PROJECT 1: LINEAR REGRESSION
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, 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 PM_{10}-particles (metric tonnes / 1000 inhabitants)

Our goal is to model how the yearly emission of 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 ε_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 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 β-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 PM_{10} emissions.

Part 2. Adding more explanatory variables

2(a). Test if there is a significant linear relationship between the log-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 β-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 β-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 β-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 β-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 β-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 β-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 validation and selection

3(a). **Leverage.** Calculate the leverage (hat values) from *Model 2(e)* and plot them against the linear predictor, adding suitable horizontal lines as visual reference. Identify the six municipalities with the highest leverage and highlight them in the plot.

Add their names to the plot using the geometry `geom_text()`:

```
geom_text(data = filter(...), aes(x = ..., y = ..., label = Kommun))
```

Determine what makes each of these municipalities have such high leverage.

Hint: Plot pairs of x-variables against each other, separately for each factor category, `facet_wrap(~NewParts)`.

3(b). **Cook's distance.** Calculate Cook's distance from *Model 2(e)* and plot them against the linear predictor, adding suitable horizontal lines as visual reference. Identify the six municipalities with the highest Cook's distance and highlight them in the plot. Are these the same municipalities that had high leverage?

Investigate the DFBETAS to find the β-parameter(s) that the municipalities with the highest Cook's distance had the largest influence on.

Then plot log-PM10 against the corresponding variable(s), highlighting the influential municipalities, and explain why they had a large influence on that parameter.

3(c). **Studentized residuals.** Calculate the studentized residuals, r_i^*, for *Model 2(e)* and plot them against the linear predictor, using suitable horizontal lines as visual guides. Highlight the six municipalities with the highest Cook's distance. Do they also have large residuals?

Also identify the municipalities were $|r_i^*| > 3$ but which did not have high Cook's distance.

Plot $\sqrt{|r_i^*|}$ against the linear predictor, adding suitable reference lines and comment on the result, e.g., does the variance appear to be constant?

According to Naturvårdsverket, domestic transports is the largest source of PM_{10} particles, contributing 40 % of Sweden's emissions. The vast majority of this is due to wear on tyres, road surfaces and brakes. Burning wood for heating individual houses contributes 13 % of the emissions while burning biomass for production of electricity and district heating only contributes 3 %. Agriculture contributes 8 % of the emissions. The second largest source of Sweden's PM_{10} emissions is industry, contributing 32 % of Sweden's emissions. Of this, 69 % comes from the the construction and demolition industry, mostly due to road construction, while 16 % comes from paper and pulp processing (see Table 1 for a list of Swedish paper mills and their locations).

A large part of the PM_{10} emissions from the rest of the industry is related to the use of fossil fuel in processes requiring high temperatures. These processes usually also result in high emissions of carbon dioxide, CO_2. In 2021, Swedish television, SVT, listed the 15 companies in Sweden that emitted the most carbon dioxide, during 2020 (see Table 2). Together they stood for 25 % of Sweden's CO_2 emissions. Note that the paper and pulp companies do not appear on this list, since they mainly use biomass as energy source.

In this project, we are more interested in the PM_{10} emissions per capita in general, and not so much if a small municipality happens to have a large PM_{10} emitting company within its borders, so we will clean up the data before we continue.

3(d). **Explain, exclude, refit.** Relate the large studentized residuals in the municipalities found in 3(c) to the information in Table 1 and 2. Then repeat the following steps:

(1) Identify the municipalities where $r_i^* > +3$ and try to identify the emission source[1].

(2) Remove the high-residual municipalities, **where the source can be identified**, from the data and refit the model on the reduced data.

- Repeat until no large residuals with identifiable emission sources are left.

List the municipalites that were identified in each step, the emission source (if identified), and whether the observation was excluded in the next step.

Present the β-estimates, with confidence intervals, for both the old version of *Model 2(e)* (full data set) and this new version, *Model 3(d)*, (reduced data set) and compare the results. Specifically, discuss which variables are/are not significant and how this might relate to the exclusion of some municipalities with large DFBETAS. Also compare how well the assumptions of normality and constant variance of the residuals are fulfilled.

3(e). **Variable selection.** Continue with the reduced data set and fit the null model using only an intercept, as well as a new version of *Model 1(b)* (only log-vehicles) and *Model 2(c)* (log-vehicles and NewParts).

Perform a stepwise selection using AIC as criterion, and another using BIC, both starting with *Model 1(b)* and using the null model as lower scope and *Model 3(d)* as upper scope. For each step, report which variable was included or excluded from the model.

Construct a table containing the number of β-parameters, the residual standard deviation, the R^2, adjusted R^2, AIC and BIC for the following six models: the null model, the refitted *Model 1(b)*, the refitted *Model 2(c)*, *Model 3(d)*, the AIC model and the BIC model. State which model you find best, and the reasons for your choice.

How much of the variability of the log-PM_{10} emissions can be explained by only the number of vehicles and how much is explained by the best model? Discuss whether the model make sense, e.g., if the variables included seem reasonable and if their β-parameters have the expected sign. Try to explain why you would expect the repationships to have the "expected" signs, including an interpretation of the NewParts parameters.

End of Project 1

[1] In addition to Table 1 and 2, also try Wikipedia. The municipality pages usually have a list of large companies. Some hints: The Swedish Air force wing F-7 Såtenäs is located in Lidköping and the Port of Gothenburg (= Göteborg) is the largest in Scandinavia.

Company	**Locations**
Ahlstrom	Askersund (Aspa$^{(1)}$), Bengtsfors (Billingsfors$^{(0)}$)
Billerud AB	Grums (Gruvön$^{(7)}$), Gävle$^{(7)}$, Lindesberg (Frövi$^{(5)}$), Kalix (Karlsborg$^{(3)}$), Norrköping (Skärblacka$^{(3)}$)
Holmen Paper AB	Hudiksvall (Iggesund$^{(3)}$)
Metsä Board Sverige AB	Örnsköldsvik (Husum$^{(7)}$)
Mondi Dynäs	Kramfors$^{(1)}$
Nordic Paper	Kristinehamn (Bäckhammar$^{(1)}$), Säffle$^{(0)}$
Rottneros	Sunne (Rottneros$^{(1)}$), Söderhamn (Vallviksbruk$^{(1)}$)
SCA	Timrå (Östrand$^{(7)}$), Piteå (Munksund$^{(3)}$), Umeå (Obbola$^{(3)}$)
Smurfit Kappa	Piteå$^{(5)}$
Stora Enso	Hammarö (Skoghall$^{(7)}$), Älvkarleby (Skutskär$^{(5)}$), Bromölla (Nymölla$^{(3)}$)
Södra Cell	Mönsterås$^{(7)}$, Varberg (Värö$^{(7)}$), Karlshamn (Mörrum$^{(5)}$)

Capacity: (0) = 0–100 000; (1) = 100 000–300 000; (3) = 300 000–500 000; (5) = 500 000–700 000; (7) = more than 700 000

Source: www-skogen.se/skogssverige/papper/massa-...

Table 1: Paper mills in Sweden

	Company	CO_2	**Description and locations**
1	SSAB	4 777 065	Steel company. Production plants in Luleå (blast furnace, swe: *masugn*), Borlänge, Oxelösund (blast furnace) and Finspång.
2	Cementa AB / Heidelberg Materials	1 900 881	Building materials company with production in Gotland (Slite), and in Skövde.
3	Preem AB	1 538 189	Petroleum and bio-fuel company with oil refineries in Lysekil and Göteborg.
4	Luossavaara-Kiirunavaara AB (LKAB)	649 747	State owned mining company. Mines iron ore in Kiruna and Gällivare (Malmberget) with ore processing plants in Kiruna (Kiruna and Svappavaara) and Gällivare (Malmberget). Owns part of SSAB.
5	St1 Refinery AB	500 033	Oil refinery in Göteborg.
6	E.ON Värme Sverige AB	372 870	District heating company. District heating plants in Malmö and several other places.
7	Stockholm Exergi AB	361 050	District heating company with plants in Stockholm.
8	Borealis AB	345 364	Chemical company producing polyethylene in Stenungsund.
9	Boliden Mineral AB	279 876	Mining company. Mines zink and copper (and gold, silver and lead) in Hedemora (Garpenberg), Gällivare (Aitik) and Lycksele, Norsjö and Skellefteå, (Skelleftefältet).
10	SMA Mineral AB	278 960	Lime (swe: *kalk*) company. Several lime stone quarries. Production plants for quicklime in Rättvik (Boda), Söderhamn (Sandarne), Luleå, Oxelösund and Rättvik (also producing slaked lime).
11	Tekniska verken i Linköping AB (publ)	266 807	District heating/electricity company with plants in Linköping and neighbouring areas.
12	Kubikenborg Aluminium AB (Kubal)	245 204	Aluminium company with a smelting plant in Sundsvall.
13	SYSAV	243 707	Garbage/waste/recycling company for south Skåne with a waste-to-energy plant in Malmö.
14	Renova AB	206 356	Garbage/waste/recycling company for west Sweden with a waste-to-energy plant in Göteborg (Sävenäs).
15	Nordkalk AB	188 926	Lime company. Several lime stone quarries. Production plants for quicklime in Gotland (Storugns) and for slaked lime in Luleå and Landskrona, and a plant for both quicklime and slaked lime in Köping.

Source: SVT, www.svt.se/nyheter/inrikes/15-foretag-...

Table 2: The 15 largest carbon dioxide emitters in Sweden 2020