第 12 回：2 値応答モデルの推定 (1)

北村　友宏

2020 年 12 月 25 日

本日の内容

1. 2 値応答モデル

2. gretl でのプロビット・モデル推定

ダミー変数

- 大きさ n の 2 変量無作為標本 $((y_1, x_1), (y_2, x_2), \cdots, (y_n, x_n))$ を用いて，y を x に回帰することを考える.
- ただし，y_i は 0 または 1 の値をとるダミー変数（dummy variable）．例えば,
  - （個人が）働くなら 1，働かないなら 0.
  - （個人が）チームを移籍するなら 1，しないなら 0.
  - （企業が）市場に参入するなら 1，しないなら 0.

線形確率モデル

被説明変数がダミー変数の場合に線形回帰モデル

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$
$$E(u_i \mid x_i) = 0,$$

を仮定すると，線形確率モデル（linear probability model）となる.

線形確率モデルの問題点

- 被説明変数の値が 1 になる確率を予測すると，0 を下回ったり 1 を上回ったりする．
- 誤差項に不均一分散が発生する．

条件付き期待値と予測値

線形確率モデルの OLS 推定量 $\hat{\beta}_0$ と $\hat{\beta}_1$ を元の式に代入し，誤差項 u_i を除くと，

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

という式で被説明変数 y_i の値を予測できる.

- $\hat{y}_i$ は y_i の予測値.
- $\hat{y}_i$ は「x_i がこの値のときに y_i はどのような値になる傾向があるか」を表す.

$\Rightarrow \hat{y}_i$ は，x_i を所与とした y_i の条件付き期待値

$$E(y_i \mid x_i) = \beta_0 + \beta_1 x_i,$$

を予測したものと解釈できる.

線形確率モデルにおける条件付き期待値

y_i がダミー変数なら，

$$\begin{aligned} E(y_i \mid x_i) &= 0 \cdot P(y_i = 0 \mid x_i) + 1 \cdot P(y_i = 1 \mid x_i) \\ &= P(y_i = 1 \mid x_i). \end{aligned}$$

$$\Downarrow$$

「y_i の条件付き期待値」が「y_i の値が 1 になる条件付き確率」と同じになる．

➥ $\hat{y}_i$ を計算すると，「y_i の値が 1 になる条件付き確率」を予測していることになる．

➥ $\hat{y}_i$, すなわち「y_i の値が 1 になる条件付き確率の予測値」は 0 を下回ったり 1 を上回ったりする（問題）．

線形確率モデルの誤差項の分散

y_i がダミー変数なら，

$$V(u_i \mid x_i) = (\beta_0 + \beta_1 x_i)\left[1 - (\beta_0 + \beta_1 x_i)\right].$$

