HW 7 Report

Yongchao Qiao

Part A

1. Build the generalized model and report the estimated parameters

Table 1: Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	97	83.9200	0.8652

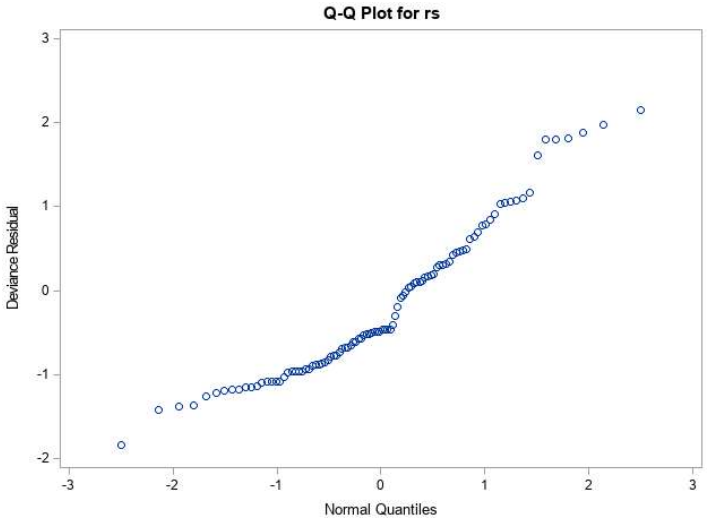
Table 2: Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > Chisq
Intercept	1	-2.8537	0.3773	-3.5931	-2.1143	57.22	<.0001
X	1	1.4299	0.3604	0.7236	2.1363	15.74	<.0001
Z	1	1.1234	0.1740	0.7823	1.4646	41.67	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Based on the introduction of the data, we build the generalized linear model following the formula: $\log(\mu) = a + bX + cZ + \log(t)$ and get the results. From table 1 we can see, the deviance is 83.92 which is less than its degree as 97, so there is no overdispersion. Then for parameters, the estimate for b is 1.4299 and that for c is 1.1234.

2. Q-Q plot and model fitting

Figure 1: Q-Q plot for deviance residuals



Based on the figure 1, we can see that these points mostly follow a diagonal line though not that ideal, which means this model generally meets the assumption for the deviance residuals. Besides, from table 1, residual deviance is not equal to its degrees of freedom, so the model can be considered as a good fitting in some degree though the performance is not

that ideal.

