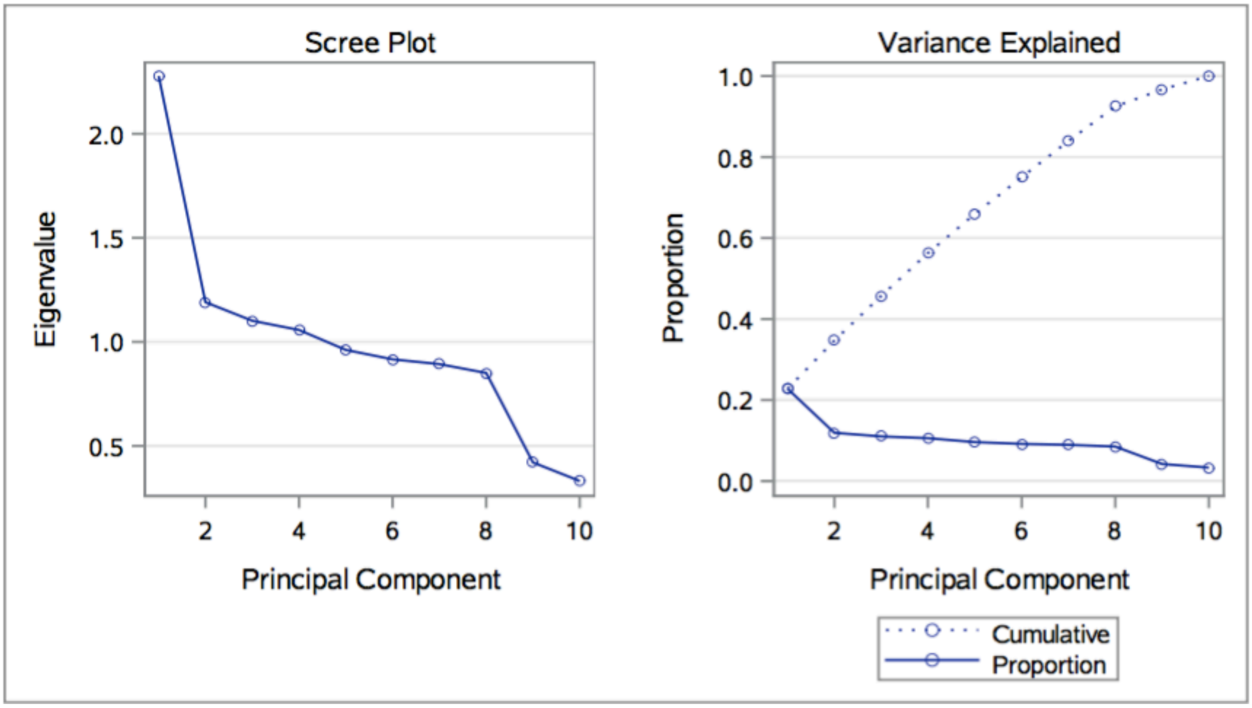
Midterm 2

Name: Zixin Ouyang

Exercise 1

Eigenvectors										
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
P8	-.166236	0.039328	-.009197	0.625540	-.294128	0.523399	-.182878	0.429786	-.006502	0.027807
P14	-.089460	-.187912	0.339155	0.472037	0.742821	-.072435	-.120778	-.160556	0.038799	0.141213
P19	0.545314	-.024147	0.076008	0.134847	-.083643	0.025923	-.157669	-.105567	0.728618	-.321766
P33	0.064589	-.515988	0.298212	-.318988	0.151838	0.465421	0.438439	0.277941	0.023017	-.170935
P37	-.547799	-.067063	0.064587	-.160692	-.108765	-.103608	0.072318	0.092574	0.651062	0.452967
P49	-.029693	0.431937	0.158808	-.467937	0.296445	0.426946	-.537417	0.086435	0.033870	0.027684
P55	0.162722	0.382532	0.438313	0.021434	0.033598	-.436718	0.181235	0.639449	-.018906	-.025742
P64	-.051561	0.519120	-.392682	0.118426	0.345019	0.240771	0.581250	-.049082	0.182651	-.094876
P70	0.574987	-.016432	-.094100	-.003945	-.000794	0.147689	0.090486	0.033741	-.051681	0.791479
P80	-.045951	0.298600	0.636881	0.081593	-.334332	0.199988	0.248893	-.522136	-.074968	0.066964

