Ablauf des Seminars

Termin Dienstag: 14:45	Inhalt
15.01.2019	Theoretische Einführung nicht-parametrische Methoden Teil I
22.01.2019	Theoretische Einführung nicht-parametrische Methoden Teil II Forschungsprozess, Prüfungsleistung, Gruppeneinteilung, Themenvergabe
29.01.2019	**Praktische Einführung in STATA Teil I (PC-Pool 4 in 46/104)**
05.02.2019	Praktische Einführung in STATA Teil II (PC-Pool in 46/104)
Ca. 12.03.2019	**!Einzeltermine** *zu: Zwischenstand, Problemen, Fragen* *(bitte eine Woche vorher vereinbaren: alexander.kaiser@unibw.de)*
29.03.2019	Deadline Exposé

Praktische Einführung in die quantitative Forschung (1/2)

1. Welche Arten von Daten gibt es?

Nach Art der Erhebung:
- Primärdaten (Erhebung neuer Daten z.B. durch Befragung)
- Sekundärdaten (Bereits vorhanden, z.B. ‚Eurostat')

Nach Art der Beschaffenheit:
- Querschnittsdaten (fixer Zeitpunkt, verschiedene Auspr.)
- Längsschnittdaten (Zeitreihen einer fixen Ausprägung)
- Paneldaten (Zeitreihe x Ausprägungen)

2. Typische Fehler in Datensätzen

Selection Bias:
- Verzerrung bei der Auswahl der Merkmalsträger

Sampling Bias:
- Nicht-zufällige Auswahl der Merkmalsträger

Measurement Bias:
- Verzerrungen aufgrund fehlerhafter Messungen

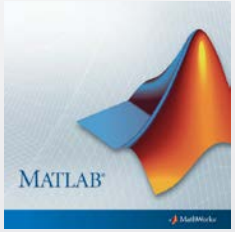
Incubation Bias:
- Nur „Erfolgsmeldungen"

Survivorship Bias:
- Konzentration auf noch existierende Merkmalsträger

Generell sind bei der Anwendung der nicht-parametrischen Methoden der „Selection Bias" und „Sampling Bias" eher unproblematisch.
(Bsp. ökol. Anwendung der DEA auf irische Farm-Level Daten: vor allem Ausreißer und (über die Zeit) Ausscheiden von Farmen aus der Messung)

Praktische Einführung in die quantitative Forschung (2/2)

Welche Statistiktools eignen sich für quantitative Forschung?



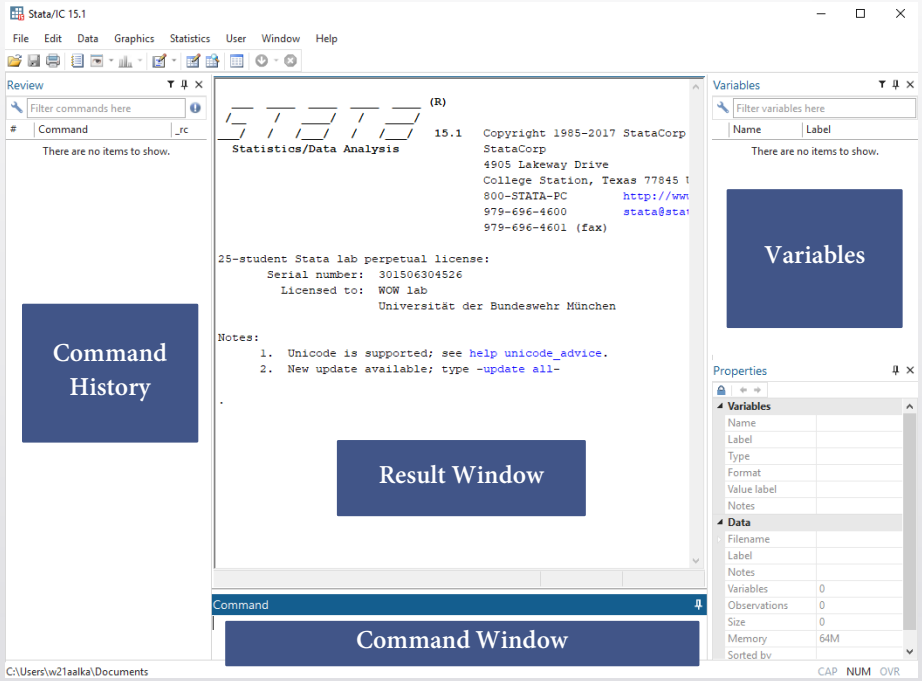
Mächtiger / Open Source

Benutzerfreundliche Bedienung (Programmiergrad)

Eine grundlegende Anwendung der vorgestellten Methoden (DEA: CCR, BCC-Modell und FDH: Orderalpha) ist mit STATA möglich. Bei komplexeren Betrachtungen (Zeitreihen und Anwendung Malmquist / Malmquist-Luenberger Index) muss auf Matlab umgestiegen werden.

