

R によるデータ分析入門：練習問題・回答例 第3章

1. 東京城南地区および川崎市の賃貸物件データ (rent-jonan-kawasaki.csv) を用いて、以下の回帰分析を実施せよ。

(1) 賃貸料 (rent_total) を被説明変数として、①占有面積 (floor), 築年数 (age), 駅からの時間距離 (walk と bus の合計, dist) の 3 つを説明変数とする回帰式, ②①にオートロックの有無 (auto_lock) を追加, ③②にケーブルテレビの有無 (catv) を追加した推計式を推定し, オートロックとケーブルテレビのどちらを追加したときに説明力がより大きく上昇するかを調べよ。結果は msummary で表にまとめること。

```
> regs <- list(  
+   "result1" = lm(rent_total ~ floor + age + dist, data=dataf),  
+   "result2" = lm(rent_total ~ floor + age + dist + auto_lock, data=dataf),  
+   "result3" = lm(rent_total ~ floor + age + dist + auto_lock + catv, data=dataf)  
+ )  
> modelsummary::msummary(regs, stars=TRUE)
```

msummary による結果は次のページ参照

自由度調整済み決定係数の変化に注目すると、

result1 → result2 0.331 - 0.293 = 0.038

result2 → result3 0.350 - 0.331 = 0.019

なのでオートロック付きダミーを追加したときの方が決定係数の上昇幅が大きい

| | result1 | result2 | result3 |
|---|-----------|-----------|-----------|
| (Intercept) | 10.230*** | 9.110*** | 8.733*** |
| | (0.346) | (0.388) | (0.393) |
| floor | 0.107*** | 0.104*** | 0.103*** |
| | (0.007) | (0.007) | (0.007) |
| age | -0.080*** | -0.053*** | -0.037** |
| | (0.013) | (0.013) | (0.014) |
| dist | -0.092** | -0.059* | -0.059* |
| | (0.028) | (0.028) | (0.028) |
| auto_lockYES | | 1.547*** | 1.293*** |
| | | (0.267) | (0.270) |
| catv | | | 1.293*** |
| | | | (0.308) |
| Num.Obs. | 584 | 584 | 584 |
| R2 | 0.297 | 0.336 | 0.355 |
| R2 Adj. | 0.293 | 0.331 | 0.350 |
| AIC | 2908.5 | 2877.5 | 2862.0 |
| BIC | 2930.4 | 2903.8 | 2892.6 |
| Log.Lik. | -1449.266 | -1432.767 | -1423.984 |
| F | 81.704 | 73.143 | 63.727 |
| RMSE | 2.89 | 2.81 | 2.77 |
| + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001 | | | |

(2) (1) で得られた回帰式を用いて、賃貸料の理論値を計算し、大森駅 (station = omori) を最寄り駅とする物件の中で、実際の賃貸料よりも賃貸料の理論値が最も大きく上回る物件 (お買い得物件) を探せ.

```
> # 3-1 (2)
> # (1) で得られた回帰式を用いて、賃貸料の理論値を計算し、大森駅
> # (station==omori) を最寄り駅とする物件の中で、実際の賃貸料よりも賃貸料の
> #理論値が最も大きく上回る物件 (お買い得物件) を探せ.
> # 予測値を yhat に格納する
> dataf <- dataf %>% dplyr::mutate(yhat=predict(result3))
> # 予測値 yhat と実績値 rent_total の差を計算
> dataf <- dataf %>% dplyr::mutate(diff=yhat-rent_total)
> # 大森に限定し、arrange() で diff を大きいほうから並べ、
> # select() で指定した変数を画面表示
> dataf %>% dplyr::filter(station=="omori") %>%
+   dplyr::arrange(-diff) %>%
+   dplyr::select(diff, yhat, rent_total, floor, age)
# A tibble: 37 × 5
   diff yhat rent_total floor age
  <dbl> <dbl>     <dbl> <dbl> <dbl>
1  2.43  12.4      10      53   32
2  2.36   8.36       6      13   35
3  2.05  10.2      8.15    27   26
4  1.94  13.8     11.9     27    1
5  1.59  11.9     10.3     38   16
6  1.55  11.2      9.7     19    1
7  1.34  10.1      8.8     29   19
8  0.832 12.4     11.6     25    0
9  0.743 12.2     11.5     46   27
10 0.709  9.31      8.6     18   14
# i 27 more rows
# i Use `print(n = ...)` to see more rows
```

