

Missing Data

Regressionsmodelle für Politikwissenschaftler

Wiederholung
Missing Data

TSCS Regression in Stata

- ▶ Struktur definieren:

```
. xtset id year
    panel variable:  id (strongly balanced)
    time variable:  year, 1996 to 2004, but with gaps
                   delta: 1 unit
```

xtdescribe

```
. xtdescribe

      id: 1, 2, ..., 209                n =      209
     year: 1996, 1998, ..., 2004        T =        5
      Delta(year) = 2 units
      Span(year) = 5 periods
      (id*year uniquely identifies each observation)

Distribution of T_i:  min      5%      25%      50%      75%      95%      max
                   5         5         5         5         5         5         5

      Freq.  Percent  Cum. | Pattern
-----+-----
      209    100.00  100.00 | 11111
-----+-----
      209    100.00      | XXXXX
```

Regression Korruption/Accountability

```
. reg cc va

      Source |          SS      df      MS                Number of obs =      917
-----+-----+-----+-----+-----+-----+-----
      Model | 483.021235      1 483.021235                F( 1, 915) = 1028.05
      Residual | 429.9069      915  .469843606                Prob > F      = 0.0000
-----+-----+-----+-----+-----+-----
      Total | 912.928135      916  .996646436                R-squared     = 0.5291
                                           Adj R-squared = 0.5286
                                           Root MSE    = .68545

-----+-----+-----+-----+-----+-----
      cc |          Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----+-----+-----+-----+-----
      va |   .733515   .0228772    32.06   0.000   .6886171   .7784128
      _cons |   .0202968   .0226481     0.90   0.370  -.0241515   .0647451
-----+-----+-----+-----+-----+-----
```

TSCS Regression

```
. xtpcse cc va
```

```
Number of gaps in sample: 714
(note: the number of observations per panel, e(n_sigma) = 1.029556650246305,
      used to compute the disturbance of covariance matrix e(Sigma)
      is less than half of the average number of observations per panel,
      e(n_avg) = 4.5172414; you may want to consider the pairwise option)
```

Linear regression, correlated panels corrected standard errors (PCSEs)

```
Group variable:  id                Number of obs    =    917
Time variable:  year              Number of groups  =    203
Panels:         correlated (unbalanced)  Obs per group: min =     1
Autocorrelation: no autocorrelation      avg = 4.517241
Sigma computed by casewise selection      max =     5
Estimated covariances =    20706        R-squared         =    0.5291
Estimated autocorrelations =     0      Wald chi2(1)     =    509.80
Estimated coefficients =     2          Prob > chi2      =    0.0000
```

	Panel-corrected					
cc	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
va	.733515	.0324868	22.58	0.000	.669842	.797188
_cons	.0202968	.0115091	1.76	0.078	-.0022606	.0428541

Regressionsmodelle für Politikwissenschaftler Missing Data (5/30)

?

```
. summ cc
```

Variable	Obs	Mean	Std. Dev.	Min	Max
cc	923	.0000867	.9981332	-2.05	2.58

```
. summ va
```

Variable	Obs	Mean	Std. Dev.	Min	Max
va	982	.0000815	.9980244	-2.32	1.76

Was meinen wir mit „missing“?

1. unit nonresponse: Ausgewählte Personen nehmen nicht an Umfrage teil
 2. item nonresponse: Antworten fehlen für einzelne Fragen
 3. missing by design: Fragen werden nur einer zufällig ausgewählten Teil-Stichprobe gestellt
- ▶ In dieser Sitzung primär item nonresponse

Was tun bei unit nonresponse?

- ▶ Typischer Fall: Zuwenig Niedriggebildete in der Stichprobe
- ▶ Standardprozedur: Repräsentativgewichtung, Anpassung der Randverteilung **bekannt**er Merkmale
- ▶ (Designgewichtung?)
- ▶ Hoffnung: Gewichtung paßt auch Verteilung unbekannter Merkmale an (Voraussetzung?)
- ▶ In der Praxis oft kaum Unterschiede, vor allem, wenn Gewichtungsvariablen als unabhängige Variablen fungieren

Welche Arten von item nonresponse gibt es?

