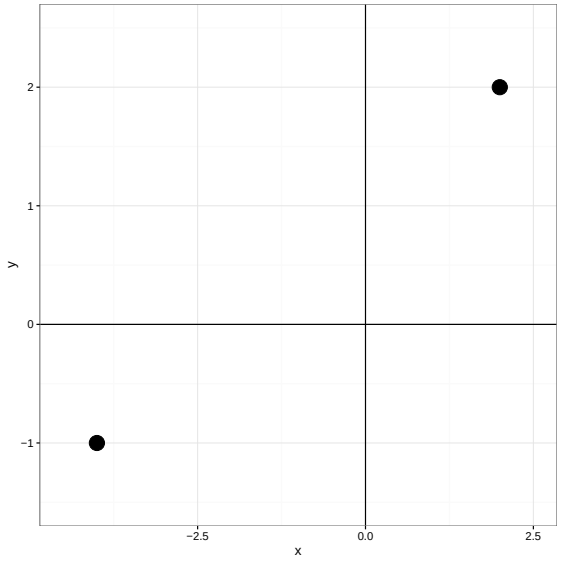
高校生のための現代数学講座　　東京大学
「確率と統計」　　玉原国際セミナーハウス
講義 (6)　鎌谷　研吾　　2016 年 7 月 23 日

