Logistic Regression

13 January 2026

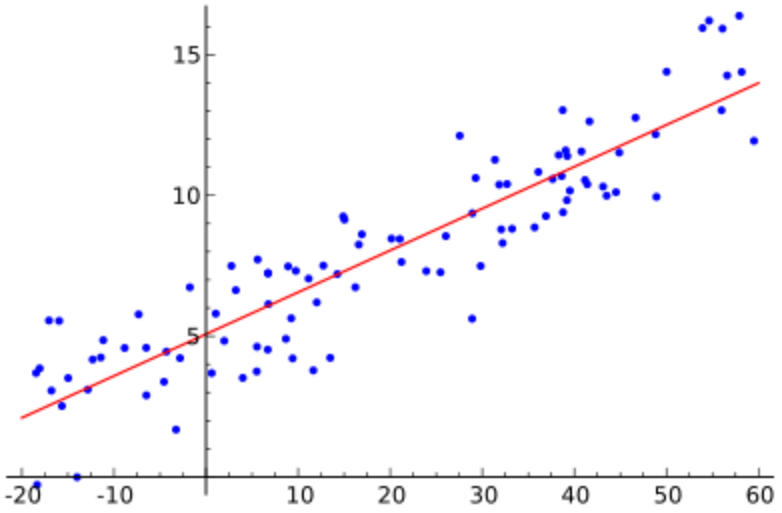
Alex Lyman

Classification via regression

- Classification: predicting the class of a given record
- Instead of predicting the class of a record we want to predict the probability of the class given the record
- The problem of predicting **continuous values** is called regression
- We will talk about regression in a couple of weeks
- General approach: find a **continuous function** that models the continuous points.

Example: Linear regression

- Given a dataset of the form $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ find a linear function that given the vector x_i predicts the y_i value as $y_i' = w^T x_i$
  - Find a vector of weights w that minimizes the sum of square errors
$$\sum_i (y_i' - y_i)^2$$
  - Several techniques for solving the problem.
  - We'll cover Linear Regression in two weeks when we cover Regression in general.

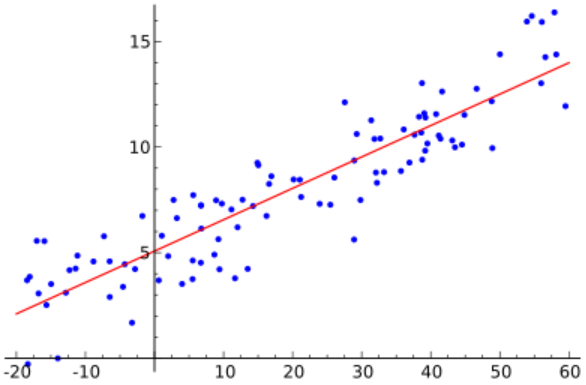


Example: Linear regression

- Data: $\{(x_1, y_1), (x_i, y_i), \dots, (x_n, y_n)\}$
- x_i : input value
- y_i : output value
- y_i' : predicted output value
- $y_i' = w^T x_i$
- w : weights

- We want a vector of weights w that minimizes the sum of square errors

$$\sum_i (y_i' - y_i)^2$$

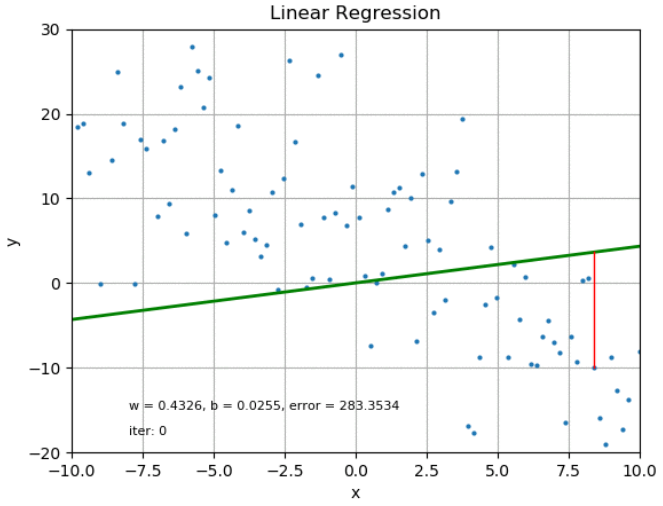


Example: Linear regression

- We want a vector of weights w that minimizes the sum of square errors

$$\sum_i (y_i' - y_i)^2$$

- We'll be doing linear regression in 3 weeks, so it's okay if you don't understand linear regression 100% right now.



What is a Vector of Weights?

Feature (x)	Value	Importance (w)	Calculation
Midterm	90	0.30	90 x 0.30 = 27
Homework	100	0.20	100 x 0.20 = 20
Final	80	0.50	80 x 0.50 = 40
Total Score			**27 + 20 + 40 = 87**

- Imagine you want to 'weight' different categories in a class.
- How would this look in Linear Algebra?
- **The Dot Product:** We store the "Values" in one list (vector x) and the "Importances" in another list (vector w)

$$x = \begin{bmatrix} 90 \\ 100 \\ 80 \end{bmatrix} \quad w = \begin{bmatrix} 0.3 \\ 0.2 \\ 0.5 \end{bmatrix} \qquad w^T x = w \cdot x$$

