WS 2018/19

Statistisches Lernen

Mitschriften bei Prof. Dr. Martin Bogdan
Universität Leipzig

An Introduction to Statistical Learning
James, G., Witten, D., Hastie, T. and Tibshirani, R.

Lehrbuch der Wahrscheinlichkeitsrechnung
Gnedenko, B.W.

Theoretical Statistics
Cox, D.R. and Hinkley, D.V.

Inhaltsverzeichnis

1
Einführung

1.1 Vorbemerkungen

Beim statistischen Lernen geht es darum, intelligente Schlüsse zu ziehen. Der Fokus dieser Vorlesung liegt auf Methoden zur Analyse, weniger auf Design, und mehr auf Beispielen aus dem Bereich der klinischen Studien. Darüberhinaus erstrecken sich die Anwendungen des statistischen Lernens aber auch auf viele andere Fachgebiete.

Beispielanwendungen:

- Unterscheidung von Behandlungen A und B
- Eigenschaften diagnostischer Tests
- Zusammenhang von Krankheiten von A und B
- Risikobewertung (Lebensmittel, Strahlung, ...)

1.2 Wahrscheinlichkeit

1.2.1 Zugänge

(a) relative Häufigkeiten (frequentistisch)

(b) Maß für Überzeugung (bayesianisch)

In dieser Vorlesung wird vor allem der frequentistische Zugang betrachtet. Dieser ist gewissermaßen intuitiv und basiert auf wiederholbaren 'Experimenten' (Münzwurf, radioaktiver Zerfall, Schwangerschaft bei Kontrazeptionsmethode A, 5-Jahres-Überleben nach Chemotherapie B, Wettervorhersagen).

Der Bayesianische Zugang liefert Werkzeuge, um mit Eingangsüberzeugungen ('Prior') und Lernen u.a. durch Daten ('Posterior') umzugehen. Dabei wird jedoch gelegentlich Willkür kritisiert.

In den ersten Vorlesungen folgen wir einem recht traditionellen Zugang, um ein solides Grundverständnis zu erlangen. Dabei werden wir aus zeitlichen Gründen nicht streng mathematisch vorgehen können (Stichwort: Kolmogorovsche Axiomatik).

1.2.2 Das Ereignisfeld

Definition (Ereignis)**.**

Als *Ereignis* bezeichnet man einen möglichen Ausgang eines Zufallsexperiments.

Beispielsweise stellt 'Zahl liegt oben' beim Münzwurf ein Ereignis dar.

Definition (Ereignisfeld)**.**

Ein System heißt *Ereignisfeld*, wenn

(a) es das sichere und unmögliche Ereignis enthält,

(b) A und B Teil des Systems sind, dann auch AB bzw. $A \cap B$. 'Produkt' von A und B bedeutet das gleichzeitige Auftreten beider Ereignisse.

(c) A und B Teil des Systems sind, dann auch $A + B$ bzw. $A \cup B$. 'Summe' von A und B bedeutet, dass mindestens eines der Ereignisse eintritt.

(d) A und B Teil des Systems sind, dann auch $A - B$ bzw. $A \setminus B$. 'Differenz' von A und B bedeutet, dass A eintritt, während B nicht eintritt.

Beispiel (Münzwurf):
Wir betrachten einen Münzwurf und das Ereignisfeld $\{A, B, \Omega, \varnothing\}$. Dann kann man folgende Modellierung vornehmen:

Variable	Beschreibung
A	Zahl liegt oben.
B	Wappen liegt oben.
Ω	Zahl oder Wappen liegen oben.
$\varnothing$	Weder Zahl noch Wappen liegen oben.

1.2.3 Gesetze der Ereignisse

(a) Kommutativität:

$$A + B = B + A$$
$$AB = BA$$

(b) Assoziativität:

$$(A + B) + C = A + (B + C)$$
$$(AB)C = A(BC)$$

(c) Distributivität:

$$A(B + C) = AB + AC$$
$$A + (BC) = (A + B)(A + C)$$

(d) Identität:

$$A + A = A$$
$$AA = A$$

1.2.4 Wahrscheinlichkeitsbegriff

Axiom 1: Jedem Ereignis A aus dem Ereignisfeld $\mathcal{F}$ ordnet man eine nichtnegative Zahl $P(A)$, die Wahrscheinlichkeit, zu.

Axiom 2: $P(\Omega) = 1$, wobei Ω das sichere Ergebnis ist.

Axiom 3: Sind Ereignisse A_i $(i = 1, \ldots, n)$ paarweise unvereinbar (d.h. $A_i A_j = \varnothing$ für $i \neq j$), so gilt $P(A_1 + A_2 + \ldots + A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$.

Eigenschaften:

(a) $P(\varnothing) = 0$

(b) $P(\bar{A}) = 1 - P(A)$ für $\bar{A} := \Omega \setminus A$

(c) $0 \leq P(A) \leq 1$

(d) Für $A \subset B$ ('A ist Teilereignis von B' bzw. 'A zieht B nach sich') folgt $P(A) \leq P(B)$.

(e) $P(A + B) = P(A) + P(B) - P(AB)$

(f) $P(A_1 + \ldots + A_n) \leq P(A_1) + \ldots + P(A_n)$

1.2.5 Bedingte Wahrscheinlichkeit

Die Wahrscheinlichkeit von A unter der Bedingung, dass B eingetreten ist, schreibt man $P(A \mid B)$. Diese ist definiert als

$$P(A \mid B) := \frac{P(AB)}{P(B)}.$$

Motivation:

Gegeben seien n unvereinbare gleichwahrscheinliche Ereignisse $A_1, \ldots, A_n$ mit m günstig für A, k günstig für B und r günstig für AB. Offensichtlich folgen $r \leq k$ und $r \leq m$. Es ergibt sich

$$P(A \mid B) = \frac{r}{k} = \frac{\frac{r}{n}}{\frac{k}{n}} = \frac{P(AB)}{P(B)}$$

Beispiel (Würfel):

Zwei Würfel werden geworfen. Wie groß ist die Wahrscheinlichkeit, die Summe 8 zu erhalten (A), falls die Summe gerade ist (B)? Zunächst bestimmt man durch einfache Überlegung

$$P(A) = \frac{5}{36},$$
$$P(B) = \frac{1}{2},$$
$$P(AB) = \frac{5}{36}.$$

Daraus ergibt sich nun

$$P(A \mid B) = \frac{\frac{5}{36}}{\frac{1}{2}} = \frac{5}{18}.$$

Satz (Bayes'sche Formel)**.**

Seien $A_1, \ldots, A_n$ unvereinbar und ein vollständiges System ($\sum_i A_i = \Omega$). Sei ferner $P(B) \neq 0$, dann folgt

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{\sum_{i=1}^{n} P(B \mid A_j) \cdot P(A_j)}.$$

1.2.6 Anwendung von bedingten Wahrscheinlichkeiten

Bedingte Wahrscheinlichkeiten spielen unter anderem eine wichtige Rolle bei diagnostischen Verfahren. Es seien D^+, D^- (Diagnose) zwei mögliche Krankheitszustände (krank bzw. gesund) und T^+, T^- (Test) die zwei möglichen Ergebnisse eines diagnostischen Tests. Unter diesen Voraussetzungen führt man folgende Bezeichnungen ein:

Variable	Beschreibung
$P(D^+)$	Prävalenz
$P(T^+ \mid D^+)$	Sensitivität
$P(T^- \mid D^-)$	Spezifität
$P(D^+ \mid T^+)$	Positiv-prädiktiver Wert (engl. PPV)
$P(D^- \mid T^-)$	Negativ-prädiktiver Wert (engl. NPV)

1.3 Zufallsvariablen und Verteilungsfunktionen

Eine Zufallsgröße ist eine Größe, deren Werte von Zufall abhängen und für die eine Wahrscheinlichkeitsverteilungsfunktion existiert. Jedem Elementarereignis $\omega \in \Omega$ (unzerteilbar) wird eine reelle Zahl zugeordnet durch $X = X(\omega)\colon \Omega \longrightarrow \mathbb{R}$. Mit

$$F_X(t) := P(X < t)$$

bezeichnet man die Verteilungsfunktion der Zufallsgröße X. Sie ist monoton nicht fallend, linksseitig stetig und gehorcht den Bedingungen $F(-\infty) = 0$ und $F(\infty) = 1$. Umgekehrt lässt sich jede Funktion mit diesen Eigenschaften als Verteilungsfunktion auffassen.

