# Klausurreport Computergestützte Datenanalyse SoSe 2024

Die Klausur zum Modul C: Computergestützte Datenanalyse im Sommersemester 2024 fand am 26.07.2024 von 10.00 - 11.00 Uhr in den Computer-Pools des Universitätsrechenzentrums (URZ) mit 53 Teilnehmer:innen als elektronische Klausur in Präsenz statt. Die Klausur bestand aus 20 Multiple Choice Aufgaben mit jeweils vier Antwortmöglichkeiten und jeweils genau einer richtigen Antwort. Sie war in 2 Teile mit jeweils 10 Fragen untergliedert. Teil 1 war closed-book und Teil 2 open-book. Für den open-book Teil konnten Prüfungsteilnehmende das Internet und R am Prüfungs-PC nutzen. Die Klausur ist diesem Bericht beigefügt, richtige Antworten sind auf der letzten Seite angegeben.

## Bewertungsschema

Die Aufteilung der zugelassenen Noten auf die erreichten Prozentpunkte wurde anhand untenstehender Tabelle vorgenommen. Diese trifft folgende Zuordnung der erreichten Prozentpunkte zu den zugelassenen Noten anhand von geschlossenen Prozentpunktintervallen.

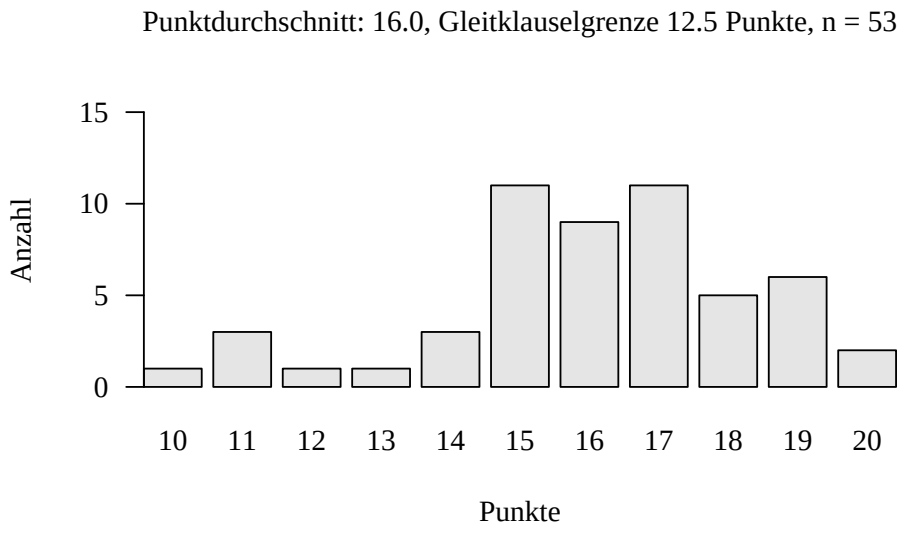
| $\leq$ | $\geq$ | Note |
|---|---|---|
| 100 | 95 | 1,0 |
| 94 | 90 | 1,3 |
| 89 | 85 | 1,7 |
| 84 | 80 | 2,0 |
| 79 | 75 | 2,3 |
| 74 | 70 | 2,7 |
| 69 | 65 | 3,0 |
| 64 | 60 | 3,3 |
| 59 | 55 | 3,7 |
| 54 | 50 | 4,0 |
| 49 | 0 | 5,0 |

Es ergibt sich folgendes Punktenotenschema, wobei $< 10$ Punkte mit 5.0 bewertet würden.