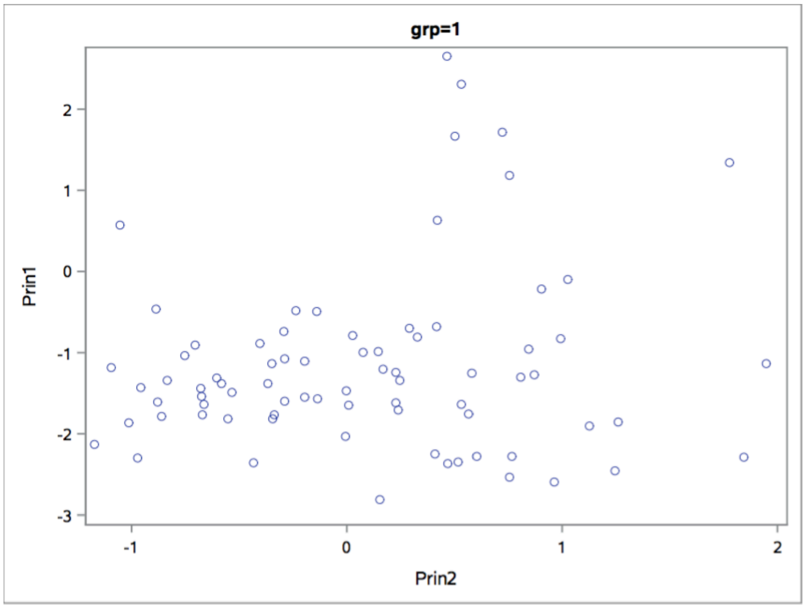
Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.27750411	1.08677128	0.2278	0.2278
2	1.19073284	0.08921380	0.1191	0.3468
3	1.10151904	0.04420932	0.1102	0.4570
4	1.05730972	0.09600690	0.1057	0.5627
5	0.96130282	0.04521848	0.0961	0.6588
6	0.91608433	0.02283080	0.0916	0.7504
7	0.89325353	0.04258669	0.0893	0.8398
8	0.85066684	0.43036599	0.0851	0.9248
9	0.42030086	0.08897494	0.0420	0.9669
10	0.33132591		0.0331	1.0000

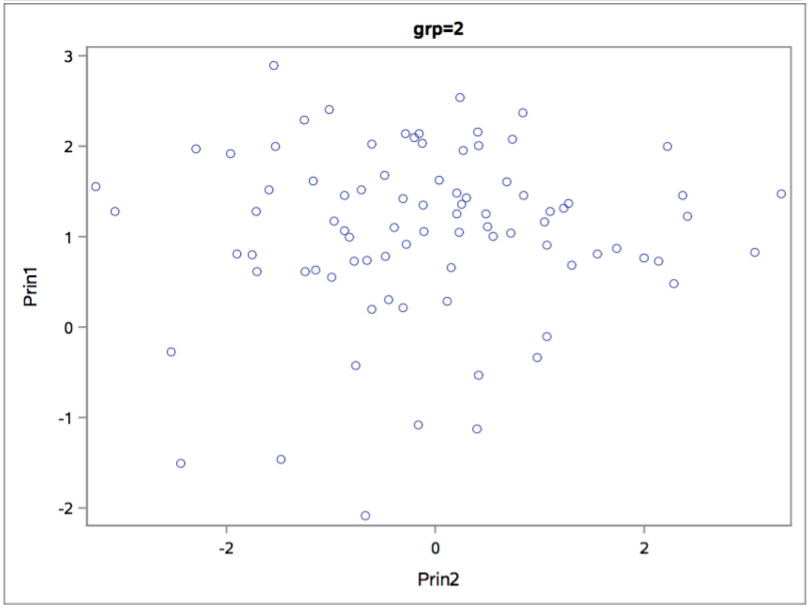


a. Firstly, in order to retain at least 50% of the total variation from the original variables, 4 principal components would need to be kept. Secondly, when we use the average eigenvalue criterion, 4 components would be kept (average eigenvalue = 1). Lastly when checking the scree plot, the “elbow” of the curve can be found at 2 and we would keep 1 component (elbow-1).

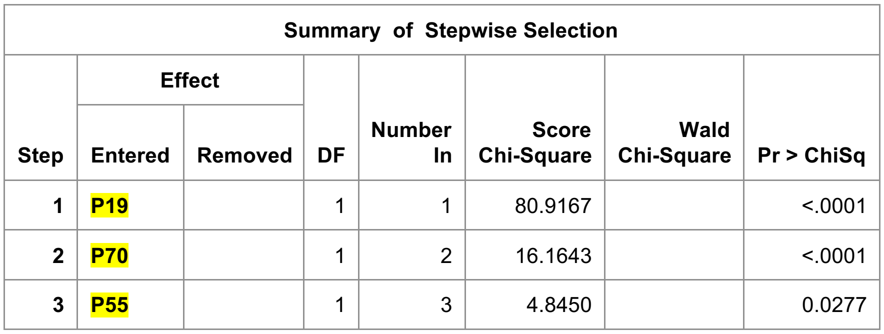
b. We proceed with the components chosen based on the 50% criterion, so we look at the first 4 components. PC1 has positive coefficients for P19 and P70 and negative coefficients for P37; PC2 has positive coefficients for P49 and P64 and negative coefficients for P33; PC3 has positive coefficients for P55 and P80; PC4 has positive coefficients for P8 and P14 and negative coefficients for P49.

c. Following are scatter plots for the first two components for group 1(healthy) and group 2(cancerous). Firstly, PC1 values of group 1(healthy) are most negative and its range is [-3, 2]. In contrast, PC1 values of group 2(cancerous) are most positive and its range is [-2, 3]. Secondly, For PC2, the range of group 1 is narrower than the range of group 2. We can say that group 2(cancerous) has larger contrasts between P49, P64 and P33.





