Table of Contents

Frage 1:

Im Rahmen einer Regressionsfragestellung schätzen Sie einen Entscheidungsbaum. Im Folgenden sehen Sie die Beobachtungen der Kriteriumswerte in den Kinderknoten zweier potentieller Splits:

Split 1:

- linker Kinderknoten: 2, 5, 1
- rechter Kinderknoten: 3, 4, 99

Split 2:

- linker Kinderknoten: 1, 2, 3, 4, 5
- rechter Kinderknoten: 99

FRAGE:

Welcher Split würde in diesem Fall gewählt werden? Begründen Sie, indem Sie für beide Splits die Impurity Reduction berechnen.

Frage 2

siehe Folie 3, S. 21,22

Sie interessieren sich für den Einfluss der stetigen Variable *Extraversion* auf die Wahrscheinlichkeit dafür, im nächsten Jahr eine depressive Episode zu erleben.

Sie stellen das folgende logistische Regressionsmodell auf:

$$P(Y_i=1|x_i)=\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$$

Hierbei steht $Y_i=1$

für das Ereignis, dass Person i im nächsten Jahr eine depressive Episode erlebt und x_i für den Extraversionswert der Person i.

Sie erhalten den folgenden R-Output:

```
summary(fit)

Call:
glm(formula = y ~ extraversion, family = "binomial", data = daten)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5509  -1.1574   0.6039   1.0884   1.6420

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.02439    0.20884   0.117   0.9070
extraversion -0.56885    0.20822  -2.732   0.0063 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 138.59  on 99  degrees of freedom
Residual deviance: 130.14  on 98  degrees of freedom
AIC: 134.14

Number of Fisher Scoring iterations: 4
```

Sie erhalten zudem den folgenden R-Output mit 95%-Konfidenzintervallen für die Parameter e^{α} und e^{β}:

```
confints
                 2.5 %    97.5 %
(Intercept)  0.6791974 1.5454069
extraversion 0.3670548 0.8359887
```

a. Zu welchem Ergebnis kommt der Hypothesentest für β bei einem Signifikanzniveau von $\alpha=0.05$? Interpretieren Sie das Ergebnis.
b. Interpretieren Sie das Konfidenzintervall für e^{β}

c. Berechnen Sie die geschätzte Wahrscheinlichkeit dafür, dass eine Person mit Extraversion $x_i=1$ im nächsten Jahr eine depressive Episode erlebt.

Frage 3

Sie führen ein Benchmark-Experiment für eine Regressions-Fragestellung mit vier verschiedenen Modellen durch:

- Dummy-Regressionsmodell (`regr.featureless`)
- Lineares Regressionsmodell (`regr.lm`)
- Entscheidungsbaum (`regr.rpart`)
- Random Forest (`regr.ranger`)

Um den erwarteten MSE der Modelle zu schätzen, verwenden Sie eine 10-Fold-Cross-Validation und erhalten den folgenden Output:

```
     task.id       learner.id mse.test.mean
1 Regression regr.featureless     12.635711
2 Regression          regr.lm      9.323205
3 Regression       regr.rpart     10.717077
4 Regression      regr.ranger      2.061365
```

Zudem trainieren sie alle Modelle auf dem Gesamtdatensatz und berechnen den MSE der so trainierten Modelle ebenfalls auf dem Gesamtdatensatz. Hierfür erhalten Sie den folgenden Output:

```
     task.id       learner.id mse.train
1 Regression regr.featureless 12.964081
2 Regression          regr.lm  8.827584
3 Regression       regr.rpart  1.502561
4 Regression      regr.ranger  1.524381
```

a. Wie erklären Sie sich den Unterschied zwischen dem Cross-Validation-Schätzwert für den erwarteten MSE und dem MSE im Gesamtdatensatz für den Entscheidungsbaum?
b. Wie erklären Sie sich den Unterschied im Cross-Validation-Schätzwert für den erwarteten MSE zwischen dem linearen Regressionsmodell und dem Random Forest?
c. Für welches Modell würden Sie sich entscheiden? Begründen Sie.

