Multiplizitätskorrektur bei Variablenselektion

Seminar: Multiples Testen
Dozent: Prof. Dr. T. Dickhaus

Referent: Maximilian Mönch
- 22.11.2010 -

Überblick

1) Einleitung
- Problem der Variablenselektion -

Lineares Modell in der Regressionsanalyse: $Y = g(X) + \varepsilon = X\beta + \varepsilon$

$Y \in \mathbb{R}^n$...Response, Zielgröße,
$X \in \mathbb{R}^{n \times m}$...Versuchsplan/Designmatrix
$X_i = (X_{ij})_{j=1,\ldots,m} \in \mathbb{R}^m$...Prädiktorenvektor
$\beta = (\beta_1, \ldots, \beta_m)'$... Regressionskoeffizienten
$\varepsilon_i \sim N(0, \sigma^2)$ iid, σ^2 unbekannt.

Aufgabe: Wähle aus {1,...,m} k Zahlen und belasse $X_{ij_1}, \ldots X_{ij_k}$ im Modell.

Aufgefasst als multiples Testproblem: Teste Hypothesensystem $H_i : \beta_i = 0, i = 1, \ldots, m$

2) Multiplizitätskorrektur

Wann treten Multiplizitätsprobleme auf?

→ Diese Problematik tritt auf, wenn verhältnismäßig viele Variablen in das Modell aufgenommen werden.
→ So erhöht sich die Irrtumswahrscheinlichkeit des multiplen Tests

Veranschaulichung:

Sei $\varphi = (\varphi_i : i = 1, 2)$ ein multipler Test für ein zweiparametriges Modell.
Seien $\varphi_i, i = 1, 2$ stoch. unabhängig.

Weiterhin sei: $\mathbb{P}(\varphi_j = 1) = 0.05, j = 1, 2$

Daraus ergibt sich: $\mathrm{FWER}(\varphi) = \mathbb{P}(\bigcup_{i=1,2} \{\varphi_i = 1\}) = 2 * 0.95 * 0.05 + 0.05^2 = 0.0975 > 0.05$

2) Multiplizitätskorrektur

Im Rahmen der Suche nach bester Vorhersage wird die a posteriori Wahrscheinlichkeit betrachtet:

$$\mathbb{P}(M_\gamma|Y) \sim \mathbb{P}(M_\gamma) f(Y|M_\gamma)$$

Wobei $f(Y|M_\gamma) = \int f(Y|\beta_\gamma, \sigma)\pi(\beta_\gamma, \sigma) d\beta_\gamma d\sigma$ die likelihood von Y, mit $\pi(\beta_\gamma, \sigma)$ a priori-Dichte der Modellparameter, ist.

Durch die Abhängigkeit der likelihood von der a priori Wahrscheinlichkeit für die Modellparameter, ist demnach ein besonderes Augenmerk auf die Wahl dieser zu legen.

Nach Scott und Berger (2010) ermöglichen der empirische und vollständige Bayes-Ansatz eine automatische Multiplizitätskorrektur.

2) Multiplizitätskorrektur - empirischer Ansatz -

Annahme: $\gamma_i \sim Ber(p), i = 1, \ldots, m$, $p \in [0,1]$

Daraus folgt für die a priori Wahrscheinlichkeit, dass Modell M_γ vorliegt: $\mathbb{P}(M_\gamma|p) = p^{k_\gamma}(1-p)^{m-k_\gamma}$

Ansatz ist es nun, p mittels Maximum Likelihood Methode zu schätzen (George und Foster (2000)): $\hat{p} = \arg\max_{p\in[0,1]} \sum_\gamma \mathbb{P}(M_\gamma|p) \cdot f(Y|M_\gamma)$

Nun wählt man das Modell mit der maximalen a posteriori Wahrscheinlichkeit

$$\mathbb{P}(M_\gamma|Y) \sim \hat{p}^{k_\gamma}(1-\hat{p})^{m-k_\gamma} \cdot f(Y|M_\gamma)$$