Wichtige Verteilungsfunktionen:

- Binomialverteilung:

$$P_n(m) = \binom{n}{m} p^m q^{n-m}$$

  Die zugehörige Verteilungsfunktion ergibt sich mit

$$F(x) = \begin{cases} 0 & \text{, falls } x \leq 0 \\ \sum_{k<x} P_k & \text{, falls } 0 < x \leq n \\ 1 & \text{, falls } x > n. \end{cases}$$

- Normalverteilung: Für $\sigma > 0$ ergibt sich die Verteilungsfunktion der Normalverteilung mit

$$F(x) = \Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(z-\sigma)^2}{2\sigma^2}} \, \mathrm{d}z.$$

1.4 Erwartungswert, Varianz und weitere Momente

Der Erwartungswert $\mathbb{E}(X)$ einer Zufallsgröße X ist im diskreten Fall definiert als

$$\mathbb{E}(X) = \sum_i x_i p_i,$$

wobei x_i die möglichen Realisierungen der Zufallsgröße und p_i die zugehörigen Eintrittswahrscheinlichkeiten bezeichnen. Im stetigen Fall lautet die Definition:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot p(x)\ \mathrm{d}x.$$

Beispiele:

- Würfel:

$$\mathbb{E}(X) = \frac{1}{6}\sum_{i=1}^{6} i = \frac{21}{6} = \frac{7}{2}.$$

- Binomialverteilung:

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=0}^{n} kP_n(k) \\
&= \sum_{k=1}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \\
&= \sum_{k=1}^{n} k\frac{n(n-1)!}{k(k-1)!(n-k)!}p^k(1-p)^{n-k} \\
&= \sum_{k=1}^{n} n\binom{n-1}{k-1}p^k(1-p)^{(n-1)-(k-1)} \\
&= np\sum_{k=1}^{n} \binom{n-1}{k-1}p^{k-1}(1-p)^{(n-1)-(k-1)} \\
&= np\sum_{k=0}^{n-1} \binom{n-1}{k}p^k(1-p)^{(n-1)-k} \\
&= np\sum_{k=0}^{m} \binom{m}{k}p^k(1-p)^{m-k} \qquad \text{mit } m := n-1 \\
&= np\cdot(p+(1-p))^m \\
&= np.
\end{aligned}$$

- Uniformverteilung auf $[a, b] \subset \mathbb{R}$:

$$\mathbb{E}(X) = \frac{1}{b-a}\int_a^b x\ \mathrm{d}x = \frac{x^2}{2(b-a)}\bigg|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2}.$$

- Normalverteilung:

$$\mathbb{E}(X) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} x\cdot e^{\frac{-(x-a)^2}{\sigma^2}}\ \mathrm{d}x$$

Sei $x' = \frac{x-a}{\sigma}$, dann folgt $x = x'\sigma + a$ und $\mathrm{d}x = \sigma\ \mathrm{d}x'$. Daraus folgt

$$\begin{aligned}
\mathbb{E}(X) &= \frac{\sigma}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(\sigma x' + a)e^{-\frac{x'^2}{2}}\ \mathrm{d}x' \\
&= \frac{a}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{x'^2}{2}}\ \mathrm{d}x' \\
&= a.
\end{aligned}$$

Varianz (Dispersion):

$$V(X) := \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

Im diskreten Fall ist

$$V(X) = \sum_i [x_i - \mathbb{E}(X)]^2 \cdot p_i.$$

Beispiele:

- Man betrachte den Wurf eines Würfels.

$$V(X) = \frac{1}{6}\sum_{i=1}^{6}(i - 3.5)^2 = \frac{1}{3}\sum_{i=1}^{3}\left(i - \frac{7}{2}\right)^2 = \frac{35}{12}$$

  Im stetigen Fall berechnet man die Varianz durch

$$V(X) = \int_{-\infty}^{\infty} (x - \mathbb{E}(x))^2 \cdot p(x)\ \mathrm{d}x = \int_{-\infty}^{\infty} x^2 \cdot p(x)\ \mathrm{d}x - \mathbb{E}(X)^2.$$

- Uniformverteiung auf $[a, b]$

$$V(X) = \frac{1}{b-a}\int_a^b x^2\ \mathrm{d}x - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$$

Definition (Moment)**.**

Wir bezeichnen m_k als das *gewöhnliche Moment (Anfangsmoment)* k-ter Ordnung,

$$m_k := \mathbb{E}(X^k).$$

Das *zentrale Moment* (auf Zentrum von $\mathbb{E}(X)$ bezogen) k-ter Ordnung ist

$$\mu_k := \mathbb{E}[(X - m)^k].$$

Die Varianz ist also das zweite Zentralmoment, $V(X) = \mu_2 = \mathbb{E}(X^2 - 2m_1X + m_1^2) = m_2 - m_1^2$. Man kann μ_k immer durch m_l $(l \leq k)$ ausdrücken.

1.5 Korrelation

Eine Erweiterung dieser Momente stellt die sogenannte Kovarianz

$$b(X, Y) = \mathbb{E}[(X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))]$$

dar. Diese wird als gemischtes Zentralmoment 2-ter Ordnung bezeichnet. Es gilt offensichtlich $b(X, X) = V(X)$. Die normierte Größe

$$\varrho(X, Y) := \frac{b(X, Y)}{\sqrt{V(X) \cdot V(Y)}}$$

bezeichnet man als *Korrelationskoeffizient*. Dieser hat folgende Eigenschaften:

- $-1 \leq \varrho \leq 1$
- $X = Y \implies \varrho = 1$ und $X = -Y \implies \varrho = -1$
- X, Y unabhängig $\implies \varrho = 0$

Wir betrachten nun eine Anwendung auf Wahrscheinlichkeiten. Seien $\mathbb{E}(X) = p_x$ und $V(X) = p_x(1-p_x)$. Dann folgt

$$\varrho_{xy} = \frac{p_{xy} - p_x p_y}{\sqrt{p_x(1-p_x)p_y(1-p_y)}}.$$

Damit folgt

$$p_{xy} = p_x p_y + \varrho\sqrt{p_x(1-p_x)p_y(1-p_y)}.$$

1.6 Wichtige Gesetze der Wahrscheinlichkeitstheorie

1.6.1 Gesetz der großen Zahlen

Man betrachte einen Bernoulli-Versuch. Sei μ die Anzahl der Ereignisse und n die Anzahl der Versuche und p die Eintrittswahrscheinlichkeit der Ereignisse. Für alle $\varepsilon > 0$ gilt dann

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

Im allgemeinen Fall sei $X = (X_1, \ldots, X_n)_n^\top$ eine Folge unabhängig und identisch verteilter Zufallsvariablen mit $\mathbb{E}(X_1) < \infty$ und $\mathrm{Var}(X_1) < \infty$. Dann gilt für alle $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X_1)\right| < \varepsilon\right\}\right) = 1.$$

Nach Tschebyschew sei $X = (X_1, \ldots, X_n)_n^\top$ eine Folge von paarweise unabhängigen Zufallsvariablen mit gleichmäßig beschränkter Varianz. Es gilt also $\mathrm{Var}(X_i) \leq C$ für $C \in \mathbb{R}_+$ und $i \in \{1, \ldots, n\}_n$. Dann gilt für alle $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i)\right| < \varepsilon\right\}\right) = 1.$$

1.6.2 Zentraler Grenzwertsatz (CLT)

Wir betrachten zunächst den *Lokalen Grenzwertsatz von Moivre-Laplace*. Sei $0 < p < 1$ die Eintrittswahrscheinlichkeit eines Ereignisses. In n Versuchen gilt

$$P_n(m) = \binom{n}{m} p^m (1-p)^{n-m}.$$

Dann gilt für $x = \frac{m-\mu}{\sigma} = \frac{m-np}{\sqrt{np(1-p)}}$:

$$\lim_{n\to\infty} \frac{\sqrt{np(1-p)}}{\frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}} = 1.$$

Im allgemeinen Fall betrachten wir eine Folge $X = (X_1, \ldots, X_n)_n^\top$ von identisch verteilten Zufallsvariablen mit $\mathbb{E}(X_1) < \infty$ und $\mathrm{Var}(X_1) = \sigma^2 < \infty$. Sei nun $S_n := \sum_{i=1}^n X_i$, dann gilt

$$\lim_{n\to\infty} P\left(\left\{\frac{S_n - n\mathbb{E}(X_1)}{\sqrt{n}\sigma} < t\right\}\right) = \lim_{n\to\infty} P\left(\left\{\sqrt{n}\cdot\frac{\frac{1}{n}S_n - \mathbb{E}(X_1)}{\sigma} < t\right\}\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-\frac{x^2}{2}}\,\mathrm{d}x.$$

2
Deskriptive Statistiken

Die Beschreibung von Daten und Kohorten ist zentral für das Verständnis einer Arbeit. Ziel ist es, mit wenigen Kenngrößen das Wesentliche zu charakterisieren.

