Sonderdruck aus den „Jahrbüchern für Nationalökonomie und Statistik"
Band 189, Heft 3/4 (1975)
Gustav Fischer Verlag Stuttgart

Beobachtungen zur Ridge-Regression

Von Nanny Wermuth, Mainz

1. Einleitung

Insbesondere für stark multikollineare Regressoren wurde Ridge-Regression als Alternative zur Methode der kleinsten Quadrate vorgeschlagen. Wir beschreiben kurz inwiefern die Methode der kleinsten Quadrate schlechte Schätzungen geben kann, stellen dann Ridge-Regression als empirisches Bayes-Schätzverfahren dar sowie die Beziehungen zu der Methode der kleinsten Quadrate, der Regression auf Hauptkomponenten und den Stein-Schätzern, um schließlich die wichtigsten Ergebnisse einer Simulationsstudie hinsichtlich der Ridge-Regression zu erwähnen.

2. Multikollinearität

Im folgenden Abschnitt werden die Modellannahmen sowie das Problem der Multikollinearität beschrieben. Der Ökonometriker ist zweifellos mit dem Gesagten vertraut und der Abschnitt kann von dem nur an Ridge-Regression interessierten Leser übersprungen werden.

Es gelte für jede von n Beobachtungen

$$(1) \qquad Y^{(i)} = \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \ldots + \beta_p X_p^{(i)} + \varepsilon^{(i)} \qquad i = 1, \ldots, n$$

$$(2) \qquad \varepsilon^{(i)} \sim N(0,\sigma^2) \text{ für alle i, und } E\,(e^{(i)}\, e^{(j)}) = 0 \text{ für } i \neq j.$$

Das bedeutet, daß die Beziehung zwischen einer abhängigen Variablen und mehreren Regressoren annähernd linear sein soll. Die nichtbeobachteten Einflüsse auf die abhängige Variable sollen durch eine Zufallsgröße beschrieben werden können, die normalverteilt ist mit Erwartungswert Null und konstanter Streuung. Darüberhinaus seien die Störkomponenten für unterschiedliche Beobachtungen unabhängig voneinander. Das Modell kann in Matrix-Schreibweise also dargestellt werden als

$$(3) \qquad Y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n),$$

wobei der nx1 Vektor Y die Beobachtungen an der abhängigen Variablen enthält, X die nxp Datenmatrix der Regressoren ist, β der px1 Vektor der zu schätzenden Regressionskoeffizienten und ε der Vektor der Störkomponenten mit multivariater Normalverteilung wie in (3) angegeben.

Das wohl älteste und bestbekannte Verfahren, die Regressionskoeffizienten β zu schätzen, ist die Methode der kleinsten Quadrate, die als Schätzer

$$b = (X'X)^{-1}X'Y \tag{4}$$

liefert. (Unter der Annahme normalverteilter Störkomponenten ist b zugleich der Maximum-Likelihood-Schätzer.) Die Inverse $(X'X)^{-1}$ und damit b sind nur dann definiert, wenn der Rang von X gleich p ist, was bedeutet, daß die Zahl der Regressoren kleiner sein muß als die Zahl der Beobachtungen, und daß keine exakte lineare Beziehung zwischen den Regressoren bestehen darf.

Der Schätzer b besitzt wünschenswerte Eigenschaften, da er unverzerrt ist und in der Klasse der unverzerrten Schätzer die kleinste Varianz hat. Dennoch können die Schätzergebnisse unbefriedigend sein, wenn die Regressoren stark multikollinear sind, das heißt, wenn die (oder einige der) Regressoren annähernd linear miteinander verbunden sind.

An dem einfachsten Beispiel von nur zwei Regressoren läßt sich das Problem bereits anschaulich darstellen. Die Regressoren seien so standardisiert, daß X'X eine Korrelationsmatrix ist. Der Korrelationskoeffizient zwischen X_1 und X_2 werde mit r bezeichnet, dann ist die Kovarianzmatrix des Schätzers b

$$\frac{\sigma^2}{1-r^2}\begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}. \tag{5}$$

Die Varianzen $\sigma^2/(1-r^2)$ sind somit um so größer, je größer der Korrelationskoeffizient ist; die Schätzung wird also ungenau bei hoher Korrelation. In diesem Fall sind b_1 und b_2 außerdem stark negativ korreliert, was bedeutet, daß das Vorzeichen eines der beiden Koeffizienten β_1 oder β_2 falsch geschätzt werden kann.