- To get the prediction, we multiply matching pairs and sum them up:

$$w \cdot x = (0.3 \times 90) + (0.2 \times 100) + (0.5 \times 80) = 87$$

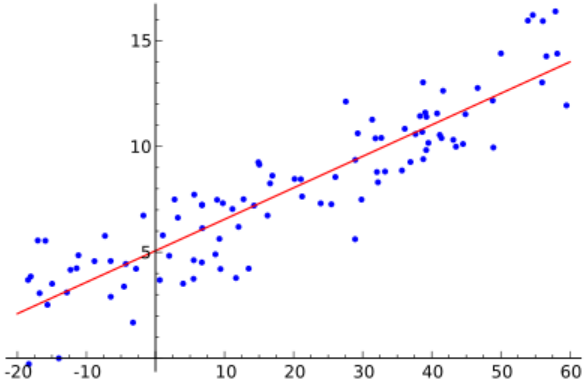
Simple Linear Regression (one feature)

$y = mx + b$

High school math

m = slope

b = intercept



$y = w_1 x + w_0$

Linear Regression

w_1= weight

w_0= bias

Classification via regression

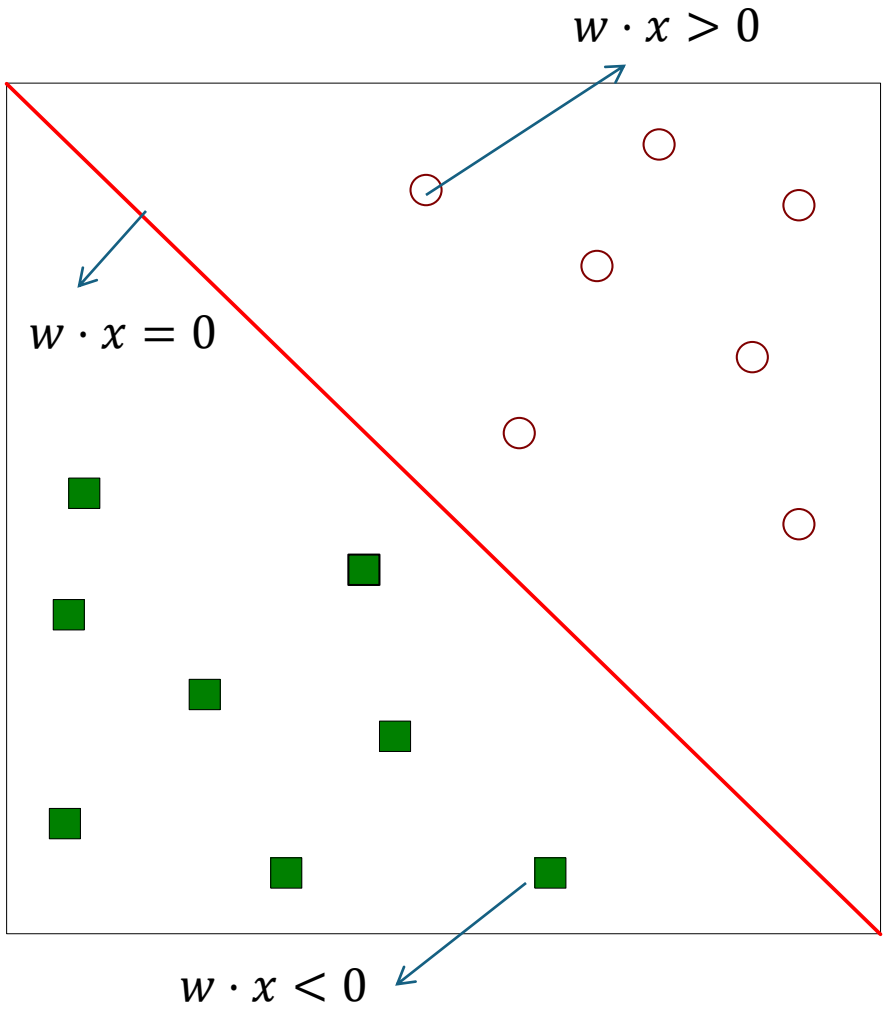
- Assume a linear classification boundary