「君は現代の預言者になれるか」

1　回帰分析



座標平面上に上の点 $(x_1, y_1), (x_2, y_2)$ が置かれている．ただし $x_1 \neq x_2$ としよう．二点を通る直線

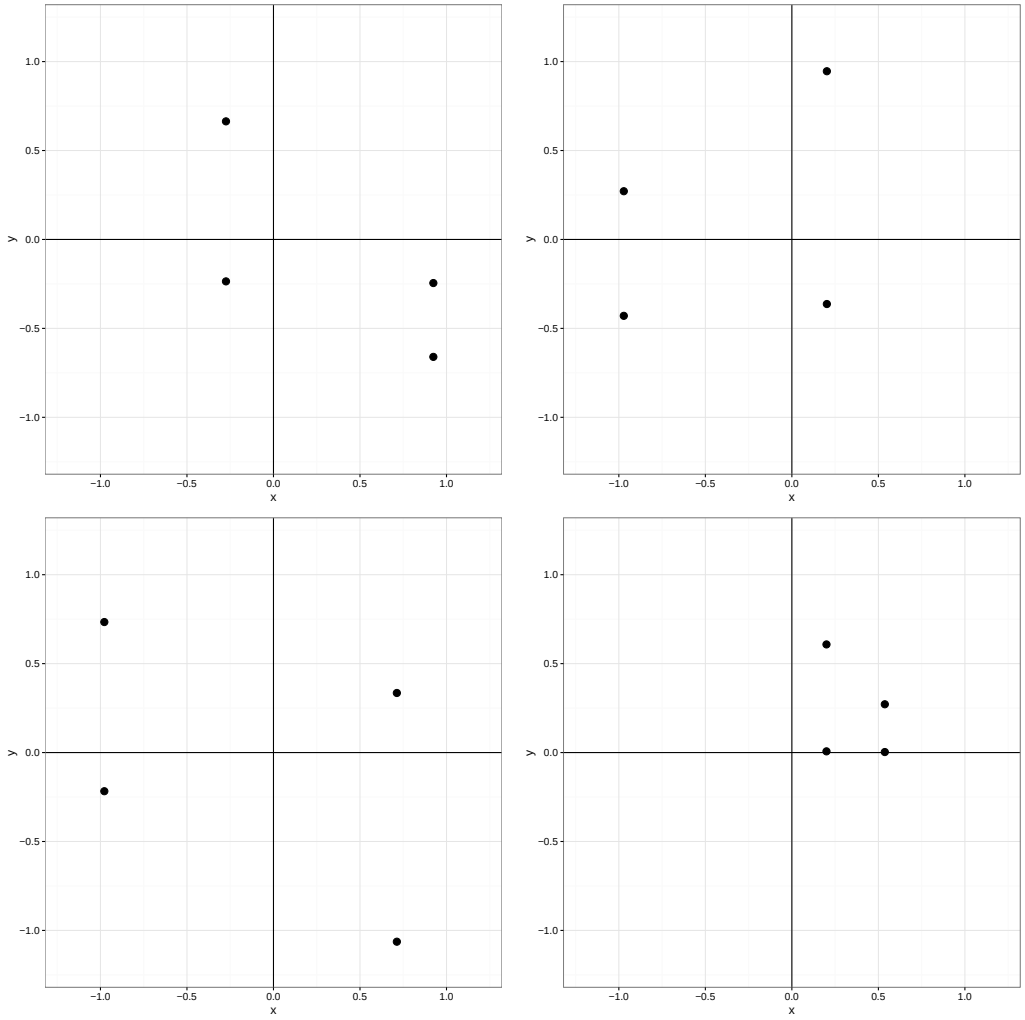
$$y = ax + b$$

を考えよう．この場合連立方程式

$$y_1 = ax_1 + b$$
$$y_2 = ax_2 + b$$

を解くことで，$a = (y_2 - y_1)/(x_2 - x_1)$ および $b = (x_1y_2 - x_2y_1)/(x_2 - x_1)$ が得られる．ここで $(x_1, y_1), (x_2, y_2)$ のことをデータ，定数 a, b をパラメータと呼ぶことにしよう．例えば二点の座標が $(-4, -1)$ および $(2, 2)$ のときは $a = 1/2, b = 1$ であるから $y = x/2 + 1$ が二点を通る直線である．

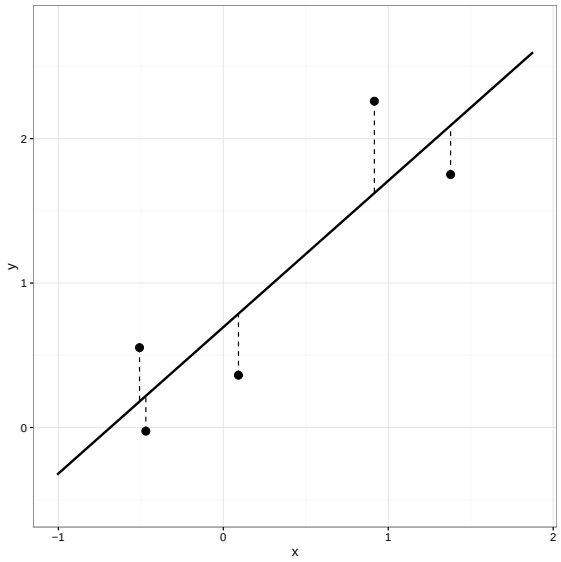
さて，データ（点）の数が増えて四点になったらどうだろうか．もはや全ての点を通る直線は存在しない．代わりに，なるべく四点の近くを通るような直線を引きたい．どのような直線が良いだろうか．データの組をいくつか例示するので，どのような直線が良いか考えて，実際に直線を書いてみよう．黒いペンかえんぴつを推奨．



直線の良し悪しに基準がないので，一つの例として以下の基準を置こう．n 個の点 $(x_1, y_1), \ldots, (x_n, y_n)$ に対し，直線 $y = ax + b$ の当てはまり具合の尺度を

$$\sum_{i=1}^{n} |y_i - ax_i - b|^2$$

で定義しよう．この尺度が小さいほどよい直線であると解釈することにする．この基準は一見わかりにくいので，次の図を見てみよう．



直線 $y = ax + b$ に対し，各点毎に，その直線との隔たりを図示した．この隔たりが $|y_i - ax_i - b|$ であることに気がつくと，先ほどの当てはまり具合の尺度は，各点の隔たりの二乗を足したものである．この事実を用いれば，先ほどの４つの図に，当てはまり具合のもっともよい直線を引くことは容易である．赤ペンで直線を引いてみよう．そして当初書いた直線との隔たり具合を見てみよう．