Bezeichnung	Bedeutung	Beispiel
„Missing Completely At Random“ (MCAR)	Ausfall von y ist unabhängig vom Wert y und vom Wert anderer Variablen ($x_1 \dots$)	Übertragungsfehler beim Eingeben der Fragebögen
„Missing At Random“ (MAR)	Ausfall von y ist unabhängig vom Wert y , wird aber vom Wert anderer Variablen ($x_1 \dots$) beeinflusst	Niedriges Politikinteresse führt zu Ausfällen bei Fragen, die sich auf Politik beziehen
„Non-Ignorable“ (NI)/ „Not Missing At Random“ (NMAR)	Ausfall von y wird vom Wert von y und/oder nicht beobachteten Variablen beeinflusst	Antwortausfall bei heiklen Fragen

Welche Arten von item nonresponse gibt es?

- ▶ Keine Prüfmöglichkeit
- ▶ MCAR in der Regel völlig unrealistisch
- ▶ $MAR \approx$ „ignorable“: Ausfallmechanismus muß nicht modelliert werden
- ▶ (Missingness selbst **nicht** ignorieren!)
- ▶ NI: Ausfallmechanismus muß Bestandteil des substantiellen Modells sein (geringe praktische Relevanz)
- ▶ In der Praxis Kontinuum zwischen MAR und NI (Einkommen)
- ▶ Alle Techniken sind nur Krücken, Missingness vermeiden

Was kann man tun?

1. Listwise Deletion
2. Pairwise Deletion
3. Dummy Variable Adjustment
4. (Konventionelle) Imputation
5. Full Information Likelihood (FIML)
6. Multiple Imputation

Was ist listwise deletion?

- ▶ Analyse vollständiger Fälle, in Standardprogrammen Voreinstellung
- ▶ „Listwise deletion is evil“?
- ▶ Korrekte Schätzungen/Standardfehler wenn MCAR (Stichprobe aus Stichprobe)
- ▶ Bei MAR Verzerrungen möglich
- ▶ Relativ unproblematisch für missingness der unabhängigen Variablen (wenn nicht vom Wert der abhängigen beeinflusst) (geschichtete Stichprobe)
- ▶ Für logistische Regression sogar missingness der Abhängigen unproblematisch, wenn nicht von unabhängigen beeinflusst
- ▶ Aber: Bei komplexeren Modellen dramatische Reduktion der Stichprobegröße
- ▶ Beispiel fünf Prozent missingness, 20 Variablen: $0,95^{20} \approx 0,36$

Was ist pairwise deletion?

- ▶ Zur Schätzung vieler linearer Modelle genügt statt Rohdaten Kovarianzmatrix
- ▶ Pairwise Deletion: Für jede Kovarianz alle verfügbaren Fälle verwenden → mehr Fälle als bei l.d. → unterschiedliche Fallzahlen
- ▶ Ambiguitäten bei Tests etc.; Standardfehler nicht korrekt
- ▶ Bei ernsthafter Missingness Inkonsistenzen möglich/wahrscheinlich; Modell nicht schätzbar

Was ist Dummy Variable Adjustment?

- ▶ Fehlende Werte für unabhängige Variable x →
- ▶ Dummy d für Missingness
- ▶ ergänzte Variable x^* mit konstantem Wert, der fehlende Werte ersetzt
- ▶ Regression von y auf d und x^*
- ▶ Alle Fälle werden genutzt, einfache Interpretation
- ▶ Leider selbst bei MCAR sehr stark verzerrte Werte möglich (Beispiel im Text)

Was ist Imputation?

- ▶ Fehlende Werte werden ersetzt
- ▶ Z. B. durch Mittelwert (ganz schlecht) oder durch Regression auf beobachtete Werte (funktioniert bei MCAR)
- ▶ Wird kompliziert, wenn mehrere Variablen betroffen sind
- ▶ Generell sind Standardfehler zu optimistisch, weil Imputation wie reale Daten betrachtet werden

Was ist das Zwischenfazit?

