Qualitative Response Model
質的反応モデル

内容

- 被説明変数がダミー変数の場合の推定方法
  - 線型確率モデル
  - Probitモデル
  - Logitモデル
- Tobitモデル
  - 被説明変数がある水準以下または以上で打ち切られている場合
  - censored regression または truncated regression ともよばれる

被説明変数がダミー変数の回帰

- 被説明変数が1または0をとるダミー変数の場合の推定方法

例）mroz.xls

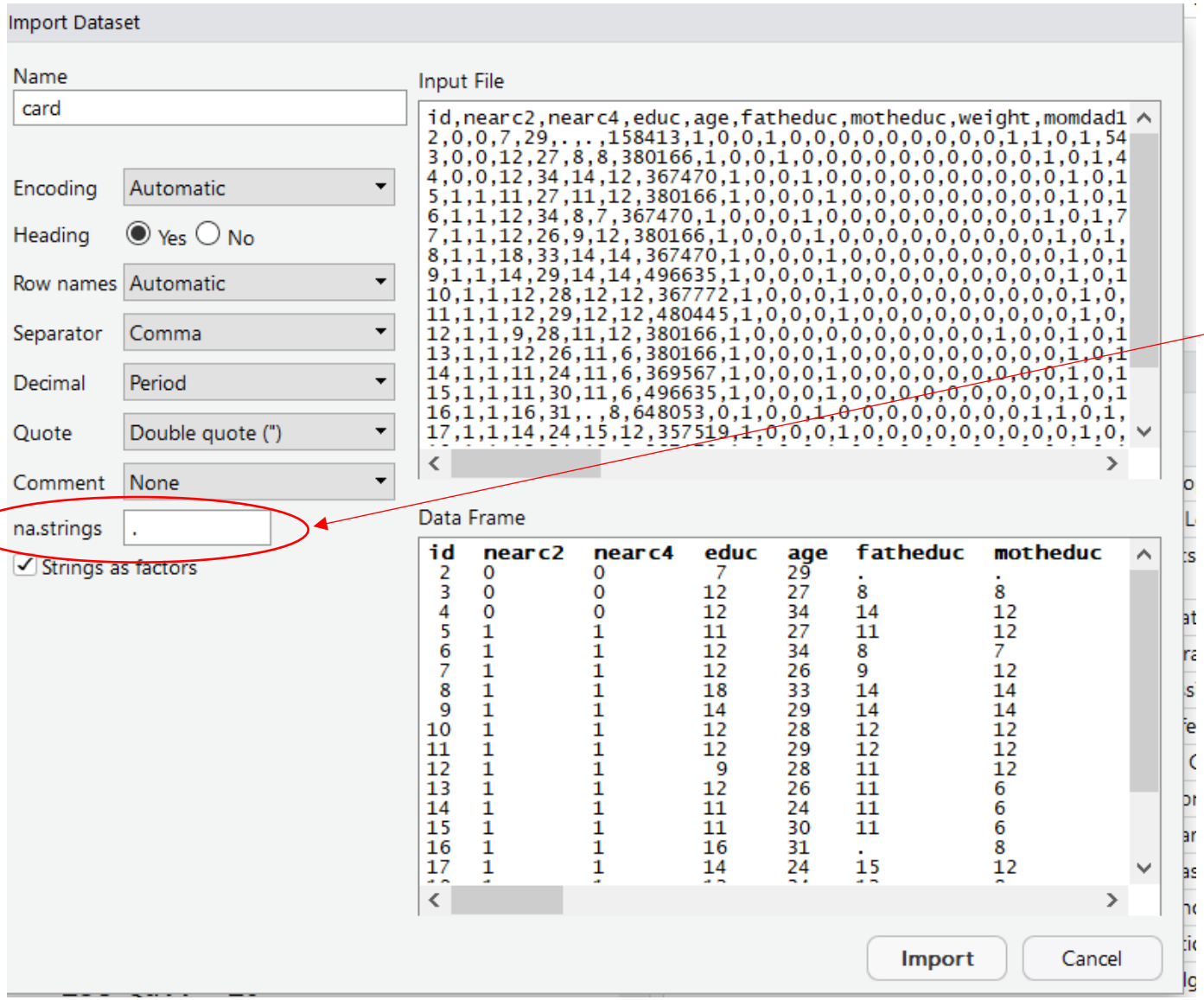
- 女性労働の決定要因の分析
- inlf :女性が外で働いていれば1，そうでなければ0
- 次のようなモデルを考える
- 被説明変数：inlf
- 説明変数：家計所得，教育年数，年齢，子育て費用
- 欠損値あり(.) importの際にna.strings=“.”の指定が必要

