Assignment #3

Katie Levesque, Sarah Unbehaun, Meerim Ruslanova

April 14, 2016

Quick Recap

This project aims to investigate the relationship between happiness levels across German federal states (Bundeslaender) and among individuals over the time horizon from 1990 to 2012. More specifically, this project explores whether the state level emissions, as well as some personal characteristics, affect life satisfaction of German citizens. The hypotheses state:

> *H1: Bundeslaender with higher emissions inversely affect reported levels of life satisfaction.*
>
> *H2: Reported individual concerns with the environment are, likewise, negatively reflected in the life satisfaction.*

Model 1 combines these two hypotheses: $satis = \beta_0 - \beta_1 * Emissions + \beta_2 * EnvironConcerns + u_j + r_i$

The relationship between the independent and dependent variables will be explored on both individual and state level via the multilevel modelling.

Data

The individual-level data is provided by the German Socio-Economic Panel Data GSOEP conducted by the German Institute for Economic Research DIW. Due to confidentiality restrictions, DIW could only supply a shortened sample with prior specified variables in a *.dta* format. Therefore, the GSOEP dataset is stored on the local drives and GitHub Climate-Happiness Repository. The original GSOEP file is cleaned and transformed into a shorter dataset with the help of the Stata Do-File. The short dataset contains the information on the main satisfaction and personal characteristic variables: reported levels of life satisfaction (on a scale from 0 to 10), subjective concerns about the environment, age, gender, employment, family status, and state residence of a respondent. Detailed labels and descriptions of the variables are given in the GSOEP codebook. All GSOEP-related files are stored on the GitHub server.

The state-level data, on the other hand, is gathered from three web-based sources: State Initiative for Core Indicators LIKI, Statista.com, Environmental-Economic Accounting of the Bundeslaender UGRdL and Agency for Renewable Agency of North Rhine-Westphalia AfEE.

A university subscription to *Statista.com* enabled access to historic state emissions from 1990 to 2012 for most of the Bundeslaender, except North Rhine-Westphalia (NRW). Since the website allows data downloads only in *Excel* and provides no unique URLs for each of them, 15 individual files were downloaded manually on a local machine, while manipulations were conducted with the help of R loops. The information on NRW involved more intensive research and data handling but were finally gathered and combined from the UGRdL (from 1990 to 2000) and *AfEE* (from 2000 to 2012) with R web-scraping functions. Fortunately, emissions are measured in the same units (annually emitted Carbon dioxide tons per capita). Hence, the yielded data frame of emissions is comprehensive and consistent, although there are missing observations on some years.

Simultaneously, the state efforts to curb their emissions and preserve local environment are reflected in their renewable energy indicators. This information is measured in percentage of renewables in the annual primary energy consumption, final energy consumption, and electricity consumption. The indicators had to be downloaded manually from *LIKI* in three separate excel files, which later on were cleaned, transformed and reshaped into suitable data frames in R.

Once the names of the Bundeslaender and the time frame of the three produced data frames match, they are easily merged in R into a final data set.

Descriptive Statistics

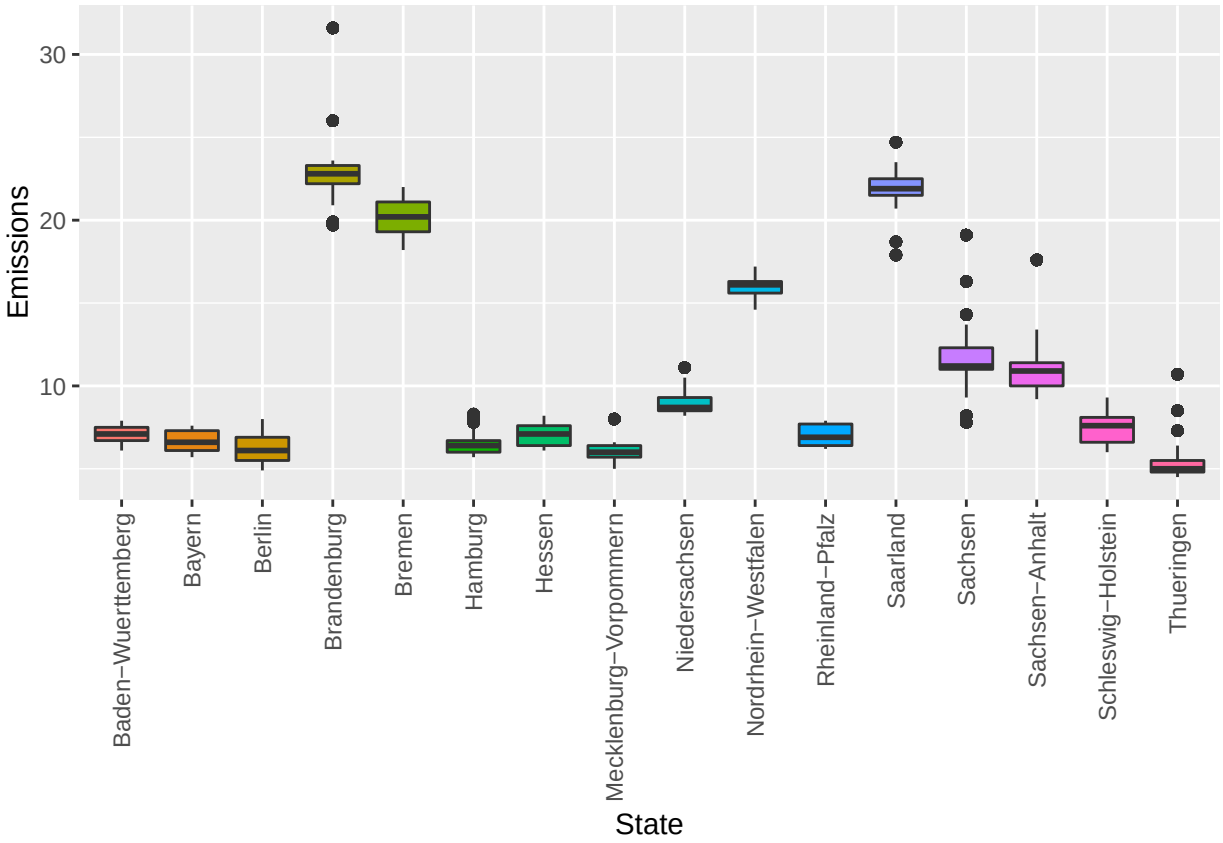
```
##     Baden-Wuerttemberg                 Bayern                 Berlin
##                  15935                  20889                   6328
##            Brandenburg                 Bremen                Hamburg
##                   8334                   1075                   1468
##                Hessen Mecklenburg-Vorpommern          Niedersachsen
##                  10886                   4073                   7930
##    Nordrhein-Westfalen        Rheinland-Pfalz               Saarland
##                  22048                   7460                   1208
##                Sachsen         Sachsen-Anhalt     Schleswig-Holstein
##                  11955                   8765                   4317
##             Thueringen
##                   8159
```

