Classification & Logistic Reg. Basics

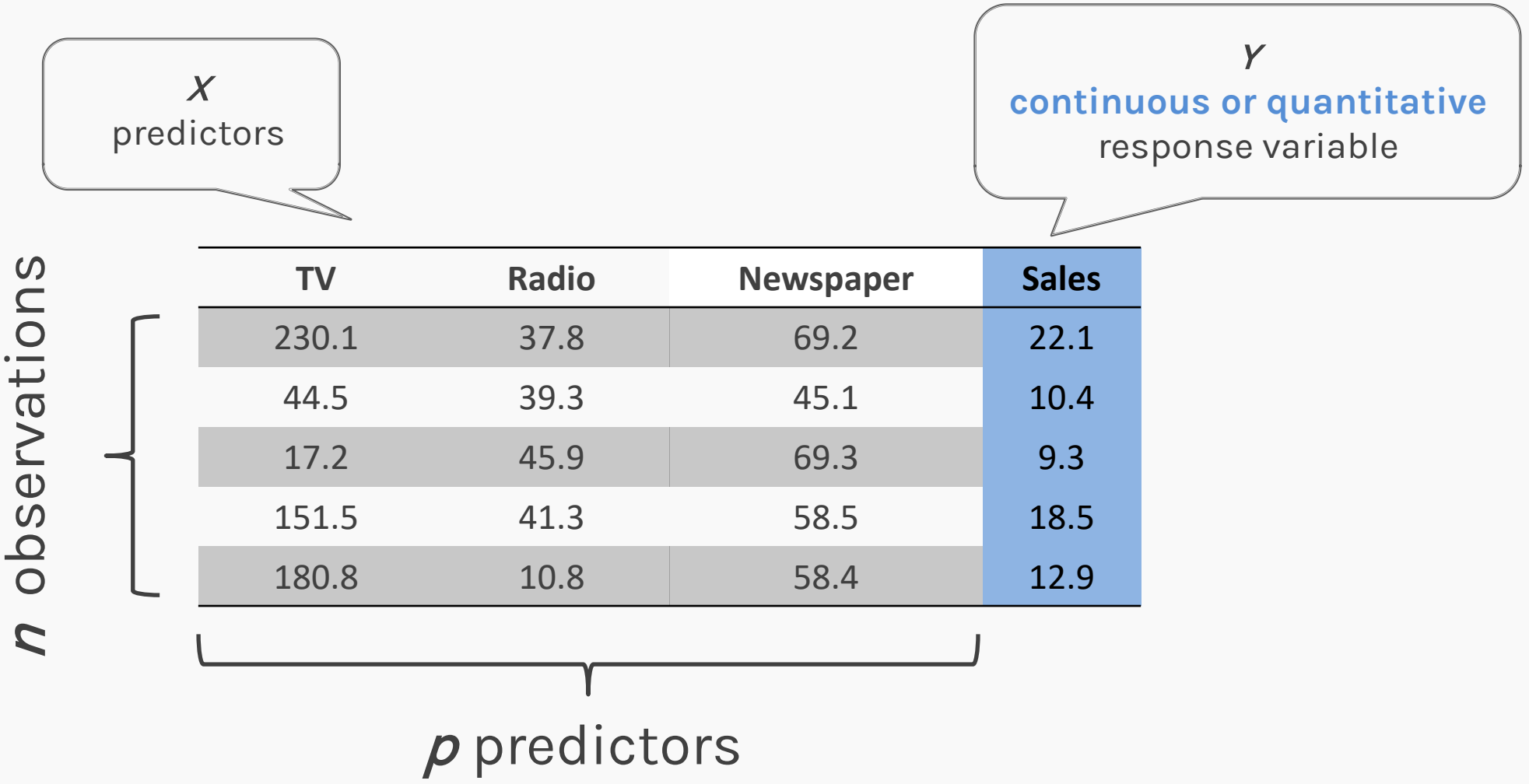
Pavlos Protopapas

Outline

- **What is Classification?**
- Why not Linear Regression?
- Estimating the Simple Logistic Model
- Inference in Logistic Regression
- Multiple Logistic Regression
- Classification Decision Boundaries

What is Classification?

Advertising Data (from earlier lectures)



TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

What is Classification?

Consider another dataset that contains a binary outcome AHD for 303 patients who presented with chest pain.

