Analysing College Admissions from SAT scores

Syed Faaris Razi

2022-08-02

In a Data Science Program from 2019 (link to **Course and my Certificate of completion**), we were given various datasets to analyse and bring insights from, where one such data was on **College Admissions based on SAT scores** of students.

Using *Python*, we applied Logistic Regression on said data to find the probability of Admittance from SAT scores. Here we shall implement the same analysis using *R*.

The Data

```
library(tidyverse)
library(knitr, quietly = T)

admittance.raw = read.csv("2.01.+Admittance.csv") # Loading the data

# Map "No"/"Yes" values to an indicator column (Group) as 0's and 1's
admittance.SAT = admittance.raw %>% mutate(Group = Admitted) %>%
            mutate(Group = replace(Group, Group == "No", 0)) %>%
            mutate(Group = replace(Group, Group == "Yes", 1)) %>%
            mutate(Group = as.numeric(Group))

# Display both our raw and modified dataframes
df1 = kable(head(admittance.raw), format = 'latex')
df2 = kable(head(admittance.SAT), format = 'latex')

cat(sprintf("Preview of our CSV data (of %g rows):", nrow(admittance.raw)),
    c("\\begin{table}[h] \\centering ", df1, "\\hspace{1cm} \\centering ", df2,
    "\\caption{Raw dataframe (left),  New dataframe (right)} \\end{table}"))
```

Preview of our CSV data (of 168 rows):

SAT	Admitted
1363	No
1792	Yes
1954	Yes
1653	No
1593	No
1755	Yes

SAT	Admitted	Group
1363	No	0
1792	Yes	1
1954	Yes	1
1653	No	0
1593	No	0
1755	Yes	1

Table 1: Raw dataframe (left), New dataframe (right)

From the above, our raw data contains a numerical *SAT* column and a textual *Admitted* column of binary “No” and “Yes” responses. We modified this dataset to have a *Group* column of “No”/“Yes” values translated as numeric 0’s and 1’s (approriate format for Logistic Regression).

Statistical summary of our Logisitic Regression:

```r
# glm()'s family = "binomial", since our dependant variable is binary (0's and 1's).
admittance.SAT.logit = glm(Group ~ SAT, data = admittance.SAT, family = binomial)
logreg_summary = summary(admittance.SAT.logit); logreg_summary
```

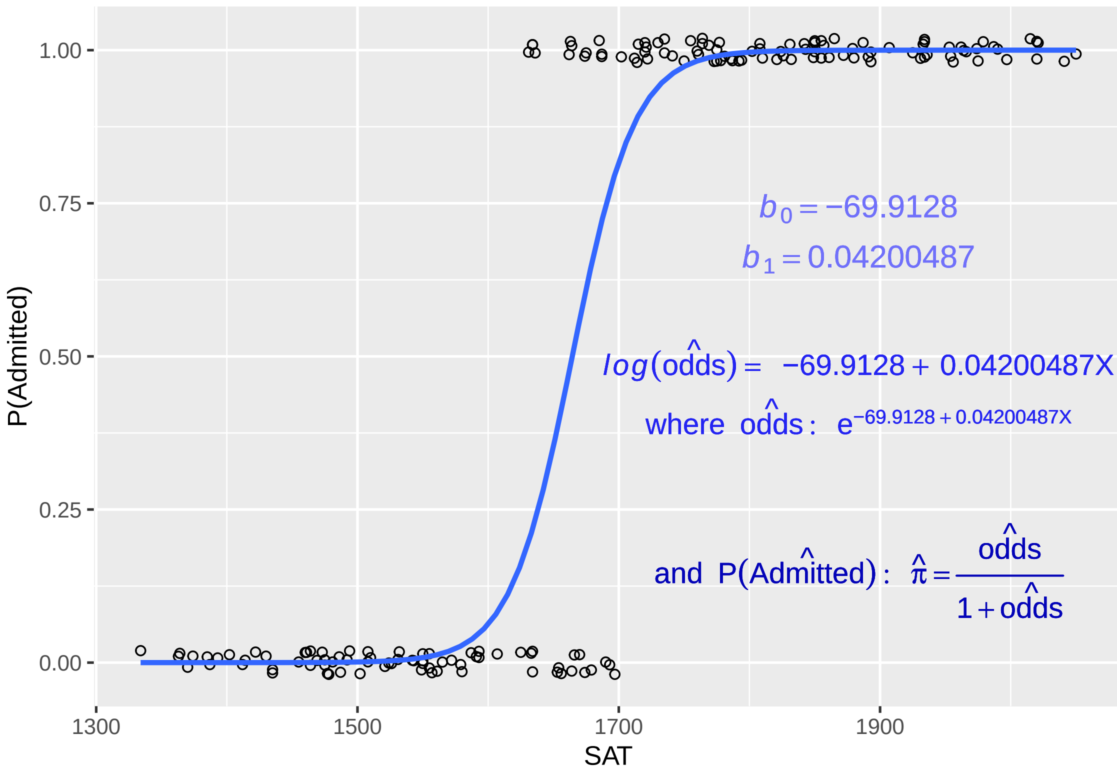
```
##
## Call:
## glm(formula = Group ~ SAT, family = binomial, data = admittance.SAT)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.78661  -0.04825   0.00199   0.07157   1.80151
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -69.912802  15.735757  -4.443 8.87e-06 ***
## SAT           0.042005   0.009431   4.454 8.43e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 230.511  on 167  degrees of freedom
## Residual deviance:  46.289  on 166  degrees of freedom
## AIC: 50.289
##
## Number of Fisher Scoring iterations: 8
```

Visualizing the Logisitic Regression

```r
b0 = logreg_summary$coefficients[1,1] # Getting the coefficients to
b1 = logreg_summary$coefficients[2,1] # be used inside our ggplot

# Setting annotations for our Plot
text_1 = bquote(atop(italic(b[0]) == .(b0), italic(b[1]) == .(b1)))
text_2 = bquote(atop(
                    italic(log)(hat(odds)) == ~~.(b0) ~ + ~ .(b1)*X,
                    where ~~ hat(odds):~~ e^{.(b0) ~ + ~ .(b1)*X}))
text_3 = bquote(and ~~ P(hat(Admitted)): ~~ hat(pi) == frac(hat(odds),1+hat(odds)))
x_text = mean(admittance.SAT$SAT)*(1 + 1/9) # key x-position for the labels

# Plotting the Logistic Regression with fitted values
ggplot(admittance.SAT, aes(x = SAT, y = Group)) +
  geom_point(shape=1, position = position_jitter(width = .02, height = .02)) +
  stat_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE) +
  labs(y = "P(Admitted)") +
  annotate("text", x = x_text, y = 0.7, label = text_1, cex = 4.5, col = "#7070fa") +
  annotate("text", x = x_text, y = 0.45, label = text_2, cex = 4.2, col = "#2424f2") +
  annotate("text", x = x_text, y = 0.15, label = text_3, cex = 4.2, col = "#0404b8")
```


$b_0 = -69.9128$
$b_1 = 0.04200487$
$log(\hat{odds}) = -69.9128 + 0.04200487X$
where $\hat{odds}$: $e^{-69.9128 + 0.04200487X}$
and P(Admitted): $\hat{\pi} = \frac{\hat{odds}}{1+\hat{odds}}$
P(Admitted)
SAT
1.00
0.75
0.50
0.25
0.00
1300
1500
1700
1900