構造研究: バランス後の比較

Data visualization

川田恵介
東京大学
keisukekawata@iss.u-tokyo.ac.jp

2025-08-03

1 社会の仕組み

1.1 記述/予測研究

- ここまで: 母分布を近似するモデルを推定する方法を学習
  - OLS/LASSO/回帰木: 母平均を近似
  - 最尤法: 母分布
- 記述モデルとも呼ばれる

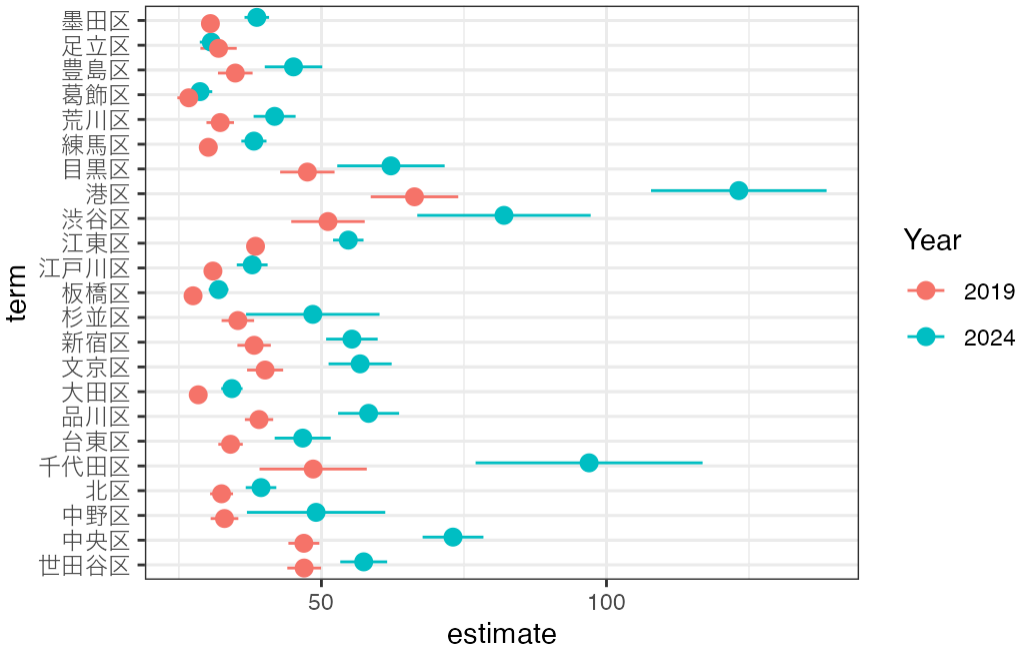
1.2 構造研究

- 研究目標 = 社会の(表面的)記述ではなく、その背後にある構造 (仕組み、理由)
- 動機
  - 仕組みによって、事象の評価が異なる
  - 仕組みを理解しないと、適切な介入が議論できない
- 一般に極めて困難な課題であり、研究者や報道機関等による”安易な”説明には注意

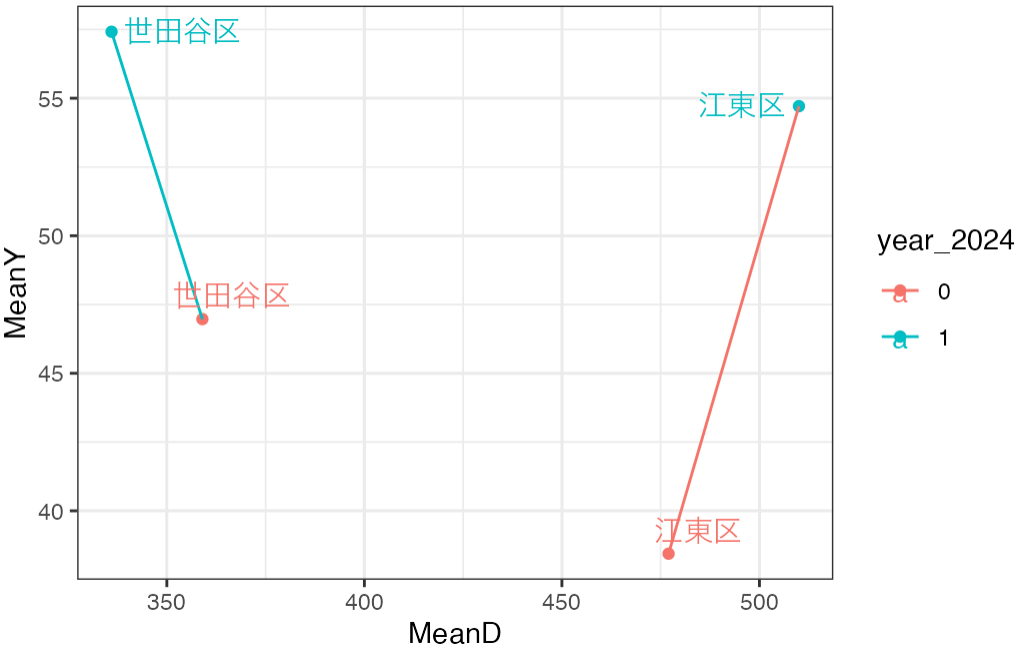
1.3 経済理論/因果推論

- 仕組みを議論する枠組みの一つ

1.4 例: 平均取引価格の記述



1.5 例: 取引価格と数量の記述



1.6 入門書的価格理論

- 需要と供給

- 江東区: 需要増加が支配的
- 世田谷区: 供給現象が支配的

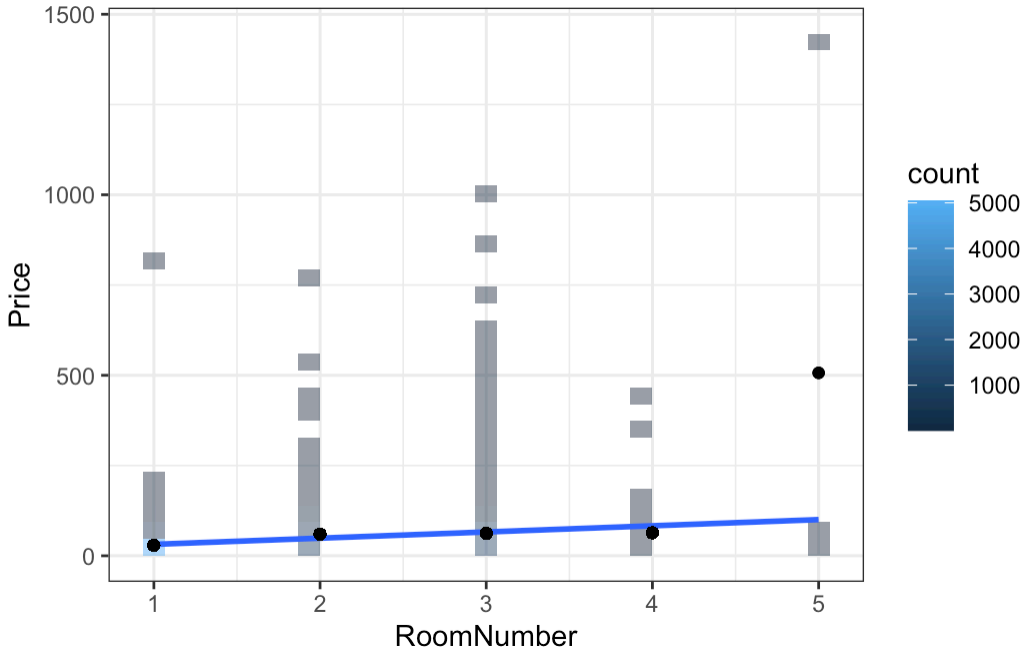
- “オルタナティブ”な説明は、無数に存在しうる
  - 例: 「日本経済の陰の支配者がそのように命じた」

