Binary choice

Introduction to binary choice

Parametric binary choice

Estimation: logit/Probit

Estimation: M-estimator

Interpetation of β

Semiparametrics: Manski (1975/85/88)

Introduction to binary choice

What is binary choice?

Many economic decisions are yes/no:

IO

- Firm: do you enter a market?
- Firm: do you export?
- Consumer: do you purchase a certain product?

Labor

- Do you participate in the labor market?
- Do you attend college?
- Do you enroll in a labor market training program?

What is binary choice? (2)

Finance

- Do you enroll in a savings program?
- Do you invest in stocks (1) or bonds (0)?

Health

- Do you smoke?
- Do you have private (1) or public (0) health insurance?

. . .

Factors influencing the binary choice

What determines whether you smoke or not?

- do your parents smoke?
- do your friends smoke?
- price of cigarettes?

Example

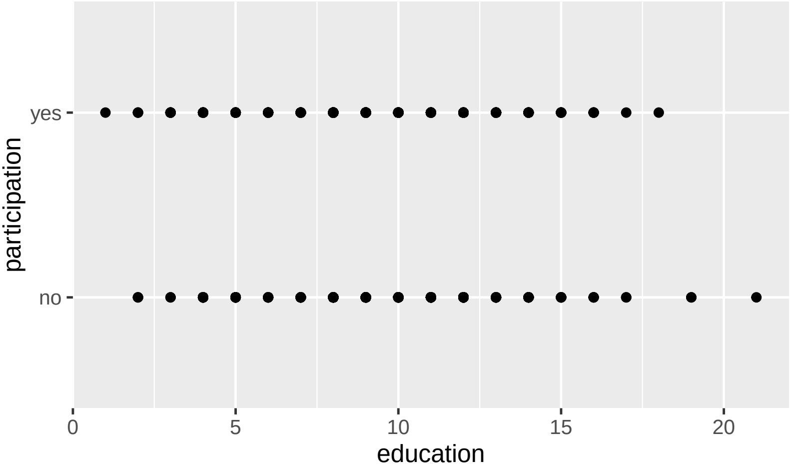
Female labor participation in Switzerland, 872 women, 1990's.

```
library(AER)
data("SwissLabor")
head(SwissLabor)
```

```
##   participation   income age education youngkids oldkids
## 1            no 10.78750 3.0         8         1       1
## 2           yes 10.52425 4.5         8         0       1
## 3            no 10.96858 4.6         9         0       0
## 4            no 11.10500 3.1        11         2       0
## 5            no 11.10847 4.4        12         0       2
## 6           yes 11.02825 4.2        12         0       1
```

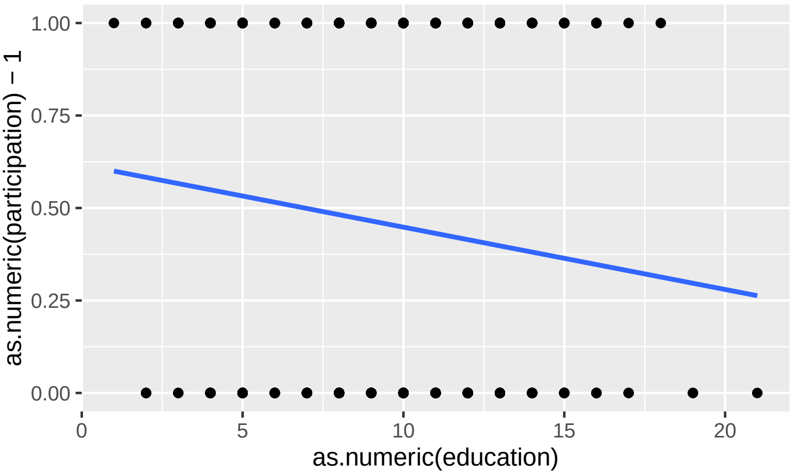
Example (2)

```
library(ggplot2)
qplot(data=SwissLabor,x=education,y=participation)
```



Example (3)

```
qplot(data=SwissLabor,x=as.numeric(education),y=as.numeric(
```



Binary choice models

- Goal: ministries of finance and health seek optimal taxation on tabacco
- We need a regression model to quantify the effect of taxes on smoking incidence
- OLS does not work
  - $Y \in \{0, 1\}$, but $\hat{Y} \notin [0, 1]$
  - predictions and causal effects from OLS are invalid
- We need a model that takes into account that Y is binary

Binary choice models.

