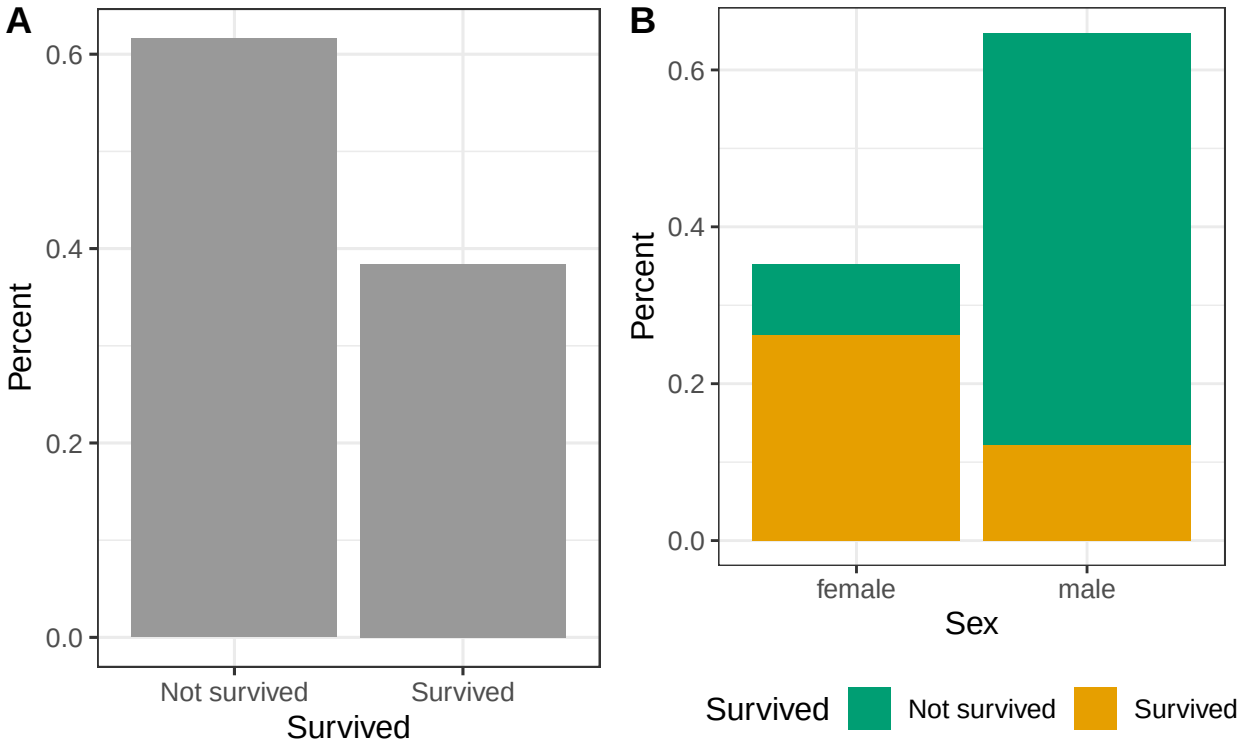
Summary: Who survived the Titanic?

