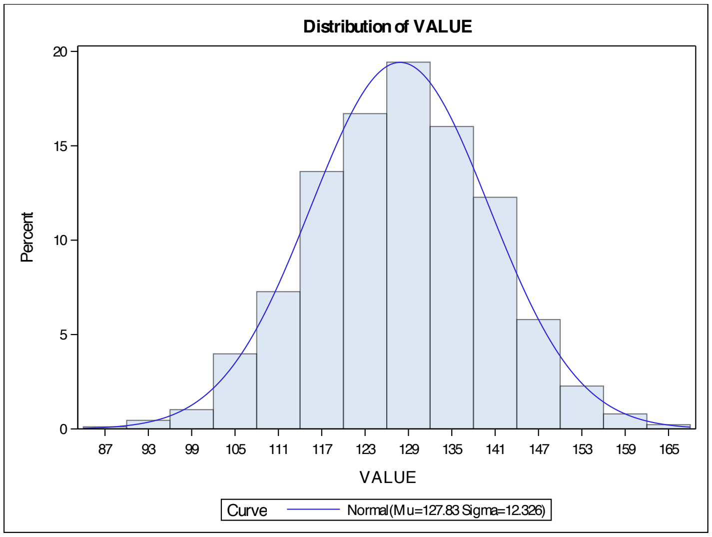
Exercise 1

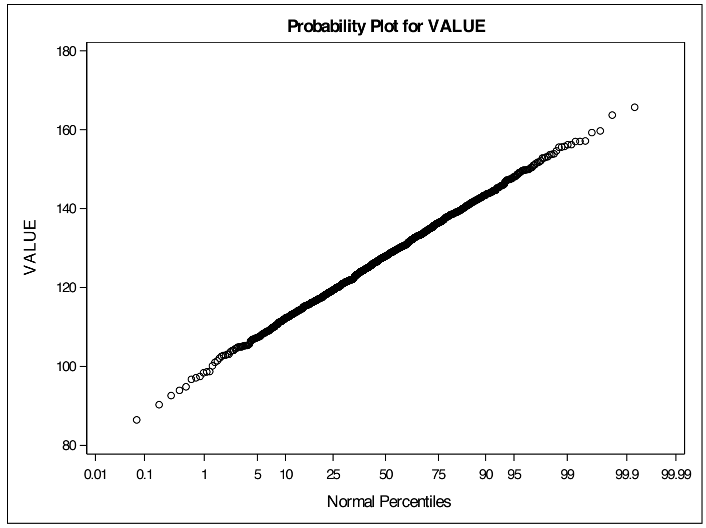
The UNIVARIATE Procedure
Variable: VALUE

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

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





1. The mean and median of systolic blood pressure is 127.83 and 128.03, with standard deviation 12.33. Since mean is less than median, the distribution is likely to have a long tail to the left. The negative skewness -0.05 supports it. Range is summarized as 79.23.

Systolic blood pressure seems to follow normal distribution. We can clearly the histogram is symmetric. The quantitative test results agree. Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling test (H0: normal distribution fits data) show p-value greater than 0.05, thus we cannot reject null hypothesis. We should assume normality for systolic blood pressure.

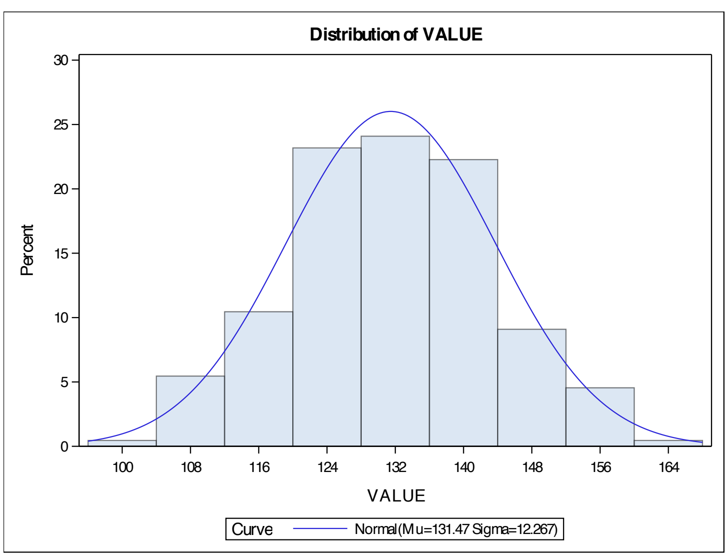
The UNIVARIATE Procedure
Variable: VALUE