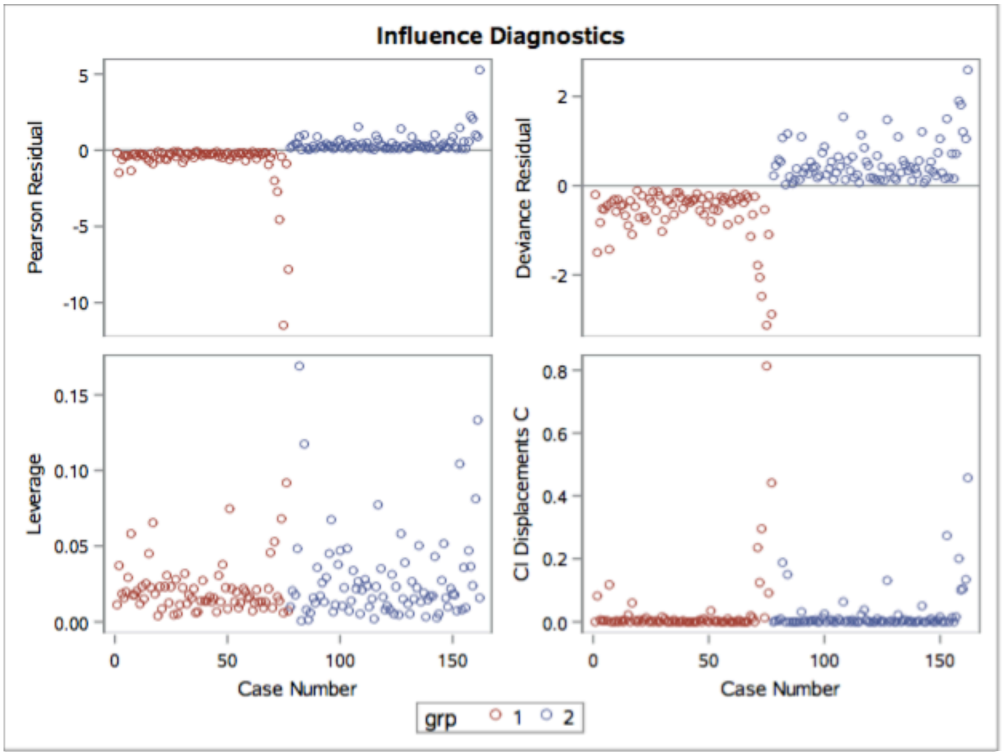
Exercise 2

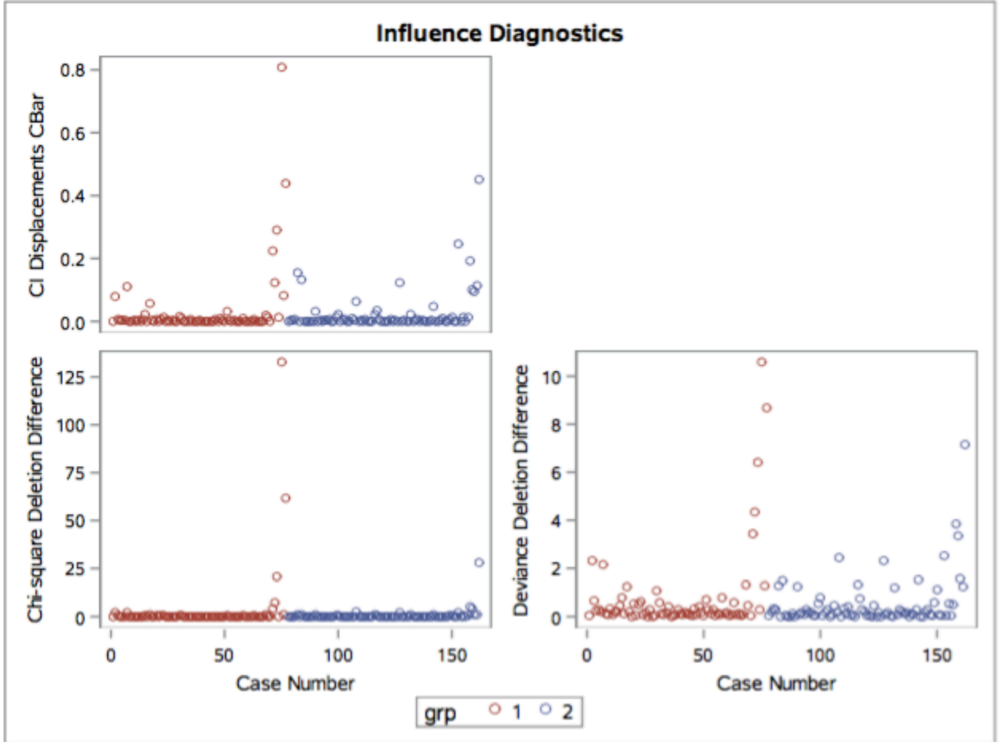


Summary of Stepwise Selection							
	Effect						
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	P19		1	1	80.9167		<.0001
2	P70		1	2	16.1643		<.0001
3	P55		1	3	4.8450		0.0277

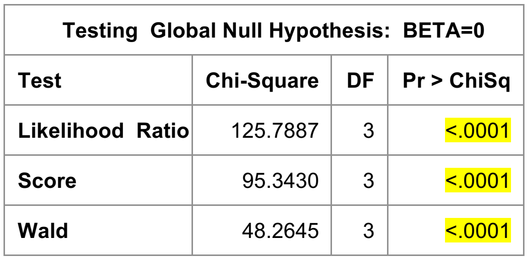
a. The predictors chosen based on the stepwise selection method are P19, P70, and P55.

