Predicting Hyponatremia Among Marathon Runners: Assessing Race Practice and Physicality

Let Us Pass This Project So We Can Major In Stats

Clara Page, Caroline Useda, Steven Yu

Abstract

For this paper, we studied the incidence of hyponatremia in marathon runners. We used a sample of Boston Marathon runners to analyze how variables related to a person's physiology, including gender and age, impact serum sodium concentration. We also looked at a number of variables related to racing, like water intake. After our premlininary anaylsis was complete, we used backward selection to construct a multiple logistic regression model. This model left us with two statistically signifcant predictors of hyponatremia affliction: gender and weight difference before and after racing.

Contents

Background and Significance

Hyponatremia occurs when a person's blood sodium levels are <135 mmol/L. Sodium is an electrolyte; the body uses it to regulate the water in and around cells. The causes of hyponatremia can rise from a multitude of factors—from underlying medical conditions, water consumption, vomiting, diarrhea, diuretics, and fever, among others. In this project, we look into exercise, specifically, marathon-related hyponatremia. Severe and potentially life-threatening hyponatremia can occur during exercise, particularly in athletes who participate in endurance events such as marathons, triathlons, and ultra-distance races.In our study, we examine the data of a cohort of Boston Marathon participants from 2002. We are using a data set collected by researchers studying hyponatremia in runners post-marathon. It contains 488 observations on 13 variables. The observational units are the different runners who participated in the 2002 Boston Marathon. There are 488 runners, of which 322 are male, and 166 are female.

Hyponatremia has the potential to lead to severe medical conditions that can result in fatality. Our studies examine the most significant factors that contribute to if a runner gets hyponatremia post-race—utilizing predictor variables such as sex, run time, weight, train pace, BMI, water load, and more. With these variables in mind, we chose to assess the effect these variables have on the serum sodium concentration (mmol per liter) immediately after marathon completion. To do so, we completed a multiple logistic regression model for our data using automated selection techniques. We hypothesize that fluid intake/outtake may be one of the most significant predictors of hyponatremia. It is also possible that other factors such as weight and body mass index may also be of significance in figuring out if a runner has hyponatremia post-marathon. Overall, we are trying to predict if a runner will have hyponatremia post-race and what factors are of significance in predicting if they will or will not have hyponatremia to decrease the risk of fatality.

Methods

Data collection

Participants over the age of 18 were selected at random while registering for the Boston Marathon and asked to participate, and then provided written consent. Before beginning the marathon, participants were asked to fill out a survey with basic information about their training, medical history, possible hydration strategies, and their demographics. After finishing the race, they were asked for both a blood sample and information about their fluid consumption and urine output during the race. The study was approved by the Committee on Clinical Investigation at Children's Hospital in Boston.

Variable creation

Variables were created that were believed to have an impact on whether or not a runner gets Hyponatremia. The predictor variables are female, and indicator variable where 1 is for female, 0 for male, howmany, the number of prior marathons completed, age, age in years, lwobup01, use of NSAIDs (non-steroidal antiinflammatory drugs) within last week of marathon, wateld01 the water loaded 24 hours prior to marathon, urinat3p, an indictator variable where 1 means they urinated 3+ times during the race and 0 means they urinated less than 3 times during the race, fluidfr3 the fluid frequency during the race, an indicator variable where 1 is equal to every mile, 2 every other mileand 3=every third mile or less, wtdiff the difference in weight (kg) post-race vs. pre-race, runtime the marathon race time in minutes, trainpse the training pace in seconds, and bmi the body mass index measured in kg/m2. The response variable is sodium: serum sodium concentration, measured in mmol per liter, immediately after marathon completion. To use sodium as the response in our multiple logistic regression, it was mutated as to have a binary response, where sodium above 135 indicated Hyponatremia, and below does not.

Analytic methods

Our analysis can be separated into three unique steps. First, we looked at all variables in the data set. We created statistical summaries of each, which included the mean, median, max, and min of the variable in question and looked at summary statistics stratified by gender. Finally, we checked VIFs to relieve any

concerns we might have about the incidence of multicollinearity. Our goal in this stage was to understand the data that we were working with.

Next, we verified that the conditions for inference were met within our dataset. As there are three central conditions, this was a three part process. It involved producing empirical logit plots to check linearity and reading background from the researchers on independence and randomness in their data selection.

