Beyond the Prediction Machines

The Role of Causal Inference in Sports Statistics

Sean Fischer, Ph.D.

Cincinnati Reds

2023-08-07 (updated: 2023-08-05)

Initial Thoughts

Sports as Applied Research Settings

- Sports provide a research setting very similar to other business domains
- Decision makers in sports face a variety of complex problems
- Just like in many other contexts, we want to know what will happen in the future

Power of Models

- Sports organizations employ teams of data scientists to build predictive models
- Forecasting is often the goal for these models
- Leads to a view of models as 'prediction machines'
- Teams and organizations are now more capable than ever of predicting performance or outcomes

Limits of Prediction Machines

- But prediction machines are inherently limited
- The limitations become apparent when we use models to address questions about cause and effect
- They will always struggle to generate estimates of debiased causal effects

Practitioners Need to Know More About the Effects of Decisions

- Coaches, support staff, and front-office staff all make decisions that generate consequences
- Data scientists can help support making better decisions by helping decision makers to understand the effects of their decisions
- Need to be able to help others make efficient and productive decisions

Causal Inference as A Solution

- By employing causal inference solutions, where appropriate, sports data scientists can give more accurate guidance to others
- But, experiments are often not feasible
- Need to produce experience with design-based methods of analysis

Case Study 1

Case Study 1: Mode of Travel Effects in Cross Country?

- The NCAA Championships in cross country are held at different venues year to year
- The NCAA pays for teams' travel
- For teams within 500 miles, covers the cost of driving
- For teams over 500 miles away, covers the cost of flying
- **Should teams consider spending limited budgets to cover the additional cost of flying?**

Case Study 1: The Available Data

- Bijan Mazaheri's LACCTiC site provides standardized race times
- Also includes predictions for what athletes should run
- Has results for the 2019, 2021, and 2022 NCAA D3 Championships
- Distances between schools and race courses estimated via Google Maps's Distance Matrix API

Case Study 1: The Usual Methods

- The most common data science approach to answering this question would be to build a regression model
- Still a regression model even if we use fancier methods to improve our estimates
- Include the treatment variable and some controls
- Perfectly fine approach if we want to describe the relationship between distance from the race course and performance

Case Study 1: The Usual Methods

```
## # A tibble: 9 × 5
##   term                        estimate std.error statistic  p.value
##   <chr>                          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                -0.240    0.0303       -7.90  8.80e-15
## 2 treat                       0.000934 0.00149       0.627 5.31e- 1
## 3 as.factor(race_year)2021    0.0118   0.00193       6.10  1.64e- 9
## 4 as.factor(race_year)2022    0.0282   0.00224      12.6   1.47e-33
## 5 yearSO                      0.00476  0.00391       1.22  2.24e- 1
## 6 yearJR                      0.00358  0.00392       0.912 3.62e- 1
## 7 yearSR                      0.00697  0.00394       1.77  7.71e- 2
## 8 yearGR                      0.0108   0.00421       2.56  1.07e- 2
## 9 lacctic_rating_2            0.000219 0.0000328     6.69  4.10e-11
```

Case Study 1: The Usual Methods

```
## # A tibble: 9 × 5
##   term                      estimate std.error statistic     p.value
##   <chr>                        <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept)              -0.261    0.0562       -4.64  0.00000690
## 2 treat                     0.00298  0.00272       1.10  0.274
## 3 as.factor(race_year)2021  0.00227  0.00353       0.641 0.522
## 4 as.factor(race_year)2022  0.0226   0.00422       5.35  0.000000288
## 5 yearSO                    0.00805  0.00589       1.37  0.174
## 6 yearJR                    0.00176  0.00607       0.290 0.772
## 7 yearSR                    0.00849  0.00587       1.45  0.150
## 8 yearGR                    0.0114   0.00637       1.80  0.0742
## 9 lacctic_rating_2          0.000250 0.0000621     4.02  0.0000874
```

Case Study 1: The Usual Methods

- This model is telling us that teams that are eligible to fly do not gain a measurable benefit over those that have to drive
- How confident are we that all the back door paths have been closed?
- We could keep adding controls or use more advanced regression techniques
- The rabbit hole of complexity
- Need to think about sources of bias, but also our ability to communicate the results

Case Study 1: A Causal Inference Approach

- But, we can also employ a design-based approach to solve this problem!
- We have a running variable in the distance variable with a natural cutoff
- Distance to the championship site has no real baring on whether teams qualify, so we have a source of as-if randomness!
- Do teams on either side of the cutoff perform differently?

