データ解析

目標：７日間でゼロから多変量解析の勘所をマスター！

夏ターム・木曜2限・3限 (10:30-12:00,13:00-14:30)

教室：A33

担当教員：瀧川 一学(たきがわ・いちがく)
情報理工学コース・大規模知識処理研究室
http://art.ist.hokudai.ac.jp/~takigawa/

連絡先：takigawa@ist.hokudai.ac.jp
情報科学研究科棟6F 6-16室

教科書

授業計画

Prologue: データを読み解くとは何なのか
(多変量解析とデータサイエンスと統計学とパターン認識と機械学習とデータマイニング)

DAY-1 6/16 (01)(02) **単回帰: 点群への直線当てはめを”真剣に”考える**
(見えない世界へようこそ)
DAY-2 6/23 (03)(04) **重回帰と線形代数: 回帰の行列計算とその意味**
(データの計算とデータの解釈)
DAY-3 6/30 (05)(06) **重回帰と確率統計: なぜ回帰に確率が必要?**
(推測統計入門:データの向こう側について語るための代償)
DAY-4 7/07 (07)(08) **多変量正規分布: 多次元の正規分布と線形代数**
(ゼロから理解する正規分布)
DAY-5 7/14 (09)(10) **マハラノビス距離と判別分析: 線形代数を使う1**
(最適な判別とは)
DAY-6 7/28 (11)(12) **固有値分解と主成分分析: 線形代数を使う2**
(高次元データがかかえる大問題)
DAY-7 8/04 (13)(14) **特異値分解と数量化: 線形代数を使う3**
(数値じゃない対象に統計を効かすには)

Epilogue: 基礎の上に在る世界(話したことと話さなかったこと)

この講義で主として扱うこと

線形代数を使った多次元の変量(多変量)の統計学
基礎となる３つの勘所：回帰分析、判別分析、主成分分析

副次的に学べること

線形代数
線形写像の像と核, 列空間と行空間, 転置行列の線形写像としての意味, ベクトル空間の直交直和分解, 射影行列, 2次形式, 固有値・固有ベクトル, 直交行列による対角化, etc.
確率統計
推定・検定, 偏回帰係数の有意性検定, 重相関と決定係数, 数量化I類-III類, etc.

この内容は以下の基礎になっています

統計学(推測統計学/数理統計学)、多変量解析、統計的信号処理、パターン認識、機械学習、統計的データ分析、データサイエンス、データマイニング、人工知能

多変量解析の主題

回帰分析 ⇒ 4,5章 ------ **数量化I類** ⇒ 6章

判別分析 ⇒ 7章 ------ **数量化II類** ⇒ 8章

主成分分析 ⇒ 9章 ------ **数量化III類** ⇒ 10章

多次元尺度構成法(12章)、クラスター分析(11章)、
因子分析・パス分析・グラフィカルモデル(13章)、
正準相関分析(13章)、多段層別分析(13章)

線形代数の技術

射影行列
線形写像の像と核
直交直和分解
(線形代数学の基本定理)

