情報管理

第４回：線形モデルによる予測
～単回帰分析と重回帰分析～

今回の講義内容

今回は１個あるいは複数個の手がかりとなる情報から，所望の情報を予測する方法を解説します。

- １個の情報から予測する・・・単回帰分析
- 複数個の情報から予測する・・・重回帰分析

前回の講義で説明した最小二乗法がここでも使われます。

今回の講義を通じてしっかりおさらいしましょう。

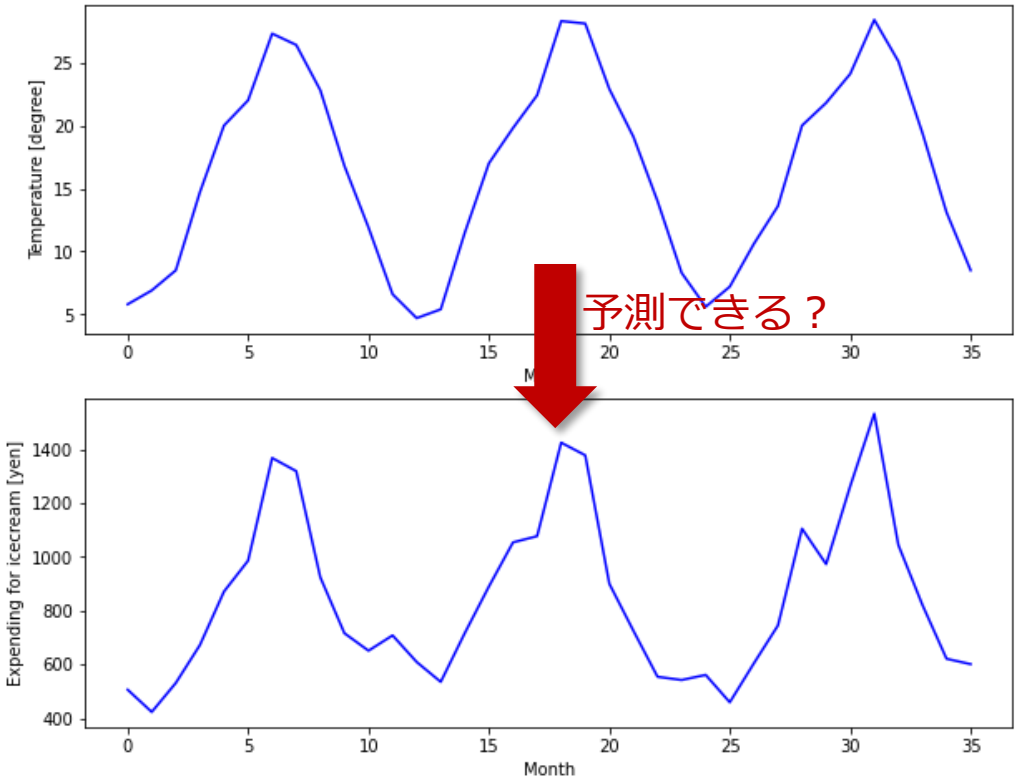
単回帰分析

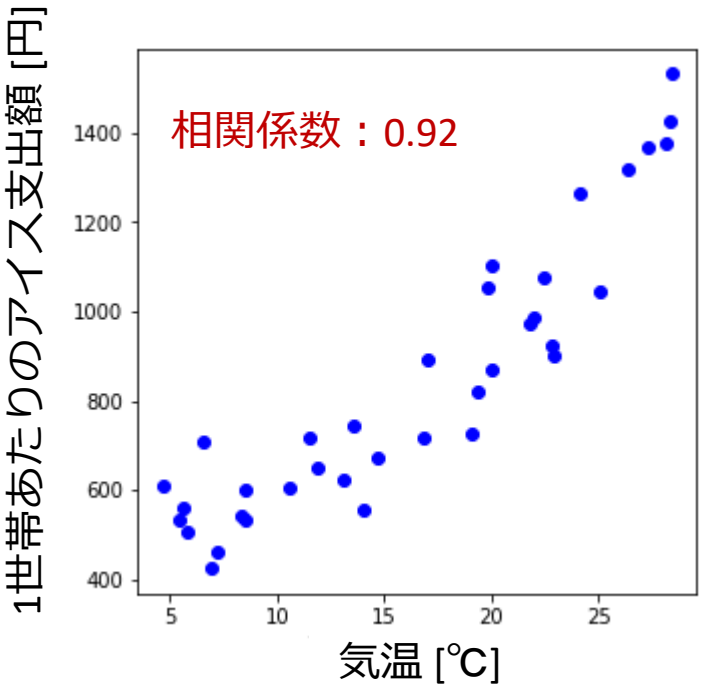
単回帰分析 – 目的

前回の講義で，気温の変化とアイス売上(正確には1世帯あたりのアイス支出額)には強い相関があることを紹介しました。

ということは，気温からアイスがどの程度売れるかを予測できそうです。

3年間の気温の変化(上)とアイスの売上(下)





単回帰分析 – 線形モデル

ある月 n における1世帯あたりのアイス支出額を y_n，気温を x_n としたとき，
アイス支出額 y_n を以下の線形関数でモデル化できると仮定します。

$$y_n = ax_n + b$$

x_n は推定をするための手がかりとなる情報で，説明変数と呼びます。
y_n は推定したい目的となる情報で，目的変数と呼びます。

目的変数を説明変数で表現する分析を回帰分析と呼びます。
説明変数が1種類（この例では気温）のみの場合，単回帰分析と呼びます。
単回帰分析の目的は，上の式におけるパラメータ a と b を求めることです。