Case Study 1: RDD Set Up

- This situation is really a regression discontinuity design!
- We can leverage the as-if randomness around the cutoff
- As long as there is balance/comparability between units on either side of the cutoff, the difference we observe is attributable to flying vs. driving
- Can use the `rdrobust` package in R to identify the best bandwidth for our analysis and to generate estimates

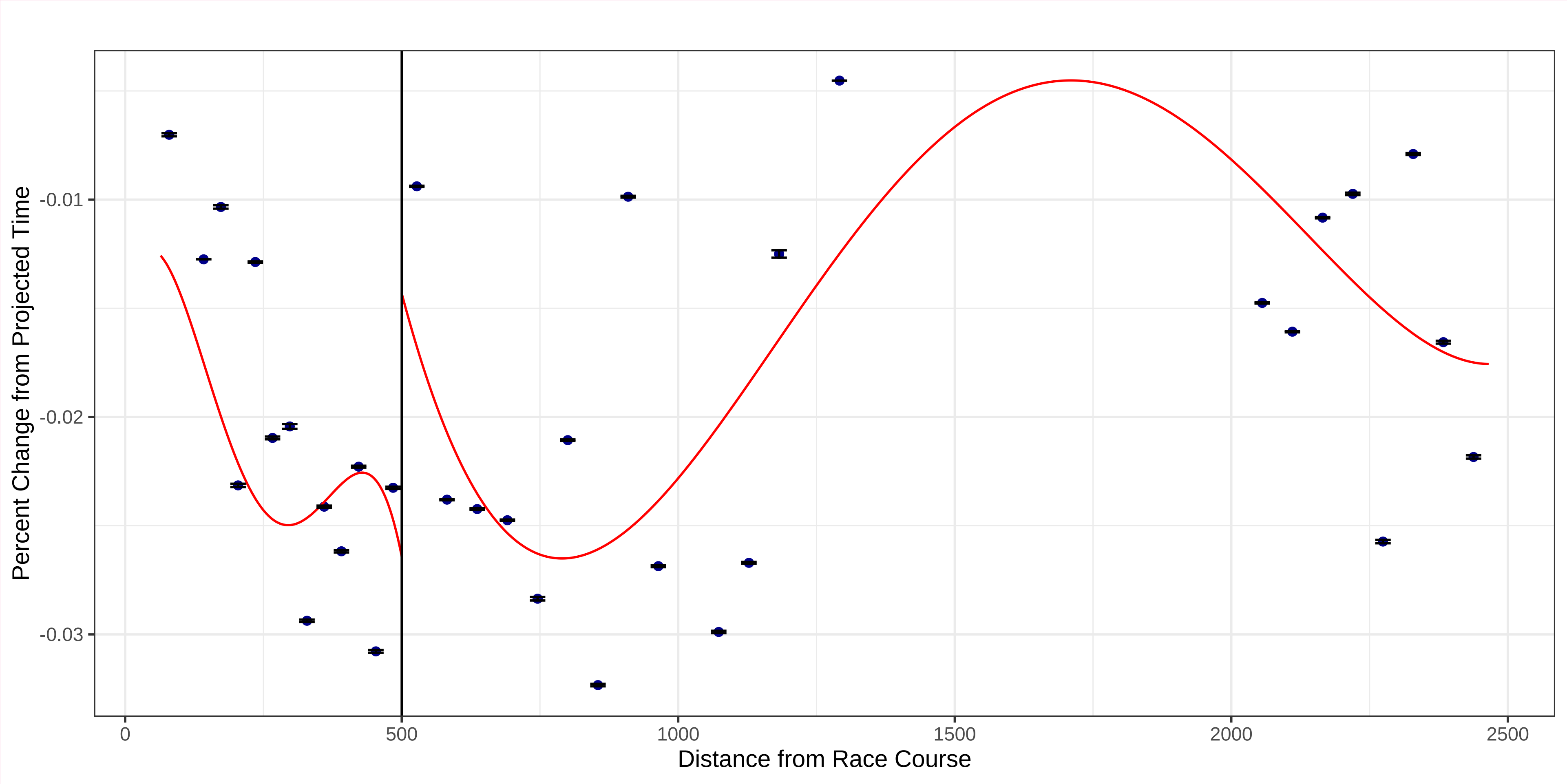
Case Study 1: RDD + Matching

- RDD methods make strong assumptions about covariate balance
- Including control variables can help meet this assumption
- But, matching methods can improve this process

Case Study 1: The RDD + Matching

- Matching implemented between a mix of genetic and exact matching
- $Treatment \sim Distance + Talent + Age + Course + Year$
- The matched dataset is then passed to the robust RDD process

Case Study 1: RDD + Matching



Case Study 1: RDD + Matching

```
## Covariate-adjusted Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                 838
## BW type                     Manual
## Kernel                  Triangular
## VCE method                      NN
##
## Number of Obs.                 272          566
## Eff. Number of Obs.             48           59
## Order est. (p)                   1            1
## Order bias  (q)                  2            2
## BW est. (h)                 81.000       81.000
## BW bias (b)                153.000      153.000
## rho (h/b)                    0.529        0.529
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional     0.018     0.004     4.800     0.000     [0.011 , 0.025]
## Bias-Corrected     0.020     0.004     5.359     0.000     [0.013 , 0.027]
##         Robust     0.020     0.004     4.975     0.000     [0.012 , 0.028]
## =============================================================================
```

Case Study 1: Results

- Teams that get to fly instead of drive do better!
