政治学 3（計量政治学）

9 回目〜ラスト

秦 正樹

京都府立大学公共政策学部　准教授

hatamasaki@kpu.ac.jp

2022/06/14

火 3 コース@一号館情報処理室

9 回目以降のテーマ

1. 9 回目以降のテーマ
   - 回帰分析の考え方
     - 量的×質的変数による平均値の差についても検定してみる
     - 多変量解析（回帰分析）についてまなぶ
     - coefficient/p-value/confidential-interval について
   - 回帰分析の考え方を理解して実践してみよう！
2. 従業準備
   - Online Rstudio はここからアクセスしてください.
   - Teams にレジュメをあげてありますのでご確認を.
   - 資料と online Rstudio を相互に動かしていくので，プログラムの方のご準備もよろしくお願いします

復習：データの種類

1. 変数と尺度
   - 量的変数（比例尺度）：0 を原点として，間隔と比率に意味があるもの

     e.g. 体重：100kg は，50kg に比べて 2 倍重いといえる
   - 質的変数（名義尺度）：数値とは無関係に，定義的に区別するためにつけられたもの

     e.g. 性別（1. 男性，2. 女性，3. その他）→数字に性別の意味はない
   - 量と質の間：順序尺度：数字の大小には意味があるが間隔が等価でないもの

     e.g. 順位（1 位・2 位・3 位…）．世論調査の多くはこれ？
2. 変数から大枠を把握する方法
   - 量的変数：記述統計を見ながら，大きな傾向を把握

     e.g. 日本人の平均身長，平均年齢（あとでやってみる）と分散など
   - 質的変数：平均値に意味はないので，度数分布を見る

     e.g. あるデータの都道府県平均が 17.2 でした→意味不明…

統計的検定の論理的基盤

- 統計的検定の考え方 1：対立仮説と帰無仮説
  - 検証したい仮説 (対立仮説) →「性別と橋下好感度の間には差があるだろう」説

  → 残念ながら，これを直接に検証する手段がない…そこで逆のパタンを考える!
  - 対立仮説と真逆の仮説 (帰無仮説) →「性別と橋下好感度の間には差がないだろう」

  → 帰無仮説（2 つの変数は関係を持たず独立している場合）を基準に考える!
- 統計的検定の考え方 2：「敵の敵は味方」理論で考える
  - 帰無仮説の方が正しい確率が超高い場合＝対立仮説が間違ってる確率が超高い

  → 「X と Y には差がない」確率が 99%なら「X と Y に差がある」確率は 1%…
  - 帰無仮説の方が正しい確率が超低い場合＝対立仮説が間違ってる確率が超低い

  → 「X と Y には差がない」確率が 1%なら「X と Y に差がある」確率は 99%!

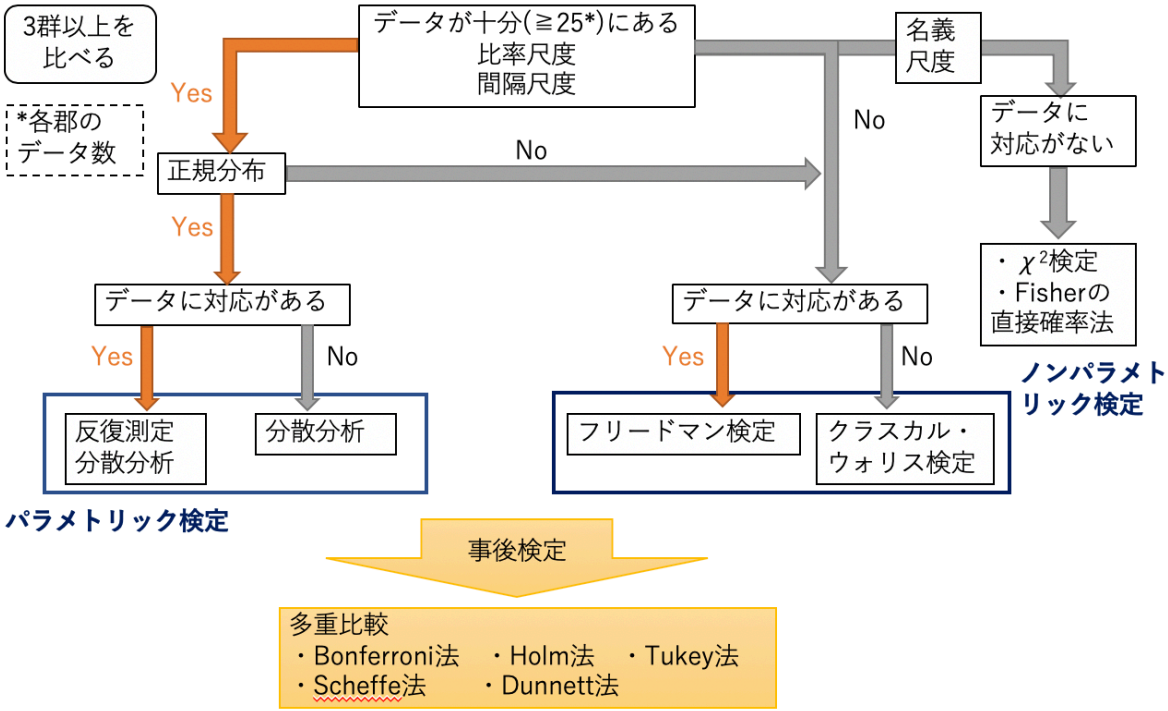
  * 敵（帰無仮説）の方が間違ってる確率が高いやんけ!（ということは，自分の仮説 (対立仮説) が正しい確率が高いってことやろ?）という論理

