HOCHSCHULE
WEIHENSTEPHAN-TRIESDORF
UNIVERSITY OF APPLIED SCIENCES

STUDIENGANG UMWELTSICHERUNG

STATISTISCHE AUSWERTUNG ÖKOLOGISCHER DATENSÄTZE

WAHLPFLICHTMODUL

PROF. DR. MICHAEL RUDNER

ANLEITUNG ZUM ARBEITEN MIT DEM R-COMMANDER

TRIESDORF IM OKTOBER 2013

Benötigt: R mit verschiedenen Erweiterungspaketen v.a. Rcmdr, das wiederum von einigen anderen Paketen abhängt. Am besten erst R installieren (www.r-project.org) und dann von R aus über das Menü das Paket Rcmdr direkt installieren. Dann werden alle Abhängigkeiten berücksichtigt.

Das Statistikprogramm R

Das Statistikprogramm R ist ein unter der GPL2 für verschiedene Betriebssysteme frei verfügbares Programm. Das Programm ist befehlszeilenorientiert, d.h., man muss Befehle eingeben. Es gibt dazu aber auch verschiedene grafische Benutzeroberflächen (GUI), die Menüs zur Bedienung anbieten. Eine solche GUI ist der R-Commander. Er wird von R aus gestartet, indem z.B. über das Menü das Paket **`Rcmdr`** geladen wird. Der R-Commander hat ein umfangreiches Menü, das es erlaubt, Daten bequem einzulesen, aus den eingelesenen Daten einen Datensatz auszuwählen und eine ganze Reihe von Analysen auszuführen und Grafiken dazu zu erzeugen.

Starten des R-Commander

Starten Sie R über das Startmenü in Windows. (Eigener Computer)

In den EDV-Räumen der HSWT starten Sie in der Übersicht der netzbasierten Anwendungen (Novell-Netware-Fenster) im Ordner Mathe, Statistik das Programm R. Im R-Fenster wählen sie im Menü **`Pakete/Lade Paket`** aus. Im Dialog wählen Sie **`Rcmdr`** aus und bestätigen über OK.

Der R-Commander hat ein Menü, eine Zeile zur aktiven Datenmatrix und dem aktiven Modell, ein Skriptfenster, in welchem die Befehlszeilen für die Analyse stehen – diese werden von den Menüaufrufen erzeugt, können aber auch direkt eingegeben werden – und ein Ausgabefenster, das der Konsole von R entspricht, in der die aufgerufenen Befehle rot und die Ausgaben blau erscheinen. dazu gibt es unten ein Fenster mit Meldungen, z.B. Fehlermeldungen.

Für die Grafikausgabe nutzt der R-Commander das Grafikpanel von R. Es kann also sein, dass man zur RGui wechseln muss, um die erzeugten Grafiken betrachten zu können.

Arbeitsverzeichnis festlegen

Sie können innerhalb einer R-Sitzung ein Arbeitsverzeichnis festlegen. Zum Daten Einlesen und Speichern greift R dann zunächst auf dieses Verzeichnis zu bzw. bietet dieses im Dialogfenster an.

Rufen Sie über den Menüpunkt **`Datei / Change working directory…`** einen Dialog auf, in welchem Sie Ihr Arbeitsverzeichnis auswählen. Bestätigen Sie mit einem Klick auf die Schaltfläche **`OK`**. Im Skriptfenster des R-Commander erscheint der Befehl **`setwd(„Zielverzeichnis“)`**, der dann auch im Ausgabefenster als ausgeführter Befehl angezeigt wird.

Daten einladen

Im R-Commander die Daten importieren über den Menüpunkt **`Datenmanagement / Importiere Daten / aus Excel- Access- oder dBase-Dateien`**. Wichtig: Das Tabellenblatt auswählen, sonst wird nichts importiert. Geben Sie im Dialogfenster einen Namen für den Datensatz ein. Beachten Sie bitte, dass Leerzeichen, Umlaute und Sonderzeichen zu vermeiden sind.

Prüfen Sie bitte, ob der richtige Datensatz aktiv ist.