- ▶ Konventionelle Methoden machen die Sache oft noch schlimmer
 - ▶ bias
 - ▶ Falsche (optimistische) Standardfehler
 - ▶ Sehr anfällig, wenn Daten NI
- ▶ Unter konventionellen Ansätzen listwise deletion oft die am wenigsten schlechte Alternative
- ▶ Es geht auch besser

Wie kann ML hier helfen?

- ▶ ML findet gute Parameterschätzungen, indem (Log-)Likelihood-Funktion maximiert wird
- ▶ (Log-)Likelihood ist eine (modellspezifische) Funktion der Daten und der Vermutungen über den Wert der Parameter
- ▶ Fallweise Berechnung, Multiplikation (Fälle unabhängig)
- ▶ Wenn MAR gilt, kann für fehlende Werte
- ▶ die Summe beziehungsweise das Integral der Likelihood-Funktion über die möglichen Werte eingesetzt werden

Probleme/Komplikationen?

- ▶ Besondere Algorithmen, Vorkehrungen für Standardfehler
- ▶ Konkrete Vorgehensweise hängt vom Modell und vom Muster der Ausfälle ab →
- ▶ Problem: Modell der gemeinsamen Verteilung aller Variablen mit fehlenden Werten erforderlich
- ▶ Methode muß für jedes statistische Modell gesondert implementiert werden

Was sind die Vorteile der multiplen Imputation (MI)?

- ▶ Hat dieselben optimalen Eigenschaften wie ML
- ▶ Kann mit (praktisch) jedem statistischen Modell kombiniert werden
- ▶ Extrem flexibel solange MAR
- ▶ Anwendung mit Standardsoftware (SPSS, STATA) relativ leicht möglich (Zusatzmodule)

Wo ist der Haken?

- ▶ Vor der Analyse mit Standardsoftware Einsatz spezieller Programme notwendig
- ▶ (Etwas) mühsam, zeit- und rechenaufwendig
- ▶ Organisationsaufwand, Fehleranfälligkeit, wenn keine speziellen Erweiterungen für Standardsoftware genutzt werden
- ▶ Auch hier: Vorüberlegungen, Zweifelsfälle
- ▶ Kommunikation der Ergebnisse nicht unproblematisch

Wie funktioniert die einfache random imputation?

Beispiel aus dem Text:

- ▶ x und y sind bivariat standard-normalverteilt mit einer Korrelation von 0,3
- ▶ Die Hälfte der x wird zufällig gelöscht (MCAR)
- ▶ Fehlende Werte von x durch Regression auf y ersetzen → Analyse der komplettierten Daten → Korrelation von 0,42
- ▶ Warum? → Vorhergesagte Werte berücksichtigen nur systematischen Teil (deterministischer Zusammenhang)
- ▶ bias, da zuwenig Streuung für imputierte Werte
- ▶ Zufällig Werte aus der Verteilung der Residuen (von x auf y) ziehen und zu imputierten Werten addieren
- ▶ Bias verschwindet fast vollständig, da nun Verteilung der imputierten Werte mit Verteilung der beobachteten Werte identisch

Wie funktioniert die multiple random imputation?

- ▶ Problem: (Zufällig) imputierte Daten werden wie reale Daten behandelt → Standardfehler zu klein
- ▶ Wie kann man Unsicherheit über fehlende Werte berücksichtigen? → multiple Imputation, z. B. acht Datensätze
- ▶ Wegen zufälliger Komponente unterscheiden sich Datensätze
 - ▶ Wenn Unsicherheit über fehlende Werte gering, sind Datensätze fast identisch
 - ▶ Je größer die Unsicherheit, desto stärker die Differenzen

Wie kommt man zu Ergebnissen?

- ▶ Analyse jedes einzelnen Datensatzes in Standardprogramm
- ▶ Parameterschätzung: Arithmetischen Mittelwert über acht Einzelschätzungen bilden
- ▶ Standardfehler: Anwendung der „Rubin-Regel“

Wie werden die Standardfehler berechnet?

$$\sqrt{V(\bar{r})} = \sqrt{\frac{1}{M} \sum_{j=1}^M s_j^2 + \left(1 + \frac{1}{M}\right) \left(\frac{1}{M-1}\right) \sum_{j=1}^M (r_j - \bar{r})^2}$$

- ▶ $\sqrt{V(\bar{r})}$: Korrigierter Standardfehler des Parameters
- ▶ M : Zahl der imputierten Datensätze
- ▶ s_j^2 : Schätzung für Varianz des Parameters auf Basis der j -ten Imputation
- ▶ $\frac{1}{M-1} \sum_{j=1}^M (r_j - \bar{r})^2$: Varianz der Parameterschätzungen
- ▶ $1 + \frac{1}{M}$: Korrekturfaktor

Was war da mit Variation der Parameterschätzungen?

