Statistische Methoden

Dr. C.J. Luchsinger

7 Regression

Literatur Kapitel 7

* Lindgren: Kapitel 12, 14
* Cartoon Guide: Kapitel 11
* Krengel: 13.4
* Stahel: Kapitel 13

7.1 Einfache Regression

In diesem Teil werden wir eine etablierte Methode kennenlernen, mit der wir eine (lineare) Beziehung zwischen 2 Variablen untersuchen können (z.B. Punkte in Mathematik und Punkte in Physik von der gleichen Person). Zu Recht ist man an die Korrelation aus der WTS erinnert. Im (theoretischen) Modell

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

werden wir die x-Variable als fest betrachten und versuchen, die Y-Variable möglichst genau durch x vorher zu sagen oder durch x zu erklären (stören wird uns dabei der Störterm ϵ). Es wird dann mit Daten $(x_i, y_i), 1 \leq i \leq n$, darum gehen, β_0 (engl. intercept) und β_1 (engl. slope) zu schätzen und zu testen, ob nicht z.B. $\beta_1 = 0$ gilt.

Gründe für Regressionsanalyse sind jeweils:

* Y vorhersagen (Interpolation (interessierendes $x \in [x_{(1)}, x_{(n)}]$) und Extrapolation (interessierendes $x \notin [x_{(1)}, x_{(n)}]$))

* Zusammenhang erklären anhand bisheriger Daten

7.1.1 Motivierendes Beispiel - Gefahren

Wir haben in der ersten Stunde zu diesem Kapitel den folgenden Datensatz mitsamt den dazugehörigen Plots betrachtet. Die Daten liest man vom File www.luchsinger-mathematics.ch/Radio.txt mit Copy & Paste in R folgendermassen ein (bitte in einer freien Stunde diese Befehle selber nachholen):

7.1.1.1 Einlesen der Daten

Daten aus dtv-Atlas zur Ökologie, Tafeln und Texte, 1990.

ort<- c("Pripjet", "Chistogalovka", ... , "Japan", "Japan2")
bq<- c(5300, 2000, ... , 0.15, 0.85) [in Kbq]
dist<- c(3.16, 5.62, ... , 12589.25, 12589.25) [in Km]
regen<- c(0, 0, ... , 0, 1)
daten<- data.frame(ort, regen, dist, bq)
daten [zeigt Tabelle an]

7.1.1.2 Erster Plot und darauffolgende Transformation der Daten

plot(dist, bq)

Man braucht praktische Erfahrung als StatistikerIn oder theoretische Kenntnisse aus dem Fachgebiet (hier Meteorologie, Ausbreitung von Gasen), um eine sinnvolle Transformationsfunktion (hier den Logarithmus, sowohl für x-, wie auch für y-Achse) zu finden.

plot(ln(dist),ln(bq))

7.1.1.3 Versuch, Radioaktivität mit Distanz allein zu erklären (oder vorherzusagen)

tsch<- lm(ln(bq) $\sim$ ln(dist))

"lm" steht für **L**inear **M**odel, $\ln(bq)$ nennt man die "Response Variable" (abhängige Variable) und $\ln(dist)$ eine erklärende Variable oder einen "Predictor". Wir werden in 7.1.2 die Methode kennenlernen, mit der hier R die Parameter schätzt.
tsch

$$\ln(bq) = 9.803 - 1.091 * \ln(dist) + \epsilon$$

abline(tsch) [zeichnet die Regressionsgerade in bereits bestehenden Plot]
aov(tsch)
Residual standard error: 1.279721 [geschätzte Standardabweichung des Fehlerterms, engl. Residuals]

Wegen des negativen Koeffizienten (-1.091) haben wir also einen negativen (linearen) Zusammenhang: je grösser die Distanz zum Unglücksreaktor, desto kleiner die Radioaktivität.

7.1.1.4 Versuch, Radioaktivität mit Regen allein zu erklären (oder vorherzusagen)

lm(ln(bq)∼regen)

$$\ln(bq) = 3.7784 - 0.8247 * \text{regen} + \epsilon$$

plot(regen, ln(bq))

Entgegen unseren Erwartungen haben wir auch hier einen negativen (linearen) Zusammenhang. Wir haben in der ersten Stunde gesehen, dass dies auf eine Besonderheit der Stichprobe zurückzuführen ist. Es hat bei unserem Datensatz nur dort geregnet, wo man schon weit vom Reaktor weg ist. Wenn man dann nur den Regen als erklärende Variable zur Verfügung hat, entsteht deshalb ein falsches Bild (Danger!!!).

7.1.1.5 Versuch, Radioaktivität mit Distanz *und* Regen zu erklären (oder vorherzusagen)

tschdue<- lm(ln(bq)∼ ln(dist)+regen)

$$\ln(bq) = 10.522 - 1.360 * \ln(dist) + 2.723 * \text{regen} + \epsilon$$

Residual standard error: 0.6166197

Dies ist *auf die Schnelle* wohl die beste Datenanalyse: der Regen hat (wie in der ersten Stunde vermutet und theoretisch erwartet) einen "Wash out" mit erhöhter Radioaktivität zur Folge.

Die geschätzte Varianz des Fehlerterms ist hier kleiner als bei 7.1.1.3, da noch ein Predictor dazugekommen ist.

Wir werden in 7.1 die theoretischen Grundlagen lediglich für den Fall der einfachen Regression liefern. In 7.1.1.5 haben wir mit 2 erklärenden Variablen bereits die multiple Regression - sie wird in Teil 7.4 vertieft behandelt.

7.1.1.6 Getrennte Untersuchung der Daten mit und ohne Niederschlag

<u>mit Niederschlag</u>

distR<- daten[daten[,2]==1, 3]
bqR<- daten[daten[,2]==1, 4]
lm(ln(bqR)∼ ln(distR))

$$\ln(bqR) = 15.691 - 1.684 * \ln(distR) + \epsilon$$

<u>ohne Niederschlag</u>

distNR<- daten[daten[,2]==0, 3]
bqNR<- daten[daten[,2]==0, 4]
lm(ln(bqNR)∼ ln(distNR))

$$\ln(bqNR) = 10.445 - 1.345 * \ln(distNR) + \epsilon$$

Generell ist noch anzumerken, dass die aufgeführten Orte (mit/ohne Regen) ein sehr unvollständiges Bild (schlimmer: tendenziöses) Bild der Lage geben. Es gibt Orte sehr nahe

beim Unglücksreaktor (in unserer Liste nicht erwähnt), wo die Radioaktivität relativ tief ist, weil der Wind am Anfang vor allem in die andere Richtung blies. Die Regressionsebene aus 7.1.1.5 taugt am ehesten bei Orten, welche weiter weg liegen.

7.1.1.7 Welches ist jetzt das richtige Modell?

Nach diesen verschiedenen Modellen fragen wir uns vielleicht: "Welches ist jetzt *das richtige Modell*?" Ausser in Simulationen, wo man genau weiss, woher die Daten stammen, ist diese Frage nicht zu beantworten. Die verschiedenen möglichen Methoden und Modelle sollten uns aber zumindest dahingehend vorsichtig machen, dass wir uns bewusst sind, dass man mit anderen Modellen andere Resultate erhält.

In den Sozialwissenschaften (Soziologie, Psychologie und auch Ökonomie) gilt diese Relativierung sehr stark (wann heiraten Menschen?). In denjenigen Naturwissenschaften, in denen die Versuchsbedingungen einigermassen vollständig kontrolliert werden können, ist jedoch zu erwarten, dass man so etwas wie ein richtiges Modell finden kann (Atomzerfall, Wachstum einer Bakterienpopulation im Labor). Bei der Medizin ist zu bemerken, dass die untersuchten Menschen in permanent ändernden Bedingungen leben (und nicht im Labor). Deshalb wird es dort je nach Untersuchungsgegenstand kaum universelle Modelle geben.

