Applied Data Science (Prof. Dr. Kauffeldt)

Inhalt

1 Deskriptive Methoden

1.1 Statistiken

```
Analysen -> Exploration -> Deskriptivstatistik
```

Deskriptivstatistik

Deskriptivstatistik

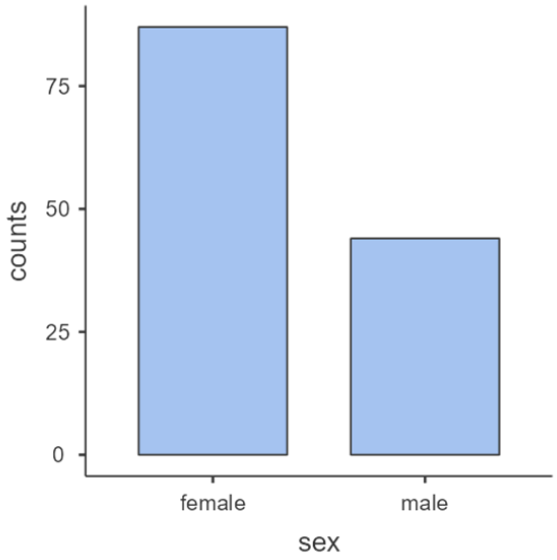
	spend_food
N	128
Fehlend	5
Mittelwert	183
Median	150
Modalwert	200
Standardabweichung	129
Varianz	16642
IQR	150
Wertebereich	800
Minimum	0
Maximum	800

Kann auch nach einer Gruppenvariable (bspw. Geschlecht) aufgeteilt werden.

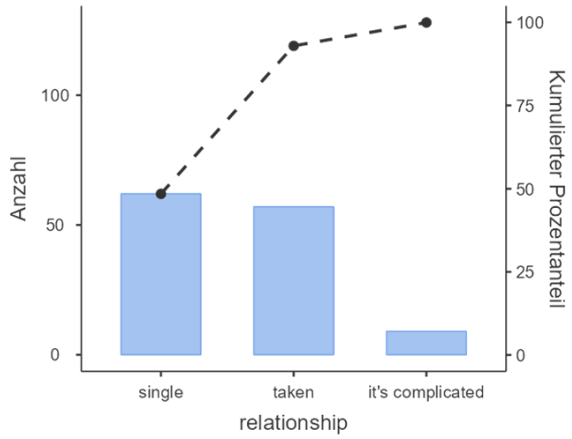
1.2 Graphiken

Nominale und Ordinale Daten: Häufigkeiten

`Analysen -> Exploration -> Deskriptivstatistik -> Diagramme -> Balkendiagramm`

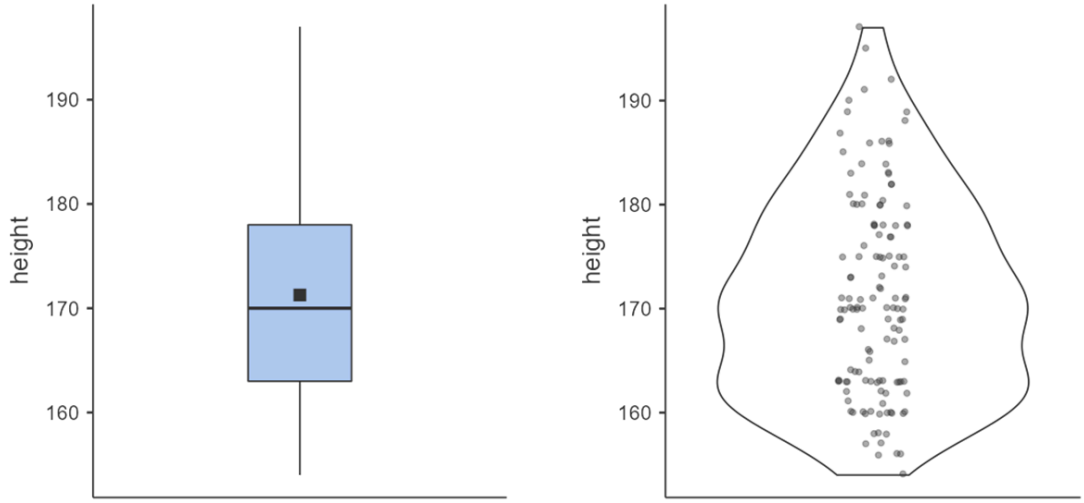


`Analysen -> Exploration -> Deskriptivstatistik -> Pareto-Diagramm`



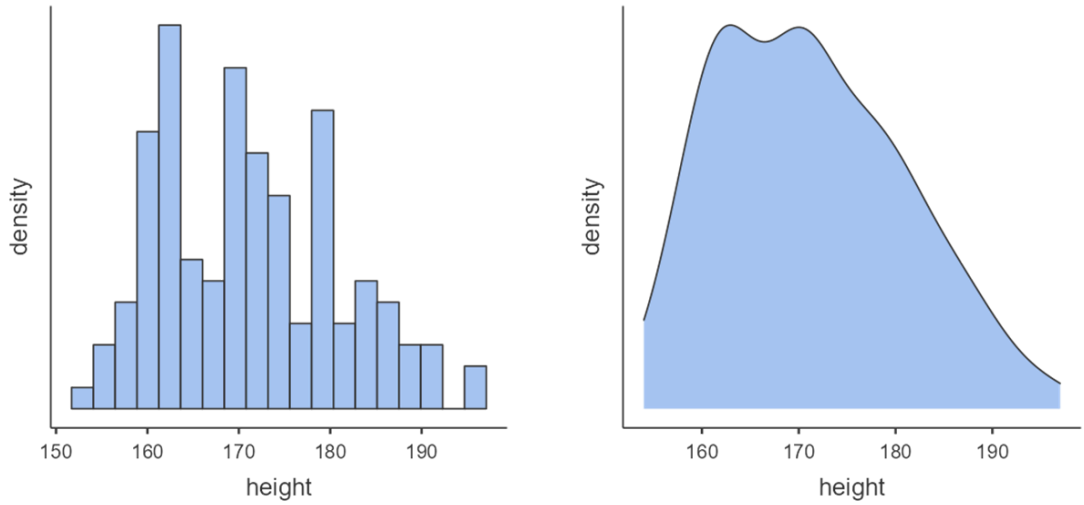
Numerische Daten und Ordinale Daten: Boxplot und Violinplot

`Analysen -> Exploration -> Deskriptivstatistik -> Diagramme -> Boxplots`



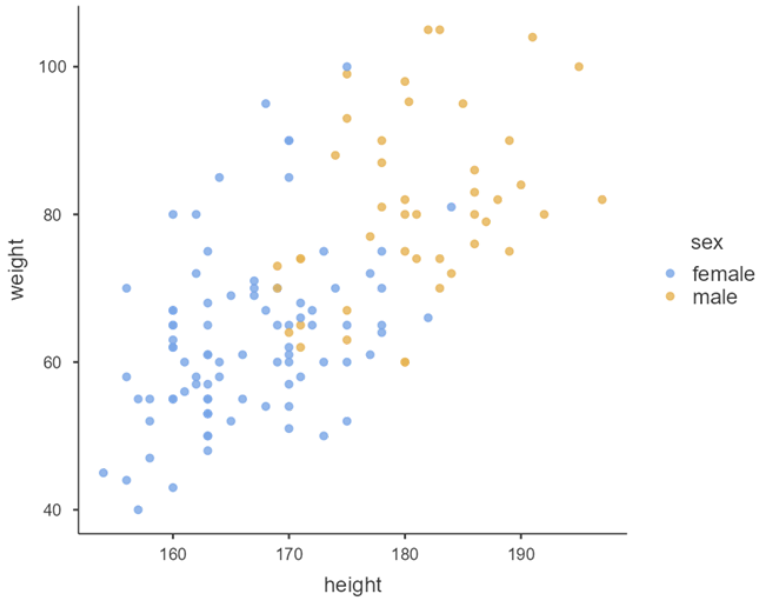
Numerische Daten: Histogramm und Dichte

`Analysen -> Exploration -> Deskriptivstatistik -> Diagramme -> Histogramme`



Bivariate numerische Daten: Streudiagramm

`Analysen -> Exploration -> Deskriptivstatistik -> Streudiagramm`



2 Testmethoden

2.1 Ablauf statistischer Test

1. **Problemstellung und Hypothesen formulieren**
   Nullhypothese H_0 ("Status Quo") und Alternativhypothese H_1 ("Forschungshypothese")
2. **Passenden statistischen Test auswählen**
3. **Voraussetzungen des Tests prüfen**
   bspw. Varianzhomogenität, Normalverteilung
4. **Ggf. Voranalyse**
5. **Ggf. Data Engineering**
   bspw. Codierung
6. **Test durchführen und interpretieren**

2.2 Testen von Lageparametern

Übersicht:

Messniveau	Test auf	Einstichprobentest	Zweistichprobentest	
			Unabhängig	*Abhängig*
Numerisch	Mittel-wert	t-Test	t-Test (Varianz-homogenität) Welch-Test	Gepaarter t-Test
Ordinal	Median	Vorzeichen-Test (Wilcoxon W) (*)	Mann-Whitney-U-Test (*)	Wilcoxon-Vorzeichen-Rang-Test (*)

(*) Nichtparametrische Tests

Beispiel: Zweistichproben t-Test (unabhängig)

