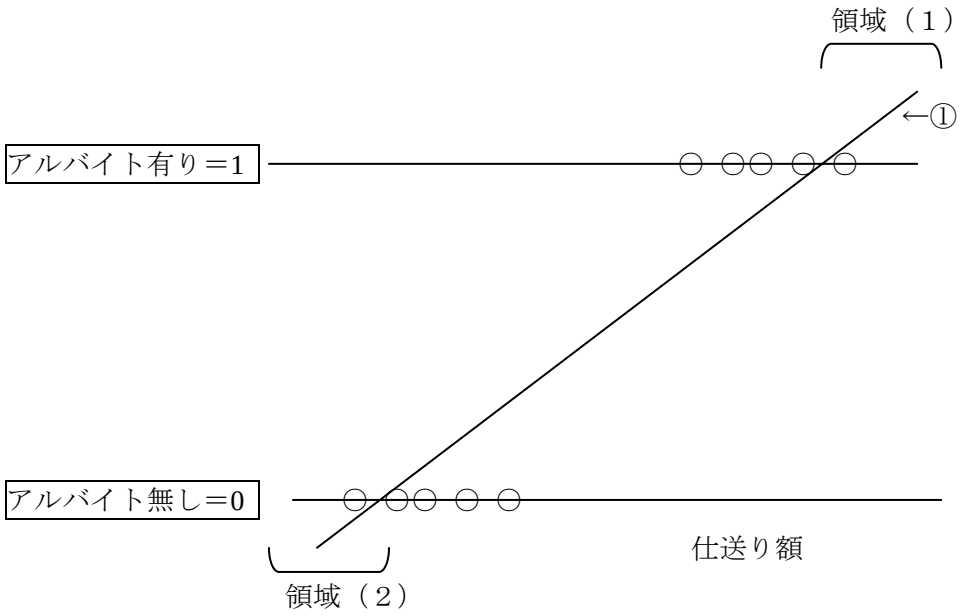
Stataによる離散選択モデル

一橋大学経済研究所

松浦寿幸

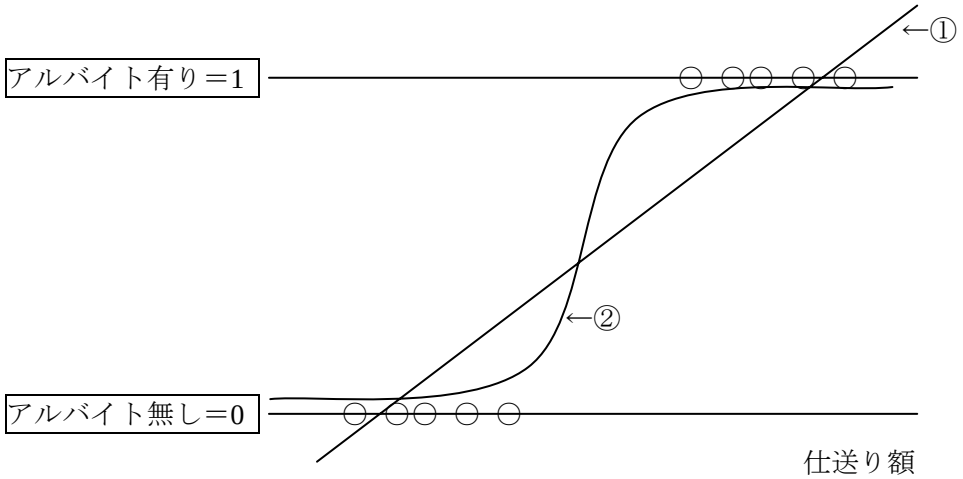
本章では、質的な情報を回帰分析で取り扱う際にダミー変数を説明変数として利用する方法を説明しました。しかし、以下のようにテーマのようにダミー変数を被説明変数として分析する場合、特別な処置を講じる必要が出てきます。

