船舶海洋システム工学コース「計算工学演習第一」

EXCELによる行列計算・多重回帰

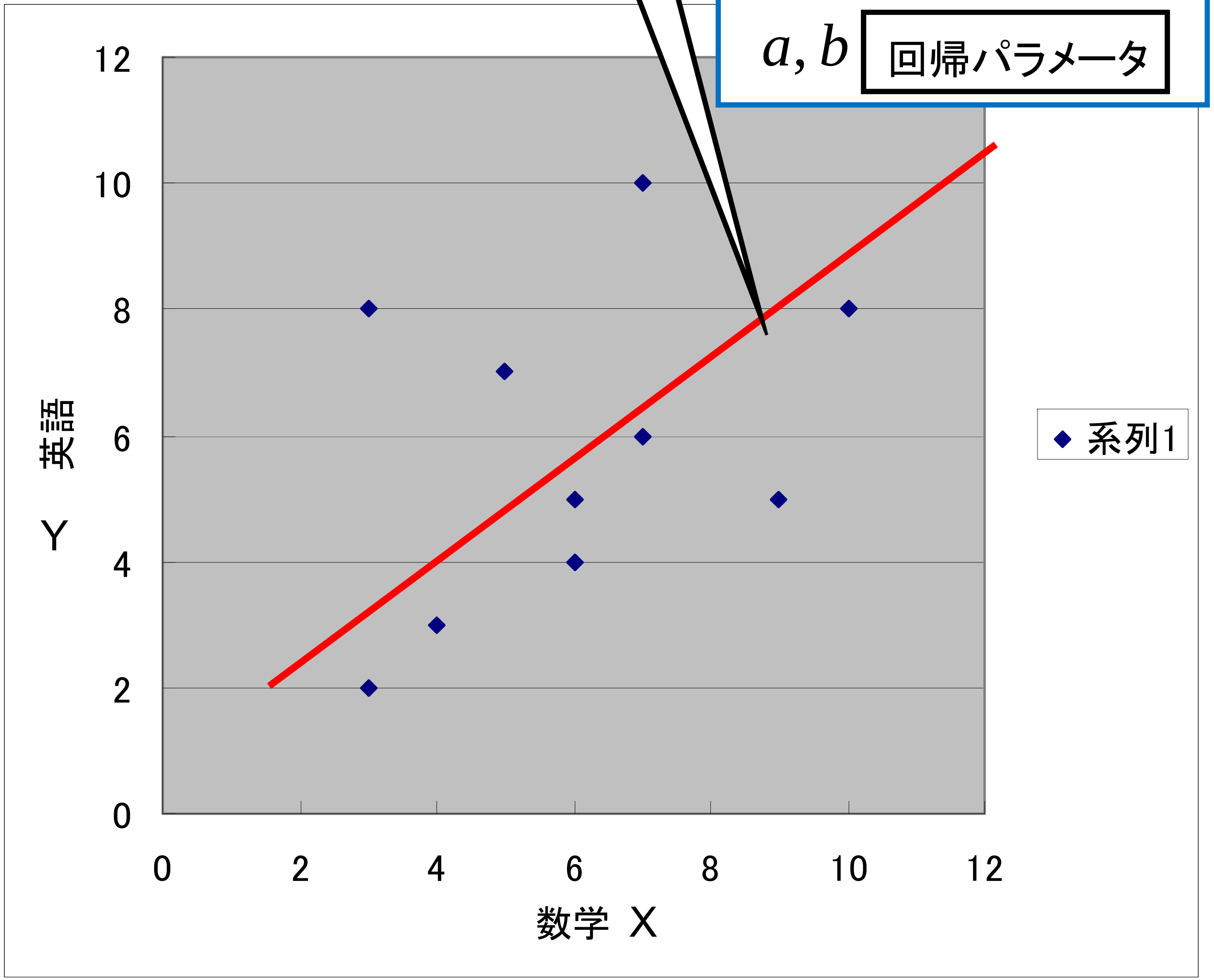
海洋システム工学部門
木村 元

(H28.12.10)

