Informationsveranstaltung zur Anwendung und Relevanz von STATA

WueDive, 17.05.2023

Patrick Sturm, M. Sc.

Agenda

Themen der heutigen Veranstaltung:

1. Einführung in STATA: Relevanz und Verbreitung
2. Einführung in die Datenanalyse mit STATA
3. Praktische Anwendungsbeispiele
4. STATA-Community und Hilfe im Netz
5. Vorstellung des Moduls „Empirische Personalforschung mit STATA“

1. Einführung in STATA: Relevanz und Verbreitung

Allgemeine Hintergrundinfos zur Software:

- Statistikprogramm mit einer großen Bandbreite an Funktionen für die Datenaufbereitung, statistische Analysen und Grafiken
  → Mächtig für fortgeschrittene ökonometrische Methoden

- Entwickelt von StataCorp und zum ersten Mal 1985 erschienen
  → Version 18.0 seit April 2023 verfügbar

- Verschiedene Versionen für unterschiedliche Anforderungen (z.B. STATA BE, SE, MP)

- Regelmäßige Aktualisierungen und großes Angebot von nutzergeschriebenen Programmen

1. Einführung in STATA: Relevanz und Verbreitung

Relevanz und Verbreitung:

- **Universitäre Forschung:**
  - Große Bedeutung in den Bereichen Statistik, Wirtschafts- und Sozialwissenschaften sowie Epidemiologie
  - Hohe Verbreitung in der Lehre und in Forschungsprojekten an Universitäten/Hochschulen
  - Sehr häufig genutzt in Publikationen von wissenschaftlichen Fachzeitschriften (darunter häufig auch Quarterly Journal of Economics, American Economic Review etc.)

- **Staatliche Forschungsinstitutionen und Behörden:**
  - Beliebt durch die Möglichkeit einer effizienten und transparenten Datenverwaltung sowie durch die gute Durchführbarkeit von statistischen Analysen mit aktuellen ökonometrischen Methoden
  - Z.B.: Statistisches Bundesamt, Institut für Arbeitsmarkt- und Berufsforschung (IAB), Deutsches Institut für Wirtschaftsforschung (DIW), Leibniz-Institut für Wirtschaftsforschung (RWI)

- **Unternehmen:**
  - Beliebt bei Beratungsunternehmen für Marktanalysen, Finanzprognosen etc.
  - Unterstützt Unternehmen bei der Entscheidungsfindung und Evaluierung bisheriger Managementpraktiken

1. Einführung in STATA: Relevanz und Verbreitung

Vor- und Nachteile von STATA

- **Vorteile:**
  - \+ Recht nutzerfreundlich und intuitiv
  - \+ Sehr großer Umfang an statistischen Funktionen, darunter alle wichtigen deskriptive Statistiken, Regressionen, Zeitreihenverfahren etc.
  - \+ Verfügbarkeit von aktuellen fortgeschrittenen ökonometrischen Methoden (z.B. aufgrund von Beiträgen im *STATA Journal*)
  - \+ Nützliche und mächtige Funktionen für das Datenmanagement
  - \+ Gute Reproduzierbarkeit und hohe Transparenz der Analysen

