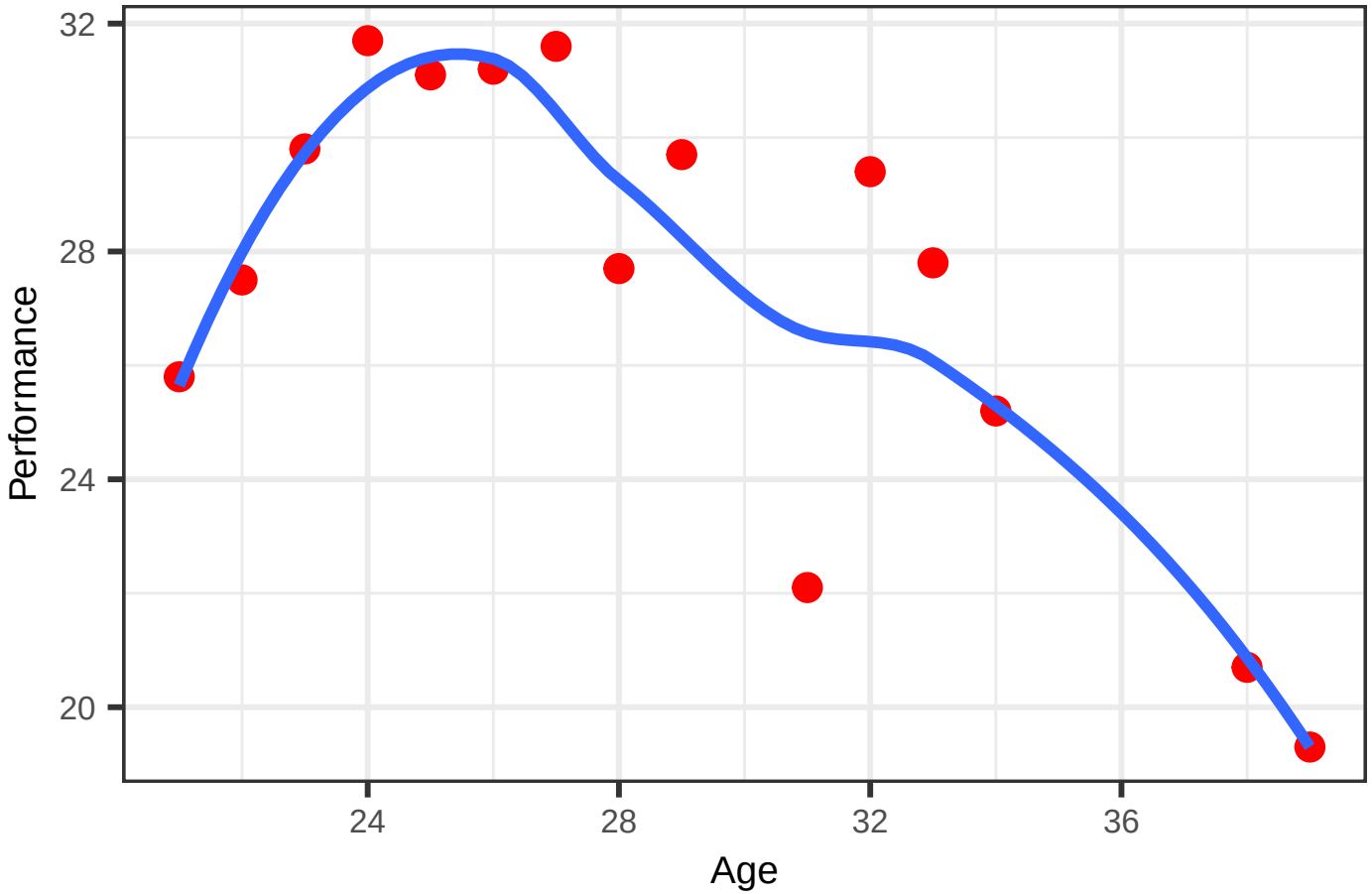
Aging Curve and Player Drop-outs

Quang Nguyen

February 26, 2021

Background

- Aging curve: performance throughout a career
- Ability vs. age
- Topics to explore: peak, rate of change, shape,. . .



Aging Curve in Sports Literature

- Morris (1983): Used parametric empirical Bayes procedures to estimate Ty Cobb's BA trajectory
- Albert (1992): Used a Poisson random effects model to smooth the career trajectory of a batter's HR rates.
- Albert (1999): Used a quadratic aging function to compare the best home run hitters.
- Berry et. al. (1999): Compared abilities of athletes from different eras in baseball, hockey, and golf.
- Schall and Smith (2000): Investigated baseball hitters and pitchers performance during the course of their careers. Used logit model to predict survival probabilities of players.