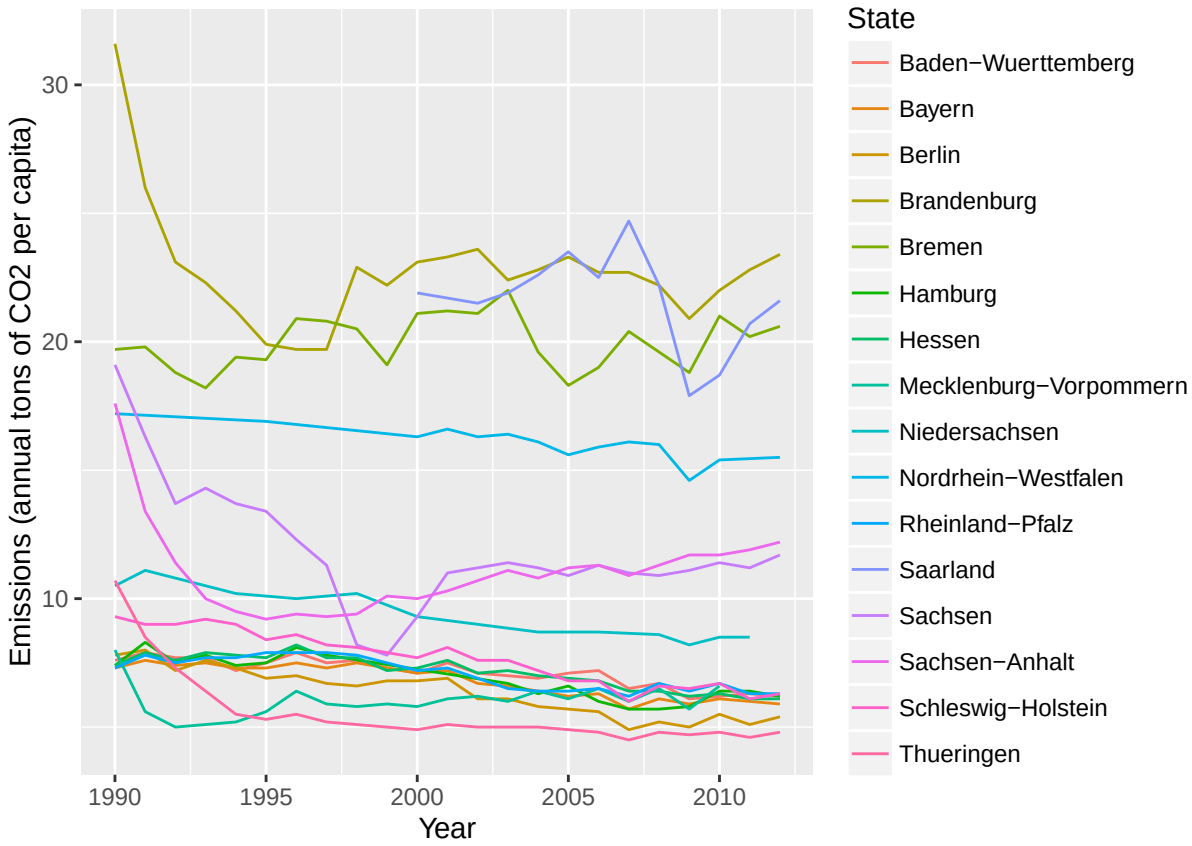
The data has a wide range in terms of the number of observations for each federal state. If we look a little closer, separating them by year, we see that some states are missing observations for a few years. Using Saarland, Nordrhein-Westfalen, and Hamburg in our analysis may require some adjustment with the missing years in mind.

```
##                         Year
## State                    1990 1991 1992 1993 1994 1995 1996 1997 1998 1999
##   Baden-Wuerttemberg      407  378  344  335  355  378  537  524  570  530
##   Bayern                  536  517  512  492  515  561  643  596  692  652
##   Berlin                  251  233  234  233  226  233  225  215  222  212
##   Brandenburg             371  337  315  307  294  276  273  262  275  268
##   Bremen                   34   32   33   30   32   35   32   31   34   24
##   Hamburg                  41   37   37   33   36   33   42   41    0    0
##   Hessen                  361  336  329  312  300  325  360  330  390  353
##   Mecklenburg-Vorpommern  219  204  178  170  175  172  157  153  171  165
##   Niedersachsen           368  347    0    0  310    0  403    0  462    0
##   Nordrhein-Westfalen     768    0    0    0    0  738    0    0    0    0
##   Rheinland-Pfalz         263  252  241  241  262  284  287  293  330  300
##   Saarland                  0    0    0    0    0    0    0    0    0    0
##   Sachsen                 568  522  491  472  457  428  419  395  432  419
##   Sachsen-Anhalt          433  407  396  372  353  335  309  308  318  317
##   Schleswig-Holstein      101  109   98   97  100   93  100   96  129  111
##   Thueringen              382  356  346  329  327  314  317  321  347  331
##                         Year
## State                    2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
##   Baden-Wuerttemberg     1116  970 1055  957  887  862  965  847  789  851
##   Bayern                 1449 1243 1387 1265 1189 1088 1187 1098 1010 1142
##   Berlin                  377  344  361  346  323  305  300  291  276  299
##   Brandenburg             462  451  463  453  434  399  416  394  366  409
##   Bremen                   71   53   74   74   69   54   47   48   48   64
##   Hamburg                   0    0    0  127  125  122  127  117  111  110
##   Hessen                  750  625  726  629  585  541  587  528  505  558
##   Mecklenburg-Vorpommern  239  207  223  221  207  195  235  204  200  202
##   Niedersachsen           843    0  849    0  743    0  778    0  694  773
##   Nordrhein-Westfalen    2156 1895 2034 1859 1747 1616 1719 1575 1482 1567
##   Rheinland-Pfalz         489  432  442  392  382  353  370  322  310  317
##   Saarland                133  113  114  108  100   97   97   86   73   73
##   Sachsen                 716  632  640  614  595  552  586  543  510  550
##   Sachsen-Anhalt          503  481  469  436  424  406  410  399  357  361
```

