L03 線形回帰モデル - 単回帰

樋口さぶろお　https://hig3.net

龍谷大学 先端理工学部 数理・情報科学課程

多変量解析☆演習　L03(2021-10-14 Thu)

最終更新: Time-stamp: "2021-10-13 Wed 15:41 JST hig"

今日の目標

- 線形回帰モデル, 単回帰とは何か説明できる
- 与えられたデータを単回帰するのが適切か判断できる, 結果を評価できる
- 与えられたデータに対し単回帰を Jupyter Notebook で

L03-Q1

Quiz 解答:2 つの連続型確率変数の和の分布

$$f_{\mathrm{S}}(s) = \int_{-\infty}^{+\infty} f_{\mathrm{X}}(x) f_{\mathrm{X}}(s-x)\,\mathrm{d}x = \begin{cases} \frac{1}{36}(s-2) & (2 \le s < 8) \\ \frac{1}{36}(14-s) & (8 \le s < 14) \\ 0 & (\text{他}) \end{cases}$$

L03-Q2

Quiz 解答:2 つの連続型確率変数の和の分布

$$f_{\mathrm{S}}(s) = \int_{-\infty}^{+\infty} f(x, s-x)\,\mathrm{d}x = \begin{cases} \frac{1}{18}s & (0 \le s < 6) \\ 0 & (\text{他}) \end{cases}$$

ここまで来たよ

2 確率変数の和

2 線形回帰モデル - 単回帰
- 線形回帰モデルとは
- 回帰分析の手順

線形回帰 linear regression (単回帰) モデルとは

データ分析 (2020)L08-11 | 確率統計 I(2021)L12 | 岩薙林 確率・統計 §9 | 永田棟方 多変量解析法入門 §4

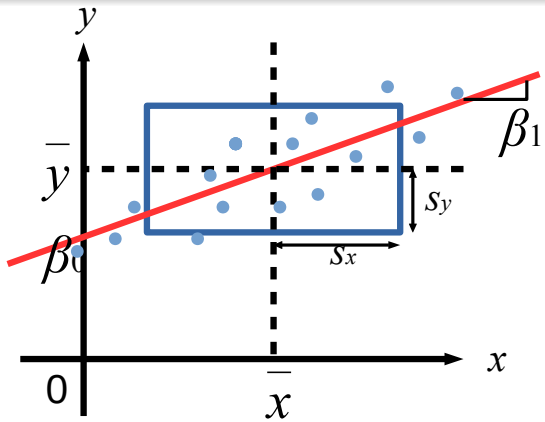
このドーナツ製造機の作るドーナツの重さ Y_i は, 温度 x_i から決まる $(i = 1, \ldots, n)$ らしい. 次の線形回帰モデル (単回帰) を仮定する.

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i, \qquad \epsilon_i \sim \mathrm{N}(0, \sigma^2), \quad \text{(独立同分布)}$$

Y, ϵ: 連続型確率定数, β_0, β_1:回帰係数, $\sigma > 0$:定数.

Y: 目的変数 (従属変数) 確率変数
x: 説明変数 (独立変数) 確率変数でない
データ $(x_i, y_i)(i = 1, 2, \ldots, n)$ から β_0, β_1 を推定するのが回帰分析. 残差 (residual)

岩薙林 確率・統計 p.205



$$e_i = y_i - (\beta_0 + \beta_1 \cdot x_i)$$

の残差平方和 岩薙林 確率・統計 p.206 $S(\beta_0, \beta_1)$ =

β_0, β_1: 本当の値
$\hat{\beta}_0, \hat{\beta}_1$: データ $(x_i, y_i)(i = 1, 2, \ldots, n)$ から求まる推定値. Python がやってくれる.

