Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

CAS Datenanalyse
Modul Regressionsanalyse
Kapitel 2: Variablenauswahl

Prof. Dr. Raúl Gimeno
FRM,CAIA, PRM

Inhalt

- ✓ Auslassen relevanter Variablen
- ✓ Verwendung irrelevanter Variablen
- ✓ Adjustiertes Bestimmtheitsmass

Variablenauswahl

Zusammenhang zwischen Lohn und seinen Bestimmungsgrössen.
Drei konkurrierende Regressionsmodelle:

Modell 1: $y_i = \beta_1 + \beta_2 x_{2i} + u_i'$

Modell 2: $y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$ (korrektes Modell)

Modell 3: $y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i^*$

y_i: Höhe des Lohnes x_2: Ausbildungszeit
x_3: Alter des Mitarbeiters x_4: Firmenzugehörigkeit

Schätzergebnisse

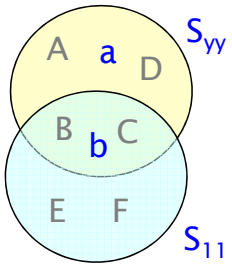
Mod	Variable	Koeff	se()	t-Wert	p-Wert
1	Konstante	1354.7	94.2	14.377	0.001
	Ausbildung	89.3	19.8	4.505	0.001
2	Konstante	1027.8	164.5	6.249	0.001
	Ausbildung	62.6	21.2	2.953	0.009
	Alter	10.6	4.6	2.317	0.033
3	Konstante	1000.5	225.7	4.432	0.001
	Ausbildung	62.4	21.8	2.859	0.011
	Alter	12.4	10.7	1.159	0.263
	Firmenzuge.	-2.6	14.3	-0.183	0.857

- Parameter β_2 (Ausbildung) und β_3 (Alter): positives Vorzeichen
- Parameter β_4 (Firmenzugehörigkeit): negatives Vorzeichen
- H_0: $\beta_4 = 0$ kann auf Signifikanzniveau von 5% nicht abgelehnt werden.

Bestimmtheitsmass und Venn-Diagramme

Einfachregression (1)

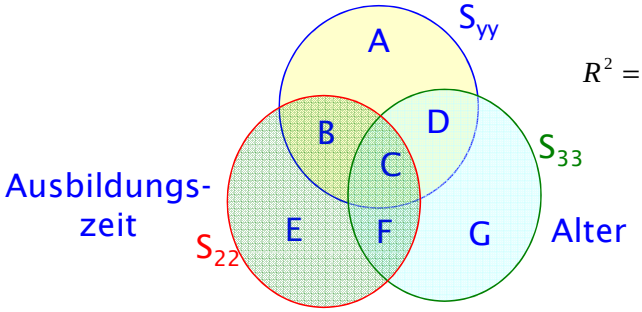
$y_t = \beta_1 + \beta_2 x_{2t} + u'_t$



$$R^2 = \frac{b}{a+b}$$

Zweifachregression (2)

$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$



$$R^2 = \frac{B+C+D}{A+B+C+D}$$

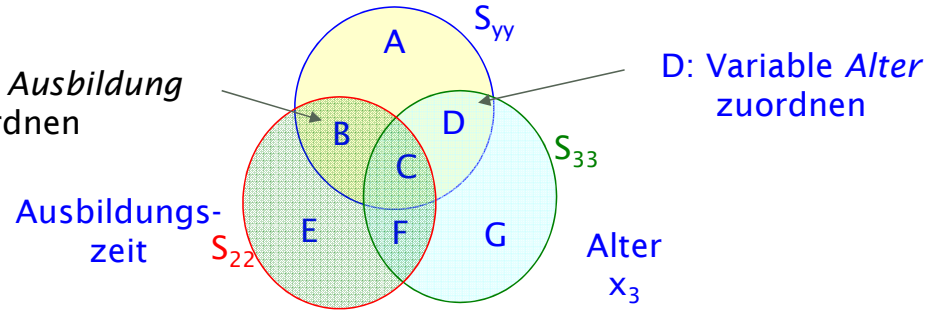
- Oberer Kreis: Variation der endogenen Variable → $S_{yy} = \sum_{i=1}^{N} (y_i - \bar{y})^2$
- Untere Kreise: Variation der exogenen Variablen $S_{kk} = \sum_{i=1}^{N} (x_{ki} - \bar{x}_k)^2$
- Überschneidungsfläche: Entsprechende Variablen korrelieren miteinander
- Je höher die Korrelation zwischen zwei Variablen, desto grösser die Überschneidungsfläche.
- S_{yy} : Variation von y, die auf die exogenen Variablen zurückzuführen ist
  (1) Fläche b (2) Fläche B+C+D
- S_{ee}: unerklärte Variation (Störeinflüsse)
  (1) Fläche a (2) Fläche A

