Statistische Methoden der Lebensverlaufsforschung

Prof. Dr. Johannes Giesecke
Humboldt-Universität zu Berlin
Institut für Sozialwissenschaften

Lebensverlauf

Deutschland , ca 1870

Analyseziele

1. Beschreibung individueller bzw. gruppen-spezifischer Verläufe
2. Zusammenhänge der Verläufe mit anderen (sozio-strukturellen) Merkmalen
3. Schätzung von Kausaleffekten

Beispiel für Forschungsfrage

Führen Armutserfahrungen zu schlechterer Gesundheit?

damit verbundene Unterfragen (Auswahl):

- Wie entwickelt sich Gesundheit über den Lebensverlauf?
- Wie hängen Armutserfahrungen mit Gesundheit und deren Entwicklung zusammen? Unterscheiden sich diese Zusammenhänge zwischen bestimmten Bevölkerungsgruppen?
- Wirkt die Erfahrung von Armut kausal auf Gesundheit?

Quer- vs. Längsschnittdesign

Querschnittdesign ungeeignet, diese Fragestellung adäquat zu beantworten

1. keine Analyse von (individuellen oder kohorten-spezifischen) Gesundheitsverläufen möglich
2. simultane Messung von Armutserfahrung und Gesundheit

Quer- vs. Längsschnittdesign

Längsschnittdesign besser geeignet, diese Fragestellung adäquat zu beantworten

Quellen für Längsschnittdaten

1. Retrospektive Längsschnittstudien
2. Prospektive Längsschnittstudien (Panel)
3. Trendstudien (wiederholte Querschnitts-untersuchungen)

Vorteile von Längsschnittdaten

1. Betrachtung dynamischer Zusammenhänge

- Analyse von Stabilität/Veränderung
- kausale Analyse möglich (zeitliche Vorlagerung der kausalen Variablen)
- dynamische Modelle möglich, z.B. $y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \gamma y_{it-1} + u_{it}$

Vorteile von Längsschnittdaten

1. Betrachtung dynamischer Zusammenhänge

- typische Verfahren:
  - Beschreibung von (zeitlicher) Variation
  - Sequenzanalyse
  - Ereignisdatenanalyse
  - dynamische Modelle
  - Growth Curve Models

Vorteile von Längsschnittdaten

1. Betrachtung dynamischer Zusammenhänge

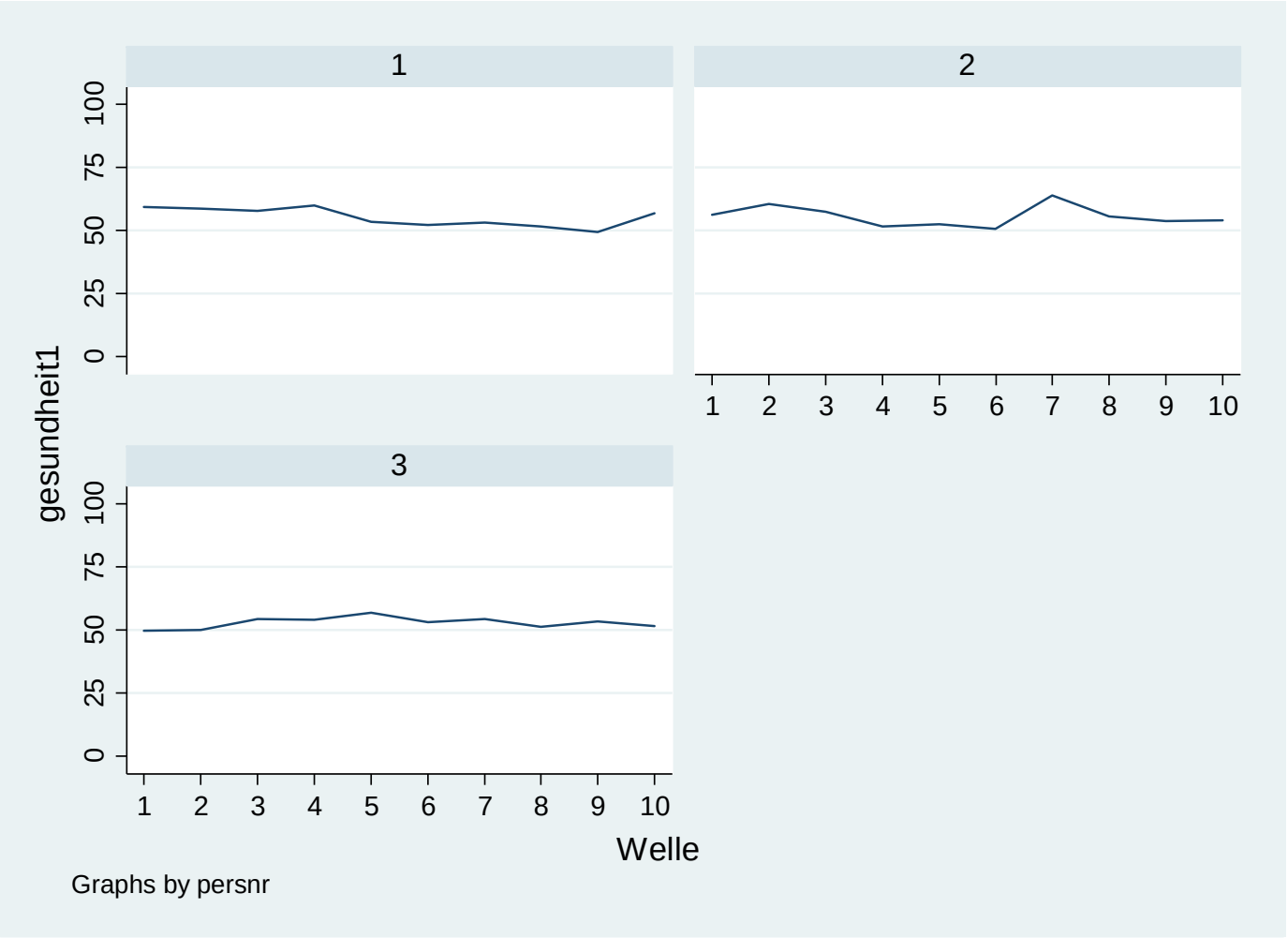
- typische Verfahren:
  - Beschreibung von (zeitlicher) Variation
  - Sequenzanalyse
  - Ereignisdatenanalyse
  - dynamische Modelle
  - Growth Curve Models

Beispiel

```
list persnr Welle  gesundheit1 gesundheit2

     +---------------------------------------+
     | persnr   Welle   gesund~1   gesund~2 |
     |---------------------------------------|
  1. |      1       1   59.40709   39.55777 |
  2. |      1       2   58.61234    74.1507 |
  3. |      1       3   57.85548   59.44157 |
  4. |      1       4   59.82533   63.38122 |
  5. |      1       5   53.50386   64.31553 |
     |---------------------------------------|
  6. |      1       6   52.20303   55.35596 |
  7. |      1       7   53.22916   49.87989 |
  8. |      1       8   51.64936    47.4058 |
  9. |      1       9   49.46273   52.93686 |
 10. |      1      10    56.9523   46.07153 |
     |---------------------------------------|
 11. |      2       1   56.18912   50.62788 |
 12. |      2       2   60.57314   43.20216 |
 13. |      2       3   57.56074   51.93335 |
     ...
```

