

# Ridge und Shrinkage Schätzer

Seminar: Ausgewählte Kapitel der Regressionsanalyse  
Prof. Dr. Götz Trenkler

Christoph Nöllenheidt, Sebastian Appelbaum

20. November 2008

- 1 Einleitung Motivation
- 2 Methoden
  - Ridge Schätzer
  - Shrinkage Schätzer
- 3 Zusammenfassung
- 4 Literatur

# Klassische lineare Regressionsmodell

- Modellgleichung:  $\mathbf{y} = \mathbf{X}\beta + \epsilon$
- (i)  $\mathbf{X}$  ist eine nicht-stochastische  $(n \times p)$ -Matrix mit  $p < n$
- (ii) Die Matrix  $\mathbf{X}$  hat Rang  $p$ , d.h.  $\mathbf{X}$  besitzt vollen Spaltenrang
- (iii)  $\mathbf{y}$  ist ein beobachtbarer  $(n \times 1)$ -dimensionaler Zufallsvektor
- (iv)  $\epsilon$  ist ein  $(n \times 1)$ -dimensionaler, nicht beobachtbarer Zufallsvektor, mit Erwartungswert  $E(\epsilon) = 0$   
und Kovarianzmatrix  $\text{Cov}(\epsilon) = \sigma^2 \mathbf{I}_n$ ,  $\sigma^2 > 0$

# Multikollinearität

- Häufig in der Ökonometrie
- Schlechte Datenlage (Multikollinearität)
- Einzelne oder mehrere Spalten der Designmatrix  $\mathbf{X}$  sind korreliert
- Einzelne oder mehrere Spalten der Designmatrix  $\mathbf{X}$  sind fast linear abhängig
- Mindestens ein Eigenwert von  $\mathbf{X}'\mathbf{X}$  liegt nahe 0
- Annahme (ii) ist nur formal erfüllt

# Beispiel 1

- Regressionsgerade  $\mathbf{y} = \mathbf{X}\beta + \epsilon$  mit

$$\mathbf{y} = \begin{pmatrix} 6.0521 \\ 7.0280 \\ 7.1230 \\ 4.4441 \\ 5.0813 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1.9 \\ 1 & 2.1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 1.8 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

- $\mathbf{y}$  simuliert mit  $\beta = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$  und  $\epsilon$  normalverteilt mit  $E(\epsilon) = 0$  und  $\text{Var}(\epsilon) = 1$

# Beispiel 1

- $\mathbf{X}'\mathbf{X} = \begin{pmatrix} 5 & 9.8 \\ 9.8 & 19.26 \end{pmatrix}$  und  $(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 74.0769 & -37.6923 \\ -37.6923 & 19.2308 \end{pmatrix}$
- Eine geringe Veränderung in  $\mathbf{X}$   
beispielsweise

$$\tilde{\mathbf{X}} = \begin{pmatrix} 1 & 1.9 \\ 1 & 2.05 \\ 1 & 2 \\ 1 & 2 \\ 1 & 1.85 \end{pmatrix}$$

führt zu,

$$\tilde{\mathbf{X}}'\tilde{\mathbf{X}} = \begin{pmatrix} 5 & 9.8 \\ 9.8 & 19.235 \end{pmatrix} \quad \text{und} \quad (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} = \begin{pmatrix} 142.4815 & -72.5926 \\ -72.5926 & 37.0370 \end{pmatrix}$$

# Kondition einer Matrix

- Matrix mit diesem Verhalten heißt schlecht konditioniert
- Eine Matrix  $\mathbf{X}$  heißt schlecht konditioniert, falls geringe Veränderungen in  $\mathbf{X}$  zu großen Veränderungen in  $(\mathbf{X}'\mathbf{X})^{-1}$  führen
- Maßzahl für Multikollinearität  
Konditionszahl:

$$\kappa(\mathbf{A}) = \sqrt{\frac{\lambda_{\max}(\mathbf{A}'\mathbf{A})}{\lambda_{\min}(\mathbf{A}'\mathbf{A})}}, \quad \kappa(\mathbf{A}) \in [1, \infty)$$

# Kleinste Quadrate Schätzer

- $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
- $\text{Cov}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
- Erwartete quadrierte Länge des Schätzers:

$$E(\|\hat{\beta}\|^2) = \sigma^2 \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}] + \|\beta\|^2$$

