Anwendungsorientierte Analyseverfahren

Lineare Regression mit metrischen Prädiktoren

Prof. Dr. Michael Scharkow

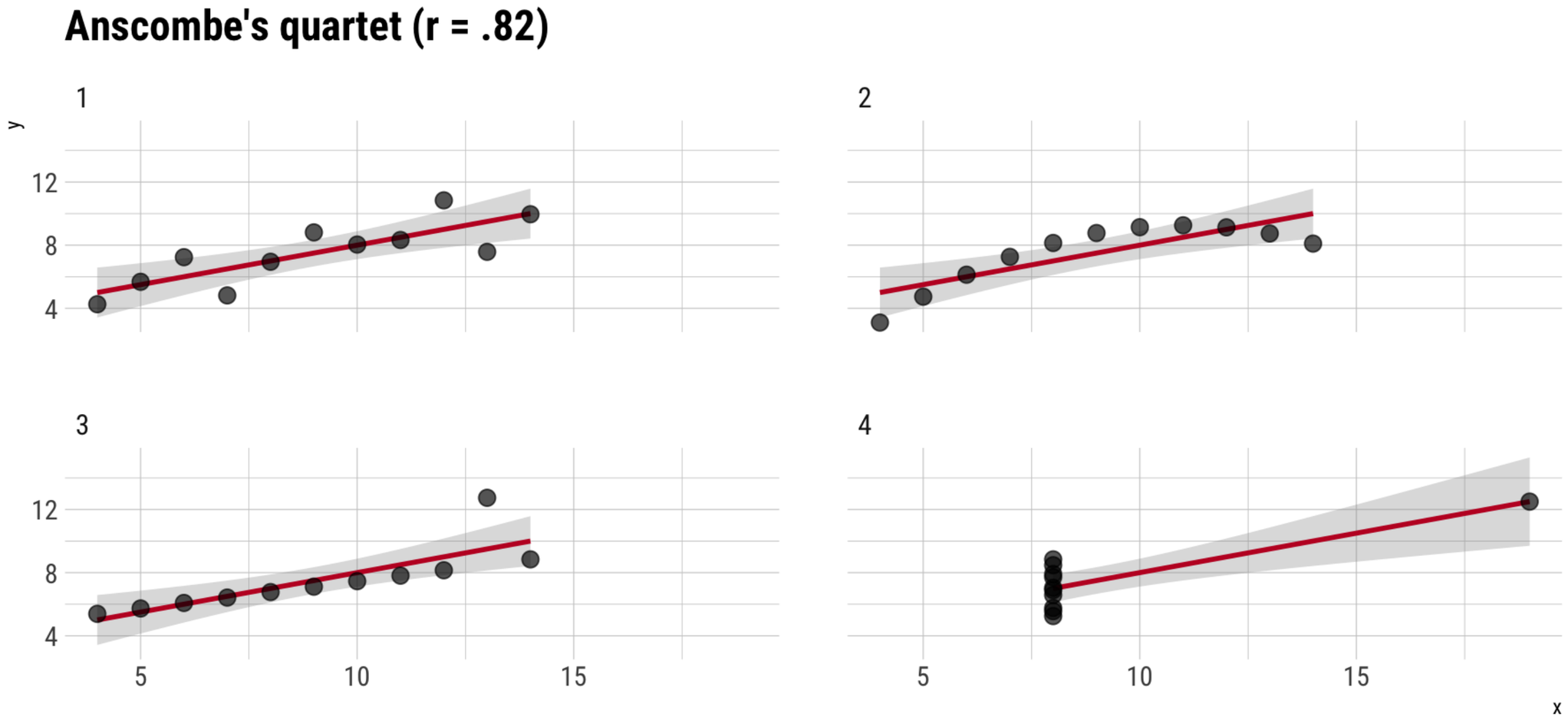
Sommersemester 2024

Fragen zur praktischen Übung?

