Logistic Regression for Y with 2 Categories

Logistic Regression

Based on Chew C. H. (2020) textbook: AI, Analytics and Data Science. Vol 1., Chap 7.

Logistic Regression Model for Binary Y

$$Y = 0 \text{ or } 1$$

$$Z = b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m$$

$$P(Y = 1) = \frac{1}{1 + e^{-z}}$$

Example: Predicting Pass/Fail Exam

Source: Wikipedia

- A group of 20 students sat for an exam.

  Question: How does the number of hours spent studying affect the outcome of the exam (Pass/Fail)?

- The dataset (passexam.csv) shows the number of hours each student spent studying, and whether they passed (1) or failed (0).

	A	B
1	Hours	Outcome
2	0.5	0
3	0.75	0
4	1	0
5	1.25	0
6	1.5	0
7	1.75	0
8	1.75	1
9	2	0
10	2.25	1
11	2.5	0
12	2.75	1
13	3	0
14	3.25	1
15	3.5	0
16	4	1
17	4.25	1
18	4.5	1
19	4.75	1
20	5	1
21	5.5	1

Base R: glm() function with family = binomial

Rscript: passexam.R

```
10  library(data.table)
11
12  setwd('D:/Dropbox/Datasets/ADA1/7_Logistic_Reg')
13
14  passexam.dt <- fread("passexam.csv")
15  passexam.dt$Outcome <- factor(passexam.dt$Outcome)
16
17  summary(passexam.dt)
18
19  pass.m1 <- glm(Outcome ~ Hours , family = binomial, data = passexam.dt)
20
21  summary(pass.m1)
```

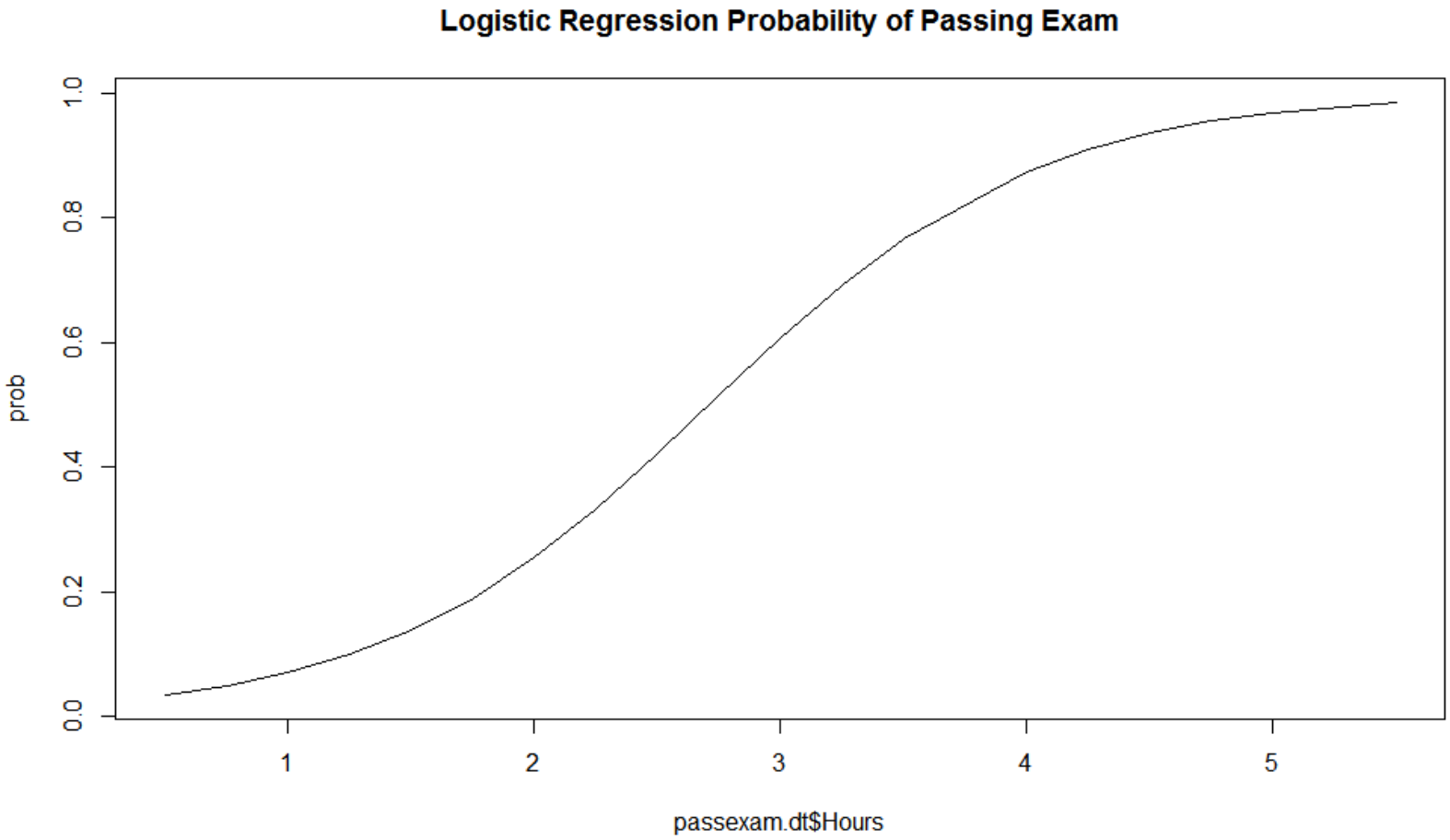

Results from summary() function

```
> summary(pass.m1)

Call:
glm(formula = Pass ~ Hours, family = binomial, data = passexam.dt
)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.70557  -0.57357  -0.04654   0.45470   1.82008

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0777     1.7610  -2.316   0.0206 *
Hours         1.5046     0.6287   2.393   0.0167 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$z = -4.0777 + 1.5046(Hours)$$

$$P(Y = 1) = \frac{1}{1 + e^{-z}}$$

Hours is statistically significant factor.




```r
# Output the probability from the logistic function for all cases in the data.
prob <- predict(pass.m1, type = 'response')
# See the S curve
plot(x = passexam.dt$Hours, y = prob, type = "l", main = 'Logistic Regression Probability of Passing Exam')
```

Outputs the logistic function P(Y = 1)

Get model predicted Y (i.e. y.hat) by comparing P(Y = 1) against threshold.
Then get Confusion Matrix by comparing y.hat vs actual Y.

```r
# Set the threshold for predicting Y = 1 based on probability.
threshold <- 0.5

# If probability > threshold, then predict Y = 1, else predict Y = 0.
y.hat <- ifelse(prob > threshold, 1, 0)

# Create a confusion matrix with actuals on rows and predictions on columns.
table(passexam.dt$Outcome, y.hat, deparse.level = 2)

# Overall Accuracy
mean(y.hat == passexam.dt$Outcome)
```

```
> table(passexam.dt$Outcome, y.hat, deparse.level = 2)
                    y.hat
passexam.dt$Outcome 0 1
                  0 8 2
                  1 2 8
> # Overall Accuracy
> mean(y.hat == passexam.dt$Outcome)
[1] 0.8
```

Display Row name and Column name

What is the meaning of the model coefficient?

$$z = -4.0777 + 1.5046(Hours)$$

$$P(Y = 1) = \frac{1}{1 + e^{-z}}$$

Next: Odds and Odds Ratio.