Lineare Regression

Gegeben seien Paare von Messdaten $(x_i, y_i),\ i = 1, \ldots, n$, geometrisch eine Punktwolke in der Ebene. Dabei können die x_i und y_i durchaus mehrfach auftreten, also auch zu gegebenem x_i mehrere Messwerte $y_{i1}, \ldots, y_{ip}$ vorliegen. Die Standardaufgabe der linearen Regression ist es, ein *lineares Modell*

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

an die Messdaten anzupassen, also eine *beste Gerade* durch die Punktwolke zu legen.

Beispiel 1: Eine Stichprobe von $n = 44$ Bauingenieurstudenten an der Universität Innsbruck ergab im Jahr 1998 die in Abbildung 1 dargestellten Werte für x = Körpergröße [cm] und y = Gewicht [kg] (vgl. Datensatz 1 [Biometrik 1] im Applet). Die beste Gerade postuliert und beschreibt einen linearen Zusammenhang zwischen Größe und Gewicht.

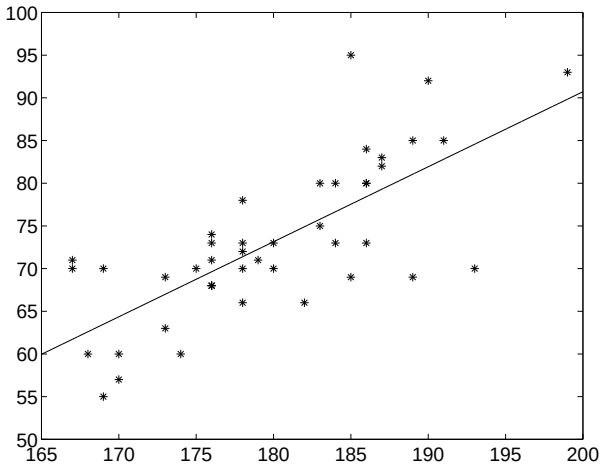

Abbildung 1: Streudiagramm Größe/Gewicht und beste Gerade.

Eine Variante der Standardaufgabe erhält man durch Variablentransformation

$$\xi = \varphi(x),\ \eta = \psi(y)$$

und einem linearen Modellansatz

$$\eta = \beta_0 + \beta_1\, \xi.$$

Rechnerisch ist diese Aufgabe identisch mit der Standardaufgabe der linearen Regression, jedoch durch die transformierten Daten

$$(\xi_i, \eta_i) = \Big(\varphi(x_i), \psi(y_i)\Big).$$

Ein typisches Beispiel ist die *loglineare Regression* $\xi = \log x,\ \eta = \log y$:

$$\log y = \beta_0 + \beta_1 \log x\,,$$

was rücktransformiert dem Ansatz

$$y = \mathrm{e}^{\beta_0} x^{\beta_1}$$

entspricht. Besteht die Variable x ihrerseits aus mehreren Komponenten, die linear kombiniert werden, so spricht man von *multipler linearer Regression* (siehe den zugehörigen Abschnitt unten).

Wir beginnen nun mit der Modellbildung des Regressionsansatzes. Das postulierte Modell für den Zusammenhang zwischen x und y ist das lineare Modell

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

mit unbekannten Koeffizienten β_0 und β_1. Die vorliegenden Daten liegen jedoch nicht exakt auf der entsprechenden Gerade, sondern zeigen Abweichungen ε_i , $i = 1, \ldots, n$:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Wir wollen aus den vorliegenden Daten Schätzwerte $\hat\beta_0, \hat\beta_1$ für β_0, β_1 gewinnen. Dies erfolgt durch Minimierung der Summe der Fehlerquadrate (Abbildung 2)

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

sodass also $\hat\beta_0, \hat\beta_1$ Lösung des Minimierungsproblems

$$L(\hat\beta_0, \hat\beta_1) = \min\Big(L(\beta_0, \beta_1) : \beta_0 \in \mathbb{R},\, \beta_1 \in \mathbb{R}\Big)$$

ist.

