Central European Journal of Economic Modelling and Econometrics

Improving the Effectiveness of Maximum Score Estimators for Binary Regression Models

Marcin Owczarczuk*

Submitted: 14.09.2015, Accepted: 3.11.2015

Abstract

Maximum score estimation is a class of semiparametric methods for the coefficients of regression models. Estimates are obtained by the maximization of the special function, called the score. In case of binary regression models it is the fraction of correctly classified observations. The aim of this article is to propose a modification to the score function. The modification allows to obtain smaller variances of estimators than the standard maximum score method without impacting other properties like consistency. The study consists of extensive Monte Carlo experiments.

Keywords: maximum score estimation, Monte Carlo experiments, effectiveness

JEL Classification: C01, C14, C15

*Warsaw School of Economics; e-mail: mo23628@sgh.waw.pl

1 Introduction

Maximum score estimators belong to the class of semiparametric estimation techniques and are developed for regression models. Their advantage over parametric methods like maximum likelihood is that they require less strict assumptions about data generating process in order to obtain required properties, for example consistency. Maximum score estimators were first introduced by Manski (1975, 1985) for binary regression models. The term maximum score refers to the construction of the estimator - estimated coefficients maximize a certain score which in case of binary regression is a fraction of correctly classified observations in a sample.
The idea of Manski was later expanded in numerous ways. Horowitz (1992) introduced a kernel function to the maximized score function, which allowed to achieve asymptotic normality of estimates, whereas Kim and Pollard (1990) showed that the version of Manski provided non-normal asymptotic distributions. Huang and Abrevaya (2005) proved that the bootstrap technique cannot be used for Manski's version whereas Horowitz (2002) proved that it can be used for his variant. Moon (2004) analyzed maximum score estimators for non-stationary data. Owczarczuk (2009) modified the score function of Manski and Horowitz giving the possibility to estimate linear, binary, tobit and truncated regression models which provided improvement over previous techniques which were designed only for binary regression. However his version requires an introduction of additional calibrating constants that must be set during the estimation process.
The contribution of this paper is twofold. We show that the selected criterion, i.e. maximal fraction of correctly classified observations in Manski and Horowitz version, although very natural in binary regression models, is not the optimal one in terms of the precision of estimates. The deviation from optimality is especially visible for highly imbalanced samples, i.e. where a fraction of observation from certain class of explained variable, say $Y = 1$, is large.
Additionally, we show that the modification which is a subject of this paper is equivalent to optimal selection of calibrating constant within Owczarczuk's (2009) framework. In this article we analyze the impact of a calibrating constant in the Owczarczuk (2009) version of the estimator on the variance of estimates and provide its optimal value.
It is worth to underscore that the proposed modification inherits major properties of the versions of Manski and Horowitz, i.e. the modification in case of using indicator function generates non-normal asymptotic distributions of estimates and their confidence intervals cannot be approximated by the bootstrap technique. On the contrary using kernel functions instead of indicators has normal asymptotic distribution and their small-sample properties may be approximated by the bootstrap. The focus of the modification is to reduce variance.
The structure of this article is as follows. Section 2 provides overview of the construction of maximum score estimators and briefly describes their properties,

section 3 describes the idea of the proposed modification. Sections 4 and 5 provide the setup and results of Monte Carlo experiments. Last section concludes the article.

2 Maximum score estimators

The following data generating process is usually assumed for binary regression.

$$y_i = \mathbf{1}\left(\beta_0 + \beta^T x_i + \varepsilon_i\right), \tag{1}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function, β_0 is a constant term, β is the vector of coefficients by a vector of explanatory variables x_i and ε_i is the error term. For convenience the constant term and remaining coefficients are split.
For binary regression models usually the following prediction rule is used

$$\hat{y}_i = \begin{cases} 1 \text{ if } \hat{\beta}_0 + \hat{\beta}^T x_i \geq 0 \\ 0 \text{ if } \hat{\beta}_0 + \hat{\beta}^T x_i < 0, \end{cases} \tag{2}$$

what may be written as $\hat{y}_i = \mathbf{1}(\hat{\beta}_0 + \hat{\beta}^T x_i \geq 0)$.
The idea of Manski (1975, 1985) estimator is to look for the values of estimates $\hat{\beta}_0$, $\hat{\beta}$ that provide the highest fraction of correctly classified observations for a given sample. This rule may be written as the following maximization problem

$$\max_{b_0, b} \frac{1}{N} \sum_{i=1}^{N} [2y_i - 1] \left[2\mathbf{1}\left(b_0 + b^T x_i \geq 0\right) - 1\right] \tag{3}$$

