

Note: The mode displayed is the smallest of 3 modes with a count of 2.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	161.3533	**Pr > |t|**	<.0001
Sign	**M**	110	**Pr >= |M|**	<.0001
Signed Rank	**S**	12155	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	155.790
99%	149.820
95%	143.825
90%	141.620
75% Q3	134.520
50% Median	126.550
25% Q1	118.135
10%	112.700
5%	106.070

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 1

The UNIVARIATE Procedure
Variable: VALUE

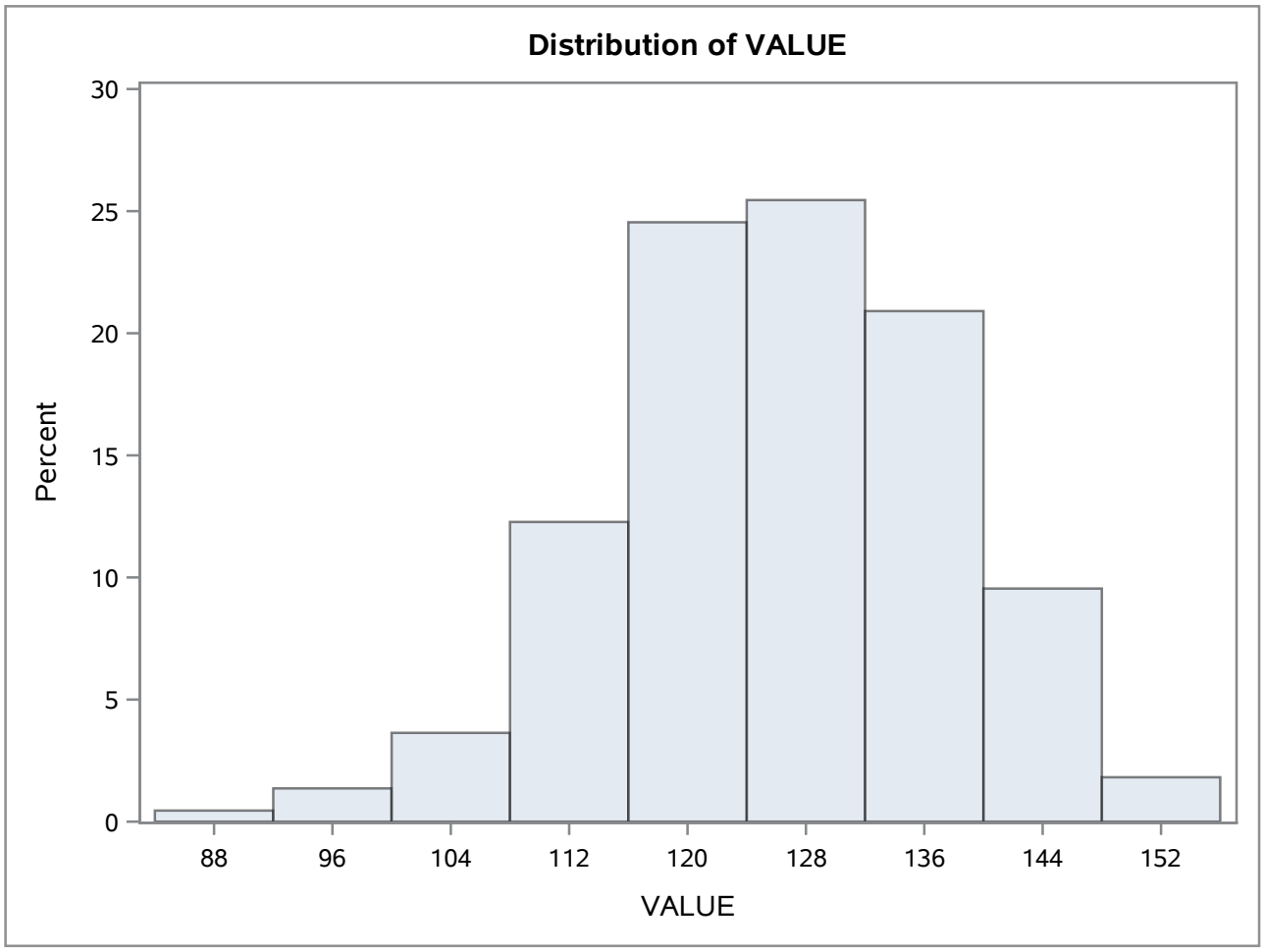
RTRTN=3

Quantiles (Definition 5)	
Level	**Quantile**
1%	98.630
0% Min	86.470

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
86.47	580	147.34	487
92.63	581	149.71	443
98.63	607	149.82	540
98.72	596	153.15	526
102.81	558	155.79	497

The UNIVARIATE Procedure

RTRTN=3



STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 1

The UNIVARIATE Procedure
Variable: VALUE

RTRTN=4

Moments			
N	220	**Sum Weights**	220
Mean	125.952909	**Sum Observations**	27709.64
Std Deviation	12.0077179	**Variance**	144.18529
Skewness	-0.1825049	**Kurtosis**	-0.0691965
Uncorrected SS	3521686.35	**Corrected SS**	31576.5785
Coeff Variation	9.5334979	**Std Error Mean**	0.80956018

Basic Statistical Measures			
Location		**Variability**	
Mean	125.9529	**Std Deviation**	12.00772
Median	126.7850	**Variance**	144.18529
Mode	117.8500	**Range**	62.53000
		Interquartile Range	15.56500

Note: The mode displayed is the smallest of 3 modes with a count of 2.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	155.5819	**Pr > |t|**	<.0001
Sign	**M**	110	**Pr >= |M|**	<.0001
Signed Rank	**S**	12155	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	152.850
99%	151.760
95%	145.660
90%	141.155
75% Q3	133.965
50% Median	126.785
25% Q1	118.400
10%	110.000
5%	105.250

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 1

The UNIVARIATE Procedure
Variable: VALUE

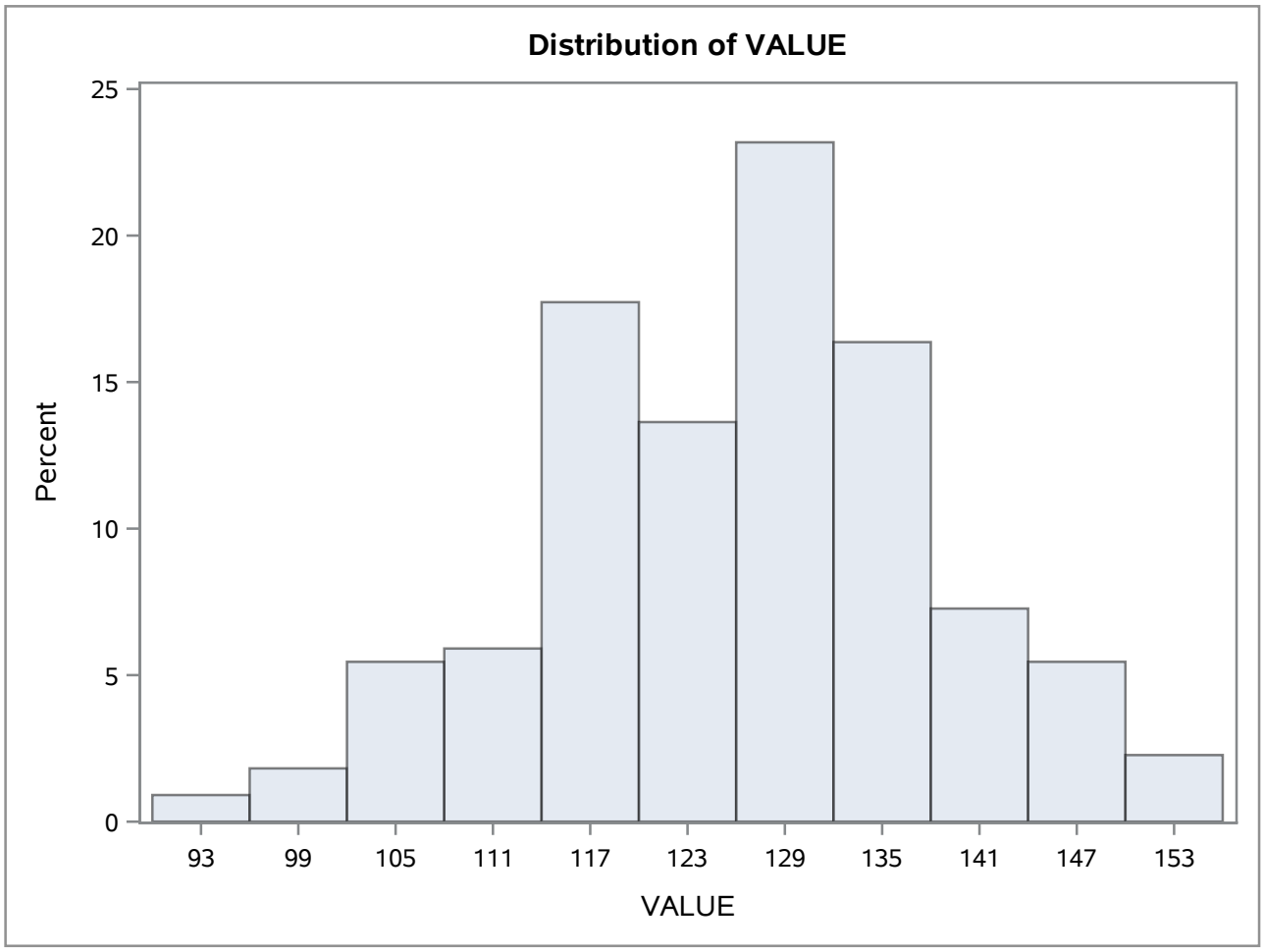
RTRTN=4

Quantiles (Definition 5)	
Level	**Quantile**
1%	97.140
0% Min	90.320

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
90.32	807	151.12	802
94.86	844	151.53	875
97.14	770	151.76	817
97.47	784	152.12	685
100.16	740	152.85	695

The UNIVARIATE Procedure

RTRTN=4



STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 1

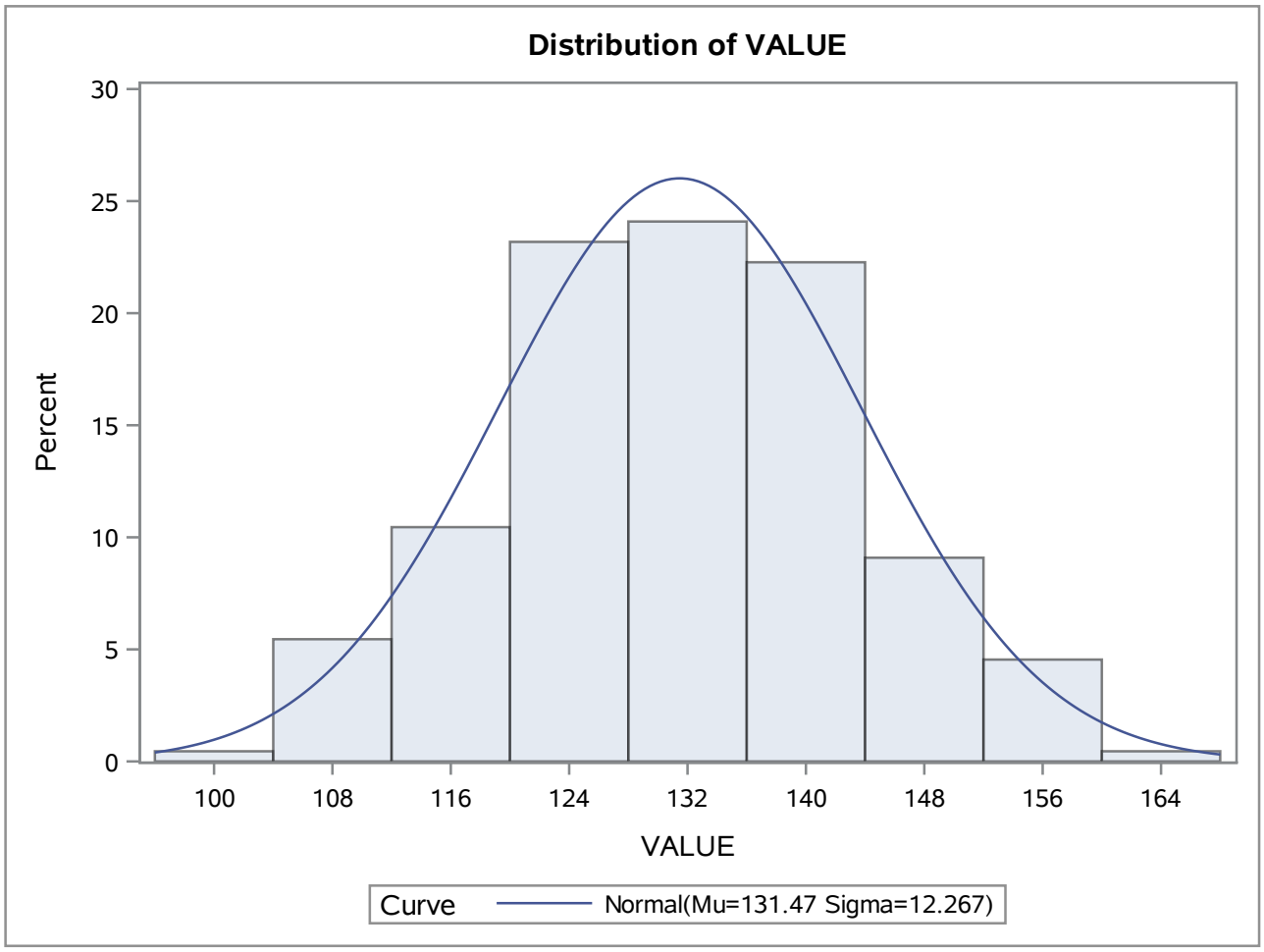
The UNIVARIATE Procedure
Variable: VALUE

