Predicting league outcomes in soccer: Combining ML, non-linear econometrics & simulation

Michael Lechner (jointly with **Michael Knaus, Alex Krumer, Daniel Goller**)
Swiss Institute for Empirical Economic Research (SEW)
University of St. Gallen | Switzerland | August 2017

Swiss Institute for
Empirical Economic Research

University of St.Gallen

1 | Introduction & general set-up

2 | Econometrics: 3 components

3 | Data: BL 1

4 | Results: Past & future

5 | Summary & outlook

Swiss Institute for
Empirical Economic Research

University of St.Gallen

1 | Introduction & general set-up

2 | Econometrics: 3 components

3 | Data: BL 1

4 | Results: Past & future

5 | Conclusions & further research

Goal

Use econometrics to predict outcomes of soccer league table

- Goal is the ranking at the end of the season (and its uncertainty)
  - not the outcome of a particular match (by-product)
- Update simulations every week to incorporate new information about past match outcome
- Update estimation and simulation at end of the 2 transfer windows

Why are we doing this?

Show potential of modern econometric methods in a difficult predictive situation

- Soccer games are notoriously difficult to predict
- Dynamics exist
- Difficult variable selection problems

Fun

Useful for illustration of various data science methods in teaching

General set-up of prediction exercise

Set-up

- Predict outcome of match
- Aggregate matches over season
- Account for randomness of outcomes by simulation methods
- Produce probabilities of a team reaching a certain rank

Econometrics

- Ordered logit model for predicting matches
- Machine learning for variable selection in ordered logit

The plan of the talk

Introduction

Econometrics

Data

Results & predictions

Conclusions & Outlook

Swiss Institute for
Empirical Economic Research

University of St.Gallen

1 | Introduction & general set-up

2 | Econometrics: 3 components

3 | Data: BL 1

4 | Results: Past & future

5 | Conclusions & further research

Prediction model |1

What to predict?

- Match outcome: 'Unit of observation' (34 rounds x 9 matches x years)
  - Alternative: Past ranking for specific club → too few observations (18 clubs x years)
- Target of interest: Expected points
  - Goal difference is ignored

What outcome to predict?

- Goal difference of home and away team
  - Most difficult to estimate, but most informative for predicting expected points
- Points
- As a compromise between complexity and information loss, we use 5 categories: high win (2+), narrow win, draw, narrow loss, high loss

Prediction model |2

How to predict match outcome with 5 categories?

- Nonparametric approach
  - Too flexible for prediction excercises (→ curse of dimensionality)
- Direct machine learning approach: RF or similar may work, but untested for this approach
- Ordered probit / logit model with unknown bounds is obvious candidate and used here
  - Big advantage: Easy computation of *E(points=3/1/0|X)*
  - Problem: Functional form dependent → make it flexible by including many interactions, powers, dummies, etc.
  - Needs formal variable selection procedure

Variable selection for prediction model|1

We have very many variables plus lots of interactions, powers, log's etc. → need formal procedure for variable selection

Use procedure for ordered logit: (currently) not available

Two possible approximations

- Every ordered model can be split into binary models (by aggregating categories) → use procedures for binary models several times
- Ordered probits/logits with 5 cat's are not so different to linear models → we use linear model for variable selection and then feed selected variables into ordered model to get predictions

A brief review of shrinkage and variable selection

There are many different ways to deal with this problem, …

- …
- Least absolute shrinkage and selection estimator (LASSO)
  - Adds a linear penalty term to OLS objective function penalising abs. large coefficients
  - Shrinks coefficients of less important variables to zero → explicit variable selection
- Ridge regression: No variable selection, but useful to start with as it can be combined with LASSO

Ridge regression|1

$$\hat{\beta}^{Ridge} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - x_i \beta \right)^2 + \lambda \underbrace{\sum_{j=1}^{p} \beta_j^2}_{= \|\beta\|_2 \quad \ell_2 - norm} \quad ; \quad \lambda \geq 0$$

λ is penalty term

- If $\lambda = 0$: OLS
- If $\lambda \rightarrow \infty$: Only constant term remains in model
- Penalty term does not apply to constant term
  - Estimate constant directly and recenter covariates

$$\hat{\beta}_{const}^{Ridge} = \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i - \bar{x} \hat{\beta}_{1:p}^{Ridge}; \quad \hat{\beta}_{1:p}^{Ridge} = (\tilde{X}' \tilde{X} + \lambda I)^{-1} \tilde{X}' \tilde{y};$$

$$\tilde{x}_j = x_j - \bar{x}_j, \quad \tilde{y} = y - \bar{y} \quad .$$

Ridge regression |2

Result of ridge regression is measurement dependent (not scale invariant like OLS)

- Usually plausible to standardise covariates by dividing them by the standard deviation

