Andreas Bender, Philipp Kopper
Institut für Statistik

Statistische Software (R) – Hausarbeit 3

Sommersemester 2021, 28.07.2021 - 18.08.2021

Name: _____

Immatrikulationsnummer: _____

Studiengang: _____

Hiermit bestätige ich, dass ich die Anweisungen auf diesem Blatt gelesen und verstanden habe. Ich bestätige, dass die abgegebene Lösung vollständig und alleinig von mir bearbeitet und erstellt worden ist, ohne Hilfe von anderen in Anspruch zu nehmen. Ich bestätige, dass ich über die Vorlesungsmaterialien hinausgehende Quellen wie Bücher oder Internetseiten im Code angegeben und falls zutreffend verlinkt sind.

Unterschrift: _____

Prüfungshinweise:

1. Überprüfgen sie ob die heruntergeladene Angabe vollständig ist. Sie sollte 4 Aufgaben enthalten. Einzelne Aufgaben können aus mehreren Teilaufgaben bestehen.
2. Insgesamt können (ohne Bonuspunkte) 80 Punkte erreicht werden. Die Aufteilung der Punkte auf die einzelnen Aufgaben kann der Angabe entnommen werden.
3. **Die Lösung soll in Form von `.Rmd` Dateien abgegeben werden**. Für jede Aufgabe soll dabei eine separate `.Rmd` datei erstellt werden. Nutzen Sie Markdown für Lösungen die keinen Code erfordern und um Beginn und Ende einzelner Teilaufgaben zu kennzeichnen. Ist die Zugehörigkeit von Code zu einer der (Teil-)Aufgaben nicht eindeutig deklariert, kann es passieren, dass Sie dafür keine Punkte bekommen.
4. Achten Sie darauf, dass ihre `.Rmd` Dateien in einem frischen Environment kompilieren. **Sollten die Dateien nicht kompilieren, erfolgt die Bewertung mit 0 Punkten**. Achten Sie insbensondere darauf, dass alle zum kompilieren notwendigen Dateien (z.B. zusätzliche `.R` files oder Datensätze) in Ihrer Abgabe vorhanden sind. Verwenden Sie keine absoluten Pfade.
5. Sollten Sie technische oder andere Schwierigkeiten haben, kontaktieren Sie Bitte die Kursleiter. E-mail: andreas.bender@stat.uni-muenchen.de, philipp.kopper@stat.uni-muenchen.de. (Bitte die Emails an alle gelisteten Personen schicken!)
6. Die Aufgaben müssen alle eigenständig bearbeitet werden. Insbesondere sind keine Arbeitsgruppen erlaubt und sonstige Diskussion der Aufgaben und Lösungen mit anderen Personen (egal ob diese Statistik studieren oder nicht) nicht zulässig.
7. Das Internet kann passiv genutzt werden. D.h. es dürfen Internetseiten oder Foren aufgerufen und gelesen werden, das aktive Stellen von Fragen, die relevant zur Lösung der Aufgaben sind, ist allerdings nicht zulässig. Ebenso dürfen keine Aufgaben oder Lösungsvorschläge und anderen Hinweise im Internet gepostet oder per Chat, Email und anderen Kommunikationswegen diskutiert oder verteilt werden.

8. Sollte der Verdacht auf Plagiat, Betrug oder anderweitig unzulässiges Verhalten bestehen, können zusätzliche (mündliche) Prüfungen einberufen werden um die eigenständige Bearbeitung der Aufgaben zu prüfen.
9. Zweifel an der eigenständigen Bearbeitung ihrer Abgabe führen zum nicht-bestehen der Prüfung und dem Einschalten des Prüfungsausschusses.
10. Die Abgabe erfolgt bis Mitternacht (23:59 Uhr) am 18.08.2021.

Achten Sie bei der Bearbeitung insgesamt darauf, dass alle top-level Funktionen gut dokumentiert sind und zumindest Basis-checks für alle Inputs der Funktionen durchzuführen. Achten Sie bei Ihren Outputs darauf, dass diese gut leserlich sind, nicht über den Rand hinausgehen (wenn man die `.Rmd` Datei zu einer PDF kompiliert) und dass Graphiken sinvolle und gut leserliche Beschriftungen und Legenden haben. Sollte dies nicht der Fall sein, kann es zu Punktabzügen kommen. Für high-level Funktionen ist die Signatur meist vorgegeben. Es können aber durchaus kleinere Hilfsfunktionen geschrieben werden, die in den high-level Funktionen aufgerufen werden.

