TAセッション1: イントロダクション

～統計学・計量経済学の復習を添えて～

TA: 北川 梨津

最終更新日: 2022-04-11

大湾秀雄ゼミTAセッションにようこそ

TAについて

- 名前: 北川 梨津
- 所属：早稲田大学大学院経済学研究科博士後期課程
- 関心：健康経営・人材採用・データ科学一般
- 論文：*Working from home and productivity under the COVID-19 pandemic: Using survey data of four manufacturing firms*（クリックして論文をチェック！）

このTAセッションについて

- シラバスを読んでください.
- 資料等は以下のウェブページとgithubレポジトリで管理.
  - https://ritsu1997.github.io/owanseminar/
  - https://github.com/ritsu1997/owanseminar

人事データ分析の重要性

- データ分析一般の重要性はもはやビジネスパーソンの常識.
- 人事もご多分に漏れない.
  - 勘や経験，先例主義，HRコンサルにだけ頼っていては，これからの時代，競争力を失う.
  - 企業の主要な生産要素は労働（人材）！→エビデンスに基づく最適な活用を目指したい！
- 人手不足も進む中で，なるべく従業員ひとりひとりの生産性を高め，また魅力的な職場をつくって優秀な人材を引きつけることの重要性も高まる一方．→人事データ分析の付加価値も高まる.
- 人事データ分析は**ピープルアナリティクス**，**HRアナリティクス**，**タレントアナリティクス**などのバズワードもあって，日本でも業界が盛り上がりつつある.
- ちなみに，計量経済学の人事データへの応用は**インサイダーエコノメトリクス**と呼ばれる.

人事データ分析の特徴

- 通常どんな企業にも必ず人事データがある．→人事データ分析というスキルの普遍性.
- 基本的にはビッグデータというほど大きくはない．→必ずしも予測タスクとは親和性が高くない（予測は倫理的な問題も！cf. **プロファイリング**）.
- 人事実務においては**因果関係の理解**が重要になる．→**計量経済学**や**統計的因果推論**が力を発揮.
  - 例　どういう職務経験がハイパフォーマンスに繋がるのか？現在行っている人事施策に効果はあるのか？
- 因果関係を理解すれば，人事上の意思決定の質向上に繋がる.
  - 例　現行の人事施策に効果が認められない場合，廃止するなど.
- もちろん，記述的な分析や予測的な分析も場合によっては有用.

因果推論と予測の違い

- 因果推論は統計モデルのパラメータに関心がある.
  - パラメータが施策などの因果効果に相当する.
- 予測は統計モデル自体には関心がなく（道具にすぎない），あくまでサンプルから学習したパターンを使って未知のデータの予測をしたい.
- 回帰モデルでのたとえ話：
  - $Y = \beta_0 + \beta_1 X + \varepsilon$ という真のモデルがあるとする.
  - データから，$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ という回帰式が推定できたとする.
  - この時，$\hat{\beta}_1$ に基づいた意思決定を志向するのが因果推論（e.g., 政策の実施・廃止）.
  - 一方で，$\hat{Y}$ に基づいた意思決定を志向するのが予測（e.g., レコメンデーション）.

統計学の復習（とてもざっくり！）

基本1

- **事象**（event）とは，何かしらのできごと．
  - 例「サイコロを振って偶数の目が出る」
- **確率**（probability）は，事象の起きやすさ．事象 A が起きる確率を，$\Pr(A)$ と書く．
- **条件付き確率**（conditional probability）は，ある事象が起きたということを所与とするときに，ある別な事象が起きる確率．事象 A を所与としたときの事象 B の条件付き確率を，$\Pr(B \mid A)$ と書く．なお，$\Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)}$ と容易に確かめられる．
- **確率変数**（random variables）とは，試行（$\simeq$ ランダムな行為）がもたらす事象を実数に対応させる関数．
  - 例「サイコロを振って偶数の目がでたら 1 の値をとり，そうでなければ 0 の値をとる確率変数 X 」

基本2

- **確率（質量）関数**（probability mass function; PMF）は，確率変数が具体的な値をとる確率を記述する関数．つまり，確率変数の確率質量関数 $p_X(x)$ は，

$$p_X(x) = \Pr(X = x)$$

と定義される．

  - 注意！「X」は確率変数，「x」はある具体的な値，「$X = x$」は事象．
  - さっきのサイコロの例だと，$p_X(1) = \Pr(X = x) = \Pr(\text{偶数の目が出る}) = 0.5$.

基本3

- **期待値**（expectated value）は，確率変数がとりうる値をその確率で重みづけて足し合わせたもの．つまり，

$$\mathbf{E}[X] = \sum_{x \in \mathcal{X}} x \times p_X(x)$$

と定義される．

  - さっきのサイコロの例だと，$\mathbf{E}[X] = 0 \times p_X(0) + 1 \times p_X(1) = 0.5$.
  - **平均**（mean）とも呼び，$\mu = \mathbf{E}[X]$ と表記されることが多い．

基本4

- **分散**（variance）は，確率変数がとりうる値と平均との差（$=$ **偏差**）の二乗をその値が実現する確率で重みづけて足し合わせたもの．つまり，

$$\mathbf{Var}[X] = \sum_{x \in \mathcal{X}} (x - \mu)^2 \times p_X(x)$$

と定義される．

  - さっきのサイコロの例だと，
    $\mathbf{Var}[X] = (0 - 0.5)^2 \times p_X(0) + (1 - 0.5)^2 \times p_X(1) = 0.25.$
  - $\sigma^2 = \mathbf{Var}[X]$ と表記されることが多い．
  - $\sigma = \sqrt{\mathbf{Var}[X]}$ を**標準偏差**（standard deviation）と呼ぶ．

基本5

- **同時確率関数**（joint probability function）とは，複数の確率変数の組が具体的な値の組をとる確率を記述する関数．つまり，2つの確率変数 X と Y があるとき，その同時確率関数 $p_{X,Y}(x,y)$ は，

$$p_{X,Y}(x,y) = \Pr(X = x, Y = y)$$

  と定義される．

  - つまり，$X = x$ という事象と $Y = y$ という事象が同時に起きる確率をすべての x, y の組み合わせについて記述している．

基本6

- **条件付き確率関数**（conditional probability function）は，$X = x$ が与えられたときの $Y = y$ となる確率を記述する関数．つまり，同時確率関数 $p_{Y|X}(y \mid x)$ は，

$$p_{Y|X}(y \mid x) = \Pr(Y = y \mid X = x)$$

  と定義される．

  - 条件付き確率の性質から，簡単に $p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$ と確かめられる．

基本7

- 以下の条件が成り立つとき，2つの確率変数 X と Y は互いに**独立**である（independent）と言う.

$$p_{X,Y}(x,y) = p_X(x)\,p_Y(y)$$

.

  - このことは，条件付き確率が条件に依存しないことと同じ（ $p_{Y|X}(y \mid x) = p_Y(y)$ ）.
  - つまり，一方が他方について何も追加の情報をもたらさないということ.
  - 2つの確率変数 X と Y が独立であるということを， $X \perp\!\!\!\perp Y$ と表記する.

基本8

- **条件付き期待値**（conditional expectated value）は，ある事象に条件付けたときに確率変数がとりうる値をその条件付き確率で重みづけて足し合わせたもの．つまり，$X = x$ で条件つけたときの確率変数 Y の条件付き期待値 $\mathbf{E}[Y \mid X = x]$ は，

$$\mathbf{E}[Y \mid X = x] = \sum_{y \in \mathcal{Y}} y \times p_{Y|X}(y \mid x)$$

  と定義される．

- **重要な命題**： $X \perp\!\!\!\perp Y \Rightarrow \mathbf{E}[Y \mid X = x] = \mathbf{E}[Y], \forall x.$
  - 独立ならば，$p_{Y|X}(y \mid x) = p_Y(y)$ であることから明らか．
  - 統計的因果推論の枠組みでは頻繁に使われるので，その意味をよく理解すること．

基本9

- ここまで**離散型確率変数**に限定して議論してきた.
  - つまり，とびとびの値をとるような確率変数.
  - 例　ダミー変数，カテゴリー変数.
- **連続型確率変数**もある.
  - 連続的な値をとる確率変数.
  - 例　身長，年収.
- 連続型確率変数の場合も直観はだいたい同じなので割愛.
  - $p(\cdot)$（確率質量関数）が $f(\cdot)$（**確率密度関数**）になって，$\sum$ が $\int$ になるくらい.

回帰分析の復習（とてもざっくり！）

回帰分析とは1

- ある確率変数 Y と1つ以上の確率変数 $X_1, X_2, \ldots, X_k$ の間の関係をデータから推定する分析.
- Y を $X_1, X_2, \ldots, X_k$ の関数として捉えるようなイメージ.
- つまり,

$$Y = f(X_1, X_2, \ldots, X_k)$$

  というような関係があるとして，この f の形をデータから推定する作業.

- Y を**従属変数**（dependent variable），**非説明変数**（explained variable），**目的変数**（objective variable）などと呼ぶ.
- $X_1, X_2, \ldots, X_k$ を**独立変数**（independent variables），**説明変数**（explanatory variables），**特徴量**（features）などと呼ぶ.

回帰分析とは2

- Y と $X_1, X_2, \ldots, X_k$ はデータから観察される変数.
- Y は観察される変数だけですべて決定されることは少なく，観察できない変数の影響も受けると考えるのが自然.
  - 例) 年収 $=$ $f($ 教育年数 $)$ としたときに，教育年数が年収を部分的を決定することは考えられるが，完全にそれだけで説明できるとは思われない.
- そのような観察されない変数の影響をひとまとめに ε と書くことにして,

$$Y = f(X_1, X_2, \ldots, X_k) + \varepsilon$$

  として観察されない変数の影響を明示的に書くことにする.
- ε は**誤差項**（error term）や**撹乱項**（disturbance term）と呼ばれる.

回帰分析とは3

- $f(X_1, X_2, \ldots, X_k)$ の形を規定するパラメータ（のベクトル）として，β を導入する．つまり，

$$Y = f(X_1, X_2, \ldots, X_k \mid \beta) + \varepsilon$$

として，パラメータを明示的に書く．

- このような式を，**回帰式**（regression equation）や**回帰モデル**（regression model）と呼ぶ．
- 特に，次の関数形を仮定するモデルを，**線形回帰モデル**（linear regression model）と呼ぶ．

$$\begin{aligned} Y &= f(X_1, X_2, \ldots, X_k \mid \beta) + \varepsilon \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon. \end{aligned}$$

回帰分析とは4

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

- 線形回帰モデルは実証分析で非常によく使われる.
- まずこの線形回帰モデルによる分析をマスターすることが，立派なデータサイエンティストへの近道.
- β_0 は**切片**（intercept）と呼ばれる.
- $\beta_1, \beta_2, \ldots, \beta_k$ は**回帰係数**（regression coefficients）や**傾き**（slope）と呼ばれる.
- 回帰係数 β_l は，説明変数 X_l が 1 単位増えたときに，被説明変数 Y が何単位増えるかに対応すると解釈される. （後で詳述）
  - 線形回帰モデルの両辺を X_l について偏微分すると，$\frac{\partial Y}{\partial X_l} = \beta_l$ となる.

回帰分析とは5

- 説明変数が1つしかない線形回帰モデルを特に**線形単回帰モデル**（simple linear regression model）と呼ぶ.
- 説明変数が2つ以上ある線形回帰モデルを特に**線形重回帰モデル**（multiple linear regression model）と呼ぶ.
- ここからしばらくは，議論の見通しをよくするために次のような線形単回帰モデルを考える.（直観は説明変数が複数の場合も同じ）

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon.$$

- 回帰モデルを具体的に定式化すること（どの説明変数を入れるか？どのような関数形にするか？）を**モデルの特定化**（model specification）と言う.