7.1.1.8 Welche Modellannahmen werden wir machen?

Wie überall in der Statistik werden wir ein intensives Wechselspiel zwischen Daten (x_i, y_i) und (vermuteten) Zufallsgrössen haben. Wenn wir uns auf der Ebene der Zufallsgrössen befinden, machen wir folgende Modellannahmen $(1 \leq i \leq n)$:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{7.1}$$

Dabei sind β_0, β_1 zu schätzende Parameter, $x_1, \ldots, x_n$ sind fest und $\epsilon_1, \ldots, \epsilon_n$ sind iid $\mathcal{N}(0, \sigma^2)$-verteilt. Auch σ^2 ist ein zu schätzender Parameter. Die Normalverteilung ist

nicht zwingend und je nach Untersuchungsgegenstand sogar schlecht. In vielen Anwendungen subsumiert man aber in dieser residualen Grösse kleine Effekte wie

* Messfehler
* Rundungsfehler
* zufällige Schwankungen
* kleinste Einflüsse, welche man nicht in das Modell einbauen will, um es einfach zu halten

7.1.2 Schätzen von β_0, β_1 und σ^2: OLS und MLE

Wir werden jetzt 2 Methoden kennenlernen (OLS und MLE), mit denen man die unbekannten Parameter in (7.1) schätzen kann. In diesem einfachen Modell werden die Schätzungen $(\hat{\beta}_0, \hat{\beta}_1)$ für (β_0, β_1) gleich sein, egal welche Methode wir anwenden. Es gibt jedoch einen wichtigen Unterschied: bei der OLS-Schätzung werden wir gar keine Modellannahmen (ausser linear) machen müssen. Wir definieren $\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i, 1 \leq i \leq n$ (geschätzte Gerade). Die Aufgabe ist lediglich: wir suchen eine Gerade durch die Punktwolke $(x_i, y_i)_{i=1}^n$ derart, dass die Summe der quadrierten Fehler (Sum of Squared Errors)

$$SSE := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 := \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

minimal ist (OLS=Ordinary Least Squares). Wir definieren weiter gleich auf Vorrat wichtige Summen:

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

"R" steht dabei für Regression, dann noch die 3 Summen

$$SS_{xx} := \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2,$$

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

und

$$SS_{xy} := \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.$$

In der Theorie zur Regression sind die absoluten Werte von x_i und y_i nicht weiter von Interesse (in den Anwendungen schon (vgl. Radioaktivität)). Man zieht deshalb meist von allem Anfang an den Mittelwert $\bar{x}$ bzw. $\bar{y}$ ab. Überraschenderweise gilt jetzt eine Beziehung, welche an den Pythagoras erinnert (den Beweis machen Sie bitte in den Übungen, wobei Sie die nachfolgenden Schätzungen (7.5) für β_0 und (7.4) für β_1 benutzen):

$$SS_{yy} = SSR + SSE,$$

also

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{7.2}$$

Jargon: Statistiker sagen dann: "Die Variation in den y (SS_{yy}) lässt sich aufspalten in einen Anteil, der durch die Regression erklärt wird (SSR) und eine residuale Summe (SSE)."

OLS-Schätzung

Schreiten wir jetzt zur OLS-Schätzung: Wir haben die Summe

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2 \tag{7.3}$$

bzgl. β_0 und β_1 zu minimieren. Von der Mittelschule her wissen wir, dass man

* bei eindimensionalen Optimierungsproblemen die erste Ableitung gleich 0 setzt (und die zweite Ableitung noch überprüft).

* Dies ist auch bei mehrdimensionalen Problemen so (bei der zweiten Ableitung ist es ein bisschen schwieriger - siehe Analysis II).

Die Ableitung von (7.3) nach β_0 bzw. β_1 gleich 0 gesetzt ergibt:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0; \qquad \sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0.$$

Dies ist äquivalent zu

$$\bar{y} - \beta_0 - \beta_1 \bar{x} = 0; \qquad SS_{xy} + n\bar{x}\bar{y} - \beta_0 n\bar{x} - \beta_1(SS_{xx} + n\bar{x}^2) = 0.$$

Nach einfachen Umformungen erhalten wir

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \tag{7.4}$$

und

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{7.5}$$

Wenn wir $\hat{\beta}_1$ anschauen und voraussetzen (Spezialfall), dass die Varianzen von x und y gleich 1 sind, so ist dies einfach die Korrelation zwischen x und y. Regression und Korrelation haben also sehr wohl etwas miteinander zu tun.

Offensichtlich geht die Regressionsgerade immer durch den Punkt $(\bar{x}, \bar{y})$: von Gleichung (7.5): $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$.

Uns fehlt noch ein Schätzer für σ^2. Eine Möglichkeit ist, dass wir mit den $e_i := y_i - \hat{y}_i, 1 \leq i \leq n$, den sogenannten *beobachteten Residuen* (die wahren sind ja unbeobachtet), arbeiten:

$$\hat{\sigma}^2 := \frac{1}{n-2} \sum_{i=1}^{n} e_i^2. \tag{7.6}$$

Warum "$n-2$" im Nenner? Angelernte StatistikerInnen sagen dazu: wir haben ja 2 Parameter (β_0, β_1) geschätzt, es gehen also 2 Freiheitsgrade verloren. Dieser Schätzer für σ^2 ist erwartungstreu (siehe Teil 7.5). Wir beweisen in den Übungen, dass auch $\hat{\beta}_0, \hat{\beta}_1$ erwartungstreu sind.

ML-Schätzung

Wie bereits angekündigt, sind die MLE für β_0 und β_1 gleich wie die OLS-Schätzer. Wir werden die Berechnung der MLE auf die Berechnung der OLS-Schätzer zurückführen.

Wegen (7.1) ist Y_i normalverteilt mit Erwartungswert $\beta_0 + \beta_1 x_i$ und Varianz σ^2:

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$

Die Likelihood (siehe Definition 5.7) ist also

$$\prod_{i=1}^{n} \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big) \exp\Big[-\frac{1}{2\sigma^2}(y_i - (\beta_0 + \beta_1 x_i))^2\Big].$$

Wir berechnen die Log-Likelihood

$$-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 - \frac{n}{2} \ln \sigma^2 + c, \tag{7.7}$$

wobei c eine Konstante ist. Da wir diesen Ausdruck maximieren wollen, können wir einfach (egal, wie wir σ^2 schätzen!) für die Schätzung von β_0, β_1 den Ausdruck

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

minimieren. Aber das ist ja unser Ausgangsproblem bei der OLS-Schätzung. Deshalb müssen die Schätzer gleich sein wie bei der OLS-Methode. Als freiwillige Übung kann man jetzt $\hat{\beta}_0$ und $\hat{\beta}_1$ in (7.7) einsetzen und bezüglich σ^2 maximieren. Man erhält dann den MLE für σ^2:

$$(\hat{\sigma}^2)^{MLE} = \frac{1}{n}\sum_{i=1}^{n} e_i^2,$$

zu vergleichen mit (7.6). Zusammengefasst: die MLE und OLS-Schätzung für β_0, β_1 ist gleich und die Schätzung für σ^2 ist bis auf den Faktor $n/(n-2)$ gleich.

Darstellung zu (beobachtete) Residuen; wahre Gerade vs geschätzte Gerade:

7.1.3 Testen ob $\beta_1 = 0$

Wenn wir uns im Spezialfall der einfachen Regression nur für die Steigung der Geraden interessieren, gibt es eine einfache Herleitung eines statistischen Tests ob

$$\mathcal{H}_0 : \beta_1 = 0$$

gilt oder nicht (alternative Frage: $\beta_1 = c$, c eine beliebige Zahl, oder nicht). Wieso treten solche Fragen überhaupt auf? Dazu drei Skizzen, alle drei unter $\mathcal{H}_0$, "The Good, the Bad, and the Ugly" - der "Ugly" lässt uns einen Fehler 1. Art machen:

Im Teil 7.4 werden wir in der multiplen Regression viel eleganter mit Matrizen die folgenden Rechnungen durchführen. Für die jetzige Untersuchung setzen wir voraus, dass (7.1) gilt. Dann können wir folgendermassen argumentieren:

1. Die Schätzformel für β_1 lautet (vgl. (7.4)):

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

2. Wir substituieren $\forall\, 1 \leq i \leq n$:

$$x_i' := x_i - \bar{x}.$$

Wir erhalten damit die einfachere Formel

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum_{i=1}^{n} x_i'(y_i - \bar{y})}{\sum_{i=1}^{n} {x_i'}^2} = \frac{\sum_{i=1}^{n} x_i' y_i}{\sum_{i=1}^{n} {x_i'}^2},$$

wobei wir im letzten Gleichheitszeichen gebraucht haben, dass $\sum_{i=1}^{n} x_i' = 0$.

3. Schätzer sind (vor der Realisation) Zufallsgrössen. Wir setzen für die Datenpunkte y jetzt die Zufallsgrössen Y ein:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i' Y_i}{\sum_{i=1}^{n} {x_i'}^2} = \frac{1}{\sum_{i=1}^{n} {x_i'}^2} \sum_{i=1}^{n} x_i' Y_i. \tag{7.8}$$

Da $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), 1 \leq i \leq n$, unabhängig verteilt, können wir die Verteilung von $\hat{\beta}_1$ als Zufallsgrösse einfach herleiten:

$$\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n {x'_i}^2}).$$

Das kleine Problem ist, dass wir σ^2 (wieder mal) nicht kennen. Aber genau wie früher können wir ja eine Schätzung von σ^2 (hier Formel (7.6)) zu Hilfe nehmen. Wir formulieren die Nullhypothese gleich allgemein: $\mathcal{H}_0 : \beta_1 = c$, c eine beliebige Zahl; dann gilt unter dieser Nullhypothese, dass

$$T_{n-2} := \frac{\hat{\beta}_1 - c}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n {x'_i}^2}}} = \frac{\hat{\beta}_1 - c}{\sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n {x'_i}^2}}}$$

die t-Verteilung mit $n-2$ Freiheitsgraden hat. Wir werden die $\mathcal{H}_0$-Hypothese verwerfen, wenn diese Teststatistik Werte annimmt, welche weiter als die kritischen Werte von 0 entfernt sind (siehe Tabelle I im Krengel).

