

# データ分析基礎 回帰分析 + 演習の手順

京都大学 国際高等教育院 附属データ科学イノベーション教育研究センター

せきど ひろと  
關戸 啓人

sekido.hiroto.7a@kyoto-u.ac.jp

# 回帰分析と最小二乗法

# 回帰分析と回帰曲線

## ★ 多変量解析

- ★ 複数の確率変数の関係を調べる
  - ★ 「身長」と「体重」の関係
  - ★ 「気温」と「ビールの売上」の関係
  - ★ 「朝食を食べる割合」と「テストの点数」の関係

## ★ 回帰分析

- ★ 回帰曲線（回帰曲面）を推定することで複数の確率変数の関係を調べる

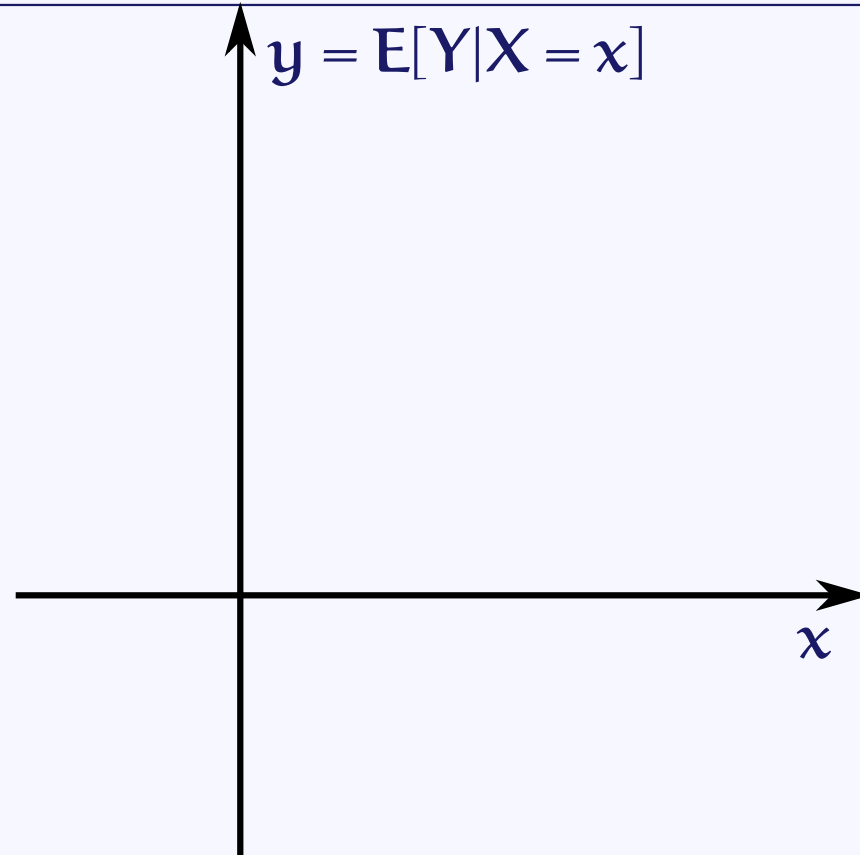
## ★ 回帰曲線

- ★ 2つの確率変数  $X$  と  $Y$  を考える
- ★  $X = x$  という条件下での  $Y$  の平均  $E[Y|X = x]$  を  $x$  の関数と思ったとき、それを回帰曲線という

# 回帰分析と回帰曲線

★  $X = x$  という条件下での  $Y$  の平均  $E[Y|X = x]$  を  $x$  の関数と思ったとき、それを回帰曲線という

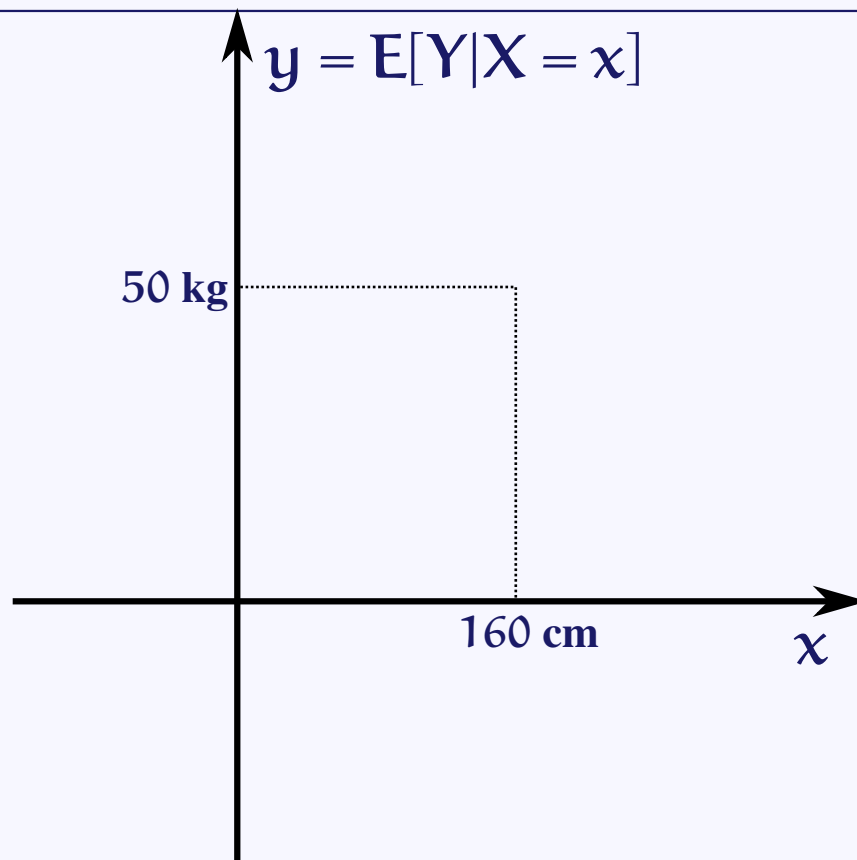
★ 例)  $X$  は身長,  $Y$  は体重を表すとする



# 回帰分析と回帰曲線

★  $X = x$  という条件下での  $Y$  の平均  $E[Y|X = x]$  を  $x$  の関数と思ったとき、それを回帰曲線という

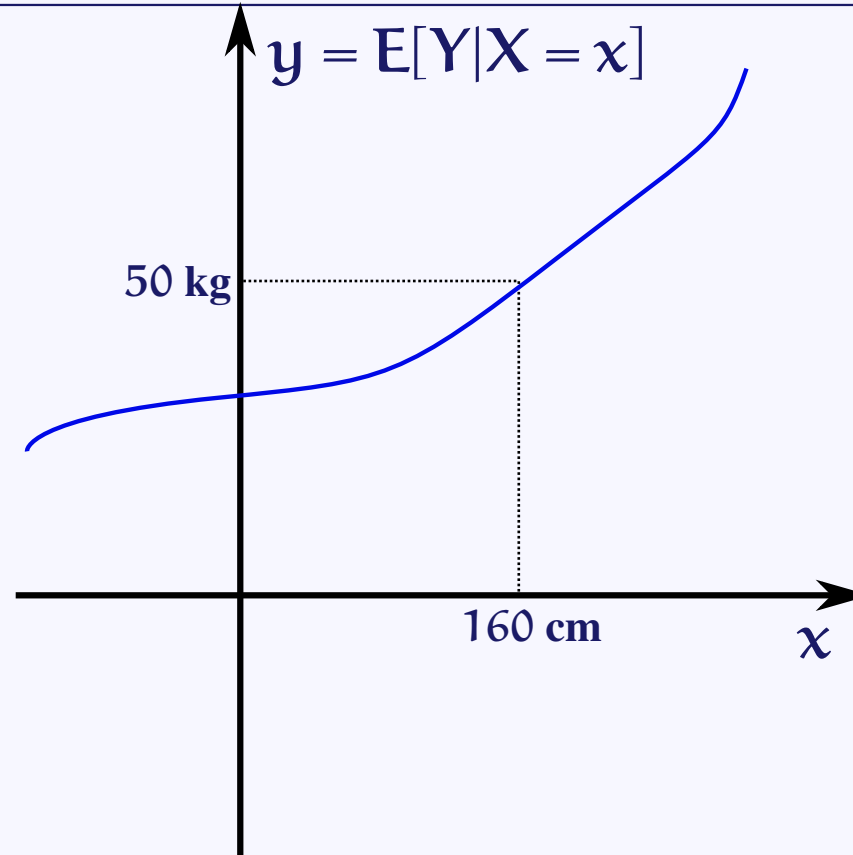
★ 例)  $X$  は身長,  $Y$  は体重を表すとする



# 回帰分析と回帰曲線

★  $X = x$  という条件下での  $Y$  の平均  $E[Y|X = x]$  を  $x$  の関数と思ったとき、それを回帰曲線という

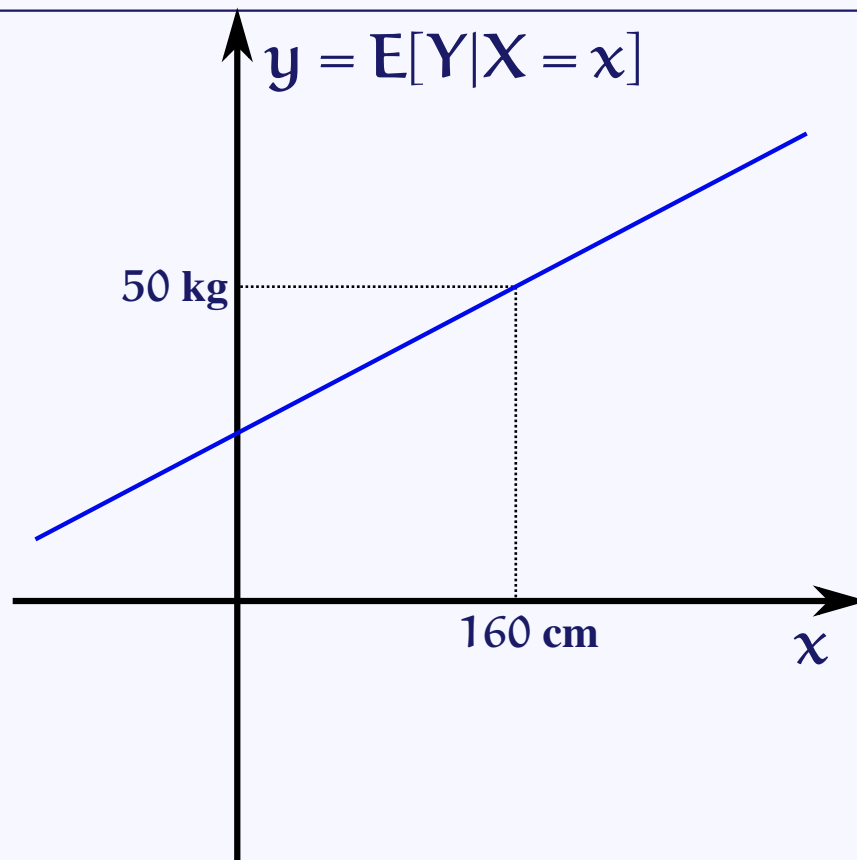
★ 例)  $X$  は身長,  $Y$  は体重を表すとする



# 回帰分析と回帰曲線

★  $X = x$  という条件下での  $Y$  の平均  $E[Y|X = x]$  を  $x$  の関数と思ったとき、それを回帰曲線という

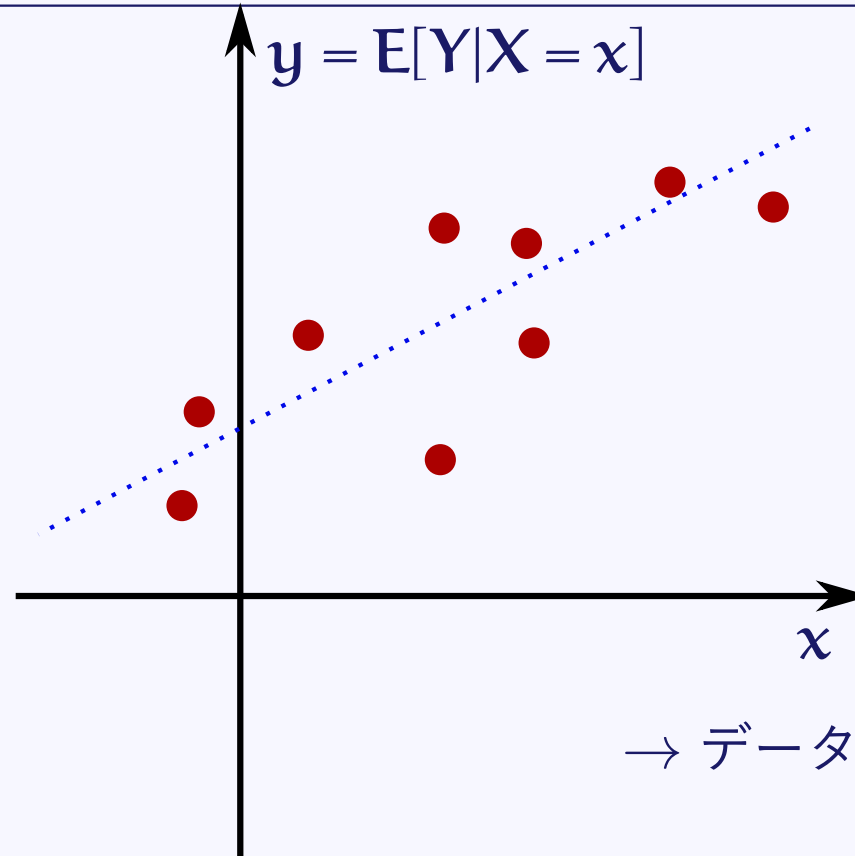
★ 簡単のため直線  $y = ax + b$  ( $a, b \in \mathbb{R}$ ) であると仮定することが多い



# 回帰分析と回帰曲線

★  $X = x$  という条件下での  $Y$  の平均  $E[Y|X = x]$  を  $x$  の関数と思ったとき, それを回帰曲線という

★ 簡単のため直線  $y = ax + b$  ( $a, b \in \mathbb{R}$ ) であると仮定することが多い



→ データから  $a, b$  を推定する



# 単回帰分析と重回帰分析

- ★  $E[Y|X = x]$  を推定するときは、 $X$  は説明変数、 $Y$  は被説明変数（目的変数）と呼ばれる
- ★ つまり、 $Y$  がどのような値を取るかは  $X$  によって決まる、と考えている
  - ★  $Y$ : ビールの売上,  $X$ : 気温
  - ★  $Y$ : テストの点数,  $X$ : 朝食を食べる割合

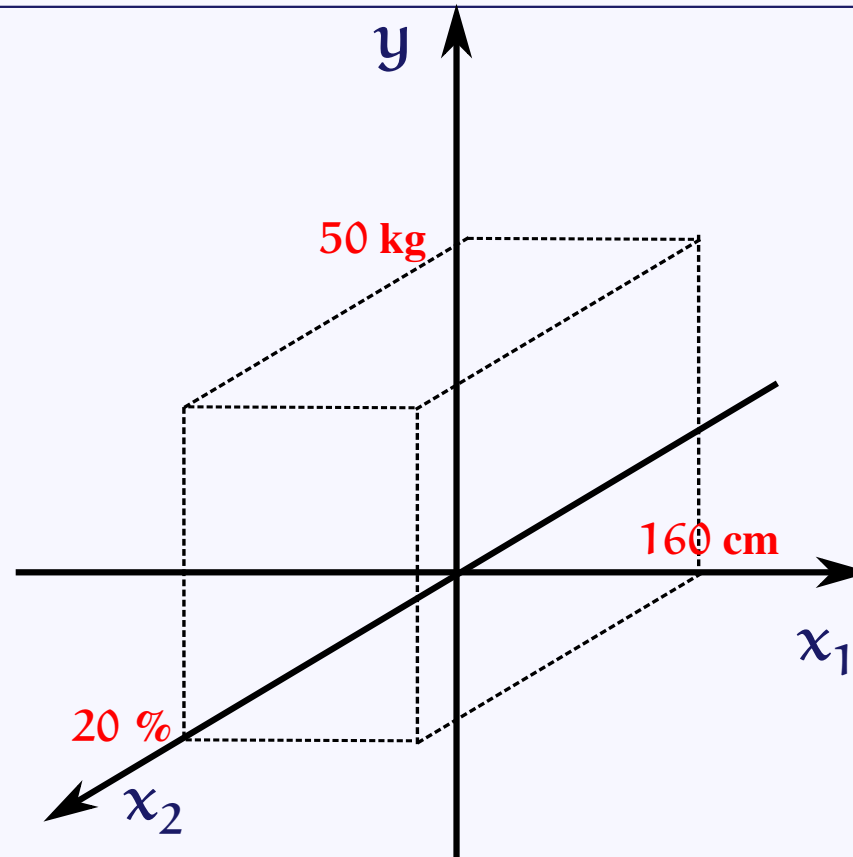
---

- ★ 説明変数は複数あっても良い
- ★ 説明変数が  $X_1, X_2, \dots, X_n$  で、 $E[Y|X_1 = x_1, X_2 = x_2, \dots, X_n = x_n]$  を考えても良い
  - ★ 説明変数が1個の場合を単回帰分析、複数の場合を重回帰分析という

# 重回帰分析の例

★ 説明変数が2個の場合： $y = E[Y|X_1 = x_1, X_2 = x_2]$

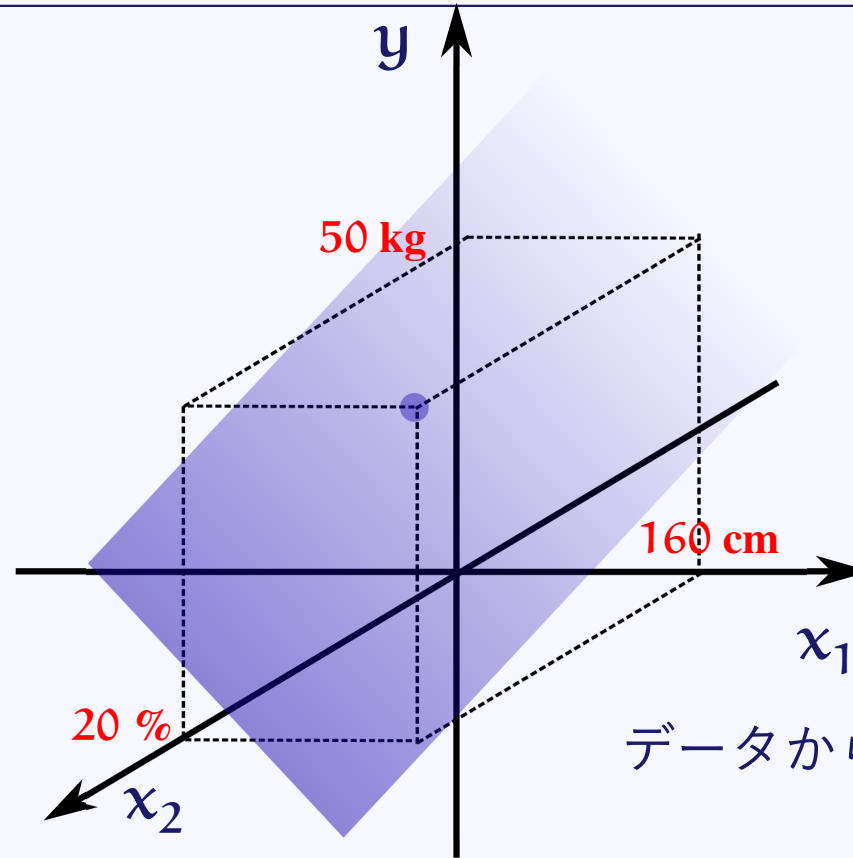
★ 例)  $X_1$  は身長,  $X_2$  は体脂肪率,  $Y$  は体重を表すとする



# 重回帰分析の例

★ 説明変数が2個の場合： $y = E[Y|X_1 = x_1, X_2 = x_2]$

★ 回帰曲面は平面  $y = a_1x_1 + a_2x_2 + b$  ( $a_1, a_2, b \in \mathbb{R}$ ) であると仮定することが多い

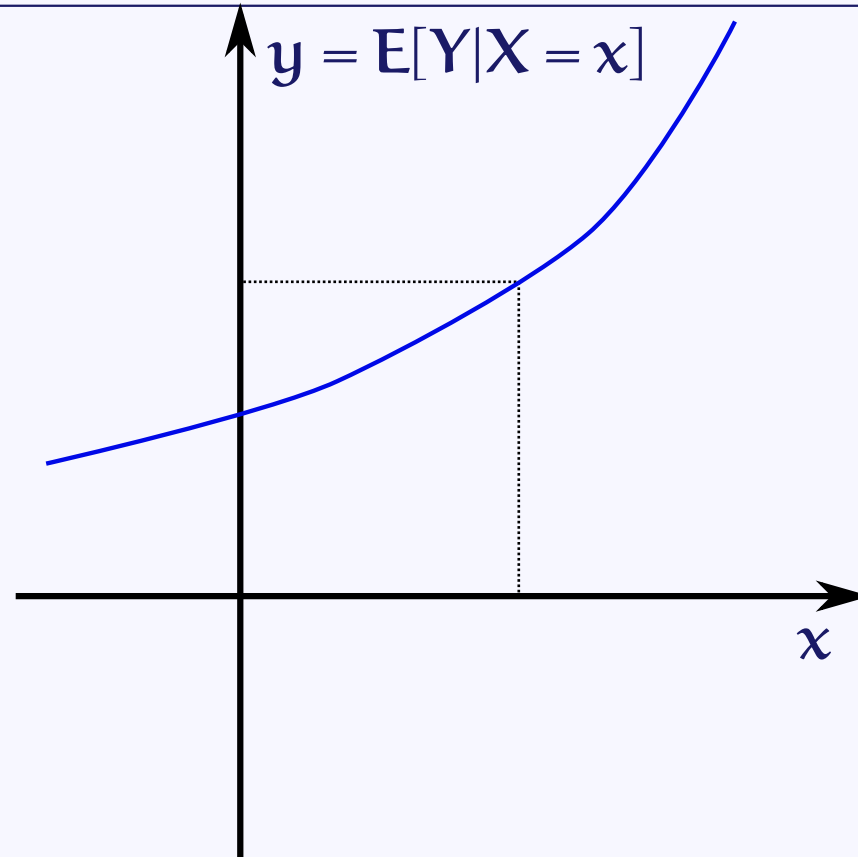


データから  $a_1, a_2, b$  を推定する

## 重回帰分析の例: 二次関数

★「身長 (X)」と「体重 (Y)」の関係は直線なのか？

★ BMIなどを考慮すると二次関数  $y = ax^2 + bx + c$  と仮定したほうが良いのでは？

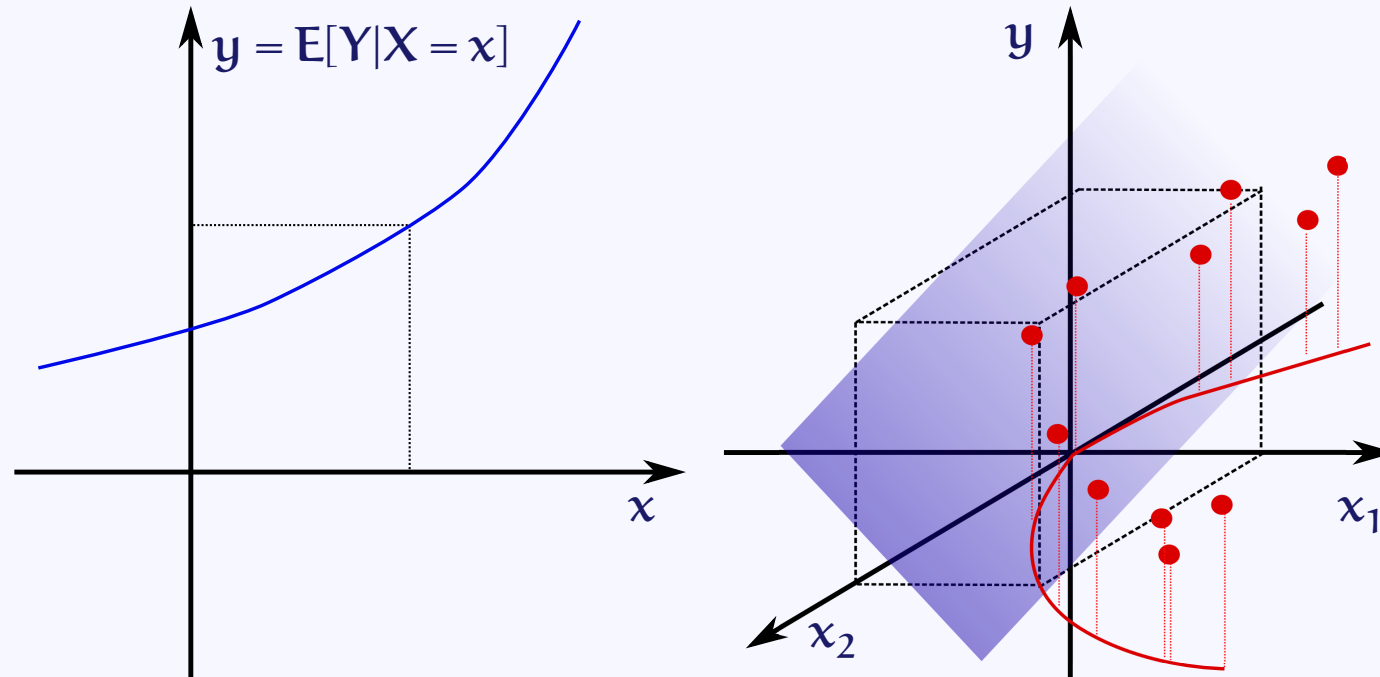


# 重回帰分析の例: 二次関数

★「身長 (X)」と「体重 (Y)」の関係は直線なのか？

★ BMIなどを考慮すると二次関数  $y = ax^2 + bx + c$  と仮定したほうが良いのでは？  
→ 重回帰分析

★  $y = a_1x_1 + a_2x_2 + b$  において  $(a_1, a_2, b, x_1, x_2) \rightarrow (a, b, c, x^2, x)$  と読み替えれば良い



## 問題

- ★ 確率変数  $X$  は血圧を表すとし、 $Y$  は年収を表すとする
- ★ 「血圧」と「年収」の関係を回帰分析で調べた場合どうなるか？

★ 回帰直線は右肩上がりになる

★  $y = ax + b$  とすると  $a > 0$

★ 年収を上げるには血圧を上げれば良い！

★ と考えるのは危険

# 解説

★「年収」と「血圧」には確かに正の相関があるが因果関係などは何も言っていない

★ 年収が多い人は、ストレスが掛かる仕事をしており、血圧が高いかもしれない

★ 実はこの場合はこれもほぼ正しくない

★「年収」も「血圧」も「年齢」と正の相関がある

★ 確率変数  $X_1$  は血圧を、 $X_2$  は年齢を、 $Y$  は年収を表すとする

★ 重回帰分析をすると  $y = a_1x_1 + a_2x_2 + b$  において  $a_2 > 0$  だが  $a_1 > 0$  とは限らない

★ 仮定が良くなかった

# 朝食を食べる割合の例について検証

★「朝食を食べる割合 (X)」と「テストの点数 (Y)」の関係を回帰分析で調べた場合はどうなるか？

★ 回帰直線は右肩上がりになる

★  $y = ax + b$  とすると  $a > 0$

★ テストの点数を上げるには朝食を食べれば良い！

★ 栄養がある状態のほうが頭が働いて勉強できる

★ 朝食を食べる割合が多い家庭はしつけができてるだけなのでは…，無理やり朝食を食べてもテストの点数は変わらないよ

★ 例え，朝食を食べることとテストの点数に直接的な因果関係がなくても，無理やり朝食を食べたら生活環境とかの影響でテストの点数上がるかも

★ よくわからない

★ 多角的に分析し，良さそうなら実際に試してみる



# 最小二乗法の概要

★ 未知な関数を得られたデータから推定したい

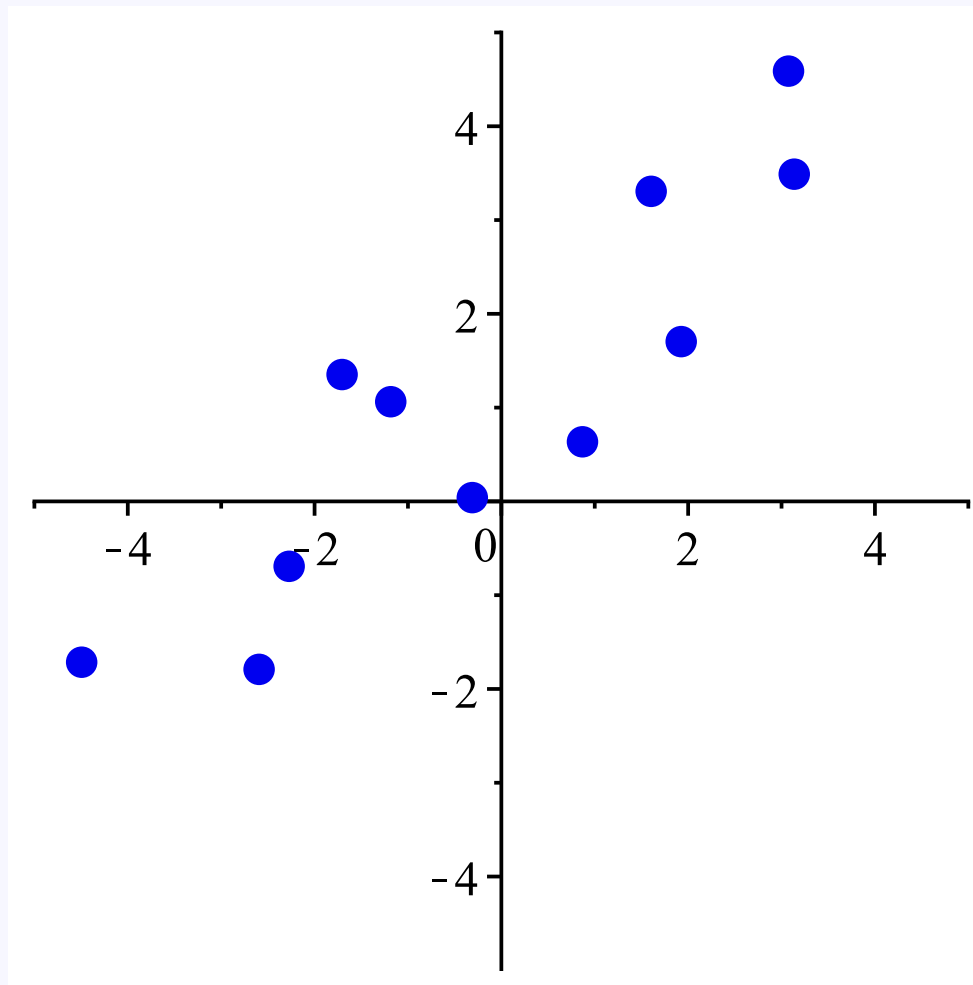
★ 未知関数  $f(x)$  の形はわかっている、未知パラメータを含む形で書かれる

