HW 8 Report

Yongchao Qiao

Part A

I Logistic regression

1. Y ~ X

Table 1: Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.5735	1	0.4489

Table 2: Analysis of Maximum Likelihood Estimates

			Standard	Wald	
Parameter	DF	Estimate	Error	Chi-Square	Pr > ChiSq
Intercept	1	-1.6001	0.5031	10.1147	0.0015
X	1	0.00341	0.00452	0.5704	0.4501

From table 1 and table 2 we can see, for the logistic model Y ~ X, 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 X is also greater than 0.05, which means variable X is not significant to explain the change of variable Y.

2. Y ~ Z

Table 3: Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	15.2029	1	<0.0001

Table 4: Analysis of Maximum Likelihood Estimates

			Standard	Wald	
Parameter	DF	Estimate	Error	Chi-Square	Pr > ChiSq
Intercept	1	0.8507	0.6127	1.9274	0.1650
Z	1	-0.2147	0.0597	12.9404	0.0003

From table 3 and table 4 we can see, for the logistic model Y ~ Z, 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 Z is also less than 0.05, which means variable Z is significant to explain the change of variable Y.

3. Y ~ X & Z

Table 5: Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	16.7329	2	0.0002

Table 6: Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.3855	0.7184	0.2880	0.5915
X	1	0.00642	0.00528	1.4742	0.2247
Z	1	-0.2276	0.0622	13.3723	0.0003

From table 5 and table 6 we can see, for the logistic model Y ~ X Z, 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 Z is also less than 0.05, which means variable Z is significant to explain the change of variable Y. However, the P-value of Wald Chi-Square test for variable X is greater than 0.05, which means variable X is not significant to explain the change of variable Y. This model result is consistent with that of previous single logistic models.

II GAM (For logistic regression)

1. Degree freedom: 6

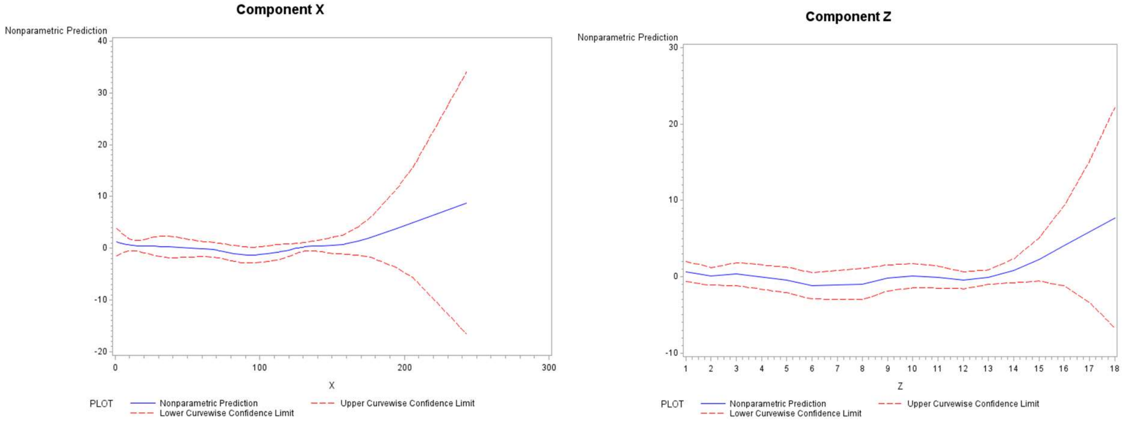
Table 7: Regression Model Analysis Parameter Estimates

Parameter	Estimate	Standard Error	t-value	Pr > |t|
Intercept	0.16019	0.86091	0.19	0.8530
Linear(X)	-0.01186	0.00758	-1.56	0.1228
Linear(Z)	0.17848	0.07811	2.28	0.0255

Table 8: Smoothing Model Analysis

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(X)	5	10.068597	10.0686	0.0733
Spline(Z)	5	10.515543	10.5155	0.0619

Figure 1: Components plot with 95% CI



From table 7 and table 8 we can see, for the GAM model, the P-Values of linear component of variables X, non-linear component of variable X and non-linear component of variable Z are greater than 0.05 which means the linear component of variable X, non-linear component of variable X and non-linear component of variable Z are not significant to explain the change of variable Y. However, the P-value of linear component of variable Z is less than 0.05, which means the linear component of variable Z is significant to explain the change of variable Y.

