

# Titel

Victor Todd-Moir  
Københavns Universitet

dd-mm-åå

## Data analysis

We would like to predict the prices of games using a regression model. We will in the following scale our features to zero mean and unit variance. Afterwards we will perform a principal component analysis (PCA) in order to reduce the amount of features taken into consideration and thus we will account for a degree of overfitting. We will like to keep 90% of the variance in our training data. This is done by looking at the explained variance ratio of each feature and keeping the features with most explained variance ratio until we reach our threshold of 0.90. Afterwards we will fit an elastic net-model to our training data using 10-fold cross validation and perform a grid search to tune our hyperparameters of the elastic net-model. The hyperparameters are  $\lambda$  and  $\alpha$ .  $\lambda$  is the penalty term that penalizes the amount of features in our model and  $\alpha$  determines the convex combination of LASSO penalization and ridge penalization.

First we scale our training data. This is done to make sure that features of different magnitude are taken equally account for when fitting our model. As we have ## features to begin with one can easily imagine that a lot of our features would be made redundant without scaling, as we have many binary features in our data and we penalize the amount and scale of our  $\beta$ 's in our final model.

First of all we have to consider our features and our target. Our taret is the current price (as of 18th August). Among our features we have the amount of days since release. This feature will be positive if the game has been released and negative if the game is yet to be released. If we take a look at our model, our predictions will look like the following:

$$\hat{y} = \beta^T x$$

## Conclusion

As we don't have data for the initial prices of games that has been released, we will assume a linear relationship between initial price, current price and days since release, hence

$$y_{initial} + \beta_{daysold} \cdot x_{daysold} = y_{current}$$

The number of games that has not been released but already have a price is 24. This is fairly low but nonetheless we will test our model under this linear hypothesis.