Homework 3

Jiangyan Feng (NetID: jf8)

Exercise 1

a) The process of building the logistic model is:

Step 1. We first use backward selection to optimize the set the predictors. The selection results are shown as below. The selected predictors are: Age DB Alamine TP ALB.

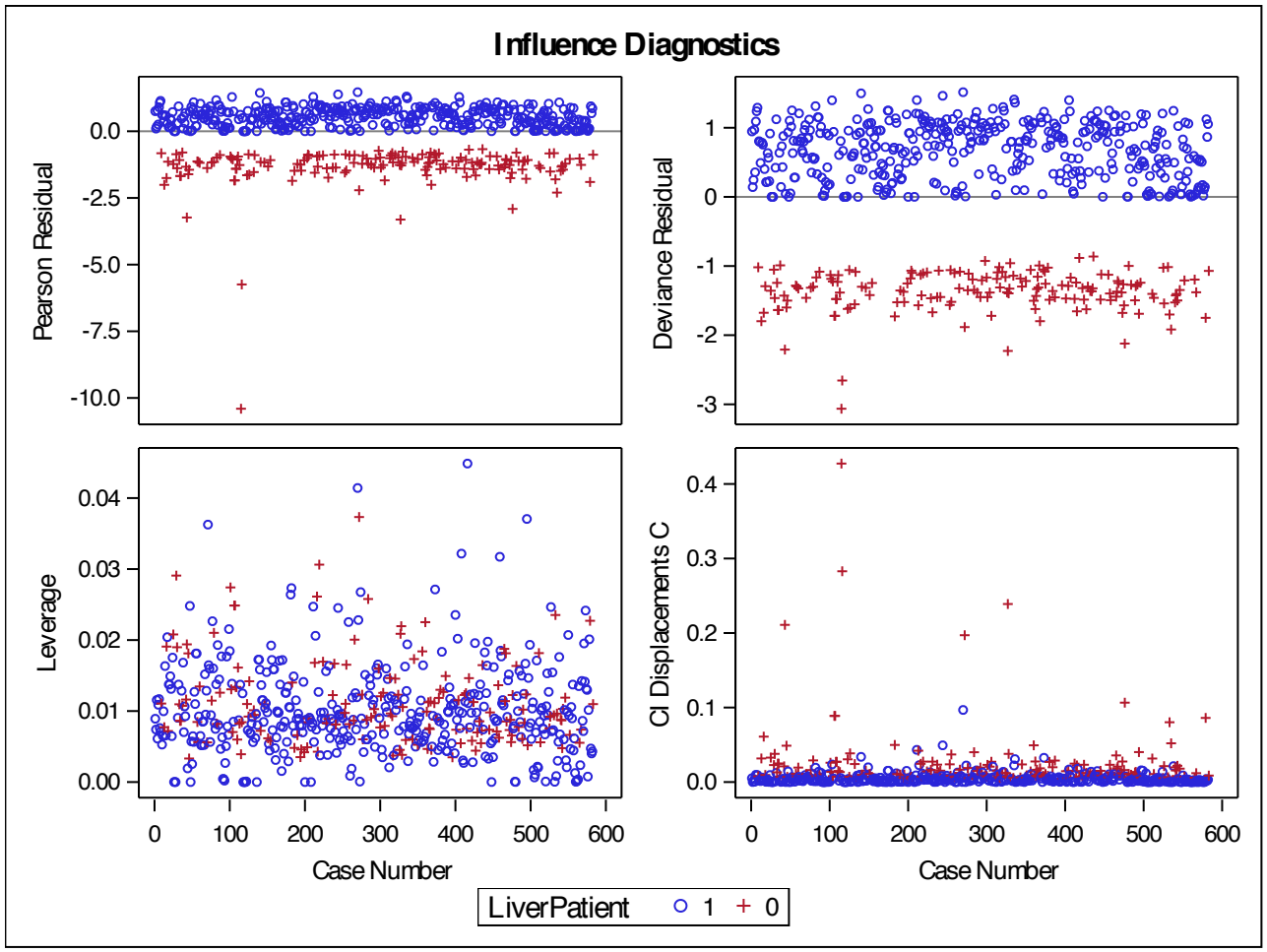
Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	TB	1	9	0.0167	0.8972
2	Gender	1	8	0.0178	0.8938
3	Aspartate	1	7	0.6317	0.4267
4	Alkphos	1	6	2.5213	0.1123
5	AGRatio	1	5	2.3791	0.1230