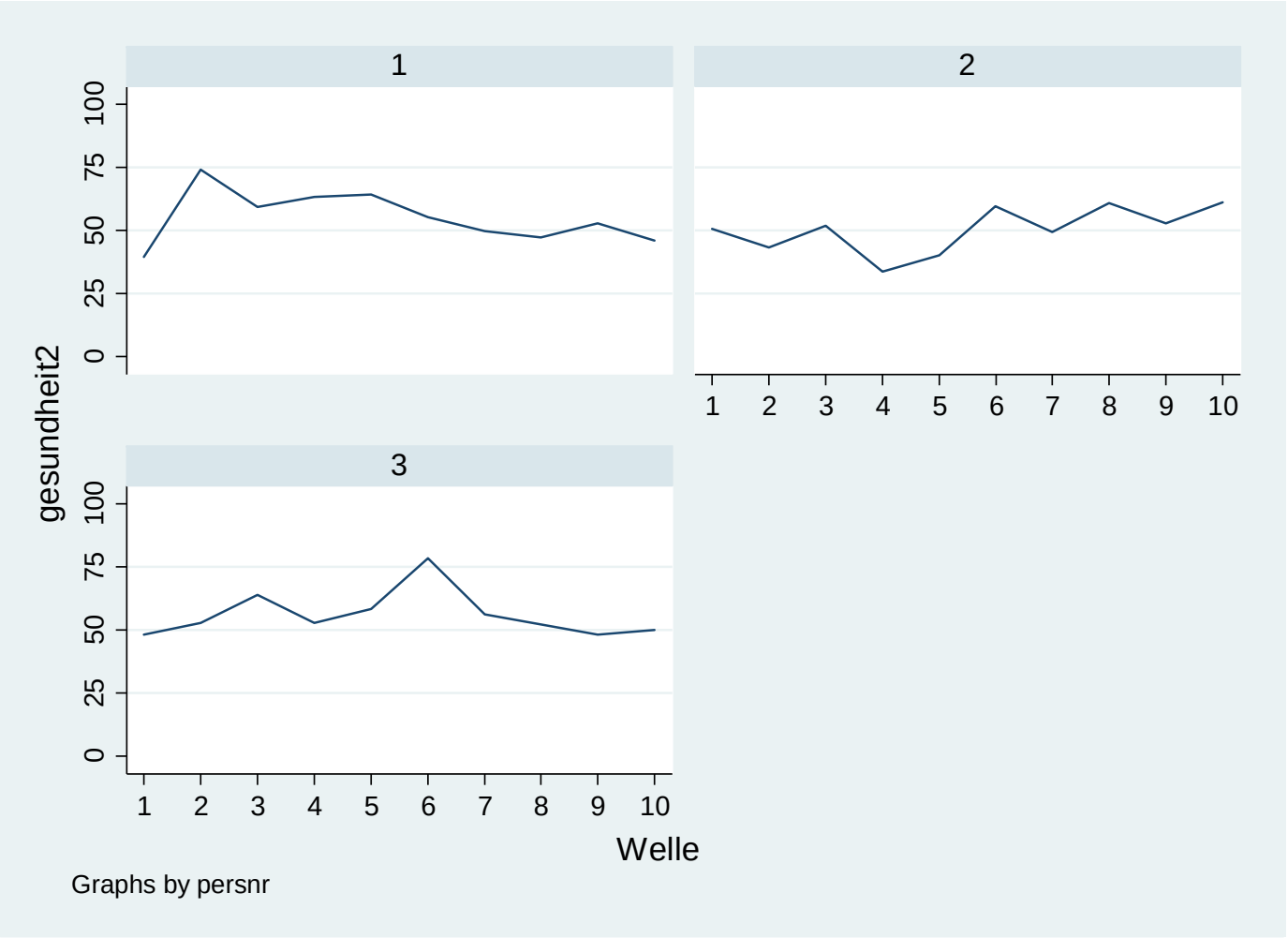
Beispiel

```
xtline  gesundheit1 in 1/30, ylabel(0(25)100) xlabel(1(1)10)
```



Beispiel

```
xtline  gesundheit2 in 1/30, ylabel(0(25)100) xlabel(1(1)10)
```



Beispiel

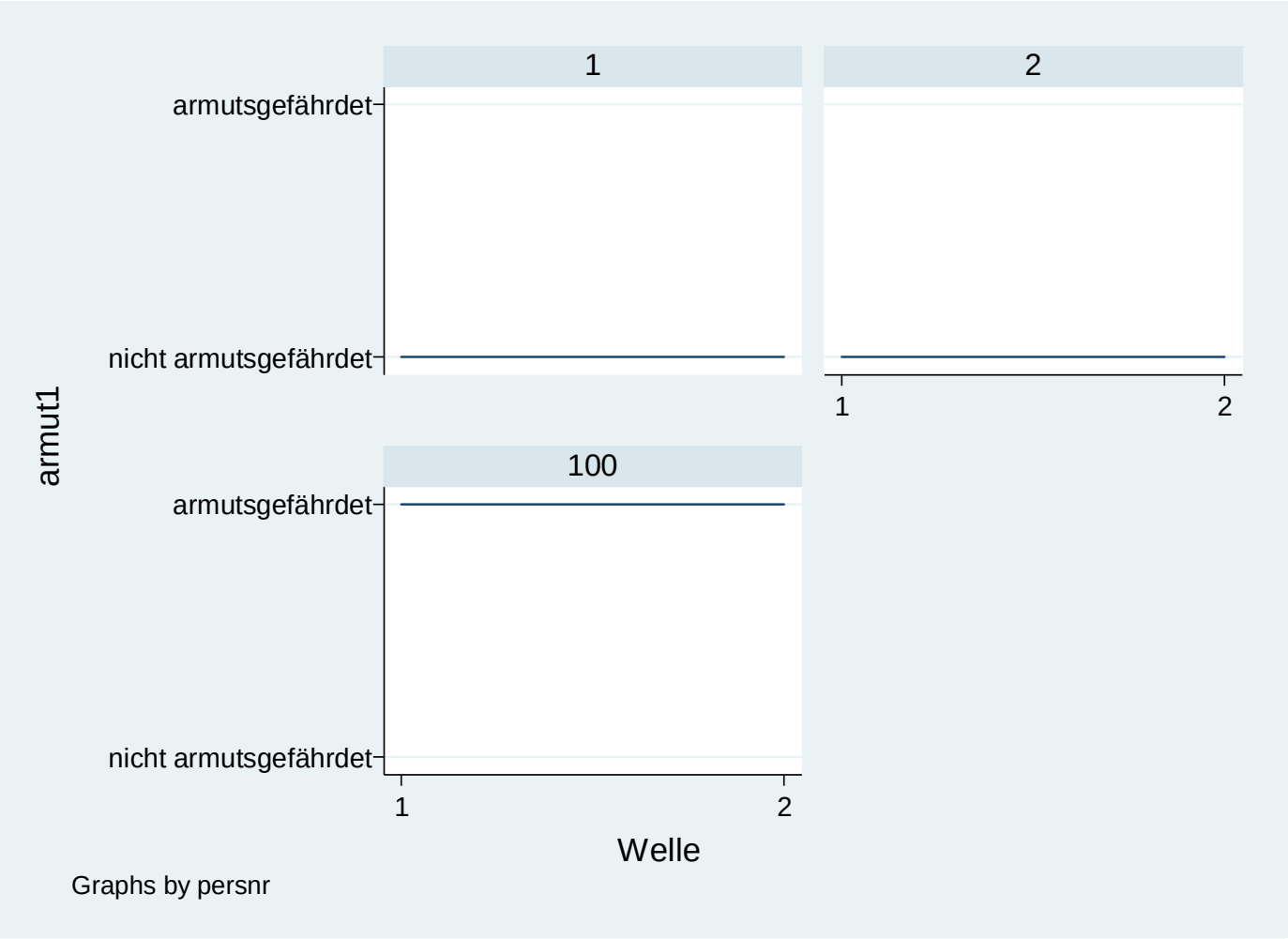
```
. xtsum gesundheit1

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
gesund~1 overall |  43.50789   7.734563   24.55437   63.77614 |     N =    1000
         between |             7.22073    31.2498    56.9231 |     n =     100
         within  |            2.855563   34.90658   52.51814 |     T =      10

. xtsum gesundheit2

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
gesund~2 overall |  43.29218   11.93548   4.454057   80.17522 |     N =    1000
         between |            7.666015   28.42707   59.17701 |     n =     100
         within  |            9.176995   15.12066    75.3243 |     T =      10
```

Beispiel

```
. list persnr Welle armut1 armut2

     +-----------------------------------------------------------------+
     | persnr   Welle                  armut1                   armut2 |
     |-----------------------------------------------------------------|
  1. |      1       1   nicht armutsgefährdet   nicht armutsgefährdet |
  2. |      1       2   nicht armutsgefährdet         armutsgefährdet |
  3. |      2       1   nicht armutsgefährdet   nicht armutsgefährdet |
  4. |      2       2   nicht armutsgefährdet         armutsgefährdet |
  5. |      3       1   nicht armutsgefährdet   nicht armutsgefährdet |
     |-----------------------------------------------------------------|
  6. |      3       2   nicht armutsgefährdet         armutsgefährdet |
     ...

196. |     98       2         armutsgefährdet   nicht armutsgefährdet |
197. |     99       1         armutsgefährdet         armutsgefährdet |
198. |     99       2         armutsgefährdet   nicht armutsgefährdet |
199. |    100       1         armutsgefährdet         armutsgefährdet |
200. |    100       2         armutsgefährdet   nicht armutsgefährdet |
     +-----------------------------------------------------------------+
```

