Andrea Kuhn & Alex Nguyen
Stat 6950

Data Project

Proposal

The data set contains results for the marathon, 50 km race-walk, 20 km race-walk, 10000 m, 5000 m, and 3000 m-steeplechase for major competitions (Commonwealth Games, Diamond League, World Athletics Continental Cup, World Athletics Gold Label Marathons, Olympic Games, World Athletics Race Walking Team Championships, and World Championships) from 2000-2019. The standing event record and world record were also included for each event. These results were obtained from the official websites and collected between February 2016 and March 2024.

The original data set contained 1258 races ranging from 1936 to 2019 after screening out races at the beginning of the creation due to inconsistencies in community interest and runner participation. For our uses, we will only be considering race-walking, track, and marathon championships races from 2000-2010 and marathon Gold Label Races from 2010-2019. Additionally, since our focus will be only on the top three results (elite athletes) rather than the 25th, 50th, 100th, and 300th place finishers (well-trained runners) we removed several Gold Label Races (marathons) that are historically not attended by elite athletes (Athens, Gold Coast, Ottawa, Philadelphia, etc.). These races were selected by reviewing the results from 2010-2019 and by utilizing personal knowledge and experience. This choice to reduce the original data set was also driven by the necessity to manually compete the data for the remaining Gold Label Races. Because most Gold Label Races (marathons) have a combined male and female start, the original data set did not distinguish between the genders for these races. Therefore, we completed our data set by copying the weather and location values from the male race and then adding the top three female finishers.

Additionally, the 4th through 10th place finishing times were not provided for a large portion of races, especially the marathon distance, and it was determined to be beyond the scope of this project to collect these results. Therefore, we cannot confidently use these variables or any function of these variables as the response for our model, due to the large number of observations with missing values. However, for all races included in our final data set, the top 3 finishing times were available. Therefore, we are able to use these variables or a function of these variables as the response. We also found 7 races with a missing meteorological station ID and missing weather data. These races were removed from our data set. Our current data set has 650 races ranging from September 2000 to December 2019. For the process used to collect the weather data, see Mantzios et al..

Question of Interest

Our goal is to assess the impact of weather, specifically temperature or heat related variables, on elite athlete performance in endurance running events. We hope to determine which, if any, weather variables influence performance, and whether the gender of participants, competition format, or race length also play a role.

Exploratory Data Analysis

Our data set contains 12 race identifier variables, 16 weather variables, and 20 race result variables with 8 being generated or derived from the original unique 12 race result variables. The race identifier variables include the purpose of the race, distance, and time and location details. Most of the weather variables are numeric and can be considered continuous. Similarly, the race result variables are numeric and can be considered continuous. To simplify calculations with time variables, we transform the records and finishing times to seconds rather than hours, minutes, and seconds.

Additionally, we separated `dist_from_loc` into a categorical variable. While the majority of races are situated within a 10-mile radius of a meteorological station, there are instances where races are located considerably farther away (491 races within 10-miles vs. 159 races further than 10-miles). By categorizing races based on their proximity to the closest meteorological station, we can account for potential inaccuracies in weather data, particularly for events located further from monitoring infrastructure.

Response Variable

Our response variable `percent_standing_avg` represents the percent difference between the average top three finishing times and the event record going into the start of the event. A negative value indicates the finisher set a new event record, while a positive value indicates the finisher had a time slower than the event record. The values range from -6.7005% to 16.2757% with an average of 2.1905%. Figure 1 shows the distribution is uni-modal and roughly symmetric with a slight right skew. It does not appear that any transformations of the response variable are necessary.

There appear to be three potential outliers in our response variable with values above 12%. These observations correspond to the 2018 Women's Boston Marathon and the 50K and 20K race-walk at the 2019 World Championships in Doha, Qatar. The 2018 Boston Marathon experienced severe wind and rain, as well as cold weather that forced many of the top competitors to drop out with hypothermia symptoms. The 2019 World Championship race-walks started just before midnight in Doha in an attempt to avoid the blistering desert heat. However, even at midnight, the high temperature and humidity produced brutal conditions. Therefore, we will not remove these observations as outliers. The results of the two observations with response values below -5% were also confirmed to be correct.

Weather Variables

We expect collinearity between several of the weather predictors as many are calculated using one another. For example, the heat index was calculated using either air temperature and dew point or air temperature and relative humidity; the adjusted wind speed was calculated using wind speed and then adjusting for buildings and elevation; and the wet bulb globe temperature (WBGT) and simple WBGT were calculated using previous methodology that relies on standard meteorological data (air temperature, dew point, etc.) to predict the WBGT without completing a true measurement. Other implicit relationships also occur without direct dependence. For example, when cloud coverage is high, solar radiation will likely be low.

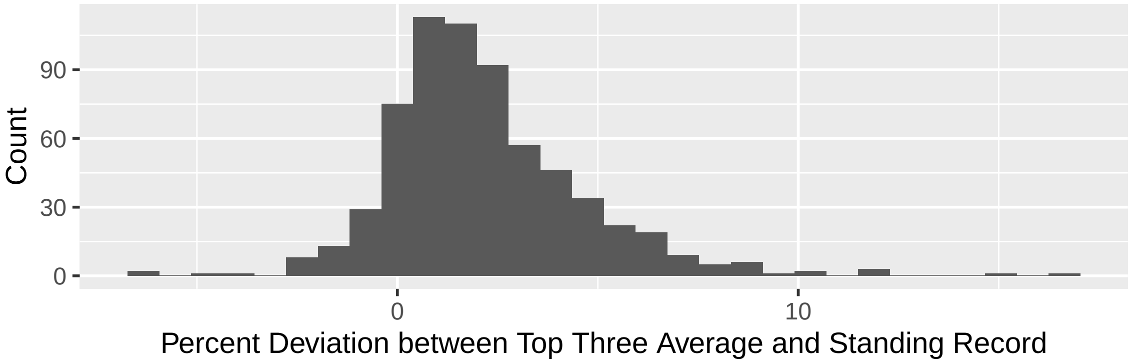
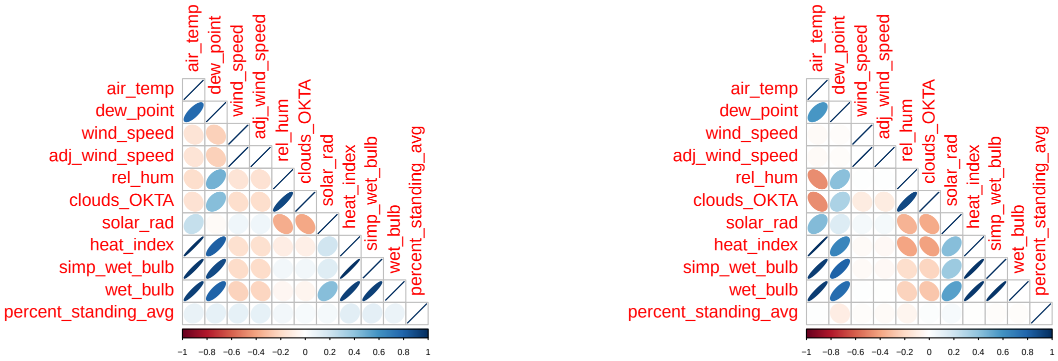


