Problem Set 4

MOEC0021 Empirical Methods

Fenqi Guo Wenjie Tu

Fall Semester 2020

Endogeneity and Panel Data

1. Theory - Endogeneity

1(a)

$$\ln u_i = \beta_0 + \beta_1 \ln b_i + \nu_i \tag{1}$$

- u_i is individual i's unemployment duration.
- b_i denotes the level of unemployment benefits for which individual i was eligible upon losing her job.
- ν_i is the error term.

We are interested in the **elasticity** of unemployment duration to the benefit level. Hence it makes sense to estimate the model in logs since

$$\epsilon_{u,b} = \frac{\frac{\partial u}{u}}{\frac{\partial b}{b}} = \frac{\partial \ln u}{\partial \ln b} = \beta_1$$

1(b)

Unemployment duration is likely to depend on other factors besides unemployment benefits. In particular, we may well believe that we cannot directly compare unemployed workers eligible for high benefits with those eligible for low benefits: working experience and skills affect both the level of benefits and unemployment duration.

Because of the endogeneity, our model will yield a biased estimte of the **causal effect** of an increase in unemployment benefits on unemployment duration (the unemployment duration elasticity).

1(c)

$$\ln u_i = \beta_0 + \beta_1 \ln \bar{b}_s + \beta_2 \mathrm{X}_{i,s} + \nu_{i,s} \tag{2}$$

- s indexes the state in which individual i lives.
- $\bar{b}_s$ is the average level of unemployment benefit in state s.
- $\mathrm{X}_{i,s}$ are individual controls (scuh as age, past employment history or education).

When using the average level of unemployment benefit $\bar{b}_s$, we ignore the variation in benefit eligibility that stems from an individual's previous working history. Because no individual will be able to influence the benefit generosity in a given state, we hope that using variation at the state level only solves the endogeneity problem that was caused by the use of $b_{i,s}$.

Yet, it might still be true that the pool of unemployed workers is quite different from one area to another (e.g., unemployed workers in Zurich are more likely to have a university diploma than unemployed workers in Uri). These differences are likely to have an impact on the job finding rate and to be correlated with the average benefit level (high-skilled workers will both receive higher benefits and find another job more easily). Hence, if we don not control for individual characteristics, our model suffers from an omitted variable problem again. By controlling for individual characteristics, we hope to obtain an unbiased estimate of β_1.

1(d)

If policy-makers set benefit generosity as a function of the unemployment level, this model would be affected by a **simultaneity bias**.

1(e)

The estimates from regression (2) are not sufficient to recover the unemployment duration elasticity by themselves. The state-level averages in benefit generosity do not fully capture the different levels of individuals benefits for which workers are actually eligible.

1(f)

$$\ln \bar{b}_s = \gamma_0 + \gamma_1 \ln \hat{u}_s + \nu_s \tag{3}$$

- $\hat{u}_s$ denotes the predicted average unemployment duration.
- ν_s is an approximation error satisfying $\mathbb{E}(\nu_s) = 0$.

$$\hat{u}_s = \bar{u}_s \eta_s \tag{4}$$

- $\bar{u}_s$ is the true average unemployment duration (which is only observed after the government sets $\bar{b}_s$).
- $\eta_s > 0$ and $\mathbb{E}(\ln \eta_s) = 0$.

$$\ln \bar{u}_s = \beta_0 + \beta_1 \ln \bar{b}_s + \mathrm{X}_s \Gamma + \varepsilon_s \tag{5}$$

- X_s contains other economic variables that determine the unemployment rate duration.
- ν_s is the approximation error.
- η_s is the prediction error.
- The approximation error and the prediction error are independent from each other and from X_s and ε_s.