2. Degree freedom: 4

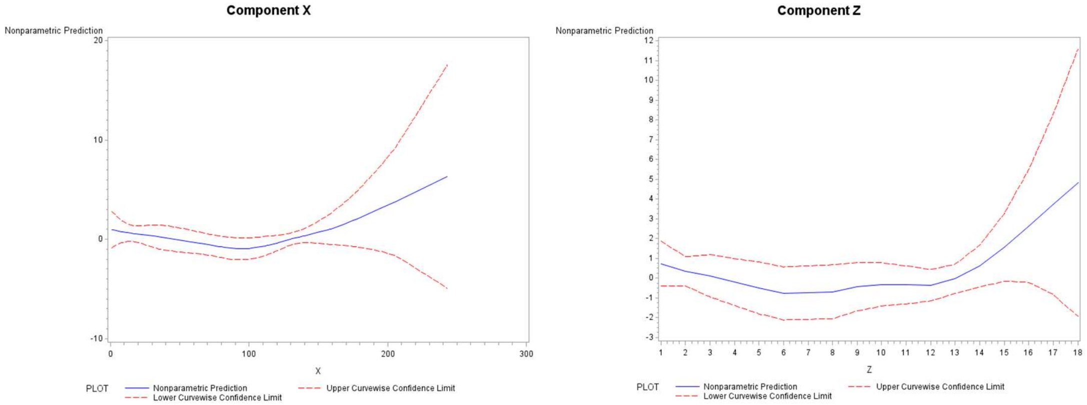
Table 9: Regression Model Analysis Parameter Estimates

Parameter	Estimate	Standard Error	t-value	Pr > |t|
Intercept	-0.01234	0.85973	-0.01	0.9886
Linear(X)	-0.01032	0.00733	-1.41	0.1635
Linear(Z)	0.18874	0.07569	2.49	0.0150

Table 10: Smoothing Model Analysis

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(X)	3	8.410317	8.4103	0.0383
Spline(Z)	3	7.577227	7.5772	0.0556

Figure 2: Components plot with 95% CI



From table 9 and table 10 we can see, for the GAM model, the P-Values of linear component of variables X and non-linear component of variable Z are greater than 0.05 which means the linear component of variable X and non-linear component of variable Z are not significant to explain the change of variable Y. However, the P-values of linear component of variable Z and non-linear component of variable X are less than 0.05, which means the linear component of variable Z and non-linear component of variable X are significant to explain the change of variable Y.

3. Degree freedom: 2

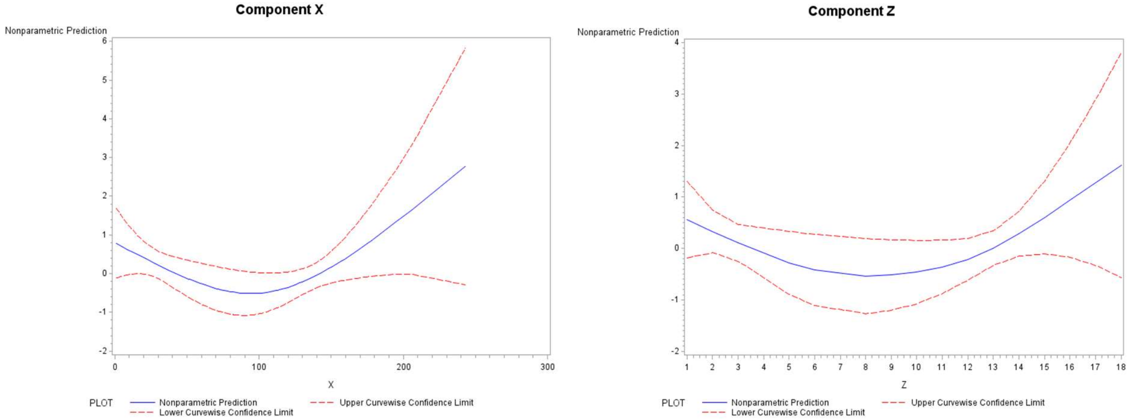
Table 11: Regression Model Analysis Parameter Estimates