【復習】回帰分析

２変量の関係として**直線をあてはめる**

数学 x	英語 y
3	8
6	4
10	8
4	3
7	6
7	10
3	2
9	5
6	5
5	7



$y = ax + b$

x 回帰変数

y 被回帰変数

a, b 回帰パラメータ

【復習】回帰分析

回帰直線

$y = ax + b$　　yのxに対する回帰直線

回帰パラメータa,bの求め方

n 個のデータの組を $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ と表す

データの各点から同じ x_i の回帰直線までの距離を d_i とし、

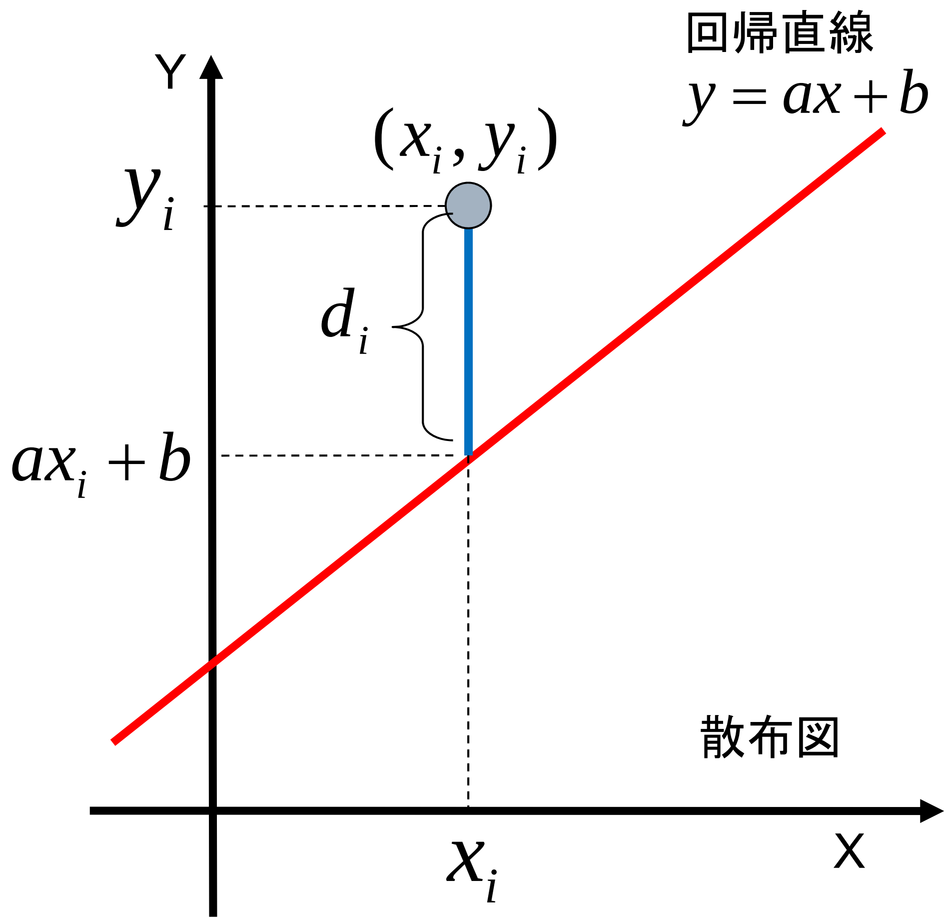
この長さの2乗和 L を最小にするように

回帰パラメータ a, b を決める
（**最小2乗法**）

$$L = \sum_{i=1}^{n} d_i^{\,2} = \boxed{\sum_{i=1}^{n} \left(y_i - (ax_i + b)\right)^2}$$

$$\frac{\partial L}{\partial a} = 0, \quad \frac{\partial L}{\partial b} = 0$$

の連立1次方程式を解いて a, b を求める



x の平均 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ y の平均 $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ とすると、

$$a = \boxed{\frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}}{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2}} \qquad b = \boxed{\bar{y} - a\bar{x}}$$