- 推定方法

  以下のようなモデルがある

  - 線型確率モデル　linear probability model
  - プロビットモデル　probit model
  - ロジットモデ ル　logit model

注意：Rでの欠損値の扱い



データセットの中に欠損値が含まれている場合(mroz.xlsに欠損値あり)

データのimportの画面で,
n.a.stringsの欄に欠損値の数値（文字列）を指定する
左図は欠損値が”.”の場合

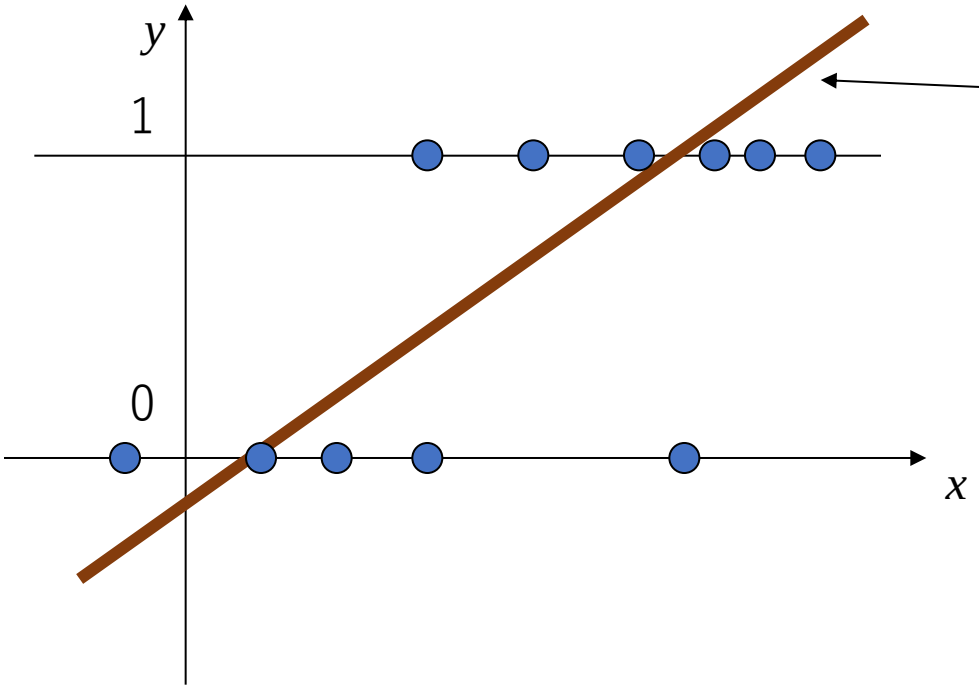
欠損値としてよく使われるのは
　-999
のようなありえない数値

線型確率モデル Linear Probability Model

被説明変数 y は1または0の値をとるダミー変数

線型モデルをそのまま当てはめる

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + u_i$$



当てはめられた直線

yの予測値（当てはめられた直線）は，説明変数が特定の値をとるときに，yが1となる確率を表すと解釈することができる

係数bはxの増加がyが1となる確率をどのくらい増加させるかを表す

線型確率(LP)モデルの問題点

LPモデル　$\Pr(y=1|x)=\beta' x$

- yの予測値が0と1の間に収まらない

  もっと良い定式化　→ $\Pr(y=1|x)=F(\beta' x)$

  $F(\ )$ に確率分布関数を当てはめるとこの問題は回避できる

- 分散不均一性の問題

y=1をとる確率 $=\beta' x$

y=0をとる確率 $=1-\beta' x$

$$\mathrm{Var}(u|x)=F(\beta' x)[1-\beta' x]^2+[1-F(\beta' x)][0-\beta' x]^2$$
$$=\beta' x[1-\beta' x]$$