Estimates are shrunken towards zero

- Leads to bias, but also reduced variance

It works even if variables are perfectly correlated

- They will then ‘share’ the effect
- This makes the interpretation of the Ridge coefficients more difficult

Ridge regression |3

No explicit variable selection

- None of the coefficients will be exactly zero

Estimates are biased towards 0 but have smaller variance than OLS

- Therefore, Ridge regression may have a smaller RMSE than OLS

Optimal value of penalty term can be determined with grid search by CV

- There are also asymptotic formulae available that maximise some sort of likelihood based information criteria

Lasso |1

$$\hat{\beta}^{Lasso} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - x_i \beta \right)^2 + \lambda \underbrace{\sum_{j=1}^{p} \left| \beta_j \right|}_{= \|\beta\|_1 \quad \ell_1 - norm} \quad ; \quad \lambda \geq 0$$

λ is penalty term

- If $\lambda = 0$: OLS
- If $\lambda \rightarrow \infty$: Only constant term remains in model
- Penalty term does not apply to constant term
  - Estimate constant directly and recenter covariates

$$\hat{\beta}_{const}^{Lasso} = \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i - \bar{x} \hat{\beta}_{1:p}^{Lasso}$$

No closed-form solution for slope coefficients

- Computation of slope coefficient is more involved than for Ridge regression
- Estimator becomes non-linear in y → not covered by the Gauss-Markov theorem

Lasso |2

Result of Lasso regression is measurement dependent (not scale invariant like OLS)

- Standardise covariates by dividing them by their standard deviation

Estimates are shrunken towards zero

- Many coefficients will be exactly zero: *Explicit variable selection*
  - Lasso yields *sparse* models
  - This is different to Ridge regression where none of the coefficients becomes 0
- Eases interpretability of estimates to some extent

If variables are perfectly correlated: One variable will receive non-zero coefficient, the others shrunken to zero.

Lasso |3

p >> *N* (ultra-high dimensional models) possible

Estimates are biased towards 0 but have smaller variance than OLS

- Therefore, Lasso regression may have a smaller RMSE than OLS

Optimal value of penalty term can be determined with grid search by CV

- Alternatively, formulas based on various likelihood-based information criteria are available

Alternative interpretation of Lasso & Ridge regr.

$$\hat{\beta}^{Ridge} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - x_i\beta\right)^2 \quad subject \quad to \quad \sum_{j=1}^{p} \beta_j^2 \leq s; \quad s \geq 0$$

$$\hat{\beta}^{Lasso} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - x_i\beta\right)^2 \quad subject \quad to \quad \sum_{j=1}^{p} \left|\beta_j\right| \leq s; \quad s \geq 0$$

There is exactly one value of *s* that corresponds to one value of λ

Lasso and Ridge regression are least squares estimates in a subspace of the coefficient space.

Lasso: Assumptions and properties |1

Under an assumption called 'sparsity' (& some more regularity assumptions), Lasso identifies the correct set of non-zero coefficient (even if $p >> N$) ➔ Oracle property

- LASSO does not necessarily find the 'true' model (asymptotically), but the true model will be a submodel of the model with non-zero LASSO coefficients

Non-technical formulation of the sparsity assumption

- OLS coefficients of variables not belonging to the model are not too large (in regressions NOT containing all variables that belong to the true model)
- This essentially means that variables not belonging to the model should NOT be too highly correlated with variables that belong to the true model

Adaptive Lasso |1

$$\hat{\beta}^{Adaptive\ Lasso} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - x_i\beta\right)^2 + \sum_{j=1}^{p} \underbrace{\lambda_j}_{\cdots \left|\frac{\lambda}{\hat{\beta}_j^{Lasso}}\right|} \left|\beta_j\right|$$

Adaptive Lasso adresses issue that Lasso usually selects too many variables

- Penalty covariate specific (larger penalty for coefficients with 'true values' close to 0)

Numerically, it can be easily implemented as a 2-stage procedure

- Estimate a standard Lasso
- Rescale covariates using the (Post-) Lasso coefficients
- Estimate a standard Lasso with the rescaled variables

Under (strong) sparsity assumption, ALASSO will find exactly the true model (asymptotically)

Post-Lasso |1

(Adaptive) Lasso coefficients are biased towards zero, but Lasso will identify the correct set of non-zero coefficients (Oracle property)

This suggests using the following two stage estimator

- 1st stage: Standard Lasso (or Adaptive Lasso)
- 2nd stage: OLS using the variables with non-zero coefficients of 1st stage

Elastic Net |1

Lasso is expected to do better than Ridge regression in *sparse* models, while Ridge regression does better in 'richer' models

Elastic Net combines these properties to allow for 'relatively' sparse settings by combining the penalty terms

$$\hat{\beta}^{EN} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - x_i\beta\right)^2 + \lambda_1 \sum_{j=1}^{p} \left|\beta_j\right| + \lambda_2 \sum_{j=1}^{p} \beta_j^2; \quad \lambda_1, \lambda_2 \geq 0$$

Find optimal values of λ_1, λ_2 by CV using grid search on a 2-dimensional grid

Simulation model|1

How do we use the estimated model?