KORRELATION, REGRESSION UND GLM

- historisch zwei unterschiedliche Ansätze, den Zusammenhang zweier metrischer Variablen zu analysieren: Korrelation und Regression
- Korrelation basiert auf der Idee der Kovarianz, d.h. dem “gemeinsamen Variieren” zweier Variablen
- bivariate Regression als GLM, bei dem ein metrisches Outcome Y durch eine Prädiktorvariable X vorhergesagt werden soll
- beide sind (natürlich!) miteinander verwandt, d.h. es unterscheiden sich vor allem die Konventionen des Berichtens

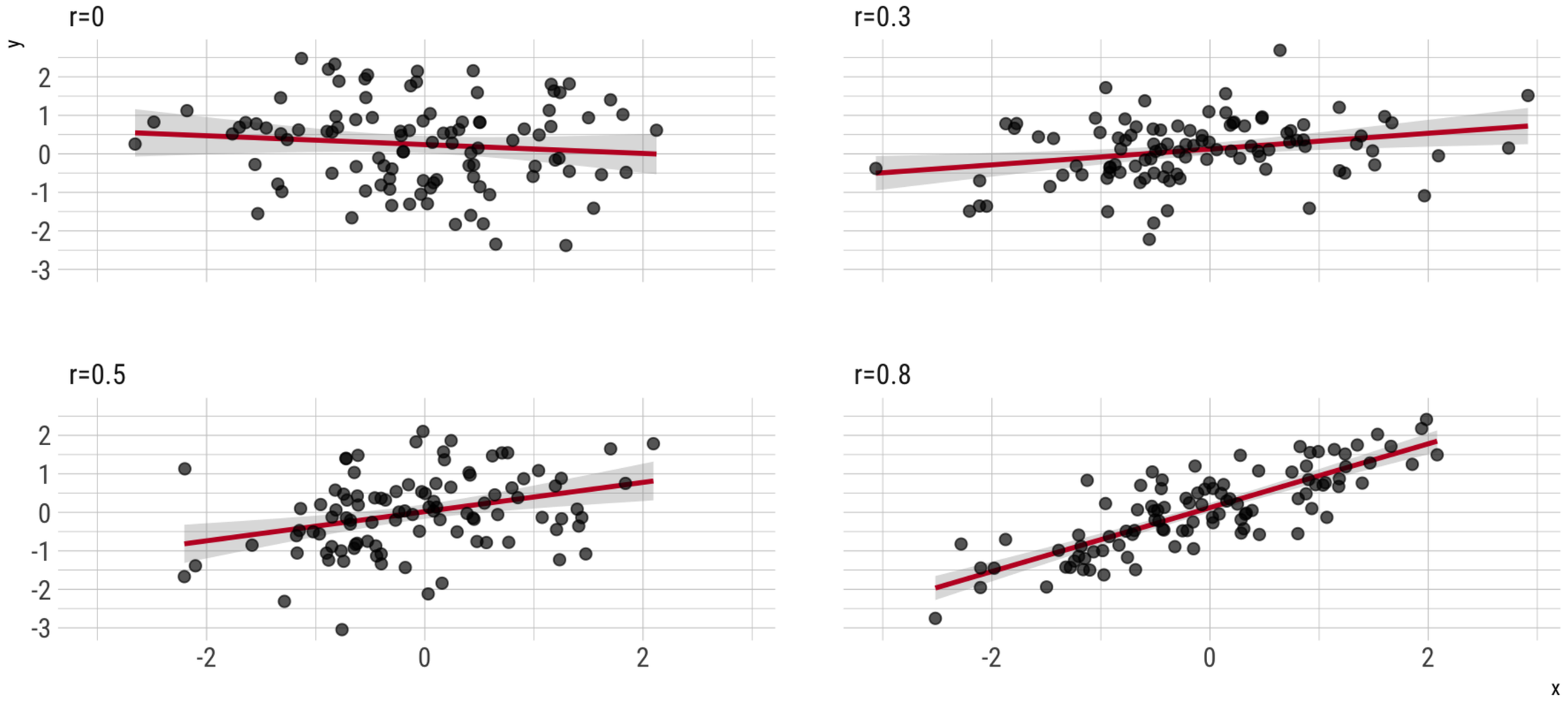
FORM DES ZUSAMMENHANGS



REGRESSIONS- VS. KORRELATIONSKOEFFIZIENTEN

- unstandardisierter Regressionskoeffizient B als Anstieg, d.h. zunächst Richtung des Zusammenhangs
- Korrelationskoeffizient r als Effektstärke, d.h. wie nahe die Werte von Y der Regressionsgeraden sind
- starker Zusammenhang = wenig Residualvarianz = hohe Korrelation r

UNTERSCHIEDLICH STARKE ZUSAMMENHÄNGE



KORRELATIONSMATRIZEN

Table 4
Means, standard deviations and zero-order correlations between genre preferences.