$$
\begin{aligned}
\ln \bar{u}_s &= \beta_0 + \beta_1 \ln \bar{b}_s + \mathrm{X}_s\Gamma + \varepsilon_s \\
\ln \bar{u}_s &= \beta_0 + \beta_1(\gamma_0 + \gamma_1 \ln \hat{u}_s + \nu_s) + \mathrm{X}_s\Gamma + \varepsilon_s \\
\ln \bar{u}_s &= \beta_0 + \beta_1(\gamma_0 + \gamma_1 \ln(\bar{u}_s\eta_s) + \nu_s) + \mathrm{X}_s\Gamma + \varepsilon_s \\
\ln \bar{u}_s &= \beta_0 + \beta_1\gamma_0 + \beta_1\gamma_1 \ln \bar{u}_s + \beta_1\gamma_1 \ln \eta_s + \beta_1\nu_s + \mathrm{X}_s\Gamma + \varepsilon_s \\
(1 - \beta_1\gamma_1) \ln \bar{u}_s &= \beta_0 + \beta_1\gamma_0 + \beta_1\gamma_1 \ln \eta_s + \beta_1\nu_s + \mathrm{X}_s\Gamma + \varepsilon_s \\
\ln \bar{u}_s &= \underbrace{\frac{\beta_0 + \beta_1\gamma_0}{1 - \beta_1\gamma_1}}_{\alpha_0} + \mathrm{X}_s \underbrace{\frac{\Gamma}{1 - \beta_1\gamma_1}}_{\alpha_1} + \underbrace{\frac{\beta_1\gamma_1 \ln \eta_s + \beta_1\nu_s + \varepsilon_s}{1 - \beta_1\gamma_1}}_{\mu_s} \\
\ln \bar{u}_s &= \alpha_0 + \mathrm{X}_s\alpha_1 + \mu_s
\end{aligned}
$$

Note that

$$
\begin{aligned}
\mathbb{E}(\mu_s|\mathrm{X}_s) &= \mathbb{E}\left(\frac{\beta_1\gamma_1 \ln \eta_s + \beta_1\nu_s + \varepsilon_s}{1 - \beta_1\gamma_1}\Big|\mathrm{X}_s\right) \\
&= \frac{1}{1 - \beta_1\gamma_1}\left(\beta_1\gamma_1 \underbrace{\mathbb{E}(\ln \eta_s|\mathrm{X}_s)}_{0} + \beta_1 \underbrace{\mathbb{E}(\nu_s|\mathrm{X}_s)}_{0} + \underbrace{\mathbb{E}(\varepsilon_s|\mathrm{X}_s)}_{0}\right) \\
&= 0
\end{aligned}
$$

Since $\mathbb{E}(\mu_s|\mathrm{X}_s) = 0$, running OLS on the reduced form yields consistent estimates of α_0 and α_1.

1(g)

$$
\begin{aligned}
\ln \bar{b}_s &= \gamma_0 + \gamma_1 \ln \hat{u}_s + \nu_s \\
&= \gamma_0 + \gamma_1 \ln(\bar{u}_s\eta_s) + \nu_s \\
&= \gamma_0 + \gamma_1 \ln \bar{u}_s + \gamma_1 \ln \eta_s + \nu_s \\
&= \gamma_0 + \gamma_1(\beta_0 + \beta_1 \ln \bar{b}_s + \mathrm{X}_s\Gamma + \varepsilon_s) + \gamma_1 \ln \eta_s + \nu_s
\end{aligned}
$$

This makes clear that $\ln \bar{b}_s$ **depends on** ε_s as well. Therefore $Cov(\ln \bar{b}, \varepsilon_s) \neq 0$, which leads the OLS estimate to be inconsistent.