The first term in the above sum is equal to $+1$ or -1 depending if y_i is equal to 1 or 0. The second term is equal to $+1$ or -1 depending if $\mathbf{1}(b_0 + b^T x_i \geq 0)$, i.e. the prediction, is equal to 1 or 0. So if prediction matches the actual value of y_i the sum increases by 1 and decreases by 1 otherwise.
The above optimization problem has infinite number of solutions. This is due to the fact, that the expression $b_0 + b^T x_i \geq 0$ has the same logical value when being multiplied by a positive constant, i.e.

$$b_0 + b^T x_i \geq 0 \Leftrightarrow cb_0 + cb^T x_i \geq 0, \text{for } c > 0. \tag{4}$$

So to obtain the unique solution, the normalizing condition must be implied. Manski assumed $\|[\beta_0, \beta]\| = 1$. So finally the estimator of Manski has the following form

$$[\hat{\beta}_0, \hat{\beta}] = \underset{[b,b_0]:\ \|[b,b_0]\|=1}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} [2y_i - 1] \left[2\mathbf{1}\left(b_0 + b^T x_i \geq 0\right) - 1\right] \tag{5}$$

Manski managed to prove consistency of his estimator. His initial idea of maximum score estimators was later expanded in numerous ways. An significant improvement

was done by Horowitz (1992). He noted that the optimized function is not a continuous one and replaced the indicator function by a smooth kernel function (cumulative distribution function). This modification allowed to achieve required asymptotic properties, i.e. asymptotic normality which in turn allowed statistical inference on significance of predictors. Horowitz formula is as follows

$$[\hat{\beta}_0, \hat{\beta}] = \underset{[b,b_0]:\ \|b_0\|=1}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} [2y_i - 1] \left[2K \frac{b_0 + b^T x_i}{h} - 1 \right], \tag{6}$$

where h is a bandwidth parameter, $K(\cdot)$ is a smooth kernel function satisfying $\lim_{t\to-\infty} K(t) = 0$ and $\lim_{t\to\infty} K(t) = 1$. Horowitz used a different but equivalent normalization $\|b_0\| = 1$.

The distinction between smooth version of Horowitz and non-smooth version of Manski has many consequences. In case of a non-smooth variant the asymptotic distribution in non normal as it was proved by Kim and Pollard (1990) and in case of a smooth formulation the asymptotic distribution is normal as proved by Horowitz (1992). What is more, small sample distributions cannot be approximated by the bootstrap technique for a Manski estimator (Huang, Abrevaya 2005) and may be for a Horowitz version (Horowitz 1992).

Owczarczuk (2009) proposed a generalization of maximum score estimators for a wider class of models than just binary regression, i.e. for linear, binary, tobit and truncated models. His idea was to replace the function that is maximized, i.e. a fraction of correctly classified observations, which makes sense only for binary models, by an average of explained variable which may be calculated for a large number of models. Introducing an average required implying an additional normalizing condition. Owczarczuk proposed the condition on the fraction of observations over which the average is calculated. The fraction parameter is denoted by $\tau \in (0, 1)$. Owczarczuk estimator may be expressed using indicator function or kernel function. The estimator is given by the following formula with indicators

$$[\beta_N, \beta_{0N}] = \underset{[b,b_0]:\ \|[b,b_0]\|=1}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} y_i \mathbf{1}\left(b^T x_i \geq b_0\right) - \mu \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(b^T x_i \geq b_0\right) - \tau \right)^2 \tag{7}$$

and with kernels

$$[\beta_N, \beta_{0N}] = \underset{[b,b_0]:\ \|[b,b_0]\|=1}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} y_i K\left(\frac{b^T x_i}{h}\right) - \mu \left(\frac{1}{N} \sum_{i=1}^{N} K \frac{b^T x_i}{h} - \tau \right)^2 \tag{8}$$

In the above formulas the normalization $\|[b, b_0]\| = 1$ was used, but any arbitrary normalization giving uniqueness may be applied.

The term $\mu(\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(b^T x_i \geq b_0) - \tau)^2$ is a penalty that ensures that only a fraction of τ observations gives the contribution to the average. The condition uses b values

which form regression parameters. As it was proved by Owczarczuk (2009), optimal values of b which maximize average and minimize penalty are close to real values β, giving consistency. The constant μ must be arbitrary large. The value of τ does not influence the consistency of the estimator for linear and binary regression. In case of tobit and truncated models, the smaller the value the higher the variance but smaller bias of the estimators (Owczarczuk 2009). In this setting value of the constant term β_0 is not estimated, so only the vector β is subject to estimation.