Kleiner Einschub/Nachtrag: Wenn man die Definition der t-Verteilung anschaut, wird man feststellen, dass Zähler und Nenner unabhängig sein müssen. Dies ist hier erfüllt (siehe 7.9).

In Statistikpaketen ist der Default derart, dass immer getestet wird, ob $\beta_1 = 0$ oder nicht (in einer Aufgabe kann man zuerst die richtige Gerade abziehen, und dann auf $\beta_1 = 0$ testen).

Mit den Tschernobyl-Daten erhalten wir in R mit den Befehlen "aov" und "summary" die Werte auf dem Datenblatt.

7.1.4 Probleme & Diagnostic Checking

Wir geben hier nur einen kurzen Überblick über mögliche Probleme, Gefahren und die Methoden, welche man unter "Diagnostic Checking" zusammenfasst.

* Ausreisser

* ungleiche Varianzen

* verbleibende Muster (z.B. quadratisch) in den e_i's

* ϵ_i's nicht unabhängig

* ϵ_i's nicht normalverteilt

7.1.5 Warum ist die lineare Regression mit OLS so wichtig, bekannt und erfolgreich?

* wird auch von Nicht-MathematikerInnen/Nicht-StatistikerInnen verstanden

* theoretisch einfach zu berechnen

* einfach auch zur multiplen Regression erweiterbar

* früher war in Statistik-Paketen oft nur diese Regression programmiert (heute kaum mehr als Argument relevant)

warum speziell linear

* Mensch kann nur lineare Zusammenhänge gut erfassen

* viele nichtlineare Abhängigkeiten können durch Transformation zu linearen Problemen gemacht werden (ist aber auch umstritten: "Man foltert die Daten bis sie gestehen"). Vor allem: viele Phänomene mit exponentiellem (oder geometrischem, falls diskret) Wachstum: Wirtschaft (gesamte Volkswirtschaft und einzelne Firmen), Pflanzen (Zellteilung), Ausbreitung Bekanntheitsgrad von Websites, Ausbreitung von Epidemien

warum speziell OLS

* früher EDV-Probleme bei alternativen Vorschlägen (heute kaum relevant)

* OLS ist auch der BLUE (Best Linear Unbiased Estimator), siehe Teil 7.7.

Bis hier ist die Vorlesung obligatorischer Prüfungsstoff - was jetzt kommt ist freiwilliger Prüfungsstoff

7.2 Vorbereitung I: Die multivariate Normalverteilung $\mathrm{MVN}_n(\mu, \mathbf{\Sigma})$

Wir werden die multivariate Normalverteilung schrittweise einführen. Am Schluss gilt: Wir nennen einen Vektor von Zufallsgrössen $(X_1, \ldots, X_n)^t$ mit Mittelwertvektor μ und Kovarianzmatrix $\mathbf{\Sigma}$ ($\Sigma_{ij} := Cov(X_i, X_j)$) multivariat normalverteilt der Dimension n, wenn die Dichte die Form

$$f(x_1, \ldots, x_n) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{1}{\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^t \Sigma^{-1}(\mathbf{x}-\mu)}$$

hat.

Kreative Personen könnten versucht sein, die MVN_n so zu definieren, dass man einfach n $\mathcal{N}(0,1)$-Zufallsgrössen zu einem Vektor zusammenfassen kann. Dies ist so nicht nur falsch, sondern im Risk-Management von Finanzkonzernen sehr gefährlich, wie Kontrastbeispiel 7.2 zeigen wird. Wenn wir hingegen von den einzelnen Zufallsgrössen noch fordern, dass sie unabhängig sind, so erhalten wir den einfachsten Fall einer $\mathrm{MVN}_n(\mu, \mathbf{\Sigma})$-Verteilung, aus dem dann der allgemeine Fall aufgebaut wird:

Definition 7.1 [Plain-Vanilla $\mathrm{MVN}_n(\mathbf{0}, \mathbf{I_n})$] *Sei $Y := (Y_1, \ldots, Y_n)^t$ ein Vektor von iid $\mathcal{N}(0,1)$-Zufallsgrössen. Dann nennen wir Y Standard-Multivariat-Normalverteilt und schreiben dafür $\mathrm{MVN}_n(\mathbf{0}, \mathbf{I_n})$. Die Matrix $\mathbf{I_n}$ ist die (hier triviale) Kovarianzmatrix:*

$$(\mathbf{I_n})_{ij} := E[(Y_i - E[Y_i])(Y_j - E[Y_j])].$$

Die Dichte von Y bzgl. Lebesgue- oder Riemann-Mass ist

$$f(y_1, \ldots, y_n) := \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_i^2} \right) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} e^{-\frac{1}{2}y^t y}. \tag{7.9}$$

Kontrastbeispiel 7.2 Das folgende Beispiel dient als Warnung, was die $\mathrm{MVN}_n(\mu, \mathbf{\Sigma})$ *nicht* ist. In Finanzkonzernen kann nämlich folgende Horror-Situation auftreten und unter Umständen von schlechten Risk-Managern nicht erkannt werden: Viele statistische Analyse-Methoden setzen eine $\mathrm{MVN}_n(\mu, \mathbf{\Sigma})$ voraus. Wenn Daten nicht $\mathrm{MVN}_n(\mu, \mathbf{\Sigma})$-verteilt sind, liefern die Methoden eventuell völlig irreführende Empfehlungen. Man testet

deshalb die Daten zuerst darauf, ob diese $\mathrm{MVN}_n(\mu, \mathbf{\Sigma})$-verteilt sind. Es gibt mindestens in der Theorie folgenden Fall: $(X, Y)^t$ ein zweidimensionaler Zufallsvektor, beide einzelnen Koordinaten je $\mathcal{N}(0, 1)$-verteilt und X und Y sind *unkorreliert*. Der Vektor muss deshalb jedoch nicht zwingend $\mathrm{MVN}_2(\mathbf{0}, \mathbf{I_2})$-verteilt sein. Im Gegenteil kann man Situationen konstruieren, in denen X und Y 100 % von einander abhängig sind und - jetzt kommt die Katastophe - beide nehmen extreme Werte gleichzeitig an (z.B. beide Finanzpapiere machen gleichzeitig Riesenverluste). Wie ist so etwas möglich?

* Konstruktion?

* Skizze der Verteilung im $\mathbb{R}^2$

Diese Verteilung hat sicher nicht die Dichte aus (7.9).

Nach diesem warnenden Kontrastbeispiel werden wir schrittweise die MVN aus Definition 7.1 verallgemeinern:

Invertierbare Transformation von Y

Sei $A \in GL(n, \mathbb{R})$ (Gruppe der invertierbaren $n \times n$-Matrizen mit Koeffizienten aus $\mathbb{R}$), Y sei $\mathrm{MVN}_n(\mathbf{0}, \mathbf{I_n})$-verteilt. Dann hat $Z := AY$ per Definition eine $\mathrm{MVN}_n(\mathbf{0}, \mathbf{AA^t})$-Verteilung. Die Dichte muss wegen Satz 2 (Transformationsformel), § 2, Forster Analysis III, gleich

$$f(z_1, \ldots, z_n) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{1}{\sqrt{\det(AA^t)}} e^{-\frac{1}{2} z^t (AA^t)^{-1} z}$$

sein; beachte $\sqrt{\det(AA^t)} = \sqrt{\det(A)\det(A^t)} = \sqrt{\det(A)\det(A)} = |\det(A)|$, $\det(A^{-1}) = 1/\det(A)$ und $(A^{-1})^t A^{-1} = (AA^t)^{-1}$.

An dieser Stelle vereinbaren wir noch, dass bei Zufalls-Vektoren (wie Z) der Erwartungswert koordinatenweise definiert wird; analog wird bei Zufalls-Vektoren unter der Varianz (z.B. $V[Z]$) eine Kovarianzmatrix verstanden:

$$V[Z] := E[ZZ^t] - E[Z]E[Z^t] \quad (= E[(Z - E[Z])(Z - E[Z])^t]). \tag{7.10}$$

Auf der Diagonalen finden wir also die einzelnen Varianzen ($(V[Z])_{ii} = V[Z_i]$) und neben der Diagonalen die Kovarianzen ($(V[Z])_{ij} = Cov(Z_i, Z_j)$). Die Kovarianzmatrix ist natürlich symmetrisch und wegen (7.10) auch positiv-semi-definit; falls $\mathrm{Rang}(V[Z]) = n$ sogar positiv definit.