2 思考実験

2.1 アプローチ

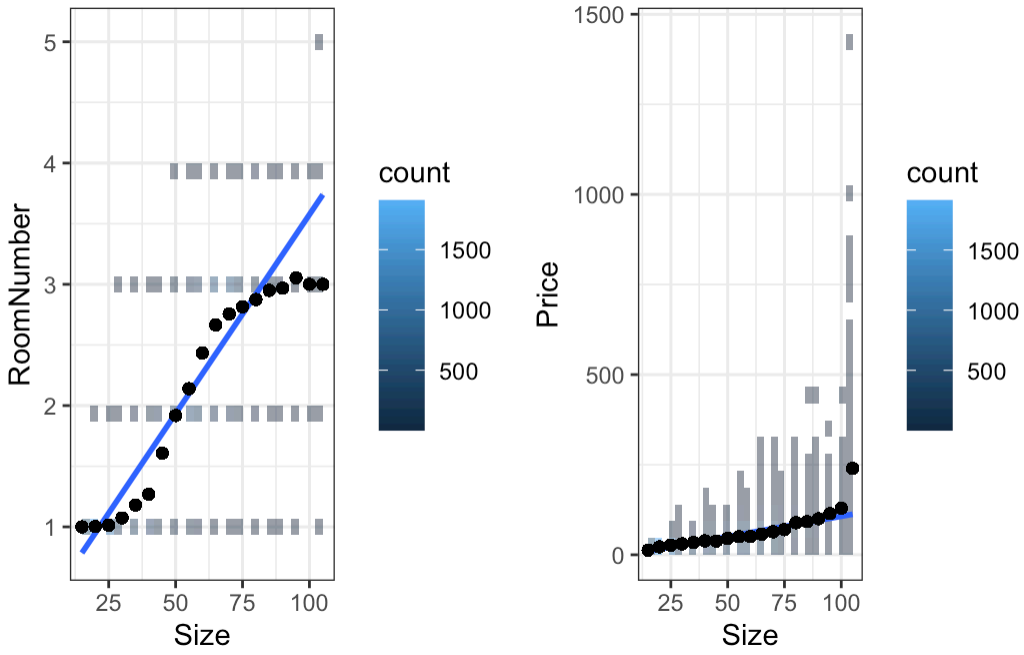
- 社会の仕組みを論じる有力な手段は、思考実験
- What if 分析はその一つ
  - 社会の仕組みを理解する上で、極めて重要
- 例: “もし部屋の数が増えるように設計すれば、市場価格は上昇するか?”

2.2 例: 部屋の数 (D) と取引価格 (Y)



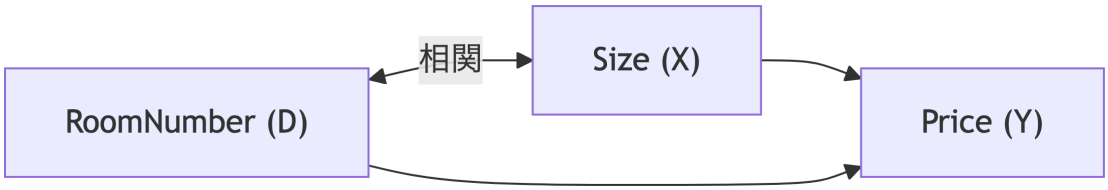
- 部屋数を多く設計すれば、市場価格が上がる?

2.3 例: 部屋の広さ (X)



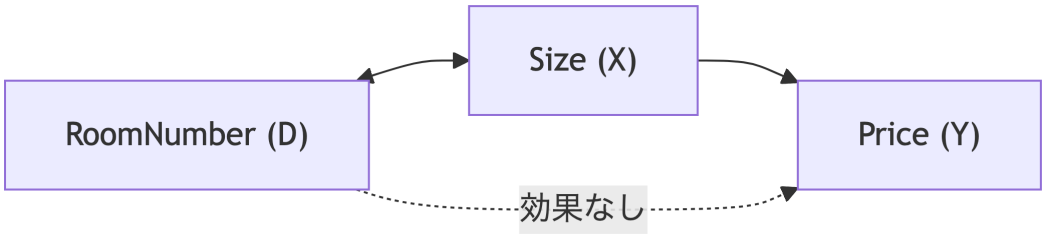
2.4 例: 可能な説明

- 矢印(実線) = “因果的影響” (後述)

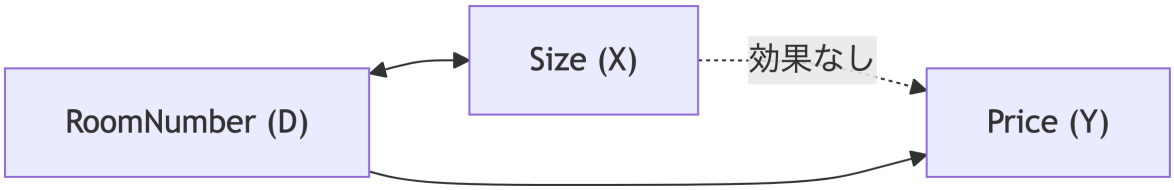


2.5 例: 可能な説明

- 矢印(点線) = “因果的影響なし”



