Exercise 1a):

		Cholesterol		
		Mean	**Std**	**N**
Blood Pressure Status	**Weight Status**			
High	**Normal**	224.37	45.05	206
	Overweight	232.99	44.93	854
	Underweight	209.85	56.30	13
Normal	**Normal**	213.68	36.79	486
	Overweight	225.14	43.30	905
	Underweight	206.17	38.09	63
Optimal	**Normal**	207.21	40.56	292
	Overweight	213.36	37.94	278
	Underweight	195.86	32.50	35

The mean value of cholesterol for high blood pressure status appears to be higher than normal and optimal blood pressure status. We also can see that overweight weight status have higher mean values of cholesterol in all bp_status. While underweight weight status have lower mean values of cholesterol in general. The standard deviation ranges between 32-56.

Exercise 1b):

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
BP_Status	3	High Normal Optimal
Weight_Status	3	Normal Overweight Underweight

Number of Observations Read	3132
Number of Observations Used	3132

The GLM Procedure

Dependent Variable: Cholesterol

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	273230.983	34153.873	19.29	<.0001
Error	3123	5530707.496	1770.960		
Corrected Total	3131	5803938.479			

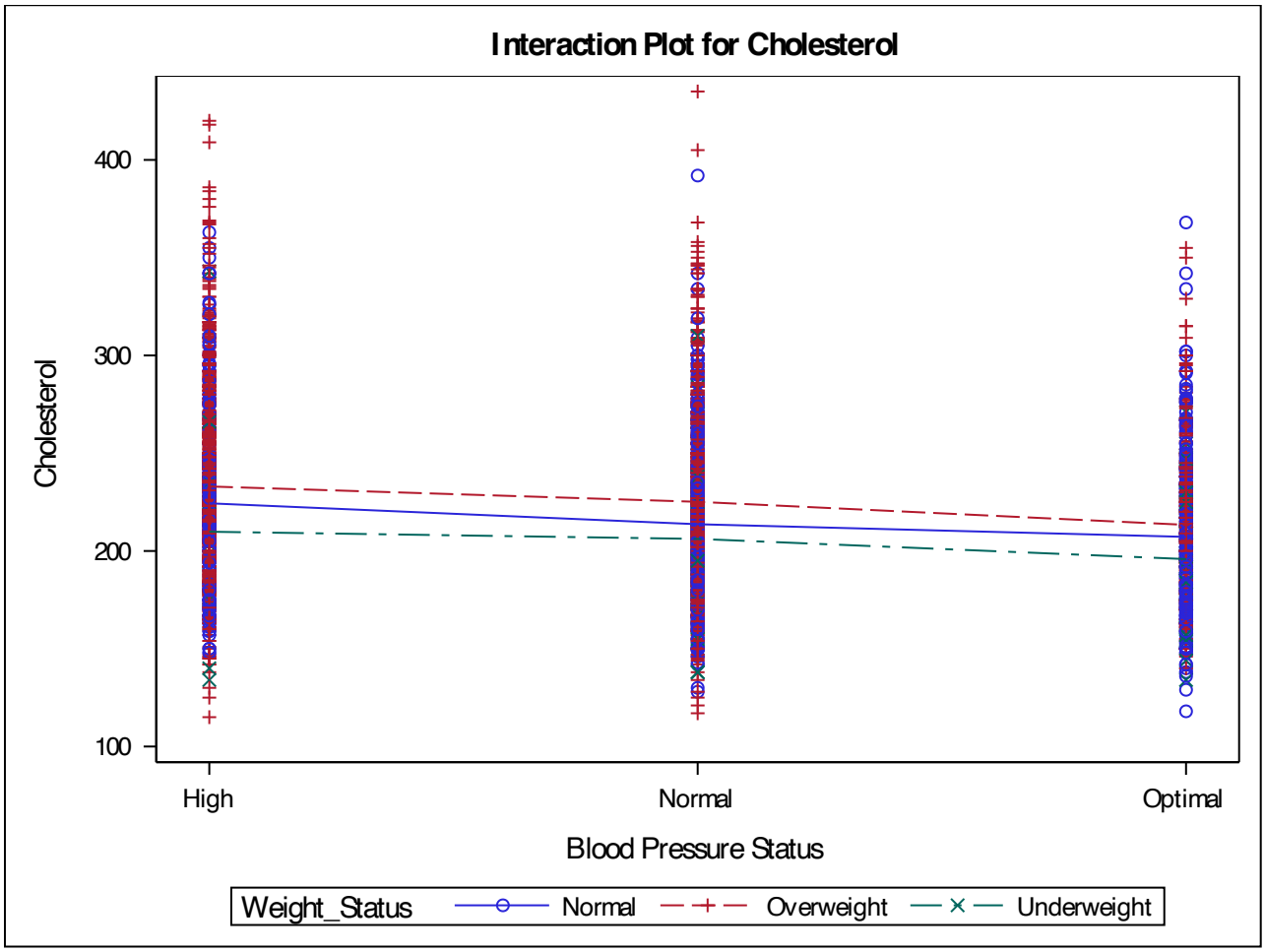
R-Square	Coeff Var	Root MSE	Cholesterol Mean
0.047077	18.95945	42.08277	221.9620

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BP_Status	2	187763.8412	93881.9206	53.01	<.0001
Weight_Status	2	82257.2664	41128.6332	23.22	<.0001
BP_Status*Weight_Sta	4	3209.8750	802.4687	0.45	0.7702

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BP_Status	2	25801.62453	12900.81226	7.28	0.0007
Weight_Status	2	62530.55513	31265.27757	17.65	<.0001
BP_Status*Weight_Sta	4	3209.87499	802.46875	0.45	0.7702

The GLM Procedure

Dependent Variable: Cholesterol



We can see from the first part that the data is unbalanced. Hence ,we will be using proc glm procedure. Based on the result of Type I SS, we can see that we might want to keep bp_status and weight_status, since we have p-value 0.7702 for the interaction term (bp_status*weight_status) (not significant). It is also can be observed from Type III table.

Exercise 1c):

The GLM Procedure

Dependent Variable: Cholesterol

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	270021.108	67505.277	38.14	<.0001
Error	3127	5533917.371	1769.721		
Corrected Total	3131	5803938.479			

R-Square	Coeff Var	Root MSE	Cholesterol Mean
0.046524	18.95282	42.06805	221.9620

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BP_Status	2	187763.8412	93881.9206	53.05	<.0001
Weight_Status	2	82257.2664	41128.6332	23.24	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BP_Status	2	120974.9920	60487.4960	34.18	<.0001
Weight_Status	2	82257.2664	41128.6332	23.24	<.0001

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

BP_Status	Cholesterol LSMEAN	LSMEAN Number
High	223.550434	1
Normal	214.932851	2
Optimal	205.510170	3

Least Squares Means for Effect BP_Status				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	**2**	8.617583	4.590265	12.644902
1	**3**	18.040265	12.861518	23.219012
2	**3**	9.422681	4.609891	14.235471

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

Weight_Status	Cholesterol LSMEAN	LSMEAN Number
Normal	214.719643	1
Overweight	224.228978	2
Underweight	205.044835	3

Least Squares Means for Effect Weight_Status				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	**2**	-9.509335	-13.452955	-5.565715
1	**3**	9.674808	-0.208145	19.557761
2	**3**	19.184143	9.489859	28.878428

So, we chose the model with bp_status and weight_status. The F statistics of just over 38 has a p-value of <.0001 and therefore it is determined that the bp_status and weight_status main effects together describe significantly more variation than expected due to chance. The R-Square value indicates that the percentage of variation described is not terribly large at just over 4.6. Given that only a mean is in the model already, the model sum of squares would increase by 93881.9 if bp_status is added. Given that a mean and origin are already included in the model, adding weight_status to the model increases the model sum of squares by 41128.6. The F statistics for both of the sources are statistically significant. The associated confidence interval for the difference of means for effect of bp_status indicates that the difference is statistically significant as 0 is not in the interval. While for effect of weight_status it is not the case, only for the last interval it is significant. People with high blood pressure status were expected to have 223.6 cholesterol, with normal bp 214.9 and with optimal 205.5. People with normal weight were expected to have 214.7 cholesterol level, with overweight 224.2 and underweight 205.04.

Exercise 2a):

