# 第 17 章: 制限従属変数モデルと標本選択修正

Jeffrey Wooldridge (2018).
Introductory Econometrics: A Modern Approach
Seventh Edition. Cengage Learning.

2026-03-14

# 準備

# 必要なパッケージの読み込み

- wooldridge: データセット

```
library(wooldridge)
```

- AER: Tobit モデル推定 tobit()

```
install.packages("AER")
library(AER)
```

- survival: 打ち切り回帰 survreg()

```
install.packages("survival")
library(survival)
```

# 17-1 二値応答のロジットとプロビット・モデル

# 線形確率モデル (LPM) の限界

- 第 7 章で学んだ LPM: 二値従属変数への重回帰の適用
- **問題点**:
  1. 推定確率が 0 未満・1 超になりうる
  2. 説明変数の偏効果が一定（非線形性を無視）
- 解決策: **二値応答モデル (binary response models)**

応答確率 [17.1, 17.2]:

$$P(y = 1|\mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) = G(\beta_0 + \mathbf{x})$$

# 17-1a ロジットとプロビット・モデルの定式化

$G$ は $(0,1)$ の値をとる関数:

**ロジット・モデル** [17.3]:

$$G(z) = \frac{\exp(z)}{1+\exp(z)} = \Lambda(z)$$

(標準ロジスティック分布の CDF)

**プロビット・モデル** [17.4]:

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \phi(v)\,dv$$

(標準正規分布の CDF)

# 潜在変数モデルによる導出

**潜在変数モデル** [17.6]:

$$y^* = \beta_0 + \mathbf{x} + e, \quad y = \mathbf{1}[y^* > 0]$$

- $e$ がロジスティック分布 $\Rightarrow$ **ロジット**
- $e$ が標準正規分布 $\Rightarrow$ **プロビット**
- $\beta_j$ の大きさ自体は通常解釈困難（$y^*$ の単位が不明）
- ただし符号と統計的有意性は解釈可能

# 偏効果

連続変数 $x_j$ の応答確率への偏効果 [17.7]:

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = g(\beta_0 + \mathbf{x})\beta_j, \quad g(z) \equiv \frac{dG}{dz}(z)$$

- $g(z) > 0$ なので偏効果の**符号**は $\beta_j$ の符号と同じ
- 効果の**大きさ**は全変数の値に依存

二値変数 $x_1$ の効果 [17.8]:

$$G(\beta_0 + \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k) - G(\beta_0 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

# 17-1b ロジット・プロビットの最尤推定

- LPM は OLS 推定可能だが、ロジット・プロビットは OLS 不可
- **最尤推定法 (MLE)** を使用

対数尤度関数 [17.11]:

$$\ell_i(\beta) = y_i \log[G(\mathbf{x}_i\beta)] + (1 - y_i) \log[1 - G(\mathbf{x}_i\beta)]$$

$\mathcal{L}(\beta) = \sum_{i=1}^{n} \ell_i(\beta)$ を最大化 $\Rightarrow$ MLE は一致・漸近正規・漸近効率

# 17-1c 複数の仮説の検定

**尤度比 (LR) 統計量** [17.12]:

$$LR = 2(\mathcal{L}_{ur} - \mathcal{L}_r) \overset{a}{\sim} \chi^2_q$$

- $q$: 制約数
- $\mathcal{L}_{ur}$: 非制約モデルの対数尤度
- $\mathcal{L}_r$: 制約モデルの対数尤度（対数尤度は常に負なので $LR \geq 0$）

**Wald 統計量**: 線形モデルの $F$ 統計量に相当（計量ソフトが自動計算）

# 17-1d 推定値の解釈: スケール因子

ロジット・プロビット係数の大きさは直接解釈困難 →**スケール因子**で調整

**平均における偏効果 (PEA)** [17.14]:

$$g(\hat{\beta}_0 + \bar{\mathbf{x}}\hat{\beta}) = g\left(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \cdots + \hat{\beta}_k\bar{x}_k\right)$$

