MASM22/FMSN30/FMSN40 SPRING 2020
PROJECT 2: LOGISTIC REGRESSION
LOW MONTHLY PRECIPITATION IN SWEDEN

Short instructions

Follow the instructions on the website under *Project 2: Instructions*. There you will also find the data and updated versions of this text.

Remember: The report must be written in English and submitted twice.

First in a preliminary version for peer assessment on 11.00 on Wednesday 13 May (Assignment: *Report2 - peer version*). At 11.30 you will get a random report for peer assessment. Read it and give feedback to its authors by 10.00 on Thursday 14 May. And get corresponding feedback from whoever read your report. Then you must submit a final version for the teachers' eyes at 16.00 on Friday 15 May (Assignment: *Report2 - final version*).

Introduction

We will return to the Swedish weather data that we used in Project 1 and now concentrate on the probability of an anusually low amount of rain.

The data is, as before, located in the file `weather.rda` located on the *Project 2: Instructions* page and consists of the following variables:

`month`	a text (factor) variable of the format `"yyyy-mm"`
`year`	the year (numeric)
`monthnr`	the month of the year (numeric)
`location`	the location of the weather station (text): Lund, Uppsala or Abisko.
`rain`	total monthly precipitation (mm)
`temp`	average monthly temperature (°C)
`pressure`	average monthly air pressure (hPa)

We want to estimate the probability of low precipitation during a month. We define low precipitation as a total monthly precipitation that lies below 25 mm (this is the lower quartile in our data set). Create a new variable, where `as.numeric()` turns the value TRUE into 1 and FALSE into 0:

```
weather$lowrain <- as.numeric(weather$rain < 25)
```

1 The null model

1.(a). Calculate the proportion of the months that have low precipitation, by, e.g., `mean(weather$lowrain)`. Use this to estimate the odds of a month having low precipitation, as well as the log-odds.

1.(b). Fit the null logistic regression model with `lowrain` as Y-variable and only an intercept, β_0. Report the β_0-estimate together with its 95 % confidence interval.

Use the β_0-estimate to estimate the odds of a month having low precipitation. Also report a 95 % confidence interval for the odds.

Use the-estimated odds to estimate the probability of a month having low precipitation. Also report a 95 % confidence interval for the probability.

Compare with 1.(a) and make sure you understand how they relate to each other.

2 Temperature . . .

2.(a). Plot `lowrain` against `temp` and add a moving average with `geom_smooth(method = loess)`. Does it seem reasonable to use `temp` as covariate?

2.(b). Fit a simple logistic regression model using `temp` as covariate. Report the β-estimates with 95 % confidence intervals, as well as the odds(ratio) e^{β}-estimates with confidence intervals. Test if there is a significant relationship with temperature.

How does the odds of low rain change when the temperature is increased by 1 °C? How does the odds of low rain change when the temperature is *decreased* by 1 °C?

2.(c). Use the model to estimate the probability of low rain when the temperature is -10 °C, -9 °C, $+9$ °C and when it is $+10$ °C and calculate 95 % confidence intervals for the estimates. Try to explain why the difference in the probability of low rain is different when we change the temparature from -10 °C to -9 °C, compared with when we change from $+9$ °C to $+10$ °C.

2.(d). Calculate the predicted probabilities and add them to the plot in 2.(a) together with a 95 % confidence interval.

2.(e). Calculate the leverage and plot them against temperature, adding horizontal lines at $1/n$, the minimal value, and $2(p+1)/n$. Also make sure the y-axis covers 0. Are there any observations with extreme leverage?

2.(f). Calculate the standardized deviance residuals for the model in 2.(b) and plot them against temperature. Add horizontal lines at 0, ± 2 and ± 4. Are there any alarmingly large residuals?

2.(g). Calculate Cook's distance and plot them against temperature, using different colors for low and not-low rain, `aes(color = as.factor(lowrain)`. Also add a horizontal line at $4/n$. Are there any observations that have had a large influence on the estimates? Are these observations the same as those that had the highest leverages?

3 . . . or pressure. . .

3.(a). Plot `lowrain` against `pressure` and add a moving average. Fit a simple logistic regression model using `I(pressure - 1012)` as covariate. Report the β-estimates with 95 % confidence intervals, as well as the odds(ratio) e^{β}-estimates with confidence intervals.

Calculate the predicted probabilities and add them to the plot together with a 95 % confidence interval.