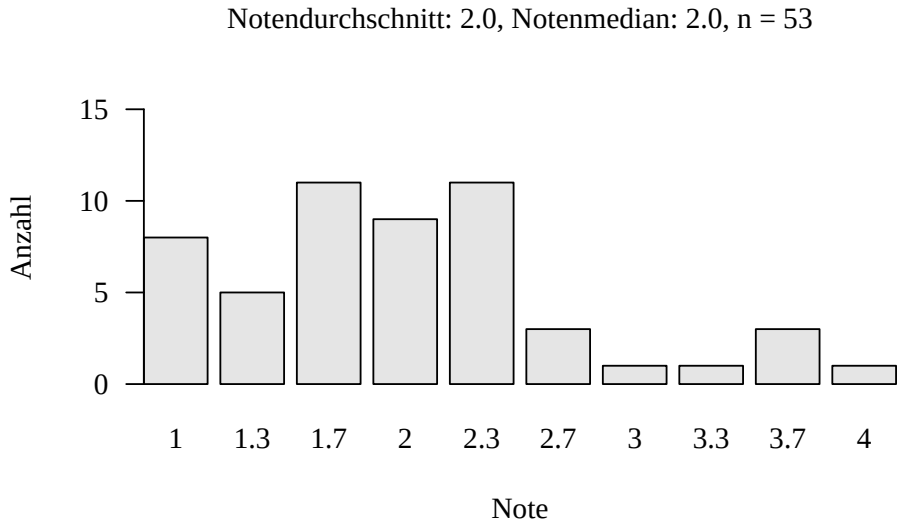
| Punkte | Prozent | Note |
|---|---|---|
| 20 | 100 | 1,0 |
| 19 | 95 | 1,0 |
| 18 | 90 | 1,3 |
| 17 | 85 | 1,7 |
| 16 | 80 | 2,0 |
| 15 | 75 | 2,3 |
| 14 | 70 | 2,7 |
| 13 | 65 | 3,0 |
| 12 | 60 | 3,3 |
| 11 | 55 | 3,7 |
| 10 | 50 | 4,0 |

## Ergebnisse

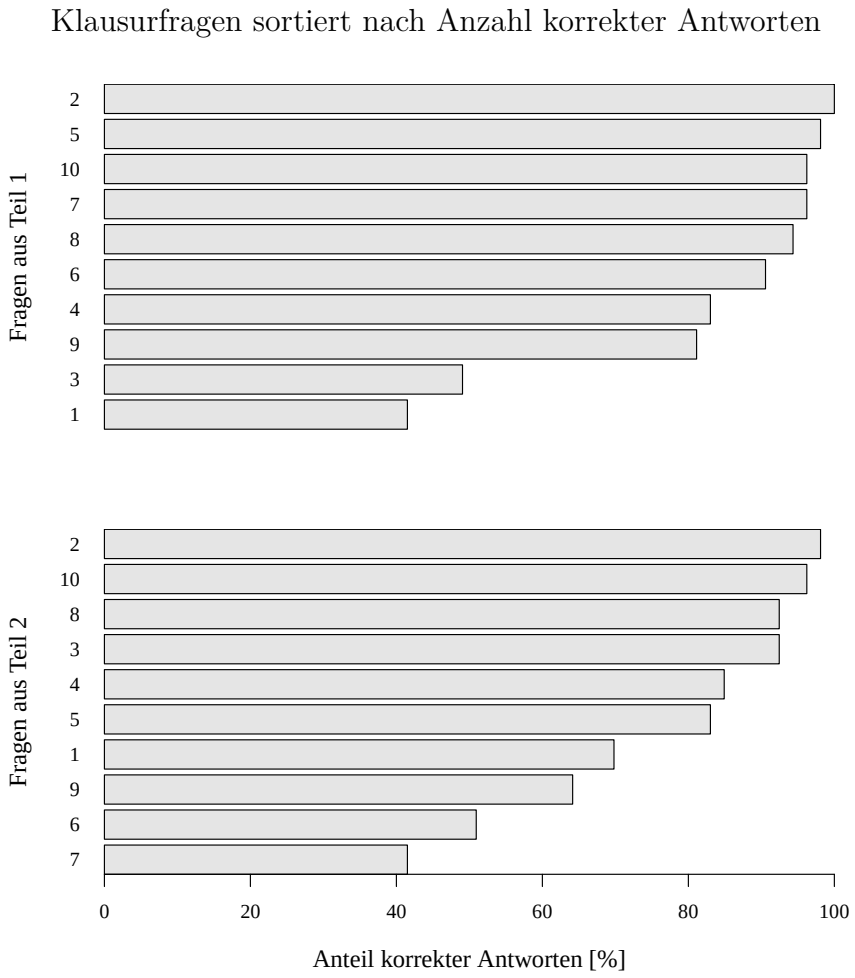
Die nachfolgende Abbildung zeigt die absolute Häufigkeitsverteilung der erzielten Punkte.




Die nachfolgende Abbildung zeigt die absolute Häufigkeitsverteilung der erreichten Noten.



Die zwei nachfolgenden Abbildungen zeigen jeweils die Reihenfolge der Klausurfragen, sortiert nach Anteil korrekter Antworten über alle Teilnehmenden für beide Teile der Klausur.



Otto-Von-Guericke-Universität Magdeburg
Institut für Psychologie
Abteilung Methodenlehre I: Methoden der experimentellen und neurowissenschaftlichen Psychologie
Belinda Fleischmann

Klausur Modul C Einführung in empirisch-wissenschaftliches Arbeiten
Termin: 26.07.2024

Name, Vorname: _____

Matrikelnummer: _____

Bearbeitungshinweise

- Die Klausur besteht aus **20 Aufgaben** und ist in **2 Teile** mit jeweils 10 Fragen untergliedert.
- Teil 1 ist closed-book. Teil 2 ist open-book. Im open-book Teil können Sie das Internet und R am Prüfungs-PC nutzen.
- Sie haben zur Bearbeitung insgesamt **60 Minuten** Zeit.
- Bei jeder Aufgabe sind jeweils **vier Antwortmöglichkeiten** vorgegeben, es trifft **immer genau eine** Antwort zu.
- Bitte kreuzen Sie bei jeder Aufgabe die zutreffende Antwort an.
- Für jede richtig gelöste Aufgabe erhalten Sie einen Punkt.
- Die Kommunikation mit anderen Klausurteilnehmenden ist nicht gestattet.
- Die Klausur ist bestanden, wenn Sie mindestens 10 Punkte erreichen.

Viel Erfolg!

# Teil 1 - Inhalte aus Einheiten (1) bis (3)

Alle Fragen zu Ethik und ethischen Formalitäten beziehen sich auf die „Deklaration von Helsinki“ der *World Medical Association (WMA)*, die „Ethical Principles of Psychologists and Code of Conduct (EPPCC)“ der *American Psychological Association (APA)* und den Leitfaden „Ethisches Handeln in der psychologischen Forschung“ der *Deutsche Gesellschaft für Psychologie (DGPs)*.

1. Welche Aussage bezüglich ethischer Grundsätze entstammt der „Deklaration von Helsinki“?

   a) Die ärztliche Pflicht, Leben, Gesundheit, Würde, Integrität, Selbstbestimmungsrecht und Privatsphäre von Patient:innen zu fördern und zu erhalten, soll auf die psychologische Forschung erweitert werden.

   b) Medizinische Forschung am Menschen darf nur durchgeführt werden, wenn die Bedeutung des Ziels die Risiken und Belastungen für die Versuchspersonen ausgleichen.

   c) Unterrepräsentierte Gruppen sollen angemessenen Zugang zu Forschungsstudien erhalten.

   d) Begonnene Studien sollten weder modifziert oder abgebrochen werden.

2. Welche Aussage bezüglich des Prinzips des „Respekt vor Selbstbestimmung“ ist gemäß der Leitlinien der DGPs zutreffend?

   a) Die Freiwilligkeit der Teilnahme ist nur für zwei Aspekte einer psychologischen Studie relevant: Die Rekrutierung von Studienteilnehmenden und die informierte Einwilligung.

   b) Bei der Höhe einer monetären Aufwandsentschädigung gilt je höher, desto besser.

   c) Die Verwendung einer Coverstory (Täuschung) ist grundsätzlich unvereinbar mit dem Prinzip der informierten Einwilligung, aber in bestimmten Fällen ethisch vertretbar.

   d) In der Teilnehmendeninformation sollten Fachbegriffe verwendet werden, um potentiellen Studienteilnehmenden das Forschungsvorhaben gegenüber möglichst sachgemäß und transparent zu kommunizieren.

3. Welche Aussage bezüglich Ethikvoten entstammt den Leitlinien der DGPs?

   a) Ein positives Ethikvotum kann nur dann ausgestellt werden, wenn alle vier Prinzpien ethischen Handelns in der psychologischen Forschung befolgt werden.

   b) Ein positives Ethikvotum kann nur dann ausgestellt werden, wenn die Dokumentvorlagen für die allgemeine Information für Studienteilnehmende sowie für spezifische Studien (z.B. MRT-Studien, EEG-Studie) der DGPs verwendet werden.

   c) Nach positivem Ethikvotum darf der Studienablauf nicht mehr verändert werden.

   d) Nach positivem Ethikvotum wird von Seiten einer Ethikkommission nicht weiter überprüft, ob alle Kriterien zur Einhaltung ethischer Richtlinien von Studiendurchführenden eingehalten werden.

4. Während der Planung einer psychologischen Studie überlegt eine Universitätsprofessorin, Studierende aus ihrer Vorlesung zu rekrutieren. Es sei jedem und jeder Kursteilnehmer:in freigestellt, ob sie an der Studie teilnehmen möchte oder nicht. Für die Teilnahme ist eine großzügige Aufwandsentschädigung geplant. Welche ethischen Überlegungen sind in diesem Fall angebracht?

   a) Wenn sichergestellt wird, dass die Entscheidung über die Teilnahme anonymisiert erfasst wird, ist die Freiwilligkeit der Teilnahme sichergestellt.

   b) Die Aufwandsentschädigung sollte so hoch wie möglich angesetzt werden, um dem Prinzip der Nutzen-Risiko-Abwägung gerecht zu werden.

   c) Eine zu hohe Aufwandsentschädigung könnte die Freiwilligkeit der Studienteilnahme verzerren.

   d) Da die Studierenden mit den Forschungsthemen der Forschenden vertraut sind, kann auf eine ausführliche allgemeine Information für Teilnehmende verzichtet werden.

5. Was ist der Unterschied zwischen anonymisierten und pseudonymisierten Daten?

   a) Bei Anonymisierung ist eine indirekte fehlerfreie Zuordnung von Daten zu einer Person noch möglich, bei Pseudonymisierung nicht.

   b) Beide Varianten enthalten Pseudonyme, aber nur anomyisierte Daten enthalten auch Anonmye.

   c) Pseudonymisierte Daten enthalten auch Dummy-Variablen, anonymisierte Daten nicht.

   d) Anonymisierte Daten können nicht zu einer Person zugeordnet werden, pseudonymisierte Daten können prinzipiell einer Person zugeordnet werden.

6. Welcher der folgenden ist **kein** Bestandteil des „Goldenen Wegs“ wissenschaftlichen Publizierens?

   a) Die Zweitveröffentlichung des Artikels in einem Online-Repositorium.

   b) Der Artikel wird für alle frei zugänglich gemacht.

   c) Der oder die Autor:in oder ihre Institution bezahlt eine Gebühr an den Verlag, um Kosten für Redaktionsprozess, technische Struktur und Marketing zu decken.

   d) Der oder die Editor:in des Verlags schickt den Artikel an andere Wissenschaftler:innen zur Überprüfung.

7. Was ist **keine** zentrale Aussage der San Francisco Declaration on Research Assessment?

   a) Bei der Forschungsevaluation sollten auf Fachzeitschriften basierende Kennzahlen nur dann einbezogen werden, wenn es sich um Entscheidungen über Einstellung oder Beförderung von Wissenschaftler:innen handelt.

   b) Bei der Forschungsevalution im Rahmen von Entscheidungen über Finanzierung, Einstellung oder Beförderung sollte auf Kennzahlen, die auf Fachzeitschriften basieren, verzichtet werden.

   c) Es sollten neue Kennzahlen für Signifikanz und Bedeutung untersucht werden.

   d) Möglichkeiten der Online-Veröffentlichung sollten verstärkt genutzt werden, um gegebenfalls unnötige Einschränkungen der Anzahl an Wörtern, Abbildungen und Literaturangaben in Artikeln zu vermeiden.

8. Was ist **kein** Unterschied zwischen Replizierbarkeit und Reproduzierbarkeit?

   a) Replizierbarkeit liegt vor, wenn andere Wissenschaftler:innen andere (neue) Daten und möglicherweise andere Methoden nutzen und zu den gleichen Schlussfolgerungen kommen wie eine Originalstudie, während Reproduzierbarkeit dann vorliegt, wenn andere Wissenschaftler:innen die gleichen Daten und Methoden verwenden und die gleichen Resultate erhalten.

   b) Replizierbarkeit ist abhängig von den unterstellten wahren, aber unbekannten, Parametern. Reproduzierbarkeit ist abhängig von der Sorgfältigkeit der wissenschaftlichen Arbeit.

   c) Für die Überprüfung der Reproduzierbarkeit benötigen Wissenschaftler:innen Zugriff auf die Daten der Originalstudie, für die Überprüfung der Replizierbarkeit nicht.

   d) Für Replizierbarkeit ist größtmögliche Methoden- und Datentransparenz wichtig, für Reproduzierbarkeit nicht.