Frage 4

Im Rahmen einer 2-fold Cross Validation zur Schätzung des erwarteten Vorhersagefehlers eines logistischen Regressionsmodells erhalten Sie für die beiden Testssets die folgenden Konfusionsmatrizen:

Testset 1:

```
   y_dach
y    0 1
  0  0 1
  1  2 0
```

Testset 2:

```
   y_dach
y    0 1
  0 0 2
  1 1 0
```

In beiden Fällen bezeichnet `y` die wahren Kriteriumswerte und `y_dach` die vorhergesagten Kriteriumswerte. Diese wurden durch das logistische Regressionsmodell vorhergesagt, welches im Trainingsset trainiert wurde.

a. Berechnen Sie den Cross-Validation-Schätzwert für den erwarteten MMCE des logistischen Regressionsmodells.
b. Berechnen Sie den Cross-Validation-Schätzwert für den erwarteten MMCE eines Dummy Klassifikationsmodells.

Frage 5

Sie betrachten den folgenden Ausschnitt aus einem Output mit Schätzwerten für die Itemparameter eines probabilistischen Testmodells:

```
Coefficients:
 $Item1
        value std.err z.value
Catgr.1  -0.5    0.25   -2.00
Catgr.2   0.0    0.17    0.00
Catgr.3   0.5    0.13    3.85
Dscrmn    0.2    0.39    0.51

$Item2
        value std.err z.value
Catgr.1  -0.5    0.25   -2.00
Catgr.2   0.5    0.24    2.08
Catgr.3   1.0    0.14    7.14
Dscrmn    0.7    0.10    7.00
```

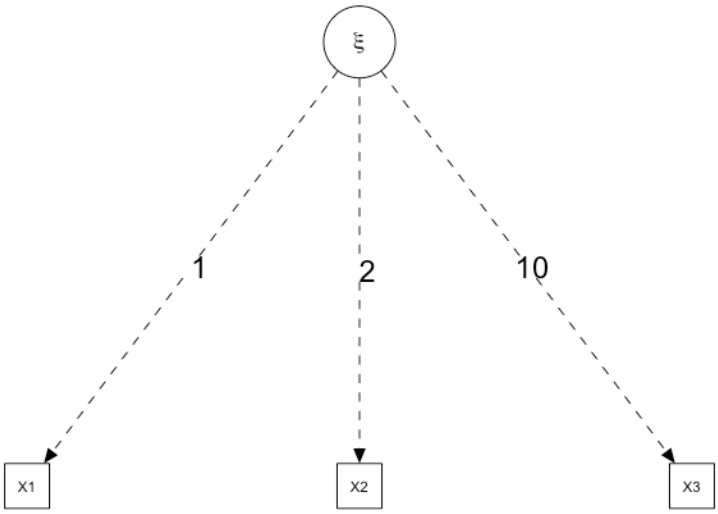
- Wie viele Antwortkategorien hat Item 1? Begründen Sie.
- Die Parameter welchen Modells wurden hier geschätzt? Begründen Sie.
- Interpretieren Sie den Schätzwert für den mit `Catgr.1` bezeichneten Parameter des ersten Items.

Frage 6

Sie betrachten einen psychologischen Test, dessen drei Items einem Rasch-Modell mit Itemparametern $\sigma_1=1$, $\sigma_2=0$ und $\sigma_3=-1$ folgen. Berechnen Sie für eine Person p mit Antwortmuster $x_{-p}=(0;1;0)$ die UML-Likelihood des Personenparameters $\theta_p=0$.