	M (SD)	1	2	3	4	5	6	7	8
1. Strategy (STR)	3.4 (1.4)	–							
2. Puzzle (PUZ)	2.9 (1.5)	.07*	–						
3. Sport (SPO)	2.5 (1.2)	.03	−.08*	–					
4. Adventure (ADV)	2.5 (1.3)	.16*	.09*	.12*	–				
5. Roleplaying (RPG)	2.5 (1.5)	.20*	−.11*	−.03	.39*	–			
6. Platform (PLA)	2.4 (1.3)	.05	.12*	.30*	.37*	.16*	–		
7. Simulation (SIM)	2.3 (1.3)	.25*	−.04*	.19*	.29*	.18*	.12*	–	
8. Music (MUS)	2.3 (1.4)	.03	.20*	.29*	.18*	.07*	.29*	.09*	–
9. Action (ACT)	2.1 (1.2)	.09*	−.28*	.21*	.31*	.33*	.26*	.20*	.05*

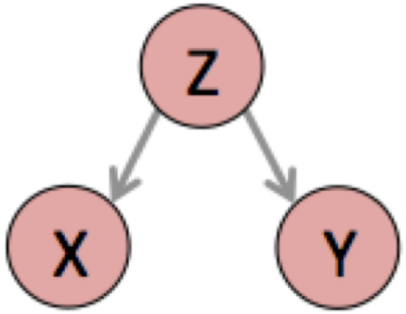
Note. Responses range from 1 ("don't like it at all") to 5 ("like it very much").
* $p \leqslant .05$, $n = 4500$.

Quelle: Scharkow, Festl, Vogelgesang & Quandt, 2013

MAL WIEDER: KORRELATION UND KAUSALITÄT

X causes Y

Y causes X



Z causes X and Y

random chance!

Quelle: https://www.cjr.org/tow_center_reports/the_curious_journalists_guide_to_data.php

BIVARIATE REGRESSION

- linearer Zusammenhang zwischen X und Y wieder, wobei *wir* definieren, was Prädiktor X und was Outcome Y ist
- die Modellformel wie immer:

$$Y_i = b_0 + b_1 X_i + \epsilon_i$$

- b_0 ist der vorhergesagte Wert von Y, wenn $x = 0$
- b_1 ist der vorhergesagte Anstieg von Y, wenn X um eine Einheit steigt (d.h. der Ansteig in Originalmetrik)
- ändert sich die Metrik von X oder Y, ändert sich die Interpretation von b_1

INTERCEPTS UND ZENTRIERUNG

- Intercept bzw. Konstante als vorausgesagte Wert von Y, wenn $X = 0$
- nur sinnvoll zu interpretieren, wenn X auch die Ausprägung 0 haben kann
- man kann X *zentrieren*, z.B. von allen Werten x_i eine Konstante c subtrahieren, dann ist Intercept der vorausgesagte Wert für $x = c$
- häufigste Zentrierung ist die **Mittelwertzentrierung**, d.h. $c = x$, der Intercept bezieht sich auf den durchschnittlichen Wert von X
- Zentrierung ändert nichts an den Regressionskoeffizienten oder am Globalfit, sondern nur am Intercept

EFFEKTGRÖSSEN UND MODELLGÜTE

- jeder Koeffizient B hat einen Standardfehler SE(B) und ein Konfindenzintervall
- mit beiden lässt sich H_0 prüfen können, dass *kein* linearer Zusammenhang existiert - bzw. ob die Daten sich mit $B = 0$ decken
- B lässt sich substantiell in der Metrik von Y interpretieren (eine Einheit mehr/weniger X entspricht B Einheiten mehr/weniger Y), er sagt aber nichts über die Stärke des Zusammenhangs oder Modellgüte
- wie gut das lineare Modell (im Vergleich zum Nullmodell ohne Prädiktoren) vorhersagt, kann man am R^2 erkennen
- im bivariaten Fall entspricht das exakt dem quadrierten Korrelationskoeffizienten r_{XY}

