STAT 318 Final Project: Boston Marathon Analysis

Marisa Papagelis and Peyton Wang
Fall 2021

https://www.telegraph.co.uk/travel/destinations/north-america/united-states/boston/articles/boston-travel-guide/

P

TABLE OF CONTENTS

P

Introduction

Motivation, data set, visualizations, &
research question

Why are we interested in this topic?

★ Joint interest in sports

★ Proximity of the Boston Marathon to Wellesley College

★ Ethics of collecting / utilizing data from a sporting event

https://www.nytimes.com/2020/05/28/sports/boston-marathon-canceled.html

Our Data Set

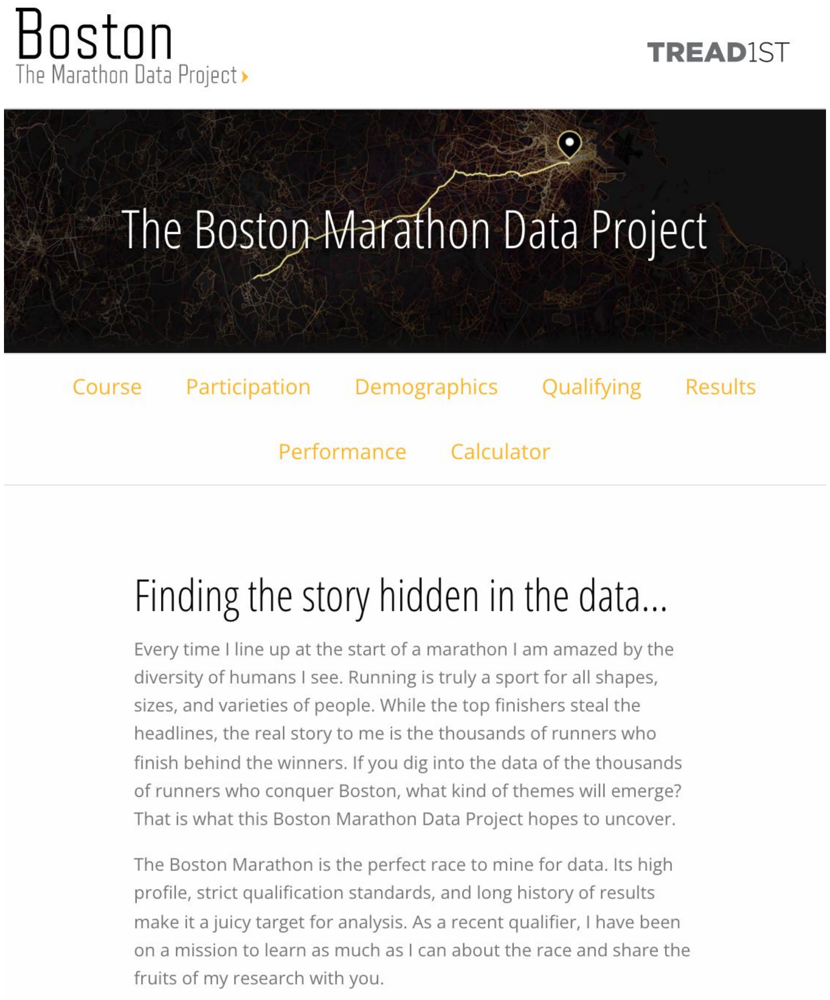
★ 2017 Boston Marathon Data collected by Adrian Hanft as part of [The Boston Marathon Data Project](https://tread1st.com/boston.html#/)

★ 26,410 observations of Boston Marathon participants

Main variables of interest:

<u>Predictors:</u> age, gender, country_residence, split times in increments of 5k (i.e. 5k, 10k, etc.) & pace