（証明は省略）

$$\Downarrow$$

誤差項 u_i の分散が説明変数 x_i に応じて変化する．

➥ 不均一分散発生（問題）．

- 仮説検定の際に，不均一分散に対して頑健な標準誤差を用いることである程度対処可能．

$\Downarrow$

これらの問題を解決するには，2 値応答モデル（binary response model）を仮定する．

- 2 値応答モデルは質的選択モデル（qualitative choice model）の 1 つ．

2 値応答モデルは,

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

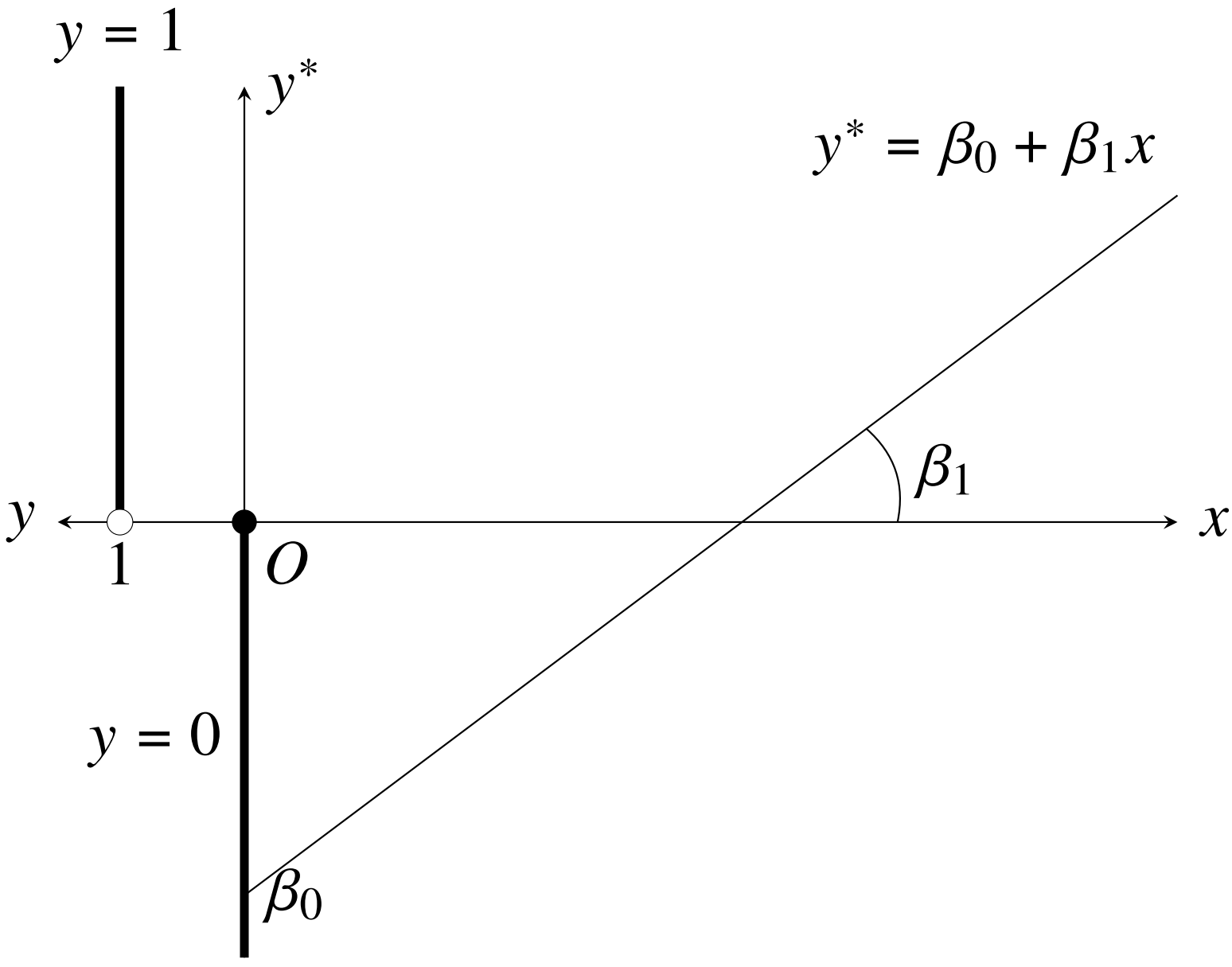
$$y_i^* = \beta_0 + \beta_1 x_i + u_i,$$