3 Modification of the score function for binary regression models

The modification of the score function for binary regression models that is subject of this paper is as follows: to maximize the fraction of correctly classified observations as previously but with implying the condition to keep the split between observations classified as $\hat{Y} = 1$ and $\hat{Y} = 0$ to 50%:50%. In other words we still want to maximize the correctness of predictions but we require exactly half of the population to be classified as $\hat{Y} = 1$.

The rationale is as follows. If a sample is balanced, i.e. number of observations from class $Y = 1$ is approximately equal to the number of observations from class $Y = 0$ then Manski and Horowitz versions, the goal of which is to maximize fraction of correctly classified observations, classify approximately half of the observations into class $\hat{Y} = 1$ and half of observations into class $\hat{Y} = 0$. If a sample is imbalanced, i.e. when a fraction of observations from one class, say $Y = 1$, increases, then these methods classify more and more observations into this majority class, i.e. the fraction of observations with prediction $\hat{Y} = 1$ increases. An extreme situation may be, say, when a sample contains of 99% observations from class $Y = 1$. Then a model that classifies all observations into this class, achieves the fraction of correctly classified observations of 0.99. Any improvements over this quantity will be only slight as 0.99 is already very large and close to maximum value of 1. So the method loses sensitivity when the imbalance increases. It will results in higher standard deviations as it will be showed in section 5.

On the contrary, in case of applying the additional condition of keeping the number of observations classified as $\hat{Y} = 1$ fixed, the situation looks quite differently. Since the new, proposed method forces to keep 50% to 50% split between $\hat{Y} = 1$ and $\hat{Y} = 0$, so it does not classify more and more observations into one class, so the sensitivity is not lost and the method obtains smaller standard deviations.

As far as the statistical underpinning is concerned, it is sufficient to note that the proposed modification is equivalent to the Owczarczuk (2009) version with $\tau = 0.5$, since the maximization of the mean value of explained variable in a certain subpopulation is equivalent to maximization of the number of observations from class $Y = 1$ in the population classified as $\hat{Y} = 1$.

A cost that the proposed modification brings is that the estimation of the constant term is biased, so only vector of coefficients by explanatory variables is estimated. However usually the constant term is of little interest in applications.
It should also be noted that the modification due to the implied split condition worsens the in-sample prediction in order to increase precision of the estimates by variables and it additionally introduces bias to the intercept. However these drawbacks may be easily overcome as after the estimation the biased intercept estimate may be replaced by its unbiased version. A typical procedure may be as follows: parameters by variables are estimated by the modified procedure and then an intercept is calibrated so that the in-sample prediction precision is maximized. This gives the final vector of estimates.

4 Monte Carlo experiments setup

In order to illustrate the properties of the proposed modification, the Monte Carlo experiments were conducted. The design is as follows. The following equation will be estimated

$$y_i = \mathbf{1}(a_0 + a_1 x_{1i} + a_2 x_{2i} + \varepsilon) \tag{9}$$

Values of a_1 and a_2 are set to 1. The value of a_0 is selected so that desired fraction of observations from the class $Y = 1$ is achieved and it differs for different setting.
Due to the normalization that must be implied on the vector of coefficients and for convenience of the interpretation of the results, the normalization $\|a_1\| = 1$ was used for estimation by maximum score. To sum up, it is sufficient to observe the estimation results for a_2 coefficient which true value is equal to 1. The following aspects of data generating process will be investigated.

1. Fraction of observations with $y_i = 1$ in the sample. The following values are considered: $\{0.2, 0.4, 0.5, 0.6, 0.8\}$. This aspect will be controlled by the constant term a_0.

2. Distribution of x_1 and x_2. The study uses normal distribution with zero expected value and unit variance. Experiments (not shown in this paper) with Student's t with 3 degrees of freedom , uniform and exponential were also conducted providing similar results.

3. Distribution of ε. Similarly to the distribution of x_1 and x_2, normal distribution was used. Experiments for Student's t with 3 degrees of freedom, uniform and exponential were also conducted, providing similar results..

4. Type of heteroscedasticity. In the study presented no heteroscedasticity was applied. Experiments with heteroscedasticity of the form $\varepsilon_{i\ het.} = \varepsilon_i \sqrt{|x_{1i} + x_{2i}|}$ where ε_i is homoskedastic, were also conducted.

