Willkommen zur Vorlesung Statistik (Master)

Thema dieser Vorlesung: Inferenzstatistik in Regressionsmodellen

Prof. Dr. Wolfgang Ludwig-Mayerhofer

Universität Siegen – Philosophische Fakultät, Seminar für Sozialwissenschaften

Tests für Regressionsmodelle

- Einführung
- Das lineare Regressionsmodell
- Regressionsmodelle auf der Basis von Maximum-Likelihood-Schätzung

Tests in Regressionsmodellen

Typische statistische Tests in Regressionsmodellen sind folgende:

1. „Gesamttest“ oder „Overall-Test“ (Fahrmeir et al.): Test der H_0, dass alle Regressionskoeffizienten gleich Null sind, sprich, dass das Modell in seiner Gesamtheit nichts zur Erklärung der abhängigen Variablen beiträgt.
2. Tests der einzelnen Koeffizienten β_k, üblicherweise gegen die H_0: $\beta_k = 0$
3. (In der Praxis selten:) Tests, die sich auf zwei oder mehr der Regressionskoeffizienten beziehen (sog. Restriktionen); etwa, dass zwei oder mehr Koeffizienten den gleichen Betrag aufweisen, oder dass sie sich um einen bestimmten Betrag unterscheiden.

Das lineare Regressionsmodell

$$y_i = \widehat{\alpha} + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \cdots + \widehat{\beta}_k x_{ki} + e_i \quad (1)$$

bzw.

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \cdots + \widehat{\beta}_k x_{ki} \quad (2)$$

- $\widehat{\alpha}$: Geschätzte Regressionskonstante (oft auch als $\widehat{\beta}_0$ geschrieben)
- $\widehat{\beta}_K$: Geschätzter Regressionskoeffizient der k-ten unabhängigen Variablen X_k
- $e_i = y_i - \widehat{y}_i$: Residuen
- i: Index für die Beobachtungen
- k: Index für die unabhängigen Variablen

Test für das Gesamtmodell

In einem linearen (OLS-)Regressionsmodell mit k unabhängigen Variablen (Prädiktoren, Regressoren) prüft der F-Test

$$F = \frac{MQS_{\text{Schätzung}}}{MQS_{\text{Residuen}}} = \frac{QS_{\text{Schätzung}}}{QS_{\text{Residuen}}} \frac{n-k-1}{k} = \frac{R^2}{1-R^2} \frac{n-k-1}{k} \tag{3}$$

die H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$.

Zugrunde liegt: $QS_{\text{Gesamt}} = QS_{\text{Schätzung}} + QS_{\text{Residuen}}$ oder

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

sowie

$$MQS_{\text{Schätzung}} = \frac{QS_{\text{Schätzung}}}{k} \quad \text{sowie} \quad MQS_{\text{Residuen}} = \frac{QS_{\text{Residuen}}}{n-k-1}$$

Wieder wird die Abhängigkeit von n deutlich.

Der F-Test im Regressionsmodell I

Der F-Test für das Gesamtmodell ist nur eine spezielle Variante des allgemeinen F-Tests, mit dem verschiedenste Modellrestriktionen getestet werden können, namentlich

- dass zwei (oder mehr) Regressionsoeffizienten den gleichen Betrag aufweisen,
- dass eine Teilmenge der Regressionskoeffizienten gleich Null ist,

Dazu wird die QS_{Residuen} des aktuellen (vollen) Modells mit der QS_{Residuen/H_0} in Beziehung gesetzt, d. h. mit der QS_{Residuen}, die sich bei Schätzung des Modells unter den Restriktionen ergibt. Von Interesse ist die relative Differenz zwischen den Residuenquadratsummen:

$$\frac{\Delta QS_{\text{Residuen}}}{QS_{\text{Residuen}}} = \frac{QS_{\text{Residuen}/H_0} - QS_{\text{Residuen}}}{QS_{\text{Residuen}}}$$

Der F-Test im Regressionsmodell II

Die Statistik

$$F = \frac{\frac{1}{r}\Delta QS_{\text{Residuen}}}{\frac{1}{n-k-1}QS_{\text{Residuen}}} = \frac{n-k-1}{r}\frac{\Delta QS_{\text{Residuen}}}{QS_{\text{Residuen}}}$$

(mit r = Zahl der Restriktionen im Modell) folgt einer F-Verteilung mit r und $n-k-1$ Freiheitsgraden. Gute Statistik-Software erlaubt die problemlose Berechnung dieser Statistik für beliebige Restriktionen.

Für den speziellen Fall, dass *ein* Regressionskoeffizient β_k gegen die H_0 $\beta_k = 0$ getestet werden soll, entspricht die Größe

$$F = \frac{\widehat{\beta}_k^2}{\text{Var}(\widehat{\beta}_k)}$$

dem bekannten und gängigen t-Test

$$t = \frac{\widehat{\beta}_k}{\text{SE}_{\beta_k}}.$$

Test für einzelne Regressionskoeffizienten

Die Varianz der geschätzten Koeffizienten kann wie folgt angegeben werden (Fahrmeir et al., Regression, S. 101):

$$\text{Var}(\widehat{\beta}_k) = \frac{\widehat{\sigma}^2_{\text{Res}}}{(1 - R^2_k) \sum_{i=1}^{n} (x_{ki} - \bar{x}_k)^2} \tag{4}$$

mit $\widehat{\sigma}^2_{\text{Res}}$ als der Residualvarianz ($\text{MQS}_{\text{Residuen}}$) und R^2_k als dem R^2, das sich bei Regression der k-ten uV auf die übrigen uV ergibt (entspricht 1/VIF, auch *tolerance* genannt).

Die Größe der Standardfehler (Wurzel aus der der Varianz!) hängt also von der Varianz der Residuen, der Abhängigkeit der uV untereinander und der Streuung der jeweiligen uV ab.

Robuste Standardfehler

Sind die inferenzstatistischen Voraussetzungen für das Standard-Regressionmodell (Normalverteilung der Residuen, Homoskedastizität, Unabhängigkeit) nicht gegeben, können für die Tests der Regressionskoeffizienten sog. robuste Standardfehler geschätzt werden. Diese sind meist (aber nicht immer) größer als die „Standard“-Standardfehler.

White, H. (1980). „A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity“ Econometrica, 48, 817-838.

Maximum-Likelihood-Schätzung: Vorbemerkung

Als Einstieg in die vertiefende Literatur für die nachfolgenden Ausführungen empfiehlt sich:

- Thomas Gautschi: Maximum-Likelihood Schätztheorie
- Henning Best/Christof Wolf: Logistische Regression

Beide in: Christof Wolf/Henning Best (Hrsg.): Handbuch der sozialwissenschaftlichen Datenanalyse, Wiesbaden: VS-Verlag 2010.

Maximum-Likelihood-Schätzung

Die meisten multivariaten statistischen Modelle jenseits des linearen Regressionsmodells basieren auf Maximum-Likelihood-Schätzungen.

Die Grundidee: Wir „probieren“ verschiedene mögliche Werte der Grundgesamtheit aus, bis der- oder diejenigen Wert(e) gefunden wurde(n), die das Stichprobenergebnis am besten erklären können.

Beispiel: Eine Stichprobe von n=4 ergibt bei einem binären Merkmal (0/1), dass einer der vier Fälle die Ausprägung 1 ausweist. Zu welcher Binomialverteilung „passt“ dieser Wert am besten?

π	P(X=1)	π	P(X=1)
0.1	0.2916	0.26	0.4214
0.2	0.4096	0.28	0.4180
0.22	0.4176	0.3	0.4116
0.24	0.4214	0.4	0.3456
0.25	0.4219		

Die Wahrscheinlichkeit, in genau 1 von 4 Fällen den Wert 1 zu erhalten, ist am größten bei $\pi = 0{,}25$.

Maximum-Likelihood-Schätzung theoretisch

Wir haben anfangs gelernt, die Wahrscheinlichkeit für unterschiedliche Stichprobenergebnisse (Daten) bei gegebenen Parametern zu bestimmen. Formal:

$$P(y|\theta) = f(y|\theta) \tag{5}$$

wobei $f(y|\theta)$ eine Wahrscheinlichkeitsdichte bzw. Wahrscheinlichkeitsfunktion für einen Parameter bzw. Parametervektor θ ist (θ wird oft als Platzhalter für unterschiedliche Parameter verwendet).