ここで、

$$S_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

x の分散

$$S_{yy} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

y の分散

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \boxed{\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \bar{x}\,\bar{y}}$$

共分散 Covariance

とおくと、

$$a = \boxed{\frac{S_{xy}}{S_{xx}}}$$

となり、求める直線は

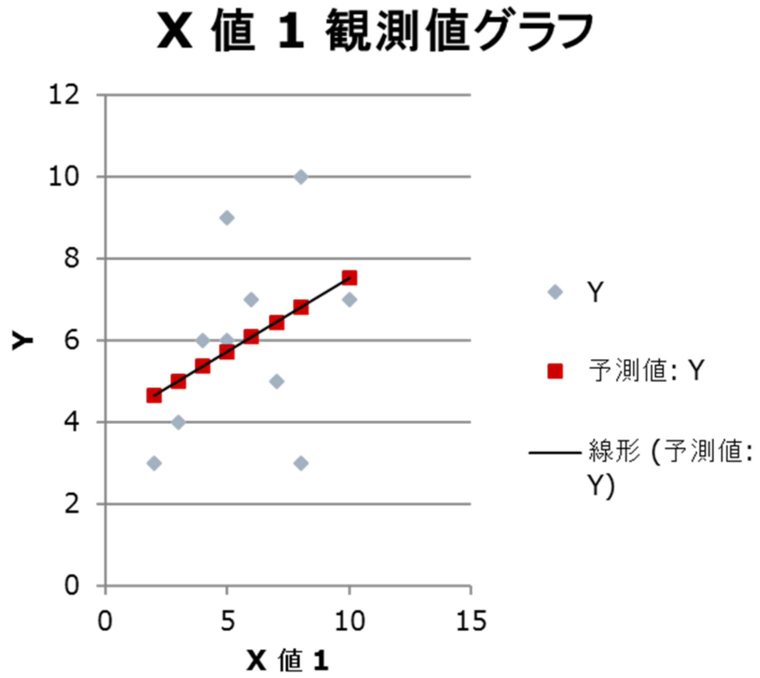
$$y - \bar{y} = \frac{S_{xy}}{S_{xx}}(x - \bar{x})$$

すなわち

点 $(\bar{x}, \bar{y})$ を通り、傾き $\frac{S_{xy}}{S_{xx}}$ の直線である

□ 演習問題(1)

以下のデータをエクセルに入力し、
散布図と**回帰直線**を表示せよ。



数学 x	英語 y
3	8
6	4
10	8
4	3
7	6
7	10
3	2
9	5
6	5
5	7

【エクセルで回帰分析を行うには】

(1) アドインを組み込む
[ファイル]→[オプション]→[アドイン]を開き、
「分析ツール」にチェックを入れて[設定]ボタンを押す
ここでアドイン一覧が出るので[分析ツール]をチェックして[OKボタン]を押す

(2) [データ]→[データ分析]→[回帰分析]を選択

多重回帰

観測値 y を変数 $x_1, x_2, \cdots x_K$ を用いて以下の式で説明する:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_K x_K + e$$

確率変動・誤差

このとき、n 個の観測値 $(y_1, x_{11}, x_{12}, \cdots x_{1K}), (y_2, x_{21}, x_{22}, \cdots x_{2K}), (y_n, x_{n1}, x_{n2}, \cdots x_{nK})$
によって係数 $b_0, b_1, b_2, \cdots b_K$ の最小2乗推定量を求める。ここで、

目的変数行列

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

説明変数行列

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1K} \\ 1 & x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nK} \end{bmatrix}$$

回帰係数行列

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_K \end{bmatrix}$$

誤差変数行列

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

と表すと、

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

線形表現

誤差変数行列 $\mathbf{e}$ の平方和 $\|\mathbf{e}\|^2$ を最小にする $\mathbf{b}$ を求める

$\rightarrow$ 回帰推定(最小2乗法)　回帰モデル

データから回帰モデルを得て何がうれしいか？
▪ 回帰モデルによる推定

未知の説明変数（回帰変数）の値が $(x_{q1}, x_{q2}, \cdots x_{qK})$ で与えられたときの

目的変数（被回帰変数）の値 y_q をデータから**推定**できる！

$$y_q = b_0 + b_1 x_{q1} + b_2 x_{q2} + \cdots + b_K x_{qK}$$

推定値

誤差eの項はゼロで計算

それでは、
回帰係数行列 $\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_K \end{bmatrix}$ をデータからどのように求めるか？

