STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 1a

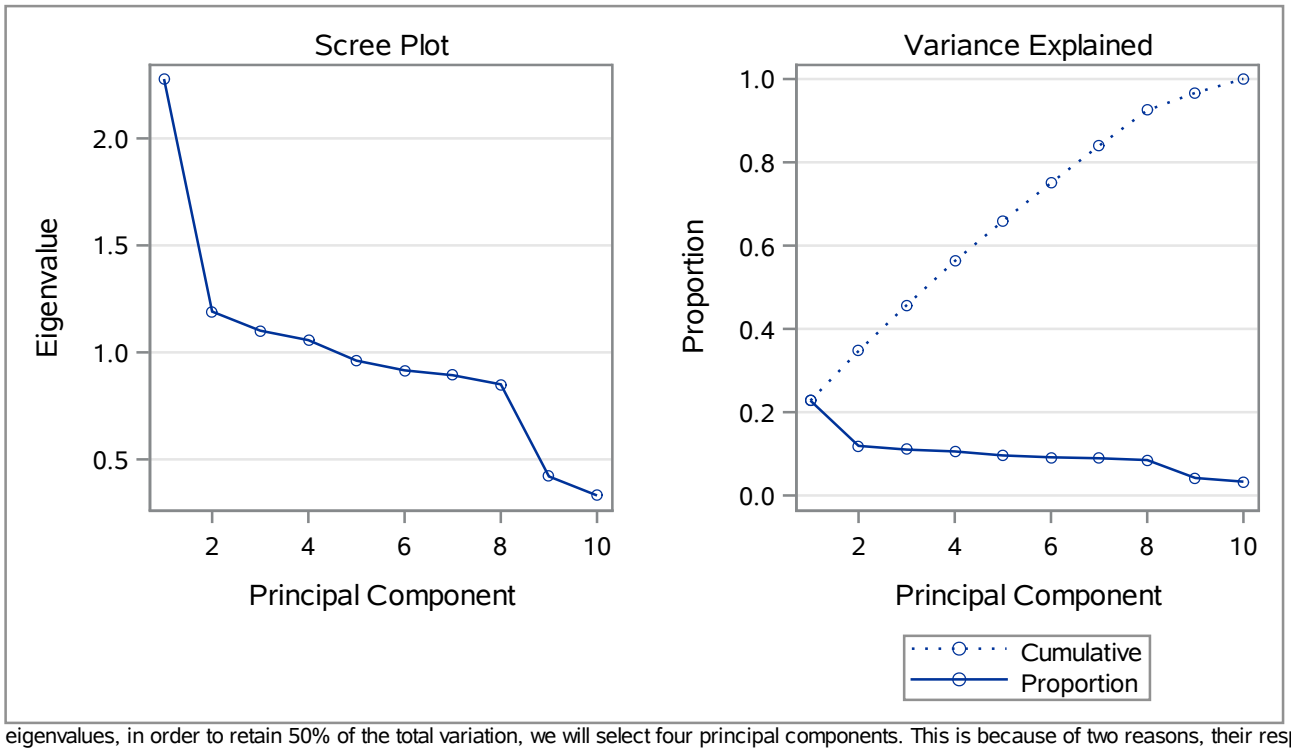
The PRINCOMP Procedure

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

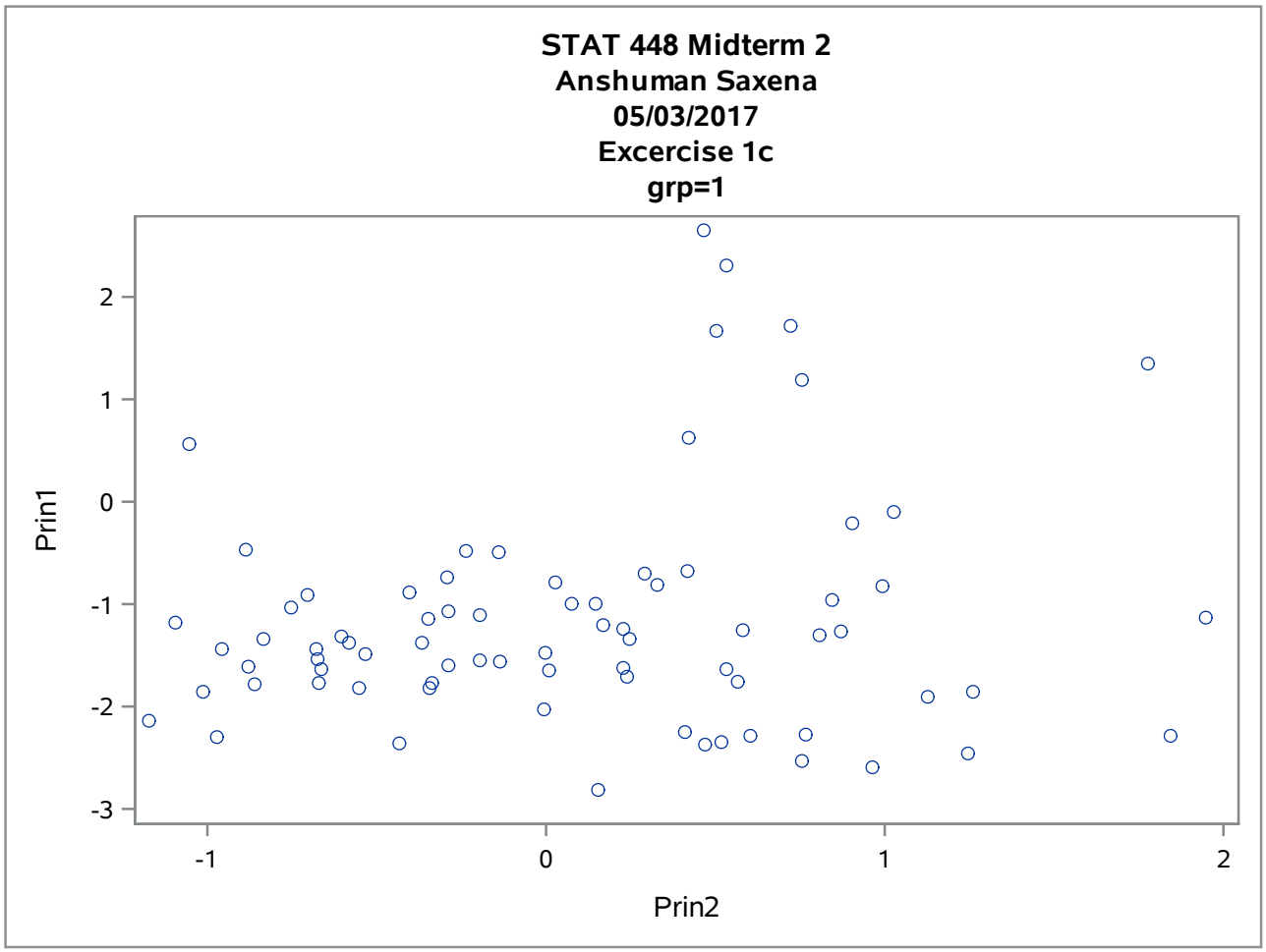
Eigenvectors										
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
P8	-.166236	0.039328	-.009197	0.625540	-.294128	0.523399	-.182878	0.429786	-.006502	0.027807
P14	-.089460	-.187912	0.339155	0.472037	0.742821	-.072435	-.120778	-.160556	0.038799	0.141213
P19	0.545314	-.024147	0.076008	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

