2.5 Lineare Regressionsmodelle

2.5.1 Wiederholung aus Statistik I

Gegeben Datenpunkte (Y_i, X_i) schätze die beste Gerade $Y_i = \beta_0 + \beta_1 X_i,\ i = 1, \ldots, n.$

Bsp. 2.27. [Kaffeeverkauf auf drei Flohmärkten]

X Anzahl verkaufter Tassen Kaffee

Y zugehöriger Gewinn (Preis Verhandlungssache)

i	x_i	y_i	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	10	9	-1	0	0
2	15	21	11	5	25
3	5	0	-10	-5	25
	$\bar{x} = 10$	$\bar{y} = 10$			

Man bestimme die Regressionsgerade und interpretiere die erhaltenen KQ-Schätzungen!

Welcher Gewinn ist bei zwölf verkauften Tassen zu erwarten?

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ &= \frac{0\cdot(-1)+5\cdot 11+(-5)\cdot(-10)}{0+25+25} = \frac{105}{50} = 2.1\end{aligned}$$

Mit der Erhöhung der Menge X um eine Einheit erhöht sich der Gewinn Y um 2.1 Einheiten, also ist $\hat{b}$ so etwas wie der durchschnittliche Gewinn pro Tasse.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} = 10 - 2.1 \cdot 10 = -11$$

„Grundlevel“, Gewinn bei 0 Tassen (Fixkosten).

Vorhergesagte Werte und Residuen:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i, \qquad \hat{\varepsilon}_i = y_i - \hat{y}_i$$

$$\begin{aligned}
\hat{y}_1 &= -11 + 2.1 \cdot 10 = 10 & \Rightarrow \hat{\varepsilon}_1 &= -1 \\
\hat{y}_2 &= -11 + 2.1 \cdot 15 = 20.5 & \Rightarrow \hat{\varepsilon}_2 &= 0.5 \\
\hat{y}_3 &= -11 + 2.1 \cdot 5 = -0.5 & \Rightarrow \hat{\varepsilon}_3 &= 0.5
\end{aligned}$$

Zur Kontrolle: $\hat{\epsilon_1} + \hat{\epsilon_2} + \hat{\epsilon_3} = 0$

Prognose: $x^* = 12 \quad \Longrightarrow \hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 \cdot x^* = -11 + 2.1 \cdot 12 = 14.2$

Bsp. 2.28. [Arbeitszeit und Einkommen]

Multiples Regressionsmodell:

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

mit

$$\begin{aligned} X_1 &= \begin{cases} 1 & \text{männlich} \\ 0 & \text{weiblich} \end{cases} \\ X_2 &= \text{(vertragliche) Arbeitszeit} \\ Y &= \text{Einkommen} \end{aligned}$$

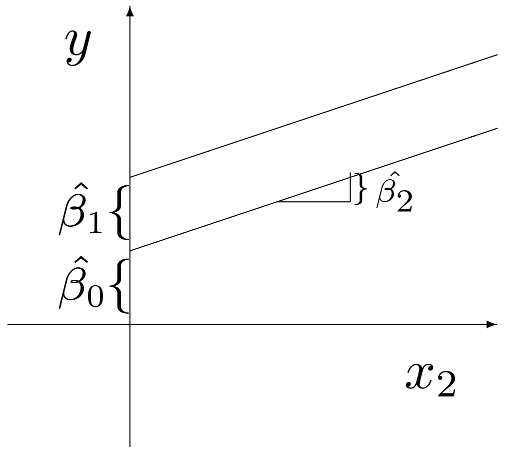
Interpretation:

Die geschätzte Gerade für die Männer lautet

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot x_{2i}$$

für die Frauen hingegen erhält man

$$\begin{aligned}\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot x_{2i} \\ &= \hat{\beta}_0 + \hat{\beta}_2 \cdot x_{2i}\end{aligned}$$



β_0 Grundlevel

β_2 durchschnittlicher Stundenlohn

β_1 Zusatzeffekt des Geschlechts zum Grundlevel.