Sie könenn bei diesem Assignment folgende Bonuspunkte sammeln:

(a) Abgabe via Github Classroom (BONUS: 2P)

(b) Rmarkdown Output wohl formatiert (BONUS: 2P)

(c) R Code folgt dem Advanced R Style guide (http://adv-r.had.co.nz/Style.html) (BONUS: 1P)

Aufgabe 1	20 Punkte

Am Anfang vieler Datenanalysen steht die Vorverarbeitung der Daten und deskriptive Analysen. Um sich in Zukunft Arbeit zu sparen, wollen Sie nun zwei Funktionen schreiben, die dabei hilfreich sind:

(a) Eine Funktion zur Imputation fehlender Werte. Genauer sollte die Funktion einen Datensatz übernehmen, in dem womöglich für einzelne Zeilen und Spalten fehlende Werte (`NA`s) vorkommen. Schreiben Sie eine Funktion, die den Datensatz als Input hat und in allen Spalten nach fehlenden Werten sucht. Wenn fehlende Werte vorhanden sind, sollen die fehlenden Werte mit dem Median der Variable (berechnet aus allen Datenpunkten, die nicht NA sind) überschrieben werden, wenn die Variable metrisch ist, und mit dem Modus, wenn die Variable kategorial ist (hierzu zählen Variablen des Typs `factor` und `character` aber auch binäre Variablen (0/1 oder `TRUE`/`FALSE`)). Der Output der Funktion soll der imputierte Datensatz sein.

```
impute <- function(data) {
  # TODO
}
```

(b) Eine Funktion die den paarweisen Zusammenhang von zwei Variablen beschreibt. Für Paare von metrischen Variablen soll dabei eine Pearson Korrelation berechnet werden. Für Paare von kategorialen Variablen soll eine Kreuztabelle mit relativen Häufigkeiten berechnet werden. Paare von metrischen und kategorialen Variablen können hier ignoriert werden. Der Output der Funktion soll eine 2-elementige Liste sein. Das erste Element soll die paarweisen Korrelationen der metrischen Variablen als Matrix enthalten, das zweite eine Liste der Kreuztabellen.
Hinweis: Die Funktion `combn` könnte hier hilfreich sein.

```
pairwise_correlation <- function(data) {
  # TODO
}
```

(c) Der folgende Code lädt den `patient` Datensatz aus dem Paket `pammtools`:

```
data("patient", package = "pammtools")
```

Machen Sie sich mit dem Datensatz vertraut. Nutzen Sie anschließend Ihre Funktionen um zunächst fehlende Werte zu imputieren und anschließend paarweise Korrelationen und Kreuztabellen zu berechnen (die Variablen `CombinedicuID` und `CombinedID` sollen dabei ignoriert werden).

(d) Wie schätzen Sie anhand Ihrer Berechnungen den Zusammenhang zwischen dem Alter der Patienten (`Age`) und der Zeit bis zur Entlassung aus dem Krankenhaus (`survhosp`) sowie der Diagnose (`DiagID2`) und ob der Patient verstorben ist (`PatientDied`) ein.

Aufgabe 2	25 Punkte

In der Übung zu dieser Lehrveranstaltung haben Sie bereits eine Funktion geschrieben, die eine "optimale" Gerade für den Zusammenhang von zwei metrischen Variablen (x und y) schätzen kann (`line_estimator`). Diese Funktion ist als R-File in der Angabe enthalten.

Abbildung 1 zeigt Daten die bei einem "Crash Test" erhoben worden sind, bei dem ein Motorrad Unfall simuliert worden ist. Konkret, wurde ein Crash-Test-Dummy auf einem Motorrad gegen eine Wand gefahren und die Beschleunigung des Kopfs (y; in $g = 9.81\frac{m}{s^2}$) seit dem Zeitpunkt des Aufpralls (x; in Millisekunden) gemessen worden (negative Beschleunigung bedeutet Verlangsamung). In Abbildung 1 ist zusätzlich die "optimale" Gerade eingezeichnet. Offensichtlich, beschreibt diese Gerade die Daten (also den Zusammenhang zwischen x und y) nicht besonders gut.