★ データ  $(x_j, y_j)$  は  $f(x_j)$  での値が  $y_j$  であることを「示唆」する

★ データは厳密に「正しい」訳ではない。つまり厳密に  $f(x_j) = y_j$  とは限らない（測定誤差などが含まれている）

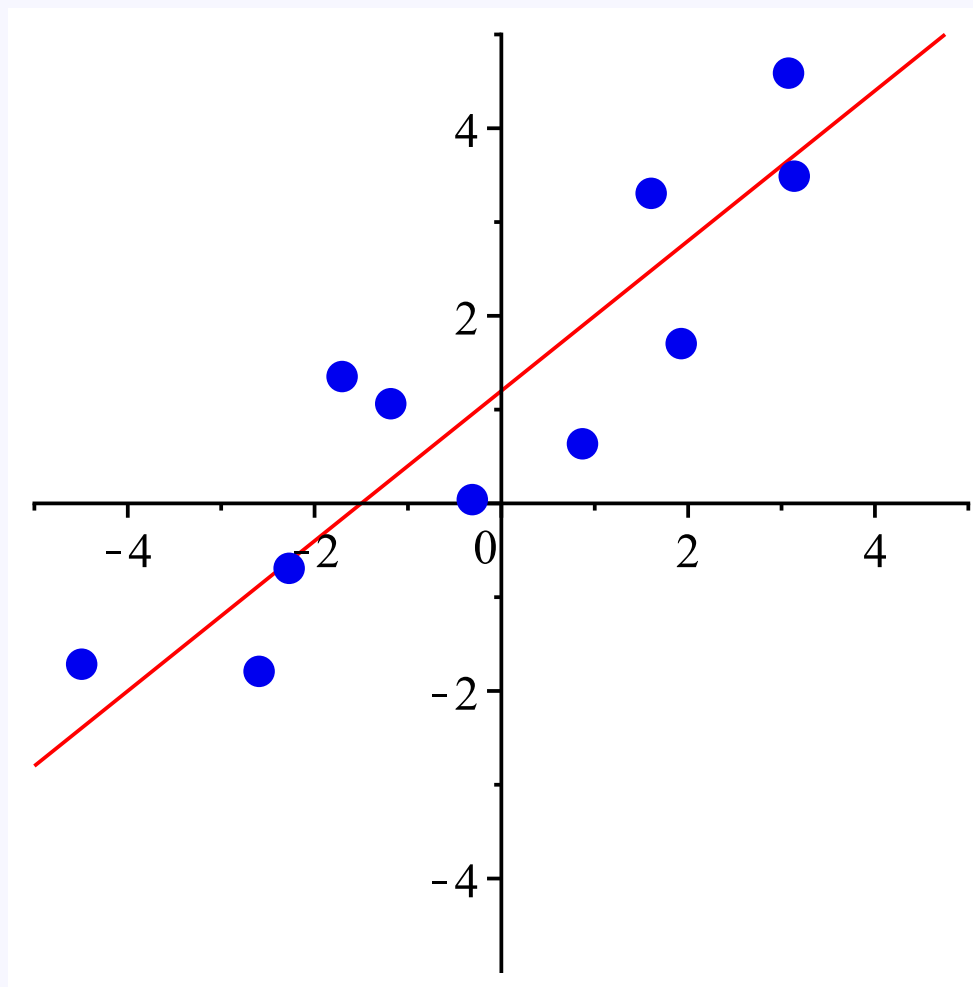
★ 2変数以上の場合は  $x$  はベクトルだと思えば良い

# 最小二乗法の例 (その1)



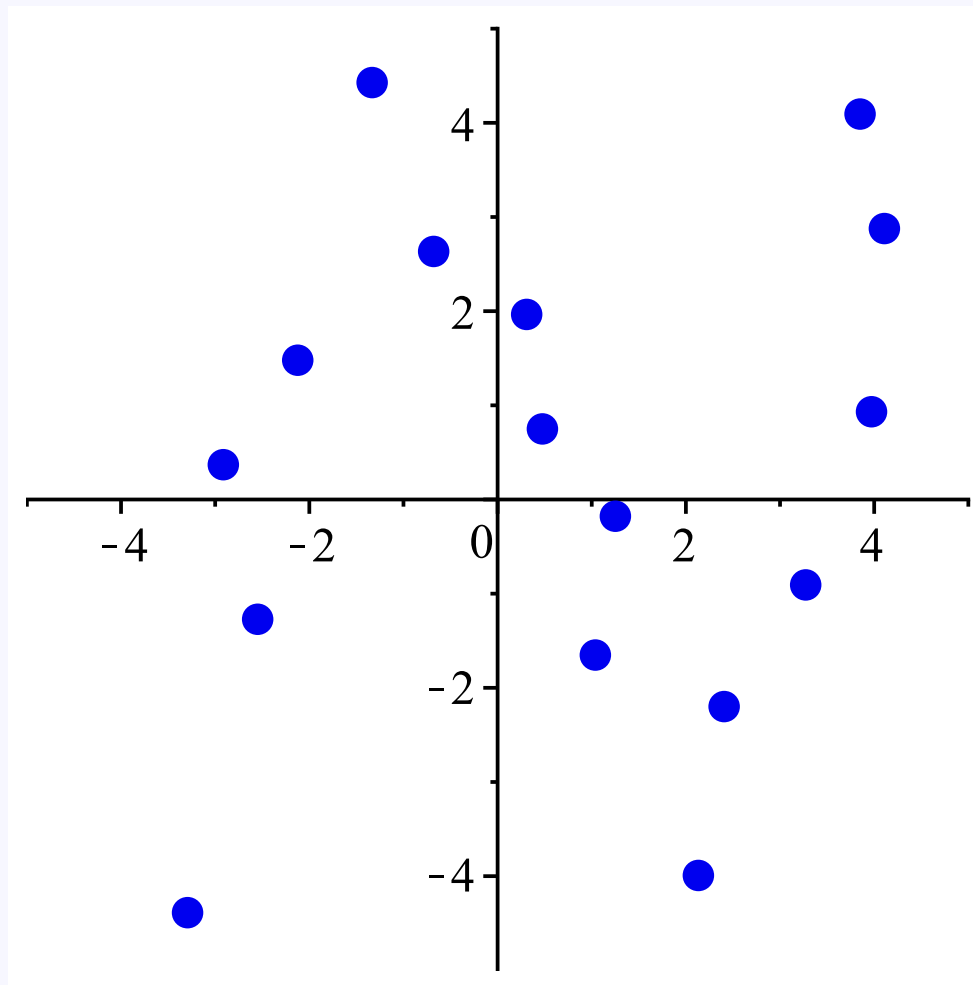
$$f(x) = \theta_1 x + \theta_0$$

# 最小二乗法の例 (その1)



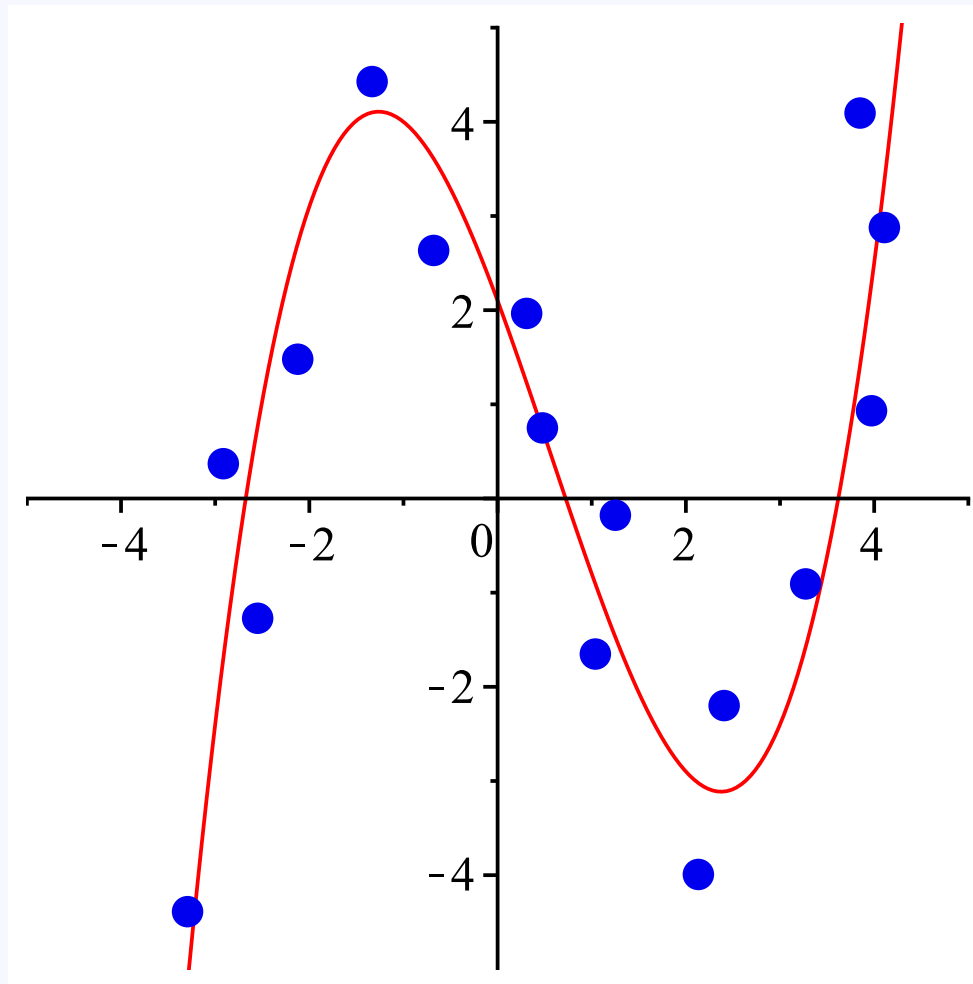
$$f(x) = 0.8x + 1.2$$

# 最小二乗法の例 (その2)



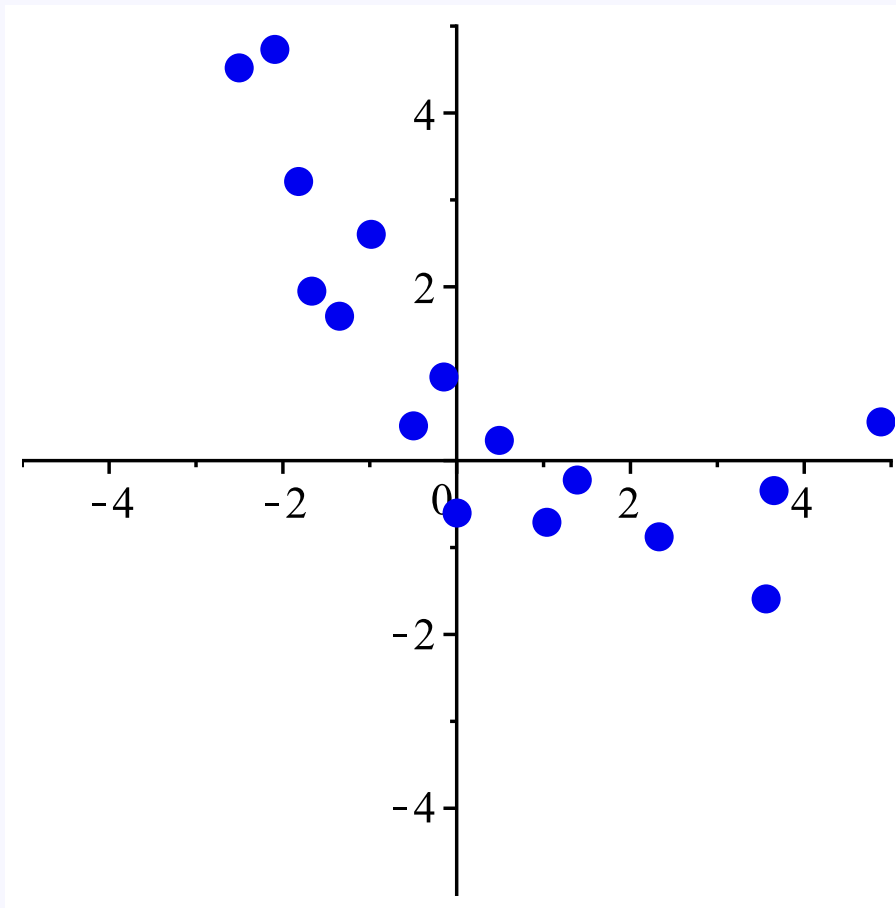
$$f(x) = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x + \theta_0$$

# 最小二乗法の例 (その2)



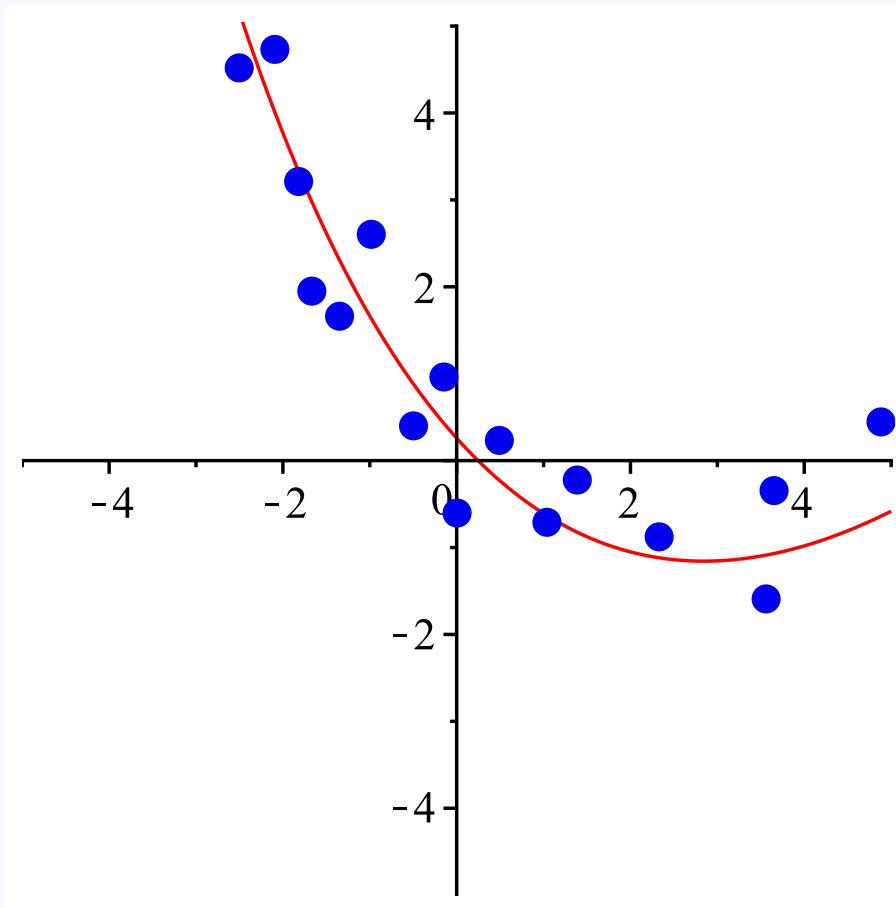
$$f(x) = 0.3x^3 - 0.5x^2 - 2.7x + 2.1$$

# 最小二乗法の例 (その3)



$$f(x) = \frac{\theta_2 x^2 + \theta_1 x + \theta_0}{x + \theta_3}$$

# 最小二乗法の例 (その3)



$$f(x) = \frac{2.1x^2 - 13.1x + 3.1}{x + 12.0}$$

# 最小二乗法の例

## ★ その1: 直線で近似する場合

★  $f(x) = \theta_0 + \theta_1 x$

★ (単純な) 単回帰分析

## ★ その2: 未知関数がパラメータについて線形 (線形最小二乗法)

★  $f(x) = \theta_0 f_0(x) + \theta_1 f_1(x) + \dots + \theta_{m-1} f_{m-1}(x)$

★ (単純な) 重回帰分析, 以下では主にこれを説明する

## ★ その3: 未知関数がパラメータについて非線形 (非線形最小二乗法)

★  $f(x) = f(x; \theta_0, \theta_1, \dots, \theta_{m-1})$

★ 複雑な式の形を指定した場合, 解く場合は最適化の理論を用いる



# 回帰モデルの例 (1) — 単回帰モデル

- ★ 体重を意味する確率変数を  $W$
- ★ 身長を意味する確率変数を  $H$
- ★ モデル：  $W = \theta_1 H + \theta_0 + \varepsilon$

★ データは、例えば

	体重 (kg)	切片	身長 (cm)
A 氏	56.8	1	163.3
B 氏	52.1	1	160.2
C 氏	52.6	1	158.0
D 氏	23.4	1	129.0
E 氏	32.1	1	139.7
F 氏	40.6	1	141.4

## 回帰モデルの例 (2-1) — 重回帰モデル

- ★ 体重を意味する確率変数を  $W$
- ★ 身長を意味する確率変数を  $H$
- ★ モデル： $W = \theta_2 H^2 + \theta_1 H + \theta_0 + \varepsilon$

★ データは、例えば

	体重 (kg)	切片	身長 (cm)	身長 <sup>2</sup> (cm <sup>2</sup> )
A氏	56.8	1	163.3	26666.89
B氏	52.1	1	160.2	25664.04
C氏	52.6	1	158.0	24964.00
D氏	23.4	1	129.0	16641.00
E氏	32.1	1	139.7	19516.09
F氏	40.6	1	141.4	19993.96

## 回帰モデルの例 (2-2) — 重回帰モデル

- ★ 体重を  $W$ , 身長  $H$ , 体脂肪率を  $F$ , 性別を  $S$
- ★ 性別は女性を 1, 男性を 0 で表す
- ★ モデル:  $W = \theta_3 S + \theta_2 F + \theta_1 H + \theta_0 + \varepsilon$

★ データは, 例えば

	体重 (kg)	切片	身長 (cm)	体脂肪率 (%)	性別
A 氏	56.8	1	163.3	14.3	0
B 氏	52.1	1	160.2	15.3	0
C 氏	52.6	1	158.0	21.2	1
D 氏	23.4	1	129.0	13.3	1
E 氏	32.1	1	139.7	16.8	0
F 氏	40.6	1	141.4	19.6	1

# 線形最小二乗法の定義, および, 性質 1

## ★ 観測と応答の関係

$$Y = \sum_{k=0}^{m-1} \theta_k f_k(x) + \varepsilon = f(x, \theta) + \varepsilon$$

は線形回帰モデルと呼ばれる

★  $f_k(x)$  は既知の関数

★  $\theta_k$  は未知のパラメータ,  $\theta = (\theta_0, \theta_1, \dots, \theta_{m-1})^T$

★  $\varepsilon$  は確率変数で平均0 ( $E[\varepsilon] = 0$ )

★ 実際に  $n$  個のデータ  $(x_1, y_1), \dots, (x_n, y_n)$  を用いて

$$y_j = f(x_j, \theta) + \varepsilon_j, \quad j = 1, 2, \dots, n$$

とする

★  $y_j, \varepsilon_j$  は確率変数

★  $\varepsilon_j$  は  $j$  回目の観測における誤差

## 線形最小二乗法の定義, および, 性質 2

★  $y_j = f(x_j, \theta) + \varepsilon_j, \quad j = 1, 2, \dots, n$

★ 今回は, 誤差  $\varepsilon_j$  に対して以下の仮定を置く

★ 平均は0. つまり,  $E[\varepsilon_j] = 0$

★ 誤差の分散は等しく, 正. つまり,  $V[\varepsilon_j] = \sigma^2 > 0$

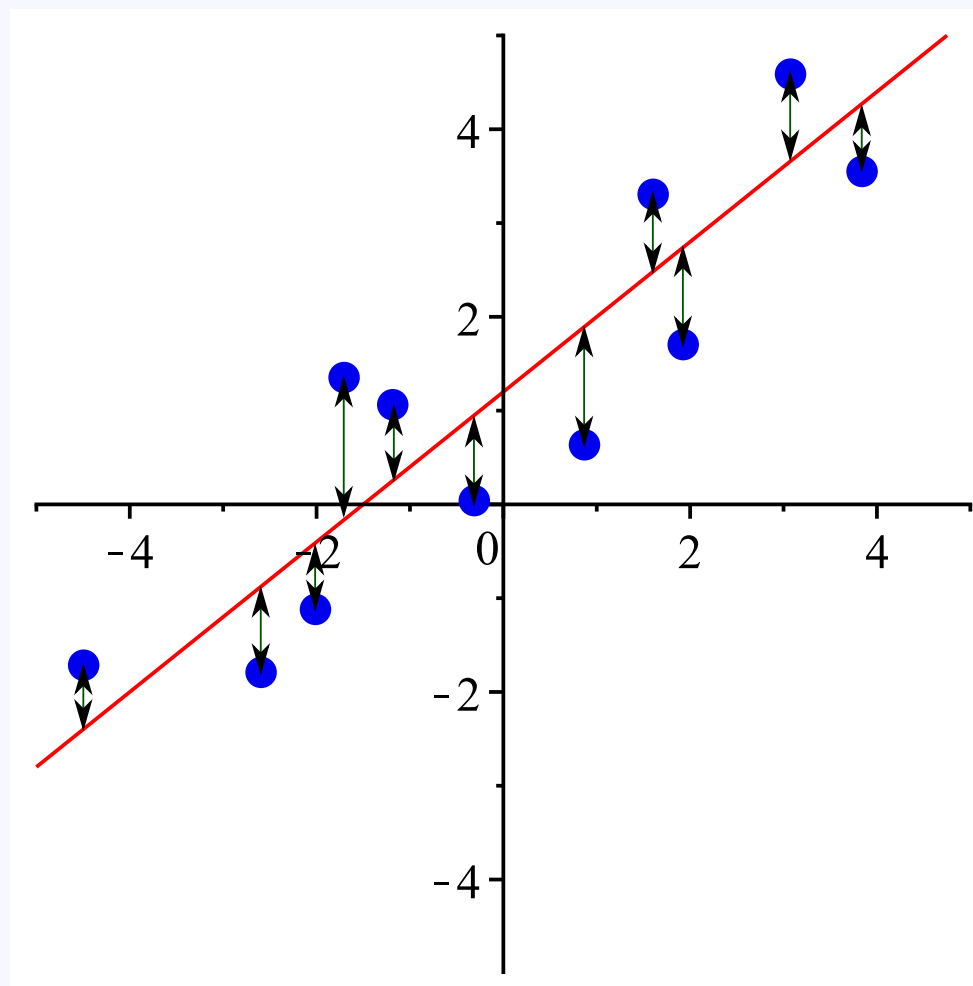
★ 誤差は互いに無相関. つまり,  $E[\varepsilon_i \varepsilon_j] = 0, \quad i \neq j$

★ 残差二乗和

$$S(\beta) = \sum_{k=1}^n (y_k - f(x_k, \beta))^2$$

を最小化する未知パラメータベクトル  $\beta$  を最小二乗推定量  $\hat{\theta}$  とする

# 絵で見る最小二乗法



緑の線の長さの二乗和を最小化するように、未知パラメータ  $\theta$  を推定

## 線形最小二乗法の定義, および, 性質 3

★ 最小二乗推定量  $\hat{\theta}$  は, 最良線形不偏推定量である

★  $E[\hat{\theta}] = \theta$  (不偏)

★  $\hat{\theta}$  は,  $y_j$  について線形の式で書ける (線形)

★ その中で, 分散がある意味で最小 (最良)

★ 任意の不偏性と線形性を満たす  $\beta$  に対して,  $\text{Cov}[\beta] - \text{Cov}[\hat{\theta}]$  が非負定値

★ 誤差  $\varepsilon$  が正規分布に従うとき, 最小二乗推定量  $\hat{\theta}$  は, 最尤推定量である

★ つまり,  $x_1, \dots, x_n$  を固定して, 測定結果として  $y_1, \dots, y_n$  が得られる確率を  $\theta$  の関数として考えたとき, その確率の値が最大となるのが  $\theta = \hat{\theta}$  のとき

# 最小二乗法推定量 (その1)

## ★ 方針

### ★ 残差二乗和

$$S(\beta) = \sum_{k=1}^n (y_k - f(x_k, \beta))^2$$

を最小化したいのだから,  $\beta_0, \beta_1, \dots, \beta_{m-1}$  で偏微分して0になる  $\beta$  を見つければ良い



# 最小二乗法推定量 (その1)

★  $f(x, \beta) = \beta_1 x + \beta_0$  の場合

★  $S(\beta) = \sum_{k=1}^n (y_k - \beta_1 x_k - \beta_0)^2$  であるから

$$\star \frac{\partial}{\partial \beta_1} S(\beta) = 2 \sum_{k=1}^n (x_k^2 \beta_1 + x_k \beta_0 - x_k y_k) = 0$$

$$\star \frac{\partial}{\partial \beta_0} S(\beta) = 2 \sum_{k=1}^n (x_k \beta_1 + \beta_0 - y_k) = 0$$

★ つまり、次の連立一次方程式を解けば良い

$$\star \begin{pmatrix} \sum x_k^2 & \sum x_k \\ \sum x_k & n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_0 \end{pmatrix} = \begin{pmatrix} \sum x_k y_k \\ \sum y_k \end{pmatrix}$$

$$\star \beta_1 = \frac{n \sum x_j y_j - \sum x_j \sum y_j}{n \sum x_j^2 - (\sum x_j)^2}, \quad \beta_0 = \frac{\sum x_j^2 \sum y_j - \sum x_j y_j \sum x_j}{n \sum x_j^2 - (\sum x_j)^2}$$

## 最小二乗法推定量 (その2)

★  $f(x, \beta) = \beta_0 f_0(x) + \beta_1 f_1(x) + \cdots + \beta_{m-1} f_{m-1}(x)$  の場合

★  $S(\beta) = \sum_{k=1}^n \left( y_k - \sum_{j=0}^{m-1} \beta_j f_j(x_k) \right)^2$  であるから

★  $\frac{\partial}{\partial \beta_i} S(\beta) = 2 \sum_{k=1}^n f_i(x_k) \left( \left( \sum_{j=0}^{m-1} f_j(x_k) \beta_j \right) - y_k \right) = 0$

★  $\sum_{j=0}^{m-1} \sum_{k=1}^n f_i(x_k) f_j(x_k) \beta_j = \sum_{k=1}^n f_i(x_k) y_k$

# 正規方程式

★ つまり，連立一次方程式  $B\beta = b$  を解けば良い

$$\star B \in M_m(\mathbb{R}), \quad B_{ij} = \sum_{k=1}^n f_i(x_k) f_j(x_k)$$

$$\star b \in \mathbb{R}^m, \quad b_i = \sum_{k=1}^n f_i(x_k) y_k$$

★ 行列  $B$  がフルランクであれば，最小二乗推定量が一意に定まる

★  $B\beta = b$  は正規方程式と呼ばれる

★ 数値計算する際は，この方程式を直接解くよりも高精度な方法が存在する

# 正規方程式

★ 行列  $A \in M_{n,m}(\mathbb{R})$  を以下で定義 (ヤコビアン, データ行列)

$$★ A_{ij} = f_j(x_i) = \frac{\partial}{\partial \beta_j} f(x_i)$$

$$★ B = A^T A$$

$$★ b = A^T y \quad (\text{ただし } y = (y_1 \cdots y_n)^T)$$

★ **正規方程式**は以下のように書き直される

$$★ A^T A \beta = A^T y$$

★ 行列  $A$  が列フルランクの場合

$$★ \text{最小二乗推定量は } \hat{\theta} = (A^T A)^{-1} A^T y$$

## 補足：そもそも最初から行列とベクトルで

★ 最小化したい残差二乗和は

$$S(\beta) = \sum_{k=1}^n (y_k - f(x_k, \beta))^2 = (A\beta - y, A\beta - y) = \|A\beta - y\|_2^2$$

★  $\beta$  で微分すると以下：これが0になるとおくと、正規方程式を得る

$$2A^T A\beta - 2A^T y$$

★ 補足1：

$$\begin{aligned} (A\beta - y, A\beta - y) &= (A\beta, A\beta) - 2(A\beta, y) + (y, y) \\ &= \beta^T A^T A\beta - 2(A^T y)^T \beta + y^T y \end{aligned}$$

