Models do what they're told

Case Study: the Time Through the Order Penalty in Baseball

In Game 6 of the 2020 World Series, the Tampa Bay Rays' manager, Kevin Cash, pulled his starting pitcher, Blake Snell, midway through the sixth inning. When he was pulled, Snell had been pitching extremely well; he had allowed just two hits and struck out nine batters on 73 pitches. Moreover, the Rays had a one run lead. Snell's replacement, Nick Anderson, promptly gave up two runs, which ultimately proved decisive: the Rays went on to lose the game and the World Series. After the game, Cash justified his decision to pull Snell, remarking that he "didn't want Mookie [Betts] or [Corey] Seager seeing Blake a third time" (Rivera, 2020).

In his justification, Cash cites the third *Time Through the Order Penalty* (TTOP), which was first formally identified in Tango et al. (2007, pp. 187–190) and recently popularized by Lichtman (2013). It has long been observed that, on average, batters tend to perform better the more times they face a pitcher; for instance, they tend to get on base more often on their third time facing a pitcher than their second. Tango et al. (2007) quantified the corresponding drop-off in pitcher performance as increases in *weighted on-base average* (wOBA; see Section 2.4 for details). They observed that the average wOBA of a plate appearance in the first time through the order (1TTO) is about 9 wOBA points less than that in the second TTO (2TTO). Further, the average wOBA of a plate appearance in the second TTO is about 8 wOBA points less than that in the third TTO (3TTO) (Tango et al., 2007, Table 81).

wOBA overcomes many limitations of traditional metrics like batting average, on-base percentage, and slugging percentage. Briefly, batting average and on-base percentage treat all hits equally, with singles being worth as much as triples. Slugging percentage attempts to reward different types of hits differently, but does so in too simplistic of a fashion: in computing slugging percentage, a triple is worth three times what a single is worth. Such weighting is arbitrary, and is not tied to the relative impact of a triple over a single with regard to, say, run scoring or win probability. wOBA combines the different aspects of offensive production into one metric, weighing each offensive action in proportion to its actual run value (Slowinski, 2010). The wOBA of a plate appearance is simply the weight associated with the offensive action of the outcome. Specifically, the 2019 wOBA weight of each offensive action in decreasing order is 1.940 for a home run (HR), 1.529 for a triple (3B), 1.217 for a double (2B), 0.870 for a single (1B), 0.719 for hit-by-pitch (HBP), 0.690 for unintentional walks (uBB), and 0 for an out (OUT) (Fangraphs, 2021). wOBA is rescaled so that the league average wOBA equals the league average on-base percentage. Throughout this paper, we use 2019 wOBA weights for each season. Additionally, we usually refer to *wOBA points*, which is wOBA multiplied by 1000, to be consistent with the baseball community's use of wOBA.

The TTOP is considered canon by much of the baseball community. Announcers routinely mention the 3TTOP during broadcasts and several managers regularly use the 3TTOP to justify their decisions to pull starting pitchers at the start of the third TTO. For instance, A.J. Hinch, who managed the Houston Astros from 2015 to 2019, noted "the third time through is very difficult for a certain caliber of pitchers to get through." Brad Ausmus, who managed the Detroit Tigers from 2014 to 2017, explained "the more times a hitter sees a pitcher, the more success that hitter is going to have" (Laurila, 2015).

Let's dig into the data ourselves!

* Dataset: Every plate appearance i from 2018-2019 featuring a starting pitcher in the first 3 times through the order (214,386 plate appearances).

Outcome y_i = wOBA of i^{th} plate appearance (using 2019 wOBA weights).

batter sequence number $t_i \in \{1, 2, \dots, 27\}$

1st time through the order $t_i \in \{1, \dots, 9\}$

2nd TTO $t_i \in \{10, \dots, 18\}$

3rd TTO $t_i \in \{19, \dots, 27\}$

* EDA — Bin and average by TTO:

	ORDER_CT	mean_woba
	<dbl>	<dbl>
1	1	0.304
2	2	0.318
3	3	0.333

We observe that a starting pitcher performs worse on average in 2TTO than in 1TTO and worse on average in 3TTO than in 2TTO.