Parameter	Estimate	Standard Error	t-value	Pr > |t|
Intercept	-0.36416	0.77872	-0.47	0.6414
Linear(X)	-0.00675	0.00626	-1.08	0.2845
Linear(Z)	0.21605	0.06916	3.12	0.0025

Table 12: Smoothing Model Analysis

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(X)	1	6.882906	6.8829	0.0087
Spline(Z)	1	4.361982	4.3620	0.0367

Figure 3: Components plot with 95% CI



From table 11 and table 12 we can see, for the GAM model, the P-Value of linear component of variables X is greater than 0.05 which means the linear component of variable X is not significant to explain the change of variable Y. However, the P-values of linear component of variable Z, non-linear component of variable X and non-linear component of variable Z are less than 0.05, which means the linear component of variable Z, non-linear component of variable X and non-linear component of variable Z are significant to explain the change of variable Y.

III Conclusion

Based on the results from logistic regression analysis, we can see there is an association between the response variable Y and explanatory variable Z.

Based on the results from generalized additive model analysis, we can see that with degree freedom as 6, there is only an association between response variable Y and explanatory variable Z and the association pattern is linear pattern. With degree freedom as 4, there is an association between response variable Y and explanatory variable X as well as the explanatory variable Z and the association between variable Y and variable X is a non-linear pattern while the association between variable Y and variable Z is a linear pattern. With degree freedom as 2, there is an association between response variable Y and explanatory variable X as well as the explanatory variable Z and the association between variable Y and variable X is a non-linear pattern while the association between variable Y and variable Z contains bother linear pattern and non-linear pattern.

Therefore, from the results of three degree freedoms, the third one with degree freedom as 2 shows most significant association patterns among variable Y and variable X & Z. So, the degree freedom as 2 is a better choice.

Part B

I GLM (Poisson regression)

Table 13: Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	45	65.7642	1.4614

Table 14: Analysis of Maximum Likelihood Parameter Estimates

				Wald 95% Confidence		Wald Chi-	Pr >
Parameter	DF	Estimate	Standard Error	Limits		Square	Chisq
Intercept	1	1.3311	0.2358	0.8690	1.7933	31.87	<.0001
X	1	0.0038	0.0371	-0.0690	0.0766	0.01	0.9191
Z	1	0.0210	0.0241	-0.0262	0.0682	0.76	0.3833
Scale	0	1.2089	0.0000	1.2089	1.2089		

From table 13 we can see, the deviance is 83.92 which is less than its degree as 97, so the effect of over dispersion should be considered. Then for parameters in table 14, the P-Values of variable X and variable Z are greater than 0.05 which means variable X and variable Z are not significant to explain the change of variable Y.

II GAM (Poisson regression)

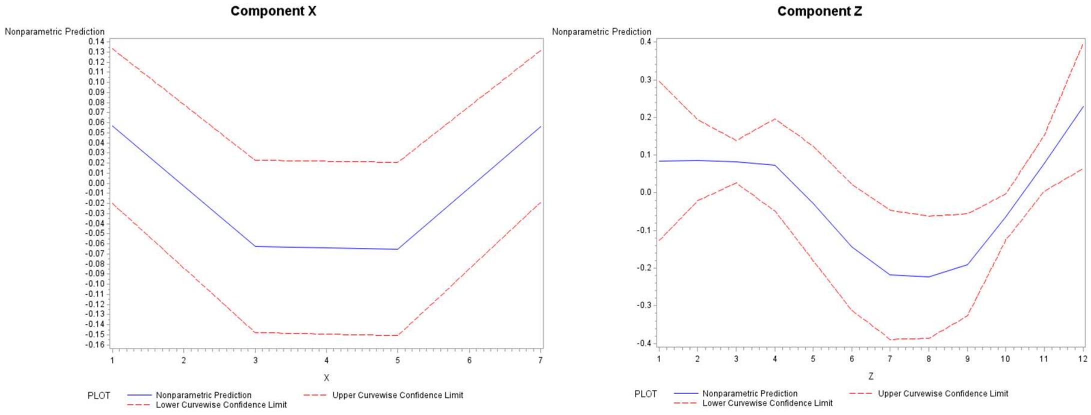
Table 15: Regression Model Analysis Parameter Estimates

Parameter	Estimate	Standard Error	t-value	Pr > |t|
Intercept	1.33319	0.18871	7.06	<0.0001
Linear(X)	0.00378	0.03000	0.13	0.9004
Linear(Z)	0.02248	0.01895	1.19	0.2419

Table 16: Smoothing Model Analysis

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Loess(X)	0.40487	1.877409	1.8774	.
Loess(Z)	1.08087	7.587817	7.8578	0.0058

Figure 3: Components plot with 95% CI



From table 15 and table 16 we can see, for the GAM model, only the P-Value of non-linear component of variables Z is less than 0.05 which means only the non-linear component of variable Z is significant to explain the change of variable Y. Since there is no p-value for the non-linear component of variable X, we need to only contain the non-linear component of variable Z and the result is shown below. From table 17 and table 18 we can see, for the GAM model, only the P-Value of non-linear component of variables Z is less than 0.05 which means only the non-linear component of variable Z is significant to explain the change of variable Y.

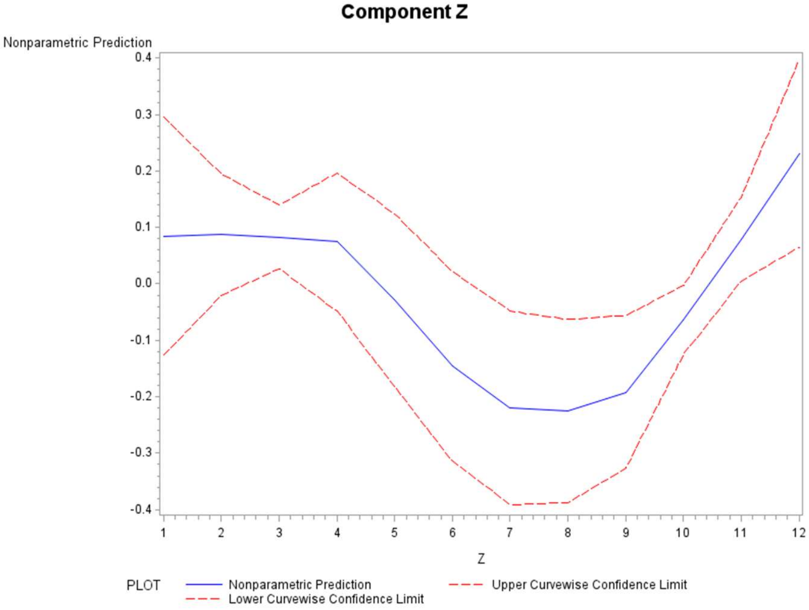
Table 17: Regression Model Analysis Parameter Estimates

Parameter	Estimate	Standard Error	t-value	Pr > |t|
Intercept	1.33145	0.19053	6.99	<0.0001
X	0.00377	0.03072	0.12	0.9028
Linear(Z)	0.02248	0.01894	1.19	0.2416

Table 18: Smoothing Model Analysis

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Loess(Z)	1.07917	7.896429	7.8964	0.0057

Figure 4: Components plot with 95% CI



III Conclusion

Based on the results from GLM Poisson regression analysis, we can see there is no association between the response variable Y and explanatory variable X as well as variable Z.

Based on the results from generalized additive model analysis, we can see that with gcv as the method, there is only an association between response variable Y and explanatory variable Z and the association pattern is non-linear pattern.

Appendix

```
/* Part A */
proc import datafile = 'E:/GW/Textbook/Data Analysis/HW8/HW8a.csv' /*read the file into
sas*/
dbms = csv /*specify the format of the file*/
out=work.HW8a;      /*specify the saved dataset in sas*/
getnames=yes;      /*get the name of the variables from the original file*/
run;    /*run this procedure*/
/* Logistic Regression */
/* Y ~ X */
proc logistic data=HW8a desc;
  model Y = X;
run;

/* Y ~ Z */
proc logistic data=HW8a desc;
  model Y = Z;
run;

/* Y ~ X Z */
proc logistic data=HW8a desc;
  model Y = X Z;
run;

/* GAM */
/* Degree 6 */
proc gam data=HW8a;
    model Y = spline(X, df=6) spline(Z, df=6) / dist=binary;
    output out=gamout all;
run;
/* Plot the component plot for each variable*/
proc sort data=gamout; by X; run;
goptions reset=all;
title "Component X";
symbol1 i=join v=none l=1 c=blue;
symbol2 i=join l=3 c=red;
symbol3 i=join l=3 c=red;
proc gplot data=gamout;
    plot (p_X uclm_X lclm_X)*X /overlay legend;
run;proc sort data=gamout; by Z; run;
title "Component Z";
proc gplot data=gamout;
    plot (p_Z uclm_Z lclm_Z)*Z /overlay legend;
```

```
run;

/* Degree 4 */
proc gam data=HW8a;
   model Y = spline(X, df=4) spline(Z, df=4) / dist=binary;
   output out=gamout all;
run;
/* Plot the component plot for each variable*/
proc sort data=gamout; by X; run;
goptions reset=all;
title "Component X";
symbol1 i=join v=none l=1 c=blue;
symbol2 i=join l=3 c=red;
symbol3 i=join l=3 c=red;
proc gplot data=gamout;
   plot (p_X uclm_X lclm_X)*X /overlay legend;
run;proc sort data=gamout; by Z; run;
title "Component Z";
proc gplot data=gamout;
   plot (p_Z uclm_Z lclm_Z)*Z /overlay legend;
run;

/* Degree 2 */
proc gam data=HW8a;
   model Y = spline(X, df=2) spline(Z, df=2) / dist=binary;
   output out=gamout all;
run;
/* Plot the component plot for each variable*/
proc sort data=gamout; by X; run;
goptions reset=all;
title "Component X";
symbol1 i=join v=none l=1 c=blue;
symbol2 i=join l=3 c=red;
symbol3 i=join l=3 c=red;
proc gplot data=gamout;
   plot (p_X uclm_X lclm_X)*X /overlay legend;
run;proc sort data=gamout; by Z; run;
title "Component Z";
proc gplot data=gamout;
   plot (p_Z uclm_Z lclm_Z)*Z /overlay legend;
run;


/* Part B */
```

```
proc import datafile = 'E:/GW/Textbook/Data Analysis/HW8/HW8b.csv' /*read the file into
sas*/
dbms = csv /*specify the format of the file*/
out=work.HW8b;      /*specify the saved dataset in sas*/
getnames=yes;      /*get the name of the variables from the original file*/
run;    /*run this procedure*/
/* GLM (Poisson regression) */
proc genmod data=HW8b;
model Y=X Z/dist=p scale=d;
run;
/* GAM */
proc gam data=HW8b;
   model Y= loess(X) loess(Z) / dist=poisson method=gcv;
   output out=gamout all;
run;

/* Plot the component plot for each variable*/
proc sort data=gamout; by X; run;
goptions reset=all;
title "Component X";
symbol1 i=join v=none l=1 c=blue;
symbol2 i=join l=3 c=red;
symbol3 i=join l=3 c=red;
proc gplot data=gamout;
   plot (p_X uclm_X lclm_X)*X /overlay legend;
run;proc sort data=gamout; by Z; run;
title "Component Z";
proc gplot data=gamout;
   plot (p_Z uclm_Z lclm_Z)*Z /overlay legend;
run;
/* Only non-linear conponent of variable Z */
proc gam data=HW8b;
   model Y= param(X)loess(Z) / dist=poisson method=gcv;
   output out=gamout all;
run;
/* Plot the component plot for each variable*/
proc sort data=gamout; by Z; run;
title "Component Z";
symbol1 i=join v=none l=1 c=blue;
symbol2 i=join l=3 c=red;
symbol3 i=join l=3 c=red;
proc gplot data=gamout;
   plot (p_Z uclm_Z lclm_Z)*Z /overlay legend;
run;
```