GV300 Quantitative Political Analysis

Week 22
Panel data models

Dominik Duell (University of Essex)

February 26, 2020

Terminology

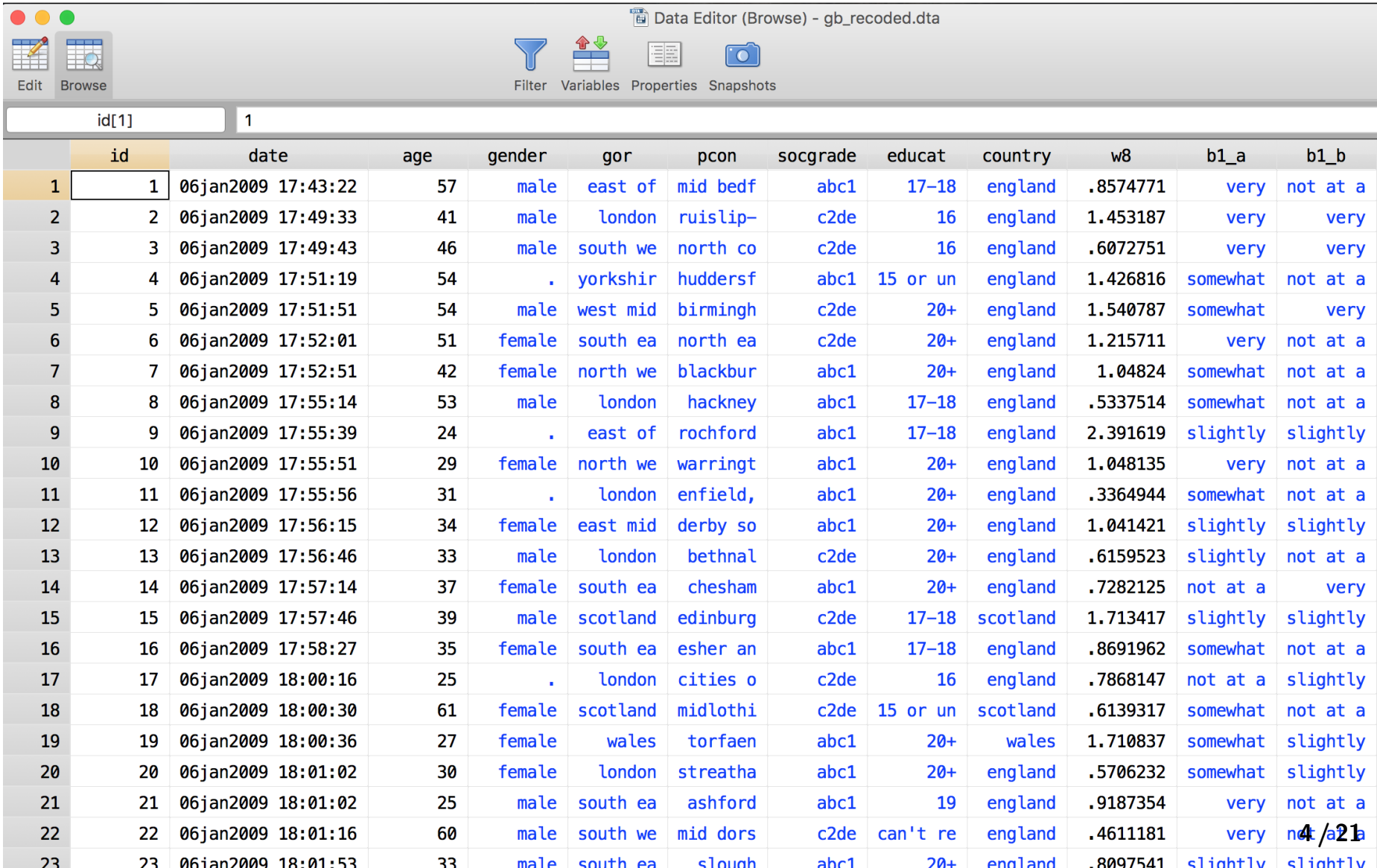
- Different models of the relationship between y and X:
  - **linear regression model**: $y_i = a + X'b$
  - non-linear regression model: $y_i = f(X, b)$ – e.g., logistic regression, poisson regression – Week 23: Maximum likelihood estimation, GLM
- Different forms of data
  - cross-sectional
  - time series – Week 24