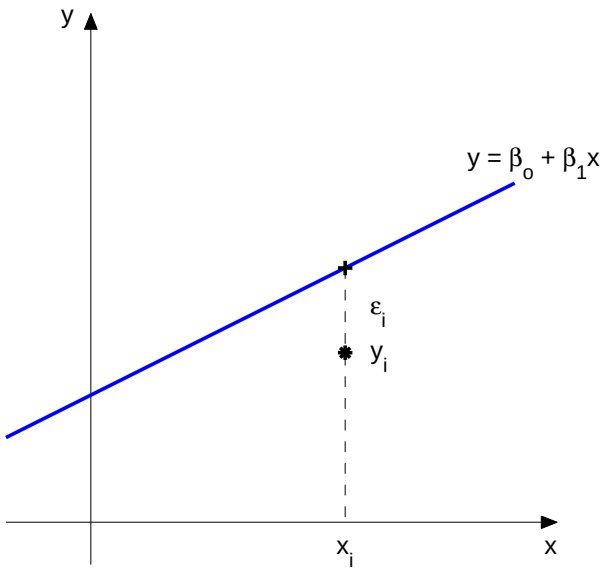

Abbildung 2: Lineares Modell und Fehler.

Wir erhalten $\hat{\beta}_0$ und $\hat{\beta}_1$, indem wir die partiellen Ableitungen von L nach β_0 und nach β_1 Null setzen:

$$\frac{\partial L}{\partial \beta_0}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \quad = 0$$
$$\frac{\partial L}{\partial \beta_1}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Dies führt auf das lineare Gleichungssystem für $\hat{\beta}_0, \hat{\beta}_1$ (die *Normalgleichungen*):

$$n\ \hat{\beta}_0 + \left(\sum x_i\right)\hat{\beta}_1 = \sum y_i$$
$$\left(\sum x_i\right)\hat{\beta}_0 + \left(\sum x_i^2\right)\hat{\beta}_1 = \sum x_i y_i$$

mit der Lösung

$$\hat{\beta}_0 = \left(\frac{1}{n}\sum y_i\right) - \left(\frac{1}{n}\sum x_i\right)\hat{\beta}_1$$
$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

Das *Ergebnis der Regression* ist die *geschätzte Regressionsgerade*

$$y = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Die *durch das Modell prognostizierten Werte* sind dann

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i\,,\ i = 1, \ldots, n.$$

Deren Abweichungen von den Messwerten y_i bezeichnet man als *Residuen*

$$e_i \ = \ y_i - \hat{y}_i \ = \ y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i,\ i \ = \ 1, \ldots, n,$$

siehe Abbildung 3.

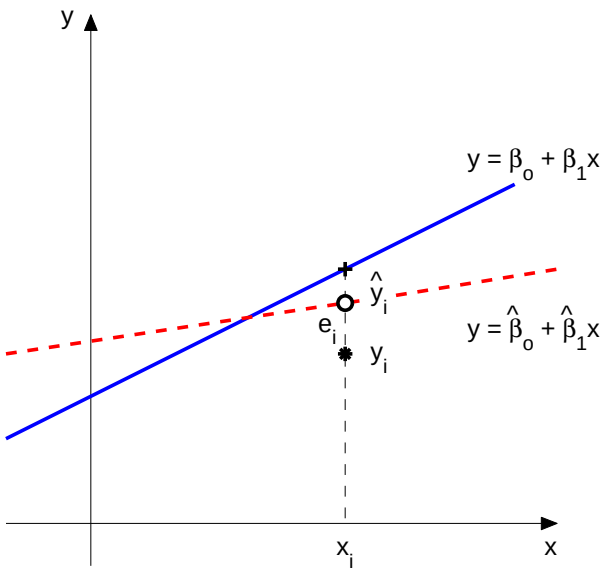

Abbildung 3: Modell, Schätzung, Fehler, Residuum.

Damit ist das *deterministische Regressionsmodell* spezifiziert. Im *probabilistischen Regressionsmodell* werden die Fehler ε_i als Zufallsgrößen mit Erwartungswert Null interpretiert. Unter weiteren Zusatzannahmen (Unabhängigkeit, konstante Varianz, Normalverteilung) wird das Modell dann statistischen Test- und Diagnoseverfahren zugänglich gemacht; dafür verweisen wir auf die einschlägige Literatur, etwa [5]. Den Standpunkt, den wir hier vertreten, ist der der *neodiskriptiven* Statistik. Wir werden keinerlei wahrscheinlichkeitstheoretische Annahmen machen und die Beurteilung der Anpassungsgüte

der Regression sowie die Modellwahl ausschließlich auf deskriptiven Statistiken der Daten aufbauen.