★ 補足2（ベクトルで微分する）：

$$\frac{df}{d\beta} = \left( \frac{df}{d\beta_0} \cdots \frac{df}{d\beta_{m-1}} \right)^T$$

$$\star \frac{d}{dx}(a^T x) = a, \quad \frac{d}{dx}(x^T A x) = (A + A^T)x$$

# QR分解を用いて解く

★ 行列  $A$  は列フルランクでQR分解できたとする

★  $A = QR$

★  $Q \in M_{n,m}(\mathbb{R})$  は列ベクトルが長さ1で互いに直交

★  $R \in M_m(\mathbb{R})$  は正則な上三角行列

★ このとき、正規方程式は

★  $A^T A \beta = A^T y$

★  $(QR)^T QR \beta = (QR)^T y$

★  $R^T Q^T QR \beta = R^T Q^T y$

★  $R^T R \beta = R^T Q^T y$

$(Q^T Q = I)$

★  $R \beta = Q^T y$

$(R^T \text{は正則})$

★  $R$  は上三角行列であるから、これは簡単に解ける

## 行列 $A$ が列フルランクでない場合

★ 行列  $A$  が列フルランクでない場合は、最小二乗推定量は一意に定まらない  
(これは**そもそもナンセンスな場合が多い**)

★ 最小二乗推定量の中で、 $\|\beta\|_2$  を最小とするものを求めることが多い

$$\star \|\beta\|_2 = \|\beta\| = \sqrt{\beta_0^2 + \beta_1^2 + \cdots + \beta_{m-1}^2} = \sqrt{\beta^T \beta}$$

★ 結論を言うと、 $A$  の Moore–Penrose の一般逆行列を  $A^+$  と書くと  $A^+ y = R^+ Q^T y$  が答え

★ ある程度ロバストに計算できる方法は特異値分解

★ 高速に計算するなら完全ピボット選択付き  $QR$  分解をして直交変換

# 一般逆行列

- ★ 正則でなくても、長方形でも良い行列  $A \in M_{mn}(\mathbb{R})$  に対して、 $AXA = A$  を満たす行列  $X \in M_{nm}(\mathbb{R})$  を一般逆行列といい  $A^-$  で表す
- ★  $A^-$  は必ず存在し、一般的には  $A^-$  は一意ではなく複数存在する
- ★ 連立一次方程式  $Ax = b$  の解の一つは、存在するならば  $x = A^-b$  と書ける
- ★ 連立一次方程式  $Ax = b$  の解は、存在するならば、任意のベクトル  $y$  を用いて  $x = A^-b + (I - A^-A)y$  と書ける
- ★ 連立一次方程式  $Ax = b$  は  $(I - AA^-)b = 0$  ならば解が存在する



# Moore–Penrose の一般逆行列

- ★ 正則でなくても、長方形でも良い行列  $A \in M_{mn}(\mathbb{R})$  に対して、 $AXA = A, XAX = X, (AX)^T = AX, (XA)^T = XA$  を満たす行列  $X \in M_{nm}(\mathbb{R})$  を Moore–Penrose の一般逆行列といい  $A^+$  で表す
- ★  $A^+$  は必ず存在し、一意である
- ★ 連立一次方程式  $Ax = b$  の解が存在するならば、その中で  $\|x\|_2$  が最小となるものは  $x = A^+b$  となる
- ★ 連立一次方程式  $Ax = b$  の解が存在しなければ、 $\|Ax - b\|_2$  が最小とするのは  $x = A^+b$  となる

## 演習 - ビールの売上の予測

# 概要と目的

- ★ 気温とビールの売上の関係を調べる
  - ★ 気象庁の長期予報などと組合せて将来のビールの売上を予想するのが目的
  - ★ 製造や仕入れなどに有効活用できる可能性

# 使用するデータ

★ 東京の日平均気温の月平均値

★ [http://www.data.jma.go.jp/obd/stats/etrn/view/monthly\\_s3.php?prec\\_no=44&block\\_no=47662](http://www.data.jma.go.jp/obd/stats/etrn/view/monthly_s3.php?prec_no=44&block_no=47662)

★ 京都の日平均気温の月平均値

★ [http://www.data.jma.go.jp/obd/stats/etrn/view/monthly\\_s3.php?prec\\_no=61&block\\_no=47759](http://www.data.jma.go.jp/obd/stats/etrn/view/monthly_s3.php?prec_no=61&block_no=47759)

★ アサヒグループホールディングスの月次販売情報

★ [https://www.asahigroup-holdings.com/ir/financial\\_data/monthly\\_data.html](https://www.asahigroup-holdings.com/ir/financial_data/monthly_data.html)

★ 上のデータを整形して作ったcsvファイルその1（このファイルを使用して演習を行います）

★ [http://ds.k.kyoto-u.ac.jp/e-learning\\_files/data\\_analysis\\_basic/jma\\_001.csv](http://ds.k.kyoto-u.ac.jp/e-learning_files/data_analysis_basic/jma_001.csv)

★ PandAのリソースにも置いてあります

# ファイルを開いてみましょう

★ 前ページの csv ファイルを Excel で開いて内容を確認してみましょう

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		ビール	東京												
2	2011年1月	475	5.1												
3	2011年2月	625	7												
4	2011年3月	800	8.1												
5	2011年4月	960	14.5												
6	2011年5月	730	18.5												
7	2011年6月	980	22.8												
8	2011年7月	1295	27.3												
9	2011年8月	1135	27.5												
10	2011年9月	830	25.1												
11	2011年10月	805	19.5												
12	2011年11月	840	14.9												
13	2011年12月	1375	7.5												
14	2012年1月	480	4.8												
15	2012年2月	610	5.4												
76	2017年3月	796	8.5												
77	2017年4月	784	14.7												
78															

## 単回帰分析を試みよう

★ 単回帰分析を行うことで、ビールの売上と東京の気温との関係を調べてみましょう

★ ビールの売上を  $B$ 、東京の気温を  $T$  として、 $B = aT + b + \varepsilon$  という回帰モデル

---

★ Excel を用いて回帰分析を行う方法はいくつかあるが、ここではアドインの「分析ツール」を用いる

★ GUI で操作できる、結果に色々表示される

## アドインの追加

★ アドインの追加は例えば以下の手順で行います

★ ファイル → オプション → アドイン → 設定 → 分析ツールにチェックを入れて OK を押す

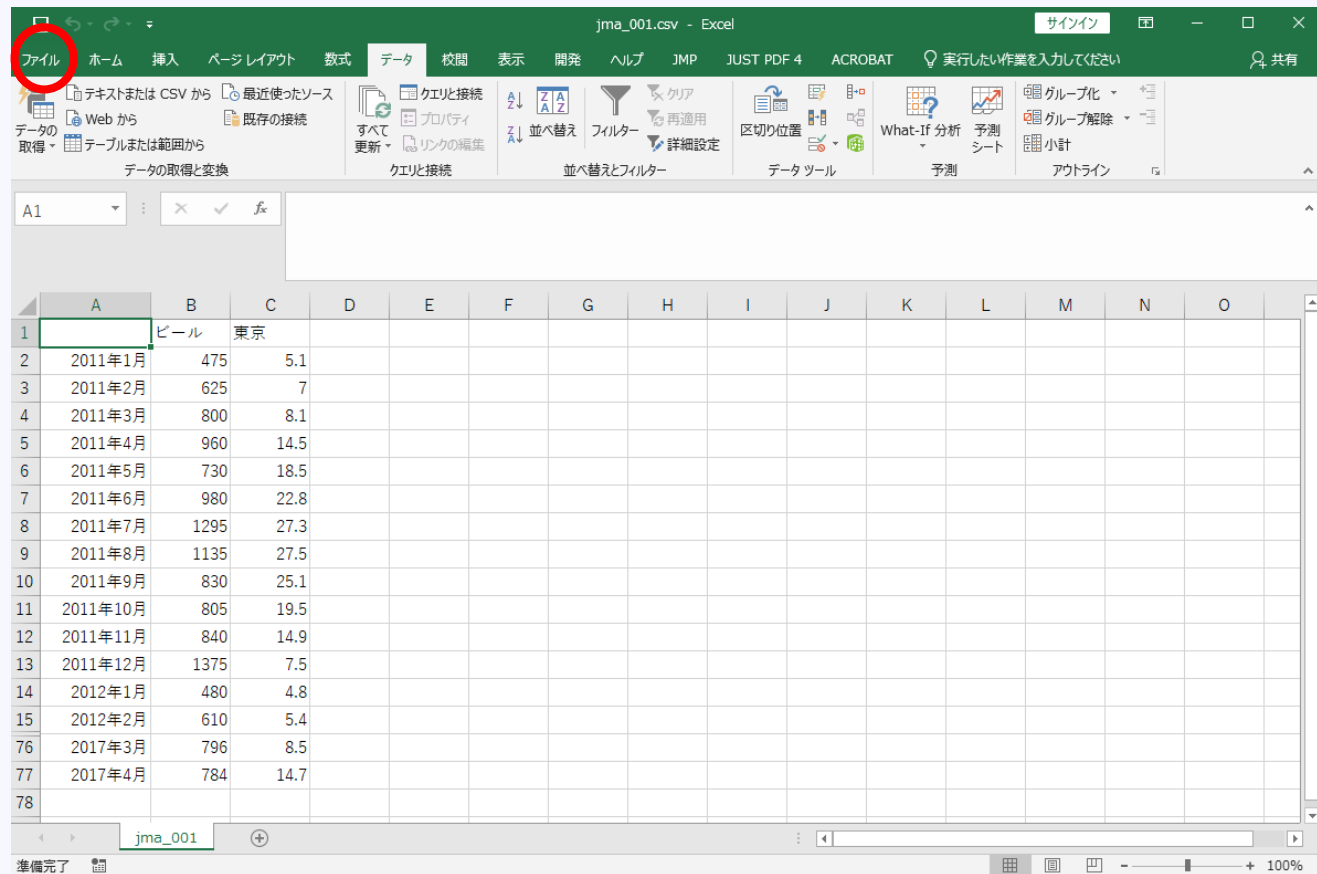
★ 成功するとリボンのデータのタブにデータ分析が表示されます

# アドインの追加

★ アドインの追加は例えば以下の手順で行います

★ **ファイル** → オプション → アドイン → 設定 → 分析ツールにチェックを入れて OK を押す

★ 成功するとリボンのデータのタブにデータ分析が表示されます





# アドインの追加

★ アドインの追加は例えば以下の手順で行います

★ ファイル → オプション → アドイン → 設定 → 分析ツールにチェックを入れて OK を押す

★ 成功するとリボンのデータのタブにデータ分析が表示されます



The screenshot shows the Adobe Analytics interface for a document named 'jma\_001'. On the left, a green sidebar contains navigation options, with 'オプション' (Options) highlighted by a red circle. The main area is titled '情報' (Information) and lists several settings:

- ブックの保護** (Book Protection): This block allows managing the types of changes users can make to this book.
- ブックの検査** (Book Check): Before publishing a file, check the following items:
  - Document properties, absolute paths
  - Non-display
  - Current file format may cause issues related to accessibility that cannot be confirmed.
- ブックの管理** (Book Management): Last updated today at 9:56 (auto-recovery).
- ブラウザーの表示オプション** (Browser Display Options): Choose the content to be displayed when the book is viewed in a browser.

On the right, document properties are listed:

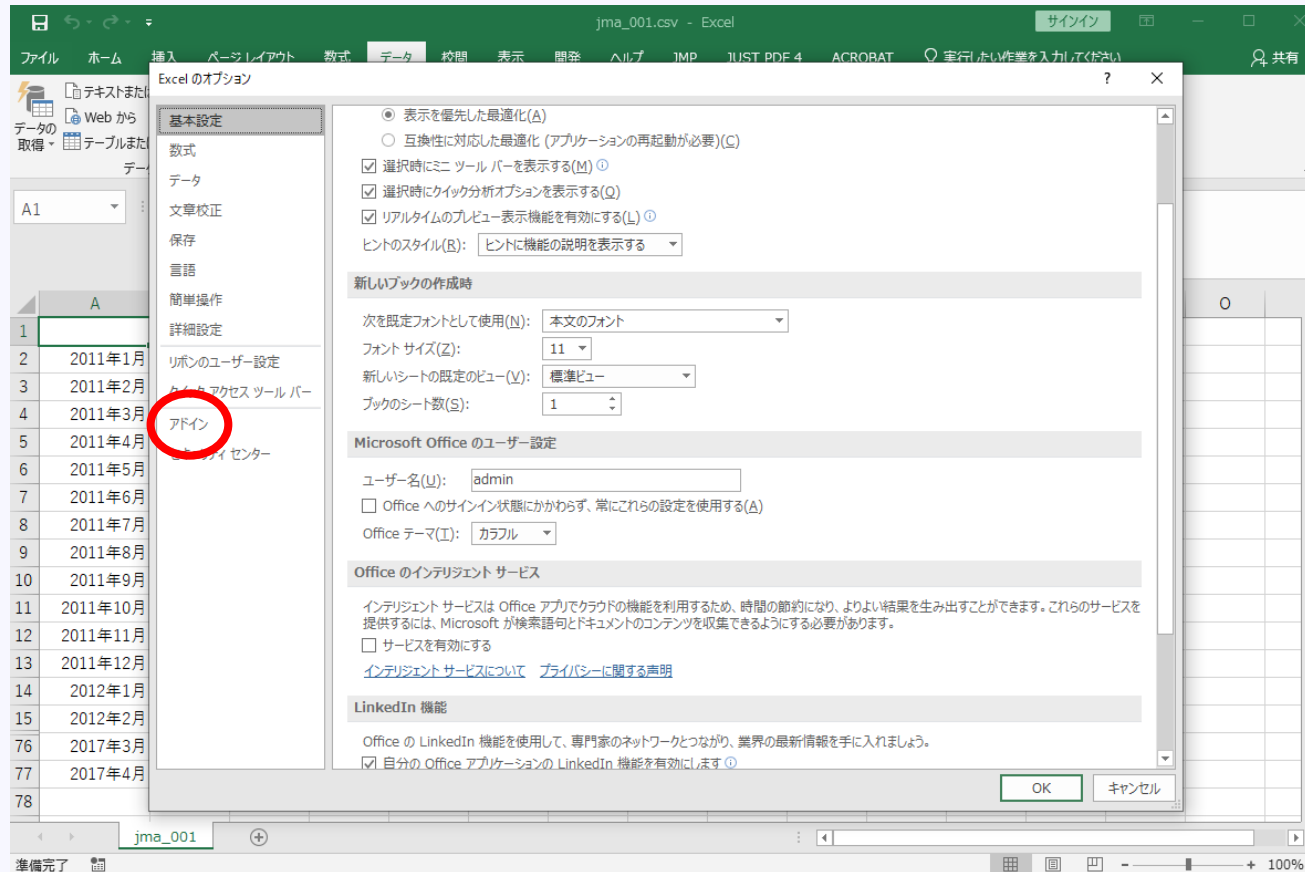
- Size: 1.50KB
- Title: Add title
- Tags: Add tags
- Categories: Add categories
- Related Date: Updated today at 9:56
- Created Date:
- Final Print Date:
- Related User: Created by admin, Last updated by admin
- Related Documents:  Open file storage location, Show all properties