rent_total が 10 万円の物件が理論家賃 12.4 万円で 2.4 万円割安である。賃貸料 10 万円前後の他の物件と比較すると、たとえば 5 行目の 10.3 万円の物件では専有面積 38 平米だが、賃貸料 10 万は 53 平米と広いことがわかる。

(3) 鉄道路線ダミー (JR 線ダミーと東急ダミー) を作成し, 東急沿線の物件の家賃が割高であるかどうか検討したい. 回帰式には, 賃貸料 (rent_total) を被説明変数とし, 鉄道路線ダミーと占有面積 (floor), 築年数 (age), 駅からの時間距離 (dist), ターミナルからの時間距離 (terminal) を説明変数とする回帰式を推定せよ.

```
lm(formula = rent_total ~ floor + age + auto_lock + dist + terminal +
    d_tokyu + d_jr, data = dataf)
```

Residuals:

| | Min | 1Q | Median | 3Q | Max |
|--|----------|---------|---------|--------|--------|
| | -22.5794 | -1.6188 | -0.2666 | 1.6019 | 9.3770 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|--------------|-----------|------------|---------|--------------|
| (Intercept) | 8.486724 | 0.472191 | 17.973 | < 2e-16 *** |
| floor | 0.114280 | 0.006837 | 16.714 | < 2e-16 *** |
| age | -0.067644 | 0.012884 | -5.250 | 2.14e-07 *** |
| auto_lockYES | 1.424180 | 0.257072 | 5.540 | 4.61e-08 *** |
| dist | -0.075741 | 0.027972 | -2.708 | 0.00698 ** |
| terminal | -0.061529 | 0.023621 | -2.605 | 0.00943 ** |
| d_tokyu | 2.120934 | 0.346850 | 6.115 | 1.78e-09 *** |
| d_jr | 0.741169 | 0.364799 | 2.032 | 0.04264 * |

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

Residual standard error: 2.711 on 576 degrees of freedom

Multiple R-squared: 0.3917, Adjusted R-squared: 0.3843

東急ダミーの係数は「(基準である) 京急沿線物件と比べて賃貸料がどの程度高いか」を示すので, 2.12 万円高いといえる. JR 沿線物件は京急沿線物件よりも 0.74 万円高い.

(4) 駅から徒歩圏内にある物件に比べて, バスを利用する物件は不便なので賃貸料が低くなると考えられる. そこで, バスを利用する物件 (bus>0 の物件) であれば 1 をとるダミー変数 (d_bus) を作成し, (3) で得られた回帰式に説明変数として追加せよ. 得られた係

数が期待される符号になっているか、統計的に有意かを検討せよ。

```
lm(formula = rent_total ~ floor + age + auto_lock + dist + d_bus +  
    terminal + d_tokyu + d_jr, data = dataf)
```

Residuals:

| | Min | 1Q | Median | 3Q | Max |
|--|----------|---------|---------|--------|--------|
| | -22.9142 | -1.6412 | -0.2379 | 1.5806 | 9.4202 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|--------------|-----------|------------|---------|--------------|
| (Intercept) | 8.290889 | 0.482806 | 17.172 | < 2e-16 *** |
| floor | 0.116024 | 0.006887 | 16.847 | < 2e-16 *** |
| age | -0.068990 | 0.012876 | -5.358 | 1.22e-07 *** |
| auto_lockYES | 1.368947 | 0.258238 | 5.301 | 1.64e-07 *** |
| dist | -0.063127 | 0.028725 | -2.198 | 0.0284 * |
| d_bus | -0.872204 | 0.468918 | -1.860 | 0.0634 . |
| terminal | -0.052033 | 0.024118 | -2.157 | 0.0314 * |
| d_tokyu | 2.106170 | 0.346202 | 6.084 | 2.15e-09 *** |
| d_jr | 0.884360 | 0.372074 | 2.377 | 0.0178 * |

