



# Informationsveranstaltung zur Anwendung und Relevanz von STATA

WueDive, 17.05.2023

Patrick Sturm, M. Sc.



## 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 ökonomische 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)



## 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“



## 2. Einführung in die Datenanalyse mit STATA

### STATA Layout

Do-File öffnen

Datensatz anzeigen

The screenshot shows the STATA 12.1 interface with the following components and labels:

- Review-Fenster:** A small window on the left showing the command history with the entry: `1 use G:\Woolridge\Datafi...`
- Ergebnis-Fenster:** The main central window displaying the STATA startup screen, including the logo, version (12.1), copyright information, and license details for the Chair of Business Administration at the University of Würzburg.
- Kommando-Fenster:** A window at the bottom for entering commands, currently showing the command `use G:\Woolridge\Datafiles\WAGE2.DTA`.
- Variablen-Fenster:** A window on the right listing available variables and their labels, such as `wage` (monthly earnings), `hours` (average weekly hours), and `IQ` (IQ score).
- Eigenschafts-Fenster:** A window at the bottom right showing the properties of the selected variable `wage`, including its name, label, type (int), format (%9.0g), and value label.



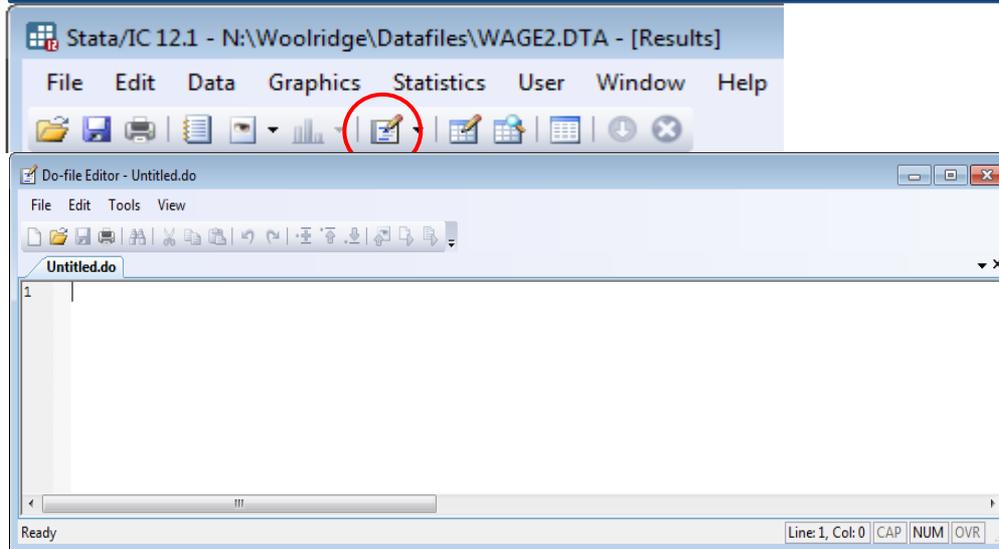
## 2. Einführung in die Datenanalyse mit STATA

### 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

## 2. Einführung in die Datenanalyse mit STATA

### 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:

The screenshot displays the STATA software interface with three main windows:

- Command Window (Left):** Shows the execution of a do-file. It starts with `frau = 0` and `frau = 1`, followed by a `tttest zeit, by(frau)` command. The output shows a two-sample t-test with equal variances. The command window also shows the `end of do-file` message and the `command` window.
- Variables Window (Middle):** Lists the variables in the dataset: `id`, `frau`, `net`, `alter`, `bildung`, `zeit`, `lohn`, `wisek`, `bereich`, and `stdlohn`.
- Do-file Editor (Right):** Shows the source code for the do-file. It includes comments in German explaining the steps:
  - Line 82: `*iv) In welchem Wirtschaftssektor sind vor allem Frauen vertreten?`
  - Line 83: `tab wisek frau, column nofreq label`
  - Line 84: `* -> Frauen sind ueberdurchschnittlich stark im Dienstleistungssektor vertreten (64,71% vs. 42,77% bei Maennern)`
  - Line 87: `*v) Unterscheiden sich Maenner und Frauen hinsichtlich der woechentlichen Arbeitszeit?`
  - Line 88: `sort frau`
  - Line 89: `by frau: sum zeit`
  - Line 91: `*Alternative:`
  - Line 92: `bysort frau: sum zeit`
  - Line 93: `* -> Maenner und Frauen unterscheiden sich kaum hinsichtlich woechentlichen Arbeitszeit`
  - Line 94: `tttest zeit, by(frau)`
  - Line 95: `* Ist der Unterschied statistisch signifikant?`
  - Line 96: `test zeit, by(frau)`
  - Line 97: `* -> ja t-Wert = 4,2951`
  - Line 101: `*vi) Erstellen einer neuen Variable Stundenlohn`
  - Line 102: `gen stdlohn = (lohn/52)/(zeit/60)`
  - Line 103: `label var stdlohn "Stundenlohn"`
  - Line 107: `*****`
  - Line 108: `*** Aufgabe 3 ***`
  - Line 109: `*****`
  - Line 111: `****`
  - Line 112: `****`
  - Line 113: `****`
  - Line 114: `****`
  - Line 115: `*** Grafische Veranschaulichung ***`
  - Line 117: `*i) Erstellen Sie ein Histogramm fuer die Variable stdlohn`
  - Line 118: `histogram stdlohn // Dichte`
  - Line 119: `histogram stdlohn, normal fraction // relative Haeufigkeiten mit Normalverteilung`
  - Line 120: `* -> die Variable stdlohn ist normalverteilt`
  - Line 121: `* -> die meisten Personen weisen einen Stundenlohn um 20-30 Einheiten auf`
  - Line 122: `graph save output/stdlohn.gph, replace`
  - Line 124: `*ii) Erstellen Sie eine Punktwolke fuer den Stundenlohn bezogen auf die Anzahl der Schuljahre`
  - Line 125: `scatter stdlohn bildung`
  - Line 126: `graph save output/scatter_stdlohn_bildung.gph, replace`
  - Line 127: `* -> Ueber je mehr Schuljahre eine Person verfügt, desto grosser ist der Stundenlohn`
  - Line 128: `* -> auch die Streuung des Stundenlohnes nimmt mit der Anzahl der Schuljahre zu`
  - Line 129: `* -> Interpretation: manche Personen haben hier einen extrem hohen Lohn, manche aber auch einen durchschnittlichen`
  - Line 132: `*iii) Stellen Sie die gleiche Punktwolke, nun aber getrennt nach Geschlecht. Fuegen Sie in die Grafik adaequate Ueberschriften und Achsenbezeichnungen ein.`
  - Line 133: `scatter stdlohn bildung, by(frau) title("Stundenlohn nach Geschlecht")`
  - Line 134: `graph save output/scatter_stdlohn_geschlecht.gph, replace`
  - Line 135: `* -> Frauen haben geringere Varianz in der Verteilung der Stundenloehne`
  - Line 136: `* -> Frauen haben durch die Reihe niedrigere Stundenloehne`
  - Line 139: `*iv) Zeigen Sie die unterschiedliche Verteilung der Stundenloehne in den drei Wirtschaftssektoren, getrennt fuer Maenner und Frauen. Finden Sie eine angemessene`
  - Line 140: `graph bar stdlohn, over(frau, relabel(1="Maenner" 2 "Frauen*)) over(wisek) ytitle("durchschnittl. Stundenlohn") title("Gehaltsuebersicht nach Wirtschaftssektoren")`