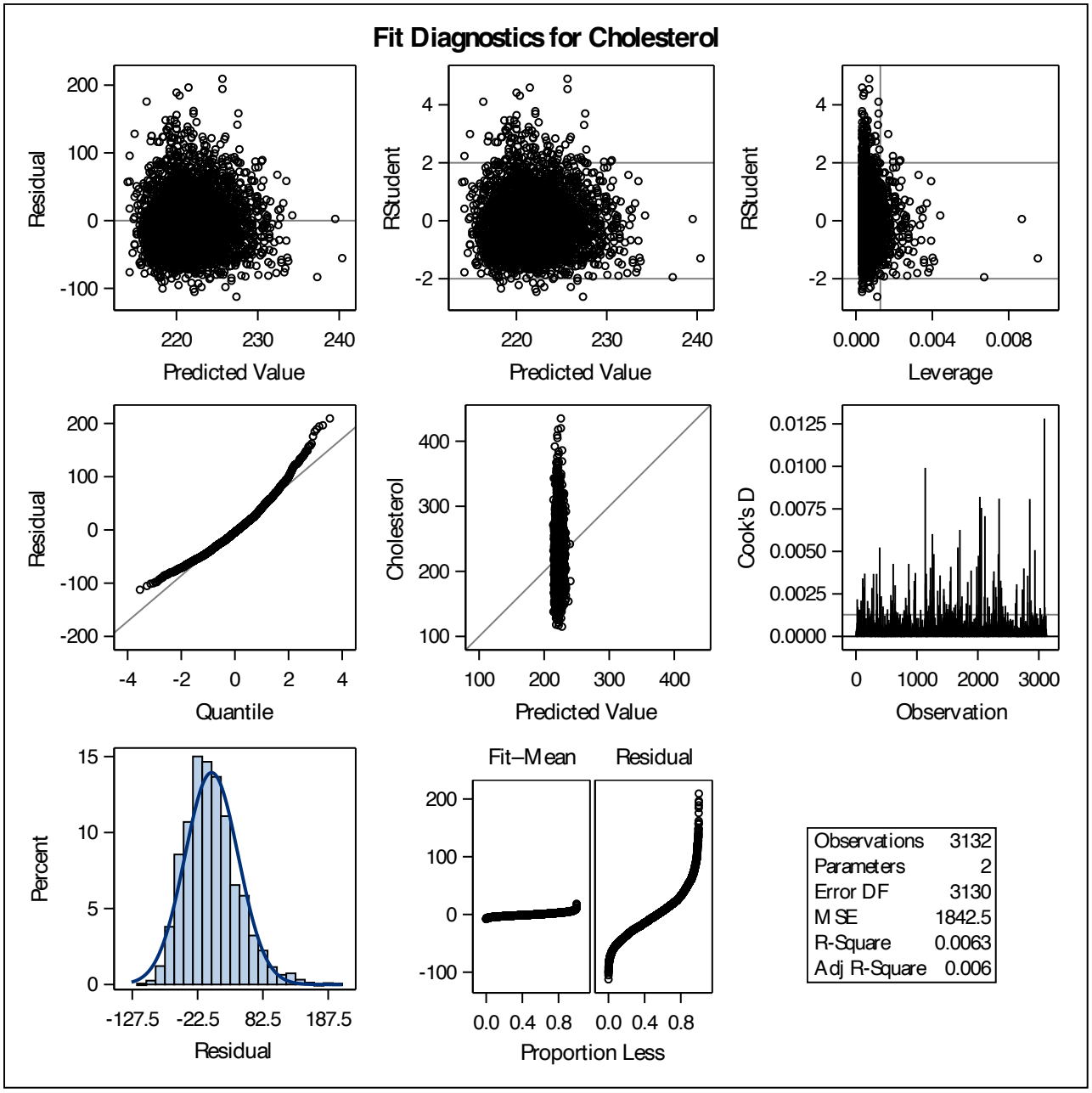
The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	36791	36791	19.97	<.0001
Error	3130	5767147	1842.53906		
Corrected Total	3131	5803938			

Root MSE	42.92481	**R-Square**	0.0063
Dependent Mean	221.96201	**Adj R-Sq**	0.0060
Coeff Var	19.33881		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	203.57605	4.18543	48.64	<.0001
Weight	1	0.12264	0.02745	4.47	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol



We can see that there are no observations with Cook's distance larger than 0.015. Hence, we are not removing any points from the dataset.

Exercise 2b):

The results for fitting a simple linear regression model of cholesterol as a function of weight follow (part a). The model is statistically significant, but it only describes 0.63% of the variation in cholesterol. The intercept β_0 is estimated to be about 203.57 and the slope β_1 is estimated to be just 0.12, and both are statistically significant as indicated by the p-values less than 0.05. The slope indicates an expected increase of about only 0.12 cholesterol level for each increase of weight. We can now interpret some of the diagnostic plots. Residuals look normal; however, we see that there are right end and left end points lie above the diagonal line in quantile plot. It also can be seen from histogram that it is slightly positively skewed. The first two residuals plot look fine, so no need to worry about constant variance. However, the plot (predicted value vs cholesterol) in the middle does not look good, so it does not seem to fit the data well, as it is not linear. I would not say that this model will be good: although diagnostics is good, but we have very small percentage of variation explained. So, weight could not be a good predictor for cholesterol.

Exercise 3a):

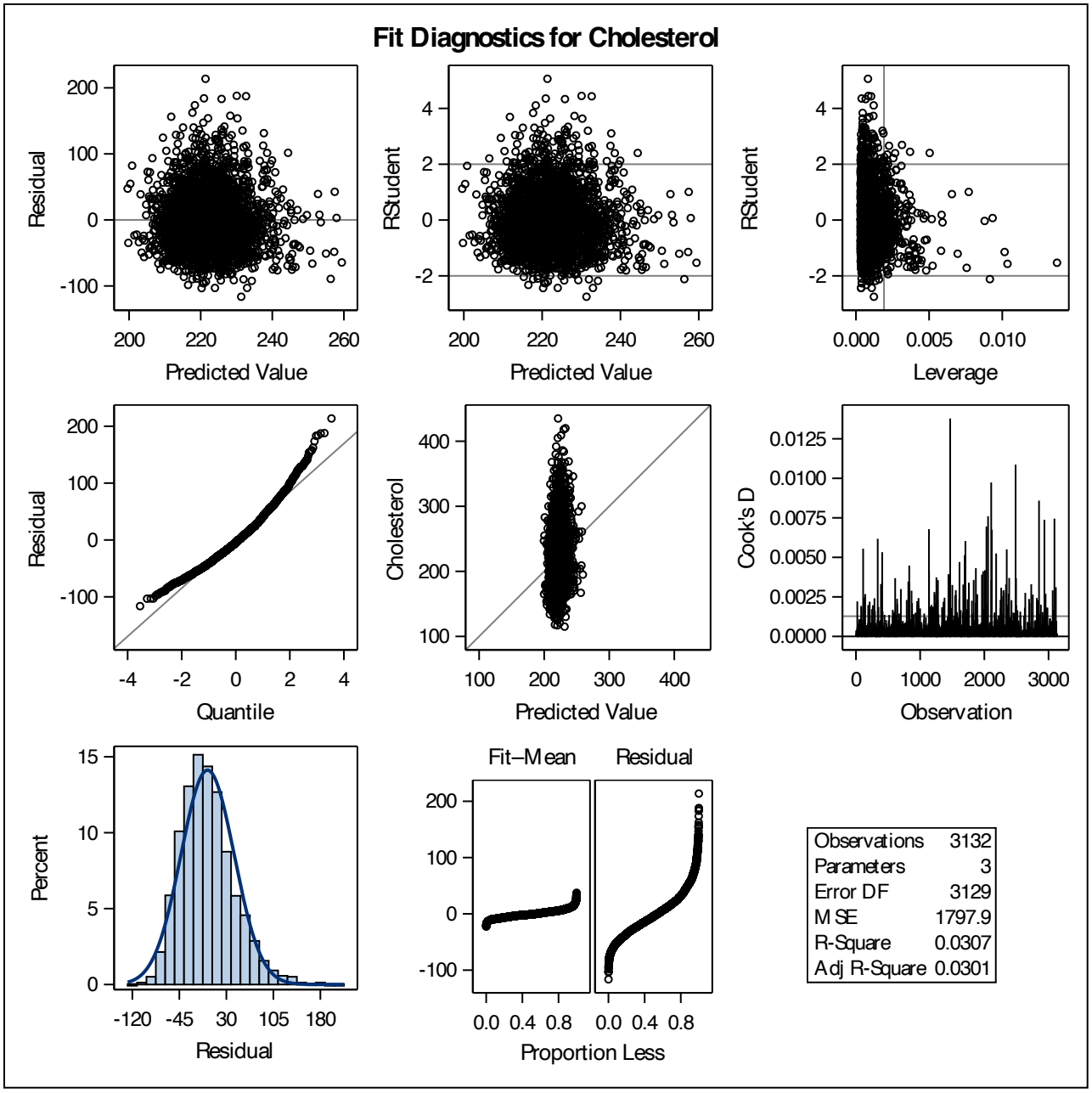
The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	178454	89227	49.63	<.0001
Error	3129	5625485	1797.85381		
Corrected Total	3131	5803938			

Root MSE	42.40111	**R-Square**	0.0307
Dependent Mean	221.96201	**Adj R-Sq**	0.0301
Coeff Var	19.10287		

