Data Analysis for Economic Policy 3

To what extent does economic activity cause CO2 emission?

Luca Keresztesi

February 13th, 2017

Consider the balanced panel (keep countries with full/almost full coverage over time)

Data on GDP per capita, PPP (constant 2005 international $) and on CO2 emission was downloaded through the World bank Data API. The time period for the dataset was from 1992 to 2016.

```r
# SEARCHING FOR DATA: GDP per capita
gdp_inds <- WDIsearch("gdp")
grep("2005", gdp_inds, value = TRUE)
gdppppCode <- gdp_inds[match("GDP per capita, PPP (constant 2005 international $)",
    gdp_inds[, 2]), 1]

# DATA DOWNLOAD: GDP per capita
dat = WDI(indicator = gdppppCode, start = 1992, end = 2016)

# FILTERING OUT REGIONS
dt <- data.table(dat)
exclusionList <- dt[, .(itemCnt = .N), by = .(code = dt$iso2c)][1:47,
    1]
gdpData <- subset(dt, !(dt$iso2c %in% exclusionList$code))

# SEARCHING FOR DATA: CO2
co2_inds <- WDIsearch("co2")
# CO2 emissions (metric tons per capita) CO2 emissions (kt)
co2code <- co2_inds[match("CO2 emissions (metric tons per capita)",
    co2_inds[, 2]), 1]

# DATA DOWNLOAD: CO2
co2dat = WDI(indicator = co2code, start = 1992, end = 2016)

# FILTERING OUT REGIONS
co2dt <- data.table(co2dat)
co2Data <- subset(co2dt, !(co2dt$iso2c %in% exclusionList$code))
```

One dataset was created from the merged datasets on GDP per capita and CO2 emission. Only countries with at least 22 yearly observations were kept in order to receive a balanced dataset.

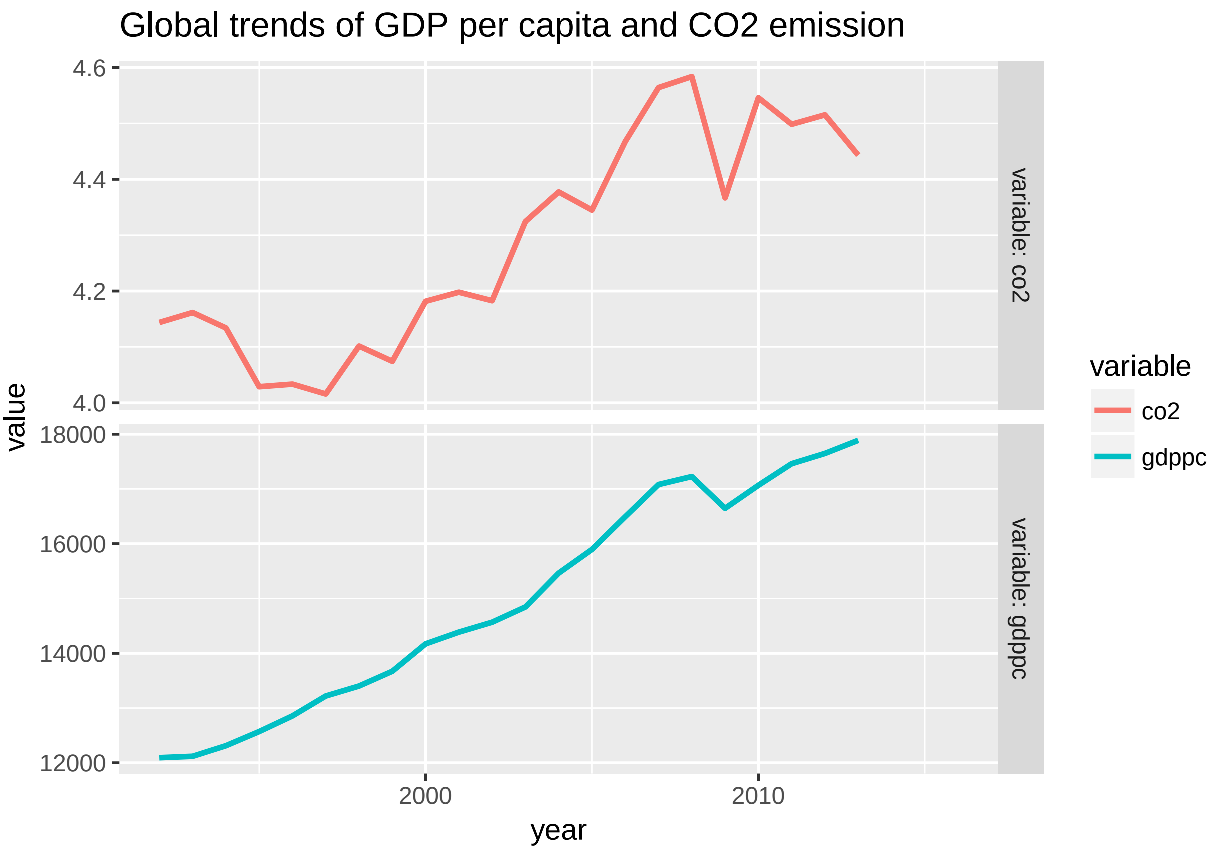
```r
# MERGING DATA INTO ORIGINAL PANEL
panelData <- merge(gdpData, co2Data, by.x = c("iso2c", "year",
    "country"), by.y = c("iso2c", "year", "country"))

dt <- NULL
dat <- NULL
co2dat <- NULL
co2dt <- NULL
exclusionList <- NULL
gdp_inds <- NULL
co2_inds <- NULL
```

```
panelData$iso2c <- NULL
names(panelData) <- c("year", "country", "gdppc", "co2")

# KEEPING ONLY COUNTRIES WITH AT LEAST 22 YEARLY OBSERVATIONS
dt <- panelData[!is.na(panelData$gdppc) & !is.na(panelData$co2),
    .(count = .N), by = country]
dta <- dt[count > 21]
panelData <- subset(panelData, (panelData$country %in% dta$country))
dt <- NULL
dta <- NULL
```

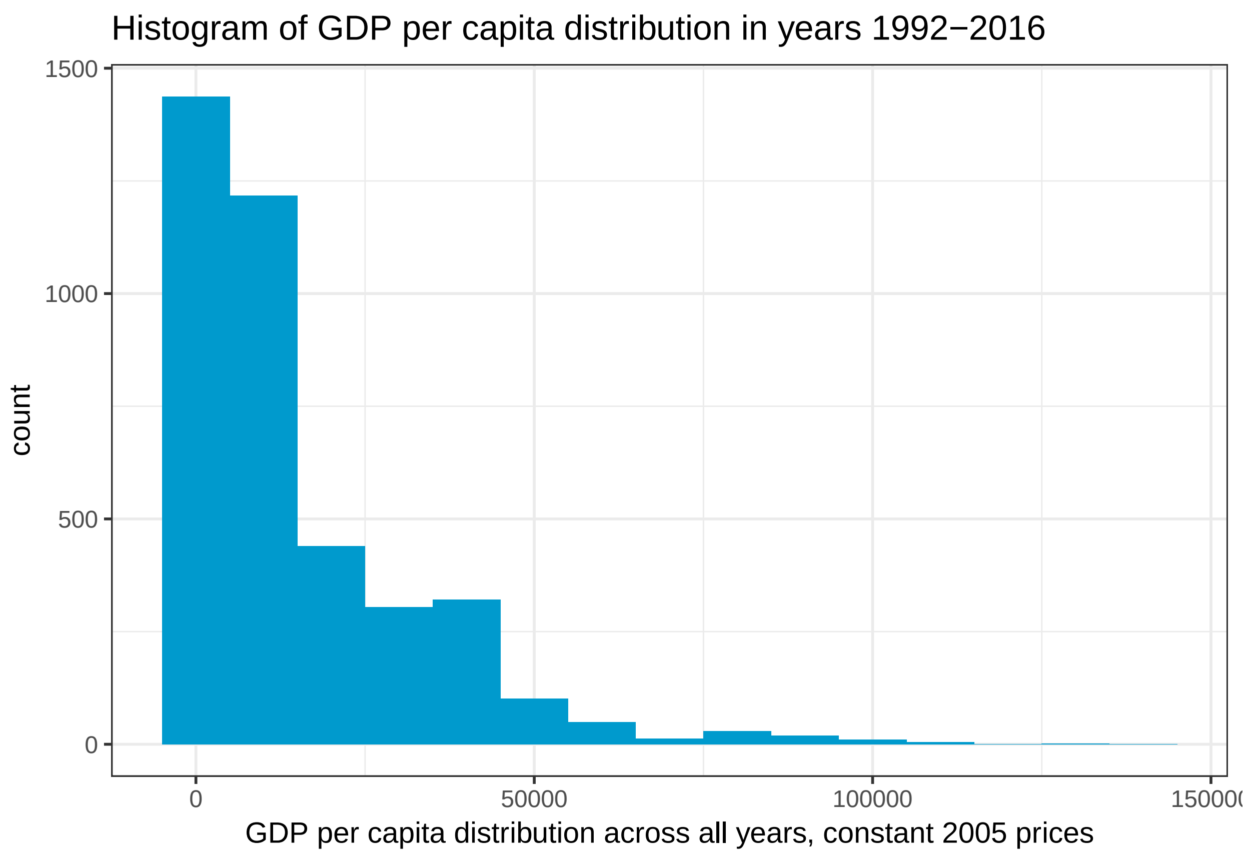
Descriptive statistics

It is visible on the global trend of GDP per capita and CO2 emission plotted as yearly averages of all countries in the dataset that the level of both variables increased on average on the long term, between 1992 and 2016.

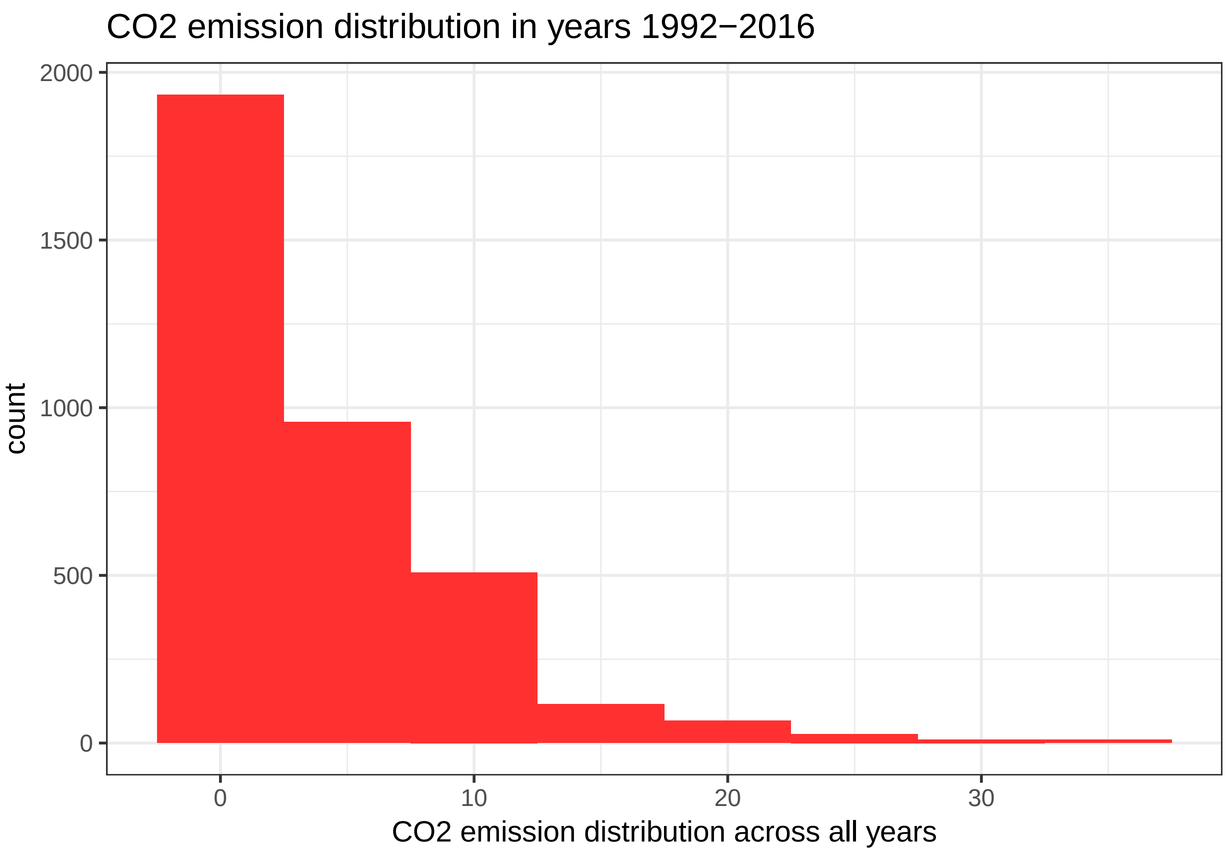
```
panelData %>% group_by(year) %>% summarize(gdppc = mean(gdppc),
    co2 = mean(co2)) %>% gather(variable, value, -year) %>% ggplot(aes(x = year,
    y = value)) + geom_line(aes(color = variable), size = 1) +
    facet_grid(variable ~ ., scales = "free", labeller = label_both) +
    ggtitle("Global trends of GDP per capita and CO2 emission")
```



```
ggplot(panelData) + aes(x = gdppc) + geom_histogram(binwidth = 10000,
    fill = "deepskyblue3") + labs(x = "GDP per capita distribution across all years, constant 2005 pric
    title = "Histogram of GDP per capita distribution in years 1992-2016") +
    theme_bw()
```



```r
ggplot(panelData) + aes(x = co2) + geom_histogram(binwidth = 5,
    fill = "firebrick1") + labs(x = "CO2 emission distribution across all years",
    title = "CO2 emission distribution in years 1992-2016") +
    theme_bw()
```



