

# Logistic Regression

## CS5056 Data Analytics

Francisco J. Cantú, Héctor Ceballos

Tecnológico de Monterrey

March 17, 2021

February-June, 2021

# Logistic Regression

- The Logistic Model (or logit model) is used to model the probability of a binary class or event such as pass/fail, win/lose, alive/dead or healthy/sick
- This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one
- It uses a **logistic function** to model a binary dependent variable, although many more complex extensions exist
- In regression analysis, **logistic regression** (or logit regression) is estimating the parameters of a logistic model (a form of binary regression)

[https://en.wikipedia.org/wiki/Logistic\\_regression](https://en.wikipedia.org/wiki/Logistic_regression)

# Logistic Function

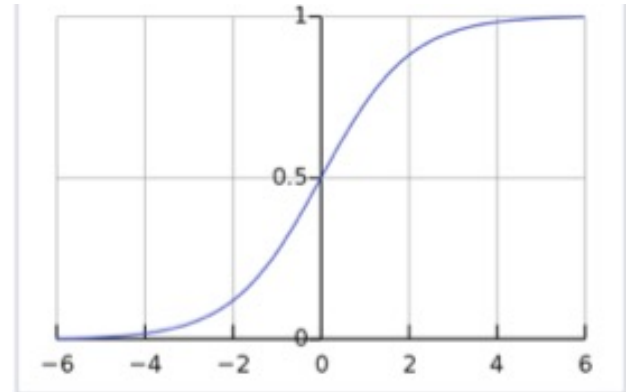
A **logistic function** or **logistic curve** is a common "S" shape (**sigmoid curve**), with equation:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

where

- $e$  = the **natural logarithm** base (also known as **Euler's number**),
- $x_0$  = the  $x$ -value of the sigmoid's midpoint,
- $L$  = the curve's maximum value, and
- $k$  = the logistic growth rate or steepness of the curve.<sup>[1]</sup>

For values of  $x$  in the domain of **real numbers** from  $-\infty$  to  $+\infty$ , the S-curve shown on the right is obtained, with the graph of  $f$  approaching  $L$  as  $x$  approaches  $+\infty$  and approaching zero as  $x$  approaches  $-\infty$ .



Standard logistic sigmoid function i.e.  
 $L = 1, k = 1, x_0 = 0$

$$f(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$

[https://en.wikipedia.org/wiki/Logistic\\_function](https://en.wikipedia.org/wiki/Logistic_function)

# Logistic Regression

- Consider a model with two predictors,  $X_1$  and  $X_2$  and one binary (Bernoulli) response variable  $Y$ , which we denote  $p = P(Y = 1)$
- We assume a linear relationship between the predictor variables and the log-odds of the event that  $Y=1$
- This linear relationship can be written in the following mathematical form where  $\ell$  is the log-odds,  $b$  is the base of the logarithm, and betas the coefficient parameters of the model:

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

We can recover the **odds** by exponentiating the log-odds:

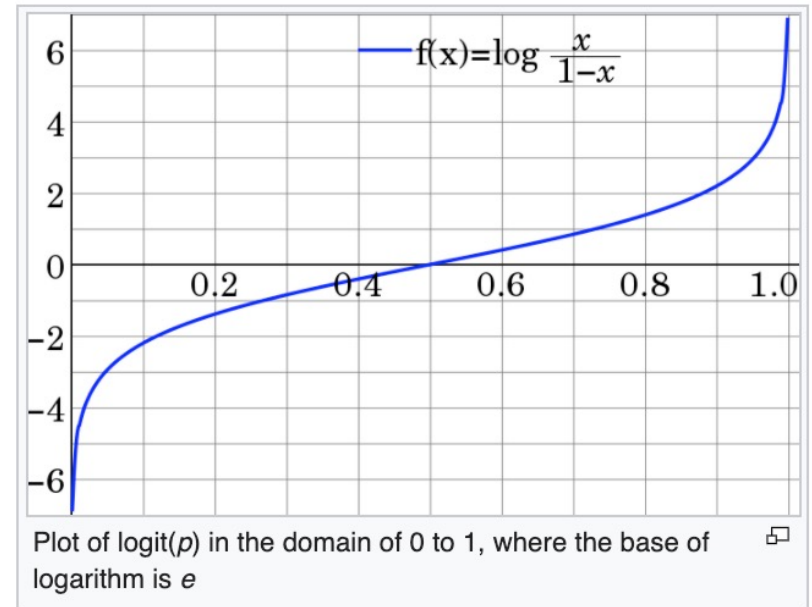
$$\frac{p}{1-p} = b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}.$$