母比率の差の検定

- 朝日新聞と読売新聞の世論調査結果が違う？
  - 朝日新聞が実施した 2022 年 5 月の世論調査の岸田内閣支持率→ 59%
  - 朝日 2022 年 5 月調査のサンプルサイズ→ N=1432 ；支持者 N=_____
  - 読売新聞が実施した 2022 年 5 月の世論調査の岸田内閣支持率→ 63%
  - 読売 2022 年 5 月調査のサンプルサイズ→ N=1052 ；支持者 N=_____
  - * ＜朝日新聞調査＞はここから，＜読売新聞調査＞はここから見れます
- 対応関係のない群間の比率（割合）の違いを統計的に検定しよう！
  - prop.test(c（var1, var2），c(var1_allsample, var2_allsample))
  - p-value はどうなっているでしょうか？
  - * 朝日新聞の 4 月調査 (N=1365) の内閣支持率は 55%でしたが，5 月は 59%となり，4%支持率が上昇しました．岸田政権の物価上昇の対策などが有権者心理に好印象であったためと考えられます．→ ホンマか？検定してみよう！

平均値の差の検定

- 平均値の検定は「カテゴリカル変数（質）*量的変数」で用いる
- カテゴリカル変数間の「対応関係」にもとづいて適切な検定方法を選ぶ
- 個人的には，2 群間の場合（oneway）は t 検定や Scheffe の検定，3 群以上の場合は Turkey か Bonderroni を使うことが多い



平均値の差の検定をやってみる

- まずは，これまでやってきた TukeyHSD を用いて，政党支持態度 (pid) と自民党／立憲民主党／共産党への感情温度のそれぞれ平均値の差を Tukey で検定をしてみる
- 群間の差・95%信頼区間（上限/下限）の横に p-value があるので，5%水準で統計的有意かどうかを判断　→ p<.05 か否かで判断する
- ＜やってみよう！＞ 3 世代ごとの US/日本/韓国の好感度の平均値の差を Turkey で検定！
- group_by 関数と summarize 関数を使って平均値も見てみよう！
- ggplot を使って，箱ひげ図を出してみよう！

```
> TukeyHSD(aov(自民党~pid_n))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = 自民党 ~ pid_n)

$pid_n
```

	群間の差分 diff	95%信頼区間 下限 lwr	上限 upr	p値 (<.05基準) p adj
independent-government_party	-24.667858	-28.79493	-20.540786	0.0000000
opposit_party-government_party	-31.292415	-35.93463	-26.650202	0.0000000
opposit_party-independent	-6.624557	-11.15383	-2.095278	0.0017913