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

Residual standard error: 2.705 on 575 degrees of freedom

Multiple R-squared: 0.3954, Adjusted R-squared: 0.3869

F-statistic: 47 on 8 and 575 DF, p-value: < 2.2e-16

d_bus が「バスを利用する場合 1」をとるダミー変数。係数はマイナスで t 値が-1.86、p 値が 0.0634 なので、10%水準で統計的に有意である。この係数の意味は、バスを利用する物件の場合 0.87 万円賃貸料が安いと解釈できる。

2. wage-census2022-by-ind.csv は、厚生労働省「賃金センサス」(2022)の年齢階級別、学歴歴別、企業規模別、産業別の所定内給与額である。これを使って以下の問いに答えよ。

(1) 被説明変数を賃金そのものとする回帰式と賃金の対数値とする回帰式を推定せよ。説明変数は、年齢、企業規模ダミー、教育水準ダミー、産業ダミーとする。また、各々回帰式の年齢の係数の意味を説明せよ。

被説明変数を賃金そのものとする回帰式

```
lm(formula = wage ~ age + factor(size) + factor(education) +
    factor(ind), data = dataf)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|-------|---------|
| -398.26 | -67.14 | -17.24 | 59.60 | 1345.41 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|--------------------|----------|------------|---------|--------------|
| (Intercept) | 230.3202 | 15.4175 | 14.939 | < 2e-16 *** |
| age | 0.9471 | 0.2309 | 4.103 | 4.35e-05 *** |
| factor(size)2 | -22.7903 | 8.6072 | -2.648 | 0.008202 ** |
| factor(size)3 | -28.7658 | 8.6225 | -3.336 | 0.000874 *** |
| factor(education)2 | -6.3279 | 12.0523 | -0.525 | 0.599652 |
| factor(education)3 | 13.3396 | 12.2146 | 1.092 | 0.274996 |
| factor(education)4 | 36.0792 | 12.2805 | 2.938 | 0.003364 ** |
| factor(education)5 | 73.3418 | 12.2403 | 5.992 | 2.70e-09 *** |
| factor(education)6 | 196.1199 | 12.5117 | 15.675 | < 2e-16 *** |
| factor(ind)2 | -3.1650 | 8.5062 | -0.372 | 0.709892 |
| factor(ind)3 | 84.2617 | 8.6833 | 9.704 | < 2e-16 *** |

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

Residual standard error: 125.4 on 1261 degrees of freedom

Multiple R-squared: 0.3002, Adjusted R-squared: 0.2947

F-statistic: 54.11 on 10 and 1261 DF, p-value: < 2.2e-16

被説明変数を賃金の対数値とする回帰式