- We can also see that the farther teams have to drive the worst they have to do
- Teams that can afford to fly would gain a marginal benefit from doing so
- But maybe not worth the additional expense?

Final Thoughts

Conclusions

- Professional sports statistics relies heavily on regression methods
- But, drawing causal conclusions from simple or advance regression methods is statistically risky
- More complex causal modeling is also plagued by obfuscation
- Stakeholders have a hard time penetrating the methodological black box

Conclusions

- Design-based methods should get more use by practioners
- Design-based methods give practioners the ability estimate debiased effects
- Design-based methods reflect simple processes
- Design-based methods are generally easier to communicate to stakeholders

Beyond the Prediction *Machines*

The Role of Causal Inference in Sports Statistics

Sean Fischer, Ph.D.

Cincinnati Reds

sfischer@reds.com

References

References

- Baumer, B. S., Matthews, G. J., & Nguyen, Q. (2023). Big ideas in sports analytics and statistical tools for their investigation. Wiley Interdisciplinary Reviews: Computational Statistics, e1612.
- Calonico, S., Cattaneo, M. D., & Titiunik, R. (2015). Rdrobust: an R package for robust nonparametric inference in regression-discontinuity designs. R J., 7(1), 38.
- Cunningham, Scott. (2022). "The Ongoing Role of Machine Learning Engineers and Data Scientists in Industry's Lucas Critique". Substack.

References

- Fischer, K., Reade, J. J., & Schmal, W. B. (2022). What cannot be cured must be endured: The long-lasting effect of a COVID-19 infection on workplace productivity. Labour Economics, 79, 102281.
- Fischer, S. (2021). The Causal Effects of Early-Career Playing Time on the Fourth-Year Performance of NBA Players.
- Gibbs, C. P., Elmore, R., & Fosdick, B. K. (2022). The causal effect of a timeout at stopping an opposing run in the NBA. The Annals of Applied Statistics, 16(3), 1359-1379.