Why are binary choice models important?

Why are binary choice models important?

1. Lots of examples of binary dependent variables
2. Example of a structural model (Roy model)
3. Program evaluation methods
4. Extensions: semiparametric versions

Parametric binary choice

Model: Words

- Individual deciding whether to enter labor force
- Receives utility

$$Y_i^* = X_i\beta + u_i$$

  if they enter, 0 otherwise.
- Individual preferences are u_i, unknown to the researcher
- **Utility maximization**: enter labor force if and only if $Y_i^* > 0$
- Q: What is X_i? What sign does β have?

Statistical model

$$Y_i^* = X_i\beta + u_i$$

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases}$$

$$u_i | X_i \sim F(u|X_i)$$

Statistical model: classification

Parametric models

- If F is standard logistic, this is the **logit** model
- If F is standard normal, this is the **probit** model

Semiparametric models

If F is unspecified, this model is semiparametric.

- Manski's BC model: $med(u_i|X_i) = 0$
- Klein and Spady's model: $F(u|X_i) = F(u)$

Probit model

The probit model assumes

$$F(u|X_i) = \Phi(u)$$

. That means:

1. $u_i \perp X_i$
2. (mean, variance) of u_i are (0,1)
3. $P(u_i \leq u) = \Phi(u)$

Logit model

Same, but

$$
\begin{aligned}
P(u_i \leq u) &= \Lambda(u) \\
&= \frac{\exp(u)}{1 + \exp(u)}
\end{aligned}
$$

Model probabilities

- The research does not observe Y_i^*
- The research observes (Y_i, X_i)
- The distribution of these observable quantities is determined by

$$P(Y_i = y, X_i = x) = P(Y_i = y | X_i = x) P(X_i = x)$$

- Ignore $P(X_i = x)$ (functional independence)
- **Model** leads to an expression for $P(Y_i = y | X_i = x)$

Identification

Assume the logit or probit model, or some other **invertible link function**.

`sketch-2.1.png`

Non-identification

Weaken the assumption a little bit: assume that

$$u_i | X_i \sim \mathcal{N}(0, \sigma^2)$$

with unknown σ.

- Q: Are β and σ identifiable?

Non-identification: Answer

- No!
- Knowledge about (β, σ) comes to us exclusively through

$$P(Y_i = 1 | X_i) = \Phi(X_i \beta / \sigma)$$

- Consider $(\beta, \sigma) = (3, 1)$ and $(\beta', \sigma') = (6, 2)$ ...

Non-identification: interpretation

Q: What does this mean?

Non-identification: interpretation: answer

The underlying utility scale is ordinal. Changing σ stretches this scale.

Estimation: logit/Probit

Example (4): Logit

```
##
## Call:
## glm(formula = participation ~ education + age + foreign,
##     data = SwissLabor)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6105  -1.0196  -0.8995   1.3156   1.4997
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.180378   0.429586   0.420   0.6746
## education   -0.007972   0.026219  -0.304   0.7611
## age         -0.138386   0.068891  -2.009   0.0446 *
## foreignyes   1.161069   0.185449   6.261 3.83e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
```

Example (5)

1. How did we obtain these estimates, i.e. what is the underlying statistical procedure?
2. How do we interpret these estimates?

Example (6): Probit / identification

```
##
## Call:
## glm(formula = participation ~ education + age + foreign,
##     data = SwissLabor)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6148  -1.0200  -0.8983   1.3154   1.5023
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.116127   0.265048   0.438    0.661
## education   -0.005189   0.016154  -0.321    0.748
## age         -0.086719   0.042442  -2.043    0.041 *
## foreignyes   0.721083   0.113658   6.344 2.23e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
```

Intro to ML

- Consider a sequence of coin tosses

$$(X_1, ..., X_4) = (H, T, T, H)$$

with $P(X_i = H) = p$ unknown.

Intro to ML

- Consider a sequence of coin tosses

$$(X_1, ..., X_4) = (H, T, T, H)$$

  with $P(X_i = H) = p$ unknown.
- If you had to guess that p is one of $p \in \{0.5, 0.8\}$, what would you say?

