Multiple lineare Regression

Worum geht es in diesem Modul?

- Das Konzept der multiplen linearen Regression
- Spezialfälle
- Eigenschaften des Residuums
- Darstellung in Matrixnotation
- Identifikation der Regressionskoeffizienten
- Der multiple Determinationskoeffizient
- Multiple lineare Quasi-Regression
- Statistische Modelle zur multiplen linearen Regression
- Das Allgemeine Lineare Modell

Multiple lineare Regression: Definition

Seien Y und $X_1, ..., X_m$ numerische Zufallsvariablen auf demselben Wahrscheinlichkeitsraum mit endlichen Erwartungswerten, positiven, endlichen Varianzen, sowie invertierbarer Kovarianzmatrix S_{xx}. Dann heißt die Regression $E(Y| X_1, ..., X_m)$ *linear in* $(X_1, ..., X_m)$, falls

$$E(Y| X_1, ..., X_m) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m.$$

Multiple lineare Regression: Spezialfälle I

Bei der obigen Definition setzen wir also noch nicht voraus, dass die Regressoren $X_1, ..., X_m$ unabhängig voneinander definiert sind. In Kapitel 9 haben wir bereits darauf hingewiesen dass z. B. auch die einfache quadratische Regression $E(Y|X)$ linear in (X, X^2) ist, falls also gilt:

$$E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2.$$

Multiple lineare Regression: Spezialfälle II

Ein weiterer Spezialfall, den wir bereits im Kapitel über die bedingte lineare Regression kennen gelernt haben, ist

$$E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2.$$

Auch hier ist die Regression $E(Y|X_1, X_2)$ linear in $(X_1, X_2, X_1 \cdot X_2)$ und es gilt:

$$E(Y|X_1, X_2) = E(Y|X_1, X_2, X_1 \cdot X_2).$$

Multiple lineare Regression: Spezialfälle III

Schließlich sei explizit auch noch einmal auf den Fall eines qualitativen Regressors X mit n Werten $x_1, \ldots, x_n$ hingewiesen. Folgendes Zellenmittelwertmodell

$$E(Y \mid X) = \beta_0 + \beta_1 \cdot I_1 + \beta_2 \cdot I_2 + \ldots + \beta_n \cdot I_n, \quad \text{mit } \beta_0 = 0,$$

ist ebenfalls ein Spezialfall der multiplen linearen Regression. Die Regression $E(Y \mid X)$ ist stets linear in $(I_1, \ldots, I_n)$, auch dann, wenn die Regression $E(Y \mid X)$ nicht linear in (X) ist.

Eigenschaften des Residuums

Das Residuum

$$e := Y - E(Y \mid X_1, \ldots, X_m)$$

besitzt die bekannten Eigenschaften

$$E(e) = 0,$$

$$E(e \mid X_1, \ldots, X_m) = 0,$$

$$Cov[e, f(X_1, \ldots, X_m)] = 0,$$

$$Cov(e, X_i) = 0 \quad \text{für } i = 1, \ldots, m,$$

wobei $f(X_1, \ldots, X_m)$ eine beliebige numerische Funktion der Regressoren bezeichnet.

Darstellung in Matrixnotation I

Fasst man die Regressoren zu dem Zeilenvektor $\boldsymbol{x}' = (X_1 \ldots X_m)$ und die Regressionskoeffizienten $\beta_1, \ldots, \beta_m$ zu einem m-dimensionalen Spaltenvektor $\mathrm{b} = (\beta_1 \ldots \beta_m)'$ zusammen, so lässt sich die multiple lineare Regression nun in Matrix- bzw. Vektornotation auch folgendermaßen schreiben:

$$E(\boldsymbol{y} \mid \boldsymbol{x}) = \beta_0 + \boldsymbol{x}' \mathrm{b} = \beta_0 + (X_1 \ldots X_m) \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}.$$

Auch hier wird der Regressand als Vektor $\boldsymbol{y} = (Y)$ aufgefasst, der eben nur aus einer einzigen Komponente, nämlich Y, besteht. Daher ist auch β_0 eine reelle Zahl.