Umformulierung der Normalgleichungen zur besseren numerischen Handhabbarkeit. Wir führen die folgenden Vektoren und Matrizen ein:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

Die Relation Daten-Modell

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ , \ i = 1, \ldots, n,$$

schreibt sich dann in Matrixform einfacher zu

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

Weiters ist

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \begin{bmatrix} 1 & 1 \ \ldots \ 1 \\ x_1 & x_2 \ldots x_n \end{bmatrix} \cdot \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

$$\mathbf{X}^\mathsf{T}\mathbf{y} = \begin{bmatrix} 1 & 1 \ \ldots \ 1 \\ x_1 & x_2 \ldots x_n \end{bmatrix} \cdot \begin{bmatrix} y_i \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix},$$

sodass die Normalgleichungen sich als

$$\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^\mathsf{T}\mathbf{y}$$

darstellen. Die Lösung ist somit

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\,\mathbf{X}^\mathsf{T}\mathbf{y};$$

die Prognosewerte und Residuen sind

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}, \quad \mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}.$$

Beispiel 1, Fortsetzung. Die Daten für $x =$ Körpergröße und $y =$ Gewicht finden Sie im Applet, Datensatz 1 [Biometrik 1]. Die Koeffizienten ergeben sich zu

$$\begin{aligned} \hat{\beta}_0 &= -85.0209, \\ \hat{\beta}_1 &= 0.8787 \end{aligned}$$

mit der in Abbildung 1 dargestellten Regressionsgerade.

Rudimente der ANOVA (ANalysis Of VAriance). Ein erstes Indiz für die Anpassungsqualität des linearen Modells liefert die *Varianzanalyse*. Das arithmetische Mittel der y-Werte ist

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Die Abweichung des Messwertes y_i vom Mittelwert $\bar{y}$ ist $y_i - \bar{y}$. Die *totale quadratische Abweichung* oder *Gesamtvariabilität* der Daten ist

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Diese wird in zwei Teilkomponenten zerlegt, und zwar in

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

Die Gültigkeit dieser Beziehung werden wir weiter unten beweisen. Dies wird folgendermaßen interpretiert: $\hat{y}_i - \bar{y}$ ist die Abweichung des Prognosewertes vom Mittelwert, und

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

die *durch die Regression beschriebene Datenvariabilität* (Sum of Squares - Regression). Andererseits sind $e_i = y_i - \hat{y}_i$ gerade die Residuen, und

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

die Quadratsumme der Residuen, welche als *Restvariabilität* interpretiert wird, die durch das lineare Modell unerklärt verbleibt (Sum of Squares-Error). Wir haben also die Varianzzerlegung

$$S_{yy} = SS_R + SS_E$$

erhalten. Die Größe

$$R^2 = \frac{SS_R}{S_{yy}}$$

wird als *Bestimmtheitsmaß* bezeichnet und misst den Anteil der durch die Regression erklärten Variabilität an der Gesamtvariabilität. Im Grenzfall einer exakten Anpassung, wenn die Regressionsgerade genau durch alle Datenpunkte geht, ist $SS_E = 0$ und damit $R^2 = 1$. Ein kleines R^2 ist ein Indiz dafür, dass das lineare Modell weniger gut an die Daten passt. $R = \sqrt{R^2}$ ist übrigens der Absolutbetrag des empirische Korrelationskoeffizienten von $\mathbf{x}$ und $\mathbf{y}$.

Kurze Herleitung der Varianzzerlegung: Es ist

$$\begin{aligned} S_{yy} &= (\mathbf{y} - \bar{y}\mathbf{1})^\mathsf{T}(\mathbf{y} - \bar{y}\mathbf{1}) = \mathbf{y}^\mathsf{T}\mathbf{y} - \bar{y}(\mathbf{1}^\mathsf{T}\mathbf{y}) - (\mathbf{y}^\mathsf{T}\mathbf{1})\bar{y} + n\bar{y}^2 \\ &= \mathbf{y}^\mathsf{T}\mathbf{y} - n\bar{y}^2 - n\bar{y}^2 + n\bar{y}^2 = \mathbf{y}^\mathsf{T}\mathbf{y} - n\bar{y}^2; \\ SS_E &= \mathbf{e}^\mathsf{T}\mathbf{e} = (\mathbf{y} - \widehat{\mathbf{y}})^\mathsf{T}(\mathbf{y} - \widehat{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\mathsf{T}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= \mathbf{y}^\mathsf{T}\mathbf{y} - \widehat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{y} - \mathbf{y}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{y}^\mathsf{T}\mathbf{y} - \widehat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{y}, \end{aligned}$$

wobei wir für das letzte Gleichheitszeichen die Normalgleichungen $\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^\mathsf{T}\mathbf{y}$ und die Transpositionsformel $\widehat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{y} = (\mathbf{y}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol{\beta}})^\mathsf{T} = \mathbf{y}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol{\beta}}$ verwendet haben. Aus der Beziehung $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ folgt insbesondere $\mathbf{X}^\mathsf{T}\widehat{\mathbf{y}} = \mathbf{X}^\mathsf{T}\mathbf{y}$. Da die erste Zeile von $\mathbf{X}^\mathsf{T}$ aus lauter Einsern besteht, folgt daraus weiter $\mathbf{1}^\mathsf{T}\widehat{\mathbf{y}} = \mathbf{1}^\mathsf{T}\mathbf{y}$. Somit ist

$$\begin{aligned} SS_R &= (\widehat{\mathbf{y}} - \bar{y}\mathbf{1})^\mathsf{T}(\widehat{\mathbf{y}} - \bar{y}\mathbf{1}) = \widehat{\mathbf{y}}^\mathsf{T}\widehat{\mathbf{y}} - \bar{y}(\mathbf{1}^\mathsf{T}\widehat{\mathbf{y}}) - (\widehat{\mathbf{y}}^\mathsf{T}\mathbf{1})\bar{y} + n\bar{y}^2 \\ &= \widehat{\mathbf{y}}^\mathsf{T}\widehat{\mathbf{y}} - n\bar{y}^2 - n\bar{y}^2 + n\bar{y}^2 = \widehat{\boldsymbol{\beta}}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol{\beta}}) - n\bar{y}^2 = \widehat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{y} - n\bar{y}^2. \end{aligned}$$

Addition der erhaltenen Ausdrücke für SS_E und SS_R ergibt die gesuchte Formel.

Bemerkung: Im Falle der Regression mittels einer Geraden durch den Ursprung, also für das konstantenfreie Modell $y = \beta_1 x$, ist die Varianzzerlegungsformel ungültig. Für Regressionsmodelle ohne Konstante β_0 kann das Bestimmtheitsmaß keine Aussage machen. Insbesondere für die verschiedenen sequentiellen Bestimmtheitsmaße bei den multivariaten Modellen des

nächsten Abschnittes ist stets vorausgesetzt, dass die Konstante β_0 einer der Modellparameter ist.

Beispiel 1, fortgesetzt. Es ergibt sich

$$S_{yy} = 3535.9,\ SS_E = 1637.7,\ SS_R = 1898.2$$

und

$$R^2 = 0.5368,\ R = 0.7327,$$

also eine eher schlechte Anpassung (bzw. ein Indiz dafür, dass Körpergröße und Gewicht eher keinen linearen Zusammenhang haben).

Multiple lineare Regression

Bei der multiplen (mehrfachen, multivariaten) Regression hängt die Variable y nicht nur von einer Regressorvariablen x ab, sondern von mehreren, etwa $x_1, x_2, \ldots, x_k$ (man beachte die Änderung der Notation im Vergleich zum vorigen Abschnitt; die Messwerte der i-ten Regressorvariablen x_i werden mit zwei Indizes versehen, $x_{i1}, x_{i2}, \ldots, x_{in}$). Gesucht ist wieder ein lineares Modell

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

mit noch unbekannten Koeffizienten $\beta_0, \beta_1, \ldots, \beta_k$.

Bemerkung: Das allgemeine multiple lineare Modell $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ beinhaltet als Spezialfall die einfache lineare Regression mit mehreren nichtlinearen Formfunktionen, also etwa

$$y = \beta_0 + \beta_1 \varphi_1(x) + \beta_2 \varphi_2(x) + \cdots + \beta_k \varphi_k(x),$$

wobei $x_1 = \varphi_1(x)$, $x_2 = \varphi_2(x), \ldots, x_k = \varphi_k(x)$ als Regressorvariablen betrachtet werden. Insbesondere kann man polynomiale Ansätze

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$$

zulassen, oder noch allgemeiner Interaktionen zwischen mehreren Variablen, etwa

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

Alle diese Fälle werden rechnerisch genauso wie die Standardaufgabe der multiplen linearen Regression behandelt, nach Umbenennung zu den Variablen $x_1, x_2, \ldots, x_k$.

Die *Messdaten* (je n Stück) für die einzelnen Variablen stellen sich schematisch wie folgt dar:

Variable	y	x_1	x_2	$\ldots$	x_k
Messung Nr. 1	y_1	x_{11}	x_{21}	$\ldots$	x_{k1}
Messung Nr. 2	y_2	x_{12}	x_{22}	$\ldots$	x_{k2}
$\cdot$	$\cdot$	$\cdot$	$\cdot$		$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$		$\cdot$
Messung Nr. n	y_n	x_{1n}	x_{2n}	$\ldots$	x_{kn}

Jeder Wert y_i ist zu approximieren durch

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i,\ i = 1, \ldots, n$$

mit den Abweichungen ε_i. Die geschätzten Koeffizienten $\hat{\beta}_0,\ \hat{\beta}_1, \ldots, \hat{\beta}_k$ werden wieder als Lösung der Minimierungsaufgabe

$$L(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \varepsilon_i^2 \to \min!$$

gewonnen. Wir führen wieder die Vektor- und Matrixnotation ein:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix},\ \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{k1} \\ 1 & x_{12} & x_{22} & \ldots & x_{k2} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 1 & x_{1n} & x_{2n} & \ldots & x_{kn} \end{bmatrix},\ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix},\ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

In Kurzform lautet das lineare Modell an die Daten wieder

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Die bestangepassten Koeffizienten erhält man wie im vorigen Abschnitt nach der Formel

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

mit den Prognosewerten und Residuen

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}, \quad \mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}.$$

Die Varianzzerlegung

$$S_{yy} = SS_R + SS_E$$

ist weiterhin gültig; das *multiple Bestimmtheitsmaß*

$$R^2 = SS_R / S_{yy}$$

ist ein Indikator der Anpassungsgüte des Modells.

Beispiel 2: Eine Getränkeautomatenfirma möchte die Servicezeiten analysieren, also die Zeitspanne y, die ein Fahrer benötigt, um einen Automaten nachzufüllen und kurz zu warten. Als einflussreichste Parameter werden die Anzahl x_1 der nachgefüllten Produkteinheiten und die Distanz x_2, die der Fahrer zu Fuß zurücklegen muss, angesehen. Die Ergebnisse einer Beobachtung von 25 Servicevorgängen sind in Datensatz 3 [Lieferzeit] im Applet angegeben (Quelle: [5]); man beachte die geänderten Variablennamen $y \sim x_1, x_1 \sim x_2, x_2 \sim x_3$ in der Bezeichnungsweise des Applets.

Das Ergebnis der Regression ist

$$\boldsymbol{\beta} = \begin{bmatrix} 2.3412 \\ 1.6159 \\ 0.0144 \end{bmatrix},$$

also

$$\hat{y} = 2.3412 + 1.6159x_1 + 0.0144x_2$$

mit einem multiplen Bestimmtheitsmaß von

$$R^2 = 0.9596$$

und der Varianzzerlegung

$$S_{yy} = 5784.5,\ SS_R = 5550.8,\ SS_E = 233.7$$

Es werden also lediglich $(1 - R^2) \cdot 100\% \approx 4\%$ der Datenvariabilität durch die Regression nicht erklärt.

Bemerkung: Im Applet wird bewusst nicht zwischen einer abhängigen Variablen y und den unabhängigen Variablen unterschieden, sondern alle Variablen gleichwertig mit $x_1, x_2, x_3, \ldots$ durchnummeriert. Dies soll betonen, dass in der explorativen Datenanalyse in der Regel weder feststeht, welche Variable als die abhängige zu betrachten ist, noch welche Variablen aufzunehmen sind (die überdies erst durch Transformation aus den Daten hervorgehen können).

Modellanpassung und Variablenwahl. Ein stets schwieriges Problem ist die Entscheidung, welche Variablen ins Modell aufgenommen werden sollen. Hätte man vielleicht besser noch $x_3 = x_2^2$ und $x_4 = x_1x_2$ dazunehmen sollen, also das Modell

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_2^2 + \beta_4x_1x_2$$

verwenden sollen, oder vielleicht anschließend den Term β_2x_2 wieder streichen sollen? Es wäre schlecht, zu viele Variablen im Modell zu haben (sind es gleich

viele wie Messwerte, so kann man die Regression exakt durch die Messwerte legen - dann hätte das Modell keine Erklärungskraft mehr). Ein Kriterium wird sicher sein, ein möglichst großes R^2 zu erzielen. Ein anderes ist es, Variablen zu streichen, wenn deren Beibehaltung den Erklärungsanteil der Regression an der Gesamtvariabilität nicht wesentlich verändert.

Sequentielle Zerlegung von SS_R: Wir fügen schrittweise Variablen hinzu, betrachten also sequentiell die Modelle (mit zugehörigem SS_R):

$$\begin{aligned}
y = \beta_0 &\quad \dots \quad SS_R(\beta_0) = 0 \quad (\text{da } \beta_0 = \bar{y} \text{ ist}) \\
y = \beta_0 + \beta_1 x_1 &\quad \dots \quad SS_R(\beta_0, \beta_1) \\
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 &\quad \dots \quad SS_R(\beta_0, \beta_1, \beta_2) \\
&\quad \vdots \quad \vdots \\
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k &\quad \dots \quad SS_R(\beta_0, \beta_1, \dots, \beta_k) = SS_R\,.
\end{aligned}$$

Der zusätzliche Erklärungsanteil der Variablen x_1 ist

$$SS_R(\beta_1|\beta_0) = SS_R(\beta_0, \beta_1) - 0,$$

derjenige der Variablen x_2 (wenn x_1 im Modell aufgenommen ist) ist

$$SS_R(\beta_2|\beta_0, \beta_1) = SS_R(\beta_0, \beta_1, \beta_2) - SS_R(\beta_0, \beta_1),$$

derjenige der Variablen x_k (wenn $x_1, x_2, \dots, x_{k-1}$ im Modell sind) ist

$$SS_R(\beta_k|\beta_0, \beta_1, \dots, \beta_{k-1}) = SS_R(\beta_0, \beta_1, \dots, \beta_k) - SS_R(\beta_0, \beta_1, \dots, \beta_{k-1}).$$

Offensichtlich gilt

$$\begin{aligned}
SS_R(\beta_1|\beta_0) &+ SS_R(\beta_2|\beta_0, \beta_1) + SS_R(\beta_3|\beta_0, \beta_1, \beta_2) + \dots \\
&+ SS_R(\beta_k|\beta_0, \beta_1, \beta_2, \dots, \beta_{k-1}) = SS_R\,.
\end{aligned}$$

Dies zeigt, dass man das *sequentielle, partielle Bestimmtheitsmaß*

$$\frac{SS_R(\beta_j|\beta_0, \beta_1, \dots, \beta_{j-1})}{S_{yy}}$$

als Erklärungsanteil der Variablen x_j interpretieren kann, unter der Bedingung, dass die Variablen $x_1, x_2, \dots, x_{j-1}$ bereits im Modell sind. Dieses partielle Bestimmtheitsmaß hängt von der Reihenfolge ab, in der man die Variablen dazu nimmt. Die Abhängigkeit kann man eliminieren, indem man über alle möglichen Modellfolgen mittelt.

Mittlerer Erklärungsanteil einzelner Koeffizienten

Bildet man alle möglichen sequentiellen, partiellen Bestimmtheitsmaße, die durch Hinzunahme der Variablen x_j zu allen möglichen Kombinationen bereits aufgenommener Variablen erzielbar sind, und dividiert durch deren Anzahl, so erhält man ein Maß für den Einfluss der Variablen x_j auf die Erklärungskraft des Modells.

Dieses Konzept wurde unter anderem von [2] vorgeschlagen; wir empfehlen das Studium der Ausführungen in [1, 3, 4] dazu. Es verwendet keine wahrscheinlichkeitstheoretisch motivierten Indikatoren, sondern ausschließlich die auf Kombinatorik beruhende empirische Varianzzerlegung, also deskriptive Statistik. Insofern unterscheidet sich der Zugang von den weit verbreiteten Standardmethoden (F-Test, t-Test), die eine Normalverteilungshypothese über die Daten verlangen, und wird daher von einigen Autoren als *neodeskriptiv* bezeichnet.

Beispiel 2, fortgesetzt. Wir berechnen zunächst die Modelle

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

und erhalten daraus

$$SS_R(\beta_0, \beta_1) = 5382.4,\ SS_R(\beta_0, \beta_2) = 4599.1$$

(mit den Koeffizienten $\hat{\beta}_0 = 3.3208$, $\hat{\beta}_1 = 2.1762$ im ersten, $\hat{\beta}_0 = 4.9612$, $\hat{\beta}_2 = 0.0426$ im zweiten Fall). Mit den bereits berechneten Werten

$$SS_R(\beta_0, \beta_1, \beta_2) = SS_R = 5550.8,\ S_{yy} = 5784.5$$

erhalten wir die beiden Sequenzen

$$\begin{aligned} SS_R(\beta_1|\beta_0) &= 5382.4 \approx 93.05\% \text{ von } S_{yy} \\ SS_R(\beta_2|\beta_0, \beta_1) &= 168.4 \approx 2.91\% \text{ von } S_{yy} \end{aligned}$$

und

$$\begin{aligned} SS_R(\beta_2|\beta_0) &= 4599.1 \approx 79.51\% \text{ von } S_{yy} \\ SS_R(\beta_1|\beta_0, \beta_2) &= 951.7 \approx 16.45\% \text{ von } S_{yy}\,. \end{aligned}$$

Der mittlere Erklärungsanteil der Variable x_1 (oder des Koeffizienten β_1) ist

$$\frac{1}{2}\Big(93.05 + 16.45\Big)\% = 54.75\%,$$

jener der Variablen x_2 ist

$$\frac{1}{2}\Big(2.91 + 79.51\Big)\% = 41.21\%.$$

Das Ergebnis ist in Abbildung 4 dargestellt:

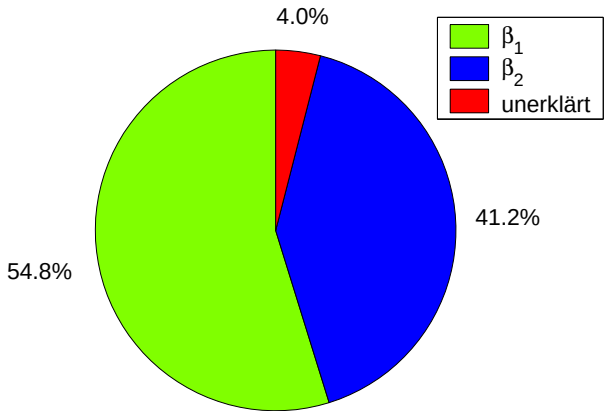

Abbildung 4: Mittlere Erklärungsanteile der einzelnen Koeffizienten.

Numerische Berechnung der mittleren Erklärungsanteile. Im Falle von mehr als zwei unabhängigen Variablen ist zu beachten, dass alle möglichen Sequenzen (dargestellt durch Permutationen der Variablennummern) berücksichtigt werden. Dies soll beispielhaft mit drei Variablen x_1, x_2, x_3 vorgeführt werden. In der Tabelle finden sich links die $3! = 6$ Permutationen von $\{1, 2, 3\}$, rechts jeweils die drei sequentiell gewonnenen Erklärungsanteile.

1	2	3	$SS_R(\beta_1\|\beta_0)$	$SS_R(\beta_2\|\beta_0,\beta_1)$	$SS_R(\beta_3\|\beta_0,\beta_1,\beta_2)$
1	3	2	$SS_R(\beta_1\|\beta_0)$	$SS_R(\beta_3\|\beta_0,\beta_1)$	$SS_R(\beta_2\|\beta_0,\beta_1,\beta_3)$
2	1	3	$SS_R(\beta_2\|\beta_0)$	$SS_R(\beta_1\|\beta_0,\beta_2)$	$SS_R(\beta_3\|\beta_0,\beta_2,\beta_1)$
2	3	1	$SS_R(\beta_2\|\beta_0)$	$SS_R(\beta_3\|\beta_0,\beta_2)$	$SS_R(\beta_1\|\beta_0,\beta_2,\beta_3)$
3	1	2	$SS_R(\beta_3\|\beta_0)$	$SS_R(\beta_1\|\beta_0,\beta_3)$	$SS_R(\beta_2\|\beta_0,\beta_3,\beta_1)$
3	2	1	$SS_R(\beta_3\|\beta_0)$	$SS_R(\beta_2\|\beta_0,\beta_3)$	$SS_R(\beta_1\|\beta_0,\beta_3,\beta_2)$

Offensichtlich sind die Zeilensummen jeweils gleich SS_R, sodass die Summe aller Einträge gleich $6 \cdot SS_R$ ist. Man beachte, dass unter den 18 SS_R-Werten tatsächlich nur 12 verschiedene sind.

Der mittlere Erklärungsanteil der Variablen x_1 ist definiert durch

$$\begin{aligned} M_1 &= \frac{1}{6}\Big(SS_R(\beta_1|\beta_0) + SS_R(\beta_1|\beta_0) + SS_R(\beta_1|\beta_0,\beta_2) + SS_R(\beta_1|\beta_0,\beta_3) \\ &+ SS_R(\beta_1|\beta_0,\beta_2,\beta_3) + SS_R(\beta_1|\beta_0,\beta_3,\beta_2)\Big) \end{aligned}$$

und analog für die anderen Variablen.

Damit ist garantiert, dass

$$M(1) + M(2) + M(3) = SS_R$$

ist, womit die Gesamtzerlegung korrekt auf Eins summiert:

$$\frac{M(1)}{S_{yy}} + \frac{M(2)}{S_{yy}} + \frac{M(3)}{S_{yy}} + \frac{SS_E}{S_{yy}} = 1.$$

Für eine genauere Analyse der zu Grunde liegenden Kombinatorik, nötigen Modifikationen im Falle der Kollinearität der Daten (linearer Abhängigkeit der Spalten der Matrix $\mathbf{X}$) und für eine Diskussion der Aussagekraft des mittleren Erklärungsanteils verweisen wir auf die oben zitierte Literatur.

Literatur

[1] A. Chevan, M. Sutherland, Hierarchical partitioning. The American Statistician **45**(1991), 90 - 96.

[2] W. Kruskal, Relative importance by averaging over orderings. The American Statistician **41**(1987), 6 - 10.

[3] N. Fickel, Partition of the coefficient of determination in multiple regression. In: K. Inderfurth, G. Schwödiauer, W. Domschke, F. Juhnke, P. Kleinschmidt, G. Wäscher (Hrsg.), Operations Research Proceedings 1999. Springer, Berlin 2000, 154 - 159.

[4] N. Fickel, Sequential regression: a neodescriptive approach to multicollinearity. Paper ewp-em/0004009, EconWPA (2000), http://econwpa.wustl.edu/eprints/em/papers/0004/0004009.abs

[5] D.C. Montgomery, E.A. Peck, Introduction to Linear Regression Analysis. John Wiley & Sons, New York 1982.

[6] M. Oberguggenberger, A. Ostermann, Analysis für Informatiker. Springer-Verlag, Berlin 2005.