Modeling the Progression of Collegiate Track and Field Performances with Quantile Regression

Riley Collins

Department of Mathematical Sciences
Montana State University

December 14, 2023

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Riley Collins

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date	Writing Project Advisor
Date	Writing Projects Coordinator

Contents

Abstract

Many Division I Track and Field athletes have the goal of qualifying for the West or East Regional meets. Over the past few years, particularly after 2020, the marks required to get into these meets have gotten substantially tougher after many years of being relatively stagnant. This is especially true in middle and long distance running events. This paper looks at the men's 800 meters in particular. By using quantile regression, which is a variation of linear regression in which we model a quantile of a distribution instead of its mean, we seek to determine the size of this effect. By comparing different models for the changes over recent years, we are able to assess the changes in qualifying times before and after 2020.

Introduction

In NCAA Division I Track and Field, a goal of many athletes is to qualify for either the East or West Regional meet in one or more of their events. The regions are split by the Mississippi River, except that the East region contains Louisiana and the West contains Wisconsin and Illinois. In each year and region, 48 men and 48 women qualify per event. In cases where an athlete either qualifies for multiple events and chooses not to participate in all of them or qualifies and gets injured later in the season, their spots are filled by the next highest performers until the field of 48 is complete.

The 2020 Outdoor Track and Field season was cancelled due to COVID-19. Since then, many events have required higher, faster or farther marks to qualify for the regional meets. This seems to be the case especially in middle and long distance events. From 2010 and 2019, those qualifying marks were relatively similar between years. In the track and field community, there are two primary reasons for the changes after 2020 that many people point to. The first is that not having a season in 2020 gave athletes the opportunity to train for a full year without having to fit races into their schedule. This allowed them to run a higher number of miles during the week and/or complete more difficult workouts. Through communication with

coaches, teammates and other athletes, many people believe that the long-term benefits of this outweighed what they would have gained from having a season in 2020. The other factor is that both the racing and training shoes that many athletes use were redesigned in such a way between 2020 and 2021 that allows athletes to distribute force into the ground more efficiently and with less effort as they are running, thus leading to faster performances (Jones and Joubert 2021).

Although this paper does not seek to isolate these two effects, it does seek to understand and estimate how the slowest qualifying time in the men's 800 has changed with respect to region and year and also looks to predict what those marks may be in future years for both regions. This paper answers that question using quantile regression. This process works similar to ordinary least squares regression, but instead of modeling the mean of the responses at different covariate values, quantile regression models a particular quantile of the responses.

Data

Dating back to 2010, a website called *Track and Field Results Reporting System* (TFRRS) has collected all collegiate track meet results across the United States (TFRRS 2023). This allows a user to search a particular athlete and see their results across their career. It also compiles performance lists for conferences, regional and national lists for each college division (I, II, III, NAIA and NJCAA) and for each season (Cross Country, Indoor Track and Outdoor Track) in a given year. Of particular interest, the men's 800 meter performance lists for the East and West Region in Division I Outdoor Track are analyzed in this paper. These are the lists that are used to determine the qualifiers of the East and West Regional meets and they were obtained via web scraping using the *rvest* package (Wickham 2022).

The aforementioned performance lists contain the name, year of eligibility, team, and the time, distance or height of someone's strongest performance in a particular event and season. The name and date of the meet at which they got that mark is also included. Only the strongest mark for a particular athlete is included; no athlete appears more than once on a list in a given year. Table 1 shows how many athletes are on each performance list corresponding to region and year for the 800 meters.

Table 1: The number of athletes in the performance list by region and year for the 800 meters.

Year	West	East
2010	500	500
2011	500	500
2012	258	269
2013	218	224
2014	500	500
2015	376	311
2016	300	311
2017	200	200
2018	499	500
2019	500	500
2021	500	500
2022	500	500
2023	500	501

For the years in which the full performance list of 500 athletes was available, Figures 1 and 2 provide non-parametric density curves of the distributions of the fastest times for each athlete on those performance lists, separated by year (Wickham 2016). It is worth noting that even the list of 500 does not represent a complete list of athletes that ran the 800 meters in a given season, with the slowest athletes not being on those lists. This is more evident in the East region, due to the steep drops that are consistently seen in the right side of the distribution. This likely indicates that the total number of 800 meter athletes per year tends to be larger in the East region. However, since that we are not directly interested in the right tail of

the distribution and a consistent sample size makes quantile regression able to address the percentiles where the qualifying cutoff is imposed, this is not necessarily problematic.