Wir berechnen, falls Y eine $\mathrm{MVN}_n(\mathbf{0}, \mathbf{I_n})$-Verteilung besitzt, und $Z = AY$

$E[Z] =$

und auch

$V[Z] =$

Also:

Invertierbare Transformation von Y, Spezialfall der orthogonalen Transformation

Wenn wir ein orthogonales $A^{(o)}$ wählen, so ist ja $A^{(o)}(A^{(o)})^t = \mathbf{I_n}$. Das heisst aber, die Koordinaten von $A^{(o)}Y$ sind stochastisch unkorreliert - in Lemma 7.3 werden wir sehen, dass sie dann sogar unabhängig sind. Wenn wir für $A^{(o)}$ eine Drehung um den Winkel α nehmen:

$$A^{(o)} := \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix},$$

so erhalten wir, dass wenn Y_1, Y_2 zwei iid $\mathcal{N}(0,1)$-Zufallsgrössen sind, dann sind (wähle $Z = A^{(o)}Y$)

$$Z_1 := (\cos\alpha)Y_1 - (\sin\alpha)Y_2$$

und

$$Z_2 := (\sin\alpha)Y_1 + (\cos\alpha)Y_2$$

unabhängig voneinander. Freiwillige Übung: probieren Sie das mit Stichproben in R/S-PLUS aus, indem Sie zumindest die Korrelationen von Z_1 und Z_2 untersuchen. Machen Sie auch Plots dazu (vgl. Aufgabe in den Übungen).

Parallele Verschiebung von AY

Wieder habe Y eine $\mathrm{MVN}_n(\mathbf{0}, \mathbf{I_n})$-Verteilung; A sei invertierbar. Mit $\mu \in \mathbb{R}^n$ fest definieren wir, dass $Z := \mu + AY$ eine $\mathrm{MVN}_n(\mu, \mathbf{AA^t})$-Verteilung besitzt.

Nicht mehr zwingende Invertierbarkeit der Transformation von Y

Wieder sei Y eine $\mathrm{MVN}_n(\mathbf{0}, \mathbf{I_n})$-Zufallsgrösse. Jetzt ist A aber nicht mehr invertierbar, sondern eine $r \times n$-Matrix mit vollem Rang $r, r < n$.

A hat r linear unabhängige Zeilen. Mit $< A >$ bezeichnen wir den von A erzeugten Raum ($\subset \mathbb{R}^n$). Wir können für $< A >$ eine Orthonormalbasis angeben, $u_1, \ldots, u_r$. Diese kann mit $u_{r+1}, \ldots, u_n$ (erzeugt $< A >^\perp$, orthogonales Komplement) zu einer orthonormalen Basis des $\mathbb{R}^n$ ergänzt werden:

$$\mathbb{R}^n = < A > \oplus < A >^\perp .$$

Wir bauen jetzt die $n \times n$-Matrix Q derart auf, dass Zeile j aus u_j^t besteht, $1 \leq j \leq n$. Es existiert eine $r \times n$-Matrix $L = (\text{etwas}|0)$, wo bis zum "|" r-Spalten sind, sodass

$$A = LQ.$$

Für die i-te Zeile von A gilt also zum Beispiel (a_i liegend, da Zeile)

$$a_i = \sum_{s=1}^{r} \lambda_{is} u_s^t.$$

Jetzt gilt aber:

$$AY = LQY = L[QY],$$

wobei QY eine $\text{MVN}_n(\mathbf{0}, \mathbf{I_n})$-Zufallsgrösse ist. Behauptung: AY hat eine $\text{MVN}_r(\mathbf{0}, \mathbf{LL^t})$-Verteilung. Warum?

Diese Kovarianzmatrix ist aber denkbar unpraktisch: sie müsste jedesmal berechnet werden. Gott sei Dank gilt aber:

$$AA^t = (LQ)(LQ)^t = LQQ^tL^t = L\mathbf{I_n}L^t = LL^t,$$

damit gilt abschliessend: AY hat eine $\text{MVN}_r(\mathbf{0}, \mathbf{AA^t})$-Verteilung.

*

* *

Nachdem wir damit den allgemeinsten Fall der MVN definiert haben, fügen wir noch wichtige Resultate hinzu. Die Beweise benutzen immer wieder die gleichen Methoden, sodass wir nicht alle Sätze beweisen - dies gilt auch für Teil 7.3. Die Beweise finden sich in C.R. Rao: Linear Statistical Inference and Its Applications, pp 185-189 & 519-527.

Lemma 7.3 [unkorreliert $\Leftrightarrow$ unabhängig, wenn MVN!] *Sei Y MVN-verteilt. Dann sind die einzelnen Koordinaten von Y genau dann alle unkorreliert, wenn sie alle unabhängig sind.*

Beweis von Lemma 7.3: In der Klasse:

□

Satz 7.4 [Orthogonalität und Unabhängigkeit] *Y habe eine* $\mathrm{MVN}_n(\mathbf{0}, \mathbf{I_n})$*-Verteilung. Sei $(A_i)_{i=1}^k$ eine Folge von k Matrizen mit jeweils n Spalten; Matrix A_i habe r_i Zeilen und Rang r_i, $1 \leq i \leq k$. Die Matrizen seien orthogonal zueinander in dem Sinne, dass für alle $1 \leq i \neq j \leq k$ gilt:*

$$A_i A_j^t = 0,$$

dabei ist "0" die $r_i \times r_j$-Nullmatrix. Dann sind die k Zufallsvektoren

$$A_1 Y, A_2 Y, \ldots, A_k Y$$

alle unabhängig voneinander (nicht jedoch zwingend auch die Koordinaten in den einzelnen $A_i Y$).

Beweis von Satz 7.4: Wieder können wir den $\mathbb{R}^n$ als direkte Summe der von den einzelnen Matrizen aufgespannten Räume darstellen (allenfalls noch notwendige orthogonale Ergänzung ($< A_* >$):

$$\mathbb{R}^n = < A_1 > \oplus < A_2 > \ldots \oplus < A_k > \oplus < A_* > .$$

Wir wählen zudem für jedes $< A_i >$ eine orthonormale Basis; total für den ganzen $\mathbb{R}^n$ die $u_1, \ldots, u_n$. Die Matrix Q bauen wir mit Zeilen u_i auf: $Q_{i\cdot} := u_i^t$. Dann gilt:

$$\begin{pmatrix} A_1 \\ A_2 \\ . \\ . \\ A_k \end{pmatrix} = \begin{pmatrix} M_1 & 0 & \ldots & 0 \\ 0 & M_2 & \ldots & 0 \\ 0 & \ldots & \ldots & 0 \\ 0 & \ldots & \ldots & 0 \\ 0 & \ldots & M_k & 0 \end{pmatrix} Q$$

mit M_i eine $r_i \times r_i$-Matrix. Es gilt für jedes $1 \leq i \leq k$: $A_i Y = M_i (QY)^{(i)}$, wobei

$$QY = \begin{pmatrix} (QY)_1 \\ (QY)_2 \\ . \\ . \\ (QY)_{r_1} \\ (QY)_{r_1+1} \\ . \\ . \\ . \\ (QY)_n \end{pmatrix}.$$

Mit $(QY)^{(i)}$ bezeichnen wir jetzt den Vektor $((QY)_{r_1+\ldots+r_{i-1}+1}, \ldots, (QY)_{r_1+\ldots+r_i})^t$. Weil Q orthogonal ist, ist QY eine $\mathrm{MVN}_n(\mathbf{0}, \mathbf{I_n})$-Zufallsgrösse. Also sind auch die

$$A_1 Y, A_2 Y, \ldots, A_k Y = M_1 (QY)^{(1)}, M_2 (QY)^{(2)}, \ldots, M_k (QY)^{(k)}$$

jeweils unabhängig - sie greifen auf verschiedene, unabhängige Zufallsgrössen $((QY)_i)_{i=1}^n$ zurück.

□

Wir bringen noch eine lose Sammlung von Resultaten - ohne Beweis:

Satz 7.5 [Äquivalente Definition der MVN] *Ein Vektor Y ist genau dann* MVN_n*, wenn für jeden Vektor $a \in \mathbb{R}^n$ gilt: $a^t Y$ ist eindimensional normalverteilt.*

Wegen Satz 7.5 ist der bivariate Zufallsvektor aus Kontrastbeispiel 7.2 nicht MVN_2.

