Choose Your Own Project

Greg Baguhin

28/07/2021

INTRODUCTION This project attempts to fit a linear & logistical regression model to a clothes size dataset. It will try to correlate the effectiveness of the predictors in giving accurate predictions for cloth size.

OBJECTIVE: I will be using the cloth_size dataset downloaded from Kaggle to identify the most significant predictor that would produce the most accurate cloth size for a given variable.

DATA PREPARATION Splitting the dataset into training and testing sets at a 50-50 ratio. This was chosen due to the relatively small amount of data used so the test set can be just as random as the training set.

```r
#install.packages("tidyverse")
library(tidyverse)
#install.packages("tidyr")
library(tidyr)
#install.packages("caret")
library(caret)


#Data Preparation:
# Loading the csv file into RStudio
url <- "D:/Desktop/Harvard Online/Data Science/Capstone/archive/cloth_size.csv"
clothsize_data <- read_csv(url)

# View first 6 rows
head(clothsize_data)
```

```
## # A tibble: 6 x 4
##   weight   age height size
##    <dbl> <dbl>  <dbl> <chr>
## 1     62    28   173. XL
## 2     59    36   168. L
## 3     61    34   165. M
## 4     65    27   175. L
## 5     62    45   173. M
## 6     50    27   160. S
```

```r
# inspect data properties
str(clothsize_data)
```

```
## tibble [119,153 x 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ weight: num [1:119153] 62 59 61 65 62 50 53 51 54 53 ...
##  $ age   : num [1:119153] 28 36 34 27 45 27 65 33 26 32 ...
##  $ height: num [1:119153] 173 168 165 175 173 ...
##  $ size  : chr [1:119153] "XL" "L" "M" "L" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   weight = col_double(),
##   ..   age = col_double(),
##   ..   height = col_double(),
##   ..   size = col_character()
##   .. )
```

```r
# inspect data for statistics
summary(clothsize_data)
```

```
##      weight            age              height           size          
##  Min.   : 22.00   Min.   :  0.00   Min.   :137.2   Length:119153     
##  1st Qu.: 55.00   1st Qu.: 29.00   1st Qu.:160.0   Class :character  
##  Median : 61.00   Median : 32.00   Median :165.1   Mode  :character  
##  Mean   : 61.76   Mean   : 34.03   Mean   :165.8                     
##  3rd Qu.: 67.00   3rd Qu.: 37.00   3rd Qu.:170.2                     
##  Max.   :136.00   Max.   :117.00   Max.   :193.0
```

```r
# Find out how many rows with NAs
sum(is.na(clothsize_data))
```

```
## [1] 0
```

```r
# converting size to a factor
clothsize_data <- clothsize_data %>% mutate(size=as_factor(size))

# define the outcome and predictors
y <- clothsize_data$size
a <- clothsize_data$age
b <- clothsize_data$weight
c <- clothsize_data$height


# generate training and test sets
set.seed(1, sample.kind = "Rounding") # if using R 3.5 or earlier, remove the sample.kind argument
test_index <- createDataPartition(y, times = 1, p = 0.5, list = FALSE)
test_set <- clothsize_data[test_index, ]
train_set <- clothsize_data[-test_index, ]
```