RTRTN=1

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.99743	**Pr < W**	0.9785
Kolmogorov-Smirnov	**D**	0.026652	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.014424	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.115705	**Pr > A-Sq**	>0.2500

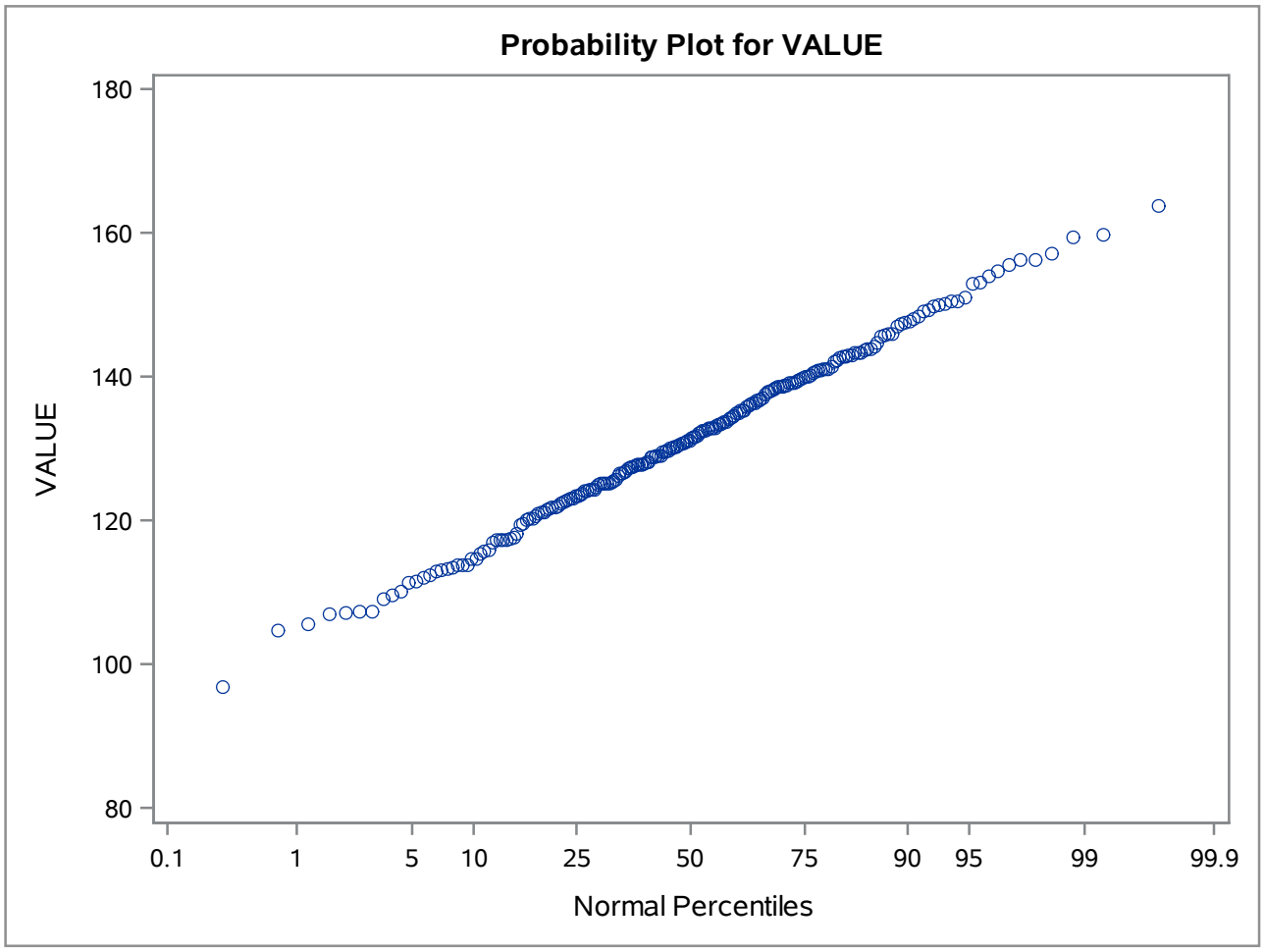
The UNIVARIATE Procedure

RTRTN=1



The UNIVARIATE Procedure

RTRTN=1



STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 1

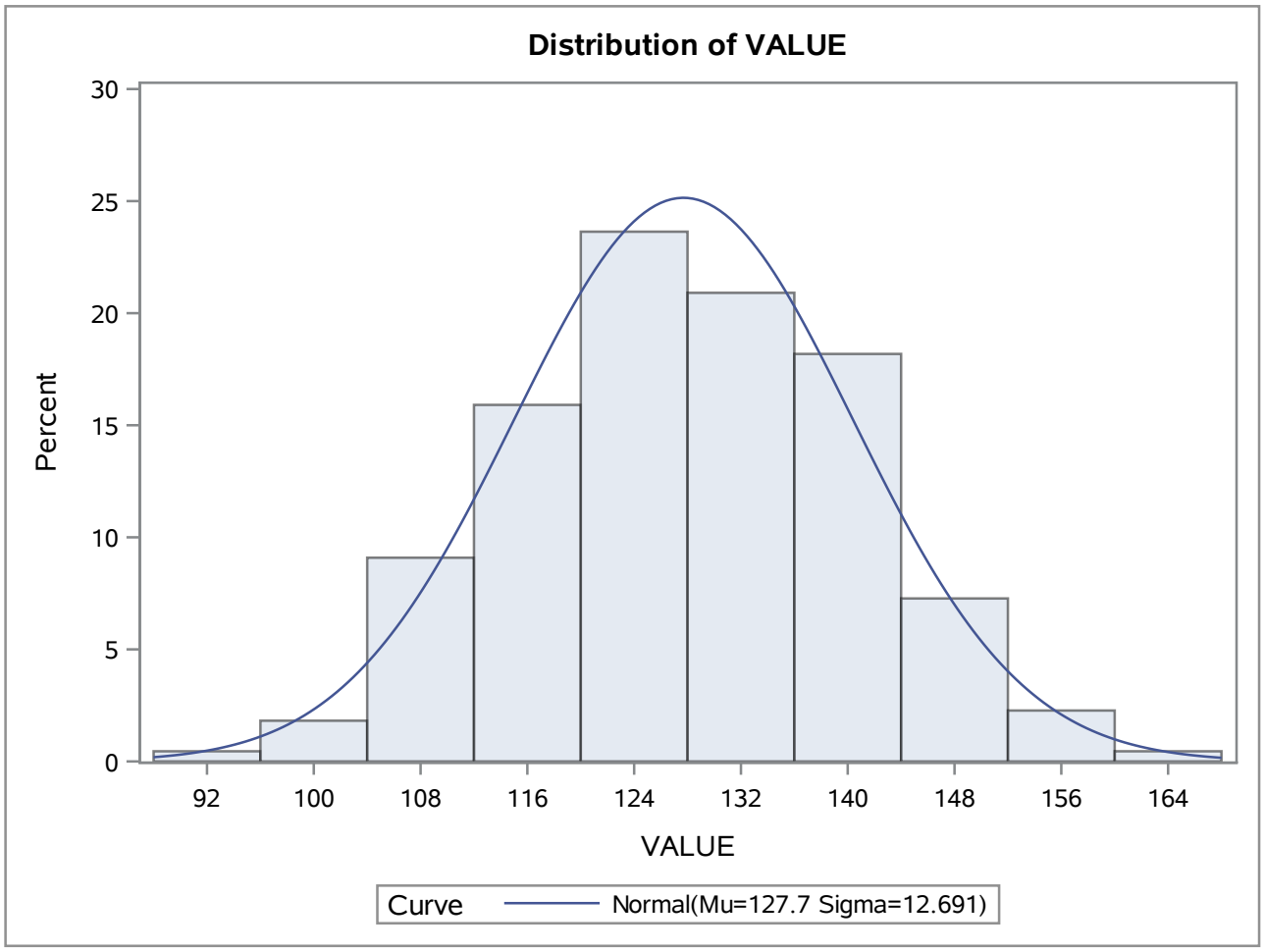
The UNIVARIATE Procedure
Variable: VALUE

RTRTN=2

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.997674	**Pr < W**	0.9879
Kolmogorov-Smirnov	**D**	0.031239	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.022743	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.142688	**Pr > A-Sq**	>0.2500