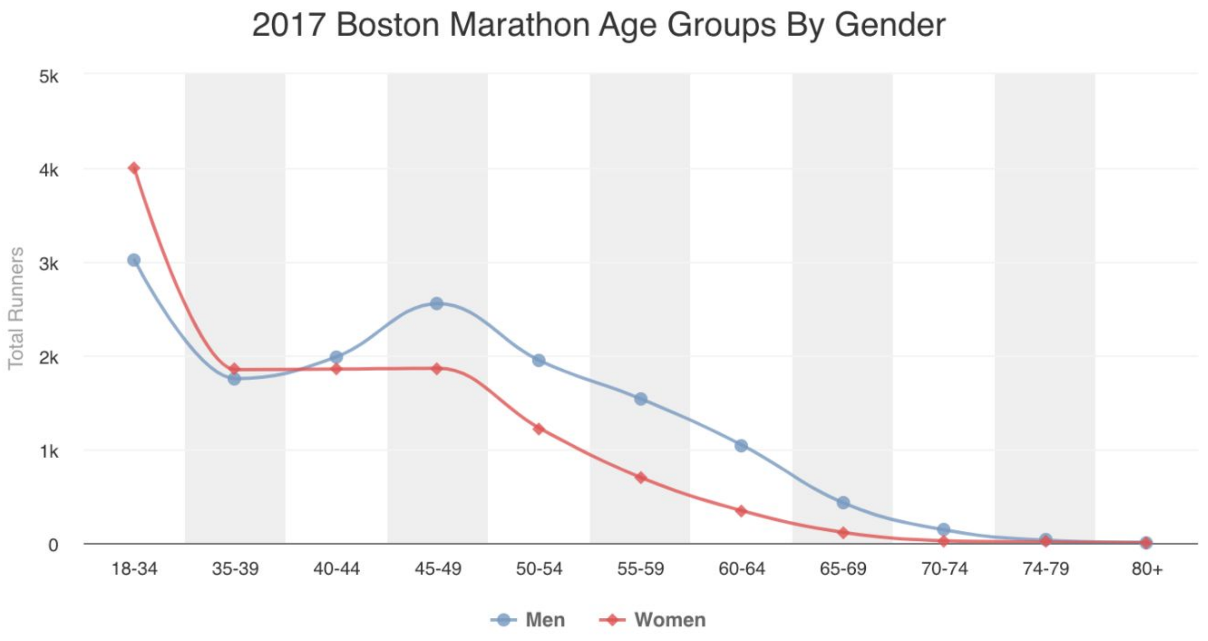
<u>Response:</u> final_time

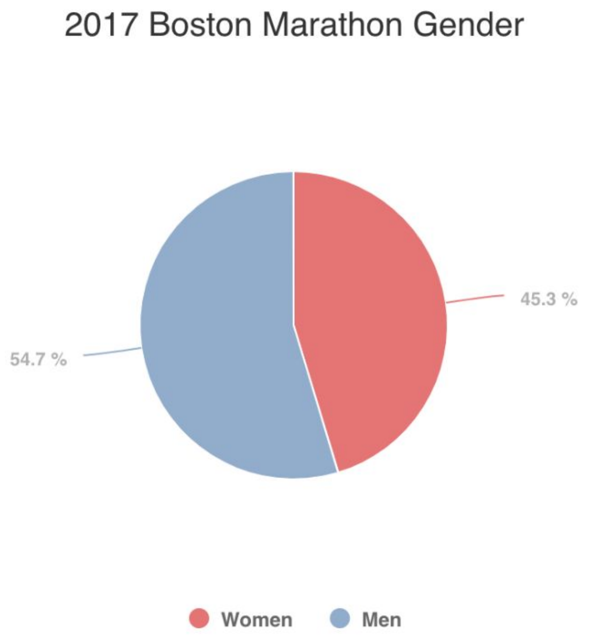


https://tread1st.com/boston.html#/

M

Data Visualizations

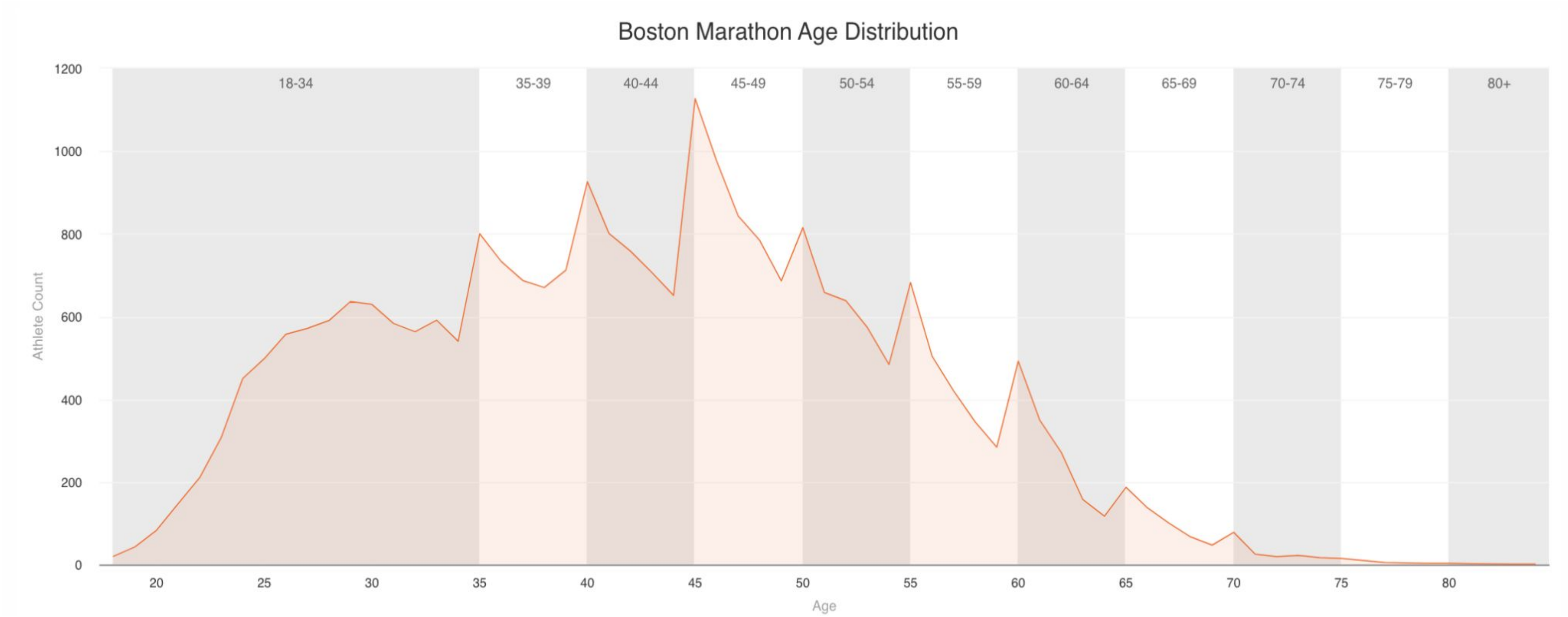




M

https://tread1st.com/boston.html#/demographics

Data Visualizations (cont.)



M

https://tread1st.com/boston.html#/demographics

Research Question

Which factors contribute to best predicting the finish time of marathon participants?

M

02

Data Cleaning

Handling missingness, identifying multicollinearity, & creating indicator variables

P

Initial Cleaning

Removing Predictors

Initially, we removed variables that didn't provide any useful information (i.e. name and bib number).

We also omitted variables that gave us the same information as the response variable (i.e. overall_place, official_time in HH:MM:SS format).

Missingness & Imputation

Variables with missingness greater than 15% of the sample size (i.e. country_residence and projected_time) were removed as well.

We performed median imputation on split times (i.e. 5k, 10k, etc.), which were the only remaining variables with missing values. *Note: in our final paper, we plan to use regression imputation to decrease bias.*

P

Additional Cleaning

Identifying Multicollinearity

Initially, we used the correlation matrix of the remaining quantitative variables in the model to identify any abnormally high correlation coefficients.

Using VIF scores and a threshold of 7, we determined which predictors to keep and remove from our data set.

