

MAT220 Lecture 19: Linear Regression (線形回帰)

Contents

1	線形回帰	2
1.1	相関係数	2
1.2	例：距離と通学時間	2
2	最小二乗法	3
2.1	アルゴリズムとしての回帰	3
2.2	最小二乗基準	4
2.3	回帰直線の公式	6
2.4	影響点	7
2.5	例：回帰直線の計算	8
3	回帰直線を用いた予測	8
3.1	予測を行う	8
3.2	内挿と外挿	9
3.3	決定係数	10
3.4	決定係数の別の形	12
3.5	r^2 としての決定係数	12
3.6	例：通学データの決定係数	13

第18回の講義では、観測値 (x_i, y_i) を散布図上にプロットすることで、対応のある量的データについて学びました。次に、標本相関係数 r を用いて、線形関係の符号と強さを表しました。相関は、点が直線におおよそ当てはまるかどうかを示すだけであり、実際にその直線を作るものではありません。今日は、最小二乗回帰直線と呼ばれるこの直線を作ります。まず回帰直線の重要な特徴を確認し、その後、回帰直線を用いて予測を行います。最後に、予測モデルとしての回帰直線の良さを測る決定係数を導入します。

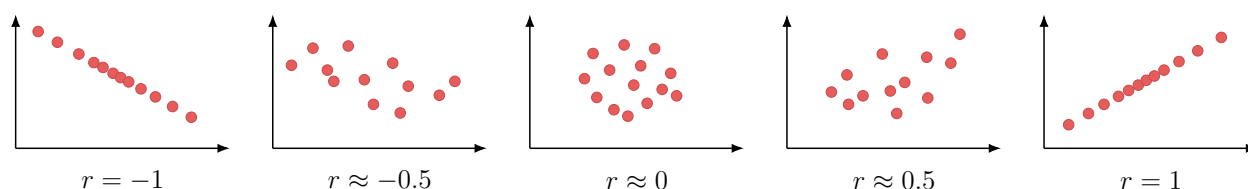
1 線形回帰

1.1 相関係数

n 組の観測値があるとします：

$$(x_1, y_1), \quad (x_2, y_2), \quad \dots, \quad (x_n, y_n).$$

前回の講義では、対応のあるデータの記述統計について学び始めました。特に、対応のあるデータがどの程度直線に沿っているかを表す線形相関について考えました。その過程で、 x と y の間の線形相関を数値化する相関係数を導入しました。この数を r と表し、その値は -1 から 1 の範囲をとります。 r の値によって、線形相関の強さと符号が決まります：



代数的には、標本相関係数は次の公式で計算できます：

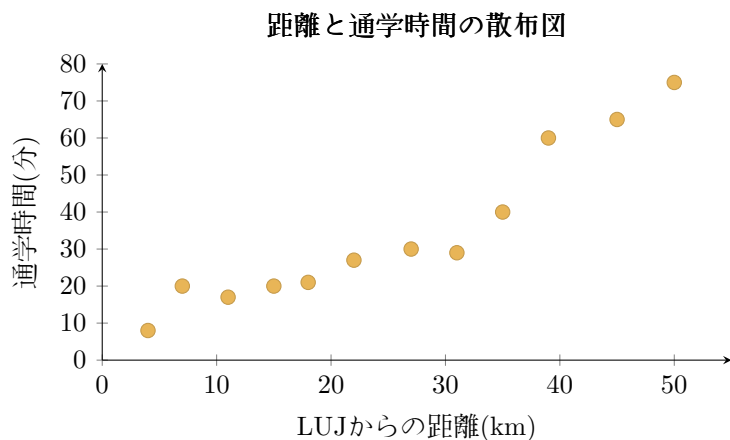
$$r = \frac{\text{cov}(x, y)}{s_x s_y},$$

ここで、分子は共分散であり、2つの変数が互いにどのように関係して変化するかを表します。また、 s_x と s_y は、それぞれ x 方向と y 方向の標本標準偏差です。

1.2 例：距離と通学時間

この講義では、次のデータを継続して例として使用します。12人の学生が、LUJから自宅までの距離と、キャンパスまでの通学時間を記録したとします。

学生	距離 (km)	時間 (分)
1	4	8
2	7	20
3	11	17
4	15	20
5	18	21
6	22	27
7	27	30
8	31	29
9	35	40
10	39	60
11	45	65
12	50	75



このデータについて、前回の講義で次の値を求めました：

$$\bar{x} \approx 25.333, \quad \bar{y} \approx 34.333, \quad s_x \approx 15.011, \quad s_y \approx 21.253, \quad \text{および} \quad \text{cov}(x, y) \approx 301.606.$$