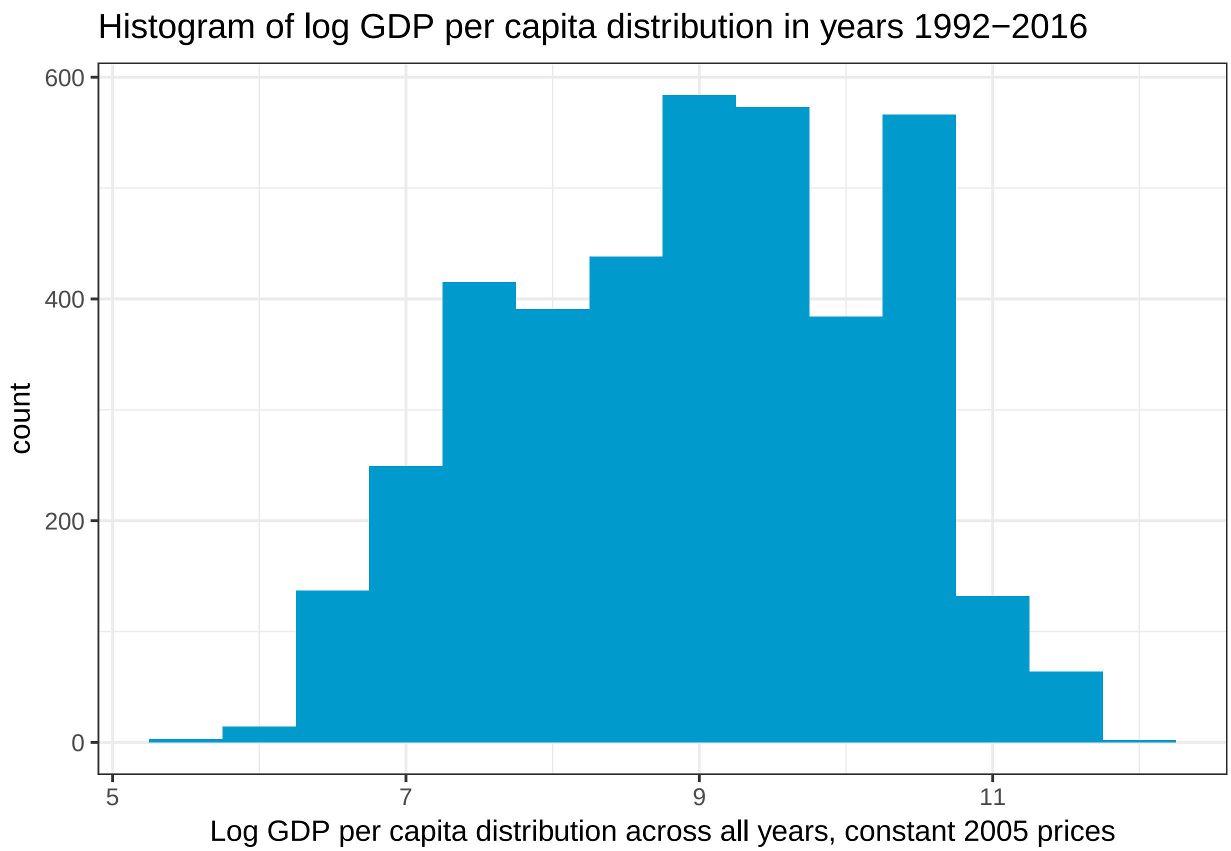
```r
dt <- data.frame(panelData)
stargazer(dt, out = "summary.html", header = FALSE, type = "latex",
    omit = "year", title = "Descriptive Statistics for the original variables")
```

Table 1: Descriptive Statistics for the original variables

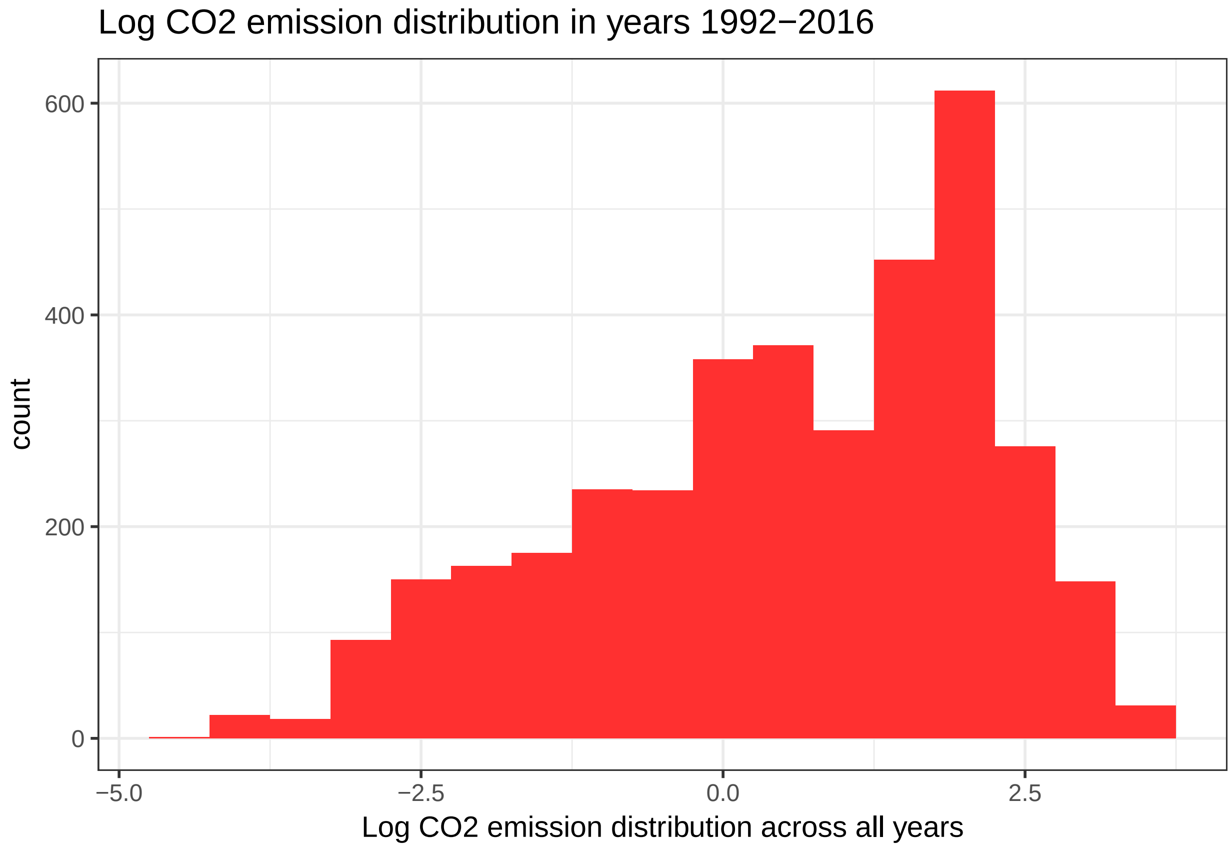
Statistic	N	Mean	St. Dev.	Min	Max
gdppc	3,952	15,209.020	17,649.080	246.671	137,164.400
co2	3,630	4.286	5.243	0.014	36.904

Both GDP per capita and CO2 emission are positively skewed and their distributions have a long right tail. Differences for both variables make more sense in relative than in absolute terms in the current context of the main question. Therefore log variables were created both for GDP per capita and CO2 emission.

```r
panelData$lngdppc <- log(panelData$gdppc)
panelData$lnco2 <- log(panelData$co2)

ggplot(panelData) + aes(x = lngdppc) + geom_histogram(binwidth = 0.5,
    fill = "deepskyblue3") + labs(x = "Log GDP per capita distribution across all years, constant 2005
    title = "Histogram of log GDP per capita distribution in years 1992-2016") +
    theme_bw()
```



```r
ggplot(panelData) + aes(x = lnco2) + geom_histogram(binwidth = 0.5,
    fill = "firebrick1") + labs(x = "Log CO2 emission distribution across all years",
    title = "Log CO2 emission distribution in years 1992-2016") +
    theme_bw()
```

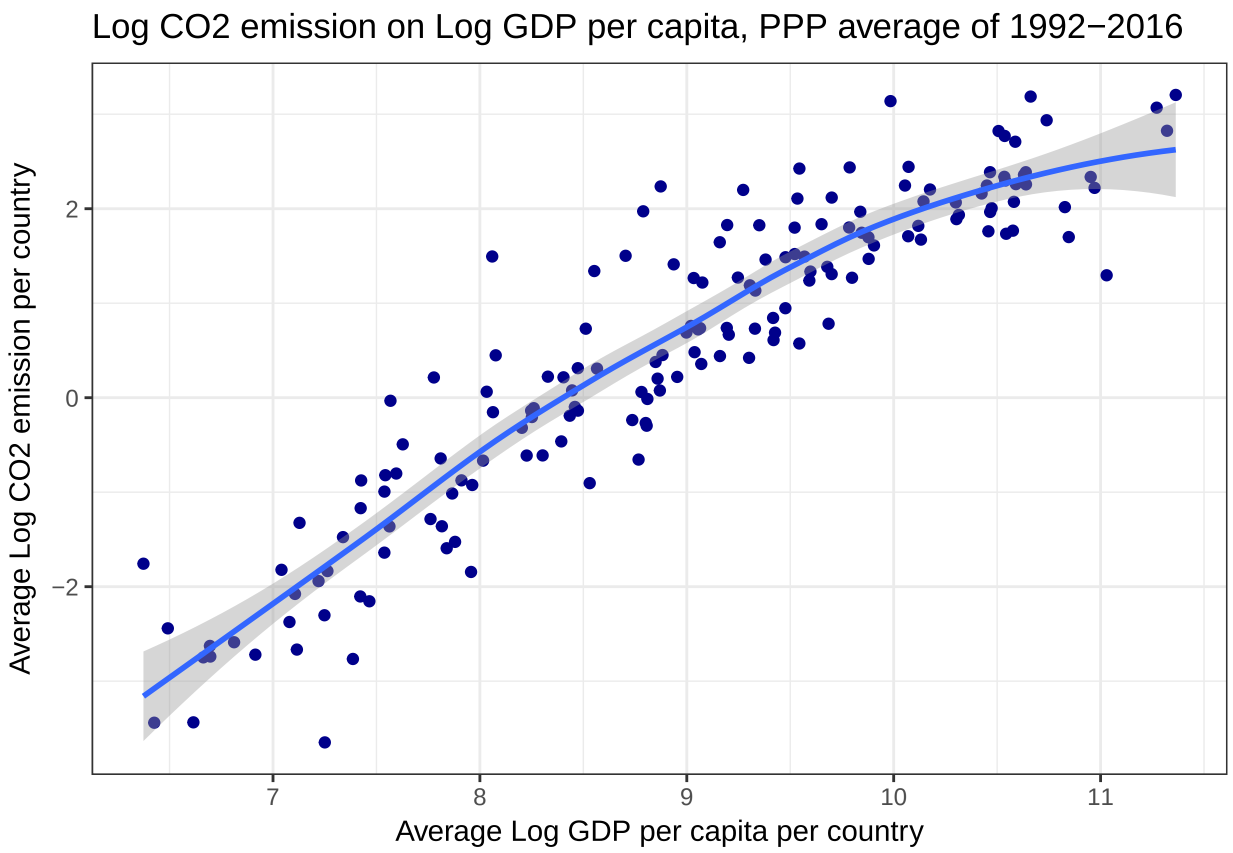


```
dt <- data.frame(panelData)
stargazer(dt, out = "summary_new.html", header = FALSE, type = "latex",
    omit = "year", title = "Descriptive Statistics for the transformed and original variables")
```

Table 2: Descriptive Statistics for the transformed and original variables

Statistic	N	Mean	St. Dev.	Min	Max
gdppc	3,952	15,209.020	17,649.080	246.671	137,164.400
co2	3,630	4.286	5.243	0.014	36.904
lngdppc	3,952	8.961	1.244	5.508	11.829
lnco2	3,630	0.510	1.650	−4.279	3.608

```
dt <- panelData[, .(m_lngdppc = mean(lngdppc, na.rm = TRUE),
    m_lnco2 = mean(lnco2, na.rm = TRUE)), by = .(country)]

ggplot(dt, aes(x = m_lngdppc, y = m_lnco2)) + geom_point(color = "darkblue") +
    geom_smooth() + labs(x = "Average Log GDP per capita per country",
    y = "Average Log CO2 emission per country", title = "Log CO2 emission on Log GDP per capita, PPP ave
    theme_bw()
```



It is visible on the scatterplot with a loess regression that in the 1992-2016 period average CO2 emission was higher in countries where average GDP per capita was higher, on average. This does not necessarily mean though that an increase in GDP per capita led to higher CO2 emission on the short or longer term. There might have been other events and other differences in countries that had an effect on CO2 emission.

Estimate FD with a few different lag lenghts.

