Panel Regression Analysis

Eduard Bukin

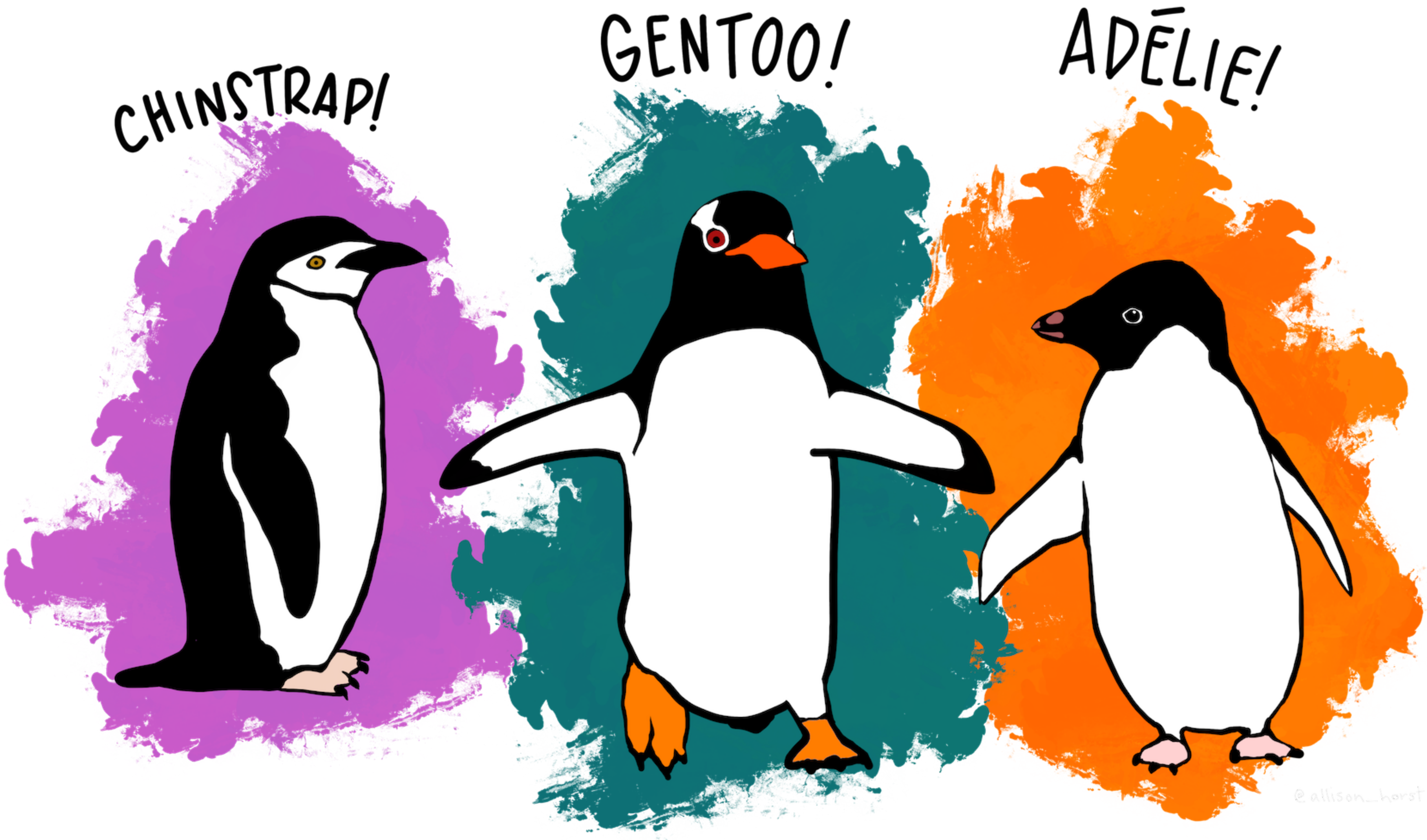
Recap

- Ceteris paribus!?
  - Why multiple regression is “good”?
  - What variables are important when establishing a causal effect of a treatment (key variable)?
  - What if we do not have an important variable?
- Selection bias = OVB! In multiple regression analysis.
  - What does OVB to our regression estimates?
  - Bias (inconsistency) of estimates!

Simpson’s paradox

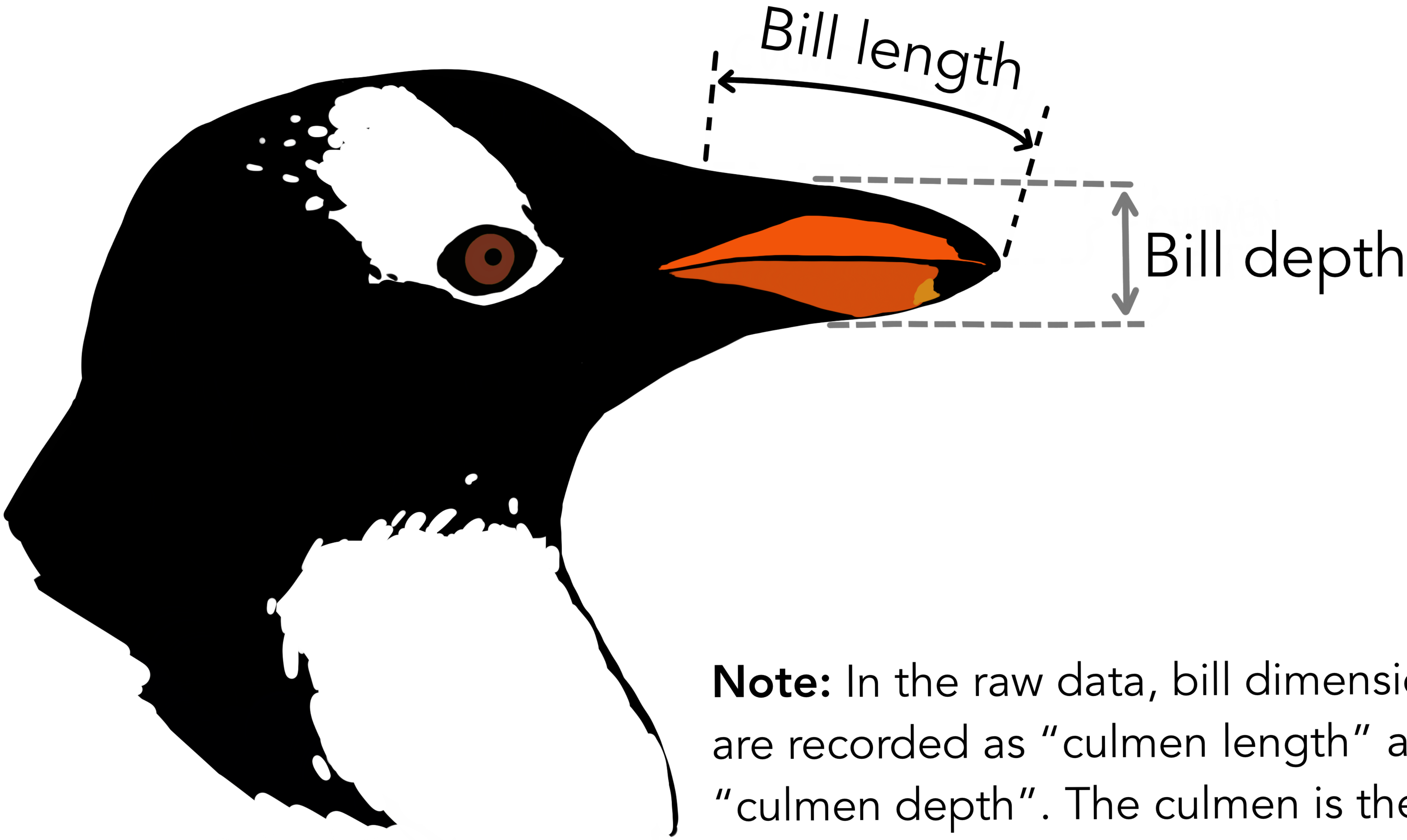
See: Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society: Series B (Methodological), 13(2), 238–241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x

Simpson’s paradox (with penguins)



