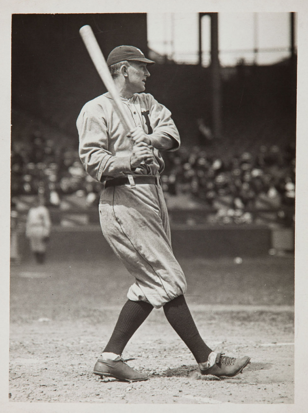
THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Model Building - Part 2

Lecture 15

STA 371G

Predicting Baseball Player Batting Averages



- load baseball.csv
- install the packages car, leaps, and corrplot if you haven't already

What predicts a player’s batting average

- All of the data here came from http://seanlahman.com/baseball-archive/statistics/

What predicts a player's batting average

- All of the data here came from http://seanlahman.com/baseball-archive/statistics/
- Some data cleaning, to calculate averages mostly, was done.

What predicts a player's batting average

- All of the data here came from http://seanlahman.com/baseball-archive/statistics/
- Some data cleaning, to calculate averages mostly, was done.
- We are going to explore this dataset with best subsets regression

The 10 potential x variables

- YEAR: Year this entry calculated for
- LG: League, either NL or AL
- AVG: Batting average
- OBP: On base percentage
- SLG: Slugging average
- EXP: Years of experience
- PAYR: Plate appearances per year
- MLAVG: Batting average for the leauge for the year
- MLOBP: On base percentage for the leaugue for the year
- MLSLG: Slugging percentage for the leaugue for the year
- AVGcumLag1: Player's cumulative batting average for previous years
- OBPcumLag1: Player's cumulative on base percentage for previous years
- SLGcumLag1: Player's cumulative slugging percentage for previous years
- G: Games played (must have been at least 98)
- YRINDEX: Number of years since 1958

Build model full and check for multicollinearity

```
full <- lm(AVG ~ OBP + SLG + EXP + PAYR + MLAVG
+ MLOBP + MLSLG + AVGcumLag1 + OBPcumLag1
+ SLGcumLag1 + G + YRINDEX, data=baseball)

round(vif(full),2)

       OBP        SLG        EXP       PAYR      MLAVG      MLOBP
      3.71       4.32       1.20       1.37      11.07      12.69
     MLSLG AVGcumLag1 OBPcumLag1 SLGcumLag1          G    YRINDEX
      7.39       2.09       3.95       3.82       1.12       2.18
```

Build model full and check for multicollinearity

```
full <- lm(AVG ~ OBP + SLG + EXP + PAYR + MLAVG
+ MLOBP + MLSLG + AVGcumLag1 + OBPcumLag1
+ SLGcumLag1 + G + YRINDEX, data=baseball)

round(vif(full),2)
```

```
       OBP        SLG        EXP       PAYR      MLAVG      MLOBP
      3.71       4.32       1.20       1.37      11.07      12.69
     MLSLG AVGcumLag1 OBPcumLag1 SLGcumLag1          G    YRINDEX
      7.39       2.09       3.95       3.82       1.12       2.18
```

Uh oh. Houston, we have a problem!

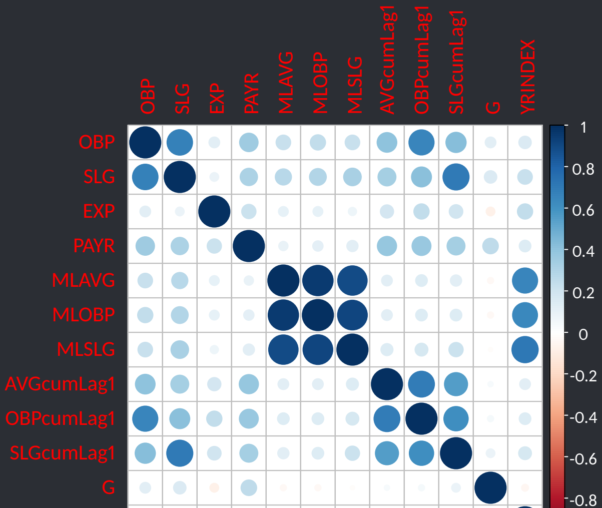
Look at the correlations to find the problem

This matrix is hard to read

```
numericpredictors <- baseball[,8:19]
M <- round(cor(numericpredictors),2) # calculate correlations

# print M by just typing M

# This table is confusing!
# There is a much better way to see this using corrplot
# So make the library available
library(corrplot)
```

Plot the correlations to better see the problem

```r
corrplot(M, method = "circle") #plot matrix
```