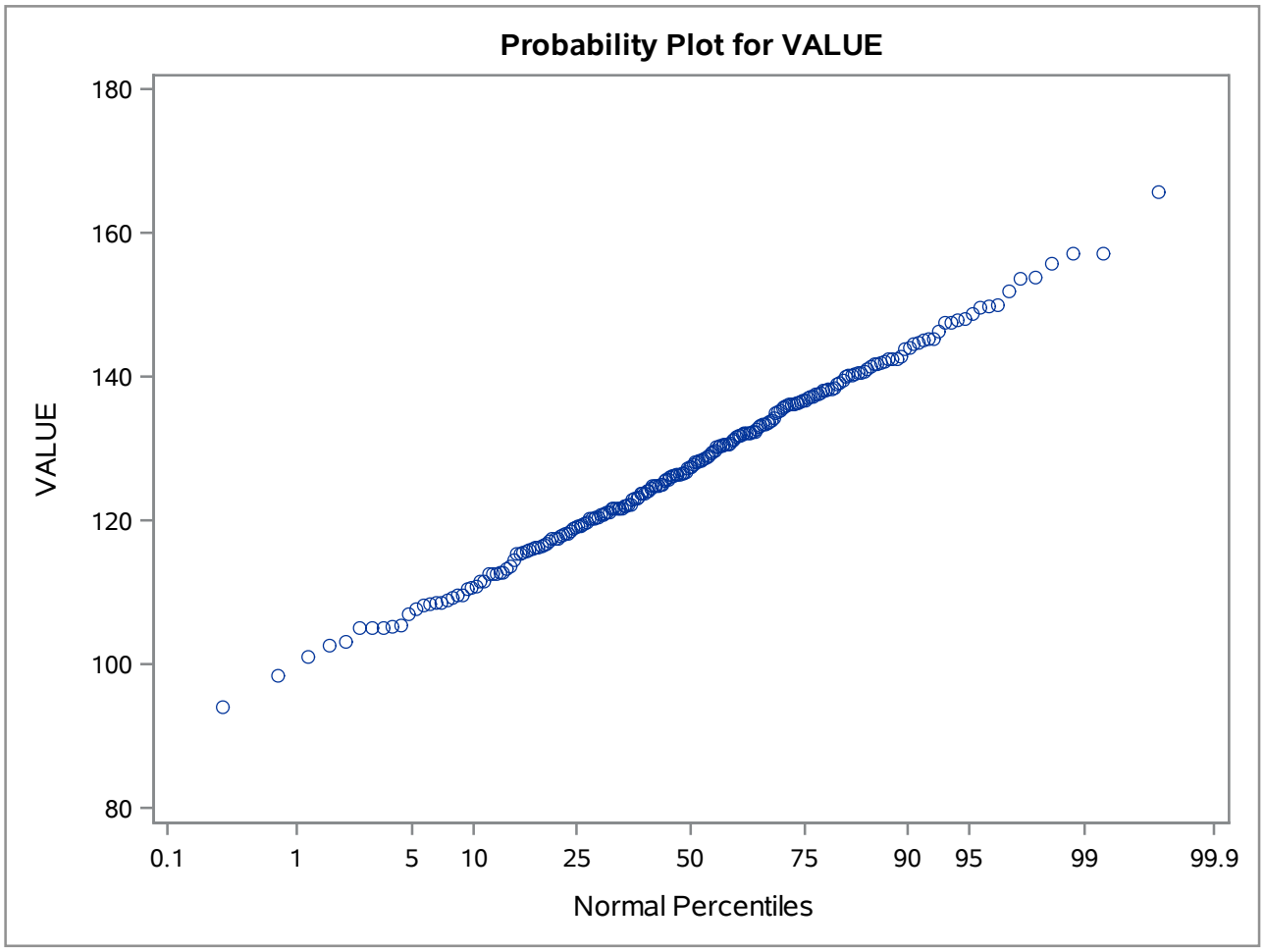
The UNIVARIATE Procedure

RTRTN=2



The UNIVARIATE Procedure

RTRTN=2



STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 1

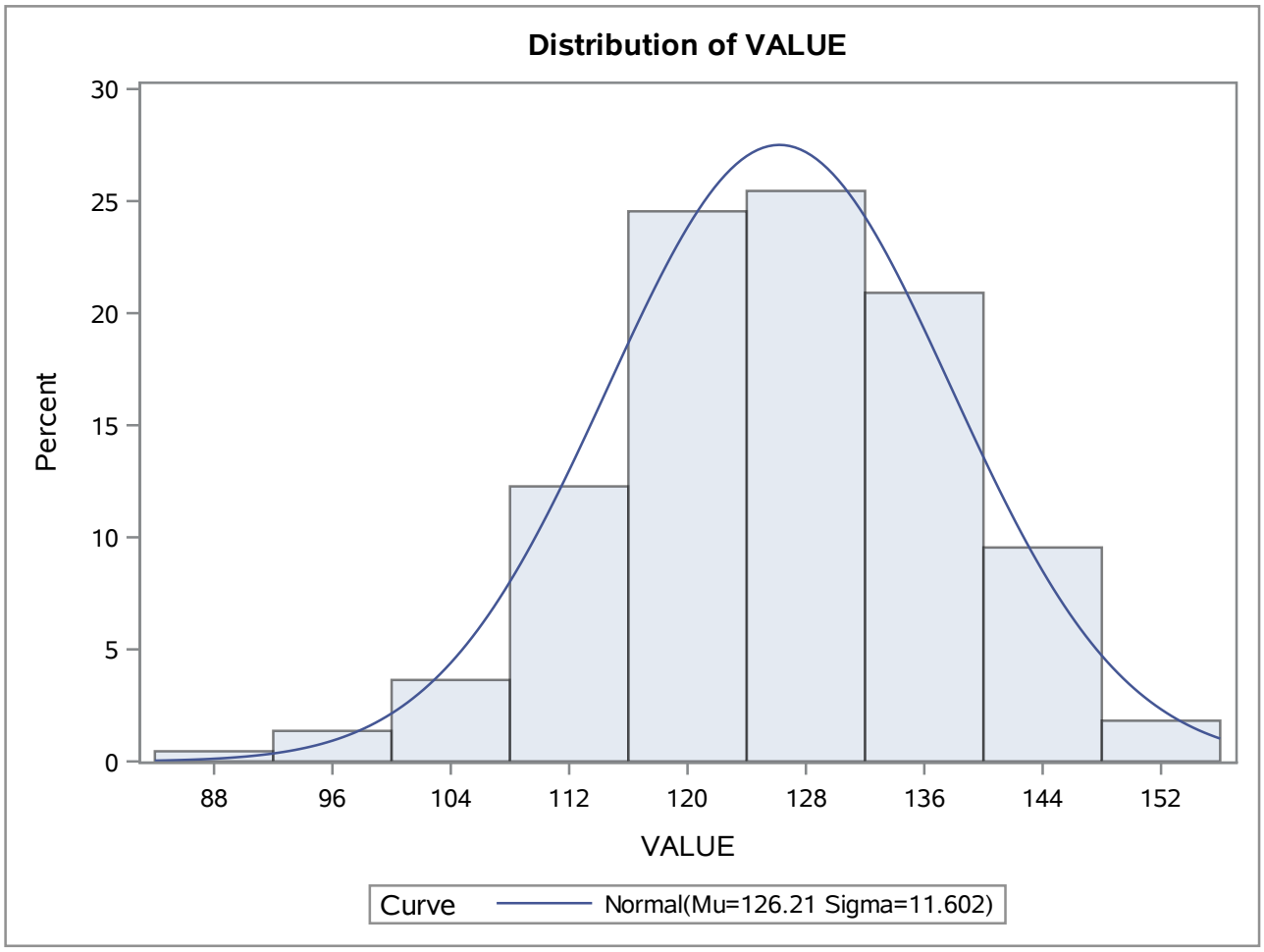
The UNIVARIATE Procedure
Variable: VALUE

RTRTN=3

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.993893	**Pr < W**	0.5093
Kolmogorov-Smirnov	**D**	0.028123	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.028136	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.24813	**Pr > A-Sq**	>0.2500

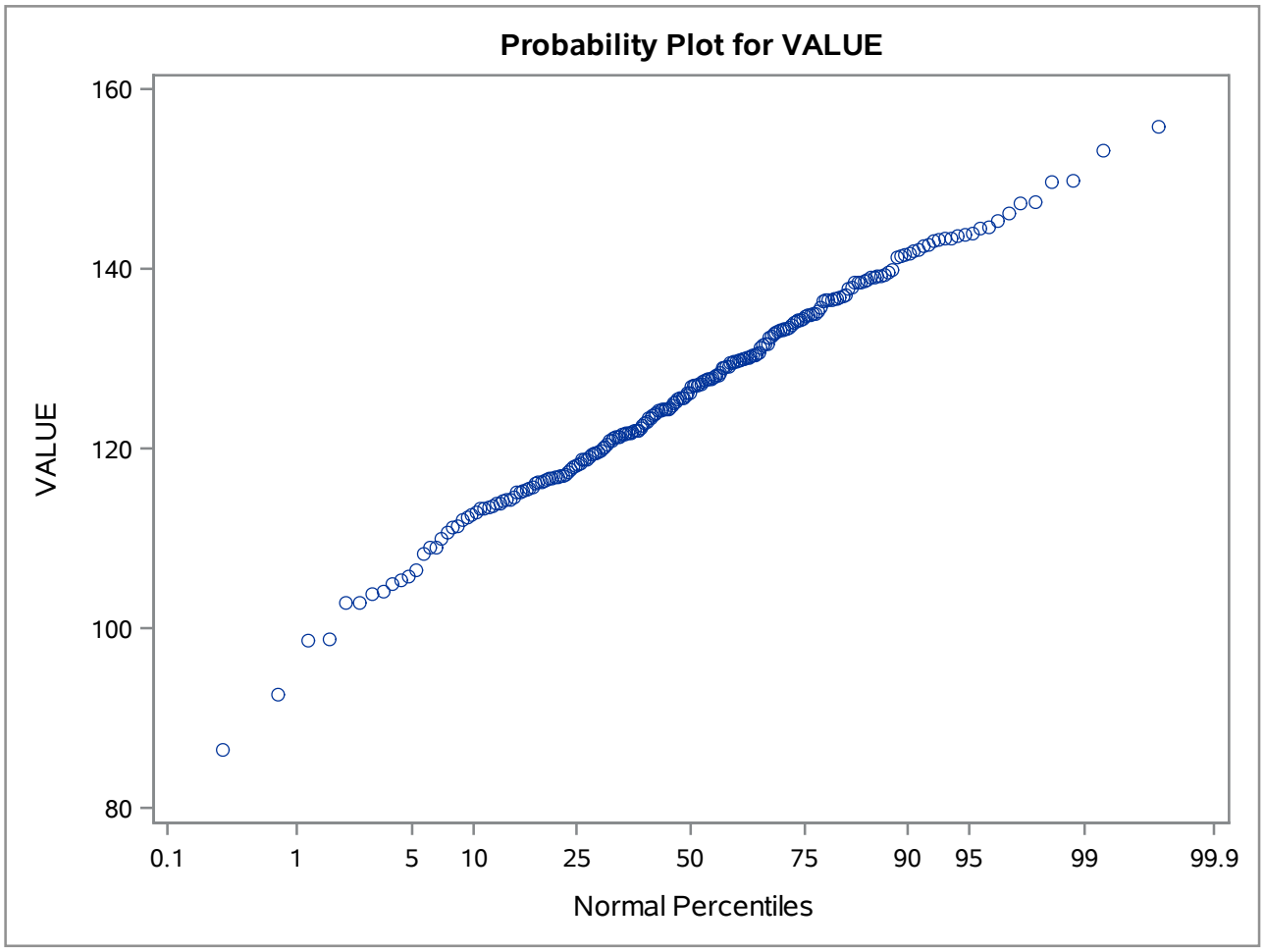
The UNIVARIATE Procedure

RTRTN=3



The UNIVARIATE Procedure

RTRTN=3



STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 1

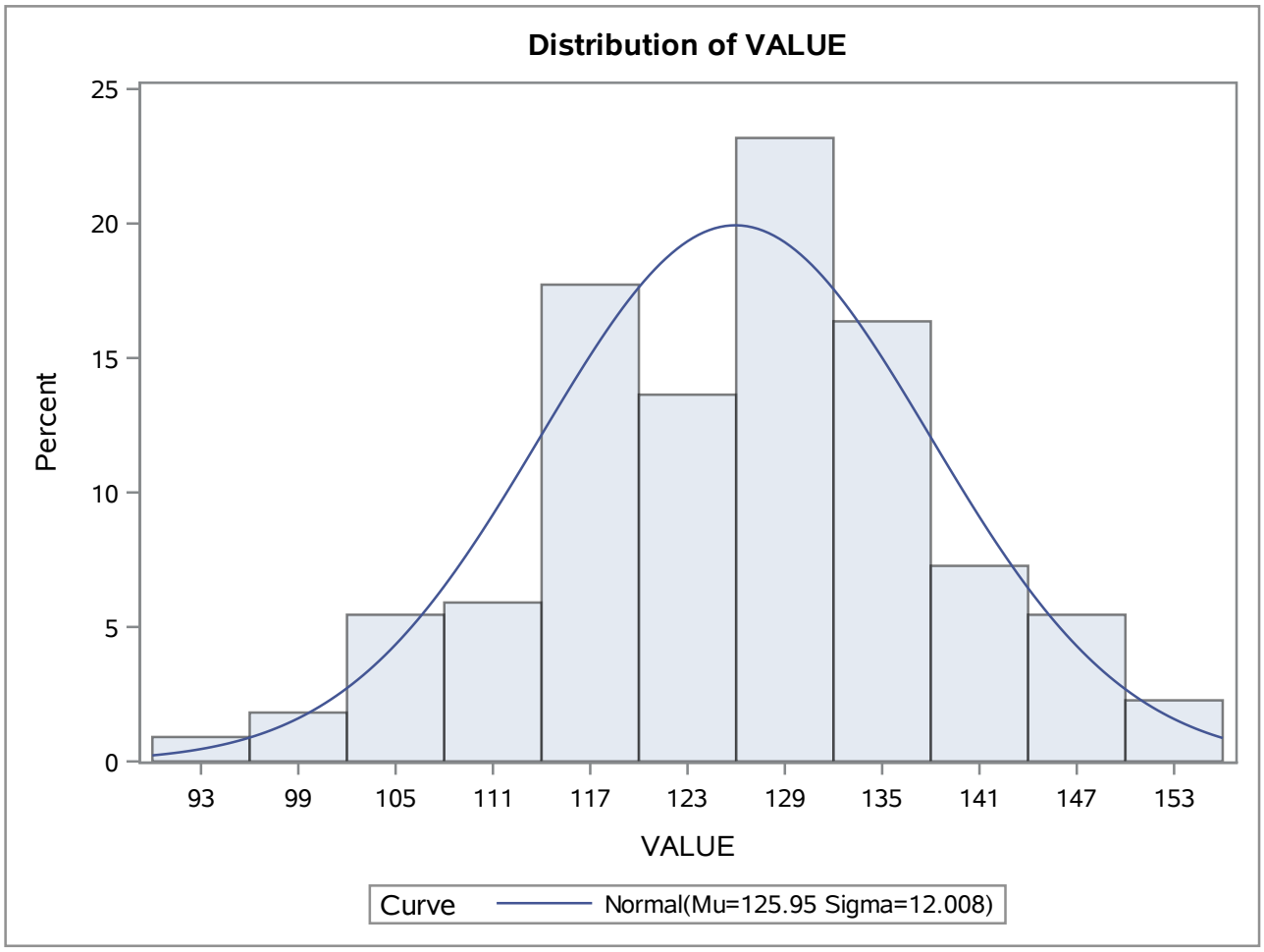
The UNIVARIATE Procedure
Variable: VALUE

RTRTN=4

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.99425	**Pr < W**	0.5645
Kolmogorov-Smirnov	**D**	0.049239	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.052537	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.283503	**Pr > A-Sq**	>0.2500

