

Rによるデータ分析入門：練習問題・回答例 第7章

OECD が収集した個人レベルのデータ PIAAC を用いて、就業者のリスクリングの効果について分析してみましょう。PIAAC とは、Programme for the International Assessment of Adult Competencies の略で、OECD 加盟国等 24 か国・地域が参加する 16 ～ 65 歳までの男女個人を対象とした調査です。年齢や性別、学歴、職歴などに加えて「読解力」や「数的思考力」「IT を活用した問題解決能力」などが調査されています。PIAAC データには「過去 12 カ月にトレーニングを受けたか」という質問があるので、日本の就業者のデータを用いてトレーニング・プログラムを受講することで賃金があがるかを検証してみましょう。

(1) トレーニング・プログラムの効果を通常の最小二乗法で推計とどんな問題に直面すると考えられるか説明せよ。

トレーニング・プログラムを受ける人は意欲的な人なので能力が高く賃金が高い人である可能性がある。逆に、職務能力を補うためにトレーニング・プログラムを受ける人は、元々能力が低く賃金が低い人である可能性がある。こうした状況でトレーニング・プログラムの受講の有無と賃金を比べても、賃金はその人の意欲や職務能力の違いを反映している可能性があるため、トレーニング・プログラム受講の有無の因果効果を測れない。

(2) トレーニングをするか否かで、年齢、学歴、雇用の状況、結婚しているかどうか、子どもの有無に違いがみられるか、CreateTableOne () 関数を使って調べよ。

```
> dataf %>% tableone::CreateTableOne(vars=c("wage", "age", "educ_cat", "child", "couple", "empstat_edt"), strata="training")
```

Stratified by training					
	0	1		p	test
n	1201	1161			
wage (mean (SD))	1979.52 (8677.02)	2462.31 (3476.11)		0.103	
age (mean (SD))	49.11 (10.39)	46.99 (9.59)		<0.001	
educ_cat (mean (SD))	13.16 (1.87)	14.28 (1.86)		<0.001	
child (mean (SD))	2.12 (0.81)	2.02 (0.94)		0.011	
couple (mean (SD))	0.90 (0.30)	0.91 (0.28)		0.154	
empstat_edt (%)				<0.001	
full	720 (60.0)	903 (77.8)			
not employed	54 (4.5)	31 (2.7)			

part

427 (35.6)

227 (19.6)

P 値が小さく統計的に有意に差がある変数をみていくと、トレーニングプログラム受講者は、そうでない人に比べて age が低い (46.99 vs. 49.11, $p < 0.001$)、つまり若くて、教育水準が高く (educ_cat が大きい, 14.28 vs. 13.16, $p < 0.001$)、正規雇用の人が多い (empstat_edt, 77.8% vs. 60%, $p < 0.001$) ことがわかる。

(3) トレーニングへの参加ダミーを被説明変数に、説明変数に年齢、年齢の 2 乗値、学歴ダミー、雇用の状況ダミーを用いたロジット・モデルを推定し、どのような人がトレーニング・プログラムに参加する確率が高いかを調べよ。

```
> result_logit1<-glm(training~age+I(age^2)+factor(educ_cat)+factor(empstat_edt)+
+                      +couple+child,family = binomial(link = "logit"),data=datas)
> summary(result_logit1)
```

Call:

```
glm(formula = training ~ age + I(age^2) + factor(educ_cat) +
    factor(empstat_edt) + +couple + child, family = binomial(link = "logit"),
    data = datas)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.9597844	0.9358519	-2.094	0.03625	*
age	0.0792920	0.0411809	1.925	0.05417	.
I(age^2)	-0.0009718	0.0004358	-2.230	0.02576	*
factor(educ_cat)12	0.3982134	0.1567372	2.541	0.01106	*
factor(educ_cat)15	1.0653928	0.1668098	6.387	1.69e-10	***
factor(educ_cat)16	1.4974738	0.1649742	9.077	< 2e-16	***
factor(empstat_edt)not employed	-0.6531116	0.2419467	-2.699	0.00695	**
factor(empstat_edt)part	-0.6771408	0.1016382	-6.662	2.70e-11	***
couple	-0.0581049	0.1507581	-0.385	0.69993	
child	-0.0501944	0.0512126	-0.980	0.32703	

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3273.7 on 2361 degrees of freedom
Residual deviance: 3002.0 on 2352 degrees of freedom
AIC: 3022

Number of Fisher Scoring iterations: 4

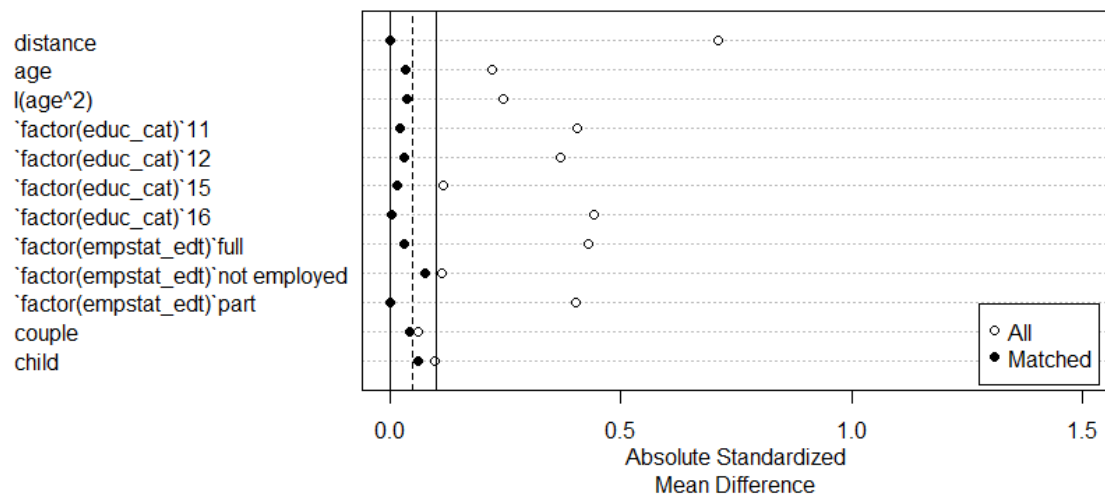
```
> DescTools::PseudoR2(result_logit1)
```

