Einfache Modelle für Paneldaten

Statistik II

Zum Nachlesen

- *Einfache* Modelle für Paneldaten
- Wooldridge ch. 13.1-13.4 (im Reader)

Exkurs: Trenddesigns

- Im Text: independently pooled cross-sections
- Querschnittsdaten aus verschiedenen Zeitpunkten
- Z. B. Allbus-Daten über die Zeit
- Trends in der Bevölkerung, aber keine individuellen Veränderungen

Was sind Paneldaten?

- „Pooled Cross-Sectional Time-Series“
- Verbindung von Querschnitts- und Zeitreihen-Design
- Dieselben Objekte (Personen, Parteien, Staaten . . .) werden mehrfach untersucht $\rightarrow$ individuelle Veränderungen
- Kombination verschiedener Zeitreihen bzw. verschiedener Querschnitte

Was sind Paneldaten?

⋱	⋮	⋮	⋮	⋮
...	Land A	Land B	Land C	1980
...	Land A	Land B	Land C	1981
...	Land A	Land B	Land C	1982
...	⋮	⋮	⋮	⋮

Was sind Paneldaten?

- „Pooled Cross-Sectional Time-Series“
- Verbindung von Querschnitts- und Zeitreihen-Design
- Dieselben Objekte (Personen, Parteien, Staaten . . .) werden mehrfach untersucht $\rightarrow$ individuelle Veränderungen
- Kombination verschiedener Zeitreihen bzw. verschiedener Querschnitte
- Panel-Daten sind doppelt indiziert: $y_{i,t}$
  - Index i über Objekte
  - Index t über Zeit
- Z. B. y-Wert für drittes Land zum zweiten Zeitpunkt: $y_{3,2}$
- Fünf Länder $\times$ fünf Zeitpunkte $= 25$ Beobachtungen von y

Warum sind Panel-Daten besonders?

- Datenmatrix normalerweise *nicht* quadratisch
  - Zeitdimension dominant: Mehr Beobachtungszeitpunkte als Objekte, z. B. Makro-Studien
  - Objekt-/Querschnittsdimension dominant: (sehr viel) mehr Objekte als Zeitpunkt, z. B. Umfragen
- Unterschiedliche Modelle und Probleme

Warum sind Panel-Daten besonders?

- Datenmatrix normalerweise *nicht* quadratisch
  - Zeitdimension dominant: Mehr Beobachtungszeitpunkte als Objekte, z. B. Makro-Studien
  - Objekt-/Querschnittsdimension dominant: (sehr viel) mehr Objekte als Zeitpunkt, z. B. Umfragen
- Unterschiedliche Modelle und Probleme
- Beobachtungen bzw. ϵ nicht unabhängig voneinander
  - Temporale Korrelation von ϵ, z. B. innerhalb eines Landes über die Zeit
  - Kontemporäre Korrelation von ϵ zum selben Zeitpunkt z. B. bei Nachbarn