**平均偏効果 (APE) / 平均限界効果 (AME)** [17.15]:

$$n^{-1}\sum_{i=1}^{n} g(\hat{\beta}_0 + \mathbf{x}_i\hat{\beta})$$

- APE は離散変数の平均の問題を回避するためより推奨

# 適合度指標

**正確な予測割合 (percent correctly predicted)**:

- $\hat{y}_i = 1$ if $\hat{p}_i \geq 0.5$、$\hat{y}_i = 0$ if $\hat{p}_i < 0.5$
- $\hat{y}_i = y_i$ となる割合

**疑似 $R^2$ (McFadden)**:

$$1 - \mathcal{L}_{ur}/\mathcal{L}_o$$

- $\mathcal{L}_o$: 定数項のみのモデルの対数尤度
- $0.2 \sim 0.4$ で「優れた適合」とされる目安

# 17-1d 例 17.1: 既婚女性の労働参加

# 例 17.1: LPM・ロジット・プロビットの比較 (MROZ)

- データ: 既婚女性 753 人、被説明変数 `inlf`（就業参加ダミー）
- 表 17.1 の再現

```
data(mroz)
# LPM (OLS)
res_lpm <- lm(inlf ~ nwifeinc + educ + exper + expersq +
                 age + kidslt6 + kidsge6, data = mroz)
# ロジット
res_logit <- glm(inlf ~ nwifeinc + educ + exper + expersq +
                   age + kidslt6 + kidsge6,
                 data = mroz, family = binomial(link = "logit"))
# プロビット
res_probit <- glm(inlf ~ nwifeinc + educ + exper + expersq +
                    age + kidslt6 + kidsge6,
                  data = mroz, family = binomial(link = "probit"))
```

# 例 17.1: 係数の比較（表 17.1）

```
round(cbind(
  LPM    = coef(res_lpm),
  Logit  = coef(res_logit),
  Probit = coef(res_probit)
), 3)
```

```
##                LPM  Logit Probit
## (Intercept)  0.586  0.425  0.270
## nwifeinc    -0.003 -0.021 -0.012
## educ         0.038  0.221  0.131
## exper        0.039  0.206  0.123
## expersq     -0.001 -0.003 -0.002
## age         -0.016 -0.088 -0.053
## kidslt6     -0.262 -1.443 -0.868
## kidsge6      0.013  0.060  0.036
```

# 例 17.1: 正確な予測割合

```
round(100 * c(
  LPM   = mean((fitted(res_lpm)    >= 0.5) == mroz$inlf),
  Logit = mean((fitted(res_logit)  >= 0.5) == mroz$inlf),
  Probit= mean((fitted(res_probit) >= 0.5) == mroz$inlf)
), 1)
```

```
##    LPM  Logit Probit
##   73.4   73.6   73.4
```

- LPM・ロジット・プロビットで正確な予測割合はほぼ同じ

# 例 17.1: 疑似 $R^2$（McFadden）

```
L0 <- logLik(glm(inlf ~ 1, data = mroz, family = binomial))

round(c(
  Logit_擬似R2  = 1 - logLik(res_logit)  / L0,
  Probit_擬似R2 = 1 - logLik(res_probit) / L0,
  LPM_R2        = summary(res_lpm)$r.squared
), 3)
```

```
##  Logit_擬似R2 Probit_擬似R2        LPM_R2
##         0.220          0.221         0.264
```

- ロジット・プロビットの疑似 $R^2 \approx 0.22$（LPM の $R^2 \approx 0.26$ に近い）
- 3 モデルとも同程度の説明力

# 例 17.1: APE スケール因子の計算

```r
# プロビット: g(z) = ϕ (z)
xb_hat <- predict(res_probit, type = "link")
ape_scale_probit <- mean(dnorm(xb_hat))
# ロジット: g(z) = Λ (z)[1-Λ (z)]
p_hat <- fitted(res_logit)
ape_scale_logit <- mean(p_hat * (1 - p_hat))
# APE スケール因子の比較
round(c(
  Probit = ape_scale_probit,
  Logit  = ape_scale_logit
), 3)
```

```
## Probit  Logit
##  0.301  0.179
```

