Statistische Signifikanz in hochdimensionalen linearen Modellen - Handout

Stella Preußler

15. Januar 2015

Einleitung

Thema: Statistische Signifikanz in hochdimensionalen linearen Modellen...
... durch **p-Werte**, also
Ziel: Konstruktion von p-Werten für Hypothesen in hochdimensionalen linearen Modellen.

Definition 1 p-Wert $= \mathbb{P}\left[|T| \geq t\right]$ $\underline{\text{unter } H_0}$,
wobei T Teststatistik und t Experimentausgang.
Der p-Wert ist also die Wahrscheinlichkeit ein solches Stichprobenergebnis oder ein noch extremeres zu erhalten, wenn die Nullhypothese wahr ist.

1 Grundlegendes: Modell und Hypothesen

Wir betrachten ein **hochdimensionales lineares System** von der Form

$$\mathbf{Y} = \mathbf{X}\beta^0 + \epsilon, \tag{1}$$

mit Ergebnisvektor $\mathbf{Y} \in \mathbb{R}^{n \times 1}$,
fest vorgegebene Designmatrix $\mathbf{X} \in \mathbb{R}^{n \times p}$,
wahrem Paramtervektor $\beta^0 \in \mathbb{R}^{p \times 1}$ und
stochastischem Fehlervektor $\epsilon \in \mathbb{R}^{n \times 1}$ mit $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$,
im Fall $p \gg n$ und beschäftigen uns mit **Nullhypothesen** der Form:

$$H_{0,j} : \beta_j^0 = 0 \quad j = 1, ..., p. \tag{2}$$

2 Ein angemessener Schätzer

2.1 Die Ridge Regression

Die **Ridge Regression**

$$\hat{\beta} = argmin_{\beta} \left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 / n + \lambda \|\beta\|_2^2 \right), \tag{3}$$

wo $\lambda = \lambda_n$ der Regularisierungsparameter und Ω die Kovarianzenmatrix des Ridgeschätzers und wir annehmen, dass

$$\min_{j \in \{1,\dots,p\}} \Omega_{jj}(\lambda) = \min_{j \in \{1,\dots,p\}} \mathrm{Var}\left[\hat{\beta}_j\right](\lambda) > 0, \tag{4}$$

schätz eigentlich $\theta^0 = P_{\mathbf{X}}\beta^0 \in \mathcal{R}(\mathbf{X})$, wobei $\mathcal{R}(\mathbf{X}) \subset \mathbb{R}^p$ der lineare Raum, der durch die n Zeilen von X aufgespannt wird und $P_{\mathbf{X}}$ die **Projektion von $\mathbb{R}^p$ auf $\mathcal{R}(\mathbf{X})$** ist.

Es gilt

$$0 < L_C \leq \liminf_{\lambda \in (0,C]} \min_{j \in \{1,\dots,p\}} \mathrm{Var}[\hat{\beta}_j] \leq M_C \tag{5}$$

für ein $0 < C < \infty$ und Konstanten mit $0 < L_C < M_C < \infty$ abhängig von C und der Designmatrix $\mathbf{X}$, und

$$\max_{j \in \{1,\dots,p\}} \left(\mathbb{E}[\hat{\beta}_j] - \theta_j^0 \right)^2 \leq \min_{j \in \{1,\dots,p\}} \mathrm{Var}[\hat{\beta}_j], \tag{6}$$

für einen Regularisierungsparameter $\lambda > 0$ der folgende Eigenschaft* hat:

$$\lambda \left(\min_{j \in \{1,\dots,p\}} \mathrm{Var}[\hat{\beta}_j] \right)^{-1/2} \leq n^{-1/2} \sigma \left\|\theta^0\right\|_2^{-1} \lambda_{min \neq 0}(\hat{\Sigma}) \tag{7}$$

wo $\lambda_{min \neq 0}(\hat{\Sigma})$ kleinster Eigenwert $\neq 0$ der Kovarianzmatrix $\hat{\Sigma} = n^{-1}\mathbf{X}^T\mathbf{X}$.

2.2 Die korrigierte Ridge Regression stochastisch modelliert

Die **korrigierte Ridge Regression** $\hat{\beta}_{corr;j}$ mit Regularisierungsparameter $\lambda > 0$ zum Testen von $H_{0,j}$

$$\hat{\beta}_{corr;j} = \hat{\beta}_j - \hat{B}_{H_{0;j}} = \hat{\beta}_j - \sum_{k \neq j} (P_{\mathbf{X}})_{jk} \hat{\beta}_{init;k} \tag{8}$$