以下のグラフは、*Y*=大学生のアルバイトの有無（アルバイトをしているとき１をとるダミー変数）と *X*=仕送り額の関係を示すグラフです。



この２つの変数の関係を最小二乗法で分析しようとすると、①のような近似直線を当てはめることになります。この場合、領域（１）や領域（２）に示されるように、予測値（理論値）が 0 から 1 の範囲を逸脱する領域が出てきてしまうため、正確な予測ができないという問題が生じます[1]。そこで、非線形の近似曲線をあてはめる方法が採用されます。たとえば、図２の②のような近似曲線をあてはめた場合、予測値（理論値）が 0 から 1 の範囲に収まります。この非線形の近似曲線として、ロジスティック曲線をあてはめたものを**ロジット・モデル**、正規分布の分布関数（累積密度関数）をあてはめたものを**プロビット・モデル**と呼びます。

[1] これに加えて、領域（１）・領域（２）では、*X* とともに誤差が拡大していくという性質により、誤差項が均一に分布でなければならないという最小二乗法の前提条件が満たされないという問題が生じます。



ロジット・モデルやプロビット・モデルでは、潜在変数モデルと呼ばれる選択行動モデルを理論背景として考えます。たとえば、Y^*をアルバイトに対する意欲（潜在変数）とし、親からの仕送り額（X）との関係を以下のような線形関数で表します。

$$Y^*=a+bX+u$$

そして、潜在変数 Y^*が 0 を超えると、Y が 1 になり、Y^*が 0 を下回るとき Y は 0 になるとします。つまり、アルバイトをする（Y=1）ときの条件は、

$$Y^*>0 \quad \Leftrightarrow \quad a+bX+u>0$$
$$\Leftrightarrow \quad u> \ -a \ -bX$$

Y が 1 をとる確率を P(Y=1)とすると、

$$\mathrm{P}(Y=1)=\mathrm{P}(Y^*>0)=\mathrm{P}(u>-a-bX)$$

ある変数が一定の条件を満たす確率は、適当な確率分布を当てはめることにより、その分布関数（累積密度関数）から計算することができます。この確率分布として、ロジスティック分布を適用したものがロジット・モデル、標準正規分布をあてはめたものがプロビット・モデルです。推定については、最尤法と呼ばれる推定法が用いられ、大抵の統計パッケージにはコマンドが用意されています。

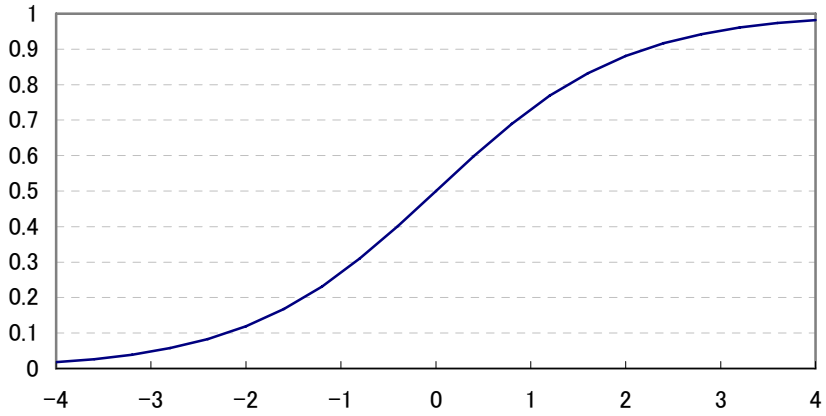
なお、ここでの係数 b は、X と Y^*の関係を示すものであって、X と確率の関係を表しているわけではないということに注意が必要です。したがって、X が変化したときに確率がどの程度変化するかを知りたいときには、係数とデータから計算される**限界効果**に注目します。

