

統計学II

早稲田大学政治経済学術院

西郷 浩

本日の講義の目標

- (線形)回帰モデル
 - 回帰モデルの仕組み

回帰モデルの仕組み(1)

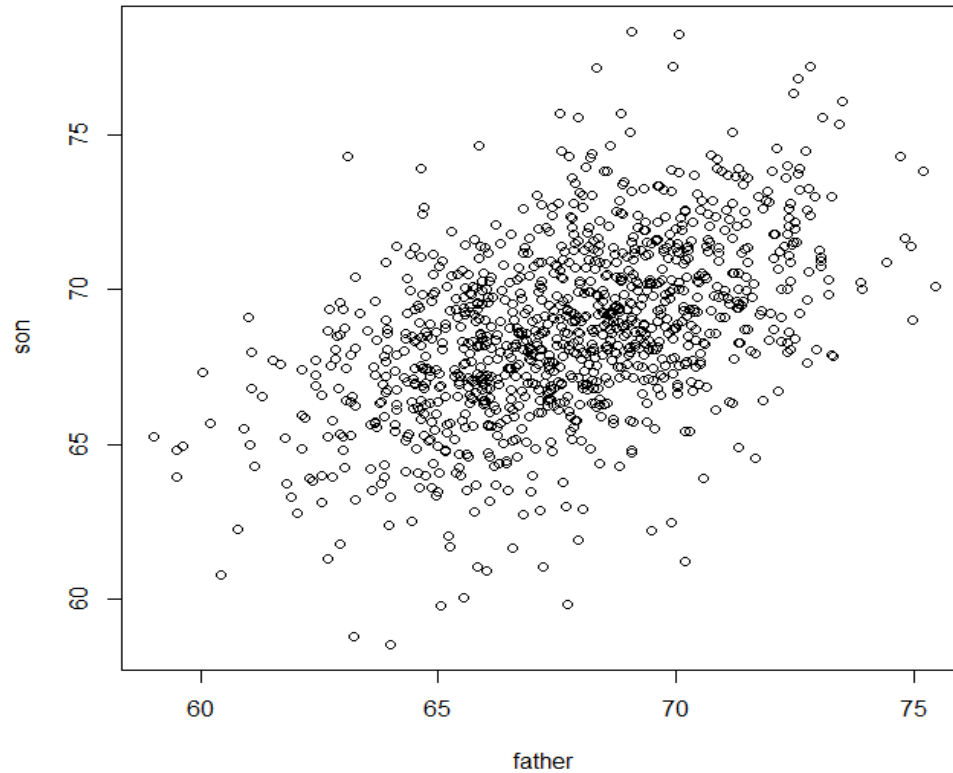


図1: 父親の身長
と息子の身長の
散布図

野口・西郷(2014)
『基本 統計学』
p. 196

回帰モデルの仕組み(2)

- 図1から読み取れる特徴(2つ)
 - 一般的な傾向
 - おおよそ右上がり
 - 父親の身長(x)が高ければ、息子の身長(Y)も高いという傾向がある。
 - » 原因: 遺伝の影響、その他
 - 個々の観察点にみられる現象
 - 縦軸方向のバラつき
 - 父親の身長(x)が同じでも、息子の身長(Y)が同じであるとはかぎらない。
 - » 原因: 父親の身長以外の要因
- これら2つの特徴を同時に表現できる、データ発生の仕組み(統計モデル)
 - 回帰モデル

回帰モデルの仕組み(3)

- (一般の)回帰モデル

- $Y_i = f(x_i) + u_i \quad (i = 1, 2, \dots, n)$

- Y_i : 被説明変数(従属変数、応答変数、結果変数、...)
 - x_i : 説明変数(独立変数、回帰子、原因変数、...)
 - u_i : 誤差項(攪乱項、攪乱項、[残差項]、...)

