Panel Data Basics - Solutions

Exercise A - (20 min)

Download `fake-panel-data.csv` from `https://ditraglia.com/data`. This dataset was simulated according to the one-way error components model described above. It contains six columns: `person` is a unique person identifier (name), `year` is a year index (1-5), `x` and `y` are the regressor and outcome variable, and `epsilon` and `eta` are the error terms. (In real data you wouldn't have the errors, but this is a simulation!)

1. Use `lm` to regress `y` on `x` with "classical" standard errors. Repeat with standard errors clustered by `person` using `lm_robust()`. Discuss your results.
2. Plot `y` against `x` along with the regression line from part 1.
3. Repeat 2, but use a different color for the points that correspond to each person in the dataset and plot a *separate* regression line for each person.
4. What does the plot you made in part 3 suggest? Use the columns `epsilon` and `eta` to check your conjecture.
5. Finally, use `lm_robust()` to regress `y` on `x` *and* a dummy variable for each `person`, clustering the standard errors by `person`. Discuss your results.

Solution

Part 1

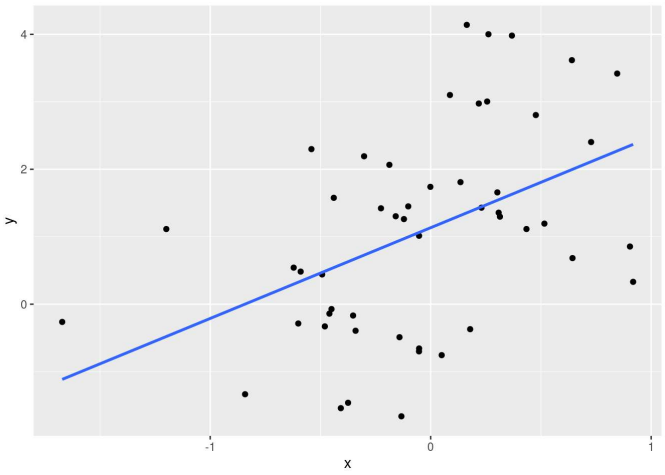
```
library(tidyverse)
library(estimatr)
library(modelsummary)
fake_panel <- read_csv('https://ditraglia.com/data/fake-panel-data.csv')

reg_classical <- lm(y~ x, fake_panel)
reg_cluster <- lm_robust(y ~ x, fake_panel, clusters = person)

modelsummary(list(Classical = reg_classical,
                  Clustered = reg_cluster),
             gof_omit = 'AIC|BIC|F|RMSE|R2|Log.Lik.')
```

	Classical	Clustered
(Intercept)	1.134	1.134
	(0.196)	(0.398)
x	1.345	1.345
	(0.378)	(0.595)
Num.Obs.	50	50
Std.Errors		by: person

Part 2

```
fake_panel |>
  ggplot(aes(x, y)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE)
```

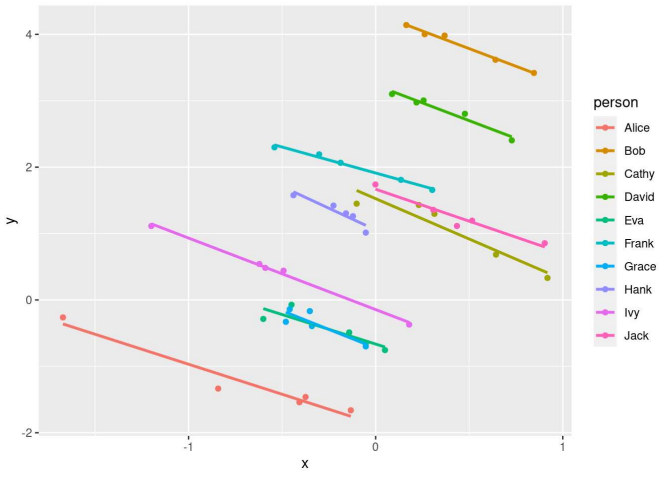
```
`geom_smooth()` using formula = 'y ~ x'
```



Part 3

```
fake_panel |>
  ggplot(aes(x, y, color = person)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Part 4

Each of the person-specific regression lines from the previous part slopes downwards, and the slopes appear to be quite similar. In contrast, the pooled regression line slopes upwards. From the plot in part 4, we see that people with higher values of `x` have systematically higher values of `y`. In other words, the *intercepts* of the person-specific regression lines appear to be correlated with `x`. This would occur if η_i were strongly correlated with x_{it} and indeed, we find that it is:

```
fake_panel |>
  summarize(cor(x, eta))
```

```
# A tibble: 1 × 1
  `cor(x, eta)`
          <dbl>
1         0.666
```

Part 5

```
reg_cluster_dummies <- lm_robust(y ~ x + person - 1, fake_panel,
                                 clusters = person)

modelsummary(list(Classical = reg_classical,
                  Clustered = reg_cluster,
                  `Clustered/Dummies` = reg_cluster_dummies),
             gof_omit = 'AIC|BIC|F|RMSE|R2|Log.Lik.',
             coef_omit = 'person')
```

	Classical	Clustered	Clustered/Dummies
(Intercept)	1.134	1.134	
	(0.196)	(0.398)	
x	1.345	1.345	−0.999
	(0.378)	(0.595)	(0.052)
Num.Obs.	50	50	50
Std.Errors		by: person	by: person

Exercise B - (10 min)

1. Use `dplyr` to subtract the individual time averages from `x` and `y` in the simulated dataset from above. Then run OLS on the demeaned dataset with classical SEs.
2. Compare the point estimates and standard errors from 1 to those from an OLS regression of `y` on `x` and a full set of `person` dummies, again with classical SEs.
3. Consult `?lm_robust()` to find out how to use the `fixed_effects` option. Use what you learn to regress `y` on `x` with `person` fixed effects, clustering by `person`.
4. Compare your results from 3 to mine computed using `feols()` above *and* to your calculations with `lm_robust()` and clustered standard errors from Exercise A above.

Solution

```
library(fixest)

reg_demeaned <- fake_panel |>
  group_by(person) |>
  mutate(x = x - mean(x), y = y - mean(y)) |>
  ungroup() |>
  lm(y ~ x, data = _)

reg_dummies <- lm(y ~ x + person - 1, fake_panel)

reg_robust_FE <- lm_robust(y ~ x, data = fake_panel,
          clusters = person, fixed_effects = ~ person)

reg_robust_dummies <- lm_robust(y ~ x + person - 1, data = fake_panel,
          clusters = person)

reg_FE <- feols(y ~ x | person, fake_panel)

modelsummary(list(Demeaned = reg_demeaned,
                  `lm` = reg_dummies,
                  `lm_robust/FE` = reg_robust_FE,
                  `lm_robust/dummies` = reg_robust_dummies,
```

```
    `feols` = reg_FE),
  gof_omit = 'AIC|BIC|F|RMSE|R2|Log.Lik.',
  coef_omit = 'person')
```

	Demeaned	lm	lm_robust/FE	lm_robust/dummies	feols
(Intercept)	0.000				
	(0.014)				
x	−0.999	−0.999	−0.999	−0.999	−0.999
	(0.044)	(0.049)	(0.052)	(0.052)	(0.050)
Num.Obs.	50	50	50	50	50
Std.Errors			by: person	by: person	by: person

Exercise C - (∞ min)

1. Install the `wooldridge` package and read the help file for `wagepan`.
2. Run an OLS regression of `lwage` on `educ`, `black`, `hisp`, `exper`, `exper` squared, married, union, and year. Use classical standard errors.
3. Repeat 2, but use `plm()` to estimate a random effects specification of the same model.
4. Repeat 3, but use `feols()` to estimate a fixed-effects specification with clustered standard errors. Can you include the same variables as in parts 2 and 3? Explain.
5. How do your estimates and standard errors of the effects of union membership vary across these three specifications? Discuss briefly.

Solution

```
library(wooldridge)
library(fixest)
library(plm)
library(modelsummary)

wagepan <- wagepan |>
  mutate(year = factor(year))

ols_formula <- lwage ~ educ + black + hisp + exper + I(exper^2) + married +
  union + year

pooled_ols <- lm(ols_formula, wagepan)

random_effects <- plm(ols_formula, data = wagepan,
                      index = c('nr', 'year'),
                      model = 'random')

# removed time-invariant regressors (married varies over time)
fe_formula <- lwage ~ exper + I(exper^2) + married + union + year | nr

# person id is `nr`
```

```
fixed_effects <- feols(fe_formula, wagepan) # Defaults to clustering by nf

modelsummary(list(OLS = pooled_ols, RE = random_effects, FE = fixed_effects),
             coef_omit = 'year|Intercept',
             gof_omit = 'AIC|BIC|F|RMSE|R2|Log.Lik.')
```

	OLS	RE	FE
educ	0.091	0.092	
	(0.005)	(0.011)	
black	−0.139	−0.139	
	(0.024)	(0.048)	
hisp	0.016	0.022	
	(0.021)	(0.043)	
exper	0.067	0.106	0.132
	(0.014)	(0.015)	(0.012)
I(exper^2)	−0.002	−0.005	−0.005
	(0.001)	(0.001)	(0.001)
married	0.108	0.064	0.047
	(0.016)	(0.017)	(0.021)
union	0.182	0.106	0.080
	(0.017)	(0.018)	(0.023)
Num.Obs.	4360	4360	4360
Std.Errors			by: nr