For the actual analysis, we relied on backward selection to produce a multiple logistic regression model. We determined the cutoff for hyponatremia to be a level of sodium less than 135. Transforming the data in this way allowed us to study probabilites and odds ratios in terms of our coefficents and put all results in terms of probability of hyponatremia. Otherwise, we would've been studying variable's effects on sodium concentration. We were interested in the illness.

Backward selection with required transformations included by variable led us to a model with two predictors. We used that model to calculate odds ratios from our predictor's slopes.

We also looked at each predictor individually, in simple logistic regression.

Results

Our sample consisted of approximately 166 women and 322 men, all 18 years or older. Their experience levels and levels of race preparation varied; we were provided with data on these factors by particpant. Our first analytic step was to take the data make sure no variable had a VIF above 5. None did, so we didn't concern ourselves with the possibility of multiplecollinearity.

Next, we produced empirical logit plots for every non-categorical variable. These plots indicated issues with linearity for runtime and BMI. Both variables needed a quadratic term if they were to meet the condition – this term was included in the subsequent backward selection. Once linearity was checked, we moved onto independence. This was easy to verify. None of the runners were selected twice. Subject's were approached randomly by the people who collected for the data so, while their might be some non-response bias, there is reason enough to believe randomness is satisfied.

Finally, we used backward selection to make our model. In the end, the model we selected was logit(pi) = -2.910 + 1.006(Female) + 0.856(wtdiff). This gave us two central predictors of hyponatremia. The predictors and their p values (all less than aplha = .05) are shown below. The low p values mean that we consider this slope to be different from 0 and reject the null hypothesis that it is 0.

```
results <- matrix(c(1.006, 0.0066, 0.856, 4.3e-10),ncol=2,byrow=TRUE)
colnames(results) <- c("Coefficent","P Value")
rownames(results) <- c("Female","Wtdiff")
results <- as.table(results)
results
```

```
       Coefficent   P Value
Female  1.006e+00 6.600e-03
Wtdiff  8.560e-01 4.300e-10
```

Based on these slopes, we can conclude that after adjusting for weight difference before and after the race, being female is associated with an odds ratio of 2.73464. This means being a woman is associated with being 173.464% more likely to contract this after or during a race.

Next, we can conclude that after adjusting for gender, weight difference is associated with an odds ratio of 2.35373. This means that a one unit change in weight difference is associated with being 135.373% more likely to contract this after or during a race.

Alone, some variables that are not statistically significant after adjusting for age and weight difference become significant. Note that quadratic terms were included for the quadratic associated variables (runtime, BMI).

```
resultsone <- matrix(c(1.292, 6.7e-05, 3.64006, 0.869, 5e-11, 2.38453, 0.01115, 0.00098, 1.01121, 0.008
colnames(resultsone) <- c("Coefficent","P Value","Odds Ratio")
rownames(resultsone) <- c("Female","Wtdiff", "Runtime", "Trainpse")
resultsone <- as.table(resultsone)
resultsone
```

```
          Coefficent     P Value  Odds Ratio
Female   1.29200e+00 6.70000e-05 3.64006e+00
Wtdiff   8.69000e-01 5.00000e-11 2.38453e+00
Runtime  1.11500e-02 9.80000e-04 1.01121e+00
Trainpse 8.03000e-03 4.50000e-04 1.00806e+00
```

Conclusion

Our objective was to study the incidence of hyponatremia among marathon runners. In the end, we developed a multiple logistic model to predict the likelihood of incidence of the disease. The model included two predictors, which is less than we might have hoped would be significant in a multiple predictor model. Nonetheless, gender and weight different did serve to show real risk factors for hyponatremia. Additionally, by using simple logistic regression, we were able to identifiy more potential predictors of illness.

Hyponatremia remains an incredbly dangerous ailment, one that is not fully understood by researches. Our reports has shown reason to believe gender and weight loss during races, as well as training pace and runtime influence the probability of being struck by the disease. Our findings do not consider preexisting conditions among participants. However, comorbid medical conditions could cause a person to be at higher risk for an ailment like hyponatremia. Additionally, treatments were not assigned in the data we used, so we can only conclude association. Assigning treatments and looking at medical history would be interesting areas for research to expand into.

In the end, we provided research that should help runners prepare for marathons and prevent illness. That does not mean, however, that there isn't much more work to be done before we fully understand hyponatremia, its causes, and how to prevent it.