Die 0-1 Variable dient als Schalter, mit dem man den Männereffekt an/abschaltet.

Bsp. 2.29. [Dummykodierung]

Nominales Merkmal mit q Kategorien, z.B. $X =$ Parteipräferenz mit

$$X = \begin{cases} 1 & \text{CDU/CSU oder FDP} \\ 2 & \text{SPD oder Grüne} \\ 3 & \text{Sonstige} \end{cases}$$

Man darf X nicht einfach mit Werten 1 bis 3 besetzen, da es sich um ein nominales Merkmal handelt.

Idee: Mache aus der einen Variable mit q (hier 3) Ausprägungen $q-1$ (hier 2) Variablen mit den Ausprägungen ja/nein ($\hat{=} 0/1$). Diese *Dummyvariablen* dürfen dann in der Regression verwendet werden.

$$X_1 = \begin{cases} 1 & \text{CDU/CSU oder FDP} \\ 0 & \text{andere} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{SPD, Grüne} \\ 0 & \text{andere} \end{cases}$$

Durch die Ausprägungen von X_1 und X_2 sind alle möglichen Ausprägungen von X vollständig beschrieben:

X	Text	X_1	X_2
1	CDU/CSU, FDP	1	0
2	SPD, Grüne	0	1
3	Sonstige	0	0

Beispiel zur Interpretation:

- Y: Score auf Autoritarismusskala
- X bzw. X_1, X_2: Parteienpräferenz
- X_3: Einkommen

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

- β_0: Grundniveau
- β_1: ceteris paribus Effekt (Erhöhung des Grundniveaus) von CDU/CSU und FDP
- β_2: ceteris paribus Effekt (Erhöhung des Grundniveaus) von SPD und Grünen
- β_3: ceteris paribus Effekt des Einkommens

Multiples Regressionsmodell:

$$
\begin{array}{ccl}
 & & X_{i1} \\
 & & X_{i2} \\
Y_i & \longleftarrow & \vdots \\
 & & X_{ip}
\end{array}
$$

abhängige Variable	unabhängige Variablen
metrisch/quasistetig	metrische/quasistetige oder dichotome (0/1) Variablen (kategoriale Variablen mit mehr Kategorien $\rightarrow$ Dummy-Kodierung)

Ansatz:

- linearer Zusammenhang.
- Ermittle aus den Daten „Wirkungsstärke“ der einzelnen Variablen.
- Im Folgenden: Probabilistische Modelle in Analogie zu den deskriptiven Modellen aus Statistik I (damit Verallgemeinerung auf die Grundgesamtheit möglich).

2.5.2 Lineare Einfachregression

Zunächst Modelle mit nur $einer$ unabhängigen Variable.

Statistische Sichtweise:

- Wahres Modell

$$y_i = \beta_0 + \beta_1 x_i$$

β_0 Grundniveau

β_1 „Elastizität“: Wirkung der Änderung von X_i um eine Einheit

- gestört durch zufällige Fehler ϵ_i Man beobachtet Datenpaare, (X_i, Y_i), $i = 1, \ldots, n$ mit

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

wobei

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

σ^2 für alle i gleich

$\epsilon_{i1}, \epsilon_{i2}$ stochastisch unabhängig für $i_1 \neq i_2$

Nach den Modellannahmen gilt für die bedingte Verteilung von Y_i gegeben $X_i = x_i$:

$$Y_i | X_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), \quad i = 1, \ldots, n.$$

Interpretation: verschiedene Normalverteilungen jeweils mit verschobenem Mittelwert $\mu_i = \beta_0 + \beta_1 x_i$, aber gleicher Varianz.

Aufgabe: Schätze die Parameter β_0, β_1 und σ^2. Die Schätzwerte und Schätzfunktionen werden üblicherweise mit $\hat{\beta}_0, \hat{\beta}_1$ und $\hat{\sigma}^2$ bezeichnet.