回帰分析とは6

- 線形回帰モデルにいくつかの仮定をおく.
- まず，なんでもない仮定として $\mathbf{E}[\varepsilon] = 0$ をおく．（innocuousな仮定）
- 次に，**平均独立の仮定**（mean independence）をおく．つまり，$\mathbf{E}[\varepsilon \mid X_1] = \mathbf{E}[\varepsilon]$.
- 誤差項の期待値はゼロという仮定とあわせて，**条件付き期待値ゼロの仮定**という．つまり，$\mathbf{E}[\varepsilon \mid X_1] = 0$.
- このとき，回帰式の両辺の条件付き期待値をとると，

$$\mathbf{E}[Y \mid X_1 = x] = \mathbf{E}[\beta_0 + \beta_1 X_1 \mid X_1 = x] + \mathbf{E}[\varepsilon \mid X_1 = x]$$

$$= \beta_0 + \beta_1 x.$$

- これを**母回帰関数**（pupulation regression function）と呼ぶ.

最小二乗法について

- 統計学的な問題は，**「線形回帰モデルのパラメータ β を手元のデータからどうやって推定するか」**というものに帰着する．
  - 例えば，$Engagement = \beta_0 + \beta_1 HoursWorked + \varepsilon$ という回帰モデルのパラメータ β_0, β_1 をどうやってデータから推定するか？
  - 我々はモデルが $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ という形であるということを仮定しているが，β_0, β_1 の値は未知であることに注意．
- $\{(Y_i, X_{1i})\}_{i=1}^n$ というデータ（サンプル）が得られたとき，そこから回帰モデルのパラメータ β_0, β_1 を直接観察することはできないので，一定の手順に従って**推定**（estimate）する必要がある．

- $\{(Y_i, X_{1i})\}_{i=1}^n$ というデータ（サンプル）が得られたとき，

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

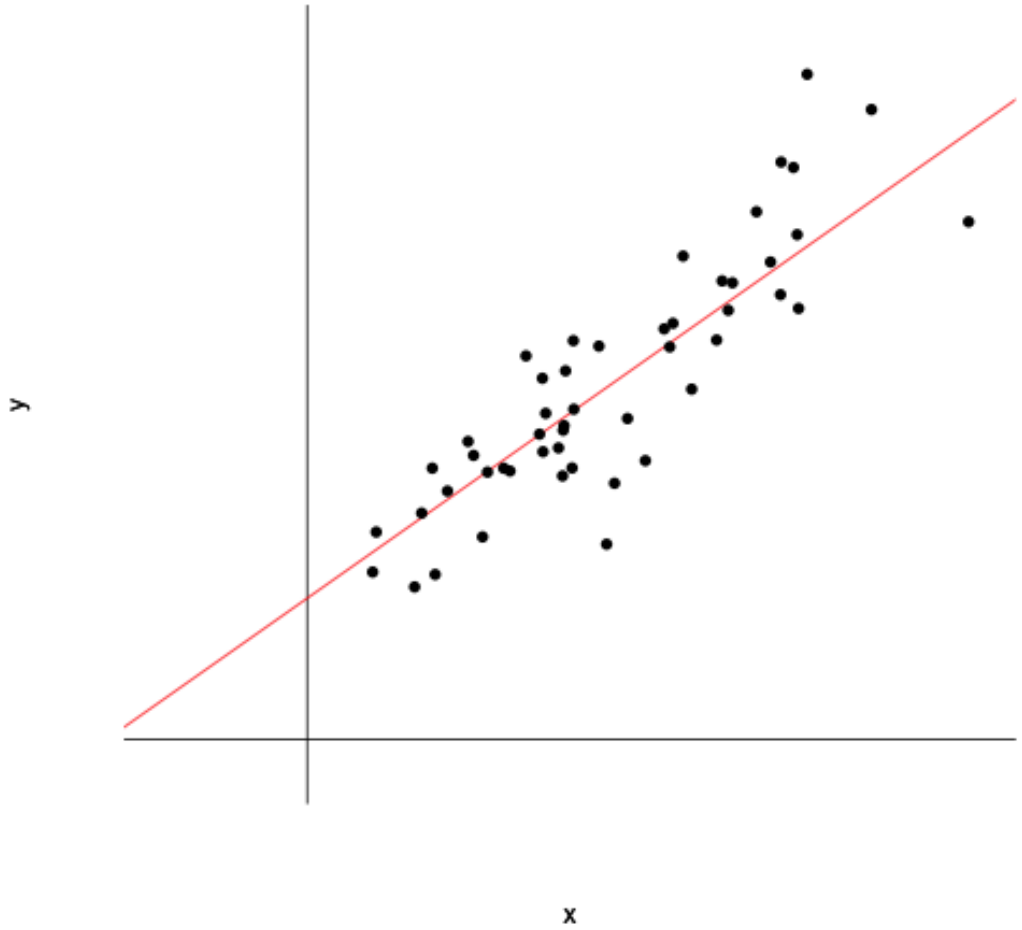
という線形回帰モデルを仮定すると，データについても次のような式が成り立つはずである：

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

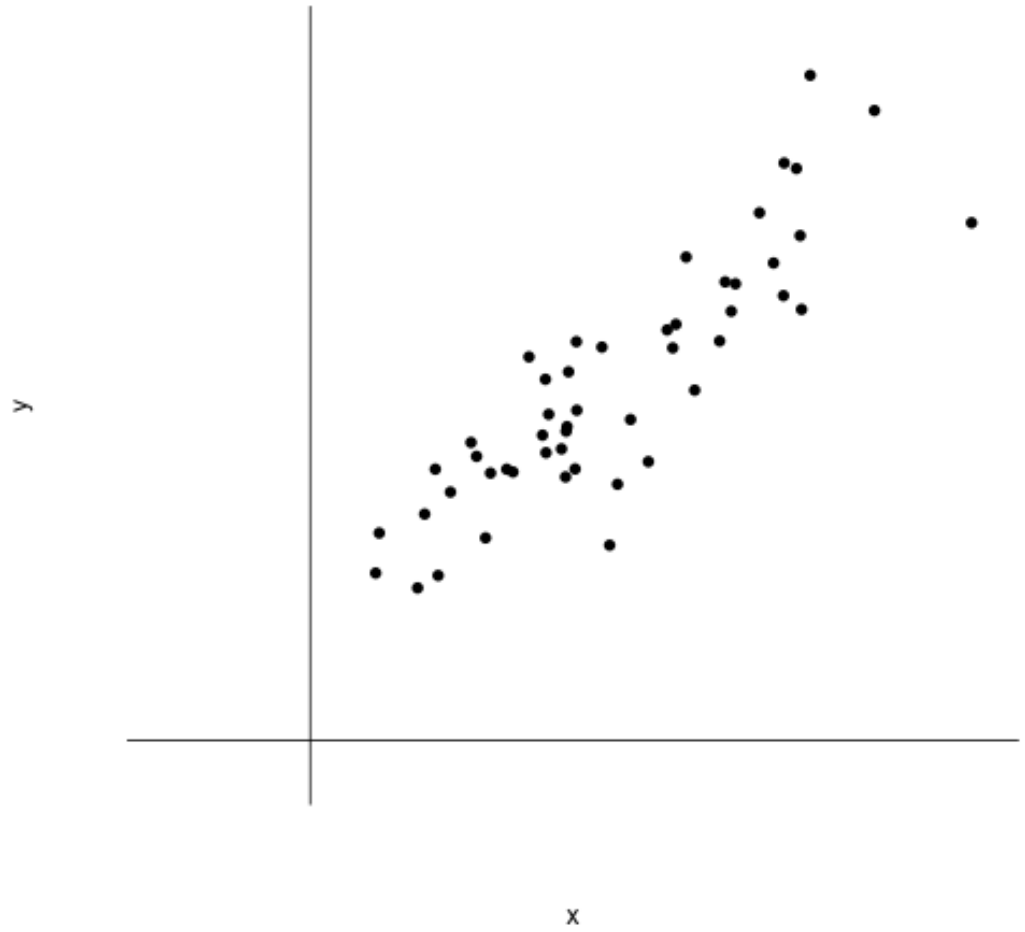
- Y_i と X_i 以外はデータから直接観測できないことに注意.
- 未知のパラメータ β_0 と β_1 の値を決めるというのは X_i と Y_i の散布図上に直線を引くことに相当する.
- イメージ： $Y = \beta_0 + \beta_1 X + \varepsilon \to \{Y_i, X_i\}_{i=1}^n \to Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \to \hat{\beta}_0, \hat{\beta}_1$.

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

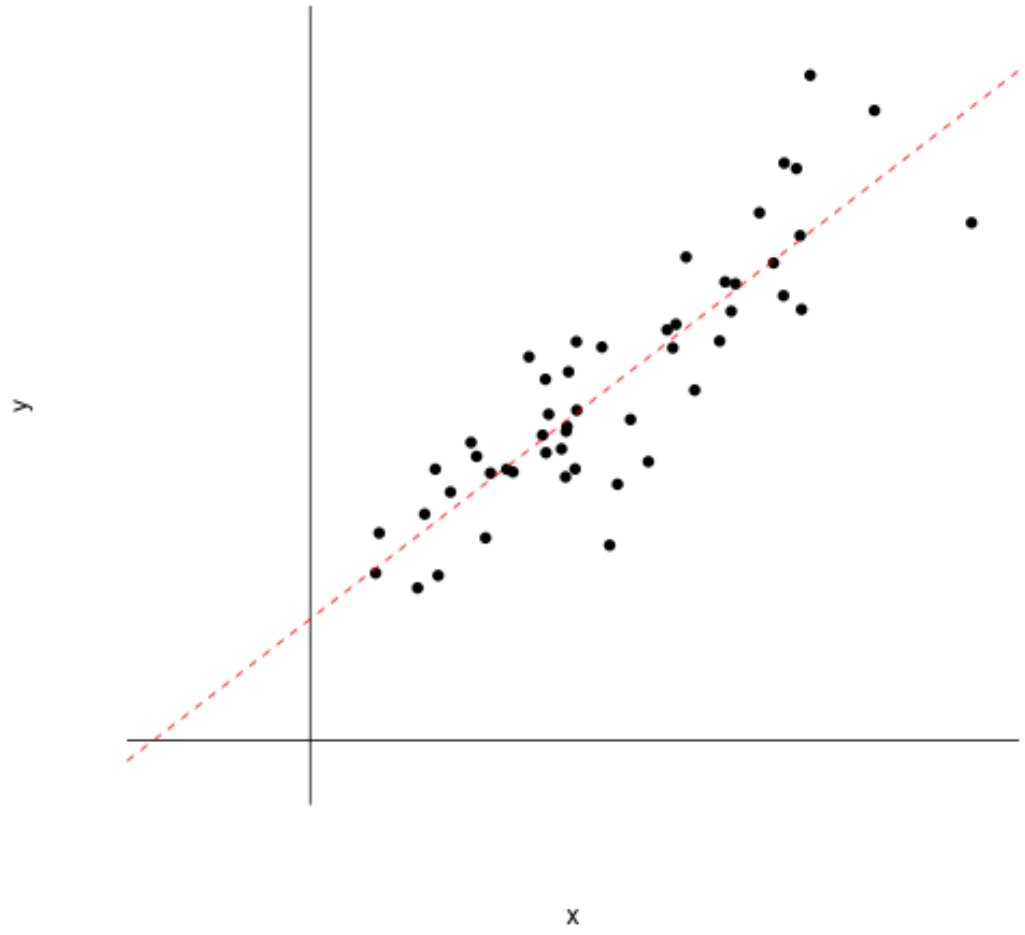
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$



$$\{Y_i, X_i\}_{i=1}^{n}.$$



$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \varepsilon_i.$$



- では，データからどのように直線を引けばよいか？
- データからパラメータを推定する手法の一つが**最小二乗法**（ordinary least squares; **OLS**）.
- 直線と各点からのズレが平均的にできるだけ小さくなるように定めたもの.
- 候補となる直線を $Y_i = b_0 + b_1 X_i$ とすると，直線と各点の縦方向に関するズレは $Y_i - b_0 - b_1 X_i$ となる.
- これの二乗和（**残差平方和**（sum of squared residuals）とも呼ぶ）

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

を最小化する b_0^*, b_1^* を**OLS推定量**と呼び，$\hat{\beta}_0, \hat{\beta}_1$ と表記する.

  - 記号 ^ は推定量（または推定値）を表すときによく用いられる．特に，OLS推定量であることを強調する必要があるとき，$\hat{\beta}_1^{OLS}$ などと表記されることもある.

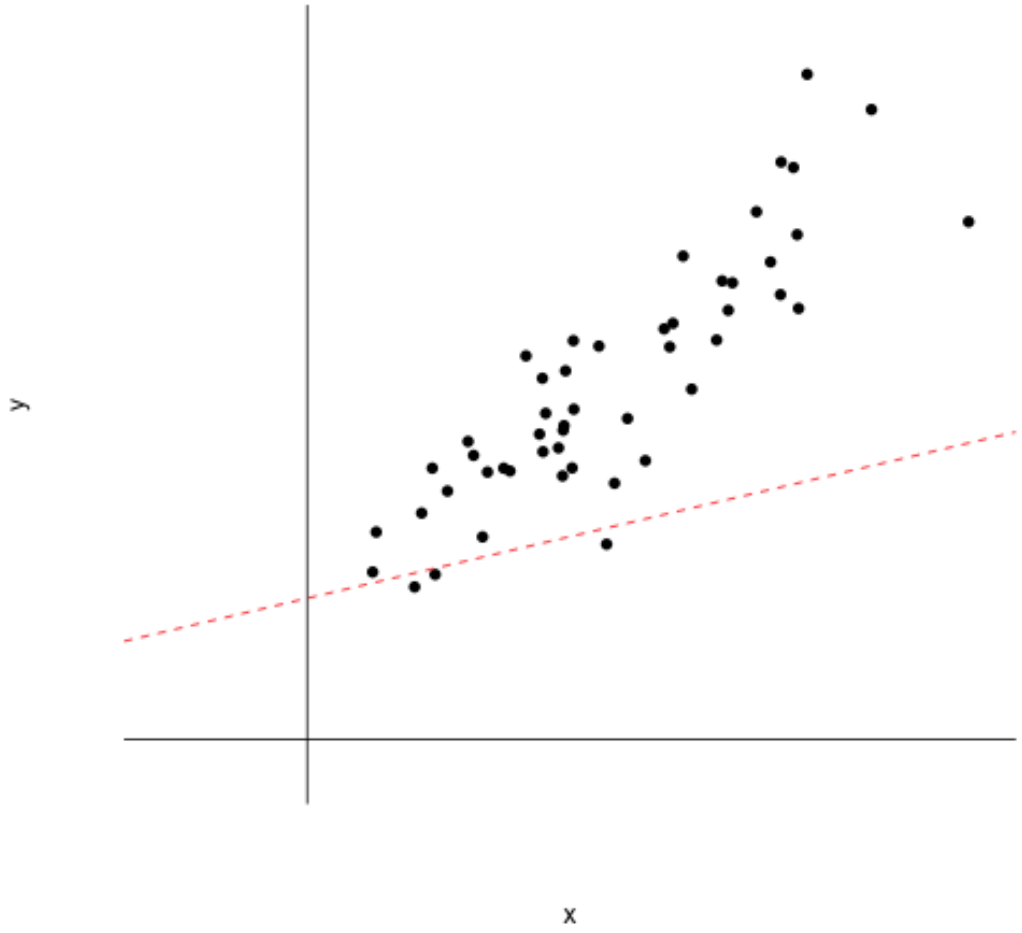
- $\min_{(b_0,b_1)} \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$ の最小化問題を解くと，OLS推定量が次の通り求まることがわかる：

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

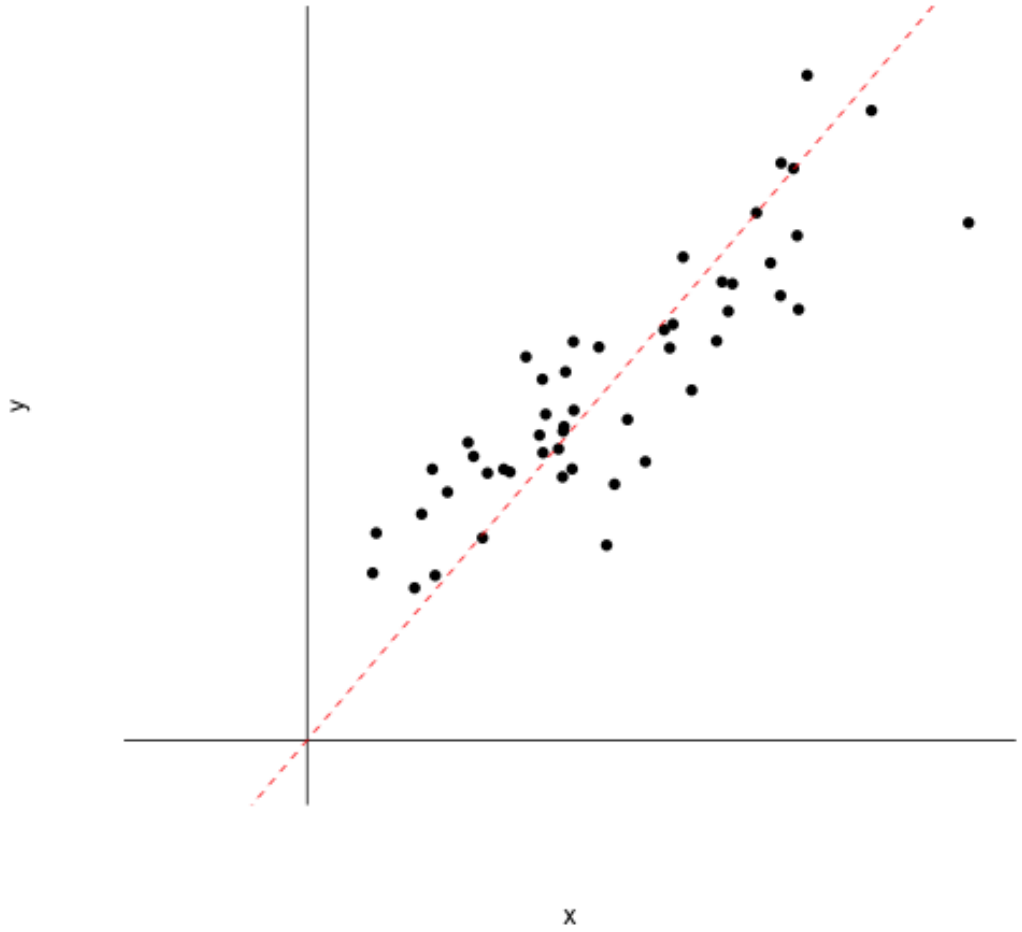
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

- ただし，$\bar{X}$ と $\bar{Y}$ はそれぞれ標本平均.
- $\hat{\beta}_1$ について，標本共分散 ÷ 標本分散になっていることがわかる.
- OLS推定により得られたパラメータの推定値から定められる回帰直線 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ を**標本回帰関数**（sample regression function）や**予測線**（fitted line）と呼ぶ.
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ を Y_i の**予測値**（fitted value）と呼び，また $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ を**残差**（residuals）と呼ぶ.

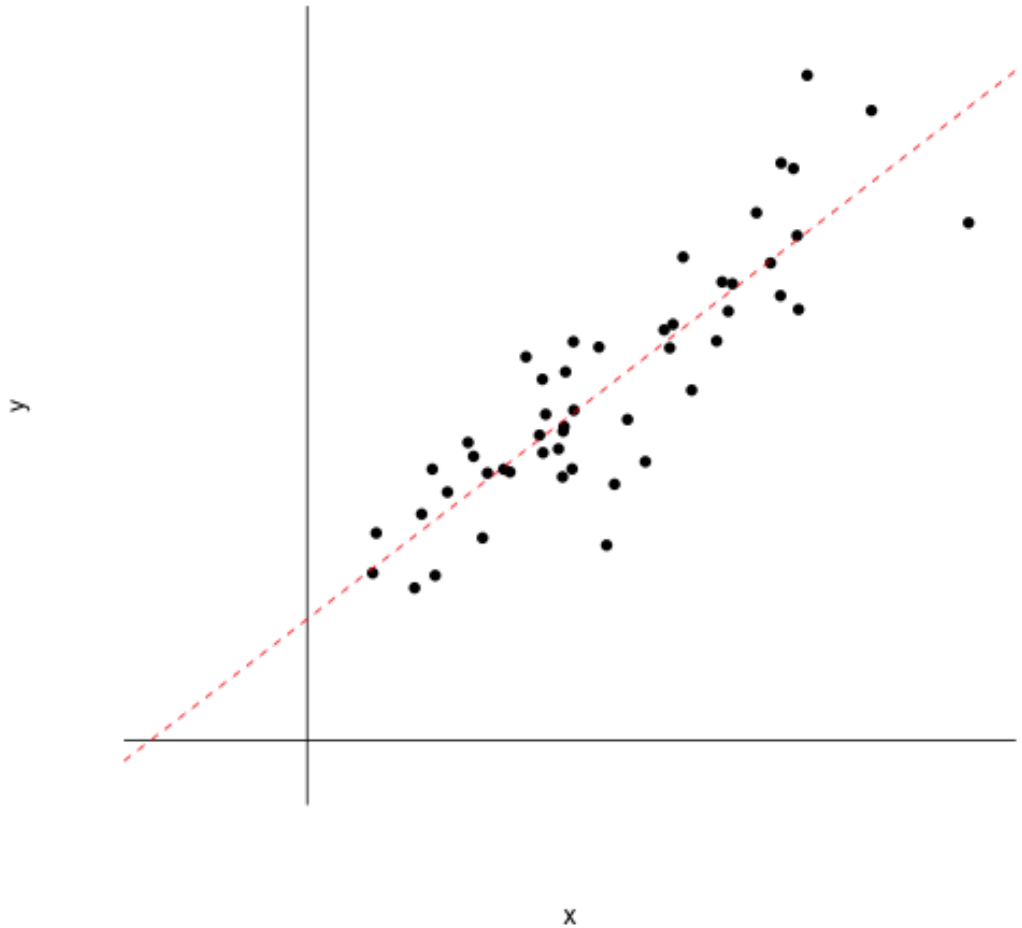
- $b_0 = 4, b_1 = 1.$



- $\beta_0 = 0, \beta_1 = 5$.



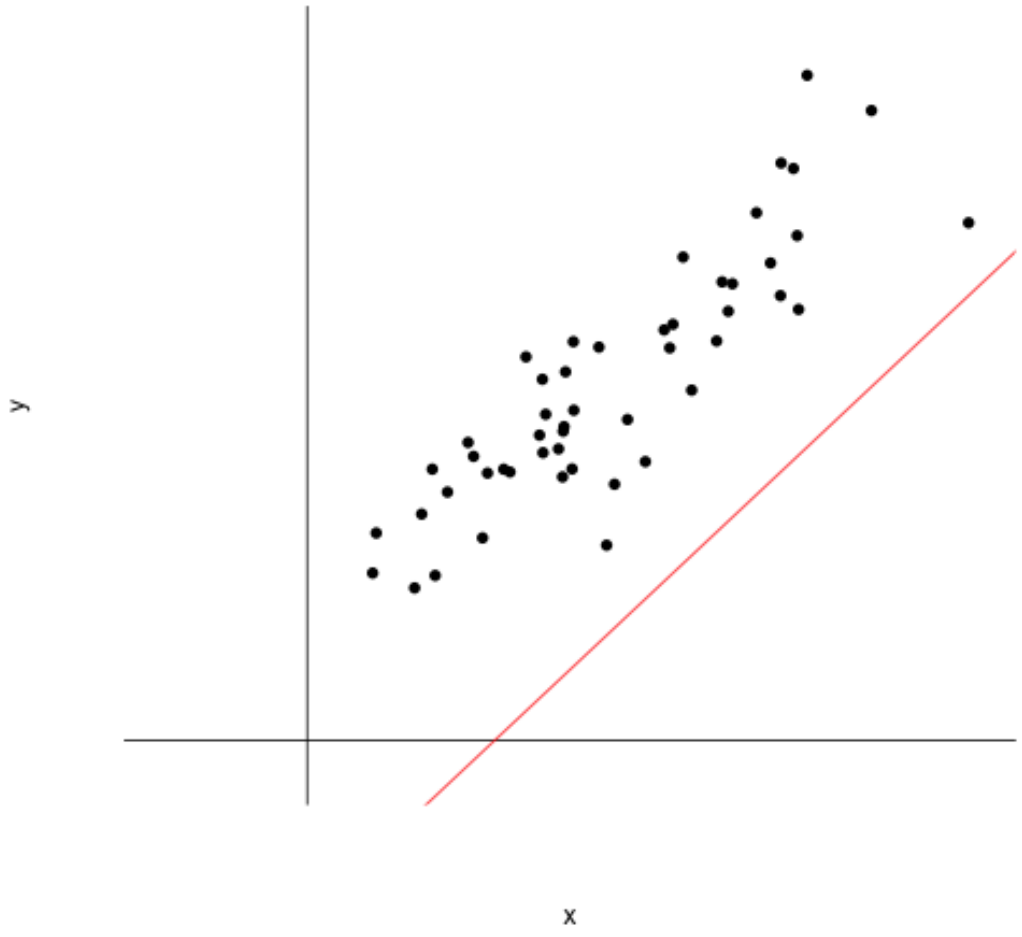
- $b_0 = \hat{\beta}_0^{OLS}, b_1 = \hat{\beta}_1^{OLS}$.



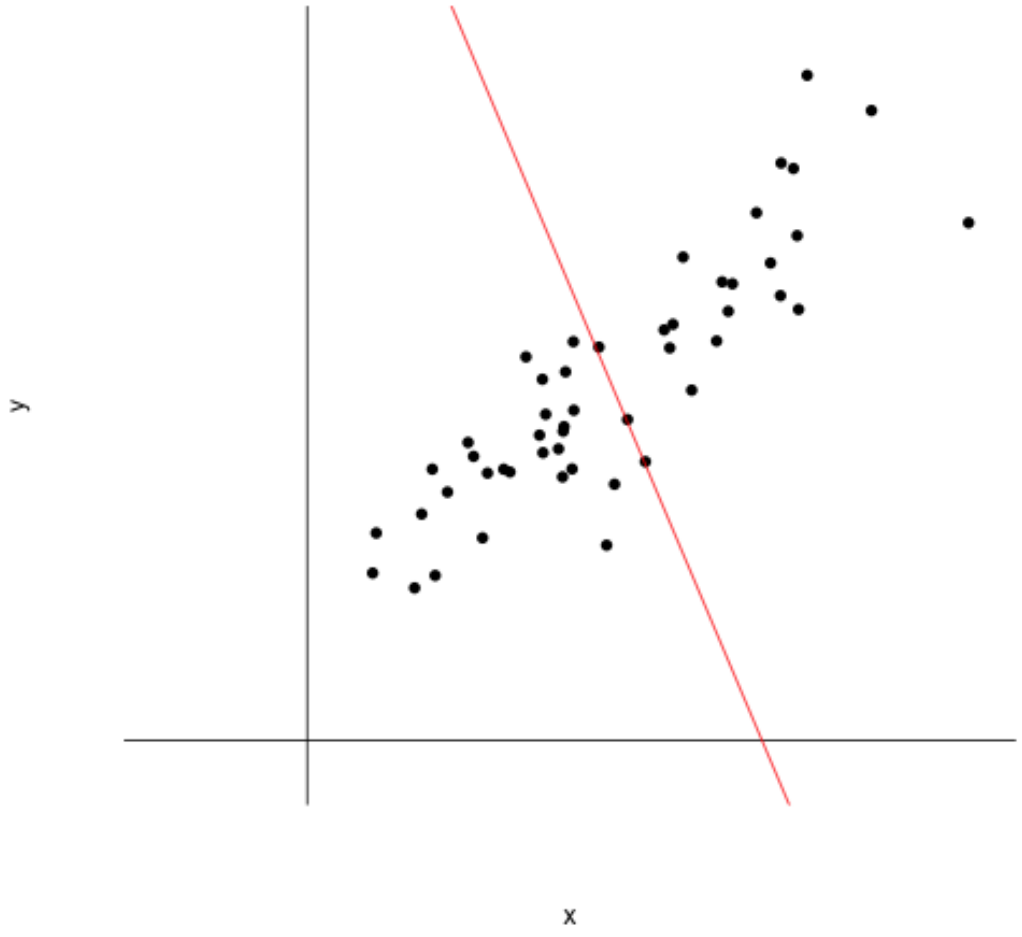
よくある疑問

- なぜ二乗するのか？
  - 二乗しないまま和をとると，目的関数の最小化（最大化も同じ）が解を持たなくなる($\simeq$ b_0, b_1 の値によっていくらでも小さくしたり大きくしたりできる).
- 二乗しないまま和をとって，$\sum_{i=1}^{n} Y_i - b_0 - b_1 X_i = 0$ となるような b_0, b_1 を使うのは？
  - そのような b_0, b_1 の組み合わせは無数に存在する.
- 二乗の代わりに絶対値をとるのは？
  - それは**最小絶対偏差法**（least absolute deviations; **LAD**）と呼ばれる推定方法で，外れ値に強いという性質がある.
  - あまり使うことはないと思うので，あまり気にしなくてOK.

直線を下にずらし続ければ，$\sum_{i=1}^{n} Y_i - b_0 - b_1 X_i$ をいくらでも小さくできる.



$\sum_{i=1}^{n} Y_i - b_0 - b_1 X_i \simeq 0$ となる直線の例（他にも無数にある）.



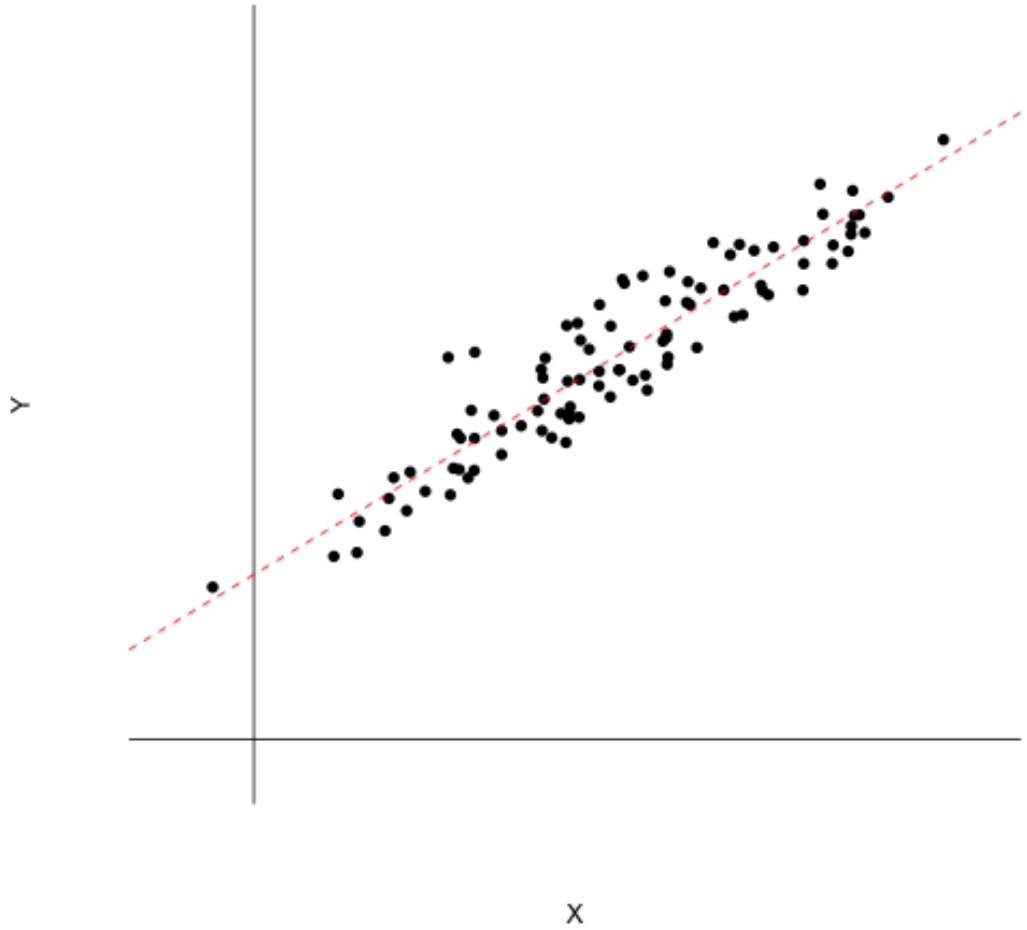
決定係数

- 次の量 R^2 を**決定係数**（coefficient of determination）と呼ぶ：

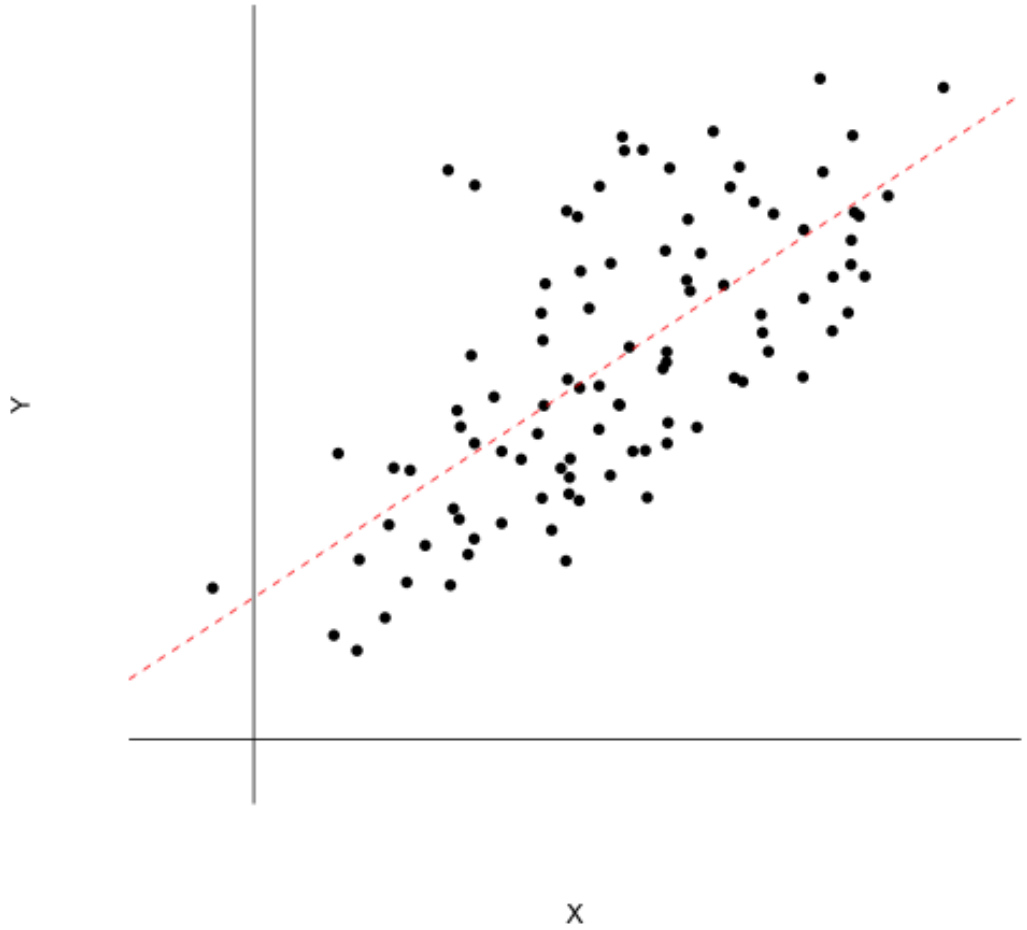
$$R^2 = \frac{\mathrm{SSE}}{\mathrm{SST}}.$$

- ただし，$\mathrm{SSE} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$, $\mathrm{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$.
- つまり，従属変数の変動のうち，どれくらいの変動を説明できているかという比率.
- モデルの**あてはまりの良さ**（goodness of fit）の指標として使われる.
- 決定係数は説明変数の数について増加関数になっているという問題がある．（関係ない説明変数でも増やせば R^2 を大きくできる．）
- 代わりに，説明変数の数についてペナルティを与えるような**調整済み決定係数**（adjusted R-squared）もよく使われる.

$R^2 = 0.89$



$R^2 = 0.52$



決定係数

- 決定係数が小さくても推定量の性質（不偏性や一致性）には関係しない.
- 推論という観点からは，実務上これ自体に特別な意味がある指標ではない.
  - 欠落変数の存在を診断するときに使われたりする (Oster, 2019).
- 予測という観点からも，モデルパフォーマンスの指標としてはあまり使えない. （cf. こちらの記事）
- 結論：あまり気にしなくてもいい.

OLS推定量の性質

OLS推定量はいくつかの条件が満たされるとき，推定する上で望ましい性質を持つ.

- **ガウス・マルコフの仮定**：
  - 仮定1. **パラメータについて線形**（linear in parameters）.
  - 仮定2. **無作為抽出**（random sampling）.
  - 仮定3. **完全な共線性がない**（no perfect collinearity）.
  - 仮定4. **条件付き期待値ゼロ**（zero conditional mean）.
  - 仮定5. **均一分散**（homoskedasticity）.
- 仮定1 - 4が満たされるとき，OLS推定量は**不偏性**と**一致性**を満たす.
- 仮定1 - 5が満たされるとき，OLS推定量は**最良線形不偏推定量**（best linear unbiased estimator; **BLUE**）となる. （**ガウス・マルコフの定理**）

仮定1. パラメータについて線形

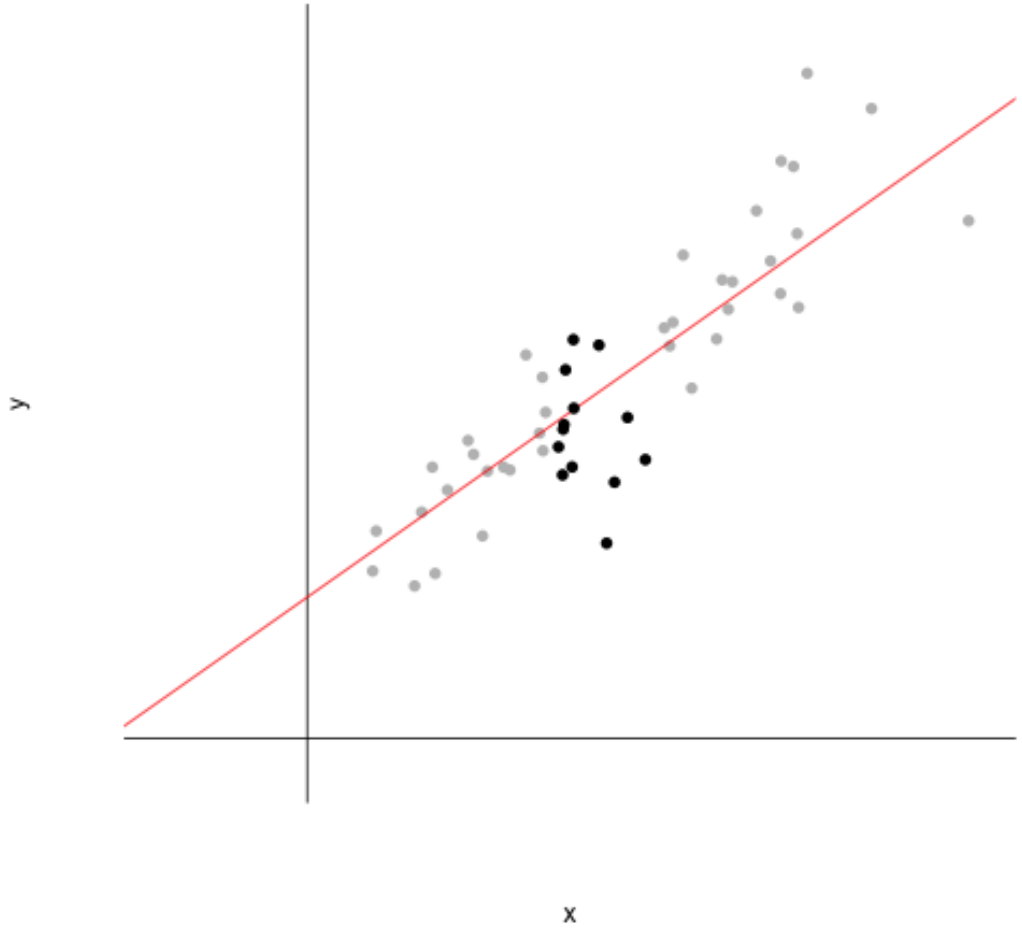
- 母集団における回帰モデル（population model）が,

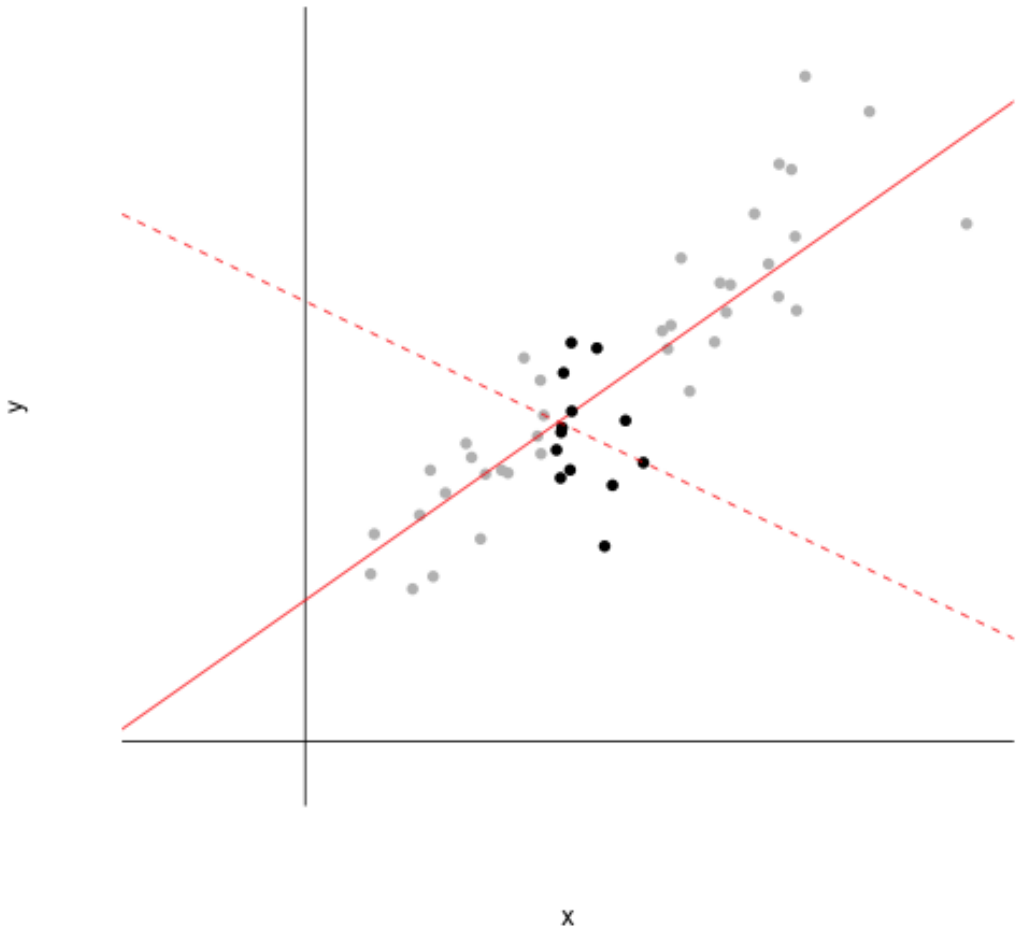
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

  と書けるという仮定.
- 母集団における回帰モデルを**真のモデル**（true model）と呼ぶこともある.
- 「ooがxxについて線形」とは，ざっくり言うと「ooとxxを2次元平面に書くと直線になっている」ということに相当する．つまり，Y と β のグラフは直線になるということ.
- パラメータについて線形ではない例：
  - $Y = \beta_0 + \frac{\beta_1}{\beta_1 - \beta_2} X_1 + \beta_2 X_2 + \varepsilon$
- パラメータについて線形な例：
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$
  - $Y = \beta_0 + \beta_1 \log(X_1) + \beta_2 \exp(X_2) + \varepsilon$
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$

仮定2. 無作為抽出

- 得られたサンプル $\{Y_i, X_1, X_2, \ldots, X_k\}_{i=1}^n$ が仮定1のモデルから得られたランダムサンプルであるという仮定.
- 偏ったサンプルに基づいて回帰モデルを推定すると，真のモデルからはかけ離れた推定値になる.


y
x


y
x

仮定3. 完全な共線性がない

- サンプルと母集団において次がなりたつという仮定：

  1. 定数の変数がない，かつ,
  2. 厳密に線形な関係が変数の間にない.

- 2はモデルの変数が他の変数の線形結合になっていないことを要請する.
- 変数どうし（またはある変数と他の変数の線形結合）の相関係数が 1 や -1 になってはいけないということ.
- 変数どうしが相関することを**許容しないわけではない**ことに注意.
- 完全に相関するのがダメというだけで，相関するのは別に良い.
- 完全に相関すると，OLS推定量は計算できなくなる.　（変数をどれか諦めてドロップするしかない.　）

仮定4. 条件付き期待値ゼロ

- 説明変数と誤差項に関係がないことを要求する仮定．つまり，

$$\mathbf{E}[\varepsilon \mid X_1, X_2, \ldots, X_k] = 0.$$

- 説明変数と誤差項に関連がないということ．（相関がないと言うと厳密さを欠くが，その理解でもよい．）
- この仮定が満たされる時，説明変数は**外生的**（exogenous）であると言われ，そのような変数を**外生変数**（exogenous variables）と呼ぶ．
- 誤差項 u と j 番目の説明変数 X_j が相関するとき（i.e., $\mathrm{Cov}(X_l, u) \neq 0$ ），X_j を**内生的**（endogenous）であると言われ，そのような変数を内生変数（endogenous variable）と呼ぶ．

OLS推定量の不偏性

- 仮定1から仮定4が満たされるとき,

$$\mathbf{E}[\hat{\beta}_j] = \beta_j, \quad (j = 0, 1, \ldots, k)$$

  である.

- つまり，OLS推定量は**不偏性**を満たす.

[証明] $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ であり，$\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{\varepsilon}$ であることに注意して，

$$\hat{\beta}_1^{OLS} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$= \frac{\sum_{i=1}^n (X_i - \bar{X})[(X_i - \bar{X})\beta_1 + (\varepsilon - \bar{\varepsilon})]}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$= \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X})\varepsilon_i}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

したがって，

$$\mathbf{E}[\hat{\beta}_1^{OLS}] = \beta_1 + \mathbf{E}\left[\frac{\sum_{i=1}^n (X_i - \bar{X})\varepsilon_i}{\sum_{i=1}^n (X_i - \bar{X})^2}\right].$$

仮定2と仮定4が成り立てば，繰り替えし期待値の法則により，

$$\mathbf{E}\left[\frac{\sum_{i=1}^{n}(X_i-\bar{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right] = \mathbf{E}\left[\mathbf{E}\left[\frac{\sum_{i=1}^{n}(X_i-\bar{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i-\bar{X})^2} \,\middle|\, X_1, X_2, \ldots, X_n\right]\right]$$

$$= \mathbf{E}\left[\frac{\sum_{i=1}^{n}(X_i-\bar{X})\mathbf{E}[\varepsilon_i \mid X_i]}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]$$

$$= 0.$$

したがって，

$$\mathbf{E}[\hat{\beta}_1^{OLS}] = \beta_1$$

を得る．

欠落変数バイアス

- 仮定4が満たされないことを，**内生性**（endogeneity）の問題と呼ぶ.
- その典型例が**欠落変数バイアス**（omitted variable bias; **OVB**）というもの.
- 真のモデルに含まれる変数の一部を空いて推定する際のモデルに含めないことによるバイアス.
- たとえば，$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ が真のモデル（仮定1-4を満たすとする）なのに，誤って X_2 を含めずに推定して，$\tilde{Y} = \tilde{\beta}_0 + \tilde{\beta}_1 X_1$ という標本回帰関数を求めてしまうということ.
- その場合，$\nu = \beta_2 X_2 + \varepsilon$ として $Y = \beta_0 + \beta_1 X_1 + \nu$ というモデルを推定していることになり，新たな誤差項 η には $\beta_2 X_2$ が含まれる.
- このとき，$\mathrm{Bias}[\tilde{\beta}_1] = \mathbf{E}[\tilde{\beta}_1] - \beta_1 = \beta_2 \tilde{\delta}_1$ であることが示せる. （**重要！**）
- ただし，$\tilde{\delta}_1$ は $X_1 = \delta_0 + \delta_1 X_2 + \eta$ というモデルの傾きの推定量である.

- $\tilde{\delta_1}$ というのは X_1 と X_2 の標本共分散 $\div$ X_2 の標本分散であるから，X_1 と X_2 の標本相関係数がゼロであることが $\tilde{\delta_1} = 0$ の必要十分条件.
- つまり，X_1 と X_2 が相関するとき $\mathbf{E}[\tilde{\beta}_1] \neq \beta_1$ ということになる.
  - X_1 と X_2 が相関するならば，当然，X_1 と $\nu = \beta_2 X_2 + \varepsilon$ も相関するため，仮定4が満たされない.
- 逆に，X_1 と X_2 が無相関ならば $\mathbf{E}[\tilde{\beta}_1] = \beta_1$ ということになり，欠落変数は無害となる（X_1 だけに関心があるならば，X_2 の影響は単なるランダムなノイズとみなせる）.
- バイアス（i.e., $\beta_2 \hat{\delta}$ ）が正であるとき，$\mathbf{E}[\tilde{\beta}_1] > \beta_1$ となるため **上方バイアス**（upward bias）と呼び，負であるとき $\mathbf{E}[\tilde{\beta}_1] < \beta_1$ となるため**下方バイアス**（upward bias）と呼ぶ.
- 欠落変数バイアスの方向は，β_2 と $\mathrm{Corr}(X_1, X_2)$ $(\text{i.e., } \tilde{\delta_1})$ の符号に依存する.

欠落変数バイアスの方向

	$\mathrm{Corr}(X_1, X_2) > 0$	$\mathrm{Corr}(X_1, X_2) < 0$
$\beta_2 > 0$	上方バイアス	下方バイアス
$\beta_2 < 0$	下方バイアス	上方バイアス

- $\mathbf{E}[\tilde{\beta}_1]$ が β_1 よりもゼロに近いとき，**ゼロに向かってバイアスがかかっている**（biased toward zero）と表現する.
- $\beta_1 > 0$ のとき下方バイアスなら，$\beta_1 < 0$ のとき上方バイアスなら，ゼロに向かってバイアスがかかっているということになる.

モデルの誤った特定化

- 欠落変数バイアスはモデルの**誤った特定化**（misspecification）によって生じる.
- 真のモデルが $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ なのに $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ と誤って特定化.
- 逆に，不要な変数を誤って入れてしまう場合はどうか？
  - 例：真のモデルが $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ なのに $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ と誤って特定化.
  - これを**過剰特定化**（overspecification）と呼ぶ.
- 過剰特定化は問題にならない．真のモデルにおいて，$\beta_2 = 0$ であるというだけで，過剰特定化は誤った特定化ではない.
- OLS推定量も不偏性を満たし，$\mathbf{E}[\hat{\beta}_1] = \beta_1$, $\mathbf{E}[\hat{\beta}_2] = 0$ となる.

仮定5. 均一分散

- 誤差項 ε の分散が $X_1, X_2, \ldots, X_k$ の値によらず一定であるという仮定．すなわち，

$$\mathrm{Var}(\varepsilon \mid X_1, X_2, \ldots, X_k) = \sigma^2.$$

- 導出は省略するが，仮定1から仮定5のもとで，サンプルの値に条件づけたとき，すべての $j = 1, 2, \ldots, k$ について

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\mathrm{SST}_j(1 - R_j^2)}$$

となる．

  - ただし，$\mathrm{SST}_j = \sum_{i=1}^{n}(X_{ij} - \bar{X})^2$ は X_j の標本総変動であり，R_j^2 は X_j を他のすべての独立変数に回帰したときの決定係数である．

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\mathrm{SST}_j(1 - R_j^2)}$$

- 推定量の分散は推定精度に相当し，小さければ小さいほど望ましい（推定値の信頼区間が狭まる）．
- **OLS推定量の分散の決定要素**：
  - 誤差項の分散 σ^2. 小さくすると， $\mathrm{Var}(\hat{\beta}_j)$ は小さくなる．モデルが従属変数をよく説明できるほど推定量の分散が小さくなるということ．統制変数を入れることで分散を小さくできる場合があることを示唆している．
  - X_j の標本総変動 SST_j. 大きくすると， $\mathrm{Var}(\hat{\beta}_j)$ は小さくなる． SST_j はサンプルサイズ n について増加するので，サンプルサイズを増やすことで推定量の分散を小さくできることが示唆される．
  - X_j を他の説明変数に回帰したときの決定係数 R_j^2. 共線性に関係する．説明変数どうしの相関が高いほどOLS推定量の分散が大きくなる．

誤差項の分散の推定

- 誤差項の分散 $\sigma^2 = \mathbf{E}[\varepsilon^2]$ はデータからは直接観察できないので，推定する必要がある.
- サンプルの各誤差 ε_i もデータからは直接観察できない.
- **残差**（residual） $\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{1k}$ を使うのが自然.
- 誤差項の分散の推定量として，次が使われる.

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

- 仮定1から5のもとで， $\mathbf{E}[\hat{\sigma}^2] = \sigma^2$ であることが示せる.
- $\hat{\sigma}^2$ を推定量の分散に代入して平方根をとったものを**標準誤差**（standard deviation）と呼ぶ.

$$\mathrm{se}(\hat{\beta}_j) = \hat{\sigma} / [\mathrm{SST}_j (1 - R_j^2)]^{1/2}.$$

OLSの効率性：ガウス・マルコフの定理

- 仮定1から5のもとで，OLS推定量は**最良線形不偏推定量**（best linear unbiased estimator; **BLUE**）である（ガウス・マルコフの定理）.
- 線形推定量とは $\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} Y_i$ として表すことができる推定量のことを指す.
- OLS推定量も実は $\hat{\beta}_j^{OLS} = \left(\sum_{i=1}^{n} r_{ij} Y_i \right) / \left(\sum_{i=1}^{n} r_{ij}^2 \right)$ みたいな感じに書ける（cf. **partialling out**, **Frisch-Waugh定理**）.
- つまり，任意の線形不偏推定量 $\tilde{\beta}_j$ について $\mathrm{Var}(\hat{\beta}_j^{OLS}) \leq \mathrm{Var}(\tilde{\beta}_j)$ が成り立つ.
- OLSは単純な割に結構良い推定量ということ.

追加の仮定6. 正規性

- 誤差項 ε が説明変数 $X_1, \ldots, X_k$ と独立であり，かつ，$\varepsilon \sim N(0, \sigma^2)$ であるという仮定.
- 仮定1から6を**古典的線形モデルの仮定**（classical linear model assumptions）と呼ぶ.
- 古典的線形モデル仮定のもとでは，OLS推定量が**最良不偏推定量**（best unbiased estimator; **BUE**）となる.
- また，このとき,

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_j^0}{\mathrm{se}(\hat{\beta}_j)} \sim t_{n-k-1}$$

  であることが示せる.

- この t **統計量**が自由度 $n-k-1$ の t 分布に従うことを利用して，t 検定を行って推定値の有意性を検定する.

係数の推定値の有意性検定と頑健標準誤差，正規近似

- $\beta_j^0 = 0$ として $t_{\hat{\beta}_j} = \dfrac{\hat{\beta}_j}{\mathrm{se}(\hat{\beta}_j)}$ という検定統計量を使うときの帰無仮説と対立仮説はそれぞれ，

$$H_0 : \beta_j = 0.$$

$$H_1 : \beta_j \neq 0.$$

- t 統計量から p 値を計算して，有意水準（10%, 5%, 1%）を下回れば，帰無仮説を棄却して**統計的に有意にゼロから離れている**と結論する．
- 実際には仮定5（均一分散）が満たされることはあまりないため，標準誤差を修正した**不均一分散に頑健な標準誤差**（heterosekedasticity-robust standard error）を用いる必要がある．
- さらに，実際には仮定6（正規性）も満たされると断言できることは少ない．（t 統計量の分布がわからなくなる．）

- しかし，仮定6が満たされなくても t 統計量が漸近的に正規分布に従うことが知られており，実際上は問題ない.
- 頑健標準誤差については*The Mixtape*でも詳しくやるので今回は詳述しない.

> People will try to scare you by challenging how you constructed your standard errors. (*The Mixtape, p.77)

- Rのビルトイン関数の`lm`はデフォルトで頑健標準誤差を計算してくれないが難点.
- **estimatr**パッケージの`lm_robust`や**sandwich**パッケージの`vocvHC`を使う必要がある.

おまけ

- 最近，Bruce Hansenというすごくえらい経済学者がOLSはガウス・マルコフの仮定のもとで，**最良不偏推定量（BUE）**だと発表した (Hansen, Forthcoming, ECTA).
- しかし，色々と議論がある模様．たとえば，こちらのWP.
- 正直よくわからないので，しばらくは従来の教科書の記述に従おう.

A Modern Gauss-Markov Theorem

Bruce E. Hansen*

University of Wisconsin†

December, 2020
Revised: September 2021

Abstract

This paper presents finite sample efficiency bounds for the core econometric problem of estimation of linear regression coefficients. We show that the classical Gauss-Markov Theorem can be restated omitting the unnatural restriction to linear estimators, without adding any extra conditions. Our results are lower bounds on the variances of unbiased estimators. These lower bounds correspond to the variances of the the least squares estimator and the generalized least squares estimator, depending on the assumption on the error covariances. These results show that we can drop the label "linear estimator" from the pedagogy of the Gauss-Markov Theorem. Instead of referring to these estimators as BLUE, they can legitimately be called BUE (best unbiased estimators).

A Modern Gauss-Markov Theorem? Really?*

Benedikt M. Pötscher and David Preinerstorfer
Department of Statistics, University of Vienna
SEPS-SEW, University of St. Gallen

First version: February 2022
Second version: March 2022

Abstract

We show that the theorems in Hansen (2021a) (the version accepted by Econometrica), except for one, are not new as they coincide with classical theorems like the good old Gauss-Markov or Aitken Theorem, respectively; the exceptional theorem is incorrect. Hansen (2021b) corrects this theorem. As a result, all theorems in the latter version coincide with the above mentioned classical theorems. Furthermore, we also show that the theorems in Hansen (2022) (the version forthcoming in Econometrica) either coincide with the classical theorems just mentioned, or contain extra assumptions that are alien to the Gauss-Markov or Aitken Theorem.

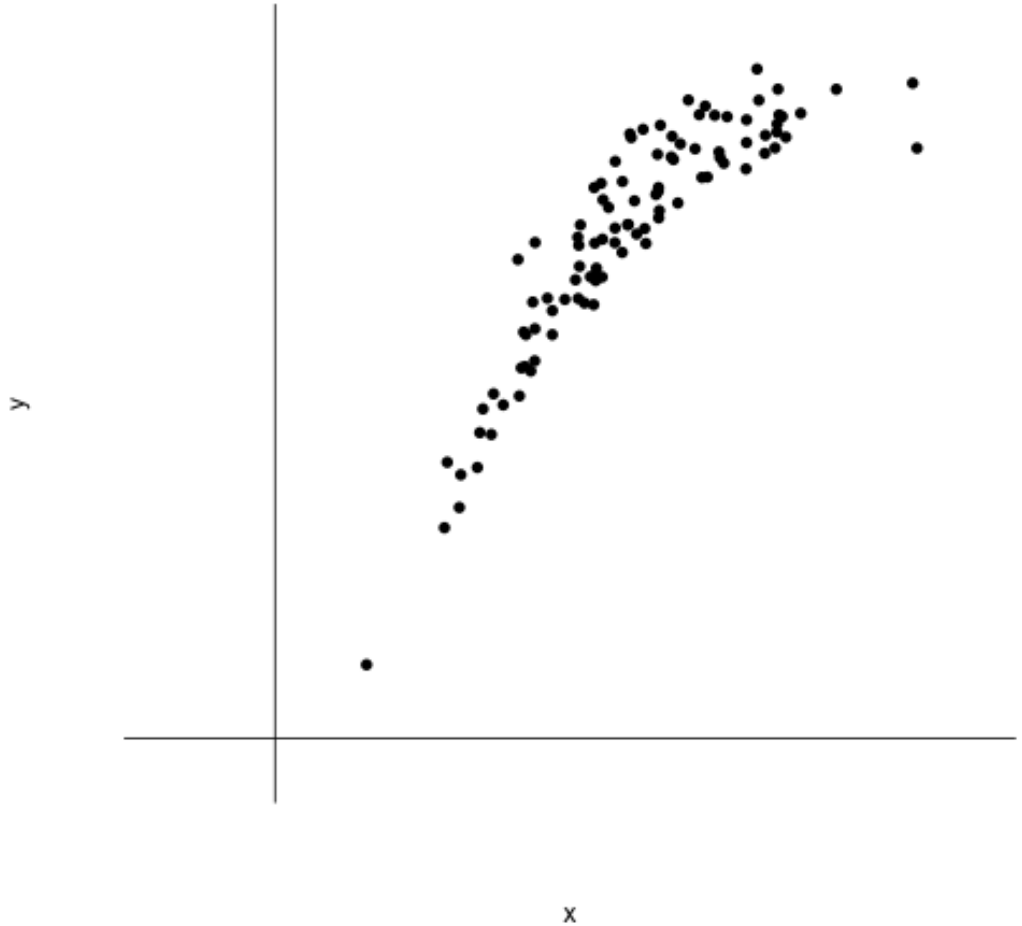
さまざまな関数形

- 線形回帰モデルの"線形"につられて，被説明変数と説明変数の間の直線的な関係しか分析できないという誤解をする人が多い.
- 線形回帰モデルの線形性はパラメータについての要求であって，説明変数に対してではない.
- その実，線形回帰モデルはかなり柔軟な関係を表現できるし，OLS推定量もシンプルな割にそういった関係をちゃんと把握できる.
- よく使われるもの
  - **多項式回帰**
  - **ダミー変数**
  - **交差項**
  - **対数**

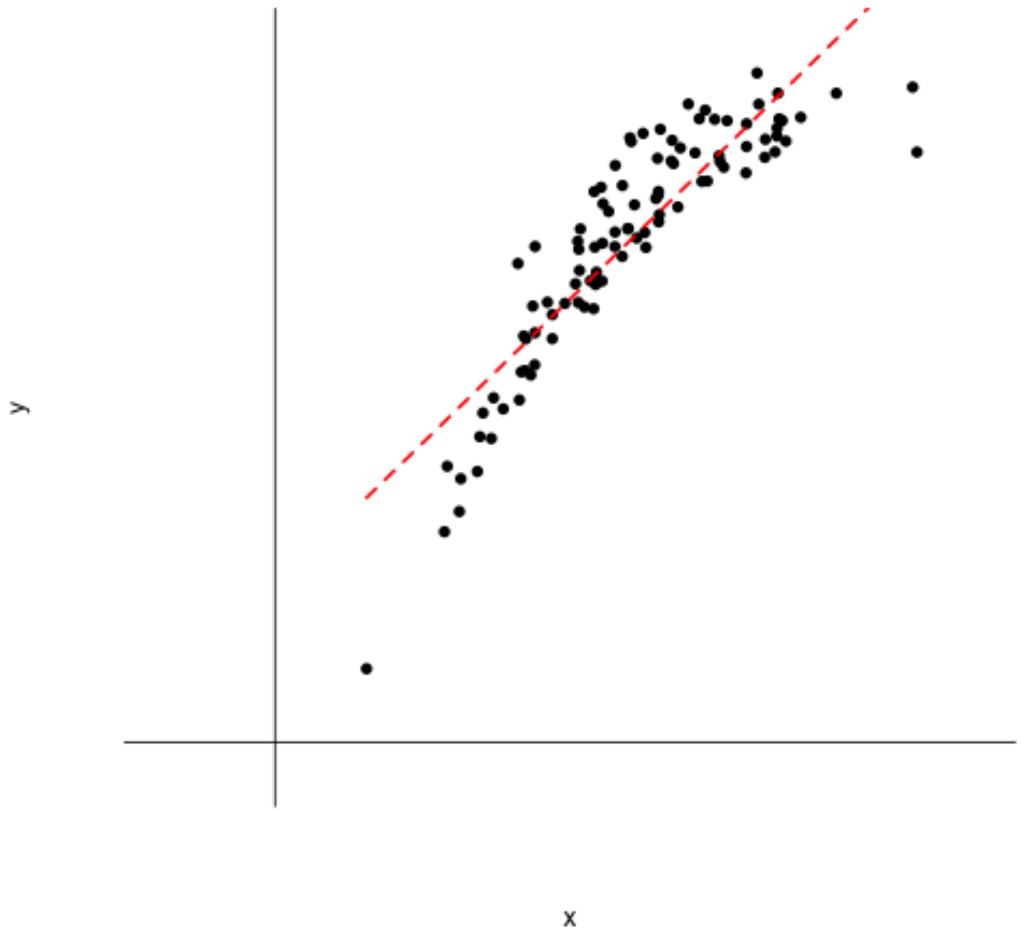
多項式回帰

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \cdots + \beta_p X_1^p + \varepsilon.$$

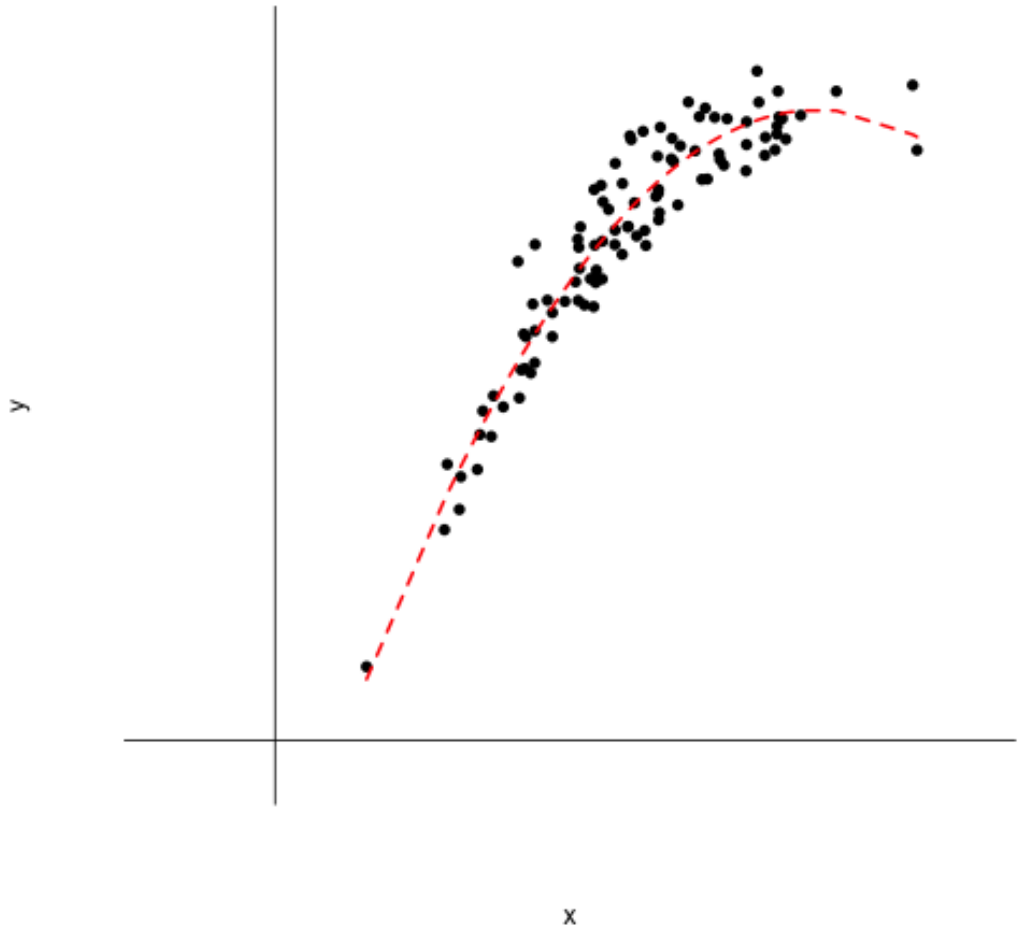
- 上のように説明変数を累乗項を含めて回帰することを**多項式回帰**（polynomial regression）と呼ぶ.
- どれくらい高次の累乗項を入れるかは分析者の判断（入れすぎると**過学習**（overfitting）になって解釈しづらくなる）.
- 限界効果が逓減または逓増すると予想されるときに使われる.
- 例えば，職務経験年数が年収に与える影響は，はじめのうちは新しい経験や知識が増えて賃金が上がっていくと予想されるが，一定のピークを迎えるとその後は新しい経験や知識があまり増えなくなり賃金への影響が小さくなり始めると考えられる.
  - $salary = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \varepsilon.$
  - $\beta_1 > 0, \beta_2 < 0$ と予想される.


y
x

1次：$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$.



2次： $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_1 X_1^2.$



- X_1 の限界効果は，$\beta_1 + \beta_2 X_1$ となり，X_1 の現在の水準に依存している.

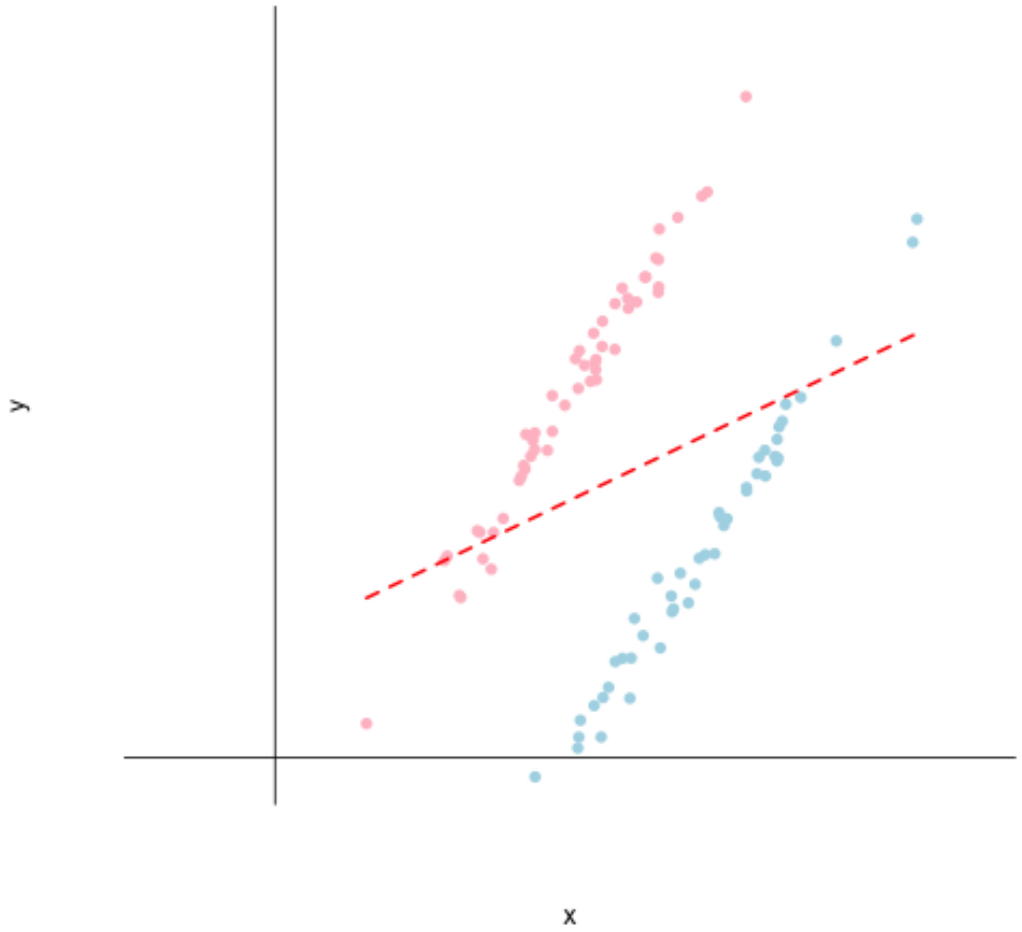
ダミー変数

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i.$$

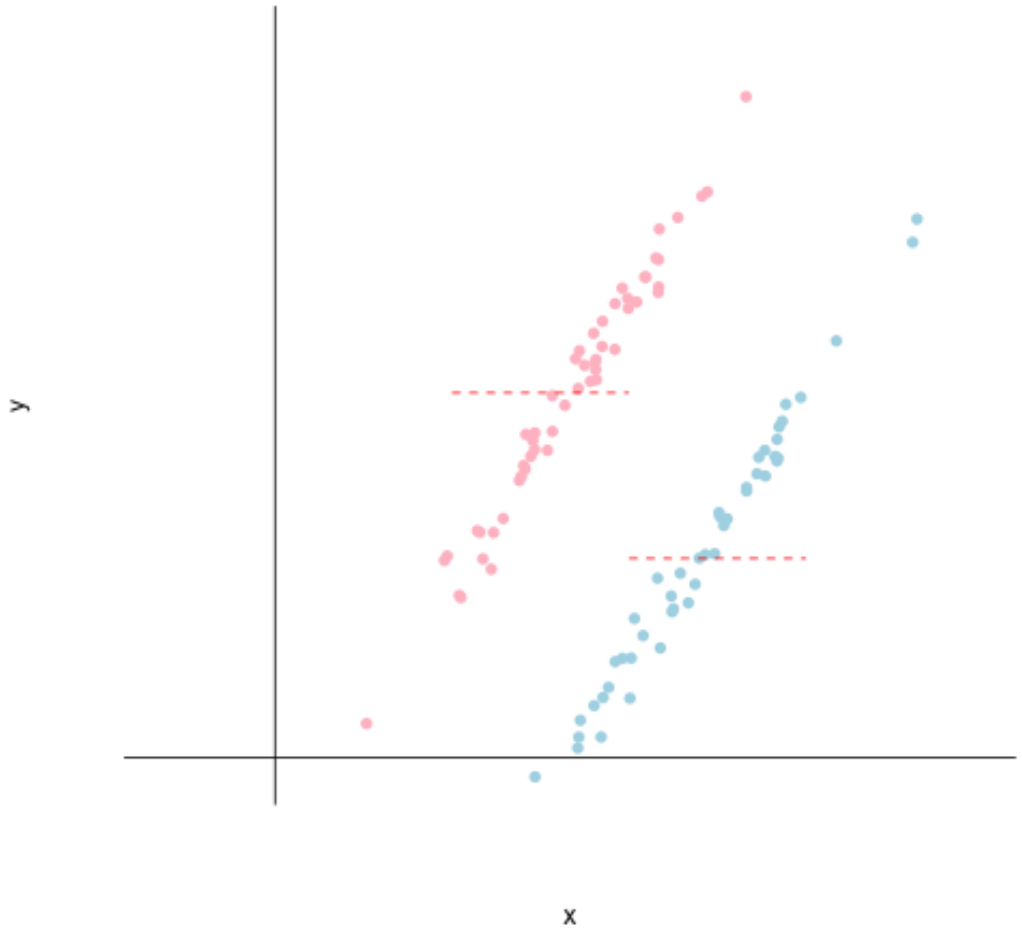
- 個人 i があるカテゴリに含まれるときに $D_i = 1$，そうでないときに $D_i = 0$ となるような変数を**ダミー変数**（dummy variable）と呼ぶ.
- カテゴリによって切片が異なるということを表現できる.
- ダミー変数と普通の変数を組み合わせることもできる.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \varepsilon_i.$$

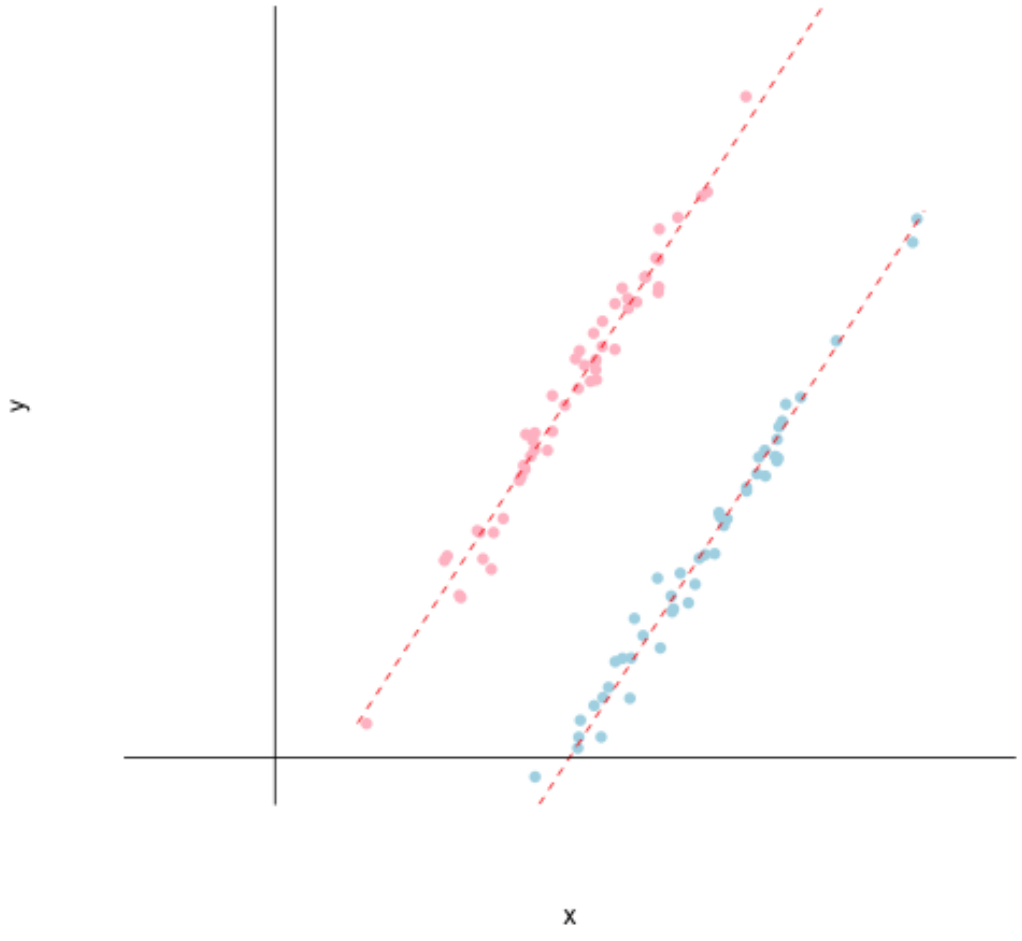
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$



$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i.$$



$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \varepsilon_i.$$



交差項

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon.$$

- 複数の変数を掛け合わせたものを**交差項**（interaction term）と呼び，その係数はそれらの変数の交互作用を捉える.
- X_1 を連続的な変数だとすると：
  - X_2 がダミー変数なら，グループごとの傾きの違いを捉える.
  - X_2 が連続変数なら，値ごとの傾きの違いを捉える.
- X_1 の限界効果は，$\beta_1 + \beta_3 X_2$ となり，X_2 に依存している.
- X_2 の限界効果は，$\beta_2 + \beta_3 X_1$ となり，X_1 に依存している.

対数