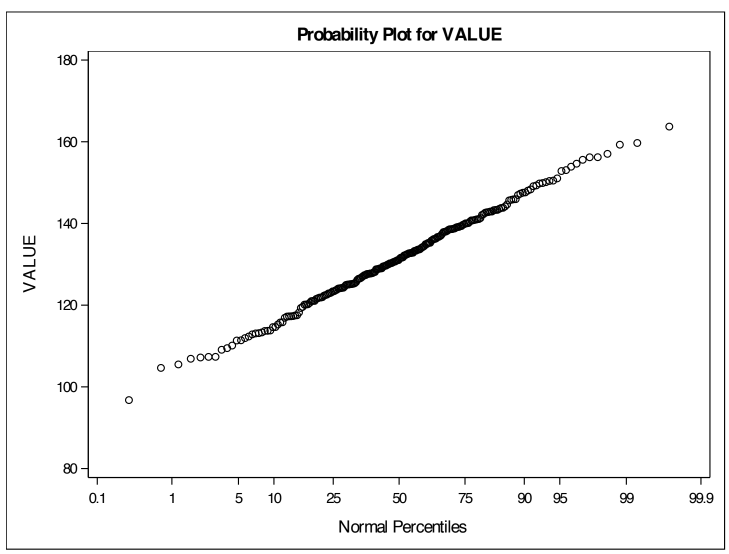
RTRTN=1

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

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

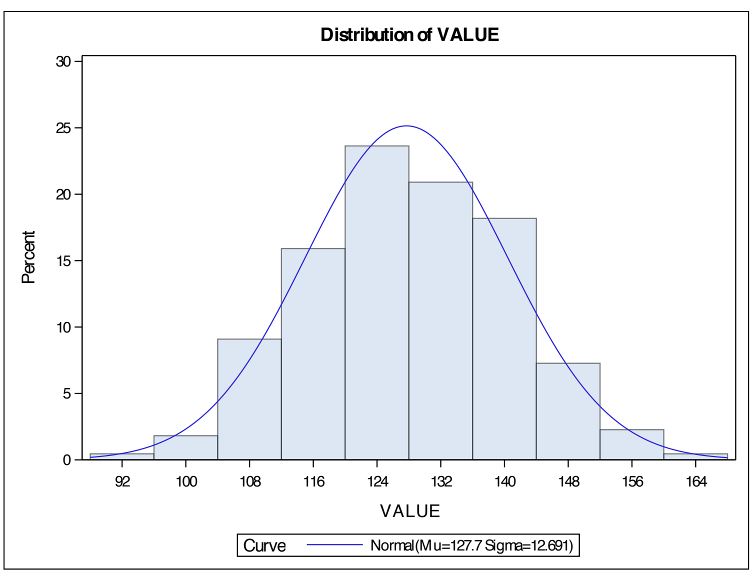


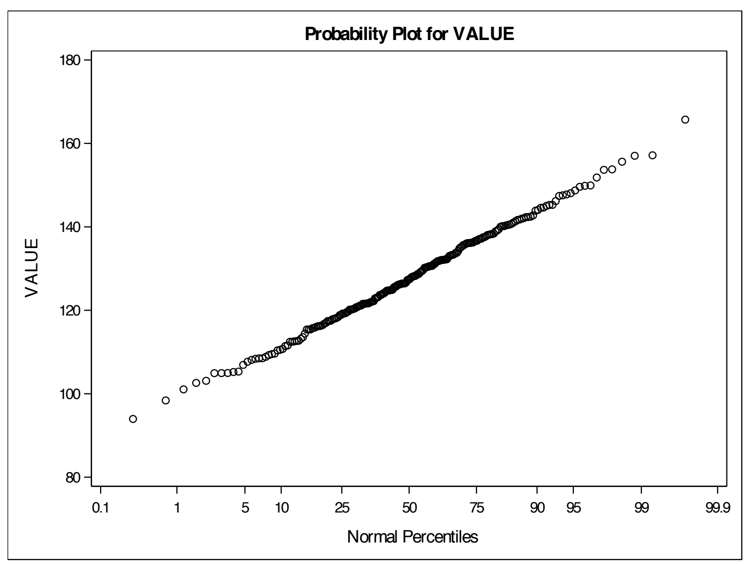


Moments			
N	220	**Sum Weights**	220
Mean	127.699136	**Sum Observations**	28093.81
Std Deviation	12.6911749	**Variance**	161.065919
Skewness	0.07224663	**Kurtosis**	-0.2093769
Uncorrected SS	3622828.71	**Corrected SS**	35273.4363
Coeff Variation	9.9383404	**Std Error Mean**	0.85563883

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

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





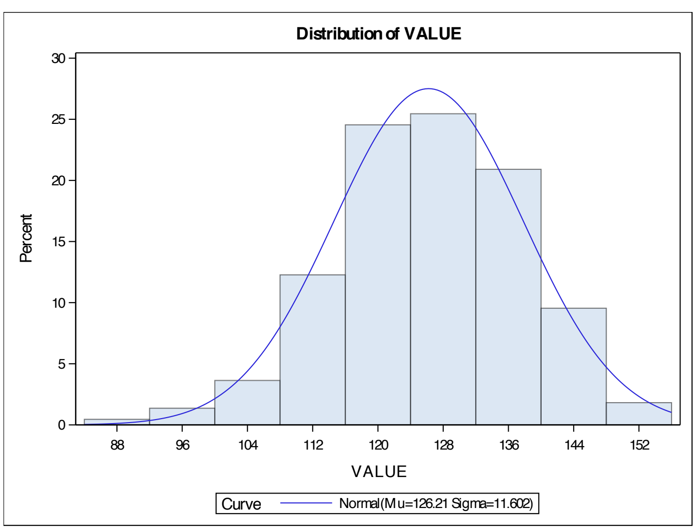
The UNIVARIATE Procedure
Variable: VALUE