X
predictors

Y
Yes or No
response variable

Age	Sex	ChestPain	RestBP	Chol	MaxHR	ExAng	Thal	AHD
63	1	typical	145	233	150	0	fixed	No
67	1	asymptomatic	160	286	108	1	normal	Yes
67	1	asymptomatic	120	229	129	1	reversable	Yes
37	1	nonanginal	130	250	187	0	normal	No

What is Classification?

Consider another dataset that contains a binary outcome AHD for 303 patients who presented with chest pain.

X
predictors

Age	Sex	ChestPain	RestBP	Chol	MaxHR	ExAng	Thal	AHD
63	1	typical	145	233	150	0	fixed	No
67	1	asymptomatic	160	286	108	1	normal	Yes
67	1	asymptomatic	120	229	129	1	reversable	Yes
37	1	nonanginal	130	250	187	0	normal	No

'Yes' indicates **presence** of heart disease

What is Classification?

Consider another dataset that contains a binary outcome AHD for 303 patients who presented with chest pain.

X predictors

Age	Sex	ChestPain	RestBP	Chol	MaxHR	ExAng	Thal	AHD
63	1	typical	145	233	150	0	fixed	No
67	1	asymptomatic	160	286	108	1	normal	Yes
67	1	asymptomatic	120	229	129	1	reversable	Yes
37	1	nonanginal	130	250	187	0	normal	No

'No' indicates **absence** of heart disease

What is Classification?

In summary,

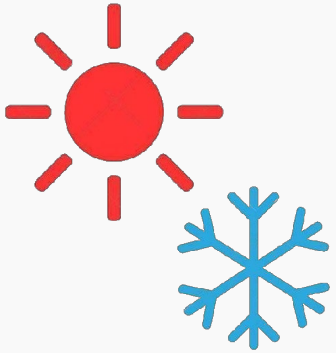
Regression

Performs well on tasks that require the prediction of a quantitative response variable.

What is the temperature going to be tomorrow?

Classification

Performs well on tasks that require the prediction of a categorical or qualitative response variable. It classifies an observation into a category or class labeled by Y.



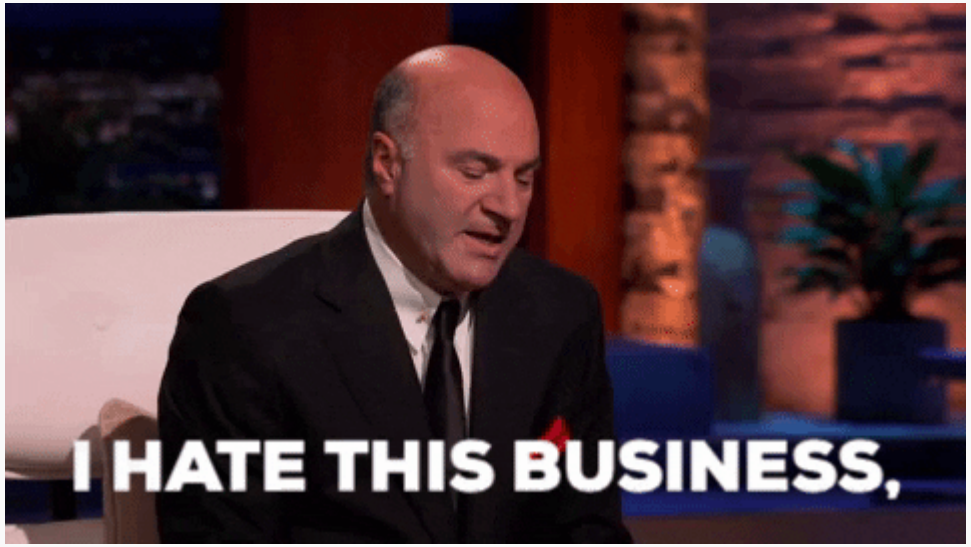
Is it going to be hot or cold tomorrow?

Typical Classification Examples

Classification problems are ubiquitous in many domains, such as healthcare, finance, sports.

Some examples of classification problems are:

- To determine whether a startup is worth investing in.



Typical Classification Examples

Classification problems are ubiquitous in many domains, such as healthcare, finance, sports.

Some examples of classification problems are:

- To determine whether a startup is worth investing in.
- To determine if a user is more likely to click on an advertisement.