In der eben beschriebenen Situation gilt:

1. Die Maximum Likelihood Schätzer lauten:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}, \\
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}, \\
\hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2
\end{aligned}$$

mit den geschätzten Residuen

$$\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i.$$

2. Mit

$$\hat{\sigma}_{\hat{\beta}_0} := \frac{\hat{\sigma}\sqrt{\sum_{i=1}^{n} X_i^2}}{\sqrt{n\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

gilt

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t^{(n-2)}$$

und analog mit

$$\hat{\sigma}_{\hat{\beta}_1} := \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

gilt

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t^{(n-2)}.$$

Bem. 2.30.

- $\hat{\beta}_0$ und $\hat{\beta}_1$ sind, wenn man die Realisationen (x_i, y_i) von (X_i, Y_i) einsetzt, die KQ-Schätzer aus Statistik I. Unter Normalverteilung fällt hier das ML- mit dem KQ-Prinzip zusammen.

- Man kann unmittelbar Tests und Konfidenzintervalle ermitteln (völlig analog zum Vorgehen in Kapitel 2.3 und 2.4).

  Konfidenzintervalle zum Sicherheitsgrad γ:

$$\text{für } \beta_0 : \qquad [\hat{\beta}_0 \pm \hat{\sigma}_{\hat{\beta}_0} \cdot t^{(n-2)}_{1+\frac{\gamma}{2}}]$$

$$\text{für } \beta_1 : \qquad [\hat{\beta}_1 \pm \hat{\sigma}_{\hat{\beta}_1} \cdot t^{(n-2)}_{\frac{1+\gamma}{2}}]$$

  Mit der Teststatistik

$$T_{\beta_1^*} = \frac{\hat{\beta}_1 - \beta_1^*}{\hat{\sigma}_{\hat{\beta}_1}}$$

ergibt sich

	Hypothesen			kritische Region
I.	$H_0: \beta_1 \leq \beta_1^*$	gegen	$\beta_1 > \beta_1^*$	$T \geq t_{1-\alpha}^{(n-2)}$
II.	$H_0: \beta_1 \geq \beta_1^*$	gegen	$\beta_1 < \beta_1^*$	$T \leq t_{1-\alpha}^{(n-2)}$
III.	$H_0: \beta_1 = \beta_1^*$	gegen	$\beta_1 \neq \beta_1^*$	$\|T\| \geq t_{1-\frac{\alpha}{2}}^{(n-2)}$

(analog für $\hat{\beta}_0$).

Von besonderem Interesse ist der Fall $\beta_1^* = 0$:

- Typischer SPSS-Output

Koeffizienten[a]

			Standardisierte Koeffizienten		
	β	Standardfehler	Beta	T	Signifikanz
Konstante	$\hat{\beta}_0$	$\hat{\sigma}_{\hat{\beta}_0}$	5)	1)	3)
Unabhängige Variable	$\hat{\beta}_1$	$\hat{\sigma}_{\hat{\beta}_1}$	6)	2)	4)

[a] abhängige Variable

1) Wert der Teststatistik

$$T_{\beta_0^*} = \frac{\hat{\beta}_0}{\hat{\sigma}_{\hat{\beta}_0}}.$$

zum Testen von H_0: $\beta_0 = 0$ gegen H_1: $\beta_0 \neq 0$.

2) Analog: Wert von

$$T_{\beta_1^*} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$$

zum Testen von H_0: $\beta_1 = 0$ gegen H_1: $\beta_1 \neq 0$.

3) p-Wert zu 1)

4) p-Wert zu 2)

5), 6) hier nicht von Interesse.

- Die Testentscheidung „$\hat{\beta}_1$ signifikant von 0 verschieden“ entspricht dem statistischen Nachweis eines Einflusses von X.
- Man kann analog zu Kap. 2.4.7.1 auch einseitige Hypothesen testen

2.5.3 Multiple lineare Regression

- Analoger Modellierungsansatz, aber mit mehreren erklärenden Variablen:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i$$

- Schätzung von $\beta_0, \beta_1, \ldots, \beta_p$ und σ^2 sinnvollerweise über Matrixrechnung bzw. Software.

