Lösungen zu den Aufgaben

1. **Aufgabe**

   Im Hinblick auf die lineare Regression: Welche der folgenden Aussage passt am besten?

   a. Die einfache Regression - $y = \alpha + \beta_1 x_1 + \epsilon$ - prüft, inwieweit zwei Variablen zusammenhängen (linear oder anderweitig).
   b. Obwohl statistische Zusammenhänge nicht ohne Weiteres Kausalschlüsse erlauben, kann man die Regression für Vorhersagen gut nutzen.
   c. Regressionskoeffizienten kann man so interpretieren: “Erhöht man X um eine 1 Einheit, so steigt daraufhin Y um β_1 Einheiten” (β_1 sei der entsprechende Regressionskoeffizient).
   d. “Lineare Regression” bedeutet, dass z.B. keine Polynome wie $y = \alpha + \beta_1 x_1^2 + \beta_2 x_1 + \epsilon$ berechnet werden dürfen, bzw. nicht zur *linearen* Regression zählen.
   e. Zentrieren der Prädiktoren ist bei der linearen Regression nicht zulässig.

   Lösung

   a. Falsch. Die lineare Regression $y = \alpha + \beta_1 x_1 + \epsilon$ untersucht, wie die Korrelation, den Grad des linearen Zusammenhangs. Allerdings sind auch nicht-lineare Zusammenhänge von y und den Prädiktoren erlaubt, etwa $y = \alpha + \beta_1 x_1^2 + \beta_2 x_2 + \epsilon$. *Linear* ist dabei so zu verstehen, dass y eine additive Funktion der Prädiktoren ist. Vielleicht wäre es daher besser, anstelle von “linearen” Modellen von “additiven” Modellen zu sprechen.
   b. Richtig. Für Vorhersagen ist Kenntnis einer Kausalstruktur nicht unbedingt nötig, kann aber sehr hilfreich sein.
   c. Falsch. Diese Interpretation suggeriert einen Kausaleffekt. Besser ist die Interpretation “Vergleicht man zwei Beobachtungen, die sich um 1 Einheit in X unterscheiden, so findet man im Durchschnitt einen Unterschied von β_1 in Y”.
   d. Falsch.Die Gleichung $y = \alpha + \beta_1 x_1^2 + \beta_2 x_2 + \epsilon$ ist linear in ihren Summanden.
   e. Falsch. Zentrieren der Prädiktoren ist bei der linearen Regression zulässig und oft sinnvoll.

2. **Aufgabe**

   Welche Aussage zur *multiplen* Regression ist korrekt?

   a. Es sind mehrere Prädiktoren erlaubt, genau dann wenn diese metrisch kontinuierlich sind.
   b. Es sind mehrere Prädiktoren erlaubt, genau dann wenn diese metrisch stetig sind.
   c. Es sind mehrere Prädiktoren erlaubt, genau dann wenn diese *nicht* metrisch kontinuierlich sind.
   d. Es sind mehrere Prädiktoren erlaubt, genau dann wenn diese *nicht* metrisch kontinuierlich sind.
   e. Keine der genannten.

   Lösung

   Keine der genannten. In der multiplen Regression sind jegliche Skalenniveaus bei den Prädiktoren möglich. “Hinter den Kulissen” werden aber nominale Prädiktoren in metrische umgewandelt.

   a. Falsch
   b. Falsch
   c. Falsch
   d. Falsch
   e. Wahr

3. **Aufgabe**

Wir betrachten einen Datensatz, der Kredite analysiert. Hier ist ein Auszug:

```
library(openintro)
data("loans_full_schema")
```

```
Rows: 1,292
Columns: 10
$ emp_title                 <chr> "security supervisor "…
$ emp_length                <dbl> 10, 10, 1, 9, 10, 10, …
$ state                     <fct> CA, MI, NV, AR, NJ, GA…
$ homeownership             <fct> RENT, MORTGAGE, MORTGA…
$ annual_income             <dbl> 35000, 35000, 42000, 5…
$ verified_income           <fct> Verified, Source Verif…
$ debt_to_income            <dbl> 57.96, 23.66, 32.00, 3…
$ annual_income_joint       <dbl> 57000, 155000, 95000, …
$ verification_income_joint <fct> Verified, Not Verified…
$ debt_to_income_joint      <dbl> 37.66, 13.12, 16.12, 2…
```

Quelle

Eine Analystin möchte den Zinssatz (`interest_rate`) auf Basis dieses Datensatzes vorhersagen.