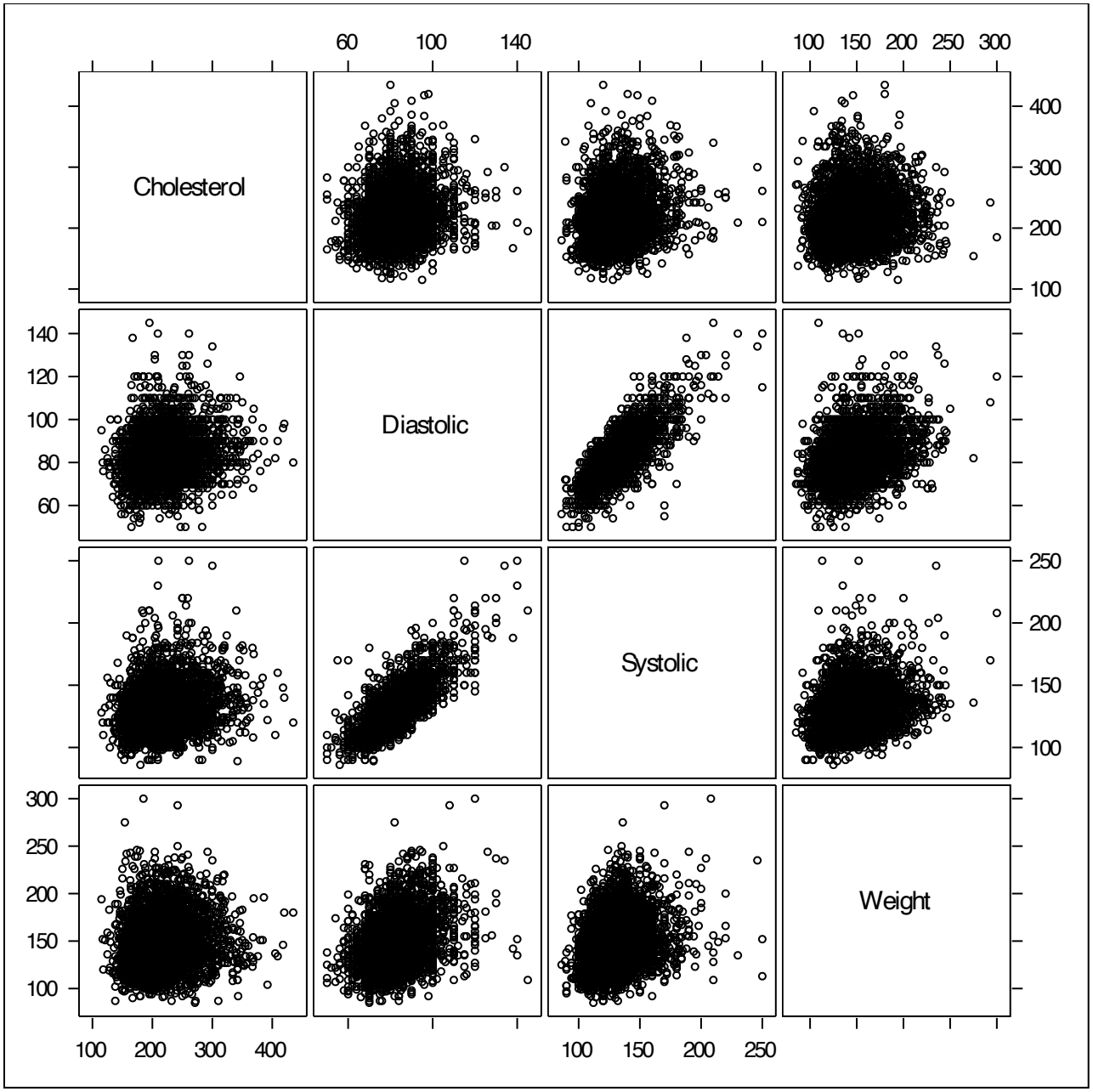
Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	164.01126	6.07941	26.98	<.0001
Diastolic	1	0.62852	0.07081	8.88	<.0001
Weight	1	0.03919	0.02869	1.37	0.1721

The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol



The results for fitting a simple linear regression model of cholesterol as a function of diastolic and weight follow. The model is statistically significant, and it describes 3% of the variation in cholesterol. It is slightly lower than the variation in Exercise 1(categorical variables). But it is better comparing to the previous model in Exercise 2. Intercept and diastolic are statistically significant, however weight is not statistically significant (by looking at p-values). We can now interpret some of the diagnostic plots. Residuals look normal; however, we see that there are right end and left end points lie above the diagonal line in quantile plot. It also can be seen from histogram that it is slightly positively skewed. There are no observations with Cook's distance larger than 0.015. Residual plots look fine.

Exercise 3b):


Cholesterol
Diastolic
Systolic
Weight
60
100
140
100
150
200
250
300
100
200
300
400
140
120
100
80
60
250
200
150
100
300
250
200
150
100
100
150
200
250

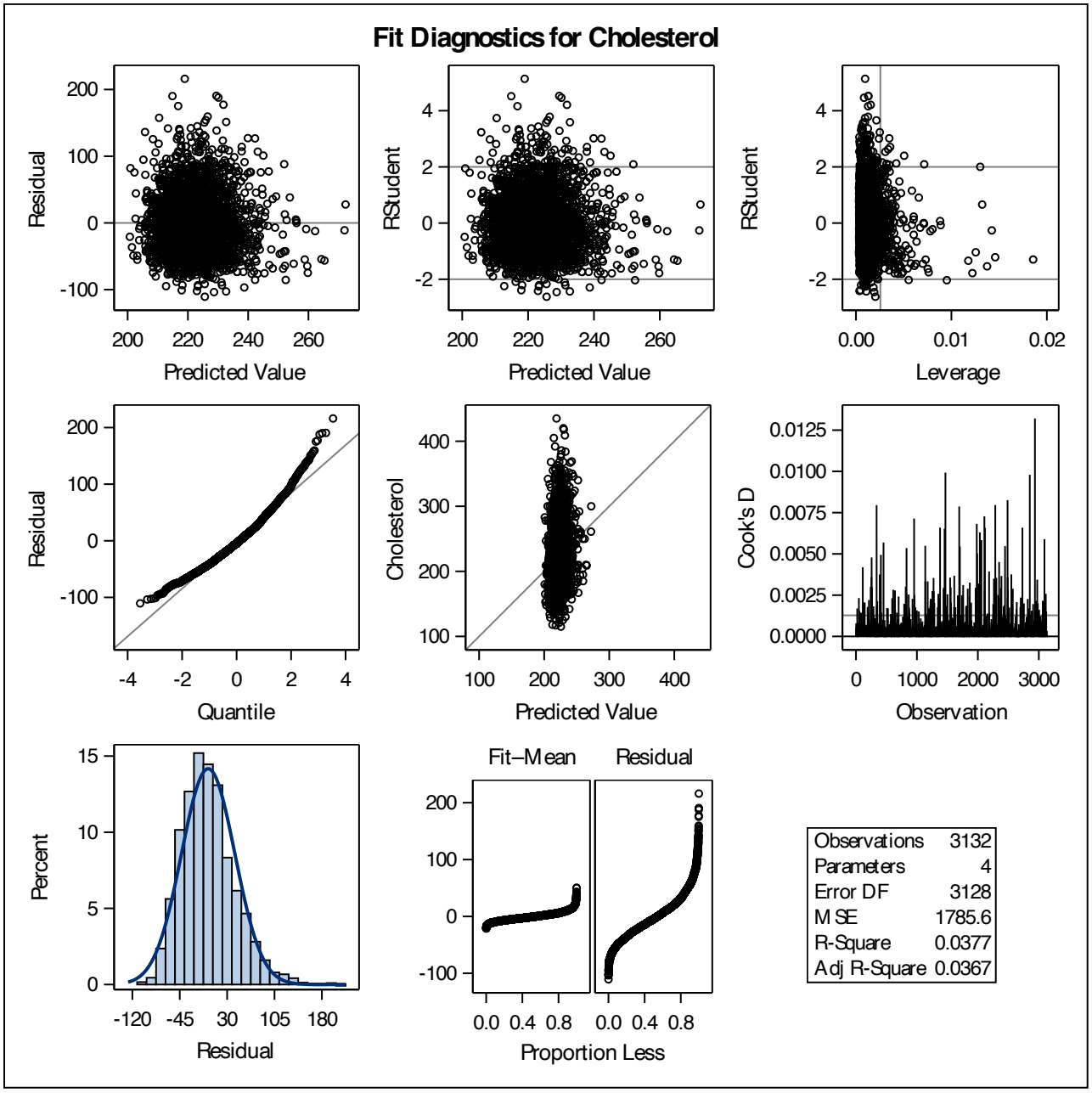
The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	218628	72876	40.81	<.0001
Error	3128	5585310	1785.58508		
Corrected Total	3131	5803938			

Root MSE	42.25618	**R-Square**	0.0377
Dependent Mean	221.96201	**Adj R-Sq**	0.0367
Coeff Var	19.03758		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	156.32618	6.27153	24.93	<.0001
Diastolic	1	0.24922	0.10665	2.34	0.0195
Systolic	1	0.30073	0.06340	4.74	<.0001
Weight	1	0.03671	0.02860	1.28	0.1994

The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol



Before fitting the model, we plotted pairwise scatter plot to see if there are any correlations between variables. It looks like there is a very strong positive linear trend for diastolic and systolic predictors. The results for fitting a simple linear regression model of cholesterol as a function of diastolic, systolic and weight follow. The model is statistically significant, and it describes 3.77% of the variation in cholesterol. It is slightly better comparing to the previous model. Intercept, diastolic and systolic are statistically significant, however weight is not statistically significant (by looking at p-values). We can now interpret some of the diagnostic plots. Residuals look normal; however, we see that there are right end and left end points lie above the diagonal line in quantile plot. It also can be seen from histogram that it is slightly positively skewed. There are no observations with Cook's distance larger than 0.015. Residual plots look fine. The intercept $\beta 0$ is estimated to be about 156.3, the coefficient for diastolic is estimated to be just 0.25, the coefficient for systolic is estimated to be 0.3 and for weight it is about 0.038. As we can see they are all positive. As the diastolic level increases by 1, about 0.25 cholesterol is expected. And as systolic level increases by 1, an increase of about 0.3 cholesterol is expected, as the weight increases by 1, an increase of about 0.038 cholesterol is expected.