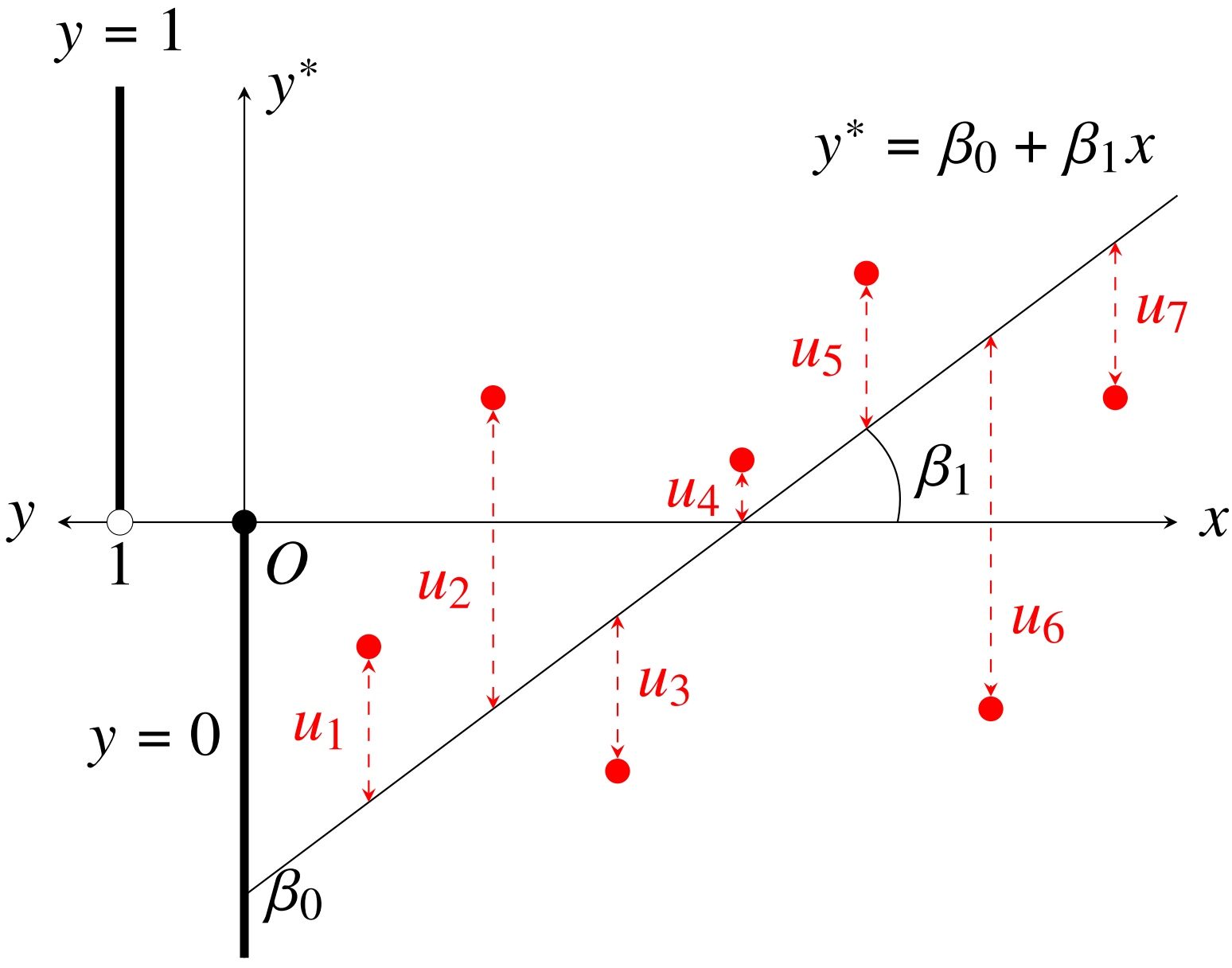
$$u_i \mid x_i \sim F(.).$$

y_i^* は潜在変数（latent variable）．観測不可能な変数で，y_i の値を決定づける．

- 観測可能：y_i, x_i
- 観測不可能：y_i^*, β_0, β_1, u_i
- 推定するもの：β_0, β_1

各変数，パラメータを図示すると？


$y = 1$
y^*
$y^* = \beta_0 + \beta_1 x$
β_1
y
x
1
O
$y = 0$
β_0


$y = 1$
y^*
$y^* = \beta_0 + \beta_1 x$
u_7
u_5
β_1
u_4
y
x
1
O
u_2
u_6
u_3
u_1
$y = 0$
β_0

- 誤差項 u_i には，説明変数 x_i を所与とした条件付き分布を仮定する．
  - e.g., 標準正規分布，ロジスティック分布
  - 誤差項の条件付き分布を標準正規分布と仮定した 2 値応答モデルを 2 値プロビット・モデル（binary probit model）という．
  - 誤差項の条件付き分布をロジスティック分布と仮定した 2 値応答モデルを 2 値ロジット・モデル（binary logit model）という（後の授業で説明）．

注：“$y_i = \cdots$” の式ではなく “$y_i^* = \cdots$” の式の誤差項の分布を仮定している．

2 値プロビット・モデルの定式化

2 値プロビット・モデルは,

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

$$y_i^* = \beta_0 + \beta_1 x_i + u_i,$$

$$u_i \mid x_i \sim N(0, 1).$$

$$\Downarrow$$

最尤（maximum likelihood）法を用いて，β_0 と β_1 を推定する.