Figure 1: Histogram of response variable.

We may note the differences in the correlation matrix between the weather variables and the percent deviation when races are separated by those within 10-miles of the meteorological station (Figure 2a) and those further than 10-miles (Figure 2b). We observe higher overall correlation between percent deviation and the weather data for races that are within 10 miles of a meteorological station as compared to races further away. Additionally, some relationships between predictors and the response appear to flip direction. This suggests that weather conditions may play a more influential role in races held closer to meteorological stations.



(a) Weather station within 10-miles.

(b) Weather station beyond 10-miles.

Figure 2: Correlation plot between weather predictors and response.

Regardless of proximity, visually assessing scatterplots and calculating correlation coefficients, it appears that air temperature, dew point, wind speed, heat index, relative humidity, and simplified WBGT are likely to be important in predicting the percent deviation between finishing time and event record. The scatterplot matrix with these predictors and all races is shown in Figure 3. However, due to the high collinearity between these predictors, it is also possible that they are explaining the same variation in percent deviation. Formal tests will need to be completed to determine which predictors provide the best model fit and whether proximity to the meteorological station influences the effect of the weather predictors. For several weather parameters, there appears to be a potential quadratic relationship with `percent_standing_avg` such that extreme high and low values of the weather parameter are associated with large percent deviations. However, this relationship may be a result of extreme cases skewing our visual inspection (i.e. 2018 Boston Marathon and 2019 World Championship Race-Walks).

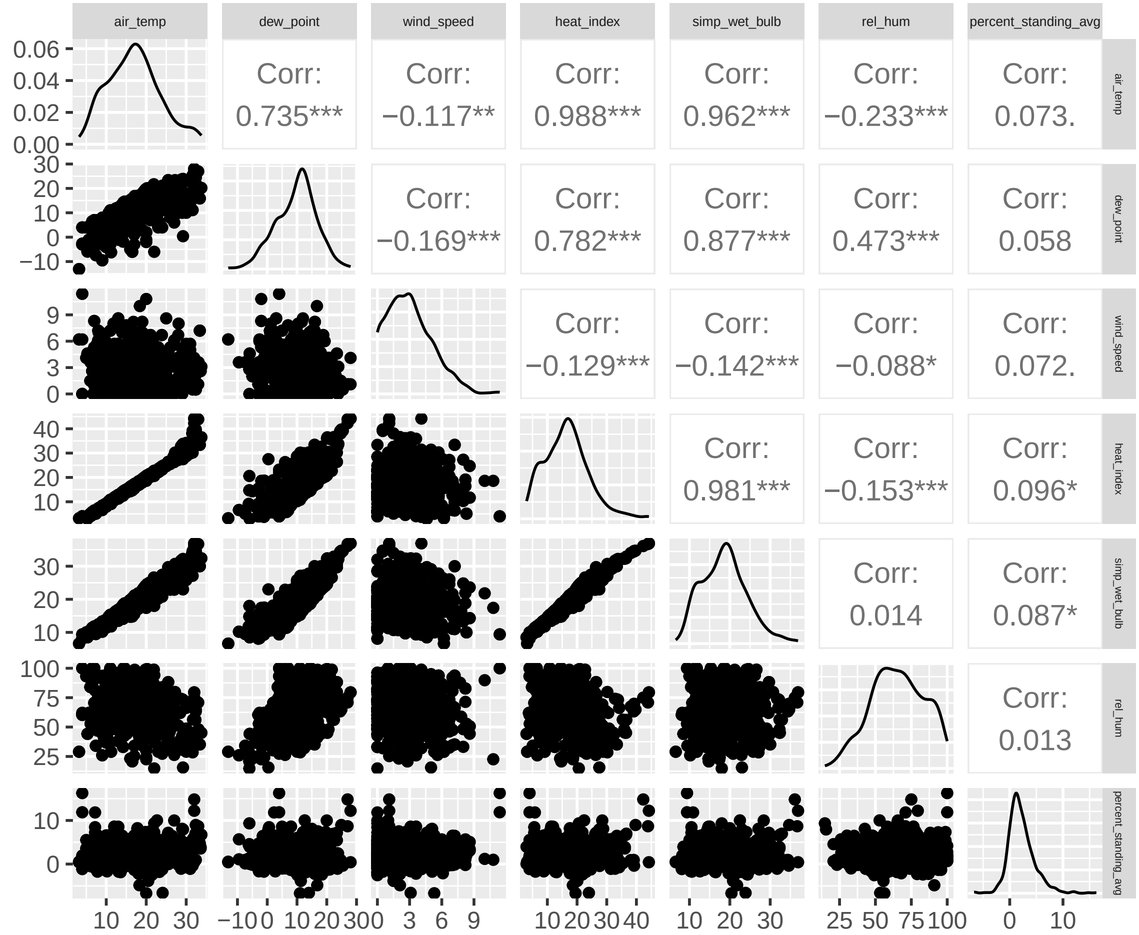


Figure 3: Scatterplot matrix of significant weather predictors.

Categorical Variables

While our primary goal is to assess the impact of weather variables on the performance of elite endurance athletes, several categorical variables display a potential relationship with percent deviation from the event record.