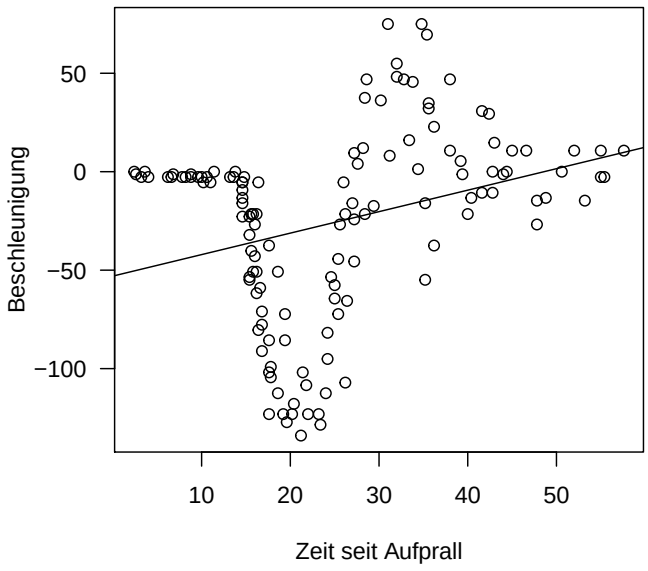

Figure 1: Zusammenhang zwischen Zeit seit Aufprall und Beschleunigung

.

(a) Schreiben Sie zunächst eine Funktion, die den Wertebereich von x in Abschnitte unterteilt und in jedem Abschnitt den `line_estimator` anwendet. Verwenden Sie hier unter anderem die Funktion `cut`. Nennen Sie die Funktion `piece_wise_line_estimator`. Die Funktion sollte dieselben Inputs haben, wie die Funktion `line_estimator`; zusätzlich sollten noch der `cut` Funktion das `breaks` Argument übergeben werden (also entweder die Anzahl Bruchpunkte oder die Bruchpunkte selbst). Der Output soll eine Liste mit dem Output des `line_estimator` in jedem Abschnitt sein.

```
source("line_estimator.R")
piece_wise_line_estimator <- function(data, x_name, y_name,
                                      breaks = NULL) {
  # TODO
}
```

(b) Schreiben Sie darauf aufbauend eine weitere Funktion, die diese Berechnung visualisiert. Die Funktion soll demnach ein Streudiagramm von x und y zeigen, sowie die berechnete Gerade im jeweiligen Abschnitt. Machen Sie zusätzlich die Bruchpunkte durch gestrichelte, vertikale Linien kenntlich.

```
plot_piecewise_line <- function(data, x_name, y_name, piecewise, ...) {
  # TODO
}
```

(c) Wählen Sie nun anhand von Abbildung 1 konkrete Bruchpunkte an denen Sie die Berechnung einer neuen Gerade für sinnvoll halten und nutzen Sie Ihre Funktion um stückweise Geraden zu berechnen.

(d) Visualisieren Sie ihre Schätzung.

(e) Schreiben Sie eine Funktion, die aus dem Output von `piece_wise_line_estimator` den Gesamtfehler der stückweisen Geraden berechnet. In jedem Abschnitt, ist der Fehler dabei wieder als die quadratische Differenz zwischen der Geraden und den y aus den Daten definiert (im Output von `line_estimator` enthalten). Der Gesamtfehler ist die Summe der abschnittsweisen Fehler. Ist der Gesamtfehler der stückweisen Geraden kleiner als der globalen Gerade?

```
total_error <- function(piecewise) {
  # TODO
}
```

(f) *Bonus*: Gehen Sie von genau 3 (inneren) Bruchpunkten aus (4 Abschnitte) aus. Schreiben Sie eine Funktion, die für einen gegebenen Datensatz die optimale Platzierung der Bruchpunkte findet, sodass keine andere Anordnung der Bruchpunkte zu einem kleineren Fehler führt.

Aufgabe 3	30 Punkte

In dieser Aufgabe werden Sie eine Portfolio Analyse durchführen.