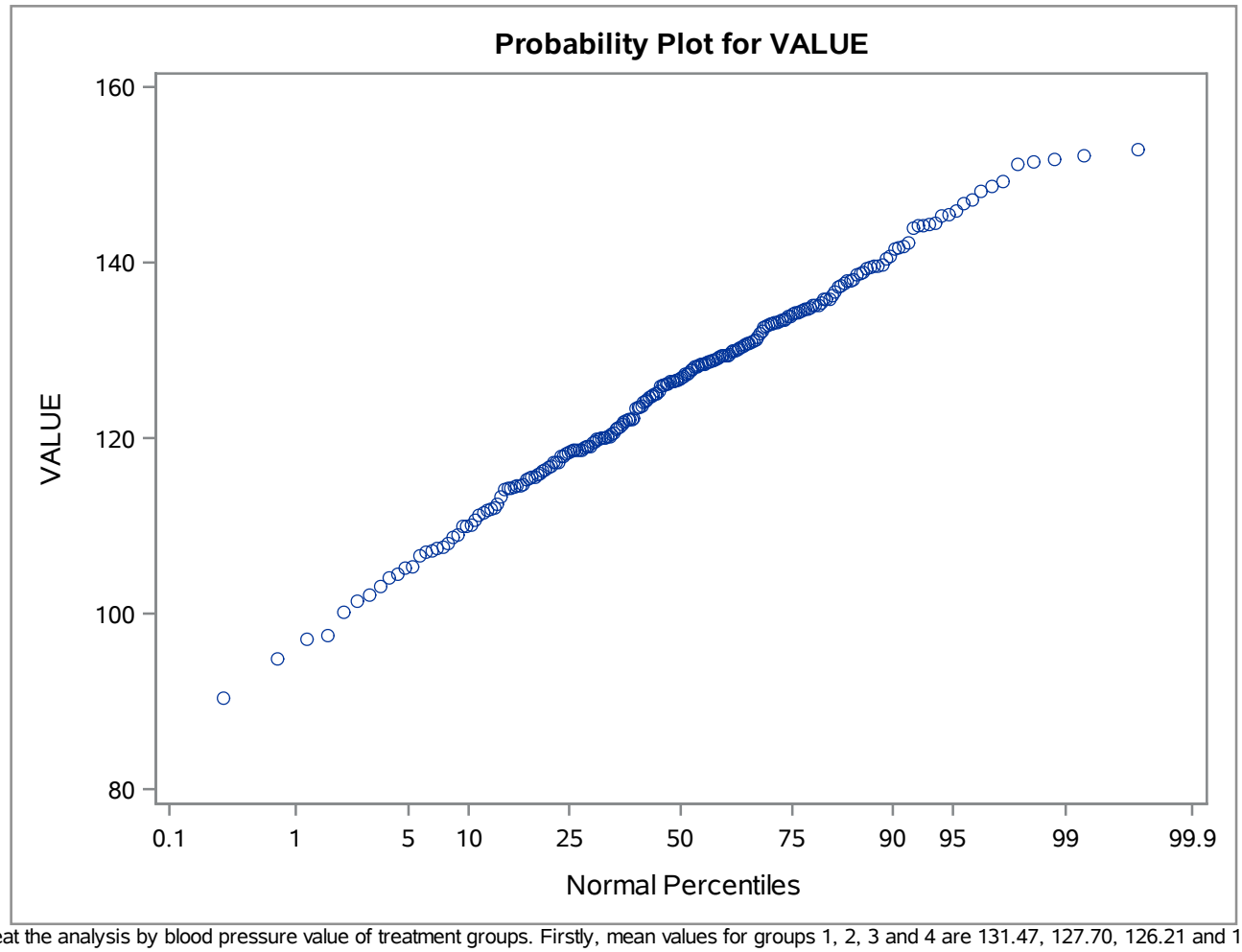
The UNIVARIATE Procedure

RTRTN=4



The UNIVARIATE Procedure

RTRTN=4



Now we repeat the analysis by blood pressure value of treatment groups. Firstly, mean values for groups 1, 2, 3 and 4 are 131.47, 127.70, 126.21 and 125.95, respectively. The mean blood pressure values decreases as the groups progress. The standard deviation and range for groups 1, 2, 3 and 4 are (131.21, 66.95), (127.43, 71.74), (126.55, 69.32) and (126.79, 62.53), respectively. We can see that there is not too large a difference in standard deviation and range between groups.According to the four tests that check for normality, all four treatment groups of blood pressure values are normally distributed. The p-value for all tests is greater than 0.05, indicating that we fail to reject H0, which is the data is normally distributed.

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 2

The TTEST Procedure

Variable: VALUE

RTRTN	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		131.5	129.8	133.1	12.2674	11.2183	13.5348
4		126.0	124.4	127.5	12.0077	10.9807	13.2483
Diff (1-2)	Pooled	5.5159	3.6082	Infty	12.1383	11.3850	12.9992
Diff (1-2)	Satterthwaite	5.5159	3.6082	Infty			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	438	4.77	<.0001
Satterthwaite	Unequal	437.8	4.77	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	219	219	1.04	0.7518

According to Excercise 1, we found that all the four treatment groups are normally distributed, hence in order to compare two population means, we will have to run the proc ttest procedure. And since we want to find out whether the reference group (rtrtn=1) has a significantly larger mean than ABC123 80mg (rtrtn=4) as the alternative hypothesis, we will have to run the upper tail test. The p-value of Folded F (equality of variances) is 0.75, which is much larger than 0.05. Hence, we refer to the Pooled section of the table above. The Pooled p-value is <0.05, therefore we reject the null hypothesis and conclude that the mean of the reference group (rtrtn=1) is significantly larger than the mean of ABC123 80mg (rtrtn=4).

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 3

The FREQ Procedure

Frequency Expected	Table of RTRTN by responder		
	responder		
RTRTN	0	1	Total
1	184 161.75	36 58.25	220
2	160 161.75	60 58.25	220
3	153 161.75	67 58.25	220
4	150 161.75	70 58.25	220
Total	647	233	880

Statistics for Table of RTRTN by responder

Statistic	DF	Value	Prob
Chi-Square	3	16.6425	0.0008
Likelihood Ratio Chi-Square	3	17.6875	0.0005
Mantel-Haenszel Chi-Square	1	13.8552	0.0002
Phi Coefficient		0.1375	
Contingency Coefficient		0.1362	
Cramer's V		0.1375	

Sample Size = 880

From the contingency table between 'rtrtn' and 'responder', we can see that the expected value of the four groups is the same based on the responder group. Responder group 1 has a smaller expectated values and frequency compared to responder group 0. We can check association between the two categorical variables through the asymptotic tests. The Chi-Square and the Likelihood Ratio Chi-Square p-values are <0.05, therefore we can reject H0, which is the two are independent. We can conclude that there exists significant associations between the two variables. We cannot apply the Mantel-Haenzel Test because 'rtrtn' is an ordinal variable (the differences between each category are not constant). The magnitutde of the relationship between the two variables can be found in the Phi Coefficient, Contigency Coefficient and Cramer's V. The values of each can be rounded to 0.14. We conclude that there is a small association between 'rtrtn' and 'responder'.

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 3

The FREQ Procedure

Frequency Expected Row Pct

Table of RTRTN by responder			
	responder		
RTRTN	0	1	Total
1	184 167 83.64	36 53 16.36	220
4	150 167 68.18	70 53 31.82	220
Total	334	106	440

Statistics for Table of RTRTN by responder

Statistic	DF	Value	Prob
Chi-Square	1	14.3667	0.0002
Likelihood Ratio Chi-Square	1	14.5679	0.0001
Continuity Adj. Chi-Square	1	13.5341	0.0002
Mantel-Haenszel Chi-Square	1	14.3341	0.0002
Phi Coefficient		0.1807	
Contingency Coefficient		0.1778	
Cramer's V		0.1807	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	184
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	0.0001
Table Probability (P)	<.0001
Two-sided Pr <= P	0.0002

Column 1 Risk Estimates						
	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.8364	0.0249	0.7875	0.8852	0.7807	0.8827
Row 2	0.6818	0.0314	0.6203	0.7434	0.6158	0.7428
Total	0.7591	0.0204	0.7191	0.7990	0.7163	0.7983
Difference	0.1545	0.0401	0.0759	0.2331		
Difference is (Row 1 - Row 2)						

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 3

The FREQ Procedure

Statistics for Table of RTRTN by responder

Column 2 Risk Estimates						
	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.1636	0.0249	0.1148	0.2125	0.1173	0.2193
Row 2	0.3182	0.0314	0.2566	0.3797	0.2572	0.3842
Total	0.2409	0.0204	0.2010	0.2809	0.2017	0.2837
Difference	-0.1545	0.0401	-0.2331	-0.0759		
Difference is (Row 1 - Row 2)						

Sample Size = 440

Now, we just compare the reference group and ABC123 80mg with the responder groups. The proportional difference between responder group 0 is 15.46 and between responder group 1 is -15.46. We can see if the difference is significant using the confidence intervals. The confidence interval (0.076, 0.23) does not include 0, hence we can conclude that ABC123 80mg has larger proportions in responder group 1 than the reference group.

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 4

The GLMSELECT Procedure

Data Set	WORK.BLOOD
Dependent Variable	VALUE
Selection Method	Stepwise
Select Criterion	Significance Level
Stop Criterion	Significance Level
Entry Significance Level (SLE)	0.05
Stay Significance Level (SLS)	0.05
Effect Hierarchy Enforced	None

Number of Observations Read	880
Number of Observations Used	880

Class Level Information		
Class	**Levels**	**Values**
SEX	2	1 2
RACE	4	Asian Black Native A White
RTRTN	4	1 2 3 4
SITE	20	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Dimensions	
Number of Effects	5
Number of Parameters	31

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 4

The GLMSELECT Procedure

Stepwise Selection Summary						
Step	**Effect Entered**	**Effect Removed**	**Number Effects In**	**Number Parms In**	**F Value**	**Pr > F**
0	**Intercept**		1	1	0.00	1.0000
1	**RTRTN**		2	4	9.64	<.0001
2	**SEX**		3	5	12.86	0.0004
3	**SITE**		4	24	1.73	0.0272

Selection stopped because the candidate for entry has SLE > 0.05 and the candidate for removal has SLS < 0.05.

Stop Details					
Candidate For	**Effect**	**Candidate Significance**		**Compare Significance**	
Entry	RACE	0.6086	>	0.0500	(SLE)
Removal	SITE	0.0272	<	0.0500	(SLS)

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 4

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 3).

Effects:	Intercept SEX RTRTN SITE

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	23	10845	471.52787	3.29
Error	856	122709	143.35143	
Corrected Total	879	133554		

Root MSE	11.97295
Dependent Mean	127.83352
R-Square	0.0812
Adj R-Sq	0.0565
AIC	5275.12990
AICC	5276.65215
SBC	4507.84802

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	129.512394	1.958473	66.13
SEX 1	1	-2.732233	0.817532	-3.34
SEX 2	0	0	.	.
RTRTN 1	1	5.602844	1.141872	4.91
RTRTN 2	1	1.858000	1.142065	1.63
RTRTN 3	1	0.545961	1.144771	0.48
RTRTN 4	0	0	.	.
SITE 1	1	1.875834	2.553249	0.73
SITE 2	1	-5.840580	2.552911	-2.29
SITE 3	1	1.419698	2.553249	0.56
SITE 4	1	-3.362196	2.555074	-1.32
SITE 5	1	0.057198	2.553249	0.02
SITE 6	1	-1.004064	2.554331	-0.39
SITE 7	1	-0.042701	2.554331	-0.02

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 4

The GLMSELECT Procedure
Selected Model

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
SITE 8	1	-1.768583	2.559394	-0.69
SITE 9	1	-4.583939	2.553249	-1.80
SITE 10	1	-2.999367	2.556965	-1.17
SITE 11	1	-0.366540	2.552708	-0.14
SITE 12	1	-4.978786	2.555074	-1.95
SITE 13	1	-5.733180	2.558112	-2.24
SITE 14	1	-1.965655	2.554331	-0.77
SITE 15	1	-1.474925	2.553249	-0.58
SITE 16	1	-5.664115	2.553722	-2.22
SITE 17	1	-4.013105	2.555074	-1.57
SITE 18	1	-3.003080	2.552911	-1.18
SITE 19	1	-2.903131	2.552708	-1.14
SITE 20	0	0	.	.

We used stepwise selection to investigate the main effects. It suggested that we drop the variable 'race', since it is not significant. The remaining variables are 'sex', 'rtrtn' and 'site'. Now, we check the interaction between those 3 statistically significant variables.

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 4

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
SEX	2	1 2
RTRTN	4	1 2 3 4
SITE	20	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Number of Observations Read	880
Number of Observations Used	880

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 4

The GLM Procedure

Dependent Variable: VALUE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	159	31688.5010	199.2987	1.41	0.0019
Error	720	101865.4611	141.4798		
Corrected Total	879	133553.9621			

R-Square	Coeff Var	Root MSE	VALUE Mean
0.237271	9.304702	11.89453	127.8335

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SEX	1	1937.736947	1937.736947	13.70	0.0002
RTRTN	3	4201.225734	1400.408578	9.90	<.0001
SEX*RTRTN	3	162.122603	54.040868	0.38	0.7660
SITE	19	4751.550870	250.081625	1.77	0.0227
SEX*SITE	19	2164.555128	113.923954	0.81	0.7022
RTRTN*SITE	57	8585.708923	150.626472	1.06	0.3521
SEX*RTRTN*SITE	57	9885.600761	173.431592	1.23	0.1290

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SEX	1	2127.554532	2127.554532	15.04	0.0001
RTRTN	3	3354.476431	1118.158810	7.90	<.0001
SEX*RTRTN	3	256.319003	85.439668	0.60	0.6126
SITE	19	4631.274706	243.751300	1.72	0.0283
SEX*SITE	19	1783.048464	93.844656	0.66	0.8564
RTRTN*SITE	57	8153.740626	143.048081	1.01	0.4554
SEX*RTRTN*SITE	57	9885.600761	173.431592	1.23	0.1290

From Type 1 SS and Type 3 SS, we can conclude that the interaction term is not significant. We now focus on the final model.