Welche der oben gezeigten Variablen muss in der Regression *nicht* in Indikatorvariablen umgewandelt werden?

a. `emp_title`
b. `state`
c. `annual_income`
d. `verified_income`
e. `verification_income_joint`

Lösung

`annual_income` muss nicht in eine Indikatorvariable umgewandelt werden, da es eine numerische Variable ist.

a. Falsch
b. Falsch
c. Wahr
d. Falsch
e. Falsch

4. **Aufgabe**

Wir betrachten einen Datensatz, der Kredite analysiert:

```
library(openintro)
data("loans_full_schema")
```

Quelle

Eine Analystin möchte den Zinssatz (`interest_rate`) auf Basis dieses Datensatzes vorhersagen.

Welche der Variablen vom Typ `factor` im Datensatz hat genau zwei Stufen (d.h. verschiedene Werte)?

a. `state`
b. `homeownership`
c. `loan_purpose`
d. `application_type`

e. `loan_status`

Lösung

`application_type`

Hier ist eine Auflistung der Anzahl der Stufen aller Faktor-Variablen des Datensatzes:

```
## $state
## [1] 50
##
## $homeownership
## [1] 3
##
## $verified_income
## [1] 3
##
## $verification_income_joint
## [1] 4
##
## $loan_purpose
## [1] 12
##
## $application_type
## [1] 2
##
## $grade
## [1] 7
##
## $sub_grade
## [1] 32
##
## $issue_month
## [1] 3
##
## $loan_status
## [1] 6
##
## $initial_listing_status
## [1] 2
##
## $disbursement_method
## [1] 2
```

a. Falsch
b. Falsch
c. Falsch
d. Wahr
e. Falsch

5. **Aufgabe**

Wir betrachten einen Datensatz, der Kredite analysiert:

```
library(openintro)
data("loans_full_schema")
```

Quelle

Eine Analystin möchte den Zinssatz (`interest_rate`) auf Basis dieses Datensatzes vorhersagen.

Sie berechnet folgendes Regressionsmodell:

$$R = \beta_0 + \beta_1 I + \beta_2 T$$

Hier steht R für `interest_rate`, I für `annual_income` und T für `application_type`.

Wie lautet der R-Befehl, um diese Regression zu berechnen?

a. `lm(interest_rate ~ annual_income + application_type)`
b. `lm(interest_rate ~ annual_income + application_type, data = loans_full_schema)`
c. `lm(R ~ I + T, data = loans_full_schema)`
d. `lm(interest_rate ~ beta0 + beta1 * annual_income + beta2* application_type, data = loans_full_schema)`
e. `lm(interest_rate ~ beta0 + beta1 * I + beta2* T, data = loans_full_schema)`

Lösung

Der korrekte R-Befehl lautet:

```
lm(interest_rate ~ annual_income + application_type, data = loans_full_schema)
```

a. Falsch
b. Wahr
c. Falsch
d. Falsch
e. Falsch

6. **Aufgabe**

Wir betrachten einen Datensatz, der Kredite analysiert:

```
library(openintro)
data("loans_full_schema")
```

Quelle

Hier ist ein Überblick über den Datensatz:

```
## tibble [10,000 × 3] (S3: tbl_df/tbl/data.frame)
##  $ interest_rate   : num [1:10000] 14.07 12.61 17.09 6.72 14.07 ...
##  $ annual_income   : num [1:10000] 90000 40000 40000 30000 35000 34000 35000 110000 65000 30000 ...
##  $ application_type: Factor w/ 2 levels "individual","joint": 1 1 1 1 2 1 2 1 1 1 ...
```

Eine Analystin möchte den Zinssatz (`interest_rate`) auf Basis dieses Datensatzes vorhersagen.