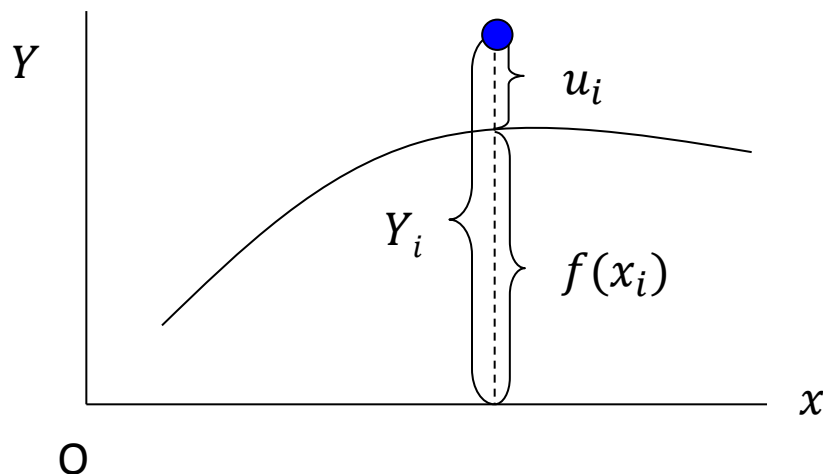


図2: 回帰モデル

回帰モデルの仕組み(4)

－ 誤差項の必要性

- 被説明変数 Y_i に影響を及ぼす要因
 - － 多数ある。
 - » 例：母親の身長、栄養状態、...
 - － すべてを取り入れて測定することは不可能である。
 - － 主要なものは説明要因に取り入れる。
 - － それ以外の種々雑多な説明要因を誤差項に一括する。
 - » 上記はひとつの説明である。他の説明もある。実際問題として、被説明変数を説明変数の関数(説明変数の値が定まれば、被説明変数の値がひとつに定まる)では表現できない。

回帰モデルの仕組み(5)

- (線形)回帰モデル
 - 回帰関数 $f(x_i)$ の選択
 - $E(Y_i|x_i) = f(x_i)$
 - ただし、 $E(u_i|x_i) = 0$ を仮定している。
 - 回帰関数 は、 x の値を条件としたときに、 Y の条件付き期待値と解釈できる。
 - どんな関数形を想定すべきか
 - 近似的な関係式をあらかじめ指定する。
 - 1次式: $f(x_i) = \beta_0 + \beta_1 x_i$
 - 2次式、指数関数、対数関数、その他。
 - ある程度単純で、かつ、データの全体的な傾向を捉える
 - » 父親の身長と息子の身長データについては、1次式で十分であるように見える。
 - 近似的な関係式をデータから探す。
 - この講義ではあつかわない。

回帰モデルの仕組み(6)

- 線形回帰モデル

- $- Y_i = \beta_0 + \beta_1 x_i + u_i \quad (i = 1, 2, \dots, n)$

- 回帰係数

- 一定数項: β_0 、傾き: β_1

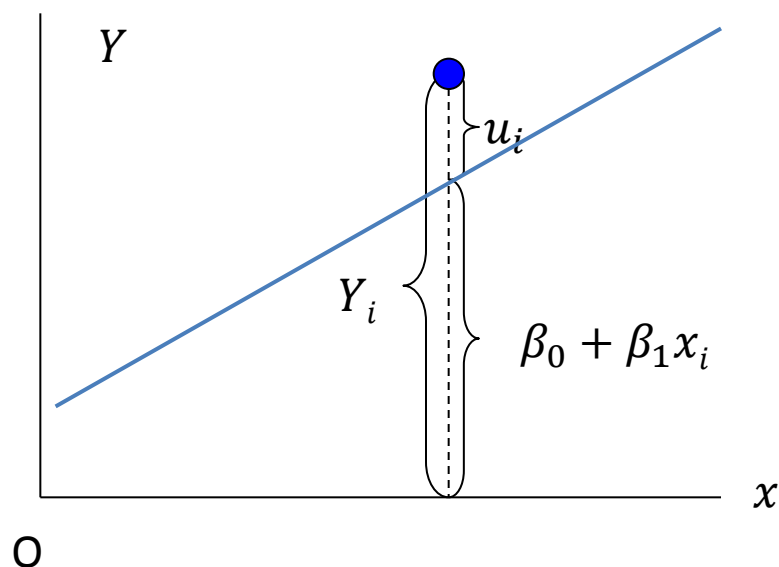


図3: 線形回帰モデル

最小2乗法(復習)(1)

- 最小2乗法

- $\sum_{i=1}^n \{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\}^2$ が最小になるような $\hat{\beta}_0$ と $\hat{\beta}_1$ を回帰係数の最小2乗推定量と呼ぶ。

- 解

- $$\begin{cases} \sum_{i=1}^n \{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\} = 0 \\ \sum_{i=1}^n x_i \{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\} = 0 \end{cases} \Leftrightarrow \begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \end{cases}$$

最小2乗法(復習)(2)