Aging Curve in Sports Literature

- Fair (2007): Estimated age effects in athletic events (track and field, swimming) and chess
- Fair (2008): Estimated age effects in baseball
- Wakim and Jin (2014): Functional Data Analysis of Aging Curves in Sports
- Vaci et. al (2019): Large data and Bayesian modeling - aging curves of NBA players

Player drop-out as a missing data problem

- Missing data: MCAR, MAR, MNAR
- Where could missingness occur for a player?
  - Beginning of career
  - End of career

Data

- Lahman database: `http://www.seanlahman.com/baseball-archive/statistics`
  - `Lahman` R package
  - `Batting` and `People` tables

```
library(Lahman)
```

```
colnames(Batting)
```

```
 [1] "playerID" "yearID"   "stint"    "teamID"   "lgID"     "G"
 [7] "AB"       "R"        "H"        "X2B"      "X3B"      "HR"
[13] "RBI"      "SB"       "CS"       "BB"       "SO"       "IBB"
[19] "HBP"      "SH"       "SF"       "GIDP"
```

```
colnames(People)
```

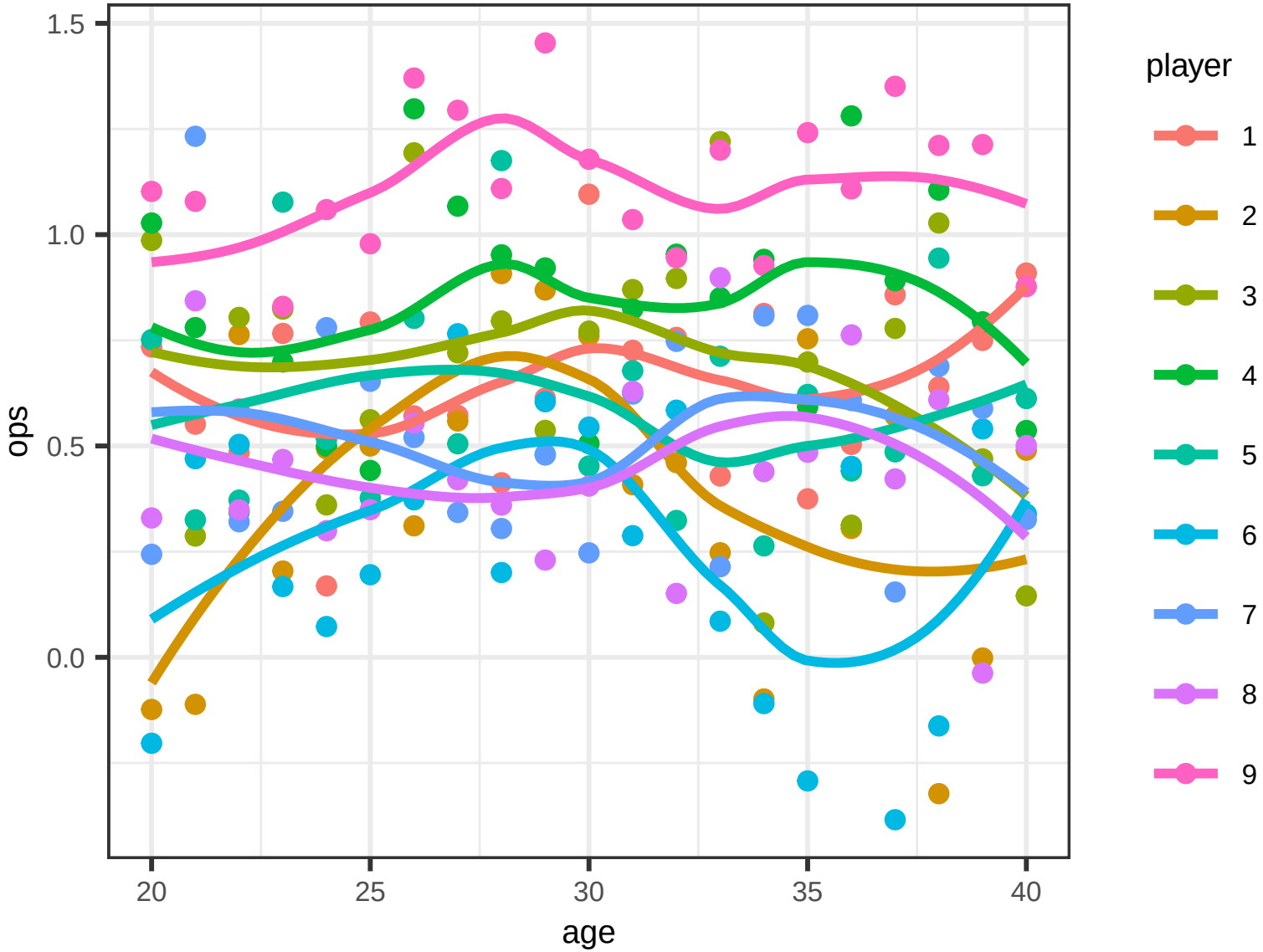
```
 [1] "playerID"     "birthYear"    "birthMonth"   "birthDay"     "birthCountry"
 [6] "birthState"   "birthCity"    "deathYear"    "deathMonth"   "deathDay"
[11] "deathCountry" "deathState"   "deathCity"    "nameFirst"    "nameLast"
[16] "nameGiven"    "weight"       "height"       "bats"         "throws"
[21] "debut"        "finalGame"    "retroID"      "bbrefID"      "deathDate"
[26] "birthDate"
```

