Confidential

統計解析 – Day4

回帰モデル

AIJobColle

統計解析

【Day4】講義内容

- 回帰モデル
  - 回帰モデル
  - モデルのパラメーターの推定(最小二乗法)
  - 結果の確認方法

- 最尤推定
  - 最尤推定とは
  - 最尤推定と最小二乗法

- ダミー変数、多重共線性、AIC、変数選択法
  - ダミー変数
  - 多重共線性
  - AIC
  - 変数減少法、増加法、増減法

線形回帰モデルと最小二乗法

説明変数/独立変数(x)から、連続値である**目的変数/従属変数**(y)を予測する

単回帰モデル：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ , \quad \varepsilon_i \sim N(0, \sigma^2)$$

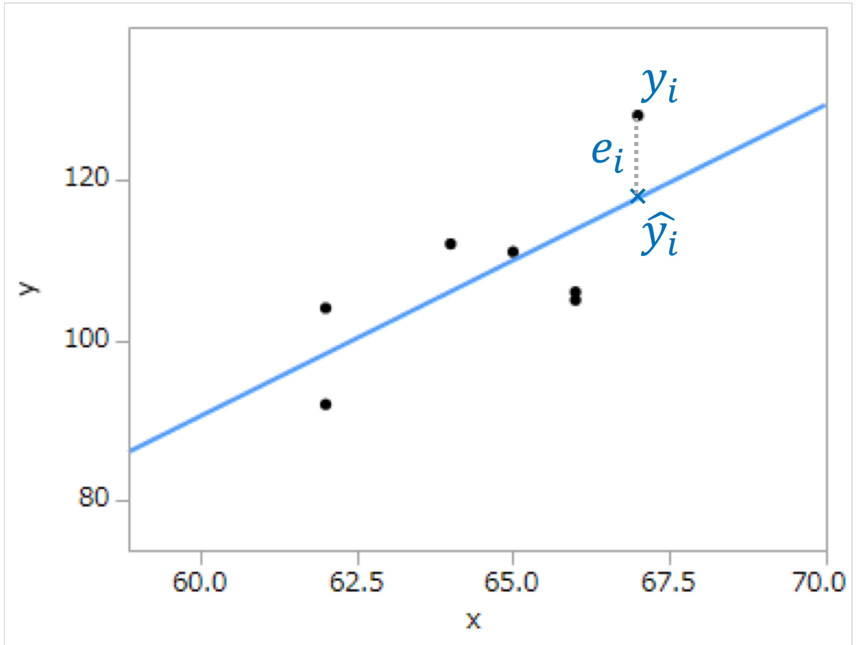
モデルの前提
- 説明変数と目的変数は線形の関係にある
- 残差は正規分布に従う
- 残差の分散(σ^2)はすべてのデータで共通

回帰係数(β_0, β_1)の推定値を$\widehat{\beta_0}, \widehat{\beta_1}$とする
y_iの予測値が $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$ となる

$e_i = y_i - \widehat{y_i}$ を**残差**と呼び、残差の2乗和を最小にすること
により、回帰係数を推定するアプローチを**最小二乗法**と呼ぶ

$\sum_i {e_i}^2 = \sum_i (y_i - \widehat{y_i})^2 = \sum_i (y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i))^2$ ⇒ 最小となる回帰係数を求める

また、σ^2の推定値は$\widehat{\sigma^2} = {\sum_i {e_i}^2}/_{n-2}$となる



説明変数が複数の場合

重回帰モデル：

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i \ , \quad \varepsilon_i \sim N(0, \sigma^2)$$

(参考) 最小二乗法の計算

残差の2乗和(E)が最小となる回帰係数(β_0,β_1)を推定する
推定結果を$\widehat{\beta_0}$, $\widehat{\beta_1}$と表記

$$E = \sum_i (y_i - (\beta_0 + \beta_1 x_i))^2 \qquad i = 1,2,\dots,n$$

β_1でEを微分。0とおく

$$\frac{\partial E}{\partial \beta_1} = -2\sum_i \big(y_i - (\beta_0 + \beta_1 x_i)\big) x_i = 0$$

$$\Leftrightarrow \quad \sum_i x_i y_i = \beta_0 \sum_i x_i + \beta_1 \sum_i {x_i}^2 \ \cdots \ ①$$

β_0でEを微分。0とおく

$$\frac{\partial E}{\partial \beta_0} = -2\sum_i \big(y_i - (\beta_0 + \beta_1 x_i)\big) = 0$$