Indicator Variables

When separating the categorical variables (gender and country_residence) by Female/Male and USA/international, we obtained similar slopes, but different intercepts.

Consequently, we created the following binary categorical variables for both pairs: Female = 0 and Male = 1, USA = 0 and not USA = 1.

Remaining Variables

Response Variable:

★ final_time: the runner's official marathon time (seconds)

Predictor Variables:

★ age: the runner's age (years)
★ gender: the runner's gender (0 or 1)
    ○ female or male encoded as binary indicator variable
★ country_residence: which country the runner represents (0 or 1)
    ○ USA or not USA encoded as binary indicator variable
★ X5k: the runner's time at 5k (seconds)
★ half: the runner's time at halfway point (seconds)

P

Model Selection & Validation

All-subset comparison, automatic selection, & cross validation

All-Subset Selection

Criterion: Mallow's C_p and adjusted R^2

- ★ Model using Mallow's Cp:

$$\widehat{final\_time} = age + gender + country\_residence + X5k + half$$

- ★ Model using adjusted R^2 criterion:

$$\widehat{final\_time} = age + gender + country\_residence + X5k + half$$

Note: we obtained the exact same models for all-subset selection with both Mallow's C_p and adjusted R^2 criterion.

Automatic Selection: Stepwise Regression

Model using AIC criterion:

$$\widehat{final\_time} = age + gender + country\_residence + X5k + half$$

Model using BIC criterion:

$$\widehat{final\_time} = gender + country\_residence + X5k + half$$

M

Model Validation: t-test

```
> summary(fit.Cp.R2.AIC)

Call:
lm(formula = final_time ~ age + gender + country_residence +
    X5k + half)

Residuals:
     Min       1Q   Median       3Q      Max
-11383.0   -479.5   -144.7    316.3  11162.6

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -185.68509   36.03957  -5.152 2.59e-07 ***
age                 -0.90314    0.45800  -1.972   0.0486 *
gender             337.77436   10.71984  31.509  < 2e-16 ***
country_residence  -61.46694   12.11669  -5.073 3.94e-07 ***
X5k                 -2.97014    0.06829 -43.495  < 2e-16 ***
half                 2.84919    0.01498 190.230  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 781.7 on 26404 degrees of freedom
Multiple R-squared:  0.9045,  Adjusted R-squared:  0.9044
F-statistic: 4.999e+04 on 5 and 26404 DF,  p-value: < 2.2e-16
```

```
> summary(fit.BIC)

Call:
lm(formula = final_time ~ gender + country_residence + X5k +
    half)

Residuals:
     Min       1Q   Median       3Q      Max
-11347.1   -480.6   -145.6    316.5  11172.3

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -200.05303   35.29729  -5.668 1.46e-08 ***
gender             331.63277   10.25799  32.329  < 2e-16 ***
country_residence  -64.84583   11.99558  -5.406 6.51e-08 ***
X5k                 -2.96808    0.06828 -43.468  < 2e-16 ***
half                 2.84569    0.01487 191.334  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 781.8 on 26405 degrees of freedom
Multiple R-squared:  0.9044,  Adjusted R-squared:  0.9044
F-statistic: 6.248e+04 on 4 and 26405 DF,  p-value: < 2.2e-16
```

M

Model Validation: ANOVA

```
> anova(fit.Cp.R2.AIC)
Analysis of Variance Table

Response: final_time
                     Df     Sum Sq    Mean Sq  F value    Pr(>F)    
age                   1 8.5682e+09 8.5682e+09  14020.4 < 2.2e-16 ***
gender                1 1.4420e+10 1.4420e+10  23595.6 < 2.2e-16 ***
country_residence     1 1.1227e+09 1.1227e+09   1837.1 < 2.2e-16 ***
X5k                   1 1.0653e+11 1.0653e+11 174320.5 < 2.2e-16 ***
half                  1 2.2115e+10 2.2115e+10  36187.3 < 2.2e-16 ***
Residuals         26404 1.6136e+10 6.1112e+05                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> anova(fit.BIC)
Analysis of Variance Table

Response: final_time
                     Df     Sum Sq    Mean Sq   F value    Pr(>F)    
gender                1 9.5950e+09 9.5950e+09  15698.89 < 2.2e-16 ***
country_residence     1 3.6906e+08 3.6906e+08    603.84 < 2.2e-16 ***
X5k                   1 1.2042e+11 1.2042e+11 197018.44 < 2.2e-16 ***
half                  1 2.2375e+10 2.2375e+10  36608.58 < 2.2e-16 ***
Residuals         26405 1.6138e+10 6.1119e+05                        
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P

Overview

★ 1st Model: Mallow's C_p, adjusted R^2, AIC

$$\widehat{final\_time} = age + gender + country\_residence + X5k + half$$

★ 2nd Model: BIC

$$\widehat{final\_time} = gender + country\_residence + X5k + half$$

Measure	1st Model	2nd Model