Part B

I. Loess and Logistic procedure

1. Loess procedure

Figure 2: Response vs. Age with 95% CI by Loess method

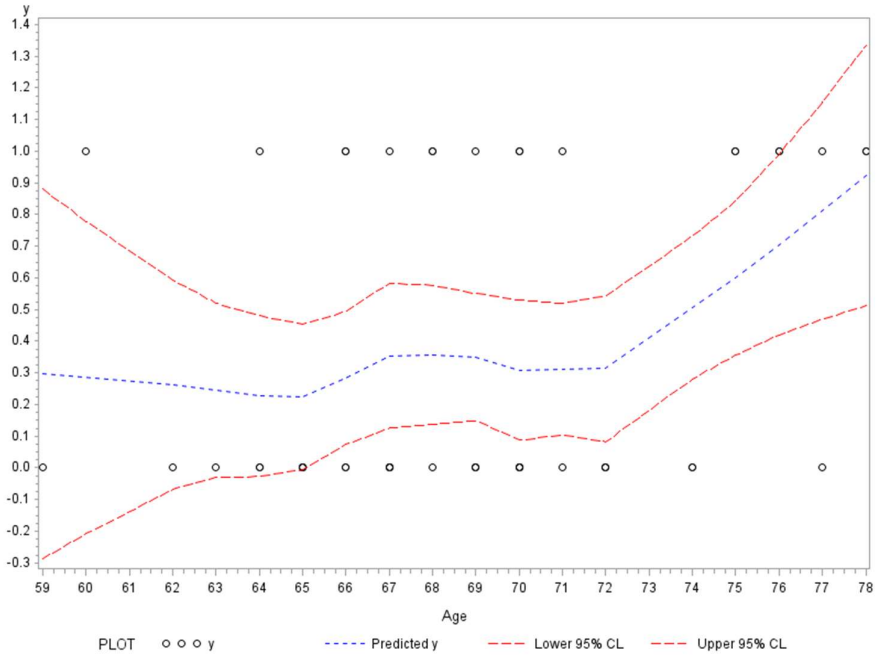
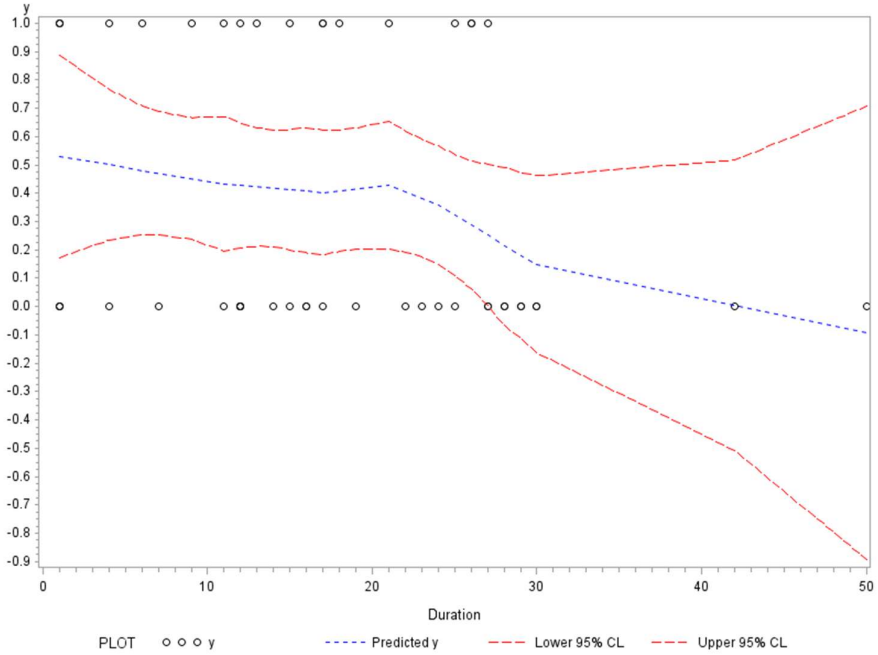


Figure 3: Response vs. Duration with 95% CI by Loess method



2. Logistic procedure

Response (Pain) ~ Age

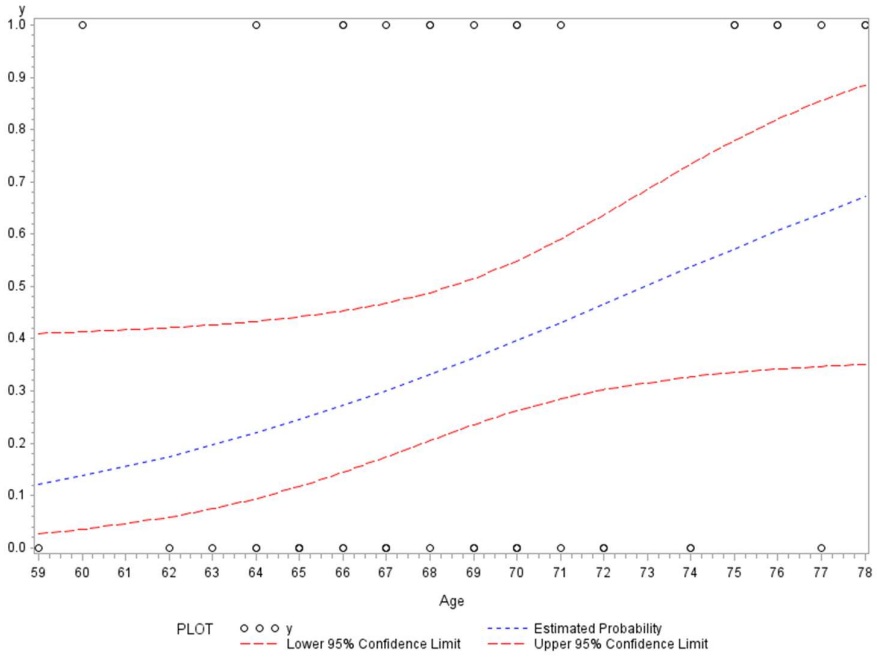
Table 3: Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4.3281	1	0.0375