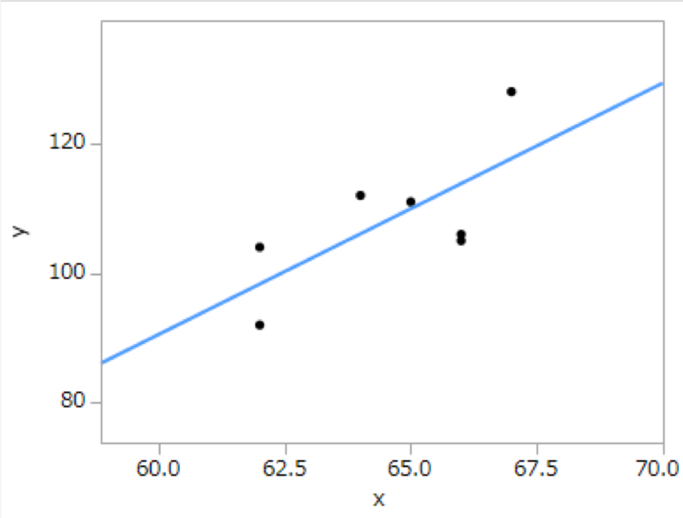
$$\Leftrightarrow \quad \sum_i y_i = n\beta_0 + \beta_1 \sum_i x_i \ \cdots \ ②$$

①,②をβ_0,β_1に関して解き

$$\widehat{\beta_1} = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i {x_i}^2 - (\sum_i x_i)^2}$$

$$\widehat{\beta_0} = \frac{\sum_i y_i}{n} - \frac{\sum_i x_i}{n} \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i {x_i}^2 - \left(\sum_i x_i\right)^2} = \bar{y} - \bar{x}\widehat{\beta_1}$$

結果の確認（1）



結果

```
Call:
lm(formula = y ~ x, data = df)

Residuals:
     1      2      3      4      5      6      7
-6.325  5.928 10.307  1.054 -8.819  5.675 -7.819

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -141.831    111.557  -1.271   0.2595
x              3.873      1.727   2.243   0.0749 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.41 on 5 degrees of freedom
Multiple R-squared:  0.5015,	Adjusted R-squared:  0.4019
F-statistic: 5.031 on 1 and 5 DF,  p-value: 0.07493
```

”Estimate”の個所
$\widehat{\beta_0}$=-141.831
$\widehat{\beta_1}$=3.873
よって、$\hat{y}_i = -141.831 + 3.873x_i$

“Std. Error”の個所
回帰係数の推定量($\widehat{\beta_0}$, $\widehat{\beta_1}$)の標準誤差（推定量のばらつき）

“Estimate”割る”Std. Error”が”t value”となり、自由度が「観測数－回帰係数の数」のt分布に従う
この性質を利用して検定を行った結果が、”Pr(>|t|)”の個所

H0: $\beta_0 = 0$、H1: $\beta_0 \neq 0$に対応するp値:0.2595
H0: $\beta_1 = 0$、H1: $\beta_1 \neq 0$に対応するp値:0.0749

通常、傾き(β_1)の検定に興味がある

信頼区間の計算には、confint()関数に結果を渡す

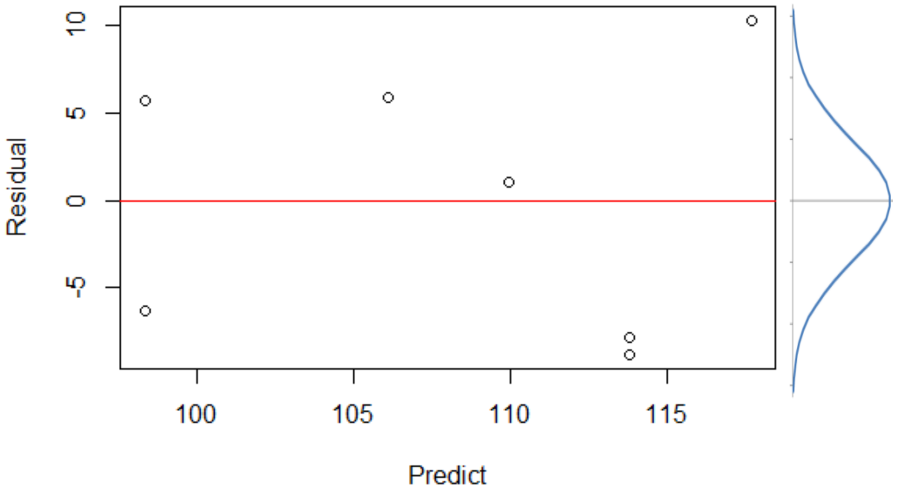
“Residual Standard Error”: $\hat{\sigma}$=8.41

結果の確認（2）

モデルの前提条件の確認

- 残差は正規分布に従う
- 残差の分散(σ^2)はすべてのデータで共通

残差プロット



横軸が予測値($\hat{y}_i$)
縦軸が残差(e_i)