References

- Hünermund, P., Kaminski, J., & Schmitt, C. (2022). Causal machine learning and business decision making. Available at SSRN 3867326.
- Keele, L., Titiunik, R., & Zubizarreta, J. R. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. Journal of the Royal Statistical Society Series A: Statistics in Society, 178(1), 223-239.
- Langen, H., & Huber, M. (2023). How causal machine learning can leverage marketing strategies: Assessing and improving the

References

- Reade, J. J., Schreyer, D., & Singleton, C. (2022). Eliminating supportive crowds reduces referee bias. Economic Inquiry, 60(3), 1416-1436.
- Rubin, D. B. (2008). For objective causal inference, design trumps analysis.
- Toumi, A., Zhao, H., Chhatwal, J., Linas, B. P., & Ayer, T. (2021). The effect of NFL and NCAA football games on the spread of COVID-19 in the United States: an empirical analysis. medRxiv, 2021-02.

References

- Weimer, L., Steinert-Threlkeld, Z. C., & Coltin, K. (2023). A causal approach for detecting team-level momentum in NBA games. Journal of Sports Analytics, 9(2), 117-132.
- Yam, D. R., & Lopez, M. J. (2019). What was lost? A causal estimate of fourth down behavior in the National Football League. Journal of Sports Analytics, 5(3), 153-167.

Case Study 2

Case Study 2: Should Athletes Go Out Faster than Target Pace?

- Track and field athletes have a limited number of opportunities to run fast times during the college season
- Coaches are asked to submit a target time for their athletes
- Coaches can enter faster times than their athletes have run
- **Should coaches enter overly aggressive entry times?**

Case Study 2: The Available Data

- Collected data from the 2023 David Hemery Invitational at Boston University
- Considered only the 5000m run

Case Study 2: The Usual Methods

- Consider a proxy-treatment variable
- **First-half race pace**
- Does running faster in the first half of the race lead to running faster for the whole race?
- Can answer this with simple regression model

Case Study 2: The Usual Methods

```
## # A tibble: 5 × 5
##   term                 estimate std.error statistic       p.value
##   <chr>                   <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept)         -0.0106     0.00303   -3.50   0.000572
## 2 AgeJR                0.00117    0.00365    0.321  0.749
## 3 AgeSO                0.00366    0.00365    1.00   0.316
## 4 AgeSR                0.000151   0.00362    0.0416 0.967
## 5 rel_pace_halfway_pct -0.886      0.140     -6.35   0.00000000139
```

Case Study 2: The Usual Methods

- These results cannot get us away from the usual concerns with observational causal inference
- We have to worry about the common sources of bias
- What haven't we measured here?

Case Study 2: A Causal Inference Approach

- Are there sources of natural randomness that we can leverage?
- **Yes!**
- The cutoffs between heats are established independent of the seed times
- That is, the cutoff between heats creates comparable groups at the bottom and top of back-to-back heats