Schritt 1: Problemstellung und Hypothesen formulieren

$$H_0 : Durschnittsgr\ddot{o}\ss eMann \leq Durchschnittsgr\ddot{o}\ss eFrau$$

$$H_1 : Durschnittsgr\ddot{o}\ss eMann > Durchschnittsgr\ddot{o}\ss eFrau$$

Schritt 2: Passenden Test auswählen
Unabhängiger Zweistichproben t-Test

Schritt 3: Voraussetzungen des Tests überprüfen

Voraussetzungen t-Test

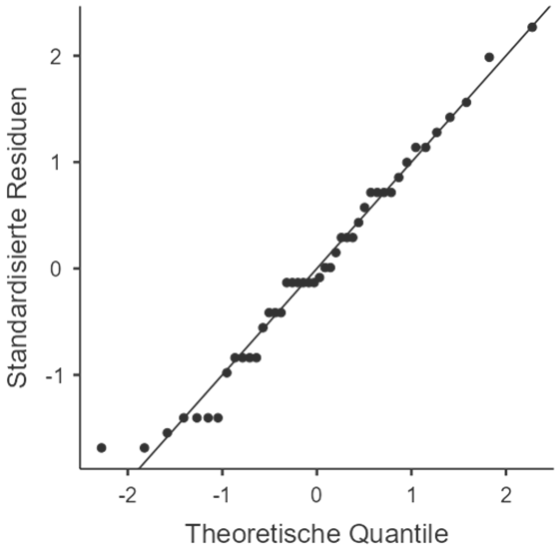
- **T1. Numerische abhängige Variable.**
- **T2. Normalität.** Die Population(en) sind normalverteilt.
- **T3. Unabhängigkeit.** Die Messungen innerhalb und zwischen den Gruppen sind unabhängig.
- **T4. Binäre Gruppenvariable.** Es werden genau zwei Gruppen verglichen. [*]
- **T5. Homoskedastizität.** Varianzhomogenität: Varianz Gruppe 1 = Varianz Gruppe 2. [*]

[*] Nur für Zweistichprobentest

T1.
Körpergröße ist numerisch. ✔

T2.
Überprüfung: Shapiro-Wilk-Test (H_0 : Normalverteilung, H_1 : Keine Normalverteilung) und QQ-Plot:
Erst nach Gruppe filtern, dann `Analysen -> Exploration -> Deskriptivstatistik -> Shapiro-Wilk und Q-Q`

Gruppe Männer:

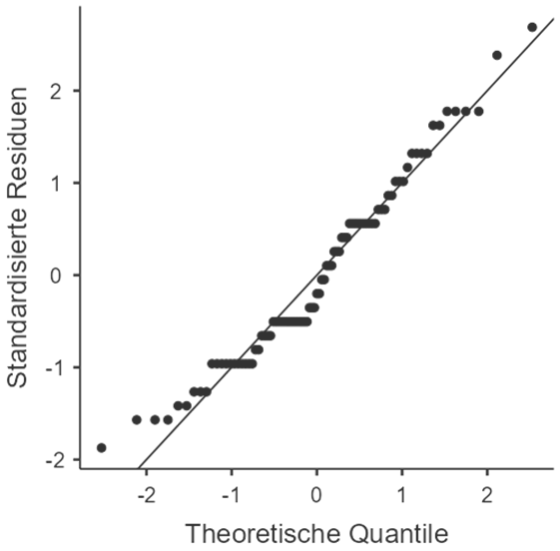


Deskriptivstatistik

	height
N	44
Fehlend	0
Mittelwert	181
Median	180
Standardabweichung	7.08
Minimum	169
Maximum	197
Shapiro-Wilk W	0.975
Shapiro-Wilk p	0.464

→ Erfüllt ✔

Gruppe Frauen:



Deskriptivstatistik

	height
N	87
Fehlend	0
Mittelwert	166
Median	165
Standardabweichung	6.58
Minimum	154
Maximum	184
Shapiro-Wilk W	0.964
Shapiro-Wilk p	0.017

→ Nicht erfüllt ✗

T3.

Messungen sind unabhängig. ✔

T4.

Nur 2 Gruppen. ✔

T5.

Überprüfung: Levenes Test (H_0 : Varianzen aller Gruppen sind gleich,

H_1 : Varianzen mindestens zweier Gruppen unterscheiden sich)

`Analysen -> t-Test für unabhängige Stichproben -> Homogenitätstest`

Levene's Test auf Varianzhomogenität

	F	df	df2	p
height	0.0233	1	129	0.879

Anmerkung. Ein niedriger p-Wert deutet auf eine Verletzung der Annahme gleicher Varianzen hin

→ Erfüllt ✔

Was tun, wenn die Voraussetzungen des Tests verletzt sind?

Abhängig von der Art der Verletzung:

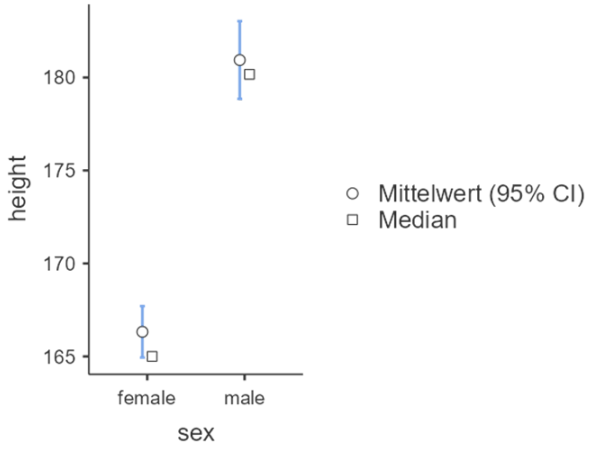
- Bei gewissen Verletzungen (bspw. abhängige Variable nicht-numerisch) kann der Test nicht durchgeführt werden
  - Bspw. bei ordinaler Variable Mann-Whitney-U-Test verwenden.
- Bei anderen Verletzungen erhalten wir weniger robuste Resultate.
  - Bei Verletzungen der Verteilungsannahme (Normalität), verwenden eines nichtparametrischen Tests zur Überprüfung der Resultate

Schritt 4: Voranalyse

`Analysen -> t-Test für unabhängige Stichproben -> Deskriptivstatistik und Deskriptive Diagramme`

Deskriptivstatistik für die Gruppen

	Gruppe	N	Mittelwert	Median	Std.-abw.	Std.-fehler
height	female	87	166	165	6.58	0.705
	male	44	181	180	7.08	1.07



Schritt 5: Test durchführen und interpretieren

`Analysen -> t-Test für unabhängige Stichproben`

t-Test für unabhängige Stichproben

								95% Konfidenzintervall			
		Statistik	±%	df	p	Mittlere Differenz	Std.-fehler der Differenz	Untere	Obere		Effektstärke
height	Student's t	-11.7		129	< .001	-14.6	1.25	-Inf	-12.5	Cohens d	-2.17
	Bayes-Faktor$_{10}$	1.12e+19	NaN								

Anmerkung. H$_a$ $\mu_{female} < \mu_{male}$

→ p-Wert < 5% → H_0 kann abgelehnt werden → Statistisch signifikant → Beleg für H_1

Effektstärke:

$$Cohens\ d = \frac{Mittelwert_1 - Mittelwert_2}{gepoolte\ Standardabweichung}$$

Cohen´s d effect size	Interpretation	Differences in SD
d = .0 – .19	Trivial effect	<1/5 from a SD
d = .20	Small effect	1/5 from a SD
d = .50	Medium effect	1/2 from a SD
d = .80 or higher	Large effect	8/10 from a SD

Bayes-Faktor:

$$BF_{10} = \frac{P(\text{Beobachtete Daten} \mid \text{H1 wahr})}{P(\text{Beobachtete Daten} \mid \text{H0 wahr})}$$

2.3 Testen von Zusammenhängen

Übersicht:

Variable 2 \ Variable 1	*Numerisch*	*Ordinal*	*Nominal*	
			nicht-binär	*binär*
Numerisch	Pearson Korrelation	Spearman ρ Kendall τ	Eta Quadrat	t-Test Punkt-Biseriale Korrelation
Ordinal		Spearman ρ Kendall τ	Chi2-Test	Mann-Whitney-U Test Cramers V
nicht-binär			Chi2-Test	Chi2-Test
Nominal / *binär*				Chi2-Test Exakter Test nach Fisher

2.3.1 Pearson Korrelationskoeffizient

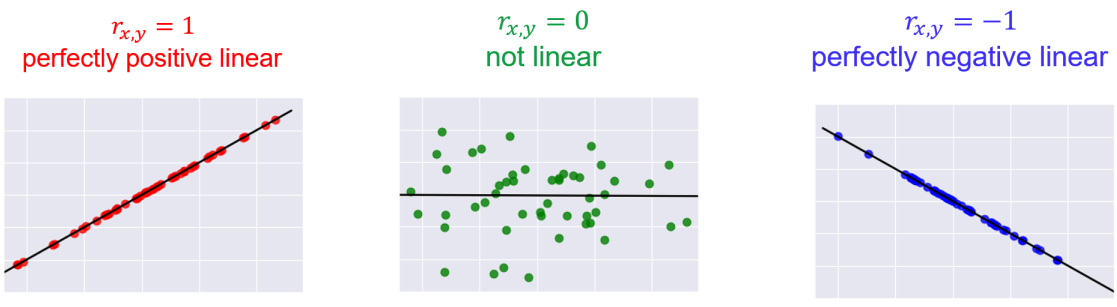
Die Kovarianz misst die lineare Beziehung zwischen zwei Variablen X und Y:

$$cov(X,Y) = \frac{(x_1 - Mittelwert_x)(y_1 - Mittelwert_y) + \ldots + (x_n - Mittelwert_x)(y_n - Mi}{n-1}$$

Da die Kovarianz von der Einheit der Messungen abhängt wird in der Praxis eine normierte Variante der Kovarianz verwendet - der *Pearson Korrelationskoeffizient r*:

$$r = \frac{cov(X,Y)}{Standardabweichung_X \cdot Standardabweichung_Y}$$

Der Pearson Korrelationskoeffizient kann nur Werte zwischen -1 und +1 annehmen, wobei -1 eine perfekte negative lineare Beziehung und +1 eine perfekte positive lineare Beziehung anzeigt:



Beispiel: Korrelationstest mit Pearson Korrelationskoeffizient (Pearsons r)

Schritt 1: Problemstellung und Hypothesen formulieren
Wir wollen wissen, ob Größe und Gewicht positiv korreliert sind. Wir müssen also testen, ob der Korrelationskoeffizient signifikant positiv ist.

$$H_0 : r_{GrößeGewicht} \leq 0$$

$$H_1 : r_{GrößeGewicht} > 0$$

Schritt 2: Passenden Test auswählen
Da Größe und Gewicht beide numerisch → Pearson Korrelationskoeffizient