2.6 例: 可能な説明



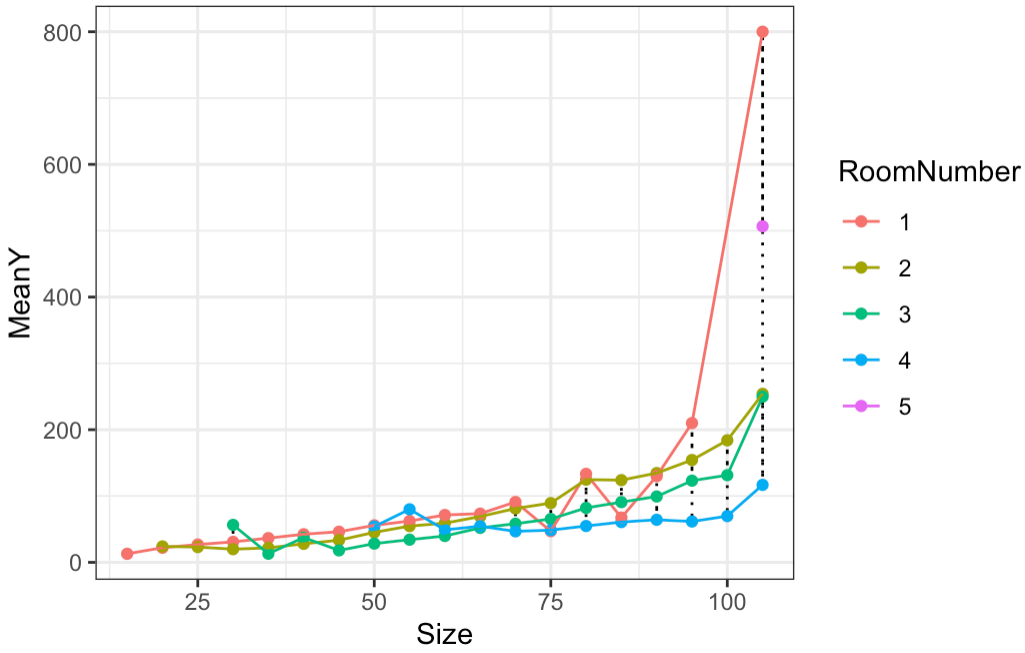
2.7 有益な推定対象

- 推定対象: 同じ部屋の広さの物件で、部屋の数ごとに平均取引価格を比較すれば、どうなるか?
  - 「バランス後の比較」とも呼ばれる

2.8 平均値による比較

- 部屋の広さ $(= X)$ と 部屋の数 $(= D)$ ごとに、取引価格 $(= Y)$ の平均値を計算し、比較する

2.9 例: 平均値による比較



2.10 実践での応用例

- 既存店前年比
  - $X =$ 既存店に絞って、 Yを比較する
- X の分布を”均一”する調整も可能
  - 合計特殊出生率

  - 女性の年齢比率が均質になるように調整し、一年間に生まれた子供の数を算出

3 モデルの活用

3.1 平均値による比較の問題点

- X の数が増えると、
  - サブサンプル数が極端に少なくなり、推定の信頼性が低下する
    - 信頼区間の計算すら難しくなる
  - 大量の平均値を特徴を理解する必要があり、難しい

3.2 モデルの利用

- モデルは、予測モデルとしてだけではなく、What if 分析の補助にも活用できる
  - 経済学における伝統的な活用方法
- 経済モデルの活用例:
  - もし生産性が上昇すれば、何が起きるか?
  - もし最低賃金を引き上げれば、何が起きるか?

3.3 政府支出の What if

- もし政府支出 G を増やせば、民間消費 C はどのように変化するか?
- 単純な IS モデル:

$$\underbrace{Y^S}_{\text{総供給}} = \underbrace{Y}_{\text{総収入}} = \underbrace{C+G}_{\text{総支出}}$$

$$C = \underbrace{c}_{\text{パラメタ}} \times Y$$

- G をモデル上で増やすと、C は増加する

3.4 仮定の検討

- 先の議論は、需要不足に直面しており、総支出が増えれば、Y^S が増えることが前提
- 供給不足 (Y^S が一定)であれば、G の増加はCを減少させる
  - 限られた生産物を、民間/政府部門が”奪い合っている”
- 現状が供給不足か需要不足かで、What if の結論は大きく異なる

3.5 数理モデルの利点

- 前提が正しければ、(証明ミスがない限り)、What if への回答は必ず正しい
  - 前提の妥当性に議論を集中できる

- データを活用した What if 分析を可能にする

4 線型モデルのバランス後の比較分析への活用

4.1 例

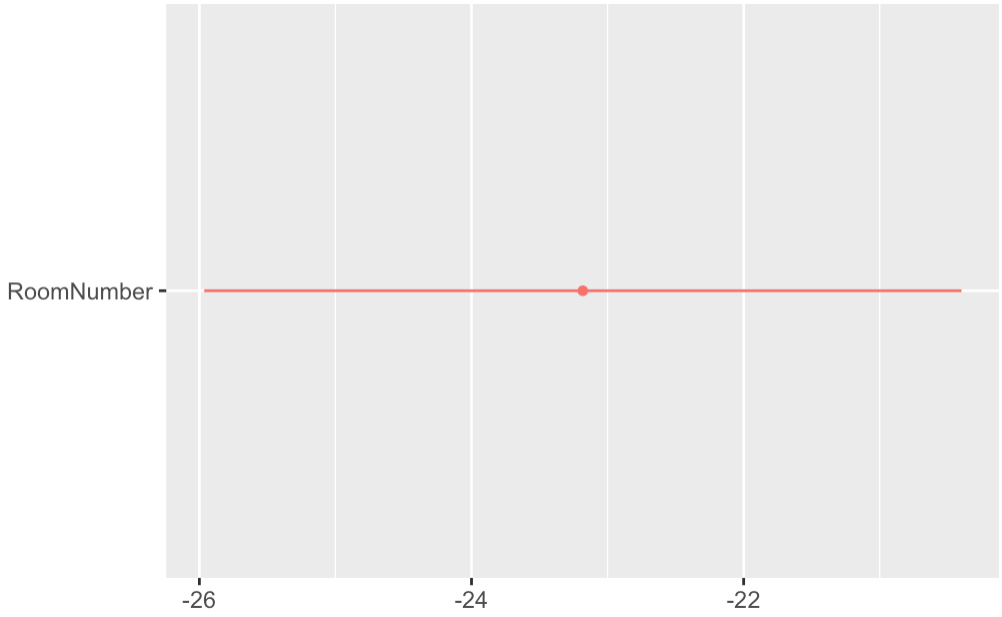
- 以下の式で、平均取引価格は**正確に**捉えられるとする

$$\text{平均取引価格} = \underbrace{\beta_D \times \text{部屋の数}}_{\text{関心}} + \underbrace{\beta_0 + \beta_1 \times \text{部屋の広さ}}_{\text{局外}(Nuisance)}$$

- β_D = 部屋の広さを”一定のまま固定し”、部屋の数についてのみ比較した結果
  - OLS で信頼区間付きで推定できる

4.2 例

```
model <- estimatr::lm_robust(Price ~ RoomNumber + Size, data)

dotwhisker::dwplot(
  list(model),
  vars_order = c("RoomNumber")
  )
```



4.3 複数変数のバランス

- X が複数の変数であったとしても、活用できる

- $X = [Size, District]$ をバランスさせたいのであれば、以下を OLS 推定

$$\text{平均取引価格} = \beta_0 + \beta_D \times \text{部屋の数} + \beta_1 \times Size$$

$$+\beta_2 \times \text{中央区} + ...$$

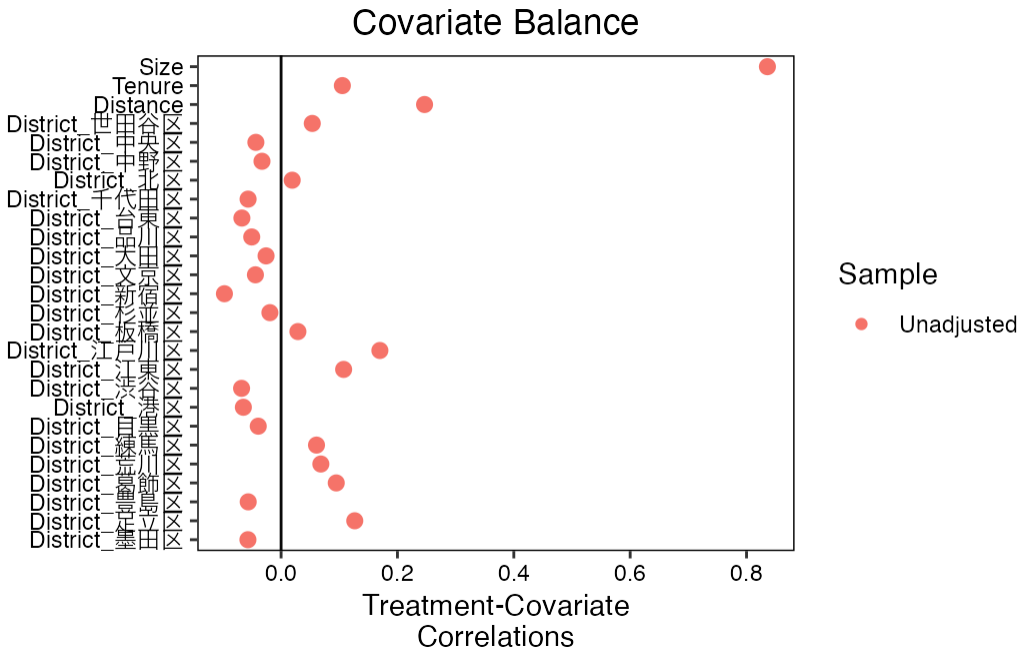
4.4 例: cobalt を用いたバランスの確認

```
Table <- cobalt::bal.tab(
  RoomNumber ~ Size + Tenure + Distance + District,
  data
)

Table
```

```
Balance Measures
                      Type Corr.Un
Size                Contin.  0.8359
Tenure              Contin.  0.1053
Distance            Contin.  0.2466
District_世田谷区   Binary  0.0536
District_中央区     Binary -0.0434
District_中野区     Binary -0.0329
District_北区       Binary  0.0191
District_千代田区   Binary -0.0567
District_台東区     Binary -0.0674
District_品川区     Binary -0.0505
District_大田区     Binary -0.0258
District_文京区     Binary -0.0441
District_新宿区     Binary -0.0974
District_杉並区     Binary -0.0192
District_板橋区     Binary  0.0289
District_江戸川区   Binary  0.1698
District_江東区     Binary  0.1075
District_渋谷区     Binary -0.0679
District_港区       Binary -0.0648
District_目黒区     Binary -0.0393
District_練馬区     Binary  0.0609
District_荒川区     Binary  0.0683
District_葛飾区     Binary  0.0948
District_豊島区     Binary -0.0565
District_足立区     Binary  0.1266
District_墨田区     Binary -0.0570

Sample sizes
    Total
All 11311
```

4.5 例

```
cobalt::love.plot(Table)
```



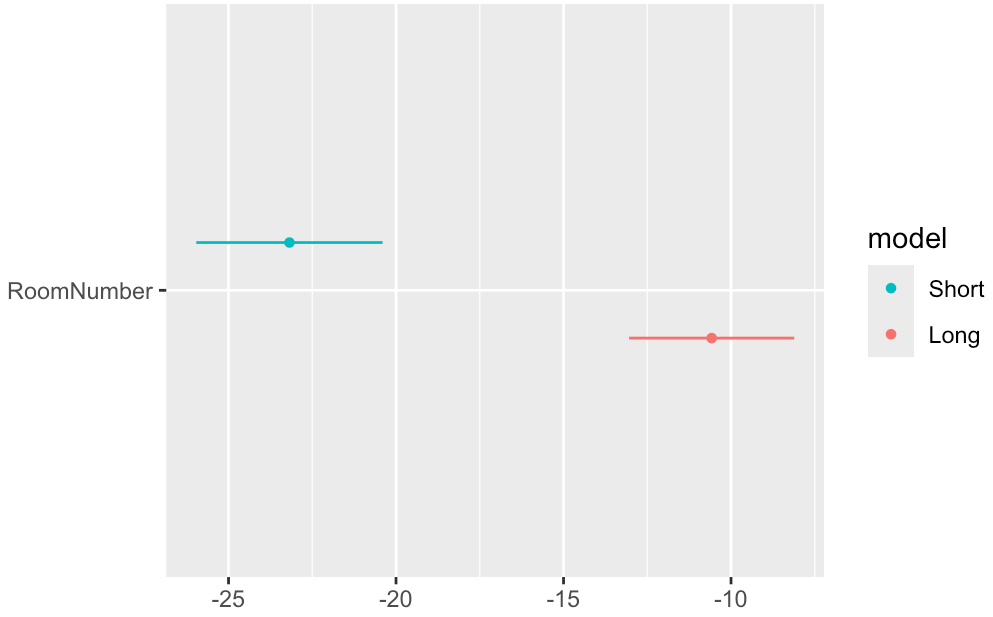
4.6 例

- 重回帰によるバランス

```
model_long <- estimatr::lm_robust(
  Price ~ RoomNumber + Size + District +Tenure + Distance, data)
```

4.7 例

```
dotwhisker::dwplot(
  list(
    Short = model,
    Long = model_long),
  vars_order = c("RoomNumber"))
```



