確率と統計　第 6 回

線形回帰への入門

5 月 25 日

Table of contents

線形回帰＝直線への当てはめ

線形回帰＝直線への当てはめ

2 つの変数 X と Y があるとき，これらの相関関係を以下の線形関係:

線形回帰＝直線への当てはめ

2 つの変数 X と Y があるとき，これらの相関関係を以下の線形関係:

$$y = \beta_0 + \beta_1 x + \epsilon$$

線形回帰＝直線への当てはめ

2 つの変数 X と Y があるとき，これらの相関関係を以下の線形関係:

$$y = \beta_0 + \beta_1 x + \epsilon$$

で近似する**線形回帰**は基本的かつ強力な統計的方法だ．

線形回帰＝直線への当てはめ

このような直線を x と y の散布図上に描くために，この直線の切片 β_0 と傾き β_1 (さらに誤差項 ϵ) を観測されたデータ $\{(x_i, y_i)\}$ から推定する必要がある．

線形回帰＝直線への当てはめ

このような直線を x と y の散布図上に描くために, この直線の切片 β_0 と傾き β_1 (さらに誤差項 ϵ) を観測されたデータ $\{(x_i, y_i)\}$ から推定する必要がある.

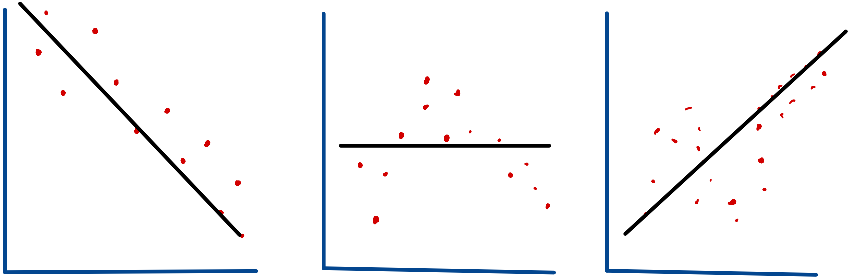
これらの定数 β_0, β_1 を母数 (パラメータ)と呼び, パラメータの推定値をそれぞれ b_0, b_1 と表すことにする.

散布図と線形回帰

つまり線形回帰では以下のような点群で表されるデータを近似する一本の直線を引くことを考える.

散布図と線形回帰

つまり線形回帰では以下のような点群で表されるデータを近似する一本の直線を引くことを考える.



ボッサムの体長と頭の長さ

ボッサムの体長と頭の長さ

オーストラリアに棲息するフクロギツネ (ボッサム) の体長 (頭からしっぽの先までの長さ) と頭の長さの関係を考える.

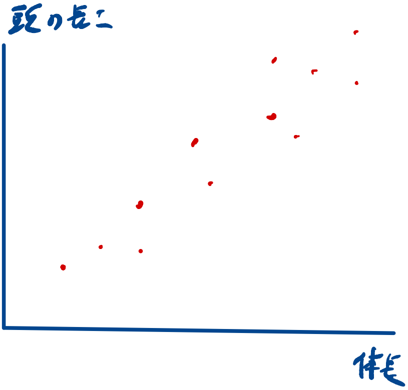
ボッサムの体長と頭の長さ

オーストラリアに棲息するフクロギツネ (ボッサム) の体長 (頭からしっぽの先までの長さ) と頭の長さの関係を考える.

体長を X(cm), 頭の長さを Y(mm) とし実際の観測値を以下のデータ行列と散布図にまとめる.

ボッサムの体長と頭の長さ

	x	y
	体長 (cm)	頭長 (mm)
1	69.3	71.0
2	73.0	75.7
⋮	⋮	⋮



残差

この散布図からそれらしい曲線

$$\hat{y} = 41 + 0.59x$$

を勘で引いたとする.

残差

この散布図からそれらしい曲線

$$\hat{y} = 41 + 0.59x$$

を勘で引いたとする.ここで $\hat{y}$ は予想値を表すので y の上にハットを付けている.

残差

この散布図からそれらしい曲線

$$\hat{y} = 41 + 0.59x$$

を勘で引いたとする.ここで $\hat{y}$ は予想値を表すので y の上にハットを付けている.

例えば観測値 $(x, y) = (77.0, 85.3)$ にたいして, その予測値 $\hat{y}$ は

残差

この散布図からそれらしい曲線

$$\hat{y} = 41 + 0.59x$$

を勘で引いたとする. ここで $\hat{y}$ は予想値を表すので y の上にハットを付けている.

例えば観測値 $(x, y) = (77.0, 85.3)$ にたいして, その予測値 $\hat{y}$ は

$$41 + 0.59 \times 77.0 = 86.4$$

となり, 実測値 y との差は

残差

この散布図からそれらしい曲線

$$\hat{y} = 41 + 0.59x$$

を勘で引いたとする.ここで $\hat{y}$ は予想値を表すので y の上にハットを付けている.