Schritt 3: Voraussetzungen des Tests überprüfen

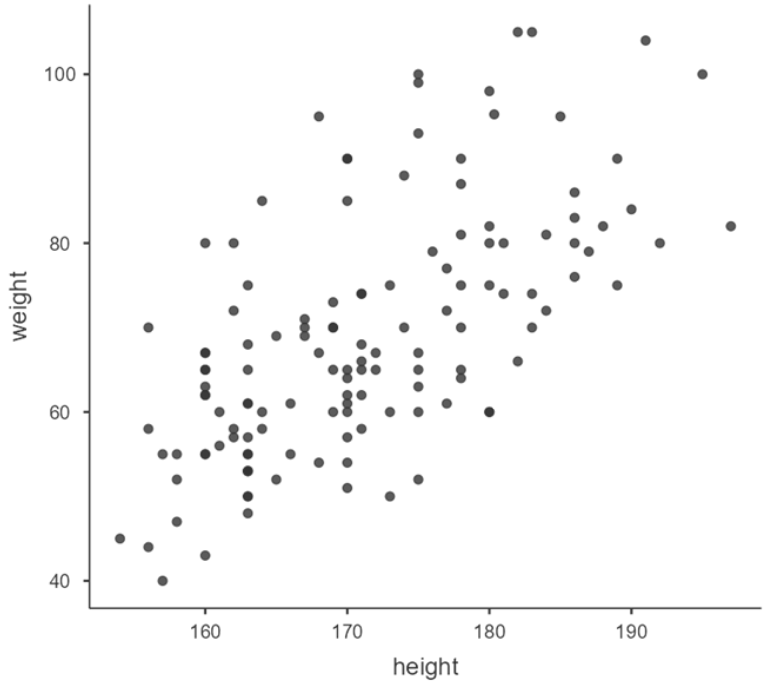
Vorraussetzungen Pearson Korrelationstest

- **PK1. Beide Variablen sind numerisch.**
- **PK2. Normalität.** Die Variablen sind normalverteilt.
- **PK3. Unabhängigkeit.** Die Messungen sind unabhängig.

Besprechung der Voraussetzungen → Vorlesung.

Schritt 4: Voranalyse

```
Analysen -> Exploration -> Streudiagramm
```



Schritt 5: Test durchführen und interpretieren

```
Analysen -> Regression -> Korrelationsmatrix
```

Korrelationsmatrix

		height	weight
height	Pearson's r	—	
	df	—	
	p-Wert	—	
	N	—	
weight	Pearson's r	0.640 ***	—
	df	130	—
	p-Wert	< .001	—
	N	132	—

Anmerkung. H_a ist eine positive Korrelation

Anmerkung. * p < .05, ** p < .01, *** p < .001, einseitig

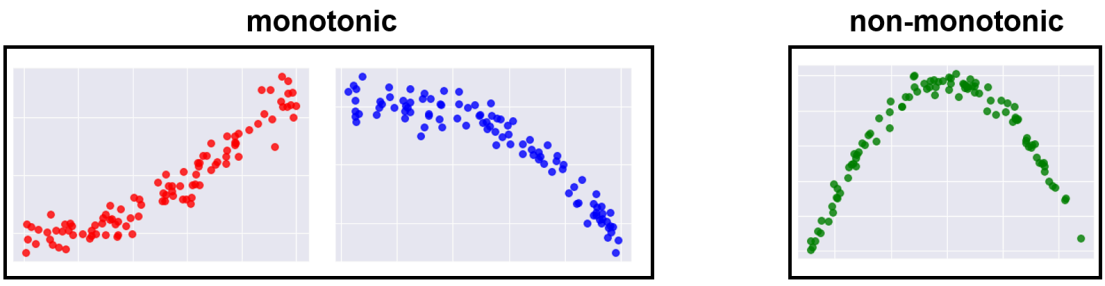
Interpretation	Correlation value
Small correlation	0.10 to 0.29
Medium correlation	0.30 to 0.49
Large correlation	0.50 to 1.0

Außerdem ist $r^2 = 0,64^2 = 40,96\%$ der Anteil der Varianz, den die Variablen teilen und der somit erklärt wird.

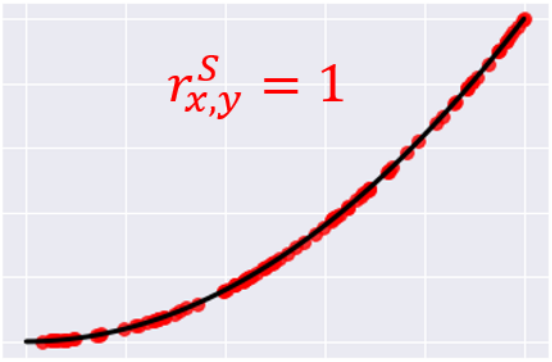
2.3.2 Spearman Rho und Kendall Tau Korrelationskoeffizient

Im Falle ordinaler Variablen können wir die Kovarianz nicht berechnen, da wir keinen Mittelwert berechnen können. Korrelationskoeffizienten werden mithilfe von *rangbasierten* Ansätzen bestimmt, die die Daten der Größe nach ordnen und jeder Beobachtung entsprechend ihrer Position einen Rang zuweisen. Beliebte rangbasierte Korrelationskoeffizienten sind: Spearmans Rho und Kendalls Tau.

Rangbasierte Ansätze identifizieren allgemeinere monotone Zusammenhänge:



Ein Koeffizient von 1 zeigt eine perfekt positive monotone Beziehung an:



Voraussetzungen rangbasierte Korrelationstests

- **RK1. Ordinal.** Beide variablen sind mindestens ordinal.
- **RK2. Unabhängigkeit.** Die Messungen sind unabhängig.

Spearmans Rho:
Dieser Koeffizient funktioniert genauso wie Pearsons r, mit dem Unterschied, dass er die Kovarianz und die Standardabweichungen in Bezug auf die Ränge anstelle der Werte der Variablen berechnet.

$$r^S = \frac{cov(Rang(X), Rang(Y))}{Standardabweichung_{Rang(X)} \cdot Standardabweichung_{Rang(Y)}}$$

Kendall Tau:
"Diese Koeffizienten basieren auf der Anzahl der konkordanten und diskordanten Paare in einem Datensatz. Gegeben zwei Variablen X und Y, sind zwei Beobachtungspaare (x_i, y_i) und (x_j, y_j)

- *konkordant* wenn $x_i > x_j$ and $y_i > y_j$ oder if $x_i*diskordant* wenn $x_i > x_j$ and y_ix_j $und y_i Beispiele:
  - (1,3) und (6,9) sind konkordant
  - (3,1) und (6,9) sind diskordant

Beziehung zwischen dem Spearman und Kendall Koeffizient:

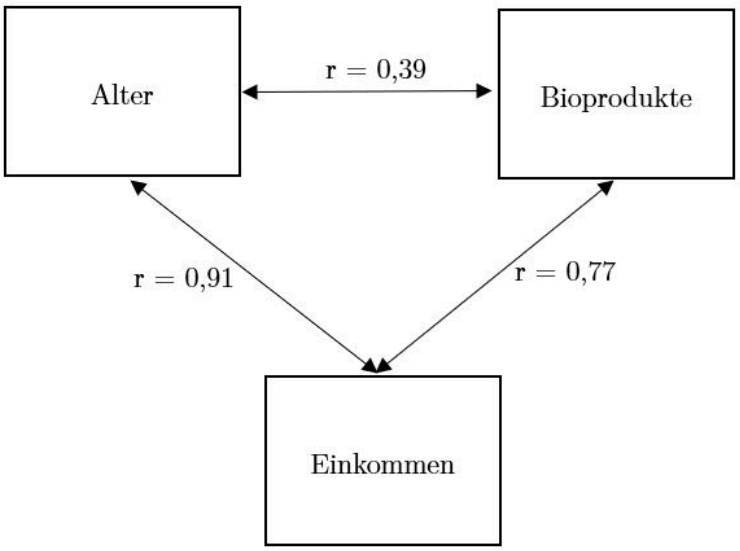
$$Kendall \approx 0.7 \cdot Spearman$$

Ausführlichere Beispiele zu Spearman Rho und Kendall Tau → Vorlesung.

2.3.3 Partielle Korrelation

Beim Testen von Korrelationen müssen wir potenzielle Störvariablen berücksichtigen. Angenommen, wir möchten testen, ob das Alter mit dem Kauf von Bio-Produkten korreliert.

Dann müssen wir auch berücksichtigen, dass das Alter mit dem Einkommen korreliert, das wiederum mit dem Kauf der (teureren) Bio-Produkte korreliert sein könnte.



Die Korrelation von 0,39 könnte zum Teil auf die positive Korrelation zwischen Alter und Einkommen zurückzuführen sein. Daher müssen wir den Effekt des Einkommens eliminieren. Die partielle Korrelationsanalyse bietet eine Möglichkeit, dies zu tun.

