Exercise 1

a) Results for a logistic regression for high cholesterol status follow. The global tests are all significant indicating that at least one of the predictors is significant. The type 3 analysis indicates that we would lose significantly more information about the response than expected due to chance if we remove the AgeAtStart term. If we were to remove any of the other terms from the full model, we would not lose significantly more information than expected due to chance. The sex and smoking terms are highly insignificant in the type 3 analysis and weight status is slightly insignificant.

The parameter estimates lead to similar conclusions about the predictors. Age at start is definitely significant (and has a positive relationship with high cholesterol status), and the other parameter estimates are all insignificant.

Based on these results, we would want to keep AgeAtStart in the model and will want to remove one or more of the other terms, though we could not remove them all at once because significant highly correlated predictors could be insignificant in a Type 3 sense simply because they are providing similar information. We would want to add or remove terms one at a time to get to a final model.

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	27.8623	5	<.0001
Score	26.8925	5	<.0001
Wald	25.4685	5	0.0001

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Sex	1	0.5167	0.4722
Weight_Status	2	5.1887	0.0747
AgeAtStart	1	16.7764	<.0001
Smoking	1	0.0937	0.7595

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-3.1345	0.6524	23.0817	<.0001
Sex	**Male**	1	0.0724	0.1007	0.5167	0.4722
Weight_Status	**Underweight**	1	-0.7067	0.7225	0.9567	0.3280
Weight_Status	**Overweight**	1	0.5780	0.3728	2.4043	0.1210
AgeAtStart		1	0.0515	0.0126	16.7764	<.0001
Smoking		1	0.00272	0.00890	0.0937	0.7595

b) Using stepwise selection, we find that age at start is the only significant predictor of high cholesterol status.

Summary of Stepwise Selection								
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	AgeAtStart		1	1	21.1401		<.0001	Age at Start

Looking at the parameter estimates, we see a statistically significant positive relationship between age at the start and having high cholesterol. Being one year older at the start would correspond to an expected increase of .0528 in the log-odds of having high cholesterol. The odds ratio shows this corresponds to a multiplicative increase of 1.054 in the odds of having high cholesterol for each year older an individual is at the start.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.7122	0.5173	27.4878	<.0001
AgeAtStart	1	0.0528	0.0117	20.4218	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
AgeAtStart	1.054	1.030	1.079

Exercise 2

a) Starting with all previous predictors and cholesterol, we model high blood pressure status. The global tests are all significant, telling us that at least one of predictors has some predictive power.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	47.5962	6	<.0001
Score	44.9630	6	<.0001
Wald	41.1667	6	<.0001

The type 3 analysis indicates that removing weight status or age at start would result in significant loss of information, but removal of any one of the other terms from the full model would not.

The parameter estimates for age at start and for comparing overweight to normal weight are both significant and positive, while sex, underweight compared to normal, smoking, and cholesterol all have insignificant parameter estimates.

Based on these results, we would want to keep age at start and weight status in the model, and would want to remove one or more of the other predictors.

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Sex	1	0.0293	0.8641
Weight_Status	2	18.3451	0.0001
AgeAtStart	1	7.8153	0.0052
Smoking	1	0.4649	0.4953
Cholesterol	1	3.0630	0.0801

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-3.3034	0.7431	19.7600	<.0001
Sex	**Male**	1	-0.0173	0.1012	0.0293	0.8641
Weight_Status	**Underweight**	1	-0.6759	0.7227	0.8749	0.3496
Weight_Status	**Overweight**	1	0.8255	0.3736	4.8822	0.0271
AgeAtStart		1	0.0361	0.0129	7.8153	0.0052
Smoking		1	-0.00613	0.00899	0.4649	0.4953
Cholesterol		1	0.00379	0.00216	3.0630	0.0801

b) Using stepwise selection, we see that age at start and weight status are the only significant predictors of high blood pressure status.

Summary of Stepwise Selection								
	Effect							
Step	**Entered**	**Removed**	**DF**	**Number In**	**Score Chi-Square**	**Wald Chi-Square**	**Pr > ChiSq**	**Variable Label**
1	**Weight_Status**		2	1	28.2459		<.0001	Weight Status
2	**AgeAtStart**		1	2	13.9317		0.0002	Age at Start

The parameter estimates for age at start and overweight compared to normal are again significant and positive, while underweight compared to normal weight is not significant. The odds ratios show that we would expect statistically significant increases in odds of high blood pressure as age increases and from normal to overweight weight status. In particular, the expected odds of high blood pressure would be

multiplied by 1.045 for every year increase in age at start, and the expected odds for overweight individuals to have high blood pressure would be 2.774 times that of normal weight individuals. The odds is not significantly different for normal and underweight individuals.