Frage 7

Sie untersuchen folgendes Messmodell (Fehlerterme werden nicht angezeigt!):



a. Bestimmen Sie die Freiheitsgrade des Modells.
b. Bestimmen Sie die modell-implizierte Kovarianz der Variablen X_2 und X_3 sowie die modell-implizierte Varianz der Variable X_2, gegeben der Parameterwerte $\sigma^2(\xi) = 2.51$, $\sigma^2(\varepsilon_1) = 3.95$, $\sigma^2(\varepsilon_2) = 0.76$ und $\sigma^2(\varepsilon_3) = 0.49$.
c. Angenommen die Fehlervarianzen sind gleich groß. Welche manifeste Variable sollte die größte Varianz aufweisen?

Frage 8

Sie untersuchen die Zusammenhangstrukturen mehrerer manifester Variablen mit einer konfirmatorischen Faktorenanalyse und nehmen an, dass drei unkorrelierte latente Variablen A, B und C den manifesten Variablen zugrunde liegen.

Sie erhalten den folgenden Output:

```
lavaan::summary(fit, fit.measures = TRUE)
## lavaan 0.6-7 ended normally after 34 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of free parameters                         38
##
##   Number of observations                           414
##
## Model Test User Model:
##
##   Test statistic                               177.129
##   Degrees of freedom                               152
##   P-value (Chi-square)                           0.080
##
## Model Test Baseline Model:
##
##   Test statistic                              3332.574
##   Degrees of freedom                               171
##   P-value                                        0.000
```

```
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.992
##   Tucker-Lewis Index (TLI)                       0.991
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)             -12428.945
##   Loglikelihood unrestricted model (H1)     -12340.380
##
##   Akaike (AIC)                               24933.890
##   Bayesian (BIC)                             25086.873
##   Sample-size adjusted Bayesian (BIC)        24966.290
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.020
##   90 Percent confidence interval - lower         0.000
##   90 Percent confidence interval - upper         0.031
##   P-value RMSEA <= 0.05                          1.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.107
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   A =~
##     X1                1.000
##     X2                0.901    0.066   13.566    0.000
##     X3                0.877    0.066   13.212    0.000
##     X4                0.895    0.066   13.668    0.000
##     X5                1.021    0.072   14.135    0.000
##     X6                0.931    0.069   13.456    0.000
##     X7                0.938    0.067   14.012    0.000
##     X8                0.930    0.066   14.135    0.000
##   B =~
##     Y1                1.000
##     Y2                0.993    0.086   11.602    0.000
##     Y3                0.952    0.080   11.856    0.000
##     Y4                0.947    0.083   11.391    0.000
##   C =~
##     Z1                1.000
##     Z2                0.979    0.066   14.887    0.000
##     Z3                0.946    0.064   14.745    0.000
##     Z4                0.943    0.068   13.952    0.000
##     Z5                0.899    0.064   14.070    0.000
##     Z6                0.936    0.065   14.377    0.000
##     Z7                1.007    0.065   15.518    0.000
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   A ~~
##     B                 0.000
##     C                 0.000
##   B ~~
```

```
##     C                 0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .X1                0.986    0.080   12.261    0.000
##    .X2                1.014    0.080   12.711    0.000
##    .X3                1.057    0.082   12.867    0.000
##    .X4                0.973    0.077   12.662    0.000
##    .X5                1.108    0.089   12.416    0.000
##    .X6                1.115    0.087   12.761    0.000
##    .X7                0.969    0.078   12.485    0.000
##    .X8                0.920    0.074   12.416    0.000
##    .Y1                0.968    0.095   10.147    0.000
##    .Y2                1.136    0.105   10.840    0.000
##    .Y3                0.915    0.089   10.324    0.000
##    .Y4                1.135    0.102   11.171    0.000
##    .Z1                0.919    0.077   11.987    0.000
##    .Z2                0.974    0.080   12.218    0.000
##    .Z3                0.945    0.077   12.302    0.000
##    .Z4                1.158    0.091   12.699    0.000
##    .Z5                1.021    0.081   12.646    0.000
##    .Z6                1.021    0.082   12.499    0.000
##    .Z7                0.862    0.073   11.790    0.000
##     A                 1.160    0.139    8.345    0.000
##     B                 1.055    0.141    7.495    0.000
##     C                 1.222    0.141    8.672    0.000
```