Beginning with analyzing our available categorical variables and their individual distributions, we see that the split between male and female races is roughly even with 357 male races and 293 female races. This difference is largely due to a greater number of Gold Marathons, 50K race-walks, and 3K steeplechase results for males. The 3K steeplechase was not introduced to major female competitions until the late 2000s, and the 50K race-walk had no major female results during our data collection time period. Additionally, we see there is a relationship between competition and distance. For example, the Diamond League only includes results at 5K and 10K, while Gold Marathon only includes results for the marathon. We see these counts in Table 1.

Table 1: Count data of primary categorical variables.

	Men						Women					
Comp.	10K	20K	3K	50K	5K	Mar.	10K	20K	3K	50K	5K	Mar.
Common Wealth	5	4	5	2	4	5	4	3	3	0	5	5
Diamond League	0	0	15	0	48	0	0	0	14	0	36	0
Gold Marathon	0	0	0	0	0	166	0	0	0	0	0	146
Olympic	5	2	5	2	4	5	5	2	2	0	5	3
World Champ.	10	9	8	8	9	10	10	8	6	0	8	12
World Cup	0	8	5	9	4	0	0	9	3	0	4	0

Now, assessing the relationships between our categorical variables and our response variable, we see that the percent deviation from the event record does not appear to show significant variability across competition type, although some variability is present (Figure 4).

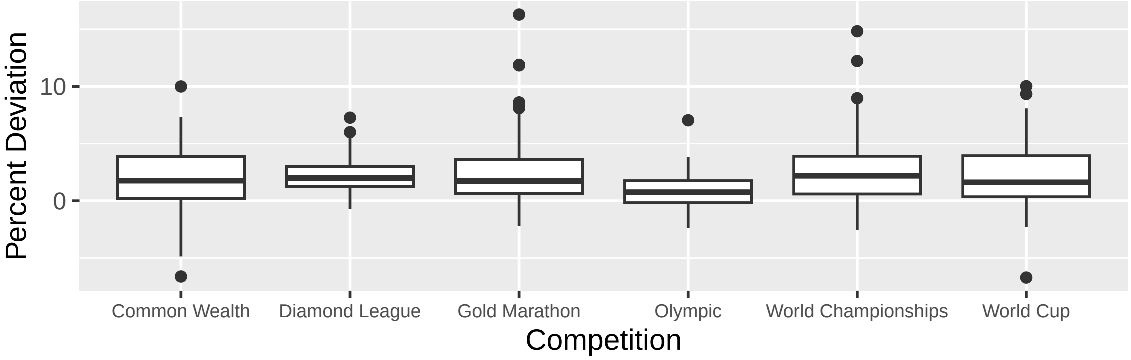


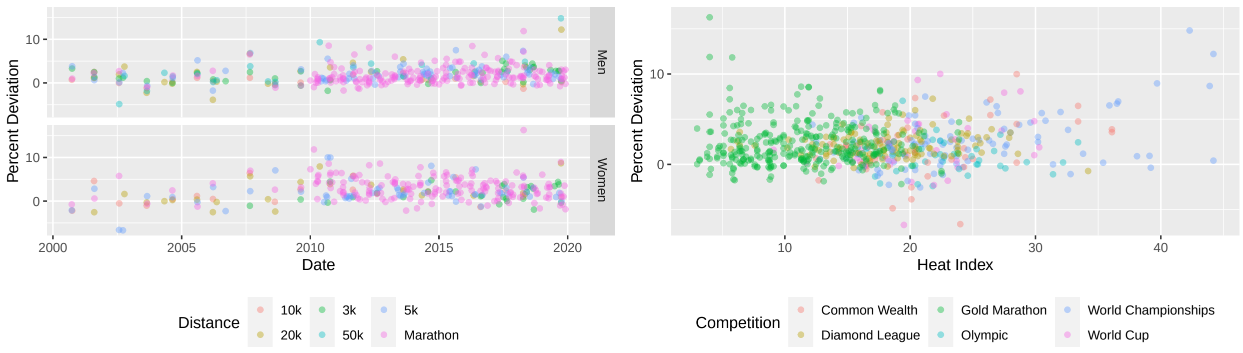
Figure 4: Percent Deviation by Competition Type.

Figure 5(a) shows the relationships between date, distance, gender, and percent deviation from the standing record.

Gender does not appear to have a significant relationship with percent deviation from the event record by itself. However, in addition to possibly accounting for the discrepancy in sample sizes for distances and competitions across genders, including gender as part of an interaction term would allow us to assess the differences in the effects of weather on endurance performance between males and females. Other interaction terms between our categorical variables and weather predictors may also be beneficial for model fit.

In Figure 5(a), we also see that both genders had a slightly higher proportion of record setting races in the mid 2000s before returning to a more stable rate of records by 2010. Similar to competition type, distance does not appear to display an overwhelmingly strong relationship with percent deviation; however, there may be some relationship. For example, marathons appear to have a higher frequency of relatively slow finishing times. Another observation to note is the increase in the number of race results after 2010. This is due to our manual entry of Gold Marathon results for women between 2010 and 2019.

To demonstrate the possibility of an interaction term between a weather predictor and a categorical variable, Figure 5(b) shows a relationship between heat index and distance. It is possible that the



(a) Percent Deviation by Date, Distance, and Gender. (b) Percent Deviation by Heat Index and Competition.

Figure 5: Percent Deviation vs. Categorical and Continuous Predictors

slope of a regression line for `percent_standing_avg` on `heat_index` will differ for each level of `distance`. Further, we see the slight curvature recognized in the earlier scatterplot matrix.

After exploring and analyzing our data, we propose to fit the data with multiple linear regression. Our focus will be on the distance of the race, competition type, gender of the participant, and proximity to the meteorological station categorical variables, as well as the air temperature, heat index, wind speed, relative humidity, dew point, and date of the race continuous predictors. We will also consider interaction terms between both categorical and continuous predictors; however, as our goal is to describe the relationship between weather and elite athlete performance rather than make predictions, we will aim to keep our model simple for interpretation.

Our preliminary model selection will be guided by `stepAIC()`. We will then complete further analysis by conducting significance tests for individual predictors, reviewing residuals for model fit, and identifying any influential points.