Sie berechnet folgendes Regressionsmodell:

```
lm1 <- lm(interest_rate ~ annual_income + application_type, data = loans_full_schema)
```

Folgende Ergebnisse bekommt Sie zurück geliefert:

term	estimate	std_error
intercept	12.90	0.083
annual_income	0.00	0.000
application_type: joint	0.71	0.140

Welche Aussage ist korrekt?

a. `Estimate` liefert eine Schätzung zur Modellgüte.
b. Das Verhältnis von Signal zu Rauschen für `application_typejoint` ist kleiner als 1.
c. Es liegt ein Fehler vor, denn `application_typejoint` hat neben `joint` noch eine weitere Stufe (`individual`), diese ist aber nicht aufgeführt.
d. Der Wert bei `Intercept` gibt den Wert der abhängigen Variable an, bei Fällen mit dem Wert `individual` bei `application_type` und ohne Jahreseinkommen.
e. Der Wert bei `Intercept` gibt den Wert der abhängigen Variable an, bei Fällen mit dem Wert `joint` bei `application_type` und ohne Jahreseinkommen.

Lösung

Der Wert bei `Intercept` gibt den Wert der abhängigen Variable an, bei Fällen mit dem Wert `individual` bei `application_type` und ohne Jahreseinkommen.

```
predict(lm1, newdata = data.frame(annual_income = 0,
                                  application_type = "individual"))
```

```
##  1
## 13
```

a. Falsch
b. Falsch
c. Falsch
d. Wahr
e. Falsch

7. **Aufgabe**

Wir betrachten einen Datensatz, der Kredite analysiert:

```
library(openintro)
data("loans_full_schema")
```

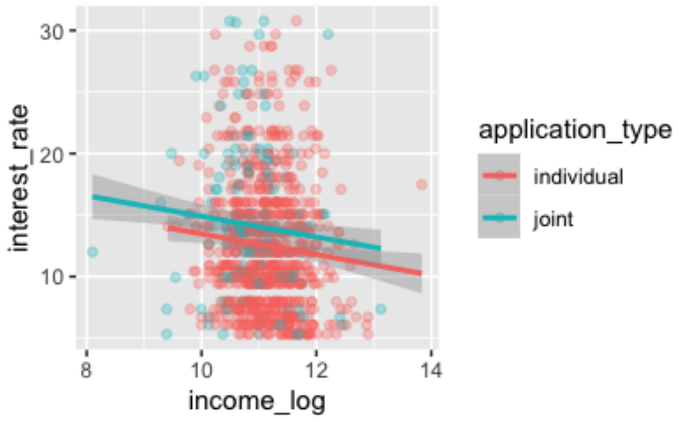
Quelle

Eine Analystin möchte den Zinssatz (`interest_rate`) auf Basis dieses Datensatzes vorhersagen.

Sie berechnet folgendes Regressionsmodell (auf einem Teil des Datensatzes):

```
lm1 <- lm(interest_rate ~ annual_income + application_type, data = loans_full_schema)
```

Grafisch aufbereitet, sieht ihr Ergebnis so aus:



Welche Aussage ist korrekt?

a. Für beide Gruppen von `application_type` (`individual` und `joint`) ist die Steigung der Regressionsgerade (annähernd) gleich.
b. Einkommen wurde logarithmiert; das ist keine sinnvolle Transformation im Allgemeinen.
c. In diesem Modell gibt es zwei Variablen: Zinssatz und logarithmiertes Einkommen.
d. In diesem Modell gibt es drei Variablen: Zinssatz, Einkommen und logarithmiertes Einkommen.
e. In diesem Modell gibt es vier Variablen: Zinssatz, Einkommen, `application_type individual` und `application_type joint`.

Lösung

Für beide Gruppen von `application_type` (`individual` und `joint`) ist die Steigung der Regressionsgerade (annähernd) gleich.