Exchange trade funds (ETFs) existieren seit dem frühen Zweitausendern und sind seitdem zu einer der bedeutendsten Anlageformen für private und institutionelle Anleger geworden. Die Idee eines ETFs ist es einen bestimmten Markt oder Index (z.B. DAX) nachzubilden. Der Vorteil für den Anleger ist es, dass man prozentual an der Entwicklung des Index partizipieren kann, ohne tatsächlich von jedem Wert (z.B. einem Unternehmen) aus dem der Index besteht eine Aktie kaufen zu müssen, was teuer werden kann. Stattdessen können Anleger in ETFs auch kleine Beträge investieren (z.B. über Wertpapiersparpläne). Der Aussteller des ETFs sammelt alle diese (Klein)Investitionen und nutzt das Gesamtvolumen um den Zielindex zu replizieren. Der Anleger partizipiert dann prozentual an der Entwicklung des ETFs.

Der bedeutendste Index, der MSCI World, umfasst mehr als 1500 Unternehmen aus über 20 Ländern (er bildet mehr oder weniger das Wachstum der Weltwirtschaft ab). ETFs, denen der MSCI World zugrunde liegt, versuchen, diesen durch eigene Investments in die Unternehmen im Index zu replizieren.

Während MSCI World ETFs die bekanntesten sind, gibt es mittlerweile viele ETFs mit verschiedenen Ausrichtungen und Spezialisierungen.

Hobbyinvestor P.K. ist ebenfalls von den Vorteilen von ETFs überzeugt und hat folgende ETFs des Anbieters iShares gekauft:

- 200 Anteile MSCI World SRI: ein ETF, der breitgefächert in Industrieländern in Unternehmen, die SRI-Kriterien (Socially responsible investment) Anforderungen entsprechen (also eine Mindestanforderung an Nachhaltigkeit erfüllen).
- 180 Anteile MSCI EM SRI: ein ETF, der breitgefächert in Emerging Markets (=EM) (Schwellenländer) investiert. Alle Investment erfüllen auch hier die SRI-Kriterien
- 170 Anteile MSCI World Small Cap: ein ETF, der breitgefächert in Industrieländern in Unternehmen mit kleiner Marktkapitalisierung investiert.
- 15 Anteile MSCI Europe SRI: ein ETF, der in SRI-konforme Unternehmen in Europa investiert.
- 12 Anteile Core MSCI World: ein ETF, der breitgefächert in Unternehmen in Industrieländern investiert.
- 17 Anteile STOXX Europe Small 200: Ein ETF, der in kleine Unternehmen (Marktkapitalisierung) in Europa investiert.
- 14 Anteile MSCI EM IMI: Ein ETF, der in Unternehmen in Schwellenländern investiert.
- 11 Anteile Global Clean Energy: Ein ETF, der in erneuerbare Energien weltweit investiert.

Diese Daten, sowie der jeweilige Kaufkurs pro Anteil (in Euro) kann dem Datensatz `etf_overview` (siehe Datei `etf-overview.Rds`) entnommen werden.

```
etf_overview <- readRDS("etf-overview.Rds")
etf_overview
```

```
##   ETF_ID                   ETF Anteile  Kurs
## 1      1         MSCI World SRI     200  8.53
```

```
## 2      2            MSCI EM SRI     180  7.81
## 3      3   MSCI World Small Cap     170  6.21
## 4      4        MSCI Europe SRI      15 59.12
## 5      5        Core MSCI World      12 70.08
## 6      6 STOXX Europe Small 200      17 36.87
## 7      7            MSCI EM IMI      14 32.48
## 8      8    Global Clean Energy      11 11.33
```

In der folgenden Aufgabe sollen Sie das Portfolio von P.K. analysieren. Die Aufgaben sollen mit Hilfe der Pakete `readr`, `dplyr`, `tidyr` und `ggplot2` bearbeitet werden.

(a) Ergänzen sie Datensatz `etf_overview` um eine neue Spalte `Investment`, welche das Gesamtinvestment pro ETF in Euro angibt. Visualisieren Sie den investierten Betrag pro ETF mit einem Balkendiagramm. Die Balken sollten vom größten zum kleinsten Investment sortiert sein

(b) Die zip Datei `ETFs.zip` enthält Daten zur Zusammensetzung der einzelnen ETFs (d.h. Informationen zu den einzelnen Unternehmen und mit welcher Gewichtung diese in den ETF einfließen). Für jeden der in `etf_overview` enthaltenen ETFs liegt eine `.csv` Datei in diesem zip Ordner. Entpacken Sie diese Dateien in einen Ordner mit dem Namen `data`. Dieser `data` Ordner soll im selben Verzeichnis liegen, wie ihre `.Rmd` Datei (wenn Sie GitHub Classroom nutzen, ist der `data` Ordner bereits im Repository vorhanden).

