5 Multivariate Regression

5.1 Das Modell

a In der multiplen linearen Regression wurde der Zusammenhang von mehreren Ausgangsvariablen oder Regressoren mit einer kontinuierlichen Zielgrösse untersucht. Nun sollen **mehrere Zielgrössen** gleichzeitig betrachtet werden.

▷ ***Beispiel Fossilien.*** *Aus Fossilien, die man in verschiedenen Schichten von Meeres-Ablagerungen findet, will man auf Umweltbedingungen (Temperatur, Salzgehalt) der entsprechenden Zeitperioden schliessen.*

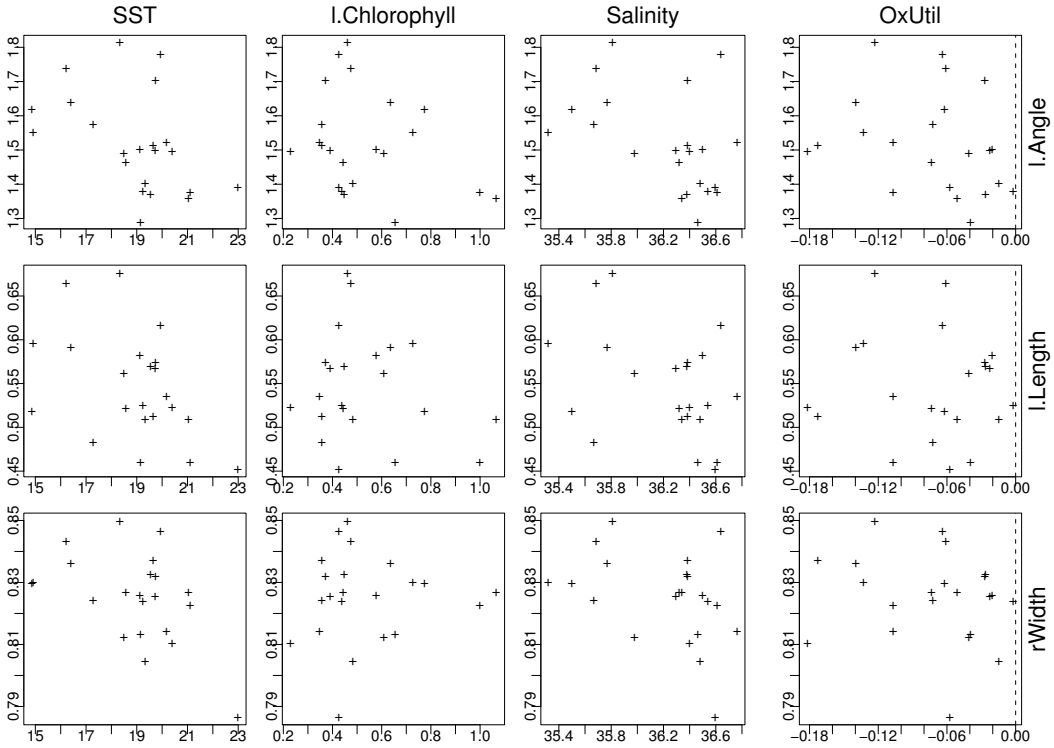

Abbildung 5.1.a: Umweltvariable und Form-Merkmale im Beispiel der Fossilien

Dazu werden an verschiedenen Stellen des Atlantischen Ozeans Messungen an „coccoliths“ der Art Gephyrocapsa *der obersten Ablagerungen vorgenommen und mit den heutigen Umweltbedingungen in Beziehung gesetzt. In Abbildung 5.1.a sind die Beziehungen zwischen den einzelnen Umweltvariablen und den Form-Merkmalen der Fossilien dargestellt.*

Das entsprechende Modell soll nachher dazu benützt werden, anhand von coccoliths aus tieferen Schichten auf die Umweltbedingungen in den entsprechenden Zeitperioden zurückzuschliessen. Für diesen Schluss muss man von der Annahme ausgehen, dass sich diese Beziehungen seither nicht geändert haben. Genaueres steht in Bollmann, Henderiks and Brabec (2002).

b **Modell.** Das Modell der multiplen linearen Regression mit einer einzigen Zielgrösse war $Y_i = \beta_0 + \sum_k \beta_k x_i^{(k)} + E_i$. Wenn nun der Zusammenhang mehrerer Zielgrössen $Y^{(j)}$, $j = 1, ..., m$, von den Ausgangsgrössen (oder erklärenden Variablen) $X^{(k)}$ untersucht werden soll, dann können wir zunächst für jede ein solches Modell aufstellen, also

$$Y_i^{(j)} = \beta_0^{(j)} + \sum_k \beta_k^{(j)} x_i^{(k)} + E_i^{(j)}$$

Das soll wieder mit Matrizen zusammengefasst werden,

$$\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{E} \ .$$

Die einzelnen Modelle, in Matrix-Schreibweise, erhalten wir, wenn wir jeweils die j-te Spalte von $\boldsymbol{Y}$, $\boldsymbol{\beta}$ und $\boldsymbol{E}$ auswählen: $\underline{Y}^{(j)} = \boldsymbol{X}\,\underline{\beta}^{(j)} + \underline{E}^{(j)}$. Wie in der Matrixschreibweise der **univariaten Regression**, also der Regression mit einer einzigen Zielgrösse, erscheint der **Achsenabschnitt** β_0 in der Matrixform nicht mehr; er wird dadurch berücksichtigt, dass eine Spalte mit lauter Einsen in die Design-Matrix $\boldsymbol{X}$ eingeschlossen wird.

c **Ausgangsgrössen, Regressoren und Terme.** Mit Regressionsmodellen wird allgemein ein Zusammenhang zwischen Zielgrössen und Ausgangsgrössen untersucht. Die Ausgangsgrössen werden oft als erklärende Variable bezeichnet, was sicher gerechtfertigt ist, wenn ein Ursache-Wirkungs-Zusammenhang besteht. Da Regression auch sinnvoll ist, wenn das nicht postuliert werden kann, soll der neutrale Ausdruck **Ausgangsgrösse statt erklärende Variable** benützt werden. Die ebenfalls übliche Bezeichnung „unabhängige Variable“ wird vermieden, da das Adjektiv „unabhängig“ nur verwirrt: Die X-Variablen müssen in keiner Weise unabhängig voneinander sein.