a. Wie viele bekannte Parameter gibt es in diesem Fall?
b. Interpretieren Sie die Ladung der Variable `Y3` inhaltlich.
c. Begründen Sie mithilfe der Fit-Indizes, wie sich die Modellpassung verbessern ließe.
d. Angenommen der Wert der χ^2-Teststatistik wäre um 14 kleiner, welchen Wert hätte der CFI angenommen (geben Sie mindestens vier Nachkommastellen an)?

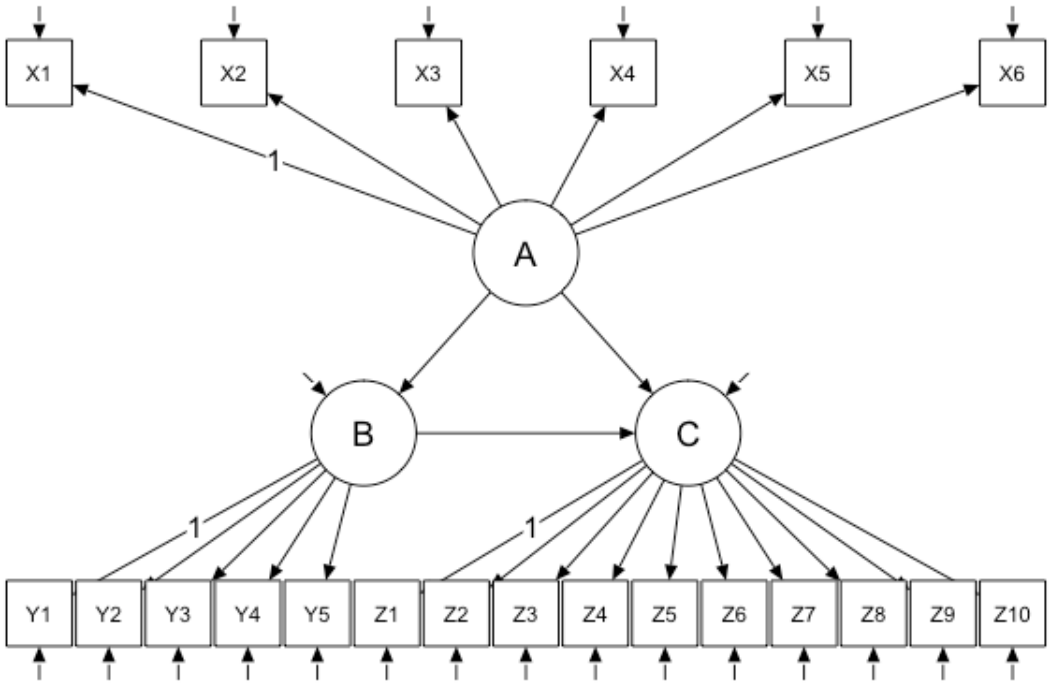
d. Sie vergleichen die Modellpassung mit einem komplexeren Alternativmodell (`fit2`):

```
AIC(fit, fit2)
##      df      AIC
## fit  38 24933.89
## fit2 41 24873.61
```

Ist das komplexere Modell vorzuziehen?

Frage 9

Welche der folgenden Aussagen treffen gegeben der Grafik und des Outputs zu?