  Aus dem SPSS-Output sind $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ sowie $\hat{\sigma}_{\hat{\beta}_0}, \hat{\sigma}_{\hat{\beta}_1}, \ldots, \hat{\sigma}_{\hat{\beta}_p}$ ablesbar.

  (Outputs lesen können ist absolut klausurrelevant! Matrixrechnung wird nicht verlangt.)

- Es gilt für jedes $j = 0, \ldots, p$

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t^{(n-p-1)}$$

und man erhält wieder Konfidenzintervalle für β_j:

$$[\hat{\beta}_j \pm \hat{\sigma}_{\hat{\beta}_j} \cdot t_{1+\frac{\gamma}{2}}^{(n-p-1)}]$$

sowie entsprechende Tests.

Von besonderem Interesse ist wieder der Test

$$H_0 : \beta_j = 0,\ H_1 : \beta_j \neq 0.$$

Der zugehörige p-Wert findet sich im SPSS-Ausdruck (Vorsicht mit Problematik des multiplen Testens!).

- Man kann auch simultan testen, z.B.

$$\beta_1 = \beta_2 = \ldots = \beta_p = 0.$$

Dies führt zu einem sogenannten F-Test ($\longrightarrow$ Software).

2.5.4 Varianzanalyse (Analysis of Variance, ANOVA)

- Sind alle X_{ij} $0/1$-wertig, so erhält man die sogenannte *Varianzanalyse*, was dem Vergleich von mehreren Mittelwerten entspricht.
  - Für Befragte mit $X_{ij} = 0$ für alle j gilt:

$$\mathrm{E}(Y) = \beta_0$$

  - Ist $X_{i1} = 1$ und $X_{ij} = 0$ für $j \geq 2$, so gilt

$$\mathrm{E}(Y) = \beta_0 + \beta_1$$

  - Ist $X_{i1} = 1$ und $X_{i2} = 1$, sowie $X_{ij} = 0$ für $j \geq 3$, so gilt

$$\mathrm{E}(Y) = \beta_0 + \beta_1 + \beta_2$$

  - etc.

- Vor allem in der angewandten Literatur, etwa in der Psychologie, wird die Varianzanalyse unabhängig vom Regressionsmodell entwickelt. Diese Sichtweise soll auch hier jetzt kurz besprochen werden.
- Ziel: Mittelwertvergleiche in mehreren Gruppen, häufig in (quasi-) experimentellen Situationen.
- Verallgemeinerung des t-Tests. Dort nur zwei Gruppen.
- Hier nur *einfaktorielle Varianzanalyse* (*Eine* Gruppierungsvariable).

Bsp. 2.31.

Einstellung zu Atomkraft anhand eines Scores, nachdem ein Film gezeigt wurde.

3 Gruppen („Faktorstufen“):

- Pro-Atomkraft-Film
- Contra-Atomkraft-Film
- ausgewogener Film

Varianzanalyse: Vergleich der Variabilität in und zwischen den Gruppen

Beobachtungen: Y_{ij}

$j = 1, \ldots, J$ Faktorstufen

$i = 1, \ldots, n_j$ Personenindex in der j-ten Faktorstufe

Zwei äquivalente Modellformulierungen:

a) Modell in Mittelwertdarstellung:

$$Y_{ij} = \mu_j + \epsilon_{ij} \qquad j = 1, \ldots, J,\ i = 1, \ldots, n_j,$$

mit

μ_j	faktorspezifischer Mittelwert
ϵ_{ij}	zufällige Störgröße
ϵ_{ij}	$\sim \mathcal{N}(0, \sigma^2)$, $\epsilon_{11}, \epsilon_{12}, \ldots, \epsilon_{Jn_J}$ unabhängig.