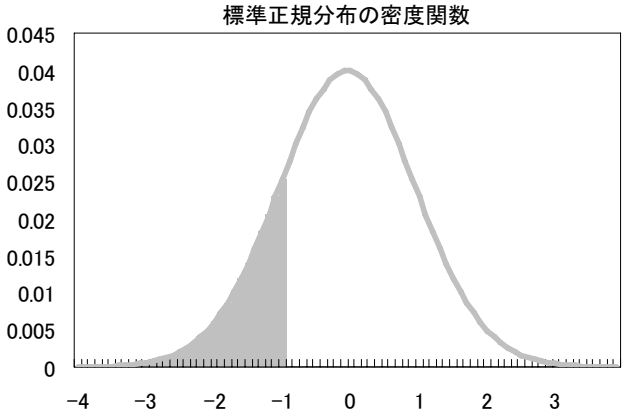
ロジスティック関数と標準積分関数

ロジスティック曲線とは、以下のように表されます。

$$Y = \frac{a}{1 + b\exp(cX)}$$

たとえば、a=1, b=1, c=1 のとき、X を－4 から＋4 まで変化させると Y は次のグラフのように変化します。

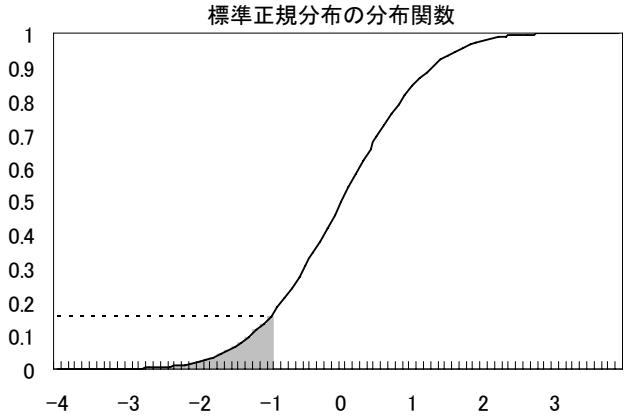




標準正規分布の分布関数（累積密度関数）も、ロジスティック曲線と似た形状をしています。

上の図は標準正規分布の密度関数で、たとえば、$-\infty < x \leqq -1$ の確率は図中のグレーの面積として表されます。

下の分布関数は、各々の$-\infty < x \leqq X$ に対応する確率を示します。



限界効果について

Y^*をアルバイトに対する意欲（潜在変数）とし、アルバイトによって得られる賃金（X）との関係を以下のような線形関数で表します。

$$Y^*=a+bX+u$$

そして、潜在変数 Y^*が 0 を超えると、Y が 1 になり、Y^*が 0 を下回るとき Y は 0 になるとします。つまり、アルバイトをする（Y=1）ときの条件は、

$$Y^*>0 \quad \Leftrightarrow \quad a+bX+u>0$$
$$\Leftrightarrow \quad u> \ -a \ -bX$$

Y が 1 をとる確率を P(Y=1)とすると、

$$P(Y=1)=P(Y^*>0) = P(\ u\ >-a-bX) = 1-F(-a-bX) \qquad (※1)$$
$$P(Y=0)=P(Y^*\leqq 0) = P(\ u\leqq -a-bX) = F(-a-bX) \qquad (※2)$$

と表せます。ここで、F は分布関数（累積密度関数）で、$F(Z)$と標記するとき、z が-∞<z≦Z の範囲をとるときの確率、P(-∞< z ≦ Z)を示します。

ここで、X が 1 増えたときに確率がどの程度変化するかは、（※1）を X で微分することで求められます。分布関数（累積密度関数）は、密度関数 f を積分したものであることに注意すると、

$$\frac{dP(Y=1)}{dX}=\frac{d(1-F(-a-bX))}{dX}=bf(-a-bX)$$

となります。確率 P と X の関係を知るためには、密度関数 f で(-a-bX)を変換した上で、係数 b をかける必要があることがわかります。なお、Ｓｔａｔａをはじめとする大抵の統計パッケージでは、簡単に限界効果を出力できるようになっています。

Ｓｔａｔａによる質的選択モデル

次の表は、縄田(1997)で紹介されている米国の人口センサス(U.S. Bureau of the Census の Current Population Survey)から抽出した 50 人の既婚女性の労働に関するデータです。このデータを用いて、女性の就業についての決定要因を分析してみましょう。