Rearrange above equation:

$$
\begin{aligned}
\ln \bar{b}_s &= \gamma_0 + \gamma_1(\beta_0 + \beta_1 \ln \bar{b}_s + \mathrm{X}_s\Gamma + \varepsilon_s) + \gamma_1 \ln \eta_s + \nu_s \\
(1 - \gamma_1\beta_1) \ln \bar{b}_s &= \gamma_0 + \gamma_1\beta_0 + \gamma_1\mathrm{X}_s\Gamma + \gamma_1\varepsilon_s + \gamma_1 \ln \eta_s + \nu_s \\
\ln \bar{b}_s &= \frac{\gamma_0 + \gamma_1\beta_0}{1 - \gamma_1\beta_1} + \mathrm{X}_s\frac{\gamma_1\Gamma}{1 - \gamma_1\beta_1} + \frac{\gamma_1\varepsilon_s + \gamma_1 \ln \eta_s + \nu_s}{1 - \gamma_1\beta_1}
\end{aligned}
$$

For which we derive that

$$
Cov(\ln \bar{b}_s, \varepsilon_s) = \frac{\gamma_1}{1 - \gamma_1\beta_1} Var(\varepsilon_s)
$$

Because $Cov(\mathrm{X}, \varepsilon_s) = 0$, the sign of $Cov(\ln \bar{b}_s, \varepsilon_s)$ determines the sign of the asymptotic bias. It is reasonable to assume that $\gamma_1 > 0$ (the benefits are increasing in the unemployment level) and $\beta_1\gamma_1 < 1$ (a stability condition that prevents the system of equations from "exploding" when it is shocked). Hence we expect a **positive asymptotic bias**.

1(h)

We can use η_s as an **instrument** for $\ln \bar{b}_s$ in equation (5). Theoretically, it satisfies both the **relevance condition** and the **exclusion restriction** (it only impacts $\ln \bar{u}_s$ through $\ln \bar{b}_s$ since η_s is assumed to be independent from the other error terms and from X_s).

In practice, the variation in $\ln \bar{b}_s$ stemming from η_s is likely to be relatively small (i.e., $\frac{Var(\eta_s)}{Var(\ln \bar{b}_s)} \approx 0$) such that we might worry about the **strength of the instrument**.

1(i)

The optimal unemployment benefit level has to trade off the benefits of insurance against the risk of job loss and the model hazard effect on job search effort. When there is a high level of unemployment in economy, workers have a hard time finding a job even if they search hard for it. Hence the moral hazard component of providing more generous unemployment benefits is likely to be lower during recessions than in times of expansion. This can motivate setting higher benefits when the unemployment level is high.

2. Empirical Application - Panel Data

```
# import libraries
library(plm)
library(lfe)
library(stargazer)
library(ggplot2)
library(dplyr)
```

```
# read data
d.mort <- read.csv("mortality_temp.csv")

cols <- c("year", "stfips", "month")
d.mort[cols] <- lapply(d.mort[cols], factor)
```

```
## 'data.frame':    26460 obs. of  18 variables:
##  $ year       : Factor w/ 45 levels "1960","1961",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ stfips     : Factor w/ 49 levels "1","4","5","6",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ month      : Factor w/ 12 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ bin_1      : num  0 0 0.00122 0 0 ...
##  $ bin_2      : num  0.051 0.0163 0.1594 0 0 ...
##  $ bin_3      : num  2.45 1.23 1.67 0 0 ...
##  $ bin_4      : num  6.881854 6.757008 8.775334 0.028558 0.000359 ...
##  $ bin_5      : num  11.931 13.211 10.352 1.467 0.505 ...
##  $ bin_6      : num  6.21 5.75 5.81 6.24 5.96 ...
##  $ bin_7      : num  3.464 0.989 3.949 14.629 9.689 ...
##  $ bin_8      : num  0.0159 0.0486 0.2835 7.6359 14.3154 ...
##  $ bin_9      : num  0 0 0 0 0.533 ...
##  $ bin_10     : num  0 0 0 0 0 ...
##  $ devp25     : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ devp75     : int  0 0 0 0 0 0 0 1 1 0 ...
##  $ lndrate    : num  4.44 4.52 4.53 4.26 4.33 ...
##  $ lndrate_cvd: num  -3.15 -3.06 -3.03 -3.34 -3.25 ...
##  $ lndrate_mva: num  -6.25 -6.61 -6.15 -6.25 -5.77 ...
```

2(a)

```
bins <- paste("bin_", c(1:6, 8:10), sep = "")
reg.ols <- formula(paste("lndrate ~ ", paste(bins, collapse = " + ")))
reg.ols.cluster <- formula(paste("lndrate ~ ",
                                 paste(bins, collapse = " + "),
                                 " | 0 | 0 | stfips + month : year "))
bins
```

```
## [1] "bin_1"  "bin_2"  "bin_3"  "bin_4"  "bin_5"  "bin_6"  "bin_8"  "bin_9"
## [9] "bin_10"
```

```
reg.ols
```

```
## lndrate ~ bin_1 + bin_2 + bin_3 + bin_4 + bin_5 + bin_6 + bin_8 +
##     bin_9 + bin_10
```

```
reg.ols.cluster
```

```
## lndrate ~ bin_1 + bin_2 + bin_3 + bin_4 + bin_5 + bin_6 + bin_8 +
##     bin_9 + bin_10 | 0 | 0 | stfips + month:year
```

```
# standard errors
pooled.ols <- lm(reg.ols, data = d.mort)

# cluster robust standard errors
pooled.ols.cluster <- felm(reg.ols.cluster, data = d.mort)

stargazer(pooled.ols, pooled.ols.cluster, header = F,
          title = "Pooled OLS regressions in 2(a)",
          keep.stat = c("n", "rsq"), digits = 4, single.row = T)
```

The relationship between log mortality rate and temperature is likely to be **non-linear**. By discretizing the temperature om several bins, we can identify these non-linear effects. (In the limit, as the number of bins goes to infinity, we are going fully non-parametric and allow for a completely flexible relationship). The coefficient is $\hat{\theta}_{10} = -0.0104$. This indicates an **unexpected negative** partial correlation between the number of hot days and the mortality rate.

2(b)

Writing the estimating equation as

$$lndrate_{i,m,y} = \sum_{j=1, j\neq 7}^{10} \theta_j binj_{i,m,y} + \epsilon_{i,m,y}$$

We need

$$\mathbb{E}(binj_{i,m,y} \times \epsilon_{i,m,y}) = 0 \quad \forall j = 1, \cdots 6, 8, \cdots, 10$$

Table 1: Pooled OLS regressions in 2(a)

	Dependent variable:	
	lndrate	
	OLS	*felm*
	(1)	(2)
bin_1	0.0038*** (0.0007)	0.0038** (0.0017)
bin_2	0.0119*** (0.0010)	0.0119*** (0.0021)
bin_3	0.0032*** (0.0007)	0.0032** (0.0015)
bin_4	0.0064*** (0.0005)	0.0064*** (0.0011)
bin_5	0.0035*** (0.0003)	0.0035*** (0.0011)
bin_6	0.0039*** (0.0004)	0.0039*** (0.0005)
bin_8	0.0032*** (0.0004)	0.0032*** (0.0010)
bin_9	0.0024*** (0.0003)	0.0024* (0.0014)
bin_10	−0.0104*** (0.0009)	−0.0104*** (0.0020)
Constant	4.1981*** (0.0067)	4.1981*** (0.0259)
Observations	26,460	26,460
R^2	0.0893	0.0893
Note:		*p<0.1; **p<0.05; ***p<0.01

for consistency of the OLS estimator.

States in the south of the United States tend to be poorer and to experience higher temperatures. Hence, income is a confounding variable that prevents us from giving a causal interpretation to our estimates. We could not use the RE estimator to solve this problem since **RE requires the same assumptions as OLS for consistency**.

2(c)

```
# state + month fixed effects
reg.fe1 <- formula(paste("lndrate ~ ",
                         paste(bins, collapse = " + "),
                         " | stfips + month | 0 | stfips + month : year "))
reg.fe1
```

```
## lndrate ~ bin_1 + bin_2 + bin_3 + bin_4 + bin_5 + bin_6 + bin_8 +
##     bin_9 + bin_10 | stfips + month | 0 | stfips + month:year
```

```
fe.state.and.month <- felm(reg.fe1, data = d.mort)
stargazer(fe.state.and.month, header = F, single.row = T, digits = 4,
          keep.stat = c("n", "rsq"), title = "Additive fixed effects regression")
```

The coefficient ($\hat{\theta}_{10} = -0.0012$) is the partial correlation between the number of days in the hottest temperature bin and the log mortality rate. State dummies control for persistent differences in temperature and mortality across states throughout the sample period. Month dummies account for systematic variations in temperature and mortality in different months. The climate fluctuation over the year varies by state. This might lead to an omitted variable bias again if this fluctuation happens to be correlated with within-state fluctuation in mortality that is not caused by the temperature shocks.

Table 2: Additive fixed effects regression

	Dependent variable:
	lndrate
bin_1	0.0014 (0.0009)
bin_2	0.0051*** (0.0008)
bin_3	0.0019* (0.0010)
bin_4	0.0021*** (0.0006)
bin_5	0.0023*** (0.0004)
bin_6	0.0013*** (0.0004)
bin_8	−0.0008** (0.0003)
bin_9	−0.0012* (0.0006)
bin_10	−0.0012 (0.0009)
Observations	26,460
R^2	0.7704
Note:	*p<0.1; **p<0.05; ***p<0.01

2(d)

```
# state x month fixed effects
reg.fe2 <- formula(paste("lndrate ~ ",
                         paste(bins, collapse = " + "),
                         " | stfips : month | 0 | stfips + month:year "))
reg.fe2
```

```
## lndrate ~ bin_1 + bin_2 + bin_3 + bin_4 + bin_5 + bin_6 + bin_8 +
##     bin_9 + bin_10 | stfips:month | 0 | stfips + month:year
```

```
fe.state.by.month <- felm(reg.fe2, data = d.mort)
stargazer(pooled.ols.cluster, fe.state.and.month, fe.state.by.month,
          header = F, single.row = T, digits = 4,
          keep.stat = c("n", "rsq"), keep = "bin_10",
          column.labels = c("Pooled OLS", "Additive FE" ,"Multiplicative FE"),
          title = "Rregressions results comparison")
```

Table 3: Rregressions results comparison

	Dependent variable:		
	lndrate		
	Pooled OLS	Additive FE	Multiplicative FE
	(1)	(2)	(3)
bin_10	−0.0104*** (0.0020)	−0.0012 (0.0009)	0.0025** (0.0011)
Observations	26,460	26,460	26,460
R^2	0.0893	0.7704	0.7772
Note:			*p<0.1; **p<0.05; ***p<0.01

We are now identifying the coefficient uniquely from unusual temperatures in a given state-by-month cell (e.g., an unusual number of hot days in Arizona for the month of July). The coefficient has the expected sign now: having experienced an unusual temperature relative to average is expected to increase mortality rate by 0.25% on average.

2(e)

The estimating equation is

$$lndrate_{i,m,y} = \sum_{j=1, j\neq 7}^{10} \theta_j binj_{i,m,y} + \gamma_{i,m} + \nu_{m,y} + \epsilon_{i,m,y}$$

Defining the within deviations as

$$\widetilde{binj_{i,m,y}} = binj_{i,m,y} - \frac{1}{45}\sum_{y=1960}^{2004} binj_{i,m,y}$$

We need

$$\mathbb{E}(\widetilde{binj_{i,m,y}} \times \epsilon_{i,m,y}) = 0 \quad \forall j = 1, \cdots 6, 8, \cdots, 10$$

