10. Grundlagen der linearen Regressionsanalyse

10.1 Formulierung linearer Regressionsmodelle

Einfaches lineares Regressionsmodell:

Das einfache lineare Regressionsmodell ist die simpelste Form eines ökonometrischen Modells und untersucht lediglich den Zusammenhang zwischen zwei Variablen, d.h. den Effekt einer Variablen auf eine andere Variable.

Regressionsgleichung:

$$y = \beta_0 + \beta_1 x + u$$

Unterstellte Beziehung in der Regression:

$$\boxed{X} \longrightarrow \boxed{Y}$$

Zur Terminologie:

y	x	u	β
Abhängige Variable	Erklärende Variable	Störterm	Regressionsparameter
Zu erklärende Variable	Unabhängige Variable	Fehlerterm	β_0: Konstante (Achsenabschnitt)
Endogene Variable	Exogene Variable		β_1: Steigungsparameter
Regressand	Regressor		

Der Steigungsparameter β_1 beschreibt den linearen Zusammenhang zwischen x und y, falls alle Faktoren in u konstant gehalten werden:

$$\Delta y = \beta_1 \Delta x \quad \text{falls } \Delta u = 0$$

Problem:

Es gibt viele weitere Faktoren, die einen Effekt auf y haben können, so dass es zu Problemen kommen kann, wenn nur das restriktive Modell betrachtet wird.

Deshalb:

Der Störterm u und die erklärende Variable x dürfen nicht miteinander in Beziehung stehen, wenngleich die Unkorreliertheit bei der Betrachtung des Korrelationskoeffizienten nicht ausreichend ist.

Generell ist aber schon die Annahme der Unkorreliertheit zwischen der erklärenden Variablen x und den unbeobachtbaren Einflussfaktoren im Störterm u unrealistisch, so dass die Gleichung in der Regel verletzt ist. Aus diesem Grund sollten weitere (d.h. sämtliche relevante) erklärenden Variablen in das lineare Regressionsmodell aufgenommen werden.

Einfachste Form eines multiplen linearen Regressionsmodells (mit zwei erklärenden Variablen):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Allgemeine Form eines multiplen linearen Regressionsmodells (mit k erklärenden Variablen):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k + u$$

Dabei gilt:

x_1, x_2, x_3, …, x_{k-1}, x_k: Erklärende Variablen
β_0: Konstante
β_1: Dieser Parameter misst den Effekt einer Veränderung von x_1 auf y, falls alle anderen beobachteten und unbeobachteten Faktoren konstant sind
β_2: Dieser Parameter misst den Effekt einer Veränderung von x_2 auf y, falls alle anderen beobachteten und unbeobachteten Faktoren konstant sind
:
β_k: Dieser Parameter misst den Effekt einer Veränderung von x_k auf y, falls alle anderen beobachteten und unbeobachteten Faktoren konstant sind
u: Störterm

10.2 Schätzung der unbekannten Parameter mit der OLS-Methode

Für die weitere Analyse linearer Regressionsmodelle benötigt man eine Stichprobe vom Umfang n aus der Grundgesamtheit.

Einfaches lineares Regressionsmodell:

$\{(x_i, y_i), i = 1,\ldots, n\}$

Multiples lineares Regressionsmodell mit k erklärenden Variablen:

$\{(x_{i1}, x_{i2},\ldots, x_{ik}, y_i), i = 1,\ldots, n\}$

Unter Einbeziehung der Beobachtungen i = 1,…, n ergeben sich folgende lineare Regressionsmodelle:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i$$

Dabei ist z.B. x_{ik} der Wert der erklärenden Variablen k bei Beobachtung i.

Wesentliche Aufgabe der Regressionsanalyse:

Schätzung der unbekannten Regressionsparameter $\beta_0, \beta_1, \beta_2,\ldots$

Optimierungsproblem bei der Methode der kleinsten Quadrate („OLS, ordinary least squares“) im einfachen linearen Regressionsmodell:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Daraus folgen die Bedingungen erster Ordnung für die beiden geschätzten Regressionsparameter:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Daraus folgen die OLS-Schätzer für die beiden unbekannten Parameter:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = \frac{1}{n} \sum_{i=1}^{n} y_i - \hat{\beta}_1 \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Optimierungsproblem bei der Methode der kleinsten Quadrate im multiplen linearen Regressionsmodell:

$$\min_{\beta_0, \beta_1, \ldots} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik})^2$$

Daraus folgen die Bedingungen erster Ordnung für die k+1 geschätzten Regressionsparameter:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik}) = 0$$

$$\sum_{i=1}^{n} x_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik}) = 0$$

$$\vdots$$

$$\sum_{i=1}^{n} x_{ik} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik}) = 0$$