2 値プロビット・モデルの推定方法

x_i を所与として，$y_i = 1$ となる条件付き確率は，

$$\begin{aligned} P(y_i = 1 \mid x_i) &= P(y_i^* > 0 \mid x_i) \\ &= P(\beta_0 + \beta_1 x_i + u_i > 0 \mid x_i) \\ &= P\left(u_i > -(\beta_0 + \beta_1 x_i) \mid x_i\right). \end{aligned}$$

標準正規分布は 0 で対称な分布なので，

$$P\left(u_i > -(\beta_0 + \beta_1 x_i) \mid x_i\right) = P\left(u_i < \beta_0 + \beta_1 x_i \mid x_i\right).$$

よって，

$$\begin{aligned} P(y_i = 1 \mid x_i) &= P\left(u_i < \beta_0 + \beta_1 x_i \mid x_i\right) \\ &= \Phi(\beta_0 + \beta_1 x_i). \end{aligned}$$

$\Phi(.)$ は標準正規分布の累積分布関数.

- 前スライドの式では,

$$\Phi(\beta_0 + \beta_1 x_i) = \int_{-\infty}^{\beta_0+\beta_1 x_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz.$$

また，x_i を所与として，$y_i = 0$ となる条件付き確率は,

$$\begin{aligned} P(y_i = 0 \mid x_i) &= 1 - P(y_i = 1 \mid x_i) \\ &= 1 - \Phi(\beta_0 + \beta_1 x_i). \end{aligned}$$

よって，x_i を所与とした y_i の条件付き確率関数は，

$$
\begin{aligned}
f(y_i \mid x_i; \beta_0, \beta_1) &= \begin{cases} \Phi(\beta_0 + \beta_1 x_i) & \text{for } y_i = 1, \\ 1 - \Phi(\beta_0 + \beta_1 x_i) & \text{for } y_i = 0, \\ 0 & \text{elsewhere} \end{cases} \\
&= [\Phi(\beta_0 + \beta_1 x_i)]^{y_i} \, [1 - \Phi(\beta_0 + \beta_1 x_i)]^{1 - y_i} \, .
\end{aligned}
$$

無作為標本なので $y_1, y_2, \cdots, y_n$ は互いに独立.
$x_1, x_2, \cdots, x_n$ を所与とした，$y_1, y_2, \cdots, y_n$ の同時確率関数は,

$$
\begin{aligned}
& f(y_1, y_2, \cdots, y_n \mid x_1, x_2, \cdots, x_n; \beta_0, \beta_1) \\
&= \prod_{i=1}^{n} f(y_i \mid x_1, x_2, \cdots, x_n; \beta_0, \beta_1) \\
&= \prod_{i=1}^{n} f(y_i \mid x_i; \beta_0, \beta_1) \\
&= \prod_{i=1}^{n} [\Phi(\beta_0 + \beta_1 x_i)]^{y_i} [1 - \Phi(\beta_0 + \beta_1 x_i)]^{1-y_i} .
\end{aligned}
$$

尤度関数（likelihood function）は，

$$
\begin{aligned}
&L(\beta_0, \beta_1; y_1, y_2, \cdots, y_n, x_1, x_2, \cdots, x_n) \\
&= \prod_{i=1}^{n} \left[\Phi(\beta_0 + \beta_1 x_i)\right]^{y_i} \left[1 - \Phi(\beta_0 + \beta_1 x_i)\right]^{1-y_i} .
\end{aligned}
$$

対数尤度関数（log-likelihood function）は，

$$
\begin{aligned}
&\ln L(\beta_0, \beta_1; y_1, y_2, \cdots, y_n, x_1, x_2, \cdots, x_n) \\
&= \sum_{i=1}^{n} \Big[y_i \ln \{\Phi(\beta_0 + \beta_1 x_i)\} \\
&\qquad + (1 - y_i) \ln \{1 - \Phi(\beta_0 + \beta_1 x_i)\} \Big] .
\end{aligned}
$$

これが最大になるような β_0 と β_1 を求める．

ML 問題は，

$$\max_{\beta_0,\beta_1} \sum_{i=1}^{n} \Big[y_i \ln \{\Phi(\beta_0 + \beta_1 x_i)\} + (1 - y_i) \ln \{1 - \Phi(\beta_0 + \beta_1 x_i)\} \Big] .$$

(β_0, β_1) の最尤推定量（maximum likelihood estimator, MLE）を $(\hat{\beta}_0, \hat{\beta}_1)$ とする．