- 例:
 - Pearson の親子の身長データ
 - 回帰係数の推定値
 - $\hat{\beta}_1 = 0.51, \hat{\beta}_0 = 33.9$
 - 決定係数
 - $R^2 = 0.25$
 - 記述統計学(統計学Iの学習範囲)では、回帰係数の推定値と決定係数を計算して終わり。
 - 推測統計学(統計学IIの学習範囲)では、その先を考える。

最小2乗法(復習)(3)

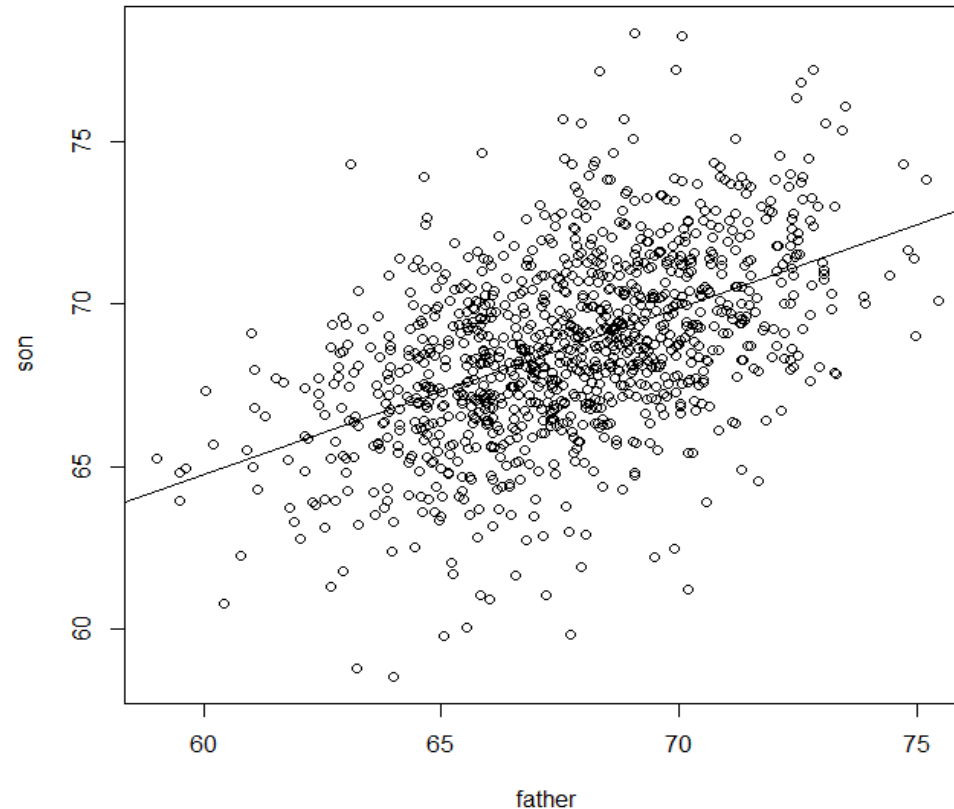


図4: 推定された回帰直線

野口・西郷(2014)
『基本 統計学』
p. 196

回帰係数とその推定量の関係(1)