Satz 7.6 [Summe von MVN] *Sei X eine* $\mathrm{MVN}_n(\mu_\mathbf{1}, \mathbf{\Sigma_1})$*-Zufallsgrösse und Y eine* $\mathrm{MVN}_n(\mu_\mathbf{2}, \mathbf{\Sigma_2})$*-Zufallsgrösse; $X \coprod Y$. Dann gilt: $X + Y$ ist* $\mathrm{MVN}_n(\mu_\mathbf{1} + \mu_\mathbf{2}, \mathbf{\Sigma_1} + \mathbf{\Sigma_2})$*-verteilt.*

Satz 7.7 [nochmals mit Matrix ran] *Sei X eine $\mathrm{MVN}_n(\mu, \mathbf{\Sigma})$-Zufallsgrösse. Sei B eine $k \times n$-Matrix und η ein $k \times 1$-Vektor. Dann hat*

$$Y := \eta + BX$$

eine $\mathrm{MVN}_k(\eta + \mathbf{B}\mu, \mathbf{B\Sigma B^t})$-Verteilung.

*

* *

Wir wollen die Niveaulinien im Fall $n = 2$ ein bisschen üben; mehr dazu in einer Aufgabe.

7.3 Vorbereitung II: Quadratische Formen und $\mathrm{MVN}_n(\mu, \mathbf{\Sigma})$

In diesem Teil werden wir uns fragen, wie Ausdrücke der Art $Y^t UY$ verteilt sind, wenn U symmetrisch und idempotent ist.

Wir repetieren kurz von 1.4.2.6: Wenn $(X_i)_{i=1}^n$ iid $\mathcal{N}(0,1)$-verteilt sind, dann ist

$$\sum_{i=1}^{n} X_i^2$$

χ_n^2-verteilt. $E[\chi_n^2] = n; V[\chi_n^2] = 2n$. Dazu noch $E[X_1^1] = E[X_1^3] = 0; E[X_1^2] = 1, E[X_1^4] = 3$ (letzteres wurde nicht gezeigt). Kleine Kontrolle: $V[\chi_1^2] = E[(X_1^2)^2] - (E[X_1^2])^2 = 3 - 1^2 = 2$; allgemein $V[\chi_n^2] = 2n$ (wegen "Varianz der Summe ist Summe der Varianzen" bei Unabhängigkeit). Wir definieren jetzt

Definition 7.8 [Nichtzentrale χ^2-Verteilung] *Sei Y_1 eine $\mathcal{N}(\mu, 1)$-Zufallsgrösse und $Y_2, \ldots, Y_n$ seien je iid $\mathcal{N}(0,1)$-Zufallsgrössen. Y_1 sei auch unabhängig von den restlichen $Y_i, 2 \leq i \leq n$. Dann nennen wir die Verteilung von*

$$\sum_{i=1}^{n} Y_i^2$$

nichtzentral χ_n^2 und schreiben dafür χ_{n,μ^2}^2. Die herkömmliche χ_n^2-Verteilung erhalten wir wenn $\mu = 0$.

Berechnen Sie:

$E[\chi_{n,\mu^2}^2] =$

und $V[\chi_{n,\mu^2}^2] =$

Satz 7.9 [Allgemeinere Definition hat auch χ^2_{n,μ^2}-Verteilung] *Sei $(Z_i)_{i=1}^n$ eine Folge von unabhängigen Zufallsgrössen, sodass Z_i eine $\mathcal{N}(\mu_i, 1)$-Verteilung hat. Die Zufallsgrösse*

$$Y := \sum_{i=1}^{n} Z_i^2$$

hat dann eine $\chi^2_{n,\mu^t\mu}$-Verteilung.

Beweis von Satz 7.9: Für $1 \leq i \leq n$ gilt $Z_i = \mu_i + N_i$, wo die N_i iid $\mathcal{N}(0,1)$-verteilt sind. Sei Q eine orthogonale $n \times n$-Matrix (wird später noch speziell gewählt). Wir haben

$$Y = Z^t Z = Z^t(Q^tQ)Z = (Z^tQ^t)(QZ) = (QZ)^t(QZ).$$

Wie ist die Verteilung von QZ? $Z = \mu + N$, wo N eine $\mathrm{MVN}_n(\mathbf{0}, \mathbf{I_n})$-Zufallsgrösse ist.

$$QZ = Q(\mu + N) = Q\mu + QN = Q\mu + N',$$

wo N' ebenfalls eine $\mathrm{MVN}_n(\mathbf{0}, \mathbf{I_n})$-Zufallsgrösse ist. Jetzt wählen wir Q so, dass $Q\mu = |\mu|\vec{e}_1$ ($|\mu| := \sqrt{\mu_1^2 + \ldots + \mu_n^2}$). Damit ist aber

$$Q\mu + N' = \begin{pmatrix} Z_1' \\ Z_2' \\ . \\ . \\ Z_n' \end{pmatrix},$$

wo $Z_1' = |\mu| + N_1'$ und für $2 \leq i \leq n$ gilt $Z_i' = N_i'$. Damit haben wir

$$Y = Z^t Z = Z^t(Q^tQ)Z = (Z^tQ^t)(QZ) = (QZ)^t(QZ) = \sum_{i=1}^{n}(Z_i')^2.$$

Dieser Ausdruck hat aber genau eine $\chi^2_{n,|\mu|^2}$-Verteilung und mit $|\mu|^2 = \mu^t\mu$ erhalten wir das gewünschte Resultat.

□

Satz 7.10 [Verteilung von Y^tUY, U symmetrisch und idempotent] *Sei Y eine $\mathrm{MVN}_n(\mu, \mathbf{I_n})$-Zufallsgrösse. Die $n \times n$-Matrix U sei symmetrisch und idempotent mit $Rang(U) = k \leq n$. Dann hat*

$$Y^tUY$$

eine $\chi^2_{k,\mu^tU\mu}$-Verteilung.

Bemerkung zu Satz 7.10: Dies ist eine Verallgemeinerung von Satz 7.9. Wenn der $\mathrm{Rang}(U) = n$, so haben wir zwangsweise $U = \mathbf{I_n}$ und damit den Fall von Satz 7.9. Warum gilt bei $\mathrm{Rang}(U) = n$ immer $U = \mathbf{I_n}$?

Beweis von Satz 7.10: Von einer Aufgabe wissen wir, dass die Eigenwerte von idempotenten Matrizen entweder 0 oder 1 sind. Von einer anderen Aufgabe und 6.6.3 folgt, dass wir k $(= Rang(U))$ mal Eigenwert 1 haben und $n - k$ mal Eigenwert 0.

Wegen 6.6.1 finden wir eine orthogonale Matrix Q, so dass

$$Q^tUQ = \Lambda := \begin{pmatrix} \mathbf{I_k} & 0 \\ 0 & 0 \end{pmatrix}.$$

Damit haben wir

$$Y^tUY = Y^tQQ^tUQQ^tY = (Y^tQ)Q^tUQ(Q^tY) = (Q^tY)^t\Lambda(Q^tY) = \sum_{i=1}^{k}[(Q^tY)_i]^2. \quad (7.11)$$

Wegen Satz 7.7 (oder direkt wegen Entwicklung allgemeinster Fall MVN) hat Q^tY eine $\mathrm{MVN}_n(Q^t\mu, Q^tQ)$-Verteilung; wegen $Q^tQ = \mathbf{I_n}$ sogar eine $\mathrm{MVN}_n(Q^t\mu, \mathbf{I_n})$-Verteilung. Damit ist die rechte Seite von (7.11) eine Summe von *unabhängigen*, quadrierten $\mathcal{N}((Q^t\mu)_i, 1)$-Zufallsgrössen. Wegen Satz 7.9 hat (7.11) deshalb eine $\chi^2_{k,\sum_{i=1}^{k}[(Q^t\mu)_i]^2}$-Verteilung. Weil

$$\mu^tU\mu = \mu^tQQ^tUQQ^t\mu = (\mu^tQ)Q^tUQ(Q^t\mu) = (Q^t\mu)^t\Lambda(Q^t\mu) = \sum_{i=1}^{k}[(Q^t\mu)_i]^2$$

ist der Satz bewiesen.

□

Mit ähnlichen Argumenten beweist man noch:

Satz 7.11 [Orthogonalität und Unabhängigkeit bei quadratischen Formen in MVN] *Sei Y eine $\mathrm{MVN}_n(\mu, \mathbf{I}_n)$-Zufallsgrösse. Sei $U = U_1 + \ldots + U_m$ eine Summe von $n \times n$-Matrizen derart, dass die $U_j, 1 \leq j \leq m$, symmetrisch und idempotent sind und $U_i U_l = 0$ für alle $1 \leq i \neq l \leq m$. Dann gilt:*

$$Y^t U_1 Y, \ldots, Y^t U_m Y$$

sind alle unabhängig voneinander und für $1 \leq j \leq m$ gilt:

$$Y^t U_j Y$$

ist $\chi^2_{\mathrm{Rang}(U_j), \mu^t U_j \mu}$-verteilt. U kommt in den Anwendungen vor; für den Satz selber ist U nicht von Belang.