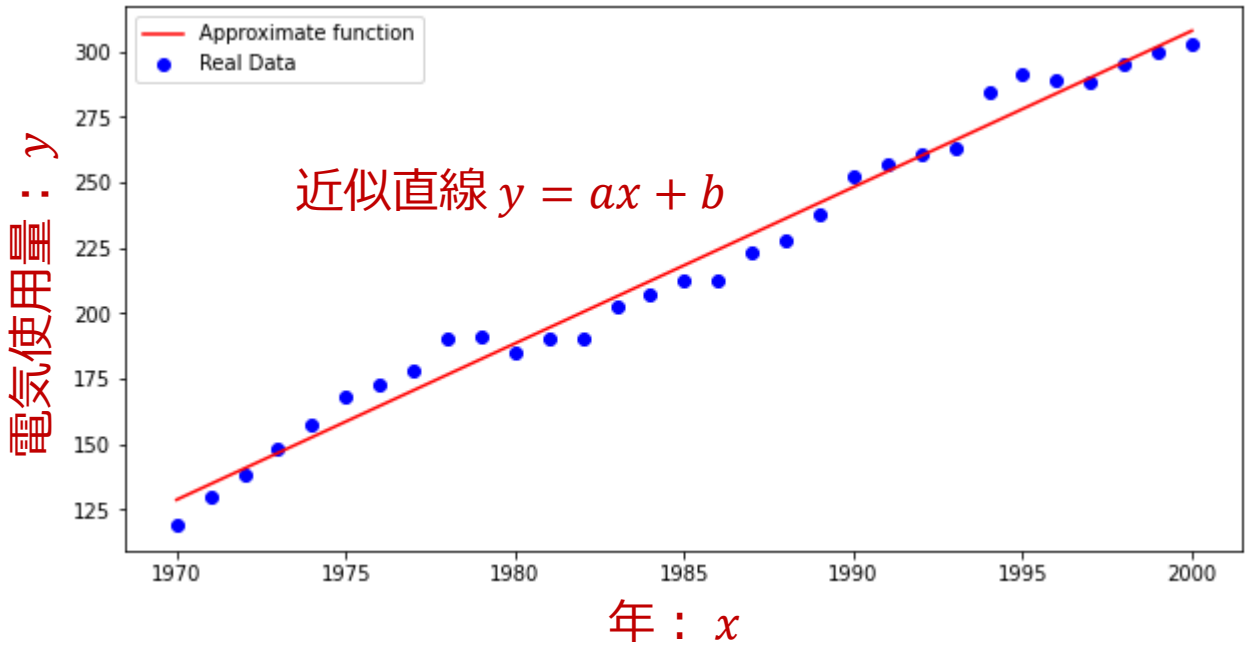
単回帰分析 – パラメータの求め方

単回帰分析のパラメータa, b の求め方は，前回解説した時系列データの直線近似と同じです。

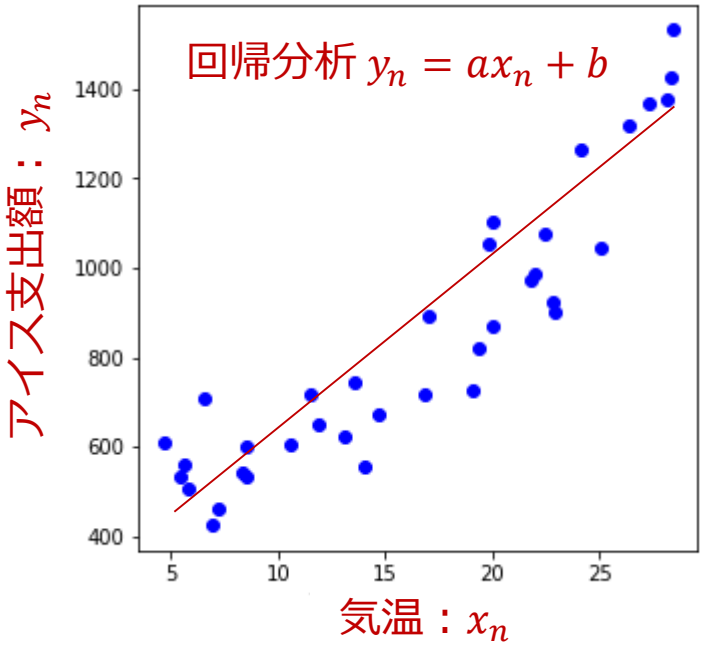
直線近似では x が時系列データの時間であったのに対して，単回帰分析では別の情報の値になっている点が異なります。

つまり，直線近似は説明変数 x が時間である単回帰分析の一例と言えます。

年ごとの電気使用量の直線近似（前回資料参照）



回帰分析



単回帰分析 – パラメータの求め方

直線近似の時と同様，**最小二乗法**を用いて a と b を求めます。

$$L = \sum_n \left(y_n - (ax_n + b)\right)^2 = \sum_n (y_n^2 + a^2 x_n^2 + b^2 - 2ax_n y_n + 2abx_n - 2by_n) \rightarrow min$$

L は a に対する下に凸の2次関数であり，また b に対しても下に凸の2次関数である。
よって，最小値を求めるにはa, b で偏微分し，=0を解けばよい。

L を a に対して偏微分し，= 0とおく

$$\frac{\partial L}{\partial a} = 2a\sum_n x_n^2 - 2\sum_n x_n y_n + 2b\sum_n x_n = 0$$