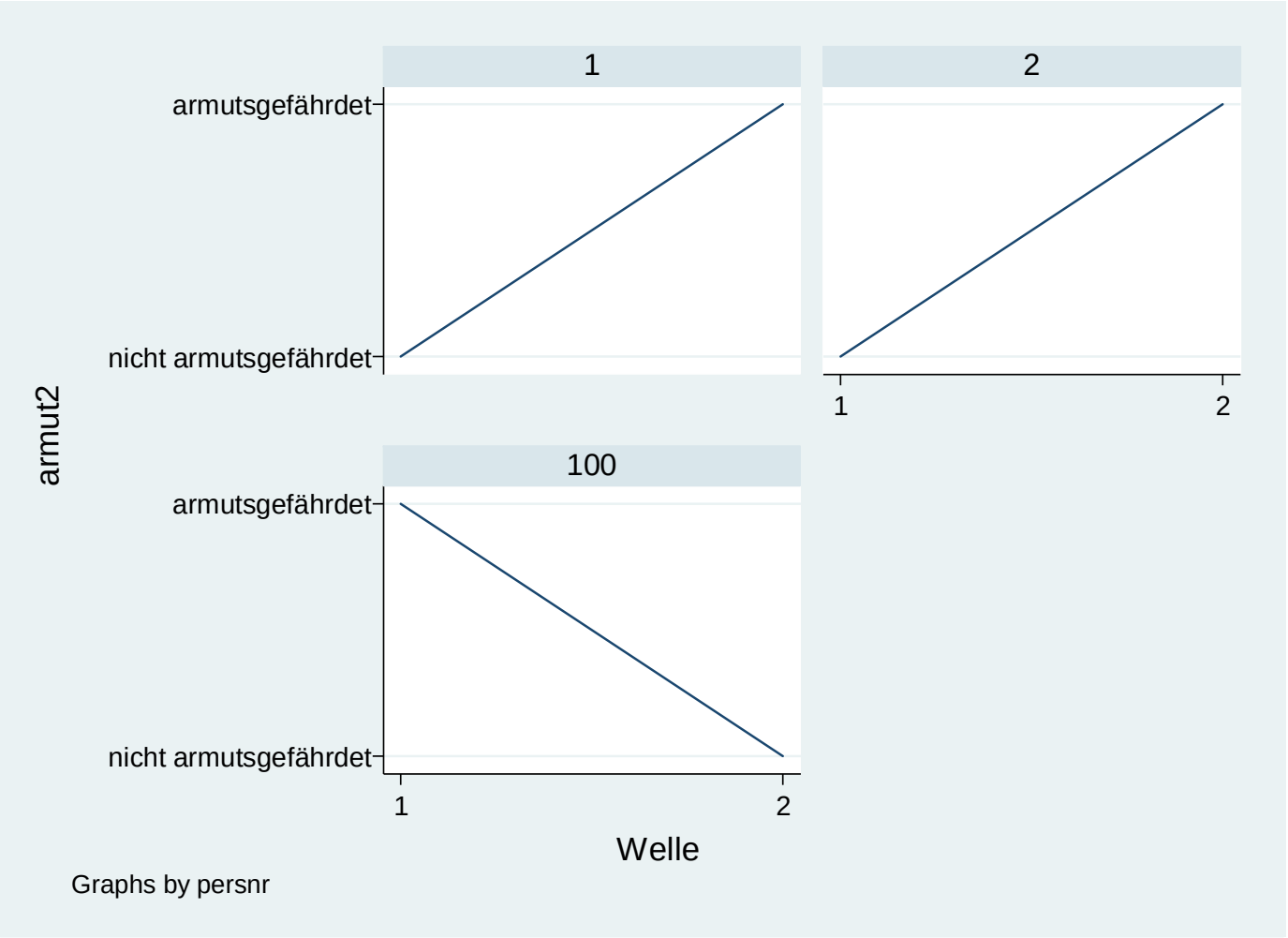
Beispiel

```
xtline  armut1 if persnr<3 | persnr==_N/2, ylabel(0 1, valuelabel
   angle(horizontal)) xlabel(1 2)
```



Beispiel

```
xtline  armut2 if persnr<3 | persnr==_N/2, ylabel(0 1, valuelabel
   angle(horizontal)) xlabel(1 2)
```



Beispiel

```
. xttrans armut1, freq
           |        armut1
    armut1 |         0          1 |     Total
-----------+----------------------+----------
         0 |        50          0 |        50
           |    100.00       0.00 |    100.00
-----------+----------------------+----------
         1 |         0         50 |        50
           |      0.00     100.00 |    100.00
-----------+----------------------+----------
     Total |        50         50 |       100
           |     50.00      50.00 |    100.00

. xttrans armut2, freq
           |        armut2
    armut2 |         0          1 |     Total
-----------+----------------------+----------
         0 |         0         50 |        50
           |      0.00     100.00 |    100.00
-----------+----------------------+----------
         1 |        50          0 |        50
           |    100.00       0.00 |    100.00
-----------+----------------------+----------
     Total |        50         50 |       100
           |     50.00      50.00 |    100.00
```

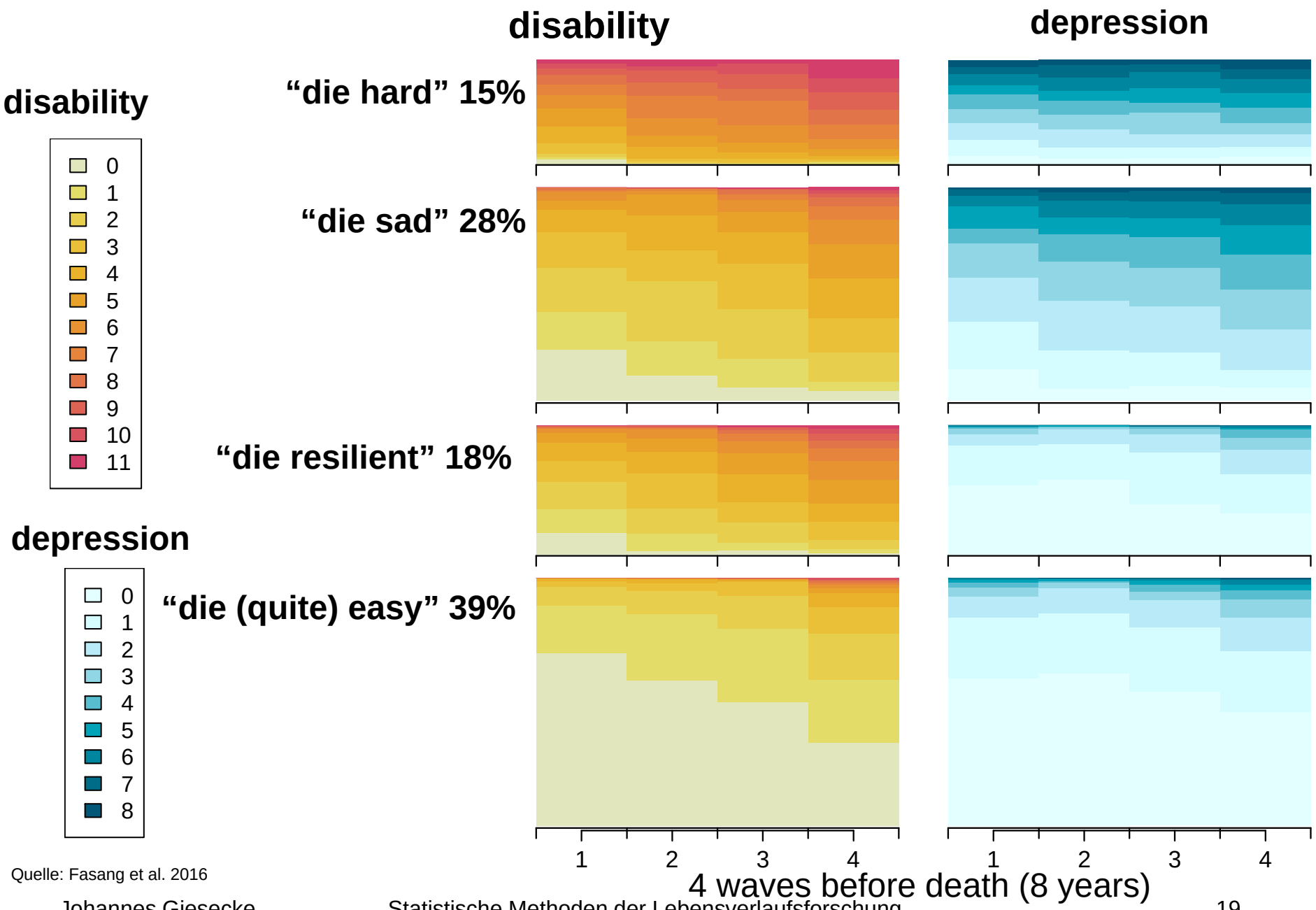
Vorteile von Längsschnittdaten

1. Betrachtung dynamischer Zusammenhänge

- typische Verfahren:
  - Beschreibung von (zeitlicher) Variation
  - Sequenzanalyse

  Versuch, Typen von Verläufen aus den Daten zu extrahieren (z.B. „Stabil Gesunde“, „dauerhaft Krankgewordene“, „Krankheit mit (dauerhafter) Genesung“ etc.)