Partieller Korrelationskoeffizient:
Seien X, Y und Z drei Variablen. Angenommen, wir möchten die Korrelation zwischen X und Y untersuchen, während wir für Z kontrollieren. Der angepasste Korrelationskoeffizient ist dann:

$$r_{XY,Z} = \frac{r_{XY} - r_{XZ} \cdot r_{YZ}}{\sqrt{1 - r_{XZ}^2} \cdot \sqrt{1 - r_{YZ}^2}}$$

2.3.4 Unabhängigkeitstests

Um zu testen, ob eine nominale und eine nominale oder ordinale Variable miteinander assoziiert sind, können wir einen Unabhängigkeitstest verwenden:

Chi2 (χ^2) Test oder exakter Test nach Fisher (bei 2×2 Kontingenztafeln)

Hypothesen:

H_0 : Die Variablen X und Y sind unabhängig und H_1 : Die Variablen X und Y sind abh

Beispiel: Unabhängigkeitstest

Schritt 1: Problemstellung und Hypothesen formulieren
Wir wollen wissen, ob Haarfarbe und Augenfarbe voneinander abhängen.

$$H_0 : \text{Haarfarbe und Augenfarbe sind unabhängig}$$

$$H_1 : \text{Haarfarbe und Augenfarbe sind abhängig}$$

Schritt 2: Passenden Test auswählen
Da Haarfarbe (black, blonde, brown, red) und Augenfarbe (blue, brown, green) zu einer 4×3-Tafel führen → Chi2-Test

Schritt 3: Voraussetzungen des Tests überprüfen

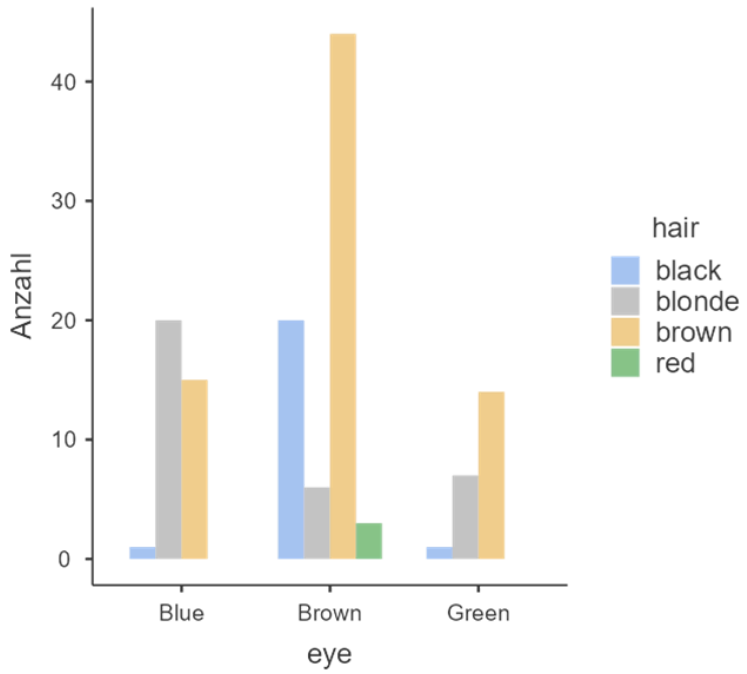
Vorraussetzungen Chi2-Unabhängigkeitstest

- **C1. Beide Variablen sind kategorial.**
- **C2. Große Stichprobe.** Daumenregel: n > 50.
- **C3. Hinreichend große erwartete Häufigkeiten.** Alle erwarteten Häufigkeiten > 5.
- **C4. Unabhängigkeit.** Messungen sind unabhängig.

Besprechung der Voraussetzungen → Vorlesung.

Schritt 4: Voranalyse

```
Analysen -> Häufigkeiten -> Unabhängige Stichproben -> Balkendiagramm
```




```
Analysen -> Häufigkeiten -> Unabhängige Stichproben -> Anzahl Beobachtet / Erwartet
```

Kreuztabellen

eye	hair				Insgesamt
	black	blonde	brown	red	
Blue	1	20	15	0	36
Brown	20	6	44	3	73
Green	1	7	14	0	22
Insgesamt	22	33	73	3	131

Kreuztabellen

eye		hair				Insgesamt
		black	blonde	brown	red	
Blue	Erwartet	6.05	9.07	20.1	0.824	36.0
Brown	Erwartet	12.26	18.39	40.7	1.672	73.0
Green	Erwartet	3.69	5.54	12.3	0.504	22.0
Insgesamt	Erwartet	22.00	33.00	73.0	3.000	131.0

Wie berechnet man die erwarteten Häufigkeiten E?

$$E_{Zeile\ i,Spalte\ j} = \frac{(Beobachtet\ Zeile\ i) \times (Beobachtet\ Spalte\ j)}{Beobachtungen\ Gesamt}$$

Beispiel:

$$E_{eye\ blue,hair\ black} = \frac{36 \times 22}{131} \approx 6,05$$

Wie berechnet man die Chi2-Teststatistik?

Allgemein für eine $n \times m$ Kontingenztafel mit B = beobachtete Häufigkeit und E = erwartete Häufigkeit:

$$\chi^2 = \frac{(B_{1,1} - E_{1,1})^2}{E_{1,1}} + \cdots + \frac{(B_{n,m} - E_{n,m})^2}{E_{n,m}}$$

Im Beispiel:

$$\chi^2 = \frac{(1 - 6,05)^2}{6,05} + \cdots + \frac{(0 - 0,504)^2}{0,504} \approx 37,1$$

Wie bestimmt man die Freiheitsgrade eines Chi2-Tests?

$$Freiheitsgrade = (AnzahlZeilen - 1) \times (AnzahlSpalten - 1)$$

Im Beispiel:

$$Freiheitsgrade = 2 \times 3 = 6$$

Schritt 5: Test durchführen und interpretieren

```
Analysen -> Häufigkeiten -> Unabhängige Stichproben -> Tests Chi2
```

χ^2-Tests

	Wert	df	p
χ^2	37.1	6	< .001
N	131		

Effektstärke:

`Analysen -> Häufigkeiten -> Unabhängige Stichproben -> Phi und Cramers V`

Nominal	
	Wert
Phi-Koeffizient	NaN
Cramer's V	0.377

$$Cramers\ V = \sqrt{\frac{\chi^2/n}{min(Z-1, S-1)}},$$

wobei n = Stichprobengröße, Z = Anzahl Zeilen, S = Anzahl Spalten

Value of φ or Cramer's V	Description
.00 and under .10	Negligible association
.10 and under .20	Weak association
.20 and under .40	Moderate association
.40 and under .60	Relatively strong association
.60 and under .80	Strong association
.80 to 1.00	Very strong association

Quelle: Rea, L. M., and Parker, R. A. (1992). Designing and conducting survey research. San Francisco: Jossey-Boss.

2×2-Kontingenztafel

Wenn jede Variable 2 Kategorien hat, kann man entweder einen Chi2-Test mit Kontinuitätskorrektur oder einen exakten Test nach Fisher (Voraussetzungen siehe unten) durchführen:

`Analysen -> Häufigkeiten -> Unabhängige Stichproben -> Tests`

Voraussetzungen exakter Test nach Fisher

- **E1. Binäre kategoriale Variablen.**
- **E2. Unabhängigkeit.**

Effektstärken bei 2×2-Kontingengenztafeln

`Analysen -> Häufigkeiten -> Unabhängige Stichproben -> Verglichene Maße`

Gegeben folgende Kontingenztafel mit beobachteten Häufigkeiten a, b, c, d:

a	b
c	d

$$Odds - Ratio(OR) = \frac{a/b}{c/d}$$

$$RelativesRisiko(RR) = \frac{a/(a+b)}{c/(c+d)}$$

Interpretation → Vorlesung.

2.4 Multiples Testen

Problem: Multiple Tests führen zu einer Alphafehler-Inflation

Mögliche Korrekturen (p-Wert-Anpassungen) bei multiplen Tests:

Bonferroni-Korrektur:

$$p_{bonf} = p_{unangepasst} \times (\text{Anzahl Tests})$$

Sidak-Korrektur:

$$p_{sid} = 1 - (1 - p_{unangepasst})^{(\text{Anzahl Tests})}$$

Holm-Bonferroni-Korrektur:

- Sortiere die unangepassten p-Werte von niedrig nach hoch: $p(1)<\dots Passe den iten p-Wert wie folgt an:

$$p_{hbonf} = (\text{Anzahl Tests} - i + 1) \cdot p_{unangepasst}$$

Benjamini Hochberg:

- Sortiere die unangepassten p-Werte von niedrig nach hoch: $p(1)<\dots Multipliziere jeden p-Wert mit der Anzahl der Tests und dividiere ihn durch seinen Rang. ($\frac{p \cdot t}{i}$)
- Die resultierende Sequenz sollte nicht abnehmen. Falls sie abnimmt, setze den vorherigen p-Wert gleich dem nachfolgenden. Wiederhole diesen Schritt, bis die Sequenz nicht mehr abnimmt.

Beispiele → Vorlesung.

Die Anpassungen können von konservativ (lehnen die Nullhypothese seltener) ab nach liberal geordnet werden:

$$\text{konservativ -- Bonferroni -- Sidak -- Holm-Bonferroni -- Benjamini-Hochberg}$$

3 Regressionsmodelle

Überblick:

Messniveau Abhängige Varible	**Regression**
Numerisch	Linear
Kategorial (2 Kategorien)	Logistisch
Ordinal (>2 Kategorien)	Ordinal
Nominal (>2 Kategorien)	Multinomial

3.1 Lineare Regression

3.1.1 Grundlegende Idee

In den Sozialwissenschaften möchten wir oft untersuchen, ob eine unabhängige Variable (X) eine abhängige Variable (Y) beeinflusst. Zum Beispiel: Erhöhen Marketingausgaben (X) den Umsatz (Y)?

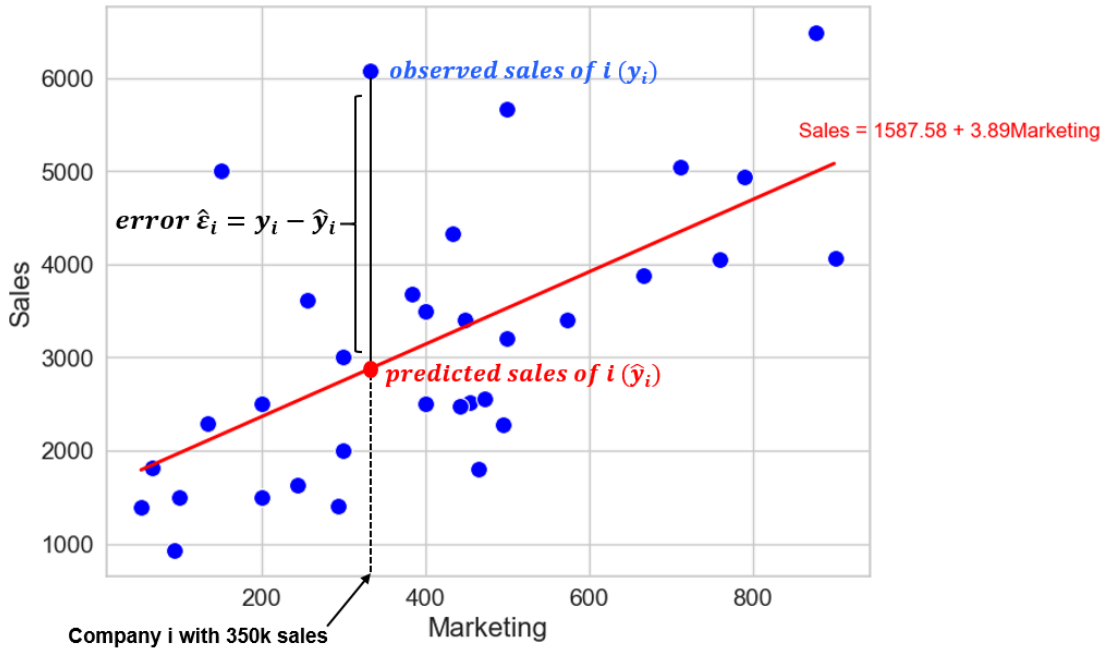
Die Idee der **linearen Regression** besteht darin, eine Gerade an die Daten anzupassen. Die theoretische Gleichung eines solchen Modells lautet:

$$Sales = \beta_0 + \beta_1 \cdot Marketing + \varepsilon,$$

wobei

- β_0, β_1 = Achsenabschnitt, Steigung
- ε = Fehlerterm, der Variablen berücksichtigt, die nicht in der Gleichung enthalten sind (z. B. Reputation des Unternehmens, Produktqualität)

Wie findet man die optimale Gerade?