9. Welche der folgenden ist **keine** Maßnahme zur Erhöhung der Methodentransparenz?

   a) Verfügbarkeit der datenanalytischen Skripte zum Zeitpunkt der ersten Veröffentlichung des korrespondiernden Artikels.

   b) Die Formatierung der digitalen Daten anhand des Datenstandards Brain Imaging Data Structure (BIDS).

   c) Verweis auf Datenanalyseskripte in Methoden- und Ergebnisabschnitten des veröffentlichten Artikels.

   d) Veröffentlichung aller zur Datenanalyse verwendeten Computercodes in gut strukturierter und extensiv kommentierter Form.

10. Welche Art des Data-Sharings hat den Nachteil, dass sie unwiderruflich ist?

    a) Public Sharing

    b) Restricted Sharing

    c) Dynamic Sharing

    d) Open Science Framework

# Teil 2 - Praktische Anwendung ALM

In den nachfolgenden Aufgaben werden Regressionsanalysen in **R** durchgeführt. Dazu sei durch untenstehenden **R** Code das **R** Dataframe `D` definiert, das einen Datensatz von Werten zweier unabhängigen Variablen `X1` und `X2` und einer abhängigen Variablen `Y` repräsentiert. Alle Referenzierungen des Datensatzes `D` im Folgenden beziehen sich auf das hier definierte Dataframe.

```
D <- data.frame(
   Y  = c(-0.4, 1, -2, 2, 0.2, -0.3, 0.1, -0.4, -1.8, -1.3, 1.3, 3),
   X1 = c(0.6, 2.2, -1.1, 0, 0, -0.9, -0.8, -0.6, -0.9, -0.8, -0.1, 2),
   X2 = c(-0.6, 0.2, -0.8, 1.6, 0.3, -0.8, 0.5, 0.7, 0.6, -0.3, 1.5, 0.4)
)
```

1. Welcher der folgenden **R** Befehle bestimmt die Anzahl `n` der Datenpunkte in der abhängigen Variable?

   a) `n <- length(D)`

   b) `n <- ncol(D$X2)`

   c) `n <- length(D[Y])`

   d) `n <- nrow(D)`

2. `n` sei die Anzahl der Datenpunkte in der abhängigen Variable. Welcher der folgenden **R** Befehle erstellt eine Designmatrix, die Teil eines multiplen Regressionsmodells mit einem Interzeptparameter und zwei kontinuierlichen Variablen als Regressoren für den vorliegenden Datensatz sein kann?

   (a) `X <- matrix(c(rep(1,n), D$X1, D$X2), nrow = n)`

   (b) `X <- matrix(c(D$Y, D$X1, D$X2), nrow = n)`

   (c) `X <- matrix(c(D$X1, D$X2), ncol = n)`

   (d) `X <- matrix(c(rep(1,n), D$Y, D$X1, D$X2), nrow = n)`

3. `beta_hat` sei die **R** Repräsentation des Betaparameterschätzers und `X` die **R** Repräsentation der Designmatrix des Modells einer multiplen Regression mit einem Interzeptparameter und zwei kontinuierlichen Variablen als Regressoren. Welcher der folgenden **R** Befehle ergibt die Ausgabe `TRUE`?

   (a) `ncol(X) == length(beta_hat)`

   (b) `nrow(X) == ncol(D)`

   (c) `ncol(X) == length(D$Y)`

   (d) `length(D$Y) == ncol(X %*% beta_hat)`

4. `X` sei die **R** Repräsentation der Designmatrix des Modells einer multiplen Regression mit einem Interzeptparameter und zwei kontinuierlichen Variablen als Regressoren. Welcher der folgenden **R** Befehle bestimmt die Anzahl `p` der Betaparameter?

   (a) `p <- ncol(X) - 1`

   (b) `p <- nrow(X) - 1`

   (c) `p <- ncol(X)`

   (d) `p <- ncol(X) + 1`

5. `X` sei die **R** Repräsentation der Designmatrix des Modells einer multiplen Regression für obigen Datensatz. Welcher der folgenden **R** Befehle berechnet den Betaparameterschätzer dieses Modells für die gegebenen Werte der abhängigen Variable?