したがって、

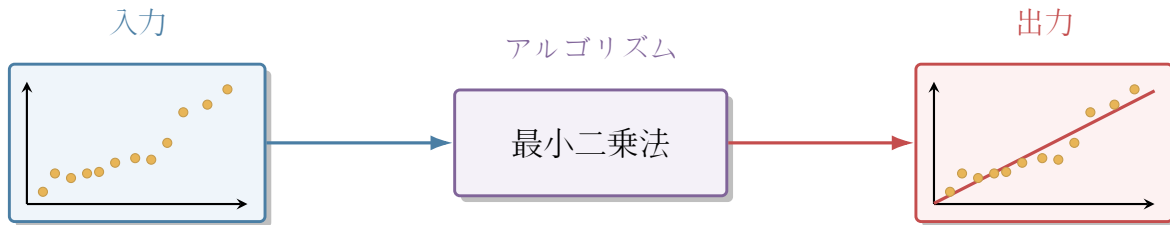
$$r = \frac{301.606}{(15.011)(21.253)} \approx 0.9454,$$

となり、2つの変数の間に強い正の線形関係があることが分かります。

2 最小二乗法

2.1 アルゴリズムとしての回帰

アルゴリズムとは、入力を受け取り、決まった手順に従って確実に出力を生成する方法です。今日の講義では、対応のあるデータを入力として受け取り、「最適直線」を出力するアルゴリズムを紹介します。後で明らかになる理由から、この方法を最小二乗法と呼びます。図式的には、次のようになります：



垂直でない直線は、2つの数によって完全に表すことができることを思い出してください：

- 傾き：直線の変化率、つまり直線の急さを表します。
- y 切片： $x = 0$ のときの直線の値です。

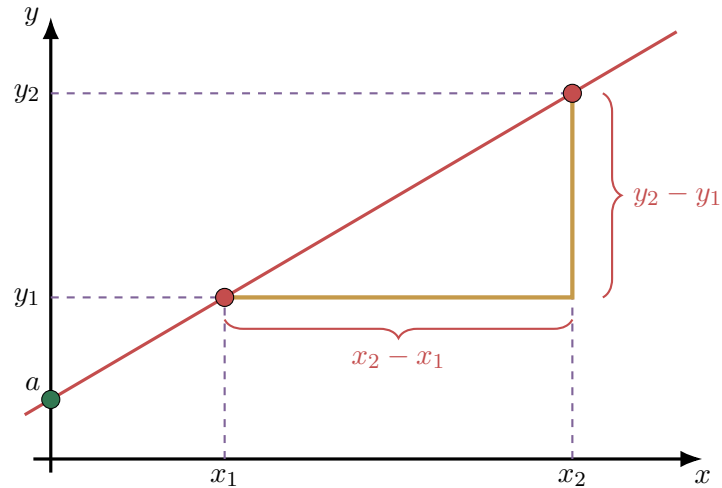
直線を次のように表します：

$$\hat{y} = a + bx,$$

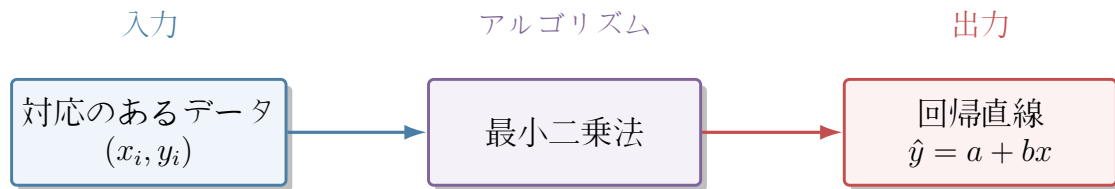
y に付いているハット記号の意味は後で明らかになります。ここで、 a は y 切片の値であり、 b は直線の傾きです。傾きは、直線上の2点を選び、次の公式を用いて計算できます：

