Assignment Overview

**Week 3 Assignment**

In this week, we generated a forecast model for the EPL and compared its performance to that of the bookmakers’ odds. While the bookmakers’ predictions were slightly better, the difference was not great, whether measured by number of correct predictions of the result, or Brier score which measures the overall accuracy of the probabilities.

For the assignment we will repeat this exercise for the Bundesliga in the 2019/20 season. We will use the data for the season up to March 11 (the last game before the COVID-19 lockdown began) to predict the results for the games played when the season resumed in May.

To answer the questions you need to complete the following steps:

**Step 1: Data preparation**

1. Load the data
2. Define the result (H, D, A) and the probabilities associated with bookmaker odds
3. Define a winvalue = 2 if the home team wins, 1 if the game is a draw and zero otherwise
4. Define a variable equal to H if the home team wins and A if the visiting team wins
5. Generate a crosstab to compare the predictions of the bookmaker with the actual outcomes
6. Define a variable equal to the logarithm of the ratio of the home team TMvalue to the visiting team TMvalue.

**Step 2: Estimate a logit model of home time wins depending on the log TMvalue ratio, using the data for gamenos 1 to 224 as the “training data”.**

1. Define a subset for games 1 to 224
2. Import the ordered logistic regression package
3. Run the ordered logistic regression of winvalue on the log TMvalue ratio

**Step 3: Define the predicted probabilities and the predicted results, using the entire data set**

1. Define the predicted probabilities from the logit regression.
2. Based on the predicted probability, define the predicted result (H, D or A).
3. Based on the predicted outcome, create a dummy variable = 1 when the prediction is correct, and zero otherwise.
4. Define three dummy variables for actual outcomes (H, D, A)

**Step 4: For games played from May 2020, compare the bookmaker probabilities and model probabilities in terms of the mean number of successfully predicted outcomes and the Brier scores.**

1. Define the subset of games played from May onwards
2. Define a dummy variable equal to 1 when the bookmaker result prediction is correct, and zero otherwise.
3. Calculate the means for each of these variables
4. Define the Brier score for the bookmaker probabilities and the Brier score for the logit model probabilities
5. Calculate the mean of each Brier score

**Beware**: even though your code might get you to the correct answer at a given point, it is sometimes possible that the way you write it might interfere with completing a further step. So even if you get the answer right, you should look at the code we supply to check if you are going the same way. In practice, there are often many ways to get to answer in Python, and we do not insist that you follow our approach exactly – but simply warn you to be aware that differences could turn out to be problematic later.