   a) `beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y`

   b) `beta_hat <- solve(t(X) %*% X) %*% t(X) %*% D$Y`

   c) `beta_hat <- coef(lm(Y ~ X1 + X2))`

   d) `beta_hat <- coef(lm(y ~ D$X1 + D$X2))`

6. `X` sei die **R** Repräsentation der Designmatrix des Modells einer multiplen Regression. Der untenstehende **R** Code bestimmt die Stichprobenkorrelationsmatrix `r`, die partiellen Stichprobenkorrelationsmatrix `pr` und den Vektor `s` mit den Standarabweichungen für jede Spalte des obigen Datensatzes. Welcher der folgenden **R** Befehle berechnet den Betaparameterschätzer für den Regressor X2 für die gegebenen Werte der abhängigen Variable in obigem Datensatz?

```
library(ppcor)
s  <- apply(D, 2, sd)
r  <- cor(D)
pr <- pcor(D)$estimate
```

   a) `beta_hat_2 <- pr[1,2]*sqrt((1-r[1,3]^2)/(1-r[2,3]^2))*(s[1]/s[2])`

   b) `beta_hat_2 <- pr[1,3]*sqrt((1-r[1,2]^2)/(1-r[3,2]^2))*(s[1]/s[3])`

   c) `beta_hat_2 <- coef(lm(Y ~ X1 + X2, data = D))[2]`

   d) `beta_hat_2 <- (solve(t(X) %*% X) %*% t(X) %*% D$Y)[2]`

7. Die **R** Variable `beta_hat` sei der Betaparameterschätzer des Modells einer multiplen Regression für obigen Datensatz. Was erzeugt der folgende **R** Befehl?

```
plot(D$X1, D$Y)
abline(coef = c(beta_hat[1], beta_hat[2]))
```

   a) Ein Streudiagramm und eine Ausgleichsgerade, die den linearen Zusammenhang zwischen der abhängigen Variable `Y` und dem Regressor `X1` im Rahmen einer multiplen Regression visualisieren.
   b) Ein Streudiagramm und eine Ausgleichsgerade, die den linearen Zusammenhang zwischen der abhängigen Variable `Y` und dem Regressor `X1` im Rahmen einer einfachen linearen Regression visualisieren.
   c) Eine Fehlermeldung, weil sich im Rahmen einer multiplen Regression keine zweidimensionalen Zusammenhänge darstellen lassen.
   d) Eine Fehlermeldung, weil die Eingabe der Daten eines zweiten Regressors fehlt.

8. Die **R** Variablen `X` und `beta_hat` seien die **R** Repräsentationen der Designmatrix und des Betaparameterschätzers für das Modell der multiplen linearen Regression bezüglich obigen Datensatzes. Welcher der folgenden **R** Befehle berechnet den Residuenvektor hinsichtlich der Werte der abhängigen Variablen?

   a) `eps_hat <- D$AV - X %*% beta_hat`
   b) `eps_hat <- D$AV - D$X %*% beta_hat`
   c) `eps_hat <- D$Y - c(X1 + X2) %*% beta_hat`
   d) `eps_hat <- D$Y - X %*% beta_hat`

9. Die **R** Variable `eps_hat` sei der Residuenvektor des multiplen Regressionsmodels mit einem Interzeptparameter und zwei kontinuierlichen Variablen als Regressoren bezüglich obigen Datensatzes. Welcher der folgenden **R** Befehle berechnet den Varianzparameterschätzer des multiplen Regressionsmodells bezüglich obigen Datensatzes?

   a) `sigsqr_hat <- (t(eps_hat) %*% eps_hat) / (10 - 2)`
   b) `sigsqr_hat <- (t(eps_hat) %*% eps_hat) / (n - 1)`
   c) `sigsqr_hat <- (1 / 9) * (t(eps_hat) %*% eps_hat)`
   d) `sigsqr_hat <- (t(eps_hat) %*% eps_hat) / (8)`

10. Welcher der folgenden **R** Vektoren ist ein geeigneter Kontrastgewichtsvektor zur Bestimmung der T-Statistik für den Interzeptparameter des Modells der multiplen Regression bezüglich obigen Datensatzes?

    a) `matrix(c(1,0), nrow = 3)`
    b) `matrix(c(1,0), nrow = 2)`
    c) `matrix(c(1,0,0), nrow = 3)`
    d) `matrix(c(1,0,0), nrow = 2)`

**Lösungen:**

Teil 1

1. c)
2. c)
3. d)
4. c)
5. d)
6. a)
7. a)
8. d)
9. b)
10. a)

Teil 2

1. d)
2. a)
3. a)
4. c)
5. b)
6. b)
7. a)
8. d)
9. c)
10. c)