２次形式
基底変換
固有値・固有ベクトル
直交行列による対角化

多変量正規分布

標準化、マハラノビス距離

多変量データの具体例

体重 w および身長 h を n 人に調査

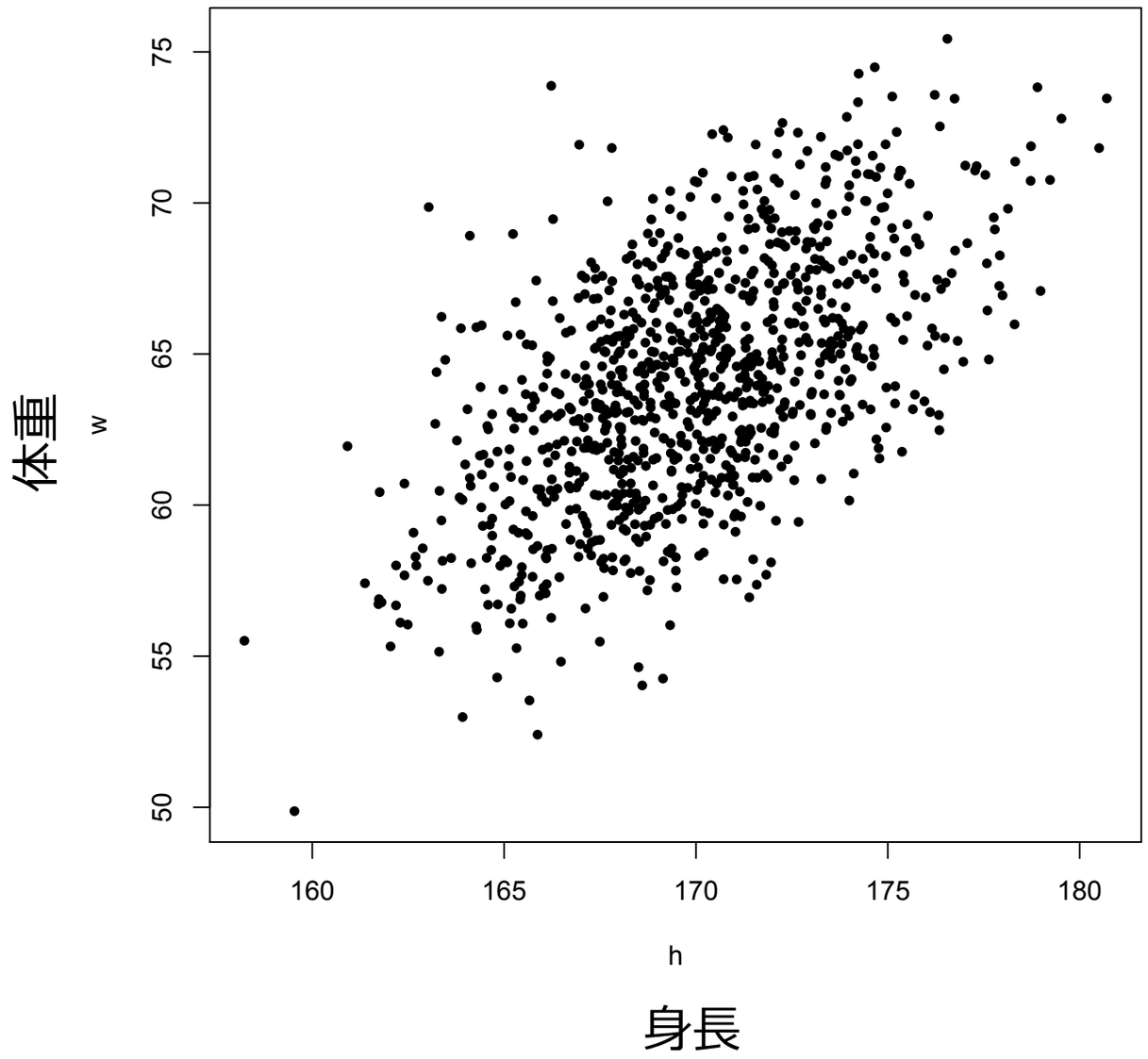
集まったデータ $\rightarrow$ $w_1, w_2, w_3, \ldots, w_n$
$h_1, h_2, h_3, \ldots, h_n$

これを各々個人ごとに (w, h) という対データを得ると見て、2変量のベクトル値データとみなす

$$\begin{pmatrix} w_1 \\ h_1 \end{pmatrix}, \begin{pmatrix} w_2 \\ h_2 \end{pmatrix}, \begin{pmatrix} w_3 \\ h_3 \end{pmatrix}, \ldots, \begin{pmatrix} w_n \\ h_n \end{pmatrix} \qquad \in \mathbb{R}^2$$

多変量のデータのイメージ (2変量の例)