Beispiel: Die Binomialverteilung

$$P(y|\pi) = \binom{n}{n_1} \cdot \pi^{n_1} \cdot (1-\pi)^{n-n_1} \tag{6}$$

ML-Schätzung theoretisch II

Die ML-Schätzung fragt genau umgekehrt: Welche „Wahrscheinlichkeiten“ ergeben sich für unterschiedliche Parameter bei gegebenen Daten:

$$\mathcal{L}(\theta|\mathbf{y}) = f(y_1, y_2, \ldots, y_n|\theta) \tag{7}$$

Bei Gültigkeit der i. i. d.-Annahme können wir diese Größe, die sog. Likelihood-Funktion, auch als Produkt der Wahrscheinlichkeitsfunktionen (oder -dichten) für die einzelnen Datenwerte schreiben:

$$\begin{aligned} \mathcal{L}(\theta|\mathbf{y}) &= f(y_1, y_2, \ldots, y_n|\theta) \\ &= f(y_1|\theta) \cdot f(y_2|\theta) \cdot \ldots \cdot f(y_n|\theta) \\ &= \prod_{i=1}^{n} f(\theta|y_i) \end{aligned} \tag{8}$$

Gesucht wird der Parameter(-vektor) θ, der $\mathcal{L}(\theta|\mathbf{y})$ maximiert.

ML-Schätzung theoretisch III

Am Beispiel der Binomialverteilung:

$$\mathcal{L}(\pi|\mathbf{y}) = \binom{n}{k} \cdot \pi^k \cdot (1-\pi)^{n-k} \tag{9}$$

Da allerdings der Ausdruck $\binom{n}{k}$ nicht von π abhängt, kann er auch weggelassen werden, so dass sich der Ausdruck

$$\mathcal{L}(\pi|\mathbf{y}) = \pi^k \cdot (1-\pi)^{n-k} \tag{10}$$

ergibt. Diesen bezeichnet man auch als Kern (engl.: Kernel) der Likelihood-Funktion.

ML-Schätzung theoretisch IV

Das Beispiel Binomialverteilung visualisiert:

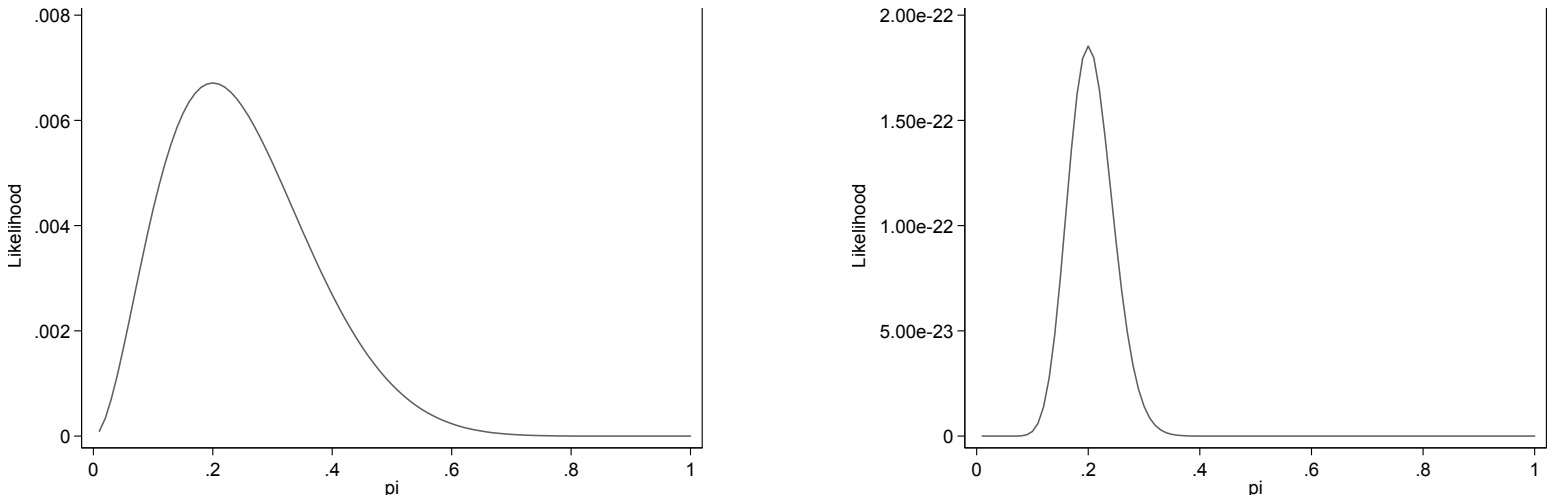


Abbildung: links: 2 mal 1 – 8 mal 0, rechts: 20 mal 1 – 80 mal 0

Es zeigt sich: Bei größeren Fallzahlen liegt die Masse wesentlich enger um das jeweilige Maximum, das bei $\pi = 0{,}2$ liegt.

ML-Schätzung theoretisch V

Wie lässt sich das Maximum der Likelihood-Funktion bestimmen?

