STAT448 - Advanced Data Analysis

Homework 3

Name: Xinyan Yang

Exercise 1 Solution:

(a).

		salary		
		Mean	Std	N
sex	rank			
F	Assist	36.83	3.06	6
	Assoc	42.25	1.71	4
M	Assist	40.25	2.63	4
	Assoc	46.38	3.54	8

According to the table above, we can find that the male have greater salary than the female at the same rank level and associate's salary is greater than the assistant's for both male and female. Moreover, we can find that these cells don't have the same variance. Finally, the count for each cell is different so it is not balanced.

(b)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	325.6098485	108.5366162	11.90	0.0002
Error	18	164.2083333	9.1226852		
Corrected Total	21	489.8181818			

R-Square	Coeff Var	Root MSE	salary Mean
0.664757	7.206976	3.020378	41.90909

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sex	1	155.1515152	155.1515152	17.01	0.0006
rank	1	169.8245614	169.8245614	18.62	0.0004
sex*rank	1	0.6337719	0.6337719	0.07	0.7951

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sex	1	71.8442982	71.8442982	7.88	0.0117
rank	1	168.2653509	168.2653509	18.44	0.0004
sex*rank	1	0.6337719	0.6337719	0.07	0.7951

About the Type I sums of squares, we can find that when we add the sex to the model, model can explain variation of 155.15, and then after we add the rank to the model, it can explain additional variation of 169.82. However, if we add the interaction term to the model with sex and rank, there will be only 0.63 additional variation explained.

About the Type III SS, if we add sex to the model with rank and interaction term, the model can explain additional variation of 71.84. Similarly, if we add rank to the model with sex and interaction term, the model can explain additional variation of 168.27. However, if we add interaction term to the model with sex and rank, we can only get additional variation of 0.63, which is very small compared to the other two.

Therefore, both two types of error tell us that the sex and rank are significant and the interaction term isn't significant.

We can also check that the P value for the interaction term is 0.7951, which indicates that this interaction term is not significant under the significance level of 5%.

(c)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	324.9760766	162.4880383	18.73	<.0001
Error	19	164.8421053	8.6759003		
Corrected Total	21	489.8181818			

R-Square	Coeff Var	Root MSE	salary Mean
0.663463	7.028280	2.945488	41.90909

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sex	1	155.1515152	155.1515152	17.88	0.0005
rank	1	169.8245614	169.8245614	19.57	0.0003

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sex	1	72.7578947	72.7578947	8.39	0.0093
rank	1	169.8245614	169.8245614	19.57	0.0003

The P value of anova model is less than 0.001, which means that the model is significant. According to the R-Square, the model can explain 66.35% of the data variation.

The P-value for sex and rank in both Type I test and Type III test are less than 0.01, which means that the main effect of these two variables are all significant under the significance level of 5%.

(d).

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

sex	salary LSMEAN	H0:LSMean1=LSMean2 Pr > |t|
F	39.5789474	0.0093
M	43.3684211	

Least Squares Means for Effect sex				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-3.789474	-6.528263	-1.050685

rank	salary LSMEAN	H0:LSMean1=LSMean2 Pr > |t|
Assist	38.5789474	0.0003
Assoc	44.3684211	

Least Squares Means for Effect rank				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-5.789474	-8.528263	-3.050685

According to the results in (b) and (c), I choose the model with only two terms: rank and sex. So the anova model result is the same with (c). The P value of anova model is less than 0.001, which means that the model is significant. According to the R-Square, the model can explain 66.35% of the data variation.

Since the least squares means confidence interval for sex doesn't include 0, we can conclude that there is significant difference for the salary between male and female teacher. Similarly, the least squares means confidence interval for rank doesn't include 0, we can conclude that there is significant difference for the salary between the associate and assistant.

