Potlačení kolinearity

- získání dalších dat
- změna formulace modelu
  - - někdy pomůže centrování proměnných
  - - vynechání proměnných (může vést k vychýleným odhadům)
- kolinearita $\Rightarrow$ velké rozptyly odhadů
  - - Gauss-Markov říká, že OLS je BLUE parametru β
  - - tzn. zmenšené rozptylu $\widehat{\beta}$ je možné jen použitím nelineárních nebo vychýlených odhadů

Hřebenová regrese (Ridge regression)

- zůstaneme u standardizovaného modelu, tzn. uvažujeme $\boldsymbol{X}^*, \boldsymbol{Y}^*$
- pro jednoduchost značení budeme používat $\boldsymbol{X}, \boldsymbol{Y}$ a předpokládat $h(\boldsymbol{X}) = m$
- hřebenová regrese - uměle se zvedne diagonála $\boldsymbol{X}^T\boldsymbol{X}$ a tím se potlačí kolinearita

- uvažujme $\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m$, kde $\delta > 0$ je malé

  - $\boldsymbol{X}^T\boldsymbol{X}$ je symetrická, tzn. ex. OG matice $\boldsymbol{Q}$ taková, že $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{S}^2\boldsymbol{Q}^T$,

  kde $\boldsymbol{S}^2$ je diagonální matice, mající na diagonále vlastní čísla, λ_i, matice $\boldsymbol{X}^T\boldsymbol{X}$

  - protože $h(\boldsymbol{X}^T\boldsymbol{X}) = m$, je $\boldsymbol{X}^T\boldsymbol{X}$ PD a $\lambda_i > 0$, $\forall i \in \widehat{m}$

  - $\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m = \boldsymbol{Q}(\boldsymbol{S}^2 + \delta\boldsymbol{I}_m)\boldsymbol{Q}^T$ to je ale regulární matice

  - odtud plyne

$$(\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m)\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{S}^2 + \delta\boldsymbol{I}_m)$$

  tzn. čísla na diagonále $\boldsymbol{S}^2 + \delta\boldsymbol{I}_m$ jsou vlastní čísla matice $\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m$

- neboli $\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m$ je regulární a má vlastní čísla $\lambda_i + \delta$
- jejich velikost i poměry lze měnit volbou δ - hlavní myšlenka HR
- místo odhadu $\widehat{\beta}^{LS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ budeme studovat odhad

$$\widehat{\beta}^{R,\delta} = (\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

VĚTA 6.3

Vychýlení odhadu $\widehat{\beta}^{R,\delta}$ je

$$bias(\widehat{\beta}^{R,\delta}) = -\delta(\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m)^{-1}\beta.$$

Pro střední kvadratickou chybu odhadu $\widehat{\beta}^{R,\delta}$ platí

$$MSE(\widehat{\beta}^{R,\delta}) = (\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m)^{-1}\left[\sigma^2\boldsymbol{X}^T\boldsymbol{X} + \delta^2\beta\beta^T\right](\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m)^{-1}.$$

Důkaz.

Věta 6.4

Ve standardizovaném modelu s plnou hodností pro $0 < \delta < \frac{2\sigma^2}{\|\boldsymbol{\beta}\|^2}$ platí, že

$$\mathrm{Cov}\widehat{\boldsymbol{\beta}}^{LS} - MSE(\widehat{\boldsymbol{\beta}}^{R,\delta}) \quad \text{je PD matice.}$$

Poznámka 6.7

- ukázali jsme $\mathrm{E}\|\widehat{\boldsymbol{\beta}}\|^2 > \|\boldsymbol{\beta}\|^2$ $\qquad (+\sigma^2 \mathrm{tr}(\boldsymbol{X}^T\boldsymbol{X})^{-1})$
- můžeme se snažit $\widehat{\boldsymbol{\beta}}$ nějak zkrátit $\qquad$ (za cenu ztráty nestrannosti)
- jiná interpretace hřebenového odhadu:

  - hledejme $\boldsymbol{\beta}$, které pro dané δ minimalizuje

$$\varphi(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \delta\|\boldsymbol{\beta}\|^2$$

  - dá se ukázat, že řešením je $\widehat{\boldsymbol{\beta}}^{R,\delta}$

VĚTA 6.5

Nechť $\boldsymbol{PSQ}^T$ je singulární rozklad matice $\boldsymbol{X}$. Potom platí

$$\mathrm{Cov}(\widehat{\beta}^{R,\delta}) = \sigma^2 \sum_{i=1}^{m} \left(\frac{s_i}{s_i^2 + \delta} \right)^2 \boldsymbol{q}_i \boldsymbol{q}_i^T = \sigma^2 \sum_{i=1}^{m} \frac{\lambda_i}{(\lambda_i + \delta)^2} \boldsymbol{q}_i \boldsymbol{q}_i^T .$$

Důkaz.

Poznámka 6.8

- pokud je $\boldsymbol{X}$ špatně podmíněná, bude min. jedno s_i^2 malé, tedy $\frac{1}{s_i^2}$ vystupující ve $\mathrm{Var}(\widehat{\beta}_j^{LS})$ bude velké
- ale $\frac{s_i^2}{(s_i^2+\delta)^2}$ může být zase malé!

Volba parametru δ

1) **Hřebenová stopa (ridge trace):**

   - vypočítají se $\widehat{\boldsymbol{\beta}}^{R,\delta}$ pro různé hodnoty δ

   - vykreslí se graf jednotlivých složek $\widehat{\boldsymbol{\beta}}^{R,\delta}$ v závislosti na δ

   - doporučuje se volit takové δ, pro které se grafy „stabilizují“ (subjekivní)

2) **Harmonic mean estimator:** $\widehat{\delta} = \frac{m s_n^2}{\|\widehat{\boldsymbol{\beta}}\|^2}$,

   kde $\widehat{\boldsymbol{\beta}}$ a s_n^2 jsou klasické (OLS) odhady parametrů $\boldsymbol{\beta}, \sigma^2$

Poznámka 6.9

- existuje spousta dalších metod pro odhad δ, vlastnosti většinou zkoumány jen pomocí simulačních studií
- pokud nemáme nějakou apriorní informaci o β, použití hřebenové regrese nezaručí zlepšení OLS

Příklad 6.1 (Data Cement)

```
XX <- cbind(scale(x1),scale(x2),scale(x3),scale(x4))
kappa(XX,exact = TRUE)

## 37.10634

ridge <- lm.ridge(scale(y) ~ XX, lambda = seq(0,0.3,0.001))

plot(ridge)

select(ridge)

## modified HKB estimator is 0.08499604
## modified L-W estimator is 0.05830686
## mallest value of GCV  at 0.3
```