3.(b). Calculate the leverage and plot them against air pressure, adding horizontal lines at $1/n$, the minimal value, and $2(p+1)/n$. Also make sure the y-axis covers 0. Are there any observations with extreme leverage?

3.(c). Calculate the standardized deviance residuals for the model in 3.(a) and plot them against air pressure. Add horizontal lines at 0, ± 2 and ± 4. Are there any alarmingly large residuals?

3.(d). Calculate Cook's distance and plot them against air pressure, using different colors for low and not-low rain. Also add a horizontal line at $4/n$. Are there any observations that have had a large influence on the estimates? Are these observations the same as those that had the highest leverages?

3.(e). Compare the leverage, residual, and Cook's distance plots with those for the temperature model from 2.(b). Which model seems best?

3.(f). Calculate $R^2_{\text{Cox-Snell}}$ and $R^2_{\text{Nagelkerke}}$ for both models. Which model is best?

3.(g). Calculate AIC and BIC for both models. Which model is best?

4 . . . or both with location?

4.(a). Fit a model with both temperature, pressure-1012, an interaction between them, as well as the location variable (without interaction). Don't forget to set a suitable reference category.

Test, using a suitable test, whether this model is significantly better than the model with only temperature, 2.(b). Also test whether this model is significantly better than the model with only air pressure, 3.(a).

4.(b). Calculated the predicted probabilities, `phat`, using model 4.(a). Plot `lowrain` against `temp`, in subplots by `location` and add the predicted probabilities to the plot, using different colours depending on the value of `pressure` by

```
+ geom_point(aes(y = phat, color = pressure)) +
scale_color_viridis_c()
```

Also do a plot with `pressure` on the x-axis and colours set by `temp` instead.

Which variable causes the largest variability in the probability of low rain: temperature or air pressure? In which location does the temperature add the most extra information, in addition to pressure?

4.(c). If you want to predict the probability of low rain in Lund, which variable seems more useful, according to 4.(b): temperature or air pressure?

Fit a model to the data from Lund, using temperature and air pressure, and their interaction. Present the β-estimates and e^{β} together with 95 % confidence intervals. Are all variables significant?

Perform a backward elimination, with BIC as criterion, in order to reduce the model. Does the resulting model agree with your conclusion of which variable was more useful?

4.(d). Calculate the standardized deviance residuals for model 4.(a) and plot them against $\mathbf{x}_i\hat{\boldsymbol{\beta}}$. Plot them against temperature, with colours according to pressure, separately for each location. Also plot them against pressure, coloured by temperature. Add suitable horizontal lines to each plot as visual guides.

Any problematic residuals? Also compare with the plots in 2.(f) and 3.(c). Have the residuals improved?

4.(e). Calculate Cook's distance for model 4.(a) and plot them in the same ways as the residuals. Compare with the plots in 2.(g) and 3.(d). Have they improved?

4.(f). Calculate $R^2_{\text{Cox-Snell}}$ and $R^2_{\text{Nagelkerke}}$ for model 4.(a) and compare with 2.(b) and 3.(a).

4.(g). Calculate AIC and BIC for model 4.(a) and compare with 2.(b) and 3.(a). Which model is best?

5 Goodness-of fit

5.(a). Classify the observations with $\hat{p}_i \leq 0.5$ as failures = "should not have low rain" and whose with $\hat{p}_i > 0.5$ as successes = "should have low rain", for each of the three models 2.(b), 3.(a) and 4.(a). Calculate the confusion matrices, sensitivity, specificity, accuracy and precision for the models. Is any of the models out-performing the others in all, or some, of the aspects?

5.(b). Calculate and plot the ROC-curves for the three models, and calculate their AUC, with 95 % confidence intervals. Also perform a test comparing the AUC for model 2.(b) and 3.(a). Comment on the results.

5.(c). For each of the three models, find and report the optimal threshold making the sensitivity and specificity approximately equal, and as large as possible.

Calculate new confusion matrices, sensitivity, etc, using the new thresholds and compare with the result in 5.(a). Did the increase in sensitivity come with a price?

5.(d). Perform a Hosmer-Lemeshow goodness of fit test for each of the three models. Use a handful of different number of groups, g, starting at $g = p + 2$, to see how sensitive the test is to different choices. Also report the smallest expected number in a group, for each g (avoid going much below 5).

For each of the models, pick a suitable number of groups and plot the expected and observed number of successes and failures together, using group number on the x-axis. Relate the result of the HL-tests to the appearance of the plots.