Typical Classification Examples

Classification problems are ubiquitous in many domains, such as healthcare, finance, sports.

Some examples of classification problems are:

- To determine whether a startup is worth investing in.
- To determine if a user is more likely to click on an advertisement.
- To determine if a given image is a real or a fake one.

Outline

- What is Classification?
- **Why not Linear Regression?**
- Estimating the Simple Logistic Model
- Inference in Logistic Regression
- Multiple Logistic Regression
- Classification Decision Boundaries

Why not Linear Regression?

Assume you are given a dataset containing information of different students and your task is to predict whether their major is Computer Science, Statistics or otherwise.

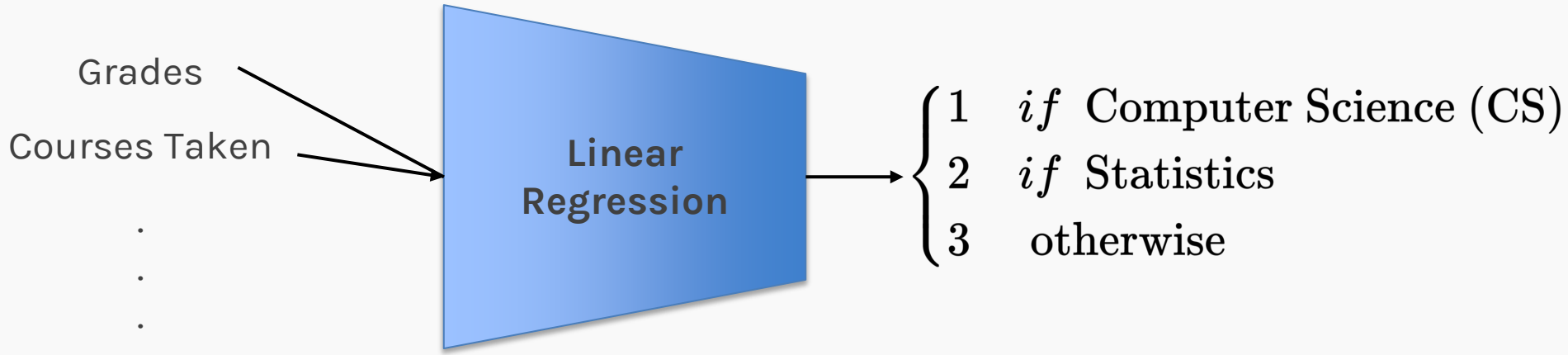
This categorical variable can't be used as is, but it could be encoded to be quantitative.

If y represents majors, then it could take on the values:

$$y = \begin{cases} 1 & if \ \text{Computer Science (CS)} \\ 2 & if \ \text{Statistics} \\ 3 & \text{otherwise} \end{cases}$$

Why not Linear Regression?

Now that we have encoded the values, a linear regression could be used to predict y from x.



But what is the problem here?

Why not Linear Regression?

This model would imply a specific ordering of the outcome.

For example, a change from $y = 1$ to $y = 2$ (Computer Science to Statistics) is considered the same as a change from $y = 2$ to $y = 3$ (Statistics to everyone else). However, this change should not be interpreted as the same.

Why not Linear Regression?

If a categorical response variable is **ordinal** (has a natural ordering, like Freshman, Sophomore, etc.), then a linear regression model would make some sense but is still not ideal.

Additionally, if the ordering of the response variable is changed, the model estimates and predictions would be fundamentally different.

For example, a model trained with $y = 1$ represents Statistics and $y = 2$ represents CS is different from a model trained with the original ordering.

Why not Linear Regression?

- Consider a simpler problem where the response variable y has only two categories. Here, there is a natural ordering of the categories.