2.1 Einzelne Merkmale

2.1.1 Nominale und ordinale Größen

- absolute und relative Häufigkeiten (siehe Häufigkeitstabellen)
- grafisch
  - Balkendiagramme (mit Konfidenzintervall oder Standardfehler)
  - Kreisdiagramme (unüblich und wenig geeignet)

Beispiel (Fehlgeburtsraten):
Man betrachte eine Kontrollgruppe und gewisse Fälle (z.B. Medikamenteneinnahme während der Schwangerschaft). Die Fehlgeburtsrate kann man schätzen durch

$$\hat{p} = \frac{r}{n}$$

für r als Anzahl der Fehlgeburten und n als Gesamtzahl der Geburten. Daraus folgt

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(a-\hat{p})}{n}}.$$

Damit ergibt sich das Konfidenzintervall $\mathrm{KI} = \hat{p} \pm 2\mathrm{SE}$.

Definition (Konfidenzintervall)**.**

Seien θ ein zu schätzender Parameter und $\hat{\theta}$ ein Schätzer für θ. Dann definiert man ein $(1-\alpha)$–Konfidenzintervall für θ, sodass gilt:

$$P(a < \theta < b) = 1 - \alpha.$$

Das heißt, dass das Intervall $[a, b]$ ein $(1-\alpha)$–Konfidenzintervall für θ ist.

Interpretation:
θ wird durch $\hat{\theta}$ geschätzt. Ein 95%–Konfidenzintervall wird durch wiederholte Konstruktion von Konfidenzintervallen bestimmt, von denen im 95% den tatsächlichen Wert θ enthalten.
Hier Grafik

Beispiel (durchschnittliche Körpergröße):
Sei θ die zu schätzende durchschnittliche Körpergröße von Frauen in Deutschland und $\hat{\theta}$ ein entsprechender Schätzer auf Grundlage einer Stichprobe. Bei kleinen Stichproben wird das Konfidenzintervall unter Umständen sehr groß. Beispielsweise könnte eine Stichprobe folgende Ergebnisse liefern:

- $\hat{\theta} = 1.68m$
- 95%–KI$(\hat{\theta}) = [1.50; 1.79]$

Bei größeren Stichproben wird das Konfidenzintervall üblicherweise schmaler und es könnten folgende Ergebnisse vorliegen:

- $\hat{\theta} = 1.69m$
- 95%–KI$(\hat{\theta}) = [1.63; 1.72]$

2.1.2 Metrische Daten

- Lagemaße:
  - Mittelwert
  - Quantile (insb. Median)
- Streumaße:
  - Standardabweichung (empirisch)
  - Interquartilabstand (IQR): |(75%–Quantil)−(25%–Quantil)|
  - Spannweite (Range): |Max − Min|
- Grafiken
  - Histogramme
  - Fehlerbalken
  - Boxplot

Übersicht von molekularbiologischen Hochdurchsatzverfahren:

- Genom (SNPs)
- Epigenom
- Transkryptom
- Metabelom
- Proteom

Mögliche Fragestellungen:

(a) Welchen Einfluss/Effekt hat die Gabe eines bestimmten Medikaments X auf die Expression eines Gens Y_i?

(b) Gibt es Unterschiede in der Expressionsrate von Gen Y_i zwischen gesunden Probanden und erkrankten Probanden?

(c) Wie stark ist der Effekt/Unterschied?

2.2 Zusammenhang zweier Merkmale

Nominale Zusammenhänge:

- Kontigenztafel
  - odds ratio
  - relatives Risiko
- grafisch: Forest plot

Metrische Zusammenhänge:

- Korrelationskoeffizient (mit KI)
- Streudiagramm

Beispiel (Simpson Paradoxon):

Durch das Simpson Paradoxon kann verdeutlicht werden, dass ein Effekt, der in der Gesamtgruppe beobachtet wird, in Subgruppen anders ausfallen kann. Man betrachte beispielsweise zwei Gruppen A, B von Fahrschulprüflingen und deren Erfolg bzw. Misserfolg in der Prüfung:

Gruppe	A	B
Erfolg	70	50
Misserfolg	160	182
Gesamt	230	232

Hiernach könnte man folgern, dass Schüler in Gruppe A mehr Erfolg hatten als Schüler in Gruppe B. Unterteilt man die Gruppen jedoch nach Geschlecht, so lässt sich dieser Effekt unter gewissen Umständen nicht auf alle Teilgruppen verallgemeinern:

	Gruppe	A	B
M	Erfolg	7	45
	Misserfolg	28	172
W	Erfolg	63	5
	Misserfolg	132	10
	Gesamt	230	232

3

Statistisches Testen

3.1 Die Logik des Testens

Die Logik des Testens weist eine gewisse Analogie zum Beweis durch Widerspruch auf, die man häufig nutzen kann. Beispielsweise könnte man sich fragen, was der Zusammenhang zwischen dem Testergebnis und der Wahrheit der Nullhypothese ist und folgende Überlegungen anstellen:

	Statistisches Testen	**Beweis durch Widerspruch**
Annahme	H_0: $\mu_1 = \mu_2$ bzw. der Mittelwert der Gruppe 1 entspricht dem Mittelwert der Gruppe 2. Es wird vermutet, dass dies falsch ist.	$\sqrt{2}$ ist rational. Es wird vermutet, dass dies falsch ist.
Folge	Man nimmt an, dass die Annahme doch stimmt.	Man nimmt an, dass die Annahme doch stimmt.
Ergebnis a)	Ist das das Ergebnis sehr unwahrscheinlich, so ist die Annahme nicht plausibel (Korrelation $< 5\%$, H_0 wird abgelehnt).	Kommt man auf einen Widerspruch, so muss die Annahmen falsch sein.
Ergebnis b)	Ist das Ergebnis plausibel, weiß man wenig über die Annahme, ein Konfidenzintervall kann jedoch weiterhelfen.	Kommt man auf keinen Widerspruch erhält man geringe Informationen über die Annahme.

Fehlertypen bei Hypothesentests:

	H_0 stimmt	H_0 stimmt nicht
H_0 abgelehnt	Typ-I-Fehler α (Korrelation $\alpha \leq 0.05$)	Power: $1 - \beta$
H_0 nicht abgelehnt	kein Fehler	Typ-II-Fehler β (Planung: $\beta = 0.1$)

3.2 Der T-Test

Beim T-Test werden zwei Mittelwerte verglichen. Die Nullhypothese lautet $H_0 : \mu_1 = \mu_2$. Unter der Annahme gleicher Varianz schätzt man die sogenannte t-Statistik durch

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1^2} + \frac{1}{n_2^2}}},$$

wobei

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Dabei bezeichne n_i die Stichprobengröße der i-ten Gruppe und $\bar{x}_i$ den Mittelwert dieser Gruppe. Die grundlegende Struktur der t-Statistik ist $T = \frac{\Delta x}{\mathrm{SE}}$. Unter der Annahme, dass H_0 stimmt, hat T eine t-Verteilung mit $f = n_1 + n_2 - 2$ Freiheitsgraden.

