The George Washington University
6210,2018 spring

Report of Homework8

Jiaying Liu

April 10, 2018

1 Part A

1.1 dataset

The data set contain one response variable y and two explanatory variable x and z.Note that the variable y is categorical variable which only have two value 0 and 1. We are going to fit this data set by the model

Table 1: Original Data set.

Obs	x	z	y
1	148	16	0
2	18	2	0
3	1	12	0
4	243	8	0
5	168	18	0
6	1	16	0
7	78	15	0
8	175	13	0
9	80	16	0
10	27	9	0
...	...	...	

1.2 Logistical regression

Since the variable y is only two response levels 1 and 2, so we can use the logistical regression to fit the model.For binary response models, the response y, of an individual or an experimental unit can take on one of two possible values, denoted for convenience by 1 and

2 Suppose is a vector of explanatory variables and is the response probability to be modeled. The linear logistic model has the form

$$logit(\pi) = log(\pi/1-\pi) = \alpha + \beta * x$$

From the table 1, it show the estimate result of logistic regression. We can get the formula

$$logit(\pi) = log(\pi/1-\pi) = 0.3855 + 0.00642 * x - 0.2276 * z$$

However when we check the p value of x and z, we find the p value of x is greater than 0.05 which mean the x variable is not significant to variable y. But when we check the p value of z, it is smaller than 0.05, so we can said the variable z have significant impact on y.

Table 2: Analysis of Maximum Likelihood Estimates.

Source	**DF**	**Estimate**	**SE**	**Wald chi-square**	**P > 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

And then we plot the component plot. From the figure 1 and figure 2, it show the relation between predict value, lower confident interval, the upper confident interval and variable x or z.From the plot we can see the plot is not that smooth. So we need to find another way to fit a more smoother plot.

Figure 1: Scatter Plot of x vs. estimated probability

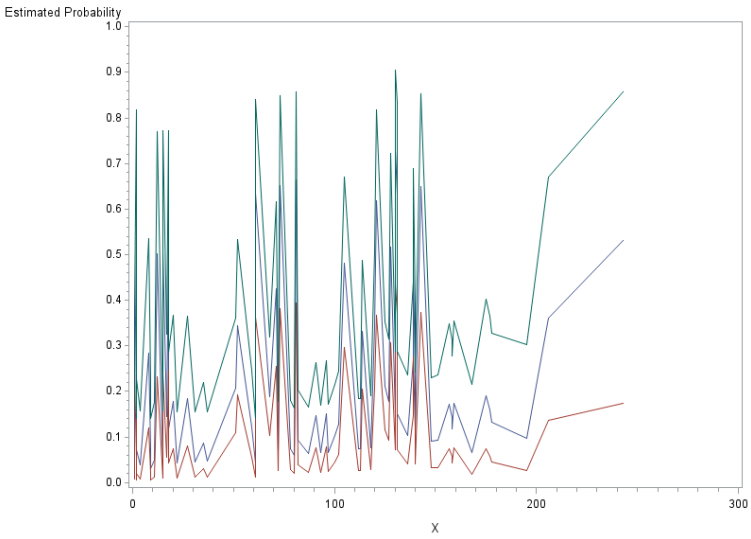
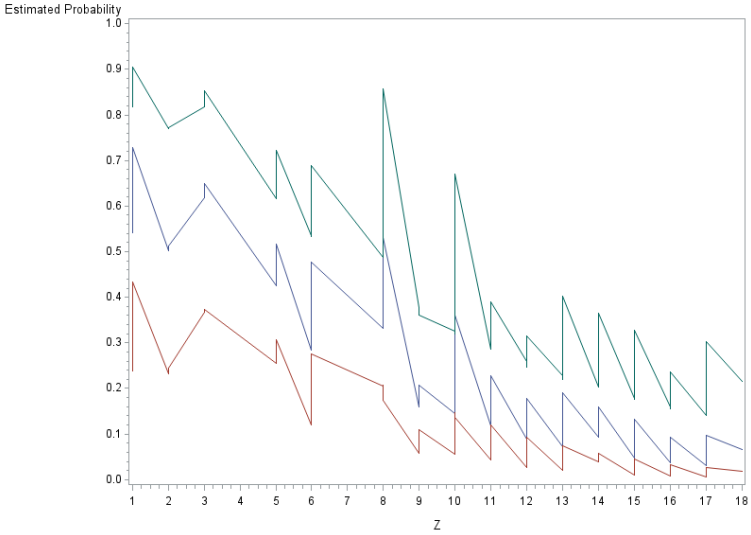