例えば観測値 $(x, y) = (77.0, 85.3)$ にたいして, その予測値 $\hat{y}$ は

$$41 + 0.59 \times 77.0 = 86.4$$

となり, 実測値 y との差は

$$y - \hat{y} = 85.3 - 86.4 = -1.1$$

となる.

残差

一般に n 個のデータの集まり $\{(x_i, y_i)\}$ を近似する直線

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

が与えられているとき，

残差

一般に n 個のデータの集まり $\{(x_i, y_i)\}$ を近似する直線

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

が与えられているとき,その各 x_i に対する予測値 $\hat{y}_i$ と実測値 y_i との差

$$e_i = y_i - \hat{y}_i$$

を残差と呼ぶ.

残差

一般に n 個のデータの集まり $\{(x_i, y_i)\}$ を近似する直線

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

が与えられているとき,その各 x_i に対する予測値 $\hat{y}_i$ と実測値 y_i との差

$$e_i = y_i - \hat{y}_i$$

を残差と呼ぶ.残差は実際のデータと直線とのずれを表す.

線形関係と相関係数

n 個のデータの集まり $\{(x_i, y_i)\}$ を考える.

線形関係と相関係数

n 個のデータの集まり $\{(x_i, y_i)\}$ を考える.

この散布図がどれだけ直線に近いかを表す数値に**相関係数** R がある.

線形関係と相関係数

n 個のデータの集まり $\{(x_i, y_i)\}$ を考える.

この散布図がどれだけ直線に近いかを表す数値に相関係数 R がある.

(a) 相関係数 R は $-1 \leq R \leq 1$ の値をとる.

線形関係と相関係数

n 個のデータの集まり $\{(x_i, y_i)\}$ を考える.

この散布図がどれだけ直線に近いかを表す数値に相関係数R がある.

(a) 相関係数 R は $-1 \leq R \leq 1$ の値をとる.
(b) $R = \pm 1$ のとき, 散布図は完全な直線になる.

線形関係と相関係数

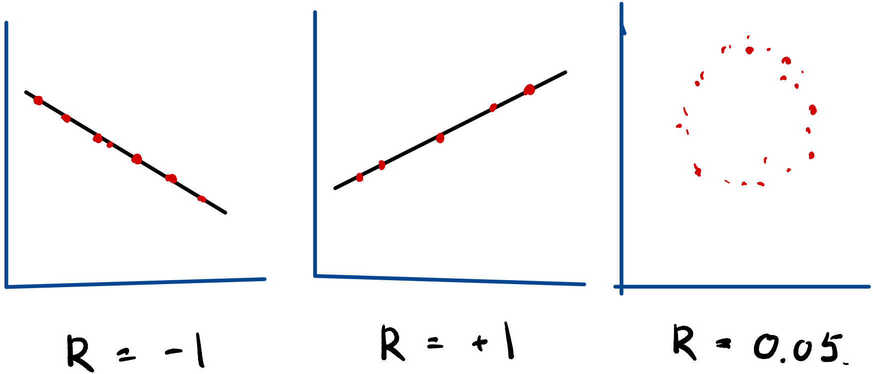
n 個のデータの集まり $\{(x_i, y_i)\}$ を考える.

この散布図がどれだけ直線に近いかを表す数値に相関係数R がある.

(a) 相関係数 R は $-1 \leq R \leq 1$ の値をとる.
(b) $R = \pm 1$ のとき, 散布図は完全な直線になる.
(c) R が 0 に近いとき, 直線関係から離れているが, 相関関係がないとは言い切れない.

線形関係と相関関係

線形関係と相関関係



線形関係と相関関係

$\{x_i\}$ と $\{y_i\}$ の標準偏差

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

を使って, 相関係数は以下の式で定義される:

線形関係と相関関係

$\{x_i\}$ と $\{y_i\}$ の標準偏差

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

を使って, 相関係数は以下の式で定義される:

$$R = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{s_x}\,\frac{y_i - \bar{y}}{s_y}.$$

ここで, n はデータの個数, $\bar{x}, \bar{y}$ は平均値を表す.

最適な直線を見つけるための指標

最適な直線を見つけるための指標

一般にデータ $\{(x_i, y_i)\}$ に対して，これを近似する最適な直線

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon$$

をどうやって見つければいいだろうか?

最適な直線を見つけるための指標

一般にデータ $\{(x_i, y_i)\}$ に対して，これを近似する最適な直線

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon$$

をどうやって見つければいいだろうか?

そのために残差 $e_i = y_i - \hat{y}_i$ に着目する．

最適な直線を見つけるための指標

最適な直線の見つけるための指標として, 以下の 2 つの量が考えられる.

最適な直線を見つけるための指標

最適な直線の見つけるための指標として, 以下の 2 つの量が考えられる.

(a) $|e_1| + |e_2| + \ldots + |e_n|$

最適な直線を見つけるための指標

最適な直線の見つけるための指標として，以下の 2 つの量が考えられる．

(a) $|e_1| + |e_2| + \ldots + |e_n|$

(b) $e_1^2 + e_2^2 + \ldots + e_n^2$

最適な直線を見つけるための指標

最適な直線の見つけるための指標として, 以下の 2 つの量が考えられる.