Jede Gerade führt zu spezifischen Fehlern. Wir möchten diejenige Gerade finden, die die Fehler minimiert – also die Linie, die den Datenpunkten am nächsten liegt.

Wie aggregieren wir die Fehler?

- Die totale Summe der Fehler ($\varepsilon_1 + \cdots + \varepsilon_n$) hat den Nachteil, dass sich negative und positive Abweichungen teilweise gegenseitig aufheben.

Die Lösung besteht darin, die gesamte Summe der *quadrierten* Fehler (SSR = Sum of Squared Residuals) zu minimieren (*Ordinary Least Squares (OLS)*-Ansatz):

$$SSR = \varepsilon_1^2 + \cdots + \varepsilon_n^2$$

Im Fall einer einfachen linearen Regression (eine abhängige Variable Y, eine unabhängige Variable X) können die geschätzten Regressionskoeffizienten wie folgt berechnet werden:

$$\hat{\beta}_1 = \frac{cov(X,Y)}{var(X)} \ und \ \hat{\beta}_0 = \frac{\Sigma y_i - \hat{\beta}_1 \Sigma x_i}{n}.$$

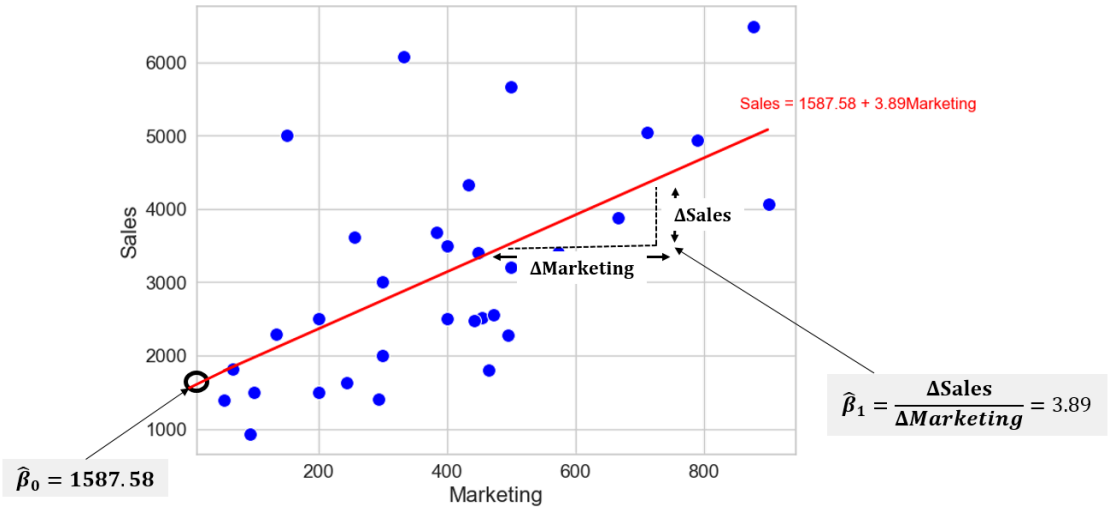
Im obigen Beispiel:

- Kovarianz(Marketing, Umsatz) = 207797,65
- Varianz(Marketing) = 53472,15
- Summe Umsatz = 104432,81
- Summe Marketing = 13392
- n = 33

Daher (Unterschiede zur geschätzten Geraden oben ergeben sich durch Rundungsdifferenzen):

$$\hat{\beta}_1 = \frac{207797.65}{53472.15} \approx 3.89 \ \ und \ \ \hat{\beta}_0 = \frac{104432.81 - 3.89 \cdot 13392}{33} \approx 1585.99.$$

Interpretation der Koeffizienten:



- Achsenabschnitt ($\hat{\beta}_0$): Der geschätzte Umsatz eines Unternehmens mit 0 Marketingausgaben beträgt 1'587'580 $
- Steigung ($\hat{\beta}_1$): Eine Einheitserhöhung der Marketingausgaben erhöht den Umsatz um ca. 3,89 $

3.1.2 Multiple Lineare Regression

Wir möchten möglicherweise weitere unabhängige Variablen zu unserem Modell hinzufügen. Die theoretische Regressionsgleichung mit k unabhängigen Variablen lautet wie folgt:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_k \cdot X_k + \varepsilon$$

Konzeptionell unterscheiden wir zwischen Zielvariablen und Kontrollvariablen. Angenommen, X_1 und X_2 sind die Zielvariablen. Wir untersuchen ihren Einfluss auf Y, während wir für $X_3, \ldots, X_k$ kontrollieren.

Die Mathematik funktioniert ähnlich wie beim einfachen Modell, aber nun müssen wir uns eine Regressionsebene (oder einen Hyperebene) vorstellen.

Beispiel:

$$Sales = \beta_0 + \beta_1 \cdot Marketing + \beta_2 \cdot Quality + \varepsilon,$$

wobei *Quality* durch die durchschnittliche Produktlebensdauer gemessen wird.

3.1.3 Dummy-/One-Hot-Codierung bei kategorialen unabhängigen Variablen

Kategoriale Variablen (nominal oder ordinal) nehmen Kategorien als Werte an. Eine Regressionsgleichung kann jedoch nur mit Zahlen arbeiten.
Da nominale Varialen keine Ordnung besitzen, müssen wir diese in Indikatorvariablen umwandeln. Ordinale Variablen kann man entweder entsprechend deren Ordnung kodieren oder ebenfalls Indikatorvariablen verwenden. Bei der Kodierung wird implizit angenommen, dass die Abstände von aufeinanderfolgenden Kategorien gleich sind. Dies kann problematisch sein. Im Falle von Indikatorvariablen verliert man die Ordnung bei ordinalen Variablen.

Beispiel: Variable „Location", die die Kategorien *Asia, Europe, USA* annehmen kann.

Indikatorvariablen:

	Indikator Asia	Indikator Europe	Indikator USA
Asia	1	0	0
Europe	0	1	0
USA	0	0	1

Wir können jedoch nicht alle 3 Indikatorvariablen verwenden, da zwei immer die dritte perfekt vorhersagen. Diese Abhängigkeit würde dazu führen, dass das Regressionsmodell kollabiert. Daher müssen wir eine der Kategorien weglassen (welche ist egal). Die ausgelassene Kategorie dient als Referenzkategorie: Alle Effekte werden im Verhältnis zu dieser Kategorie gemessen. Im Allgemeinen hat man bei K Kategorien K-1 Indikatorvariablen.

3.1.4 Beispiel Lineare Regression

Schritte Regressionsanalyse

1. Schreibe vermutete Ursache-Wirkungs-Beziehungen mit Kontrollvariablen auf.
2. Überprüfen der Voraussetzungen.
3. Schreiben Sie die geschätzte Regressionsgleichung auf und interpretieren Sie das Ergebnis.
4. Ggf. weitere Robustheitsprüfungen.

Schritt 1: Theoretische Regressionsgleichung

Wir möchten untersuchen, wie sich Marketingausgaben auf den Umsatz auswirken, während wir für den Standort und die Reputation eines Unternehmens kontrollieren. Reputation ist eine ordinale Variable und nimmt die Werte *High, Medium, Low* an. Wir entscheiden uns diese zu kodieren (2, 1, 0), statt Indikatorvariablen zu verwenden.

$$Sales = \beta_0 + \beta_1 \cdot Marketing + \beta_2 \cdot Quality + \beta_3 \cdot Location_{Asia} + \beta_4 \cdot Location_{Europe} +$$

Schritt 2: Voraussetzungen überprüfen

Voraussetzungen Linearen Regression

- **LR1. Numerische abhängige Variable.**
- **LR2. Linearität.** Es besteht eine lineare Beziehung zwischen der abhängigen und den unabhängigen Variablen.
- **LR3. Fehlen perfekter (Multi)kollinearität.** Es besteht keine perfekte lineare Beziehung zwischen unabhängigen Variablen.
- **LR4. Strikte Exogenität.** Die bedingten Mittelwerte der Fehler sind null ($E[\varepsilon_i \mid x_i] = 0$).
- **LR5. Homoskedastizität.** Die Fehler haben für alle i,j gleiche bedingte Varianzen ($var(\varepsilon_i \mid x_i) = var(\varepsilon_j \mid x_j)$).
- **LR6. Keine Autokorrelation.** Die Fehler sind für alle i,j nicht korreliert ($cov(\varepsilon_i, \varepsilon_j) = 0$).
- **LR7. Normalität.** Die Fehler folgen einer multivariaten Normalverteilung.

LR1 bis LR3 beziehen sich auf die Variablen, die restlichen Voraussetzungen auf die Fehlerterme.