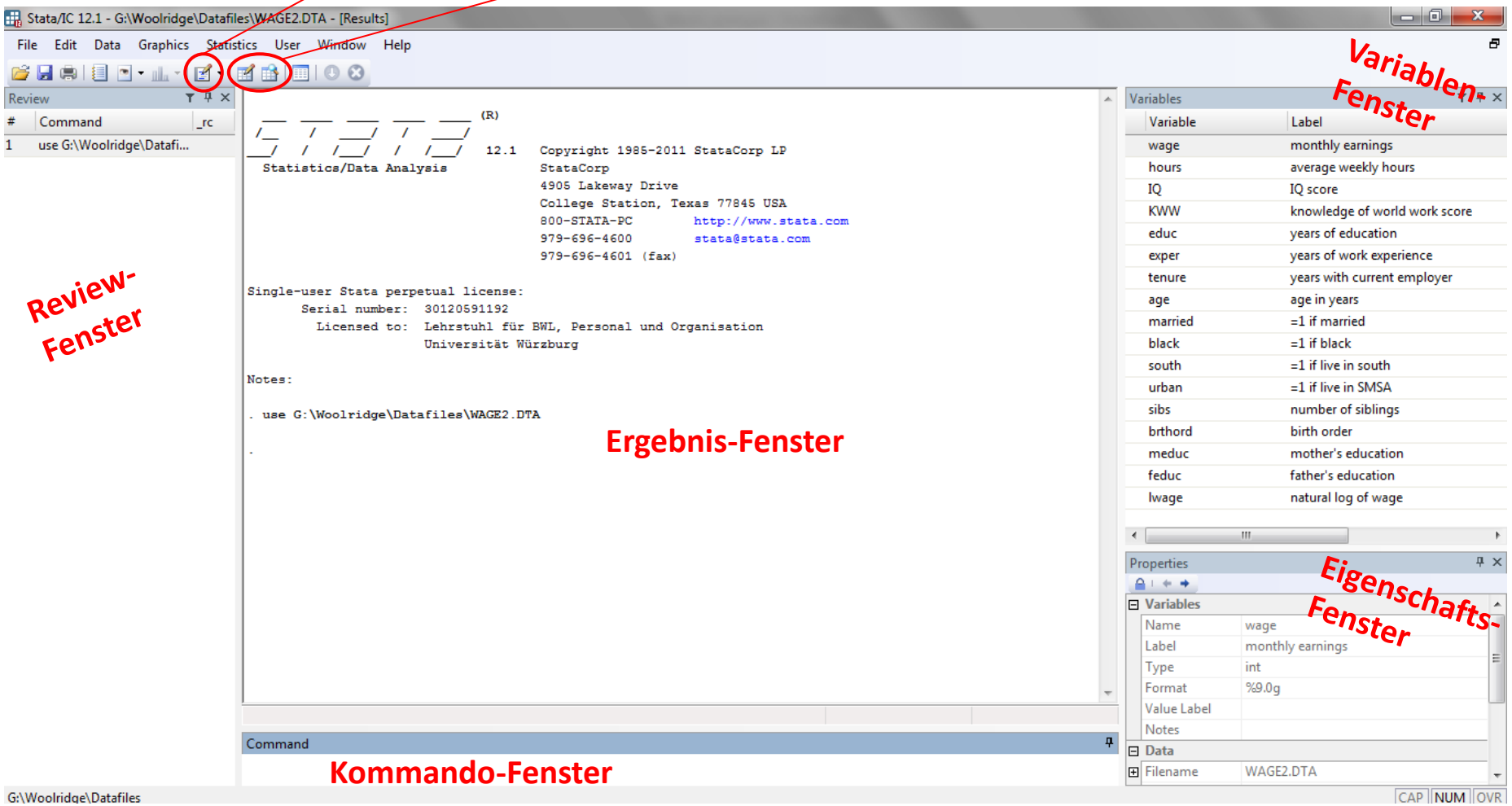
- **Nachteile**
  - – Teuer (100 – 500 € für jährliche individuelle Lizenzen)
  - – Bei bestimmten spezifischen Programmieranwendungen möglicherweise weniger flexibel als Python oder R
  - – In bestimmten Bereichen stärker vertreten als in anderen (hohe Verbreitung überwiegend in Sozial- und Wirtschaftswissenschaften)

Agenda

Themen der heutigen Veranstaltung:

1. Einführung in STATA: Relevanz und Verbreitung
2. **Einführung in die Datenanalyse mit STATA**
3. Praktische Anwendungsbeispiele
4. STATA-Community und Hilfe im Netz
5. Vorstellung des Moduls „Empirische Personalforschung mit STATA“

STATA Layout

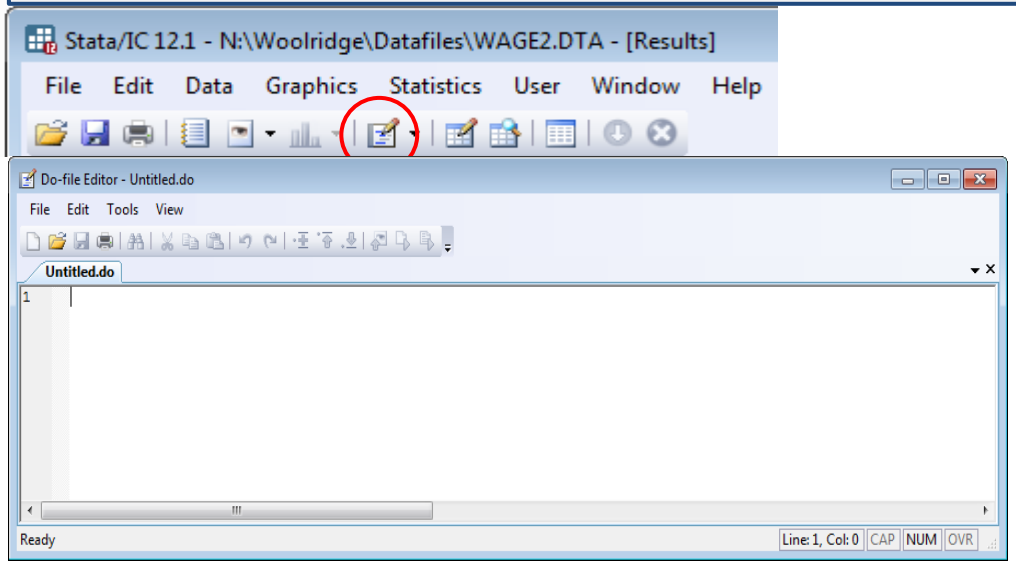


Do-Files

- Bearbeitung eines Datensatzes über Kommando-Fenster verursacht Probleme:
  - Veränderungen des Datensatzes, die nicht gespeichert werden, gehen verloren
  - Datentransformation kann nicht zurückverfolgt werden
  - In der Retrospektive ist Datenanalyse oder Datenmanipulation kaum nachvollziehbar

➡ Zur Datenanalyse immer den Do-file-Editor nutzen

➡ Veränderte Daten immer unter einem anderen Namen abspeichern



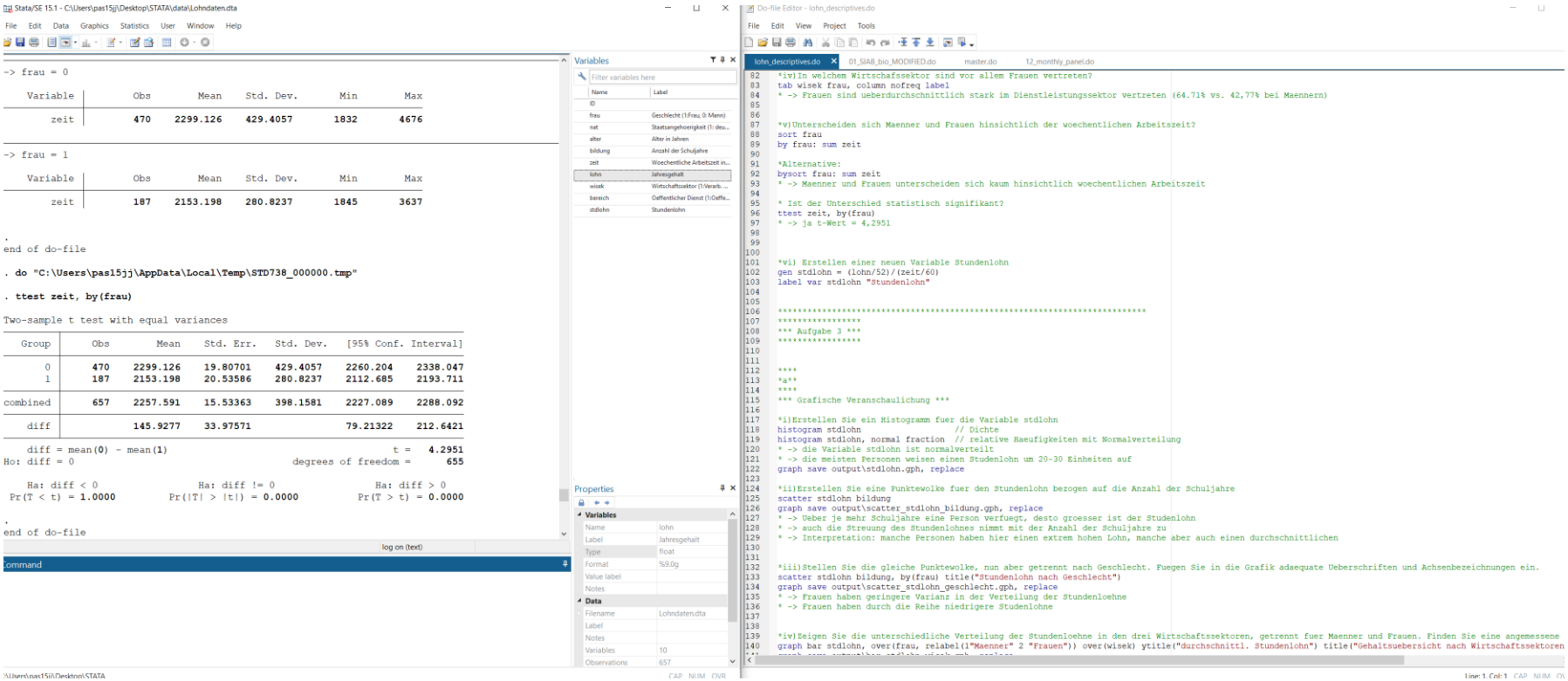
- Do-File beinhaltet Befehle, die der Reihe nach ausgeführt werden
- Erleichtert die Dokumentation
- Fehler können leicht korrigiert und das Do-File erneut ausgeführt werden

Art der Programmierung bei der Anwendung:

- Programmierung erfolgt hauptsächlich über Befehle, die nacheinander nach dem Starten der „Do-files“ durchlaufen
- Verschiedene Möglichkeiten für das Automatisieren der Anwendungen: „Macros“, „Loops“ mit Konditionen
- Spezifische mathematische Anwendungen mit „Mata“, der Matrixprogrammiersprache von STATA
- Nutzung von User-geschriebenen Programmen („ado-files“), die viele Abläufe erleichtern können

2. Einführung in die Datenanalyse mit STATA

Übliche Arbeitsfenster:



```
-> frau = 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        zeit |        470    2299.126    429.4057       1832       4676

-> frau = 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        zeit |        187    2153.198    280.8237       1845       3637

.
end of do-file

. do "C:\Users\pas15jj\AppData\Local\Temp\STD738_000000.tmp"

. ttest zeit, by(frau)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     470    2299.126    19.80701    429.4057    2260.204    2338.047
       1 |     187    2153.198    20.53586    280.8237    2112.685    2193.711
---------+--------------------------------------------------------------------
combined |     657    2257.591    15.53363    398.1581    2227.089    2288.092
---------+--------------------------------------------------------------------
    diff |            145.9277    33.97571                79.21322    212.6421
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   4.2951
Ho: diff = 0                     degrees of freedom =      655

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000

.
end of do-file
```

```
*iv)In welchem Wirtschafssektor sind vor allem Frauen vertreten?
tab wisek frau, column nofreq label
* -> Frauen sind ueberdurchschnittlich stark im Dienstleistungssektor vertreten (64.71% vs. 42,77% bei Maennern)


*v)Unterscheiden sich Maenner und Frauen hinsichtlich der woechentlichen Arbeitszeit?
sort frau
by frau: sum zeit

*Alternative:
bysort frau: sum zeit
* -> Maenner und Frauen unterscheiden sich kaum hinsichtlich woechentlichen Arbeitszeit

* Ist der Unterschied statistisch signifikant?
ttest zeit, by(frau)
* -> ja t-Wert = 4,2951



*vi) Erstellen einer neuen Variable Stundenlohn
gen stdlohn = (lohn/52)/(zeit/60)
label var stdlohn "Stundenlohn"


**********************************************************************
****************
*** Aufgabe 3 ***
****************


****
*a**
****
*** Grafische Veranschaulichung ***

*i)Erstellen Sie ein Histogramm fuer die Variable stdlohn
histogram stdlohn                    // Dichte
histogram stdlohn, normal fraction  // relative Haeufigkeiten mit Normalverteilung
* -> die Variable stdlohn ist normalverteilt
* -> die meisten Personen weisen einen Stundenlohn um 20-30 Einheiten auf
graph save output\stdlohn.gph, replace

*ii)Erstellen Sie eine Punktewolke fuer den Stundenlohn bezogen auf die Anzahl der Schuljahre
scatter stdlohn bildung
graph save output\scatter_stdlohn_bildung.gph, replace
* -> Ueber je mehr Schuljahre eine Person verfuegt, desto groesser ist der Studenlohn
* -> auch die Streuung des Stundenlohnes nimmt mit der Anzahl der Schuljahre zu
* -> Interpretation: manche Personen haben hier einen extrem hohen Lohn, manche aber auch einen durchschnittlichen


*iii)Stellen Sie die gleiche Punktewolke, nun aber getrennt nach Geschlecht. Fuegen Sie in die Grafik adaequate Ueberschriften und Achsenbezeichnungen ein.
scatter stdlohn bildung, by(frau) title("Stundenlohn nach Geschlecht")
graph save output\scatter_stdlohn_geschlecht.gph, replace
* -> Frauen haben geringere Varianz in der Verteilung der Stundenloehne
* -> Frauen haben durch die Reihe niedrigere Studenlohne


*iv)Zeigen Sie die unterschiedliche Verteilung der Stundenloehne in den drei Wirtschaftssektoren, getrennt fuer Maenner und Frauen. Finden Sie eine angemessene
graph bar stdlohn, over(frau, relabel(1"Maenner" 2 "Frauen")) over(wisek) ytitle("durchschnittl. Stundenlohn") title("Gehaltsuebersicht nach Wirtschaftssektoren
```

Agenda

Themen der heutigen Veranstaltung:

1. Einführung in STATA: Relevanz und Verbreitung
2. Einführung in die Datenanalyse mit STATA
3. **Praktische Anwendungsbeispiele**
4. STATA-Community und Hilfe im Netz
5. Vorstellung des Moduls „Empirische Personalforschung mit STATA“

Deskriptive Statistik

Univariate Statistik

sort wfl_60
by wfl_60: sum nmqm, detail
Alternativ: bysort wfl_60: sum nmqm, detail

//Alternative zu vorherigem Befehl mit allen Ausprägungen: getrennte Deskription der Nettomiete pro m² für unter und über 60 m² Wohnfläche

```
-> wfl_60 = Wohnfläche kleiner 60m²

                          nmqm
-------------------------------------------------------------
      Percentiles      Smallest
 1%           2.6           1.8
 5%          4.31          1.89
10%          5.38          2.16       Obs                 743
25%          7.52          2.16       Sum of Wgt.         743

50%          9.31                     Mean           9.025114
                       Largest       Std. Dev.      2.560931
75%         10.73         16.81
90%         12.02         16.93       Variance       6.558368
95%         12.58         17.45       Skewness      -.2314117
99%         14.34         20.09       Kurtosis       3.497972
```

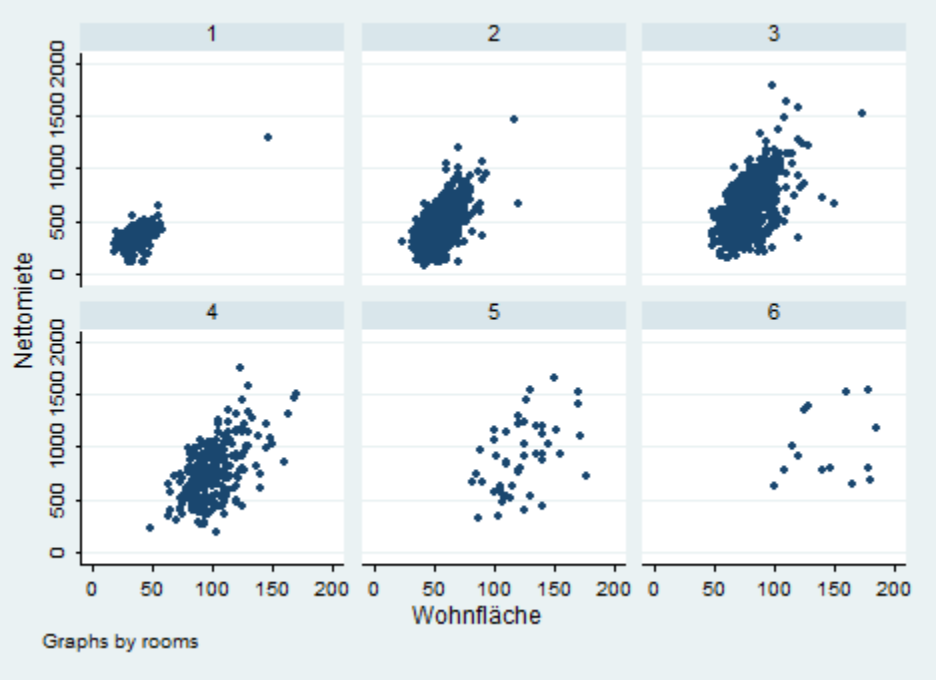
```
-> wfl_60 = Wohnfläche größer gleich 60 m²

                          nmqm
-------------------------------------------------------------
      Percentiles      Smallest
 1%          2.88          1.47
 5%          4.15          1.89
10%          4.95          2.24       Obs                1,310
25%          6.45          2.29       Sum of Wgt.        1,310

50%           8.1                     Mean           8.035893
                       Largest       Std. Dev.      2.33808
75%          9.58         16.45
90%         10.98         17.09       Variance       5.466617
95%         11.75         17.27       Skewness       .1228333
99%         13.47         18.26       Kurtosis       3.254621
```

Grafische Darstellungen

Gruppierte Grafiken

sort [groupvarname]
scatter bzw. bar [varnames], by (groupvarname) [options]

Beispiel:
sort rooms
scatter nm wfl, by(rooms) ytitle("Nettomiete") xtitle("Wohnfläche")



// gruppierte Streudiagramme über die beiden Variablen Nettomiete in Euro und Wohnfläche ausgegeben für die unterschiedliche Anzahl an Räumen

Grafische Darstellungen

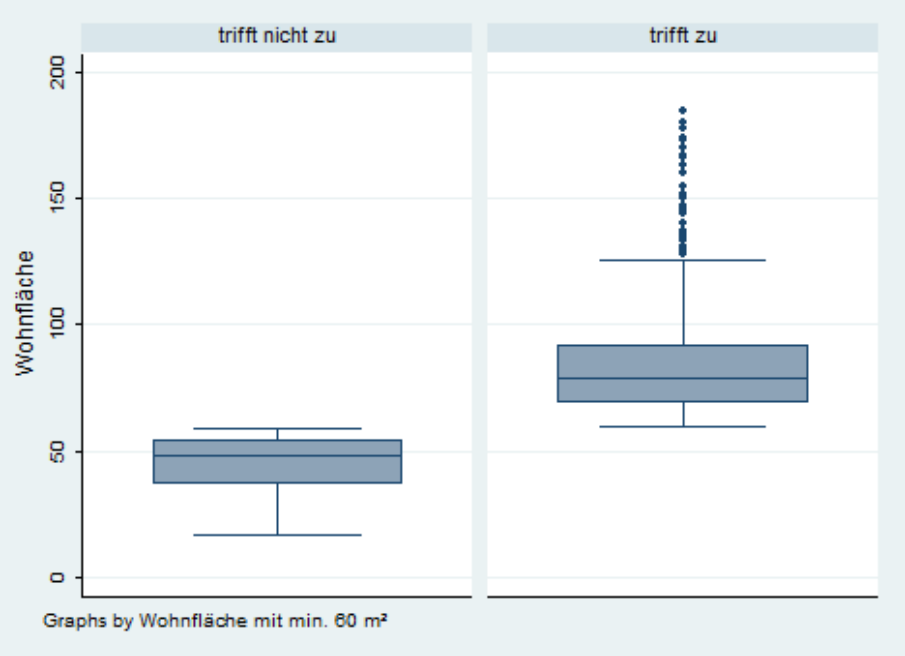
Boxplot

- grafische Darstellung robuster Streuungs- und Lagemaße

graph box [varname], by(groupvarname) [options]

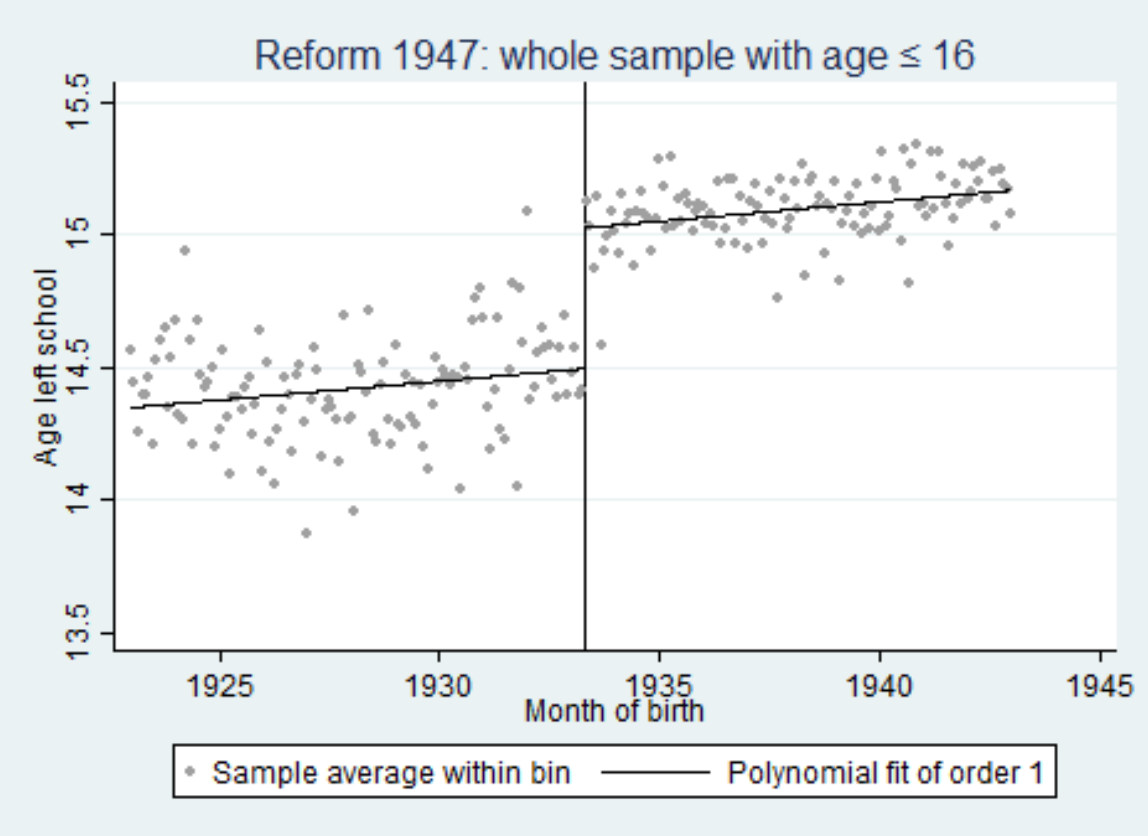
Beispiel:

graph box wfl, by(wfl_60) ytitle("Wohnfläche") // Boxplot der Variable Wohnfläche
graph save output/box_wfl.gph, replace



Grafische Darstellungen

Grafische Analysen der Auswirkungen von Arbeitsmarkt- und Bildungsreformen



Darstellung von Regressionsergebnissen

Regression Mietspiegel

VARIABLES	(1) Modell 1	(2) Modell 2	(3) Modell 3
Altbauwohnung	-1.265***	-1.029***	-1.125***
	(0.117)	(0.118)	(0.113)
wfl		-0.018***	-0.005
		(0.002)	(0.004)
bez			-0.035***
			(0.007)
wohngut			0.751***
			(0.106)
wohnbest			2.028***
			(0.334)
badkach0			-0.999***
			(0.125)
kueche			1.688***
			(0.188)
rooms			-0.510***
			(0.092)
Constant	8.762***	9.932***	10.521***
	(0.063)	(0.153)	(0.179)
Observations	2,053	2,053	2,053
R-squared	0.054	0.085	0.219

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 9: Non-parametric estimations: 1972 reform

	1972 reform					
Sample	Mother-Offspring			Father-Offspring		
Kernel	Triangular	Epanech-nikov	Uniform	Triangular	Epanech-nikov	Uniform
First Stage						
Treatment	0.453***	0.430***	0.464***	0.320***	0.320***	0.279***
	(0.066)	(0.065)	(0.0685)	(0.061)	(0.062)	(0.080)
Second Stage						
GHQ-12	0.917	0.929	0.835	-0.463	-0.319	-0.593
	(0.797)	(0.758)	(0.785)	(1.039)	(1.020)	(1.243)
Observations	41017	41017	41017	25954	25954	25954

Notes: Local linear estimation, optimal bandwidth estimated using Calonico et al. (2014)'s method. Standard errors are clustered at the month-year of birth level. ***, ** and * indicate significance at the 1%, 5% and 10% level, respectively. All regressions include the child's sex, survey-year dummies, a third order age polynomial of the child and the respective parent.
Source: Own calculations based on BHPS and UKHLS.

Agenda

Themen der heutigen Veranstaltung:

1. Einführung in STATA: Relevanz und Verbreitung
2. Einführung in die Datenanalyse mit STATA
3. Praktische Anwendungsbeispiele
4. **STATA-Community und Hilfe im Netz**
5. Vorstellung des Moduls „Empirische Personalforschung mit STATA“

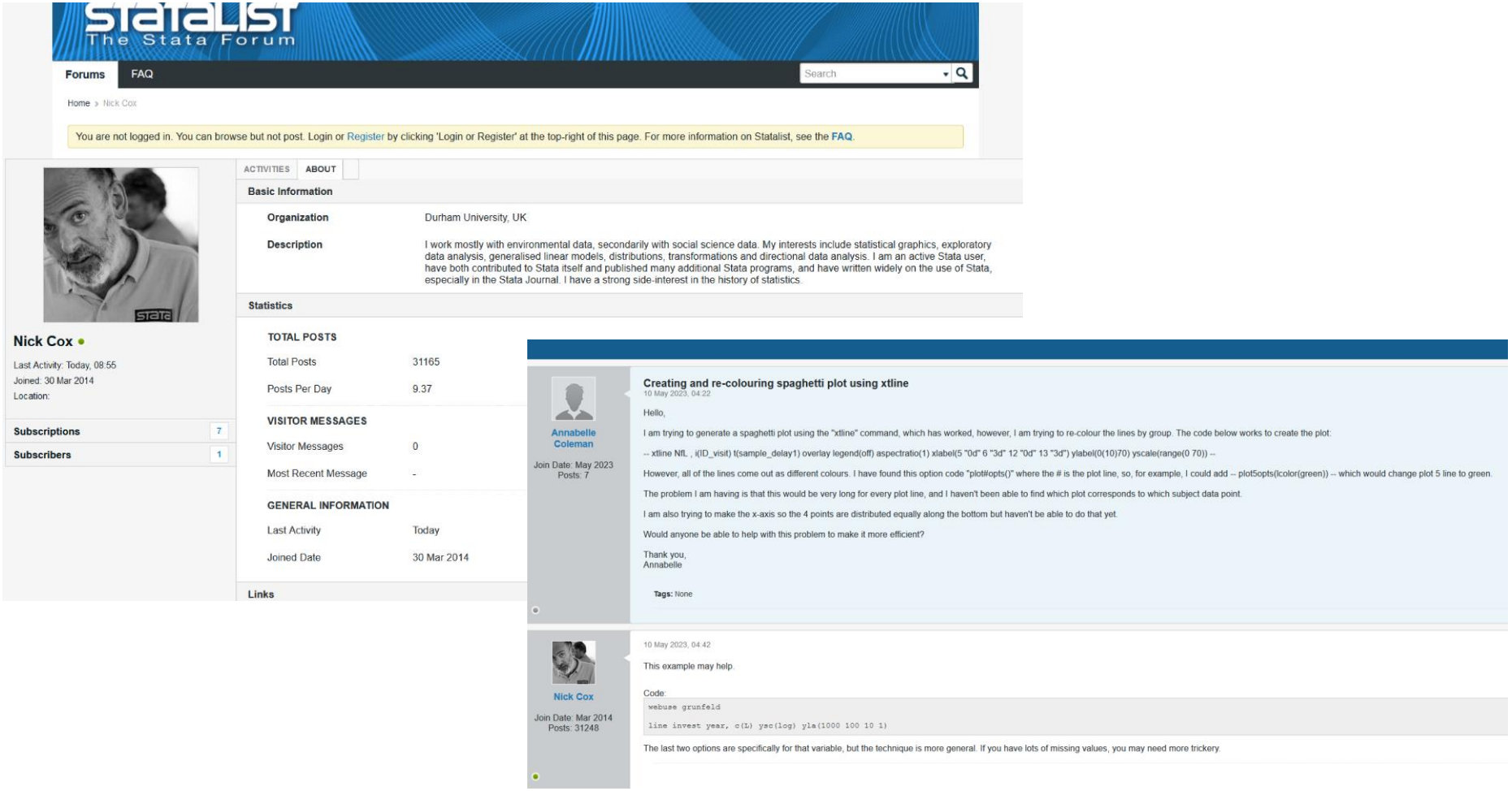
4. STATA-Community und Hilfe im Netz

- **STATALIST (STATA Forum):**
  - Fragen & Antworten zu Themen über Datenaufbereitung, statistische Analysen und Programmierungsmöglichkeiten

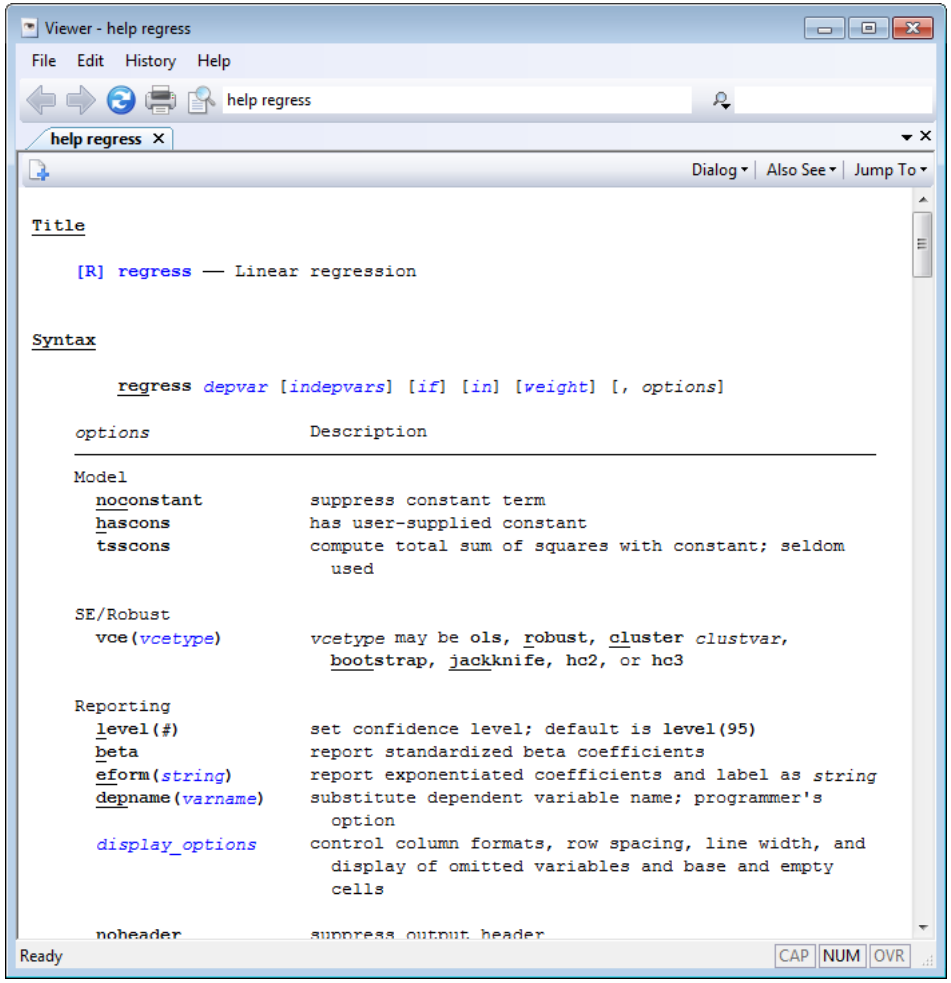
- Ausführliche und gut strukturierte Dokumentation der Anwendungen:
  - Direkter Zugriff auf Anleitungen und Handbücher
  - Dokumentationen und nützliche Hinweise zu den Befehlen in der Software selbst verfügbar

