Central European Journal of Economic Modelling and Econometrics

Maximum Score Type Estimators

Marcin Owczarczuk*

Submitted: 31.08.2008, Accepted: 5.03.2009

Abstract

This paper presents maximum score type estimators for linear, binomial, tobit and truncated regression models. These estimators estimate the normalized vector of slopes and do not provide the estimator of intercept, although it may appear in the model. Strong consistency is proved. In addition, in the case of truncated and tobit regression models, maximum score estimators allow restriction of the sample in order to make ordinary least squares method consistent.

Key Words: maximum score estimation, tobit, truncated, binomial, semiparametric

JEL Classification: C24, C25, C21

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

1 Introduction

Parametric methods of estimation, for example maximum likelihood method, require strong assumptions about the data-generating process, for example normality of the error term. Often, if these assumptions are not fulfilled, these methods provide inconsistent estimators of the model parameters. This fact lead researchers to look for semiparametric estimators, which rely on very weak assumptions about the data, especially about the error term, but assume certain functional form of the dependency between explained and explanatory variables, for example linear form.
Manski (1975, 1985) introduced the semiparametric maximum score estimator of parameters in the binary and multinomial choice model which is consistent under very mild assumptions. It allows for heteroskedasticity of the unknown form and arbitrary distribution of the error term. The idea of this method is based on considering proper "score function". Because in applications we are often interested in prediction which maximizes the accuracy, we may search for vector of parameters which maximizes the fraction of correctly predicted observations in the sample. This is the "score function" for this case. It turns out that vector which maximizes the score function is a consistent estimator of parameters of the model.
Properties of the maximum score estimator are well known. Manski (1985) proved consistency. Kim and Pollard (1990) derived its asymptotic distribution and showed that it is nonnormal. Huang, Abrevaya (2005) showed that the bootstrap method does not provide consistent estimators of the confidence intervals. Moon (2004) analyzed the problem of nonstationary regressors and showed consistency in that case.
Horowitz (1992) replaced nondifferentiable indicator function in estimator of Manski by a smooth cumulative distribution function and achieved a smoothed maximum score estimator. Horowitz (1992, 2002) showed that the smoothed maximum score estimator has the normal limit distribution and that the bootstrap provides consistent estimates of the confidence intervals.
In this article we generalize the score function so that it encompasses other regression models and allows consistent estimation with the advantages of the estimator of Manski. The article is organized as follows: in the next section we introduce the generalized score function and informally present its motivation. The next five sections (2-6) present formal construction of the models and theorems of consistency. The seventh section addresses the problem of estimating using OLS to a sample restricted by maximum score. The eighth section presents the results of a Monte Carlo study. In the last section we present conclusions. Proofs of theorems are in the appendices.

2 The generalized score function

Informally, the idea of our maximum score type estimators is based on searching for the subset of observations where the mean of the explained variable is as high as possible with additional restriction on the number of observations in this subset. This

subset is determined by the linear restriction involving explanatory variables.
It turns out that the parameters of this restriction are consistent estimators of the parameters of the model. Instead of the number of observations of this subset, we may consider its measure expressed as a fraction or a percentage of the whole sample. This parameter will be denoted by τ. For different values of this parameter different estimators can be obtained and the quality of estimation may depend on choice of this parameter. This problem is analyzed later in this paper.
Let us consider the linear model $y = \beta_0 + \beta^T x + u$. Then we consider the subset of the (x, y) in which variable y has a value greater than c. We have

$$\{(x,y) : y > c\} = \{(x,y) : \beta_0 + \beta^T x + u > c\} = \{(x,y) : \beta^T x + u > c - \beta_0\} \qquad (1)$$

If we assume that the measure of this subset is equal to τ, then $c + \beta_0$ is a quantile of the order $1 - \tau$ of the variable $\beta^T x + u$.
If $E[u] = 0$, then the subset of size τ having the highest expected value of y is described by condition $\beta^T x > c + \beta_0 = \beta_\tau$. So, if we search for a subset bounded by linear restriction and having the highest mean of the explained variable y, we may expect that the parameters of this condition are close to β and β_τ.
More precisely, we are able to estimate parameters standing by the explanatory variables and we are not able to estimate the intercept. The reason is that we estimate $\beta_\tau = c + \beta_0$ not knowing c.
Similarly to models for binary choice, parameters are estimated up to a multiplicative constant. In practice some additional normalizing condition is added, which guaranties identification, for example, in probit regression the unit variance of the error term is imposed. We, similarly as in Manski (1985) assume unit length of the parameter vector, so we estimate $\beta^* = \frac{\beta}{\|\beta\|}$, where $\|\beta\|$ is the Euclidean norm of β.
Estimating the model without an intercept and up to the multiplicative constant is satisfactory in some applications. It provides information about the direction of the relations between variables (sign of the parameters) and allows verifying hypotheses about the significance of variables. It also allows ranking of observations with respect to the predicted value of the explained variable which is sufficient, for example, in credit scoring applications.
In the case of applications where estimating the intercept is required, there is a solution based on a maximum score. In the case of the tobit and truncated regression model, it turns out that applying ordinary least squares to a subset separated by our maximum score estimator, that is, to a subset determined by the restriction $\beta_N^T x > \beta_{0N}$, where β_N and β_{0N} are maximum score type estimators, provides a consistent estimator of the original parameters of the model, including the intercept. Note that the OLS estimator applied to the whole sample is inconsistent. So the maximum score allows deletion of observations which cause inconsistency.
The consistent estimators β_N^* of the normalized parameter vector β^* are achieved as the solution of certain maximization problems. Now we present the outline of the construction of the estimator and of the proof of consistency for linear regression.

The formal description of the model and the proof of consistency for linear regression and other models are presented in next sections. We consider the following linear regression model

$$y = \beta_0 + \beta^T x + u, \tag{2}$$

where x is the k-dimensional vector of the explanatory variables, y is the explained variable and u is the error term. We also assume, that there is given the sample of length N, (x_i, y_i) $i = 1, \ldots, N$, drawn from (x, y).
The estimator β_N^* is given by the following formulae

$$[\beta_N, \beta_{0N}] = \underset{[b,b_0]:\ \|[b,b_0]\|=1}{\operatorname{argmax}} \frac{1}{N}\sum_{i=1}^{N} y_i \mathbf{1}(b^T x_i \geq b_0) - \mu(\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(b^T x_i \geq b_0) - \tau)^2 \tag{3}$$

$$\beta_N^* = \frac{\beta_N}{\|\beta_N\|}. \tag{4}$$

Its heuristic derivation is as follows. Our goal is to find a subset of observations of fixed size τ and bounded by the hyperplane $b^T x \geq b_0$ where the mean of the explained variable y is maximal in comparison to other subsets of size τ and bounded by linear restriction. Since condition $b^T x \geq b_0$ is fulfilled when multiplied by positive constant c, that is

$$b^T x \geq b_0 \Leftrightarrow cb^T x \geq cb_0 \tag{5}$$

we add the condition which guarantees identification $\|[b, b_0]\| = 1$. As a result we obtain the following optimization problem:

$$\max_{[b,b_0]:\ \|[b,b_0]\|=1} \frac{1}{N}\sum_{i=1}^{N} y_i \mathbf{1}(b^T x_i \geq b_0) \quad \text{subject to} \quad \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(b^T x_i \geq b_0) = \tau. \tag{6}$$

In order to prove consistency, we want to apply the theorem of consistency of M-estimators (Engle, McFadden (1999), p. 2121). We cannot use it directly, because it involves obtaining estimators as a solution to certain optimization problems under deterministic restriction $\theta \in \Theta$. In our case we have an additional restriction which depends on the explanatory variables $\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(b^T x_i \geq b_0) = \tau$. We may construct the following modified maximization problem:

$$\max_{[b,b_0]:\ \|[b,b_0]\|=1} \frac{1}{N}\sum_{i=1}^{N} y_i \mathbf{1}(b^T x_i \geq b_0) - \mu[\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(b^T x_i \geq b_0) - \tau]^2, \tag{7}$$

where $\mu > 0$ is constant and is chosen by the researcher. Due to the theorem of convergence of the quadratic penalty method (Nocedal, Wright (1999), p. 494), the solution of this problem is close to the solution of the problem (6) for high values of μ. We simply substituted the restriction by the quadratic penalty.
Next, we may apply the theorem of consistency of M-estimators to the problem (7).