$$\rightarrow a\sum_n x_n^2 - \sum_n x_n y_n + b\sum_n x_n = 0 \quad \cdots\cdots \quad ①$$

L を b に対して偏微分し，= 0とおく

$$\frac{\partial L}{\partial b} = 2Nb + 2a\sum_n x_n - 2\sum_n y_n = 0$$

$$\rightarrow a\sum_n x_n + Nb - \sum_n y_n = 0 \quad \cdots\cdots \quad ②$$

単回帰分析 – パラメータの求め方

$$a\sum_n x_n^2 - \sum_n x_n y_n + b\sum_n x_n = 0 \quad \cdots\cdots \quad ①$$

$$a\sum_n x_n + Nb - \sum_n y_n = 0 \quad \cdots\cdots \quad ②$$

上記の連立方程式を解くと，以下の結果が得られます（導出は省略）

単回帰分析のパラメータ

x_n: n番目のデータの x 軸の値
y_n: n番目のデータの y 軸の値
N: データの総数

$$a = \frac{N\sum_n x_n y_n - \sum_n x_n \sum_n y_n}{N\sum_n x_n^2 - (\sum_n x_n)^2}$$

$$b = \frac{\sum_n x_n^2 \sum_n y_n - \sum_n x_n y_n \sum_n y_n}{N\sum_n x_n^2 - (\sum_n x_n)^2}$$