残差が正規分布に従っているか？
予測値の大小によって残差の分布に偏りがないか？

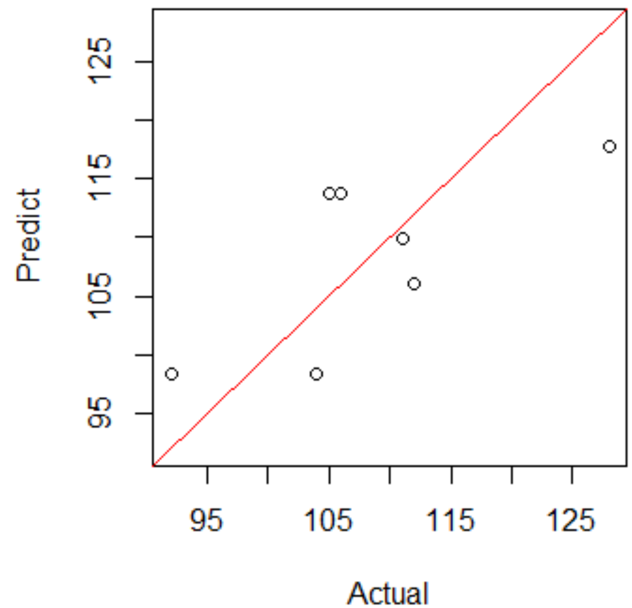
当てはまりの精度の確認

決定係数によって、モデルのデータへの当てはまりの良さが評価できる

$R^2 = 1 - {\sum_i e_i{}^2}/{\sum_i y_i{}^2}$

ただし、説明変数を加えれば加えるほど決定係数は上昇するので注意

予測値,実測値プロット



横軸が実測値(y_i)
縦軸が予測値($\hat{y}_i$)

斜め45度線にデータ点が近いほど当てはまりが高い

予測と信頼区間

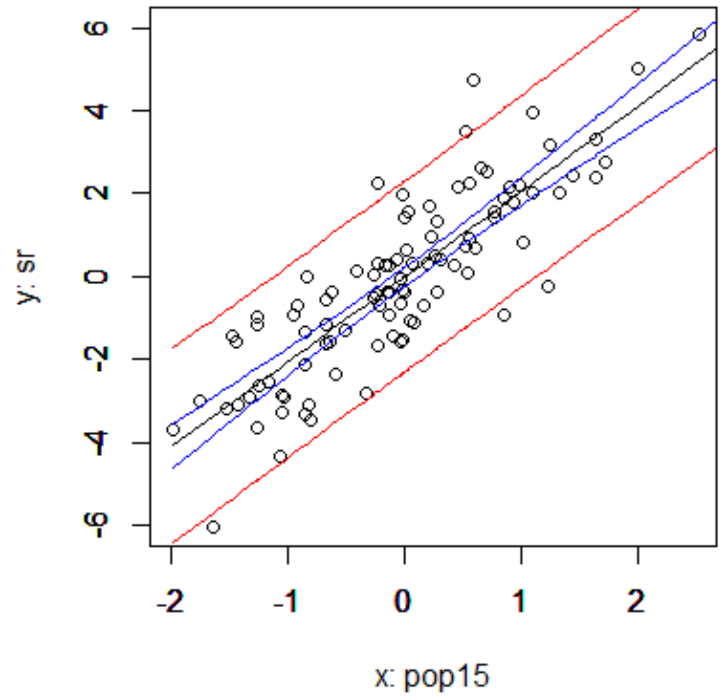
モデルを推定することにより、予測が実施できる
推定したモデルが $\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$ の際（$\widehat{\beta_0}$と$\widehat{\beta_1}$の値が与えられているので）、x_iを代入すると$\hat{y}_i$が計算できる

回帰分析を実行した際、2種類の信頼区間を計算することができる

回帰の信頼区間
- 回帰直線の信頼区間

予測の信頼区間
- 将来観測される予測値が入るであろうと考えられる範囲
- プロットから約5%が区間外なのが読み取れる（データは100個）
- 観測値の誤差(ε_i)を考慮する分、幅が広くなる

100個のデータでのサンプル



— 回帰直線
— 回帰の(95%)信頼区間
— 予測の(95%)信頼区間

演習【Day4-Exercise1】回帰モデル

- 回帰分析をlm()関数で実行する
- 結果の確認方法に慣れる
- プロットであてはめ結果を確認する

演習の解説

サンプルデータ – LifeCycleSavings

sr
- aggregate personal savings

pop15
- % of population under 15

pop75
- % of population over 75

dpi
- real per-capita disposable income

ddpi
- % growth rate of dpi

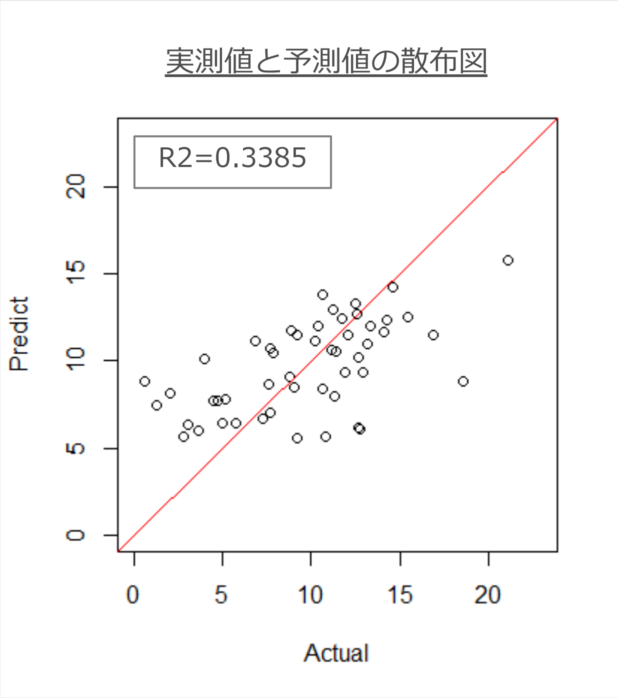
	sr	pop15	pop75	dpi	ddpi
Australia	11.43	29.35	2.87	2329.68	2.87
Austria	12.07	23.32	4.41	1507.99	3.93
Belgium	13.17	23.80	4.43	2108.47	3.82
Bolivia	5.75	41.89	1.67	189.13	0.22
Brazil	12.88	42.19	0.83	728.47	4.56
Canada	8.79	31.72	2.85	2982.88	2.43
Chile	0.60	39.74	1.34	662.86	2.67
China	11.90	44.75	0.67	289.52	6.51
Colombia	4.98	46.64	1.06	276.65	3.08