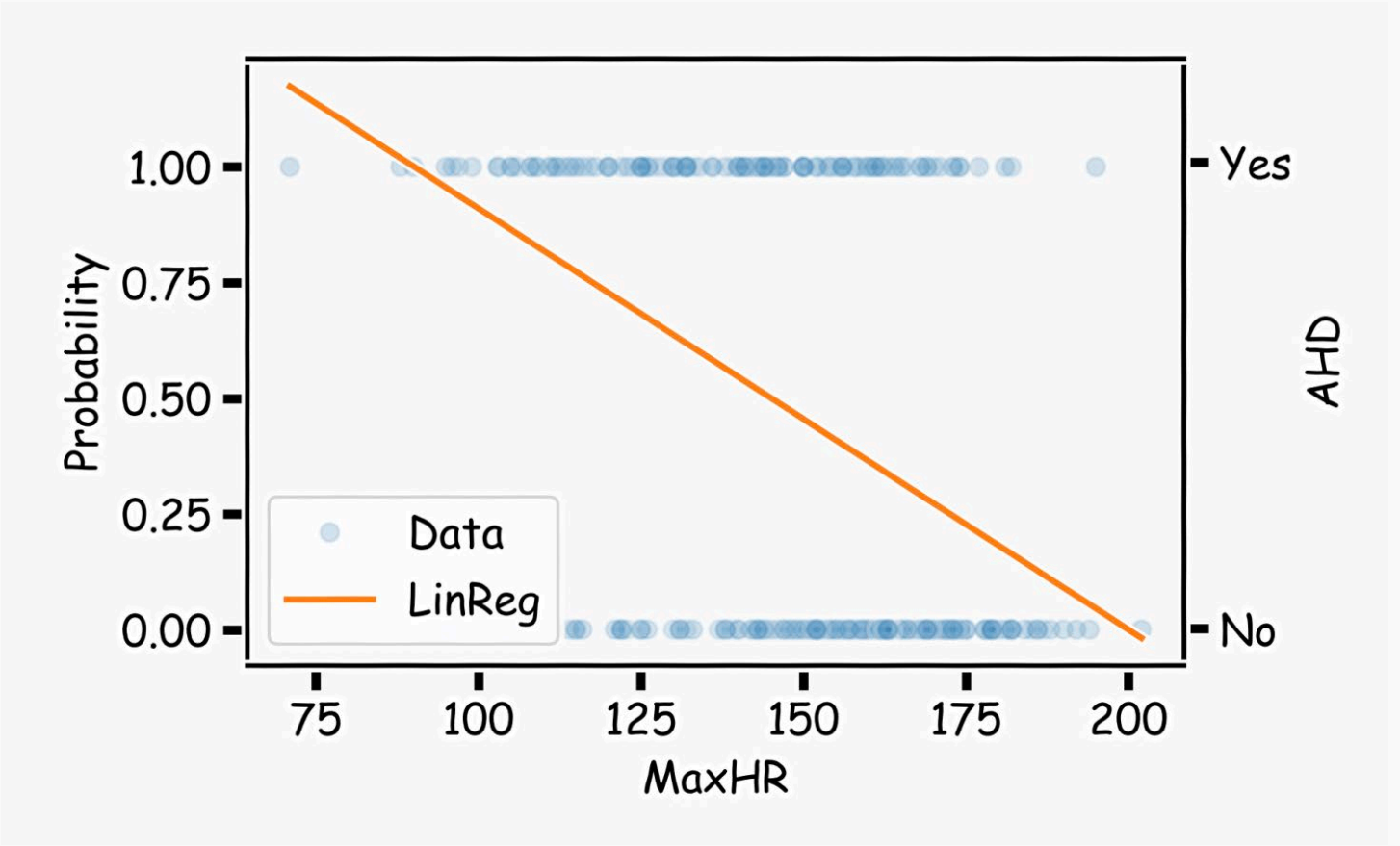
$$y = \begin{cases} 1, & has\ heart\ disease \\ 0, & no\ heart\ disease \end{cases}$$

- Linear regression could be used to predict the probability $P(y = 1)$ directly from a set of predictors such as sex, cholesterol levels, etc.