Variable		Obs	Mean	Std. Dev.	Min	Max
obs	標本番号	50	26	14.58	1	50
work	就業中=1, 非就業=0	50	1	0.48	0	1
c18	18歳未満の子供の数	50	1	1.31	0	5
age	年齢	50	44	13.40	22	70
ed	教育年数	50	13	2.75	2.5	21
hi	夫の所得	50	20508	21747.60	0	99999

プロビット・モデルの推定は、以下のコマンドにより行います。

probit [従属変数]　[独立変数]

```
. probit  work c18 age ed hi age

Probit regression                                 Number of obs   =         50
                                                  LR chi2(4)      =      11.62
                                                  Prob > chi2     =     0.0204
Log likelihood = -26.862914                       Pseudo R2       =     0.1778

------------------------------------------------------------------------------
        work |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         c18 |  -.3882481   .1935349    -2.01   0.045    -.7675696   -.0089266
         age |  -.0576319   .0208064    -2.77   0.006    -.0984116   -.0168522
          ed |    .052782   .0913544     0.58   0.563    -.1262694    .2318334
          hi |   9.50e-06   .0000102     0.93   0.350    -.0000104    .0000294
       _cons |    2.50193   1.500945     1.67   0.096    -.4398681    5.443728
------------------------------------------------------------------------------
```

z は、係数（Coef.）を標準誤差（Std. Err.）で割ったもので、通常の回帰分析の t 値に対応します。質的選択モデルにおける当てはまり具合の指標は、Pseudo R2（擬似決定係数）が用いられます。

プロビット・モデルにおける限界効果（各説明変数が変化したときに確率がどの程度変化するか）は、dprobit コマンドによって表示することができます。

```
. dprobit  work c18 age ed hi age

Probit regression, reporting marginal effects           Number of obs =     50
                                                        LR chi2(4)    =  11.62
                                                        Prob > chi2   = 0.0204
Log likelihood = -26.862914                             Pseudo R2     = 0.1778

------------------------------------------------------------------------------
    work |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
     c18 |  -.1404291   .0687614    -2.01   0.045         1  -.275199 -.005659
     age |  -.0208454   .0074181    -2.77   0.006     44.06  -.035385 -.006306
      ed |   .0190912   .0330081     0.58   0.563     12.76  -.045603  .083786
      hi |   3.44e-06   3.67e-06     0.93   0.350   20508.4  -3.8e-06  .000011
---------+--------------------------------------------------------------------
  obs. P |        .64
 pred. P |    .671018  (at x-bar)
------------------------------------------------------------------------------
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

ロジット・モデルは、以下のコマンドで推計が可能です。

logit [従属変数]　[独立変数]

```
. logit  work c18 age ed hi

Logistic regression                                     Number of obs   =         50
                                                        LR chi2(4)      =      11.58
                                                        Prob > chi2     =     0.0208
Log likelihood = -26.881062                             Pseudo R2       =     0.1772