# アドインの追加

★ アドインの追加は例えば以下の手順で行います

★ ファイル → オプション → **アドイン** → 設定 → 分析ツールにチェックを入れて OK を押す

★ 成功するとリボンのデータのタブにデータ分析が表示されます

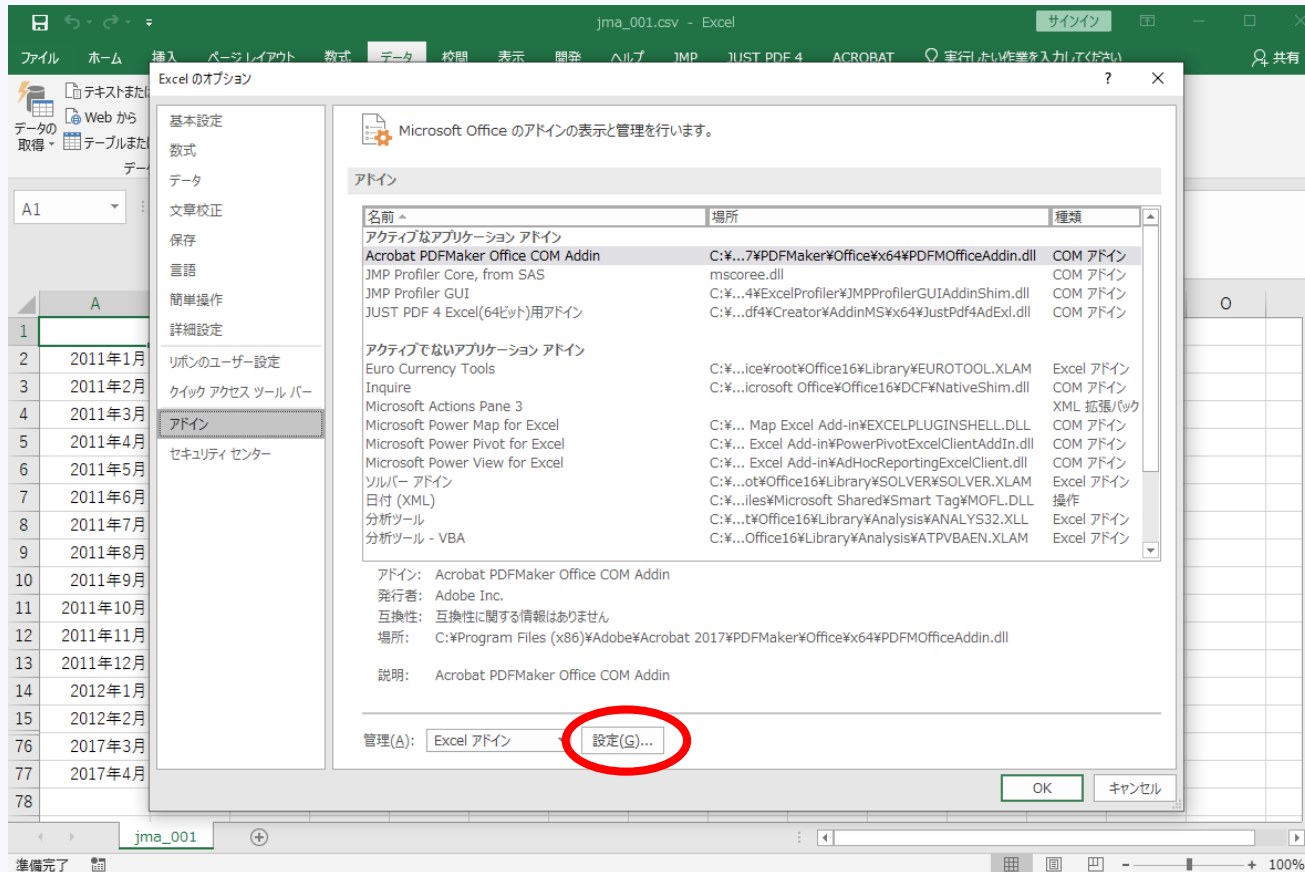


# アドインの追加

★ アドインの追加は例えば以下の手順で行います

★ ファイル → オプション → アドイン → **設定** → 分析ツールにチェックを入れて OK を押す

★ 成功するとリボンのデータのタブにデータ分析が表示されます

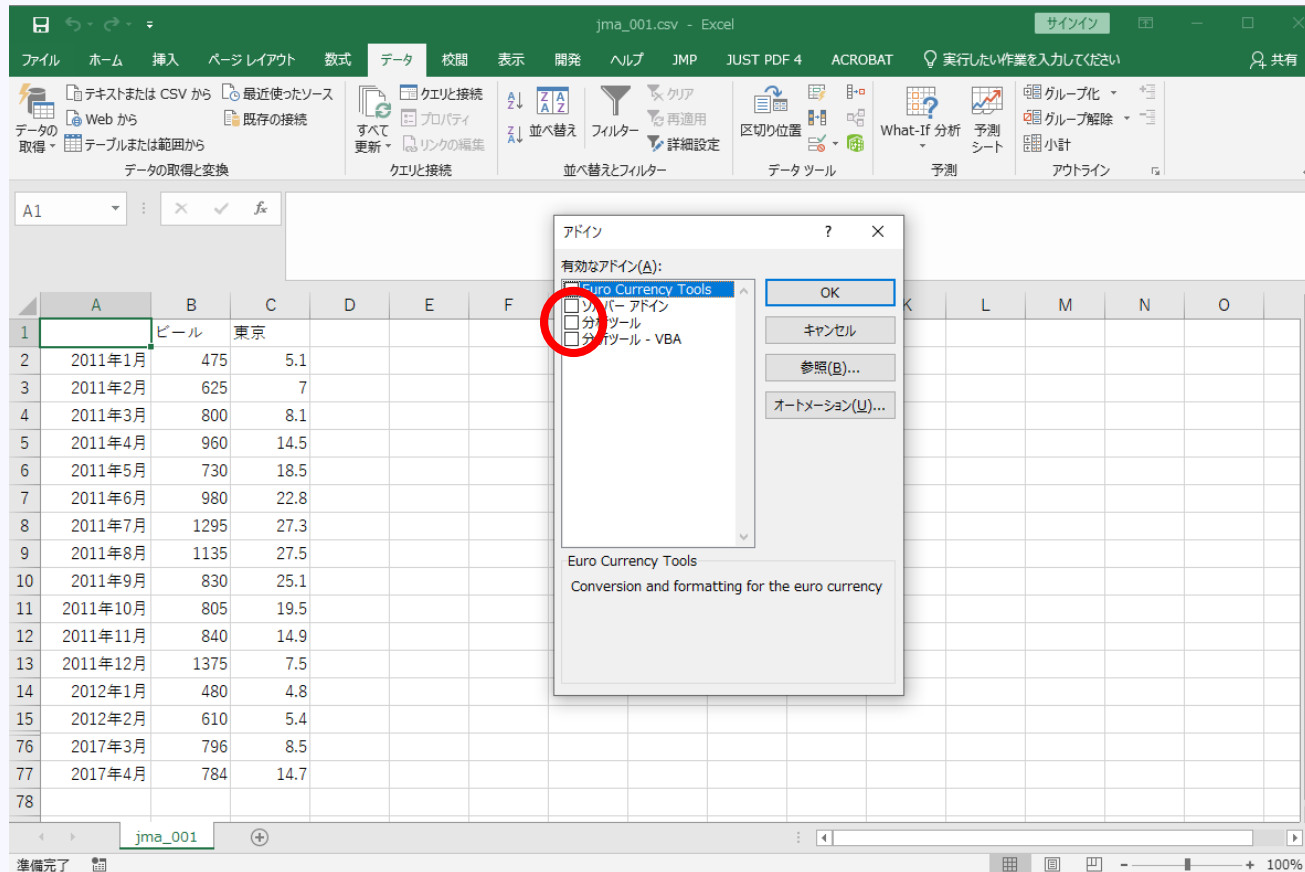


# アドインの追加

★ アドインの追加は例えば以下の手順で行います

★ ファイル → オプション → アドイン → 設定 → **分析ツールにチェック**を入れて OK を押す

★ 成功するとリボンのデータのタブにデータ分析が表示されます

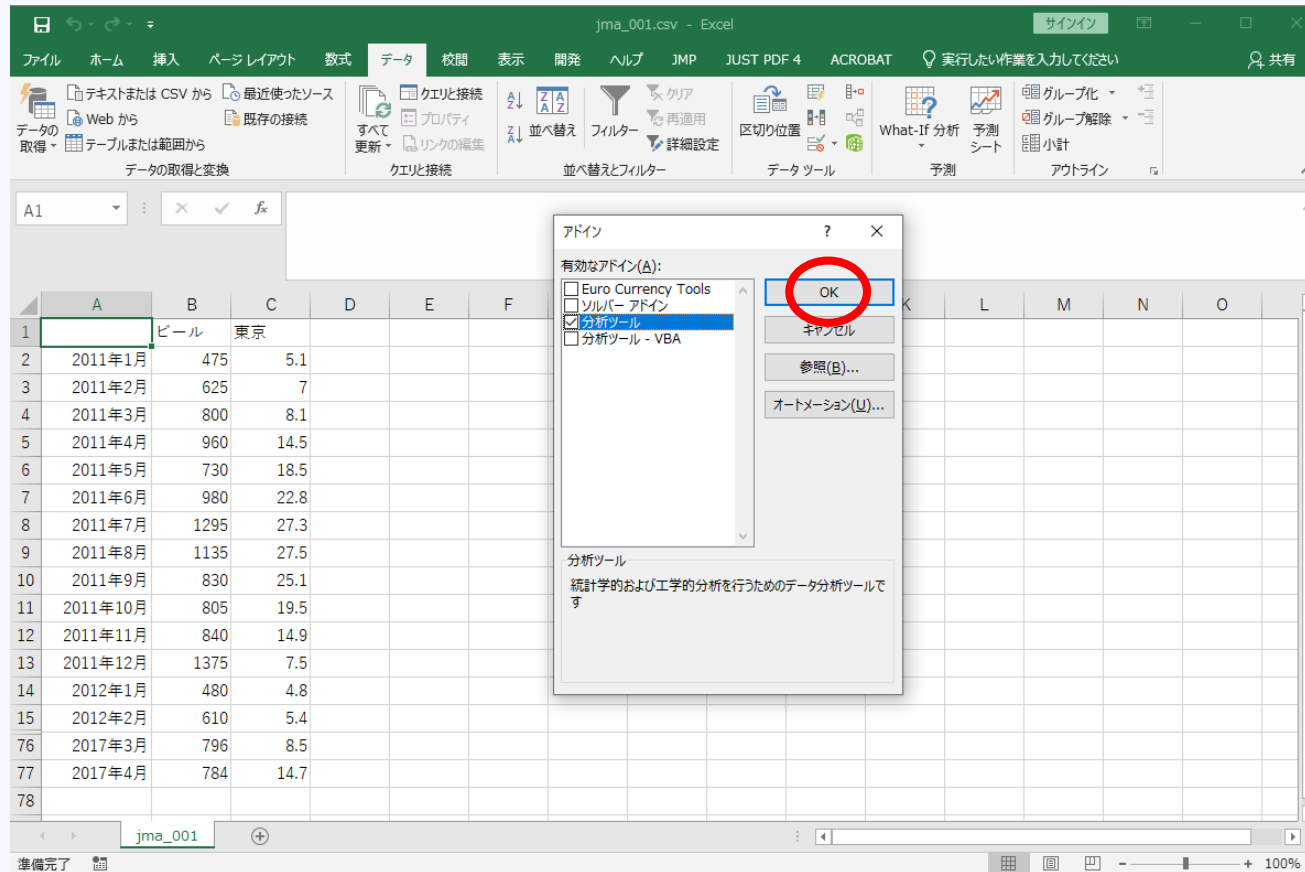


# アドインの追加

★ アドインの追加は例えば以下の手順で行います

★ ファイル → オプション → アドイン → 設定 → 分析ツールにチェックを入れて **OK** を押す

★ 成功するとリボンのデータのタブにデータ分析が表示されます

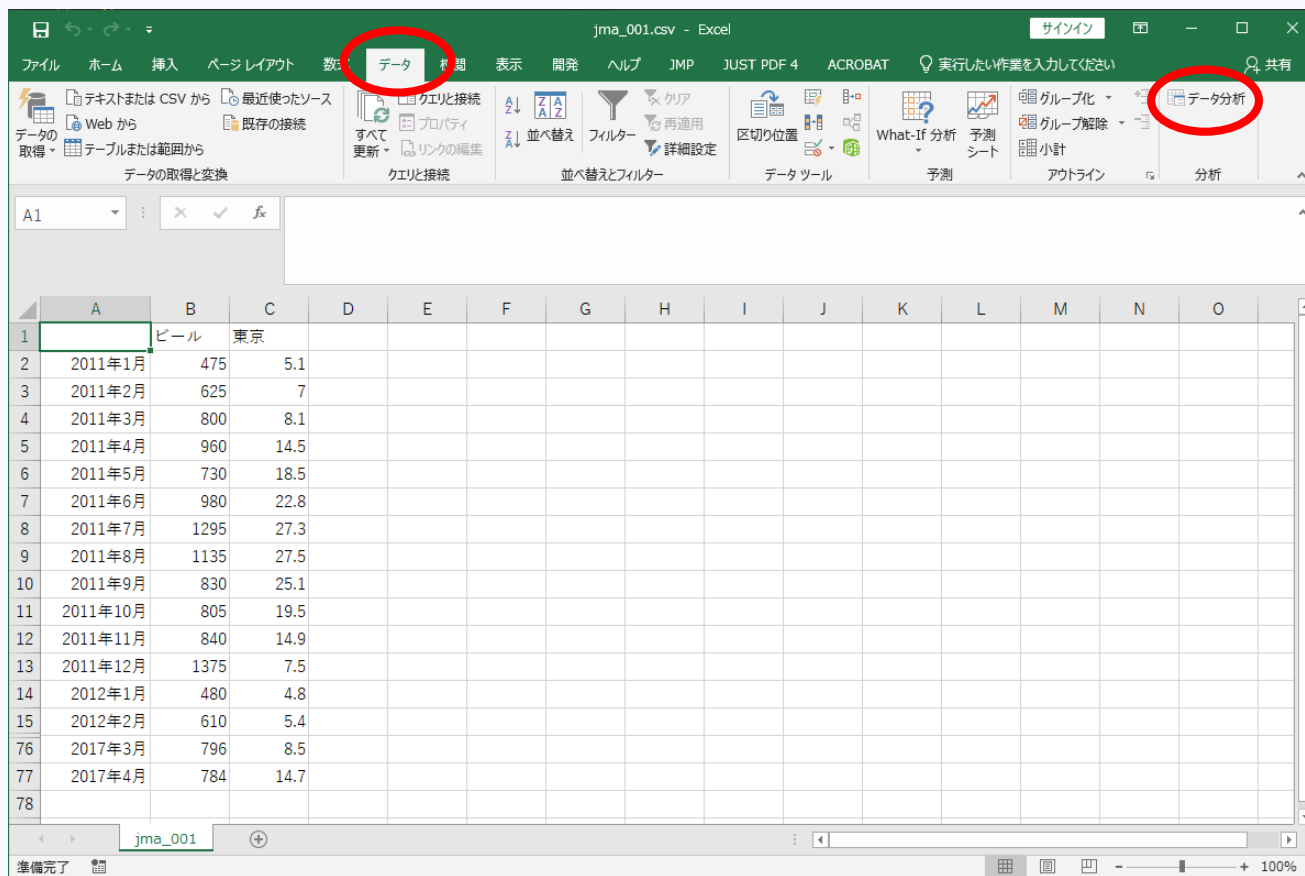


# アドインの追加

★ アドインの追加は例えば以下の手順で行います

★ ファイル → オプション → アドイン → 設定 → 分析ツールにチェックを入れて OK を押す

★ 成功するとリボンのデータのタブにデータ分析が表示されます



# 単回帰分析の実行

★ 回帰分析の実行は以下の手順で行います

★ 1. データ分析をクリックし，回帰分析を選び，OK を押す

★ 2. 入力 Y 範囲，入力 X 範囲などを適切に記入し，OK を押すことで，回帰分析を行う

★ 2-1. 入力 Y 範囲には「\$B\$1:\$B\$77」と入力（\$はなくても構いません）

★ 2-2. 入力 X 範囲には「\$C\$1:\$C\$77」と入力（\$はなくても構いません）

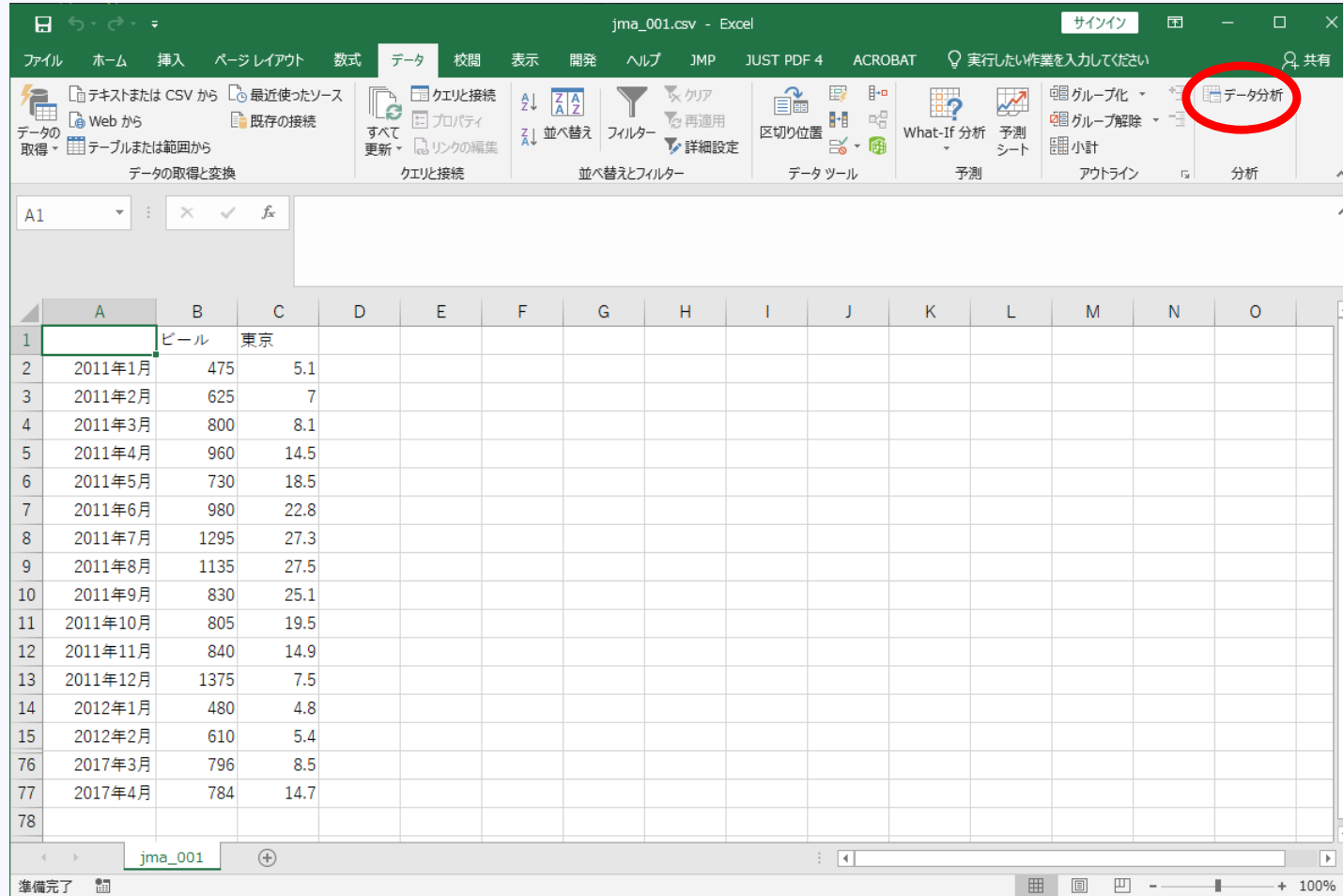
★ 2-3. 「ラベル」にチェックを入れる

★ 2-4. OK を押す

# 単回帰分析の実行

★ 回帰分析の実行は以下の手順で行います

★ 1. データ分析をクリックし，回帰分析を選び，OK を押す

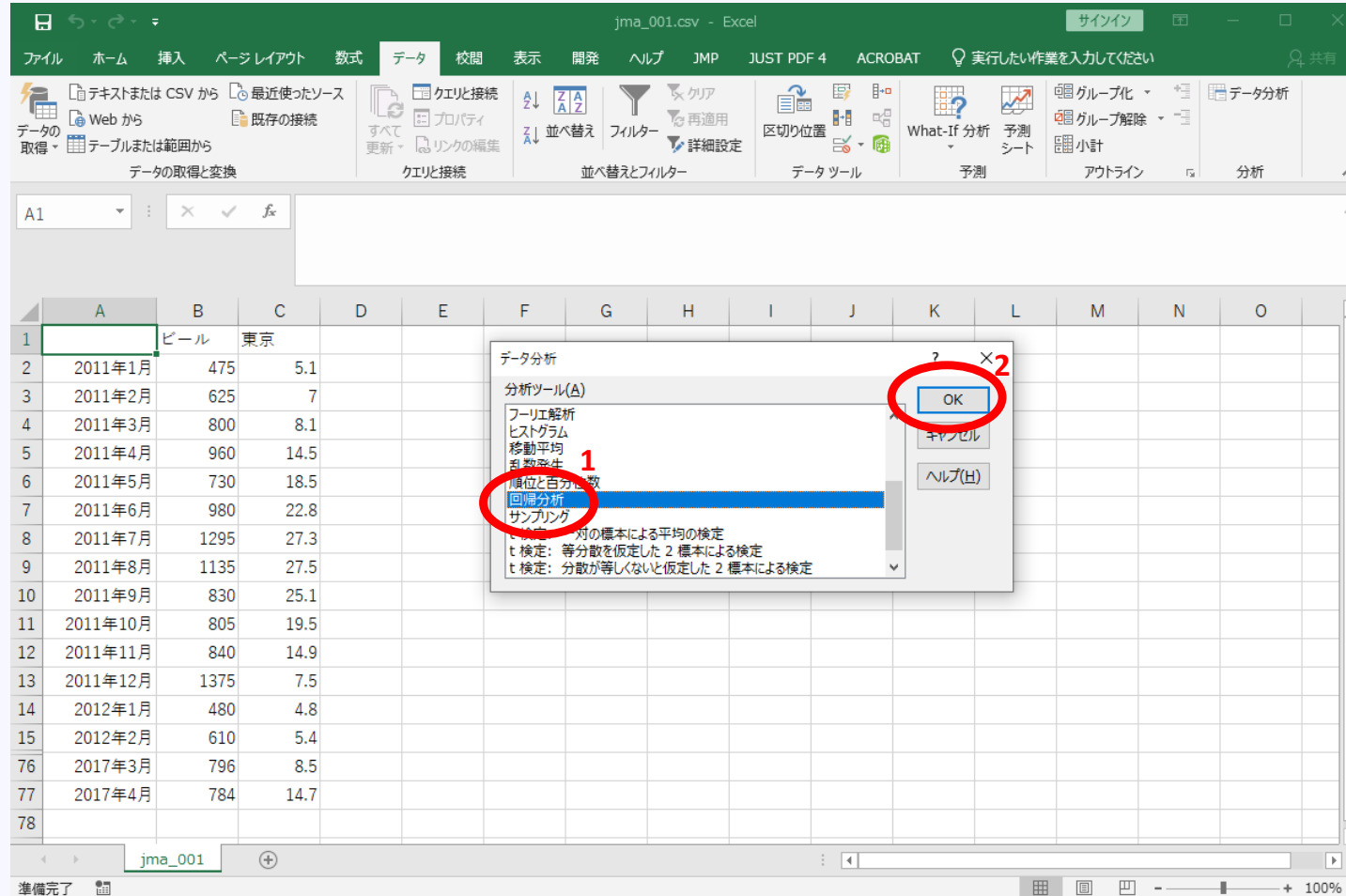




# 単回帰分析の実行

★ 回帰分析の実行は以下の手順で行います

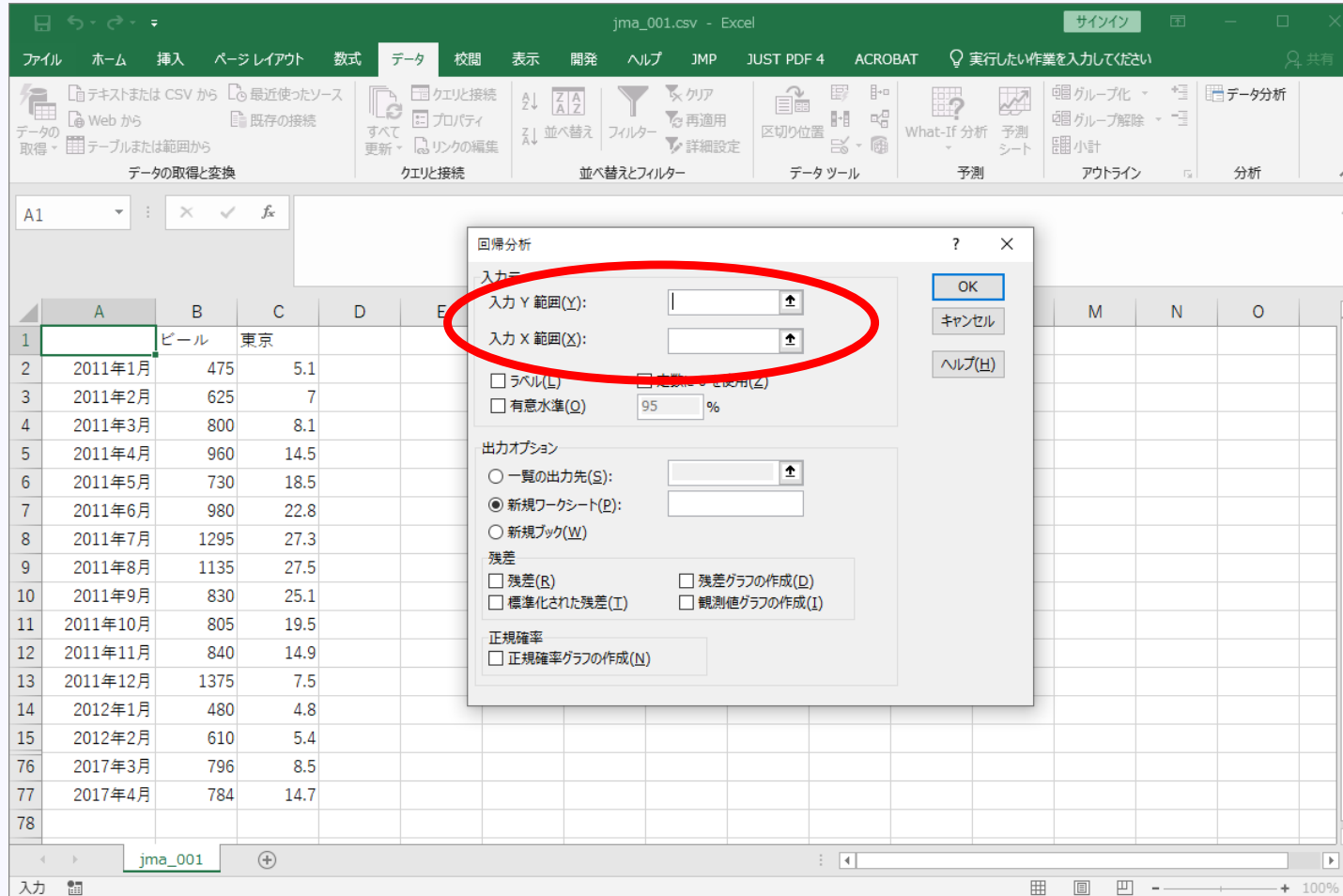
★ 1. データ分析をクリックし，回帰分析を選び，OK を押す



# 単回帰分析の実行

★ 回帰分析の実行は以下の手順で行います

★ 2. 入力 Y 範囲, 入力 X 範囲などを適切に記入し, OK を押すことで, 回帰分析を行う



# 単回帰分析の実行

★ 回帰分析の実行は以下の手順で行います

★ 2-1. 入力 Y 範囲には「\$B\$1:\$B\$77」と入力（\$はなくても構いません）

The screenshot shows the '回帰分析' (Regression Analysis) dialog box in Excel. The '入力 Y 範囲(Y):' field is set to '\$B\$1:\$B\$77', which is highlighted by a red arrow. The dialog box also includes options for '入力 X 範囲(X):', 'ラベル(L)', '有意水準(Q)', '出力オプション', '残差', and '正規確率'.

	A	B	C	D	E
1		ビール	東京		
2	2011年1月	475	5.1		
3	2011年2月	625	7		
4	2011年3月	800	8.1		
5	2011年4月	960	14.5		
6	2011年5月	730	18.5		
7	2011年6月	980	22.8		
8	2011年7月	1295	27.3		
9	2011年8月	1135	27.5		
10	2011年9月	830	25.1		
11	2011年10月	805	19.5		
12	2011年11月	840	14.9		
13	2011年12月	1375	7.5		
14	2012年1月	480	4.8		
15	2012年2月	610	5.4		
76	2017年3月	796	8.5		
77	2017年4月	784	14.7		
78					

# 単回帰分析の実行

★ 回帰分析の実行は以下の手順で行います

★ 2-2. 入力 X 範囲には「\$C\$1:\$C\$77」と入力（\$はなくても構いません）

回帰分析

入力元

入力 Y 範囲(Y): \$B\$1:\$B\$77

入力 X 範囲(X): \$C\$1:\$C\$77

ラベル(L)       定数に 0 を使用(Z)

有意水準(Q)      95 %

出力オプション

一覧の出力先(S):

新規ワークシート(P):

新規ブック(W)

残差

残差(R)       残差グラフの作成(D)

標準化された残差(I)       観測値グラフの作成(L)

正規確率

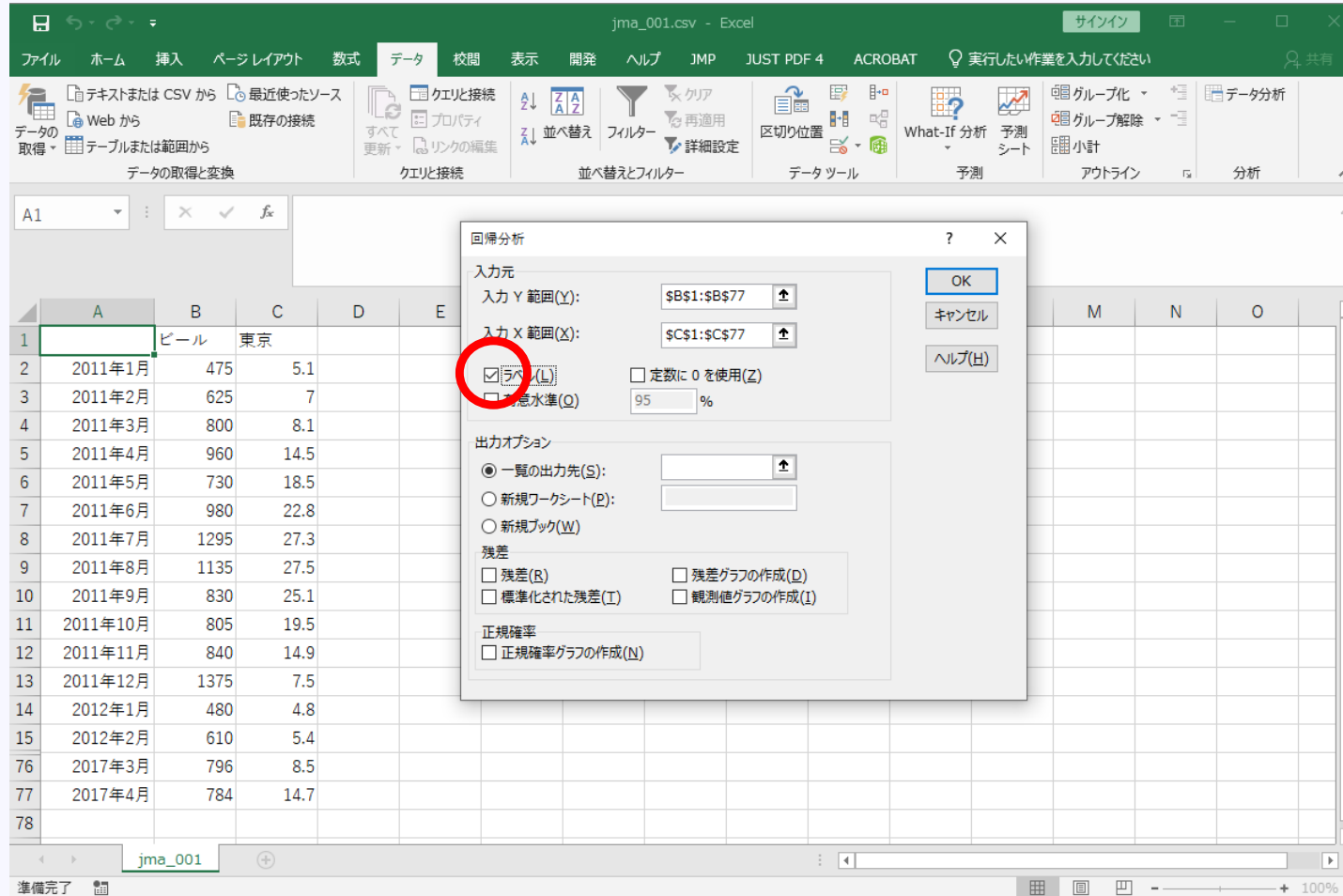
正規確率グラフの作成(N)

	A	B	C	D	E
1		ビール	東京		
2	2011年1月	475	5.1		
3	2011年2月	625	7		
4	2011年3月	800	8.1		
5	2011年4月	960	14.5		
6	2011年5月	730	18.5		
7	2011年6月	980	22.8		
8	2011年7月	1295	27.3		
9	2011年8月	1135	27.5		
10	2011年9月	830	25.1		
11	2011年10月	805	19.5		
12	2011年11月	840	14.9		
13	2011年12月	1375	7.5		
14	2012年1月	480	4.8		
15	2012年2月	610	5.4		
76	2017年3月	796	8.5		
77	2017年4月	784	14.7		
78					

# 単回帰分析の実行

★ 回帰分析の実行は以下の手順で行います

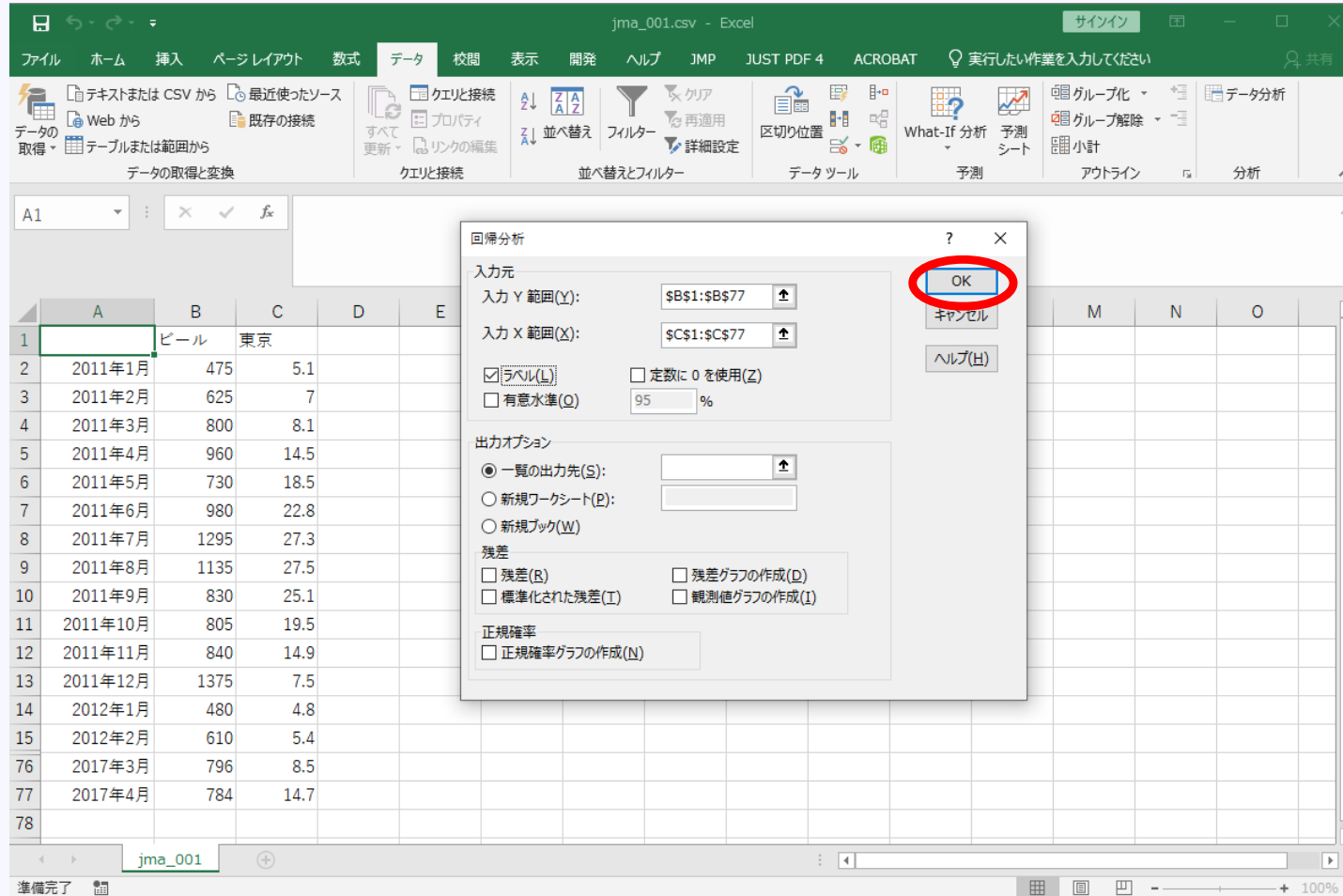
★ 2-3. 「ラベル」にチェックを入れる



# 単回帰分析の実行

★ 回帰分析の実行は以下の手順で行います

★ 2-4. OK を押す



# 単回帰分析の実行結果

★ 正しく実行できたら以下のような結果が別のシートに作られ表示されます

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	概要															
2																
3	回帰統計															
4	重相関 R	0.431346														
5	重決定 R2	0.186059														
6	補正 R2	0.17506														
7	標準誤差	212.299														
8	観測数	76														
9																
10	分散分析表															
11		自由度	変動	分散	割された分散	有意 F										
12	回帰	1	762406.1	762406.1	16.91572	0.0001										
13	残差	74	3335243	45070.86												
14	合計	75	4097650													
15																
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%							
17	切片	656.9307	56.52093	11.62279	2.33E-18	544.3103	769.5511	544.3103	769.5511							
18	東京	12.99158	3.15876	4.112873	0.0001	6.697612	19.28555	6.697612	19.28555							

# 実行結果を読み取る

★  $B = aT + b + \varepsilon$  の係数  $a, b$  が以下のように推定されたことがわかる

★  $a = 12.99158, b = 656.9307$

回帰統計						
重相関 R	0.431346					
重決定 R2	0.186059					
補正 R2	0.17506					
標準誤差	212.299					
観測数	76					

分散分析表						
	自由度	変動	分散	F	有意 F	
回帰	1	762406.1	762406.1	16.91572	0.0001	
残差	74	3335243	45070.86			
合計	75	4097650				

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	656.9307	56.52093	11.62279	2.33E-18	544.3103	769.5511	544.3103	769.5511
東京	12.99158	3.15876	4.112873	0.0001	6.697612	19.28555	6.697612	19.28555



# 実行結果を読み取る

★ この結果から、東京の気温が1度あがるとビールの売上が13万箱ぐらい増えるだろうと推定されたことがわかる

The screenshot shows the following data from the regression analysis:

回帰統計	重相関 R	0.431346
	重決定 R2	0.186059
	補正 R2	0.17506
	標準誤差	212.299
	観測数	76

分散分析表	自由度	変動	分散	F	P-値	有意 F
回帰	1	762406.1	762406.1	16.91572	0.0001	
残差	74	3335243	45070.86			
合計	75	4097650				

係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%	
切片	656.9307	56.52093	11.62279	2.33E-18	544.3103	769.5511	544.3103	769.5511
東京	12.99158	3.15876	4.112873	0.0001	6.697612	19.28555	6.697612	19.28555

# 実行結果を読み取る

★ 東京の気温に対して、 $P$ 値が0.0001と表示されています

★ これは仮説検定に置いて、帰無仮説を「東京の気温に対する係数 $a = 0$ 」としたときの $P$ 値です

	自由度	変動	分散	割された分	有意 F
回帰	1	762406.1	762406.1	16.91572	0.0001
残差	74	3335243	45070.86		
合計	75	4097650			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	656.9307	56.52093	11.62276	0.0001	544.3103	769.5511	544.3103	769.5511
東京	12.99158	3.15876	4.112873	0.0001	6.67612	19.28555	6.697612	19.28555

# 実行結果を読み取る

- ★ 東京の気温に対して、下限95%が6.697612、上限95%が19.28555と表示されています
- ★ これは、係数  $a$  に対して、95%の確率で  $6.69 \leq a \leq 19.28$  であるという意味
- ★  $X$  の95%信頼区間が  $[L, U]$  であるとは、 $P(L \leq X \leq U) \geq 0.95$

	自由度	変動	分散	割られた分散	有意 F
11 回帰	1	762406.1	762406.1	16.91572	0.0001
12 残差	74	3335243	45070.86		
13 合計	75	4097650			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17 切片	656.9307	56.52093	11.62279	2.33E-18	544.2102	769.5511	544.3103	769.5511
18 東京	12.99158	3.15876	4.112873	0.001	6.697612	19.28555	6.697612	19.28555

# 実行結果を読み取る

★ 重決定 R2 の値が 0.186059 と表示されています

★ これは、ビールの売上の変化のうち、18% ぐらいが東京の気温の変動で説明できるという意味

回帰統計	
重決定 R2	0.186059
補正 R2	0.17906
標準誤差	212.299
観測数	76

分散分析表						
	自由度	変動	分散	F	P-値	有意 F
回帰	1	762406.1	762406.1	16.91572	0.0001	
残差	74	3335243	45070.86			
合計	75	4097650				

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	656.9307	56.52093	11.62279	2.33E-18	544.3103	769.5511	544.3103	769.5511
東京	12.99158	3.15876	4.112873	0.0001	6.697612	19.28555	6.697612	19.28555

# 実行結果を読み取る

★ 重決定 R2 はビールの売上の変動のうち、回帰式で説明できた割合（説明できなかったのが残差二乗和）

★ 変動は  $\sum(\text{平均からのずれ})^2$

回帰統計								
重決定 R2	0.186059							
補正 R2	0.17908							
標準誤差	212.299							
観測数	76							
分散分析表								
	自由度	変動	分散	F	有意 F			
回帰	1	762406.1	62406.1	16.91572	0.0001			
残差	4	3335243	5070.86					
合計	5	4097650						
係数								
	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	656.9307	56.52093	11.62279	2.33E-18	544.3103	769.5511	544.3103	769.5511
東京	12.99158	3.15876	4.112873	0.0001	6.697612	19.28555	6.697612	19.28555

## 考察

- ★ 東京の気温が1度上がるとビールの売上が13万箱ぐらい増えるんじゃないか
  - ★ しかし，東京の気温だけでは，ビールの打ち上げの変化のうち18%ぐらいしか説明できていない
  - ★ もうちょっと色々な要因からビールの売上を予想した方が良いのでは
- 
- ★ 東京の気温の他に，京都の気温も追加で使って予想
    - ★ 重回帰分析（説明変数の数を1つから2つに増やす）

# ファイル

★ 京都の気温のデータも入れた以下のファイルを使用

★ [http://ds.k.kyoto-u.ac.jp/e-learning\\_files/  
data\\_analysis\\_basic/jma\\_002.csv](http://ds.k.kyoto-u.ac.jp/e-learning_files/data_analysis_basic/jma_002.csv)

★ PandA のリソースにも置いてあります

# ファイルを開いてみましょう

★ 前ページの csv ファイルを Excel で開いて内容を確認してみましょう

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		ビール	東京	京都											
2	2011年1月	475	5.1	2.8											
3	2011年2月	625	7	6.3											
4	2011年3月	800	8.1	6.8											
5	2011年4月	960	14.5	12.5											
6	2011年5月	730	18.5	19											
7	2011年6月	980	22.8	24.1											
8	2011年7月	1295	27.3	27.9											
9	2011年8月	1135	27.5	28.7											
10	2011年9月	830	25.1	24.7											
11	2011年10月	805	19.5	18.4											
12	2011年11月	840	14.9	13.8											
13	2011年12月	1375	7.5	6.5											
14	2012年1月	480	4.8	4.1											
15	2012年2月	610	5.4	4.1											
76	2017年3月	796	8.5	8.2											
77	2017年4月	784	14.7	14.8											
78															



