PHP 2550 Project 1

Peirong Hao

Introduction

Endurance exercise performance, particularly in long-distance events like marathons, is influenced by a complex interplay of physiological, environmental, and demographic factors. Previous research demonstrates that heat stress slows pacing during aerobic exercise, as it increases core temperature and cardiovascular strain, leading to fatigue and a reduction in overall performance capacity [1]. The impact of environmental heat stress is quantified through metrics like the Wet-Bulb Globe Temperature (WBGT), a weighted average of dry bulb, wet bulb, and globe temperature. [2] highlights a negative relationship between WBGT and marathon outcomes, with performance declining progressively as WBGT levels rise. These findings underscore the critical role of environmental conditions in determining marathon success.

Age significantly influences performance under heat stress. Older adults face heightened challenges in thermoregulation due to diminished sweat production, reduced cardiovascular efficiency, and lower overall heat dissipation capabilities [3]. These limitations impair their ability to maintain performance in warmer conditions, particularly in prolonged endurance activities like marathons. Understanding how age interacts with other environmental variables could provide valuable insight into older athletes' vulnerabilities and needs.

Sex differences add another layer of complexity to endurance performance. While women generally have lower aerobic fitness and maximal oxygen uptake (VO2 max) than men [4], they demonstrate distinct advantages, including more consistent pacing strategies and reduced fatigue during endurance running. Even though these characteristics can enhance women's ability to sustain performance over long distances, these benefits are counterbalanced by lower oxygen-carrying capacity and higher body fat percentages, limiting overall performance potential [5]. Exploring these differences can inform tailored strategies for improving endurance outcomes across sexes.

This project, conducted in collaboration with Dr. Brett Romano Ely and Dr. Matthew Ely from the Department of Health Sciences at Providence College, aimed to build on these findings by investigating the relationships between environmental factors, age, and sex in marathon performance. Using data from five major marathons collected over the years, this study examined how temperature, humidity, solar radiation, and wind influence marathon performance in men

Table 1: Summary table of weather parameters and participants' characteristics for all races

Characteristic	Boston N = 2,088	Chicago N = 2,553	Grandmas N = 2,000	NYC N = 2,930	TC N = 1,993	p-value
Flag						
White	1,040 (50%)	732 (30%)	0 (0%)	1,394 (50%)	587 (31%)	
Green	810 (39%)	1,459 (60%)	702 (37%)	901 (32%)	834 (44%)	
Yellow	115 (5.5%)	120 (4.9%)	945 (50%)	504 (18%)	338 (18%)	
Red	123 (5.9%)	116 (4.8%)	237 (13%)	0 (0%)	116 (6.2%)	
Black	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
Dry bulb temperature in Celcius	11.64 (5.89)	12.42 (6.05)	18.86 (3.32)	11.73 (4.67)	13.14 (5.52)	<0.001
Wet bulb temperature in Celcius	7.59 (3.81)	8.54 (5.68)	14.91 (2.45)	7.57 (4.98)	9.85 (5.41)	<0.001
Black globe temperature in Celcius	24.22 (8.38)	24.52 (6.30)	31.63 (7.86)	21.36 (5.93)	24.94 (6.52)	<0.001
Solar radiation in Watts per meter squared	649.80 (186.92)	460.48 (94.56)	676.81 (190.51)	401.15 (130.90)	435.89 (138.93)	<0.001
Dew Point in Celcius	3.32 (4.47)	4.65 (6.86)	12.43 (3.17)	2.74 (6.99)	5.96 (7.25)	<0.001
Wind speed (km/hr)	11.99 (4.47)	8.21 (3.19)	9.16 (2.87)	11.22 (4.55)	8.80 (3.20)	<0.001
Wet Bulb Globe Temperature (WBGT)	11.32 (4.53)	12.12 (5.76)	18.65 (3.22)	10.74 (4.91)	13.20 (5.35)	<0.001
Course Record(hr)	2.35 (0.01)	2.31 (0.03)	2.46 (0.02)	2.39 (0.02)	2.45 (0.01)	<0.001
Sex						0.8
F	984 (47%)	1,210 (47%)	934 (47%)	1,402 (48%)	922 (46%)	
M	1,104 (53%)	1,343 (53%)	1,066 (53%)	1,528 (52%)	1,071 (54%)	
Completion Time (hr)	3.31 (0.79)	3.50 (1.08)	3.63 (0.98)	3.71 (1.34)	3.57 (0.90)	<0.001
Age						
14 and under	0 (0%)	5 (0.2%)	12 (0.6%)	0 (0%)	7 (0.4%)	
15-19	67 (3.2%)	174 (6.8%)	164 (8.2%)	92 (3.1%)	142 (7.1%)	
20-29	360 (17%)	420 (16%)	340 (17%)	460 (16%)	339 (17%)	
30-39	360 (17%)	420 (16%)	340 (17%)	460 (16%)	340 (17%)	
40-49	360 (17%)	420 (16%)	340 (17%)	460 (16%)	340 (17%)	
50-59	359 (17%)	420 (16%)	340 (17%)	460 (16%)	339 (17%)	
60-69	337 (16%)	410 (16%)	304 (15%)	458 (16%)	314 (16%)	
70-79	215 (10%)	253 (9.9%)	142 (7.1%)	397 (14%)	154 (7.7%)	
80 and over	30 (1.4%)	31 (1.2%)	18 (0.9%)	143 (4.9%)	18 (0.9%)	

[1] n (%); Mean (SD)

[2] Kruskal-Wallis rank sum test; Pearson's Chi-squared test

and women. The overarching goal was to provide a more comprehensive understanding of the determinants of marathon success, offering insights to optimize training and environmental adaptation across diverse athlete populations.