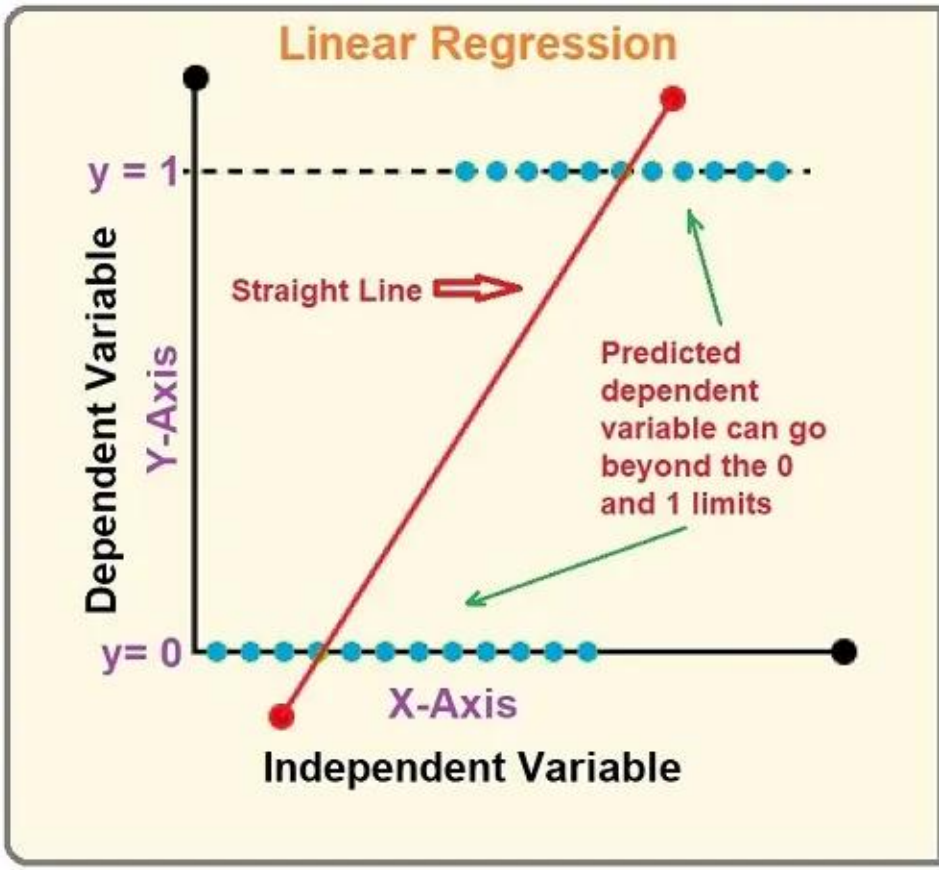
For the positive class the bigger the value of $w \cdot x$, the further the point is from the classification boundary, the higher our certainty for the membership to the positive class

- Define $P(C_+|x)$ as an increasing function of $w \cdot x$

For the negative class the smaller the value of $w \cdot x$, the further the point is from the classification boundary, the higher our certainty for the membership to the negative class

- Define $P(C_-|x)$ as a decreasing function of $w \cdot x$

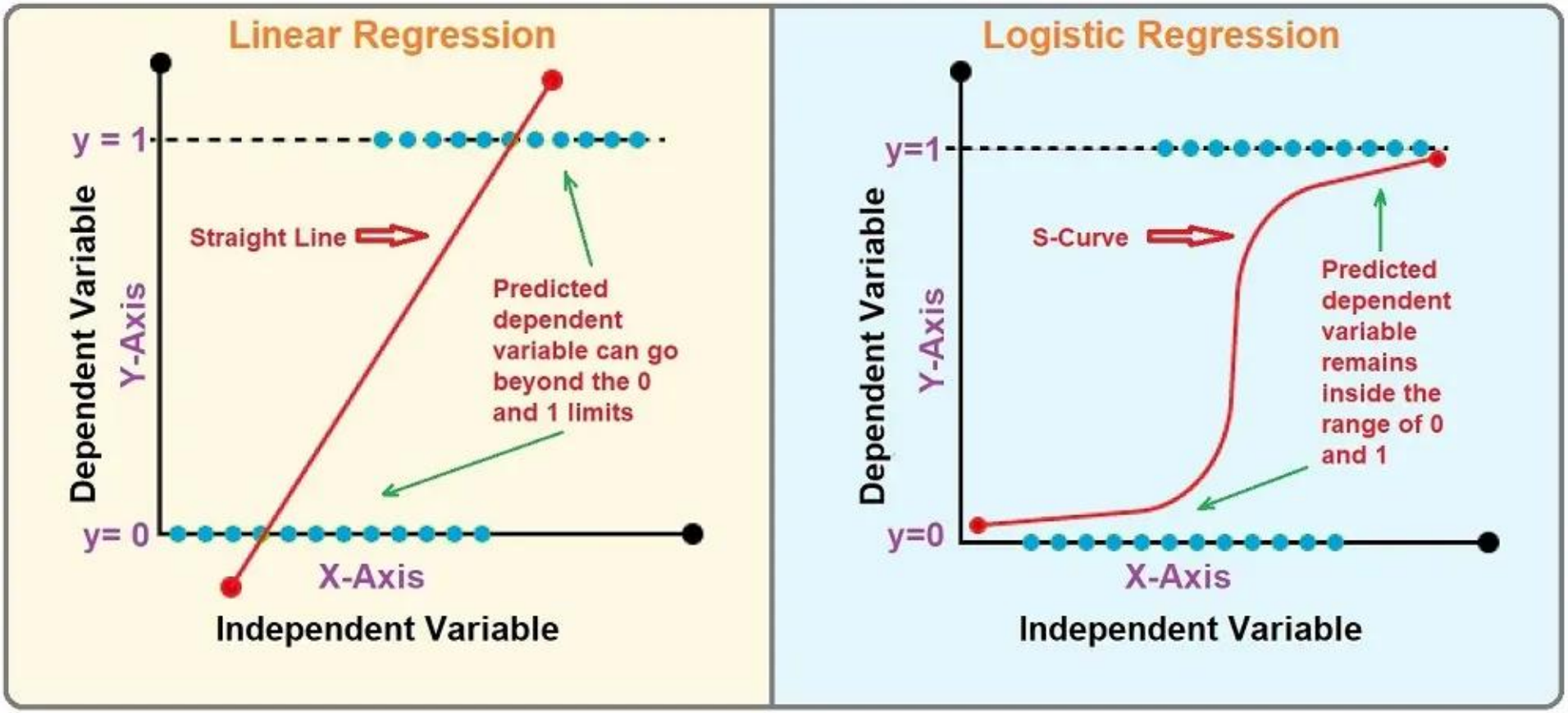




- Straight line doesn't work very well here.
- Is there something we can use instead of a linear function?
- Yes! We can use the Logistic (sigmoid) function.

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

- This squashes the line into a nice, (0,1) range.


Linear Regression
y = 1
y= 0
Dependent Variable
Y-Axis
Straight Line
Predicted dependent variable can go beyond the 0 and 1 limits
X-Axis
Independent Variable
Logistic Regression
y=1
y=0
Dependent Variable
Y-Axis
S-Curve
Predicted dependent variable remains inside the range of 0 and 1
X-Axis
Independent Variable

Logistic Regression

The logistic function (sigmoid)

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

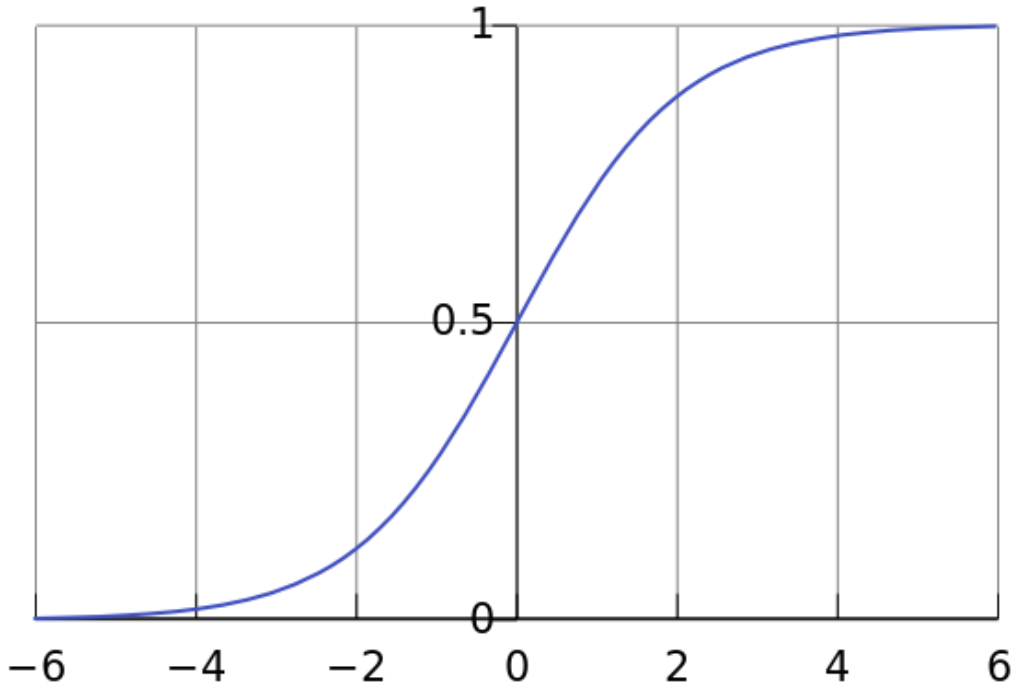
$$P(C_+|x) = \frac{1}{1 + e^{-w \cdot x}}$$

Probability of class membership

$$P(C_-|x) = \frac{e^{-w \cdot x}}{1 + e^{-w \cdot x}}$$

$$\log \frac{P(C_+|x)}{P(C_-|x)} = w^T x$$

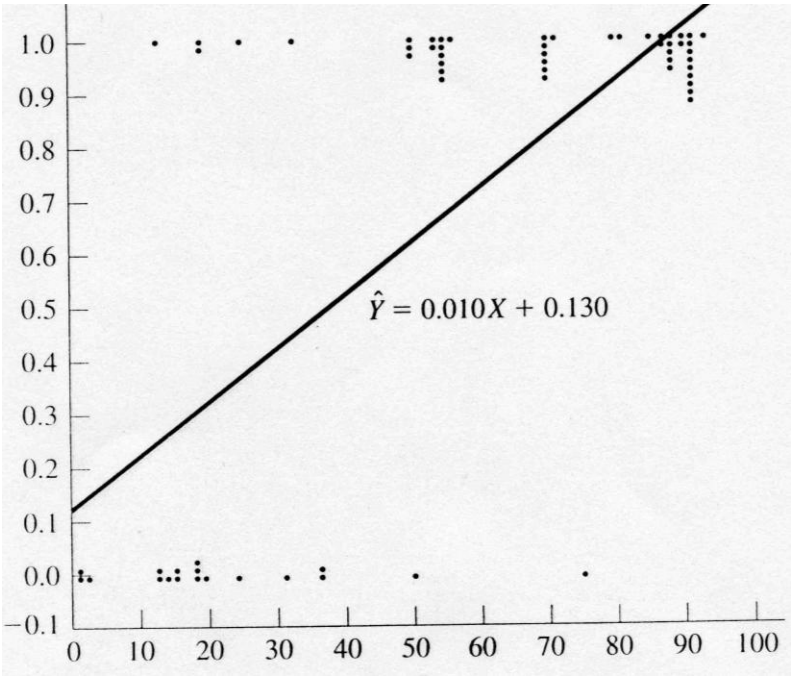
Log odds

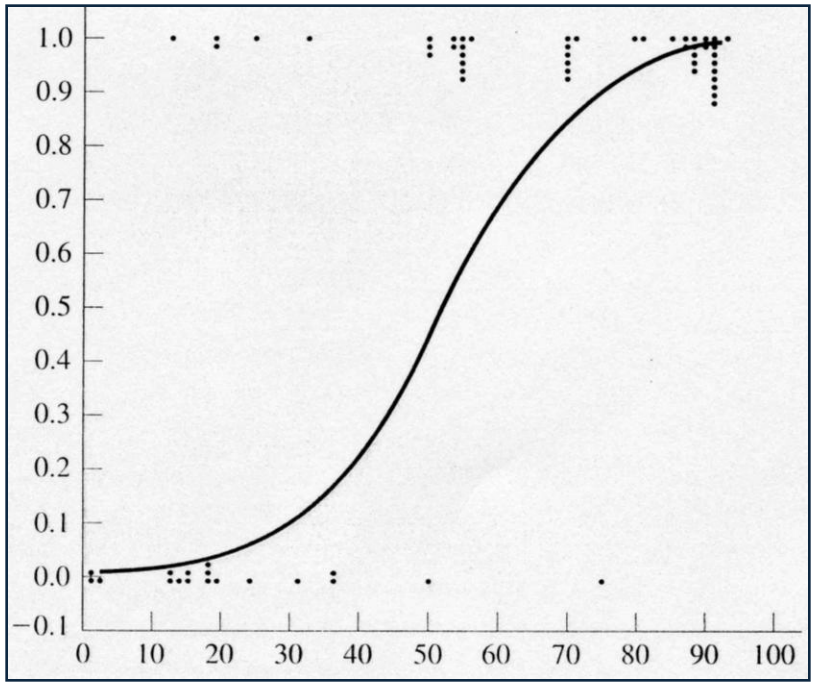


Logistic Regression: Find the vector w that maximizes the probability of the observed data

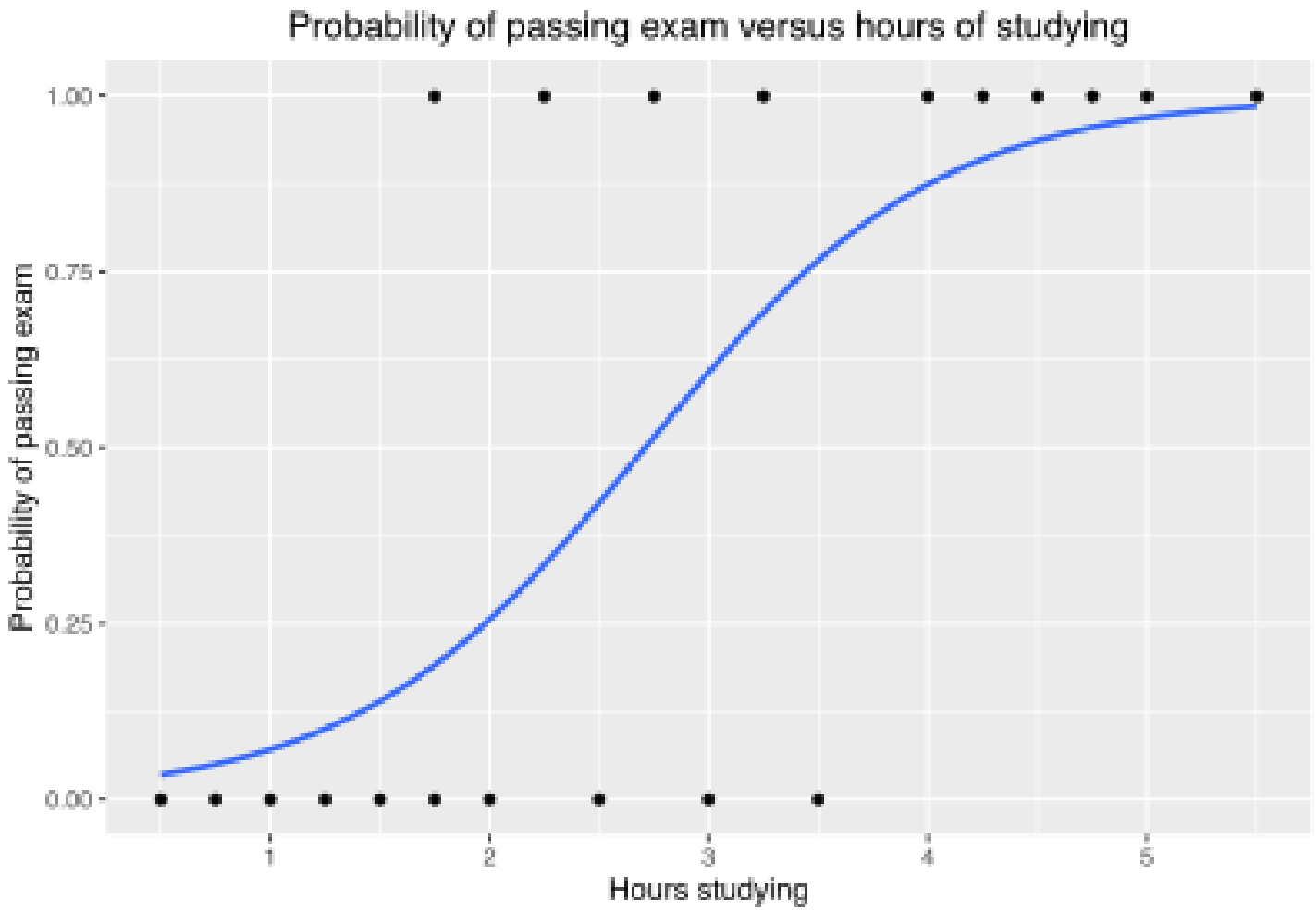
Logistic Regression Example

- Age (X axis, input variable) – Data is fictional
- Heart Failure (Y axis, 1 or 0, output variable)
- If use value of regression line as a probability approximation
  - Extrapolates outside 0-1 and not as good empirically
- Sigmoidal curve to the right gives empirically good probability approximation and is bounded between 0 and 1