- Überschätzung der wahren Parameterlänge

# Beispiel 2

- Lineares Regressionsmodell  $\mathbf{y} = \mathbf{X}\beta + \epsilon$  mit Modellannahmen i) bis iv) und Dimension  $p = 2$
- $\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}, \quad -1 \leq r \leq 1$
- $r = 0$  und  $r = 0.9$
- $\text{Cov}(\hat{\beta}) = \frac{\sigma^2}{1-r^2} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}$
- $\text{Var}(\hat{\beta}_1 + \hat{\beta}_2) = \text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) + 2\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$

# Beispiel 2 Varianzvergleich

| Schätzer                        | Varianz     |                |
|---------------------------------|-------------|----------------|
|                                 | $r = 0$     | $r = 0.9$      |
| $\hat{\beta}_1$                 | $\sigma^2$  | $5.26\sigma^2$ |
| $\hat{\beta}_2$                 | $\sigma^2$  | $5.26\sigma^2$ |
| $\hat{\beta}_1 + \hat{\beta}_2$ | $2\sigma^2$ | $1.25\sigma^2$ |

- Multikollinearität führt nicht zwangsläufig zu unpräzisen Schätzungen

# Ridge Schätzer

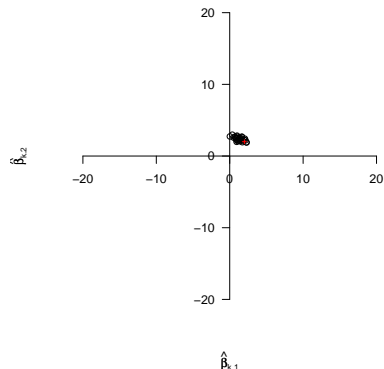
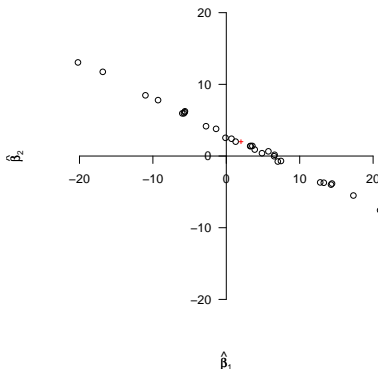
- Eingeführt von Hoerl und Kennard (1970)
- Idee: Kondition der Matrix  $\mathbf{X}'\mathbf{X}$  verbessern
- Definition:  $\hat{\beta}_k = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{y}$ ,

mit Ridge Parameter  $k \geq 0$

- (Beispiel 1)  $\hat{\beta} = \begin{pmatrix} -4.2489 \\ 5.2013 \end{pmatrix} \quad \hat{\beta}_{1/5} = \begin{pmatrix} 0.9396 \\ 2.5349 \end{pmatrix}$

# Graphische Veranschaulichung

- Vergleich KQ und Ridge Schätzer mit Daten aus Beispiel 1, Ridge Parameter  $k = 1/5$



# Eigenschaften Ridge Schätzer

- homogener linearer Schätzer, falls  $k$  fest
- Erwartungswert  $E(\hat{\beta}_k) = \beta - k(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}\beta$   
mit wachsendem  $k$  größerer Bias
- Kovarianzmatrix  $Cov(\hat{\beta}_k) = \sigma^2\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-2}$   
mit wachsendem  $k$  kleinere Varianz
- Fazit: Zielkonflikt

# Eigenschaften Ridge Schätzer

- Die Gesamtvarianz des Ridge Schätzers:

$$TV(\hat{\beta}_k) = tr(Cov(\hat{\beta}_k)) = \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^2}$$

- Die Gesamtvarianz des KQ-Schätzers:

$$TV(\hat{\beta}) = tr(Cov(\hat{\beta})) = \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i} = \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{\lambda_i^2}$$

- Fazit: Ridge Schätzer besitzt kleinere Gesamtvarianz

# Eigenschaften Ridge Schätzer

- betrachte quadratische Länge:

## Theorem 1

Ungleichung  $\|\hat{\beta}_{k_2}\| < \|\hat{\beta}_{k_1}\|$  ist für  $0 \leq k_1 < k_2$  erfüllt

- streng monoton fallende quadratische Länge
- kürzere quadratische Länge gegenüber KQ Schätzer