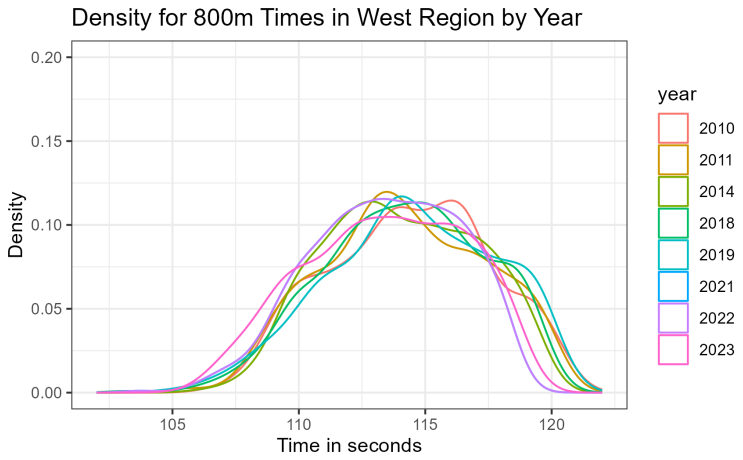

Figure 1: Density curves for 800 meter times in the West region for years in which the performance list has 500 entries.

From Table 1, we know that some years do not have the full set of 500 athletes. We were able to find the slowest qualifying times in the 800 for those years and regions, because the missing entries in those datasets corresponded to slower athletes. However, we don't know specifically which athletes and what their positions on those lists would be were those lists complete. Additionally, there are some much slower athletes that likely would not have ended up in the top 500 but were added to some of the smaller lists. These issues would require us to model a separate percentile for each of those

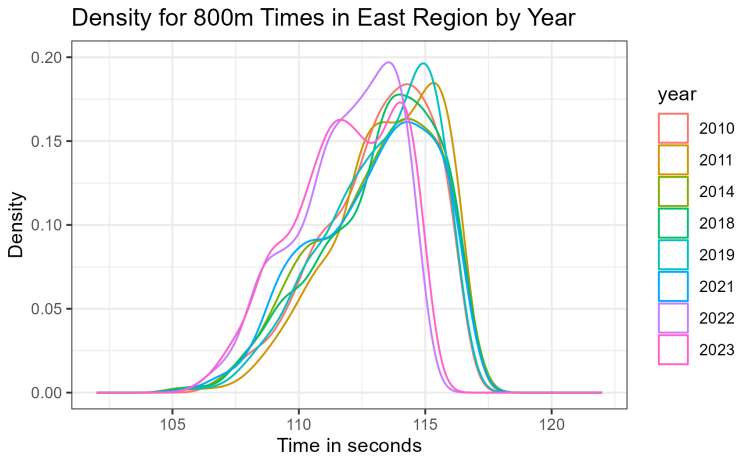

Figure 2: Density curves for 800 meter times in the East region for years in which the performance list has 500 entries.

lists individually, because they contain a different number of observations. Hence, a quantile regression model containing those years would be meaningless. As for the remaining years that do have 500 observations, we can model the time of the slowest qualifying athlete by modeling the 12th percentile of the distribution across those years. Although this isn't exact due to the 60th fastest time not being the cutoff for both regions every year, Table 2 shows that 60th place is usually very close to that mark. Furthermore, the times are dense enough around 60th place so that any differences between the times corresponding to the places in Table 2 and 60th place are usually no more than a few hundredths of a second. Thus, 60th place and the 12th percentile can be considered to be a reasonably accurate proxy for modeling

the slowest qualifying time by year.