5. Number of observations. The following values were considered: $\{500, 1500, 3000, 4500\}$

Within each setting 1000 replications were used. The aim of the experiments is to show that the split of 50%:50% between the predictions provides the smallest root mean squared error of estimates. Within the experiments the split will be parametrized by parameter τ, i.e. split τ:$(1-\tau)$ will be implied and it will be shown that the minimum is obtained by τ=0.5. The results are benchmarked against the standard Horowitz version. For each replication the following values of τ were applied $\{0.2, 0.3, \ldots, 0.7, 0.8\}$. The value of the kernel bandwidth was selected to $h = 1$ and the penalty constant to $\mu = 80$.

5 Monte Carlo experiments results

Since the simulations were conducted for various combinations of aspects of data generating process, results may be analyzed in many intersections.

Table 1: Results of Monte Carlo experiments: mean values of estimates

n	$P(Y=1)$	Horowitz	$\tau=0.2$	$\tau=0.3$	$\tau=0.4$	$\tau=0.5$	$\tau=0.6$	$\tau=0.7$	$\tau=0.8$
500	0.2	1.010	1.012	1.008	1.008	1.009	1.009	1.009	1.011
	0.4	1.008	1.011	1.009	1.008	1.008	1.008	1.010	1.015
	0.5	1.006	1.012	1.007	1.006	1.006	1.006	1.008	1.009
	0.6	1.010	1.017	1.014	1.012	1.011	1.010	1.010	1.015
	0.8	1.008	1.001	1.003	1.003	1.002	1.002	1.004	1.013
1500	0.2	1.003	1.003	1.002	1.002	1.002	1.003	1.003	1.003
	0.4	1.001	1.001	1.001	1.001	1.002	1.002	1.002	1.003
	0.5	1.002	1.002	1.002	1.002	1.003	1.003	1.005	1.007
	0.6	1.005	1.005	1.004	1.004	1.005	1.006	1.007	1.008
	0.8	1.006	1.006	1.006	1.006	1.006	1.005	1.005	1.006
3000	0.2	1.002	1.002	1.001	1.001	1.001	1.000	1.000	1.002
	0.4	0.999	1.001	1.000	0.999	0.999	0.999	0.999	0.998
	0.5	1.001	1.003	1.002	1.001	1.001	1.001	1.001	1.001
	0.6	1.003	1.001	1.002	1.002	1.002	1.003	1.003	1.005
	0.8	1.003	1.004	1.004	1.004	1.004	1.003	1.003	1.004
4500	0.2	0.999	0.999	0.999	0.999	0.999	1.000	1.000	1.001
	0.4	1.002	1.002	1.002	1.002	1.002	1.002	1.002	1.003
	0.5	0.998	0.999	0.998	0.998	0.998	0.998	0.998	0.999
	0.6	1.002	1.003	1.002	1.002	1.002	1.002	1.002	1.002
	0.8	0.999	1.001	1.001	1	1.000	0.999	0.999	0.999

Table 1 shows mean values of estimates, averaged over 1000 replications. We may observe that all variants provide approximately unbiased estimators, as all values are approximately equal to 1. We may conclude that the value of τ does not influence the

bias and estimators remain unbiased. The same conclusion applies to the Horowitz version. Estimators remain unbiased for all selected sample sizes and fractions of observations from class $Y = 1$.
Table 2 shows standard deviations of estimates, averaged over 1000 replications. Figures 1, 2, 3 and 4 provide graphical representation of values in the tables. We may observe that the value of τ strongly influences the standard deviation which is a measure of estimation precision. The optimal value is equal to $\tau = 0.5$. The higher deviation from 0.5, the higher standard deviation. This result is consistent for all selected sample sizes and fractions of observations from class $Y = 1$.

Table 2: Results of Monte Carlo experiments: standard deviations of estimates

n	$P(Y=1)$	Horowitz	$\tau=0.2$	$\tau=0.3$	$\tau=0.4$	$\tau=0.5$	$\tau=0.6$	$\tau=0.7$	$\tau=0.8$
500	0.2	0.170	0.184	0.153	0.144	0.142	0.142	0.144	0.166
	0.4	0.128	0.159	0.134	0.128	0.126	0.127	0.131	0.159
	0.5	0.122	0.155	0.130	0.124	0.123	0.124	0.131	0.159
	0.6	0.127	0.153	0.132	0.127	0.126	0.128	0.134	0.164
	0.8	0.170	0.168	0.150	0.147	0.144	0.145	0.154	0.188
1500	0.2	0.093	0.100	0.086	0.082	0.081	0.083	0.085	0.100
	0.4	0.068	0.090	0.074	0.069	0.067	0.067	0.070	0.087
	0.5	0.070	0.084	0.072	0.070	0.070	0.072	0.077	0.096
	0.6	0.073	0.088	0.074	0.071	0.071	0.073	0.078	0.094
	0.8	0.096	0.089	0.079	0.078	0.078	0.081	0.088	0.107
3000	0.2	0.064	0.069	0.059	0.056	0.057	0.058	0.059	0.067
	0.4	0.050	0.062	0.052	0.050	0.050	0.052	0.055	0.066
	0.5	0.048	0.062	0.053	0.050	0.048	0.049	0.051	0.062
	0.6	0.049	0.059	0.051	0.049	0.049	0.049	0.052	0.062
	0.8	0.068	0.065	0.058	0.057	0.057	0.058	0.062	0.075
4500	0.2	0.053	0.057	0.049	0.046	0.046	0.047	0.048	0.055
	0.4	0.042	0.052	0.044	0.042	0.042	0.042	0.044	0.053
	0.5	0.039	0.050	0.042	0.040	0.039	0.040	0.042	0.052
	0.6	0.040	0.048	0.042	0.040	0.039	0.040	0.042	0.052
	0.8	0.055	0.055	0.049	0.048	0.048	0.048	0.051	0.060