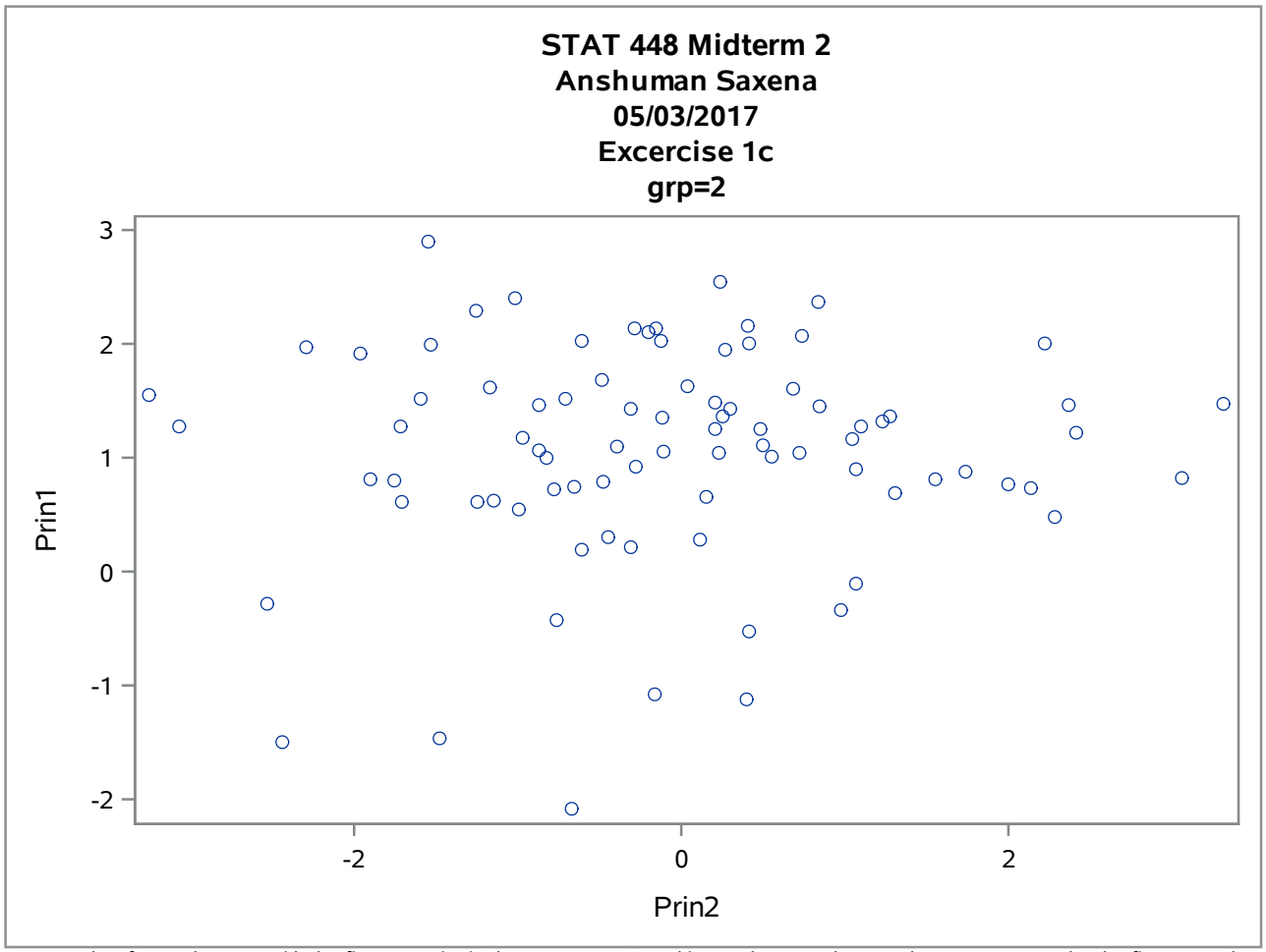
STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 1a

The PRINCOMP Procedure



Based on the eigenvalues, in order to retain 50% of the total variation, we will select four principal components. This is because of two reasons, their respective eigenvalues are over 1 and their cumulative variation is over 50%. Looking at the scree plots we can see the elbow to be at components 2 and 9. So 1 or 8 components.
We continue with taking four principal components. The first principal component has positive coefficients with P19 and P70 and a negative coefficient with P37. The second principal component has positive coefficients with P49, P55 and P64 and negative coefficient with P33. The third principal component has positive coefficients with P14, P55 and P80 and negative coefficient with P64. The fourth principal compenent has positive coefficients with P8 and P14 and negative coefficients with P33 and P49.


STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 1c
grp=1
Prin1
2
1
0
-1
-2
-3
-1
0
1
2
Prin2



Above are the scatter plots for each group with the first two principal components. By taking a glance at the two plots, we can see the the first group has values mostly towards the bottom half of the scatter plot, whereas the second group has values towards the top half of the scatter plot. The ranges for principal component 1 in both the graphs are pretty similar; (-3, 2) for group 1 and (-2, 3) for group 2. For principal component 2, the range for group 2 is wider than group 1; (-1.1, 2) for group 1 and (-3, 3) for group 2.

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 2a

The LOGISTIC Procedure

Model Information	
Data Set	WORK.CANCER
Response Variable	cancerous
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	162
Number of Observations Used	162

Response Profile		
Ordered Value	**cancerous**	**Total Frequency**
1	0	77
2	1	85

Probability modeled is cancerous=0.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	226.184	107.728
SC	229.272	141.692
-2 Log L	224.184	85.728

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	138.4563	10	<.0001
Score	102.5565	10	<.0001
Wald	48.3136	10	<.0001

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 2a

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-10.8120	9.5981	1.2689	0.2600
P8	1	-0.0296	0.0637	0.2154	0.6426
P14	1	-0.0241	0.0320	0.5662	0.4518
P19	1	-0.1879	0.0435	18.6801	<.0001
P33	1	0.0737	0.0402	3.3592	0.0668
P37	1	0.0478	0.0327	2.1375	0.1437
P49	1	0.0662	0.0405	2.6683	0.1024
P55	1	0.0831	0.0355	5.4911	0.0191
P64	1	0.0702	0.0381	3.3923	0.0655
P70	1	-0.1404	0.0479	8.6096	0.0033
P80	1	0.00397	0.0685	0.0034	0.9538

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
P8	0.971	0.857	1.100
P14	0.976	0.917	1.039
P19	0.829	0.761	0.902
P33	1.076	0.995	1.165
P37	1.049	0.984	1.118
P49	1.068	0.987	1.157
P55	1.087	1.014	1.165
P64	1.073	0.996	1.156
P70	0.869	0.791	0.954
P80	1.004	0.878	1.148

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	94.9	Somers' D	0.898
Percent Discordant	5.1	Gamma	0.898
Percent Tied	0.0	Tau-a	0.451
Pairs	6545	c	0.949

Above is a logistic regression model of the classification variable 'grp' on attributes of protein biomarkers. The global H0 tests shows all tests to have a p-value <0.05, hence we can conclude that there is at least one predictor who's coefficient is significantly different from 0. Significant predictors, with p-values <0.05 are P19, P55 and P70. The rest are insignificant at the 5% level and should be dropped from the model.

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 2a

The LOGISTIC Procedure

Summary of Stepwise Selection

Step	Effect Entered	Effect 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

The best model chosen is logit(cancerous=1)~P19 + P70 + P55.