1 階条件は,

$$\frac{\partial \ln L}{\partial \beta_0} = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \left[\frac{y_i}{\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)} \cdot \frac{d\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{dz} \right.$$

$$\left. - \frac{1 - y_i}{1 - \Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)} \cdot \frac{d\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{dz} \right] = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \left[\frac{y_i \phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)} - \frac{(1 - y_i)\phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 - \Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)} \right] = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \left[\frac{(y_i - \Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i))\phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)(1 - \Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i))} \right] = 0, \qquad (1)$$

$$\frac{\partial \ln L}{\partial \beta_1} = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \left[\frac{y_i x_i}{\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)} \cdot \frac{d\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{dz} \right.$$

$$\left. - \frac{(1 - y_i) x_i}{1 - \Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)} \cdot \frac{d\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{dz} \right] = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \left[\frac{y_i x_i \phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)} - \frac{(1 - y_i) x_i \phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 - \Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)} \right] = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \left[\frac{(y_i - \Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)) x_i \phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)(1 - \Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i))} \right] = 0. \qquad (2)$$

$\phi(.)$ は標準正規分布の確率密度関数.

- (1) と (2) において,

$$\phi(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\hat{\beta}_0 + \hat{\beta}_1 x_i)^2}{2}\right).$$

(1) と (2) からなる連立方程式は解析的に解けない.

$$\Downarrow$$

コンピューターを用いて数値的に解き，$(\hat{\beta}_0, \hat{\beta}_1)$ を求める.

実習の内容

- データ：日本プロサッカーリーグ（J リーグ）『出場記録』の，2011 年におけるヴィッセル神戸チーム所属選手の年間試合出場時間と総得点，および翌年にチームを移籍したかどうか（選手別データ）

- 分析：「サッカー選手のチーム移籍に影響を与える要因」の分析
  ➥ 推定するモデルの被説明変数がダミー変数（移籍する＝ 1，移籍しない＝ 0）

- 参考：鹿野繁樹（2015）『新しい計量経済学―データで因果関係に迫る』日本評論社.

2 値プロビット・モデルの定式化

いま整理・加工・分析しているデータセットを用いて，以下の 2 値プロビット・モデルを推定する．

$$Transfer_i = \begin{cases} 1 & \text{if } Transfer_i^* > 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$Transfer_i^* = \beta_0 + \beta_1 Timerate_i + \beta_2 Goalrate_i + u_i,$$

$$u_i \mid Timerate_i, Goalrate_i \sim N(0, 1).$$

- $Transfer_i$: 移籍ダミー
  - 翌年（2012 年）に移籍した＝ 1
  - 翌年（2012 年）に移籍しなかった（残留した）＝ 0
- $Timerate_i$: 出場時間率
- $Goalrate_i$: 得点率

「2 値プロビット・モデル」なのに,

$$Transfer_i = \beta_0 + \beta_1 Timerate_i + \beta_2 Goalrate_i + u_i,$$

と書くのは誤り.

- これは線形回帰モデル（被説明変数がダミー変数なので線形確率モデル）の書き方.

出場時間率と得点率の作成

モデルの説明変数として，以下の変数を用意する．

- 出場時間率
  - その選手は年間合計試合時間 3,060 時間のうち，何時間出場したか？
  - 新たな変数を作成し，年間出場時間を表す Time という変数を 3,060 で割ったものと定義すればよい．
- 得点率
  - その選手は年間合計 34 試合のうち，1 試合当たり平均で何点分の得点に貢献したか？
  - 新たな変数を作成し，年間総得点を表す Goal という変数を 34 で割ったものと定義すればよい．

gretl での変数の作成方法

1. gretl のメニューバーから「追加」→「新規変数の定義」と操作.
2. 出てきた「gretl: 変数の追加」ダイアログボックスの入力ボックスに

   (付けたい変数名)=(変数の定義式)

   を入力し,「OK」をクリック.

使える演算子などについては,「gretl: 変数の追加」ダイアログボックスの「ヘルプ」をクリックすれば参照できる（英語）.

実習 1

1. gretl を起動.
2. 「ファイル」→「データを開く」→「ユーザー・ファイル」と操作.
3. jleaguekobe2011.gdt を選択し，「開く」をクリック.