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 4

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
SEX	2	1 2
RTRTN	4	1 2 3 4
SITE	20	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Number of Observations Read	880
Number of Observations Used	880

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 4

The GLM Procedure

Dependent Variable: VALUE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	23	10845.1411	471.5279	3.29	<.0001
Error	856	122708.8210	143.3514		
Corrected Total	879	133553.9621			

R-Square	Coeff Var	Root MSE	VALUE Mean
0.081204	9.366045	11.97295	127.8335

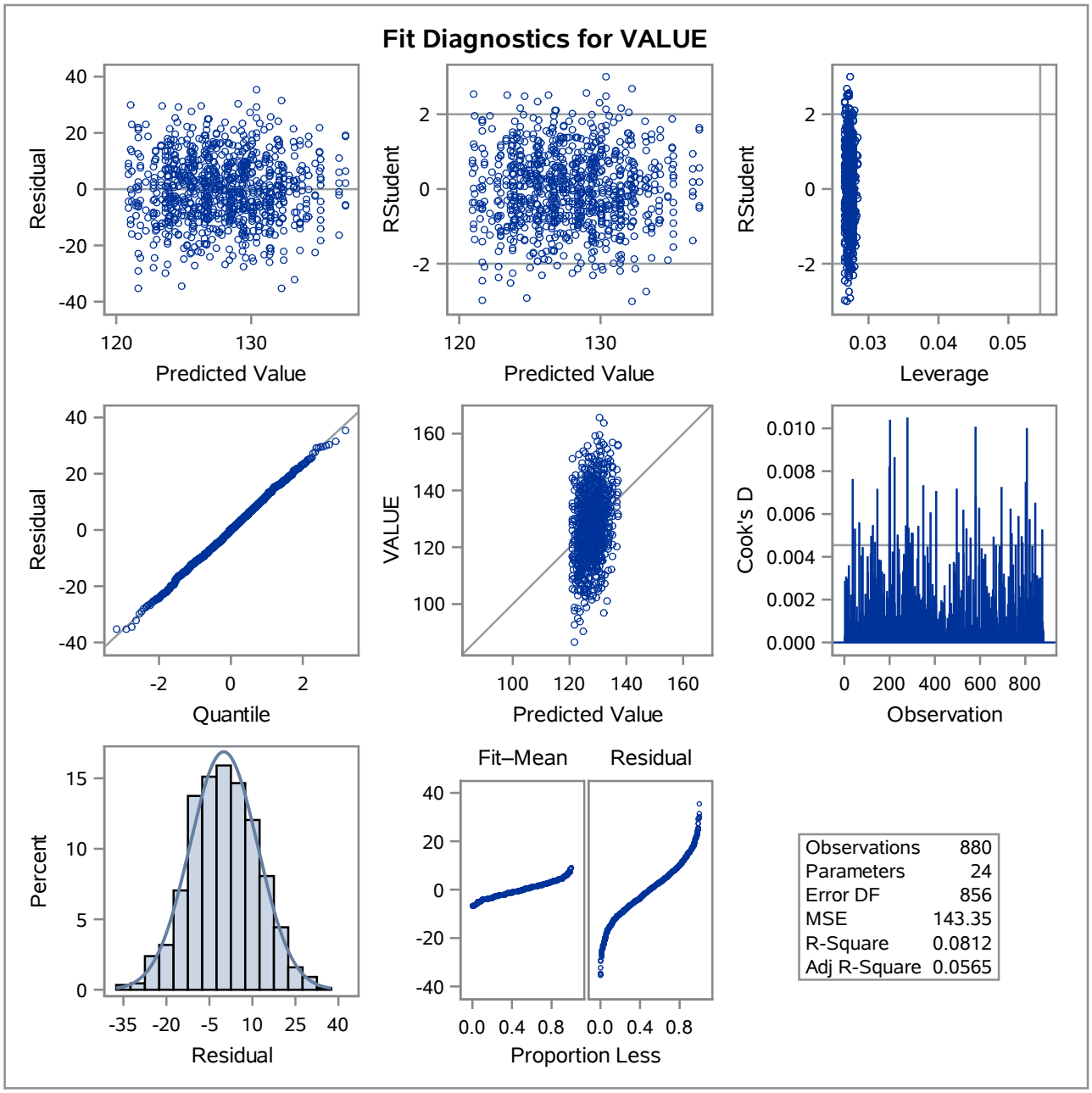
Source	DF	Type I SS	Mean Square	F Value	Pr > F
SEX	1	1937.736947	1937.736947	13.52	0.0003
RTRTN	3	4201.225734	1400.408578	9.77	<.0001
SITE	19	4706.178404	247.693600	1.73	0.0272

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SEX	1	1601.135987	1601.135987	11.17	0.0009
RTRTN	3	4204.846218	1401.615406	9.78	<.0001
SITE	19	4706.178404	247.693600	1.73	0.0272

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 4

The GLM Procedure

Dependent Variable: VALUE

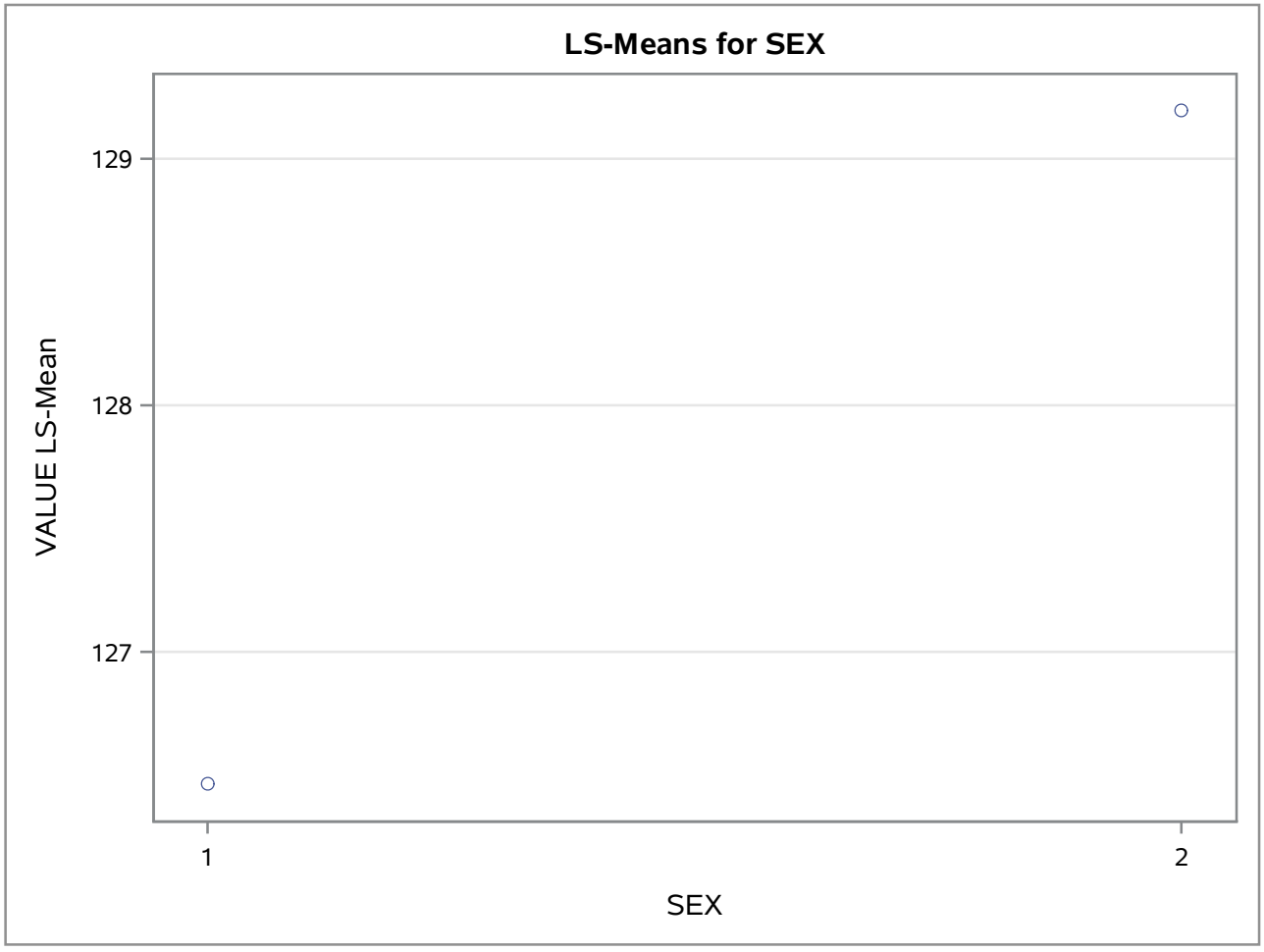


The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

SEX	VALUE LSMEAN	H0:LSMean1=LSMean2	
		t Value	Pr > |t|
1	126.464301	-3.34	0.0009
2	129.196534		

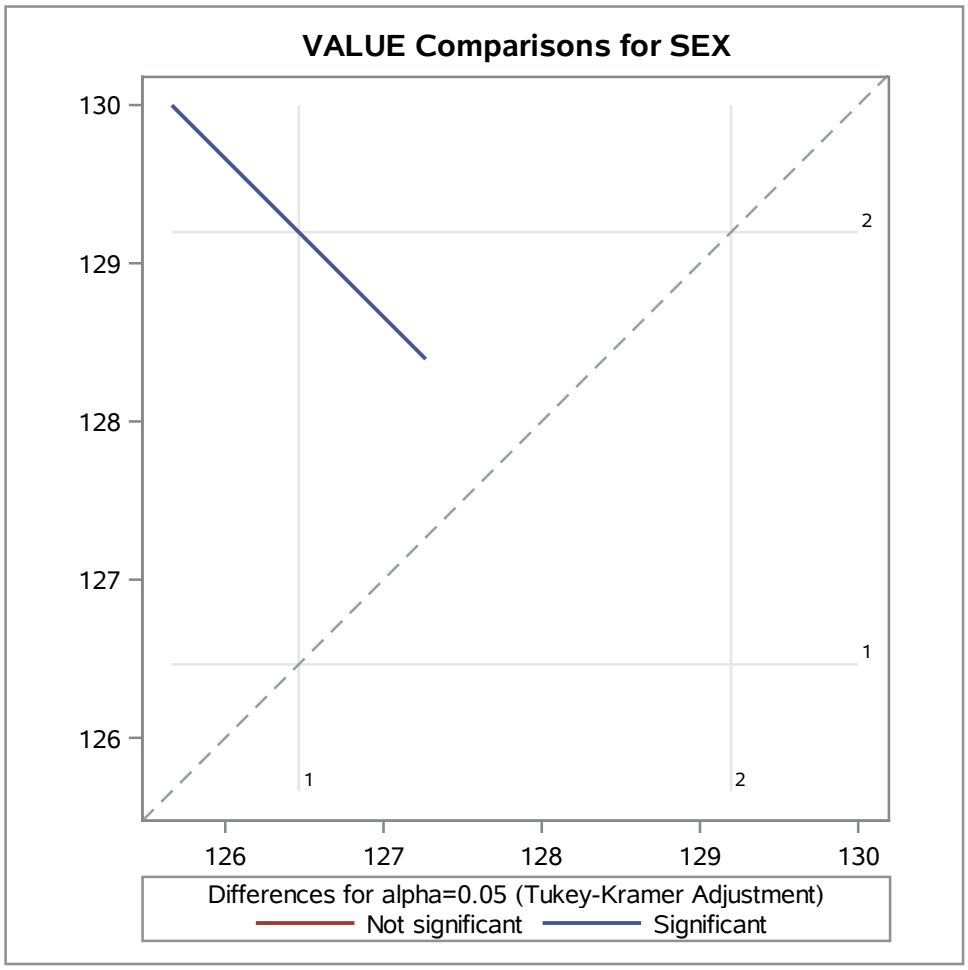
SEX	VALUE LSMEAN	95% Confidence Limits	
1	126.464301	125.335514	127.593089
2	129.196534	128.070342	130.322726

Least Squares Means for Effect SEX				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-2.732233	-4.336835	-1.127631



STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 4

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer



STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 4

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

RTRTN	VALUE LSMEAN	LSMEAN Number
1	131.431560	1
2	127.686717	2
3	126.374677	3
4	125.828717	4

Least Squares Means for Effect RTRTN
t for H0: LSMean(i)=LSMean(j) / Pr > |t|

Dependent Variable: VALUE

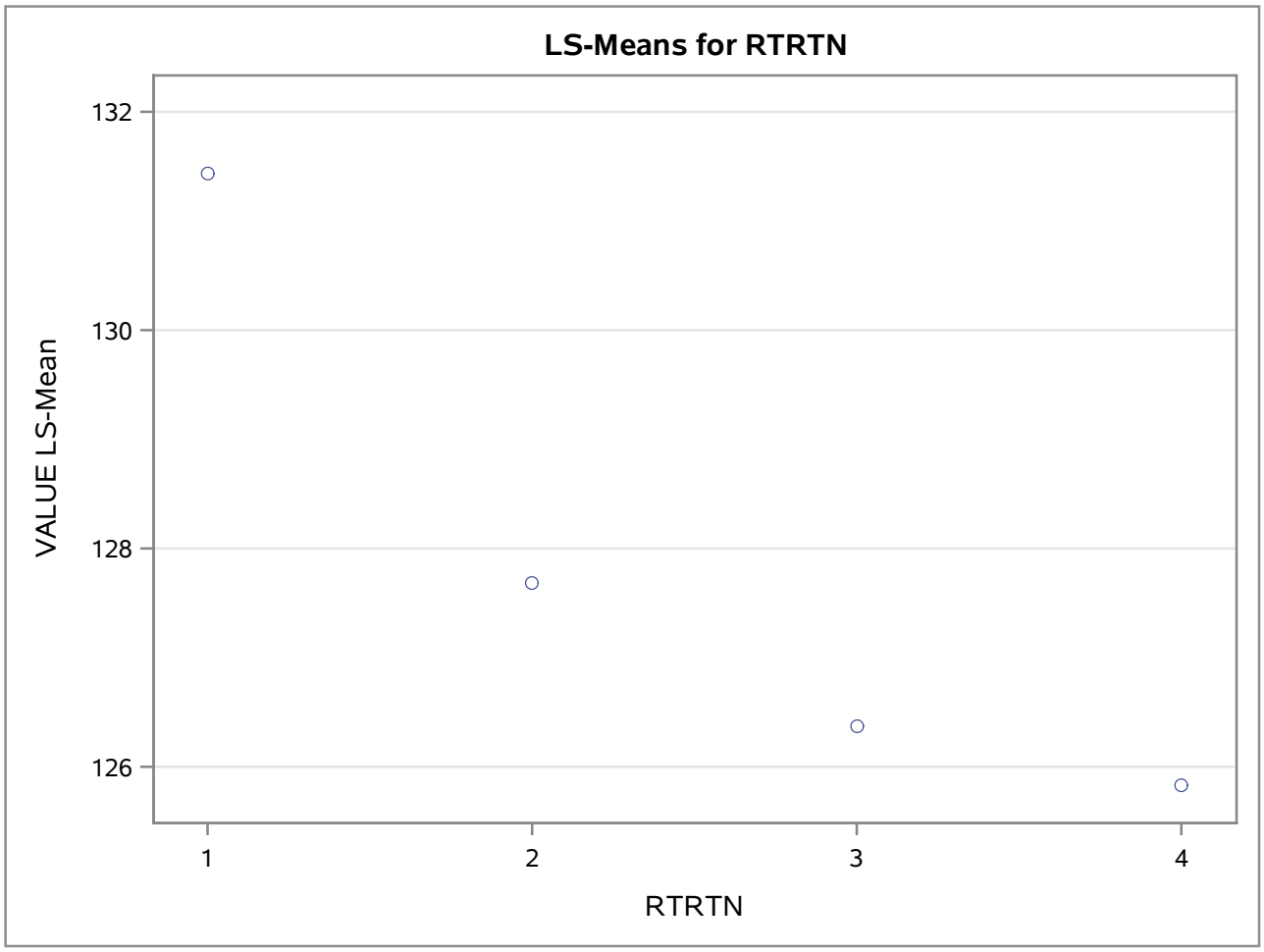
i/j	1	2	3	4
1		3.280347 0.0059	4.423744 <.0001	4.906718 <.0001
2	-3.28035 0.0059		1.148132 0.6597	1.626877 0.3640
3	-4.42374 <.0001	-1.14813 0.6597		0.476917 0.9642
4	-4.90672 <.0001	-1.62688 0.3640	-0.47692 0.9642	

