DATA 621—Assignment no. 1

Critical Thinking Group 2: All of our names, Ben H.

September XX, 2019

Contents

Just some preliminary notes on this data set:

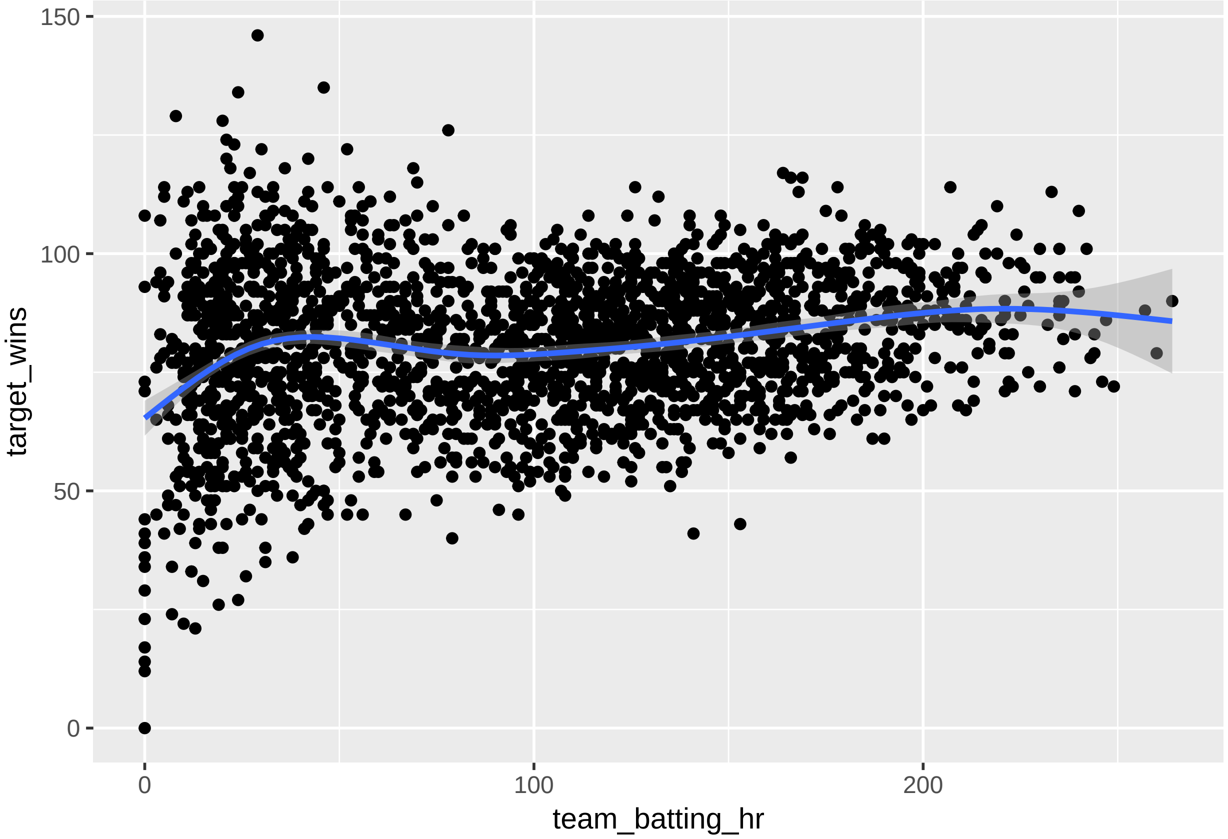
Load libraries:

```r
library(dplyr)
library(ggplot2)
library(knitr)
```

Load data:

```r
train <- read.csv('moneyball-training-data.csv', stringsAsFactors=FALSE)
colnames(train) <- tolower(colnames(train))

ggplot(train, aes(x=team_batting_hr, y=target_wins)) + geom_point() + geom_smooth()
```



- Target variable `target_wins` is normal, so that's good
- The following independent variables are noticeably skewed:
  - `team_batting_3b`

  - team_batting_bb
  - team_baserun_sb, with long right tail
  - team_baserun_cs
  - team_pitching_bb is very skewed, consider transformation
  - team_pitching_h, very skewed
  - team_pitching_e, very skewed

These variables almost appear as combinations of two distributions. If we could potentailly uncover the source of this divergence from the other variables—or maybe even use interaction variables—we could substantially improve modeling:

- team_batting_hr
- team_batting_so
- team_pitching_hr

Many of the variables are clearly related to the target variable. Some are clearly linear, others have clear non-linear relationships that should be accounted for in the model:

- All of the team_batting_XX variables
- team_batting_hbp
- team_pitching_hr
- Maybe team_fielding_e

These will have to be accounted for like `lm(y ~ poly(x, 2))`.

Outliers and missing values are definitely going to be an issue, so we'll need to come up with a strategy to understand why and deal with both of those.

Importantly – many of the independent variables are correlated with eachother, so we will have to be very careful about this or it will dramatically disturb our parameter estimates! (If we really wanted to, we could do PCA to make synthetic, non-correlated variables.)

We will have to decide how we want to evaluate each model, independently and in predicting on the training data. My vote is Root Mean Squared Error (RMSE).