PROJECT 2: LOGISTIC REGRESSION
MASM22/FMSN30/FMSN40: LINEAR AND LOGISTIC REGRESSION (WITH DATA GATHERING), 2023

Peer assessment version: **12.30 on Wednesday 10 May**
Peer assessment comments: **13.00 on Thursday 11 May**
Final version: **17.00 on Friday 12 May**

Low monthly precipitation

We will continue studying the weather data from Project 1, now with focus on low precipitation. Our goal is to model how the probability of low monthly precipitation varies as a function of one or several of the other variables, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$.

`location`	the location of the weather station (text): Lund, Uppsala or Katterjåkk.
`month`	a text variable of the format `"yyyy-mm"`
`rain`	total monthly precipitation (mm)
`pressure`	average monthly air pressure (hPa)
`temp`	average monthly temperature (°C)
`speed`	average monthly wind speed (m/s)

Part 1. Introduction to logistic regression

1(a). **Probability, odds and confidence intervals for probabilities:** We start by estimating a small probability to see what happens to the different types of confidence intervals for the probability when the sample size is too small for the estimates to be normally distributed. Lets define a dry month as a month where the precipitation is less than 0.5 mm. Create a new variable, `drymonth`, with value 1 if the precipitation is lower than 0.5 and 0 otherwise, and add it to the data set:

```
weather$drymonth - as.numeric(weather$rain < 0.5)
```

Find the total number of months n, the number of dry months Y, and estimate the probability p of a month being dry, the odds $p/(1-p)$, and the log odds. Also fit a logistic regression null model with `drymonth` as dependent variable using only an intercept. The estimate of β_0 should be identical to your log odds estimate.

If the number of dry (and not dry) months is (had been) large enough we would have the following approximate distributions:

$$\hat{p} = Y/n \sim N(,) \qquad \text{with standard error } d(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

$$\ln\frac{\hat{p}}{1-\hat{p}} = \ln\frac{Y}{n-Y} \sim N(,) \quad \text{with standard error } d(\ln\frac{\hat{p}}{1-\hat{p}}) = \sqrt{\frac{1}{Y} + \frac{1}{n-Y}}.$$

The test statistic for a likelihood ratio test for H_0: $\beta_0 = \vartheta_0$ would be approximately $\chi^2(1)$-distributed.

These three approximations have different requirements for what ”large enough” is.

- Use the normal approximation of $\hat{p}$ to construct a 95 % confidence interval for p.
- Use the normal approximation of the log odds estimate to construct a 95 % confidence interval for the log odds and transform it into an interval for p.
- Calculate the profile likelihood based 95 % confidence interval for the log odds using `confint(model)` and transform it into an interval for p.

An exact 95 % confidence interval[1] for p is given by $I_p = (0.00023,\ 0.0047)$.

Which of your three intervals is closest to this? If you cannot use the profile likelihood method (as is usual in a regression setting), which of the two normal approximation based methods, for $\hat{p}$ or for the log odds, should you use? Which of the methods should you *not* use, and why?

1(b). **Low rain by location without regression**: We now turn to a more frequent event and define low precipitation as a monthly precipitation that is smaller that the first quartile, Q_1, in our data, where Q_1 is defined by $Pr(\texttt{rain} < Q_1) = 0.25$. Use

```
Q1 <- quantile(weather$rain, 0.25)
```

Compare with the value you get from `summary(weather)` to double check.

Create a new variable, `lowrain`, with value 1 if the precipitation is lower than `Q1` and 0 otherwise, and add it to the data set:

```
weather$lowrain <- as.numeric(weather$rain < Q1)
```

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:

```
weather$low_cat <- factor(weather$lowrain,
                          levels = c(0, 1),
                          labels = c("high", "low"))
```

You can then plot with, e.g, `aes(..., y = lowrain)` and get 0 and 1 on the y-axis, while `aes(..., color = low_cat)` will give you different colours for `"high"` and `"low"`. You can use either version as dependent variable, since R uses the last category (1 or "low") as "succeess".

Start by examining the relationship between low precipitation and location by a simple cross-tabulation:

```
table(weather$location, weather$low_cat)
```

Use the values in the table to estimate the probabilities, $1 - p$ and p, and the corresponding odds, $p/(1 - p)$, of having low precipitation, for each of the locations, and add them as columns to the table. Also calculate the odds ratios for Lund and Uppsala, using Katterjåkk as reference category, and add them to the table as well.

