Logistic regression, curvilinear regression, and the interpretation of results

Amm Quamruzzaman

ISF, UC Berkeley

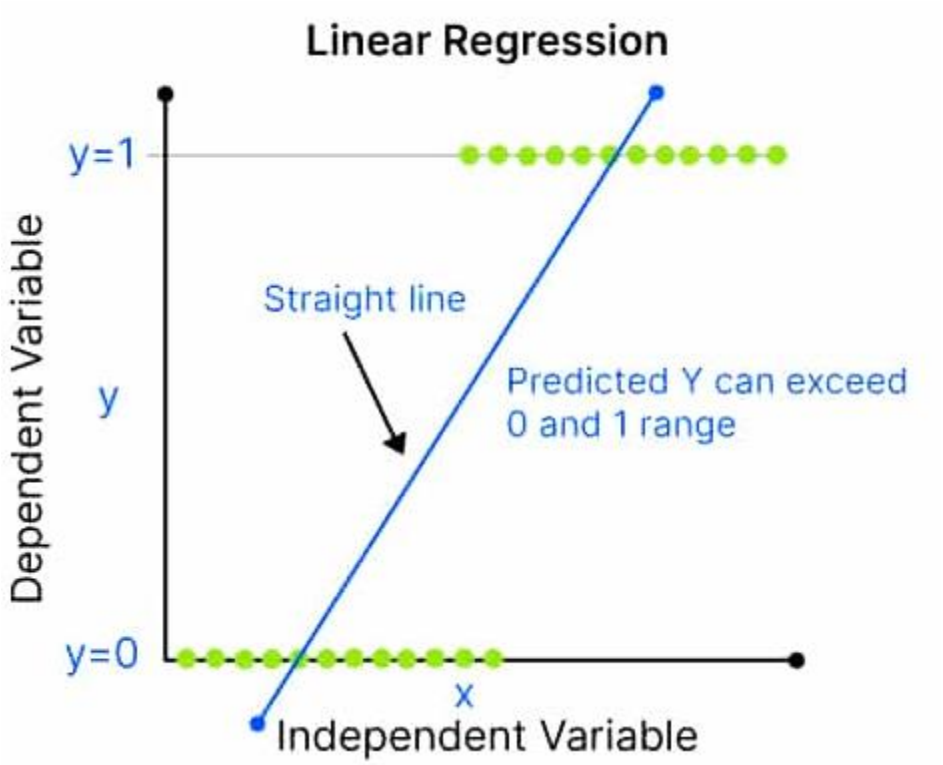
March 29, 2022

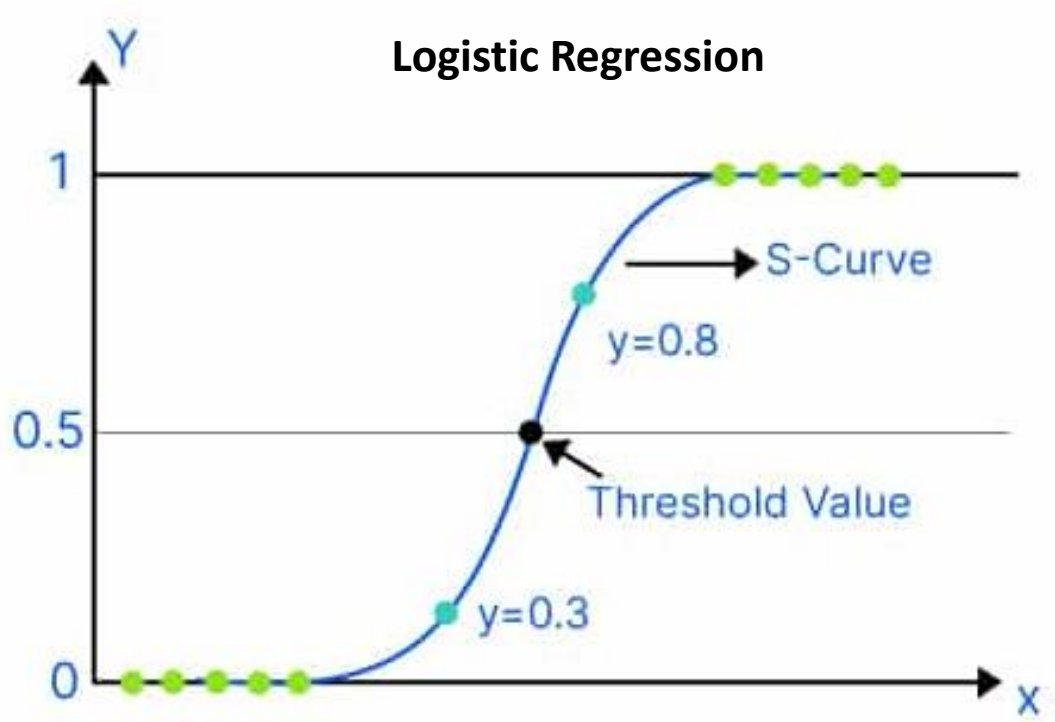
Outline

I. Introduction
II. The problem of classification
III. When to use which type of logistic regression?
IV. Simple or binomial logistic regression
V. Multinomial logistic regression
VI. Ordinal or ordered logistic regression
VII. Interpretation of results
VIII. Curvilinear or logistic?

I. Introduction

- While linear regression is good for prediction, we cannot use it to solve a problem of classification in the outcome (DV).





II. The problem of classification

- Binary classification (0 or 1): When the outcome variable has only two categories, such as success or failure, high or low, etc.
  - Use simple or binomial logistic or logit or probit regression.
- Multinomial classification: When the outcome has multiple categories but no rank order, such as you can go to college (academic), trade school (vocational), or into the workforce (general) based on your SES.