Exploratory Data Analysis (EDA)

Completion times and weather variables varied significantly across races, as reflected by p-values below 0.001 for most characteristics. For example, Grandma's Marathon had the highest average dry bulb temperature at $18.86°C$, while Boston and NYC reported lower averages of $11.64°C$ and $11.73°C$, respectively. Solar radiation also showed notable differences, with Grandma's Marathon recording the highest levels at 676.81 W/m^2 and NYC the lowest at 401.15 W/m^2. Wind speed followed a similar trend, with Boston experiencing the highest average at 11.99 km/hr, compared to Chicago, which had the lowest at 8.21 km/hr. These environmental variations appeared to influence completion times. Runners in Boston achieved the fastest average completion time of 3.3 hours, while participants in New York City had the slowest average time of 3.7 hours.

There were four datasets, including environmental conditions and participants' performances

from five major marathons (Boston, Chicago, New York, Twin Cities, and Grandma's Marathons) spanning 1993 to 2016. In total, the data contained 11,564 runners who participated in 98 marathons over the years. As shown in Table 1, weather variables and course records were statistically significantly different across races.

Table 2: Missingness by race

Year	Race	Number of observations
2011	Chicago	126
2011	NYC	131
2011	TC	118
2012	Grandmas	116

Four matches (2011 Chicago, 2011 NYC, 2011 TC, 2012 Grandmas) had missing weather parameters. Since the missingness was unrelated to other data, it was considered as missing completely at random (MCAR).

Table 3: Single variable missingness table in AQI dataset

variable	n_miss	pct_miss
88502	26	27.1
88101	21	21.9

The air quality index (AQI) dataset contained a significant number of missing values. It is important to note that although PM2.5 was coded in 88502, it had more missing values than 88101, the primary PM2.5 measurement. Because Ozone and PM2.5 measurements were recorded at different sublocations and times of the day for each city, I calculated an average of each parameter for each marathon date and site.

The first aim was to examine the effects of increasing age on marathon performance in men and women. CR is the course record for each year's marathon, obtained from last year's record. %CR refers to the percent course record. A positive %CR means the percentage more time used to complete the marathon. A negative %CR means the participant breaks the record. I calculated the completion time for each individual using CR * (1+%CR). All record-breakers were between the ages of 21 and 41. Males broke the record in a younger age range (21 to 34 years old), but females could break the record in the broader age group (23 to 41 years old).

Figure 1 reveals a clear U-shaped relationship between age and marathon completion time, with faster times observed for individuals aged 20 to 40. Men generally completed marathons faster than women, as indicated by the blue line consistently below the red line and their lower minimum completion time (2.68 hours at age 39.2 for men versus 2.83 hours at age 37.2 for women). Table 4 further highlights this trend: completion times decrease until the minimum point and then increase at a faster rate with age for both genders.

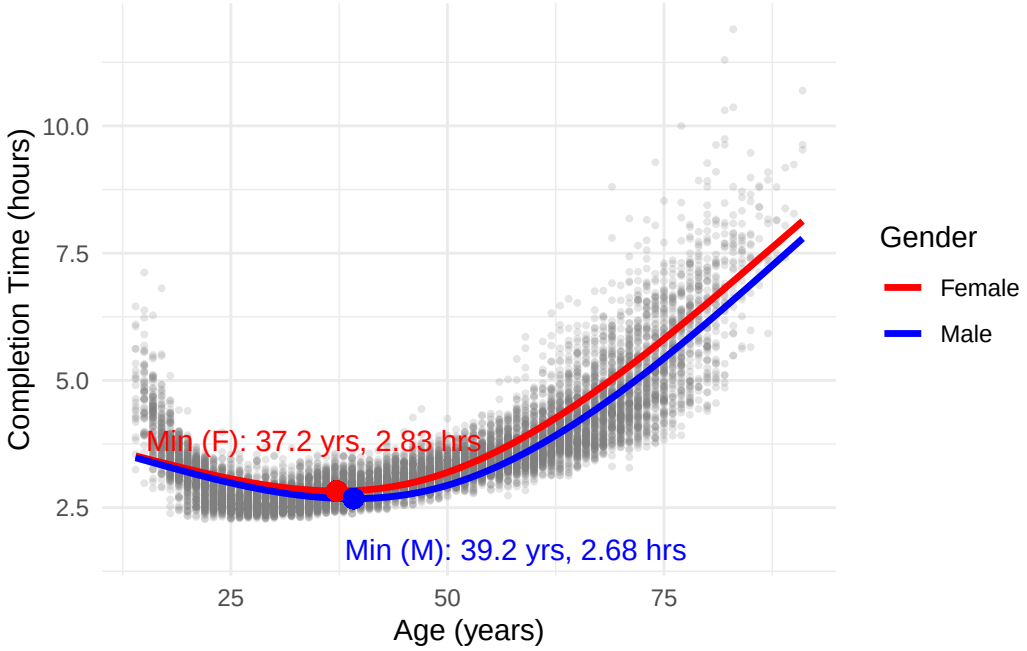

Figure 1: Spline plot of completion time vs age

Table 4: Slopes of spline plot

Age	Female	Male
25	-0.035	-0.039
50	0.060	0.051
75	0.137	0.137

Regarding the completion time of each match, some people in the elder age group took longer than 9 hours to complete, with the longest time being 11 hours and 54 minutes. These outliers could be due to some data collection error since Boston, Twin Cities, and Chicago marathons must be completed within 6 hours, 6 hours and 15 minutes, 6 and a half hours, respectively. However, Grandma's and NYC marathons allow participants to continue finishing on the sidewalk after 7 hours and 8 hours limit, respectively, without services, including cross-street protection, medical assistance, and aid stations. Thus, even though those outliers weren't official finishing times, they could be valid self-reported completion times.

The second aim was to explore the impact of environmental conditions on marathon performance and examine if the effect differs across age and gender. Wet Bulb Globe Temperature (WBGT) is derived from three different sources: Wet bulb temperature (Tw, which accounts for humidity), Black globe temperature (Tg, which accounts for solar radiation), and Dry bub temperature (Td). Flag is determined by WBGT and the chance of heat illness. The Flag

variable has five categories arranged from best to worst condition: White, Green, Yellow, Red, and Black. We did not have any data falling in the Black condition, probably because Marathons were canceled in extreme weather.