Figure 2: Scatter Plot of z vs. estimated probability



1.3 Generalized additive model(for logistic regression

Alternatively,we can fit the generalized additive model under the logistics regression.

1.3.1 df=2

Setting the degree freedom equal to two, we get following result. From the table 3 we got the result of model estimate. Since we use the generalized additive model.We can see that both p value of x variable and z variable are smaller than 0.05. So we can conclude that under the degree freedom of 2, both spline(x) and spline(z) are significant.

Table 3: Smoothing Model Analysis.

Source	**DF**	**Sum of Squares**	**chi-square**	**P > ChiSq**
Spline(X)	1.00000	6.882906	6.8829	0.0087
Spline(Z)	1.00000	4.361982	4.3620	0.0367

Figure 3: Scatter Plot of x vs. Nonparametric Prediction

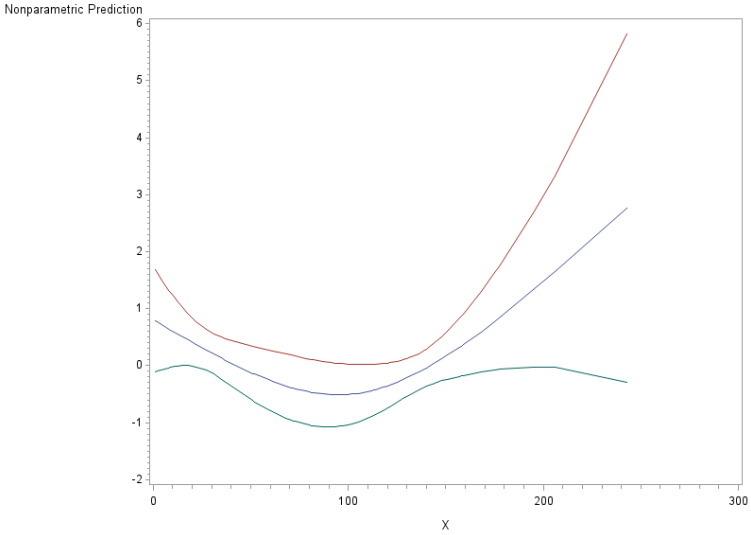
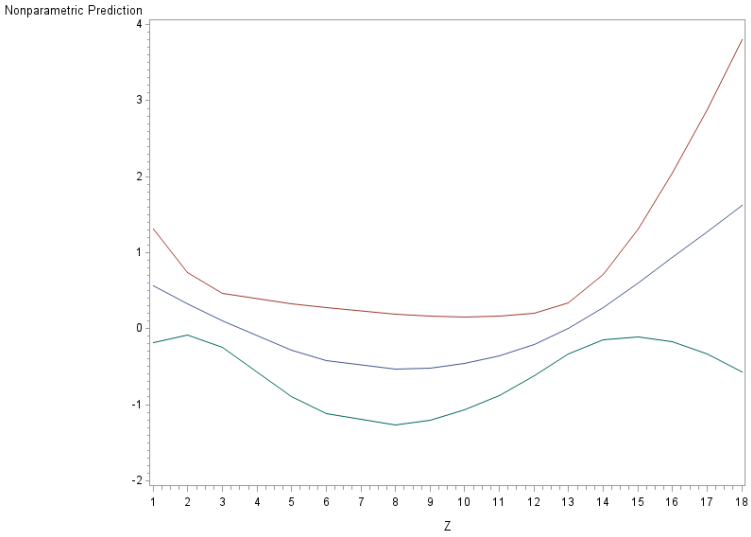


