BSc in Cognitive Science, Aarhus University

Methods 1 E2021

**PORTFOLIO 4**

**Mixed-effects models and logistic regression**

**Deadline: December 2nd, 2021**

The portfolio uses two data sets: The “**Breakage Angle of Chocolate Cakes**” data set and the “**Titanic**” data set. The data sets include the following variables:

Cake:

* **replicate**: a factor with levels 1 to 15 indicating # replication of test
* **recipe**: a factor with levels A, B, and C for each of three different recipes
* [**temperature**: disregard]
* **angle**: a numeric vector giving the angle at which the cake broke
* **temp**: a numeric value of the baking temperature (degrees F)

Titanic:

* **Survived**: a numeric value indicating whether each participant survived the incident or not
* **Pclass**: a currently numeric variable with levels 1 to 3 for 1st, 2nd, and 3rd class
* **Name**: a character variable with passenger names
* **Sex**: a character variable with two levels (male/female)
* **Age**: a numeric value indicating passenger age
* [**Siblings/Spouses Aboard**: disregard]
* [**Parents/Children Aboard**: disregard]
* [**Fare**: disregard]

**Analysis 1:** Cake breakage

To predict the angle at which cake break, I fitted a linear mixed-effect model to predict *angle* as the outcome variable. I started with 3 models and found temperature to be the predictor variable. Recipe turned out to be a random slope and replicate to be the random intercept:

*Cake\_1 = angle ~ temp + (1+recipe|replicate)*

This model got chosen as it had the lowest AIC and highest conditional R^2. This means, that the angle at which cakes break is significantly predicted by temperature (beta = 0,158, SD = 0,016, t = 9,8, p = < 0.001). When temperature increases, the angle that the cake breaks at increases.

|  |  |  |
| --- | --- | --- |
| **Models:** | **AIC** | **R2c** |
| *Cake\_1 = angle ~ temp + (1+recipe|replicate)* | *1666* | *0.702* |
| Cake\_2 = angle ~ temp + recipe + (1|replicate) | 1674 | 0.659 |
| Cake\_3 = angle ~ temp \* recipe +(1|replicate) | 1678 | 0.660 |
| Cake\_4 = angle ~ temp \* recipe +(1|replicate) +(1|recipe) | 1677 | 0.658 |

**Summary output:**

Text

Description automatically generated

**Check assumptions:**

**Chart, scatter chart

Description automatically generated**

There is compact and unsystematic spread in the plot therefore the assumptions are fulfilled.

**Analysis 2:** Titanic survival

To predict the survival rate of titanic passengers I created a generalized logistic model with binomial outcomes on the titanic data set, after testing other plausible models:

*Survived ~ Sex + Age + Passenger\_class*

As seen in figure 1 ‘*summary of GLM’*, the model has a baseline passenger of a *first-class female at age 0*, and all other predictors has a negative log-odds, meaning everyone has a smaller likelihood of surviving than the baseline passenger. All predictors have a significant p-value < 0.01.

When trained on a training dataset (seed (666) in r, p 0.8) the prediction accuracy on the remaining test dataset was 78 %, see figure 2 ‘Confusion matrix’. The training dataset had a R2 MacFadden of 0.409 and the test dataset a R2 MacFadden of 0.376.

**Chart, scatter chart

Description automatically generated**

*Text

Description automatically generated*

Figure . Summary of GLM

**Table of survival:**

|  |  |
| --- | --- |
| **Passengers (median age)** | **Probability of survival** |
| **First class female** | **92 %** |
| **Second class female** | **81 %** |
| **Third class female** | **60 %** |
| **First class male** | **41 %** |
| **Second class male** | **23 %** |
| **Third class male** | **9 %** |

**Chart, treemap chart

Description automatically generated**

Figure . Confusion matrix