Edgar Treischl

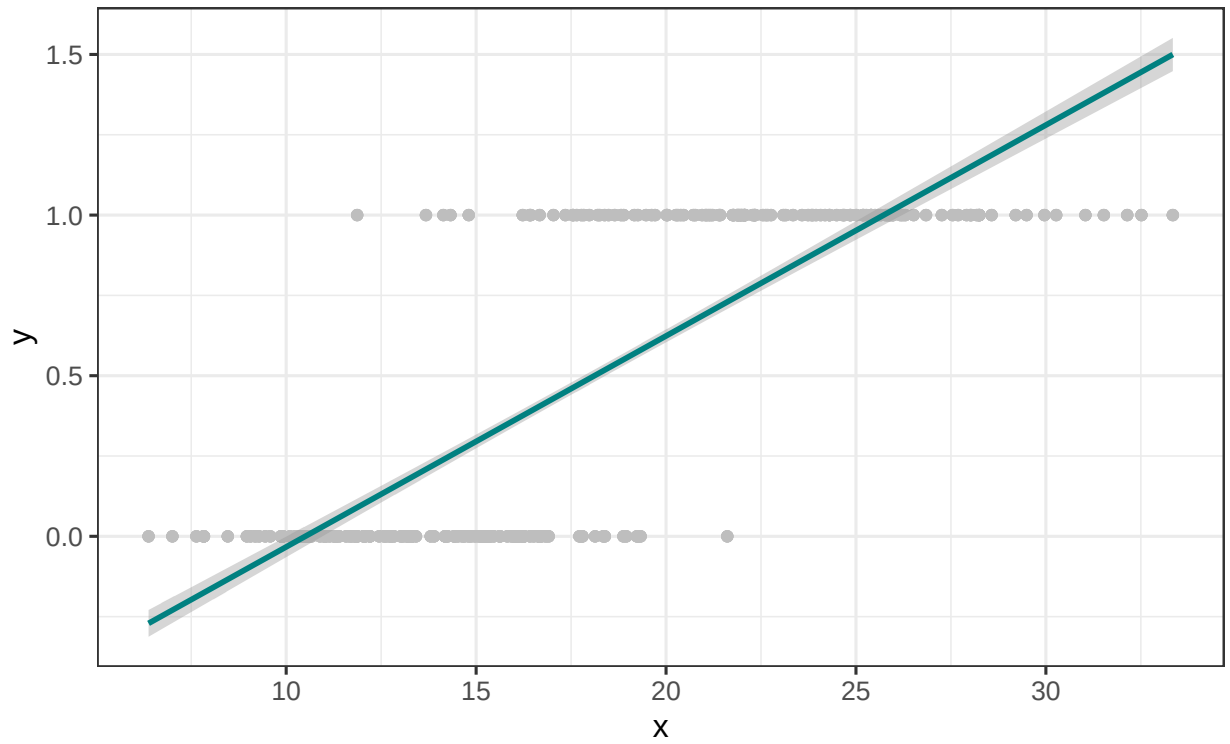
05. Mai 2021

How many people survived the Titanic? Who had a higher chance to survive, men or women? What about class and age? This app shows you some basic aspects about logistic regression. We use passenger's sex, class, and age to estimate the effect on the survival of the Titanic accident.