散布図



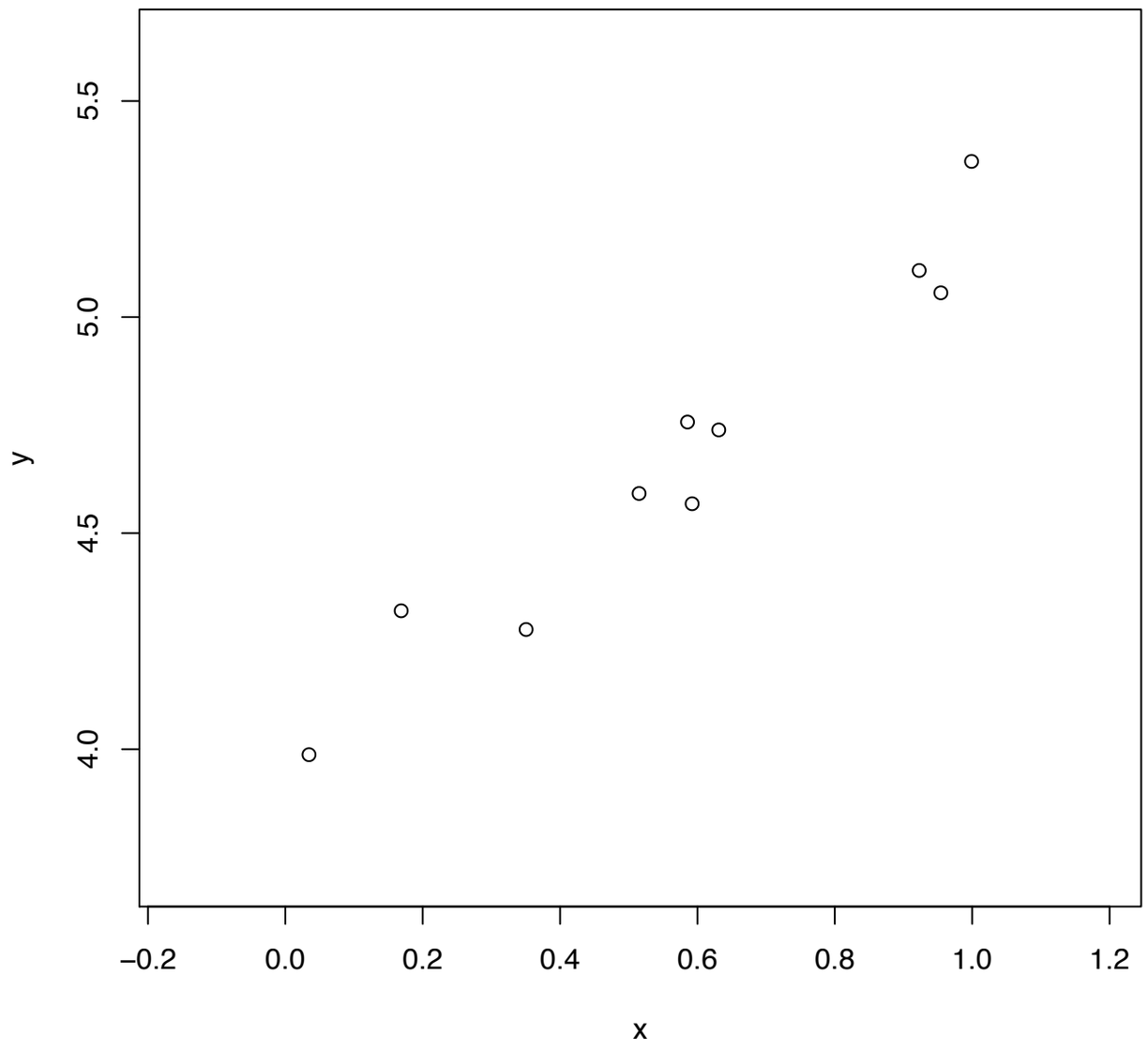
表形式データ

	身長 h	体重
1	174.0	64.1
2	169.6	65.4
3	168.4	67.4
4	171.7	63.4
5	172.1	72.3
6	167.0	63.4
7	167.0	62.5
:	:	:
999	172.7	64.9
1000	167.3	61.97