Figure 4: Scatter Plot of z vs.Nonparametric Prediction



1.3.2 df=4

Setting the degree freedom equal to four, we get following result.From the table 4 we got the result of model estimate.However when the df=4, We can see that only x variable's p value is smaller than 0.05.And the variable z's pvalue become greater than 0.05. It is no longer significant.

Table 4: Smoothing Model Analysis.

Source	**DF**	**Sum of Squares**	**chi-square**	**P > ChiSq**
Spline(X)	3.00000	8.410317	8.4103	0.0383
Spline(Z)	3.00000	7.577227	7.5772	0.0556

Figure 5: Scatter Plot of x vs. Nonparametric Prediction

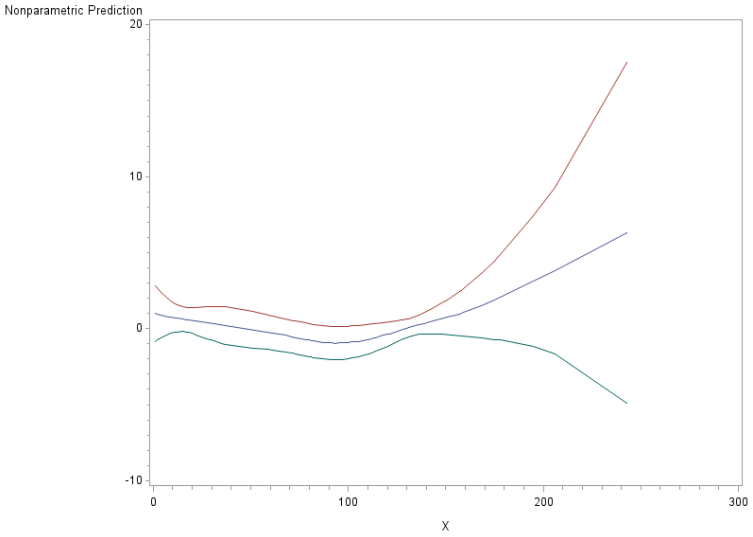
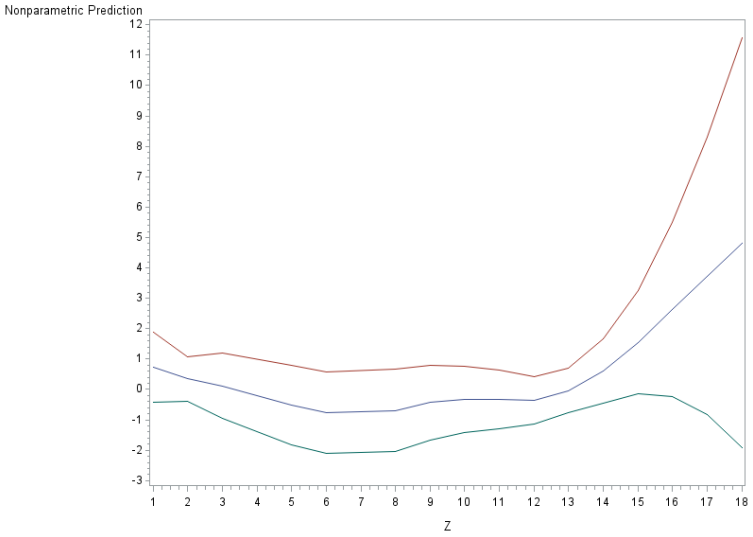


Figure 6: Scatter Plot of z vs.Nonparametric Prediction



1.3.3 df=6

Setting the degree freedom equal to four, we get following result.From the table 4 we got the result of model estimate.However when the df=6, We can see both x variable's p value and z variable's p value are greater than 0.05. They are no significant to variable y. We can conclude that with the df=6, the model is not useful.