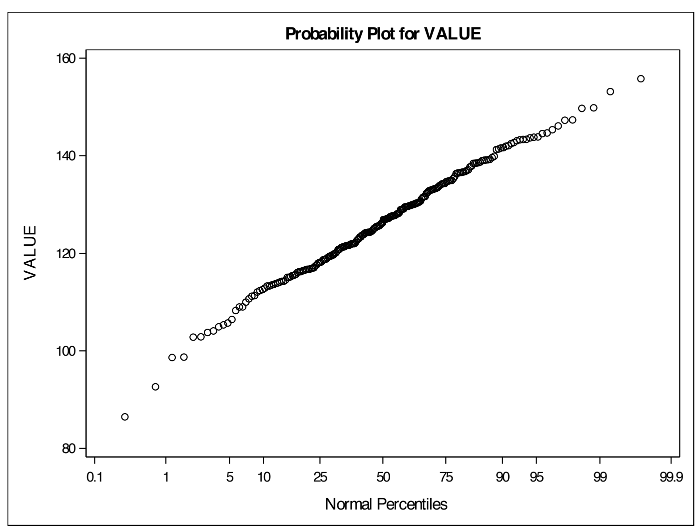
RTRTN=3

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

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





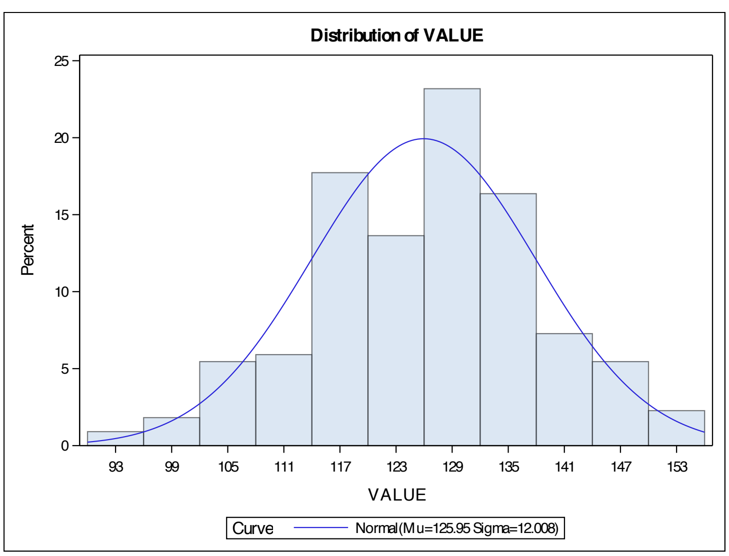
The UNIVARIATE Procedure
Variable: VALUE

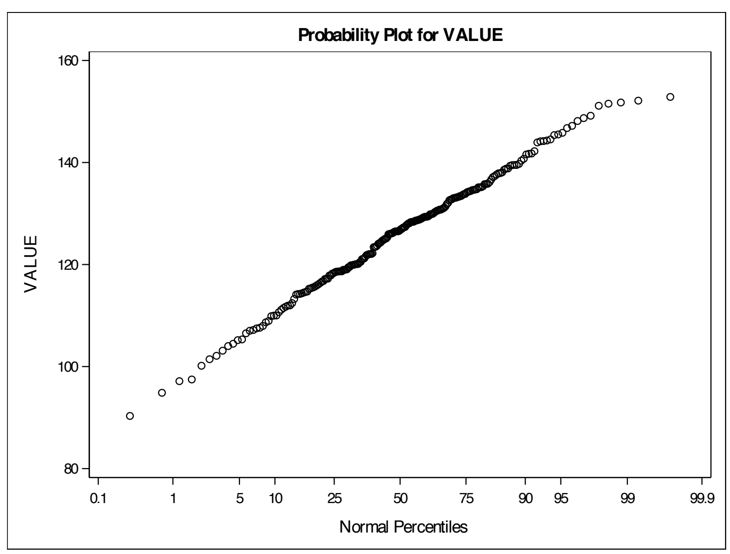
RTRTN=4

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

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





Now we repeat the analysis by treatment groups (RTRTN). Firstly, mean values for 1, 2, 3 and 4 are 131.47, 127.70, 126.21 and 125.95, and the median values are 131.21, 127.43, 126.55 and 126.79, respectively. 1 is likely to have larger systolic blood pressure compared to the other three groups, 2, 3, 4. Among 2, 3, 4, 2 has larger weight. The standard deviation and range for 1, 2, 3, and 4 are (12.27, 66.95), (12.69, 71.74), (11.60, 69.32) and (12.01, 62.53), respectively. It can be seen that systolic blood pressure of 2 has larger standard deviation and range.
For the skewness, systolic blood pressure of 1, 2, 3 and 4 have skewness measure as -0.001, 0.072, -0.229 and -0.183. Generally, positive measure indicates right long tail, and right-skewness.