Modifikationsindizes (Auswahl):

	lhs	op	rhs	mi	epc
51	A	=~	Y3	102.309	4.952
53	A	=~	Y5	21.592	-1.076
199	Y1	~~	Y5	20.692	0.534
235	Y4	~~	Y5	18.550	0.492
196	Y1	~~	Y2	14.077	0.441
49	A	=~	Y1	13.208	-0.842
223	Y3	~~	Y4	11.177	-0.483
198	Y1	~~	Y4	11.108	0.377
262	Z1	~~	Z8	9.809	0.308
100	X1	~~	Y5	9.427	-0.322
280	Z4	~~	Z5	9.219	0.234
50	A	=~	Y2	8.968	-0.704
197	Y1	~~	Y3	8.799	-0.434
211	Y2	~~	Y4	7.038	0.304
212	Y2	~~	Y5	6.152	0.295
186	X6	~~	Z1	5.922	0.215
86	C	=~	Y1	5.523	0.235
88	C	=~	Y3	5.303	-0.324
278	Z3	~~	Z9	5.209	0.181
269	Z2	~~	Z7	5.035	0.187

Parameterschätzungen (Auswahl):

	lhs	op	rhs	est	se	z	pvalue	ci.lower	ci.upper
1	A	=~	X1	1.000	0.000	NA	NA	1.000	1.000
2	A	=~	X2	0.864	0.098	8.804	0	0.672	1.057
3	A	=~	X3	0.862	0.107	8.087	0	0.653	1.071
4	A	=~	X4	0.966	0.112	8.602	0	0.746	1.187
5	A	=~	X5	1.042	0.123	8.471	0	0.801	1.283

```
6    A =~  X6 1.028 0.113 9.113      0    0.807    1.249
```

Wählen Sie eine oder mehrere Antworten:

a. Das Strukturmodell erlaubt die Überprüfung einer hypothetischen vollständigen Mediation.
b. Die Anzahl der bekannten Parameter beträgt ...?
c. Angenommen Person i hat einen latenten Variablenwert $a_i = 5$ und Person j einen latenten Variablenwert $a_j = 0$, dann würden wir einen Unterschied von ca. 4.31 auf der Variable $X3$ zwischen diesen Personen erwarten.

d. Es müssen 42 Parameter geschätzt werden.
e. Zur Schätzung des Modells müsste der Pfad zwischen A und B oder der Pfad zwischen A und C auf 1 fixiert werden.
f. Durch Aufnahme des Pfades zwischen A und der manifesten Variable $Y3$ ließe sich die $\chi 2$-Teststatistik um ca. 4.952 verringern und die Modellpassung verbessern.

Frage 10

In einer Studie soll untersucht werden, ob die Produktivität von Arbeitnehmer*innen (*productivity*) vom Typ des Unternehmens (*type*) und ihrer Zufriedenheit am Arbeitsplatz (*satisfaction*) abhängig ist. Dabei nehmen mehrere Arbeitnehmer/innen pro Firma (*company*) teil.

Die Variablen sind folgendermaßen skaliert:

- *productivity*: *z*-standardisierte Skala (grand-mean und -SD); größere Werte bedeuten mehr Produktivität
- *type*: 0 = mittelständisches Unternehmen, 1 = Industrie
- *satisfaction*: *z*-standardisierte Arbeitszufriedenheitsskala (grand-mean und -SD); größere Werte bedeuten mehr Zufriedenheit
- *company*: kategorialer Faktor (Unternehmens-ID)

Sie erhalten folgenden Output:

```
Formula: productivity ~ 1 + satisfaction  type + (1 + satisfaction |
company)

Random effects:
 Groups   Name         Variance Std.Dev. Corr
 company  (Intercept)  0.172578 0.41542
          satisfaction 0.006192 0.07869  -0.75
 Residual              3.986821 1.99670
Number of obs: 5000, groups:  company, 100

Fixed effects:
                  Estimate Std. Error       df t value Pr(>|t|)
(Intercept)       -0.13744    0.07105 97.96670  -1.934  0.05594 .
satisfaction      -0.11100    0.04170 99.14271  -2.662  0.00906
type              -0.20451    0.10048 97.96779  -2.035  0.04451
satisfaction:type  0.19465    0.05861 95.44148   3.321  0.00127
```

Geben Sie an, welche der folgenden Hypothesen Sie auf Basis der hier gezeigten Analyse beibehalten würden.