Das Diagramm wurde mit dieser Syntax erzeugt:

```
library(tidyverse)
library(moderndive)
library(mosaic)
data("loans_full_schema")

set.seed(42)
loans_full_schema %>%
    sample_n(1000) %>%
    filter(annual_income > 10) %>%
    mutate(income_log = log(annual_income)) %>%
    ggplot() +
    aes(x = income_log, y = interest_rate,
        color = application_type) +
    geom_point(alpha = .3) +
    geom_parallel_slopes()
```

a. Wahr
b. Falsch
c. Falsch
d. Falsch
e. Falsch

8. **Aufgabe**

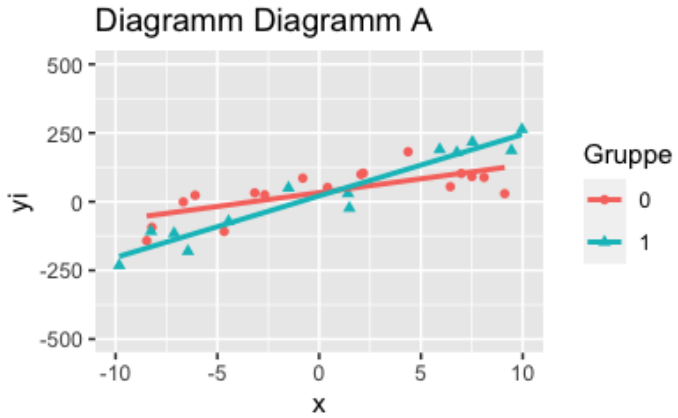


Diagramm Diagramm A
500
250
0
-250
-500
yi
-10
-5
0
5
10
x
Gruppe
0
1

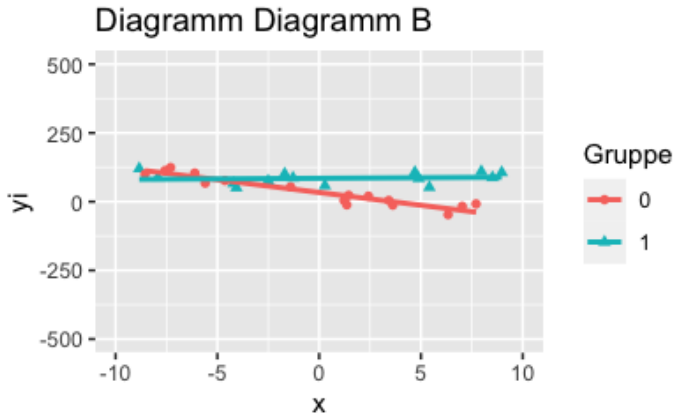


Diagramm Diagramm B
500
250
0
-250
-500
yi
-10
-5
0
5
10
x
Gruppe
0
1

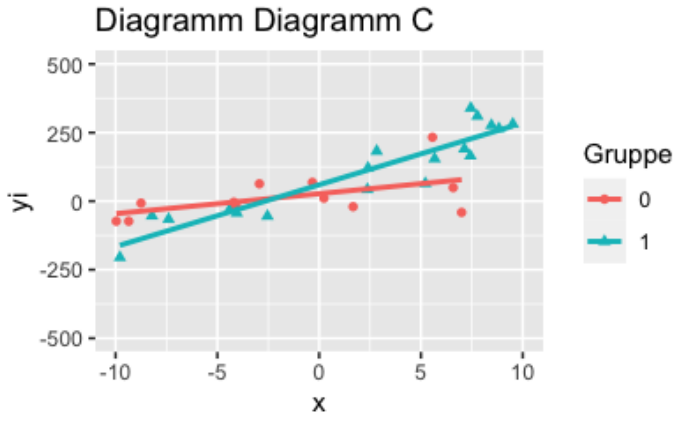
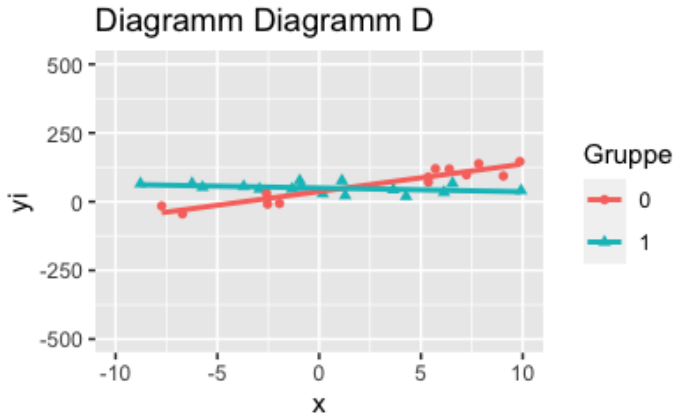
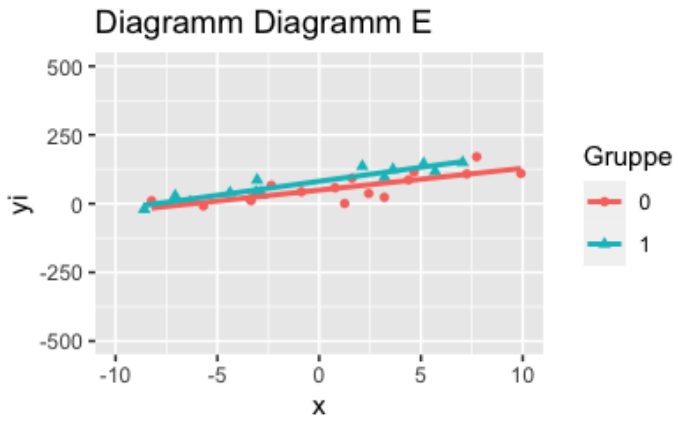