RTRTN	VALUE LSMEAN	95% Confidence Limits	
1	131.431560	129.847055	133.016066
2	127.686717	126.102346	129.271088
3	126.374677	124.787489	127.961866
4	125.828717	124.242685	127.414749

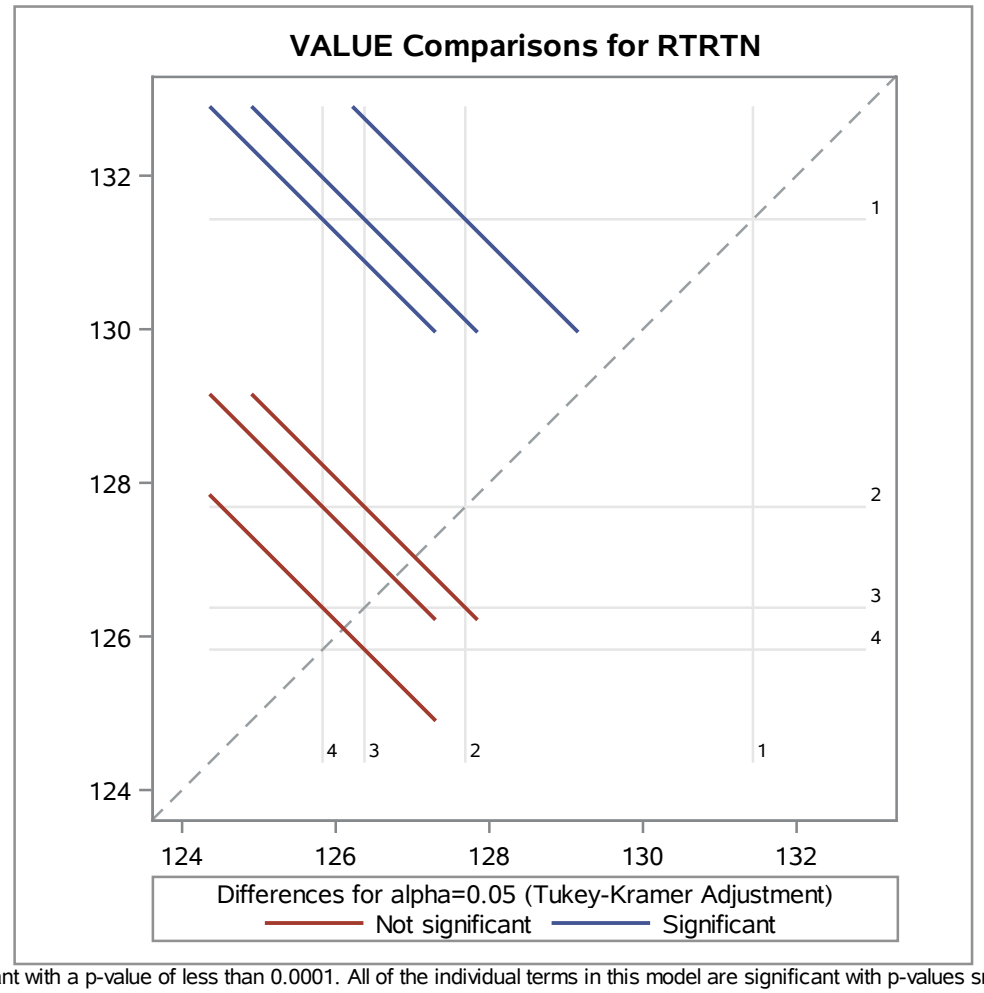
Least Squares Means for Effect RTRTN

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	3.744843	0.806323	6.683364
1	3	5.056883	2.114442	7.999324
1	4	5.602844	2.663623	8.542065
2	3	1.312040	-1.629469	4.253548
2	4	1.858000	-1.081719	4.797720
3	4	0.545961	-2.400722	3.492643

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer



The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer



The model chosen is significant with a p-value of less than 0.0001. All of the individual terms in this model are significant with p-values smaller than 0.05. The variation explained by this model is 8.12%. There is a miniscule difference in the results of Type 1 SS and Type 3 SS. Now we investigate the variables to find the significant differences. For 'sex', we can see that sex 2 has a higher value than sex 1. This pair is significant.For 'rtrtn', we can see that mean values keep decreasing as the groups progress. We can see that the pairs (rtrtn=2, rtrtn=3), (rtrtn=2, rtrtn=4) and (rtrtn=3, rtrtn=4) are not significant. Their p-values is larger 0.05. rtrtn=1 forms a significant pair with all the other 3 groups.For 'site', none of the pairs are significant (all have a p-value > 0.05).

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 4

The UNIVARIATE Procedure
Variable: resid

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.998765	**Pr < W**	0.8165
Kolmogorov-Smirnov	**D**	0.020656	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.034131	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.228641	**Pr > A-Sq**	>0.2500

The normality tests for residuals show that the residuals are normally distributed (p-value > 0.05). The model assumptions are valid.

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 5

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
SEX	2	1 2
RACE	4	Asian Black Native A White
RTRTN	4	1 2 3 4
SITE	20	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Number of Observations Read	880
Number of Observations Used	880

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 5

The GLM Procedure

Dependent Variable: VALUE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	27	56038.3279	2075.4936	22.81	<.0001
Error	852	77515.6342	90.9808		
Corrected Total	879	133553.9621			

R-Square	Coeff Var	Root MSE	VALUE Mean
0.419593	7.461568	9.538385	127.8335

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SEX	1	1937.73695	1937.73695	21.30	<.0001
RACE	3	290.03286	96.67762	1.06	0.3642
RTRTN	3	4216.63787	1405.54596	15.45	<.0001
SITE	19	4663.41908	245.44311	2.70	0.0001
BASE	1	44930.50116	44930.50116	493.85	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SEX	1	1840.71634	1840.71634	20.23	<.0001
RACE	3	188.25744	62.75248	0.69	0.5584
RTRTN	3	5105.20845	1701.73615	18.70	<.0001
SITE	19	4170.90016	219.52106	2.41	0.0007
BASE	1	44930.50116	44930.50116	493.85	<.0001

For this ANCOVA model, we fit 'sex', 'race', 'rtrtn' and 'site' with the covariate 'base'.

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 5

The GLMSELECT Procedure

Data Set	WORK.BLOOD
Dependent Variable	VALUE
Selection Method	Stepwise
Select Criterion	Significance Level
Stop Criterion	Significance Level
Entry Significance Level (SLE)	0.05
Stay Significance Level (SLS)	0.05
Effect Hierarchy Enforced	None

Number of Observations Read	880
Number of Observations Used	880

Class Level Information		
Class	**Levels**	**Values**
SEX	2	1 2
RACE	4	Asian Black Native A White
RTRTN	4	1 2 3 4
SITE	20	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Dimensions	
Number of Effects	6
Number of Parameters	32

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 5

The GLMSELECT Procedure

Stepwise Selection Summary						
Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	F Value	Pr > F
0	Intercept		1	1	0.00	1.0000
1	BASE		2	2	436.10	<.0001
2	RTRTN		3	5	17.61	<.0001
3	SEX		4	6	22.96	<.0001
4	SITE		5	25	2.49	0.0004

Selection stopped because the candidate for entry has SLE > 0.05 and the candidate for removal has SLS < 0.05.

Stop Details					
Candidate For	Effect	Candidate Significance		Compare Significance	
Entry	RACE	0.5584	>	0.0500	(SLE)
Removal	SITE	0.0004	<	0.0500	(SLS)

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 5

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 4).

Effects:	Intercept SEX RTRTN SITE BASE

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	24	55850	2327.08627	25.61
Error	855	77704	90.88174	
Corrected Total	879	133554		

Root MSE	9.53319
Dependent Mean	127.83352
R-Square	0.4182
Adj R-Sq	0.4019
AIC	4875.05007
AICC	4876.69602
SBC	4112.54811

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-13.667535	6.620411	-2.06
SEX 1	1	-2.935376	0.651005	-4.51
SEX 2	0	0	.	.
RTRTN 1	1	6.622694	0.910344	7.27
RTRTN 2	1	3.351580	0.911818	3.68
RTRTN 3	1	1.963506	0.913721	2.15
RTRTN 4	0	0	.	.
SITE 1	1	3.959005	2.035122	1.95
SITE 2	1	-2.446944	2.038411	-1.20
SITE 3	1	0.814686	2.033150	0.40
SITE 4	1	-0.491596	2.038506	-0.24
SITE 5	1	0.380479	2.033020	0.19
SITE 6	1	0.295423	2.034667	0.15
SITE 7	1	1.604897	2.035176	0.79

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 5

The GLMSELECT Procedure
Selected Model

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
SITE 8	1	1.004447	2.041667	0.49
SITE 9	1	-3.815850	2.033261	-1.88
SITE 10	1	-1.221882	2.037493	-0.60
SITE 11	1	2.771067	2.037422	1.36
SITE 12	1	-4.390241	2.034593	-2.16
SITE 13	1	-2.017133	2.043674	-0.99
SITE 14	1	-1.362388	2.034010	-0.67
SITE 15	1	1.760251	2.038159	0.86
SITE 16	1	-2.183887	2.039350	-1.07
SITE 17	1	-3.675972	2.034477	-1.81
SITE 18	1	1.976391	2.044978	0.97
SITE 19	1	-0.597120	2.035177	-0.29
SITE 20	0	0	.	.
BASE	1	0.850936	0.038239	22.25

'sex', 'rtrtn' and 'site' are retained in the final model after the stepwise procedure with a p-value <0.05. 'race' has been dropped since its p-value was too large.

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 5

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
SEX	2	1 2
RTRTN	4	1 2 3 4
SITE	20	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Number of Observations Read	880
Number of Observations Used	880

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 5

The GLM Procedure

Dependent Variable: VALUE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	24	55850.0705	2327.0863	25.61	<.0001
Error	855	77703.8916	90.8817		
Corrected Total	879	133553.9621			

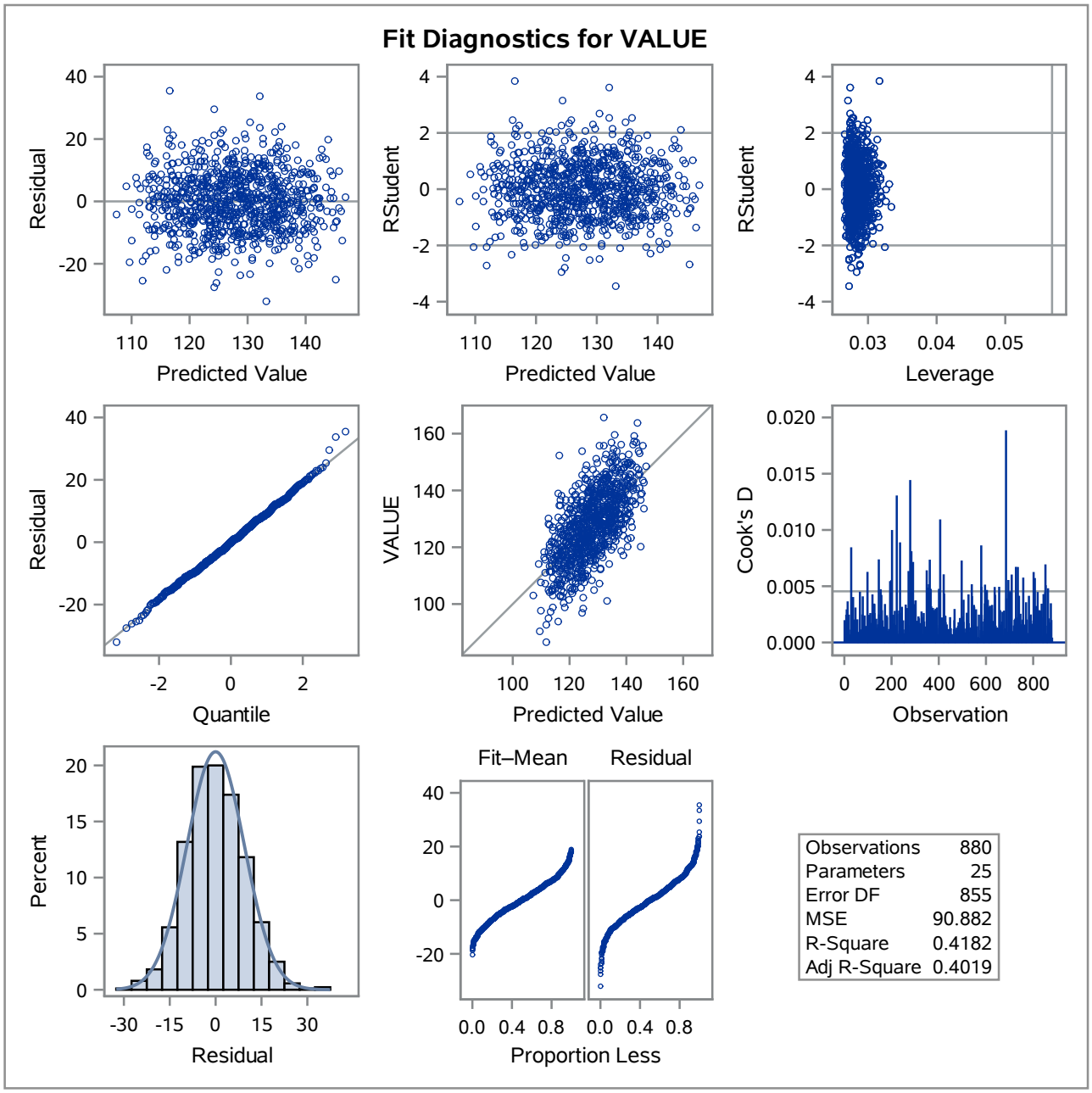
R-Square	Coeff Var	Root MSE	VALUE Mean
0.418184	7.457505	9.533192	127.8335