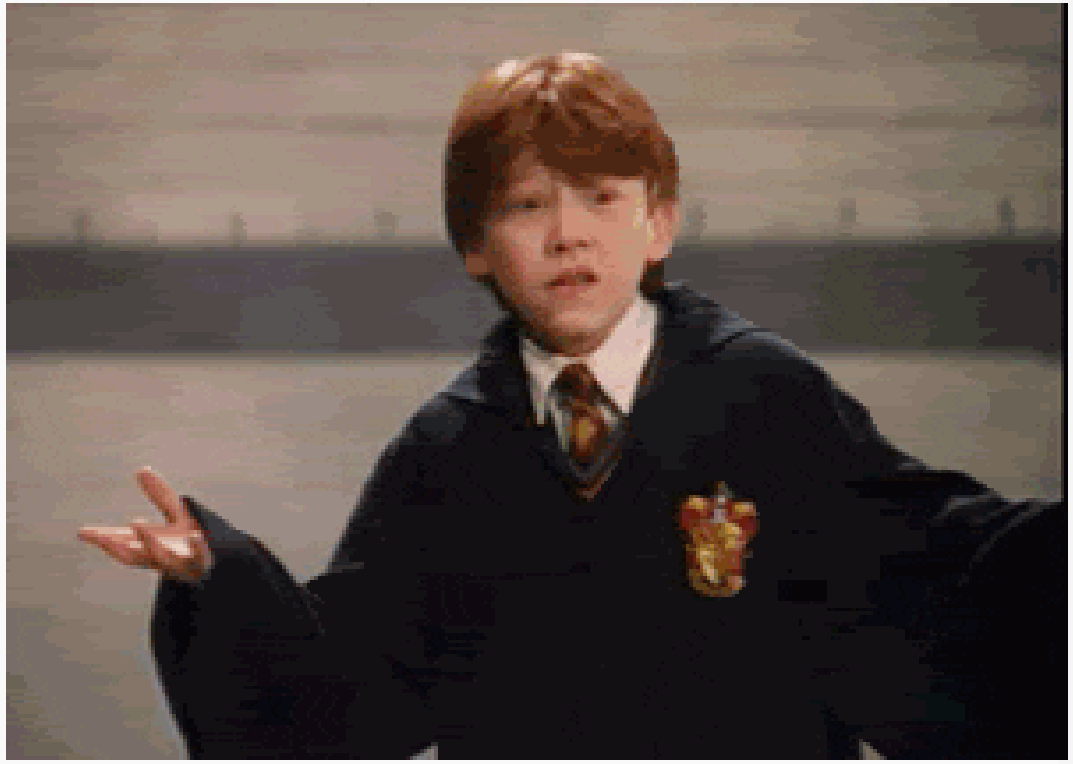
- If $P(y = 1) \geq 0.5$, we could predict that the patient has heart disease and predict otherwise if $P(y = 1) < 0.5$.

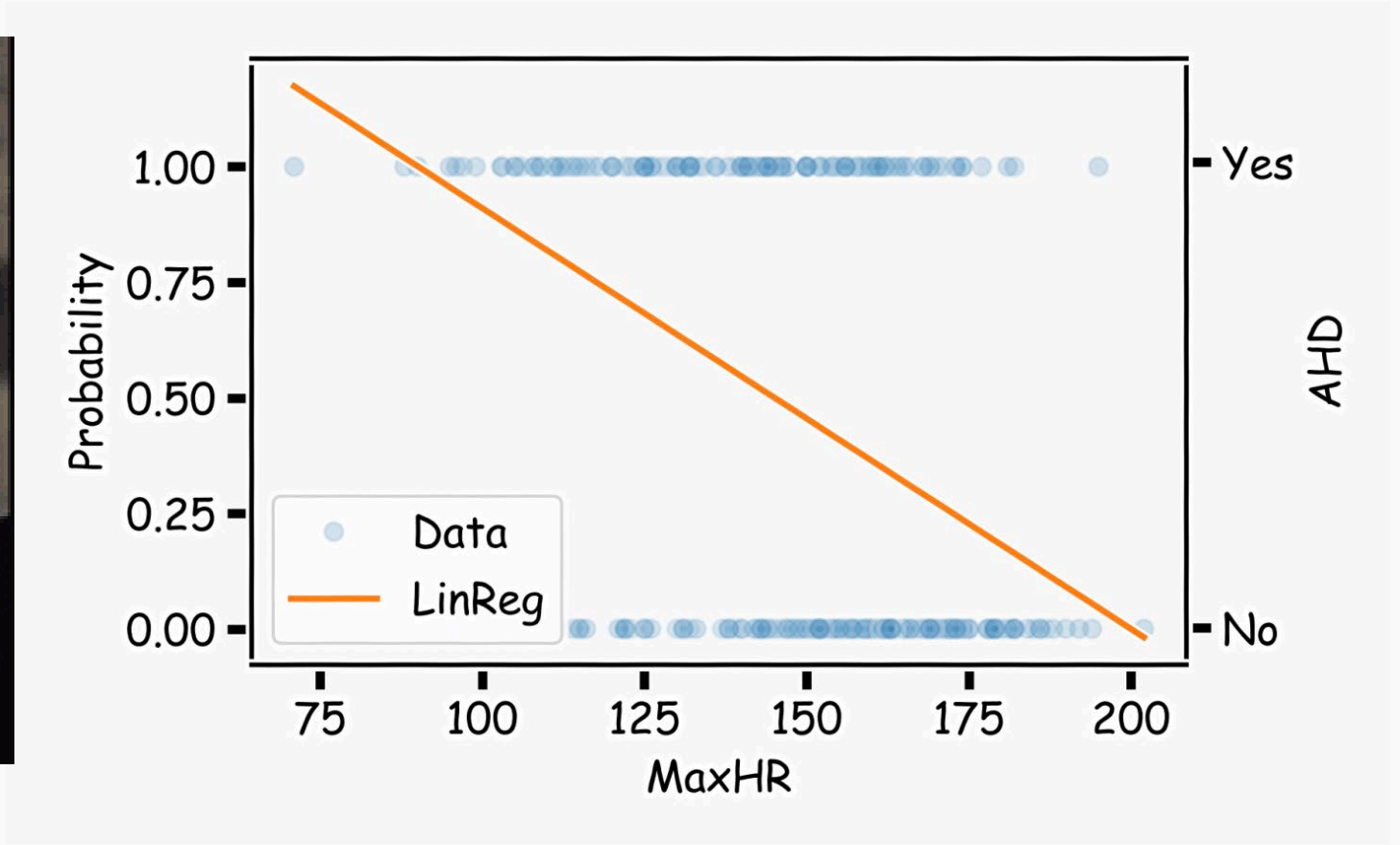
Why not Linear Regression?

What could go wrong with this linear regression model?



Why not Linear Regression?





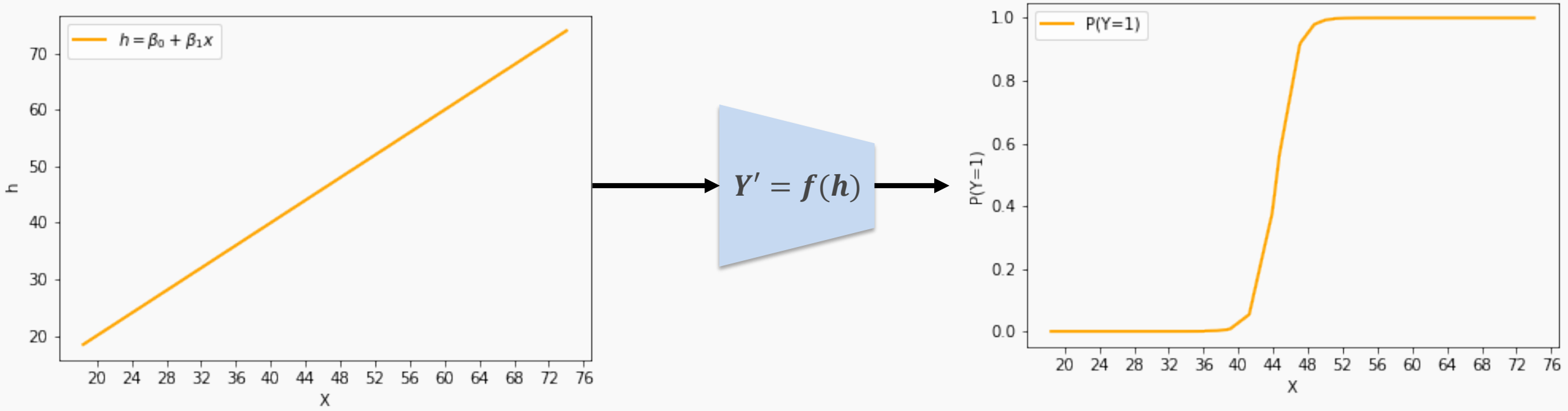
Since this is modeling $P(y = 1)$, values for $\hat{y}$ below 0 and above 1 would not make sense as a probability.

Outline

- What is Classification?
- Why not Linear Regression?
- **Estimating the Simple Logistic Model**
- Inference in Logistic Regression
- Multiple Logistic Regression
- Classification Decision Boundaries

What function should we use?

Now we know that linear regression yields values for probability that are larger than 1 or smaller than 0. So, what can we do to fix this?