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 2b

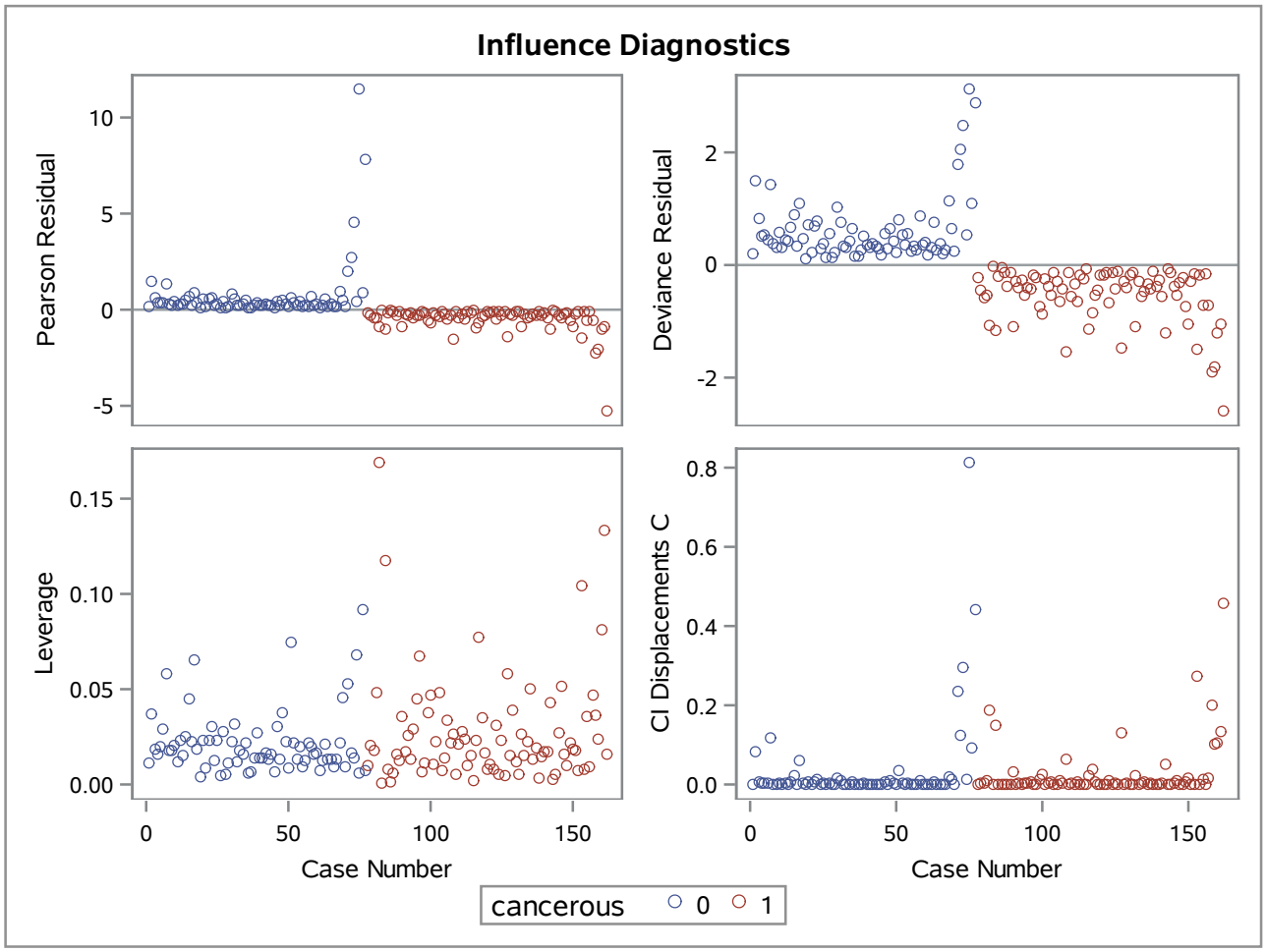
The LOGISTIC Procedure

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
P70	1	-0.1224	0.0321	14.4974	0.0001
P55	1	0.0677	0.0316	4.6037	0.0319

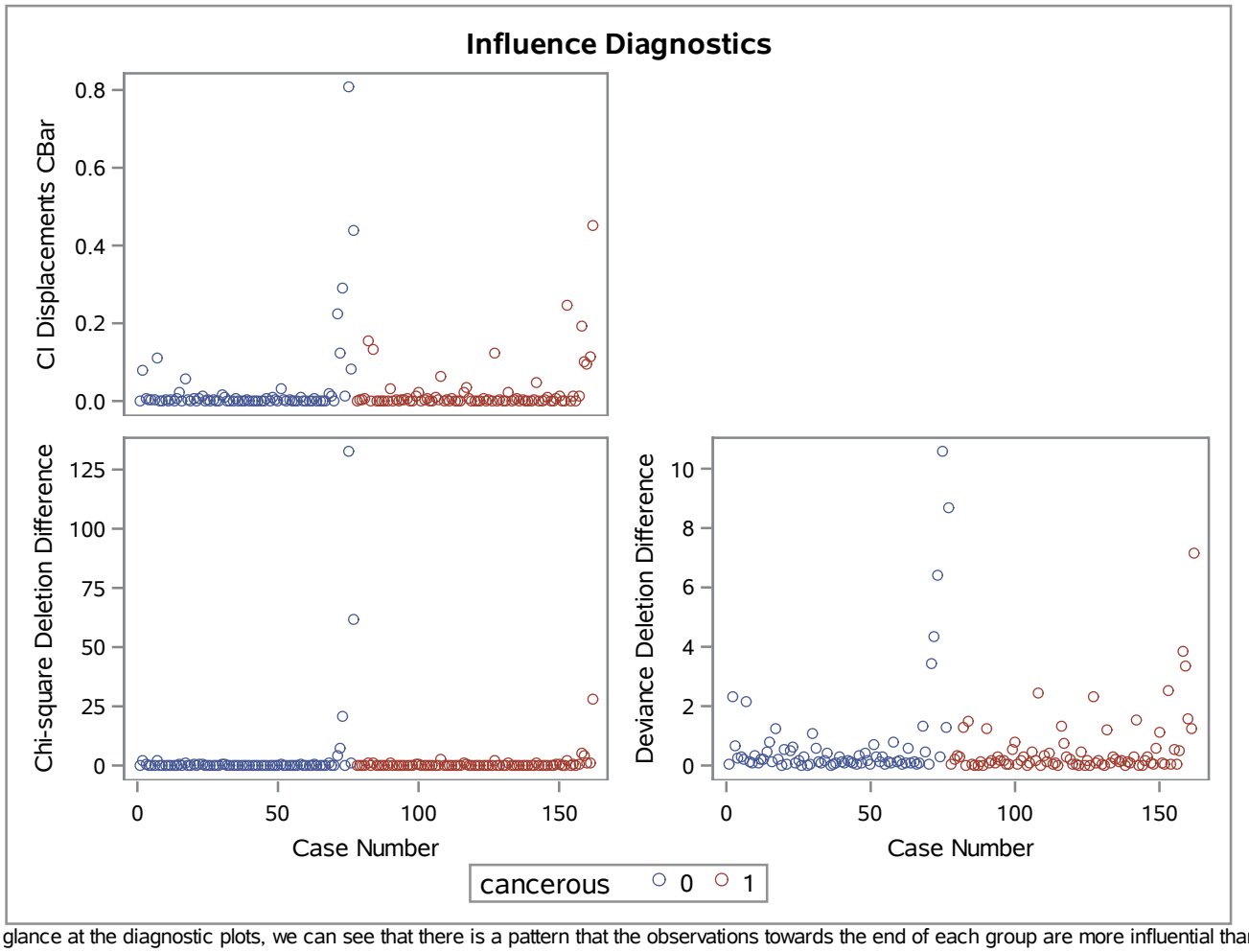
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
P19	0.801	0.735	0.874
P70	0.885	0.831	0.942
P55	1.070	1.006	1.138

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
14.8040	8	0.0631

The LOGISTIC Procedure



The LOGISTIC Procedure



After a quick glance at the diagnostic plots, we can see that there is a pattern that the observations towards the end of each group are more influential than the other observations. They are poorly accounted for and are causing instability in the model; observation 75 can be considered an extreme point. We do not omit any observations for being unduly influential since the CBar doesn't have a value greater than 1.The Hosmer-Lemeshow Goodness of Fit test is greater than 0.05, meaning we fail to reject the H0. Hence, we conclude that there isn't a lack of fit issue and the model is adequate.