By simple algebraic manipulation, the probability that  $Y = 1$  is

$$p = \frac{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} + 1} = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}.$$

# Logit Function

- The Logit function or the log-odds is the logarithm of the odds  $p/(1-p)$  where  $p$  is a probability value
- It is a type of function that creates a map of probability values from  $(0,1)$  to  $(-\infty, +\infty)$
- It is the inverse of the sigmoidal "logistic" function or logistic transform



$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p) = -\log\left(\frac{1}{p} - 1\right)$$

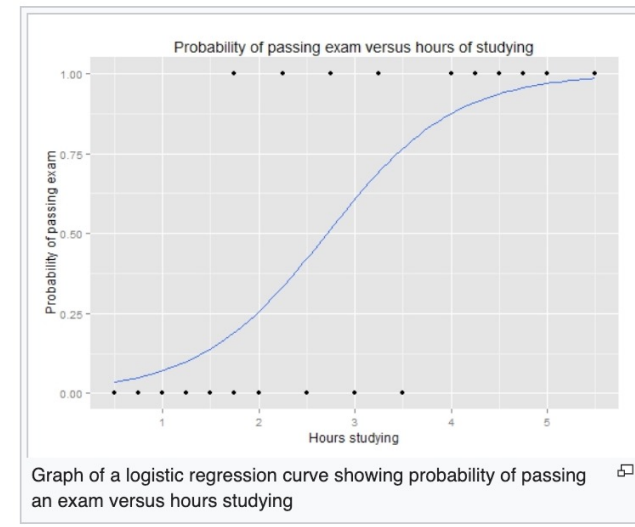
A group of 20 students spends between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability of the student passing the exam?

The reason for using logistic regression for this problem is that the values of the dependent variable, pass and fail, while represented by "1" and "0", are not **cardinal numbers**. If the problem was changed so that pass/fail was replaced with the grade 0–100 (cardinal numbers), then simple **regression analysis** could be used.

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


Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

The graph shows the probability of passing the exam versus the number of hours studying, with the logistic regression curve fitted to the data.



# Example

# Example



	Coefficient	Std.Error	z-value	P-value (Wald)
Intercept	-4.0777	1.7610	-2.316	0.0206
Hours	1.5046	0.6287	2.393	0.0167

The output indicates that hours studying is significantly associated with the probability of passing the exam ( $p = 0.0167$ , [Wald test](#)). The output also provides the coefficients for Intercept =  $-4.0777$  and Hours =  $1.5046$ . These coefficients are entered in the logistic regression equation to estimate the odds (probability) of passing the exam:

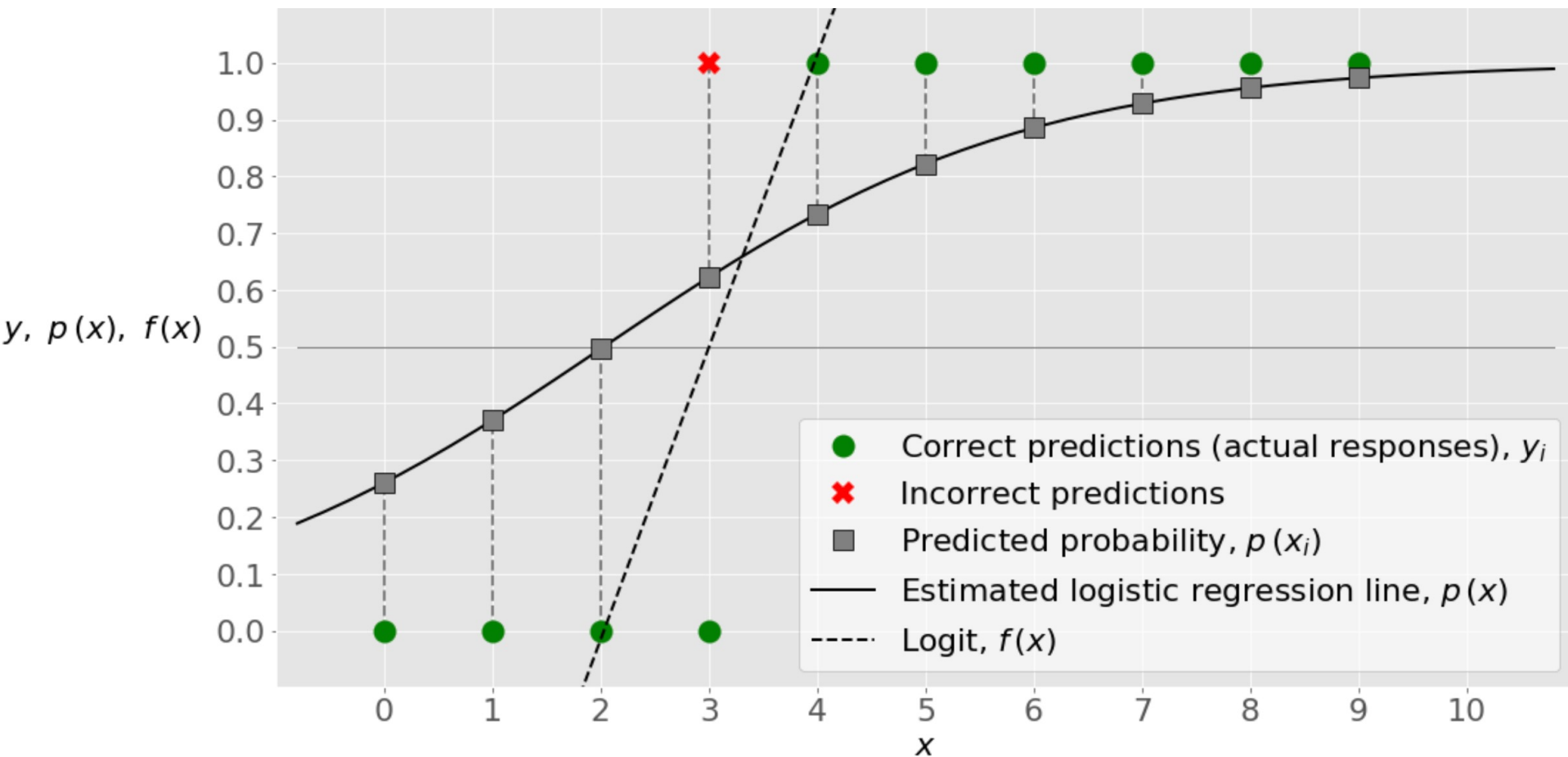
# Hands-on Exercise with Logistic Regression using Python

Load Python script logit.py

---

<https://realpython.com/logistic-regression-python/>







## Logistic Regression using R

<https://stats.idre.ucla.edu/r/dae/logit-regression/>

```
mydata <-  
read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")  
head(mydata)  
summary(mydata)  
sapply(mydata, sd)  
xtabs(~admit + rank, data = mydata)  
mydata$rank <- factor(mydata$rank)  
mylogit <- glm(admit ~ gre + gpa + rank, data =  
mydata, family = "binomial")  
summary(mylogit)
```

# Hands-on Exercise Logistic Regression

- The Call shows what the model we ran was, what options we specified, etc
- The deviance residuals are a measure of model fit. It shows the distribution of the deviance residuals for individual cases used in the model.
- The coefficients have their standard errors, the z-statistic (sometimes called a Wald z-statistic), and the associated p-values
- Both gre and gpa are statistically significant, as are the three terms for rank
- The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable:
- For every one unit change in gre, the log odds of admission (versus non-admission) increases by 0.002.
- For a one unit increase in gpa, the log odds of being admitted to graduate school increases by 0.804.
- The indicator variables for rank have a slightly different interpretation
- For example, having attended an undergraduate institution with rank of 2, versus an institution with a rank of 1, changes the log odds of admission by -0.675.
- Below the table of coefficients are fit indices, including the null and deviance residuals and the AIC.

## Hands-on Exercise Logistic Regression

- We can use the `confint` function to obtain confidence intervals for the coefficient estimates
- Note that for logistic models, confidence intervals are based on the profiled log-likelihood function
- We can also get CIs based on just the standard errors by using the default method

```
confint(mylogit)
```

```
confint.default(mylogit)
```

```
install.packages("aod")
```

```
library(aod)
```

```
library(caret)
```

```
wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms = 4:6)
```

# Hands-on Exercise

## Logistic Regression

- We can test for an overall effect of rank using the `wald.test` function of the `aod` library
- The order in which the coefficients are given in the table of coefficients is the same as the order of the terms in the model
- This is important because the `wald.test` function refers to the coefficients by their order in the model
- We use the `wald.test` function. `b` supplies the coefficients, while `Sigma` supplies the variance covariance matrix of the error terms, finally `Terms` tells R which terms in the model are to be tested, in this case, terms 4, 5, and 6, are the three terms for the levels of rank.
- The chi-squared test statistic of 20.9, with three degrees of freedom is associated with a p-value of 0.00011 indicating that the overall effect of rank is statistically significant