LR1.
Umsatz ist numerisch. ✔

LR2.
`Analysen -> Regression -> Lineare Regression -> Modellanpassung -> F-Test`

Güte der Modellanpassung

			Test des Gesamtmodells			
Modell	R	R^2	F	df1	df2	p
1	0.841	0.707	16.9	4	28	< .001

F-Test:

$$H_0 : \beta_1 = \cdots = \beta_5 \ \text{(das Modelle erklärt nichts)}$$

$$H_1 : \beta_i \neq 0 \ \text{für mindestens ein i} = 1,2,3 \ \text{(das Modelle erklärt etwas)}$$

Erfüllt ✔

LR3.
`Analysen -> Regression -> Lineare Regression -> Überprüfung der Voraussetzungen -> Kollinearitätsstatistik`

Kollinearitätsstatistik

	VIF	Toleranz
Marketing	1.13	0.883
Location	1.02	0.985
Reputation_coded	1.15	0.873

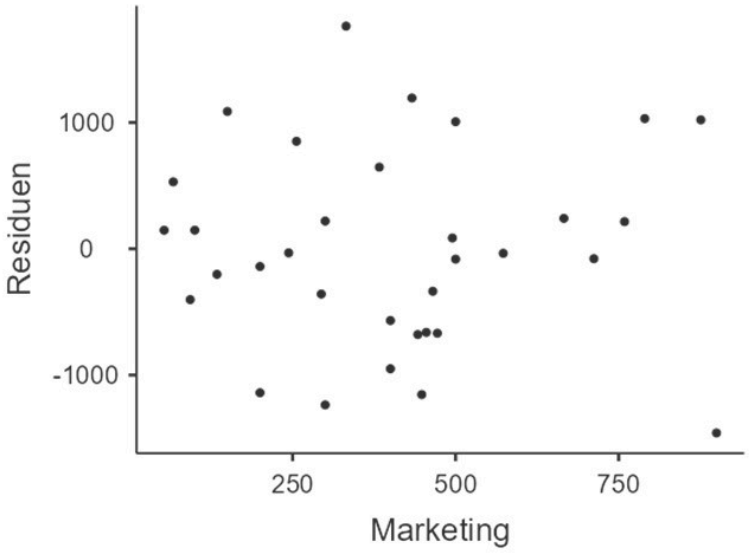
Erfüllt ✔

LR4.

Das direkte Überprüfen der strikten Exogenität kann herausfordernd sein, da es sich nicht um etwas handelt, das man direkt mit einem statistischen Test testen kann.
Oft ist der beste Ansatz, die theoretische Grundlage für strikte Exogenität zu berücksichtigen. Zum Beispiel, wenn Sie es mit Zeitreihendaten zu tun haben, überlegen Sie, ob vergangene Werte der abhängigen Variablen oder der Prädiktoren den Fehlerterm beeinflussen könnten. Ebenso, wenn Sie mit Querschnittsdaten arbeiten, überlegen Sie, ob der Fehlerterm durch unbeobachtete Faktoren, die mit den Prädiktoren zusammenhängen, beeinflusst werden könnte.
Ein nützlicher erster Schritt ist es jedoch, die Residuen (Fehler) des Regressionsmodells gegen die Prädiktorvariablen zu plotten.

- Wenn in diesen Plots ein systematisches Muster zu erkennen ist (z. B. eine gekrümmte Beziehung), könnte dies darauf hindeuten, dass die Annahme der strikten Exogenität verletzt ist, da es eine Abhängigkeit zwischen den Residuen und den Prädiktoren impliziert.
- Idealerweise sollten die Residuen beim Plotten gegen eine unabhängige Variable eine zufällige Streuung um Null zeigen (kein Muster).

`Analysen -> Regression -> Lineare Regression -> Überprüfung der Voraussetzungen -> Diagramme der Residuen`

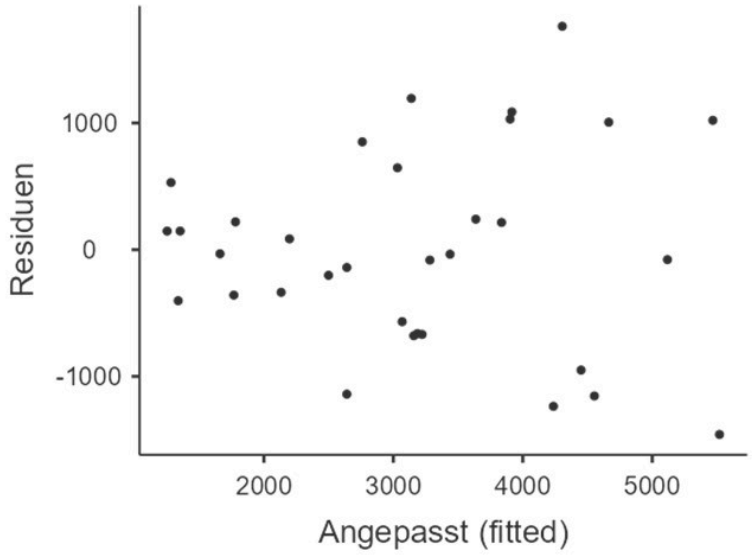


LR 5.

Visuell: Eine der einfachsten und häufigsten Methoden zur Überprüfung der Homoskedastizität besteht darin, die Residuen gegen die angepassten Werte (vorhergesagten Werte) aus Ihrem Regressionsmodell zu plotten.

- So geht's: Berechnen Sie zuerst die Residuen: Residuen = Beobachtet − Vorhergesagt. Dann plotten Sie die Residuen auf der y-Achse und die angepassten Werte auf der x-Achse.
- Worauf Sie achten sollten: Wenn der Plot eine zufällige Streuung zeigt (kein klares Muster), weist dies auf Homoskedastizität hin. Wenn der Plot ein klares Muster zeigt, wie zum Beispiel eine Trichterform (die Residuen nehmen mit den angepassten Werten zu oder ab), weist dies auf Heteroskedastizität hin.

`Analysen -> Regression -> Lineare Regression -> Überprüfung der Voraussetzungen -> Diagramme der Residuen`



LR6.

`Analysen -> Regression -> Lineare Regression -> Überprüfung der Voraussetzungen -> Autokorrelationstest`

Durbin-Watson-Autokorrelationstest

Autokorrelation	DW-Statistik	p
−0.0260	1.99	0.930

Durbin-Watson Autokorrleationstest:

$$H_0 : KeineAutokorrelation$$

$$H_1 : Autokorrelation$$

Erfüllt ✔

LR7.

`Analysen -> Regression -> Lineare Regression -> Überprüfung der Voraussetzungen -> Test auf Normalverteilung`

Test auf Normalverteilung (Shapiro-Wilk)	
Statistik	p
0.978	0.732

Shapiro-Wilk Test:

$$H_0 : Normalverteilung$$

$$H_1 : KeineNormalverteilung$$

Erfüllt ✔

Schritt 3: Schreiben Sie die geschätzte Regressionsgleichung auf und interpretieren Sie das Ergebnis.

`Analysen -> Regression -> Lineare Regression`

Modellkoeffizienten - Sales

Prädiktor	Schätzung	Std.-fehler	t	p
Interzept [a]	1011.58	364.120	2.778	0.010
Marketing	2.03	0.729	2.793	0.009
Location:				
Asia – USA	163.59	371.754	0.440	0.663
Europe – USA	112.38	356.365	0.315	0.755
Reputation_coded	1236.89	227.282	5.442	< .001

[a] Repräsentiert das Referenzniveau

Geschätzte Regressionsgleichung (Koeffizienten gerundet auf ganze Zahlen):

$$Sales = 1012 + 2 \cdot Marketing + 164 \cdot Location_{Asia} + 112 \cdot Location_{Europe} + 1237 \cdot Repu$$

Interpretation → Vorlesung.

3.1.5 Potentielle Probleme

Überblick:

1. Überanpassung des Modells an die Daten (Overfitting)
2. Moderationseffekt
3. Mediationseffekt

4. Omitted Variable Bias

Überanpassung des Modells an die Daten (Overfitting)

Das Hinzufügen zusätzlicher unabhängiger Variablen zum Modell erhöht stets R^2. Allerdings ist nicht klar, ob dies auf einen tatsächlichen kausalen Effekt oder lediglich auf einen mechanischen Zusammenhang zurückzuführen ist: Wenn wir unabhängige Variablen hinzufügen, verlieren wir Freiheitsgrade. Dies kann zu einem künstlichen Anstieg von R^2 führen.

Zum Beispiel, wenn keine Freiheitsgrade mehr vorhanden sind Anzahl der Unabhängigen Variablen (k) -1 = Anzahl der Beobachtungen (n), erhalten wir immer $R^2 = 1$, und zwar rein aus mathematisch-mechanischen Gründen.

Ein solcher künstlicher Anstieg des R^2 wird als *Overfitting* bezeichnet. Das bedeutet, dass das Modell zu stark an die vorhandenen Daten angepasst wurde.

Eine Möglichkeit, Overfitting zu erkennen, ist das adjustierte R^2_{adj}. Im Gegensatz zu R^2 berücksichtigt es die Freiheitsgrade und kann sinken, wenn zusätzliche unabhängige Variablen hinzugefügt werden:

$$R^2_{adj} = \left(\boldsymbol{R^2} - \frac{\boldsymbol{k}}{\boldsymbol{n-1}} \right) \cdot \left(\frac{\boldsymbol{n-1}}{\boldsymbol{n-k-1}} \right)$$

`Analysen -> Regression -> Lineare Regression -> Modellanpassung -> Adjustiertes R2`