Appendix

EMPERICAL LOGIT

```
Marathon<- read_csv("http://kcorreia.people.amherst.edu/S1920/boston_marathon.csv")
```

```
New names:
* `` -> ...1
```

```
Rows: 488 Columns: 13
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (13): ...1, sodium, female, howmany, age, lwobup01, wateld01, urinat3p, ...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Marathon2<-Marathon%>%mutate(sodium_low=ifelse(sodium<135, yes=1,no=0))
```

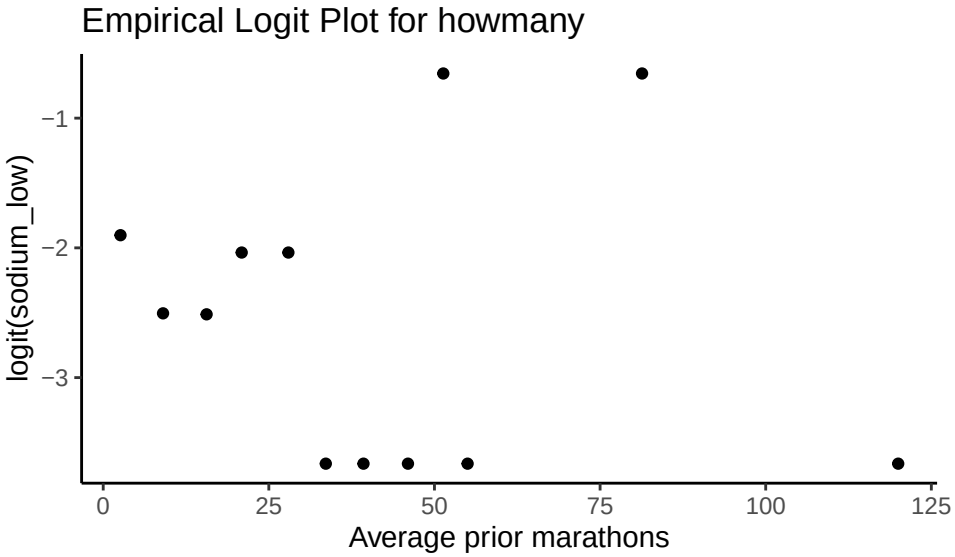
```
run2<-Marathon2%>%
  mutate(howmanygrp = cut(howmany,breaks=20))
binned.y2<-mean(sodium_low~howmanygrp, data=run2, na.rm=TRUE)

binned.x2<-mean(howmany~howmanygrp, data=run2, na.rm=TRUE)
```

```
gf_point(logit(binned.y2)~binned.x2, xlab="Average prior marathons"
         ,ylab="logit(sodium_low)",
         title="Empirical Logit Plot for howmany")
```

```
Warning in logit(binned.y2): proportions remapped to (0.025, 0.975)
```

```
Warning in logit(binned.y2): proportions remapped to (0.025, 0.975)
```

```
Warning: Removed 8 rows containing missing values (geom_point).
```



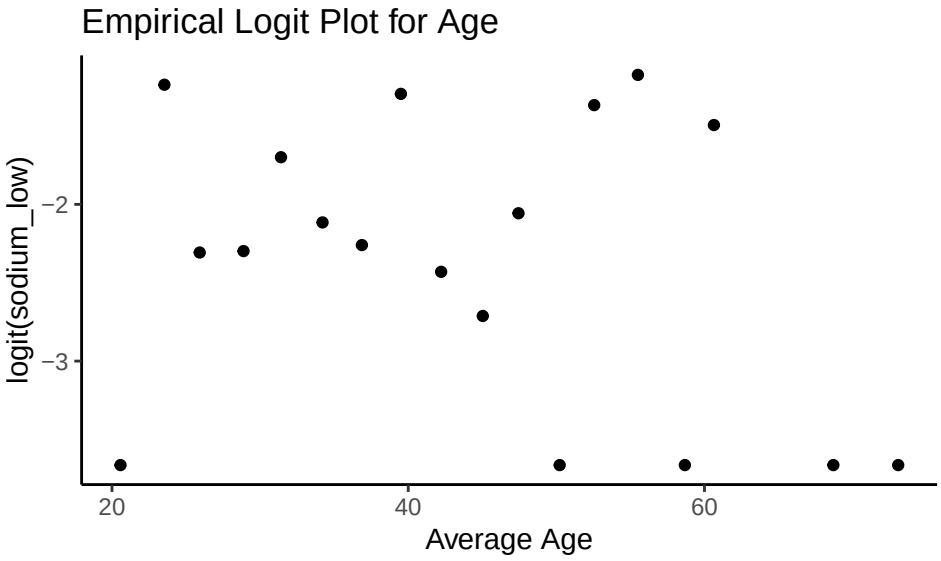
Age is also not looking like an ideal predictor. However, there is some correlation.

```
run1<-Marathon2%>%
  mutate(agegrp = cut(age,breaks=20))
binned.y <- mean (sodium_low ~ agegrp, data=run1, na.rm=TRUE)
binned.x <- mean (age ~ agegrp, data = run1, na.rm=TRUE)
gf_point(logit(binned.y) ~ binned.x, xlab="Average Age"
         ,ylab="logit(sodium_low)",
         title="Empirical Logit Plot for Age")
```

```
Warning in logit(binned.y): proportions remapped to (0.025, 0.975)
```

```
Warning in logit(binned.y): proportions remapped to (0.025, 0.975)
```