Darstellung in Matrixnotation II

Definieren wir den Zeilenvektor $\boldsymbol{z}' := (1 \ X_1 \ldots X_m)$ und den Spaltenvektor $\mathrm{g} := (\beta_0 \ \beta_1 \ \ldots \ \beta_m)'$, so können wir die letzte Gleichung auch wie folgt schreiben:

$$E(\boldsymbol{y} \mid \boldsymbol{x}) = E(\boldsymbol{y} \mid \boldsymbol{z}) = \boldsymbol{z}' \mathrm{g} = (1 \ X_1 \ldots X_m) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}.$$

Identifikation der Regressionskoeffizienten I

Zur Identifikation von β_0 und der Komponenten von $b = (\beta_1 \ldots \beta_m)'$ greift man auf die Erwartungswerte des Regressanden und der Regressoren sowie die Kovarianzmatrizen S_{xx} und S_{xy} zurück. Für die Konstante β_0 ergibt sich:

$$\beta_0 = E(y) - E(x)'\, b = E(Y) - [E(X_1) \ \ldots \ E(X_m)] \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}.$$

Identifikation der Regressionskoeffizienten II

Multipliziert man beide Seiten der folgenden Gleichung

$$S_{xy} = Cov(x, y) = Cov(x, \beta_0 + x'\, b + e) = Cov(x, \beta_0 + b'x + e)$$

$$= Cov(x, x)\, b = S_{xx}\, b$$

mit S_{xx}^{-1} vor, so folgt

$$S_{xx}^{-1} S_{xx}\, b = S_{xx}^{-1} S_{xy}.$$

Da $S_{xx}^{-1} S_{xx} = \mathbf{I}$ die Einheitsmatrix ist, folgt daraus

$$b = S_{xx}^{-1} S_{xy}.$$

Damit haben wir eine Formel zur Berechnung der Regressionskoeffizienten aus den Varianzen und Kovarianzen der Regressoren und des Regressanden.

Identifikation der Regressionskoeffizienten bei zwei Regressoren I

Im Fall mit zwei Regressoren X_1 und X_2 erhält man die bereits aus Kapitel 9 bekannte Gleichung

$$\beta_0 = E(Y) - E(\mathbf{x})' \mathrm{b} = E(Y) - [\beta_1 E(X_1) + \beta_2 E(X_2)] = E(Y) - \beta_1 E(X_1) - \beta_2 E(X_2).$$

Die Kovarianz und die Varianz-Kovarianzmatrix lauten

$$\mathrm{S}_{\boldsymbol{xx}} = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & Var(X_2) \end{pmatrix} \text{ und } \mathrm{S}_{\boldsymbol{xy}} = \begin{pmatrix} Cov(X_1, Y) \\ Cov(X_2, Y) \end{pmatrix}.$$

Die Kramersche Regel aus dem letzten Kapitel liefert:

$$\mathrm{S}_{\boldsymbol{xx}}^{-1} = \frac{1}{Var(X_1)\,Var(X_2) - Cov(X_1, X_2)^2} \begin{pmatrix} Var(X_2) & -Cov(X_1, X_2) \\ -Cov(X_1, X_2) & Var(X_1) \end{pmatrix}.$$

Identifikation der Regressionskoeffizienten bei zwei Regressoren II

Multipliziert man $\mathrm{S}_{\boldsymbol{xx}}^{-1}$ mit $\mathrm{S}_{\boldsymbol{xy}}$ folgen:

$$\beta_1 = \frac{Var(X_2)\ Cov(X_1, Y) - Cov(X_2, Y)\ Cov(X_1, X_2)}{Var(X_1)\ Var(X_2) - Cov(X_1, X_2)^2},$$

$$\beta_2 = \frac{Var(X_1)\ Cov(X_2, Y) - Cov(X_1, Y)\ Cov(X_1, X_2)}{Var(X_1)\ Var(X_2) - Cov(X_1, X_2)^2}.$$

Der multiple Determinationskoeffizient

Für die Varianz $Var[E(\boldsymbol{y}|\boldsymbol{x})]$ der Regression gilt:

$$Var[E(\boldsymbol{y}|\boldsymbol{x})] = Var(\beta_0 + \boldsymbol{x}'\mathrm{b}) = Var(\boldsymbol{x}'\mathrm{b}) = Var(\mathrm{b}'\boldsymbol{x}) = = \mathrm{b}'\, Var(\boldsymbol{x})\, \mathrm{b} =$$

$$= \mathrm{b}'\, \mathrm{S}_{\boldsymbol{xx}}\, \mathrm{b},$$

Dabei sind $Var(\boldsymbol{x}) = \mathrm{S}_{\boldsymbol{xx}}$ die $(m \times m)$ Varianz-Kovarianzmatrix der Regressoren X_1, ..., X_m und b der m-dimensionale Spaltenvektor der Regressionskoeffizienten β_1, ..., β_m. Der multiple Determinationskoeffizient ergibt sich dann wie folgt:

$$R^2_{Y|X_1,\ldots,X_m} = [\mathrm{b}'\, \mathrm{S}_{\boldsymbol{xx}}\, \mathrm{b}] \,/\, Var(Y).$$

Der multiple Determinationskoeffizient: Spezialfälle

Im Fall mit zwei Regressoren X_1 und X_2 erhält man die bereits aus Kapitel 9 bekannte Gleichung

$$R^2_{Y|X_1,X_2} = [ß_1^2\, Var(X_1) + ß_2^2\, Var(X_2) + 2\, ß_1\, ß_2\, Cov(X_1, X_2)] \,/\, Var(Y).$$

Für den Spezialfall, dass alle Regressoren *paarweise unkorreliert* sind, vereinfacht sich der Ausdruck für den multiplen Determinationskoeffizienten zu

$$R^2_{Y|X_1,\ldots,X_m} = \left(\sum_{i=1}^{m} ß_i^2\, Var(X_i)\right) \frac{1}{Var(Y)}.$$

Multiple lineare Quasi-Regression: Definition

Definition 14.2. Unter den gleichen Voraussetzungen wie in Definition 14.1 definieren wir die *multiple lineare Quasi-Regression*, die wir mit $Q(Y|X_1, ..., X_m)$ oder $Q(\boldsymbol{y}|\boldsymbol{x})$ bezeichnen, als diejenige Linearkombination $\beta_0 + \beta_1 X_1 + ... + \beta_m X_m = \beta_0 + \mathrm{b}'\boldsymbol{x}$ der Komponenten von $\boldsymbol{x} = (X_1 ... X_m)'$, die folgendes erfüllt:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_m X_m + \boldsymbol{n}$$

mit

$$E(\boldsymbol{n}) = 0,$$

und

$$Cov(\boldsymbol{n}, X_1) = ... = Cov(\boldsymbol{n}, X_m) = 0.$$

Multiple lineare Quasi-Regression: Alternative Definition

Definition 14.3. Unter den gleichen Voraussetzungen wie zuvor, können wir $Q(\boldsymbol{y}|\boldsymbol{x})$ auch als diejenige Linearkombination $\beta_0 + \boldsymbol{x}'\mathrm{b}$ definieren, welche die folgende Funktion von b_0 und $\boldsymbol{b}$, das *Kleinst-Quadrat-Kriterium*, minimiert:

$$LS(b_0, \boldsymbol{b}) = E[[Y - (b_0 + \boldsymbol{x}'\boldsymbol{b})]^2].$$

Diejenige Zahl b_0 und derjenige Vektor $\boldsymbol{b}$, für welche die Funktion $LS(b_0, \boldsymbol{b})$ ein Minimum annimmt, seien mit β_0 und b, respektive, bezeichnet. Die multiple lineare Quasi-Regression ist dann definiert durch:

$$Q(\boldsymbol{y}|\boldsymbol{x}) = \beta_0 + \boldsymbol{x}'\mathrm{b}.$$

Der Determinationskoeffizient der multiplen linearen Quasi-Regression

Für die Bestimmung der Regressionskoeffizienten der multiplen linearen Quasi-Regression gelten übrigens analog die gleichen Formeln wie für die entsprechenden Koeffizienten der echten multiplen linearen Regression. Ebenfalls gleich ist die Berechnungsformel für den Determinationskoeffizienten der linearen Quasi-Regression, d. h. es gilt

$$Q^2_{Y|X_1,\ldots,X_m} := Var[Q(\boldsymbol{y}\,|\,\boldsymbol{x})] \,/\, Var(Y) = [\mathrm{b}^{\mathrm{t}}\, \mathrm{S}_{\boldsymbol{xx}}\, \mathrm{b}] \,/\, Var(Y).$$

Statistische Modelle zur multiplen linearen Regression

Bisher haben wir nur ein Einzelexperiment betrachtet: Ziehen einer Beobachtungseinheit u aus der Population und Registrierung der Werte des Regressanden und der Regressoren. Statistische Modelle beziehen sich jedoch auf N Zufallsexperimente, in denen Informationen über die zu schätzenden Parameter gesammelt werden.

Modelle mit stochastischen Regressoren

Modelle mit *stochastischen Regressoren* bestehen aus der N-maligen Wiederholung unseres bisher betrachteten Einzelexperiments. Dies führt dazu, dass man nicht mehr nur einen einzigen Regressanden Y und m Regressoren betrachten muss, sondern N Vektoren $(Y_i\, X_{i1}\, \ldots\, X_{im})$, $i = 1, \ldots, N$, die jeweils das Ergebnis des i-ten Zufallsexperiments repräsentieren. Über diese Vektoren kann man unterschiedliche Verteilungsannahmen machen, z. B. dass die $(Y_i\, X_{i1}\, \ldots\, X_{im})$ unabhängig sind und jeder dieser Vektoren $(m + 1)$-*variat normalverteilt* ist.