Cross-sectional data

$$\mathbf{X} = \begin{bmatrix} y_1 & x_{11} & x_{12} & \dots & x_{1K} \\ \dots & x_{21} & x_{22} & \dots & \vdots \\ \dots & \vdots & \vdots & \ddots & \vdots \\ y_N & x_{N1} & \dots & \dots & x_{NK} \end{bmatrix}$$

- A data point is a pair indicating observation i for variable k , a typical observation is x_{ik}

Cross-sectional data



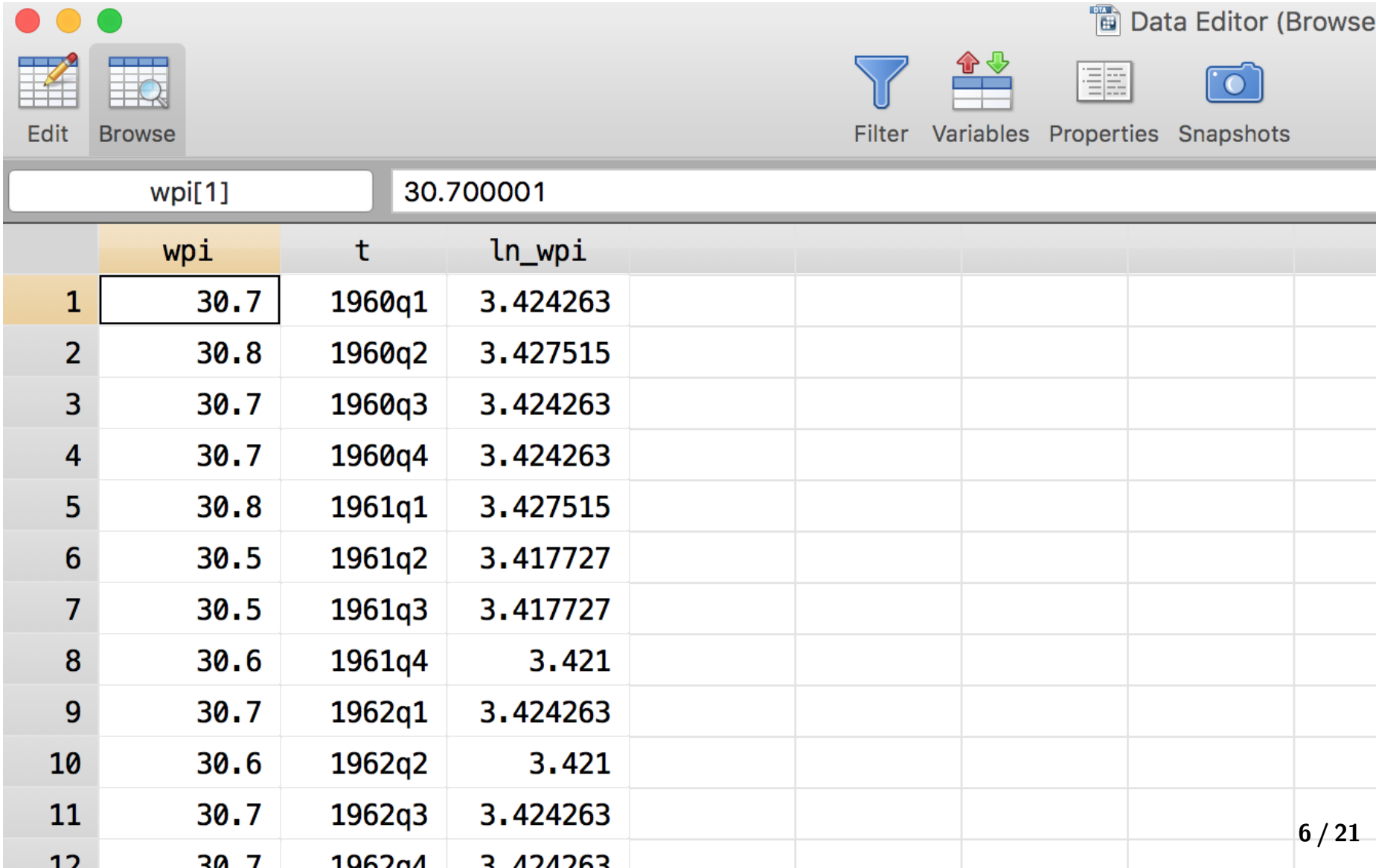
	id	date	age	gender	gor	pcon	socgrade	educat	country	w8	b1_a	b1_b
1	1	06jan2009 17:43:22	57	male	east of	mid bedf	abc1	17-18	england	.8574771	very	not at a
2	2	06jan2009 17:49:33	41	male	london	ruislip-	c2de	16	england	1.453187	very	very
3	3	06jan2009 17:49:43	46	male	south we	north co	c2de	16	england	.6072751	very	very
4	4	06jan2009 17:51:19	54	.	yorkshir	huddersf	abc1	15 or un	england	1.426816	somewhat	not at a
5	5	06jan2009 17:51:51	54	male	west mid	birmingh	c2de	20+	england	1.540787	somewhat	very
6	6	06jan2009 17:52:01	51	female	south ea	north ea	c2de	20+	england	1.215711	very	not at a
7	7	06jan2009 17:52:51	42	female	north we	blackbur	abc1	20+	england	1.04824	somewhat	not at a
8	8	06jan2009 17:55:14	53	male	london	hackney	abc1	17-18	england	.5337514	somewhat	not at a
9	9	06jan2009 17:55:39	24	.	east of	rochford	abc1	17-18	england	2.391619	slightly	slightly
10	10	06jan2009 17:55:51	29	female	north we	warringt	abc1	20+	england	1.048135	very	not at a
11	11	06jan2009 17:55:56	31	.	london	enfield,	abc1	20+	england	.3364944	somewhat	not at a
12	12	06jan2009 17:56:15	34	female	east mid	derby so	abc1	20+	england	1.041421	slightly	slightly
13	13	06jan2009 17:56:46	33	male	london	bethnal	c2de	20+	england	.6159523	slightly	not at a
14	14	06jan2009 17:57:14	37	female	south ea	chesham	abc1	20+	england	.7282125	not at a	very
15	15	06jan2009 17:57:46	39	male	scotland	edinburg	c2de	17-18	scotland	1.713417	slightly	slightly
16	16	06jan2009 17:58:27	35	female	south ea	esher an	abc1	17-18	england	.8691962	somewhat	not at a
17	17	06jan2009 18:00:16	25	.	london	cities o	c2de	16	england	.7868147	not at a	slightly
18	18	06jan2009 18:00:30	61	female	scotland	midlothi	c2de	15 or un	scotland	.6139317	somewhat	not at a
19	19	06jan2009 18:00:36	27	female	wales	torfaen	abc1	20+	wales	1.710837	somewhat	slightly
20	20	06jan2009 18:01:02	30	female	london	streatha	abc1	20+	england	.5706232	somewhat	slightly
21	21	06jan2009 18:01:02	25	male	south ea	ashford	abc1	19	england	.9187354	very	not at a
22	22	06jan2009 18:01:16	60	male	south we	mid dors	c2de	can't re	england	.4611181	very	not at a
23	23	06jan2009 18:01:53	33	male	south ea	slough	abc1	20+	england	.8097541	slightly	slightly