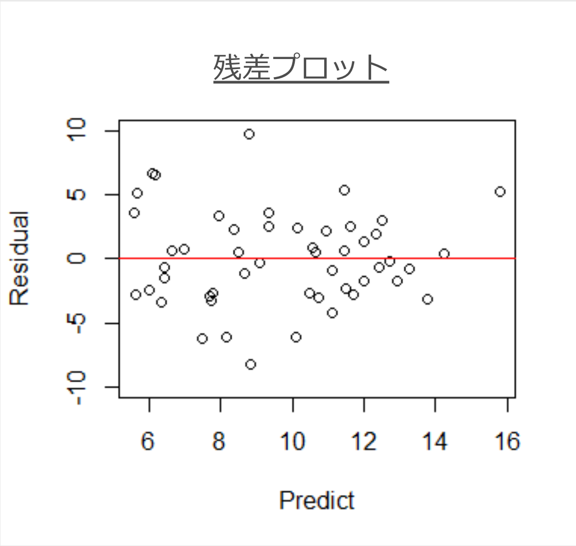
- ✓ 単回帰と重回帰の結果を比べる
- ✓ 有意な変数は？推定値の大きさは？信頼区間は？
- ✓ あてはまりの結果を可視化（実測値と予測値の散布図、残差プロット）

演習の解説

重回帰の結果

	推定値	95%信頼区間(下側)	95%信頼区間(上側)	p値
切片	28.566	13.753	43.379	0.0003
pop15	-0.461	-0.753	-0.170	0.0026
pop75	-1.691	-3.874	0.491	0.1255
dpi	0.000	-0.002	0.002	0.7192
ddpi	0.410	0.015	0.805	0.0425





最尤推定

回帰モデルは、 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$であった
目的変数(y_i)は平均($\beta_0 + \beta_1 x_i$)、分散(σ^2)の正規分布から生成される確率変数と言える $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

x	y
x_1	y_1
x_2	y_2
:	:
:	:
x_n	y_n

← $N(\beta_0 + \beta_1 x_1, \sigma^2)$ から生成
← $N(\beta_0 + \beta_1 x_2, \sigma^2)$ から生成
⋮

各yが観測される確率は

$$P(y_1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-{(y_1-(\beta_0+\beta_1 x_1))^2}/{2\sigma^2}}$$

$$P(y_2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-{(y_2-(\beta_0+\beta_1 x_2))^2}/{2\sigma^2}}$$

⋮

すべてのyが同時に観測される確率は

$$L = P(y_1, y_2, \dots, y_n) = \prod_i P(y_i) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-{(y_i-(\beta_0+\beta_1 x_i))^2}/{2\sigma^2}}$$

これを**尤度関数**と呼ぶ
この尤度関数を最大にするようモデルパラメータ(β_0, β_1, σ)を推定する手法を**最尤法**と呼ぶ
（目的変数の仮定する確率分布を用い、データが生成される確率が最大となるようなパラメータを推定）

与えられたデータの元、パラメータを動かして尤度が最大になる値を探す
言い換えると、そのパラメータの元、そのデータが最も観測されやすい、と言える

最尤推定と最小二乗法

回帰モデルにおいて、最尤推定と最小二乗推定が一致することを示す

$$L = \prod_i \frac{1}{\sqrt{2\pi\sigma}} e^{-(y_i-(\beta_0+\beta_1 x_i)^2 / 2\sigma^2} = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_i (y_i-(\beta_0+\beta_1 x_i))^2}$$

対数を取り、積から和へ変換。**対数尤度関数**と呼ばれる

$$l = ln(L) = n\, ln\left(\frac{1}{\sqrt{2\pi\sigma}}\right) - \frac{1}{2\sigma^2}\sum_i (y_i - (\beta_0 + \beta_1 x_i))^2$$

Lの最大化もlの最大化も意味的には同じ

lの最大化 = 右辺の第二項の最小化

よって、 $\sum_i (y_i - (\beta_0 + \beta_1 x_i))^2$の最小化

最小二乗法に一致

ダミー変数

ダミー変数(Dummy Variable, Indicator Variable)とは、カテゴリカル変数を0/1で表示したもの
Rなどの統計ツールの中では、実際はカテゴリカルデータはダミー変数に変換されて計算が行われている