For 1, 2, 3 and 4, all test p-values are greater than 0.05, so conclude that we do not have evidence to reject normality assumption. It is clear that histograms of 1 and 2 are symmetric. Although histograms of 3 and 4 do not have perfect symmetric shapes, they are not very bad. The observations in the QQ-plots almost follow the diagonal line. In conclusion, the distributions of systolic blood pressure for all four treatment groups are normal.

Exercise 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

2. In Exercise 1 we found that systolic blood pressure for reference (RTRTN=1) and ABC123 80mg (RTRTN=4) be assumed as normal. Since we have normality for two groups, we perform the T-test to compare the systolic blood pressure of two groups. The "Equality of Variances" test reveals insufficient evidence of unequal variances (the Folded F statistic F'=1.04, with p=0.7518.), so we use Satterthwaite adjustment. One-sided two-sample T-test gives p-value less than 0.05, and we can conclude that ABC123 80mg significantly reduces the blood pressure compared to the reference drug.

Exercise 3

The FREQ Procedure

Table of RTRTN by responder			
RTRTN	**responder**		
Frequency Expected	**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

Now we perform a hypothesis test of whether there is a significant relationship between treatment groups (RTRTN) and responder categories (responder). We find no cell presenting expected less than 5. Accordingly, we can use asymptotic tests. Both Chi-Square and Likelihood Ratio tests give p-values less than 0.05, thus we reject the null hypothesis (null: two variables are independent; alternative: there is some significant association) and conclude that there exists significant association between peanut allergy and early assumption. Note that we cannot apply Mantel-Haenszel test because treatment group is not ordered variable. Since the sample size is large, we cannot get the result of Fisher's exact test for the exact result.

The information about the magnitude of the relationship between two variables can be obtained from Phi Coefficient, Contingency Coefficient, and Cramer's V. The Phi Coefficient and Cramer's V give the same value, 0.1375, and Contingency Coefficient gives the similar magnitude 0.1362. We conclude that there's a mild association between treatment groups and responder categories. For larger dimensional tables, Cramer's V is standardized and its upper bound is 1, and therefore Cramer's V is preferred for larger dimensional tables.

The FREQ Procedure

Table of RTRTN by responder			
RTRTN	**responder**		
Frequency Expected	**0**	**1**	**Total**
1	184 167	36 53	220
4	150 167	70 53	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

We can test if two categorical variables are significantly associated via various types of asymptotic tests. Note that asymptotic tests can be performed, because all cells have expected values greater than 5. Chi-Square and Likelihood Ration Chi-Square tests give p-value less than 0.05, thus we reject the null hypothesis (H0: two are independent), and conclude that there exists a significant association between treatment groups and responder categories. Furthermore, we cannot apply Mantel-Haenszel test because treatment group is not ordinal variable. We can always refer to Fisher's exact test for the exact result. The Fisher's test also gives p-value less than 0.05 (2-sided p-value=0.0002), thus we can make the same conclusion.

The information about the magnitude of the relationship between two variables can be obtained from Phi Coefficient, Contingency Coefficient, and Cramer's V. The Phi Coefficient and Cramer's V give the same value, 0.1807, and Contingency Coefficient gives the similar magnitude 0.1778. For 2x2 tables, we focus more on the Phi and Cramer's V coefficients because they are bounded between -1 and 1, and therefore are analogues of correlation coefficients. We conclude that there's a mild association between peanut allergy and early consumption. We conclude that there's a mild association between treatment groups and responder categories.

In the following table Row1 and Row2 correspond to reference and ABC123 80mg groups, respectively. The proportional difference in the responder category is -0.1545. We can test whether the difference is significant by using the confidence interval. The table provides 95% confidence limit as (-0.2331, -0.0759), and zero is not in the interval. Thus, we can conclude there are more responders in the ABC123 80mg group compared to the reference group.

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)

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

Exercise 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)

Effects:	Intercept SEX RTRTN SITE

We first use stepwise selection to investigate main effects. By using GLMSELECT, only SEX, RTRTN, and SITE are selected in the model.

The GLM Procedure

Dependent Variable: VALUE

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

We then investigate whether the interaction terms are significant or not. From Type III SS, we could conclude interaction terms are not significant. Meanwhile from the plot we could also conclude there is no significant interaction. Then we focus on model with only SEX, RTRTN, and SITE as three main effects.

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			

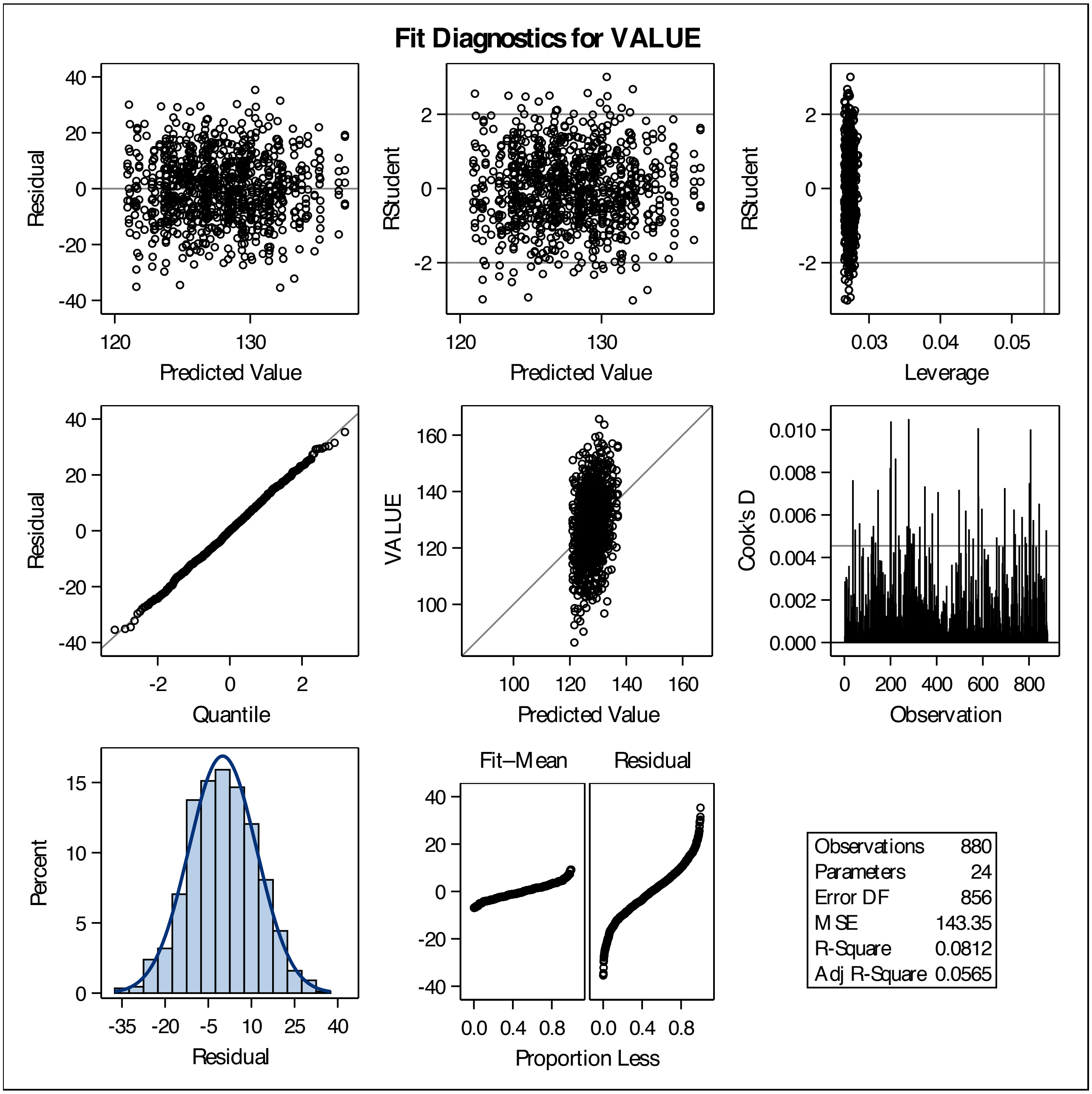
R-Square	Coeff Var	Root MSE	VALUE Mean
0.081204	9.366045	11.97295	127.8335

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

The GLM Procedure

Dependent Variable: VALUE



Then we focus on model with only SEX, RTRTN, and SITE as three main effects. The ANOVA model is significant with p-value less than .0001.The proportion of variation in blood pressure value explained by the model is 8.12%, which is R-square in the output.

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 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

Now we investigate each main effect to see where we can find the significant differences. Firstly, we rank the treatment groups based on efficacy at reducing blood pressure. For blood pressure, reference group has the largest value, the ABC123 20mg group has the second largest value, the ABC123 40mg group has the third largest value, and the ABC123 80mg group has the smallest value. Second, differences of LS means, 95% CI and p-values are highlighted. Differences between group 1 and 2, between group 1 and 3, and between group 1 and 4 are significant. Differences between group 2 and 3, between group 2 and 4, and between group 3 and 4 are insignificant. At last we test for the normality. The normality tests for residuals show p-value greater than 0.05, thus we can say that residuals are normally distributed. It implies that model assumption is valid.