Welch-Test:

Im Gegensatz zum normalen T-Test ist der Welch-Test eine Variante, die keine Annahme über die Gleichheit der Varianzen macht. In diesem Fall betrachtet man:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{1}{n_1^2} + \frac{1}{n_2^2}}},$$
$$f = \frac{(\tilde{s}_1^2 + \tilde{s}_2^2)^2}{\frac{\tilde{s}_1^4}{n_1 - 1} + \frac{\tilde{s}_2^4}{n_2 - 1}},$$
$$\tilde{s}_i = \frac{s_i}{n_i}.$$

Man kann dann ein Konfidenzintervall für $\Delta\mu$ bestimmen:

$$\text{KI} = \Delta\bar{x} \pm \underbrace{t_{\alpha/2,f}}_{\approx 2 \text{ für } \alpha = 0.05} \text{SE}.$$

Weiter bezeichnet man als p-Wert die Wahrscheinlichkeit, den Wert T zu beobachten unter der Annahme, dass H_0 stimmt.

3.3 Kontingenztafel

Kontingenztafeln sind Tabellen, die die absoluten oder relativen Häufigkeiten (Häufigkeitstabellen) von Kombinationen bestimmter Merkmalsausprägungen enthalten. Kontingenz hat dabei die Bedeutung des gemeinsamen Auftretens von zwei Merkmalen. Das bedeutet, es werden Häufigkeiten für mehrere miteinander durch „und“ oder „sowie“ (Konjunktion) verknüpfte Merkmale dargestellt. Diese Häufigkeiten werden ergänzt durch deren Randsummen, die die sogenannten Randhäufigkeiten bilden. Der häufige Spezialfall einer Kontingenztabelle mit zwei Merkmalen ist eine Konfusionsmatrix.

Beispiel (Odds Ratios):

Gegeben sei die folgende Kontingenztafel:

	A	B	Randhäufigkeit
I	n_{11}	n_{12}	$n_{1\bullet}$
II	n_{21}	n_{22}	$n_{2\bullet}$
Randhäufigkeit	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

In diesem Beispiel schätzt $\frac{n_{11}}{n_{21}}$ die Odds von I im Vergleich zu II bei Gruppe A. Das Odds Ratio ist gegeben durch

$$\widehat{\text{OR}} = \frac{\frac{n_{11}}{n_{21}}}{\frac{n_{22}}{n_{12}}}.$$

Weiter erhält man das Konfidenzintervall

$$\text{KI} = \widehat{\text{OR}} \cdot e^{z_{\alpha/2} \cdot \text{SE}},$$

wobei

$$\text{SE} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

Hier stellt sich die Frage, ob 1 in dem Intervall liegt, da kein Unterschied zwischen A und B existiert, wenn $\widehat{\text{OR}} = 1$.

Fisher-Test:

Der Fisher-Test ist ein exakter Test, da ein genauer Wert aus der Kombinatorik berechnet werden kann gemäß

$$P = \frac{\binom{n_{1\bullet}}{n_{11}}\binom{n_{2\bullet}}{n_{22}}}{\binom{n_{\bullet\bullet}}{n_{\bullet 1}}}.$$

Der Zusammenhang kann mit hoher „Power“ mit dem Chi-Quadrat-Test getesten werden. Bezeichne e_{ij} die erwartete Anzahl

$$e_{ij} = \frac{n_{i\bullet}n_{\bullet i}}{n_{\bullet\bullet}}$$

und betrachte

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}.$$

χ^2 hat eine χ^2_f-Verteilung mit $f = (l-1)(m-1)$ Freiheitsgraden für $i = 1, \ldots, l$ und $j = 1, \ldots, m$. Als Faustregel gilt, wenn $n_{ij} \neq 0$ für alle i, j gilt und $< 25\%$ der Zellen haben $l_{ij} < 5$, dann kann der χ^2-Test angewandt werden.

4

Statistische Modelle

Definition (Statistisches Modell)**.**

Ein statistisches Modell stellt eine Zufallsvariable Y in Beziehung zu einem oder mehreren Kovariablen:

$$Y = f(X) + \varepsilon$$

Bezeichnungen:

Variable	Beschreibung
Y	abhängige Zufallsvariable
X	Kovariaten, unabhängige Variablen
$f(X)$	unbekannte Funktion, die den systematischen Effekt von X auf Y modelliert
ε	zufälliger Fehler, gibt den Anteil von Y an, der nicht durch $f(X)$ erklärt werden kann

Anwendungen von statistischen Modellen:

- Inferenz:
  - Ziel ist es, die Art des Zusammenhangs zwischen Y und X zu verstehen. Es ist also von Interesse, die Form von $f(X)$ zu kennen.
- Vorhersage:
  - Ziel ist es, den Wert von Y so genau wie möglich vorherzusagen. Hier ist es nicht von Interesse, die exakte Form von $f(X)$ zu kennen.

Schätzen der Funktion:

Die Funktion $f(X)$ wird mittels einer statistischen Lernmethode anhand einer Menge von Traningsdaten geschätzt. Die geschätzte Funktion wird mit $\hat{f}(X)$ gekennzeichnet. Die Trainingsdaten liegen üblicherweise (beim überwachten Lernen) in der Form

$$(X, Y) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

vor, wobei n die Anzahl der Messungen/Beobachtungen bezeichne.

4.1 Klassifikation von statistischen Lernmethoden

4.1.1 Parametrische Methoden

Modellwahl:

Für $f(X)$ wird eine bestimmte Form angenommen und oft wird die Anzahl der Kovariablen schon vorher festgelegt. Verfahren zur Modellselektion legen die Wahl der Kovariablen fest.

Training des Modells:
Es werden die Gewichte der Kovariablen anhand der vorhandenen Trainingsdaten geschätzt.

Fazit:

- Modelle sind weniger flexibel, können daher oft besser interpretiert werden.
- Gewähltes Modell entspricht oft nicht der wahren Form von $f(X)$.
- Da nur Parameter gelernt werden, ist eine geringere Stichprobengröße ausreichend.

4.1.2 Nichtparametrische Modelle

Es wird keine bestimmte Form von $f(X)$ festgelegt. Es werden die Form und die Parameter einer beliebig komplexen Funktion $f(X)$ anhand der Trainingsdaten geschätzt.

- Modelle sind sehr flexibel, aber oft weniger gut interpretierbar.
- Form von $f(X)$ wird anhand der Trainingsdaten gelernt.
- Oftmals eine größere Stichprobengröße notwendig.

4.2 Lineare Regression

Die Lineare Regression gehört zu den linearen statistischen Modellen. Als Form von $f(X)$ wird ein annähernd linearer Zusammenhang zwischen X und Y angenommen. Dabei nimmt die Zufallsvariable Y quantitative Werte an und die Kovariablen X können qualitative oder quantitative Werte annehmen.

4.2.1 Univariate lineare Regression

Definition (Univariate lineare Regression)**.**

Die univariate lineare Regression ist ein statistisches Modell, das den Wert der Zielvariablen Y auf Basis der Werte einer einzigen Kovariablen X unter Annahme eines annähernd linearen Zusammenhangs vorhersagt:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Bezeichnungen:

Variable	Beschreibung
β_0	Mittelwert von Y, falls es keine Zusammenhang zwischen Y und X gibt. Sonst ist β_0 der Schnittpunkt mit der y-Achse.
β_1	Regressionskoeffizient, Effekt der Kovariablen X auf Y, mittlerer Anstieg in Y, wenn sich X um eine Einheit ändert
ε	normalverteilter Fehler $\varepsilon \sim N(0, \sigma^2)$

Annahmen über den zufälligen Fehler:

- Alle Störungen haben die gleiche Varianz $\mathrm{Var}(\varepsilon_i) = \sigma^2$ (Homoskedastizität).
- Alle Störungen sind um 0 verteilt (zentriert) bzw. $\mathbb{E}(\varepsilon_i) = 0$.

- Störgrößen sind unabhängig voneinander: $\mathrm{Cor}(\varepsilon_i, \varepsilon_j) = 0$.