Sie können auf diesem Weg die Dateien im Excel-Format öffnen, ohne vorher ins CSV-Format exportieren zu müssen (wie das sonst in R meist der Fall ist).

Mit Klick auf die Schaltfläche **`Datenmatrix betrachten`** können sie überprüfen, ob der gewünschte Datensatz richtig importiert wurde.

Datenanalyse

Deskriptive Statistik
Über den Menüpunkt **Statistik / Deskriptive Statistik / Aktive Datenmatrix** können Sie die Kennwerte zur Verteilung der Variablen im Datensatz aufrufen: Minimum, Maximum, 1. und 3. Quartil, Median und arithmetischer Mittelwert. Die Ergebnisse werden im Ausgabefenster angezeigt. Dieser Aufruf entspricht dem Befehl **summary(Datensatz)** im Skriptfenster.

Korrelationsanalyse
Die Korrelationsanalyse können Sie über den Menüpunkt **Statistik / Deskriptive Statistik / Korrelation** oder **Statistik / Deskriptive Statistik / Test auf Signifikanz der Korrelation** aufrufen. Im Dialogfenster können Sie die Variablen auswählen, die Sie einbeziehen wollen. Hinzufügen zur Auswahl mit gedrückter STRG-Taste. Wählen Sie das Korrelationsmaß aus. Auch partielle Korrelation kann direkt hier berechnet werden. Hierzu müssen Sie wenigstens 3 Variablen auswählen.

Falls Sie die partielle Korrelation schrittweise berechnen wollen, können Sie eine Regression rechnen und anschließend mit den Residuen weiterarbeiten. Die Residuen erhalten Sie mit der Funktion **resid()**. In die Klammer müssen sie die Bezeichnung Ihres Regressionsmodells eingeben. Das Ergebnis der Funktion müssen Sie einem Objekt zuweisen [z.B. **N.resid <- resid(RegMod.1)**].

Dann können Sie über die Funktion **cbind()** die beiden Residuenreihen zusammenfügen, als Datensatz auswählen und die Korrelation berechnen

[z.B. **dat.resid <- cbind(N.resid, P.resid)**].

Die Ergebnisse werden im Ausgabefenster angezeigt.

Mittelwertvergleich
Für die Mittelwertvergleiche stehen Ihnen im R-Commander mehre Funktionen zur Verfügung. Den t-Test können Sie für eine Stichprobe ausführen (**Statistik / Mittelwerte vergleichen / t-Test für eine Stichprobe**) oder für 2 unabhängige Stichproben (**… / t-Test für unabhängige Stichproben**). Hierfür benötigen Sie einen Datensatz mit einer Spalte, die nur 2 Gruppen bezeichnet. Sie können den t-Test auch für verbundene Stichproben durchführen (**… / t-Test für gepaarte Stichproben**). In diesem Fall sollten die beiden Stichproben in einem Datensatz in zwei Spalten stehen mit den gepaarten werten jeweils in der gleichen Zeile.

Es kann auch der Wilcoxon-Test angewandt werden. Sie finden die entsprechenden Funktionasufrufe unter **Statistik / Nichtparametrische** Tests.

Lineare Regression

Einfache lineare Regression
Die Regressionsanalyse aufrufen über den Menüpunkt **Statistik / Regressionsmodelle / Lineare Regression.** Die abhängige und die unabhängige Variable auswählen. Analyse starten. Ergebnis aus dem Ausgabefenster auslesen.

Das Diagramm können Sie über **Grafik / Streudiagramm** erzeugen. Die unabhängige Variable sollte auf der x-Achse zu liegen kommen. Sie können dazu die Kleinst-Quadrat-Linie auswählen, das ist die Regressionskurve.

Berechnen Sie die Regressionskoeffizienten für das Konfidenzintervall über den Menüpunkt **Modelle / Konfidenzintervalle**. Stellen Sie eine Reihe von Datenpunkten entlang der x-Achse auf, für die Sie die y-Werte des Konfidenzintervalls berechnen wollen, auf. [z.B. **xWerte <- seq(0.2, 1.0, length=50)**] Beachten Sie bitte, dass innerhalb von R das Dezimaltrennzeichen immer ein Punkt ist!