Ausgangsgrössen gehen oft nicht in der ursprünglichen Form ins Regressionsmodell ein, sondern werden zunächst transformiert – einzeln, beispielsweise mit einer Logarithmus-Transformation, oder gemeinsam, indem beispielsweise die eine als Prozentzahl einer anderen ausgedrückt wird. Diese transformierten Grössen, die als X-Variable ins Regressionsmodell eingehen, nennen wir **Regressoren**. Analoges kann mit den Zielgrössen geschehen. Man könnte dann die transformierten Zielgrössen als „Regressanden“ bezeichnen. Die Unterscheidung zu den Zielgrössen ist weniger wichtig: in der Residuenanalyse spielen die untransformierten Ausgangsgrössen eine Rolle, untransformierte Zielgrössen dagegen nicht. Da zudem „Regressand“ zu ähnlich tönt wie „Regressor“, bleiben wir beim Ausdruck „Zielgrösse“, der auch für transformierte Zielgrössen gelten soll.

Bei der Festlegung des Modells kommt zusätzlich der Begriff „**Term**“ ins Spiel. Die dummy-Variablen, die zu einem Faktor (s. unten, 5.1.f) oder einer Interaktion zwischen Variablen gehören, bilden jeweils einen Term. Bei der Modellwahl wird jeder Term ins Modell gesamthaft einbezogen oder weggelassen.

d **Zufallsabweichungen.** Die Annahmen über die Verteilung der Zufallsabweichungen $E_i^{(j)}$ bilden die naheliegende Verallgemeinerung der Annahmen im Fall einer einzigen Zielgrösse. Es sei $\underline{E}_i$ die $i$te Zeile von $\boldsymbol{E}$, also der Vektor der Zufallsabweichungen aller Zielgrössen für die Beobachtung i. Die Annahmen sind:

- Erwartungswert $\mathcal{E}\langle\underline{E}_i\rangle = \underline{0}$. Diese Festlegung ist für die Identifizierbarkeit von $\underline{\beta}$ nötig und sagt auch, dass die (lineare) Regressionsfunktion richtig ist.
- Die Zufallsabweichungen $E_i^{(j)}$ haben Varianzen σ_j^2, die für alle Beobachtungen gleich sind. Ausserdem können die $E_i^{(j)}$ für verschiedene Zielgrössen zusammenhängen. Beides zusammen wird durch die Kovarianzmatrix $\mathrm{var}\langle\underline{E}_i\rangle = \boldsymbol{\Sigma}$ charakterisiert, von der wir annehmen, dass sie gleich ist für alle Beobachtungen i.
- Die Zufallsabweichungen der *verschiedenen Beobachtungen* sind unabhängig (oder wenigstens unkorreliert), $\mathcal{E}\langle\underline{E}_h\underline{E}_i^T\rangle = \boldsymbol{0}$, falls $h \neq i$.
- Die Zufallsabweichungen sind gemeinsam normalverteilt.

Man kann das alles zusammenfassen zu

$$\underline{E}_i \sim \mathcal{N}_m\langle\underline{0}, \boldsymbol{\Sigma}\rangle\,, \quad \text{unabhängig}\,.$$

Das Modell mit der gemeinsamen Verteilung der Fehlerterme ist das Modell der **multivariaten Regression**. Sie ist auch eine *multiple Regression*, soweit sie mehrere Regressoren $X^{(j)}$ umfasst.

e Die Modelle für die einzelnen Zielgrössen haben wir zunächst einfach formal in eine einzige Matrizen-Formel geschrieben. Durch die Annahme einer gemeinsamen Normalverteilung der Fehlerterme erhalten sie jetzt auch inhaltlich eine Verbindung.

Die Tatsache, dass die Design-Matrix $\boldsymbol{X}$ für alle Zielgrössen die gleiche ist, muss nicht zwingend einen inhaltlichen Zusammenhang angeben: Wenn die Koeffizienten-Matrix $\boldsymbol{\beta}$ in jeder Zeile nur ein einziges von Null verschiedenes Element enthält, dann reagieren die Zielgrössen eben auf verschiedene Regressoren, die man nur formell zu einer Matrix zusammengefasst hat.

f **Varianzanalyse.** Die Ausgangsgrössen können, wie in der univariaten Regression, auch Faktoren (nominale oder kategorielle Variable) sein, die als „dummy variables“ in die Design-Matrix $\boldsymbol{X}$ eingehen.

Die multivariate Varianzanalyse mit festen Effekten (MANOVA) kann deshalb als Spezialfall der Regression behandelt werden. Wie bei einer einzigen Zielgrösse gibt es aber interessante zusätzliche methodische Aspekte.