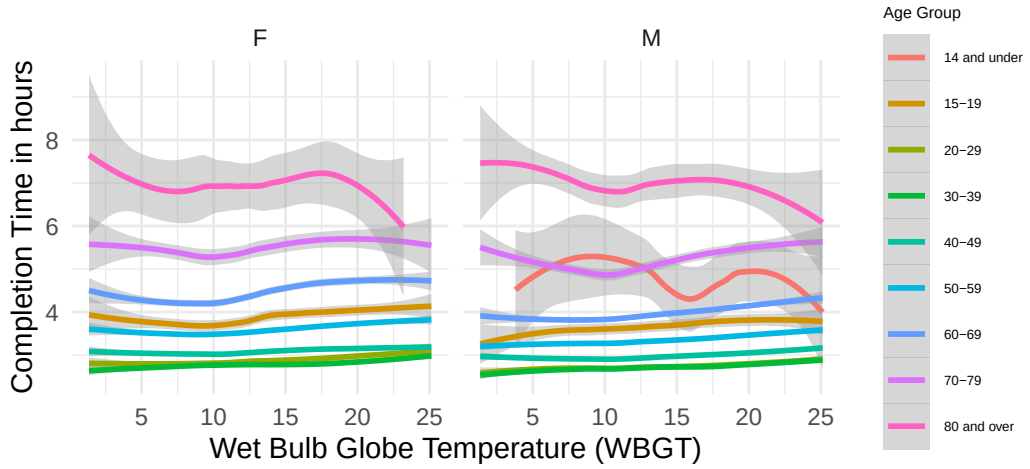

Figure 2: Impact of temperature on completion time

As shown in Figure 2, heat stress, measured by WBGT, significantly impacts marathon performance, with older runners exhibiting more fluctuations in completion times. In contrast, younger and middle-aged runners were less affected, showing relatively stable completion times, though with a slight increasing trend as WBGT rises. While men generally outperform women, both genders exhibit similar patterns of performance.

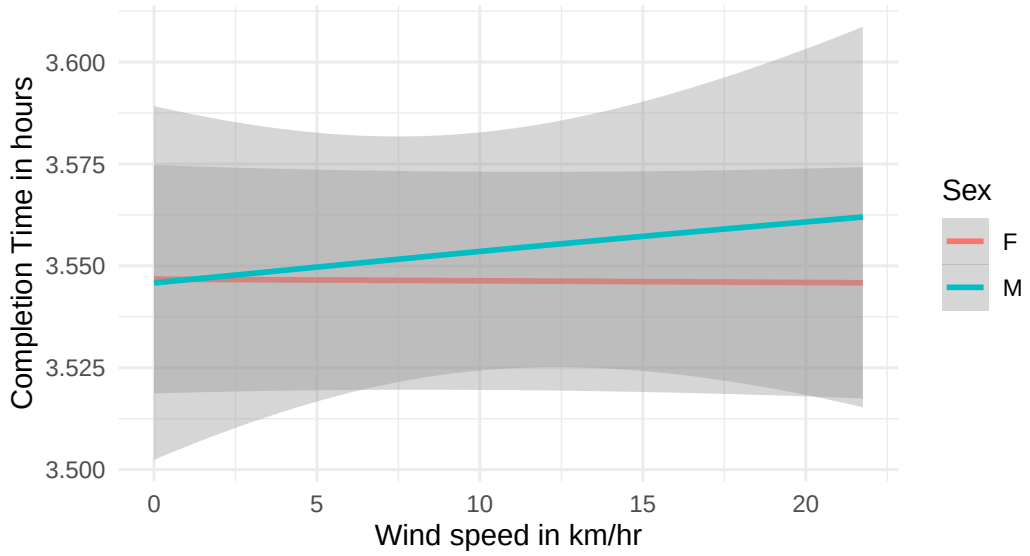

Figure 3: Impact of wind on completion time

Another important weather parameter is wind speed, measured in km/h. In Figure 3, there is a slight difference between the effects of wind speed on completion time for men and women. Women were less affected by changes in wind speed, as indicated by their relatively flat trend line.

