Case 02 EC

Michael Li

March 29, 2021

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
df <- read_csv('case_02_data.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   G = col_double(),
##   AB = col_double(),
##   H = col_double(),
##   AVG = col_double(),
##   salary = col_double(),
##   allstar = col_double(),
##   birthYear = col_double(),
##   birthCountry = col_character(),
##   weight = col_double(),
##   height = col_double(),
##   bats = col_character(),
##   debutYear = col_double(),
##   ageDebut = col_double()
## )
```

```
df
```

```
## # A tibble: 1,053 x 13
##        G    AB     H   AVG salary allstar birthYear birthCountry weight height
##    <dbl> <dbl> <dbl> <dbl>  <dbl>   <dbl>     <dbl> <chr>         <dbl>  <dbl>
##  1   626  1990   524 0.263 8.00e6       1      1974 Other           220     72
##  2   165   413   102 0.247 5.50e5       0      1984 Other           200     70
##  3   176   328    70 0.213 5.10e5       0      1991 Other           217     71
##  4   818  2336   609 0.261 7.98e5       0      1988 USA             245     75
##  5   151   396    92 0.232 5.01e5       0      1985 Other           210     74
##  6   422   990   246 0.248 5.07e5       0      1989 Other           195     73
##  7   613  2014   475 0.236 5.15e5       0      1990 USA             200     74
```

```
##  8   167   428    81 0.189 5.10e5       0      1991 Other           170     70
##  9   116   344    70 0.203 4.82e5       0      1986 USA             230     74
## 10    16    29     3 0.103 4.14e5       0      1978 Other           245     74
## # ... with 1,043 more rows, and 3 more variables: bats <chr>, debutYear <dbl>,
## #   ageDebut <dbl>
```

```
# create variable to indicate whether or not player had batting average above 0.3
df_mod <- df %>%
  mutate(bat30 = if_else(AVG >= 0.3, 1, 0))
```

```
# calculate probabilities
picalc <- function(X,beta){
  pi <- 1:nrow(X)
  expn <- 1:nrow(X)
  for (i in 1:nrow(X)){
    expn[i] <- 0
    for (j in 1:ncol(X)){
      expo <- X[i,j] * beta[j]
      expn[i] <- expo + expn[i]
    }
    pi[i] <- exp(expn[i])/(1+exp(expn[i]))
    }
  return(pi)
}
```

```
# find W
Wcalc <- function(pi){
  W <- matrix(0,length(pi),length(pi))
  for (i in 1:length(pi)){
    W[i,i] <- pi[i]*(1-pi[i])
    }
  return(W)
}
```

```
# logistic function
myglm <- function(X,Y,covs, obs, dif) {
  beta <- rep(0, (covs+1))
  intercept <- rep(1, obs)
  X_n <- cbind(intercept,X)
  deriv <- 1:(covs+1)
  diff <- 100000
  while(diff > dif) { # Newton Raphson method
    pi <- as.vector(picalc(X_n,beta))
    W <- Wcalc(pi)
    deriv <- (solve(t(X_n)%*%W%*%as.matrix(X_n))) %*% (t(X_n)%*%(Y - pi))
    beta = beta + deriv
    diff <- sum(deriv^2)
    }
  return(beta)
}
```

```
myglm(df_mod[,c(9:10)], df_mod$bat30, 2, nrow(df_mod), 0.0000001)
```

```
##                   [,1]
## intercept  1.447244466
## weight    -0.009926185
```

```
## height    -0.055298269
```

We can use the above binary logistic regression model to predict whether or not a player would have a batting average above 0.300. The model above uses weight and height as covariates, allowing you to plug in values and use the coefficient point estimates to predict whether or not a players has over a 0.300 batting average.

References

[1] Srivastava, T. https://www.analyticsvidhya.com/blog/2015/10/basics-logistic-regression/.