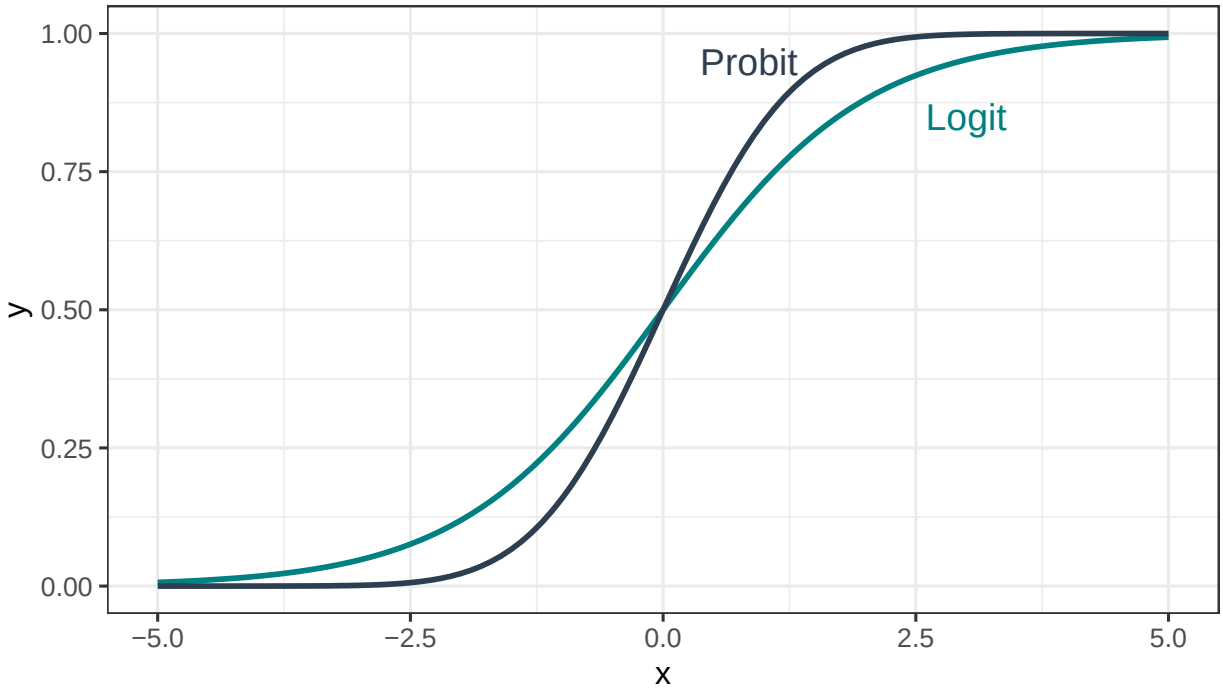


Idea: Logistic regression, but why?

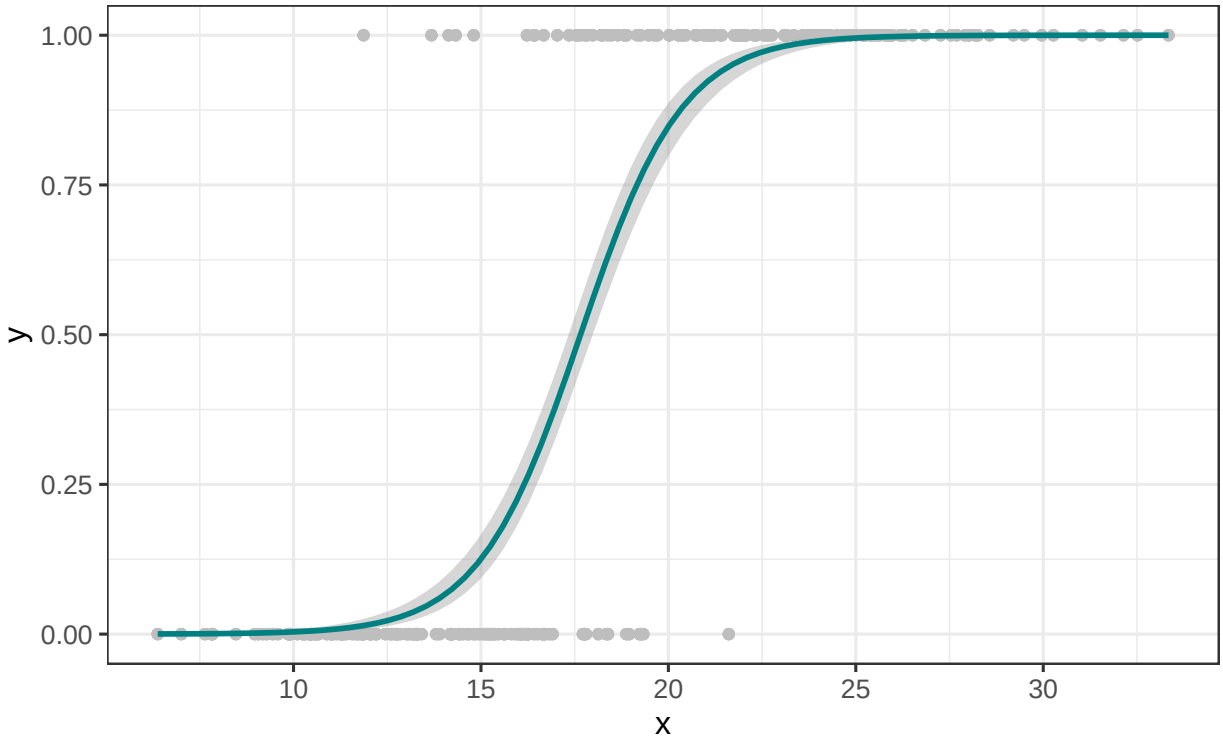
There are several reasony why logisitic regressions were invented to model binary outcomes. You can see the most obvious reason in the figure below. Imagine that we insert a regression line to model a binary outcome. Look how a scatter plot would look like in such a situation.