Wählen Sie eine oder mehrere Antworten:
a. Durchschnittlich zufriedene Arbeitnehmer*innen sind in der Industrie weniger produktiv als in mittelständischen Unternehmen.
b. In der Industrie ist Produktivität stärker von der Zufriedenheit der Arbeitnehmer*innen abhängig als in mittelständischen Unternehmen.
c. Bei mittelständischen Unternehmen hängt die Zufriedenheit am Arbeitsplatz positiv mit der Produktivität der Arbeitnehmer*innen zusammen.

Frage 11

Sie arbeiten mit einer hierarchischen Datenstruktur, in der Personen verschiedenen Firmen zugeordnet werden. Sie wollen herausfinden, wie stark die Zufriedenheit von Personen von ihrer Anzahl der gesammelten Turnschuhe, sowie der jeweiligen Branche (Pflege vs. Unternehmensberatung) und Größe ihrer Firma abhängt. Geben Sie die Variable/n an, die als Prädiktor/en auf Level 1 in die Modellgleichung einfließt/einfließen.

Wählen Sie eine oder mehrere Antworten:
a. Branche (Pflege vs. Unternehmensberatung)
b. keine der Variablen
c. Größe
d. Anzahl der gesammelten Turnschuhe
e. Zufriedenheit
f. Firma

Frage 12

Sie haben ein gemischtes lineares Modell mit zwei Leveln aufgestellt, bei dem tägliche Messzeitpunkte über 5 Wochen in Personen geschachtelt sind. Auf Level 1 haben Sie sowohl die Prädiktorvariable "Heutige Lerndauer Statistik (in Stunden)" (*lerndauer*), als auch die abhängige Variable "Zufriedenheit mit dem Tag" (auf einer Skala von -10 bis +10) gemessen (*daysat*). Es gibt keine weiteren Variablen in Modell. Sie lassen sowohl die Intercepts als auch die Slopes zufällig zwischen Personen variieren. Die Berechnung des Modells ergibt einen festen Effekt für *lerndauer* von 0.53, ein Intercept von 2.4, τ_{00}=0.67, τ_{01}=0.33, sowie τ_{11} von 0.02.

Interpretieren Sie den geschätzten Wert von τTau_00 inhaltlich.

Frage 13

Sie haben bei Schüler*innen aus unterschiedlichen Klassen einer Schule verschiedene Variablen erhoben. Auf Ebene der Schüler die sportliche Ausdauer (`aus`), die Konzentrationsfähigkeit (`kon`) und die Sprachkompetenz (`kom`), sowie auf Klassenebene die Lehrerfahrung der Klassenleitung (`erf`). Sie wollen die umfangreich theoretisch hergeleitete Hypothese prüfen, dass die Sprachkompetenz (`kom`) für Schüler*innen die ansonsten im Klassendurchschnitt liegen, einen Einfluss auf die Konzentrationsfähigkeit (`kon`) hat. Ihnen liegen folgende Rohdaten vor. Nehmen Sie die erforderliche Zentrierung vor.

Person	Klasse	aus	kon	kom	erf
1	1	8	7	14	8
2	1	9	10	17	8
3	1	7	6	14	8
4	2	10	9	14	10
5	2	6	9	10	10
6	3	10	7	15	7
7	3	6	10	13	7

Tragen Sie bitte die Zahlen bei Bedarf mit maximal zwei Dezimalstellen gerundet und mit Dezimalkomma (nicht -punkt) ein.

a. Welchen Wert hat Person 1 auf der zentrierten Variable?
b. Welchen Wert hat Person 5 auf der zentrierten Variable?

Frage 14

Sie haben eine Experience-Sampling-Studie durchgeführt, bei der die Teilnehmer*innen eines eintägigen Kletterkurses 8x nach dem Klettern verschiedener Routen die Flow-Kurzskala (`FKS`) ausgefüllt haben. Die Forschungsfrage war, ob das Flowgefühl (`FLOW`, mit der Flow-Kurzskala gemessen) durch den Schwierigkeitsgrad der jeweiligen Kletterroute (`SCHW`; auch 8x gemessen) zusammenhängt, und ob dieser Zusammenhang bei Männern oder Frauen größer ist (`SEX`).