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 2c

The LOGISTIC Procedure

Model Information	
Data Set	WORK.CANCER
Response Variable	cancerous
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	162
Number of Observations Used	162

Response Profile		
Ordered Value	**cancerous**	**Total Frequency**
1	0	77
2	1	85

Probability modeled is cancerous=0.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	226.184	106.396
SC	229.272	118.746
-2 Log L	224.184	98.396

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

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 2c

The LOGISTIC Procedure

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
P70	1	-0.1224	0.0321	14.4974	0.0001
P55	1	0.0677	0.0316	4.6037	0.0319

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
P19	0.801	0.735	0.874
P70	0.885	0.831	0.942
P55	1.070	1.006	1.138

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	94.2	Somers' D	0.884
Percent Discordant	5.8	Gamma	0.884
Percent Tied	0.0	Tau-a	0.444
Pairs	6545	c	0.942

Above is a logistic regression model of the classification variable 'grp' on the chosen attributes of protein biomarkers through stepwise selection (P19, P70 and P55). The global H0 tests shows all tests to have a p-value <0.05, hence we can conclude that there is at least one predictor who's coefficient is significantly different from 0. The parameter estimates show that all variables are significant. The odds ratio are very close to 1 (~0.8), indicating that if one has that protein biomarker, having cancer probability increases by ~0.8 - 1.

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 2d

The LOGISTIC Procedure

Model Information	
Data Set	WORK.CANCER
Response Variable	cancerous
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	162
Number of Observations Used	162

Response Profile		
Ordered Value	**cancerous**	**Total Frequency**
1	0	77
2	1	85

Probability modeled is cancerous=0.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	226.184	106.396
SC	229.272	118.746
-2 Log L	224.184	98.396

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

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 2d

The LOGISTIC Procedure

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
P70	1	-0.1224	0.0321	14.4974	0.0001
P55	1	0.0677	0.0316	4.6037	0.0319

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
P19	0.801	0.735	0.874
P70	0.885	0.831	0.942
P55	1.070	1.006	1.138

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	94.2	Somers' D	0.884
Percent Discordant	5.8	Gamma	0.884
Percent Tied	0.0	Tau-a	0.444
Pairs	6545	c	0.942

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 2d

The FREQ Procedure

Frequency	Table of cancerous by _INTO_		
	INTO(Formatted Value of the Predicted Response)		
cancerous	0	1	Total
0	70	7	77
1	8	77	85
Total	78	84	162

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 2d

Obs	P8	P14	P19	P33	P37	P49	P55	P64	P70	P80	grp	id	cancerous	_FROM_	_INTO_	IP_0	IP_1
1	12.49	24.73	29.65	49.89	47.98	62.11	93.42	93.35	26.97	102.71	1	1	0	0	0	0.97910	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	0	0	1	0.32398	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	0	0	0	0.71366	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	0	0	0	0.87859	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	0	0	0	0.87076	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	0	0	0	0.90226	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	0	0	1	0.36132	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	0	0	0	0.93406	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	0	0	0	0.95252	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	0	0	0	0.84686	0.15314

15 out of 162 observations are misclassified, which 9.26% of the observations.

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3a

The DISCRIM Procedure

Total Sample Size	162	**DF Total**	161
Variables	10	**DF Within Classes**	160
Classes	2	**DF Between Classes**	1

Number of Observations Read	162
Number of Observations Used	162

Class Level Information

grp	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	1	77	77.0000	0.475309	0.475309
2	2	85	85.0000	0.524691	0.524691

Within Covariance Matrix Information

grp	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
1	10	34.04216
2	10	39.42889
Pooled	10	38.94920

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3a

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

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

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.

Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3a

The DISCRIM Procedure

Generalized Squared Distance to grp		
From grp	1	2
1	35.52974	48.88079
2	43.61754	40.71878

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3a

The DISCRIM Procedure

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

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3a

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CANCER
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into grp			
From grp	1	2	Total
1	70 90.91	7 9.09	77 100.00
2	7 8.24	78 91.76	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	0.0909	0.0824	0.0864
Priors	0.4753	0.5247	

The Test of Homogeneity of Within Covariance Matrix is significant (<0.05), hence we reject H0 and conclude that the two groups do not have the same covariance and quadratic discriminant analysis (QDA) needs to be implemented. Based on the MANOVA tests we can conclude that there is a linear relationship between 'grp' and the continuous variables. The null was rejected by all tests, hence the canonical correlations are not zero.

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3b

The STEPDISC Procedure

The Method for Selecting Variables is STEPWISE			
Total Sample Size	162	**Variable(s) in the Analysis**	10
Class Levels	2	**Variable(s) Will Be Included**	0
		Significance Level to Enter	0.05
		Significance Level to Stay	0.05

Number of Observations Read	162
Number of Observations Used	162

Class Level Information				
grp	**Variable Name**	**Frequency**	**Weight**	**Proportion**
1	1	77	77.0000	0.475309
2	2	85	85.0000	0.524691

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3b

The STEPDISC Procedure
Stepwise Selection: Step 1

Statistics for Entry, DF = 1, 160				
Variable	**R-Square**	**F Value**	**Pr > F**	**Tolerance**
P8	0.0101	1.64	0.2025	1.0000
P14	0.0016	0.26	0.6115	1.0000
P19	0.4995	159.67	<.0001	1.0000
P33	0.0025	0.40	0.5280	1.0000
P37	0.3623	90.90	<.0001	1.0000
P49	0.0173	2.82	0.0951	1.0000
P55	0.0001	0.01	0.9271	1.0000
P64	0.0126	2.04	0.1547	1.0000
P70	0.4156	113.76	<.0001	1.0000
P80	0.0015	0.24	0.6245	1.0000

Variable P19 will be entered.

Variable(s) That Have Been Entered
P19

Multivariate Statistics					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.500514	159.67	1	160	<.0001
Pillai's Trace	0.499486	159.67	1	160	<.0001
Average Squared Canonical Correlation	0.499486				

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3b

The STEPDISC Procedure
Stepwise Selection: Step 2

Statistics for Removal, DF = 1, 160			
Variable	**R-Square**	**F Value**	**Pr > F**
P19	0.4995	159.67	<.0001

No variables can be removed.

Statistics for Entry, DF = 1, 159				
Variable	**Partial R-Square**	**F Value**	**Pr > F**	**Tolerance**
P8	0.0017	0.28	0.5993	0.9898
P14	0.0002	0.03	0.8632	0.9981
P33	0.0068	1.09	0.2980	0.9999
P37	0.1256	22.84	<.0001	0.6897
P49	0.0211	3.42	0.0663	0.9983
P55	0.0150	2.42	0.1217	0.9826
P64	0.0009	0.15	0.6985	0.9835
P70	0.1562	29.44	<.0001	0.6475
P80	0.0045	0.73	0.3954	0.9998

Variable P70 will be entered.

Variable(s) That Have Been Entered	
P19	P70

Multivariate Statistics					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.422313	108.75	2	159	<.0001
Pillai's Trace	0.577687	108.75	2	159	<.0001
Average Squared Canonical Correlation	0.577687				

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3b

The STEPDISC Procedure
Stepwise Selection: Step 3

Statistics for Removal, DF = 1, 159			
Variable	Partial R-Square	F Value	Pr > F
P19	0.2774	61.04	<.0001
P70	0.1562	29.44	<.0001

No variables can be removed.