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-2.8487	0.6125	21.6327	<.0001
Weight_Status	**Underweight**	1	-0.7115	0.7218	0.9715	0.3243
Weight_Status	**Overweight**	1	0.8659	0.3726	5.4016	0.0201
AgeAtStart		1	0.0442	0.0120	13.6604	0.0002

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Weight_Status Underweight vs Normal	0.573	0.067	4.912
Weight_Status Overweight vs Normal	2.774	1.750	4.397
AgeAtStart	1.045	1.021	1.070

Exercise 3

a) Fitting with all predictors, the scaled deviance is about 2.7 so we estimate a scale to account for overdispersion.

Model Information	
Data Set	WORK.EPIL
Distribution	Poisson
Link Function	Log
Dependent Variable	P4

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	54	145.8141	2.7003
Scaled Deviance	54	145.8141	2.7003
Pearson Chi-Square	54	135.5868	2.5109
Scaled Pearson X2	54	135.5868	2.5109
Log Likelihood		591.2912	
Full Log Likelihood		-166.7912	
AIC (smaller is better)		343.5825	
AICC (smaller is better)		344.7146	
BIC (smaller is better)		353.9702	

After estimating the scale, we obtain the following results. Only the parameter estimate for baseline (BL) is significant, and it is positive indicating an expected increase in seizure count for increased baseline seizure count.

Since we have estimated the scale, we should refer to the F statistics for the type 1 and type 3 analyses. The type 1 analysis indicates that, in the order given, the count after one period of treatment (P1), the treatment indicator, and the baseline value are all significant. Type 3 analysis indicates that only removing the baseline count from the full model would result in a significant loss of information. We will definitely want to keep baseline count in the model, but at least one of the other terms can be removed. We will have to check which if any of the other terms should be kept.

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	54	145.8141	2.7003
Scaled Deviance	54	54.0000	1.0000
Pearson Chi-Square	54	135.5868	2.5109
Scaled Pearson X2	54	50.2125	0.9299
Log Likelihood		218.9756	
Full Log Likelihood		-166.7912	
AIC (smaller is better)		343.5825	
AICC (smaller is better)		344.7146	
BIC (smaller is better)		353.9702	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.9023	0.5019	-0.0814	1.8861	3.23	0.0722
P1	1	0.0039	0.0058	-0.0076	0.0154	0.44	0.5049
Treat	1	-0.3100	0.1782	-0.6593	0.0392	3.03	0.0819
BL	1	0.0195	0.0042	0.0113	0.0277	21.85	<.0001
Age	1	0.0120	0.0144	-0.0162	0.0402	0.70	0.4032
Scale	0	1.6432	0.0000	1.6432	1.6432		

Note: The scale parameter was estimated by the square root of DEVIANCE/DOF.

LR Statistics For Type 1 Analysis							
Source	Deviance	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
Intercept	476.2487						
P1	217.7231	1	54	95.74	<.0001	95.74	<.0001
Treat	200.8566	1	54	6.25	0.0155	6.25	0.0124
BL	147.6894	1	54	19.69	<.0001	19.69	<.0001
Age	145.8141	1	54	0.69	0.4083	0.69	0.4046

LR Statistics For Type 3 Analysis						
Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
P1	1	54	0.45	0.5065	0.45	0.5037
Treat	1	54	3.06	0.0860	3.06	0.0803
BL	1	54	20.18	<.0001	20.18	<.0001
Age	1	54	0.69	0.4083	0.69	0.4046

b) Sequentially removing terms, we find that baseline is significant and treatment has a type 3 p-value of .055 and its parameter estimate has a p-value slightly greater than .05, so we would have baseline count as the only predictor in the model. (Note: It would probably be better to treat baseline as an offset – a known fixed value for each observation – rather than a predictor in this case. This would give a more accurate modeling of change in seizure count, which is really what we are interest in, and would potentially result in a significant result for the treatment effect. However, we did not discuss offsets in class).