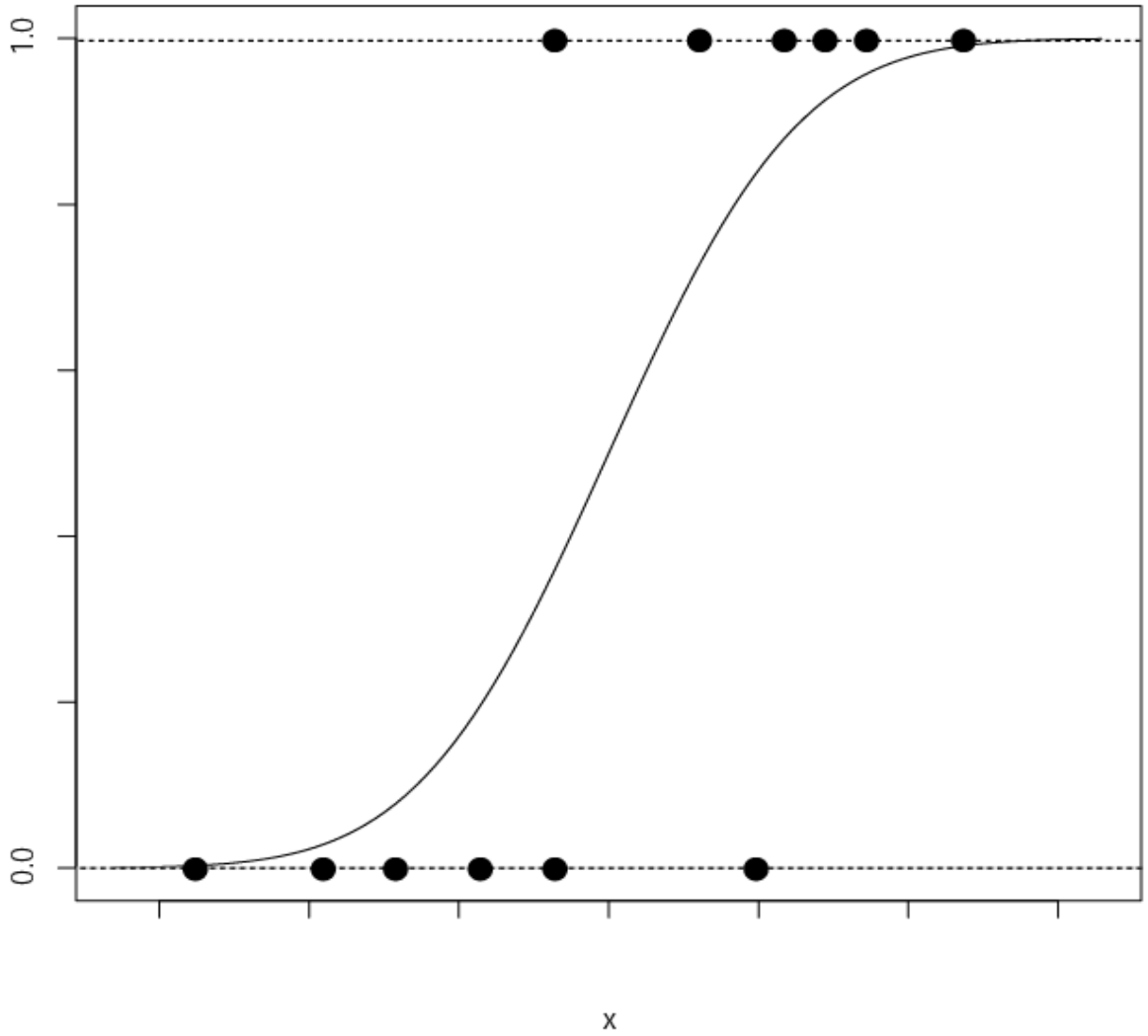
誤差項の分散がxの関数（均一でない）→分散不均一性

最後の等式は，LPモデルにおいて$F(\beta' x)=\beta' x$が成立することを用いた

「yの実現値（１または0）マイナス期待値」の平方

分布関数の当てはめ

xが与えられた場合にy=1となる確率の予測値を0と1の間に収めるためには，分布関数を用いればよい



probit model, logit model

- probit model も logit model も次のようなモデルを想定する（$F(\ )$は確率分布関数）

$$\Pr(y = 1|x) = F(\beta' x) = F(\hat{y})$$

ただし，$\hat{y} \equiv \beta' x = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$