## 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“



# 3. Praktische Anwendungsbeispiele

## 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



## 3. Praktische Anwendungsbeispiele

# 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

## 3. Praktische Anwendungsbeispiele

### Grafische Darstellungen

#### Boxplot

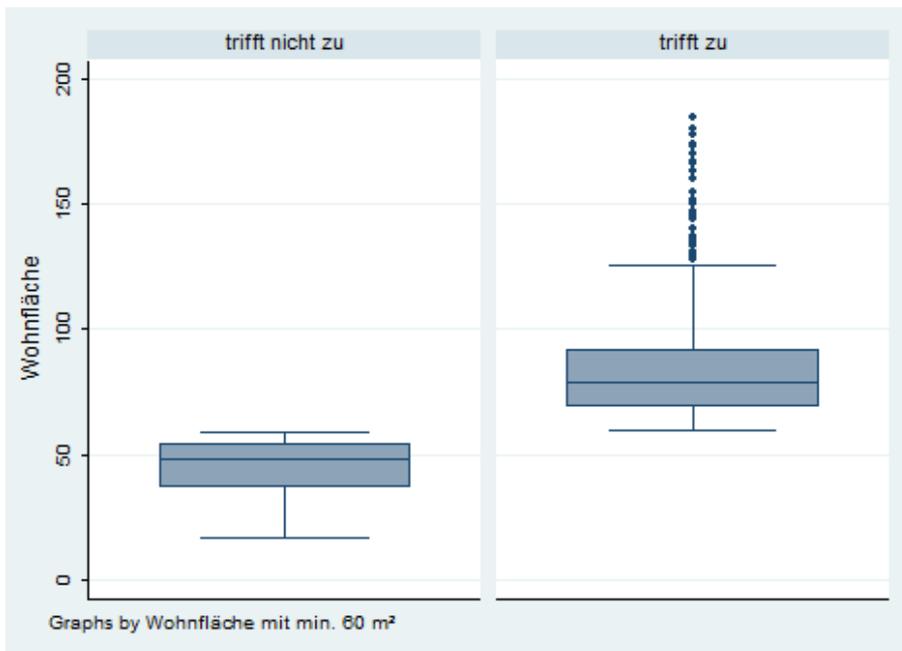
- grafische Darstellung robuster Streuungs- und Lagemaße