Intro to ML (2)

- If $p = 0.5$, then the probability of seeing the sequence is

```
0.5^4
```

```
## [1] 0.0625
```

- If $p = 0.8$, then the probability of seeing the sequence is

```
0.8*(1-0.8)*(1-0.8)*0.8
```

```
## [1] 0.0256
```

- Conclusion?

General coin flip formulation

If we allow for $p \in [0,1]$, then

- the *likelihood contribution* for a given observation $X_i \in \{0,1\}$ depends on p, and is
$$L_i(p) = p^{X_i}(1-p)^{1-X_i}.$$
- the *likelihood* of the sample is
$$L_n(p) = \prod_i L_i(p) = p^{\sum_i X_i}(1-p)^{\sum_i (1-X_i)}$$
- Letting $\sum_i X_i = n_1$, we can write this as
$$L_n(p) = \prod_i L_i(p) = p^{n_1}(1-p)^{n-n_1}$$
- The log-likelihood is
$$\mathcal{L}_n(p) = 1/n \log(L_n(p)) = n_1/n \log(p) + (n-n_1)/n \log(1-p)$$
- The derivative of the log-likelihood is
$$s_n(p) = \sum_i s_i(p)$$

log/likelihood estimator

- The maximum likelihood estimator is defined as

$$\hat{p} = argmax_{p \in [0,1]} L_n(p)$$

- Because increasing transformations preserve the location of the maximum,

$$\hat{p} = argmax_{p \in [0,1]} \mathcal{L}_n(p)$$

- Under smoothness and identification conditions, it is defined through

$$s_n(\hat{p}) = 0$$

Coin flip solution

First order condition is . . .

whiteboard

Practice

Quasi-homework:

1. Let X_i be Poisson(λ). Obtain the MLE for λ for a random sample $(X_1, \cdots, X_n)$.
2. Let $Y_i|X_i \sim \mathcal{N}(X_i\beta, \sigma^2)$. Deliver $\hat{\beta}$ and $\hat{\sigma}$. What about $\widehat{\sigma^2}$?

Logit conditional likelihood

- The contribution to the log-likelihood is

$$log(P(Y = y|X = x)) + log(P(X = x))$$

  so we can focus on the first term
- Conditional log-likelihood contribution is ...

whiteboard

Solution

For symmetric F, the score is

$$s_i(p) = \frac{Y_i}{F(X_i\beta)} f(X_i\beta) - \frac{1 - Y_i}{1 - F(X_i\beta)} f(X_i\beta)$$

so that the MLE $\hat{\beta}$ solves

$$\sum_i f(X_i\hat{\beta}) \left(\frac{Y_i}{F(X_i\hat{\beta})} - \frac{1 - Y_i}{1 - F(X_i\hat{\beta})} \right)$$

- No explicit solution
- Interpret the FOC?

Estimation: M-estimator

Setup

The maximum likelihood estimator is an example of an **extremum estimator**. Ingredients:

- a criterion function $Q_n(\theta)$
- a parameter space Θ known to contain true value θ_0
- an estimator defined through Q_n, i.e.

$$\hat{\theta}_n = argmax_{\Theta} Q_n(\theta)$$

Examples

- For the BC estimator above, $Q_n = \mathcal{L}_n(\beta)$
- OLS: $Q_n(\beta) = \sum_i (Y_i - X_i\beta)^2$
- Moment conditions: if $E(m(Z;\theta_0)) = 0$ then build

$$Q_n(\theta) = 1/n \sum m(Z_i,\theta)' W m(Z_i,\theta),$$

  - covers ML, NLS, OLS, IV, ...
  - analog principe: *whiteboard*

Result

In the context of the M-estimation setup, assume that there exists a function $Q_0(.)$ such that

1. $Q_0(\theta) = 0 \Leftrightarrow \theta = \theta_0$
2. Q_0 is continuous
3. Q_n converges uniformly to Q_0

Finally, assume that

4.Θ is compact.

Under conditions 1-4, $\hat{\theta}_n$ converges to θ_0 in probability.

Uniform convergence

- Uniform convergence says:

$$sup_{\Theta}|Q_n(\theta) - Q_0(\theta)| \rightarrow 0$$

  in probability, as $n \rightarrow \infty$.
- More details in Newey and McFadden, and when we come to nonsmooth objective functions

Proof sketch

whiteboard

Details

- My note pdf
- Newey and McFadden, Section 2

Application to MLE

Homework. Link the results in Newey and McFadden to the general extremum estimator results.