単回帰分析を実行してみよう

04_01_single_linear_regression.ipynb を動かして，単回帰分析によるアイス売上予測を行ってみましょう。

予測結果の評価について

04_01_single_linear_regression.ipynb では3種類の指標を使って予測性能を評価していました。

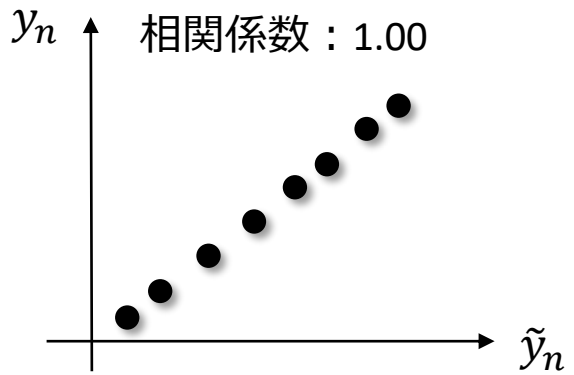
単回帰分析によって予測されたアイス支出額を $\tilde{y}_n = ax_n + b$ ，実際のアイス支出額を y_n とします。

- 平均二乗誤差（Mean square error: MSE）= $\frac{1}{\mathrm{N}}\sum_{n=1}^{N}(y_n - \tilde{y}_n)^2$

  最小二乗法において最小化している式そのもの（前回解説済み）。
- 平方根平均二乗誤差 (Root mean square error: RMSE) = $\sqrt{\mathrm{MSE}}$

  平均二乗誤差の平方根を取ることで，目的変数と同じ単位で誤差が測れる（前回解説済み）。
- 相関係数 = $\frac{\sum_n(\tilde{y}_n-\bar{\tilde{y}})(y_n-\bar{y})}{\sqrt{\sum_n(\tilde{y}_n-\bar{\tilde{y}})^2}\sqrt{\sum_n(y_n-\bar{y})^2}}$

  $y_n = \tilde{y}_n$ の場合(右図)，相関係数は 1 になる。



x_n と y_n の相関 = $\tilde{y}_n$ と y_n の相関

説明変数 x_n と目的変数 y_n の相関は，単回帰分析の式 $\tilde{y}_n = ax_n + b$ と y_n の相関と一致します。

そのため，単回帰分析を実際に行わなくても，説明変数と目的変数の相関を見るだけで，単回帰分析をしたときの性能が評価できてしまいます。

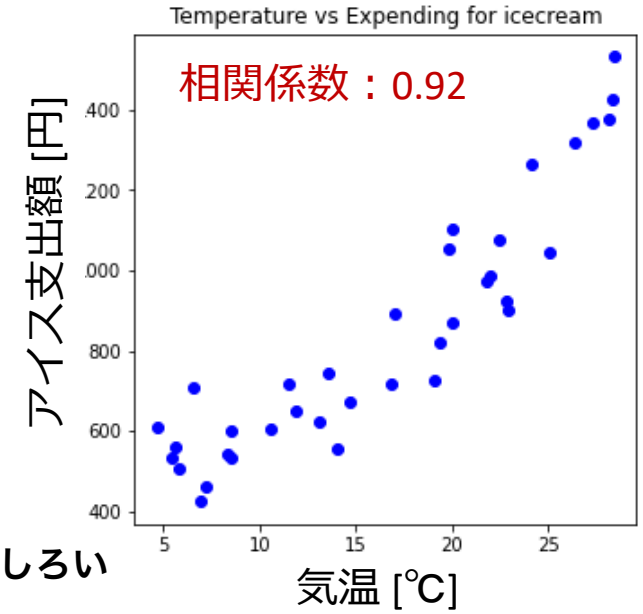
$$R = \frac{\sum_n (ax_n + b - \overline{ax_n + b})(y_n - \bar{y})}{\sqrt{\sum_n \left(ax_n + b - \overline{ax_n + b}\right)^2}\sqrt{\sum_n (y_n - \bar{y})^2}}$$