Exercise 4a):

The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol

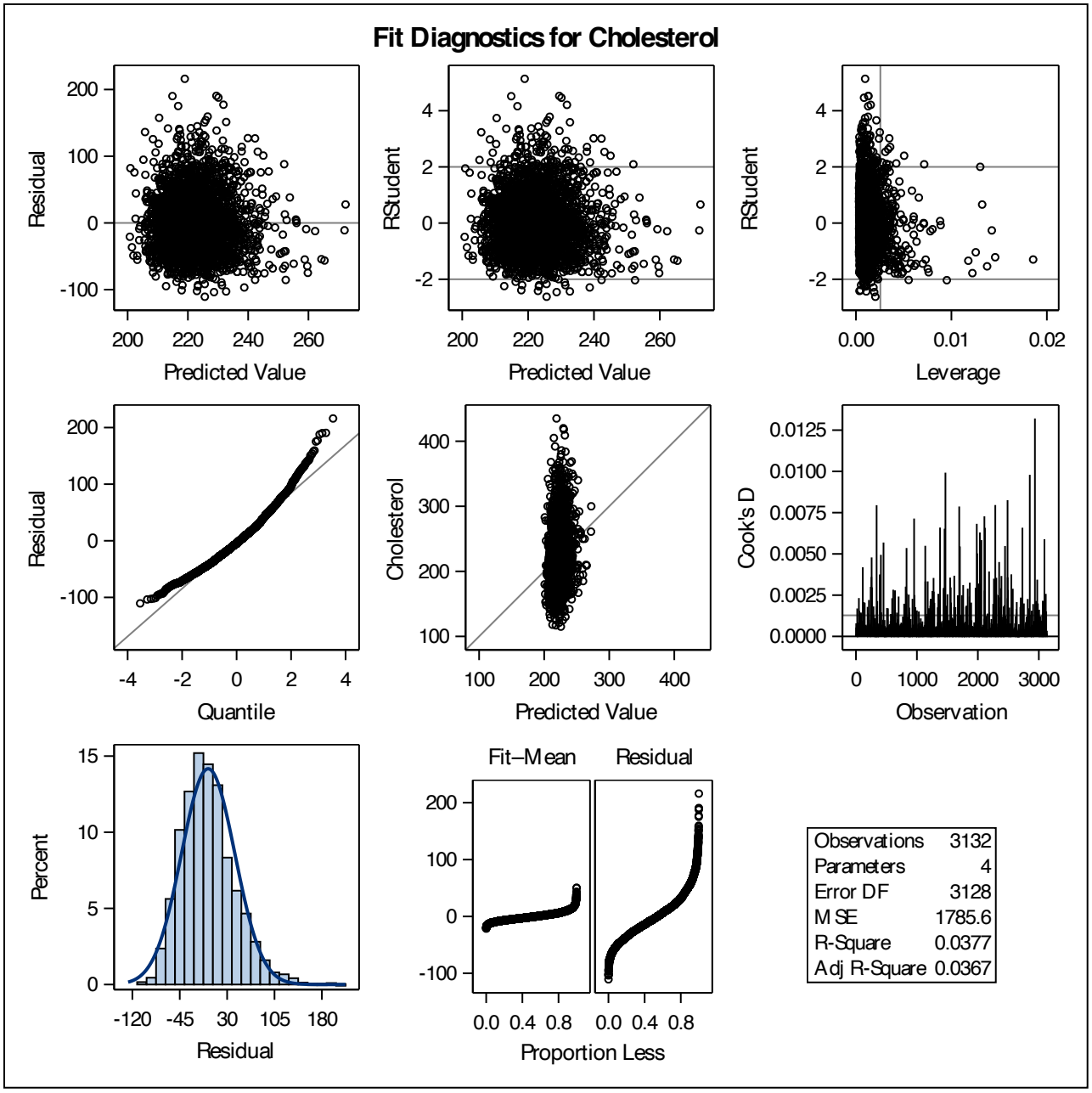
Number of Observations Read	3132
Number of Observations Used	3132

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	218628	72876	40.81	<.0001
Error	3128	5585310	1785.58508		
Corrected Total	3131	5803938			

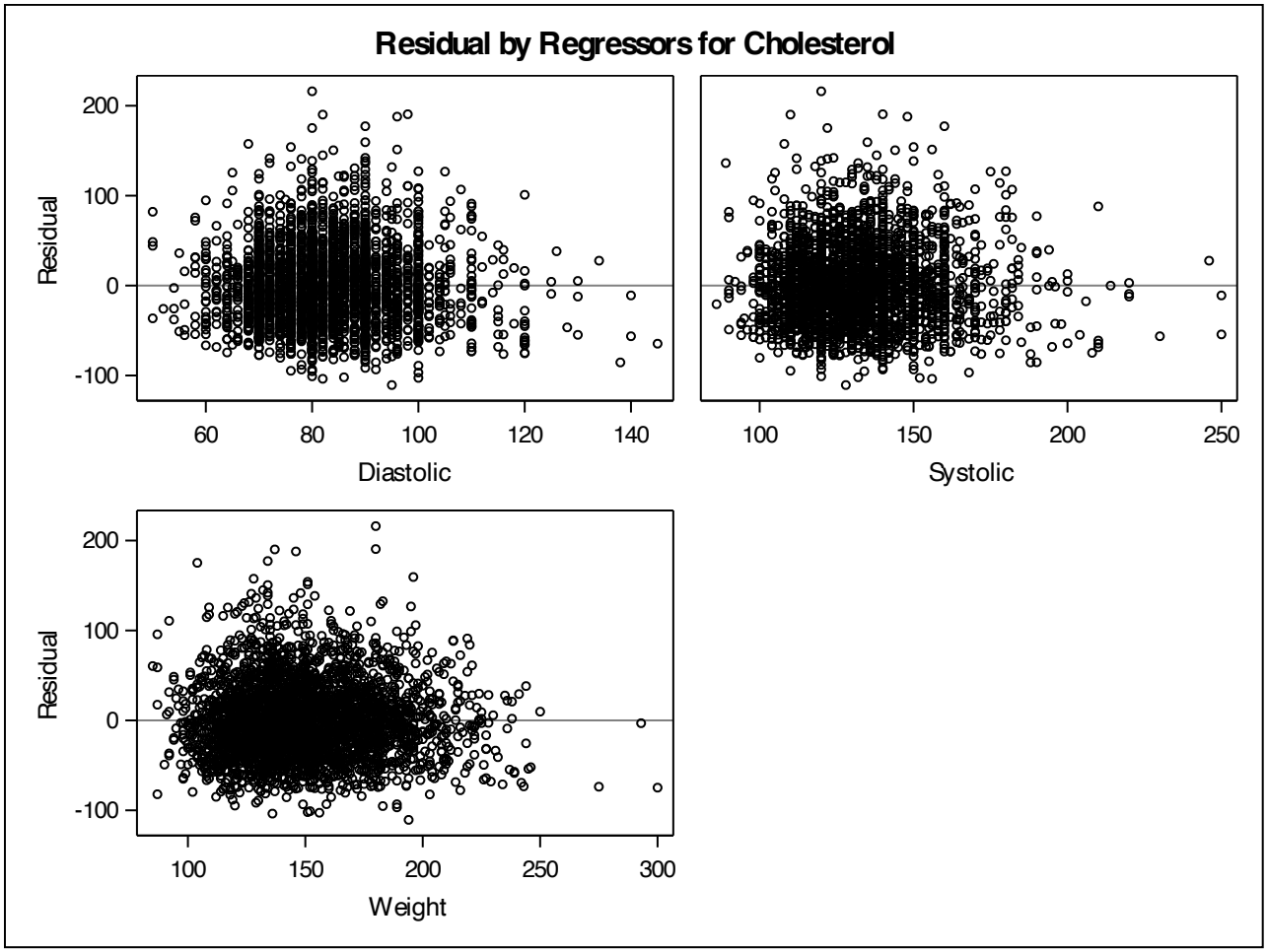
Root MSE	42.25618	**R-Square**	0.0377
Dependent Mean	221.96201	**Adj R-Sq**	0.0367
Coeff Var	19.03758		

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	156.32618	6.27153	24.93	<.0001	0
Diastolic	1	0.24922	0.10665	2.34	0.0195	2.55891
Systolic	1	0.30073	0.06340	4.74	<.0001	2.45421
Weight	1	0.03671	0.02860	1.28	0.1994	1.12063

The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol



The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol



The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol

Number of Observations Read	3132
Number of Observations Used	3132

Stepwise Selection: Step 1

Variable Systolic Entered: R-Square = 0.0350 and C(p) = 8.6847

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	203121	203121	113.51	<.0001
Error	3130	5600817	1789.39853		
Corrected Total	3131	5803938			

Variable	**Parameter Estimate**	**Standard Error**	**Type II SS**	**F Value**	**Pr > F**
Intercept	165.29119	5.37250	1693764	946.55	<.0001
Systolic	0.43164	0.04051	203121	113.51	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable Diastolic Entered: R-Square = 0.0372 and C(p) = 3.6475

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	215687	107843	60.38	<.0001
Error	3129	5588252	1785.95461		
Corrected Total	3131	5803938			

The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol

Stepwise Selection: Step 2

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	159.33174	5.81860	1339178	749.84	<.0001
Diastolic	0.27702	0.10444	12565	7.04	0.0080
Systolic	0.30221	0.06340	40586	22.72	<.0001