なお，当てはまり具合の尺度を最小にする a, b は具体的に解くことができて，

$$a = \overline{y} - b\overline{x},\ b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

となる．ただし，$\overline{x} = \sum_{i=1}^{n} x_i/n, \overline{y} = \sum_{i=1}^{n} y_i/n$. 得られた直線を回帰直線という．

2　予測対決

実際にデータを取って回帰直線を描いてみよう．グループに分かれてデータを集計し，直線を引いて予測を行う．予測を行う対象は講義中に示す．データの集計についてはチーム内で相談すること．

予測の方法を明示しておこう．例えば A さんの右手のひらの大きさがわかっている時，右腕の長さを当てたいとする．チーム内のメンバーの右手のひらの大きさを x 軸に，右腕の長さを y 軸に取り，座標平面に書き込む．そして先程の回帰直線の書き方を参考にして，最も当てはまりの良さそうな直線を引く．y 軸に平行に，x 座標が A さんの手のひらの大きさとなる直線を引く．回帰直線との交点の y 座標の値が，A さんの右腕の長さの予測値になる．

3　モデル

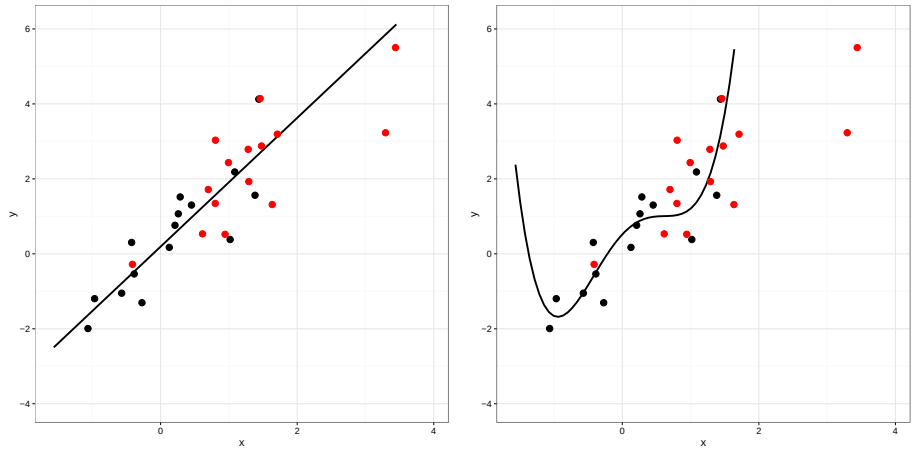
各チームに分かれて予測を行ったが，集め得るすべてのデータを使って予測を行うとどうなるだろうか．どんどん情報が増えて予測はより正確になりそうだ．しかし，情報をうまく扱わないと，却って予測を悪くする．問題を解きほぐして，より単純な問題を考えてみよう．先程の回帰直線は実際には

$$y = ax + b + \epsilon$$

なる，誤差 ϵ を許すモデルを考えていた．ここで ϵ は平均的には 0 であるようなノイズである．一方で，たくさんの情報を利用できるモデルとして，

$$y = a_1 x + a_2 x^2 + \cdots + a_k x^k + b + \epsilon$$

を用意しよう．データ y を説明するのに x だけでなく，$x^2, \ldots, x^k$ も使える．明らかに前者のほうが予測力に優れているように感じるが，果たしてそうだろうか．この議論は統計的仮説検定や赤池情報量基準に結びつくが，ここでは感覚的に，大きすぎるモデルの弊害を説明する．



左の図は黒い点で示されるデータを使って，モデル $y = ax + b + \epsilon$ から得られる回帰直線，左の図はモデル $y = a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + b + \epsilon$ から得られる，回帰直線に対応する回帰曲線である．予測したい，新しいデータが赤い点で示されている．観測はモデル $y = ax + b + \epsilon$ から生成しているので，単純なモデルが予測の点で優れているのは自然だが，一方で複雑なモデルは単純なモデルを含んでいる，より柔軟性の高いものである．予測力に大きな差がでたが，その背景の理論について少しだけ講義中に解説する．