(a) $|e_1| + |e_2| + \ldots + |e_n|$
(b) $e_1^2 + e_2^2 + \ldots + e_n^2$

どちらも重要な量だけれども, 今回は後者 (b) の残差の二乗和を最小にする直線を最適な直線とする.

最適な直線を見つけるための指標

最適な直線の見つけるための指標として, 以下の 2 つの量が考えられる.

(a) $|e_1| + |e_2| + \ldots + |e_n|$
(b) $e_1^2 + e_2^2 + \ldots + e_n^2$

どちらも重要な量だけれども, 今回は後者 (b) の残差の二乗和を最小にする直線を最適な直線とする. このようにして求まる直線を最小二乗線と呼ぶ.

統計量から最小二乗線を求める

統計量から最小二乗線を求める

n 個のデータ $\{(x_i, y_i)\}$ の統計量 $\bar{x}$, $\bar{y}$, s_x, s_y, R から,

統計量から最小二乗線を求める

n 個のデータ $\{(x_i, y_i)\}$ の統計量 $\bar{x}$, $\bar{y}$, s_x, s_y, R から, 最小二乗線

$$\hat{y} = b_0 + b_1 x$$

の切片 b_0 と傾き b_1 が以下の式によって求まる:

統計量から最小二乗線を求める

n 個のデータ $\{(x_i, y_i)\}$ の統計量 $\bar{x}$, $\bar{y}$, s_x, s_y, R から, 最小二乗線

$$\hat{y} = b_0 + b_1 x$$

の切片 b_0 と傾き b_1 が以下の式によって求まる:

$$(a)\ b_1 = \frac{s_y}{s_x} R,$$
$$(b)\ \hat{y} - \bar{y} = b_1 (x - \bar{x}),$$
$$(c)\ b_0 = \bar{y} - b_1 \bar{x}.$$

統計量から最小二乗線を求める

n 個のデータ $\{(x_i, y_i)\}$ の統計量 $\bar{x}$, $\bar{y}$, s_x, s_y, R から, 最小二乗線

$$\hat{y} = b_0 + b_1 x$$

の切片 b_0 と傾き b_1 が以下の式によって求まる:

$$(a)\ b_1 = \frac{s_y}{s_x} R,$$
$$(b)\ \hat{y} - \bar{y} = b_1 (x - \bar{x}),$$
$$(c)\ b_0 = \bar{y} - b_1 \bar{x}.$$

(b) から最小二乗線は $(\bar{x}, \bar{y})$ を通ることが分かる.

フィットの良さ

与えられたデータ $\{(x_i, y_i)\}$ の何割が最小二乗線で説明できているかは, 相関係数の二乗 R^2 を計算するとわかる.

フィットの良さ

与えられたデータ $\{(x_i, y_i)\}$ の何割が最小二乗線で説明できているかは, 相関係数の二乗 R^2 を計算するとわかる.

例えば $R = -0.7$ の場合, $R^2 = 0.49$ だから分布の 49%が最小二乗線によって説明できている.

フィットの良さ

与えられたデータ $\{(x_i, y_i)\}$ の何割が最小二乗線で説明できているかは, 相関係数の二乗 R^2 を計算するとわかる.

例えば $R = -0.7$ の場合, $R^2 = 0.49$ だから分布の 49%が最小二乗線によって説明できている.

このことから R^2 は決定係数と呼ばれる.

最小二乗回帰が適用できる場合

データ $\{(x_i, y_i)\}$ の分布が以下の全てを満たす場合には, 最小二乗線が分布の良い近似となる.

最小二乗回帰が適用できる場合

データ $\{(x_i, y_i)\}$ の分布が以下の全てを満たす場合には, 最小二乗線が分布の良い近似となる.

(a) 線形性

最小二乗回帰が適用できる場合

データ $\{(x_i, y_i)\}$ の分布が以下の全てを満たす場合には, 最小二乗線が分布の良い近似となる.

(a) 線形性
(b) ほぼ正規分布の残差

最小二乗回帰が適用できる場合

データ $\{(x_i, y_i)\}$ の分布が以下の全てを満たす場合には, 最小二乗線が分布の良い近似となる.

(a) 線形性
(b) ほぼ正規分布の残差
(c) 一定の変動性

最小二乗回帰が適用できる場合

データ $\{(x_i, y_i)\}$ の分布が以下の全てを満たす場合には, 最小二乗線が分布の良い近似となる.

(a) 線形性
(b) ほぼ正規分布の残差
(c) 一定の変動性
(d) 独立な観測値

最小二乗回帰が適用できる場合

以下は上記の (a) - (d) のどれかが満たされていない場合:

最小二乗回帰が適用できる場合

以下は上記の (a) - (d) のどれかが満たされていない場合: