Statistische Signifikanz in hochdimensionalen linearen Modellen

Hauptseminar Erweiterungen des linearen Regressionsmodells und genomische Anwendungen in der Biomedizin WS 2014/2015

Stella Preußler

12. Januar 2015

Inhaltsverzeichnis

Hochdimensionales lineares Modell

Wir betrachten ein hochdimensionales lineares Modell von der Form

$$\mathbf{Y} = \mathbf{X}\beta^0 + \epsilon, \quad (1)$$

mit Ergebnisvektor $\mathbf{Y} \in \mathbb{R}^{n\times 1}$,
fest vorgegebener 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 $p \gg n$.

Hypothesen

Das Paper beschäftigt sich mit einer oder vielen Nullhypothesen der Form

$$H_{0,G} : \beta_j^0 = 0 \ \ \forall j \in G,$$

wo $G \subset \{1, ..., p\}$ eine Teilmenge aller Indizes der Kovariablen ist.

Hypothesen

Das Paper beschäftigt sich mit einer oder vielen Nullhypothesen der Form

$$H_{0,G} : \beta_j^0 = 0 \ \ \forall j \in G,$$

wo $G \subset \{1, ..., p\}$ eine Teilmenge aller Indizes der Kovariablen ist. Wir interessieren uns heute allerdings nur für die **Individuellen Nullhypothesen** der Form

$G = \{j\}$: für den j-ten Regressionsparameter für $j = 1, ..., p$, d.h.:

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

„Statistische Signifikanz in hochdimensionalen linearen Modellen"...

... durch p-Werte, also beschäftigen wir uns heute mit der **Konstruktion von p-Werten für Hypothesen in hochdimensionalen linearen Modellen.**

„Statistische Signifikanz in hochdimensionalen linearen Modellen"...

... durch p-Werte, also beschäftigen wir uns heute mit der **Konstruktion von p-Werten für Hypothesen in hochdimensionalen linearen Modellen.**

Vorgehensweise:

„Statistische Signifikanz in hochdimensionalen linearen Modellen"...

... durch p-Werte, also beschäftigen wir uns heute mit der **Konstruktion von p-Werten für Hypothesen in hochdimensionalen linearen Modellen.**

Vorgehensweise:

einen angemessenen Schätzer für β^0 finden

„Statistische Signifikanz in hochdimensionalen linearen Modellen"...

... durch p-Werte, also beschäftigen wir uns heute mit der **Konstruktion von p-Werten für Hypothesen in hochdimensionalen linearen Modellen.**

Vorgehensweise:

einen angemessenen Schätzer für β^0 finden

$\Rightarrow$ asymptotische Verteilung für diesen Schätzer

„Statistische Signifikanz in hochdimensionalen linearen Modellen“...

... durch p-Werte, also beschäftigen wir uns heute mit der **Konstruktion von p-Werten für Hypothesen in hochdimensionalen linearen Modellen.**

Vorgehensweise:

einen angemessenen Schätzer für β^0 finden

$\Rightarrow$ asymptotische Verteilung für diesen Schätzer

$\Rightarrow$ p-Wert

„Statistische Signifikanz in hochdimensionalen linearen Modellen"...

... durch p-Werte, also beschäftigen wir uns heute mit der **Konstruktion von p-Werten für Hypothesen in hochdimensionalen linearen Modellen.**

Vorgehensweise:

einen angemessenen Schätzer für β^0 finden

$\Rightarrow$ asymptotische Verteilung für diesen Schätzer

$\Rightarrow$ p-Wert $= \mathbb{P}[|T| \geq t]$ $\underline{\text{unter } H_0}$,

T Teststatistik, t Experimentausgang.

„Statistische Signifikanz in hochdimensionalen linearen Modellen"...

... durch p-Werte, also beschäftigen wir uns heute mit der **Konstruktion von p-Werten für Hypothesen in hochdimensionalen linearen Modellen.**

Vorgehensweise:

einen angemessenen Schätzer für β^0 finden

$\Rightarrow$ asymptotische Verteilung für diesen Schätzer

$\Rightarrow$ p-Wert $= \mathbb{P}[|T| \geq t]$ $\underline{\text{unter } H_0}$,

T Teststatistik, t Experimentausgang.