Statistics for Entry, DF = 1, 158				
Variable	Partial R-Square	F Value	Pr > F	Tolerance
P8	0.0000	0.00	0.9985	0.6403
P14	0.0021	0.34	0.5609	0.6346
P33	0.0241	3.90	0.0501	0.6333
P37	0.0483	8.02	0.0052	0.5237
P49	0.0226	3.65	0.0578	0.6472
P55	0.0257	4.17	0.0429	0.6443
P64	0.0066	1.05	0.3066	0.6292
P80	0.0007	0.11	0.7407	0.6397

Variable P37 will be entered.

Variable(s) That Have Been Entered		
P19	P37	P70

Multivariate Statistics					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.401908	78.38	3	158	<.0001
Pillai's Trace	0.598092	78.38	3	158	<.0001
Average Squared Canonical Correlation	0.598092				

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3b

The STEPDISC Procedure
Stepwise Selection: Step 4

Statistics for Removal, DF = 1, 158			
Variable	Partial R-Square	F Value	Pr > F
P19	0.2254	45.97	<.0001
P37	0.0483	8.02	0.0052
P70	0.0817	14.05	0.0002

No variables can be removed.

Statistics for Entry, DF = 1, 157				
Variable	Partial R-Square	F Value	Pr > F	Tolerance
P8	0.0000	0.00	0.9753	0.5195
P14	0.0008	0.12	0.7259	0.5096
P33	0.0194	3.10	0.0803	0.5082
P49	0.0258	4.17	0.0429	0.5233
P55	0.0290	4.69	0.0319	0.5227
P64	0.0096	1.52	0.2199	0.5210
P80	0.0005	0.08	0.7717	0.5194

Variable P55 will be entered.

Variable(s) That Have Been Entered			
P19	P37	P55	P70

Multivariate Statistics					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.390256	61.32	4	157	<.0001
Pillai's Trace	0.609744	61.32	4	157	<.0001
Average Squared Canonical Correlation	0.609744				

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3b

The STEPDISC Procedure
Stepwise Selection: Step 5

Statistics for Removal, DF = 1, 157			
Variable	Partial R-Square	F Value	Pr > F
P19	0.2366	48.65	<.0001
P37	0.0515	8.53	0.0040
P55	0.0290	4.69	0.0319
P70	0.0877	15.09	0.0002

No variables can be removed.

Statistics for Entry, DF = 1, 156				
Variable	Partial R-Square	F Value	Pr > F	Tolerance
P8	0.0000	0.00	0.9743	0.5186
P14	0.0009	0.15	0.7024	0.5085
P33	0.0221	3.52	0.0626	0.5069
P49	0.0236	3.77	0.0539	0.5222
P64	0.0091	1.44	0.2320	0.5201
P80	0.0000	0.01	0.9393	0.5179

No variables can be entered.

No further steps are possible.

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3b

The STEPDISC Procedure

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

Based on the stepwise selection, the variables we retain are P19, P70, P37 and P55.

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3b

The DISCRIM Procedure

Total Sample Size	162	**DF Total**	161
Variables	4	**DF Within Classes**	160
Classes	2	**DF Between Classes**	1

Number of Observations Read	162
Number of Observations Used	162

Class Level Information

grp	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	1	77	77.0000	0.475309	0.475309
2	2	85	85.0000	0.524691	0.524691

Within Covariance Matrix Information

grp	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
1	4	15.53695
2	4	16.99771
Pooled	4	16.49642

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3b

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
29.980274	10	0.0009

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.

Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3b

The DISCRIM Procedure

Generalized Squared Distance to grp		
From grp	1	2
1	17.02454	25.27966
2	23.09375	18.28761

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3b

The DISCRIM Procedure

Multivariate Statistics and Exact F Statistics					
S=1 M=1 N=77.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.39025604	61.32	4	157	<.0001
Pillai's Trace	0.60974396	61.32	4	157	<.0001
Hotelling-Lawley Trace	1.56242027	61.32	4	157	<.0001
Roy's Greatest Root	1.56242027	61.32	4	157	<.0001

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3b

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CANCER
Cross-validation Summary using Quadratic Discriminant Function

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	

The Test of Homogeneity of Within Covariance Matrix is significant (<0.05), hence we reject H0 and conclude that the two groups do not have the same covariance and quadratic discriminant analysis (QDA) needs to be implemented. Based on the MANOVA tests we can conclude that there is a linear relationship between 'grp' and the selected variables. The null was rejected by all tests, hence the canonical correlations are not zero.The cross-validation error is 0.099, with 8+8=16 misclassified observations. This is 4 more than the full model. The seperation performance is roughly the same as part a).

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3c

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	CANCER
Random Number Seed	123456789
Sample Size	50
Selection Probability	0.308642
Sampling Weight	3.24
Number of Replicates	1
Total Sample Size	50
Output Data Set	TEST

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3c

The DISCRIM Procedure

Total Sample Size	112	**DF Total**	111
Variables	4	**DF Within Classes**	110
Classes	2	**DF Between Classes**	1

Number of Observations Read	112
Number of Observations Used	112

Class Level Information

grp	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	1	54	54.0000	0.482143	0.482143
2	2	58	58.0000	0.517857	0.517857

Within Covariance Matrix Information

grp	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
1	4	15.66600
2	4	16.81650

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3c

The DISCRIM Procedure

Generalized Squared Distance to grp		
From grp	1	2
1	17.12503	27.76494
2	21.68422	18.13261

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 3c

The DISCRIM Procedure
Classification Summary for Test Data: WORK.TEST
Classification Summary using Quadratic Discriminant Function

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	

The following are results from a quadratic discriminant analysis based on training and test set. Among 50 observations assigned to the test set, 6 observations aremisclassified and the total error rate is observed as 0.1169, which is larger than thecross-validation error rate of 0.099 in Exercise 2. The performance is good. According to the percentage of misclassified observation, the logistic regression was the best method to fit this model. In this logistic regression, 15 of 162 observations were misclassified, whereas in the discriminant analysis, 16 of 162 observations were misclassified. And the test data misclassified 6 of 50 observations.

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 5a

The GENMOD Procedure

Model Information	
Data Set	WORK.HOUSING
Distribution	Gamma
Link Function	Log
Dependent Variable	medv

Number of Observations Read	506
Number of Observations Used	506

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	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	4.2057	0.1985	3.8166	4.5947	448.89	<.0001
crim	1	-0.0100	0.0013	-0.0125	-0.0075	60.10	<.0001
zn	1	0.0013	0.0005	0.0002	0.0024	5.82	0.0158
indus	1	0.0022	0.0025	-0.0026	0.0070	0.79	0.3729
chas	1	0.1093	0.0344	0.0420	0.1767	10.12	0.0015
nox	1	-0.8495	0.1501	-1.1437	-0.5553	32.03	<.0001
rm	1	0.0847	0.0157	0.0539	0.1155	29.06	<.0001
age	1	0.0002	0.0005	-0.0008	0.0012	0.22	0.6427
dis	1	-0.0537	0.0077	-0.0688	-0.0385	48.21	<.0001

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 5a

The GENMOD Procedure

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
rad	1	0.0152	0.0026	0.0100	0.0204	33.19	<.0001
tax	1	-0.0006	0.0002	-0.0009	-0.0003	15.86	<.0001
ptratio	1	-0.0392	0.0052	-0.0494	-0.0289	56.23	<.0001
b	1	0.0004	0.0001	0.0002	0.0006	14.40	0.0001
lstat	1	-0.0286	0.0019	-0.0323	-0.0250	232.80	<.0001
Scale	1	27.7923	1.7369	24.5882	31.4138		

Note: The scale parameter was estimated by maximum likelihood.

