第2回：線形回帰のモデル選択・回帰診断・外挿

Code

1. 前準備

今回は、線形回帰にまつわる話題から

- モデル選択（変数選択）
- 残差分析
- 外挿

を解説していきます。デモデータには前回同様 salary.csv 、ある会社の社員に関する
月給 、 勤続年数 、 仕事の達成度 、 欠勤日数 、 特殊免許の有無 の情報が記録されたデータを用います。

社員の 月給 の傾向を 勤続年数 、 仕事の達成度 、 欠勤日数 、 特殊免許の有無 の4変数で説明する線形回帰は

$$\text{月給} = \beta_0 + \beta_1 \times \text{勤続年数} + \beta_2 \times \text{仕事の達成度} + \beta_3 \times \text{欠勤日数} + \beta_4 \times \text{特殊免許の有無} + \text{誤差}$$

と表すことができ、偏回帰係数 $\beta_0, \beta_1, \cdots, \beta_4$ の推定値はR言語の lm 関数を用いて次のように推定することができるのでした。（復習されたい方は、第1回の資料をご覧ください。）

Hide

```
# 線形回帰の計算結果
dat <- read.csv(file = "./data/salary.csv",
                fileEncoding = "utf-8")
result <- lm(formula = 月給 ~ ., data = dat)
summary(result)
```

```
Call:
lm(formula = 月給 ~ ., data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-15413  -3884     -5   3004  88154

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    213082.90    9511.38  22.403  < 2e-16 ***
勤務年数         4761.56     271.51  17.538  < 2e-16 ***
仕事の達成度       93.07      35.25   2.640  0.00969 **
欠勤日数          262.71     975.75   0.269  0.78833
特殊免許の有無   3977.66    2150.76   1.849  0.06751 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10630 on 95 degrees of freedom
Multiple R-squared:  0.7746,    Adjusted R-squared:  0.7651
F-statistic: 81.61 on 4 and 95 DF,  p-value: < 2.2e-16
```

2. 線形回帰のモデル選択

変数 $x_1, \cdots, x_D$ を用いて変数 y を説明するような線形回帰

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_D x_D + \text{誤差}$$

を作るとき、説明変数 $x_1, \cdots, x_D$ の組み合わせは事前知識によって入念に検討されて決まることが一般的です。しかし実際には、変数 x_d を説明変数に含めた方が良いかいなか悩ましい変数もあるでしょう。そこで、何らかの基準に基づいて変数を選択する方法がいくつか提案されてきました。

- 偏回帰係数の統計的仮説検定
- 自由度調整済み決定係数
- 赤池情報量規準（AIC）

これらの手法を、以下の疑問に答える形で解説します。

Remark : 他にもMallows' Cp規準やベイズ情報量規準、正則化を用いる方法などがよく知られています。

問題 : 変数 `欠勤日数` を説明変数に含めるかいなか、上記の3通りの方法で検討してみてください。

A. 偏回帰係数の統計的仮説検定

`欠勤日数` を説明変数に含めるかいなかを検討する方法の一つに、 `欠勤日数` の偏回帰係数が 0 かいなかを検討するという考え方があります。これは、偏回帰係数の統計的仮設検定によって実現できます。

解答 : 前回紹介した偏回帰係数の統計的仮説検定を、モデル選択に用いてみましょう。ここで考える帰無仮説と対立仮説は、

$$H_0 : \beta_3 = 0 \text{ v.s. } H_1 : \beta_3 \neq 0$$

です。このとき、検定統計量の値（t値）は

$$t = \frac{263 - 0}{976} \sim 0.269$$

です。

Hide

```
# Rでt値を計算してみよう。
t_value <- 263 / 976
t_value
```

```
[1] 0.2694672
```

両側検定であることに注意すると、有意水準 5% のとき棄却域は$|t| \geq 1.99$ （R言語を用いて `qt(p=0.975, df=100-(4+1))` と計算できます）です。