The comparison to the standard deviation of Horowitz version provides interesting conclusions. Namely if the fraction of observations from class $Y = 1$ is equal to 0.5, i.e. the sample is balanced, standard deviation of Horowitz version is approximately equal to standard deviation of the proposed modification. The higher the imbalance in the sample, the higher the difference in standard deviation between the modification and Horowitz version.
Equivalently we have shown that $\tau = 0.5$ is the optimal calibrating constant within Owczarczuk (2009) framework.

Figure 1: Standard deviation of estimates of maximum score estimators as a function of τ. Vertical line represents standard deviation of Horowitz estimator. Number of observations is equal to 500

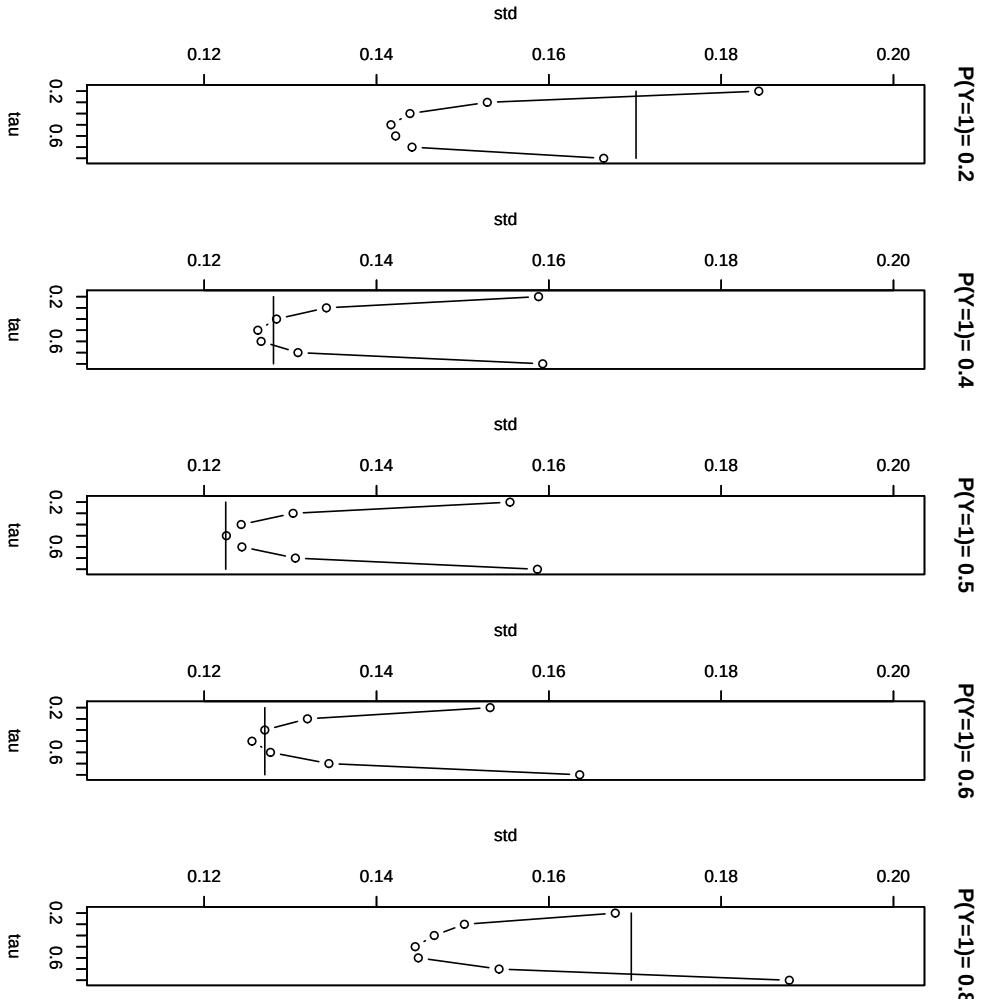


Figure 2: Standard deviation of estimates of maximum score estimators as a function of τ. Vertical line represents standard deviation of Horowitz estimator. Number of observations is equal to 1500