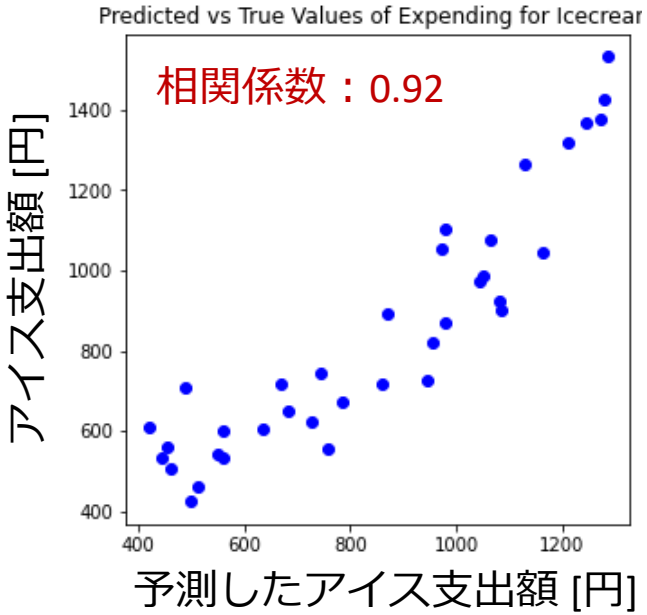
$$= \frac{\sum_n (ax_n + b - a\bar{x} + b)(y_n - \bar{y})}{\sqrt{\sum_n (ax_n + b - a\bar{x} + b)^2}\sqrt{\sum_n (y_n - \bar{y})^2}}$$

$$= \frac{a\sum_n (x_n - \bar{x})(y_n - \bar{y})}{a\sqrt{\sum_n (x_n - \bar{x})^2}\sqrt{\sum_n (y_n - \bar{y})^2}}$$

$$= \frac{\sum_n (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\sum_n (x_n - \bar{x})^2}\sqrt{\sum_n (y_n - \bar{y})^2}}$$

これおもしろい

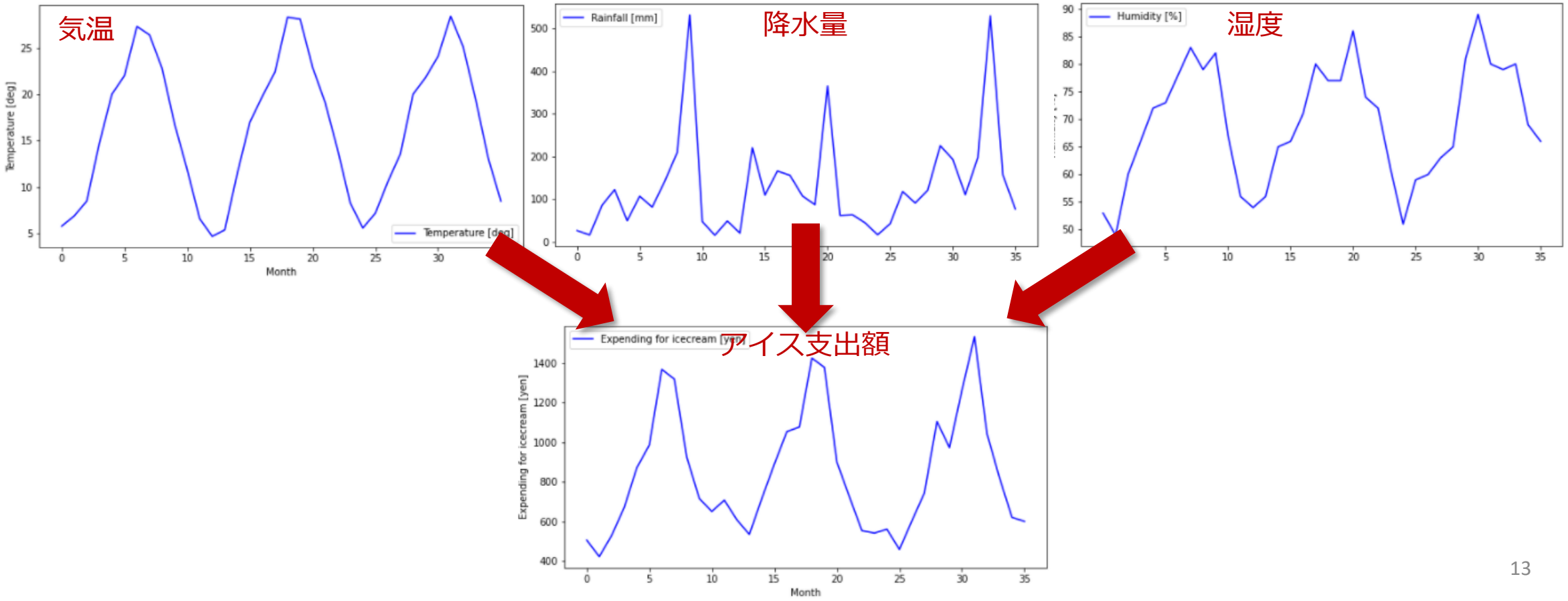




重回帰分析

重回帰分析 – 目的

単回帰分析では，気温情報のみからアイスの売上を予測しようとしていました。
しかし，アイスの売り上げに関係する情報は，気温だけではないはずです。
そこで，気温以外の情報も用いて，アイス売上を予測してみます。