4. 出場時間率の変数を作成する．gretl のメニューバーから「追加」→「新規変数の定義」と操作.
5. 出てきたダイアログボックスの入力ボックスに

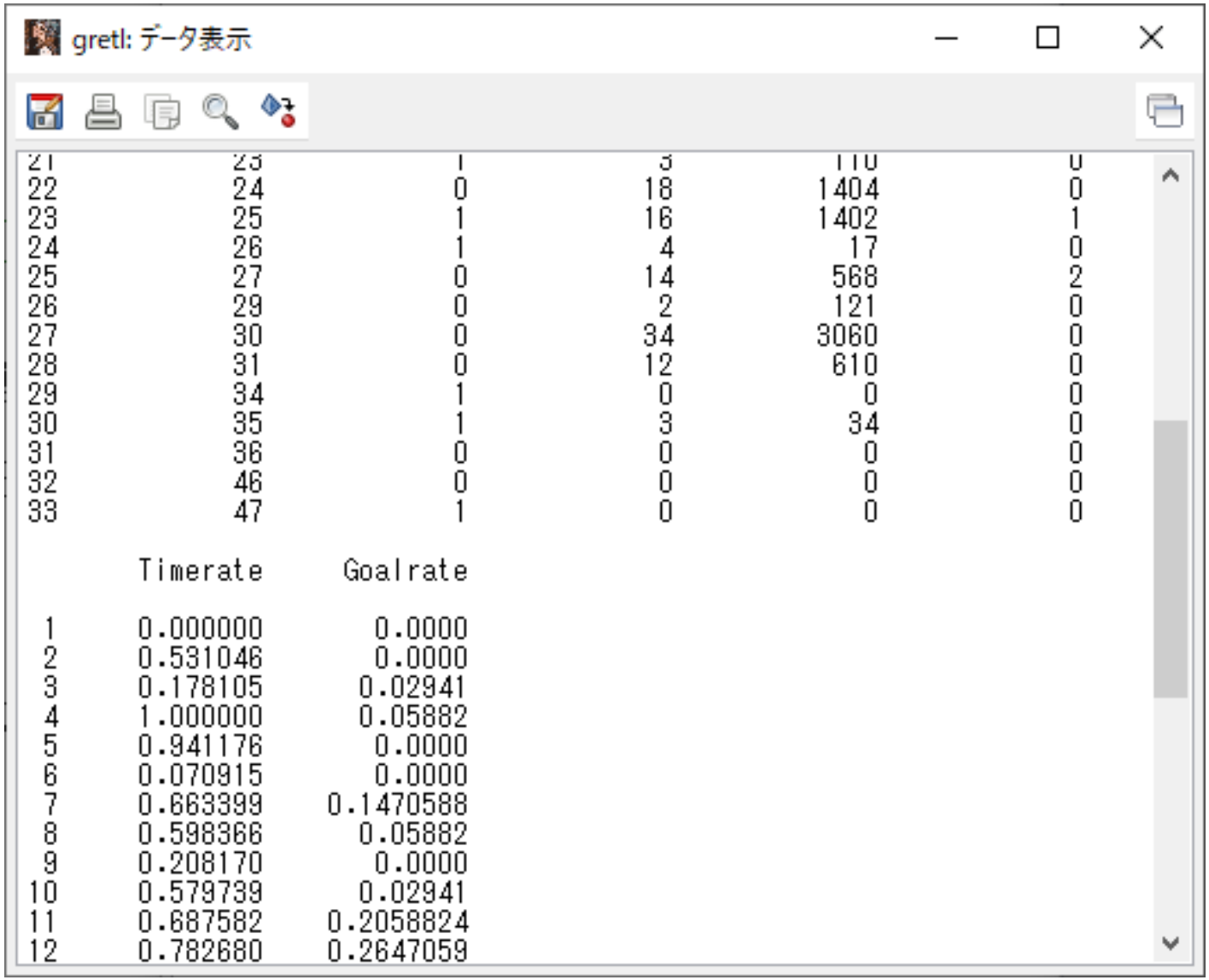
   `Timerate=Time/3060`

   と入力し，「OK」をクリック.
   - 「Timerate」という変数が作成され，「Time を 3,060 で割ったもの」と定義される.
6. 得点率の変数を作成する．gretl のメニューバーから「追加」→「新規変数の定義」と操作.
7. 出てきたダイアログボックスの入力ボックスに

   `Goalrate=Goal/34`

   と入力し，「OK」をクリック.
   - 「Goalrate」という変数が作成され，「Goal を 34 で割ったもの」と定義される.