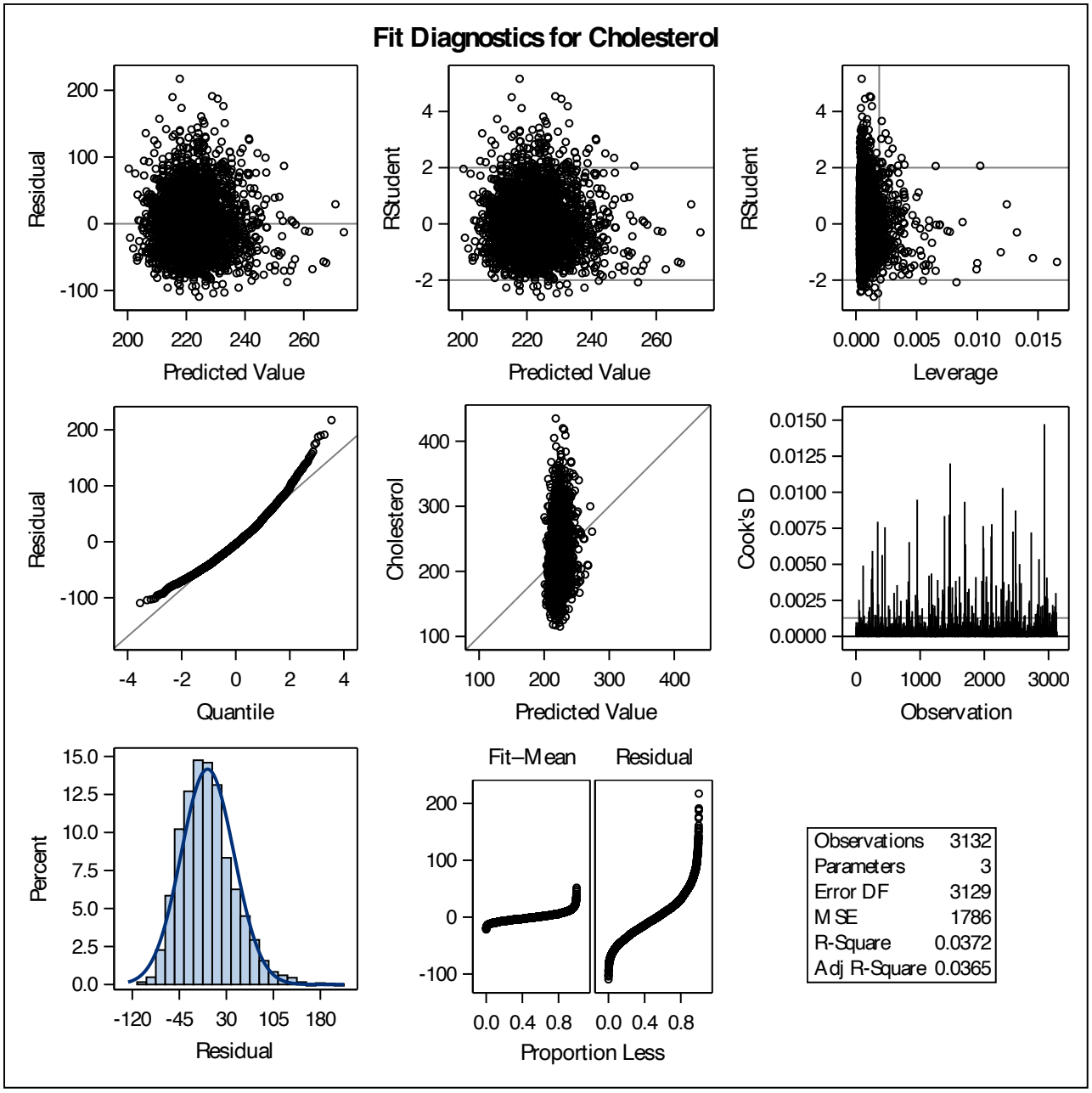
Bounds on condition number: 2.4534, 9.8136

All variables left in the model are significant at the 0.0500 level.

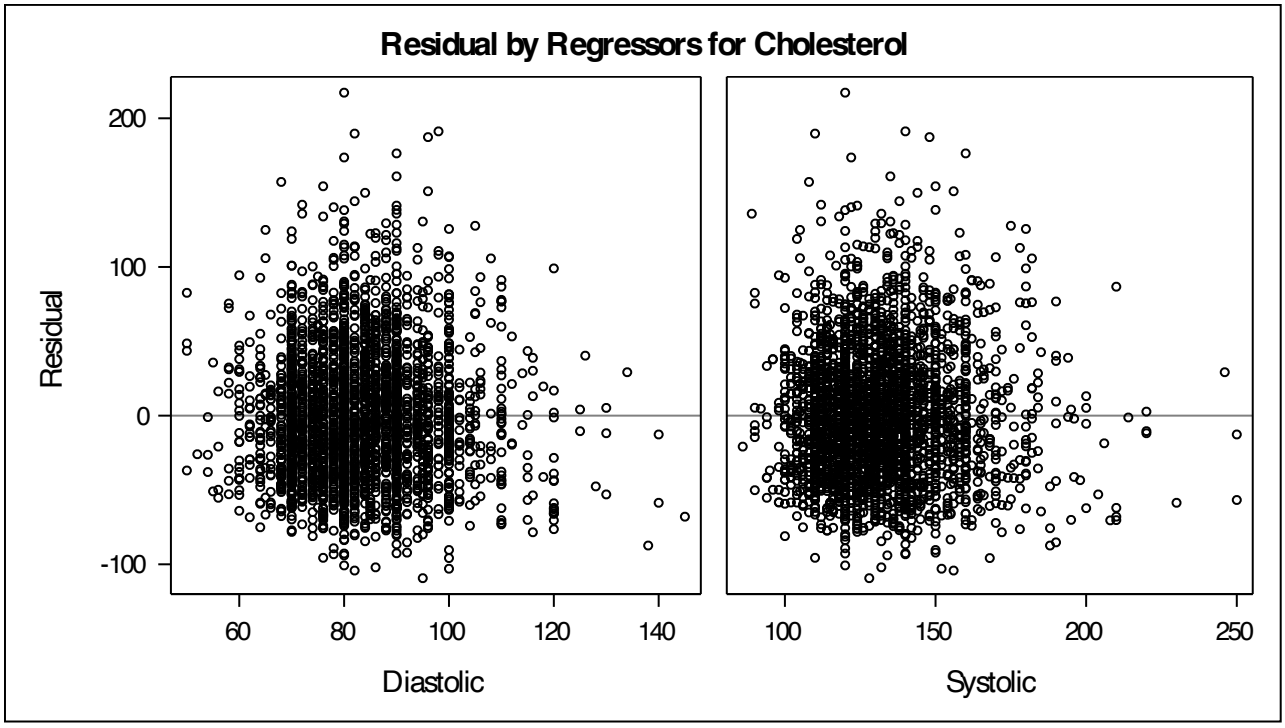
No other variable met the 0.0500 significance level for entry into the model.

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	Systolic		1	0.0350	0.0350	8.6847	113.51	<.0001
2	Diastolic		2	0.0022	0.0372	3.6475	7.04	0.0080

The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol



The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol



The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol

Number of Observations Read	3132
Number of Observations Used	3132

Forward Selection: Step 1

Variable Systolic Entered: R-Square = 0.0350 and C(p) = 8.6847

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	203121	203121	113.51	<.0001
Error	3130	5600817	1789.39853		
Corrected Total	3131	5803938			

Variable	**Parameter Estimate**	**Standard Error**	**Type II SS**	**F Value**	**Pr > F**
Intercept	165.29119	5.37250	1693764	946.55	<.0001
Systolic	0.43164	0.04051	203121	113.51	<.0001

Bounds on condition number: 1, 1

Forward Selection: Step 2

Variable Diastolic Entered: R-Square = 0.0372 and C(p) = 3.6475

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	215687	107843	60.38	<.0001
Error	3129	5588252	1785.95461		
Corrected Total	3131	5803938			

The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol

Forward Selection: Step 2

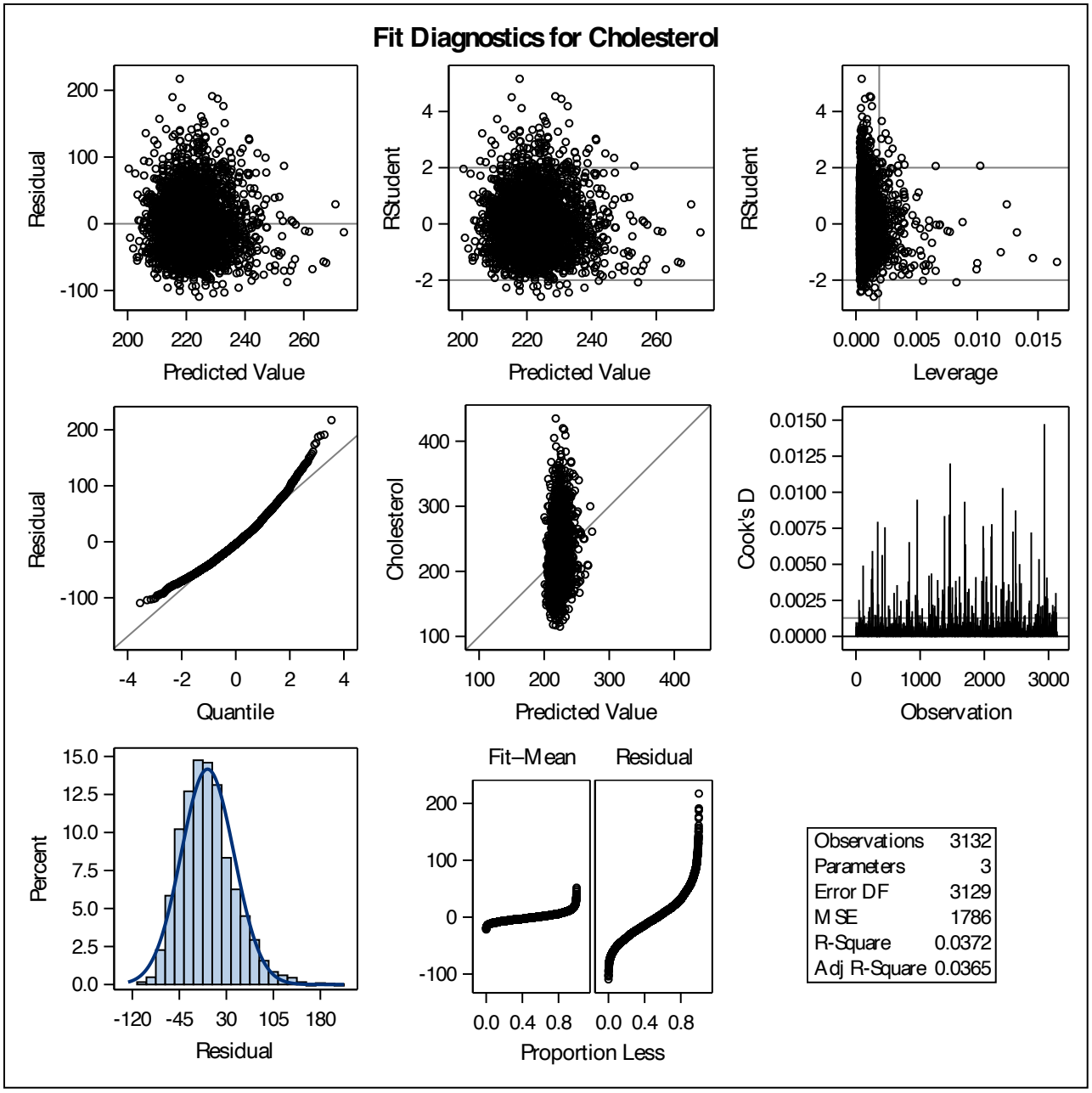
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	159.33174	5.81860	1339178	749.84	<.0001
Diastolic	0.27702	0.10444	12565	7.04	0.0080
Systolic	0.30221	0.06340	40586	22.72	<.0001