Methods and Model Building

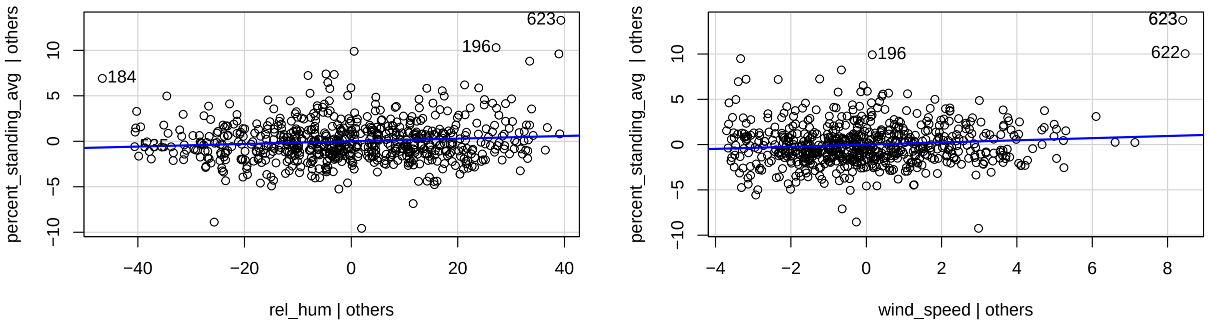
We start by fitting a simple model with all of our previously noted potential parameters and perform step-wise, forward, and backward model selection using AIC. Each selection process returned the same predictors with the only variation coming from the forward selection model which excluded `rel_hum`, `simp_wet_bulb` and `proximity` as a predictor while the step-wise and backward selection models excluded `air_temp` and `simp_wet_bulb`. Therefore, we will begin our analysis with the step-wise/backward selection model and consider more stringent parameter selection criteria due to multi-collinearity in predictors. Additionally, we will explore adding interaction terms to determine whether relationships exist between weather predictors and our categorical variables.

To assess potential multi-collinearity, we will look at the VIF (GVIF) values for our step-wise model. As expected, we see that dew point, relative humidity, and heat index show high collinearity.. Therefore, given the significance of each predictor and our previously observed correlations, we will remove dew point from the model. Air temperature showed strong collinearity with heat index, which we desire to maintain in the model for inference and interpretation purposes.

After refitting our model, our VIF values do not suggest evidence of multi-collinearity. However, to explore whether relative humidity or wind speed explain any additional variability in the percent

	GVIF	Df	GVIF^(1/(2*Df))
sex	1.032903	1	1.016318
competition	13.995586	5	1.301964
distance	10.354361	5	1.263317
dew_point	36.869081	1	6.071992
rel_hum	14.505507	1	3.808610
wind_speed	1.072295	1	1.035517
heat_index	30.085432	1	5.485019
proximity	1.105044	1	1.051211

deviation from event record after accounting for all other parameters in the model, we may review added variable plots for `rel_hum` and `wind_speed`. Figure 6(a) and Figure 6(b) both appear to be null plots, which suggests that relative humidity and wind speed do not explain any additional variability. This is supported by the results of ANOVA tests on each parameter at a family-wise error rate of 5%. Further, the ANOVA tests (Table 2) suggest that distance and proximity are also not significant. We will remove proximity as a predictor, as we only hoped to explore whether there was a relationship between distance to meteorological station and percent deviation, but it is possible that the relationship only lies in the integrity of the weather data itself. We will maintain distance in the model as we hope to consider interaction terms with this parameter.



(a) Relative Humidity AVP

(b) Wind Speed AVP

Figure 6: Added Variable Plots

As we hope to consider whether the heat index effects different race distances differently, we will add an interaction term between the continuous predictor `heat_index` and the categorical predictor `distance`. From the ANOVA tests at a family-wise error rate of 5%, we see that the new interaction term is considered significant (Table 3). While the main effect `distance` is not significant, we will keep it in the model to maintain the principal of marginality.

We will now assess the fit of our model. The residual plots in Figure 7 do not suggest any serious violations from the assumptions of equal variance and independence of errors in the multiple linear regression model. The Normal QQ-Plot does suggest that the normality assumption may be violated as the residuals distribution has heavier tails. A Yeo-Johnson Box-Cox transformation was considered to resolve this issue. However, because the optimizing value of $\lambda = 0.8282$ was near 1 and did not show significant improvement in model fit, we chose to maintain our non-transformed

Table 2: VIF for Reduced Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sex	1	80.27799	80.277989	14.626516	0.0001440
competition	5	94.26698	18.853397	3.435058	0.0045250
distance	5	82.80521	16.561043	3.017394	0.0105980
wind_speed	1	17.82424	17.824239	3.247547	0.0720053
heat_index	1	127.95270	127.952697	23.312768	0.0000017
rel_hum	1	31.91861	31.918610	5.815518	0.0161686
proximity	1	15.19605	15.196050	2.768695	0.0966186
Residuals	634	3479.72449	5.488524	NA	NA