  Analyse der Verteilung dieser Typen mittels multivariater (Regressions)Modelle



Quelle: Fasang et al. 2016

Vorteile von Längsschnittdaten

1. Betrachtung dynamischer Zusammenhänge

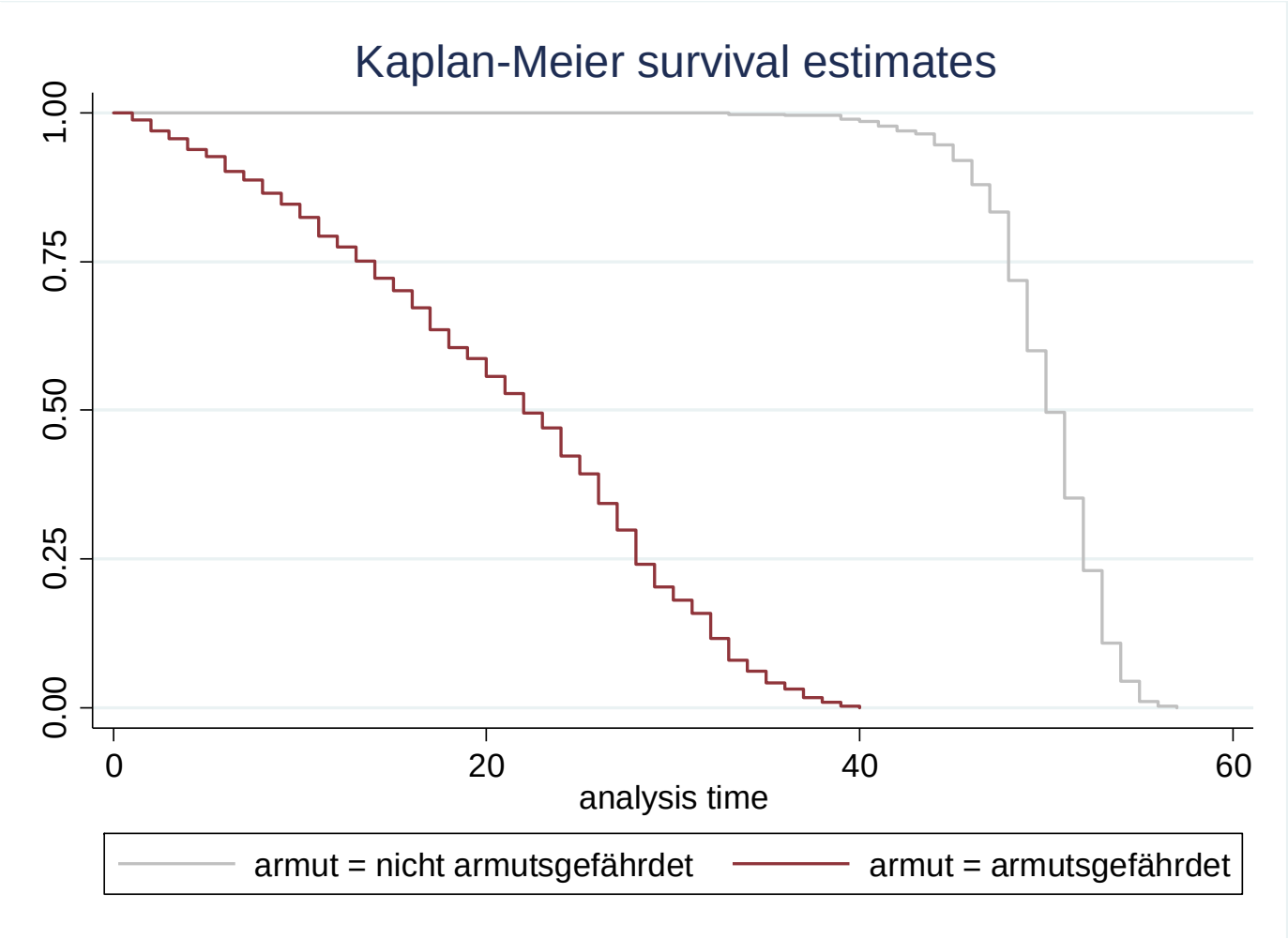
- typische Verfahren:
  - Beschreibung von (zeitlicher) Variation
  - Sequenzanalyse
  - Ereignisdatenanalyse

  Analyse der Zeit bis zum Eintritt eines Ereignisses bzw. der (zeitlichen) Verteilung dieser Ereignisse (z.B. Ereignis: substantielle Verschlechterung des Gesundheitszustandes)

  Vorteile: Nutzung der zeitlichen Informationen, Inklusion (rechts-) zensierter Fälle

  Kaplan-Meier-Schätzer und (semi-)parametrische Modelle z.B. Cox-Ph-Modell (inklusive zeitabhängiger Effekte)

Beispiel


Kaplan-Meier survival estimates
1.00
0.75
0.50
0.25
0.00
0
20
40
60
analysis time
armut = nicht armutsgefährdet
armut = armutsgefährdet

Beispiel

```
. stcox armut
        failure _d:  event
  analysis time _t:  (alter-origin)
            origin:  time alter
                id:  persnr

No. of subjects =          989                 Number of obs    =      35,270
No. of failures =          989
Time at risk    =        35270
                                               LR chi2(1)       =     1282.06
Log likelihood  =   -5280.7453                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       armut |   476.9942   197.5612    14.89   0.000      211.818    1074.146
------------------------------------------------------------------------------
```

Vorteile von Längsschnittdaten

1. Betrachtung dynamischer Zusammenhänge

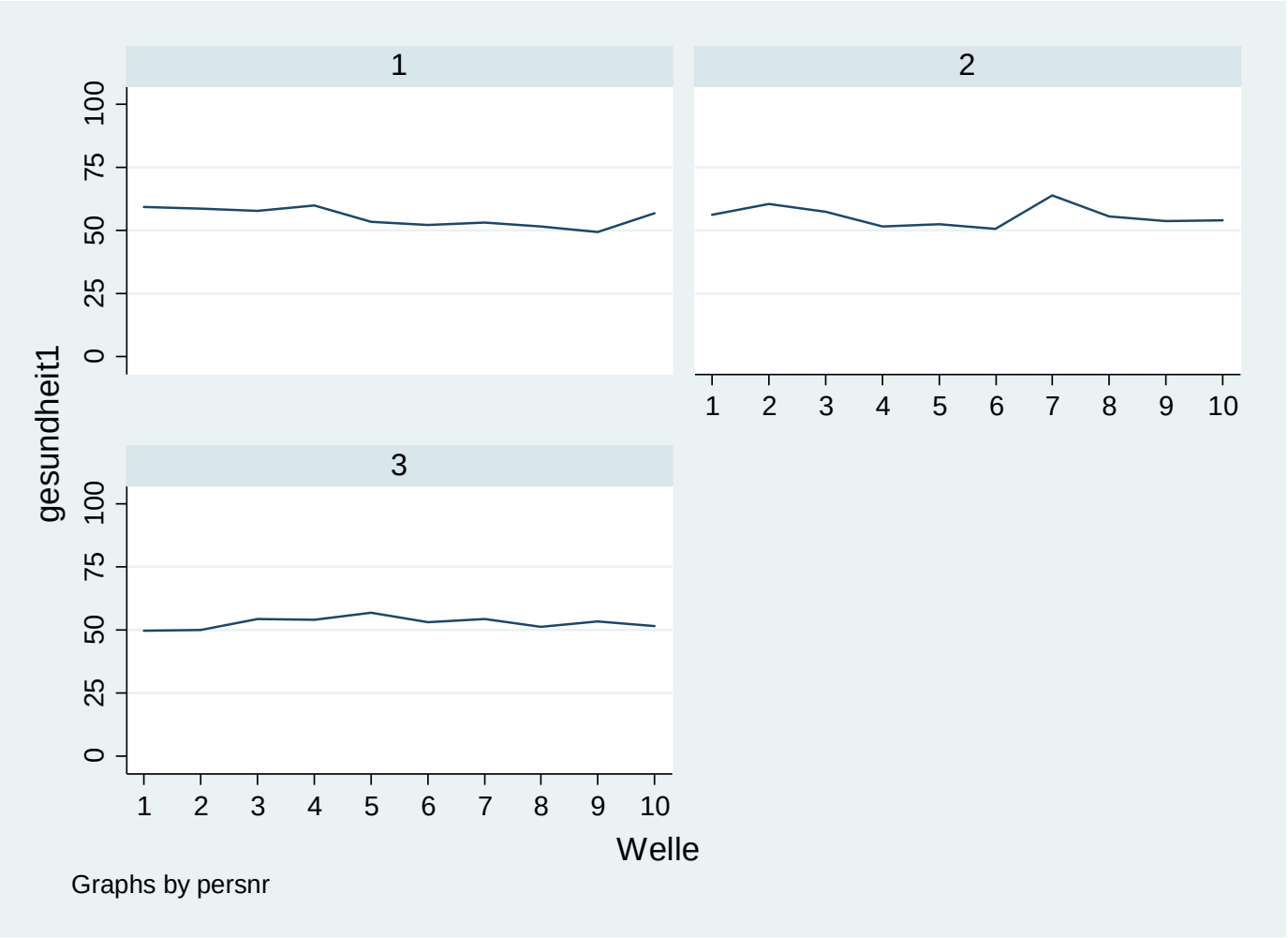
- typische Verfahren:
  - Beschreibung von (zeitlicher) Variation
  - Sequenzanalyse
  - Ereignisdatenanalyse
  - dynamische Modelle

  Frage: Wie stark beeinflussen vergangene Werte der abhängigen Variable den aktuellen Zustand (z.B. Welche Persistenz hat Gesundheitszustand?)

