DA2 Assignment 1

Istvan Janco #2003877

11/28/2020

Introduction

Research Question: Do the number of recorded daily deaths due to Covid 19 correlate with the number of daily recorded cases?

Introduction of the data

- Outcome variable: Recorded deaths per capita due to Covid 19 for the day of 8/10/2020
- Explanatory variable: Recorded cases per capita due to Covid 19 for the day of 8/10/2020
- Population: Total number of Covid 19 deaths and cases per capita in the world. The sample represents a snapshot of the general pattern of association for he day of 8/10/2020.
- Potential data quality issues and cleaning:
  - After the data was downloaded, all non-country observations (e.g. cruise ships) were removed, which yields incomplete coverage.
  - Some observations which contained missing values for population, confirmed cases of Covid 19 or number of deaths due to Covid 19 were dropped.
  - The variables were scaled up by 1M in order to enable a meaningful interpretation.

Variables Summary and Distribution

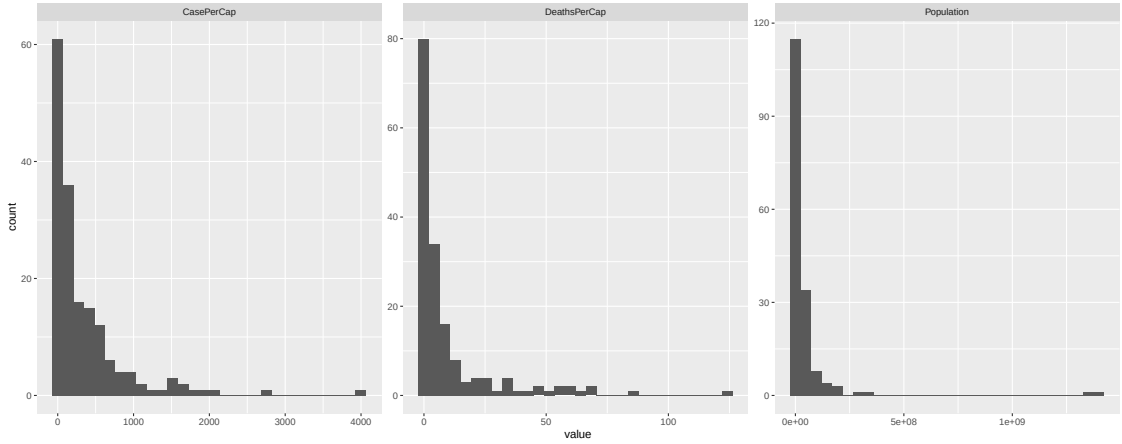


Table 1: Variable Summary

CasePerCap	DeathsPerCap	Population
Min. : 0.878	Min. : 0.00867	Min. :3.386e+04
1st Qu.: 38.659	1st Qu.: 0.68558	1st Qu.:2.955e+06

CasePerCap	DeathsPerCap	Population
Median : 128.912	Median : 2.78911	Median :1.028e+07
Mean : 354.122	Mean : 10.27219	Mean :4.473e+07
3rd Qu.: 446.627	3rd Qu.: 10.29217	3rd Qu.:3.209e+07

- The the distribution for all of the variables seems to be log normal.
- There are some extreme values on the left hand side. These reflect the population of observed coutries.

Transformation of Variables

Log-log transformation (appendix 1.4.) is the most optimal due to the following reasons:
- Substantive: it allows for a somewhat meaningful interpretation - Statistical: better approximation makes distribution is close to linear, thus it's easier to fit a proper model It is important to mention that some of the observations were dropped due to yielding negative values after the log transformation.

Model Comparison Results

Based on model comparison (Table 2) our chosen model is reg1 - ln_CPC ~ ln_DPC. The reasoning outlined below.

- Substantive: the model works well with log-log transformed variables. In addition, the magnitude of coefficients seems to be meaningful.
- Statistical: the model is simple model, which enables easy to interpretation. It can offset the effect of log-log transformation.

Hypothesis Testinf on Beta

- The test hypothesis is the following $H_0 : \beta = 0$, $H_A : \beta \neq 0$ or not in our model.
  - The estimated t-statistics is 9.73, with p-value: $1.5153875 \times 10^{-16}$.
  - Choosing a significance level of p = 0.05.
  - Thus we reject the H_0, which means the number of daily recorded deaths per capita due to Covid 19 is not uncorrelated with daily number to recorded cases of Covid 19 per capita.
  - Based on the p-value being less then the the significance level, the conclusion can be made that the sample data provides enough evidence to reject the null hypothesis. The data favors the hypothesis that there is a non-zero correlation. Changes in the independent variable are associated with changes in the dependent variable.

	Linear	Cubic	P.L.S	WOLS
Intercept	-2.41^{***}	7.15	0.56	-3.54^{***}
	(0.43)	(4.09)	(0.34)	(0.62)
ln(Cases/capita)	0.78^{***}	-5.26^{*}		1.00^{***}
	(0.08)	(2.35)		(0.10)
ln(Cases/capita)^2		1.20^{**}		
		(0.45)		
ln(Cases/capita)^3		-0.08^{**}		
		(0.03)		
ln(Cases/capita<=50)			-0.02	
			(0.11)	
ln(Cases/capita>50)			0.85^{***}	
			(0.08)	
R^2	0.55	0.61	0.57	0.73
Num. obs.	113	113	113	113

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2: ln Recorded Deaths due to Covid 19 and ln Recorded Cases of Covid 19

Residuals Analysis

- The summary of residual analysis is provided below:
  - The largest negative deviance from the predicted value is found in Qatar with predicted number of fatalities of 4, however the real value is 1.9.
  - The largest positive deviance from the predicted value is found in United Kingdom with deaths due to Covid 19 estimate of 2.4, however the real value is 4.2.

Executive Summary

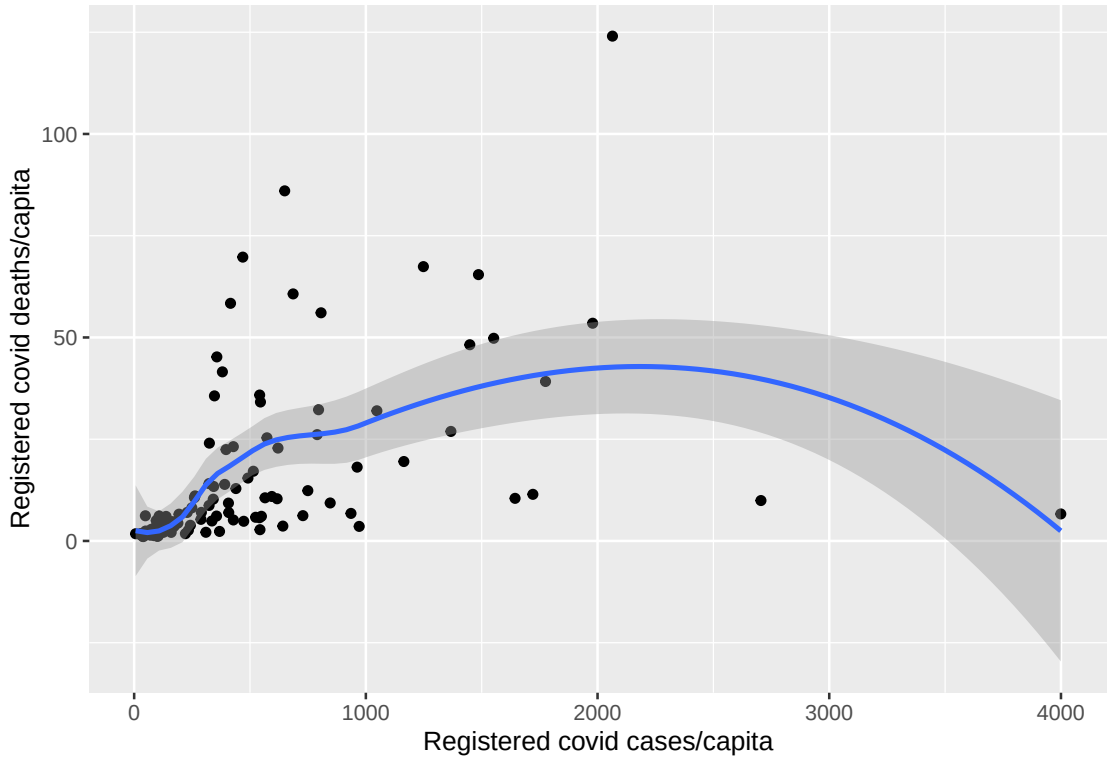
The correlation between the number of recorded daily deaths due to Covid 19 and number of daily recorded cases was investigated. The pattern of association resembled linear after log-log transformation, thus the decision to use a linear model was made. Based on the model, we can assume that there is a possible correlation between X and Y.

- The conducted analysis can be strengthened by adding observations for a longer time frame (e.g. a month, half a year, etc.)
- The analysis is weakened by the missing observations for countries that don't report Covid 19 cases with high frequeny.

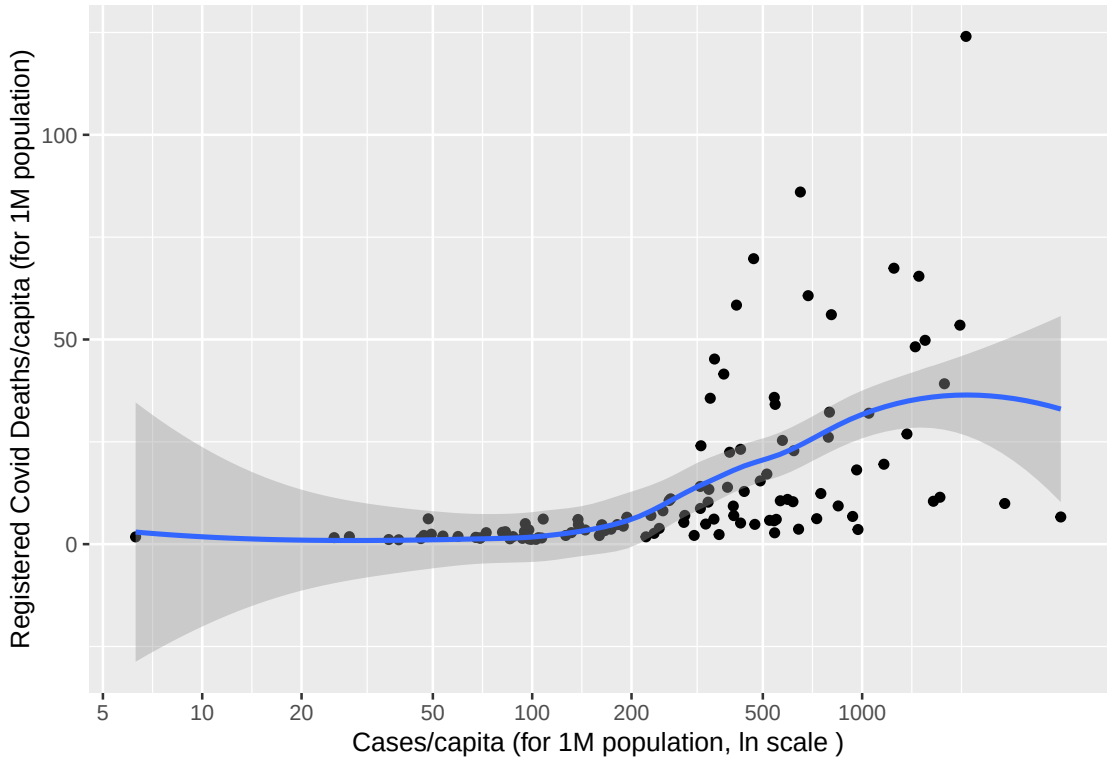
Appendix

1. Distributions based on variable transformation

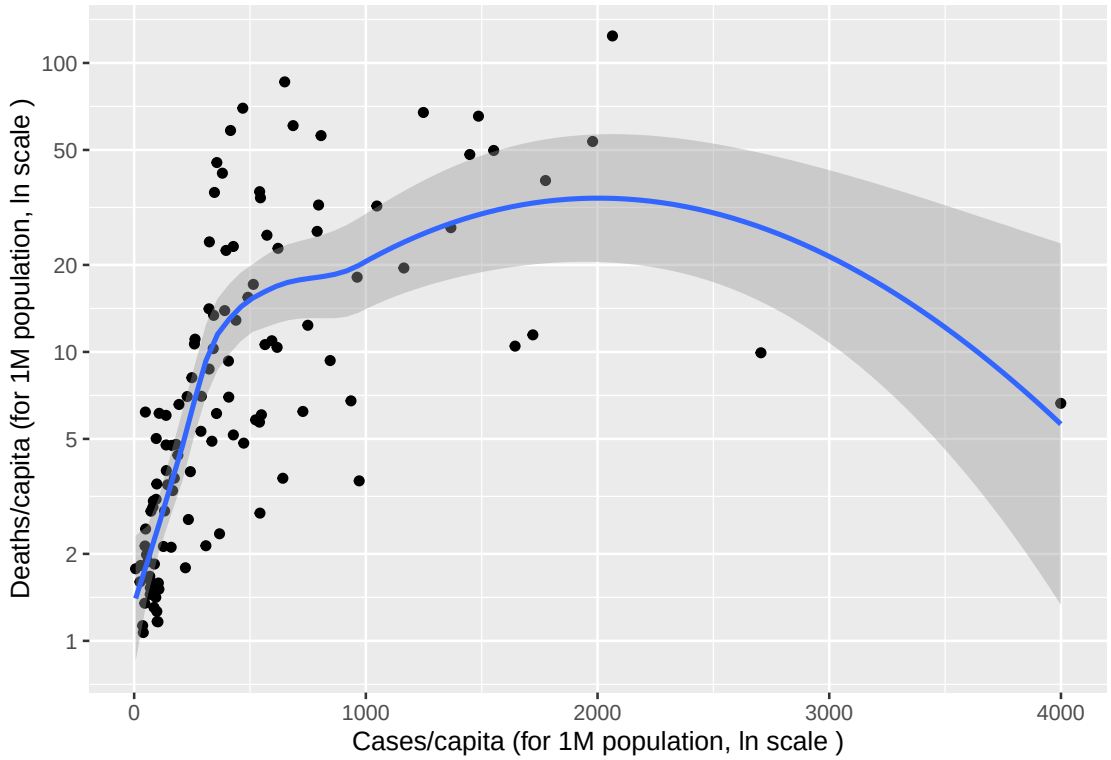
1.1. **Registered number of deaths - Registered number of cases: level-level model without scaling**



1.2. **Registered number of deaths/capita - ln(Registered number of cases/capita): log-transformation applied for registered cases/capita**

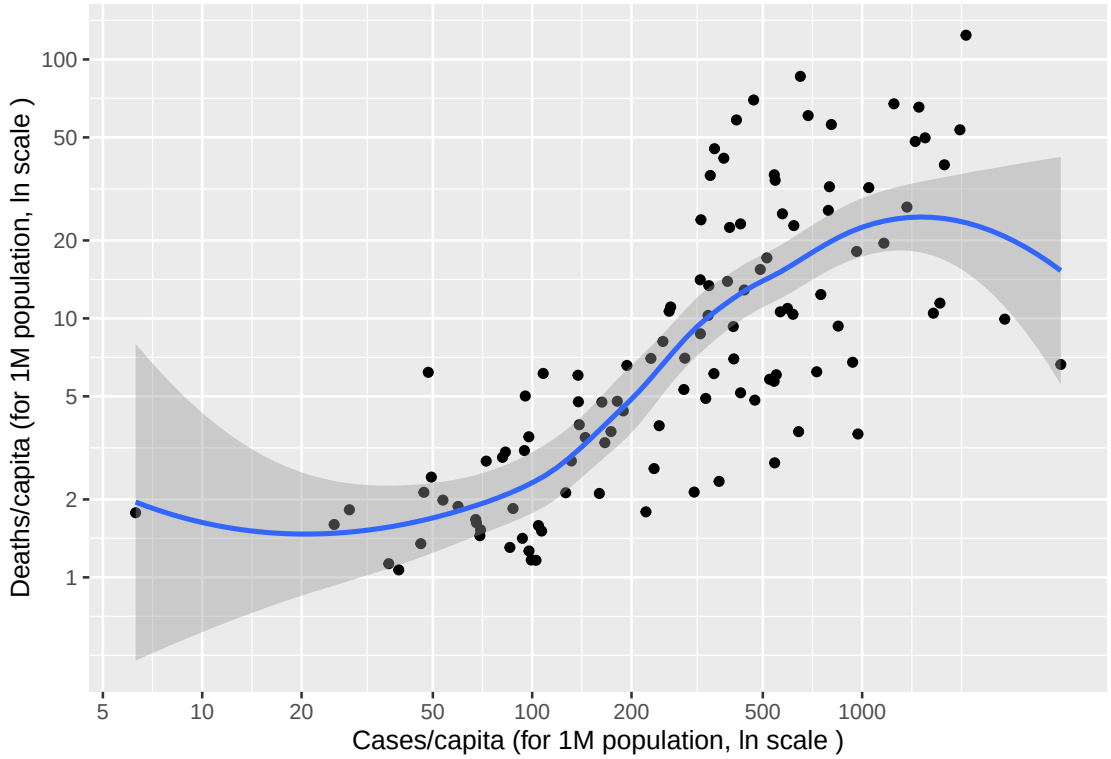


1.3. **ln(Registered number of deaths/capita) - Registered number of cases/capita: log-transformation applied for registered deaths/capita**



1.4. **ln(Registered number of deaths/capita) - ln(Registered number of cases/capita): log-**

transformation applied for registered deaths/capita and registered cases/capita



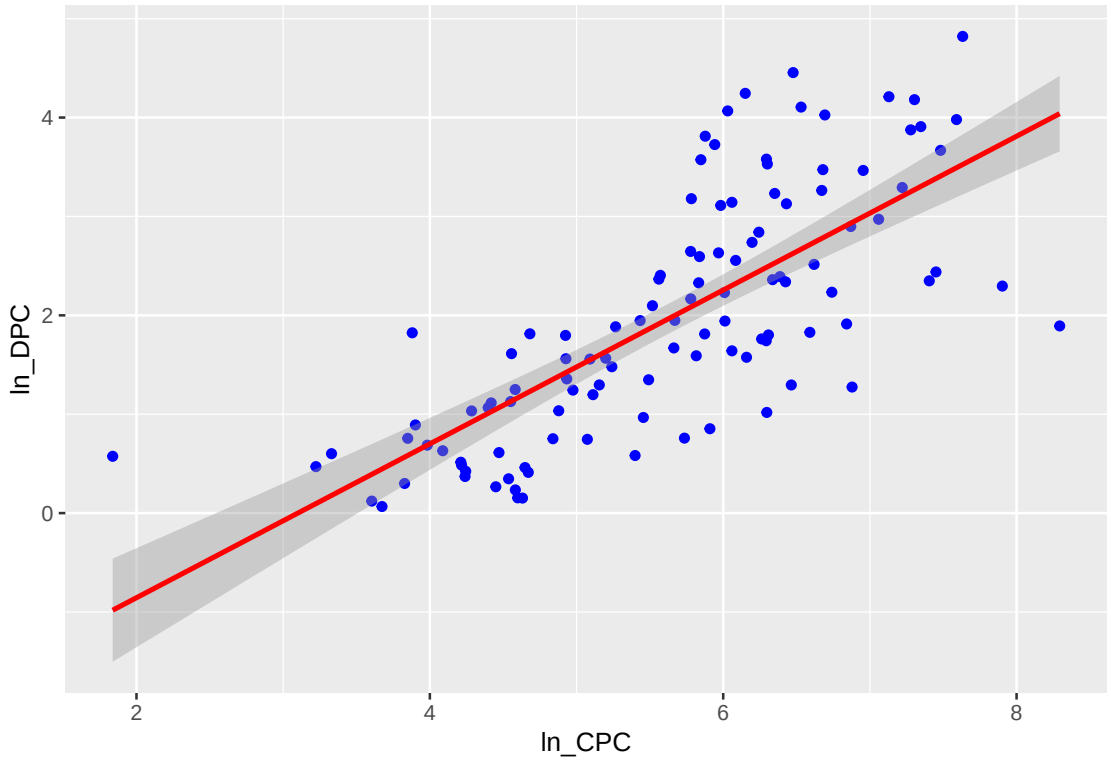
2. Explored Models:

Regressions

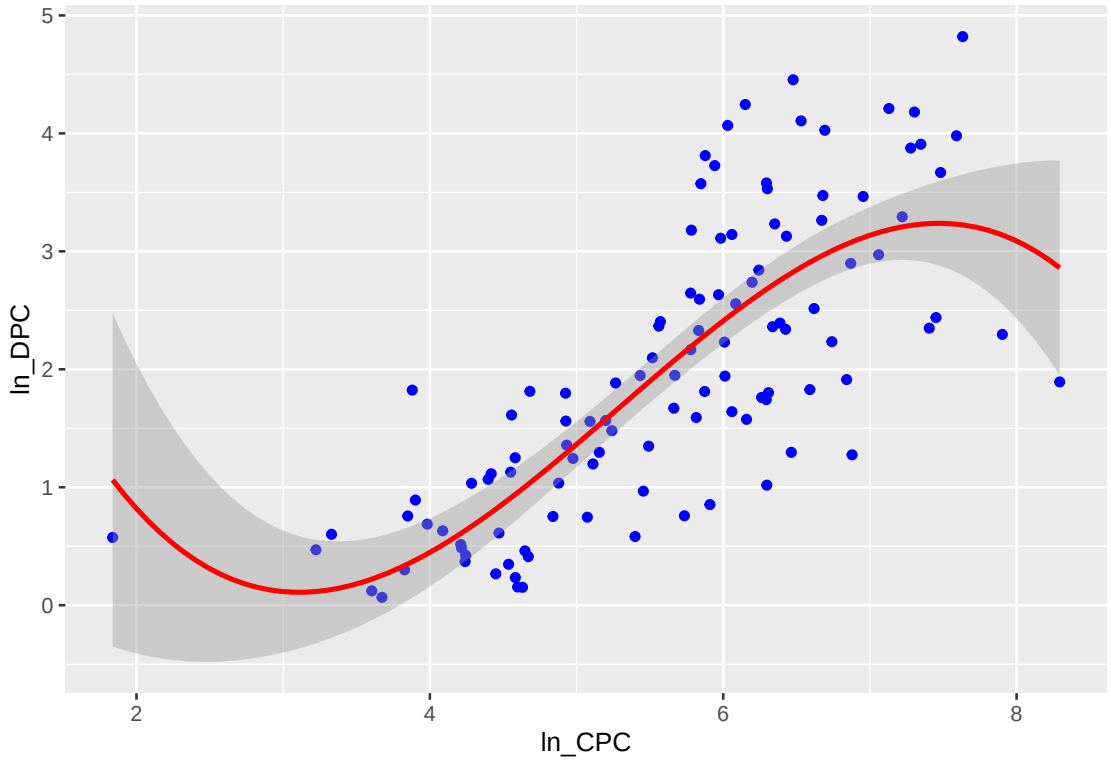
Investigating which regression model would enable a meaningful analysis. Using lm_robust to tackle issues associated heteroscedasticity.

2.1. reg1: ln_DPC = alpha + beta * ln_CPC

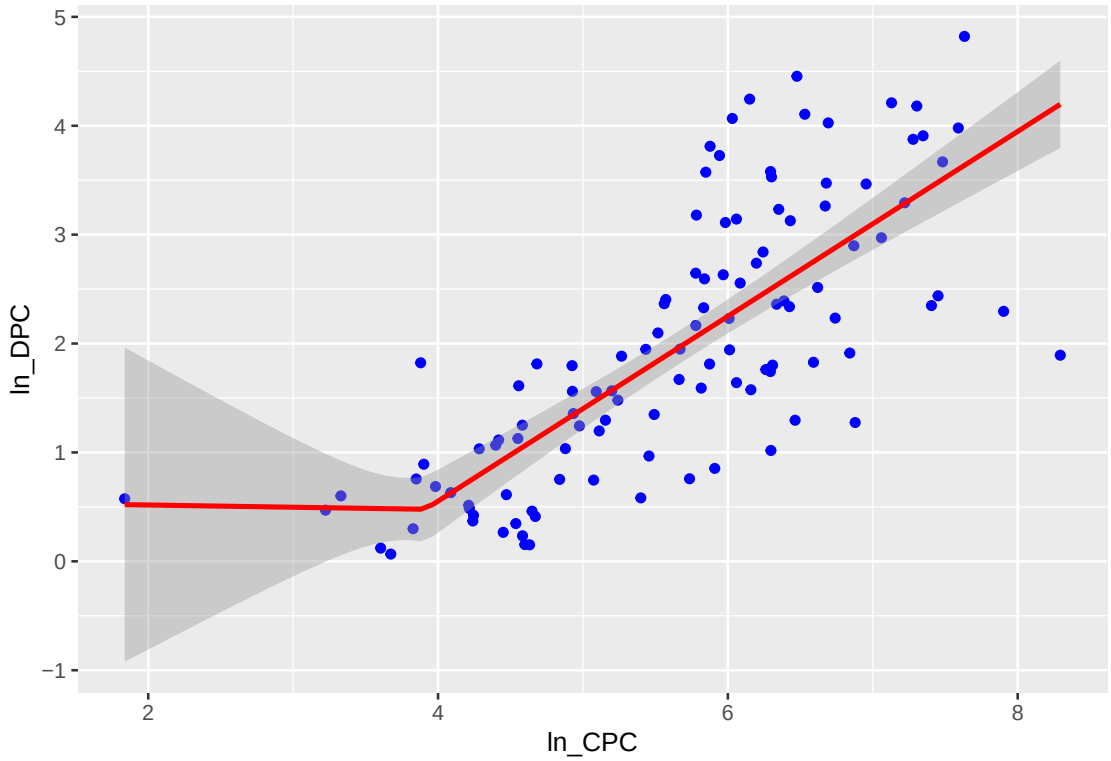
```
## `geom_smooth()` using formula 'y ~ x'
```



2.2. **reg2: ln_DPC = alpha + beta_1 * ln_CPC + beta_2 * ln_CPCˆ2 + beta_3 *ln_CPCˆ3**



2.3. **reg4: ln_DPC = alpha + beta_1 * ln_CPC * 1(ln_CPC < 50) + beta_2 * ln_CPC * 1(ln_CPC >= 50)**



2.4. **reg5: ln_DPC = alpha + beta * ln_CPC, weights: population (weighted-ols)**

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