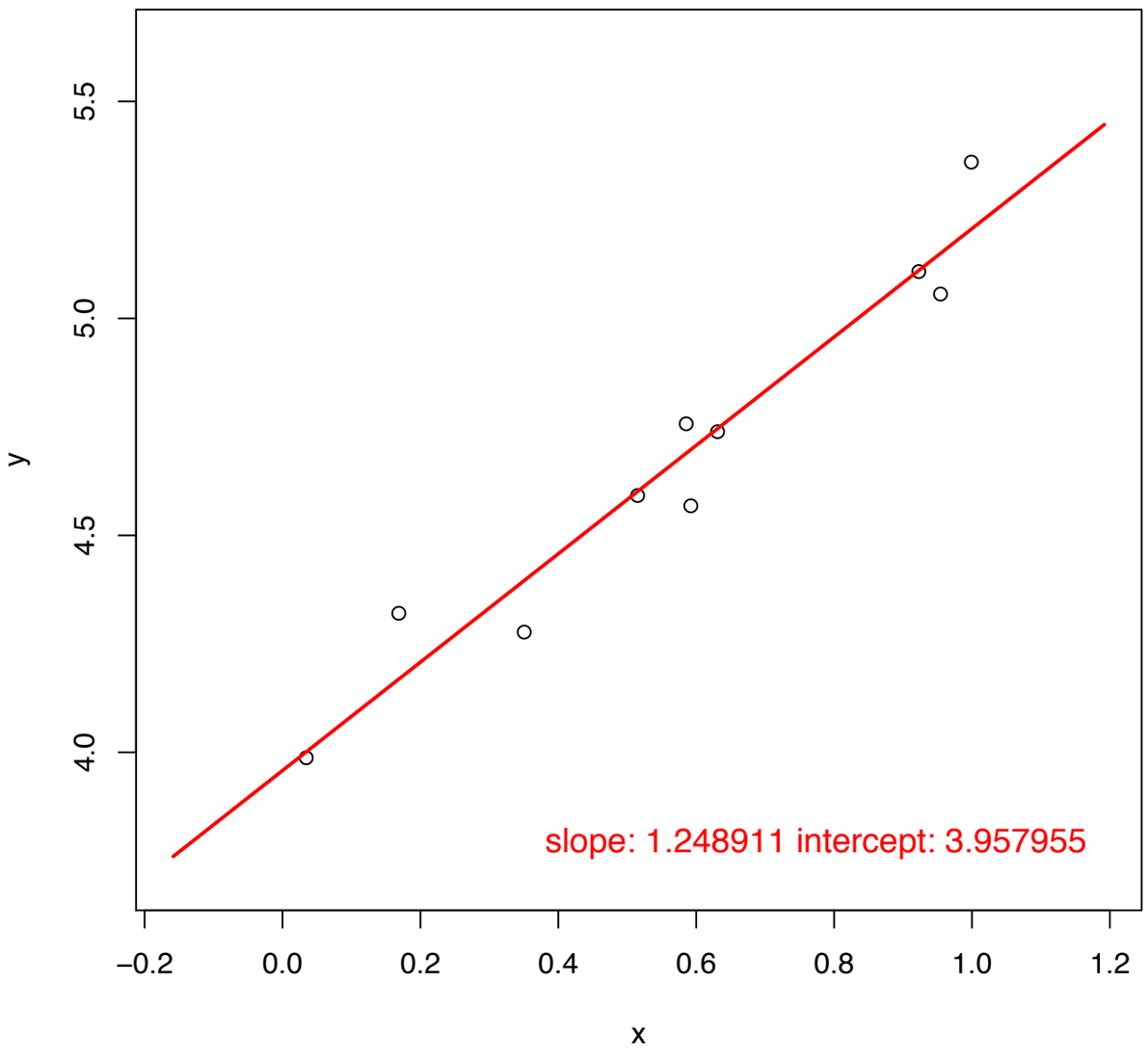
データ(表形式)

x	y
0.954347545048222	5.05622523205918
0.999102738685906	5.36031043090979
0.0344490522984415	3.98743160130578
0.35056076082401	4.27713727492932
0.585548606002703	4.7572216692572
0.631030546268448	4.73894391078217
0.592243858613074	4.56811837127514
0.168670440558344	4.32047600620958
0.515009876806289	4.59182198538531
0.922892721602693	5.10791569839468

データの散布図



回帰直線



予測に使う

x=0.85のとき、yの予測値は
いくらくらいだろうか?

