FDS Week 9 Workshop: instructors' answers

1. Identify what question or prediction problem the regression was being used to address. What are the independent and dependent variables?

Question: can we predict the market value of football forward players based on various attributes?

Response/target/dependent variable: market value of forward player

Predictors/features/independent variables (found in the Methodology section, p 4):

- Club
- Age
- Height
- Dominant legs
- Nationality
- Outfitter (e.g. Nike or Addidas)
- Matches played in season
- Number of direct contributions to goals
- Card scores (I.e. number of red and yellow cards)
- Total time played in season

Many of these variables, which might naturally be thought of as numeric (e.g. height), have been converted to categorical variables. The authors argue that there are likely to be nonlinear relationships, e.g. salary will increase with age and then decrease, and this conversion allows the multiple regression model to learn the nonlinear relationships.

In the lecture on multiple regression, we discussed how a binary categorical variable can be represented as a 0/1 indicator variable. As mentioned in the lecture notes on Data the term **one-hot encoding** is also used to describe how categorical variables with more than two values (for example colours) can be encoded by creating an indicator variable corresponding to each category. Only the variable corresponding to the category being encoded is set to 1, and the others are all set to 0, leading to the name "one-hot encoding" for this type of encoding.

Sometimes in machine learning, one-hot encoding can be used to encode a numeric variable, since it allows nonlinear relationships to be predicted: The mean salary in the 26-27 years old range will be higher than the salary in the 18-19 range or the 34-35 years old range. A very simple prediction algorithm (based just on age) could predict a salary based on the average salary in the age band of the player - I.e. what players of a similar age earn.

In machine learning the term **feature engineering** refers to creating new variables ("features") from data, as the authors have done in this paper.

2. Identify any numerical or visual diagnostics that were reported for the fit, e.g. coefficient of determination, RMSE, residual plot. (See the Week 5 and week 6 lectures for the diagnostics).

There is both the coefficient of determination and adjusted coefficient of determination. In the full (105 predictor) model, the adjusted R^2 is a lot lower than the R^2 – a sign that there are too many variables included (Fig 1). The 55-variable model is better (Fig 4).

There are plots to test heteroscedasticity, which we covered in the lectures on linear regression. There's also the Bruesch-Pagan test for heteroscedasticity. Essentially this involves regressing the squared residuals on the independent variables. If that regression explains a lot of the variance, then it indicates heteroscedasticity.

3. Identify any problems you see in the analysis, e.g. lurking variables, low adjusted coefficient of determination

Number of parameters and amount of data

For *large* amounts of data, one-hot encoding can work well. But here there are only 105 observations, so perhaps about 10 players in each age band. Suppose one age category contains a superstar player. The mean in that band will be massively inflated by the superstar's salary, and so predictions of players in this age group are likely to be inflated.

Note that after the one-hot encoding, there are 105 independent variables and 105 observations! This is almost certainly "over-fitting" (which we cover in the lecture no on k-Nearest neighbours and multiple regression). Bly, the model is too flexible, and so is learning the noise in the data.

The authors recognise over-fitting is a problem and trim the model down to 52 parameters using a "backward elimination" method. The authors don't specify the details of how coefficients are eliminated, but it looks like they use the p-value of each coefficient (a concept covered in the part of the course on Inferential statistics). The chance that assuming there is no underlying relationship between two variables (age and salary), randomness in the data gives rise to the regression coefficient we obtain. So a p-value equal to 0.1 means that if there were no relationship in the data, there would be a 10% change of obtaining a regression coefficient as big (in magnitude) as the one returned by the regression. They first get rid of the parameter that looks most likely to have arisen by chance (largest p-value), and run the model again and repeat.

It would have been interesting to see a linear regression without the discretisation of the ranges. What is the R^2 of such a model compared to their feature engineering? Following this linear regression, they could have tried creating variables (features) such as "Age squared" (as we considered in the lectures on linear regression), which would have allowed a nonlinear relationship to be predicted.

The low adjusted coefficient of determination in the 105 variable model is definitely problematic. With 55 parameters the model is still likely to be over-parameterised. It would be interesting to test the model on unseen data.

The final selection of variables seems a bit weird. E.g. some clubs are included, but others are not - though probably Man City and Barca do explain a lot of variance in salaries.

Display of results

The display of the results was not good, with the tables being poorly formatted, and not referring to real variable names. This makes it difficult for the reader (and the authors) to understand the meaning of the model, and whether it makes sense when compared with our prior knowledge, in other words to critically evaluate the model.

It would have been useful to see the features ranked by the size of the coefficient, assuming that we had first standardised the data (as we have covered in the topics on descriptive statistics and linear regression). In such a linear model the size and sign of the coefficient indicates the influence of the variable in the model. The first thing is the sign -- e.g. does it make sense that being mid-range in age is positive, and being older is negative? Or that being at Man City is positive?

Lurking variables

We don’t think there’s room for any variables to lurk here, though we might think about this being a model of forward players – presumably goalies would need a different model.