  - Must use multinomial logistic regression.
- Ordinal classification: The DV has multiple categories with a rank order, such as on a 0-5 scale; low, medium, high; unemployed, part-time employed, full-time employed, etc.
  - Use ordinal or ordered logistic regression.

III. When to use which type of logistic regression?

- Depends on:

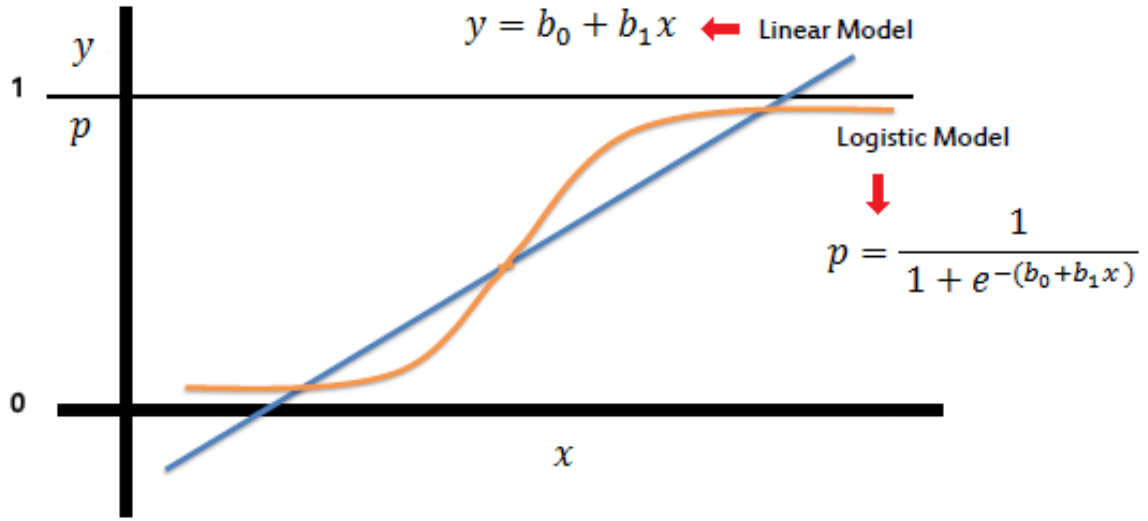
Types of Logistic Regression Models

	Binomial Logistic Regression	Multinomial Logistic Regression	Ordinal Logistic Regression
Number of Categories for Response Variable	2	3 or more	3 or more
Does Order of Categories Matter?	No	No	Yes

- Sometimes, logistic regression is problematic when the IVs are categorical based on some arbitrary categories (e.g., age from continuous to uneven categories like young, adult, old).
- In this case, curvilinear regression can be a better fit to the data [example at the end].

