BSc in Cognitive Science, Aarhus University
Methods 1 E2021

PORTFOLIO 4
Mixed-effects models and logistic regression
Deadline: December 2^{nd}, 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 1^{st}, 2^{nd}, and 3^{rd} 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)

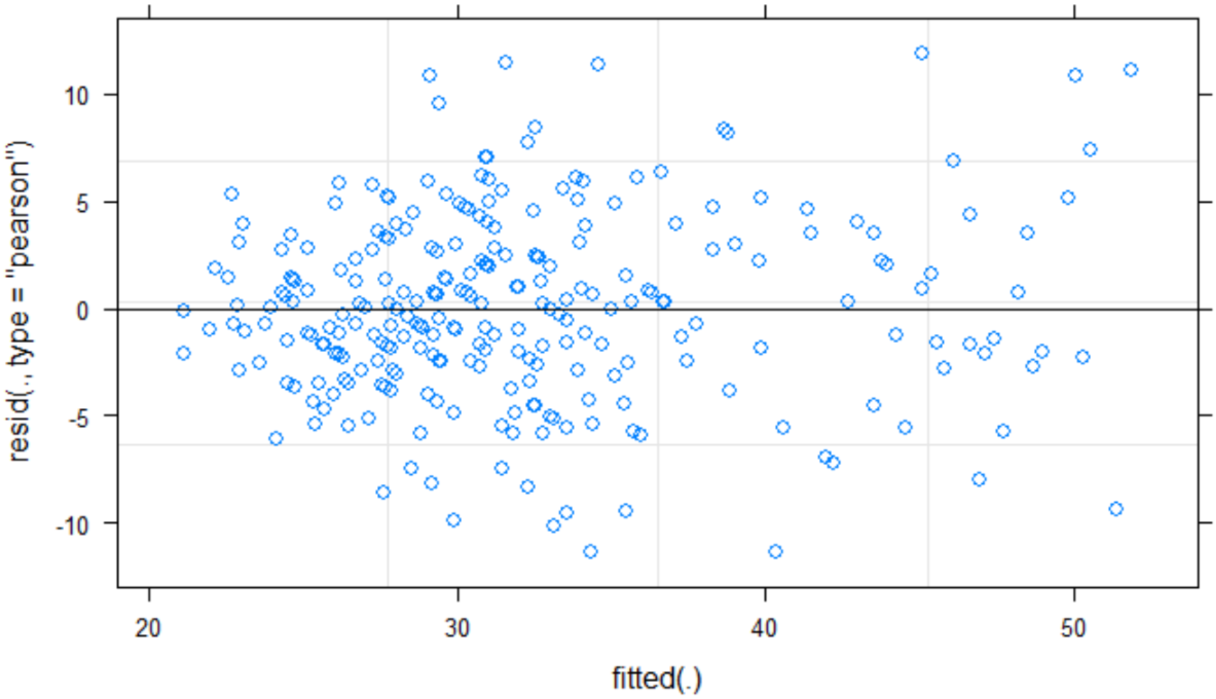
This model got chosen as it had the lowest AIC and highest conditional R^2. This means, that the angle at which cakes break is significantly predicted by temperature (beta = 0,158, SD = 0,016, t = 9,8, p = < 0.001). When temperature increases, the angle that the cake breaks at increases.

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:

```
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: angle ~ temp + (1 + recipe | replicate)
   Data: cake

     AIC      BIC   logLik deviance df.resid
  1666.2   1698.6   -824.1   1648.2      261

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.51095 -0.56465 -0.01979  0.62483  2.62895

Random effects:
 Groups    Name        Variance Std.Dev. Corr
 replicate (Intercept) 24.981   4.998
           recipeB      8.513   2.918    0.42
           recipeC     15.347   3.918    0.31 0.99
 Residual              20.477   4.525
Number of obs: 270, groups:  replicate, 15

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept)   1.77214    3.50194 219.36537   0.506    0.613
temp          0.15803    0.01613 239.97848   9.800   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
     (Intr)
temp -0.921
```

Check assumptions:



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

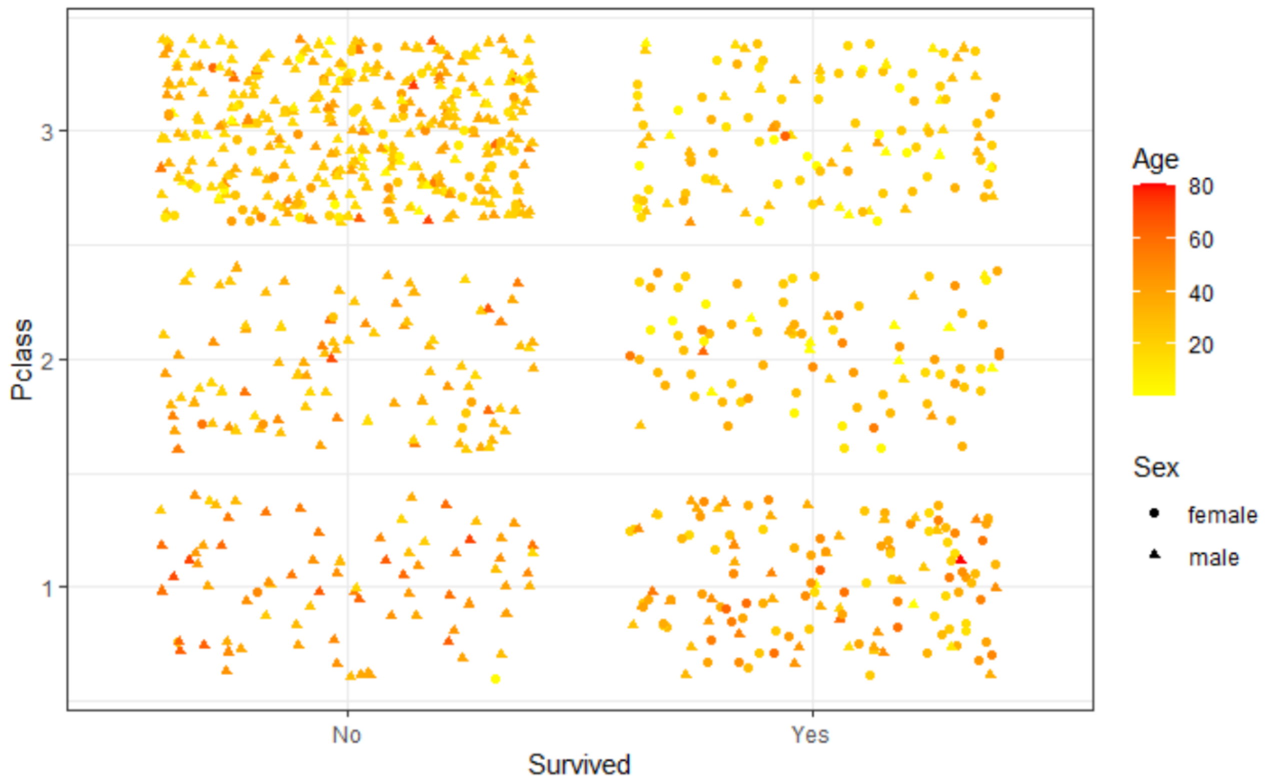
Analysis 2: Titanic survival

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:

$$Survived \sim Sex + Age + Passenger\_class$$

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.



```
Call:
glm(formula = Survived ~ Sex + Age + Pclass, family = binomial,
    data = titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6811  -0.6653  -0.4137   0.6367   2.4505

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.63492    0.37045   9.812  < 2e-16 ***
Sexmale     -2.58872    0.18701 -13.843  < 2e-16 ***
Age         -0.03427    0.00716  -4.787 1.69e-06 ***
Pclass2     -1.19911    0.26158  -4.584 4.56e-06 ***
Pclass3     -2.45544    0.25322  -9.697  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.77  on 886  degrees of freedom
Residual deviance:  801.59  on 882  degrees of freedom
AIC: 811.59

Number of Fisher Scoring iterations: 5

       GVIF Df GVIF^(1/(2*Df))
Sex    1.09  1            1.04
Age    1.35  1            1.16
Pclass 1.45  2            1.10
```

Figure 1. 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 %**

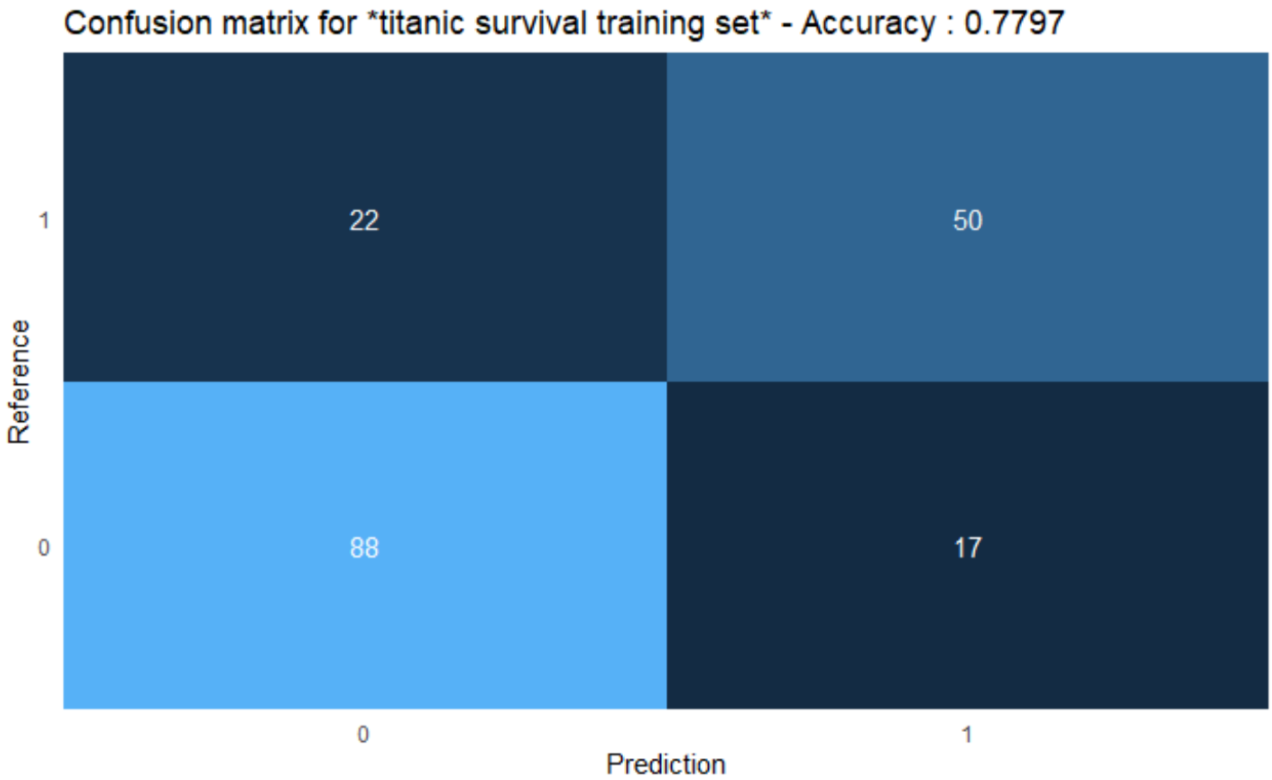


Figure 2. Confusion matrix