LR Statistics For Type 1 Analysis				
Source	2*LogLikelihood	DF	Chi-Square	Pr > ChiSq
Intercept	-3595.8885			
crim	-3470.7025	1	125.19	<.0001
zn	-3422.7082	1	47.99	<.0001
indus	-3376.0131	1	46.70	<.0001
chas	-3343.9931	1	32.02	<.0001
nox	-3338.7742	1	5.22	0.0223
rm	-3149.1101	1	189.66	<.0001
age	-3146.1330	1	2.98	0.0844
dis	-3100.3981	1	45.73	<.0001
rad	-3100.2438	1	0.15	0.6945
tax	-3088.7519	1	11.49	0.0007
ptratio	-3047.5538	1	41.20	<.0001
b	-3023.3105	1	24.24	<.0001
lstat	-2830.6340	1	192.68	<.0001

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

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 5a

The GENMOD Procedure

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
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

From the Type 3 analysis, we can see that the insignificant variables are 'indus' and 'age'. They have a p-value greater than 0.05. We can conclude that these variables are not related to median value.

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 5b

The GENMOD Procedure

Model Information	
Data Set	WORK.HOUSING
Distribution	Gamma
Link Function	Log
Dependent Variable	medv

Number of Observations Read	506
Number of Observations Used	506

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	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	4.2007	0.1982	3.8123	4.5891	449.32	<.0001
crim	1	-0.0100	0.0013	-0.0125	-0.0075	59.98	<.0001
zn	1	0.0013	0.0005	0.0002	0.0023	5.64	0.0176
indus	1	0.0022	0.0025	-0.0026	0.0070	0.80	0.3700
chas	1	0.1103	0.0343	0.0431	0.1776	10.35	0.0013
nox	1	-0.8317	0.1452	-1.1163	-0.5472	32.82	<.0001
rm	1	0.0860	0.0155	0.0557	0.1163	30.95	<.0001
dis	1	-0.0547	0.0074	-0.0692	-0.0403	54.95	<.0001
rad	1	0.0151	0.0026	0.0100	0.0203	32.99	<.0001

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 5b

The GENMOD Procedure

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
tax	1	-0.0006	0.0002	-0.0009	-0.0003	15.76	<.0001
ptratio	1	-0.0390	0.0052	-0.0492	-0.0288	56.06	<.0001
b	1	0.0004	0.0001	0.0002	0.0006	14.55	0.0001
lstat	1	-0.0284	0.0018	-0.0319	-0.0249	251.61	<.0001
Scale	1	27.7806	1.7362	24.5779	31.4006		

Note: The scale parameter was estimated by maximum likelihood.

LR Statistics For Type 1 Analysis				
Source	2*LogLikelihood	DF	Chi-Square	Pr > ChiSq
Intercept	-3595.8885			
crim	-3470.7025	1	125.19	<.0001
zn	-3422.7082	1	47.99	<.0001
indus	-3376.0131	1	46.70	<.0001
chas	-3343.9931	1	32.02	<.0001
nox	-3338.7742	1	5.22	0.0223
rm	-3149.1101	1	189.66	<.0001
dis	-3115.0432	1	34.07	<.0001
rad	-3114.8020	1	0.24	0.6233
tax	-3103.0227	1	11.78	0.0006
ptratio	-3059.1583	1	43.86	<.0001
b	-3035.0488	1	24.11	<.0001
lstat	-2830.8489	1	204.20	<.0001

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

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 5b

The GENMOD Procedure

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
tax	1	15.44	<.0001
ptratio	1	53.17	<.0001
b	1	14.10	0.0002
lstat	1	204.20	<.0001

We first removed 'age', since it had the largest p-value. The AIC value is 2858.8, which is smaller than the previous model. There is one variable with a large p-value, 'indus', hence we remove it and refit the gamma model again.

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 5b

The GENMOD Procedure

Model Information	
Data Set	WORK.HOUSING
Distribution	Gamma
Link Function	Log
Dependent Variable	medv

Number of Observations Read	506
Number of Observations Used	506

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	

Algorithm converged.

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

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 5b

The GENMOD Procedure

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
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		

Note: The scale parameter was estimated by maximum likelihood.

LR Statistics For Type 1 Analysis				
Source	2*LogLikelihood	DF	Chi-Square	Pr > ChiSq
Intercept	-3595.8885			
crim	-3470.7025	1	125.19	<.0001
zn	-3422.7082	1	47.99	<.0001
chas	-3401.9568	1	20.75	<.0001
nox	-3362.6136	1	39.34	<.0001
rm	-3153.1939	1	209.42	<.0001
dis	-3125.4178	1	27.78	<.0001
rad	-3125.3260	1	0.09	0.7619
tax	-3104.7498	1	20.58	<.0001
ptratio	-3059.4536	1	45.30	<.0001
b	-3035.1682	1	24.29	<.0001
lstat	-2831.6547	1	203.51	<.0001

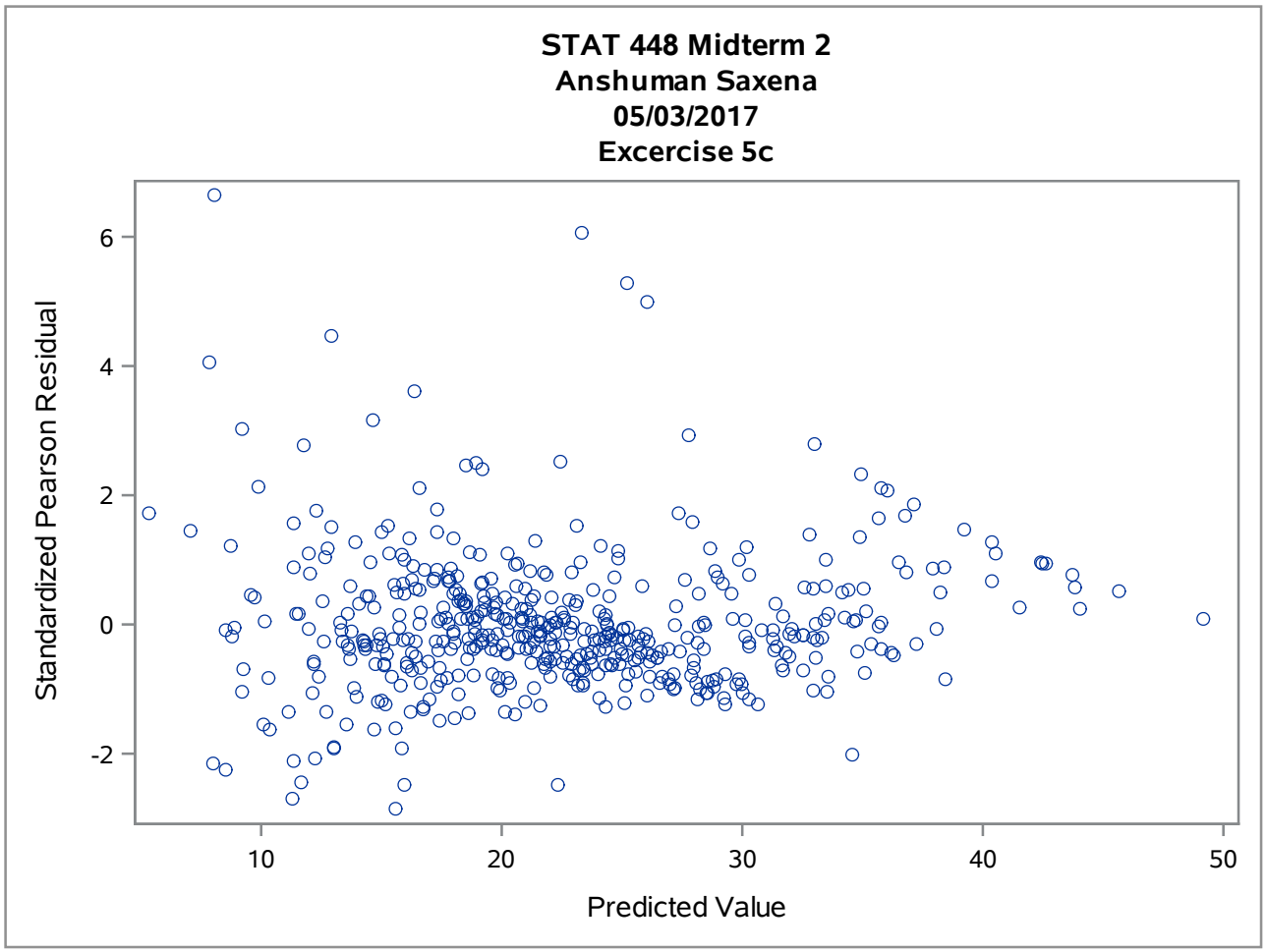
LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
crim	1	49.87	<.0001
zn	1	5.23	0.0222
chas	1	11.25	0.0008
nox	1	31.11	<.0001
rm	1	29.21	<.0001
dis	1	56.62	<.0001
rad	1	31.06	<.0001
tax	1	15.42	<.0001
ptratio	1	52.37	<.0001