UNSTANDARDISIERTE B VS. STANDARDISIERTE BETA

- Interpretation von unstandardisiertem B setzt voraus, dass wir die Metriken von X und vor allem Y kennen
- oft wird (z.B. für Vergleiche oder Meta-Analysen) ein standardisiertes Maß gewünscht, das unabhängig von X und Y ist
- wir können Regressionskoeffizienten standardisieren, in dem wir
  - entweder die Daten X und Y *vor der Analyse* z-standardisieren (M = 0, SD = 1)
  - oder den Koeffizienten selbst standardisieren, durch $\beta = b\frac{s_x}{s_y}$
- im bivariaten Fall (nur dort!) entspricht β dem Korrelationskoeffizienten r

KORRELATION VS. BIVARIATE REGRESSION/GLM

- r als standardisierte Größe, d.h. auch ohne Kenntnis der Skalen von X und Y interpretierbar
- Korrelationsanalyse verführt ggf. weniger zu kausalen (Fehl-)Interpretationen als ein GLM mit unabhängiger und abhängiger Variable
- bei Korrelationen sind alternative Verfahren für nicht-metrische Daten (z.B. Spearmans Rangkorrelation) verbreitet
- beim GLM bekommt mehr Informationen: unstandardisierte Effektgrößen (inkl. Intercept)
- beide Verfahren liefern dieselben Schätzer und dieselben Testentscheidungen, basieren auf denselben Annahmen

BEISPIELSTUDIE: JOHANNES ET AL. (2022)

No effect of different types of media on well-being

Niklas Johannes[1]✉, Tobias Dienlin[2], Hasan Bakhshi[3] & Andrew K. Przybylski[1]

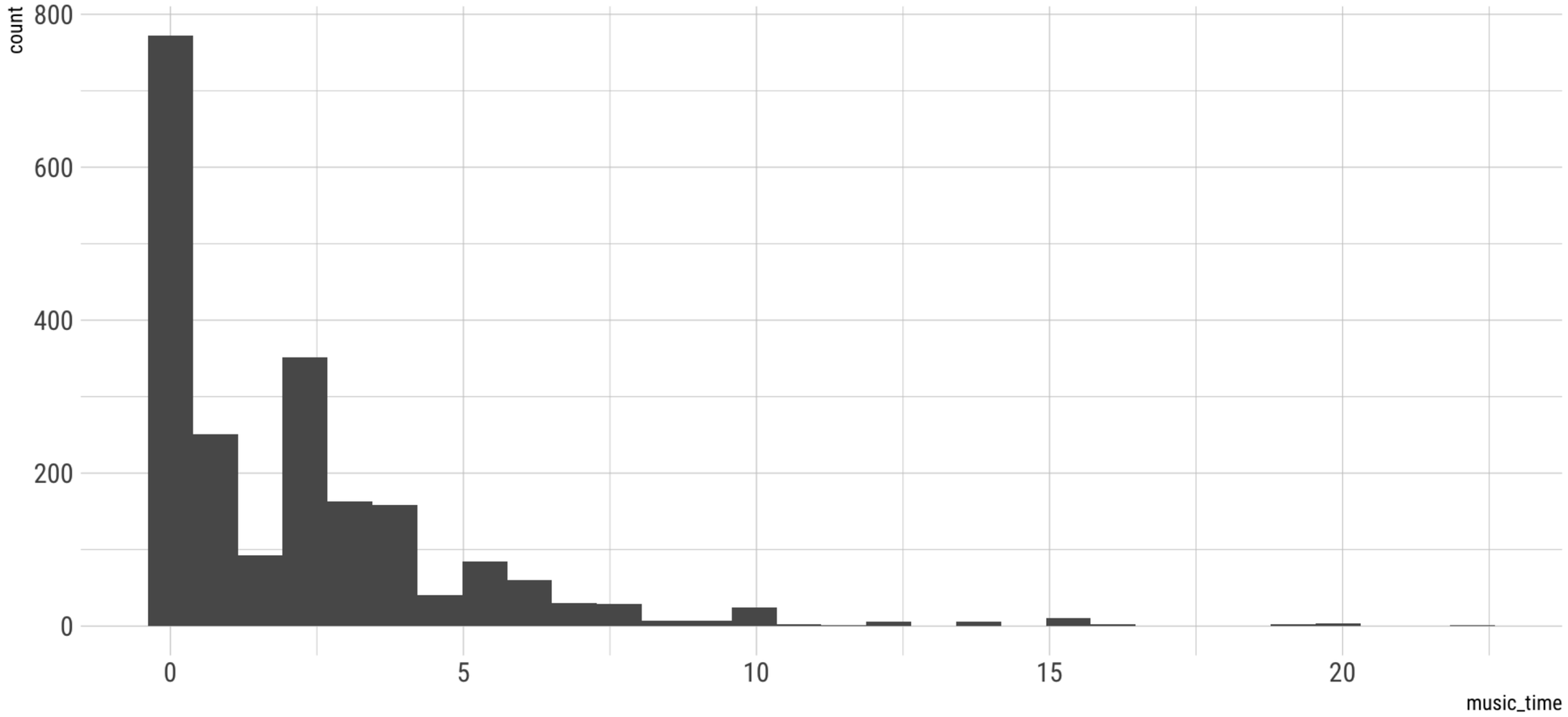
It is often assumed that traditional forms of media such as books enhance well-being, whereas new media do not. However, we lack evidence for such claims and media research is mainly focused on how much time people spend with a medium, but not whether someone used a medium or not. We explored the effect of media use during one week on well-being at the end of the week, differentiating time spent with a medium and use versus nonuse, over a wide range of different media types: music, TV, films, video games, (e-)books, (digital) magazines, and audiobooks. Results from a six-week longitudinal study representative of the UK population 16 years and older (N = 2159) showed that effects were generally small; between-person relations but rarely within-person effects; mostly for use versus nonuse and not time spent with a medium; and on affective well-being, not life satisfaction.