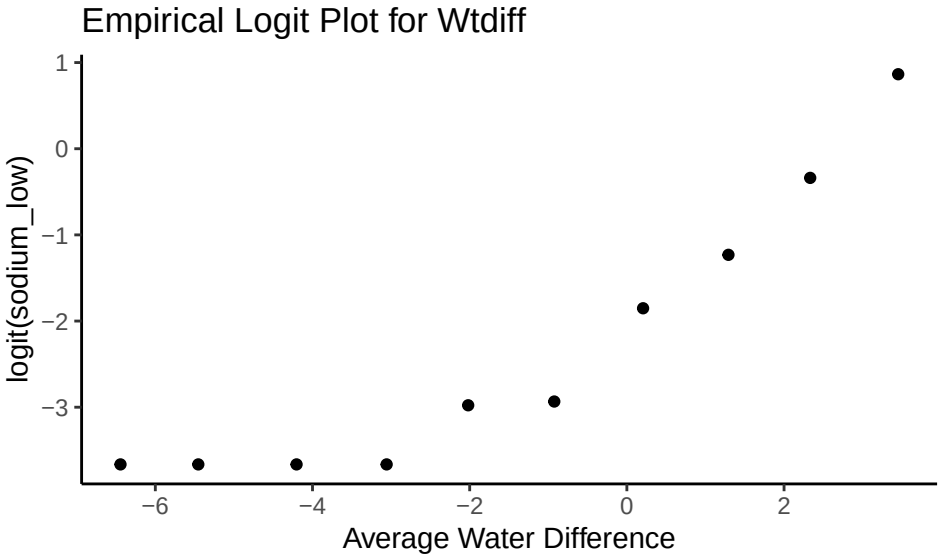
```
Warning: Removed 2 rows containing missing values (geom_point).
```



This looks like it needs a log transformation.

```
run3<-Marathon2%>%
  mutate(wtdiffgrp = cut(wtdiff,breaks=10))
binned.y3<-mean(sodium_low~wtdiffgrp, data=run3, na.rm=TRUE)

binned.x3<-mean(wtdiff~wtdiffgrp, data=run3, na.rm=TRUE)
gf_point(logit(binned.y3)~binned.x3, xlab="Average Water Difference"
         ,ylab="logit(sodium_low)",
         title="Empirical Logit Plot for Wtdiff")
```

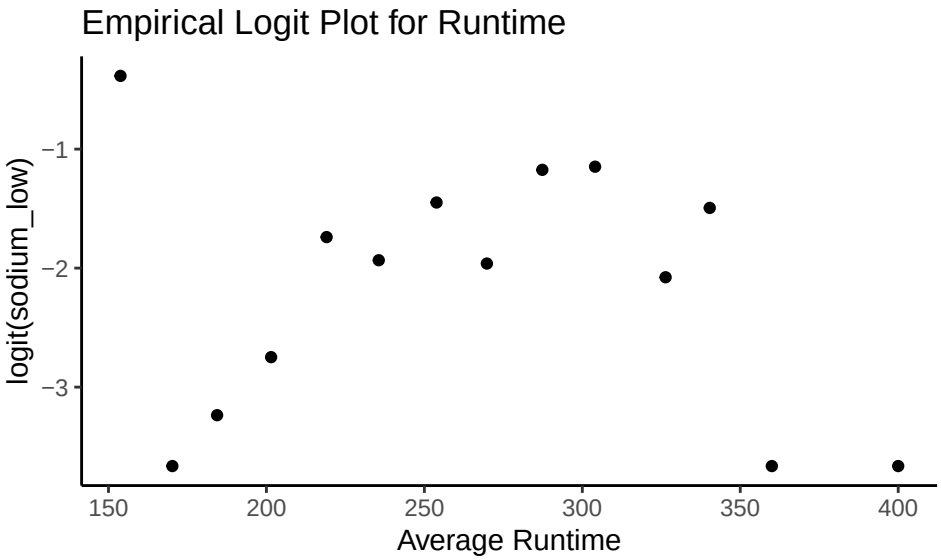
```
Warning in logit(binned.y3): proportions remapped to (0.025, 0.975)

Warning in logit(binned.y3): proportions remapped to (0.025, 0.975)
```