# Gütekriterium $MSE$

- $\tilde{\beta}$  beliebiger Schätzer für  $\beta$
- (1) die Matrix der mittleren quadratischen Fehler (Matrixrisiko)

$$MSE(\beta, \tilde{\beta}) = E[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)']$$

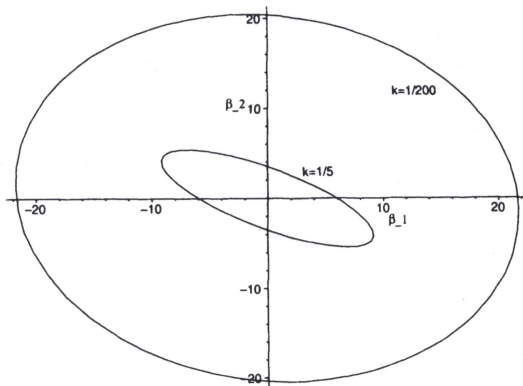
- (2) den reellwertigen mittleren quadratischen Fehler (Risiko)

$$\rho(\beta, \tilde{\beta}) = \text{tr}(MSE(\beta, \tilde{\beta}))$$

# Risikovergleich gegenüber KQ Schätzer

- matrixwertiger Vergleich
- betrachte Differenzenmatrix  $\Delta = MSE(\beta, \hat{\beta}) - MSE(\beta, \hat{\beta}_k)$
- Verbesserungsregionen des Ridge Schätzer gegenüber KQ Schätzer
- Gestalt: Ellipsoid im  $\mathbb{R}^p$

# Risikovergleich gegenüber KQ Schätzer



- Verbesserungsregionen von  $\hat{\beta}_k$  gegenüber  $\hat{\beta}$  bezüglich des MSE

# Risikovergleich gegenüber KQ Schätzer

## Theorem 2

Seien im linearen Regressionsmodell die Annahmen (i) bis (iv) erfüllt und  $k > 0$  eine feste Zahl

Die Differenzenmatrix

$$\Delta = \text{MSE}(\beta, \hat{\beta}) - \text{MSE}(\beta, \hat{\beta}_k)$$

ist genau dann nicht-negativ definit, wenn die Ungleichung

$$\beta' \left[ \frac{2}{k} \mathbf{I}_p + (\mathbf{X}'\mathbf{X})^{-1} \right]^{-1} \beta \leq \sigma^2$$

gilt

# Risikovergleich gegenüber KQ Schätzer

## Korollar 1

Seien im linearen Regressionsmodell die Annahmen (i) bis (iv) erfüllt.

Falls die Parameter ( $\beta \neq 0, \sigma^2 > 0$ ) die Ungleichung

$$0 < k \leq \frac{2\sigma^2}{\beta'\beta}$$

erfüllen, so ist  $\Delta = MSE(\beta, \hat{\beta}) - MSE(\beta, \hat{\beta}_k)$  nicht-negativ definit

# Wahl des Ridge Parameters $k$

- Zentraler Aspekt: Wahl des Parameters  $k$
- Zielkonflikt: Kompromiss zwischen kleiner Varianz  
und geringem Bias
- Zwei verschiedene Methoden, ein  $k$  zu finden

# Die Ridge Spur

- Graphisches Verfahren zur Bestimmung von  $k$
- Komponenten des Schätzers in Abhängigkeit von  $k$
- Stabilisierung der Funktionen ab einem bestimmten Wert
- Subjektive Methode

