An empirical comparison of Bianry&Probit&Logit regression

Zhang Zhiyuan 15220162202517

April 11, 2019

1 Introduction

People smoke all over the world, and some working areas enforce a smoke ban to save nonsmokers from second-hand smoke. Some people think that smoking ban will not only bring health benefits, but also reduce the number of smokers in that it prevent smokers from smoke to a certain extent. And in this article I use three different regression methods to capture the effect of smoke ban on the number of smokers, and further compare their performances.

The data is a cross-sectional data set[1] with observations on 10,000 indoor workers, which is a subset of a 18,090-observation data set collected as part of the National Health Interview Survey in 1991. And I divide it into a 8000&2000 combination separately for testing and forcasting.

2 Regression[2]

For the regression, I use "smoker" as the dependent variable, "smkban", "age", "hsdrop", "hsgrad", "colsome", "colgrad", "black" and "female" as independent variable. And all the three regressions share the same variable set.

2.1 Binary regression[3]

For the binary regression:

regress smoker smkban age hsdrop hsgrad colsome colgrad black female

[1] see appendix 1 for detailed information of dataset

[2] see appendix 2 for detailed description of variables

[3] you can find the link to download all the codes and the dataset in the appendix 3

Source	SS	df	MS	Number of obs	=	8,000
				F(8, 7991)	=	50.01
Model	68.7484853	8	8.59356067	Prob > F	=	0.0000
Residual	1373.15539	7,991	.171837741	R-squared	=	0.0477
				Adj R-squared	=	0.0467
Total	1441.90388	7,999	.180260517	Root MSE	=	.41453

smoker	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
smkban	-.0415469	.0097331	-4.27	0.000	-.0606263	-.0224675
age	-.0007811	.000385	-2.03	0.043	-.0015358	-.0000263
hsdrop	.2575922	.0206991	12.44	0.000	.2170165	.2981678
hsgrad	.2157892	.0163891	13.17	0.000	.1836624	.2479161
colsome	.1461983	.0166569	8.78	0.000	.1135464	.1788502
colgrad	.0354787	.0175453	2.02	0.043	.0010853	.0698722
black	-.0195565	.0177263	-1.10	0.270	-.0543047	.0151916
female	-.0308917	.0094898	-3.26	0.001	-.0494941	-.0122893
_cons	.1694339	.0225216	7.52	0.000	.1252857	.213582

2.2 Probit regression

For the probit regression:

probit smoker smkban age hsdrop hsgrad colsome colgrad black female

Probit regression

Number of obs	=	8,000	
LR chi2(8)	=	407.92	
Prob > chi2	=	0.0000	
Log likelihood = -4166.2652	Pseudo R2	=	0.0467

smoker	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
smkban	-.1407694	.0325567	-4.32	0.000	-.2045793	-.0769595
age	-.0023167	.0013001	-1.78	0.075	-.0048648	.0002314
hsdrop	.9417677	.0757855	12.43	0.000	.7932308	1.090305
hsgrad	.8312891	.0651962	12.75	0.000	.7035069	.9590712
colsome	.6228782	.0664485	9.37	0.000	.4926415	.7531149
colgrad	.1973913	.0719083	2.75	0.006	.0564537	.338329
black	-.0576871	.0595988	-0.97	0.333	-.1744985	.0591244
female	-.1028728	.0322451	-3.19	0.001	-.166072	-.0396736
_cons	-1.089706	.0833228	-13.08	0.000	-1.253016	-.9263963

2.3 Logit regression

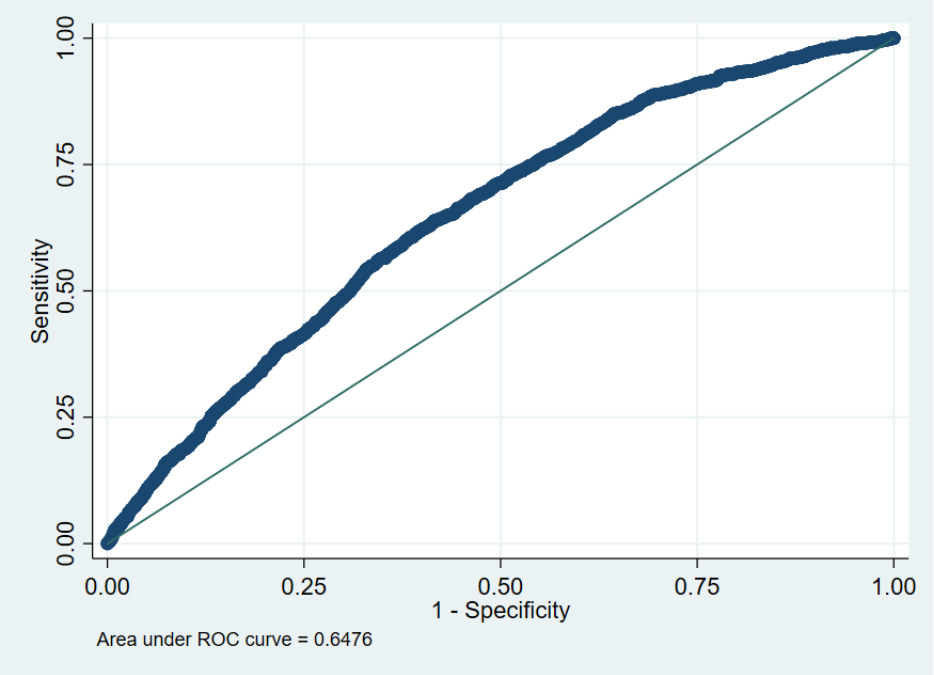
For the logit regression:

logit smoker smkban age hsdrop hsgrad colsome colgrad black female
##

lroc

```
Logistic regression                             Number of obs     =      8,000
                                                LR chi2(8)        =     407.74
                                                Prob > chi2       =     0.0000
Log likelihood = -4166.3542                     Pseudo R2         =     0.0467
```

smoker	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
smkban	-.2325941	.0554123	-4.20	0.000	-.3412003	-.1239879
age	-.0041837	.0021994	-1.90	0.057	-.0084945	.0001272
hsdrop	1.66565	.1400377	11.89	0.000	1.391181	1.940119
hsgrad	1.487921	.1253793	11.87	0.000	1.242182	1.73366
colsome	1.13577	.1277366	8.89	0.000	.8854106	1.386129
colgrad	.3787724	.1395671	2.71	0.007	.105226	.6523188
black	-.1098985	.1022587	-1.07	0.283	-.3103219	.0905249
female	-.1784903	.0551325	-3.24	0.001	-.286548	-.0704327
_cons	-1.875903	.1530701	-12.26	0.000	-2.175915	-1.575891



3 Testing

After the regression, I use the left 2000-observation sample to do the fitting

test.

3.1 Binary

For the binary:

gen prediction_binary=-.0415469*smkban-.0007811*age+.2575922*hsdrop+.2157892*hsgrad+.1461983 .0195565*black-.0308917*female+ .1694339

gen predictedsmoker_binary = 0

replace predictedsmoker_binary = 1 if prediction_binary >= 0.5

gen error_binary = abs (smoker- predictedsmoker_binary)

sum error_binary

Variable	Obs	Mean	Std. Dev.	Min	Max
error_binary	2,000	.268	.4430284	0	1

3.2 Probit

For the probit:

gen prediction_probit= -.1407694*smkban-.0023167*age+.9417677*hsdrop+.8312891*hsgrad+.6228782 .0576871*black-.1028728*female-1.089706

egen prediction_probit_pr=std(prediction_probit)

gen predictedsmoker_probit=0

replace predictedsmoker_probit_pr=1 if prediction_probit>=0.5

gen error_probit=abs(smoker - predictedsmoker_probit)

sum error_probit

Variable	Obs	Mean	Std. Dev.	Min	Max
error_probit	2,000	.369	.4826546	0	1

3.3 Logit

For the logit:

gen prediction_logit=-.2325941*smkban-.0041837*age+ 1.66565*hsdrop+1.487921*hsgrad+ 1.13577*colsome+.3787724*colgrad -.1098985*black -.1784903*female-1.875903

gen prediction_logit_pr=(1+exp(-1*prediction_logit))^-1

gen predictedsmoker_logit=0

replace predictedsmoker_logit=1 if prediction_logit_pr >=0.5

gen error_logit=abs(smoker - predictedsmoker_logit)

sum error_logit

Variable	Obs	Mean	Std. Dev.	Min	Max
error_logit	2,000	.268	.4430284	0	1

4 Comparison

From the above results we can see, in the regresson part, all the three regressions showed similiar significance and R-square, and they also demonstrate valid economic rules. Smoking ban does have its positive effect in reducing the number of smokers, and the education variables of higher level has smaller coefficient, which means highly educated employees are less likely to smoke. But in the testing part, we can see that both binary and logit regression model made 26.8% wrong predictions, while the probit regession model made 36.9% wrong predictions. But it does not mean binary and logit are better classification method than the probit, it simply means that the probit are less appropriate in this very specific empirical application. And when we do real classification tasks, we should also consider adequate candidate method, compare them to find the optimal solution.

5 Appedix

5.1 Information for dataset:

Smoking is a cross-sectional data set with observations on 10,000 indoor workers, which is a subset of a 18,090-observation data set collected as part of the National Health Interview Survey in 1991 and then again (with different respondents) in 1993. The data set contains information on whether individuals were, or were not, subject to a workplace smoking ban, whether or not the individuals smoked and other individual characteristics. These data were provided by Professor William Evans of the University of Maryland and were used in his paper with Matthew Farrelly and Edward Montgomery "Do Workplace Smoking Bans Reduce Smoking?" American Economic Review, September 1999, Vol. 89, No. 4, 728-747.

5.2 Description of variables

---

smoker: =1 if current smoker, =0 otherwise
smkban: =1 if there is a work area smoking ban, =0 otherwise
age: age in years
hsdrop: =1 if high school dropout, =0 otherwise
hsgrad: =1 if high school graduate, =0 otherwise

colsome: =1 if some college, =0 otherwise
colgrad: =1 if college graduate, =0 otherwise
black: =1 if black, =0 otherwise
female: =1 if female, =0 otherwise

5.3 link to the codes & dataset

---

codes:stata-do-file
datasets:stata dataset_regression partstata dataset_testing part