R^2	0.9045	0.9044
Adjusted R^2	0.9044	0.9044
5-Fold CV Score	30.36673	30.44347

** 5-Fold CV Score is higher than we expected, likely due to discrepancies in our imputation methods and outliers. We plan to look into this further in the final paper.*

04

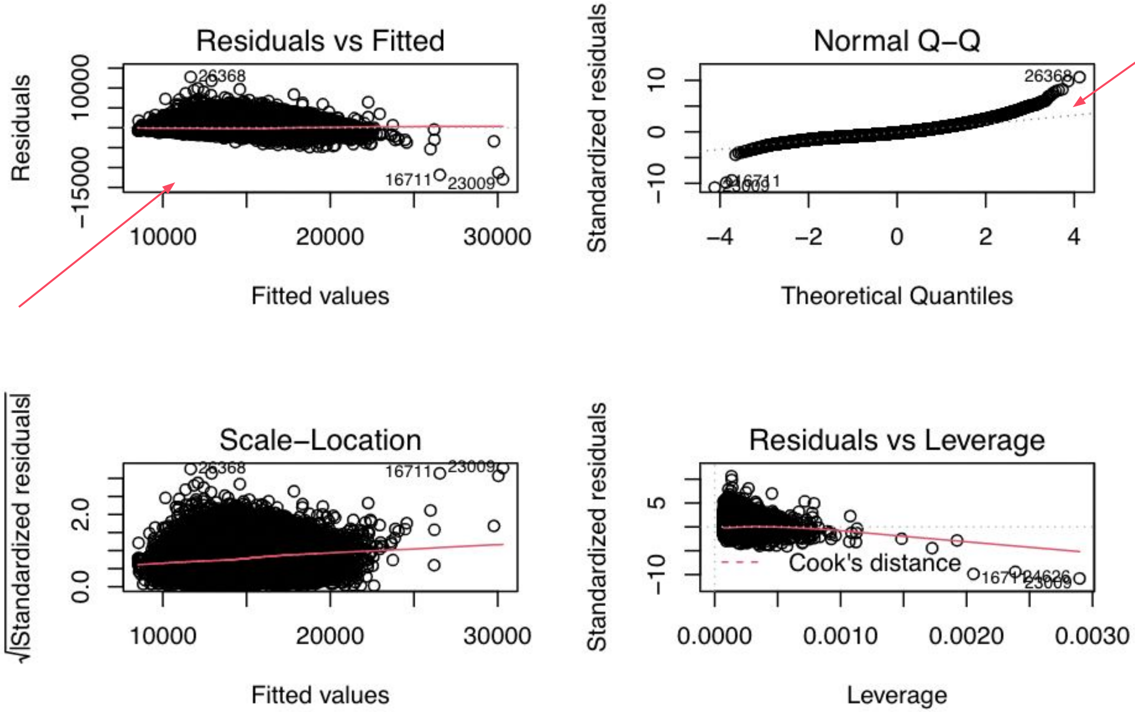
Model Diagnostics

Model significance, influential outliers,
data transformation, & assumptions

P

Initial Model Diagnostics

possible violation of
normality assumption

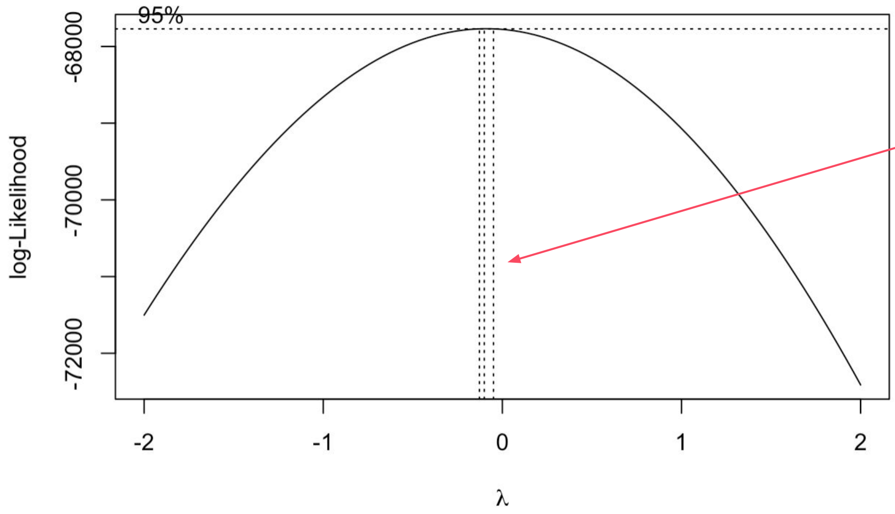


possible violation of
constant-variance
assumption

4 influential or
outlying
observations
(26368, 16711,
23009, 24626)

Model Diagnostics

★ transform Y to help correct the non-constant error variance and departure from normality

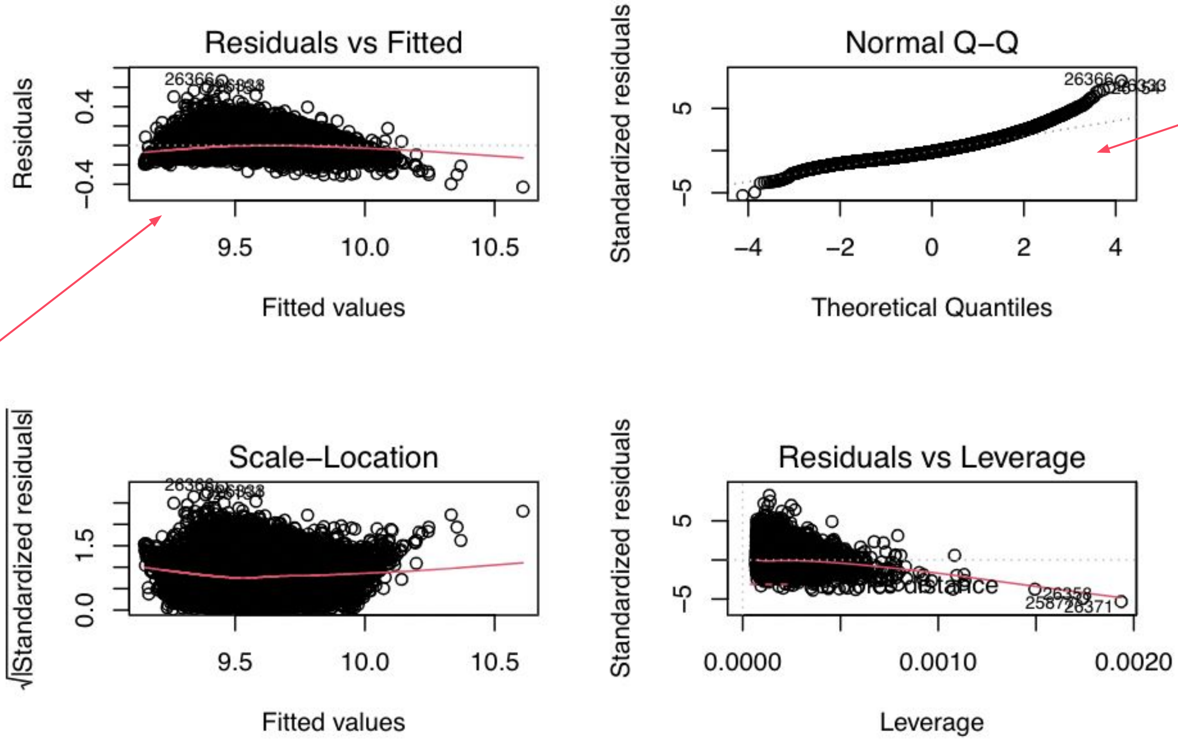


λ = 0 → natural log transformation on response variable

★ transformed model: $log(\widehat{final\_time}) = age + gender + country\_residence + X5k + half$

P

Final Model Diagnostics



better normality

better constant variance

5 influential or outlying observations (26355, 26212, 24200, 25949, 23127)

Which factors contribute to best predicting the finish time of marathon participants?

- ★ age, gender, country residence, X5k, half
  - ○ high significance in summary (t-test) and ANOVA output (F-test)
- ★ R^2: 0.8903; adjusted R^2: 0.8902
  - ○ model explains a good amount (89%) of variation in the data