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	57	159.9413	2.8060
Scaled Deviance	57	57.0000	1.0000
Pearson Chi-Square	57	151.2008	2.6526
Scaled Pearson X2	57	53.8851	0.9454
Log Likelihood		208.2074	
Full Log Likelihood		-173.8548	
AIC (smaller is better)		351.7097	
AICC (smaller is better)		351.9240	
BIC (smaller is better)		355.8648	

With the baseline count model, we find that baseline is significant in both the type 1 and type 3 analyses. The baseline parameter estimate is .0209 indicating that the seizure count after four periods of treatment increases slightly as baseline count increases. Specifically, for an increase of one in the baseline count, we would expect the count after 4 periods of treatment to be multiplied by $e^{.0209} = 1.021$, so we would expect a 2.1% increase in the number of seizures after four treatment periods for each additional seizure prior to treatment. In this model, we do not see a significant effect from the treatment at a .05 level, though it would be significant at a .1 level and might be significant if the baseline count were treated as an offset as mentioned before.

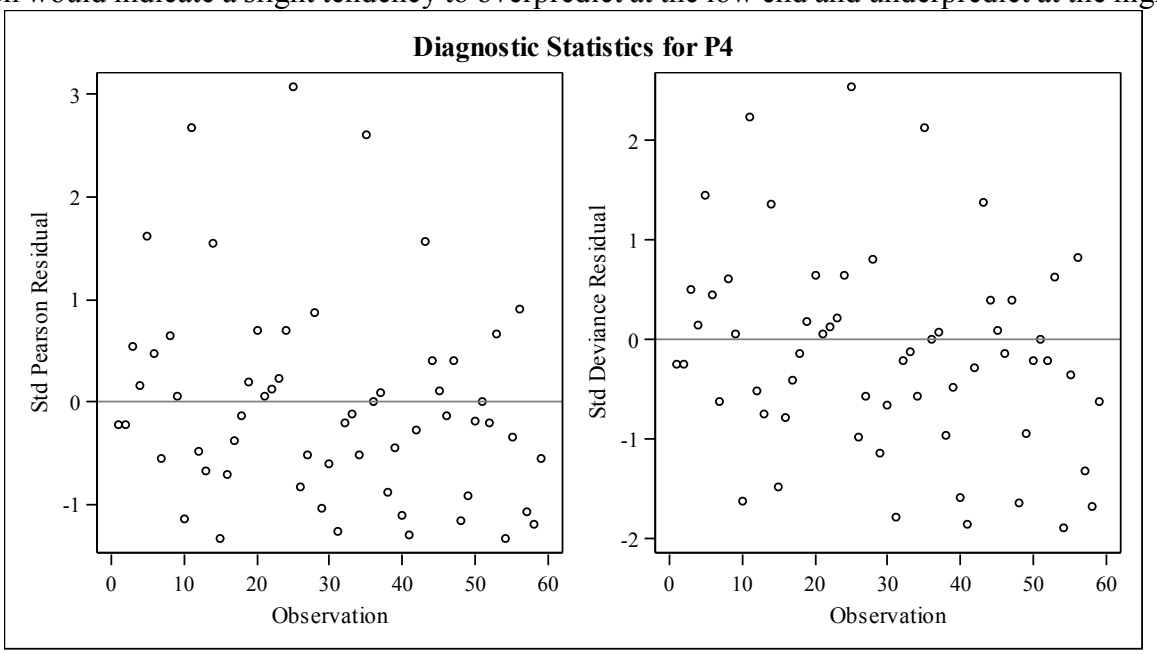
Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	1.0897	0.1295	0.8360	1.3435	70.84	<.0001
BL	1	0.0209	0.0017	0.0176	0.0242	156.27	<.0001
Scale	0	1.6751	0.0000	1.6751	1.6751		