Nutzen Sie nun das `readr` Paket, um die einzelnen `.csv` Dateien in `R` einzulesen und fügen Sie jedem Datensatz die ETF ID hinzu die im `etf_overview` Datensatz gegeben ist. Schreiben Sie hierfür zunächst eine Funktion `import_etf` mit zwei Argumenten (`path`: dem Pfad zur Datei; und `id`: Die ID des ETF wie er in der Datei `etf_overview` angegeben ist). Die Funktion soll die in `path` spezifizierte Datei einlesen (machen Sie sich hierfür mit der Funktion `read_delim` vertraut) und eine neue Spalte `ETF` anlegen, mit der ID (`ETF_ID`) des ETF als Elementen. Am Ende soll der eingelesene und modifizierte Datensatz zurückgegeben werden. Für den Datensatz `CoreMSCIWorld.csv` sollte dies z.B. wie folgt aussehen:

```
library(readr)
core_msci_world <- import_etf(
  path = "data/CoreMSCIWorld.csv",
  id = 5)
core_msci_world %>%
  select(ETF_ID, ISIN, Name, Kurs)
```

```
## # A tibble: 1,616 x 4
##    ETF_ID ISIN         Name                 Kurs
##     <dbl> <chr>        <chr>               <dbl>
##  1      5 US0378331005 APPLE INC            136.
##  2      5 US5949181045 MICROSOFT CORP       271.
##  3      5 US0231351067 AMAZON COM INC      3448.
##  4      5 US30303M1027 FACEBOOK CLASS A INC  352.
##  5      5 US02079K1079 ALPHABET INC CLASS C 2520.
##  6      5 US02079K3059 ALPHABET INC CLASS A 2445.
##  7      5 US88160R1014 TESLA INC            681.
```

```
##  8      5 US67066G1040 NVIDIA CORP           801.
##  9      5 US46625H1005 JPMORGAN CHASE & CO   154.
## 10      5 US4781601046 JOHNSON & JOHNSON     164.
## # ... with 1,606 more rows
```

(c) Verwenden Sie die Funktion aus der vorherigen Aufgabe in Kombination der Funktion `list.files` und der `map*` Familie von Funktionen aus dem `purrr` package um die Daten aller ETFs einzulesen und zu einem Datensatz zusammenzufügen.

(d) Modifizieren Sie, wenn nötig, Spaltennamen von `etf_data`, sodass die Spaltennamen keine Leer- oder Sonderzeichen, sowie Umlaute enthalten.

(e) Legen Sie eine neue Spalte im Datensatz `etf_data` an, die angibt, wie viel Euro P.K. in die einzelnen Unternehmen investiert hat. Nutzen Sie zur Berechnungen die Informationen aus den Datensätzen `etf_overview` und `etf_data`.

(f) Ersetzen Sie mögliche `NA`s in der Spalte `Investment` mit einer 0.

(g) Bestimmen Sie, aus wie vielen Einzelwerten (ISIN) jedes ETF zusammengesetzt ist.

(h) Entfernen Sie alle Beobachtungen bei denen ISIN keine 12-stellige Zeichenfolge ist. Überprüfen Sie nun mit Hilfe der Spalte `ISIN` ob einzelne Unternehmen mehrfach vorkommen. Wenn ja, berechnen Sie pro Aktie (definiert durch die ISIN) die Gesamtinvestition pro Aktie.

(i) Sie interessieren sich dafür, welche Unternehmen wie stark in die einzelnen ETFs einfließen. Berechnen Sie hierfür pro ETF ID den Anteil des Investments an den einzelnen Aktien relativ zum Gesamtvolumen des jeweiligen ETF. Extrahieren Sie pro ETF die Top 3 Investments. Welches Land (`Standort`) ist in diesen Daten am häufigsten vertreten? Welcher Sektor (`Sektor`) ist am häufigsten Vertreten.