2水準の例

Gender
F
M

Gender_M	Gender_F
0	1
1	0

3水準の例

Color
Red
Green
Yellow

Color_Red	Color_Green	Color_Yellow
1	0	0
0	1	0
0	0	1

回帰分析においては、各カテゴリカル変数に対し「水準数-1」のダミー変数を投入する
（水準すべての変数を投入しても、同じ情報を持っている変数どうしなので意味がない）

ダミー変数に変換してから分析を実施すると、効果を把握したい水準の結果が取得できる
※ factor()関数のlevels引数に順序を指定するやり方もある

演習【Day4-Exercise2】
ダミー変数

- dummiesパッケージの利用
- ダミー変数化した場合としなかった場合の結果の確認
- Factor型は裏ではダミー変数として分析が実行されていることを理解する

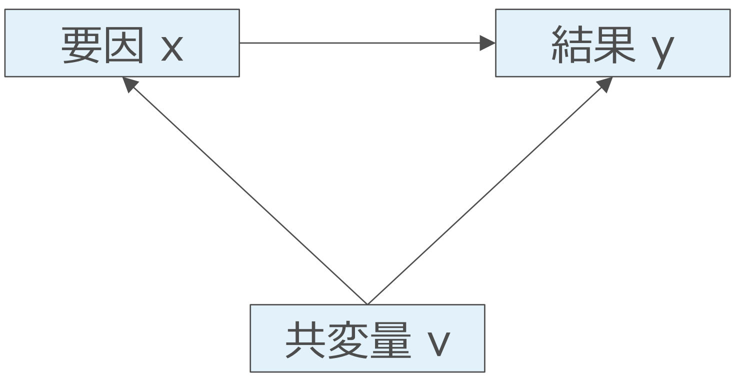
共分散分析

ある結果に与える要因の大きさを分析したい
ただし、結果と要因ともに関連のある第三の要因が存在する

結果と要因ともに関連する第三の要因を**共変量(Covariate)**と呼ぶ

共変量の影響を除去して、結果と要因の関連を分析を行いたい

共分散分析(ANCOVA:ANalysis of COVariance)は、要因がカテゴリカル変数、共変量が連続変数の場合、共変量の影響を除去して、結果と要因の関連を分析する回帰モデルを用いた手法



例1）
要因：　投薬
結果：　病状の変化
共変量：年齢

例2）
要因：　広告媒体
結果：　売り上げ
共変量：広告費用

モデル：

$$y_i = \mu + \beta_x x_i + \beta_v v_i + \varepsilon_i \ , \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$x_i = 0, 1$$

要因xの結果yへの影響を検定（H0： $\beta_x = 0$）

演習【Day4-Exercise3】共分散分析

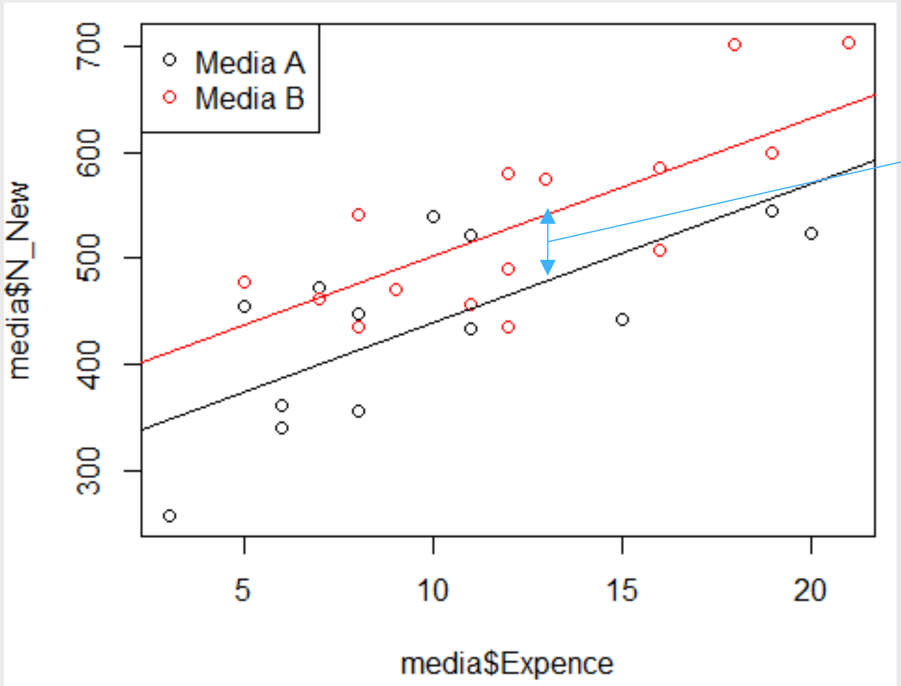
- 共分散分析の実施

演習の解説

広告媒体(Media)AとBの違いによる入会者数(N_New)への寄与を分析したい
広告媒体とも入会者数とも関連のある共変量である広告費用(Expense)を考慮して分析を実施する
MediaEffect.csv

広告媒体(Media) … カテゴリカル変数
広告費用(Expense) … 数値変数

- 共変量を考慮しなかった場合の媒体効果？
- 共変量を考慮した場合の媒体効果？



2直線の差（切片の差）が費用を考慮したうえでの媒体の効果

多重共線性

多重共線性(Multicollinearity)は、相関が高い説明変数どうし（正確には説明変数間に線形結合の関係がある場合）を用いて重回帰を実行した場合に起こる、モデルが不安定になる現象

特に、推定した回帰係数が不安定になり、信頼のおける解釈ができなくなる

x_1とx_2の相関が高い（ほぼ同じ）場合、下の3つの異なるモデルは同じ予測値yを出力する

$$y = 1x_1 + 1x_2 \approx 0.5x_1 + 1.5x_2 \approx -1x_1 + 3x_2$$

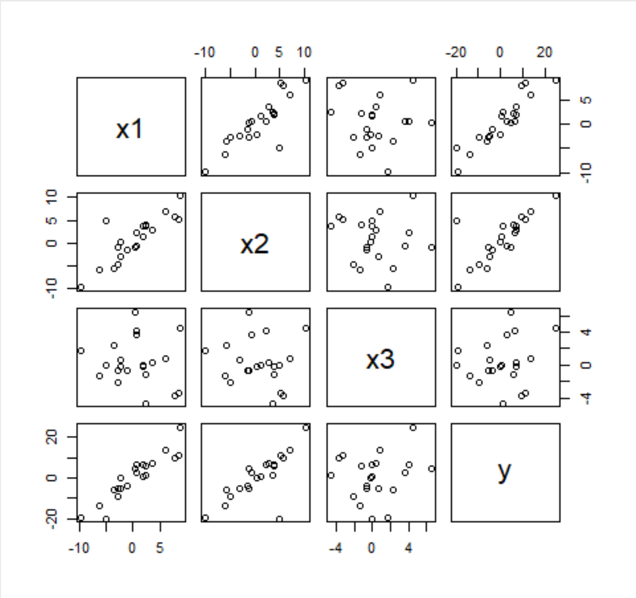
対処方法

1. 相関関係の高い説明変数を含まない
   - 目的変数とより相関の高い説明変数を残す
   - ドメイン知識より残す変数を判断
   - 変数選択法（後述）
2. 手法で対処する
   - 主成分回帰
   - PLS
   - 正則化

演習【Day4-Exercise4】多重共線性

- 多重共線性の理解

演習の解説



Multicollinearity.csv

yとx1,x2間の線形の関係が強そう
また、x1とx2間にも強い線形の関係が見られる

利用変数	x1,x2,x3	x1,x2	x1,x3	x2,x3	x1	x2	x3
	Estimate						
(Intercept)	0.6964	-0.3178	-0.749	-1.1698	-0.4006	-0.8569	0.299
x1	2.2762	2.1993	2.1072		2.0386		
x2	-0.1972	-0.1877		1.6986		1.6515	
x3	1.0483		1.0467	0.9021			0.6238
R2	0.9296	0.8568	0.9272	0.6286	0.8547	0.5744	0.02612
AIC	112.8278	125.7331	111.5096	145.7426	124.0397	146.601	163.9853