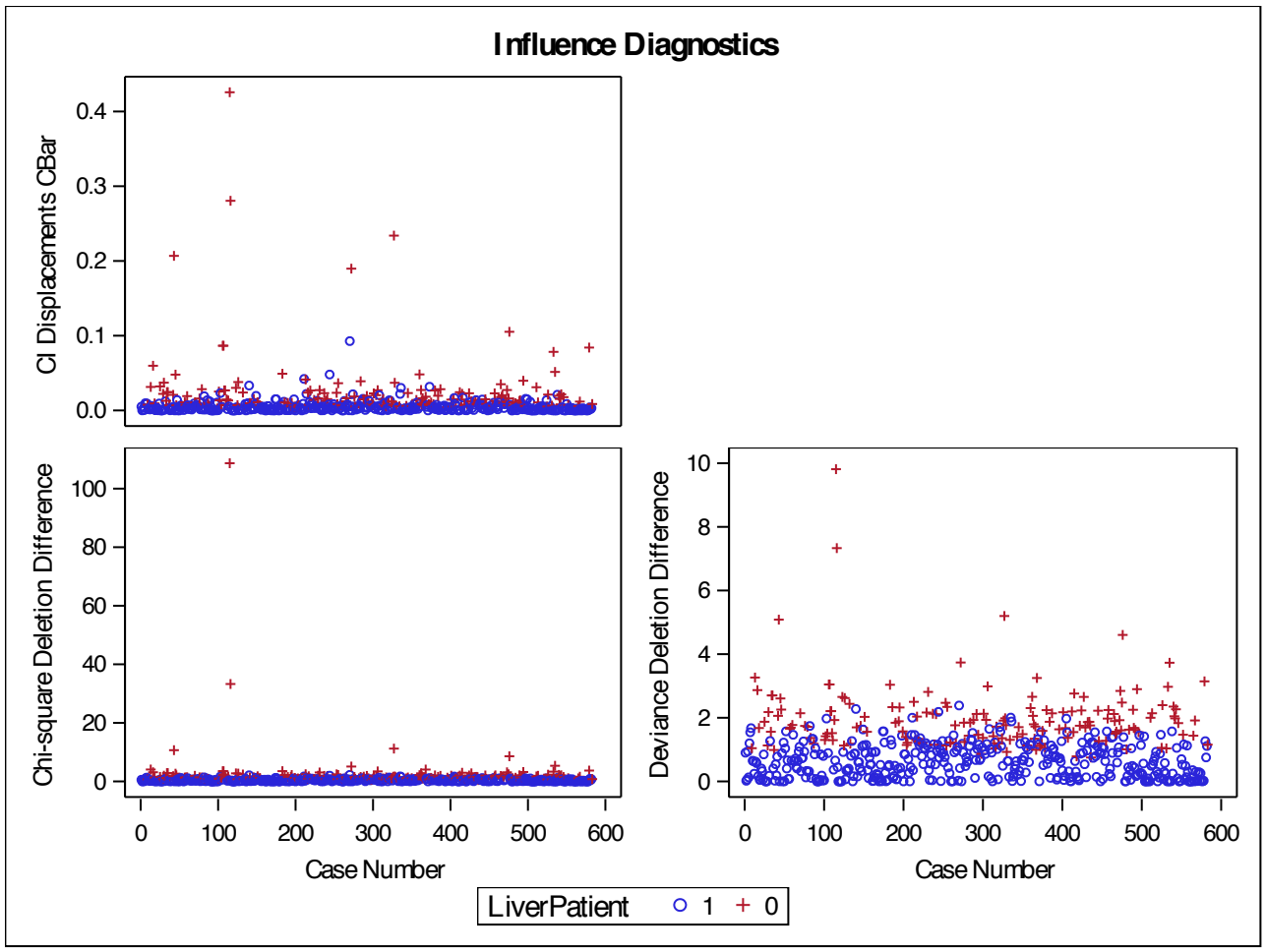
Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	7.9456	0.0048
DB	1	9.9749	0.0016
Alamine	1	16.0693	<.0001
TP	1	6.0789	0.0137
ALB	1	7.0729	0.0078

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.6635	0.7762	4.5934	0.0321
Age	1	0.0179	0.00636	7.9456	0.0048
DB	1	0.5534	0.1752	9.9749	0.0016
Alamine	1	0.0157	0.00392	16.0693	<.0001
TP	1	0.4333	0.1757	6.0789	0.0137
ALB	1	-0.6653	0.2502	7.0729	0.0078

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	1.018	1.005	1.031
DB	1.739	1.234	2.452
Alamine	1.016	1.008	1.024
TP	1.542	1.093	2.176
ALB	0.514	0.315	0.839

Step 2. We then fit the model with the selected predictors and check the influence of each data point. The influence is shown as below. Based on CBar results, there are few points deviating a lot from the rest.





Step 3. We then print out the unduly influential data points as shown below. And reselect the predictors using backward selection on the remaining data. The selected predictors are: Age DB Alkphos Alamine TP ALB.