(j) Aggregieren Sie `etf_data` auf Länderebene (`Standort`) und Sektorebene (`Sektor`) bzgl. des Investments pro Unternehmen und visualisieren Sie dann das Investmentvolumen nach Standort mit einem Balkendiagramm. Jeder Sektor soll dabei ein eigenes Diagramm bekommen. Achten Sie darauf, dass der Plot in jedem Fall gut leserlich ist. Fassen Sie hierzu alle bis auf die zehn größten Länder/Standorte als "Sonstige" zusammen und entfernen Sie diese.

Aufgabe 4 — 5 Punkte

Lesen Sie in R das Objekt `umfragen.Rds` ein. Bei dem Objekt handelt es sich um einen `data.frame`, der durch eine Befragung von Medizinern entstanden ist, bei der einflussreiche Faktoren für Herzinfarkte bestimmt werden sollten. Die Spalten stellen die einzelnen Einflussfaktoren dar und die Zeilen die Anzahl der Faktoren, die die Mediziner nennen durften. Die jeweilige Zahl in einer Zelle bezieht sich auf den Anteil (zwischen 0 und 1) der Mediziner, die den jeweiligen Einflussfaktor genannt haben. D.h. der Eintrag in Zeile 2, Spalte 4 kann so interpretiert werden: Wenn die Mediziner nur zwei wichtige Einflussfaktor nennen durften, haben 6.49 Prozent `age` gewählt.

Replizieren Sie basierend auf diesen Daten Abbildung 2.

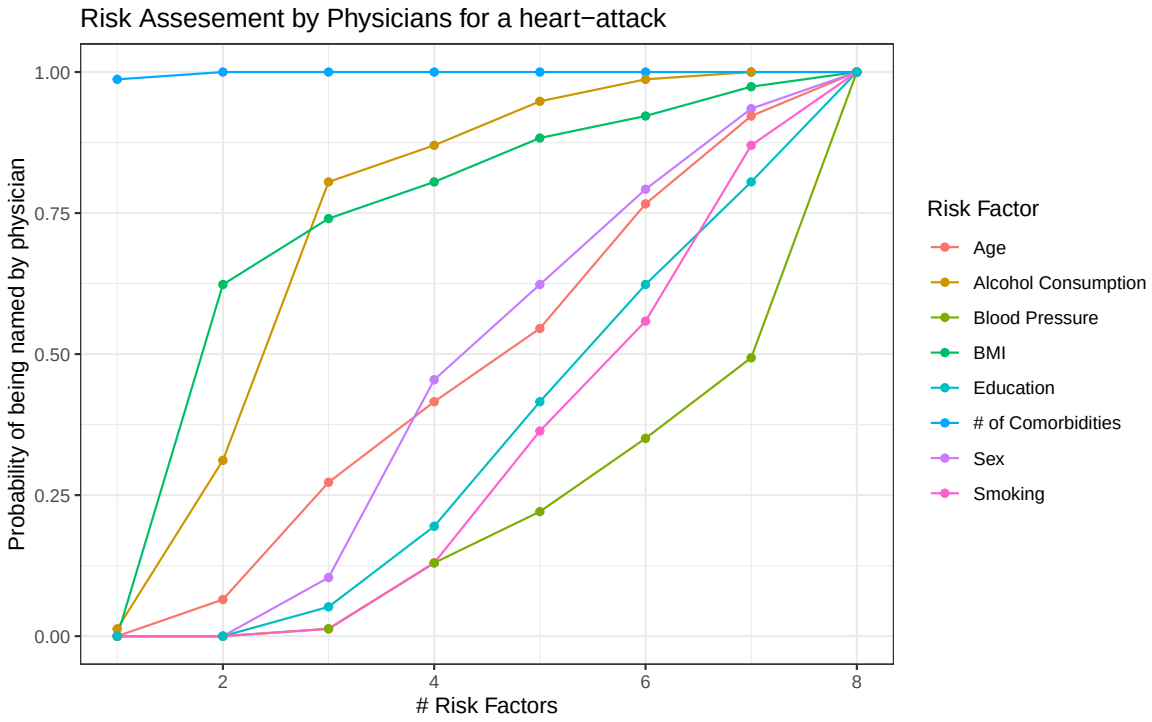

Figure 2: Risk Assesement by Physicians for heartattack.