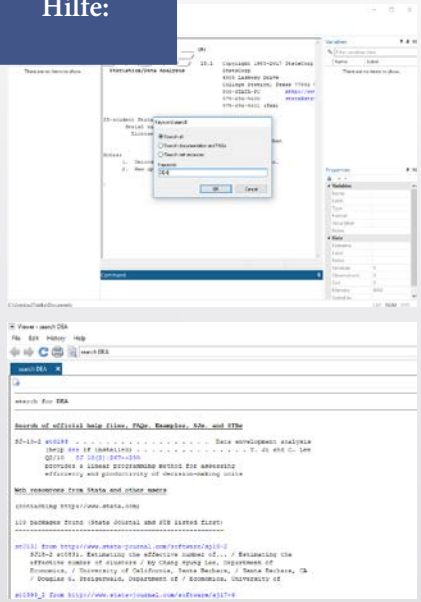
Einführung in die Arbeit mit Stata (1/2)

Benutzer-oberfläche:



Einführung in die Arbeit mit Stata (2/2)

Hilfe:



Im Menüreiter **‚help'** können Suchbegriffe, z.B. ein Befehl wie > dea < eingegeben werden

➔

dann werden u.a. ‚help files', ‚packages' und ‚paper' zu dem Suchwort angezeigt.

Zudem finden sich zu den meisten gängigen Problemen Lösungen im **‚The Stata Blog'.**

Genauso können Paper im **‚The Stata Journal'** den Umgang mit externen packages erleichtern. (https://www.stata-journal.com/article.html?article=st0193)

Kein Tutorial und kein Seminar kann vorbeugen, dass bei der Arbeit mit Statistiksoftware plötzlich ein Fehler auftritt für den es scheinbar keine Lösung gibt. In diesem Fall hilft es, sich per Suchfunktion, Stata-Blog und Journal zu behelfen.

Import von Daten (1/4)

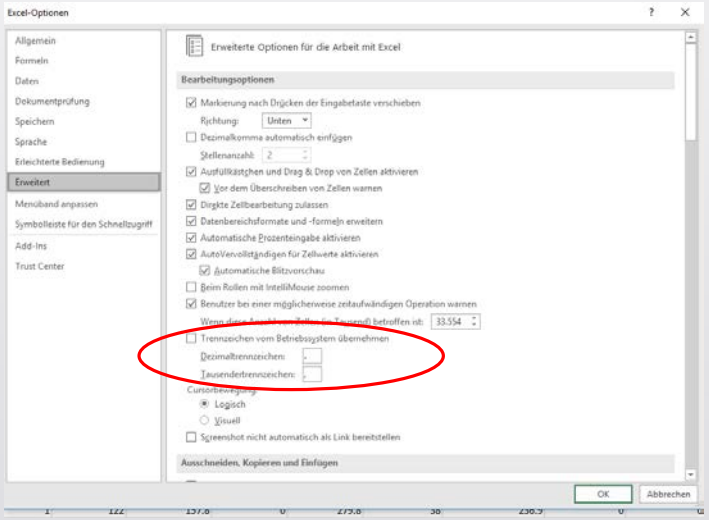
Der Import folgender Dateiformate ist möglich:

1. .dta (Stata Standard)
2. .xls & .xlsx (Excel)
3. .csv
4. .txt

Wenn nicht .dta genutzt wird müssen die Datensätze vorher aufbereitet werden

Bsp: Import von Daten aus Excel-Arbeitsmappe:

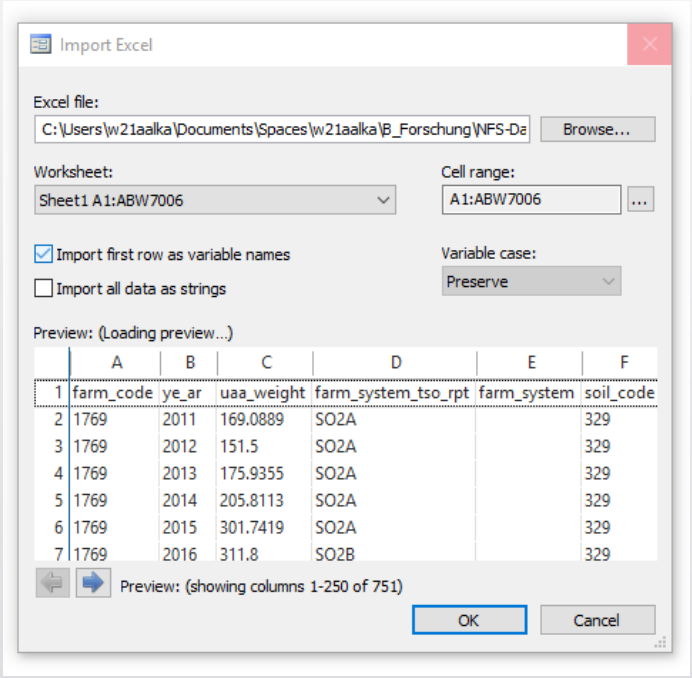
Stata kann nur Zahlen als solche interpretieren, wenn sie mit dem Dezimaltrennzeichen >.< und dem Tausender-trennzeichen >,< gegeben sind.



Datei > Optionen > Erweitert

Import von Daten (2/4)

In STATA: File > Import > Excel Spreadsheet > Browse > Öffnen



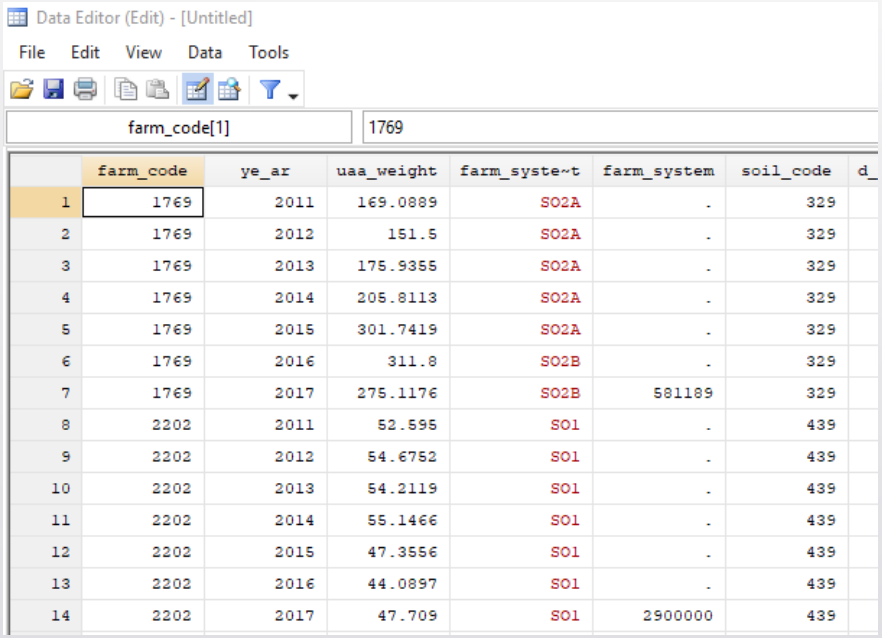
	A	B	C	D	E	F
1	farm_code	ye_ar	uaa_weight	farm_system_tso_rpt	farm_system	soil_code
2	1769	2011	169.0889	SO2A		329
3	1769	2012	151.5	SO2A		329
4	1769	2013	175.9355	SO2A		329
5	1769	2014	205.8113	SO2A		329
6	1769	2015	301.7419	SO2A		329
7	1769	2016	311.8	SO2B		329

Als 1. macht es Sinn das richtige **Worksheet** zu bestimmen. Möglicherweise habt ihr in Excel die Daten nach Jahren auf mehrere „Arbeitsblätter“ aufgeteilt.