5.2 Schätzungen und Tests

a **Schätzung der Koeffizienten.** Die Spalten von $\boldsymbol{\beta}$ können separat durch Kleinste Quadrate, also mit je einer (univariate) multiplen Regressionrechnung geschätzt werden. Das lässt sich aber auch zusammengefasst schreiben als

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

Die angepassten Werte und die Residuen sind auch wie früher definiert und werden zusammengefasst zur Matrix $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\,\widehat{\boldsymbol{\beta}}$ und zur Residuen-Matrix $\boldsymbol{R} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$.

Die **Schätzung der Kovarianzmatrix** der Zufallsabweichungen erfolgt durch die empirische Kovarianzmatrix der Residuen unter Berücksichtigung der Anzahl $n-p$ der Freiheitsgrade,

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-p}\boldsymbol{R}^T\boldsymbol{R}$$

b **Verteilung der geschätzten Koeffizienten.** Wie zu erwarten ist, sind die Koeffizienten erwartungstreu und normalverteilt. Die Kovarianzmatrix der geschätzten Koeffizienten wird schon irgendwie zu berechnen sein; überlassen wir das getrost den Programmen!

* Wollen Sie es etwas genauer wissen? Da stossen wir auf eine Schwierigkeit in der Notation: Die geschätzten Koeffizienten bilden eine zufällige Matrix. Wer brauchen nicht nur die Verteilung jedes einzelnen Elementes dieser Matrix, sondern auch die gemeinsame Verteilung der Elemente. Insbesondere interessieren uns auch die Kovarianzen zwischen den Elementen. Sie werden $\operatorname{cov}\langle\widehat{\beta}_h^{(j)}, \widehat{\beta}_k^{(\ell)}\rangle = ((\boldsymbol{X}^T\boldsymbol{X})^{-1})_{hk}\,\boldsymbol{\Sigma}_{j\ell}$. Das lässt sich nicht direkt als Matrix schreiben, denn es variieren vier Indices!

Im Übrigen ist die Herleitung nicht schwierig: Man setzt $\boldsymbol{C} = \boldsymbol{X}^T\boldsymbol{X}$ und rechnet

$$\begin{aligned}\operatorname{cov}\langle\underline{\widehat{\beta}}^{(j)}, \underline{\widehat{\beta}}^{(k)}\rangle &= \operatorname{cov}\langle\boldsymbol{C}^{-1}\boldsymbol{X}^T\underline{Y}^{(j)}, \boldsymbol{C}^{-1}\boldsymbol{X}^T\underline{Y}^{(k)}\rangle = \boldsymbol{C}^{-1}\boldsymbol{X}^T\operatorname{cov}\langle\underline{Y}^{(j)}, \underline{Y}^{(k)}\rangle(\boldsymbol{C}^{-1}\boldsymbol{X}^T)^T \\ &= \boldsymbol{C}^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{j\ell}\boldsymbol{X}(\boldsymbol{C}^{-1})^T = \boldsymbol{\Sigma}_{j\ell}\boldsymbol{C}^{-1}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{C}^{-1} = \boldsymbol{\Sigma}_{j\ell}\boldsymbol{C}^{-1}\end{aligned}$$

c ▷ *Im* ***Beispiel der Fossilien*** *sind die Ergebnisse für die einzelnen Zielgrössen in Tabelle 5.2.c zusammengestellt. Sie sind nicht ermutigend, liefert doch der Gesamttest für keine Zielgrösse ein signifikantes Resultat!*

	l.Angle		l.Length		rWidth	
	coef	p-value	coef	p-value	coef	p-value
(Intercept)	447.787	0.347	1.6590	0.471	0.59265	0.283
SST	-0.721	0.800	-0.0102	0.463	-0.00493	0.147
l.Chlorophyll	-19.202	0.155	-0.0765	0.238	0.00208	0.890
Salinity	-10.756	0.452	-0.0242	0.726	0.00888	0.590
OxUtil	-23.770	0.662	0.0433	0.869	-0.04368	0.489
R^2	0.285	0.260	0.2571	0.198	0.24889	0.255

Tabelle 5.2.c: Regressionskoeffizienten und Bestimmtheitsmasse mit p-Werten für die einzelnen Form-Variablen als Zielgrössen im Beispiel der Fossilien

d **Gemeinsame Tests.** Wir wissen, wie wir für jede einzelne Zielgrösse $Y^{(j)}$ testen, ob sie von den Ausgangsgrössen abhängt. Aus der gemeinsamen Betrachtung ergibt sich auch die gemeinsame Nullhypothese, dass keine der Zielgrössen von einem Regressor $X^{(k)}$ abhängt, dass also $\beta_k^{(j)} = 0$ ist für alle j oder, noch umfassender, dass zwischen keiner Zielgrösse und keinem Regressor ein Zusammenhang besteht, dass also alle $\beta_k^{(j)} = 0$ sind. Dazwischen liegen, wie in der univariaten linearen Regression, die Vergleiche von hierarchisch geschachtelten Modellen.

Die naheliegendste Art, eine solche Hypothese zu testen, besteht in der Verwendung des entsprechenden Likelihood-Ratio-Tests. Diese Teststatistik wird als Wilks' Λ (grosses griechisches Lambda) bezeichnet.

* In der univariaten Regression setzt die Teststatistik des F-Tests im Wesentlichen die „between group sum of squares" ins Verhältnis zur „within group sum of squares". Im multivariaten Fall werden beide Grössen zu „sum of squares and cross products" *Matrizen*, bezeichnet mit $\boldsymbol{B}$ und $\boldsymbol{W}$. Entscheidend ist wieder die Grösse des Quotienten. Teststatistiken sind deshalb Funktionen der Eigenwerte λ_k von $\boldsymbol{W}^{-1}\boldsymbol{B}$.