I might like to see a quadratic predictor for runtime. It appears very fast and very slow runners aren't at high risk – while average ones are.

```
run4<-Marathon2%>%
  mutate(runtimegrp = cut(runtime,breaks=15))
binned.y4<-mean(sodium_low~runtimegrp, data=run4, na.rm=TRUE)

binned.x4<-mean(runtime~runtimegrp, data=run4, na.rm=TRUE)
gf_point(logit(binned.y4)~binned.x4, xlab="Average Runtime"
         ,ylab="logit(sodium_low)",
         title="Empirical Logit Plot for Runtime")
```

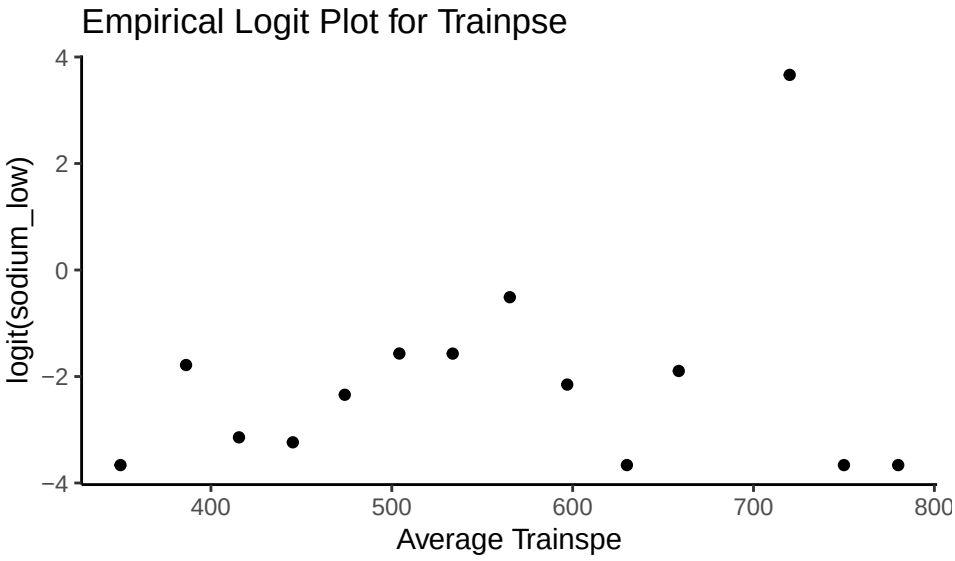
```
Warning in logit(binned.y4): proportions remapped to (0.025, 0.975)

Warning in logit(binned.y4): proportions remapped to (0.025, 0.975)

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



This looks like it has a reasonable, positive association.

```
run5<-Marathon2%>%
  mutate(trainpsegrp = cut(trainpse,breaks=15))
binned.y5<-mean(sodium_low~trainpsegrp, data=run5, na.rm=TRUE)