Testproblem:

$$H_0 \quad : \quad \mu_1 = \mu_2 = \ldots \mu_J$$

gegen

$$H_1 \quad : \quad \mu_l \neq \mu_q \quad \text{für mindestens ein Paar } (l, q)$$

b) Modell in Effektdarstellung:

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

wobei α_j so, dass

$$\sum_{j=1}^{J} n_j \alpha_j = 0.$$

μ	globaler Erwartungswert
α_j	Effekt in der j-ten Faktorstufe, faktorspezifische systematische Abweichung vom gemeinsamen Mittelwert μ

Testproblem:

$$H_0 \quad : \quad \alpha_1 = \alpha_2 = \ldots \alpha_J = 0$$

gegen

$$H_1 \quad : \quad \alpha_j \neq 0 \text{ für mindestens ein } j$$

Die beiden Modelle sind äquivalent: setze $\mu_j := \mu + \alpha_j$.

Streuungszerlegung

Mittelwerte:

$\bar{Y}_{\bullet\bullet}$	Gesamtmittelwert in der Stichprobe
$\bar{Y}_{\bullet j}$	Mittelwert in der j-ten Faktorstufe

Es gilt (vgl. Statistik I) die Streuungszerlegung:

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{\bullet\bullet})^2 = \sum_{j=1}^{J}\underbrace{n_j(\bar{Y}_{\bullet j}-\bar{Y}_{\bullet\bullet})^2} + \underbrace{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{\bullet j})^2}$$

Die Testgröße

$$F = \frac{SQE/(J-1)}{SQR/(n-J)}$$

ist geeignet zum Testen der Hypothesen

$$H_0 \quad : \quad \mu_1 = \mu_2 = \ldots \mu_J$$

gegen

$$H_1 \quad : \quad \mu_l \neq \mu_q \quad \text{für mindestens ein Paar } (l,q)$$

beziehungsweise

$$H_0 \quad : \quad \alpha_1 = \alpha_2 = \ldots \alpha_J = 0$$

gegen

$$H_1 \quad : \quad \alpha_j \neq 0 \text{ für mindestens ein } j$$

Sie besitzt eine sog. F-*Verteilung* mit $(J-1)$ und $(n-J)$ Freiheitsgraden.

Die kritische Region besteht aus den *großen* Werten von F (Vorsicht: obwohl H_0 von „Gleichheitsform").

Also H_0 ablehnen falls

$$T > F_{1-\alpha}(J-1, n-J),$$

mit dem entsprechenden $(1-\alpha)$-Quantil der F-Verteilung mit $(J-1)$ und $(n-J)$ Freiheitsgraden.

(Je größer die Variabilität zwischen den Gruppen im Vergleich zu der Variabilität in den Gruppen, desto unplausibler ist die Nullhypothese, dass alle Gruppenmittelwerte gleich sind.)

Bei Ablehnung des globalen Tests ist dann oft von Interesse, welche Gruppen sich unterscheiden.

$\Rightarrow$ Testen spezifischer Hypothesen über die Effekte α_j bzw. die Mittelwerte μ_j. Dabei tritt allerdings wieder Problematik des multiplen Testens auf.

2.6 Fallstudie: Determinanten der Unterernährung in Sambia

Wichtiger Hinweis:
Wenn Sie diese Seiten „extern" besuchen, also ohne an der Veranstaltung teilzunehmen, so beachten Sie bitte:
Die hier durchgeführten Berechnungen dienten vorwiegend didaktischen Zwecken und erhebten keinen Anspruch auf substanzwissenschaftliche Relevanz aller „Ergebnisse"; So wurden sogar absichtlich häufig in der Praxis vorkommende Fehler mit eingebaut. Ein Teil des Outputs produziert also inhaltlich absolut sinnlose Aussagen.