Bei mehr als zwei Regressoren können komplizierte, fast lineare Abhängigkeiten auftreten der Art, daß man nicht mehr notwendig von den Korrelationskoeffizienten allein auf das Vorhandensein des Problems schließen kann, wohl aber von den Eigenwerten der Korrelationsmatrix oder der Determinante (dem Produkt der Eigenwerte)[1]).

Eine Determinante nahe Null hat zur Folge, daß der Schätzer b die strukturelle Beziehung zwischen der abhängigen Variablen und den Regressoren nur ungenau oder sogar falsch abschätzt. Der erwartete quadrierte Fehler zeigt dies. Es gilt

$$E(\beta - b)'(\beta - b) = \Sigma\, \sigma^2/\lambda_i, \tag{6}$$

wobei λ_i die Eigenwerte von X'X sind. Für standardisierte Regressoren bedeutet (6), daß der erwartete quadrierte Fehler den Wert σ^2/p hat, wenn die Regressoren unkorreliert sind, dagegen ist der Fehler (bei gleichen β's) um so größer, je höher der Grad der Multikollinearität ist.

---

[1]) Bei nur zwei standardisierten Regressoren sind $\lambda_1 = 1 + r$, $\lambda_2 = 1 - r$.

Es sei noch erwähnt, daß die Vorhersage der abhängigen Variablen nicht in ähnlicher Weise beeinflußt wird. Für $\hat{Y} = Xb$ gilt

$$E \sum_{i=1}^{n} (\hat{Y}^{(i)} - E\, Y^{(i)})^2 = E\,(\beta - b)'X'X(\beta - b) = p\,\sigma^2 \tag{7}$$

Der erwartete quadrierte Vorhersagefehler ist somit unabhängig von der Struktur der Matrix X'X.

Zusammenfassend kann man sagen: es ist kennzeichnend für die Methode der kleinsten Quadrate, daß die Regressionskoeffizienten um so schlechter (im Sinne des quadrierten Fehlers) geschätzt werden, je abhängiger die Regressoren untereinander sind.

3. Ridge-Regression

Von Anwendungen der Regressionsanalyse in der Chemie motiviert, schlugen Hoerl und Kennard (1970) für multikollineare Regressoren

$$b^* = (X'X + kI)^{-1}X'Y \tag{8}$$

als Schätzer für β vor, wobei k eine positive Zahl ist. Die Autoren wählen k so, daß sich b^* stabilisiert, d. h. für ansteigende Werte von k wenig verändert.

b^* kann geschrieben werden als $(I + k(X'X)^{-1})^{-1}b$, wobei b der Schätzer nach der Methode der kleinsten Quadrate ist. Um zu verstehen, wie b bei Ridge-Regression verändert wird, nehme man eine Hauptachsentransformation vor. Es wird wieder angenommen, daß die Regressoren so standardisiert sind, daß X'X eine Korrelationsmatrix ist. Das bedeutet, daß die Regressoren in vergleichbaren Maßstäben gemessen werden (also die absolute Größe der β_i die Wichtigkeit der einzelnen Regressoren angeben), und daß die im folgenden dargestellte Transformation nicht von den ursprünglichen Maßeinheiten der Regressoren abhängt. Nach einer Hauptachsentransformation entspricht dem Modell (3):

$$Y = X^*\alpha + \varepsilon\,. \tag{9}$$

Y und ε sind unverändert, X^* sind die Hauptachsenregressoren und α deren Regressionskoeffizienten. Im einzelnen gilt

$$X^* = XC', \qquad \alpha = C\beta \tag{10}$$

und C ist eine orthogonale Matrix (also $C'C = I$), so daß

$$CX'XC' = X^{*\prime}X^* = \operatorname{diag}(\lambda, \ldots, \lambda_p) \tag{11}$$

eine diagonale Matrix mit den Eigenwerten oder Hauptkomponenten (λ_i) der Matrix X'X ist.

Mit der Methode der kleinsten Quadrate wird α geschätzt als

$$a^* = (X^{*\prime}X^*)^{-1}X^{*\prime}Y, \tag{12}$$

woraus folgt, daß

$$a_i \sim N(\alpha_i, \sigma^2/\lambda_i)\,. \tag{13}$$

Eine Hauptkomponente λ_i kann somit interpretiert werden einerseits als Varianz des Hauptachsenregressor X_i^*, und andererseits als Element der Varianz des Schätzers a_i. Eine sehr kleine Hauptkomponente λ_i bedeutet also, daß X_i^* nur wenig zur Gesamtvarianz der Regressoren $\Sigma\lambda_i = p$ beiträgt, und daß der Koeffizient α_i durch a_i nur ungenau geschätzt werden kann.

Dem Ridge-Regressions-Schätzer entspricht nach der Transformation

$$a^* = Cb^* = (X^{*\prime}X^* + kI_p)^{-1}X^{*\prime}Y \tag{14}$$

oder

$$a_i^* = \frac{\lambda_i}{\lambda_i + k}\, a_i. \tag{15}$$

Während die Komponenten des Ridge-Regressions-Schätzers b^* größer oder kleiner sein können als die entsprechenden b_i, werden alle Schätzer a_i verkleinert oder gedämpft. Jene Koeffizienten, die großen (grob gesagt: wichtigen) Hauptkomponenten entsprechen, werden kaum verändert; diejenigen, die kleinen λ's zugeordnet sind, werden dagegen stark gedämpft.

Da aus (13) und (15) folgt, daß

$$a_i^* \sim N\left(\frac{\lambda_i}{\lambda_i + k}\,\alpha_i, \quad \frac{\lambda_i}{(\lambda_i + k)^2}\,\sigma^2\right), \tag{16}$$

ist gleichzeitig die in Kauf genommene relative Verzerrung $\dfrac{E(a_i^*) - \alpha_i}{\alpha_i} = \dfrac{k}{\lambda_i + k}$ nur groß für kleine Hauptkomponenten.

Die bisherige Beschreibung des Ridge-Schätzers macht das Verhältnis zu der bisher für multikollineare Regressoren häufig verwendeten „Regression auf Hauptkomponenten" durchsichtig. Bei diesem Verfahren ist der Schätzer für β

$$\begin{aligned} b^{**} &= Ca^{**} && \text{mit} \\ a_i^{**} &\begin{cases} = a_i & \text{für große } \lambda_i \\ = 0 & \text{andernfalls} \end{cases} \end{aligned} \tag{17}$$

Es wird dabei mehr oder minder willkürlich entschieden, welche Hauptkomponenten groß genug sind. Während bei dieser Methode die Hauptachsen-Regressoren in nur zwei Gruppen aufgeteilt wurden, nämlich wichtige ($a_i^{**} = a_i$) und unwichtige ($a_i^{**} = 0$), wird durch Ridge-Regression jedes a_i unterschiedlich gedämpft, und nur in Extremfällen (etwa $\lambda_i = 0{,}00001$) wird ein Koeffizient praktisch gleich Null geschätzt.

Unter der Annahme austauschbarer a-priori-Normalverteilungen für die einzelnen Regressionskoeffizienten resultiert eine Bayes-Schätzer der

Form $b = (X'X + kI_p)^{-1}X'Y$, wie von Lindley (1972) gezeigt wurde. Eine zusätzliche Überlegung von Dempster (beschrieben in Wermuth (1972)) ergibt eine Methode, k zu schätzen: k wird nunmehr als Parameter der a-priori-Verteilung der Regressionskoeffizienten angesehen.

Aus

$$\begin{aligned} a_i \mid \alpha_i &\sim N(\alpha_i, \sigma^2/\lambda_i) \\ \alpha_i &\sim N(0, \sigma^2/k) \end{aligned} \tag{18}$$

folgt (wie im Anhang gezeigt wird) erstens die a-posteriori-Verteilung der α_i:

$$\alpha_i \mid a_i \sim N\left(\frac{\lambda_i}{\lambda_i + k} a_i, \quad \frac{\lambda_i + k}{\sigma^2}\right), \tag{19}$$

zweitens die unkonditionale Verteilung der a_i, die zur Schätzung von k verwendet werden kann:

$$a_i \sim N(0, \sigma^2/\lambda_i + \sigma^2/k). \tag{20}$$

Daraus ergibt sich als Summe von p quadrierten N(0,1)-Variablen eine χ^2-Verteilung mit p Freiheitsgraden

$$\Sigma \frac{a_i^2}{\sigma^2\left(\frac{1}{\lambda_i} + \frac{1}{k}\right)} \sim \chi_p^2. \tag{21}$$

Mittels der Momentenmethode kann k daher aus folgender Beziehung bestimmt werden:

$$\Sigma \frac{a_i^2}{\left(\frac{1}{\lambda_i} + \frac{1}{k}\right)} = p\sigma^2. \tag{22}$$

Rechentechnisch findet man k etwa durch Bisektion. Den kleinsten und größten Eigenwert der Korrelationsmatrix bezeichne man mit λ_{min} und λ_{max}, und $\widehat{\sigma^2}$ sei $(Y - Xb)'(Y - Xb)/(n - p)$. Der Wert von $\frac{1}{k}$, der die Bedingung (22) erfüllt (mit σ^2 ersetzt durch $\widehat{\sigma^2}$) liegt dann zwischen

$$\Sigma \frac{a_i^2/p}{\widehat{\sigma^2}} - \frac{1}{\lambda_{min}} \text{ und } \frac{\Sigma a_i^2/p}{\widehat{\sigma^2}} - \frac{1}{\lambda_{max}}. \tag{23}$$

Zur Beurteilung dieser Methode k zu schätzen ist folgendes zu sagen:

— k ist für jede gegebene Datenmenge eindeutig bestimmt (im Gegensatz zu Hoerl und Kennards (1970) Verfahren);

— k muß nicht durch Iterationen ermittelt werden (wie etwa bei Befolgung von Lindleys Vorschlag (1972));

— der (22) entsprechende Schätzer b^* schließt als Spezialfall den Stein-Schätzer ein, für den James und Stein (1961) bewiesen, daß sein qua-

drierter Fehler immer kleiner ist als der des Schätzers der kleinsten Quadrate.

Wie im ersten Abschnitt gezeigt wurde, sind bei starker Multikollinearität schlechte Schätzungen der Regressionskoeffizienten zu erwarten. Das James-und-Stein-Ergebnis bedeutet nun, daß es sogar im Falle von unkorrelierten Regressoren Schätzer gibt, die im Sinne des erwarteten quadrierten Fehlers besser sind als der Schätzer nach der Methode der kleinsten Quadrate. Im einzelnen wurde folgendes bewiesen (dargestellt in unseren Symbolen): für $p \geqslant 3$, σ^2 bekannt und für alle α_i gibt

$$(24) \qquad E \sum_{i=1}^{p} (\alpha_i - \left(1 - \frac{p\sigma^2}{\sum a_i^2}\right) a_i)^2 < E \sum_{i=1}^{p} (\alpha_i - a_i)^2 = p\,\sigma^2.$$

Wie oben behauptet, ergibt sich $1 - \frac{p\sigma^2}{\sum a_i^2}\, a_i$ aus (22) und (8), wenn σ^2 bekannt und alle $\lambda_i = 1$ sind. Letzteres bedeutet, daß alle Regressoren unkorreliert sind.

Für unbekanntes σ^2 oder multikollineare Regressoren gibt es unseres Wissens nach kein theoretisches Resultat, das (24) ähnlich ist, wohl aber die empirischen Ergebnisse einer Simulationsstudie (Wermuth 1972).

4. Simulationsergebnisse

Datenmengen des Modells (3) wurden gemäß faktorieller Versuchspläne so simuliert, daß insbesondere die Wirkung verschiedener Multikollinearitätsgrade und unterschiedlicher Wichtigkeit der Regressoren studiert werden konnten. Der empirische Bayes-Schätzer unter der Annahme austauschbarer a-priori-Normalverteilungen der Regressionskoeffizienten wird als „RIDGM" bezeichnet und war oben durch (8), (22) und (23) gekennzeichnet. Als Maßstab zur Beurteilung eines Regressionsverfahrens wurden standardisierte quadrierte Fehler verwendet, also $(\beta - b)'(\beta - b)\sigma^2$ und $(\beta - b)'X'X(\beta - b)/\sigma^2$ (siehe dazu (6) und (7)).

Die Ergebnisse hinsichtlich RIDGM können wie folgt zusammengefaßt werden:

— Die Regressionskoeffizienten wurden mit RIDGM wesentlich besser geschätzt als mit dem Schätzer b nach der Methode der kleinsten Quadrate, insbesondere bei starker Multikollinearität; aber auch die Vorhersageergebnisse waren im allgemeinen besser.

— Während wir mit Regression auf die Hauptachsenkomponenten (17) häufig schlechtere Vorhersageergebnisse als mit der Methode der kleinsten Qadrate beobachteten, trat dieser Nachteil bei RIDGM nicht auf.

— Für sehr unterschiedlich wichtige Regressoren ergab RIDGM immer noch bessere Ergebnisse als b. Das deutet darauf hin, daß die Annahme der Austauschbarkeit der a-priori-Verteilungen der β_i (und der α_i wie in (18)) für viele Anwendungen vertretbar sein wird. Allerdings lieferte

im Falle sehr unterschiedlich wichtiger Regressoren ein (im nächsten Beitrag beschriebener) Bayes-Ansatz zur Regressoren-Selektion bessere Schätzungen der Regressionskoeffizienten als RIDGM.

Anhang

Obwohl die Ergebnisse in (18) und (19) leicht aus allgemeineren in Raiffa und Schlaifer (1961) folgen, wird hier der Beweis mit unseren Symbolen gegeben.

Aus

$$\begin{aligned} a_i \mid \alpha_i &\sim N(\alpha_i,\ \sigma^2/\lambda_i) \\ \alpha_i &\sim N(0,\ \sigma^2/k) \end{aligned}$$

folgt, daß

$$\begin{aligned} f(a, \alpha) &\propto \exp\left[-\frac{1}{2\sigma^2}\Sigma(a_i-\alpha_i)^2\ \lambda_i\right] \\ &\times \exp\left[-\frac{1}{2\sigma^2}\Sigma k\,\alpha_i^2\right] \\ &= \exp\left[-\frac{1}{2\sigma^2}\Sigma\left((k+\lambda_i)\alpha_i^2 - 2(k+\lambda_i)\frac{\lambda_i a_i}{\lambda_i+k}\alpha_i\right.\right. \\ &\qquad \left.\left.+(k+\lambda_i)\frac{\lambda_i^2 a_i^2}{(\lambda_i+k)^2}\right)\right] \\ &\times \exp\left[\frac{1}{2\sigma^2}\Sigma\left(\lambda_i a_i^2 - \frac{\lambda_i}{(\lambda_i+k)}\lambda_i a_i^2\right)\right] \\ &= \exp\left[\frac{1}{2\sigma^2}\Sigma(k+\lambda_i)\left(\alpha_i - \frac{\lambda_i}{\lambda_i+k}\alpha_i\right)^2\right] \\ &\times \exp\left[\frac{1}{2\sigma^2}\Sigma\frac{k\lambda_i}{\lambda_i+k}a_i^2\right]. \end{aligned}$$

Die gemeinsame Verteilung ist somit faktorisiert in folgende

$$\begin{aligned} \alpha_i \mid a_i &\sim N\left(\frac{\lambda_i}{\lambda_i+k}a_i,\ \frac{\sigma^2}{\lambda_i+k}\right] \\ a_i &\sim N\left[0, \frac{\lambda_i+k}{k\lambda_i}\sigma^2\right], \end{aligned}$$

wie zu zeigen war.

Zusammenfassung

Ein Schätzer der Form $b^* = (X'X + kI_p)^{-1}X'Y$ (Ridge-Regression) wird als empirischer Bayes-Schätzer dargestellt, und die Beziehungen zu bisher üblichen Schätzverfahren im linearen Regressionsmodell werden aufgezeigt. Das Verfahren

ist von praktischer Bedeutung, da es, wie eine Simulationsstudie ergab, selbst im Falle von stark multikollinearen Regressoren gute Schätzungen der Regressionskoeffizienten liefert.

Summary

In the case of highly multicollinear regressors the estimation procedure called ridge regression yields excellent point estimates of the regression coefficients. In particular, an empirical Bayes variant of ridge regression proved to be superior to least squares estimation in an extensive simulation study. We describe briefly in which sense least squares estimation can give poor estimates and discuss how ridge regression estimators relate to those obtained by least squares estimation regression on principal components, and Stein-type estimation.

Referenzen

Dempster, A. P.: Elements of Continuous Multivariate Analysis. Addison-Wesley, Reading, Mass. (1969).

Efron, B., and Morris, C.: Limiting the Risk of Bayes and Empirical Bayes Estimators, Part II: The Empirical Bayes Case. J. Am. Stat. Assoc., Vol. 67, 130—139 (1972).

Hoerl, A. E., and Kennard, R. W.: Ridge-Regression: Biased Estimation for Nonorthogonal problems. Technometrics, Vol. 8, 27—51 (1970).

James, W., and Stein, C.: Estimation with Quadratic Loss. Proceedings of the Fourth Berkeley Symposium, Vol. 1. University of California Press (1961).

Lindley, D. V., and Smith, A. F.: Bayes Estimates for the Linear Model. J. Royal Stat. Soc., Vol. 34, 1—46 (1972).

Raiffa, H., and Schlaifer, R.: Applied Statistical Decision Theory. Harvard Business School, Cambridge, Mass. (1961).

Wermuth, N.: An Empirical Comparison of Regression Methods. Ph. D. thesis, Department of Statistics, Harvard University, Cambridge, Mass. (1972).

Dr. Nanny Wermuth, Institut für Medizinische Statistik und Dokumentation der Universität Mainz, 6500 Mainz, Langenbeckstr. 1.