## 重回帰分析をしてみよう

- ★ 今度は、重回帰分析を行うことで、ビールの売上と東京の気温と京都の気温の関係を調べてみましょう
  - ★ ビールの売上を  $B$ 、東京の気温を  $T$ 、京都の気温を  $K$  として、 $B = aT + bK + c + \varepsilon$  という回帰モデル
- 
- ★ 実行の仕方は、単回帰分析とほぼ同様で、 $X$  の範囲として複数の行を指定すれば良い

# 重回帰分析の実行

★ 回帰分析の実行は以下の手順で行います

★ 1. データ分析をクリックし，回帰分析を選び，OK を押す

★ 2. 入力 Y 範囲，入力 X 範囲などを適切に記入し，OK を押すことで，回帰分析を行う

★ 2-1. 入力 Y 範囲には「 $\$B\$1:\$B\$77$ 」と入力（\$はなくても構いません）

★ 2-2. 入力 X 範囲には「 $\$C\$1:\$D\$77$ 」と入力（\$はなくても構いません）

★ 2-3. 「ラベル」にチェックを入れる

★ 2-4. OK を押す

# 重回帰分析の実行結果

★ 正しく実行できたら以下のような結果が別のシートに作られ表示されます

The screenshot shows the Excel interface with the 'データ' (Data) tab selected. The 'データ分析' (Data Analysis) toolpak is active. The '概要' (Summary) sheet is displayed, showing the following regression statistics:

重回帰統計	
重回帰 R	0.431415
重決定 R <sup>2</sup>	0.186119
補正 R <sup>2</sup>	0.163821
標準誤差	213.7403
観測数	76

Below the summary, an ANOVA table is shown:

	自由度	変動	分散	割された分散	有意 F
回帰	2	762649.4	381324.7	8.346836	0.000544
残差	73	3335000	45684.94		
合計	75	4097650			

At the bottom, the coefficients table is displayed:

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	653.1112	77.32065	8.446788	2.04E-12	499.0113	807.211	499.0113	807.211
東京	15.34582	32.42195	0.473316	0.6374	-49.271	79.96268	-49.271	79.96268
京都	-2.15779	29.57313	-0.07296	0.942034	-61.097	56.78137	-61.097	56.78137

# 実行結果を読み取る

★  $B = aT + bK + c + \varepsilon$  の係数  $a, b, c$  が以下のように推定されたことがわかる

★  $a = 15.23582, b = -2.15779, c = 653.1112$

The screenshot shows the following data in the regression analysis table:

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	653.1112	7.32065	8.446788	2.04E-12	499.0113	807.211	499.0113	807.211
東京	15.34582	3.42195	0.473316	0.6374	-49.271	79.96268	-49.271	79.96268
京都	-2.15779	2.157313	-0.07296	0.942034	-61.097	56.78137	-61.097	56.78137

# 実行結果を読み取る

★ この結果から以下のことが示唆される

★ 東京の気温が1度あがるとビールの売上が15万箱ぐらい増える

★ 京都の気温が1度あがるとビールの売上が2万箱ぐらい減る

★ 直感的には、京都の気温が上がるほどビールの売上が減るのは変な気がする

★ 重相関R2の値も0.186119とほとんど改善していない

★ ビールの売上の変化は、東京の気温と京都の気温を使っても18.6%ぐらいしか説明できていない

★ 東京の気温のみを使用した場合は0.186059でした

## ★ 課題1

★ 直感的に反する結果となったのは何故か、また改善していないのは何故か？ 考えてみてください

# 多重共線性と主成分回帰

# 多重共線性

★ 重回帰モデル  $B = aT + bK + c + \varepsilon$

★  $B$  : ビールの売上

★  $T$  : 東京の気温

★  $K$  : 京都の気温

★ このように、説明変数間で相関がある場合、多重共線性という問題が起こり、最小二乗推定量が不安定になる

★ より正確には、データ行列の条件数が大きいと不安定になる

★  $\text{cond}(A) = \sigma_{\max}(A) / \sigma_{\min}(A)$

★ 条件数が大きいと、ちょっとした摂動で連立一次方程式の解が大きく変わる

# 多重共線性

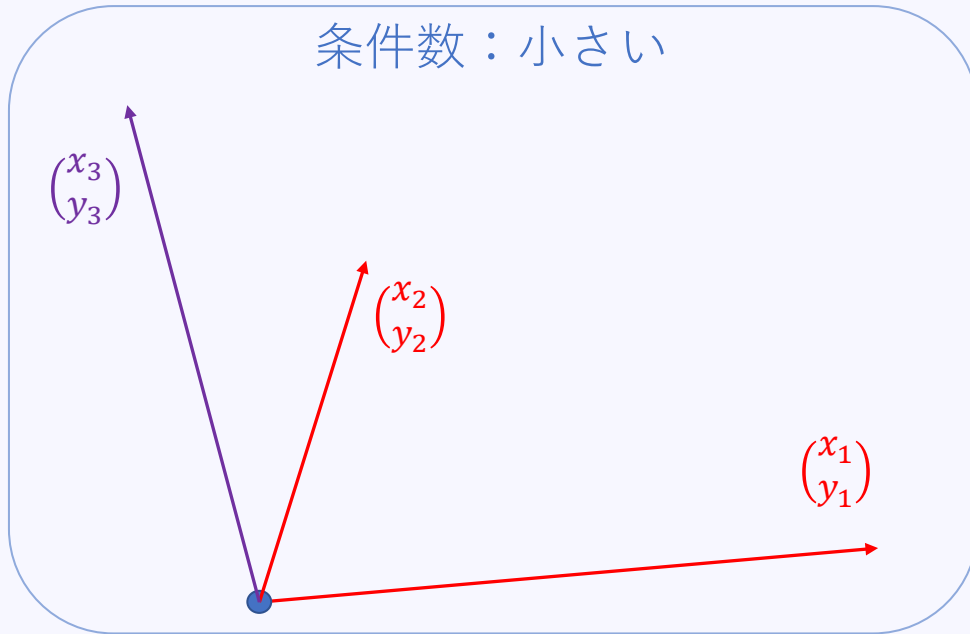
★ 連立一次方程式を次のように解釈する

$$\begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} x_3 \\ y_3 \end{pmatrix} \iff a \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + b \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_3 \\ y_3 \end{pmatrix}$$

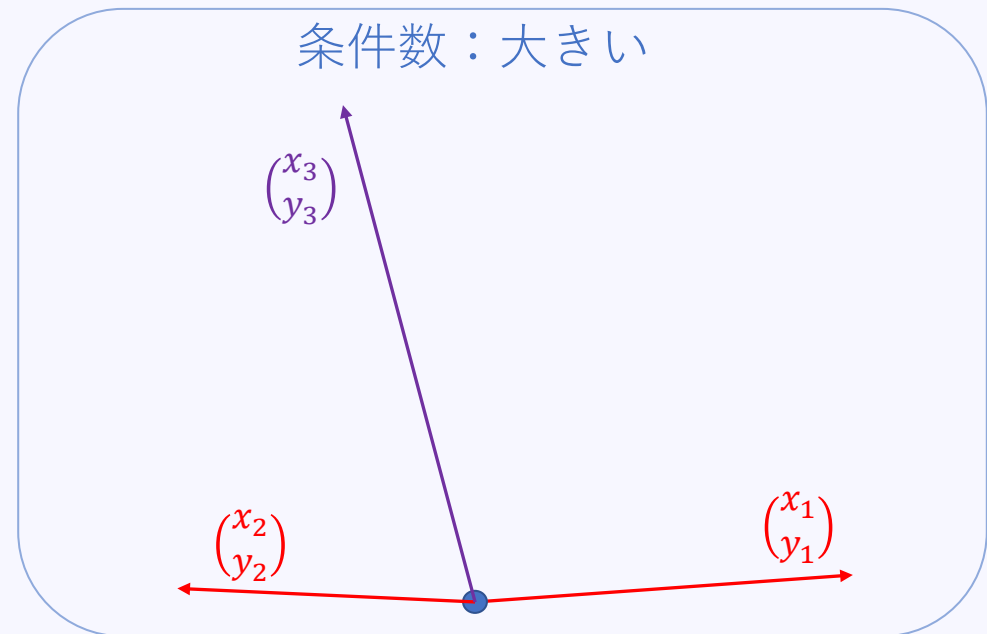
★ 2つのベクトルをどのように足せば、ベクトル  $\begin{pmatrix} x_3 \\ y_3 \end{pmatrix}^T$  を作れるか？

★ このとき、条件数が大きいというのは、2つのベクトルがほぼ線形従属であることを意味する

条件数：小さい



条件数：大きい





# 多重共線性

- ★ 回帰分析に於いても、強い相関のあるデータ（似たようなデータ、ほぼ同じ方向を向いたデータ）を用いると結果が不安定になる

- ★ 重回帰モデルを次のように書き換える

$$\begin{aligned} B &= aT + bK + c + \varepsilon \\ &= a'(T + K) + b'(T - K) + c + \varepsilon \end{aligned}$$

- ★ ただし、 $a' + b' = 2a$ ,  $a' - b' = 2b$
- ★  $T, K$  はほぼ同じ傾向を示すから、 $T - K$  は0に近い値ばかり取る
  - ★ その状況で、 $T - K$  の情報を使って「データの近くを通ろうと」すると  $T - K$  の係数  $b'$  の推定結果は絶対値が大きく不安定になる
  - ★ それが伝搬して、結果的に  $a, b$  の推定結果も不安定になる

# 多重共線性

## ★ 多重共線性の回避方法

- ★ 不安定になっているように見えたら説明変数を減らす
- ★ 説明変数を無相関にする（連立一次方程式に於いてベクトルが直交するようにする）

## ★ 説明変数を以下に取り替えて、 $a', b', c$ を推定することにする

- ★  $T + K$ ：全国的な気温
- ★  $T - K$ ：関東と開催での気温の差

## ★ こうすると、2つの説明変数 $T + K$ と $T - K$ はあまり相関は強くない

- ★ 互いに影響を及ぼし、全体的に推定結果が不安定になることはない
- ★  $T - K$  は全体的に値が小さく、ビールの売上をうまく説明できないであろうため、この係数に対してはうまく行かない
  - ★  $T - K$  は説明変数として不要

# 多重共線性

## ★ 説明変数の無相関化

- ★ 説明変数を無相関にする方法として主成分分析を行う方法がある
- ★ 主成分分析した結果を用いて回帰分析を行うことを主成分回帰と呼ぶ

# 多重共線性と主成分回帰

★  $A = UDV^T$  と特異値分解されるとする

★  $U^T U = I, V^T V = V V^T = I$  で  $D$  は対角行列で対角成分は全て正

★ 最小二乗推定量は  $A^T A \beta = A^T y$  を満たす  $\beta$  のことだから,

$$A^T A \beta = A^T y$$

$$(UDV^T)^T (UDV^T) \beta = (UDV^T)^T y$$

$$VDU^T UDV^T \beta = VDU^T y$$

$$VD^2 V^T \beta = VDU^T y$$

$$\beta = VD^{-1} U^T y$$

★ ここで,  $D^{-1}$  の要素が大きい部分が不安定になる

# 多重共線性と主成分回帰

★  $A = UDV^T$  と特異値分解されるとする

★  $U^T U = I, V^T V = VV^T = I$  で  $D$  は対角行列で対角成分は全て正

★ 説明変数を主成分  $AV = UD$  とする

★ 最小二乗推定量は  $(AV)^T(AV)\beta = (AV)^T y$  を満たす  $\beta$  のことだから、

$$A^T A \beta = A^T y$$

$$(UD)^T(UD)\beta = (UD)^T y$$

$$DU^T UD \beta = DU^T y$$

$$D^2 \beta = DU^T y$$

$$\beta = D^{-1} U^T y$$

★ ここで、 $D^{-1}$  の要素が大きい部分、下位の主成分の係数が不安定になる

## 演習 - ビールの売上の予測

## 考察

- ★ 東京の気温のデータと京都の気温のデータは強い相関がある
  - ★ そのため推定結果が不安定になる（多重共線性）
  - ★ その上，ほぼ同じデータなので，説明能力がほぼ上がらない

---

- ★ ビールの売上が気温だけからあまり説明できない要因を見つけよう

## 可視化してみる

★ 東京の気温とビールの売上のデータを可視化してみよう

★ ここでは散布図を利用する



# 散布図の描き方

★ 散布図を描きたいデータを選択する

★ 挿入 → 散布図 と選択する

The screenshot shows the Microsoft Excel interface with the 'Insert' tab selected. The 'Charts' group in the ribbon is expanded, and the 'Scatter' icon is highlighted with a red circle. A tooltip is displayed over the 'Scatter' icon, providing instructions on how to insert a scatter chart.

散布図 (X, Y) またはバブル チャートの挿入  
この種類のグラフは、値のセットの関係を表示するのに使います。

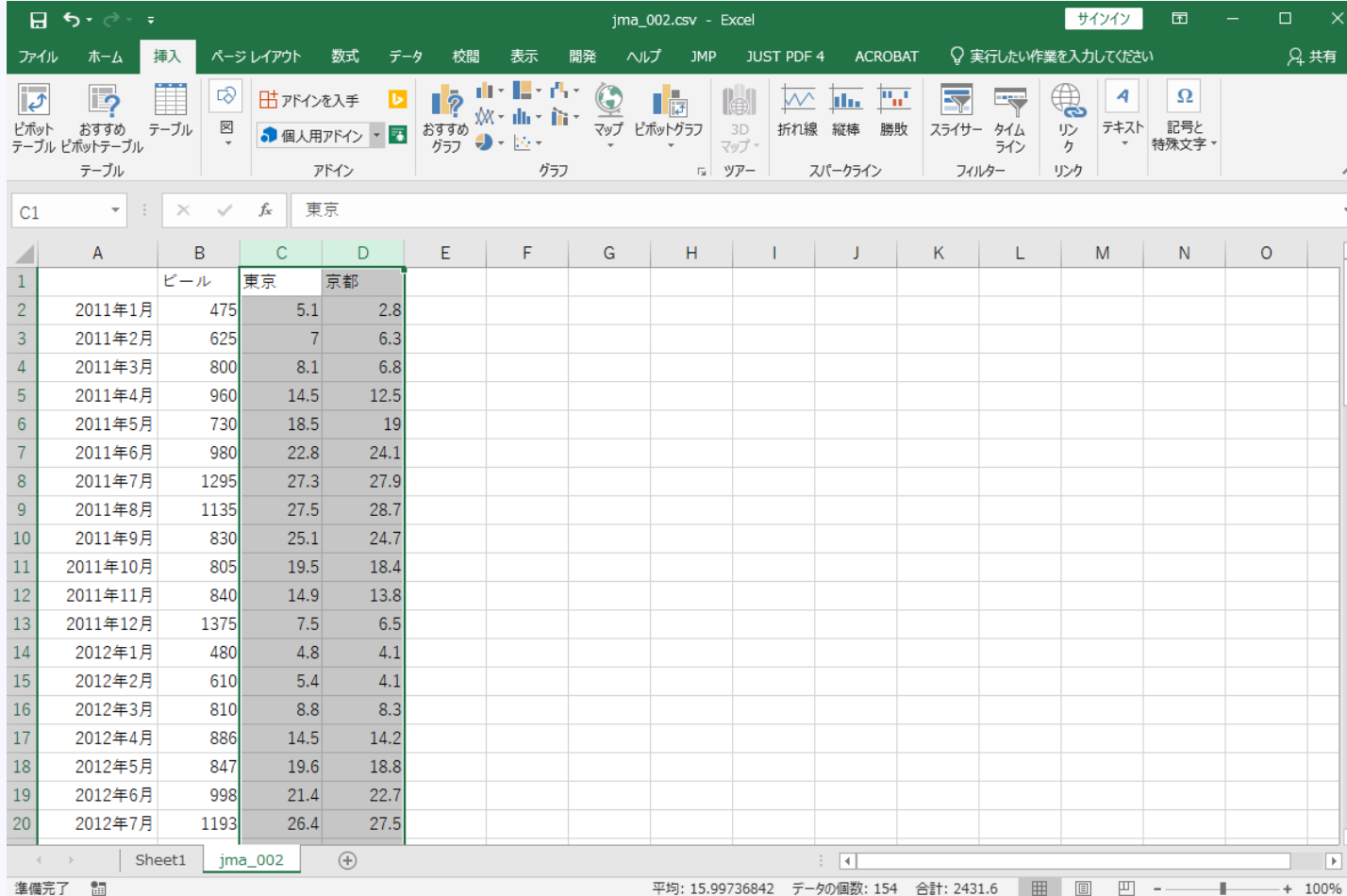
別の種類の散布図グラフとバブル チャートを表示するには、矢印をクリックします。ドキュメント内でプレビューを表示するには、ポインターをアイコンの上に置きます。

	A	B	C	D	E	J	K	L	M	N	O
1		ビール	東京	京都							
2	2011年1月	475	5.1	2.8							
3	2011年2月	625	7	6.3							
4	2011年3月	800	8.1	6.8							
5	2011年4月	960	14.5	12.5							
6	2011年5月	730	18.5	19							
7	2011年6月	980	22.8	24.1							
8	2011年7月	1295	27.3	27.9							
9	2011年8月	1135	27.5	28.7							
10	2011年9月	830	25.1	24.7							
11	2011年10月	805	19.5	18.4							
12	2011年11月	840	14.9	13.8							
13	2011年12月	1375	7.5	6.5							
14	2012年1月	480	4.8	4.1							
15	2012年2月	610	5.4	4.1							
76	2017年3月	796	8.5	8.2							
77	2017年4月	784	14.7	14.8							
78											
79											
80											

# 散布図1

★ 本題の前に、東京の気温と京都の気温の散布図を見てみましょう

★ 強い正の相関があることが目で見てわかります



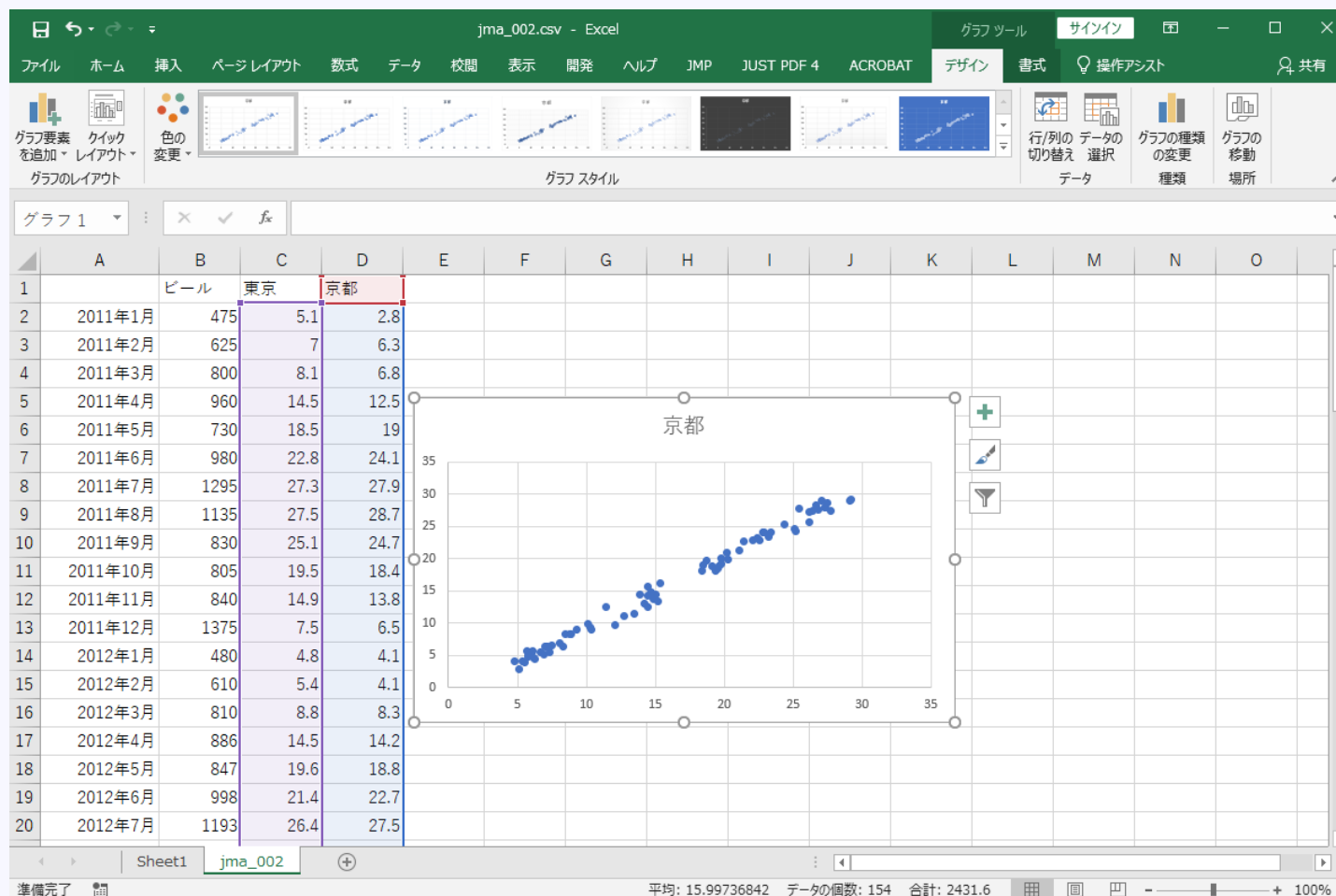
The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		ビール	東京	京都											
2	2011年1月	475	5.1	2.8											
3	2011年2月	625	7	6.3											
4	2011年3月	800	8.1	6.8											
5	2011年4月	960	14.5	12.5											
6	2011年5月	730	18.5	19											
7	2011年6月	980	22.8	24.1											
8	2011年7月	1295	27.3	27.9											
9	2011年8月	1135	27.5	28.7											
10	2011年9月	830	25.1	24.7											
11	2011年10月	805	19.5	18.4											
12	2011年11月	840	14.9	13.8											
13	2011年12月	1375	7.5	6.5											
14	2012年1月	480	4.8	4.1											
15	2012年2月	610	5.4	4.1											
16	2012年3月	810	8.8	8.3											
17	2012年4月	886	14.5	14.2											
18	2012年5月	847	19.6	18.8											
19	2012年6月	998	21.4	22.7											
20	2012年7月	1193	26.4	27.5											

# 散布図1

★ 本題の前に、東京の気温と京都の気温の散布図を見てみましょう

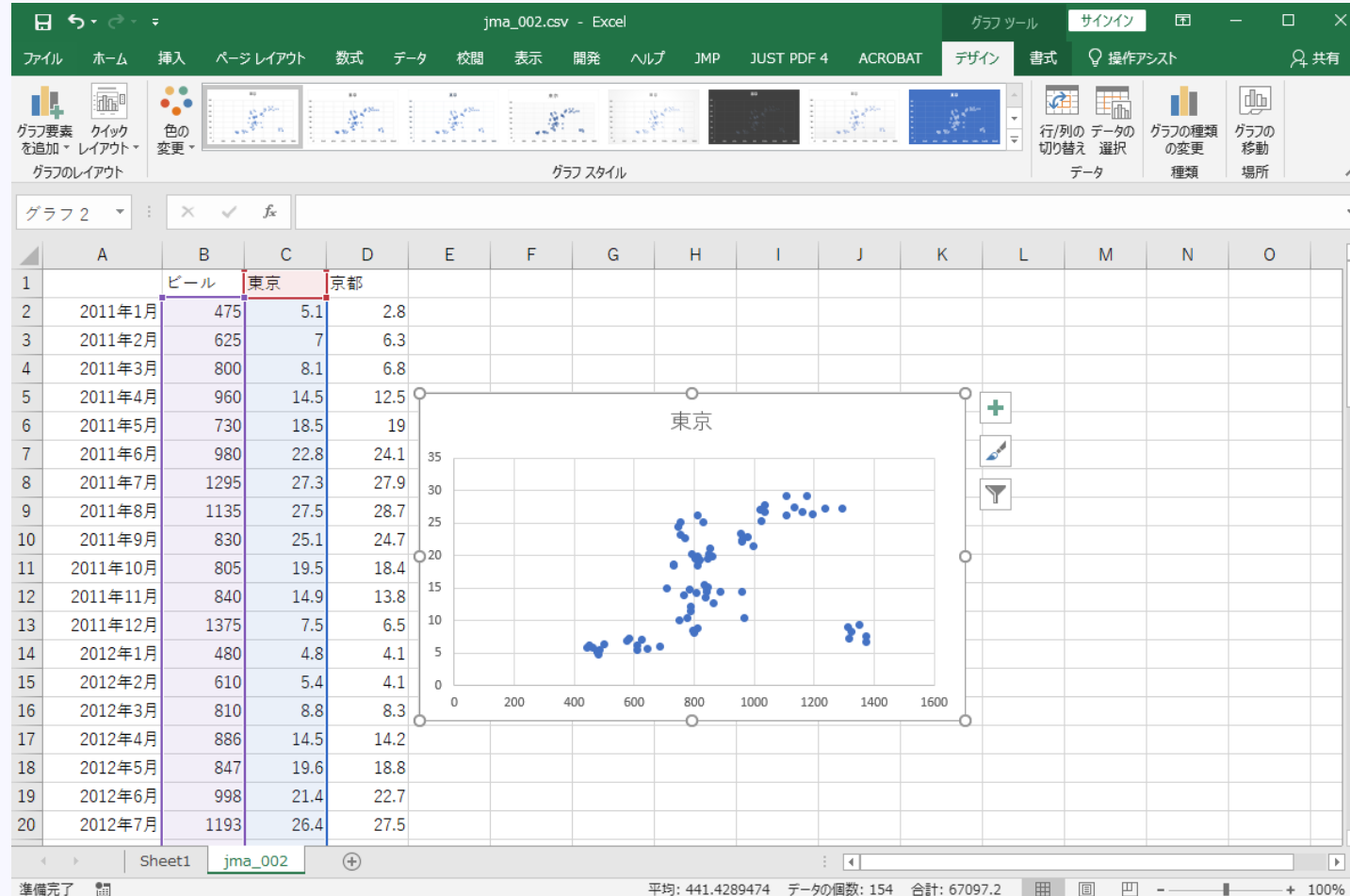
★ 強い正の相関があることが目で見てわかります



# 散布図2

★ 東京の気温とビールの売上の散布図を見てみましょう

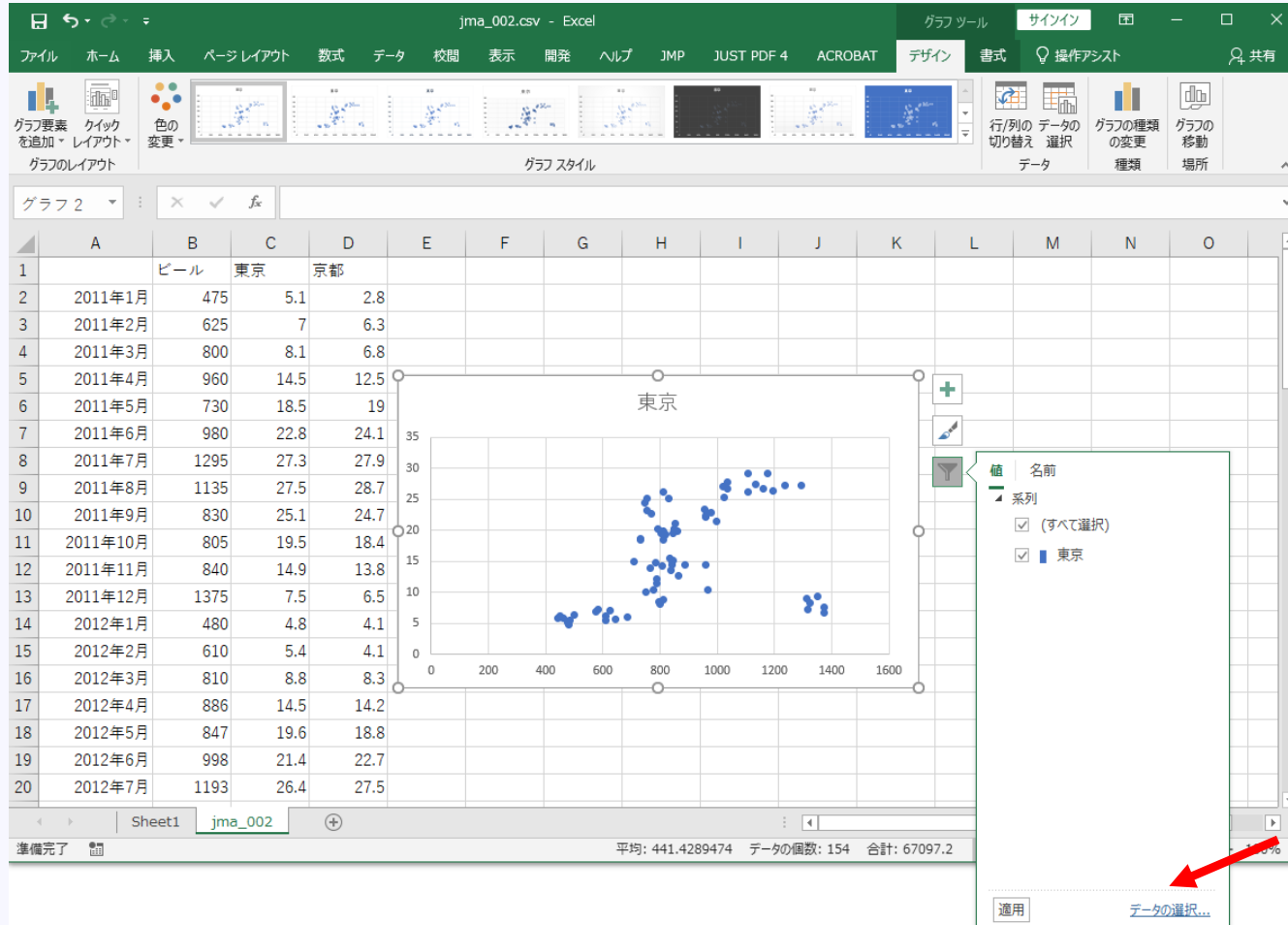
★ 何か感じることはあるでしょうか



# 散布図2

★ 東京の気温とビールの売上の散布図を見てみましょう

★ 何か感じることはあるでしょうか



# 散布図2

★ 東京の気温とビールの売上の散布図を見てみましょう

★ 何か感じることはあるでしょうか

The screenshot shows the Excel interface with a scatter plot of beer sales in Tokyo. The data is as follows:

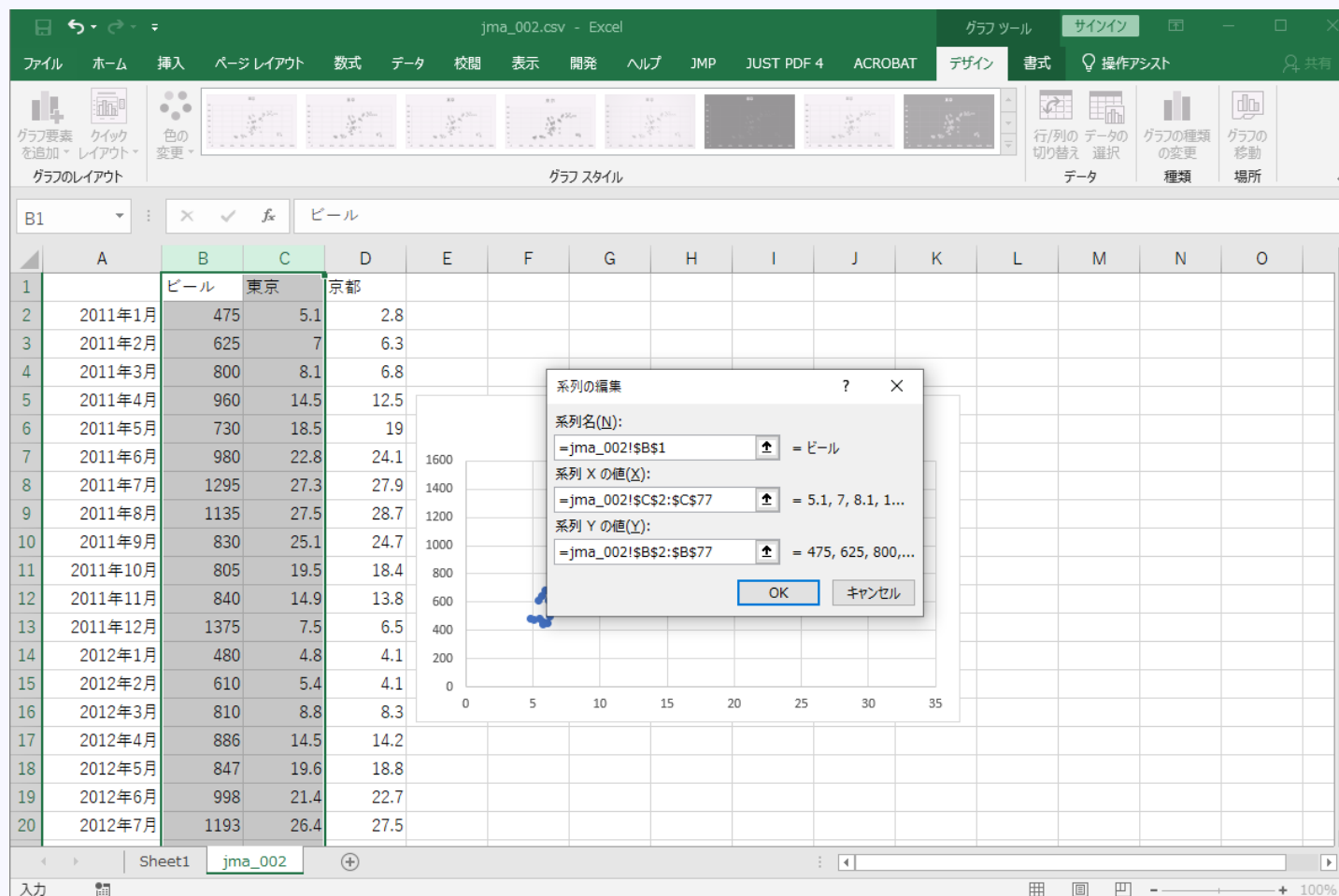
Year	Beer Sales	Temperature
2011年1月	475	5.1
2011年2月	625	7
2011年3月	800	8.1
2011年4月	960	14.5
2011年5月	730	18.5
2011年6月	980	22.8
2011年7月	1295	27.3
2011年8月	1135	27.5
2011年9月	830	25.1
2011年10月	805	19.5
2011年11月	840	14.9
2011年12月	1375	7.5
2012年1月	480	4.8
2012年2月	610	5.4
2012年3月	810	8.8
2012年4月	886	14.5
2012年5月	847	19.6
2012年6月	998	21.4
2012年7月	1193	26.4

The dialog box 'データのソースの選択' (Select Data Source) is open, showing the data range as '=jma\_002!\$B\$1:\$C\$77'. The 'Series in Groups' (凡例項目) list includes '東京' (Tokyo) with a checked box. The 'Horizontal Axis Labels' (横軸ラベル) list includes the values 475, 625, 800, 960, and 730.

# 散布図2

★ 東京の気温とビールの売上の散布図を見てみましょう

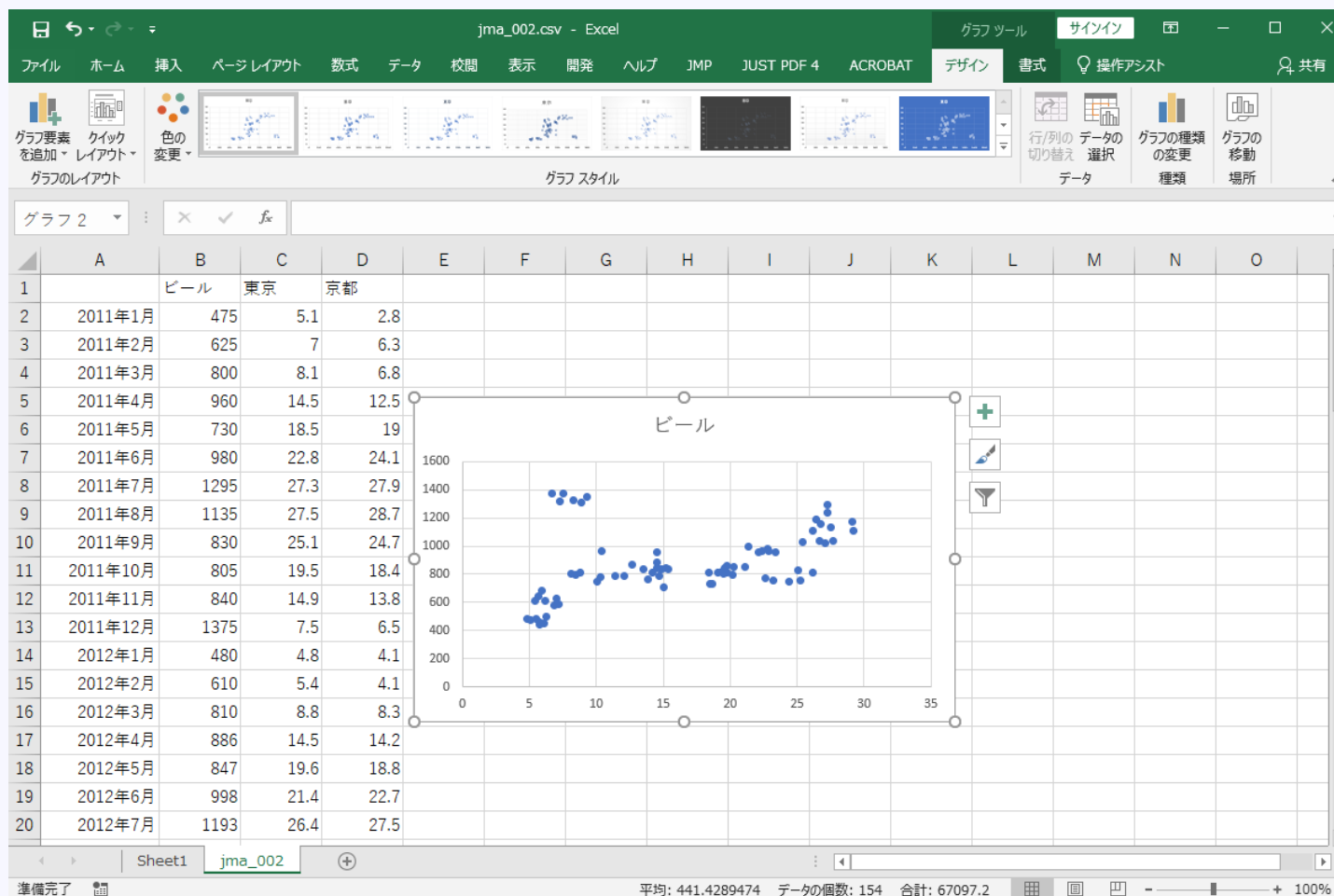
★ 何か感じることはあるでしょうか



# 散布図2

★ 東京の気温とビールの売上の散布図を見てみましょう

★ 何か感じることはあるでしょうか

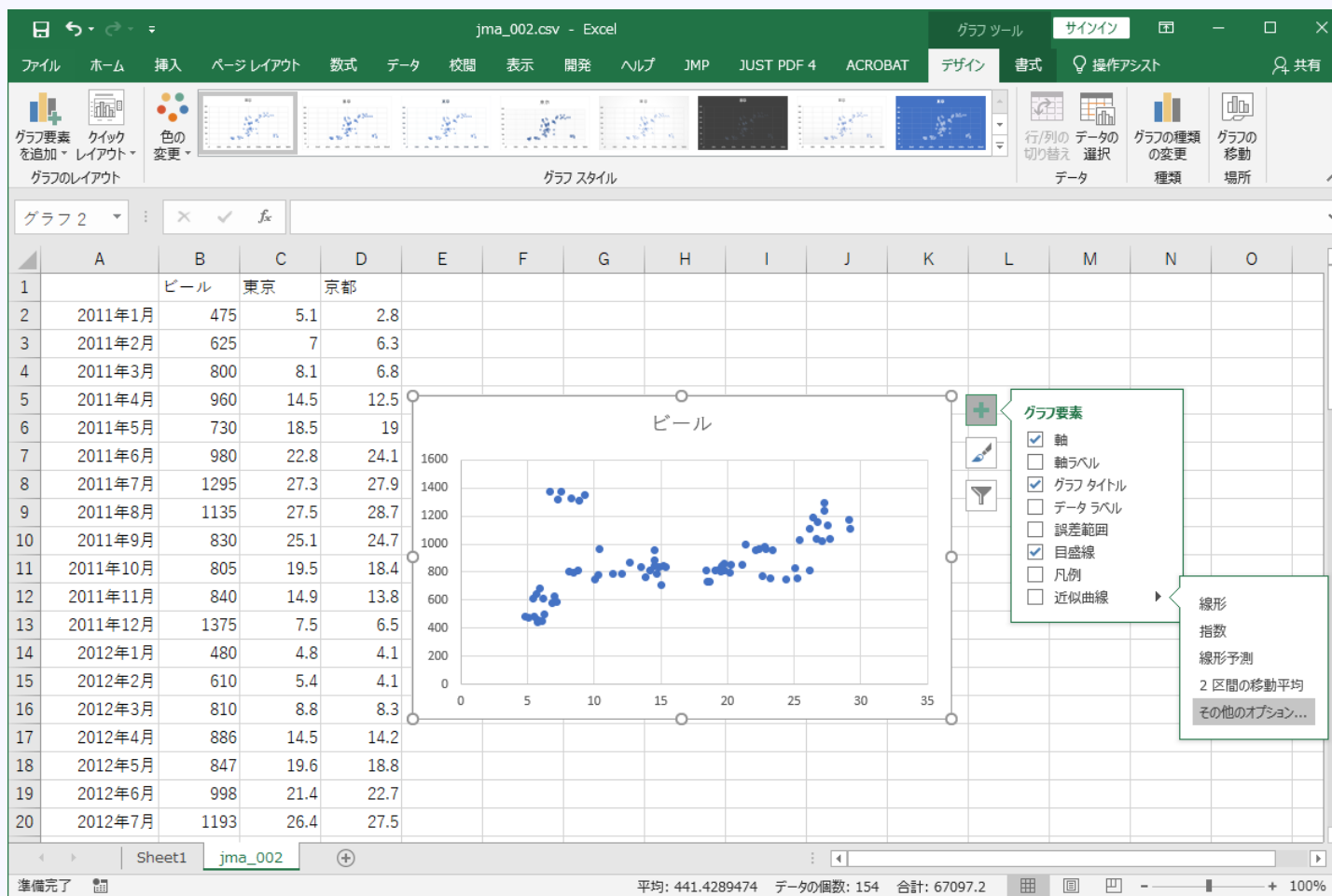




# 散布図2

★ 東京の気温とビールの売上の散布図を見てみましょう

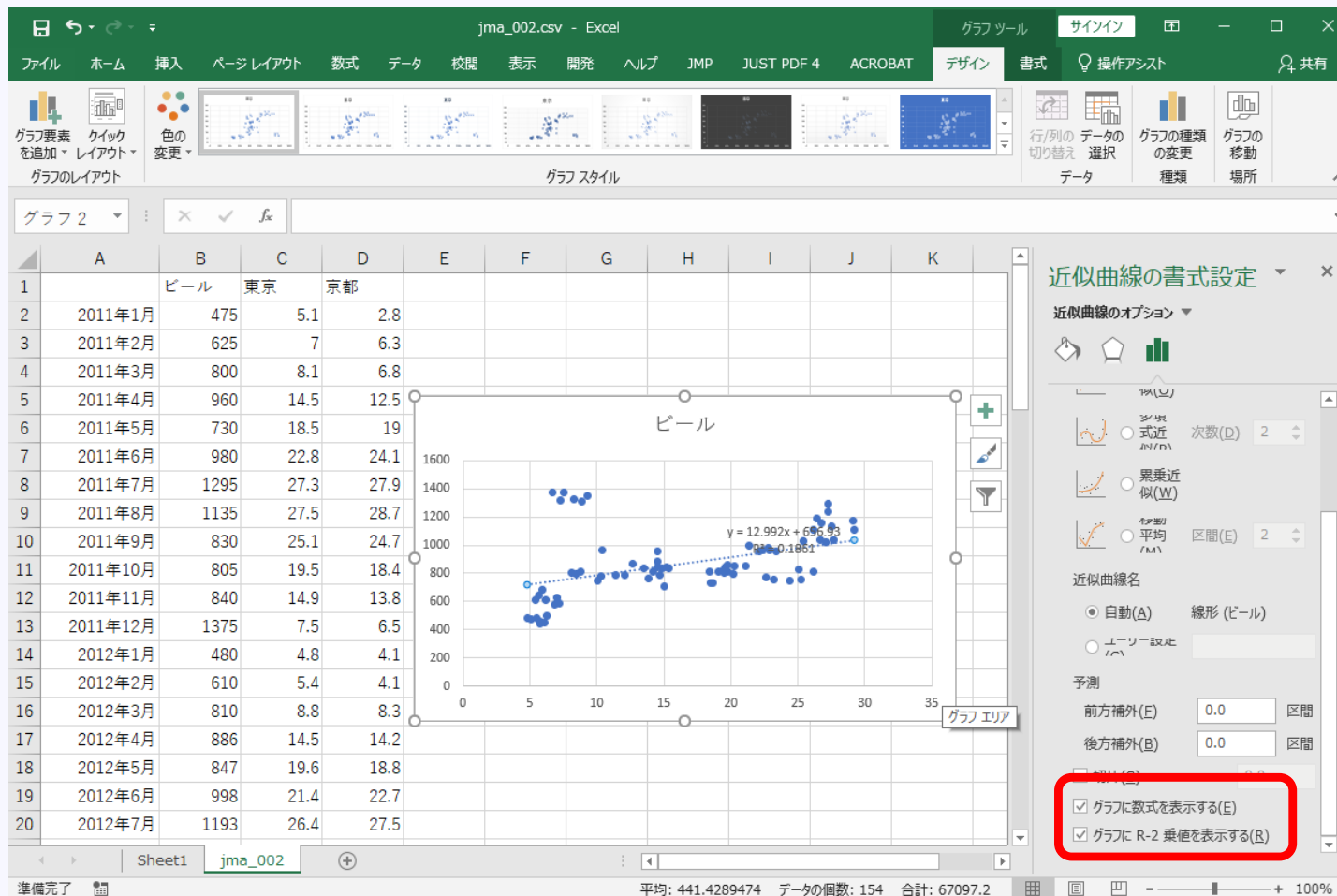
★ 何か感じることはあるでしょうか



# 散布図2

★ 東京の気温とビールの売上の散布図を見てみましょう

★ 何か感じることはあるでしょうか



# 残差をしてみる

★ 残差を何らかの散布図などにプロットしてみるのも有効です

★ 分析ツールで回帰分析を行うときに「残差グラフの作成」にチェックを入れてやってみましょう

回帰分析

入力元  
入力 Y 範囲(Y):   
入力 X 範囲(X):   
 ラベル(L)  定数に 0 を使用(Z)  
 有意水準(Q)  %

出力オプション  
 一覧の出力先(S):   
 新規ワークシート(P):   
 新規ブック(W)

残差  
 残差(R)  残差グラフの作成(D)  
 標準化された残差(I)  観測値グラフの作成(L)

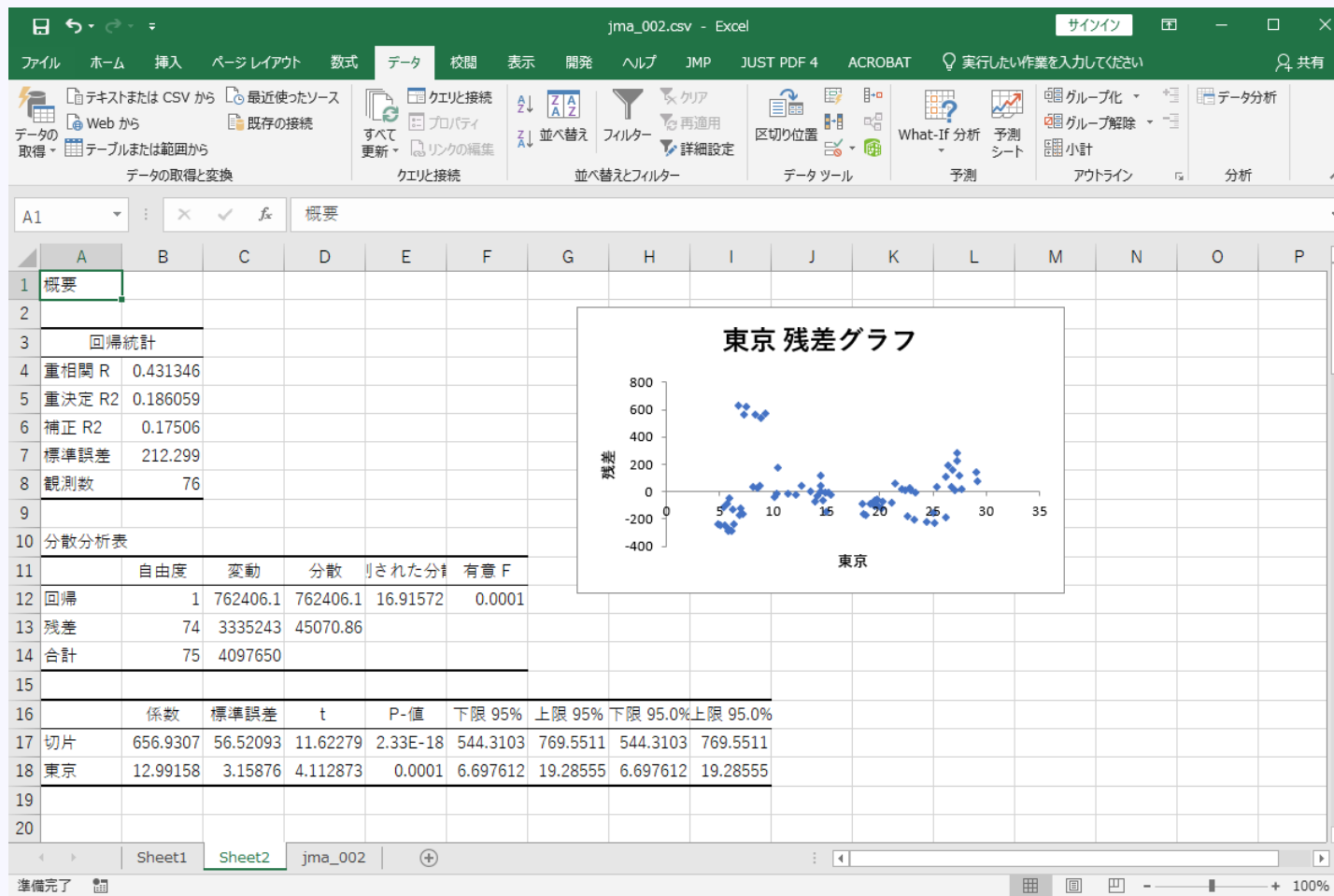
正規確率  
 正規確率グラフの作成(N)

	A	B	C	D
1		ビール	東京	京都
2	2011年1月	475	5.1	2.8
3	2011年2月	625	7	6.3
4	2011年3月	800	8.1	6.8
5	2011年4月	960	14.5	12.5
6	2011年5月	730	18.5	19
7	2011年6月	980	22.8	24.1
8	2011年7月	1295	27.3	27.9
9	2011年8月	1135	27.5	28.7
10	2011年9月	830	25.1	24.7
11	2011年10月	805	19.5	18.4
12	2011年11月	840	14.9	13.8
13	2011年12月	1375	7.5	6.5
14	2012年1月	480	4.8	4.1
15	2012年2月	610	5.4	4.1
16	2012年3月	810	8.8	8.3
17	2012年4月	886	14.5	14.2
18	2012年5月	847	19.6	18.8
19	2012年6月	998	21.4	22.7
20	2012年7月	1193	26.4	27.5

# 残差をしてみる

★ 残差を何らかの散布図などにプロットしてみるのも有効です

★ 分析ツールで回帰分析を行うときに「残差グラフの作成」にチェックを入れてやってみましょう



## 課題2

### ★ 課題2

- ★ ビールの売上を東京の気温のみであまり説明できていなかった
- ★ 散布図などを見てみて、その理由や改善案を考えてみてください

## 考察

- ★ 散布図を見てみると残差の絶対値が大きいグループが見て取れます
  - ★ 左上のグループ
- ★ 対応するデータ番号を見てみると12, 24, 36, …であり, 12月のデータ
  - ★ 改めてデータを確認すると, 12月は気温が低いですが売上が多い
    - ★ 忘年会・お歳暮・新年会 (の準備)

## 課題3

### ★ 課題3

- ★ 気温だけで考えると、12月がイレギュラーな振る舞いをしているので、上手く行っていない（と考えられる）
- ★ 回帰分析の範囲で、その問題に対応するには、どのようにすれば良いだろうか？

# 解決策

★ 色々考えられると思いますが一例

★ 12月だけ変な傾向があるので、12月のデータを削除して回帰分析を行う

★ 月ごとにデータを分解して、12種類のデータにして、それぞれ回帰分析を行う

★ 12月ならば1, 12月でなければ0というダミー変数を導入して回帰分析を行う

---

★ ここでは、12月かどうかを表すダミー変数を導入してみよう



# ファイル

★ ダミー変数のデータも入れた以下のファイルを使用

★ [http://ds.k.kyoto-u.ac.jp/e-learning\\_files/  
data\\_analysis\\_basic/jma\\_003.csv](http://ds.k.kyoto-u.ac.jp/e-learning_files/data_analysis_basic/jma_003.csv)

★ PandAのリソースにも置いてあります

★ 自分でこのデータを作成することも簡単にできます

# ファイルを開いてみましょう

★ 前ページの csv ファイルを Excel で開いて内容を確認してみましょう

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	I
1		ビール	東京	12月												
2	2011年1月	475	5.1	0												
3	2011年2月	625	7	0												
4	2011年3月	800	8.1	0												
5	2011年4月	960	14.5	0												
6	2011年5月	730	18.5	0												
7	2011年6月	980	22.8	0												
8	2011年7月	1295	27.3	0												
9	2011年8月	1135	27.5	0												
10	2011年9月	830	25.1	0												
11	2011年10月	805	19.5	0												
12	2011年11月	840	14.9	0												
13	2011年12月	1375	7.5	1												
14	2012年1月	480	4.8	0												
15	2012年2月	610	5.4	0												
16	2012年3月	810	8.8	0												
17	2012年4月	886	14.5	0												
18	2012年5月	847	19.6	0												
19	2012年6月	998	21.4	0												
20	2012年7月	1193	26.4	0												

# 重回帰分析を行う

★ 今までと同様の手順で重回帰分析を行う

★ 1. データ分析をクリックし，回帰分析を選び，OK を押す

★ 2. 入力 Y 範囲，入力 X 範囲などを適切に記入し，OK を押すことで，回帰分析を行う

★ 2-1. 入力 Y 範囲には「 $\$B\$1:\$B\$77$ 」と入力（\$はなくても構いません）

★ 2-2. 入力 X 範囲には「 $\$C\$1:\$D\$77$ 」と入力（\$はなくても構いません）

★ 2-3. 「ラベル」にチェックを入れる

★ 2-4. OK を押す

# 重回帰分析の実行結果

★ 正しく実行できたら以下のような結果が別のシートに作られ表示されます

The screenshot shows the Excel interface with the 'データ' (Data) ribbon selected. The 'データ分析' (Data Analysis) task pane is open, and the '重回帰分析' (Multiple Regression Analysis) tool has been used. The results are displayed on the 'Sheet1' worksheet, starting from cell A1. The results are organized into two main sections: a summary table and a dispersion analysis table.