Table 5: Smoothing Model Analysis.

Source	**DF**	**Sum of Squares**	**chi-square**	**P > ChiSq**
Spline(X)	5.00000	10.068597	10.0686	0.0733
Spline(Z)	5.00000	10.515543	10.5155	0.0619

Figure 7: Scatter Plot of x vs. Nonparametric Prediction

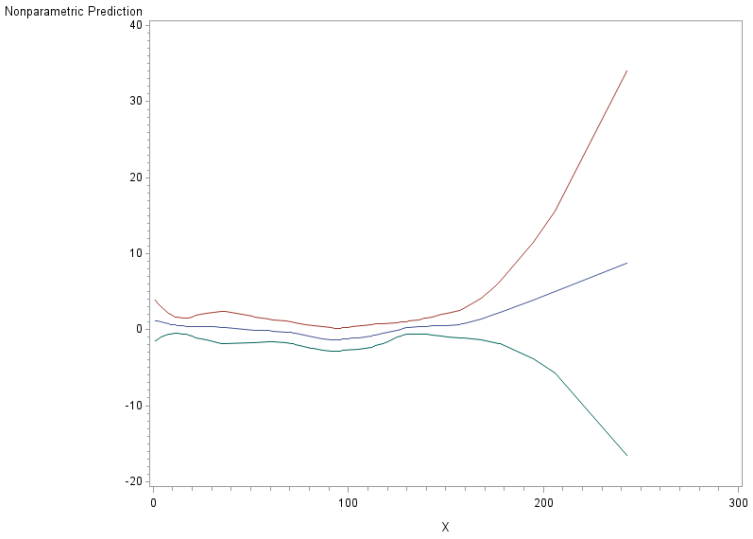
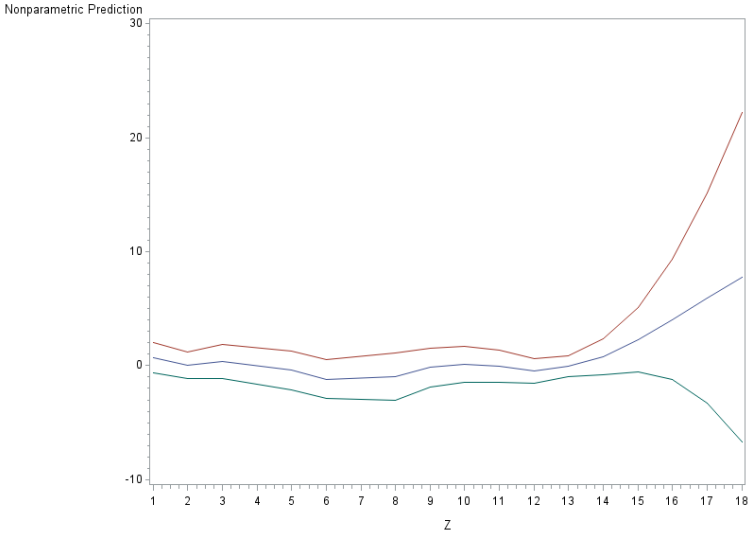


Figure 8: Scatter Plot of z vs.Nonparametric Prediction



1.4 conclusion

From above analysis, by analyze the p value of variable under different degree of freedom. We can say that the model under df=2 is the best model.

2 Part B

The dataset contain one response variable y and two explanatory variable x and z. We assume that y follows a poisson distribution with mean λ and $log(\lambda) = a + bx + cz$

2.1 dataset

Table 6: Original Data set.

Obs	x	z	y
1	5	12	10
2	3	12	3
3	1	12	10
4	7	12	6
5	5	11	1
6	3	11	6
7	1	11	4
8	7	11	6
9	5	10	5
10	3	10	3
...	...	...	