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
10.4627	8	0.2340



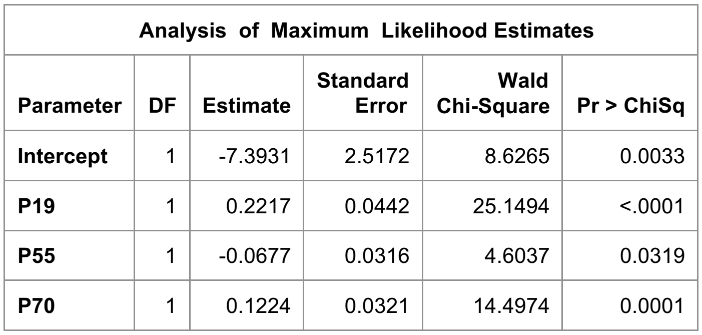


b. In diagnostic plots, we find no unduly influential points. All Cbar measures are less than 1. Also we can see that all absolute values of deviance residuals are around or less than 2 with no pattern. There is no issue on diagnostic plots. To test goodness of fit for a model we refer to the result from the Hosmer and Lemeshow test. The p- value is 0.2340 which is greater than 0.05, thus we conclude that there is no lack of fit issue and our fitted model is adequate.



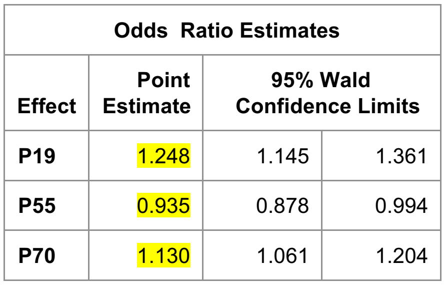
Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	125.7887	3	<.0001
Score	95.3430	3	<.0001
Wald	48.2645	3	<.0001



Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7.3931	2.5172	8.6265	0.0033
P19	1	0.2217	0.0442	25.1494	<.0001
P55	1	-0.0677	0.0316	4.6037	0.0319
P70	1	0.1224	0.0321	14.4974	0.0001



Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
P19	1.248	1.145	1.361
P55	0.935	0.878	0.994
P70	1.130	1.061	1.204

c. The global test shows p-values less than 0.05 for three kinds of asymptotic tests, Likelihood Ratio, Score and Wald test, thus we can conclude that there exists at least one predictor whose coefficient is significantly different from zero. The odds ratio of P19, P55 and P70 are 1.248, 0.935 and 1.130, respectively. We can say that one-unit increase in P19 changes the odds of having grp=2 by a multiplicative factor of 1.248 (or one-unit increase in P19 leads to 24.8% (=124.8%-100.0%) increase in odds of having grp=2). One-unit increase in P55 changes the odds of having grp=2 by a multiplicative factor of 0.935 (or one-unit increase in P55 leads to 6.5% (=100%-93.5%) decrease in odds of having grp=2). One-unit increase in P70 changes the odds of having grp=2 by a multiplicative factor of 1.13 (or one-unit increase in P70 leads to 13% (=113%-100%) increase in odds of having grp=2).

Obs	P8	P14	P19	P33	P37	P49	P55	P64	P70	P80	grp	id	_LEVEL_	pred
1	12.49	24.73	29.65	49.89	47.98	62.11	93.42	93.35	26.97	102.71	1	1	2	0.02090
2	21.70	21.05	35.01	50.76	40.50	65.59	78.67	88.94	46.54	92.25	1	2	2	0.67602
3	15.53	14.78	35.31	49.32	47.47	72.40	82.64	94.45	34.72	104.90	1	3	2	0.28634
4	17.65	23.51	31.71	49.72	62.91	69.46	77.56	97.51	29.72	101.46	1	4	2	0.12141
5	23.63	18.31	31.13	51.01	54.72	56.55	84.47	86.03	35.18	105.45	1	5	2	0.12924
6	19.63	12.86	24.82	49.64	57.99	59.66	75.11	88.78	38.86	97.43	1	6	2	0.09774
7	19.33	8.45	38.80	50.05	69.26	67.28	67.73	92.04	32.26	96.12	1	7	2	0.63868
8	19.48	26.56	33.48	48.25	65.94	59.75	84.58	89.24	24.91	100.52	1	8	2	0.06594
9	25.55	21.07	32.56	50.26	42.81	76.13	91.15	86.40	27.37	97.66	1	9	2	0.04748
10	19.25	7.78	33.11	49.90	65.33	70.53	75.39	84.12	28.18	97.78	1	10	2	0.15314

Frequency	Table of grp by _INTO_		
	INTO(Formatted Value of the Predicted Response)		
grp	1	2	Total
1	70	7	77
2	8	77	85
Total	78	84	162

d. If the predicted probability is larger than 0.5, the patient is classified as cancerous. Therefore, patient 1, 3, 4, 5, 6, 8, 9, 10 are healthy, patient 2 and patient 7 is cancerous. The total number of misclassified observations is 7+8=15. The misclassification error is (7+8)/162 =0.0926. It means that 9.26% of the total number of observations is misclassified. The separation performance of the logistic regression model is good.