```
McFadden  
0.08301793
```

統計的に有意な変数をみていくと、年齢の係数がプラス、二乗項がマイナスなので、年齢が上がると参加確率は上昇するものの、年齢が一定の水準を超えると参加確率は下がっていく。学歴は高いほど、就業形態については正規雇用だとトレーニング・プログラムへの参加確率は高くなるのがわかる。

(4) 傾向スコア・マッチング法でトレーニングへの参加の有無によって賃金 (wage) に差がみられるかどうかを調べよ。賃金是对数をとること。

まず、ラブ・プロットでバランス・テストの結果をみる。●が破線の両脇の実線の内側に位置していれば、マッチングにより「よく似たペア」が選ばれていると判断できる。



マッチさせたデータで、トレーニングの受講の有無と賃金の関係を分析したのが以下の回帰分析である。トレーニングの受講の有無ダミー (training) の係数は0.16でトレーニングの受講すると16%賃金が高くなると解釈できる。

```
lm(formula = log(wage) ~ training, data = matched_data, weights = weights)
```

Weighted Residuals:

Min	1Q	Median	3Q	Max
-9.4093	-0.4642	-0.0473	0.4118	3.7414

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.36173	0.03819	192.762	< 2e-16 ***
training	0.16029	0.04602	3.483	0.000511 ***

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

Residual standard error: 0.8091 on 1437 degrees of freedom

(256 個の観測値が欠損のため削除されました)

Multiple R-squared: 0.008371, Adjusted R-squared: 0.007681

F-statistic: 12.13 on 1 and 1437 DF, p-value: 0.0005107

(5) 傾向スコア回帰でトレーニングへの参加の有無によって賃金 (wage) の対数値に差がみられるかどうかを調べよ。

(3) のロジットモデルの理論確率をウエイトとして、傾向スコア回帰を行ったのが以下の結果である。トレーニングの受講の有無 (training) の係数が 0.151 で統計的に有意となった。

(4) とほぼ同じ結果が得られていることがわかる。

```
lm(formula = log(wage) ~ training + age + I(age^2) + factor(educ_cat) +  
    factor(empstat_edt) + couple + child, data = dataf, weights = weight_ATT)
```

Weighted Residuals:

Min	1Q	Median	3Q	Max
-9.0838	-0.2460	-0.0096	0.2509	4.1457

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.0817246	0.3598410	16.901	< 2e-16 ***

training	0.1512367	0.0321112	4.710	2.65e-06 ***
age	0.0421583	0.0158663	2.657	0.00795 **
I (age^2)	-0.0003764	0.0001701	-2.213	0.02701 *
factor(educ_cat)12	0.0646290	0.0728413	0.887	0.37505
factor(educ_cat)15	0.1453679	0.0740634	1.963	0.04982 *
factor(educ_cat)16	0.4526270	0.0721636	6.272	4.36e-10 ***
factor(empstat_edt)not employed	-0.2698905	0.1216662	-2.218	0.02665 *
factor(empstat_edt)part	-0.6149071	0.0405199	-15.175	< 2e-16 ***
couple	0.1336841	0.0586668	2.279	0.02279 *
child	-0.0232190	0.0188987	-1.229	0.21937

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

Residual standard error: 0.7081 on 1971 degrees of freedom

(380 個の観測値が欠損のため削除されました)

Multiple R-squared: 0.2083, Adjusted R-squared: 0.2043

F-statistic: 51.85 on 10 and 1971 DF, p-value: < 2.2e-16