In a linear regression, we try to fit a line that minimizes the error, but in the case of a binary outcome, the observed error is not homoscedastic. Moreover, the variance of the error term depends on the particular value of X, but we observe only 0 or 1. There are no no observations between zero and one, even though we use a regression line to model between the two outcome values. The next outshows shows you how the distribution of a logistic and the probit function looks like.



Both distributions are often used to model binary outcomes in the social sciences. Of course, we can adjust the first scatter plot and use a logit function to describe the relationsship between X and Y instead of a regression line.

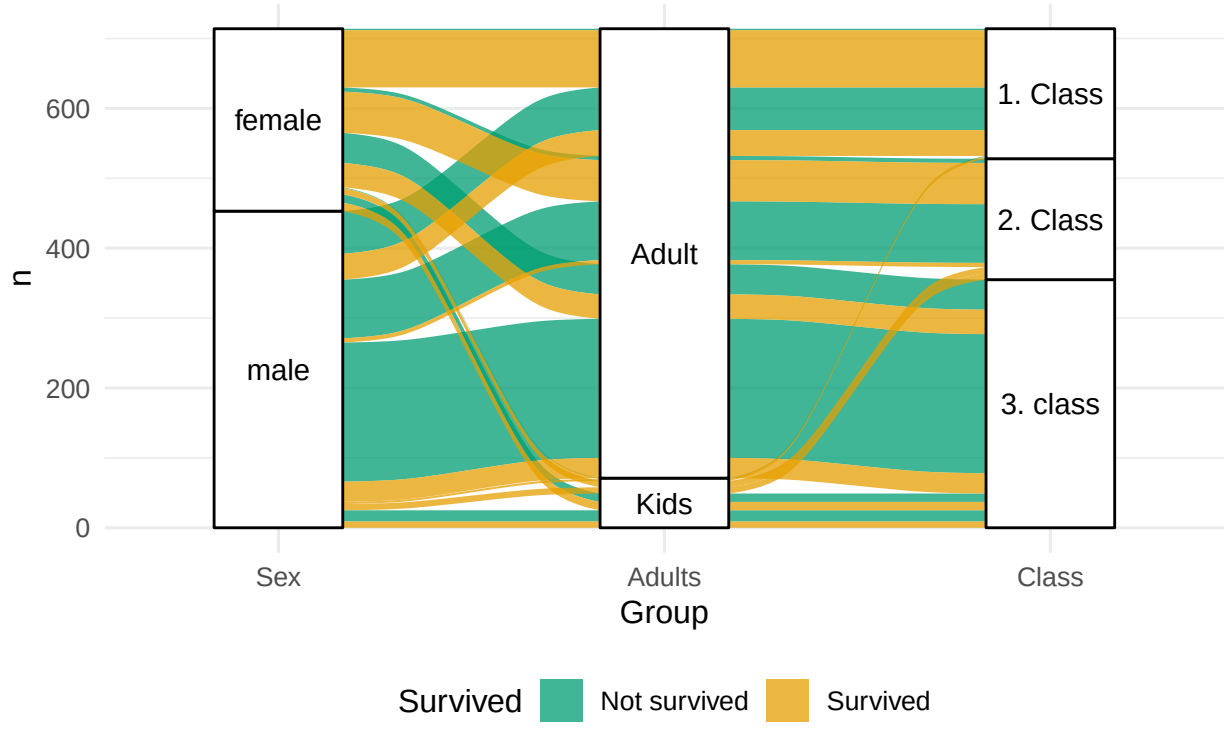


The data for the scatter plot was simulated, that is reason why it looks nice and smooth, but I hope you get

a first impression about the difference between linear and logistic regression.