Logistic Regression

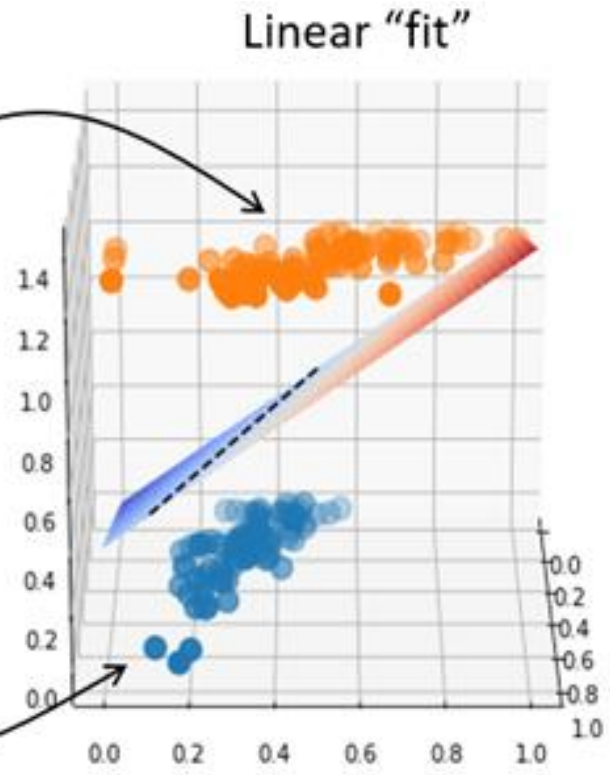


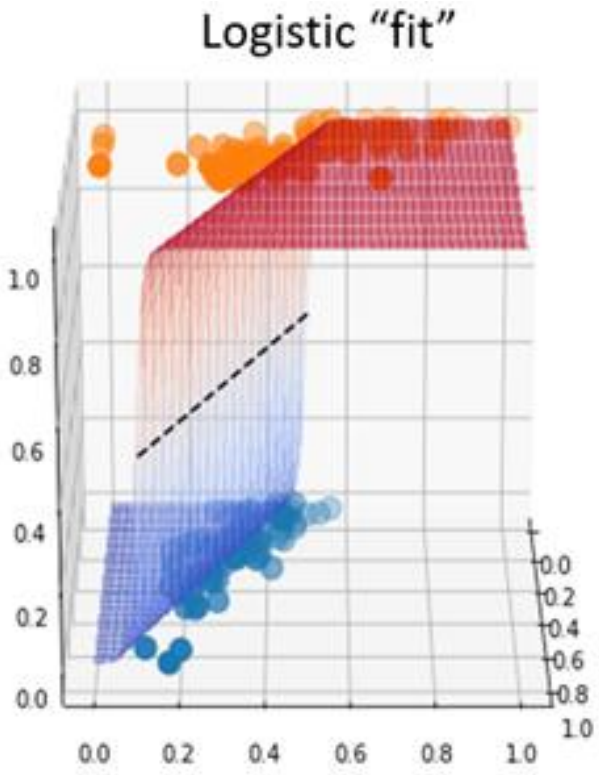
Logistic Regression

- Produces a probability estimate for the class membership which is often very useful.
- The weights can be useful for understanding the feature importance.
- Works for relatively large datasets
- Fast to apply.

BMWs
(*mileage*,*price*,1)

Priuses
(*mileage*,*price*,0)





Logistic Regression Approach

Learning

1. Transform initial input probabilities into log odds (logit)
2. Do a standard linear regression using the logit values
   - **This effectively fits a logistic curve to the data, while still just doing a linear regression with the transformed input** (ala quadric machine, etc.)