Table 3: ANOVA Test for Interaction Model

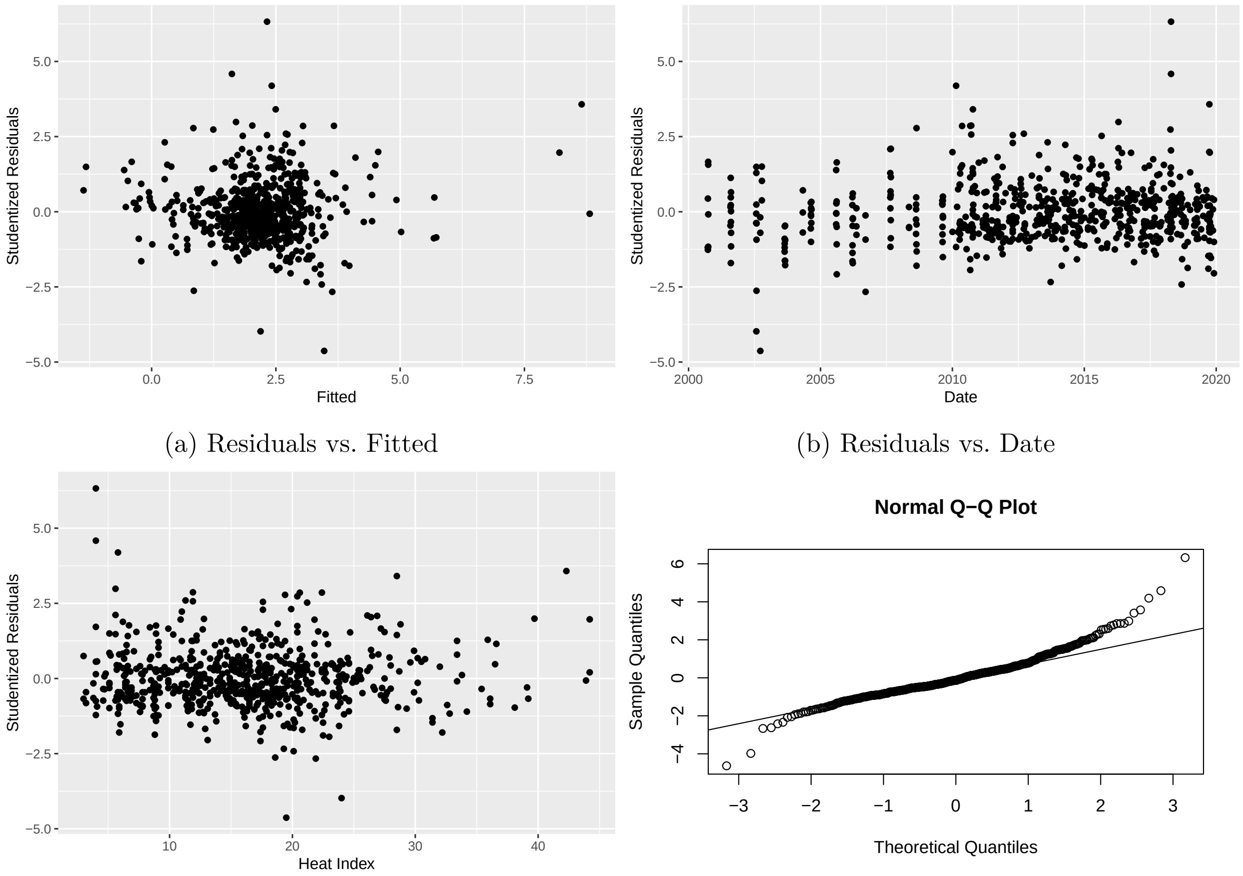
	DF	SS	MS	F-Value	P-Value
sex	1	80.27799	80.277989	15.348651	0.0000991
competition	5	94.26698	18.853397	3.604652	0.0031900
distance	5	82.80521	16.561043	3.166368	0.0078398
heat_index	1	118.65599	118.655991	22.686286	0.0000024
distance:heat_index	5	248.41307	49.682613	9.499006	0.0000000
Residuals	632	3305.54703	5.230296	NA	NA

data.

Because our initial exploratory analysis found potential high leverage and outlier cases, we will test for outliers using the studentized residuals and identify highly influential cases using Cook's distance. In Figure 8, we see that many cases have leverages greater than $3(p/n)$ and may be considered potentially influential. However, only one case has a significant Cook's distance (near 1), indicating that this case has a larger influence on the estimated parameter coefficients. Evaluating the studentized residuals, we see that three cases have residuals with absolute values greater than $F_{1-\alpha^*,n-p-1}$ (dotted lines), where α^* is the Bonferroni corrected error rate to account for testing all n observations. Reviewing each of the four identified outliers, we find that the 2019 World Championship 50K race-walk and the 2018 Men's and Women's Boston Marathons are included. We will remove these observations as they reflect extreme and uncommon weather conditions in which many of the top performers opted to drop out of the race. The fourth race did not appear to have an obvious or sound reason for removal, therefore we will not remove it from our data set.

```
Warning: Removed 1 rows containing missing values (`geom_point()`).
```

After refitting our model without the influential cases, we reviewed the updated model and residual diagnostics. The plots did not suggest any significant changes in model fit. After observing the slight curvature in the relationship between heat index and percent deviation, we fit a model including the squared heat index as an additional term. While this parameter did improve model fit, we did not believe the improvement outweighed the cost in interpretability. We also acknowledge that the quadratic relationship appears to be more pronounced for the 20K and 50K race-walks and marathons, as evident in Figure 5(b). Given more time to complete our data set, we would have liked to have expanded our model to individual distances and improve our predictive performance.