The independent variables

The surivival of the Titanic is a binary outcome and on the left side you can see how many people surived, based on a series of simple bar plot. However, If we want to explore the effect of several independent variables simultaneously, we can use a sankey plot. A sankey plot shows you how these variable work together. You can literally see how many people of each of group survived(1) or did not survive (0).



What would you say? Which one has the strongest effect on the survival? It looks like sex and class have a strong effect on survival. Instead of guessing, we can use a logisitic regression to estimate the effect of passengers' sex, age, and class simultaneously. You never applied a linear regression? Well, check out the regression in a nutshell app first because on this page I assume that you are familar with the principals of a linear regression analysis.

The Model

Let's run a logistic regression. Which of the following independent variables do you want to include to estimate the effect on survival?

```
summary(glm(Survived ~ Sex + Pclass + Age,
            family = binomial(link = 'logit'),
            data = train_df))
```

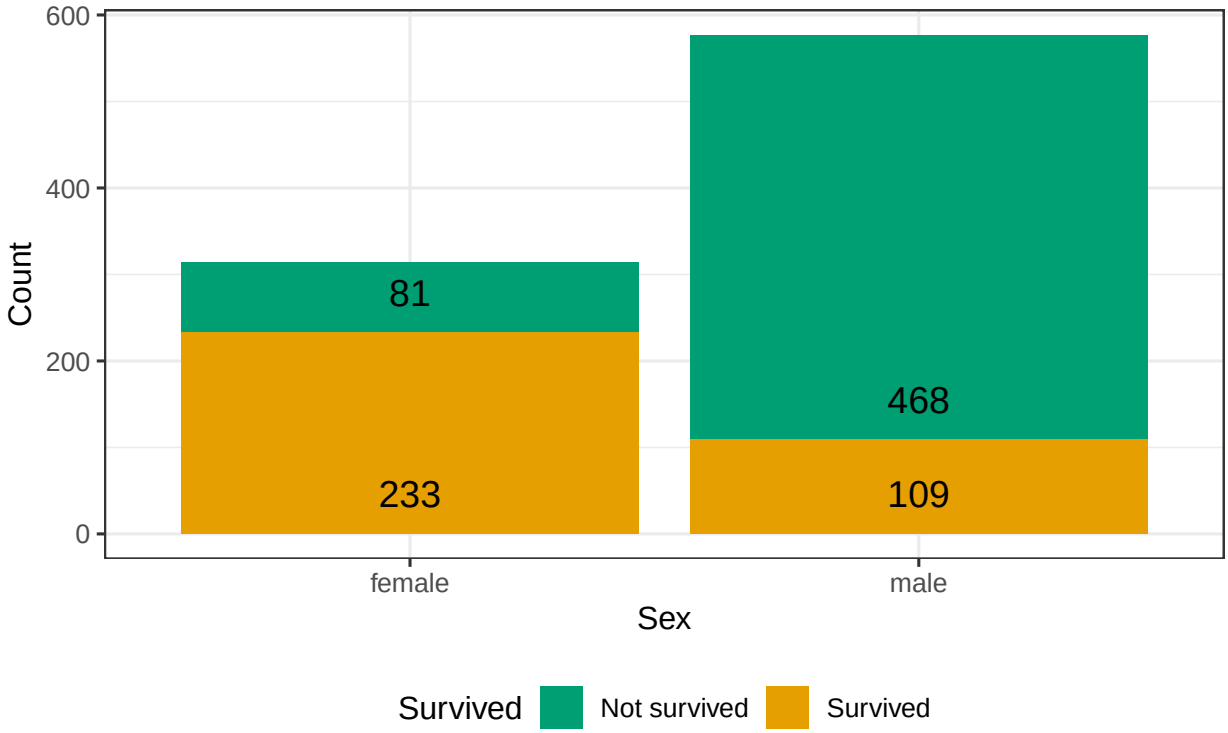
```
##
## Call:
## glm(formula = Survived ~ Sex + Pclass + Age, family = binomial(link = "logit"),
##     data = train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.7303  -0.6780  -0.3953   0.6485   2.4657
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          3.777013   0.401123   9.416  < 2e-16 ***
## Sexmale             -2.522781   0.207391 -12.164  < 2e-16 ***
## PclassSecond class -1.309799   0.278066  -4.710 2.47e-06 ***
## PclassThird class  -2.580625   0.281442  -9.169  < 2e-16 ***
## Age                -0.036985   0.007656  -4.831 1.36e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 647.28  on 709  degrees of freedom
##   (177 observations deleted due to missingness)
## AIC: 657.28
##
## Number of Fisher Scoring iterations: 5
```

