PROJECT 2: LOGISTIC REGRESSION
MASM22/FMSN30/FMSN30F/FMSN40: LINEAR AND LOGISTIC REGRESSION (WITH DATA GATHERING), 2026

Peer assessment version: **17.00 on Monday 11 May**
Peer assessment comments: **17.00 on Tuesday 12 May**
Final version: **17.00 on Wednesday 13 May**

Introduction

As we found out in Project 1, there were a few municipalities that had a large influence on the linear regression model, suggesting that we should not use them when fitting the model for the log-PM_{10} particle emissions.

In this project we will use all the original data, including the problematic ones, and instead model the probability of a municipality having an unusually high (in the top third) particle emission rate using binary logistic regression.

The data is still located in `Project1data.xlsx`.

Part 1. Introduction to logistic regression

Lec.7 1(a). Define a high emmision rate as a PM_{10}-value lying in the upper tertile (upper third) of the data. Create a new variable, `pm10_high`, with value 1 if PM10 is lies the upper tertile and 0 otherwise. *Hint:* use `quantile(PM10, probs = )` to find the cut-off value.

In order to make it easier to colour the observations according to the values of this response variable, as well as for model validation later, it will be convenient to also create a separate factor version. Add both new variables to the data set. You can now plot with, e.g, `aes(..., y = pm10_high)` and get 0 and 1 on the y-axis, while `aes(..., color = pm10_cat)` will give you different colours for `"High"` and `"Low"`. You can use either version as dependent variable in a logistic regression since R uses the last category (1 or "High") as "succeess".

Report the cut-off value and present a frequency table of the number of municipalities with Low and High emissions, respectively.

Lec.7 1(b). Turn `Part` into a factor variable (again), using "Götaland" as reference category.

Examine the relationship between a high emission level and whether the municipality is located in Götaland, Svealand or Norrland by counting the number of observations in each of the six combinations.

Use these raw numbers (not a regression model) to estimate the probability p and the corresponding odds, $p/(1-p)$, and log odds, of having a high emission level for each of the three Parts. Also calculate the odds ratios for Svealand and Norrland using Götaland as reference category, and the corresponding log odds ratios. Present the results in a table (Table 1(b)).

Describe how the odds of having a high emission rate changes when we change from Götaland to Norrland.

Lec.8 1(c). Now fit a logistic regression model, *Model.1(c)*, with `Part` as explanatory variable. Present a table with the β-estimates, their standard errors, their 95 % profile likelihood confidence intervals, the corresponding e^{β}, and their 95 % confidence intervals.

Identify the odds and odds ratios in Table 1(b) that are connected to the different e^{β} from the model.

Use *Model.1(c)* to estimate the log-odds of high emissions for Götaland, Svealand and Norrland, together with their standard errors and 95 % confidence intervals. Also calculate the corresponding probabilities and 95 % confidence intervals.

Use a suitable test to determine whether there are any significant differences in the probability of a high level of emissions between the three parts ot Sweden. Report the type of test you use, the null hypothesis, the value of the test statistic, its distribution, the P-value and the conclusion.

Lec.8 1(d). Ignore `Part` and turn back to the fuel consumption instead.

Plot the 0/1 variable `pm10_high` against `Fuel` and add a moving average with `geom_smooth()`. Also plot a version using log-Fuel instead.

Comment on the two plots. Should we still use the log-transfomed Fuel, as in Project 1, and does it seem reasonable to use (log) Fuel consumption as an explanatory variable?

Fit a simple logistic regression, *Model.1(d)*, using `log(Fuel)` as explanatory variable. Report the β-estimates with 95 % confidence intervals, as well as the e^{β}-estimates and their confidence intervals.

Add the estimated probability of a high emission rate, and its confidence interval, to the plot.

Use a suitable test to determine if there is a significant relationsship between a high emission rate and the fuel consumption in a municipality. State the null hypothesis H_0, what type of test you use , and why you choose that type, the value of the test statistic, the distribution of the test statistic when H_0 is true, the P-value and the conclusion.

How does the Fuel consumption change when the log-fuel consumption increases by one unit? How does the odds of having a high emission rate change when the log-fuel consumption increases by 1 unit, according to the model? How does the odds change when the fuel consumption increases by 10 %? Also present 95 % confidence intervals for these changes.

Lec.8 1(e). Calculate the leverage values for *Model.1(d)* and plot them against log-`Fuel`, against the linear predictor $x_i\hat{\beta}$, and against the predicted probabilities $\hat{p}_i$. Add horizontal reference lines at the minimal value $1/n$ and at $2(p+1)/n$ and make sure the y-axes includes zero.

Relate the general behaviour of the leverage to the behaviour of the estimated probabilities. Why are the two "bumps" in the leverage located where they are? *Hint*: Where does the S-curve change its slope?

Lec.8 1(f). Calculate McFadden's adjusted pseudo R^2, AIC and BIC for *Model.1(c)* and *Model.1(d)* and decide whether part of Sweden or fuel consumption seems more important for explaining the differences in the probability of having a high emission rate in a municipality.

Part 2. Variable selection and influential observations

Lec.8 2(a). We will now use stepwise variable selection in order to find a suitable set of variables for explaing the probability of a high emission rate.

Start by fitting a full logistic regression model (*Model.full*) with all the explanatory variables we used in the full model in Project 1, including the interactions with `Part`. Do not use the high VIF variable identified in Project 1 or `Coastal`. Don't forget to log-transform all variables you log-transformed in Project 1.

Perform two stepwise selections, one using AIC (*Model.AIC*) and one using BIC (*Model.BIC*) as criterion, starting with *Model 1(c)* (using `Part` as explanatory variable), with the null model as the smallest model allowed and the full model as the largest model allowed.

For each of the two models, report which variable is included or excluded in each step. For the final models, present a table with the β-estimates, as well as e^{β}-estimates, with confidence intervals.

Lec.8 2(b). If the two models, AIC and BIC, are nested, perform a suitable test for whether any of the additional variables in the larger model are significant. Report the null hypothesis, the type of test, the test statistic, its distribution when H_0 is true, the P-value and the conclusion.

For all three models, the full model, the AIC model, and the BIC model, report Fadden's adjusted R^2, AIC and BIC and motivate which of the models seems best. Call this *Model.2(b)*.

Lec.8 2(c). Plot the leverage for *Model 2(b)* against, e.g., the linear predictor, using suitable horizontal lines for reference, and identify any municipalities with an unusually high leverage.

Plot Cook's distance for *Model 2(b)* against the linear predictor with suitable reference lines and identify any municipality with a worryingly high Cook's D.

Use the DFBETAS to identify which parameters were most affected by the observation with the highest Cook's D, and produce suitable plots that explain why this observation had high DFBETAS for these parameters.

Lec.8 2(d). Plot the standardised deviance residuals for *Model 2(b)* against the linear predictor, with colour coding (low or high number of cars) and suitable reference lines, and highlight the observation with the highest Cook's D you identified before. Also identify any observations with a large deviance residual, $|d_i| > 3$. Do these observations also have high leverage and/or Cook's D?

Lec.8 2(e). Is there anything in the results in 2(c) or 2(d) that would make you regret your choice of best model in 2(b)?

Lec.8 2(f). Compare the coefficients for Svealand and Norrland in *Model.1(c)* and *Model.AIC*. Explain why they are so different.

Hint: look at the signs of the β-coefficients for the additional variables in the AIC model and the distributions, e.g.
`ggplot(df, aes(x = Part, y = Fuel)) + geom_boxplot() + scale_y_log10()`.

Part 3. Goodness-of-fit

Lec.9 3(a). Use the threshold value 0.5, classifying observations with $\hat{p}_i \leq 0.5$ as "should have low emissions", and observations with $\hat{p}_i > 0.5$ as "should have high emissions", for *Model.null*, *Model 1(c)* and *Model 1(d)*, as well as *Model.BIC*, *Model.AIC* and *Model.full* from 2(b).

Present the resulting confusion matrices as well as a table, *Table 3(a)*, collecting the Accuracy, the P-value for Acc > NIR, Cohen's $\varkappa$, the P-value for McNemar's test, Sensitivity and Specificity for all six models.

State which of the models are significantly better than always predicting that the emissions will be low. Also state which of the models are predicting significanty incorrect proportions of low and high emissions, and in which direction they err.

Lec.9 3(b). Plot the ROC-curves for all six models in the same plot, and present a table with their AUC-values, including 95 % confidence intervals.

Perform pair-wise tests comparing the AUC for the "best" model, *Model.2(b)*, against each of the other models and discuss the result. Does it agree with the conclusion in 2(b)?

Note: these tests are not independent but we perform them here as a crude way of determining whether the performance of the models are significantly different.

Lec.9 3(c). For each of the five models (not the null model), find the optimal threshold for $\hat{p}_i$, where the distance to the ideal model is minimized. Use these new thresholds to calculate new confusion matrices and a new version of *Table 3(a)*, with the optimal thresholds added, *Table 3(c)*.

Comment on any interesting differences between the conclusions that can be drawn from the two tables, 3(a) and 3(c).

Do the conclusions confirm or contradict your decision of which model is "best"?

Lec.9 3(d). Taking all the results into account, select the model you would prefer as the overall "best" model. Describe the reasons behind your decision.

Comment on the variables included and the possible reasons for why they would have that positive/negative effect on the emission rates in a municipality.

End of Project 2