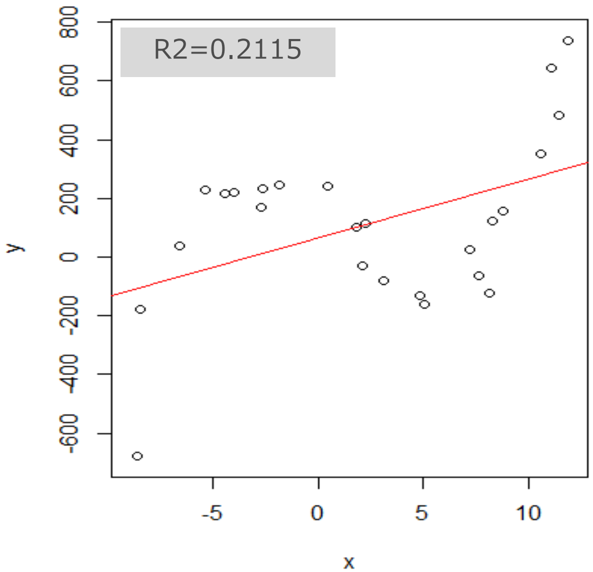
値が大きく変わる

モデルの複雑性

モデルは複雑であるべきか、シンプルであるべきか

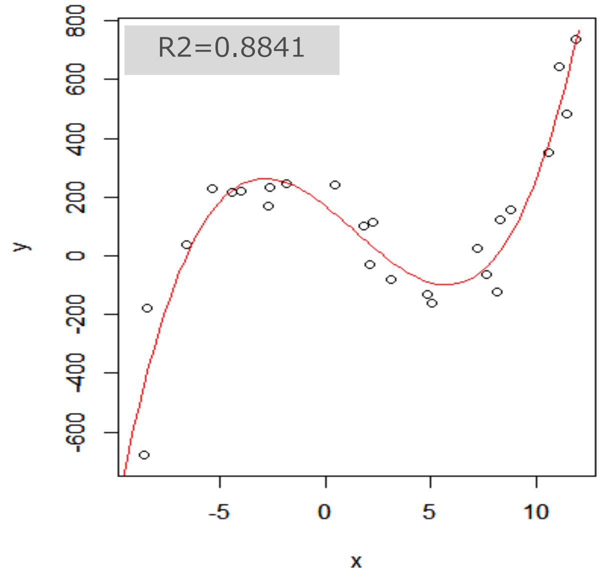
回帰モデルの場合、
- 沢山の説明変数 ⇒ 複雑なモデル
- 少数の説明変数 ⇒ シンプルなモデル

多項式モデルを用いた例

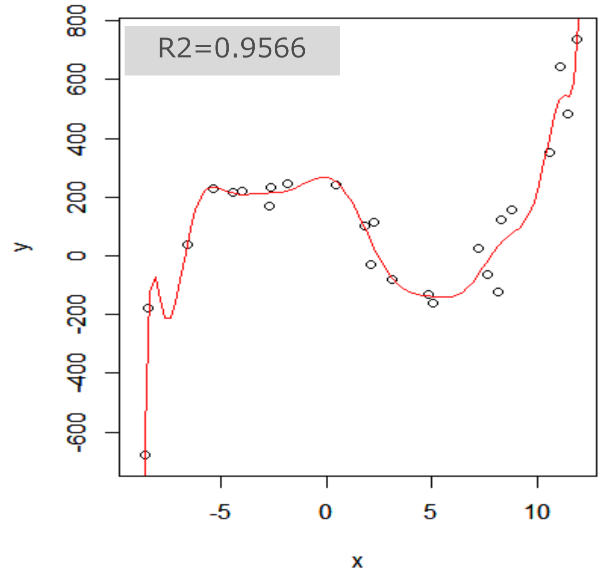


$y = \widehat{\beta_0} + \widehat{\beta_1}x$

Under Fitting



$y = \widehat{\beta_0} + \widehat{\beta_1}x + \widehat{\beta_2}x^2 + \widehat{\beta_3}x^3$



$y = \widehat{\beta_0} + \widehat{\beta_1}x + \widehat{\beta_2}x^2 + \cdots + \widehat{\beta_{15}}x^{15}$

過剰適合(Over Fitting)
- 既存のデータには高く適合しているように見えるが、未知のデータには対応できない

AIC

AIC(Akaike's Information Criteria)は、統計モデルの良さを評価するための指標
「モデルの複雑さと、データとの適合度とのバランスを取る」ために使用
統計モデルを作成する場合
- パラメータの数や次数を増やせば増やすほど、その測定データとの適合度を高めることができる
- その反面、ノイズなどの偶発的な（測定対象の構造と無関係な）変動にも無理にあわせてしまうため、未知のデータには合わなくなる（過適合問題、Overfitting）

$$AIC = \underbrace{-2\ ln(L)}_{\text{適合度の高さ}} + \underbrace{2k}_{\text{モデルの複雑性によるペナルティ}}$$

適合度の高さ
適合度が高いと小さくなる

モデルの複雑性による
ペナルティ

$ln(L)$：尤度Lの自然対数
k ：モデルのパラメータ数（回帰分析の場合、説明変数の数+2）

AICが小さければ小さいほど良いモデル

(Wiki) https://ja.wikipedia.org/wiki/%E8%B5%A4%E6%B1%A0%E6%83%85%E5%A0%B1%E9%87%8F%E8%A6%8F%E6%BA%96

変数選択法

過剰適合や多重共線性でみたように、説明変数の数が多すぎるとモデル自体の信頼性や解釈において不都合が生じる
変数選択法により、有用な少数の説明変数に絞り、実用的なモデルの作成を目指す

変数選択法の種類
- 変数増加法
  - 目的変数と関連性の高い説明変数を順次取り込む
- 変数減少法
  - 最初に全説明変数を取り込み、目的変数と関連性の低い説明変数を順次削除
- 変数増減法（ステップワイズ法）
  - 増加法と減少法を組み合わせたアプローチ。説明変数を取り込んだ後、重要度が入れ替わり重要でなくなった説明変数を削除。それを繰り返しながら説明変数を取り込んで行く