Bounds on condition number: 2.4534, 9.8136

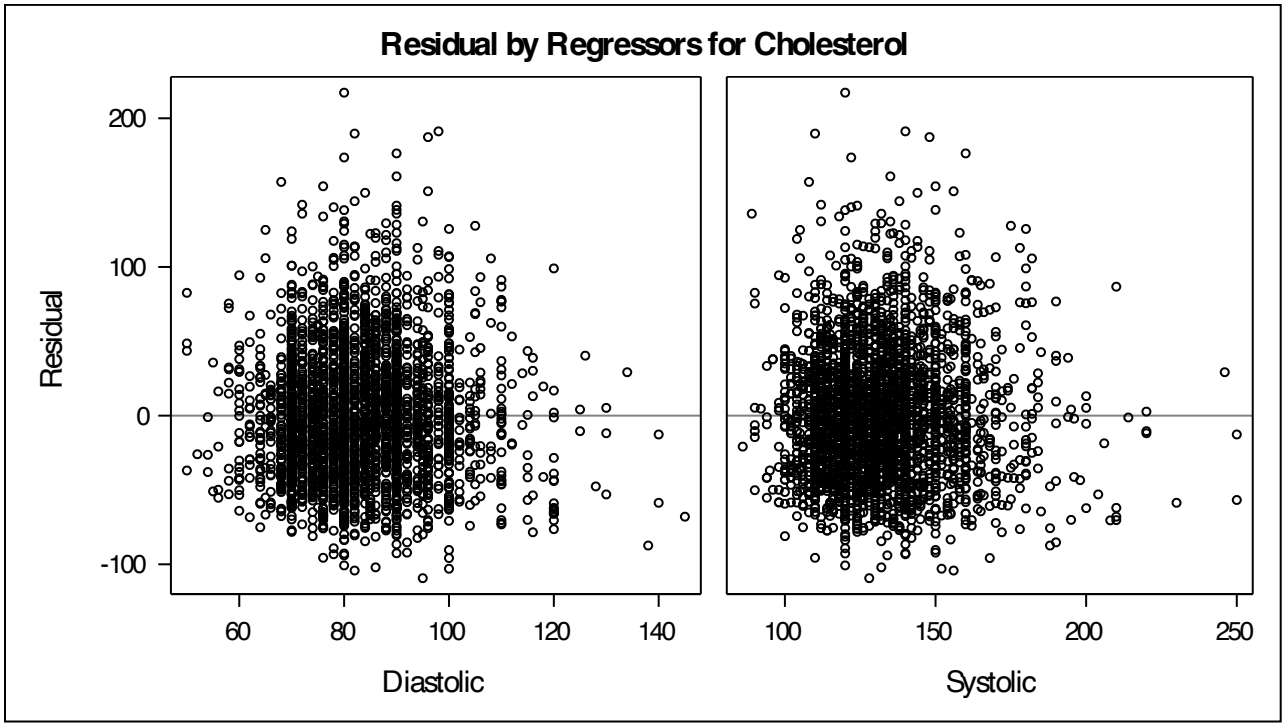
No other variable met the 0.0500 significance level for entry into the model.

Summary of Forward Selection							
Step	**Variable Entered**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Systolic	1	0.0350	0.0350	8.6847	113.51	<.0001
2	Diastolic	2	0.0022	0.0372	3.6475	7.04	0.0080

The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol



The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol



The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol

Number of Observations Read	3132
Number of Observations Used	3132

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.0377 and C(p) = 4.0000

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	218628	72876	40.81	<.0001
Error	3128	5585310	1785.58508		
Corrected Total	3131	5803938			

Variable	**Parameter Estimate**	**Standard Error**	**Type II SS**	**F Value**	**Pr > F**
Intercept	156.32618	6.27153	1109424	621.32	<.0001
Diastolic	0.24922	0.10665	9750.75172	5.46	0.0195
Systolic	0.30073	0.06340	40174	22.50	<.0001
Weight	0.03671	0.02860	2941.81959	1.65	0.1994

Bounds on condition number: 2.5589, 18.401

Backward Elimination: Step 1

Variable Weight Removed: R-Square = 0.0372 and C(p) = 3.6475

The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol

Backward Elimination: Step 1

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	215687	107843	60.38	<.0001
Error	3129	5588252	1785.95461		
Corrected Total	3131	5803938			

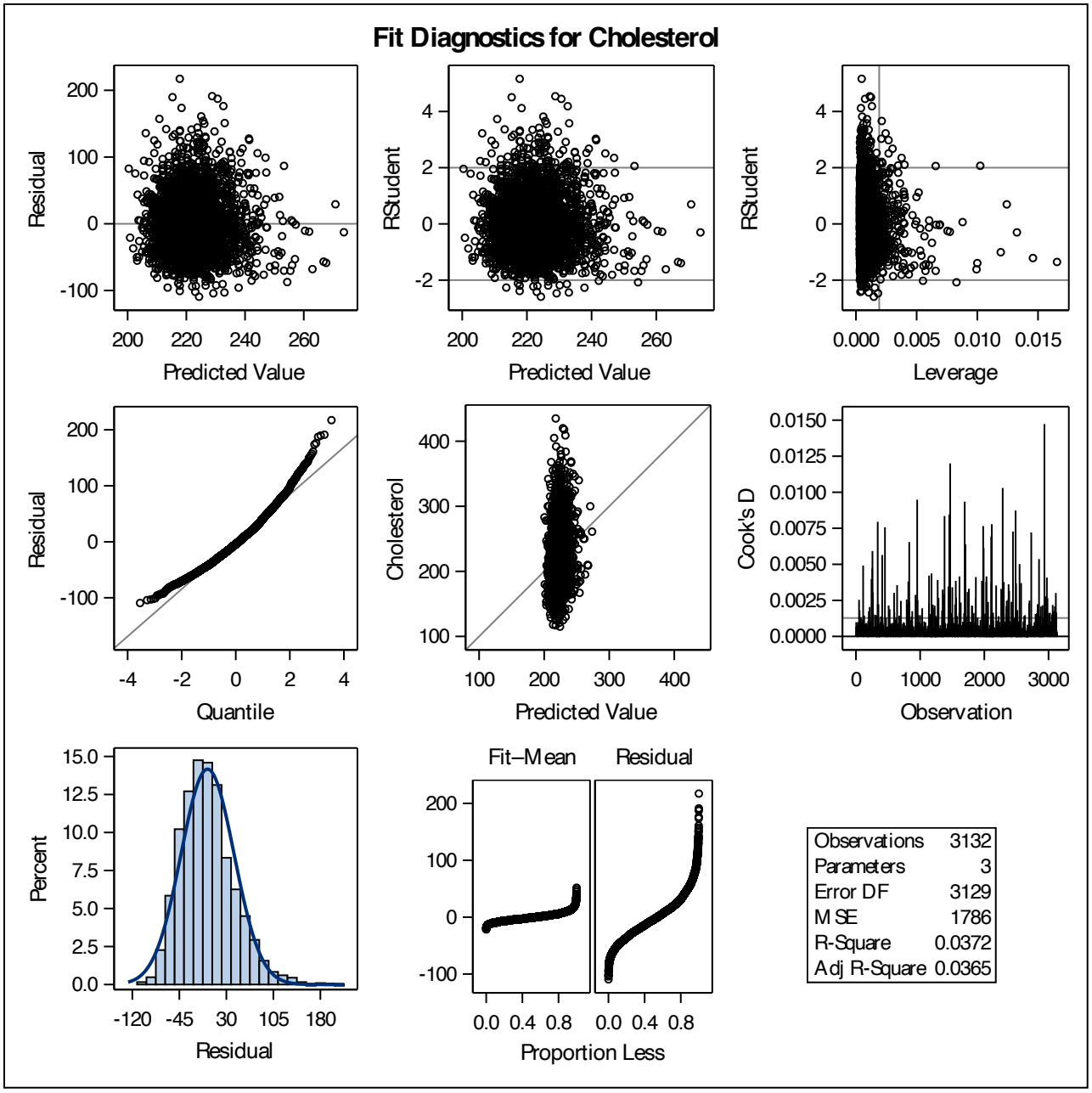
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	159.33174	5.81860	1339178	749.84	<.0001
Diastolic	0.27702	0.10444	12565	7.04	0.0080
Systolic	0.30221	0.06340	40586	22.72	<.0001