Varianzdekomposition:

$$\begin{array}{ccccc} \sum_{i=1}^{n}(y_i-\bar{y})^2 & = & \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 & + & \sum_{i=1}^{n}(\hat{y}_i-y_i)^2 \\ \underbrace{\textit{Total sum of squares}}_{\mathrm{TSS}} & = & \underbrace{\textit{Explained sum of squares}}_{\mathrm{ESS}} & + & \underbrace{\textit{Residual sum of squares}}_{\mathrm{RSS}} \end{array}$$

Schätzen der Parameter (Methode der kleinsten Quadrate):
Bei der Methode der kleinsten Quadrate möchte man die Differenz zwischen den Werten der Zufallsvariablen Y_i und den geschätzten bzw. vorhergesagten Werten $\hat{y}_i$ für alle Beobachtungen reduzieren:

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2 \quad \longrightarrow \quad \min.$$

Durch Bilden der ersten partiellen Ableitung und Null setzen erhält man folgende explizite Lösungsvorschriften:

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\mathrm{Cor}(X,Y)}{\mathrm{Var}(X)}. \end{aligned}$$

Güte der Schätzungen:
Der wahre Zusammenhang zwischen Y und X ist unbekannt. Das heißt, die wahre Regressionsgerade ist unbekannt, da die Schätzungen $\hat{\beta}_0, \hat{\beta}_1$ von dem gewählten Trainingsdatensatz abhängen. Folglich schwanken die trainierten Regressionsgeraden um die die wahre Regressionsgerade. Die Koeffizienten β_i sind Zufallsvariablen, welche auf Basis von Beobachtungen (Stichprobe) geschätzt werden. Es gelten

$$\begin{aligned} \mathrm{Var}(\beta_1) &= \frac{\delta_\varepsilon^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \\ \mathrm{Var}(\beta_0) &= \frac{\delta_\varepsilon^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}, \end{aligned}$$

wobei δ_ε^2 die Varianz der Residuen ist.

Bestimmtheitsmaß:
Das Bestimmtheitsmaß R^2 gibt an, wie gut die Schätzung der beobachteten Daten anhand der trainierten Regressionsgerade ist. Eine Schätzung ist dann besonders gut, wenn in der Varianzdekomposition der Anteil von RSS an TSS klein ist und ESS möglichst groß ist. Daraus gewinnt man das Bestimmtheitmaß

$$R^2 = \frac{\mathrm{TSS}-\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}.$$

Offensichtlich gilt $0 \leq R^2 \leq 1$. Ist $R^2 \approx 1$, so kann ein großer Anteil der Variabilität, die wir in der Zielvariablen beobachten, durch die trainierte Regressionsgerade erklärt werden.

Testen auf Zusammenhang:
Es stellt sich die Frage, ob ein linearer Zusammenhang $Y = \beta_0 + \beta_1 X + \varepsilon$ zwischen Y und X besteht. Zur Beantwortung kann man einen t-Test nutzen. Die passende Nullhypothese H_0 behauptet, dass kein Zusammenhang existiert ($\beta_1 = 0$), während die Alternativhypothese H_1 behauptet, dass ein Zusammenhang

existiert ($\beta_1 \neq 0$). Es ist dann

$$T = \frac{\hat{\beta}_1}{\mathrm{SD}\sqrt{\hat{\beta}_1}} \sim A_{1-\alpha/2,n-2}.$$

Dabei gilt

$$\mathrm{SD}(\hat{\beta}_1) = \sqrt{\mathrm{Var}(\beta_1)} = \sqrt{\frac{\frac{\mathrm{RSS}}{n-2}}{\mathrm{Var}(X)}}.$$

4.2.2 Multivariate lineare Regressionsmodelle

Die multivariate lineare Regression ist die Erweiterung der univariaten Regression auf mehrere Variablen $X = \{X_1, \ldots, X_p\}$.

Ziele:

- Identifiziere ein verbessertes Modell, als es auf Basis einer einzigen Variable möglich ist.
- Modelliere den individuellen Effekt einer Variable X_j über den Effekt aller anderen gegebenen Kovariablen hinaus.

Individuelle Effekte von Kovariablen X_j können durch spezifische Regressionsgeraden modelliert werden. Jede partielle Regressionsgerade modelliert den Effekt von X_j, während alle anderen Kovariablen X_i für $i \neq j$ ihren enstprechenden Mittelwert annehmen.

Fragen:

- Hat mindestens eine der Kovariablen einen Effekt auf Y?
- Welche der Kovariablen hat einen Effekt auf Y?
- Wie gut entspricht das gefittete Modell den Daten?

Additives Modell:

Beim additiven Modell nimmt man an, dass der Effekt einer Kovariablen X_j auf Y unabhängig von allen anderen Kovariablen ist:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon.$$

Dabei gelten folgende Bezeichnungen:

Variable	Beschreibung
$\beta_1, \ldots, \beta_p$	Effekt der Kovariablen X_j für $1 \leq j \leq p$ auf Y, wenn alle Koeffizienten β_i für $i \neq j$ konstant gehalten werden.
k	Anzahl der Kovariablen
$x_i^{(j)}$	Wert der Kovariablen j für Individuen i
ε_i	zufällige Störung (häufig: $\varepsilon_i \sim N(0, \sigma^2),\ \sigma^2 > 0,\ i \in \{1, \ldots, n\}$, unabhängig)
β_j	unbekannte, wahre Parameter

Multiplikatives Modell:

Im multiplikativen Modell wird angenommen, dass Interaktionen zwischen den Kovariablen möglich sind. Sei

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \ldots + \beta_{p+1} X_p$$

ein multiplikatives Modell, dann gibt β_3 an, inwieweit X_1 von X_2 abhängt.

Schätzen der Koeffizienten:
Über die Methode der kleinsten Quadrate kann man Schätzer $\hat{\beta}_1, \ldots, \hat{\beta}_p$ für die Koeffizienten $\beta_1, \ldots, \beta_p$ finden. Mit diesen Schätzern kann man schließlich Vorhersagen berechnen nach der Formel

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p.$$

Die Schätzer sind dann genau diejenigen Parameter, für die die folgende Summe minimal wird:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \ldots - \hat{\beta}_p x_p)^2.$$

Wichtige Fragen:

(a) Hat mindestens eine der Kovariaten einen Effekt auf Y?

Um diese Frage zu beantworten, testen wir

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0,$$
$$H_A : \beta_j \neq 0.$$

Dazu berechnen wir die F-Statistik

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)}.$$

H_0 kann nicht abgelehnt werden, wenn $F \approx 1$. Gilt $F > 1$, dann wird H_0 abgelehnt.

(b) Welche Kovariaten haben einen Effekt?

Dies kann für jeden einzelnen Koeffizienten mittels t-Test beantwortet werden. Wir testen $H_0 : \beta_j = 0$ gegen $H_A : \beta_j \neq 0$ und berechnen

$$t = \frac{\beta_j}{\text{SD}(\hat{\beta}_2)} \sim A_{1-\alpha/2, n-p-1}.$$

Beachte: Ist p groß, so ist die Anzahl der einzelnen t-Tests groß und damit auch die Family-wise Error Rate. Das heißt, dass H_0 für mindestens ein β_j abgelehnt wird, obwohl H_0 wahr ist.

4.3 Nichtlineare Regression

Rückblick:
Eine lineare Regression kann mathematisch beschrieben werden durch:

$$Y = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_k x_i^{(k)} + \varepsilon_i,$$

$i \in \{1, \ldots, n\}$. Dabei gelten folgende Bezeichnungen:
Die Parameter β_j müssen geschätzt werden, wofür sich unter anderem die Methode der kleinsten Quadrate eignet:

$$\sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_k x_i^{(k)} \right) \right)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \quad \underset{\beta_0, \ldots, \beta_k}{\longrightarrow} \quad \min$$

Variable	Beschreibung
n	Anzahl der Messungen/Individuen/Objekte
k	Anzahl der Kovariablen
$x_i^{(j)}$	Wert der Kovariablen j für Individuen i
ε_i	zufällige Störung (häufig: $\varepsilon_i \sim N(0, \sigma^2),\ \sigma^2 > 0,\ i \in \{1, \ldots, n\}$, unabhängig)
β_j	unbekannte, wahre Parameter

Um zu testen, ob $x^{(j)}$ einen Einfluss auf Y hat, testet man

$$H_0^{(j)} : B_j = 0 \text{ vs. } H_A^{(j)} : B_j \neq 0.$$

4.3.1 Nichtlineare Zusammenhänge

Häufig bestehen nichtlineare Zusammenhänge zwischen abhängiger und unabhängigen Variablen. Manchmal ist dies aus theoretischen Wissen oder empirischer Beobachtung bekannt (Bsp.: Wachstum von Kindern). Oft hat man einen beschränkten Wertebereich, z.B. $R = [0, 1]$ oder einen diskreten Wertebereich, z.B. $R = \{0, 1\}$. Ein nichtlineares Modell kann mathematisch beschrieben werden durch:

$$Y_i = h\left(x_i^{(1)}, \ldots, x_i^{(m)}; \theta_1, \ldots, \theta_p\right) + \varepsilon_i,$$

$i \in \{1, \ldots, n\}$. Dabei ist h im Allgemeinen eine nichtlineare Funktion und $\theta_1, \ldots, \theta_p$ sind die unbekannten, wahren Parameter.