Satz 7.12 [Verteilung des Rests] *Sei Y eine $\mathrm{MVN}_n(\mu, \mathbf{I}_n)$-Zufallsgrösse. Seien U und V je symmetrische und idempotente $n \times n$-Matrizen derart, dass $UV = V$. Wir definieren $Z := Y^t U Y - Y^t V Y = Y^t (U - V) Y$ (gebraucht als $Y^t U Y = Y^t V Y + Z$). Dann gilt: Z ist unabhängig von $Y^t V Y$ und Z ist $\chi^2_{\mathrm{Rang}(U) - \mathrm{Rang}(V), \mu^t U \mu - \mu^t V \mu}$-verteilt.*

Corollar zu Satz 7.12 [$\sum_{i=1}^n (X_i - \bar{X})^2$ ist χ^2_{n-1}-verteilt] *Sei $X_1, \ldots X_n$ eine Folge von iid $\mathcal{N}(\mu, \sigma^2)$-Zufallsgrössen. Dann hat*

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}$$

eine χ^2_{n-1}-Verteilung.

Beweis von Corollar zu Satz 7.12: (in Klasse)

□

7.4 Multiple Regression

7.4.1 Die Modellannahmen

Wir werden jetzt eine etablierte Methode kennenlernen, mit der wir

einfache Regression: eine (lineare) Beziehung zwischen 2 Variablen untersuchen können (z.B. Punkte in Mathematik und Punkte in Physik von der gleichen Person). Zu Recht ist man an die Korrelation von 1.3.3 erinnert. Im (theoretischen) Modell

$$Y_i = \alpha + \beta x_i + \epsilon_i \qquad (7.12 \text{ light})$$

werden wir die x-Variable als fest betrachten und versuchen, die Y-Variable möglichst genau durch x vorher zu sagen oder durch x zu erklären (stören wird uns dabei der Störterm ϵ). Es wird dann mit Daten $(x_i, y_i), 1 \leq i \leq n$, darum gehen, α, β und $\sigma^2 (= V[\epsilon])$ zu schätzen und zu testen, ob nicht z.B. $\beta = 0$ gilt.

multiple Regression: in Verallgemeinerung der einfachen Regression die Y-Variable (Response-Variable) durch mehrere erklärende Variablen (Regressors, Predictors) möglichst genau beschreiben können. Das theoretische Modell wird dann zu

$$Y = A\beta + \epsilon. \qquad (7.12)$$

Dabei ist Y ein n-dimensionaler Zufallsvektor, A eine $n \times k$-Matrix (Struktur-, Design- oder Daten-Matrix) von Rang k (Idealfall), β der k-dimensionale Parametervektor und ϵ ein $\text{MVN}_n(0, \sigma^2 \mathbf{I_n})$-Zufallsvektor (vgl. 7.1; die $\epsilon_i, 1 \leq i \leq n$, sind iid $\mathcal{N}(0, \sigma^2)$-Zufallsgrössen). Bekannt sind in (7.12) Realisationen $y_1, \ldots, y_n$, die Datenmatrix A; unbekannt ist β und die Realisation des Fehlers ϵ sowie das σ^2. Auch hier geht es darum, β, σ^2 zu schätzen (Teil 7.5) und zu testen, ob ganz β oder (häufiger) bestimmte β_i Null sind oder nicht (Teil 7.6).

Häufig wird die erste Spalte von A aus lauter 1-ern bestehen. Damit wird die Verallgemeinerung von (7.12 light) zu (7.12) deutlich. Es gilt nämlich für Datenpunkt Y_i wo $x_{ij} := A_{ij}$

$$Y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i.$$

Was ist ϵ? Im Fehlerterm ϵ (auch Residual) fassen wir verschiedene Effekte wie Rundungsfehler, Messfehler, zufällige Schwankungen und kleinste Einflüsse, welche man nicht in's Modell einbauen will, zusammen.

Kleine Aufgaben:

* Wie sieht $A, \vec{\beta}$ aus, damit wir eine einfache Regression haben?

* Wie sieht A, β aus, damit wir eine einfache Varianz-Analyse haben?

* Wie ist $E[Y_i]$?

* Wie ist die Verteilung von Y_i?

7.4.2 "Bottom Up" und "Top Down"

Schema zur Wahl von Modellen / Welche Regressoren soll man berücksichtigen?

7.4.3 Welches ist "Das richtige Modell"?

Nach diesen verschiedenen Modellen fragen wir uns vielleicht: "Welches ist jetzt *das richtige Modell*?" Ausser in Simulationen, wo man genau weiss, woher die Daten stammen, ist diese Frage nicht zu beantworten. Die verschiedenen möglichen Methoden und Modelle sollten uns aber zumindest dahingehend vorsichtig machen, dass wir uns bewusst sind, dass man mit anderen Modellen andere Resultate erhält.

In den Sozialwissenschaften (Soziologie, Psychologie und auch Ökonomie) gilt diese Relativierung sehr stark (wann heiraten Menschen?). In denjenigen Naturwissenschaften, in denen die Versuchsbedingungen einigermassen vollständig kontrolliert werden können, ist jedoch zu erwarten, dass man so etwas wie ein richtiges Modell finden kann (Atomzerfall, Wachstum einer Bakterienpopulation im Labor). Bei der Medizin ist zu bemerken, dass die untersuchten Menschen in permanent ändernden Bedingungen leben (und nicht im Labor). Deshalb wird es dort je nach Untersuchungsgegenstand kaum universelle Modelle geben.

7.5 Schätzen im Regressionsmodell: OLS und MLE

In der Vorlesung "Einführung in die Numerik" im 2. Semester haben Sie bereits (nach der Konditionszahl eines linearen Gleichungssystems) das lineare Ausgleichsproblem behandelt (Methode der kleinsten Fehlerquadrate (engl. Ordinary Least Squares OLS)). Die Aussagen dort bleiben natürlich richtig (insbesondere bzgl. numerischer Probleme (Konditionszahl) etc.); wir ändern ein bisschen Sichtweise und Motivation.

Wir werden jetzt 2 Methoden kennenlernen (OLS und MLE), mit denen man die unbekannten Parameter in (7.12) schätzen kann. In diesem einfachen Modell werden die Schätzungen $\hat{\beta}$ für β gleich sein, egal welche Methode wir anwenden. Es gibt jedoch einen wichtigen (philosophischen) Unterschied: bei der OLS-Schätzung werden wir gar keine Modellannahmen (ausser linear) machen müssen. Wir *definieren* $\hat{Y} := A\hat{\beta}$. Die Aufgabe ist lediglich: suche $\hat{\beta}_1, \ldots, \hat{\beta}_k$ derart, dass die Summe der quadrierten Fehler (Sum of Squared Errors)

$$SSE := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 := \sum_{i=1}^{n} (y_i - (A\hat{\beta})_i)^2$$

minimal ist.

OLS-Schätzung

Schreiten wir jetzt zur OLS-Schätzung: Wir haben die Summe

$$\sum_{i=1}^{n} (y_i - (A\beta)_i)^2$$

bzgl. β zu minimieren. Ich verweise hier auf die bereits genannte Vorlesung "Einführung in die Numerik". Man erhält durch ableiten nach den einzelnen $\beta_j, 1 \leq j \leq k$, erstmal

$$\frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} (y_i - (A\beta)_i)^2 = 2 \sum_{i=1}^{n} (y_i - (A\beta)_i) x_{ij} \quad .$$

Durch Nullsetzen und Aggregieren erhalten wir die Normalgleichungen (Synonymgebrauch von Zufallsgrössen Y und Datenpunkten y)

$$A^t Y = A^t A\hat{\beta}. \qquad \text{(Normalgleichungen)}$$

Von 6.5: wenn A vollen Rang k hat, so ist A^tA invertierbar und wir erhalten

$$\hat{\beta} = (A^tA)^{-1}A^tY.$$

In der Vorlesung "Einführung in die Numerik" wurde gezeigt, dass diese Lösung eindeutiges Minimum ist.

Uns fehlt noch ein Schätzer für σ^2. Eine Möglichkeit ist, dass wir mit den $e_i := y_i - \hat{y}_i, 1 \leq i \leq n$, den sogenannten *beobachteten Residuen* (die wahren sind ja unbeobachtet), arbeiten:

$$\hat{\sigma}^2 := \frac{1}{n-k}\sum_{i=1}^{n} e_i^2. \tag{7.13}$$

Warum "$n-k$" im Nenner? Angelernte StatistikerInnen sagen dazu: wir haben ja k Parameter (im β) geschätzt, es gehen also k Freiheitsgrade verloren. Dieser Schätzer für σ^2 ist erwartungstreu, wie wir weiter hinten noch zeigen werden. Wir beweisen in einer Aufgabe, dass auch $\hat{\beta}$ erwartungstreu ist.