Details: PS1.

Interpetation of β

Recap

So far:

- why we need binary choice
- model
- estimator + consistency

Now: what do I do with the results?

Partial effects

- The partial or marginal effect is the quantity of interest to the applied microeconometrician.
- Simply stated: how does Y change with X
- For now, that will mean

$$\frac{\partial E(Y|X)}{\partial X_k}$$

or

$$E(Y|X = x') - E(Y|X = x)$$

- When is this a causal effect? See: *program evaluation*
- What about other properties of the conditional distribution? See: *quantile regression*

Linearity

- The linear model is special because

$$PE_k = \beta_k$$

- The regression coefficient contains all the relevant information
- Nonlinear models do not work like that

Binary choice partial effect

For the binary choice model

$$E\left(Y_i|X_i\right) = 1 \times P(Y_i = 1|X_i) + 0 \times P(Y_i = 0|X_i).$$

If F is symmetric, then $E\left(Y_i|X_i\right) = F(X_i\beta)$ the partial effect for individual i with respect to regressor k is

$$\frac{\partial E(Y_i|X_i = x)}{\partial x_k} = \beta_k f(x\beta)$$

PE - notes

1. PE has the same sign as β_k
2. magnitude depends on X_i
   - vanishes as $X_i\beta \to \pm\infty$
   - maximized at the center

whiteboard:2.5

PE - Logit

Back to the logit model. We know that

$$\Lambda(u) = \exp(u)/(1 + \exp(u))$$

and that its derivative is

$$\lambda(u) = \Lambda(u)[1 - \Lambda(u)]$$

so that the partial effect is equal to

$$\beta_k \Lambda()[1 - \Lambda()] \leq \beta_k/4$$

Logit / Probit

For logit and probit to give you similar results in the center, we would need

$$\hat{\beta}_1 \lambda(0) = \hat{\beta}_2 \phi(0)$$

so that their ratio should be close to

```r
dnorm(0)/dlogis(0)
```

```
## [1] 1.595769
```

APE v PEA

Three effects.

1. Individual, $\hat{\tau}_i = \hat{\beta}_k f(X_i \hat{\beta})$
2. Partial effect at the average, $PEA = \hat{\beta}_k f(\bar{X} \hat{\beta})$
3. Average partial effect $\hat{\tau} = 1/n \sum_i \hat{\tau}_i$

Wooldridge has in-depth discussions.

Alt interpretation

Log odds. See Wooldridge.

Application: Swiss Labor

```
##
## Call:
## glm(formula = participation ~ youngkids + education + ag
##     family = binomial(link = "logit"), data = SwissLabor
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9230  -0.9775  -0.5934   1.1245   2.1916
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.8764670  0.4969653   3.776 0.000159 ***
## youngkids   -1.2657010  0.1658454  -7.632 2.31e-14 ***
## education   -0.0001148  0.0274396  -0.004 0.996661
## age         -0.5000419  0.0852590  -5.865 4.49e-09 ***
## foreignyes   1.3522275  0.1983002   6.819 9.16e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
```

Application: having a child

```
baseline <- predict.glm(model,type="response")
df_cf <- SwissLabor
df_cf$youngkids  <- df_cf$youngkids + 1
newval <- predict.glm(model,newdata = df_cf,type="response'
mean(newval-baseline)
```

```
## [1] -0.2380156
```

Having an additional child

```
df_cf <- subset(SwissLabor,youngkids>0)
baseline <- predict.glm(model,newdata=df_cf,type="response'
df_cf$youngkids  <- df_cf$youngkids + 1
newval <- predict.glm(model,newdata = df_cf,type="response'
mean(newval-baseline)
```

```
## [1] -0.1926308
```

Semiparametrics: Manski (1975/85/88)

Overview: binary choice

$$Y_i^* = X_i\beta + u_i$$

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* <= 0 \end{cases}$$

$$u_i | X_i \sim F$$

- If F is the standard logistic, this is the **logit** model
- If F is the standard normal, this is the **probit** model
- If F is unspecified, this model is semiparametric

Identification

1. For identification of β the **linear model**, all we need is
   - $E(u_i|X_i) = 0$
   - $E(X_i'X_i)$ is invertible
2. For identification of β in the parametric binary choice model, we imposed
   - $u_i|X_i$ has known distribution function (mean, variance, shape)
   - $E(X_i'X_i)$ invertible

Q: Can we relax the distributional assumption?