Reduce multicollinearity by dropping variables

- The Major League averages are highly correlated with each other
- Let's keep just MLAVG and drop MLOBP and MLSLG

```
full <- lm(AVG ~ OBP + SLG + EXP + PAYR + MLAVG
+ AVGcumLag1 + OBPcumLag1
+ SLGcumLag1 + G + YRINDEX, data=baseball)

round(vif(full), 2)
```

```
       OBP        SLG        EXP       PAYR      MLAVG AVGcumLag1 
      3.62       4.29       1.16       1.37       1.86       2.09 
OBPcumLag1 SLGcumLag1          G    YRINDEX 
      3.92       3.79       1.12       1.85 
```

Much better!

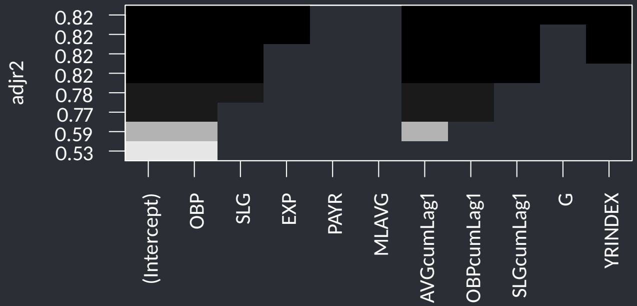
Run resubsets to get a sense of the best predictors

```
library(leaps)
bestsubsets <- regsubsets(AVG ~ OBP + SLG + EXP + PAYR + MLAVG
+ AVGcumLag1 + OBPcumLag1
+ SLGcumLag1 + G + YRINDEX, data=baseball)
```

Now let’s plot and identify the important predictors

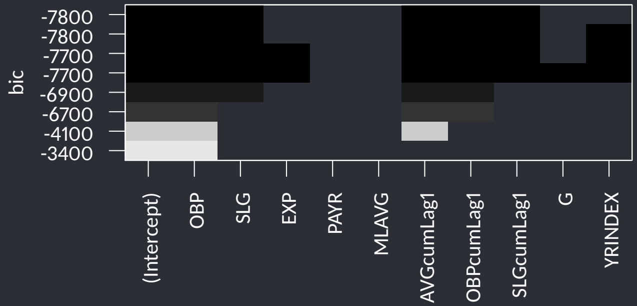
Use Adj R^2 to compare models

```
plot(bestsubsets, scale="adjr2")  # use adjusted R^2
```



Use BIC to compare models

```r
plot(bestsubsets, scale="bic")  # use BIC
```



Generate the best candidate model

```
model <- lm(AVG ~ OBP + SLG + AVGcumLag1 + OBPcumLag1
+ SLGcumLag1, data=baseball)

summary(model)
```

```
Call:
lm(formula = AVG ~ OBP + SLG + AVGcumLag1 + OBPcumLag1 + SLGcumLag1,
    data = baseball)

Residuals:
      Min        1Q    Median        3Q       Max
-0.056014 -0.007723  0.000263  0.008180  0.040508

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.027871   0.002497   11.16   <2e-16 ***
OBP          0.498209   0.009089   54.81   <2e-16 ***
SLG          0.160826   0.004697   34.24   <2e-16 ***
AVGcumLag1   0.880348   0.011950   73.67   <2e-16 ***
OBPcumLag1  -0.476261   0.012105  -39.34   <2e-16 ***
SLGcumLag1  -0.171831   0.005547  -30.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01213 on 4529 degrees of freedom
Multiple R-squared:  0.821,Adjusted R-squared:  0.8208
F-statistic:  4154 on 5 and 4529 DF,  p-value: < 2.2e-16
```

Does the National League’s Designated Hitter Rule matter?

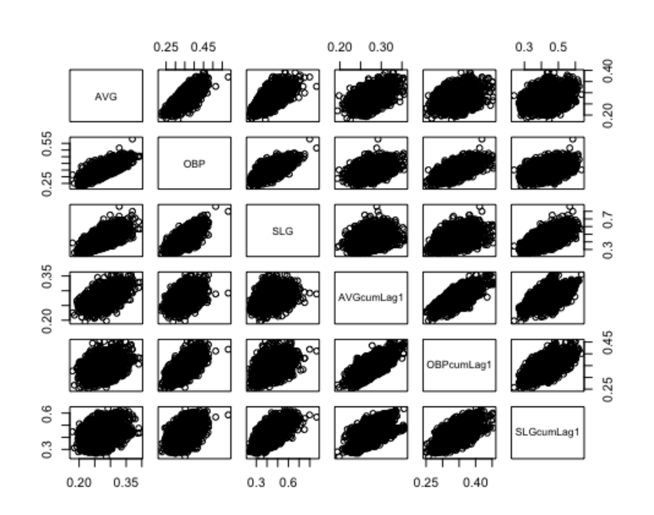
```
#Add a the categorical variable LG and find out!

model <- lm(AVG ~ OBP + SLG + AVGcumLag1 + OBPcumLag1
+ SLGcumLag1 + LG, data=baseball)
```