* Binning and averaging is equivalent to fitting the following regression model

$$\mathbb{E}(y_i|t_i) = \beta_1 \cdot \mathbb{1}\{t_i \in 1TTO\} + \beta_2 \cdot \mathbb{1}\{t_i \in 2TTO\} + \beta_3 \cdot \mathbb{1}\{t_i \in 3TTO\}.$$

$$\mathbb{1}\{z\} = \begin{cases} 1 & \text{if } z \text{ is True} \\ 0 & \text{if } z \text{ is False} \end{cases}$$

Math HW: Prove it via $\hat{\beta} = (X^T X)^{-1} X^T y$

```
> ### binning and averaging is equivalent to a fixed effects regression
> m1 = lm(EVENT_WOBA_19 ~ 0 + factor(ORDER_CT), data=df0)
> m1

Call:
lm(formula = EVENT_WOBA_19 ~ 0 + factor(ORDER_CT), data = df0)

Coefficients:
factor(ORDER_CT)1  factor(ORDER_CT)2  factor(ORDER_CT)3
           0.3044             0.3183             0.3326
```

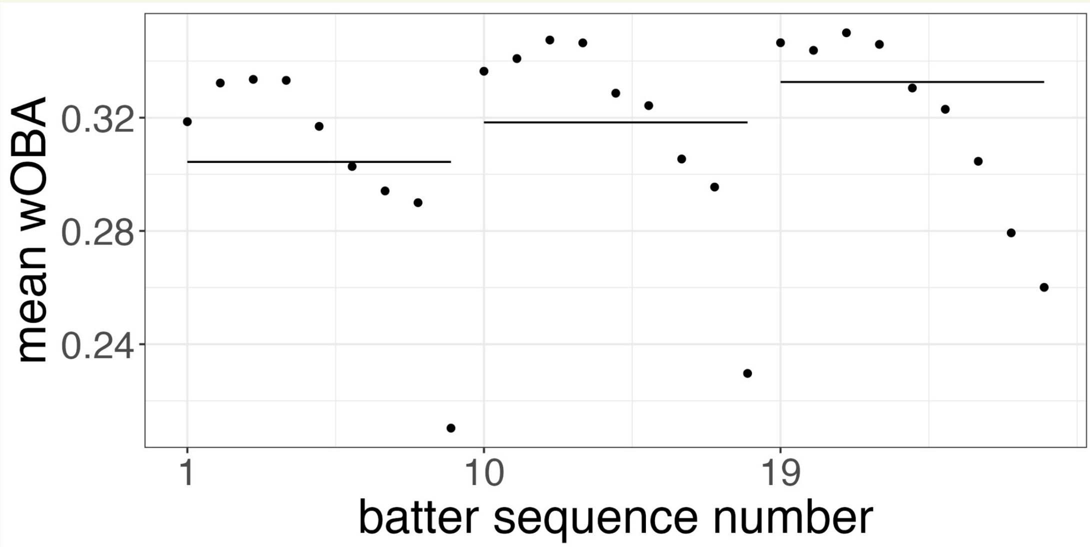
* Rephrase the model slightly so that β_2 represents pitcher decline from 1TTO to 2TTO and β_3 represents pitcher decline from 2TTO to 3TTO.

$$\mathbb{E}(y_i|t_i) = \beta_1 + \beta_2 \cdot \mathbb{1}\{t_i \geq 2TTO\} + \beta_3 \cdot \mathbb{1}\{t_i \geq 3TTO\}.$$

* What about the trajectory across the batter sequence number $t=1,\dots,27$?

Bin and average for each $t=1,\dots,27$,

or equivalently fit $\mathbb{E}(y_i|t_i) = \sum_{t=1}^{27} \beta_t \cdot \mathbb{1}\{t_i = t\}$.



Explain this shape.

What's wrong?

* Regression fits observed data; it is Not causation!

we need to adjust for confounders!

we need to disentangle the effect of pitcher decline within a game from

- batter quality BQ_i
- pitcher quality PQ_i
- handedness match $hand_i$
- home field advantage $home_i$

* For now, define a batter's quality by his end-of-season wOBA averaged across all his plate appearances and likewise for pitchers.

Later in this course we will learn a better way to estimate batter & pitcher quality that doesn't use future data: Empirical Bayes.