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SEX	1	1937.73695	1937.73695	21.32	<.0001
RTRTN	3	4201.22573	1400.40858	15.41	<.0001
SITE	19	4706.17840	247.69360	2.73	0.0001
BASE	1	45004.92938	45004.92938	495.20	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SEX	1	1847.71487	1847.71487	20.33	<.0001
RTRTN	3	5122.02942	1707.34314	18.79	<.0001
SITE	19	4292.96137	225.94534	2.49	0.0004
BASE	1	45004.92938	45004.92938	495.20	<.0001

The GLM Procedure

Dependent Variable: VALUE



STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 5

The UNIVARIATE Procedure
Variable: resid

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.998249	**Pr < W**	0.5170
Kolmogorov-Smirnov	**D**	0.020189	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.047935	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.316642	**Pr > A-Sq**	>0.2500

The final model is significant with a p-value of <0.0001. Both the categorical predictors and covariate are significant. 41.82% of the variation is explained by the model. This is much larger than the ANOVA model, thus we can say this model is better in terms of variation explained. The residuals in the Q-Q plot follow the diagonal line. Using the normality tests, we can see that the p-value is greater than 0.05 suggesting that the residuals satisfy the normality assumption.

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 6

The REG Procedure
Model: MODEL1
Dependent Variable: medv

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	lstat		1	0.5441	0.5441	309.376	601.62	<.0001
2	rm		2	0.0944	0.6386	143.326	131.39	<.0001
3	ptratio		3	0.0401	0.6786	74.0183	62.58	<.0001
4	dis		4	0.0117	0.6903	55.2227	18.90	<.0001
5	nox		5	0.0178	0.7081	25.5732	30.46	<.0001
6	bb		6	0.0073	0.7154	14.5796	12.80	0.0004
7	zn		7	0.0042	0.7196	9.1297	7.43	0.0066
8	crim		8	0.0026	0.7222	6.5133	4.64	0.0317

Here, we run a stepwise selection producer on the regression model. The procedure suggests that we drop the variables 'indus', 'age', and 'tax' since they are insignificant.

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 6

The REG Procedure
Model: MODEL1
Dependent Variable: medv

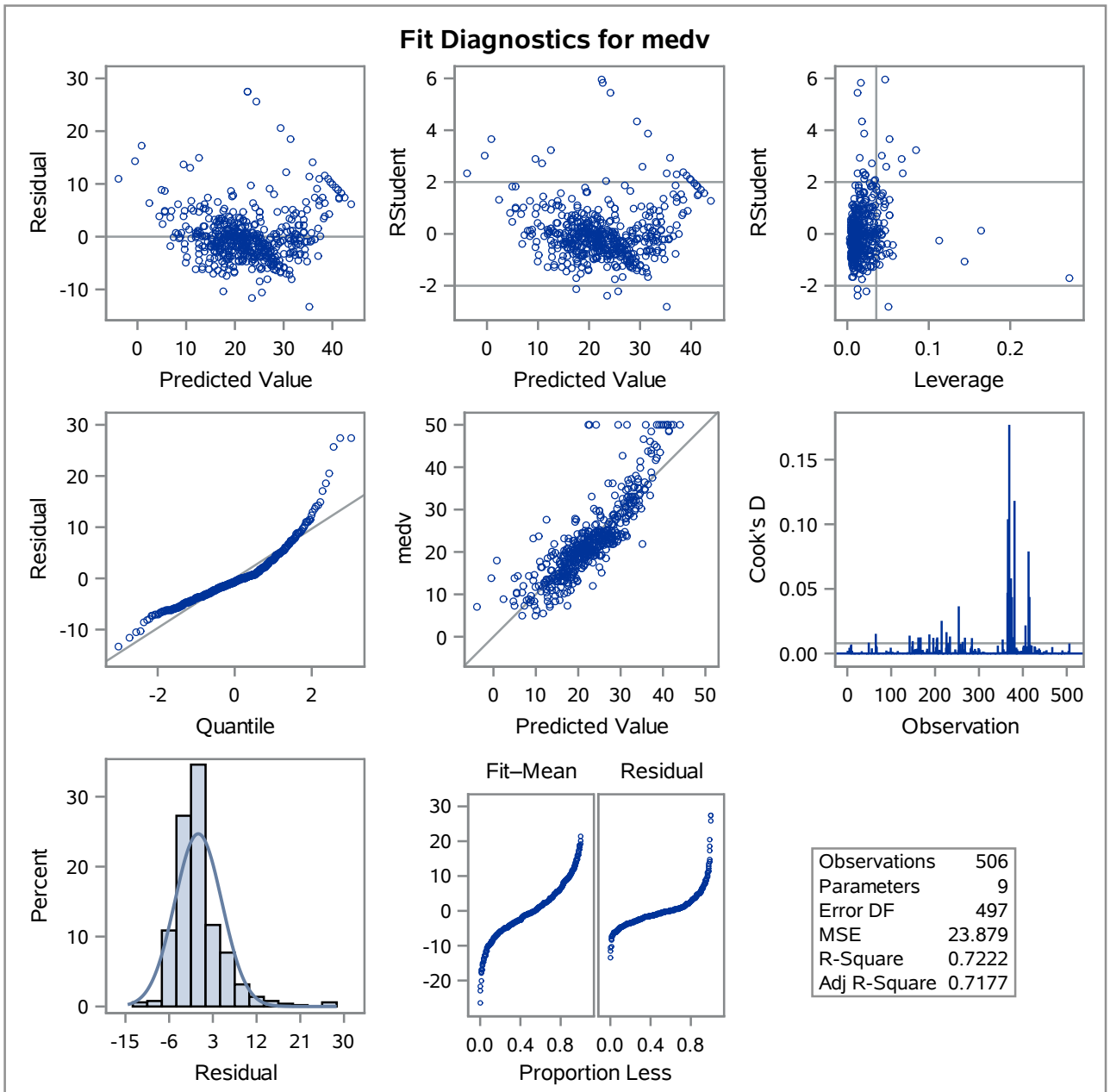
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	30848	3856.04614	161.48	<.0001
Error	497	11868	23.87913		
Corrected Total	505	42716			

Root MSE	4.88663	R-Square	0.7222
Dependent Mean	22.53281	Adj R-Sq	0.7177
Coeff Var	21.68672		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	29.54971	4.92700	6.00	<.0001	0
crim	1	-0.06609	0.03068	-2.15	0.0317	1.47310
zn	1	0.04127	0.01357	3.04	0.0025	2.11847
nox	1	-15.21364	3.25900	-4.67	<.0001	3.01606
rm	1	4.21741	0.41178	10.24	<.0001	1.77024
dis	1	-1.46380	0.19048	-7.68	<.0001	3.40240
ptratio	1	-0.87583	0.11816	-7.41	<.0001	1.38399
bb	1	0.00878	0.00271	3.24	0.0013	1.29887
lstat	1	-0.53163	0.04885	-10.88	<.0001	2.57395

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 6

The REG Procedure
Model: MODEL1
Dependent Variable: medv



We performed the variance inflation factor (VIF) test to check if any of the variables violate the multicollinearity assumption. Since we are taking a reference value of VIF < 10 is not correlated, none of the variables exceed it. Therefore, there aren't any correlations.

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 6

The REG Procedure
Model: MODEL1
Dependent Variable: medv

Number of Observations Read	506
Number of Observations Used	506

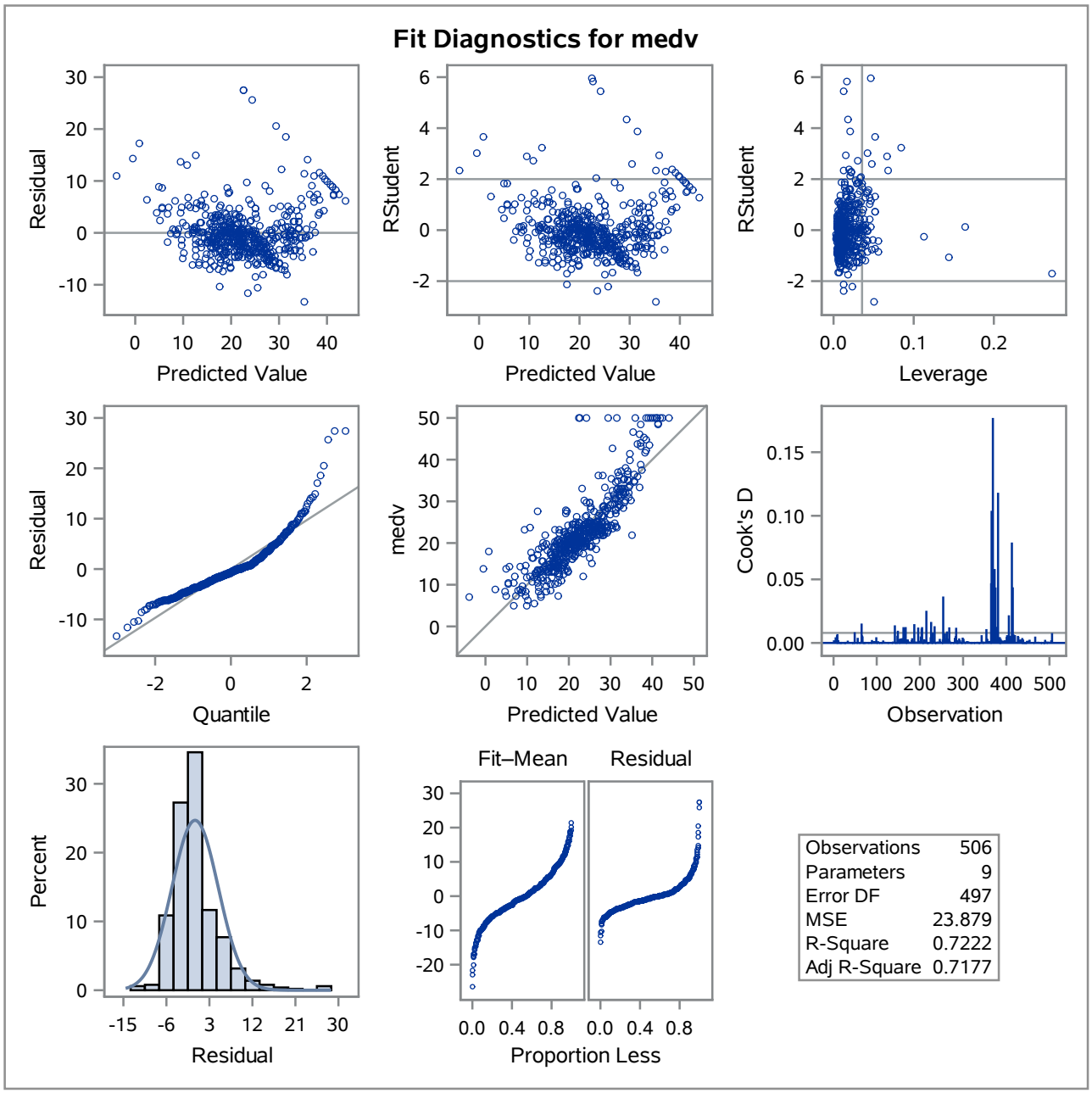
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	30848	3856.04614	161.48	<.0001
Error	497	11868	23.87913		
Corrected Total	505	42716			

Root MSE	4.88663	R-Square	0.7222
Dependent Mean	22.53281	Adj R-Sq	0.7177
Coeff Var	21.68672		