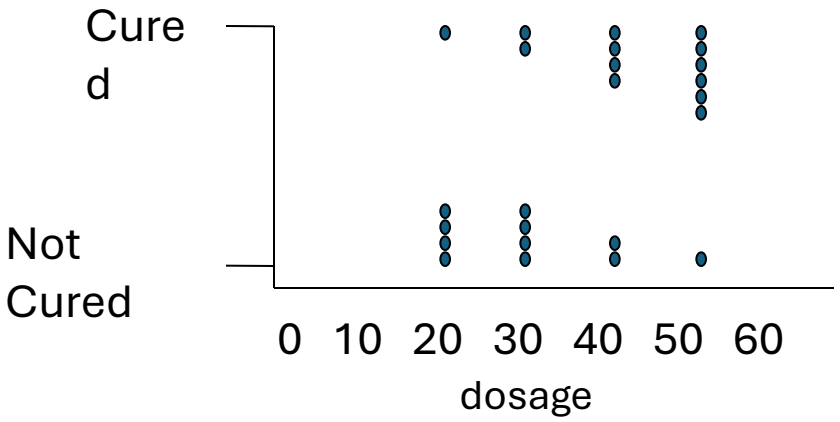
Generalization

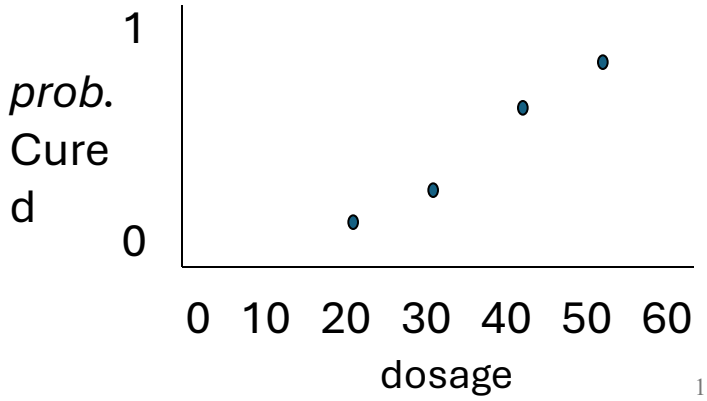
1. Find the value for the new input on the logit line
2. Transform that logit value back into a probability

Logistic Regression Approach

Medication Dosage	# Cured	Total Patients	Probability: # Cured/Total Patients
20	1	5	.20
30	2	6	.33
40	4	6	.67
50	6	7	.86

First convert to probabilities

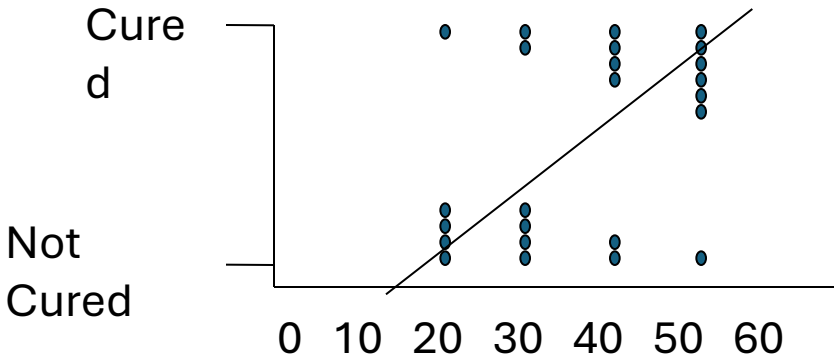


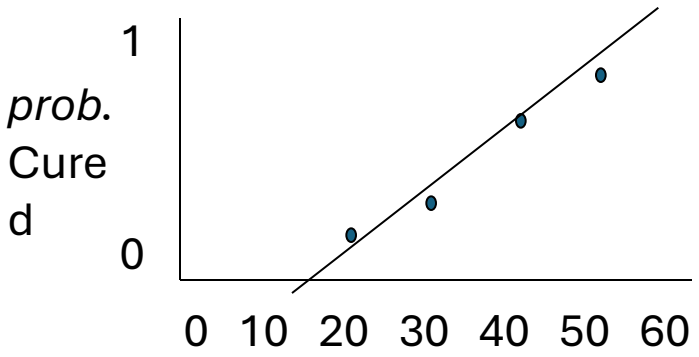


Logistic Regression Approach

Medication Dosage	# Cured	Total Patients	Probability: # Cured/Total Patients
20	1	5	.20
30	2	6	.33
40	4	6	.67
50	6	7	.86

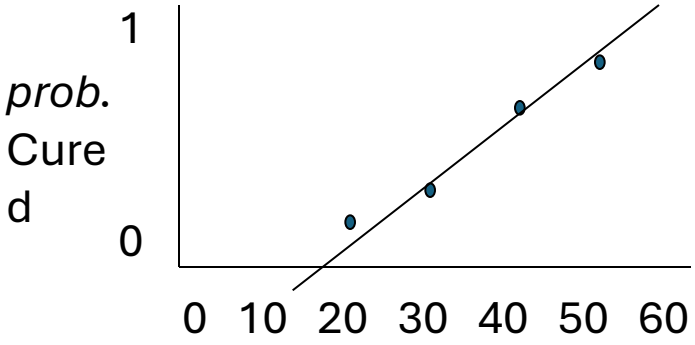
Could just do linear regression on probabilities, but…

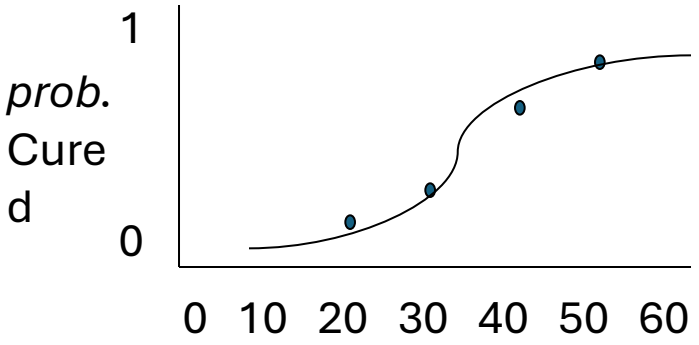




Logistic Regression Approach

- Could use linear regression with the probability points, but that would not extrapolate well
  - Generalized probabilities could be less than 0 or greater than 1
- Logistic version is better but how do we get it?
- Wdo a non-linear pre-process of the input and then do linear regression on the transformed values – do a linear regression on the log odds - Logit