Let us investigate…

The relationship between bill length and depth in penguins…



Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

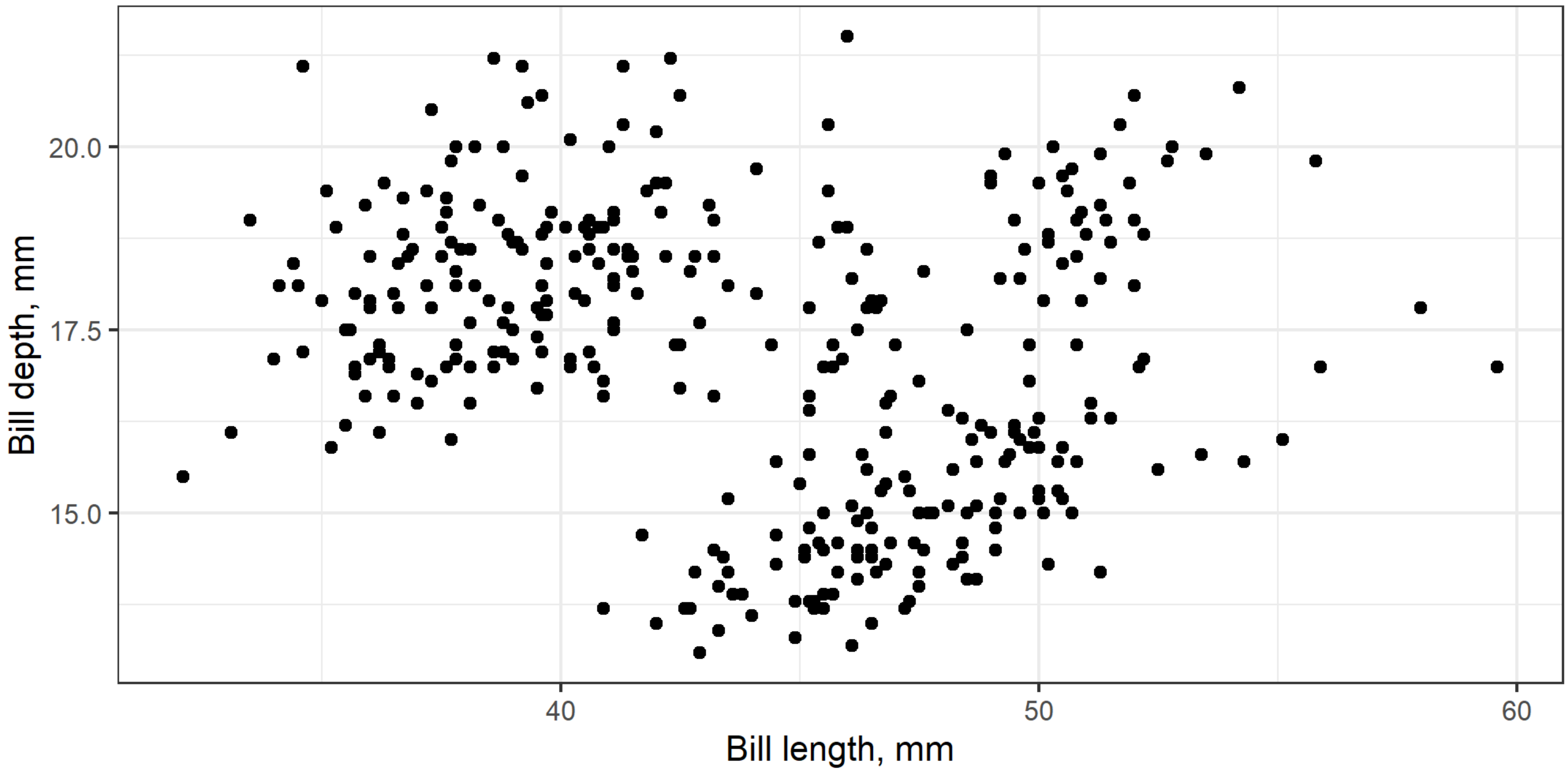
The data

```r
library(tidyverse)
library(modelsummary)
penguins <- read_csv("penguins")
penguins %>% glimpse()
```

```
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel…
$ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse…
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, …
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, …
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186…
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, …
$ sex               <fct> male, female, female, NA, female, male, female, male…
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007…
```

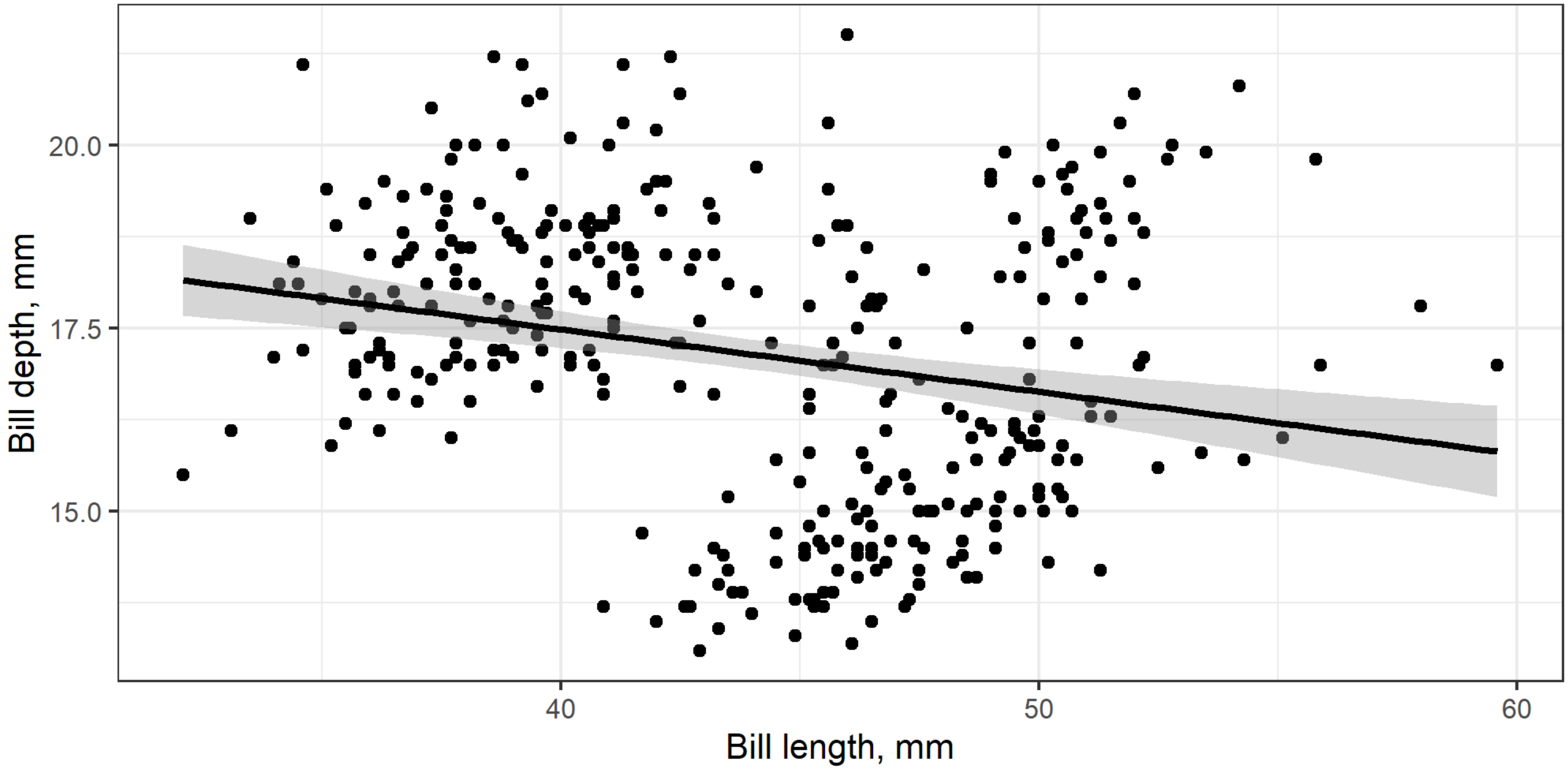
The relationship

```r
gg_bill <-
  penguins %>% ggplot() +
  aes(x = bill_length_mm, y = bill_depth_mm) +
  xlab("Bill length, mm") + ylab("Bill depth, mm") +
  geom_point()
gg_bill
```



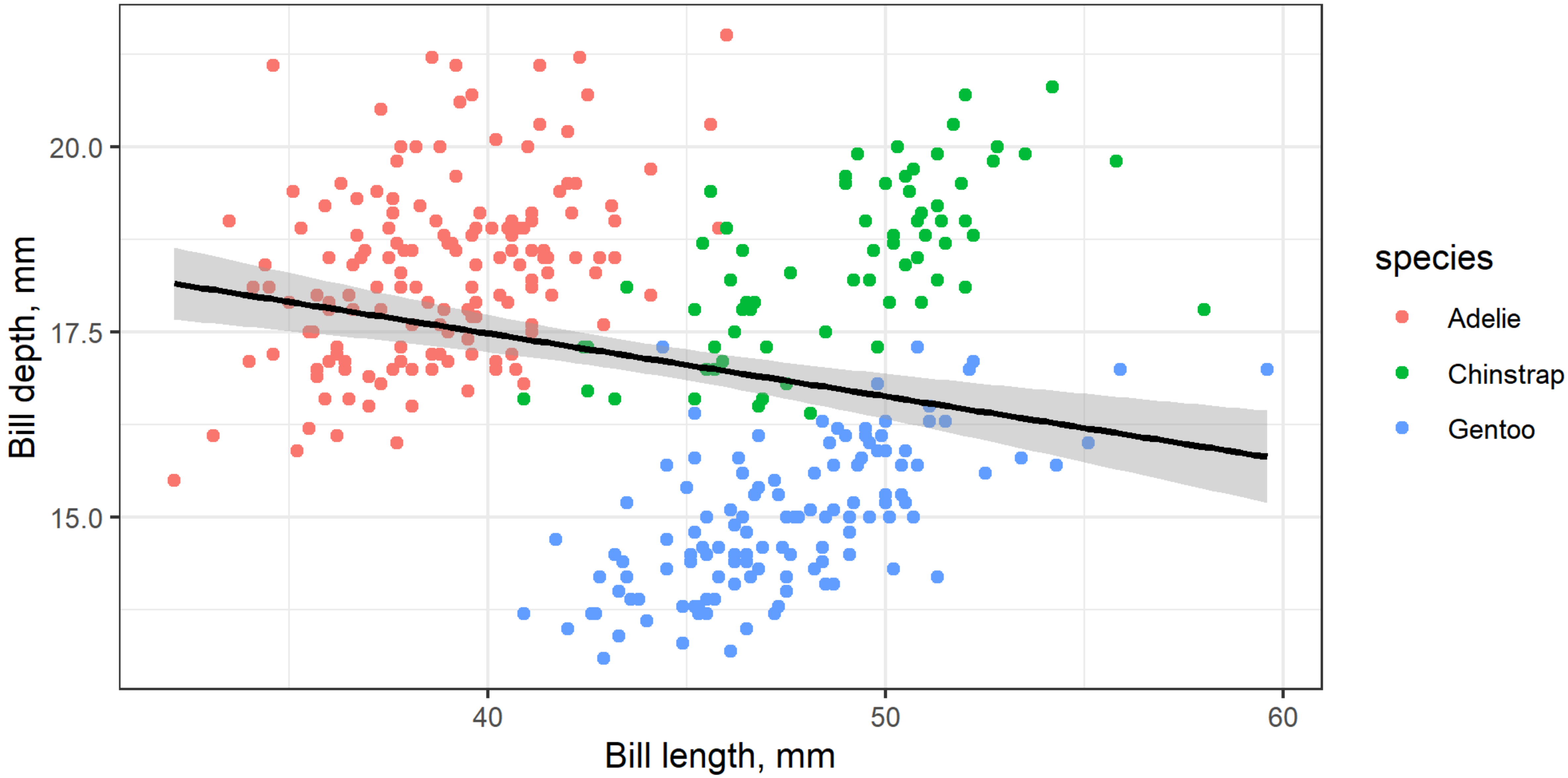
The trend

```r
gg_bill +
  geom_smooth(method = "lm", formula = y ~ x, colour = "black")
```



Is this the true trend?

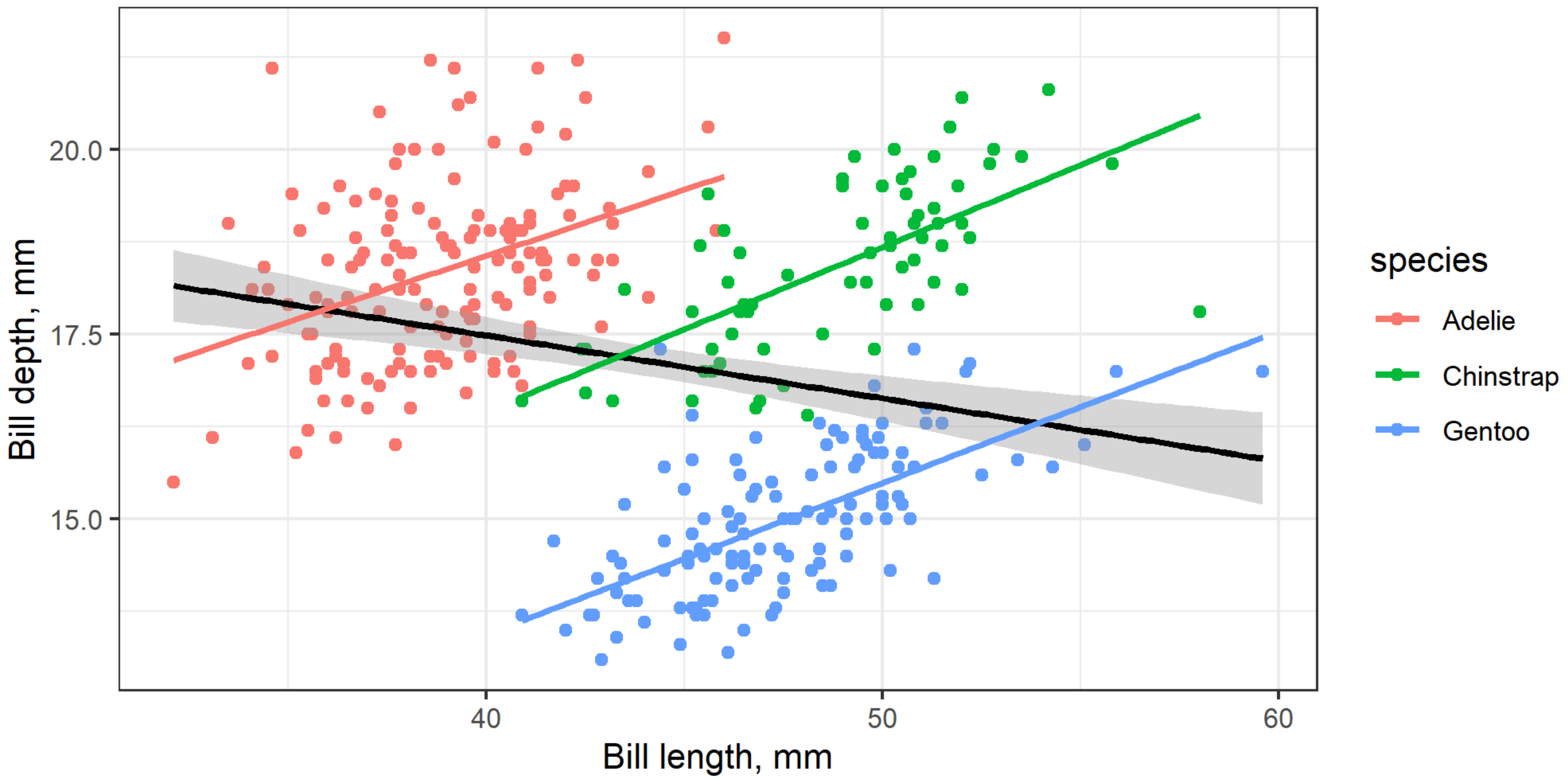
```
gg_bill +
  geom_smooth(method = "lm", formula = y ~ x, colour = "black") +
  aes(colour = species)
```



JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

The true trends

```r
gg_bill +
  geom_smooth(method = "lm", formula = y ~ x, colour = "black") +
  aes(colour = species) +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE)
```



Regression results

```r
fit1 <- lm(bill_depth_mm ~ bill_length_mm, data = penguins)
fit2 <- lm(bill_depth_mm ~ bill_length_mm + species, data = penguins)
fit3 <- lm(bill_depth_mm ~ -1 + bill_length_mm + species, data = penguins)
modelsummary(
  list(fit1, fit2, fit3),
  estimate = "{estimate}{stars} ({std.error})",
  statistic = NULL,
  gof_map = c("nobs", "adj.r.squared")
)
```

	(1)	(2)	(3)
(Intercept)	20.885*** (0.844)	10.592*** (0.683)	
bill_length_mm	−0.085*** (0.019)	0.200*** (0.017)	0.200*** (0.017)
speciesChinstrap		−1.933*** (0.224)	8.659*** (0.862)
speciesGentoo		−5.106*** (0.191)	5.486*** (0.835)
speciesAdelie			10.592*** (0.683)
Num.Obs.	342	342	342
R2 Adj.	0.052	0.767	0.997

Simpson’s paradox conclusion

Trends or relationships are observed in the whole population, but they reverse or disappear, when each group is treated separately.

Causes:

- Unobserved heterogeneity/differences between groups.
- Underlining processes that are different between parts of the population.

Resolutions to the paradox:

- Control variables in the MRL.
- Panel data.

Data Types

Cross-sectional data

ID	Y	X1	X2
1	y_1	x_1^1	x_1^2
2	y_2	x_2^1	x_2^2
3	y_3	x_3^1	x_3^2
4	y_4	x_4^1	x_4^2
5	y_5	x_5^1	x_5^2
6	y_6	x_6^1	x_6^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	y_N	x_N^1	x_N^2

- Data that we usually collect in a single data collection.
  - Each individual is represented by one observation.
- Could be repeatedly collected multiple times (repeated cross-section),
  - but, in every repetition, there are different individuals!

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Panel data

ID	Time	Y	X1	X2
1	1	y_{11}	x^1_{11}	x^2_{11}
1	2	y_{12}	x^1_{12}	x^2_{12}
1	3	y_{13}	x^1_{13}	x^2_{13}
2	2	y_{22}	x^1_{22}	x^2_{22}
2	3	y_{23}	x^1_{23}	x^2_{23}
3	1	y_{31}	x^1_{31}	x^2_{31}
3	2	y_{32}	x^1_{32}	x^2_{32}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	1	y_{N1}	x^1_{N1}	x^1_{N1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	T	y_{NT}	x^1_{NT}	x^2_{NT}

- table with data, where
- **each individual** (cohort, e.i. region, country)
- is represented by **multiple observations** at **different time periods**.

Panel data: Balanced and Unbalanced

Balanced

Each individual is represented in all time periods.

ID	Time	Y	X
1	1	Y_{11}	X_{11}
1	2	Y_{12}	X_{12}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	T	Y_{1T}	X_{1T}
2	1	Y_{21}	X_{21}
2	2	Y_{22}	X_{22}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	T	Y_{2T}	X_{2T}
3	1	Y_{31}	X_{31}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	T	Y_{NT}	X_{NT}

Un balanced

Each individual only appears in some time periods (not all).

ID	Time	Y	X
1	1	Y_{11}	X_{11}
1	2	Y_{12}	X_{12}
2	2	Y_{22}	X_{22}
2	3	Y_{23}	X_{23}
3	3	Y_{33}	X_{33}
4	1	Y_{41}	X_{41}
5	2	Y_{52}	X_{52}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	T	Y_{NT}	X_{NT}

Regressions with Panel Data

Important

is a strategy to **control** for unobserved/omitted but fixed heterogeneity using **time** or **cohort (individual)** dimensions.

There are:

1. Pooled regression
2. Least-squares dummy variable (LSDV) model
3. **Fixed Effect** Panel Regression (within, first-difference and between)
4. **Random Effect** Panel Regression

Example 1: Effect of an employee's union membership on wage

Problem setting

Does the collective bargaining (union membership) has any effect on wages?

- See: (Card, 1996; Freeman, 1984)

$$log(\text{Wage}_{it}) = \beta_0 + \beta_1 \cdot \text{Union}_{it} + \beta_2 \cdot X_{it} + \beta_3 \cdot \text{Ability}_i + \epsilon_{it}$$

where i is the individual and t is the time dimension;

Is there an endogeneity problem?

$$log(\text{Wage}_{it}) = \beta_0 + \beta_1 \cdot \text{Union}_{it} + \beta_2 \cdot X_{it} + \beta_3 \cdot \text{Ability}_i + \epsilon_{it}$$

- Is there a source of endogeneity / selection bias here?
  - Any ideas?
- Ability_i not observable and not measurable;
  - time invariant;
  - correlates with X and Y;
- Omitting ability causes bias

One of the solutions:

- Ability are **time-invariant** and **unique to each individual**;
  - If we have multiple observation per each individual (**panel data**),
  - we can introduce dummy variables for each individual, to approximate ability.
- This is also called Fixed Effect - regression model
  - or a **within transformation** model
  - or **Difference in Difference**

Other solutions:

- Any ideas?
- Introduce **control variables** that are **proxy** of ability.
- Employ specific **research design**:
    - RCT
    - RDD

Empirical example

```r
library(tidyverse)
library(modelsummary)
wage_dta <- read_csv("wage_unon_panel.csv")
glimpse(wage_dta)
```

```
Rows: 4,165
Columns: 15
$ id      <dbl> 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3,…
$ year    <dbl> 82, 83, 84, 85, 86, 87, 88, 82, 83, 84, 85, 86, 87, 88, 82, 83…
$ exper   <dbl> 3, 4, 5, 6, 7, 8, 9, 30, 31, 32, 33, 34, 35, 36, 6, 7, 8, 9, 1…
$ hours   <dbl> 32, 43, 40, 39, 42, 35, 32, 34, 27, 33, 30, 30, 37, 30, 50, 51…
$ bluecol <chr> "no", "no", "no", "no", "no", "no", "no", "yes", "yes", "yes",…
$ ind     <dbl> 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
$ south   <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "no", "no", "…
$ smsa    <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no", "n…
$ married <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes",…
$ union   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1,…
$ educ    <dbl> 9, 9, 9, 9, 9, 9, 9, 11, 11, 11, 11, 11, 11, 11, 12, 12, 12, 1…
$ black   <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no", "n…
$ lwage   <dbl> 5.56068, 5.72031, 5.99645, 5.99645, 6.06146, 6.17379, 6.24417,…
```

The data

id	year	exper	hours	bluecol	ind	south	smsa	married	union	educ	black	lwage	fer
5	82	10	50	yes	0	no	no	yes	1	16	no	6.43775	
5	83	11	46	yes	0	no	no	yes	1	16	no	6.62007	
5	84	12	40	yes	0	no	no	yes	1	16	no	6.63332	
5	85	13	50	no	0	no	no	yes	0	16	no	6.98286	
5	86	14	47	yes	0	no	yes	yes	0	16	no	7.04752	
5	87	15	47	no	0	no	no	yes	0	16	no	7.31322	
5	88	16	49	no	0	no	no	yes	0	16	no	7.29574	
168	82	3	40	no	0	no	yes	yes	0	17	no	6.23245	
168	83	4	42	no	0	no	yes	yes	0	17	no	6.57925	
168	84	5	44	no	0	no	yes	yes	1	17	no	6.65286	
168	85	6	48	no	0	no	yes	yes	1	17	no	6.74524	
168	86	7	48	no	0	no	yes	yes	0	17	no	7.49554	
168	87	8	48	no	0	no	yes	yes	0	17	no	8.16052	
168	88	9	50	no	0	no	yes	yes	0	17	no	8.30820	

Pooled Regression

$$log(\text{Wage}_{it}) = \beta_0 + \beta_1 \cdot \text{Union}_{it} + \beta_2 \cdot X_{it} + \epsilon_{it}$$

- Regression model on all observations in the **panel data set** without any individual effects.

```
union_fit_0 <- lm(log(wage) ~ union + educ + exper + I(exper^2) + hours ,
                  data = wage_dta)
union_fit_0
```

```
Call:
lm(formula = log(wage) ~ union + educ + exper + I(exper^2) + 
    hours, data = wage_dta)

Coefficients:
(Intercept)        union         educ        exper   I(exper^2)        hours  
  4.7054380    0.1261467    0.0819744    0.0437155   -0.0006932    0.0077042  
```

Least-squares dummy variable (LSDV)

$$log(\mathrm{Wage}_{it}) = \beta_0 + \beta_1 \cdot \mathrm{Union}_{it} + \beta_2 \cdot X_{it} + \beta_3 \cdot \delta_i + \epsilon_{it}$$

- Pooled regression plus dummy variable for each individual.
- This is not a Fixed Effect Panel Regression!

```r
union_fit_1 <- lm(log(wage) ~ union + educ + exper + I(exper^2) + hours + factor(id),
                  data = wage_dta)
union_fit_1
```

```
Call:
lm(formula = log(wage) ~ union + educ + exper + I(exper^2) + 
    hours + factor(id), data = wage_dta)

Coefficients:
  (Intercept)          union           educ          exper     I(exper^2)  
    4.3485732      0.0300295      0.1023231      0.1137052     -0.0004234  
        hours    factor(id)2    factor(id)3    factor(id)4    factor(id)5  
    0.0007980     -2.2946317     -0.1234092     -2.3978310     -0.5366805  
  factor(id)6    factor(id)7    factor(id)8    factor(id)9   factor(id)10  
   -1.4597761     -1.2411420     -1.5180049      0.2078530     -0.1734390  
 factor(id)11   factor(id)12   factor(id)13   factor(id)14   factor(id)15  
   -1.4704689     -1.3608691     -1.3890339     -1.0439450     -0.9439428  
 factor(id)16   factor(id)17   factor(id)18   factor(id)19   factor(id)20  
```

Data structure in the LSDV

ID	Time	Y	X1	X2	δ_1	δ_2	δ_N
1	1	y_{11}	x_{11}^1	x_{11}^2	1	0	0
1	2	y_{12}	x_{12}^1	x_{12}^2	1	0	0
1	3	y_{13}	x_{13}^1	x_{13}^2	1	0	0
2	2	y_{22}	x_{22}^1	x_{22}^2	0	1	0
2	3	y_{23}	x_{23}^1	x_{23}^2	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	1	y_{N1}	x_{N1}^1	x_{N1}^1	0	0	1
N	2	y_{N2}	x_{N2}^1	x_{N2}^2	0	0	1

Results

```
modelsummary(
  list(
    `Pooled` = union_fit_0,
    `Least-squares dummy variable` = union_fit_1),
  estimate = "{estimate}{stars} ({std.error})",
  statistic = NULL,
  coef_map = c("(Intercept)", "union", "educ", "exper", "hours", "tenure"),
  gof_map = c("nobs", "adj.r.squared" , "df"),
  notes = "In the Least-squares dummy variable model we omitted all individual-related variables"
)
```

	Pooled	Least-squares dummy variable
(Intercept)	4.705*** (0.070)	4.349*** (0.289)
union	0.126*** (0.013)	0.030* (0.015)
educ	0.082*** (0.002)	0.102*** (0.027)
exper	0.044*** (0.002)	0.114*** (0.002)
hours	0.008*** (0.001)	0.001 (0.001)
Num.Obs.	4165	4165
R2 Adj.	0.298	0.891
DF	5	598

In the Least-squares dummy variable model we omitted all individual-related variables

Cross-sectional data and LSDV (1)

Can we run a LSDV model with the cross-sectional data?

- Any ideas?
- Why?….
- NO…
- Because the **number of independent variables have to be less then or equal to the number of observations**.

Cross-sectional data and LSDV (2)

ID	Y	X1	X2	δ_1	δ_2	δ_3	δ_N
1	y_1	x_1^1	x_1^2	1	0	0	0
2	y_2	x_2^1	x_2^2	0	1	0	0
3	y_3	x_3^1	x_3^2	0	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	y_N	x_N^1	x_N^1	0	0	0	1

Panel data and LSDV

LSDV model works with the panel data, but…

it is inefficient! Any ideas why?…

- Number of dummy variables is equal to the number of individuals + control variables.
  - If we have 5,000 individuals, we have 5,000+ regression coefficients.
  - What if we have 100,000 individuals?
- Having too many regressors remains unbiased, but complicates inference:
  - number of degrees of freedom increases;
  - adjusted R^2 may shrink to zero;

Panel regression: brief theory

Readings

Key readings:

- Mundlak (1961)
- Angrist & Pischke (2009) Ch. 5
- J. M. Wooldridge (2010);
- M. J. Wooldridge (2020);
- Söderbom, Teal, & Eberhardt (2014), Ch. 9-11

Other readings:

- Croissant & Millo (2018)

Terminology

Panel data has:

- i individuals (groups);
- t time periods **for each individual**; and
- k independent variables x

Panel Regression could be:

- **Pooled** OLS (regression without any panel structure);
- **Fixed Effect**:
  - **Least-squares dummy variable** (Pooled OLS + individual dummies);
  - **Within**-transformation panel regression **most commonly used**
  - **First-difference**, Between transformation panel regressions (look it up in (Croissant & Millo, 2018))
- **Random Effect** panel regression

Pooled OLS

OLS regression on the entire data set with panel structure.

$$y_{it} = \beta_0 + \beta_1 \cdot x_{1it} + \beta_2 \cdot x_{2it} + \cdots + \beta_k \cdot x_{kit} + \epsilon_{it}$$

- Estimates are biased because of the OVB.
- We assume the OVB to be time-invariant.

Least-squares dummy variable model

$$y_{it} = \beta_0 + \beta_1 \cdot x_{1it} + \beta_2 \cdot x_{2it} + \cdots + \beta_k \cdot x_{kit} + \gamma_i \cdot \delta_i + \epsilon_{it}$$

- Introduces a vector of dummy variables δ and estimated coefficients γ_i for each dummy variable.
- Estimates $\hat{\beta}$ and $\hat{\gamma}$ are unbiased (consistent) but inefficient.
- When there are too many δ_i (5000 or more), computer will have difficulties with estimating the coefficients…

Fixed Effect Panel Regression Model

Individual Fixed effect model:

$$y_{it} = \beta_1 \cdot x_{1it} + \beta_2 \cdot x_{1it} + \cdots + \beta_k \cdot x_{kit} + \color{red}{\alpha_i} + \epsilon_{it}$$

- $\color{red}{\alpha_i}$ are the individual-specific (i) fixed effect;
- usually without the intercept β_0;

Two-ways fixed effect model (individual + time effect):

$$\begin{gathered} y_{it} = \beta_1 \cdot x_{1it} + \beta_2 \cdot x_{1it} + \cdots + \beta_k \cdot x_{kit} \\ + \color{red}{\alpha_i} + \color{blue}{\eta_t} + \epsilon_{it} \end{gathered}$$

Time Fixed Effect model:

$$\begin{gathered} y_{it} = \beta_1 \cdot x_{1it} + \beta_2 \cdot x_{1it} + \cdots + \beta_k \cdot x_{kit} \\ + \color{blue}{\eta_t} + \epsilon_{it} \end{gathered}$$

Fixed Effect Model: Within transformation (Step 1)

Within-transformation subtracts **group means** from each observation and estimates β on transformed data using OLS.

$$
\begin{gathered}
y_{it} - \overline{y_i} = \beta_1 (x_{1it} - \overline{x_{1i}}) + \beta_2 (x_{2it} - \overline{x_{2i}}) \\
+ \beta_3 (x_{3i} - \overline{x_{3i}}) + \color{red}{\alpha_i} + \epsilon_{it} \, ,
\end{gathered}
$$

$\overline{y_i}$ and $\overline{x_{ki}}$ are group i-specific means computed as: $\overline{x_{ki}} = \frac{1}{N_i} \sum_t x_{kit}$, where N_i is the number of observations (time periods t) in the group i.

$\ddot{y}_i = y_{it} - \overline{y_i}, \ddot{x}_i = x_{it} - \overline{x_i}$ are **de-meaned** regressand and regressors.

- Note! $\beta_3 = 0$, because any **time-invariant** x_{ki} (x_k without t index) will become zero: $x_i - \overline{x_i} = 0$.
  - Such x are: gender, race, individual characteristics …

JUSTUS-LIEBIG-UNIVERSITAT GIESSEN

Fixed Effect Model: Within transformation (Step 2)

Based on the demeaned data without time-invariant effects, OLS method is used to estimate $\hat{\beta}$ for all k variables:

$$\ddot{y}_i = \hat{\beta}_1 \ddot{x}_{1it} + \hat{\beta}_2 \ddot{x}_{2it} + \cdots + \hat{\beta}_k \ddot{x}_{kit} + \epsilon_{it}$$

Estimated $\hat{\beta}$ are identical to the one obtain using LSDV model!

Fixed Effect Model: Within transformation (Step 3)

Individual Fixed Effects α_i are computed as:

$$\alpha_i = \overline{y}_i - (\hat{\beta}_1 \overline{x}_{1i} + \hat{\beta}_2 \overline{x}_{2i} + \cdots + \hat{\beta}_k \overline{x}_{ki})$$

Individual fixed effects are identical to δ_i from LSDV model:

$$\alpha_i = \beta_0 + \delta_i$$

Ignoring FE causes bias to the estimates.

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Fixed Effect Model: assumptions (1)

- **NOT ZERO** correlation between fixed effects α_i and (not de-meaned) regressors x_{kit}:
  - $Cov(\alpha_i, x_{kit}) \neq 0$
- **Strict exogeneity (No endogeneity):**
  - $E[\epsilon_{is} | x_{kit}, \alpha_i] = 0$
  - $Cov(\epsilon_{is}, x_{kit}) = 0$ and $Cov(\epsilon_{it}, x_{kjt}) = 0$, where $j \neq i$ and $s \neq t$;
  - Residuals (ϵ) do not correlate with all explanatory variable (x_k) in all time periods (t) and for all individuals (i).
- Variance homogeneity:
  - No autocorrelation/serial correlation: $Cov(\epsilon_{it}, X_{i,t-1}) = 0$;
  - No cross-sectional dependence: $Cov(\epsilon_{it}, X_{j,t}) = 0$ (when individual observations react similarly to the common shocks or correlate in space);

Panel Regression FE model not less important assumptions (2)

- All Gauss-Markov assumptions
  - Linearity
  - Random sampling
  - No endogeneity
  - No collinearity
- Homoscedasticity of error terms: $Var(\delta_i|X_{it}) = \sigma^2_{\delta}$
- Normality of the residuals

Fixed effect application: literature

Seminal papers: (Mundlak, 1961)

Climate and agriculture: Bozzola, Massetti, Mendelsohn, & Capitanio (2017)

Choice of irrigation: Chatzopoulos & Lippert (2015)

Crop choice: Seo & Mendelsohn (2008b)

Livestock choice: Seo & Mendelsohn (2008a)

Cross-sectional dependence: (Conley, 1999)

Random Effect Model (individual, time and two-ways)

- Introduce random components v_i and/or u_t

$$y_{it} = \beta_0 + \beta_1 \cdot x_{1it} + \cdots + \beta_k \cdot x_{kit} \\ + v_i + u_t + \epsilon_{it}$$

- Difference from the fixed effect model:
  - Assumes NO CORRELATION (ZERO CORRELATION) between random effects and regressors:
    - $Cov(v_i, X_{it}) = 0$
  - Ignoring RE causes no bias to the estimates;

Summary on the Panel Regression

Fixed Effect (within transformation)

- Assumes that Fixed Effects correlate with regressors!
- Partially resolves the OVB.
- Ignoring FE (using pooled regression) causes bias of estimates.