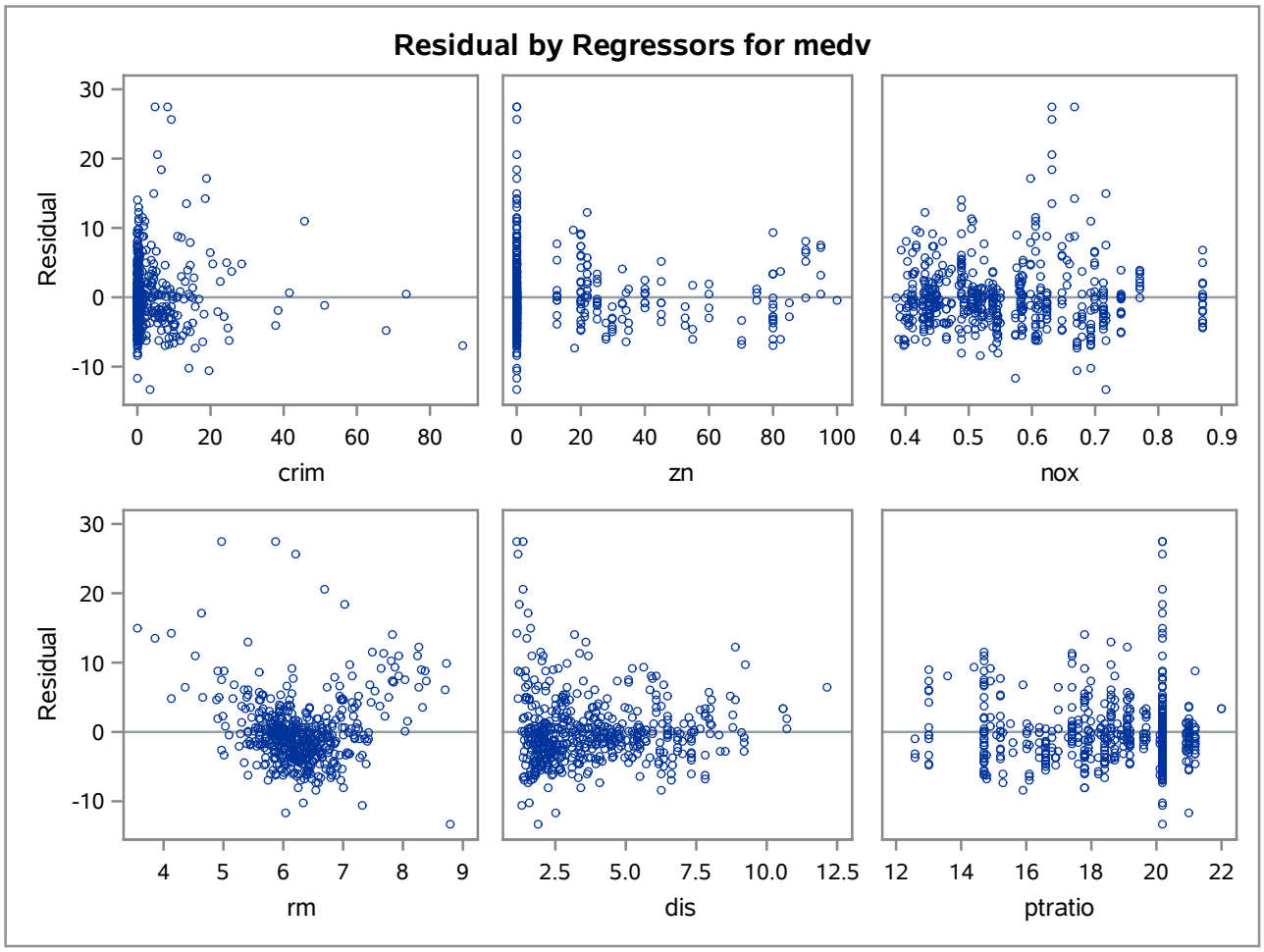
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	29.54971	4.92700	6.00	<.0001
crim	1	-0.06609	0.03068	-2.15	0.0317
zn	1	0.04127	0.01357	3.04	0.0025
nox	1	-15.21364	3.25900	-4.67	<.0001
rm	1	4.21741	0.41178	10.24	<.0001
dis	1	-1.46380	0.19048	-7.68	<.0001
ptratio	1	-0.87583	0.11816	-7.41	<.0001
bb	1	0.00878	0.00271	3.24	0.0013
lstat	1	-0.53163	0.04885	-10.88	<.0001

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 6

The REG Procedure
Model: MODEL1
Dependent Variable: medv

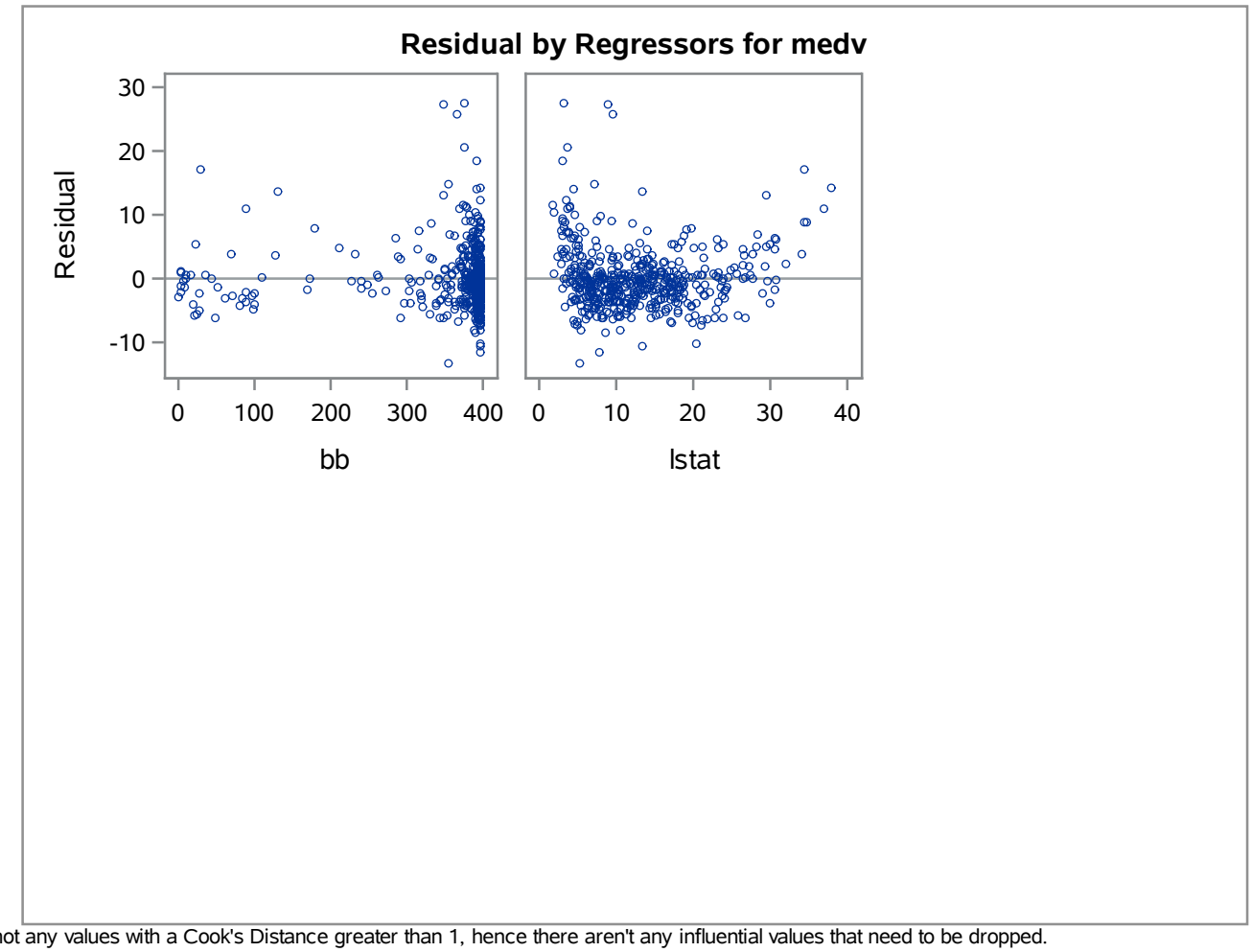


The REG Procedure
Model: MODEL1
Dependent Variable: medv



STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 6

The REG Procedure
Model: MODEL1
Dependent Variable: medv



There are not any values with a Cook's Distance greater than 1, hence there aren't any influential values that need to be dropped.

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 6

The REG Procedure
Model: MODEL1
Dependent Variable: medv

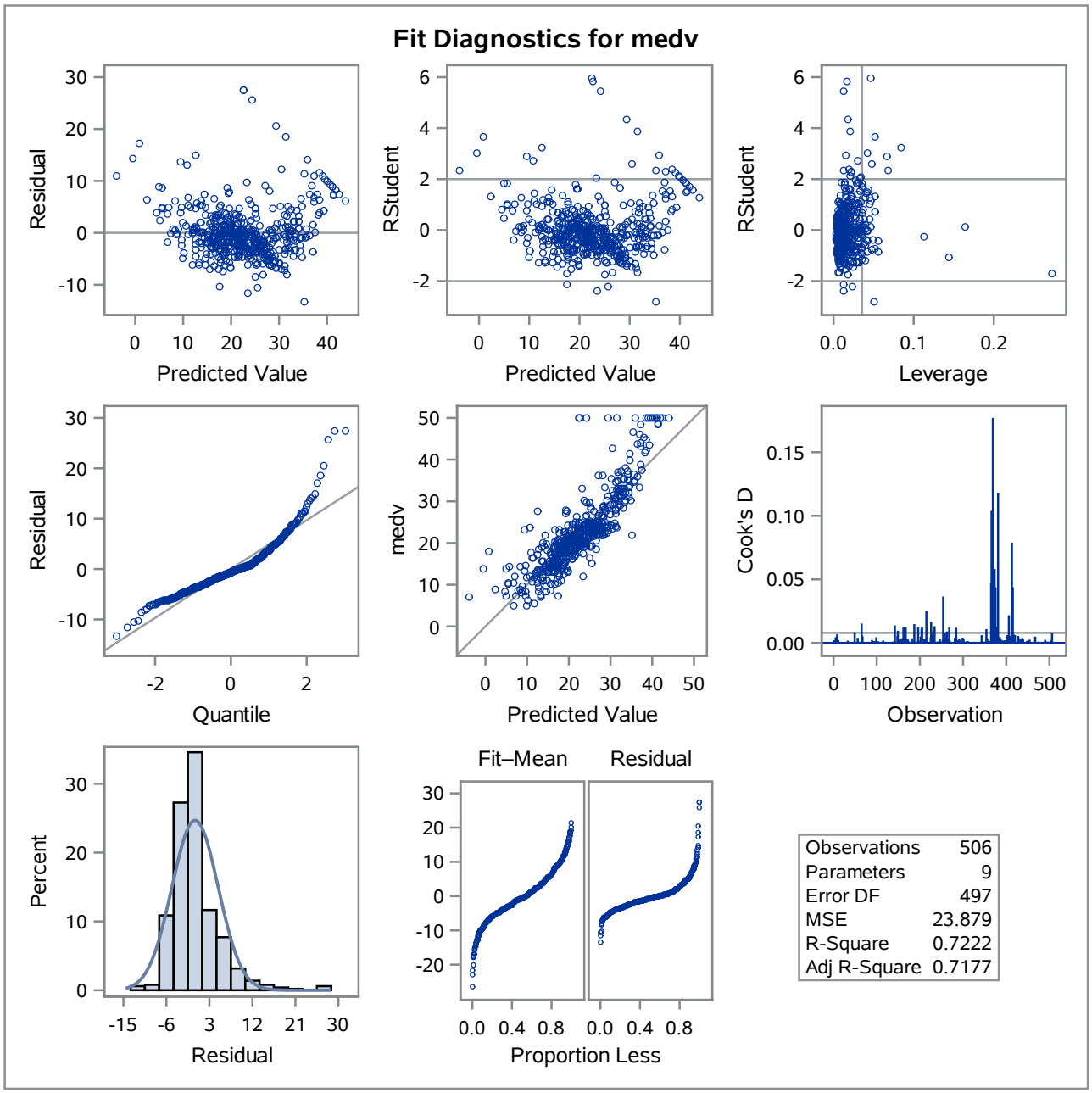
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	30848	3856.04614	161.48	<.0001
Error	497	11868	23.87913		
Corrected Total	505	42716			

Root MSE	4.88663	R-Square	0.7222
Dependent Mean	22.53281	Adj R-Sq	0.7177
Coeff Var	21.68672		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	29.54971	4.92700	6.00	<.0001
crim	1	-0.06609	0.03068	-2.15	0.0317
zn	1	0.04127	0.01357	3.04	0.0025
nox	1	-15.21364	3.25900	-4.67	<.0001
rm	1	4.21741	0.41178	10.24	<.0001
dis	1	-1.46380	0.19048	-7.68	<.0001
ptratio	1	-0.87583	0.11816	-7.41	<.0001
bb	1	0.00878	0.00271	3.24	0.0013
lstat	1	-0.53163	0.04885	-10.88	<.0001

STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 6

The REG Procedure
Model: MODEL1
Dependent Variable: medv



STAT 448 Midterm 1
Anshuman Saxena
03/28/2017
Excercise 6

The UNIVARIATE Procedure
Variable: resid (Residual)

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.886384	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.136664	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	2.154266	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	12.02091	**Pr > A-Sq**	<0.0050

The model is significant with a p-value less than 0.0001. The variation explained by the model is 72.22%. All the parameters are significant. The normality tests have a p-value less than 0.05 suggesting that the residuals do not satisfy the normality assumption.