Beispiele:

(a) Biochemischer Sauerstoffverbrauch von Mikroorganismen:

$$h\left(x; \theta_1, \theta_2\right) = \theta_1(1 - \exp(-\theta_2 x))$$

(b) Cobb-Douglas-Funktion:

$$h\left(x^{(1)}, x^{(2)}; \theta_1, \theta_2, \theta_3\right) = \theta_1 \left(x^{(1)}\right)^{\theta_2} \left(x^{(2)}\right)^{\theta_3}$$

(c) Polynomiale Regression:

$$h(x; \theta_0, \ldots, \theta_p) = \theta_0 + \theta_1 + \ldots + \theta_p x^p$$

4.3.2 Linearisierung

Manchmal (praktisch eher häufig) lässt sich h in einen Ausdruck umwandeln, der linear in den transformierten Variablen ist.

Beispiel:
Sei $h(x; \theta_1, \theta_2) = \theta_1 x^{\theta_2}$. Dann gilt

$$\begin{aligned} \log h(x; \theta_1, \theta_2) &= \log\left(\theta_1 x^{\theta_2}\right) \\ &= \log \theta_1 + \log x^{\theta_2} \\ &= \log \theta_1 + \theta_2 \log x. \end{aligned}$$

Damit erhalten wir $\tilde{h}(\tilde{x};\tilde{\theta}_1,\tilde{\theta}_2) = \log\theta_1 + \theta_2 \log x = \tilde{\theta}_1 + \tilde{\theta}_2\tilde{x}$. Das zugehörige Modell ist

$$\tilde{Y}_i = \tilde{\theta}_1 + \tilde{\theta}_2\tilde{x} + \varepsilon_i,$$

$i \in \{1,\dots,n\}$. Die Parameter $\tilde{\theta}_1$ und $\tilde{\theta}_2$ können nun über die Methode der kleinsten Quadrate geschätzt werden. Eine Rücktransformation liefert:

$$Y_i = \exp\tilde{Y}_i = \exp\tilde{\theta}_1 \exp\left(\tilde{\theta}_2\tilde{x}\right) \exp\varepsilon_i = \theta_1 x^{\theta_2} \exp\varepsilon_i.$$

Zufällige Störungen wirken in diesem Fall multiplikativ und (falls $\varepsilon \sim N(\cdot,\cdot)$) sind log-normalverteilt. Daraus folgern wir, dass wir Linearisierung nur anwenden, falls sich die Fehler tatsächlich so wie in der transformierten Variable verhalten. Durch eine Residuenanalyse könnte man jedoch auch in anderen Fällen eine Linearisierung in Erwägung ziehen.

4.3.3 Spezielle nichtlineare Situationen

$$Y_i = \sum_{j=1}^{p} \theta_j h_j\left(x_i^{(1)},\dots,x_i^{(m)}\right) + \varepsilon_i,$$

$i \in \{1,\dots,n\}$. Beispiele für h_j:

(a) lineare Modell: $h_j\left(x_i^{(1)},\dots,x_i^{(m)}\right) = x_i^{(j)}$

(b) polynomielle Terme: $h_j(x_i) = x_i^j$

(c) Indikatorfunktion: $h_j\left(x_i^{(1)},\dots,x_i^{(m)}\right) = 1_{[\alpha_j,\alpha_{j+1})}$

Polynomiale Regression:

$$Y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \ldots + \theta_p x_i^p + \varepsilon_i,$$

$i \in \{1,\dots,n\}$. Dabei ist die Anpassung an die Daten umso besser, je größer p wird. Gleichzeitig verschlechtert sich damit aber auch das Randverhalten.

Stückweise Regression:

$$Y_i = \sum_{j=1}^{p} \theta_j 1_{[\alpha_j,\alpha_{j+1})}\left(x^{(j)}\right) + \varepsilon_i,$$

$i \in \{1,\dots,n\}$. Diese Art der Regression ermöglicht eine stückweise Approximation mit kostanten Funktionen. Erweiterungen sind mit linearen, quadratischen oder kubischen Funktionen möglich.

Beispiel:
Seien $m = 1$ und

$$Y_i = \sum_{j=1}^{p} \left(a_j x_i^3 + b_j x_i^2 + c_j x_i + d_j\right) 1_{[\alpha_j,\alpha_{j+1})}(x_i) + \varepsilon_i.$$

Bei stückweisen Regressionen können folgende Probleme auftreten:

- Unstetigkeiten an Intervallgrenzen

- Asymptotisches Verhalten an den Intervallgrenzen kann (insbesondere bei Polynomen höherer Ordnung) sehr unpassend sein.

Gewünscht sind jedoch stetige und idealerweise glatte Übergänge an den Intervallgrenzen.

4.4 Multiples Testen

Rückblick:

Bisher haben wir eine Hypothese bezüglich eines Parameters θ betrachtet (bspw. $\theta \in \{\mu, \sigma^2, \beta_j\}$). Mit Θ bezeichnen wir den Parameterraum für θ (bspw. $\Theta \in \{\mathbb{R}, \mathbb{R}_+, [0,1]\}$). Man kann den Parameterraum Θ in Θ_0 und Θ_A zerlegen, sodass gilt:

$$\Theta_0 \cup \Theta_A = \Theta,$$
$$\Theta_0 \cap \Theta_A = \varnothing.$$

Entsprechend betrachtet man $H_0 : \theta \in \Theta_0$ vs. $H_A : \theta \in \Theta_A$ (bspw. seien $\theta = \mu,\ \Theta_A = \{0\},\ \Theta = \mathbb{R} \setminus \{0\}$). Um nun zu testen, betrachten wir den Raum S^n aller n-elementigen Stichproben $X = (X_1, \dots, X_n)^\top$. Weiter sei

$$\varphi \colon S^n \longrightarrow \{0,1\}$$

ein Test. Für diesen Test gelten:

- $\varphi(X) = 1 \iff H_0$ wird verworfen,
- $\varphi(X) = 0 \iff H_A$ wird nicht verworfen.

Bei derartigen Tests können Fehler auftreten. Ein Fehler 1. Art beschreibt die Ablehnung von H_0, obwohl H_0 gilt. Dies ist gleichbedeutend mit $\varphi(X) = 1$, obwohl $\theta \in \Theta_0$. Ein Fehler 2. Art beschreibt den Fall, dass H_0 nicht abgelehnt wird, obwohl H_A gilt. Dies ist gleichbedeutend mit $\varphi(X) = 0$, obwohl $\theta \in \Theta_A$.

Vorgehen:

(a) Festlegen einer oberen Schranke α für Fehler 1. Art (bspw. $\alpha \in \{10\%, 5\%, 1\%\}$)

(b) Minimierung der Wahrscheinlichkeit β

4.4.1 Multiples Testproblem

Gegeben sei eine Stichprobe $X = (X_1, \dots, X_n)^\top \in S^n$. Ein multipler Test ist $\varphi = (\varphi_1, \dots, \varphi_m)^\top$, wobei jedes $\varphi_j \colon S^n \longrightarrow \{0,1\}$ für $j \in \{1, \dots, m\}$ ein Test auf Grundlage der Stichprobe X ist. Sei nun $\theta \in \Theta$ der wahre Parameter. Dann gilt $H_0^{(j)}$ genau dann, wenn $\theta \in \Theta_0^{(j)}$. Wir betrachten nun die Menge

$$I_0(\theta) := \left\{ j \in \{1, \dots, m\} \mid \theta \in \Theta_0^{(j)} \right\}$$

der unter θ wahren Nullhypothesen. Entsprechend definieren wir die Menge der unter θ falschen Nullhypothesen gemäß

$$I_A(\theta) := \left\{ j \in \{1, \dots, m\} \mid \theta \in \Theta_A^{(j)} \right\}.$$

Family-wise Error Rate (FWER):

Die Family-wise Error Rate ist definiert gemäß

$$\mathrm{FWER}_\theta(\varphi) := P\left(\bigcup_{j \in I_0(\theta)} \{\varphi_i = 1\}\right)$$

und entspricht der Wahrscheinlichkeit für multiple Fehler 1. Art. Eine Abbildung φ heißt Test zum multiplen Niveau α, falls $\mathrm{FWER}_\theta(\varphi) \leq \alpha$ für alle $\theta \in \Theta$ gilt.