Random Effect

- Assumes that Random Effects do NOT correlate with regressors
- Do NOT resolved any OVB.
- Provides additional control strategy, but ignoring RE causes NO bias.

- Both require valid Gauss–Markov assumptions.

Limitations of the Fixed and Random effect models

- NOT the ultimate solution to Endogeneity.
- OVB may still remain after applying the fixed effects.
- Measurement error is a problem in panel data.

Panel Regression Example: Union and wages

Problem setting

Does the collective bargaining (union membership) has any effect on wages?

- See: (Card, 1996; Freeman, 1984)

$$log(\text{Wage}_{it}) = \beta_0 + \beta_1 \cdot \text{Union}_{it} + \beta_2 \cdot X_{it} + \beta_3 \cdot \text{Ability}_i + \epsilon_{it}$$

where i is the individual and t is the time dimension;

General algorithm

1. Pooled OLS
   - Choose an appropriate functional form (log/level);
   - Validate gauss-Markov assumption validation: Linearity, Collinearity, Random Sampling; Homoscedasticity;
   - Note on the ‘No endogeneity’ assumption (if not validated, shows importance of the FE model)
2. FE: Fixed Effect. Within-transformation. Individual, Time or Two-ways effects;
   - `F-test` on FE consistency against pooled.
   - `LM test` on FE Individual, Time or Two-ways effects consistency against each other.
   - If tests suggest the pooled model, but the theory emphasizes FE, discus and reason your choice.
3. RE: Random Effect;
   - `Hausman test` on effects’ correlation with regressors of RE consistency against the FE;
   - Similar `Chamberlain test`, `Angrist and Newey` tests.
4. Serial correlation and cross-sectional dependence tests;
   - `Wooldridge's`, `Locally–Robust LM Test`, `Breusch–Godfrey Test`,
   - t > 3, we may have a serial correlation problem. Check it with a test.
   - Could individuals be affected by common shocks? We might have a cross-sectional dependence problem.
5. Use robust standard errors to correct for serial correlation and/or cross-sectional dependence:
   - Clustered SE and/or heteroscedasticity and/or autocorrelation robust SE;
6. Summary and interpretation;

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Step 1.a Pooled OLS

```r
library(tidyverse)
library(modelsummary)
library(parameters)
library(performance)
library(lmtest)
wage_dta <- read_csv("wage_unon_panel.csv")
glimpse(wage_dta)
```

```r
union_fit_0 <-
  lm(log(wage) ~ union + educ + exper + I(exper^2) + hours ,
     data = wage_dta)
union_fit_0
```

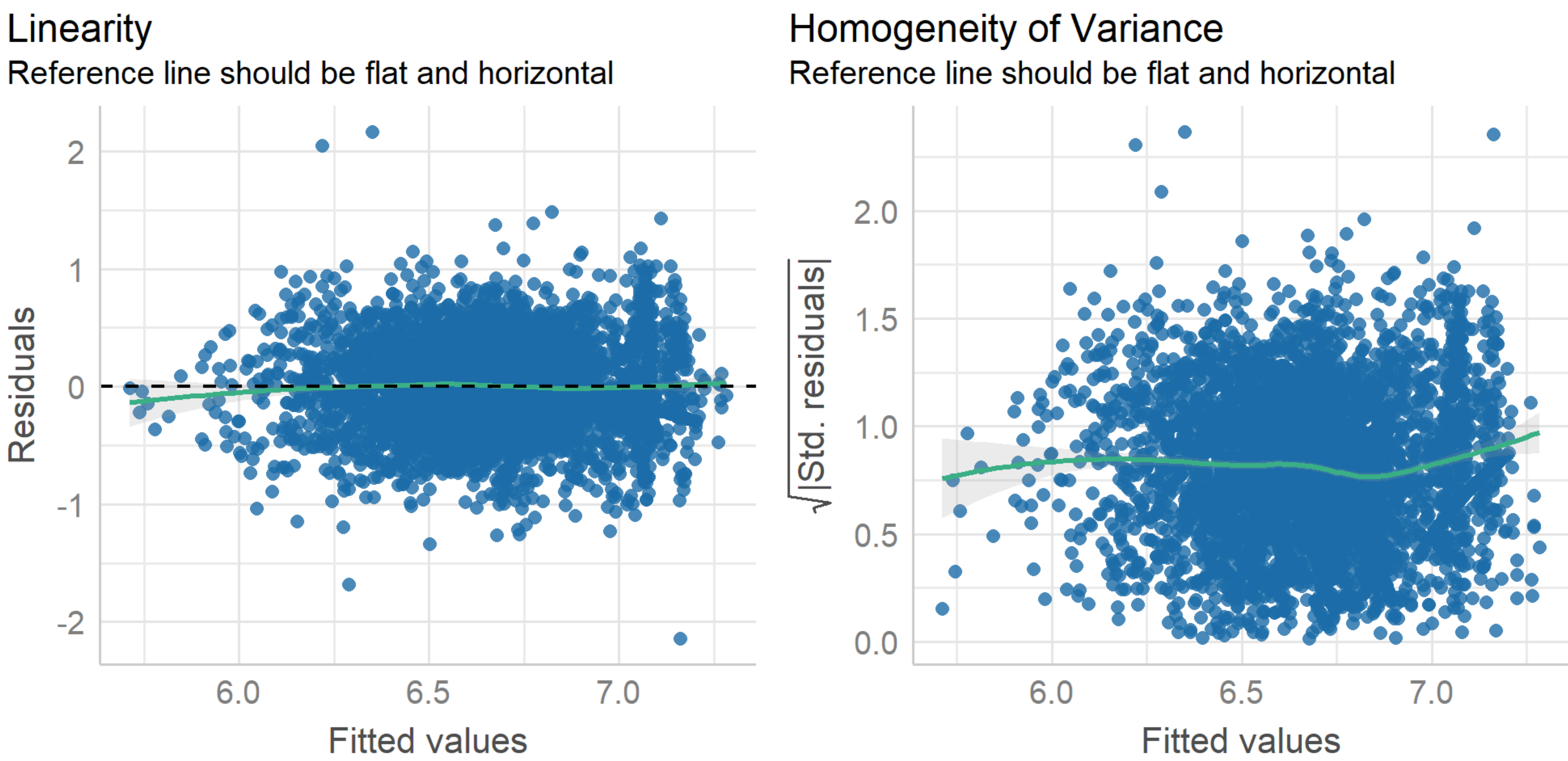
```

Call:
lm(formula = log(wage) ~ union + educ + exper + I(exper^2) + 
    hours, data = wage_dta)

Coefficients:
(Intercept)        union         educ        exper   I(exper^2)        hours  
  4.7054380    0.1261467    0.0819744    0.0437155   -0.0006932    0.0077042  
```

Step 1.b Assumptions (Linearity + Homoscedasticity)

```r
check_model(union_fit_0, check = c("linearity", "homogeneity"))
```



JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Step 1.b Assumptions (Homoscedasticity)

```r
check_heteroscedasticity(union_fit_0)
```

```
OK: Error variance appears to be homoscedastic (p = 0.647).
```

```r
bptest(union_fit_0)
```

```

	studentized Breusch-Pagan test

data:  union_fit_0
BP = 92.71, df = 5, p-value < 2.2e-16
```


JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Step 1.b Assumptions (Collinearity)

```r
check_collinearity(union_fit_0)
```

```
# Check for Multicollinearity

Low Correlation

  Term  VIF     VIF 95% CI Increased SE Tolerance Tolerance 95% CI
 union 1.11 [ 1.08,  1.15]         1.05      0.90     [0.87, 0.93]
  educ 1.13 [ 1.10,  1.18]         1.06      0.88     [0.85, 0.91]
 hours 1.03 [ 1.01,  1.09]         1.02      0.97     [0.92, 0.99]

High Correlation

       Term   VIF     VIF 95% CI Increased SE Tolerance Tolerance 95% CI
      exper 18.81 [17.73, 19.95]         4.34      0.05     [0.05, 0.06]
 I(exper^2) 18.81 [17.73, 19.95]         4.34      0.05     [0.05, 0.06]
```

Step 1.b Assumptions (No endogeneity)

$$log(\text{Wage}_{it}) = \beta_0 + \beta_1 \cdot \text{Union}_{it} + \beta_2 \cdot X_{it} + \beta_3 \cdot \text{Ability}_i + \epsilon_{it}$$

Ability_i not observable and not measurable.

Omitting the ability may cause the OVB.

- No endogeneity assumption cannot be satisfied.
- We should exploit the panel data structure.

Step 2. FE: Fixed Effect (within)

Note, the new package: `plm` used for running panel regressions.

```r
library(plm)
```

Declare data to be panel.

```r
wage_dta_pan <- pdata.frame(wage_dta, index = c("id", "year"))
```

Check panel dimensions.

```r
pdim(wage_dta_pan)
```

```
Balanced Panel: n = 595, T = 7, N = 4165
```

Step 2. FE: Fixed Effect (within) (1)

Rerun the pooled regression with `plm`:

```
union_pooled <-
  plm(log(wage) ~ union + educ + exper + I(exper^2) + hours ,
      data = wage_dta, model = "pooling")
union_pooled
```

```
Model Formula: log(wage) ~ union + educ + exper + I(exper^2) + hours

Coefficients:
(Intercept)       union        educ       exper  I(exper^2)       hours 
 4.70543801  0.12614668  0.08197441  0.04371549 -0.00069316  0.00770422 
```

Fixed Effect (individual) model

```
union_fe_ind <-
  plm(log(wage)  ~ union + educ + exper + I(exper^2) + hours ,
      data = wage_dta, model = "within", effect = "individual")
union_fe_ind
```

```
Model Formula: log(wage) ~ union + educ + exper + I(exper^2) + hours

Coefficients:
      union       exper  I(exper^2)       hours 
 0.03002946  0.11370518 -0.00042343  0.00079804 
```

Step 2. FE: Fixed Effect (within) (2)

Fixed Effect (time) model

```
union_fe_time <-
  plm(log(wage) ~ union + educ + exper + I(exper^2) + hours ,
      data = wage_dta, model = "within", effect = "time")
union_fe_time
```

```
Model Formula: log(wage) ~ union + educ + exper + I(exper^2) + hours

Coefficients:
      union        educ       exper  I(exper^2)       hours
 0.12428314  0.07944664  0.03582803 -0.00058079  0.00756342
```

Fixed Effect (Two-ways) model

```
union_fe_twoways <-
  plm(log(wage) ~ union + educ + exper + I(exper^2) + hours ,
      data = wage_dta, model = "within", effect = "twoways")
union_fe_twoways
```

```
Model Formula: log(wage) ~ union + educ + exper + I(exper^2) + hours

Coefficients:
      union  I(exper^2)       hours
 0.02712246 -0.00040431  0.00064611
```

Step 2. `F-test` (1)

Which model to choose: Pooled or FE?

- Compares FE models (individual, time, two-ways) vs pooled
  - Pooled is always consistent vs FE
- Test logic:
  - H0: One model is inconsistent. (no individual/time/two-way effects)
  - H1: Both models are equally consistent.

- Run the test. Check the p-value
  - p-value < 0.05: FE is as good as pooled. Not using the FE model may lead to the bias.
  - p-value >= 0.05: Pooled is better than the FE model. Use pooled for interpretation.

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Step 2. F-test (2)

```
1 pFtest(union_fe_ind, union_pooled)
```

```
	F test for individual effects

data:  log(wage) ~ union + educ + exper + I(exper^2) + hours
F = 39.274, df1 = 593, df2 = 3566, p-value < 2.2e-16
alternative hypothesis: significant effects
```

- FE is preferred (pooled is biased)

```
1 pFtest(union_fe_twoways, union_pooled)
```

```
	F test for twoways effects

data:  log(wage) ~ union + educ + exper + I(exper^2) + hours
F = 39.309, df1 = 598, df2 = 3561, p-value < 2.2e-16
alternative hypothesis: significant effects
```

- Two-ways is preferred (pooled is biased)

```
1 pFtest(union_fe_time, union_pooled)
```

```
	F test for time effects

data:  log(wage) ~ union + educ + exper + I(exper^2) + hours
F = 154.34, df1 = 6, df2 = 4153, p-value < 2.2e-16
alternative hypothesis: significant effects
```

- Time FE is preferred (pooled is biased)

- `F-test` leads us to stick with the FE individual, two-ways or time regression.

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Step 2. `LM test`: Lagrange multiplier test (2)

Which FE model to choose: individual, time or two-way?

- Exist to compare FE models between each other assuming that:
  - **Pooled is always consistent** in pooled vs individual FE
  - **Individual FE always consistent** in individual FE vs time or two-way FE
- Test logic:
  - H0: One model is inconsistent.
  - H1: Both models are equally consistent.
- Run the test (one or another or both). Check p-value:
  - p-value < 0.05:
    - Individual FE is as good as pooled;
    - Time or Two-ways model is as good as individual FE;
  - p-value >= 0.05: Pooled or individual FE is better than the alternative

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Step 2. `LM test`Lagrange multiplier (2)

```
plmtest(union_pooled, effect = "individual")
```

```
	Lagrange Multiplier Test - (Honda)

data:  log(wage) ~ union + educ + exper + I(exper^2) + hours
normal = 70.727, p-value < 2.2e-16
alternative hypothesis: significant effects
```

```
plmtest(union_pooled, effect = "twoway")
```

```
	Lagrange Multiplier Test - two-ways effects (Honda)

data:  log(wage) ~ union + educ + exper + I(exper^2) + hours
normal = 186.47, p-value < 2.2e-16
alternative hypothesis: significant effects
```

```
plmtest(union_pooled, effect = "time")
```

```
	Lagrange Multiplier Test - time effects (Honda)

data:  log(wage) ~ union + educ + exper + I(exper^2) + hours
normal = 192.98, p-value < 2.2e-16
alternative hypothesis: significant effects
```

- Individual FE is preferred (pooled is biased)
- Individual FE and two-ways are both consistent. We can choose any of those two.
- Individual FE and time FE are both consistent. We can choose any of those two.

- All tests suggest that individual, time and two-ways fixed effect models are equally consistent.

Step 3. Random Effect model (individual)

```r
union_rand_ind <-
  plm(log(wage) ~ union + educ + exper + I(exper^2) + hours ,
      data = wage_dta, model = "random", effect = "individual")
union_rand_ind
```

```
Model Formula: log(wage) ~ union + educ + exper + I(exper^2) + hours

Coefficients:
(Intercept)       union        educ       exper  I(exper^2)       hours 
 3.80500063  0.05488827  0.11346029  0.08804403 -0.00077520  0.00095463 
```

Step 3. Hausman test

Which model to choose: Fixed effect or Random effect?

- Compares Fixed Effect model with Random Effect:
  - Fixed effect model is always consistent
- Test logic:
  - H0: One model is inconsistent. Use FE!
  - H1: Both models are equally consistent. RE is as good as FE.
- Run the test. Check the p-value.
  - p-value < 0.05: Use FE or RE, both are good.
  - p-value >= 0.05: Use FE, discard RE.

```r
phtest(union_fe_ind, union_rand_ind)
```

```
	Hausman Test

data:  log(wage) ~ union + educ + exper + I(exper^2) + hours
chisq = 6183.7, df = 4, p-value < 2.2e-16
alternative hypothesis: one model is inconsistent
```

- FE is preferred instead of the RE model.

Step 4.1 Wooldridge's test (1)

Is there serial correlation / cross-sectional dependence in the data?

- Wooldridge's test for unobserved individual effects
  - H0: no unobserved effects
  - H1: some effects exist due to cross-sectional dependence and/or serial correlation

- Run the test Check the p-value.
  - p-value < 0.05: cross-sectional dependence and/or serial correlation are present
  - p-value >= 0.05: No cross-sectional dependency and/or serial correlation

Step 4.1 Wooldridge's test (2)

```r
pwtest(union_pooled, effect = "individual")
```

```
    Wooldridge's test for unobserved individual effects

data:  formula
z = 13.865, p-value < 2.2e-16
alternative hypothesis: unobserved effect
```

```r
pwtest(union_pooled, effect = "time")
```

```
    Wooldridge's test for unobserved time effects

data:  formula
z = 2.015, p-value = 0.04391
alternative hypothesis: unobserved effect
```

- cross-sectional dependence is present
- serial correlation is present

Step 4.2 Lagrange-Multiplier tests (1)

Is there serial correlation in the data?

- Locally–Robust Lagrange Multiplier Tests for serial correlation
  - H0: serial correlation is zero
  - H1: some serial correlation is present
- Run the test Check the p-value.
  - p-value < 0.05: serial correlation need to be addressed
  - p-value >= 0.05: no serial correlation

Step 4.2 Lagrange-Multiplier tests (2)

```r
pbsytest(union_pooled, test = "ar")
```

```
    Bera, Sosa-Escudero and Yoon locally robust test

data:  formula
chisq = 608.98, df = 1, p-value < 2.2e-16
alternative hypothesis: AR(1) errors sub random effects
```

- serial correlation is present

Step 5. Robust inference

Serial correlation and/or cross-sectional dependence render our Standard errors useless.

- Cross-sectional dependence and/or serial correlation violate the variance homogeneity assumption:
  - Estimates are unbiased, but inefficient.
  - Standard errors need to be corrected.

We need to use:

- **Robust Standard Errors**, and/or
- Clustered SE at the individual (group) level

Step 5. Robust Standard Error (1)

```r
library(lmtest)
library(car)
library(sandwich)
options(digits = 3, scipen = 6)
union_fe_ind
```

```
Model Formula: log(wage) ~ union + educ + exper + I(exper^2) + hours

Coefficients:
     union      exper I(exper^2)      hours 
  0.030029   0.113705  -0.000423   0.000798 
```

Correcting cross-sectional dependence:

Step 5. Robust Standard Error (2)

We produce new Variance-covariance matrix:

```
vcovHC(union_fe_ind,
       method = "white1",
       type = "HC0",
       cluster = "group")
```

```
                   union          exper       I(exper^2)            hours
union         0.0002534526 -0.0000028527  0.000000054371 -0.000000443094
exper        -0.0000028527  0.0000067329 -0.000000126361  0.000000043707
I(exper^2)    0.0000000544 -0.0000001264  0.000000002902  0.000000000683
hours        -0.0000004431  0.0000000437  0.000000000683  0.000000566054
attr(,"cluster")
[1] "group"
```

- `methods` for cross–sectional dependence “white1” and “white2” and for cross–sectional dependence and autocorrelation “arellano”;
- `type` for sample size correction: “HC0”, “sss”, “HC1”, “HC2”, “HC3”, “HC4” (“HC3” is recommended);
- `cluster` enabled by default (“group” or “time”);

Step 6. Reporting results (1)

```r
pooled_robust <-
  coeftest(union_pooled,
           vcov. = vcovHC(union_pooled, method = "arellano",
                          type = "HC3", cluster = "group"))

pooled_cs_robust <-
  coeftest(union_fe_ind,
           vcov. = vcovHC(union_fe_ind, method = "white1",
                          type = "HC0", cluster = "group"))

pooled_csac_robust <-
  coeftest(union_fe_ind,
           vcov. = vcovHC(union_fe_ind, method = "arellano",
                          type = "HC3", cluster = "group"))
```

```r
modelsummary(
  list(
    `Pooled (no SE correction)` = coeftest(union_pooled),
    `Pooled (c/s dep. and aut.)` = pooled_robust,
    `Ind. FE (no SE correction)` = coeftest(union_fe_ind),
    `Ind. FE (c/s dep.)` = pooled_cs_robust,
    `Ind. FE (c/s dep. and aut.)` = pooled_csac_robust
    ),
  fmt = 4, statistic = NULL,
  estimate = "{estimate}{stars} ({std.error})")
```

Step 6. Reporting results (1)

	Pooled (no SE correction)	Pooled (c/s dep. and aut.)	Ind. FE (no SE correction)	Ind. FE (c/s dep.)	Ind. FE (c/s dep. and aut.)
(Intercept)	4.7054*** (0.0699)	4.7054*** (0.1383)			
union	0.1261*** (0.0131)	0.1261*** (0.0255)	0.0300* (0.0148)	0.0300+ (0.0159)	0.0300 (0.0256)
educ	0.0820*** (0.0023)	0.0820*** (0.0052)			
exper	0.0437*** (0.0024)	0.0437*** (0.0053)	0.1137*** (0.0025)	0.1137*** (0.0026)	0.1137*** (0.0040)
I(exper^2)	−0.0007*** (0.0001)	−0.0007*** (0.0001)	−0.0004*** (0.0001)	−0.0004*** (0.0001)	−0.0004*** (0.0001)
hours	0.0077*** (0.0012)	0.0077*** (0.0019)	0.0008 (0.0006)	0.0008 (0.0008)	0.0008 (0.0009)
Num.Obs.	4165	4165	4165	4165	4165
AIC	3911.1	3911.1	−4502.1	−4502.1	−4502.1
BIC	3955.5	3955.5	−4470.5	−4470.5	−4470.5
Log.Lik.	-1948.564	-1948.564	2256.066	2256.066	2256.066

Step 6. Reporting GOF (1)

```
library(performance)
compare_performance(list(Pooled = union_pooled, FE = union_fe_ind))
```

```
# Comparison of Model Performance Indices

Name    | Model |   AIC (weights) |  AICc (weights) |   BIC (weights) |    R2 | R2 (adj.) |  RMSE
| Sigma
---------------------------------------------------------------------------------------------------
--------
Pooled |   plm | 59525.1 (<.001) | 59525.1 (<.001) | 59569.4 (<.001) | 0.299 |     0.298 | 0.386
| 0.387
FE     |   plm | 51111.8 (>.999) | 51111.8 (>.999) | 51143.5 (>.999) | 0.657 |     0.600 | 0.141
| 0.141
```

Takeaways

Takeaways for the exam

1. Simpson's paradox. What are the causes of it and solutions.
2. Data types (cross-section, repeated cross-section, balanced panel, unbalanced panel)
3. Panel Regression
   - Pooled;
   - Least Squared Dummy Variable model;
   - Fixed effect (within transformation);
   - Why FE is so important?
   - What is the key difference between FE and RE?
   - When FE and when RE are appropriate?
4. Panel Regression tests `F-test`, `LM-test`, `Hausman test`
5. Robust and Clustered SE:
   - Why these are important and when do we need to use one?

Homework

1. Reproduce code from the slides
2. Perform practical exercises.

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics*. Princeton University Press. http://doi.org/10.1515/9781400829828

Blanc, E., & Schlenker, W. (2017). The use of panel models in assessments of climate impacts on agriculture. *Review of Environmental Economics and Policy*, *11*(2), 258–279. http://doi.org/10.1093/reep/rex016

Bozzola, M., Massetti, E., Mendelsohn, R., & Capitanio, F. (2017). A ricardian analysis of the impact of climate change on italian agriculture. *European Review of Agricultural Economics*, *45*(1), 57–79. http://doi.org/10.1093/erae/jbx023

Card, D. (1996). The effect of unions on the structure of wages: A longitudinal analysis. *Econometrica*, *64*(4), 957. http://doi.org/10.2307/2171852

Chatzopoulos, T., & Lippert, C. (2015). Endogenous farm-type selection, endogenous irrigation, and spatial effects in ricardian models of climate change. *European Review of Agricultural Economics*, *43*(2), 217–235. http://doi.org/10.1093/erae/jbv014

Conley, T. G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics*, *92*(1), 1–45. http://doi.org/10.1016/s0304-4076(98)00084-0

Croissant, Y., & Millo, G. (2018). *Panel data econometrics with r*. John Wiley & Sons.

Freeman, R. B. (1984). Longitudinal analyses of the effects of trade unions. *Journal of Labor Economics*, *2*(1), 1–26. http://doi.org/10.1086/298021

Kurukulasuriya, P., Kala, N., & Mendelsohn, R. (2011). Adaptation and climate change impacts: A structural ricardian model of irrigation and farm income in africa. *Climate Change Economics*, *2*(02), 149–174.

Kurukulasuriya, P., Mendelsohn, R., Kurukulasuriya, P., & Mendelsohn, R. (2008). Crop switching as a strategy for adapting to climate change. http://doi.org/10.22004/AG.ECON.56970

Mendelsohn, R., Nordhaus, W. D., & Shaw, D. (1994). The impact of global warming on agriculture: A ricardian analysis. *The American Economic Review*, 753–771.

Mundlak, Y. (1961). Empirical production function free of management bias. *Journal of Farm Economics*, *43*(1), 44. http://doi.org/10.2307/1235460

Seo, S. N., & Mendelsohn, R. (2008a). An analysis of crop choice: Adapting to climate change in south american farms. *Ecological Economics*, *67*(1), 109–116. http://doi.org/10.1016/j.ecolecon.2007.12.007

Seo, S. N., & Mendelsohn, R. (2008b). Measuring impacts and adaptations to climate change: A structural ricardian model of african livestock management. *Agricultural Economics*, *38*(2), 151–165. http://doi.org/10.1111/j.1574-0862.2008.00289.x

Seo, S. N., Mendelsohn, R., Seo, S. N., & Mendelsohn, R. (2008). Animal husbandry in africa: Climate change impacts and adaptations. http://doi.org/10.22004/AG.ECON.56968

Söderbom, M., Teal, F., & Eberhardt, M. (2014). *Empirical development economics*. ROUTLEDGE. Retrieved from https://www.ebook.de/de/product/21466458/mans_soederbom_francis_teal_markus_eberhardt_empirical

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wooldridge, M. J. (2020). *Introductory econometrics: A modern approach*. South-Western. Retrieved from https://www.cengage.uk/shop/isbn/9781337558860