Hide

```
# Rで両側検定の棄却限界を計算してみよう。
critical <- qt(p = 0.975, df = 100-(4+1))
critical
```

```
[1] 1.985251
```

すなわち、帰無仮説は棄却されません。すなわち「欠勤日数は説明変数として含めなくて良いだろう」という結果になります。なおこの結果は、 `summary` 関数の出力 `t-value` , `Pr(>|t|)` によって確認することもできます。

B. 自由度調整済み決定係数

欠勤日数 を説明変数に含めるかいなかを検討するもう一つの方法に、目的変数を精度よく予測する式が作れるかいなかという考え方があります。偏回帰係数を推定して得られる式の予測の精度を測る指標に、自由度調整済み決定係数があります。

自由度調整済み決定係数（adjusted R-squared）は、目的変数に対する予測の精度を、線形回帰によるものと標本平均によるものとで比較したものです。$\hat{y_i}$ を標本点 i の目的変数の値 y_i に対する線形回帰の予測値、$\bar{y}$ を目的変数の標本平均とします。このとき、自由度調整済み決定係数は次のように定義されます。

$$R^{*2} = 1 - \frac{\sum_{i=1}^{n}(\hat{y_i} - y_i)^2/\{n-(D+1)\}}{\sum_{i=1}^{n}(\bar{y} - y_i)^2/(n-1)}$$

自由度調整済み決定係数は、1 以下の値をとります。なお、負の数になることもあります。この値が 1 に近い線形回帰のほうが予測の精度が高いと推定できます。

Remark : 自由度調整済み決定係数の分子は線形回帰モデルが正しいと仮定したときの誤差の分散の不偏推定量、分母はなんら説明変数を用いなかったときの誤差の分散の不偏推定量になっています。このため、自由度調整済み決定係数は決定係数に比べて、母集団の決定係数を偏りなく推定することができます。つまり、決定係数は偏回帰係数の推定に用いたデータへのあてはまりの良さ、自由度調整済み決定係数は未知のデータへのあてはまりの良さの指標だと解釈できます。

解答 : 自由度調整済み決定係数は、 summary 関数の Adjusted R-squared に書かれています。今回、変数 欠勤日数 を含む線形回帰の自由度調整済み決定係数は 0.7651 です。一方、 欠勤日数 を含まない自由度調整済み決定係数は、以下のスクリプトから 0.7674 とわかります。

Hide

```
result2 <- lm(formula = 月給 ~ 勤務年数+仕事の達成度+特殊免許の有無,
              data = dat)
summary(result2)
```

```

Call:
lm(formula = 月給 ~ 勤務年数 + 仕事の達成度 + 特殊免許の有無,
    data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-15764  -4305    115   2984  88058

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    214630.76    7540.68  28.463  < 2e-16 ***
勤務年数         4744.69     262.89  18.048  < 2e-16 ***
仕事の達成度       92.59      35.04   2.643  0.00961 **
特殊免許の有無   3936.54    2134.94   1.844  0.06829 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10570 on 96 degrees of freedom
Multiple R-squared:  0.7744,    Adjusted R-squared:  0.7674
F-statistic: 109.9 on 3 and 96 DF,  p-value: < 2.2e-16
```

変数 `欠勤日数` を含む線形回帰より、含まない線形回帰のほうが自由度調整済み決定係数が高く、 `欠勤日数` を含まない場合の線形回帰のほうが未知のデータに対する予測の精度が高いと推定されます。すなわち「欠勤日数は説明変数として含めなくて良いだろう」という結果になります。

C. 赤池情報量規準

未知のデータへの予測の精度を推定する方法には、自由度調整済み決定係数の他にも**赤池情報量規準**（Akaike Information Criterion, AIC）を使う方法があります。赤池情報量規準は、次のように定義されます。

$$AIC = n \log \sum_{i=1}^{n} (\hat{y_i} - y_i)^2 + 2(D+1)$$

この値が小さい線形回帰のほうが予測の精度が高いと推定できます。

Remark : 赤池情報量規準は、Kullback-Leibler divergenceとよばれる母集団と推定によって得た線形回帰の間の近さを測る指標の（漸近）不偏推定量になっています。