4.8 単純なモデルの限界

- 誤定式化が存在すると、“X を固定した”比較結果とは解釈できない
- 例: 実際の関係性は、

$$\text{平均取引価格} = 5 \times \text{部屋の数} + 10 \times \text{部屋の広さ}^2$$

- $$\text{平均取引価格} = \beta_0 + \beta_D \text{部屋の数} + \beta_1 \text{部屋の広さ}$$

  を推定し得られるβ_D の信頼区間は、95 % よりも低い確率でしか、真の値 5 を含まない

4.9 仮定の緩和

- X に関する部分を複雑化すれば、誤定式化の影響を減らせる

$$\begin{aligned}\text{平均取引価格} = \beta_0 &+ \beta_D \text{部屋の数} \\ &+ \beta_1 \text{部屋の広さ} + \beta_2 \text{駅からの距離} \\ &+ \beta_3 \text{部屋の広さ}^2 \\ &+ \beta_4 \text{部屋の広さ} \times \text{駅からの距離}\end{aligned}$$

4.10 実践への推奨

- X について、以下を導入
  - 連続変数については二乗項

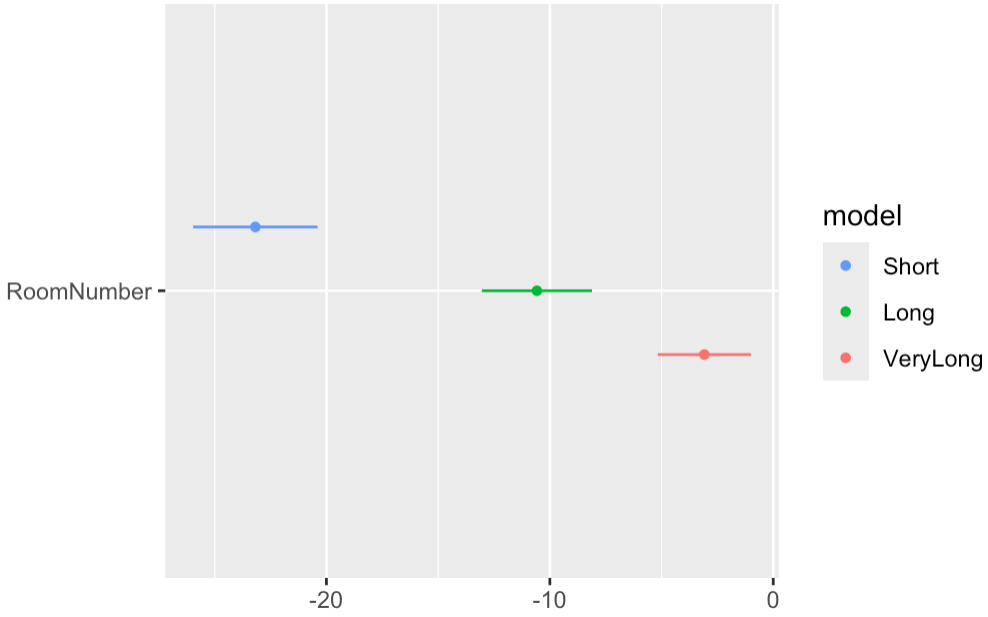
- 可能な限り全ての変数の交差項

4.11 例

```
model_very_long <- estimatr::lm_robust(
  Price ~ RoomNumber + (Size + District +Tenure + Distance)^2 +
    I(Size^2) + I(Tenure^2) + I(Distance^2), data)
```

4.12 例

```
dotwhisker::dwplot(
  list(
    Short = model,
    Long = model_long,
    VeryLong = model_very_long),
  vars_order = c("RoomNumber"))
```



5 LASSO の活用

5.1 OLS の限界

- X の数が多いと、十分に複雑化することが困難になる
  - ざっくり、モデルが含むパラメタ β の数が、事例数の 1/3 を超えると、推定精度が急速に悪化する

5.2 アイディア

- 全てのX (含む高次項/交差項)が”重要なわけではない”
  - 主たる関心である Y/D と、関係がほとんどない変数が含まれているかもしれない
- LASSO などによって重要な変数を選び、OLS などの伝統的な方法で最終的な推定を行えば良いのでは?
  - 変数選択などについても、サンプリング誤差が生じるので、適切な対処が必要

5.3 Single-selection OLS

0. 元々の nuisance variables に二乗項などを加えて、Xを作成する

1. $Y \sim X$ を LASSO で推定し、重要ではない変数を除外

2. 除外されなかったX $(= Z)$とDのみを用いて、重回帰 $Y \sim D + Z$を行う

- **除外される変数**も、データに依存してしまい、信頼区間の近似計算ができない

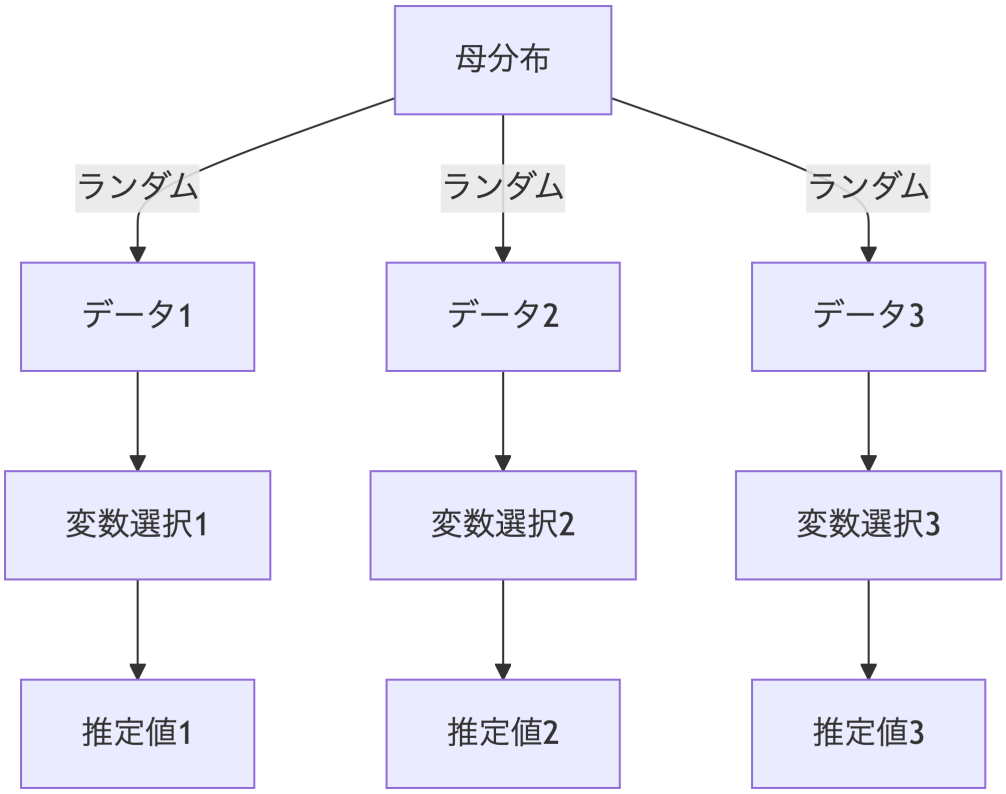
5.4 推定値の分布

- 事例数が十分に大きくなると、β_Dの推定値の分布 $=$

$$\underbrace{\underbrace{Step1 : \textbf{変数選択の誤差}}_{?} + \underbrace{Step2 : OLS\textbf{の誤差}}_{\Rightarrow \textbf{正規分布}}}_{?}$$

- 変数選択がβ_Dの分布に影響を与えてしまい、正規分布に収束しない

5.5 イメージ



5.6 Double-selection

1. Y および D を予測するモデルを、LASSO で推定し、選択された変数を記録
2. **どちらかの**予測モデルで選択された変数 (Z) を用いて、 $Y \sim D + Z$ を回帰

- 重要な変数を誤って除外しないように、Yの”予測 AI”とDの”予測 AI”に”ダブルチェック”を行わせている

5.7 推定値の分布

- 仮定: (Approximately) sparsity: 事例数に比べて、十分に少ない変数数で、母平均をうまく近似できる
  - Xの中には、“trivial”な変数も含まれている
- 事例数が十分に大きくなると、β_Dの推定値の分布 $=$

$$\underbrace{\underbrace{Step1:\textbf{変数選択の誤差}}_{\to 0}+\underbrace{Step2:OLS\textbf{の誤差}}_{\Rightarrow \textbf{正規分布}}}_{\Rightarrow \textbf{正規分布}}$$

- 変数選択による、推定値の分散を削減できる

5.8 直感

- $Y \sim X$ で変数選択すると、Y とそこそこ相関がある変数も除外される可能性がある
  - バランス後の比較のおいては、D との相関も重要
    - D 間で分布が大きく異なる変数ならば、バランス後の比較結果に大きな影響を与える
- $D \sim X$ での変数選択結果も活用し、上記のリスクを軽減する

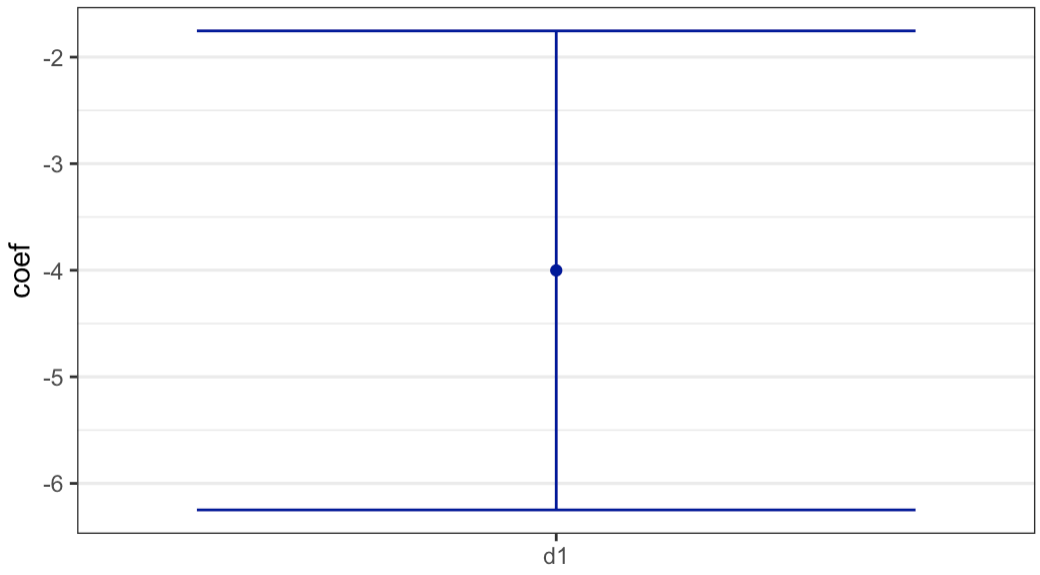
5.9 例

5.10 例

```
PDS <- hdm::rlassoEffect(
  y = Y,
  d = D,
  x = X
)
```

5.11 例

```
plot(PDS)
```



5.12 例

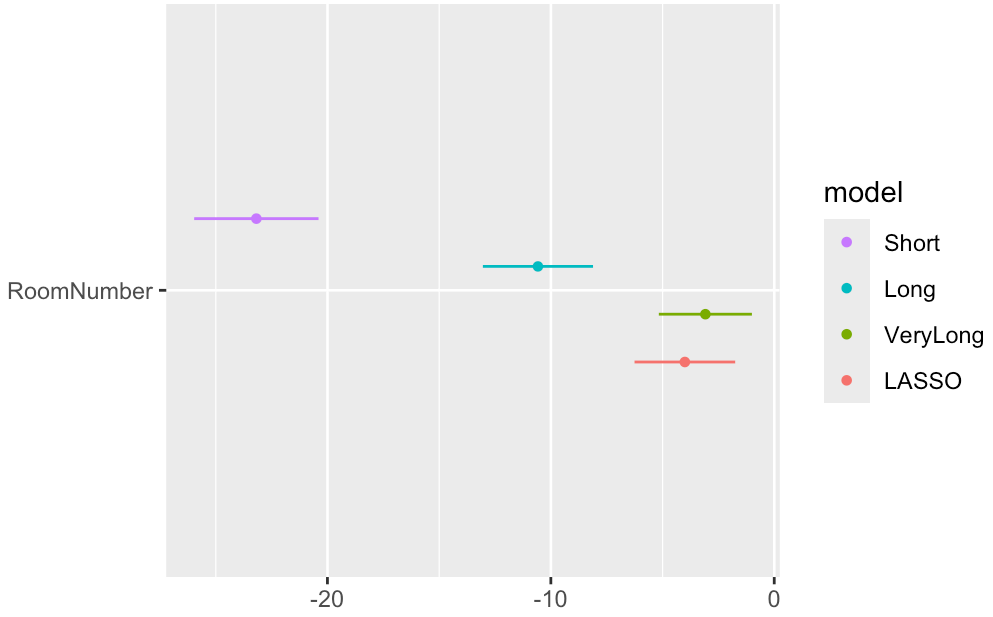
```
PDS$selection.index
```

```
                    Size            District中央区            District中野区
                    TRUE                     TRUE                     FALSE
            District北区          District千代田区            District台東区
                   FALSE                    FALSE                     FALSE
          District品川区            District大田区            District文京区
                   FALSE                    FALSE                     FALSE
          District新宿区            District杉並区            District板橋区
                    TRUE                    FALSE                     FALSE
        District江戸川区            District江東区            District渋谷区
                   FALSE                    FALSE                      TRUE
            District港区            District目黒区            District練馬区
                   FALSE                     TRUE                     FALSE
          District荒川区            District葛飾区            District豊島区
                   FALSE                    FALSE                     FALSE
          District足立区            District墨田区                    Tenure
                   FALSE                     TRUE                      TRUE
                Distance                 I(Size^2)              I(Tenure^2)
                   FALSE                     TRUE                     FALSE
           I(Distance^2)       Size:District中央区       Size:District中野区
                    TRUE                     TRUE                     FALSE
       Size:District北区     Size:District千代田区       Size:District台東区
                    TRUE                     TRUE                     FALSE
     Size:District品川区       Size:District大田区       Size:District文京区
                   FALSE                     TRUE                     FALSE
     Size:District新宿区       Size:District杉並区       Size:District板橋区
                    TRUE                    FALSE                      TRUE
   Size:District江戸川区       Size:District江東区       Size:District渋谷区
                    TRUE                     TRUE                      TRUE
       Size:District港区       Size:District目黒区       Size:District練馬区
                    TRUE                     TRUE                      TRUE
     Size:District荒川区       Size:District葛飾区       Size:District豊島区
                    TRUE                     TRUE                     FALSE
     Size:District足立区       Size:District墨田区              Size:Tenure
                    TRUE                     TRUE                      TRUE
           Size:Distance     District中央区:Tenure     District中野区:Tenure
                    TRUE                     TRUE                     FALSE
     District北区:Tenure   District千代田区:Tenure     District台東区:Tenure
                   FALSE                    FALSE                     FALSE
   District品川区:Tenure     District大田区:Tenure     District文京区:Tenure
                   FALSE                    FALSE                     FALSE
   District新宿区:Tenure     District杉並区:Tenure     District板橋区:Tenure
                   FALSE                    FALSE                     FALSE
District江戸川区:Tenure     District江東区:Tenure     District渋谷区:Tenure
                   FALSE                     TRUE                     FALSE
     District港区:Tenure     District目黒区:Tenure     District練馬区:Tenure
```

```
                  FALSE                      FALSE                      FALSE
   District荒川区:Tenure     District葛飾区:Tenure     District豊島区:Tenure
                  FALSE                      FALSE                      FALSE
   District足立区:Tenure     District墨田区:Tenure   District中央区:Distance
                  FALSE                      FALSE                      FALSE
 District中野区:Distance     District北区:Distance District千代田区:Distance
                  FALSE                      FALSE                      FALSE
 District台東区:Distance   District品川区:Distance   District大田区:Distance
                   TRUE                      FALSE                      FALSE
 District文京区:Distance   District新宿区:Distance   District杉並区:Distance
                  FALSE                       TRUE                      FALSE
 District板橋区:Distance District江戸川区:Distance   District江東区:Distance
                  FALSE                      FALSE                      FALSE
 District渋谷区:Distance     District港区:Distance   District目黒区:Distance
                   TRUE                      FALSE                      FALSE
 District練馬区:Distance   District荒川区:Distance   District葛飾区:Distance
                  FALSE                      FALSE                      FALSE
 District豊島区:Distance   District足立区:Distance   District墨田区:Distance
                  FALSE                      FALSE                      FALSE
        Tenure:Distance
                  FALSE
```

5.13 例

```
dotwhisker::dwplot(
  list(
    Short = model,
    Long = model_long,
    VeryLong = model_very_long,
    LASSO = LASSO),
  vars_order = c("RoomNumber"))
```



6 重回帰の画像診断

6.1 重回帰の問題点

- 一般に、X が複数存在するケースで、OLS 推定を誤解なくイメージすることは難しい
- FWL 定理 から、残差回帰として解釈し、visualization できる

6.2 残差回帰

- 以下の手順で推定したとしても、重回帰と全く同じ β_D が推定される

1. Y と D を X から OLS 推定し、Y/D の”予測値”を算出する
2. “予測誤差” $Y^* = Y -$ **予測値** と $D^* = D -$ **予測値** を算出する
3. $Y^* \sim D^*$ を OLS 回帰する

6.3 例

```
estimatr::lm_robust(
  Price ~ RoomNumber + Size,
  data)
```

```
               Estimate Std. Error     t value      Pr(>|t|)   CI Lower
(Intercept)   0.3175006 1.08132352   0.2936222  7.690520e-01  -1.802081
RoomNumber  -23.1817649 1.41987399 -16.3266354  3.031162e-59 -25.964965
Size          1.8938610 0.06708868  28.2292197 1.771707e-169   1.762356
```

```
              CI Upper    DF
(Intercept)   2.437083 11308
RoomNumber  -20.398565 11308
Size          2.025366 11308
```

```
data$Res_D <- lm(RoomNumber ~ Size, data)$residuals
data$Res_Y <- lm(Price ~ Size, data)$residuals

estimatr::lm_robust(
  Res_Y ~ Res_D,
  data)
```

```
                Estimate Std. Error       t value    Pr(>|t|)    CI Lower
(Intercept)  1.106809e-13  0.3356026  3.297976e-13 1.00000e+00  -0.6578393
Res_D       -2.318176e+01  1.4194971 -1.633097e+01 2.82752e-59 -25.9642259
               CI Upper    DF
(Intercept)   0.6578393 11309
Res_D       -20.3993039 11309
```