DATEN

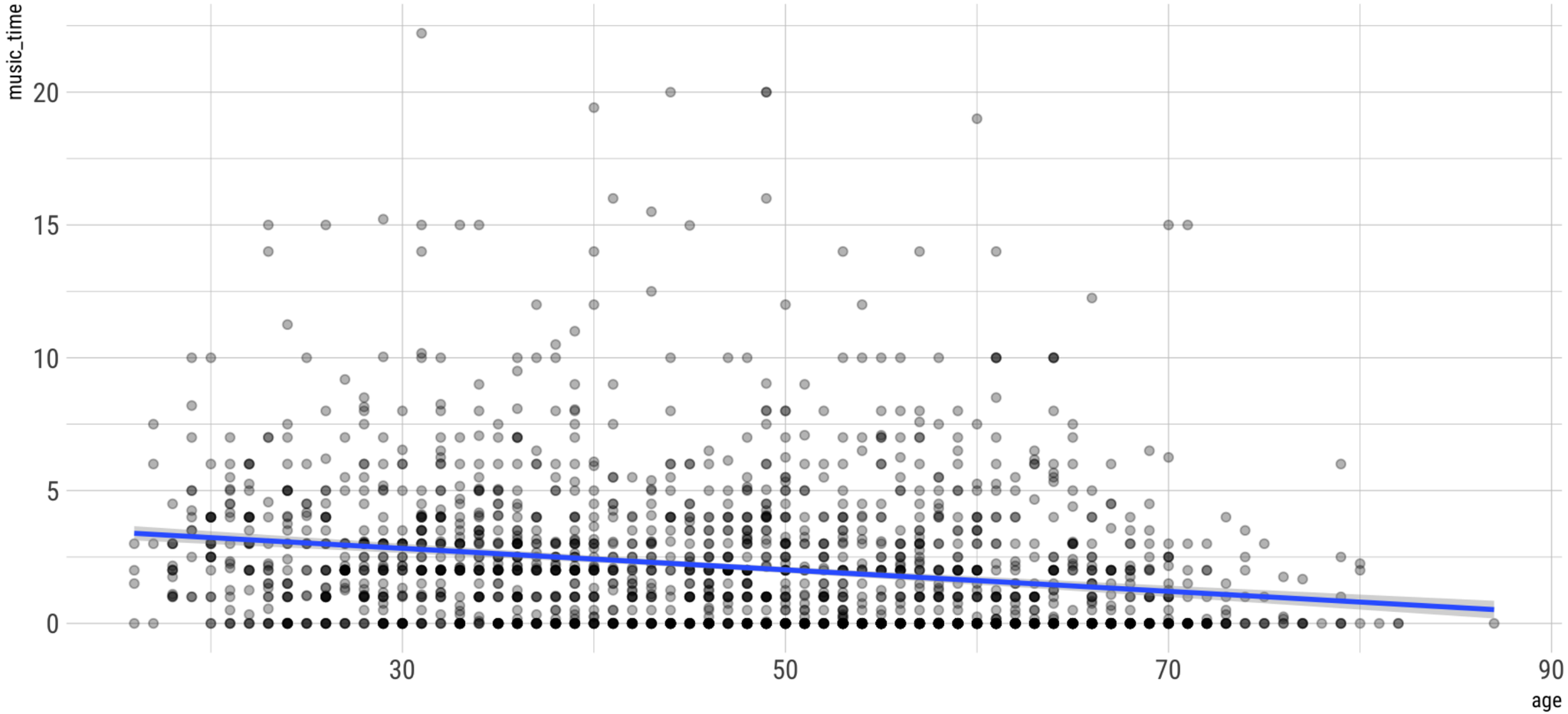
tv_time	age	games_time	music_time
0.0	22	0	4.00
0.0	43	0	2.50
2.0	38	0	0.17
5.0	30	0	2.00
1.5	29	1	0.75
2.0	57	0	0.00