(Kandala, Lang, Klasen, Fahrmeir (2001, Research in Official Statistics))

In Abstimmung mit der WHO werden in Entwicklungsländern regelmäßig DHS (Demographic and Health Surveys)-Erhebungen mittels repräsentativer Stichproben durchgeführt. Sie erhalten insbesondere Informationen zu Unterernährung, Sterblichkeit und Krankheitsrisiken für Kinder. Als Beispiel betrachten wir DHS-Daten für Sambia für das Jahr 1992. Ziel dieser Analyse ist es, Determinanten der Unterernährung von neugeborenen Kindern zu bestimmen. Unter mehreren möglichen anthropometrischen Indikatoren wählen wir die Maßzahl „Z-Score“ für chronische Unterernährung („Stunting“). Als erklärende Variablen, die den Z-Score beeinflussen, kommen in der DHS-Erhebung enthaltene Merkmale zum sozio-ökonomischen Status der Mutter bzw. des Haushalts sowie zur hygienischen und gesundheitlichen Situation in Frage. Zusätzlich ist als geographische Information die Region bzw. der Distrikt, in dem der Wohnort der Mutter liegt, bekannt.

Variable	**Beschreibung**
sex	Geschlecht des Kindes 1 = männlich 0 = weiblich
reg	Wohnort der Mutter 1 = Central 2 = Copperbelt 3 = Eastern 4 = Luapula 5 = Lusaka 6 = Northern 7 = North-Western 8 = Southern 9 = Western
dist	Wohnort der Mutter (genauere georaphische Unterteilung)

Variable	**Beschreibung**
zscore	Z-Score des Kindes
bmi	Body Mass Index der Mutter des Kindes
alter	Alter des Kindes in Monaten
erw	Erwerbsstatus der Mutter 1 = Mutter arbeitet 0 = Mutter arebitet nicht
edu	Ausbildungsstatus der Mutter 0 = keine Ausbildung 1 = incomplete primary education 2 = complete primary education 3 = incomplete secundary education 4 = complete secundary education 5 = higher education
sta	Stadt/Land 1 = Mutter lebt in der Stadt 0 = Mutter lebt auf dem Land

Maß für Unterernährung: Z-Score definiert als

$$z_i = \frac{AI_i - MAI}{\sigma} \cdot 100$$

AI_i Größe eines Kindes in einem bestimmten Alter

MAI Median der Größe für eine Referenzpopulation

σ Standardabweichung der Referenzpopulation

Lösen Sie folgende Fragen anhand beiliegender SPSS-Outputs. Geben Sie dabei auch jeweils das verwendete statistische Modell (z.B. zugrundegelegte Annahmen, genaue Formulierung der Hypothesen und des Tests, Modellgleichungen bei Regressionsanalysen) an und interpretieren Sie Ihre Ergebnisse sorgfältig!

Frage 1:

i) Geben Sie einen Punktschätzer für das Geschlechterverhältnis der Kinder in der Grundgesamtheit an. Durch welche theoretischen Eigenschaften wird Ihre Wahl gestützt?

ii) Testen Sie die Hypothese

1) „Weder Jungen noch Mädchen sind überrepräsentiert“

zum Niveau 5%.

iii) Wie lautet ein Konfidenzintervall zum Niveau 5% für den Anteil der Jungen? Wie kann man aus den Angaben ein Konfidenzintervall zum Niveau 90% bestimmen?

iv) Wie hätte man i) mit den Ergebnissen von ii) lösen können? (Begründung!)

Frage 2:

Testen Sie die Hypothesen

2) „Die Ernährungslage ist schlechter als in der Referenzpopulation.“

3) „Es gibt höchstens 5% Frauen mit hoher formaler Bildung.“

4) „Der Anteil chronisch Unterernährter (Z-Score ≤ -200) ist höchstens 25%.“

jeweils zum Niveau 5%. (Dabei wurde hohe formale Bildung festgelegt als „weiterführende Schulbildung abgeschlossen“.)

Frage 3:

Testen Sie die Hypothese

5) „Der Ernährungszustand von Kindern, deren Mütter hohe formale Bildung besitzen, ist besser, als Ernährungszustand von Kindern, deren Mütter keine hohe formale Bildung besitzen.“

zum Niveau 5%.

Frage 4:

Besitzen folgende Größen einen signifikanten Einfluss auf den Ernährungsscore?