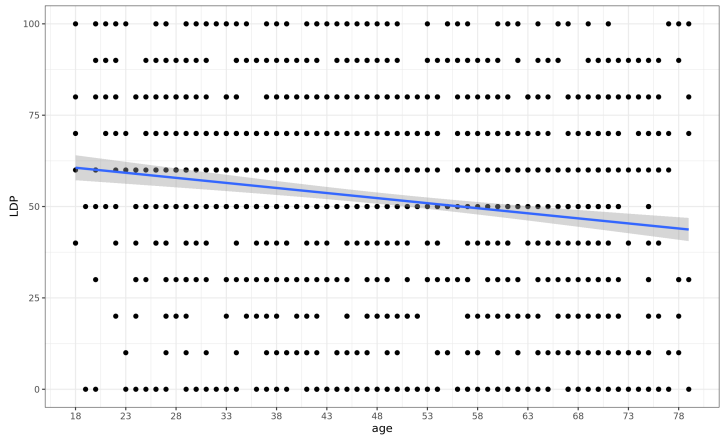
```
> TukeyHSD(aov(立憲民主党~pid_n))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = 立憲民主党 ~ pid_n)

$pid_n
```

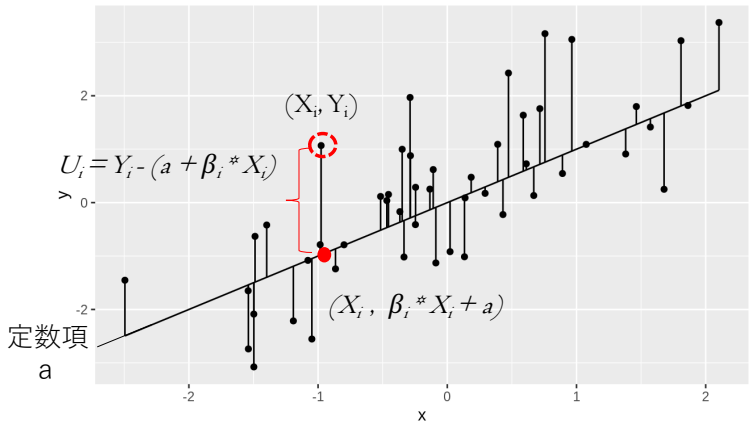
	群間の差分 diff	95%信頼区間 下限 lwr	上限 upr	p値 (<.05基準) p adj
independent-government_party	7.658127	3.970104	11.34615	3.8e-06
opposit_party-government_party	21.400095	17.321312	25.47888	0.0e+00
opposit_party-independent	13.741968	9.757817	17.72612	0.0e+00

注: "e" 以下は 10 の指数を表す (e.g. 3.8e-06=3.8 10^{-6}=3.8*0.000001=0.0000038)

最小二乗法の考え方

- たとえば，自民党の感情温度と年齢の関係を考えてみよう
  - 散布図：年齢と自民党の感情温度の間には関係がある（右肩下がり）ように見える
  - 漸近線：各実測値からのズレが最も小さくなるように線形で予測するもの
  - 回帰式：$y = a + \beta_1 x_1 (+\beta_2 x_2 + \beta_3 x_3 + \cdots + \varepsilon)$（以下は予測値 $\hat{y}$ で表す）
  - 最小二乗法：誤差 $u_i = y_i - (\beta_0 + \beta_1 x_1)$ の総和を最小化する
  - とてもシンプルな式で表せる：$\sum_{k=1}^{n}(e_i^2) = \sum_{k=1}^{n}\{y_i - (a + \beta_1 x_1)\}^2$





最小二乗法の考え方

- 最小二乗法 (Ordinary Least Square:OLS）の推定原理
  - OLS はアウトカムが量的変数の場合のみ（質的変数の場合は別の推定方法）
  - 推定式：$\hat{y} = a + \beta_1 x_1 (+\beta_2 x_2 + \beta_3 x_3 + \cdots)$
  - 有意（significance)：説明変数がアウトカムに影響がある確率（この授業では 95%）
  - 効果（effect)：説明変数が 1 単位変化したときにアウトカムに与えるインパクト
- 統計的推定とこれまでの考え方の違い（秦的に…）
  - 高校：$y = 3x + 4$ の時，$x = 5$ ならば y はいくつ？的な感じでしたよね？
  - 統計的推定：$19 = \beta_1 * 5 + 4$ の時の β_1 をデータを使って逆算して推論する感じ？
    - A さんは：$24 = \beta_1 * 5 + 4$ なので，β_1=4
    - B さん：$36 = \beta_1 * 4 + 4$ なので，β_1=8
    - C さん：$7 = \beta_1 * 1 + 4$ なので，β_1=3
    - β_1 の平均値は 5 なので，$\hat{y} = 4 + 5 * x_1$ である程度，X のあてはまり度を計算可能