- (母)回帰直線

- $y = \beta_0 + \beta_1 x$

- 確定(偶然的な変動をふくまない)。

- しかし、観察不可能(未知母数をふくむ)。

- 最小2乗法で推定された回帰直線

- $y = \hat{\beta}_0 + \hat{\beta}_1 x$

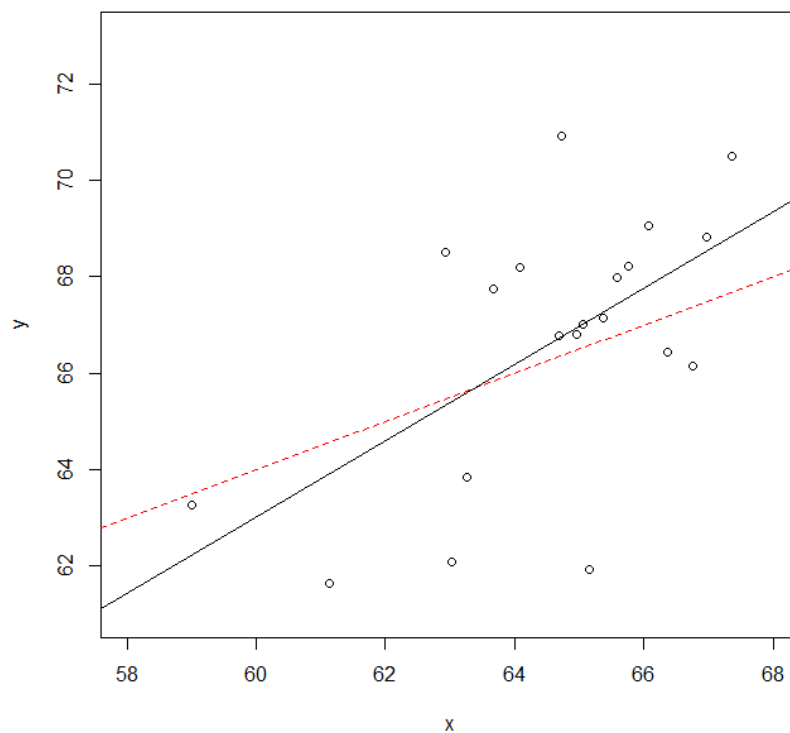
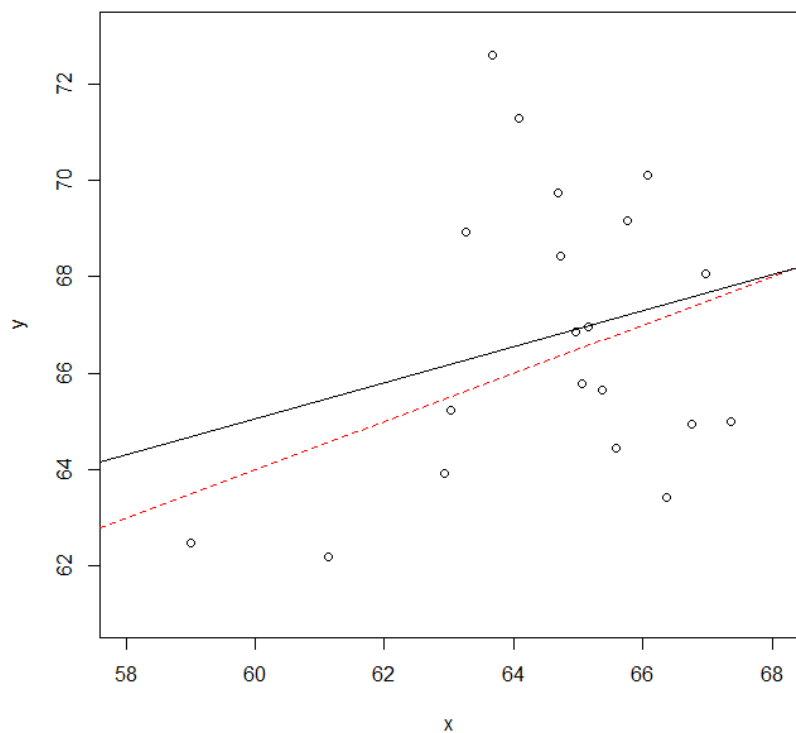
- 観察可能。

- しかし、不確定(偶然的な変動をふくむ)。

- 両者の関係は？

回帰係数とその推定量の関係(2)

図5: (母)回帰直線と推定された回帰直線
(赤:母回帰直線 黒:推定回帰直線)



回帰係数とその推定量の関係(3)

- 誤差の出方
 - 推定された回帰直線の影響
 - つまり、最小2乗推定量 $\hat{\beta}_0$ と $\hat{\beta}_1$ に影響
 - 誤差項の確率的な性質によって、最小2乗推定量 $\hat{\beta}_0$ と $\hat{\beta}_1$ の性質が決まる。
- 誤差項の性質(それに課される条件)を明示する必要あり。

誤差項に課される条件(1)

1. 期待値が0である: $E(u_i) = 0$
2. 分散が均一である: $V(u_i) = \sigma^2$
3. 相互に相関をもたない: $Cov(u_i, u_j) = E(u_i u_j) = 0$
4. 正規分布にしたがう。
5. 説明変数 x と無関係である。

誤差項に課される条件(2)

1. 期待値が0である:

- $E(u_i) = 0$

- $E(Y_i) = E(\beta_0 + \beta_1 x_i + u_i)$
 $= E(\beta_0 + \beta_1 x_i) + E(u_i) = \beta_0 + \beta_1 x_i$