- Alter des Kindes
- Body-Mass-Index der Mutter
- hohe formale Bildung
- die geographische Herkunft

i) Betrachten Sie zunächst die Variablen Alter und BMI jeweils isoliert. Wie kann man mit den Ergebnissen ein Konfidenzintervall für den Effekt des Alters ermitteln?
Zusatzfrage: Wie testet man die Nullhypothese „Der Effekt des Alters ist gleich -0.1“ zum Niveau 5%?

ii) Weitere Ergänzung: Man könnte argumentieren, dass man die in i) betrachteten

Variablen vorher transformieren sollte. Welche Transformationen bieten sich warum an?

iii) Betrachten Sie nun das Modell mit der Variable ‚Alter'(in ihrer ursprünglichen Form) und der Variable ‚hohe formale Bildung'. Interpretieren Sie den Output.

iv) Nehmen Sie nun die geographische Herkunft als weitere Variable hinzu und wiederholen Sie obige Analyse.

Frage 5:

Untersuchen Sie den Einfluss der geographischen Herkunft auch mittels einer Varianzanalyse.

Regression

========================= Frage 4 ===================================

----------------------------Teil i ---------------------------

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten B	Standardfehler	Standardisierte Koeffizienten Beta	T	Signifikanz
1	(Konstante)	-115,807	3,553		-32,597	,000
	alter	-2,029	,112	-,252	-18,098	,000

a. Abhängige Variable: zscore

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten B	Standardfehler	Standardisierte Koeffizienten Beta	T	Signifikanz
1	(Konstante)	-272,230	13,295		-20,476	,000
	bmi	4,662	,599	,111	7,781	,000

a. Abhängige Variable: zscore

------------------- Teil i) zur Kontrolle -----------------------

Koeffizienten[a]

Modell		95%-Konfidenzintervall für B Untergrenze	Obergrenze
1	(Konstante)	-122,772	-108,842
	alter	-2,249	-1,809

a. Abhängige Variable: zscore

Koeffizienten[a]

Modell		95%-Konfidenzintervall für B Untergrenze	Obergrenze
1	(Konstante)	-298,295	-246,165
	bmi	3,487	5,837

a. Abhängige Variable: zscore

--------------------- Teil ii) -------------------------------------

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten B	Standardfehler	Standardisierte Koeffizienten Beta	T	Signifikanz
1	(Konstante)	-17,149	6,694		-2,562	,010
	ln_alter	-50,431	2,121	-,323	-23,775	,000

a. Abhängige Variable: zscore

Seite 1

-------------------- Teil iii)----------------------

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten B	Standardfehler	Standardisierte Koeffizienten Beta	T	Signifikanz
1	(Konstante)	-118,544	3,557		-33,325	,000
	alter	-2,027	,112	-,251	-18,170	,000
	hohe_bildung	68,504	9,837	,096	6,964	,000

a. Abhängige Variable: zscore

--------------- Teil iv-----------------------

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten B	Standardfehler	Standardisierte Koeffizienten Beta	T	Signifikanz
1	(Konstante)	-109,653	4,981		-22,013	,000
	alter	-2,026	,112	-,251	-18,163	,000
	hohe_bildung	67,086	9,847	,094	6,813	,000
	reg	-1,900	,746	-,035	-2,548	,011

a. Abhängige Variable: zscore

Aufgenommene/Entfernte Variablen[b]

Modell	Aufgenommene Variablen	Entfernte Variablen	Methode
1	reg_9, alter, hohe_bildung, reg_7, reg_4, reg_3, reg_6, reg_5, reg_8, reg_2[a]	.	Eingeben

a. Alle gewünschten Variablen wurden aufgenommen.

b. Abhängige Variable: zscore

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten B	Standardfehler	Standardisierte Koeffizienten Beta	T	Signifikanz
1	(Konstante)	-107,151	6,984		-15,342	,000
	alter	-2,036	,109	-,252	-18,642	,000
	hohe_bildung	53,027	9,765	,075	5,430	,000
	reg_2	15,812	7,470	,047	2,117	,034
	reg_3	-39,377	8,859	-,082	-4,445	,000
	reg_4	-60,170	8,981	-,122	-6,700	,000
	reg_5	14,502	8,044	,037	1,803	,071
	reg_6	-51,784	8,806	-,109	-5,880	,000
	reg_7	-40,899	10,037	-,069	-4,075	,000
	reg_8	5,913	7,954	,015	,743	,457
	reg_9	-15,131	9,344	-,029	-1,619	,105

a. Abhängige Variable: zscore

```
---------------------- Frage v) -------------
```

