Ökonometrie: Wie Ökonomen an statistische Analysen herangehen

1

In diesem Kapitel

- Ziele der ökonometrischen Analyse entdecken
- Herangehensweise und Methodik der Ökonometrie verstehen
- Sich mit ökonometrischer Software vertraut machen

Willkommen zum Studium der Ökonometrie! Die 1930 gegründete *Econometric Society* (Ökonometrische Gesellschaft) definiert die Ökonometrie als ein Gebiet, das auf der »theoretisch-quantitativen und empirisch-quantitativen Herangehensweise an ökonomische Probleme« basiert. Dieser Zungenbrecher bedeutet, dass Ökonometriker – manchmal – Mathematiker sind und komplexe Algorithmen sowie analytische Tools verwenden, um verschiedene Schätzungen abzuleiten und Verfahren zu testen. In anderen Fällen sind Ökonometriker als angewandte Ökonomen tätig und verwenden Tools, die von anderen Ökonometrikern entwickelt wurden, um ökonomische Phänomene zu untersuchen.

In diesem Kapitel werden Sie sehen, dass eine charakteristische Aufgabe der Ökonometrie aus der Entwicklung von Techniken zur Analyse von Daten besteht, die nicht aus kontrollierten Experimenten stammen und folglich gegen viele standardmäßige statistische Annahmen verstoßen. Sie werden auch zu verstehen beginnen, dass gute quantitative Ergebnisse wesentlich von zuverlässigen und geeigneten Daten sowie einer vernünftigen ökonomischen Theorie abhängen.

Und da Computer und ökonometrische Software heutzutage allgemein bei einführenden ökonometrischen Lehrveranstaltungen verwendet werden, widmen auch wir einen Teil dieses Kapitels grundlegenden Anweisungen in R (Version 3.1.2), einem gängigen Statistikprogramm. Diese Software ermöglicht Ihnen, theoretische Konzepte unmittelbar anzuwenden und Ihr Verständnis der Daten zu verbessern.

Ökonomische Beziehungen auswerten

Die Ökonomie liefert die theoretischen Methoden, die Sie zum Auswerten ökonomischer Beziehungen verwenden, und trifft qualitative Vorhersagen zu ökonomischen Phänomenen unter Verwendung der *Ceteris-Paribus-Annahme*. Vielleicht erinnern Sie sich aus Ihren Vorlesungen daran, dass die Ceteris-Paribus-Annahme bedeutet, dass alle anderen Bedingungen unverändert bleiben. Zwei Beispiele unter zahllosen Möglichkeiten sind:

- ✔ In der mikroökonomischen Theorie würden Sie erwarten, dass ökonomische Profite in einem Wettbewerbsmarkt mehr Firmen dazu bewegen, in diesen Markt einzusteigen, *ceteris paribus*.

✔ In der makroökonomischen Theorie würden Sie erwarten, dass höhere Zinssätze Investitionen verringern, *ceteris paribus*.

Ökonometrie unterstützt die ökonomische Theorie, indem sie die nötigen Werkzeuge liefert, um qualitative Aussagen zu quantifizieren, die Sie (oder andere) unter Verwendung der Theorie machen. Unbekannte oder vermutete Beziehungen aus der abstrakten Theorie können durch Verwendung realer Daten und von Ökonometrikern entwickelten Methoden quantifiziert werden.

Der folgende Abschnitt erklärt, wie mithilfe der Ökonometrie die Zukunft prognostiziert und ökonomische Phänomene quantitativ dargestellt werden können. Anschließend lesen Sie, warum es für Ökonometriker so wichtig ist, stets sinnvolle Annahmen zu treffen.

Mittels ökonomischer Theorie Zusammenhänge beschreiben und Vorhersagen treffen

Eine der Eigenschaften, die die angewandte Forschung in der Ökonometrie von anderen Vorgehensweisen bei der statistischen Analyse unterscheidet, ist das Theoriegebäude, das die empirische Arbeit unterstützt.

Ökonometrie kommt in der Regel zum Einsatz, um zu erklären, wie Faktoren ein bestimmtes Ergebnis beeinflussen, oder um zukünftige Ereignisse vorherzusagen. Unabhängig davon, was Ihr Hauptziel ist, muss Ihre ökonometrische Analyse auf einem ökonomischen Modell beruhen. Ihr Modell sollte aus einem interessierenden Ergebnis, der abhängigen Variable Y, und kausalen Faktoren, den unabhängigen Variablen X_1 bis X_n, bestehen, die theoretisch oder logisch mit dem Ergebnis verknüpft sind.

Sinnvolle Annahmen sind der Grundstein

Ein bekannter Witz über Ökonomen geht wie folgt: Ein Physiker, ein Chemiker und ein Ökonom sind auf einer einsamen Insel gestrandet und haben nichts zu essen. Eine Dosensuppe wird angespült. Der Physiker sagt: »Lasst uns die Dose mit einem Stein aufschlagen.« Der Chemiker sagt: »Lasst uns zuerst ein Feuer machen und die Dose erhitzen.« Der Ökonom überlegt kurz und sagt: »Nehmen wir mal an, wir hätten einen Dosenöffner …«. Das ist zwar als Witz gemeint und doch kann es helfen, Annahmen über die Realität zu machen, daraus Schlüsse zu ziehen und Vorhersagen für den Fall abzugeben, dass bestimmte Bedingungen zutreffen. In der Ökonometrie kann es jedoch mitunter gefährlich sein, Annahmen zu treffen, ohne deren Realisierbarkeit zu überprüfen. Da hat der Witz dann doch wieder recht.

Zu viele Annahmen über gegebene Bedingungen, funktionale Form und statistische Eigenschaften zu treffen, kann zu verzerrten Ergebnissen führen und die Genauigkeit der Schätzung untergraben, die Sie durchführen wollen. Obwohl Sie einige Annahmen treffen müssen, um Ihre ökonometrische Arbeit durchzuführen, sollten Sie die meisten Annahmen überprüfen und ehrlich sein, welche potenziellen Auswirkungen die Annahmen, die Sie nicht testen können, auf Ihre Ergebnisse haben können.

Das Überprüfen von Prognosen, die auf ökonomischer Theorie oder logischen Überlegungen beruhen, ist in der Regel alles andere als einfach. Beobachtete Daten stammen eher selten aus einem kontrollierten Experiment, und dadurch ist es schwierig, sicherzustellen, dass die Ceteris-Paribus-Annahme zutrifft. Widmen Sie also den (unabhängigen) Variablen, die Sie in die Analyse integrieren, um die ceteris paribus Situation (so nah wie möglich) zu simulieren, hinreichend Aufmerksamkeit.

Mit statistischen Methoden ökonomische Probleme angehen

Bücher über Ökonometrie gehen gern davon aus, dass Sie genügend Statistik gelernt haben, um ökonometrische Modelle zu erzeugen, Schätzungen vornehmen und Hypothesen testen zu können. Wir haben jedoch festgestellt, dass Studenten es stets schätzen, jene statistischen Konzepte noch einmal durchzugehen, die für den Erfolg in der Ökonometrie am wichtigsten sind. Vor allem sollten Sie mit Wahrscheinlichkeitsverteilungen und Hypothesentests zurechtkommen. (Falls Ihre Kenntnisse in diesen Bereichen etwas eingerostet sind, sollten Sie unbedingt Kapitel 2 und 3 lesen.)

Wie exakt Sie ökonomische Beziehungen quantifizieren können, hängt nicht nur von Ihren Fähigkeiten ab, ökonometrische Modelle zu erzeugen, sondern auch von der Qualität der für die Analyse verwendeten Daten und Ihrem Geschick, geeignete Methoden zum Schätzen von Modellen zu finden, deren Voraussetzungen durch die Daten nicht vollständig erfüllt sind. Die Daten müssen nicht nur aus einem zuverlässigen Erfassungsprozess stammen; Sie sollten außerdem die Augen offen halten hinsichtlich weiterer Einschränkungen oder Erfordernisse. Dazu zählen unter anderem:

- ✔ **Aggregation von Daten:** Eine Information, die möglicherweise aus einem Haushalt, von einem Individuum oder aus einem Unternehmen stammt, wird in Ihren Daten auf der Ebene einer Stadt, eines Landes oder Staates gemessen.
- ✔ **Statistisch korrelierte, aber ökonomisch irrelevante Daten:** Einige Datensätze enthalten zwar eine Fülle von Informationen, aber viele der Variablen haben möglicherweise nichts mit der ökonomischen Frage zu tun, der Sie sich widmen wollen.
- ✔ **Kategoriale oder qualitative Daten:** Ergiebige Datensätze enthalten zwar in der Regel qualitative Variablen (geografische Informationen, ethnischer Hintergrund und so weiter), aber die Informationen bedürfen einer speziellen Behandlung, bevor sie in einem ökonometrischen Modell verwendet werden können.
- ✔ **Verletzung einer dem klassischen linearen Regressionsmodell zugrundeliegenden Annahme:** Die Legitimität Ihrer ökonometrischen Herangehensweise beruht stets auf einem Satz statistischer Annahmen, aber Sie werden sehr wahrscheinlich feststellen, dass mindestens eine dieser Annahmen nicht gilt (für Ihre Daten nicht zutrifft).

Ökonometriker wollen sich von Statistikern unterscheiden, indem sie sich offen zu Verstößen gegen statistische Annahmen bekennen, die sonst oft als gegeben hingenommen werden. Die gängigste Methode des Schätzens eines ökonometrischen Modells ist die gewöhnliche Methode der kleinsten Quadrate (GKQ oder englisch *Ordinary Least Squares, OLS*), die wir in Kapitel 5 behandeln. Wie wir jedoch in Kapitel 6 und 7 sehen werden, muss eine Reihe von Annahmen des klassischen linearen Regressionsmodells für die GKQ-Methode zutreffen, um zuverlässige Schätzwerte zu erhalten. In der Praxis hängen die Annahmen, gegen die am wahrscheinlichsten verstoßen wird, von Ihren konkreten Daten und der konkreten Verwendung ab. In den Kapiteln 10, 11 und 12 lernen Sie, die häufigsten Verstöße gegen Annahmen zu erkennen und damit umzugehen.

In den folgenden Abschnitten beschreiben wir, wie Ihnen die Kenntnis bestimmter Eigenschaften Ihrer Daten helfen kann, bessere ökonometrische Modelle zu erzeugen. Besondere Aufmerksamkeit sollten Sie der Struktur Ihrer Daten widmen, der Art, wie Variablen gemessen werden, und wie quantitative Daten durch qualitative oder kategoriale Informationen vervollständigt werden können.

Die Bedeutung des Datentyps, der Häufigkeit und der Aggregation erkennen

Die Daten, die Sie zum Schätzen und Testen Ihres ökonometrischen Modells verwenden, werden in der Regel in drei mögliche Typen unterteilt (für weitere Details siehe Kapitel 4):

- ✔ **Querschnittsdaten:** Dieser Datentyp besteht aus Messungen zu einzelnen Beobachtungen (für Personen, Haushalte, Unternehmen, Gemeinden, Länder oder was auch immer) zu einem *bestimmten Zeitpunkt*.
- ✔ **Zeitreihen:** Dieser Datentyp besteht aus Messungen von einer oder mehrerer Variablen (wie dem Bruttoinlandsprodukt oder der Arbeitslosenquote) an *verschiedenen Zeitpunkten* in einem definierten Bereich (zum Beispiel ein bestimmtes Bundesland oder Land).
- ✔ **Panel- oder Längsschnittdaten:** Dieser Datentyp besteht aus einer *Zeitreihe für jede Querschnittseinheit* in der Stichprobe. Die Daten enthalten Messungen zu einzelnen Beobachtungen (Personen, Haushalte, Unternehmen, Gemeinden, Bundesländer, Länder und so weiter) über einen Zeitraum (Tage, Monate, Quartale oder Jahre).

Welchen Datentyp Sie verwenden, kann Einfluss darauf haben, wie Sie Ihr ökonometrisches Modell schätzen. Vor allem für den Umgang mit Zeitreihen und Paneldaten sind in der Regel spezielle Methoden erforderlich. Wir befassen uns in Kapitel 12 mit Zeitreihen-Methoden und besprechen Panel-Methoden in den Kapiteln 16 und 17.

Mit der Zeit bekommen Sie ein Gefühl dafür, welchen Schwierigkeiten Sie bei Ihrer Analyse begegnen werden, da für bestimmte Typen von Daten bestimmte Verletzungen bestimmter Annahmen des klassischen linearen Regressionsmodells wahrscheinlicher sind als andere. Zwei typische Fälle von Annahmeverlet-

zungen betreffen Heteroskedastizität (die häufig bei Querschnittsdaten auftritt) sowie Autokorrelation (die gern bei Zeitreihendaten auftritt). Für eine ausführliche Darstellung von Heteroskedastizität und Autokorrelation lesen Sie bitte Kapitel 11 beziehungsweise 12.

Sie sollten nicht nur wissen, mit welchem Datentyp Sie arbeiten, sondern auch stets folgende Informationen haben:

- ✔ **Die beim Messen der Variablen verwendete *Aggregation*sebene:** Die Aggregationsebene bezieht sich auf die Einheit (Zusammenfassung der Einzelgrößen), in der die Daten gewonnen werden. Anders ausgedrückt können die Variablenmessungen von einer niedrigeren Aggregationsebene stammen (zum Beispiel einem Individuum, Haushalt oder Unternehmen) oder einer höheren Aggregationsstufe (wie einer Stadt, Gemeinde oder einem Land).
- ✔ **Die *Frequenz*, mit der die Daten erfasst werden:** Die Häufigkeit bezieht sich auf die Anzahl der Messungen pro Zeitintervall. Zeitreihendaten können mit höherer Frequenz erfasst werden (wie stündlich, täglich oder wöchentlich) oder mit einer niedrigeren Frequenz (wie monatlich, vierteljährlich oder jährlich).

Selbst mit sämtlichen Daten dieser Welt können Sie keine überzeugenden Ergebnisse hervorbringen, wenn die Aggregationsebene oder die Frequenz für Ihr Problem ungeeignet ist. Wenn Sie zum Beispiel daran interessiert sind, wie die finanziellen Aufwendungen pro Schüler die akademischen Leistungen beeinflussen, werden Daten auf Bundesebene vermutlich ungeeignet sein, da Aufwendungen und Schülereigenschaften von Land zu Land so stark schwanken, dass Ihre Ergebnisse vermutlich irreführend sein werden.

Tappen Sie nicht in die Data-Mining-Falle

Je mehr Werkzeuge zur Datenanalyse Sie beherrschen, desto eher könnten Sie versucht sein, Ihren Datenbestand ohne Modellvorgaben nach Beziehungen zwischen den einzelnen Variablen zu durchsuchen. Mit ausreichend Statistikwissen würden Sie so zweifellos Modelle finden, die Ihre Daten ziemlich gut beschreiben. Diese Praxis wird als *Data-Mining* (bei dem aus dem Datenberg etwas Wertvolles herausgezogen werden soll) bezeichnet. Aus Sicht der Ökonometriker ist dies jedoch unzulässig, da am Anfang der Analyse immer ein ökonomisches Modell stehen soll – und keines, welches sich erst nachträglich aus den Daten formt.

Data-Mining kann sehr sinnvoll sein in Bereichen, in denen es keine Rolle spielt, welche zugrunde liegenden Mechanismen die Ergebnisse generieren. Für Ökonomen ist diese Herangehensweise jedoch bedenklich. In der Ökonometrie ist es viel wichtiger, ein Modell zu erstellen, das Sinn ergibt und von Kollegen nachvollziehbar ist, als nach einem Modell zu suchen, das perfekt angepasst ist. Die Bedeutung vernünftiger Modelle wird in Kapitel 4 dargelegt und konkrete Beispiele für gängige ökonomische Modelle liefert Kapitel 8.

Quantitative und qualitative Informationen einbeziehen

Ökonomische Ergebnisse können sowohl durch quantitative (intervall- oder verhältnisskalierte, also kontinuierliche) als auch qualitative (nominale oder auch kategoriale) Daten beeinflusst werden. Im Allgemeinen ist die Verwendung und Interpretation quantitativer Information in ökonometrischen Modellen einfach.

Qualitative Variable gehen mit Merkmalen einher, die keine natürliche Darstellung als Zahlen haben (z. B. die Augenfarbe). Allerdings können qualitative Eigenschaften von Ausgangsdaten durch einen numerischen Wert repräsentiert werden. Zum Beispiel kann eine geographische Region kodiert sein mit einer 1 für Westen, 2 für Süden, 3 für Osten und 4 für Norden. Die Zuordnung der spezifischen Werte ist jedoch beliebig und hat keine besondere Bedeutung. Um die in den qualitativen Variablen enthaltene Information zu nutzen, werden Sie in vorliegendem Buch in der Regel in *Dummy-Variablen* umgewandelt – dichotome (zweiwertige) Variablen, die den Wert 1 annehmen, wenn eine bestimmte Eigenschaft vorhanden ist und 0, wenn nicht. Wir werden die Verwendung von Dummy-Variablen als unabhängige Variable in Kapitel 9 in einem ökonometrischen Modell veranschaulichen.

Manchmal ist das ökonomische Ergebnis selbst qualitativ oder enthält zensierte (= eingeschränkte) Wertebereiche. Zum Beispiel könnte Ihre abhängige Variable unter Einbeziehung verschiedener Unternehmenseigenschaften als unabhängige Variable messen, ob ein Unternehmen in einem bestimmten Jahr in Konkurs geht oder nicht. Obwohl Standardmethoden bei qualitativen oder diskreten abhängigen Variablen manchmal akzeptabel sind, münden sie für gewöhnlich in Annahmeverstößen und erfordern eine besondere ökonometrische Vorgehensweise. Blättern Sie zu den Kapiteln 13 und 14, um geeignete Methoden für Situationen zu finden, in denen Ihre abhängige Variable nicht-kontinuierlich ist.

Mit ökonometrischer Software arbeiten: Eine Einführung in R

Spezialsoftware macht die Anwendung ökonometrischer Methoden für jeden möglich, auch wenn er kein Programmiergenie ist. Behalten Sie im Hinterkopf, dass mehrere gute Softwarealternativen erhältlich sind und dass Sie, als guter Ökonom, Kosten und Nutzen jeder einzelnen abwägen sollten. Natürlich hängt die Art der Software, mit der Sie schließlich in Ihrem Ökonometrie-Einführungskurs arbeiten werden, davon ab, was Ihr Professor für seine wissenschaftliche Arbeit verwendet oder für den Kurs als sinnvoll erachtet. Roberto stützt sich bei seiner wissenschaftlichen Arbeit hauptsächlich auf STATA und nutzt es in seinen Ökonometrie-Kursen ausschließlich, Karl-Kuno schwört auf R, aber Ihr Professor kann genauso gut EVIEWS, SAS oder ein anderes Programm verwenden.

Gerade für Studenten halten wir die einfache und möglichst kostenlose Verfügbarkeit von Software für ein sehr hohes Gut. Da das Softwarepaket R (www.r-project.org) nicht nur kostenlos ist, sondern auch erstklassige Qualität verspricht und hält, wird es in dieser Ausgabe des Buches ausschließlich verwendet. In Verbindung mit der Software RStudio (www.rstudio.com) bietet R eine ausgezeichnete Kombination aus einem benutzerfreundlichen Interface, konsistenter Struktur in der Syntax sowie einfachen Anweisungen, um sämtliche Metho-

den zu implementieren, die Sie in der Ökonometrie kennenlernen. Außerdem ist es für eine Reihe von Plattformen und Betriebssystemen verfügbar.

R lässt sich mithilfe von Paketen, sogenannten *packages* auf schier unendlich verschiedene Arten erweitern. Siehe zum Beispiel: `http://cran.r-project.org/web/packages/`. Stellen Sie sich bei der Recherche am besten einen Wecker, damit Sie nicht die Zeit vergessen!

Wir werden R als *kommandogesteuertes* Programm verwenden, weil das schnell und einfach ist. Für jede Aufgabe, die R für Sie erledigen soll, geben Sie eine Anweisung in R ein und erhalten nach kurzer Zeit das Resultat unterhalb der Eingabezeile oder im Grafik-Fenster.

Die folgenden Abschnitte zeigen einige R Anweisungen, die Ihnen den Einstieg in das Arbeiten mit dieser Software ermöglichen.

Die Einführung in R ist keinesfalls erschöpfend. Das »Handbuch« bei R besteht aus der Online-Hilfe sowie der Dokumentation auf `www.r-project.org` (Tausende von Seiten). Wir können also unmöglich sämtliche Facetten von R behandeln, die Sie in der Ökonometrie (oder anderswo) möglicherweise anwenden werden. Jedoch ist die Dokumentation verständlich geschrieben und bietet (vor allem) gute Beispiele für das jeweilige Thema. Eine sehr gute Einführung in R bietet das Buch »R für Dummies« von Joris Meys und Andrie de Vries.

Sich mit R vertraut machen

In diesem Abschnitt des Kapitels lesen Sie, wie Sie Datensätze anlegen oder einlesen, Veränderungen vornehmen und Dateien speichern.

Datensätze in R erstellen

Nachdem R gestartet wurde, erwartet es Ihre Anweisungen nach folgendem Zeichen:

```
>
```

Sie können hier zum Beispiel eingeben:

```
> print("Hallo Welt!")
```

nachdem Sie ›Enter‹ gedrückt haben, antwortet R mit

```
[1] "Hallo Welt!"
```

Herzlichen Glückwunsch! R hat Sie verstanden.

Ihre Eingaben erwartet R nach dem >. Danach erwartet es ein ›Enter‹. Dann weiß es, dass es jetzt selbst an der Reihe ist. Vor Ausgaben steht bei R kein >. Daran können Sie sehen, dass es eine Ausgabe ist. Meist fassen wir Ein- und Ausgabe in den Listings zusammen, etwa so:

```
> print("Hallo Leserin!")
[1] "Hallo Leserin!"
```

Jetzt können wir Daten einlesen. Der Dreh- und Angelpunkt für Daten in R ist der sogenannte `data.frame` (deutsch etwa ›Datensatz‹). Um einen solchen aufzubauen, geben Sie folgendes ein (und drücken Sie nach jeder Zeile ›Enter‹):

```
> datensatz <- data.frame()
> datensatz <- edit(datensatz)
```

Hier passiert folgendes: Sie weisen zunächst der neuen Variablen `datensatz` einen leeren `data.frame` zu. Anschließend rufen Sie den Dateneditor mit dem noch leeren `data.frame` namens `datensatz` auf. Nun öffnet sich ein Fenster, in welches Sie Ihre Informationen eingeben können. Das sehen Sie in Abbildung 1.1. In unserem Beispiel haben wir zwei Variablen und drei Beobachtungen pro Variable. Klicken Sie in das Feld mit Namen `var1` und geben Sie in das sich öffnende Fenster ›Stadt‹ ein, den Typ können Sie bei *`character`* belassen. Schließen Sie das Fenster wieder und klicken Sie in das Feld mit Namen `var2`, dem Sie den Namen ›Einwohner‹ geben und den Typ auf *`numeric`* ändern. Anschließend können Sie die Informationen eingeben, etwa so:

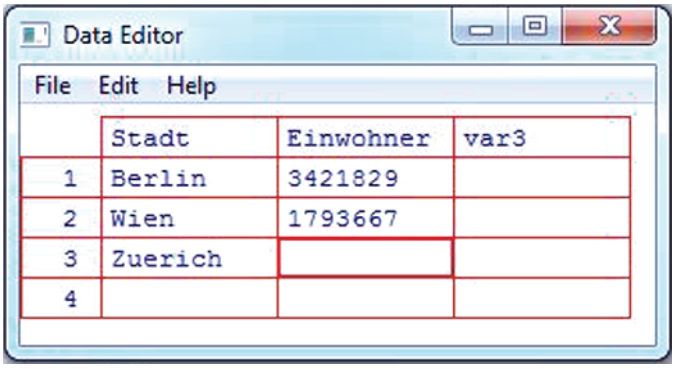

Abbildung 1.1: Der Data Editor zum Editieren von `data.frames`

Jetzt schließen Sie den Dateneditor mit dem roten Kreuz rechts oben. Das Ergebnis wird wieder der Variablen `datensatz` zugewiesen. Was Sie mit dieser anstellen können, wird im nächsten Abschnitt beschrieben.

Daten anzeigen und beschreiben

Den Inhalt einer Variablen zeigen Sie an, indem Sie ihren Namen – gefolgt von ›Enter‹ – eingeben. Etwa so:

```
> datensatz
```

und R antwortet mit:

```
    Stadt Einwohner
1  Berlin   3421829
2    Wien   1793667
3 Zuerich    402275
```

Mithilfe der vier folgenden Anweisungen können Sie sich nun einige Informationen zu Ihrem Datensatz anzeigen lassen:

✔ Mit der Funktion **`head()`** lassen Sie sich die ersten Daten eines (großen) `data.frame` anzeigen; die Zahl nach dem Komma legt fest, wie viele Datensätze ausgegeben werden sollen:

```
> head(datensatz,1)
   Stadt Einwohner
1 Berlin   3421829
```

✔ Mit der Funktion **`tail()`** verhält es sich analog, nur eben von unten – oder hinten: ganz wie Sie wollen:

```
> tail(datensatz,1)
    Stadt Einwohner
3 Zuerich    402275
```

✔ Die Funktion **`str()`** gibt Auskunft über die Struktur eines Objekts :

```
str(datensatz)
'data.frame': 3 obs. of  2 variables:
 $ Stadt    : chr  "Berlin" "Wien" "Zuerich"
 $ Einwohner: num  3421829 1793667 402275
```

Übersetzt bedeutet dies: Es handelt sich um einen `data.frame` mit drei Beobachtungen von zwei Variablen. Die erste heißt `Stadt` und die Beobachtungen sind Zeichenketten. Die ersten Einträge werden angezeigt. Die zweite heißt `Einwohner` und ihre Einträge bestehen aus Zahlen. Auch hier werden die ersten Einträge angezeigt.

✔ Mit **`summary()`** erhalten Sie ein paar Informationen über den Inhalt der Observablen:

```
> summary(datensatz)
    Stadt             Einwohner
 Length:3           Min.   : 402275
 Class :character   1st Qu.:1097971
 Mode  :character   Median :1793667
                    Mean   :1872590
                    3rd Qu.:2607748
                    Max.   :3421829
```

Je nach Typ erhalten Sie einen ersten Überblick über Ihre Daten. Dazu später mehr.

Hilfe zu einer Funktion erhalten Sie mit **`?funktion()`**, also zum Beispiel `?head()` für Hilfe zur Anweisung `head()`.

Datensätze in R speichern und einlesen

Sicher wollen Sie Ihre Daten nicht jedes Mal wieder eingeben, wenn Sie R neu starten. Das müssen Sie auch nicht. Speichern Sie zum Beispiel die Variable `datensatz` einfach mit:

```
> save(datensatz, file = "mein_datensatz.rda")
```

Zum erneuten Einlesen verwenden Sie

```
> load("mein_datensatz.rda")
```

Sie können den Datensatz auch als *csv* – Datei speichern:

```
> write.csv(datensatz, file = "mein_datensatz.csv", row.names = FALSE)
```

Das Argument `row.names` legt fest, ob beim Speichern zusätzliche Zeilennamen vergeben werden sollen. Unsere Zeilen haben jedoch schon Namen, so dass dies nicht nötig ist. Mit `read.csv()` lesen Sie die Daten wieder ein:

```
> read.csv(file = "mein_datensatz.csv")
    Stadt Einwohner
1  Berlin   3421829
2    Wien   1793667
3 Zuerich    402275
```

Achten Sie darauf, die eingelesenen Daten einer Variablen zuzuweisen. Zum Beispiel so:

```
> datensatz_neu <- read.csv(file = "mein_datensatz.csv")
```

Andernfalls werden die Daten nur angezeigt, jedoch keiner Variablen zugewiesen.

Datenquellen und Formate

Es gibt unendlich viele Datenquellen, die fast schon einen eigenen Studiengang ›Datenbeschaffung‹ rechtfertigen würden. Schauen Sie bei Gelegenheit mal auf folgenden Seiten nach:

`http://de.wikipedia.org/wiki/Portal:Statistik/Datensaetze`

`http://ec.europa.eu/economy_finance/db_indicators/ameco/index_en.htm`

`http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home`

Dort finden Sie sicher spannende Daten für eine neue Analyse. Häufig können Sie das Format der Daten vor dem Herunterladen selbst festlegen. Für den Import eignet sich am besten das kommaseparierte Format *(csv)*. Dieses können Sie mit jedem beliebigen Texteditor anzeigen lassen und mithilfe der Anweisung `read.csv()` in einen `data.frame` in R importieren. Die Weiterverarbeitung und Aufbereitung der Daten ist nicht immer einfach. Konsultieren Sie gern *R für Dummies* von Andrie de Vries und Joris Meys für eine prall gefüllte Trickkiste, mit der sie (fast) alle Daten in die gewünschte Form bringen. Natürlich gibt es auch weitere gute Bücher, zum Beispiel *Programmieren mit R* von Uwe Ligges (Springer).

R hat schon in der Basisausführung einige eingebaute Datensätze, mit denen Sie herumexperimentieren können. Hier sparen Sie sich den Aufwand der Datenaufbereitung. Häufig bezieht sich R-Literatur auch auf diese eingebauten Datensätze, wie zum Beispiel den berühmten Datensatz `iris`. Was R alles für Sie bereithält, können Sie sich mit `data()` anzeigen lassen. Informationen zum jeweiligen Datensatz erhalten Sie mit `?datensatz`. Für den Datensatz `longley` wäre dies die Eingabe `?longley`. Dieser begegnet uns später noch ein paar Mal; er ist zwar recht kurz, doch hat er es in sich!

Fehlermeldungen interpretieren

Nicht immer läuft der Ökonometriker-Alltag reibungslos ab und hin und wieder will R nicht so, wie Sie wollen. Immerhin ist R äußerst gesprächig, wenn etwas nicht klappt. Wenn wir zum Beispiel den Namen einer Variablen nicht korrekt eingeben, reagiert R so:

```
> datensat
Error: object 'datensat' not found
```

Da tappt man tatsächlich nicht lange im Dunkeln. Anhand der Informationen können wir leicht erkennen, wo das Problem liegt und wie wir es beheben können.

R anhalten und beenden

Wenn ein Prozess für Ihren Geschmack zu lange läuft, klicken Sie einfach auf das Stopp-Schild, welches in der Menüleiste angezeigt wird. Nach kurzer Zeit sollte die Eingabeaufforderung > wieder erscheinen.

Zum Beenden von R haben Sie mehrere Möglichkeiten. In der Kommandozeile verwenden Sie die Anweisungen `quit()` oder `q()`. In der Menüzeile verwenden Sie File|Quit oder Datei|Beenden, je nach Sprachversion. Im Anschluss daran werden Sie gefragt, ob Sie den sogenannten `workspace` speichern wollen. Dies bietet sich an, denn so können Sie beim nächsten Start wieder dort einsetzen, wo Sie aufgehört haben. Der `workspace` enthält alle Variablen und Funktionen, die Sie im Laufe der Sitzung erzeugt haben.

Spalten zu einem Datensatz hinzufügen und entfernen

Zuweilen wollen Sie Ihrem Datensatz weitere Variable hinzufügen. Nehmen wir an, Sie wollen dem Datensatz aus unserem Beispiel die Variable `Land` hinzufügen. Dafür geben Sie folgendes ein:

```
> datensatz$Land <- c("DE","AT","CH")
> datensatz
    Stadt Einwohner Land
1  Berlin   3421829   DE
2    Wien   1793667   AT
3 Zuerich    402275   CH
```

Durch Anhängen von `$Land` an `datensatz` erzeugen Sie die neue Spalte Land im Datensatz. Dieser weisen Sie auch gleich neue Werte zu, welche Sie mithilfe der Funktion `c()` zunächst in einen Vektor verwandeln, der dann in einem Stück an den Datensatz angehängt wird.

Manchmal möchte man eine Spalte auch löschen. Das geht so:

```
> datensatz$Land <- NULL
> datensatz
    Stadt Einwohner
1  Berlin   3421829
2    Wien   1793667
3 Zuerich    402275
```

Wenn Sie nur eine Spalte von einem `data.frame` anzeigen lassen wollen, so geben Sie zum Beispiel ein:

```
> datensatz$Stadt
[1] "Berlin"  "Wien"    "Zuerich"
```

Schätzen, Testen und Vorhersagen

Nachdem Sie Daten gesammelt und die für Ihre Analyse eventuell zusätzlich benötigten Variablen erstellt haben, sind Sie bereit, ihr ökonometrisches Modell zu schätzen und Hypothesentests durchzuführen.

Die geeignete Schätzmethode hängt von der Natur Ihres ökonometrischen Modells ab. Arbeitspferde der Ökonometrie in R sind die Funktion `lm()` und ihre Verwandten (zum Beispiel `glm()`) Achtung, manche erfordern gesonderte Pakete, siehe den Tipp weiter oben. Innerhalb der der Schätzfunktion kommt der sogenannten Formelschnittstelle *(formula interface)* besondere Bedeutung zu. Hier geben Sie durch Formeln wie `x ~ y` das jeweilige Schätzmodell vor. Mehr dazu im Verlauf der nächsten Kapitel.

Folgende Tabelle gibt einen Überblick über Funktionen zur Diagnose oder Weiterverwendung von geschätzten Modellen.

Funktion	Beschreibung
`print()`	Gibt die Regressionsgleichung und Punktschätzer der Koeffizienten aus
`summary()`	Gibt umfangreiche Regressionsdaten aus
`coef()`	Gibt die Punktschätzer der Koeffizienten aus
`residuals()`	Gibt die Residuen aus
`fitted()`	Gibt die geschätzten Werte der abhängigen Variablen aus
`anova()`	Gibt die Varianzanalyse für ein oder mehrere geschätzte Modelle aus
`predict()`	Gibt Vorhersagen anhand des geschätzten Modells aus
`plot()`	Gibt diagnostische Grafiken aus
`confint()`	Gibt Konfidenzintervalle aus
`deviance()`	Gibt die Summe der Fehlerquadrate *(residual sum of squares)* aus
`vcov()`	Gibt die Varianz-Kovarianz-Matrix aus
`logLik()`	Gibt die log-likelihood (unter Normalverteilungsannahme) aus
`AIC()`	Informationskriterien (Akaike, …) jeweils unter Normalverteilungsannahme

Tabelle 1.1: (Diagnose-)Funktionen für geschätzte Modelle (nach Kleiber, Zeileis: Applied Econometrics with R (Springer))

Nehmen wir als einführendes Beispiel den `longley` Datensatz noch einmal etwas genauer unter die Lupe. Wie bereits erwähnt, steht dieser Datensatz nach dem Starten von R zur Verfügung. Die jeweils erste und letzte Zeile sind:

```
> head(longley,1)
     GNP.deflator     GNP Unemployed Armed.Forces Population Year Employed
1947           83 234.289      235.6          159    107.608 1947   60.323
> tail(longley,1)
     GNP.deflator     GNP Unemployed Armed.Forces Population Year Employed
1962        116.9 554.894      400.7        282.7    130.081 1962   70.551
```

Details zum Datensatz finden Sie unter `?longley`. Um die Struktur des Datensatzes kennen zu lernen geben Sie ein:

```
> str(longley)
'data.frame': 16 obs. of  7 variables:
 $ GNP.deflator: num  83 88.5 88.2 89.5 96.2 ...
 $ GNP         : num  234 259 258 285 329 ...
 $ Unemployed  : num  236 232 368 335 210 ...
 $ Armed.Forces: num  159 146 162 165 310 ...
 $ Population  : num  108 109 110 111 112 ...
 $ Year        : int  1947 1948 1949 1950 1951 1952 1953 1954 1955 1956 ...
 $ Employed    : num  60.3 61.1 60.2 61.2 63.2 ...
```

Es handelt sich also um einen `data.frame` mit 7 Variablen und je 16 Beobachtungen. Sie sehen die Variablennamen, deren Typ und die ersten Beobachtungen. Nehmen wir an, uns interessiert der Zusammenhang zwischen Bruttosozialprodukt *(Gross National Product, GNP)* und beschäftigten Personen *(Employed)*. Dann geben wir ein:

```
> longley.lm <- lm(Employed ~ GNP, data = longley)
> print(longley.lm)

Call:
lm(formula = Employed ~ GNP, data = longley)

Coefficients:
(Intercept)          GNP
   51.84359      0.03475
```

Zunächst schätzen wir die Modellgleichung `Employed ~ GNP` mithilfe der Funktion `lm()` und weisen das Ergebnis der Variablen `longley.lm` zu. Anschließend lassen wir uns mit `print()` grundlegende Informationen zur Schätzung ausgeben.

Probieren Sie einmal alle Anweisungen aus Tabelle 1.1 mit dem Regressionsobjekt – im Beispiel `longley.lm` – aus. Es lohnt sich!

Mit summary() erhalten wir schon einige Details mehr:

```
> summary(longley.lm)

Call:
lm(formula = Employed ~ GNP, data = longley)

Residuals:
     Min       1Q   Median       3Q      Max
–0.77958 –0.55440 –0.00944  0.34361  1.44594

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 51.843590   0.681372   76.09  < 2e–16 ***
GNP          0.034752   0.001706   20.37 8.36e–12 ***
–––
Signif. codes:  0 ‹***’ 0.001 ‹**’ 0.01 ‹*’ 0.05 ‹.’ 0.1 ‹ ’ 1

Residual standard error: 0.6566 on 14 degrees of freedom
Multiple R–squared:  0.9674, Adjusted R–squared:  0.965
F–statistic: 415.1 on 1 and 14 DF,  p–value: 8.363e–12
```

Das sind die Informationen, die man sich von einer Regression in etwa erwartet.