$$b = \frac{y_2 - y_1}{x_2 - x_1},$$

つまり、垂直方向の変化を水平方向の変化で割ります。

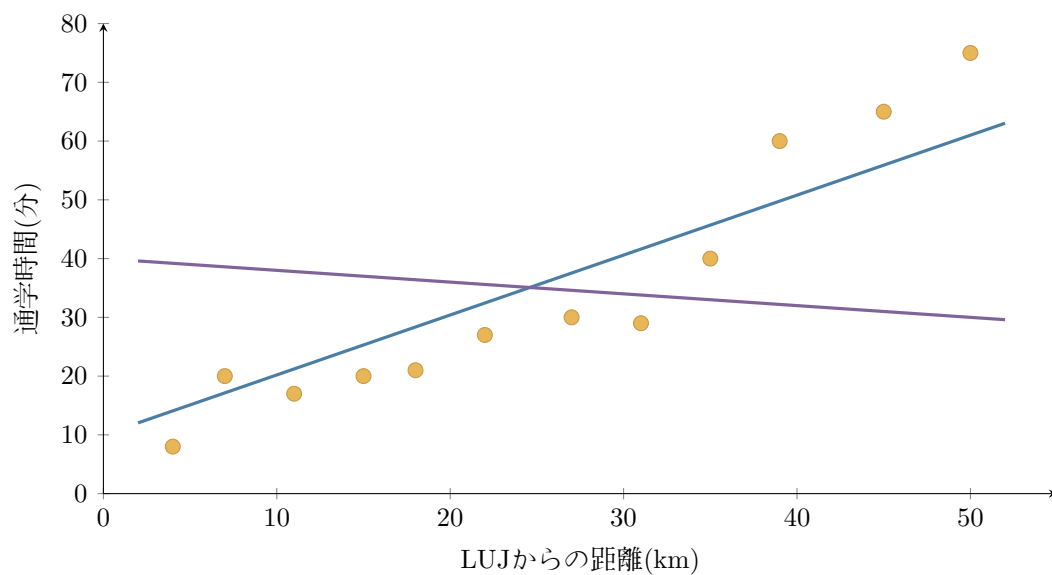


この記法を用いると、最小二乗法の出力は、入力データに最もよく当てはまる直線を作る2つの数 a と b であることが分かります。このような直線を回帰直線と呼びます。したがって、アルゴリズムをより詳しく表すと次のようになります：



2.2 最小二乗基準

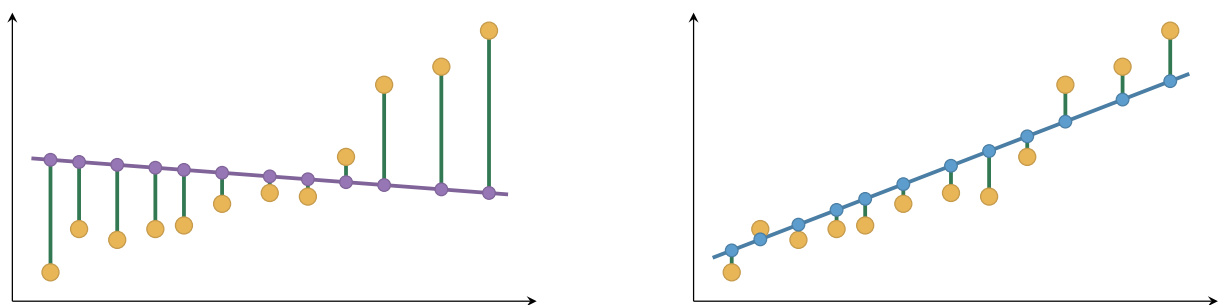
第1節のデータを用いて、2本の直線を考えてみましょう：



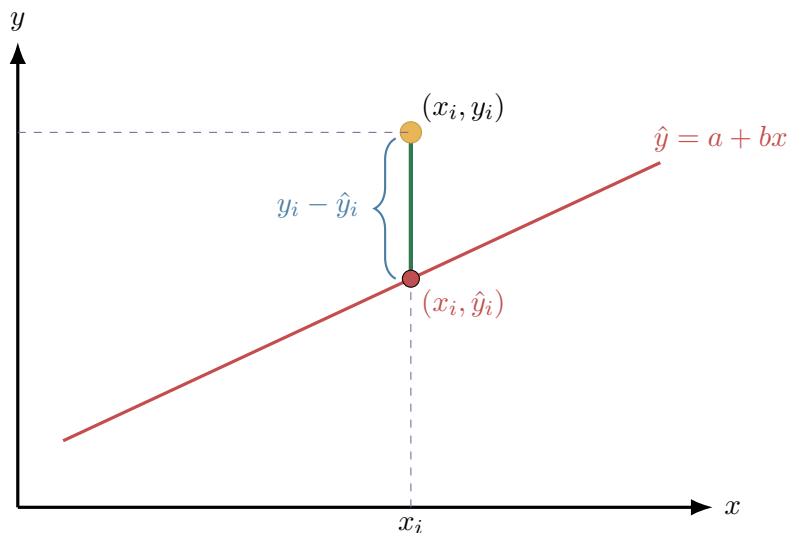
明らかに、青い直線は紫の直線よりもデータによく当てはまっています。しかし、数学的には、なぜそう言えるのでしょうか。言い換えると、青い直線の何か紫の直線より優れているのでしょうか。この問いに答える一つの方法は、データと回帰直線の間の垂直距離を調べることです。各データ点は (x_i, y_i) の形で表されることを思い出してください。したがって、各点から直線までの垂直距離を表すために、回帰直線 $\hat{y}$ の式に x_i を代入し、直線上の出力を求めます。 x_i に対応する回帰直線上の値は次の数です：