Methods

Quantile Regression

In standard linear models, we are interested in the mean response value for a given set of covariate values. A standard linear model can be written as

$$y_i = x_i'\beta + \epsilon_i,$$

where

$$\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

In matrix notation, $Y = X\beta + \epsilon$. Y is an $n * 1$ vector of response values (y_i), x_i' represents the values any predictor variables take for observation i and is the i^{th} row of $n * p$ design matrix X. β is a $p * 1$ parameter vector, ϵ is an $n * 1$ vector representing the residuals (ϵ_i) of the model, n represents the number of observations, and p represents the number of parameters in

the model. The β values are estimated by minimizing

$$\sum_{i=1}^{n} |y_i - x_i'\beta|^2,$$

which is the ordinary least squares criterion.

The primary modeling method this paper uses is quantile regression. With quantile regression, rather than modeling the mean response value as in conventional linear models, we are interested in modeling a particular quantile of the response value for a given set of explanatory values (Fahrmeir et al. 2023). The median is usually better suited for modeling the center of a distribution for skewed responses and can sometimes be used as a proxy for the mean. Quantile models for the median should be very similar to the coefficients from conventional linear models if the responses are normally distributed or at least the residuals are symmetric in the linear model. By using a quantile close to 0 or 1, one can get a better idea of how much variance a distribution has (Hardle and Linton 1994).

A quantile regression model for quantile τ can be written as

$$y_i = x_i'\beta_\tau + \epsilon_{i\tau}.$$

β_τ is estimated by minimizing

$$\sum_{i=1}^{n} w_\tau(y_i, \phi_{i\tau})|y_i - \phi_{i\tau}|,$$

where $\phi_{i\tau} = x_i'\beta_\tau$ and $w_\tau(y_i, \phi_{i\tau})$ is a weight function that is denoted as

$$w_\tau(y_i, \phi_{i\tau}) = \begin{cases} 1-\tau, & y_i < \phi_{i\tau} \\ 0, & y_i = \phi_{i\tau} \\ \tau, & y_i > \phi_{i\tau} \end{cases}.$$

The *rq* function in the *quantreg* package in R was used to estimate the quantile regression models (Koenker 2023). The algorithm that is used can be changed by specifying the *method* argument, but the default value is *br* and represents the Barrodale and Roberts simplex algorithm (Barrodale and Roberts 1974). This algorithm is generally useful where there are less than a few thousand observations (Koenker, n.d.). For larger sample sizes, the Frisch-Newton algorithm can be used by specifying *fn* or *pfn* (Portnoy and Koenker 1997).

In order to develop our quantile regression models, the data were modified and combined as follows. First we took the complete years of data

Table 2: The position on the performance list of the slowest qualifying athlete by region and year is shown.

Year	**West**	**East**
2010	59	54
2011	57	58
2012	57	55
2013	54	54
2014	59	58
2015	58	52
2016	62	58
2017	60	66
2018	61	63
2019	62	59
2021	64	61
2022	61	61
2023	55	56

(2010, 2011, 2014, 2018, 2019 and 2021-23) and merged their respective performance lists into a single dataset (Wickham et al. 2021). This was done separately for each region. The design matrix also includes an indicator variable for whether a season took place before or after 2020, denoted *post2020*. This variable takes on a value of 1 for seasons after 2020 and 0 for seasons before 2020. The value for *year* was transformed to be the actual year minus 2020. This was done so that the models that include an effect for before or after 2020 could generate an interpretable parameter value that is the estimate for 2020. The times and regions of each mark are included in the dataset, but the additional variables that were present in the original performance

lists are not present in this merged dataset, since they are not relevant to our research question. With the *lubridate* package, the times were converted to numerical objects so that they could be modeled more easily (e.g. 1:46.00 would be converted to 106.00) (Grolemund and Wickham 2011).

Five models for each region were run, each using the 12th percentile as a proxy for the slowest qualifying time. First, an intercept-only model was estimated, which corresponds to no change over time. *Year* and *post2020* were modeled separately against the 12th percentile in the next two models. The *year* model would suggest a change over time but no intercept shift after 2020, whereas the *post2020* model would suggest there was a one-time shift after 2020, but no change between years. The fourth model included *year* and *post2020*, without an interaction between them. The fifth model also included the interaction, which allows us to assess whether the linear change between years is different before and after 2020.