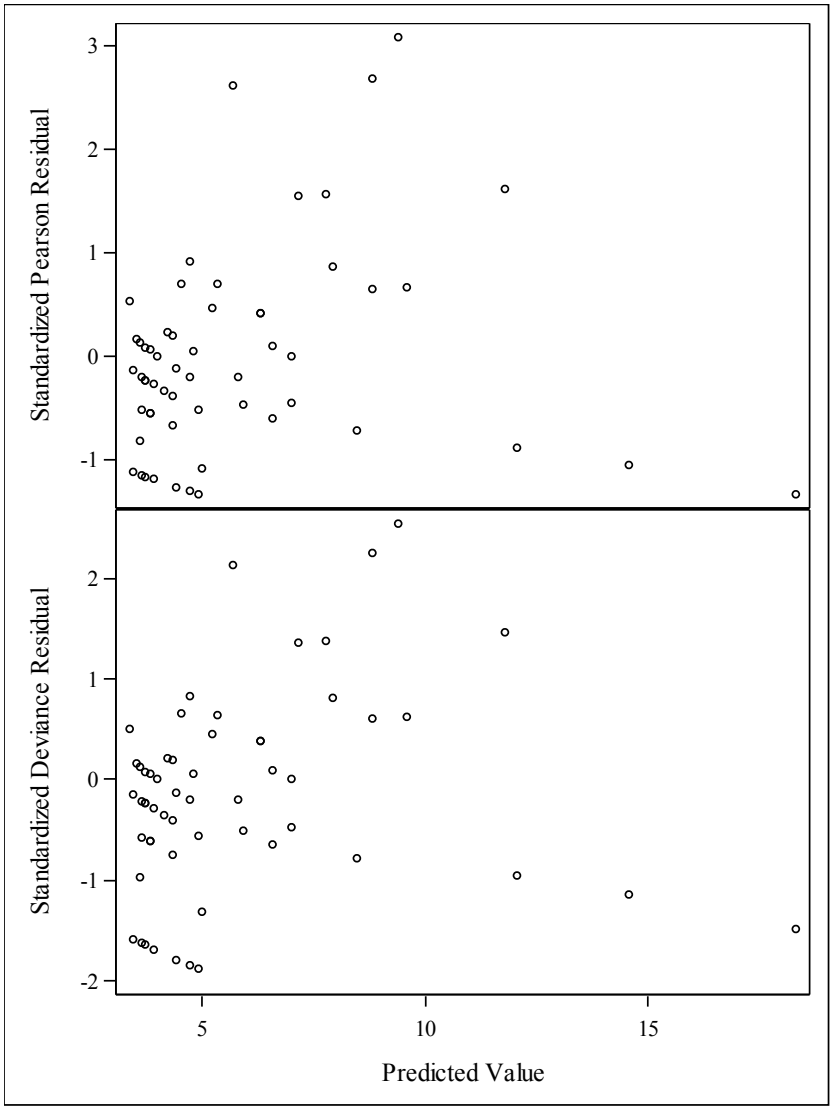
Note: The scale parameter was estimated by the square root of DEVIANCE/DOF.

LR Statistics For Type 1 Analysis							
Source	**Deviance**	**Num DF**	**Den DF**	**F Value**	**Pr > F**	**Chi-Square**	**Pr > ChiSq**
Intercept	476.2487						
BL	159.9413	1	57	112.73	<.0001	112.73	<.0001

LR Statistics For Type 3 Analysis						
Source	**Num DF**	**Den DF**	**F Value**	**Pr > F**	**Chi-Square**	**Pr > ChiSq**
BL	1	57	112.73	<.0001	112.73	<.0001

The residuals for the model generally look OK. The distribution of residuals versus observation number shows no major issues. Plotted against the predicted values, it appears there may a slight upward trend which would indicate a slight tendency to overpredict at the low end and underpredict at the higher end.





Exercise 4

a) Fitting count after one treatment period as a function of treatment, baseline count and age, the scaled deviance is about 3.4 so we estimate a scale to account for overdispersion.

Model Information	
Data Set	WORK.EPIL
Distribution	Poisson
Link Function	Log
Dependent Variable	P1

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	55	187.3802	3.4069
Scaled Deviance	55	187.3802	3.4069
Pearson Chi-Square	55	199.4665	3.6267
Scaled Pearson X2	55	199.4665	3.6267
Log Likelihood		908.6710	
Full Log Likelihood		-191.9378	
AIC (smaller is better)		391.8756	
AICC (smaller is better)		392.6163	
BIC (smaller is better)		400.1857	

After estimating the scale, we obtain the following results. Both the parameter estimate for age and the estimate for baseline are statistically significant and positive.