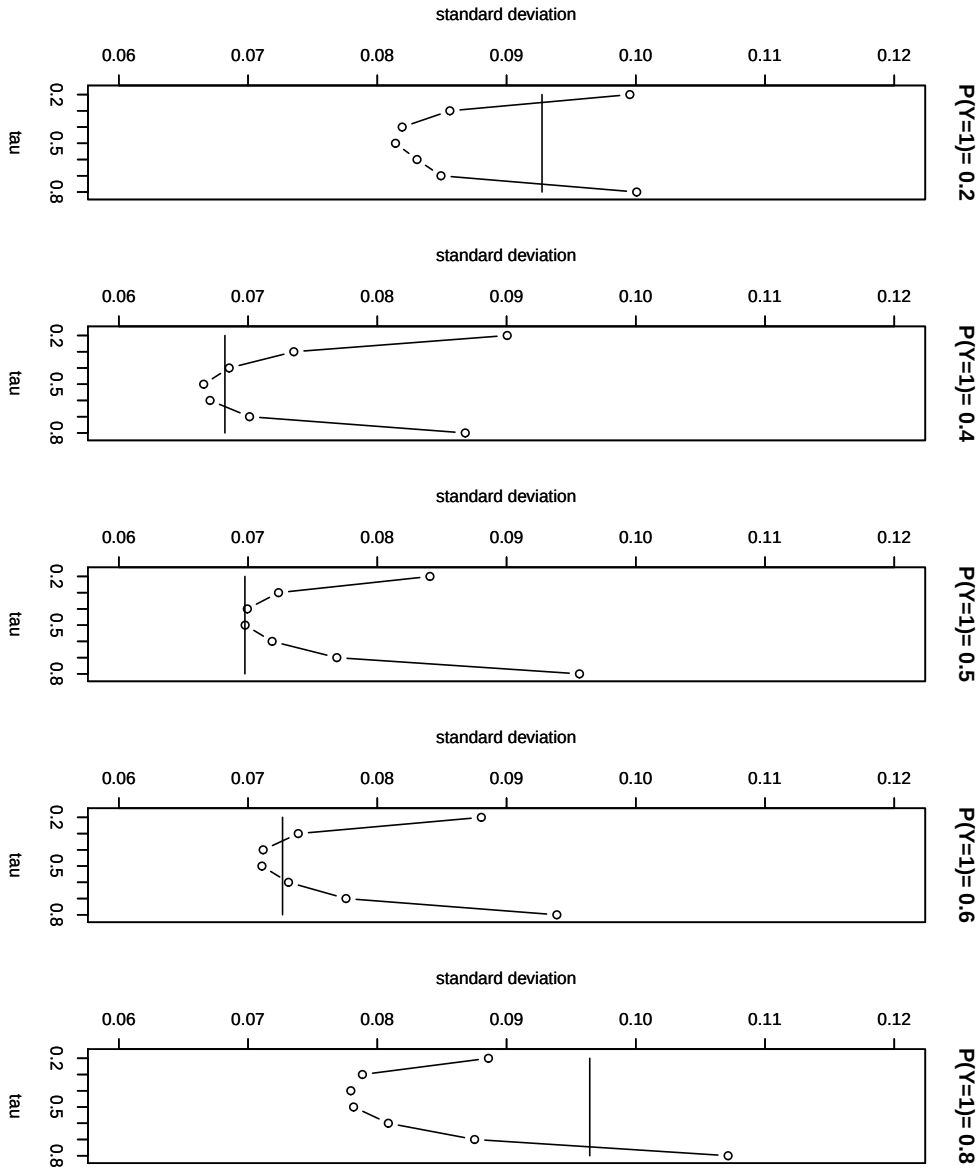


Figure 3: Standard deviation of estimates of maximum score estimators as a function of τ. Vertical line represents standard deviation of Horowitz estimator. Number of observations is equal to 3000