[1] with Blaker's method based on Fisher's Exact Test

1(c). **Low rain by location with regression**: Fit a logistic regression model, *Model.1(c)*, with `location` as explanatory variable (don't forget to turn it into a factor variable with a suitable reference category). Present a table with the β-estimates, their standard errors, their 95 % 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 low precititaion, for each of the locations, together with their standard errors and 95 % confidence intervals. Also calculated the corresponding probabilities and 95 % confidence intervals.

Use a suitable test to determine whether there are any significant differences in the probability of low precipitation between the locations. 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(d). **Low rain and air pressure**: We will now look at air pressure instead of location. Plot the 0/1 variable `lowrain` against `pressure` and add a moving average with `geom_smooth()`. Does it seem reasonable to use air pressure as an explanatory variable?

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

Use a suitable test to determine if there is a significant relationsship between low precipitation and air pressure. 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 low precipitation change when the air pressure increases by 1 hPa? Increases by 5 hPa? Decreases by 1 hPa? Decreases by 5 hPa?

Add the estimated probability of low precipitation, and its confidence interval, to the plot.

1(e). **Leverage:** Calculate the leverage values for *Model.1(d)* and plot them against air pressure. 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. Why are the two "bumps" in the leverage located where they are?

1(f). Calculate McFadden's adjusted pseudo R^2, AIC and BIC for *Model.1(c)* and *Model.1(d)* and decide whether location or air pressure seems more important for the probability of having a low precipitation.

Part 2. Influential observations and variable selection

2(a). **Variable selection:** We will now, as in Project 1: Part 3(a), use stepwise variable selection in order to find a suitable set of variables for explaing the probability of low rain.

Start by fitting the full logistic regression model with all explanatory variables and their interactions: `pressure*location*speed*temp*monthnr`, where `monthnr` is the factor variable taking the values "01", "02", . . . , "12".

Are there any problems here? Why might that be?

Ignore the problems (let's hope we don't need the full model) and perform two stepwise selections, both using BIC as criterion, with the null model as the smallest model allowed and the full model as the largest model allowed. For the first model, *Model.2(a).Fw*, start with the null model, for the second model, *Model.2(a).Bw*, start with the full model instead.

For each of the two models, report the variables and interactions included in the final model. Are there any interesting differences between the variables included in these models, and those that were included in the corresponding linear regression in Project 1?

For both models, calculate McFadden's adjusted R^2, AIC and BIC and decide which of the models seems best.

2(b). **Influential observations:** Plot Cook's distance for *Model.2(a).Bw* against the linear predictor, $\mathbf{x}_i\hat{\beta}$, using different colours for low (1) and high (0) precipitation. Also add a suitable reference line.

Identify the observation with the highest Cook's D among those with low precipitation and the one with the highest Cook's D with high precipitation and highlight them in the plot.

Also plot the standardized deviance residuals against the linear predictor, with colour coding and suitable reference lines, and highlight the two observations with high Cook's D you identified before. Explain why these two observations have both a high Cook's D and a large residual. Hint: think about the combination (contradiction?) of observed value and predicted value.

Part 3. Goodness-of-fit

3(a). **Confusion:** Use *Model.2(a).Bw* and the threshold value 0.5, classifying observations with $\hat{p}_i \leq 0.5$ as "should not have low precipitation", and observations with $\hat{p}_i > 0.5$ as "should have low precipitation".

Present the resulting confusion matrix for the model and calculate the sensitivity, specificity, accuracy, and precision. Comment on the result.

3(b). **ROC-curves and AUC:** Plot the ROC-curves for all five models, *Model.1(c)*, *Model.1(d)*, *Model.2(a).Fw*, and *Model.2(a).Bw*, in the same plot and present a table with their AUC-values, including 95 % confidence intervals. Perform pair-wise tests comparing the AUC for *Model.2(a).Bw* against each of the other models and discuss the result. Does it agree with the conclusion in 2(a)?

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 threshold:** For *Model.2(a).Bw*, find the optimal threshold for p, where the sensitivity and the specificity are approximately equal, and as large as possible. Use the threshold to calculate a new confusion matrix, sensitivity, specificity, accuracy, and precision, and compare with the result in 3(a).

3(d). **Goodness-of-fit test:** Perform a Hosmer-Lemeshow goodness-of-fit test for model *Model.2(a).Bw*. Use a couple of different number of groups, g, starting at $g = p + 2$, in order to se how sensitive the test is to different choices. For each choice of g, report the smallest expected number of observations in any group.

Plot the observed and expected number of successes and failures, using group number on the x-axis. Relate the result of the HL-test to the appearance of the plot and comment on the result.

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

End of Project 2