解答 : 赤池情報量規準は、 `AIC` 関数を用いて計算することができます。

Hide

```
# 欠勤日数を含む場合のAICと、含まない場合のAIC
AIC(result); AIC(result2)
```

```
[1] 2144.875
[1] 2142.952
```

この結果、欠勤日数を含む場合の赤池情報量規準の値として 2144.875、含まない場合の赤池情報量規準の値として 2142.952 が得られ、 `欠勤日数` を含まない場合の線形回帰のほうが未知のデータに対する予測の精度が高いと推定されます。すなわち「欠勤日数は説明変数として含めなくて良いだろう」という結果になります。

3. 回帰診断

線形回帰モデルの偏回帰係数を最小2乗法で推定することは、以下の場合に「適切な」推定になっていることが数理統計学で確認されてきました。

- 誤差 ϵ の分散が説明変数によらず一定な場合。
- 誤差が正規分布に従っていると仮定する。

深くは立ち入りませんが、前者の場合はGauss-Markovの定理とよばれるもの、後者の場合は最尤法と最小2乗法が等しくなるというものです。（詳しくは数理統計学の講座を受講されてみてください。）

また、最小2乗法は残差の「二乗和」を小さくするように偏回帰係数を求めるため、**外れ値**に敏感であることが知られています。そこで、線形回帰を計算した後は

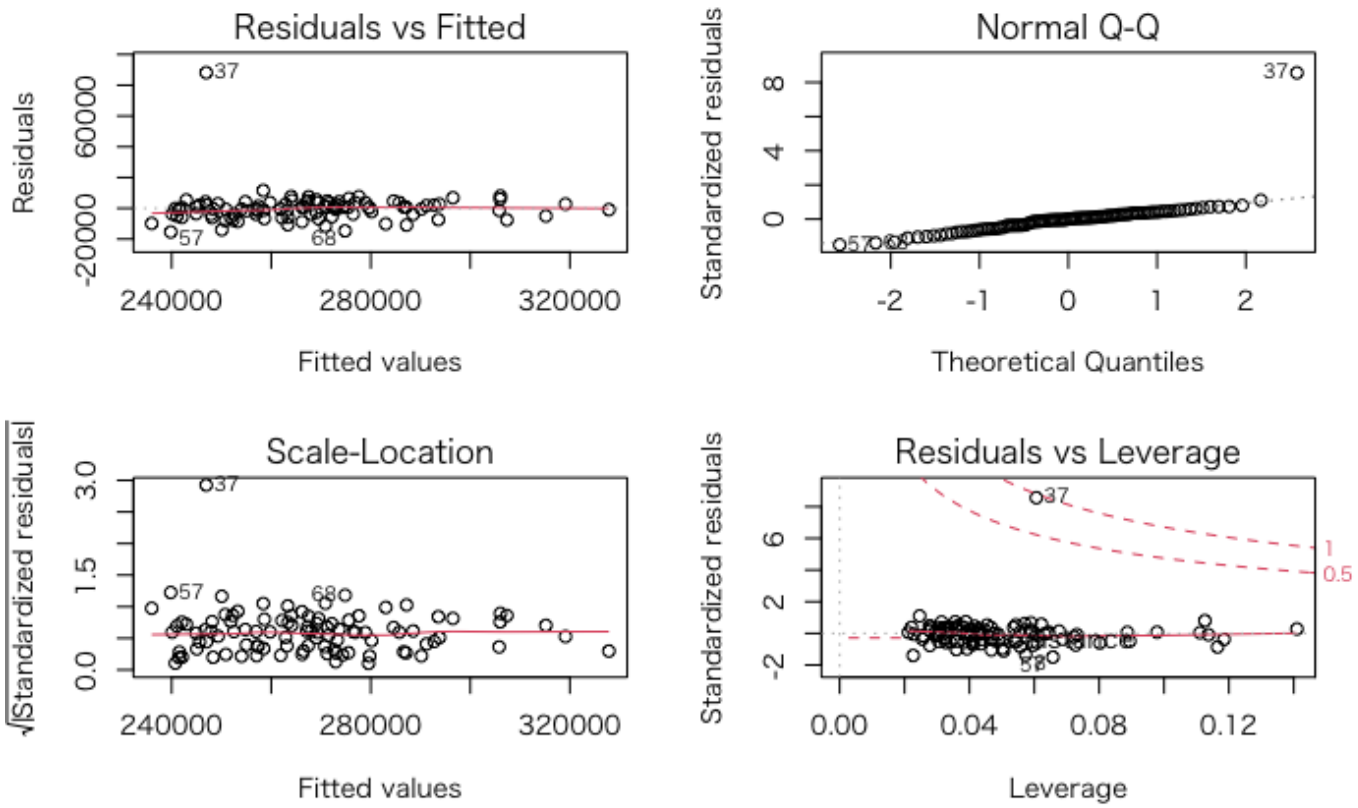
- 誤差が説明変数の値によらず一定か？
- 誤差が正規分布に従っているか？
- 外れ値がないか？