Wenn ihr euren Variablen bereits im Excel Sheet einen Namen zugewiesen habt, könnt ihr mit
> Import first row as variable names <
etwas Arbeit sparen und habt zudem keine ungewünschten ‚string‘ Variablen (nicht als Zahlen erkannte), welche beim Ausführen eines Befehls eine Fehlermeldung generieren würden

Import von Daten (3/4)

In STATA: Data > Data Editor > Data Editor (Edit)



	farm_code	ye_ar	uaa_weight	farm_syste~t	farm_system	soil_code	d_
1	1769	2011	169.0889	SO2A	.	329	
2	1769	2012	151.5	SO2A	.	329	
3	1769	2013	175.9355	SO2A	.	329	
4	1769	2014	205.8113	SO2A	.	329	
5	1769	2015	301.7419	SO2A	.	329	
6	1769	2016	311.8	SO2B	.	329	
7	1769	2017	275.1176	SO2B	581189	329	
8	2202	2011	52.595	SO1	.	439	
9	2202	2012	54.6752	SO1	.	439	
10	2202	2013	54.2119	SO1	.	439	
11	2202	2014	55.1466	SO1	.	439	
12	2202	2015	47.3556	SO1	.	439	
13	2202	2016	44.0897	SO1	.	439	
14	2202	2017	47.709	SO1	2900000	439	

Im Data Editor könnt ihr eure erfolgreich importierten Daten noch einmal auf Korrektheit überprüfen.

Im Bild links sieht alles korrekt aus.

Die einzigen ‚Strings' finden wir unter ‚farm syste~t'. Das ist auch gut so, denn es handelt sich hier um einen Code aus Buchstaben.

Für die leeren Beobachtungen unter ‚farm_system' setzt STATA einen >.< Diese Zellen werden bei Berechnungen, z.B. des Mittelwertes dann auch als leere Zellen behandelt.

Praktische Einführung in die quantitative Forschung (4/4)

Aufgabe 1: Importiert eure mitgebrachten Datensätze in STATA

Befehle zum Umgang mit Datenreihen und arithmetische Manipulationen (1/2)

Befehle zum Umgang mit Datenreihen

Alle der folgenden Befehle beziehen sich auf die Variable ‚farm_system', welche durch jede andere syntaktisch korrekte ‚Variablenbezeichnung' substituiert werden kann

- Löschen einer Datenreihe:
  > drop farm_system <
- Datenreihe umbenennen:
  > rename farm_system betriebsart
- Datenreihen aufsteigend nach Namen sortieren:
  > aorder_all <
- Datenreihen nach einer Datenreihe sortieren
  > sort farm_system
- Alle Datenreihen des Datensatzes löschen:
  > clear < oder > drop _all <
- Einzelne Variablen behalten:
  > keep farm_system

Befehle zu arithmetischen Manipulationen

Erstellen einer neuen Variable:
> gen neu = <

- Addition > = alt1 + alt2
- Subtraktion > = alt1 – alt2
- Multiplikation > = alt1 * 100
- Division > = alt1 / alt2
- Potenzrechnung > = alt1^2

Außerdem:

Mit dem Befehl > sysuse auto.dta < kann ein Datensatz geladen werden, welcher häufig in Foren oder Papern als Beispiel dient

Befehle zum Umgang mit Datenreihen und arithmetische Manipulationen (2/2)

Aufgabe 2:

1. Löscht euren Datensatz
2. Ladet die ‚auto.dta' Daten
3. Wandelt mpg (Miles per Gallon) in Liter je 100 km um
4. ..und generiert dabei die entsprechende Variable
5. Sortiert den Datensatz nach der neuen Variable
6. Welcher Wagen hat die größte Reichweite und wie viele Liter benötigt dieser für 100 km?

Voranalyse von Datensätzen (1/3)

Deskriptive Statistiken

Instrumente:

- Anzahl der Beobachtungen (Prüfen auf Vollständigkeit)
- Mittelwert, Median (Lage)
- Minimum, Maximum (Spannweite)
- Quantile, Varianz, Standardabweichung zur Bestimmung der Streuung

Diese „einfachen“ deskriptiven Statistiken können für eine Variable mit dem Befehl > summarize price < bzw. > summarize price, detail < generiert werden (für alle Variablen: einfach Variablenbezeichnung weglassen)

Grafische Analysen

Instrumente:

- Box-Whisker-Plot zur Bestimmung von Lage, Spannweite und Streuung der Beobachtungen
- Histogramm (Symmetrie und Form der Verteilung)
- Scatterplots zur Bestimmung von Zusammenhängen (Korrelationen)
- Liniendiagramme zum Abbilden von Trends bei Zeitreihen

Voranalyse von Datensätzen (2/3)

Aufgabe 3:

1. Ladet euren eigenen Datensatz
2. Führt deskriptive Statistiken zu euren Daten durch
3. Wie sind eure Datensätze beschaffen?
   (Normalverteilt, gestreut etc. etc.)