Die Skalierung der Variablen war folgendermaßen:

- `MZP`: Messzeitpunkt, von 0 (erste Befragung) bis 7 (letzte Befragung)
- `FLOW`: Fragebogenskala von 0 bis 18
- `SCHW`: Schwierigkeitsgrad der aktuellen Route, von 0 (=Kindergarten) bis 10 (=Schlucht des Todes)
- `SEX`: -1 = Frau, 1 = Mann

Sie stellen folgende Modellgleichungen auf:

$$\text{FLOW}_{ti}=\pi_{0i}+\pi_{1i}*\text{SCHW}_{ti}+r_{ti}$$

$$\pi_{0i}=\beta_{00}+\beta_{01}*\text{SEX}_i+u_{0i}$$

$$\pi_{1i}=\beta_{10}+\beta_{11}*\text{SEX}_i+u_{1i}$$

Das Modell ergab folgende Parameterschätzung:

$$\beta_{00}=2.4,\ \beta_{01}=-1.1,\ \beta_{10}=0.7,\ \beta_{11}=0.4,\ \hat{\tau}_{00}=1.48,\ \hat{\tau}_{11}=0.03,\ \hat{\sigma}_2=2.30.$$

Eine Frau steigt an Messzeitpunkt 6 in einer Route der Schwierigkeit 4 ein.
Was ist ihr prognostizierter Flow-Wert?

Berechnen Sie den Wert auf zwei Nachkommastellen gerundet.

Frage 15

Sie haben die Hypothese, dass die Negative Selbstbewertung (`sel`) von Patient*innen (Level 1) und die Empathiefähigkeit (`emp`) ihrer jeweiligen Therapeut*in (Level 2) einen Einfluss auf die Häufigkeit dysfunktionaler Kognitionen (`dys`) der Betroffenen hat und dass es hier auch einen Interaktionseffekt gibt.

Ihr Kollege hat zur Untersuchung der Frage folgende gemischte Modellgleichung aufgestellt.

$$dys = \gamma_{00} + \gamma_{01} * sel + \gamma_{10} * emp + \gamma_{11} * sel * emp + u_{0j} + u_{1j} * emp + r_{ij}$$

Er hat dann anhand der aufgestellten Modellgleichung für die Berechnung mit `lmer` in R folgenden `lmer`-Aufruf aufgestellt:

```
lmer(dys ~ 1 + sel + emp + sel:emp + (1 + emp|therapeut), data = df).
```

a. Können Sie anhand dieses Modells Ihre Hypothese überprüfen? Wenn nicht, warum?

b. Unabhängig von Ihrer Antwort auf die erste Frage: Ist die Übersetzung der Formel in den R-Code korrekt? Wenn nicht, welcher Fehler ist hier aufgetreten?

Frage 16

Sie arbeiten mit einer hierarchischen Datenstruktur, in der Personen verschiedenen Firmen zugeordnet werden. Sie wollen herausfinden, wie stark die Produktivität von Personen von ihrer Anzahl der gesammelten Turnschuhe abhängt. Um diesen Zusammenhang möglichst präzise zu schätzen, möchten sie berücksichtigen, dass in jeder Firma, der die Personen zugeordnet werden können, dieser Zusammenhang anders sein könnte. Geben Sie die Variable/n an, die Sie zu diesem Zweck als Prädiktor/en auf Level 2 in die Modellgleichung einfließen lassen.

Wählen Sie eine oder mehrere Antworten:
a. Anzahl der gesammelten Turnschuhe
b. Branche der Firma (Pflege vs. Unternehmensberatung)
c. Produktivität
d. keine der Variablen
e. Firma