説明変数$x1,\ldots,x\mathrm{k}$の一次関数$\beta' x$が$\mathrm{y}=1$となる確率を決定するという定式化

- Probit model → 標準正規分布

$$F(\hat{y}) = \Phi(\hat{y})$$

$\Phi(\ )$ は標準正規分布の分布関数を表す

- logit model → logistic分布

$$F(\hat{y}) = \frac{\exp(\hat{y})}{1 + \exp(\hat{y})}$$

probit model, logit modelの考え方

$$y_i^* = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + u_i$$

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

yは質的反応を表す変数で，観察不可能な変数y*によって規定されている

　例）女性の労働参加を決定するある観察不可能な変数がある（y*）。

観察不可能な変数y*は，説明変数xの線型関数+誤差項で決定される。

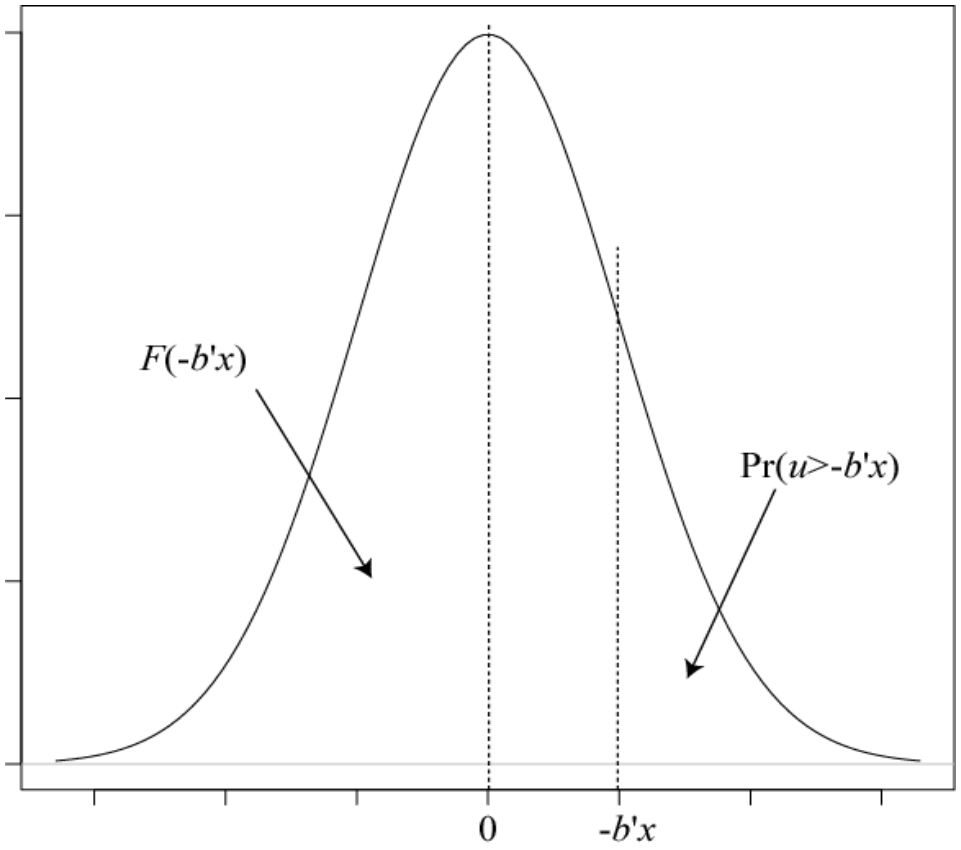
y*がある閾値(critical value)を超えると女性は労働に参加する(y=1)。

しかし，y*が閾値を越えなければ女性は労働に参加しない(y=0)。

probit model, logit modelの考え方(続き)

$$\begin{aligned}\Pr(y=1) &= \Pr(y^* > 0) \\ &= \Pr(\beta' x + u > 0) \\ &= \Pr(u > -\beta' x) \\ &= 1 - F(-\beta' x) \\ &= F(\beta' x)\end{aligned}$$

- 最後の等式は，確率密度関数がx=0に関して対称的な場合に成立 (右の図を参照)
- 標準正規分布, ロジスティック分布の密度関数はx=0に関し対称



probit and logit model : R

- glm( )という関数を用いる

probit :

```
glm(formula,family=binomial(link="probit"))
```

logit :

```
glm(formula,family=binomial(link="logit"))
```

- object <- glm( ) で結果を保存し，summary(object) で結果の要約を出力

例)

```
inlf_probit <- glm(inlf ~ nwifeinc + educ +
exper + age + kidslt6 + kidsge6 ,
family=binomial(link="probit"))
summary(inlf_probit)
```

probit and logit model : R (2)

- 対数尤度　`logLik( object 名)`
- deviance　`-2lnL`
- MacFdden pseudo R2
  `1- object$deviance/object$null.deviance` で求める

- null.deviance 係数=0という制約を課した場合のdeviance

- MacFdden pseudo $R^2 = 1 - L/L_0$
- L: 対数尤度，L_0：定数項のみで当てはめた場合の対数尤度
- OLSのR2にあたる尺度（当てはまりの良さを表す）
  （詳細は計量経済学の教科書を参照してください）

probit model： Rの結果（mroz.xls)