Voranalyse von Datensätzen (3/3)

Deskriptive Statistiken

Instrumente:
Anzahl der Beobachtungen (Prüfen auf Vollständigkeit)
Mittelwert, Median (Lage)
Minimum, Maximum (Spannweite)
Quantile, Varianz, Standardabweichung zur Bestimmung der Streuung

Diese „einfachen“ deskriptiven Statistiken können für eine Variable mit dem Befehl > summarize farm_system < bzw. > summarize farm_system, detail < generiert werden (für alle Variablen: einfach Variablenbezeichnung weglassen)

Grafische Analysen

Instrumente:

- Box-Whisker-Plot zur Bestimmung von Lage, Spannweite und Streuung der Beobachtungen
  > graph box price <
- Histogramm (Symmetrie und Form der Verteilung)
  > histogram price <
- Scatterplots zur Bestimmung von Zusammenhängen (Korrelationen) > scatter var_y var_x <
- Liniendiagramme zum Abbilden von Trends bei Zeitreihen > graph twoway line var1 var2 <

Voranalyse von Datensätzen (2/3)

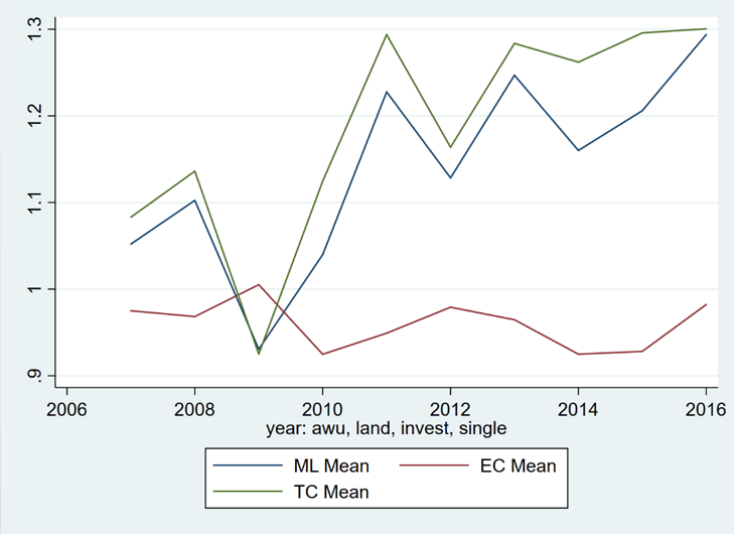
Aufgabe 4:

1. Erstellt Scatterplots für eure Inputs und Outputs
2. Welchen Zusammenhang vermutet ihr? Wie wird sich der Zusammenhang auf eure DEA Ergebnisse auswirken?

Voranalyse von Datensätzen: Exkurs (1/2)

Bsp: graph twoway line anhand Malmquist-Luenberger Indizes für Irish Dairy Farms von 2006 - 2016

- Frage: Ist die Effizienz Irischer Milchbetriebe zwischen 2006 und 2016 gestiegen, wenn man Nitratüberschüsse als ‚bad output' integriert
- Anwendung nicht-parametrischer Methoden
- DDF erlaubt die Integration von ‚ecological pressures' als output
- Malmquist-Luenberger Index erlaubt Analyse über die Zeit
- n = 30 (Dairy farms mit vollständigen Beobachtungen zu den 4 Inputs: „annual working units, land, investment in machinery, single farm payments" und vollständigen Beobachtungen zu dem Output: „farm gross output" und dem negativen Output: „nutrient surplus"



Voranalyse von Datensätzen: Exkurs (2/2)

Bsp: graph twoway line anhand Malmquist-Luenberger Indizes für Irish Dairy Farms von 2006 - 2016

- Bei der Berechnung wird die Technologie „Directional Distance Function“ genutzt
- Vorgehensweise aber analog zur DEA: Die Produktivitäten der 30 Farms werden für eine Beobachtungsperiode (ein Jahr) verglichen. Die Effizientesten konstituieren die Efficiency Frontier und dienen als Referenzpunkt.
- Der MLI vergleicht nun die Änderungen in den Efficiency Scores aller DMUs über zwei Perioden (2007 zu 2006 ; 2008 zu 2006 , …, 2016 zu 2006)
- Für MLI > 1 hat die Effizienz zugenommen
- EC beschreibt dabei den „Catch-Up“ Effekt, also wie viele DMUs zur Frontier aufschließen konnten
- TC beschreibt dabei den „Frontier-Shift“ Effekt, also ob sich die Effizientesten weiter (durch Technological Change) verbessern konnten