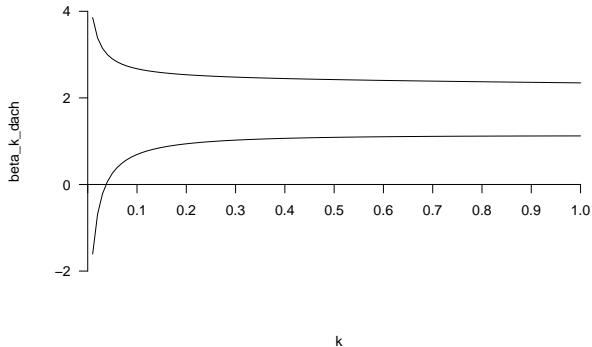
# Die Ridge Spur Beispiel 1

- Regressionsgerade  $\mathbf{y} = \mathbf{X}\beta + \epsilon$  mit

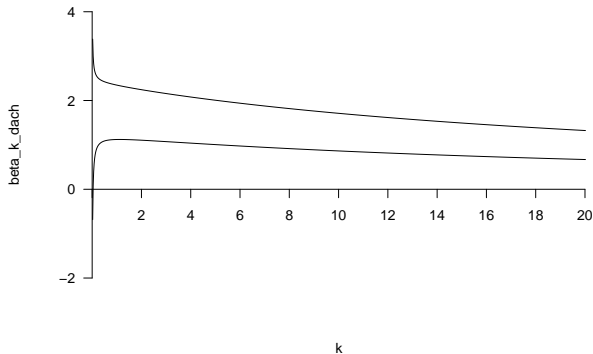
$$\mathbf{y} = \begin{pmatrix} 6.0521 \\ 7.0280 \\ 7.1230 \\ 4.4441 \\ 5.0813 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1.9 \\ 1 & 2.1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 1.8 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

- betrachte  $\hat{\beta}_k = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_2)^{-1}\mathbf{X}'\mathbf{y}$  und fasse  
beide Komponenten von  $\hat{\beta}_k$  als Funktionen von  $k$  auf

# Die Ridge Spur für $k \in [0, 1]$



# Die Ridge Spur für $k \in [0, 20]$



# Objektive Bestimmung des Parameters

- Datenbasierte Bestimmung von  $k$
- Optimalitätskriterium:  
reellwertiger mittlerer quadratischer Fehler
- Optimalitätseigenschaft:  $\rho(\beta, \hat{\beta}_{k_{opt}}) \leq \rho(\beta, \hat{\beta}_k)$ , für jedes  $k \geq 0$
- Für jedes  $(\beta \neq 0, \sigma^2)$  existiert ein  $k_{opt}$

# Objektive Bestimmung des Parameters

## Theorem 3

Seien im linearen Regressionsmodell die Annahmen (i) bis (iv) und  $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$  erfüllt.

Die Ungleichung

$$\rho(\beta, \hat{\beta}_{k_{opt}}) \leq \rho(\beta, \hat{\beta}_k), \forall k \geq 0$$

ist wahr für

$$k_{opt} = \frac{p\sigma^2}{\beta'\beta}$$

# Objektive Bestimmung des Parameters

- Betrachte ad-hoc Schätzer

$$\hat{k}_{HKB} = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$$

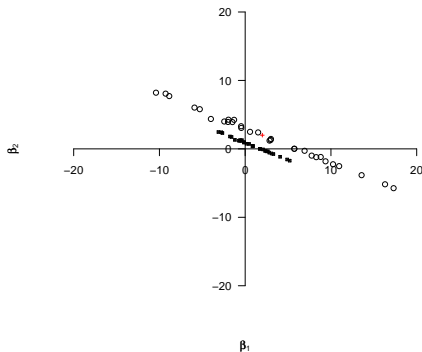
- Nachteil:  $\hat{k}_{HKB}$  oftmals zu klein
- Lösung: iteratives Schätzverfahren nach Hoerl und Kennard

# Shrinkage Schätzer

- Eingeführt von Mayer und Willke
- Auch Kontraktionsschätzer genannt
- Idee: Überschätzung der Länge des wahren Parameters  $\beta$  durch KQ Schätzer korrigieren
- Definition:  $\hat{\beta}(\varrho) = \frac{1}{1+\varrho}\hat{\beta}$ ,  
mit Shrinkage Parameter  $\varrho \geq 0$

# Graphische Veranschaulichung

- Vergleich KQ und Shrinkage Schätzer mit Daten aus Beispiel 1, Shrinkage Parameter  $\varrho = 7/3$



# Eigenschaften Shrinkage Schätzer

- homogener linearer Schätzer, falls  $\varrho$  fest
- Erwartungswert  $E(\hat{\beta}(\varrho)) = \frac{1}{1+\varrho}\beta$
- Kovarianzmatrix  $Cov(\hat{\beta}(\varrho)) = \frac{\sigma^2}{(1+\varrho)^2}(\mathbf{X}'\mathbf{X})^{-1}$
- Länge der Schätzung  $E(\|\hat{\beta}(\varrho)\|^2) = (\frac{1}{1+\varrho})^2\sigma^2 tr[(\mathbf{X}'\mathbf{X})^{-1}] + (\frac{1}{1+\varrho})^2\|\beta\|^2$
- Schrumpfung Richtung Nullvektor

