Analyzing Sports Data Pt. 2

Ken McAlinn
Joe Futoma

Predicting Results

- Predicting wins is the ultimate goal for most analysts
- A good predictive model should be accurate and robust
- What kind of method should one use?
  - What kind of model?
    - Linear regression? Multiple regression?
  - What kind of data should one use?
    - What should the data length be?
    - How do we define a win?
  - Which stats are useful in predicting wins?
    - Offensive stats or defensive stats?

Multiple Regression

- For your last HW you used a basic linear regression
  - Can we do this for predicting wins?
- We shouldn't be able to answer something as complex as a team's performance based on any one statistic
- (One) solution: Multiple regression

Model Building

- Which factors best represent a team's strength?
- Given N stats, we have 2^N number of linear models
- How can we think about model building without going through every single model (you can if you want)?
- Optimal model for a given set of data:
  - Selection criteria: Adj-R^2, AIC, …
- Optimal model for predictions:
  - Selection criterion: RMSE, …

AIC – Akaike Information Criterion

$$AIC = -2\ln L + 2k$$

- Used for model selection, lower the better
- *L*: likelihood of the model – likelihood of seeing these data given the model parameters
- *ln*: natural log
- Applies a penalty for number of parameters in the model, k
- R: `AIC(model)`

RMSE – root mean squared error

$$\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} = \sqrt{\frac{RSS}{n}}$$

- Standard in predictive analysis
- RSS (residual sum of squares)
- Can lead to over-fitting if used as a criterion for model selection
- Why do we prefer RMSE over MSE?
  - Why do we prefer standard deviation over variance?
- R: Remember `lm()` calculates the `fitted` values of the response variable.

Modeling Temporal Data

- The effect of a statistic may change over time
- How relevant is the data from 2005 for predicting tomorrow's game?
- Rolling Window:
  - Use past k games for each prediction
  - To predict game T, just use some of the variables from games T-1,T-2,…,T-k
- Weighted Average:
  - Use a weighted average of the stats from past k games for prediction
    - Same as rolling window, but instead of using the actual variables from last game, now take a weighted average
    - Weight all games equally?
    - Weight for game T-1 > T-2 > … > T-k (downweight older games)?

Defining a Win

- How should we define a win?
  - Win/loss or point spread
- Predicting win/loss (i.e., 1/0)
  - Logistic regression
  - Hard to predict
  - No information on how much they won by
- Predicting point spread
  - Linear regression
  - Fairly easy to predict
  - Probability of winning can be defined as p[point spread>0]

Caution

- Make sure you are using past stats to predict!
  - Do NOT build a model using that game's stats
  - After game T, the stats for game T+1 are unknown
- How to build a good predictive model
  1. Split your data into two: training + validation
  2. Fit models on set of training data
  3. Then test your models on the validation set (e.g. using RMSE)
  4. Repeat steps 2 & 3 until you have a model you like
  5. Train your final model on the training + validation sets
  6. Evaluate final model on a new test set, distinct from both the training and validation sets from step 1

Application Exercise 14

- **Task:** Build and compare models
  - Compute the point spread for that season for the 76ers (PHI) and the Celtics (BOS)
  - Use the point spread as the response variable
    - Will be unable to predict game 1, possibly more depending on stats chosen
  - Build three models: (1) using offensive stats (o), (2) using defensive stats (d), and (3) using what you think is the best combination of stats
  - Compare the three models (using RMSE, Adj-R^2, AIC)
  - Determine which model is best according to each of these 3 evaluation metrics. Is it always the same one?
  - Report which stats were good predictors for the two teams
    - Any stats useful for one team but not the other?
- **Data:** https://stat.duke.edu/courses/Fall14/sta112.01/data/NBA1011.html

HW 5

2/3 through the season (immediately after game 62), you are hired by the Boston Celtics to predict the point spreads for the remaining 20 games in the season

- Build a model using the variables given
- Consider augmenting the data for the task (weighted average or rolling window)
- Use games 1-52 as a training set, games 53-62 as a validation set, and games 63-82 as your final test set
- Report:
  - RMSE, Adj-R^2,AIC of final model on the training set
  - RMSE on validation set
  - RMSE on test set
  - Accuracy of win/loss predictions on test set (treat positive point spread as win)
- Be creative! Use opponent stats, outside data etc.
- Winning team (team with lowest RMSE on test set) gets StatSci t-shirts!

HW5 & Office Hours

- HW5 can be found at https://stat.duke.edu/courses/Fall14/sta112.01/hw/hw5.html and is due next Thursday.
- Joe & Ken, Old Chem 211:
  - Wednesday 4:45 - 5:45pm
- Dr. Çetinkaya-Rundel (adjusted to not overlap):
  - Wednesday 11:30am – 12:30pm
  - by appointment