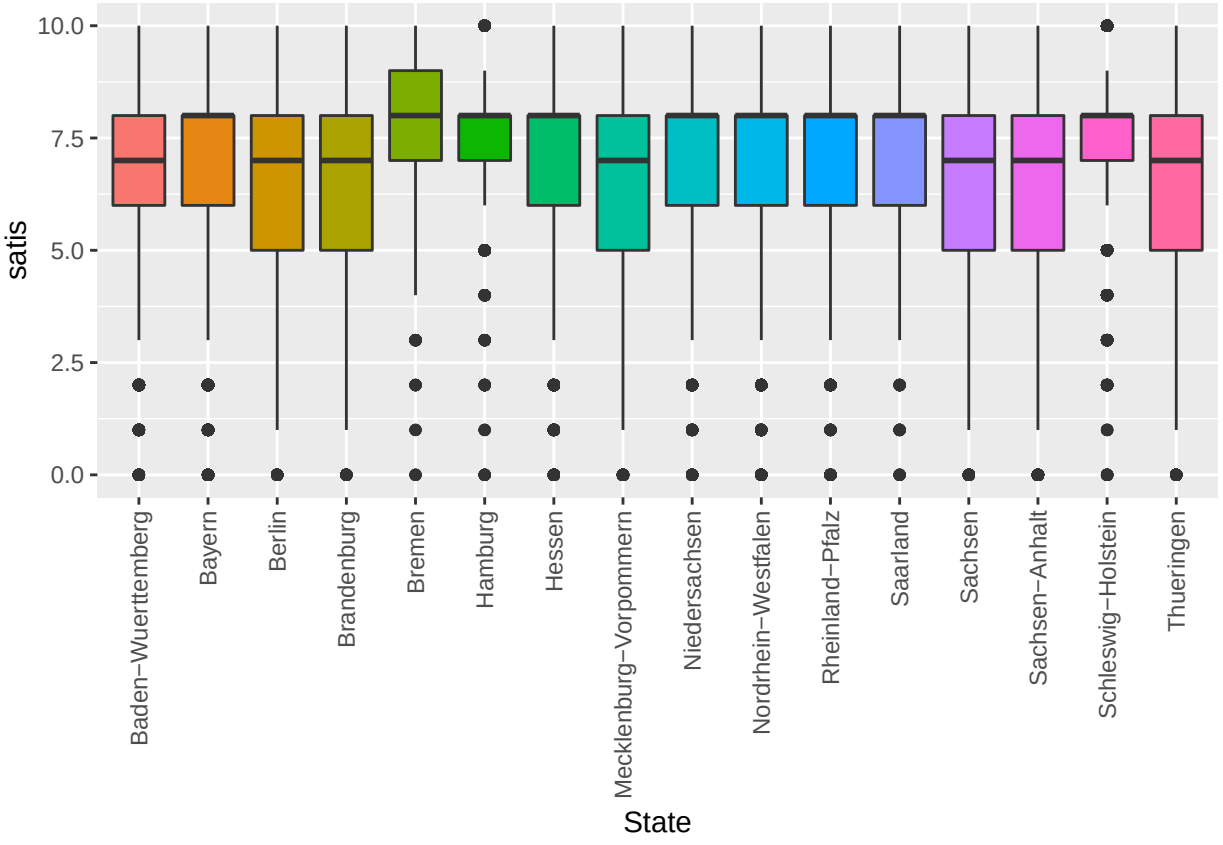
```
##    Schleswig-Holstein      295  266  318  272  274  256  271  246  209  222
##    Thueringen              496  436  420  404  390  357  367  341  324  338
##                          Year
## State                     2010 2011 2012
##    Baden-Wuerttemberg      730  731  817
##    Bayern                  998  995 1122
##    Berlin                  268  259  295
##    Brandenburg             358  363  388
##    Bremen                   58   46   52
##    Hamburg                 109  106  114
##    Hessen                  488  450  518
##    Mecklenburg-Vorpommern  176    0    0
##    Niedersachsen           700  660    0
##    Nordrhein-Westfalen    1410    0 1482
##    Rheinland-Pfalz         281  289  328
##    Saarland                 66   72   76
##    Sachsen                 475  453  486
##    Sachsen-Anhalt          319  314  338
##    Schleswig-Holstein      198  211  245
##    Thueringen              304  291  321
```

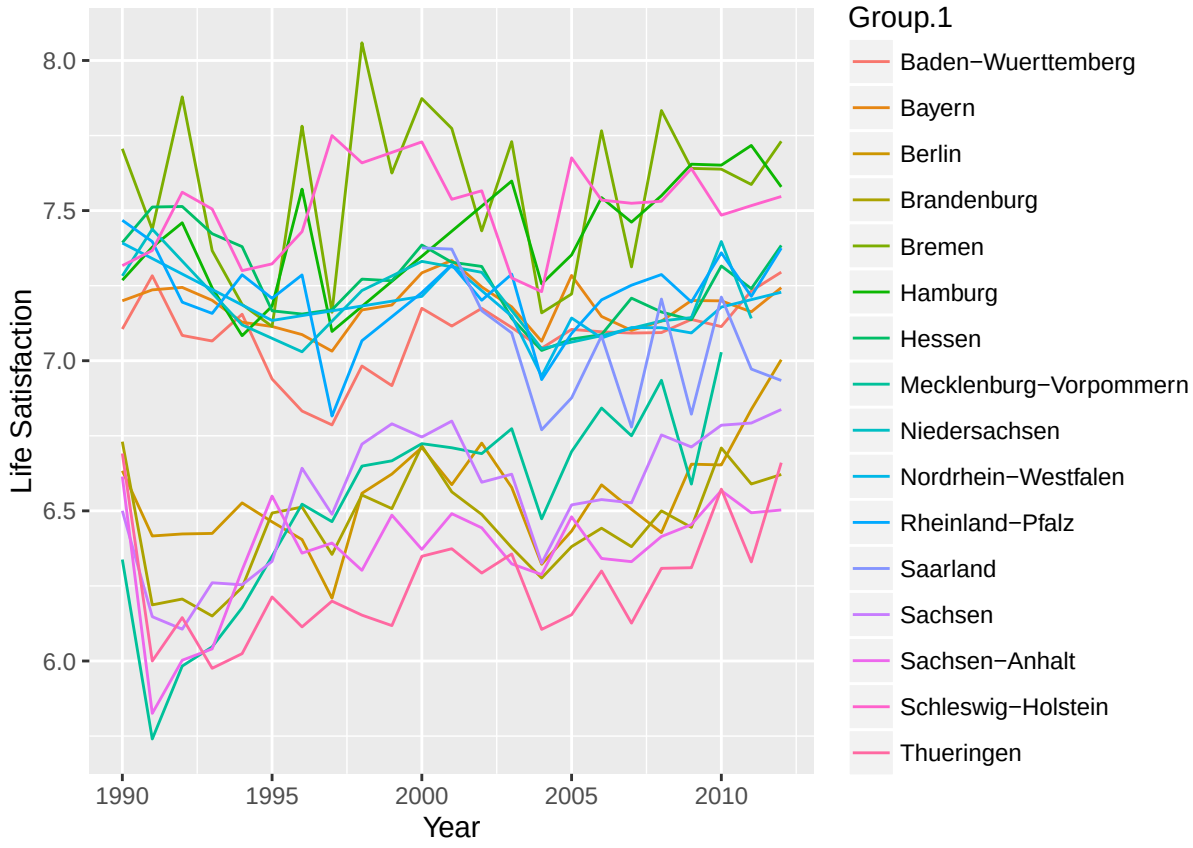
Because we are looking at emissions and happiness over a period of years, the values of emissions for each state also vary, which we can see by state and over time. The colors are coordinated between the box plots and line graphs.





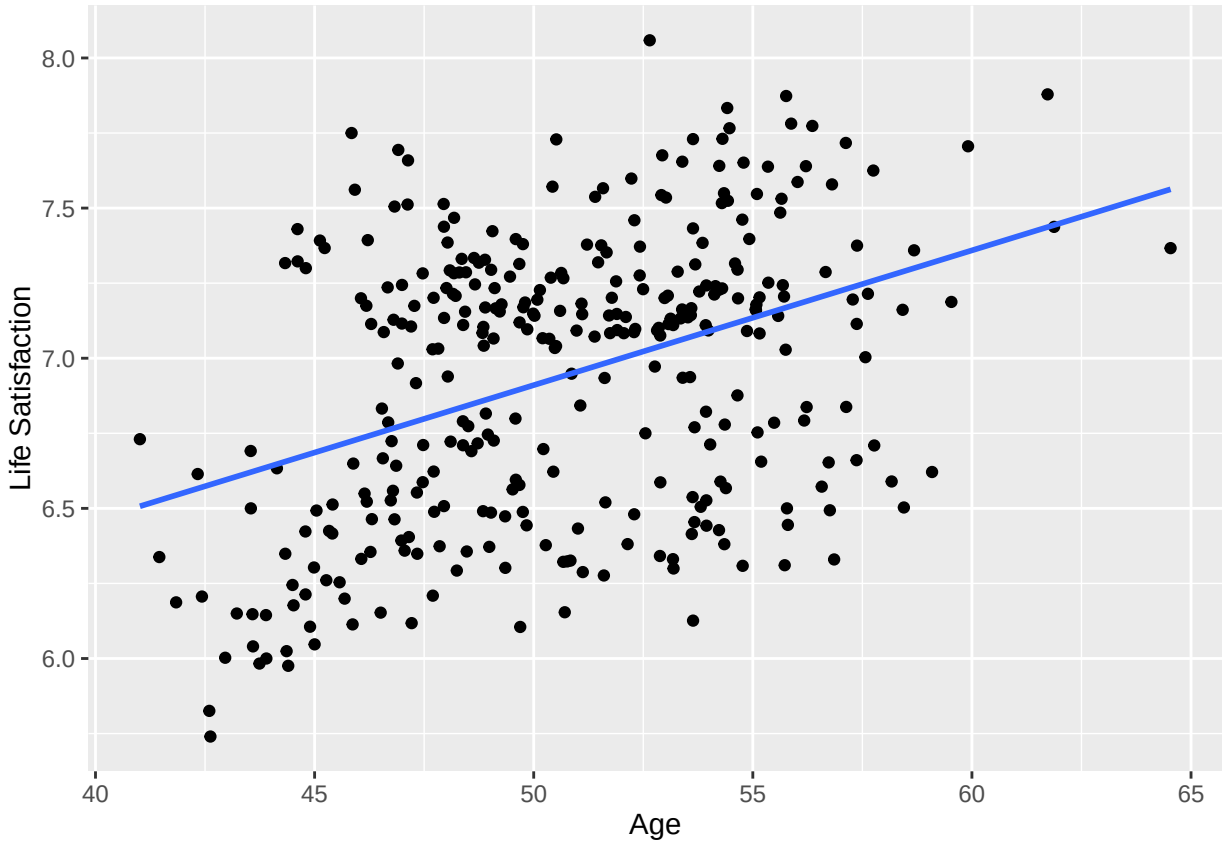
The main explanatory variable of interest is life satisfaction. As with emissions, there is variation within each state and the variation in life satisfaction across years. The following charts use the mean of the individual observations from each state.

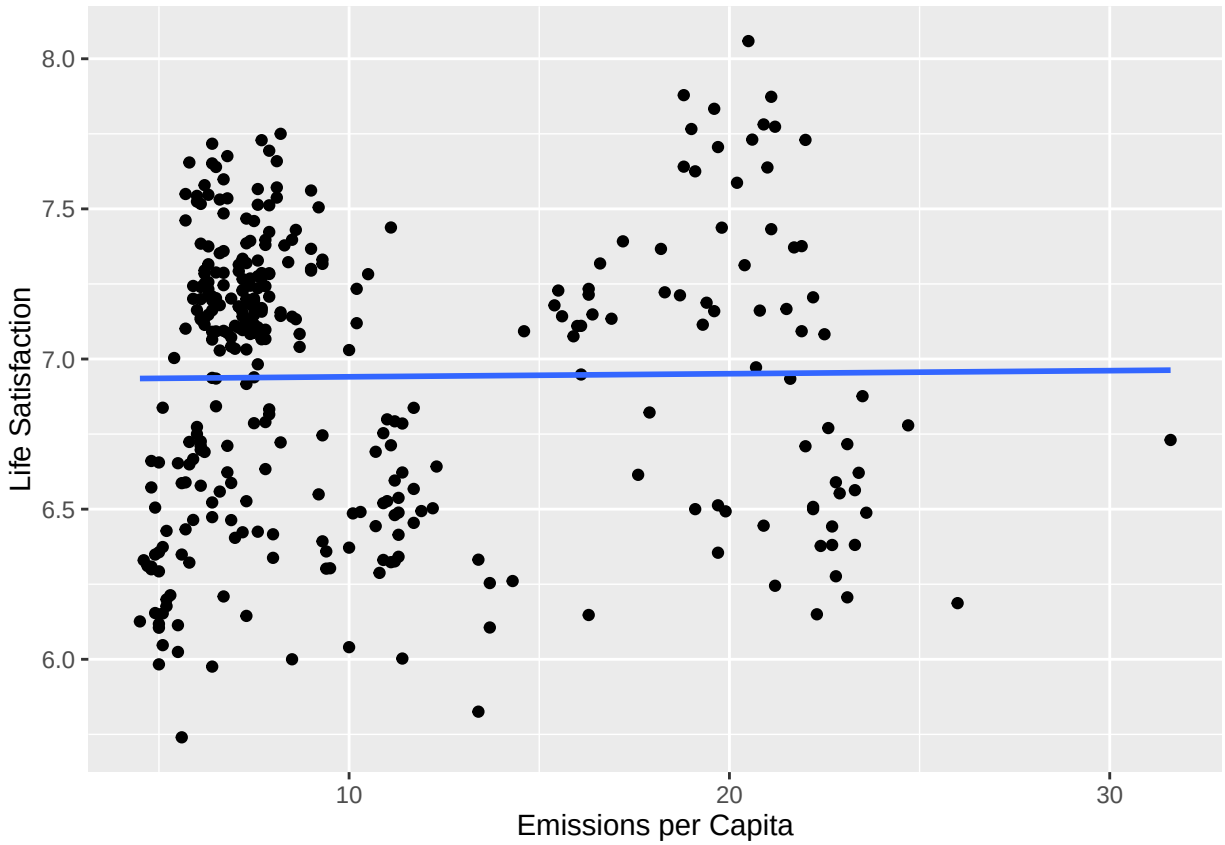


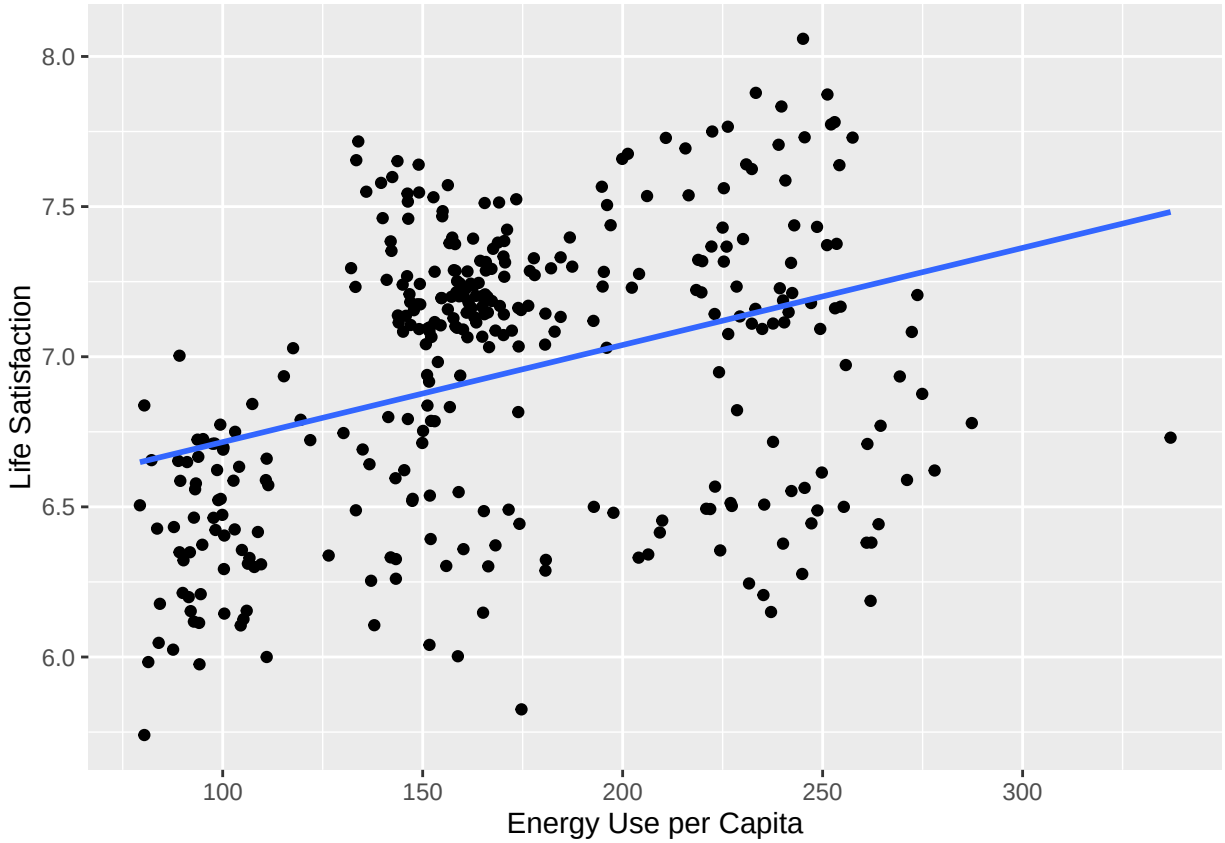


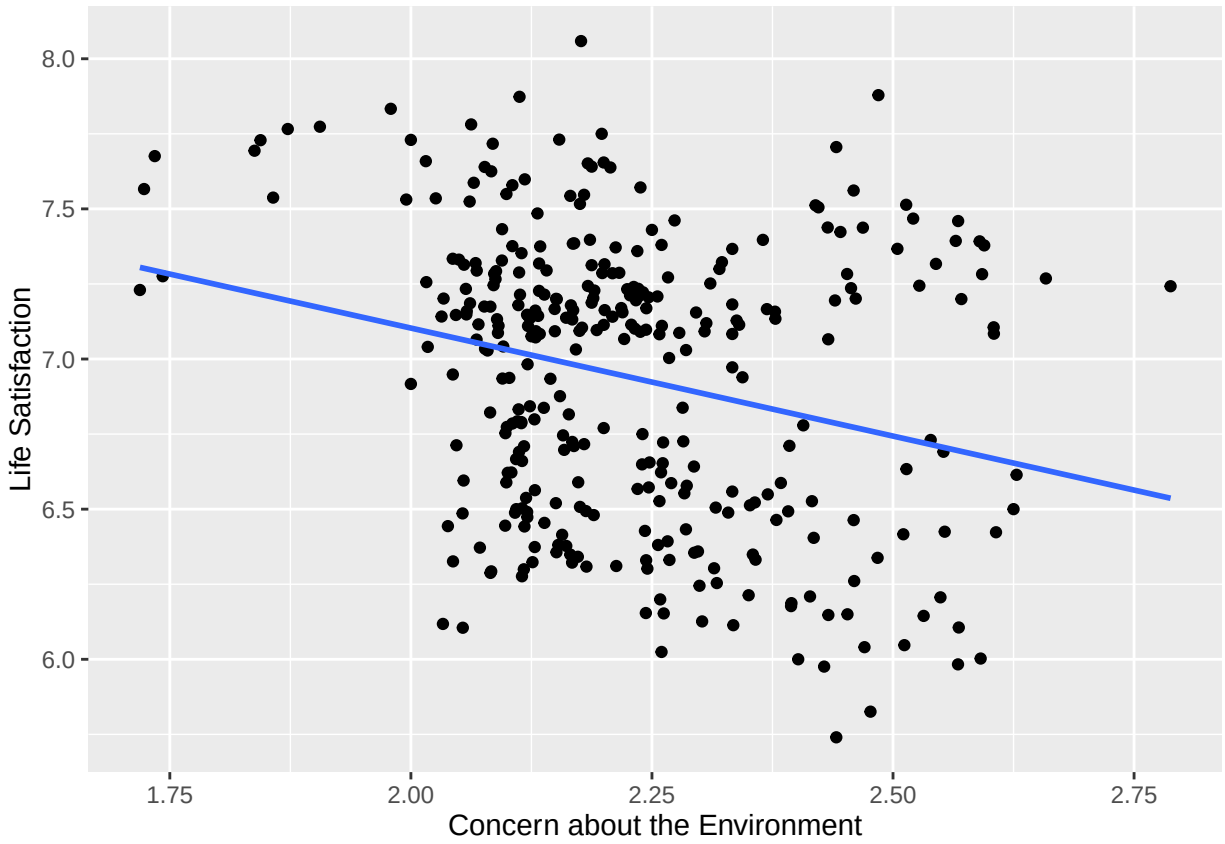
Again using the means of each state, we can create some scatterplots showing correlation between variables.

Life satisfaction appears to be positively correlated with age, energy use, and concern about the environment, but does not show a strong correlation with emissions.




8.0
7.5
7.0
6.5
6.0
Life Satisfaction
10
20
30
Emissions per Capita


8.0
7.5
7.0
6.5
6.0
Life Satisfaction
100
150
200
250
300
Energy Use per Capita



Inferential Statistics

To run inferential statistics, we first create "pdataind" as panel data. For the models below, X1 represents the explanatory variables of interest: emissions, energy use, concern for the environment, gender, and age. Emissions has a wide range: from 4.5 up to 31.6 annual tons of CO2 per capita. Observations in the upper range may be outliers that need to be removed; half of the observations fall between 6.5 and 13.7 tons, and the median is 7.6 compared to a mean of 10.2 tons. Energy use also varies widely, from 79.3 to 337.0 annual gigajoules (GJ) per capita, with a median of 164.0 and mean of 172.3 GJ. Concern for the environment, measured on a scale from 1 (very) to 3 (not very), has an average of 1.8. Gender is almost evenly split between males and females, and the average age is about 50 years old.

```
##    Emissions          Use          environ         gender          age       
##  Min.   : 4.5   Min.   : 79   Min.   :1.0   Min.   :1.0   Min.   : 17  
##  1st Qu.: 6.5   1st Qu.:149   1st Qu.:2.0   1st Qu.:1.0   1st Qu.: 38  
##  Median : 7.6   Median :164   Median :2.0   Median :2.0   Median : 51  
##  Mean   :10.2   Mean   :172   Mean   :2.2   Mean   :1.5   Mean   : 50  
##  3rd Qu.:13.7   3rd Qu.:211   3rd Qu.:3.0   3rd Qu.:2.0   3rd Qu.: 64  
##  Max.   :31.6   Max.   :337   Max.   :3.0   Max.   :2.0   Max.   :102
```

Y1 represents "satis" (life satisfaction), measured on a scale ranging from 0 (low) to 10 (high). The data is skewed toward the upper range, with the majority of responses being more than 6 and with a mean of 6.9.

```
##      satis      
##  Min.   : 0.0  
##  1st Qu.: 6.0
```

```
##  Median : 7.0
##  Mean   : 6.9
##  3rd Qu.: 8.0
##  Max.   :10.0
```

Below, we tested various types of panel data models on our individual-level data. The data set is unbalanced because we don't have information for all years in all the states.

```
## Oneway (individual) effect Pooling Model
##
## Call:
## plm(formula = Y1 ~ X1, data = pdataind, model = "pooling")
##
## Unbalanced Panel: n=22285, T=1-23, N=140830
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
##  -7.700  -1.110   0.342   1.110   4.020
##
## Coefficients :
##                Estimate  Std. Error  t-value Pr(>|t|)
## (Intercept)  6.43347006  0.03328325 193.2945   <2e-16 ***
## X1Emissions -0.09340835  0.00167109 -55.8968   <2e-16 ***
## X1Use        0.01104070  0.00018611  59.3239   <2e-16 ***
## X1environ   -0.06628195  0.00769927  -8.6089   <2e-16 ***
## X1gender    -0.01355593  0.00958703  -1.4140   0.1574
## X1age       -0.00534866  0.00027374 -19.5394   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    465260
## Residual Sum of Squares: 452070
## R-Squared:      0.028347
## Adj. R-Squared: 0.028346
## F-statistic: 821.692 on 5 and 140824 DF, p-value: < 2.22e-16
```

```
## Oneway (individual) effect Between Model
##
## Call:
## plm(formula = Y1 ~ X1, data = pdataind, model = "between")
##
## Unbalanced Panel: n=22285, T=1-23, N=140830
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -7.3400 -0.5390  0.0669  0.7110  3.8700
##
## Coefficients :
##                Estimate  Std. Error  t-value  Pr(>|t|)
## (Intercept)  6.58648647  0.07294836  90.2897 < 2.2e-16 ***
## X1Emissions -0.09277665  0.00282401 -32.8528 < 2.2e-16 ***
## X1Use        0.01029657  0.00030811  33.4187 < 2.2e-16 ***
## X1environ   -0.12037222  0.01892220  -6.3614 2.038e-10 ***
```

```
## X1gender     0.02326979  0.02441911   0.9529    0.3406
## X1age       -0.00426204  0.00066807  -6.3796 1.810e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    31990
## Residual Sum of Squares: 30206
## R-Squared:      0.05576
## Adj. R-Squared: 0.055745
## F-statistic: 263.128 on 5 and 22279 DF, p-value: < 2.22e-16

## Oneway (individual) effect First-Difference Model
##
## Call:
## plm(formula = Y1 ~ X1, data = pdataind, model = "fd")
##
## Unbalanced Panel: n=22285, T=1-23, N=140830
##
## Residuals :
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -10.30000  -1.77000   0.00251   1.75000  10.20000
##
## Coefficients :
##                Estimate  Std. Error  t-value  Pr(>|t|)
## (intercept) -0.00558269  0.00728448  -0.7664   0.44345
## X1Emissions -0.07466170  0.03434678  -2.1738   0.02973 *
## X1Use        0.01404905  0.00324545   4.3288 1.500e-05 ***
## X1environ   -0.05554390  0.00851167  -6.5256 6.799e-11 ***
## X1gender    -0.02587608  0.01032822  -2.5054   0.01223 *
## X1age       -0.00586130  0.00030099 -19.4736 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    748530
## Residual Sum of Squares: 745660
## R-Squared:      0.0038401
## Adj. R-Squared: 0.0038399
## F-statistic: 91.391 on 5 and 118539 DF, p-value: < 2.22e-16

## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = Y1 ~ X1, data = pdataind, model = "within")
##
## Unbalanced Panel: n=22285, T=1-23, N=140830
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
##  -7.880  -0.903   0.128   1.080   5.170
##
## Coefficients :
##                Estimate  Std. Error  t-value  Pr(>|t|)
## X1Emissions -0.08841148  0.01626967  -5.4341 5.517e-08 ***
## X1Use        0.01276929  0.00152923   8.3501 < 2.2e-16 ***
```

```
## X1environ   -0.05866414  0.00851204  -6.8919 5.533e-12 ***
## X1gender    -0.01951174  0.01032560  -1.8896   0.05881 .
## X1age       -0.00587612  0.00030079 -19.5359 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    374420
## Residual Sum of Squares: 372810
## R-Squared:      0.0043051
## Adj. R-Squared: 0.0036237
## F-statistic: 102.507 on 5 and 118540 DF, p-value: < 2.22e-16
```

In terms of explanatory power, the first difference and within estimators might be less useful than the between or pooled OLS estimators, based on their low R-squared and adjusted R-squared Values. Interestingly, gender does not appear to have a significant effect on reported life satisfaction in any of the models. The other variables are statistically significant in all the models. For variables other than gender, the directions of the relationships are also the same across models: emissions and age have negative coefficients, while concern about the environment and energy use have positive coefficients.

We can compare the fixed effects/within and pooled OLS models using an F-test for individual and/or time effects:

```
##
##  F test for individual effects
##
## data:  Y1 ~ X1
## F = 1.131, df1 = 22284, df2 = 118540, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

The test is set up so that the null hypothesis is that OLS pooled is better than the within estimator. The small resulting p-value indicates that despite the larger R-squared values with OLS pooling, we should reject the null hypothesis that OLS pooling is better.

Multilevel Analysis: Basic Steps

Multilevel Coefficient Models (MCM, *multilevel* and *nlme* packages) appear as one of the options to proceed with the multilevel analysis. MCM examines individual variation within and across groups, as well as permits non-independence of individual factors from the group level variables. The first step of the MCM investigates whether statistically significant variation between Bundeslaender in terms of mean individual satisfaction exists. Otherwise, simpler OLS and panel-data models are more applicable for the given research.

An unconditional model, *Null.Model*, serves the first purpose: the model only controls for the State that specifies that the variation intercept is a function of residence.

```
## Stateid = pdLogChol(1)
##             Variance  StdDev
## (Intercept) 0.1909161 0.4369395
## Residual    3.1680043 1.7798888
```

The model examines how much of the average individual life satisfaction is explained by the residence of a respondent through R's general purpose optimization routine (opt="optim"). According to the *Null.Model*, the Bundesland variation (intercept variance) is 0.19, while the within-State residual is 3.17.

```
## [1] "ICC"     "Group"   "GrpSize" "MeanRel"
```

```
## [1] 0.9958663
```

```
## 'log Lik.' 567965.4 (df=2)
```

```
## 'log Lik.' 562146.4 (df=3)
```

Furthermore, the *GmeanRel* function (*multilevel* package) yields the mean reliability of 16 Bundeslaender, which, in this case, is substantially high (0.996). As the cut-off point for acceptable reliability is 0.7, Bindeslaeander group reliability meets the threshold. The difference between the -2 log likelihood values of *Null.Model* and *Null.Model.2* tests whether the between-State effect is present compared to the random variation in life satisfaction without any control variables. According to the results, the difference is significantly large based on the Chi-Squared distribution with 2 degrees of freedom (5819.047). These results suggest that there is significant State variation in happiness level, which justifies the usage of the MCM.

The second step of the MCM looks into how both Bundesland emissions and subjective concerns about the environment influence reported individual life satisfaction. *Model.1* (*lme* function) for the time being does not control for individual characteristics, such as gender, age, employment and family status. Moreover, percentage of the renewables (which is anticipated to directly relate to the life satisfaction) will be also examined in the later phase of the research. *Model.1* serves as the basic tool to understand relationship between the group- and individual-level factors representing the environment.

```
## Linear mixed-effects model fit by REML
##  Data: finaldata
##        AIC      BIC    logLik
##   611544.6 611593.9 -305767.3
##
## Random effects:
##  Formula: ~1 | Stateid
##         (Intercept) Residual
## StdDev:   0.2752213 2.121016
##
## Fixed effects: satis ~ environ + Emissions
##                 Value  Std.Error     DF   t-value p-value
## (Intercept)  8.550986 0.07522865 140812 113.66662  0.0000
## environ     -0.044112 0.00909175 140812  -4.85189  0.0000
## Emissions   -0.000055 0.00027583 140812  -0.19851  0.8426
##  Correlation:
##           (Intr) envirn
## environ   -0.266
## Emissions -0.284  0.003
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -3.7759454 -0.5416871  0.2151472  0.6835521  1.4327174
##
## Number of Observations: 140830
## Number of Groups: 16
```

The output demonstrates that the State emissions are indeed negatively related to the individual happiness. However, after controlling for the individual environmental concerns, the coefficient (-0.005) is not statistically significant. Consequently, the group-level effect does not significantly differ from the individual-level. On

the other hand, the coefficient of environmental concerns statistically differs from 0, but the direction is oddly positive counter the expectations. These results point at the omitted variables (absence of the demographic characteristics) that should be included in the further model. Similarly, the surprising direction of the environmental coefficient demands investigation of endogeneity between reported happiness and environmental concerns.

A brief overview of *Model.2*, which covers demographic characteristics, reassures the anticipation about the State emissions: the coefficient became significant after controlling for other individual factors. As a result, the difference between State- and individual-level slopes exists and permit multilevel analysis. However, the coefficient of environmental concerns is still positive, which is to be discussed together with other factors in the final paper.

```
## Linear mixed-effects model fit by REML
##  Data: finaldata
##        AIC      BIC    logLik
##   609547.6 609636.3 -304764.8
##
## Random effects:
##  Formula: ~1 | Stateid
##         (Intercept) Residual
## StdDev:   0.2799221 2.105731
##
## Fixed effects: satis ~ environ + Emissions + age + fam + gender + emp
##                 Value  Std.Error     DF   t-value p-value
## (Intercept)  8.599703 0.07863972 140808 109.35572  0.0000
## environ     -0.050416 0.00905164 140808  -5.56977  0.0000
## Emissions   -0.000356 0.00027410 140808  -1.29811  0.1943
## age         -0.007657 0.00037522 140808 -20.40679  0.0000
## fam2         0.201748 0.01242550 140808  16.23662  0.0000
## gender2     -0.030859 0.01137446 140808  -2.71303  0.0067
## emp2         0.310966 0.01279972 140808  24.29476  0.0000
##  Correlation:
##           (Intr) environ Emissns age    fam2   gendr2
## environ   -0.252
## Emissions -0.276  0.004
## age       -0.205  0.029  0.033
## fam2      -0.021 -0.026 -0.012 -0.316
## gender2   -0.074 -0.067  0.000  0.004  0.043
## emp2      -0.163  0.003  0.006  0.456 -0.182  0.119
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -3.8912559 -0.4798106  0.2541150  0.7047883  1.6879033
##
## Number of Observations: 140830
## Number of Groups: 16
```