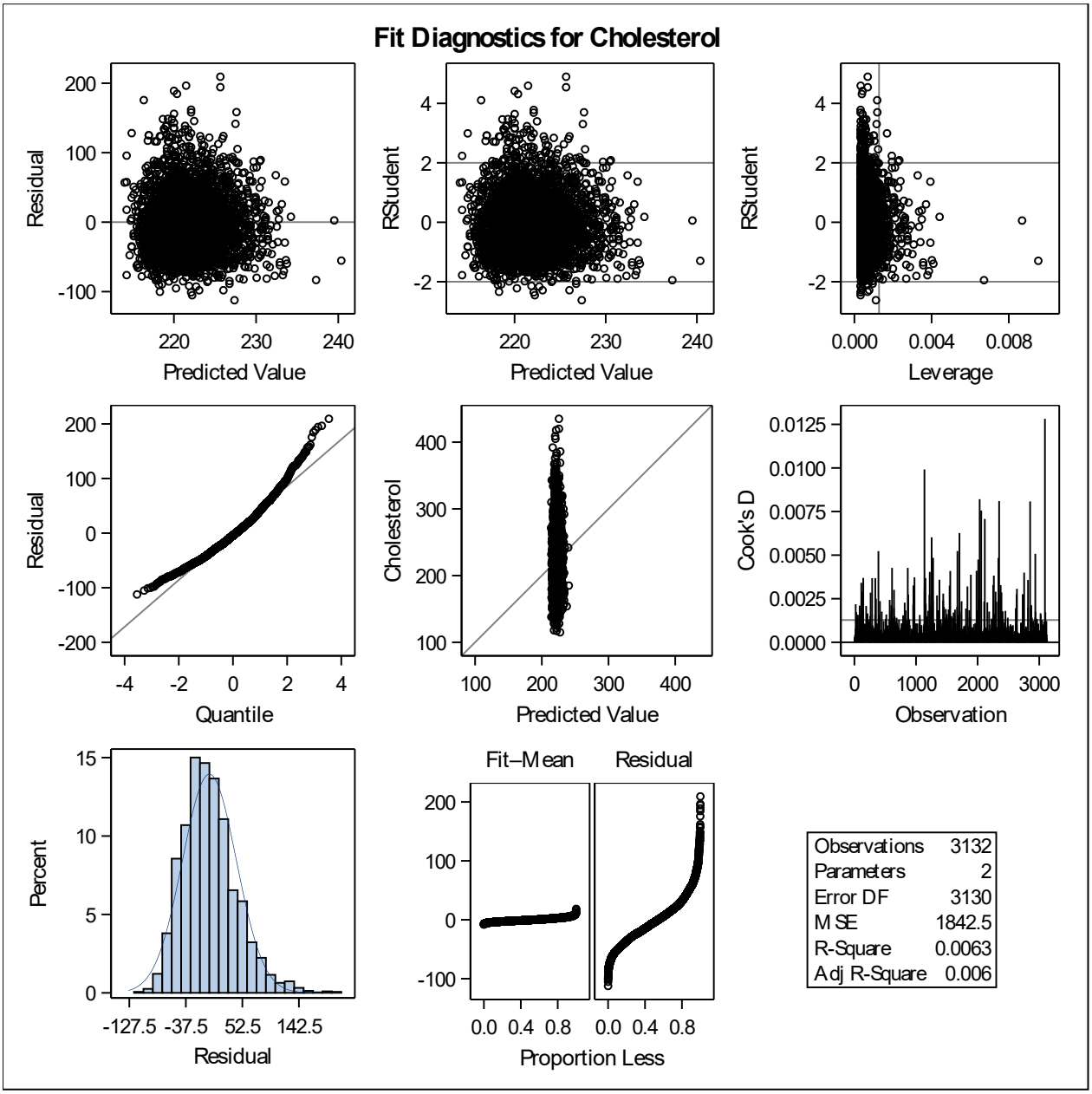
About the model diagnostics, I have coded in SAS, but somehow always got the warning message that "Output 'DiagnosticsPanel' was not created". And I feel really confusing about that.

Exercise 2 Solution:

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



From the above regression results, we can get the following conclusions.

The model's P-value is less than 0.0001, so it is significant under the significance level of 5%. Also, the parameter's and intercept's P-value are both less than 0.0001, so the weight term is significant too. The R square is 0.0063, therefore the model can explain only 0.63% of the data variation. From the diagnostics panel, we can find that the residual nearly follows normal distribution, but the head and tail is a little deviated.

Since the parameter estimate for the weight is 0.123, we can say that when weight increases by 1 unit, the cholesterol level would increase by 0.123 unit. They are positively related.