最小二乗法の前提

- 最良線形不偏推定量 (Best (Linear) Unbiased Estimaton:B(L)UE)
  1. 不偏性：推定量の期待値が真の値（θ) に等しくなること
  2. 効率性：推定量の分散が最小であること
  3. 一致性：サンプルサイズを大きくするほど，推定量は一定の幅で収束すること
  - * (線形性：アウトカムに対する説明変数は線形（リニア）に変動すること)
- ガウス・マルコフの仮定
  1. 誤差項の分散が均一であること（誤差項の分散均一性の仮定）
  2. 説明変数の間で相関が生じていないこと（共分散なしの仮定）
  3. 誤差項と説明変数の間に相関が生じていないこと（誤差項の独立性）

OLS 仮定を満たさないとき

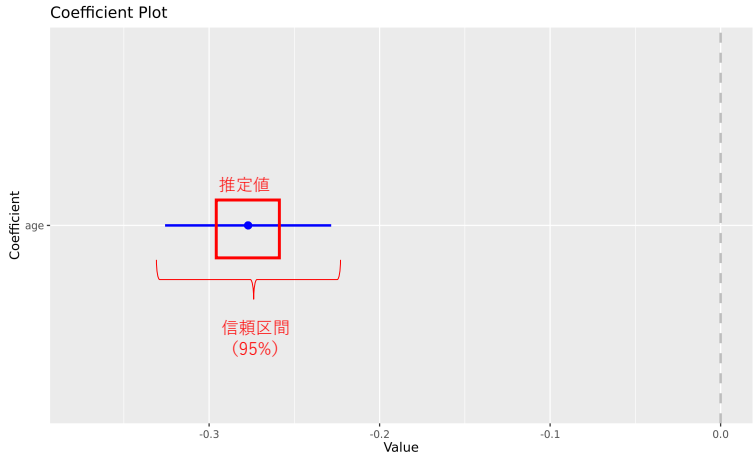
- これらの仮定が満たれないとき…
  - 誤差項と相関する説明変数（内生変数）の場合，OLS 推定量は一致性を持たない
  - 推定量の期待値が真の値（θ) が等しくならないときのバイアスを考える必要がある
    - ✓ 欠落変数バイアス (omitted variable bias)
    - ✓ 測定誤差（measurement error）
    - ✓ 同時性の問題（simultaneity）

OLS 推定の実践

- R 上では lm 関数を使えば簡単に推定できる
  - > result <- lm(outcome ~exp_var1 , data = df)
  - > summary(result)
- 自民党感情温度に対する年齢の効果を推定する単回帰分析
  - > result1 <- lm(LDP ~age , data = data) → summary(result1)
  - > coefplot(result1, intercept = FALSE)

```
> #自民党の感情温度に与える年齢の効果を検証
> r1 <- lm(LDP ~ age, data = data2020)
> summary(r1)

Call:
lm(formula = LDP ~ age, data = data2020)

Residuals:
    Min      1Q  Median      3Q     Max
-60.346 -20.623   2.829  22.142  56.281

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.61086    2.55657  25.664  < 2e-16 ***
age         -0.27711    0.04871  -5.689 1.57e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.5 on 1332 degrees of freedom
  (175 observations deleted due to missingness)
Multiple R-squared:  0.02372,   Adjusted R-squared:  0.02298
F-statistic: 32.36 on 1 and 1332 DF,  p-value: 1.573e-08
```