2) Multiplizitätskorrektur
- empirischer Ansatz -

Die Multiplizitätskorrektur erfolgt hierbei folgenderrmaßen:

Angenommen, wir haben k wahre Parameter β_i . Dann folgt aus $m \rightarrow \infty$: $\hat{p} \rightarrow 0$

Grund: $m \rightarrow \infty \Rightarrow \mathbb{P}(M_\gamma | p) \rightarrow 0 \Rightarrow \hat{p} \rightarrow 0$

Das heißt, dass Modelle mit einer hohen Anzahl an Variablen unwahrscheinlich sind, wenn es darum geht, das Modell mit der höchsten a posteriori Wahrscheinlichkeit zu wählen.

2) Multiplizitätskorrektur
- vollst. Bayes-Ansatz -

Annahme: $p \sim Beta(a,b)$

Für die Wahrscheinlichkeit, dass Modell M_γ vorliegt,
gilt dann: $\mathbb{P}(M_\gamma) = \int\limits_0^1 f(M_\gamma|p)\pi(p)dp = \frac{\beta(a+k_\gamma,b+m-k_\gamma)}{\beta(a,b)}$

Im Spezialfall a=b=1, d.h. Parameter p ist gleichverteilt, vereinfacht sich die a priori Wahrscheinlichkeit:

$$\beta(1+k_\gamma, 1+m-k_\gamma) = \frac{\Gamma(1+k_\gamma)\Gamma(1+k_\gamma)}{\Gamma(2+m)} = \frac{(k_\gamma)!(m+k_\gamma)!}{(m+1)!} = \frac{(k_\gamma)!(m-k_\gamma)!}{(m+1)(m!)} = \frac{1}{m+1}\begin{pmatrix} m \\ k_\gamma \end{pmatrix}^{-1}$$

2) Multiplizitätskorrektur
- vollst. Bayes-Ansatz -

Das heißt, die für die a posteriori Wahrscheinlichkeit, die es zu maximieren gilt, gilt dann:

$$\mathbb{P}(M_\gamma|Y) \sim \frac{1}{m+1}\begin{pmatrix} m \\ k_\gamma \end{pmatrix}^{-1} \cdot f(Y|M_\gamma)$$

Auch wenn der Parameter p nicht mehr direkt vorhanden ist, so erfolgt dennoch eine Multiplizitätskorrektur. Denn offensichtlich gilt bei fester Anzahl wahrer Parameter und wachsendem m, dass die a posteriori Wahrscheinlichkeit gegen 0 konvergiert.

2) Multiplizitätskorrektur
- Nachteil des empir. Bayes Ansatz -

Trotz Konvergenz der empirischen Bayes-Lösung gegen den wahren Parameter, hilft der Ansatz nicht immer, das optimale Vorhersagemodell zu finden:

Lemma 1. :
Beim Problem der Variablenselektion gilt:
Hat das Nullmodell die echt größte likelihood, so ist der MLE-Schätzer $\hat{p}$ von p gleich 0.
Hat das vollständige Modell die echt größte likelihood, so ist der MLE-Schätzer $\hat{p}$ von p gleich 1.

3) Median probability model

Ziel: Finden einer Theorie eines optimalen Vorhersagemodells

Barbieri und Berger (2004) wiesen nach, dass dies das sogenannte „Median probability model“ ist.

Definition 1. *:*
Existiert die Gesamt-a-posteriori-Inklusionswahrscheinlichkeit

$$p_i \equiv \sum_{\gamma:\gamma_i=1} \mathbb{P}(M_\gamma|y)$$

für Variable i, so ist das Median probability model M_{γ^*} als das Modell definiert, welches aus denjenigen Variablen besteht, deren a-posteriori-Inklusions-Wahrscheinlichkeit mindestens $\frac{1}{2}$ beträgt.

3) Median probability model

Formal lässt sich das Modell wie folgt ausdrücken:

$$\gamma_i^* = \begin{cases} 1 & , \quad p_i \geq 1/2 \\ 0 & , \quad \text{sonst} \end{cases}$$

3) Median probability model

Existenz:

Es muss gewährleistet sein, dass die Menge der somit gewählten Kovariate zu der Menge der betrachteten Modelle und somit zum Hypothesensystem H gehört.

3) Median probability model

Dies ist gewährleistet, wenn:

1) Es werden alle möglichen Modelle H_i betrachtet.
2) Die betrachteten Modelle gehören zur Klasse der graphischen Modelle

Was ist ein graphisches Modell?

Definition 2. *:*
Sei $I(i)$ die zu einer Variablen x_i gehörenden Menge von Parameter-Indizes. Eine Teilklasse linearer Modelle besitzt graphische Modellstruktur, wenn es aus allen Modellen besteht, die die folgende Bedingung erfüllen:
„ Wenn für jedes i die Variable x_i zu dem Modell gehört, dann gehören die Variablen $x_j, j \in I(i)$ ebenfalls dazu.“

4) Optimalität

- Theorie zum besten Vorhersagemodell -

Nach dem Bayesianischen Ansatz müsste das Modell mit der höchsten a posteriori Wahrscheinlichkeit die beste Datenprädiktion ermöglichen.

Doch das ist nicht immer der Fall, da diese Wahrscheinlichkeiten oft sehr klein sind.
Barbieri und Berger (2004) nutzten die a posteriori Inklusionswahrscheinlichkeiten, eine Summation der a posteriori Wahrscheinlichkeiten, um das zuvor genannte Median probability model zu definieren.

Dieses weist die gewünschte beste Prädiktion auf.

4) Optimalität

- Theorie zum besten Vorhersagemodell -

Betrachten die Kovarianzmatrix $Q = \mathbb{E}[(X'X)]$

Ist Q Diagonalmatrix, so erfüllen die a posteriori Mittelwerte $\tilde{\beta}'_\gamma = H'_\gamma \tilde{\beta}$

mit
$$(H)_{ij} = \begin{cases} 1 & , \quad \gamma_i = 1 \text{ und } j = \sum_{r=1}^{i} \gamma_r \\ 0 & , \quad \text{sonst} \end{cases}$$

Beste Vorhersage bedeutet auch, das quadratische Risiko zu minimieren.
In diesem Fall gilt:

$$R(M_\gamma) = \sum_{i=1}^{k} \tilde{\beta}_i^{\,2} q_i (\gamma_i - p_i)^2$$

4) Optimalität

- Theorie zum besten Vorhersagemodell -

Fall 1: Es wird nur eine Auswahl an Modellen betrachtet

Theorem 1. *:*
Sei Q Diagonalmatrix mit Diagonaleinträgen $q_i > 0$, es gelte $\tilde{\beta}_\gamma = H'_\gamma \tilde{\beta}$ und die betrachteten Modellhypothesen besitzen graphische Modellstruktur.
Dann ist das Median probability model das beste Vorhersagemodell.

4) Optimalität

- Theorie zum besten Vorhersagemodell -

Fall 2: Es werden alle möglichen Modelle betrachtet:

Korollar 1. *:*
Sei Q diagonal mit $q_i > 0$, es gelte $\tilde{\beta}_\gamma = H'_\gamma \tilde{\beta}$ und alle Modelle M_γ werden betrachtet.
Dann ist das beste Vorhersagemodell das Median probability model.
Ist σ^2 bekannt und die a priori-Wahrscheinlichkeiten der Modelle erfüllen

$$\mathbb{P}(M_\gamma) = \prod_{i=1}^{k} (p_i^0)^{\gamma_i} (1 - p_i^0)^{1-\gamma_i}$$

, wobei p_i^0 die a priori-Wahrscheinlichkeit dafür ist, dass x_i im Modell ist, dann gilt
$\mathbb{P}(M_\gamma|y) = \prod_{i=1}^{k} (p_i)^{\gamma_i}(1-p_i)^{1-\gamma_i}$, wobei p_i die a posteriori Inklusionswahrscheinlichkeit ist, und das Median probability model ist das Modell mit der größten a posteriori Wahrscheinlichkeit.