を確認し、偏回帰係数の推定の適切さを見積もったり、外れ値と考えた標本点を除去して偏回帰係数の推定値が適切な値になるかを実際にやってみて検討したりします。これを**回帰診断**（regression diagnostics）といいます。

残差分析はR言語の場合、 `plot` 関数に線形回帰の計算結果を渡すことで実行できます。

Hide

```
# 残差分析
par(family = "ヒラギノ角ゴシック W3")
par(mfrow = c(2, 2))
plot(result)
```



問題 : 上の図を参考に、以下の事項について検討してください。

1. 除去するべき外れ値はあるか？
2. 誤差は説明変数によらず一定か？
3. 誤差は正規分布に従っているとみなせるか？

解答 : 上の図には、 Fitted values , Residual , Standardized residual , Theoretical Quantities , Leverage , Cook's distance という6つの単語が出てくるため、これを解説します。

- Fitted values ：目的変数の予測値のこと。
- Residuals ：目的変数の実測値と予測値の差のこと。
- Standardized residuals ：各標本点における誤差 ϵ_i のz得点の推定値のこと。
- Theoretical Quantities ： Standardized residuals を小さい順に並べたとき、対応する標準正規分布の分位点の値のこと。
- Leverage ：その標本点の目的変数の実現値を変化させたとき、予測値がどれだけ変化するか。数式で表すと $\partial \hat{y_i} / \partial y_i$ のこと。日本語では「てこ比」という。
- Cook's distance ：次のように定義される外れ値の評価指標のひとつ。 Standardized Residual が大きいほど、また Leverage が大きいほど外れ値だと考えることができるため、このような式になっている。

$$\text{Cook's distance} = \text{Standardized Residual} * \frac{\text{Leverage}}{1 - \text{Leverage}}$$

1. 除去するべき外れ値はあるか？: 4番目の図から、37行目の標本点のCook距離が大きく、外れ値であろうと推測できます。また、1番目の図と3番目の図からも、37行目の標本点だけが大きな誤差の実現値を持っていると推定できる。ゆえに、37行目の標本点が外れ値であろうと考えられる。

2. 誤差は説明変数によらず一定か？：1番目の図と3番目の図はいずれも、誤差の推定値と予測値の間の関係を表したものです。いずれも、予測値によらず誤差の値は同じ範囲を散布していることから、誤差は説明変数によらず一定であろうと推測できます。
3. 誤差は正規分布に従っているとみなせるか？：2番目の図は誤差のz得点の推定値と対応する標準正規分布の推定値を比較したものです。37行目を除いて、直線 $y = x$ に点が並んでいることからから、誤差は正規分布に従っていると仮定して良いだろうことがわかります。

以上で解答は終わりです。■

なお、今回の問題で外れ値と考えた37行目を除去して改めて線形回帰を行うと、次のような結果になります。回帰診断の結果、外れ値は十分に除去できており、また偏回帰係数の推定も適切に行えているだろうことが推測できます。