誤差ベクトル $\mathbf{e}$ の平方和 $\|\mathbf{e}\|^2$ を最小にする $\mathbf{b}$ を最尤推定値 $\hat{\mathbf{b}}$ と表すと、

単純回帰の場合と同様に、回帰係数の各要素で誤差ベクトルの平方和を偏微分し、これらが全てゼロとした連立方程式を立てて解くことにより、回帰係数ベクトルは以下の式で計算される:

$$\hat{\mathbf{b}} = \underbrace{\left(\mathbf{X}^{\mathrm{Trans}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{Trans}}}_{\text{Xの擬似逆行列 } \mathbf{X}^{+}}\mathbf{y}$$

pseudo-inverse matrix

ただしXはm行n列、m > n

ただし $\mathbf{X}^{\mathrm{Trans}}$ は $\mathbf{X}$ の転置行列を表す。

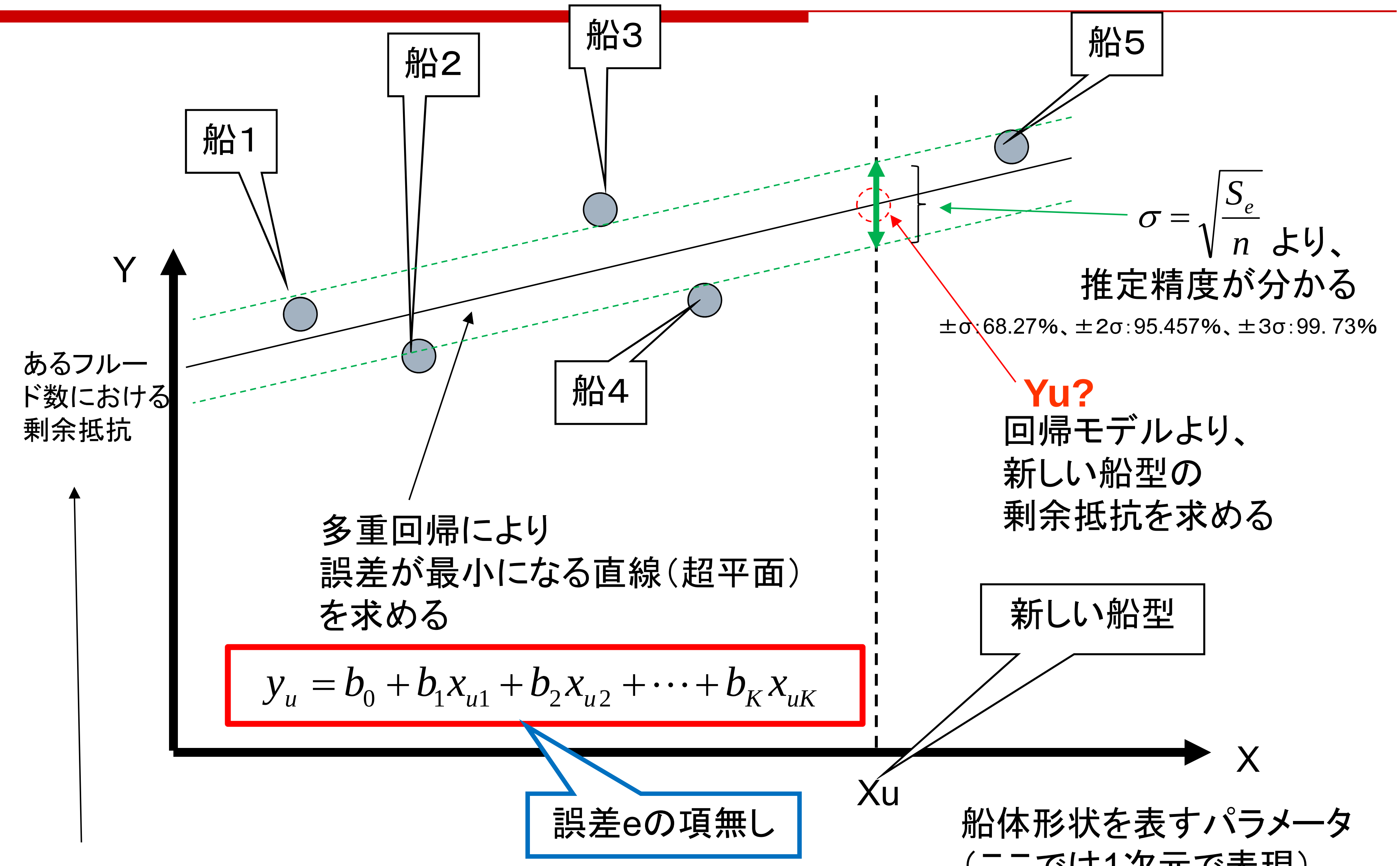
$\|\mathbf{e}\|^2$ の最小値 S_e を残差平方和といい、

$$S_e = \mathbf{y}^{\mathrm{Trans}}\left\{\mathbf{I} - \mathbf{X}\left(\mathbf{X}^{\mathrm{Trans}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{Trans}}\right\}\mathbf{y}$$

±σ の範囲内に68.27%のデータが存在

で与えられる。 ➡ $\sigma = \sqrt{\dfrac{S_e}{n}}$ より、回帰で推定する場合の精度が分かる