ML-Schätzung

Wie bereits angekündigt, ist der MLE für β gleich wie beim OLS-Schätzer. Wir werden die Berechnung des MLE von β auf die Berechnung des OLS-Schätzers zurückführen.

Wegen (7.12) ist Y_i normalverteilt mit Erwartungswert $(A\beta)_i$ und Varianz σ^2:

$$Y_i \sim \mathcal{N}((A\beta)_i, \sigma^2).$$

Die Likelihood (vgl. Definition 5.7) ist also

$$\prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\exp\left[-\frac{1}{2\sigma^2}(y_i - (A\beta)_i)^2\right].$$

Wir berechnen die Log-Likelihood

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (A\beta)_i)^2 - \frac{n}{2}\log\sigma^2 + c, \tag{7.14}$$

wobei c eine Konstante ist. Da wir diesen Ausdruck maximieren wollen, können wir einfach (egal, wie wir σ^2 schätzen!) für die Schätzung von β den Ausdruck

$$\sum_{i=1}^{n}(y_i - (A\beta)_i)^2$$

minimieren. Aber das ist ja unser Ausgangsproblem bei der OLS-Schätzung. Deshalb müssen die Schätzer gleich sein wie bei der OLS-Methode. Aber wie ist der MLE von σ^2 in Modell (7.12)? Kleine HA: setzen Sie dazu obiges $\hat{\beta}$ in (7.14) ein und maximieren Sie bzgl. σ^2. Man erhält dann den MLE für σ^2:

$$(\hat{\sigma}^2)^{MLE} = \frac{1}{n}\sum_{i=1}^{n} e_i^2,$$

zu vergleichen mit (7.13). Zusammengefasst: die MLE und OLS-Schätzung für β ist gleich und die Schätzung für σ^2 ist bis auf den Faktor $n/(n-k)$ gleich.

Geometrische Interpretation für Schätzproblem:

Rechtfertigung dieser Interpretation durch folgende Turnübungen **T**

Mit **Notationen** (siehe auch Aufgabenblatt und Kapitel 6):

$H := A(A^tA)^{-1}A^t$
$M := \mathbf{I_n} - H$
$\hat{\beta} := (A^tA)^{-1}A^tY$
$\hat{Y} := A\hat{\beta}$
$e := Y - \hat{Y}$

folgende kleine **Turnübungen**:

T 1: H und M sind idempotent und symmetrisch, $HA = A$ und H und M stehen orthogonal zueinander: $HM = 0$ (Aufgabe in den Übungen).

T 2:

$$\hat{\beta} = (A^tA)^{-1}A^tY = (A^tA)^{-1}A^t(A\beta + \epsilon) = \beta + (A^tA)^{-1}A^t\epsilon$$

T 3: bin schon in A

$$MA = (\mathbf{I_n} - H)A = A - A = 0$$

T 4:

$$\begin{aligned} e = Y - \hat{Y} = Y - A\hat{\beta} = Y - A(A^tA)^{-1}A^tY = (\mathbf{I_n} - H)Y &= MY \\ &= M(A\beta + \epsilon) \\ &= M\epsilon \end{aligned}$$

T 5: H projiziert Y auf die Ebene $A\mathbb{R}^k$

$$\hat{Y} = A\hat{\beta} = A(A^tA)^{-1}A^tY = HY$$

T 6:

$$e^te = (M\epsilon)^tM\epsilon = \epsilon^tM^tM\epsilon = \epsilon^tM\epsilon$$

T 7:

$$e^t e = (MY)^t MY = Y^t M^t MY = Y^t MY$$

T 8:

$$(\hat{Y})^t \hat{Y} = (HY)^t HY = Y^t H^t HY = Y^t HY$$

T 9: Geschätzte Residuen orthogonal zu A. Wir haben alles aus A herausgeholt, um Y zu schätzen. Wäre **T 9** verletzt, wären wir noch nicht ganz optimal.

$$e^t A = (M\epsilon)^t A = \epsilon^t MA = 0$$

T 10:

$$e^t \hat{Y} = \epsilon^t MA\hat{\beta} = 0$$

T 11: **Jargon:** "Die Variation in den Y ($Y^t Y$) lässt sich aufspalten in einen Anteil, der durch die Regression erklärt wird ($\hat{Y}^t \hat{Y}$) und eine residuale Summe ($e^t e$)."

$$Y^t Y = (\hat{Y} + e)^t (\hat{Y} + e) = \hat{Y}^t \hat{Y} + \hat{Y}^t e + e^t \hat{Y} + e^t e = \hat{Y}^t \hat{Y} + e^t e$$

T 12: H ist die orthogonale Projektion auf $A\mathbb{R}^k$ (vgl. **T 5**). Sei dazu $y^{(1)} \in A\mathbb{R}^k$, d.h. $\exists u \in \mathbb{R}^k : y^{(1)} = Au$. Dann gilt:

$$Hy^{(1)} = A(A^t A)^{-1} A^t Au = Au = y^{(1)}. \qquad \text{(T 12.1)}$$

Sei $y^{(2)} \in (A\mathbb{R}^k)^\perp$, d.h. $(y^{(2)})^t Au = 0 \ \forall u \in \mathbb{R}^k$, d.h. $u^t A^t y^{(2)} = 0 \ \forall u \in \mathbb{R}^k$, d.h. $A^t y^{(2)} = 0$. Dann gilt:

$$Hy^{(2)} = A(A^t A)^{-1} A^t y^{(2)} = 0. \qquad \text{(T 12.2)}$$

Zum Schluss (auch als Vorbereitung für 7.6) wollen wir noch kurz den Schätzer $\hat{\beta}$ und $(\hat{\sigma}^2)^{MLE}$ als Zufallsgrösse auffassen:

$\hat{\beta}$:

$(\hat{\sigma}^2)^{MLE}$:

7.6 Testen im Regressionsmodell

7.6.1 Gesamter Vektor $\vec{\beta} = \vec{0}$?

Repetition **T 11**: **Jargon:** "Die Variation in den Y (Y^tY) lässt sich aufspalten in einen Anteil, der durch die Regression erklärt wird ($\hat{Y}^t\hat{Y}$) und eine residuale Summe (e^te)."

$$Y^tY = \hat{Y}^t\hat{Y} + e^te$$

Mit Hilfe dieser Turnübung lässt sich einfach ein Test konstruieren. Dazu betrachten wir nochmals die Skizze mit der Interpretation der OLS-Schätzung von β als orthogonale Projektion von Y auf die Ebene $A\mathbb{R}^k$:

Die Entwicklung der Teststatistik geschieht dann in 3 Schritten (vgl. auch Kapitel 4.5 ANOVA):

1. Die Ausspaltung in **T 11** ist derart, dass wir auf der rechten Seite *unabhängige* Summanden haben.
2. Die einzelnen Summanden sind χ^2-verteilt (mit noch zu bestimmenden Parametern)
3. Die beiden Summanden auf der rechten Seite von **T 11** werden wir (richtig normiert) in Zähler und Nenner einer Statistik einsetzen, welche im Fall $\vec{\beta} = 0$ F-verteilt ist (mit noch zu bestimmenden Parametern). Ein allfälliges $\sigma^2 \neq 1$ kürzt sich heraus.

Wir werden **T 11** zuerst noch umschreiben; unter Verwendung von $H := A(A^tA)^{-1}A^t$, **T 7** und **T 8** folgt:

$$Y^tY = Y^tHY + Y^tMY = Y^tHY + Y^t(\mathbf{I_n} - H)Y. \qquad \text{(Streuungszerlegung)}$$

Unabhängige Summanden: Dies folgt aus Satz 7.12 (wir wählen $U := \mathbf{I_n}, V := H$ und teilen Y durch σ).

χ^2-verteilte Summanden mit Parametern...: Wenn wir den ersten Summanden durch σ^2 teilen, ist er wegen Satz 7.10 (wir wählen $U := H$) $\chi^2_{k,\frac{\vec{\beta}^tA^tHA\vec{\beta}}{\sigma^2}} = \chi^2_{k,\frac{\vec{\beta}^tA^tA\vec{\beta}}{\sigma^2}}$-verteilt. Wenn wir den zweiten Summanden durch σ^2 teilen, ist er χ^2_{n-k}-verteilt (siehe Schluss von 7.5).

Der F-Test: Wenn $\vec{\beta} = \vec{0}$ ($\mathcal{H}_0$-Hypothese), dann ist Y^tHY (der noch nicht normierte Zähler) bis auf den Faktor σ^2 χ^2_k-verteilt, sonst Nichtzentral χ^2; also tendenziell mit grösseren Werten. Wir werden also die Teststatistik

$$F_{k,n-k} := \frac{Y^tHY/k}{Y^t(\mathbf{I_n} - H)Y/(n-k)}$$

benutzen und im Fall von grösseren Werten $\vec{\beta} = \vec{0}$ verwerfen.