# 例 17.1: APE の LPM との比較（表 17.2）

```
# 各変数の APE 比較（表 17.2 より educ, exper, kidslt6）
vars_ape <- c("educ", "exper", "kidslt6")
ape_tbl <- cbind(
  LPM    = coef(res_lpm)[vars_ape],
  Logit  = ape_scale_logit  * coef(res_logit)[vars_ape],
  Probit = ape_scale_probit * coef(res_probit)[vars_ape]
)
round(ape_tbl, 4)
```

```
##              LPM   Logit  Probit
## educ      0.0380  0.0395  0.0394
## exper     0.0395  0.0368  0.0371
## kidslt6 -0.2618 -0.2578 -0.2612
```

- APE は LPM 推定値に非常に近い（3 モデルで整合的）

# 17-2 コーナー解応答のための Tobit モデル

# コーナー解応答とは

- **制限従属変数**のもう 1 つの重要な型:
  - 正の確率でゼロをとる（人口の相当割合がゼロを選択）
  - 正の値をとるときは連続的
  - 例: 家計の慈善寄付額、年間就業時間、年金積立額
- **コーナー解 (corner solution)**: 最適化行動の結果としてのゼロ
- LPM やロジット・プロビットとは異なる問題設定

# Tobit モデルの定式化

潜在変数モデル [17.18, 17.19]:

$$y^* = \beta_0 + \mathbf{x} + u, \quad u|\mathbf{x} \sim \text{Normal}(0, \sigma^2)$$

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

$y_i = 0$ の確率 [17.21]:

$$P(y = 0|\mathbf{x}) = 1 - \Phi(\mathbf{x}/\sigma)$$

対数尤度 [17.22]:

$$\ell_i = \mathbf{1}(y_i = 0) \log[1 - \Phi(\mathbf{x}_i\beta/\sigma)] + \mathbf{1}(y_i > 0) \log\left\{ \frac{1}{\sigma}\phi\left[\frac{y_i - \mathbf{x}_i\beta}{\sigma}\right] \right\}$$

# 17-2a Tobit 推定値の解釈

**条件付き期待値** ($y > 0$ のとき) [17.24]:

$$E(y|y > 0, \mathbf{x}) = \mathbf{x} + \sigma\lambda(\mathbf{x}/\sigma)$$

**逆ミルズ比**: $\lambda(c) = \phi(c)/\Phi(c)$ （常に正）

**無条件期待値** [17.25]:

$$E(y|\mathbf{x}) = \Phi(\mathbf{x}/\sigma)\mathbf{x} + \sigma\phi(\mathbf{x}/\sigma)$$

$E(y|\mathbf{x})$ への偏効果 [17.30]:

$$\frac{\partial E(y|\mathbf{x})}{\partial x_j} = \beta_j \, \Phi(\mathbf{x}/\sigma)$$

# Tobit 偏効果と OLS の比較

- OLS 係数 $\tilde{\gamma}_j$ との比較には調整が必要

**PEA スケール因子**: $\Phi(\bar{\mathbf{x}}\hat{\beta}/\hat{\sigma})$ - $\hat{\beta}_j \times \Phi(\bar{\mathbf{x}}\hat{\beta}/\hat{\sigma}) \approx \tilde{\gamma}_j$

**APE スケール因子**: $n^{-1}\sum_{i=1}^{n} \Phi(\mathbf{x}_i\hat{\beta}/\hat{\sigma})$

コーナー解では $\sigma$ も経済的に重要（補助的パラメータではない）

# 例 17.2: 既婚女性の年間就業時間 (MROZ)

- 753 人中 325 人が就業時間ゼロ → Tobit が適切
- 正の就業時間は 12 〜 4,950 時間と広範囲
- 表 17.3 の再現

```
data(mroz)
# OLS（全観測）
res_ols172 <- lm(hours ~ nwifeinc + educ + exper + expersq +
                   age + kidslt6 + kidsge6, data = mroz)
# Tobit MLE（AER::tobit, 左打ち切り at 0）
res_tobit172 <- tobit(hours ~ nwifeinc + educ + exper + expersq +
                        age + kidslt6 + kidsge6,
                      data = mroz, left = 0)
```