概要					
1	概要				
2					
3	重回帰統計				
4	重相関 R	0.883829			
5	重決定 R2	0.781153			
6	補正 R2	0.775157			
7	標準誤差	110.8348			
8	観測数	76			
9					
10	分散分析表				
11		自由度	変動	分散	判された分散
12	回帰	2	3200891	1600446	130.2832
13	残差	73	896758.4	12284.36	8.22E-25
14	合計	75	4097650		
15					
16		係数	標準誤差	t	P-値
17	切片	479.715	32.07686	14.95518	6.34E-24
18	東京	20.55118	1.734185	11.85062	1.12E-18
19	12月	698.5422	49.58027	14.08912	1.63E-22
20					

	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	415.7859	543.6441	415.7859	543.6441
東京	17.09495	24.0074	17.09495	24.0074
12月	599.7289	797.3556	599.7289	797.3556

# 実行結果を読み取る

★  $B = aT + bM_{12} + c + \varepsilon$  の係数  $a, b, c$  が以下のように推定されたことがわかる

★  $a = 20.55118, b = 698.5422, c = 479.715$

★  $B$  はビールの売上,  $T$  は東京の気温,  $M_{12}$  は12月かどうか

	自由度	変動	分散	F	有意 F
回帰	2	3200891	1600446	130.2832	8.22E-25
残差	73	896758.4	12284.36		
合計	75	4097650			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	479.715	32.07686	14.95518	6.34E-24	415.7859	543.6441	415.7859	543.6441
東京	20.55118	1.734185	11.85062	1.12E-18	17.09495	24.0074	17.09495	24.0074
12月	698.5422	40.58027	14.08912	1.63E-22	599.7289	797.3556	599.7289	797.3556

# 実行結果を読み取る

★ 重決定 R2 の値が 0.781153 と表示

★ 説明能力が飛躍的に向上

The screenshot shows an Excel spreadsheet with the following data tables:

回帰統計					
重決定 R2	0.781153				
補正 R2	0.775157				
標準誤差	110.8348				
観測数	76				

分散分析表					
	自由度	変動	分散	F 値	有意 F
回帰	2	3200891	1600446	130.2832	8.22E-25
残差	73	896758.4	12284.36		
合計	75	4097650			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	479.715	32.07686	14.95518	6.34E-24	415.7859	543.6441	415.7859	543.6441
東京	20.55118	1.734185	11.85062	1.12E-18	17.09495	24.0074	17.09495	24.0074
12月	698.5422	49.58027	14.08912	1.63E-22	599.7289	797.3556	599.7289	797.3556

## 課題4：更なる説明変数の検討

### ★ 課題4

★ その他に説明変数を導入することで説明能力を向上することができないだろうか？

## 課題4：更なる説明変数の検討

### ★ 課題4

★ その他に説明変数を導入することで説明能力を向上することができないだろうか？

### ★ 解答例

- ★ 12月だけ特別視するのは違和感がある。他の月にも色々イベント等があるし、暑いからビールの売上が上がるのではなく、夏だからなんとなくビールの売上が上がるなど季節的な効果があるかもしれない。12月以外の他の月に対してもダミー変数を導入してみたらどうか。
- ★ 横軸を気温にして残差をプロットしてみると傾向が見える。または、気温とビールの売上の散布図を見ると12月のデータを除いても直線的な関係より3次関数的な関係があるように見える。気温の3乗の値を説明変数として加えるのはどうか。
- ★ その他、時代を表す指標、景気を表す指標、ビールを飲むことができる20歳以上の人口、対応する月の平日・祝日の日数、などの導入。



# 方針

- ★ ここでは、以下の説明変数を採用して回帰分析を行ってみよう
  - ★ 東京の気温，京都の気温
  - ★ 次の月の東京の気温
  - ★ とある月からの経過した月の数
  - ★ 1月～12月までの，それぞれの月を表すダミー変数

---

- ★ Excelの分析ツールの回帰分析では説明変数の数は16まで

# ファイル

- ★ 前スライドの説明変数を整形した以下のファイルを使用
  - ★ [http://ds.k.kyoto-u.ac.jp/e-learning\\_files/data\\_analysis\\_basic/jma\\_004.csv](http://ds.k.kyoto-u.ac.jp/e-learning_files/data_analysis_basic/jma_004.csv)
  - ★ PandAのリソースにも置いてあります

# ファイルを開いてみましょう

★ 前ページの csv ファイルを Excel で開いて内容を確認してみましょう

The screenshot shows the Excel interface with the following data in the spreadsheet:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		ビール	東京	京都	東京1月後	時間	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月		
2	2011年1月	475	5.1	2.8	7	1	1	0	0	0	0	0	0	0	0	0	0	0		
3	2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0		
4	2011年3月	800	8.1	6.8	14.5	3	0	0	1	0	0	0	0	0	0	0	0	0		
5	2011年4月	960	14.5	12.5	18.5	4	0	0	0	1	0	0	0	0	0	0	0	0		
6	2011年5月	730	18.5	19	22.8	5	0	0	0	0	1	0	0	0	0	0	0	0		
7	2011年6月	980	22.8	24.1	27.3	6	0	0	0	0	0	1	0	0	0	0	0	0		
8	2011年7月	1295	27.3	27.9	27.5	7	0	0	0	0	0	0	1	0	0	0	0	0		
9	2011年8月	1135	27.5	28.7	25.1	8	0	0	0	0	0	0	0	1	0	0	0	0		
10	2011年9月	830	25.1	24.7	19.5	9	0	0	0	0	0	0	0	0	1	0	0	0		
11	2011年10月	805	19.5	18.4	14.9	10	0	0	0	0	0	0	0	0	0	1	0	0		
12	2011年11月	840	14.9	13.8	7.5	11	0	0	0	0	0	0	0	0	0	0	1	0		
13	2011年12月	1375	7.5	6.5	4.8	12	0	0	0	0	0	0	0	0	0	0	0	1		
14	2012年1月	480	4.8	4.1	5.4	13	1	0	0	0	0	0	0	0	0	0	0	0		
15	2012年2月	610	5.4	4.1	8.8	14	0	1	0	0	0	0	0	0	0	0	0	0		
16	2012年3月	810	8.8	8.3	14.5	15	0	0	1	0	0	0	0	0	0	0	0	0		
17	2012年4月	886	14.5	14.2	19.6	16	0	0	0	1	0	0	0	0	0	0	0	0		
18	2012年5月	847	19.6	18.8	21.4	17	0	0	0	0	1	0	0	0	0	0	0	0		
19	2012年6月	998	21.4	22.7	26.4	18	0	0	0	0	0	1	0	0	0	0	0	0		
20	2012年7月	1193	26.4	27.5	29.1	19	0	0	0	0	0	0	1	0	0	0	0	0		

# データの説明

★ ビール・東京・京都・12月は今まで利用したのと同じデータです

The screenshot shows an Excel spreadsheet with the following data structure:

	ビール	東京	京都	東京1月後	時間	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
2011年1月	475	5.1	2.8	7	1	1	0	0	0	0	0	0	0	0	0	0	0
2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0
2011年3月	800	8.1	6.8	14.5	3	0	0	1	0	0	0	0	0	0	0	0	0
2011年4月	960	14.5	12.5	18.5	4	0	0	0	1	0	0	0	0	0	0	0	0
2011年5月	730	18.5	19	22.8	5	0	0	0	0	1	0	0	0	0	0	0	0
2011年6月	980	22.8	24.1	27.3	6	0	0	0	0	0	1	0	0	0	0	0	0
2011年7月	1295	27.3	27.9	27.5	7	0	0	0	0	0	0	1	0	0	0	0	0
2011年8月	1135	27.5	28.7	25.1	8	0	0	0	0	0	0	0	1	0	0	0	0
2011年9月	830	25.1	24.7	19.5	9	0	0	0	0	0	0	0	0	1	0	0	0
2011年10月	805	19.5	18.4	14.9	10	0	0	0	0	0	0	0	0	0	1	0	0
2011年11月	840	14.9	13.8	7.5	11	0	0	0	0	0	0	0	0	0	0	1	0
2011年12月	1375	7.5	6.5	4.8	12	0	0	0	0	0	0	0	0	0	0	0	1
2012年1月	480	4.8	4.1	5.4	13	1	0	0	0	0	0	0	0	0	0	0	0
2012年2月	610	5.4	4.1	8.8	14	0	1	0	0	0	0	0	0	0	0	0	0
2012年3月	810	8.8	8.3	14.5	15	0	0	1	0	0	0	0	0	0	0	0	0
2012年4月	886	14.5	14.2	19.6	16	0	0	0	1	0	0	0	0	0	0	0	0
2012年5月	847	19.6	18.8	21.4	17	0	0	0	0	1	0	0	0	0	0	0	0
2012年6月	998	21.4	22.7	26.4	18	0	0	0	0	0	1	0	0	0	0	0	0
2012年7月	1193	26.4	27.5	29.1	19	0	0	0	0	0	0	1	0	0	0	0	0

# データの説明

- ★ 1月から11月までのデータは、その月であれば1, その月でなければ0を表すダミー変数
  - ★ 4月は歓迎会のシーズンなどのイベントがあるので要因を説明してくれるかも
  - ★ 夏は季節的にビールを飲むものだ, という季節的な要因も説明してくれるかも

The screenshot shows an Excel spreadsheet titled 'jma\_004.csv - Excel'. The data is organized as follows:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		ビール	東京	京都	東京1月後 時間	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月			
2	2011年1月	475	5.1	2.8	7	1	0	0	0	0	0	0	0	0	0	0	0	0		
3	2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0		
4	2011年3月	800	8.1	6.8	14.5	3	0	0	1	0	0	0	0	0	0	0	0	0		
5	2011年4月	960	14.5	12.5	18.5	4	0	0	0	1	0	0	0	0	0	0	0	0		
6	2011年5月	730	18.5	19	22.8	5	0	0	0	0	1	0	0	0	0	0	0	0		
7	2011年6月	980	22.8	24.1	27.3	6	0	0	0	0	0	1	0	0	0	0	0	0		
8	2011年7月	1295	27.3	27.9	27.5	7	0	0	0	0	0	0	1	0	0	0	0	0		
9	2011年8月	1135	27.5	28.7	25.1	8	0	0	0	0	0	0	0	1	0	0	0	0		
10	2011年9月	830	25.1	24.7	19.5	9	0	0	0	0	0	0	0	0	1	0	0	0		
11	2011年10月	805	19.5	18.4	14.9	10	0	0	0	0	0	0	0	0	0	1	0	0		
12	2011年11月	840	14.9	13.8	7.5	11	0	0	0	0	0	0	0	0	0	0	1	0		
13	2011年12月	1375	7.5	6.5	4.8	12	0	0	0	0	0	0	0	0	0	0	0	1		
14	2012年1月	480	4.8	4.1	5.4	13	1	0	0	0	0	0	0	0	0	0	0	0		
15	2012年2月	610	5.4	4.1	8.8	14	0	1	0	0	0	0	0	0	0	0	0	0		
16	2012年3月	810	8.8	8.3	14.5	15	0	0	1	0	0	0	0	0	0	0	0	0		
17	2012年4月	886	14.5	14.2	19.6	16	0	0	0	1	0	0	0	0	0	0	0	0		
18	2012年5月	847	19.6	18.8	21.4	17	0	0	0	0	1	0	0	0	0	0	0	0		
19	2012年6月	998	21.4	22.7	26.4	18	0	0	0	0	0	1	0	0	0	0	0	0		
20	2012年7月	1193	26.4	27.5	29.1	19	0	0	0	0	0	0	1	0	0	0	0	0		

# データの説明

- ★ 時間は、データの最初の月を1として、次の月は2、その次は3と番号を振ったもの
- ★ 最近では若者のビール離れが進んでいる、など時代的な要因を説明してくれるかも

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		ビール	東京	京都	東京1月	時間	月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月		
2	2011年1月	475	5.1	2.8	7	1	0	0	0	0	0	0	0	0	0	0	0	0		
3	2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0		
4	2011年3月	800	8.1	6.8	14.5	3	0	0	1	0	0	0	0	0	0	0	0	0		
5	2011年4月	960	14.5	12.5	18.5	4	0	0	0	1	0	0	0	0	0	0	0	0		
6	2011年5月	730	18.5	19	22.8	5	0	0	0	0	1	0	0	0	0	0	0	0		
7	2011年6月	980	22.8	24.1	27.3	6	0	0	0	0	0	1	0	0	0	0	0	0		
8	2011年7月	1295	27.3	27.9	27.5	7	0	0	0	0	0	0	1	0	0	0	0	0		
9	2011年8月	1135	27.5	28.7	25.1	8	0	0	0	0	0	0	0	1	0	0	0	0		
10	2011年9月	830	25.1	24.7	19.5	9	0	0	0	0	0	0	0	0	1	0	0	0		
11	2011年10月	805	19.5	18.4	14.9	10	0	0	0	0	0	0	0	0	0	1	0	0		
12	2011年11月	840	14.9	13.8	7.5	11	0	0	0	0	0	0	0	0	0	0	1	0		
13	2011年12月	1375	7.5	6.5	4.8	12	0	0	0	0	0	0	0	0	0	0	0	1		
14	2012年1月	480	4.8	4.1	5.4	13	1	0	0	0	0	0	0	0	0	0	0	0		
15	2012年2月	610	5.4	4.1	8.8	14	0	1	0	0	0	0	0	0	0	0	0	0		
16	2012年3月	810	8.8	8.3	14.5	15	0	0	1	0	0	0	0	0	0	0	0	0		
17	2012年4月	886	14.5	14.2	19.6	16	0	0	0	1	0	0	0	0	0	0	0	0		
18	2012年5月	847	19.6	18.8	21.4	17	0	0	0	0	1	0	0	0	0	0	0	0		
19	2012年6月	998	21.4	22.7	26.4	18	0	0	0	0	0	1	0	0	0	0	0	0		
20	2012年7月	1193	26.4	27.5	29.1	19	0	0	0	0	0	0	1	0	0	0	0	0		

# データの説明

★ 東京1ヶ月後は東京の次の月の気温のデータ

★ 12月にビールの売上が上がる要因として、新年会の準備が考えられるのでした

★ このように次の月の情報で説明できることもあるかも

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		ビール	東京	京都	東京1ヶ月後	時間	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月		
2	2011年1月	475	5.1	2.8	7	1	1	0	0	0	0	0	0	0	0	0	0	0		
3	2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0		
4	2011年3月	800	8.1	6.8	14.5	3	0	0	1	0	0	0	0	0	0	0	0	0		
5	2011年4月	960	14.5	12.5	18.5	4	0	0	0	1	0	0	0	0	0	0	0	0		
6	2011年5月	730	18.5	19	22.8	5	0	0	0	0	1	0	0	0	0	0	0	0		
7	2011年6月	980	22.8	24.1	27.3	6	0	0	0	0	0	1	0	0	0	0	0	0		
8	2011年7月	1295	27.3	27.9	27.5	7	0	0	0	0	0	0	1	0	0	0	0	0		
9	2011年8月	1135	27.5	28.7	25.1	8	0	0	0	0	0	0	0	1	0	0	0	0		
10	2011年9月	830	25.1	24.7	19.5	9	0	0	0	0	0	0	0	0	1	0	0	0		
11	2011年10月	805	19.5	18.4	14.9	10	0	0	0	0	0	0	0	0	0	1	0	0		
12	2011年11月	840	14.9	13.8	7.5	11	0	0	0	0	0	0	0	0	0	0	1	0		
13	2011年12月	1375	7.5	6.5	4.8	12	0	0	0	0	0	0	0	0	0	0	0	1		
14	2012年1月	480	4.8	4.1	5.4	13	1	0	0	0	0	0	0	0	0	0	0	0		
15	2012年2月	610	5.4	4.1	8.8	14	0	1	0	0	0	0	0	0	0	0	0	0		
16	2012年3月	810	8.8	8.3	14.5	15	0	0	1	0	0	0	0	0	0	0	0	0		
17	2012年4月	886	14.5	14.2	19.6	16	0	0	0	1	0	0	0	0	0	0	0	0		
18	2012年5月	847	19.6	18.8	21.4	17	0	0	0	0	1	0	0	0	0	0	0	0		
19	2012年6月	998	21.4	22.7	26.4	18	0	0	0	0	0	1	0	0	0	0	0	0		
20	2012年7月	1193	26.4	27.5	29.1	19	0	0	0	0	0	0	1	0	0	0	0	0		

# データの説明

★ 東京1ヶ月後は東京の次の月の気温のデータ

★ このデータは東京の気温データを使って、1行ずらして生成できます

★ 最後の月のデータはありませんので、合計のデータの個数が1つ減らしています

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		ビール	東京	京都	東京1ヶ月後	時間	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月		
2	2011年1月	475	5.1	2.8	7	1	1	0	0	0	0	0	0	0	0	0	0	0		
3	2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0		
4	2011年3月	800	8.1	6.8	14.5	3	0	0	1	0	0	0	0	0	0	0	0	0		
5	2011年4月	960	14.5	12.5	18.5	4	0	0	0	1	0	0	0	0	0	0	0	0		
6	2011年5月	730	18.5	19	22.8	5	0	0	0	0	1	0	0	0	0	0	0	0		
7	2011年6月	980	22.8	24.1	27.3	6	0	0	0	0	0	1	0	0	0	0	0	0		
8	2011年7月	1295	27.3	27.9	27.5	7	0	0	0	0	0	0	1	0	0	0	0	0		
9	2011年8月	1135	27.5	28.7	25.1	8	0	0	0	0	0	0	0	1	0	0	0	0		
10	2011年9月	830	25.1	24.7	19.5	9	0	0	0	0	0	0	0	0	1	0	0	0		
11	2011年10月	805	19.5	18.4	14.9	10	0	0	0	0	0	0	0	0	0	1	0	0		
12	2011年11月	840	14.9	13.8	7.5	11	0	0	0	0	0	0	0	0	0	0	1	0		
13	2011年12月	1375	7.5	6.5	4.8	12	0	0	0	0	0	0	0	0	0	0	0	1		
14	2012年1月	480	4.8	4.1	5.4	13	1	0	0	0	0	0	0	0	0	0	0	0		
15	2012年2月	610	5.4	4.1	8.8	14	0	1	0	0	0	0	0	0	0	0	0	0		
75	2017年2月	575	6.9	5.1	8.5	74	0	1	0	0	0	0	0	0	0	0	0	0		
76	2017年3月	796	8.5	8.2	14.7	75	0	0	1	0	0	0	0	0	0	0	0	0		
77																				
78																				
79																				



# 重回帰分析を行う

★ 今までと同様の手順で重回帰分析を行う

★ 1. データ分析をクリックし，回帰分析を選び，OK を押す

★ 2. 入力 Y 範囲，入力 X 範囲などを適切に記入し，OK を押すことで，回帰分析を行う

★ 2-1. 入力 Y 範囲には「**\$B\$1:\$B\$76**」と入力（\$はなくても構いません）

★ 2-2. 入力 X 範囲には「**\$C\$1:\$R\$76**」と入力（\$はなくても構いません）

★ 2-3. 「ラベル」にチェックを入れる

★ 2-4. OK を押す

# 重回帰分析の実行結果1

★ 正しく実行できたら以下のような結果が別のシートに作られ表示されます

The screenshot shows an Excel spreadsheet with the following data:

分散分析表									
	自由度	変動	分散	割された分散	有意 F				
11									
12	回帰	16	3964305	247769	123.3494	1.12E-38			
13	残差	59	126412.7	2142.588					
14	合計	75	4090717						
15									
係数									
	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%	
16	切片	684.3485	105.1918	6.505719	1.85E-08	473.8601	894.8368	473.8601	894.8368
17	東京	17.56614	9.091125	1.93223	0.058137	-0.62515	35.75744	-0.62515	35.75744
18	京都	-9.73623	9.133442	-1.066	0.290768	-28.0122	8.539748	-28.0122	8.539748
19	東京1月後	5.127993	6.523588	0.786069	0.434973	-7.92568	18.18166	-7.92568	18.18166

# 重回帰分析の実行結果2

★ 正しく実行できたら以下のような結果が別のシートに作られ表示されます

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	684.3485	105.1918	6.505719	1.85E-08	473.8601	894.8368	473.8601	894.8368
東京	17.56614	9.091125	1.93223	0.058137	-0.62515	35.75744	-0.62515	35.75744
京都	-9.73623	9.133442	-1.066	0.290768	-28.0122	8.539748	-28.0122	8.539748
東京1月後	5.127993	6.523588	0.786069	0.434973	-7.92568	18.18166	-7.92568	18.18166
時間	-0.87056	0.296181	-2.93927	0.00469	-1.46321	-0.2779	-1.46321	-0.2779
1月	-270.33	61.10781	-4.42383	4.24E-05	-392.607	-148.054	-392.607	-148.054
2月	-141.868	43.39417	-3.2693	0.001802	-228.7	-55.037	-228.7	-55.037
3月	0	0	65535	#NUM!	0	0	0	0
4月	-31.3622	50.58889	-0.61994	#NUM!	-132.59	69.86595	-132.59	69.86595
5月	-103.683	82.74835	-1.25299	0.215153	-269.262	61.89637	-269.262	61.89637
6月	13.65431	110.5861	0.123472	0.902153	-207.628	234.9365	-207.628	234.9365
7月	177.2383	135.3375	1.309603	0.195407	-93.5713	448.048	-93.5713	448.048
8月	97.66896	134.6737	0.725226	0.47118	-171.813	367.1505	-171.813	367.1505
9月	-163.846	105.6793	-1.55041	0.126391	-375.31	47.61743	-375.31	47.61743
10月	-72.7804	75.65567	-0.962	0.339978	-224.167	78.60622	-224.167	78.60622
11月	15.48542	63.00444	0.245783	0.806703	-110.586	141.557	-110.586	141.557
12月	591.406	62.96016	9.393337	2.54E-13	465.423	717.389	465.423	717.389

# 考察

- ★ 東京の気温、京都の気温の係数は多重共線性の影響で単体だと信用できない
  - ★ 足すと正ではあるので、やっぱり季節的な問題ではなく、同じ月でも気温が高いほうがビールの売上が上がる傾向があるだろう
- ★ 重決定  $R^2$  の値が 0.969098 とかなり 1 に近づいた。これだけ説明変数があればビールの売上の大部分が説明できている
  - ★ ただし、これはあくまで今あるデータ（標本）に対して説明できているだけで、母集団について説明できているかどうかは別物
  - ★ 母集団について上手く説明できているかどうかを調べるには別の指標を用いると良い
- ★ 3月・4月を表すダミー変数の結果あたりが変なことになっている

## ダミー変数に対する考察

★ 3月・4月のダミー変数が上手く推定できていない問題

★  $k$ 月を表すダミー変数  $M_k$  とすると、全てのデータに対して以下が成立.

$$M_1 + M_2 + \cdots + M_{12} = 1$$

★  $M_k$  に対する係数  $c_k$  を全て1増やして、切片を1減らせば、全てのデータに対して残差が一致

★ 最小二乗推定量が一意に定まらない！

## ダミー変数に対する考察

- ★ 解決するためには、ダミー変数を1つ削除すれば良い
- ★ 1月に対応するダミー変数  $M_1$  を削除した場合、係数  $c_k$  が意味するのは、1月に比べて  $k$  月のビールの売上はどれぐらい多いか？
  - ★ 基準を決めて、そこからの差異を表すことになる
- ★ 基準は1月が良いのか？
  - ★ 例えば、基準を12月にすると、全ての月はかなり影響が大きい
  - ★ 例えば、基準を1月にすると、12月以外の月はあまり影響が大きい
    - ★ 各変数の重要性が変わってくる
    - ★ 「良い」モデルを考えるときには影響する可能性

# 良いモデルとは何か？

★ 一般的に、母集団（標本ではない）をうまく説明できるのが良いモデル

★ 個別の目的がある場合はその限りではない

★ モデルの自由度・複雑度が低いにもかかわらず、標本（データ）をうまく説明できているモデルが良い

★ 自由度が高くと、無理やりデータの近くを通ろうとし、結果不自然なモデルへ

★ 自由度・複雑度を抑えるために

★ パタメータに制限やペナルティを与え、あまり自由な値を取れなくする

★ Ridge 回帰・Lasso 回帰

★ 考えられる説明変数のうち、本当に必要と思われるものの荷を採用（説明変数を減らす）

★ モデル選択の一部手法

# 正則化付き最小二乗法



# Ridge回帰

★ 通常の回帰分析 (最小二乗法)

$$★ \sum_{k=1}^n (y_k - f(x_k; \beta))^2 \rightarrow \text{minimize}$$

★ Ridge回帰

$$★ \sum_{k=1}^n (y_k - f(x_k; \beta))^2 + \lambda \|\beta\|_2^2 \rightarrow \text{minimize}$$

$$★ \sum_{k=1}^n (y_k - f(x_k; \beta))^2 + \lambda (\beta_1^2 + \beta_2^2 + \dots + \beta_m^2) \rightarrow \text{minimize}$$

$$★ \text{解くべき連立一次方程式 } (A^T A + \lambda I)\beta = A^T y$$

★ 通常は切片にはペナルティを課さないことが多い

★  $\lambda > 0$  はパラメータ

## ★ 通常の回帰分析（最小二乗法）

$$★ \sum_{k=1}^n (y_k - f(x_k; \beta))^2 \rightarrow \text{minimize}$$

## ★ Lasso 回帰

$$★ \sum_{k=1}^n (y_k - f(x_k; \beta))^2 + \lambda \|\beta\|_1 \rightarrow \text{minimize}$$

$$★ \sum_{k=1}^n (y_k - f(x_k; \beta))^2 + \lambda (|\beta_1| + |\beta_2| + \dots + |\beta_m|) \rightarrow \text{minimize}$$

## ★ スパース推定

★ 係数の推定結果が0になりやすい

★ 0になった場合，その説明変数は不要だったのではないかと考えられる

# モデル選択

# モデル選択

## ★ 変数指定法

- ★ 理論的に正しいモデルが証明できる

## ★ 総当たり法

- ★ 候補になるモデルを全て試し、何らかの基準で最も良いと思われるものを採用

- ★ AIC,  $C_p$  基準, 自由度調整済み決定係数, クロスバリデーション

## ★ 逐次変数選択法

- ★ 一定の規則に従い、変数を追加したり削除する
- ★ 総当たり法における局所探索にあたる場合もある

## ★ ...

# 赤池情報量基準 AIC

★ 回帰分析の場合 (定数の差は気にしないことにして)

★  $AIC = n \log(S/n) + 2p$

★  $n$ : データ数

★  $S$ : 残差の二乗和

★  $p$ : パラメータの数

★ この値が小さい方が「良い」モデル

★ KL 情報量と最尤推定の理論から出てくるもの

★ 通常の線形回帰分析にのみ利用可能

★ WAIC (広く使える情報量規準)

## クロスバリデーション（交差検証）

- ★ 訓練用のデータと、テスト用のデータに分けて、訓練用のデータで学習（係数を推定）し、テスト用のデータを用いてどの程度良いかを評価する
- ★ 訓練用のデータと、テスト用のデータを変更し、それを繰り返す

### ★ leave-one-out 交差検証（LOOCV）

- ★ データ数を  $n$  とし、訓練用データ  $n - 1$  個、テスト用データ 1 個とする
- ★ テスト用データを取り替えながら  $n$  回繰り返す

### ★ K-fold 交差検証（K分割交差検証）

- ★ データを  $K$  個のグループに分け、訓練用データ  $K - 1$  グループ、テスト用データ 1 グループのデータとする
- ★ テスト用データを取り替えながら  $K$  回繰り返す

# 逐次変数選択法

## ★ 変数増加法

- ★ 最初，説明変数はなし
- ★ 説明変数を1つ加えたとき，その係数に対する検定の $P$ 値が最も小さいものを調べる
- ★ その $P$ 値の値が，一定値より小さければ加えて，同じ手順を繰り返す．一定値より大きければ終了する

## ★ 変数減少法

- ★ 最初，全ての説明変数を使ったモデルから考える
- ★ 説明変数の係数に対する検定の $P$ 値が最も大きいものを調べる
- ★ その $P$ 値の値が，一定値より大きければ説明変数から外し，同じ手順を繰り返す．一定値より小さければ終了する

## ★ 変数増減法

## 演習 - ビールの売上の予測



## 次にやること

- ★ データ数の割に説明変数をかなり増やしたのでRidge回帰やLasso回帰を試してみよう
  - ★ ここから先は、正直Excelでやるよりは、RやPythonなど他のソフトウェアなどを使用して行った方が楽で効率的にできる
- ★ Excelには汎用的に最適化問題を解いてくれるソルバーというアドインがあるのでそれを利用する

# ファイル

★ Ridge 回帰を見据えて整形したファイルを使用

★ [http://ds.k.kyoto-u.ac.jp/e-learning\\_files/  
data\\_analysis\\_basic/jma\\_005.xlsx](http://ds.k.kyoto-u.ac.jp/e-learning_files/data_analysis_basic/jma_005.xlsx)

★ PandA のリソースにも置いてあります

---

★ ワークシート関数を利用しているため xlsx 形式にしています

# ファイルを開いてみましょう

★ 前ページのxlsx ファイルを Excel で開いて内容を確認してみましょう

The screenshot shows an Excel spreadsheet titled "jma\_005.xlsx". The data is organized as follows:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		ビール	東京	京都	東京1月後	時間	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	定数項	残差			目的関数	
2	2011年1月	475	5.1	2.8	7	1	1	0	0	0	0	0	0	0	0	0	0	0	1		225625		60573216	
3	2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0	1		390625			
4	2011年3月	800	8.1	6.8	14.5	3	0	0	1	0	0	0	0	0	0	0	0	0	1		640000			
5	2011年4月	960	14.5	12.5	18.5	4	0	0	0	1	0	0	0	0	0	0	0	0	1		921600			
6	2011年5月	730	18.5	19	22.8	5	0	0	0	0	1	0	0	0	0	0	0	0	1		532900			
7	2011年6月	980	22.8	24.1	27.3	6	0	0	0	0	0	1	0	0	0	0	0	0	1		960400			
8	2011年7月	1295	27.3	27.9	27.5	7	0	0	0	0	0	0	1	0	0	0	0	0	1		1677025			
9	2011年8月	1135	27.5	28.7	25.1	8	0	0	0	0	0	0	0	1	0	0	0	0	1		1288225			
10	2011年9月	830	25.1	24.7	19.5	9	0	0	0	0	0	0	0	0	1	0	0	0	1		688900			
11	2011年10月	805	19.5	18.4	14.9	10	0	0	0	0	0	0	0	0	0	1	0	0	1		648025			
12	2011年11月	840	14.9	13.8	7.5	11	0	0	0	0	0	0	0	0	0	0	1	0	1		705600			
13	2011年12月	1375	7.5	6.5	4.8	12	0	0	0	0	0	0	0	0	0	0	0	1	1		1890625			
74	2017年1月	441	5.8	4.8	6.9	73	1	0	0	0	0	0	0	0	0	0	0	0	1		194481			
75	2017年2月	575	6.9	5.1	8.5	74	0	1	0	0	0	0	0	0	0	0	0	0	1		330625			
76	2017年3月	796	8.5	8.2	14.7	75	0	0	1	0	0	0	0	0	0	0	0	0	1		633616			
77																								
78	係数		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	lambda			
79	係数の二乗		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			0		
80																								
81																								

## アドインの追加

★ アドインの追加は例えば以下の手順で行います

★ ファイル → オプション → アドイン → 設定 → ソルバーにチェックを入れて OK を押す

★ 成功するとリボンのデータのタブにデータ分析が表示されます

# 最適化問題について

★ 高校の数学で以下のような問題を解いたことがあるかと思います

★  $3x + 4y$  の最大値を求めてください。

★ ただし、変数  $x, y$  は非負の実数で、 $x + y \leq 7, 2x + y \leq 9$  を満たす。

★ 最適化問題とは一般的に以下のような形で定義される問題

★ このような変数があるとある条件を満たしながら動くとき、とある関数の値の最大値（最小値）を求めてください

★ 最大化（最小化）する関数を **目的関数** という

★ 上の例では  $3x + 4y$

★ 条件を **制約条件** という

★ 上の例では  $x \geq 0, y \geq 0, x + y \leq 7, 2x + y \leq 9$

## 最適化問題について

★ ソルバーアドインでは汎用的に最適化問題を解いてくれる

★ いろいろな状況において便利なので是非使ってください

---

★ ただし、最適化問題は一般的には非常に解くのが難しい

★ ソルバーアドインが必ずしも正しい答えを求めているかどうかは怪しいので要検証

★ 特殊な条件を満たす、解きやすい問題、というクラスが色々研究されている

# 最適化問題について

★ 最小二乗法も最適化問題の一種

★ 連立一次方程式に帰着され、QR分解で解けるので、最適化問題の中では非常に解きやすい問題

★ Ridge回帰, Lasso回帰も最適化問題の一種

★ 主成分分析なども最適化問題として定式化され、データ分析に関する問題でも最適化問題は非常に多く登場する

## ソルバーアドインの使い方：準備

- ★ ソルバーアドインを使うために、Excel上で最適化問題を作る
  - ★ 残差の二乗和 + 正則化項の値を表現するセルを作る
- ★ まずは $\lambda$ を0に設定して、普通に最小二乗法を行ってみる