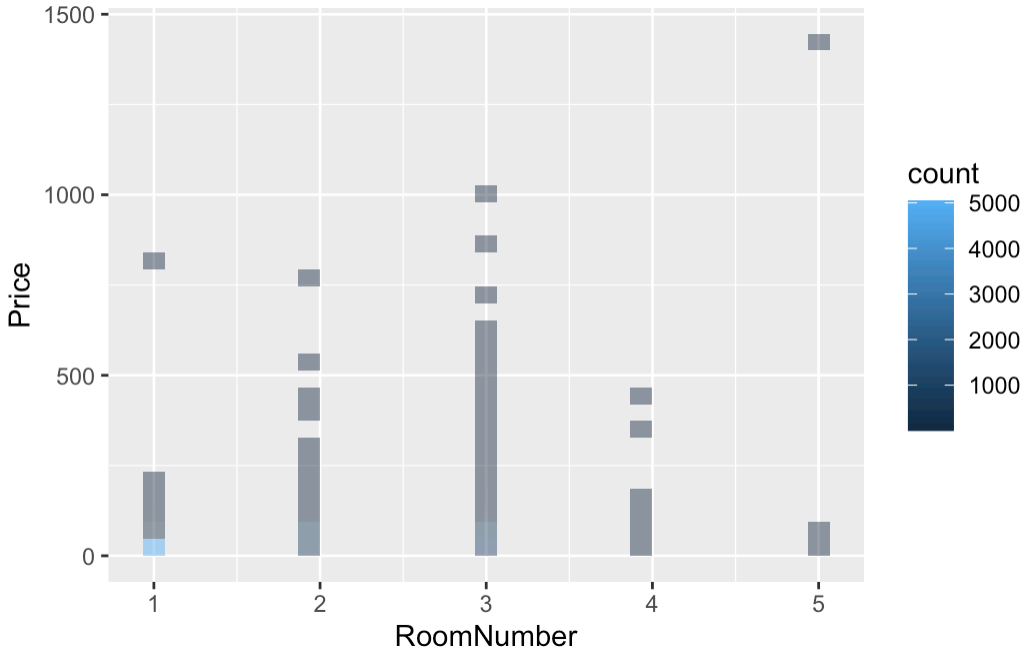
6.4 可視化

- 残差回帰は、最終的にY^* を D^* のみで回帰するため、容易に可視化できる
  - ヒートプロット + 平均 + OLS

6.5 例: 単純比較

```
Fig1 <- data |>
  ggplot(
    aes(
      x = RoomNumber,
      y = Price
    )
  ) +
  geom_bin2d(
    alpha = 0.5
  )
```

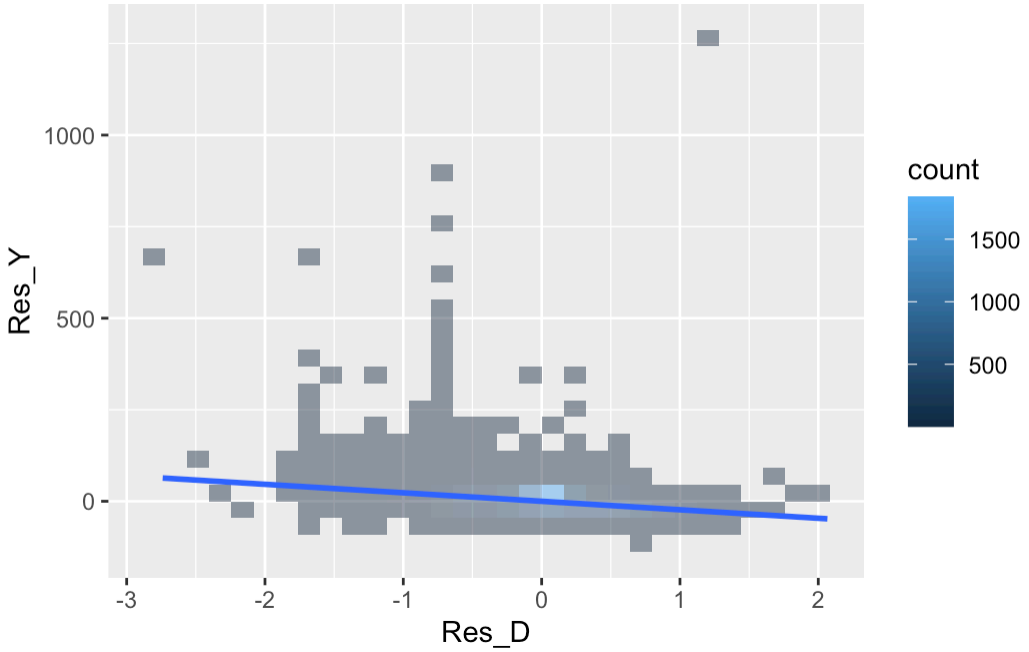
6.6 例: 単純比較



6.7 例: Size のバランス

```
Fig2 <- data |>
  ggplot(
    aes(
      x = Res_D,
      y = Res_Y
    )
  ) +
  geom_bin2d(
    alpha = 0.5
  ) +
  geom_smooth(
    method = "lm",
    se = FALSE
  )
```

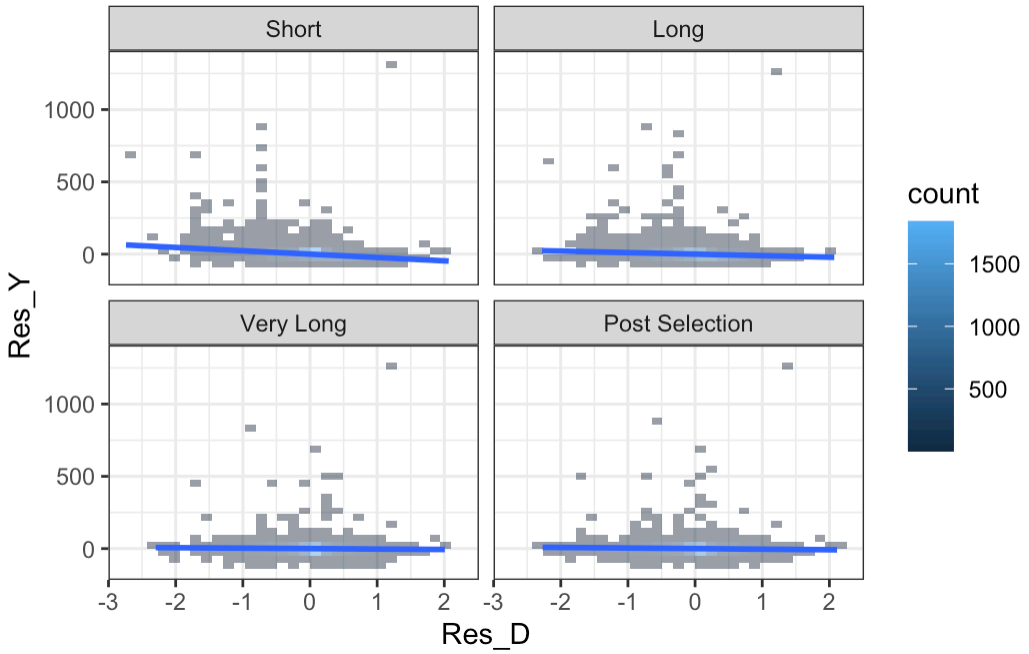
6.8 例: Size のバランス



6.9 例: 比較

- Short: $Price \sim RoomNumber + Size$
- Long: $Price \sim RoomNumber + Size + District + Tenure + Distance$
- Very Long: $Price \sim RoomNumber + (Size + District + Tenure + Distance)^2 + I(Size^2) + I(Tenure^2) + I(Distance^2)$
- Post Selection: Very Long を Double-selection で推定

6.10 例



6.11 Takeaway

- X に関する違いをバランスさせる分析は、データ分析の中核的な関心の一つ
- $$Y \sim \underbrace{\textbf{関心となる部分}}_{\beta_D D} + \underbrace{\textbf{局外}}_{\beta_0+\beta_1 X_1+..}$$
  - 局外部分を十分に複雑化し、OLS モデルを推定することで達成できる
  - LASSO による変数選択も応用できる
    - AI によるミスの影響を緩和する工夫が必要

6.12 Reference

Bibliography