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

#### Beispiel:

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

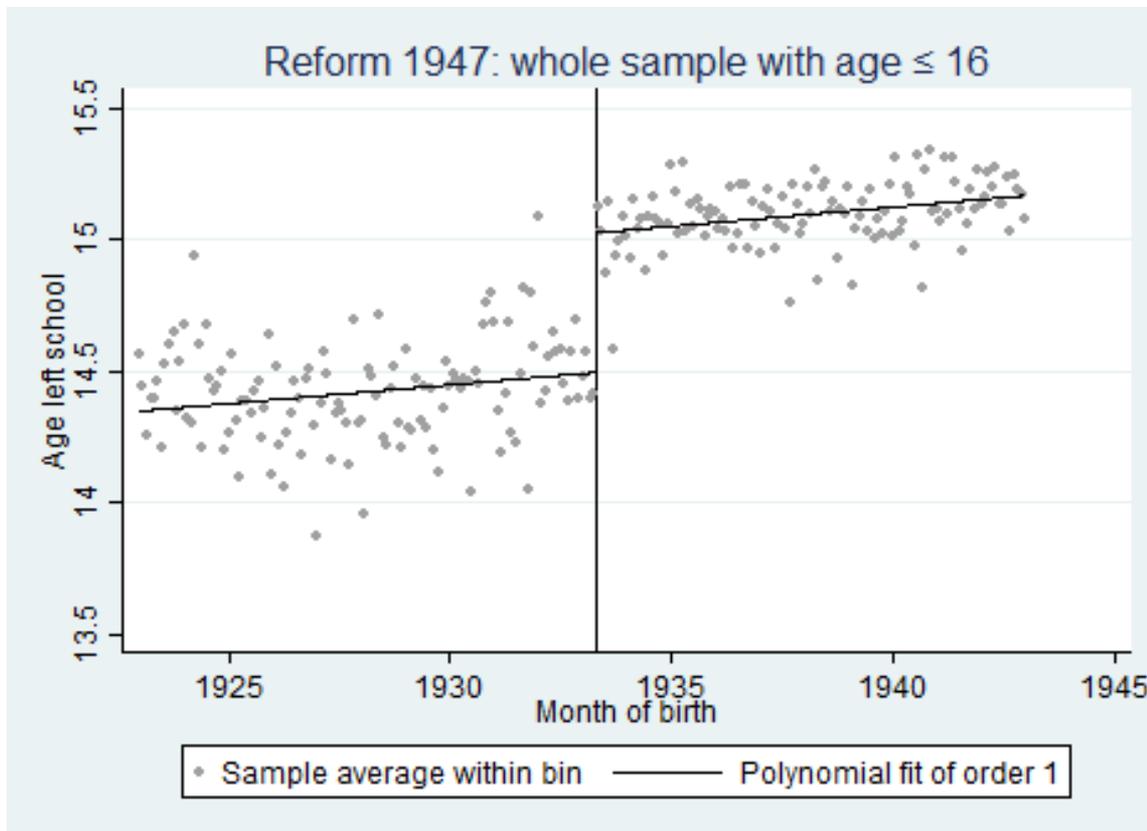
// Boxplot der Variable Wohnfläche



### 3. Praktische Anwendungsbeispiele

## Grafische Darstellungen

Grafische Analysen der Auswirkungen von Arbeitsmarkt- und Bildungsreformen



### 3. Praktische Anwendungsbeispiele

## Darstellung von Regressionsergebnissen

Regression Mietspiegel			
VARIABLES	(1) Modell 1	(2) Modell 2	(3) Modell 3
Altbauwohnung	-1.265*** (0.117)	-1.029*** (0.118)	-1.125*** (0.113)
wfl		-0.018*** (0.002)	-0.005 (0.004)
bez			-0.035*** (0.007)
wohngut			0.751*** (0.106)
wohnbst			2.028*** (0.334)
badkach0			-0.999*** (0.125)
kueche			1.688*** (0.188)
rooms			-0.510*** (0.092)
Constant	8.762*** (0.063)	9.932*** (0.153)	10.521*** (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

Sample	1972 reform					
	Mother-Offspring			Father-Offspring		
Kernel	Triangular	Epanechnikov	Uniform	Triangular	Epanechnikov	Uniform
<b>First Stage</b>						
Treatment	0.453*** (0.066)	0.430*** (0.065)	0.464*** (0.0685)	0.320*** (0.061)	0.320*** (0.062)	0.279*** (0.080)
<b>Second Stage</b>						
GHQ-12	0.917 (0.797)	0.929 (0.758)	0.835 (0.785)	-0.463 (1.039)	-0.319 (1.020)	-0.593 (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.



## 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



Forums    FAQ    Search

Home > Nick Cox

You are not logged in. You can browse but not post. Login or Register by clicking 'Login or Register' at the top-right of this page. For more information on Statalist, see the FAQ.



**Nick Cox**

Last Activity: Today, 08:55  
 Joined: 30 Mar 2014  
 Location:

Subscriptions: 7  
 Subscribers: 1

ACTIVITIES    ABOUT

**Basic Information**

**Organization**      Durham University, UK

**Description**      I work mostly with environmental data, secondarily with social science data. My interests include statistical graphics, exploratory data analysis, generalised linear models, distributions, transformations and directional data analysis. I am an active Stata user, have both contributed to Stata itself and published many additional Stata programs, and have written widely on the use of Stata, especially in the Stata Journal. I have a strong side-interest in the history of statistics.