Call:

```
glm(formula = inlf ~ nwifeinc + educ + exper + age + kidslt6 +
kidsge6, family = binomial(link = "probit"))


Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.579574   0.495537   1.170   0.2422
nwifeinc    -0.011565   0.004858  -2.380   0.0173 *
educ         0.133690   0.025254   5.294 1.20e-07 ***
exper        0.070217   0.007693   9.127  < 2e-16 ***
age         -0.055555   0.008305  -6.689 2.24e-11 ***
kidslt6     -0.874290   0.117359  -7.450 9.35e-14 ***
kidsge6      0.034546   0.043376   0.796   0.4258
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  812.44  on 746  degrees of freedom
AIC: 826.44
```

logit model: Rの結果 (mroz.xls)

```
glm(formula = inlf ~ nwifeinc + educ + exper + age +
kidslt6 + kidsge6, family = binomial(link = "logit"))
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.837909   0.840933   0.996   0.3191
nwifeinc    -0.020216   0.008264  -2.446   0.0144 *
educ         0.226977   0.043295   5.243 1.58e-07 ***
exper        0.119746   0.013626   8.788  < 2e-16 ***
age         -0.091088   0.014321  -6.361 2.01e-10 ***
kidslt6     -1.439393   0.201498  -7.143 9.10e-13 ***
kidsge6      0.058174   0.073380   0.793   0.4279
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  812.29  on 746  degrees of freedom
AIC: 826.29
```

係数の比較

	ols		probit		logit	
	coef	s.e.	coef	s.e.	coef	s.e.
Const.	0.707	0.150	0.580	0.496	0.838	0.841
NWIFEINC	-0.003	0.001	-0.012	0.005	-0.020	0.008
EDUC	0.040	0.007	0.134	0.025	0.227	0.043
EXPER	0.023	0.002	0.070	0.008	0.120	0.014
AGE	-0.018	0.002	-0.056	0.008	-0.091	0.014
KIDSLT6	-0.272	0.034	-0.874	0.118	-1.439	0.201
KIDSGE6	0.013	0.013	0.035	0.043	0.058	0.073

probit, logit model の推定方法
最尤法による推定

尤度関数 likelihood function

$$L = \prod_{y_i=0} \left(1 - F(b'x_i)\right) \prod_{y_i=1} F(b'x_i)$$

$$= \prod_{i=1}^{n} [F(b'x_i)]^{y_i} [1 - F(b'x_i)]^{1-y_i}$$

対数変換して

$$\ln L = \sum_{i=1}^{n} \left[y_i \ln F\left(b'x_i\right) + (1 - y_i) \ln\left(1 - F(b'x_i)\right) \right]$$

最尤法: 対数尤度を最大にするようにパラメータを決定

（最尤法：回帰分析の推計方法のもう一つの手法で，実現した(x,y)の組を生成する最もありそうなパラメータは何かを考えて係数を推計する方法。非線形モデルの推計に多く用いられます）

係数の意味: marginal effects

- probit modelやlogit modelの係数の意味

まず，説明変数x_jが1単位増加した場合のyの期待値(y=1となる確率）の変化 -- x_jのmarginal effect -- を求めると次のようになる。

$$\frac{\partial}{\partial x_j}\mathrm{E}[y|x] = \frac{\partial F(\hat{y})}{\partial x_j} = \frac{\partial F(\hat{y})}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial x_j} = f(\hat{y})\beta_j$$

$F(\ )$ ：分布関数, $f(\ )$：確率密度関数

probit model の場合：$f(y) = \phi(y)$

$\phi(\ )$ は標準正規分布の密度関数

logit modelの場合： $f(y) = F(y)[1 - F(y)]$

$F(y) = \frac{\exp(y)}{1+\exp(y)}$ より導かれる

* marginal effectsはxの水準に依存→係数の解釈はやや複雑

marginal effects $f(\hat{y})b_j$ の求め方

marginal effectsはオブザベーション毎に異なるので，全てのケースについて報告するのではなく，代表的なケースを報告する→通常，次の結果を報告する

1. 説明変数xの平均値$(\bar{x})$で評価した値

$$f(b'\bar{x})b_j$$

2. 密度関数$f(b'x)$の平均値を求めて評価した値

$$\left[\frac{1}{n}\sum_{i=1}^{n} f(b'x_i)\right] b_j$$

ここで，x_iはi番目のオブザベーションのx

3. 簡便な方法(相対的な大きさのみを報告)

$$\frac{\partial E[y|x]/\partial x_i}{\partial E[y|x]/\partial x_j} = \frac{b_i}{b_j}$$