Maybe you picked passenger's sex, but what tells you the estimate(s) of your model? The estimate for Male compared to women is negative and -2.51? Due to the assumptions of the logistic regression, we get the logarithm of the odds to survive as a results. In case of log(odds), we can only say whether an effect is positiv or negative and check the significane. Thus, such a estimate is hard to explain what it really means. Instead of the log odds, we can estimate odds ratios and make prediction about the probability to survive. Both are easier to interpret.

Odds Ratio?

What would be the chance to survive for men if they had the same odds (chance) to surive compare to women? The odds would be one, since we expect that the same amount of men and women would surived. You can calculate the odds ratio with the help of the logisitic regression. However, let's try to calculate it by hand to get a better intuition what a OR means. In order to do so, the next plot how many men and women have survived.



Look at the bar graph and the numbers of each group. We get the men's odds to survive if we divide the number of survived men (109) by the number of men who did not surive (468). Women's odds to survive are calculated the very same way (233/81). In the last step, divide men's odds by women's odds and you get the odds ratio for men to survive.

We don't have to work this out in our own head, just use your statistics software as a calcultor, as the next console shows:

```
Oddswoman <- 233/81
Oddswoman
```

```
## [1] 2.876543
```

```
Oddsman <- 109/468
Oddsman
```

```
## [1] 0.232906
```

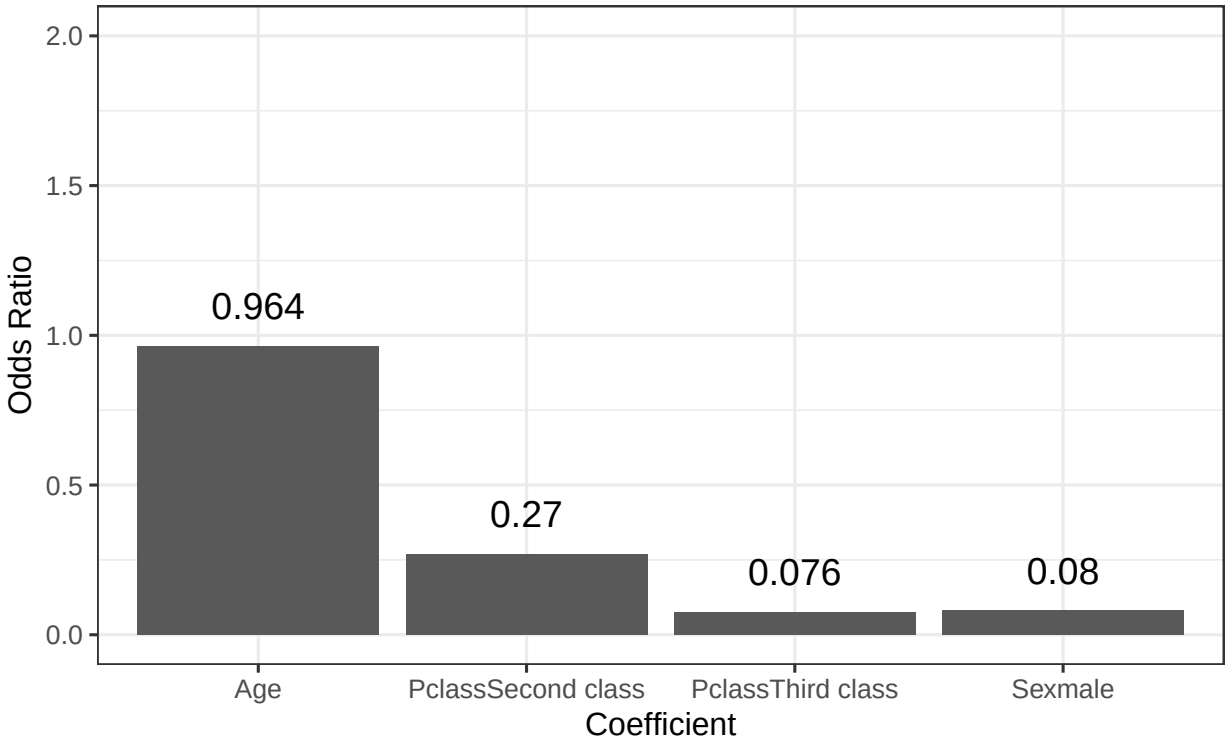
```
ORman <- round(Oddsman/Oddswoman, 3)
ORman
```

```
## [1] 0.081
```