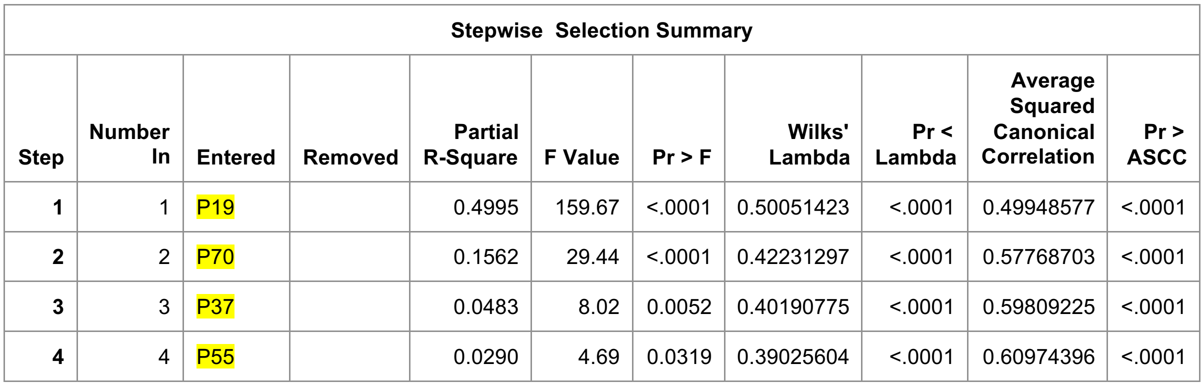
Exercise 3

Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
310.928330	55	<.0001

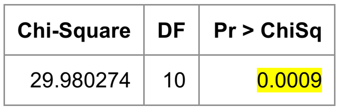
Multivariate Statistics and Exact F Statistics					
S=1 M=4 N=74.5					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.36693488	26.05	10	151	<.0001
Pillai's Trace	0.63306512	26.05	10	151	<.0001
Hotelling-Lawley Trace	1.72527918	26.05	10	151	<.0001
Roy's Greatest Root	1.72527918	26.05	10	151	<.0001

a. The p-value for the test of homogeneity of within covariance is less than 0.1. Thus we can conclude that two groups have the different covariance and quadratic discriminant analysis needs to be implemented. The MANOVA tests show p-values less than 0.05, thus we can conclude that there are significant differences in some attributes between healthy and cancerous group. This implies that discrimination between groups based on these variables should be a reasonable approach and provide some separation between groups.

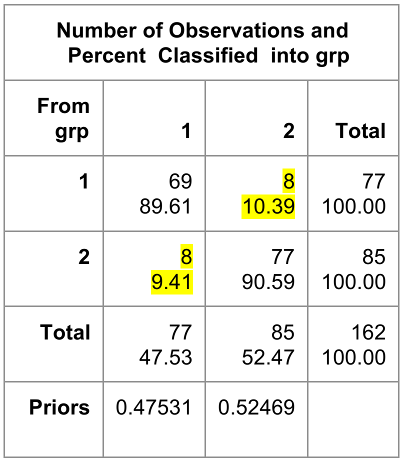


Stepwise Selection Summary										
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	P19		0.4995	159.67	<.0001	0.50051423	<.0001	0.49948577	<.0001
2	2	P70		0.1562	29.44	<.0001	0.42231297	<.0001	0.57768703	<.0001
3	3	P37		0.0483	8.02	0.0052	0.40190775	<.0001	0.59809225	<.0001
4	4	P55		0.0290	4.69	0.0319	0.39025604	<.0001	0.60974396	<.0001

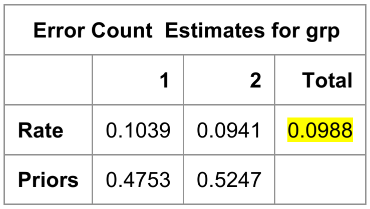
Test of Homogeneity of Within Covariance Matrices



Chi-Square	DF	Pr > ChiSq
29.980274	10	0.0009



Number of Observations and Percent Classified into grp			
From grp	1	2	Total
1	69 89.61	8 10.39	77 100.00
2	8 9.41	77 90.59	85 100.00
Total	77 47.53	85 52.47	162 100.00
Priors	0.47531	0.52469	



Error Count Estimates for grp			
	1	2	Total
Rate	0.1039	0.0941	0.0988
Priors	0.4753	0.5247	

b. Based on the stepwise discrimination procedure, the predictors selected are P19, P70, P37 and P55. The p-value for the test of homogeneity of within covariance is less than 0.1. Thus we can conclude that two groups have the different covariance and quadratic discriminant analysis needs to be implemented. The cross-validation estimated overall error rate is 0.0988 based on proportional-prior discriminant analysis and the individual group error rate estimates are 10.39% and 9.41% respectively. It is easy to misclassify data point in healthy group compared to cancerous group. The total number of misclassified observations is 8+8=16.

Observation Profile for Test Data	
Number of Observations Read	50
Number of Observations Used	50

Number of Observations and Percent Classified into grp

From grp	1	2	Total
1	22 95.65	1 4.35	23 100.00
2	5 18.52	22 81.48	27 100.00
Total	27 54.00	23 46.00	50 100.00
Priors	0.48214	0.51786	

Error Count Estimates for grp

	1	2	Total
Rate	0.0435	0.1852	0.1169
Priors	0.4821	0.5179	

c. The cross-validation estimated overall error rate is 0.1169 based on proportional-prior discriminant analysis and the individual group error rate estimates are 4.35% and 18.52% respectively. It is easy to misclassify data point in cancerous group compared to healthy group. The total number of misclassified observations is 1+5=6.

Exercise 4

Based on the results from exercises 2-3, logistic regression best classifies cancerous vs healthy population in this particular dataset. According to the logistic regression in exercise 2, the total number of misclassified observations is 7+8=15. The misclassification error 0.0926. Based on the quadratic discriminant analysis in exercise 3b, the total number of misclassified observations is 8+8=16, and the cross-validation estimated overall error rate is 0.0988. Based on the quadratic discriminant analysis in exercise 3c, the cross-validation estimated overall error rate is 0.1169.

