Petra Stein, Monika Pavetic, Marcel Noack

Multivariate Analyseverfahren

INHALTSVERZEICHNIS

KAPITEL 1

MULTIPLE REGRESSION

Bei der multiplen Regression handelt es sich um ein Standardverfahren der multivariaten Statistik. Hierbei wird von einem linearen Zusammenhang zwischen einer abhängigen, zu erklärenden Variablen mit einer -oder mehreren- unabhängigen, erklärenden Variablen ausgegangen. Um die Idee besser erfassen zu können wenden wir uns zunächst der einfachen linearen Regression zu, die aus dem Grundstudium noch in vager Erinnerung sein sollte.
In der einfachen linearen Regression versuchen wir, eine metrische abhängige Variable y durch eine unabhängige metrische Variable x vorherzusagen. Hierzu benötigen wir eine Gerade (daher der Name *lineare* Regression), die sogenannte Regressionsgerade, auf der die vorhergesagten Punkte liegen. Ihre Gleichung lautet $\hat{y}_i = b_0 + b_1 x_i$. Diese Gerade ist die Optimale Gerade durch die von den Variablen x und y gebildete Punktwolke. Sehen wir uns als Beispiel den Scatterplot für folgende Daten an:

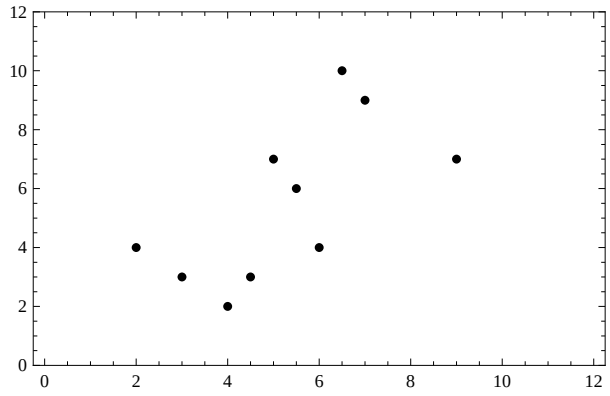


Abbildung 1.1: Scatterplot

x	2	3	4	4.5	5	5.5	6	6.5	7	9
y	4	3	2	3	7	6	4	10	9	7

Wie erhalten wir unsere Regressionsgerade für die obige Graphik? Wir benötigen eigentlich nur eine Handvoll Dinge:

1. Das arithmetische Mittel von $\overline{y}$, dies beträgt 5.5.
2. Die Regressionsgerade $\hat{y} = b_0 + b_1 x_i$. Die Parameter a und b sind momentan noch unbekannt, dies ist aber nicht weiter schlimm.
3. Das Konzept der Varianzzerlegung.
4. Die Daten, natürlich.

1.1 Least Squares & Varianzzerlegung

Sehen wir uns einmal den Scatterplot inklusive $\overline{y}$ und $\hat{y}_i$ an:

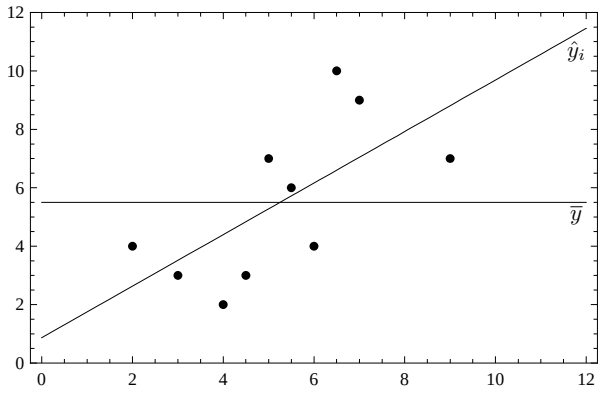


Abbildung 1.2:

Hier treffen wir auf das Verfahren der kleinsten Fehlerquadrate (oder englisch: least squares). Entwickelt wurde es vom dem Mathematiker Carl Friedrich Gauss, der uns von den letzten 10 DM Scheinen bekannt ist. Mit diesem Verfahren lässt sich die perfekte Regressionsgerade mit mathematischer Genauigkeit bestimmen.
Bevor wir uns ansehen, wie die Regressionsgleichung ($\widehat{y_i} = b_0 + b_1 \cdot x_i$) bestimmet wird, müssen wir uns noch mit einigen Grundlagen vertraut machen, die in diesem Zusammenhang eine Rolle spielen.

- Die beste Vorhersage von y *ohne Kenntnis von* x ist das arithmetische Mittel von y, also $\overline{y}$. Dies gilt ebenso, wenn zwischen x und y *kein* Zusammenhang besteht, also $r = b = 0$.

- Die beste Vorhersage *mit Kenntnis von* x ist nicht mehr $\bar{y}$ sondern $\hat{y}$, also die Regressionsgerade.

Für jeden einzelnen Fall sind 3 Werte von Bedeutung. Einmal der wirklich gemessene Wert y_i, der von der Regression vorhergesagte Wert $\hat{y}_i$ sowie $\bar{y}$

- Die Abweichung der gemessenen Werte y von $\bar{y}$ wird gesamte Abweichung, gesamte Streuung oder gesamter Fehler genannt. Wir werden später den Zusammenhang mit der Varianz von y sehen.
- Die Abweichung der vorhergesagten Werte $\hat{y}$ von $\bar{y}$ wird erklärter Fehler genannt. Dies ist die Verbesserung der Vorhersage, die die Regressionsgerade gegenüber $\bar{y}$ bietet.
- Die Abweichung der gemessenen y-Werte von den vorhergesagten $\hat{y}$-Werten der Regressionsgeraden ist der nicht erklärte Fehler von y.

Die Gesamte Streuung von y setzt sich aus zwei Komponenten zusammen: dem Teil, der durch x erklärt wird, und dem Teil, der nicht durch x erklärt wird. In Graphik 1.2 sehen wir einen Scatterplot mit eingezeichneter Regressionsgerade $\hat{y}$ und arithmetischem Mittel $\bar{y}$. In Graphik 1.3 sind zusätzlich die eben angeführten Begriffe für einen einzelnen Datenpunkt eingetragen.

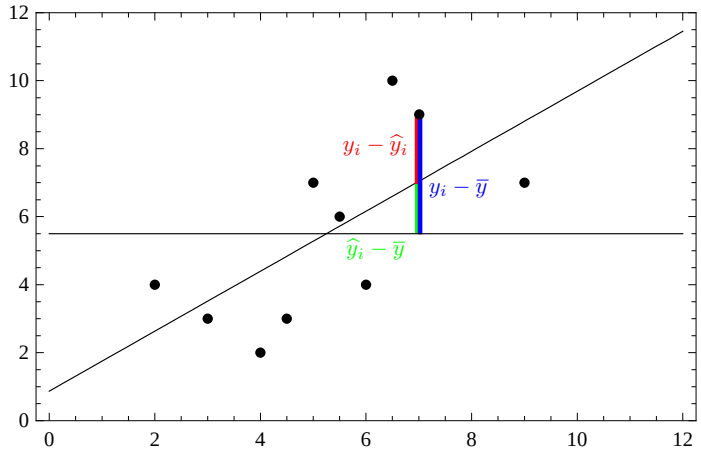


Abbildung 1.3: Beispiel mit einem Datenpunkt

Was bedeuten diese Begriffe nun?
Die blaue Strecke ist die Abweichung des Punktes y von seinem arithmetischen Mittel $\bar{y}$. Dies ist der individuelle Beitrag des Punktes zum Gesamten Fehler von y. Die rote Strecke ist der Beitrag zum nicht-erklärten Fehler, also der Abweichung von der Regressionsgerade. Die grüne Strecke der Beitrag zum erklärten Fehler, also was die Regressionsgerade "besser" vorhersagt als das arithmetische Mittel $\bar{y}$.
Wie kommen wir an die Werte für die verschiedenen Komponenten? Für einen einzigen Fall gilt folgendes:

- $\widehat{y}_i - \bar{y}$ = erklärter Fehler

- $y_i - \widehat{y}_i$ = nicht erklärter Fehler
- $y_i - \overline{y}$ = gesamter Fehler

Da wir es aber nicht nur mit *einem einzelnen* Fall zu tun haben, sondern mit *mehreren*, nämlich n Fällen, müssen wir mehrere Datenpunkte mit jeweils einem erklärten Fehler, nicht erklärten Fehler und gesamten Fehler berücksichtigen.

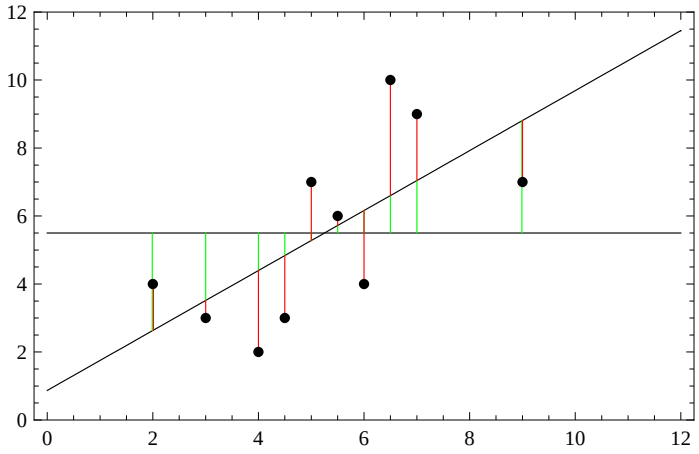


Abbildung 1.4: Beispiel mit allen Datenpunkten

Für alle Datenpunkte werden nun die einzelnen Fehlerwerte aufsummiert. Wir haben es also hier nicht mit

$$y_i - \overline{y}, \widehat{y}_i - \overline{y}, y_i - \widehat{y}_i$$

für einen einzelnen Datenpunkt, sondern mit der Summe mehrerer -nämlich n- Datenpunkten, unterteilt nach den verschiedenen Fehlerarten zu tun:

$$\sum_{i=1}^{n} y_i - \overline{y}, \sum_{i=1}^{n} \widehat{y}_i - \overline{y}, \sum_{i=1}^{n} y_i - \widehat{y}_i$$

Hierbei stoßen wir auf ein Problem. Die verschiedenen Summen ergeben Null. Diese Eigenschaft tritt ebenfalls bei der Berechnung der Varianz auf. Wir bedienen uns hier eines Tricks um dies zu vermeiden: Wir quadrieren die Differenzen bevor wir sie summieren. Es resultiert nun:

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

Diese 3 Terme werden "Fehlerquadratsummen" genannt. Hoffentlich ist deutlich geworden, warum. Es handelt sich hierbei um die quadrierten Summen der Differenzen, oder anders gesagt die quadrierten Summen der Fehler. Es besteht folgender Zusammenhang zwischen diesen Termen:

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

Teilen wir alle 3 Terme durch $n-1$ erkennen wir etwas. Die totale Streuung wird durch die gleiche Formel ausgedrückt, wie die Varianz von y. Somit ist die Varianz von y in 2 Teile zerlegbar. Einen Teil, der durch die Regression *erklärt* wird, und ein Teil der durch die Regression *nicht erklärt* wird.

$$\frac{\sum(y_i-\overline{y})^2}{n-1} = \frac{\sum(\widehat{y}_i-\overline{y})^2}{n-1} + \frac{\sum(y_i-\widehat{y}_i)^2}{n-1}$$

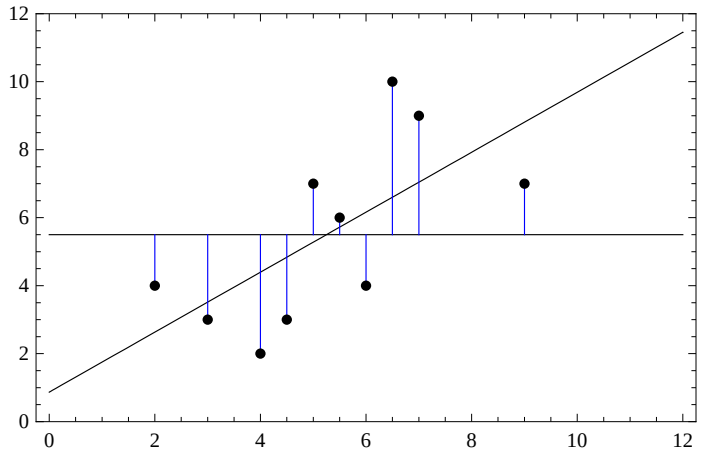


Abbildung 1.5: Gesamter Fehler

In der folgenden Tabelle sind die einzelnen Fehler berechnet.

Tabelle 1.1: Fehlerberechnung

x	y	$\hat{y}$	ef $\hat{y}-\overline{y}$	ef^2 $(\hat{y}-\overline{y})^2$	nef $y-\hat{y}$	nef^2 $(y-\hat{y})^2$	gf $y-\overline{y}$	gf^2 $(y-\overline{y})^2$
2	4	2.633	−2.876	8.220	1.367	1.869	−1.5	2.25
3	3	3.515	−1.985	3.940	−0.515	0.265	−2.5	6.25
4	2	4.400	−1.103	1.216	−2.400	5.747	−3.5	12.25
4.5	3	4.838	−0.662	0.438	−1.838	3.380	−2.5	6.25
5	7	5.279	−0.221	0.048	1.721	2.960	1.5	2.25
5.5	6	5.720	0.221	0.048	0.280	0.078	0.5	0.25
6	4	6.161	0.662	0.438	−2.161	4.673	−1.5	2.25
6.5	10	6.603	1.103	1.216	3.397	11.542	4.5	20.25
7	9	7.044	1.544	2.383	1.956	3.827	3.5	12.25
9	7	8.808	3.308	10.943	−1.808	3.269	1.5	2.25
52.5	55		−0.003	28.890	−0.001	37.610	0	66.5

Lassen wir uns die Varianz von y auszurechnen erhalten wir 7.389 als Ergebnis. Teilen wir die Quadratsumme des gesamten Fehlers (66.5) durch $n-1$, in unserm Falle also 9 erhalten wir ebenfalls die Varianz:

$$\frac{66.5}{9} = 7.389$$

Was das Verfahren der kleinsten Fehlerquadrate tut, ist die Regressionsgerade aus der unendlichen Anzahl von möglichen Regressionsgeraden zu bestimmen, für die die Quadratsumme des nicht erklärten Fehlers *minimal* ist. Es gibt keine Regressionsgerade für die dieser Wert geringer ist. Der erklärte Teil ist also analog maximal.

1.2 OLS mathematisch

Wie bereits erwähnt lautet die Gleichung der Regressionsgerade:

$$\hat{y}_i = b_0 + b_1 x_i$$

Die *nicht erklärte Streuung* ist wie oben dargelegt bestimmt, und soll minimal sein:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \frac{1}{n} \stackrel{!}{=} min$$

In folgender Graphik sehen wir, was dies bedeutet: Auf der x und y-Achse sind jeweils mögliche Werte für b_0 und b_1 abgetragen. Am tiefsten Punkt des Graphen befindet sich die OLS-Lösung für b_0 und b_1.

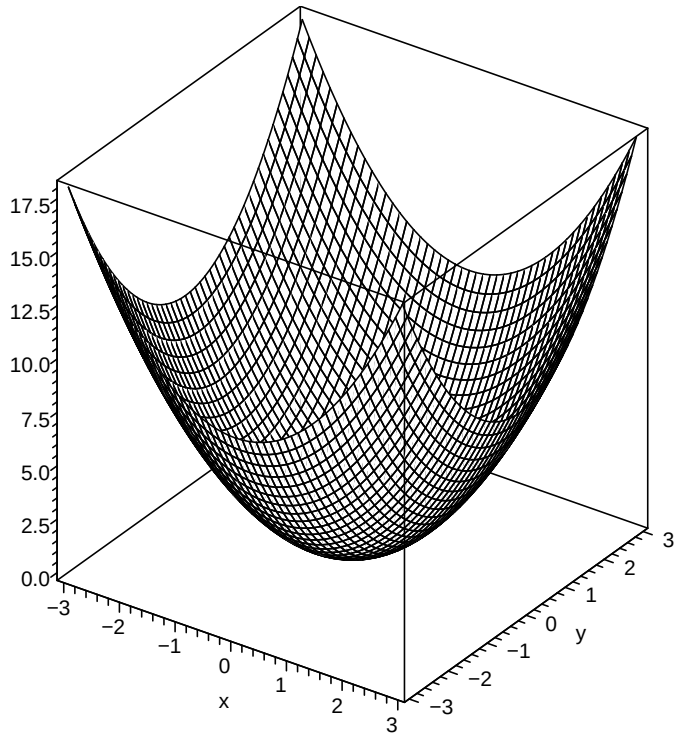


Abbildung 1.6: Bestimmung des kleinsten Fehlerquadrates

Als ersten Schritt substituieren wir $\hat{y}_i$ in $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \frac{1}{n}$ durch $b_0 + b_1 x_i$:

$$\sum_{i=1}^{n} (y_i - (a + bx_i))^2 \frac{1}{n}$$

Ausmultiplizieren

$$\sum_{i=1}^{n} (y_i^2 - 2(b_0 + b_1 x_i) y_i + (b_0 + b_1 x_i)^2) \frac{1}{n}$$

$$\sum_{i=1}^{n}(y_i^2 - 2y_ib_0 - 2y_ib_1x_i + b_0^2 + 2b_0b_1x_i + b_1^2x_i^2)\frac{1}{n}$$

Partielle Ableitung nach b_0	Partielle Ableitung nach b_1
$\frac{\partial f(b_0,b_1)}{\partial b_0}$	$\frac{\partial f(b_0,b_1)}{\partial b_1}$
$\sum \frac{-2y_i+2b_0+2b_1x_i}{n} = 0$	$\sum \frac{-2y_ix_i+2b_0x_i+2b_1x_i^2}{n} = 0$
$2\frac{-\sum y_i+\sum b_0+\sum b_1\sum x_i}{n} = 0$	$2\frac{-\sum y_ix_i+\sum b_0\sum x_i+\sum b_1\sum x_i^2}{n} = 0$
$-\frac{\sum y_i}{n} + \frac{\sum b_0}{n} + \frac{\sum b_1\sum x_i}{n} = 0$	$-\frac{\sum x_iy_i}{n} + \frac{\sum b_0\sum x_i}{n} + \frac{\sum b_1\sum x_i^2}{n} = 0$
$-\bar{y} + \frac{nb_0}{n} + \frac{nb_1}{n}\bar{x} = 0$	$-\overline{xy} + \frac{nb_0}{n}\bar{x} + \frac{nb_1}{n}\bar{x^2} = 0$
$-\bar{y} + b_0 + b_1\bar{x} = 0$	$-\overline{xy} + b_0\bar{x} + b_1\bar{x^2} = 0$
$b_0 + b_1\bar{x} = \bar{y}$	$-\overline{xy} + (\bar{y} - b_1\bar{x})\bar{x} + b_1\bar{x^2}$
$b_0 = \bar{y} - b_1\bar{x}$	$-\overline{xy} + \bar{x}\bar{y} - b_1\bar{x}\bar{x} + b_1\bar{x^2} = 0$
	$-\overline{xy} + \bar{x}\bar{y} - b_1\bar{x}^2 + b_1\bar{x^2} = 0$

Hier ist die linke partielle Ableitung (nach b_0) beendet, die Gleichung für die Regressionskonstante (das Interzept) im bivariaten Fall lautet also: $b_0 = \bar{y} - b_1\bar{x}$. Die rechte partielle Ableitung nach b_1 ist noch nicht vollendet, wir müssen also noch weiter rechnen.

	Partielle Ableitung nach b
	$-\overline{xy} + \bar{x}\bar{y} - b_1(\bar{x}^2 - \bar{x^2}) = 0$
	$-b_1(-\bar{x^2} + \bar{x}^2) = \overline{xy} - \bar{x}\bar{y}$
	$b_1(\bar{x^2} - \bar{x}^2) = \overline{xy} - \bar{x}\bar{y}$
	$b_1 = \frac{\overline{xy}-\bar{x}\bar{y}}{\bar{x^2}-\bar{x}^2}$
$b_0 = \bar{y} - b_1\bar{x}$	$b_1 = \frac{s_{xy}}{s_x^2}$

So ergeben sich also die Parameterschätzer für die Regressionsgerade:

$$\hat{y_i} = b_0 + b_1x_i, \text{ mit}$$
$$b_0 = \bar{y} - b_1\bar{x} \text{ und } b_1 = \frac{s_{xy}}{s_x^2}$$

An dieser Stelle wird auf die Inspektion der zweiten Ableitung verzichtet, die anzeigt ob es sich bei dem Extremwert um ein Minimum oder Maximum handelt. Es handelt sich an dieser Stelle um ein Minimum.

1.2.1 Multipler Fall

Während die bivariate Regression einen Schätzer für den Achsenabschnitt b_0 und einen Schätzer für den Steigungsparameter b_1 benötigt, sind in der multiplen Regression mehrere Steiungsparameter b_j zu schätzen. Für b_0 führt dies im Falle einer Regression y auf x_1, x_2 und x_3 zur Schätzung

$$b_0 = \bar{y} - |b_{yx_1-x_2x_3} \cdot \bar{x}_1 + b_{yx_2-x_1x_3} \cdot \bar{x}_2 + b_{yx_3-x_1x_2} \cdot \bar{x}_3|$$

mit den partiellen Regressionskoeffizienten

$$\begin{array}{rcccl} b_1 & = & b_{yx_1-x_2x_3} & = & b_{y(1-23)} \\ b_2 & = & b_{yx_2-x_1x_3} & = & b_{y(2-13)} \\ b_3 & = & b_{yx_3-x_1x_2} & = & b_{y(3-12)} \end{array}$$

1.2.2 Matrixnotation

Im multiplen Fall bietet es sich an, die Regressionsgleichung in Matrixnotation zu notieren:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$

Durch Umstellen erhalten wir,

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}$$

was minimiert werden soll. Das führt zu

$$\boldsymbol{e'e} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \stackrel{!}{=} \min$$

Die 1. Ableitung wird gleich Null gesetzt, um den potentiellen Extremwert dieser Funktion zu ermitteln. Danach wird die 2. Ableitung gleich Null gesetzt, um zu ermitteln, ob es sich um ein Minimum, Maximum oder einen Sattelpunkt handelt, darauf wird an dieser Stelle verzichtet. Für ein Minimum muss die 2. Ableitung positiv sein.
Nullsetzen der 1. Ableitung

$$\frac{\partial}{\partial \boldsymbol{b}}(\boldsymbol{e'e}) = 0$$

Substituieren

$$\frac{\partial}{\partial \boldsymbol{b}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = 0$$

Transponieren

$$\frac{\partial}{\partial \boldsymbol{b}}(\boldsymbol{y'} - \boldsymbol{b'X'})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = 0$$

Ausmultiplizieren

$$\frac{\partial}{\partial \boldsymbol{b}}(\boldsymbol{y'y} - \boldsymbol{y'X}\boldsymbol{b} - \boldsymbol{b'X'y} + \boldsymbol{b'X'X}\boldsymbol{b}) = 0$$

Weil $\boldsymbol{y'X}\boldsymbol{b}$ und $\boldsymbol{b'X'y}$ Skalare sind, gilt $\boldsymbol{y'X}\boldsymbol{b} = (\boldsymbol{y'X}\boldsymbol{b})' = \boldsymbol{b'X'y}$, da $\varphi = \varphi'$:

$$\frac{\partial}{\partial \boldsymbol{b}}(\boldsymbol{y'y} - 2\boldsymbol{b'X'y} + \boldsymbol{b'X'X}\boldsymbol{b}) = 0$$

$$\frac{\partial(\boldsymbol{y'y})}{\partial \boldsymbol{b}} - \frac{\partial(2\boldsymbol{b'X'y})}{\partial \boldsymbol{b}} + \frac{\partial(\boldsymbol{b'X'Xb})}{\partial \boldsymbol{b}} = 0$$

Ableiten

$$0 - 2\boldsymbol{X'y} + 2\boldsymbol{X'Xb} = 0$$

Ergibt

$$2\boldsymbol{X'Xb} = 2\boldsymbol{X'y}$$

$$\boldsymbol{X'Xb} = \boldsymbol{X'y}$$

Linksseitig mit $(\boldsymbol{X'X})^{-1}$ multiplizieren, also durch $(\boldsymbol{X'X})^{-1}$ teilen. ("$\frac{1}{\boldsymbol{X'X}}$")

$$(\boldsymbol{X'X})^{-1}\boldsymbol{X'Xb} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$$

Wir erhalten das Ergebnis für $\boldsymbol{b}$

$$\boldsymbol{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$$

1.3 Partielle Korrelation und Regression

Um die Begriffe partielle Korrelation und partielle Regression zu verdeutlichen betrachten wir ein Beispiel: Eine Untersuchung ergab, dass die Fahrleistung (x_1) mit steigendem Alter (x_2) abnimmt, obwohl zu erwarten ist, dass sie zunimmt. Hier spielt aber eine andere Variable auch eine wichtige Rolle, das Alter des Führerscheinerwerbs (x_3). Partialisiert oder kontrolliert man nun die Variablen x_1 und x_2 mit x_3, dann ergibt sich der erwartete positive Zusammenhang.Das Konstanthalten oder Herauspartialisieren von Einflüssen dritter Variablen erfolgt mittels Regression.Wir führen eine einfache Regression von Fahrleistung (x_1) auf Alter Führerscheinerwerb (x_3) durch:

$$\hat{x}_1 = a_{13} + b_{13}x_3 \text{ mit } x_1 = a_{13} + b_{13}x_3 + \varepsilon_1$$

und eine einfache Regression von Alter (x_2) auf Alter Führerscheinerwerb (x_3)

$$\hat{x}_2 = a_{23} + b_{23}x_3 \text{ mit } x_2 = a_{23} + b_{23}x_3 + \varepsilon_2$$

Die Fehlervarianzen $s^2_{\varepsilon_1}$ und $s^2_{\varepsilon_2}$ sind die Anteile von x_1 und x_2, die durch x_3 nicht geklärt werden. Führen wir nun mit diesen Restvarianzen eine Korrelation oder Regression durch,

$$\hat{x}_{1-3} = a_{12-3} + b_{12-3}x_{2-3}$$

analysieren wir x_1 und x_2 unter Konstanthalten von x_3. Diese wurde mittels Regression herauspartialisiert.

Konventionen

- Partialvariable 1. Ordnung
  x_{1-3} ist die Variable x_1 ohne x_3
  x_{2-3} ist die Variable x_2 ohne x_3
- Partialvariable 2. Ordnung
  x_{1-34} ist die Variable x_1 ohne x_3 und x_4
  x_{2-34} ist die Variable x_2 ohne x_3 und x_4
- Partialvariable n-ter Ordnung
  $x_{1-34\ldots n}$ ist die Variable x_1 ohne x_3 und $x_4 \ldots x_n$
  $x_{2-34\ldots n}$ ist die Variable x_2 ohne x_3 und $x_4 \ldots x_n$

Partielle Korrelation

$$r_{12-3} = \frac{\mathrm{cov}(x_{1-3}, x_{2-3})}{\sqrt{\mathrm{Var}(x_{1-3})\mathrm{Var}(x_{2-3})}} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Partielle Regression

$$b_{z_1 z_2 - z3} = \frac{\mathrm{cov}(z_{1-3}, z_{2-3})}{\mathrm{Var}(z_{2-3})} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}$$

beziehungsweise

$$b_{x_1 x_2 - x3} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \cdot \frac{s_1}{s_2}$$

1.3.1 Semipartielle Korrelation und Regression

Möchte man den Einfluss der Variablen x_3 nur aus einer Variablen herauspartialisieren, z.B. nur x_2, so ergibt sich für die Korrelation:

$$r_{1(2-3)} = \frac{\mathrm{cov}(z_1, z_{2-3})}{\sqrt{\mathrm{Var}(z_1)}\sqrt{\mathrm{Var}(z_{2-3})}} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}$$

Der semipartielle Regressionskoeffizient entspricht dem partiellen Regressionskoeffizienten. Die Korrelation von x_1 mit x_2 ohne x_3 ist identisch mit der Regression von x_1 auf x_2 ohne x_3.

$$r_{1(2-3)} = b_{1(2-3)}$$

Die Regressionskoeffizienten können durch

- Simultane Schätzung (partielle Regressionskoeffizienten)
- Schrittweise Schätzung (semipartiell Regressionskoeffizienten)

erfolgen.

Simultane Schätzung

Beispiel vierfache multiple Regression $\mathrm{x}_1 = f(x_2,\ x_3,\ x_4)$
Die Regressionskoeffizienten geben Auskunft darüber, wie viel $\hat{x}_1$ sich verändert, wenn die unabhängige Variable x_j unter Konstanthalten der übrigen unabhängigen Variablen $x_2, \ldots x_{j-1},\ x_{j+1}, \ldots,\ x_k$ um eine Einheit wächst.

$$\begin{aligned} \hat{x}_1 &= b_0 + b_2x_2 + b_3x_3 + b_4x_4 \\ b_1 &= b_{12-34} \\ b_2 &= b_{13-24} \\ b_3 &= b_{14-23} \end{aligned}$$

Schrittweise Schätzung

Die Funktion $x_1 = f(x_2,\ x_3,\ x_4)$ ist als Funktion fortschreitender Partialvariablen anzusehen. Die Regressionskoeffizienten sind semipartielle Koeffizienten. Variablen werden schrittweise hinzugenommen.

$$\begin{aligned} \hat{x}_1 &= b_0 + b_2x_2 + b_3x_3 + b_4x_4 \\ b_2 &= b_{12} \\ b_3 &= b_{1(3-2)} \\ b_4 &= b_{1(4-2,3)} \end{aligned}$$

1.4 Relevante Koeffizienten

In der multiplen Regression wird von einem *linearen Zusammenhang* zwischen einer abhängigen Variablen y und $J \geq 2$ unabhängigen Variablen x ausgegangen. Die Modellgleichung für die Grundgesamtheit lautet:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

oder in kürzerer Schreibweise

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} + \varepsilon_i.$$

In der uns zugänglichen Stichprobe lautet die Modellgleichung

$$\hat{y}_i = b_0 + \sum_{j=1}^{J} b_j x_{ij} + e_i$$

Da wir von b_j auf β_j rückschliessen wollen. Im Zusammenhang mit der Regression existieren eine Reihe von Koeffizienten, die die Regression bestimmen und die Güte der Anpassung beschreiben. Es folgt ein kurzer Überblick.

1.4.1 Determinationskoeffizient r^2

Im bivariaten Fall gibt $r^2 \cdot 100$ den prozentualen Anteil der Varianz der abhängigen Variablen y an, der durch die unabhängige Variable x erklärt/vorhergesagt wird. Er berechnet sich über:

$$r^2 = \frac{\mathrm{QS}_{\text{gesamt}} - \mathrm{QS}_{\text{nicht erklärt}}}{\mathrm{QS}_{\text{gesamt}}} = \frac{\mathrm{QS}_{\text{erklärt}}}{\mathrm{QS}_{\text{gesamt}}}$$

$$r^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2 - \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Je größer der Wert der erklärten Quadratsumme, desto größer wird r^2. Analog wird r^2 kleiner, wenn die Quadratsumme der Residuen größer wird. Ziehen wir die Wurzel aus r^2 erhalten wir den Korrelationskoeffizienten nach Pearson. Um zu wissen, ob es sich um eine negative oder positive Korrelation handelt, müssen wir inspizieren, ob der Regressionskoeffizient ein positives oder negatives Vorzeichen besitzt.