Remember the interpretation

- OR = 1: No effect
- OR > 1: Positive effect
- 0 < OR < 1: Negative effect

Thus, men's chance to surive is reduced by the factor 0.08 compared to women. What about age and the other variables in your model? Go back to the Model tab if you did not choose any independent variable for the analysis. And all the OR from the full model:



A lot of people argue that OR are also not very intuitive and they provide several good reasons why this might be the case. For instance, include age in your model. What would you say regarding the odds ratio for age? Has age no or at least only a small effect? Nope, age has a substantial effect on the chance to surive! A OR is intuitive if we compare groups, in the case of age it is easier to examine the effect size if we calculate probabilities for the entire range of age. Go and grab your wand, on the next page you can make predictions and see how each variable affects the probability to survive.

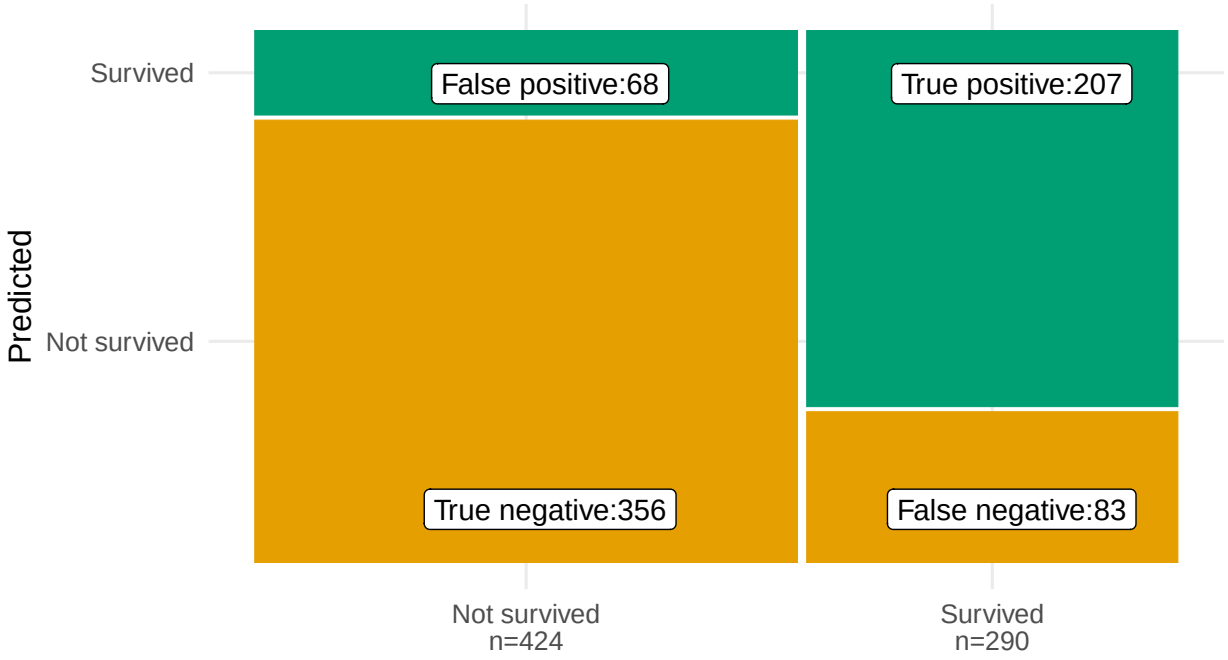
Predictions

I guess this is the most intuitive interpretation of a logistic regression, since we want to know about the probability to survive, not log odds or a ratio. We can use the model to predict values! For instance, see how the prediction of survival drops if you switch from women to men. Below you can provide values for age and class as well.