Method

- Fit a model
- Generate “fake” players careers
- Create different drop-out mechanisms
- Impute the data
- Go back to the real MLB data

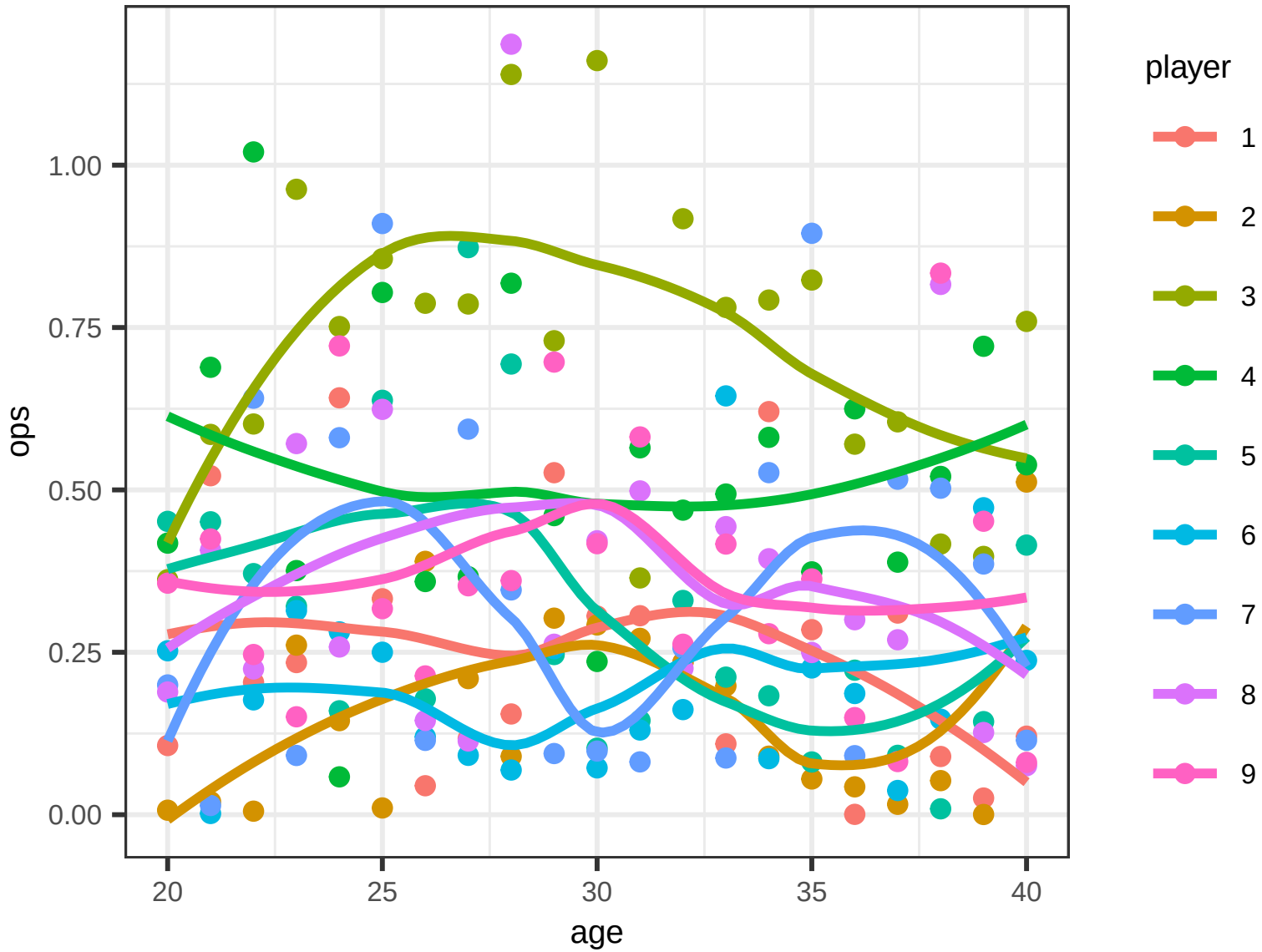