Satz (Bonferroni)**.**

Sei $\varphi = (\varphi_1, \ldots, \varphi_m)^\top$ *ein multipler Test und es gelte*

$$P(\{\varphi_j = 1\}) \leq \frac{\alpha}{m}$$

für alle $\theta \in \Theta_0^{(j)}$ *für alle* $j \in \{1, \ldots, m\}$*. Dann folgt* $\mathrm{FWER}_\theta(\varphi) \leq \alpha$ *für alle* $\theta \in \Theta$*.*

Satz (Sidak)**.**

Sei $\varphi = (\varphi_1, \ldots, \varphi_m)^\top$ *ein multipler Test und es seien* $\varphi_j(X)$ *stochastisch unabhängig für alle* $j \in \{1, \ldots, m\}$*. Es gelte*

$$P(\{\varphi_j = 1\}) = 1 - (1 - \alpha)^{\frac{1}{m}}$$

für alle $\theta \in \Theta_0^{(j)}$*. Dann folgt* $\mathrm{FWER}_\theta(\varphi) \leq \alpha$ *für alle* $\theta \in \Theta$*. Ist also jedes* φ_j *ein Test zum Niveau* $\tilde{\alpha} = 1 - (1 - \alpha)^{\frac{1}{m}}$ *und sind alle Tests voneinander unabhängig, so ist* φ *ein multipler Test zum multiplen Niveau* α*.*

Bemerkung:
Es gilt $\frac{\alpha}{m} \leq 1-(1-\alpha)^{\frac{1}{m}}$. Deshalb ist Sidak weniger konservativ als Bonferroni, wenn die Tests unabhängig voneinander sind.

p-Wert:
Der p-Wert ist die Wahrscheinlichkeit, dass die Prüfgröße T (statistischer Test) den Wert $T(X)$ oder einen extremeren Wert annimmt, wobei $x = (x_1, \ldots, x_n)^\top$ die Realisierung der zugrunde liegenden Stichprobe $X = (X_1, \ldots, X_n)^\top$ ist. Man betrachtet also

$$\sup_{\vartheta \in \Theta} P(\text{„}T(X) \text{ ist extremer als } T(x).\text{“}).$$

Bonferroni-Holm-Test:
Seien $\varphi = (\varphi_1, \ldots, \varphi_m)^\top$ ein multipler Test, $p = (p_1, \ldots, p_m)^\top$ die zum Test gehörigen p-Werte mit ...

5
Modellwahl und Regularisierung

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_n x_i^{(k)} + \varepsilon_i,$$

$i \in \{1, \ldots, n\}$. Schätzer für β_j über Methode der kleinsten Quadrate.

Probleme:

(a) $n \approx k$

(b) $k > n$ bzw. $k \gg n \implies$ keine eindeutige Lösung

Lösungsideen:

(a) Einfügen zusätzlicher Bedingungen in das Modell

(b) Eliminierung der irrelevanten Variablen

5.1 Modellbewertung

Residual Sum of Squares:

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

Mallow's C_p:

$$C_p := \frac{1}{n}(\text{RSS} + 2k\hat{\sigma}^2)$$

mit $\hat{\sigma}^2$ als Varianzschätzer für ε_i, $i \in \{1, \ldots, n\}$.

Akaike Informationskriterium (AIC):

$$\text{AIC} := n \log \hat{\sigma}^2 + 2k$$

Bayes'sches Informationskriterium (BIC):

$$\text{BIC} := n \log \hat{\sigma}^2 + k \log n$$

Bestimmtheitsmaß (R^2):

$$R^2 := 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_n)^2}$$

Korrigiertes Bestimmtheitsmaß (R^2_{adj}):

$$R^2_{\text{adj}} := 1 - \frac{\frac{\text{RSS}}{n-k-1}}{\frac{\text{TSS}}{n-1}} \in [0, 1]$$

Bemerkung:

- RSS und R^2 hängen stark von der Anzahl der Parameter ab.
- RSS und R^2 sind nur zum Vergleich von Modellen mit der gleichen Anzahl an Einflussgrößen geeignet.

5.2 Modellwahlverfahren

Best Subset Selection:

- Vergleich aller möglichen Submodelle:

  $$\left\{x^{(j_1)}, \ldots, x^{(j_l)}\right\} \subseteq \left\{x^{(1)}, \ldots, x^{(k)}\right\}$$

  Dann betrachtet man alle Modelle

  $$Y_i = \beta_0 + \beta_1 x_i^{(j_1)} + \ldots + \beta_l x_i^{(j_l)} + \varepsilon_i,$$

  für $i \in \{1, \ldots, n\}$. Dafür ergeben sich 2^k Möglichkeiten.
- Vorgehen:
  (a) Berechne das Null-Modell M_0:

  $$Y_i = \beta_0 + \varepsilon_i.$$

  (b) Für $l = 1, \ldots, k$:
    (1) Berechne alle $\binom{k}{l}$ Modelle mit genau l Einflussgrößen.
    (2) Wähle das Modell mit dem größten R^2, erhalte M_l.

  (c) Wähle unter $M_0, \ldots, M_k$ das beste Modell gemäß:
    - R^2_{adj}
    - AIC
    - kreuzvalidierte Vorhersagefehler
- Falls k zu groß ist, wird die Best Subset Selection zu teuer. Beispielsweise sei $k = 20$. Dann müssten bereits $> 10^6$ Modelle berechnet werden.

Vorwärtsselektion:

- Vorgehen:
  (a) Berechne das Null-Modell M_0.
  (b) Für $l = 0, \ldots, k-1$:

(1) Berechne alle $k-l$ Modelle, welche das Modell M_l mit einem zusätzlichen Parameter betrachten.
(2) Wähle unter all diesen $k-l$ Modellen das mit dem größten R^2, erhalte M_{l+1}.

(c) Wähle unter $M_0, \ldots, M_k$ das beste Modell gemäß:
- R^2_{adj}
- AIC
- kreuzvalidierte Vorhersagefehler

- Vorteil: Das Verfahren ist weniger rechenintensiv, die Anzahl der zu berechnenden Modelle ist

$$1 + \sum_{l=0}^{k-1} (k-l) = 1 + \frac{k(k+1)}{2}.$$

Für $k = 20$ müssen also 211 Modelle berechnet werden.

- Nachteil: Die Vorwärtsselektion übersieht viele Modelle.
- Für $k > n$ muss der Algorithmus bei $l = n-1$ abgebrochen werden, da sonst keine eindeutigen Lösungen enstehen.

Rückwärtsselektion:

- Vorgehen:

(a) Berechne das volle Modell M_k.

(b) Für $l = k, \ldots, 1$:
(1) Berechne alle l Modelle, welche alle Einflussgrößen aus M_l außer einem enthalten.
(2) Wähle das Modell mit dem größten R^2, erhalte M_{l-1}.

(c) Wähle unter $M_k, \ldots, M_0$ das beste Modell gemäß:
- R^2_{adj}
- AIC
- kreuzvalidierte Vorhersagefehler

- Die Vor- und Nachteile entsprechen denen der Vorwärtsselektion. Es muss $k \leq n$ gelten.