推定値　誤差　p値（<.05基準）

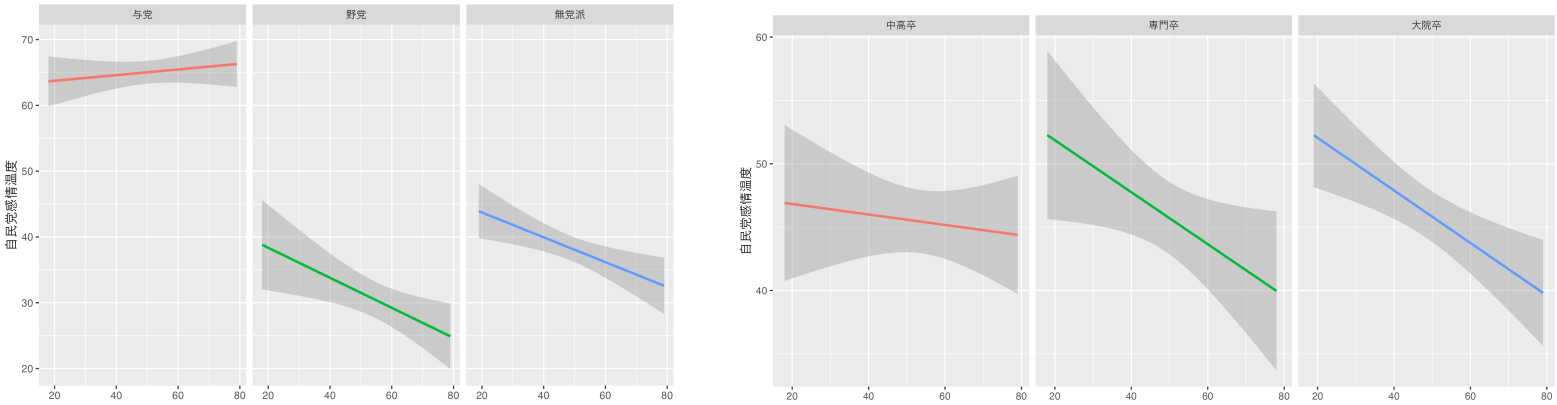


因果関係に迫るために

- 「ターゲットとする説明変数→アウトカム」の因果効果を推定するために
  - 「ターゲットとする説明変数（キー変数と呼ぶ）→アウトカム」の因果推定において，一変数だけでアウトカム変数の分散を説明することは実質的に不可能
  - 疑似相関：第三変数（交絡変数）による効果が働いている可能性を無視
  - アウトカムに影響を与える要因は，キー変数以外にもあることを無視（共変量）

  e.g. アイスの売上→溺死者数の関係を考えてみて

→ 他の条件がまったく一緒であったときのキー変数の変動が重要



* 支持政党／教育程度と自民党感情温度との関連

潜在的結果にもとづく因果推論

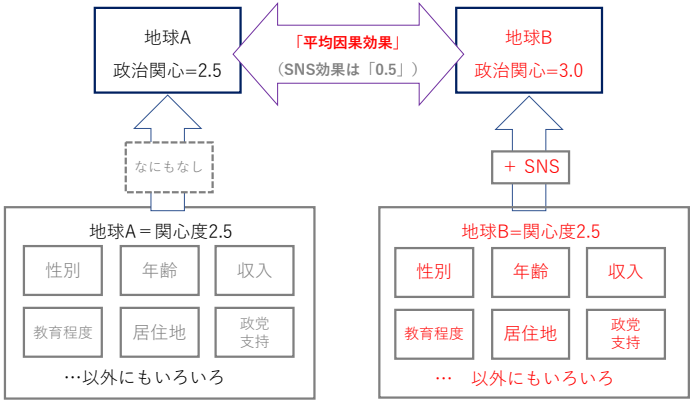
- 潜在的結果 (potential outcome) にもとづく因果推論
  - 反実仮想 (counterfactual)：「もし～だったら○○になるだろう」