- Predict expected points for all matches in round t+1 given the data up to round t
- Do the same for t+2 given information up to t+1
  - If t+1 is not yet available, use expected values → dynamic simulation
- Aggregate the points over all matches for all clubs to derive ranking after each round until the end of the season

Simulation model|2

A note on past outcome variables in the dynamic simulation

- We only take into account the predicted previous game outcomes (in terms of point) as they are a direct outcome of the model anyway
- The other 'outcome' variables (red cards, etc.) are taken from the previous season and are thus fixed in the simulation
  - Predicting them would require a prediction model for each of them, all of these model needs to link as all these outcomes are related
  - This leads to issue of likely biased prediction for these variables, but in particularly it will substantially inflate the variability of the prediction (which is very undesirable)

Simulation model|3

How do we capture that game outcomes cannot be perfectly predicted ('luck' is important in soccer)? ➔ stochastic simulation

- Model allows to predict probabilities for win/draw/loss
- Draw random number (3/1/0) according to these probabilities
- Aggreate these simulated game outcomes over the season
- Compute final ranks
- Do this many, many times (10'000+)
- Compute rank distribution for each club based on these simulated ranks

Final outcome of simulations

Expected number of points for each club for each round

Expected rank distributed showing the intrinsic uncertainty in the soccer game

- Allows also to derive probabilities for a club to become champion, qualify for the champions league or get relegated, etc.
- Caveat: Uncertainty in the estimation process is ignored
  - Observed games are treated as population and not as sample
  - Model selection uncertainty (tuning parameter selection etc.) is ignored

Swiss Institute for
Empirical Economic Research

University of St.Gallen

1 | Introduction & general set-up

2 | Econometrics: 3 components

3 | Data: BL1

4 | Results: Past & future

5 | Conclusions & further research

Data: BL1

The model is implemented for the German Bundesliga 1

Could also be applied to other/lower leagues …

BL1 schedule

BL1 consists of 18 clubs

- Worst 2 or 3 clubs relegated to BL2 and substituted by best BL2 clubs

Round-robin system

- 34 rounds per season, 9 matches each round
  - break approx. in the middle (around Christmas, about 1 month)
  - Games in rounds 18-34 same as in rounds 1-17, but home games & away games are switched

Usually, 1 game on Friday (20:30), 5 games on Saturday (15:30), 1 game Saturday at 18:30, 2 games on Sunday (15:30 & 17:30)

Some games midweek on Tuesday & Wednesday at 20:00

- Reasons: Breaks for summer and winter 'holidays', international games (incl. tournaments over the summer)

Data

All Bundesliga games from season 2006/7 to 2016/7

- 2006/7 only used to build control variables
- 3060 observations before start of season 2017/8 (306 each season)
- Data base is updated weekly to include the latest game
- The schedule of 2017/8 is always part of the simulation (and has an impact via the dynamics)

Variables available | 1

Game level

- Timing
  - Game day, games after international break, round fixed effects, season fixed effects, etc.
- Attendance
  - Only used from last season
- Final score
  - Continuously used in various forms, updated in dynamic simulation
- Shots, shots on target, fouls, corners, cards for each team
  - Only used from last season
- Distance between cities of teams

Variables available | 2

Team level (for each season)

- Values of player and its distribution (www.transfermarkt.com)
- Age, height, and preferred feet of players
- TV revenues according to formula by DFL
- Performance in past season (points, shots, etc.)
- Change of coach (not predicted, although we easily (!!) could)
- Capacity of stadium
- Regional economic indicators

Interactions between team and game level data

Variables are usually computed as value for

- home team minus value for away team (or ratio)
- home team

Swiss Institute for
Empirical Economic Research

University of St.Gallen

1 | Introduction & general set-up

2 | Econometrics: 3 components

3 | Data: BL 1

4 | Results: Past & future

5 | Conclusions & further research

Results

Live estimation for current season with discussion of intermediate results

Go to website www.sew.unisg.ch/soccer_analytics

- Current predictions
- Validation
- Coaches

1 | Introduction & general set-up

2 | Econometrics: 3 components

3 | Data: BL 1

4 | Results: Past & future

5 | Conclusions & further research

Conclusions

Many, many possible improvements

- Alternative ML methods
- Better selection of tuning parameters

Many possible extensions

- Other leagues
- Other countries
- Other sports

Thank you for your attention!

Michael Lechner
University of St. Gallen | SEW
Michael.Lechner@unisg.ch
www.michael-lechner.eu

www.sew.unisg.ch/soccer_analytics