4) Optimalität
- Theorie zum besten Vorhersagemodell -

Bemerkung:

1)
Viele a priori Wahrscheinlichkeiten erfüllen die Bedingung im Korollar.

Jeffreys Vorschlag (1961), dass die Wahrscheinlichkeit für die Modellordnung j ungefähr 1/j sein kann, erfüllt diese Bedingung allerdings nicht.

2)
Ist allen Modellen ein Intercept gemein, so wird die zugehörige a priori Wahrscheinlichkeit p_i^0 gleich 1 gesetzt.

4) Optimalität
- Anwendungsbeispiel -

ANOVA: 2-Faktor-Modell

Sei $y_{ijk} = \mu + a_i + b_j + ab_{ij} + \varepsilon_{ijk}, i = 1,2, j = 1,2, k = 1,2,\ldots,K, \varepsilon \sim N(0,\sigma^2)$ iid

Kurz: $y = X\beta + \varepsilon$

$$y = (y_{111},\ldots,y_{11K},y_{121},\ldots,y_{12K},y_{211},\ldots,y_{21K},y_{221},y_{22K})'$$

$$\beta = (\mu, a_1, b_1, ab_{11})', a_1 = -b_2, ab_{11} = ab_{22} = -ab_{12} = -ab_{21}$$

4) Optimalität
- Anwendungsbeispiel -

Mit der Designmatrix

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

folgt für die Kovarianzmatrix: $Q = 4 * K * I_4$

Wir können also die obige Theorie anwenden.

4) Optimalität
- Anwendungsbeispiel -

Fall 1:

Modellsystem: $\{M_{1000}, M_{1100}, M_{1010}, M_{1001}, M_{1101}, M_{1011}, M_{1110}, M_{1111}\}$
Im Rahmen des multiplen Testens mit $H_\gamma = \beta_{\gamma_j} = 0, j = 1, \ldots, k$:
$\{H_{123}, H_{23}, H_{13}, H_{12}, H_1, H_2, H_3\}$.
D.h. wir testen, ob die indizierten Parameter von $\tilde{\beta} = (a_1, b_1, ab_{11})$ gleich Null sind oder nicht.

Fall 2:
Annahme: Interaktion nur dann möglich, wenn beide Haupteffekte vorhanden sind.

Modellsystem: $\{M_{1000}, M_{1100}, M_{1010}, M_{1110}, M_{1111}\}$

Dieses Modell besitzt graphische Modellstruktur.

In beiden Fällen bietet das Median probability model die beste Datenvorhersage.

5) Fazit

Es werden starke Bedingungen an das Median probability model gestellt, damit es eine optimale Vorhersage ermöglicht.
Daher hilft dieses Modell nicht immer.

Beispiel:

Modelle mit Intercept α und einer Prädiktorenvariable $x_i, i = 1, \ldots k, k \geq 3$
sowie das Nullmodell.

Nehme als a priori Wahrscheinlichkeit 1/(k+1) für Modellordnung.
Weiterhin seien alle Kovariate hoch miteinander sowie mit y korelliert.

Damit ist die a posteriori Wahrscheinlichkeit für das Nullmodell fast gleich 0 und die für die anderen Modelle stimmen mit den a posteriori Inklusionswahrscheinlichkeiten überein. Diese sind aber kleiner als ½ und damit ist das beste Vorhersagemodell das Nullmodell.

6) Literatur

- Scott, James G., Berger, James O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem, The Annals of Statistics
- Barbieri, Maria M., Berger, James O.(2004). Optimal predictive model selection. The Annals of Statistics, Vol. 32, 870-897