binned.x5<-mean(trainpse~trainpsegrp, data=run5, na.rm=TRUE)
gf_point(logit(binned.y5)~binned.x5, xlab="Average Trainspe"
         ,ylab="logit(sodium_low)",
         title="Empirical Logit Plot for Trainpse")
```

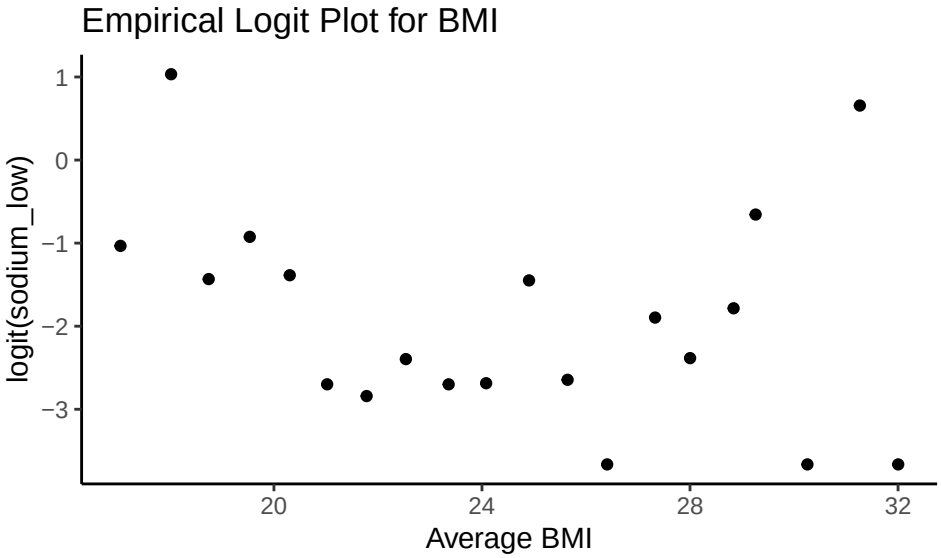
```
Warning in logit(binned.y5): proportions remapped to (0.025, 0.975)

Warning in logit(binned.y5): proportions remapped to (0.025, 0.975)

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



This is quadratic.

```
run6<-Marathon2%>%
  mutate(bmigrp = cut(bmi,breaks=20))
binned.y6<-mean(sodium_low~bmigrp, data=run6, na.rm=TRUE)

binned.x6<-mean(bmi~bmigrp, data=run6, na.rm=TRUE)
gf_point(logit(binned.y6)~binned.x6, xlab="Average BMI"
         ,ylab="logit(sodium_low)",
         title="Empirical Logit Plot for BMI")
```

```
Warning in logit(binned.y6): proportions remapped to (0.025, 0.975)

Warning in logit(binned.y6): proportions remapped to (0.025, 0.975)
```



BACKWARD SELECTION PROCRESS

```
#Kitchen Sink Model
MarathonCorrect <- dplyr::select(Marathon2, female, howmany, age, lwobup01, wateld01, urinat3p, fluidfr
```

```
MarathonFull <- glm(sodium_low ~  age + trainpse + as.factor(female) + as.factor(lwobup01) + as.factor(
#msummary(MarathonFull)

#The first model isn't satisfactory. I'm going to remove urinat3p, which has the highest p value.

MarathonFull2 <- glm(sodium_low ~  age + trainpse + as.factor(female) + as.factor(lwobup01) + as.factor
#msummary(MarathonFull2)

#Most p values are still high. Removing lwobup01.
MarathonFull3 <- glm(sodium_low ~  age + trainpse + female + as.factor(wateld01) + as.factor(fluidfr3) +
                     (runtime^2) + bmi + (bmi^2), data = MarathonCorrect, family = "binomial")
#msummary(MarathonFull3)

#Removing fluidfr3.
MarathonFull4 <- glm(sodium_low ~  age + trainpse + female + as.factor(wateld01) + wtdiff + runtime + (
#msummary(MarathonFull4)

#Removing Training Pace, same logic as above.
MarathonFull5 <- glm(sodium_low ~  age + female + as.factor(wateld01) + wtdiff + runtime + (runtime^2) +
#msummary(MarathonFull5)

#Removing Age, above logic again.
MarathonFull6 <- glm(sodium_low ~  female + as.factor(wateld01) + wtdiff + runtime + (runtime^2) + bmi +
#msummary(MarathonFull6)

#Removing BMI
MarathonFull7 <- glm(sodium_low ~  female + as.factor(wateld01) + wtdiff + runtime + (runtime^2), data =
#msummary(MarathonFull7)

#Wateld01.
MarathonFull8 <- glm(sodium_low ~  female + wtdiff + runtime + (runtime^2), data = MarathonCorrect, fami
#msummary(MarathonFull8)

#Runtime
MarathonFull9 <- glm(sodium_low ~  female + wtdiff, data = MarathonCorrect, family = "binomial")
msummary(MarathonFull9)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9098     0.3039  -9.575  < 2e-16 ***
female        1.0062     0.3705   2.715  0.00662 **
wtdiff        0.8558     0.1371   6.242 4.33e-10 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 275.54  on 454  degrees of freedom
Residual deviance: 211.42  on 452  degrees of freedom
  (33 observations deleted due to missingness)
AIC: 217.42

Number of Fisher Scoring iterations: 6
```

References