Diagramm Diagramm C
500
250
0
-250
-500
yi
-10
-5
0
5
10
x
Gruppe
0
1





Wählen Sie das Diagramm, in dem *kein* Interaktionseffekt (in der Population) vorhanden ist (bzw. wählen Sie Diagramm, dass dies am ehesten darstellt).

a. Diagramm A
b. Diagramm B
c. Diagramm C
d. Diagramm D
e. Diagramm E

Lösung

Das Streudiagramm `Diagramm E` zeigt *keinen* Interaktionseffekt.

a. Falsch
b. Falsch
c. Falsch
d. Falsch
e. Wahr

9. **Aufgabe**

Die Korrelation prüft, ob zwei Merkmale linear zusammenhängen.

Wie viele andere Verfahren kann die Korrelation als ein Spezialfall der Regression bzw. des linearen Modells $y = \beta_0 + \beta_1 + \ldots \beta_n + \epsilon$ betrachtet werden.

Als ein spezielles Beispiel betrachten wir die Frage, ob das Gewicht eines Diamanten (`carat`) mit dem Preis (`price`) zusammenhängt (Datensatz `diamonds`).

Den Datensatz können Sie so laden:

```
library(tidyverse)
data(diamonds)
```

a. Geben Sie das Skalenniveau beider Variablen an!

b. Betrachten Sie die Ausgabe von R:

```
lm1 <- lm(price ~ carat, data = diamonds)
summary(lm1)
```

```
##
## Call:
## lm(formula = price ~ carat, data = diamonds)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18585   -805    -19    537  12732
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2256.4       13.1    -173   <2e-16 ***
## carat         7756.4       14.1     551   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1550 on 53938 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.849
## F-statistic: 3.04e+05 on 1 and 53938 DF,  p-value: <2e-16
```

Wie (bzw. wo) ist aus dieser Ausgabe die Korrelation herauszulesen?

c. Macht es einen Unterschied, ob man Preis mit Karat bzw. Karat mit Preis korreliert?

d. In der klassischen Inferenzstatistik ist der p-Wert eine zentrale Größe; ist er klein ($p < .05$) so nennt man die zugehörige Statistik *signifikant* und verwirft die getestete Hypothese.

e. Im Folgenden sehen Sie einen Korrelationstest auf statistische Signifikanz, mit R durchgeführt. Zeigt der Test ein (statistisch) signifikantes Ergebnis? Wie groß ist der “Unsicherheitskorridor”, um den Korrelationswert (zugleich Punktschätzer für den Populationswert)?

```
library(rstatix)
diamonds %>%
  sample_n(30) %>%
  select(price, carat) %>%
  rstatix::cor_test() %>%
  gt()
```

var1	var2	cor	statistic	p	conf.low	conf.high	method
price	carat	0.84	8.3	5.6e-09	0.69	0.92	Pearson