**`wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms = 4:6)`**

# Hands-on Exercise

## Logistic Regression

- We can also test additional hypotheses about the differences in the coefficients for the different levels of rank
- Below we test that the coefficient for rank=2 is equal to the coefficient for rank=3. The first line of code below creates a vector `l` that defines the test we want to perform. In this case, we want to test the difference (subtraction) of the terms for rank=2 and rank=3 (i.e., the 4th and 5th terms in the model)
- To contrast these two terms, we multiply one of them by 1, and the other by -1. The other terms in the model are not involved in the test, so they are multiplied by 0.
- The second line of code below uses `L=l` to tell R that we wish to base the test on the vector `l` (rather than using the `Terms` option as we did above)
- The chi-squared test statistic of 5.5 with 1 degree of freedom is associated with a p-value of 0.019, indicating that the difference between the coefficient for rank=2 and the coefficient for rank=3 is statistically significant

```
l <- cbind(0, 0, 0, 1, -1, 0)
```

```
wald.test(b = coef(mylogit), Sigma = vcov(mylogit), L = l)
```

# Hands-on Exercise

## Logistic Regression

- You can also exponentiate the coefficients and interpret them as odds-ratios. R will do this computation for you
- To get the exponentiated coefficients, you tell R that you want to exponentiate (`exp`), and that the object you want to exponentiate is called coefficients and it is part of mylogit (`coef(mylogit)`)
- We can use the same logic to get odds ratios and their confidence intervals, by exponentiating the confidence intervals from before. To put it all in one table, we use `cbind` to bind the coefficients and confidence intervals column-wise

**`exp(coef(mylogit))`**

**`exp(cbind(OR = coef(mylogit), confint(mylogit)))`**

- Now we can say that for a one unit increase in `gpa`, the odds of being admitted to graduate school (versus not being admitted) increase by a factor of 2.23

# Hands-on Exercise

## Logistic Regression

```
newdata1 <- with(mydata, data.frame(gre = mean(gre), gpa = mean(gpa), rank = factor(1:4)))  
newdata1
```

```
newdata1$rankP <- predict(mylogit, newdata = newdata1, type = "response") newdata1
```

```
newdata2 <- with(mydata, data.frame(gre = rep(seq(from = 200, to = 800, length.out = 100), 4), gpa =  
mean(gpa), rank = factor(rep(1:4, each = 100))))
```

```
newdata3 <- cbind(newdata2, predict(mylogit, newdata = newdata2, type = "link", se = TRUE))  
newdata3 <- within(newdata3, { PredictedProb <- plogis(fit) LL <- plogis(fit - (1.96 * se.fit)) UL <-  
plogis(fit + (1.96 * se.fit)) })
```

```
head(newdata3)
```

```
ggplot(newdata3, aes(x = gre, y = PredictedProb)) + geom_ribbon(aes(ymin = LL, ymax = UL, fill = rank),  
alpha = 0.2) + geom_line(aes(colour = rank), size = 1)
```

```
with(mylogit, null.deviance - deviance)
```

```
with(mylogit, df.null - df.residual)
```

```
with(mylogit, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))
```

```
logLik(mylogit)
```



# Francisco J. Cantú-Ortiz, PhD

Professor of Computer Science and Artificial Intelligence  
Tecnológico de Monterrey  
Enago-Academy Advisor for Strategic Alliances

E-mail: [fcantu@tec.mx](mailto:fcantu@tec.mx), [fjcantor@gmail.com](mailto:fjcantor@gmail.com)

Cel: +52 81 1050 8294, SNI-2 CVU: 9804

Personal Page: <http://semtech.mty.itesm.mx/fcantu/>

Facebook: [fcantu](#); Twitter: [@fjcantor](#); Skype: [fjcantor](#)

Orcid: 0000-0002-2015-0562

Scopus ID:6701563520

Researcher ID: B-8457-2009

[https://www.researchgate.net/profile/Francisco\\_Cantu-Ortiz](https://www.researchgate.net/profile/Francisco_Cantu-Ortiz)

<https://scholar.google.com.mx/citations?hl=es&user=45-uuK4AAAAJ>

<https://itesm.academia.edu/FranciscoJavierCantuOrtiz>

Ave. Eugenio Garza Sada No. 2501, Monterrey N.L., C.P. 64849, México