LINEAR REGRESSION MODEL Let's look at the predictor averages for each size

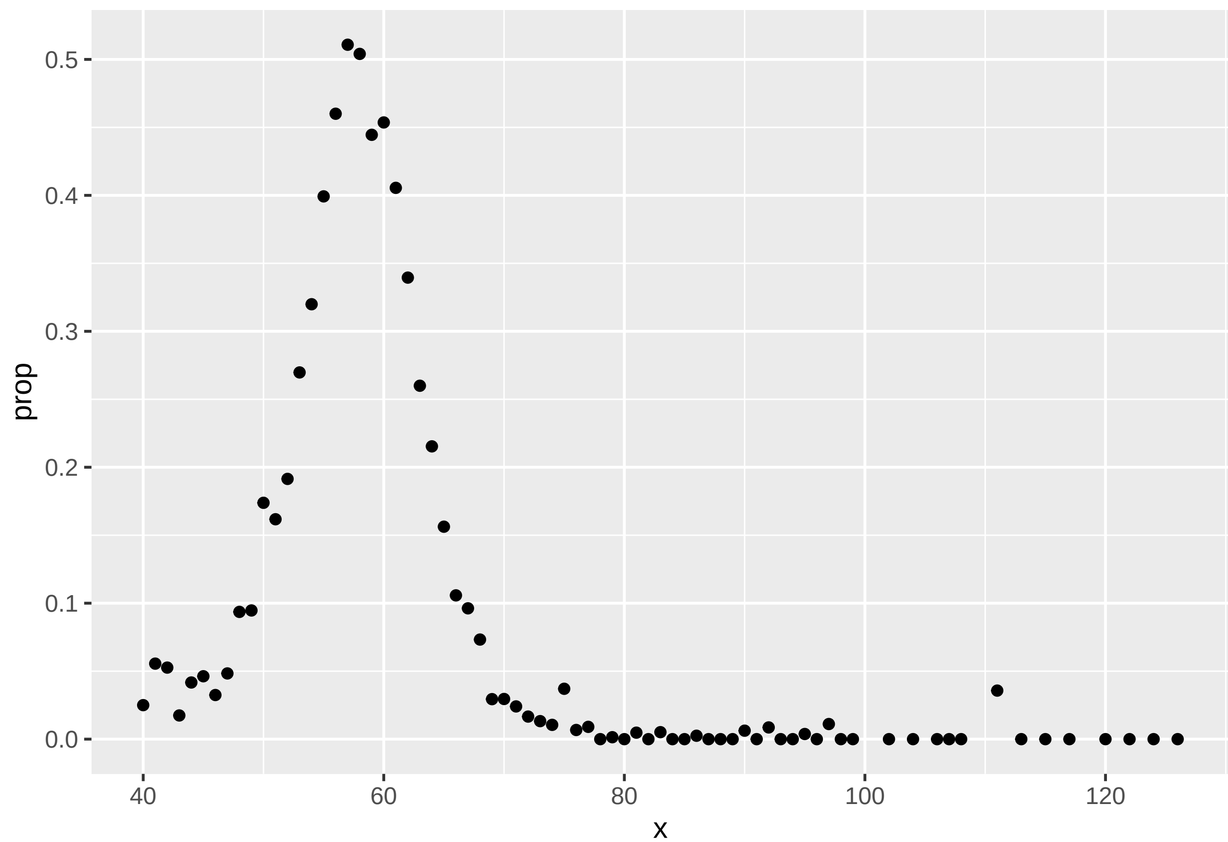
```r
clothsize_data %>% group_by(size) %>%
  summarise(avg_wt = mean(weight), avg_ht = mean(height), avg_age = mean(age))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 7 x 4
##   size  avg_wt avg_ht avg_age
##   <fct>  <dbl>  <dbl>   <dbl>
## 1 XL      65.6   168.    34.9
## 2 L       62.2   167.    34.2
## 3 M       58.2   165.    33.5
## 4 S       54.1   164.    32.6
## 5 XXS     50.5   161.    31.6
## 6 XXXL    75.9   168.    36.4
## 7 XXL     66.4   160.    36.3
```

Plotting probability of a size "M" vs weight

```
clothsize_data %>%
  mutate(x = round(weight)) %>%
  group_by(x) %>%
  filter(n() >= 10) %>%
  summarize(prop = mean(size == "M")) %>%
  ggplot(aes(x, prop)) +
  geom_point()
```



We see that a person who is 58kgs has a 50% probability of wearing a size "M"

```
train_set %>%
  filter(round(weight)==58) %>%
  summarize(y_hat = mean(size=="M"))
```

```
## # A tibble: 1 x 1
##   y_hat
##   <dbl>
## 1 0.503
```

LOGICTIS REGRESSION MODEL Fit logistic regression model using weight as a predictor for size M, the size with the highest frequency in the dataset.

```r
glm_fit <- train_set %>%
  mutate(y = as.numeric(weight==58)) %>%
  glm(y ~ size, data=., family = "binomial")

p_hat_logit <- predict(glm_fit, newdata = test_set, type = "response")
```

CONCLUSION: We were able to show that we can use linear & logistic regression to model predictions for cloth size using the dataset. A more thorough testing algorithm could be built in future projects applying the same principles learned from this project.