**Statistics**

**TOTAL POSTS**

Total Posts            31165  
 Posts Per Day          9.37

**VISITOR MESSAGES**

Visitor Messages        0  
 Most Recent Message   -

**GENERAL INFORMATION**

Last Activity            Today  
 Joined Date            30 Mar 2014

**Links**



**Annabelle Coleman**  
 Join Date: May 2023  
 Posts: 7

**Creating and re-colouring spaghetti plot using xtline**  
 10 May 2023, 04:22

Hello,

I am trying to generate a spaghetti plot using the "xtline" command, which has worked, however, I am trying to re-colour the lines by group. The code below works to create the plot:

```
-- xtline N1L , i(ID_visit) t(sample_delay1) overlay legend(off) aspectratio(1) xlabel(5 "0d" 6 "3d" 12 "0d" 13 "3d") ylabel(0(10)70) yscale(range(0 70)) --
```

However, all of the lines come out as different colours. I have found this option code "plotlopts(\*)" where the # is the plot line, so, for example, I could add -- plot5opts(icolor(green)) -- which would change plot 5 line to green.

The problem I am having is that this would be very long for every plot line, and I haven't been able to find which plot corresponds to which subject data point.

I am also trying to make the x-axis so the 4 points are distributed equally along the bottom but haven't been able to do that yet.

Would anyone be able to help with this problem to make it more efficient?

Thank you,  
 Annabelle

**Tags:** None

---



**Nick Cox**  
 Join Date: Mar 2014  
 Posts: 31248

10 May 2023, 04:42

This example may help.

Code:

```
webuse grunfeld
xtline invest year, o(L) yac(log) yla(1000 100 10 1)
```

The last two options are specifically for that variable, but the technique is more general. If you have lots of missing values, you may need more trickery.



## 4. STATA-Community und Hilfe im Netz

### Hilfsfunktion

Viewer - help regress

File Edit History Help

help regress

help regress X

Dialog Also See Jump To

**Title**

[R] regress — Linear regression

**Syntax**

regress depvar [indepvars] [if] [in] [weight] [, options]

options	Description
<b>Model</b>	
noconstant	suppress constant term
hascons	has user-supplied constant
tsscons	compute total sum of squares with constant; seldom used
<b>SE/Robust</b>	
vce(vcetype)	vcetype may be ols, robust, cluster clustvar, bootstrap, jackknife, hc2, or hc3
<b>Reporting</b>	
level(#)	set confidence level; default is level(95)
beta	report standardized beta coefficients
eform(string)	report exponentiated coefficients and label as string
depname(varname)	substitute dependent variable name; programmer's option
display_options	control column formats, row spacing, line width, and display of omitted variables and base and empty cells
noheader	suppress output header

Ready CAP NUM OVR

- 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

# 4. STATA-Community und Hilfe im Netz

## 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
...	...

## 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

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
<p><b>Innerhalb von 4 – 5 Wochen</b></p>	<p><b>Block 1</b></p>	<ul style="list-style-type: none"> <li>▪ Was ist Ökonometrie?</li> <li>▪ Einführung in die empirische Datenanalyse mit STATA</li> </ul>
	<p><b>Block 2</b></p>	<ul style="list-style-type: none"> <li>▪ Das einfache Regressionsmodell</li> <li>▪ Das multiple Regressionsmodell</li> <li>▪ Omitted Variable Bias</li> <li>▪ Inferenzstatistik</li> </ul>
	<p><b>Block 3</b></p>	<ul style="list-style-type: none"> <li>▪ Datentransformation</li> <li>▪ Dummyvariablen</li> <li>▪ Schätzer für binäre Zielvariablen</li> </ul>
	<p><b>Block 4</b></p>	<ul style="list-style-type: none"> <li>▪ Endogenität</li> <li>▪ Instrumentalvariablen</li> <li>▪ Wiederholung</li> </ul>
<p><b>Einwöchiger Prüfungszeitraum</b></p>	<p><b>Home-Assignment</b></p>	<p>Selbständige Bearbeitung des Home-Assignments (freie Orts- und Zeiteinteilung) in <b><u>Teams mit bis zu drei Personen</u></b></p>



**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](mailto:patrick.sturm@uni-wuerzburg.de)

Telefon: 0931/ 31-83476