- Wilks: $\prod_\ell 1/(1+\lambda_\ell)$
- Pillai: $\sum_\ell \lambda_\ell/(1+\lambda_\ell)$
- Lawley-Hotelling: $\sum_\ell \lambda_\ell$
- Roy (union-intersection): λ_1 (resp. $\lambda_1/(1+\lambda_1)$)

Im multivariaten Fall gibt es also **mehrere gebräuchliche Tests,** die für den Fall einer einzigen Zielgrösse in den üblichen F-Test (oder t-Test) übergehen. Wenn der Einfluss einer einzigen kontinuierlichen Variablen getestet wird, liefern alle diese Tests das gleiche Ergebnis (* da $\boldsymbol{B}$ nur einen Freiheitsgrad hat und deshalb nur der erste Eigenwert λ_1 von 0 verschieden ist).

e ▷ *Im **Beispiel der Fossilien** zeigt der globale Test, der die Nullhypothese prüft, dass kein Zusammenhang zwischen den Form- und den Umwelt-Variablen bestehe, keine Signifikanz! Die Sache scheint also hoffnungslos. – Genauere Analysen ergaben die Möglichkeit, aus der Verteilung des Winkels (Angle) und der Länge (l.Length) drei Gruppen zu identifizieren und die Anteile dieser Gruppen in den Stichproben als neue Zielvariable einzuführen. Tabelle 5.2.e gibt in der mit „.total." bezeichneten Zeile an, dass die Umweltvariablen gesamthaft auf diese Gruppen einen signifikanten Einfluss haben. Die anderen Teststatistiken führen zu P-Werten von 0.0388 (Pillai), 0.0163 (Hotelling-Lawley) und 0.00381 (Roy).*

	Df	Wilks	approx F	num Df	den Df	p value
SST	1	0.564	5.405	2	14	0.0182
l.Chlorophyll	1	0.886	0.905	2	14	0.4271
Salinity	1	0.847	1.267	2	14	0.3122
OxUtil	1	0.890	0.863	2	14	0.4431
.total.	4	0.417	1.922	8	28	0.0961
Residuals	15					

Tabelle 5.2.e: Gesamttests für den Einfluss der einzelnen Regressoren auf die Gruppenanteile, sowie für alle Regressoren zusammen im Beispiel der Fossilien

Gemäss Tabelle haben die Variablen SST und OxUtil einen signifikanten Einfluss. Wie in der univariaten Regression ist es aber denkbar, dass die höheren P-Werte der anderen beiden Variablen von Kollinearität verursacht sind. Um dies zu überprüfen, rechnen wir wie früher das Modell ohne die am wenigsten signifikante Variable durch, also ohne `l.Chlorophyll`*. Die Variable* `Salinity` *erhält dann einen P-Wert von 0.177, während die andern beiden mit 0.018 und 0.053 ähnliche P-Werte behalten.*

f **Bedeutung der multivariaten Regression.** Von der Interpretation her sind meistens die Koeffizienten $\boldsymbol{\beta}$ von Interesse.

Schätzung und Vertrauensintervall für ein $\beta_k^{(j)}$ aus der multivariaten Regression sind identisch mit denen aus der multiplen Regression von $Y^{(j)}$ auf die Regressoren – die anderen Zielgrössen haben keinen Einfluss.

Wenn man einen Lauf mit einem Programm für multivariate Regression macht, erhält man also als Hauptsache das, was auch m Läufe eines Programms für multiple Regression liefert. Zusätzlich erhält man:

- die Kovarianzmatrix der Zufallsfehler. Die Korrelation zwischen den Zufallsabweichungen $E^{(j)}$ und $E^{(\ell)}$ von den linearen Regressionen von $Y^{(j)}$ und $Y^{(\ell)}$ auf die Regressoren $X^{(k)}$, $k = 1, ..., p$ nennt man auch **partielle Korrelation** zwischen $Y^{(j)}$ und $Y^{(\ell)}$, gegeben die X-Variablen.
- gemeinsame Tests für die vorher genannten Fragen, ob die Zielgrössen alle von gewissen oder allen Regressoren unabhängig sind – genauer, ob sie linear mit ihnen zusammenhängen.

g **Residuen-Analyse.** Residuen-Analyse zur Prüfung der Modellannahmen ist, wie in allen Regressionsmodellen, ein unverzichtbarer Bestandteil einer seriösen Datenanalyse. Zuerst sollen die Regressionen für alle einzelnen Zielgrössen mit den bekannten Methoden überprüft werden.

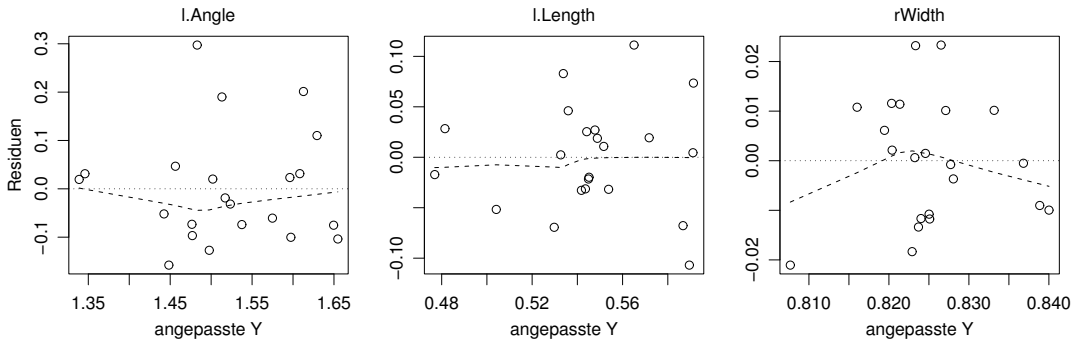

Abbildung 5.2.g (i): Tukey-Anscombe-Diagramme für das Beispiel der Fossilien

Abbildung 5.2.g (i) zeigt für das Beispiel die Zusammenstellung der Tukey-Anscombe-Diagramme, die zur Gesamtüberprüfung des Modells und insbesondere für Hinweise auf die Nützlichkeit einer Transformation der Zielgrössen dienen. Die Streudiagramme der Residuen gegen die Hebelarm-Werte (leverages) (Abbildung 5.2.g (ii)), aus denen man einflussreiche Beobachtungen erkennt. Die Streudiagramme der Residuen gegen die Ausgangsgrössen (Abbildung 5.2.g (iii)) sollen vor allem Hinweise Nichtlinearitäten