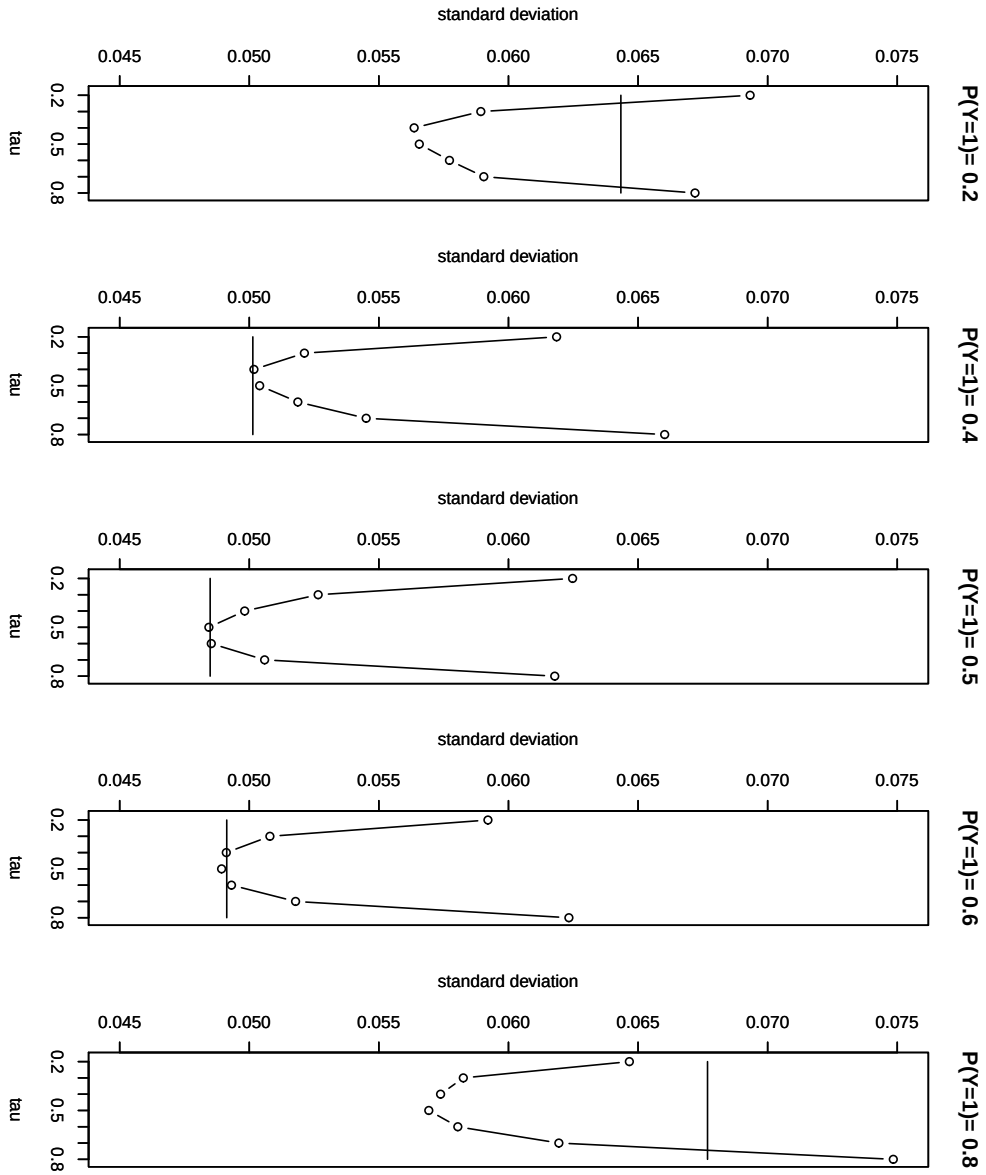
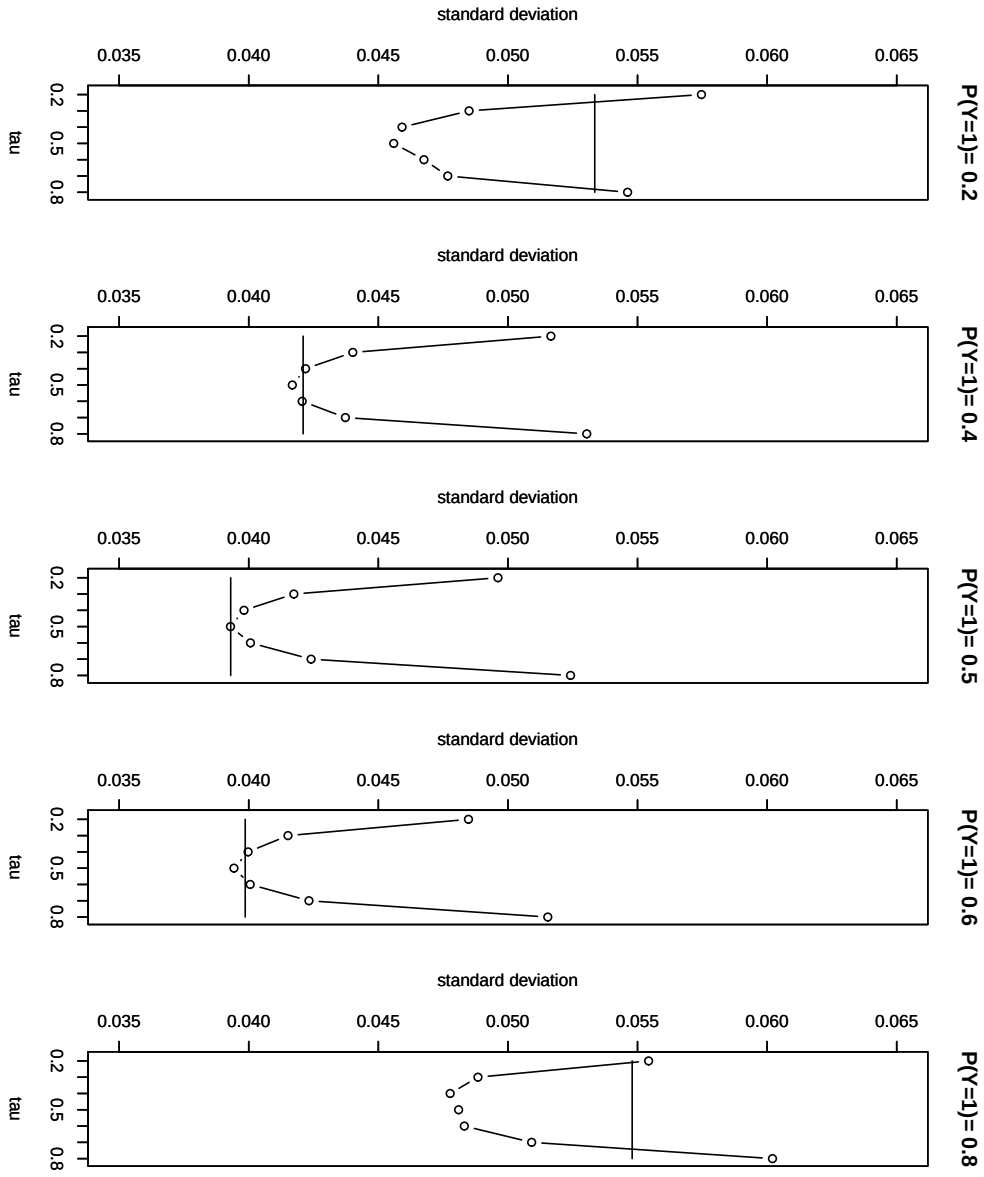


Figure 4: Standard deviation of estimates of maximum score estimators as a function of τ. Vertical line represents standard deviation of Horowitz estimator. Number of observations is equal to 4500



6 Conclusions

In this article we proposed a modification of the score function within the maximum score estimation for binary regression models. The proposed modification gives smaller variances of estimation than the standard maximum score technique. The advantage of the proposed approach is especially visible for imbalanced samples. Equivalently we provided the optimal value of the calibrating constant to the Owczarczuk (2009) framework of maximum score estimation.

References

[1] Horowitz J. L., (1992), A smoothed maximum score estimator for the binary response model, *Econometrica* 60(3), 505–531.

[2] Horowitz J. L., (2002), Bootstrap critical values for tests based on the smoothed maximum score estimator, *Journal of Econometrics* 111, 141–167.

[3] Huang J., Abrevaya J., (2005), On the Bootstrap of the Maximum Score Estimator, *Econometrica* 73(4), 1175–1204.

[4] Kim J., Pollard D., (1990), Cube root asymptotics, *Annals of Statistics* 18, 191-219.

[5] Manski C. F., (1975), Maximum score estimation of the stochastic utility model of choice, *Journal of Econometrics* 3, 205–228.

[6] Manski C. F., (1985), Semiparametric analysis of the discrete response. Asymptotic properties of the maximum score estimator, *Journal of Econometrics* 27, 313–333.

[7] Moon H. R., (2004), Maximum score estimation of a nonstationary binary choice model, *Journal of Econometrics* 122, 385–403.

[8] Owczarczuk M. *Maximum score type estimators*, Central European Journal of Economic Modelling and Econometrics 1 2009, 7–34.