- 注意）説明変数にダミー変数が含まれている場合，平均値から1単位増えるというのは変な想定

R： marginal effects

例）　probitモデルの結果をinlf.prに保存し，線形部分のインデックスの予測値をxbに保存する（各オブザベーションごと）。xbの平均値を求め，その点での密度関数（標準正規分布の密度関数はdnorm( )）を求め，回帰係数をかけてmarginal effectを求める。

```
inlf.pr <- glm(inlf ~ nwifeinc + educ + exper + age +
kidslt6 + kidsge6,  family=binomial(link="probit"))
xb <- predict(inlf.pr)
f <-  dnorm(mean(xb))     #fは説明変数の平均値で評価したf(y)
mfx1 <-  f * coef(inlf.pr) #coef( )で係数を取り出す
-------
f2 <- mean(dnorm(xb))     #f2はf(y)の平均値
mfx2 <- f2 * coef(inlf.pr)
```

とすれば，密度関数の平均値で評価したmarginal effectが求まる。

logit モデルの場合は dlogis( ) 関数または f(y)=exp(y)/[1+exp(y)]^2　を用いる

Marginal effects

probit1, logit1は　$f(b'\bar{x})b_j$

probit2, logit2は$\left[\frac{1}{n}\sum_{i=1}^{n} f(b'x_i)\right] b_j$で計算

	probit1	probit2	logit1	logit2
(Intercept)	0.2261	0.1770	0.2030	0.1518
nwifeinc	-0.0045	-0.0035	-0.0049	-0.0037
educ	0.0522	0.0408	0.0550	0.0411
exper	0.0274	0.0214	0.0290	0.0217
age	-0.0217	-0.0170	-0.0221	-0.0165
kidslt6	-0.3411	-0.2670	-0.3488	-0.2608
kidsge6	0.0135	0.0106	0.0141	0.0105

fの値 ：probit1 0.3901 ； probit2 0.3054 ; logit1 0.2423 ; logit2 0.1812
kidslt6, kidsge6の平均値は0.24, 1.35（ここから１増えるという想定は問題あり）

問題 1

- データ：mroz.xls
- 女性の労働参加を，線型確率モデル，logit model, probit model で推計し，係数を解釈せよ。

  被説明変数：inlf　（労働力であれば1）

  説明変数： nwifeinc(non wife income), educ(教育年数），exper(実際に働いた年数），age（年齢），kidslt6（6歳未満の子供の数），kidsge6（6-18歳の子供の数）