# Risikovergleich gegenüber KQ Schätzer

- Betrachtung des Matrixrisikos und der Differenzenmatrix  $\Delta$
- Verbesserungsregion des Shrinkage Schätzers gegenüber KQ Schätzer
- Gestalt: Ellipsoid im  $\mathbb{R}^p$
- Ellipsoid vergrößert sich für wachsendes  $\varrho$

# Risikovergleich gegenüber KQ Schätzer

## Theorem 4

Unter den Annahmen i) bis iv) des linearen Regressionsmodells sei  $\varrho > 0$  eine (nicht-stochastische) Zahl. Dann ist die Differenzenmatrix

$$\Delta = \text{MSE}(\beta, \hat{\beta}) - \text{MSE}(\beta, \hat{\beta}(\varrho)),$$

genau dann nicht-negativ definit, wenn

$$\beta' \mathbf{X}' \mathbf{X} \beta \leq \frac{\varrho + 2}{\varrho} \sigma^2,$$

wobei  $\Delta \neq \mathbf{0}$ ,  $\varrho > 0$  und  $p > 1$

# Risikovergleich gegenüber KQ Schätzer

## Korollar 2

Unter den Annahmen i) bis iv) des linearen Regressionsmodells und den Parametern ( $\beta \neq \mathbf{0}, \sigma^2$ ) ist die Differenzenmatrix

$$\Delta = \text{MSE}(\beta, \hat{\beta}) - \text{MSE}(\beta, \hat{\beta}(\varrho)),$$

nicht-negativ definit, falls die Ungleichung

$$0 < \varrho \leq \frac{2\sigma^2}{\beta' \mathbf{X}' \mathbf{X} \beta}$$

erfüllt ist

# Bestimmung des Shrinkage Parameters

## Theorem 5

Unter den Annahmen i) bis iv) des linearen Regressionsmodells für  $(\beta \neq \mathbf{0}, \sigma^2)$  ist die Ungleichung

$$\rho(\beta, \hat{\beta}(\varrho_{opt})) \leq \rho(\beta, \hat{\beta}(\varrho)) \quad \forall \varrho \geq 0,$$

wahr für

$$\varrho_{opt} = \sigma^2 \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}] / \beta' \beta$$

# Bestimmung des Shrinkage Parameters

- Theorem 5 liefert Idee für Schätzer
- Ersetze wahre Parameter durch erwartungstreue Schätzungen
- Schätzstatistik:  $\hat{\varrho} = \hat{\sigma}^2 \frac{\text{tr}[(\mathbf{X}'\mathbf{X})^{-1}]}{\hat{\beta}'\hat{\beta}}$
- Kein linearer Schätzer

# Zusammenfassung

- Ergebnis: Eignung beider Schätzer bei Multikollinearität
- Varianten beider Schätzer:
  - Generalisierte Ridge Schätzer
  - Richtungsmodifizierte Shrinkage Schätzer
- weitere Möglichkeit: Hauptkomponentenschätzer bei Multikollinearität

# Literatur

-  Hackl, P. (2005): *Einführung in die Ökonometrie*. Pearson Studium, München.
-  Groß, J., (2003): *Linear Regression*. Springer-Verlag, Berlin Heidelberg New York.
-  Gruber, M.H.J., (1998): *Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators*. Marcel Dekker, New York.
-  Hoerl, A.E., Kennard, R.W. (1970): Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.
-  Hoerl, Kennard (1976): Ridge regression iterative estimation of the biasing parameter. *Communications in Statistics, Theory and Methods* **5**, 77-88.

# Literatur



Hoerl, A.E., Kennard, R.W., Baldwin, K.F. (1975): Ridge regression: some simulations. *Communications in Statistics, Theory and Methods* **4**, 105-123.



Mayer, L.S., Willke, T.A. (1973): On biased estimation in linear models. *Technometrics* **15**, 497-508.



Ohtani, K. (1996): On an adjustment of degrees of freedom in the minimum mean squared error estimator. *Communications in Statistics, Theory and Methods* **25**, 3049-3058.



Schmidt, K., Trenkler, G. (2006). *Einführung in die moderne Matrix-Algebra*, Springer-Verlag, Berlin Heidelberg.

Vielen Dank für Ihre Aufmerksamkeit!