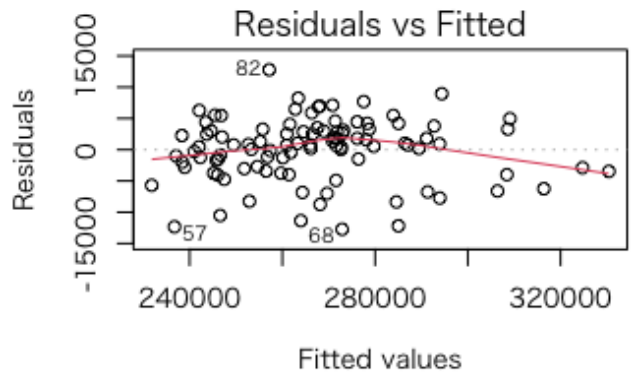
Hide

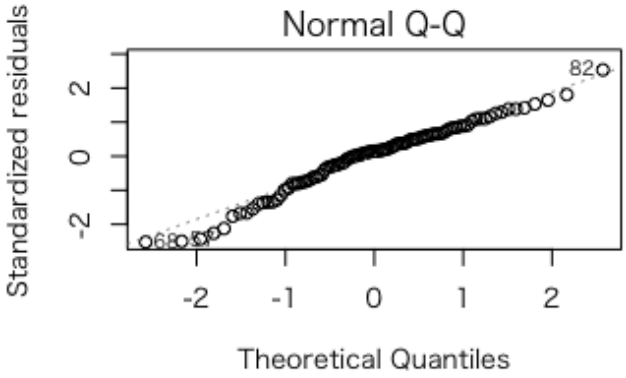
```
result3 <- lm(formula = 月給 ~ .,
              data = dat[-37, ])
summary(result3)
```

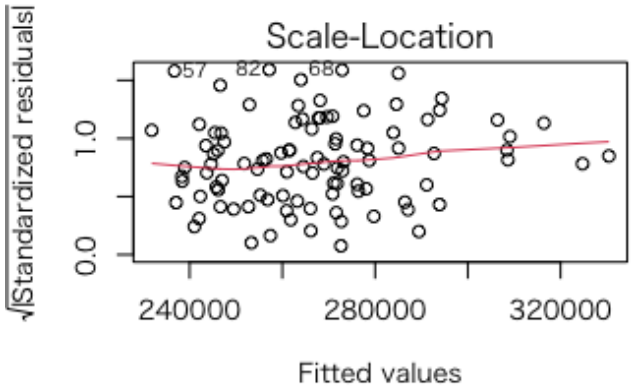
```

Call:
lm(formula = 月給 ~ ., data = dat[-37, ])

Residuals:
     Min       1Q   Median       3Q      Max 
-12731.8  -2960.8    739.8   3174.7  12787.0 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    196501.81    4667.92  42.096  < 2e-16 ***
勤務年数         4994.37     131.22  38.061  < 2e-16 ***
仕事の達成度      151.81      17.27   8.790 6.88e-14 ***
欠勤日数          551.78     469.52   1.175    0.243    
特殊免許の有無   5399.52    1037.38   5.205 1.14e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5110 on 94 degrees of freedom
Multiple R-squared:  0.9429,    Adjusted R-squared:  0.9405 
F-statistic: 388.3 on 4 and 94 DF,  p-value: < 2.2e-16
```

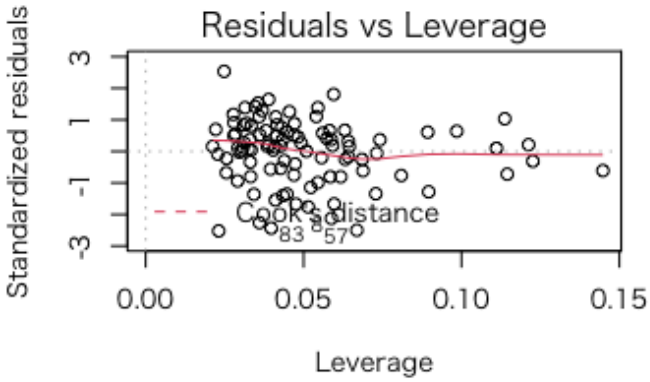
Hide

```
# 残差分析
par(family = "ヒラギノ角ゴシック W3")
par(mfrow = c(2, 2))
plot(result3)
```