  Modelle mit zeitlich verzögerten Effekten der abhängigen Variable, z.B. $y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \gamma y_{it-1} + u_{it}$

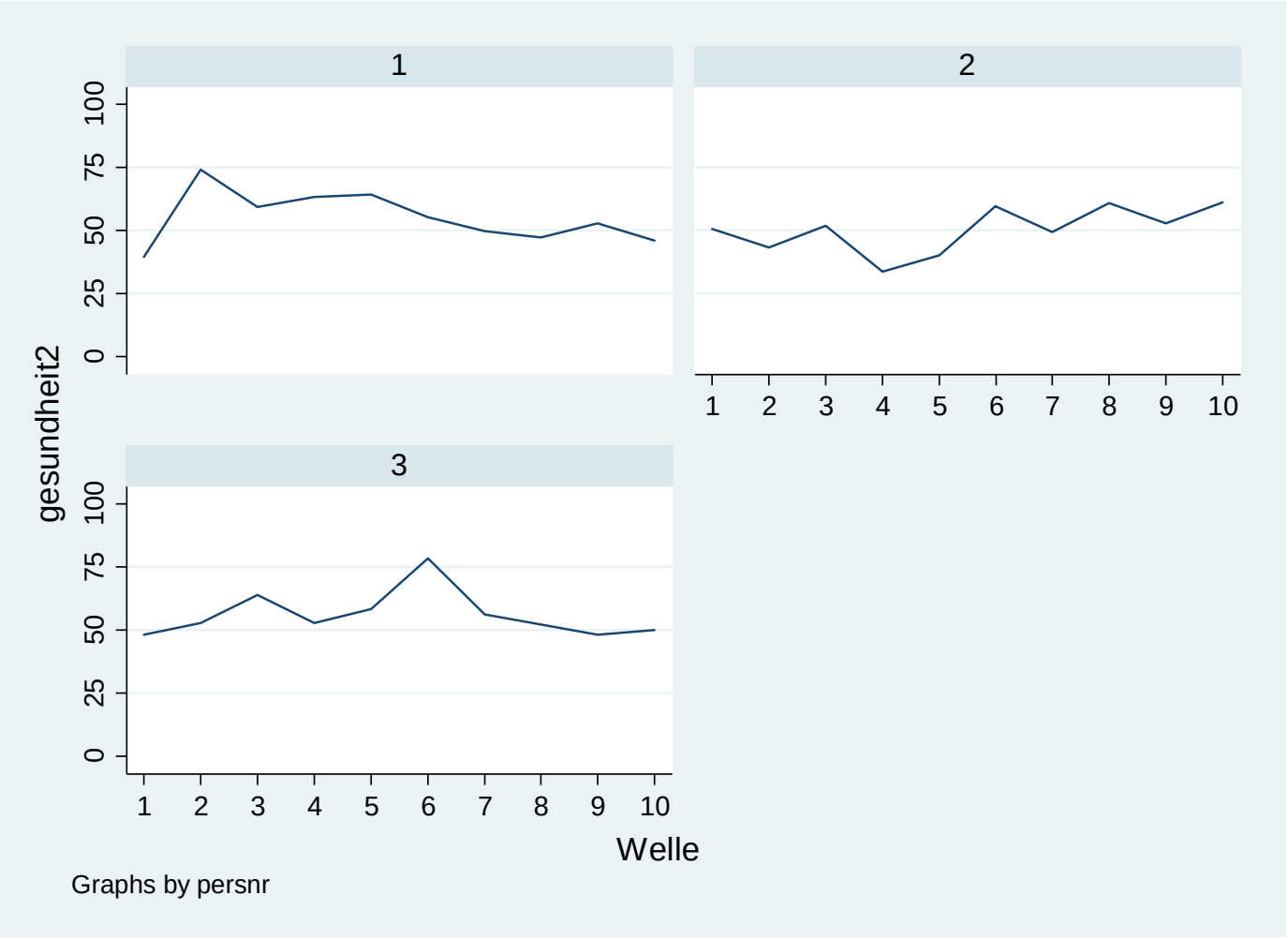
Beispiel

```
xtline  gesundheit1 in 1/30, ylabel(0(25)100) xlabel(1(1)10)
```



Beispiel

```
xtline  gesundheit2 in 1/30, ylabel(0(25)100) xlabel(1(1)10)
```



Beispiel

```
. reg gesundheit1 L1.gesundheit1 L1.armut1

      Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(2, 97)        =    116.77
       Model |  4718.98946         2  2359.49473   Prob > F        =    0.0000
    Residual |  1959.97809        97  20.2059597   R-squared       =    0.7065
-------------+----------------------------------   Adj R-squared   =    0.7005
       Total |  6678.96755        99  67.4643187   Root MSE        =    4.4951

------------------------------------------------------------------------------
 gesundheit1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 gesundheit1 |
         L1. |   .8804392   .0576353    15.28   0.000      .766049    .9948293
             |
      armut1 |
         L1. |   .4062324   .9005259     0.45   0.653    -1.381062    2.193527
             |
       _cons |   5.223281   2.604517     2.01   0.048     .0540352    10.39253
------------------------------------------------------------------------------
```

Beispiel

```
. reg gesundheit2 L1.gesundheit2 L1.armut2

      Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(2, 97)        =      5.13
       Model |  1247.43698         2   623.71849   Prob > F        =    0.0077
    Residual |  11802.2446        97  121.672624   R-squared       =    0.0956
-------------+----------------------------------   Adj R-squared   =    0.0769
       Total |  13049.6815        99  131.814965   Root MSE        =    11.031

------------------------------------------------------------------------------
 gesundheit2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 gesundheit2 |
         L1. |   .3240929   .1102609     2.94   0.004     .1052555    .5429303
             |
      armut2 |
         L1. |   -1.29893   2.264569    -0.57   0.568    -5.793473    3.195613
             |
       _cons |   31.50178   5.216199     6.04   0.000     21.14907    41.85449
------------------------------------------------------------------------------
```

Vorteile von Längsschnittdaten

1. Betrachtung dynamischer Zusammenhänge

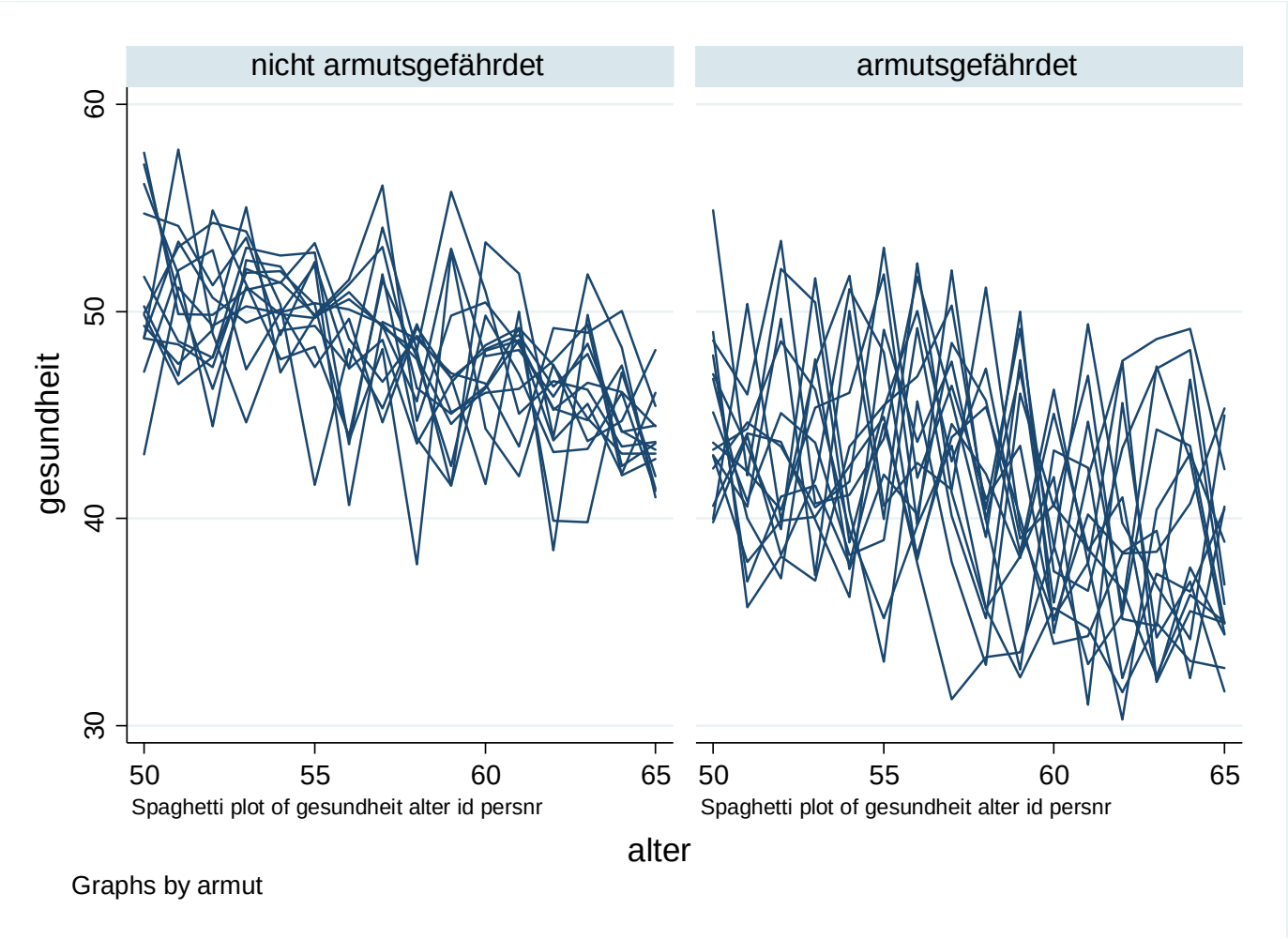
- typische Verfahren:
  - Beschreibung von (zeitlicher) Variation
  - Sequenzanalyse
  - Ereignisdatenanalyse
  - dynamische Modelle
  - Growth Curve Models

  analysiert werden within-Veränderungen und between-Differenzen in diesen Veränderungen