$$\hat{y}_i = a + bx_i.$$

x_i と $\hat{y}_i$ を組み合わせると、直線上の点 $(x_i, \hat{y}_i)$ が得られます。この点は x_i の真上にあります。理想的には、この点は観測されたデータ点 (x_i, y_i) に近いはずですが、図から分かるように、紫の直線の場合、実際のデータまでの垂直距離は、青い直線の場合よりはるかに大きくなっています：



この意味で、青い直線の方がよりよく当てはまっていると直感的に考えます。青い直線に対するこれらの垂直線の長さの合計は、紫の直線に対する合計より小さいからです。一般に、これらの符号付きの差を残差と呼び、単に差 $y_i - \hat{y}_i$ として表されます：



値 y_i が直線より上または下にある場合、 $y_i - \hat{y}_i$ はそれぞれ正または負になります。したがって、ある種の「総誤差」を作るためにこれらの残差を単純にすべて足すと、互いに打ち消し合

い、「当てはまりの良さ」を表す尺度が損なわれる可能性があります。この問題を解決するために、 $y_i - \hat{y}_i$ をすべて二乗してから足し合わせます。¹ したがって、「総誤差」の適切な尺度は、実際には総平方誤差です：

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2.$$

a と b の値が異なれば、異なる直線が得られ、この和の値も異なります。上の例では、紫の直線の総平方誤差は、青い直線の総平方誤差よりはるかに大きくなっています。このことから、「最適直線」を、この総平方誤差を最小にする直線として定義できます。

最小二乗基準

最小二乗回帰直線 $\hat{y} = a + bx$ は、

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

が、すべての直線の中で可能な限り小さくなるように選ばれます。

2.3 回帰直線の公式

a と b の値を正確に導出するには、微積分を使用する必要があります。この授業は入門科目であるため、ここでは回帰直線の導出には時間を使いません。² ここでは、 a と b の公式が次のようになることを受け入れます：

$$b = \frac{\text{cov}(x, y)}{s_x^2} \quad \text{および} \quad a = \bar{y} - b\bar{x}.$$

これら2つの公式を用いることで、回帰直線 $\hat{y} = a + bx$ を書くことができます。

回帰直線の計算

対応のあるデータ (x_i, y_i) が与えられたとき、

$$b = \frac{\text{cov}(x, y)}{s_x^2} \quad \text{および} \quad a = \bar{y} - b\bar{x}$$

を計算します。その後、最小二乗回帰直線を

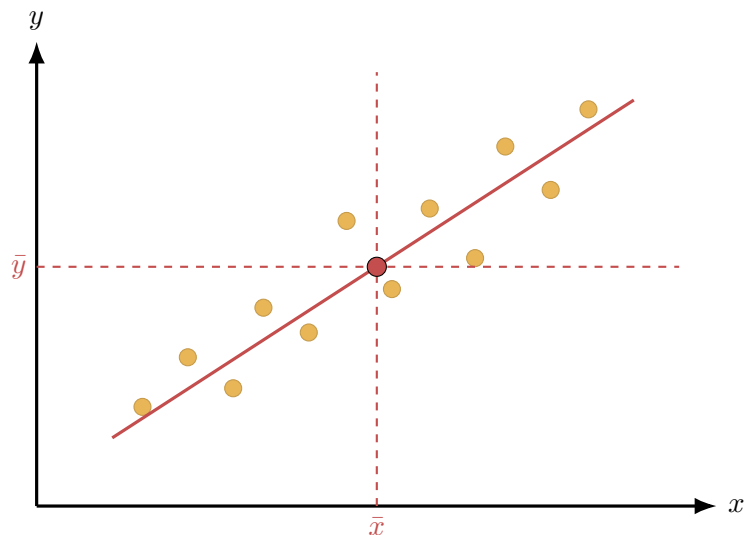
$$\hat{y} = a + bx$$

と書きます。

小さな系として、点 $(\bar{x}, \bar{y})$ は常に最小二乗回帰直線上にあることが分かります：

¹これは、標本分散 s_y^2 の公式で $y_i - \bar{y}$ を二乗する理由と同じです。

²興味のある人のために説明すると、 $\sum_{i=1}^n (y_i - (a + bx_i))^2$ を a と b について微分できます。最小値を求めるために、これらの導関数を0に等しいとおき、得られる2本の一次方程式を解きます。



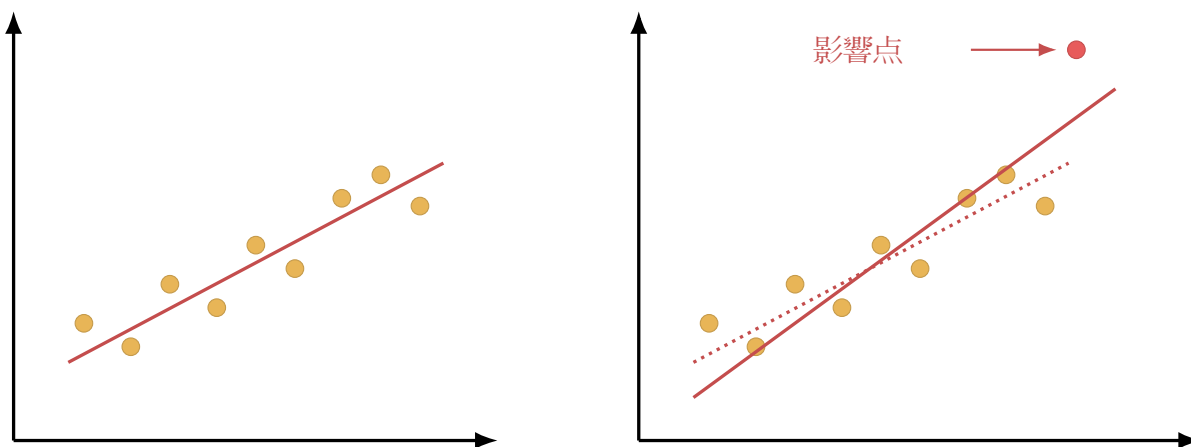
これを確認するには、式に $x = \bar{x}$ を代入します：

$$\hat{y} = a + b\bar{x} = (\bar{y} - b\bar{x}) + b\bar{x} = \bar{y}.$$

どのデータから始めても結果は変わりません。これは単に回帰直線の構成から得られる性質です。

2.4 影響点

ある点がデータの主要な集まりから遠く、特に水平方向に離れている場合、回帰直線に大きな影響を与えることがあります。このような観測値を影響点と呼びます。



影響点が存在すること自体は、必ずしも問題ではありません。しかし、影響点は回帰直線を大きく変化させることがあるため、回帰直線をモデルとして用いる予測に影響を与える可能性があります。したがって、このような点を扱う場合には常に注意が必要です。

2.5 例：回帰直線の計算

第1節の通学データについて、最小二乗回帰直線を計算します。第18回の講義から、すでに次の値が分かっています：

$$\bar{x} \approx 25.333, \quad \bar{y} \approx 34.333, \quad s_x \approx 15.011, \quad \text{cov}(x, y) \approx 301.606.$$

まず、傾き b を計算します：

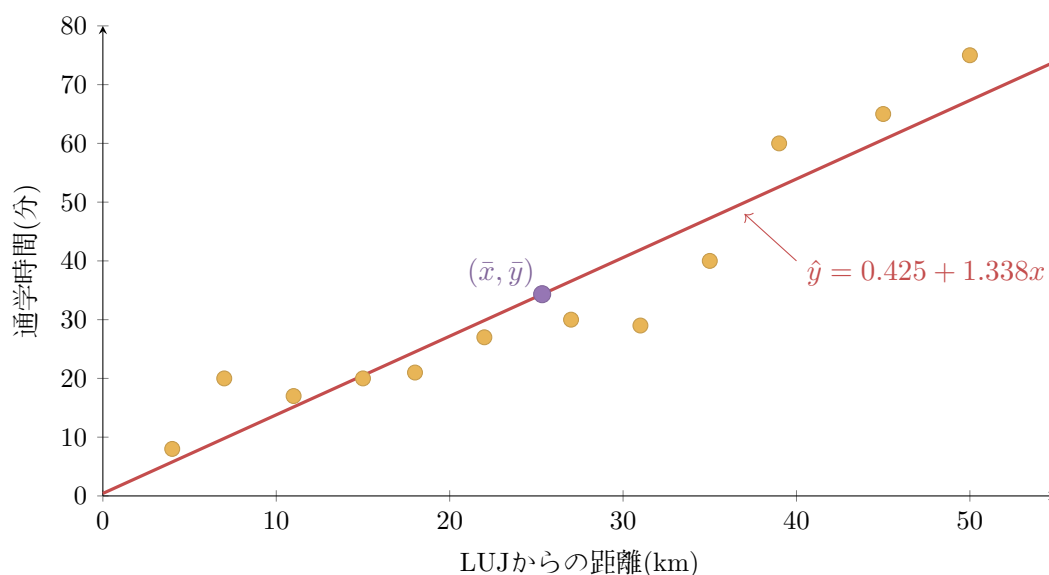
$$b = \frac{\text{cov}(x, y)}{s_x^2} = \frac{301.606}{(15.011)^2} \approx 1.33849.$$