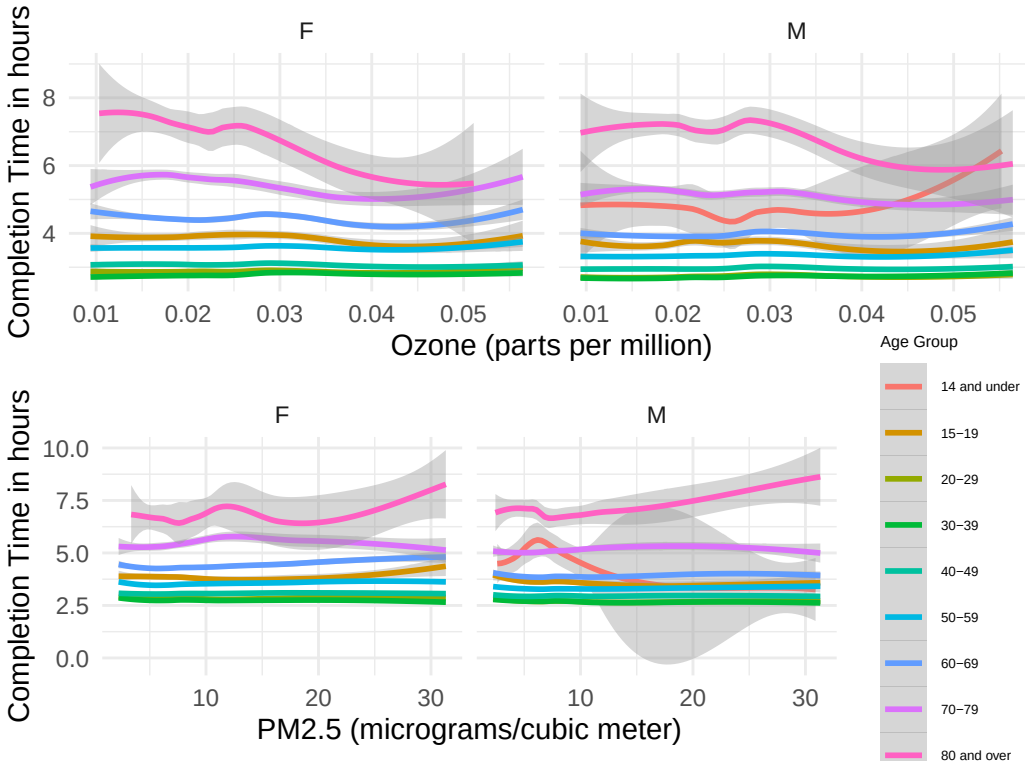

Figure 4: Impact of air pollutants' condition on completion time

Ozone Solutions suggested the safe limit for ozone is 0.1 ppm, and our Ozone data were well below that threshold. The relationships between Completion Time and Ozone held constant for most age groups but varied slightly for the youngest and eldest groups in Figure 6. According to Indoor Air Hygiene Institute, PM2.5 at or under 12 micrograms/cubic meter is considered healthy, but it is deemed unhealthful when it reaches or surpasses 35 micrograms/cubic meter. Because the recorded PM2.5 data were below 35 micrograms/cubic, there was not much health concern about it. Concerning variation in completion time with PM2.5, there was little change for the majority but slight variation for the youngest and eldest groups. These relations held for both sex groups.

The final aim was to identify the weather parameters most affecting marathon performance. No weather parameter showed a strong correlation with completion time, nor were there particularly strong correlations among the weather parameters themselves. Due to the sex differences found earlier, I built two hierarchical random intercept (differ by race) models for females and males. The outcome of interest was `completion time`, and I aimed to find its linear

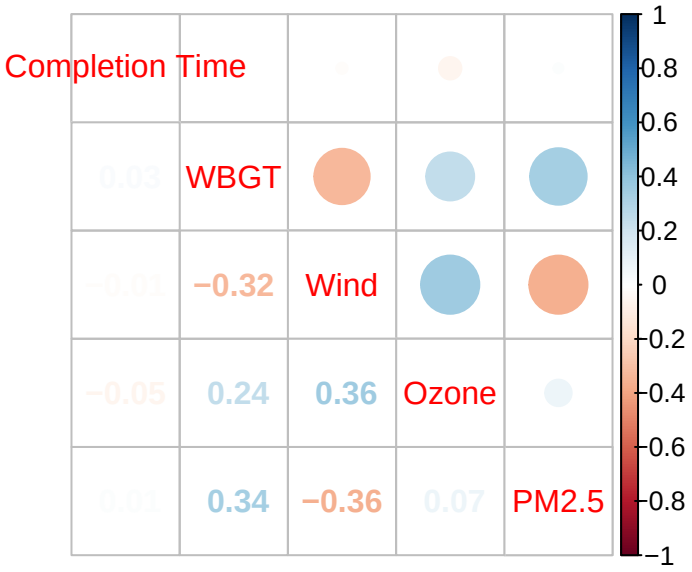

Figure 5: Correlation plot

relationship with weather variables, age, and age squared.

Table 5: Fixed Effects - Female

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	12048.356	264.542	6.368	45.544	0.000
poly(age, 2)1	157477.160	1615.057	4057.567	97.506	0.000
poly(age, 2)2	122611.676	1609.757	4056.761	76.168	0.000
WBGT	23.264	6.489	3881.919	3.585	0.000
Wind	30.506	7.857	4055.833	3.883	0.000
PM2.5	31.153	5.151	3917.514	6.048	0.000
Ozone	-6126.445	4613.218	3315.682	-1.328	0.184

Table 6: Random Effects - Female

	(Intercept)
Boston	-663.126
Chicago	-24.806
Grandmas	559.812
NYC	-319.359
TC	447.477

Table 7: Fixed Effects - Male

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	12049.490	279.817	5.966	43.062	0.000
poly(age, 2)1	193570.023	1684.789	4517.006	114.893	0.000
poly(age, 2)2	151253.002	1679.439	4516.309	90.062	0.000
WBGT	18.743	6.450	4364.965	2.906	0.004
Wind	25.558	7.804	4516.416	3.275	0.001
PM2.5	19.316	5.155	4393.981	3.747	0.000
Ozone	2514.153	4577.486	3854.872	0.549	0.583

Table 8: Random Effects - Male

	(Intercept)
Boston	-796.355
Chicago	-189.350
Grandmas	543.717
NYC	-67.987
TC	509.975

In terms of fixed effects, the intercept and coefficients for age, age squared, WBGT, wind speed, and PM2.5 are similar in both the female and male models, with the male model showing slightly smaller coefficient values. While the sign of the ozone coefficient differs between the two models, the p-values indicate that these estimates are not statistically significant. The interpretation of the significant weather parameters for the female model is as follows: for each additional unit increase in WBGT, the completion time is expected to increase by 23.3 seconds, holding all other variables constant. For each additional unit increase in wind speed, the completion time is expected to increase by 30.5 seconds, holding all other variables constant. For each additional unit increase in PM2.5, the completion time is expected to increase by 31.2 seconds, holding all other variables constant. The male model shows comparable patterns in the relationships between weather factors and completion times. The random effects structure reveals similar variations across race locations for both gender-specific models, though the male model exhibits a slightly higher variance in the random intercepts (312832) compared to the female model (269360).

Conclusion

People aged between 20 and 40 run much faster than the other age groups and could break course records. Age and weather parameters have a positive linear relationship with completion

time: older runners finish the marathon later; severe weather, such as strong winds and higher temperatures, also delays completion. Additionally, we found that men, on average, run faster than women. Ozone is not s statistically significant parameter in the random intercept models. This could result from the observed ozone level falling between 0 and 0.054 ppm, which is thought to be good air quality (Kaiterra). Four marathons had incomplete weather data. If we conduct imputation for missing values, we might get more robust results. It is also worth examining more marathons in different regions to investigate the impact of weather on performance further.