ギモン

(1)データ点から傾きと切片を
どうやって計算するの？

(2)良し悪しをどう考えれば
良いのだろう？

単回帰分析 (最小二乗推定)

予測したい

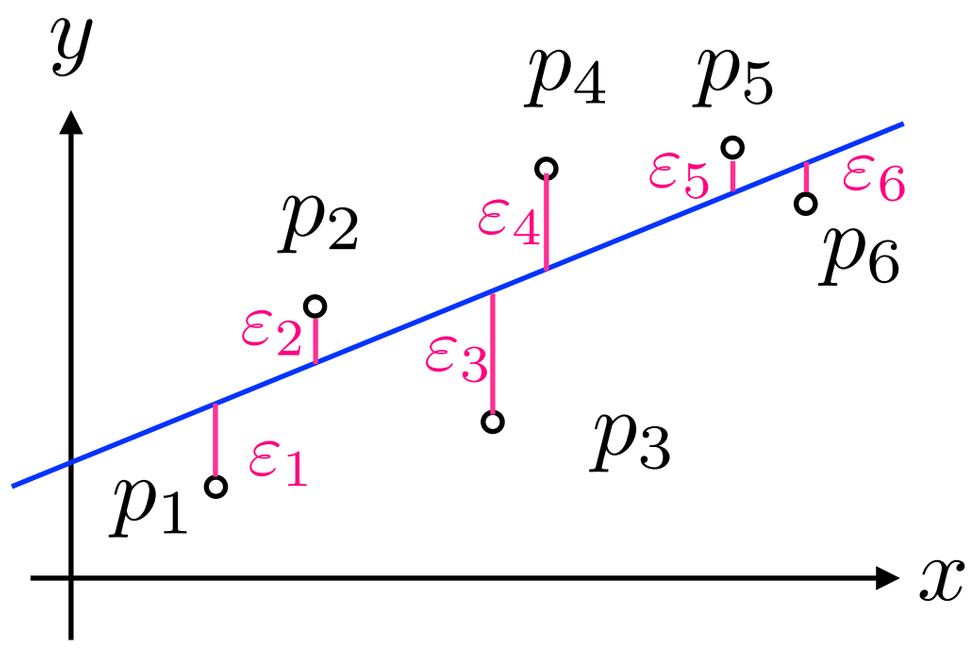
説明変数 目的変数

	x	y
p_1	x_1	y_1
p_2	x_2	y_2
p_3	x_3	y_3
p_4	x_4	y_4
p_5	x_5	y_5
p_6	x_6	y_6

二乗誤差を最小にする直線当てはめ

$$\varepsilon_i = y_i - (ax_i + b)$$

$$\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \varepsilon_4^2 + \varepsilon_5^2 + \varepsilon_6^2 \to \min$$



$$y = ax + b$$

ここをいじる

最小二乗推定の計算法(関数の極値問題に)

