Modelling umpire misclassification of balls and strikes using pitchFX data

Justin Post and Jason A. Osborne
Dept. of Statistics, N.C. State University

2013 New England Symposium on Statistics in Sports at Harvard University, Cambridge, MA

NC State University
DEPARTMENT OF STATISTICS

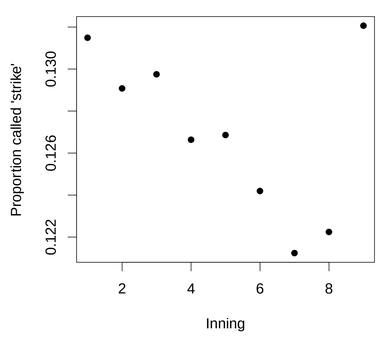
Introduction

- Watching MLB games on TV with pitch location graphics suggests that umpires occasionally call 'STEE-RIKE!' on balls outside the strikezone.
- Candidate factors affecting left-right "misclassication" according to Pitch FX determination of pitch location:
  - Horizontal location(`offplate`), • Pitchtype,
  - Velocity(`endspeed`), • inning, • umpire
  - (LHB/RHB) × (Inside/outside) combo (`stand*inout`)
- PITCH/fx data from almost all games of 2010-2012
- Approx 500000 pitches off the plate ($|px| > .85$)
- Logistic regression models (fit with SAS PROC GLIMMIX)

$$\pi_i = P(\text{"strike"}|\text{ball is inside or outside})$$
$$\log\frac{\pi_i}{1-\pi} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

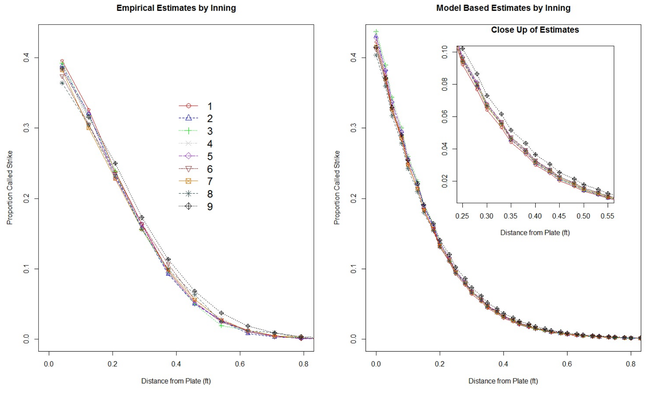
- ROC curves with PROC LOGISTIC.

Inning effects



$SE(\hat{p}) \approx .02$ What explains the ninth inning effect?

After building a model to control for other effects, (velocity, umpire, handedness, etc) effect of 9th inning diminished, but still highly significant ($p < .0001$):



Data Retrieval

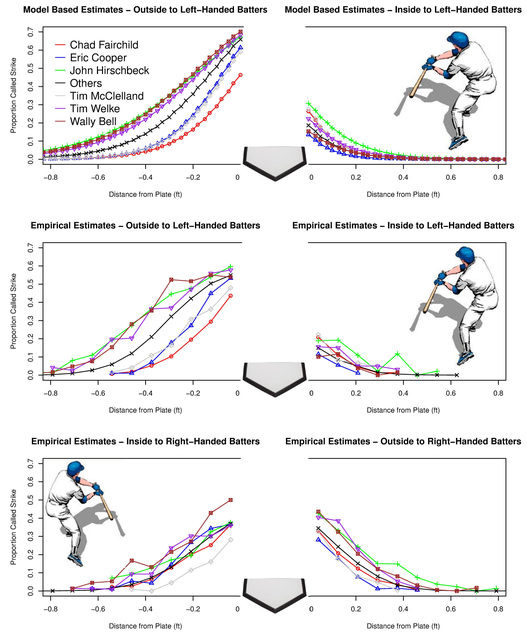
- Major League Baseball PITCHf/x data (pitch-tracking information developed by Sportvision)
- R code posted by user apeescape at `www.r-bloggers.com`. uses `xml` package to parse html

```
<inning num="4" away_team="sea" home_team="tex" next="Y">
 <top>
  <atbat num="33" b="2" s="1" o="0" start_tfs="202210" start_tfs_zulu="2010-04-11T20:22:10Z" batter="429711" stand="R"
   b_height="6-2" pitcher="444857" p_throws="R" des="Franklin Gutierrez singles on a line drive to center fielder Josh Hamilton. "
   des_es="Franklin Gutierrez pega sencillo con linea a jardinero central Josh Hamilton. " event="Single">
    <pitch des="Ball" des_es="Bola mala" id="287" type="B" tfs="202203" tfs_zulu="2010-04-11T20:22:03Z" x="62.66" y="142.47"
    sv_id="100411_152143" start_speed="87.1" end_speed="82.1" sz_top="3.31" sz_bot="1.33" pfx_x="-0.19" pfx_z="1.28" px="1.108"
    pz="2.626" x0="-1.919" y0="50.0" z0="6.116" vx0="7.75" vy0="-127.524" vz0="-2.972" ax="-0.311" ay="20.977" az="-29.941"
    break_y="23.9" break_angle="-1.6" break_length="7.3" pitch_type="FC" type_confidence=".861" zone="12" nasty="66"
    spin_dir="187.927" spin_rate="256.490" cc="" mt=""/>
    <pitch des="Ball" des_es="Bola mala" id="288" type="B" tfs="202219" tfs_zulu="2010-04-11T20:22:19Z" x="100.43" y="183.05"
    sv_id="100411_152159" start_speed="89.5" end_speed="85.2" sz_top="3.35" sz_bot="1.58" pfx_x="-6.38" pfx_z="2.07" px="-0.0090"
    pz="0.89" x0="-2.043" y0="50.0" z0="5.913" vx0="7.507" vy0="-130.779" vz0="-7.723" ax="-11.426" ay="18.936" az="-28.401"
    break_y="24.0" break_angle="19.6" break_length="7.0" pitch_type="SI" type_confidence=".756" zone="13" nasty="7"
    spin_dir="251.726" spin_rate="1334.849" cc="" mt=""/>
    <pitch des="Called Strike" des_es="Strike cantado" id="289" type="S" tfs="202238" tfs_zulu="2010-04-11T20:22:38Z" x="80.69"
    y="135.56" sv_id="100411_152218" start_speed="88.5" end_speed="85.3" sz_top="3.24" sz_bot="1.34" pfx_x="-3.38" pfx_z="4.23"
    px="0.585" pz="2.861" x0="-2.237" y0="50.0" z0="6.016" vx0="8.524" vy0="-129.407" vz0="-3.534" ax="-6.031" ay="13.528"
    az="-24.542" break_y="24.1" break_angle="11.9" break_length="5.7" pitch_type="FC" type_confidence=".663" zone="3" nasty="44"
    spin_dir="218.316" spin_rate="1090.528" cc="" mt=""/>
```