Exercise 5

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	492	18.3157	0.0372
Scaled Deviance	492	509.0340	1.0346
Pearson Chi-Square	492	20.2702	0.0412
Scaled Pearson X2	492	563.3557	1.1450
Log Likelihood		-1415.3170	
Full Log Likelihood		-1415.3170	
AIC (smaller is better)		2860.6340	
AICC (smaller is better)		2861.6135	
BIC (smaller is better)		2924.0320	

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
crim	1	49.60	<.0001
zn	1	5.83	0.0157
indus	1	0.80	0.3722
chas	1	10.29	0.0013
nox	1	30.67	<.0001
rm	1	28.03	<.0001
age	1	0.21	0.6429
dis	1	45.51	<.0001
rad	1	31.57	<.0001
tax	1	15.54	<.0001
ptratio	1	53.38	<.0001
b	1	13.96	0.0002
lstat	1	192.68	<.0001

a. From the Type 3 Analysis table, we can see crim, an, chas, nox, rm, dis, rad, tax, ptratio, b and lstat are significant predictors with p-values less than 0.05, while indus and age are not significant in the model.

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	493	18.3234	0.0372

Scaled Deviance	493	509.0353	1.0325
Pearson Chi-Square	493	20.2842	0.0411
Scaled Pearson X2	493	563.5070	1.1430
Log Likelihood		-1415.4245	
Full Log Likelihood		-1415.4245	
AIC (smaller is better)		2858.8489	
AICC (smaller is better)		2859.7043	
BIC (smaller is better)		2918.0205	

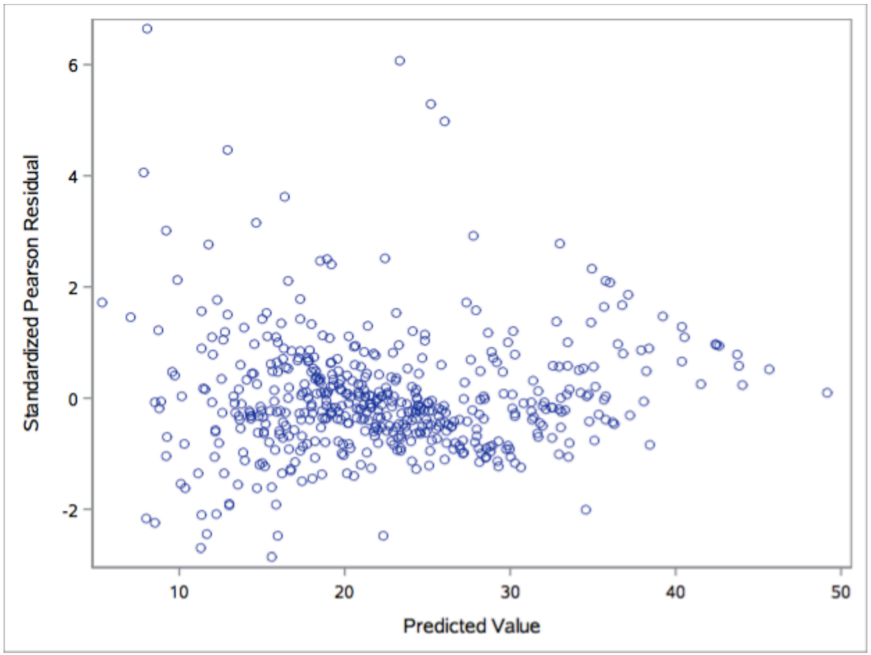
LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
crim	1	49.52	<.0001
zn	1	5.65	0.0175
indus	1	0.81	0.3694
chas	1	10.53	0.0012
nox	1	31.40	<.0001
rm	1	29.85	<.0001
dis	1	51.62	<.0001
rad	1	31.38	<.0001
tax	1	15.44	<.0001
ptratio	1	53.17	<.0001
b	1	14.10	0.0002
lstat	1	204.20	<.0001

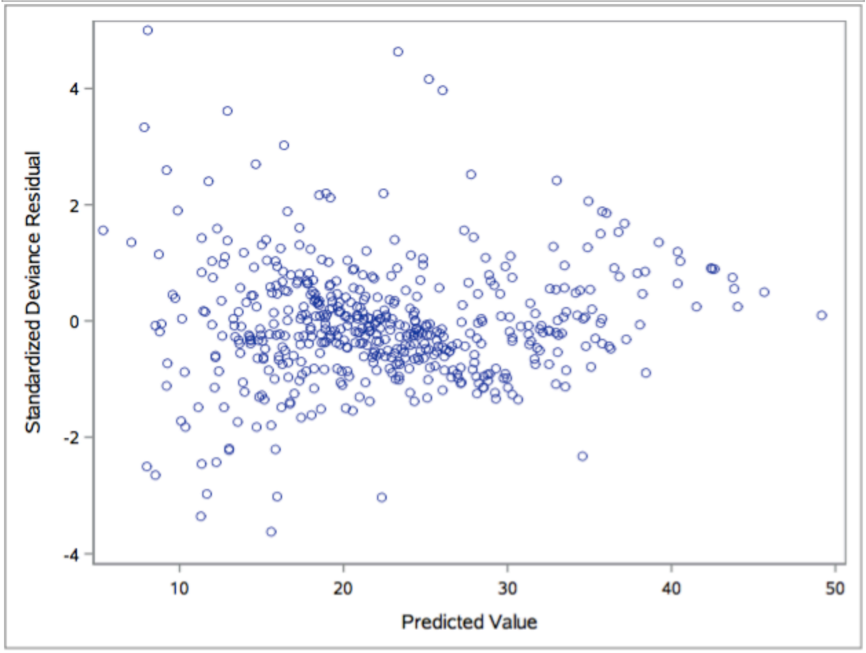
Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	494	18.3524	0.0372
Scaled Deviance	494	509.0401	1.0304
Pearson Chi-Square	494	20.3011	0.0411
Scaled Pearson X2	494	563.0906	1.1399
Log Likelihood		-1415.8274	
Full Log Likelihood		-1415.8274	
AIC (smaller is better)		2857.6547	
AICC (smaller is better)		2858.3946	
BIC (smaller is better)		2912.5997	