It turns out that having a sufficiently large sample and choosing a sufficiently large value of μ we may achieve, with an arbitrary high probability, estimators which are arbitrary close to the true values of the parameters $\frac{[\beta,\beta_\tau]}{\|[\beta,\beta_\tau]\|}$. Since we are interested only in $\beta^* = \frac{\beta}{\|\beta\|}$, we take $\beta_N^* = \frac{\beta_N}{\|\beta_N\|}$.
In equation (3) the value β_{0N} is not the estimator of β_0. It is an additional parameter which depends on τ. It is a quantile of order $1-\tau$ of the variable $\beta_N^T x$. So we are interested only in the first element of the pair $[\beta_N, \beta_{0N}]$.

Comment 2.1 *The indicator function* $\mathbf{1}(\cdot)$ *can be replaced by a smooth cumulative distribution function* $K(\cdot)$ *and we can achieve a smoothed maximum score type estimator, similarly to Horowitz (1992). In this case we must solve the following optimization problem*

$$\max_{[b,b_0]:\ \|[b,b_0]\|=1} \frac{1}{N}\sum_{i=1}^{N} y_i K(\frac{b^T x_i - b_0}{h}) \quad \textit{subject to} \quad \frac{1}{N}\sum_{i=1}^{N} K(\frac{b^T x_i - b_0}{h}) = \tau. \tag{8}$$

3 The linear regression model

We consider the following linear regression model

$$y = \beta_0 + \beta^T x + u. \tag{9}$$

Assumption 3.1

1. $y = \beta_0 + \beta^T x + u$, $x \in R^K$ $(K \geq 1)$, u *is random scalar,* $\beta_0 \in R$ *and* $\beta \in R^K$ *are constant;*
2. *The support of* x *is not contained in any proper linear subspace of* R^K*;*
3. *There exist at least one* $k \in \{1, \ldots, K\}$ *such that* $\beta_k \neq 0$ *and for almost every value of* $\tilde{x} = (x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_K)$ *the conditional distribution of* x_k *conditional on* $\tilde{x}$ *has everywhere positive density with respect to the Lebesgue measure;*
4. $E(u|x) = 0$ *for almost every* x*;*
5. $E[x] < \infty$*;*
6. $\{y_n, x_n : n = 1, \ldots, N\}$ *is random sample from* (y, x).

The estimator β_N^* of the parameter vector $\beta^* = \frac{\beta}{\|\beta\|}$ is given by the following formula

$$[\beta_N, \beta_{0N}] = \operatorname*{argmax}_{[b,b_0]:\ \|[b,b_0]\|=1} \frac{1}{N}\sum_{i=1}^{N} y_i \mathbf{1}(b^T x_i \geq b_0) - \mu(\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(b^T x_i \geq b_0) - \tau)^2, \tag{10}$$

$$\beta_N^* = \frac{\beta_N}{\|\beta_N\|}, \tag{11}$$

where parameter $\tau \in (0,1)$ is arbitrary. The following theorem is satisfied:

Theorem 3.1 *Under Assumption 3.1 the estimator β_N^* defined by formulae (10) and (11) is a strongly consistent estimator of the vector of parameters $\beta^* = \frac{\beta}{\|\beta\|}$ in model (9).*

The proof of this theorem is presented in Appendix A.

4 The binary model

We consider the following binary regression model

$$y = \begin{cases} 1 & \text{when} \quad \beta_0 + \beta^T x + u \geq 0 \\ 0 & \text{when} \quad \beta_0 + \beta^T x + u < 0. \end{cases} \tag{12}$$

Assumption 4.1

1. *$y = \mathbf{1}(\beta_0 + \beta^T x + u \geq 0)$, $x \in R^K$ $(K \geq 1)$, u is random scalar and $\beta_0 \in R$ i $\beta \in R^K$ is constant;*
2. *The support of x is not contained in any proper linear subspace of R^K;*
3. *$0 < P(y = 1|x) < 1$ almost everywhere;*
4. *There exists at least one $k \in \{1, \ldots, K\}$ such that $\beta_k \neq 0$ and for almost every value of $\tilde{x} = (x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_K)$ the conditional distribution x_k conditional on $\tilde{x}$ has everywhere positive density with respect to the Lebesgue measure;*
5. *$g(E[y]) = \beta^T x + \beta_0$, where $g : (0,1) \to R$ is increasing;*
6. *$\{y_n, x_n : n = 1, \ldots, N\}$ is random sample from (y, x).*

The estimator β_N^* of the vector of parameters $\beta^* = \frac{\beta}{\|\beta\|}$ is given by the following formula

$$[\beta_N, \beta_{0N}] = \underset{[b,b_0]:\ \|[b,b_0]\|=1}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} y_i \mathbf{1}(b^T x_i \geq b_0) - \mu(\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(b^T x_i \geq b_0) - \tau)^2, \tag{13}$$

$$\beta_N^* = \frac{\beta_N}{\|\beta_N\|}, \tag{14}$$

where the parameter $\tau \in (0,1)$ can be arbitrary.The following theorem is satisfied.

Theorem 4.1 *Under assumption 4.1 the estimator defined by formulae (13) and (14) is a strongly consistent estimator of the vector of parameters* $\beta^* = \frac{\beta}{\|\beta\|}$ *in model (12).*

The proof of this theorem is presented in Appendix B.

5 Truncated regression

We consider the following truncated regression model

$$y = \beta_0 + \beta^T x + u, \tag{15}$$

Assumption 5.1

1. $y = \beta_0 + \beta^T x + u$, $x \in R^K$ $(K \geq 1)$,u *is random scalar,* $\beta_0 \in R$, $\beta \in R^K$, C *is a known constant;*
2. *The support of* x *is not contained in any proper linear subspace of* R^K.
3. *There exist at least one* $k \in \{1, \ldots, K\}$ *such that* $\beta_k \neq 0$ *and for almost every value of* $\tilde{x} = (x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_K)$ *the conditional distribution of* x_k *conditional on* $\tilde{x}$ *has everywhere positive density with respect to the Lebesgue measure;*
4. $E(u|x) = 0$ *for almost every* x*;*
5. $E[x] < \infty$*;*
6. $\frac{1}{1-F_{u|x}(C-\beta_0-\beta^T x)} \int_{C-\beta_0-\beta^T x}^{\infty} u dF_{u|x} \to 0$ *for* $\beta_0 + \beta^T x \to \infty$*;*
7. $\{y_n, x_n : n = 1, \ldots, N\}$ *is random sample from* $(y, x)|y \geq C$.

The estimator β_N^* of the vector of parameters $\beta^* = \frac{\beta}{\|\beta\|}$ is given by the following formula

$$[\beta_N, \beta_{0N}] = \underset{[b,b_0]:\ \|[b,b_0]\|=1}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} y_i \mathbf{1}(b^T x_i \geq b_0) - \mu(\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(b^T x_i \geq b_0) - \tau)^2, \tag{16}$$

$$\beta_N^* = \frac{\beta_N}{\|\beta_N\|}. \tag{17}$$

Parameter $\tau \in (0, 1)$ is a function of N such that $\tau \overset{N\to\infty}{\longrightarrow} 0$. The following theorem is satisfied

Theorem 5.1 *Under assumption 5.1 the estimator given by the formula (16) and (17) is a strongly consistent estimator of the vector of parameters* β^* *in model (15).*

The proof of this theorem is presented in Appendix C.

6 The tobit model

We consider the following tobit model

$$y = \begin{cases} \beta_0 + \beta^T x + u & \text{for} \quad \beta_0 + \beta^T x + u \geq C \\ 0 & \text{for} \quad \beta_0 + \beta^T x + u < C \end{cases} \tag{18}$$

Assumption 6.1

1. *$y = \max(C, \beta_0 + \beta^T x + u)$, $x \in R^K$ $(K \geq 1)$, u is random scalar, $\beta_0 \in R$, $\beta \in R^K$, C is a known censoring constant;*
2. *The support of x is not contained in any proper linear subspace of R^K;*
3. *There exist at least one $k \in \{1, \ldots, K\}$ such that $\beta_k \neq 0$ and for almost every value of $\tilde{x} = (x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_K)$ the conditional distribution of x_k conditional on $\tilde{x}$ has everywhere positive density with respect to the Lebesgue measure;*
4. *$E(u|x) = 0$ for almost every x;*
5. *$E[x] < \infty$;*
6. *$\frac{1}{1-F_{u|x}(C-\beta_0-\beta^T x)} \int_{C-\beta_0-\beta^T x}^{\infty} u dF_{u|x} \to 0$ for $\beta_0 + \beta^T x \to \infty$;*
7. *$\{y_n, x_n : n = 1, \ldots, N\}$ is random sample from (y, x).*

The estimator β_N^* of the vector of parameters $\beta^* = \frac{\beta}{\|\beta\|}$ is given by the following formula

$$[\beta_N, \beta_{0N}] = \underset{[b,b_0]:\ \|[b,b_0]\|=1}{\operatorname{argmax}} \frac{1}{N^*} \sum_{i=1}^{N^*} y_i \mathbf{1}(b^T x_i \geq b_0) - \mu(\frac{1}{N^*} \sum_{i=1}^{N^*} \mathbf{1}(b^T x_i \geq b_0) - \tau)^2, \tag{19}$$

$$\beta_N^* = \frac{\beta_N}{\|\beta_N\|}. \tag{20}$$

Here $1, \ldots, N^* < N$ are indices of the noncensored observations (i.e. $(x, y) : y > C$) observations. Parameter $\tau \in (0, 1)$ is a function of N such that $\tau \overset{N\to\infty}{\longrightarrow} 0$.

Comment 6.1 *We reduced the problem of estimation of the tobit model to a problem of estimation of a truncated regression model.*

The following theorem is satisfied

Theorem 6.1 *Under assumption 6.1 the estimator defined by (19) and (20) is strongly consistent estimator of the vector of parameters β^* in the model (18).*

Proof. The proof of this theorem is the same as proof of the theorem for truncated regression, according to the Comment 6.1.

7 Applying OLS to the restricted sample

In the case of tobit and truncated regression the least squares method used to the whole sample gives inconsistent estimates of the parameters. However, the OLS estimation applied to a subsample satisfying the restriction implied by *maximum score* gives a consistent estimator of the unknown parameters.

Theorem 7.1 *Under assumption 5.1, the OLS estimator applied to the sample satisfying condition $\beta_N x \geq \beta_{0N}$, where parameters β_N and β_{0N} are given by formula (16), is a consistent estimator of the parameters β and β_0 in model (15).*

The proof of this theorem is presented in Appendix D. A similar theorem is valid for tobit model, because we may reduce the problem of estimating the tobit model to the problem of estimating truncated regression.
The just described reasoning can be easily illustrated graphically. Here we present it for tobit regression. We generated a sample of 500 observations according to the following scheme

$$y = 2 + x + \epsilon, \quad \epsilon \sim N(0,1), \quad x \sim N(0,4), \tag{21}$$

$$y^* = \max(0, y). \tag{22}$$

The scatter plots for this sample are presented on Figures 1, 2 , 3 and 4.
In Figure 1, the true relationship between variables x and y is presented. Estimating the regression of the form $y = ax + b + \epsilon$ using OLS gives a consistent estimator of the unknown parameters.
However, we have censoring and we observe a relation presented in Figure 2. The OLS estimator is inconsistent. We obtain a line that has a smaller slope than the true regression line. The reason for this is the fact that points on the left side of the graph turn the line - this is the effect of censoring.
However, further restriction on the sample - deleting points on the left side of the graph, that is points on the left of the vertical line on Figure 3, makes the OLS estimator consistent. We simply delete observations causing the smaller slope of the regression line.

8 Monte Carlo simulation

In this section we present Monte Carlo simulations illustrating small sample properties of the maximum score type estimators. Our theoretical results are asymptotic, so it is interesting to verify whether these estimators have also good properties in finite samples.
The main goal of the simulations is to answer the following questions:

1. Are the bias and variance of the maximum score estimators comparable to the bias and variance of the classical estimators in the normal homoskedastic case?

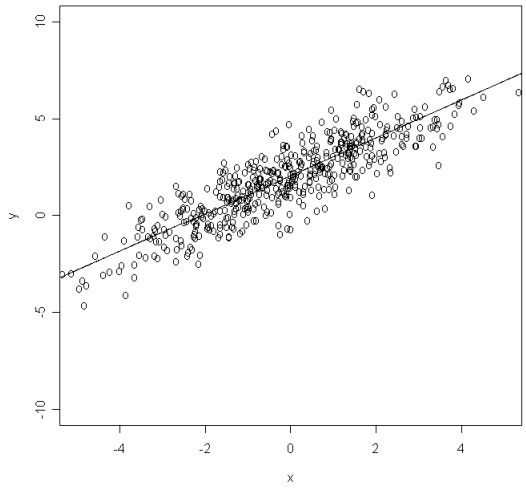

Figure 1: True, unobserved relation between variables y and x with the regression line estimated by OLS.

2. Are the maximum score estimators robust to heteroskedasticity when the size of the sample is moderate? Are they more robust that the classical methods, i.e. maximum likelihood in the case of tobit and truncated regression, logistic regression in the case of binomial regression and OLS in the case of linear regression?

The scheme of experiments was as follows. We tested 4 models: linear, binomial, tobit and truncated regression. In all experiments the number of observations was set to 1500. There were 1000 replications per experiment. For the error term we used the following distributions: uniform, normal and Student with 3 degrees of freedom and unit variance.
We used the moderate size of the sample, but it cannot be small when using maximum score. Maximum score separates a relatively small subset of observations and calculates mean of the explained variable in this subset. So in order to keep the precision of the estimates, this mean must be estimated precisely. A moderate size of the sample implies a relatively moderate size of the separated subset and precision of the estimator of the mean.
We used Student distribution of the error term in order to check the quality of the estimates when the distribution of the error term has fat tails. The uniform distribution represents situation where the error term is bounded.

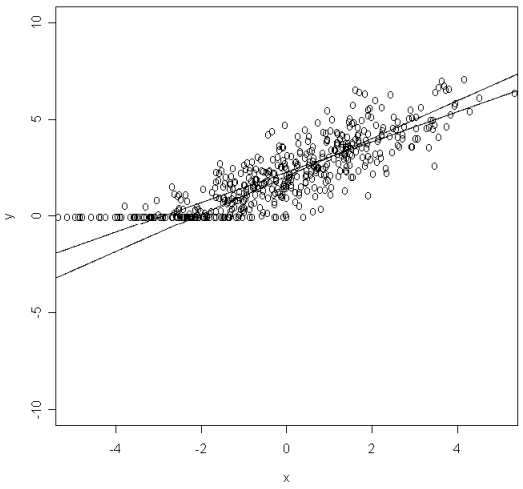

Figure 2: Observed relation between y^* and x with the OLS regression line (the line with the smaller slope) and the true regression line (the line with the greater slope).

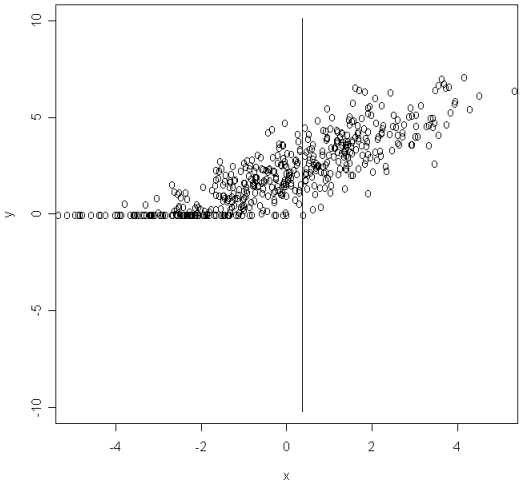

Figure 3: Deleting observations on the left side of the line cancels the OLS bias.

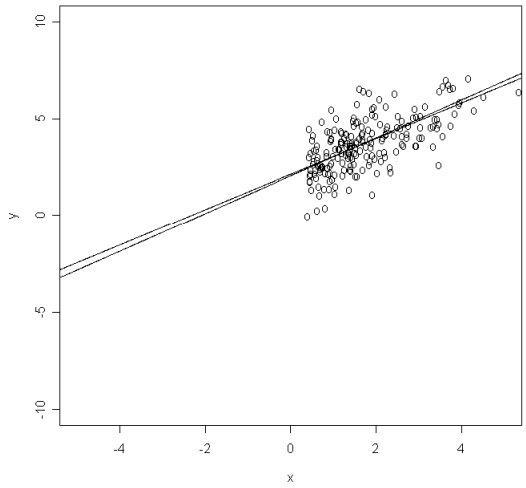

Figure 4: The true regression line and the OLS regression line estimated on a restricted sample. The lines match almost perfectly.

The number of replications was set to 1000 which gave the stability of results. We used popular methods like logistic regression or OLS as benchmarks. The data generating process is as follows

$$y = 1 + x_1 + x_2 + \epsilon, \tag{23}$$

$$x_1 \sim N(0,1), \tag{24}$$

$$x_2 \sim N(0,1), \tag{25}$$

$$\epsilon \in \{N(0,1), t_3, U[-1,1]\}, \tag{26}$$

$$y^* = \begin{cases} y & \text{for linear regression} \\ \mathbf{1}(y \geq 0) & \text{for binomial regression} \\ \max(y,0) & \text{for tobit regression} \\ y|y>0 & \text{for truncated regression.} \end{cases} \tag{27}$$

So the following models were estimated

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \tag{28}$$

assuming that (y^*, x_1, x_2) is observed. We used distributions without and with heteroskedasticity. In case of heteroskedasticity, we introduced it in the following form:

$$\epsilon_{heteroskedastic} = \epsilon\sqrt{|x_1 + x_2|} \tag{29}$$

Table 1: Linear regression, bias of normalized slope

distribution	maximum score	OLS
normal	0.0028	0.0018
normal het.	-0.0006	-0.0009
t_3	0.0007	0.0005
t_3 het.	0.0034	0.0025
uniform	0.0003	0.0000
uniform het.	0.0019	0.0036

Table 2: Linear regression, rmse of normalized slope

distribution	maximum score	OLS
normal	0.0507	0.0367
normal het.	0.0449	0.0395
t_3	0.0507	0.0365
t_3 het.	0.0457	0.0369
uniform	0.0526	0.0367
uniform het.	0.0458	0.0382

8.1 Experiment 1 - estimating the standardized vector of slopes

In this experiment we are interested in estimating the vector $\beta^* = \frac{\beta}{\|\beta\|} = \frac{[\beta_1,\beta_2]}{\|[\beta_1,\beta_2]\|}$. In the proofs of consistency we used normalization $\|[\beta_\tau, \beta]\| = 1$. Here, similarly to Horowitz (1992) we use a slightly different normalization, namely $\|\beta_1\| = 1$. The reason is that we are not interested in estimating the intercept, and since there are only two slopes, one of them is identified up to a sign knowing the second one. Using normalization $\|\beta_1\| = 1$ we may focus on small sample properties of estimates of only one parameter, β_2^*. Its true value is equal 1. We set $\tau = 0.5$ for linear and binomial regression and $\tau = 0.2$ for truncated and tobit regression. Because of numerical complexity, we used the smoothed maximum score version with the normal kernel and the bandwidth parameter $h = 1$.

In the case of linear regression we compared the maximum score estimator with OLS. In the case of binomial regression we compared the maximum score estimator with logistic regression. In the case of tobit and truncated regression we compared the maximum score estimator with OLS and the maximum likelihood method.

Despite this fact that estimating the tobit model may be reduced to truncated regression estimation by deleting censored observations, in our experiments we used the full sample for the estimation purposes. It resulted in slightly smaller mean square errors of the estimators.

We may observe that in the case of estimating the normalized vector of coefficients, all methods give approximately unbiased estimates. This is very interesting, because as far as estimating the full vector of coefficients of truncated and tobit regression is concerned, maximum likelihood is consistent only in case of normal distribution without heteroskedasticity. Monte Carlo experiments show that, despite this fact, this method provides approximately unbiased estimates of normalized vector of coef-

Table 3: Binomial regression, bias of normalized slope

distribution	maximum score	logit
normal	0.0048	0.0062
normal het.	0.0027	0.0015
t_3	0.0015	0.0017
t_3 het.	0.0006	0.0029
uniform	0.0047	0.0023
uniform het.	0.0023	0.0031

Table 4: Binomial regression, rmse of normalized slope

distribution	maximum score	logit
normal	0.0751	0.0639
normal het.	0.0692	0.0654
t_3	0.0668	0.0534
t_3 het.	0.0657	0.0573
uniform	0.0766	0.0644
uniform het.	0.0692	0.0642

Table 5: Tobit regression, bias of normalized slope

distribution	maximum score	OLS	ML
normal	0.0026	0.0009	0.0007
normal het.	0.0046	0.0031	0.0019
t_3	0.0025	0.0032	0.0020
t_3 het.	0.0022	-0.0007	-0.0010
uniform	0.0020	0.0023	0.0027
uniform het.	0.0052	0.0030	0.0028

Table 6: Tobit regression, rmse of normalized slope

distribution	maximum score	OLS	ML
normal	0.0566	0.0442	0.0400
normal het.	0.0554	0.0452	0.0402
t_3	0.0544	0.0422	0.0366
t_3 het.	0.0524	0.0446	0.0381
uniform	0.0583	0.0440	0.0407
uniform het.	0.0580	0.0468	0.0424

Table 7: Truncated regression, bias of normalized slope

distribution	maximum score	OLS	ML
normal	0.0045	0.0035	0.0036
normal het.	0.0038	0.0014	0.0009
t_3	0.0009	0.0001	0.0002
t_3 het.	0.0017	-0.0002	-0.0002
uniform	0.0048	0.0029	0.0023
uniform het.	0.0040	0.0017	0.0005

Table 8: Truncated regression, rmse of normalized slope

distribution	maximum score	OLS	ML
normal	0.0734	0.0486	0.0502
normal het.	0.0731	0.0513	0.0508
t_3	0.0717	0.0441	0.0462
t_3 het.	0.0690	0.0462	0.0461
uniform	0.0838	0.0530	0.0538
uniform het.	0.0769	0.0549	0.0539

Table 9: Tobit regression, bias of full vector of parameters

distribution	maximum score-OLS			OLS			ML		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
normal	0.0081	-0.0011	-0.0058	0.3033	-0.2821	-0.2822	-0.0002	0.0004	0.0002
normal het.	0.0188	-0.0085	-0.0018	0.2946	-0.2975	-0.2960	0.0491	-0.0601	-0.0591
t_3	0.0286	-0.0100	-0.0079	0.2843	-0.2706	-0.2689	-0.0146	0.0228	0.0242
t_3 het.	0.0134	0.0107	0.0072	0.2797	-0.2769	-0.2782	0.0170	-0.0126	-0.0143
uniform	-0.0031	0.0032	0.0007	0.3073	-0.2859	-0.2849	0.0024	-0.0078	-0.0059
uniform het.	0.0078	-0.0040	-0.0039	0.2978	-0.3033	-0.3021	0.0693	-0.0828	-0.0811

ficients. A similar situation can be observed for OLS and logistic regression. Greene (1981) analyzed estimating tobit and truncated regression using OLS. He proved consistency of OLS under normality of the vector of the explanatory variables and showed (by Monte Carlo experiments) robustness to this distributional assumption.
The advantage of maximum score is that its robustness is proved not only empirically but also theoretically for a greater class of data-generating processes. Maximum score has a slightly greater variance than other methods.

8.2 Experiment 2 - estimating the full vector of parameters

As it was mentioned earlier, in the case of tobit and truncated regression, OLS applied to subsample separated by maximum score, gives a consistent estimator of the unknown parameters of the model. In this experiment we compared OLS applied to a subsample separated by maximum score to OLS applied to the whole sample and maximum likelihood method. We set parameter $\tau = 0.2$.
In the case of estimating the full vector of coefficients of the tobit or truncated regression model, OLS is always biased and Monte Carlo experiments confirm this fact.

Table 10: Tobit regression. rmse of full vector of parameters

distribution	maximum score-OLS			OLS			ML		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
normal	0.1786	0.0957	0.0954	0.3042	0.2832	0.2832	0.0285	0.0309	0.0298
normal het.	0.2775	0.1563	0.1575	0.2955	0.2988	0.2973	0.0558	0.0696	0.0690
t_3	0.1746	0.0916	0.0907	0.2851	0.2715	0.2699	0.0355	0.0420	0.0446
t_3 het.	0.2519	0.1456	0.1423	0.2805	0.2783	0.2796	0.0343	0.0446	0.0453
uniform	0.1901	0.0993	0.0976	0.3081	0.2868	0.2859	0.0305	0.0319	0.0319
uniform het.	0.2849	0.1578	0.1589	0.2986	0.3046	0.3033	0.0742	0.0905	0.0882

Table 11: Truncated regression. bias of full vector of parameters

distribution	maximum score-OLS			OLS			ML		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
normal	0.0256	-0.0108	-0.0081	0.4046	-0.2461	-0.2443	0.0005	-0.0018	0.0005
normal het.	0.0723	-0.0064	-0.0040	0.3378	-0.2451	-0.2452	-0.3619	0.1675	0.1668
t_3	0.0622	-0.0123	-0.0144	0.3053	-0.1816	-0.1823	-0.1815	0.0958	0.0947
t_3 het.	0.0621	0.0105	0.0107	0.2735	-0.1820	-0.1829	-0.4523	0.2233	0.2214
uniform	0.0062	-0.0041	-0.0001	0.4335	-0.2692	-0.2681	0.0337	-0.0280	-0.0272
uniform het.	-0.0053	0.0048	0.0007	0.3451	-0.2812	-0.2811	-0.4214	0.1635	0.1625

Table 12: Truncated regression. rmse of full vector of parameters

distribution	maximum score-OLS			OLS			ML		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
normal	0.2462	0.1165	0.1165	0.4055	0.2478	0.2461	0.0556	0.0448	0.0437
normal het.	0.3800	0.1897	0.1861	0.3388	0.2483	0.2480	0.3722	0.1813	0.1788
t_3	0.2407	0.1116	0.1150	0.3067	0.1845	0.1852	0.7607	0.2206	0.2208
t_3 het.	0.4052	0.1981	0.2041	0.2751	0.1867	0.1880	0.9706	0.3286	0.3210
uniform	0.2582	0.1212	0.1198	0.4343	0.2707	0.2697	0.0670	0.0517	0.0510
uniform het.	0.3860	0.1928	0.1982	0.3460	0.2838	0.2837	0.4293	0.1756	0.1754

Maximum likelihood is consistent only in the case of the normal distribution without heteroskedasticity, but Monte Carlo experiments show that it is robust to this assumption for tobit regression. In the case of the truncated model, maximum likelihood gives strongly biased estimates when normality and homoskedasticity assumption do not hold.

Meanwhile, OLS applied to a subset separated by a maximum score always gives unbiased estimates. Its variance is high due to the fact that the model is estimated on a smaller sample.

It would be interesting to analyze how to choose τ in order to minimize root mean square error. This problem is not addressed in this paper.

9 Conclusions

In this paper we presented *maximum score estimators* for broad classes of regression models. The concept of our method is based on generalizing score function of Manski (1975, 1985). We proved consistency when estimating the normalized vector of slopes under very mild conditions. Our estimators are robust to heteroskedasticity and consistent regardless of the shape of the distribution. We also showed how the maximum score can be used to remove the bias of OLS when estimating tobit and truncated regression models.

Monte Carlo simulations confirmed that these estimators are approximately unbiased in moderate samples. Also estimating models using OLS on the sample restricted by the maximum score generates approximately unbiased coefficients, but the variance is larger that the variance of classical estimators.

As far as future work is concerned, it would be interesting to derive the asymptotic

distribution for our class of maximum score estimators. It would also be interesting to verify if the bootstrap gives consistent estimates of confidence intervals. The problem of nonstationary regressors can also be addressed.

References

[1] Bierens, H. J., (2004), *Introduction to the Mathematical and Statistical Foundations of Econometrics*, Cambridge University Press

[2] Engle R. F., McFadden D. L., (1994), *Handbook of Econometrics*, vol. IV, Elsevier Science

[3] Ferguson T. S., (1996), *A Course in Large Sample Theory*, Chapman and Hall

[4] Greene W. H., (2003), *Econometric Analysis*, Prentice Hall

[5] Greene W. H., (1981), On the Asymptotic Bias of the Ordinary Least Squares Estimator of the Tobit Model, *Econometrica* 49, 505-513.

[6] Horowitz J. L., (1992), A Smoothed Maximum Score Estimator for the Binary Response model, *Econometrica* 60, 505-531

[7] Horowitz J. L., (2002), Bootstrap Critical Values for Tests Based on the Smoothed Maximum Score Estimator, *Journal of Econometrics* 111, 141-167

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

[9] Kim J., Pollard D., (1990), Cube Root Asymptotics, *Annals of Statistics* 18, 191-219.

[10] Manski, C. F., (1975), Maximum Score Estimation of the Stochastic Utility Model of Choice, *Journal of Econometrics* 3, 205-228

[11] Manski, C. F., (1985), Semiparametric Analysis of the Discrete Response. Asymptotic Properties of the Maximum Score Estimator, *Journal of Econometrics* 27, 313-333

[12] Moon, H. R., (2004), Maximum Score Estimation of a Nonstationary Binary Choice Model, *Journal of Econometrics* 122, 385-403

[13] Nocedal J., Wright S. J., (1999), *Numerical Optimization*, Springer-Verlag

[14] Rao R. R., (1962), Relations Between Weak and Uniform Convergence of Measures with Applications, *Annals of Mathematical Statistics* 33, 659-680

10 Appendix - general note on the proofs of consistency

Proofs of consistency use a similar technique as the proof of consistency of the estimator of Manski (1985). The proofs are conducted in several steps and are based on verifying assumptions of the theorem of consistency of M-estimators. The proofs of assumptions: (1) compactness of the parameter space, (2) the continuity of the limit function and (3) the uniform convergence of the sample function mimic analogous proofs of Manski (1985). The proofs about the maxima of the limit function are the author's own work.

11 Appendix A - consistency of maximum score for linear regression

First we show that $[\beta_N, \beta_0]$ is a strongly consistent estimator of $[\tilde{\beta}, \tilde{\beta}_\tau] = \frac{[\beta,\beta_\tau]}{\|[\beta,\beta_\tau]\|}$. Then, on the basis of this fact we show consistency of β_N^*.
The proof of this theorem is conducted in several steps and generally is based on verifying assumptions of the theorem of consistency of M-estimators.
The natural choice of function $Q_0(\theta)$ from this theorem is the expected value of function $\hat{Q}_n(\theta)$ defined by (10).

$$Q_0(b, b_0) = E[y\mathbf{1}(b^T x_i \geq b_0)] - \mu(E[\mathbf{1}(b^T x_i \geq b_0)] - \tau)^2 \tag{30}$$

11.1 The maxima of $Q_0(\theta)$

We prove in Lemma 11.1 , that maxima of $Q_0(\theta)$ are contained in arbitrary small neighborhood of point $[\tilde{\beta}, \tilde{\beta}_\tau]$, where β_τ is quantile of order $1-\tau$ of variable $\beta^T x$. Due to this lemma we may use the theorem of consistency of M-estimators with comment to this theorem about multiple maxima of the objective function (Engle, McFadden (1999), p. 2122).

Lemma 11.1 *Let*

$$Q_0^{mod}(b, b_0) = E[y\mathbf{1}(b^T x \geq b_0)] \quad \textit{subject to} \quad E[\mathbf{1}(b^T x \geq b_0)] = \tau \tag{31}$$

Function $Q_0^{mod}(b, b_0)$ has exactly one maximum, which is located in point $[\tilde{\beta}, \tilde{\beta}_\tau]$, where $\tilde{\beta}_\tau$ is quantile of order $1-\tau$ of $\tilde{\beta}^T x$.

Proof. Let us consider a different, arbitrary vector $[\delta, \delta_\tau]$, which also satisfies condition $P[\delta^T x \geq \delta_0] = \tau$ and $\|[\delta, \delta_\tau]\| = 1$. We show that it has lower value of the

criterion function.

$$
\begin{aligned}
&E[y|\tilde{\beta}^T x \geq \tilde{\beta}_\tau] - E[y|\delta^T x \geq \delta_0] \\
&= E[y|\beta^T x \geq \beta_\tau] - E[y|\delta^T x \geq \delta_0] \\
&= E[\beta_0 + \beta^T x + u|\beta^T x \geq \beta_\tau] - E[\beta_0 + \beta^T x + u|\delta^T x \geq \delta_0] \\
&= E[\beta_0|\beta^T x \geq \beta_\tau] + E[\beta^T x|\beta^T x \geq \beta_\tau] \\
&+ E[u|\beta^T x \geq \beta_\tau] - E[\beta_0|\delta^T x \geq \delta_0] \\
&- E[\beta^T x|\delta^T x \geq \delta_0] - E[u|\delta^T x \geq \delta_0] \\
&= \beta_0 + E[\beta^T x|\beta^T x \geq \beta_\tau] + 0 - \beta_0 - E[\beta^T x|\delta^T x \geq \delta_0] - 0 \\
&= E[\beta^T x|\beta^T x \geq \beta_\tau] - E[\beta^T x|\delta^T x \geq \delta_0] \qquad (32)
\end{aligned}
$$

There are following cases to consider

1. $\{x:\ \beta^T x \geq \beta_\tau\} = \{x:\ \delta^T x \geq \delta_0\}$
   Then $[\tilde{\beta}^T, \tilde{\beta}_\tau] = [\delta, \delta_0]$, due to:
   (a) condition of identification: $\|[\tilde{\beta}, \tilde{\beta}_\tau]\| = \|[\delta, \delta_0]\| = 1$, which excludes $[\delta, \delta_0] = c[\tilde{\beta}, \tilde{\beta}_\tau]$ for some c.
   (b) condition that the distribution of x is not contained in any proper linear subspace of R^K, which excludes collinear elements of vector x and parameters being linear combination of other parameters.
2. $\{x:\ \beta^T x \geq \beta_\tau\} \neq \{x:\ \delta^T x \geq \delta_0\}$
   We have

$$
\begin{aligned}
&E[\beta^T x|\beta^T x \geq \beta_\tau] - E[\beta^T x|\delta^T x \geq \delta_0] \\
&= \frac{1}{P[\beta^T x \geq \beta_\tau]} E[\beta^T x \mathbf{1}(\beta^T x \geq \beta_\tau)] - \frac{1}{P[\delta^T x \geq \delta_0]} E[\beta^T x \mathbf{1}(\delta^T x \geq \delta_0)] \\
&= \frac{1}{\tau} E\left[\beta^T x \left(\mathbf{1}(\beta^T x \geq \beta_\tau) - \mathbf{1}(\delta^T x \geq \delta_0)\right)\right] \\
&= \frac{1}{\tau} E\Big[\beta^T x \Big(\mathbf{1}(\beta^T x \geq \beta_\tau \wedge \delta^T x \geq \delta_0) + \mathbf{1}(\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0) \\
&- \mathbf{1}(\delta^T x \geq \delta_0 \wedge \beta^T x \geq \beta_\tau) - \mathbf{1}(\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau)\Big)\Big] \\
&= \frac{1}{\tau} E\left[\beta^T x \left(\mathbf{1}(\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0) - \mathbf{1}(\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau)\right)\right] \\
&= \frac{1}{\tau} \Big(E\left[\beta^T x \mathbf{1}(\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0)\right] \\
&- E\left[\beta^T x \mathbf{1}(\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau)\right]\Big) \qquad (33)
\end{aligned}
$$

Note that (the proof of this fact is shown later)

$$P[\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0] = P[\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau] = c \tag{34}$$

The constant c cannot be equal to zero, because the measure of this set is nonzero, due to the assumption that the conditional distribution of x has everywhere positive density. So

$$\begin{aligned}&\frac{1}{\tau}\left(E\left[\beta^T x \mathbf{1}(\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0)\right] - E\left[\beta^T x \mathbf{1}(\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau)\right]\right)\\ &\quad = \frac{c}{\tau}\left(E\left[\beta^T x | (\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0)\right] - E\left[\beta^T x | (\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau)\right]\right)\end{aligned} \tag{35}$$

Note that

$$\beta^T x | (\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0) > \beta^T x | (\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau) \tag{36}$$

because on the left side there is condition $\beta^T x \geq \beta_\tau$ and on the right $\beta^T x < \beta_\tau$. So

$$E\left[\beta^T x | (\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0)\right] > E\left[\beta^T x | (\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau)\right] \tag{37}$$

So finally

$$E[y | \beta^T x \geq \beta_\tau] - E[y | \delta^T x \geq \delta_0] > 0 \tag{38}$$

so

$$E[y | \tilde{\beta}^T x \geq \tilde{\beta}_\tau] - E[y | \delta^T x \geq \delta_0] > 0 \tag{39}$$

So $[\tilde{\beta}^T, \tilde{\beta}_\tau]$ is the only one maximum of (31).

Proof. Now we prove the following fact

$$P[\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0] = P[\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau] \tag{40}$$

We have

$$P[\beta^T x \geq \beta_\tau] = P[\delta^T x \geq \delta_0] = \tau \tag{41}$$

$$\begin{aligned}&P[\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0] + P[\beta^T x \geq \beta_\tau \wedge \delta^T x \geq \delta_0]\\ &= P[\delta^T x \geq \delta_0 \wedge \beta^T x \geq \beta_\tau] + P[\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau]\\ &= P[\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0]\\ &= P[\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau]\end{aligned} \tag{42}$$

Lemma 11.2 *M - the set of maxima of $Q_0(b, b_0)$ - is contained in arbitrary small neighborhood of point $[\tilde{\beta}, \tilde{\beta}_\tau]$, that is*

$$\forall_{\epsilon>0}\exists_{\mu>0} M \subset B([\tilde{\beta}, \tilde{\beta}_\tau], \epsilon), \tag{43}$$

where $B(x, r)$ is a ball centered in x and radius r.

Proof. Due to the theorem of convergence of quadratic penalty method we know that the set of maxima of $Q_0(b, b_0)$ converges to the set of maxima of $Q_0^{mod}(b, b_0)$, and due to the Lemma 11.1 we know that the set of maxima of $Q_0^{mod}(b, b_0)$ is equal $\{[\tilde{\beta}, \tilde{\beta}_\tau]\}$, which completes the proof.

11.2 Compactness of Θ

The set $[b, b_0] : \|[b, b_0]\| = 1$ is sphere in R^{K+1}, which is closed and bounded so it is compact.

11.3 Continuity of $Q_0(\theta)$

We have

$$Q_0(b, b_0) = E[y\mathbf{1}(b^T x_i \geq b_0)] - \mu(E[\mathbf{1}(b^T x_i \geq b_0)] - \tau)^2 \tag{44}$$

We show that the first component is continuous with respect b, b_0 to for $b_k \neq 0$ and that the expression in the square is also continuous with respect b, b_0 to for $b_k \neq 0$, which gives the continuity of the whole function which is a superposition of continuous functions.

Lemma 11.3 *$E[\mathbf{1}(b^T x_i \geq b_0)]$ under assumption 3.1 is continuous function of b, b_0 for $b_k \neq 0$.*

Proof. We prove this lemma for $b_k > 0$. The case when $b_k < 0$ is similar.

$$\begin{aligned} E[\mathbf{1}(b^T x \geq b_0)] &= \int_{R^K} \mathbf{1}(b^T x \geq b_0) dF_x = \int_{R^K} \mathbf{1}(\tilde{b}^T \tilde{x} + b_k x_k \geq b_0) dF_x \\ &= \int_{R^{K-1}} \int_R (x_k \geq \frac{b_0}{b_k} - \frac{\tilde{b}^T \tilde{x}}{b_k}) f_k(x_k|\tilde{x}) dx_k dF_{\tilde{x}} = \\ &= \int_{R^{K-1}} \int_{\frac{b_0}{b_k} - \frac{\tilde{b}^T \tilde{x}}{b_k}}^{\infty} f_k(x_k|\tilde{x}) dx_k dF_{\tilde{x}} \end{aligned} \tag{45}$$

The inner integral is a function of $\tilde{x}$, b i b_0 which is continuous with respect to b and b_0, measurable with respect to $\tilde{x}$ and uniformly bounded. Due to the Lebesgue convergence theorem $E[\mathbf{1}(b^T x_i \geq b_0)]$ is continuous function of b and b_0 when $b_k > 0$ and when $b_k < 0$.

Lemma 11.4 *$E[y\mathbf{1}(b^T x_i \geq b_0)]$ under assumption 3.1 is continuous function of b, b_0 for $b_k \neq 0$.*

Proof. The proof is similar to proof of Lemma 11.3. We prove lemma for $b_k > 0$.

$$\begin{aligned} &E[y\mathbf{1}(b^T x \geq b_0)] = E[(\beta_0 + \beta^T x + u)\mathbf{1}(b^T x_i \geq b_0)] \\ &= E[\beta_0 \mathbf{1}(b^T x_i \geq b_0)] + E[\beta^T x \mathbf{1}(b^T x_i \geq b_0)] + E[u\mathbf{1}(b^T x_i \geq b_0)] \\ &= \beta_0 E[\mathbf{1}(b^T x_i \geq b_0)] + E[\beta^T x \mathbf{1}(b^T x_i \geq b_0)] + 0 \end{aligned} \tag{46}$$

The continuity of the first component is proved in Lemma 11.3. We prove continuity of $E[\beta^T x \mathbf{1}(b^T x_i \geq b_0)]$. We have

$$E[\beta^T x \mathbf{1}(b^T x \geq b_0)] = \int_{R^K} \beta^T x \mathbf{1}(b^T x \geq b_0) dF_x = \tag{47}$$

$$= \int_{R^K} \beta^T x \mathbf{1}(\tilde{b}^T \tilde{x} + b_k x_k \geq b_0) dF_x$$

$$= \int_{R^{K-1}} \int_R \beta^T x \mathbf{1}(x_k \geq \frac{b_0}{b_k} - \frac{\tilde{b}^T \tilde{x}}{b_k}) f_k(x_k|\tilde{x}) dx_k dF_{\tilde{x}} = \tag{48}$$

$$= \int_{R^{K-1}} \int_{\frac{b_0}{b_k} - \frac{\tilde{b}^T \tilde{x}}{b_k}}^{\infty} \beta^T x f_k(x_k|\tilde{x}) dx_k dF_{\tilde{x}} \tag{49}$$

The inner integral is function of $\tilde{x}$, b and b_0, which is continuous with respect to b and b_0, measurable with $\tilde{x}$ and uniformly bounded (because the expected value of x exists). So due to the Lebesgue dominated convergence theorem, $E[y\mathbf{1}(b^T x_i \geq b_0)]$ is continuous function of b and b_0, when $b_k > 0$ and when $b_k < 0$.

11.4 Uniform convergence of $\hat{Q}_n(\theta)$ to $Q_0(\theta)$

We prove that $\hat{Q}_n(\theta_0) \to Q(\theta_0)$ in probability and for every $\epsilon > 0$ we have $\hat{Q}_n(\theta) < \hat{Q}_0(\theta) + \epsilon$ for every $\theta \in \Theta$ with probability approaching 1. These are the conditions stated in comment to the theorem of convergence of M-estimators (Engle, McFadden (1999), p. 2122). First

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^{N} y_i \mathbf{1}(\tilde{\beta}^T x_i \geq \tilde{\beta}_\tau) - \mu(\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\tilde{\beta}^T x_i \geq \tilde{\beta}_\tau) - \tau)^2 \overset{N\to\infty}{\longrightarrow} \\ &E[y\mathbf{1}(\tilde{\beta}^T x_i \geq \tilde{\beta}_\tau)] - \mu(E[\mathbf{1}(\tilde{\beta}^T x_i \geq \tilde{\beta}_\tau)] - \tau)^2 \ a.s. \end{aligned} \tag{50}$$

from the Kolmogorov strong law of large numbers and from the fact that if the sequence converges almost surely, then its continuous function converges almost surely to the function of its limit. So $\hat{Q}_n(\theta_0) \to Q(\theta_0)$ almost surely.
Now we may use the uniform law of large numbers for upper semicontinuous functions

(Ferguson 1996, p. 109). Function $\mathbf{1}(b^T x \geq b_0)$ is upper semicontinuous and it has only two values - 0 i 1. So the assumptions of the uniform law of large numbers for upper semicontinuous functions are met. So for every $\epsilon > 0$, $\hat{Q}_n(\theta) < \hat{Q}_0(\theta) + \epsilon$ for every $\theta \in \Theta$ almost surely.

12 Appendix B - consistency of maximum score for binomial regression

The proof is also based on verifying assumptions of the theorem of consistency of M-estimators. The natural choice of the function $Q_0(\theta)$ is expected value of the function $\hat{Q}_n(\theta)$ defined by equation(13).

$$Q_0(b, b_0) = E[y\mathbf{1}(b^T x \geq b_0)] - \mu(E[\mathbf{1}(b^T x \geq b_0)] - \tau)^2 \tag{51}$$

12.1 The maxima of $Q_0(\theta)$

We prove in Lemma 12.1 that maxima of function $Q_0(\theta)$ are contained in an arbitrary small neighborhood of point $[\tilde{\beta}, \tilde{\beta}_\tau] = \frac{[\beta, \beta_\tau]}{\|[\beta, \beta_\tau]\|}$, where $\tilde{\beta}_\tau$ is quantile of order $1 - \tau$ of variable $\tilde{\beta}^T x$. Due to this we may use the theorem of consistency of M-estimators with comment about multiple maxima of the objective function (Engle, McFadden (1999), p. 2124). First we show auxiliary lemma

Lemma 12.1 *Let*

$$Q_0^{mod}(b, b_0) = E[y\mathbf{1}(b^T x \geq b_0)] \quad \textit{subject to} \quad E[\mathbf{1}(b^T x \geq b_0)] = \tau \tag{52}$$

Function$Q_0^{mod}(b, b_0)$ *has exactly one maximum, which is in point* $[\tilde{\beta}, \tilde{\beta}_\tau]$*, where* $\tilde{\beta}_\tau$ *is quantile of order* $1 - \tau$ *of* $\tilde{\beta}^T x$*.*

Proof. Let us consider other vector $[\delta, \delta_\tau]$, which also satisfies condition $P[\delta^T x \geq \delta_0] = \tau$ and $\|[\delta, \delta_\tau]\| = 1$. We show that it has the lower value of the criterion function.

$$\begin{aligned} &E[y|\tilde{\beta}^T x \geq \tilde{\beta}_\tau] - E[y|\delta^T x \geq \delta_0] \\ &= E[y|\beta^T x \geq \beta_\tau] - E[y|\delta^T x \geq \delta_0] \\ &= E[g^{-1}(\beta^T x)|\beta^T x \geq \beta_\tau] - E[g^{-1}(\beta^T x)|\delta^T x \geq \delta_0] \end{aligned} \tag{53}$$

The following cases must be considered

1. $\{x: \ \beta^T x \geq \beta_\tau\} = \{x: \ \delta^T x \geq \delta_0\}$
   Then $[\tilde{\beta}^T x, \tilde{\beta}_\tau] = [\delta, \delta_0]$, due to:
   (a) condition of identification: $\|[\tilde{\beta}, \tilde{\beta}_\tau]\| = \|[\delta, \delta_0]\| = 1$, which excludes $[\delta, \delta_0] = c[\tilde{\beta}, \tilde{\beta}_\tau]$ for some c.

(b) condition, that the distribution x is not contained in any proper linear subspace of R^K, which excludes collinear elements of x and parameters being a linear combination of other parameters.

2. $\{x:\ \beta^T x \geq \beta_\tau\} \neq \{x:\ \delta^T x \geq \delta_0\}$
We have

$$
\begin{aligned}
&E[g^{-1}(\beta^T x)|\beta^T x \geq \beta_\tau] - E[g^{-1}(\beta^T x)|\delta^T x \geq \delta_0] \\
&= \frac{1}{P[g^{-1}(\beta^T x) \geq \beta_\tau]} E[g^{-1}(\beta^T x)\mathbf{1}(\beta^T x \geq \beta_\tau)] - \\
&\frac{1}{P[\delta^T x \geq \delta_0]} E[g^{-1}(\beta^T x)\mathbf{1}(\delta^T x \geq \delta_0)] \\
&= \frac{1}{\tau} E\left[g^{-1}(\beta^T x)\Big(\mathbf{1}(\beta^T x \geq \beta_\tau) - \mathbf{1}(\delta^T x \geq \delta_0)\Big)\right] \\
&= \frac{1}{\tau} E\Big[g^{-1}(\beta^T x)\Big(\mathbf{1}(\beta^T x \geq \beta_\tau \wedge \delta^T x \geq \delta_0) + \mathbf{1}(\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0) \\
&- \mathbf{1}(\delta^T x \geq \delta_0 \wedge \beta^T x \geq \beta_\tau) - \mathbf{1}(\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau)\Big)\Big] \\
&= \frac{1}{\tau} E\left[g^{-1}(\beta^T x)\Big(\mathbf{1}(\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0) - \mathbf{1}(\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau)\Big)\right] \\
&= \frac{1}{\tau}\Big(E\Big[g^{-1}(\beta^T x)\mathbf{1}(\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0)\Big] \\
&- E\Big[g^{-1}(\beta^T x)\mathbf{1}(\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau)\Big]\Big) \qquad (54)
\end{aligned}
$$

Note that

$$P[\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0] = P[\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau] = c \qquad (55)$$

The constant c cannot be equal to zero because the measure of this set is nonzero, due to the assumption that the conditional distribution of x has everywhere positive density. So

$$
\begin{aligned}
&\frac{1}{\tau}\Big(E\Big[g^{-1}(\beta^T x)\mathbf{1}(\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0)\Big] \\
&- E\Big[g^{-1}(\beta^T x)\mathbf{1}(\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau)\Big]\Big) \\
&= \frac{c}{\tau}\Big(E\Big[g^{-1}(\beta^T x)|(\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0)\Big] \\
&- E\Big[g^{-1}(\beta^T x)|(\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau)\Big]\Big) \qquad (56)
\end{aligned}
$$

Note that

$$g^{-1}(\beta^T x)|(\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0) > g^{-1}(\beta^T x)|(\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau) \tag{57}$$

because on the left side we have condition $\beta^T x \geq \beta_\tau$ and on the right $\beta^T x < \beta_\tau$ and function $g^{-1}(\cdot)$ is increasing. So

$$E\left[g^{-1}(\beta^T x)|(\beta^T x \geq \beta_\tau \wedge \delta^T x < \delta_0)\right] > E\left[g^{-1}(\beta^T x)|(\delta^T x \geq \delta_0 \wedge \beta^T x < \beta_\tau)\right] \tag{58}$$

So finally

$$E[y|\beta^T x \geq \beta_\tau] - E[y|\delta^T x \geq \delta_0] > 0 \tag{59}$$

so

$$E[y|\tilde{\beta}^T x \geq \tilde{\beta}_\tau] - E[y|\delta^T x \geq \delta_0] > 0 \tag{60}$$

So $[\tilde{\beta}^T, \tilde{\beta}_\tau]$ is the only one maximum of function (31).