Table 4: Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-10.3317	4.9993	4.2708	0.0388
Age	1	0.1416	0.0717	3.9039	0.0482

Figure 4: Response vs. Age with 95% CI by Logistic method



From table 3 and table 4 we can see, the for logistic model Response (Pain) ~ Age, the P-Value of Likelihood Ratio Test is less than 0.05 which means the model is significant. Besides, the P-value of Wald Chi-Square test for variable Age is also less than 0.05, which means variable Age is significant to explain the change of variable Pain.

Response (Pain) ~ Duration

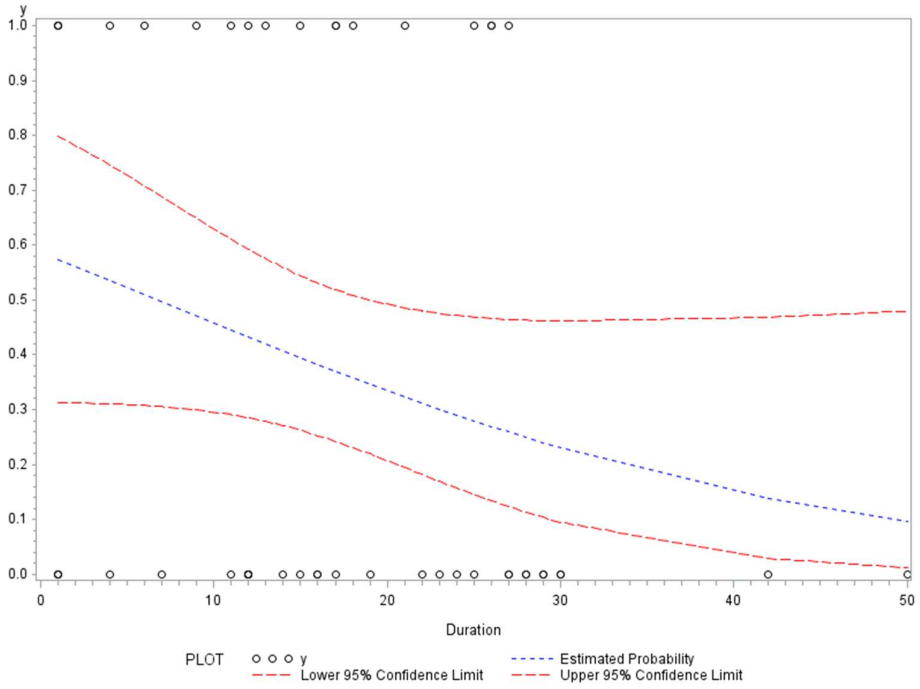
Table 5: Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3.1158	1	0.0775

Table 6: Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.3479	0.5783	0.3618	0.5475
Duration	1	-0.0518	0.0310	2.7983	0.0944

Figure 5: Response vs. Duration with 95% CI by Logistic method



From table 5 and table 6 we can see, the for logistic model Response (Pain) ~ Duration, the P-Value of Likelihood Ratio Test is greater than 0.05 which means the model is not significant. Besides, the P-value of Wald Chi-Square test for variable Duration is also greater than 0.05, which means variable Duration is not significant to explain the change of variable Pain.