Exercise 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.

Effects:	Intercept SEX RTRTN SITE BASE

Base score, rtrtn, sex and site are retained in the final ANCOVA model after stepwise selection procedure with significance level 0.05.

The GLMSELECT Procedure
Selected Model

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
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

One-unit increase in the baseline score increases the final blood pressure value by 0.850936.

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

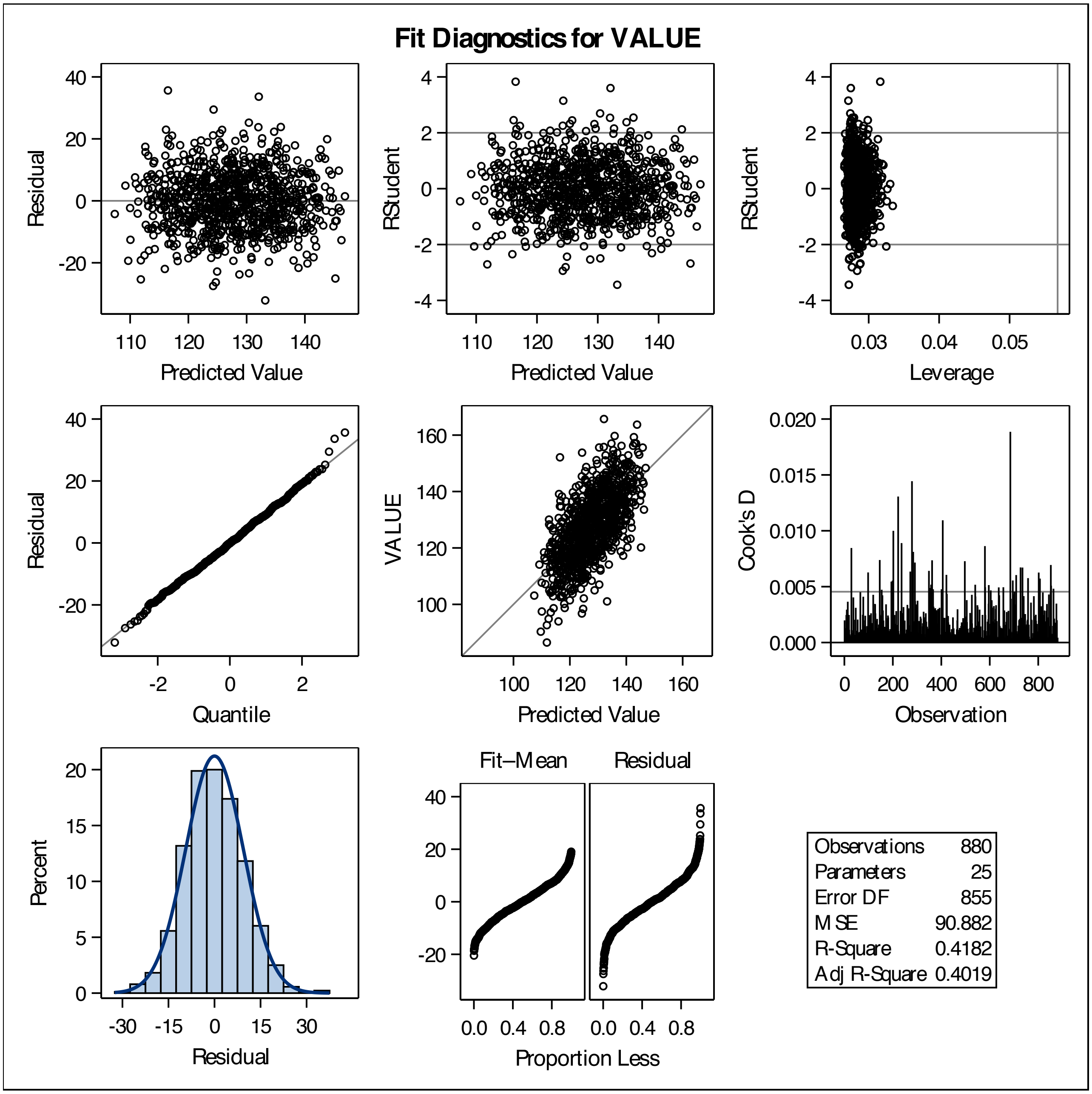
Source	DF	Type III SS	Mean Square	F Value	Pr > F
SEX	1	1847.71487	1847.71487	20.33	<.0001
RTRTN	3	5122.02942	1707.34314	18.79	<.0001
SITE	19	4292.96137	225.94534	2.49	0.0004
BASE	1	45004.92938	45004.92938	495.20	<.0001

The predictors in the ANOVA model from Exercise 4 does not contain base score as in the ANCOVA model. The proportion of variation in blood pressure value explained by the ANCOVA model is 41.82%, which is much higher than the R-square in the ANOVA model from Exercise 4 (R-square=8.14%).