8. gretl のメニューバーから「ファイル」→「データを保存」と操作し，jleaguekobe2011.gdt を上書き保存.
9. Ctrl キーを押しながら「No」「Transfer」「Attend」「Time」「Goal」「Timerate」「Goalrate」の 7 つを左クリックして選択し，その上で右クリック→「データ（値）を表示」と操作すると，Player を除く 7 変数の観測値リストが新規ウィンドウにて表示される.



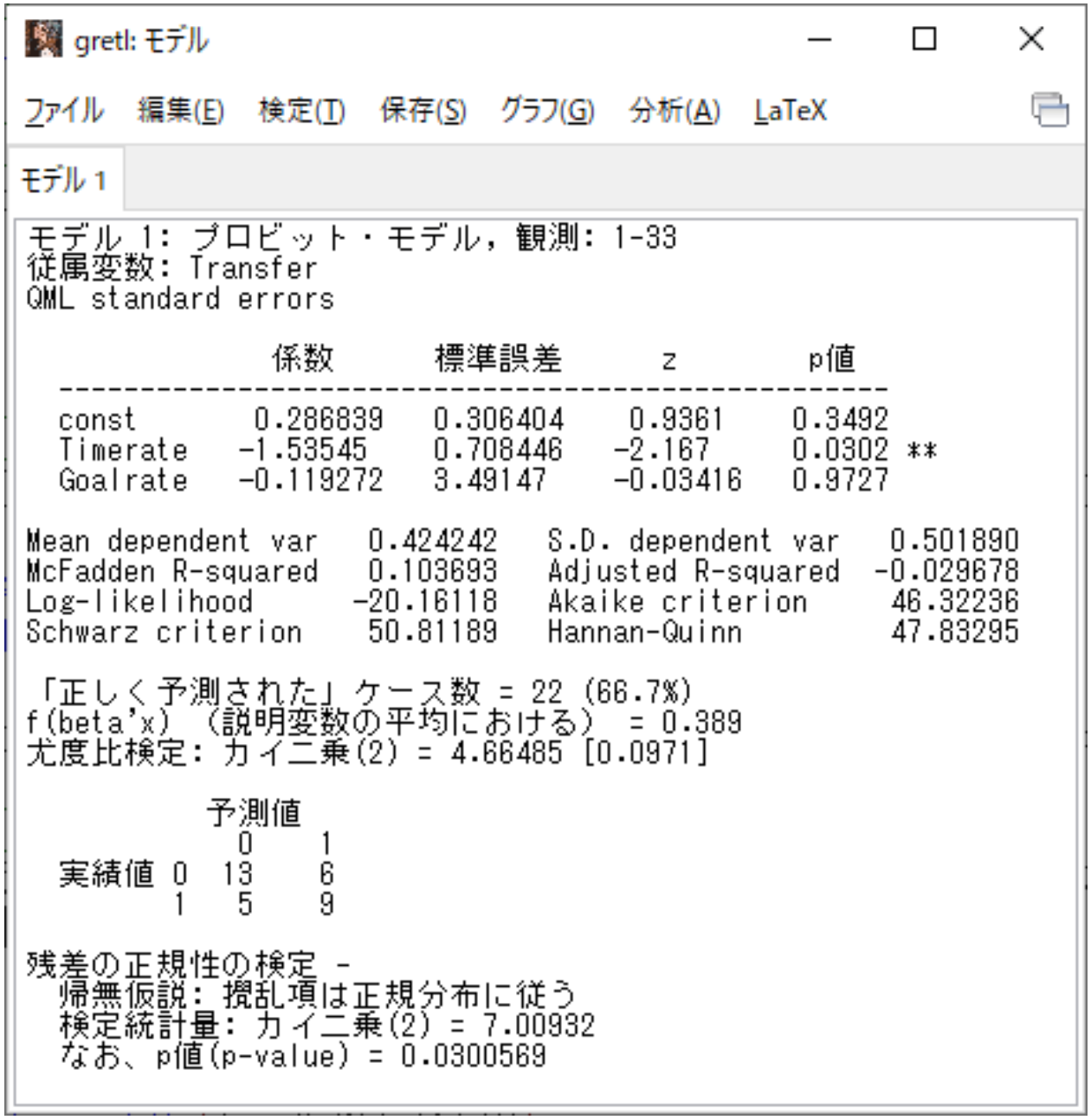
このような画面が表示されれば成功．Timerate と Goalrate の観測値リストは，下のほうに表示されている．確認したら閉じる．

実習 2

「サッカー選手のチーム移籍に影響を与える要因」を分析するための 2 値プロビット・モデルを推定する.