Tippen Sie die Anweisungen in die letzte Zeile des Skriptfensters und klicken Sie anschließend die Schaltfläche **Befehl ausführen**.

Berechnen Sie die untere Wertereihe und die obere Wertereihe und verwenden Sie dabei die erhaltenen Regressionskoeffizienten und die erzeugte Reihe von x-Werten. [**`konf.min <- Intercept(2.5%) + Regr.koeffizient(2.5%) * xWerte`**, ,

etwa: **`konf.min <- 0.44 + 3.85*xWerte`**; entsprechend auch für die obere Grenze]

Zeichnen Sie die Linien in das vorhandene Diagramm ein. [z.B. **`lines(xwerte,konf.min, lty=2, col="blue"`**) und entsprechend für die obere Grenze)].

Speichern Sie die Abbildung über **`Grafiken/Speichere Abbildung in Datei`**.

Multiple lineare Regression
Die Berechnung erfolgt wie bei der einfachen linearen Regression. Sie müssen im Dialogfenster mehrere unabhängige Variablen auswählen (Mehrfachauswahl bei gedrückter STRG-Taste).

Sie können eine automatisierte Variablenauswahl vornehmen, die sich nach dem Akaike-Informationskriterium AIC richtet (oder nach BIC). Wählen Sie dazu den Menüpunkt **`Modelle / Stepwise Model Selection…`**. Achten sie bitte darauf, dass das richtige Modell aktiv ist.

Sie können eine schrittweise Variablenauswahl vornehmen lassen. Es wird allgemein empfohlen, mit dem Modell zu beginnen, das alle sinnvollen Variablen enthält, die untereinander nicht zu stark korreliert sind. Dann sollte man schrittweise rückwärts oder rückwärts/vorwärts die Auswahl treffen lassen.

Logistische Regression
Wählen Sie den Menüpunkt **`Statistik/Regressionsmodelle/Generalisiertes lineares Modell…`** . Im Dialogfenster stellen Sie die Regressionsformel zusammen: Eine abhängige Variable (mit 0 oder 1 als möglichen Werten) und mehrere unabhängige Variablen. Die unabhängigen Variablen werden in der Formelschreibweise mit "+" verknüpft. Falls Sei einen Interaktionsterm mit einbeziehen möchten, müssen Sie zwischen den entsprechenden Variablen das "+" durch "*" ersetzen. Darunter können Sie ggf. eine Auswahl für die Probepunkte treffen. Für die logistische Regression stellen Sie bei Family bitte "binomial" ein. Die passende Link-.Funktion ist der Logit.

Grafische Analyse eines Regressionsmodells
Ihr fertiges Modell sollten Sie grafisch analysieren. Hierzu können Sie die diagnostischen Plots über den Menüpunkt **`Modelle / Grafiken / grundlegende diagnostische Grafiken`** erzeugen.

Beenden des R-Commanders

Der R-Commander wird auf die in Windows übliche Weise geschlossen. Dabei werden Sie gefragt, ob Sie das Skript speichern möchten. Das Skript enthält alle ausgeführten Befehle und kann sehr nützlich sein. Mit leichten Modifikationen kann man damit auch andere Datensätze in gleicher Weise analysieren.

Sie werden weiterhin gefragt, ob Sie den Inhalt des Ausgabefensters speichern möchten. das sind Ihre Ergebnisse. Das Speichern wird empfohlen.

Die Grafiken sollten Sie jeweils nach der Erstellung direkt speichern.

Anschließend sollten Sie R schließen. Sie werden gefragt, ob Sie den Workspace sichern möchten. Falls Sie zu einem späteren Zeitpunkt an dieser Stelle weiterarbeiten möchten, ohne alle Analyseschritte erneut auszuführen sollten Sie den Workspace speichern. Nach dem Laden des Workspace (geht über das Menü von R oder R-Commander) sind alle Objekte (Datensätze und Ergebnisse) wieder verfügbar.