Does the National League's Designated Hitter Rule Matter?

```r
# Find the rows in baseball where LG is not either NL or AL
# and remove them so we can focus on the difference
# between NL and AL

base1 <- baseball[baseball$LG == "NL" | baseball$LG == "AL",]

modelLG <- lm(AVG ~ OBP + SLG + AVGcumLag1 + OBPcumLag1
+ SLGcumLag1 + LG, data=base1)

# Look at the summary, LG is not statistically significant
```

Check for linear relationships

```
# Depending on your computer, this command may run slowly
#pairs(~ AVG + OBP + SLG + AVGcumLag1 + OBPcumLag1 + SLGcumLag1, data=baseball)
```



Is this model really useful?

- Automated regression model selection methods cannot make something out of nothing.

Is this model really useful?

- Automated regression model selection methods cannot make something out of nothing.
- If you omit some important variables or fail to use data transformations when they are needed, or if the assumption of linear or linearizable relationships is simply wrong, the model is a bad one, no matter what the R^2.

Is this model really useful?

- Automated regression model selection methods cannot make something out of nothing.
- If you omit some important variables or fail to use data transformations when they are needed, or if the assumption of linear or linearizable relationships is simply wrong, the model is a bad one, no matter what the R^2.
- Use your own judgment and intuition about your data to try to fine-tune whatever the computer comes up with.

A challenge

```
# Load the dataset challenge and run the following regression

model <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7
     + x8 + x9 + x10 + x11 + x12, data=challenge)

    # Generate a summary and examine R^2
    # 0.21 of the variance is explained
```

Surprise!

```
# I created this data with a random number generator
# You may have to run it a couple of times to get significance

 y <- rnorm(100)
 x1 <- rnorm(100)
 x2 <- rnorm(100)
 x3 <- rnorm(100)
 x4 <- rnorm(100)
 x5 <- rnorm(100)
 x6 <- rnorm(100)
 x7 <- rnorm(100)
 x8 <- rnorm(100)
 x9 <- rnorm(100)
 x10 <- rnorm(100)
 x11 <- rnorm(100)
 x12 <- rnorm(100)
 summary(lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7
 + x8 + x9 + x10 + x11 + x12))


Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12)
```

Be careful of spurious correlations and overfitting!

- If you have more than 1 predictor for 10-15 y values, you are likely to see spurious correlations.

Be careful of spurious correlations and overfitting!

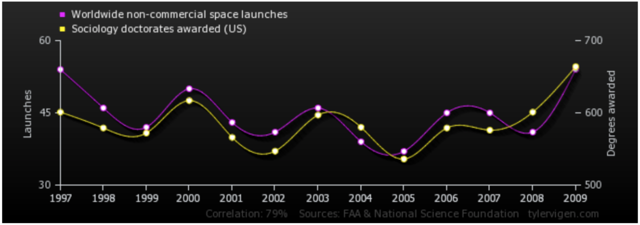
- If you have more than 1 predictor for 10-15 y values, you are likely to see spurious correlations.
- If you fit models with meaningless variables, you are fitting noise and will end up with an overfitted model that is not predictive going forward.

Be careful of spurious correlations and overfitting!

- If you have more than 1 predictor for 10-15 y values, you are likely to see spurious correlations.
- If you fit models with meaningless variables, you are fitting noise and will end up with an overfitted model that is not predictive going forward.
- You could even end up int he American Statistical Association's Hall of Shame!

www.tylervigen.com/spurious-correlations

Worldwide non-commercial space launches
correlates with
Sociology doctorates awarded (US)



	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Worldwide non-commercial space launches *Launches (FAA)*	54	46	42	50	43	41	46	39	37	45	45	41	54
Sociology doctorates awarded (US) *Degrees awarded (National Science Foundation)*	601	579	572	617	566	547	597	580	536	579	576	601	664

Correlation: 0.78915

- Don't fall for these!
- Have a great spring break!