Grundsätzlich ist das Vorgehen aus der Schulmathematik bekannt: Die erste Ableitung der Funktion muss gleich 0 (relatives Extremum) und die zweite Ableitung negativ sein (letzteres die Bedingung für ein Maximum [und nicht Minimum]). Was heißt das konkret?

- In einfachen Fällen können erste und zweite Ableitung analytisch bestimmt werden.
- In der Forschungspraxis ist das häufig weder möglich noch nötig: Für die Maximierung gibt es geeignete Algorithmen, die sich iterativ an das Maximum „herantasten“.

Eine wichtige Erleichterung ist dabei die Verwendung der Log-Likelihood, genauer: der logarithmierten Likelihood (die Likelihood führt wegen der [u. U. sehr vielen] Produkte „meist zu unfreundlichen Ausdrücken“ [Fahrmeir et al., 1. Auflage, S. 373]; siehe bereits oben die Graphik für zur Binomialverteilung mit $n = 100$).

ML-Schätzung: Binomial-Verteilung

Bleiben wir weiter beim Beispiel der Binomialverteilung.

Die Likelihood-Funktion $\mathcal{L}(\pi|\mathbf{y}) = \pi^k \cdot (1-\pi)^{n-k}$ können wir auch wie folgt schreiben:

$$\mathcal{L} = \prod_{i=1}^{n} \pi^{y_i} (1-\pi)^{1-y_i} \tag{11}$$

Damit ergibt sich die Log-Likelihood als

$$LL = \sum_{1=1}^{n} [y_i \cdot \ln(\pi)] + [(1-y_i) \cdot \ln(1-\pi)] \tag{12}$$

ML-Schätzung: Binomialverteilung/logistische Regression

In der logistischen Regression wird die Auftretenswahrscheinlichkeit π als $\left(\frac{\exp^{\mathbf{x}_i'\beta}}{1+\exp^{\mathbf{x}_i'\beta}}\right)$ modelliert ($\mathbf{x}_i'\beta$ steht für $\beta_0 + X_1\beta_1 + \cdots + X_k\beta_k$).

Die Likelihood-Funktion ist dann

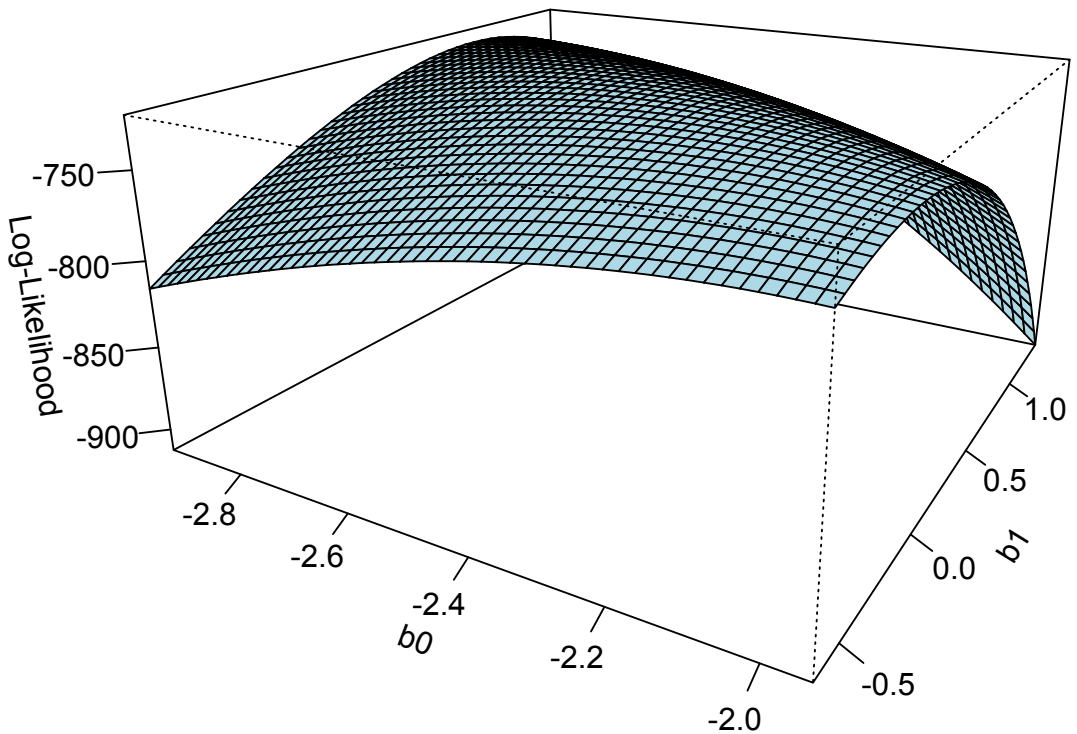
$$\mathcal{L} = \prod_{i=1}^{n} \left(\frac{\exp^{\mathbf{x}_i'\beta}}{1+\exp^{\mathbf{x}_i'\beta}}\right)^{y_i} \left(1 - \frac{\exp^{\mathbf{x}_i'\beta}}{1+\exp^{\mathbf{x}_i'\beta}}\right)^{1-y_i} \tag{13}$$