- 所与の x のもとで、回帰直線が観察値の平均的な値をあらわす。

- つまり、「回帰直線によって、被説明変数 Y と説明変数 x との関係が適切に捉えられている」ということを

誤差項に課される条件(3)

図6: $E(u_i) = 0$ の場合

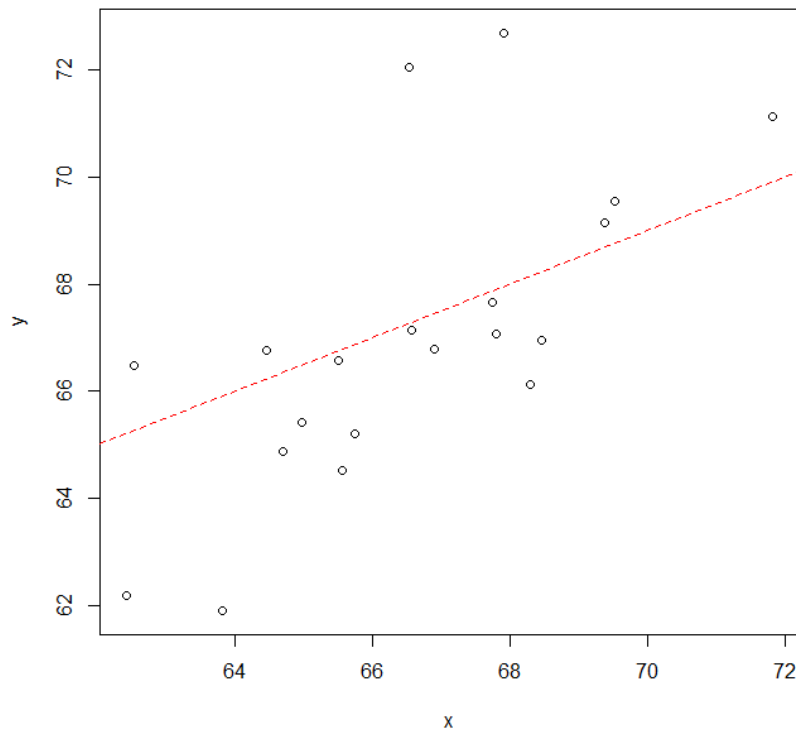
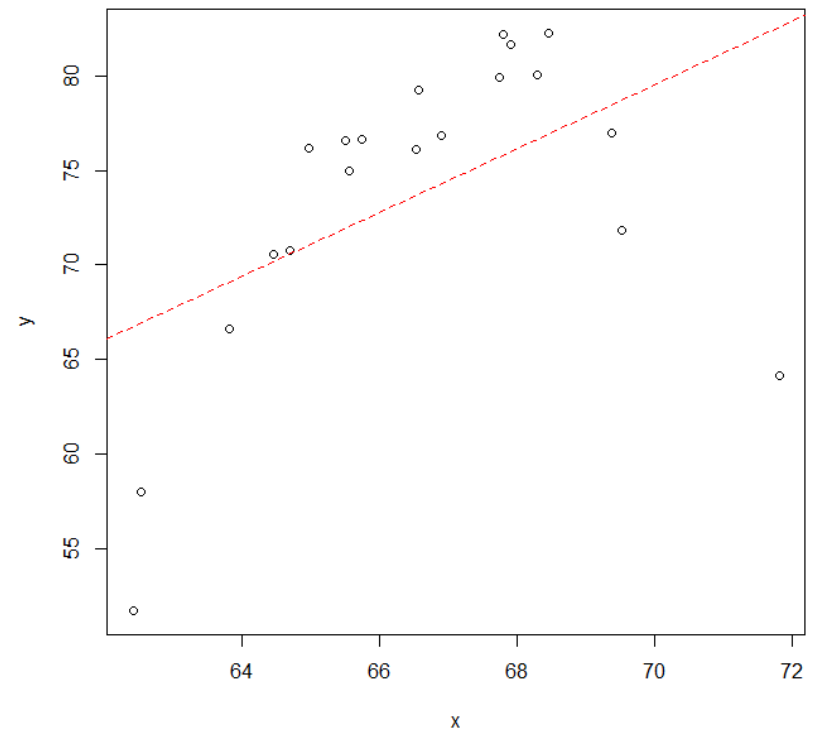


図7: $E(u_i) \neq 0$ の場合



誤差項に課される条件(4)

2. 分散が均一である:

- $V(u_i) = \sigma^2$

- $V(Y_i) = V(\beta_0 + \beta_1 x_i + u_i) = V(u_i) = \sigma^2$

- 説明変数 x の値によらず、縦軸方向の散らばりが一定である。
 - どの誤差も、回帰直線からの乖離の度合いとして、直接的に大小を比較できる。
 - » 不均一分散の場合を参照

誤差項に課される条件(5)

図8: 均一分散の場合

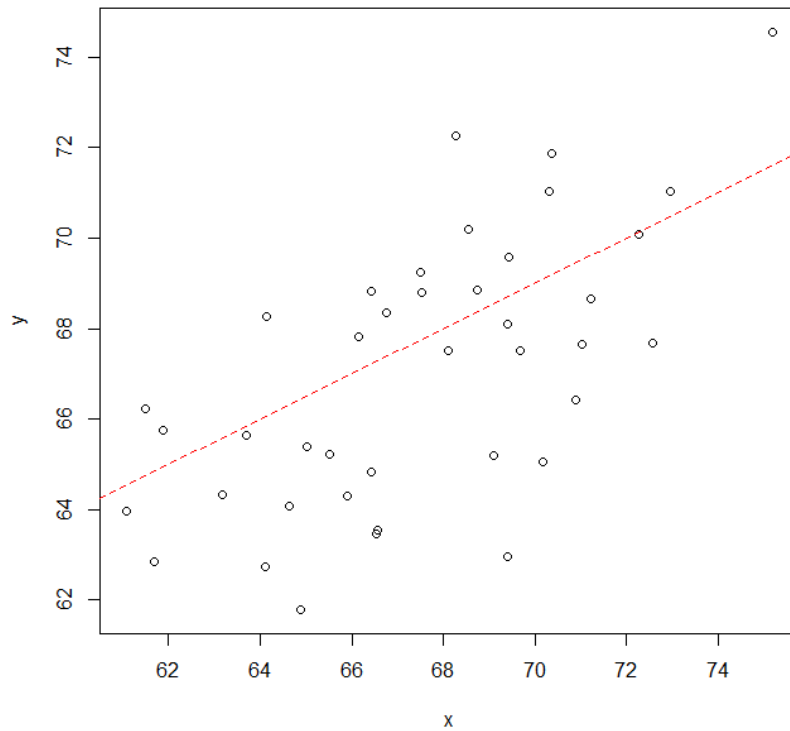
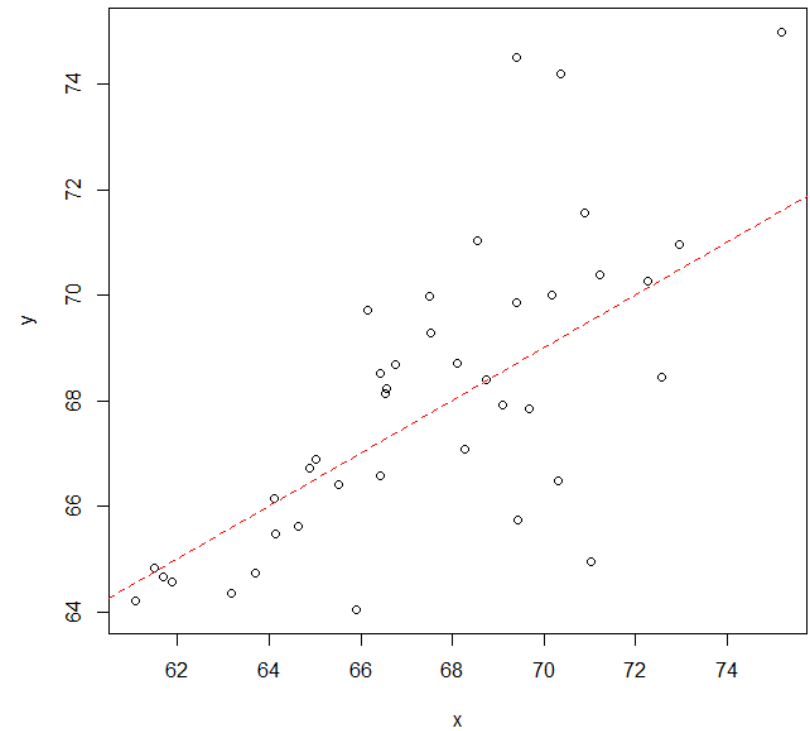


図9: 不均一分散の場合



誤差項に課される条件(6)