Model and Simulation

- Model: `OPS ~ poly(age, 3) + (1|playerID)`



A better approach...

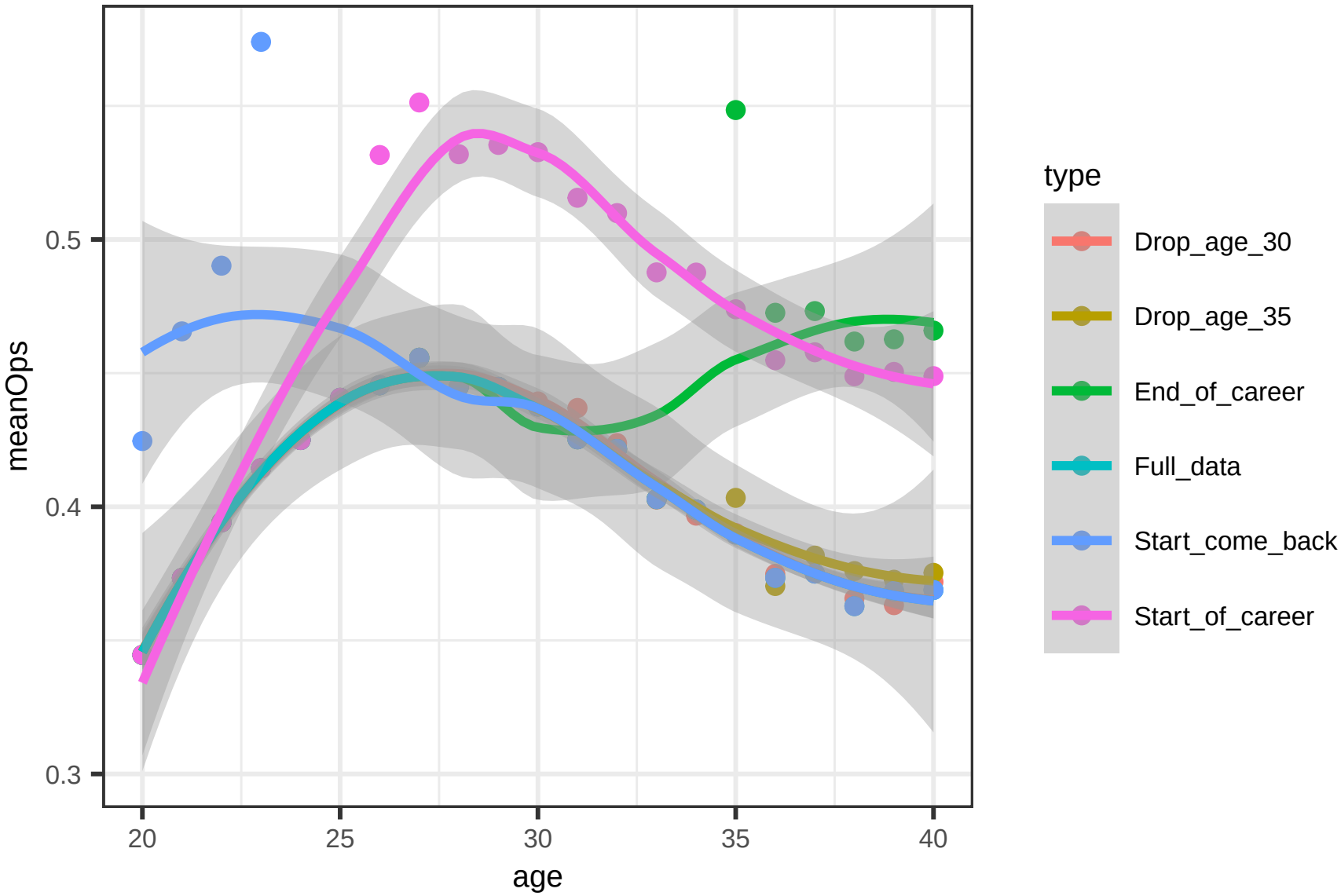
- Use a transformation: arcsin transformation