# 例 17.2: OLS vs Tobit（表 17.3）

```
# 係数と σ の比較
b_ols   <- c(coef(res_ols172),   sigma = summary(res_ols172)$sigma)
b_tobit <- c(coef(res_tobit172)[1:8], sigma = res_tobit172$scale)
round(cbind(OLS = b_ols, Tobit = b_tobit), 2)
```

```
##                    OLS   Tobit
## (Intercept) 1330.48  965.31
## nwifeinc      -3.45   -8.81
## educ          28.76   80.65
## exper         65.67  131.56
## expersq       -0.70   -1.86
## age          -30.51  -54.41
## kidslt6     -442.09 -894.02
## kidsge6      -32.78  -16.22
## sigma        750.18 1122.02
```

# 例 17.2: APE スケール因子

```
xb_lin    <- predict(res_tobit172)
sigma_hat <- res_tobit172$scale
ape_scale_tobit <- mean(pnorm(xb_lin / sigma_hat))
cat("APE スケール因子 (Tobit):", round(ape_scale_tobit, 3), "\n")
```

```
## APE スケール因子 (Tobit): 0.589
```

# 例 17.2: 変数別 APE 比較（表 17.4）

```
vars172 <- c("nwifeinc","educ","exper","kidslt6","kidsge6","age")
ape_tbl172 <- cbind(
  Linear    = coef(res_ols172)[vars172],
  Tobit_APE = ape_scale_tobit * coef(res_tobit172)[vars172]
)
round(ape_tbl172, 2)
```

```
##           Linear Tobit_APE
## nwifeinc   -3.45     -5.19
## educ       28.76     47.47
## exper      65.67     77.45
## kidslt6  -442.09   -526.28
## kidsge6   -32.78     -9.55
## age       -30.51    -32.03
```

- Tobit APE は OLS 係数より全体的に大きい（特に educ, kidslt6）

# 17-2b Tobit モデルの定式化の問題

- Tobit は正規性・均一分散を仮定 → 違反時に MLE は不一致

**簡易チェック**: 二値変数 $w = \mathbf{1}[y > 0]$ をプロビット推定

- プロビット係数 $\hat{\gamma}_j$ は Tobit 係数 $\hat{\beta}_j/\hat{\sigma}$ に近いはず
- 符号が異なる・大きさが極端に異なる → Tobit 不適

**ハードル（2 部分）モデル**:

- $P(y > 0|\mathbf{x})$ と $E(y|y > 0, \mathbf{x})$ に別々のパラメータを使用
- コーナー解に対してより柔軟（ただし推定は複雑）

# 17-3 ポアソン回帰モデル

# カウント変数とポアソン回帰

- **カウント変数**: 非負の整数値 $\{0, 1, 2, ...\}$ をとる従属変数
  - 例: 年間逮捕回数、年間特許申請数、子どもの数
- ゼロが多い場合、OLS は適切でない
- 条件付き期待値を指数関数でモデル化 [17.31]:

$$E(y|x_1, \dots, x_k) = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

両辺の対数 [17.32]:

$$\log[E(y|\mathbf{x})] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

$x_j$ が 1 単位増加 $\rightarrow$ $E(y|\mathbf{x})$ が約 $100\beta_j\%$ 変化

# ポアソン回帰の推定

**ポアソン分布**:

$$P(y = h \mid \mathbf{x}) = \frac{\exp(-\exp(\mathbf{x}\beta))[\exp(\mathbf{x}\beta)]^h}{h!}$$

**ポアソン対数尤度** [17.33]:

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \{y_i \mathbf{x}_i \beta - \exp(\mathbf{x}_i \beta)\}$$

# ポアソン分布の等分散制約

ポアソン分布の等分散制約 [17.34]:

$$\mathrm{Var}(y|\mathbf{x}) = E(y|\mathbf{x})$$

- 制約が成り立たなくても **QMLE（Quasi Maximum Likelihood Estimator：準最尤推定量）**として $\beta_j$ は一致推定
  - QMLE: 尤度の分布仮定が完全に正しくなくても、最尤法の形で推定する方法
- **過分散** ($\mathrm{Var} > E$) の場合 → 修正標準誤差を使用