IV. Simple or binomial logistic regression

- Simple logistic regression is like simple linear regression.
- But the curve is constructed using the natural logarithm of the "odds" of the DV, rather than the probability.
- By this transformation, the logistic regression equation can be written in terms of an odds ratio.
- Taking the natural log of both sides, we can write the equation in terms of log-odds (logit).
- The coefficient (b_1) is the amount the logit (log-odds) changes with a one unit change in *x*.



$$\frac{p}{1-p} = \exp(b_0 + b_1 x) \qquad logit(p) = \ln\left(\frac{p}{1-p}\right)$$

$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x \qquad p = \frac{1}{1 + e^{-logit(p)}}$$

V. Multinomial logistic regression

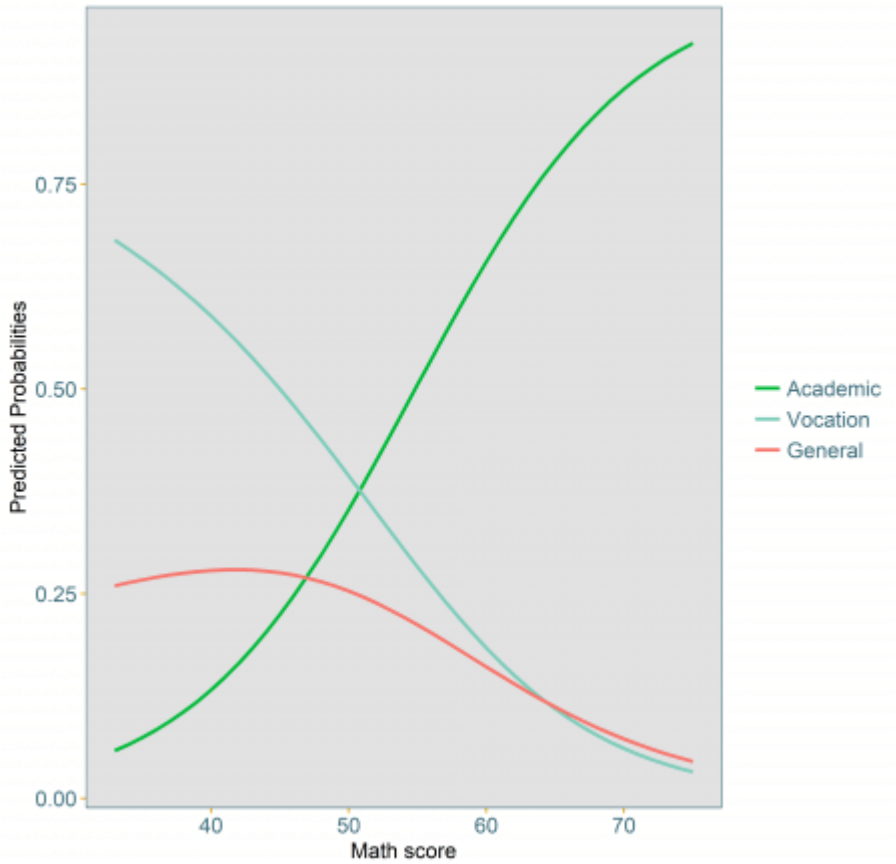
- Like multiple linear regression, logistic regression can handle any number of numerical and/or categorical variables.
- If the DV is a categorical variable with more than two categories where order does not matter, we can write:

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)}}$$

$$\hat{P}(Y_i = \text{academic}) = \frac{\exp[-5.0391 + 0.1099x_i]}{1 + \exp[-5.0391 + 0.1099x_i] + \exp[2.8996 - 0.0599x_i]}$$

$$\hat{P}(Y_i = \text{vocational}) = \frac{\exp[2.8996 - 0.0599x_i]}{1 + \exp[-5.0391 + 0.1099x_i] + \exp[2.8996 - 0.0599x_i]}$$

$$\hat{P}(Y_i = \text{general}) = \frac{1}{1 + \exp[-5.0391 + 0.1099x_i] + \exp[2.8996 - 0.0599x_i]}$$