多重回帰を利用した新船型の剰余抵抗値の推定



営業フルード数周辺を0.01刻みで
それぞれ多重回帰モデルを生成する

多重回帰のデータ

自動車の燃費について以下のようなデータがある。「クラウン」から「ギャランΣ」までのデータを利用して
車種「ルーチェ」の10モード走行性能 y を予測する。

車名	x1	x2	x3	X4	x5	x6	y
クラウン	1.360	4.778	2.4251	125	17.5	8.8	8.7
マークII	1.245	4.100	2.4082	125	17.5	8.8	9.5
カムリ	1.070	3.214	2.3575	120	17.6	8.7	10.6
ソアラ	1.235	4.100	2.3052	125	17.5	8.8	9.2
セドリック	1.420	4.625	2.4251	130	17.5	9.5	8.9
ローレル	1.175	3.889	2.3660	125	17.0	9.1	9.2
スカイライン	1.175	4.111	2.3198	125	17.0	9.1	9.2
レパード	1.220	3.900	2.2899	125	17.0	9.1	9.4
カペラ	1.030	3.450	2.3829	120	17.0	8.6	10.2
ギャランΣ	1.180	3.665	2.3645	110	16.7	8.5	10.6
ルーチェ	1.150	3.909	2.3829	120	17.0	8.6	?

x1:車体重量(1000kg)， x2:減速比， x3:幅×高さ(m^2)， x4:最大出力(ps)，
x5:最大トルク(kgm)， x6:圧縮比， y:10モード走行(km/ℓ)

早川 毅 著「回帰分析の基礎」朝倉書店（1986）より引用

□ 演習問題(2)