```
lm(formula = lwage ~ age + factor(size) + factor(education) +  
    factor(ind), data = dataf)
```

Residuals:

| | Min | 1Q | Median | 3Q | Max |
|--|----------|----------|----------|---------|---------|
| | -1.20337 | -0.18781 | -0.01847 | 0.21362 | 1.75038 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|--------------------|------------|------------|---------|--------------|
| (Intercept) | 5.4723205 | 0.0373768 | 146.410 | < 2e-16 *** |
| age | 0.0021436 | 0.0005597 | 3.830 | 0.000134 *** |
| factor(size)2 | -0.0712230 | 0.0208663 | -3.413 | 0.000662 *** |
| factor(size)3 | -0.0944290 | 0.0209035 | -4.517 | 6.85e-06 *** |
| factor(education)2 | 0.0095700 | 0.0292184 | 0.328 | 0.743319 |
| factor(education)3 | 0.0857947 | 0.0296118 | 2.897 | 0.003829 ** |
| factor(education)4 | 0.1500202 | 0.0297715 | 5.039 | 5.36e-07 *** |
| factor(education)5 | 0.2584297 | 0.0296743 | 8.709 | < 2e-16 *** |
| factor(education)6 | 0.5042492 | 0.0303322 | 16.624 | < 2e-16 *** |
| factor(ind)2 | -0.0163008 | 0.0206216 | -0.790 | 0.429401 |
| factor(ind)3 | 0.1878870 | 0.0210510 | 8.925 | < 2e-16 *** |

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

Residual standard error: 0.3041 on 1261 degrees of freedom

Multiple R-squared: 0.3092, Adjusted R-squared: 0.3038

F-statistic: 56.45 on 10 and 1261 DF, p-value: < 2.2e-16

係数の意味の違いについて

たとえば企業規模 size3 のダミーの場合、この係数は大企業に比べて小企業がどの程度賃金が異なるかを示す。被説明変数を賃金そのものとする回帰式の場合、その係数は-28.8、被説明変数は月額給与で単位が千円なので2万8800円月給が低いと解釈できる。被説明変数を賃金の対数値とする回帰式の場合、係数は-0.09である。この場合、片対数モデルなので9%賃金が低いと解釈できる ($-\exp(\beta) - 1$ で計算すると-8.6%)。

(2) 産業によって年齢と賃金の関係がどのように異なるかを調べたい。被説明変数 を賃金
そのもの、説明変数には 1) の変数に年齢の 2 乗値、年齢と産業ダミーの交差項、年齢の 2
乗値と産業ダミーの交差項を加えた回帰式を推定せよ。製造業 (ind : 1)、卸小売業 (ind :
2)、金融・保険業 (ind : 3) で、賃金が最大となる年齢、その年齢における賃金を計算せよ。

```
lm(formula = wage ~ age * factor(ind) + I(age^2) * factor(ind) +  
    factor(size) + factor(education), data = dataf)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|-------|---------|
| -383.88 | -61.08 | -3.03 | 45.49 | 1264.63 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|-----------------------|------------|------------|---------|----------|-----|
| (Intercept) | -158.25571 | 48.80151 | -3.243 | 0.001215 | ** |
| age | 20.13639 | 2.18822 | 9.202 | < 2e-16 | *** |
| factor(ind)2 | -15.83654 | 67.18998 | -0.236 | 0.813705 | |
| factor(ind)3 | -263.85240 | 70.93842 | -3.719 | 0.000208 | *** |
| I(age^2) | -0.20907 | 0.02306 | -9.068 | < 2e-16 | *** |
| factor(size)2 | -18.36399 | 7.54953 | -2.432 | 0.015135 | * |
| factor(size)3 | -27.03881 | 7.56195 | -3.576 | 0.000363 | *** |
| factor(education)2 | -4.63382 | 10.56587 | -0.439 | 0.661052 | |
| factor(education)3 | 7.85832 | 10.71482 | 0.733 | 0.463448 | |
| factor(education)4 | 26.94417 | 10.77793 | 2.500 | 0.012548 | * |
| factor(education)5 | 67.44833 | 10.73860 | 6.281 | 4.63e-10 | *** |
| factor(education)6 | 186.10738 | 10.98359 | 16.944 | < 2e-16 | *** |
| age:factor(ind)2 | 0.52274 | 3.03805 | 0.172 | 0.863415 | |
| age:factor(ind)3 | 14.41947 | 3.19551 | 4.512 | 7.01e-06 | *** |
| factor(ind)2:I(age^2) | -0.00442 | 0.03190 | -0.139 | 0.889809 | |
| factor(ind)3:I(age^2) | -0.13596 | 0.03352 | -4.055 | 5.31e-05 | *** |

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

Residual standard error: 110 on 1256 degrees of freedom

Multiple R-squared: 0.4644, Adjusted R-squared: 0.458

F-statistic: 72.61 on 15 and 1256 DF, p-value: < 2.2e-16

賃金が最大となる年齢= age の係数 / ($2 \times \text{age}$ の 2 乗値の係数)