First the lagged variables were created for CO2 emission, and for GDP per capita with one to five lags. Based on the lagged variables, the difference of differenced variables were also defined. Six FD regressions were run with zero to five lagged variables added. Standard error estimates were adjusted to the FD regressions to ensure robustness.

```r
panelData_reg <- panelData %>% group_by(country) %>% mutate(lag_co2 = lag(co2),
    lag_gdppc = lag(gdppc)) %>% mutate(dlco2 = log(co2) - log(lag_co2),
    dlgdppc = log(gdppc) - (log(lag_gdppc)))

panelData_reg_1 <- panelData_reg %>% group_by(country) %>% mutate(lag_gdppc_1 = lag(lag_gdppc)) %>%
    dplyr::filter(!is.na(lag_gdppc_1)) %>% mutate(dlgdppc_1 = log(lag_gdppc) -
    log(lag_gdppc_1))

panelData_reg_2 <- panelData_reg_1 %>% group_by(country) %>%
    mutate(lag_gdppc_2 = lag(lag_gdppc_1)) %>% dplyr::filter(!is.na(lag_gdppc_2)) %>%
    mutate(dlgdppc_2 = log(lag_gdppc_1) - log(lag_gdppc_2))

panelData_reg_3 <- panelData_reg_2 %>% group_by(country) %>%
    mutate(lag_gdppc_3 = lag(lag_gdppc_2)) %>% dplyr::filter(!is.na(lag_gdppc_3)) %>%
    mutate(dlgdppc_3 = log(lag_gdppc_2) - log(lag_gdppc_3))

panelData_reg_4 <- panelData_reg_3 %>% group_by(country) %>%
    mutate(lag_gdppc_4 = lag(lag_gdppc_3)) %>% dplyr::filter(!is.na(lag_gdppc_4)) %>%
    mutate(dlgdppc_4 = log(lag_gdppc_3) - log(lag_gdppc_4))

panelData_reg_5 <- panelData_reg_4 %>% group_by(country) %>%
    mutate(lag_gdppc_5 = lag(lag_gdppc_4)) %>% dplyr::filter(!is.na(lag_gdppc_5)) %>%
    mutate(dlgdppc_5 = log(lag_gdppc_4) - log(lag_gdppc_5))

fd1 <- plm(dlco2 ~ dlgdppc, data = panelData_reg, model = "fd")
fd2 <- plm(dlco2 ~ dlgdppc + dlgdppc_1, data = panelData_reg_1,
    model = "fd")
fd3 <- plm(dlco2 ~ dlgdppc + dlgdppc_1 + dlgdppc_2, data = panelData_reg_2,
    model = "fd")
fd4 <- plm(dlco2 ~ dlgdppc + dlgdppc_1 + dlgdppc_2 + dlgdppc_3,
    data = panelData_reg_3, model = "fd")
fd5 <- plm(dlco2 ~ dlgdppc + dlgdppc_1 + dlgdppc_2 + dlgdppc_3 +
    dlgdppc_4, data = panelData_reg_4, model = "fd")
fd6 <- plm(dlco2 ~ dlgdppc + dlgdppc_1 + dlgdppc_2 + dlgdppc_3 +
    dlgdppc_4 + +dlgdppc_5, data = panelData_reg_5, model = "fd")
```

```r
# Standard errors were adjusted to make sure robust standard
# errors are used
cov1 <- vcovSCC(fd1, type = "HC1")
robust_se_1 <- sqrt(diag(cov1))
cov2 <- vcovSCC(fd2, type = "HC1")
robust_se_2 <- sqrt(diag(cov2))
cov3 <- vcovSCC(fd3, type = "HC1")
robust_se_3 <- sqrt(diag(cov3))
cov4 <- vcovSCC(fd4, type = "HC1")
robust_se_4 <- sqrt(diag(cov4))
cov5 <- vcovSCC(fd5, type = "HC1")
robust_se_5 <- sqrt(diag(cov5))
cov6 <- vcovSCC(fd6, type = "HC1")
robust_se_6 <- sqrt(diag(cov6))
```

```r
stargazer(title = "First Differences with different Lags", list(fd1,
    fd2, fd3, fd4, fd5, fd6), digits = 2, column.labels = c("No Lag",
    "Lag:1", "Lag:2", "Lag:3", "Lag:4", "Lag:5"), model.names = FALSE,
    omit.stat = c("adj.rsq", "f"), dep.var.caption = "Dependent variable: Log CO2",
    out = "Reg_1.html", notes.align = "l", se = list(robust_se_1,
        robust_se_2, robust_se_3, robust_se_4, robust_se_5, robust_se_6),
    add.lines = list(c("Cumulative Coeff", round(sum(fd1$coefficients),
        2), round(sum(fd2$coefficients), 2), round(sum(fd3$coefficients),
        2), round(sum(fd4$coefficients), 2), round(sum(fd5$coefficients),
        2), round(sum(fd6$coefficients), 2))), header = FALSE,
    type = "latex")
```

Table 3: First Differences with different Lags

	Dependent variable: Log CO2					
	dlco2					
	No Lag	Lag:1	Lag:2	Lag:3	Lag:4	Lag:5
	(1)	(2)	(3)	(4)	(5)	(6)
dlgdppc	0.49***	0.43***	0.49***	0.49***	0.51***	0.50***
	(0.10)	(0.12)	(0.13)	(0.14)	(0.14)	(0.11)
dlgdppc_1		0.13	0.01	−0.09	−0.12	−0.16
		(0.08)	(0.10)	(0.11)	(0.13)	(0.15)
dlgdppc_2			0.08	0.10	0.06	0.07
			(0.11)	(0.10)	(0.12)	(0.13)
dlgdppc_3				0.05	−0.06	−0.06
				(0.05)	(0.05)	(0.06)
dlgdppc_4					0.37***	0.31***
					(0.10)	(0.11)
dlgdppc_5						0.05
						(0.07)
Constant	−0.0001	−0.0001	−0.0002	−0.0002	−0.0000	−0.0002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Cumulative Coeff	0.49	0.56	0.58	0.55	0.76	0.7
Observations	3,444	3,280	3,116	2,952	2,788	2,624
R^2	0.04	0.04	0.04	0.04	0.07	0.06
Note:	*p<0.1; **p<0.05; ***p<0.01					

Choose one model, FD(x), where x is number of lags you picked. (x may be 0.) Use that FD(x) model thereafter. (compare models in Table 2)

The largest cumulative coefficient belongs to the FD estimate with 4 lags, which has a cumulative beta of 0.76. Therefore FD with 4 lags was used in the next step.

Estimate OLS for 1995, 2013, 2007, as well as POLS, FE, FD, LD (compare models in Table 3)

Multiple regressions were estimated: OLS for three different years, the pooling model, the fixed effect model, the first difference model and the long difference model. Again, standard error estimates were adjusted to ensure robustness of all regression estimates.

```r
ols1995 <- lm(data = panelData[year == 1995], lnco2 ~ lngdppc)
ols2007 <- lm(data = panelData[year == 2007], lnco2 ~ lngdppc)
ols2013 <- lm(data = panelData[year == 2013], lnco2 ~ lngdppc)
pools <- lm(data = panelData, lnco2 ~ lngdppc)
fe <- plm(data = panelData, lnco2 ~ lngdppc + year, model = "within")
fd_4lag <- plm(data = panelData_reg_4, dlco2 ~ dlgdppc + dlgdppc_1 +
    dlgdppc_2 + dlgdppc_3 + dlgdppc_4)
ld_gdp <- panelData[year == 2013]$lngdppc - panelData[year ==
    1992]$lngdppc
ld_co2 <- panelData[year == 2013]$lnco2 - panelData[year == 1992]$lnco2
ld <- lm(ld_co2 ~ ld_gdp)

# Standard errors were adjusted to make sure robust standard
# errors are used
cov2_1 <- vcovHC(ols1995, type = "HC1")
rob_se_1 <- sqrt(diag(cov2_1))
cov2_2 <- vcovHC(ols2007, type = "HC1")
rob_se_2 <- sqrt(diag(cov2_2))
cov2_3 <- vcovHC(ols2013, type = "HC1")
rob_se_3 <- sqrt(diag(cov2_3))
cov2_4 <- vcovHC(pools, type = "HC1")
rob_se_4 <- sqrt(diag(cov2_4))
cov2_5 <- vcovSCC(fe, type = "HC1")
rob_se_5 <- sqrt(diag(cov2_5))
cov2_6 <- vcovSCC(fd_4lag, type = "HC1")
rob_se_6 <- sqrt(diag(cov2_6))
cov2_7 <- vcovHC(ld, type = "HC1")
rob_se_7 <- sqrt(diag(cov2_7))
```

```r
stargazer(title = "Comparing multiple models", list(ols1995,
    ols2007, ols2013, pools, fe, fd_4lag, ld), digits = 2, column.labels = c("OLS1995",
    "OLS2007", "OLS2013", "Pools", "FE", "FD Lag: 4", "LD"),
    model.names = FALSE, omit.stat = c("adj.rsq", "f", "ser"),
    dep.var.caption = "Dependent variable: Log CO2", out = "Reg_2.html",
    notes.align = "l", se = list(rob_se_1, rob_se_2, rob_se_3,
        rob_se_4, rob_se_5, rob_se_6, rob_se_7), dep.var.labels.include = FALSE,
    header = FALSE, type = "latex")
```

Table 4: Comparing multiple models

	Dependent variable: Log CO2						
	OLS1995	OLS2007	OLS2013	Pools	FE	FD Lag: 4	LD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
lngdppc	1.22***	1.20***	1.15***	1.21***	1.22***		
	(0.05)	(0.05)	(0.05)	(0.01)	(0.04)		
dlgdppc						0.54***	
						(0.17)	
dlgdppc_1						−0.10	
						(0.14)	
dlgdppc_2						0.06	
						(0.12)	
dlgdppc_3						−0.04	
						(0.06)	
dlgdppc_4						0.35***	
						(0.11)	
ld_gdp							0.63***
							(0.08)
Constant	−10.28***	−10.35***	−9.93***	−10.30***			−0.01
	(0.44)	(0.42)	(0.45)	(0.09)			(0.04)
Observations	165	165	165	3,630	3,630	2,805	165
R^2	0.81	0.84	0.83	0.83	0.83	0.07	0.27

Note: *p<0.1; **p<0.05; ***p<0.01

Interpret the slope coefficient for all models. Are they different? Can you interpret results in a causal way? Discuss.

In OLS1995, comparing two countries, the country with a 1% higher ln GDP per capita was expected to have a 1.22% higher CO2 emission, on average. In OLS2007, comparing two countries, the country with a 1% higher ln GDP per capita was expected to have a 1.20% higher CO2 emission, on average. In OLS2013, comparing two countries, the country with a 1% higher ln GDP per capita was expected to have a 1.15% higher CO2 emission, on average. The OLS regressions practically provide the same results.

In POOLS, comparing any two observations in the period of 1992 and 2016, the observation with a 1% higher ln GDP per capita was expected to have a 1.21% higher CO2 emission, on average. This 1.21% can also be interpreted as the weighted average of all slope coefficients of each year (OLS), where weights are the number of observations is each year. This coefficient is also practically the same as the OLS coefficients.

In FE, comparing two countries that have different levels of GDP per capita relative to its mean in country i, but are the same in everything else that does not change in time, CO2 emission is expected to be higher by 1.22%, on average, relative to its mean value in country i, where or when GDP per capita is higher by 1% than its long-term average in counry i.

In FD with 4 lags the coefficient of the contemporaneous right-hand-side variable is 0.56. In years when GDP per capita increases by 1% more, CO2 emission increases by 0.56% more, on average. The average decrease in CO2 emission is 0.14% the year after. The average increase two years later is 0.06%, -0.04% after three years, and it is 0.36% after four years. The cumulative coefficient is 0.80, which means that comparing countries, or years, with a 1% difference in GDP per capita, CO2 emission increases by 0.80% more on average after four years in where, or when, GDP per capita increases by 1% more.

In LD, comparing two countries with different changes in GDP per capita across the long time horizon, CO2 emission is expected to increase by 0.63% more where or when GDP per capita increases by 1% more, on average.

Estimates cannot be interpreted as the measure of the causal effect unless it is assumed that reverse causality is not an issue and the fixed effects take care of all kinds of selection and other confounders. Cross-sectional FE do take care of selection on fixed country characteristics (countries with more income select themselves to use more CO2 or trade CO2 quotas). Time FE would take care of positive trends in both GDP per capita and CO2 emission as long as these trends are the same across countries, but might miss potential time varying selection and country-specific trends. While some observable measures of time-varying selection can be added to the regression, we can never be sure that we included all that matters. Therefore by including FE and time-varying control variables we can get closer to causal interpretation but can never be sure whether we get there.

Which result would you have as your best estimate? Why? Interpret the point estimate and its 95% confidence interval.

The FD estimate captures how CO2 emission changes after GDP per capita is changed from one year to another, and the impact of this change is tracked through the subsequent 4 years. The FE estimate shows the extent to which CO2 emission is higher in years when GDP per capita is higher within countries. Both FD and FE control for aggregate trends in the given time period. The LD estimate shows the difference in the change of CO2 emission from 1992 to 2013 for countries that experienced higher GDP per capita in this period, and it also controls for aggregate trends.

The FD and LD coefficient estimates are not very different in mangitude, however, the FE coefficient is closer to the OLS and Pooled coefficients, being the double FD and LD estimates. The best estimate is the one for which the common trends assumption is the most likely to hold. + Since LD looks at changes over a long time horizon, the common trends assumption is less likely to hold. Countries with very different changes in GDP may have had different experiences in other relevant dimensions as well. + FD looks at changes happening in certain years (same year or a subsequent year), but might miss the effect of the change in other years, for example if increase in GDP in 1991 led to increase in emission in any subsequent year in the given period. + FE looks at levels not changes relative to the long-term mean in a country. This has the advantage that fixed cross-country differences don't confound the estimates, and that also long-term effects are captured. Looking at levels also has the disadvantage that it needs stationary series for the time series regression run in each country.

Based on these arguments I would choose FE as the best estimate, because it is the one for which the common trends assumption is the most likely to hold (fixed cross-country differences don't confound the estimates, and that also long-term effects are captured).

The point estimate is 1.22, which means that comparing two countries that have different levels of GDP per capita relative to its mean in country i, but are the same in everything else that does not change in time, CO2 emission is expected to be higher by 1.22%, on average, relative to its mean value in country i, where or when GDP per capita is higher by 1% than its long-term average in councry i. The 95% confidence interval around the coefficient is rather narrow, [1.14, 1.30]. We can be 95% confident that in the general pattern represented by our data CO2 emission is 1.14% to 1.30% higher than the country average in years when GDP per capita is 1% higher than the country average.

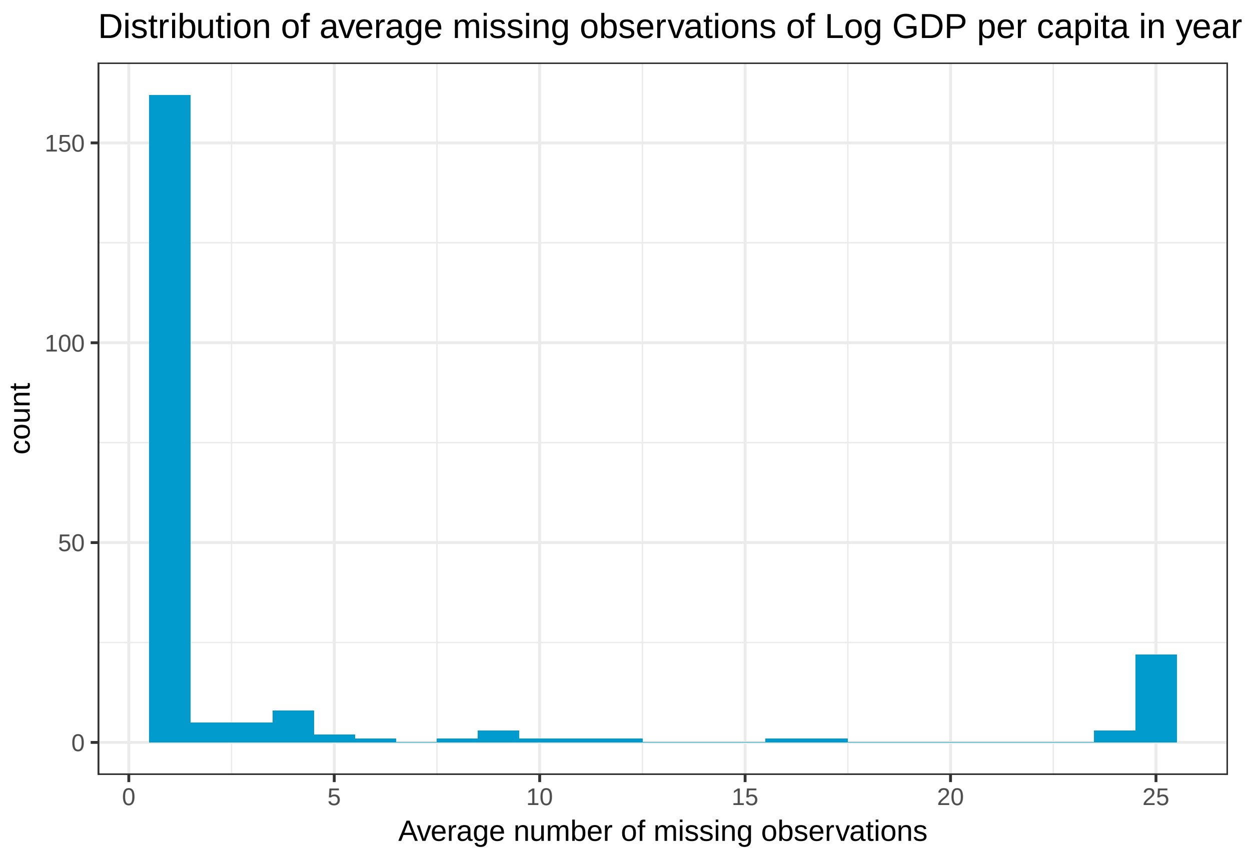
Now consider the full, unbalanced panel. Discuss missing observations.

For the unbalanced panel I created a new dataset which includes all observations, including all yearly country data with missing values as well.

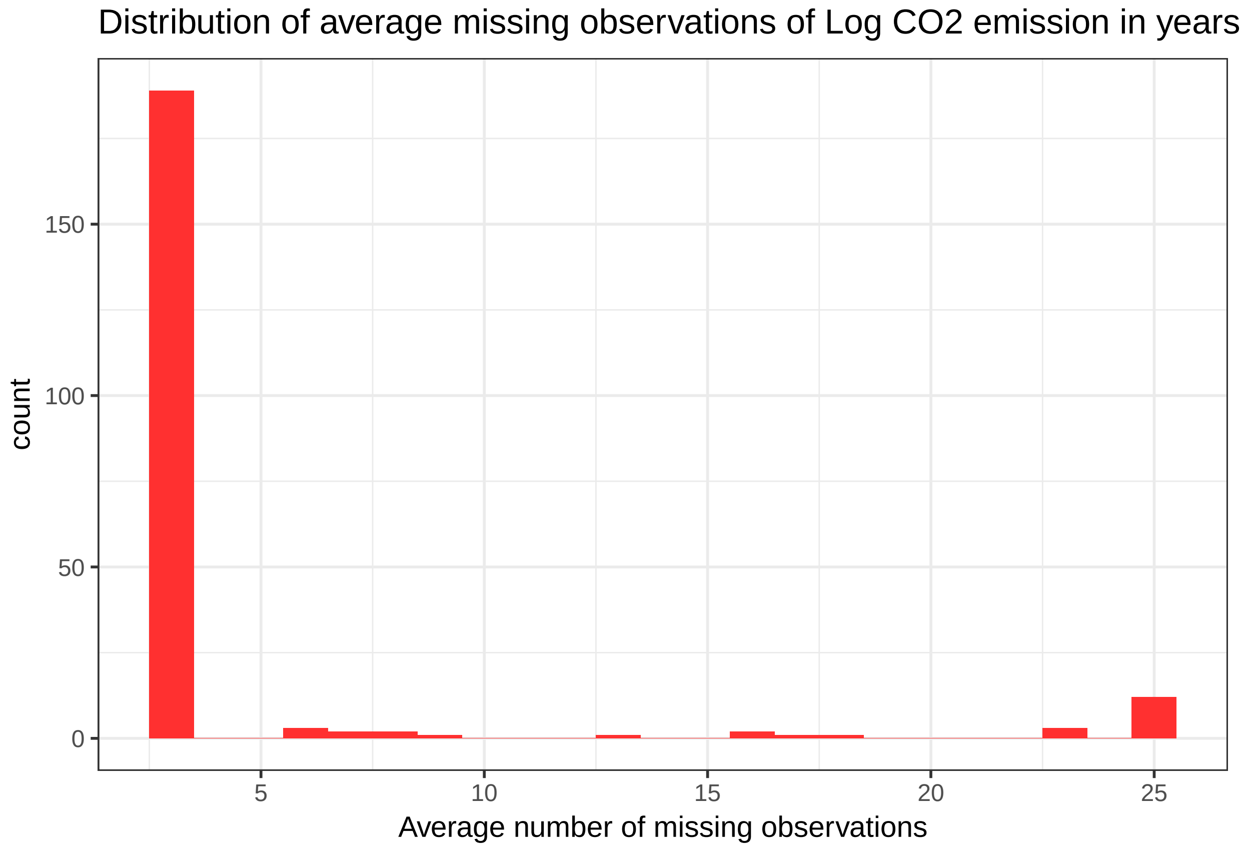
Distributions of observations with missing values for the Log GDP per capita and Log CO2 emission variables were visualized in histograms below.

By plotting the number of missing observations and the average country data in the given time period we can see that for both variables countries with more missing values tend to have higher GDP per capita and CO2 emission, however, there are relatively few observations with a high number of missing values

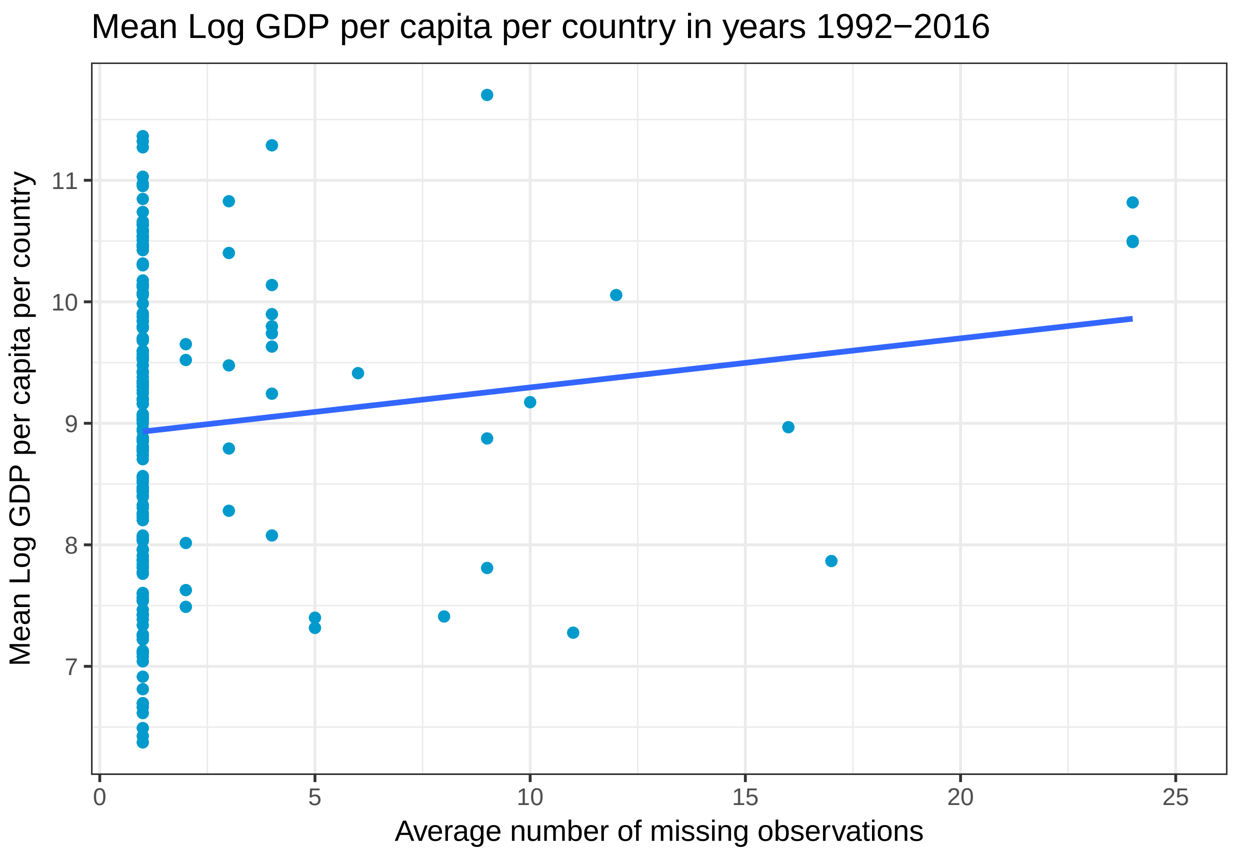
```r
unbData <- merge(gdpData, co2Data, by.x = c("iso2c", "year",
    "country"), by.y = c("iso2c", "year", "country"))

unbData$iso2c <- NULL
names(unbData) <- c("year", "country", "gdppc", "co2")

unbData$lngdppc <- log(unbData$gdppc)
unbData$lnco2 <- log(unbData$co2)

miss_lngdppc <- merge(unbData[, mean(lngdppc, na.rm = TRUE),
    by = country], unbData[is.na(lngdppc), .N, by = country],
    by = "country", all = TRUE)

miss_lnco2 <- merge(unbData[, mean(lnco2, na.rm = TRUE), by = country],
    unbData[is.na(lnco2), .N, by = country], by = "country",
    all = TRUE)

ggplot(miss_lngdppc) + aes(x = N) + geom_histogram(binwidth = 1,
    fill = "deepskyblue3") + labs(x = "Average number of missing observations",
    title = "Distribution of average missing observations of Log GDP per capita in years 1992-2016") +
    theme_bw()
```



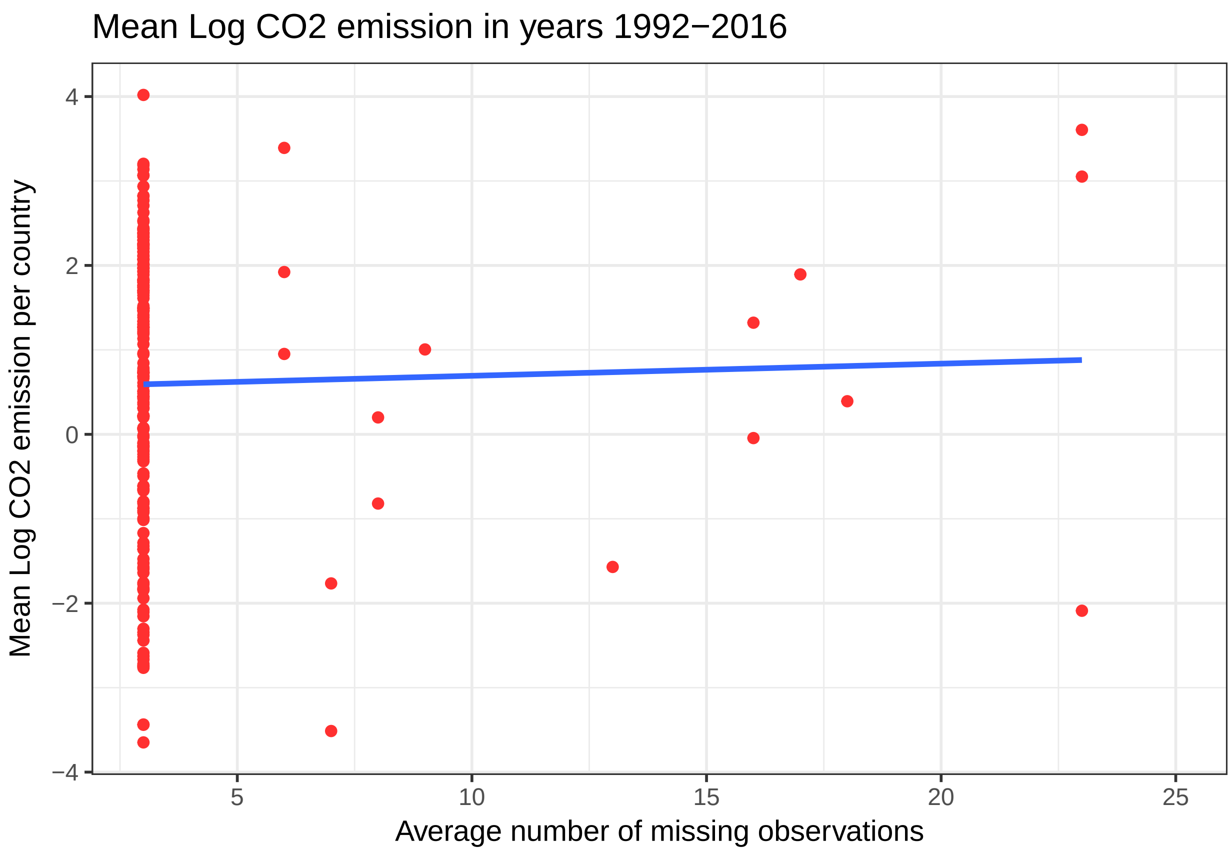
```r
ggplot(miss_lnco2) + aes(x = N) + geom_histogram(binwidth = 1,
    fill = "firebrick1") + labs(x = "Average number of missing observations",
    title = "Distribution of average missing observations of Log CO2 emission in years 1992-2016") +
    theme_bw()
```



```
ggplot(miss_lngdppc, aes(y = V1, x = N)) + geom_point(color = "deepskyblue3") +
    geom_smooth(method = "lm", se = FALSE) + labs(x = "Average number of missing observations",
    y = "Mean Log GDP per capita per country", title = "Mean Log GDP per capita per country in years 19
    theme_bw()
```



```
ggplot(miss_lnco2, aes(y = V1, x = N)) + geom_point(color = "firebrick1") +
    geom_smooth(method = "lm", se = FALSE) + labs(x = "Average number of missing observations",
    y = "Mean Log CO2 emission per country", title = "Mean Log CO2 emission in years 1992-2016") +
    theme_bw()
```



Estimate OLS for 2007, FE and FD(x), LD (Table 3). Compare results to balanced panel. Discuss your finding.

The results show that the estimates of the balanced and unbalanced panel are very similar.
+ For OLS2007 the slope coefficient and its standard error are the same and the intercept is a slightly smaller negative number for the unbalanced panel. + For FE the slope coefficient and its standard error are the same. + For FD with 4 lags, both the coefficient of the first difference and the of lag 4 are slightly smaller for the unbalanced panel. + For LD the slope coefficient and its standard error are the same. These results indicate that the countries with missing observations are not so much different on average from the rest of the countries, since the estimates are not significantly different when including these countries in the models.

```
unbData_reg <- unbData %>% group_by(country) %>% mutate(lag_co2 = lag(co2),
    lag_gdppc = lag(gdppc)) %>% mutate(dlco2 = log(co2) - log(lag_co2),
    dlgdppc = log(gdppc) - (log(lag_gdppc)))

unbData_reg_1 <- unbData_reg %>% group_by(country) %>% mutate(lag_gdppc_1 = lag(lag_gdppc)) %>%
    dplyr::filter(!is.na(lag_gdppc_1)) %>% mutate(dlgdppc_1 = log(lag_gdppc) -
    log(lag_gdppc_1))

unbData_reg_2 <- unbData_reg_1 %>% group_by(country) %>% mutate(lag_gdppc_2 = lag(lag_gdppc_1)) %>%
    dplyr::filter(!is.na(lag_gdppc_2)) %>% mutate(dlgdppc_2 = log(lag_gdppc_1) -
    log(lag_gdppc_2))

unbData_reg_3 <- unbData_reg_2 %>% group_by(country) %>% mutate(lag_gdppc_3 = lag(lag_gdppc_2)) %>%
    dplyr::filter(!is.na(lag_gdppc_3)) %>% mutate(dlgdppc_3 = log(lag_gdppc_2) -
    log(lag_gdppc_3))

unbData_reg_4 <- unbData_reg_3 %>% group_by(country) %>% mutate(lag_gdppc_4 = lag(lag_gdppc_3)) %>%
    dplyr::filter(!is.na(lag_gdppc_4)) %>% mutate(dlgdppc_4 = log(lag_gdppc_3) -
    log(lag_gdppc_4))

u_ols2007 <- lm(data = unbData[year == 2007], lnco2 ~ lngdppc)
u_fe <- plm(data = unbData, lnco2 ~ lngdppc + year, model = "within")
u_fd_4lag <- plm(data = unbData_reg_4, dlco2 ~ dlgdppc + dlgdppc_1 +
    dlgdppc_2 + dlgdppc_3 + dlgdppc_4)
u_ld_gdp <- unbData[year == 2013]$lngdppc - unbData[year == 1992]$lngdppc
u_ld_co2 <- unbData[year == 2013]$lnco2 - unbData[year == 1992]$lnco2
u_ld <- lm(u_ld_co2 ~ u_ld_gdp)

# Standard errors were adjusted to make sure robust standard
# errors are used
u_cov2_2 <- vcovHC(u_ols2007, type = "HC1")
u_rob_se_2 <- sqrt(diag(u_cov2_2))
u_cov2_5 <- vcovSCC(u_fe, type = "HC1")
u_rob_se_5 <- sqrt(diag(u_cov2_5))
u_cov2_6 <- vcovSCC(u_fd_4lag, type = "HC1")
u_rob_se_6 <- sqrt(diag(u_cov2_6))
u_cov2_7 <- vcovHC(u_ld, type = "HC1")
u_rob_se_7 <- sqrt(diag(u_cov2_7))
```

```
stargazer(title = "Comparing multiple models", list(u_ols2007,
    u_fe, u_fd_4lag, u_ld), digits = 2, column.labels = c("OLS2007",
    "FE", "FD Lag: 4", "LD"), model.names = FALSE, omit.stat = c("adj.rsq",
    "f", "ser"), dep.var.caption = "Dependent variable: Log CO2",
    out = "Reg_3.html", notes.align = "l", se = list(rob_se_2,
        rob_se_5, rob_se_6, rob_se_7), dep.var.labels.include = FALSE,
    header = FALSE, type = "latex")
```

Think about a potential confounder and add a variable. Not: GDP, population. Estimate OLS for 2007, FE and FD(x), LD (Table 4). Discuss your finding.

I added the variable "Fossil fuel energy consumption (% of total)" as a confounder as a potential confounder. It impacts GDP per capita in various ways, since the price of fossil fuel influences basically all industries. It affects CO2 emission, since fossil fuel energy consumption is one of the main reasons of CO2 emission in a country. The variable also potentially effects the mechanism between the changes of GDP per capita and CO2 emission levels, because the increase in GDP per capita does not necessarily have to lead to increase in CO2 emission, in the ratio of fossil fuels does not remain at least the same in a country. If fossil fuels are substituted with renewable energy sources, the causal mechanism might become different.

```
# SEARCHING FOR DATA: Fossil fuel energy consumption
fossil <- WDIsearch("fossil")
fossilcode <- fossil[match("Fossil fuel energy consumption (% of total)",
    fossil[, 2]), 1]

# DATA DOWNLOAD: fossil
fossil_dat = WDI(indicator = fossilcode, start = 1992, end = 2016)

# FILTERING OUT REGIONS
```

Table 5: Comparing multiple models

	Dependent variable: Log CO2			
	OLS2007	FE	FD Lag: 4	LD
	(1)	(2)	(3)	(4)
lngdppc	1.20^{***}	1.22^{***}		
	(0.05)	(0.04)		
dlgdppc			0.53^{***}	
			(0.17)	
dlgdppc_1			−0.10	
			(0.14)	
dlgdppc_2			0.09	
			(0.12)	
dlgdppc_3			−0.03	
			(0.06)	
dlgdppc_4			0.30^{***}	
			(0.11)	
u_ld_gdp				0.63
Constant	-10.31^{***}			−0.01
	(0.42)			(0.04)
Observations	188	3,989	3,070	165
R^2	0.85	0.84	0.08	0.27
Note:	*p<0.1; **p<0.05; ***p<0.01			

```
fossil_dt <- data.table(fossil_dat)
exclusionList <- fossil_dt[, .(itemCnt = .N), by = .(code = fossil_dt$iso2c)][1:47,
    1]
fossil_Data <- subset(fossil_dt, !(fossil_dt$iso2c %in% exclusionList$code))

# MERGING DATA INTO ORIGINAL PANEL
unbData_new <- merge(unbData, fossil_Data, by.x = c("year", "country"),
    by.y = c("year", "country"))

unbData_new$iso2c <- NULL
names(unbData_new) <- c("year", "country", "gdppc", "co2", "lngdppc",
    "lnco2", "fossil")
```

I created only first lags for CO2 emission and Fossil fuel consumption, and four lags for GDP per capital. The OLS2007, FE, FD with 4 lags and LD models were run for the period of 1992 to 2016.

```
unbData_new$lnfos <- log(unbData_new$fossil + 1)

unbData_fos <- unbData_new %>% group_by(country) %>% mutate(lag_co2 = lag(co2),
    lag_gdppc = lag(gdppc), lag_fos = lag(fossil)) %>% mutate(dlco2 = log(co2) -
    log(lag_co2), dlgdppc = log(gdppc) - (log(lag_gdppc)), dlfos = log(fossil) -
    (log(lag_fos)))

unbData_fos_1 <- unbData_fos %>% group_by(country) %>% mutate(lag_gdppc_1 = lag(lag_gdppc)) %>%
    dplyr::filter(!is.na(lag_gdppc_1)) %>% mutate(dlgdppc_1 = log(lag_gdppc) -
    log(lag_gdppc_1))

unbData_fos_2 <- unbData_fos_1 %>% group_by(country) %>% mutate(lag_gdppc_2 = lag(lag_gdppc_1)) %>%
    dplyr::filter(!is.na(lag_gdppc_2)) %>% mutate(dlgdppc_2 = log(lag_gdppc_1) -
    log(lag_gdppc_2))

unbData_fos_3 <- unbData_fos_2 %>% group_by(country) %>% mutate(lag_gdppc_3 = lag(lag_gdppc_2)) %>%
    dplyr::filter(!is.na(lag_gdppc_3)) %>% mutate(dlgdppc_3 = log(lag_gdppc_2) -
    log(lag_gdppc_3))

unbData_fos_4 <- unbData_fos_3 %>% group_by(country) %>% mutate(lag_gdppc_4 = lag(lag_gdppc_3)) %>%
    dplyr::filter(!is.na(lag_gdppc_4)) %>% mutate(dlgdppc_4 = log(lag_gdppc_3) -
    log(lag_gdppc_4))

unbData_fos$lnco2[which(is.nan(unbData_fos$lnco2))] = NA
unbData_fos$lnco2[which(unbData_fos$lnco2 == Inf)] = NA
unbData_fos$lngdppc[which(is.nan(unbData_fos$lngdppc))] = NA
unbData_fos$lngdppc[which(unbData_fos$lngdppc == Inf)] = NA
unbData_fos$lnfos[which(is.nan(unbData_fos$lnfos))] = NA
unbData_fos$lnfos[which(unbData_fos$lnfos == Inf)] = NA

data2007 <- subset(unbData_fos, year == 2007)
data2013 <- subset(unbData_fos, year == 2013)
data1992 <- subset(unbData_fos, year == 1992)
fos_ols2007 <- lm(data = data2007, lnco2 ~ lngdppc + lnfos)
fos_fe <- plm(data = unbData_fos, lnco2 ~ lngdppc + lnfos, model = "within")
fos_fd_4lag <- plm(data = unbData_fos_4, dlco2 ~ dlgdppc + dlgdppc_1 +
    dlgdppc_2 + dlgdppc_3 + dlgdppc_4 + dlfos)
fos_ld_gdp <- data2013$lngdppc - data1992$lngdppc
```

```r
fos_ld_co2 <- data2013$lnco2 - data1992$lnco2
fos_ld_fos <- data2013$lnfos - data1992$lnfos
fos_ld <- lm(fos_ld_co2 ~ fos_ld_gdp + fos_ld_fos)

# Standard errors were adjusted to make sure robust standard
# errors are used
fos_cov2_2 <- vcovHC(fos_ols2007, type = "HC1")
fos_rob_se_2 <- sqrt(diag(fos_cov2_2))
fos_cov2_5 <- vcovSCC(fos_fe, type = "HC1")
fos_rob_se_5 <- sqrt(diag(fos_cov2_5))
fos_cov2_6 <- vcovSCC(fos_fd_4lag, type = "HC1")
fos_rob_se_6 <- sqrt(diag(fos_cov2_6))
fos_cov2_7 <- vcovHC(fos_ld, type = "HC1")
fos_rob_se_7 <- sqrt(diag(fos_cov2_7))
```

```r
stargazer(title = "Comparing multiple models", list(fos_ols2007,
    fos_fe, fos_fd_4lag, fos_ld), digits = 2, column.labels = c("OLS2007",
    "FE", "FD Lag: 4", "LD"), model.names = FALSE, omit.stat = c("adj.rsq",
    "f", "ser"), dep.var.caption = "Dependent variable: Log CO2",
    out = "Reg_4.html", notes.align = "l", se = list(fos_rob_se_2,
        fos_rob_se_5, fos_rob_se_6, fos_rob_se_7), dep.var.labels.include = FALSE,
    header = FALSE, type = "latex")
```

The results show that all coefficient estimates became smaller in magnitude by controlling for fossil fuel energy consumption. This means that the control variable explains some of the effect of GDP per capita changes on CO2 emission changes, which mechanism was previously unobserved in the regressions. The slope coefficient of fossil fuel energy consumption is statistically significant in all regressions at the 1% level - just like the slope coefficient of the CO2 emission variable, except for OLS2007, where the slope coefficient of fossil fuel energy consumption is statistically significant at the 5% level. These results validate that fossil fuel consumption affects both GDP per capita and CO2 emission, therefore is a relevant confounder, which can help us get closer to the causal effect.