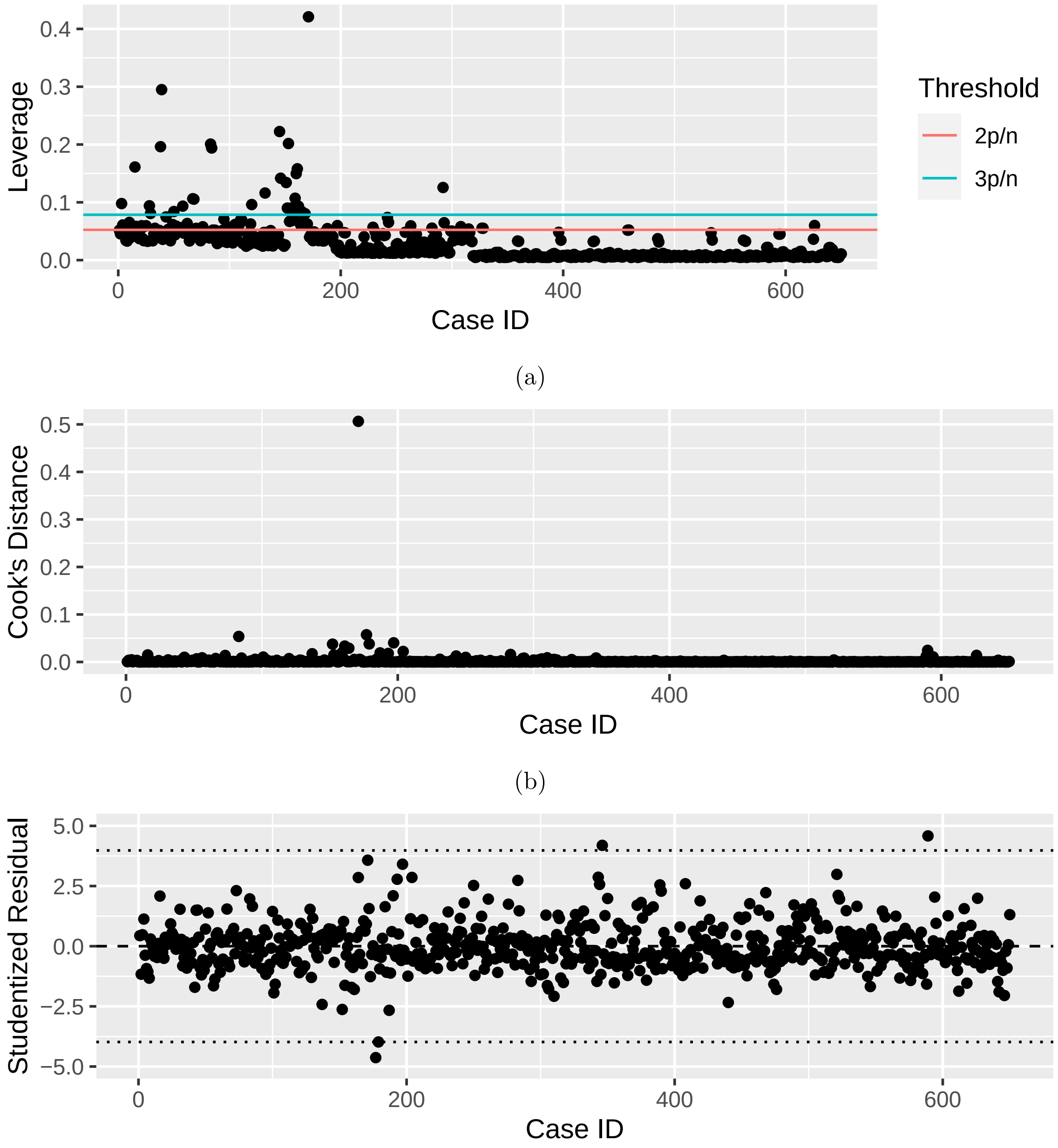
(a) Residuals vs. Fitted

(b) Residuals vs. Date

(c) Residuals vs. Heat Index

(d) Normal QQ-Plot

Figure 7: Model Diagnostics



(a)

(b)

(c)

Figure 8: Residual Diagnostics

Table 4: Final Model Summaries

(a) Coefficient Estimates

	Estimate	SE	T-Statistic	P-Value
(Intercept)	1.2184	1.3374	0.9110	0.3627
sexWomen	0.6902	0.1733	3.9836	0.0001
competitionDiamond League	0.7543	0.4438	1.6996	0.0897
competitionGold Marathon	0.3077	0.5347	0.5755	0.5652
competitionOlympic	-0.9857	0.4726	-2.0858	0.0374
competitionWorld Championships	0.4694	0.3990	1.1765	0.2399
competitionWorld Cup	1.4288	0.5220	2.7373	0.0064
distance20k	-7.8195	1.7979	-4.3492	0.0000
distance3k	1.3245	1.6935	0.7821	0.4344
distance50k	-2.1033	2.3541	-0.8935	0.3720
distance5k	-1.3179	1.5841	-0.8320	0.4057
distanceMarathon	-0.4357	1.4916	-0.2921	0.7703
heat_index	-0.0223	0.0622	-0.3576	0.7208
distance20k:heat_index	0.3453	0.0771	4.4791	0.0000
distance3k:heat_index	-0.0455	0.0759	-0.5993	0.5492
distance50k:heat_index	0.1307	0.1057	1.2365	0.2167
distance5k:heat_index	0.0905	0.0730	1.2395	0.2156
distanceMarathon:heat_index	0.0954	0.0663	1.4389	0.1507

(b) Model Fit

R^2	sigma	DF	N
0.156	2.1598	17	647

Results and Discussion

Our final selected model includes sex, competition, distance, heat index, and the interaction between distance and heat index to predict the percent deviation between the top three average finishing times and the standing event record. In Table 4(a), we see that while our initial ANOVA tests suggested statistically significant parameters, many of the individual categorical parameters are not significant. However, we will continue with this model and note its limitations.

From our model, we may observe that females typically have a mean "top-three average" percent deviation 0.6902 percentage units greater than males, when holding distance, competition, and heat index constant. This indicates that females underperform–relative to the standing event record–by greater margins than males.

When holding distance, gender, and heat index constant, we may be interested in the difference in average percent deviation between championship races that require qualification and include the highest level athletes for a given year and non-championship races that may not require qualification or be the primary focus of an athlete's racing campaign. We will construct a 95% confidence interval for the difference in the effect of an even being an Olympic competition versus being a World Cup competition. Using the results in Table 4 and the `vcov()` function, our interval will take the form

$$\left[(-0.9857 - 1.4288) \pm t_{629,0.975} \cdot \sqrt{(0.4726)^2 + (0.5220)^2 - 2(0.1008)}\right] \implies (-2.8675, -1.9616)\,.$$

Therefore, we are 95% confident that an Olympic race typically has an average percent deviation 1.9616 to 2.8675 percentage units lower than a World Cup race. In other words, finishers in Olympic races typically perform closer to (or better than) the standing Olympic record compared to finishers in World Cup races.

Two of the most statistically significant parameters in our model include the 20K race-walk distance as a component. We see that relative to the baseline 10K running distance, a one °C increase in heat index is associated with a 0.345 percentage unit greater increase in average percent deviation. Without conducting formal tests, we see that in general, the 20K race-walk appears to be effected greater by the heat index compared to the other distance categories. Additionally, without considering the interaction with race distance, heat index by itself does not appear to be a significant contributor to average percent deviation. This leads back to our interest in exploring separate models for each race distance and conducting polynomial regression to explore more complex relationships between heat index (or other weather predictors) and performance.

Unfortunately, our current model is limited due to both the lack of certain race distances and the lack of depth in race results. It is possible that elite athletes (top three finishers) are not as affected by the weather as average runners and including mass participation results of Gold Label marathons could provide insight into general runner performance. We believe these limitations may also partially contribute to the low explanatory power of our model. In addition to polynomial regression, we may also explore generalized linear models or nonparametric regression to deal with the violation of the normality assumption.

