第 14 回：2 値応答モデルの推定 (3)

北村　友宏

2021 年 1 月 21 日

本日の内容

1. 2 値ロジット・モデルとは
2. gretl でのロジット・モデル推定
3. 2 値プロビット・モデルと 2 値ロジット・モデル

2 値ロジット・モデル

- 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.

2 値ロジット・モデルの定式化

2 値ロジット・モデルは,

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

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

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

$$\Downarrow$$

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

2 値ロジット・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，

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

ロジスティック分布は 0 で対称な分布．よって，

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

したがって，

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

$\Lambda(.)$ はロジスティックス分布の累積分布関数.

- 前スライドの式では,

$$\Lambda(\beta_0 + \beta_1 x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

よって,

$$P(y_i = 1 \mid x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

また，x_i を所与として，$y_i = 0$ となる条件付き確率は，

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

よって，x_i を所与とした y_i の条件付き確率関数は，

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

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,

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

尤度関数（likelihood function）は，

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

対数尤度関数（log-likelihood function）は，

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

これが最大になるような β_0 と β_1 を求める.
ML 問題は,

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

(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,

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

$$\Leftrightarrow \sum_{i=1}^{n} \left[y_i \cdot \left[1 - \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right] \right.$$

$$\left. + (1 - y_i) \cdot \left[-\frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right] \right] = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \left[\frac{y_i}{1 + \exp(\beta_0 + \beta_1 x_i)} - \frac{(1 - y_i)\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right] = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \left[\frac{y_i - (1 - y_i)\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right] = 0, \qquad (1)$$

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

$$\Leftrightarrow \sum_{i=1}^{n} \left[y_i \cdot \left[x_i - \frac{x_i \exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right] \right.$$

$$\left. + (1 - y_i) \cdot \left[- \frac{x_i \exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right] \right] = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \left[\frac{x_i y_i - x_i (1 - y_i) \exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right] = 0. \qquad (2)$$

(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.

2 値ロジット・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 値ロジット・モデルを推定する.

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

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

$$u_i \mid Timerate_i, Goalrate_i \sim \Lambda(.).$$

- $Transfer_i$: 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 値ロジット・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.

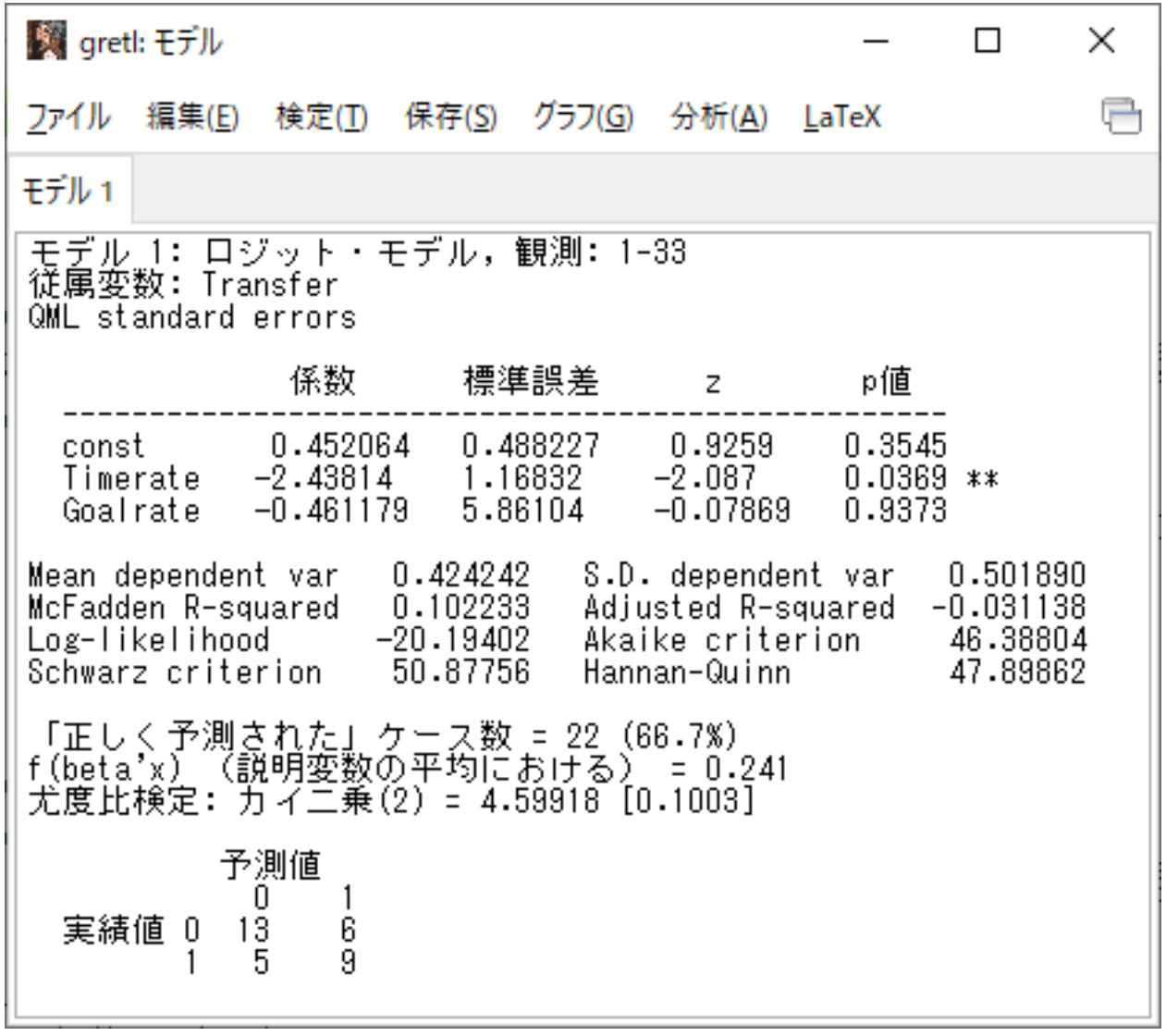
実習 1

「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 値ロジット・モデルを推定する.

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. gretl のメニューバーから「モデル」→「制限従属変数」→「ロジット」→「二項 (Binary)」と操作.
5. 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.
6. 「デフォルトとして設定」にチェック.
   - gretl を終了するまでの間，次回以降モデルの推定を行う際に，いま選択した変数が自動的に被説明変数（従属変数）に入力される.

7. 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．
8. 「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．
9. 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．
10. 「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 値ロジット・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値ロジット・モデルの場合，対数尤度関数は，

$$\ln L(\beta_0, \beta_1; y_1, y_2, \cdots, y_n, x_1, x_2, \cdots, x_n)$$
$$= \sum_{i=1}^{n} \Big[y_i \left[(\beta_0 + \beta_1 x_i) - \ln \{1 + \exp(\beta_0 + \beta_1 x_i)\} \right]$$
$$+ (1 - y_i) \left[- \ln\{1 + \exp(\beta_0 + \beta_1 x_i)\} \right] \Big],$$

なので，それに係数推定値と変数の値を代入したものが対数尤度（Log-likelihood）となる．

モデル推定結果

- 出場時間率の係数
  - −2.43814
  - 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
  - −0.461179
  - 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
  - 0.452064
  - 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
  - −20.19402

実習 2

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. ロジ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.

限界効果

2 値ロジット・モデルにおける，x_i の限界効果（marginal effect）は，

$$\frac{\partial P(y_i = 1 \mid x_i)}{\partial x_i}$$
$$= \frac{\partial}{\partial x_i}\left[\frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}\right]$$
$$= \frac{\{\exp(\beta_0 + \beta_1 x_i)\}\beta_1}{\{1 + \exp(\beta_0 + \beta_1 x_i)\}^2}.$$

$$\Downarrow$$

β_1 そのものではなく $\dfrac{\{\exp(\beta_0 + \beta_1 x_i)\}\beta_1}{\{1 + \exp(\beta_0 + \beta_1 x_i)\}^2}$ が，「x_i が 1 単位増加したときに $y_i = 1$ となる確率がどの程度変化する傾向があるか」を表す．

- $\exp(.)$ は指数関数なので正の値.
  $\Rightarrow \dfrac{\{\exp(\beta_0 + \beta_1 x_i)\}}{\{1 + \exp(\beta_0 + \beta_1 x_i)\}^2}$ も正の値.
  $\Rightarrow$ 限界効果の符号は β_1 の符号と同じ.

- 説明変数 x_i の値は各個体によって異なる.
  ➥ 限界効果

$$\frac{\partial P(y_i = 1 \mid x_i)}{\partial x_i} = \frac{\{\exp(\beta_0 + \beta_1 x_i)\}\beta_1}{\{1 + \exp(\beta_0 + \beta_1 x_i)\}^2},$$

  の値も各個体によって異なる.
  $\Rightarrow$ 説明変数 x_i をその平均 $\bar{x}$ で置き換えた，平均における限界効果（marginal effect at mean, slope at mean）を計算する.

- 2 値ロジット・モデルの，x_i の平均における限界効果（marginal effect at the mean）は，

$$\frac{\{\exp(\hat{\beta}_0 + \hat{\beta}\bar{x})\}\beta_1}{\{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1\bar{x})\}^2}.$$

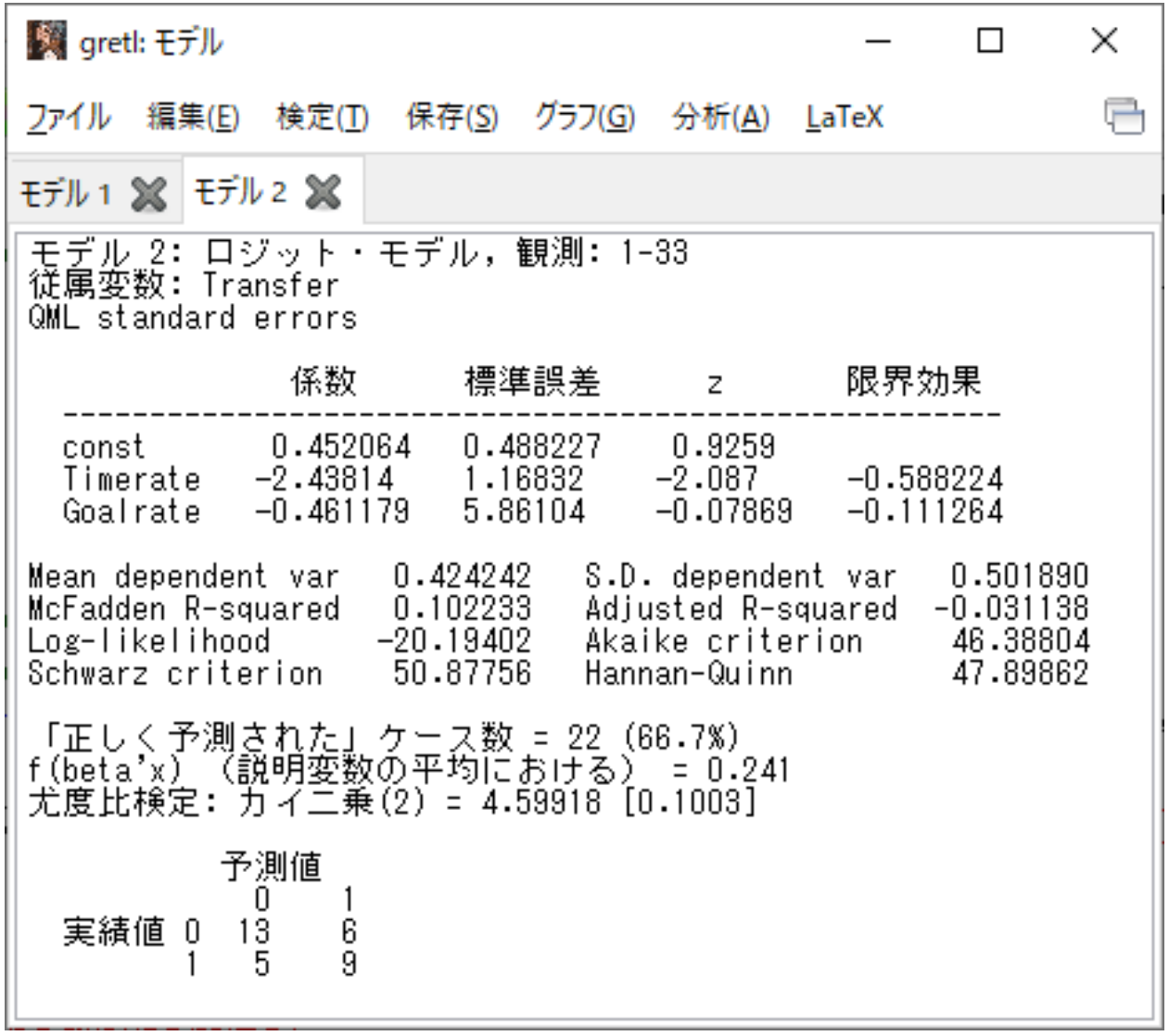
  - $\hat{\beta}_0, \hat{\beta}_1$ はそれぞれ β_0, β_1 の最尤推定値.
  - $\bar{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i.$
  - 定数項以外に説明変数が複数個ある場合は，それらを全てそれぞれの標本平均で置き換える.

実習 3

限界効果を表示させる.

1. gretl のメニューバーから「モデル」→「制限従属変数」→「ロジット」→「二項 (Binary)」と操作．説明変数（回帰変数）は必ず前回の選択内容が記録されており，被説明変数（従属変数）は前回「デフォルトとして設定」にチェックしていれば前回の選択内容が記録されている.
2. 従属変数の入力ボックスに Transfer が入力されていなければ，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. 「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．
4. ラジオボタンの「平均での限界効果（slope at mean）を表示する」をクリック．
   - 各説明変数の，「平均における限界効果」が表示されるようになる．
5. 「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.

限界効果推定結果

- 出場時間率の限界効果
  - −0.588224
    ➥ 出場時間率が 0.01 高くなると（1 パーセントポイント高くなると），チームを移籍する確率が 0.00588224 低くなる（0.588224 パーセントポイント低くなる）.
  - すでに推定した，2 値プロビット・モデルの出場時間率の限界効果（−0.596640）に近い値.

- 得点率の限界効果
  - −0.111264
    ➥ 得点率が 0.01 高くなると（1 パーセントポイント高くなると），チームを移籍する確率が 0.00111264 低くなる（0.111264 パーセントポイント低くなる）.

2 値ロジット・モデルを仮定した分析においても，仮説検定で，出場時間率のみ，係数ゼロの帰無仮説が棄却されたことから，出場機会に恵まれないサッカー選手がチームを移籍する傾向がある．

実習 4

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. ロジットモデル推定結果 2.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.

2 値プロビット・モデルと 2 値ロジット・モデル

- 2 値プロビット・モデルと 2 値ロジット・モデルは，偏回帰係数推定値の大きさにある程度の差が生じるが，係数の統計的有意性や限界効果は両者で似たような結果になる場合が多い.
- 2 値プロビット・モデルと 2 値ロジット・モデルのどちらを採択するかを検定することは不可能.

⇓

レポートや論文では，2 値プロビット・モデルと 2 値ロジット・モデルのどちらか一方のみの推定結果を載せるか，両者の結果を並列して載せて比較するとよい.

レポートや論文に，2 値プロビット・モデルと 2 値ロジット・モデルの推定結果を並列して載せて比較したいときは，例えば以下のような表を載せればよい.

表 1：2 値応答モデル推定結果

	2 値プロビット・モデル				2 値ロジット・モデル			
	偏回帰係数	限界効果	z 値		偏回帰係数	限界効果	z 値	
出場時間率	−1.54	−0.60	−2.17	**	−2.44	−0.59	−2.09	**
得点率	−0.12	−0.05	−0.03		−0.46	−0.11	−0.08	
定数項	0.29		0.94		0.45		0.93	
対数尤度	−20.16				−20.19			

（注 1）表中の**は有意水準 5%で統計的に有意であることを表す.
（注 2）モデルの定式化に対して頑健な標準誤差を用いている.
（注 3）観測値数は 33 である.

本日の作業はここまで.
今回は gretl のデータセットに変更を加えていないので，gretl のデータセット（jleaguekobe2011.gdt）を上書き保存する必要はない.