Heterogeneity of relationship. Create country groups by your definition by GPD per capita in year 2007. Argue for your choice. For the balanced data countries, estimate OLS 2007, FE and FD(x), LD, for both subsamples (Table 5). Discuss findings.

By visualizing the distribution of GDP per capita in 2007 it becomes visible that a significant part of countries is between 0 and 15,000 USD, constant 2005 prices. The group with the most number of observations with a binwidth of 2,000 is the one with a GDP per capita between 1,000 and 3,000 USD, therefore I will break up the groups for countries with a GDP per capita above and below 3,000 USD. I expect this grouping to provide information on how much low-income countries are different from the rest of the dataset in terms of the impact of GDP per capita changes on CO2 emission change. My presumption is that in low-income countries theeffect of change in GDP per capita is bigger impact on the change of CO2 emission, than in countries with higher income.

```r
panel2007 <- subset(unbData_fos, year == 2007)
dt_2007 <- data.frame(panel2007)
```

```r
stargazer(dt_2007, out = "summary_2007.html", header = FALSE,
    type = "latex", omit = "year", title = "Descriptive Statistics for panel 2007")
```

```r
ggplot(panel2007) + aes(x = gdppc) + geom_histogram(binwidth = 2000,
    fill = "deepskyblue3") + labs(x = "GDP per capita distribution in 2007, constant 2005 prices",
    title = "Histogram of GDP per capita distribution in 2007") +
```

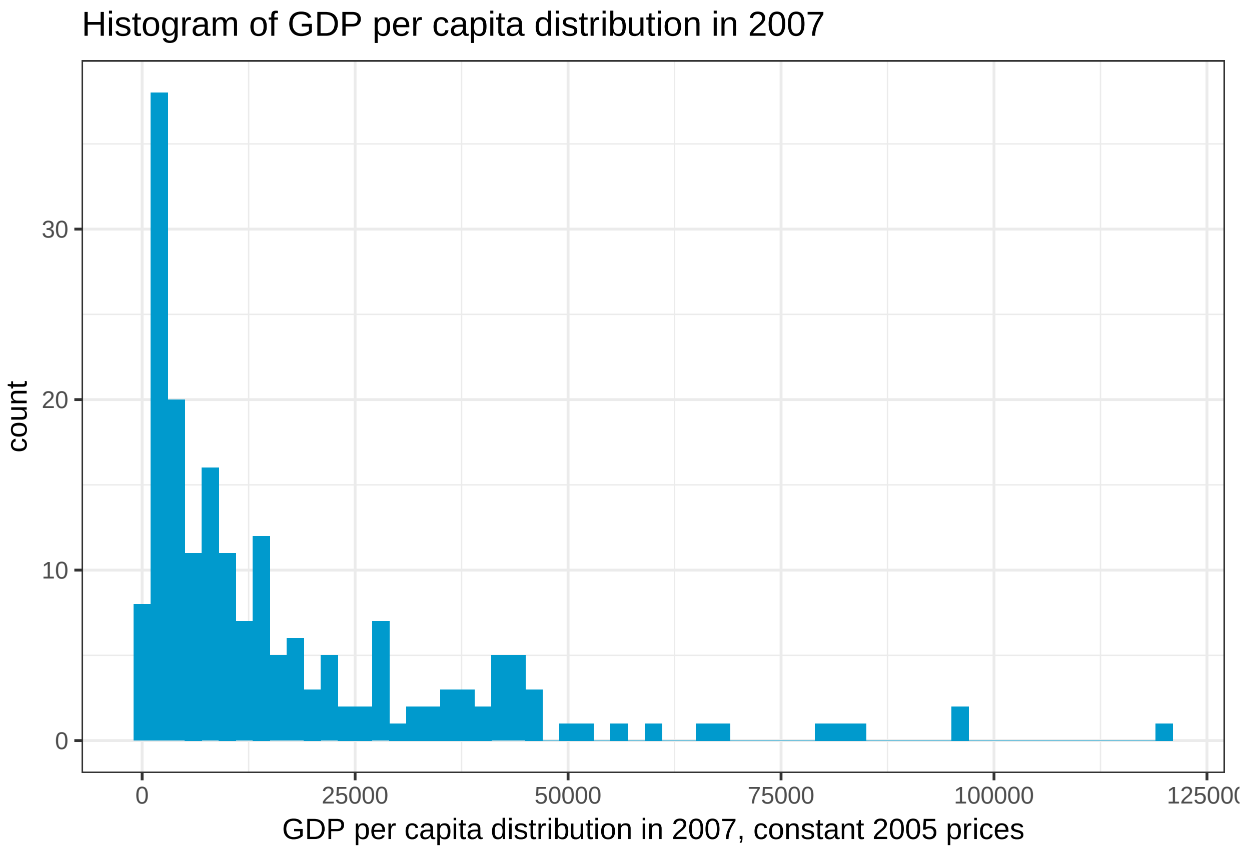
Table 6: Comparing multiple models

	Dependent variable: Log CO2			
	OLS2007	FE	FD Lag: 4	LD
	(1)	(2)	(3)	(4)
lngdppc	1.12^{***}	1.06^{***}		
	(0.04)	(0.05)		
lnfos	0.06^{**}	0.26^{***}		
	(0.02)	(0.05)		
dlgdppc			0.42^{***}	
			(0.05)	
dlgdppc_1			0.03	
			(0.07)	
dlgdppc_2			−0.07	
			(0.08)	
dlgdppc_3			−0.03	
			(0.05)	
dlgdppc_4			0.20^{**}	
			(0.09)	
dlfos			0.76^{***}	
			(0.06)	
fos_ld_gdp				0.48^{***}
				(0.10)
fos_ld_fos				0.81^{***}
				(0.13)
Constant	-9.71^{***}			-0.12^{***}
	(0.38)			(0.04)
Observations	165	2,945	2,170	115
R^2	0.86	0.85	0.26	0.56
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$			

Table 7: Descriptive Statistics for panel 2007

Statistic	N	Mean	St. Dev.	Min	Max
gdppc	191	17,303.640	20,425.380	561.704	119,907.900
co2	202	5.079	6.857	0.022	53.673
lngdppc	191	9.073	1.262	6.331	11.694
lnco2	202	0.668	1.658	−3.818	3.983
fossil	170	56.990	36.818	0.000	100.000
lnfos	170	3.367	1.697	0.000	4.615
lag_co2	201	5.047	7.101	0.023	64.218
lag_gdppc	190	16,799.310	20,212.290	546.027	121,216.600
lag_fos	168	56.185	36.869	0.000	100.000
dlco2	201	0.019	0.131	−0.551	0.706
dlgdppc	190	0.042	0.041	−0.119	0.212
dlfos	137	0.010	0.070	−0.322	0.528

```
    theme_bw()
```



```
# Creating balanced dataset of countries
dt_2007 <- unbData_fos[!is.na(unbData_fos$gdppc) & !is.na(unbData_fos$co2) &
    !is.na(unbData_fos$fossil), ]
dt_2007_4 <- unbData_fos_4[!is.na(unbData_fos$gdppc) & !is.na(unbData_fos$co2) &
    !is.na(unbData_fos$fossil), ]
```

```
# Group 1
dt_2007$group <- ifelse(dt_2007$gdppc <= 3000, 1, 2)
dt_2007_4$group <- ifelse(dt_2007_4$gdppc <= 3000, 1, 2)
dt_2007_1_2007 <- data.frame(subset(dt_2007, group == 2 & year ==
    2007))
dt_2007_1_2016 <- data.frame(subset(dt_2007, group == 2 & year ==
    2016))
dt_2007_1_1992 <- data.frame(subset(dt_2007, group == 2 & year ==
    1992))
dt_2007_1 <- data.frame(subset(dt_2007_4, group == 2))

p_ols2007 <- lm(data = dt_2007_1_2007, lnco2 ~ lngdppc + lnfos)
p_fe <- plm(data = dt_2007_1, lnco2 ~ lngdppc + lnfos, model = "within")
```

```
## This series is constant and has been removed: group
```

```
p_fd_4lag <- plm(data = dt_2007_1, dlco2 ~ dlgdppc + dlgdppc_1 +
    dlgdppc_2 + dlgdppc_3 + dlgdppc_4 + dlfos)
```

```
## This series is constant and has been removed: group
```

```
# Standard errors were adjusted to make sure robust standard
# errors are used
p_cov2_2 <- vcovHC(p_ols2007, type = "HC1")
p_rob_se_2 <- sqrt(diag(p_cov2_2))
p_cov2_5 <- vcovSCC(p_fe, type = "HC1")
p_rob_se_5 <- sqrt(diag(p_cov2_5))
p_cov2_6 <- vcovSCC(p_fd_4lag, type = "HC1")
p_rob_se_6 <- sqrt(diag(p_cov2_6))

# Group 2
dt_2007_0_2007 <- data.frame(subset(dt_2007, group == 1 & year ==
    2007))
dt_2007_0_2016 <- data.frame(subset(dt_2007, group == 1 & year ==
    2016))
dt_2007_0_1992 <- data.frame(subset(dt_2007, group == 1 & year ==
    1992))
dt_2007_0 <- data.frame(subset(dt_2007_4, group == 1))

r_ols2007 <- lm(data = dt_2007_0_2007, lnco2 ~ lngdppc + lnfos)
r_fe <- plm(data = dt_2007_0, lnco2 ~ lngdppc + lnfos, model = "within")
```

```
## This series is constant and has been removed: group
```

```
r_fd_4lag <- plm(data = dt_2007_0, dlco2 ~ dlgdppc + dlgdppc_1 +
    dlgdppc_2 + dlgdppc_3 + dlgdppc_4 + dlfos)
```

```
## This series is constant and has been removed: group
```

```
# Standard errors were adjusted to make sure robust standard
# errors are used
r_cov2_2 <- vcovHC(r_ols2007, type = "HC1")
r_rob_se_2 <- sqrt(diag(r_cov2_2))
r_cov2_5 <- vcovSCC(r_fe, type = "HC1")
r_rob_se_5 <- sqrt(diag(r_cov2_5))
r_cov2_6 <- vcovSCC(r_fd_4lag, type = "HC1")
```

```
r_rob_se_6 <- sqrt(diag(r_cov2_6))
```

```
stargazer(title = "Comparing multiple models", list(p_ols2007,
    r_ols2007, p_fe, r_fe, p_fd_4lag, r_fd_4lag), digits = 2,
    column.labels = c("OLS2007:2", "OLS2007:1", "FE:2", "FE:1",
        "FD Lag4:2", "FD Lag4:1"), model.names = FALSE, omit.stat = c("adj.rsq",
        "f", "ser"), dep.var.caption = "Dependent variable: Log CO2",
    out = "Reg_5.html", notes.align = "l", se = list(p_rob_se_2,
        r_rob_se_2, p_rob_se_5, r_rob_se_5, p_rob_se_6, r_rob_se_6),
    dep.var.labels.include = FALSE, header = FALSE, type = "latex")
```

Table 8: Comparing multiple models

	Dependent variable: Log CO2					
	OLS2007:2	OLS2007:1	FE:2	FE:1	FD Lag4:2	FD Lag4:1
	(1)	(2)	(3)	(4)	(5)	(6)
lngdppc	0.98***	1.45***	0.97***	1.27***		
	(0.06)	(0.23)	(0.05)	(0.18)		
lnfos	0.07***	−0.03	0.17***	0.23***		
	(0.02)	(0.06)	(0.04)	(0.08)		
dlgdppc					0.42***	0.55***
					(0.05)	(0.20)
dlgdppc_1					0.14	−0.18
					(0.09)	(0.19)
dlgdppc_2					0.05	0.19
					(0.13)	(0.13)
dlgdppc_3					−0.15	−0.08
					(0.11)	(0.11)
dlgdppc_4					0.27***	0.15
					(0.10)	(0.12)
dlfos					0.89***	0.67***
					(0.14)	(0.09)
Constant	−8.39***	−12.15***				
	(0.58)	(1.66)				
Observations	133	32	1,150	260	1,089	238
R^2	0.75	0.53	0.73	0.57	0.23	0.45
Note:	*p<0.1; **p<0.05; ***p<0.01					

I filtered for countries with non-missing values for all variables. The original dataset had 5,425 observations and the new dataset had 2,945 observations. By creating group 2 (GDP per capita larger or equal to 3,000 USD, contant 2005 prices) and group 1, the difference in the mechanism between GDP per capita changes and CO2 emission changes for low- and high-income countries became visible. By comparing the slope coefficient estimates we can see that for all regression models the coefficient estimate is larger for group 1, the low-income countries. This means that in countries, where GDP per capita was below 3,000 USD in 2007 at constant 2005 prices, the change in GDP per capita had a larger impact on the change in CO2 emission, on average.

In the OLS2007 model, there is a relatively large difference between the slope coefficients, and for high-income countries the slope coefficient on fossil fuel consumption is statistically significant at the 1% level.

In the FE model, there is also a relatively large difference between the slop coefficients, with all coefficients of Log GDP per capita and Log Fossil fuel consumption being statistically significant at the 1% level.

In the FD model with 4 lags, the contemporous coefficient is smaller for high-income countries, while the slope coefficient of lag 4 is only statistically significant at any given level for high-income countries.

This means overall that the change in GDP per capita affects low-income countries with a larger magnitude, on average, than high-income countries. This conclusion confirms my initial presumption, based on data from 2007.

In one paragraph, summarize all your findings for a policymaker.

Overall, we can say that GDP per capita has a very strong impact on CO2 emission. The slope coefficients in both the balanced and unbalanced panels in fixed effects model, the first difference model and the long difference model were statistically significant at the 1% level. By adding a control variable (fossil fuel consumption) the estimates did not become ver different, and as a result we could not explain the relationship fully in this analysis even by adding a confounding variable. By examining unobserved heterogenity in the low-income and high-income groups of countries we could conclude that there is a visible difference in the how GDP per capita change effects CO2 emission between the two groups. As a conclusion we can say that very likely there is a causal relationship between the two variables.

Appendix

```r
knitr::opts_chunk$set(echo = TRUE)

rm(list = ls())
library(WDI)
library(data.table)
library(stringr)
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(gridExtra)
library(plm)
library(stargazer)
library(fBasics)
library(pander)
library(knitr)

opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)

# SEARCHING FOR DATA: GDP per capita
gdp_inds <- WDIsearch("gdp")
grep("2005", gdp_inds, value = TRUE)
gdpppCode <- gdp_inds[match("GDP per capita, PPP (constant 2005 international $)",
    gdp_inds[, 2]), 1]

# DATA DOWNLOAD: GDP per capita
dat = WDI(indicator = gdpppCode, start = 1992, end = 2016)

# FILTERING OUT REGIONS
dt <- data.table(dat)
exclusionList <- dt[, .(itemCnt = .N), by = .(code = dt$iso2c)][1:47,
    1]
gdpData <- subset(dt, !(dt$iso2c %in% exclusionList$code))

# SEARCHING FOR DATA: CO2
co2_inds <- WDIsearch("co2")
# CO2 emissions (metric tons per capita) CO2 emissions (kt)
co2code <- co2_inds[match("CO2 emissions (metric tons per capita)",
    co2_inds[, 2]), 1]

# DATA DOWNLOAD: CO2
co2dat = WDI(indicator = co2code, start = 1992, end = 2016)

# FILTERING OUT REGIONS
co2dt <- data.table(co2dat)
co2Data <- subset(co2dt, !(co2dt$iso2c %in% exclusionList$code))

# MERGING DATA INTO ORIGINAL PANEL
panelData <- merge(gdpData, co2Data, by.x = c("iso2c", "year",
    "country"), by.y = c("iso2c", "year", "country"))

dt <- NULL
dat <- NULL
co2dat <- NULL
co2dt <- NULL
exclusionList <- NULL
gdp_inds <- NULL
co2_inds <- NULL
panelData$iso2c <- NULL
names(panelData) <- c("year", "country", "gdppc", "co2")

# KEEPING ONLY COUNTRIES WITH AT LEAST 22 YEARLY OBSERVATIONS
dt <- panelData[!is.na(panelData$gdppc) & !is.na(panelData$co2),
    .(count = .N), by = country]
dta <- dt[count > 21]
panelData <- subset(panelData, (panelData$country %in% dta$country))
dt <- NULL
dta <- NULL

library(pander)
library(ggthemes)

panelData %>% group_by(year) %>% summarize(gdppc = mean(gdppc),
    co2 = mean(co2)) %>% gather(variable, value, -year) %>% ggplot(aes(x = year,
    y = value)) + geom_line(aes(color = variable), size = 1) +
    facet_grid(variable ~ ., scales = "free", labeller = label_both) +
    ggtitle("Global trends of GDP per capita and CO2 emission")

ggplot(panelData) + aes(x = gdppc) + geom_histogram(binwidth = 10000,
    fill = "deepskyblue3") + labs(x = "GDP per capita distribution across all years, constant 2005 pric
    title = "Histogram of GDP per capita distribution in years 1992-2016") +
    theme_bw()

ggplot(panelData) + aes(x = co2) + geom_histogram(binwidth = 5,
    fill = "firebrick1") + labs(x = "CO2 emission distribution across all years",
    title = "CO2 emission distribution in years 1992-2016") +
    theme_bw()

dt <- data.frame(panelData)
stargazer(dt, out = "summary.html", header = FALSE, type = "latex",
    omit = "year", title = "Descriptive Statistics for the original variables")

panelData$lngdppc <- log(panelData$gdppc)
panelData$lnco2 <- log(panelData$co2)

ggplot(panelData) + aes(x = lngdppc) + geom_histogram(binwidth = 0.5,
    fill = "deepskyblue3") + labs(x = "Log GDP per capita distribution across all years, constant 2005 p
    title = "Histogram of log GDP per capita distribution in years 1992-2016") +
    theme_bw()

ggplot(panelData) + aes(x = lnco2) + geom_histogram(binwidth = 0.5,
    fill = "firebrick1") + labs(x = "Log CO2 emission distribution across all years",
    title = "Log CO2 emission distribution in years 1992-2016") +
    theme_bw()

dt <- data.frame(panelData)
stargazer(dt, out = "summary_new.html", header = FALSE, type = "latex",
    omit = "year", title = "Descriptive Statistics for the transformed and original variables")

dt <- panelData[, .(m_lngdppc = mean(lngdppc, na.rm = TRUE),
    m_lnco2 = mean(lnco2, na.rm = TRUE)), by = .(country)]

ggplot(dt, aes(x = m_lngdppc, y = m_lnco2)) + geom_point(color = "darkblue") +
    geom_smooth() + labs(x = "Average Log GDP per capita per country",
    y = "Average Log CO2 emission per country", title = "Log CO2 emission on Log GDP per capita, PPP av
    theme_bw()

library(plm)

panelData_reg <- panelData %>% group_by(country) %>% mutate(lag_co2 = lag(co2),
    lag_gdppc = lag(gdppc)) %>% mutate(dlco2 = log(co2) - log(lag_co2),
    dlgdppc = log(gdppc) - (log(lag_gdppc)))

panelData_reg_1 <- panelData_reg %>% group_by(country) %>% mutate(lag_gdppc_1 = lag(lag_gdppc)) %>%
    dplyr::filter(!is.na(lag_gdppc_1)) %>% mutate(dlgdppc_1 = log(lag_gdppc) -
    log(lag_gdppc_1))

panelData_reg_2 <- panelData_reg_1 %>% group_by(country) %>%
    mutate(lag_gdppc_2 = lag(lag_gdppc_1)) %>% dplyr::filter(!is.na(lag_gdppc_2)) %>%
```

```r
    mutate(dlgdppc_2 = log(lag_gdppc_1) - log(lag_gdppc_2))

panelData_reg_3 <- panelData_reg_2 %>% group_by(country) %>%
    mutate(lag_gdppc_3 = lag(lag_gdppc_2)) %>% dplyr::filter(!is.na(lag_gdppc_3)) %>%
    mutate(dlgdppc_3 = log(lag_gdppc_2) - log(lag_gdppc_3))

panelData_reg_4 <- panelData_reg_3 %>% group_by(country) %>%
    mutate(lag_gdppc_4 = lag(lag_gdppc_3)) %>% dplyr::filter(!is.na(lag_gdppc_4)) %>%
    mutate(dlgdppc_4 = log(lag_gdppc_3) - log(lag_gdppc_4))

panelData_reg_5 <- panelData_reg_4 %>% group_by(country) %>%
    mutate(lag_gdppc_5 = lag(lag_gdppc_4)) %>% dplyr::filter(!is.na(lag_gdppc_5)) %>%
    mutate(dlgdppc_5 = log(lag_gdppc_4) - log(lag_gdppc_5))

fd1 <- plm(dlco2 ~ dlgdppc, data = panelData_reg, model = "fd")
fd2 <- plm(dlco2 ~ dlgdppc + dlgdppc_1, data = panelData_reg_1,
    model = "fd")
fd3 <- plm(dlco2 ~ dlgdppc + dlgdppc_1 + dlgdppc_2, data = panelData_reg_2,
    model = "fd")
fd4 <- plm(dlco2 ~ dlgdppc + dlgdppc_1 + dlgdppc_2 + dlgdppc_3,
    data = panelData_reg_3, model = "fd")
fd5 <- plm(dlco2 ~ dlgdppc + dlgdppc_1 + dlgdppc_2 + dlgdppc_3 +
    dlgdppc_4, data = panelData_reg_4, model = "fd")
fd6 <- plm(dlco2 ~ dlgdppc + dlgdppc_1 + dlgdppc_2 + dlgdppc_3 +
    dlgdppc_4 + +dlgdppc_5, data = panelData_reg_5, model = "fd")

library(stargazer)
library(sandwich)
library(lmtest)

# Standard errors were adjusted to make sure robust standard
# errors are used
cov1 <- vcovSCC(fd1, type = "HC1")
robust_se_1 <- sqrt(diag(cov1))
cov2 <- vcovSCC(fd2, type = "HC1")
robust_se_2 <- sqrt(diag(cov2))
cov3 <- vcovSCC(fd3, type = "HC1")
robust_se_3 <- sqrt(diag(cov3))
cov4 <- vcovSCC(fd4, type = "HC1")
robust_se_4 <- sqrt(diag(cov4))
cov5 <- vcovSCC(fd5, type = "HC1")
robust_se_5 <- sqrt(diag(cov5))
cov6 <- vcovSCC(fd6, type = "HC1")
robust_se_6 <- sqrt(diag(cov6))

stargazer(title = "First Differences with different Lags", list(fd1,
    fd2, fd3, fd4, fd5, fd6), digits = 2, column.labels = c("No Lag",
    "Lag:1", "Lag:2", "Lag:3", "Lag:4", "Lag:5"), model.names = FALSE,
    omit.stat = c("adj.rsq", "f"), dep.var.caption = "Dependent variable: Log CO2",
    out = "Reg_1.html", notes.align = "l", se = list(robust_se_1,
        robust_se_2, robust_se_3, robust_se_4, robust_se_5, robust_se_6),
    add.lines = list(c("Cumulative Coeff", round(sum(fd1$coefficients),
        2), round(sum(fd2$coefficients), 2), round(sum(fd3$coefficients),
        2), round(sum(fd4$coefficients), 2), round(sum(fd5$coefficients),
        2), round(sum(fd6$coefficients), 2))), header = FALSE,
    type = "latex")

ols1995 <- lm(data = panelData[year == 1995], lnco2 ~ lngdppc)
ols2007 <- lm(data = panelData[year == 2007], lnco2 ~ lngdppc)
ols2013 <- lm(data = panelData[year == 2013], lnco2 ~ lngdppc)
pools <- lm(data = panelData, lnco2 ~ lngdppc)
fe <- plm(data = panelData, lnco2 ~ lngdppc + year, model = "within")
fd_4lag <- plm(data = panelData_reg_4, dlco2 ~ dlgdppc + dlgdppc_1 +
    dlgdppc_2 + dlgdppc_3 + dlgdppc_4)
ld_gdp <- panelData[year == 2013]$lngdppc - panelData[year ==
    1992]$lngdppc
ld_co2 <- panelData[year == 2013]$lnco2 - panelData[year == 1992]$lnco2
ld <- lm(ld_co2 ~ ld_gdp)

# Standard errors were adjusted to make sure robust standard
# errors are used
cov2_1 <- vcovHC(ols1995, type = "HC1")
rob_se_1 <- sqrt(diag(cov2_1))
cov2_2 <- vcovHC(ols2007, type = "HC1")
rob_se_2 <- sqrt(diag(cov2_2))
cov2_3 <- vcovHC(ols2013, type = "HC1")
rob_se_3 <- sqrt(diag(cov2_3))
cov2_4 <- vcovHC(pools, type = "HC1")
rob_se_4 <- sqrt(diag(cov2_4))
cov2_5 <- vcovSCC(fe, type = "HC1")
rob_se_5 <- sqrt(diag(cov2_5))
cov2_6 <- vcovSCC(fd_4lag, type = "HC1")
rob_se_6 <- sqrt(diag(cov2_6))
cov2_7 <- vcovHC(ld, type = "HC1")
rob_se_7 <- sqrt(diag(cov2_7))

stargazer(title = "Comparing multiple models", list(ols1995,
    ols2007, ols2013, pools, fe, fd_4lag, ld), digits = 2, column.labels = c("OLS1995",
    "OLS2007", "OLS2013", "Pools", "FE", "FD Lag: 4", "LD"),
    model.names = FALSE, omit.stat = c("adj.rsq", "f", "ser"),
    dep.var.caption = "Dependent variable: Log CO2", out = "Reg_2.html",
    notes.align = "l", se = list(rob_se_1, rob_se_2, rob_se_3,
        rob_se_4, rob_se_5, rob_se_6, rob_se_7), dep.var.labels.include = FALSE,
    header = FALSE, type = "latex")

unbData <- merge(gdpData, co2Data, by.x = c("iso2c", "year",
    "country"), by.y = c("iso2c", "year", "country"))

unbData$iso2c <- NULL
names(unbData) <- c("year", "country", "gdppc", "co2")

unbData$lngdppc <- log(unbData$gdppc)
unbData$lnco2 <- log(unbData$co2)

miss_lngdppc <- merge(unbData[, mean(lngdppc, na.rm = TRUE),
    by = country], unbData[is.na(lngdppc), .N, by = country],
    by = "country", all = TRUE)

miss_lnco2 <- merge(unbData[, mean(lnco2, na.rm = TRUE), by = country],
    unbData[is.na(lnco2), .N, by = country], by = "country",
    all = TRUE)

ggplot(miss_lngdppc) + aes(x = N) + geom_histogram(binwidth = 1,
    fill = "deepskyblue3") + labs(x = "Average number of missing observations",
    title = "Distribution of average missing observations of Log GDP per capita in years 1992-2016") +
    theme_bw()

ggplot(miss_lnco2) + aes(x = N) + geom_histogram(binwidth = 1,
    fill = "firebrick1") + labs(x = "Average number of missing observations",
    title = "Distribution of average missing observations of Log CO2 emission in years 1992-2016") +
    theme_bw()

ggplot(miss_lngdppc, aes(y = V1, x = N)) + geom_point(color = "deepskyblue3") +
    geom_smooth(method = "lm", se = FALSE) + labs(x = "Average number of missing observations",
    y = "Mean Log GDP per capita per country", title = "Mean Log GDP per capita per country in years 19
    theme_bw()

ggplot(miss_lnco2, aes(y = V1, x = N)) + geom_point(color = "firebrick1") +
    geom_smooth(method = "lm", se = FALSE) + labs(x = "Average number of missing observations",
    y = "Mean Log CO2 emission per country", title = "Mean Log CO2 emission in years 1992-2016") +
    theme_bw()

unbData_reg <- unbData %>% group_by(country) %>% mutate(lag_co2 = lag(co2),
    lag_gdppc = lag(gdppc)) %>% mutate(dlco2 = log(co2) - log(lag_co2),
    dlgdppc = log(gdppc) - (log(lag_gdppc)))

unbData_reg_1 <- unbData_reg %>% group_by(country) %>% mutate(lag_gdppc_1 = lag(lag_gdppc)) %>%
    dplyr::filter(!is.na(lag_gdppc_1)) %>% mutate(dlgdppc_1 = log(lag_gdppc) -
    log(lag_gdppc_1))

unbData_reg_2 <- unbData_reg_1 %>% group_by(country) %>% mutate(lag_gdppc_2 = lag(lag_gdppc_1)) %>%
    dplyr::filter(!is.na(lag_gdppc_2)) %>% mutate(dlgdppc_2 = log(lag_gdppc_1) -
    log(lag_gdppc_2))

unbData_reg_3 <- unbData_reg_2 %>% group_by(country) %>% mutate(lag_gdppc_3 = lag(lag_gdppc_2)) %>%
    dplyr::filter(!is.na(lag_gdppc_3)) %>% mutate(dlgdppc_3 = log(lag_gdppc_2) -
    log(lag_gdppc_3))

unbData_reg_4 <- unbData_reg_3 %>% group_by(country) %>% mutate(lag_gdppc_4 = lag(lag_gdppc_3)) %>%
    dplyr::filter(!is.na(lag_gdppc_4)) %>% mutate(dlgdppc_4 = log(lag_gdppc_3) -
    log(lag_gdppc_4))

u_ols2007 <- lm(data = unbData[year == 2007], lnco2 ~ lngdppc)
u_fe <- plm(data = unbData, lnco2 ~ lngdppc + year, model = "within")
u_fd_4lag <- plm(data = unbData_reg_4, dlco2 ~ dlgdppc + dlgdppc_1 +
    dlgdppc_2 + dlgdppc_3 + dlgdppc_4)
u_ld_gdp <- unbData[year == 2013]$lngdppc - unbData[year == 1992]$lngdppc
u_ld_co2 <- unbData[year == 2013]$lnco2 - unbData[year == 1992]$lnco2
u_ld <- lm(u_ld_co2 ~ u_ld_gdp)
```

```
# Standard errors were adjusted to make sure robust standard
# errors are used
u_cov2_2 <- vcovHC(u_ols2007, type = "HC1")
u_rob_se_2 <- sqrt(diag(u_cov2_2))
u_cov2_5 <- vcovSCC(u_fe, type = "HC1")
u_rob_se_5 <- sqrt(diag(u_cov2_5))
u_cov2_6 <- vcovSCC(u_fd_4lag, type = "HC1")
u_rob_se_6 <- sqrt(diag(u_cov2_6))
u_cov2_7 <- vcovHC(u_ld, type = "HC1")
u_rob_se_7 <- sqrt(diag(u_cov2_7))

stargazer(title = "Comparing multiple models", list(u_ols2007,
    u_fe, u_fd_4lag, u_ld), digits = 2, column.labels = c("OLS2007",
    "FE", "FD Lag: 4", "LD"), model.names = FALSE, omit.stat = c("adj.rsq",
    "f", "ser"), dep.var.caption = "Dependent variable: Log CO2",
    out = "Reg_3.html", notes.align = "l", se = list(rob_se_2,
        rob_se_5, rob_se_6, rob_se_7), dep.var.labels.include = FALSE,
    header = FALSE, type = "latex")

# SEARCHING FOR DATA: Fossil fuel energy consumption
fossil <- WDIsearch("fossil")
fossilcode <- fossil[match("Fossil fuel energy consumption (% of total)",
    fossil[, 2]), 1]

# DATA DOWNLOAD: fossil
fossil_dat = WDI(indicator = fossilcode, start = 1992, end = 2016)

# FILTERING OUT REGIONS
fossil_dt <- data.table(fossil_dat)
exclusionList <- fossil_dt[, .(itemCnt = .N), by = .(code = fossil_dt$iso2c)][1:47,
    1]
fossil_Data <- subset(fossil_dt, !(fossil_dt$iso2c %in% exclusionList$code))

# MERGING DATA INTO ORIGINAL PANEL
unbData_new <- merge(unbData, fossil_Data, by.x = c("year", "country"),
    by.y = c("year", "country"))

unbData_new$iso2c <- NULL
names(unbData_new) <- c("year", "country", "gdppc", "co2", "lngdppc",
    "lnco2", "fossil")

unbData_new$lnfos <- log(unbData_new$fossil + 1)

unbData_fos <- unbData_new %>% group_by(country) %>% mutate(lag_co2 = lag(co2),
    lag_gdppc = lag(gdppc), lag_fos = lag(fossil)) %>% mutate(dlco2 = log(co2) -
    log(lag_co2), dlgdppc = log(gdppc) - (log(lag_gdppc)), dlfos = log(fossil) -
    (log(lag_fos)))

unbData_fos_1 <- unbData_fos %>% group_by(country) %>% mutate(lag_gdppc_1 = lag(lag_gdppc)) %>%
    dplyr::filter(!is.na(lag_gdppc_1)) %>% mutate(dlgdppc_1 = log(lag_gdppc) -
    log(lag_gdppc_1))

unbData_fos_2 <- unbData_fos_1 %>% group_by(country) %>% mutate(lag_gdppc_2 = lag(lag_gdppc_1)) %>%
    dplyr::filter(!is.na(lag_gdppc_2)) %>% mutate(dlgdppc_2 = log(lag_gdppc_1) -
    log(lag_gdppc_2))

unbData_fos_3 <- unbData_fos_2 %>% group_by(country) %>% mutate(lag_gdppc_3 = lag(lag_gdppc_2)) %>%
    dplyr::filter(!is.na(lag_gdppc_3)) %>% mutate(dlgdppc_3 = log(lag_gdppc_2) -
    log(lag_gdppc_3))

unbData_fos_4 <- unbData_fos_3 %>% group_by(country) %>% mutate(lag_gdppc_4 = lag(lag_gdppc_3)) %>%
    dplyr::filter(!is.na(lag_gdppc_4)) %>% mutate(dlgdppc_4 = log(lag_gdppc_3) -
    log(lag_gdppc_4))

unbData_fos$lnco2[which(is.nan(unbData_fos$lnco2))] = NA
unbData_fos$lnco2[which(unbData_fos$lnco2 == Inf)] = NA
unbData_fos$lngdppc[which(is.nan(unbData_fos$lngdppc))] = NA
unbData_fos$lngdppc[which(unbData_fos$lngdppc == Inf)] = NA
unbData_fos$lnfos[which(is.nan(unbData_fos$lnfos))] = NA
unbData_fos$lnfos[which(unbData_fos$lnfos == Inf)] = NA

data2007 <- subset(unbData_fos, year == 2007)
data2013 <- subset(unbData_fos, year == 2013)
data1992 <- subset(unbData_fos, year == 1992)
fos_ols2007 <- lm(data = data2007, lnco2 ~ lngdppc + lnfos)
fos_fe <- plm(data = unbData_fos, lnco2 ~ lngdppc + lnfos, model = "within")
fos_fd_4lag <- plm(data = unbData_fos_4, dlco2 ~ dlgdppc + dlgdppc_1 +
    dlgdppc_2 + dlgdppc_3 + dlgdppc_4 + dlfos)
fos_ld_gdp <- data2013$lngdppc - data1992$lngdppc
fos_ld_co2 <- data2013$lnco2 - data1992$lnco2
fos_ld_fos <- data2013$lnfos - data1992$lnfos
fos_ld <- lm(fos_ld_co2 ~ fos_ld_gdp + fos_ld_fos)

# Standard errors were adjusted to make sure robust standard
# errors are used
fos_cov2_2 <- vcovHC(fos_ols2007, type = "HC1")
fos_rob_se_2 <- sqrt(diag(fos_cov2_2))
fos_cov2_5 <- vcovSCC(fos_fe, type = "HC1")
fos_rob_se_5 <- sqrt(diag(fos_cov2_5))
fos_cov2_6 <- vcovSCC(fos_fd_4lag, type = "HC1")
fos_rob_se_6 <- sqrt(diag(fos_cov2_6))
fos_cov2_7 <- vcovHC(fos_ld, type = "HC1")
fos_rob_se_7 <- sqrt(diag(fos_cov2_7))

stargazer(title = "Comparing multiple models", list(fos_ols2007,
    fos_fe, fos_fd_4lag, fos_ld), digits = 2, column.labels = c("OLS2007",
    "FE", "FD Lag: 4", "LD"), model.names = FALSE, omit.stat = c("adj.rsq",
    "f", "ser"), dep.var.caption = "Dependent variable: Log CO2",
    out = "Reg_4.html", notes.align = "l", se = list(fos_rob_se_2,
        fos_rob_se_5, fos_rob_se_6, fos_rob_se_7), dep.var.labels.include = FALSE,
    header = FALSE, type = "latex")

panel2007 <- subset(unbData_fos, year == 2007)
dt_2007 <- data.frame(panel2007)

stargazer(dt_2007, out = "summary_2007.html", header = FALSE,
```

```r
    type = "latex", omit = "year", title = "Descriptive Statistics for panel 2007")

ggplot(panel2007) + aes(x = gdppc) + geom_histogram(binwidth = 2000,
    fill = "deepskyblue3") + labs(x = "GDP per capita distribution in 2007, constant 2005 prices",
    title = "Histogram of GDP per capita distribution in 2007") +
    theme_bw()

# Creating balanced dataset of countries
dt_2007 <- unbData_fos[!is.na(unbData_fos$gdppc) & !is.na(unbData_fos$co2) &
    !is.na(unbData_fos$fossil), ]
dt_2007_4 <- unbData_fos_4[!is.na(unbData_fos$gdppc) & !is.na(unbData_fos$co2) &
    !is.na(unbData_fos$fossil), ]

# Group 1
dt_2007$group <- ifelse(dt_2007$gdppc <= 3000, 0, 1)
dt_2007_4$group <- ifelse(dt_2007_4$gdppc <= 3000, 0, 1)
dt_2007_1_2007 <- data.frame(subset(dt_2007, group == 1 & year ==
    2007))
dt_2007_1_2016 <- data.frame(subset(dt_2007, group == 1 & year ==
    2016))
dt_2007_1_1992 <- data.frame(subset(dt_2007, group == 1 & year ==
    1992))
dt_2007_1 <- data.frame(subset(dt_2007_4, group == 1))

p_ols2007 <- lm(data = dt_2007_1_2007, lnco2 ~ lngdppc + lnfos)
p_fe <- plm(data = dt_2007_1, lnco2 ~ lngdppc + lnfos, model = "within")
p_fd_4lag <- plm(data = dt_2007_1, dlco2 ~ dlgdppc + dlgdppc_1 +
    dlgdppc_2 + dlgdppc_3 + dlgdppc_4 + dlfos)

# Standard errors were adjusted to make sure robust standard
# errors are used
p_cov2_2 <- vcovHC(p_ols2007, type = "HC1")
p_rob_se_2 <- sqrt(diag(p_cov2_2))
p_cov2_5 <- vcovSCC(p_fe, type = "HC1")
p_rob_se_5 <- sqrt(diag(p_cov2_5))
p_cov2_6 <- vcovSCC(p_fd_4lag, type = "HC1")
p_rob_se_6 <- sqrt(diag(p_cov2_6))

# Group 2
dt_2007_0_2007 <- data.frame(subset(dt_2007, group == 0 & year ==
    2007))
dt_2007_0_2016 <- data.frame(subset(dt_2007, group == 0 & year ==
    2016))
dt_2007_0_1992 <- data.frame(subset(dt_2007, group == 0 & year ==
    1992))
dt_2007_0 <- data.frame(subset(dt_2007_4, group == 0))

r_ols2007 <- lm(data = dt_2007_0_2007, lnco2 ~ lngdppc + lnfos)
r_fe <- plm(data = dt_2007_0, lnco2 ~ lngdppc + lnfos, model = "within")
r_fd_4lag <- plm(data = dt_2007_0, dlco2 ~ dlgdppc + dlgdppc_1 +
    dlgdppc_2 + dlgdppc_3 + dlgdppc_4 + dlfos)

# Standard errors were adjusted to make sure robust standard
# errors are used
r_cov2_2 <- vcovHC(r_ols2007, type = "HC1")
r_rob_se_2 <- sqrt(diag(r_cov2_2))
r_cov2_5 <- vcovSCC(r_fe, type = "HC1")
r_rob_se_5 <- sqrt(diag(r_cov2_5))
r_cov2_6 <- vcovSCC(r_fd_4lag, type = "HC1")
r_rob_se_6 <- sqrt(diag(r_cov2_6))

stargazer(title = "Comparing multiple models", list(p_ols2007,
    r_ols2007, p_fe, r_fe, p_fd_4lag, r_fd_4lag), digits = 2,
    column.labels = c("OLS2007:1", "OLS2007:0", "FE:1", "FE:0",
        "FD Lag4:1", "FD Lag4:0"), model.names = FALSE, omit.stat = c("adj.rsq",
        "f", "ser"), dep.var.caption = "Dependent variable: Log CO2",
    out = "Reg_5.html", notes.align = "l", se = list(p_rob_se_2,
        r_rob_se_2, p_rob_se_5, r_rob_se_5, p_rob_se_6, r_rob_se_6),
    dep.var.labels.include = FALSE, header = FALSE, type = "latex")
```