From the diagnostic plots below, we see all have cook's D less than 1. No observation is highly influential. The normality tests for residuals show p-value greater than 0.05, thus we can say that residuals are normally distributed. It implies that model assumption is valid.

The GLM Procedure

Dependent Variable: VALUE



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

Exercise 6

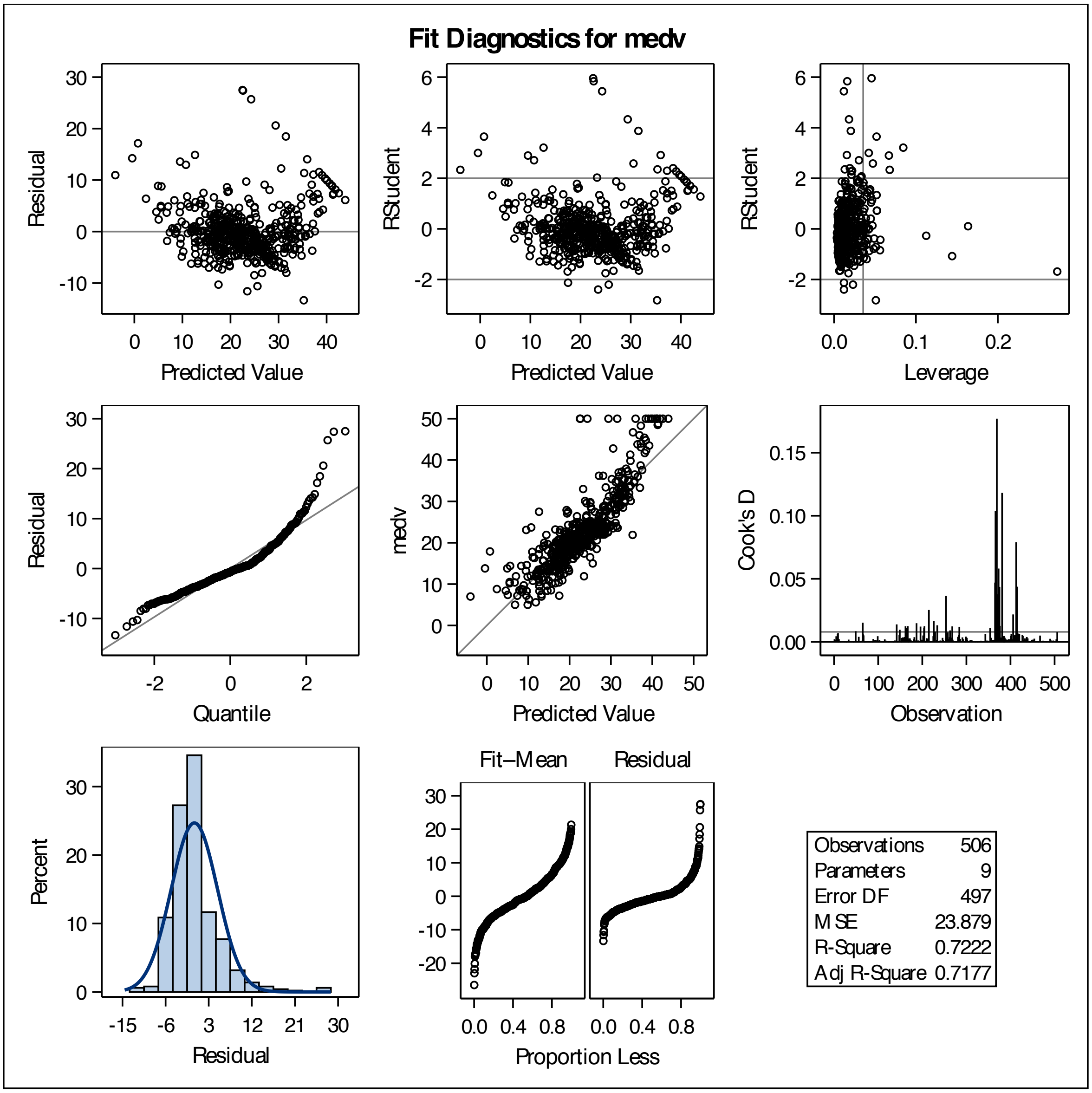
The REG Procedure
Model: MODEL1
Dependent Variable: medv

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

6. Based on stepwise selection, indus, age and tax are removed, and keep lstat, rm, ptratio, dis, nox, bb, zn and crim. We need to examine VIF to check multicollinearity problems. All VIFs are less than 10, thus we would not remove variables and keep all significant predictors.

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

We need to examine VIF to check multicollinearity problems. All VIFs are less than 10, thus we would not remove variables and keep all significant predictors.



Now we need to identify any influential observations. From the diagnostic plots above, we see all have cook's D less than 1. No observation is highly influential.