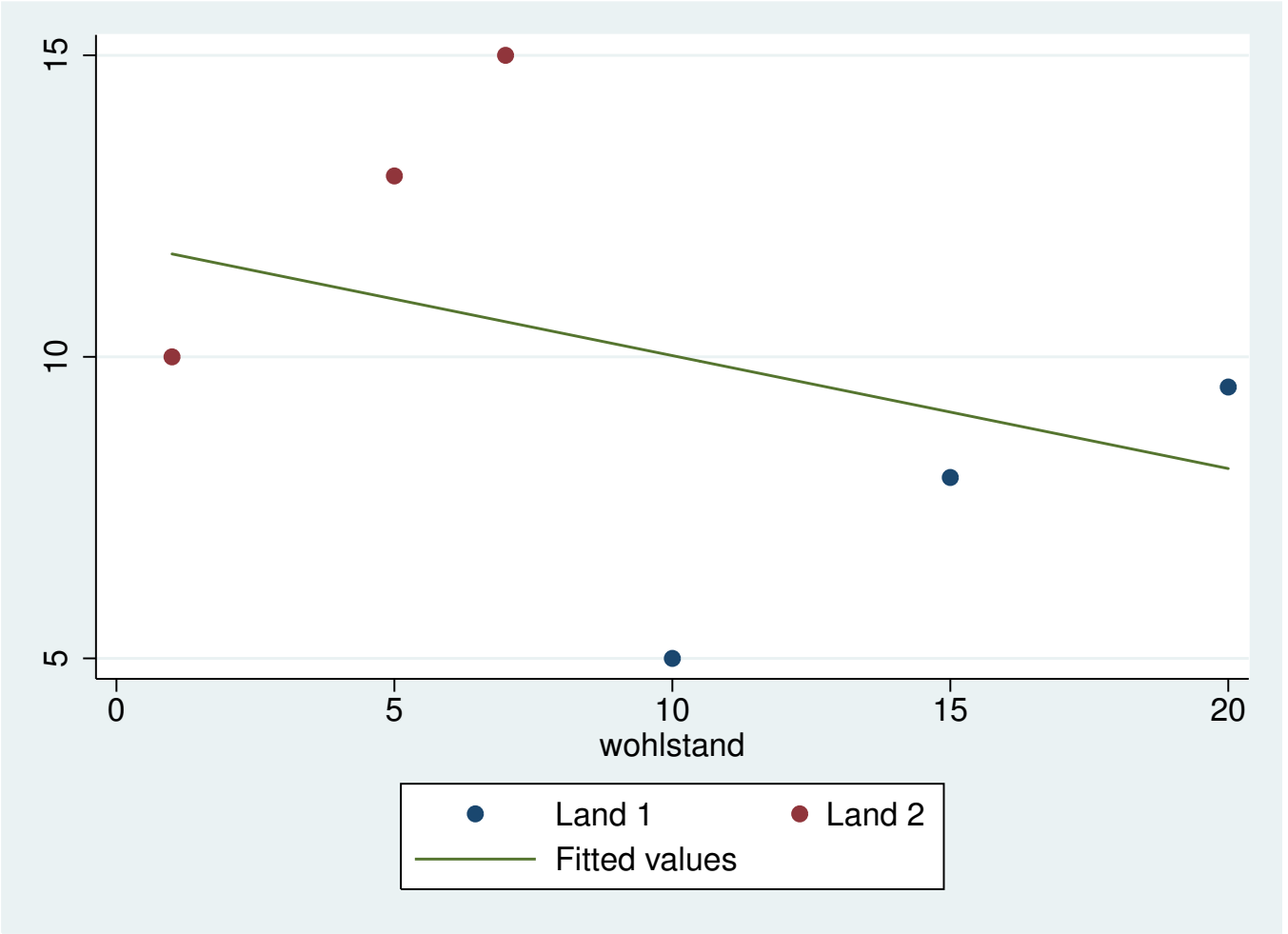
Warum sind Panel-Daten besonders?

- Datenmatrix normalerweise *nicht* quadratisch
  - Zeitdimension dominant: Mehr Beobachtungszeitpunkte als Objekte, z. B. Makro-Studien
  - Objekt-/Querschnittsdimension dominant: (sehr viel) mehr Objekte als Zeitpunkt, z. B. Umfragen
- Unterschiedliche Modelle und Probleme
- Beobachtungen bzw. ϵ nicht unabhängig voneinander
  - Temporale Korrelation von ϵ, z. B. innerhalb eines Landes über die Zeit
  - Kontemporäre Korrelation von ϵ zum selben Zeitpunkt z. B. bei Nachbarn
- Korrekte Standardfehler?
- Bias, wenn verschiedene Regressionsmodelle für verschiedene Objekte (verborgene Heterogenität)

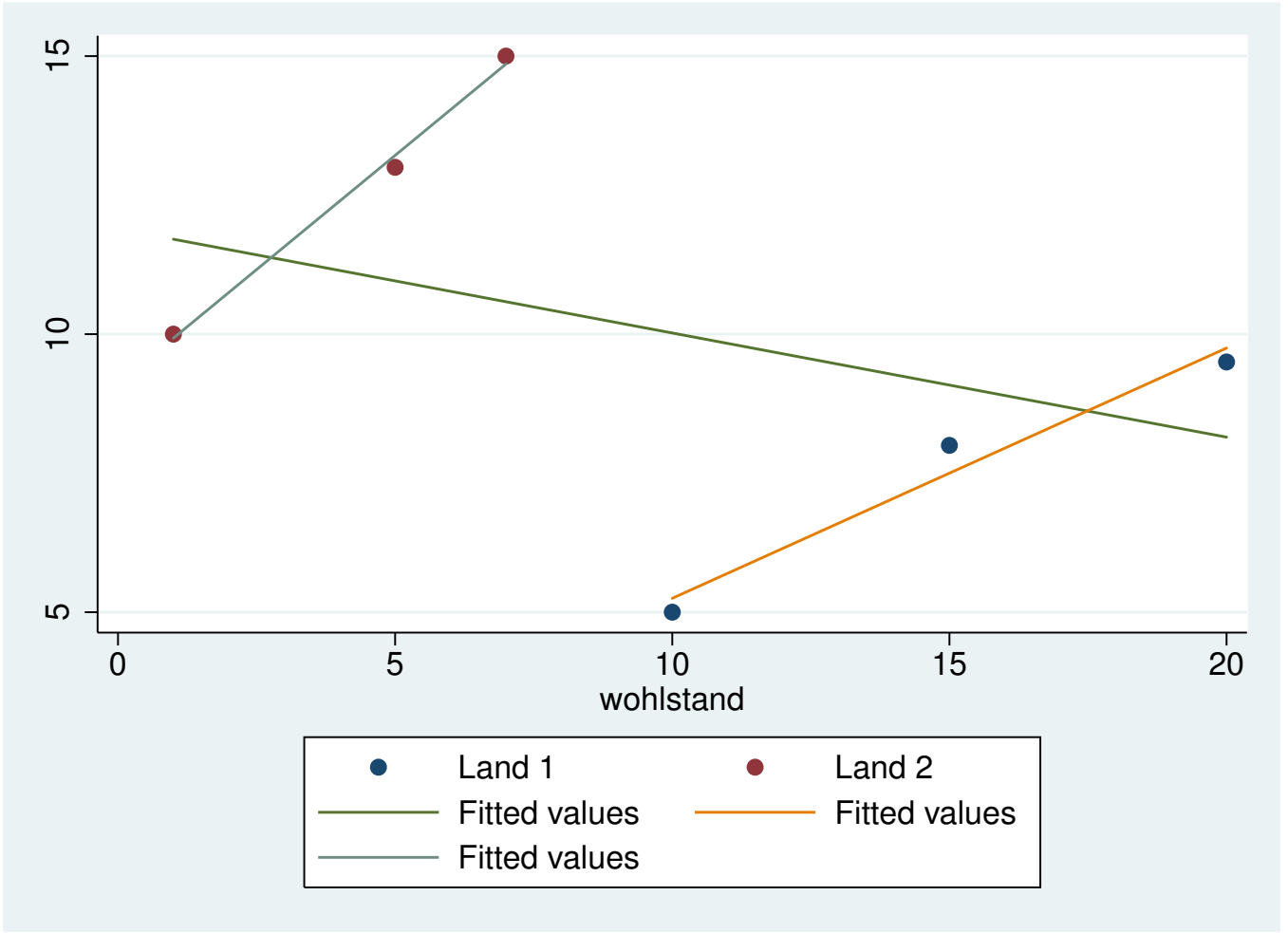
Wie war das mit der Heterogenität?

- Fiktives Beispiel mit zwei Ländern zu drei Zeitpunkten
- In beiden positiver Zusammenhang zwischen Wohlstand und Demokratieniveau
- Wird insgesamt positiver Zusammenhang geschätzt? (Simpson's Paradox)

Wie war das mit der Heterogenität?



Wie war das mit der Heterogenität?



Wie war das mit der Heterogenität?

```
. reg demokratie wohlstand
```

Source	SS	df	MS		
Model	8.39844475	1	8.39844475	Number of obs	= 6
Residual	54.8098886	4	13.7024721	F(1, 4)	= 0.61
				Prob > F	= 0.4774
Total	63.2083333	5	12.6416667	R-squared	= 0.1329
				Adj R-squared	= -0.0839
				Root MSE	= 3.7017

demokratie	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
wohlstand	-.1873259	.2392753	-0.78	0.477	-.8516606	.4770088
_cons	11.89415	2.762913	4.30	0.013	4.223074	19.56523

```
. reg demokratie wohlstand land2
```

Source	SS	df	MS		
Model	60.8867314	2	30.4433657	Number of obs	= 6
Residual	2.32160194	3	.773867314	F(2, 3)	= 39.34
				Prob > F	= 0.0070
Total	63.2083333	5	12.6416667	R-squared	= 0.9633
				Adj R-squared	= 0.9388
				Root MSE	= .8797

demokratie	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
wohlstand	.5509709	.1061598	5.19	0.014	.2131229	.8888189
land2	11.04369	1.340961	8.24	0.004	6.776152	15.31123
_cons	-.7645631	1.671432	-0.46	0.678	-6.083807	4.554681

Was ist ein Quasi-Experiment?

- Politische Veränderung, die im voraus bekannt ist (z. B. Einführung neue Schulart)
- Quasi-Experiment, natürliches Experiment
- Wissenschaft kann vor und nach Veränderung Daten sammeln
  - Kein echtes Experiment, da keine zufällige Aufteilung $\rightarrow$ kausale Aussagen schwierig wg. möglicher Drittvariablen
  - Aber Vorher-Nachher-Messung $\rightarrow$ aussagekräftiger als reine Querschnittsuntersuchung
- Im Text: Bau einer Müllverbrennungsanlage $\rightarrow$ sinkende Hauspreise
- Keine echten Panel-Daten, sondern Kombination von Querschnitten (nicht diesselben Häuser)

Müllverbrennung und Hauspreise

- Nach 1978 wurde bekannt, daß in einem Vorort von Boston eine Müllverbrennungsanlage gebaut wird
- 1981 begannen die Bauarbeiten
- Negative Auswirkung auf Hauspreise in der Nähe der Anlage?
- Allgemeiner: Negative Konsequenzen von Politik für Bürger (bzw. für Subgruppe von Bürgern)?

Naives Modell: Nähe zur Anlage

1981

```
. reg rprice nearinc if year == 1981
```

Source	SS	df	MS
Model	2.7059e+10	1	2.7059e+10
Residual	1.3661e+11	140	975815069
Total	1.6367e+11	141	1.1608e+09

Number of obs	142
F(1, 140)	27.73
Prob > F	0.0000
R-squared	0.1653
Adj R-squared	0.1594
Root MSE	31238

rprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
nearinc	-30688.27	5827.709	-5.27	0.000	-42209.97	-19166.58
_cons	101307.5	3093.027	32.75	0.000	95192.43	107422.6

- Niedrigeres Preisniveau durch Bau verursacht?
- Andere Gründe?

Naives Modell: Nähe zur Anlage

1978

```
. reg rprice nearinc if year == 1978
```

Source	SS	df	MS
Model	1.3636e+10	1	1.3636e+10
Residual	1.5332e+11	177	866239953
Total	1.6696e+11	178	937979126

Number of obs	179
F(1, 177)	15.74
Prob > F	0.0001
R-squared	0.0817
Adj R-squared	0.0765
Root MSE	29432

rprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
nearinc	-18824.37	4744.594	-3.97	0.000	-28187.62	-9461.117
_cons	82517.23	2653.79	31.09	0.000	77280.09	87754.37

- Preisniveau bereits vor Bau niedriger
- Kausalitätsproblem: Bau dort geplant, wo ökonomische Verluste geringer (und weniger Widerstand zu erwarten)?
- Außerdem: Hauspreise insgesamt gestiegen
- (Vermutlich) kausaler Effekt: *Differenz zwischen den beiden Differenzen*