# 過分散の対処

分散が平均に比例する仮定 [17.35]:

$$\mathrm{Var}(y|\mathbf{x}) = \sigma^2 E(y|\mathbf{x})$$

$\sigma^2 > 1$: 過分散、$\sigma^2 < 1$: 過少分散

一致推定量 [テキスト p.581]:

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\frac{\hat{u}_i^2}{\hat{y}_i}$$

- $\hat{\sigma} > 1$ なら通常のポアソン標準誤差に $\hat{\sigma}$ を掛けて修正

# 例 17.3: 若い男性の逮捕回数 (CRIME1)

- データ: 2,725 人、被説明変数 narr86（1986 年逮捕回数）
- 1,970 人が逮捕ゼロ（0 が 72.3%）→ ポアソン回帰が適切
- 表 17.5 の再現

```
data(crime1)
# OLS
res_ols173 <- lm(narr86 ~ pcnv + avgsen + tottime + ptime86 +
                   qemp86 + inc86 + black + hispan + born60,
                 data = crime1)
# ポアソン QMLE
res_poisson <- glm(narr86 ~ pcnv + avgsen + tottime + ptime86 +
                     qemp86 + inc86 + black + hispan + born60,
                   data = crime1, family = poisson)
```

# 例 17.3: 過分散パラメータの推定

```
n    <- nrow(crime1)
k1   <- length(coef(res_poisson))
u_hat <- residuals(res_poisson, type = "response")
y_hat <- fitted(res_poisson)
sigma2_hat <- sum(u_hat^2 / y_hat) / (n - k1)
# 過分散パラメータ σ^²
round(sigma2_hat, 3)
```

```
## [1] 1.517
```

```
# QMLE 標準誤差の修正倍率 σ^
round(sqrt(sigma2_hat), 3)
```

```
## [1] 1.232
```

- $\hat{\sigma} = 1.232 > 1 \rightarrow$ 過分散が存在 $\rightarrow$ 通常の MLE 標準誤差を $\hat{\sigma}$ 倍に修正

# 例 17.3: OLS vs ポアソン回帰（表 17.5）

```
# 比較する変数を指定
vars173 <- c("pcnv","black")
```

# 例 17.3: OLS vs ポアソン回帰（表 17.5）

```
# 係数の比較
round(cbind(
  OLS     = coef(res_ols173)[vars173],
  Poisson = coef(res_poisson)[vars173]
), 3)
```

```
##           OLS Poisson
## pcnv   -0.132  -0.402
## black   0.327   0.661
```

## Poisson 回帰の係数の解釈

- 有罪判決率（pcnv）が 10 ポイント上がると（$\Delta pcnv = 0.10$）、逮捕回数は $100[\exp(-0.402 \times 0.10) - 1]$ より約-3.9%減少
- black 係数 0.661: 黒人男性の期待逮捕回数は他条件一定で $100[\exp(0.661) - 1]$ より約 93.6%高い

# 例 17.3: OLS vs ポアソン回帰の解釈比較

| 変数 | OLS の解釈 | Poisson の解釈 |
|---|---|---|
| `pcnv` | 有罪判決率が **10 ポイント上昇**すると、逮捕回数は**約-0.013 回減少** | 有罪判決率が **10 ポイント上昇**すると、期待逮捕回数は**約-3.9%減少** |
| `black` | 黒人男性は、他条件一定で**平均逮捕回数が約 0.327 回多い** | 黒人男性の期待逮捕回数は、他条件一定で **約 93.6%多い** |

# 17-4 打ち切り回帰モデルと切断回帰モデル

# 打ち切りデータと切断データの違い

**打ち切り回帰 (censored regression)**:

- 説明変数は全観測で利用可能
- 被説明変数 $y_i$ が閾値を超えると観測不可
- 例: トップコーディング、存続期間分析
- Tobit の数学的構造と同じだが**解釈が異なる**