References

[1] Ely, B. R., Cheuvront, S. N., Kenefick, R. W., & Sawka, M. N. (2010). Aerobic performance is degraded, despite modest hyperthermia, in hot environments. Med Sci Sports Exerc, 42(1), 135-41.

[2] Ely, M. R., Cheuvront, S. N., Roberts, W. O., & Montain, S. J. (2007). Impact of weather on marathon-running performance. Medicine and science in sports and exercise, 39(3), 487-493.

[3] Kenney, W. L., & Munce, T. A. (2003). Invited review: aging and human temperature regulation. Journal of applied physiology, 95(6), 2598-2603.

[4] Yanovich, R., Ketko, I., & Charkoudian, N. (2020). Sex differences in human thermoregulation: relevance for 2020 and beyond. Physiology, 35(3), 177-184.

[5] Besson, T., Macchi, R., Rossi, J., Morio, C. Y., Kunimasa, Y., Nicol, C., … & Millet, G. Y. (2022). Sex differences in endurance running. Sports medicine, 52(6), 1235-1257.

Code Appendix

```
set.seed(123456)
library(tidyverse)
library(knitr)
library(tidyr)
library(dplyr)
library(kableExtra)
library(readr)
library(visdat)
library(gtsummary)
library(naniar)
#library(patchwork)
library(gridExtra)
library(corrplot)
#library(olsrr)# model selection for linear model
library(splines)
library(lme4)
library(lmerTest)  #p-values for lmer models
library(ggplot2)
library(gridExtra)

#read in data files
df.aqi<-read.csv("~/Documents/GitHub/PHP2550-PDA-projects/data/aqi_values.csv")
df.course.record<-read.csv("~/Documents/GitHub/PHP2550-PDA-projects/data/course_record.csv
df.marathon.dates<-read.csv("~/Documents/GitHub/PHP2550-PDA-projects/data/marathon_dates.c
df.project1.main<-read.csv("~/Documents/GitHub/PHP2550-PDA-projects/data/project1.csv")
df.aqi<-df.aqi %>%
  mutate(race = case_when(
          marathon == "Boston" ~ "Boston",
          marathon == "Chicago" ~ "Chicago",
          marathon == "Grandmas" ~ "Grandmas",
          marathon == "NYC" ~ "NYC",
          marathon == "Twin Cities" ~ "TC",
          .default = NA)) %>%
  select(-c(marathon))

df.course.record<-df.course.record %>%
  mutate(race = case_when(
          Race == "B" ~ "Boston",
          Race == "C" ~ "Chicago",
```

```
          Race == "D" ~ "Grandmas",
          Race == "NY" ~ "NYC",
          Race == "TC" ~ "TC",
          .default = NA)) %>%
  select(-c(Race))

df.marathon.dates<-df.marathon.dates %>%
  mutate(race = case_when(
          marathon == "Boston" ~ "Boston",
          marathon == "Chicago" ~ "Chicago",
          marathon == "Grandmas" ~ "Grandmas",
          marathon == "NYC" ~ "NYC",
          marathon == "Twin Cities" ~ "TC",
          .default = NA)) %>%
  select(-c(marathon))

df.project1.main<-df.project1.main %>%
  mutate(race = case_when(
          Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D. == 0 ~ "Boston",
          Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D. == 1 ~ "Chicago",
          Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D. == 4 ~
 ↪  "Grandmas",
          Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D. == 2 ~ "NYC",
          Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D. == 3 ~ "TC",
          .default = NA)) %>%
  select(-c(Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.))
# Visualize missing data pattern for all dfs

#aqi column has 16% missingness
df.aqi %>% abbreviate_vars() %>% vis_miss() + theme(text =
 ↪  element_text(size = 7)) + ggtitle("Missing Data")

#no missingness
df.course.record %>% abbreviate_vars() %>% vis_miss() + theme(text =
 ↪  element_text(size = 7)) + ggtitle("Missing Data")

#no missingness
df.marathon.dates %>% abbreviate_vars() %>% vis_miss() + theme(text =
 ↪  element_text(size = 7)) + ggtitle("Missing Data")

#last 8 columns have missingness
```

```r
df.project1.main %>% abbreviate_vars() %>% vis_miss() + theme(text =
 ↪  element_text(size = 7)) + ggtitle("Missing Data")
#start to join datasets
df.all <- left_join(df.project1.main, df.course.record,
 ↪  by=c("Year","race"))%>%
  arrange(Year, race)

#rename sex, age column
df.all<-df.all %>%
  mutate(sex = case_when(
          Sex..0.F..1.M. == 0 ~ "F",
          Sex..0.F..1.M. == 1 ~ "M",
          .default = NA))  %>%
  select(-c(Sex..0.F..1.M.))  %>%
  rename(age = Age..yr.)

#actual time = CR + %CR * CR -> marathon performance
#CR, unique value for each race, each year
#convert CR from character to time

#https://www.runnersworld.com/races-places/a20794726/age-and-weight-groups-com/
#create age divisions
df.all<-df.all %>%
  mutate(CR = as.numeric(hms(CR)),
    actual.time = CR * (1+ X.CR/100),
    age.division = case_when(
          age <= 14 ~ "14 and under",
          age >= 15 & age <= 19 ~ "15-19",
          age >= 20 & age <= 29 ~ "20-29",
          age >= 30 & age <= 39 ~ "30-39",
          age >= 40 & age <= 49 ~ "40-49",
          age >= 50 & age <= 59 ~ "50-59",
          age >= 60 & age <= 69 ~ "60-69",
          age >= 70 & age <= 79 ~ "70-79",
          age >= 80 ~ "80 and over",