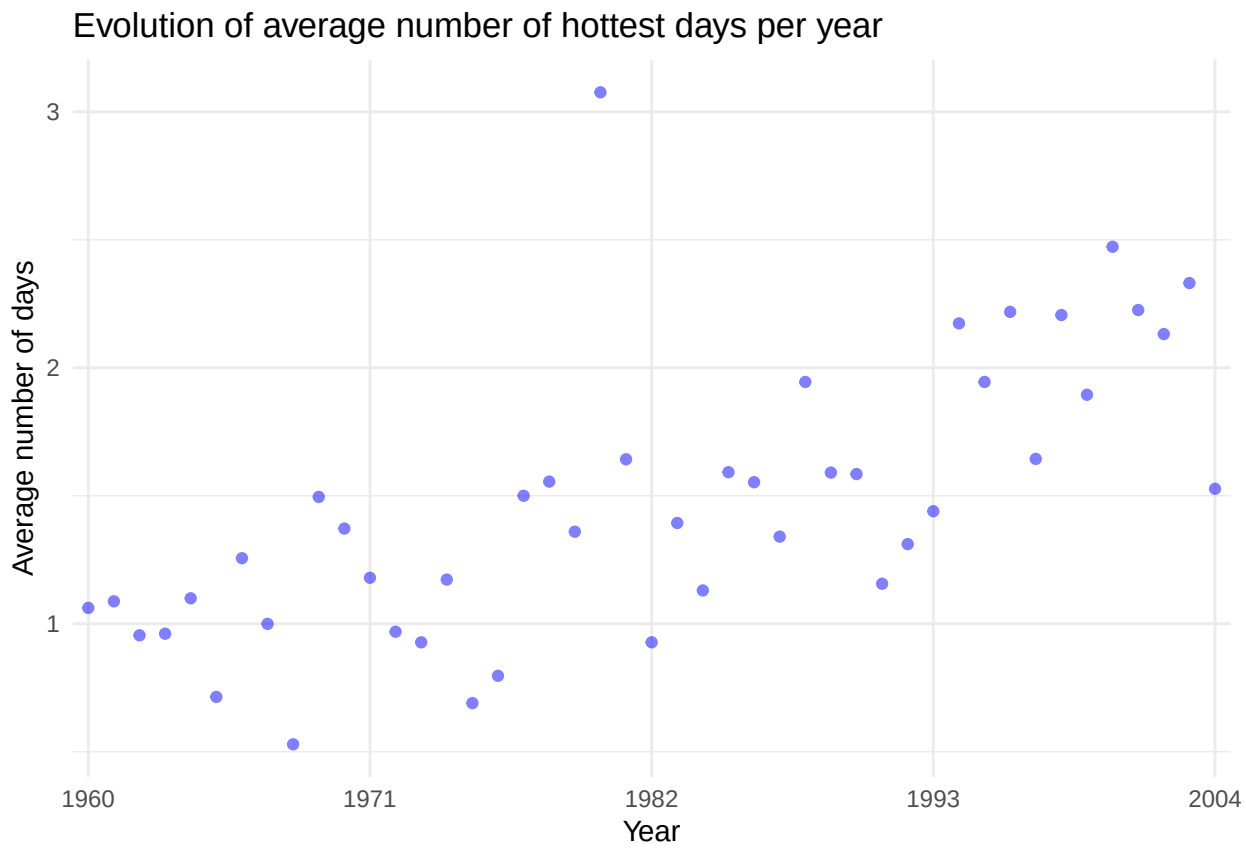
for consistency of the OLS estimator.

At the first sight, this assumption seems quite credible. It will be satisfied if the deviations from the "usual" number of days in each temperature bin for a given month and state are exogenous. Because year-to-year weather variations can be thought of as random, this exogeneity assumption seems to hold. One concern, however, is that temperature shocks may be accompanied by precipitations shocks (e.g., droughts or heavy rainfalls). Hence, coefficients on temperature fluctuations might pick up the effect of usual precipitations on mortality, leading to an omitted variable bias.

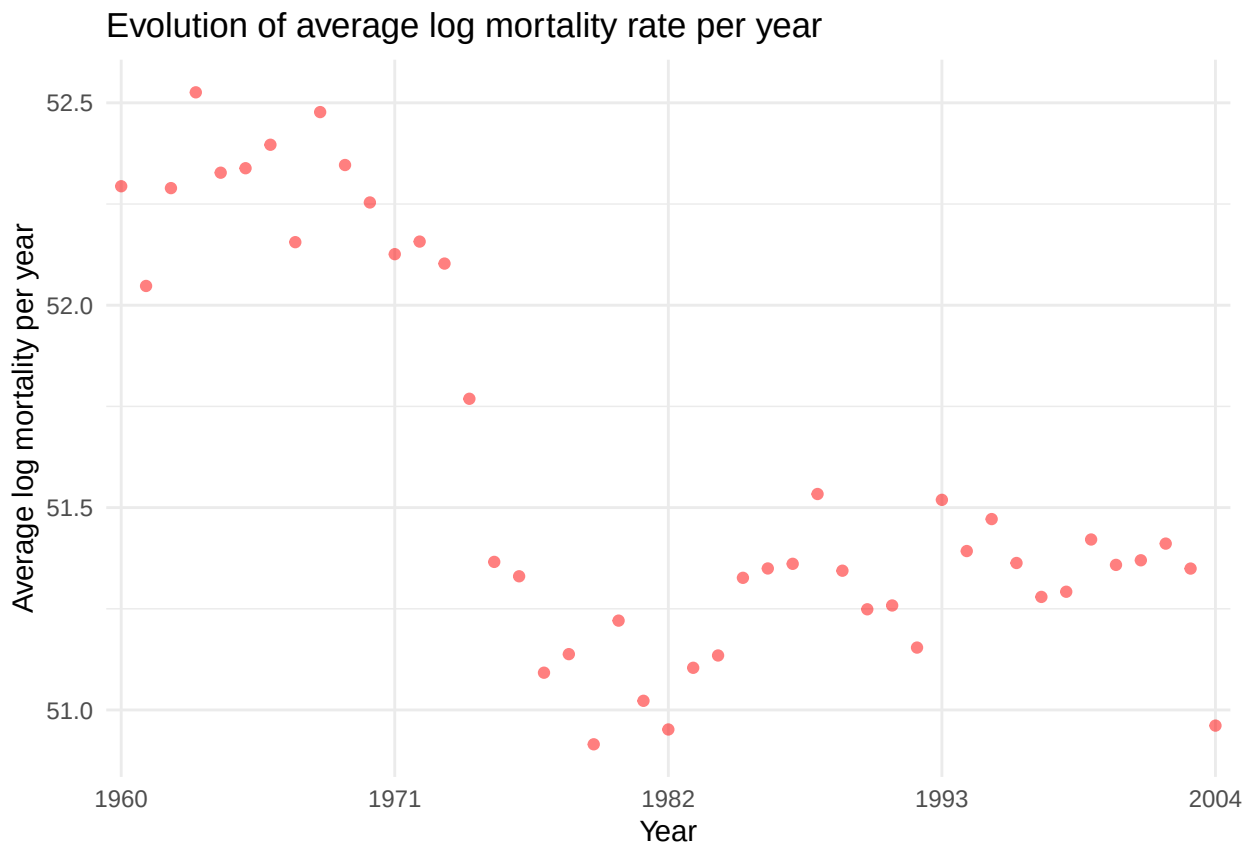
2(f)

```
d.plot <- d.mort %>% group_by(year) %>%
  summarise(days.ave = mean(bin_10) * 12,
            lndrate.ave = mean(lndrate) * 12,
            .groups = "drop")
```

```
ggplot(d.plot, aes(x = year, y = days.ave)) +
  geom_point(color = "blue", alpha = 0.5) +
  xlab("Year") + ylab("Average number of days") +
  ggtitle("Evolution of average number of hottest days per year") +
  theme_minimal() + scale_x_discrete(breaks = seq(1960, 2004, 11))
```




```
ggplot(d.plot, aes(x = year, y = lndrate.ave)) +
  geom_point(color = "red", alpha = 0.5) +
  xlab("Year") + ylab("Average log mortality per year") +
  ggtitle("Evolution of average log mortality rate per year") +
  theme_minimal() + scale_x_discrete(breaks = seq(1960, 2004, 11))
```



2(g)

```
# add a quadratic time trend
d.mort$year <- as.integer(d.mort$year)
d.mort$trend <- (d.mort$year - mean(d.mort$year))^2

d.mort[c("devp25", "devp75")] <- lapply(d.mort[c("devp25", "devp75")], factor)
```

```
vars <- c(bins, "devp25", "devp75", "year", "trend")
reg.fe3 <- formula(paste("lndrate ~ ", paste(vars, collapse = " + "),
                         " | stfips : month | 0 | stfips + month:year "))
reg.fe3
```

```
## lndrate ~ bin_1 + bin_2 + bin_3 + bin_4 + bin_5 + bin_6 + bin_8 +
##     bin_9 + bin_10 + devp25 + devp75 + year + trend | stfips:month |
##     0 | stfips + month:year
```

```
model.fe <- felm(reg.fe3, data = d.mort)
stargazer(model.fe, title = "Regression resulst in 2(g)",
          header = F, single.row = T, digits = 4,
          keep.stat = c("n", "rsq"), keep = "bin_10")
```

The coefficient is $\hat{\theta}_{10} = 0.0065$, almost three times larger than the coefficient obtained in 2(d). Because of global warming, a disproportionately large number of days in the hottest bin are observed in the last decades

Table 4: Regression resulst in 2(g)

	Dependent variable:
	lndrate
bin_10	0.0065*** (0.0011)
Observations	26,460
R^2	0.8271
Note:	*p<0.1; **p<0.05; ***p<0.01

of our sample. Since mortality rate declined throughout the years, the increase in the number of hot days biases the coefficient in 2(d) negatively.

2(h)

```
reg.fe4 <- formula(paste("lndrate_mva ~ ", paste(vars, collapse = " + "),
                         " | stfips : month | 0 | stfips + month:year "))
mva.fe <- felm(reg.fe4, data = d.mort)

reg.fe5 <- formula(paste("lndrate_cvd ~ ", paste(vars, collapse = " + "),
                         " | stfips : month | 0 | stfips + month:year "))
cvd.fe <- felm(reg.fe5, data = d.mort)
```

```
model.fe <- felm(reg.fe3, data = d.mort)
stargazer(model.fe, mva.fe, cvd.fe, title = "Regression results in 2(h)", header = F,
          single.row = T, digits = 4, keep.stat = c("n", "rsq"), keep = "bin_10")
```

Table 5: Regression results in 2(h)

	Dependent variable:		
	lndrate	lndrate_mva	lndrate_cvd
	(1)	(2)	(3)
bin_10	0.0065*** (0.0011)	−0.0060 (0.0061)	0.0133*** (0.0018)
Observations	26,460	26,453	26,460
R^2	0.8271	0.6784	0.8930
Note:			*p<0.1; **p<0.05; ***p<0.01

We obtain $\hat{\theta}_{10}^{mva} = -0.006$ when using the log mortality rate due to motor-vehicle accidents as dependent variable and $\hat{\theta}_{10}^{cvd} = -0.0133$ when using cardiovascular diseases. These results make us more inclined to attribute a causal interpretation to our estimate in 2(d).

Indeed, medical studies tell us that hotter temperatures should affect death hazard from cardiovascular diseases, while it should not matter for vehicle accidents. Choosing a variable that we think is not affected by the treatment as a dependent variable is a good test for the exogeneity of the treatment. It is akin to a **placebo test**.

2(i)

```
beta_10 <- coef(model.fe)["bin_10"]
bin_10.sd <- sd(d.mort$bin_10)
lndrate.sd <- sd(d.mort$lndrate)

magnitude_pct <- beta_10 * bin_10.sd * 100
magnitude_sd <- beta_10 * bin_10.sd / lndrate.sd
```

Arguably, the magnitude is quite small. One standard-deviation increase in the number of days in the hottest bins (≈ 1.08) leads to a $1.08 \times 0.0065 \approx 0.7\%$ increase in the mortality rate (log-level specification). This corresponds to an increase of about 0.05 standard deviation in mortality rate.