で計算できます。計算結果は、

製造業 48.2 歳

卸小売 48.4 歳

金融保険 50.1 歳

となる。

その年齢における賃金

大企業 (size: 1)、大卒 (education: 5) の賃金を計算すると、

製造業 39.4 万円

卸小売 41.0 万円

金融保険 77.4 万円

となる。

※計算の詳細は別添の”練習問題 2-2.xlsx”を参照のこと。

3. 早生まれは不利か？ 1 ~ 3 月生まれの人を「早生まれ」といいますが、幼稚園・保育園や小学校 低学年では「早生まれ」の子供は運動能力や認知能力で劣っているといわれている。こうした「早生まれ」の損失は成人後も続くものなのでしょうか。これを、第 2 章 の練習問題でも使用した東大社研若年者パネル非制限公開データ (today-shaken.csv) を用います。このデータはアンケートの回答がそのまま記録されているので、まず分析目的に沿った変数を作成し回帰分析を実施します。

(1) 省略 (ex3-3.R 参照)

(2) 回帰分析から born_early の係数が有意になるか確認せよ

回帰式 1

```
> # 3-3 (2)
```

```
> # 回帰分析
```

```
> result1 <- lm(educ ~ born_early, data = dataf)
```

```
> summary(result1)
```

Call:

```
lm(formula = educ ~ born_early, data = dataf)
```

Residuals:

| | Min | 1Q | Median | 3Q | Max |
|--|---------|---------|---------|--------|--------|
| | -5.2437 | -1.9052 | -0.2437 | 1.7563 | 4.0948 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|------------|
| (Intercept) | 14.24373 | 0.07614 | 187.081 | <2e-16 *** |
| born_early | -0.33851 | 0.14535 | -2.329 | 0.0201 * |

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

Residual standard error: 1.799 on 767 degrees of freedom

(231 個の観測値が欠損のため削除されました)

Multiple R-squared: 0.007022, Adjusted R-squared: 0.005727

F-statistic: 5.424 on 1 and 767 DF, p-value: 0.02012

回帰式 2

```
> result2 <-lm(educ~born_early+age+educ_pa+n_siblings, data=ataf)
> summary(result2)
```

Call:

```
lm(formula = educ ~ born_early + age + educ_pa + n_siblings,
    data = dataf)
```

Residuals:

| | Min | 1Q | Median | 3Q | Max |
|--|---------|---------|--------|--------|--------|
| | -5.7378 | -1.2522 | 0.0371 | 1.0549 | 4.6700 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|------------|
| (Intercept) | 9.94517 | 0.62302 | 15.963 | <2e-16 *** |
| born_early | -0.20937 | 0.14057 | -1.489 | 0.1369 |
| age | 0.04069 | 0.01646 | 2.472 | 0.0137 * |
| educ_pa | 0.27460 | 0.02515 | 10.917 | <2e-16 *** |
| n_siblings | -0.18502 | 0.07884 | -2.347 | 0.0192 * |

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

Residual standard error: 1.613 on 668 degrees of freedom

(327 個の観測値が欠損のため削除されました)

Multiple R-squared: 0.1636, Adjusted R-squared: 0.1586

F-statistic: 32.67 on 4 and 668 DF, p-value: $< 2.2e-16$

回帰式 1 の born_early の係数

born_early -0.33851 0.14535 -2.329 0.0201 *

回帰式 2 の born_early の係数

born_early -0.20937 0.14057 -1.489 0.1369

回帰式 1 では born_early の係数は-0.33 とマイナスで統計的に有意となった。この結果から、他の説明変数を考慮しない状況では早生まれの方が、教育年数が短いことを示唆する。一方、年齢、父親の教育年数や兄弟姉妹の数などの他の説明変数を導入した回帰式 2 では、係数は依然としてマイナスであるが、t 値は-1.489、p 値は 0.1369 で係数は統計的に非有意となった。つまり、今日行く年数は、年齢、ならびに父親の教育年数や兄弟姉妹の数などの家庭環境との相関が強く、これらの変数を考慮すると、早生まれと教育年数の間には関係性が見いだせなくなると結論付けることができる。