重回帰分析

説明変数が複数ある場合，重回帰分析と呼びます。
重回帰分析は以下の式で定義されます。

$$y_n = a_1 x_{n,1} + a_2 x_{n,2} + \cdots + a_M x_{n,M} + b$$
$$= \boldsymbol{x}_n \boldsymbol{a}$$

$$\boldsymbol{x}_n = \left[x_{n,1}, x_{n,2}, \ldots, x_{n,M}, 1\right], \qquad \boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_M \\ b \end{bmatrix}$$

$x_{n,1}, x_{n,2}, \ldots, x_{n,M}$はそれぞれ $1, 2, \ldots, M$ 個目の説明変数です。
重回帰分析の目的は，上式におけるベクトル $\boldsymbol{a}$ を求めることです。
（$M = 1$ のとき，単回帰分析）

重回帰分析 – パラメータの求め方

単回帰分析と同じく，最小二乗法を使って求めます。

$$
\begin{aligned}
L &= \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \boldsymbol{a})^2 \\
&= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}) \\
&= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{a} - \boldsymbol{a}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{a}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{a} \\
&= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{a} + \boldsymbol{a}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{a} \\
&\rightarrow min
\end{aligned}
$$

$$
\boldsymbol{y} = [y_1\ y_2\ \cdots\ y_N]
$$

$$
\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M} & 1 \\ x_{2,1} & x_{2,2} & \cdots & x_{2,M} & 1 \\ & & \vdots & & \\ x_{N,1} & x_{N,2} & \cdots & x_{N,M} & 1 \end{bmatrix}
$$

L を $\boldsymbol{a}$ で偏微分し，$= 0$ を解く

$$
\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{a}} &= -2\frac{\partial}{\partial \boldsymbol{a}}(\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{a}) + \frac{\partial}{\partial \boldsymbol{a}}(\boldsymbol{a}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{a}) \\
&= -2\boldsymbol{X}^T\boldsymbol{y} + (\boldsymbol{X}^T\boldsymbol{X} + (\boldsymbol{X}^T\boldsymbol{X})^T)\boldsymbol{a} \\
&= -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{a} \\
&= 0
\end{aligned}
$$

これを解いて以下の式を得る。

$$
\boldsymbol{a} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}
$$

（$\boldsymbol{X}^{-1}$ は $\boldsymbol{X}$ の逆行列）

重回帰分析を実行してみよう

04_02_multiple_linear_regression.ipynb を動かして，重回帰分析によるアイス売上予測を行ってみて，単回帰分析の結果と比較してみましょう。

おわりに

今回は，情報を予測する方法として単回帰分析と重回帰分析について解説しました。

重回帰分析はプロモーション分析（売上に貢献している広報活動の分析）などでも使われる基礎的な分析方法ですので，覚えておくと良いでしょう。

演習プログラムでは，データに対して回帰分析を適用して終わりでしたが，分析結果の正確さをきちんと検証するためには，回帰分析に用いたデータ（学習データ）と異なるデータ（テストデータ）に対して予測式を適用し，精度を測る必要があります。これについては今回のレポートで実施してもらいます。

レポート課題

今回のレポートは report_04.ipynb に加筆する形で行ってもらいます。

(report_04.ipynb はBEEFの第四回ファイル一式に入っています。)

時系列データを直線近似 $y = ax + b$ した場合と，5次多項式近似 $y = a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + b$ したときの挙動を比較・考察してもらいます。

report_04.ipynb では学習データと呼ばれるデータ “train.csv” を使って回帰分析を行い，求まったパラメータを使って，テストデータと呼ばれるデータ “test.csv” に対して目的変数の推定を行います。

詳しくは， report_04.ipynb を参照してください。

レポート提出期限：5/17(火) AM10:30，ipynbファイルをhtmlファイルに変換して提出