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}, \qquad y = \frac{s_{xy}}{s_{xx}}x + \overline{y} - \frac{s_{xy}}{s_{xx}}\overline{x}$$

$$\hat{\beta}_0 = \overline{y} - \frac{s_{xy}}{s_{xx}}\overline{x} \qquad y - \overline{y} = \frac{s_{xy}}{s_{xx}}(x - \overline{x})$$

ここで,

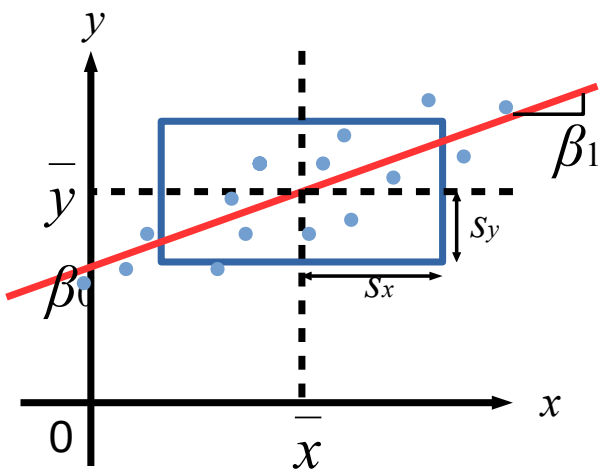
$$\overline{x} = \frac{1}{n}\sum_i x_i$$ 平均値ぽい

$$\overline{y} = \frac{1}{n}\sum_i y_i$$ 平均値ぽい

$$s_{xy} = \frac{1}{n}\sum_i x_i y_i - \overline{x}\cdot\overline{y} \overset{\text{岩薩林 確率・統計 定理 1.5}}{=} \frac{1}{n}\sum_i (x_i - \overline{x})(y_i - \overline{y})$$ 共分散ぽい

$$s_{xx} = \frac{1}{n}\sum_{i}^{n} x_i^2 - \overline{x}^2 \overset{\text{岩薩林 確率・統計 定理 1.2}}{=} \frac{1}{n}\sum (x_i - \overline{x})^2$$ 分散ぽい

β_i の推定は，母平均値の推定と似ている/違う



線形回帰モデルの中の単回帰の位置づけ

- 一般化線形モデル
  - 線形回帰モデル
    - ★ 単回帰 永田棟方 多変量解析法入門 §4
    - ★ 重回帰 永田棟方 多変量解析法入門 §5
  - ロジスティック回帰モデル
  - ⋮

ここまで来たよ

2 確率変数の和

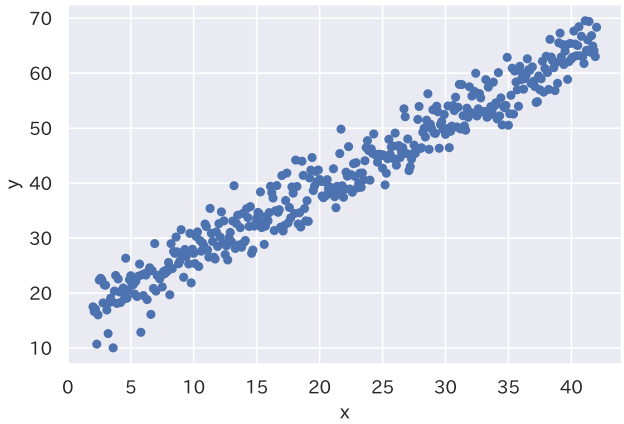
2 線形回帰モデル - 単回帰
- 線形回帰モデルとは
- 回帰分析の手順

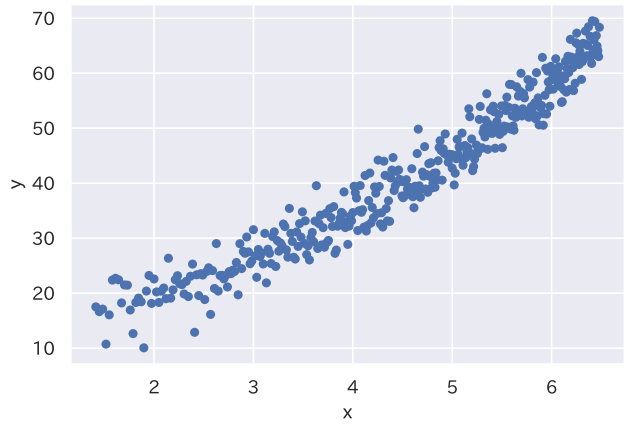
回帰分析の手順

1. やっていいのか? 現象・データは線形回帰に向いている (ようにできる) か?
2. やる ソフトウェアで, 推定値 $\hat{\beta}_0, \hat{\beta}_1$ などを推定
3. 本当にやってよかったか? ソフトウェアの出力で, 線形回帰がよくあてはまっているか, どう改善できるか検討
4. 何がわかったか? 回帰係数の信頼区間, 仮説検定, 予測値

データは線形回帰に向いてる (ようにできる) か?