Univariat

ONEWAY ANOVA

zscore

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	4114848,0	8	514356,003	28,230	,000
Innerhalb der Gruppen	88150212	4838	18220,383		
Gesamt	92265060	4846			

=================================Frage 1 ===============================

T-Test

Statistik bei einer Stichprobe

	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
sex	4847	,4943	,50002	,00718

Test bei einer Sichprobe

	Testwert = 0.5					
					95% Konfidenzintervall der Differenz	
	T	df	Sig. (2-seitig)	Mittlere Differenz	Untere	Obere
sex	-,790	4846	,430	-,00567	-,0198	,0084

===========================Frage 2====================================

T-Test

Statistik bei einer Stichprobe

	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
zscore	4847	-169,9251	137,98343	1,98194

Test bei einer Sichprobe

	Testwert = 0					
					95% Konfidenzintervall der Differenz	
	T	df	Sig. (2-seitig)	Mittlere Differenz	Untere	Obere
zscore	-85,737	4846	,000	-169,92511	-173,8106	-166,0396

Seite 1

T-Test

Statistik bei einer Stichprobe

	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
hohe_bildung	4847	,0392	,19409	,00279

Test bei einer Sichprobe

	Testwert = 0.05					
					95% Konfidenzintervall der Differenz	
	T	df	Sig. (2-seitig)	Mittlere Differenz	Untere	Obere
hohe_bildung	-3,874	4846	,000	-,01080	-,0163	-,0053

T-Test

Statistik bei einer Stichprobe

	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
chronisch	4847	,4068	,49130	,00706

Test bei einer Sichprobe

	Testwert = 0.25					
					95% Konfidenzintervall der Differenz	
	T	df	Sig. (2-seitig)	Mittlere Differenz	Untere	Obere
chronisch	22,227	4846	,000	,15685	,1430	,1707

Seite 2

============================== Frage 3 ===========================

T-Test

Statistik bei gepaarten Stichproben

		Mittelwert	N	Standardabweichung	Standardfehler des Mittelwertes
Paaren 1	zscore	-169,9251	4847	137,98343	1,98194
	hohe_bildung	,0392	4847	,19409	,00279

Korrelationen bei gepaarten Stichproben

		N	Korrelation	Signifikanz
Paaren 1	zscore & hohe_bildung	4847	,097	,000

Test bei gepaarten Stichproben

		Gepaarte Differenzen					T	df	Sig. (2-seitig)
		Mittelwert	Standardabweichung	Standardfehler des Mittelwertes	95% Konfidenzintervall der Differenz Untere	95% Konfidenzintervall der Differenz Obere			
Paaren 1	zscore - hohe_bildung	-169,96431	137,96474	1,98167	-173,84928	-166,07933	-85,768	4846	,000

Seite 3

T-Test

Gruppenstatistiken

	hohe_bildung	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
zscore	,00	4657	-172,6279	138,02756	2,02261
	1,00	190	-103,6789	119,36617	8,65973

Test bei unabhängigen Stichproben

		Levene-Test der Varianzgleichheit		T-Test für die Mittelwertgleichheit						
									95% Konfidenzintervall der Differenz	
		F	Signifikanz	T	df	Sig. (2-seitig)	Mittlere Differenz	Standardfehler der Differenz	Untere	Obere
zscore	Varianzen sind gleich	5,909	,015	-6,783	4845	,000	-68,94892	10,16544	-88,87780	-49,02005
	Varianzen sind nicht gleich			-7,753	210,158	,000	-68,94892	8,89280	-86,47945	-51,41840

Seite 4