Lemma 12.2 *M - the set of maxima of $Q_0(b, b_0)$ - is contained in arbitrary small neighborhood of $[\tilde{\beta}, \tilde{\beta}_\tau]$*

$$\forall_{\epsilon>0}\exists_{\mu>0} M \subset B([\tilde{\beta}, \tilde{\beta}_\tau], \epsilon) \tag{61}$$

Proof. Due to the theorem of convergence of quadratic penalty method, we know that the set of maxima of $Q_0(b, b_0)$ converges to the set of maxima of $Q_0^{mod}(b, b_0)$, and due to the Lemma 12.1 we know that the set of maxima of $Q_0^{mod}(b, b_0)$ is equal $\{[\tilde{\beta}, \tilde{\beta}_\tau]\}$, which completes proof.

12.2 Compactness Θ

The set $[b, b_0] : \|[b, b_0]\| = 1$ is a sphere in R^{K+1}, which is closed and bounded, so compact.

12.3 Continuity of $Q_0(\theta)$

We have

$$Q_0(b, b_0) = E[y\mathbf{1}(b^T x_i \geq b_0)] - \mu(E[\mathbf{1}(b^T x_i \geq b_0)] - \tau)^2 \tag{62}$$

We show that the first component is continuous with respect to b, b_0 for $b_k \neq 0$. The expression in square is also continuous with respect to b, b_0 for $b_k \neq 0$ (this fact was shown in the proof for linear regression), which gives continuity of the whole function as superposition of the continuous functions.

Lemma 12.3 *$E[y\mathbf{1}(b^T x_i \geq b_0)]$ under assumption 4.1 is continuous function of b, b_0 for $b_k \neq 0$.*

Proof. The proof is similar to the proof of Lemma 11.3. Similarly we prove lemma for $b_k > 0$.

$$E[y\mathbf{1}(b^T x \geq b_0)] = E[g^{-1}(\beta^T x)\mathbf{1}(b^T x_i \geq b_0)] \tag{63}$$

We have

$$\begin{aligned} E[g^{-1}(\beta^T x)\mathbf{1}(b^T x \geq b_0)] &= \int_{R^K} g^{-1}(\beta^T x)\mathbf{1}(b^T x \geq b_0) dF_x \\ &= \int_{R^K} g^{-1}(\beta^T x)\mathbf{1}(\tilde{b}^T \tilde{x} + b_k x_k \geq b_0) dF_x \\ &= \int_{R^{K-1}} \int_R g^{-1}(\beta^T x)\mathbf{1}(x_k \geq \frac{b_0}{b_k} - \frac{\tilde{b}^T \tilde{x}}{b_k}) f_k(x_k|\tilde{x}) dx_k dF_{\tilde{x}} \\ &= \int_{R^{K-1}} \int_{\frac{b_0}{b_k} - \frac{\tilde{b}^T \tilde{x}}{b_k}}^{\infty} g^{-1}(\beta^T x) f_k(x_k|\tilde{x}) dx_k dF_{\tilde{x}} \end{aligned} \tag{64}$$

The inner integral is a function of $\tilde{x}$, b i b_0 which is continuous with respect to b and b_0, measurable with respect to $\tilde{x}$ and uniformly bounded, because g^{-1} is bounded and measurable. Due to the Lebesgue convergence theorem $E[\mathbf{1}(b^T x_i \geq b_0)]$ is a continuous function of b and b_0 when $b_k > 0$ and when $b_k < 0$.

12.4 Uniform convergence of $\hat{Q}_n(\theta)$ to $Q_0(\theta)$

We have

$$\begin{aligned} &E[y\mathbf{1}(b^T x_i \geq b_0)] - \mu(E[\mathbf{1}(b^T x_i \geq b_0)] - \tau)^2 \\ &= \frac{P[\beta_0 + \beta^T x + u \geq 0 \wedge b^T x \geq b_0]}{P[b^T x \geq b_0]} - \mu(P[b^T x_i \geq b_0] - \tau)^2 \end{aligned} \tag{65}$$

and

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^N y_i \mathbf{1}(b^T x_i \geq b_0) - \mu(\frac{1}{N}\sum_{i=1}^N \mathbf{1}(b^T x_i \geq b_0) - \tau)^2 \\ &= \frac{P_N[\beta_0 + \beta^T x + u \geq 0 \wedge b^T x \geq b_0]}{P_N[b^T x \geq b_0]} - \mu(P_N[b^T x_i \geq b_0] - \tau)^2 \end{aligned} \tag{66}$$

Every set, the probability of which is calculated above, is a halfspace or a sum of two halfspaces. We may use the theorem of Rao (Rao R. R. (1962), p. 675), and obtain the result that the elements of (66) converge to corresponding elements of (65) uniformly in b and b_0.

13 Appendix C - consistency of maximum score for truncated regression

The proof of this theorem is similar to the proof of consistency for linear regression. However it must be commented.

Comment 13.1 *The conditional expected value of y in the truncated regression model is given by*

$$\begin{aligned}
&E[y|y>C] \\
&= \frac{1}{P[y>C]}E[y\mathbf{1}(y>C)] \\
&= \frac{1}{P[\beta_0+\beta^T x+u>C]}E[(\beta_0+\beta^T x+u)\mathbf{1}(\beta_0+\beta^T x+u>C)] \\
&= \frac{1}{P[\beta_0+\beta^T x+u>C]}\Big(E[(\beta_0+\beta^T x)\mathbf{1}(\beta_0+\beta^T x+u>C)] \\
&+ E[u\mathbf{1}(\beta_0+\beta^T x+u>C)]\Big) \\
&= \frac{1}{P[\beta_0+\beta^T x+u>C]}\Big((\beta_0+\beta^T x)P[\beta_0+\beta^T x+u>C] \\
&+ E[u\mathbf{1}(\beta_0+\beta^T x+u>C)]\Big) \\
&= \beta_0+\beta^T x+\frac{1}{1-F_u(C-\beta_0-\beta^T x)}\int_{C-\beta_0-\beta^T x}^{\infty} udF_u \qquad (67)
\end{aligned}$$

Note that for fixed F_u we have

$$F_u(C-\beta_0-\beta^T x)\to 0 \quad for \quad \beta_0+\beta^T x\to\infty \qquad (68)$$

$$\int_{C-\beta_0-\beta^T x}^{\infty} udF_u \to E[u]=0 \quad for \quad \beta_0+\beta^T x\to\infty \qquad (69)$$

Note that the distribution F_u may depend on x (heteroskedasticity) so not always the last term in (67) converges to zero when $\beta_0+\beta^T x$ increases. The additional set of assumptions is necessary to ensure this property. However if

$$\frac{1}{1-F_{u|x}(C-\beta_0-\beta^T x)}\int_{C-\beta_0-\beta^T x}^{\infty} udF_{u|x}\to 0 \quad for \quad \beta_0+\beta^T x\to\infty \qquad (70)$$

then $argmaxQ_0(\theta)$ when τ decreases, so for sufficiently large values of $\beta_0+\beta^T x$, converges to $[\beta,\beta_\tau]$, like in case of linear regression.

14 Appendix D - consistency of OLS applied to a restricted sample

Proof. We use Comment 13.1. Because β_N i β_{0N} are consistent estimators of $\tilde{\beta}$ and $\tilde{\beta}_\tau$ then

$$\frac{1}{1-F_{u|x}(C-\beta_{0N}-\beta_N^T x)}\int_{C-\beta_{0N}-\beta_N^T x}^{\infty} udF_{u|x} \overset{n\to\infty}{\longrightarrow} \frac{1}{1-F_{u|x}(C-\beta_0-\beta^T x)}\int_{C-\beta_0-\beta^T x}^{\infty} udF_{u|x} \tag{71}$$

with probability 1.
Then

$$\frac{1}{1-F_{u|x}(C-\beta_0-\beta^T x)}\int_{C-\beta_0-\beta^T x}^{\infty} udF_{u|x} \to 0 \quad \text{for} \quad \beta_0+\beta^T x \to \infty \tag{72}$$

So using (71) and (72) we obtain

$$\lim_{N\to\infty}\lim_{\beta_{0N}+\beta_N^T x\to\infty}\frac{1}{1-F_{u|x}(C-\beta_{0N}-\beta_N^T x)}\int_{C-\beta_{0N}-\beta_N^T x}^{\infty} udF_{u|x}=0 \tag{73}$$

with probability 1.
We may use the theorem for omitted variable problem (Greene W. (2003), p. 148) and as omitted variable $\mathbf{x}_2$ from this theorem we take $\frac{1}{1-F_{u|x}(C-\beta_0-\beta^T x)}\int_{C-\beta_0-\beta^T x}^{\infty} udF_{u|x}$. with parameter $\beta_2=1$. So the bias converges to zero for $N\to\infty$ and $\tau\to 0$.