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	55	187.3802	3.4069
Scaled Deviance	55	55.0000	1.0000
Pearson Chi-Square	55	199.4665	3.6267
Scaled Pearson X2	55	58.5476	1.0645
Log Likelihood		266.7139	
Full Log Likelihood		-191.9378	
AIC (smaller is better)		391.8756	
AICC (smaller is better)		392.6163	
BIC (smaller is better)		400.1857	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-0.2411	0.5045	-1.2299	0.7477	0.23	0.6328
Treat	1	-0.1189	0.1710	-0.4540	0.2163	0.48	0.4869
BL	1	0.0257	0.0018	0.0222	0.0293	204.16	<.0001
Age	1	0.0465	0.0144	0.0182	0.0748	10.40	0.0013
Scale	0	1.8458	0.0000	1.8458	1.8458		

Note: The scale parameter was estimated by the square root of DEVIANCE/DOF.

Since we have estimated the scale, we should refer to the F statistics. The type 1 analysis indicates that, in the order given, the baseline count and age are significant. Type 3 analysis indicates the same, so our final model would contain those two terms.

LR Statistics For Type 1 Analysis							
Source	**Deviance**	**Num DF**	**Den DF**	**F Value**	**Pr > F**	**Chi-Square**	**Pr > ChiSq**
Intercept	746.4361						
Treat	745.4460	1	55	0.29	0.5920	0.29	0.5898
BL	222.2496	1	55	153.57	<.0001	153.57	<.0001
Age	187.3802	1	55	10.23	0.0023	10.23	0.0014

LR Statistics For Type 3 Analysis						
Source	**Num DF**	**Den DF**	**F Value**	**Pr > F**	**Chi-Square**	**Pr > ChiSq**
Treat	1	55	0.48	0.4896	0.48	0.4867
BL	1	55	163.79	<.0001	163.79	<.0001
Age	1	55	10.23	0.0023	10.23	0.0014

b) Fitting the model with baseline count and age as predictors, we get the following results.

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	56	189.0289	3.3755
Scaled Deviance	56	56.0000	1.0000
Pearson Chi-Square	56	200.1801	3.5746
Scaled Pearson X2	56	59.3036	1.0590
Log Likelihood		268.9505	
Full Log Likelihood		-192.7621	
AIC (smaller is better)		391.5243	
AICC (smaller is better)		391.9606	
BIC (smaller is better)		397.7569	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.3815	0.4630	-1.2890	0.5259	0.68	0.4099
BL	1	0.0257	0.0018	0.0222	0.0292	205.86	<.0001
Age	1	0.0494	0.0138	0.0225	0.0764	12.92	0.0003
Scale	0	1.8373	0.0000	1.8373	1.8373		

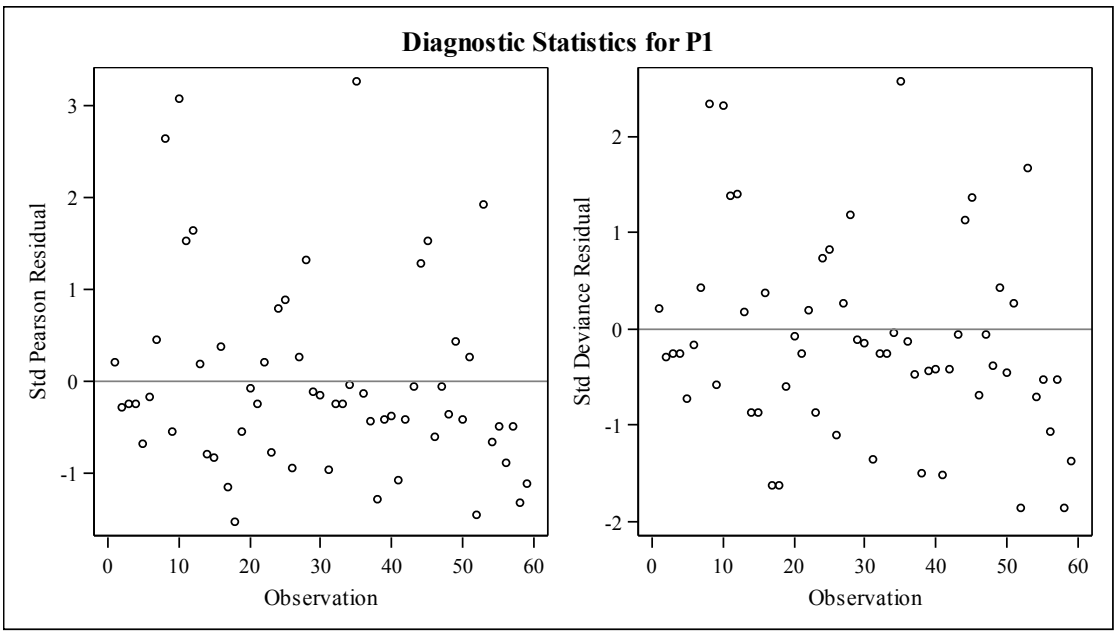
Note: The scale parameter was estimated by the square root of DEVIANCE/DOF.