Screen-scraper loops over

- Years
- Months
- Days
- Games
- Innings
- At-bats
- Pitches

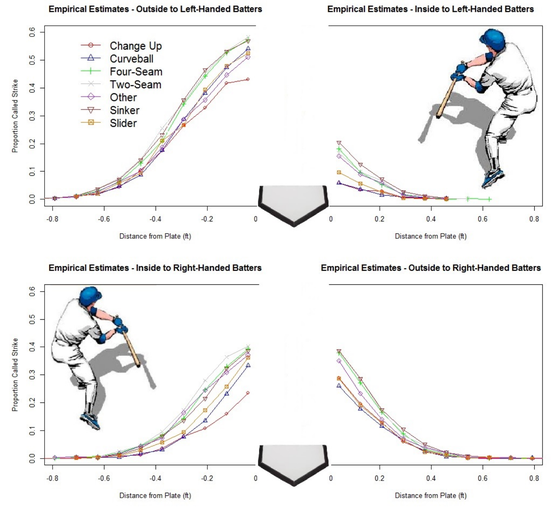
Umpire Effects



Umpire	Slope	SE	Left-handed bat: Pitches in Strikezone	Left-handed bat: Proportion called strike	Right-handed bat: Pitches in Strikezone	Right-handed bat: Proportion called strike
Eric Cooper	-6.5942^a	0.3001	1655	0.915	2351	0.917
Tim McClelland	-6.4030^a	0.3041	1630	0.890	2474	0.915
Chad Fairchild	-6.2003^a	0.3094	1603	0.893	2192	0.904
Others	-5.8268^{ab}	0.0470	108956	0.891	158879	0.907
Tim Welke	-5.1817^b	0.2026	1492	0.874	2302	0.908
Wally Bell	-5.0426^b	0.1920	1715	0.885	2251	0.917
John Hirschbeck	-4.7471^b	0.2386	917	0.893	1416	0.909

Cooper, McClelland and Fairchild call balls correctly most quickly as location moves away from plate, Welke, Bell and Hirschbeck not as quickly. Slope estimates average over levels of other factors like stand, inout, etc. Slopes w/ same letter not sig. diff.

Pitch Type Effects



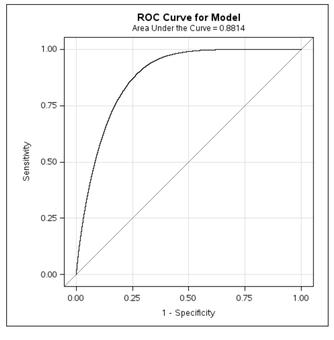
Pitchtype Slopes and intercepts

pitch	Intercept	SE	pitch	Slope	SE
CH	-1.0502^a	0.03875	FF	-6.1503^a	0.1031
CU	-0.9999^a	0.04085	SI	-6.0135^a	0.1200
SL	-0.7830^b	0.03499	FT	-5.9820^a	0.1177
other	-0.5452^c	0.03898	oth	-5.6957^b	0.1370
SI	-0.1658^d	0.03439	CU	-5.5371^b	0.1333
FF	-0.1644^d	0.02956	SL	-5.4708^b	0.1204
FT	-0.1416^d	0.03373	CH	-5.1463^c	0.1291

(Adjusted to avg inning, stand, inout, umpire, velocity)

Findings

- Fitted logistic regression model produces good ROC curve.
- All factors investigated exhibited highly significant associations with misclassification of balls horizontally outside of the strike zone as strikes, even inning, suggesting the possibility of umpire fatigue.
- Confirmed earlier findings regarding strikezone shift to the outside for LH batters.
- $\hat{\beta}_{vel} = -.026(SE = .002)$. Faster balls of given type less likely to be called strike.



Source	df	F
Inning	8	4.90
endspeed	1	276.15
pitch	6	209.98
stand	1	65.07
inout	1	1733.75
inout*stand	1	2488.98
inout*stand*ump3	308	7.41
offplate	1	10057.1
offplate*Inning	8	6.13
offplate*pitch	6	18.80
offplate*stand	1	30.77
offplate*inout	1	13.42
offplate*inout*stand	1	18.03
offplate*ump3	77	3.80

* each entry has $p < .0002$