Bounds on condition number: 2.4534, 9.8136

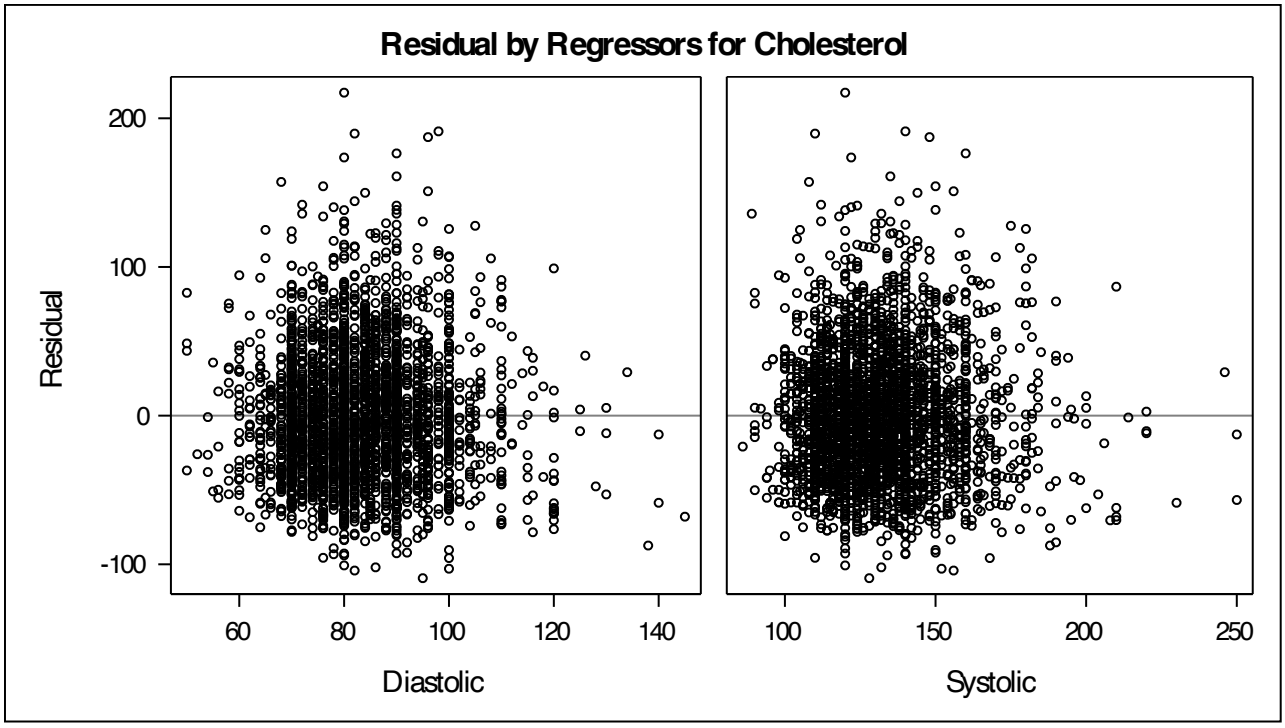
All variables left in the model are significant at the 0.0500 level.

Summary of Backward Elimination							
Step	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Weight	2	0.0005	0.0372	3.6475	1.65	0.1994

The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol



The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol



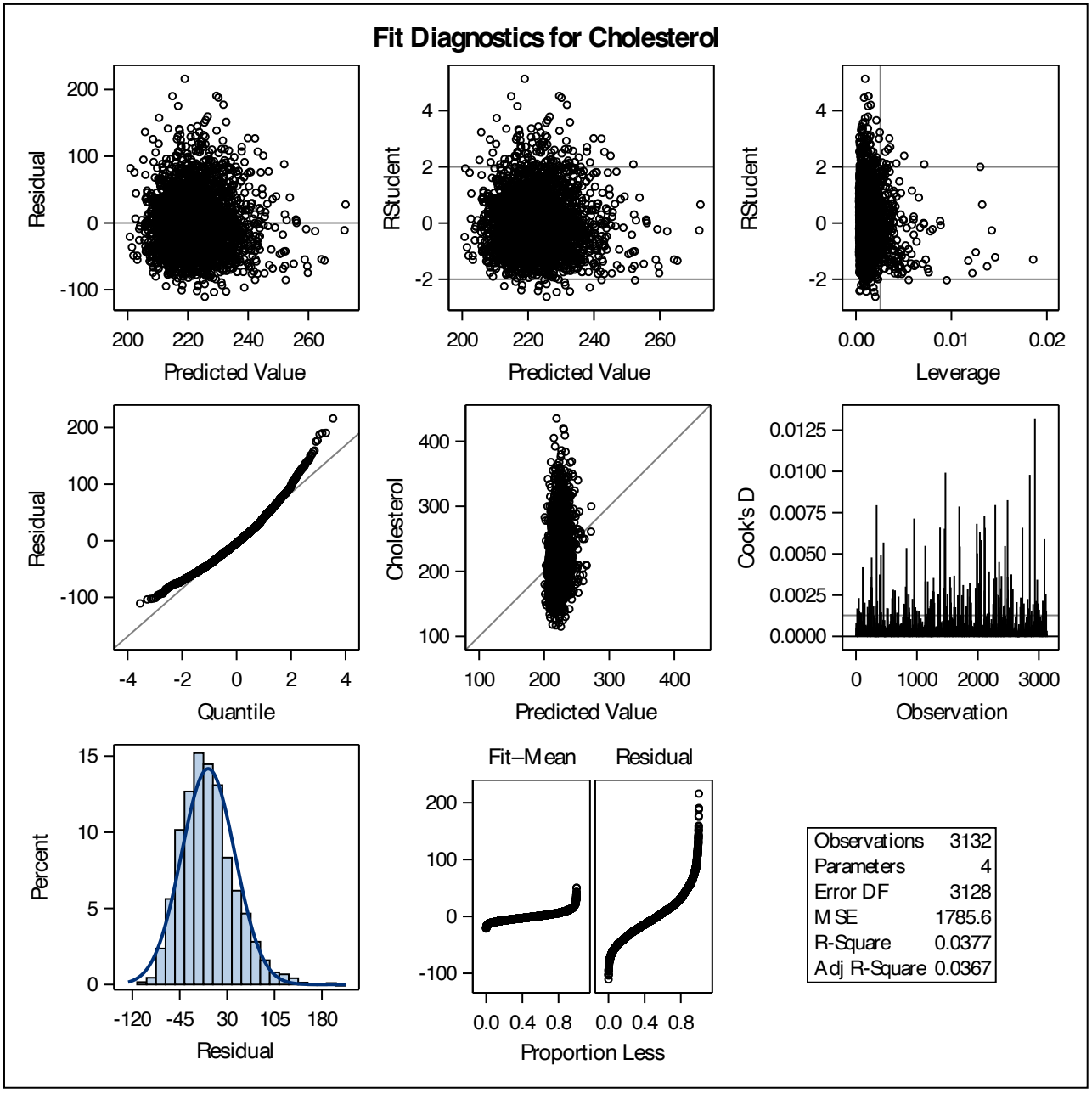
The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol

Adjusted R-Square Selection Method

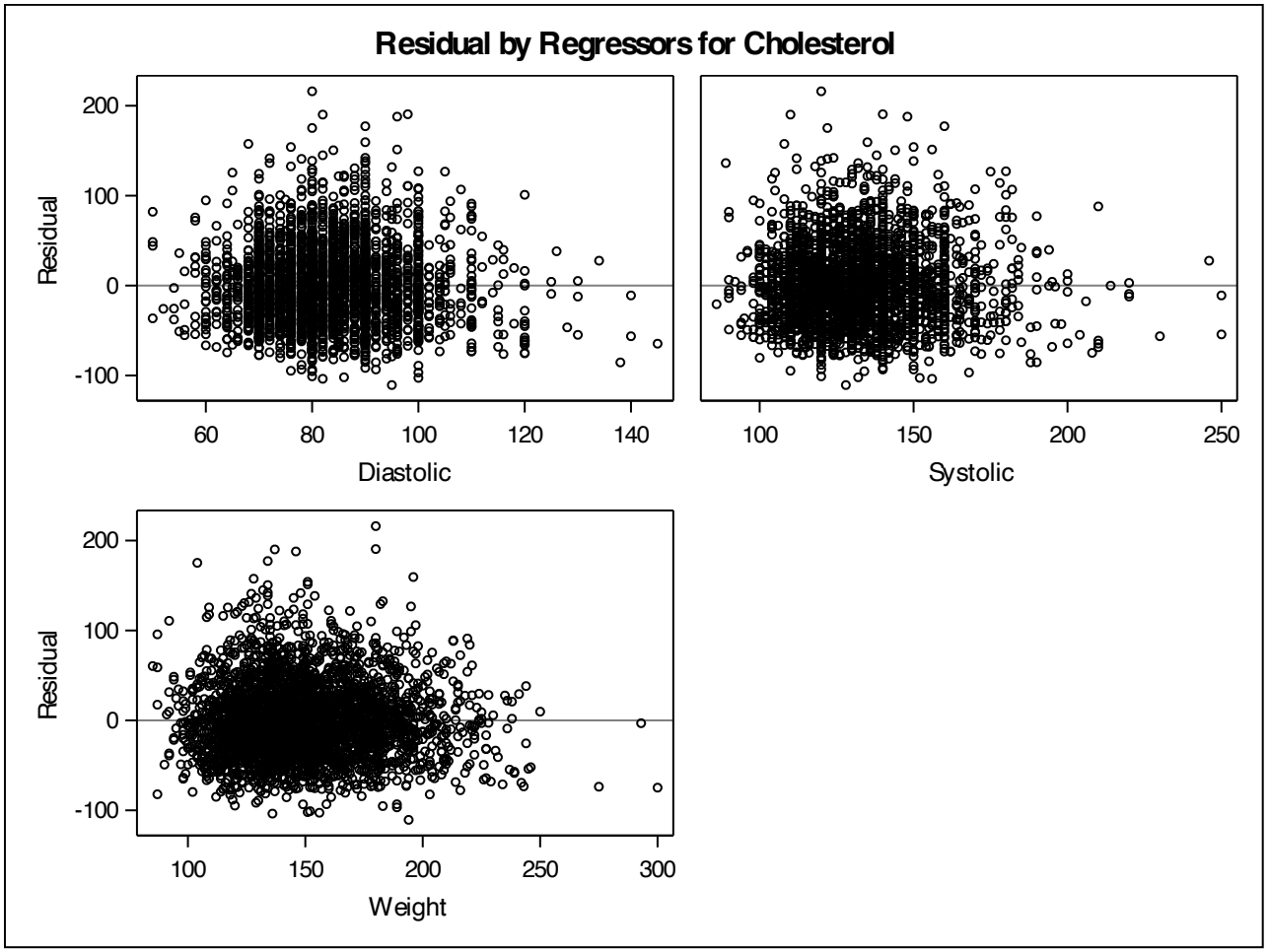
Number of Observations Read	3132
Number of Observations Used	3132