次に、この値を用いて y 切片 a を計算します：

$$a = \bar{y} - b\bar{x} = 34.333 - (1.33849)(25.333) \approx 0.425.$$

これらを組み合わせると、回帰直線は次のようになります：

$$\hat{y} = 0.425 + 1.338x.$$



傾きの解釈

傾きは約1.338 です。当てはめられた線形モデルによると、LUJからの距離が1キロメートル増えるごとに、予測通学時間は約1.34 分増加します。

切片は約0.425 分です。この例では、切片は直線を正しい位置に置くために必要となる理論上の値であり、必ずしも現実的な意味を与えるべきではありません。

3 回帰直線を用いた予測

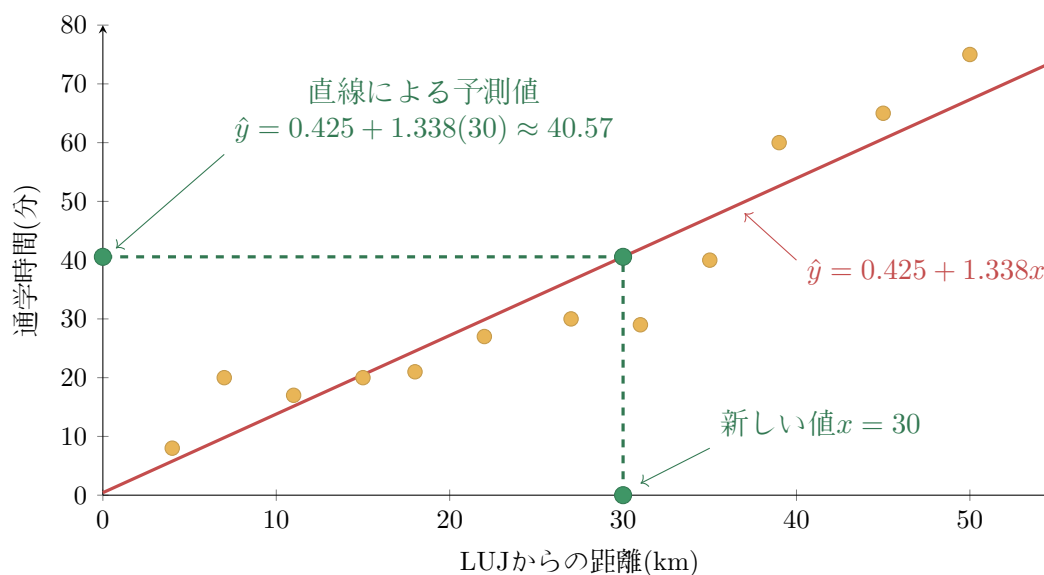
3.1 予測を行う

視覚的な要約に加えて、回帰直線の重要な利点の一つは、予測を行うためのモデルとして使用できることです。通学時間と移動距離の例を用いて、学生が引っ越し、新しいアパートがキャ