------------------------------------------------------------------------------
        work |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         c18 |  -.6577524   .3525673    -1.87   0.062    -1.348772    .0332668
         age |  -.0953201    .035904    -2.65   0.008    -.1656907   -.0249496
          ed |   .0814968   .1579706     0.52   0.606    -.2281199    .3911134
          hi |   .0000173   .0000177     0.98   0.327    -.0000173    .0000519
       _cons |   4.214012    2.63546     1.60   0.110    -.9513944    9.379419
------------------------------------------------------------------------------
```

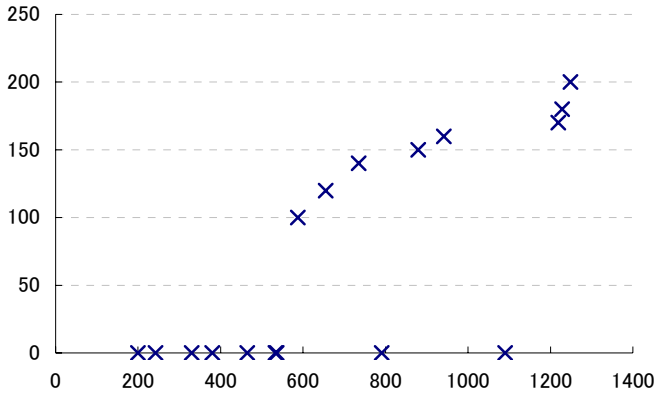
既婚女性の労働データ

obs	WORK	C18	AGE	ED	HI
1	0	0	69	16	0
2	1	0	27	12	37400
3	0	0	58	12	30000
4	1	2	29	12	18000
5	1	0	58	12	60000
6	0	1	36	12	55000
7	1	0	52	13	33000
8	1	0	29	16	28000
9	1	0	46	14	33000
10	0	0	67	7.5	0
11	0	0	65	12	0
12	1	0	51	12	29650
13	1	2	36	13	0
14	1	0	22	2.5	12000
15	1	1	30	14	45000
16	1	2	34	12	39000
17	0	3	38	16	39750
18	1	5	34	11	1200
19	0	0	48	11	0
20	0	3	27	12	14500
21	1	1	43	13	16887
22	1	2	33	12	28320
23	0	0	58	12	500
24	1	0	46	13	1000
25	1	0	52	21	99999
26	1	2	23	11	2300
27	0	2	32	14	11000
28	1	1	34	20	8809
29	0	1	37	11	32800
30	1	0	53	11	0
31	1	0	26	12	15704
32	0	5	42	13	41000
33	1	2	47	12	48200
34	1	1	43	14	0
35	0	0	62	12	0
36	1	1	29	12	0
37	0	0	63	13	0
38	0	0	57	10	0
39	0	3	34	16	20000
40	1	3	32	16	60000
41	1	0	60	12	33000
42	1	0	53	12	0
43	1	1	37	12	45000
44	0	0	70	12	25400
45	1	3	28	12	0
46	1	0	52	11	24000
47	0	1	38	13	0
48	1	0	57	16	14000
49	1	1	52	16	0
50	1	1	54	12	22000

縄田 (1997) P.249 より

トービット・モデル

トービット・モデルは、所得と自動車購入額や株の保有高のように、所有していなければ 0、所有しているときは実数値をとる変数を被説明変数として分析する際に用いられます。こうした変数は、0 以下にならないという意味で質的な側面を持つ変数で、0 以上のとき実数値をとるという意味では量的な側面を持つ変数です。図は、Y 自動車購入額（縦軸）と X 所得（横軸）の関係を示すものですが、低所得階層では自動車購入額がゼロとなっている世帯が多いことが分かります。こうした変数に、回帰直線を当てはめると、質的選択モデルの分析と同様の問題が生じます。こうしたデータのことを 0 で切断されたデータと呼びます。



トービット・モデルは、以下のようなモデルを推定します。

$$Y = \begin{cases} Y^* & Y^* > 0 \\ 0 & Y^* \le 0 \end{cases}$$

$$Y^* = a + bX + u$$

トービット・モデルの推定についての理論的背景については本書の範囲を超えるため割愛します。

事例：Theory of Extramarital Affairs～不倫の理論～

Fair(1976)は、余暇時間の使い方に関する分析を報告している。余暇時間に関する分析は古くから議論されているが、Fair の分析では、余暇時間のうち、家族と伴に過ごす時間と家族以外の者と過ごす時間の配分について分析し、その事例として、不倫行動の決定要因分析を紹介している。データは、1969 年 6 月の Psychology Today 誌によるアンケート調査である。サンプルには、未婚の人や離婚・再婚を繰り返している人も含まれるが、ここでは、

離婚経験の無い有配偶者に限定されている。

なお、この分析は、米国の一流経済学術誌である Journal of Political Economy の 1978 年２月号 P.45-P.61 に掲載されたものである。

Variable		Obs	Mean	Std. Dev.	Min	Max
ypt	不倫回数	601	1.455907	3.298758	0	12
sex	性別	601	0.475874	0.499834	0	1
age	年齢	601	32.48752	9.288762	17.5	57
years_married	結婚年数	601	8.177696	5.571303	0.125	15
children	子供の数	601	0.715474	0.451564	0	1
religious	信仰心(1がantireligion)	601	3.116473	1.167509	1	5
education	教育年数	601	16.16639	2.402555	9	20
marriage_happiness	結婚満足度(5がvery happy)	601	3.93178	1.103179	1	5

Ｓｔａｔａによるトービット・モデルの推定

トービット・モデルの推定は tobit コマンドを用います。

tobit [被説明変数] [説明変数], ll(0)

オプションの ll(0)は、被説明変数が、0 を下限（”ll”は、lower limit を意味する）として切断されていることを意味します。ある数値 x を上限として切断されている場合は、ul(x)と入力します。
推定結果の下には、切断されたサンプル数が表示されています。ここから、601 件中 451 件が 0 の値をとっていることがわかります。