**切断回帰 (truncated regression)**:

- $y$ が閾値以上（または以下）の観測のみがサンプルに含まれる
- 説明変数の情報もない
- 例: 一定所得以下の家計のみを対象とした調査

# 17-4a 打ち切り正規回帰モデル

母集団モデル [17.36, 17.37]:

$$y_i = \beta_0 + \mathbf{x}_i\beta + u_i, \quad u_i|\mathbf{x}_i, c_i \sim \text{Normal}(0, \sigma^2)$$

$$w_i = \min(y_i, c_i)$$

($c_i$: 打ち切り閾値; 右打ち切りの場合)

- 打ち切り観測の密度 [17.38]:
  $f(w|\mathbf{x}_i, c_i) = 1 - \Phi[(c_i - \mathbf{x}_i\beta)/\sigma]$ $(w = c_i)$
- 打ち切りなし観測の密度 [17.39]:
  $f(w|\mathbf{x}_i, c_i) = (1/\sigma)\phi[(w - \mathbf{x}_i\beta)/\sigma]$ $(w < c_i)$
- **注意**: Tobit と違い $\beta_j$ は線形回帰と同様に解釈可能

# 打ち切りと Tobit の比較

| | Tobit (コーナー解) | 打ち切り回帰 |
|---|---|---|
| ゼロの意味 | 経済的最適化の結果 | データ収集の問題 |
| $\beta_j$ の解釈 | $E(y^*\|\mathbf{x})$ への偏効果 | $E(y\|\mathbf{x})$ への直接的効果 |
| OLS 適用 | 不一致（バイアスあり） | 不一致（バイアスあり） |
| 代表例 | 年間就業時間 | 上限打ち切りされた賃金 |

# 例 17.4: 再犯の存続期間分析 (RECID)

- データ: 北カロライナ州刑務所出所者 1,445 人
- 被説明変数: $\log(durat)$（再逮捕までの期間の対数）
- 893 人が観察期間内に再逮捕されず（右打ち切り: `cens = 1`）
- 表 17.6 の再現

```
library(survival)
data(recid)
# 打ち切り正規回帰（右打ち切り）
# cens=1: 打ち切り観測 → 事象指標 = 1 - cens
res_cens <- survreg(
  Surv(ldurat, 1 - cens) ~ workprg + priors + tserved +
    felon + alcohol + drugs + black + married + educ + age,
  data = recid, dist = "gaussian")
```

# 例 17.4: 推定結果（表 17.6）前半

```
coef_surv <- coef(res_cens)
se_surv   <- sqrt(diag(vcov(res_cens)))
result174 <- cbind(`係数` = round(coef_surv, 3),
                   SE   = round(se_surv,   3))
print(result174[1:6, ])
```

```
##                係数    SE
## (Intercept)  4.099 0.348
## workprg     -0.063 0.120
## priors      -0.137 0.021
## tserved     -0.019 0.003
## felon        0.444 0.145
## alcohol     -0.635 0.144
```

# 例 17.4: 推定結果（表 17.6）後半

```
print(result174[7:11, ])
```

```
##              係数    SE
## drugs    -0.298 0.133
## black    -0.543 0.117
## married   0.341 0.140
## educ      0.023 0.025
## age       0.004 0.001
```

```
cat("σ^ =", round(res_cens$scale, 3), "\n")
```

```
## σ^ = 1.81
```

# 例 17.4: 結果の解釈

- `priors`（過去の有罪歴）：$-0.137 \to 1$ 回増加ごとに再犯までの期間が約 $14\%$ 短縮
- `tserved`（服役期間）：$-0.019 \to 12$ ヶ月増加で約 $22.8\%$ 短縮（抑止ではなく再犯傾向を反映）
- `felon`（重罪犯）：$0.444 \to$ 非重罪犯より期間が約 $56\%$ $[\exp(0.444) - 1]$ 長い
- `workprg`（刑務所内作業プログラム）：$-0.063$（非有意）
- OLS で打ち切りを無視すると係数が全体的に過小推定