Wirkungszusammenhang

Regressionsmodell: $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$

B: Variable *Ausbildung* zuordnen

D: Variable *Alter* zuordnen



- Einflussfläche C: keine eindeutige Zuordnung
- Falls vollständig dem Alter (x_3) zurechnen → b_3 überschätzt
- Falls vollständig der Ausbildung (x_2) zurechnen → b_2 überschätzt
- Modell 1: $y_i = \beta_1 + \beta_2 x_{2i} + u'_i$ berücksichtigt fälschlicherweise die relevante Variable x_3 *Alter* nicht → Kreis S_{33} wird nicht wahrgenommen → b_2 überschätzt

Auswirkungen auf den Erwartungswert der Störgrösse

- Modell 1: $y_i = \beta_1 + \beta_2 x_{2i} + u'_i$ unvollständiges Modell
- Ausgelassene relevante Variable: Alter (x_3)
- Störgrösse u' → Einfluss der wahren Störgrösse u
  → Einfluss der ausgelassenen Variable x_3

$$u'_i = \beta_3 x_{3i} + u_i \qquad E(u'_i) = E(\beta_3 x_{3i} + u_i) = \beta_3 x_{3i} + 0 \neq 0$$

- Annahme A5 verletzt: E(u) = 0
- Konsequenz für Punktschätzer: $b'_2 = b_2 + b_3 \frac{S_{23}}{S_{22}}$
- b_2 und b_3: Punktschätzer auf Basis des korrekten Modells 2

$$E(b'_2) = E\left(b_2 + b_3 \frac{S_{23}}{S_{22}}\right) = \beta_2 + \beta_3 \frac{S_{23}}{S_{22}}$$

- Schätzer b'_2 ist verzerrt → Verzerrungsterm = $\beta_3(S_{23}/S_{22})$

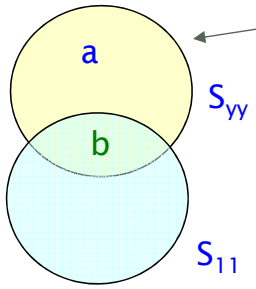
Scheinbarer Einfluss

- Regression x_3 Alter auf exogene Variable x_2: $x_3 = b_1 + b_2 x_2 + u$
- Einfachregression: Schätzer $b_2 = S_{23}/S_{22}$
- Quotient S_{23}/S_{22} quantifiziert Einfluss der Ausbildung auf das Alter
- Scheinbarer Einfluss der Ausbildung auf Höhe des Lohnes (keine Kausalität)
- Variablen Ausbildung und Alter sind positiv korreliert
- $S_{23} > 0 \rightarrow S_{23}/S_{22} > 0$
- Verzerrungsterm: $\beta_3 \frac{S_{23}}{S_{22}}$ indirekter scheinbarer Einfluss der Ausbildung auf die Lohnhöhe
- Lohnbeispiel: $\frac{S_{23}}{S_{22}} = \frac{448.9}{178.2} = 2.52$
- Modelle 1: $b'_2 = b_2 + b_3 \frac{S_{23}}{S_{22}} = 62.6 + 10.6 \cdot 2.52 = 89.3$

Bestimmtheitsmass und Venn-Diagramme

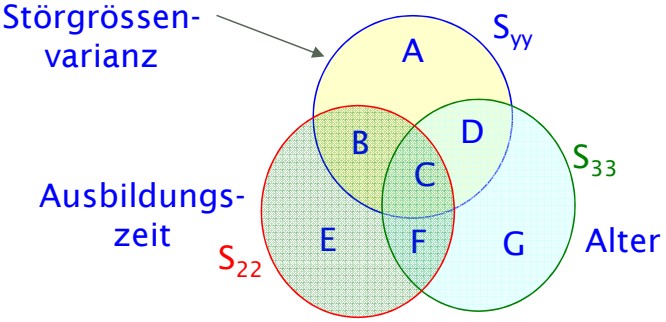
Einfachregression (1)

$y_t = \beta_1 + \beta_2 x_{2t} + u'_t$



Zweifachregression (2)

$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$



Korrektes Modell 2:

- Summe der Residuenquadrate: $S_{ee} = \sum_{i=1}^{N} (e_i - \bar{e})^2 = \sum_{i=1}^{N} e_i^2 = 957'698$
- Unverzerrte geschätzte Störgrössenvarianz: $s_e^2 = \frac{S_{ee}}{N-3} = 56'335$

Unvollständiges Modell 1:

Summe der Residuenquadrate: $S_{e'e'} = \sum_{i=1}^{N} (e'_i - \bar{e}')^2 = \sum_{i=1}^{N} e'^2_i = 1'260'028$

Unverzerrte geschätzte Störgrössenvarianz: $s^2_{e'} = \frac{S_{e'e'}}{N-2} = 70'001$

Auswirkungen

Konsequenzen für Intervallschätzer des unvollständigen Modells 1

- Intervallschätzer für b_2 (Ausbildung): $\left[b'_2 - t_{c,\alpha/2,df} \cdot se(b'_2), b'_2 + t_{c,\alpha/2,df} \cdot se(b'_2)\right]$

Verzerrtes Intervall:

- Zentrum b'_2 liegt bei wiederholten Stichproben im Mittel nicht auf den wahren Wert b_2.
- Zu grosse Breite wegen $se(b'_2)$

Das Auslassen relevanter Variablen führt zu:

- verzerrten Punktschätzern
- verzerrten Intervallschätzern
- wertlosen Hypothesentests

Verwendung irrelevanter Variablen

- Wenn die Variable x_4 (Firmenzugehörigkeit) irrelevant ist, dann $\beta_4 = 0$

$$u_i^* = u_i - \beta_4 x_{4i}$$

$$E(u_i^*) = E(u_i) = 0$$

Konsequenzen für die Punktschätzer

- Unverzerrte Schätzer: $E(\beta_i^*) = \beta_i$ i = 1, 2, 3, 4
- Höhere Varianz der Schätzer: β_i^* i = 1,2,3,4
- Nicht effiziente Schätzer

Die Aufnahme irrelevanter Variablen führt zu:

- unverzerrten, aber ineffizienten Punktschätzern
- unverzerrten, aber ineffizienten Intervallschätzern
- verwendbaren, aber unscharfen Hypothesentests

Die Konsequenzen sind weit weniger gravierend als beim Auslassen relevanter Variablen.

Bestimmtheitsmass

Bestimmtheitsmass: $$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{S_{ee}}{S_{yy}}$$

Sinnvolles Kriterium wenn drei Bedingungen erfüllt sind:

1. Die endogene Variable der Modelle ist identisch
2. Die Anzahl der exogenen Variablen ist identisch
3. Die Modelle besitzen einen Niveauparameter (Interzept)

Lohnbeispiel: Bedingung 2 nicht erfüllt → R^2 nicht sinnvoll als Kriterium

R^2 erhöht sich durch zusätzliche Parameter → jede zusätzlich aufgenommene Variable verringert die Summe der Residuenquadrate S_{ee}, oftmals nur sehr geringfügig.

Modell	R^2	se(b_1)	se(b_2)	se(b_3)	se(b_4)
1	52.99	8.877	392.824		
2	64.27	164.473	21.19	4.576	
3	64.35	225.72	21.83	10.65	14.29

Adjustiertes Bestimmtheitsmass

Erweiterung eines Modells um einen Regressor: R^2 wird grösser
Zunahme von R^2 bedeutet nicht notwendigerweise, dass der neue Regressor zur Erklärung von y beiträgt!

Adjustiertes Bestimmtheitsmass: $\bar{R}^2 = 1 - \frac{N-1}{N-k}\frac{S_{ee}}{S_{yy}}$

Das Hinzufügen eines Regressors verkleinert den Quotienten RSS/TSS, vergrössert aber den Faktor (N–1)/(N–k). Mit wachsendem k wird der Faktor (N–1)/(N–k) grösser und kompensiert dafür, dass RSS tendenziell kleiner wird.

$\bar{R}^2$: Dient zum Vergleichen von konkurrierenden Regressionsmodellen mit unterschiedlicher Anzahl von exogenen Variablen

Zusammenhang: $\bar{R}^2 < R^2$

Bei grossem N ist (N–1)/(N–k) ≈ 1 und $R^2 \approx \bar{R}^2$

Adjustiertes Bestimmtheitsmass

Adjustiertes Bestimmtheitsmass: $\bar{R}^2 = 1 - \frac{N-1}{N-k}\frac{S_{ee}}{S_{yy}}$

R^2 erhöht sich durch zusätzliche Parameter, aber $\bar{R}^2$ nicht

Modell	K	N-K	(N-1)/(N-K)	R^2 (%)	$\bar{R}^2$ (%)
1	2	18	1.055	52.99	50.38
2	3	17	1.117	64.27	60.06
3	4	16	1.187	64.35	57.66

$R^2 > \bar{R}^2$

Zusammenhang: $\bar{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-K}$

Modell 2: $\bar{R}^2 = 1 - (1 - 0.6427)\frac{19}{17} = 0.6$