ンパスから30 km 離れているとします。この場合、回帰直線の式に $x = 30$ を代入することで、おおよその通学時間を予測できます：

$$\hat{y} = 0.425 + 1.338(30) \approx 40.57.$$

したがって、このモデルは通学時間を約41 分と予測します。図では、この予測は次のように表されます：

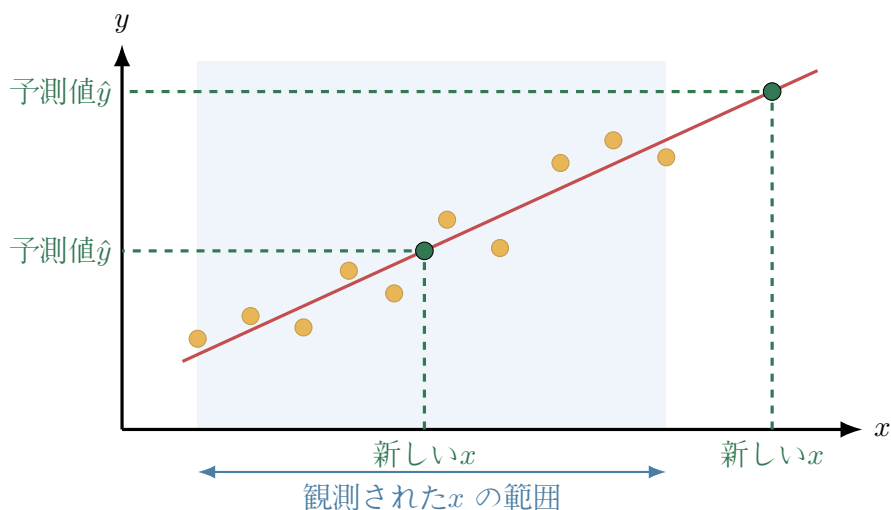


3.2 内挿と外挿

予測には、内挿と外挿と呼ばれる、根本的に異なる2つの種類があります。

内挿と外挿

内挿とは、観測された x の範囲内にある新しい値に基づいて予測を行うことです。一方、**外挿**とは、観測された x の範囲外にある新しい値に基づいて予測を行うことです。



一般に、内挿は回帰直線を使用する際のより安全な方法です。これは、実際にデータが観測された領域でモデルを使用するためです。一方、外挿では、適切な範囲の外側でモデルを使用する危険があります。 x が極端な値をとる場合には、データに影響を与える他の変数が存在する可能性があります、それらは回帰直線だけでは考慮されません。

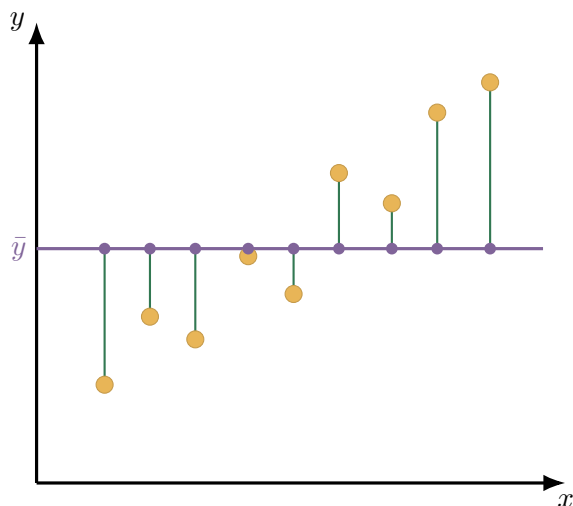
外挿に関する注意

ある x の範囲で観測された線形パターンが、その範囲外でも続くとは限りません。したがって、元の相関が強い場合であっても、外挿による予測は信頼できないことがあります。

3.3 決定係数

回帰直線は散布図を要約し、新しい x の値に基づいて y の値を予測することを可能にします。ここでは、この当てはまりがどの程度よいかを数値化する数を導入します。そのために、まず基準となる予測方法を導入します。

x のデータを完全に無視したとします。この場合、任意の y の値に対する最も一般的な予測は、標本平均 $\bar{y}$ を使用することです。もちろん、これは多くの場合に正確ではありませんが、平均的には最もよい予測になります。すべての y の値を平均 $\bar{y}$ で予測することは、水平線 $y = \bar{y}$ を選ぶことに相当します：

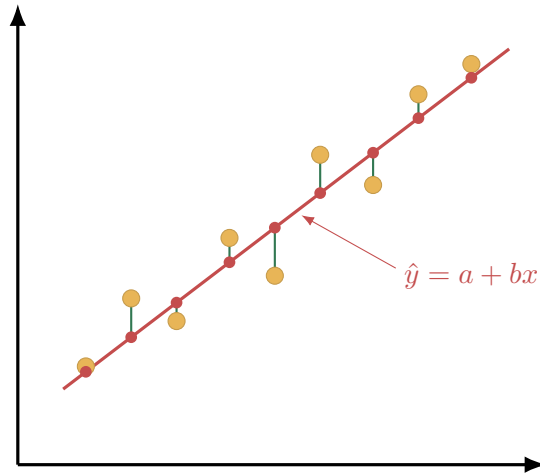


この予測方法の良さは、直線と観測されたデータ点との間に生じる誤差の合計を調べることで測ることができます。ここでも、誤差が打ち消し合うことを避けるために、まず値を二乗します。単純な予測によって生じる総平方誤差は次のようになります：

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

この量を総平方誤差と呼び、基準として使用します。

次に、回帰直線の誤差と単純な予測の誤差を比較することで、回帰直線の予測の良さを測ります。すべての値を単に $\bar{y}$ と予測する代わりに、入力値 x_i と回帰直線を用いるとします。この場合にも、いくつかの誤差が生じます：



回帰直線を予測に使用したときの総平方誤差は、次の量で表されます：

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

最小二乗基準の構成により、これは最小化される量です。したがって、常に

$$\sum_{i=1}^n (y_i - \bar{y})^2 \geq \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