$\Rightarrow$ Aussagen wie „$\mathrm{H}_{0,3} : \beta_3^0 = 0$ kann zum Signifikanzniveau α abgelehnt werden $\Rightarrow$ der Effekt der 3. Kovariate von $\mathbf{X}$ ist statistisch signifikant" möglich.

Ridge Regression

Sei

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

die **Ridge Regression**, wobei
$\lambda = \lambda_n$ der Regularisierungsparameter, Ω die Kovarianzenmatrix und wir annehmen, dass

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

Ridge Regression

Sei

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

die **Ridge Regression**, wobei
$\lambda = \lambda_n$ der Regularisierungsparameter, Ω die Kovarianzenmatrix und wir annehmen, dass

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

Eigentlich schätzt die Ridge Regression jedoch $\theta^0 = P_{\mathbf{X}}\beta^0 \in \mathcal{R}(\mathbf{X})$, 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.

Angemessenheit

Es gilt

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

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

Angemessenheit

Es gilt

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

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

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

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

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

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

Projektionsverzerrung

Ridge Regression schätzt den Parameter $\theta^0 = P_{\mathbf{X}}\beta^0$, d. h. es tritt neben der Schätzungsverzerrung noch eine zusätzliche Projektionsverzerrung B_j (j=1, ..., p) auf

Projektionsverzerrung

Ridge Regression schätzt den Parameter $\theta^0 = P_{\mathbf{X}}\beta^0$, d. h. es tritt neben der Schätzungsverzerrung noch eine zusätzliche Projektionsverzerrung B_j (j=1, ..., p) auf :

$$B_j = \theta_j^0 - \beta_j^0$$

Projektionsverzerrung

Ridge Regression schätzt den Parameter $\theta^0 = P_{\mathbf{X}}\beta^0$, d. h. es tritt neben der Schätzungsverzerrung noch eine zusätzliche Projektionsverzerrung B_j (j=1, ..., p) auf :

$$B_j = \theta_j^0 - \beta_j^0 = \left(P_{\mathbf{X}}\beta^0\right)_j - \beta_j^0 = (P_{\mathbf{X}})_{jj}\,\beta_j^0 - \beta_j^0 + \sum_{k \neq j} (P_{\mathbf{X}})_{jk}\,\beta_k^0.$$

Projektionsverzerrung

Ridge Regression schätzt den Parameter $\theta^0 = P_{\mathbf{X}}\beta^0$, d. h. es tritt neben der Schätzungsverzerrung noch eine zusätzliche Projektionsverzerrung B_j (j=1, ..., p) auf :

$$B_j = \theta_j^0 - \beta_j^0 = \left(P_{\mathbf{X}}\beta^0\right)_j - \beta_j^0 = \left(P_{\mathbf{X}}\right)_{jj}\beta_j^0 - \beta_j^0 + \sum_{k\neq j}\left(P_{\mathbf{X}}\right)_{jk}\beta_k^0.$$

Da wir p-Werte konstruieren wollen, müssen wir die Verzerrung nur unter H_0 berücksichtigen:

$$B_{H_{0;j}} = \sum_{k\neq j}\left(P_{\mathbf{X}}\right)_{jk}\beta_k^0 ...$$

Korrektur der Projektionsverzerrung

... und können sie mit Hilfe eines Initialschätzers $\hat{\beta}_{init}$ (z. B. Lasso) schätzen:

$$\hat{B}_{H_{0;j}} = \sum_{k \neq j} (P_{\mathbf{X}})_{jk} \, \hat{\beta}_{init;k}$$

um schließlich die **korrigierte Ridge Regression** $\hat{\beta}_{corr;j}$ zum Testen von $H_{0,j}$ zu erhalten:

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

Proposition 1

Proposition 1

Die korrigierte Ridge Regression $\hat{\beta}_{corr;j}$ mit Regularisierungsparameter $\lambda > 0$ können wir darstellen als:

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

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

Scharenmodell und Annahme

Für diese Verteilung soll nun eine asymptotisch stochastische Schranke unter der Nullhypothese gefunden werden.