6
Resampling

6.1 Cross Validation

Beispiel:
Seien $(x_i, y_i)_{i=1,\dots,n}$ eine Stichprobe, x_i die Körpergröße und y_i das Körpergewicht. Wir betrachten das Modell

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

für $i \in \{1, \dots, n\}$. Wir möchten nun $\hat{\beta}_0$ und $\hat{\beta}_1$ bestimmen und dann einschätzen, wie gut die beiden Schätzungen sind. Wir betrachten

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2.$$

Dieser Fehler gilt jedoch nur für die vorliegende Stichprobe und nicht zwangläufig für zukünftige Daten.

Validation Set Approach:

- Stichprobenaufteilung in Trainingsdatensatz und Validierungsdatensatz
- Schätzen der Parameter durch Trainingsdatensatz
- Überprüfung der Vorhersagegüte durch Validierungsdatensatz

Beispiel (Fortsetzung):
Seien $(x_i, y_i)_{i \in I_T}$ der Trainingsdatensatz und $(x_i, y_i)_{i \in I_V}$ der Validierungsdatensatz. Dabei gelte $I_T \cap I_V = \varnothing$ und $I_T \cup I_V = \{1, \dots, n\}$. Seien weiter $\hat{\beta}_0^{(T)}, \hat{\beta}_1^{(T)}$ Schätzer für β_0, β_1 unter Verwendung der Trainingsdaten. Dann berechnen wir den mittleren quadratischen Fehler über die Validierungsdaten:

$$\text{MSE}_V := \frac{1}{|I_V|} \sum_{i \in I_V} \left(y_i - \left(\hat{\beta}_0^{(T)} + \hat{\beta}_1^{(T)} x_i \right) \right)^2.$$

Problem:

- Bias durch ungünstige Aufteilung in Trainings- und Validierungsdaten
- Fehler kann sowohl Schätzungen betreffen als auch die Fehleinschätzung des mittleren quadratischen Fehlers

Lösung:

- M-malige zufällige Aufteilung der Daten
- Schätzen der Parameter und Berechnen der mittleren quadratischen Fehler für alle $m \in \{1, \dots, M\}$ und anschließende Mittelung

Beispiel (Fortsetzung):

Aufteilung in M Trainings- und Validierungsdatensätze:

$$(x_i, y_i)_{i \in I_T^{(1)}} \ , \ (x_i, y_i)_{i \in I_V^{(1)}}$$
$$\vdots$$
$$(x_i, y_i)_{i \in I_T^{(M)}} \ , \ (x_i, y_i)_{i \in I_V^{(M)}}$$

Wir bestimmen dann jeweils $\hat{\beta}_0^{(T,m)}, \hat{\beta}_1^{(T,m)}$ und

$$\mathrm{MSE}_V^{(m)} := \frac{1}{|I_V^{(m)}|} \sum_{i \in I_V^{(m)}} \left(y_i - \left(\hat{\beta}_0^{(T,m)} + \hat{\beta}_1^{(T,m)} x_i \right) \right)^2$$

für $m \in \{1, \dots, M\}$. Wir bestimmen dann den gemittelten Fehler

$$\mathrm{MSE}_V^{M} := \frac{1}{M} \sum_{m=1}^{M} \mathrm{MSE}_V^{(m)}.$$

Leave One Out Cross Validation:
Ein Element (Tupel) der Stichprobe dient als Validierungsdatensatz und der Rest als Validierungsdatensatz. Dies wird für jedes Element im Gesamtdatensatz durchgeführt.

Beispiel (Fortsetzung):
Für alle $m \in \{1, \dots, n\}$ teilen wir die Daten in Trainingsdatensatz $(x_i, y_i)_{i \in \{1,\dots,n\} \setminus \{m\}}$ und Validierungsdatensatz (x_m, y_m). Dann berechnen wir

$$\mathrm{MSE}_V^{(m)} = \left(y_m - \left(\hat{\beta}_0^{(T,m)} + \hat{\beta}_1^{(T,m)} x_m \right) \right)^2.$$

Diese Fehler mitteln wir und erhalten

$$\mathrm{MSE}_V^{n} = \frac{1}{n} \sum_{m=1}^{n} \mathrm{MSE}_V^{(m)}.$$

K-fold Cross Validation:
Die Daten werden in K möglichst gleichgroße Blöcke aufgeteilt. Es wird dann jeweils ein Block als Validierungsdatensatz und die restlichen $K-1$ Blöcke als Trainingsdatensatz genutzt. Üblich sind $K \in \{5, 10\}$. Zusätzlich kann man die Aufteilung in die Blöcke wiederholt durchführen.

Beispiel (Fortsetzung):
Wir definieren K Blöcke $(x_i, y_i)_{i \in I^{(k)}}$ für $k \in \{1, \dots, K\}$, wobei $I^{(k_1)} \cap I^{(k_2)} = \varnothing$ für $k_1 \neq k_2$ und $\bigcup_{k \in \{1,\dots,K\}} I^{(k)} = \{1, \dots, n\}$ gelten. Wir teilen nun für alle $k \in \{1, \dots, K\}$ unsere Daten in Trainingsdatensatz $(x_i, y_i)_{i \in I^{(l)}}$ für $l \in \{1, \dots, K\} \setminus \{k\}$ und Validierungsdatensatz $(x_i, y_i)_{i \in I^{(k)}}$ auf. Dann berechnen wir unsere Fehler

$$\mathrm{MSE}_V^{(k)} = \frac{1}{|I^{(k)}|} \sum_{i \in I^{(k)}} \left(y_i - \left(\hat{\beta}_0^{(T,k)} + \hat{\beta}_1^{(T,k)} x_i \right) \right)^2.$$

Diese Fehler mitteln wir und erhalten

$$\mathrm{MSE}_V^{K} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{MSE}_V^{(k)}.$$

Bias-Varianz-Dilemma:
Bei unterschiedlicher Nutzung der Daten zum Trainieren und Validieren können folgende Probleme auftreten:

	50:50	K-fold CV	LOOCV
Bias	$\uparrow$	0	$\downarrow$
Varianz	$\downarrow$	0	$\uparrow$

6.2 Bootstrapping

Bootstrapping wird dann angewandt, wenn Unsicherheit über die zugrunde liegende Verteilung herrscht.

Beispiel:
Seien $X = (X_1, \ldots, X_n)^\top$ eine Stichprobe mit X_i unabhängig und $\bar{X}_n$ das Stichprobenmittel. Wir wollen nun zusätzlich Konfidenzintervall bzw. Standardfehler des Schätzers angeben. Das ist wichtig für die Deskription der Daten und die Inferenz (statistische Tests). Wir wissen im Allgemeinen jedoch nicht, wie X verteilt ist.

$$\mathrm{SE}(\bar{X}_n) = \sqrt{\mathrm{Var}(\bar{X}_n)} = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i) = ?$$

Lösung:
Eine Idee wäre, eine Simulation durchzuführen. Dazu ziehen wir wiederholt Stichproben mit Zurücklegen aus der Stichprobe (Ersetzen der unbekannten Verteilung durch die empirische Verteilung).

Beispiel (Fortsetzung):
Wir ziehen nun B Stichproben (mit Zurücklegen) aus $X = (X_1, \ldots, X_n)^\top$. Wir erhalten

$$X^{(b)} = (X_{b_1}, \ldots, X_{b_n})$$

für $b \in \{1, \ldots, B\}$, $b_j \in \{1, \ldots, n\}$ und $j \in \{1, \ldots, n\}$. Wir berechnen für jede Bootstrap-Strichprobe den Mittelwert:

$$\bar{X}_n^{(b)} = \frac{1}{n} \sum_{j=1}^{n} X_{b_j}.$$

Wir bestimmen nun den Mittelwert über alle Bootstrap-Mittelwerte:

$$\bar{X}_n^B = \frac{1}{B} \sum_{b=1}^{B} \bar{X}_n^{(b)}.$$

Damit schätzen wir den Standardfehler über einen Varianzschätzer (korrigierte Stichprobenvarianz):

$$\widehat{\mathrm{SE}}(\bar{X}_n) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\bar{X}_n^{(b)} - \bar{X}_n^B \right)^2}.$$

Bemerkungen:

- Verteilung $\Longrightarrow$ empirische Verteilung
- Erwartungswert $\Longrightarrow$ Mittelwert
- Mittelwert $\Longrightarrow$ Bootstrap-Mittelwert