Lösung

a. `carat` ist metrisch (verhältnisskaliert) und `price` ist metrisch (verhältnisskaliert)

b. R^2 kann bei einer einfachen (univariaten) Regression als das Quadrat von r berechnet werden. Daher $r = \sqrt{R^2}$.

```
sqrt(0.8493)
```

```
## [1] 0.92
```

Zum Vergleich

```
diamonds %>%
  summarise(r = cor(price, carat))
```

r
0.92

Man kann den Wert der Korrelation auch noch anderweitig berechnen (β umrechnen in ρ).

c. Nein. Die Korrelation ist eine symmetrische Relation.

d. Ja; die Zahl “3.81e-14” bezeichnet eine positive Zahl kleiner eins mit 13 Nullern vor der ersten Ziffer, die nicht Null ist (3.81 in diesem Fall). Der “Unsicherheitskorridor” reicht von etwa 0.87 bis 0.97.

10. **Aufgabe**

Laden Sie den Datensatz `mtcars` aus dieser Quelle.

Berechnen Sie eine Regression mit `mpg` als Ausgabevariable und `hp` aus Eingabevariable!

Welche Aussage ist für diese Analyse richtig?

a. `mpg` und `hp` sind positiv korreliert laut dem Modell.
b. Der Achsenabschnitt ist nahe Null.
c. Die Analyse beinhaltet einen nominal skalierten Prädiktor.
d. Das geschätzte Betagewicht für `hp` liegt bei `30.099`.
e. Das geschätzte Betagewicht für `hp` liegt bei `-0.068`.

Lösung

Das geschätzte Betagewicht für `hp` liegt bei `-0.068`.

Die Analyse könnte so aussehen:

```
library(tidyverse)
library(moderndive)
mtcars <- read_csv("https://vincentarelbundock.github.io/Rdatasets/csv/datasets/mtcars.csv")
```

```
## New names:
## Rows: 32 Columns: 12
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: "," chr
## (1): ...1 dbl (11): mpg, cyl, disp, hp, drat, wt, qsec, vs, am, gear, carb
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## • `` -> `...1`
```

```
lm1 <- lm (mpg ~ hp, data = mtcars)

get_regression_table(lm1)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	30.10	1.63	18.4	0	26.76	33.44
hp	-0.07	0.01	-6.7	0	-0.09	-0.05

a. Falsch
b. Falsch
c. Falsch
d. Falsch
e. Richtig

11. **Aufgabe**

Ist es möglich, kategorial skalierte Prädiktoren in eine Regressionsanalyse (lineare Modell) aufzunehmen?

a. Ja
b. Nein
c. Nur nominal skalierte, nicht ordinal skalierte
d. Nur ordinal skalierte, nicht nominal skalierte
e. Ja, aber nur eine

Lösung

Ja; diese werden aber in Dummy-Variablen umgerechnet (also in zweistufige Variablen mit den Stufen `0` und `1`), bevor die Regression berechnet wird.

a. Wahr
b. Falsch
c. Falsch
d. Falsch
e. Falsch

12. **Aufgabe**

Betrachten wir folgende Regressionsmodell:

$y = \alpha + \beta_1 x + \epsilon$

Geben Sie eine mathematische Formel an zur Zentrierung der Prädiktoren bzw. des Prädiktors!

Hinweise: - Geben Sie nur eine Formel ein, keinen Text, keine Leerzeichen und keine Sonderzeichen. - Verwenden Sie das Suffix "_c", um eine zentrierte Variable zu benennen. - Auf Funktionen wie den Mittelwert dürfen Sie zurückgreifen. Um den Mittelwert der Variablen `var` zu spezifizieren, kennzeichnen Sie dies mit `mw(var)`. - Verzichten Sie auf ein Malzeichen bei Multiplikationen. - Beispiel: "y_c = 2 mw(x) - 1".

Lösung

x_c=x-mw(x)

13. **Aufgabe**

Zwei Modelle, `m1` und `m2` produzieren jeweils die gleiche Vorhersage (den gleichen Punktschätzer).

`m1`:

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2124 -0.0581 -0.0011  0.0651  0.3414
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.000201   0.009498    0.02     0.98
```

```
## x           0.994614   0.009127  108.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.093 on 98 degrees of freedom
## Multiple R-squared:  0.992,  Adjusted R-squared:  0.992
## F-statistic: 1.19e+04 on 1 and 98 DF,  p-value: <2e-16
```