Log Odds

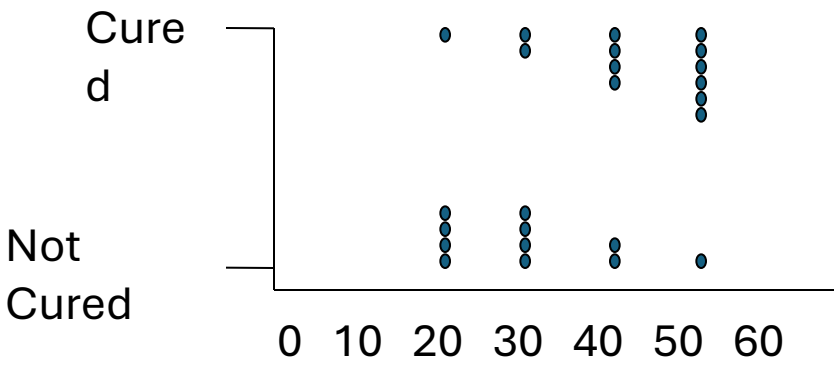
Log Odds

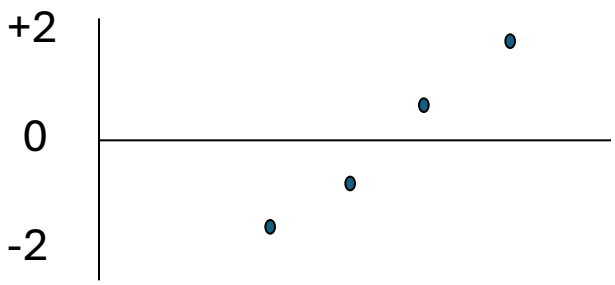
- The probability of an event occurring is a fraction of how many times you would expect to see that event occurring on a set number of occasions.
  - A probability is always expressed as a number between 0 and 1.
- Odds are expressed as the likelihood of an event occurring divided by the likelihood of it not occurring.
  - Generally a ratio (for example: 2 to 1 odds)
  - Odds = p / (1 – p)

- Logistic regression works in the odds space which converts the model from probabilities to likelihoods
- Then we take the logarithm so we don't have precision issues

Non-Linear Pre-Process to Logit (Log Odds)

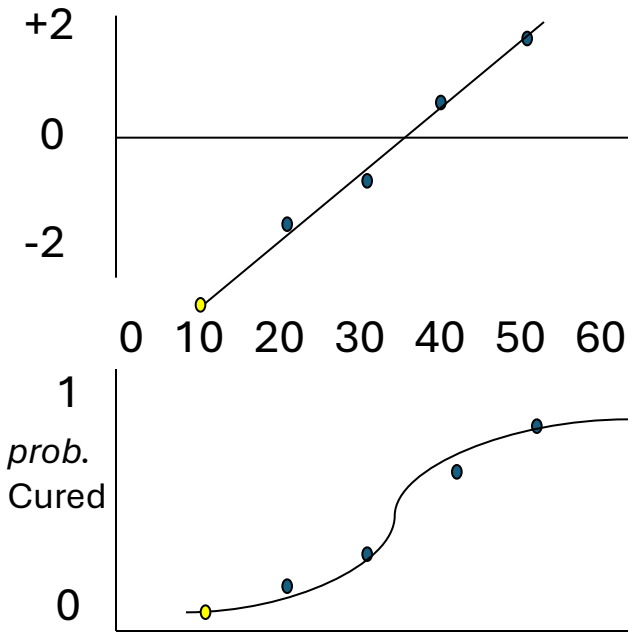
Medication Dosage	# Cured	Total Patients	Probability: # Cured/Total Patients	Odds: $p/(1-p)$ = # cured/ # not cured	Logit Log Odds: ln(Odds)
20	1	5	.20	.25	-1.39
30	2	6	.33	.50	-0.69
40	4	6	.67	2.0	0.69
50	6	7	.86	6.0	1.79





Regression of Log Odds

Medication Dosage	# Cured	Total Patients	Probability: # Cured/Total Patients	Odds: $p/(1-p)$ = # cured/ # not cured	Log Odds: ln(Odds)
20	1	5	.20	.25	-1.39
30	2	6	.33	.50	-0.69
40	4	6	.67	2.0	0.69
50	6	7	.86	6.0	1.79



- $y = .11x - 3.8$ - Logit regression equation
- Now we have a regression line for log odds (logit)
- To generalize, we use the log odds value for the new data point
- Then we transform that log odds point to a probability: $p = e^{\text{logit}(x)}/(1+e^{\text{logit}(x)})$
- For example assume we want p for dosage = 10

$$\text{Logit}(10) = .11(10) - 3.8 = -2.7$$

$p(10) = e^{-2.7}/(1+e^{-2.7}) = .06$ [note that we just work backwards from logit to p]

- These p values make up the sigmoidal regression curve (which we never have to actually plot)

Classification via regression

- Assume a linear classification boundary

For the positive class the bigger the value of $w \cdot x$, the further the point is from the classification boundary, the higher our certainty for the membership to the positive class

- Define $P(C_+|x)$ as an increasing function of $w \cdot x$

For the negative class the smaller the value of $w \cdot x$, the further the point is from the classification boundary, the higher our certainty for the membership to the negative class

- Define $P(C_-|x)$ as a decreasing function of $w \cdot x$