

Rによるデータ分析入門：練習問題・回答例 第4章

1. 東大社研若年者パネル調査の公開データには支持政党に関する調査項目が含まれている。このデータを用いて自民党支持の決定要因を分析したい。

(1) 被説明変数に LDP (自民党支持ダミー), 説明変数に女性ダミー (female), 教育年数 (educ), 結婚ダミー (marriage), 年収 (income), 年齢 (age), 伝統的家族感 (value_family), 社会階層 (social_class) を説明変数として最小二乗法で回帰分析を実施せよ。また, 予測値を計算し, 予測値の最大値・最小値を調べよ。

```
lm(formula = LDP ~ female + educ + marriage + income + age +  
    value_family + social_class, data = dataf)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.38250	-0.22101	-0.16094	-0.07925	0.94797

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3.006e-01	1.602e-01	-1.876	0.06100 .
female	-7.766e-03	3.173e-02	-0.245	0.80669
educ	2.090e-03	7.723e-03	0.271	0.78678
marriage	-2.154e-02	3.210e-02	-0.671	0.50244
income	6.786e-05	8.538e-05	0.795	0.42694
age	8.773e-03	4.079e-03	2.151	0.03182 *
value_family	2.691e-02	1.174e-02	2.293	0.02214 *
social_class	2.660e-02	9.085e-03	2.928	0.00352 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3894 on 751 degrees of freedom

Multiple R-squared: 0.0345, Adjusted R-squared: 0.0255

F-statistic: 3.834 on 7 and 751 DF, p-value: 0.0004178

```
> dataf <- dataf %>% dplyr::mutate(yhat1=predict(result1))  
> summary(dataf$yhat1)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.02151	0.14250	0.19382	0.19236	0.24245	0.40993

予測値の最小値は-0.021、最大値は0.410。被説明変数は0, または1の値をとるが、予測値がマイナスの値になるは不自然な結果といえる。

(2) (1) と同じ式を Logit モデルで推計せよ。その予測確率を計算し、最大値・最小値を確認せよ。

```
glm(formula = LDP ~ female + educ + marriage + income + age +
     value_family + social_class, family = binomial(link = logit),
     data = dataf)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.9312531	1.1298581	-4.364	1.27e-05 ***
female	-0.0347096	0.2110617	-0.164	0.86937
educ	0.0200538	0.0506872	0.396	0.69237
marriage	-0.1398017	0.2079702	-0.672	0.50144
income	0.0003625	0.0005193	0.698	0.48511
age	0.0603269	0.0278536	2.166	0.03032 *
value_family	0.1780589	0.0782641	2.275	0.02290 *
social_class	0.1805256	0.0615708	2.932	0.00337 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 743.25 on 758 degrees of freedom
 Residual deviance: 716.68 on 751 degrees of freedom
 AIC: 732.68

Number of Fisher Scoring iterations: 4

```
> dataf <- dataf %>% dplyr::mutate(yhat2=predict(result2, type="response"))
> summary(dataf$yhat2)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05059	0.13683	0.18449	0.19236	0.23687	0.45163

Logit モデルの予測値は、最小値が 0.05、最大値が 0.452 と 0 と 1 の間に収まっていることがわかる。

(3) (2) の推計結果から、どのような人が自民党を支持していると考えられるか。

Logit モデルの推計結果より、年齢 (age)、伝統的家族感 (value_family)、社会階層 (social_class) の係数がいずれもプラスで統計的に有意となった。年齢については、年齢が高い人ほど自民党を支持する人が多いことを意味する。伝統的家族観と社会階層については『男性は収入を得て、女性は家庭と家族の面倒をみるべき』と強く考える人ほど、自身が社会階層の上位に位置していると感じている人ほど、自民党を支持していると解釈できる。

2. 企業は従業員に給与以外にもフリンジ・ベネフィット (fringe benefit) と呼ばれる福利厚生を支給します。この練習問題では、福利厚生に恵まれた仕事に就く人の属性を調べます。使用するデータは 1977 年に米国で実施された福利厚生に関するサーベイ調査のデータ 616 件で、Wooldridge (2010) の演習用データとして配布されているデータです。

(1) hrbens (労働時間当たりの福利厚生支給額) と pension (福利厚生支給額のうち年金) がゼロのサンプル数を調べよ。

```
> # fringe benefit を受けているダミー
> fringe <- fringe %>% dplyr::mutate(d_bens=if_else(hrbens>0,1,0))
> # ダミー変数が 0 の数→table() で出力
> table(fringe$d_bens)

 0    1
41 575
>
> # sum() 関数を使うほうが簡単に計算できます
> sum(fringe$hrbens==0)
[1] 41
```

(2) 被説明変数を hrbens, 説明変数には age (年齢), educ (教育年数), married (既婚ダミー), white (白人ダミー), male (男性ダミー) とする式を最小二乗法とトービットモ

デルで推定し、その係数を比較せよ。

```
> # 最小二乗法
```

```
> result1 <-lm(hrbens~age+educ+married+male+white, data=fringe)
```

```
> summary(result1)
```

Call:

```
lm(formula = hrbens ~ age + educ + married + male + white, data = fringe)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.41772	-0.43816	-0.06977	0.36431	2.14174

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-0.676054	0.161979	-4.174	3.43e-05	***
age	0.009503	0.002014	4.719	2.94e-06	***
educ	0.065203	0.009071	7.188	1.93e-12	***
married	0.158819	0.056756	2.798	0.0053	**
male	0.351265	0.052868	6.644	6.77e-11	***
white	0.085787	0.083385	1.029	0.3040	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.597 on 610 degrees of freedom

Multiple R-squared: 0.2005, Adjusted R-squared: 0.194

F-statistic: 30.6 on 5 and 610 DF, p-value: < 2.2e-16