Korrigiertes r^2

Durch Hinzunahme weiterer unabhängiger Variablen kann das normale r^2 allenfalls steigen, aber nicht sinken, unabhängig, ob die weiteren Variablen einen Erklärungsbeitrag leisten, oder nicht. Um dieses Problem zu beheben wurde der korrigierte r^2-Koeffizient für den multivariaten Fall entwickelt. Er berechnet sich über folgende Formel:

$$r_k^2 = 1 - \frac{n-1}{n-k}(1 - r^2)$$

Wobei k die Anzahl der Parameter und n die Anzahl der Fälle angibt. Es bleibt jedoch anzumerken, dass gegen falsch spezifizierte Modelle nur theoretische Überlegungen und sorgfältige Diagnostik hilft.

Multiples r^2

Beim multiplen r^2 handelt es sich um den Anteil erklärter Varianz relativ zur Gesamtvarianz, wie auch in der bivariaten Regression. Schreibt man nun die mittels semipartieller Regressionskoeffizienten ermittelte erklärte Varianz von x

$$s_{\hat{x}}^2 = b_{12}^2 \cdot s_2^2 + b_{1(3-2)}^2 \cdot s_{3-2}^2 + \ldots + b_1^2 k - 2,\ 3, \ldots,\ k-1 \cdot s_{k-2,\ 3,\ldots,\ k-1}^2$$

und dividiert durch die Gesamtvarianz von x, so erhält man den multiplen Determinationskoeffizienten als Summe semipartieller Determinationskoeffizienten fortschreitend höherer Ordnung.

$$r_k^2 = \frac{b_{12}^2 \cdot s_2^2}{s_1^2} + \frac{b_{1(3-2)}^2 \cdot s_{3-2}^2}{s_1^2} + \ldots + \frac{b_{1(k-2,\ 3,\ldots,\ k-1)}^2 \cdot s_{k-2,\ 3,\ldots,\ k-1}^2}{s_1^2}$$

$$r_k^2 = r_{12}^2 + r_{1(3-2)}^2 + \cdots + r_{1(k-2,\ 3,\ldots,\ k-1)}^2$$

Damit wird eine Aussage über die zusätzliche Erklärungskraft durch Hinzunahme einer weiteren Variablen möglich.

1.4.2 Der Standardfehler des Schätzers: RMSE

Hierbei handelt es sich um die Quadratwurzel aus den durchschnittlichen residuen des Modells.

$$\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\widehat{y}_i\right)^2}{n-k}}$$

Diese Maßzahl ist in der Einheit der abhängigen Variablen angegeben, sie gibt den durchschnittlichen Wert an, den unsere Schätzung im Schnitt *daneben liegt*. Je kleiner dieser Wert, desto besser ist die Anpassungsgüte des Modells.

1.4.3 Der F-Test

Beim F-Test handelt es sich um einen Signifikanztest des Gesamtmodells. Der F-Wert berechnet sich über:

$$\frac{k-1}{n-k}\cdot\frac{r^2}{1-r^2}$$

Getestet werden die Hypothesen:

$$H_0 : \beta_i = 0$$

Kann die H_0 nicht verworfen werden, so ist das Ergebnis der Regression als Zufallsergebnis zu bewerten.

1.4.4 Regressionskoeffizienten

Die Regressionskoeffizienten der multiplen Regression werden so berechnet, dass die Werte jeder x_j-Variable um diejenigen Anteile bereinigt werden, die durch lineare Effekte der anderen x_j-Variablen verursacht werden. Man spricht auch von *herauspartialisieren*. Es wird also eine Regression der abhängigen Variable y auf die nun *kontrollierten* Variablen x_j durchgeführt.

Beispiel: Nehmen wir die Regression $y_i = b_0+b_1x_{i1}+b_2x_{i2}+b_3x_{i3}$: Betrachten wir den Koeffizienten b_1, so gibt er uns den Wert der Veränderung für y_i an, wenn sich x_{i1} um eine Einheit ändert, alle anderen unabhängigen Variablen, also x_{i2} und x_{i3}, konstant gehalten werden.

Wird also eine Einflussreiche x-Variable fälschlicherweise *nicht* in das Regressionsmodell aufgenommen, so kann auch ihr linearer Effekt nicht aus den übrigen x_j Variablen herausgerechnet werden. Er wird sich also in den Residuen wiederfinden.

b-Koeffizienten

Bei den b-Koeffizienten handelt es sich um die unstandardisierten Regressionskoeffizienten. Sie geben an, um wieviel -nämlich um b_j- sich y verändert, wenn sich die zugehörige x_j Variable um eine Einheit verändert. Die b_j Koeffizienten können hinsichtlich ihrer Bedeutung *nicht* miteinander verglichen werden. Es ist offensichtlich, dass die Veränderung des Einkommens um einen (oder 100) Euro eine andere Bedeutung besitzt als eine Veränderung des Alters um ein Jahr (oder 100).

β-Koeffizienten

Bei den β_j-Koeffizienten handelt es sich um die standardisierten Regressionskoeffizienten. Hier ist der Hinweis angebracht, dass es sich, trotz der gleichen Bezeichnung, *nicht* um die Regressionskoeffizienten der Population handelt. Es gibt mehrere Möglichkeiten sie zu berechnen: Entweder indem man die Ursprungsvariablen x_j und y z-Standardisiert, also:

$$z_x = \frac{x_{ij} - \bar{x_j}}{s_{x_j}} \quad \text{sowie} \quad z_y = \frac{y_i - \bar{y}}{s_y}$$

oder über:

$$\beta_j = b_j \frac{s_{x_j}}{s_y}$$

Hier lautet die Interpretation folgendermaßen: y verändert sich um β Standardabweichungen, wenn sich x um eine Standardabweichung verändert. Wir haben es hier also nicht mehr mit den ursprünglichen Einheiten der Variablen - beispielsweise Euro, Alter in Jahren oder ähnlichem- zu tun. Die β_j-Koeffizienten können untereinander hinsichtlich ihrer Bedeutung nach verglichen werden, da sie sich auf Standardbweichungseinheiten beziehen.

1.5 Statistische Tests bei multipler Regression

1.5.1 Der Test der multiplen Regressionskoeffizienten

Die Berechnung von Konfidenzintervallen für die partiellen Regressionskoeffizienten erfolgt, wie im bivariaten Fall

$$\beta_j = b_j \pm s_{b_j} \cdot z_{\frac{\alpha}{2}}$$

oder

$$z = \frac{b_j - \beta_j}{s_{b_j}}$$

mit

$$s_{b_j} = \frac{s_u^2}{\sum(x_i - \hat{x}_j)} = \frac{s_u^2}{(n-1)s_j^2(1-r_j^2)}$$

wobei s_u^2 die durch die j Variablen ($j = 2, ..., J$) nicht erklärte Varianz von x_1 (abhängige / Kriteriumsvariable) ist.
$\sum(x_i - \hat{x}_j)$ ist die durch die Variablen x_1, x_2, ..., x_j nicht erklärte Varianz von x_j (lineare Unabhängigkeit von x_j von anderen unabhängigen Variablen)
$(1 - r_j^2)$ wird als "Toleranz" bezeichnet und ist der Anteil der nicht erklärten Varianz einer unabhängigen Variablen, gegeben die anderen unabhängigen Variablen. Die Toleranz gibt Auskunft über das Ausmaß der Unabhängigkeit oder Abhängigkeit der unabhängigen Variablen. Bei 0 besteht Multikollinearität.

1.5.2 Test der Korrelationskoeffizienten

Es ist zu testen, ob sich die multiplen Korrelationskoeffizienten signifikant von 0 unterscheiden. Dies kann entweder durch einen Gesamttest oder durch partielle Tests erfolgen.

Der Gesamttest

Wenn nur einer der partiellen Korrelationskoeffizienten von 0 abweicht, ist H_0 abzulehnen.

$$H_0 : \rho_{12-3\ldots k} = \rho_{13-2,4\ldots k} = \ldots = \rho_{1k-2,3\ldots k-1} = 0 \Leftrightarrow \rho_k^2 = 0$$

Der multiple Determinationskoeffizient r_k^2 wird mittels F-Test der Form

$$F = \frac{n-k}{k-1} \cdot \frac{r_k^2}{1-r_k^2}$$

geprüft. Es folgen durch diesen Test fast immer hohe Werte für F und damit signifikante Modelle. Dieser Test gibt jedoch keine Auskunft darüber, welche Variablen einen Beitrag leisten.

Der partielle Test

Bei diesem Test ist es möglich, den Beitrag der zuletzt aufgenommenen Variablen (1.) zum Modell auf Signifikanz zuprüfen (Die Variable x_k, aus der alle anderen unabhängigen Variablen $X2$, $\ldots X_{k-1}$ herauspartialisiert sind). Man kann jedoch auch den Beitrag einer beliebigen Partialvariablen (2.) zum Modell testen.

1. Wir testen den Beitrag der letzten Variablen x_k zum Modell auf Signifikanz, d.h. wir testen $H_0 : \rho_{1k-2,3,\ldots,k-1} = 0$ mittels F-Test.

   $$F = \frac{\frac{1}{2-1} \cdot r^2_{1(k-2,3,\ldots,k-1)}}{\frac{1}{n-k} \cdot (1-r_k^2)}$$

   indem wir die durch x_k erklärte Varianz relativ zur nicht erklärten Gesamtvarianz setzen.

2. Getestet wird die H_0, dass eine beliebige Variable X_j keinen signifikanten Beitrag leistet mit

   $$F = \frac{r^2_{1(j-2,3,\ldots,j-1,j+1,\ldots,k)}}{\frac{1}{n-k} \cdot (1-r_k^2)}$$

1.5.3 Test des multiplen Regressionsmodells

Blockweise Regression

Es erfolgt eine simultane Schätzung der Koeffizienten, alle Variablen werden simultan in das Modell aufgenommen. Die Einzelbeiträge der Variablen zum Regressionsmodell werden nicht getestet. Lediglich die Erklärungskraft des Gesamtmodells wird geprüft.

Schrittweise Regression

Die Variablen werden nacheinander in Abhängigkeit ihrer einfachen Korrelationskoeffizienten in die Analyse einbezogen. Es bestehen 2 Möglichkeiten:

Vorwärts Die Variable mit der höchsten Korrelation geht als erste Variable in das Modell ein. Für die anderen Variablen werden nun partielle Korrelationen berechnet und die Variable mit dem nächsthöchsten Erklärungsbeitrag geht in das Modell ein. Das Ziel ist, aus einer gegebenen Menge von x_j Variablen diejenige Menge an x_j herauszufinden, deren linearkombination bei geringster Anzahl von x_j die beste Schätzung liefert. Ist der r^2-Zuwachs bei Hinzunahme nicht mehr signifikant, wird abgebrochen. Als Kritikpunkt ist zu erwähnen, dass das Modell über die Relevanz der Variablen entscheidet, nicht der Forscher mit theoretischen Argumenten. Es ist möglich, dass Suppressorvariablen den Effekt einer wichtigen x-Variablen verdecken, die somit nicht ins Modell mit aufgenommen wird. Dies kann zu dem Vorwurf führen, dass man theoretische Probleme in formal-statistische aufgelöst hat, und es sich somit bei den Ergebnissen um künstlich verursachte Fehlschätzungen handelt. In der Literatur wird ihre allzuhäufig sinnlose Anwendung angemerkt.

Rückwärts Hier werden die Variablen nicht vom höchsten r^2 an einbezogen. Wie bei der blockweisen Regression werden zuerst alle x_j-Variablen aufgenommen, danach wird diejenige x_j-Variable mit dem niedrigsten, nicht signifikanten, Beitrag zu r^2 aus dem Modell entfernt. Dies wird so lange wiederholt, bis nur noch Variablen mit einem signifikanten Beitrag zu r^2 vorhanden sind.

Es bleibt anzumerken, dass die schrittweisen Regressionen "vorwärts" und "rückwärts" nicht notwendigerweise zu identischen Ergebnissen führen müssen.

Hierarchische Regression

Die Einbeziehung der Variablen ist theoretisch untermauert. Es gibt zwischen ihnen hierarchische Beziehungen, die in dem Modell modelliert werden. Es wird folglich festgelegt, in welcher Reihenfolge die Variablen in das Modell eingehen. Als Analyse in mehreren Stufen ist die hierarchische (oder auch sequentielle oder kumulative) Regression der schrittweisen Regression recht ähnlich. Der gewichtige Unterschied ist, dass es hier der Forscher ist, der die Reihenfolge der Variablenaufnahme bestimmt, nicht das statistische Modell über r^2. Der Vorteil gegenüber der klassischen, simultanen Analyse ist der, dass wir die Abhängigkeit der Schätzung einzelner Variableneffekte von anderen, im Modell enthaltenen Variableneffekten kontrollieren können.

1.6 Spezielle Modelle und Erweiterungen

1.6.1 Interaktionseffekte

Möchte man prüfen, ob 2 Variablen nicht nur einen separaten, sondern auch einen gemeinsamen Effekt haben, testen wir dies über Interaktionseffekte: Für die Variablen x_1 und x_2 und das Modell

$$y_i = b_0 + x_1 b_1 + x_2 b_2$$

bilden wir die neue Variable $(x_1 \cdot x_2)$ und nehmen sie in das Modell auf.

$$y_i = b_0 + x_1 b_1 + x_2 b_2 + (x_1 \cdot x_2) b_3$$

Setzen wir $(x_1 \cdot x_2) = x_3^*$ resultiert

$$y_i = b_0 + x_1 b_1 + x_2 b_2 + x_3^* b_3$$

Betrachten wir beispielsweise die beiden Variablen Familiengröße x_1 und Einkommen x_2, und wollen wissen ob sie gemeinsam einen Effekt auf das Sparverhalten y haben, so generieren wir Familiengröße $\times$ Einkommen $(x_1 \cdot x_2)$ und können nun so testen, ob sie gemeinsam wirken, also ob beispielsweise große Familien mit geringem Einkommen stärker sparen, als sich durch beide Variablen einzeln erklären lässt.
Wichtig an diesem Modell ist, dass nicht nur die Interaktionsvariable in das Modell einfliesst, sondern auch die beiden "Haupteffekte". Nur so wird statistisch kontrolliert (partielle Regressionskoeffizienten), ob der Interaktionseffekt unabhängig von den Einzeleffekten seiner Komponenten einen eigenständigen Einfluß auf y ausübt. Ein auftretendes Problem ist die resultierende Multikollinearität, da x_3^* hoch mit x_1 und x_2 korrelieren wird, da es ja aus ihnen gebildet wurde. Um dieses Problem zu beheben werden x_1 und x_2 vor Bildung des Interaktionsterms Mittelwertzentriert. Dies geschieht wie folgt:

$$x_i^{\bar{x}} = x_i - \bar{x}$$

Die Werte der neu gebildeten Variable sind wie folgt zu interpretieren:

- $-m = m$ Einheiten unterhalb des Mittelwertes
- 0 = entspricht genau dem Mittelwert
- $+m = m$ Einheiten oberhalb des Mittelwert

Haben wir also für eine Person auf der zentrierten Variablen Einkommen einen Wert von 145, so ist das Einkommen um 145 Euro höher, als das durchschnittliche Einkommen. Wichtig ist hier der Unterschied der Mittelwertzentrierung zur Z-Transformation:

- Die Mittelwertzentrierung ist immer noch in der Ursprungsskala (2 = 2 Euro über dem Durschschnitt) gemessen, nur der Mittelwert ist nun = 0
- Bei der Z-Transformation ist der neue Mittelwert = 0, sowie die Skala in Standardabweichungen transformiert (2=2 Standardabweichungen über dem Durchschnitt).

Achtung!

Die Berechnung der standardisierten Regressionskoeffizienten β_j ist über die Form $\beta_j = b_j \frac{s_{x_j}}{s_y}$ in Gegenwart von Interaktionseffekten nicht zulässig (so beispielsweise in Stata), sie sind nicht interpretierbar. Zur Ermittlung der β_j-Koeffizienten müssen die an der Bildung der Interaktionsterme beteiligten Variablen x_j im Vorfeld z-transformiert werden.

1.6.2 Dummy-Regression

Es besteht auch die Möglichkeit, Dummy-Variablen in die Regression einzubinden. Wir können Dummys jedoch nur als unabhängige Variablen nutzen, eine Dummy-Variable als abhängige Variable ist in der linearen Regression nicht ratsam. Hierzu benötigt man logistische Ansätze, die später diskutiert werden sollen.
Wie also funktioniert nun eine lineare Regression mit einer unabhängigen Dummy-Variablen? Eine Besonderheit ist, dass sich die Interpretation des Regressionskoeffizienten verändert. Hier macht die generelle Aussage "wenn sich x um eine Einheit ändert, dann ändert sich y um b" nur bedingt Sinn, da sich die x-Variable nämlich (da sie ein Dummy ist) nur *ein einziges Mal* um eine Einheit ändern kann, nämlich von 0 zu 1.

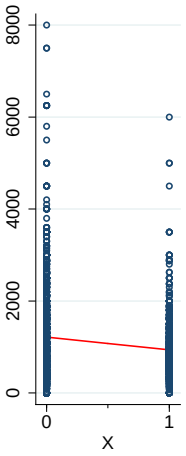


Abbildung 1.7: Dummy als abhängige Variable

Die Interpretation der Parameter gestaltet sich wie folgt:

- b_0, also die Konstanten bezeichnet den Mittelwert der y-Variable für die mit 0 codierten Fälle.
- $b_0 + b_1$ gibt den Mittelwert der mit 1 codierten Fälle an.

- b_1 steht für den Effekt, den das besitzen, bzw. das nichtbesitzen des Merkmals hat. Ist dieser Effekt, also b_1 signifikant, so besteht ein signifikanter Unterschied zwischen den beiden Gruppenmittelwerten.

Achtung!

Die Verwendung standardisierter Regressionskoeffizienten bei dichotomen x_j ist nicht zulässig. Da es sich bei der Standardabweichung eines Dummys um eine Funktion seiner Schiefe handelt, werden die β_j-Koeffizienten umso kleiner, je Schiefer der Dummy ist.

1.7 Voraussetzungen

1. Die wahre Beziehung zwischen den erklärenden Variablen x und der zu erklärenden Variable y (d.h. die "Population Regression Function") ist linear in den Parametern. Wenn wir k unabhängige Variablen haben:

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

   Die Parameter der Grundgesamtheit $\beta_0, \beta_1, \beta_2, \ldots, \beta_K$ sind für alle N Beobachtungen konstant.

2. Das Regressionsmodell ist korrekt spezifiziert, d.h. es fehlen keine relevanten unabhängigen Variablen, und die verwendeten unabhängigen Variablen sind nicht irrelevant.

3. Die Störterme ε_i der Grundgesamtheit haben einen Erwartungswert $= 0$:

   $$E(\varepsilon_i) = 0$$

4. Homoskedastizität: alle ε_i haben die gleiche konstante Varianz σ^2

   $$\mathrm{Var}(\varepsilon_i) = \sigma^2$$

   Wenn die Residuen diese Annahme verletzen spricht man von Heteroskedastizität.

5. Die Störterme ε_i der Grundgesamtheit sind nicht autokorreliert,ε darf nicht hoch oder niedrig sein, weil sein vorhergehender Wert dies war. D.h. für jedes Paar x_i und x_j, $(i \neq j)$ ist die Korrelation zwischen den Störtermen ε_i und ε_j gleich Null.:

   $$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0; \; i \neq j$$

6. Keine perfekte Multikollinearität (d.h. die x-Variablen sind linear unabhängig).

7. Die Anzahl der Beobachtungen n ist größer als die Anzahl der zuschätzenden Parameter k.

8. Keine Korrelation zwischen den Störgrößen und den erklärenden Variablen:

   $$\mathrm{cov}(\varepsilon_i, x_i)$$

   ε_i darf nicht klein sein, nur weil x_i klein ist

9. Die Störgrößen sind normalverteilt:

   $$\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

Für die prinzipiell unbeobachtbaren Störgrößen ε werden die Residuen $e_i = y_i - \hat{y}_i$ herangezogen.

1.7.1 Das Anscombe-Quartett

Das Anscombe-Quartett ist ein klassisches Beispiel für die Notwendigkeit der Regressionsdiagnostik. Wir sehen 4 Regressionen für 4 Datensätze. Das Besondere an diesen Regressionen ist, dass jeweils $\bar{x}$, $\bar{y}$, b_0, b_1, r^2 sowie σ_{b_1} identisch sind. Inspizieren wir nur diese Werte, so entgeht uns, ob es sich um einen

- Angemessen linearen Zusammenhang
- Quadratischen Zusammenhang
- Bis auf einen Ausreißer perfekt linearen Zusammenhang
- Ohne den Ausreißer gar keinen Zusammenhang

handelt.

Parameter im Anscombe-Quartett

b_0	=	3.0	r^2	=	0.67
b_1	=	0.5	$\overline{x}$	=	9.0
σ_{b_1}	=	0.118	$\overline{y}$	=	7.5

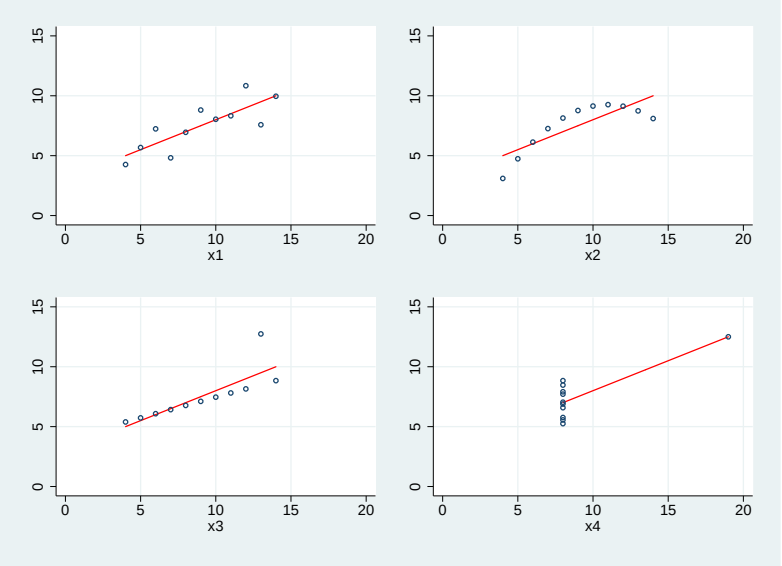


Abbildung 1.8: Anscombe-Quartett

In solchen bivariaten Fällen ist es noch möglich, sich über die Graphiken einen Überblick zu verschaffen. Dies ist in höherdimensionalen Regressionen leider nicht mehr möglich. An dieser Stelle setzt die Regressionsdiagnostik an.

1.7.2 Heteroskedastizität

Auswirkung

Die OLSMethode gewichtet die Beobachtungen mit großer Varianz stärker als jene mit kleinen Varianzen. Aufgrund dieser impliziten Gewichtung sind die OLSParameter zwar weiterhin erwartungstreu und konsistent, aber nicht mehr effizient. Außerdem sind die geschätzten Standardabweichungen der Parameter verzerrt. Deshalb sind die statistischen Tests und Konfidenzintervalle ungültig, selbst wenn die Störterme unabhängig und normalverteilt sind.

Tests

- GoldfeldQuandt Test
- BreuschPagan Test
- WhiteTest

Heteroskedastizität (auch (Residuen)-Varianzheterogenität) bedeutet in der Statistik unterschiedliche Streuung innerhalb einer Datenmessung. Die Streuung der Fehlerwerte variiert in Abhängigkeit der Ausprägungen der unabhängigen x-Variablen. Heteroskedastizität kann beispielsweise durch systematische Meßfehler der y-Variable, oder durch ein fehlerhaft spezifiziertes Regressionsmodell entstehen.

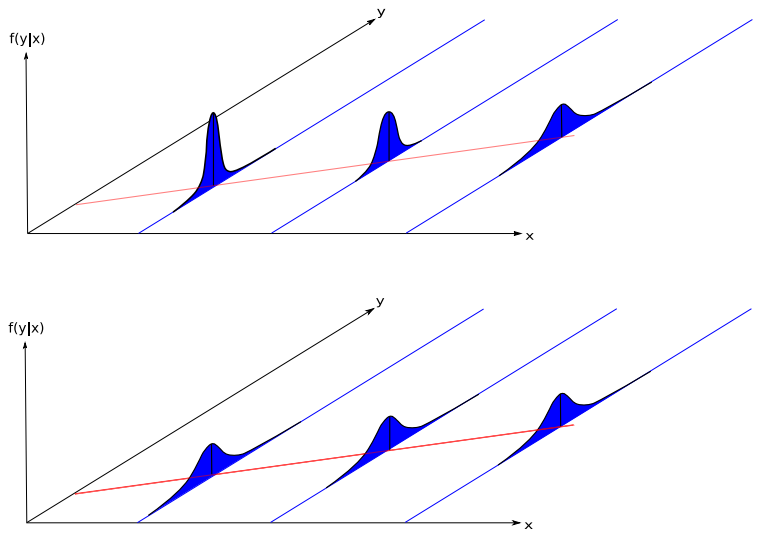


Abbildung 1.9: Hetero- und Homoskedastizität

Wenn man zum Beispiel die Urlaubsausgaben von Haushalten (y) in Abhängigkeit vom Einkommen (x) untersucht ist zu erwarten, daß die Varianz bei reicheren Haushalten größer ist als bei weniger wohlhabenden Haushalten. Wenn die

Varianz der Residuen (und somit die Varianz der erklärten Variablen selbst) für alle Ausprägungen der anderen (Prädiktor-) Variablen nicht signifikant unterschiedlich sind, liegt Homoskedastizität ((Residuen-) Varianzhomogenität) vor. Ist Homoskedastizität, also sind die Gleichheit der Residuumsvarianzen für unterschiedliche x-Werte nicht gegeben, so haben die Regressionskoeffizienteb verzerrte Varianzen. Die Varianzen von b entsprechen nicht mehr jenen von β. Dies verfälscht den Standardfehler der Regressionskoeffizienten. Allerdings bleiben die geschätzten b-Koeffizienten an sich unverzerrt. Die Konfidenzintervalle sind jedoch von der Verzerrung betroffen. Heteroskedastizität kann häufig schon in einem Streudiagramm (Scatterplott) erkannt werden.Zur Inspektion bietet sich eine visuelle Analyse der Residuen an, indem man die vorhergesagten $\hat{y}_i$-Werte gegen die Residuen plottet. Das Streudiagramm der geschätzten Werte der abhängigen Variablen (ZPRED) und der Residuen (ZRESID) darf kein Dreiecksmuster aufweisen.

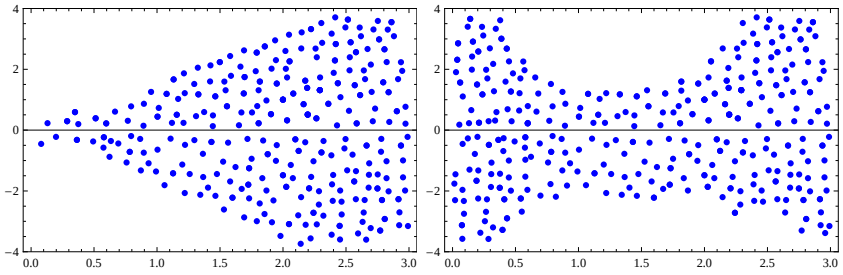


Abbildung 1.10: Heteroskedastische Residuen

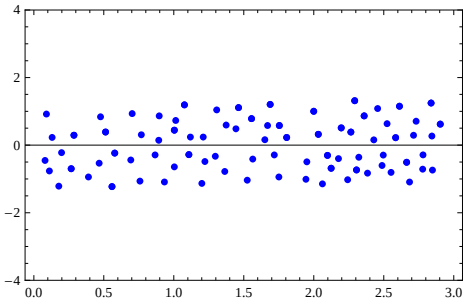


Abbildung 1.11: Homoskedastische Residuen

Ein statistischer Test zur Aufdeckung von Heteroskedastizität ist der Goldfeldt-Quandt Test. Hier wird die Stichprobe in zwei Subgruppen aufgespalten. Es werden nun die beiden Varianzen ins Verhältnis (Quotient) gesetzt. Gilt $s_1^2 = s_2^2$, so beträgt der Quotient 1. Je weiter sich der Wert von 1 entfernt, so stärker ist die Tendenz zur Heteroskedastizität. Sind die Residuen normalverteilt und trifft die Annahme der Homoskedastizität zu, so folgt das Verhältnis der Varianzen einer F-Verteilung. Heteroskedastizität ist auch eine Folge von Nichtlinearität und nichtlineare Transformationen können somit Homoskedastizität herstellen.

1.7.3 Multikollinearität

Auswirkung

Perfekte Multikollinearität führt dazu, dass die Regression nicht mehr berechenbar ist (Division durch 0). Auch wenn dies in der sozialwissenschaftlichen forschung seltenst vorkommt, so werden doch die Schätzungen der Regressionsparameter unzuverlässiger durch einen größeren Standardfehler.

Tests

- Toleranz
- VIF

In folgender Graphik sehen wir, dass bei steigender Multikollinearität ein immer größerer Teil der Daten redundant wird. Es lässt sich die redundante Information auch nicht mehr eindeutig einer Variable zuordnen. Der rote bereich kann nicht zur Bestimmung der Koeffizienten der Regressoren enutzt werden. Sie geht aber trotzdem in die Berechnung des standardfehlers ein, vermindert ihn und trägt somit zur Verbesserung der Prognose und Steigerung von r^2 bei. Es kann vorkommen, dass r^2 signifikant ist, obwohl kein Regressor dies ist.

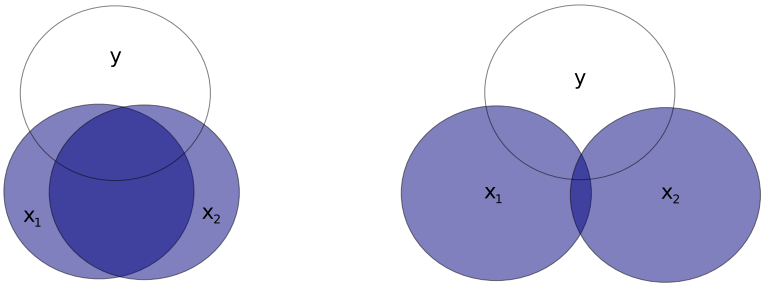


Abbildung 1.12: Hohe und geringe Multikollinearität

Es kann vorkommen, dass sich die Regressoren stark verändern, wenn eine Variable hinzugenommen wird, die die Multikollinearität stark erhöht, oder eine herausgenommen wird, die sie stark senkt. Ein Maß für Multikollinearität sind hohe Korrelationen zwischen den unabhängigen Variablen. Hier sind Werte nahe ± 1 ein Indiz. Allerdings ist diese Untersuchung nur bivariat.
Ein anderes Maß zur Inspektion ist die *Toleranz*. Als Toleranz bezeichne wir den Term $1 - r_j^2$ Hierbei bezeichnet r_j^2 die Regression der unabhängigen Variable x_j auf alle übrigen unabhängigen Variablen. Da ein hoher Wert von r^2 auf eine starke Erklärung der unabhängigen Variablen durch die anderen unabhängigen Variablen hinweist, können wir sagen, dass die Toleranz uns immer stärkere Probleme anzeigt, je niedriger ihr Wert wird. Ein darauf aufbauendes Maß ist der *Variance Inflation Factor*, auch VIF genannt. Er ist der simple Kehrwert

der Toleranz, also:

$$\frac{1}{1-r_j^2}$$

Toleranz und VIF liegen bei nicht vorhandener Multikollinearität bei 1. Bei einem VIF-Wert größer als 10 ist größte Vorsicht geboten. Der Name Variance Inflation Factor resultiert daraus, dass sich mit zunehmender Multikollinearität die Varianzen der Regressionskoeffizienten um diesen Betrag vergrößern. Die Schätzungen werden also mit zunehmender Multikollinearität immer schlecher.

1.7.4 Autokorrelation

Auswirkung
Autokorrelation führt zu Verzerrung bei der Ermittlung des Standardfehlers der Regressionskoeffizienten, demnach also auch zu Problemen bei der Bestimmung des Konfidenzintervalls
Tests
• Durbin-Watson-Test

Eine weitere Vorraussetzung für die Verwendung der Regression ist, dass die Residuen nicht miteinander korrelieren. Bei auftretender Autokorrelation sind die Abweichungen von der Regressionsgeraden nicht mehr zufällig, sondern von den Abweichungen der vorangehenden Werte abhängig. Die Werte beziehen sich also auf die ihnen vorhergehenden Werte. Diese Verletzung der Prämisse führt zu einem verzerrten Standardfehler des entsprechenden Regressionskoeffizienten und damit auch zu einem fehlerhaften Konfidenzintervall.

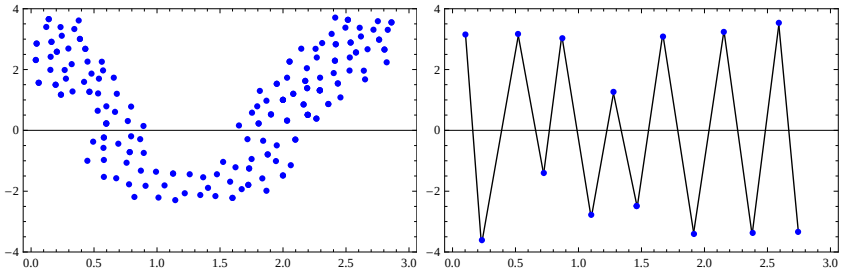
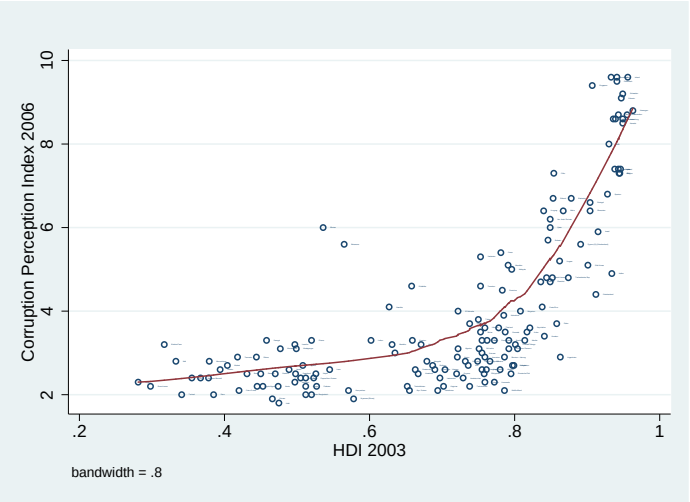


Abbildung 1.13: Autokorrelation, positiv und negativ

Ist ein Meßwert hoch, weil sein Vorgänger hoch ist, so sprechen wir von positiver Autokorrelation. Ist ein Meßwert hoch weil sein vorhergehender Wert niedrig ist, so sprechen wir von negativer Autokorrelation. Bei positiver Autokorrelation geht d gegen null, bei negativer Autokorrelation geht d gegen +4. Wir testen die Hypothese $H_0 : \rho = 0$, dass die Autokorrelation null ist. Fehlspezifikation, oder fehlende Variablen können zu Autokorrelation führen. Der Störterm repräsentiert den Einfluß aller nicht berücksichtigten erklärenden Variablen. Wir erwarten, daß der Einfluß dieser Variablen gering ist und daß sie sich in ihrer Wirkung im Durchschnitt gegenseitig aufheben. Wenn sich die "ausgelassenen" Variablen aber sehr ähnlich verhalten kann dies zu Autokorrelation führen.
Weitere Anhaltspunkte für eine eventuell vorhandene Autokorrelation liefert das Streuungsdiagramm der geschätzten abhängigen Variablen und der Residuen, das schon zur Beurteilung der Homoskedastizität herangezogen wird. Positive Autokorrelation ist erkennbar, wenn aufeinander folgende Residualwerte nahe beieinander stehen, negative daran, dass die Schwankungen sehr groß sind.
Autokorrelation tritt häufig bei aufeinanderfolgenden Beobachtungen in Zeitreihen (serielle Autokorrelation) auf, man trifft aber auch bei räumlich nahe beieinanderliegenden Erhebungseinheiten (spatial correlation) auf Autokorrelation.

1.7.5 Nichtlinearität

Das lineare Regressionsmodell fordert, dass die Beziehung zwischen den X und der Y-Variablen linear in den Parametern ist. Es ist daher ohne weiteres möglich eine Variable X durch Transformation in eine Variable $X_t = f(x)$ zu überführen, für die der Scatterplot eine liineare Beziehung ausweist.



$f(x)$ kann dabei jede beliebige nichtlineare Funktion bezeichnen, beispielsweise $f(x) = e^x$, $f(x) = \ln x$ oder $f(x) = x^2$.

1.7.6 Erwartungswert der Störgrößen ungleich Null

Wenn alle systematischen Einflußgrößen als unabhängige Variablen erfasst sind, dann erfasst die Störvariable nur zufällige Effekte, die Schwankungen gleichen sich im Mittel aus. Eine Verletzung dieser Annahme besteht dann, wenn die Stichprobenauswahl nict zufällig war, wichtige unabhängig Variablen vernachlässigt werden oder die Meßwerte von y systematisch zu hoch oder zu niedrig gemessen werden. Dann enthält die Störgröße nicht nur zufällige Abweichungen, sondern einen systematischen Effekt. Durch die OLS-Methode wird der Mittelwert auf Null "gezwungen", der systematische Fehler geht in die Berechnung des Intercepts b_0 ein. b_0 wird bei konstant zu groß gemessenen y ebenfalls verzerrt zu hoch sein.

1.7.7 Residuen nicht Normalverteilt

Die Überprüfung dieser Annahme steht am Schluss der Residualanalsye, da eine Verletzung dieser Annahme oftmals durch Verletzungen der anderen Annahmen verursacht wird. Sie hebt sich oftmals auf, wenn die anderen Verletzungen behoben werden. Diese Annahme muss nicht eingehalten werden, damit die Regressionsparameter nach der OLS-Methode als BLUE angesehen werden können. Sie bezieht sich viel mehr -und nur- auf die durchführbarkeit statistischer Signifikanzests, wie den F-Test oder den T-Test. Hierbei wird unterstellt, dass die Regressionsparameter b_0 und b_j normalverteilt sind. Hier bieten sich

der Kolmogoroff-Smirnoff Test, der Skewness-Kurtosis Test oder der Shapiro-Wilk W Test an. An Graphiken lassen sich ein Histogramm mit eingezeichneter Normalverteilung, Kerndichteschätzungen oder P-P-Plots anwenden.

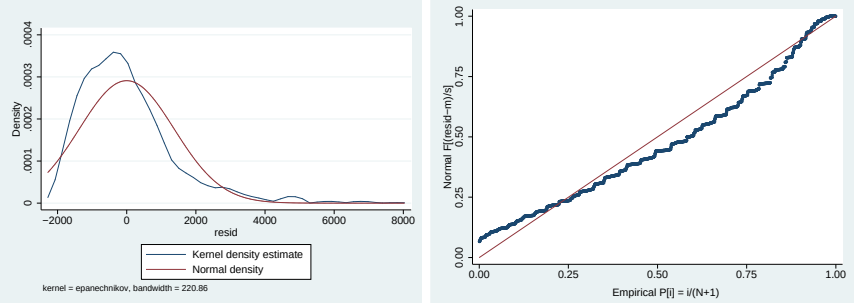


Abbildung 1.14: Überprüfung auf Normalverteilung

KAPITEL 2

VARIANZANALYSE

In der Varianzanalyse wird geprüft, ob eine nominalskalierte unabhängige Variable x oder mehrere nominalskalierte unabhängige Variablen x_j einen Einfluss auf eine metrische skalierte abhängige Variable y ausüben. Die Ausprägungen dieser x_j-Variablen definieren die verschiedenen Gruppen, für die getestet wird, ob sich die Mittelwerte dieser Gruppen signifikant voneinander Unterscheiden.
Die nominalskalierten unabhängigen Variablen werden als *Faktoren* bezeichnet, aber Achtung: Diese Faktoren haben mit den Faktoren der Faktorenanalyse nichts gemeinsam. Die Ausprägungen der unabhängigen Variablen werden als *Faktorstufen* bezeichnet. Die Varianzanalyse hat ihren Ursprung in der Landwirtschaft, wo geprüft wurde, ob Felder, die mit verschiedenen Düngern behandelt wurden auch unterschiedliche Erträge erzielten. Heute ist die Varianzanalyse ein gängiges Verfahren, insbesondere zur Auswertung von Experimenten in der Psychologie, aber auch der Medizin, der Biologie und natürlich auch der Sozialwissenschaft. Im Zusammenhang mit der Varianzanalyse wird oft auch der Begriff *ANOVA* verwendet, der **AN**alysis **O**f **VA**riance bedeutet.
Je nach Anzahl der Faktoren sprechen wir von einer

- Einfaktoriellen *ANOVA* (eine UV)
- Zweifaktoriellen *ANOVA* (zwei UVen)
- Dreifaktoriellen *ANOVA* (drei UVen)
- ...

2.1 Einfaktorielle *ANOVA*

Es wird eine Stichprobe vom Umfang n aus einer Grundgesamtheit gezogen, die sich auf Grund der j Stufen des ersten Faktors in ebensoviele, nämlich j Gruppen einteilen lassen. Bei den Untersuchungseinheiten wird jeweils der Wert der abhängigen Variablen y ermittelt.
Als Beispiel für eine *einfaktorielle ANOVA* könnten wir ein Experiment mit dem dreistufigen Faktor "Medikamentendosierungen" ansehen:

Faktor 1		
Placebo	Dosierung $\times$ 1	Dosierung $\times$ 2
y_{11}	y_{21}	y_{31}
$\dots$	$\dots$	$\dots$
y_{1n_1}	y_{2n_2}	y_{3n_3}
$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$

Wir gehen von folgendem Modell aus: Die Werte der abhängigen Variable y ergeben sich systematisch als Summe des Gesamtmittelwerts der Grundgesamtheit μ und dem Effekt des Faktors α_j. Alle anderen nicht beachteten Einflußgrößen sind in der Störgröße ε_{ij} enthalten, die als normalverteil angenommen wird.

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

Die Mittelwerte der j Stufen des ersten Faktors, μ_j, sind in der Grundgesamtheit durch

$$\mu_j = \mu + \alpha_j$$

gegeben. Die Größe α_j gibt somit den Effekt des Faktors in Form einer Veränderung des Gesamtmittels μ wieder.
Die Nullhypothese, die in der einfaktoriellen Varianzanalyse geprüft wird lautet

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_j$$

Wobei j die Anzahl der Faktorstufen bezeichnet. Es wird also unterstellt, dass der Mittelwert über alle Faktorstufen gleich ist. Ist mindestens ein $\mu_i \neq \mu_j$ mit $i \neq j$, so ist die H_0 zu verwerfen.

2.1.1 Voraussetzungen

Um eine Varianzanalyse berechnen zu dürfen, müssen einige Voraussetzungen erfüllt sein.

Normalverteilungsannahme Die Residuen der Gruppen müssen in der Population normalverteilt sein. Vielfach wird die Normalverteilung der Residuen fälschlicherweise mit der Normalverteilung der Ausgangswerte gleichgesetzt. Liegen beispielsweise bei zwei Gruppen normalverteilte Residuen und ein signifikanter Mittelwertsunterschied zwischen den beiden voneinander unabhängigen Gruppen vor, so können die die zusammengefassten Daten der beiden Gruppen nicht wiederum normalverteilt sein. Die mengentheoretische Zusammenfassung zweier Normalverteilungen darf nicht mit der Linearkombination zweier Normalverteilungen verwechselt werden. Die Forderung nach Normalverteilung bezieht sich nicht auf die y-Werte, sondern auf die Residuen, deren Inspektion diese Forderung auch prüft.

Homogenitätsannahme Die Varianzen der Residuen innerhalb der Gruppen des Designs müssen homogen sein (Homoskedastizität).

Unabhängigkeitsannahme Die Residuen innerhalb der Gruppen müssen unabhängig voneinander sein.

Sind die Vorraussetzungen nicht erfüllt, so bietet sich der Kruskal-Wallis H-Test an. Hierbei handelt es sich um eine Erweiterung des Mann-Whitney U-Test für mehr als drei Gruppen. Der Test ist nicht-parametrisch, und testet auf Gleichheit der Populationsmediane innerhalb der Gruppen, wobei die Daten durch ihre Ränge ersetzt werden. Für vergleiche der Gruppenmediane untereinander wird der Dunn-Test verwendet.

2.1.2 Varianzzerlegung

Die Gesamtquadratsumme der Messwerte kann zerlegt werden in die Treatmentsumme (SS_b /erklärter Anteil), also den Anteil an Unterschiedlichkeit der Beobachtungen, der auf die verschiedenen Faktorstufen zurückzuführen ist, sowie die Fehlersumme (SS_w /nicht erklärter Anteil), die so nicht erklärt werden kann. In Abbildung 2.1 sehen wir, was dies bedeutet. Die Grundidee ist derjenigen der Regression äusserst ähnlich. Es werden die Abweichungen vom Gesamtmittelwert $\hat{y}_G$ gemessen. Dieser Abstand lässt sich in zwei Teile zerlegen:

1. Den erklärbaren Abstand vom Gesamtmittelwert $\hat{y}_G$ zu den durch die j Faktorstufen gebildeten Mittelwerten $\hat{y}_j$. Hier sehen wir 4 Gruppen, die durch die Faktorstufen des Faktors gebildet werden. Dies erlaubt uns, den Effekt des *treatments* zu messen. Diese Streuung bezeichnet die Streuung *zwischen den Gruppen*.

2. Den Abstand der Messwerte y_{ji} -hier exemplarisch nur eine Beobachtung pro Gruppe- von ihren Gruppenmittelwerten $\hat{y}_j$. Diese Streuung geht *nicht* auf den Effekt des Faktors zurück. Je geringer die Relevanz des Faktors, desto stärker streuuen die Werte um ihre Gruppenmittelwerte $\hat{y}_j$. Man nennt dies auch die Streuung *innerhalb der Gruppen*.

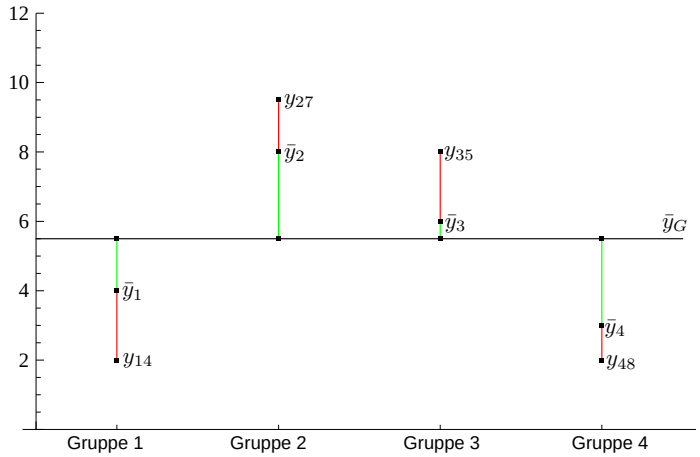


Abbildung 2.1: Group Means

Gesamte Abweichung	=	Erklärte Abweichung	+	Nicht erklärte Abweichung
Summe der quadrierten Gesamtabweichung	=	Summe der quadrierten Abweichungen innerhalb der Faktorstufen	+	Summe der quadrierten Abweichungen zwischen den Faktorstufen
$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ji}-\bar{y})^2$	=	$\sum_{j=1}^{J} n_j(\bar{y}_j-\bar{y})^2$	+	$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ji}-\bar{y}_j)^2$
SS_t	=	SS_b	+	SS_w
$SS_{t(otal)}$	=	$SS_{b(etween)}$	+	$SS_{w(ithin)}$

In Abbildung 2.2 sehen wir, wie man sich die SS_w vorstellen kann: Links ist die Streuung *innerhalb der Gruppen* gering, die Gesamte Streuung geht also größtenteils auf die Streuung *zwischen den Gruppen* zurück. die rechte Abbildung zeigt eine große Streuung *innerhalb der Gruppen*. Die gesamte streuung wird hier viel weniger durch die Streuung *zwischen den Gruppen* gebildet, die Gruppen sind weit weniger eindeutig als in der linken Abbildung. An dieser Stelle wird hoffentlich der Begriff *Varianzanalyse* klarer, da die Entscheidung, ob es sich hier um signifikant Unterschiedliche Mittelwerte handelt nicht allein von der Lage der Mittelwerte ($\bar{x}_1 = 2$, $\bar{x}_2 = 5$ und $\bar{x}_3 = 8$) ausgeht. Die Varianzen der einzelnen Gruppen in Relation zur Gesamten Varianz steht hier im Mittelpunkt. Links sind die Varianzen innerhalb der gruppen klein, die Gruppen sind gut getrennt, wohingegen die Gruppen rechts grosse Überlappungsbereiche aufweisen, sie sind somit schlechter getrennt, und würden als nicht signifikant unterschiedlich angesehen. Der Effekt des Faktors wäre also nicht signifikant.

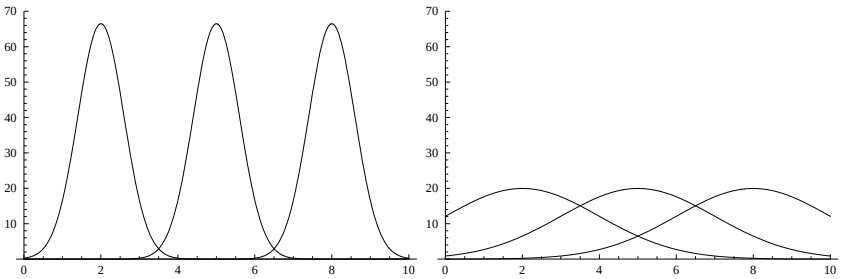


Abbildung 2.2: SS_w klein vs. SS_w gross

Varianzaufklärung

Ähnlich der Regressionsanalyse kann auch bei der Varianzanalyse eine Varianzaufklärung durchgeführt werden. Der Quotient η^2 ist in seiner Berechnung r^2 sehr ähnlich:

$$\eta^2 = \frac{QS_{\text{Treat}}}{QS_{\text{Total}}} \cdot 100$$

stellt jedoch eher ein deskriptives Maß dar, da er die Varianzaufklärung der Population überschätzt.

Signifikanz

Für die Population bleibt zu prüfen, ob die "deskriptiv" erklärte Varianz ein Zufallseffekt der Stichprobe ist, oder ob der Faktor auch in der Grundgesamtheit signifikante Mittelwertunterschiede verursacht. Wir testen $H_0 : \mu_1 = \mu_2 = \ldots = \mu_j$, das ist äquivalent zu $H_0 : \sigma^2_{\text{Treat}} = \sigma^2_{\text{Fehler}}$. Wenn wir H_0 ablehnen möchten, muss $\sigma^2_{\text{Treat}} > \sigma^2_{\text{Fehler}}$ sein, wir überprüfen das mit dem F-Test

$$F = \frac{\sigma^2_{\text{Treat}}}{\sigma^2_{\text{Fehler}}}$$

2.1.3 Ungleiche Stichprobengrössen

Bisher sind wir von gleichgroßen Stichproben je Faktorstufe ausgegangen. Das ist bei geplanten Experimenten sicherlich sinnvoll, entspricht aber nicht der Realität einer Varianzanalyse bei gegebenem Datensatz. Die unterschiedlichen Stichprobengrößen haben Änderungen in den Berechnungen der Quadratsummen zur Folge. Während wir vorher ein einheitliches n hatten gibt es nun mehrere n_i für jede Faktorstufe (Stichprobe) j. Die Gesamtzahl der Untersuchungseinheiten N ist nun nicht mehr

$$N = p \cdot n \text{ sondern } N = \sum_{j=1}^{J} n_j$$

Dies hat eine Änderung in der Berechnung der Quadratsummen zur Folge. Die Berechnung

$$QS_{\text{Total}} = QS_{\text{Treatment}} + QS_{\text{Fehler}}$$

ergibt sich für $n_i = n_j$, $i \neq j$

$$\sum_{j=1}^{J}\sum_{i=1}^{n}(y_{ji} - \bar{y})^2 = \sum_{j=1}^{J}(\bar{y}_j - \bar{y})^2 + \sum_{j=1}^{J}\sum_{i=1}^{n}(y_{ji} - \bar{y}_j)^2$$

für $n_i \neq n_j$, $i \neq j$

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ji} - \bar{y})^2 = \sum_{j=1}^{J} n_j(\bar{y}_j - \bar{y})^2 + \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ji} - \bar{y}_j)^2$$

Die Freiheitsgrade der Gesamtvarianz ergeben sich jeweils durch

$$df_{\text{Total}} = df_{\text{Treatment}} + df_{\text{Fehler}}$$

mit

Freiheitsgrade, gleiche Stichprobengrößen			Freiheitsgrade, ungleiche Stichprobengrößen		
df_{Total}	=	$p \cdot n - 1$	df_{Total}	=	$N - 1$
df_{Fehler}	=	$p(n-1)$	df_{Fehler}	=	$p - 1$
df_{Treat}	=	$p - 1$	df_{Treat}	=	$N - p$

2.1.4 Einzelvergleiche

Wenn wir einen signifikanten F-Wert haben können wir schließen, dass sich die Mittelwerte unterscheiden. Wir können allerdings nicht wissen, ob sich alle oder einige unterscheiden oder sogar vielleicht nur ein einziger Mittelwert die Gesamtsignifikanz verursachte. Erst durch Einzelvergleiche (Kontraste) finden wir heraus, welche Mittelwerte sich signifikant voneinander unterscheiden. Wir gehen vom Fall eines Faktors mit 2 Faktorstufen aus, dieser lässt bei Signifikanz eine eindeutige Aussage über die Unterschiedlichkeit der beiden Mittelwerte zu. Man kann diesen Ansatz aber auf Faktoren mit mehr Faktorstufen übertragen. Wir vergleichen immer paarweise 2 Mittelwerte. Dies erreichen wir durch Gewichtungen. Für die Gewichtungskoeffizienten c_j gilt:

$$\sum_{j=1}^{J} c_j = 0$$

Prüfung des Einzelvergleichs mittels F-Verteilung

$$F = \frac{D^2}{var(D)}$$

mit

$$var(D) = \frac{\left(\sum_{j=1}^{J} c_j^2\right) \cdot \hat{\sigma}^2_{\text{Fehler}}}{n} \quad \text{sowie} \quad D^2 = \sum_{j=1}^{J} (c_j \cdot \bar{A}_j)^2$$

Beispiel Drei Behandlungsmethoden und eine Kontrollbedingung werden gestestet (p=4, n=20) in Bezug auf den Behandlungserfolg. Folgende Mittelwerte finden sich für die Gruppen:

$$\bar{A}_1 = 16,\ \bar{A}_2 = 14,\ \bar{A}_3 = 18,\ \bar{A}_4 = 15$$

Die nicht durch die Behandlungen erklärte Varianz beträgt

$$\hat{\sigma}^2_{\text{Fehler}} = 5$$

Wir testen, ob sich die Behandlungsgruppen signifikant von der Kontrollgruppe unterscheiden (α=0,05).

$$D = \left(\frac{1}{3} \cdot 16 + \frac{1}{3} \cdot 14 + \frac{1}{3} \cdot 18\right) - 15$$

Dann ist

$$F = \frac{20 \cdot 1^2}{\frac{1}{3^2} + \frac{1}{3^2} + \frac{1}{3^2} + (-1)^2} = 3$$

Mit $F_{\text{theoretisch}} = 3.98$ und $df_z = 1,\ df_n = p \cdot (n-1) = 76$ (Fehler). Damit ist der Unterschied nicht signifikant.

Orthogonale Einzelvergleiche

Man kann verschieden Mittelwerte bei einem mehrstufigen Faktor vergleichen. So ist also nicht nur der Vergleich von Behandlungsmethodeneffekt gegenüber der Kontrollgruppe interessant, sondern vielleicht diverse andere Vergleiche. Möchte man diverse voneinander unabhängige Vergleiche durchführen, also solche ohne redundante Testung, nennt man dies orthogonale Einzelvergleiche.

Beispiel Wir führen bei p=4 Faktorstufen 6 Tests durch:

$$\begin{aligned} D_1 &= \frac{\bar{A}_1+\bar{A}_2+\bar{A}_3}{3-\bar{A}_4} \\ D_2 &= \bar{A}_1 - \bar{A}_4 \\ D_3 &= \bar{A}_1 - \bar{A}_2 \\ D_4 &= \bar{A}_3 - \bar{A}_4 \\ D_5 &= \frac{\bar{A}_1+\bar{A}_2}{2} - \frac{\bar{A}_3+\bar{A}_4}{2} \\ D_6 &= \frac{\bar{A}_1+\bar{A}_3}{2} - \frac{\bar{A}_2+\bar{A}_4}{2} \end{aligned}$$

Sind nun redundante Informationen in den Einzeltests enthalten oder haben wir hier einen vollständigen Satz orthogonaler Einzelvergleiche?

2. Feststellung Orthogonalitätsbedingung für Kontraste

$$c_{1k} \cdot c_{1l} + c_{2k} \cdot c_{2l} + \ldots + c_{jk} \cdot c_{jl} = \sum_{j=1}^{J} c_{jk} \cdot c_{jl} = 0$$

Vergleich D_1/D_2 :	$\frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 0 + (-1) \cdot (-1) = \frac{4}{3}$	$\rightarrow$ nicht erfüllt
Vergleich D_3/D_4 :	$1 \cdot 0 + (-1) \cdot 0 + 0 \cdot 1 + 0 \cdot (-1) = 0$	$\rightarrow$ erfüllt
Vergleich D_5/D_6 :	$\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{-1}{2} + \frac{-1}{2} \cdot \frac{1}{2} + \frac{-1}{2} \cdot \frac{-1}{2} = 0$	$\rightarrow$ erfüllt
Vergleich D_2/D_5 :	$1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} + 0 \cdot \frac{-1}{2} + (-1) \cdot \frac{-1}{2} = 1$	$\rightarrow$ nicht erfüllt

Es ist nun zu prüfen, ob auch D_2/D_3 wechselseitig orthogonal sind, oder D_1/D_3, um einen kompletten Satz wechselseitig othogonaler Einzelvergleiche zu finden. Ein kompletter Satz besteht aus $p-1$ Einzelvergleichen.Wenn wir die restlichen Tests durchführen, stellen wir fest, dass $D_3/D_4/D_5$ einen kompletten Satz orthogonal wechselseitiger Einzelvergleiche darstellen. Einen vollständigen Satz kann man sich auch mittels der Regeln für *Helmert-Kontraste / umgekehrter Helmert-Kontraste* erstellen:

Helmert-Kontraste			umgekehrte Helmert-Kontraste		
D_1	=	$\bar{A}_1 - \frac{\bar{A}_2+\bar{A}_3+\ldots+\bar{A}_p}{p-1}$	D_1	=	$\bar{A}_1$
D_2	=	$\bar{A}_2 - \frac{\bar{A}_3+\bar{A}_4+\ldots+\bar{A}_p}{p-2}$	D_2	=	$\bar{A}_3 - \frac{\bar{A}_1+\bar{A}_2}{2}$
$\ldots$	$\ldots$	$\ldots$	D_3	=	$\bar{A}_4 - \frac{\bar{A}_1+\bar{A}_2+\bar{A}_3}{3}$
D_{p-2}	=	$\bar{A}_{p-2} - \frac{\bar{A}_{p-1}+\bar{A}_p}{2}$	D_{p-2}	=	$\bar{A}_{p-1} - \frac{\bar{A}_1+\bar{A}_2+\ldots+\bar{A}_{p-2}}{p-2}$
D_{p-1}	=	$\bar{A}_{p-1}\bar{A}_p$	D_{p-1}	=	$\bar{A}_p - \frac{\bar{A}_1+\bar{A}_2+\ldots+\bar{A}_{p-1}}{p-1}$

Wir erhalten aus den zwei Sätzen orthogonaler Einzelvergleiche folgende Gewichtungsübersicht:

Set 1:				
D_1	1	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$
D_2	0	1	$-\frac{1}{2}$	$-\frac{1}{2}$
D_3	0	0	1	-1

Set 2:				
D_1	-1	1	0	0
D_2	$-\frac{1}{2}$	$-\frac{1}{2}$	1	0
D_3	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$	0

Zerlegung der QS_{treat} bei Einzelvergleichen QS_{treat} setzt sich bei vollständig orthogonalen Sätzen von Einzelvergleichen durch Addition aus diesen zusammen

$$QS_{\text{treat}} = QS_{D_1} + QS_{D_2} + \ldots + QS_{D_{p-q}}$$

Die QS_D sind definert als

$$QS_D = \frac{n \cdot D^2}{\sum_{j=1}^{J} c_j^2} \Leftrightarrow \hat{\sigma}_D^2, \text{ da } df = 1$$

und auf Signifikanz werden die Einzelvergleiche dann mittels F-Test geprüft

$$F = \frac{\hat{\sigma}_D^2}{\hat{\sigma}_{Fehler}^2}$$

Einzelvergleiche bei Stichproben unterschiedlicher Größe

Keine Gewichtung Werden die Mittelwerte zu Vergleichen zusammengefasst, geschieht dies als Durchschnittsbildung der Mittelwerte. Die unterschiedlichen Stichprobengrößen bleiben unberücksichtigt. Bei Experimenten mit unterschiedlichen Gruppengrößen ist das praktikabel, da die VPs den Gruppen zufällig zugeordnet werden.

Gewichtung In die Durchschnittbildung der Mittelwerte geht eine Gewichtung ein, die der Stichprobengröße berücksichtigt. Entsprechen die unterschiedlichen Stichprobenumfänge den Populationsgegebenheiten, ist das mittels Gewichtung zu berücksichtigen. Bei den meisten sozialwissenschaftlichen Analysen

ist dieser Vorgehensweise Vorrang zu gewähren, da die Datensätze keine experimentellen Versuchsanordnungen abbilden. Gewichtete Durchschnittsbildung:

$$D = \sum_{j=1}^{J} n_j \cdot c_j \cdot \bar{A}_j \text{ mit } \sum_{j=1}^{J} n_j \cdot c_j = 0$$

Die Schätzung der Gewichte für die Einzelvergleiche, ist nun etwas komplizierter. Mittels spezieller Regeln jedoch durchführbar. Mittels Kontrastierung besteht die Möglichkeit, unterschiedliche Effekte wechselseitig zu testen. Was bedeutet das für die Signifikanzberechnungen?
Standardmäßig testet die Varianzanalyse $H_0 : \mu_1 = \mu_2 = \ldots = \mu_j$, den Globalvergleich. Bei einer Irrtumswahrscheinlichkeit $\alpha = 0.05$ wird bei diesem Test in 5% irrtümlich H_0 abgelehnt, bzw. H_1 angenommen. Korrekterweise beibehalten wird H_0 in $1 - \alpha = 0.95$, also 95%.
Nun führen wir 2 orthogonale Einzeltests durch. Die Wahrscheinlichkeit, H_0 korrekterweise beizubehalten reduziert sich in der Folge auf $0,95 \cdot 0,95 = 0,9025$ (Multiplikationssatz bei gemeinsamen Auftreten unabhängiger Ereignisse: $2 \cdot H_0$ korrekterweise ablehnen bei 2 Versuchen). Bei einem Satz wechselseitig orthogonale Einzelvergleiche reduziert sich die Wahrscheinlichkeit, H korrekterweise anzunehmen auf

$$\pi = (1 - \alpha)^{j-1}$$

Oder anders ausgedrückt: wenn in 5% der Fälle H_0 zu Unrecht abgelehnt wird, steigt bei zunehmender Testzahl die Wahrscheinlichkeit, dass H_0 zu Unrecht abgelehnt wird.

Bonferoni-Korrektur

Werden mehrere Tests durchgeführt, man möchte aber gewährleisten, dass $\alpha = 0.05$ für alle Tests nicht überschritten wird, so sind die Irrtumswahrscheinlichkeiten der Einzeltests dementsprechend festzulegen. Die einfache Approximation von Bonferoni

$$\alpha' = \frac{\alpha}{m}$$

ist konservativ, d.h. dass die angepassten α' etwas niedriger ausfallen, als sie müssten.

Beispiel Wir führen m=4 orthogonale Einzelvergleiche mit $\alpha = 0.05$ durch. $\alpha' = \frac{0,05}{4} = 0,0125$. Jeder Einzeltest darf folglich den Wert 0,0125 nicht überschreiten.

2.1.5 A priori-Tests vs. a posteriori-Tests

Liegt einer Untersuchung eine eindeutige Hypothese über Wirkungen zugrunde, die aufgrund von Voruntersuchungen o.ä. begründet wird, können Einzeltestungen (i.d. Regel max. 3) ohne Fehlerkorrektur auskommen. A posteriori bedeutet in diesem Zusammenhang, dass man eine globale Signifikanz für die Faktorstufen erhalten hat und nun nicht hypothesengeleitet sehen möchte, welche Effekte dafür verantwortlich sind. Für A-Posteriori-Vergleiche kann man Verfahren

nach Duncan, Tukey,.., Scheffe benutzen. Der Scheffe-Test garantiert ,dass das Overall-Signifikanzniveau nicht überschritten wird.

2.2 Zweifaktorielle *ANOVA*

Bei Zweifaktoriellen *ANOVA*s gilt es einige Begrifflichkeiten zu unterscheiden. Erstens, ob eine Interaktion zwischen dem ersten und dem zweiten Faktor erlaub sein soll, zweitens, ob alle Gruppen identisch groß sind (orthogonale *ANOVA*), wie es in experimentellen Designs oft der Fall ist, oder ob es sich um ungleich große Gruppen handelt (nonorthogonale *ANOVA*)
Während bei experimentellen Daten eine gleichmäßige Zellbesetzung im Allgemeinen herstellbar ist, so ist dies bei Beobachtungsdaten nicht unbedingt der Fall. Dort müssen bestimmte Anpassungen vorgenommen werden, da sich hier die einzelnen Abweichungsquadratsummen SS_b und SS_w -im Gegensatz zum orthogonalen Fall- nicht von vornherein zur totalen Abweichungsquadratsumme SS_t aufsummieren.
Das Modell, von dem wir ausgehen, lautet wie folgt:

$$y_{ijk} = \mu + \underbrace{\alpha_j + \beta_k}_{\text{HE}} + \overbrace{\alpha\beta_{jk}}^{\text{IE 1. Ordnung}} + \underbrace{e_{ijk}}_{\text{Residuum}}$$

Bei α_j und β_k handelt es sich um die Haupteffekte der beiden Faktoren, wobei der erste Faktor über j Stufen verfügt, der zweite über k. Sind Interaktionen erlaubt, so handelt es sich bei $\alpha\beta_{jk}$ um den Effekt, den bestimmte Kombinationen der Faktorstufen j des ersten Faktors gemeinsam mit den Faktorstufen k des zweiten Faktors über die Haupteffekte hinaus auf y ausübt. Da in diesem Fall alle möglichen Einflüsse auf die abhängige Variable y, die direkt oder indirekt durch die beiden Faktoren hervorgerufen werden können im Modell enthalten sind, spricht man auch von einem gesättigten oder saturierten Modell.
Sind Interaktionen nicht erlaubt, so gilt folgendes Modell

$$y_{ijk} = \mu + \alpha_j + \beta_k + \varepsilon_{ijk}$$

in dem sich die Interaktion in den Störgrößen ε_{ijk}, wie alle anderen nicht beachteten Einflüße, bemerkbar macht. In folgender Graphik sehen wir, was man sich unter Interaktionseffekten zwischen den Faktoren A und B, sowie dem Interaktionseffekt $A \times B$ vorstellen könnte:

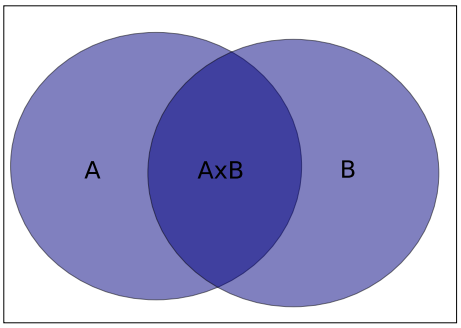


Abbildung 2.3: *ANOVA* zweifaktoriell

2.2.1 Beispiel

Als Beispiel für eine *zweifaktorielle ANOVA* erweitern wir das Experiment um den Faktor "Geschlecht"

		Faktor 1			
		Placebo	einfache Dosis	doppelte Dosis	
F	♂	y_{111} … $y_{11n_{11}}$	y_{121} … $y_{12n_{12}}$	y_{131} … $y_{13n_{13}}$	$\bar{y}_{1.}$
a	♀	y_{211} … $y_{21n_{21}}$	y_{221} … $y_{22n_{22}}$	y_{231} … $y_{23n_{23}}$	$\bar{y}_{2.}$
k		$\bar{y}_{.1}$	$\bar{y}_{.2}$	$\bar{y}_{.3}$	$\bar{y}_G$
t					
o					
r					
2					

Mit den Werten:

		Faktor A			
		Placebo	einfache Dosis	doppelte Dosis	B_j
F	♂	22	16	13	
a		25	16	12	
k		22	16	12	16.8
t		21	15	13	
o		22	15	12	
r	♀	18	19	16	
		19	20	14	
B		17	17	16	17.0
		21	16	13	
		19	16	14	
A_i	20.6	16.6	13.5	16.9	

Dnn berechnen wir die einfaktorielle Varianzanalyse für den Faktor A.

Einfaktorielle Varianzanalyse (nur Faktor A)					
QS_{tot}	$=$	348.7	QS_{tot}	$=$	$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{m=1}^{n}(x_{ijm}-\bar{G})^2$
QS_A	$=$	95.3	QS_A	$=$	$\sum_{i=1}^{p} n\cdot q\cdot(\bar{A}_i-\bar{G})^2$
QS_{Fehler}	$=$	253.4	QS_{Fehler}	$=$	$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{m=1}^{n}(x_{ijm}-\bar{A}_i)^2$

Sowie die einfaktorielle Varianzanalyse für den Faktor B.

Einfaktorielle Varianzanalyse (nur Faktor B)					
QS_B	$=$	0.3	QS_B	$=$	$n\cdot p\cdot\sum_{i=1}^{p}(\bar{B}_i-\bar{G})^2$

Zuletzt die Zelleneffekte:

Zellenquadrate: Die Effekte hängen von beiden Faktoren A/B ab:					
QS_{tot}	$=$	$307.90+40.8=348.7$	QS_{tot}	$=$	$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{m=1}^{n}(x_{ijm}-\bar{G})^2$
QS_{Zellen}	$=$	307.9	QS_{Zellen}	$=$	$n\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{A}\bar{B}_{ij}-\bar{G})^2$
QS_{Fehler}	$=$	40.8	QS_{Fehler}	$=$	$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{m=1}^{n}(x_{ijm}-\bar{A}\bar{B}_{ij})^2$

A und B erklären zusammen einen größeren Teil der abhängigen Variablen als A alleine. Geschlecht und Behandlungsart bestimmen also die Wirkung des Medikaments, ist der Zelleneffekt A/B die Summe aus Faktor A und Faktor B ($QS_{\text{Zellen}} = QS_{\text{SA}} + QS_{\text{SB}}$)?

$$QS_{\text{Zellen}} = QS_{\text{SA}} + QS_{\text{SB}} = 253.4 + 0.30 = 253.70 < 307.90$$

Die Differenz von 54.20 ist der Teil, der durch die Wechselwirkung von A und B entsteht: Die Behandlungsarten wirken geschlechtsspezifisch. Diese Wechselwirkung wird als Interaktionseffekt bezeichnet:

$$n\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{A}\bar{B}'_{ij}-\bar{A}\bar{B}_{ij})^2 \text{ mit } \bar{A}\bar{B}'_{ij}=\bar{A}_i-\bar{B}_j-\bar{G} \text{ (nur Haupteffekte)}$$

$QS_{A\times B}$	$=$	$n\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{A}\bar{B}'_{ij} - \bar{A}\bar{B}_{ij})^2$	$df_{A\times B}$	$=$	$(p-1)\cdot(q-1)$
QS_{tot}	$=$	$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{m=1}^{n}(x_{ijm} - \bar{G})^2$	df_{tot}	$=$	$p\cdot q\cdot n - 1$
QS_{Zellen}	$=$	$n\cdot\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{A}\bar{B}_{ij} - \bar{G})^2$	df_{Zellen}	$=$	$p\cdot q - 1$
QS_{Fehler}	$=$	$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{m=1}^{n}(x_{ijm} - \bar{A}\bar{B}_{ij})^2$	df_{Fehler}	$=$	$p\cdot 1\cdot(n-1)$
QS_A	$=$	$n\cdot q\cdot\sum_{i=1}^{p}(\bar{A}_i - \bar{G})^2$	df_A	$=$	$p-1$
QS_B	$=$	$n\cdot p\cdot\sum_{j=1}^{q}(\bar{B}_j - \bar{G})^2$	df_B	$=$	$q-1$
QS_{tot}	$=$	$QS_{\text{Zellen}} + QS_{\text{Fehler}}$	df_{tot}	$=$	$df_A + df_B + df_{A\times B} + df_{\text{Fehler}}$
QS_{Zellen}	$=$	$QS_A + QS_B + QS_{A\times B}$	df_{Zellen}	$=$	$df_A + df_B + df_{A\times B}$

2.2.2 Hypothesen

In einer zweifaktoriellen Varianzanalyse werden folgende Hypothesen getestet:

1. Die Untersuchungseinheiten aus dem ersten Faktor entstammen einer Population mit gleichem Mittelwert

$$H_0: \ \mu_{1.} = \mu_{2.} = \ldots = \mu_{j.}$$

2. Die Untersuchungseinheiten aus dem zweiten Faktor entstammen einer Population mit gleichem Mittelwert

$$H_0: \ \mu_{.1} = \mu_{.2} = \ldots = \mu_{.k}$$

3. Die Zellenmittelwerte sind lediglich die Summe der Haupteffekte

$$H_0: \ \mu_{jk} = \mu_j + \mu_k - \mu$$

Die Prüfung der Nullhypothesen erfolgt über den F-Test für die entsprechenden Varianzen:

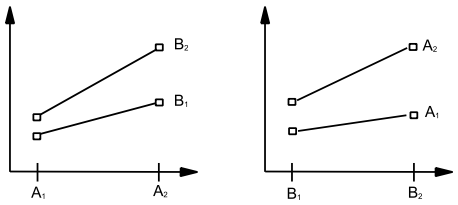
$\hat{\sigma}_{\text{Fehler}}$	$=$	$\frac{QS_{\text{Fehler}}}{df_{\text{Fehler}}}$	$=$	$\frac{40.8}{24}$					
$\hat{\sigma}_A$	$=$	$\frac{QS_A}{df_A}$	$=$	$\frac{253.4}{2}$	$\rightarrow$	F_A	$=$	$\frac{126.7}{1.7}$	$=$ 74.53
$\hat{\sigma}_B$	$=$	$\frac{QS_B}{df_B}$	$=$	$\frac{0.3}{1}$	$\rightarrow$	F_B	$=$	$\frac{0.3}{1.7}$	$=$ 0.18
$\hat{\sigma}_{A\times B}$	$=$	$\frac{QS_{A\times B}}{df_{A\times B}}$	$=$	$\frac{54.2}{2}$	$\rightarrow$	$F_{A\times B}$	$=$	$\frac{27.1}{1.7}$	$=$ 15.94

Bei $F_A/F_{A\times B\ (2,24,0.99)} = 5.61$ sind A und A signifikant, B ist bei $F_{B\ (1,24,0.99)} = 7.82$ nicht signifikant.

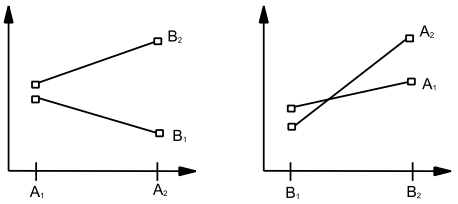
2.2.3 Wichtige Interaktionsformen

Interaktionen lassen sich in Form von Diagrammen darstellen und erleichtern die Interpretation der Ergebnisse.

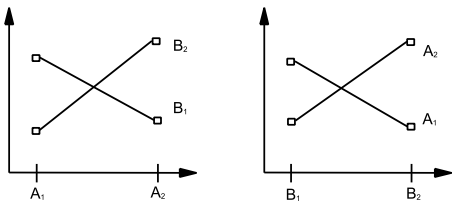
1. **Ordinale Interaktion:** Es gibt in beiden möglichen Interaktionsdiagrammen keine Überschneidungen und die Effekte sind stets gleichgerichtet (es liegen eindeutige Haupteffekte vor). Gleicher Trend der Linien für zwei mögliche Darstellungen, die Haupteffekte sind eindeutig interpretierbar



2. **Hybride Interaktion:** In einem Diagramm gibt es gegengerichtete Trends, daher überschneiden sich die Linien in dem anderen Diagramm. Haupteffekte sind mit Vorsicht zu interpretieren, in einem Faktor hängt die Reihenfolge der Stufen ja von dem anderen Faktor ab (Überschneidung). Nicht interpretierbar für Faktor A.



3. **Disordinale Interaktion:** In beiden Diagrammen gibt es starke Überschneidungen, die Haupteffekte sind nicht eindeutig bzw. nicht interpretierbar. Die Haupteffekte alleine sind bedeutungslos Die Interaktion ist die bestimmende Größe der Werte. Die Unterschiede zwischen a_1 und a_2 sind nur in Verbindung mit den Stufen von Faktor B interpretierbar, gleiches gilt für b_1 und b_2.



2.2.4 Feste und zufällige Effekte

Bei *festen Effekten* sind alle möglichen Faktorstufen Teil der Untersuchung.

Bei *zufällige Effekten* sind nicht alle möglichen Faktorstufen Teil der Untersuchung. Es werden bspw. Therapeuten zufällig ausgewählt, wenn man testen möchte, ob die Person des Therapeuten Einfluss auf den Verlauf der Therapie nimmt.
Die Unterscheidung von zufälligen und festen Effekten wird erst mit der zweifaktoriellen VA rechnerisch notwendig. Für die einfaktorielle VA ändert sich nur die Interpretation. Die Auswirkung der Hinzunahme zufälliger Effekte in die VA besteht in der Vorraussetzung, dass alle Treatmenteffekte normalverteilt sein müssen. Bei Hinzunahme von zufälligen Effekten ändern sich die Prüfvarianzen im F-Test.

2.2.5 Einzelvergleiche

Einfache Einzelvergleiche

Vergleich von bspw. Placebo gegen Medikamente (Faktor B) oder Vergleich von Psychoanalyse gegen Verhaltenstherapie (Faktor A)

$$QS_{D(A)} = \frac{n \cdot q \left(\sum_{i=1}^{p} c_i - \bar{A}_i \right)^2}{\sum_{i=1}^{p} c_i^2}$$

$$QS_{D(B)} = \frac{n \cdot p \left(\sum_{j=1}^{q} c_j - \bar{B}_j \right)^2}{\sum_{j=1}^{q} c_j^2}$$

Bedingte Haupteffekte

$$QS_{A|b_j} = n \cdot \sum_{i=1}^{p} (\bar{A}\bar{B}_{ij} - \bar{B}_j),\ df_{A|b_j} = p - 1$$

$$QS_{B|a_i} = n \cdot \sum_{i=1}^{p} (\bar{A}\bar{B}_{ij} - \bar{B}_j),\ df_{B|a_i} = q - 1$$

Fragestellung: wirken die Therapien nur bei bestimmter Medikamentendosierung?

Bedingte Einzelvergleiche

Nicht der gesamte bedingte Haupteffekt A wird mittels Einzeltests verglichen, sondern Einzeltests je Faktorstufe B werden durchgeführt.

$$D_s(A|b_j) = \sum_{i=1}^{p} c_{is} \cdot \bar{A}\bar{B}_{ij} \text{ mit } s = \text{Anzahl der Einzelvergleiche}$$

Vergleich von bspw. Psychoanalyse mit anderen Therapien gegeben ein bestimmtes Medikament. Ergebnis bspw. bei Placeboeinnahme erzielt die Psychoanalyse bessere Ergebnisse als andere Therapieformen.

Interaktionseinzelvergleiche

Z.B. Vergleich Placebowirkung bei Kontrollgruppe gegen Placebowirkung Therapie.

$$D_w(D(A) \times D(B)) = \sum_{j=1}^{q} c_{ju} \times D_s(A|b_j)$$

2.3 Dreifaktorielle *ANOVA*

In einer dreifaktoriellen *ANOVA* gehen wir von folgendem Modell aus:

$$y_{ijkl} = \mu + \underbrace{\alpha_j + \beta_k + \gamma_l}_{\text{HE}} + \overbrace{\alpha\beta_{jk} + \alpha\gamma_{jl} + \beta\gamma_{kl}}^{\text{IE 1. Ordnung}} + \underbrace{\alpha\beta\gamma_{jkl}}_{\text{IE 2. Ordnung}} + \overbrace{e_{ijk}}^{\text{Residuum}}$$

Mit 3 Haupteffekten (A, B, C), 3 Interaktionseffekten 1. Ordnung ($A \times B$, $A \times C$, $B \times C$) und einem Interaktionseffekt 2. Ordnung ($A \times B \times C$), die man sich in etwa so vorstellen kann:

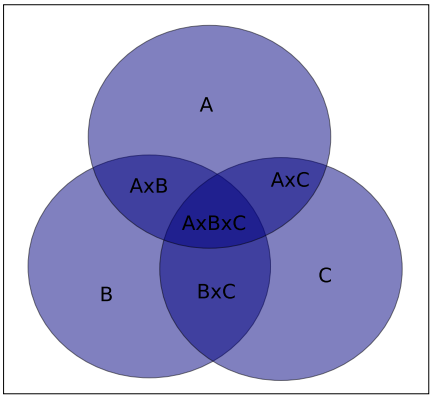


Abbildung 2.4: *ANOVA* dreifaktoriell

2.3.1 Hypothesen

Es werden folgende Hypothesen getestet:

Faktor A	:	$\mu_1 = \mu_2 = \ldots = \mu_p$
Faktor B	:	$\mu_1 = \mu_2 = \ldots = \mu_q$
Faktor C	:	$\mu_1 = \mu_2 = \ldots = \mu_r$
Interaktion $A \times B$	:	$\mu_{ij} = \mu_i + \mu_j - \mu$
Interaktion $A \times C$	:	$\mu_{ik} = \mu_i + \mu_k - \mu$
Interaktion $B \times C$	:	$\mu_{jk} = \mu_j + \mu_k - \mu$
Interaktion 2. Ord. $A \times B \times C$	:	$\mu_{ijk} = \mu_{ij} + \mu_{ik} + \mu_{jk} - \mu_i - \mu_j - \mu_k + \mu$

2.3.2 Quasi-F-Brüche/Pooling-Prozeduren

Wird für nicht direkt testbare (F-Test) Effekte benutzt, da hier die Fehlervarianzen nicht verfügbar sind. Dieser Fall ist in einem Modell mit mehr als 2 zufälligen Effekten gegeben. Zwei Strategien damit umzugehen bestehen in der Berechnung von Quasi-F-Brüchen oder Pooling-Prozeduren. Anmerkung: Interaktionen 2. Ordnung sind testbar, jedoch schwer interpretierbar. Regeln der 3-faktoriellen VA sind auf die mehrfaktorielle übertragbar.

2.3.3 Nonorthogonale *ANOVA*

Mehrfaktorielle Varianzanalysen mit ungleich großen Stichproben verletzen die Voraussetzung der Orthogonalität von Haupt- und Interaktionseffekten. Varianzanalysen mit ungleich großen Stichproben werden als *nicht orthogonale Varianzanalysen* bezeichnet.
Lösungsansätze:

1. Missing Data Techniken: Werden nur eingesetzt, wenn die Werte auch tatsächlich fehlen. Das bedeutet, wenn ursprünglich größere Stichproben geplant waren, es aber zu Datenausfällen gekommen ist.

2. *ANOVA* mit proportional geschichteten Stichproben: Die Stichproben entsprechen den Populationen und sind zeilen- und spaltenweise zueinander proportional. Die Berechnung ist fast identisch mit der *ANOVA* mit gleich großen Stichprobenumfängen.

Faktor B		Faktor A		
		A_1	A_2	A_3
	B_1	$n_{11} = 5$	$n_{12} = 15$	$n_{13} = 10$
	B_2	$n_{21} = 20$	$n_{22} = 60$	$n_{23} = 40$
	B_3	$n_{31} = 10$	$n_{32} = 30$	$n_{33} = 20$
	B_3	$n_{41} = 15$	$n_{42} = 45$	$n_{43} = 30$

3. *ANOVA* mit harmonischem Mittel der Stichprobenumfänge: Für ungleich große Stichproben, die nicht proportional geschichtet sind.

Quadratsummenberechnung / Prüfgrößen für den F-Test

QS	df
$QS_A \cdot \bar{n}_h$	$p-1$
$QS_B \cdot \bar{n}_h$	$q-1$
$QS_{A\times B} \cdot \bar{n}_h$	$(p-1)(q-1)$
$QS_{\text{Fehler}} \cdot \bar{n}_h$	$N - p \cdot q$

Der Einsatz des Harmonischen Mittels setzt voraus, das ursprünglich gleich große Stichproben geplant waren. Ersetzung der Stichprobenumfänge durch das harmonische Mittel aller Stichprobengrößen (zweifaktorieller Fall):

$$HM = \bar{n}_h = \frac{j \cdot k}{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \ldots + \frac{1}{n_{jk}}} = \frac{j \cdot k}{\sum_{j=1}^{J} \sum_{k=1}^{K} \frac{1}{n_{jk}}}$$

4. *ANOVA* nach dem ALM

KAPITEL 3

LOGISTISCHE REGRESSION

Ist die uns interessierende abhängige Variable metrisch, so hilft uns der bereits besprochene lineare Regressionsansatz weiter. Es ist allerdings auch möglich, dass die abhängige Variable nominales (also dichotom oder multinomial) Meßniveau aufweist. In diesem Fall kommen wir mit dem klassischen Regressionsansatz nicht weiter. Die Methode, die hier besprochen werden soll ermöglicht es uns, nominale Variablen als abhängige Variablen zu nutzen. Wir werden uns den Fall einer *dichotomen* abhängigen Variablen ansehen.
Folgend besprechen wir die Probleme, die in solchen Fällen die logistische Regression zur Methode der Wahl machen, und nicht die lineare Regression:

1. Allgemein handelt es sich bei den vorhergesagten $\hat{y}$-Werten um Schätzungen des bedingten Mittelwertes der abhängigen Variable. Denken wir uns als Beispiel 10 Personen, von denen jeweils das Alter und der Familienstatus erfragt worden sind:

Alter	18	25	27	29	34	35	42	42	51	60
Verheiratet	0	0	0	1	1	0	0	1	1	1

   Betrachten wir den Mittelwert eines Dummies, so können wir ihn als Anteil der mit 1 codierten Fälle betrachten. Bei unseren 10 Fällen sind 5 ledig (mit 0 codiert) und 5 verheiratet (mit 1 codiert). Es ergibt sich folgender Mittelwert:

   $$\frac{0+0+0+0+0+1+1+1+1+1}{10} = \frac{5}{10} = 0.5$$

   Den vorhergesagten Wert einer dichotomen Variablen kann man also so interpretieren, dass ein Wert 0.5 einer Wahrscheinlichkeit für das Vorhandensein des Merkmals von 50% entspricht. Wir haben hier also eine 50%-Wahrscheinlichkeit, eine verheiratete Person zu erwischen.

   Das Problem besteht nun darin, dass wir bei entsprechend großen bzw. kleinen x-Werten vorhergesagte Werte größer 1 oder kleiner 0 erhalten. Ein vorhergesagter Mittelwert von 1.2 entspräche einer 120% Wahrscheinlichkeit, einen Verheirateten zu treffen. Dies ist natürlich Unsinn, ergibt

sich aber zwingend bei einer linearen Vorhersage. Es sind also nicht alle vorhergesagten Werte inhaltlich interpretierbar.

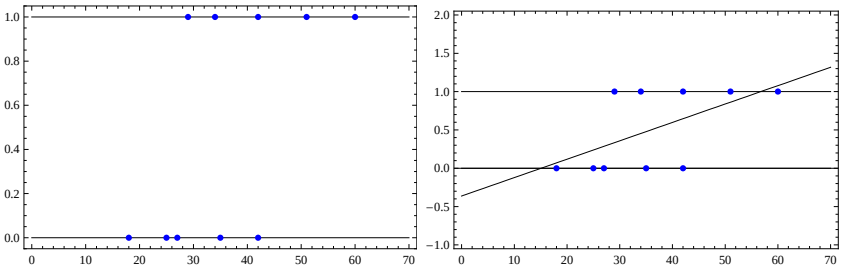


Abbildung 3.1: Linearer Regressionsansatz

Das logistische Regressionsmodell wurde entwickelt, um dieses Problem zu beheben. Es ersetzt die Regressionsgerade durch eine $S-$förmige Kurve die sich den Werten 0 und 1 asymptotisch nähert. Der Wertebereich ist also hier auf $[0;1]$ festgesetzt, während er in der linearen Regression mit $[-\infty;+\infty]$ unbeschränkt ist

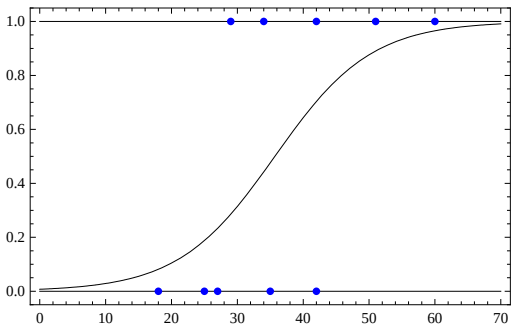


Abbildung 3.2: Logistischer Regressionsansatz

2. Es ergeben sich Probleme mit der Homoskedastizitätsannahme der linearen Regression. Hierbei soll die Varianz der Fehler für alle $\hat{y}$ konstant sein. Sagen wir für einen Verheirateten (1) auf Grund einer unabhängigen Variablen wie beispielsweise des Alters eine Wahrscheinlichkeit von 0.6 dafür vorher, dass er verheiratet ist, so liegt ein Residuum von $y_i - \hat{y}_i = 1 - 0.6 = 0.4$ vor. Sagen wir auf Grund des Alters für eine ledige Person einen Wert von 0.6 vorher, verheiratet zu sein, so ergibt sich ein Residuum von $y_i - \hat{y}_i = 0 - 0.6 = -0.6$. Es sind also nur $1 - \hat{y}_i$ sowie $-\hat{y}_i$ als Residuen möglich. Die bedingte Varianz $\hat{y}_i \times (1 - \hat{y}_i)$ der Residuen ist umso größer, je näher die vorhergesagten Werte an 0.5 herangehen.

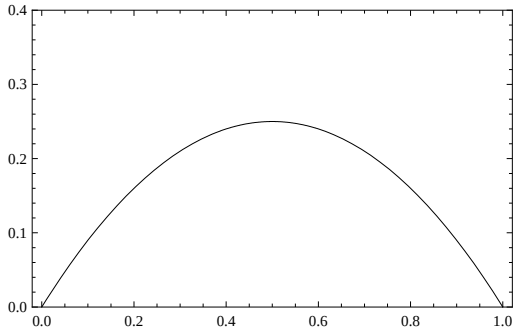


Abbildung 3.3: $\hat{y}_i \times (1 - \hat{y}_i)$

Sie sind demnach *per definitionem* heteroskedatisch. Dies führt zu Problemen mit der Berechnung der Konfidenzintervalle der Regressionskoeffizienten. Sie sind nicht mehr zuverlässig.

3.1 Grundidee

Im Gegensatz zur linearen Regression führt die logistische Regression keine Vorhersage der (wie in der linearen Regression) metrischen y_i-Werte der abhängigen Variablen durch, sondern eine Vorhersage der *Eintrittswahrscheinlichkeit* der (dichotomen) y_i-Werte. Hierzu wird der Ansatz der linearen Regression verändert, so dass sich keine Regressionsgerade, sondern die oben erwähnte, für die logistische Regression charakteristische S-Kurve ergibt. Um die Eintrittswahrscheinlichkeit des Ereignisses bestimmen zu können, wird eine latente Variable z angenommen, die die y-Ausürägungen in Abhängigkeit der unabhängigen Variablen x_j erzeugen kann. Es gilt:

$$y_k = \begin{cases} 1 & \text{falls } z_k > 0 \\ 0 & \text{falls } z_k \leq 0 \end{cases}$$

mit

$$z_k = b_0 + \sum_{j=1}^{J} b_j x_{ij} + \varepsilon_i = \text{Logit}$$

Um nun die Wahrscheinlichkeitsaussage bezüglich des Eintretens von y treffen zu können, benötigen wir noch eine Wahrscheinlichkeitsfunktion, die dann $y = 0$, bzw. $y = 1$ aus z_k erzeugen kann. Hier wird auf die logistische Wahrscheinlichkeitsfunktion

$$p_k(y = 1) = \frac{e^{z_k}}{1 + e^{z_k}} = \frac{1}{1 + e^{-z_k}}$$

mit

$$z_k = b_0 + \sum_{j=1}^{J} b_j x_{ij} + \varepsilon_i = \text{Logit}$$

zurückgegriffen. Bei den b_j Koeffizienten spricht man auch von Logit-Koeffizienten. Die logistische Funktion stellt die Wahrscheinlichkeitsbeziehung zwischen y und x_j her, sie wird auch *linking-function* genannt. Es ist zu beachten, dass es sich bei

$$p_k(y=1) = \frac{e^{z_k}}{1+e^{z_k}} = \frac{1}{1+e^{-z_k}}$$

um einen nicht-linearen Zusammenhang zwischen dem Eintreten von y und x_j handelt (S-Kurve), das Zustandekommen der aggregierten Einflussstärke z_k

$$z_k = b_0 + \sum_{j=1}^{J} b_j x_{ij} + \varepsilon_i = \text{Logit}$$

aber als linear unterstellt wird.

Im Gegensatz zur linearen Regression wird hier keine unmittelbare *je-desto*-Hypothese zwischen y und x aufgestellt, sondern zwischen x und der Eintrittswahrscheinlichkeit von $y = 1$.

3.2 Herleitung der logistischen Regressionsgleichung

Der Einfachheit und Übersichtlichkeit halber verkürzen wir die Schreibweise von

$$\sum_{i=1}^{n} b_i x_i \quad \text{auf} \quad bx_i$$

also auf den bivariaten Fall und

$$p(Y=1) \quad \text{auf} \quad p$$

Ebenso gilt für die Konstante $a = b_0$. Wenn wir eine Wahrscheinlichkeit durch lineare Regression vorhersagen wollen treffen wir auf Probleme: Die Wahrscheinlichkeit ist auf das Intervall von 0 bis 1 festgelegt. Sie kann nicht negativ oder grösser 1 werden, so wie es die rechte Seite der Formel $p = a + bx_i$ kann. Um dieses Problem zu lösen betrachtet man zuerst die Odds (Chance), also den Quotienten aus zwei Wahrscheinlichkeiten, nämlich einmal der Wahrscheinlichkeit des Eintretens ($p(Y=1)$) und der Wahrscheinlichkeit, dass das Ereignis nicht eintritt ($1 - p(Y=1)$).

$$\frac{p}{1-p} = a + bx_i$$

Betrachten wir ein beliebiges Ereignis, dass entweder eintreten kann, oder nicht, wie z.B. Regen. Der Odd der Wahrscheinlichkeit, dass das Ereignis x eintritt, es also regnet ($p(x) = 0.75$) beträgt $\frac{p(x)}{1-p(x)} = \frac{0.75}{0.25} = 3$. Also ist die Wahrscheinlichkeit, dass es regnet 3 mal höher als das es trocken bleibt. Das ist schon besser. Aber immer noch nicht OK, den die Odds können nich negativ werden, sie besitzen einen Wertebereich zwischen 0 und $+\infty$. Durch logarithmieren (üblicherweise mit dem logarithmus naturalis ln) erreichen wir einen Wertebereich zwischen $-\infty$ und $+\infty$. Die logarithmierten Odds werden Logits genannt.

$$\ln \frac{p}{1-p} = a + bx_i$$

Wenn wir die Gleichung nun nach p auflösen, da uns ja gerade p interessiert, gehen wir folgendermaßen vor:

$$e^{\ln \frac{p}{1-p}} = e^{a+bx_i}$$

Da gilt $e^{\ln x} = \ln e^x = x$, es sich also um die Umkehrfunktion handelt, gilt folgendes:

$$\frac{p}{1-p} = e^{a+bx_i}$$

Multiplikation mit $1 - p(x)$

$$p = e^{a+bx_i}(1-p)$$

Ausmultiplizieren

$$p = e^{a+bx_i} - pe^{a+bx_i}$$

Addition, um pe^{a+bx_i} auf die linke Seite zu bringen:

$$p + pe^{a+bx_i} = e^{a+bx_i}$$

Ausklammern von p

$$p\left(1 + e^{a+bx_i}\right) = e^{a+bx_i}$$

Dividieren durch $\left(1 + e^{a+bx_i}\right)$

$$p = \frac{e^{a+bx_i}}{1 + e^{a+bx_i}}$$

Hier ist in manchen Lehrbüchern Schluss, wir haben die Formel der logistischen Regression erreicht. Doch kann man noch weiter vereinfachen: Klammern wir unter dem Bruchstrich e^{a+bx_i} aus.

$$p = \frac{e^{a+bx_i}}{e^{a+bx_i}(\frac{1}{e^{a+bx_i}} + 1)}$$

Umschreiben, da gilt $\frac{1}{a} = a^{-1}$

$$p = \frac{e^{a+bx_i}}{e^{a+bx_i}(e^{-(a+bx_i)} + 1)}$$

Finales Kürzen

$$p(Y = 1) = \frac{1}{1 + e^{-(a+bx_i)}}$$

3.3 Maximum Likelihood-Schätzung

Es gilt

$$p_k(y) = \begin{cases} \frac{1}{1+e^{-z_k}} & \text{für } y_k = 1 \\ 1 - \frac{1}{1+e^{-z_k}} & \text{für } y_k = 0 \end{cases}$$

was sich zusammenfassen lässt zu:

$$p_k(y) = \left(\frac{1}{1+e^{-z_k}}\right)^{y_k} \left(1 - \frac{1}{1+e^{-z_k}}\right)^{1-y_k}$$

Für alle k Fälle zusammen greift man auf den Wahrscheinlichkeitssatz für unabhängige Ereignisse zurück, um die Likelihood-Funktion zu bilden, die maximiert werden soll:

$$L(\theta) = \prod_{k=1}^{K} \left(\frac{1}{1+e^{-z_k}}\right)^{y_k} \left(1 - \frac{1}{1+e^{-z_k}}\right)^{1-y_k} \stackrel{!}{=} \max$$

Es ist einfacher, die Log-Likelihood zu maximieren, als die normale Likelihood. Dies liegt daran, dass es einfacher ist, mit Summen zu arbeiten, als mit Produkten. Die Extremwerte bleiben identisch. Es gelten folgende Regeln zum Rechnen mit Logarithmen:

$$\begin{aligned} \log_a(u \cdot v) &= \log_a u + \log_a v \\ \log_a(u^r) &= r \log_a u \ (r \in \mathbb{R}) \end{aligned}$$

Hier wird der Logarithmus naturalis verwendet, also $\log_e$, der logarithmus mit der eulerschen Zahl als Basis.

$$\mathcal{L}(\theta) = \sum_{k=1}^{k} \left(y_k \cdot \ln\left(\frac{1}{1+e^{-z_k}}\right)\right) + \left((1-y_k) \cdot \ln\left(1 - \frac{1}{1+e^{-z_k}}\right)\right)$$

Maximierung erfolgt in vielen Programmpaketen durch den Newton-Raphson-Algorithmus (Annäherung an den Nullpunkt durch Iteration).

3.4 Interpretation

Die Logits sind nicht leicht zu interpretieren, da es sich bei ihnen, wie wir später sehen werden, um logarithmierte Odds (Chancen) handelt. Sie werden deshalb wieder in normale Odds zurücktransformiert, indem das entsprechende z_k in $\frac{1}{1+e^{-z_k}}$ eingesetzt wird. Die Regressionskoeffizienten a und b werden als Logit-Koeffizienten bezeichnet, und im Gegensatz zur OLS-Regression über das Maximum-Liklihood-Verfahren ermittelt.
Der Logit-Koeffizient a hat Einfluss auf die Lage der Kurve, nicht auf ihre Steigung. Bei positivem a verschiebt sich die Kurve nach links, bei negativen a nach rechts. Der Koeffizient b hat hier nicht die Eigenschaften, wie im Falle der linearen Regression, d.h. gleiche Veränderungen von x_j in unterschiedlichen Bereichen wirken sich unterschiedlich auf die Eintrittswahrscheinlichkeiten von y aus, da es sich ja um einen nichtlinearen Zusammenhang handelt.
Ferner gilt:

b=0 Für alle Beobachtungen von x_j liegen die Wahrscheinlichkeiten bei 0.5

0<b<1 Die Wahrscheinlichkeitswerte steigen in Abhängigkeit von x_j nur sehr langsam an

+b Die Wahrscheinlichkeitswerte steigen mit größer werdenden Werten x_j (nicht linear)

-b Die Wahrscheinlichkeitswerte sinken mit steigenden Beobachtungswerten x_j

Mit der Möglichkeit, die logistische Regression in ein Logitmodell zu überführen, in der Form

$$\begin{aligned} p_i &= \tfrac{e^{z_i}}{1+e^{z_i}} \\ e^{z_i} &= p_i \cdot (1 + e^{z_i}) \\ e^{z_i} &= p_i + p_i e^{z_i} \\ e^{z_i} &= e^{z_i}(1 - p_i) \\ \tfrac{p_i}{1-p_i} &= e^{z_i} \\ \ln \tfrac{p_i}{1-p_i} &= z_i \end{aligned}$$

ist eine bessere Interpretation des Einflusses der unabhängigen Variablen x_j (Mehrvariablenfall) auf die Eintrittswahrscheinlichkeiten p verbunden.

- $\frac{p_i}{1-p_i} = e^{z_k}$ wird als der Odds bezeichnet, diese drücken ein Chancenverhältnis aus. Bsp.: $p(Y = 1) = 0.8 \rightarrow \frac{0.8}{0.2} = 4$, d.h. bei einer Odds von 4 ist die Chance des Eintretens von y vier mal größer als das Nichteintreten.
- $\ln \frac{p_i}{1-p_i} = z_i$ Wird als Linkfunktion bezeichnet, die den Regressionsausdruck mit der Wahrscheinlichkeit p_i verbindet: $\ln \frac{p_i}{1-p_i} = \alpha + \beta x_i$ Die Linkfunktion ist der logarithmierte Odds, sie wird als Logit bezeichnet.

Da Informationen über die der logarithmierten Erfolgschancen etwas fremd anmuten, bedient man sich verschiedener Hilfskonstruktionen:

Vorzeicheninterpretation: Die einfachste Möglichkeit ist, sich bei der Interpretation der Koeffizienten auf die Vorzeichen und die relative Größe der Koeffizienten zu beschränken. Ein positives Vorzeichen bedeutet beispielsweise, dass die Wahrscheinlichkeit für $y = 1$ mit der entsprechenden unabhängigen Variablen ansteigt, ein negatives Vorzeichen, dass die Wahrscheinlichkeit fällt. Je größer der Betrag der Koeffizienten, desto größer das Ausmaß der Veränderung. Über das genaue Ausmaß lassen sich aber so keine Rückschlüsse ziehen.

Interpretation der Odds-Ratios: Da es sich bei den Logits um die logarithmierten Chancen (Odds) handelt, können wir sie wieder in normale Chancen (Odds) umrechnen. Der schnellste Weg, um die Odds-Ratios zu erhalten, ist das direkte exponieren der b-Koeffizienten.

$$\frac{e^{b_0+b_1(x_1+1)}}{e^{b_0+b_1x_1}} = \frac{e^{b_0+b_1x_1}e^{b_1}}{e^{b_0+b_1x_1}} = e^{b_1}$$

Wenn sich die unabhängige Variable x_j um eine Einheit erhöht, dann steigt die Chance für $y = 1$ um das e_j^b-fache. Steigt x_j um c Einheiten, so erhöht

sich die Chance für $y = 1$ um $c \cdot e_j^b$. Man spricht im Zusammenhang mit den Odds-Ratios von multiplikativen Einheitseffekt, im Gegensatz zum additiven Einheitseffekt der Regressionskoeffizienten in der linearen Regression.

Wahrscheinlichkeitsinterpretation: Die Dritte Möglichkeit ist zur Interpretation der Logitkoeffizienten liegt in der Umrechnung in Wahrscheinlichkeiten. Über

$$p(y = 1) = \frac{1}{1 + e^{-z_k}}$$

kann für jeden z_k-Wert die entsprechende Wahrscheinlichkeit $p(y = 1)$ angegeben werden, also wie hoch die Wahrscheinlichkeit ist, dass für $z_k = b_0 + \sum_{j=1}^{J} b_j x_{ij}$ gilt: $p(y = 1)$. Das Problem bei der Interpretation der Wahrscheinlichkeiten besteht nun aber darin, dass sie nicht linear mit der Erhöhung der unabhängigen Variablen ansteigen. Eine Erhöhung von x_j um eine Einheit hat also *nicht* immer den selben Effekt.

Relevant für die Interpretation ist auch der Effekt-Koeffizient e^b, der eine genauere Analyse des Einflusses der exogenen Variablen auf die Eintrittswahrscheinlichkeit erlaubt. Hierzu betrachten wir jetzt erst einmal wieder den Odds (Wahrscheinlichkeitsverhältnis von $\frac{p(Y=1)}{p(Y=0)}$ und erhöhen dabei die exogene Variablen um 1

$$\frac{p_i}{1 - p_i} = e^{a+b(x_i+1)} = e^a \cdot e^{bx_i} \cdot e^b = \frac{p_i}{1 - p_i} \cdot e^b$$

Es zeigt sich, dass man den Effekt-Koeffizienten als Faktor begreifen kann, der das Wahrscheinlichkeitsverhältnis (Odds) verändert.

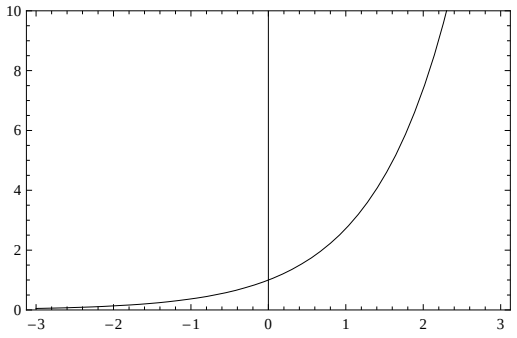


Abbildung 3.4: Exponentialfunktion $f(b) = e^b$

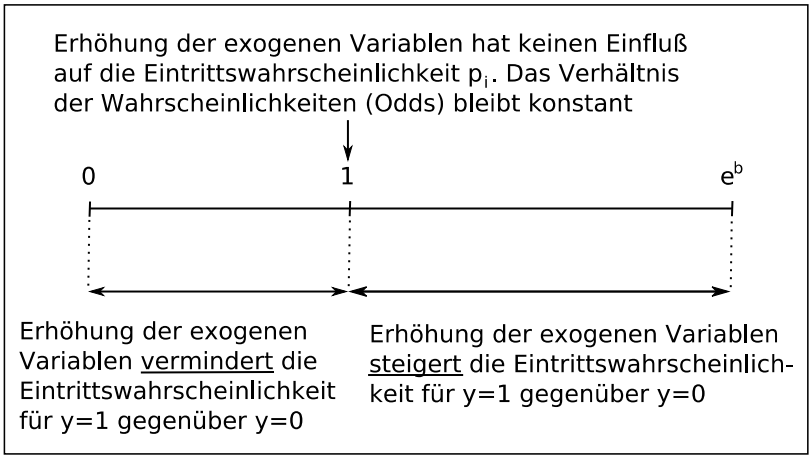
Der Effektkoeffizient e^b kann Werte zwischen 0 und $+\infty$ annehmen. Bei negativen Regressionskoeffizienten b verringert sich das Wahrscheinlichkeitsverhältnis, bei positivem b vergrößert der Faktor e^b das Wahrscheinlichkeitsverhältnis.

Beispiel:

$$\frac{p_i}{1 - p_i} \cdot e^b, \; p_i = 0.6, \; 1 - p_i = 0.4, \; \text{Odds} = \frac{0.6}{0.4} = 1.5$$

1. $b = 0,\ \frac{0.6}{0.4} \cdot e^0 = 1.5 \cdot 1 = 1.5$
2. $b = 2,\ \frac{0.6}{0.4} \cdot e^2 = 1.5 \cdot 7.39 = 11.081$
3. $b = -2,\ \frac{0.6}{0.4} \cdot e^{-2} = 1.5 \cdot 0.125 = 0.203$

Eine Erhöhung der unabhängigen Variablen um eins bewirkt in Abhängigkeit von b eine Verbesserung oder Verschlechterung des Wahrscheinlichkeitsverhältnisses. Eine Veränderung um das e^b-fache (Erhöhung der unabhängigen Variablen um eine Einheit) beeinflusst linear das Wahrscheinlichkeitsverhältnis (Odds), p_i jedoch in nicht-linearer Weise. Dies liegt daran, dass die Auswirkung einer Erhöhung von x_j durch die Linkfunktion vermittelt wird.
Man kann nun die Auswirkung der Erhöhung von e^b danach unterscheiden, ob sie den Odds zu Gunsten oder zu ungunsten von p_i verändern.



Der Logit-Koeffizient b_j ist an die Ausprägungen der x_j gebunden, so dass auch der Effekt-Koeffizient e^b von der Skalierung der x_j abhängt. Möchte man eine Vergleichbarkeit der (Variablen) Effekte erreichen, braucht es eine Standardisierung. Der normierte Logit-Koeffizient

$$\beta_j = b_j \sqrt{var(x_j)}$$

führt zum standardisierten Effektkoeffizienten

$$\beta_j = e^{b_j \sqrt{var(x_j)}}$$

3.5 Prüfung des logistischen Modells:

Für ausgewählte Fälle wird ein Vergleich von tatsächlicher (beobachteter) Gruppenzugehörigkeit und den durch die Schätzung hervorgegangenen Gruppenzuordnungen gelistet, wobei für die Zuordnung zu einer Gruppe gilt:

$$y_k = \begin{cases} \text{Gruppe } y = 1 & falls\ p_k(y = 1) > 0.5 \\ \text{Gruppe } y = 0 & falls\ p_k(y = 1) < 0.5 \end{cases}$$

3.5.1 Klassifikationsmatrix

Gegenüberstellung der richtigen und falschen Zuordnungen erfolgt über eine Kreuztabellierung in der Form, dass richtige Zuordnungen in der Hauptdiagonalen stehen und Fehlklassifikationen in den übrigen Feldern.

		Vorhergesagt		
		Gruppe $y = 1$	Gruppe $y = 0$	% richtig
Beobachtet	Gruppe $y = 1$	10	2	83.3
	Gruppe $y = 0$	2	10	83.3
	Gesamt %			83.3

Man kann nun die Trefferquote mit einer Trefferquote vergleichen, die rein zufällig entstanden wäre, d.h. durch völlig willkürliche Zuordnung der Personen zu einer der Gruppen (Münzwurf).
Anmerkung: Da die Trefferquoten sehr stark an die aktuelle Stichprobe gebunden sind, ist davon auszugehen, dass sie in einer anderen Stichprobe niedriger ist, da sie auf der Basis der Anpassung an die aktuellen Stichprobendaten entstanden ist. Die Überschätzung der Trefferquote kann durch Kreuzvalidierung des Modells überwunden werden. Dies geschieht durch eine Berechnung der Schätzfunktion auf Basis einer Stichprobe, die Zuordnung oder Klassifikation der Elemente auf Basis der Schätzfunktion erfolgt aber in einer anderen Stichprobe. In ausreichend großen Datensätzen (mit genügend großen Zellbesetzungen) kann die Kreuzvalidierung innerhalb des Datensatzes durchgeführt werden.

3.5.2 Press's Q-Test

Bezieht sich auf die Klassifikationsmatrix und überprüft die Abweichung der Trefferquote aufgrund der Berechnungen von der Trefferquote auf Basis einer zufälligen Zuordnung. Die Prüfgröße ist χ^2-verteilt mit $df = 1$. Getestet wird die H_0: Die Klassifikation der Elemente entspricht einem Zufallsprozess.
Press's Q berechnet sich über:

$$Q = \frac{[n - (n \cdot g \cdot a)]^2}{n(g-1)}$$

wobei:
n den Stichprobenumfang, g die Anzahl der Gruppen und a den Anteil der korrekt klassifizierten Elemente angibt. Liegt Q oberhalb des kritischen Wertes (bei $\alpha = 0.05 \rightarrow 3.84$) wird die H_0 abgelehnt, die Klassifikationsergebnisse sind signifikant von denen eine zufälligen Zuordnung unterschieden.

3.5.3 Hosmer-Lemeshow-Test

Prüfung der Nullhypothese, dass die Differenz zwischen vorhergesagtem und beobachtetem Wert null ist, also $H_0 : y_k - (\text{Zuordnung gemäß} p_k) = 0$. Die Fälle werden in Gruppen aufgeteilt, dann werden beobachtete und erwartete Zuordnungen verglichen. Liegt die Prüfgröße innerhalb der kritischen Grenzen, wird H_0 beibehalten.

3.5.4 Devianzanalyse

Bei der Maximum Likelihood-Methode maximieren wir die Wahrscheinlichkeit der Parameter bei gegebenen Daten. Das 2-fache der log-likelihood ist annähernd χ^2-verteilt mit $df = N - J - 1$, wobei N die Anzahl der Beobachtungen und J die Anzahl Parameter angibt. 2 wird als Devianz bezeichnet und ist mit der Fehlerquadratsumme der linearen Regression vergleichbar. Bei einem perfekten Fit ist die Devianz = 0. Es wird die H_0: "Das Modell besitzt eine perfekte Anpassung" getestet, je geringer der Wert für $-2 \ln L$, desto besser der Fit. Die Schiefe der Verteilung der Beobachtungen hat einen Einfluss auf die Devianz. Ein Schiefer Datensatz hat tendentiell eine bessere Anpassung, als ein gleichverteilter Datensatz.

3.5.5 Likelihood-Ratio-Test

Beim LR-Test vergleichen wir die Devianz des vollständigen Modells $-2 \ln L_1$ mit der Devianz des Nullmodells $-2 \ln L_0$. Beim Nullmodell handelt es sich *nicht* um das Modell mit einer Devianz $-2 \ln L = 0$, sondern um das Modell, in dem nur die Konstante vorhanden ist, und alle anderen Parameter auf 0 gesetzt werden. Bei größer die Differenz, desto mehr tragen die unabhängigen Variablen zur Unterscheidung der y-Zustände bei.
Als H_0 wird "Alle Logitkoeffizienten sind = 0" getestet.
Als Testgröße fungiert die absolute Differenz zwischen $-2 \ln L_0$ und $-2 \ln L_1$, also:

$$\chi^2_L = -2(\ln L_0 - \ln L_1)$$

Diese ist annähernd χ^2-verteilt mit $df = J$. Der χ^2-Wert kann also ähnlich dem F-Wert in der linearen Regression genutzt werden, um zu prüfen, ob alle $b_j = 0$ sind. Bei großen Werten ist die H_0 abzulehnen, was auf ein für die Daten relevantes Modell hinweist. Allerdings ist auch hier, wie beim F-Test in der linearen Regression eine einfache Zurückweisung nicht ausreichend, um mit dem Modell zufrieden zu sein.

3.6 Pseudo-r^2

3.6.1 McFaddens - r^2

Analog zum Determinationskoeffizienten der linearen Regression r^2 kann die Güte des logistischen Regressionsmodells mit McFaddens Pseudo-r^2 beurteilt werden. Hierbei handelt es sich um ein globales Gütemaß, das aus den logarithmierten Maximum-Likelihood-Schätzungen des Ausgangsmodells (nur Konstante) $-2 \ln L_0$ und $-2 \ln L_1$ des Modells mit den unabhängigen Variablen berechnet wird. Es beurteilt die Trennkraft der unabhängigen Variablen.

$$\text{McFaddens} - r^2 = 1 - \frac{-2 \ln L_1}{-2 \ln L_0} = 1 - \frac{\ln L_1}{\ln L_0} = \frac{\ln L_0 - \ln L_1}{\ln L_0}$$

Besteht kein Unterschied zwischen L_0 und L_1 wird, dann wird r^2 den Wert null annehmen, je größer der Unterschied, desto stärker geht r^2 gegen 1, ohne diesen Wert jedoch zu erreichen. Werte zwischen 0.2 und 0.4 deuten auf einen guten Modellfit hin.

Während der Likelihood-Ratio-Test ein Test auf Übertragbarkeit der Stichprobenergebnisse auf die Grundgesamtheit ist, handelt es sich bei McFaddens Pseudo-r^2 um einen Modellvergleich, der die Trennkraft der unabhängigen Variablen beurteilt.

3.6.2 Cox & Snell - r^2

$$\text{Cox \& Snell} - r^2 = 1 - \left[\frac{L_0}{L_1}\right]^{\frac{2}{n}}$$

Gegenüberstellung der nicht logarithmierten Likelihoodwerte. Der Koeffizient wird über den Stichprobenumfang gewichtet. Akzeptable Werte >0.2, gute Werte >0.4, Nachteil: erreicht nur Werte <1

3.6.3 Nagelkerke - r^2

Erreicht im Gegensatz zu den beiden anderen Pseudo-r^2-Koeffizienten den Maximalwert von 1, und sollte in der Analyse vornehmlich genutzt werden. Desweiteren soll er eine Interpretation wie die des "originalen" Determinationskoeffizienten in der linearen Regression zulassen, er gibt den Anteil der Varianzerklärung der abhängigen Variable durch die unabhängigen Variablen an. Werte ab 0.5 deuten auf einen guten Modellfit hin. Er berechnet sich über:

$$\text{Nagelkerke} - r^2 = \frac{\text{Cox \& Snell} - r^2}{1 - [L_0]^{\frac{2}{n}}}$$

oder anders

$$\text{Nagelkerke} - r^2 = \frac{r^2}{r^2_{max}}$$

mit $r^2_{max} = 1 - (L_0)^{\frac{2}{n}}$ und L_0, der Likelihood des Nullmodells, in welchem nur die Konstante geschätzt wird, aber keine unabhängigen Variablen eingegangen sind.

3.7 Diagnostik

3.7.1 Linearität

Bei der logistischen Regression muss die funktionale Form des Scatterplots nichtlinear sein, da sich die Linearitätsannahme auf die Logits bezieht, nicht auf den Zusammenhang zwischen Eintrittswahrscheinlichkeit $p(y = 1)$ und x_j, der als S-förmig angemonnem wird. Dies kann über sogenannte LOWESS (**LO**cally **WE**ighted **S**catterplot **S**moother) inspiziert werden, hier liegt eindeutig keine S-förmiger Zusammenhang vor.

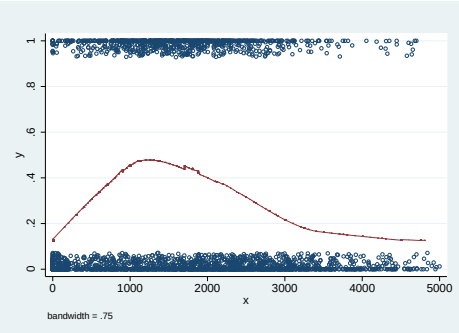


Abbildung 3.5: LOWESS

Die abhängige Variable in Abbildung 3.5 ist dichotom, um Überlagerungen zu vermeiden wurde zu jedem y-Wert jeweils ein Zufallswert addiert, damit sie nicht alle auf den Werten 0 und 1 liegen, sondern leicht um diese Werte herum "zittern" (jitter).

3.7.2 Ausreißer

Es geht um die Beurteilung des Effektes, den einzelne Personen auf die Modellgüte haben. Diese können neben einer schlechten Auswahl der unabhängigen Variablen eine schlechte Modellanpassung verursachen. Die Identifizierung von Ausreißern erfolgt über die Berechnung der Residuen, d.h. es wird die Diskrepanz zwischen empirischem Wert und geschätzter Wahrscheinlichkeit $p(y = 1)$ berechnet. Als Ausreißer gelten Personen, deren standardisierte Residuen über 0,5 liegen. Die Berechnung der standardisierten Residuen erfolgt nach

$$Z_{Resid_k} = \frac{y_k - p(y_k = 1)}{\sqrt{p(y_k = 1) \cdot (1 - p(y_k = 1))}}$$

Bei einem beobachteten Wert für Person k mit $y = 1$ und einer geschätzten Wahrscheinlichkeit $p(y_k = 0.073)$ ergibt sich dann

$$Z_{Resid_k} = \frac{1 - 0.073}{\sqrt{0.073 \cdot 0.927}} = \frac{0.927}{0.2601} = 3.564$$

Person k kann nach dem Kriterium $Z_{Resid} > 0.5$ als Ausreißer angesehen werden.

3.8 Prüfung der Merkmalsvariablen

Angaben zur Trennfähigkeit der einzelnen Variablen geben der Likelihood-Quotienten-Test und die Wald-Statistik.

3.8.1 Likelihood-Quotienten-Test

Ähnelt dem LR-Test, ist aber kein Vergleich des vollständigen Gesamtmodells $\ln L_V$ gegen das Nullmodell, sondern ein Vergleich unterschiedlicher reduzierter Modelle, wobei jeweils ein Koeffizient $b_j = 0$ gesetzt wird. Dann wird die Differenz der $-2 \ln L$ zwischen vollständigem ($\ln L_V$) und reduzierten ($\ln L_R$)Modell betrachtet.

H_0: Der Effekt von b_j ist Null ($b_j = 0$) Die Testgröße ($\ln L_R - \ln L_V$) ist χ^2-verteilt und somit kann auf dieser Basis eine Signifikanzprüfung durchgeführt werden.

3.8.2 Wald-Statistik

Testet die Nullhypothese, dass ein bestimmtes b_j Null ist, also die unabhängige Variablen nicht zur Trennung der Gruppen beiträgt. Die Wald-Statistik

$$W = \left(\frac{b_j}{s_{b_j}}\right)^2 \text{ mit } s_{b_j} = \text{Standardfehler von } b_j$$

ist ebenfalls asymptotisch χ^2-verteilt.

KAPITEL 4

DISKRIMINANZANALYSE

Bei der Diskriminanzanalyse handelt es sicg um ein multivariates Verfahren zur Analyse von Gruppen- bzw. Klassenunterschieden. Durch diese Methode ist es möglich, $G \geq 2$ Gruppen unter Berücksichtigung von x_j Variablen zu untersuchen, und dabei zu ermitteln, in wie weit sich diese Gruppen unterscheiden. Der Unterschied zur Clusteranalyse liegt darin, dass es sich bei der Diskriminanzanalyse um kein *exploratives*, sondern um ein *konfirmatorisches* Verfahren handelt. Es werden keine Gruppen gebildet, sondern es werden vorhandene Gruppierung hinsichtlich ihrer Gruppierungsqualität überprüft. Die abhängige Variable, die die Gruppenzugehörigkeit festlegt ist nominal, die unabhängigen Variablen sind metrisch.

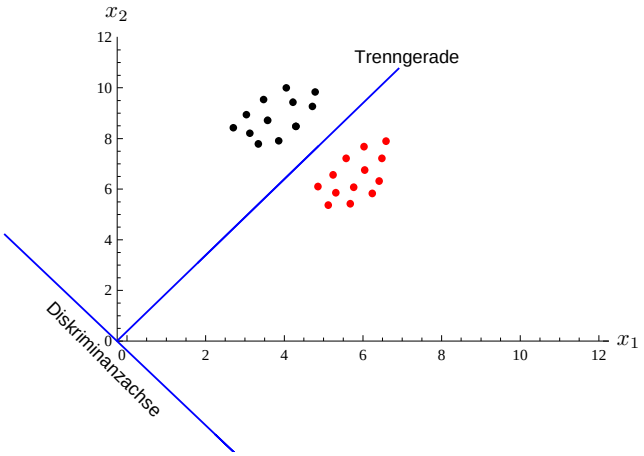


Abbildung 4.1: Diskriminanzfunktion

Durch die Diskriminanzanalyse kann geprüft werden, ob das Ergebnis einer Clusteranalyse verbesserungsfähig ist, welche Variablen für die Gruppierung besonders erklärungskräftig sind, oder in welche Gruppe ein neues Objekt eingeordnet werden sollte.

Der einfachste Fall besteht darin, wenn nur 2 Gruppen vorliegen - bsp. die Vorhersage der Zuordnung von Personen entweder zur Gruppe der SPD-Wähler oder der Gruppe der CDU/CSU-Wähler. Diese 2 Gruppen werden dann durch die sogenannte *Diskriminanzfunktion*, in die mehrere unabhängige Variablen eingehen können, getrennt. Sollen mehr Gruppen getrennt werden, so benötigen wir auch mehrere Diskriminanzfunktionen. Bei beispielsweise 3 Gruppen benötigen wir 2 Diskriminanzfunktionen.
Die Diskriminanzfunktion ähnelt derjenigen Gleichung der Regression.

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_j x_j$$

SPSS beispielsweise benötigt zur Durchführung einer Diskriminanzanalyse einen *a priori*-Wert. Als Voreinstellungen können *aus der Gruppengröße berechnen* oder *alle Gruppen gleich* gewählt werden. Möchte man andere a priori Wahrscheinlichkeiten verwenden, beispielsweise aus der amtlichen Statistik oder vorherigen Untersuchungsergebnissen, so kann dies nur über die Syntax über den Unterbefehl /PRIORS = X,Y,Z realisiert werden.
Für jeden Fall wird ein Diskriminanzwert berechnet. Mittels der Diskriminanzwerte kann jedes Objekt einer Gruppe zugeordnet werden. Die Werte der Diskriminanzfunktion sind metrisch, stellen also noch keine Gruppenzugehörigkeiten dar.

4.1 Ansatz über Bayes-Theorem

Betrachten wir die abhängige Variable *Wahlabsicht*. Die Schätzung der Koeffizienten b_j soll den Anteil der durch die Gruppenzugehörigkeit erklärten Varianz maximieren Prinzipiell lässt sich keine Funktion finden, die eine eindeutige Zuordnung der Gruppen erlaubt. Tendenziell sind nach der berechneten nach b_j maximierten Diskriminanzfunktion Personen mit niedrigen Diskriminanzwerten SPD-Wähler. Es kann jedoch auch eine Person mit niedrigen Werten ein CDU/CSU-Wähler sein. Es ist folglich nicht eindeutig möglich zu sagen: ab dem Diskriminanzwert Y^* wählt eine Person immer die CDU/CSU. Diesen Trennwert kann man dennoch im 2-Gruppenfall zur Trennnung benutzen. Sind mehr als 2 Gruppenzugehörigkeiten zu schätzen und in der Folge mehr Diskriminanzfunktion, erfolgt die Prognose der Gruppenzugehörikeit mittels Bayes-Statistik. Für jede Person lässt sich die Wahrscheinlichkeit $P(G_i|Y_j)$, bestimmen, dass sie bei gegebenem Diskriminanzwert in Gruppe G_i gehört.

4.1.1 Klassifikation der Fälle

Für alle Gruppen werden für jede Person die Wahrscheinlichkeiten berechnet, in diese Gruppen zu gehören. Über alle Gruppen hinweg ergibt sich für jedes Objekt eine Gesamtwahrscheinlichkeit von 1. Bei der Berechnung dieser Wahrscheinlichkeiten, kommt der Satz von Bayes zum Einsatz:

$$P(G_i|Y_j) = \frac{P(Y_j|G_i) \cdot P(G_i)}{\sum_{i=1}^{G} P(Y_j|G_i) \cdot P(G_j)}$$

Mittels dieses Satzes können wir die Wahrscheinlichkeit für die Gruppenzugehörigkeit aus zwei bekannten Wahrscheinlichkeiten auf Grund des Diskriminanzwertes berechnen. Man bezeichnet $P(G_i|Y_j)$ als **a posteriori- Wahrscheinlichkeit**. Diese wird auf der Basis der bedingten Wahrscheinlichkeit $P(Y_j|G_i)$ und der **a priori-Wahrscheinlichkeit** $P(G_i)$ berechnet.
Die a priori - Wahrscheinlichkeit $P(G_i)$ ist die theoretische Wahrscheinlichkeit in eine der beiden Gruppen zu fallen. Sie ist oftmals nicht bekannt, und muss vom Forscher gewählt werden. Hat man gar keine Informationen über die Verteilung in der Grundgesamtheit (z.B. aus der amtlichen Statistik), wird man eine Gleichverteilung zugrundelegen. Wir gehen davon, dass unsere relativen Häufigkeiten denen der Grundgesamtheit entsprechen und nutzen sie als a priori Eingaben.
Mit dem Vergleich der Gruppenzugehörigkeits-Wahrscheinlichkeit lässt sich auch ein Rückschluss auf die Güte der Zuweisung ziehen. Sind die Wahrscheinlichkeiten sehr unterschiedlich, ist die Einordnung eindeutig. Betrachtung der Gruppenmittelwerte bzw. Centroide, die auf der Basis der Zuweisung von Diskriminanzwerten errechnet werden. Je näher die Centroide zusammenliegen, desto schwieriger wird die Zuweisung zu einer Gruppe. Diese ist jedoch noch keinen Test darauf, ob die Unterschiede in den Gruppencentroiden auch in der Grundgesamtheit gelten.
Der Eigenwert entspricht nahezu dem F-Wert der Varianzanalyse, seine Berechnung

$$\lambda = \frac{QS_{zwischen}}{QS_{innerhalb}} = \frac{\text{erklärte Streuung}}{\text{nicht erklärte Streuung}}$$

unterscheidet sich von der des F-Wertes der Varianzanalyse dadurch, dass hier die Freiheitsgrade nicht einfließen. Je größer der Eigenwert, desto größer die durch die Diskriminanzfunktion erklärte Streuung. Werte innerhalb der Gruppen sind sich ähnlich, Werte zwischen den Gruppen unterscheiden sich deutlich. Die Diskriminanzfunktion soll die Varianz zwischen den Gruppen maximieren. Die Maximierung erfolgt über die Gewichtungsfaktoren b_j der Variablen x_j.

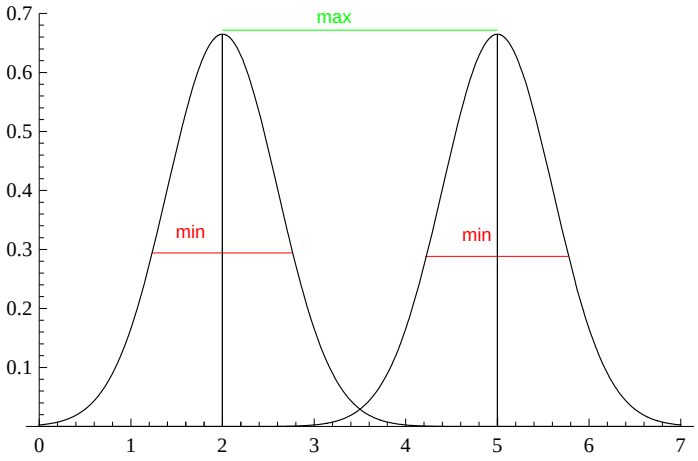


Abbildung 4.2: Verhältnis der Varianzen

Der Eigenwert berechnet sich über:

$$\lambda = \frac{(b_1(\bar{x}_{1,A_1} - \bar{x}_{1,A_2}) + b_2(\bar{x}_{2,A_1} - \bar{x}_{2,A_2}))^2}{b_1^2 s_{11} + b_2^2 s_{22} + 2b_1 b_2 s_{12}} = max!$$

Das Maximum für λ erhält man nach

$$\frac{\partial \lambda}{\partial b_j} = 0 \text{ und } \frac{\partial^2 \lambda}{\partial b_j^2} < 0$$

Die kanonische Korrelation

$$\sqrt{\frac{QS_{zwischen}}{QS_{zwischen} + QS_{innerhalb}}}$$

misst den Zusammenhang zwischen abhängigen und erklärenden Variablen und entspricht in der Definition dem η der Varianzanalyse: erklärter Anteil zu Gesamtstreuung.

Wilk's Λ

Wilk's Λ ist konträr zur kanonischen Korrelation definiert als

$$\Lambda = \frac{QS_{zwischen}}{QS_{zwischen} + QS_{innerhalb}}$$

und addiert sich somit mit dem Quadrat des kanonischen Korrelationskoeffizienten zu eins auf. Es wird eher zur Überprüfung der Modellgüte herangezogen, da hier mittels χ^2-Transformation von Λ ein Signifikanztest durchgeführt werden kann: H_0: Die Diskriminanzwerte sind in der Population zwischen den Gruppen identisch.

4.2 Mehrfache Diskriminanzanalyse

Bei g Gruppen ($l = 1, ..., g$) der abhängigen Variablen werden durch $k - 1$ Diskriminanzfunktionen getrennt, die nacheinander berechnet werden.
In Abbildung4.3 werden 3 Gruppen durch 2 Diskriminanzachsen getrennt.

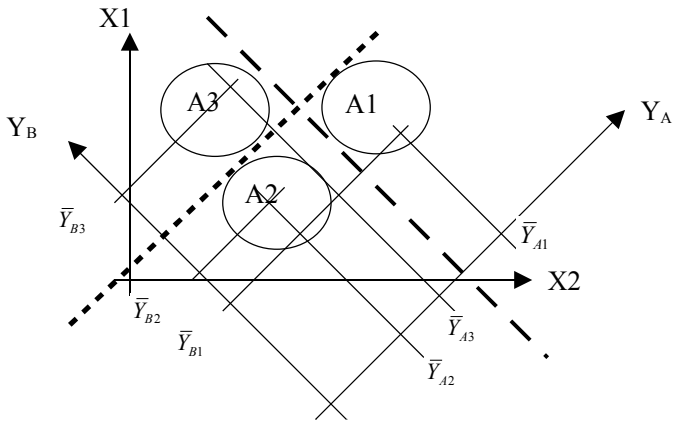


Abbildung 4.3: 3 Gruppen-Fall

Die erste Diskriminanzfunktion

$$Y_A = b_{1A}X_1 + b_{2A}X_2$$

trennt Gruppe $A1$ gegen die Gruppen $A3$ und $A2$, Y_A trennt nicht zwischen $A3$ und $A2$. Die zweite Diskriminanzfunktion

$$Y_B = b_{1B}X_1 + b_{2B}X_2$$

steht orthogonal zu Y_A und trennt Gruppe $A3$ gegen Gruppen $A2$ und $A1$. Y_A trennt nicht zwischen $A2$ und $A1$. Wir benötigen beide Diskriminanzfunktionen zur Trennung der Gruppen.

4.2.1 Prozedere

Bildung der ersten Diskriminanzfunktion als Linearkombination der Variablen $x_1, \ldots, x_k$ mit den Gewichten b_{jA}

$$Y_A = b_{1A}X_1 + b_{2A}X_2 + \ldots + b_{kA}X_k$$

Wie im 2-Gruppenfall wird auch hier der Wert für die Gewichte b_j gesucht der die Funktion maximiert.

$$\lambda_A = \frac{QS_{zwischen}}{QS_{innerhalb}} = max!$$

Die zweite Diskriminanzfunktion ist die Linearkombination der Partialvariablen erster Ordnung x_{j-A}:

$$Y_B = b_{1-A,B}X_{1-A} + \ldots + b_{k-A,B}X_{k-A}$$

Durch Transformation der Gewichte lässt sich die Funktion jedoch als Linearkombination der Ursprungsvariablen schreiben

$$Y_B = b_{1B}X_1 + b_{2B}X_2 + \ldots + b_{kB}X_k$$

und wie Y_A wird sie nach b_j maximiert. Alle weiteren Diskrimanzfunktionen werden als Linearkombinationen von Partialvariablen höherer Ordnung gebildet $x_{j-A,B,\ldots}$.
Die Klassifikation nach dem Satz von Bayes ist ab dem Fall mehrerer Gruppen erst nachvollziehbar. Wir können jetzt nicht mehr einfach mit einem Trennindex arbeiten, wie im Zweigruppenfall. Ab jetzt ist die Klassifizierung einfacher mittels anderer Verfahren.

$$P(G_I|Y_{ji}) = \frac{P(Y_{ji}|G_I) \cdot P(G_I)}{\sum_{I=2}^{g} P(Y_{ji}|G_I) \cdot P(G_I)}$$

Die Wahrscheinlichkeit wird für Person i für alle Gruppen mit allen Diskrimanzwerten berechnet. Die Zuordnung zu einer Gruppe erfolgt nach dem Maximalwert von:

$$\sum_{J=A}^{K} P(G_I|Y_{Ji})$$

4.3 Varianzzerlegung

Ähnlich der Varianzanalyse, lässt sich bei der Diskriminanzanalyse eine Varianzzerlegung durchführen

Gesamte Abweichung	=	Erklärte Abweichung	+	Nicht erklärte Abweichung
Summe der quadrierten Gesamtabweichung	=	Summe der quadrierten Abweichungen innerhalb der Faktorstufen	+	Summe der quadrierten Abweichungen zwischen den Faktorstufen
$\sum_{g=1}^{G}\sum_{i=1}^{I_g}(y_{gi} - \bar{y})^2$	=	$\sum_{g=1}^{G} I_g(\bar{y}_g - \bar{y})^2$	+	$\sum_{g=1}^{G}\sum_{i=1}^{I_g}(y_{gi} - \bar{y}_g)^2$
SS_t	=	SS_b	+	SS_w
$SS_{t(otal)}$	=	$SS_{b(etween)}$	+	$SS_{w(ithin)}$

Die Schätzung der Koeffizienten b_j soll den Anteil der durch die Gruppenzugehörigkeit erklärten Varianz maximieren. Kleine Überschneidungsbereiche der Häufigkeitsverteilungen auf der Diskriminanzachse bedeutet gute Trennung.
Jede Gruppe besitzt einen Centroid (mittleren Diskriminanzwert):

$$\bar{y}_g = \frac{\sum_{i=1}^{I_g} y_{gi}}{I_g}$$

Ein Maß für die Unterschiedlichkeit zweier Gruppen Gruppen ist beispielsweise

$$|\bar{y}_A - \bar{y}_B|$$

Die Parameter der Diskriminanzfunktion sollen nun so geschätzt werden, daß sich die Gruppen maximal unterscheiden. $|\bar{y}_A - \bar{y}_B|$ ist als Maß aber ungeeignet, da es die Streuung der Gruppen nicht berücksichtigt.
Wenn

1. nur 2 Gruppen vorliegen
2. die annähernd gleich groß sind
3. mit ungefähr gleicher Streuung s

dann ist

$$U = \frac{|\bar{y}_A - \bar{y}_B|}{s}$$

ein geeigneteres Diskriminaznmaß. Dazu ist äquivalent:

$$U^2 = \frac{(\bar{y}_A - \bar{y}_B)^2}{s^2}$$

Um die Voraussetzungen 1. und 2. aufzuheben, muß $(\bar{y}_A - \bar{y}_B)^2$ in der obigen Formel durch ein Maß für die Streuung zwischen den Gruppen ersetzt werden. Dies geschieht durch

$$SS_b = \sum_{g=1}^{G} I_g(\bar{y}_g - \bar{y})^2$$

wobei

G	=	Anzahl der Gruppen
I_g	=	Anzahl der elemente in Gruppe g
$\bar{y}g$	=	Mittlerer Diskriminanzwert in Gruppe g
$\bar{y}$	=	Gesamtmittel der Diskriminanzwerte aller Elemente

Um die Vorausssetzung 3.aufzuheben, muß s^2 in der Formel für U^2 durch ein Maß für die gesamte (gepoolte) Streuung innerhalb der zwei oder mehr Gruppen ersetzt werden. Ein Maß dafür ist:

$$SS_w = \sum_{g=1}^{G} \sum_{i=1}^{I_g} (y_{gi} - \bar{y}_g)^2$$

wobei y_{gi} den Diskriminanzwert von Element i in Gruppe g bezeichnet.

Das Diskriminanzkriterium, dass bei 2 oder mehr Gruppen verwendet wird, um die Unterschiedlichkeit der Gruppen zu messen ist:

$$\Gamma = \frac{SS_b}{SS_w} = \frac{\text{Streuung zwischen den Gruppen}}{\text{Streuung in den Gruppen}}$$

$$\Gamma = \frac{\sum_{g=1}^{G} I_g(\bar{y}_g - \bar{y})^2}{\sum_{g=1}^{G}\sum_{k=1}^{K}(y_{gk} - \bar{y}_g)^2}$$

Γ kann als Quotient aus **erklärter** zu **nicht erklärter** streuung interpretiert werden, also

$$\Gamma = \frac{\text{erklärte Streuung}}{\text{nicht erklärte Streuung}}$$

Die Diskriminanzwerte und damit auch Γ sind abhängig von den zu schätzenden Koeffizienten. Das Problem der Schätzung der Diskriminanzfunktion läßt sich nun so formulieren: Wähle die Koeffizienten $b_0,\ b_1,\ \ldots,\ b_j$ so, daß Γ maximal wird

4.4 Schätzen der Diskriminanzfunktion

Hierfür benötigen wir 2 Matrizen, die Matrix der Streuung *zwischen den Gruppen*, die Between-Matrix:

$$\boldsymbol{B} = \sum_{i=1}^{g} I_g(\bar{y}_g - \bar{y})(\bar{y}_g - \bar{y})'$$

sowie die Matrix der Streuung *innerhalb der Gruppen*, die Within-Matrix:

$$\boldsymbol{W} = \sum_{k=1}^{g} \boldsymbol{W}_k$$

$$\boldsymbol{W} = \sum_{g=1}^{g}\sum_{i=1}^{I_g}(\bar{y}_{gi} - \bar{y}_g)(\bar{y}_{gi} - \bar{y}_g)'$$

Das Diskriminanzkriterium Γ soll maximiert werden. Man kann Γ kann mit Hilfe der Matrizen $\boldsymbol{W}$ und $\boldsymbol{B}$ schreiben:

$$\Gamma = \frac{\boldsymbol{\vartheta' B \vartheta}}{\boldsymbol{\vartheta' W \vartheta}}$$

wobei ϑ der Vektor der unbekannten Parameter der Diskriminanzfunktion ist. Um Γ zu maximieren, muss die erste Ableitung nullgesetzt werden (sowie die zweite Ableitung an diesem Punkt negativ sein). Dies ergibt folgende Bedingung für den Wert λ:

$$(\boldsymbol{B} - \lambda \boldsymbol{W})\boldsymbol{\vartheta} = 0$$

Was zu

$$\boldsymbol{B\vartheta} = \lambda \boldsymbol{W\vartheta}$$

führt. Durch invertieren der Matrix $\boldsymbol{W}$ erhalten wir

$$(\boldsymbol{W^{-1}B})\boldsymbol{\vartheta} = \lambda\boldsymbol{\vartheta}$$

Es handelt sich bei λ um den maximalen Eigenwert der Matrix $\boldsymbol{W^{-1}B}$, $\boldsymbol{\vartheta}$ ist der dazugehörige Eigenvektor.
Dies führt zu der nicht-normierten Diskriminanzfunktion:

$$y = \vartheta_1 x_1 + \vartheta_2 x_2 + \ldots + \vartheta_j x_j$$

Die normierte Variante lautet:

$$b_j = \frac{1}{s}\vartheta_j \text{ mit } s = \sqrt{\frac{1}{I-G}\boldsymbol{\vartheta' W \vartheta}}$$

und $b_0 = -\sum_{j=1}^{J} b_j \bar{x}_j$

4.5 Güte der Diskriminanz

Wir wissen, dass für den maximalen λ des Diskriminanzkriteriums Γ gilt:

$$\lambda = \frac{\text{erklärte Streuung}}{\text{nicht erklärte Streuung}}$$

Dieses Unterschiedlichkeitsmaß ist allerding nicht auf den Bereich $0 \leq \lambda \leq 1$. Folgende Werte sind auf diesen Bereich genormt:

$$\frac{\lambda}{1+\lambda} = \frac{\text{erklärte Streuung}}{\text{gesamte Streuung}}$$

sowie

$$\frac{1}{1+\lambda} = \frac{\text{nicht erklärte Streuung}}{\text{gesamte Streuung}}$$

welcher *Wilk's Lambda* genannt wird.
Wilk's Λ ist ein sogenanntes "inverses Gütemaß", d.h. kleinere Werte bedeuten höhere Unterschiedlichkeit der Gruppen, bzw. höhere Trennkraft der Diskriminanzfunktion.

4.5.1 Signifikanz der Diskriminanzfunktion

Es werden folgende Hypothesen getetstet:

H_0 : Gruppen unterscheiden sich nicht
H_1 : Gruppen unterscheiden sich

Aus Wilk's Λ lässt sich eine Prüfgröße berechnen, die annähernd χ^2-verteilt ist mit $df = J \cdot (G-1)$, mit der die Hypothesen geprüft werden können.

$$\chi^2 = -\left[N - \frac{J+G}{2} - 1\right] \ln \Lambda$$

wobei
N : Gesamtzahl der Fälle
J : Anzahl der Merkmalsvariablen
G : Anzahl der Gruppen
Wenn mehr als zwei Gruppen vorhanden sind, so wird mehr als eine Diskriminanzfunktion verwendet. Bei G Gruppen: höchstens $G-1$ Diskriminanzfunktionen. Aber: nicht mehr Diskriminanzfunktionen als Merkmalsvariablen. Zu jeder Diskriminanzfunktion gehört ein Eigenwert. Für diese Eigenwerte gilt:

$$\lambda_1 > \lambda_2 > \ldots > \lambda_{g-1} >$$

Die zweite Diskriminanzfunktion wird so ermittelt, daß sie einen maximalen Anteil jener Streuung erklärt, die nach Ermittlung der ersten Diskriminanzfunktion als Rest verbleibt.
Ein Maß für die relative Wichtigkeit der $k$ten Diskriminanzfunktion ist der sogenannte Eigenwertanteil:

$$EA_k = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_K}$$

Er gibt die durch die $k-$te Diskriminanzfunktion erklärte Streuung als Anteil jener Streuung an, die durch alle K Diskriminanzfunktionen erklärt wird. Die Wichtigkeit der Diskriminanzfunktionen nimmt schnell ab. Meist genügen 23 Diskriminanzfunktionen.
Zum Prüfen der Unterschiedlichkeit der Gruppen müssen alle Diskriminanzfunktionen und deren Eigenwerte berücksichtigt werden. Man verwendet dazu das multivariate Wilks Λ:

$$\Lambda = \prod_{k=1}^{K} \frac{1}{1+\lambda_k} = \frac{1}{1+\lambda_1} \cdot \frac{1}{1+\lambda_2} \cdot \ldots \cdot \frac{1}{1+\lambda_K}$$

Wobei λ_k der Eigenwert der k-ten Diskriminanzfunktion ist. Das multivariate Wilks Λ ergibt sich als Produkt der univariaten. Es kann wiederum die χ^2-Prüfgröße gebildet werden, um auf Signifikanz zu testen.
Wilks Λ kann einem in unterschiedlichen Schreibweisen begegnen:

$$\Lambda = \frac{\det(\boldsymbol{W})}{\det(\boldsymbol{T})} = \frac{\det(\boldsymbol{W})}{\det(\boldsymbol{B}+\boldsymbol{W})} = \det(\boldsymbol{I} + \boldsymbol{W^{-1}B})^{-1} = \prod_{k=1}^{q} (1+\lambda_k)^{-1}$$

ANHANG A

MATRIX-ALGEBRA

Eine Matrix besteht aus m Zeilen und n Spalten. Matrizen werden mit fetten Großbuchstaben bezeichnet. Eine $m \times n$-Matrix sieht folgendermaßen aus:

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix}$$

A.1 Skalarmultiplikation

Um eine Matrix mit einem Skalar zu multiplizieren muss jedes Element der Matrix mit diesem Skalar multipliziert werden.

$$\boldsymbol{A} \cdot \varphi = \varphi \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} \varphi a_{11} & \varphi a_{12} & \varphi a_{13} \\ \varphi a_{21} & \varphi a_{22} & \varphi a_{23} \\ \varphi a_{31} & \varphi a_{32} & \varphi a_{33} \end{pmatrix}$$

Rechenregeln für Skalarmultiplikation

$$\begin{aligned} (\alpha + \beta)\boldsymbol{A} &= \alpha\boldsymbol{A} + \beta\boldsymbol{A} \\ \alpha(\boldsymbol{A} + \boldsymbol{B}) &= \alpha\boldsymbol{A} + \alpha\boldsymbol{B} \end{aligned}$$

$$\alpha(\beta\boldsymbol{A}) = (\alpha\beta)\boldsymbol{A} = (\beta\alpha)\boldsymbol{A} = \beta(\alpha\boldsymbol{A})$$
$$\alpha(\boldsymbol{A}\boldsymbol{B}) = (\alpha\boldsymbol{A})\boldsymbol{B} = \boldsymbol{A}(\alpha\boldsymbol{B}) = \alpha\boldsymbol{A}\boldsymbol{B}$$

A.2 Multiplikation

In der Rechnung mit Matrizen ist Einiges zu beachten, was uns auf den ersten Blick unlogisch oder verwirrend erscheint. So ist jedem von uns aus dem Alltag bekannt, dass 3 mal 5 identisch ist mit 5 mal 3. Dies gilt in der Matrizenrechnung nicht, oder nur für ganz bestimmte Matrizen $\boldsymbol{A}$ und $\boldsymbol{B}$, nämlich für idempotente Matrizen. Im "Normalfall" gilt:

$$\boldsymbol{A} \cdot \boldsymbol{B} \neq \boldsymbol{B} \cdot \boldsymbol{A}$$

Berechnen wir $\boldsymbol{A} \cdot \boldsymbol{B}$ mit

$$\boldsymbol{A} = \begin{pmatrix} 2 & -3 & 1 \\ -1 & 4 & 0 \end{pmatrix}; \boldsymbol{B} = \begin{pmatrix} 3 & 1 \\ 4 & 2 \\ 5 & -3 \end{pmatrix}$$

erhalten wir folgendes:

$$\begin{pmatrix} 2 & -3 & 1 \\ -1 & 4 & 0 \end{pmatrix} \cdot \begin{pmatrix} 3 & 1 \\ 4 & 2 \\ 5 & -3 \end{pmatrix} = \begin{pmatrix} -1 & -7 \\ 13 & 7 \end{pmatrix}$$

Berechnen wir nun $\boldsymbol{B} \cdot \boldsymbol{A}$ erhalten wir ein anderes Ergebnis:

$$\begin{pmatrix} 3 & 1 \\ 4 & 2 \\ 5 & -3 \end{pmatrix} \cdot \begin{pmatrix} 2 & -3 & 1 \\ -1 & 4 & 0 \end{pmatrix} = \begin{pmatrix} 5 & -5 & 3 \\ 6 & -4 & 4 \\ 13 & -27 & 5 \end{pmatrix}$$

Matrixmultiplikation ist also nicht kommutativ! Wann können wir überhaupt zwei Matrizen miteinander *multiplizieren*? Dies ist nicht immer möglich. Man kann zwei Matrizen nur multiplizieren, wenn die Anzahl der Spalten der ersten Matrix identisch ist mit der Anzahl der Zeilen der zweiten Matrix. Wir können als eine 4×5-Matrix mit einer 5×6-Matrix multiplizieren und erhalten so eine 5×5-Matrix als Ergebnis. Es ist jedoch nicht möglich, eine 5×6-Matrix mit einer 4×5-Matrix zu multiplizieren.

Matrixmultiplikation								
3×4	$\cdot$	4×5	$=$	3×5	4×3	$\cdot$	5×4	$=$	/
5×6	$\cdot$	6×5	$=$	5×5	6×5	$\cdot$	5×6	$=$	6×6
7×5	$\cdot$	5×3	$=$	7×3	5×7	$\cdot$	3×5	$=$	/

Als Beispiel multiplizieren wir nun die schon angesprochenen Matrizen $\boldsymbol{A}$ und $\boldsymbol{B}$, um zu sehen, wie wir an die Werte kommen, die in der resultierenden Matrix $\boldsymbol{C}$ stehen. Wir berechnen:

$$\boldsymbol{A} \cdot \boldsymbol{B} = \boldsymbol{C}$$

also

$$\begin{pmatrix} 2 & -3 & 1 \\ -1 & 4 & 0 \end{pmatrix} \cdot \begin{pmatrix} 3 & 1 \\ 4 & 2 \\ 5 & -3 \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

Welchen Wert hat nun beispielsweise der Eintrag c_{11}?

$$\begin{pmatrix} 2 & -3 & 1 \\ -1 & 4 & 0 \end{pmatrix} \cdot \begin{pmatrix} 3 & 1 \\ 4 & 2 \\ 5 & -3 \end{pmatrix}$$

Der Wert von c_{11} ergibt sich durch

$$c_{11} = 2 \cdot 3 + (-3) \cdot 4 + 1 \cdot 5 = -1$$

Es wird also das erste Element der ersten Zeile der Matrix $\boldsymbol{A}$ mit dem ersten Element der ersten Spalte der Matrix $\boldsymbol{B}$ multipliziert. Dann wird das zweite Element der ersten Zeile der Matrix $\boldsymbol{A}$ mit dem zweiten Element der ersten Spalte der $\boldsymbol{B}$ multipliziert und dieses Ergebnis zu dem vorherigen Ergebnis addiert. Zu guter Letzt wird das Ergebnis der Multiplikation des dritten Elements der ersten Zeile der Matrix $\boldsymbol{A}$ mit dem dritten Element der ersten Spalte der Matrix $\boldsymbol{B}$ zu den beiden vorherigen addiert. Wir erhalten folgende 4 Werte:

$$\begin{array}{rcrcrcrcr} c_{11} & = & 2 \cdot 3 & + & (-3) \cdot 4 & + & 1 \cdot 5 & = & -1 \\ c_{12} & = & 2 \cdot 1 & + & (-3) \cdot 2 & + & 1 \cdot (-3) & = & -7 \\ c_{21} & = & (-1) \cdot 3 & + & 4 \cdot 4 & + & 0 \cdot 5 & = & 13 \\ c_{22} & = & (-1) \cdot 1 & + & 4 \cdot 2 & + & 0 \cdot (-3) & = & 7 \end{array}$$

Also erhalten wir als Ergebnis der Multiplikation $\boldsymbol{A} \cdot \boldsymbol{B}$ die Matrix

$$\boldsymbol{C} = \begin{pmatrix} -1 & -7 \\ 13 & 7 \end{pmatrix}$$

Rechenregeln zur Matrixmultiplikation

Assoziativgesetz	$(\boldsymbol{AB})\boldsymbol{C}$	$=$	$\boldsymbol{A}(\boldsymbol{BC})$
	$(\boldsymbol{AB})\boldsymbol{C}$	$=$	$\boldsymbol{ABC}$
	$\boldsymbol{A}(\boldsymbol{BC})$	$=$	$\boldsymbol{ABC}$
linksseitiges Distributivgesetz	$\boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C})$	$=$	$\boldsymbol{AB}+\boldsymbol{AC}$
rechtsseitiges Distributivgesetz	$(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{C}$	$=$	$\boldsymbol{AC}+\boldsymbol{BC}$
im Allgemeinen	$\boldsymbol{AB}$	$\neq$	$\boldsymbol{BA}$
	$\boldsymbol{AI}$	$=$	$\boldsymbol{IA} = \boldsymbol{A}$

A.3 Addition und Subtraktion

Addition und *Subtraktion* von Matizen sind einfacher zu bewerkstelligen als Multiplikation oder Division von Matrizen. Eine Einschränkung ist jedoch, dass nur Matrizen der gleichen Ordnung addiert oder subtrahiert werden können. Schauen wir uns nun die Addition zweier Matrizen an:

$$\begin{pmatrix} 3 & 1 \\ 5 & 2 \\ 2 & 4 \end{pmatrix} + \begin{pmatrix} 5 & 4 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 3+5 & 1+4 \\ 5+1 & 2+2 \\ 2+1 & 4+3 \end{pmatrix} = \begin{pmatrix} 8 & 5 \\ 6 & 4 \\ 3 & 7 \end{pmatrix}$$

die Subtraktion zweier Matrizen verläuft analog. Auch hier ein Beispiel:

$$\begin{pmatrix} 3 & 1 \\ 5 & 2 \\ 2 & 4 \end{pmatrix} + \begin{pmatrix} 5 & 4 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 3-5 & 1-4 \\ 5-1 & 2-2 \\ 2-1 & 4-3 \end{pmatrix} = \begin{pmatrix} -2 & -3 \\ 4 & 0 \\ 1 & 1 \end{pmatrix}$$

Rechenregeln Addition und Subtraktion

$$\begin{aligned} (\boldsymbol{A}+\boldsymbol{B})+\boldsymbol{C} &= \boldsymbol{A}+(\boldsymbol{B}+\boldsymbol{C}) \\ \boldsymbol{A}+\boldsymbol{B} &= \boldsymbol{B}+\boldsymbol{A} \\ \boldsymbol{A}+\boldsymbol{0} &= \boldsymbol{A} \\ \boldsymbol{A}+(-\boldsymbol{A}) &= \boldsymbol{0} \end{aligned}$$

Wobei $\boldsymbol{A}$, $\boldsymbol{B}$ und $\boldsymbol{C} = n \times m$-Matrizen und $\boldsymbol{0}$ die $n \times m$-Nullmatrix

A.4 Transponieren

Als *transponieren* einer Matrix wird der Vorgang bezeichnet, durch den die Zeilen einer Matrix zu Spalten werden und Spalten zu Zeilen. Technischer gesprochen: Aus dem Eintrag a_{ij} wird der a_{ji}. Die transponierte Matrix von $\boldsymbol{A}$ wird mit $\boldsymbol{A'}$ oder auch $\boldsymbol{A^t}$ bezeichnet.

$$\boldsymbol{A} = \begin{pmatrix} a & b & c \\ x & y & z \end{pmatrix} \overset{transponieren}{\rightarrow} \boldsymbol{A'} bzw. \boldsymbol{A^t} = \begin{pmatrix} a & x \\ b & y \\ c & z \end{pmatrix}$$

Transponieren und multiplizieren wirken sich kombiniert folgendermaßen aus:

- $\boldsymbol{A} \times \boldsymbol{B}$ = Multiplikation Reihe mal Spalte von $\boldsymbol{A}$ und $\boldsymbol{B}$
- $\boldsymbol{A} \times \boldsymbol{B'}$ = Multiplikation Reihe mal Reihe von $\boldsymbol{A}$ und $\boldsymbol{B}$
- $\boldsymbol{A'} \times \boldsymbol{B}$ = Multiplikation Spalte mal Spalte von $\boldsymbol{A}$ und $\boldsymbol{B}$
- $\boldsymbol{A'} \times \boldsymbol{B'}$ = Multiplikation Spalte mal Reihe von $\boldsymbol{A}$ und $\boldsymbol{B}$

Rechenregeln für Transponierte

$$\begin{aligned} (\boldsymbol{A}')' &= \boldsymbol{A} \\ (\boldsymbol{A}+\boldsymbol{B})' &= \boldsymbol{A}'+\boldsymbol{B}' \\ (\alpha\boldsymbol{A})' &= \alpha\boldsymbol{A}' \\ (\boldsymbol{AB})' &= \boldsymbol{B}'\boldsymbol{A}' \\ (\boldsymbol{ABC})' &= \boldsymbol{C}'\boldsymbol{B}'\boldsymbol{A}' \end{aligned}$$

A.5 Diagonalmatrizen

Bei einer *symmetrischen* Matrix handelt es sich um einen Sonderfall einer quadratischen Matrix ($n \times n$-Matrix). Hierbei sind die Matrix und ihre Transponierte identisch, es gilt: $\boldsymbol{A} = \boldsymbol{A'}$. Eintrag a_{ij} und a_{ji} sind identisch, also a_{ij}=a_{ji}.
Bei einer *Diagonalmatrix* handelt es sich um eine Matrix, in der alle Wert, bis auf die der *Hauptdiagonalen* gleich Null sind. Es gilt also:

$$\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}$$

Bei einer Skalarmatrix handelt es sich um eine Diagonalmatrix, die durch den Skalar gebildet wird. Sie steht für den Skalar und unmegekehrt. Es gilt:

$$\boldsymbol{\omega} = \begin{pmatrix} \omega & 0 & 0 \\ 0 & \omega & 0 \\ 0 & 0 & \omega \end{pmatrix} = \omega$$

$$\boldsymbol{\omega A} = \boldsymbol{A\omega} = \omega \boldsymbol{A} = \boldsymbol{A}\omega$$

Ein Spezialfall einer Diagonalmatrix ist die sogenannte Einheits- oder Identitätsmatrix, abgekürzt mit $\boldsymbol{E}$ oder $\boldsymbol{I}$. In ihr stehen nur Einsen auf der Hauptdiagonalen.

$$\boldsymbol{E} \text{ bzw. } \boldsymbol{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

A.6 Die Spur einer Matrix

Als Spur einer Matrix $\mathrm{sp}(\boldsymbol{A})$ oder $\mathrm{tr}(\boldsymbol{A})$, Abkürzung des englischen "trace" für "Spur") bezeichnet man die Summe der Elemente der Hauptdiagonalen. Sie ist definiert als $\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii}$.

In der Matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

beträgt die Spur $1 + 1 + 1 = 3$

In der Matrix

$$\begin{pmatrix} 5 & -5 & 3 \\ 2 & 3 & 4 \\ 6 & -7 & 4 \end{pmatrix}$$

beträgt die Spur $5 + 3 + 4 = 12$

Rechenregeln für die Spur einer Matrix

$$\begin{array}{rcl|rcl} \mathrm{tr}(\boldsymbol{A}+\boldsymbol{B}) & = & \mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{B}) & \mathrm{tr}(\boldsymbol{A}') & = & \mathrm{tr}(\boldsymbol{A}) \\ \mathrm{tr}(\alpha\boldsymbol{A}) & = & \alpha\mathrm{tr}(\boldsymbol{A}) & \mathrm{tr}(\boldsymbol{AB}) & = & \mathrm{tr}(\boldsymbol{BA}) \\ \mathrm{tr}(\boldsymbol{BCA}) = \mathrm{tr}(\boldsymbol{CAB}) & = & \mathrm{tr}(\boldsymbol{ABC}) & \mathrm{tr}(\boldsymbol{B}^{-1}\boldsymbol{AB}) & = & \mathrm{tr}(\boldsymbol{A}) \\ \text{(i. Allg.) } \mathrm{tr}(\boldsymbol{AB}) & \neq & \mathrm{tr}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{B}) & & & \end{array}$$

Für partionierte Matrizen

$$\mathrm{tr}\begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix} = \mathrm{tr}(\boldsymbol{A}_{11}) + \mathrm{tr}(\boldsymbol{A}_{22})$$

A.7 Determinante

Die Determinante einer 2×2-Matrix errechnet sich wie folgt:
für die Matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \rightarrow \det(\mathrm{A}) = a_{11}a_{22} - a_{12}a_{21}$$

Wie aber berechnet man die Determinante von höher dimensionierten Matrizen? Hier als Beispiel eine 3×3-Matrix:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

Es bestehen nun mehrere Möglichkeiten, die Determinante einer 3×3-Matrix zu berechnen (Laplace'scher Entwicklungssatz). Wir können die Determinante über die gewichtete Summe der Elemente einer Reihe oder Spalte (egal welcher) bestimmen.

- Für jedes Element der gewählten Spalte oder Zeile, hier a_{11}, a_{21} und a_{31}, wird ein Gewicht berechnet.
- Dieses Gewicht ist die Determinante einer 2×2-Matrix.
- Diese 2×2-Matrix erhält man, wenn man alle Elemente streicht, die in der gleichen Spalte sowie Zeile stehen, wie das Element, für das man das Gewicht berechnen will.

Also berechnen wir wie folgt für die Matrix $\boldsymbol{A}$:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} = a_{22}a_{33} - a_{23}a_{32} = \alpha$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{pmatrix} = a_{12}a_{33} - a_{13}a_{32} = \beta$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{pmatrix} = a_{12}a_{23} - a_{13}a_{22} = \gamma$$

.
Die Determinanten der Restmatrizen werden Kofaktoren der Einzelelemente genannt. Das Vorzeichen des Kofaktors erhält man, indem man den Spalten/Zeilen-Index des Einzelelements addiert. Bei einer geraden Summe des Indexes ergibt sich ein positives Vorzeichen, bei einer ungeraden Summe ein negatives Vorzeichen. Also:

$$\begin{array}{lllll} \alpha \rightarrow a_{11} = a_{\text{gerade}} & \text{da } 1+1=2, \; \textit{also}: & + \\ \beta \rightarrow a_{12} = a_{\text{ungerade}} & \text{da } 1+2=3, \; \textit{also}: & - \\ \gamma \rightarrow a_{13} = a_{\text{gerade}} & \text{da } 1+3=4, \; \textit{also}: & + \end{array}$$

Zusammengefasst ergibt sich die Determinante der 3×3-Matrix aus:

$$\mathrm{Det}(\boldsymbol{A}) = a_{11} \cdot \alpha - a_{12} \cdot \beta + a_{13} \cdot \gamma$$

Ebenso lässt sich die Determinante einer 4×4-Matrix berechnen, allerdings ist dies verschachtelter, und somit aufwendiger. Hier ist es erforderlich, aus der 4×4-Matrix auf analoge Weise erst vier 3×3-Matrizen zu extrahieren, um danach mit der eben dargestellten Methode aus diesen 3×3-Matrizen deren Determinanten zu errechnen. Also wird dieses Verfahren sehr schnell sehr aufwendig.
Eine alternative Berechnung der Determinante einer 3×3-Matrix funktioniert folgendermaßen (Regel von Sarrus): Die Spalten der Matrix

$$\boldsymbol{A} = \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix}$$

stellt man wie folgt angeordnet dar:

$$\begin{array}{ccccc} a_1 & b_1 & c_1 & a_1 & b_1 \\ a_2 & b_2 & c_2 & a_2 & b_2 \\ a_3 & b_3 & c_3 & a_3 & b_3 \end{array}$$

Nun werden Elemente nach einem bestimmten Muster multipliziert und addiert bzw. subtrahiert.

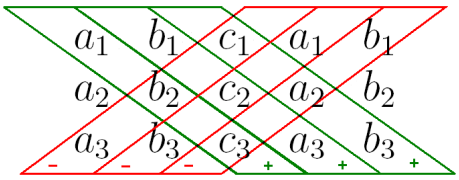


Abbildung A.1: Rechenschema

Wir rechnen:

$$\det(\boldsymbol{A}) = a_1 b_2 c_3 + a_3 b_1 c_2 + a_2 b_3 c_1 - a_3 b_2 c_1 - a_1 b_3 c_2 - a_2 b_1 c_3$$

Eine weitere Möglichkeit besteht darin, die Matrix in Stufenform zu bringen. Hierbei ist das Vertauschen von zwei Zeilen oder das Multiplizieren einer Zeile mit einer Zahl (z.B. mit (-1)) nun aber nicht erlaubt bzw. verändert den Wert der Determinante.
Beispiel:
Die Matrix $\boldsymbol{X}$

$$\boldsymbol{X} = \begin{pmatrix} 4 & 7 & 6 & 2 & 3 & 4 \\ 0 & 2 & 1 & 5 & 4 & 5 \\ 4 & 7 & 7 & 5 & 6 & 6 \\ 0 & 2 & 1 & 8 & 6 & 6 \\ 0 & 4 & 2 & 16 & 16 & 14 \\ 8 & 14 & 13 & 7 & 9 & 12 \end{pmatrix}$$

wird Schrittweise in Stufenform gebracht. Diese Matrix hat eine Determinante mit dem wert 192. Auf die ausführliche Darstellung der Berechnung der Determinante sowie der herstellung der triagonalisierten Matrix (Stufenform) wird an dieser Stelle verzichtet. Wir erhalten folgende Matrix als Ergebnis:

$$\begin{pmatrix} 4 & 7 & 6 & 2 & 3 & 4 \\ 0 & 2 & 1 & 5 & 4 & 5 \\ 0 & 0 & 1 & 3 & 3 & 2 \\ 0 & 0 & 0 & 3 & 2 & 1 \\ 0 & 0 & 0 & 0 & 4 & 2 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix}$$

Das Produkt der Hauptdiagonalen $4 \times 2 \times 1 \times 3 \times 4 \times 2 = 192$ ergibt wiederum $det(\boldsymbol{A}) = 192$.

Regeln für Determinanten

Für eine $n \times n$-Matrix $\boldsymbol{A}$ gilt:

1. Wenn alle Elemente in einer Zeile oder Spalte von $\boldsymbol{A}$ gleich 0, dann $\det(\boldsymbol{A}) = 0$
2. $\det(\boldsymbol{A}) = \det(\boldsymbol{A'})$
3. Wenn alle Elemente in einer Zeile oder Spalte von $\boldsymbol{A}$ mit φ multipliziert werden gilt $\varphi \det(\boldsymbol{A}) = 0$
4. Wenn zwei Spalten oder Zeilen von $\boldsymbol{A}$ vertauscht werden wechselt $\det(\boldsymbol{A})$ das Vorzeichen, der Absolutwert bleibt identisch.
5. Wenn zwei Zeilen oder Spalten von $\boldsymbol{A}$ proportional sind, dann $\det(\boldsymbol{A}) = 0$
6. $\det(\boldsymbol{A})$ bleibt unverändert, wenn das Vielfache einer Zeile oder Spalte zu einer anderen Zeile oder Spalte von $\boldsymbol{A}$ addiert wird.
7. Wenn $\boldsymbol{B}$ ebenfalls eine $n \times n$-Matrix ist, dann:
   $\det(\boldsymbol{AB}) = \det(\boldsymbol{A}) \cdot \det(\boldsymbol{B})$
8. Für $\varphi \in \mathbb{R}$ gilt $\det(\varphi\boldsymbol{A}) = \varphi^n \det(\boldsymbol{A})$

A.8 Adjunkte

Um die Adjunkte (abgekürzt mit: adj($\boldsymbol{A}$)) einer Matrix zu bestimmen müssen wir folgendes berechnen:

- Für jedes Matrixelement wird der Kofaktor bestimmt
- Jedes mit Element der Matrix wird durch seinen Kofaktor ersetzt
- Die Kofaktoren werden mit (+1) multipliziert, wenn die Indexsumme gerade ist, mit (-1), wenn die Indexsumme negativ ist.

- Danach wird die Matrix der Kofaktoren transponiert.

Beispiel:

$$\begin{pmatrix} 2 & 1 & 2 \\ 2 & 0 & 0 \\ 4 & 2 & 2 \end{pmatrix}$$

Wir ersetzen die ursprünglichen Elemente der Matrix durch die zugehörigen Kofaktoren. Dies geschieht durch die oben dargestellte Methode. Die Matrix der Kofaktoren sieht so aus:

$$\begin{pmatrix} 0 & 4 & 4 \\ -2 & -4 & 0 \\ 0 & -4 & -2 \end{pmatrix}$$

Nun wird für eine gerade Indexsumme mit (+1) multipliziert, mit (-1) bei ungerader Indexsumme. Also ergibt sich folgendes Schema:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \rightarrow \begin{pmatrix} \text{gerade} & \text{ungerade} & \text{gerade} \\ \text{ungerade} & \text{gerade} & \text{ungerade} \\ \text{gerade} & \text{ungerade} & \text{gerade} \end{pmatrix} \rightarrow \begin{pmatrix} + & - & + \\ - & + & - \\ + & - & + \end{pmatrix}$$

Also ergibt sich folgendes:

$$\begin{pmatrix} +(0) & -(4) & +(4) \\ -(-2) & +(-4) & -(0) \\ +(0) & -(-4) & +(-2) \end{pmatrix} \rightarrow \begin{pmatrix} 0 & -4 & 4 \\ 2 & -4 & 0 \\ 0 & 4 & -2 \end{pmatrix}$$

Schlussendlich transponieren wir diese Matrix und erhalten so die Adjunkte.

$$\begin{pmatrix} 0 & -4 & 4 \\ 2 & -4 & 0 \\ 0 & 4 & -2 \end{pmatrix} \overset{transponieren}{\rightarrow} \begin{pmatrix} 0 & 2 & 0 \\ -4 & -4 & 4 \\ 4 & 0 & -2 \end{pmatrix}$$

A.9 Inverse

Bleibt noch die *Inverse* einer Matrix, auch *Reziprokmatrix* genannt. Sie ist nur für quadratische Matrizen definiert. Die Inverse von $\boldsymbol{A}$ wird mit $\boldsymbol{A^{-1}}$ abgekürzt. Es gilt: $\boldsymbol{A \cdot A^{-1} = I}$. Wenn man sich die Zahlen, mit denen wir tagtäglich rechnen als eindimensionale, also 1×1-Matrizen vorstellt, dann ist die Inverse zu 7 1/7, oder anders geschrieben 7^{-1}, da $7 \cdot 1/7 = 1$ ist. 1 ist die Identitätsmatrix im eindimensionalen Raum. Sie besteht nur aus einem Eintrag, nämlich 1, da dieser einzige Eintrag gleichzeitig die gesammte Hauptdiagonale ist. Für höherdimensionale Räume wird die Berechnung der Inversen aufwendiger, sofern sie überhaupt existiert.

- Die Inverse einer Matrix existiert nur für *quadratische* Matrizen, da nur quadratische Matrizen eine Determinante haben, die zur Berechnung der Inversen notwendig ist. Vorsicht: nicht jede quadratische Matrix besitzt eine Inverse!

- Die Inverse existiert nur, wenn die Determinante der Matrix von Null verschieden ist. Solche Matrizen heissen regulär oder nicht-singulär. Quadratische Matrizen besitzen also nicht immer eine Inverse sondern *können* Inversen besitzen. Müssen sie aber nicht.
- Eine Matrix mit $\det(\boldsymbol{A})= 0$ heisst singulär.
- Eine $\det(\boldsymbol{A})=0$ resultiert dann, wenn man eine Zeile oder Spalte als Linearkombination einer oder mehrerer Spalten darstellen kann.

Berechnung der Inversen:

$$\boldsymbol{A}^{-1} = \frac{\mathrm{adj}(\boldsymbol{A})}{\det(\boldsymbol{A})}$$

Hier mag man sich noch einmal vor Augen führen, dass die Adjunkte einer Matrix wieder eine Matrix ist, die Determinante einer Matrix jedoch keine Matrix sondern ein Skalar. Ein Beispiel:

$$\boldsymbol{A} = \begin{pmatrix} 2 & 1 & 2 \\ 2 & 0 & 0 \\ 4 & 2 & 2 \end{pmatrix};\ \mathrm{adj}(\boldsymbol{A}) = \begin{pmatrix} 0 & 2 & 0 \\ -4 & -4 & 4 \\ 4 & 0 & -2 \end{pmatrix};\ \det(\boldsymbol{A}) = 4$$

Wir setzen diese Werte ein:

$$\boldsymbol{A}^{-1} = \frac{\mathrm{adj}(\boldsymbol{A})}{\det(\boldsymbol{A})} = \mathrm{adj}(\boldsymbol{A}) \cdot \det(\boldsymbol{A})^{-1}$$

$$\boldsymbol{A}^{-1} = \begin{pmatrix} 0 & 2 & 0 \\ -4 & -4 & 4 \\ 4 & 0 & -2 \end{pmatrix} \cdot 4^{-1} = \begin{pmatrix} 0 & 0.5 & 0 \\ -1 & -1 & 1 \\ 1 & 0 & -0.5 \end{pmatrix}$$

wir erinnern uns, dass nun gilt $\boldsymbol{A} \cdot \boldsymbol{A}^{-1} = \boldsymbol{I}$

$$\begin{pmatrix} 2 & 1 & 2 \\ 2 & 0 & 0 \\ 4 & 2 & 2 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0.5 & 0 \\ -1 & -1 & 1 \\ 1 & 0 & -0.5 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Wir sehen hier, warum nur Matrizen mit einer $\det(\boldsymbol{A}) \neq 0$ eine Inverse besitzen: ganz einfach deshalb, weil die Berechnung *möglich* ist. Matrizen mit einer $\det(\boldsymbol{A})=0$ stoßen bei der Berechnung der Inversen auf das altbekannte Problem einer Division durch Null, dem einen "großen Verbot" aus Schultagen neben dem Wurzelziehen aus negativen Zahlen. Es liegt also schlichtweg daran, das wir auf dem Rechenweg in einer Sackgasse enden.
Eine alternative Methode die Inverse einer Matrix auszurechnen funktioniert folgendermaßen: Wir schreiben links die Matrix, z.B. $\boldsymbol{A}$, die wir invertieren wollen und rechts die gleichdimensionierte Einheitsmatrix $\boldsymbol{I}$

$$(\boldsymbol{A}|\boldsymbol{I})$$

$$\left(\begin{array}{ccc|ccc} 2 & 1 & 2 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 & 1 & 0 \\ 4 & 2 & 2 & 0 & 0 & 1 \end{array}\right) \begin{array}{l} \text{Zeile I} \\ \text{Zeile II} \\ \text{Zeile III} \end{array}$$

Nun formen wir die linke Seite schrittweise so um, dass sie zur Einheitsmatrix wird. Dadurch verändert sich die Einheitsmatrix auf der rechten Seite so, dass sie zur Inversen wird. Resultiert auf der linken Seite eine komplette Nullzeile oder Nullspalte, so hat die Matrix $\boldsymbol{A}$ nicht vollen Rang, und sie ist nicht invertierbar. Als 1. Schritt subtrahieren wir Zeile I von der Zeile II

$$\left(\begin{array}{ccc|ccc} 2 & 1 & 2 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 & 1 & 0 \\ 4 & 2 & 2 & 0 & 0 & 1 \end{array}\right) \quad -I$$

Als 2. Schritt subtrahieren wir 2·I von III

$$\left(\begin{array}{ccc|ccc} 2 & 1 & 2 & 1 & 0 & 0 \\ 0 & -1 & -2 & -1 & 1 & 0 \\ 4 & 2 & 2 & 0 & 0 & 1 \end{array}\right) \quad -2 \cdot I$$

Als 3. Schritt multiplizieren wir III mit (-1)

$$\left(\begin{array}{ccc|ccc} 2 & 1 & 2 & 1 & 0 & 0 \\ 0 & -1 & -2 & -1 & 1 & 0 \\ 0 & 0 & -2 & -2 & 0 & 1 \end{array}\right) \quad (-1)$$

4. subtrahieren wir III von I

$$\left(\begin{array}{ccc|ccc} 2 & 1 & 2 & 1 & 0 & 0 \\ 0 & -1 & -2 & -1 & 1 & 0 \\ 0 & 0 & 2 & 2 & 0 & -1 \end{array}\right) \quad -III$$

5. addieren wir III zu II

$$\left(\begin{array}{ccc|ccc} 2 & 1 & 0 & -1 & 0 & 1 \\ 0 & -1 & -2 & -1 & 1 & 0 \\ 0 & 0 & 2 & 2 & 0 & -1 \end{array}\right) \quad +III$$

6. addieren wir II zu I

$$\left(\begin{array}{ccc|ccc} 2 & 1 & 0 & -1 & 0 & 1 \\ 0 & -1 & 0 & 1 & 1 & -1 \\ 0 & 0 & 2 & 2 & 0 & -1 \end{array}\right) \quad +II$$

7. dividieren wir I und III durch 2 und multiplizieren II mit (-1)

$$\left(\begin{array}{ccc|ccc} 2 & 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 & 1 & -1 \\ 0 & 0 & 2 & 2 & 0 & -1 \end{array}\right) \begin{array}{l} \div 2 \\ (-1) \\ \div 2 \end{array}$$

Das Ergebnis sieht wie folgt aus:

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 1 & 0 & -1 & -1 & 1 \\ 0 & 0 & 1 & 1 & 0 & -0.5 \end{array}\right)$$

$$(\boldsymbol{I}|\boldsymbol{A}^{-1})$$

Wir erreichen hier wiederum, wie in der vorherigen Rechnung:

$$\text{adj}(\boldsymbol{A}) = \begin{pmatrix} 0 & 0.5 & 0 \\ 1 & 1 & -1 \\ 1 & 0 & -0.5 \end{pmatrix}$$

Die *Division* zweier Matrizen kann als Multiplikation zweier Matrizen aufgefasst werden. Dazu benötigt man die Inverse.

$$\frac{\boldsymbol{A}}{\boldsymbol{B}} = \boldsymbol{A} \cdot \boldsymbol{B}^{-1}, \text{ deshalb auch } \frac{\boldsymbol{A}}{\boldsymbol{A}} = \boldsymbol{A} \cdot \boldsymbol{A}^{-1} = \boldsymbol{I}$$

Rechenregeln für Inverse

Sofern die Matrizen $\boldsymbol{A}$, $\boldsymbol{B}$ und $\boldsymbol{C}$ invertierbar sind:

$$\begin{aligned} (\boldsymbol{A}^{-1})^{-1} &= \boldsymbol{A} \\ \boldsymbol{AV} &= \boldsymbol{I}, \text{ wenn } \boldsymbol{V} = \boldsymbol{A}^{-1} \\ (\boldsymbol{AB})^{-1} &= \boldsymbol{B}^{-1}\boldsymbol{A}^{-1} \\ (\boldsymbol{ABC})^{-1} &= \boldsymbol{C}^{-1}\boldsymbol{B}^{-1}\boldsymbol{A}^{-1} \\ (\boldsymbol{A}')^{-1} &= (\boldsymbol{A}^{-1})' \\ (\varphi\boldsymbol{A})^{-1} &= \varphi^{-1}\boldsymbol{A}^{-1} \end{aligned}$$

Lösen von Gleichungen mittels Inversen:

$$\boldsymbol{AX} = \boldsymbol{B} \Leftrightarrow \boldsymbol{X} = \boldsymbol{A}^{-1}\boldsymbol{B}$$

$$\boldsymbol{XA} = \boldsymbol{B} \Leftrightarrow \boldsymbol{X} = \boldsymbol{B}\boldsymbol{A}^{-1}$$

A.10 Der Rang einer Matrix

Der *Rang* ist innerhalb der Mathematik ein Begriff aus der linearen Algebra. Man ordnet ihn einer linearen Abbildung oder einer Matrix zu. Übliche Abkürzungen sind $\text{rang}(\boldsymbol{A})$ oder $\text{rg}(\boldsymbol{A})$. Bei einer linearen Abbildung ist der Rang als Dimension des Bildes dieser Abbildung definiert. Zu einer Matrix exisitiert ein *Zeilenrang* und ein *Spaltenrang*. Der Zeilenrang ist die Dimension des von den Zeilenvektoren aufgespannten Vektorraumes und entspricht der Anzahl der unabhängigen Zeilenvektoren. Entsprechendes gilt für den Spaltenrang. Man kann zeigen, dass der Zeilenrang und der Spaltenrang identisch sind. Man spricht deshalb vom Rang einer Matrix. Fasst man eine Matrix als Abbildungsmatrix einer linearen Abbildung auf, so besitzen beide -die Matrix und die lineare Abbildung- den gleichen Rang.
Um den Rang einer Matrix zu bestimmen, formt man sie mittels gauss'schem Eliminationsverfahren in eine äquivalente Matrix in Stufenform um. Die Anzahl der von Null verschiedenen Zeilen ergibt den Rang der Matrix. Eine $n \times n$-Matrix heißt regulär, wenn sie vollen Rang aht, also wenn $\text{rg}(\boldsymbol{A}) = n$.

Beispiel :

$$\boldsymbol{A} = \begin{pmatrix} 3 & 4 & 6 \\ 0 & 3 & 2 \\ 0 & 6 & 5 \end{pmatrix} \sim \begin{pmatrix} 3 & 4 & 6 \\ 0 & 3 & 2 \\ 0 & 0 & 1 \end{pmatrix} \Rightarrow \text{rang}(\boldsymbol{A}) = 3$$

$$\boldsymbol{B} = \begin{pmatrix} 2 & 1 & 6 \\ 0 & 6 & 4 \\ 0 & 3 & 2 \end{pmatrix} \sim \begin{pmatrix} 2 & 1 & 6 \\ 0 & 6 & 4 \\ 0 & 0 & 0 \end{pmatrix} \Rightarrow \text{rang}(\boldsymbol{B}) = 2$$

Die einzige Matrix mit dem Rang 0 ist die Nullmatrix $\mathbf{0}$. Für eine $m \times n$ Matrix gilt: $\text{rang}(\boldsymbol{A}) \leq \min(m, n)$.
Alle reduzierten Korrelationsmatrizen, die von einem gemeinsamen Faktor gebildet werden haben eine Gemeinsamkeit: ihr Rang ist 1. Wenn die Matrix von zwei gemeinsamen Faktoren gebildet wird ist ihr Rang 2. Ein rg($\boldsymbol{A}$)=1 bedeutet, dass alle Spalten durch eine andere Spalte fehlerfrei reproduziert werden können. ein rg($\boldsymbol{A}$)=2 bedeutet, dass alle Spalten durch eine linearkombination von zwei anderen Spalten "vorhergesagt" werden können. Wenn wir wissen, dass k gemeinsame Faktoren gegeben sind, können wir daraus schleissen, dass der Rang der reduzierten korrelationsmatrix ebenflls k ist. Bei zwei oder mehr Faktoren sind jedoch zusätzliche Annehmen nötig, um das Modell zu formulieren. Sind die Faktoren korreliert, und mit welchen Variablen stehen die Faktoren in Beziehung? Verschmutzung der Daten durch Sampling- oder Messfehler können ebenfalls problematisch sein.

A.11 Idempotente Matrix

Eine quadratische Matrix heisst idempotent, wenn gilt: $\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}^2 = \boldsymbol{A}$.

Für idempotente Matrizen $\boldsymbol{X}$ und $\boldsymbol{Y}$ gilt:

$\boldsymbol{X}\boldsymbol{Y} = \boldsymbol{Y}\boldsymbol{X} \rightarrow \boldsymbol{X}\boldsymbol{Y}$ idempotent

$\boldsymbol{I} - \boldsymbol{X} \rightarrow$ idempotent

$\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{X}) = (\boldsymbol{I} - \boldsymbol{X})\boldsymbol{X} = \mathbf{0}$

A.12 Diverses

A.12.1 Gramian Matrix

Als *Gramian Matrix* bezeichnet man eine quadratische Matrix, wenn sie symmetrisch ist, und alle Eigenwerte ≥ 0 sind. Korrelations- und Kovarianzmatrizen sind immer gramian.

A.12.2 Spektralzerlegung einer Matrix

Die Matrix $\boldsymbol{R}$ sei Reell und Symmetrisch. Dann kann sie in die drei Matrizen $\boldsymbol{A}$, $\boldsymbol{\Lambda}$ und $\boldsymbol{A}'$ zerlegt werdn. Dabei handelt es sich bei $\boldsymbol{A}$ um eine Matrix, deren Spalten aus den Eigenvektoren von $\boldsymbol{R}$ bestehen. Die Matrix $\boldsymbol{\Lambda}$ ist eine Diagonalmatrix, deren Hauptdiagonale die Eigenwerten von $\boldsymbol{R}$ enthält. Es gilt:

$$\boldsymbol{R} = \boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}'$$

Die Matrix

$$\boldsymbol{R} = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}$$

lässt sich zerlegen in $\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}'$, also:

$$\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}' = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix}$$

mit den entsprechenden Werten:

$$\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}' = \begin{pmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 5 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{pmatrix}$$

ANHANG B

MAXIMUM LIKELIHOOD

Im Kern geht es darum, dass wir eine konkrete Stichprobe vorliegen haben, und uns fragen, welche Parameterwerte θ (z.B. Mittelwert und Varianz bei Normalverteilung) das Zustandekommen dieser konkreten Stichprobe *am wahrscheinlichsten* macht. Dazu müssen wir allerdings a-priori wissen, aus welcher Verteilung diese Stichprobe gezogen wurde. Wenn wir wissen, dass die Stichprobe aus einer normalverteilten Grundgesamtheit gezogen wurde, stellt sich die Frage der Gestalt: welcher Mittelwert μ und welche Varianz σ^2 macht die Stichprobendaten *am wahrscheinlichsten*?

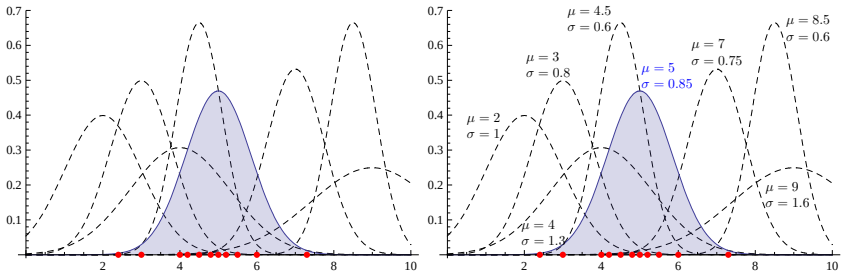


Abbildung B.1: Maximum Likelihood

In dieser Graphik sehen wir einige rot gekennzeichnete Meßwerte. Um intuitiv verstehen zu können, wie die ML-Methode funktioniert, wollen wir hier auf Berechnungen verzichten. Gestrichelt sind mehrere Normalverteilungen eingezeichnet, wir sehen jedoch recht deutlich, dass es bei einigen eher, bei anderen weniger wahrscheinlich ist, dass die Meßwerte aus einer dieser Normalverteilungen gezogen worden sind. Der Wahrscheinlichste Kandidat ist keine dieser gestrichelten Verteilungen, sondern die blau eingefärbte Normalverteilung. Für diese Normalverteilung ist die Wahrscheinlichkeit maximal, solche Meßwerte zu erreichen. Als Ergebnis erhalten wir die Parameter θ, nämlich $\mu = 5$ und $\sigma = 0.85$.

Der wichtigste Hintergedanke bei ML ist der, dass wir *nicht* die Wahrscheinlichkeitsdichte $f(Y|\theta)$ betrachten, bei der es sich um eine Funktion von Y bei

fixiertem θ handelt, sondern uns die Likelihood Funktion $L(\theta|Y)$ genauer anschauen, bei der es sich um eine Funktion von θ für fixes Y handelt. Also ziehen wir hier nicht aus einer über θ genau spezifizierten Funktion beliebige Werte Y, sondern wir betrachten die Werte Y der Zufallsstichprobe als fixiert, und schließen von diesen Werten auf ein ganz bestimmtes Set von Parametern θ.

B.1 ML formaler

Die Maximum Likelihood Methode (Größte Dichte Methode) erfordert Kenntnisse über die Verteilungsfunktion der Zufallsvariable und schätzt dann die Parameter θ dieser Verteilung. Dies geschieht so, dass das Produkt der Wahrscheinlichkeiten der Stichprobe maximal wird. Geht man von dem Parameter θ, der höherdimensional sein kann, dann gilt für den Fall n unabhängiger identischer Wiederholungen die Dichte:

$$L(y_1, \dots y_n|\theta) = f(y_1|\theta)f(y_2|\theta)\dots f(y_n|\theta)$$

Anstatt für *feste Parameter* θ die Dichte der beliebigen y_i Werte zu verändern kann man ebensogut für *feste Werte* y_i die Dichte als Funktion von θ auffassen:

$$L(\theta) = f(y_1, \dots y_n|\theta)$$

Diese Funktion heisst *Likelihoodfunktion* und besitzt als Argument den Parameter θ bei festen Realisationen von y_i. Diese Funktion ist zu maximieren:

$$L(\theta) = f(y_1|\theta)f(y_2|\theta)\dots f(y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta) \stackrel{!}{=} max$$

Diese Funktion wird partiell nach den Parametern abgeleitet und dann Null gesetzt. Also:

$$\frac{\partial L(\theta)}{\partial \theta} = 0$$

Damit hinreichende und notwendige Bedingung für ein Maximum beide erfüllt sind, muss die zweite Ableitung kleiner Null sein. Also:

$$\frac{\partial^2 L(\theta)}{\partial \theta^2} < 0$$

Da die Likelihood eines einzelnen Falles etwas ähnliches ist wie eine Wahrscheinlichkeit, kann sie einen Wertebereich von [0;1] annehmen. Das Produkt vieler Zahlen zwischen Null und Eins wird allerdings sehr klein, sodass sich ausgesprochen schlecht damit umgehen lässt. Um das Problem von Zahlen zu nahe an Null zu vermeiden, wird die Likelihood üblicherweise logarithmiert. Da Logarithmieren eine *monotone Transformation* und somit die Extremwerte bei den gleichen Werten für x vorkommen, kann ebenso gut (und einfacher!) der Logarithmus der Likelihood Funktion, die sogenannte *Log-Likelihood Funktion*, maximiert werden. Dies hat den großen Vorteil, dass sie leichter abzuleiten ist.

$$L(\theta) = f(y_1|\theta) \cdot f(y_2|\theta) \cdot \dots \cdot f(y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

Diese Funktion wird logarithmiert, da die Produkte im Term zu unerfreulichen Ergebnissen führen können und sich Summen leichter ableiten lassen. Produktterme werden beim logarithmieren in Summen, Exponenten in Produkte umgewandelt. Es resultiert also:

$$\mathcal{L}(\theta) = \ln L(\theta) = \ln f(y_1|\theta) + \ln f(y_2|\theta) + \ldots + \ln f(y_n|\theta) = \sum_{i=1}^{n} \ln f(y_i|\theta)$$

Logarithmen

Es existieren unendlich viele Logarithmen. $\log_a$ bezeichnet einen Logarithmus zur Basis a. Häufig verwendete Logarithmen sind:

ln	=	$\log_e$	=	Logarithmus naturalis, Basis $e = 2,718281828\ldots$
lg	=	$\log_{10}$	=	Dekadischer / Briggscher Logarithmus, Basis 10
ld / lb	=	$\log_2$	=	Logarithmus Dualis / Binärlogarithmus, Basis 2

$\lg 100 = \log_{10} 100$ gibt als Ergebnis, welchen Exponent für 10 man bennötigt, um 100 als Ergebnis zu erhalten. Also $\log_{10} 100 = 2$, da $10^2 = 100$
Gesetze zum Rechnen mit Logarithmen:

$$\log_a(u \cdot v) = \log_a u + \log_a v$$

$$\log_a(\tfrac{u}{v}) = \log_a u - \log_a v$$

$$\log_a(u^r) = r \log_a u \ (r \in \mathbb{R})$$

$$\log_a \sqrt[n]{u} = \tfrac{1}{n} \log_a u \ (n \in \mathbb{N}\backslash 1)$$

Es gilt übrigens $-2 \ln L(\theta) \sim \chi^2$

Wahrscheinlichkeitssätze

Multiplikationssatz für abhängige Ereignisse

$$p(A_1 \cap A_2 \cap A_3 \cap \ldots \cap A_k) =$$
$$p(A_1)p(A_2|A_1)p(A_3|A_1 \cap A_2 \cap A_3)\ldots p(A_k|A_1 \cap A_2 \cap A_3 \cap \ldots \cap A_{k-1})$$

Multiplikationssatz für unabhängige Ereignisse

$$p(A_1 \cap A_2 \cap A_3 \cap \ldots \cap A_k) =$$
$$p(A_1)p(A_2)p(A_3)\ldots p(A_k) = \prod_{i=1}^{k} p(A_k)$$

ANHANG C

BAYES STATISTIK

Bei der Bayes-Statistik handelt es sich nicht einfach um eine weitere Methode der herkömmlichen Statstik, sondern viel mehr um einen anderen *Ansatz*.

- In der *klassischen parametrischen Inferenzstatistik* werden die Daten unter der Annahme verschiedener Parameterwerte analysiert. Nach bestimmten Kriterien werden dann einige ausgewählt.
- In der *Bayesianischen Statistik* wird die Verteilung der Schätzparameter analysiert. Dies geschieht unter der Annahme einer bestimmten Verteilungsstruktur der Daten sowie der *a-priori* Verteilung der gesuchten Parameter.

Ein VORTEIL der Bayes-Statistik ist ihre Anwendbarkeit bei kleinen Fallzahlen. So können komplexe Modelle bei kleinem n berechnet werden, die mit herkömmlichen Methoden nicht zu bearbeiten wären. Ebenso kann qualitatives und quantitatives Wissen gemeinsam in die Vorannahmne der *a-priori*-Verteilung der Parameter eingehen.

Ein oft hervorgehobener NACHTEIL ist die große Bedeutung subjektiver Verteilungsannahmen der Parameter, die auf Vermutungen, früheren Erfahrungen oder -starken- Überzeugungen beruhen können. Dies ist ein Einfallstor für Kritiker. Ein anderes Problem, das mit der Leistungsfähigkeit heutiger und zukünftiger PCs an Bedeutung verliert ist die analytische Intraktabilität vieler Modelle, die sich nur durch numerische Schätzverfahren wie Jackknife, Bootstrap oder Markov Chain Monte Carlo-Simulationen lösen lassen.

C.1 Frequentisten vs. Bayesianer

Der Hauptunterschied zwischen Bayesianern und "normalen" Statistikern liegt in der grundsätzlichen Untersheidung des Begriffes der *Wahrscheinlichkeit*. Die klassische, frequentistische Sicht definiert Wahrscheinlichkeit wie folgt:

$$\lim_{n\to\infty} P\left(\left|\frac{f_a}{n} - p(A)\right| < \varepsilon\right) = 1$$

Das frequentistische Wahrscheinlichkeitskonzept ist *unabhängig* von subjektiven Wahrscheinlichkeitsvorstellungen.
Die bayesianische Sichtweise hingegen geht von *subjektiven Wahrscheinlichkeiten* aus. Nur so ist es möglich, Aussagen über die unbedingte a-priori Parameterwahrscheinlichkeit $p(\theta)$ zu machen. Wahrscheinlichkeit ist demnach die subjektive Einschätzug von Unsicherheit, z.B. durch Erfahrung, Epertisen, Forschung (qualitativ oder quantitativ) oder Aberglaube, wie mancher Kritiker polemisch meinen wird.

In der Praxis ist der bayesianische Ansatz gerade bei kleinen Stichproben dem klassischen, frequentistischen Ansatz überlegen. Zwar ist hier die Wahl der *a-prioir*-Verteilung von stärkerer Bedeutung, jedoch ist die Anwendung von bayesianischen Methoden besser -oder überhaupt- möglich als frequentistische Verfahren.

Der frequentistische Standarfehler wird durch die bayesianische Standardabweichung der *a-posteriori*-Verteilung ersetzt, das 95%-Konfidenzintervall durch das 2,5% bis 97,5%-Perzentil-Intervall der *a-posteriori*-Verteilung.

C.2 Grundlagen und Idee

Während in der *frequentistischen Statistik* z.B. über die Maximum Likelihood-Methode bei gegebenen Daten der wahrscheinlichste Parameter gesucht wird, so geht die Bayes-Statistik einen anderen Weg. Hier wird nicht nur nach dem Parameter gesucht, der die Datenwahrscheinlichkeit maximiert, sondern auch nach der tatsächlichen Wahrscheinlichkeit der verschiedenen möglichen Parameterwerte. Der Bayes Ansatz gibt sich nicht mit der konditionalen *Daten*wahrscheinlichkeit

$$p(\boldsymbol{X}|\theta)^1$$

zufrieden, sondern ermittelt die konditionale *Parameter*wahrscheinlichkeit. Diese wird auch als
a-posteriori-Wahrscheinlichkeit bezeichnet:

$$\text{a-posteriori}^2 : p(\theta|\boldsymbol{X})$$

.
Dies geschieht durch den bekannten SATZ VON BAYES:

$$p(\theta|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\theta)p(\theta)}{p(\boldsymbol{X})}$$

Bei genauerem hinsehen entdecken wir ein Problem:

$$\text{a-priori}^3 : p(\theta)$$

die sogenannte *a-priori*-Wahrscheinlichkeit, ist unbekannt. Da sie nicht objektiv gegeben ist, muss hier eine Schätzung vorgenommen werden, die -wie wir gehört haben- auf Vermutungen, Überzeugungen oder Erfahrungen beruhen kann,

[1] Wahrscheinlichkeit der Daten unter Bedingung der Parameter
[2] Wahrscheinlichkeit der Parameter unter Bedingung der Daten
[3] Wahrscheinlichkeit der Parameter

und eine argumentative Schwachstelle des Bayes-Ansatzes darstellt, da sie als subjektiv unwissenschaftlich angesehen werden könnte.

Die *unbedingte Datenwahrscheinlichkeit* $p(\boldsymbol{X})$ beruht indirekt auf der angenommen *a-priori*-Wahrscheinlichkeit. Warum ist das so? Über den SATZ DER TOTALEN WAHRSCHEINLICHKEIT kann sie als Summe aller möglichen Konditionalwahrscheinlichkeiten ermittelt werden:

$$p(\boldsymbol{X}) = \sum_{\theta} p(\boldsymbol{X}|\theta)p(\theta)$$

Im Falle stetiger Parameter ersetzen wir $\sum$ einfach durch $\int$.

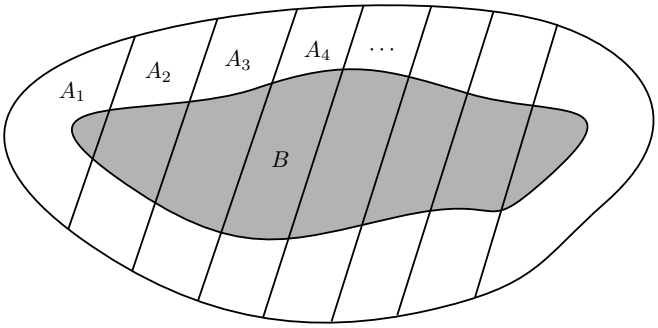


Abbildung C.1: Satz der totalen Wahrscheinlichkeit

Setzt man die Gleichung etwas anders ergibt sich:

$$p(\theta|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\theta)p(\theta)}{p(\boldsymbol{X})} \text{ in anderer Form: } p(\theta|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\theta)p(\theta)}{\sum_{\theta} p(\boldsymbol{X}|\theta)p(\theta)}$$

Der letzte Bruch zeigt uns -noch einmal auf andere Weise-, dass die a-posteriori-Wahrscheinlichkeit für jeden Parameter angibt, wie hoch seine Wahrscheinlichkeit, gegeben die Daten, ist.

BEGRIFFE			
a priori	$p(\theta)$	unbedingte	Parameterwahrscheinlichkeit
a posteriori	$p(\theta\|\boldsymbol{X})$	bedingte	Parameterwahrscheinlichkeit
	$p(\boldsymbol{X}\|\theta)$	bedingte	Datenwahrscheinlichkeit
	$p(\boldsymbol{X})$	unbedingte	Datenwahrscheinlichkeit

Beispiel: Satz der totalen Wahrscheinlichkeit

Betrachten wir Abbildung C.2: Um $p(B)$ zu berechnen, müssen wir nach dem Satz der totalen Wahrscheinlichkeit

$$p(B) = \sum_{i=1}^{m} p(B|A_i)p(A_i)$$

berechnen. Hier ergibt sich

$$p(B) = p(B|A_1)p(A_1) + p(B|A_2)p(A_2) + p(B|A_3)p(A_3) + p(B|A_4)p(A_4)$$

Warum ist dies richtig? Wir wissen, dass

$$p(A \cap B) = p(B|A)p(A)$$

gilt.
Wenn wir also substituiern, erhalten wir:

$$p(B) = p(A_1 \cap B) + p(A_2 \cap B) + p(A_3 \cap B) + p(A_4 \cap B)$$

Dies entspricht den 4 grauen Teilstücken, da es sich bei diesen um genau die Bereiche handelt, die A_i und B abdecken. Addiert man diese, so erhält man $p(B)$.

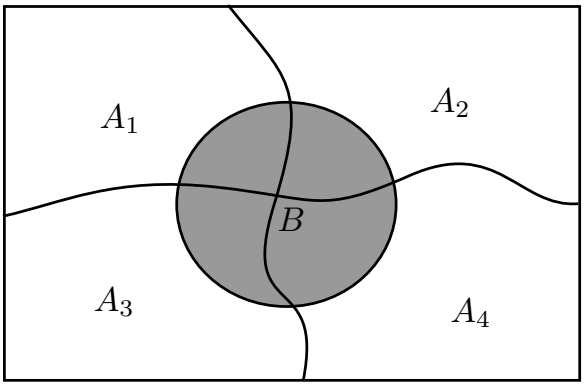


Abbildung C.2: Beispiel zum Satz der totalen Wahrscheinlichkeit

ANHANG D

DAS ALLGEMEINE LINEARE MODELL

Die Gleichung des allgemeinen linearen Modell (ALM) entspricht der Gleichung der multiplen Regression für Fall i

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_j x_{ij} + \varepsilon_i$$

Das Allgemeine Lineare Modell ist ein Ansatz, der viele varianzanalytische Verfahren verbindet. Hierzu zählen als Kernstücke die multiple Regressions- und Korrelationsanalyse und zudem die Diskriminanzanalyse, die Varianzanalyse,... Es handelt sich um einen Integrationsansatz, dem die einzelnen Verfahren vorausgingen. Im allgemeinen linearen Modell (ALM) sind die Parameter additiv verknüpft und treten (höchstens) in der ersten Potenz auf. Produkte oder Potenzen der Parameter sind nicht zulässig, sind es aber bei den Variablen.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i1} x_{i4} + \varepsilon_i$$

wobei diese Form wieder in die allgemeine Form des ALM überführbar ist, indem

$$x_{i1} x_{i4} = x_{i3} \text{ sowie } x_{i1}^2 = x_{i2}$$

gesetzt wird. So erhält man wieder die Form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

Die Prädiktorvariablen bzw. unabhängigen Variablen können sowohl Intervallskalenniveau als auch qualitatives Messniveau besitzen. Auch das Kriterium (abhängiger Teil) kann mehrdimensional sein. In diesem Falle werden sowohl die Prädiktorvariablen als auch die Kriteriumsvariablen als Linearkombinationen dargestellt.
Geht man von nur einer abhängigen Variablen und mehreren unabhängigen Variablen aus, sieht die Grundgleichung des ALM wie folgt aus:

$$y = X\beta + \varepsilon$$

oder

$$_n y_1 =_n X_{pp} \beta_n +_n \varepsilon_1$$

mit

- Dem $n \times 1$ Spaltenvektor y, der abhängigen Variablen.
- Der Matrix X, die die Werte der n Personen auf den p Variablen enthält.
- Dem Spaltenvektor β, der die Gewichte für die p Variablen enthält.
- Dem Spaltenvektor ε, der die Fehlerterme der n Personen enthält.
- Die Prädiktoren sind fest / konstant
- ε und y sind Zufallsvariablen
- $E(\varepsilon) = 0$
- $E(\varepsilon\varepsilon') = I$

$$E(\varepsilon\varepsilon') = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \dots & \varepsilon_n \end{pmatrix}$$

Multipliziert sich aus zu:

$$\begin{pmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \dots & E(\varepsilon_1\varepsilon_n) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & \dots & E(\varepsilon_2\varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(\varepsilon_n\varepsilon_1) & \dots & \dots & E(\varepsilon_n^2) \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

Also

$$e \sim \mathcal{N}(0, \sigma^2 I)$$

Dies entspricht der Varianz-Kovarianzmatrix der Fehlerterme wegen $E(\varepsilon) = 0$. Die Fehlervarianz ist konstant und es gibt keine Kovarianzen, d.h. die Fehler sind unkorreliert. Rückschlüsse auf die Population / Hypothesentests setzen die Annahme der Normalverteilung der Fehler voraus.
Werden im ALM Variablen mit nominalem Meßniveau betrachtet, müssen sie verschlüsselt werden. Die Varianzanalyse kann mit unabhängigen Variablen auf Nominalniveau durchgeführt werden. Verschlüsselungsarten sind Dummycodierung, Effektcodierung und Kontrastcodierung.

D.1 Kodierung

Nehmen wir zum beispiel die nominale Variable *Parteizugehörigkeit* mit den Ausprägungen

SPD	:	1
CDU	:	2
FDP	:	3
Grüne	:	4

Sind die Nominalvariablen codiert, können sie als Prädiktoren in einer multiple Regressionsgleichung zur Vorhersage einer abhängigen Variablen eingesetzt werden Welche Bedeutung haben die jeweiligen b-Gewichte?

Dummycodierung

In der Dummy-codierung werden aus den m Ausprägungen der ursprünglichen nominalen Variablen $m-1$ Dummyvariablen generiert. Dies geschieht deshalb, da sich ansonsten Multikollinearitätsprobleme ergeben, da die Information im letzten Dummy als redundant angesehen werden kann, da sie sich 100% aus den vorherigen $m-1$ Dummys reproduzieren lässt. Hier die drei Dummys *SPD: ja/nein*, *CDU: ja/nein* und *Grüne: ja/nein*.

Partei	x_1	x_2	x_3
SPD	1	0	0
CDU	0	1	0
FDP	0	0	1
Grüne	0	0	0

Die Dummycodierung mit der Regressionsgleichung

$$\hat{y}_i = b_0 + b_{i1}x_1 + b_{i2}x_2 + +b_{i3}x_3$$

Hat Person i die Parteipräferenz Grüne, so ergibt sich

$$\hat{y}_i = b_0 + 0 + 0 + 0, \quad a = \bar{y}_4 = \hat{y} \text{ nach KQ}$$

Eine andere Person präferiert die SPD:

$$\hat{y}_i = \bar{y}_1 = b_{i1} + 0 + 0 + \bar{y}_4,$$
$$b_{i1} = \bar{y}_1 - \bar{y}_4$$

Die Konstante y entspricht dem Mittelwert der Referenzgruppe, die Gewichte b_i, drücken die Differenzen zwischen Referenzgruppe und Gruppe i aus.

Effektcodierung

Personen, die in allen Kategorien eine Null haben, werden mit -1 codiert.

Partei	x_1	x_2	x_3
SPD	1	0	0
CDU	0	1	0
FDP	0	0	1
Grüne	-1	-1	-1

Die Regressionsgleichung lautet wieder wie folgt:

$$\hat{y}_i = b_0 + b_{i1}x_1 + b_{i2}x_2 + +b_{i3}x_3$$

Nach der KQ-Methode gilt als bester Schätzer für Personen mit Parteipräferenz Grüne $\bar{y}_4$, es ergibt sich

$$\bar{y}_4 = -b_1 - b_2 - b_3 + b_0$$

Für die übrigen Personengruppen ergeben sie die Mittelwerte

$$\begin{array}{rcccl} \bar{y}_1 & = & b_1 & + & b_0 \\ \bar{y}_2 & = & b_2 & + & b_0 \\ \bar{y}_3 & = & b_3 & + & b_0 \end{array}$$

Nach Einsetzen und Auflösen nach b_0 ergibt sich

$$b_0 = \frac{\bar{y}_1 + \bar{y}_2 + \bar{y}_3 + \bar{y}_4}{4} = \bar{G}$$

und damit

$$\begin{array}{rcl} b_1 & = & \bar{y}_1 - \bar{G} \\ b_2 & = & \bar{y}_2 - \bar{G} \\ b_3 & = & \bar{y}_3 - \bar{G} \end{array}$$

Die Konstante b_0 entspricht bei Effektcodierung dem Gesamtmittelwert der abhängigen Variablen. Die b_i entsprechen den Differenzen der Gruppenmittel zum Gesamtdurchschnitt.

Kontrastcodierung

Entspringt dem Gedanken einzelne Mittelwerte einer Varianzanalyse zu vergleichen. Dies geschieht mittels Gewichten c_i, für die gilt $\sum_i c_i = 0$

Partei	x_1	x_2	x_3
SPD	1	0	$\frac{1}{2}$
CDU	-1	0	$\frac{1}{2}$
FDP	0	1	$-\frac{1}{2}$
Grüne	0	-1	$-\frac{1}{2}$

Kontrastieren von SPD/CDU

$$c_1 = 1,\ c_2 = -1,\ c_3 = 0,\ c_4 = 0\ \rightarrow x_1$$

Kontrastieren von FDP/Grüne

$$c_1 = 0,\ c_2 = 0,\ c_3 = 1,\ c_4 = -1\ \rightarrow x_2$$

Kontrastieren von SPD/CDU mit FDP/Grüne

$$c_1 = \frac{1}{2},\ c_2 = -\frac{1}{2}\ \rightarrow x_3$$

Die Regressionsgleichung lautet wieder wie folgt:

$$\hat{y}_i = b_0 + b_{i1}x_1 + b_{i2}x_2 + +b_{i3}x_3$$

Die beste Schätzung ist auch hier

$$\begin{array}{rcccccccc}
\bar{y}_1 & = & b_0 & + & 1b_1 & + & 0b_2 & + & \frac{1}{2}b_3 \\
\bar{y}_2 & = & b_0 & + & (-1b_1) & + & 0b_2 & + & \frac{1}{2}b_3 \\
\bar{y}_3 & = & b_0 & + & 0b_1 & + & 1b_2 & + & \left(-\frac{1}{2}\right)b_3 \\
\bar{y}_4 & = & b_0 & + & 0b_1 & + & (-1)b_2 & + & \left(-\frac{1}{2}\right)b_3
\end{array}$$

- x_1 konstrastiert Zugehörigkeit zu CDU oder SPD, sonst 0.
- x_2 konstrastiert Zugehörigkeit zu FDP oder Grünen, sonst 0.
- x_3 konstrastiert Zugehörigkeit CDU/SPD oder FDP/Grüne.

Für die 4 Gleichungen mit 4 Unbekannten erhält man nach Auflösung für $b_0,\ b_1,\ b_2,\ b_3$:

$$\begin{array}{rcl}
b_1 & = & \dfrac{\bar{y}_1 - \bar{y}_2}{2} \\
b_2 & = & \dfrac{\bar{y}_3 - \bar{y}_4}{2} \\
b_3 & = & \dfrac{\bar{y}_1 + \bar{y}_2}{2} - \dfrac{\bar{y}_3 + \bar{y}_4}{2} \\
b_0 & = & \dfrac{\bar{y}_1 + \bar{y}_2 + \bar{y}_3 + \bar{y}_4}{4} = \bar{G}
\end{array}$$

b_0 entspricht dem Gesamtmittelwert der abhängigen Variablen y. Die b-Gewichte entsprechen dem jeweils codierten Kontrast.

D.2 Verallgemeinertes lineares Modell

Das verallgemeinerte Lineare Modell subsummiert auch das Allgemeine Lineare Modell. Die Schätzung für dichotome und poissonverteilte abhängige Variable lässt sich im Rahmen der verallgemeinerten linearen Modelle vereinheitlichen. Die Basis des Verallgemeinerten Linearen Modells sind die Wahrscheinlichkeitsfunktionen der "Exponentialfamilie". Entwickelt wurde die Idee der Exponentialfamilie von Fisher, wobei es darum geht, eine einheitliche allgemeine mathematische Struktur einer Funktion zu schaffen, innerhalb derer verschiedene Subfunktionen darstellbar sind. Exponentialfamilie meint nun, dass verschiedene Unterfunktionen in der Exponentenkomponente der natürlichen Exponentialfunktion ($e = 2.71828\ldots$) enthalten sind. Jede Subfunktion kann in den Exponenten gebracht werden, wobei der Transfer über den natürlichen Logarithmus geschieht.

D.2.1 Beispiele

- Für die Poisson-Verteilung

$$f(y|\mu) = \frac{e^{-\mu}\mu^y}{y!}$$

$$\rightarrow \exp\{y\log(\mu) - \mu - \log(y!)\}$$

- Für die Binomialverteilung

$$f(y|n,p) = \binom{n}{y} p^y (1-p)^{n-y}$$

$$\rightarrow \exp\left\{\log\binom{n}{y} + y\log(p) + (n-y)\log(1-p)\right\}$$

- Für die Normalverteilung

$$f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

$$\rightarrow \exp\left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y^2 - 2y\mu + \mu^2)\right\}$$

D.2.2 Generalisierung

Die Generalisierung des linearen Modells geschieht wie folgt:

1. y ist die Zufallskomponente des Modells, diese ist entsprechend einer der Wahrscheinlichkeitsfunktionen der drei Exponentialfamilien (siehe oben) verteilt
2. $\Theta = X\beta$ ist die systematische Komponente des Modells. Die erklärenden Variablen x beeinflussen y nur indirekt über die Funktion g()
3. Die Linkfunktion Θ verbindet die systematische und die Zufallskomponente des Modells
4. Die Linkfunktion sorgt dafür, dass im linearen Modell mit Variablen gearbeitet werden kann, die den Modellkriterien nicht entsprechen.

Die Schätzung der Parameter bzw. Maximierung der Liklihoodfunktion geschieht über den Newton-Raphson-Algorithmus. Zur Kategorie der verallgemeinerten linearen Modelle zählt beispielsweise die logistische Regression.