OLS-Schätzfunktion und OLS-Regressionswerte

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} \quad \text{für } i = 1, \ldots, n$$

Interpretation der geschätzten Parameter in multiplen linearen Regressionsmodellen:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \cdots + \hat{\beta}_k \Delta x_k$$

Falls die erklärenden Variablen x_2, x_3, x_4,…, x_k konstant gehalten werden, folgt:

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1$$

Der geschätzte Parameter für die erklärende Variable x_1 gibt somit in diesem Fall die Veränderung des Regressionswertes an, falls x_1 um eine Einheit steigt.

Falls x_1, x_2, x_3,…, x_{k-1} konstant gehalten werden, ergibt sich:

$$\Delta\hat{y} = \hat{\beta}_k \Delta x_k$$

Der geschätzte Parameter für die erklärende Variable x_k gibt somit in diesem Fall die Veränderung des Regressionswertes an, falls x_k um eine Einheit steigt.

Die geschätzten Parameter können somit als geschätzte partielle Effekte interpretiert werden, d.h. bei der Schätzung des Effektes einer Variablen wird für die anderen erklärenden Variablen kontrolliert. Dies ist der große Vorteil der Regressionsanalyse (bzw. allgemein ökonometrischer Analysen), d.h. es kann eine ceteris paribus Betrachtung vorgenommen werden, ohne dass ein entsprechendes kontrolliertes Experiment durchgeführt werden muss.

Beispiel: Erklärung von College-Noten

Mit Hilfe eines linearen Regressionsmodells soll der Effekt der durchschnittlichen Punktzahl in der Highschool (hsGPA) und der Punktzahl im College-Aufnahmetest (ACT) auf die durchschnittliche College-Punktzahl (colGPA) untersucht werden:

$$\text{colGPA} = \beta_0 + \beta_1 \text{hsGPA} + \beta_2 \text{ACT} + u$$

Dabei wurde mit Hilfe einer Stichprobe von n = 141 Studierenden folgende OLS-Regressionsfunktion geschätzt:

$$\widehat{\text{colGPA}} = 1{,}29 + 0{,}453\text{hsGPA} + 0{,}009\text{ACT}$$

Interpretation:

- Die geschätzte Konstante kann nicht sinnvoll interpretiert werden, da hier ACT und hsGPA realitätsfern mit null gleichgesetzt werden
- Geschätzter positiver Einfluss der Highschool-Note auf die College-Note: Falls ACT konstant gehalten wird, führt eine um einen Punkt höhere Highschool-Note zu einer geschätzten Erhöhung der College-Punktzahl um 0,453 Punkte
- Dagegen hat die Punktzahl im College-Aufnahmetest einen deutlich geringeren geschätzten Effekt, ceteris paribus

Bestimmtheitsmaß (Determinationskoeffizient): Anteil der Variation der abhängigen Variablen y_i, der durch die OLS-Regressionsfunktion erklärt wird

$$R^2 = \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})\right)^2}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\right)}$$

Eigenschaften des Bestimmtheitsmaßes:

- $0 \leq R^2 \leq 1$
- R^2 sinkt niemals, wenn eine weitere (möglicherweise irrelevante) erklärende Variable hinzugefügt wird
- Aus diesem Grund ist R^2 ein schlechtes Maß zur Beurteilung der Güte eines linearen Regressionsmodells

10.3 Erwartungswert und Varianz von OLS-Schätzern

→ Zu beachten ist, dass sich statistische Eigenschaften (d.h. Erwartungswert und Varianz) von OLS-Schätzern nicht auf eine bestimmte Stichprobe, sondern auf wiederholte Zufallsstichproben aus der Grundgesamtheit beziehen

Unter bestimmten Annahmen (z.B. Linearität in den Parametern, keine perfekte Kollinearität) sind alle mit der OLS-Methode geschätzten Parameter erwartungstreu, d.h. der Erwartungswert der Schätzer ist gleich den unbekannten Regressionsparametern:

$$E(\hat{\beta}_j) = \beta_j \quad \text{für } j = 0, 1, \ldots, k$$

Mit weiteren Annahmen (insbesondere mit der Annahme der Homoskedastizität, d.h. einer konstanten bedingten Varianz des Fehlerterms u) ergibt sich für die Varianz der mit OLS geschätzten Steigungsparameter in linearen Regressionsmodellen:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{(1-R_j^2)\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2} \quad \text{für } j = 1, \ldots, k$$

Dabei stellt R_j^2 das Bestimmtheitsmaß bei einer Regression von x_j auf alle anderen erklärenden Variablen (einschließlich einer Konstante) dar.

Unter verschiedenen Annahmen sind die OLS-Schätzer die besten linearen unverzerrten Schätzer der Regressionsparameter in linearen Regressionsmodellen („BLUE, best linear unbiased estimator“).

Bestandteile von BLUE:

- „Unverzerrt“ bedeutet, dass der Schätzer erwartungstreu ist.
- „Linear“ bedeutet, dass der Schätzer eine lineare Funktion der Daten und der abhängigen Variablen darstellt.
- „Beste“ bedeutet, dass der Schätzer die geringste Varianz besitzt.

OLS-Schätzer haben also in der Klasse aller linearen und unverzerrten Schätzer die geringste Varianz und sind somit effizient.

Erwartungstreue Schätzung der Varianz σ^2 des Fehlerterms u als Grundlage für die Schätzung der Varianz der mit OLS geschätzten Regressionsparameter:

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^{\,2}$$

Der entsprechende (zwar konsistente, aber nicht erwartungstreue) Schätzer für die Standardabweichung σ des Fehlerterms u lautet dann:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^{\,2}}$$

10.4 Tests in linearen Regressionsmodellen

Für den folgenden Test wird angenommen, dass ein normalverteilter Störterm vorliegt (wenngleich eine von der Normalverteilung abweichende Verteilung des Störterms im Hinblick auf die Durchführung von statistischen Tests bei großen Stichprobenumfängen kein Problem darstellt). In diesem Fall sind auch die mit OLS geschätzten Steigungsparameter in linearen Regressionsmodellen normalverteilt, d.h. es gilt (j = 1,…, k):

$$\hat{\beta}_j \sim N[\beta_j;\ Var(\hat{\beta}_j)] \ \text{ bzw. } \ \hat{\beta}_j \sim N\left[\beta_j;\ \frac{\sigma^2}{(1-R_j^2)\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}\right]$$

Damit ergibt sich:

$$\frac{\hat{\beta}_j-\beta_j}{\sqrt{Var(\hat{\beta}_j)}} = \frac{\hat{\beta}_j-\beta_j}{\dfrac{\sigma}{\sqrt{(1-R_j^2)\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}} \sim N(0;\ 1)$$

Allerdings sind die Varianzen und Standardabweichungen der mit OLS geschätzten Steigungsparameter in linearen Regressionsmodellen in der Regel nicht bekannt und müssen deshalb geschätzt werden. Dabei ergibt sich:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{Var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j}{\dfrac{\hat{\sigma}}{\sqrt{(1-R_j^2)\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}} \sim t_{n-k-1}$$

Dabei ist k+1 die Anzahl der unbekannten Regressionsparameter. Die t-Verteilung mit n-k-1 Freiheitsgraden ergibt sich gegenüber der Standardnormalverteilung durch die Einbeziehung der Schätzung $\hat{\sigma}$ der Standardabweichung σ des Störterms u.

Die wichtigste zu testende Nullhypothese in empirischen Anwendungen lautet:

$$H_0: \beta_j = 0 \quad \text{für } j = 1,\ldots, k$$

Diese Nullhypothese über den Steigungsparameter β_j impliziert, dass die erklärende Variable x_j keinen partiellen Effekt auf die abhängige Variable y hat. Die Alternativhypothese lautet:

$$H_1: \beta_j \neq 0 \quad \text{für } j = 1,\ldots, k$$

Als Prüfgröße wird hierzu folgende t-Statistik (t-Wert) betrachtet, die die geschätzte Standardabweichung des geschätzten Parameters einbezieht:

$$t_{\hat{\beta}_j} = t_j = \frac{\hat{\beta}_j}{\sqrt{V\hat{a}r(\hat{\beta}_j)}}$$

Bei Gültigkeit der Nullhypothese ist diese Prüfgröße t-verteilt mit n-k-1 Freiheitsgraden. Die Nullhypothese wird somit bei einem Signifikanzniveau von α verworfen, falls gilt:

$$\left| t_{\hat{\beta}_j} \right| > t_{n-k-1;1-\alpha/2}$$

Wenn z.B. die Anzahl der Freiheitsgrade n-k-1 = 25 ist, gilt bei einem Signifikanzniveau von 5% für das Quantil aus der t-Verteilung mit 25 Freiheitsgraden $t_{25;0,975}$ = 2,06. Wenn in diesem Fall der Absolutwert der t-Statistik größer als 2,06 ist, ist der entsprechende geschätzte Regressionsparameter zu diesem Signifikanzniveau von null verschieden. Damit hat die entsprechende erklärende Variable einen (statistisch) signifikanten Effekt auf die abhängige Variable.

Mit einer Zunahme der Anzahl an Freiheitsgraden nähert sich die t-Verteilung der Standardnormalverteilung an, so dass man in diesem Fall zur Ermittlung der Schrankenwerte auch die Quantile der Standardnormalverteilung verwenden kann (z.B. für n-k-1 > 120).

Beispiel: Erklärung von College-Noten

Mit Hilfe eines linearen Regressionsmodells soll der Effekt der durchschnittlichen Punktzahl in der Highschool (hsGPA), der Punktzahl im College-Aufnahmetest (ACT) und der durchschnittlichen Anzahl verpasster Vorlesungen pro Woche (skipped) auf die durchschnittliche College-Punktzahl (colGPA) untersucht werden. Dabei wurde folgende OLS-Regressionsfunktion geschätzt:

$$\widehat{colGPA} = 1{,}39 + 0{,}412hsGPA + 0{,}015ACT - 0{,}083skipped$$

$$(0{,}33)\ (0{,}094) \qquad (0{,}011) \qquad (0{,}026)$$

$$n = 141,\ R^2 = 0{,}234$$

Bei einem 5%-Signifikanzniveau ergibt sich für das Quantil aus der t-Verteilung mit n-k-1 = 141-4 = 137 Freiheitsgraden (bzw. das Quantil aus der Standardnormalverteilung) $t_{137;0,975} = 1{,}96$. Bei einem 1%-Signifikanzniveau ergibt sich ein Wert von 2,58. Die t-Werte betragen für hsGPA 4,38, für ACT 1,36 und für skipped -3,19. Damit hat hsGPA einen (statistisch) signifikant positiven Effekt, skipped einen (statistisch) signifikant negativen Effekt und ACT keinen (statistisch) signifikanten Effekt auf colGPA.

Beispiel: Erklärung von Ausschussraten eines Unternehmens

Mit Hilfe eines linearen Regressionsmodells (mit normalverteiltem Störterm) soll der Effekt der jährlichen Trainingsstunden pro Beschäftigten (hrsemp), des Logarithmus der jährlichen Umsätze in Dollar (logsales) und des Logarithmus der Anzahl der Beschäftigten (logemploy) auf den Logarithmus der Anzahl von Ausschussteilen pro 100 produzierte Stücke (logscrap) untersucht werden. Dabei wurde folgende OLS-Regressionsfunktion geschätzt (n = 29 Unternehmen, $R^2 = 0{,}262$):

$$\widehat{\log scrap} = \underset{(5{,}69)}{12{,}46} - \underset{(0{,}023)}{0{,}029}\,hrsemp - \underset{(0{,}453)}{0{,}962}\,logsales + \underset{(0{,}407)}{0{,}761}\,logemploy$$

Interpretation des Effektes von Weiterbildung:

- Eine Erhöhung der jährlichen Trainingsstunden pro Beschäftigten hrsemp um z.B. fünf Stunden führt zu einer geschätzten Senkung von logscrap um $0{,}029 \cdot 5 = 0{,}145$. Dies entspricht einer Senkung von etwa 14,5% und ist damit nicht trivial.
- Allerdings ergibt sich bei hrsemp ein t-Wert von -0,029/0,023 = -1,26, so dass auch bei einem Signifikanzniveau von 10% keine statistische Signifikanz vorliegt

Beispiel: Erklärung von Geburtsgewichten

Mit Hilfe eines linearen Regressionsmodells soll der Effekt der Anzahl der von der Mutter während der Schwangerschaft täglich gerauchten Zigaretten, des jährlichen Familieneinkommens (in 1000 Dollar), der Ausbildungszeit des Vaters (in Jahren), der Ausbildungszeit der Mutter (in Jahren), und des Zigarettenpreises (in Dollar) auf das Geburtsgewicht des Kindes (in Unzen = 28,35 Gramm) untersucht werden. Die OLS-Schätzung hat dabei zu folgenden Ergebnissen geführt (n = 1388 Geburten, R^2 = 0,034):

Abhängige Variable: Geburtsgewicht		
Erklärende Variablen	Parameterschätzwert	t-Statistik
Anzahl gerauchte Zigaretten	-0,593	-5,36
Familieneinkommen	0,051	1,39
Ausbildungszeit Vater	0,475	1,68
Ausbildungszeit Mutter	-0,435	-1,36
Zigarettenpreis	0,068	1,21
Konstante	109,54	13,90

Interpretieren Sie die Schätzergebnisse bei einem Signifikanzniveau von 5%!