1. 目的変数は連続値か?
2. 説明変数は間隔尺度か?
3. 可視化してみよう (散布図など)





1. 外れ値 $\rightsquigarrow$ 取り除く
2. 直線じゃない形 $\rightsquigarrow$ 説明変数を変換する
3. 無関係っぽい $\rightsquigarrow$ 別の説明変数を探す

ソフトウェアで推定，何がわかったか?

```
[ ]  1  formula='height ~ weight'  # formula 表記. ~ の左辺を目的変数，右辺を説明変数として線形回
     2  result=smf.ols(formula, body).fit() # ols = 普通の最小二乗法で fit せよ
     3  result.summary() # result に蓄えられた結果を取り出す
```

目的変数は height

決定係数R
(Adjusted 自由度調整済)

OLS Regression Results

Dep. Variable:	height	R-squared:	0.905
Model:	OLS	Adj. R-squared:	0.904
Method:	Least Squares	F-statistic:	935.3
Date:	Sat, 09 Oct 2021	Prob (F-statistic):	6.31e-52
Time:	00:59:55	Log-Likelihood:	-255.30
No. Observations:	100	AIC:	514.6
Df Residuals:	98	BIC:	519.8
Df Model:	1		
Covariance Type:	nonrobust		

データの個数
自由度

F統計量
F検定のp値
(有意確率)

t検定のp値
(有意確率)

95%信頼区間

値

標準誤差

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	129.4555	1.122	115.412	0.000	127.230	131.681
weight	0.8233	0.027	30.582	0.000	0.770	0.877

切片β0
説明変数weightの
係数β1

Omnibus:	1.147	Durbin-Watson:	2.218
Prob(Omnibus):	0.564	Jarque-Bera (JB):	1.053
Skew:	-0.063	Prob(JB):	0.591
Kurtosis:	2.514	Cond. No.	149.

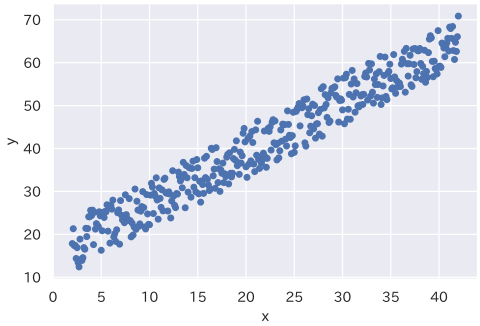
線形回帰はよくあてはまっているか? I

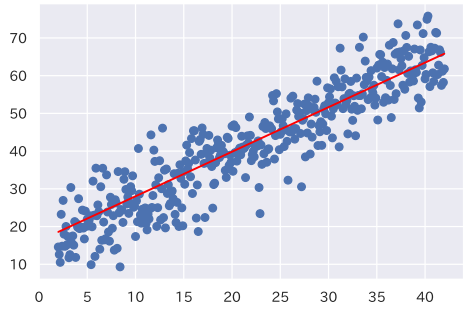
$$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i) \sim \mathrm{N}(0, \sigma^2).$$
残差 $e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).$

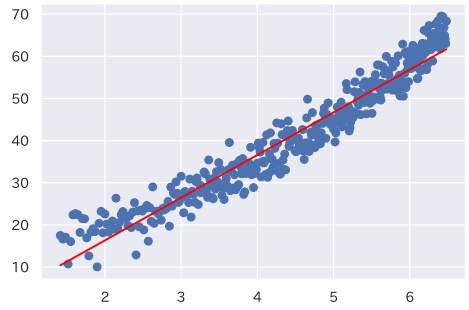
1. 決定係数 $R^2 = 1 - \dfrac{\sum \text{残差}^2}{\text{水平な直線からのずれ}^2}$ が 1 に近いか?

確率統計 I(2021)L12

データ分析 (2020)L11



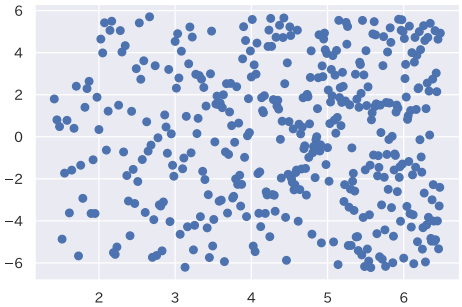
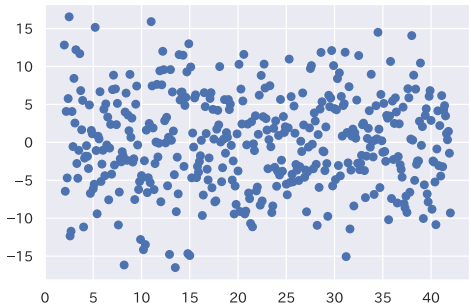
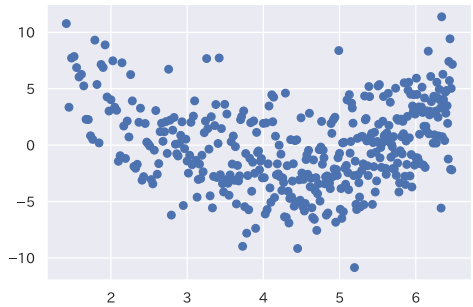




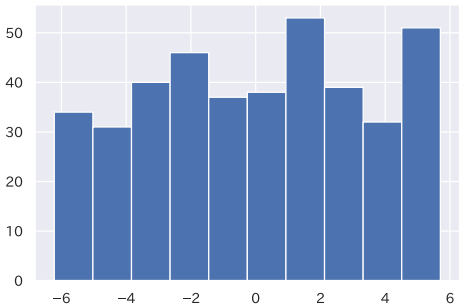
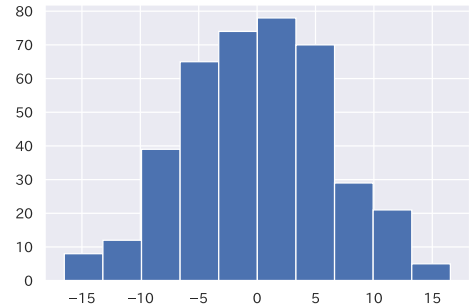
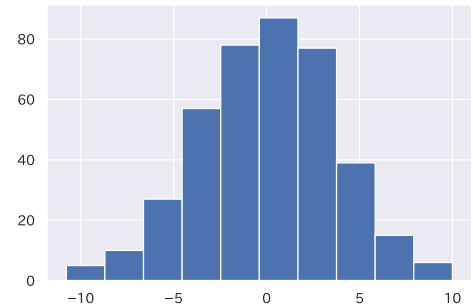
$R^2 = 0.944\ 0.828\ 0.932$

線形回帰はよくあてはまっているか? II

2. (x, e) の散布図を描く. 残差 $e_i = y_i - \hat{y}_i$ は測定値の予測値からのずれ.

3. 残差 e のヒストグラムを描く.

4. 区間推定 (β_0, β_1 の信頼区間は?)

線形回帰はよくあてはまっているか? III

5. 検定 帰無仮説 0: $\beta_0 = 0 \Leftrightarrow$ 切片はいらない
6. 検定 帰無仮説 1: $\beta_1 = 0 \Leftrightarrow Y$ は x に依存しない
7. 機械学習のりなら, 分けておいたテストデータで検証 (上はすべて, 訓練データでの検証)

機械学習としての線形回帰モデル

これは教師あり学習の一種

出力が連続的な値, 予測器が 1 次関数.

n 個の訓練データ (x_i, y_i) の $\overset{\text{学習}}{\to}$ パラメタの推定値 $\hat{\beta}_0, \hat{\beta}_1 \to$ 予測器 $\hat{y} = f(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

性能評価

機械学習のりなら, 訓練データと分けておいたテストデータを使うところだが, 統計学のりでは, 数学的仮定に基づき, 決定係数 R^2 で評価 (訓練データ=テストデータとしてしまっているようなもの)