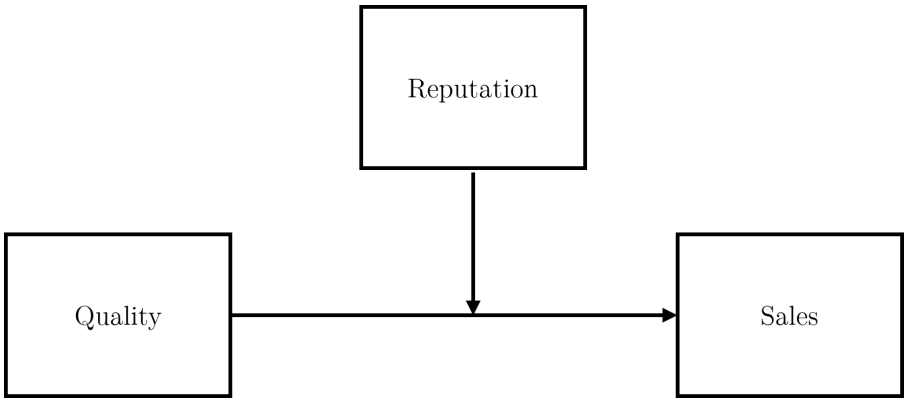
Güte der Modellanpassung

Modell	R	R^2	Adjustiertes R^2
1	0.886	0.785	0.745

Moderationseffekt

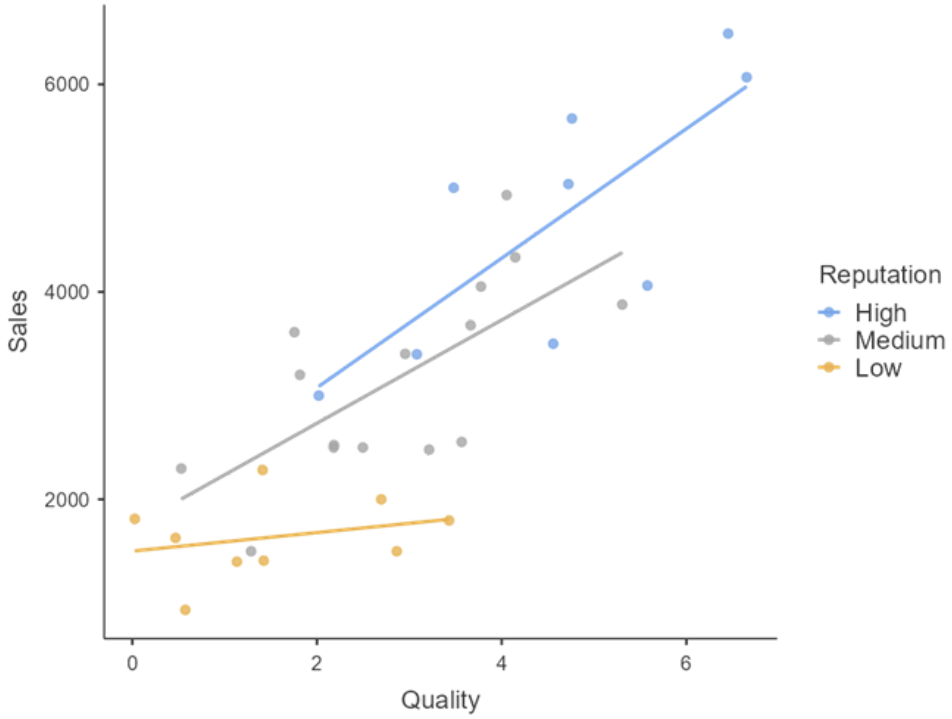
Moderator = Dritte Variable, die die Stärke des Effekts bestimmt.

Beispiel: Qualität könnte einen schwächeren Effekt auf den Umsatz haben, wenn das Unternehmen eine geringere Reputation hat.



Die Steigungen der Reputation „mittel“ und „hoch“ scheinen ziemlich ähnlich zu sein. Die Steigung von „niedrig“ ist jedoch flacher:

```
Analysen -> Regression -> Exploration -> Streudiagramm
```



Um zu testen, ob tatsächlich eine Moderation vorliegt, müssen wir die Interaktionen $quality \times reputation$ in das Modell aufnehmen. Wenn diese Interaktionen signifikant ist, haben wir einen Hinweis auf einen Moderationseffekt gefunden:

Modellkoeffizienten - Sales

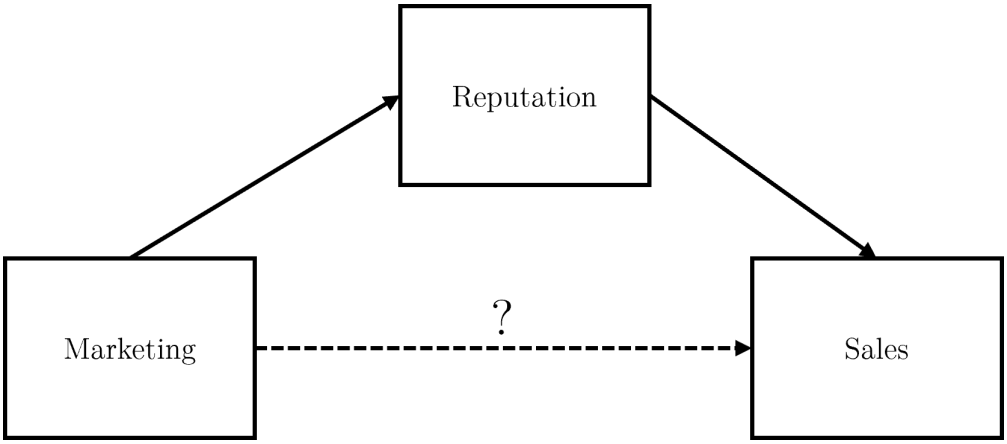
Prädiktor	Schätzung	Std.-fehler	t	p
Interzept [a]	1276.043	407.398	3.132	0.004
Marketing	0.737	0.784	0.940	0.356
Quality	220.719	191.177	1.155	0.259
Reputation_coded	438.030	379.743	1.153	0.259
Location:				
Asia – USA	−93.019	334.869	−0.278	0.783
Europe – USA	−56.067	310.943	−0.180	0.858
Interaction_qual_rep	142.712	107.578	1.327	0.196

[a] Repräsentiert das Referenzniveau

Der p-Wert des Interaktionsterms ist größer als 5 %. Daher haben wir keinen Hinweis auf einen Moderationseffekt gefunden.

Mediationseffekt

Mediator = dritte Variable, durch die der Effekt entsteht



Möglichkeiten, um zu testen, ob ein Mediationseffekt vorliegt:

1. *Jamovi medmod Modul*
   Lade medmod Modul herunter, dann `medmod -> Mediation`

Mediation Estimates

Effect	Estimate	SE	Z	p
Indirect	1.81	0.609	2.98	0.003
Direct	2.07	0.768	2.70	0.007
Total	3.89	0.769	5.06	< .001

(Es wurde ein Bootstrapverfahren verwendet, da die Stichprobe zu klein war für ein Testverfahren über den üblichen Weg)

2. *Händisch*

A. Schätze Gesamteffekt durch die Regression ohne Mediator:
$Sales = a_0 + a \cdot Marketing + e_1$ (Koeffizient a = Gesamteffekt)

Modellkoeffizienten - Sales

Prädiktor	Schätzung	Std.-fehler	t	p
Interzept	1587.58	413.703	3.84	< .001
Marketing	3.89	0.889	4.37	< .001

B. Schätze direkten Effekt durch die Regression mit Mediator:
$Sales = b_0 + a' \cdot Marketing + b \cdot Reputation + e_2$ (Koeffizient a' = direkter Effekt)

Modellkoeffizienten - Sales

Prädiktor	Schätzung	Std.-fehler	t	p
Interzept	1086.00	303.311	3.58	0.001
Marketing	2.07	0.700	2.96	0.006
Reputation_coded	1237.19	215.694	5.74	< .001

C. Indirekter Effekt. Zunächst Abschätzung des Effekts der Variable auf den Mediator:
$Reputation = c_0 + c \cdot Marketing + e_3$. Der indirekte Effekt ist dann: $b \cdot c$.

Modellkoeffizienten - Reputation_coded

Prädiktor	Schätzung	Std.-fehler	t	p
Interzept	0.40542	0.242	1.68	0.104
Marketing	0.00147	5.20e−4	2.82	0.008

D. Bestimmung der Signifikanz des indirekten Effekts mit dem Sobel-Test, dessen Teststatistik standardnormalverteilt ist:

$$Z = \frac{b \times c}{\sqrt{c^2 \cdot se_b^2 + b^2 \cdot se_c^2}},$$

wobei se die Standardfehler der Koeffizientenschätzer sind. Mit den Zahlen von oben:

$$Z = \frac{1237,19 \times 0,00147}{\sqrt{0,00147^2 \cdot 215,694^2 + 1237,19^2 \cdot 0.00052^2}} = \frac{1.8186693}{0.7172} = 2,54$$

Nun können wir mit Hilfe des R-Editors in Jamovi den p-Wert berechnen. Der R-Befehl lautet:

$$p - Wert = pnorm(2.54, lower.tail = FALSE) = 0.005611$$

Interpretation der Resultate:

Wenn der indirekte Effekt signifikant ist, liegt eine der folgenden Mediationen vor:

- Volle Mediation: Direkter Effekt ist nicht mehr signifikant.
- Partielle Mediation: Direkter Effekt bleibt signifikant, ist aber kleiner.

Omitted Variable Bias
Die Ergebnisse einer Regression sind verzerrt, wenn eine relevante unabhängige Variable weggelassen wird. Eine Variable ist relevant, wenn die folgenden Bedingungen erfüllt sind:

1. Die Variable muss einen Einfluss auf die abhängige Variable haben.
2. Die Variable muss mit einer anderen unabhängigen Variable korreliert sein.

Unser Datensatz enthält eine weitere Variable, die relevant sein könnte: die Produktqualität, gemessen an der durchschnittlichen Lebensdauer der Produkte.

Was wäre der Fall, wenn Qualität relevant wäre? Aus Vereinfachungsgründen lassen wir Reputation und Location weg.

- Wahre kausale Beziehung: $Sales = \beta_0 + \beta_1 \cdot Marketing + \beta_2 \cdot Quality$
- Unser Modell: $Sales = \gamma_0 + \gamma_1 \cdot Marketing$

Wenn Qualität und Marketingausgaben korreliert sind, gilt:

$$Quality = \alpha_0 + \alpha_1 \cdot Marketing$$

Setzen wir diese Gleichung in die wahre kausale Beziehung ein, ergibt sich:

$$\gamma_1 = \beta_1 + \beta_2 \alpha_1$$

Daher haben wir anstelle des wahren Effekts von Marketing auf den Umsatz β_1 den Wert γ_1 geschätzt, der je nach Richtung der Korrelation zwischen Marketing und Qualität größer oder kleiner als β_1 sein kann.

3.2 Logistische Regression

3.2.1 Grundlegende Idee

Bei einer kategorialen abhängigen Variablen wird nicht der Variablenwert selbst, sondern die Wahrscheinlichkeit, dass eine bestimmte Kategorie angenommen wird, geschätzt. Ein linearer Ansatz kann hier zu Problemen führen: Vorhergesagte Wahrscheinlichkeiten könnten außerhalb des Einheitsintervalls liegen, die Fehlerterme sind nicht normalverteilt und ein konstanter marginaler Effekt würde angenommen werden.