（１） 前ページの自動車の燃費データに関して多重回帰を行い、回帰係数や残差平方和を計算せよ。

（２） 次に上で求めた回帰係数を利用して「ルーチェ」の燃費ｙの推定を行い、「yの推定値±σ」がどうなるか計算せよ。

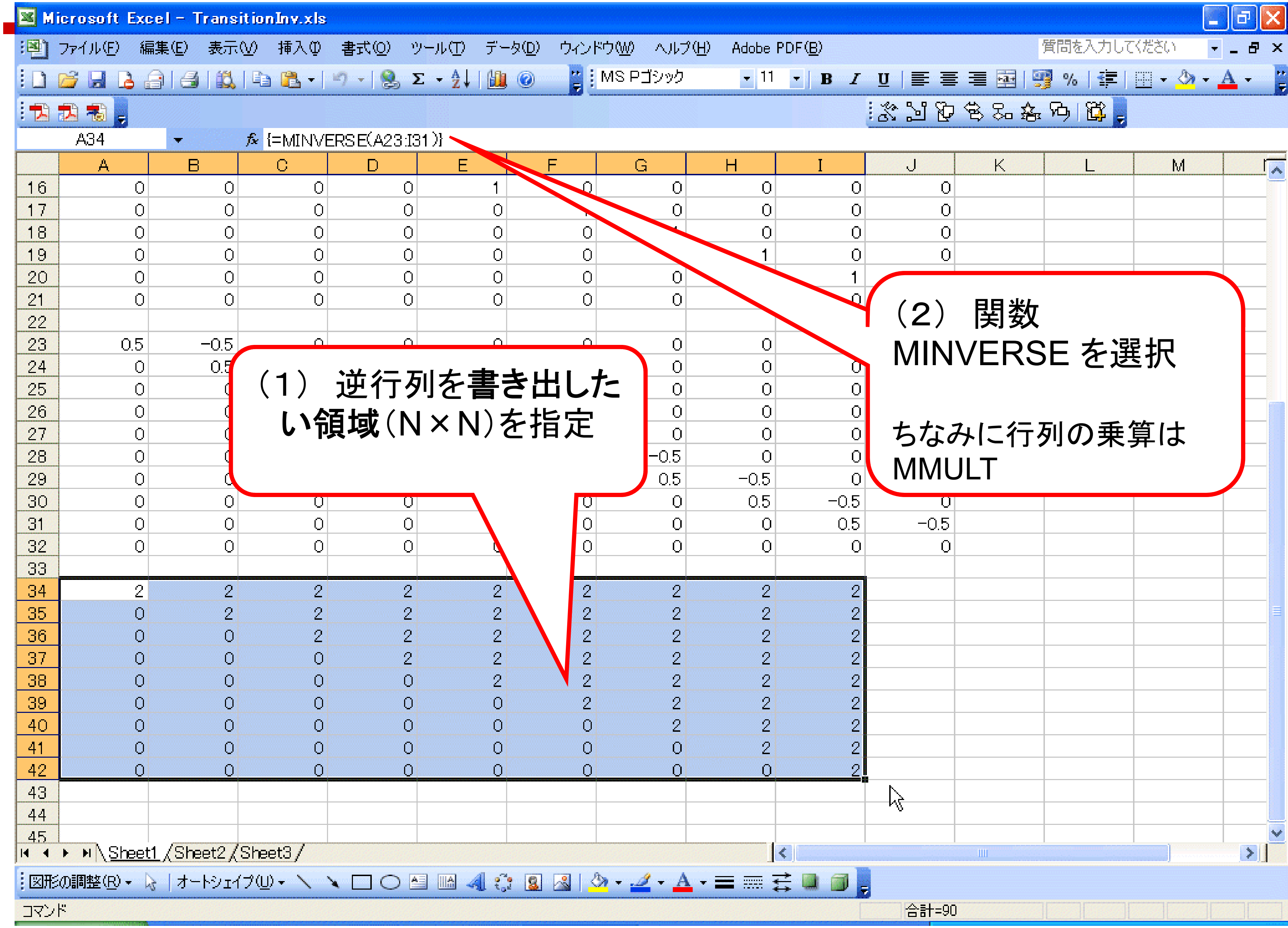
直接計算結果の数値を入力するのではなく、
データの修正に対しても自動で再計算するように関数やマクロを用いよ。