# 17-4b 切断回帰モデル

切断正規回帰モデル [17.40, 17.41]:

$$y = \beta_0 + \mathbf{x} + u, \quad u|\mathbf{x} \sim \text{Normal}(0, \sigma^2)$$

観測は $y_i \leq c_i$ のときのみ（$c_i$: 切断閾値）

条件付き密度:

$$g(y|\mathbf{x}_i, c_i) = \frac{f(y|\mathbf{x}_i, \beta, \sigma^2)}{F(c_i|\mathbf{x}_i, \beta, \sigma^2)}, \quad y \leq c_i$$

- 切断標本に OLS を適用すると $\beta_j$ をゼロ方向にバイアス（図 17.4 参照）
- 正規性・均一分散の違反時は不一致

# 17-5 標本選択修正

# 非無作為標本選択の問題

- **非無作為標本選択**: 観測されるかどうかが $y_i$ や $u_i$ に依存
- **偶発的打ち切り (incidental truncation)**: 別の変数の結果として $y$ が観測されない
  - 例: 賃金提示 $y = \log(wage^o)$ は就業者のみ観測
  - 就業決定と賃金提示は共通の観測不能要因を持つ可能性
  - これまでの賃金例は全てこの問題を抱えている

# 17-5a 選択標本で OLS が一致する条件

母集団モデル [17.42]:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u, \quad E(u|x_1, \ldots, x_k) = 0$$

選択指標 $s_i$: $s_i = 1$ のとき観測、$s_i = 0$ のとき非観測

**外生的標本選択** → OLS 一致: 1. $s$ が説明変数 $\mathbf{x}$ のみの関数 2. $s$ が $(\mathbf{x}, u)$ と独立

**内生的標本選択** → OLS 不一致: - 選択が誤差項 $u$ に直接依存
(例: 切断標本、偶発的打ち切り)

# 17-5b 偶発的打ち切り: モデル設定

人口モデル [17.46, 17.47]:

$$y = \mathbf{x} + u, \quad E(u|\mathbf{x}) = 0$$

$$s = \mathbf{1}[\mathbf{z} + v \geq 0]$$

仮定:

- $\mathbf{x}$ は $\mathbf{z}$ の真部分集合 ($\mathbf{x} \subsetneq \mathbf{z}$): **除外制約**が重要
- $v$ は標準正規分布に従う
- $(u, v)$ は $\mathbf{z}$ から独立、$\mathrm{Corr}(u, v) = \rho$

# 偶発的打ち切りと OLS バイアス

$\rho \neq 0$ のとき、選択標本 $(s = 1)$ での条件付き期待値 [17.48]:

$$E(y|\mathbf{z}, s = 1) = \mathbf{x} + \rho\lambda(\mathbf{z})$$

**逆ミルズ比**: $\lambda(\mathbf{z}) = \phi(\mathbf{z})/\Phi(\mathbf{z})$

- $\rho = 0$ なら $\lambda(\mathbf{z})$ は不要 → OLS で $\beta$ を一致推定
- $\rho \neq 0$ なら $\lambda(\mathbf{z})$ を省略することがバイアスの原因

# Heckit 法の手順

選択方程式のプロビットモデル [17.49]:

$$P(s = 1|\mathbf{z}) = \Phi(\mathbf{z})$$

**2 段階推定（Heckman, 1976）**:

(i) 全 $n$ 観測でプロビット $s_i$ on $\mathbf{z}_i$ を推定
$\Rightarrow \hat{\gamma} \Rightarrow \hat{\lambda}_i = \lambda(\mathbf{z}_i\hat{\gamma})$

(ii) 選択サンプル（$s_i = 1$）で回帰 [17.50]:

$$y_i \text{ on } \mathbf{x}_i, \hat{\lambda}_i$$

- $\hat{\lambda}_i$ の $t$ 統計量: $H_0 : \rho = 0$（選択バイアスなし）の検定
- 除外制約（$\mathbf{x} \subsetneq \mathbf{z}$）がなければ識別が困難

