回帰分析

労働経済学2

川田恵介

## 回帰分析

* 経済分析においては、複数の変数理解が求められる
  + 典型例: 政策変数(最低賃金) - 評価変数(雇用)
* 一般に非常に複雑な関係があり、適切な”近似”が必要
  + マクロ経済モデルや感染症モデル、交通流動モデル、気候変動モデルなどなど
* **より単純化しているが、”推定”が容易（信頼性の高い・透明性が高い）な方法を紹介**
  + 線形回帰モデル

# データにおける変数間の関係性理解

## 他変数間の関係性理解の基本戦略

* 可視化・記述統計両共に大量の変数間の記述には、適さない
  + 人間は4次元以上を認識できない
* 原則のある単純化を行う

## 線形近似

* ある”目的変数” と”説明変数群” との関係性を近似的に記述する関数
* 以下を最小化するように を決定

## 例

Fit <- lm(Employment ~ MinimumWage,  
 DataReal)  
  
Fit$coefficients

(Intercept) MinimumWage   
-1635.736889 4.162029

* 線形近似モデルにおける変数間の関係性は、 の値が表す
* 定数項 (Intercept)は深刻な意味を持たない

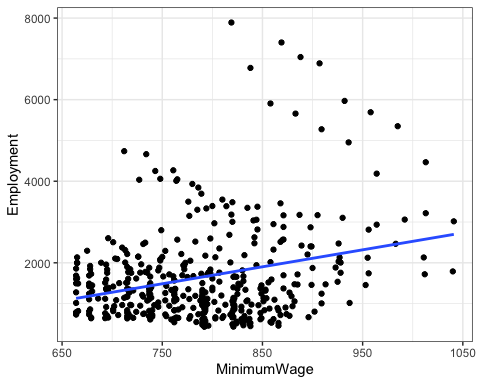
## 例

DataReal %$%  
 lm(Employment ~ MinimumWage,  
 weights = Employment) |>   
 coef()

(Intercept) MinimumWage   
-3564.394935 7.499091

## 例

DataReal |>   
 ggplot(aes(x = MinimumWage,  
 y = Employment)) +  
 geom\_point() +  
 geom\_smooth(method = "lm",  
 se = FALSE) +  
 theme\_bw()



## ダミー変数

* データには、非数値変数も大量に含まれている
  + 都道府県など
* ダミー変数 (0か1をとる変数)に変換し、活用する
  + 例: 東京ダミー (東京であれば1、それ以外であれば0)
* 参照グループも設定する
  + 例: ダミーが全て0であれば、香川県

## 例

Coef <- DataReal %$%  
 lm(Employment ~ MinimumWage + Pref + Year,  
 weights = Employment  
 ) |>   
 coef()  
  
Coef[2]

MinimumWage   
 -12.25381

## 解釈

* 係数値 近似モデルにおいて、ある変数ととの間の関係性
  + が１大きい場合の、 の値

# 母集団における平均的関係の推定

* （データではなく）社会における変数間の理解はより困難なチャレンジ
  + 誤差を考慮する必要がある
* “複雑な”特徴の推定は、より誤差が大きくなる
  + データの特徴に強く依存する

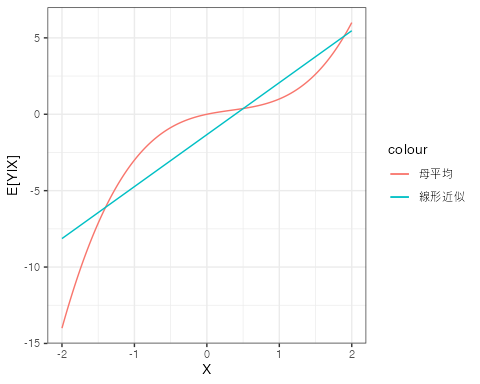
## 条件付き母平均値

* 条件つき母平均 を関心の対象とする。
* 平均はの特徴を全て表すわけではないが、良い出発点
  + 例外: が二つの値のみをとる変数であれば、 は の比率

## 線形近似

* 通常は は極めて複雑な関数であり、推定困難
  + 例外: が少数の組み合わせしか取らないのであれば、サブサンプル平均として推定できる
* 比較的シンプルなモデル、典型的には を推定対象とする。
  + は以下を最小化するように定義する

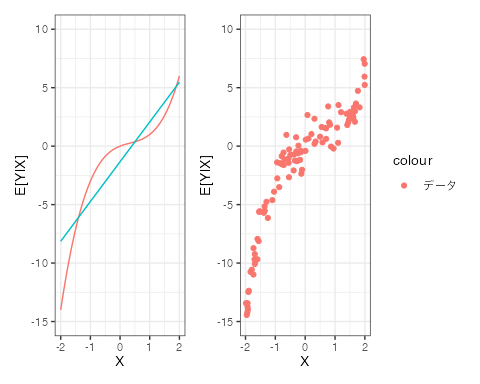
## 数値例



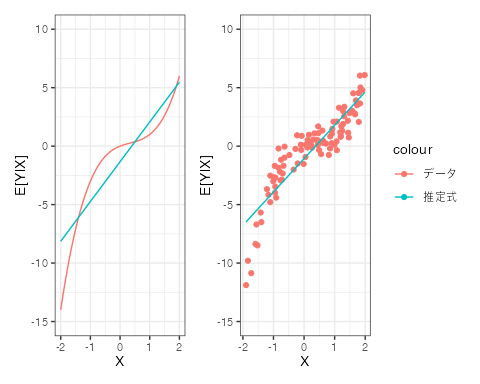
## 推定方法

* 記述統計としての回帰分析と同じ!!!!
* 線形近似モデルとデータとの乖離が最も小さくなるように を推定

## 数値例



## 数値例

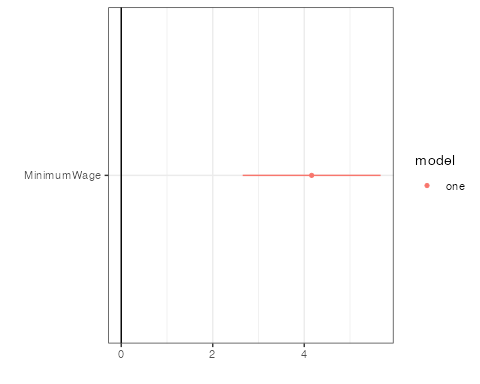


## 信頼区間

* 平均の推定と同様に の推定値もデータに依存する
  + データ 母集団である以上は、本当の線形近似は推定できない
* サンプルサイズがある程度大きければ、平均の推定と同様に、信頼区間を近似計算できる

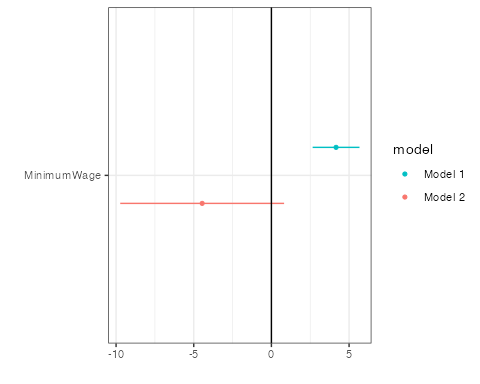
## 例

library(estimatr)  
library(dotwhisker)  
Fit <- lm\_robust(Employment ~ MinimumWage,  
 DataReal)  
dwplot(Fit) +  
 geom\_vline(xintercept = 0) +  
 theme\_bw()



## 例

FitControl <- lm\_robust(Employment ~ MinimumWage + Pref + Year,  
 DataReal)  
dwplot(list(Fit,FitControl),  
 vars\_order = c("MinimumWage"),  
 ) +  
 geom\_vline(xintercept = 0) +  
 theme\_bw()



## まとめ

* 非常に実践的な手法だがいくつか注意が必要
* サンプルサイズ: 平均にしろ条件付き平均にしろ、信頼区間は”近似的”な計算を行なっている
  + サンプルサイズがある程度ないと、近似的計算はできない (教科書的には 150 ~ 200)
  + 変数の数 サンプルサイズ となると推定精度が大幅に悪化
* あくまでも線形近似を推定しているにすぎない
  + 致命的なミスリードを伴う恐れ