Tobit model
censored (truncated) regression

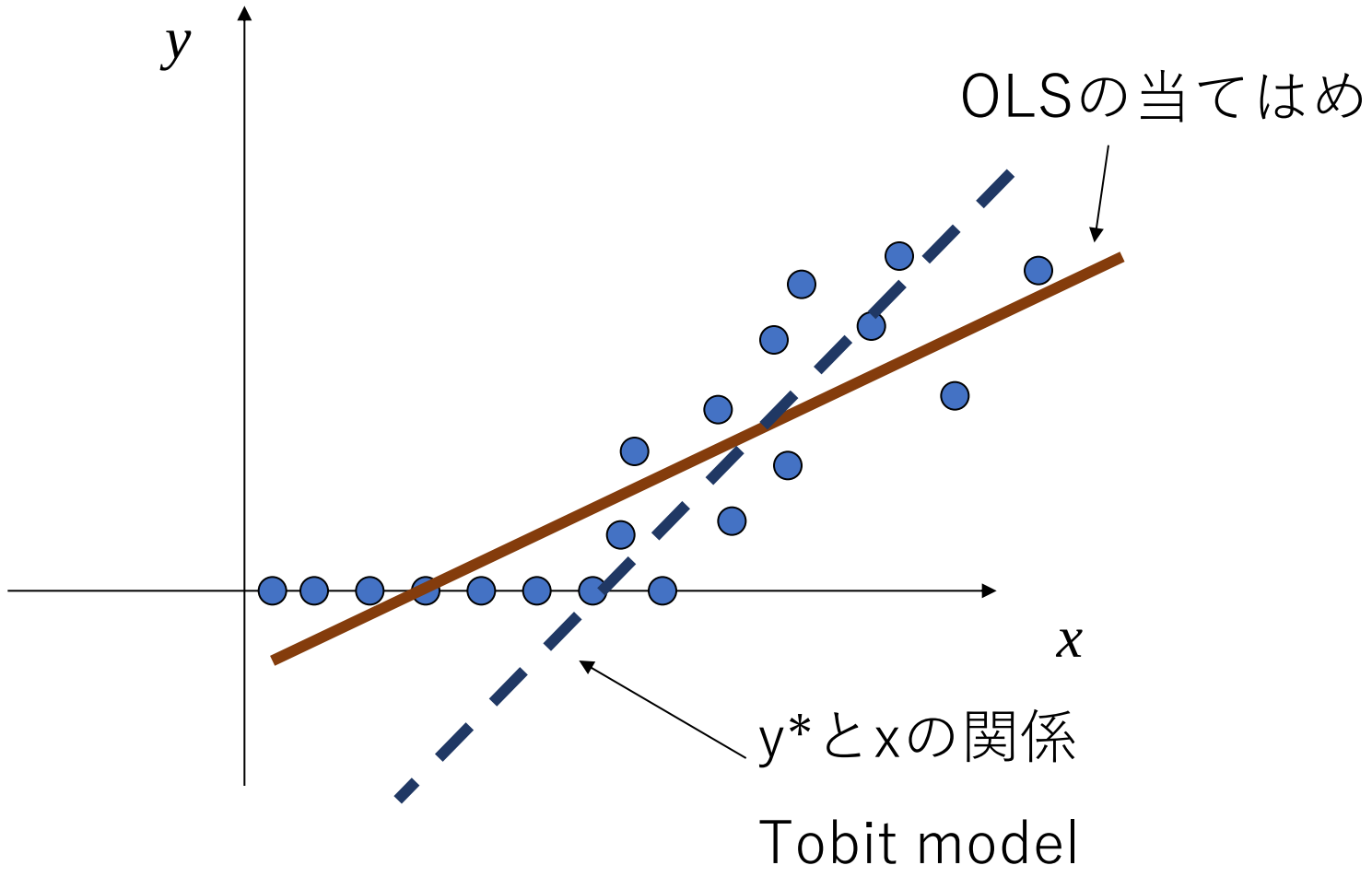
- 例）耐久財の購入量(y)の決定
  - 購入しない人(y=0)が一定数存在
  - y>0 と y=0 のみが観察される
- この場合，次のようなモデルを考える

　y* は観察不可能な変数で，耐久財の購入量(y)を決定する潜在変数とする。y*がある閾値（下の式では0）を超えると，y>0が観察されるが，閾値を超えなければy=0が観察される。

$$y^* = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

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

Tobit model の当てはめ



Tobit modelの応用

- 女性の労働供給
  - 労働参加していない女性が一定割合存在
- 低賃金労働者の労働供給
  - 働かないことを選択する人が一定数存在
- 耐久財の購入
  - 一定期間中に，耐久財購入がゼロの人が一定数存在
  - Tobit model の名前は，James Tobinが耐久財の購入に関する計量経済学的分析でここで説明するようなモデルを採用したことに由来

Tobit model の想定

- 潜在変数y*が次のようなモデルによって決定されると想定

$$y^* = x'\beta + u$$
$$u \sim N(0, \sigma^2)$$

ただし，y*は観測不能で，観測できる変数はy。そして，y*がある閾値（ここでは0と想定）を超えるとy=y*が観測されるが，y*が閾値を超えなければy=0が観測される。

$$y = \max(0, y^*)$$

- この時，y=0 およびy>0である確率は次の式で与えられる

$$\Pr(y = 0) = \Pr(x'\beta + u \leq 0) = \Pr(u \leq -x'\beta)$$
$$= \Pr(u/\sigma \leq -x'\beta/\sigma) = \Phi(-x'\beta/\sigma)$$
$$= 1 - \Phi(x'\beta/\sigma)$$
$$\Pr(y > 0) = \Phi(x'\beta/\sigma)$$