```
>
```

```
>
```

```
> # Tobit model
```

```
> # censReg パッケージのインストールと呼び出しが必要
```

```
> res_tobit1 <-censReg::censReg(hrbens~age+educ+married+male+white, left=0, data=fringe)
```

```
> summary(res_tobit1)
```

Call:

```
censReg::censReg(formula = hrbens ~ age + educ + married + male +  
  white, left = 0, data = fringe)
```

Observations:

Total	Left-censored	Uncensored	Right-censored
616	41	575	0

Coefficients:

	Estimate	Std. error	t value	Pr(> t)
(Intercept)	-0.799316	0.171828	-4.652	3.29e-06 ***
age	0.010402	0.002138	4.864	1.15e-06 ***
educ	0.068846	0.009582	7.185	6.72e-13 ***
married	0.176727	0.060250	2.933	0.00335 **
male	0.365878	0.056142	6.517	7.17e-11 ***
white	0.085991	0.088380	0.973	0.33057
logSigma	-0.466170	0.029910	-15.586	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Newton-Raphson maximisation, 5 iterations

Return code 1: gradient close to zero (gradtol)

Log-likelihood: -592.3296 on 7 Df

```
> summary(margEff(res_tobit1))
```

	Marg. Eff.	Std. Error	t value	Pr(> t)
age	0.0095852	0.0019711	4.8629	1.475e-06 ***
educ	0.0634395	0.0088406	7.1760	2.096e-12 ***
married	0.1628495	0.0555211	2.9331	0.003482 **
male	0.3371472	0.0517992	6.5087	1.584e-10 ***
white	0.0792386	0.0814429	0.9729	0.330972

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

結果の比較: 最小二乗法 (OLS) の結果と Tobit モデルの限界効果の係数を比較すると、係数の符号と統計的に有意かどうかについてほぼ同じとなったが、いくつかの係数でその大き

さに若干の違いがみられる。たとえば male の係数は OLS では 0.351 だが Tobit の限界効果では 0.337 となった。Married の係数は OLS では 0.159 だが Tobit の限界効果では 0.163 となっている。

(3) (2) の推定式の被説明変数を pension に変更して、最小二乗法とトービットモデルで推定し、その係数を比較せよ。

```
> # (2)の推定式の被説明変数を pension に変更して、  
> # 最小二乗法とトービットモデルで推定し、その係数を比較せよ。  
> result2 <-lm(pension~age+educ+married+male+white, data=fringe)  
> summary(result2)
```

Call:

```
lm(formula = pension ~ age + educ + married + male + white, data = fringe)
```

Residuals:

Min	1Q	Median	3Q	Max
-1212.19	-423.52	-31.95	354.62	2212.43

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-843.638	151.419	-5.572	3.79e-08 ***
age	8.820	1.883	4.685	3.46e-06 ***
educ	65.036	8.480	7.670	6.85e-14 ***
married	85.731	53.056	1.616	0.107
male	334.438	49.422	6.767	3.09e-11 ***
white	84.622	77.949	1.086	0.278

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 558.1 on 610 degrees of freedom

Multiple R-squared: 0.194, Adjusted R-squared: 0.1874

F-statistic: 29.36 on 5 and 610 DF, p-value: < 2.2e-16

>

```
> res_tobit2 <- censReg::censReg(pension~age+educ+married+male+white, left=0, data=fringe)
> summary(res_tobit2)
```

Call:

```
censReg::censReg(formula = pension ~ age + educ + married + male +
  white, left = 0, data = fringe)
```

Observations:

Total	Left-censored	Uncensored	Right-censored
616	172	444	0

Coefficients:

	Estimate	Std. error	t value	Pr(> t)
(Intercept)	-1.434e+03	2.081e+02	-6.890	5.57e-12 ***
age	1.158e+01	2.549e+00	4.544	5.51e-06 ***
educ	8.559e+01	1.143e+01	7.485	7.13e-14 ***
married	1.402e+02	7.191e+01	1.950	0.0512 .
male	3.840e+02	6.683e+01	5.745	9.19e-09 ***
white	1.169e+02	1.066e+02	1.097	0.2726
logSigma	6.579e+00	3.575e-02	184.024	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Newton-Raphson maximisation, 7 iterations

Return code 1: gradient close to zero (gradtol)

Log-likelihood: -3705.716 on 7 Df

```
> summary(margEff(res_tobit2))
```

	Marg. Eff.	Std. Error	t value	Pr(> t)
age	8.8697	1.9522	4.5435	6.674e-06 ***
educ	65.5448	8.7615	7.4810	2.589e-13 ***
married	107.4020	55.0424	1.9513	0.05148 .
male	294.0315	51.3975	5.7207	1.664e-08 ***
white	89.5565	81.6300	1.0971	0.27303

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

結果の比較：最小二乗法(OLS)の結果とTobitモデルの限界効果の係数を比較すると、係数の符号について同じであったが、OLSでは非有意であった married がTobitでは10%水準で統計的に有意となるなどやや変化が見られた。係数の大きさについても、marriedの係数は85.7だったがTobitの限界効果では55となった。maleの係数もOLSでは334.4だがTobitの限界効果では294と変化している。