After comparing figure 2 & figure4 and figure 3 & figure 5 respectively, given "Yes" category in Pain set to be 1 and "No" category in Pain set to be 0, we can find that although the plots from logistic method are more stable than that of loess method, the trend between response (Pain) and Age/Duration observed in the simple logistic regression with that observed in the local regression analysis are generally consistent. For example, there is a same increasing trend for Pain ~ Age in plots of both two methods and a same decreasing trend for Pain ~ Duration in plots of both two methods.

II. Backward selection

Table 7: Summary of backward selection

Step	Effect Removed	Number in	DF	Wald Chi-Square	Pr> ChiSq
1	Duration	3	1	0.3249	0.5687
2	Sex	2	1	3.6601	0.0557

Table 8: Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	16.5454	3	0.0009

Table 9: Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	6.1859	0.0129
Treatment	2	9.1565	0.0103

Table 10: Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-15.6709	6.5605	5.7059	0.0169
Age	1	0.2422	0.0974	6.1859	0.0129
Treatment: A	1	-2.1494	0.9030	5.6653	0.0173
Treatment: B	1	-3.0845	1.0720	8.2791	0.0040

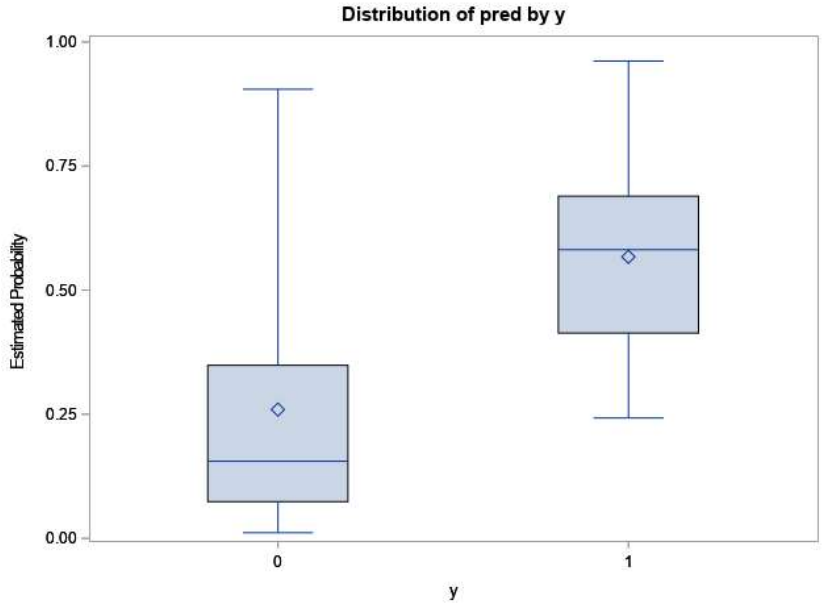
* The AUC score is 0.84.

From table 8, table 9 and table 10, we can see, the for the optimized model select by backward selection, the P-Value of Likelihood Ratio Test is less than 0.05 which means the model is significant. Besides, the P-values of Wald Chi-Square test for both variable Age and Treatment are all less than 0.05, which means variable Age and Treatment are significant to explain the change of variable Pain. Since variable Treatment is a categorical variable, more analysis will be considered. When Treatment P is the reference, both Treatment A and Treatment B are significant.