          .default = NA),
        age.division =  factor(age.division, levels = c("14 and under",
        "15-19","20-29","30-39","40-49","50-59","60-69","70-79","80 and
         ↪  over")),
        Flag = factor(Flag, levels = c("White", "Green", "Yellow",
         ↪  "Red", "Black")))
```

```r
#aqi dataset, parameter_code column: 44201: Ozone, 88101 (primary
 ↪  measure)/88502: PM2.5, multiple rows correspond to one observation
#aqi: air quality index, calculated from mean (integer)
#arithmetic_mean: actual measurement, average sample across that time
 ↪  period, use this instead of aqi; units_of_measure
#do not need "sample_duration"
#WBGT, Flag = 0.7 * Tw + 0.2 * Tg + 0.1 * Td

#merge the other two dfs
df.all <- left_join(df.all, df.marathon.dates,
 ↪  by=c('Year'='year',"race"))%>%
  arrange(Year, race)

df.aqi$date_local<-as.Date(df.aqi$date_local)
df.all$date<-as.Date(df.all$date)
date.unique.list<-unique(df.all$date)

#pivot wider df.aqi, take mean of arithmetic mean for each region
df.aqi.wider <- df.aqi %>%
  filter(date_local %in% date.unique.list) %>%
  #select(date_local, parameter_code, arithmetic_mean, race,
   ↪  units_of_measure) %>%
  group_by(date_local, race, parameter_code) %>%
  summarise(mean=mean(arithmetic_mean),.groups = 'drop') %>%
  pivot_wider(names_from = parameter_code, values_from = mean)

df.all <- left_join(df.all, df.aqi.wider,
 ↪  by=c('date'='date_local',"race"))%>%
  arrange(Year, race)

df.all <- df.all %>%
  rename(Ozone = `44201`, PM2.5 = `88101`, Sex=sex)
df.all %>%
  dplyr::select(-c(Year, X.CR, date, `88502`, age, X.rh, Ozone,
 ↪  PM2.5))%>%
  mutate(actual.time = actual.time/3600, CR=CR/3600)%>%
  tbl_summary(by = race, statistic = list(
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n} ({p}%)"),
      digits = all_continuous() ~ 2,
```

```
      label = c(actual.time ~ "Completion Time (hr)",
                Wind ~ "Wind speed (km/hr)",
                age.division ~ "Age",
                Td..C ~ "Dry bulb temperature in Celcius",
                Tw..C ~ "Wet bulb temperature in Celcius",
                Tg..C ~ "Black globe temperature in Celcius",
                SR.W.m2 ~ "Solar radiation in Watts per meter squared",
                DP ~ "Dew Point in Celcius",
                WBGT ~ "Wet Bulb Globe Temperature (WBGT)",
                CR ~ "Course Record(hr)"),
      missing="no"
      #missing_text = "(Missing)",
    ) %>%
  add_p()%>%
  bold_labels() %>%
  as_kable_extra(format = "latex", booktabs = TRUE, linesep = "") %>%
  kable_styling(latex_options = "scale_down")
summary_long <- df.all %>%
  group_by(age, Sex) %>%
  summarize(
    Avg = mean(actual.time, na.rm = TRUE)/3600,
    Min = min(actual.time, na.rm = TRUE)/3600,
    Max = max(actual.time, na.rm = TRUE)/3600,
    .groups = "drop") %>%
  pivot_longer(
    cols = c(Avg, Min, Max),
    names_to = "performance_metric",
    values_to = "performance"
  )

ggplot(summary_long, aes(x = age, y = performance, color = Sex, group =
 ↪  interaction(Sex, performance_metric))) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  facet_wrap(~ performance_metric, scales = "free_y", ncol = 1) +
  labs(
    x = "Age Division",
    y = "Course Record (Hours)",
    color = "Sex"
  ) +
  theme_minimal() +
```

```
  theme(
    text = element_text(size = 12),
    legend.position = "top",
    panel.grid.major = element_line(color = "gray90")
  )
#On average, as age increases, completion time increase. There was no
 ↪  significant difference between sexes in the overall trend of
 ↪  completion time.
miss_race<-df.project1.main %>% filter(if_any(everything(), is.na))%>%
  mutate(race = as.factor(race))%>%
  group_by(Year, race)%>%
  summarise(n = n(),.groups = 'drop')%>%
  rename(Race = race, `Number of observations` = n)
kable(miss_race, booktabs = TRUE, digits = 2)
#MCAR
miss_aqi <- miss_var_summary(df.aqi.wider)[c(1:2),]#first 2 variables
 ↪  have missing values

kable(miss_aqi, format = "latex", booktabs = TRUE, digits = 2)%>%
  kable_styling(latex_options = "hold_position")
trend.record.break <- df.all %>%
  filter(X.CR < 0) %>%
  select(Sex, X.CR, age) %>%
  group_by(age, Sex) %>%
  summarise(n = n(),.groups = 'drop') %>%
  arrange(Sex, age) %>%
  mutate(Sex = factor(Sex)) #%>%
  #pivot_wider(names_from = Sex, values_from = n)

# Trend plot - Break record
ggplot(trend.record.break)+
  geom_line(aes(x=age, y = n, color=Sex, group=Sex), linewidth=1.7)+
    scale_x_continuous(breaks = seq(21, 41, by = 2), guide =
     ↪  guide_axis(angle = 45))+
  theme_minimal()+
  ylab("Number of runners")+
  xlab("Age")
df_clean <- df.all %>%
  filter(!is.na(age) & !is.na(actual.time))

df_clean$completion_time_hours <- df_clean$actual.time / 3600
```

```r
# spline models
spline_model_F <- lm(completion_time_hours ~ ns(age, df = 2), data =
 ↪  df_clean[df_clean$Sex == "F", ])
spline_model_M <- lm(completion_time_hours ~ ns(age, df = 2), data =