# 例 17.5: 既婚女性の賃金提示方程式 (MROZ)

- 賃金方程式: $\log(wage)$ on `educ`, `exper`, `expersq`（428 人）
- 選択方程式: `inlf` on 賃金変数 + `nwifeinc`, `age`, `kidslt6`, `kidsge6`（753 人）
- **除外変数**: `nwifeinc`, `age`, `kidslt6`, `kidsge6`（賃金提示には影響しないが就業参加には影響すると仮定）
- 表 17.7 の再現

# 例 17.5: Heckit 推定（手動実装）

```r
data(mroz)
# Step 1: 選択方程式のプロビット（全 753 人）
sel_probit <- glm(inlf ~ nwifeinc + educ + exper + expersq +
                    age + kidslt6 + kidsge6,
                  data = mroz, family = binomial(link = "probit"))
# 逆ミルズ比の計算
zg_hat <- predict(sel_probit, type = "link")
imr    <- dnorm(zg_hat) / pnorm(zg_hat)
# Step 2: 選択サンプル（就業者 428 人）での OLS
mroz_work <- subset(mroz, inlf == 1)
imr_work  <- imr[mroz$inlf == 1]
res_heckit <- lm(lwage ~ educ + exper + expersq + imr_work,
                 data = mroz_work)
```

# 例 17.5: OLS vs Heckit（表 17.7）

```
# OLS（選択サンプルのみ）
res_ols175 <- lm(lwage ~ educ + exper + expersq, data = mroz_work)
# 係数比較
ols_vec <- c(coef(res_ols175), lambda = NA)
hec_vec <- coef(res_heckit)
names(hec_vec)[5] <- "lambda"
round(cbind(OLS = ols_vec, Heckit = hec_vec), 4)
```

```
##                    OLS  Heckit
## (Intercept) -0.5220 -0.5781
## educ         0.1075  0.1091
## exper        0.0416  0.0439
## expersq     -0.0008 -0.0009
## lambda           NA  0.0323
```

# 例 17.5: 選択バイアスの検定

```
# λ_hat の係数（H0: rho = 0 の検定）
round(summary(res_heckit)$coefficients["imr_work", ], 4)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##     0.0323     0.1344     0.2401     0.8104
```

- $\hat{\lambda}$ の係数 $\approx 0.032$（se $\approx 0.134$）、$t \approx 0.239$
- $H_0 : \rho = 0$ を棄却できない（選択バイアスの証拠なし）
- educ 係数の差は約 0.1 %ポイント未満（実質的にも差なし）
- **結論**: この賃金方程式では標本選択修正は不要

# まとめ

# 第 17 章のまとめ

| モデル | 状況 | 推定法 | R 関数 |
|---|---|---|---|
| ロジット | 二値応答 | MLE | `glm(..., family=binomial("logit"))` |
| プロビット | 二値応答 | MLE | `glm(..., family=binomial("probit"))` |
| Tobit | コーナー解 | MLE | `AER::tobit(..., left=0)` |
| ポアソン | カウント変数 | QMLE | `glm(..., family=poisson)` |
| 打ち切り回帰 | 打ち切りデータ | MLE | `survival::survreg()` |
| Heckit | 標本選択修正 | 2 段階 | 手動: プロビット + OLS |

# まとめ（続き）

- **ロジット・プロビット**: LPM の欠点を克服; APE で OLS 係数と比較可能
- **Tobit**: コーナー解応答（ゼロが最適行動の結果）に使用
  - OLS 係数より大きい → $\Phi(\bar{\mathbf{x}}\hat{\beta}/\hat{\sigma})$ で調整して OLS と比較
- **ポアソン**: カウント変数; Var $= E$ でなくても QMLE として一致
  - 過分散 ($\hat{\sigma}^2 > 1$) 時は QMLE 修正標準誤差を使用
- **打ち切り回帰**: データ収集上の打ち切り; $\beta_j$ は線形解釈可能
- **Heckit**: 偶発的打ち切りによる選択バイアス修正
  - 除外制約（$\mathbf{x} \subsetneq \mathbf{z}$）が識別に重要
  - $\hat{\lambda}$ の $t$ 統計量で選択バイアスを検定