# ファイルの説明

★ 係数の部分に係数を格納することにします

★ 今は暫定的に全て0にしていますが、これをうまく動かして、例えば残差二乗和を最小化したい

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		ビール	東京	京都	東京1月後	時間	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	定数項	残差		目的関数		
2	2011年1月	475	5.1	2.8	7	1	1	0	0	0	0	0	0	0	0	0	0	0	1	225625		60573216		
3	2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0	1	390625				
4	2011年3月	800	8.1	6.8	14.5	3	0	0	1	0	0	0	0	0	0	0	0	0	1	640000				
5	2011年4月	960	14.5	12.5	18.5	4	0	0	0	1	0	0	0	0	0	0	0	0	1	921600				
6	2011年5月	730	18.5	19	22.8	5	0	0	0	0	1	0	0	0	0	0	0	0	1	532900				
7	2011年6月	980	22.8	24.1	27.3	6	0	0	0	0	0	1	0	0	0	0	0	0	1	960400				
8	2011年7月	1295	27.3	27.9	27.5	7	0	0	0	0	0	0	1	0	0	0	0	0	1	1677025				
9	2011年8月	1135	27.5	28.7	25.1	8	0	0	0	0	0	0	0	1	0	0	0	0	1	1288225				
10	2011年9月	830	25.1	24.7	19.5	9	0	0	0	0	0	0	0	0	1	0	0	0	1	688900				
11	2011年10月	805	19.5	18.4	14.9	10	0	0	0	0	0	0	0	0	0	1	0	0	1	648025				
12	2011年11月	840	14.9	13.8	7.5	11	0	0	0	0	0	0	0	0	0	0	1	0	1	705600				
13	2011年12月	1375	7.5	6.5	4.8	12	0	0	0	0	0	0	0	0	0	0	0	1	1	1890625				
74	2017年1月	441	5.8	4.8	6.9	73	1	0	0	0	0	0	0	0	0	0	0	0	1	194481				
75	2017年2月	575	6.9	5.1	8.5	74	0	1	0	0	0	0	0	0	0	0	0	0	1	330625				
76	2017年3月	796	8.5	8.2	14.7	75	0	0	1	0	0	0	0	0	0	0	0	0	1	633616				
77																								
78	係数		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	lambda			
79	係数の二乗		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
80																								

# ファイルの説明

★ 残差の部分では各データに対する残差の二乗を計算しています

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		ビール	東京	京都	東京1月後	時間	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	定数項	残差	目的関数			
2	2011年1月	475	5.1	2.8	7	1	1	0	0	0	0	0	0	0	0	0	0	0	1	225625	60573216			
3	2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0	1	390625				
4	2011年3月	800	8.1	6.8	14.5	3	0	0	1	0	0	0	0	0	0	0	0	0	1	640000				
5	2011年4月	960	14.5	12.5	18.5	4	0	0	0	1	0	0	0	0	0	0	0	0	1	921600				
6	2011年5月	730	18.5	19	22.8	5	0	0	0	0	1	0	0	0	0	0	0	0	1	532900				
7	2011年6月	980	22.8	24.1	27.3	6	0	0	0	0	0	1	0	0	0	0	0	0	1	960400				
8	2011年7月	1295	27.3	27.9	27.5	7	0	0	0	0	0	0	1	0	0	0	0	0	1	1677025				
9	2011年8月	1135	27.5	28.7	25.1	8	0	0	0	0	0	0	0	1	0	0	0	0	1	1288225				
10	2011年9月	830	25.1	24.7	19.5	9	0	0	0	0	0	0	0	0	1	0	0	0	1	688900				
11	2011年10月	805	19.5	18.4	14.9	10	0	0	0	0	0	0	0	0	0	1	0	0	1	648025				
12	2011年11月	840	14.9	13.8	7.5	11	0	0	0	0	0	0	0	0	0	0	1	0	1	705600				
13	2011年12月	1375	7.5	6.5	4.8	12	0	0	0	0	0	0	0	0	0	0	0	1	1	1890625				
74	2017年1月	441	5.8	4.8	6.9	73	1	0	0	0	0	0	0	0	0	0	0	0	1	194481				
75	2017年2月	575	6.9	5.1	8.5	74	0	1	0	0	0	0	0	0	0	0	0	0	1	330625				
76	2017年3月	796	8.5	8.2	14.7	75	0	0	1	0	0	0	0	0	0	0	0	0	1	633616				
77																								
78	係数		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	lambda				
79	係数の二乗		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0			
80																								

# ファイルの説明

★ 残差の部分では各データに対する残差の二乗を計算しています

★ 計算にはExcelのワークシート関数を利用しています

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		ビール	東京	京都	東京1年後	時間	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	定数項	残差	目的関数			
2	2011年1月	475	5.1	2.8	7	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	=POWER(B2-SUMPRODUCT(C2:S2,\$C\$78:\$S\$78),2)	60573216		
3	2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0	0	1	390625			
4	2011年3月	800	8.1	6.8	14.5	3	0	0	1	0	0	0	0	0	0	0	0	0	0	1	640000			
5	2011年4月	960	14.5	12.5	18.5	4	0	0	0	1	0	0	0	0	0	0	0	0	0	1	921600			
6	2011年5月	730	18.5	19	22.8	5	0	0	0	0	1	0	0	0	0	0	0	0	0	1	532900			
7	2011年6月	980	22.8	24.1	27.3	6	0	0	0	0	0	1	0	0	0	0	0	0	0	1	960400			
8	2011年7月	1295	27.3	27.9	27.5	7	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1677025			
9	2011年8月	1135	27.5	28.7	25.1	8	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1288225			
10	2011年9月	830	25.1	24.7	19.5	9	0	0	0	0	0	0	0	0	1	0	0	0	0	1	688900			
11	2011年10月	805	19.5	18.4	14.9	10	0	0	0	0	0	0	0	0	0	1	0	0	0	1	648025			
12	2011年11月	840	14.9	13.8	7.5	11	0	0	0	0	0	0	0	0	0	0	1	0	0	1	705600			
13	2011年12月	1375	7.5	6.5	4.8	12	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1890625			
74	2017年1月	441	5.8	4.8	6.9	73	1	0	0	0	0	0	0	0	0	0	0	0	0	1	194481			
75	2017年2月	575	6.9	5.1	8.5	74	0	1	0	0	0	0	0	0	0	0	0	0	0	1	330625			
76	2017年3月	796	8.5	8.2	14.7	75	0	0	1	0	0	0	0	0	0	0	0	0	0	1	633616			
77																								
78	係数		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	lambda		
79	係数の二乗		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
80																								

# ファイルの説明

★ lambda の部分では, Ridge 回帰や Lasso 回帰の正則化パラメータ  $\lambda$  を格納

★ 最初は 0 に設定しておき, 通常の最小二乗法を行います

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		ビール	東京	京都	東京1月後	時間	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	定数項	残差		目的関数		
2	2011年1月	475	5.1	2.8	7	1	1	0	0	0	0	0	0	0	0	0	0	0	1		225625	60573216		
3	2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0	1		390625			
4	2011年3月	800	8.1	6.8	14.5	3	0	0	1	0	0	0	0	0	0	0	0	0	1		640000			
5	2011年4月	960	14.5	12.5	18.5	4	0	0	0	1	0	0	0	0	0	0	0	0	1		921600			
6	2011年5月	730	18.5	19	22.8	5	0	0	0	0	1	0	0	0	0	0	0	0	1		532900			
7	2011年6月	980	22.8	24.1	27.3	6	0	0	0	0	0	1	0	0	0	0	0	0	1		960400			
8	2011年7月	1295	27.3	27.9	27.5	7	0	0	0	0	0	0	1	0	0	0	0	0	1		1677025			
9	2011年8月	1135	27.5	28.7	25.1	8	0	0	0	0	0	0	0	1	0	0	0	0	1		1288225			
10	2011年9月	830	25.1	24.7	19.5	9	0	0	0	0	0	0	0	0	1	0	0	0	1		688900			
11	2011年10月	805	19.5	18.4	14.9	10	0	0	0	0	0	0	0	0	0	1	0	0	1		648025			
12	2011年11月	840	14.9	13.8	7.5	11	0	0	0	0	0	0	0	0	0	0	1	0	1		705600			
13	2011年12月	1375	7.5	6.5	4.8	12	0	0	0	0	0	0	0	0	0	0	0	1	1		1890625			
74	2017年1月	441	5.8	4.8	6.9	73	1	0	0	0	0	0	0	0	0	0	0	0	1		194481			
75	2017年2月	575	6.9	5.1	8.5	74	0	1	0	0	0	0	0	0	0	0	0	0	1		330625			
76	2017年3月	796	8.5	8.2	14.7	75	0	0	1	0	0	0	0	0	0	0	0	0	1		633616			
77																								
78	係数		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	lambda			
79	係数の二乗		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
80																								

# ファイルの説明

- ★ 係数の二乗の部分では、Ridge 回帰の正則化項の値を計算するために、各係数の二乗を計算します

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		ビール	東京	京都	東京1月後	時間	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	定数項	残差			目的関数	
2	2011年1月	475	5.1	2.8	7	1	1	0	0	0	0	0	0	0	0	0	0	0	1	225625			60573216	
3	2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0	1	390625				
4	2011年3月	800	8.1	6.8	14.5	3	0	0	1	0	0	0	0	0	0	0	0	0	1	640000				
5	2011年4月	960	14.5	12.5	18.5	4	0	0	0	1	0	0	0	0	0	0	0	0	1	921600				
6	2011年5月	730	18.5	19	22.8	5	0	0	0	0	1	0	0	0	0	0	0	0	1	532900				
7	2011年6月	980	22.8	24.1	27.3	6	0	0	0	0	0	1	0	0	0	0	0	0	1	960400				
8	2011年7月	1295	27.3	27.9	27.5	7	0	0	0	0	0	0	1	0	0	0	0	0	1	1677025				
9	2011年8月	1135	27.5	28.7	25.1	8	0	0	0	0	0	0	0	1	0	0	0	0	1	1288225				
10	2011年9月	830	25.1	24.7	19.5	9	0	0	0	0	0	0	0	0	1	0	0	0	1	688900				
11	2011年10月	805	19.5	18.4	14.9	10	0	0	0	0	0	0	0	0	0	1	0	0	1	648025				
12	2011年11月	840	14.9	13.8	7.5	11	0	0	0	0	0	0	0	0	0	0	1	0	1	705600				
13	2011年12月	1375	7.5	6.5	4.8	12	0	0	0	0	0	0	0	0	0	0	0	1	1	1890625				
74	2017年1月	441	5.8	4.8	6.9	73	1	0	0	0	0	0	0	0	0	0	0	0	1	194481				
75	2017年2月	575	6.9	5.1	8.5	74	0	1	0	0	0	0	0	0	0	0	0	0	1	330625				
76	2017年3月	796	8.5	8.2	14.7	75	0	0	1	0	0	0	0	0	0	0	0	0	1	633616				
77																								
78	係数		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	lambda				
79	係数の二乗		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0			
80																								

# ファイルの説明

★ 係数の二乗の部分では，Ridge 回帰の正則化項の値を計算するために，各係数の二乗を計算します

★ 計算には Excel のワークシート関数を利用しています

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		ビール	東京	京都	東京1月後	時間	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	定数項	残差		目的関数		
2	2011年1月	475	5.1	2.8	7	1	1	0	0	0	0	0	0	0	0	0	0	0	1	225625		60573216		
3	2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0	1	390625				
4	2011年3月	800	8.1	6.8	14.5	3	0	0	1	0	0	0	0	0	0	0	0	0	1	640000				
5	2011年4月	960	14.5	12.5	18.5	4	0	0	0	1	0	0	0	0	0	0	0	0	1	921600				
6	2011年5月	730	18.5	19	22.8	5	0	0	0	0	1	0	0	0	0	0	0	0	1	532900				
7	2011年6月	980	22.8	24.1	27.3	6	0	0	0	0	0	1	0	0	0	0	0	0	1	960400				
8	2011年7月	1295	27.3	27.9	27.5	7	0	0	0	0	0	0	1	0	0	0	0	0	1	1677025				
9	2011年8月	1135	27.5	28.7	25.1	8	0	0	0	0	0	0	0	1	0	0	0	0	1	1288225				
10	2011年9月	830	25.1	24.7	19.5	9	0	0	0	0	0	0	0	0	1	0	0	0	1	688900				
11	2011年10月	805	19.5	18.4	14.9	10	0	0	0	0	0	0	0	0	0	1	0	0	1	648025				
12	2011年11月	840	14.9	13.8	7.5	11	0	0	0	0	0	0	0	0	0	0	1	0	1	705600				
13	2011年12月	1375	7.5	6.5	4.8	12	0	0	0	0	0	0	0	0	0	0	0	1	1	1890625				
74	2017年1月	441	5.8	4.8	6.9	73	1	0	0	0	0	0	0	0	0	0	0	0	1	194481				
75	2017年2月	575	6.9	5.1	8.5	74	0	1	0	0	0	0	0	0	0	0	0	0	1	330625				
76	2017年3月	796	8.5	8.2	14.7	75	0	0	1	0	0	0	0	0	0	0	0	0	1	633616				
77																								
78	係数		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	lambda			
79	係数の二乗		C78,2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0			
80																								

# ファイルの説明

★最後に、目的関数の部分で、Ridge回帰の目的関数の値を計算しています

The screenshot shows an Excel spreadsheet titled 'jma\_005.xlsx'. The data is organized as follows:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		ビール	東京	京都	東京1月後	時間	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	定数項	残差			目的関数	
2	2011年1月	475	5.1	2.8	7	1	1	0	0	0	0	0	0	0	0	0	0	0	1		225625		60573216	
3	2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0	1		390625			
4	2011年3月	800	8.1	6.8	14.5	3	0	0	1	0	0	0	0	0	0	0	0	0	1		640000			
5	2011年4月	960	14.5	12.5	18.5	4	0	0	0	1	0	0	0	0	0	0	0	0	1		921600			
6	2011年5月	730	18.5	19	22.8	5	0	0	0	0	1	0	0	0	0	0	0	0	1		532900			
7	2011年6月	980	22.8	24.1	27.3	6	0	0	0	0	0	1	0	0	0	0	0	0	1		960400			
8	2011年7月	1295	27.3	27.9	27.5	7	0	0	0	0	0	0	1	0	0	0	0	0	1		1677025			
9	2011年8月	1135	27.5	28.7	25.1	8	0	0	0	0	0	0	0	1	0	0	0	0	1		1288225			
10	2011年9月	830	25.1	24.7	19.5	9	0	0	0	0	0	0	0	0	1	0	0	0	1		688900			
11	2011年10月	805	19.5	18.4	14.9	10	0	0	0	0	0	0	0	0	0	1	0	0	1		648025			
12	2011年11月	840	14.9	13.8	7.5	11	0	0	0	0	0	0	0	0	0	0	1	0	1		705600			
13	2011年12月	1375	7.5	6.5	4.8	12	0	0	0	0	0	0	0	0	0	0	0	1	1		1890625			
74	2017年1月	441	5.8	4.8	6.9	73	1	0	0	0	0	0	0	0	0	0	0	0	1		194481			
75	2017年2月	575	6.9	5.1	8.5	74	0	1	0	0	0	0	0	0	0	0	0	0	1		330625			
76	2017年3月	796	8.5	8.2	14.7	75	0	0	1	0	0	0	0	0	0	0	0	0	1		633616			
77																								
78	係数		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	lambda				
79	係数の二乗		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0			
80																								

# ファイルの説明

★ 最後に，目的関数の部分で，Ridge回帰の目的関数の値を計算しています

★ 計算にはExcelのワークシート関数を利用しています

The screenshot shows an Excel spreadsheet with the following data structure:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		ビール	東京	京都	東京1月後	時間	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	定数項	残差		目的関数		
2	2011年1月	475	5.1	2.8	7	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	225625		R79)	
3	2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0	0	1	390625			
4	2011年3月	800	8.1	6.8	14.5	3	0	0	1	0	0	0	0	0	0	0	0	0	0	1	640000			
5	2011年4月	960	14.5	12.5	18.5	4	0	0	0	1	0	0	0	0	0	0	0	0	0	1	921600			
6	2011年5月	730	18.5	19	22.8	5	0	0	0	0	1	0	0	0	0	0	0	0	0	1	532900			
7	2011年6月	980	22.8	24.1	27.3	6	0	0	0	0	0	1	0	0	0	0	0	0	0	1	960400			
8	2011年7月	1295	27.3	27.9	27.5	7	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1677025			
9	2011年8月	1135	27.5	28.7	25.1	8	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1288225			
10	2011年9月	830	25.1	24.7	19.5	9	0	0	0	0	0	0	0	0	1	0	0	0	0	1	688900			
11	2011年10月	805	19.5	18.4	14.9	10	0	0	0	0	0	0	0	0	0	1	0	0	0	1	648025			
12	2011年11月	840	14.9	13.8	7.5	11	0	0	0	0	0	0	0	0	0	0	1	0	0	1	705600			
13	2011年12月	1375	7.5	6.5	4.8	12	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1890625			
74	2017年1月	441	5.8	4.8	6.9	73	1	0	0	0	0	0	0	0	0	0	0	0	0	1	194481			
75	2017年2月	575	6.9	5.1	8.5	74	0	1	0	0	0	0	0	0	0	0	0	0	0	1	330625			
76	2017年3月	796	8.5	8.2	14.7	75	0	0	1	0	0	0	0	0	0	0	0	0	0	1	633616			
77																								
78	係数		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	lambda		
79	係数の二乗		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
80																								



## ソルバーアドインの使い方

- ★ データのリボンに追加されているソルバーを選択し
  - ★ 目的セルの設定で「\$W\$2」と入力（\$はなくても構いません）
  - ★ 目標値の設定で「最小値」にチェック
  - ★ 変数セルの変更で「\$C\$78:\$S\$78」と入力（\$はなくても構いません）
  - ★ 制約のない変数を非負数にするのチェックを外す
- ★ とやって、解決を選択すれば良い

# ソルバーアドインの使い方

★ データのリボンに追加されているソルバーを選択し

The screenshot shows the Microsoft Excel interface with the 'データ' (Data) ribbon selected. The 'ソルバー' (Solver) button is circled in red and labeled with a '2'. A red circle labeled '1' highlights the 'データ' ribbon itself. The spreadsheet below shows a table with columns for years and months, and rows for various data points and coefficients.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		ビール	東京	京都	東京1月後	時間	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	定数項		残差		目的関数	
2	2011年1月	475	5.1	2.8	7	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	225625		60573216	
3	2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0	0	1	390625			
4	2011年3月	800	8.1	6.8	14.5	3	0	0	1	0	0	0	0	0	0	0	0	0	0	1	640000			
5	2011年4月	960	14.5	12.5	18.5	4	0	0	0	1	0	0	0	0	0	0	0	0	0	1	921600			
6	2011年5月	730	18.5	19	22.8	5	0	0	0	0	1	0	0	0	0	0	0	0	0	1	532900			
7	2011年6月	980	22.8	24.1	27.3	6	0	0	0	0	0	1	0	0	0	0	0	0	0	1	960400			
8	2011年7月	1295	27.3	27.9	27.5	7	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1677025			
9	2011年8月	1135	27.5	28.7	25.1	8	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1288225			
10	2011年9月	830	25.1	24.7	19.5	9	0	0	0	0	0	0	0	0	1	0	0	0	0	1	688900			
11	2011年10月	805	19.5	18.4	14.9	10	0	0	0	0	0	0	0	0	0	1	0	0	0	1	648025			
12	2011年11月	840	14.9	13.8	7.5	11	0	0	0	0	0	0	0	0	0	0	1	0	0	1	705600			
13	2011年12月	1375	7.5	6.5	4.8	12	0	0	0	0	0	0	0	0	0	0	0	1	1	1890625				
74	2017年1月	441	5.8	4.8	6.9	73	1	0	0	0	0	0	0	0	0	0	0	0	0	1	194481			
75	2017年2月	575	6.9	5.1	8.5	74	0	1	0	0	0	0	0	0	0	0	0	0	0	1	330625			
76	2017年3月	796	8.5	8.2	14.7	75	0	0	1	0	0	0	0	0	0	0	0	0	0	1	633616			
77																								
78	係数		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	lambda			
79	係数の二乗		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
80																								
81																								

# ソルバーアドインの使い方

★ 目的セルの設定で「\$W\$2」と入力（\$はなくても構いません）

目的セルの設定:(I)

目標値:  最大値(M)  最小値(N)  指定値:(V)

変数セルの変更:(B)

制約条件の対象:(U)

制約のない変数を非負数にする(K)

解決方法の選択:  オプション(E)

解決方法  
滑らかな非線形を示すソルバー問題には GRG 非線形エンジン、線形を示すソルバー問題には LP シンプレックス エンジン、滑らかではない非線形を示すソルバー問題にはエボリューションナリー エンジンを選択してください。

ヘルプ(H)

	A	B	C	D	E
1		ビール	東京	京都	東京1月後
2	2011年1月	475	5.1	2.8	
3	2011年2月	625	7	6.3	8
4	2011年3月	800	8.1	6.8	14
5	2011年4月	960	14.5	12.5	18
6	2011年5月	730	18.5	19	22
7	2011年6月	980	22.8	24.1	27
8	2011年7月	1295	27.3	27.9	27
9	2011年8月	1135	27.5	28.7	25
10	2011年9月	830	25.1	24.7	19
11	2011年10月	805	19.5	18.4	14
12	2011年11月	840	14.9	13.8	7
13	2011年12月	1375	7.5	6.5	4
74	2017年1月	441	5.8	4.8	6
75	2017年2月	575	6.9	5.1	8
76	2017年3月	796	8.5	8.2	14
77					
78	係数		0	0	
79	係数の二乗		0	0	
80					
81					

# ソルバーアドインの使い方

## ★ 目標値の設定で「最小値」にチェック

目的セルの設定:(I)

目標値:  最大値(M)  最小値(N)  指定値:(Y)

変数セルの変更:(B)  
\$C\$78:\$S\$78

制約条件の対象:(U)

制約のない変数を非負数にする(K)

解決方法の選択:  
(E) GRG 非線形

解決方法  
滑らかな非線形を示すソルバー問題には GRG 非線形エンジン、線形を示すソルバー問題には LP シンプレックス エンジン、滑らかではない非線形を示すソルバー問題にはエボリューションナリー エンジンを選択してください。

# ソルバーアドインの使い方

★ 変数セルの変更で「\$C\$78:\$S\$78」と入力（\$はなくても構いません）

The screenshot shows the 'Solver Parameters' dialog box in Microsoft Excel. The 'Set Objective' field is set to '\$W\$2'. The 'To: Of' radio button is selected. The 'Change Variable Cells' field is set to '\$C\$78:\$S\$78', with a red arrow pointing to it. The 'GRG Nonlinear' engine is selected. The 'Solve' button is highlighted.

U1	A	B	C	D	E
1		ビール	東京	京都	東京1月後
2	2011年1月	475	5.1	2.8	
3	2011年2月	625	7	6.3	8
4	2011年3月	800	8.1	6.8	14
5	2011年4月	960	14.5	12.5	18
6	2011年5月	730	18.5	19	22
7	2011年6月	980	22.8	24.1	27
8	2011年7月	1295	27.3	27.9	27
9	2011年8月	1135	27.5	28.7	25
10	2011年9月	830	25.1	24.7	19
11	2011年10月	805	19.5	18.4	14
12	2011年11月	840	14.9	13.8	7
13	2011年12月	1375	7.5	6.5	4
74	2017年1月	441	5.8	4.8	6
75	2017年2月	575	6.9	5.1	8
76	2017年3月	796	8.5	8.2	14
77					
78	係数		0	0	
79	係数の二乗		0	0	
80					
81					

# ソルバーアドインの使い方

★ 制約のない変数を非負数にするのチェックを外す

The screenshot shows the Excel Solver Parameters dialog box. The 'Make Unconstrained Variables Non-Negative' checkbox is unchecked, as indicated by a red arrow. The dialog box is set to minimize the target cell \$W\$2, with the variable cells set to \$C\$78:\$S\$78. The GRG Nonlinear engine is selected as the solving method.

U1	A	B	C	D	E
1		ビール	東京	京都	東京1月後
2	2011年1月	475	5.1	2.8	
3	2011年2月	625	7	6.3	8
4	2011年3月	800	8.1	6.8	14
5	2011年4月	960	14.5	12.5	18
6	2011年5月	730	18.5	19	22
7	2011年6月	980	22.8	24.1	27
8	2011年7月	1295	27.3	27.9	27
9	2011年8月	1135	27.5	28.7	25
10	2011年9月	830	25.1	24.7	19
11	2011年10月	805	19.5	18.4	14
12	2011年11月	840	14.9	13.8	7
13	2011年12月	1375	7.5	6.5	4
74	2017年1月	441	5.8	4.8	6
75	2017年2月	575	6.9	5.1	8
76	2017年3月	796	8.5	8.2	14
77					
78	係数		0	0	
79	係数の二乗		0	0	
80					
81					

# ソルバーアドインの使い方

★ とやって、解決を選択すれば良い

ソルバーのパラメーター

目的セルの設定:(I) \$W\$2

目標値:  最大値(M)  最小値(N)  指定値:(Y) 0

変数セルの変更:(B) \$C\$78:\$S\$78

制約条件の対象:(U)

制約のない変数を非負数にする:(K)

解決方法の選択:(E) GRG 非線形 オプション(P)

解決方法  
滑らかな非線形を示すソルバー問題には GRG 非線形エンジン、線形を示すソルバー問題には LP シンプレックス エンジン、滑らかではない非線形を示すソルバー問題にはエボリューションナリー エンジンを選択してください。

ヘルプ(H) 解決(S) 閉じる(O)

U1	残差
1	
2	ビール 東京 京都 東京1月後
3	2011年1月 475 5.1 2.8
4	2011年2月 625 7 6.3 8
5	2011年3月 800 8.1 6.8 14
6	2011年4月 960 14.5 12.5 18
7	2011年5月 730 18.5 19 22
8	2011年6月 980 22.8 24.1 27
9	2011年7月 1295 27.3 27.9 27
10	2011年8月 1135 27.5 28.7 25
11	2011年9月 830 25.1 24.7 19
12	2011年10月 805 19.5 18.4 14
13	2011年11月 840 14.9 13.8 7
14	2011年12月 1375 7.5 6.5 4
15	2017年1月 441 5.8 4.8 6
16	2017年2月 575 6.9 5.1 8
17	2017年3月 796 8.5 8.2 14
18	
19	
20	
21	
22	
23	
24	
25	
26	
27	
28	
29	
30	
31	
32	
33	
34	
35	
36	
37	
38	
39	
40	
41	
42	
43	
44	
45	
46	
47	
48	
49	
50	
51	
52	
53	
54	
55	
56	
57	
58	
59	
60	
61	
62	
63	
64	
65	
66	
67	
68	
69	
70	
71	
72	
73	
74	
75	
76	
77	
78	
79	
80	
81	
82	
83	
84	
85	
86	
87	
88	
89	
90	
91	
92	
93	
94	
95	
96	
97	
98	
99	
100	

# 結果

- ★ 上手く実行できれば最適解が求まり、変数のセルに値が上書きされる
- ★ 実際には最適解が上手く求まっていない場合もある
- ★ 初期値（変数セルの値）を変更してから実行したり、ソルバーアドインの設定を変えると求まることもある

The screenshot shows the Microsoft Excel interface with the Solver dialog box open. The dialog box displays the message "大域解に確率収束しました。" (Global solution reached with probability convergence). The "Solver Parameters" section is visible, and the "OK" button is highlighted with a red arrow. The spreadsheet in the background shows a table with columns for months and values. A red box highlights the row containing coefficients, with some cells displaying "###".

U1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		ビール	東京	京都	東京																残差		目的関数	
2	2011年1月	475	5.1	2.8																	1016.53706		130367.9	
3	2011年2月	625	7	6.3																	623.9536641			
4	2011年3月	800	8.1	6.8																	364.0107769			
5	2011年4月	960	14.5	12.5																	8830.872461			
6	2011年5月	730	18.5	19																	9423.665962			
7	2011年6月	980	22.8	24.1																	657.984536			
8	2011年7月	1295	27.3	27.9																	7530.950702			
9	2011年8月	1135	27.5	28.7																	672.1303517			
10	2011年9月	830	25.1	24.7																	126.7319624			
11	2011年10月	805	19.5	18.4																	1719.118392			
12	2011年11月	840	14.9	13.8																	644.0903586			
13	2011年12月	1375	7.5	6.5																	407.2597721			
74	2017年1月	441	5.8	4.8																	6.604626086			
75	2017年2月	575	6.9	5.1		8.5	74	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
76	2017年3月	796	8.5	8.2		14.7	75	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	
77																								
78	係数	21.005	-5.24	7.8267003	-0.9	-120	-6.6	93	4.9	-117	-36	91	12	-204	-55	95	724	480.97		lambda				
79	係数の二乗	441.2	27.45	61.257238	0.9	###	43	###	24	###	###	###	###	###	###	###	###	###	###	###	###	###	###	
80																								

※セルのサイズの問題で一部のセルに###と表示されているがセルを広げたり、セルを選択すれば値が表示される



## 試行錯誤

★  $\lambda$  の値を変更して，Ridge 回帰を行ってみよ

★  $\lambda$  の値を変化させることで，係数の推定結果はどのように変化するか

★ 目的セルの式を変更して，Lasso 回帰を行ってみよ

★  $\lambda$  の値を大きくすると，係数の推定結果がどのように変化するか

★ Ridge 回帰，Lasso 回帰を Excel のソルバーアドインで解くと，結構な割合で最適解が正しく求まらないこともわかる

★ 最適化問題は非常に色々な場面でよく出てくるのに，一般的に解くのは難しい

★ (問題のクラスに特化した) 最適化問題の解き方を考えるのは重要

## さらなる発展（おまけ）

### ★ モデル選択

- ★ クロスバリデーションやAICなどの基準を用いる

- ★ ExcelではVBA（マクロ機能）を用いてプログラミングもできる

---

- ★ 今回の演習の資料の別バージョンのラストに多少書いてあります

- ★ [http://ds.k.kyoto-u.ac.jp/e-learning\\_files/  
data\\_analysis\\_basic/regression\\_ex.pdf](http://ds.k.kyoto-u.ac.jp/e-learning_files/data_analysis_basic/regression_ex.pdf)