Empirische Wirtschaftsforschung

Der Maximum Likelihood Schätzer

Ulrich Fritsche

Universität Hamburg

Email: Ulrich.Fritsche@wiso.uni-hamburg.de

Die Idee des ML-Schätzers

Vergleich von OLS und ML:

- OLS-Schätzer: Welcher Schätzer minimiert die quadrierten Abweichungen der Schätzgeraden von den tatsächlichen Beobachtungen?
- ML-Schätzer: Welcher Schätzer erhöht die Wahrscheinlichkeit, dass die Verteilung der Schätzwerte möglichst nahe an der Verteilung der tatsächlichen Werten liegt?
- Vorteil der ML-Methode: Schätzung nichtlinearer Modelle
- Grundvoraussetzung für ML: Verteilung der Daten ist bekannt, Annahme der Normalverteilung

Schätzung einer Cobb-Douglas-Produktionsfunktion

$$Y_t = a \cdot K_t^{\alpha} \cdot L_t^{\beta} \tag{1}$$

Diese kann mit OLS nur nach Loglinearisierung geschätzt werden:

$$logY_t = a + \alpha logK_t + \beta logL_t \tag{2}$$

Eine Schätzung von Gleichung (1) mit Maximum Likelihood und (2) mit OLS ergibt[1]

	a	α	β	MSE
OLS	1.239	0.267	0.691	171.954
ML	-0.177	0.233	0.807	105.818

Einsetzen der geschätzten Koeffizienten in die jeweilige Gleichung ergibt die geschätzte Reihe für Y_t. Die Berechnung des **mittleren quadratischen Fehlers (MSE, mean squared error)** spricht für die Schätzung des nichtlinearen Modells.

[1] Daten aus Cobb/Douglas (1928): A Theory of Production, in: American Economic Review.

Beispiel 1

Gegeben ist eine Zufallsvariable X mit zwei Beobachtungen $X_1 = 4$ und $X_2 = 6$. X ist normalverteilt mit Standardabweichung $\sigma = 1$.
Was ist der Erwartungswert von X, d.h. $E(X) = \mu$?
Die Normalverteilung ist gegeben als:

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(X - E(X))^2}{\sigma^2}\right\} \tag{3}$$

Da wir $E(X)$ nicht kennen, nehmen wir einen Startwert an, z.B. $E(X) = 3.5$. Damit erhalten wir die Wahrscheinlichkeiten für $p(4) = 0.3521$ und $p(6) = 0.0175$.
Die Wahrscheinlichkeit für das gemeinsame Eintreten von 4 und 6 ergibt sich als $L = p(4) \cdot p(6)$, mit L als **Likelihoodfunction**.

Beispiel 1, cont'd

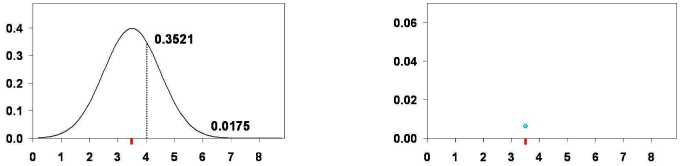

Abbildung: Einzelne und gemeinsame Wahrscheinlichkeitsfunktionen

μ	$p(4)$	$p(6)$	L
3.5	0.3521	0.0175	0.0062

Abbildung: Werte

Beispiel 1, cont'd

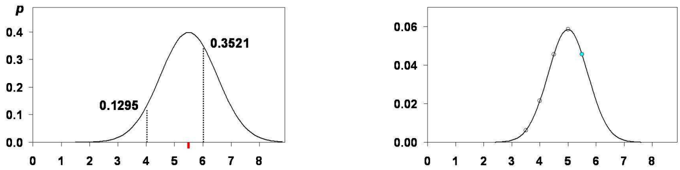

Abbildung: Einzelne und gemeinsame Wahrscheinlichkeitsfunktionen

μ	$p(4)$	$p(6)$	L
3.5	0.3521	0.0175	0.0062
4.0	0.3989	0.0540	0.0215
4.5	0.3521	0.1295	0.0456
5.0	0.2420	0.2420	0.0585
5.5	0.1295	0.3521	0.0456

Abbildung: Werte

Beispiel 2

Bernoulli-Verteilung: Eine diskrete Zufallsvariable X aus einer Grundgesamtheit T:

$$X = \begin{cases} 1 : \text{bestimmter Zustand tritt ein} \\ 0 : \text{bestimmter Zustand tritt nicht ein} \end{cases}$$

Wahrscheinlichkeit für $X = 1 : P(X = 1) = p \rightarrow \text{unbekannt}$
Wahrscheinlichkeit für $X = 0 : P(X = 0) = 1 - p = q$

Mit $T_1 = \sum_{i=1}^{T} y_i$ als Anzahl der Beobachtungen für $X = 1$ erhält man die gemeinsame Wahrscheinlichkeitsverteilung (Likelihood-Funktion) f

$$f(p; T_1) = p^{T_1}(1 - p)^{(T - T_1)} \tag{4}$$

Beispiel 2 cont'd

Logarithmieren von Gleichung (4) ergibt die **Loglikelihood-Funktion**

$$\log L(p; T_1) = T_1 \log p + (T - T_1) \log(1 - p) \tag{5}$$

Maximiere die Wahrscheinlichkeit p für den Eintritt von $X = 1$:

$$\frac{\partial \log L}{\partial p} = T_1 \frac{1}{\hat{p}} - (T - T_1) \frac{1}{1 - \hat{p}} \tag{6}$$

Auflösen nach $\hat{p}$ ergibt den Maximum-Likelihood-Schätzer:

$$\hat{p} = \frac{T_1}{T} \tag{7}$$

Idee des ML-Schätzers

1. Gegeben sind Daten T_1 als Stichprobe aus einer Grundgesamtheit T
2. Annahme: Wahrscheinlichkeitsverteilung der Daten ist bekannt, und abhängig vom unbekannten p
3. Frage: Wie hoch ist die unbekannte Wahrscheinlichkeit p, dass die angenommene Verteilung mit der tatsächlichen Verteilung übereinstimmt, dass also die Stichprobe möglichst exakt der wahren Verteilung entspricht?
4. Antwort: Wähle p so, dass die Wahrscheinlichkeit der Übereinstimmung maximal wird

Anwendung auf das lineare Regressionsmodell

Das tatsächliche lineare Schätzmodell aus dem 1.Teil der Vorlesung:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \tag{8}$$

Unter der Annahme normalverteilter Störterme

$$\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I})$$

kann man die Wahrscheinlichkeitsfunktion der Normalverteilung nutzen: Für eine Zufallsvariable X mit $X \sim N(E(X), \sigma^2)$ gilt

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(X - E(X))^2}{\sigma^2}\right\} \tag{3}$$

Mit $E(\mathbf{u}) = \mathbf{0}$ gilt $E(\mathbf{y}) = \mathbf{X}\beta$, und man kann für die Likelihood-Funktion schreiben

$$f(\beta;\sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^T} \exp\left\{-\frac{1}{2}\frac{(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)}{\sigma^2}\right\} \tag{9}$$

Durch Logarithmieren erhält man:

$$\log L(\beta;\sigma^2) = -\frac{T}{2}[\log\sigma^2 + \log(2\pi)] - \frac{1}{2\sigma^2}((\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)) \tag{10}$$

Den ML-Schätzer erhält man durch Ableitung von (10) nach β.

$$\frac{\partial \log L}{\partial \beta} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} = 0 \tag{11}$$

D.h., der ML-Schätzer für β entspricht genau dem OLS-Schätzer:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{12}$$

Ableitung von (10) nach σ^2 ergibt den Schätzer für die Varianz:

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{T}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = 0 \tag{13}$$

Auflösen führt zu

$$\hat{\sigma}^2_{ML} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{T} \tag{14}$$

Dies bedeutet, dass der ML-Schätzer der Varianz im Vergleich zu dem OLS-Schätzer einen Bias aufweist:

Aus dem 1.Teil der Vorlesung:

$$\hat{\sigma}^2_{OLS} = \frac{1}{T-m}\hat{\mathbf{u}}'\hat{\mathbf{u}} \tag{15}$$

Allerdings ist der ML-Schätzer **asymptotisch** erwartungstreu:

Für $T \rightarrow \infty$, geht die Differenz der Schätzer, $\frac{m}{(T-m)T}$ gegen Null.

Testprinzipien auf Basis des ML-Schätzers

Auf Basis des ML-Prinzips lassen sich eine Reihe von Testverfahren berechnen:
Schätze

$$\max_{\beta} \log L(\beta) \tag{16}$$

mit der Nullhypothese

$$H_0 : \mathbf{R}(\beta) = \mathbf{q} \tag{17}$$

mit $\mathbf{R}$ als $J \times m$ Matrix, $\mathbf{q}$ als Vektor der Restriktionen, m als Anzahl der Parameter und J als Anzahl der Restriktionen.

Beispiel

Test auf Nullrestriktion bei zwei Parametern, d.h. $m = J = 2$:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Das heißt, wir testen auf:

$$\begin{aligned} \beta_1 &= 0 \\ \beta_2 &= 0 \end{aligned}$$

All drei Testverfahren

- Likelihood-Ratio (LR) - Test
- Lagrange-Multiplier (LM) - Test
- Wald-Test

beruhen auf einem Vergleich des restringierten mit dem unrestringierten Modell. Asymptotisch sind all drei Tests Chi-Quadrat verteilt.

Der LR-Test

- Idee: Ist die *Differenz* der geschätzten Loglikelihood-Funktionen des restringierten und des unrestringierten Modells signifikant von Null verschieden?

$$\lambda = \frac{L(\tilde{\beta})}{L(\hat{\beta})} \tag{18}$$

- Dass heißt, man erhält die Teststatistik als

$$LR = -2\log\lambda = 2[\log L(\hat{\beta}) - \log L(\tilde{\beta})] \overset{a}{\sim} \chi^2(J) \tag{19}$$

- Wenn Differenz statistisch signifikant von Null verschieden, Ablehnung der Restriktion
- Lehne H_0 ab, wenn Differenz groß ist.

Der LM-Test

- Schätzung des Modells mit Restriktionen
- Test, ob *1.Ableitungen der Loglikelihood-Funktion signifikant von Null verschieden* sind, d.h., ob sich die Steigungen der Loglikelihood-Funktion signifikant unterscheiden
- Die Teststatistik ergibt sich somit als:

$$LM = \frac{\partial \log L'}{\partial \tilde{\beta}} [\mathbf{J}(\tilde{\beta})^{-1}] \frac{\partial \log L}{\partial \tilde{\beta}} \overset{a}{\sim} \chi^2(J) \tag{20}$$

  mit $\mathbf{J}^{-1}$ als *Informationsmatrix*, d.h. der Inversen der geschätzten Varianz-Kovarianz-Matrix des restringierten Modells.
- Lehne H_0 ab, wenn die Ableitung ungleich Null ist.
- LM-Test-Statistik als Basis für eine Reihe von Tests:
  - Test auf ausgelassene Variablen
  - Test auf Heteroskedastie (Breusch-Pagan-Test)
  - Test auf Autokorrelation (Breusch-Godrey-Test)

Der Wald-Test

- Der Wald-Test vergleicht die Schätzer des restringierten mit dem unrestringierten Modell. Da der ML-Schätzer erwartungstreu ist, sollte die Differenz Null betragen, wenn die Restriktion gültig ist.
  $\hat{\beta} - \tilde{\beta} = 0$
- Diese Differenz wird mit den *2.Ableitungen der Loglikelihood-Funktion* gewichtet, also der Krümmung. Man erhält die Teststatistik als:

$$W = (\mathbf{R}\tilde{\beta})[\mathbf{R}'[\mathbf{J}(\hat{\beta})]^{-1}\mathbf{R}](\mathbf{R}\tilde{\beta}) \overset{a}{\sim} \chi^2(J) \tag{21}$$

- Vorteil: Restringierte Schätzung muss nicht durchgeführt werden (da $H_0 : \tilde{\beta} = 0$)
- Nachteil: Testergebnisse abhängig von mathematischer Formulierung der Restriktion
- F-Statistik als Spezialfall des Wald-Tests
- Lehne H_0 ab, wenn $h(\hat{\beta})$ ungleich Null ist.

Beispiel 2 cont'd

Im Bernoulli-Beispiel war der ML-Schätzer über die 1.Ableitung berechnet als:

$$\frac{\partial \log L}{\partial p} = T_1 \frac{1}{\hat{p}} - (T - T_1) \frac{1}{1 - \hat{p}} \tag{6}$$

Die 2.Ableitung führt auf die Varianz-Kovarianz-Matrix (2. Moment der Verteilung):

$$\frac{\partial^2 \log L}{\partial^2 p} = T_1 \frac{1}{\hat{p}^2} - (T - T_1) \frac{1}{1 - \hat{p}^2} \tag{6a}$$

Dies kann man auflösen nach der Varianz-Kovarianz-Matrix V:

$$J = \frac{1}{p(1-p)} = V^{-1} \tag{6b}$$

Beispiel 2 cont'd: Teststatistiken

Der Wald-Test

$$\xi_W = T(\hat{p} - \tilde{p})[\hat{p}(1 - \hat{p})]^{-1}(\hat{p} - \tilde{p}) \tag{22}$$

Der Likelihood Ratio Test

$$\xi_{LR} = 2(logL(\hat{p}) - logL(\tilde{p})) \tag{23}$$

Der Lagrange Multiplier Test

$$\begin{aligned} \xi_{LM} &= T^{-1}\left(\frac{T_1}{\hat{p}} - \frac{T - T_1}{1 - \hat{p}}\right)[\tilde{p}(1 - \tilde{p})]\left(\frac{T_1}{\hat{p}} - \frac{T - T_1}{1 - \hat{p}}\right) \\ &= T^{-1}(T_1 - T\tilde{p})[\tilde{p}(1 - \tilde{p})]^{-1}(T_1 - T\tilde{p}) \end{aligned} \tag{24}$$

und, da $\hat{p} = T_1/T$:

$$\xi_{LM} = T(\hat{p} - \tilde{p})[\tilde{p}(1 - \tilde{p})]^{-1}(\hat{p} - \tilde{p}) \tag{24}$$

Beispiel 3: Sensitivität des Wald-Tests

Schätze

$$log(C_t) = \beta_1 + \beta_2 log(Yver_t) + \beta_3 log(Yver_{t-4}) + \beta_4 log(C_{t-4}) + u_t \quad (25)$$

mit C = reale Konsumausgaben und $Yver$ = reales verfügbares Einkommen.
Teste, ob die langfristige Einkommenselastizität des Konsums gleich 1 beträgt:[2]
Langfristig gilt: $C_t = C_{t-4} = C$ und $Yver_t = Yver_{t-4} = Yver$
Auflösen von Gleichung (25) führt zu

$$H_0 : \frac{\beta_2 + \beta_3}{1 - \beta_4} = 1$$

oder

$$H_0 : \beta_2 + \beta_3 + \beta_4 - 1 = 0$$

[2]Kointegration zwischen Konsum und Einkommen

Sensitivität des Wald-Tests

Eine Schätzung mit Daten für Großbritannien von 1955:1 bis 1998:4 ergibt:

Tabelle: Wald-Test

Hypothese	F-Statistik	p-Wert
$\frac{\beta_2+\beta_3}{1-\beta_4} = 1$	34.23852	0.0000
$\beta_2 + \beta_3 + \beta_4 - 1 = 0$	5.592360	0.0192

Gute Nachricht: GRETL erlaubt nur das Testen von linearen Restriktionen und formuliert nichtlineare automatisch um.

Erweiterungen

- Quasi-ML als allgemeinere Methode, Aufgabe der Annahme normalverteilter Störterme
- Problem: Schätzer ist nur asymptotisch effizient.
- Lösung: großer Beobachtungsumfang oder Bayesianische Priors

Zusammenfassung

- ML-Schätzer als allgemeines Schätzprinzip
- Schätzer für σ^2 nicht erwartungstreu bei kleinem Beobachtungsumfang, dafür Möglichkeit der Schätzung nichtlinearer Modelle
- Berechnungsgrundlage für eine Reihe von Testprinzipien
- **Literatur:**
  Verbeek (2008): Kapitel 6.1 - 6.3
  Dougherty (2002): Kapitel 11.6