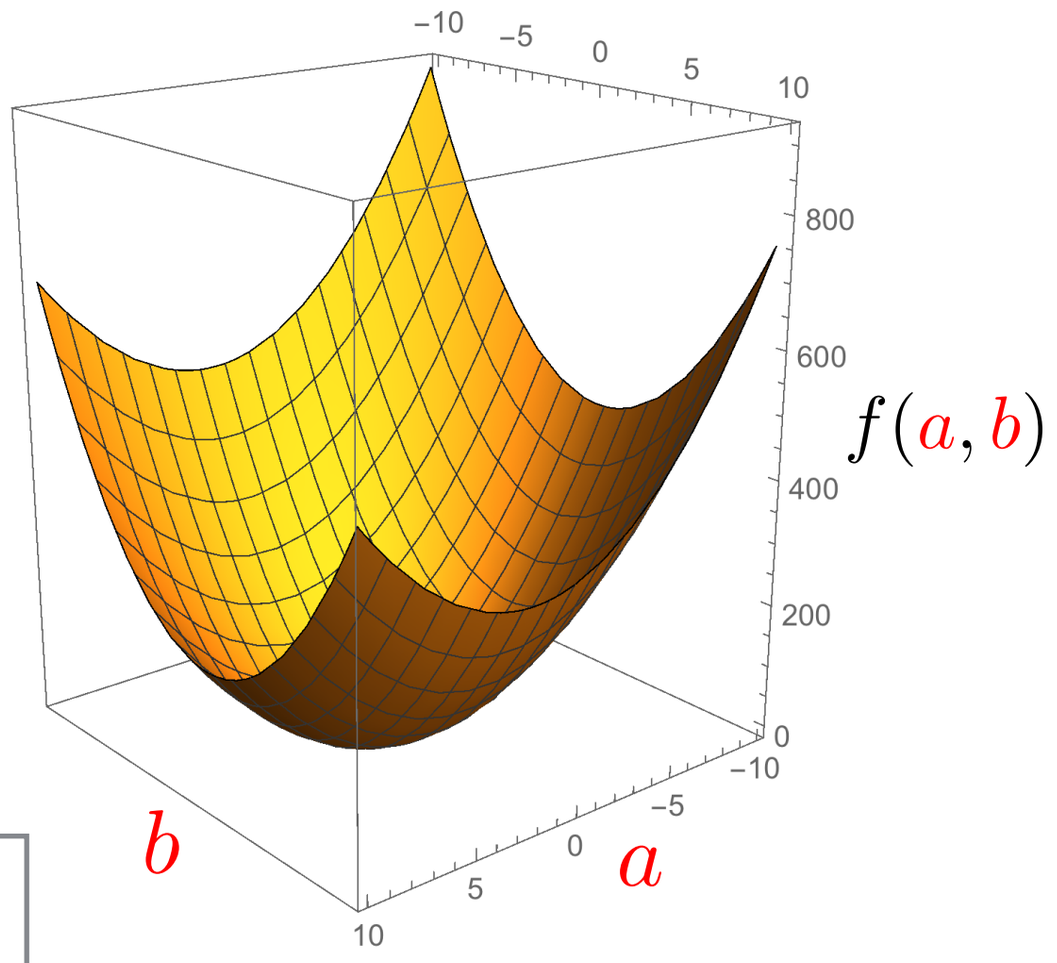
二乗誤差を最小にする直線当てはめ

$$\varepsilon_1^2+\varepsilon_2^2+\varepsilon_3^2+\varepsilon_4^2+\varepsilon_5^2+\varepsilon_6^2 \to \min$$

$$\varepsilon = y_i - (ax_i + b)$$

$$f(a,b) := \sum_{i=1}^{6} \varepsilon^2 = \sum_{i=1}^{6} (y_i - (ax_i + b))^2$$

展開すればとにかく a と b の2次関数に！



項の数が多いが、原理上は
がんばれば高校数学で解ける！

大学の微積分を使うなら
偏微分＝0の連立方程式を解く。

有名問題	2変数2次関数の最大・最小	サクシード 数学Ⅰ　$p.50$ 重要例題78 改題 参考：チャートW 数学Ⅰ　$p.119$ 重要例題78

x, yが互いに関係なく変化するとき、
$P=x^2-4xy+5y^2-6y+10$ の最小値と、そのときのx, yの値を求めよ。

《解答》

$$P=x^2-4y\cdot x+5y^2-6y+10=(x-2y)^2-(2y)^2+5y^2-6y+10$$
$$=(x-2y)^2+y^2-6y+10=(x-2y)^2+(y-3)^2-3^2+10$$
$$=(x-2y)^2+(y-3)^2+1$$

よって、Pの最小値は1
そのときのx, yの値は、$x-2y=y-3=0$ より、$x=6$, $y=3$

《別法》

$$P=5y^2-2(2x+3)\,y+x^2+10=5\left\{y^2-\frac{2(2x+3)}{5}y\right\}+x^2+10$$
$$=5\left\{\left(y-\frac{2x+3}{5}\right)^2-\left(\frac{2x+3}{5}\right)^2\right\}+x^2+10=5\left(y-\frac{2x+3}{5}\right)^2+\frac{x^2-12x+41}{5}$$
$$=5\left(y-\frac{2x+3}{5}\right)^2+\frac{(x-6)^2}{5}+1$$

よって、Pの最小値は1
そのときのx, yの値は、$y-\dfrac{2x+3}{5}=x-6=0$ より、$x=6$, $y=3$

偏微分で解く場合

$$\frac{\partial P}{\partial y}=-6-4x+10y=0$$

$$\frac{\partial P}{\partial x}=2x-4y=0$$