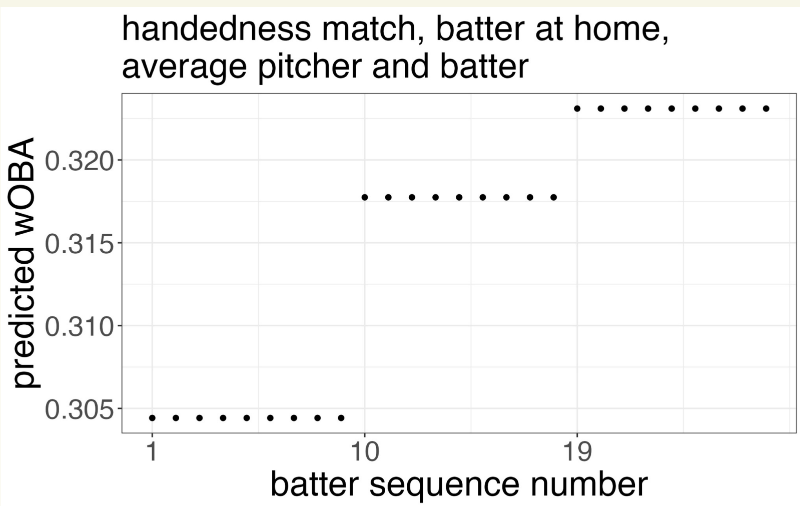
* Model that captures pitcher decline from one TTO to the next after adjusting for confounders:

$$
\begin{aligned}
\mathbb{E}(y_i|t_i) = \beta_1 &+ \beta_2 \cdot \mathbb{1}\{t_i \geq 2TTO\} \\
&+ \beta_3 \cdot \mathbb{1}\{t_i \geq 3TTO\} \\
&+ \beta_{BQ} \cdot BQ_i + \beta_{PQ} \cdot PQ_i \\
&+ \beta_{hand} \cdot hand_i + \beta_{home} \cdot home_i
\end{aligned}
$$

```
> m2 = lm(EVENT_WOBA_19 ~ 1 + as.numeric(ORDER_CT>=2) + as.numeric(ORDER_CT>=3) +
+                  HAND_MATCH + BAT_HOME_IND + WOBA_FINAL_BAT_19 + WOBA_FINAL_PIT_19,
+         data=df0)
> m2

Call:
lm(formula = EVENT_WOBA_19 ~ 1 + as.numeric(ORDER_CT >= 2) +
    as.numeric(ORDER_CT >= 3) + HAND_MATCH + BAT_HOME_IND + WOBA_FINAL_BAT_19 +
    WOBA_FINAL_PIT_19, data = df0)

Coefficients:
              (Intercept)  as.numeric(ORDER_CT >= 2)  as.numeric(ORDER_CT >= 3)              HAND_MATCH
                -0.299509                   0.013320                   0.005357               -0.016306
             BAT_HOME_IND          WOBA_FINAL_BAT_19          WOBA_FINAL_PIT_19
                 0.009988                   0.969370                   0.962359
```



* After adjusting for confounders, we still estimate that pitchers decline from one TTO to the next.

* Tom Tango's original analysis is based off of this model, and is a big part of the reason that starting pitchers are often removed in the 6th or 7th inning at the start of 3TTO.

Thoughts?

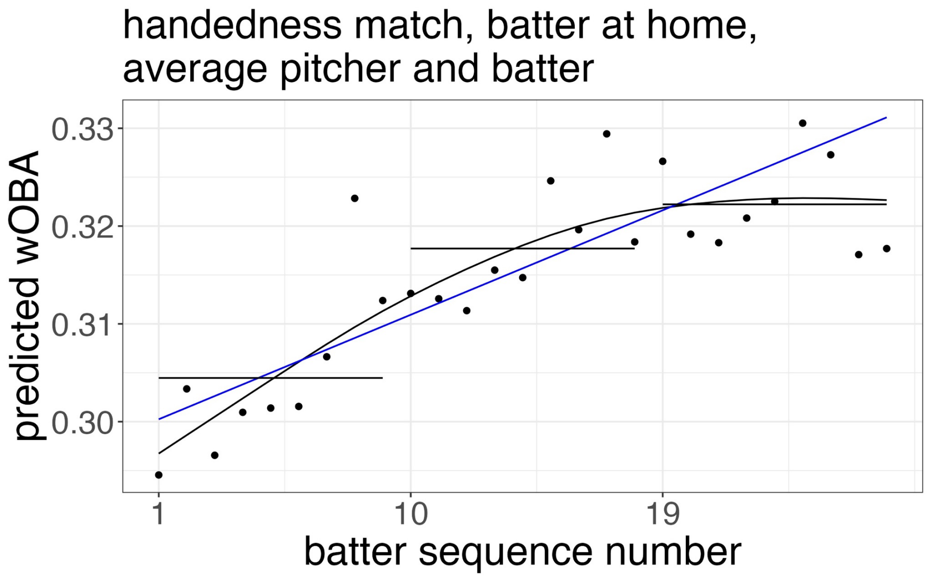
* After adjusting for confounders, what does the trajectory of pitcher decline look like across t (or over the course of the game)?