Obs	Age	Gender	TB	DB	Alkphos	Alamine	Aspartate	TP	ALB	AGRatio	LiverPatient	cbar
43	42	Male	6.8	3.2	630	25	47	6.1	2.3	0.60	0	0.20684
115	50	Male	5.8	3.0	661	181	285	5.7	2.3	0.67	0	0.42574
116	50	Male	7.3	3.6	1580	88	64	5.6	2.3	0.60	0	0.28050
272	4	Male	0.8	0.2	460	152	231	6.5	3.2	0.90	0	0.18968
327	36	Female	1.2	0.4	358	160	90	8.3	4.4	1.10	0	0.23394
476	38	Male	2.2	1.0	310	119	42	7.9	4.1	1.00	0	0.10524

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	TB	1	9	0.0005	0.9829
2	Aspartate	1	8	0.0428	0.8361
3	Gender	1	7	0.0615	0.8041
4	AGRatio	1	6	3.5205	0.0606

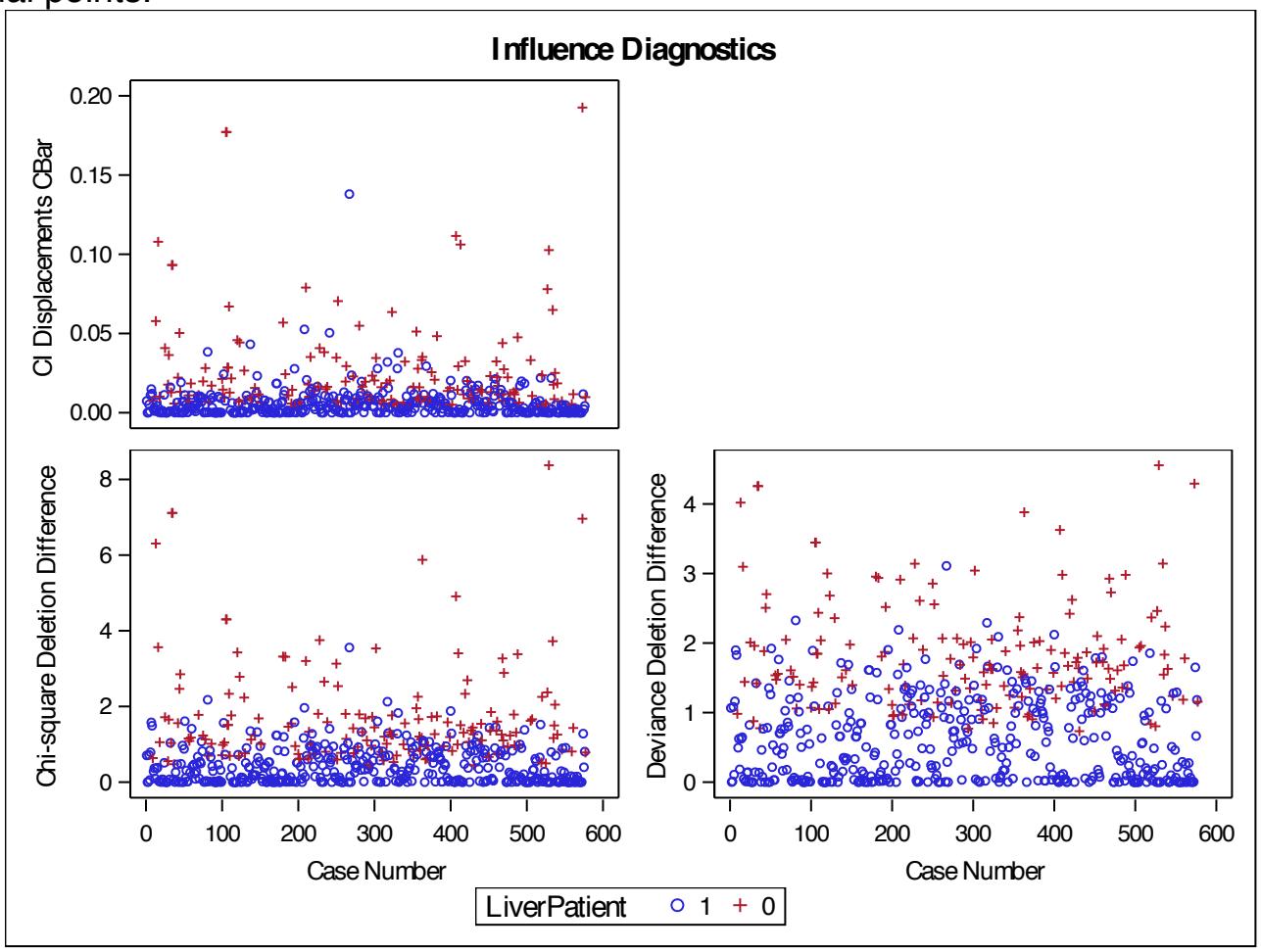
Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	7.7437	0.0054
DB	1	9.8638	0.0017
Alkphos	1	5.9723	0.0145
Alamine	1	15.2717	<.0001
TP	1	4.4005	0.0359
ALB	1	4.6622	0.0308