III. The predictive power of this optimized model

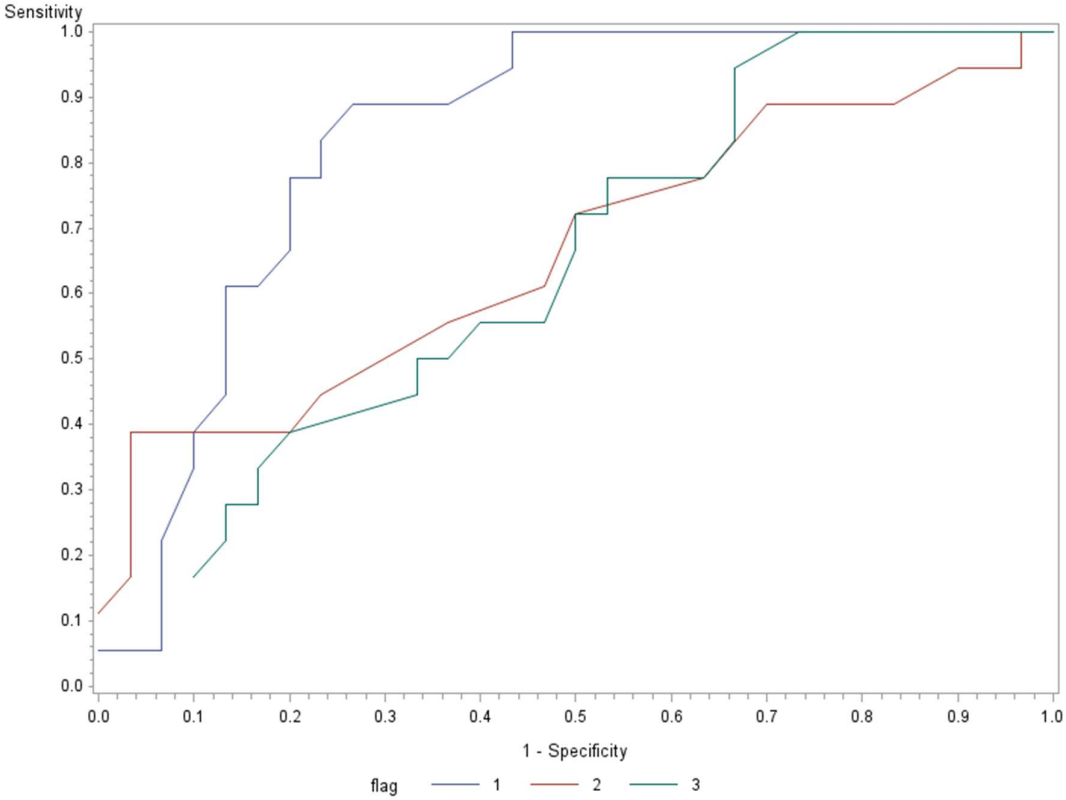
1. Boxplot

Figure 6: The boxplot for predicted probability vs. response



2. Three ROC curves

Figure 7: Three ROC curves

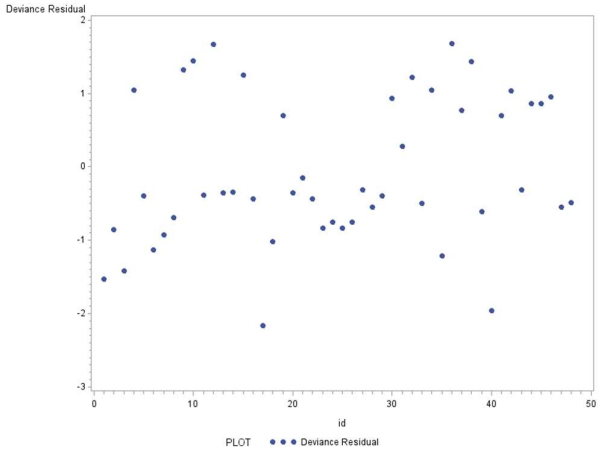


From figure 6 we can see that the boxplots of predicted probabilities for two categories of response variable generally have little overlap, specially, for the majority part greater than 25% quantile and less than 75%. In this way the optimized model has a significant predictive power. From figure 7 we can see that the optimized model has the best ROC curve with AUC score as 0.84 among these three cures which means the model has a strong predictive power.

IV. Perform the following model diagnostics

1. Index plot: residual vs. observation number

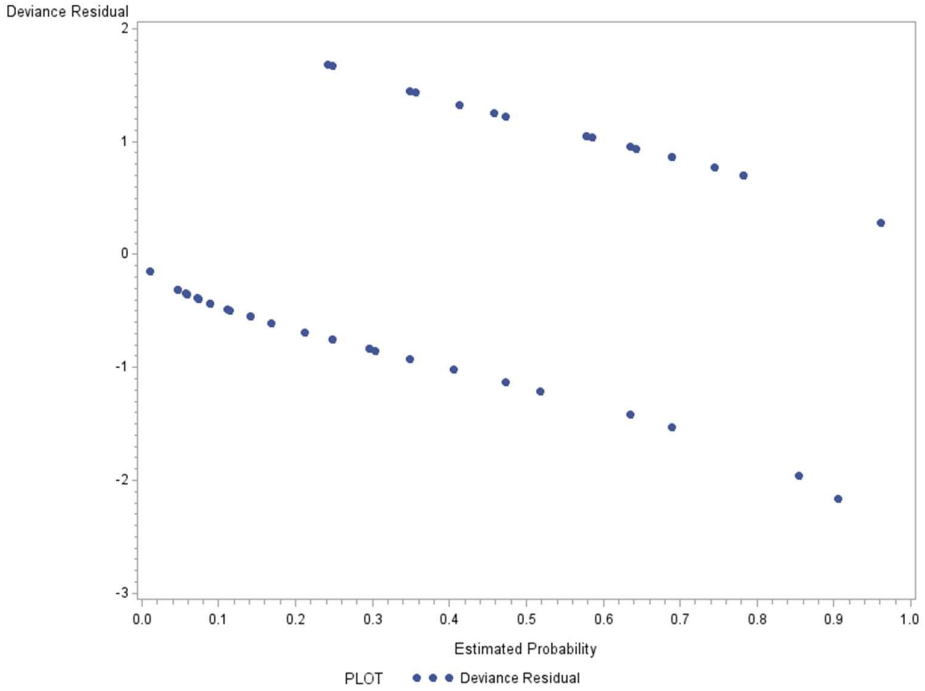
Figure 8: Index plot: residual vs. observation number



From this Index plot, we can see generally the pattern is regular while there might be two outliers for observation 18 and 40. In a word, the model assumptions are appropriate for the data.

2. Plot of residual vs. predicted probability

Figure 9: Plot of residual vs. predicted probability



3. Plot of residual vs. each explanatory variable

Figure 10: Plot of residual vs. Age

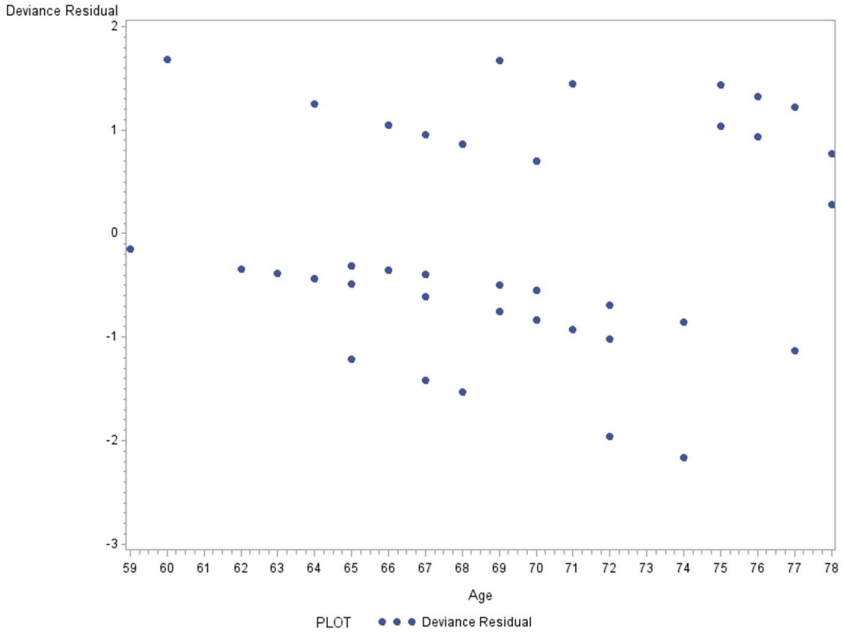
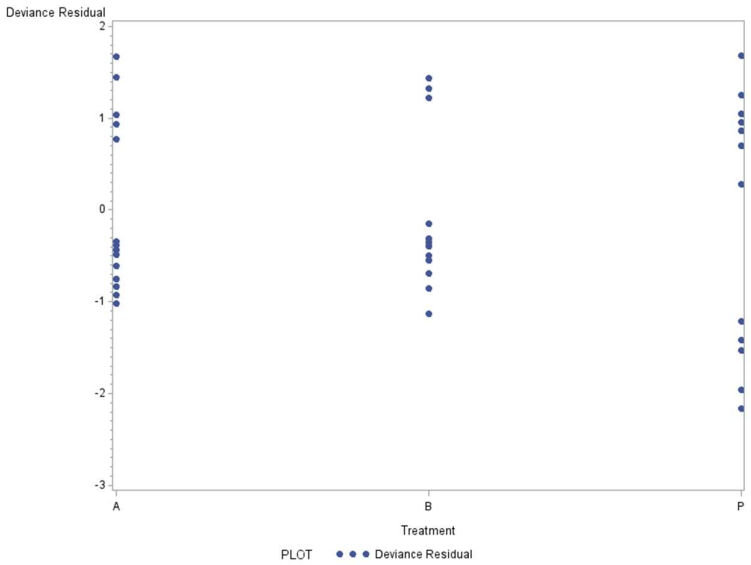