Identification (Manski, 1988)

- Parameter of interest: $(\beta, F_{u|X})$
- Assumption X1: no multicollinearity
- Assumption X3: at least one continuous regressor with non-0 coefficient

0. $F_{u|X}$ is known and X1: logit/probit
1. $E(u|X) = 0$: no identification
2. $med(u|X) = 0$ and (X1,X3): identification of β up to scale
3. $u \perp X$ and (X1,X3): identification of $(\beta, F_{u|X})$ (see nonparametrics)

Case 1: $E(u|X) = 0$

- Sufficient for identification in the linear model
- Here, the only thing you observe is $P(Y_i = 1|X_i)$ and the marginal distribution of X_i
- True values $(F(u|X), \beta)$ should satisfy:
  1. match: $P(Y_i = 1|X_i = x) = F_{u|X}(x\beta)$ for all x, for the observable $P(Y_i|X_i)$
  2. mean-zero-ness: $E(u|X = x) = \int_u udF(u|x) = 0$for all xinsupp(X)

We will construct $(b, G(u|X))$ that also satisfies (1) and (2)

Case 1 (2)

1. Pick $b \neq \beta$.
2. Then, construct $G(u|X)$ such that *match* and *meanzeroness* are satisfied.

Point 2. can be done for each x separately.

0. Consider an x such that $0 > xb > x\beta$.
1. Then $F(xb) > F(x\beta)$ because of increasingness
2. Construct $\tilde{G}$ by removing some mass on the left
3. Construct G by putting the mass back on the right

[sketch-2.2.jpg]

Case 2: Manski, 1975

$$
\begin{aligned}
Y^* &= X\beta + u \\
Y &= \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{if } Y^* \leq 0 \end{cases} \\
u|X &\sim F_{u|X}
\end{aligned}
$$

Assumption 1. There exists a unique $\beta \in \mathbb{R}^K$ such that $med(Y|X) = X\beta$.

Manski: Key insight

Ley $\tilde{Y} = sgn(Y^*)$. Because

$$\begin{aligned} E(\tilde{Y}|X) &= 1 \times P(Y^* \geq 0|X) + (-1) \times (1 - P(Y^* \geq 0|X)) \\ &= 2P(Y^* \geq 0|X) - 1 \\ &= 2P(-u \leq X\beta|X) - 1 \end{aligned}$$

it follows that

- $X\beta > 0 \Leftrightarrow E(\tilde{Y}|X) > 0$
- $X\beta = 0 \Leftrightarrow E(\tilde{Y}|X) = 0$
- $X\beta < 0 \Leftrightarrow E(\tilde{Y}|X) < 0$

Manski: takeaway

Estimating equation:

$$sgn(E(\tilde{Y}|X)) = sgn(X\beta)$$

- Left: observable
- Right: model parameters
- Identification / estimation?

Assumption 2

Assumption 2. (a) The support of F_x is not contained in any proper linear subspace of $\mathbb{R}^K$. (b) $0 < P(\tilde{Y} \geq 0|X) < 1$, a.e. F_x. (c) There exists at least one $k \in \{1, \cdots, K\}$ such that, for almost every value of $\bar{X} = (X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_K)$, the distribution of X_k conditional on X_{k-1} has everywhere positive Lebesgue density.

Identification: notation

Let $b \in \mathbb{R}^K, b \neq \beta$. For each $x \in \mathbb{R}^K$,

- b cannot be differentiated from β at x if $sgn(xb) = sgb(x\beta)$
- b can be differentiated from β at x if $sgn(xb) \neq sgb(x\beta)$

Define the set

$$X_b = \{x \in \mathbb{R}^K : sgn(xb) \neq sgn(x\beta)\}$$

and its measure:

$$R(b) = \int_{X_b} dF_x$$

Identification: definition

Definition. $g(\beta)$ is identified relative to B - the parameter space for β - if and only if $R(b) > 0$ for every $b \in B$ such that $g(b) \neq g(\beta)$.