変数を選択する際の指標
- AIC
- 各回帰係数の検定結果（p値）

演習【Day4-Exercise5】
変数選択法

- AICを用いた変数選択法
- stepAIC()関数（step()関数でも実行可能）

演習の解説

direction="both"（変数増減法）とした場合の結果

```
> resAIC <- stepAIC(unicef_lm_full, direction="both")
Start:  AIC=876.77
Child.Mortality ~ Literacy.Fem + Literacy.Ad + Drinking.Water +
    Polio.Vacc + Tetanus.Vacc.Preg + Urban.Pop + Foreign.Aid

                    Df Sum of Sq    RSS    AIC
- Literacy.Ad        1     438.3 149125 875.13
- Tetanus.Vacc.Preg  1     503.3 149190 875.18
<none>                           148686 876.77
- Foreign.Aid        1    3522.3 152209 877.60
- Urban.Pop          1    5839.4 154526 879.43
- Literacy.Fem       1    8980.6 157667 881.87
- Polio.Vacc         1   12088.2 160774 884.23
- Drinking.Water     1   24768.6 173455 893.41

Step:  AIC=875.13
Child.Mortality ~ Literacy.Fem + Drinking.Water + Polio.Vacc +
    Tetanus.Vacc.Preg + Urban.Pop + Foreign.Aid

                    Df Sum of Sq    RSS    AIC
- Tetanus.Vacc.Preg  1       383 149508 873.44
<none>                           149125 875.13
- Foreign.Aid        1      3651 152776 876.05
+ Literacy.Ad        1       438 148686 876.77
- Urban.Pop          1      6158 155283 878.02
- Polio.Vacc         1     14553 163678 884.39
- Drinking.Water     1     24644 173769 891.63
- Literacy.Fem       1     57886 207011 912.81

Step:  AIC=873.44
Child.Mortality ~ Literacy.Fem + Drinking.Water + Polio.Vacc +
    Urban.Pop + Foreign.Aid

                    Df Sum of Sq    RSS    AIC
<none>                           149508 873.44
- Foreign.Aid        1      3687 153195 874.38
+ Tetanus.Vacc.Preg  1       383 149125 875.13
+ Literacy.Ad        1       318 149190 875.18
- Urban.Pop          1      5797 155304 876.04
- Polio.Vacc         1     19684 169192 886.40
- Drinking.Water     1     25329 174837 890.37
- Literacy.Fem       1     60926 210434 912.80
```

AICの小さい順にソート

各変数を抜いた場合(-)、加えた場合(+)

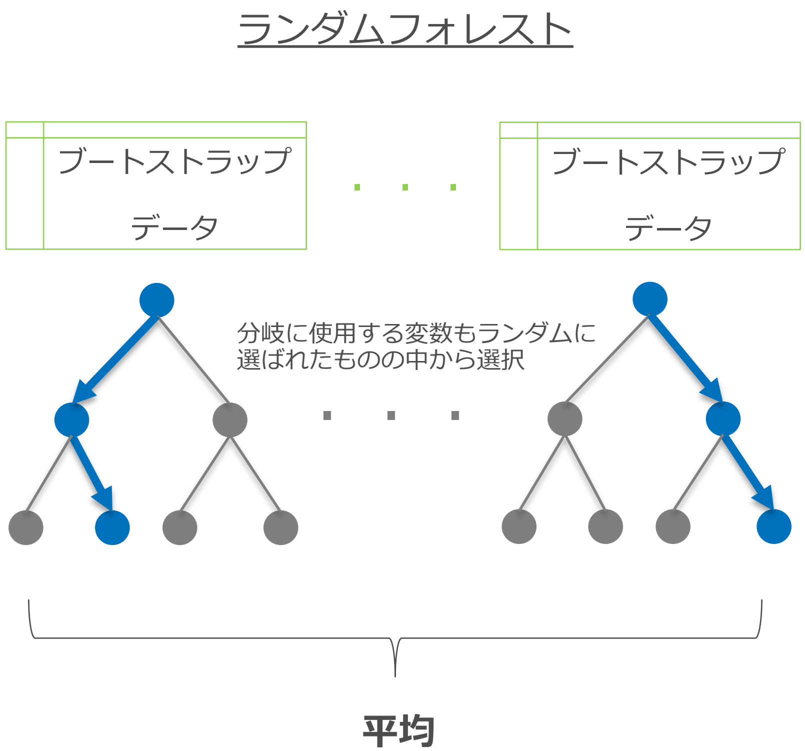
<none>は現時点でのモデル

現時点でのモデルのAICが最小になるので終了

機械学習アプローチによる変数のスクリーニング

大量の変数があり、そもそも何が効いているか事前知識があまりない場合、機械学習/データマイニングの手法を用いて、少数の変数へ絞り込みを行う場合がある

ここでは、**ランダムフォレスト(Random Forest)**の変数重要度を用いて、変数のスクリーニングを実施する方法を紹介



ランダムフォレストは、重複サンプリング(ブートストラップサンプリング)により作成した異なった各データセットに対し決定木をあてはめるアンサンブル手法

- 決定木モデルの利用（非線形な関係も捉えれる）
- 評価する説明変数のランダムな選択

といった理由から、説明変数のスクリーニングに利用するのに適している

IncNodePurityは、寄与した説明変数の貢献度を示す指標。値が大きいほど大きく貢献

	IncNodePurity
Literacy.Fem	125145.71
Literacy.Ad	97628.93
Drinking.Water	90653.86
Polio.Vacc	87741.95
Tetanus.Vacc.Preg	31203.61
Urban.Pop	52312.95
Foreign.Aid	103437.74

演習【Day4-Exercise6】
ランダムフォレストによる変数のスクリーニング

- randomForestパッケージrandomForest()関数