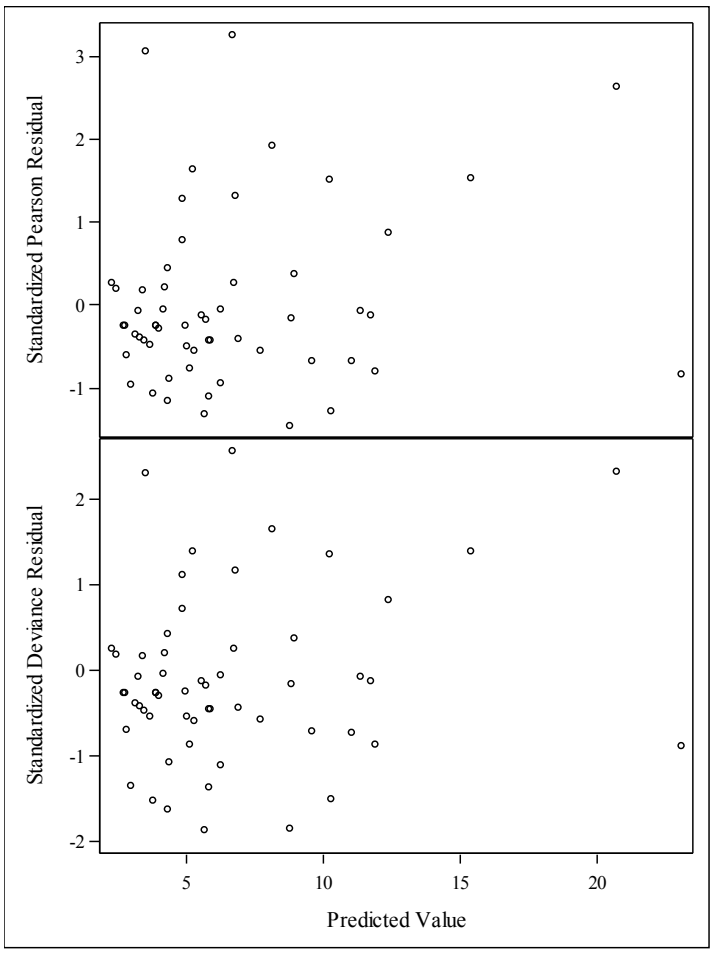
LR Statistics For Type 1 Analysis							
Source	Deviance	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
Intercept	746.4361						
BL	231.9037	1	56	152.43	<.0001	152.43	<.0001
Age	189.0289	1	56	12.70	0.0008	12.70	0.0004

LR Statistics For Type 3 Analysis						
Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
BL	1	56	165.11	<.0001	165.11	<.0001
Age	1	56	12.70	0.0008	12.70	0.0004

Baseline count and age are significant in the type 1 and type 3 analyses as expected, and their parameter estimates are as well. Age and baseline count are both positively related to seizure count after one period of treatment. For an increase of one in the baseline count, we would expect the count after one period of treatment to be multiplied by $e^{.0257} = 1.026$, so we would expect a 2.6% increase in the number of seizures after one treatment period for each additional seizure prior to treatment. For a one year increase in age, we would expect a multiplicative increase of $e^{.0494} = 1.051$ in seizures after one period of treatment, so we would expect 5.1% more seizures for each additional year of age.

Looking at the residuals for this model, there are no issues with the distributional assumption. The residuals versus observation number show no extreme large residuals, and the residuals versus predicted values seem fine as well. There may be a slight upward trend, but it is much less than in the previous model and is of no concern here.





Comparing the two models, we find that baseline count is a significant positive predictor of seizure count after one period and after four periods of treatment. Age is a significant positive predictor after one period but not after four periods. The difference between active treatment and placebo is highly insignificant after one treatment period, but barely insignificant after four periods, so there appears to be no short term impact from the active treatment, but there may be some impact over longer periods of time.