Die logistische Regression schätzt die Wahrscheinlichkeit bei einer binären Variablen (nimmt die Kategorien 0 und 1 an) wie folgt:

$$P(Y = 1) = \frac{e^z}{e^z + 1},$$

wobei $z = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$.
Alternativ:

$$log\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = z = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k,$$

wobei der Term $\left(\frac{P(Y=1)}{1-P(Y=1)}\right)$ die *Odds* genannt wird.

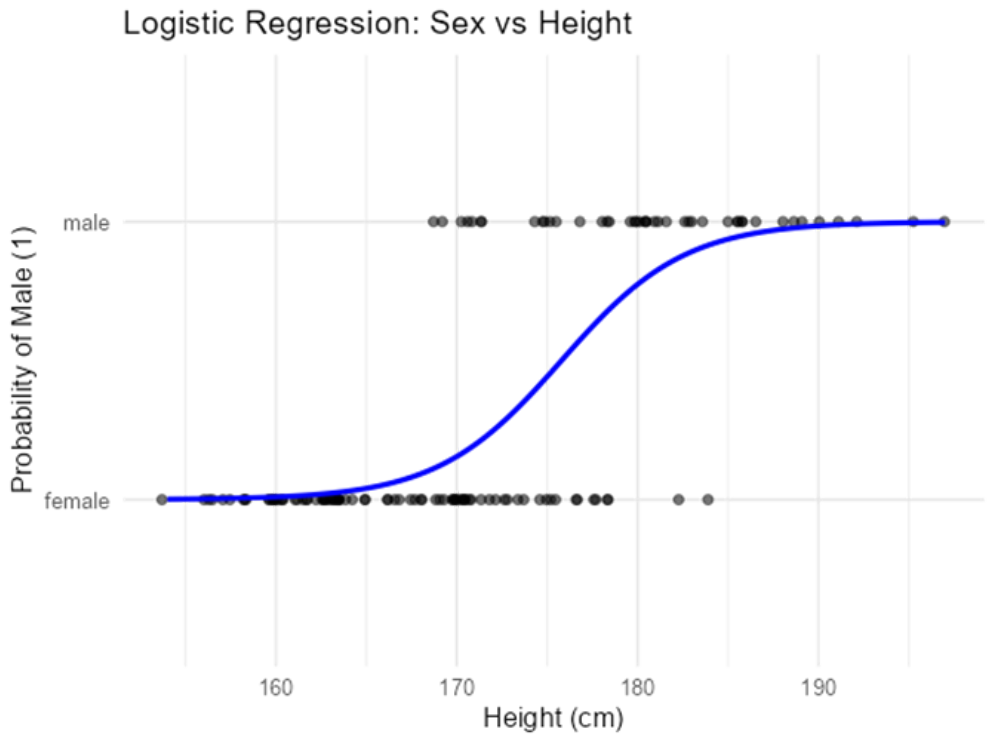
3.2.2 Beispiel

Schritt 1: Theoretische Regressionsgleichung

Können wir das Geschlecht vorhersagen basierend auf der Körpergröße?

$$P(Y = Mann) = \frac{e^z}{e^z + 1},$$

$z = \beta_0 + \beta_1 Größe$



Erzeugung der Regressionslinie der logistischen Regression entweder durch das *linear models* Modul oder anhand des folgenden R-Codes:

'Load necessary package'
`library(ggplot2)`

'Fit logistic regression model'
`model <- glm(sex ~ height, data = data, family = binomial)`

'Generate a sequence of height values for the logistic curve'
`height_seq <- seq(min(data$height), max(data$height), length.out = 300)`

'Predict probabilities using type = "response"'
`predicted_probs <- predict(model, newdata = data.frame(height = height_seq), type = "response")`
`predicted_probs <- predicted_probs+1`

'(Optional) Check the range of predicted probabilities'

`print(range(predicted_probs))  # Should be within 0 and 1`

'Create a dataframe for the logistic curve'
`curve_data <- data.frame(height = height_seq, probability = predicted_probs)`

```
'Plot scatterplot with jitter and logistic regression curve'
ggplot(data, aes(x = height, y = sex)) +
geom_jitter(width = 0.5, height = 0.00, alpha = 0.5, color = "black") +  # Binary outcome jittered
geom_line(data = curve_data, aes(x = height, y = probability+1), color = "blue", size = 1) +  # Logistic curve
labs(title = "Logistic Regression: Sex vs Height", x = "Height (cm)", y = "Probability of Male (1)") +
theme_minimal()
```

Schritt 2: Voraussetzungen überprüfen

Voraussetzungen Logistische Regression

- **LOG1. Binäre abhängige Variable.**
- **LOG2. Linearität.** Es besteht eine lineare Beziehung zwischen den log-odds der abhängigen Variablen und den unabhängigen Variablen.
- **LOG3. Fehlen perfekter (Multi)kollinearität.** Es besteht keine perfekte lineare Beziehung zwischen unabhängigen Variablen.
- **LOG4. Hinreichende Stichprobengröße.** Daumenregel: 10 bis 20 Beobachtungen pro unabhängiger Variable.
- **LOG5. Unabhängigkeit.** Beobachtungen sollten unabhängig voneinander sein (keine wiederholten Messungen an denselben Versuchspersonen).

Besprechung der Voraussetzungen → Vorlesung.

Schritt 3: Schreiben Sie die geschätzte Regressionsgleichung auf und interpretieren Sie das Ergebnis.

Model fit: Log-Likelihood-Ratio-Test ($H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$)

Loglikelihood ratio tests

	X^2	df	p
height	85.5	1	< .001

Geschätzte Regressionsgleichung:

$$P(Y = Mann) = \frac{e^z}{e^z + 1},$$

$z = -51,521 + 0,293Größe$

Parameter Estimates

Names	Estimate	SE	exp(B)	z	p
(Intercept)	−51.521	8.9370	4.22e−23	−5.76	< .001
height	0.293	0.0512	1.34	5.72	< .001

Exp(B) = 1,34 ist die Odds-Ratio und bedeutet, dass mit jeder 1 cm Zunahme der Körpergröße die Chancen männlich zu sein um 34 % steigen.

3.3 Ordinale Regression

3.3.1 Grundlegende Idee

Der grundlegende Gedanke der Ordinalregression basierend auf dem latenten Variablenansatz besteht darin, dass die beobachteten ordinalen Kategorien als diskrete Manifestationen einer kontinuierlichen latenten Variablen betrachtet werden. Dabei wird angenommen, dass es eine zugrunde liegende, nicht direkt beobachtbare Variable gibt, die die Rangordnung der Kategorien bestimmt.
Ein lineares Modell beschreibt diese latente Variable als Funktion von erklärenden Variablen und einem Fehlerterm. Die beobachteten ordinalen Kategorien ergeben sich dann durch Schwellenwerte (Cutoff-Werte), die die kontinuierliche latente Variable in diskrete Klassen unterteilen. Mathematisch kann dies wie folgt ausgedrückt werden:

$$y^* = X\beta + \varepsilon,$$

wobei y^* die latente Variable, X die unabhängigen Variablen und ε ein Fehlerterm sind. Die beobachtete ordinale Kategorie y ist dann definiert durch:

$$y = \begin{cases} 1, & \text{wenn } y^* \leq \tau_1 \\ 2, & \text{wenn } \tau_1 < y^* \leq \tau_2 \\ \vdots & \\ K, & \text{wenn } y^* > \tau_{K-1} \end{cases}$$

Da die latente Variable y^* nicht direkt beobachtbar ist, interessiert uns die Wahrscheinlichkeit, mit der eine beobachtete ordinale Kategorie y eintritt. Diese Wahrscheinlichkeit ergibt sich aus der Verteilung des Fehlerterms ε und den Schwellenwerten $\tau_1, \ldots, \tau_{K-1}$:

$$P(y = j \mid X) = P(\tau_{j-1} < y^* \leq \tau_j)$$

$$\Leftrightarrow$$

$$P(y = j \mid X) = P(\tau_{j-1} < X\beta + \varepsilon \leq \tau_j)$$

$$\Leftrightarrow$$

$$P(y = j \mid X) = P(\tau_{j-1} - X\beta < \varepsilon \leq \tau_j - X\beta)$$

Nun benötigen wir eine Annahme über die Verteilung des Fehlerterms ε, um diese Wahrscheinlichkeit zu berechnen:

1. **Probit-Modell:** Fehlerterm ist standardnormal verteilt. Die Wahrscheinlichkeiten werden dann durch die kumulative Verteilungsfunktion (CDF) der Normalverteilung $\Phi(\cdot)$ berechnet:

$$P(y = j \mid X) = \Phi(\tau_j - X\beta) - \Phi(\tau_{j-1} - X\beta)$$

2. **Logit-Modell:** Fehlerterm folgt einer logistischen Verteilung mit Varianz $\frac{\pi^2}{3}$. Dann:

$$P(y = j \mid X) = \Lambda(\tau_j - X\beta) - \Lambda(\tau_{j-1} - X\beta),$$

wobei $\Lambda(z) = \frac{e^z}{1+e^z}$.

Ob das Probit- oder Logit-Modell besser ist, hängt von der jeweiligen Anwendung und den Eigenschaften der Daten ab. Hier sind die wichtigsten Unterschiede und Entscheidungskriterien:

Kriterium	Probit	Logit
Fehlerverteilung	Normalverteilung $\mathcal{N}(0,1)$	Logistische Verteilung
Berechnung	Benötigt numerische Approximationen (z. B. für $\Phi(z)$)	Einfacher zu berechnen
Interpretierbarkeit	Schwieriger zu interpretieren	Odds Ratios sind leicht interpretierbar
Sensitivität für extreme Werte	Weniger sensitiv	Stärker sensitiv durch längere Tails
Verwendung in Sozialwissenschaften	Häufiger in der Ökonomie	Häufiger in der Psychologie und Medizin
Geschätzte Koeffizienten	Kleiner als im Logit-Modell	Größer als im Probit-Modell, aber oft ähnlich

3.3.2 Beispiel

In [ ]:

3.4 Multinomiale Regression

In [ ]: