PROJECT 2: LOGISTIC REGRESSION
MASM22/FMSN30/FMSN40: LINEAR AND LOGISTIC REGRESSION (WITH DATA GATHERING), 2024

Peer assessment version: **12.30 on Monday 13 May**
Peer assessment comments: **13.00 on Tuesday 14 May**
Final version: **17.00 on Wednesday 15 May**

Introduction

According to Project 1, the number of vehicles (cars, buses and trucks) in a municipality explained a large proportion of the variability in PM_{10} emissions between the municipalities. We will now focus on personal cars (not buses and trucks) and model how the probability of a municipality having a high number of cars per capita varies as a function of one or several variables, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$.

The data is located in `kommunerProject2.xlsx`. Note that some variables from Project 1 are no longer present and we have some other variables instead.

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
`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)
`GRP`	Gross Regional Product per capita (1000 SEK)
`Cars`	Number of passenger cars / 1000 inhabitants
`Urban`	Proportion of the population that lives in urban areas (percentage)
`Transit`	Proportion of the population that lives within 400 meters of a bus stop or other public transportation (percentage).
`Apartments`	Proportion of the apartments that are in apartment buildings (swe: *flerfamiljshus*), i.e., not in single family homes (swe: *småhus*) (percentage).
`Fertility`	Total fertility rate. Number of births per woman.
`Persperhh`	Average number of persons per household.

Note that `Fertility` will be a character variable. We will deal with that later.

Part 1. Introduction to logistic regression

1(a). **High number of cars without regression**:

Define a high number of cars as more than 600 cars per 1000 inhabitants and create a new variable, `highcars`, with value 1 if the number of cars is larger than 600 and 0 otherwise, and add it to the data set:

```
kommuner |> mutate(highcars = as.numeric(Cars > 600)) -> kommuner
```

In order to make it easier to colour the observations according to the values of this response variable, it will be convenient to also create a separate factor version:

```
kommuner <- mutate(kommuner,
                   highcars_cat = factor(highcars,
                                         levels = c(0, 1),
                                         labels = c("low, "high")))
```

You can then plot with, e.g, `aes(..., y = highcars)` and get 0 and 1 on the y-axis, while `aes(..., color = highcars_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".

Turn `Part` into a factor variable, using "Götaland" as reference category, and examine the relationship between a high number of cars and whether the municipality is located in Götaland, Svealand or Norrland by counting the number of observations in each of the six combinations:

```
kommuner |> count(Part, highcars_cat)
```

Use these numbers to estimate the probability p and the corresponding odds, $p/(1-p)$, and log odds, of having a high number of cars 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.

How does the odds of having a high number of cars change when we change from Götaland to Norrland?

1(b). **High number of cars with regression**:

Fit a logistic regression model, *Model.1(b)*, 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(a) that are connected to the different e^{β} from the model.

Use *Model.1(b)* to estimate the log-odds of high number of cars, 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 large number of cars 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.

1(c). **Access to a bus stop**: Some parts of Sweden are more densely populated and have a more extensive public transportation network which might explain some of the differences in the number of cars. We will now look at the proportion of the inhabitants that have less than 400 meters to a bus stop, `Transit`, instead of which part of Sweden the municipality is located in.

Plot the 0/1 variable `highcars` against `Transit` and add a moving average with `geom_smooth()`. Does it seem reasonable to use the proportion living close to a bus stop as an explanatory variable?

Fit a simple logistic regression, *Model.1(c)*, using `Transit` 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 number of cars, and its confidence interval, to the plot.

Use a suitable test to determine if there is a significant relationsship between a high number of cars and the proportion living close to a bus stop. 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 odds of having a high number of cars change when the proportion living close to a bus stop increases by 1 percentage unit, increases by 10 percentage units, decreases by 1 percentage unit or decreases by 10 percentage units?

1(d). **Leverage:** Calculate the leverage values for *Model.1(c)* and plot them against `Transit`. Add horizontal reference lines at the minimal value $1/n$ and at $2(p+1)/n$ and make sure the y-axis includes zero.

Relate the general behaviour of the leverage to the behaviour of the estimated probabilities. You may have to exclude the horizontal line at $2(p+1)/n$ in order to see the behaviour better. Why are the two "bumps" in the leverage located where they are? *Hint*: Where does the S-curve change its slope?

1(e). Calculate McFadden's adjusted pseudo R^2, AIC and BIC for *Model.1(b)* and *Model.1(c)* and decide whether part of Sweden or access to a bus stop seems more important for explaining the differences in the probability of having a high number of cars in a municipality.

Part 2. Variable selection and influential observations

2(a). **Imputation of missing data**.

Before we can start selecting variables we have to deal with `Fertility`. It is a character (text) variable since there are some missing data coded as "NA" in the Excel file. The reason for this is that SCB does not report the fertility if the number of women or births in a municipality is too small, in order to protect the identities of the persons involved.

Start by forcing the variable into a numerical variable:

```
kommuner |> mutate(Fertility = as.numeric(Fertility)) -> kommuner
```

Note the warning message.

Find out which municipalities are affected. The function `is.na()` tests if its input is NA (Not Available). Dorotea and Bjurholm are competing for the title of Sweden's smallest municipality. In the data set, Bjurholm was the smallest with 2392 inhabitants. The largest of the municipalities with NA Fertility is Överkalix with 3269 inhabitants[1].

We will want to use the fertility rate in our models but due to the missing values we will have problems comparing models using different variables. Our comparison measures require that we always use the same data set. We could remove the municipalities with missing fertility rates, but we will instead replace the missing values with the average fertility rate in Norrland Inland:

[1]The median population size among Sweden's municipalities is 16 000 inhabitants.

```
kommuner |> filter(Part == "Norrland" & Coastal == "No") |>
  summarise(meanfertility = mean(Fertility, na.rm = TRUE))
```

Create an index variable, `I`, with the row numbers for the municipalities and then replace their missing fertility rates by the mean fertility rate (report the value!):

```
I <- which(is.na(kommuner$Fertility))
kommuner$Fertility[I] <- ...
```

2(b). **Variable selection:**

We will now use stepwise variable selection in order to find a suitable set of variables for explaing the probability of a high number of cars.

Start by fitting a full logistic regression model (*Model.full*) with all 11 continous explanatory variables. Don't forget to log-transform Income and GRP plus any of the other variable you log-transformed in Project 1. It may also be a good idea to log-transform `Apartments` since it is skewed. Do not use `Part` or `Coastal`. Report the VIF-values for this model and comment on any issues with possible multicollinearity.

Ignore the issues and perform two stepwise selections, one using AIC (*Model.AIC*) and one using BIC (*Model.BIC*) as criterion, starting with the null model (only intercept), 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 the variables included in the final model and the corresponding VIF-values. Are there any worrying multicollinearity problems now?

If the two models 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 both models, also report Fadden's adjusted R^2, AIC and BIC and motivate which of the models seems best. Call this *Model.2(b)*.

2(c). **Influential observations:** Plot the leverage for *Model.2(b)* against the linear predictor, using suitable horizontal line for visual reference, and identify any municipalities with a worryingly 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 affected by the observations with the high Cook's D.

2(d). **Deviance residuals**: 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 any observations with high Cook's D you identified before. Identify any observations with a large deviance residual, $|d_i| > 3$. Plot the standardised deviance residuals against each of the x-variables in the model, with reference lines, and comment on any interesting patterns (or lack of patterns).

Part 3. Goodness-of-fit

3(a). **Confusion:** Use the threshold value 0.5, classifying observations with $\hat{p}_i \leq 0.5$ as "should have low number of cars", and observations with $\hat{p}_i > 0.5$ as "should have high number of cars", for *Model.null*, *Model 1(b)* and *Model 1(c)*, 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 five models.

State which of the models are significantly better than always predicting that the number of cars will be low, and which of the models are predicting the correct (or not incorrect) proportions of low and high numbers of cars.

3(b). **ROC-curves and AUC:** 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.

3(c). **Optimal thresholds:** For each of the five models (not the null model), find the optimal threshold for p, 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. Do the conclusions confirm or contradict your decision of which model is "best"?

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.

Present the model, the β-estimates and their corresponding 95 % confidence intervals. Comment on the variables and the possible reasons for why they would have that positive/negative effect on the number of cars in a municipality.

End of Project 2