`m2`:

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.506 -0.687  0.002  0.716  2.632
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0283     0.1060    0.27     0.79
## x             0.8585     0.0990    8.67  9.1e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.1 on 98 degrees of freedom
## Multiple R-squared:  0.434,  Adjusted R-squared:  0.429
## F-statistic: 75.2 on 1 and 98 DF,  p-value: 9.07e-14
```

Die Modelle unterscheiden sich aber in ihrer Ungewissheit bezüglich β, wie in der Spalte `Std. Error` ausgedrückt.

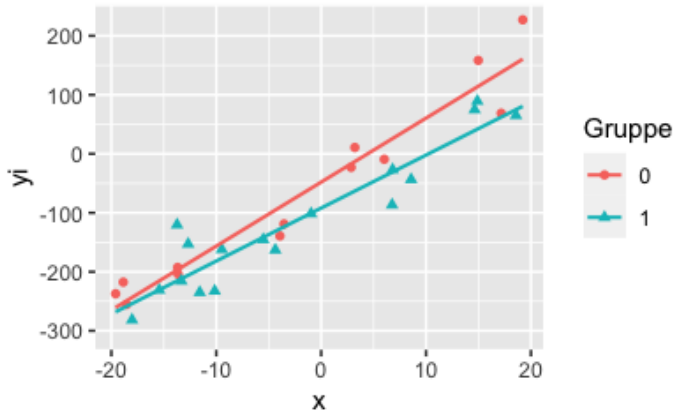
Welches der beiden Modelle ist zu bevorzugen? Begründen Sie.

Lösung

Modell `m1` hat eine *kleinere* Ungewissheit im Hinblick auf die Modellkoeffizienten β_0, β_1 und ist daher gegenüber `m2` zu bevorzugen.

14. **Aufgabe**

Ein Streudiagramm von x und y ergibt folgende Abbildung; dabei wird noch die Gruppierungsvariable g (mit den Stufen `0` und `1`) berücksichtigt (vgl. Farbe und Form der Punkte). Zur besseren Orientierung ist die Regressionsgerade pro Gruppe eingezeichnet.



Wählen Sie das (für die Population) am besten passende Modell aus der Liste aus!

Hinweis: Ein Interaktionseffekt der Variablen x und g ist mit xg gekennzeichnet.

a. $y = -40 + -10 \cdot x + 40 \cdot g + -10 \cdot xg + \epsilon$
b. $y = -40 + 10 \cdot x + 0 \cdot g + 10 \cdot xg + \epsilon$
c. $y = 40 + -10 \cdot x + 0 \cdot g + 10 \cdot xg + \epsilon$
d. $y = -40 + 10 \cdot x + -40 \cdot g + 0 \cdot xg + \epsilon$

Lösung

Das dargestellte Modell lautet $y = -40 + 10 \cdot x + -40 \cdot g + 0 \cdot xg + \epsilon$. Der Modellfehler ϵ hat den Anteil 0.3 im Vergleich zur Streuung von y.