kann durch

$$\hat{\beta}_{corr;j} = Z_j + \gamma_j \quad (j = 1, ..., p) \tag{9}$$

$Z_1, ..., Z_p \sim \mathcal{N}(0, n^{-1}\sigma^2\Omega), \Omega = \Omega(\lambda)$,
$\gamma_j = (P_{\mathbf{X}})_{jj} \beta_j^0 - \sum_{k \neq j} (P_{\mathbf{X}})_{jk} \left(\hat{\beta}_{init;k} - \beta_k^0\right) + b_j(\lambda)$,
$b_j(\lambda) = \mathbb{E}[\hat{\beta}_j(\lambda)] - \theta_j^0$ „Schätzungsverzerrung“
stochastisch dargestellt werden. Mit den **Normalisierungsfaktoren**

$$a_{n,p;j}(\sigma) = n^{1/2}\sigma^{-1}\Omega_{jj}^{-1/2} (j = 1, ..., p) \tag{10}$$

gilt $a_{n,p;j}(\sigma) Z_j \sim \mathcal{N}(0, 1)$.

2.3 Eine asymptotische Schranke: Präzise formuliert

Definition 2 Seien X und Y reelle Zufallsvariablen. X ist kleiner-gleich Y bezüglich der gewöhnlichen stochastischen Ordnung, wenn für alle $b \in \mathbb{R}$ gilt $\mathbb{P}(X \geq b) \leq \mathbb{P}(Y \geq b)$.

Satz 1 *Wir betrachte eine Schar von linearen Modellen*

$$\boldsymbol{Y}_n = \boldsymbol{X}_n \beta_n^0 + \epsilon_n, n = 1, 2, ..., \tag{11}$$

*die korrigierte Ridge Regression $\hat{\beta}_{corr}$ mit geeignetem** Regularisierungsparameter $\lambda_n > 0$*

$$\lambda_n \Omega_{min}(\lambda_n)^{-1/2} = o(\min(n^{-1/2}\sigma \left\|\theta^0\right\|_2^{-1} \lambda_{\min\neq 0}(\hat{\Sigma}))), (n \to \infty) \tag{12}$$

und machen folgende Annahme A:
Es gibt Konstanten $\Delta_j = \Delta_{j,n} > 0$ so dass

$$\mathbb{P}\left[\bigcap_{j=1}^{p_n} \left\{ \left| a_{n,p;j}(\sigma) \sum_{k\neq j} (P_{\boldsymbol{X}})_{jk} (\hat{\beta}_{init;k} - \beta_k^0) \right| \leq \Delta_{j,n} \right\}\right] \to 1 \quad (n \to \infty). \tag{13}$$

Dann gilt für $j \in \{1, ..., p_n\}$ unter $H_{0,j}$: für alle $u \in \mathbb{R}^+$,

$$\limsup_{n\to\infty} \left(\mathbb{P}\left[a_{n,p;j}(\sigma) |\hat{\beta}_{corr;j}| > u \right] - \mathbb{P}\left[|W| + \Delta_j > u \right] \right) \leq 0,$$

wobei $W \sim \mathcal{N}(0, 1)$.

3 Die Herleitung der p-Werte

3.1 p-Werte

$$\text{p-Wert} = P\left[|W + \Delta_j| \geq a_{n,p;j}(\sigma) |\hat{\beta}_{corr;j}| \right] \text{ unter } H_0 \tag{14}$$

Für die individuelle Hypothese $H_{0,j}$ definieren wir daher den p-Wert für die zweiseitige Alternative als:

$$P_j = 2(1 - \Phi((a_{n,p;j}(\sigma)|\hat{\beta}_{corr;j}| - \Delta_j)_+)). \tag{15}$$

3.2 Die Herleitung der Δ_j

Satz 2 *Betrachte (2) mit normalisierten Spalten $\hat{\Sigma}_{jj} \equiv 1$, welche die Kompatibilitätsbedingung mit Konstante $\Phi_0^2 = \Phi_{0,n}^2$ erfüllen. Nehme den Lasso als Initialschätzer $\hat{\beta}_{init}$ mit Regularisierungsparameter $\lambda_{Lasso} = 4\sigma\sqrt{C\log(p_n)/n}$ für ein $2 < C < \infty$. Nehme an, dass die Menge der aktiven Koeffizienten $s_0 = s_{0,n} = o((n/log(p_n))^{\xi}) \quad (n \to \infty)$ für ein $0 < \xi < 1/2$, und dass $\liminf_{n\to\infty} \Phi_{0,n}^2 > 0$. Dann erfüllt*

$$\Delta_j :\equiv \max_{k\neq j} |a_{n,p;j}(\sigma)(P_{\boldsymbol{X}})_{jk}| \, (\log(p)/n)^{1/2-\xi} \tag{16}$$

die Bedingung A (13).

Korollar 1 *Nehme die Annahmen von Satz 2, ohne Bedingung A und mit den Bdingungen von Satz 2, an. Dann gilt, mit dem Lasso als Inititalschätzer, die Aussage von Satz 1.*

Literatur

[B13] Peter Buhlmann, Statistical significance in high-dimensional linear models, Bernoulli, 19(4), 1212 - 1242, (2013).