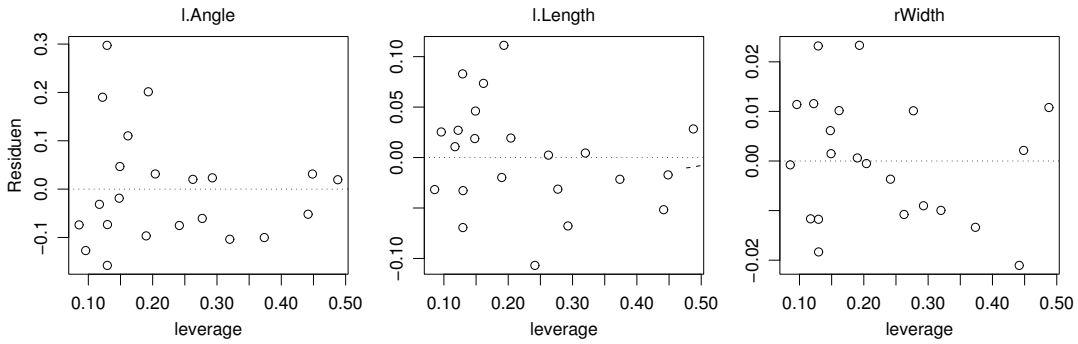

Abbildung 5.2.g (ii): Streudiagramm der Residuen gegen Hebelarmwerte für das Beispiel der Fossilien

in den Ausgangsgrössen geben. – Ausser einer schiefen Fehler-Verteilung für `l.Angle` zeigt sich im Beispiel kaum etwas Ernst zu Nehmendes.

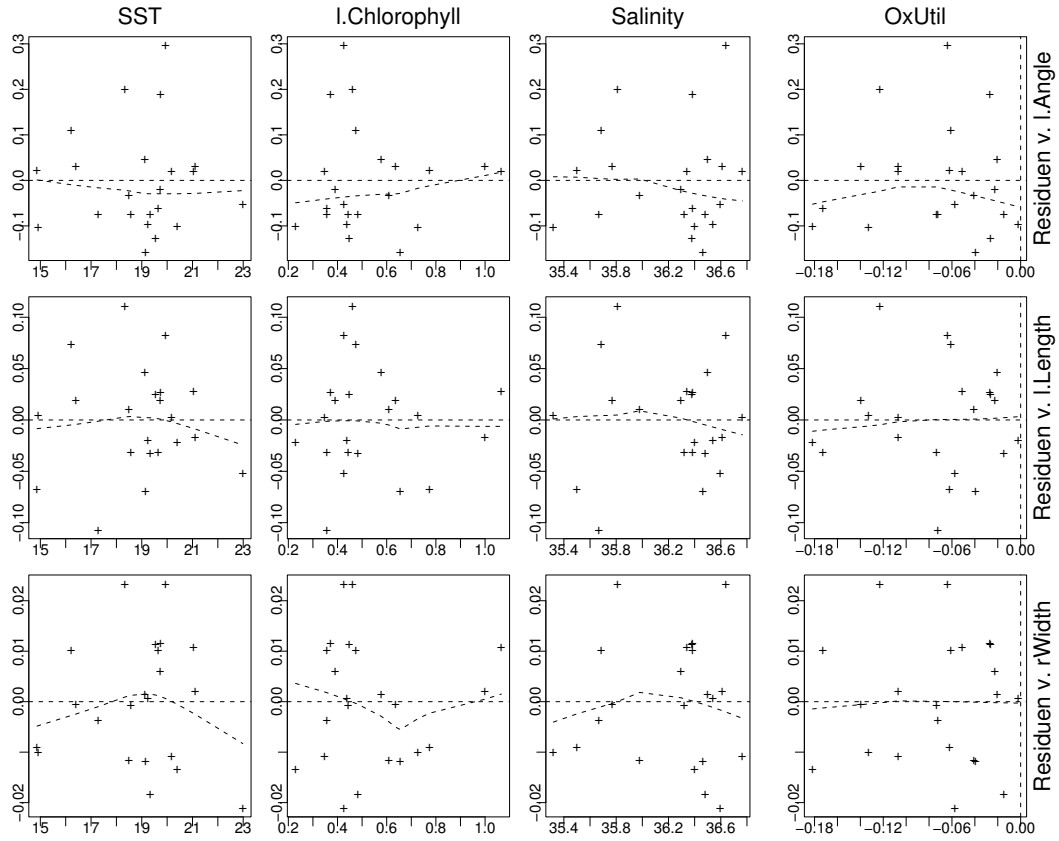

Abbildung 5.2.g (iii): Streudiagramme der Residuen gegen die Ausgangsgrössen für das Beispiel der Fossilien

h In Ergänzung zu den Residuen-Analysen für jede Zielgrösse lohnt es sich, eine Streudiagramm-Matrix der Residuen-Matrix (Abbildung 5.2.h (i)) zu betrachten. Es fallen im Beispiel (mindestens) zwei extreme Punkte mit grossen Residuen für alle Variablen auf. Wären mehr Beobachtungen vorhanden, so könnte man die Rechnungen ohne diese beiden Punkte wiederholen.