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
crim	1	49.87	<.0001
zn	1	5.23	0.0222
chas	1	11.25	0.0008
nox	1	31.11	<.0001
rm	1	29.21	<.0001
dis	1	56.62	<.0001
rad	1	31.06	<.0001
tax	1	15.42	<.0001
ptratio	1	52.37	<.0001
b	1	13.90	0.0002
lstat	1	203.51	<.0001

b. The gamma model including all predictors shows AIC as 2860.634 and type3 analysis gives age as an insignificant predictor with p-value of 0.6429. Thus we first remove age and fit the gamma model again. After removing age, the AIC is 2858.8489, and type3 analysis gives indus as an insignificant predictor with p-value of 0.3694. So we remove indus in the second step and refit the gamma model. Now the AIC is 2857.6547, and all p-values of predictor are less than 0.05. The final model contains crim, an, chas, nox, rm, dis, rad, tax, ptratio, b and lstat.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	4.1879	0.1978	3.8003	4.5755	448.45	<.0001
crim	1	-0.0100	0.0013	-0.0125	-0.0075	60.46	<.0001
zn	1	0.0012	0.0005	0.0002	0.0023	5.22	0.0224
chas	1	0.1135	0.0342	0.0466	0.1805	11.05	0.0009
nox	1	-0.7942	0.1391	-1.0669	-0.5215	32.59	<.0001
rm	1	0.0848	0.0154	0.0546	0.1150	30.26	<.0001
dis	1	-0.0562	0.0072	-0.0703	-0.0421	60.68	<.0001
rad	1	0.0145	0.0025	0.0095	0.0194	32.87	<.0001
tax	1	-0.0005	0.0001	-0.0008	-0.0003	15.94	<.0001
ptratio	1	-0.0384	0.0052	-0.0485	-0.0282	55.22	<.0001
b	1	0.0004	0.0001	0.0002	0.0006	14.34	0.0002
lstat	1	-0.0282	0.0018	-0.0317	-0.0247	250.66	<.0001
Scale	1	27.7369	1.7334	24.5393	31.3512		





c. From the residual plots (standardized deviance and standardized pearson residuals), we see no pattern and no observation with large value. Thus our model assumptions are adequate to the data. In terms of parameter estimates, crim, nox, dis, tax, ptratio and lstat have negative relationships with respect to log(medv). The estimates are -0.01, -0.7942, -0.0562, -0.0005, -0.0384 and -0.0282. It implies for one unit increase of crim, nox, dis, tax, ptratio and lstat, we expect a multiplicative change in median home values of e^-0.01, e^-0.7942, e^--0.0562, e^--0.0005, e^--0.0384 and e^--0.0282. Predictors zn, chas, rm, rad and b have positive relationships with respect to log(medv). The estimates are 0.0012, 0.1135, 0.0848, 0.0145 and 0.0004. It implies for one unit increase of zn, chas, rm, rad and b, we expect a multiplicative change in median home values of e^0.0012, e^0.1135, e^0.0848, e^0.0145 and e^0.0004.

Exercise 6

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	55	147.0216	2.6731
Scaled Deviance	55	147.0216	2.6731
Pearson Chi-Square	55	136.6408	2.4844
Scaled Pearson X2	55	136.6408	2.4844
Log Likelihood		590.6875	
Full Log Likelihood		-167.3950	
AIC (smaller is better)		342.7900	
AICC (smaller is better)		343.5307	
BIC (smaller is better)		351.1002	

Model Information

Data Set	WORK.EPI
Distribution	Poisson
Link Function	Log
Dependent Variable	Period4