In einer einfachen und freiwilligen Hausaufgabe überprüfe man bitte, dass wegen $(t_n)^2 = F_{1,n}$ im Fall $k = 1 : Y_i = \alpha + \epsilon_i$, obiger Test wie erwartet zu einem (quadrierten) t-Test wird. Damit ist die Theorie aus Kapitel 4.4 "$\mu = 0$? bei unbekannter Varianz" offenbar einfach ein Speziallfall der Regression.

Jargon: Die Summen in Gleichung (Streuungszerlegung) werden manchmal mit

$$\text{Totalquadratsumme} = \text{Behandlungsquadratsumme} + \text{Restquadratsumme}$$

oder

$$\text{Totalquadratsumme} = \text{Strukturquadratsumme} + \text{Restquadratsumme}$$

bezeichnet.

Wir haben oben zugeschnitten ein Resultat eingesetzt, welches für sich noch allgemein formuliert werden sollte:

Corollar 7.13 $[\hat{\beta} \coprod \hat{\sigma}^2]$ *In Modell (7.12) sind die Schätzer*

$$\hat{\beta} := (A^t A)^{-1} A^t Y$$

und (vgl. **T 6***)*

$$K\hat{\sigma}^2 := \sum_{i=1}^{n} e_i^2 = \epsilon^t M^t M \epsilon = |M\epsilon|^2$$

stochastisch unabhängig voneinander. Für K wählen wir (meist) entweder $(n-k)$ oder n.

Bemerkung zu Corollar 7.13: Die Schätzungen der einzelnen Komponenten in $\hat{\beta}$ sind jedoch nicht zwingend unabhängig voneinander, vgl. Schluss von 7.5.

Beweis von Corollar 7.13: Von **T 1, T 2** haben wir

$$\hat{\beta} = \beta + (A^t A)^{-1} A^t \epsilon = \beta + (A^t A)^{-1} A^t H \epsilon =: u(H\epsilon)$$

und

$$K\hat{\sigma}^2 = |M\epsilon|^2 =: v(M\epsilon)$$

für Funktionen u, v. Wir haben wegen **T 1** $HM = 0$. Es gilt $r(H) = k$ und $r(M) = n - k$ (kleine HA). Wir können jetzt ähnlich wie in Satz 7.4 vorgehen:

Damit folgt, dass $H\epsilon$ und $M\epsilon$ unabhängig sind. Damit gilt auch $\hat{\beta} \coprod \hat{\sigma}^2$.

□

7.6.2 Einzelne $\beta_i = 0$, $1 \leq i \leq k$?

Wir wollen jetzt einen Test für "$\beta_i = 0$?" entwickeln: Dazu repetieren wir vom Schluss von 7.5:

$$\hat{\beta} \sim \mathrm{MVN}_k(\beta, \sigma^2(A^tA)^{-1}).$$

Damit gilt:

$$\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2[(A^tA)^{-1}]_{ii}).$$

Dummerweise ist σ^2 unbekannt. Wir können aber unsere Schätzung aus (7.13) nehmen:

$$\hat{\sigma}^2 := \frac{1}{n-k} Y^t M Y = \frac{1}{n-k} Y^t (\mathbf{I_n} - H) Y.$$

Damit haben wir als Teststatistik

$$T_i := \frac{\hat{\beta}_i}{\sqrt{\frac{1}{n-k} Y^t (\mathbf{I_n} - H) Y [(A^tA)^{-1}]_{ii}}}.$$

Die Verteilung unter $\mathcal{H}_0 : \beta_i = 0$ ist eine t_{n-k}-Verteilung; warum (3 Sachen)?

7.7 OLS ist BLUE (Best Linear Unbiased Estimator)

Dies ist formalisiert im

Theorem 7.14 [Gauss-Markov-Theorem: "OLS ist BLUE"] *Im Modell (7.12): $Y = A\beta + \epsilon$ ist der Schätzer $\hat{\beta} := (A^tA)^{-1}A^tY$ für β optimal in dem Sinne, dass alle anderen in Y linearen, erwartungstreuen Schätzer $\tilde{\beta}$ eine "schlechtere" Kovarianzmatrix haben: für alle $a \in \mathbb{R}^k$ gilt*

$$a^tV(\hat{\beta})a \leq a^tV(\tilde{\beta})a.$$

Mit $a = \vec{e}_i, 1 \leq i \leq k$, kanonischer Einheitsvektor, ist insbesondere die Schätzung jeder Koordinate von β durch $\hat{\beta}$ von kleinstmöglicher Varianz.

Beweis von Theorem 7.14: Damit ein (Konkurrenz-) Schätzer $\tilde{\beta}$ linear in Y ist, muss er eine Darstellung mit einer $k \times n$-Matrix B (B wie besser - er meint er sei besser) haben: $\tilde{\beta} := BY$ (bei OLS gilt $B := (A^tA)^{-1}A^t$). Damit ein solcher linearer Schätzer auch erwartungstreu ist muss gelten:

$$\beta = E[\tilde{\beta}] = E[BY] = E[B(A\beta + \epsilon)] = BA\beta.$$

Dies muss für alle β gelten. Damit muss

$$BA = \mathbf{I_k} \tag{7.15}$$

sein.

Wir berechnen jetzt die Kovarianzmatrix von BY unter Verwendung von (7.15):

$$\begin{aligned} V(BY) &= E[(BY - \beta)(BY - \beta)^t] \\ &= E[(B(A\beta + \epsilon) - \beta)(B(A\beta + \epsilon) - \beta)^t] \\ &= E[(BA\beta + B\epsilon - \beta)(BA\beta + B\epsilon - \beta)^t] \\ &= E[(\beta + B\epsilon - \beta)(\beta + B\epsilon - \beta)^t] \\ &= E[(B\epsilon)(B\epsilon)^t] = BE[\epsilon\epsilon^t]B^t = \sigma^2 BB^t \end{aligned}$$

Wir definieren "Delta": $D := B - (A^tA)^{-1}A^t$ und damit $B = (A^tA)^{-1}A^t + D$. Es gilt wegen (7.15):

$$DA = BA - (A^tA)^{-1}A^tA = \mathbf{I_k} - \mathbf{I_k} = 0. \tag{7.16}$$

Wir rechnen nochmals unter Verwendung von (7.16):

$$\begin{aligned} V(BY) &= \sigma^2 BB^t = \sigma^2[(A^tA)^{-1}A^t + D][(A^tA)^{-1}A^t + D]^t \\ &= \sigma^2[(A^tA)^{-1}A^tA(A^tA)^{-1} + DA(A^tA)^{-1} + (A^tA)^{-1}A^tD^t + DD^t] \\ &= \sigma^2[(A^tA)^{-1} + DD^t] \end{aligned}$$

Vom Schluss von 7.5 her wissen wir, dass die Kovarianzmatrix von $\hat{\beta}$ gleich $\sigma^2(A^tA)^{-1}$ ist. Damit haben wir

$$V(\tilde{\beta}) := V(BY) = V(\hat{\beta}) + \sigma^2 DD^t.$$

□

7.8 Verbindung zu EDV-Ausdrücken mit abgezogenem Mittelwert

Manche EDV-Pakete ziehen vom Vektor Y zuerst den Mittelwert ab und arbeiten dann mit $Y - \bar{Y}$. Wie ist das mit der bisherigen Theorie des Testens zu vereinbaren und wieso kommen die gleichen Verteilungen (F, auch t) heraus (zum Teil mit um 1 kleinerem df)? Wir vergleichen: In unserem Beispiel aus 7.1 liefert uns der Computerausdruck sowohl die Teststatistik aus 7.6.1 (ganzer Vektor) wie auch 7.6.2 (einzelne Koordinaten).

7.9 Hängepartien von früheren Kapiteln

7.9.1 $\bar{X} \coprod \sum_{i=1}^{n}(X_i - \bar{X})^2$ (beendet 4.4.3 und 4.4.5, (t-Test))

Wenn $X_1, \ldots, X_n$ iid $\mathcal{N}(\mu, \sigma^2)$-verteilt, dann gilt:

$$\bar{X} \coprod \sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Beweis: Corollar 7.13 mit $A := \mathbf{1_n}$, **T 7**.

□

Wie in der Vorlesung schon mehrfach erwähnt, gilt sogar auch die Umkehrung. Ein Beweis hiervon findet sich in R.C. Geary, "Distribution of Student's ratio for nonnormal samples," Roy. Stat. Soc. Jour., Supp. Vol 3, no. 2.

7.9.2 Fundi ANOVA (beendet 4.5.3, (einfache Varianzanalyse))

Dies folgt aus 7.8:

□