Performance: How well does the model perform?

Probably you know that R^2 is often used to assess the performance of a linear model. Unfortunately, it is a bit harder to assess the performance of a logistic regression. There are pseudo R^2 for logistic regression to compare nested models, but we cannot interpret them as explained variance as we do it with linear model. Instead, you may encounter two terms: sensitivity and specificity.

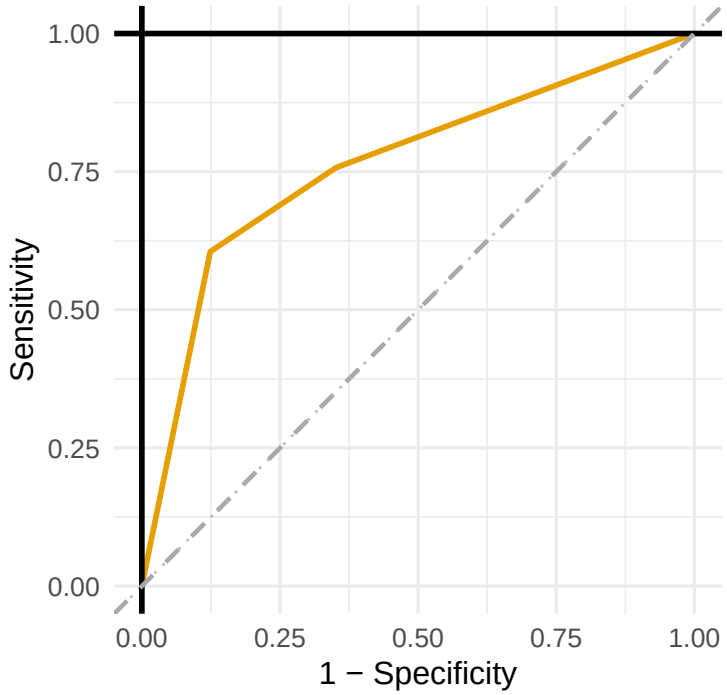
Sensitivity takes into account whether we classified the true outcome right. How many people who survived (true positives) did our model classify as survivor? The mosaic plot shows you how many passengers did (not) survive; on the x-axis as we have observed and y-axis displays our prediction. We can calculate the sensitivity by dividing the true positives by all the people who survived (207/290). Thus, the sensitivity is 0.71.



Sensitivity: 0.71, Specificity: 0.84

How many people did we classify as not-surived, who actually did not survive the Titanic (true negative)? Hence, the specificity does the same job for the negative outcome. If you divide the true negatives by all people who did not survive (356/424), you get the specificity.

A common way to combine both indicators are ROC curves. As the plot below illustrates, a ROC curve displays the sensitivity on the y-axis, and the false positive rate (1-specificity) on the x-axis. What does the ROC curve tells you? By predicting a binary outcome, we want to achieve two things simualtaneously: We want to classify the number of people who survived correctly, while we wish that the number of false positives is rather small. Thus, we wish to have a sensitivity of 1 and a false positive rate of zero (highlighted in black in the ROC curve below). However, if the model does not help to predict the outcome, the ROC curve would be a diagnal line, since a fair toss of a coin would have the same predicitive power. 50% of the time we identify people correctly, 50% of the time we make a wrong prediction.



Sensitivity and specificity are not the only measures of performance, but in terms of interpretation, we should remember the further the ROC curve is away from the diagonal (and closer to the black line), the greater the explanatory power of the model.