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.5939	0.8479	9.3588	0.0022
Age	1	0.0185	0.00665	7.7437	0.0054
DB	1	0.7672	0.2443	9.8638	0.0017
Alkphos	1	0.00288	0.00118	5.9723	0.0145
Alamine	1	0.0211	0.00541	15.2717	<.0001
TP	1	0.4000	0.1907	4.4005	0.0359
ALB	1	-0.5869	0.2718	4.6622	0.0308

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	1.019	1.005	1.032
DB	2.154	1.334	3.476
Alkphos	1.003	1.001	1.005
Alamine	1.021	1.011	1.032

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
TP	1.492	1.027	2.168
ALB	0.556	0.326	0.947

Step 4. We then refit the model and check the influence again. This time, no obvious unduly influential points.



b) The significance results and Hosmer-Lemeshow's test results are shown as below. According to AIC results, AIC for intercept and covariates is much lower than the intercept alone. This suggests that the model is significant. According to the likelihood ratio value, P value = <0.0001 suggests that at least one of these predictors are not 0. According to the maximum likelihood estimates, P value of all these 6 predictors are less than 0.05. This suggests all these 6 predictors are significant.

According to Hosmer-Lemeshow's test results, P value is 0.6028, much larger than 0.05. This suggests there is no strong evidence to reject null hypothesis. Therefore, the model fits the data well.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	685.208	544.063
SC	689.565	574.568
-2 Log L	683.208	530.063

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	153.1450	6	<.0001
Score	74.4420	6	<.0001
Wald	56.8403	6	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.6235	0.8450	9.6381	0.0019
Age	1	0.0191	0.00665	8.2855	0.0040
DB	1	0.7559	0.2424	9.7272	0.0018
Alkphos	1	0.00297	0.00119	6.2231	0.0126
Alamine	1	0.0212	0.00541	15.3953	<.0001
TP	1	0.3954	0.1902	4.3223	0.0376
ALB	1	-0.5821	0.2711	4.6106	0.0318

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
6.3974	8	0.6028

c) The significance results for odds ratios are shown as below. First of all, 1 is not in the range of 95% confidence limits for all 6 predictors. This suggests that all 6 predictors are significant.
A unit increase in Age, DB, Alkphos, Alamine, TP, ALB, increases the odds of being a liver patient by 1.019, 2.130, 1.003, 1.021, 1.485, 0.559 times, respectively.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	1.019	1.006	1.033
DB	2.130	1.324	3.425
Alkphos	1.003	1.001	1.005
Alamine	1.021	1.011	1.032
TP	1.485	1.023	2.156
ALB	0.559	0.328	0.951