a. Falsch
b. Falsch
c. Falsch
d. Richtig

15. **Aufgabe**

Berechnen Sie $\hat{y}$ für das unten ausgegeben Modell!

Nutzen Sie dafür folgende Werte:

- $g = 0$
- $x = 4$.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-34.5	4.76	-7.2	0	-44.0	-25
x	10.6	0.88	12.1	0	8.9	12
g	-35.4	6.60	-5.4	0	-48.6	-22
x:g	8.9	1.16	7.6	0	6.5	11

Hinweis: Ein Interaktionseffekt der Variablen x und g ist mit `x:g` gekennzeichnet. Runden Sie zur nächsten ganzen Zahl.

Lösung

$\hat{y}$ beträgt im Fall der vorliegenden Parameter und dem vorliegenden Modell 8.

16. **Aufgabe**

Gegeben sei ein Datensatz mit fünf Prädiktoren, wobei Studierende die Beobachtungseinheit darstellen:

- X_1: Muttersprachler (0: nein, 1: ja)
- X_2: Letzte Mathenote (z-Wert)
- X_3: Alter (z-Wert)
- X_4: Interaktion von $X1$ und $X2$

Die vorherzusagende Variable (Y; Kriterium) ist *Gehalt nach Studienabschluss*.

Wie lautet das Kriterium y für eine Person mit folgenden Werten:

- x_1: 1
- x_2: -1.51
- x_3: -0.8

Berechnen Sie dazu ein Regressionsmodell (Least Squares) anhand folgender Modellparameter:

- $\beta_0 : 30$
- $\beta_1 : 30$
- $\beta_2 : 1$
- $\beta_3 : 5$
- $\beta_4 : 1$

Geben Sie als Antwort den vorhergesagten Y-Wert an!

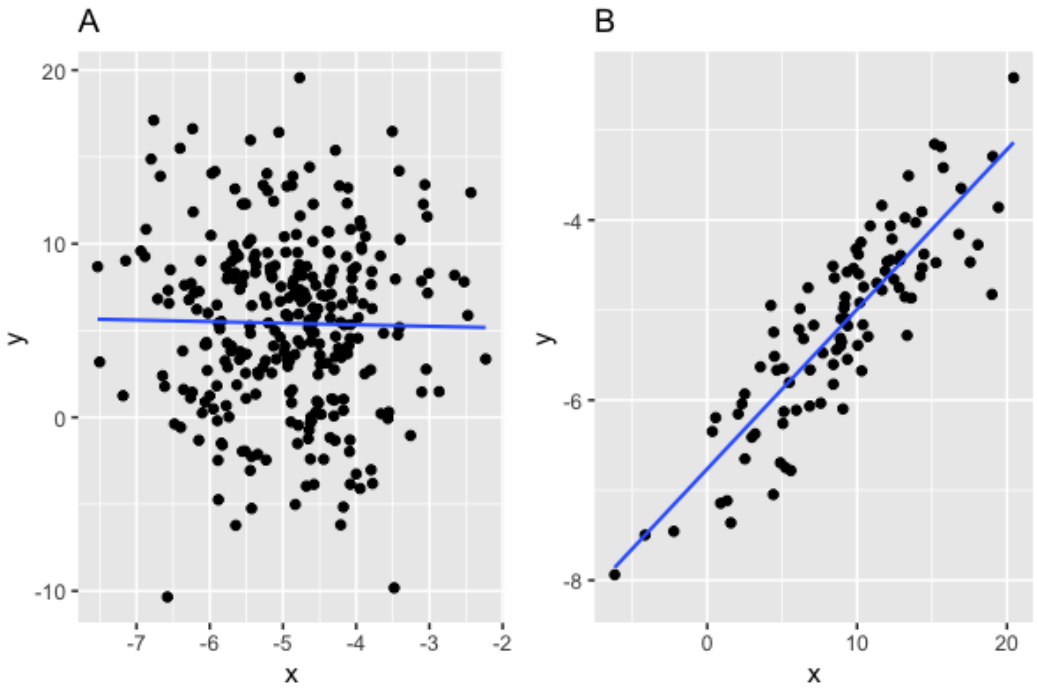
Hinweis: Runden Sie auf zwei Dezimalstellen.

Lösung

Die Antwort lautet 52.98.

17. **Aufgabe**

Die beiden folgenden Abbildungen zeigen zwei lineare Regressionen.



Welche Aussage stimmt?

a. $R_A^2 < R_B^2$
b. $R_A^2 \approx R_B^2$
c. $R_A^2 > R_B^2$

Lösung

Je enger die Punkte um die Gerade streuen, desto größer ist R^2.

a. Richtig
b. Falsch
c. Falsch