VI. Ordinal or ordered logistic regression

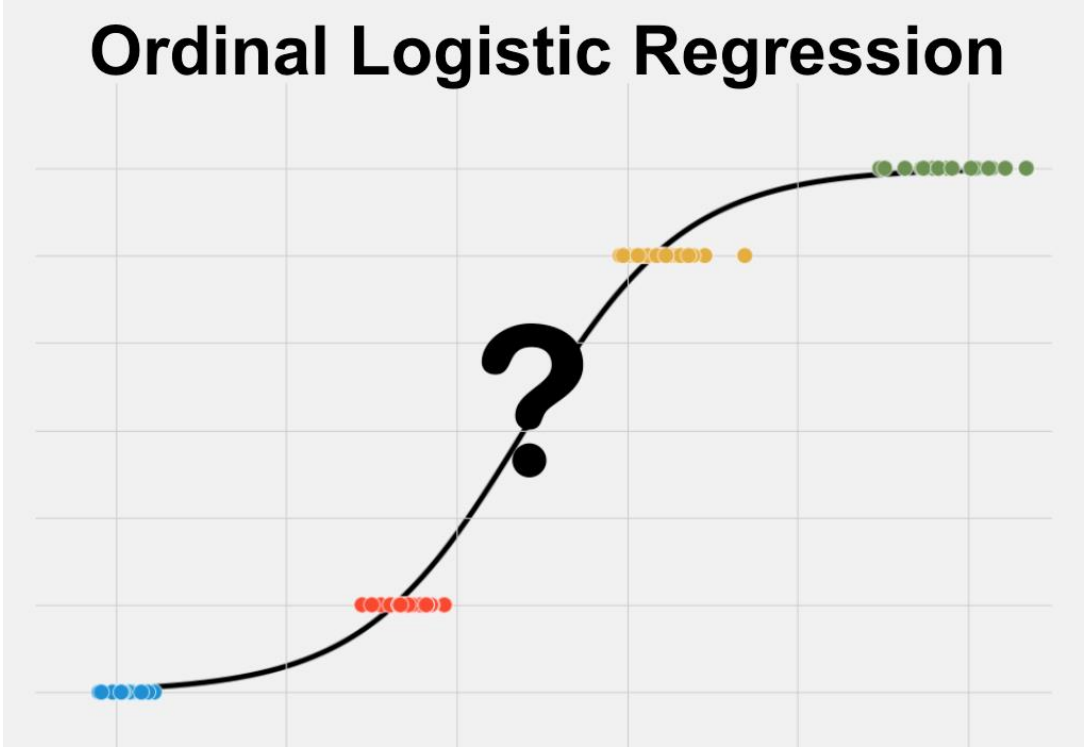
- Suppose, in a survey the proportions of respondents who would answer "poor", "fair", "good", "very good", and "excellent" health are respectively p1, p2, p3, p4, p5.
- The logarithms of the odds of answering in certain ways are:

$$\text{poor}, \quad \log \frac{p_1}{p_2+p_3+p_4+p_5}, \quad 0$$

$$\text{poor or fair}, \quad \log \frac{p_1+p_2}{p_3+p_4+p_5}, \quad 1$$

$$\text{poor, fair, or good}, \quad \log \frac{p_1+p_2+p_3}{p_4+p_5}, \quad 2$$

$$\text{poor, fair, good, or very good}, \quad \log \frac{p_1+p_2+p_3+p_4}{p_5}, \quad 3$$



VII. Interpretation of results

```
use https://stats.idre.ucla.edu/stat/data/hsb2, clear

generate honcomp = (write >=60)
logit honcomp female read science

Iteration 0:   log likelihood = -115.64441
Iteration 1:   log likelihood = -84.558481
Iteration 2:   log likelihood = -80.491449
Iteration 3:   log likelihood = -80.123052
Iteration 4:   log likelihood = -80.118181
Iteration 5:   log likelihood =  -80.11818

Logit estimates                                   Number of obs   =        200
                                                  LR chi2(3)      =      71.05
                                                  Prob > chi2     =     0.0000
Log likelihood =  -80.11818                       Pseudo R2       =     0.3072

------------------------------------------------------------------------------
     honcomp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   1.482498   .4473993     3.31   0.001     .6056111    2.359384
        read |   .1035361   .0257662     4.02   0.000     .0530354    .1540369
     science |   .0947902   .0304537     3.11   0.002      .035102    .1544784
       _cons |   -12.7772    1.97586    -6.47   0.000    -16.64982   -8.904589
------------------------------------------------------------------------------
```