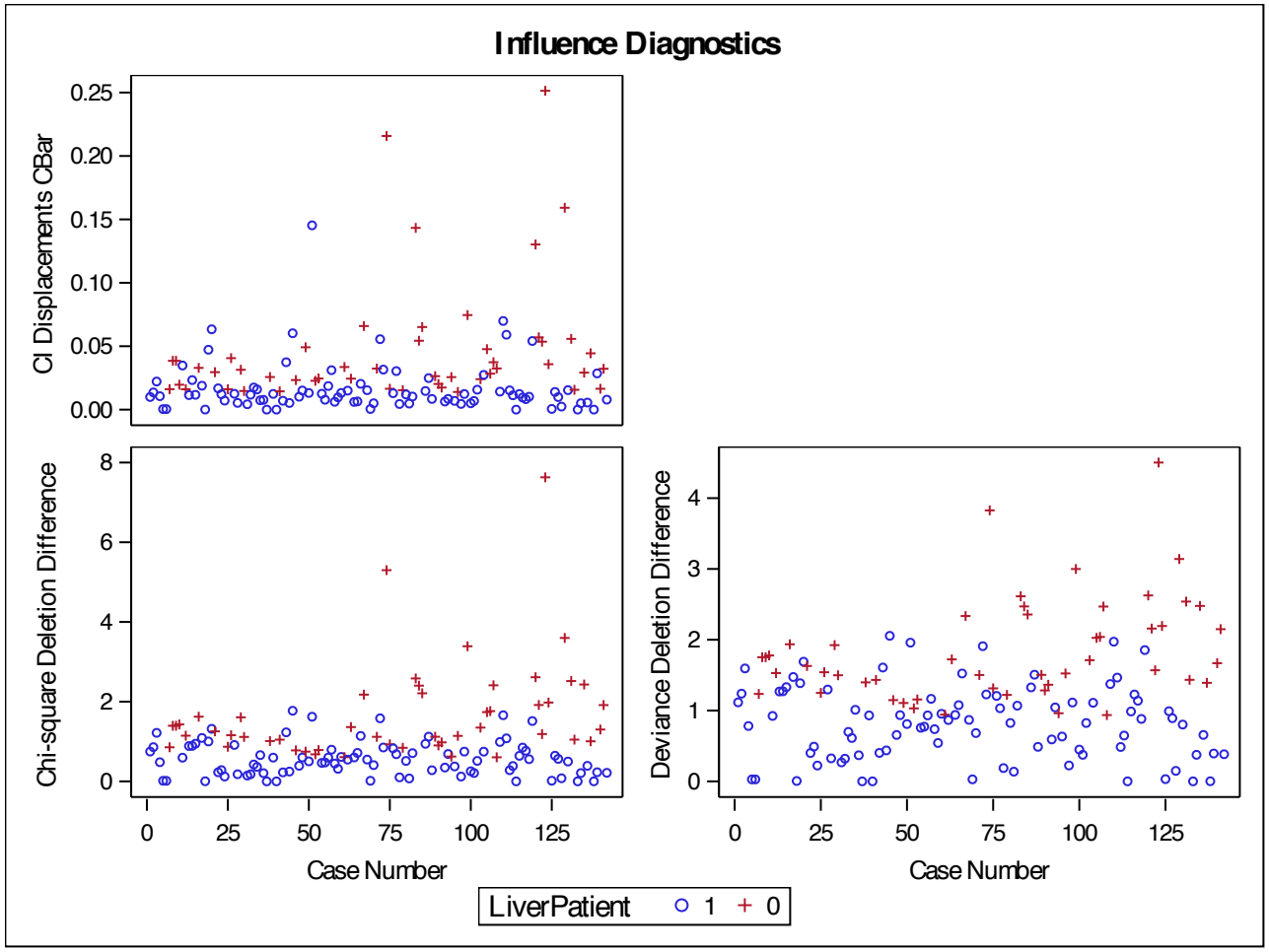
Exercise 2

a) The process of building the logistic model is:
Step 1. We first create a new dataset with only female data. We then use backward selection to optimize the set the predictors. The selection results are shown as below. The selected predictors are: Aspartate TP ALB.

Step 2: We then fit the model with the selected predictors and check the influence of each data point. The influence is shown as below. Based on CBar results, there are no points deviating a lot from the rest. Therefore, we use this model as our final model.

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	Age	1	8	0.0091	0.9242
2	Alamine	1	7	0.0187	0.8911
3	TB	1	6	0.1489	0.6996
4	Alkphos	1	5	1.0281	0.3106
5	AGRatio	1	4	1.3175	0.2510
6	DB	1	3	2.1765	0.1401

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Aspartate	1	5.5045	0.0190
TP	1	4.8919	0.0270
ALB	1	4.0556	0.0440



b) The significance results and Hosmer-Lemeshow's test results are shown as below. According to AIC results, AIC for intercept and covariates is much lower than the intercept alone. This suggests that the model is significant. According to the likelihood ratio value, P value = 0.0001 suggests that at least one of these predictors are not 0. According to the maximum likelihood estimates, P value of Asparatate and TP is less than 0.05. However, the p value of ALB is slightly above 0.0530. Considering Pvalue of ALB is significant in the backward elimination step and it is very close to 0.0530, we decide to keep this predictor.

According to Hosmer-Lemeshow's test results, P value is 0.5083, much larger than 0.05. This suggests there is no strong evidence to reject null hypothesis. Therefore, the model fits the data well.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	186.243	171.977
SC	189.199	183.800
-2 Log L	184.243	163.977

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	20.2663	3	0.0001
Score	11.4906	3	0.0093
Wald	10.7805	3	0.0130

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-2.1520	1.2672	2.8839	0.0895
Aspartate	1	0.0167	0.00702	5.6651	0.0173
TP	1	0.7811	0.3549	4.8437	0.0277
ALB	1	-0.9565	0.4943	3.7447	0.0530

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
7.2650	8	0.5083

c) The significance results for odds ratios are shown as below. First of all, 1 is not in the range of 95% confidence limits for Asparate and TP predictors. This suggests that these two predictors are significant. For ALB, the 95% confidence limits have a large range.