 ↪  df_clean[df_clean$Sex == "M", ])

# predictions for smooth spline lines
age_range <- seq(min(df_clean$age), max(df_clean$age), length.out = 200)
pred_F <- predict(spline_model_F, newdata = data.frame(age = age_range))
pred_M <- predict(spline_model_M, newdata = data.frame(age = age_range))

spline_data <- data.frame(
  age = rep(age_range, 2),
  completion_time = c(pred_F, pred_M),
  sex = rep(c("Female", "Male"), each = length(age_range))
)

# minimum completion time for each gender group
min_F_index <- which.min(pred_F)
min_F_age <- age_range[min_F_index]
min_F_time <- pred_F[min_F_index]

min_M_index <- which.min(pred_M)
min_M_age <- age_range[min_M_index]
min_M_time <- pred_M[min_M_index]

ggplot(data = df_clean, aes(x = age, y = completion_time_hours, color =
 ↪  Sex)) +
  geom_point(alpha = 0.2, size = 0.7) +
  geom_line(data = spline_data, aes(x = age, y = completion_time, color
   ↪  = sex), size = 1.2) +
  geom_point(aes(x = min_F_age, y = min_F_time), color = "red", size =
   ↪  3) +
  geom_point(aes(x = min_M_age, y = min_M_time), color = "blue", size =
   ↪  3) +
  annotate("text", x = min_F_age - 22, y = min_F_time+1,
           label = paste0("Min (F): ", round(min_F_age, 1), " yrs, ",
            ↪  round(min_F_time, 2), " hrs"),
           color = "red", hjust = 0) +
  annotate("text", x = min_M_age - 1, y = min_M_time-1,
```

```
           label = paste0("Min (M): ", round(min_M_age, 1), " yrs, ",
            ↪  round(min_M_time, 2), " hrs"),
           color = "blue", hjust = 0) +
  labs(
    x = "Age (years)",
    y = "Completion Time (hours)",
    color = "Gender"
  ) +
  scale_color_manual(values = c("Female" = "red", "Male" = "blue")) +
  theme_minimal()
# ages at which to calculate slopes
ages_to_check <- c(25, 50, 75)

# calculate slope at a specific age using finite differences
calculate_slope <- function(model, age) {
  delta <- 0.5
  y_plus <- predict(model, newdata = data.frame(age = age + delta))
  y_minus <- predict(model, newdata = data.frame(age = age - delta))
  slope <- (y_plus - y_minus) / (2 * delta)
  return(slope)
}

# slopes for both gender groups
slopes_F <- sapply(ages_to_check, function(age)
 ↪  calculate_slope(spline_model_F, age))
slopes_M <- sapply(ages_to_check, function(age)
 ↪  calculate_slope(spline_model_M, age))

slope_results <- data.frame(
  Age = ages_to_check,
  Female = slopes_F,
  Male = slopes_M
)

kable(slope_results, format = "latex", booktabs = TRUE, digits = 3)%>%
  kable_styling(latex_options = "hold_position")
ggplot(df.all, aes(x=age.division, y=actual.time, fill=Sex)) +
  geom_violin() +
  scale_y_continuous(breaks = c(3600*c(2,3,4,5,6,7,8,9)),
                     labels =
                      ↪  c("2HR","3HR","4HR","5HR","6HR","7HR","8HR","9HR"))+
```

```
  theme_minimal()+
  labs(x = "Age Group", y = "Completion Time")+
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1))
tbl.time<-t(summary(as.period(df.all$actual.time)))
kable(tbl.time, booktabs = TRUE, digits=2)
trend.age.impact.summ<-df.all %>%
  select(Sex, actual.time, age) %>%
  group_by(age, Sex) %>%
  summarise(n = n(),.groups = 'drop') %>%
  # arrange(Sex, age) %>%
  mutate(Sex = factor(Sex))

trend.age.impact.summ
#age: 14-91

ggplot(trend.age.impact.summ)+
  geom_line(aes(x=age, y = n, color=Sex, group=Sex), linewidth=1.7)+
    scale_x_continuous(breaks = seq(14, 92, by = 2), guide =
     ↪  guide_axis(angle = 45))+
  ylab("Number of participants")+
  xlab("Age")

df.all %>%
  ggplot(aes(x = age, colour = Sex)) +
  facet_wrap(vars(race), ncol = 3) +
  geom_point(stat = "count",size=1) +
  geom_line(stat = "count") +
  labs(x = "Age", y = "Number of participants")
df.all %>% ggplot(aes(x=WBGT, y=actual.time/3600, color=age.division)) +
  geom_smooth()+
  theme_minimal()+
  labs(x = "Wet Bulb Globe Temperature (WBGT)", y = "Completion Time in
   ↪  hours",
       color = "Age Group")+
  theme(
    legend.text = element_text(size = 5),
    legend.title = element_text(size = 6)
    )+
  facet_grid(.~Sex)
df.all %>%
```

```
  mutate(race = factor(race, levels=c("Boston", "Grandmas", "TC",
   ↪  "Chicago", "NYC"))) %>%
ggplot(aes(x=Wind, y=actual.time/3600, color=Sex)) +
  geom_smooth()+
  # scale_y_continuous(breaks = c(3600*c(3,4)),
  #                    labels = c("3HR","4HR"))+
  theme_minimal()+
  labs(x = "Wind speed in km/hr", y = "Completion Time in hours")#+
  #facet_grid(.~race)
p.ozone<-df.all %>% ggplot(aes(x=Ozone, y=actual.time/3600,
 ↪  color=age.division)) +
  geom_smooth()+
  # scale_y_continuous(breaks = c(3600*c(3,4)),
  #                    labels = c("3HR","4HR"))+
  theme_minimal()+
  labs(x = "Ozone (parts per million)", y = "Completion Time in hours")+
  facet_grid(.~Sex)+
  theme(legend.position="none")

p.pm2.5<-df.all %>% ggplot(aes(x=PM2.5, y=actual.time/3600,
 ↪  color=age.division)) +
  geom_smooth()+
  # scale_y_continuous(breaks = c(3600*c(3,4)),
  #                    labels = c("3HR","4HR"))+
  theme_minimal()+