Time-series data

$$\mathbf{X} = \begin{bmatrix} y_1 & x_{11} & x_{12} & \dots & x_{1K} \\ \dots & x_{22} & \dots & \vdots & \\ \dots & \vdots & \vdots & \ddots & \vdots \\ y_T & x_{T1} & \dots & \dots & x_{TK} \end{bmatrix}$$

- A data point is a pair indicating observation i at time t , a typical observation is x_{it}
- For sure, we could have K variables measured across many time units or one variable measured over time for many observations N
- T is large

Time-series data



	wpi	t	ln_wpi
1	30.7	1960q1	3.424263
2	30.8	1960q2	3.427515
3	30.7	1960q3	3.424263
4	30.7	1960q4	3.424263
5	30.8	1961q1	3.427515
6	30.5	1961q2	3.417727
7	30.5	1961q3	3.417727
8	30.6	1961q4	3.421
9	30.7	1962q1	3.424263
10	30.6	1962q2	3.421
11	30.7	1962q3	3.424263
12	30.7	1962q4	3.424263

Terminology

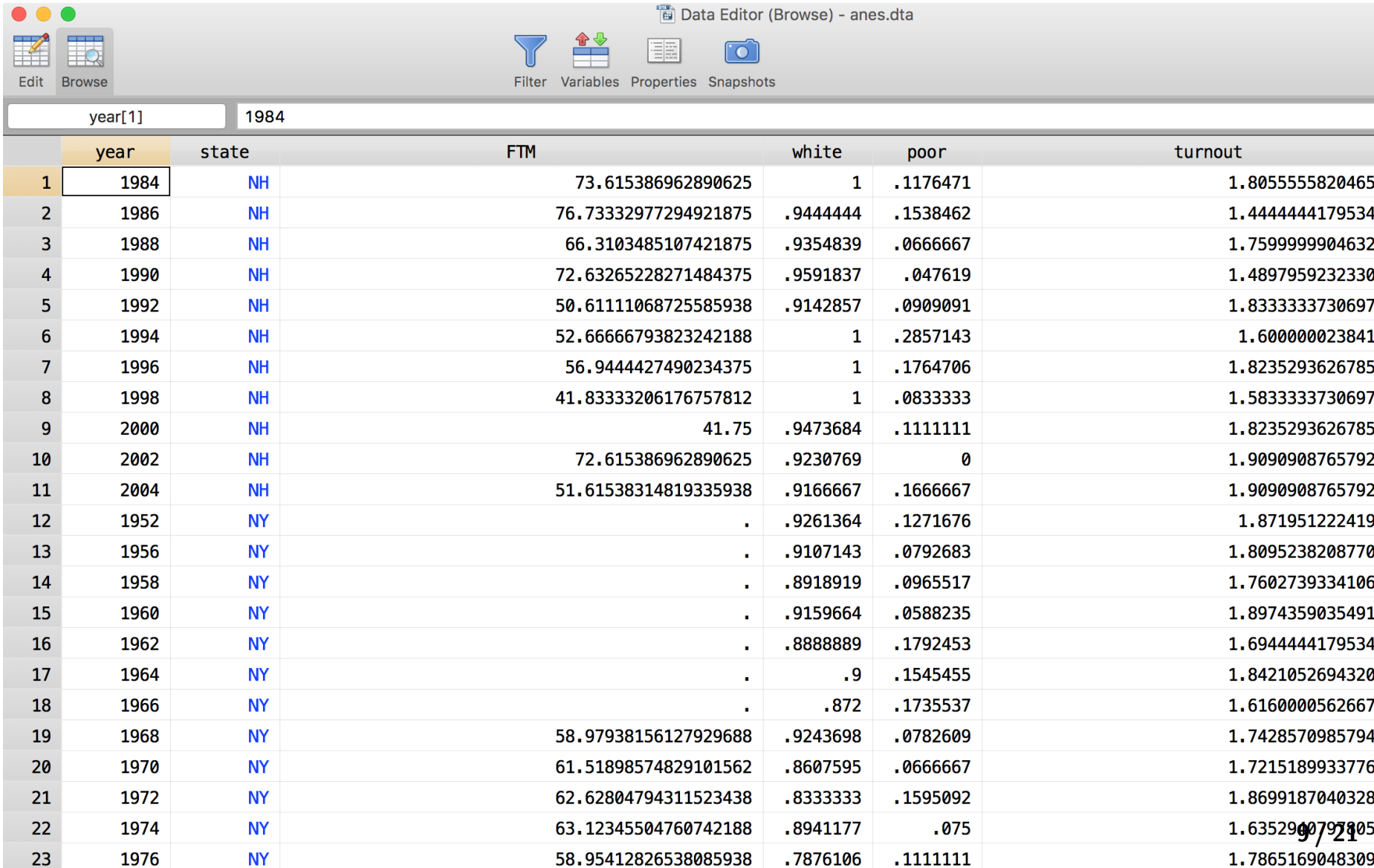
- Different models of the relationship between y and X:
  - **linear regression model**: $y_i = a + X'b$
  - **non-linear regression model**: $y_i = f(X, b)$ – e.g., logistic regression, poisson regression
- Different forms of data
  - cross-sectional
  - time - series
  - **repeated cross-sectional data**
  - **panel or longitudinal data**

Panel/longitudinal data

$$\mathbf{X} = \begin{bmatrix} y_{11} & x_{111} & x_{121} & \cdots & x_{1K1} \\ y_{12} & x_{112} & x_{122} & \cdots & x_{1K2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{1T} & x_{11T} & x_{12T} & \cdots & x_{1KT} \\ y_{21} & x_{211} & x_{221} & \cdots & x_{2K1} \\ y_{22} & x_{211} & x_{222} & \cdots & \vdots \\ y_{2T} & x_{212} & x_{222} & \cdots & x_{2KT} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{N1} & x_{N11} & \cdots & \cdots & x_{NK1} \\ y_{N2} & x_{N12} & \cdots & \cdots & x_{NK2} \\ y_{NT} & x_{N12T} & \cdots & \cdots & x_{NKT} \end{bmatrix}$$

- A data point is a triple indicating observation i at time t for variable k , a typical observation is x_{ikt}
- T is small

Panel/longitudinal data

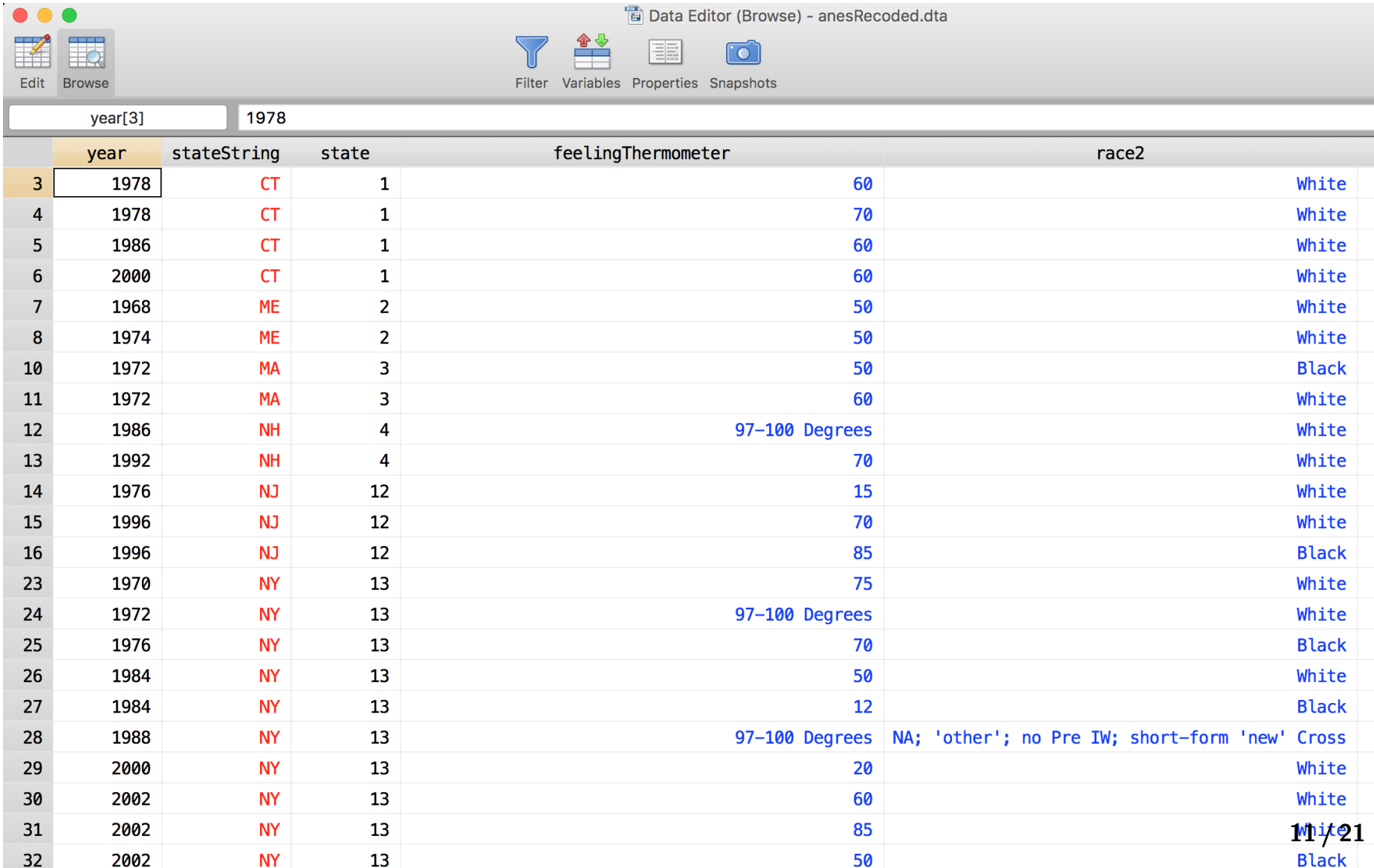


	year	state	FTM	white	poor	turnout
1	1984	NH	73.615386962890625	1	.1176471	1.8055555820465
2	1986	NH	76.73332977294921875	.9444444	.1538462	1.4444444179534
3	1988	NH	66.3103485107421875	.9354839	.0666667	1.7599999904632
4	1990	NH	72.63265228271484375	.9591837	.047619	1.4897959232330
5	1992	NH	50.61111068725585938	.9142857	.0909091	1.8333333730697
6	1994	NH	52.66666793823242188	1	.2857143	1.600000023841
7	1996	NH	56.9444427490234375	1	.1764706	1.8235293626785
8	1998	NH	41.83333206176757812	1	.0833333	1.5833333730697
9	2000	NH	41.75	.9473684	.1111111	1.8235293626785
10	2002	NH	72.615386962890625	.9230769	0	1.9090908765792
11	2004	NH	51.61538314819335938	.9166667	.1666667	1.9090908765792
12	1952	NY	.	.9261364	.1271676	1.871951222419
13	1956	NY	.	.9107143	.0792683	1.8095238208770
14	1958	NY	.	.8918919	.0965517	1.7602739334106
15	1960	NY	.	.9159664	.0588235	1.8974359035491
16	1962	NY	.	.8888889	.1792453	1.6944444179534
17	1964	NY	.	.9	.1545455	1.8421052694320
18	1966	NY	.	.872	.1735537	1.6160000562667
19	1968	NY	58.97938156127929688	.9243698	.0782609	1.7428570985794
20	1970	NY	61.51898574829101562	.8607595	.0666667	1.7215189933776
21	1972	NY	62.62804794311523438	.8333333	.1595092	1.8699187040328
22	1974	NY	63.12345504760742188	.8941177	.075	1.6352940797805
23	1976	NY	58.95412826538085938	.7876106	.1111111	1.7865169048309

Terminology

- Different models of the relationship between y and X:
  - **linear regression model**: $y_i = a + X'b$
  - **non-linear regression model**: $y_i = f(X, b)$ – e.g., logistic regression, poisson regression
- Different forms of data
  - cross-sectional
  - time - series
  - **repeated cross-sectional data**
  - **panel or longitudinal data**
  - **hierarchical or multi-level data** – where observations fall into hierarchical, completely nested levels

Hierarchical or multi-level data



	year	stateString	state	feelingThermometer	race2
3	1978	CT	1	60	White
4	1978	CT	1	70	White
5	1986	CT	1	60	White
6	2000	CT	1	60	White
7	1968	ME	2	50	White
8	1974	ME	2	50	White
10	1972	MA	3	50	Black
11	1972	MA	3	60	White
12	1986	NH	4	97-100 Degrees	White
13	1992	NH	4	70	White
14	1976	NJ	12	15	White
15	1996	NJ	12	70	White
16	1996	NJ	12	85	Black
23	1970	NY	13	75	White
24	1972	NY	13	97-100 Degrees	White
25	1976	NY	13	70	Black
26	1984	NY	13	50	White
27	1984	NY	13	12	Black
28	1988	NY	13	97-100 Degrees	NA; 'other'; no Pre IW; short-form 'new' Cross
29	2000	NY	13	20	White
30	2002	NY	13	60	White
31	2002	NY	13	85	White
32	2002	NY	13	50	Black

Terminology

- Different models of the relationship between y and X:
  - **linear regression model**: $y_i = a + X'b$
  - **non-linear regression model**: $y_i = f(X, b)$ – e.g., logistic regression, poisson regression
- Different forms of data
  - cross-sectional
  - time - series
  - repeated cross-sectional
  - panel or longitudinal data
  - Hierarchical or multi-level data
- Recall: **Consistent estimator**
  - An estimator that converges in probability to the population parameter as the sample size grows

Why do we care?

- better **control** over the DGP
  - Often, observations are not independent over time and/or linked
  - We need a way to account for
    - **between** variation $\rightarrow$ as in cross-section data
    - **within** variation $\rightarrow$ as in time-series data

How to model panel data?

- Consider a simple bivariate, cross-sectional population model:

$$y_i = a + b_1 x_1 + e_i$$

- How to represent the panel-nature of the data? Think about how regressors, coefficients enter the model
- How to represent error dependencies? Think about what kind of correlation should be described.
- Any other considerations necessary for consistent estimation of the population parameters?

How to model panel data?

- Short or long panel: few t, many i or many t, few i – Why important?
- Regressors:
  - time-variant vs
  - time-invariant (e.g. gender)
  - endogenous ($E[e|x] \neq 0$) $\rightarrow$ FE-models
  - lagged DV $\rightarrow$ dynamic models
- Coefficients: may vary across i or t
- Errors:
  - Correlation over t (within i) $\rightarrow$ clustering on i
  - Correlation over i (e.g., over is within one country, school) $\rightarrow$ hierarchical data $\rightarrow$ HLM
- Balance: $T_i = T \, \forall \, i$?
- Measurement: t equally spaced?

How to model panel data?

- We focus on short panels: T small, fixed and $N \to \infty$
- errors independent across i
- balanced panel $\to$ why could a panel be unbalanced? $\to$ attrition
- We will consider the individual-specific effects (population) model

$$y_{it} = a_i + X'_{it} b + e_{it}$$

- Estimation strategies:
  - pooled OLS model with clustered standard errors (feasible in short panels)
  - population-averaged estimator (pooled FGLS) model – Feasible Generalized Least Squares
  - fixed effects (FE) models
  - random effects (RE) models
  - mixed linear models with slope parameters varying over i or t – technically RE models, also referred to as HLM

Population-averaged estimator (pooled FGLS) model

- Rewrite individual-specific models as pooled model

$$y_{it} = a + \mathbf{x}'_{it} b + u_{it}$$

  where $u_{it} = a_i - a + e_{it}$ and any time-specific effect is assumed to be fixed and included as time dummies in the regressors $\mathbf{x}_{it}$
- Individual-specific effect a_i are now centered on zero
- For consistent estimation with OLS error term $(a_i - a + e_{it})$ needs to be uncorrelated with $\mathbf{x}_{it}$
- Achieved by specifying assumed error correlation structure

Fixed-effects model – Within estimator

- Individual-specific effect a_i allowed to correlate with regressors $\mathbf{x}_i$

$$y_{it} = \mathbf{x}'_{it} b + u_{it}$$

  where $u_{it} = a_i + e_{it}$ and $\mathbf{x}_{it}$ is allowed to be correlated with the **time-invariant** component of the error term, a_i – thus the term**fixed** effects
- We cannot get a consistent estimate for $a_1, \ldots, a_N$ and b because T is too small $\rightarrow$ **incidental parameter** problem
- We can get a consistent estimate for b, the coefficient on the time-varying regressor by removing the fixed effect a_i:
  - performing OLS on mean-differenced data
  - $(y_{it} - \overline{y}_i) = (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)'\mathbf{b} + (e_{it} - \overline{e}_i)$ – Where did a_i go?
- Downside: we cannot get predicted values of y_{it} because we have no estimate for a_i
- Assumption: $E(e_{it}|a_i, \mathbf{x}_{it}) = 0$

Random-effects model

- Individual-specific effect a_i is assumed to be uncorrelated with regressors X_i

$$y_{it} = X'_{it} b + u_{it}$$

where $u_{it} = a_i + e_{it}$, $a_i \sim (a, \sigma_a^2)$, and $e_{it} \sim (0, \sigma_e^2)$

- Then,

$$Corr(u_{it}, u_{is}) = \frac{\sigma_e^2}{(\sigma_a^2 + \sigma_e^2)} \forall\, s \neq t$$

- Here is the RE estimator:

$$(y_{it} - \hat{\theta}_i \overline{y}_i) = (1 - \hat{\theta}_i)a + (\mathbf{x}_{it} - \hat{\theta}_i \overline{\mathbf{x}_i})'\mathbf{b} + [(1 - \hat{\theta}_i)a_i + (e_{it} - \hat{\theta}_i \overline{e}_i)]$$

where $\hat{\theta}_i$ is a consistent estimate of

$$\theta_i = 1 - \sqrt{\frac{\sigma_e^2}{T_i \sigma_a^2 + \sigma_e^2}}$$

- Assumption: $E(e_{it}|\mathbf{x}_{it}) = 0$

Random-effects model

- Here is the RE estimator:

$$(y_{it} - \hat{\theta}_i \overline{y}_i) = (1 - \hat{\theta}_i)a + (\mathbf{x}_{it} - \hat{\theta}_i \overline{\mathbf{x}_i})'\mathbf{b} + [(1 - \hat{\theta}_i)a_i + (e_{it} - \hat{\theta}_i \overline{e}_i)]$$

- Special cases of the RE estimator are pooled OLS (no within variation or $\theta_i = 0$) and within estimator ($\theta_i = 1$)
- Upside: We are able to obtain predicted values and marginal effects (time-variant and in-variant regressors)
- Downside: inconsistent if a_i correlated with X_i

FE or RE models?

- RE uses within variation – if no OVB, consistent and more efficient than FE
- FE allows for OVB of time-invariant variables – FE uses subjects as their own control! – enough variation within subject needed – assuming $E(e_{it}|a_i, \mathbf{x}_{it}) = 0$ vs $E(e_{it}) = 0$ for consistent estimation
- RE often more efficient but higher danger of inconsistent estimates (asymptotic bias)
- FE cannot deliver predicted values and coefficient estimates for time-invariant variables
- Hausman test – see `gv300_stataFile_week22.do`