Aus den Residuen und ihrer geschätzten Kovarianzmatrix erhält man in der früher besprochenen Art () multivariat standardisierte Grössen, deren quadrierte Längen näherungsweise Chiquadrat-verteilt sind, falls die Annahme der Normalverteilung für die Zufallsabweichungen stimmt. Mit Hilfe eines Quantil-Quantil-Diagramms (Abbildung 5.2.h (ii)) kann man deshalb diese Voraussetzung überprüfen.

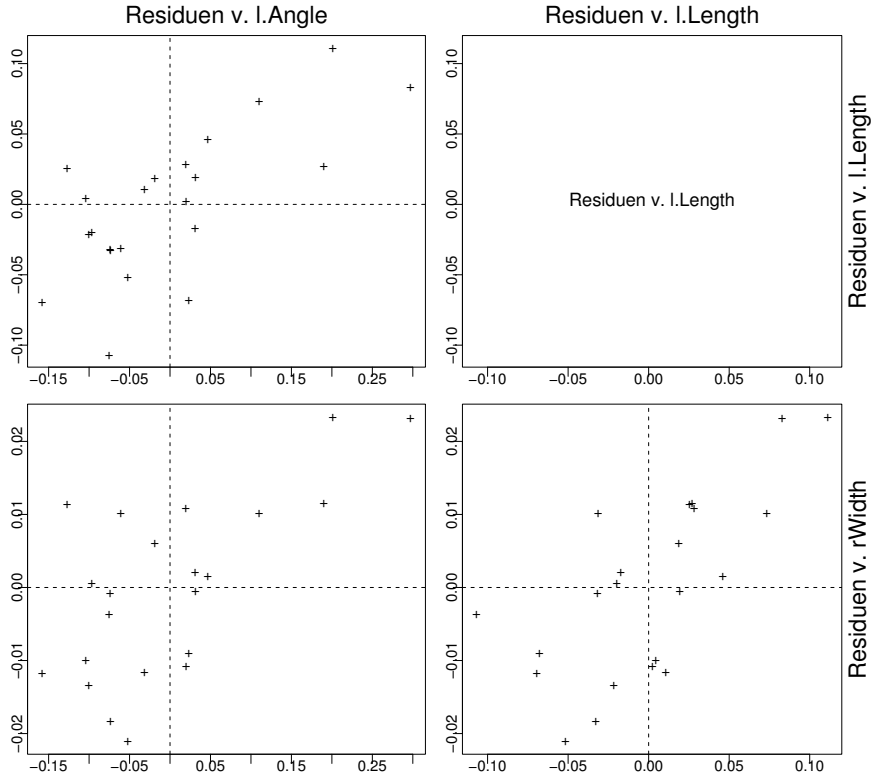

Abbildung 5.2.h (i): Streudiagramm-Matrix der Residuen für das Beispiel der Fossilien

i* **Vorhersage.** Es soll für einen gegebenen Satz $\underline{x}_0$ von Werten der Regressoren eine neue Beobachtung $\underline{Y}_0$ gemacht werden. Was können wir im Voraus über die Verteilung von $\underline{Y}_0$ sagen?

Bei bekannten Parametern ist das Problem trivial: Die gesamte Verteilung der neuen Beobachtung ist durch das Regressionsmodell 7.2.a gegeben. Die beste Vorhersage ist der Erwartungswert von $\underline{Y}_0$, $\mathcal{E}\langle \underline{Y}_0^T \rangle = \underline{x}_0^T \boldsymbol{\beta}$ (transponiert geschrieben).

In der Realität muss der Zusammenhang von $\underline{x}$ und $\underline{Y}$ aus „Trainingsdaten“ $\boldsymbol{X}$, $\boldsymbol{Y}$ geschätzt werden. Das führt zur Schätzung $\widehat{\boldsymbol{\beta}}$, die wir an Stelle von $\boldsymbol{\beta}$ einsetzen. Die beste Vorhersage wird also $\underline{\widehat{Y}}_0^T = \underline{x}_0^T \widehat{\boldsymbol{\beta}}$. Die Verteilung der Vorhersage lässt sich wie im eindimensionalen Fall aus der Verteilung der geschätzten Koeffizienten herleiten.

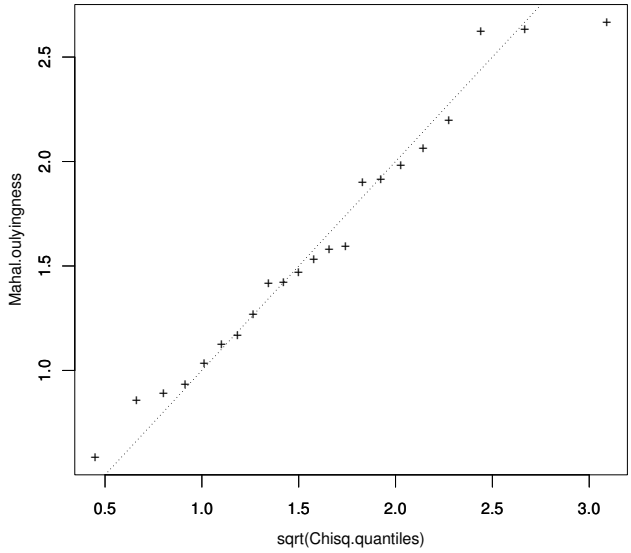

Abbildung 5.2.h (ii): Q-Q-Diagramm der Längen der multivariat standardisierten Residuen für das Beispiel der Fossilien

j* **Vorhersagebereich.** Der Vorhersagebereich soll die *Beobachtung* mit vorgegebener Wahrscheinlichkeit enthalten (während ein Vertrauensbereich einen *Parameter* mit einer solchen Wahrscheinlichkeit enthält). Analog zum Vorhersage-Intervall für eine einige Zielgrösse gibt die Summe der Kovarianzmatrizen für die geschätzte beste Vorhersage und für den Zufallsfehler der neuen Beobachtung, $\mathrm{var}\langle\underline{\widehat{Y}}\rangle + \widehat{\boldsymbol{\Sigma}}$, die Grösse und Form des gesuchten Bereiches an.