  Modelle in Form von mixed-models

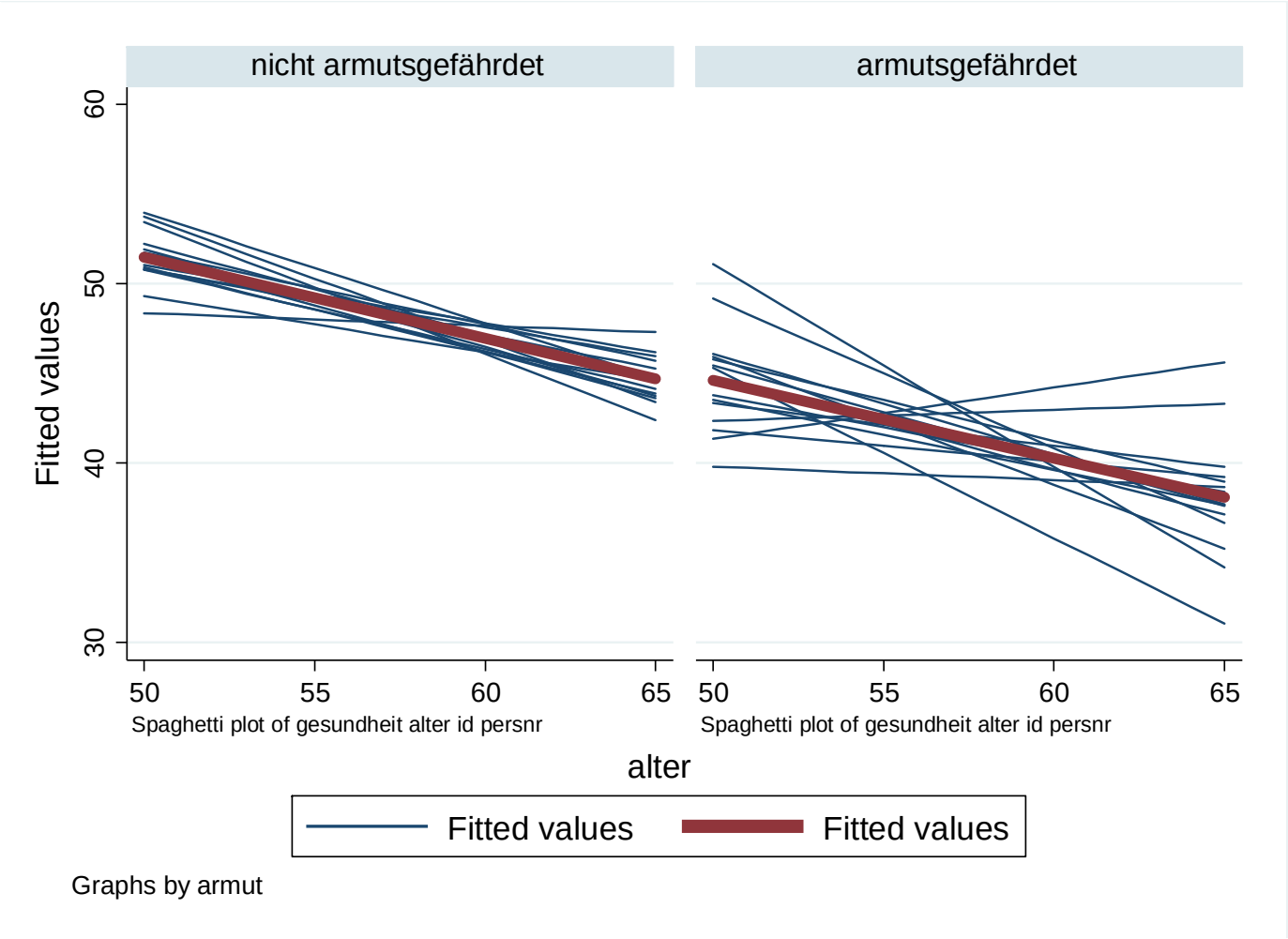
Beispiel

```
spagplot gesundheit alter, id(persnr) by(armut, graphregion(color(white))
plotregion(color(white))) nofit
```



Beispiel

```
spagplot gesundheit alter, id(persnr) by(armut, graphregion(color(white))
plotregion(color(white)))
```



Vorteile von Längsschnittdaten

2. Berücksichtigung von Heterogenität zur Schätzung von Kausaleffekten

a) Fixed-effects-Schätzung zur Verringerung des omitted-variable-bias durch unbeobachtete Heterogenität
b) Fixed-effects- oder random-effects-Schätzung zur Verringerung des Selektionsbias

Vorteile von Längsschnittdaten

2. Berücksichtigung von Heterogenität

a) Fixed-effects-Schätzung zur Verringerung des omitted-variable-bias durch unbeobachtete Heterogenität

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_{it} + u_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + e_{it}$$

bekannt: Schätzungen für β_j verzerrt,

wenn $\operatorname{cov}\left(x_{jit}, c_{it}\right) \neq 0$

bei Annahme zeitkonstanter c_i:

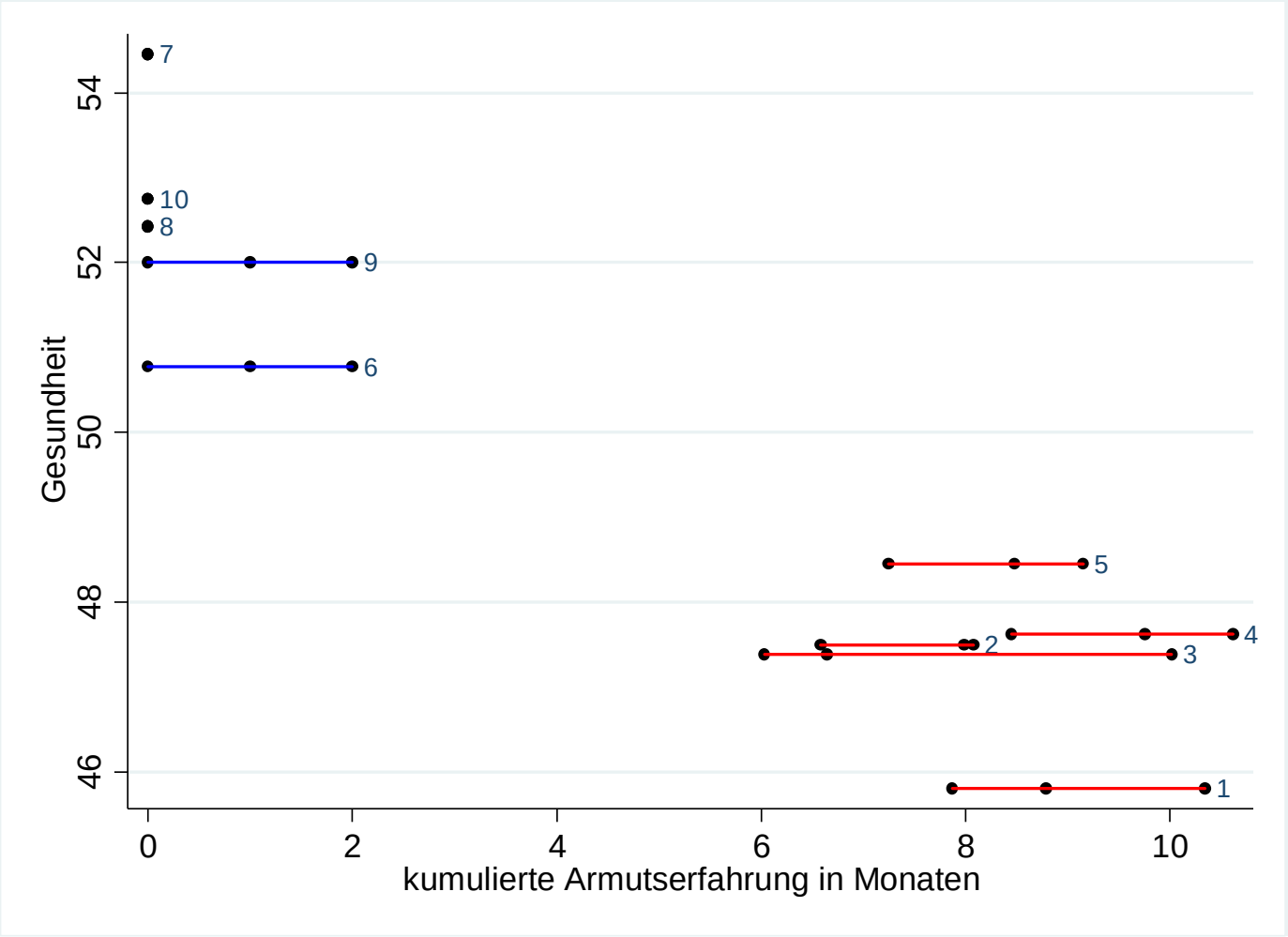
FE-Schätzung

Beispiel

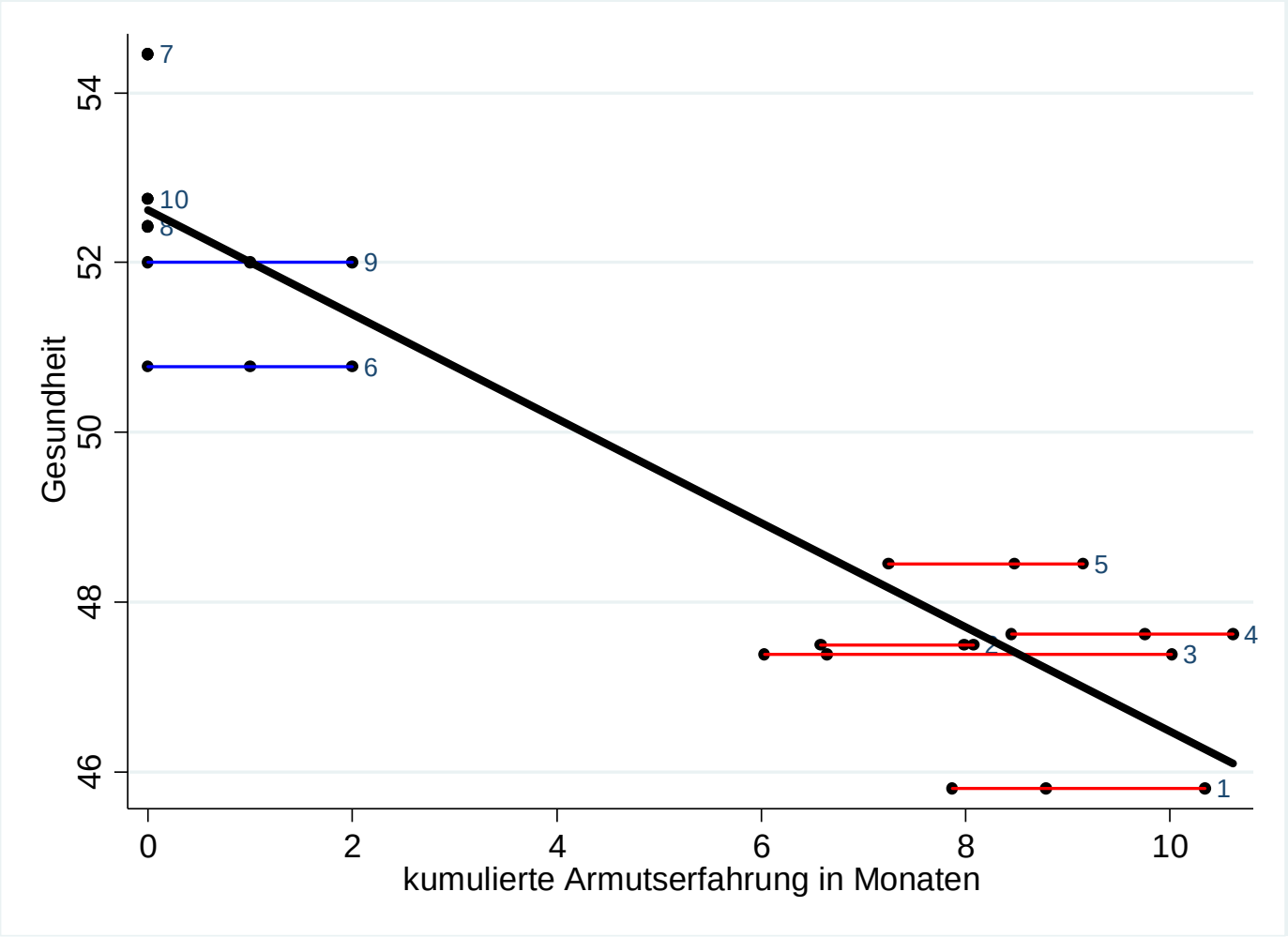
```
. list pid gesundheit x elternhaus risiko_aversion

     +-------------------------------------------------+
     | pid   gesund~t     kum_Al   eltern~s   risiko~n |
     |-------------------------------------------------|
  1. |   1   45.80452   7.870029          0    1.60905 |
  2. |   1   45.80452    8.78968          0    1.60905 |
  3. |   1   45.80452   10.34263          0    1.60905 |
    ...
 28. |  10   52.74524          0          1   5.490471 |
 29. |  10   52.74524          0          1   5.490471 |
 30. |  10   52.74524          0          1   5.490471 |
     +-------------------------------------------------+
```

```
gesundheit=45+5*elternhaus+0.5*risiko_aversion
```

Beispiel


7
10
8
9
6
5
2
4
3
1
Gesundheit
46
48
50
52
54
0
2
4
6
8
10
kumulierte Armutserfahrung in Monaten

Beispiel


Gesundheit
46
48
50
52
54
0
2
4
6
8
10
kumulierte Armutserfahrung in Monaten
7
10
8
9
6
5
4
2
3
1

Beispiel

```
. reg gesundheit x

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(1, 28)        =    156.32
       Model |  194.826379         1  194.826379   Prob > F        =    0.0000
    Residual |  34.8975837        28  1.24634228   R-squared       =    0.8481
-------------+----------------------------------   Adj R-squared   =    0.8427
       Total |  229.723963        29  7.92151597   Root MSE        =    1.1164

------------------------------------------------------------------------------
  gesundheit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -.6146164   .0491585   -12.50   0.000     -.715313   -.5139197
       _cons |   52.62176   .2972631   177.02   0.000     52.01285    53.23068
------------------------------------------------------------------------------
```

Beispiel

```
. reg gesundheit x elternhaus risiko_aversion

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(3, 26)        >  99999.00
       Model |  229.723963         3  76.5746544   Prob > F        =    0.0000
    Residual |  1.8007e-11        26  6.9258e-13   R-squared       =    1.0000
-------------+----------------------------------   Adj R-squared   =    1.0000
       Total |  229.723963        29  7.92151597   Root MSE        =    8.3e-07

------------------------------------------------------------------------------
     gesundheit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
              x |  -8.65e-08   1.45e-07    -0.59   0.557    -3.85e-07    2.12e-07
     elternhaus |   4.999999   1.20e-06  4.2e+06   0.000     4.999996    5.000001
risiko_aversion |   .5000003   7.59e-08  6.6e+06   0.000     .5000001    .5000004
          _cons |         45   1.38e-06  3.3e+07   0.000           45          45
------------------------------------------------------------------------------
```

Beispiel

```
. xtreg gesundheit x, fe i(pid)

Fixed-effects (within) regression               Number of obs     =         30
Group variable: pid                             Number of groups  =         10

R-sq:                                           Obs per group:
     within  =      .                                           min =          3
     between =      .                                           avg =        3.0
     overall =      .                                           max =          3

                                                F(1,19)           =          .
corr(u_i, Xb)  =      .                         Prob > F          =          .

------------------------------------------------------------------------------
  gesundheit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |          0  (omitted)
       _cons |   49.91641          .        .       .            .           .
-------------+----------------------------------------------------------------
     sigma_u |  2.9168982
     sigma_e |          0
         rho |          1   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(9, 19) = .                     Prob > F =      .
```

Vorteile von Längsschnittdaten

b) Fixed-effects- oder random-effects-Schätzung zur Verringerung des Selektionsbias

$$y_{it}^{*} = \mathbf{x}_{it}\boldsymbol{\beta} + c_{it} + u_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + e_{it}$$

$$y_{it} = y_{it}^{*} \text{ wenn } y_{it}^{*} > Z$$

$$y_{it} = \text{missing wenn } y_{it}^{*} \leq Z$$

$$\Rightarrow \text{Schätzungen für } \boldsymbol{\beta} \text{ verzerrt, auch}$$

$$\text{wenn } \operatorname{cov}\left(x_{jit}, c_{it}\right) = 0$$

bei Annahme zeitkonstanter c_i:

RE- oder FE-Schätzung

Beispiel

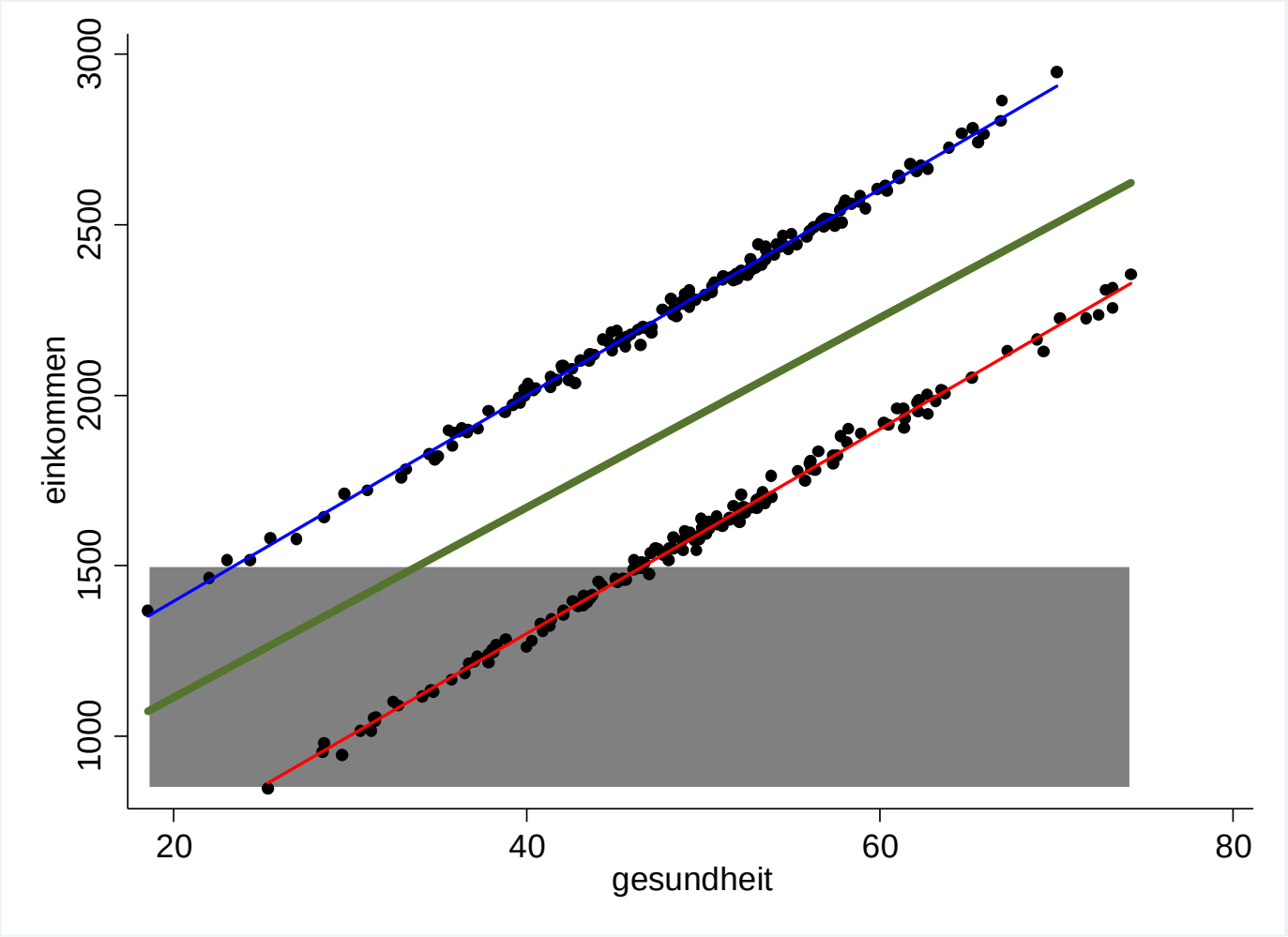
```
l  pid Welle einkommen x elternhaus
```

	pid	Welle	einkom~n	gesund~t	eltern~s
1.	1	1	1343.76	41.4315	0
2.	1	2	1640.15	51.47841	0
3.	1	3	1660.792	52.28166	0
...					
298.	100	1	2053.4	41.36801	1
299.	100	2	2337.763	51.73312	1
300.	100	3	2678.563	61.75181	1

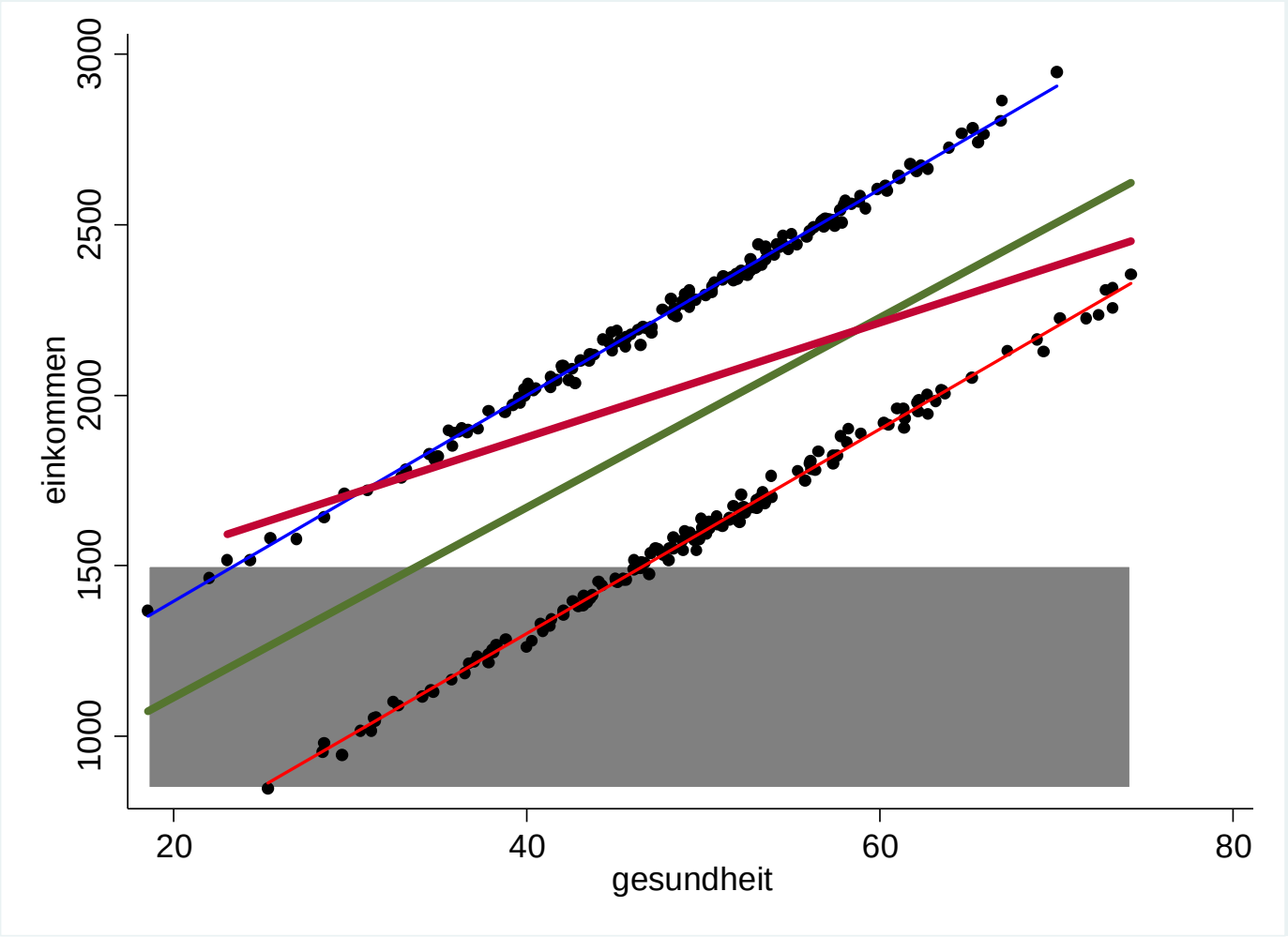
```
einkommen=100+30*gesundheit+700*elternhaus+20*invnorm(uniform())
```

Selektion für Einkommen kleiner 1500

Beispiel



Beispiel



Beispiel

```
. reg einkommen gesundheit if einkommen>1500

      Source |       SS           df       MS      Number of obs   =       244
-------------+----------------------------------   F(1, 242)       =     58.42
       Model |  6000493.77         1  6000493.77   Prob > F        =    0.0000
    Residual |  24854945.5       242  102706.386   R-squared       =    0.1945
-------------+----------------------------------   Adj R-squared   =    0.1911
       Total |  30855439.3       243  126977.116   Root MSE        =    320.48

------------------------------------------------------------------------------
   einkommen |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  gesundheit |   16.83409   2.202393     7.64   0.000     12.49578    21.17239
       _cons |   1202.838   115.0223    10.46   0.000     976.2651     1429.41
------------------------------------------------------------------------------
```

Beispiel

```
. reg einkommen gesundheit elternhaus if einkommen>1500

      Source |       SS           df       MS      Number of obs   =       244
-------------+----------------------------------   F(2, 241)       =  37543.56
       Model |  30756722.3         2  15378361.2   Prob > F        =    0.0000
    Residual |  98716.9265       241  409.613803   R-squared       =    0.9968
-------------+----------------------------------   Adj R-squared   =    0.9968
       Total |  30855439.3       243  126977.116   Root MSE        =    20.239

------------------------------------------------------------------------------
   einkommen |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  gesundheit |   30.10085   .1491882   201.76   0.000     29.80698    30.39473
  elternhaus |   699.3936   2.844898   245.84   0.000     693.7896    704.9976
       _cons |   96.85591   8.544194    11.34   0.000     80.02508    113.6867
------------------------------------------------------------------------------
```

Beispiel

```
. xtreg einkommen gesundheit if einkommen>1500, fe i(pid)

Fixed-effects (within) regression               Number of obs     =       244
Group variable: pid                             Number of groups  =        97

R-sq:                                           Obs per group:
     within  = 0.9952                                         min =         1
     between = 0.0000                                         avg =       2.5
     overall = 0.1945                                         max =         3

                                                F(1,146)          =  30158.58
corr(u_i, Xb)  = -0.3649                        Prob > F          =    0.0000

------------------------------------------------------------------------------
   einkommen |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  gesundheit |    30.2471    .174172   173.66   0.000     29.90288    30.59133
       _cons |   513.5628   9.030413    56.87   0.000     495.7156      531.41
-------------+----------------------------------------------------------------
     sigma_u |  351.41375
     sigma_e |  18.731568
         rho |  .99716679   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(96, 146) = 736.37                     Prob > F = 0.0000
```

Fazit

- empirische Lebensverlaufsforschung benötigt Längsschnittdaten
- diese Daten tragen i.d.R. mehr Informationen als reine Querschnittdaten
- ermöglichen Schätzung von verschiedenen Modellen
- Modelle beruhen auf unterschiedlichen Annahmen und haben unterschiedliche Anforderungen an Daten
- hohe Datenqualität aber immer extrem wichtig