Figure 11: Plot of residual vs. Treatment



Then in figure 9 we can see a smooth regular pattern which means there is no violation for the model assumptions. Also, regular patterns show in figure 10 and 11.

Conclusion:

Based on previous analysis, the trend between response (Pain) and Age/Duration observed in the simple logistic regression with that observed in the local regression analysis are quite consistent. Besides, the good parsimonious model for the logistic regression is the model containing Pain as response variable and Age as well as Treatment as explanatory variable, which has a very strong predictive power. And model assumptions are appropriate for the data.

Appendix:

```
/* Part A */
proc import datafile = 'E:/GW/Textbook/Data Analysis/HW7/HW7a.csv' /*read the file into
sas*/
dbms = csv /*specify the format of the file*/
out=work.HW7a;      /*specify the saved dataset in sas*/
getnames=yes;      /*get the name of the variables from the original file*/
run;    /*run this procedure*/
/* Generate log(T) */
data HW7a;
set HW7a;
logT=log(T);output;
run;
/* Perform GLM model */
proc genmod data=HW7a;
model Y=X Z/offset=logT dist =p;
output out=pout resdev=rs;
run;
/* Plot the QQ plot */
proc univariate data=pout noprint;
var rs;
qqplot rs/normal;
run;

/* Part B */
/* Read the data and save as Neuralgia*/
Data Neuralgia;
        input Treatment $ Sex $ Age Duration Pain $ @@;
        if Pain="Yes" then y=1;
        if Pain="No" then y=0;
        datalines;
    P  F  68   1  No    B  M  74  16  No  P  F  67  30  No    P  M  66  26
Yes  B  F  67  28  No  B  F  77  16  No
    A  F  71  12  No    B  F  72  50  No  B  F  76   9  Yes    A  M  71
17  Yes  A  F  63  27  No  A  F  69  18  Yes
    B  F  66  12  No    A  M  62  42  No  P  F  64   1  Yes    A  F  64
17  No    P  M  74   4  No  A  F  72  25  No
    P  M  70   1  Yes  B  M  66  19  No  B  M  59  29  No    A  F  64
30  No    A  M  70  28  No  A  M  69   1  No
    A  M  70  12  No    A  F  69  12  No  B  F  65  14  No    B  M  70
1  No    B  M  67  23  No  A  M  76  25  Yes
    P  M  78  12  Yes  B  M  77   1  Yes B  F  69  24  No    P  M  66   4
Yes  P  F  65  29  No  P  M  60  26  Yes
    A  M  78  15  Yes  B  M  75  21  Yes A  F  67  11  No    P  F  72  27
```

```
No    P   F   70   13   Yes A   M   75    6   Yes
    B   F   65    7   No    P   F   68   27   Yes P   M   68   11   Yes    P   M   67   17
Yes   B   M   70   22   No   A   M   65   15   No
    ;

/* Loess Procedure */
/* Response Y ~ Age */
proc sort data=Neuralgia;
by Age;
run;
proc loess data=Neuralgia;
   model Y =Age / all    smooth=.6;
ods output Outputstatistics=lofitAge;
run;
goptions reset=all;
symbol1 v=circle i=none c=black;
symbol2 v=none i=join c=blue l=2;