This classification is arbitrary

Iteration log or LL

VII. Interpretation of results…

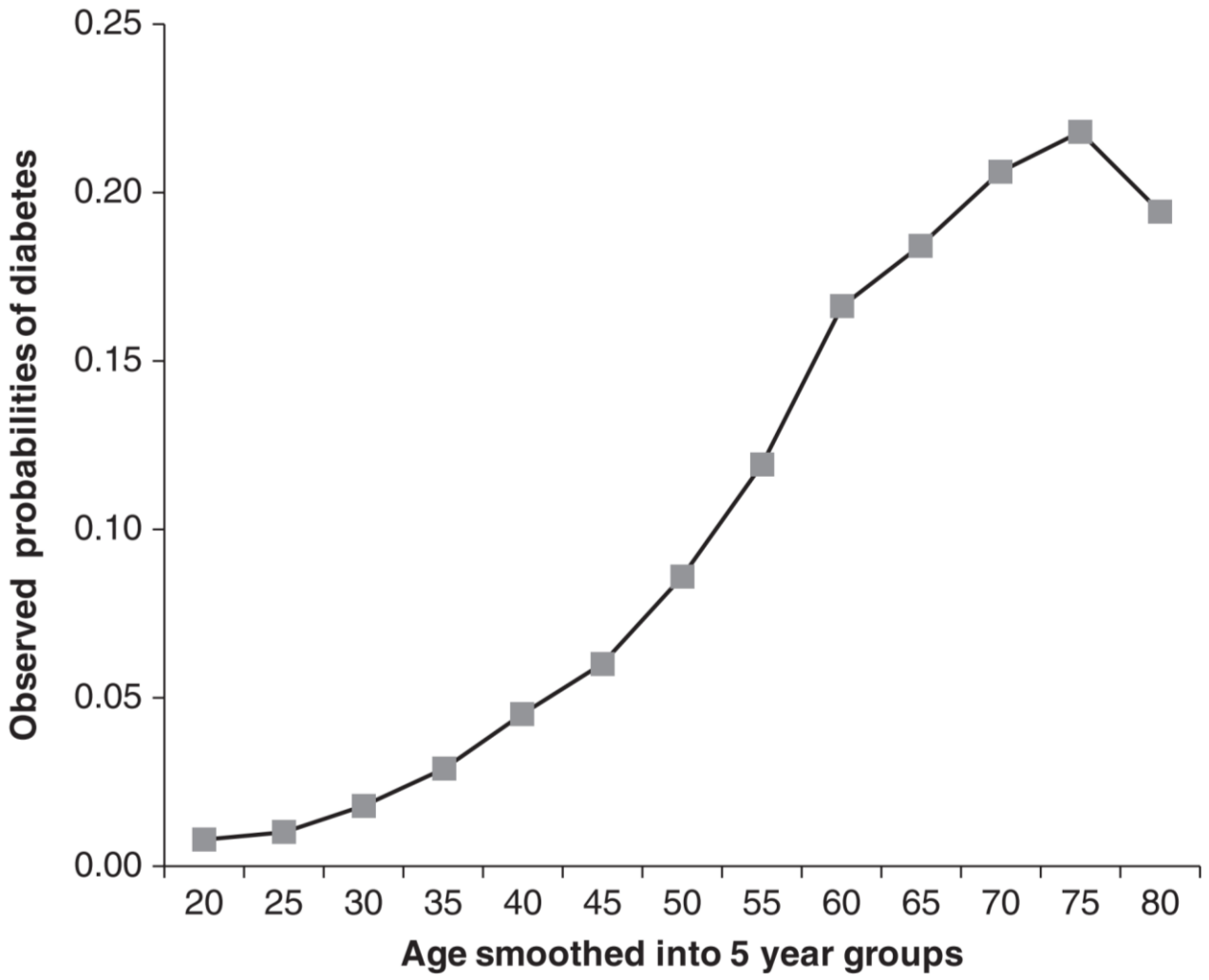
- The coefficients are given in log-odds units.
- They are often difficult to interpret, so they are often converted into odds ratios, by using “logistic” command or writing “or” option if you use “logit” command.
- Alternatively, use “probit” that gives you coefficients like OLS.

```
. logistic honcomp female

Logistic regression                               Number of obs   =        200
                                                  LR chi2(1)      =       3.94
                                                  Prob > chi2     =     0.0473
Log likelihood =  -113.6769                       Pseudo R2       =     0.0170

------------------------------------------------------------------------------
     honcomp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   1.918168   .6400451     1.95   0.051     .9973827    3.689024
------------------------------------------------------------------------------
```

Odds ratio <1 means less likely, whereas >1 means more likely. OR cannot be negative.

VIII. Curvilinear or logistic?

- If existing theory suggests a curvilinear pattern, do not use logistic function; use a curvilinear model (usually, by adding a squared or cubed term to the linear regression equation).
- Theory: The probability of being diagnosed with diabetes is low in early life, then it accelerates at later ages, finally slowing down.



VIII. Curvilinear or logistic?

- Curvilinear regression fits the observed probabilities better than logit (log-odds of linear coefficients).