1. gretl のメニューバーから「モデル」→「制限従属変数」→「プロビット」→「二項 (Binary)」と操作.
2. 出てきたウィンドウ左側の変数リストにある Transfer をクリックし，3 つの矢印のうち上の青い右向き矢印をクリック.
   - 推定式の左辺の変数（被説明変数，従属変数）が「『Transfer』が 1 になる確率（移籍する確率）」となる.
3. 独立変数: の右端のボタンをクリックして出場時間率と得点率をクリック.
   - 説明変数が「出場時間率」と「得点率」となる.

4. ウィンドウ左側の変数リストにある Timerate をクリックした後，Ctrl キーを押しながら Goalrate をクリックして，3 つの矢印のうち真ん中の緑の右向き矢印をクリック．
   - 推定式の右辺の変数（説明変数，独立変数）が Timerate（出場時間率）と Goalrate（得点率）となる．
   - 最初から説明変数リストに入っている const は推定式の切片（定数項）のこと．
5. 「頑健標準誤差を使用する」にチェックする．このデータは横断面データのため，不具合は発生しないと考えられる．
   - モデルの定式化に対して頑健な標準誤差が計算される．
6. ラジオボタンの「p 値を表示する」をクリック．
   - 各説明変数の係数がゼロという帰無仮説を検定するための p 値が出力されるようになる．
7. 「OK」をクリックすると，結果が表示される．



このような画面が表示されれば成功.

出力結果の見方

- 係数: （偏）回帰係数推定値
- 標準誤差: （偏）回帰係数の標準誤差
- z: 「（偏）回帰係数が 0」という帰無仮説の両側 z 検定における検定統計量の実現値（z 値）
  - 2 値プロビット・モデルは係数ゼロ仮説の検定統計量の従う確率分布が複雑で，通常は観測値数が十分大きいときに推定されるので，t 検定ではなく正規分布で近似して z 検定を行う.
- p 値: 両側 p 値
- Log-likelihood: 対数尤度

対数尤度

- （説明変数１つの２値プロビット・モデルの場合で説明すると,）対数尤度関数

$$\begin{aligned}
&\ln L(\beta_0, \beta_1; y_1, y_2, \cdots, y_n, x_1, x_2, \cdots, x_n) \\
&= \sum_{i=1}^{n} \Big[y_i \ln \{\Phi(\beta_0 + \beta_1 x_i)\} \\
&\qquad + (1 - y_i) \ln \{1 - \Phi(\beta_0 + \beta_1 x_i)\} \Big],
\end{aligned}$$

に係数推定値と変数の値を代入したものを対数尤度（Log-likelihood）という.

モデル推定結果

- 出場時間率の係数
  - −1.53545
  - 有意水準 5%で，係数ゼロの帰無仮説棄却.
    ➥ 出場時間率はチームを移籍する確率と統計的に有意に相関しており，出場時間率の係数はゼロでないと判断される.

- 得点率の係数
  - −0.119272
  - 有意水準 10%で，係数ゼロの帰無仮説採択.
    ➥ 得点率はチームを移籍する確率と統計的に有意に相関しておらず，得点率の係数はゼロでないとはいえないと判断される.

- 定数項
  - 0.286839
  - 有意水準 10%で，係数ゼロの帰無仮説採択.
    ➥ 定数項はゼロでないとはいえないと判断される.

- 対数尤度
  - −20.16118

係数の解釈

2 値プロビット・モデルなどの 2 値応答モデルの係数は，「被説明変数への影響度合い（説明変数が 1 単位増加すると被説明変数が何単位変化する傾向があるか）」を表さない.

$$\Downarrow$$

- 係数の値そのものに意味はない（解釈できない）.
- 係数の符号の向きと統計的有意性のみ確認できる.
- 被説明変数への定量的な影響度合いを見る方法は，次回の授業で説明する.

実習 3

1. 「gretl: モデル 1」のウィンドウのメニューバーから「ファイル」→「名前を付けて保存」と操作.
2. 「標準テキスト」を選び,「OK」をクリック。
3. プロビットモデル推定結果 1.txt という名前で「2020 ミクロデータ分析 2」フォルダに保存.
すると，表示された推定結果をそのままテキストファイルで保存できる.

本日の作業はここまで.