Appendix

```r
knitr::opts_chunk$set(fig.height = 2, fig.width = 6,
                      echo = FALSE, fig.align = 'center',
                      message = FALSE, number_sections = FALSE)
library(alr4)
library(tidyverse)
library(ggplot2)
library(cowplot)
library(kableExtra)
library(tidymodels)
library(lemon)
library(MASS)
library(GGally)
library(dplyr)
library(hms)
library(corrplot)
## pull in data and clean
results = read.csv("race_results.csv", header = TRUE)

## rename columns for easier reading
colnames(results) = c("competition", "distance", "sex", "host", "country",
                      "day", "month", "year", "time",
                      "latitude", "longitude", "NOAA_ID", "station_loc",
                      "dist_from_loc", "air_temp", "dew_point", "wind_speed",
                      "adj_wind_speed", "rel_hum", "clouds_OKTA",
                      "diff_from_req_time", "time_zone", "solar_rad",
                      "heat_index", "simp_wet_bulb", "wet_bulb",
                      "world_record", "standing_record", "time_1st",
                      "time_2nd", "time_3rd", "time_4th", "time_5th",
                      "time_6th", "time_7th", "time_8th", "time_9th",
                      "time_10th", "time_avgTop3")

## remove missing values
weather_missing = which(is.na(results$NOAA_ID) | is.na(results$rel_hum))

results_missing = which(results$world_record == "" |
                        results$standing_record == "" |
                        results$time_1st == "")
results = results[-c(weather_missing, results_missing),]

## convert time string to time object
results$world_record = as_hms(results$world_record)
results$standing_record = as_hms(results$standing_record)
results$time_1st = as_hms(results$time_1st)
results$time_avgTop3 = as_hms(results$time_avgTop3)

## convert time to seconds
results$world_record_s = as.numeric(results$world_record)
results$standing_record_s = as.numeric(results$standing_record)
results$time_1st_s = as.numeric(results$time_1st)
results$time_avgTop3_s = as.numeric(results$time_avgTop3)

## add proximity variable
results$proximity = ifelse(results$dist_from_loc <= 10, "close", "far")

## percent deviation from world and standing record
results = results %>%
  mutate(percent_world_1st = (time_1st_s - world_record_s) /
           world_record_s * 100,
         percent_standing_1st = (time_1st_s - standing_record_s) /
           standing_record_s * 100,
         percent_standing_avg = (time_avgTop3_s - standing_record_s) /
           standing_record_s * 100)

## make overall date of competition variable
results = results %>%
  mutate(date = make_date(year, month, day))
# review proximity distribution
```

```r
table(results$proximity)
#summary(results$percent_standing_1st)
ggplot(data = results, aes()) +
  geom_histogram(aes(x = percent_standing_avg)) +
  xlab("Percent Deviation between Top Three Average and Standing Record") +
  ylab("Count")
weather = c("air_temp", "dew_point", "wind_speed",
            "adj_wind_speed", "rel_hum", "clouds_OKTA", "solar_rad",
            "heat_index", "simp_wet_bulb", "wet_bulb")
temp = results[results$proximity == "close" ,c(weather, "percent_standing_avg")]
M.under = cor(temp, use = "na.or.complete")
corrplot::corrplot(M.under, method = "ellipse", type = "lower",
         tl.cex = 1, cl.cex = 0.5, mar = c(0,0,0,0))
temp = results[results$proximity == "far" ,c(weather, "percent_standing_avg")]
M.over = cor(temp, use = "na.or.complete")
corrplot::corrplot(M.over, method = "ellipse", type = "lower",
         tl.cex = 1, cl.cex = 0.5, mar = c(0,0,0,0))
ggpairs(data = results[, c("air_temp", "dew_point", "wind_speed", "heat_index", "simp_wet_bu
  theme(strip.text.x = element_text(size = 5),
        strip.text.y = element_text(size = 4))
## histogram of the distances
ggplot(data = results, aes()) + geom_boxplot(aes(x = dist_from_loc))

## proximity vs. response
ggplot(data = results, aes()) + geom_boxplot(aes(x = proximity, y = percent_standing_1st))

## subset dataframe based on values in dist_from_loc
under10 <- subset(results, proximity == "close")
above10 <- subset(results, proximity == "far")

## weather predictor scatterplots
ggpairs(data = under10[, c(weather[1:5], "percent_standing_1st")])
ggpairs(data = above10[, c(weather[1:5], "percent_standing_1st")])

ggpairs(data = under10[, c(weather[5:10], "percent_standing_1st")])
ggpairs(data = above10[, c(weather[5:10], "percent_standing_1st")])

ggpairs(data = results[, c(weather[1:5], "percent_standing_1st")])
ggpairs(data = results[, c(weather[5:10], "percent_standing_1st")])
## count table of sex, competition, and distance variables
cat.counts = as.data.frame(table(results$distance, results$competition, results$sex))
cat.counts = cat.counts %>% pivot_wider(names_from = c("Var1", "Var3"), values_from = "Freq"
cat.counts$Var2 = c("Common Wealth", "Diamond League", "Gold Marathon",
                    "Olympic", "World Champ.", "World Cup")

table = cat.counts %>% kable(col.names = c("Comp.", "10K", "20K", "3K", "50K", "5K", "Mar.",
                                           "10K", "20K", "3K", "50K", "5K", "Mar."), booktab
  add_header_above(c(" ", "Men" = 6, "Women" = 6)) %>%
  kable_styling(latex_options = "striped", full_width = FALSE)
table
ggplot(data = results, aes()) +
  geom_boxplot(aes(x = competition, y = percent_standing_avg)) +
  xlab("Competition") +
  ylab("Percent Deviation") +
  theme(axis.text.x = element_text(size = 7))
p1 = ggplot(data = results, aes()) +
  geom_point(aes(x = date, y = percent_standing_avg, color = distance),
             alpha = 0.4) +
  facet_grid(rows = vars(sex)) +
  xlab("Date") +
  ylab("Percent Deviation") +
  labs(color = "Distance") +
  theme(legend.position = "bottom")
p2 = ggplot(data = results) +
  geom_point(aes(x = heat_index, y = percent_standing_avg, color = competition),
             alpha = 0.4) +
  xlab("Heat Index") +
  ylab("Percent Deviation") +