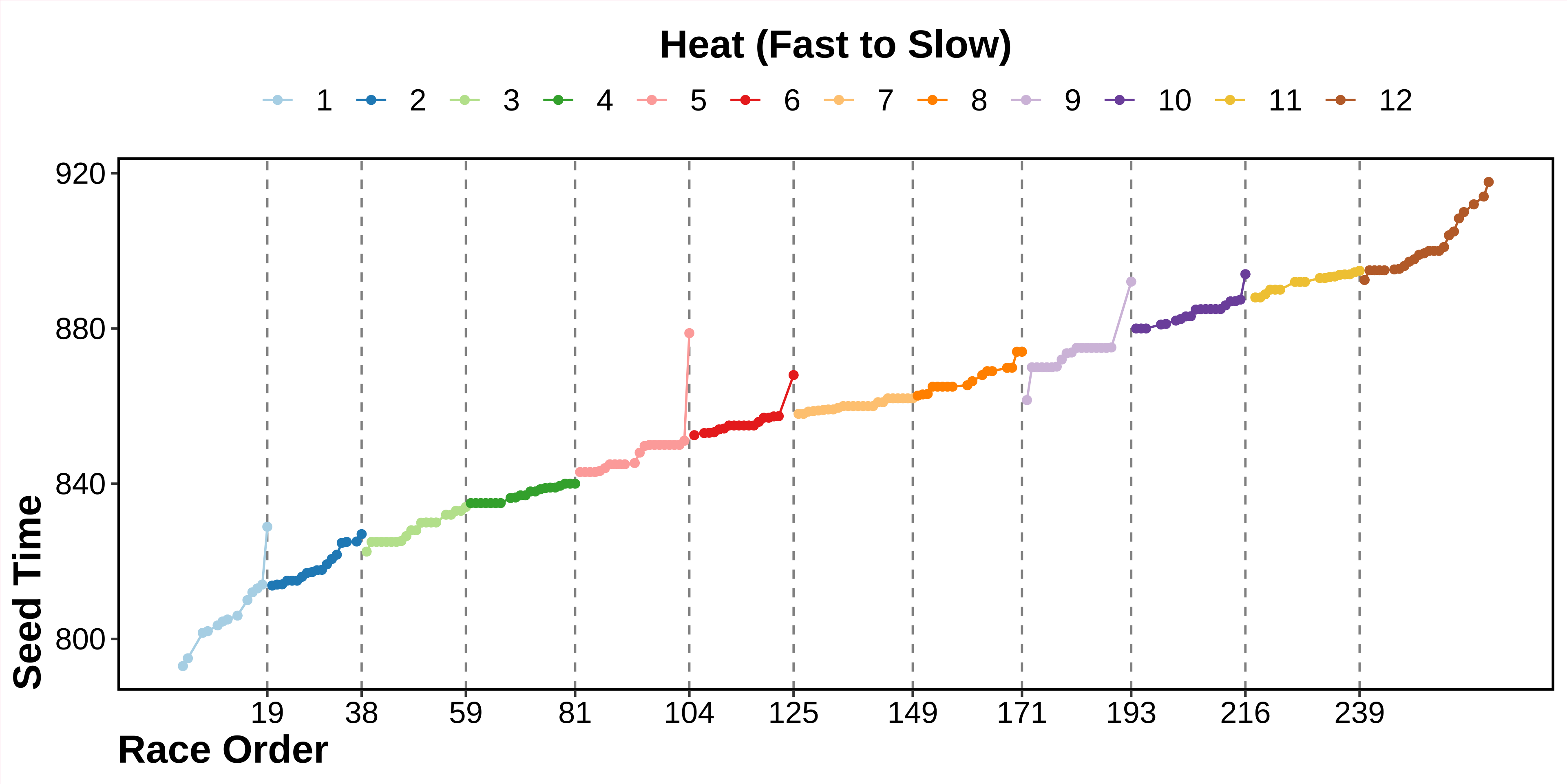
Case Study 2: RDD Set Up

- In this set up, we can apply a multi-cutpoint RDD or we can recenter all the cutpoints