2.2 Generalized linear model(Poisson regression)

First we fit the data by generalized linear model under Poisson regression. Before fit the model, we need to check the whether there are over dispersion effect.The table 7 display the result of assessing goodness of fit. We can see the pearson chi-square's value is 1.4716 and the scaled oearson x2's value is 1.000. It is obviously that the data contains over-dispersion effect. However, after scaling The effect was eliminate. So we can fit the model after scaling. And the result was show in the table 8.

$$log(\mu) = 1.3311 + 0.0038X + 0.021Z$$

which $log(\mu)$ represent the expectation of y. We can see both p value of variable x and z is greater than 0.05 which mean they are not significant to the variable y.

Table 7: Criteria for assessing goodness of fit.

Criterion	DF	Value	Value/DF
Deviance	45	65.7642	1.4614
Scaled Deviance	45	44.6903	0.9931
Pearson Chi-Square	45	66.2200	1.4716
Scaled Pearson X2	45	45.0000	1.0000

Table 8: Analysis Of Maximum Likelihood Parameter Estimates.

Parameter	DF	Estimate	SE	WALD CI		WALD chisquare	P value
Intercept	1	1.3311	0.2366	0.8674	1.7949	31.65	<.0001
X	1	0.0038	0.0373	-0.0693	0.0768	0.01	0.9193
Z	1	0.0210	0.0242	-0.0264	0.0684	0.75	0.3850
Scale	0	1.2131	0.0000	1.2131	1.2131		

2.3 Generalized Additive Model(For Poisson Regression

Alternative, we can use generalized additive model to better fit and smooth the data. We can use loess as scatter plot smoother and fit the model.Let set the model is under Poisson. From the table 9,displaying the analysis of deviance. We can see the loess(x)'s pvalue is 0.2921 greater than 0.05. So x is not significant to y. But the loess(z)'s p value is 0.0218 smaller than 0.05. So the variable z is significant to variable y. And let check the component plots of variable x and variable z.(figure 9 and 10)

Table 9: Smoothing Model Analysis.

Source	DF	Sum of Squares	chi-square	P > ChiSq
Loess(X)	2.04216	2.516293	2.5163	0.2921
Loess(Z)	2.89001	9.435180	9.4352	0.0218

Figure 9: Scatter Plot of x vs. Nonparametric Prediction

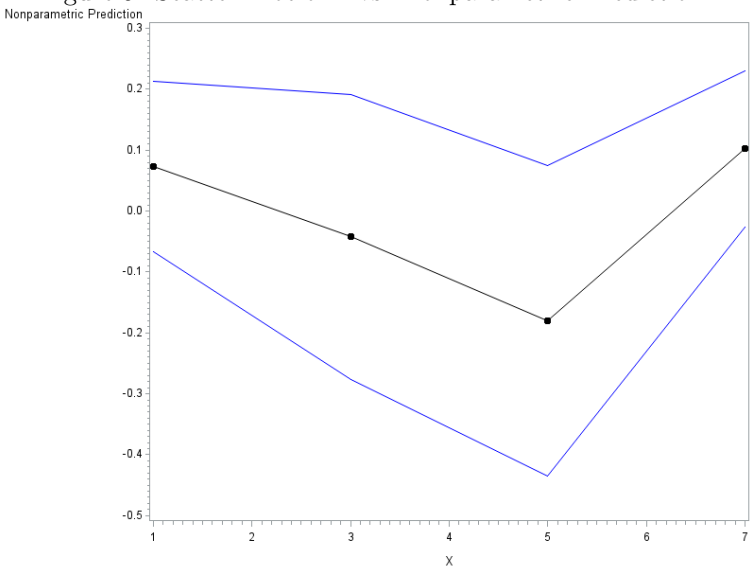


Figure 10: Scatter Plot of z vs.Nonparametric Prediction