  labs(color = "Competition") +
  theme(legend.position = "bottom")

p1
p2
potential <- c("sex", "competition", "distance",
              "air_temp", "dew_point", "rel_hum",
              "wind_speed", "heat_index", "simp_wet_bulb",
              "proximity", "percent_standing_avg")

results.potential <- results[,potential]

nullMod <- lm(percent_standing_avg ~ 1, data = results.potential)
fullMod <- lm(percent_standing_avg ~ ., data = results.potential)

stepMod <- stepAIC(object = fullMod, direction = "both", k = 2, trace = FALSE)
forwardMod <- stepAIC(object = nullMod, scope = list(upper = fullMod), direction = "forward"
backwardMod <- stepAIC(object = fullMod, direction = "backward", k = 2, trace = FALSE)
# summary(stepMod)
# summary(forwardMod)
# summary(backwardMod)

#add table with including predictors marked for each model
## vif
```

```r
kable(vif(stepMod))%>%
  kable_styling(latex_options = "striped", full_width = FALSE)
## reduced model
reducedMod <- lm(percent_standing_avg ~ sex + competition + distance + wind_speed + heat_ind
## added variable plots
avPlot(reducedMod, "rel_hum",
       main = "")
avPlot(reducedMod, "wind_speed",
       main = "")

kable(anova(reducedMod))%>%
  kable_styling(latex_options = "striped", full_width = FALSE)
# kable(vif(reducedMod))%>%
#   kable_styling(latex_options = "striped", full_width = FALSE)
## Add interaction term.
intMod <- lm(percent_standing_avg ~ sex + competition + distance + heat_index + distance:hea

# summary(intMod)
kable(anova(intMod),
      col.names = c("DF", "SS", "MS", "F-Value", "P-Value"))%>%
  kable_styling(latex_options = "striped", full_width = FALSE)
base = augment(intMod, results) %>%
  mutate(.ti = rstudent(intMod)) %>%
  rownames_to_column("case") %>%
  ggplot()

# Model diagnostics for initial interaction model.
base + geom_point(aes(x = .fitted, y = .ti)) +
  labs(x = "Fitted", y = "Studentized Residuals")
base + geom_point(aes(x = date, y = .ti)) +
  labs(x = "Date", y = "Studentized Residuals")
base + geom_point(aes(x = heat_index, y = .ti)) +
  labs(x = "Heat Index", y = "Studentized Residuals")
qqnorm(rstudent(intMod))
qqline(rstudent(intMod))


# Consider Box-Cox transformation using Yeo-Johnson family to allow for negative response va
bc <- boxCox(intMod, family="yjPower", plotit = TRUE)
## Optimal lambda value. 0.9091
lambda <- bc$x[which.max(bc$y)]


## Code from Professor Hans: `residuals_diagnostics`
base + geom_point(aes(x = as.numeric(case), y = .hat)) +
  geom_hline(data = tibble(Threshold = c("2p/n", "3p/n"),
                           ref = c(2*17/dim(results)[1], 3*17/dim(results)[1])),
             aes(yintercept = ref, color = Threshold)) +
  labs(x = "Case ID", y = "Leverage")
base + geom_point(aes(x = as.numeric(case), y = .cooksd)) +
  labs(x = "Case ID", y = "Cook's Distance")
base + geom_point(aes(x = as.numeric(case), y = .ti)) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  geom_hline(yintercept = qt(1 - (0.05/(2*dim(results)[1])), df = 632 - 1), linetype = "dott
  geom_hline(yintercept = qt((0.05/(2*dim(results)[1])), df = 632 - 1), linetype = "dotted")
  ylim(-5,5) +
  labs(x = "Case ID", y = "Studentized Residual")


# Make temporary dataframe to find the cases with the extreme Cook's distance and studentize
outlierTestDF = augment(intMod, results) %>%
  mutate(.ti = rstudent(intMod)) %>%
  rownames_to_column("case")
# Sort studentized residuals. Cases 590, 589, and 177 are our identified outliers.
head(outlierTestDF[order(outlierTestDF$.ti, decreasing = TRUE), c(".ti", "case")])
head(outlierTestDF[order(outlierTestDF$.ti, decreasing = FALSE), c(".ti", "case")])
# Sort Cook's distances. Case 171 is our identified influencer.
head(outlierTestDF[order(outlierTestDF$.cooksd, decreasing = TRUE), c(".cooksd", "case")])
# Look at "outlier" race results.
outliers = c(171,177,590,589)
results[which(outlierTestDF$case %in% outliers),c("date", "host", potential)]

# Refit model without influential case.
results.new <- results[-which(outlierTestDF$case %in% c(171,589,590)),]
intMod.new <- lm(percent_standing_avg ~ sex + competition + distance + heat_index  + distanc

# Review ANOVA, residual diagnostics, and model diagnostics.
anova(intMod.new)
base = augment(intMod.new, results.new) %>%
  mutate(.ti = rstudent(intMod.new)) %>%
  rownames_to_column("case") %>%
  ggplot()
base + geom_point(aes(x = .fitted, y = .ti))
base + geom_point(aes(x = date, y = .ti))
base + geom_point(aes(x = heat_index, y = .ti))
qqnorm(rstudent(intMod.new))
qqline(rstudent(intMod.new))

base + geom_point(aes(x = as.numeric(case), y = .hat)) +
  geom_hline(data = tibble(threshold = c("2p/n", "3p/n"),
                           ref = c(2*17/dim(results.new)[1], 3*17/dim(results)[1])),
             aes(yintercept = ref, color = threshold))
```

```
base + geom_point(aes(x = as.numeric(case), y = .cooksd))
base + geom_point(aes(x = as.numeric(case), y = .ti)) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  geom_hline(yintercept = qt(1 - (0.05/(2*dim(results.new)[1])), df = 629 - 1), linetype = "
  geom_hline(yintercept = qt((0.05/(2*dim(results.new)[1])), df = 629 - 1), linetype = "dott
  ylim(-5,5)


# Polynomial approach.
testMod <- lm(percent_standing_avg ~ sex + competition + distance + heat_index, data = resul

summary(testMod)
kable(anova(testMod))
kable(tidy(intMod.new), digits = 4,
      col.names = c("", "Estimate", "SE", "T-Statistic", "P-Value"))%>%
  kable_styling(latex_options = "striped", full_width = FALSE)
kable(glance(intMod.new)[,c("r.squared", "sigma", "df", "nobs")], digits = 4,
      col.names = c(expression(R^2), expression(sigma), "DF", "N"))%>%
  kable_styling(latex_options = "striped", full_width = FALSE)
# Competition difference confidence interval.
est = -0.9857 - 1.4288
t.mult = pt(0.975, 629)
vcov(intMod.new)
se = sqrt(0.4726^2 + 0.5220^2 - 2*0.1008)
compCI = c(est - t.mult*se,
           est + t.mult*se)
```