symbol3 v=none i=join c=red l=3;
symbol4 v=none i=join c=red l=3;
proc gplot data=lofitAge;
plot (DepVar pred lowerCL upperCL)*Age / overlay legend;
Run;

/* Response Y ~ Duration */
proc sort data=Neuralgia;
by Duration;
run;
proc loess data=Neuralgia;
   model Y =Duration / all    smooth=.6;
ods output Outputstatistics=lofitDuration;
run;
goptions reset=all;
symbol1 v=circle i=none c=black;
symbol2 v=none i=join c=blue l=2;
symbol3 v=none i=join c=red l=3;
symbol4 v=none i=join c=red l=3;
proc gplot data=lofitDuration;
plot (DepVar pred lowerCL upperCL)*Duration / overlay legend;
Run;

/* Logistic procedure */
/* Response Y ~ Age */
proc sort data=Neuralgia;
by Age;
```

```
run;
proc logistic data=Neuralgia descending;
     model Y =Age;
     output out=loutAge p=pred l=lowerCL u=upperCL;
run;
proc gplot data=loutAge;
plot (y pred lowerCL upperCL)*Age / overlay legend;
run;

/* Response Y ~ Duration */
proc sort data=Neuralgia;
by Duration;
run;
proc logistic data=Neuralgia descending;
     model Y =Duration;
     output out=loutDuration p=pred l=lowerCL u=upperCL;
run;
proc gplot data=loutDuration;
plot (y pred lowerCL upperCL)*Duration / overlay legend;
run;

/* Backward */
proc logistic data=Neuralgia descending;
   class Treatment Sex / param = ref ref=last;
   model Y = Treatment Sex Age Duration / selection=b details;
run;
/* Optimized model selected by backward */
proc logistic data=Neuralgia descending;
     class Treatment/ param = ref ref=last;
     model y = Age Treatment;
run;
/* BoxPlot*/
proc logistic data=Neuralgia descending;
     class Treatment/ param = ref ref=last;
     model y = Age Treatment;
     output out=loutop p=pred resdev=dres;
run;
proc sort data=loutop;
by y;
run;
proc boxplot data=loutop;
plot pred*y;
run;
```

```
/* 3 ROC curves */
proc logistic data=Neuralgia descending;
     class Treatment/ param = ref ref=last;
     model y = Age Treatment / outroc=roc1;
run;
proc logistic data=Neuralgia descending;
     model y = Age / outroc=roc2;
run;
proc logistic data=Neuralgia descending;
     model y = Duration / outroc=roc3;
run;

data roc1;     set roc1;     flag=1;     run;
data roc2;     set roc2;     flag=2;     run;
data roc3;     set roc3;     flag=3;     run;
data roc;     set roc1 roc2 roc3;     run;
goption reset=all;     symbol i=join;
proc gplot data=roc;
plot _sensit_ * _1mspec_ = flag / legend;
run;

/* Residuals Plots*/
proc logistic data=Neuralgia descending;
     class Treatment/ param = ref ref=last;
     model y = Age Treatment;
     output out=loutop p=pred resdev=dres;
run;

data loutop;
set loutop;
id = _N_;
run;
proc sort data=loutop;
by y;
run;
goption reset=all;     symbol v=dot;
proc gplot data=loutop;
plot dres*(id pred Age Treatment)/ legend;
run;
```