```
> summary(log.Cp.R2.AIC)

Call:
lm(formula = log(final_time) ~ age + gender + country_residence +
    X5k + half, data = removed_df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46040 -0.03518 -0.00637  0.02804  0.44828

Coefficients:
                   Estimate Std. Error  t value Pr(>|t|)
(Intercept)       8.565e+00  2.646e-03 3236.564  < 2e-16 ***
age               4.234e-04  3.347e-05   12.651  < 2e-16 ***
gender            1.140e-02  7.834e-04   14.553  < 2e-16 ***
country_residence -4.506e-03  8.840e-04   -5.097 3.48e-07 ***
X5k              -1.538e-04  5.211e-06  -29.517  < 2e-16 ***
half              1.809e-04  1.140e-06  158.600  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05701 on 26374 degrees of freedom
Multiple R-squared:  0.8903,	Adjusted R-squared:  0.8902
F-statistic: 4.279e+04 on 5 and 26374 DF,  p-value: < 2.2e-16
```

```
> anova(log.Cp.R2.AIC)
Analysis of Variance Table

Response: log(final_time)
                     Df Sum Sq Mean Sq  F value    Pr(>F)
age                   1  47.36   47.36  14572.3 < 2.2e-16 ***
gender                1  80.20   80.20  24677.3 < 2.2e-16 ***
country_residence     1   5.16    5.16   1587.6 < 2.2e-16 ***
X5k                   1 480.91  480.91 147979.4 < 2.2e-16 ***
half                  1  81.75   81.75  25153.8 < 2.2e-16 ***
Residuals         26374  85.71    0.00
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

M

05

Conclusion & Further Considerations

Model concerns, data ethics, & future improvements

M

Model Considerations and Concerns

- ★ We are limited by our data (i.e. ambiguous data collection process, no wheelchair data).
- ★ We are not sure if our model is externally valid, so we cannot apply it to outside datasets.

Data Ethics

- ★ Our dataset is pulled from a secondhand source and has no association with the Boston Athletic Association.
- ★ The collector/distributor of the data (and us) have no way of knowing if the data set is discrepancy free.

Further Considerations & Future Improvements

- ★ Use regression imputation instead of median imputation.
- ★ Assess patterns among missing values.
- ★ Similar to how the marathon handles age division, create age bins and use this as a factor.
- ★ Incorporate country residence into the model by creating continent bins.
- ★ Remove more influential outlying observations to improve model fit.
- ★ Find marathon data sets with additional factors (i.e. wheelchairs, seed time, height/weight).
- ★ Observe different Boston marathon data sets with respect to time.

P

Thank you!
Any questions?