Variable	Summary
Mean tv_time (SD)	2.73 (3.67)
Mean age (SD)	46.95 (14.67)
Mean games_time (SD)	0.93 (2.41)
Mean music_time (SD)	2.14 (2.75)

OUTCOME-VARIABLE



SCATTERPLOT



KORRELATION SAMT HYPOTHESENTEST

Parameter1	Parameter2	r	95% CI	p
age	music_time	-0.22	(-0.26, -0.17)	< .001

Alternative hypothesis: true correlation is not equal to 0

BIVARIATE LINEARE REGRESSION

Parameter	Coefficient	95% CI	t(2101)	p	Std. Coef.	Fit
(Intercept)	4.04	(3.65, 4.42)	20.53	< .001	0.00	
age	-0.04	(-0.05, -0.03)	-10.14	< .001	-0.22	
AICc						10125.19
R2						0.05
R2 (adj.)						0.05
Sigma						2.68

Reminder: t-Wert = $B/SE(B)$

ZENTRIERUNG ALTER = 18

Parameter	Coefficient	95% CI	t(2101)	p	Std. Coef.	Fit
(Intercept)	3.31	(3.06, 3.57)	25.49	< .001	0.00	
age18	-0.04	(-0.05, -0.03)	-10.14	< .001	-0.22	
AICc						10125.19
R2						0.05
R2 (adj.)						0.05
Sigma						2.68

MITTELWERTZENTRIERUNG

Parameter	Coefficient	95% CI	t(2101)	p	Std. Coef.	Fit
(Intercept)	2.14	(2.03, 2.25)	36.56	< .001	0.00	
age centered	-0.04	(-0.05, -0.03)	-10.14	< .001	-0.22	
AICc						10125.19
R2						0.05
R2 (adj.)						0.05
Sigma						2.68

TRANSFORMIERTE Y-VARIABLE (MINUTEN)

Parameter	Coefficient	95% CI	t(2101)	p	Std. Coef.	Fit
(Intercept)	128.40	(121.51, 135.28)	36.56	< .001	0.00	
age centered	-2.43	(-2.90, -1.96)	-10.14	< .001	-0.22	
AICc						27346.01
R2						0.05
R2 (adj.)						0.05
Sigma						161.06

Z-STANDARDISIERTE VARIABLEN

Parameter	Coefficient	95% CI	t(2101)	p	Std. Coef.	Fit
(Intercept)	0.00	(-0.04, 0.04)	0.08	0.936	0.00	
age zstd	-0.22	(-0.26, -0.17)	-10.14	< .001	-0.22	
AICc						5872.65
R2						0.05
R2 (adj.)						0.05
Sigma						0.98

AUFGABE: BOOKS TIME (MINUTEN)

Parameter	Coefficient	95% CI	t(2101)	p	Std. Coef.	Fit
(Intercept)	64.85	(59.45, 70.26)	23.52	< .001	0.00	
age centered	0.44	(0.07, 0.81)	2.35	0.019	0.05	
AICc						26327.51
R2						0.00
R2 (adj.)						0.00
Sigma						126.42

Fragen?

LITERATUR

Johannes, N., Dienlin, T., Bakhshi, H., & Przybylski, A. K. (2022). No effect of different types of media on well-being. *Scientific reports*, 12(1), 1-13.

Scharkow, M., Festl, R., Vogelgesang, J., & Quandt, T. (2015). Beyond the “core-gamer”: Genre preferences and gratifications in computer games. Computers in Human Behavior, 44, 293-298.