5.S S-Funktionen

a Wie bereits erwähnt (4.S), dienen die Funktionen `lm` und `manova` zur Durchführung von multivariaten Varianzanalysen und Regressionen.

```
> t.r  <- lm( cbind(Sepal.Length,Sepal.Width,Petal.Length,
Petal.Width) Species, data=iris)
```

erzeugt, da `lm` hier mit mehreren Zielgrössen aufgerufen wird, ein Objekt der Klasse `mlm`, für die

```
> summary(t.r)
```

die Resultate für alle Zielgrössen nacheinander auflistet.

b Für multivariate Tests braucht man `manova`,

```
> t.r  <- manova( cbind(Sepal.Length,Sepal.Width,Petal.Length,
Petal.Width) Species, data=iris)
```

Die Funktion `summary(t.r,test="Wilks")` führt dann den Test durch. Wenn das Modell mehrere Terme (Faktoren, Ausgangsgrössen) umfasst, werden entsprechend viele Tests durchgeführt – Vorsicht! Es sind „Type I“ Tests, die für ein schrittweise aufgebautes Modell jeweils prüfen, ob der nächste Term eine signifikante Verbesserung des Modells bringt.

c Da R zurzeit in dieser Beziehung lückenhaft ist, stellt der Autor einige Funktionen zur Verfügung. Man erhält sie über

```
> source("...")
```

Für die zusätzlichen Funktionen gibt es leider zurzeit noch keine Hilfe-Dokumentation ausser einem längeren Kommentar in jeweiligen Programm, den man erhält, indem man beispielsweise `drop1.mlm` ohne Klammern eintippt.

d **Funktion `drop1.mlm`.** Zunächst gibt es da eine Funktion `drop1.mlm`, die man aufrufen kann mit

```
> drop1(t.r)
```

Sie liefert die „Type III“ Tests, prüft also, ob die einzelnen Terme des Modells weggelassen werden können, ohne dass sich die Anpassung signifikant verschlechtert.

e **Funktion `summary.mreg`.** Eine Zusammenfassung der Koeffizienten in Form einer Tabelle, die der Matrix $\boldsymbol{\beta}$ entspricht und somit alle Zielgrössen umfasst, erhält man durch

```
> summary.mreg(t.r)
```

Die Funktion liefert zudem eine analoge Tabelle für die Standardfehler und die P-Werte, die angeben, ob ein einzelner Koeffizient signifikant von 0 verschieden ist (ob also die entsprechende Ausgangsgrösse für eine bestimmte Zielgrösse aus dem Modell weggelassen werden kann – eine univariate Betrachtungsweise).

f **Funktion `plot.regr`.** Die Funktion `plot.regr` liefert eine umfassende Residuen-Analyse – auch für univariate Regressionen.

```
> plot.regr(t.r)
```

(Die Funktion `regr` passt viele verschiedene Arten von Regressionsmodellen an – leider zurzeit noch keine multivariaten.)

In Kürze, was diese Funktion zeigt:

- Streudiagramme der Residuen gegen die angepassten Werte für alle Zielgrössen. Diese Streudiagramme dienen dazu, die generelle Form der Regressionsfunktionen zu prüfen und insbesondere Hinweise auf allfällige Transfomationen der Zielgrössen zu geben.
- Streudiagramme der Absolutwerte der Residuen gegen die angepassten Werte. Man kann gegebenenfalls Abweichungen von der Voraussetzung der gleichen Varianzen für alle Beobachtungen entdecken.
- Normalverteilungs-Diagramme.
- Streudiagramme der Residuen gegen die „Hebelarm“-Werte (leverages). Sie zeigen einflussreiche Beobachtungen an.
- Streudiagramm-Matrix der Residuen für die verschiedenen Zielgrössen.
- Streudiagramm-Matrix der Residuen gegen die Ausgangsgrössen im Modell. Sie können Hinweise auf Abweichungen von Linearitätsannahmen und Verbesserungsmöglichkeiten durch Transformation von Ausgangsgrössen geben.

Literaturverzeichnis

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, Wiley, N. Y.

Andrews, D. F. and Herzberg, A. M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, Springer-Verlag, N. Y.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988). *The S Language; A Programming Environment for Data Analysis and Graphics*, Wadsworth & Brooks/Cole, Pacific Grove.

Bilodeau, M. and Brenner, D. (1999). *Theory of Multivariate Statistics*, Springer Texts in Statistics, Springer-Verlag, New York.

Bollmann, J., Henderiks, J. and Brabec, B. (2002). Global calibration of gephyrocapsa coccolith abundance in holocene sediments for paleotemperature assessment, *Paleoceanography* **17**(3): 1035.

Bortz, J. (1977). *Lehrbuch der Statistik für Sozialwissenschaftler*, Springer Lehrbücher, Springer, Berlin.