Criteria For Assessing Goodness Of Fit

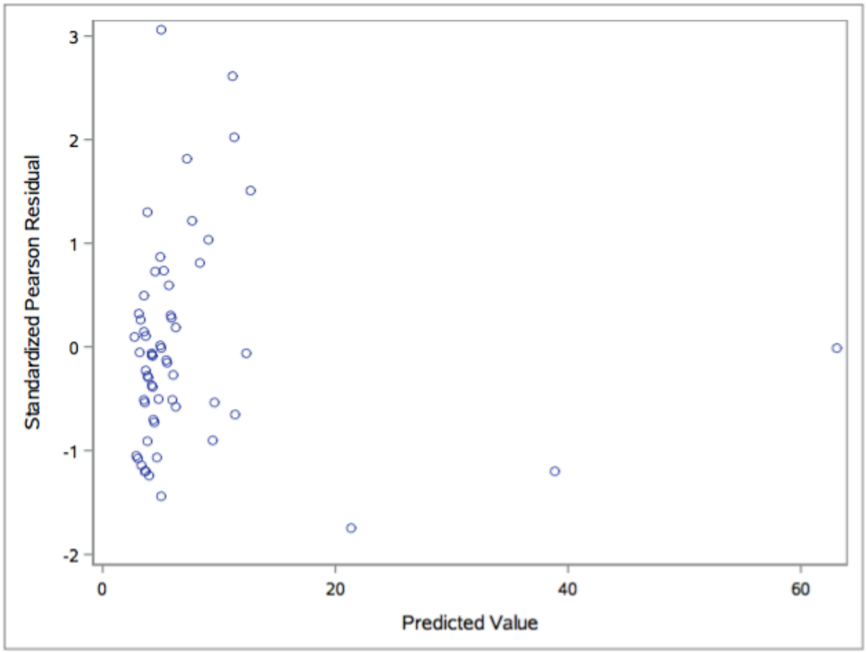
Criterion	DF	Value	Value/DF
Deviance	55	147.0216	2.6731
Scaled Deviance	55	55.0000	1.0000
Pearson Chi-Square	55	136.6408	2.4844
Scaled Pearson X2	55	51.1166	0.9294
Log Likelihood		220.9730	
Full Log Likelihood		-167.3950	
AIC (smaller is better)		342.7900	
AICC (smaller is better)		343.5307	
BIC (smaller is better)		351.1002	

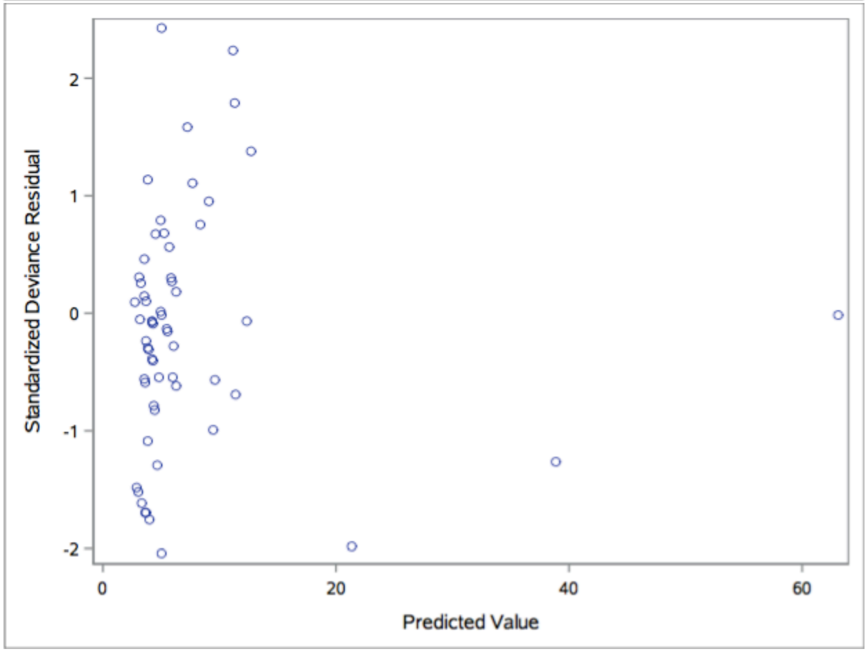
Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.7756	0.4653	-0.1364	1.6876	2.78	0.0956
Treat	1	-0.2705	0.1666	-0.5969	0.0560	2.64	0.1044
BL	1	0.0221	0.0018	0.0186	0.0255	153.66	<.0001
Age	1	0.0140	0.0140	-0.0135	0.0415	1.00	0.3168
Scale	0	1.6350	0.0000	1.6350	1.6350		

LR Statistics For Type 1 Analysis							
Source	Deviance	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
Intercept	476.2487						
Treat	473.0840	1	55	1.18	0.2813	1.18	0.2766
BL	149.6763	1	55	120.99	<.0001	120.99	<.0001
Age	147.0216	1	55	0.99	0.3233	0.99	0.3190

LR Statistics For Type 3 Analysis						
Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
Treat	1	55	2.65	0.1093	2.65	0.1036
BL	1	55	119.99	<.0001	119.99	<.0001
Age	1	55	0.99	0.3233	0.99	0.3190

a. In the above model we detect a potential problem with over-dispersion since the scale factor, e.g., Value/DF, is greater than 1. Therefore, we need to use over- dispersion with deviance scale. Both the type 1 analysis and the type 3 analysis tell us that only the predictor BL is significant with p-value less that 0.05, Treat and Age are insignificant. But we do need treatment in a model. A statistician would want to remove Treat and Age from the model, but do not remove those terms from the model.





b. According to the plots above, it seems there is a slight upward trend which would show a slight tendency to over-predict at the low end and under-predict at the higher end. The baseline parameter estimate is 0.0221 indicating that the seizure count after four periods of treatment increases slightly as baseline count increases. For an increase of one in the baseline count, we would expect the count after 4 periods to be multiplied by *e*^.0.0221. The Treatment parameter estimate is -0.2705 indicating that the seizure count after four periods of treatment progabide increases slightly as placebo. For progabide treatment instead of placebo, we would expect the count after 4 periods to be multiplied by *e*^.-0.2705.