Scharenmodell und Annahme

Für diese Verteilung soll nun eine asymptotisch stochastische Schranke unter der Nullhypothese gefunden werden.
Dazu betrachten wir eine Schar von linearen Modellen

$$\mathbf{Y}_n = \mathbf{X}_n \beta_n^0 + \epsilon_n, n = 1, 2, ..., \quad (11)$$

wobei sich alle Größen und auch die Dimension $p = p_n$ mit n verändern dürfen

Scharenmodell und Annahme

Für diese Verteilung soll nun eine asymptotisch stochastische Schranke unter der Nullhypothese gefunden werden.
Dazu betrachten wir eine Schar von linearen Modellen

$$\mathbf{Y}_n = \mathbf{X}_n \beta_n^0 + \epsilon_n, n = 1, 2, ..., \quad (11)$$

wobei sich alle Größen und auch die Dimension $p = p_n$ mit n verändern dürfen und machen folgende Annahme A (13):
Es gibt Konstanten $\Delta_j = \Delta_{j,n} > 0$ so dass

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

Asymptotische stochastische Schranke

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

Mit dieser Annahme und einem geeignet** gewählten Regularisierungsparameter $\lambda_n > 0$ kann man nun zeigen, dass unter der Nullhypothese $H_{0,j}$ für $j \in \{1, ..., p_n\}$:

$$a_{n.p;j}(\sigma)\left|\hat{\beta}_{corr;j}\right| \prec_{st} |W| + \Delta_j, \quad W \sim \mathcal{N}(0,1).$$

Asymptotische stochastische Schranke

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

Mit dieser Annahme und einem geeignet** gewählten Regularisierungsparameter $\lambda_n > 0$ kann man nun zeigen, dass unter der Nullhypothese $H_{0,j}$ für $j \in \{1, ..., p_n\}$:

$$a_{n.p;j}(\sigma)\left|\hat{\beta}_{corr;j}\right| \prec_{st} |W| + \Delta_j, \quad W \sim \mathcal{N}(0,1).$$

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

Asymptotische stochastische Schranke: Satz 1

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

Mit dieser Annahme und einem geeignet** gewählten Regularisierungsparameter $\lambda_n > 0$ kann man nun zeigen, dass unter der Nullhypothese $H_{0,j}$ für $j \in \{1, ..., p_n\}$:

$$a_{n.p;j}(\sigma)\left|\hat{\beta}_{corr;j}\right| \prec_{st} |W| + \Delta_j, \quad W \sim \mathcal{N}(0,1).$$

$\hat{\beta}_{corr;j} = Z_j + \gamma_j \quad ; \quad a_{n,p;j}(\sigma)Z_j \sim \mathcal{N}(0,1)$

$\gamma_j = (P_{\mathbf{X}})_{jj}\,\beta_j^0 - \sum_{k\neq j}(P_{\mathbf{X}})_{jk}\left(\hat{\beta}_{init;k} - \beta_k^0\right) + b_j(\lambda)$

$b_j(\lambda) = \mathbb{E}[\hat{\beta}_j(\lambda)] - \theta_j^0 \qquad (j = 1, ..., p)$

**

„Geeignet** gewählter Regularisierungsparameter" λ_n soll für uns (wieder; vgl. mit * aus (7)) heißen:

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

ist erfüllt, denn dann gilt:

$$\left\|a_{n,p} b(\lambda_n)\right\|_\infty \to 0 \quad \text{für} \quad n \to \infty,$$

**

„Geeignet** gewählter Regularisierungsparameter“ λ_n soll für uns (wieder; vgl. mit * aus (7)) heißen:

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

ist erfüllt, denn dann gilt:

$$\|a_{n,p} b(\lambda_n)\|_\infty \to 0 \quad \text{für} \quad n \to \infty,$$

denn

$$\|a_{n,p} b(\lambda_n)\|_\infty \leq \frac{\lambda_n \Omega_{min}(\lambda_n)^{-1/2}}{\min(n^{-1/2}\sigma \left\|\theta^0\right\|_2^{-1} \lambda_{\min\neq 0}(\hat{\Sigma}))} \to 0 \quad \text{für} \quad n \to \infty.$$

Herleitung der p-Werte mit Hilfe der asymptotischen Verteilung

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

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

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

Herleitung der p-Werte mit Hilfe der asymptotischen Verteilung

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

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

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

Um die p-Werte berechnen zu können, müssen wir die Δ_j kennen, wie sehen diese also aus?

Satz 2

Betrachte (11) 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 $s_0 = s_{0,n} = o((n/log(p_n))^{\xi})$ $(n \to \infty)$ für ein $0 < \xi < 1/2$, und dass $\liminf_{n\to\infty} \Phi_{0,n}^2 > 0$. Dann erfüllt

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

die Bedingung A (13).

... mit Hilfe von Korollar

Korollar

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