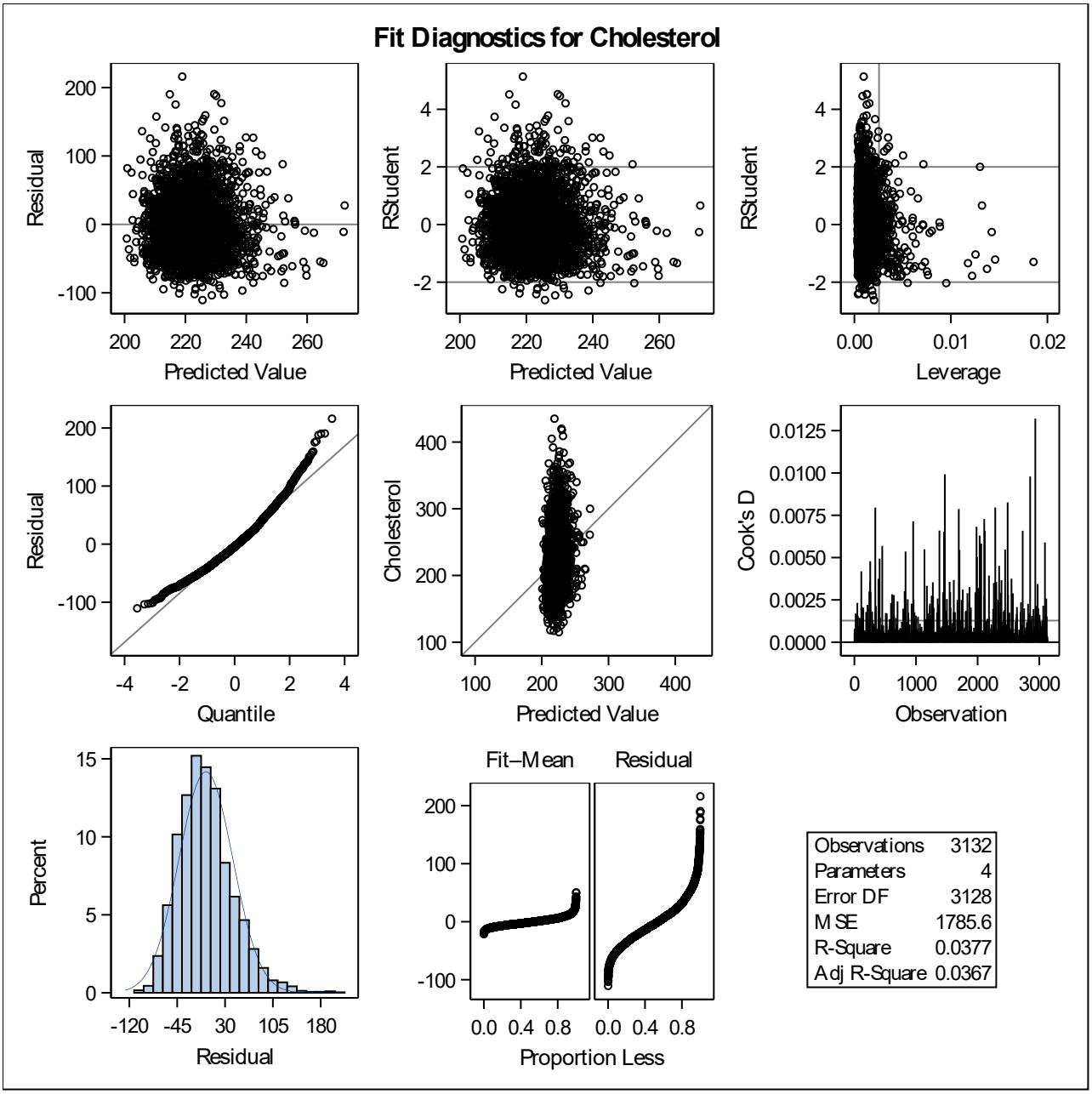
Because the variation explained by this mode is too small and the residual plot seems not very normal, this model is not a good model and it can still be improved.

Exercise 3 Solution:

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



Pearson Correlation Coefficients, N = 3132 Prob > |r| under H0: Rho=0			
	Weight	**Diastolic**	**Systolic**
Weight	1.00000	0.32764 <.0001	0.26318 <.0001
Diastolic	0.32764 <.0001	1.00000	0.76968 <.0001
Systolic	0.26318 <.0001	0.76968 <.0001	1.00000

From the above regression results, we can get the following conclusions.

From the diagnostics panel, we can find that the residual still shows a right-tail trend.

The model's P-value is less than 0.0001, so it is significant under the significance level of 5%, which means that this model is useful.

Also, the P-value for diastolic and Systolic are both less than 0.05, so these two terms are significant too. However, the P value of Weight is not significant any more.

The R square is 0.0377, therefore the model can explain only 3.77% of the data variation.

From the Pearson correlation Coefficients, we can see that the p-values are all <0.0001, therefore we can conclude that there are correlation relationship between these 3 predictors, which means that there exists a multicollinearity issue.

Since the model has a multilinearity issue and the R squared is quite small, I don't think it's a good model.

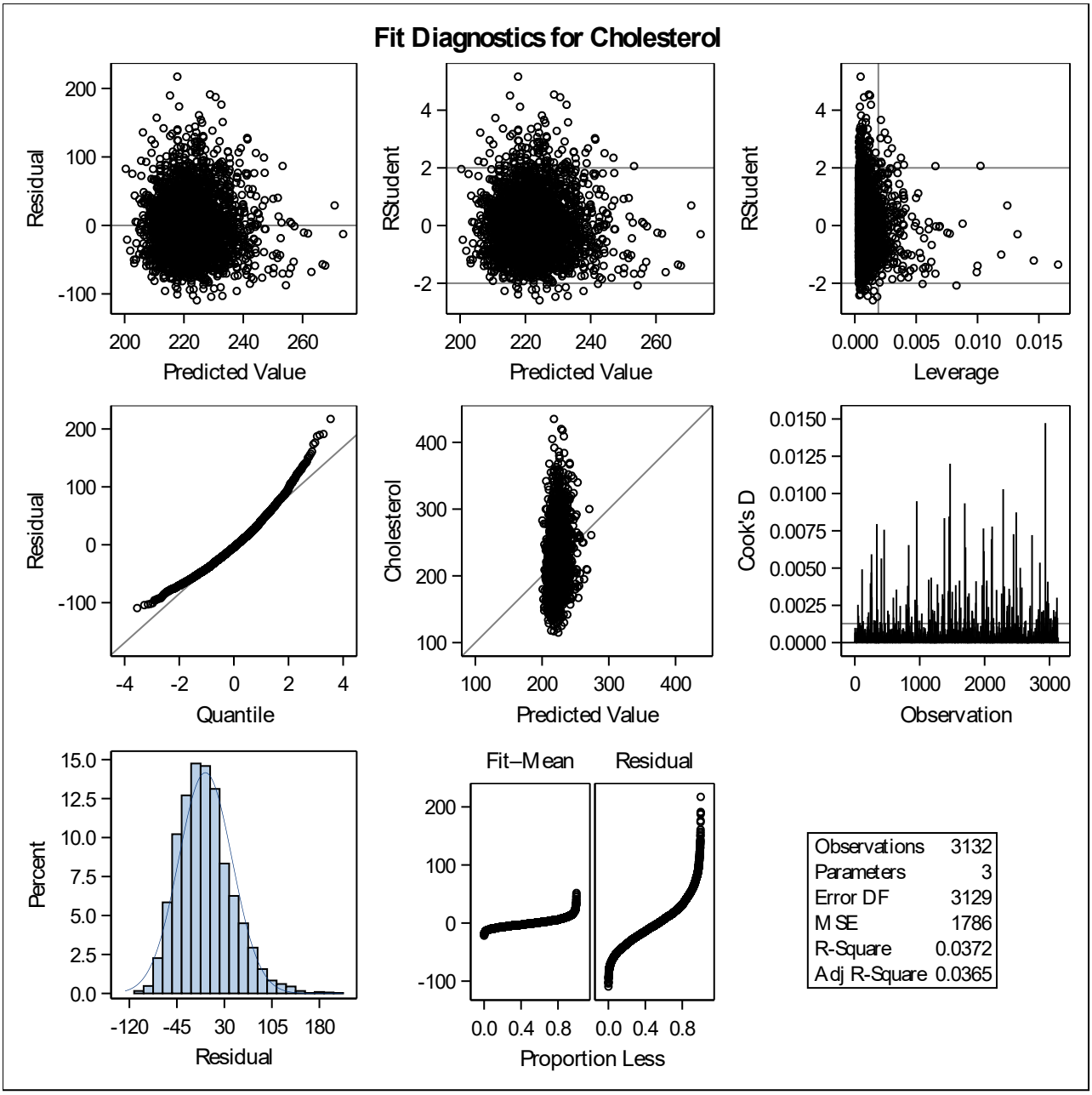
Exercise 4 Solution:

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

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



After the forward selection, the model has two predictors which are Diastolic and Systolic. Both parameter estimates for these two terms are significant.

The P-value for the whole mode is less than 0.0001, which means that we should reject the null hypothesis and conclude that this model is useful.

This model can explain about 3.72% of the variation in data, which is a little less than the model in Exercise3 and more than that of the model in Exercise2, which is not surprising since this model has fewer predictors than the model in 3 and more predictors than the model in 2.