→ Indicator model:

$$\mathbb{E}(y_i | t_i) = \sum_{t=1}^{27} \beta_t \cdot \mathbb{1}\{t_i = t\} + \beta_{BQ} \cdot BQ_i + \beta_{PQ} \cdot PQ_i + \beta_{hand} \cdot hand_i + \beta_{home} \cdot home_i$$

→ Linear model:

$$\mathbb{E}(y_i | t_i) = \beta_0 + \beta_1 \cdot t_i + \beta_{BQ} \cdot BQ_i + \beta_{PQ} \cdot PQ_i + \beta_{hand} \cdot hand_i + \beta_{home} \cdot home_i$$



Black dot: β_t from indicator model
Black curve: smoothing spline over $\beta_1, \ldots, \beta_{27}$
3 Black lines: Mean β_t in each TTO
Blue line: $\beta_0 + \beta_1 t$ from linear model

* Pitchers do decline on average from one TTO to the next after adjusting for confounders as Tango showed (3 Black Lines), but these models Reveal that pitchers on average decline continuously from one TTO to the next.

Hence the rule of thumb to pull pitchers prior to the start of 3TTO doesn't make sense.

* Please note that regression is about fitting patterns from observational data and does NOT imply causation.

We're not saying anything about the causes of pitcher decline (fatigue or batter learning), we're only saying that after adjusting for confounders, pitchers appear to predominantly decline continuously, on average.

* But, a Potential cause of continuous pitcher decline is pitcher fatigue and a Potential cause of discontinuous pitcher decline is batter learning.

* After adjusting for confounders and controlling for continuous pitcher decline within a game, do pitchers decline discontinuously from one TTO to the next?

Model:

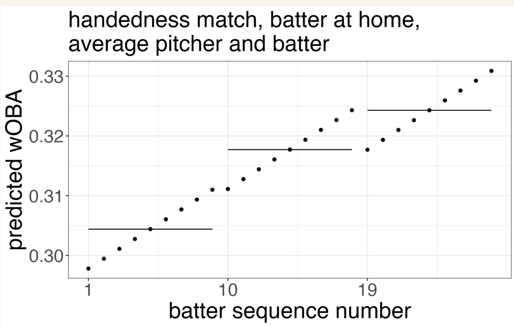
$$\mathbb{E}(y_i|t_i) = \beta_0 + \beta_1 \cdot t_i + \beta_2 \cdot \mathbb{1}\{t_i \geq 2TTO\} + \beta_3 \cdot \mathbb{1}\{t_i \geq 3TTO\} + \beta_{BQ} \cdot BQ_i + \beta_{PQ} \cdot PQ_i + \beta_{hand} \cdot hand_i + \beta_{home} \cdot home_i$$

```
> m5 = lm(EVENT_WOBA_19 ~ 1 + as.numeric(ORDER_CT>=2) + as.numeric(ORDER_CT>=3) + BATTER_SEQ_NUM + HAND_MATCH + BAT_HO
ME_IND + WOBA_FINAL_BAT_19 + WOBA_FINAL_PIT_19, data=df0)
> m5

Call:
lm(formula = EVENT_WOBA_19 ~ 1 + as.numeric(ORDER_CT >= 2) +
    as.numeric(ORDER_CT >= 3) + BATTER_SEQ_NUM + HAND_MATCH +
    BAT_HOME_IND + WOBA_FINAL_BAT_19 + WOBA_FINAL_PIT_19, data = df0)

Coefficients:
              (Intercept)  as.numeric(ORDER_CT >= 2)  as.numeric(ORDER_CT >= 3)             BATTER_SEQ_NUM
                -0.316611                  -0.001528                  -0.008262                   0.001649
               HAND_MATCH               BAT_HOME_IND          WOBA_FINAL_BAT_19          WOBA_FINAL_PIT_19
                -0.016837                   0.009994                   0.999090                   0.962273
```

We don't find statistical evidence.



Continuous Avg pitcher decline dominates discontinuous.

Takeaway: Models do what they're told!

If you only tell the model to look for discontinuous pitcher decline, then that is what you will find.

We find that the expected wOBA forecast by our model increases steadily over the course of a game and does not display sharp discontinuities between times through the order. Based on these results, we recommend managers cease pulling starting pitchers at the beginning of the 3TTO.