Number in Model	Adjusted R-Square	R-Square	Variables in Model
3	0.0367	0.0377	Diastolic Systolic Weight
2	0.0365	0.0372	Diastolic Systolic
2	0.0354	0.0360	Systolic Weight
1	0.0347	0.0350	Systolic
2	0.0301	0.0307	Diastolic Weight
1	0.0299	0.0302	Diastolic
1	0.0060	0.0063	Weight

The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol



The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol



We will first look at VIFs. We got VIF for diastolic 2.56, for systolic 2.45 and for weight 1.12. The values were not that high to remove the variables. So, hence we did stepwise selection and got the model with diastolic and systolic predictors. Based on the result, we can see that intercept, diastolic and systolic were significant (small p-values). Next, we did forward selection, the results were the same as in stepwise selection. Finally, we performed backward selection, it showed us the predictor that was removed, which is weight. In the preceding table, we can see the retained variables, diastolic and systolic. We could also select the model based on adjusted R-squared. However, there is only 0.0002 difference between the R-squared values of the first and the second model. So, we can choose more simpler model that was also chosen based on the previously mentioned selection methods.

The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol

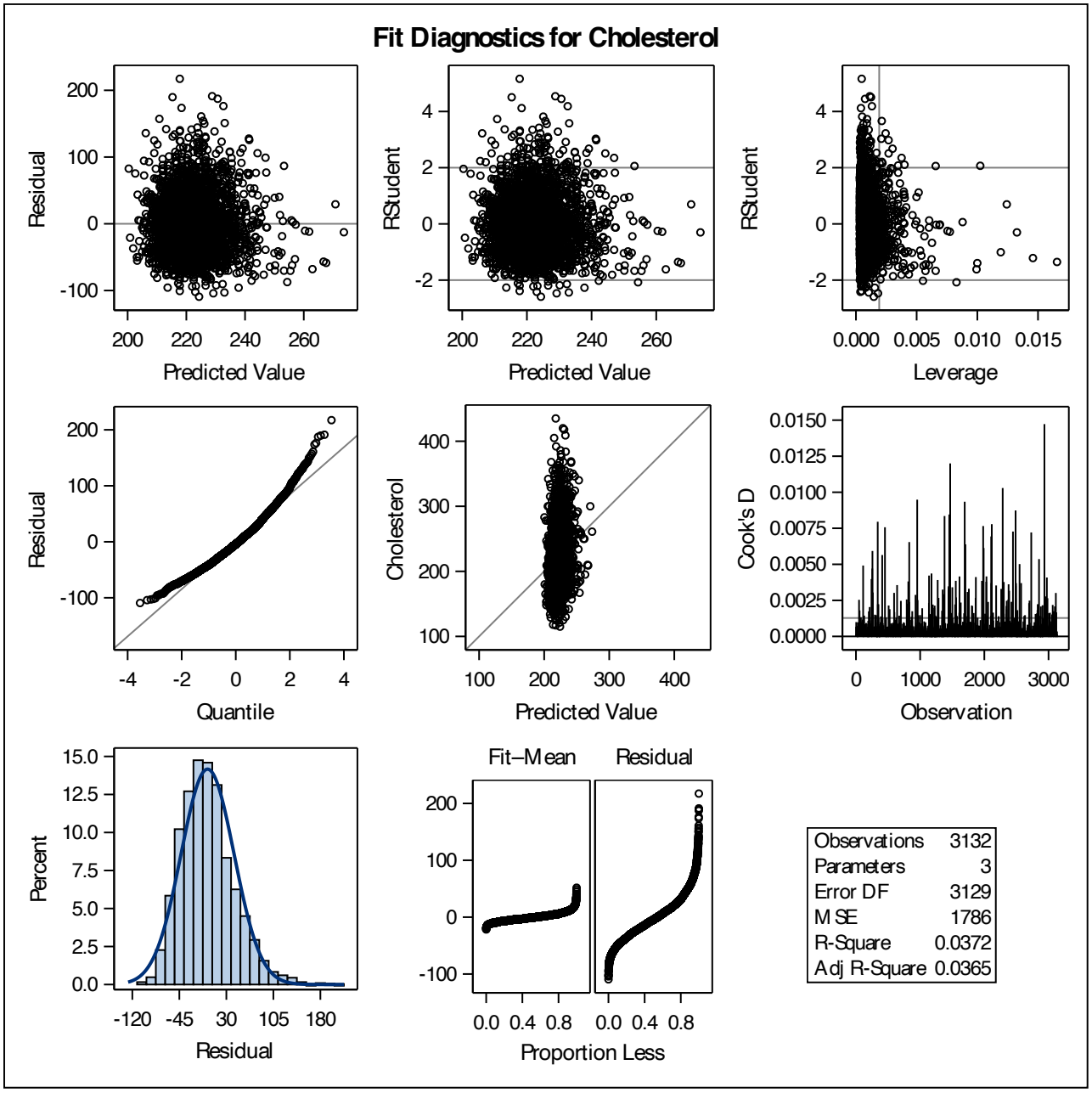
Number of Observations Read	3132
Number of Observations Used	3132

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	215687	107843	60.38	<.0001
Error	3129	5588252	1785.95461		
Corrected Total	3131	5803938			

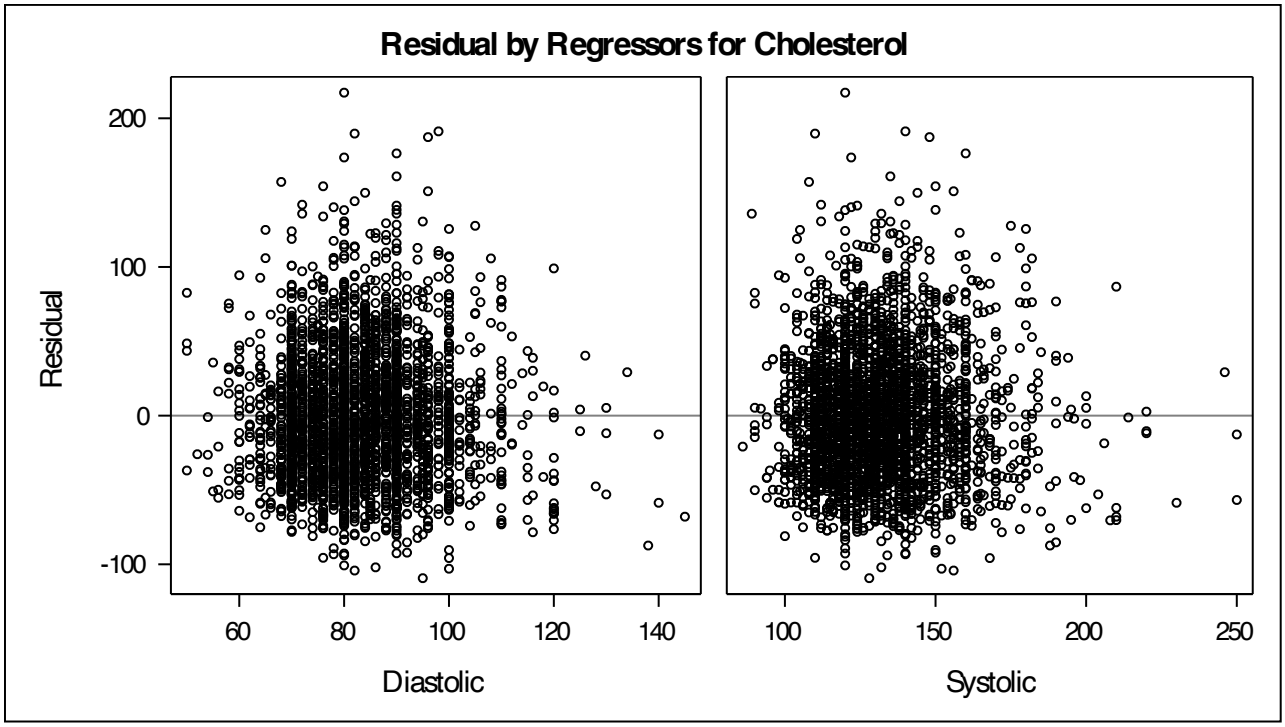
Root MSE	42.26056	**R-Square**	0.0372
Dependent Mean	221.96201	**Adj R-Sq**	0.0365
Coeff Var	19.03955		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	159.33174	5.81860	27.38	<.0001
Diastolic	1	0.27702	0.10444	2.65	0.0080
Systolic	1	0.30221	0.06340	4.77	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol



The REG Procedure
Model: MODEL1
Dependent Variable: Cholesterol



Exercise 4b):

We now selected the final model with diastolic and systolic for modeling cholesterol. The diagnostics looks fine and very similar to the previous diagnostics results. There are also no observations with Cook's distance larger than 0.015. The model is highly statistically significant (F-value = 60.38) but it only describes 3.72% of the variation in cholesterol. The intercept β_0 is estimated to be about 159.3 and the coefficient for diastolic is estimated to be just 0.28, the coefficient for systolic is estimated to be 0.3. As we can see they are all positive. As the diastolic level increases by 1, about 0.28 cholesterol is expected. And as systolic level increases by 1, an increase of about 0.3 cholesterol is expected. I would not say that this model will be appropriate to use: although diagnostics is good, but we have very small percentage of variation explained. So, there is very little of a relationship between the independent variables (diastolic, systolic) and dependent variable (cholesterol).