が成り立ちます。これら2つの数の相対的な大きさを比較するには、その比をとることができます：

$$\frac{\text{回帰直線を用いた総平方誤差}}{\text{標本平均を用いた総平方誤差}} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

回帰直線がよい予測モデルであれば、標本平均を使用した場合の誤差と比べて、小さな誤差を生じるはずですが、上のは小さくなると考えられます。一方、回帰直線が予測モデルとしてよくない場合、単に平均を使用した場合と同程度の誤差を生じるため、上の比は1に近くなります。予測モデルがよいときに大きく、よくないときに小さくなる数が必要であれば、この比を1から引きます：

$$1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

これを決定係数と呼びます。

決定係数

決定係数は

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

です。これは、回帰直線によって生じる誤差と、標本平均を用いた単純な予測によって生じる誤差の比を1から引いた値です。定義される場合、その値は0から1の間にあり、値が大きいほど回帰モデルの当てはまりがよいことを示します。

3.4 決定係数の別の形

決定係数を変形することで、この数を別の方法で表すことができます。一つの方法は、式 $y_i - \bar{y}$ に $\hat{y}_i - \bar{y}$ を挿入することです：

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}).$$

回帰直線には、次の便利な性質があります：

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n \left((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \right)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

通常であれば、 $2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$ の形の交差項が現れるはずですが、回帰直線を構成する際に使用した微積分の結果として、これらを合計すると0になります。基本的な代数計算により、次のようになります：

$$\begin{aligned} 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \end{aligned}$$

分子と分母の両方を $n-1$ で割ると、次の式が得られます：

$$1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

この式の分母は y の分散、つまり対応のあるデータの垂直方向の広がりです。分子は、予測値から $\bar{y}$ までの二乗距離のある種の平均を計算しています。これも一種の分散ですが、今回は予測値 $\hat{y}_i$ の $\bar{y}$ の周りの広がりを計算しています。したがって、決定係数をこのように理解することができます。

決定係数の別の表現

決定係数は、同値な形として

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

と表すことができます。したがって、決定係数は、 y の分散のうち、回帰直線 $\hat{y}$ によって説明される割合を測るものと理解できます。

3.5 r^2 としての決定係数

決定係数の別の形をもう一度考えます。計算を分かりやすくするために、分子と分母に $\frac{1}{n-1}$ を挿入します：

$$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

したがって、分母は y の標本分散 s_y^2 になります。回帰直線が $\hat{y}_i = a + bx_i$ で与えられ、 $a = \bar{y} - b\bar{x}$ であるため、次のようになります：

$$\hat{y}_i - \bar{y} = a + bx_i - \bar{y} = (\bar{y} - b\bar{x}) + bx_i - \bar{y} = b(x_i - \bar{x}).$$

したがって、決定係数の分子は次のようになります：

$$\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \frac{1}{n-1} \sum_{i=1}^n (b(x_i - \bar{x}))^2 = b^2 \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = b^2 s_x^2.$$

これらを組み合わせると、次の式が得られます：

$$\frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = \frac{b^2 s_x^2}{s_y^2}.$$

b の定義を用いると、次のようになります：

$$b = \frac{\text{cov}(x, y)}{s_x^2} = \frac{\text{cov}(x, y)}{s_x^2} \cdot \frac{s_y^2}{s_y^2} = \frac{\text{cov}(x, y)}{s_x s_y} \cdot \frac{s_y}{s_x} = r \cdot \frac{s_y}{s_x}.$$

したがって、決定係数は次のように大きく簡略化されます：

$$\frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = b^2 s_x^2 \div s_y^2 = \left(r \frac{s_y}{s_x} \right)^2 s_x^2 = r^2 \frac{s_y^2}{s_x^2} s_x^2 \div s_y^2 = r^2 s_y^2 \div s_y^2 = r^2.$$

実際、説明変数が1つで切片を含む単回帰では、決定係数は標本相関係数の二乗に等しくなります：

$$r^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

これには、次のような直接的な結果があります：

- r を二乗すると、その符号は消えます。
- 強さが等しい正の相関と負の相関は、同じ r^2 を与えます。
- 強い線形相関は、大きな決定係数を与えます。
- 弱い線形相関は、小さな決定係数を与えます。

3.6 例：通学データの決定係数

通学データについて、以前に $r \approx 0.9454$ と計算しました。したがって、

$$r^2 \approx (0.9454)^2 \approx 0.894.$$

したがって、この標本では、観測された通学時間の変動の約89.4% が、LUJからの距離に対する線形回帰によって説明されます。説明されない変動は、交通手段、交通状況、乗り換え、徒歩時間、測定誤差など、通学時間に影響を与える他の変数を反映している可能性があります。