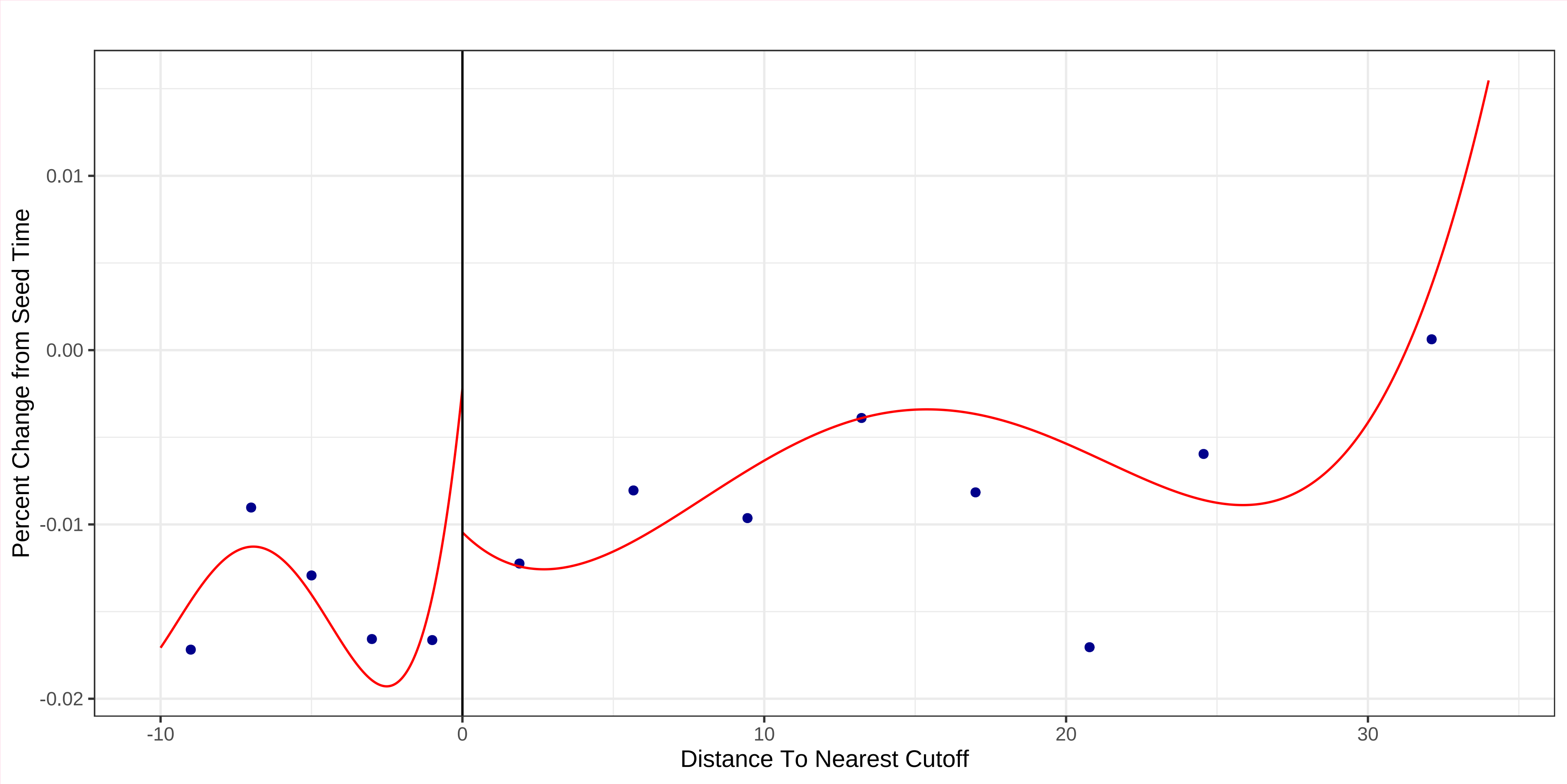
- The running variable is athletes' distance to the next cutoff

- The cutoffs are the last times to get into each heat

Case Study 2: RDD Set Up



Case Study 2: RDD Set Up



Case Study 2: RDD Set Up

```
## # A tibble: 3 × 7
##   term           estimate std.error statistic p.value conf.low conf.high
##   <chr>             <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 Conventional     0.0367    0.0693     0.530   0.596  -0.0991     0.173
## 2 Bias-Corrected   0.0355    0.0693     0.512   0.608  -0.100      0.171
## 3 Robust           0.0355    0.0709     0.500   0.617  -0.104      0.175
```

Case Study 2: Conclusion

- No significant effect of moving up a heat!
- Moving up a heat does not seem to be associated with going out too fast
- But, there also does not seem to be a benefit