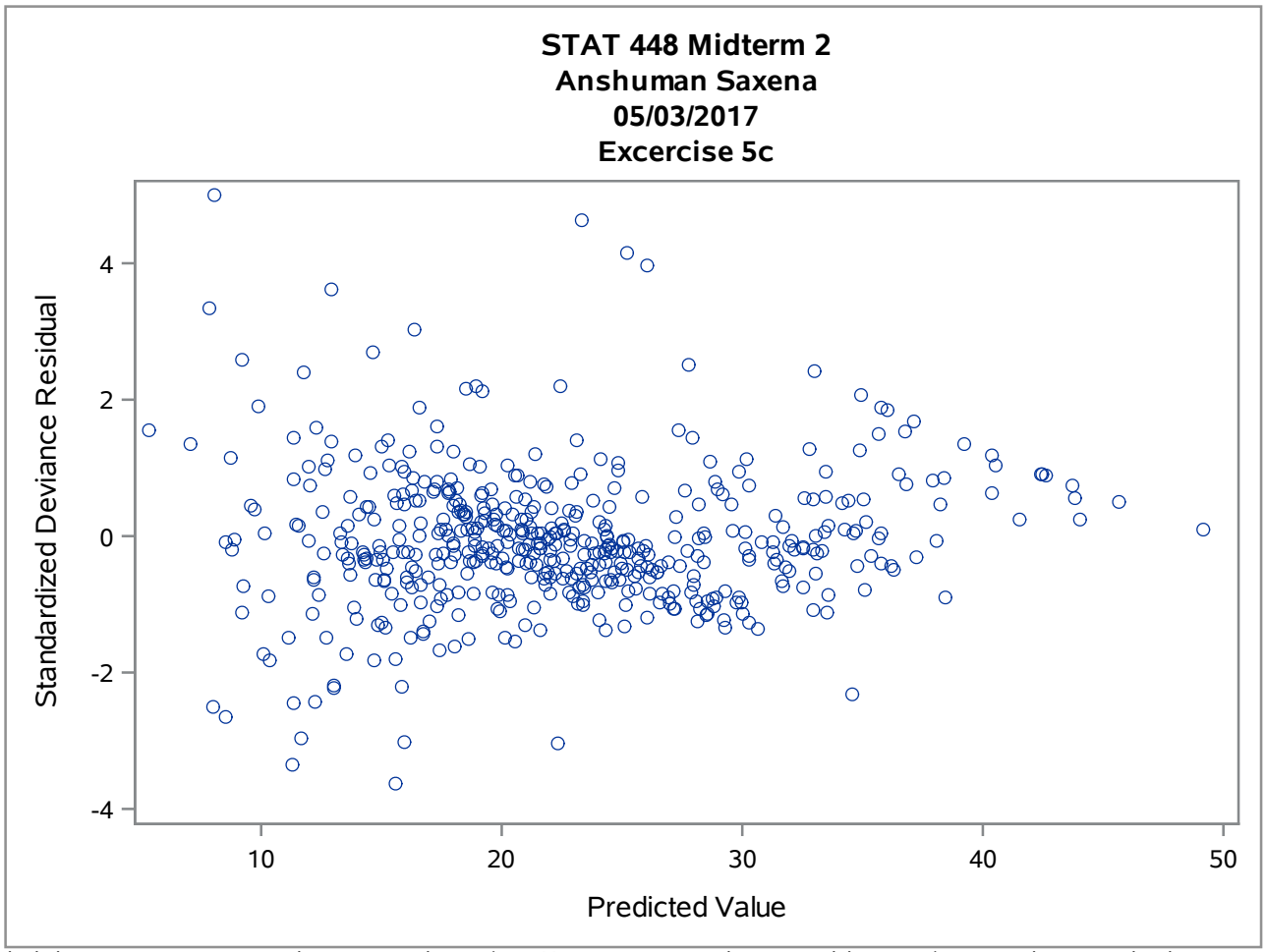
STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 5b

The GENMOD Procedure

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
b	1	13.90	0.0002
lstat	1	203.51	<.0001

After removing 'indus', the AIC value is 2857.65 and there are no more predictors which are considered insignificant (all have p-value < 0.05). Based on Type 3 analysis, this should be our final model.


STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 5c
Standardized Pearson Residual
6
4
2
0
-2
10
20
30
40
50
Predicted Value



From the residual plots, we see no pattern and no extreme observations. Hence, we assume that our model assumptions are adequate to the data.

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 6a

The GENMOD Procedure

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

Number of Observations Read	59
Number of Observations Used	59

Class Level Information		
Class	**Levels**	**Values**
Treat	2	0 1

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	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	0.5051	0.2638	-0.0119	1.0221	3.67	0.0555
Treat	0	1	0.2705	0.1019	0.0708	0.4701	7.05	0.0079
Treat	1	0	0.0000	0.0000	0.0000	0.0000	.	.
BL		1	0.0221	0.0011	0.0199	0.0242	410.75	<.0001

STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 6a

The GENMOD Procedure

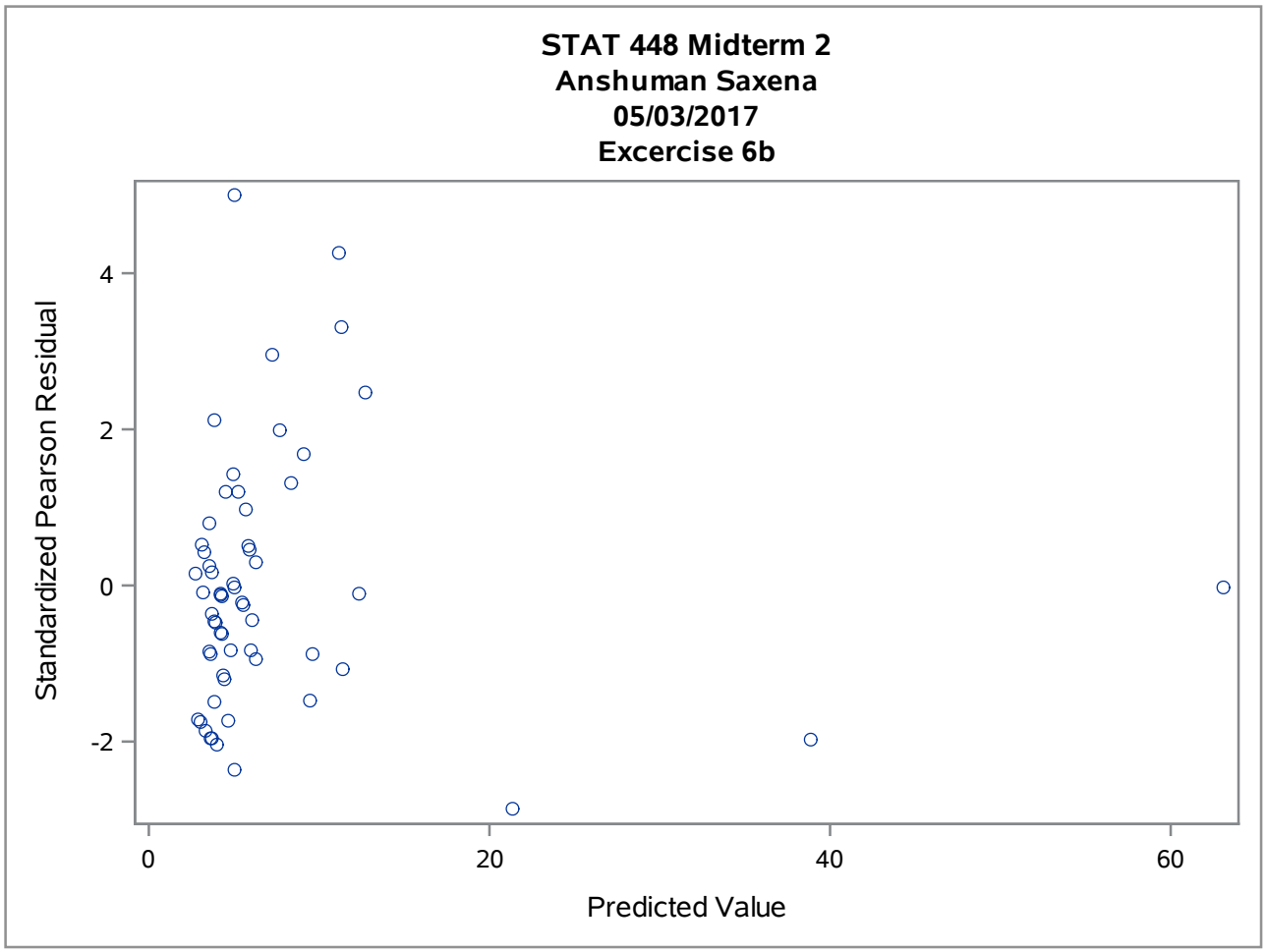
Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Age		1	0.0140	0.0086	-0.0028	0.0309	2.68	0.1017
Scale		0	1.0000	0.0000	1.0000	1.0000		

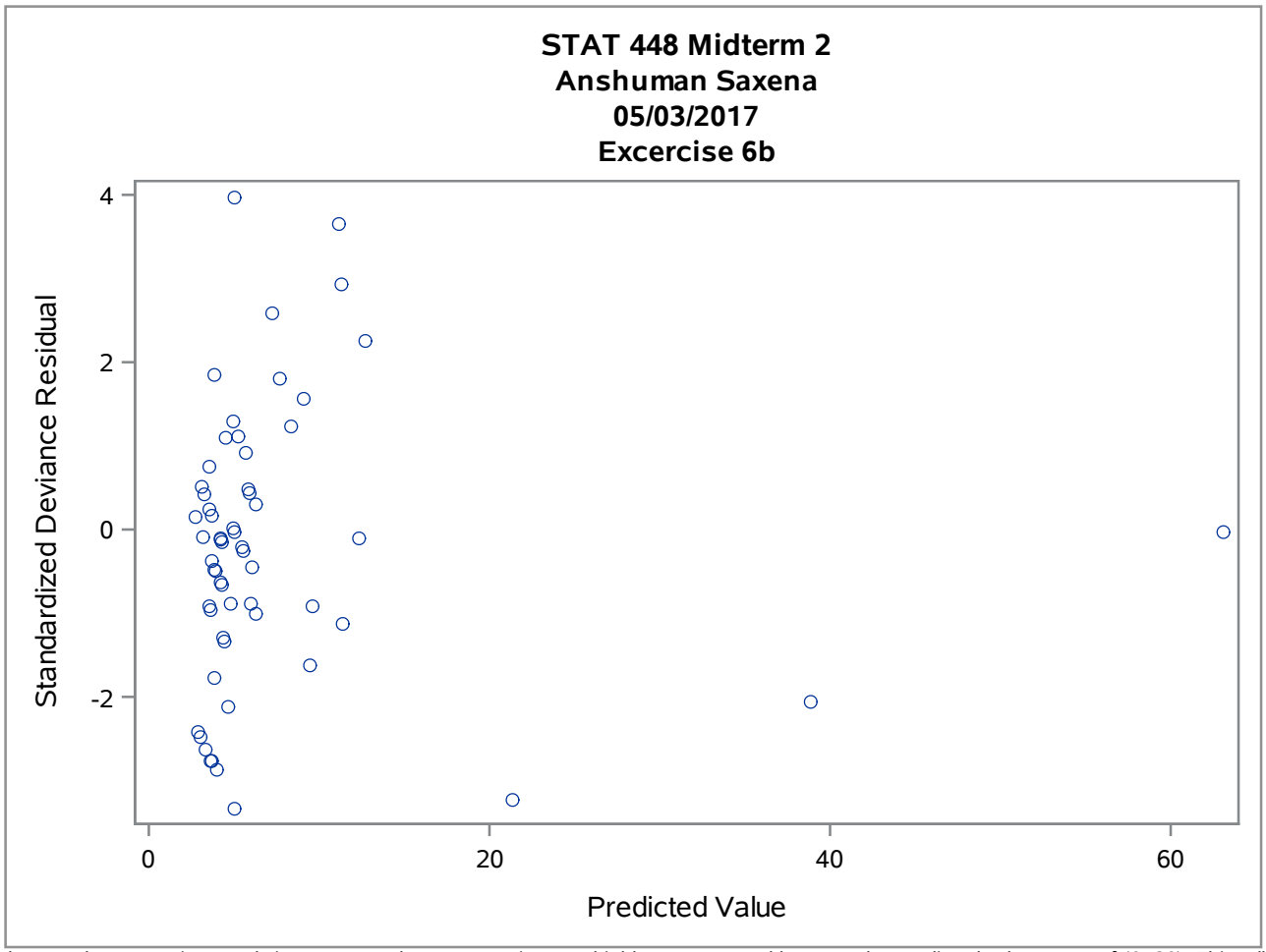
Note: The scale parameter was held fixed.

LR Statistics For Type 1 Analysis				
Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	476.2487			
Treat	473.0840	1	3.16	0.0752
BL	149.6763	1	323.41	<.0001
Age	147.0216	1	2.65	0.1032

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
Treat	1	7.08	0.0078
BL	1	320.75	<.0001
Age	1	2.65	0.1032

Based on the parameter estimates log-linear Poisson model, the variable 'Age' is insignificant with a p-value of 0.1017. All other predictor variables are significant. The same conclusion holds for the Type 3 analysis. The Type 1 analysis, however, classifies 'Treat' to be insignificant as well as 'Age'.


STAT 448 Midterm 2
Anshuman Saxena
05/03/2017
Excercise 6b
Standardized Pearson Residual
4
2
0
-2
0
20
40
60
Predicted Value



By looking at the two plots we notice an obvious pattern: the scatter points are highly concentrated between the predicted value range of (0, 20). This tells us that the predictor variables don't fit the model well because the predicted values are low but the residuals are quite high.