3. 相互に相関をもたない:

$$- Cov(u_i, u_j) = E(u_i u_j) = 0$$

- i 番目の観察値の誤差がプラスだろうとマイナスだろうと、 j 番目の観察値の誤差はそれと無関係に符号・値が定まる。

$$\begin{aligned} - Cov(Y_i, Y_j) \\ &= Cov(\beta_0 + \beta_1 x_i + u_i, \beta_0 + \beta_1 x_j + u_j) \\ &= Cov(u_i, u_j) = 0 \end{aligned}$$

誤差項に課される条件(7)

図10: 相互に無相関の場合

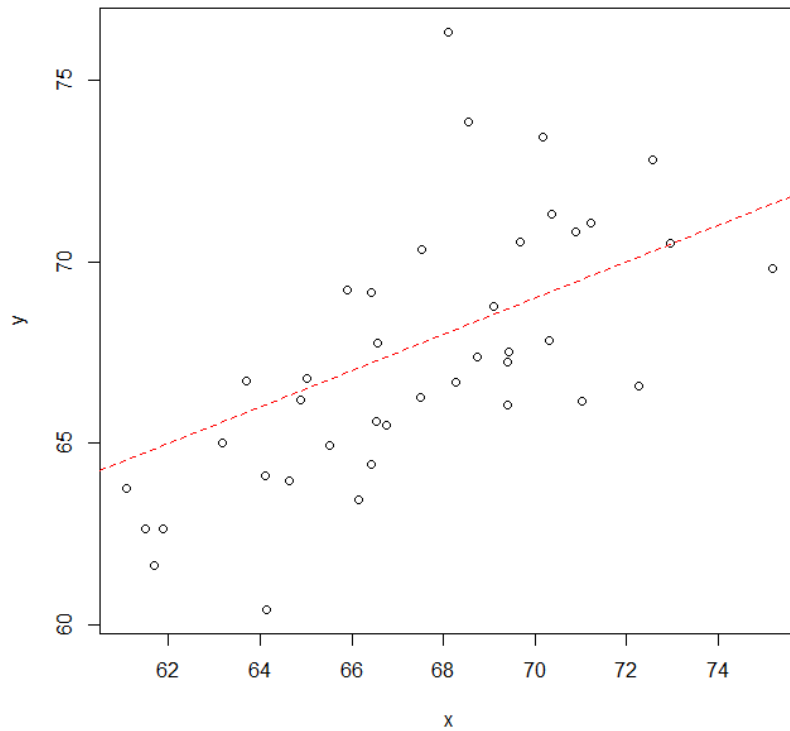
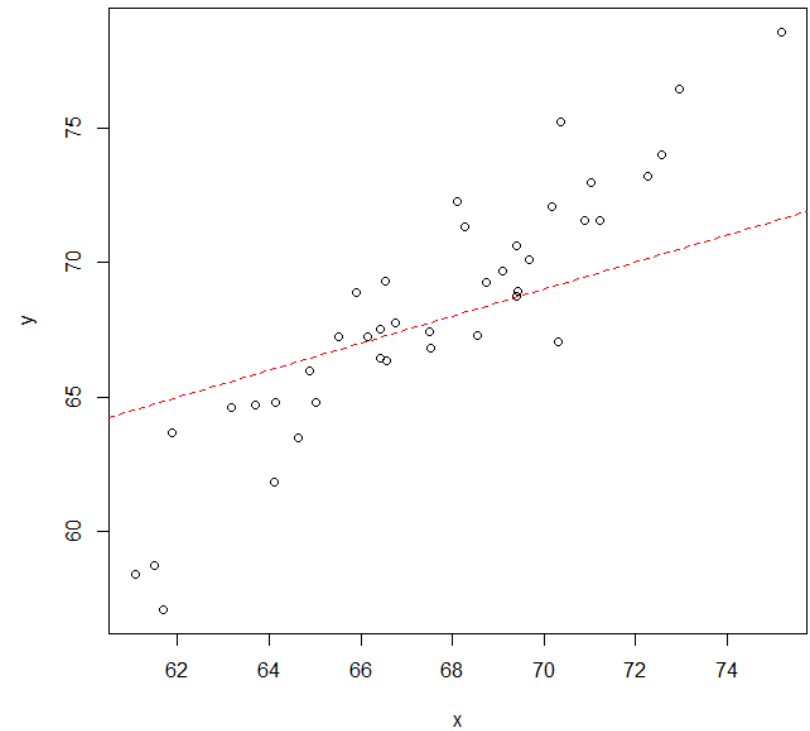