- **STATA Journal:**
  - Vierteljährlich veröffentlichte Fachzeitschrift für neue statistische Methoden und praktische Anwendungen
  - Impact Score: 4,45 (STATA Press)

4. STATA-Community und Hilfe im Netz



Hilfsfunktion



- Wird in Kommando-Fenster eingegeben – startet mit „help“
- Erleichtert Anfängern das Finden der entsprechenden STATA-Befehle
- Hilfe erhält man auch über:
  - Google-Suche
  - http://www.stata-forum.de/
- *Search*-Option für Nutzer, die das Kommando kennen, aber Zusatzinfos benötigen

The STATA Journal

The Stata Journal (2023)
23, Number 1, pp. 293–297

DOI: 10.1177/1536867X231162009

Stata tip 151: Puzzling out some logical operators

Nicholas J. Cox
Department of Geography
Durham University
Durham, U.K.
n.j.cox@durham.ac.uk

The logical operators `&` ("and") and `|` ("or") can sometimes be tricky in statistical software such as Stata. They are extremely useful, so you need to understand thoroughly how they work. Any trickiness arises mostly in translating from ordinary language to a statistical computer language. Here I survey various common confusions and explain what to do instead.

`auto.dta` in Stata will serve fine as a sandbox.

```
. sysuse auto
(1978 automobile data)
```

1 What is wrong, and why

The repair record variable `rep78` in `auto.dta` takes on values 1 (poor) to 5 (best) and also missing. You can see that with a simple tabulation:

```
. tabulate rep78, missing
```

Repair record 1978	Freq.	Percent	Cum.
1	2	2.70	2.70
2	8	10.81	13.51
3	30	40.54	54.05
4	18	24.32	78.38
5	11	14.86	93.24
.	5	6.76	100.00
Total	74	100.00	

Let's see which cars have repair record 1. You need to use the operator `==` when testing for equality. If this point is new to you, please consult `help operators`.

```
. list make rep78 if rep78 == 1
```

make	rep78

Agenda

Themen der heutigen Veranstaltung:

1. Einführung in STATA: Relevanz und Verbreitung
2. Einführung in die Datenanalyse mit STATA
3. Praktische Anwendungsbeispiele
4. STATA-Community und Hilfe im Netz
5. **Vorstellung des Moduls „Empirische Personalforschung mit STATA“**

5. Empirische Personalforschung mit STATA

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Der Schwerpunkt der Veranstaltung liegt auf dem Nachvollziehen empirischer Fragestellungen, damit verbundenen Problemen sowie der methodischen Umsetzung.

Durch die Anwendung der gelernten theoretischen Kursinhalte mit dem Statistikprogramm STATA soll ein Grundverständnis für die Arbeit mit Daten geschaffen und die Intuition für verschiedene Schätzverfahren und -probleme geschärft werden.

Inhalt:

- 1. Einführung mit STATA
- 2. Einfache deskriptive Darstellungen
- 3. Transformationen
- 4. Einfache Regressionsmodelle
- 5. Schätzer für binäre Zielvariablen und Interaktionen
- 6. Logit / Probit -Regressionen
- 7. Endogenität
- 8. Instrumentalvariablen

5. Empirische Personalforschung mit STATA

Termine		Veranstaltungsübersicht
Innerhalb von 4 – 5 Wochen	**Block 1**	▪ Was ist Ökonometrie? ▪ Einführung in die empirische Datenanalyse mit STATA
	Block 2	▪ Das einfache Regressionsmodell ▪ Das multiple Regressionsmodell ▪ Omitted Variable Bias ▪ Inferenzstatistik
	Block 3	▪ Datentransformation ▪ Dummyvariablen ▪ Schätzer für binäre Zielvariablen
	Block 4	▪ Endogenität ▪ Instrumentalvariablen ▪ Wiederholung
Einwöchiger Prüfungszeitraum	**Home-Assignment**	Selbständige Bearbeitung des Home-Assignments (freie Orts- und Zeiteinteilung) in **Teams mit bis zu drei Personen**

Vielen Dank für Ihre Aufmerksamkeit! Haben Sie noch Fragen oder Anregungen?

Patrick Sturm, M. Sc.
Zimmer 390
Sanderring 2
97070 Würzburg
Sprechstunden nach Vereinbarung per E-Mail
E-Mail: patrick.sturm@uni-wuerzburg.de
Telefon: 0931/ 31-83476