  labs(x = "PM2.5 (micrograms/cubic meter)", y = "Completion Time in
   ↪  hours",
       color = "Age Group")+
  theme(
    legend.text = element_text(size = 5),
    legend.title = element_text(size = 6)
    )+
  facet_grid(.~Sex)

grid.arrange(p.ozone, p.pm2.5)
#largest impact analysis: correlation
df.cor.subset<-df.all %>%
  dplyr::select(actual.time, WBGT, Wind, Ozone, PM2.5)

colnames(df.cor.subset) <- c("Completion Time", "WBGT", "Wind", "Ozone",
 ↪  "PM2.5")
```

```r
corrplot.mixed(cor(df.cor.subset, use = "complete.obs"))
df_hier <- df.all %>%
  filter(complete.cases(WBGT, Wind, PM2.5, Ozone, Sex, age, actual.time,
   ↪  race)) %>%
  mutate(actual.time=actual.time)

df_M <- df_hier %>% filter(Sex == "M") %>%
  select(-c(Sex))

df_F <- df_hier %>% filter(Sex == "F") %>%
  select(-c(Sex))

#random intercept models by sex
model_F <- lmer(actual.time ~ poly(age, 2) + WBGT + Wind + PM2.5 + Ozone
 ↪  +
                           (1|race), data=df_F)

model_M <- lmer(actual.time ~ poly(age, 2) + WBGT + Wind + PM2.5 + Ozone
 ↪  +
                           (1|race), data=df_M)

df_F$preds <- predict(model_F)
df_M$preds <- predict(model_M)
#summary(model_F)
kable(summary(model_F)$coefficients, booktabs = TRUE, digits = 3,
 ↪  caption="Fixed Effects - Female")
kable(ranef(model_F)$race, booktabs = TRUE, digits = 3, caption="Random
 ↪  Effects - Female")#Variance: 269360
#summary(model_M)
kable(summary(model_M)$coefficients, booktabs = TRUE, digits = 3,
 ↪  caption="Fixed Effects - Male")
kable(ranef(model_M)$race, booktabs = TRUE, digits = 3, caption="Random
 ↪  Effects - Male")#Variance: 312832
p.wbgt.f <- ggplot(df_F, aes(x = WBGT, y = preds, color = race)) +
    geom_smooth(method = "lm", fullrange = TRUE, size = 0.3) +
    geom_jitter(aes(x = WBGT, y = actual.time, group = race, color =
     ↪  race),
                alpha = 0.1) +
    labs(x = "WBGT", y = "Completion Time (sec)") +
    ggtitle("Female") +
```

```
    scale_colour_discrete('race', name="Race")


p.wbgt.m <- ggplot(df_M, aes(x = WBGT, y = preds, color = race)) +
    geom_smooth(method = "lm", fullrange = TRUE, size = 0.3) +
    geom_jitter(aes(x = WBGT, y = actual.time, group = race, color =
     ↪  race),
                alpha = 0.1) +
    labs(x = "WBGT", y = "Completion Time (sec)") +
    ggtitle("Male") +
    scale_colour_discrete('race', name="Race")


p.wind.f <- ggplot(df_F, aes(x = Wind, y = preds, color = race)) +
    geom_smooth(method = "lm", fullrange = TRUE, size = 0.3) +
    geom_jitter(aes(x = Wind , y = actual.time, group = race, color =
     ↪  race),
                alpha = 0.1) +
    labs(x = "Wind ", y = "Completion Time (sec)") +
    ggtitle("Female") +
    scale_colour_discrete('race', name="Race")


p.wind.m <- ggplot(df_M, aes(x = Wind, y = preds, color = race)) +
    geom_smooth(method = "lm", fullrange = TRUE, size = 0.3) +
    geom_jitter(aes(x = Wind, y = actual.time, group = race, color =
     ↪  race),
                alpha = 0.1) +
    labs(x = "Wind", y = "Completion Time (sec)") +
    ggtitle("Male") +
    scale_colour_discrete('race', name="Race")


p.pm2.5.f <- ggplot(df_F, aes(x = PM2.5, y = preds, color = race)) +
    geom_smooth(method = "lm", fullrange = TRUE, size = 0.3) +
    geom_jitter(aes(x = PM2.5, y = actual.time, group = race, color =
     ↪  race),
                alpha = 0.1) +
    labs(x = "PM2.5", y = "Completion Time (sec)") +
    ggtitle("Female") +
    scale_colour_discrete('race', name="Race")
```

```
p.pm2.5.m <- ggplot(df_M, aes(x = PM2.5, y = preds, color = race)) +
    geom_smooth(method = "lm", fullrange = TRUE, size = 0.3) +
    geom_jitter(aes(x = PM2.5, y = actual.time, group = race, color =
     ↪  race),
                alpha = 0.1) +
    labs(x = "PM2.5", y = "Completion Time (sec)") +
    ggtitle("Male") +
    scale_colour_discrete('race', name="Race")


p.ozone.f <- ggplot(df_F, aes(x = Ozone , y = preds, color = race)) +
    geom_smooth(method = "lm", fullrange = TRUE, size = 0.3) +
    geom_jitter(aes(x = Ozone, y = actual.time, group = race, color =
     ↪  race),
                alpha = 0.1) +
    labs(x = "Ozone", y = "Completion Time (sec)") +
    ggtitle("Female") +
    scale_colour_discrete('race', name="Race")  +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))


p.ozone.m <- ggplot(df_M, aes(x = Ozone , y = preds, color = race)) +
    geom_smooth(method = "lm", fullrange = TRUE, size = 0.3) +
    geom_jitter(aes(x = Ozone , y = actual.time, group = race, color =
     ↪  race),
                alpha = 0.1) +
    labs(x = "Ozone", y = "Completion Time (sec)") +
    ggtitle("Male") +
    scale_colour_discrete('race', name="Race")  +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

grid.arrange(p.wbgt.f, p.wbgt.m,
             p.wind.f, p.wind.m,
             p.pm2.5.f, p.pm2.5.m,
             p.ozone.f, p.ozone.m, nrow = 4)
```