# Project 1: Linear Regression — Part 1+2

## MASM22/FMSN30/FMSN40: Linear and Logistic Regression (with Data Gathering), 2024

Peer assessment version: **12.30 on Wednesday 24 April**

Peer assessment comments: **13.00 on Thursday 25 April**

Final version: **17.00 on Friday 26 April**

## Introduction

We want to model the yearly emissions of atmospheric particles with a diameter between 2.5 and 10 µm, $\text{PM}_{10}$, per capita in Swedens 290 municipalities using data from Statistics Sweden (Statistiska Centralbyrån), `www.scb.se`. All data is from 2021 and located in `kommuner.xlsx`. See Canvas for maps of the locations of municipalities and counties.

| Variable namn | Description |
|---|---|
| `Kommun` | Municipality number and name |
| `County` | County (Län) number and name |
| `Part` | Part of Sweden: 1 = Götaland; 2 = Svealand; 3 = Norrland |
| `Coastal` | Coastal = any sea area within its borders: 0 = Inland; 1 = Coastal |
| `Vehicles` | Number of passenger cars, busses and trucks / 1000 inhabitants |
| `Builton` | Area covered in buldings, roads, etc (= not nature), (hectares / 1000 inhabitants) |
| `Children` | 0–14 year olds (percentage) |
| `Seniors` | 65+ year olds (percentage) |
| `Higheds` | At least 3 years of post-secondary (eftergymnasial) education (percentage) |
| `Income` | Median yearly income (1000 SEK) |
| `BRP` | Gross Regional Product per capita (1000 SEK) |
| `PM10` | The yearly emission of $\text{PM}_{10}$-particles (metric tonnes / 1000 inhabitants) |

Our goal is to model how the yearly emission of $\text{PM}_{10}$ particles varies as a function of one or several of the other variables, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$. We will use a linear regression model $Y_i = \mathbf{x}_i\beta + \varepsilon_i$ where the random errors $\varepsilon_i$ are assumed to be pairwise independent and $N(0,\ \sigma^2)$. In order to fulfill these model assumptions we will have to use suitable transformations of both the emissions and some of the other variables.

# Part 1. Simple linear regression

Since traffic is a large contributor to $\mathrm{PM}_{10}$ particles, we start with the most obvious explanatory variable, the number of vehicles per 1000 inhabitants.

1(a). Motivate, with the help of suitable residual plots, why we should take the logarithm of the emissions:

$$\ln(\texttt{PM10}) = \beta_0 + \beta_1 x + \varepsilon.$$

Also try to determine whether we should use $x = \texttt{Vehicles}$ or $x = \ln(\texttt{Vehicles})$.

1(b). Use the model with $x = \ln(\texttt{Vehicles})$ (*Model.1(b)*) and present the $\beta$-estimates, with 95 % confidence intervals, and plot $\ln(\texttt{PM10})$ against $\ln(\texttt{Vehicles})$ together with this estimated linear relationship, its 95 % confidence interval and a 95 % prediction interval for future observations.

Then transform the relationship back to $\texttt{PM10} = \ldots$, and plot `PM10` against `Vehicles` together with the estimated relationship, its 95 % confidence interval and a 95 % prediction interval.

Comment on any problems with the model fit and how this is reflected in the behaviour of the residuals.

1(c). Express how, according to *Model.1(b)*, the expected emission of `PM10` particles would change in a municipality if the number of vehicles would decrease by 10 %. Also calculate a 95 % confidence interval for this change rate.

Also calculate, with 95 % confidence interval, how much a municipality would have to reduce the number of cars in order to half its $\mathrm{PM}_{10}$ emissions.

# Part 2. Adding more explanatory variables

2(a). Test if there is a significant linear relationship between the log-$\mathrm{PM}_{10}$ emissions and the log-vehicles, according to *Model 1(b)*. Report the type of test you use, the null hypothesis, the value of the test statistics and its distribution when the null hypothesis is true (including the degrees of freedom), the P-value of the test and the conclusion.

2(b). Turn the two categorical variables `Part` and `Coastal` into factor variables (use the label ”No” for inland and ”Yes” for coastal municipalities) and present the number of observations in each of the 6 possible combinations.

Add both variables and their interaction to *Model 1(b)*. Present the new model, *Model 2(b)*, the $\beta$-estimates and their confidence intervals. What is the reference category here? Is this a suitable reference considering the number of observations?

Test if any of the added $\beta$-parameters are significantly different from zero. Report the type of test you use, the null hypothesis, the value of the test statistics and its distribution when the null hypothesis is true (including the degrees of freedom), the P-value of the test and the conclusion.

Also test if the interaction is significant. Report the type of test you use, the null hypothesis, the value of the test statistics and its distribution when the null hypothesis is true (including the degrees of freedom), the P-value of the test and the conclusion.

Use the model to calculate 95 % confidence intervals for the expected log-$PM_{10}$ and $PM_{10}$ for each of the 6 part/coastal combinations when `Vehicles` = 1000 vehicles per 1000 inhabitants.

2(c). It might not be necessary to have all 6 part/coastal combinations. Fit a model using Coastal = ”Yes” as reference using `relevel(Coastal, "Yes")` and use the results from both versions as well as the confidence intervals for the expected values to find a suitable way to reduce the combinations. As an example, the following code will calculate a new variable with values 1 = Götaland or coastal Svea/Norrland, 2 = Inland-Svealand, 3 = Inland-Norrland.

```
kommuner |>
   mutate(NewParts =
      as.numeric(Part == "Götaland" | Coastal == "Yes") +
      2*as.numeric(Part == "Svealand" & Coastal == "No") +
      3*as.numeric(Part == "Norrland" & Coastal == "No"))
```

Modify as you see fit and turn it into a factor variable, fit a model with log-vehicles and this new category variable instead of `Parts*Coastal` and test if it is significantly different from *Model 2(b)*. Modify the categorisation until you have a model that has as few categories as possible while still not being significantly different from *Model 2(b)*. Call this *Model 2(c)*. Present the model, its $\beta$-estimates and their confidence intervals.

2(d). Now we turn to the other numerical variables. Plot log-$PM_{10}$ against each of them and determine which of these should also be log-transformed. Any variable where a relative change is more natural than an additive change, e.g. any economic variable, as well as non-negative positively skewed variables are likely candidates.

Then find the two (transformed) variables that are most strongly correlated with log-$PM_{10}$, in addition to log-vehicles, and fit a model for log-$PM_{10}$ as a linear function of these three explanatory variables. Present the $\beta$-estimates, their standard errors, confidence intervals and the P-values for their t-tests. Use a combination of plots, correlations and VIF-values to determine whether it is reasonable to use all three variables in the model.

Motivate which variable we should exclude, refit the model without it (*Model 2(d)*) and present the $\beta$-estimates and their standard errors, confidence intervals and P-values. Also calculate and comment on the new VIF-values.

Compare the standard error for the $\beta$-estimate for log-vehicles in the two models. Explain the difference in size between the two standard errors and relate it to the relevant VIF-values.

2(e). Fit a new model using the new categorisation from 2(c) and all the continuous variables (don't forget any log-transforms), *except* the one you excluded in 2(d). Examine the VIF values, explain why they are now GVIF values, and determine whether we might reasonably safely use all these variables in the same model or if there are problems.

Explain why the most problematic variable is problematic, e.g. any strong correlations with any other variables that we should (or should not) have expected to occur. Exclude the most problematic variable from the model and examine the new GVIFs. This model will be refered to as *Model 2(e)*.

## Part 3. Model selection

Coming soon. . .