Drop-out rules

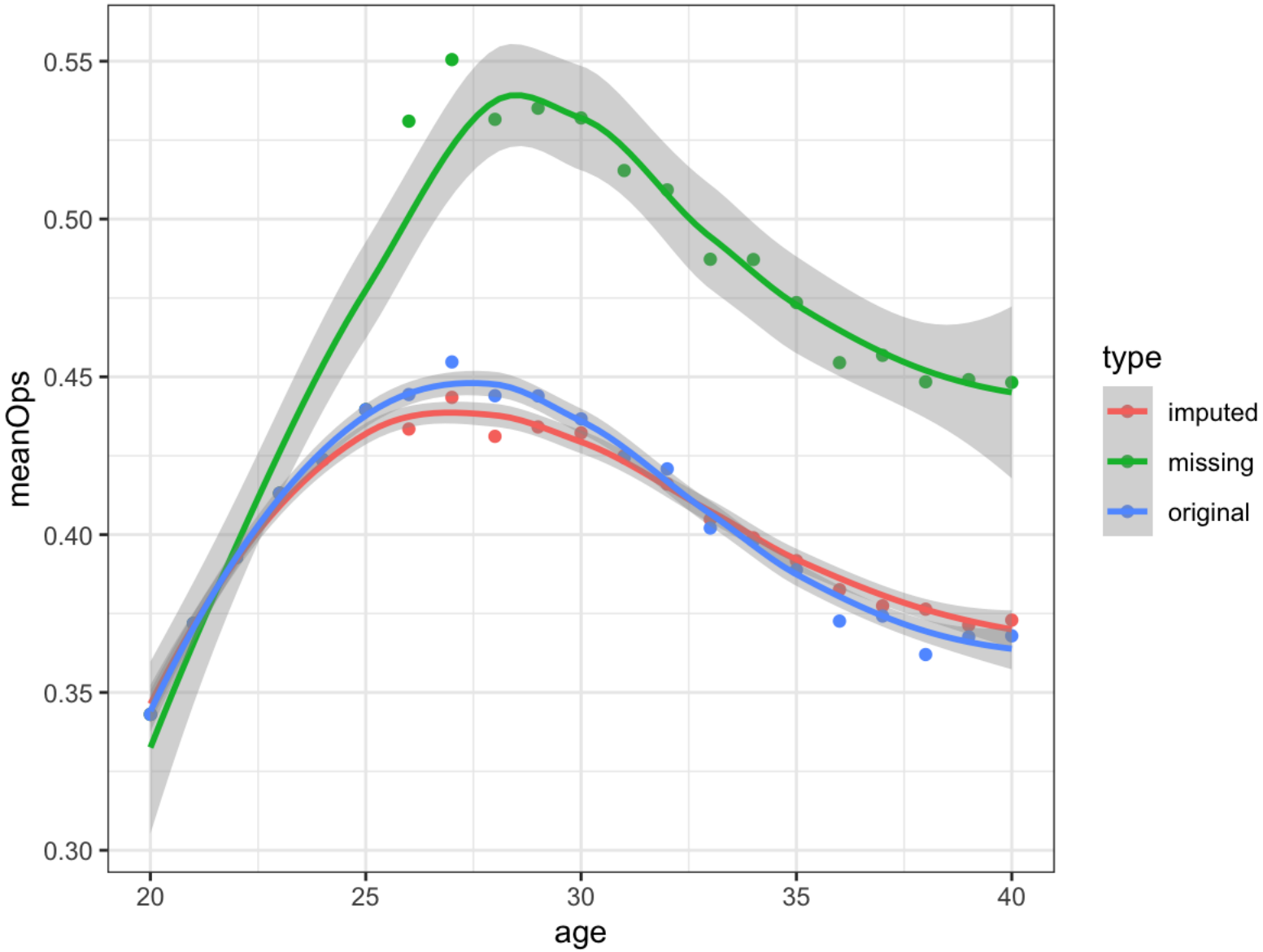
- If a player is 35 and his OPS is below a given threshold, he retires
- If a player is 25 and OPS is below a given threshold, then drop (i.e. player is not good enough to remain in the league)
- Drop players at the start of their careers due to low performance, but allow them to come back.
- At ages 30 and 35, 25% and 35% of players randomly retire.

Comparing drop-outs



Imputation

- Drop-out method: 25 low and out



More exciting things to come...

- Try with real MLB data
- Extend to other sports?
- GitHub repo: `https://github.com/qntkhvn/agingcurve`