und die Log-Likelihood

$$LL = \sum_{1=1}^{n} \left[y_i \cdot \ln\left(\frac{\exp^{\mathbf{x}_i'\beta}}{1+\exp^{\mathbf{x}_i'\beta}}\right)\right] + \left[(1-y_i) \cdot \ln\left(1 - \frac{\exp^{\mathbf{x}_i'\beta}}{1+\exp^{\mathbf{x}_i'\beta}}\right)\right] \tag{14}$$

ML-Schätzung: logistische Regression

Beispiel: Eine logistische Regression mit den Parametern $b_0 = -2{,}4373$ und $b_1 = 0{,}3488$ (entspricht dem Beispiel der Aufgabe 2 zur logistischen Regression aus der Übung Statistik):



ML-Schätzung: Begriffe

Einige Begriffe tauchen in der Literatur, möglicherweise aber auch (vor allem bei Problemen des Algorithmus) im Output von Statistik-Software auf:

- Die erste Ableitung gibt die Steigung der (Log-)Likelihood-Funktion an der jeweiligen Stelle an. Sie heißt daher auch Gradient / Gradientenvektor (gradient [vector]). Eine andere Bezeichnung ist Score-Vektor.
- Die Matrix der zweiten Ableitung(en) heißt Hesse-Matrix; diese muss „negativ definit“ sein (Fehlermeldung manchmal: „Hessian not negative definite“). Der negative Wert der Hesse-Matrix (also $-\mathbf{H}$) heißt Fisher-Information(smatrix). Die Inverse dieser Matrix enthält die Varianzen (und Kovarianzen) der Schätzung der Parameter des Modells.

  Sind also die zweiten Ableitungen sehr groß (LL ist an der betreffenden Stelle steil, siehe Graphiken weiter oben), sind die Varianzen und damit die Standardfehler entsprechend klein.

ML-Schätzung: Inferenzstatistik I

- Die Schätzer für die Regressionskoeffizienten folgen einer Normalverteilung. Die ML-Schätzung liefert effiziente, konsistente und *asymptotisch* erwartungstreue (unverzerrte) Schätzer. (Ein häufig zitiertes Beispiel ist $\widehat{\sigma}^2$ der Normalverteilung: Der ML-Schätzer entspricht s^2, also der Varianz der beobachteten Daten, die nur asymptotisch gegen den korrekten Schätzwert konvergiert.)
- Manchmal wird statt dessen ein Wald-Test durchgeführt (nach dem schwedischen Statistiker des gleichen Namens). Dieser basiert auf dem Quadrat der Varianz für den Schätzer, folgt also einer Chi-Quadrat-Verteilung.

Andere Tests wie der Score-Test (oder Lagrange-Multiplier-Test) finden sich vor allem in der theoretischen Literatur.

ML-Schätzung/Inferenzstatistik II: Likelihood-Ratio-Test

Anhand der Likelihood bzw. der Log-Likelihood lassen sich statistische Tests durchführen, die jeweils ein aktuelles Modell mit einem restringierten Modell vergleichen. Die Größe

$$LR = 2 \cdot \ln(L_1/L_0) = 2 \cdot (LL_1 - LL_0) = 2LL_1 - 2LL_0 \tag{15}$$

mit L_1 bzw. LL_1 = (Log-)Likelihood des aktuellen Modells und L_0 bzw. LL_0 = (Log-)Likelihood des reduzierten Modells folgt einer χ^2-Verteilung, deren Freiheitsgrade der Zahl der restringierten Parameter entsprechen.

Dieser Test heißt Likelihood-Quotienten- oder Likelihood-Verhältnis- oder Likelihood-Ratio-Test (LR-Test). Er bildet das Analogon zum F-Test im linearen Regressionsmodell.

Oft wird der LR-Test in folgender (äquivalenter) Formulierung vorgestellt:

$$LR = -2 \cdot \ln(L_0/L_1) = -2 \cdot (LL_0 - LL_1) = -2LL_0 + 2LL_1 \tag{16}$$