A unit increase in Asparate, TP, ALB, increases the odds of being a liver patient by 1.017, 2.184, 0.384 times, respectively.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Aspartate	1.017	1.003	1.031
TP	2.184	1.089	4.378
ALB	0.384	0.146	1.012

Comparison with the overall model:
1. For the overall model, more predictors are required. 6 (Age, DB, Alkphos, Alamine, TP, ALB) predictors are used for the overall model while only 3 (Asparate, TP, ALB) are used in the female model.
2. The overall model and the female model have two common predictors: TP, ALB. The increase of TP increases the odds of being liver patient for both the overall model and the odds in the female model. The increase of ALB decreases the odds of being liver patient for both the overall model and the odds in the female model.

Exercise 3

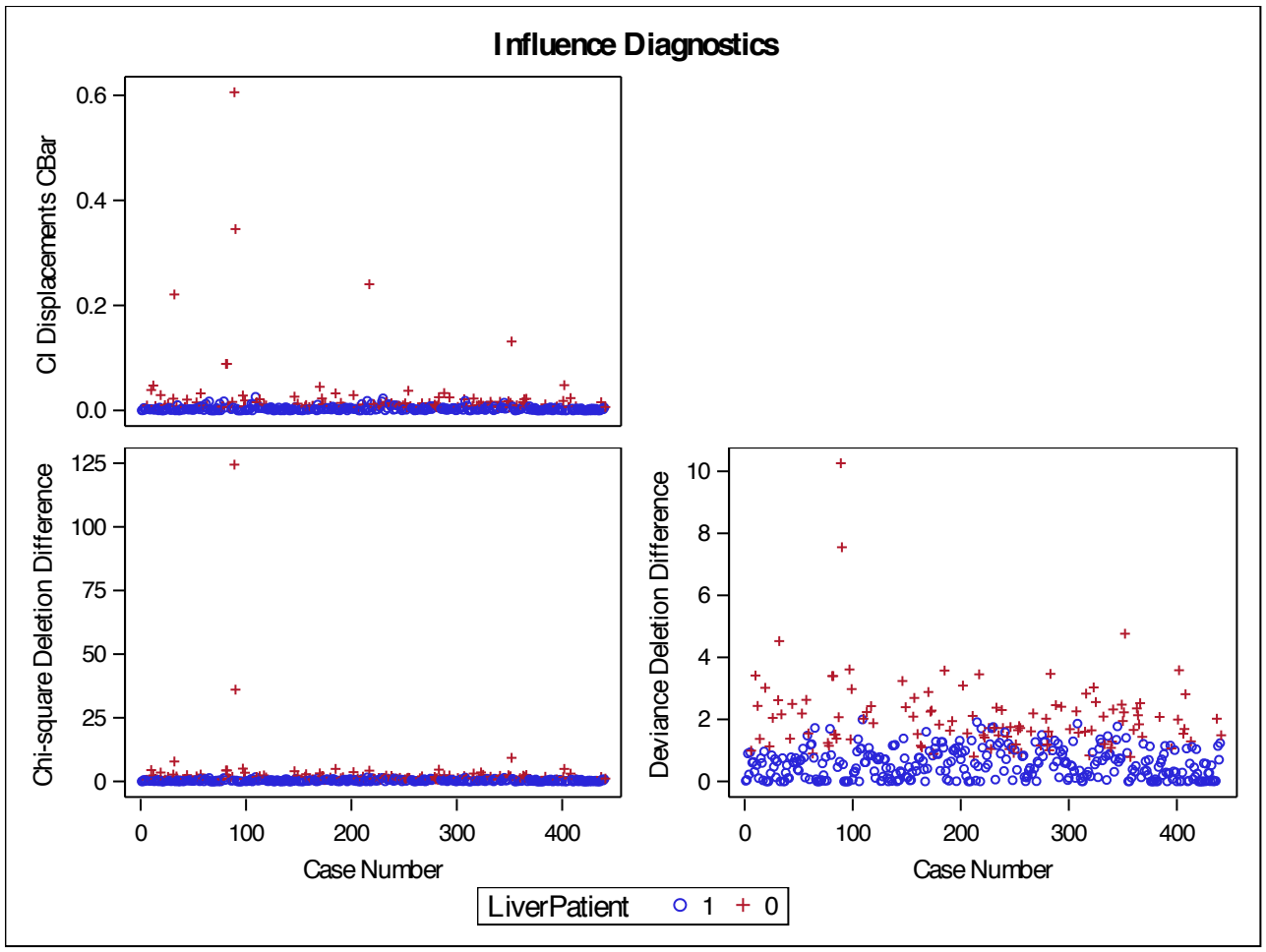
a) The process of building the logistic model is:
Step 1. We first create a new dataset with only male data. We then use backward selection to optimize the set the predictors. The selection results are shown as below. The selected predictors are: Age DB Alamine.

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	TB	1	8	0.0199	0.8879
2	Aspartate	1	7	0.1789	0.6723
3	AGRatio	1	6	1.1977	0.2738
4	Alkphos	1	5	1.0300	0.3101
5	TP	1	4	3.1623	0.0754
6	ALB	1	3	1.7323	0.1881

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	12.8939	0.0003
DB	1	10.7898	0.0010
Alamine	1	12.9962	0.0003

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.4450	0.4056	12.6919	0.0004
Age	1	0.0263	0.00731	12.8939	0.0003
DB	1	0.6136	0.1868	10.7898	0.0010
Alamine	1	0.0172	0.00477	12.9962	0.0003

Step 2. We then fit the model with the selected predictors and check the influence of each data point. The influence is shown as below. Based on CBar results, there are few points deviating a lot from the rest.



Step 3. We then print out the unduly influential data points as shown below. And reselect the predictors using backward selection on the remaining data. The selected predictors are: Age DB Alkphos Alamine.

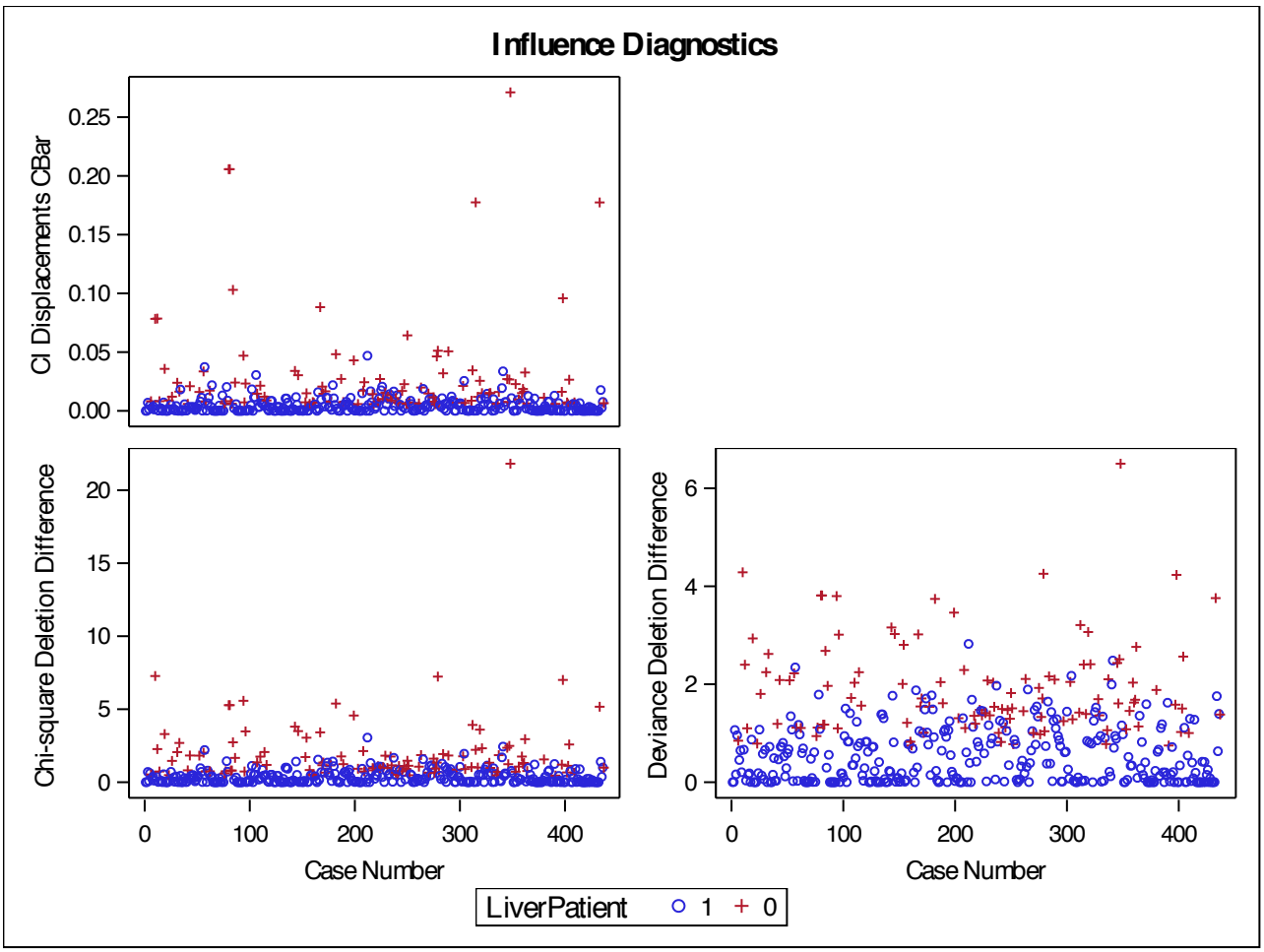
Obs	Age	Gender	TB	DB	Alkphos	Alamine	Aspartate	TP	ALB	AGRatio	LiverPatient	cbar
32	42	Male	6.8	3.2	630	25	47	6.1	2.3	0.60	0	0.22083
89	50	Male	5.8	3.0	661	181	285	5.7	2.3	0.67	0	0.60585
90	50	Male	7.3	3.6	1580	88	64	5.6	2.3	0.60	0	0.34536
217	4	Male	0.8	0.2	460	152	231	6.5	3.2	0.90	0	0.24023

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	TB	1	8	0.0002	0.9878
2	Aspartate	1	7	0.1126	0.7372
3	AGRatio	1	6	1.5197	0.2177
4	TP	1	5	1.1738	0.2786
5	ALB	1	4	1.3566	0.2441

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	13.1148	0.0003
DB	1	11.1840	0.0008
Alkphos	1	7.8276	0.0051
Alamine	1	11.6309	0.0006

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.6611	0.5589	22.6691	<.0001
Age	1	0.0280	0.00774	13.1148	0.0003
DB	1	0.8911	0.2665	11.1840	0.0008
Alkphos	1	0.00405	0.00145	7.8276	0.0051
Alamine	1	0.0211	0.00620	11.6309	0.0006

Step 4. We then refit the model and check the influence again. This time, no obvious unduly influential points.



b) The significance results and Hosmer-Lemeshow's test results are shown as below. According to AIC results, AIC for intercept and covariates is much lower than the intercept alone. This suggests that the model is significant. According to the likelihood ratio value, P value = <0.0001 suggests that at least one of these predictors are not 0. According to the maximum likelihood estimates, P value of all these 4 predictors are less than 0.05. This suggests all these 4 predictors are significant.

According to Hosmer-Lemeshow's test results, P value is 0.8380, much larger than 0.05. This suggests there is no strong evidence to reject null hypothesis. Therefore, the model fits the data well.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	501.550	388.334
SC	505.630	408.734
-2 Log L	499.550	378.334

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	121.2162	4	<.0001
Score	57.8975	4	<.0001
Wald	44.7747	4	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-2.7117	0.5589	23.5417	<.0001
Age	1	0.0287	0.00773	13.7609	0.0002
DB	1	0.8781	0.2643	11.0366	0.0009
Alkphos	1	0.00417	0.00146	8.1826	0.0042
Alamine	1	0.0212	0.00620	11.6572	0.0006

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Age	1.029	1.014	1.045
DB	2.406	1.433	4.039
Alkphos	1.004	1.001	1.007
Alamine	1.021	1.009	1.034

Partition for the Hosmer and Lemeshow Test					
		LiverPatient = 1		**LiverPatient = 0**	
Group	**Total**	**Observed**	**Expected**	**Observed**	**Expected**
1	44	15	16.96	29	27.04
2	44	24	21.65	20	22.35
3	44	27	24.75	17	19.25
4	44	27	28.25	17	15.75
5	44	33	30.85	11	13.15
6	44	31	34.78	13	9.22
7	44	39	39.25	5	4.75
8	44	43	42.64	1	1.36
9	45	45	44.87	0	0.13
10	40	40	40.00	0	0.00

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
4.2071	8	0.8380

c) The significance results for odds ratios are shown as below. First of all, 1 is not in the range of 95% confidence limits for all 4 predictors. This suggests that all 4 predictors are significant.
A unit increase in Age, DB, Alkphos, Alamine, increases the odds of being a liver patient by 1.029, 2.406, 1.004, 1.021 times, respectively.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	1.029	1.014	1.045
DB	2.406	1.433	4.039
Alkphos	1.004	1.001	1.007
Alamine	1.021	1.009	1.034

Comparison with the overall model:
1. For the overall model, more predictors are required. 6 (Age, DB, Alkphos, Alamine, TP, ALB) predictors are used for the overall model while only 4 (Age, DB, Alkphos, Alamine) are used in the male model.

2. The overall model and the male model have 4 common predictors: Age, DB, Alkphos, Alamine. And the effects of these four predictors are similar between the overall model and the male only model. In the overall model, a unit increase in Age, DB, Alkphos, Alamine, increases the odds of being a liver patient by 1.019, 2.130, 1.003, 1.021 times, respectively. In the male only model, a unit increase in Age, DB, Alkphos, Alamine, increases the odds of being a liver patient by 1.029, 2.406, 1.004, 1.021 times, respectively.

Comparison with the female only model:
1. For the male only model, more predictors are required. 4 (Age, DB, Alkphos, Alamine) are used in the male model while only 3 (Asparate, TP, ALB) are used in the female model.
2. There is no shared predictor between the female only and male only model.