    e.g. 「昨晩飲みすぎて今朝から頭痛が…」⇔「もし昨晩飲まなければ頭痛はない」
  - 時間軸は一つなので結果は「一つ」しかない⇔集団単位で「反実仮想」する

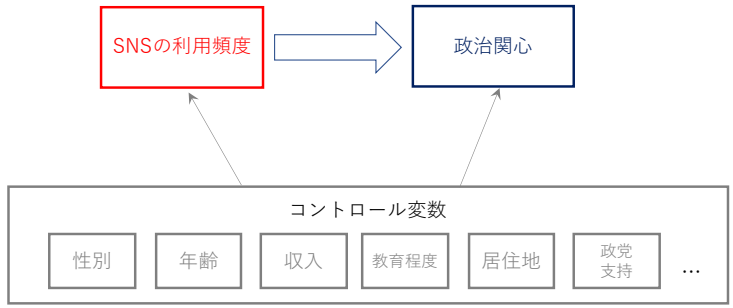
	共変量＝コントロール変数				キー変数	アウトカム
	性別	年齢	居住地	職業	SNS利用	政治関心
Aさん	男	54	東京	公務員	利用	あり
もう一つの世界のAさん	男	54	東京	公務員	未利用	？？
Bさん	女	32	大阪	OL	利用	あり
もう一つの世界のBさん	女	32	大阪	OL	未利用	？？
Cさん	女	87	広島	無職	未利用	なし
もう一つの世界のCさん	女	87	広島	無職	利用	？？
Dさん	女	51	京都	主婦	利用	あり
もう一つの世界のDさん	女	51	京都	主婦	未利用	？？
Eさん	男	21	沖縄	学生	未利用	なし
もう一つの世界のEさん	男	21	沖縄	学生	利用	？？
Fさん	男	35	宮崎	派遣社員	利用	あり
もう一つの世界のFさん	男	35	宮崎	派遣社員	未利用	？？
Gさん	女	19	京都	学生	未利用	なし
もう一つの世界のGさん	女	19	京都	学生	利用	？？
Hさん	男	49	島根	正規職	利用	あり
もう一つの世界のHさん	男	49	島根	正規職	未利用	？？
Iさん	男	69	福岡	農家	未利用	なし
Iもう一つの世界のさん	男	69	福岡	農家	利用	？？

理想的な因果推論の世界

- 統計的因果推論の考え方
  - 理想的な「条件統制」：ランダム化比較試験（RCTs）による実験
  - 調査観察データ（observational data）の限界；アウトカムと説明変数は同時に回答する（政治関心と SNS 利用は同時にしか観察できない）＝「因果推論の根本問題」
  - 回帰分析はあくまで「相関」であり「因果」とするためには強い理論が必要
  - コントロール変数（統制変数）：アウトカム／説明変数のいずれにも影響を与えると考えられる要因を回帰モデルに変数として投入→できるだけ同じ条件に…



* 理想的な因果推論モデル (ランダム化比較試験)



* 観察データを用いた擬似的な因果推論モデル

重回帰分析の実践

- 先ほどの単回帰モデルに ”+” で変数を加えるだけ

```
> result <- lm(outcome ~exp_var1 +control_var2+control_var3···) , data = df)
```

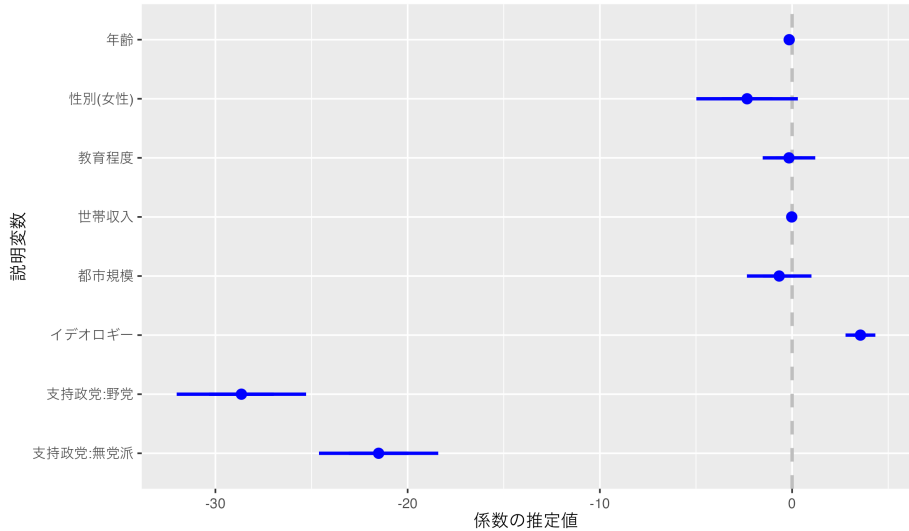
- 自民党感情温度に対する年齢の効果を推定する重回帰分析
  - 統制変数としてよく利用する変数：社会経済的要因（socio-economic factor)
  - * 性別 (gender)/収入 (income)/教育程度 (edu.n)/都市規模 (citysize.n)/政治的イデオロギー (ideology)/政党支持（pid) を投入

```
> result2 <- lm(LDP ~age + gender + edu.n + income + citysize.n + ideology +
  pid, data = data2020)
> summary(result2)
> coefplot(result2, intercept = FALSE)
```