4. 外挿

早速ですが、以下の問題を考えてみましょう。

問題 :

1. 勤続年数・仕事の達成度・特殊免許の有無の3変数を用い、また37行目を除去したデータを用いて、月給を説明する線形回帰を計算してください。
2. 1の結果を用いて、勤続年数が40年、仕事の達成度が200、特殊免許を持っている社員の月給を予測してください。
3. 2の結果をみた分析者は、一般にこのような高い月給が発生することはなく、予測が妥当ではないと考えました。ではなぜ、このような現実的ではない予測が得られたのか、その理由を考えてください。

解答 : 線形回帰は以下のように計算できます。

Hide

```
result4 <- lm(formula = 月給 ~ 勤務年数+仕事の達成度+特殊免許の有無,
              data = dat[-37, ])
summary(result4)
```

```
Call:
lm(formula = 月給 ~ 勤務年数 + 仕事の達成度 + 特殊免許の有無,
    data = dat[-37, ])

Residuals:
     Min       1Q   Median       3Q      Max
-13074.5  -3025.8    598.1   3333.9  12692.1

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    199786.79    3745.98  53.334  < 2e-16 ***
勤務年数         4958.44     127.86  38.779  < 2e-16 ***
仕事の達成度      150.65      17.28   8.719 9.00e-14 ***
特殊免許の有無   5310.01    1036.65   5.122 1.58e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5120 on 95 degrees of freedom
Multiple R-squared:  0.9421,    Adjusted R-squared:  0.9403
F-statistic: 515.2 on 3 and 95 DF,  p-value: < 2.2e-16
```

この結果、勤続年数が40年、仕事の達成度が200、特殊免許を持っている社員の月給は次のように予測できます。

$$199787 + 4958 \times 40 + 151 \times 200 + 5310 \times 1 = 433617$$

さて、分析者はこの予測について「一般にこのような高い月給が発生することはなく、予測が妥当ではない」と考察しています。このような現実的ではない予測が生じた理由は、データに含まれる説明変数の値の範囲を確認するとわかります。

Hide

```
# 説明変数の値の範囲
summary(dat)
```

```
      月給            勤務年数       仕事の達成度       欠勤日数 特殊免許の有無
 Min.   :224399   Min.   : 0.00   Min.   :108.0   Min.   :3   Min.   :0.00
 1st Qu.:251698   1st Qu.: 4.00   1st Qu.:182.8   1st Qu.:4   1st Qu.:0.00
 Median :267124   Median : 6.50   Median :207.5   Median :5   Median :0.00
 Mean   :267768   Mean   : 6.84   Mean   :203.0   Mean   :5   Mean   :0.48
 3rd Qu.:278390   3rd Qu.: 9.00   3rd Qu.:224.2   3rd Qu.:6   3rd Qu.:1.00
 Max.   :335092   Max.   :19.00   Max.   :263.0   Max.   :8   Max.   :1.00
```

勤務年数は最大で19年の社員しかデータには含まれていません。つまり今回、月給が線形回帰で予測できることを確かめることができたのは、勤務年数が19年までの社員です。実際、勤続年数が長くなるに従って、昇給額は小さくなり線形回帰があてはまらなくなる可能性があります。今回もこのような理由で、勤続年数が40年の社員には線形回帰の予測値が妥当な額にならない現象が発生したと考えられます。■

このように、線形回帰の計算に用いたデータに含まれる説明変数の値の範囲を超えた値をもつような標本点に対して、得られた線形回帰による予測を行うことを**外挿**（extrapolation）といいます。