The REG Procedure
Model: MODEL1
Dependent Variable: medv

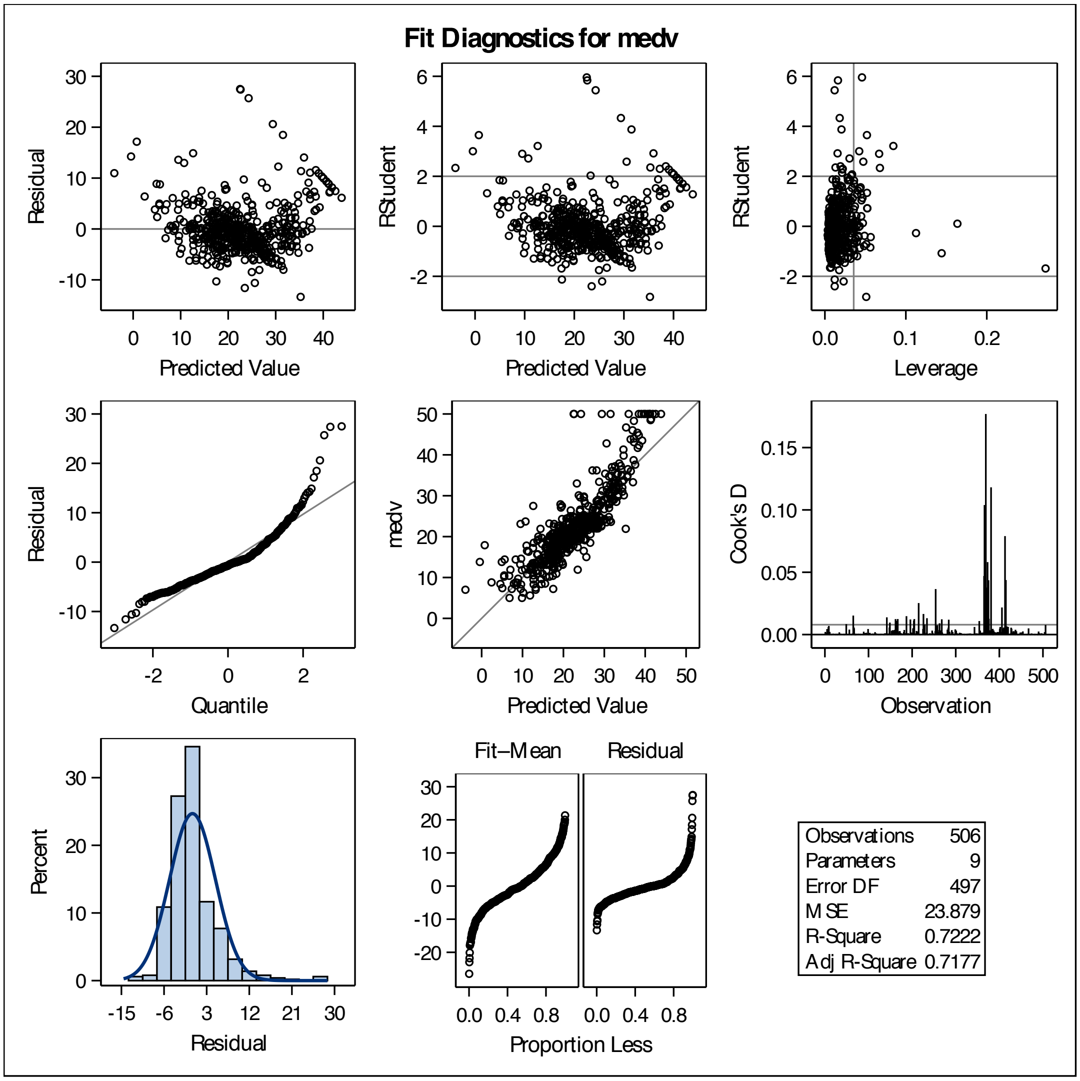
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	8	30848	3856.04614	161.48	<.0001
Error	497	11868	23.87913		
Corrected Total	505	42716			

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

The final model (medv~crim+zn+nox+rm+dis+ptratio+bb+lstat) is statistically significant model with p-value less than 0.0001. All predictors are significant in the model. The variation explained by the model is 72.22%.

The parameter estimates are -0.06609, 0.04127, -15.21364, 4.21741, -1.46380, -0.87583, 0.00878 and -0.53163 for crim, zn, nox, rm, dis, ptratio, bb and lstat, respectively. It means that medv is expected to decrease 0.06609, 15.21364, 1.46380, 0.87583 and 0.53163 for a one-unit increase in crim, nox, dis, ptratio and lstat. Medv is expected to increase 0.04127, 4.21741 and 0.00878 for a one-unit increase inzn, rm and bb.



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

From diagnostic plots, we see residuals are randomly distributed around 0. The normal QQ plot illustrate that residuals do not follow the diagonal line. The histogram supports it with right skewed shape. The normality can be investigated via normality tests. All p-values are less than 0.05, thus we can conclude that residuals are not normally distributed and there is an issue on assuming normality of the residuals.