Brown, P. J. (1993). *Measurement, Regression, and Calibration*, Clarendon Press, Oxford, U.K.

Chambers, J. M. (1998). *Programming with Data; A Guide to the S Language*, Springer-Verlag, New York.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S*, Wadsworth & Brooks/Cole.

Chatfield, C. and Collins, A. J. (1980). *Introduction to Multivariate Analysis*, Science Paperbacks, Chapman and Hall, London.

Cleveland, W. S. (1993). *Visualizing Data*, Hobart Press, Summit, New Jersey. *2 Ex.*

Cleveland, W. S. (1994). *The Elements of Graphing Data*, 2nd edn, Hobart Press, Summit, New Jersey.

Cooley, W. W. and Lohnes, P. R. (1971). *Multivariate Data Analysis*, Wiley, New York.

Everitt, B. S. (1978). *Graphical Techniques for Multivariate Data*, Heinemann Educational Books.

Fahrmeir, L., Hamerle, A. and Tutz, G. (eds) (1996). *Multivariate statistische Verfahren*, 2nd edn, de Gruyter, Berlin.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Ann. Eugenics* **7**: 179–184.

Flury, B. (1997). *A first course in multivariate statistics*, Springer texts in statistics, Springer-Verlag, NY.

Friedman, Hastie and Tibshirani (2000). Additive logistic regression: a statistical view of boosting, *Annals of Statistics* **28**: 377–386.

Fuller, W. A. (1987). *Measurement Error Models*, Wiley, N. Y.

Gnanadesikan, R. (1997). *Methods for Statistical Data Analysis of Multivariate Observations*, Series in Probability and Statistics, 2nd edn, Wiley, NY.

Green, P. E. and Caroll, J. D. (1976). *Mathematical Tools for Applied Multivariate Analysis*, Academic Press, New York.

Harman, H. H. (1960, 1967). *Modern Factor Analysis*, 2nd edn, University of Chicago Press.

Harris, R. J. (1975). *A Primer of Multivariate Statistics*, Academic Press, New York.

Hastie, T. and Tibshirani, R. (1994). Discriminant analysis by gaussian mixtures, *Journal of the Royal Statistical Society B* **?**: ?

Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis, *Annals of Statistics*.

Hastie, T., Tibshirani, R. and Buja, A. (1994). Flexible disriminant analysis by optimal scoring, *Journal of the American Statistical Association* pp. 1255–1270.

Jewell, P. L., Güsewell, S., Berry, N. R., Käuferle, D., Kreuzer, M. and Edwards, P. (2005). Vegetation patterns maintained by cattle grazing on a degraded mountain pasture. *Manuscript*

Johnson, N. L. and Kotz, S. (1972). *Continuous Multivariate Distributions*, A Wiley Publication in Applied Statistics, Wiley, New York.

Johnson, R. A. and Wichern, D. W. (1982, 1988, 1992). *Applied Multivariate Statistical Analysis*, Prentice Hall Series in Statistics, 3rd edn, Prentice Hall Int.,Englewood Cliffs,N.J.,USA.

Karson, M. J. (1982). *Multivariate Statistical Methods*, The Iowa State University Press, Ames.

Kendall, M. G. (1957, 1961). *A Course in Multivariate Analysis*, Griffin's Statistical Monographs & Courses, No.2, 2nd edn, Charles Griffin, London.

Krzanowski, W. J. (1988). *Principles of Multivariate Analysis; A User's Perspective*, Clarendon Press, Oxford.

Lawley, D. N. and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*, Butterworths Mathematical Texts, 2nd edn, Butterworths, London.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley, N. Y.

Manly, B. F. J. (1986, 1990). *Multivariate Statistical Methods: A Primer*, Chapman and Hall, London.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press, London.

Maxwell, A. E. (1977). *Multivariate Analysis in Behavioural Research*, Monographs on Applied Probability and Statistics, Chapman and Hall, London.

Morrison, D. F. (1967, 1976). *Multivariate Statistical Methods*, McGraw-Hill Series in Probability and Statistics, 2nd edn, McGraw-Hill Book Co., New York.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, Wiley, N. Y.

Rencher, A. C. (1995). *Methods of Multivariate Analysis*, Wiley, N. Y.

Rencher, A. C. (1998). *Multivariate Statistical Inference and Applications*, Wiley, N. Y.

Renner, R. M. (1993). The resolution of a compositional data set into mixtures of fixed source compositions, *Applied Statistics — Journal of the Royal Statistical Society C* **42**: 615–631.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge UK.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, number 72 in *Monographs on Statistics and Applied Probability*, Chapman and Hall.

Seber, G. A. F. (1984). *Multivariate Observations*, Wiley, N. Y.

Srivastava, M. S. and Carter, E. M. (1983). *An Introduction to Applied Multivariate Statistics*, North Holland.

Stahel, W. A. (2002). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 4. Aufl., Vieweg, Wiesbaden.

Tatsuoka, M. M. (1971). *Multivariate Analysis: Techniques for Educational and Psychological Research*, Wiley, New York.

Timm, N. H. (2002). *Applied Multivariate Analysis*, Springer-Verlag, N. Y.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*, Graphics Press, Cheshire.

Tufte, E. R. (1990). *Envisioning Information*, Graphics Press, Cheshire.

Tufte, E. R. (1997). *Visual Explanations; Images and quantities, evidence and narrative*, Graphics Press, Cheshire.

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-Plus*, Statistics and Computing, 3rd edn, Springer-Verlag, New York.

Venables, W. N. and Ripley, B. D. (2000). *S Programming*, Statistics and Computing, Springer-Verlag, New York.