- Observation: for every $a > 0$, $R(a\beta) = 0$.
- Conclusion: pick $g(b) = b/\|b\|$
- Interpretation: β is only identified up to scale

No identification without full support

Assume

- $K \geq 2$,
- B contain a neighbourhood of β,
- F_X has finite support
- there exists a $\lambda > 0$ such that $|x\beta| \geq \lambda|$ a.e.

Then identification fails. See Manski, 1988, p. 316-317.

Identification

Lemma 2. Under Assumption 2, $\beta/\|\beta\|$ identified.

Assumption 2. (a) The support of F_x is not contained in any proper linear subspace of $\mathbb{R}^K$. (b) $0 < P(\tilde{Y} \geq 0|X) < 1$, a.e. F_x. (c) There exists at least one $k \in \{1, \cdots, K\}$ such that, for almost every value of $\bar{X} = (X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_K)$, the distribution of X_k conditional on X_{k-1} has everywhere positive Lebesgue density.

[proof: `sketch-2.3.pdf`]

Estimation

Assume that a random sample $(\tilde{y}_i, X_i,\ i = 1, \cdots, n)$ is available. Define

$$S_n(b) = 1/n \sum_i \tilde{y}_i sgn(X_i b)$$

and the **maximum score estimator**

$$\hat{\beta} = argmax S_n(b)$$

defined in Manski (1975)

Consistency

Manski (1985) shows that

1. Under random sampling, S_n converges uniformly to

$$S(b) = E(\tilde{y} \times sgn(Xb))$$

2. Under Assumptions 1+2, S is uniquely maximized at β (up to scale)
3. S is continuous for all $b,\ b_k \neq 0$

Asymptotic distribution

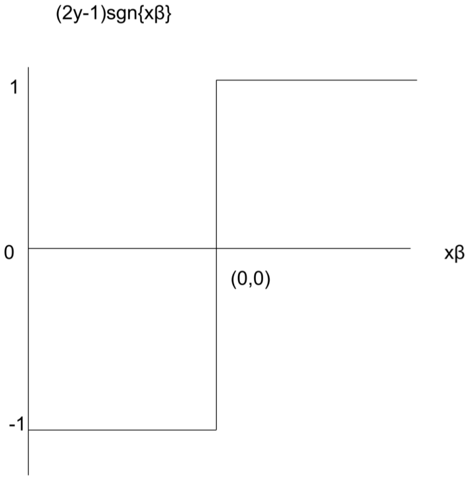
- Consistency (Manski, 1975, 1985)
- Cube-n rate (Kim and Pollard, 1990, AoS)
- Bootstrap does not work (Abrevaya and Huang, 2005, ECTA)

But:

- Smoothed version by Horowitz (1992, ECTA)
  - close to root-n
  - asymptotically Normal

Nonsmooth objective function

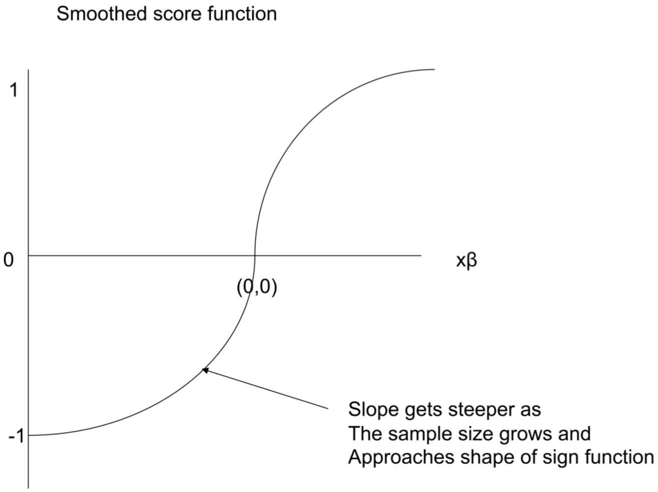
Pictures taken from Todd's 721 notes.



Nondifferentiability creates difficulties
for the asymptotic distribution theory

Smooth objective function

Pictures taken from Todd's 721 notes.



Horowitz estimator

Define

$$S_n(b) = 1/n \sum_i y_i^* K\left(\frac{X_i b}{h_n}\right)$$

where $h_n \to 0$.

The **smooth maximum score estimator** is

$$\hat{\beta} = argmax S_n(b)$$

defined in Manski (1975).

- K is a *kernel function*, which plays an important role in nonparametric estimation

Simulations

See *maxscore.R*

- Note: *thin set identification*