Komplettes Modell

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

Mit

- β_0: Konstante
- β_1: Effekt von x_1 (Jahr=1981)
- β_2: Effekt von x_2 (Nähe zur Anlage)
- β_3: Interaktion zwischen Jahr=1981 und Nähe zur Anlage = Difference in Differences

Komplettes Modell

```
. reg rprice y81 nearinc y81nrinc
```

Source	SS	df	MS
Model	6.1055e+10	3	2.0352e+10
Residual	2.8994e+11	317	914632749
Total	3.5099e+11	320	1.0969e+09

Number of obs	= 321
F(3, 317)	= 22.25
Prob > F	= 0.0000
R-squared	= 0.1739
Adj R-squared	= 0.1661
Root MSE	= 30243

rprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y81	18790.29	4050.065	4.64	0.000	10821.88	26758.69
nearinc	-18824.37	4875.322	-3.86	0.000	-28416.45	-9232.293
y81nrinc	-11863.9	7456.646	-1.59	0.113	-26534.67	2806.866
_cons	82517.23	2726.91	30.26	0.000	77152.1	87882.36

- Wie groß ist der Wertverlust durch den Bau der Verbrennungsanlage?

Komplettes Modell

```
. reg rprice y81 nearinc y81nrinc
```

Source	SS	df	MS
Model	6.1055e+10	3	2.0352e+10
Residual	2.8994e+11	317	914632749
Total	3.5099e+11	320	1.0969e+09

Number of obs	321
F(3, 317)	22.25
Prob > F	0.0000
R-squared	0.1739
Adj R-squared	0.1661
Root MSE	30243

rprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y81	18790.29	4050.065	4.64	0.000	10821.88	26758.69
nearinc	-18824.37	4875.322	-3.86	0.000	-28416.45	-9232.293
y81nrinc	-11863.9	7456.646	-1.59	0.113	-26534.67	2806.866
_cons	82517.23	2726.91	30.26	0.000	77152.1	87882.36

- Wie groß ist der Wertverlust durch den Bau der Verbrennungsanlage?
- 11 864$
- Realistischerweise weitere Kontrollvariablen berücksichtigen, da keine zufällige Aufteilung auf Experimental-/ Kontrollgruppe (*Quasi*-Experiment)

Fixed-Effects Modell mit zwei Wellen

$$y_{it} = \beta_0 + \beta_1 d2_t + \beta_2 x_{it} + a_i + \epsilon_{it}$$

- Einfachstes Panel-Modell: zwei Beobachtungszeitpunkte
- $d2$: Indikator (Dummy) für zweite Panelwelle, konstant über Objekte
- x_{it}: unabhängige Variable, variiert über Zeit und Objekte (im Beispiel Arbeitslosenquote)
- a_i: unbeobachtete objektspezifische Einflüsse (konstanter Fehler über Zeit)
- ϵ_{it}: weitere zufällige Einflüsse (variabel über Zeit und Objekte)

Fixed-Effects Modell mit zwei Wellen

$$y_{it} = \beta_0 + \beta_1 d2_t + \beta_2 x_{it} + a_i + \epsilon_{it}$$

- Einfachstes Panel-Modell: zwei Beobachtungszeitpunkte
- $d2$: Indikator (Dummy) für zweite Panelwelle, konstant über Objekte
- x_{it}: unabhängige Variable, variiert über Zeit und Objekte (im Beispiel Arbeitslosenquote)
- a_i: unbeobachtete objektspezifische Einflüsse (konstanter Fehler über Zeit)
- ϵ_{it}: weitere zufällige Einflüsse (variabel über Zeit und Objekte)
- Problem: Korrelation (wahrscheinlich!) zwischen a_i und x_{it} macht OLS verzerrt und inkonsistent

Eine Lösung: Differenzen

$$\begin{aligned} y_{i2} &= \beta_0 + \beta_1 d2_2 (= 1) + \beta_2 x_{i2} + a_i + \epsilon_{i2} \\ y_{i1} &= \beta_0 + \beta_1 d2_1 (= 0) + \beta_2 x_{i1} + a_i + \epsilon_{i1} \end{aligned}$$

$$\begin{aligned} (y_{i2} - y_{i1}) &= \beta_1 + \beta_2 (x_{i2} - x_{i1}) + (\epsilon_{i2} - \epsilon_{i1}) \\ \Delta y_i &= \beta_1 + \beta_2 \Delta x_i + \Delta \epsilon_i \end{aligned}$$

- Über die Zeit konstanten zufälligen Einflüsse aus dem Modell entfernen, indem man *Veränderungen* betrachtet
- Korrelation zwischen x und a kein Problem mehr
- Korrelationen zwischen Veränderungen von x und Veränderungen von ϵ immer noch problematisch

Fixed Effects und Policy Evaluation

- Wirken sich Gesetzesänderungen auf die Zahl der Verkehrstoten aus?
- Zwei Typen von Gesetzen
  - „Open container“: offene Flaschen etc. mit Alkohol verboten
  - „Administrative per se“: Führerschein kann schon vor Verurteilung entzogen werden
- y_{it}: Tote pro 100 Millionen Personenmeilen pro Jahr in 1985 und 1990
  - Mittelwert 1985: 2.7 (Std.abweichung 0.6)
  - Mittelwert 1990: 2.2 (Std.abweichung 0.5))
- 50+1 US-Bundestaaten
- Diverse Gesetzesänderungen

Was hat sich geändert?

```
. tab admn85 admn90
```

admn85	admn90 0	admn90 1	Total
0	21	9	30
1	1	20	21
Total	22	29	51

```
. tab open85 open90
```

open85	open90 0	open90 1	Total
0	29	3	32
1	0	19	19
Total	29	22	51

Differenzen bilden

```
.   gen deltadeath= dthrte90-dthrte85
.   gen deltaadm= admn90-admn85
.   gen deltaopen = open90-open85
.   tab deltaopen
```

deltaopen	Freq.	Percent	Cum.
0	48	94.12	94.12
1	3	5.88	100.00
Total	51	100.00	

```
.   tab deltaadm
```

deltaadm	Freq.	Percent	Cum.
-1	1	1.96	1.96
0	41	80.39	82.35
1	9	17.65	100.00
Total	51	100.00	

```
.   summ deltadeath
```

Variable	Obs	Mean	Std. Dev.	Min	Max
deltadeath	51	-.545098	.3585045	-1.9	.3

Modell für Veränderungen schätzen

```
. reg deltadeath deltaopen deltaadm
```

Source	SS	df	MS
Model	.762579785	2	.381289893
Residual	5.66369475	48	.117993641
Total	6.42627453	50	.128525491

Number of obs	51
F(2, 48)	3.23
Prob > F	0.0482
R-squared	0.1187
Adj R-squared	0.0819
Root MSE	.3435

deltadeath	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
deltaopen	-.4196787	.2055948	-2.04	0.047	-.8330547	-.0063028
deltaadm	-.1506024	.1168223	-1.29	0.204	-.3854894	.0842846
_cons	-.4967872	.0524256	-9.48	0.000	-.6021959	-.3913784

Zusammenfassung

- Panel-Daten sind eine Mischform aus Quer- und Längsschnittanalyse
- Untersuchung von individuellen Veränderungen an mehreren Objekten
- Spezifische Probleme
- Vorsicht bei kausaler Interpretation – Design der Datenerhebung

Übersicht Vorlesung S II

1. Mittelwerte, Zusammenhansmaße, Hypothesentests
2. Was ist Regression?
3. Multiple Regression und Drittvariablenkontrolle
4. Test, Gewichtung, Multikollinearität, Kohorten
5. ANOVA und Transformationen
6. Schätzverfahren und Annahmen
7. Logit und Probit
8. Ordinale und multinomiale Modelle
9. Count Data
10. Zeitreihen
11. Panel