図11: 相関のある場合



誤差項に課される条件(8)

4. 正規分布にしたがう。

- 「主要な説明要因以外の種々雑多の説明要因を一括したものが誤差」であるなら、誤差の分布は正規分布で近似できると想定される。

5. 説明変数 x と無関係である。

- 説明変数 x が大きくても小さくても、誤差の分布は同じである。

誤差項の分散の推定(1)

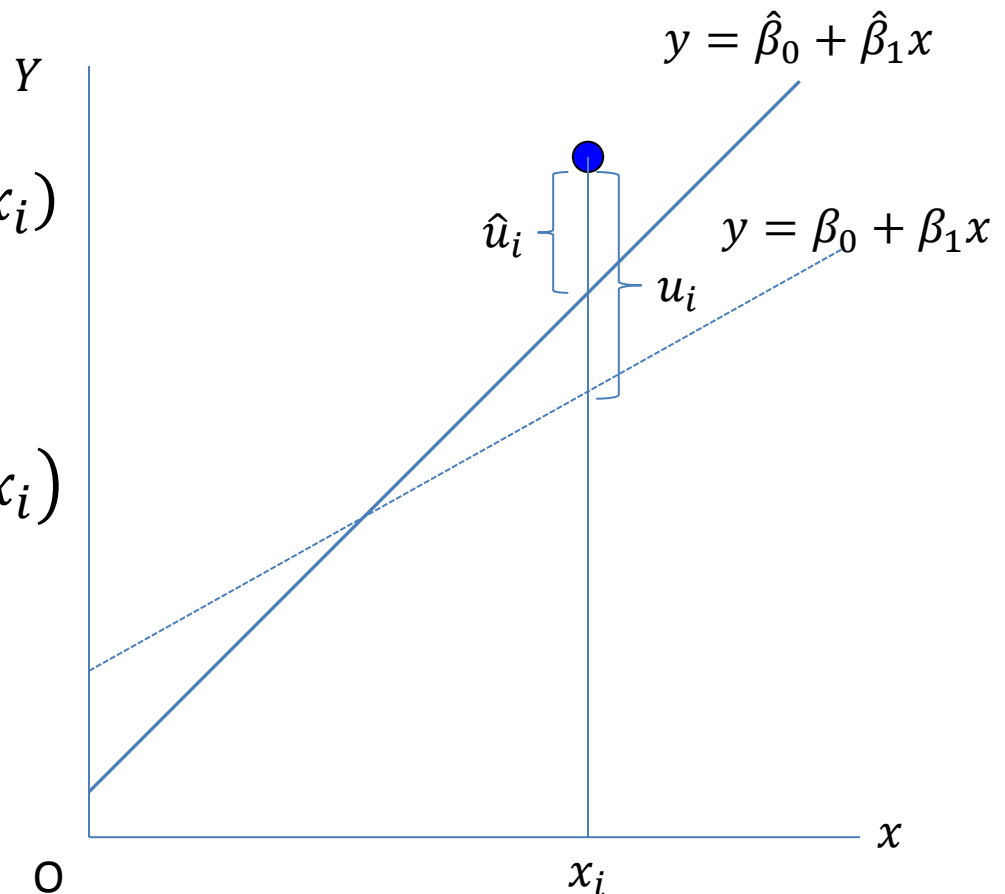
- 誤差項と残差

- 誤差項

- $u_i = Y_i - (\beta_0 + \beta_1 x_i)$
 - 観察不可能

- 残差

- $\hat{u}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$
 - 観察可能
 - 残差の性質
 - $\sum_{i=1}^n \hat{u}_i = 0$
 - $\sum_{i=1}^n x_i \hat{u}_i = 0$



誤差項の分散の推定(2)

- 誤差項の分散の推定量

- $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2$

- 誤差項の分散の推定量の性質

- 誤差項の分散の不偏推定量: $E(\hat{\sigma}^2) = \sigma^2$

- もし、誤差 u_i そのものが観察できれば、以下のように不偏推定量が求められる。

- $\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n u_i^2$

- しかし、誤差 u_i は観察不可能である。その代わりに残差 $\hat{u}_i$ を利用する。ただし、その際は、 $n - 2$ を分母に使う。

誤差項の分散の推定(3)

- 例
 - Pearson の親子の身長データ
 - $\hat{\sigma}_{obs}^2 = 5.94$