```
Obs. summary:        451  left-censored observations at ypt<=0
                     150     uncensored observations
                       0 right-censored observations
```

比較のために、最小二乗法による推定結果も示されていますが、随分と結果が異なることが分かります。

```
. tobit  ypt marriage_happiness education religious children years_married age
sex,ll(0)

Tobit regression                                  Number of obs   =        601
                                                  LR chi2(7)      =      79.57
                                                  Prob > chi2     =     0.0000
Log likelihood =  -704.9511                       Pseudo R2       =     0.0534

------------------------------------------------------------------------------
         ypt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
marriage_h~s |  -2.289961   .4155051    -5.51   0.000    -3.105999   -1.473923
   education |    .092386   .2045277     0.45   0.652    -.3092995    .4940714
   religious |  -1.709846   .4059763    -4.21   0.000     -2.50717   -.9125228
    children |   .8984287   1.268116     0.71   0.479    -1.592108    3.388965
years_marr~d |    .537995    .146667     3.67   0.000      .249946     .826044
         age |  -.1904162   .0810143    -2.35   0.019    -.3495255   -.031307
         sex |     1.1831   1.005443     1.18   0.240    -.7915568    3.157756
       _cons |   7.365336    3.89438     1.89   0.059    -.2830927    15.01376
-------------+----------------------------------------------------------------
      /sigma |   8.270711   .5553742                      7.179975    9.361447
------------------------------------------------------------------------------

  Obs. summary:        451  left-censored observations at ypt<=0
                       150     uncensored observations
                         0 right-censored observations

. reg  ypt marriage_happiness education religious children years_married age sex

      Source |       SS       df       MS              Number of obs =     601
-------------+------------------------------           F(  7,   593) =   12.62
       Model |  846.786227     7  120.969461            Prob > F      =  0.0000
    Residual |  5682.2953    593   9.5822855           R-squared     =  0.1297
-------------+------------------------------           Adj R-squared =  0.1194
       Total |  6529.08153    600  10.8818026           Root MSE      =  3.0955

------------------------------------------------------------------------------
         ypt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
marriage_h~s |  -.7167147   .1199768    -5.97   0.000    -.9523459   -.4810835
   education |   .0179886   .0582489     0.31   0.758    -.0964105    .1323878
   religious |  -.4825125   .1116888    -4.32   0.000    -.7018663   -.2631588
    children |  -.2189835   .3442829    -0.64   0.525    -.8951457    .4571787
years_marr~d |   .1712488   .0412109     4.16   0.000     .0903116    .2521859
         age |  -.0491046    .022571    -2.18   0.030    -.0934333   -.0047758
         sex |   .1696991   .2841729     0.60   0.551    -.3884087    .7278068
       _cons |   5.757585   1.133735     5.08   0.000     3.530961    7.984209
------------------------------------------------------------------------------
```