Modelle mit festen Regressoren

Mit einem anderen, weitaus häufiger verwendeten statistischen Modell, schätzt man innerhalb der Wertekombinationen $x_1, ..., x_m$ der Regressoren $X_1, ..., X_m$ die Erwartungswerte $E(Y \mid X_1 = x_1, ..., X_m = x_m)$ von Y, indem man innerhalb dieser Wertekombinationen den Regressanden Y mehrfach beobachtet. Die Werte $x_1, ..., x_m$ der Regressoren sind dabei also nicht mehr zufällig, sondern werden als feste Größen betrachtet, die das Design des Experiments charakterisieren. Man spricht daher auch von Modellen mit festen oder nichtstochastischen Regressoren. Das wichtigste dieser Modelle mit festen Regressoren ist das *Allgemeine Lineare Modell*.

Das Allgemeine Lineare Modell

Das Allgemeine Lineare Modell (ALM) ist durch die folgenden Annahmen definiert:

$$\boldsymbol{y} = \mathbf{X}\,\mathbf{ß} + \boldsymbol{e}$$

$$\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \mathrm{s}^2\,\mathbf{I}),$$

Dabei bezeichnet $\boldsymbol{y} = (Y_1 \ldots Y_i \ldots Y_N)'$ nun den Spaltenvektor der für eine Stichprobe des Gesamtumfangs N zu erhebenden „abhängigen“ Variablen. Die so genannte *Designmatrix* $\mathbf{X}$ besteht aus $N \times (m + 1)$ festen Zahlen. Dabei besteht jede Zeile von $\mathbf{X}$ aus den Vektoren $\mathbf{x}_i' := (1\ x_{i1} \ldots x_{im})$, eben den Wertekombinationen der Regressoren, innerhalb derer die Beobachtung Y_i erhoben wird und der vorangestellten Konstanten 1, die dazu führt, dass die Regressionskonstante $ß_0$ die erste Komponente von $\mathbf{ß} = (ß_0\ ß_1 \ldots ß_m)'$ ist. Der Vektor $\boldsymbol{e} = (e_1 \ldots e_i \ldots e_N)'$ schließlich enthält die Residuen $Y_i - (\mathbf{x}_i'\,\mathbf{ß})$.

Das Allgemeine Lineare Modell II

Die zweite Annahme besagt, dass $\boldsymbol{e}$ mit Erwartungswertvektor $E(\boldsymbol{e}) = \mathbf{0}$ und der $N \times N$–Kovarianzmatrix $\mathbf{S}_{\boldsymbol{ee}} = \mathrm{s}^2\,\mathbf{I}$ multivariat normalverteilt ist. Die Residuen e_i sind also unkorreliert und haben gleiche Varianzen. Letzteres ist die so genannte Homoskedastizitätsannahme.

Folgerungen aus den Annahmen des ALM sind zunächst:

$$E(\boldsymbol{y}) = \mathbf{X}\,\mathbf{ß},$$

und

$$\mathbf{S}_{yy} = \mathrm{s}^2\,\mathbf{I}.$$

Das Allgemeine Lineare Modell: Identifikationen

Ist $\mathbf{X}'\mathbf{X}$ invertierbar, so gilt

$$\hat{b} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\boldsymbol{y}.$$

zur Schätzung des Vektors der Regressionskoeffizienten. Diese Formel erhält man durch die Minimierung der Kleinst-Quadrat-Funktion

$$LS(\boldsymbol{b}) = (\boldsymbol{y} - \mathbf{X}\,\boldsymbol{b})'\,(\boldsymbol{y} - \mathbf{X}\,\boldsymbol{b}).$$

Weiter ist noch

$$S_{\hat{b}\hat{b}} = s^2\,(\mathbf{X}'\mathbf{X})^{-1}$$

von Bedeutung, die Kovarianzmatrix dieser Schätzer. Die Wurzeln aus den diagonalen Komponenten von $S_{\hat{b}\hat{b}}$ sind die Standardschätzfehler der Regressionskoeffizienten.

Schätzer des Determinationskoeffizienten

Schließlich sei noch die Formel zur Schätzung des Determinationskoeffizienten genannt:

$$\hat{R}^2 = \frac{\boldsymbol{y}'\mathbf{X}\hat{b} - \boldsymbol{N}\cdot\bar{Y}^2}{\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{N}\cdot\bar{Y}^2} = \frac{Quadratsummeder Regression}{QuadratsummeGesamt},$$

wobei $\bar{Y} = (1/N)\cdot\sum_{i=1}^{N} Y_i$.

Formulierung von Hypothesen I

Wie wir in diesem und den vorangegangenen Kapiteln gesehen haben, lassen sich mit der multiplen linearen Regression durchaus auch komplexe und nichtlineare Abhängigkeiten beschreiben. Dabei gibt es zwei grundsätzliche Strategien.

Erste Strategie: Vergleich zwischen

$$Q(Y \mid X_1, X_2, \ldots, X_{m-p}) = \gamma_0 + \gamma_1 X_1 + \ldots + \gamma_{m-p} X_{m-p}$$

und

$$E(Y \mid X_1, \ldots, X_m) = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m.$$

Die zu testende Nullypothese kann dann in drei Versionen formuliert werden:

H_0: $\beta_{m-p+1} = \beta_{m-p+2} = \ldots = \beta_m = 0$ *Version 1*

H_0: $Q(Y \mid X_1, \ldots, X_{m-p}) = E(Y \mid X_1, X_2, \ldots, X_m)$. *Version 2*

H_0: $R^2_{Y \mid X_1, \ldots, X_m} - Q^2_{Y \mid X_1, \ldots, X_{m-p}} = 0$ *Version 3*

Formulierung von Hypothesen II

Die zweite Strategie ist noch allgemeiner: Man formuliert eine H_0 in der Form

$$H_0: \mathbf{A}\, b - d = \mathbf{0},$$

der *Allgemeinen Linearen Hypothese* und testet diese mit einem Programm wie z. B. Systat oder SPSS (über Syntax) direkt, indem man die Matrix **A** und den Vektor d gemäß seiner Hypothese spezifiziert. Die Matrix **A** muss $p \leq m$ linear unabhängige Zeilen enthalten.

Signifikanztests im ALM I

Will man die Nullhypothese H_0: $\beta_{m-p+1} = \beta_{m-p+2} = ... = \beta_m = 0$, testen, schätzt man zunächst den Determinationskoeffizienten $\hat{R}_E^2$ für die Regression und dann $\hat{R}_Q^2$ für die multiple lineare Quasi-Regression. Dabei beachte man, dass dies jeweils mit unterschiedlichen Designmatrizen und unterschiedlichen Regressionskoeffizienten geschieht. Mit diesen Schätzungen $\hat{R}_E^2$ bzw. $\hat{R}_Q^2$ der beiden Determinationskoeffizienten geht man dann in die Formel

$$F = \frac{(\hat{R}_E^2 - \hat{R}_Q^2) / p}{(1 - \hat{R}_E^2) / (N - m - 1)},$$

die unter den Annahmen des ALM und der Gültigkeit der Nullhypothese eine F-verteilte Teststatistik liefert, mit den Zählerfreiheitsgraden $df_1 = p$ und den Nennerfreiheitsgraden $df_2 = N - m - 1$. Dabei sind m die Anzahl der Regressoren in der multiplen linearen Regression, p die Anzahl der Parameter, die laut Nullhypothese gleich null sein sollen und N der Stichprobenumfang.

Signifikanztests im ALM II

Bei der *zweiten Strategie* berechnet man für die jeweilige *Allgemeine Lineare Hypothese* (ALH) H_0: $\mathbf{A}\,\mathrm{b} - \mathrm{d} = \mathbf{0}$,

die Prüfgröße $$F = \frac{\hat{Q}_h / p}{\hat{Q}_e / (N - m - 1)},$$

Dabei sind p die Anzahl der (linear unabhängigen) Zeilen der Matrix $\mathbf{A}$ der ALH (und damit die Anzahl der simultan geprüften Einzelhypothesen),

$\hat{Q}_h = (\mathbf{A}\hat{b} - \mathrm{d})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{b} - \mathrm{d})$ *Hypothesenquadratsumme*

und $\hat{Q}_e = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\mathbf{X}\hat{b}$. *Fehlerquadratsumme*

Auch die letztgenannte Prüfgröße F ist unter den Annahmen des ALM und der Gültigkeit der Nullhypothese eine F-verteilte Teststatistik, mit den Zählerfreiheitsgraden $df_1 = p$ und den Nennerfreiheitsgraden $df_2 = N - m - 1$.