- ▶ Fehlende Werte von x werden durch Regressionsmodell $x = \beta_0 + \beta_1 y + \epsilon$ ersetzt, das zufällige Komponente berücksichtigt
- ▶ Aber: Modellparameter ($\beta_0, \beta_1, \sigma_\epsilon^2$) basieren ja selbst nur auf Schätzungen
- ▶ Müssen deshalb über Schätzungen variieren
- ▶ Zufällige Ziehung der Werte aus *Verteilung möglicher Modellparameter*
- ▶ Macht bei hoher Rate von missingness einen Unterschied (größere korrigierte Standardfehler)

Was passiert in komplexeren Fällen?

- ▶ Wir brauchen ein Imputationsmodell
- ▶ Meistens: Multivariates normales Modell
 - ▶ Alle Variablen sind normalverteilt
 - ▶ Jede Variable ist als Linearkombination aus den übrigen Variablen und einem homoskedastischen Fehlerterm (ϵ) darstellbar
- ▶ In der Regel völlig unrealistisch
- ▶ Aber als Imputationsmodell relativ robust; Transformationen

Was macht der Computer konkret?

- ▶ Dummerweise ist die Verteilung der Parameter nicht bekannt
- ▶ Kann wegen der fehlenden Werte nicht mal korrekt geschätzt werden
- ▶ Iterative Algorithmen die zwischen
 - ▶ Zufälligen Ziehungen aus der Verteilung der Parameter und
 - ▶ Zufälligen Ziehungen aus der Verteilung der fehlenden Werte pendeln
- ▶ Wenn der Algorithmus konvergiert, zufällige Ziehungen aus der gemeinsamen Verteilung von Daten und Parametern
- ▶ In Abhängigkeit von beobachteten Werten (MAR)
- ▶ Verteilungen in der Regel nicht analytisch darstellbar, Zugriff über Simulationsverfahren
- ▶ Immenser numerischer Aufwand

Was ist MICE?

- ▶ Multivariates normales Modell: Gemeinsames Imputationsmodell für alle Variablen
- ▶ Multiple Imputation by Chained Equations:
 - ▶ Für jede Variable mit fehlenden Werten individuelles Imputationsmodell
 - ▶ Lineare Regression, Logit, Probit, Poisson. . .
 - ▶ Starke Annahme: Individuelle Verteilungen sind miteinander kompatibel & brauchbare Approximation für *gemeinsame* Verteilung
- ▶ Sehr flexibel, Implementation in Stata, Behandlung von Dummies und transformierten Variablen

Was sind die Hauptergebnisse?

- ▶ Item nonresponse als Problem wird unterschätzt und sollte vermieden werden
- ▶ Für einfache Modelle mit wenig missingness ist i.d. eine brauchbare Lösung
- ▶ Andere konventionelle Ansätze vermeiden
- ▶ Für komplexere Modelle Likelihood-basierte Lösung oder MI um
 - ▶ Vorhandene Daten auszuschöpfen und
 - ▶ Korrekte Parameterschätzungen/Standardfehler zu erhalten
- ▶ Möglichst viel zusätzliche Information nutzen, um MAR-Bedingung realistischer zu machen (Imputationsmodell \neq Analysemodell)

Übung für heute

- ▶ Laden Sie das Stata Skript <http://www.kai-arzheimer.com/Lehre-Regression/simulation.do> von der Homepage herunter
- ▶ Sie können damit das Beispiel zur einfache random imputation aus dem Text nachvollziehen
- ▶ Lesen Sie die Kommentare und versuchen Sie, den Code zu verstehen
- ▶ Probieren Sie die Befehle aus