Assumptions

There are two primary assumptions that are made with a quantile regression model (Fahrmeir et al. 2023). One is that the cumulative distribution function of the residuals for modeling a particular quantile should be equal

to that quantile value when evaluated at 0. In other words, for our 12th percentile models above, we want 12% of the residuals to be negative and 88% of them to be positive. The weights, $w_\tau(y_i, \phi_{i\tau})$, in the optimization function that we seek to minimize help ensure this assumption is met (Koenker and Hallock 2001). To be sure, we extracted the residual values from all models that were run. The proportions of negative residuals ranged from 11.89% to 12.10%, with the rest being positive. These values may not be precisely 12% due to 12% of the total number of data points not being a whole number, but for practical purposes, this assumption is met by the way these models are estimated.

The second assumption for quantile regression models is that the error values, $\epsilon_{i\tau}$, are independent of each other. This assumption is much more difficult to assess, but it may be violated when an athlete qualifies for the regional meet in multiple years. Unfortunately, differentiating and tracking certain athletes across multiple years can be difficult. One limitation of the data is that athletes are not given a unique identifier, meaning there is no way to distinguish the 21 "Sam Smith" profiles without looking at each one of them individually (TFRRS 2023). It gets more complicated when accounting for athletes that may have transferred, which requires matching

multiple profiles to a single athlete. In cases where an athlete does transfer, it is also possible that they are listed as “Sam” at one school and “Samuel” at the other. Some athletes also change their names. This makes it virtually impossible to account for repeated measures on an individual athlete using a random effect.

This assumption may also be violated given the fact that certain programs and conferences as a whole compete at a higher level than others and an individual athlete may run faster based on being able to train and compete with faster athletes. Many of these programs also have the ability to recruit athletes that are more naturally talented, regardless of who they train and compete with. While the information to account for variation between schools and conferences is available, this is left as future work.

It is likely that one or both of these factors makes it so that the independence assumption is violated. This is a limitation of the model that must be considered.

Results

Each of the five models was fit to each region separately, then compared using Akaike Information Criteria (AIC) values.

The estimated intercept-only model can be written as

$\widehat{QualTime}_i = 110.000$ for the West region and

$\widehat{QualTime}_i = 109.830$ for the East region.

The estimated *year* model can be written as

$\widehat{QualTime}_i = 109.909 - 0.049 * year_i$ for the West region and

$\widehat{QualTime}_i = 109.593 - 0.101 * year_i$ for the East region.

The estimated *post2020* model can be written as

$\widehat{QualTime}_i = 110.270 - 0.550 * I(post2020_i = 1)$ for the West region and

$\widehat{QualTime}_i = 110.300 - 1.030 * I(post2020_i = 1)$ for the East region.

The model with slowest qualifying time against both *year* and *post2020*, but no interaction between them, can be written as

$\widehat{QualTime}_i = 110.338 + 0.008 * year_i - 0.642 * I(post2020_i = 1)$ for the West region and

$\widehat{QualTime}_i = 110.012 - 0.048 * year_i - 0.647 * I(post2020_i = 1)$ for the East region.

The model with slowest qualifying time against both *year* and

post2020, as well as the interaction between them, can be written as

$\widehat{QualTime}_i = 110.338 + 0.020 * year_i - 0.310 * I(post2020_i = 1) - 0.220 * year_i * I(post2020_i = 1)$ for the West region and

$\widehat{QualTime}_i = 110.030 - 0.043 * year_i - 0.120 * I(post2020_i = 1) - 0.297 * year_i * I(post2020_i = 1)$ for the East region.

Table 3: The AICs for each of the five models for the West are shown.

Region	**Model**	**AIC**	**Δ AIC**
West	Intercept	22130.2	38.46
West	Year	22106.56	14.82
West	Post2020	22095.48	3.74
West	Year + Post2020	22097.36	5.62
West	Year * Post2020	22091.74	0

Table 4: The AICs for each of the five models for the East are shown.

Region	**Model**	**AIC**	**Δ AIC**
East	Intercept	20414.1	149.43
East	Year	20271.32	36.65
East	Post2020	20259.04	24.37
East	Year + Post2020	20248.4	13.73
East	Year * Post2020	20234.67	0

Akaike Information Criteria (AIC) can be calculated as

$$AIC = 2k - 2ln(\hat{L}),$$

where k is the number of parameters in a model and $\hat{L}$ is the maximum of the

likelihood function of the model (R Core Team 2023). A difference in AIC of two or more points is generally considered to be strong evidence in support of the model with the smaller AIC. The AIC values for the five models are given in Tables 3 and 4.

The model that included *year* and *post2020* as well as the interaction between *year* and *post2020* provided the lowest AIC for both the West and East regions. Because of this and the large Δ AIC values, this was the primary model that was chosen for this analysis. For the West region, the interaction model is 3.74 points smaller than the shift-only model. For the East region, the interaction model is 13.73 points lower than the linear trend and shift model. Both present strong evidence in support of the interaction model versus the next best AIC model, but the next best models in each region differ. The point estimates for the parameter values, as well as the corresponding standard errors and 95% confidence intervals for the interaction models are provided in Tables 5 and 6 for the West and East regions, respectively.

The confidence intervals in Tables 5 and 6 as well as Figures 3 and

4 were calculated using the asymptotic covariance matrix of

$$V(\hat{\beta}_\tau) = \sigma_\tau^2 (X'X)^{-1},$$

where $\sigma_\tau^2 = \frac{\tau(1-\tau)}{p[P^{-1}(\tau)]}$ and $\hat{\beta}_\tau$ is a point estimate for slowest qualifying time at a given set of covariate values. $p(\cdot)$ is the probability density function of the error distribution and $P^{-1}(\cdot)$ is the quantile function for the errors. Hence, $p[P^{-1}(\tau)]$ is the density of the error distribution at quantile τ (Fox 2016). Confidence intervals for the corresponding linear combination can be found using

$$c'\hat{\beta} \pm t_{df}\sqrt{c'V(\hat{\beta})c},$$

where c is a contrast vector to form a linear combination of the slope coefficients and *df* is the residual degrees of freedom in the model (Fahrmeir et al. 2023). The *df* is equal to the number of data points minus the number of parameters in the model (3995 in the West and 3997 in the East). For τ values near 0 and 1, the finite sample performance of this method is not well understood, but this may not be an issue for $\tau = 0.12$.

An alternative way of writing the interaction models is by simplifying the models for before and after 2020 for each region. The models for

Table 5: The parameter coefficients for the interaction model for the West region are given, along with their standard errors and confidence intervals.

Coefficient	**Estimate**	**Standard Error**	**95 Percent Confidence Interval**
Intercept	110.380	0.1783	(110.0305, 110.7295)
Year	0.020	0.0258	(-0.0306, 0.0706)
Post2020	-0.310	0.3050	(-0.9078, 0.2878)
Year:Post2020	-0.220	0.1268	(-0.4685, 0.0285)

Table 6: The parameter coefficients for the interaction model for the East region are given, along with their standard errors and confidence intervals.

Coefficient	**Estimate**	**Standard Error**	**95 Percent Confidence Interval**
Intercept	110.030	0.1647	(109.7072, 110.3528)
Year	-0.043	0.0237	(-0.0894, 0.0035)
Post2020	-0.120	0.3245	(-0.7560, 0.5160)
Year:Post2020	-0.297	0.1253	(-0.5426, -0.0514)

before 2020 can be written as

$\widehat{QualTime}_i = 110.380 + 0.020 * year_i$ for the West region and

$\widehat{QualTime}_i = 110.030 - 0.043 * year_i$ for the East region.

The models for after 2020 can be written as

$\widehat{QualTime}_i = 110.070 - 0.200 * year_i$ for the West region and

$\widehat{QualTime}_i = 109.910 - 0.340 * year_i$ for the East region. In other words, the 12th percentile was estimated to be getting slightly slower in the West and slightly faster in the East each year prior to 2020. After 2020, the estimated models suggest that the 12th percentile gets 0.2 seconds faster in the West

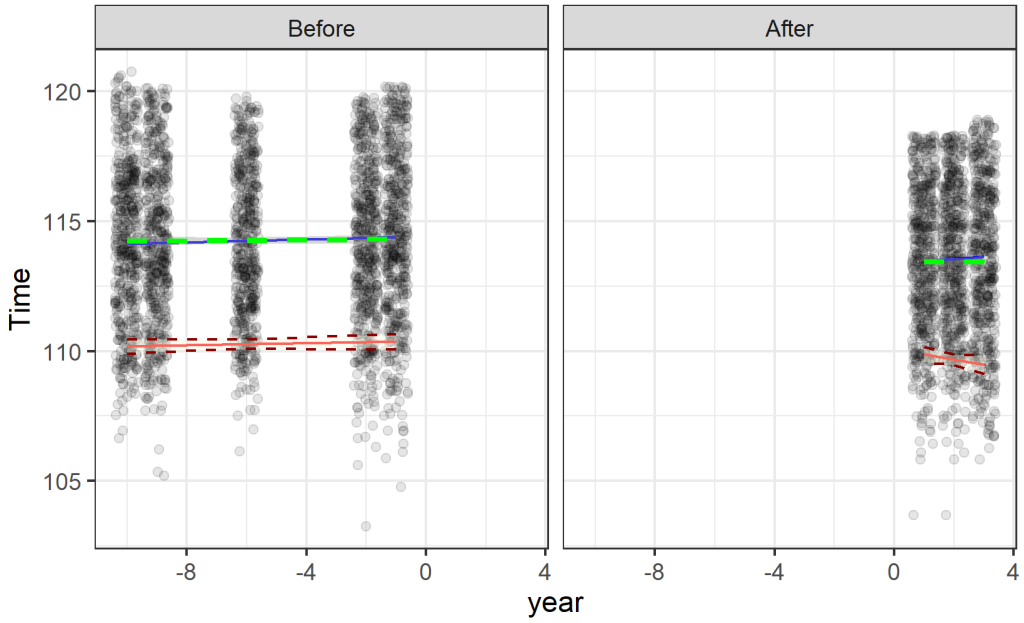

Figure 3: Quantile regression in the West region for the 12th and 50th percentiles are shown in red and blue, respectively. The ordinary least squares model is shown in green. The dark red dashed lines indicate 95 percent confidence intervals for the 12th percentile.

and 0.34 seconds faster in the East each year. Given the rate of progression of 800 meter times for many athletes, these changes between years are quite large.

Future Predictions

As for predicting future years, it is difficult to conclude whether the trends that were accounted for in this model will continue, at least to the same extent. It is likely that the shoes that have been used since 2020 will continue

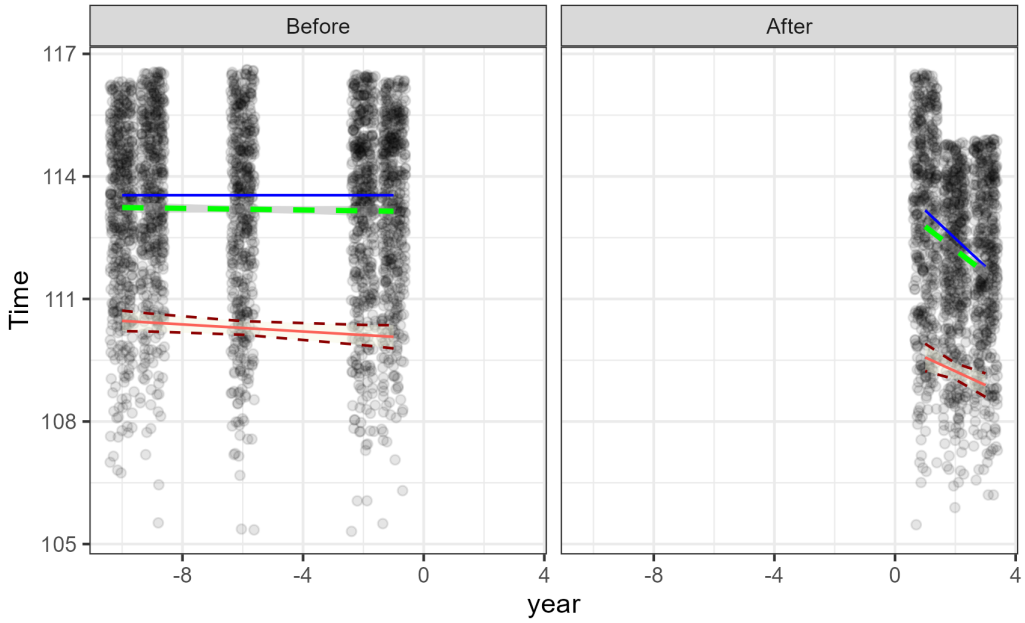

Figure 4: Quantile regression in the East region for the 12th and 50th percentiles are shown in red and blue, respectively. The ordinary least squares model is shown in green. The dark red dashed lines indicate 95 percent confidence intervals for the 12th percentile.

to be used, because athletes don't want to knowingly put themselves at a disadvantage. However, it is not known how shoe technology will continue to advance in the coming years. Furthermore, the effect that the cancelled track season has on athletes may change. An increasing number of athletes were in high school when COVID-19 started. Given that high school athletes generally train at a lower volume and they have less racing experience, it may not have been as much to their advantage to lose a season of racing. In a few more years, many of the NCAA athletes at that time will have been in

middle school and may not have been running at all when the track season was cancelled. The East region model in particular may not be robust to these changes, given that both models assume a linear trend with respect to year and the effect size in the East is so much larger than in the West.

Due to the uncertainty of if and how these effects will change in future years, this paper will only predict the slowest qualifying time in the men's 800 meters for the 2024 season. Using the *predict.rq* function in the *quantreg* package (Koenker 2023), point estimates and 95% confidence intervals were generated for each region. For the West region, the point estimate is 1:49.27, with the 95% confidence interval spanning from 1:48.71 to 1:49.83. For the East region, the point estimate is 1:48.55, with the 95% confidence interval spanning from 1:48.05 to 1:49.04.

Conclusion

Prior to 2020, the amount by which the slowest qualifying time in the men's 800 meters changed per year was minimal. This was true for both regions and their times were much more similar to each other. After 2020, both regions got faster, both from an intercept shift that applies to all years after 2020

and the decrease in qualifying time from one year to the next also became larger. This effect has been much more pronounced in the East, meaning their times have become much faster compared to those from the West. This is represented by the 12th percentile having a much steeper slope after 2020 in Figure 4 than in Figure 3. Still, the qualifying times for the West in some cases have been close to a second faster than they were in past years.

Given that many athletes can spend their entire collegiate career dropping a second from their 800 time, this is a substantial difference from the marks that could be expected a few years ago. The qualifying times for 2024 were also predicted assuming a continuation of the trends that have been observed over the past few years, although whether those trends continue remains to be seen and could be addressed in future research.

Other future explorations include looking more into AIC selection performance in quantile regression, confidence interval performance at quantiles near 0 or 1, and quantile models with random effects to account for athlete-athlete, school-school and conference-conference variation.

Bibliography

Barrodale, I., and F. Roberts. 1974. "Solution of an Overdetermined System of Equations in the L1 Norm" 17 (6): 319–20.

Fahrmeir, Ludwig, Thomas Kneib, Stefan Lang, and Brian D. Marx. 2023. *Regression Models, Methods and Applications*. Springer Berlin.

Fox, John. 2016. *Applied Regression Analysis and Generalized Linear Models*. Sage Publications.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Hardle, Wolfgang, and Oliver Linton. 1994. "Applied Nonparametric Methods." *Handbook of Econometrics*.

Jones, Garrett, and Dustin Joubert. 2021. "A Comparison of Running Economy Across Seven Carbon-Plated Racing Shoes." *Footwear Science*.

Koenker, Roger. 2023. *Quantreg: Quantile Regression*. https://CRAN.R-project.org/package=quantreg.

———. n.d. "Quantile Regression - University of Illinois Urbana-Champaign." *Quantile Regression in R: A Vignette*. University of Illinois: Urbana-Champaign.

Koenker, Roger, and Kevin Hallock. 2001. "Quantile Regression." *Journal of Economic Perspectives* 15 (4): 143–56.

Portnoy, S., and R. Koenker. 1997. "The Gaussian Hare and the LaPlace Tortoise: Computability of Squared-Error Versus Absolute-Error Estimators, with Discussion." *Stat Science* 12 (4): 279–300.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

TFRRS. 2023. https://www.tfrrs.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2022. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.

Wickham, Hadley, Romain Francois, Lionel Henry, and Kirill Muller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Appendix: R Code

The code used for scraping, wrangling and cleaning the data, as well as writing the quantile regression models, can be found at the GitHub repository *rileycollins5/QuantileRegression*.