Tobit model の解釈

Tobitモ，デルで推計された係数はy*への影響を表すが，yに与える影響はやや複雑である。まず，xが与えられた場合のy>0という条件付きのyの期待値を求めると次の通りになる（以下は，難しいので読み飛ばして構わない）。

$$\mathrm{E}(y|y>0) = x'\beta + \mathrm{E}(u|u > -x'\beta)$$
$$= x'\beta + \mathrm{E}(u|\, u/\sigma > -x'\beta/\sigma)$$
$$= x'\beta + \sigma\frac{\phi(-x'\beta/\sigma)}{1-\Phi(-x'\beta/\sigma)}$$
$$= x'\beta + \sigma\frac{\phi(x'\beta/\sigma)}{\Phi(x'\beta/\sigma)}$$
$$= x'\beta + \sigma\lambda(x'\beta/\sigma)$$

inverse Mills ratio

ここで$\Phi(\,),\phi(\,)$は，標準正規分布の分布関数と密度関数である。また，上式の導出には，標準正規分布に従う確率変数zについて次の式が成立することを用いている（導出はやや難しい）。

$$\mathrm{E}(z|z>c) = \frac{\phi(c)}{1-\Phi(c)}$$

y>0の条件付きの期待値は，x'bよりも大きくなることが重要

Tobitモデルの解釈(2)

（このページも難しいので読み飛ばして結構）

前ページの結果からyの期待値を求めると次の通りになる。

$$\mathrm{E}(y) = \Pr(y > 0) \cdot E(y|y > 0) + \Pr(y = 0) \cdot 0$$
$$= \Phi(x'\beta/\sigma) \cdot x'\beta + \sigma\phi(x'\beta/\sigma)$$

また，説明変数が1単位増加した場合の効果は次の通りになる。

$$\frac{\partial}{\partial x_j}\Pr(y > 0) = \left(\beta_j/\sigma\right)\varphi(x'\beta/\sigma)$$

$$\frac{\partial}{\partial x_j}\mathrm{E}(y) = \Phi(x'\beta/\sigma) \cdot \beta_j$$

なお，2番目の式の導出には$\phi'(z) = -z\phi(z)$を用いる

Tobit : R

tobit(y~ x1 + x2 + x3)

- AER パッケージが必要
- デフォルトでは左側0, 右側は無限大でセンサー(censor)される指定, 左側，右側を指定することもできる
- Tobitモデルは，もともとはy=0のところで切断されるモデルでしたが，現在では，それを一般化して，被説明変数yがある水準以下または以上で観察できないように拡張されています。ある水準は0でなくても構いません。また，ある水準以上のyが観察されない場合，右側がセンサーされているあるいは切断されている(truncated)といいます。そして，Tobit modelではなく，censored regression または truncatede regressionと呼ばれる場合もあります。
- Rでの例

```
tobit(y~ x1 + x2 + x3,left = 0, right = Inf, dist = "gaussian")
```

Infは無限大（infinity）, gaussian は誤差項が正規分布という指定

Tobit: Rの出力画面

データ: mroz.xls
被説明変数: hours

Coefficients:

	Estimate	Std. Error	z value	P r(>|z|)
(Intercept)	965.30530	446.43614	2.162	0.030599 *
nwifeinc	-8.81424	4.45910	-1.977	0.048077 *
educ	80.64561	21.58324	3.736	0.000187 ***
exper	131.56430	17.27939	7.614	2.66e-14 ***
expersq	-1.86416	0.53766	-3.467	0.000526 ***
age	-54.40501	7.41850	-7.334	2.24e-13 ***
kidslt6	-894.02174	111.87804	-7.991	1.34e-15 ***
kidsge6	-16.21800	38.64139	-0.420	0.674701
Log(scale)	7.02289	0.03706	189.514	< 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

σの推計値

Scale: 1122

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -3819 on 9 Df
Wald-statistic: 253.9 on 7 Df, p-value: < 2.22e-16

Tobit model とOLSの比較

Dependent var hours

	Tobit		OLS	
	Coef	s.e.	Coef	s.e.
C	965.31	446.44	1330.48	270.78
NWIFEINC	-8.81	4.46	-3.45	2.54
EDUC	80.65	21.58	28.76	12.95
EXPER	131.56	17.28	65.67	9.96
EXPERSQ	-1.86	0.54	-0.70	0.32
AGE	-54.41	7.42	-30.51	4.36
KIDSLT6	-894.02	111.88	-442.09	58.85
KIDSGE6	-16.22	38.64	-32.78	23.18
σ	1122.02		750.179	

Tobitの場合，xj の1単位の増加がy*ではなくyに与える影響をみるためには，x'b/σ を計算し，その累積分布を計算する必要あり。

問題2

- データ：mroz.xls
- 女性の労働時間の回帰分析
- 次の方程式をOLSとTobit model で推計し，結果を解釈せよ(Tobit model の場合，y*に与える影響だけでよい)
  - 被説明変数：hours
    - 40%強が労働時間0
  - 説明変数：nwifeinc, educ, exper, expersq, age, kidslt6, kidsge6