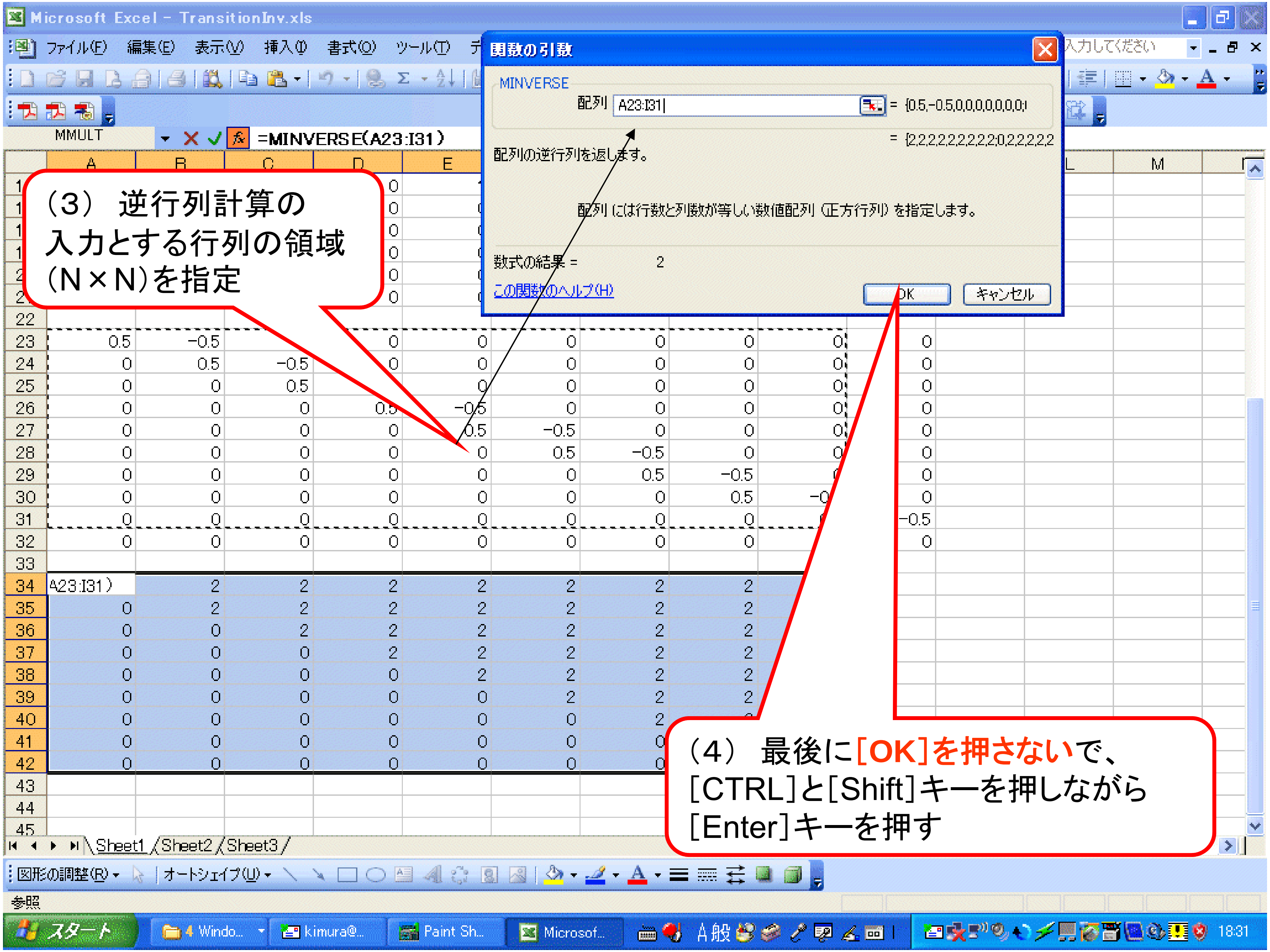
演習の提出について

作成したエクセルファイルを、前回の演習で作成した九大全学ファイル共有システム
http://www.m.kyushu-u.ac.jp/share/ の演習専用のフォルダへ追加でアップロードせよ。

前回提出した演習と区別できるよう「第2回演習.xls」等の名前を付け、
またワークシートの左上に自分の氏名と学籍番号を記入しておくこと。

【参考】 Excel2003で逆行列を求める方法





（3） 逆行列計算の
入力とする行列の領域
（N×N）を指定

（4） 最後に[OK]を押さないで、
[CTRL]と[Shift]キーを押しながら
[Enter]キーを押す