重回帰分析の結果を読む

- 分析結果の解釈
  - 推定結果の見方は単回帰 (pp.11) と同じ（coefplot も同様）
  - $\hat{Y}_{ldp} = -0.15 * age - 2.34 * gender - 0.16 * edu.n - 0.02 * income - 0.67 * citysize.n + 3.55 * ideology - 28.65 * (pid/mutoha) - 25.51 * (pid/oppo) + 53.83$
  - 年齢：p=0.00015 $<$.05 → 5%(0.1%でも) 水準で統計的に有意な効果あり

```
Call:
lm(formula = 自民 ~ age + gender + edu.n + income + citysize_n +
    ideology + pid, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-69.200 -11.305   2.159  11.582  62.460

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  53.82908    4.45353  12.087  < 2e-16 ***
age          -0.15182    0.03989  -3.806  0.00015 ***
gender女性   -2.34150    1.31979  -1.774  0.07637 .
edu.n        -0.16275    0.68355  -0.238  0.81186
income       -0.01972    0.12022  -0.164  0.86976
citysize_n   -0.67228    0.83944  -0.801  0.42342
ideology      3.55493    0.38771   9.169  < 2e-16 ***
pid野党     -28.64808    1.68283 -17.024  < 2e-16 ***
pid無党派   -21.50998    1.55108 -13.868  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.15 on 924 degrees of freedom
  (576 observations deleted due to missingness)
Multiple R-squared:  0.4166,    Adjusted R-squared:  0.4115
F-statistic: 82.47 on 8 and 924 DF,  p-value: < 2.2e-16
```

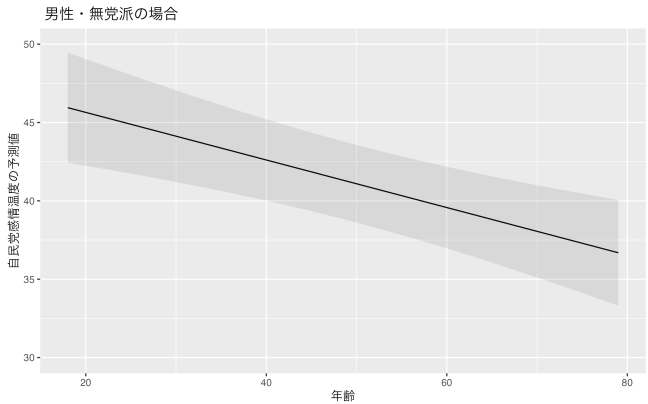


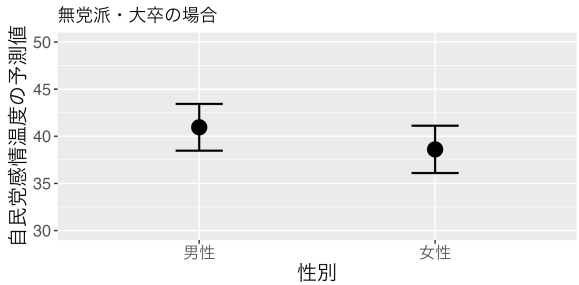
* 重回帰分析の結果（そのまま）

* coefplot を用いた分析結果の可視化

推定結果を使った予測と可視化

- 推定結果による予測（post-estimate simulation)
  - $\hat{Y}_{ldp} = -0.15 * age[18-79] - 2.34 * gender[1] - 0.16 * edu[3.168] - 0.02 * income[6.775] - 0.67 * citysize\_n[2.391] + 3.55 * ideology[5.23] - 28.65 * (pid/mutoha)[0] - 25.51 * (pid/oppo)[1] + 53.83$
  - この推定式から，キー変数以外に観察された平均値（質的変数は任意の値）を代入
  - ここでいえば，age 以外の変数に平均値を代入して，age を 18〜79 歳まで動かしたとき，自民党感情温度がどの程度変動するか予測できる＝効果量 (effect size)





* 年齢（量的変数）が自民党感情温度に与える効果

* 性別（質的変数）の違いが自民党感情温度に与える効果