What function should we use?

We can use the sigmoid function:

$$h = \beta_0 + \beta_1 X \longrightarrow \mathrm{p} = \frac{1}{1 + e^{-h}} \longrightarrow P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Logistic Regression

- Logistic Regression addresses the problem of estimating a probability, $P(y = 1)$, to be outside the range of $[0,1]$.

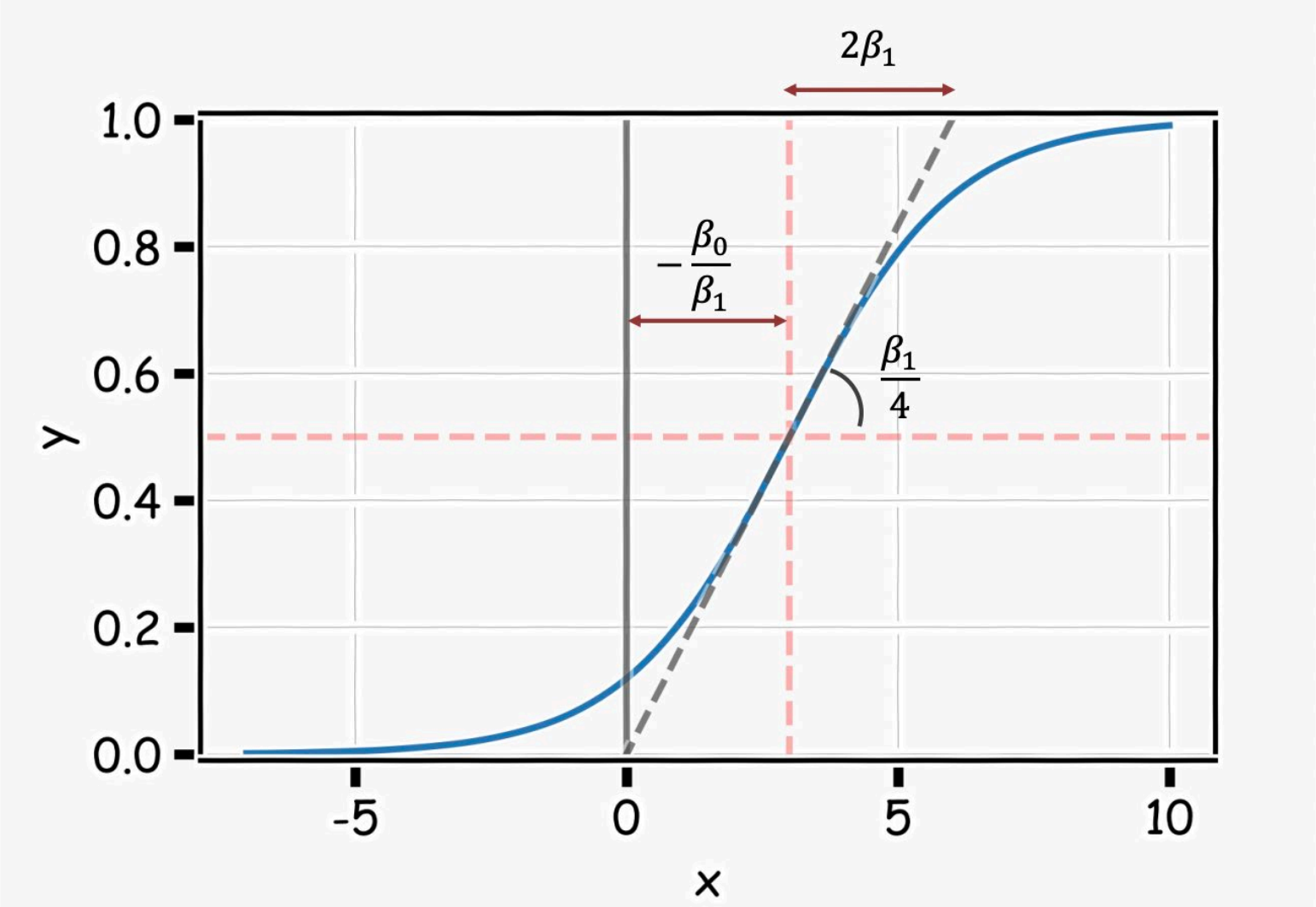
- The logistic regression model uses a function, called the **logistic** function, to model $P(y = 1)$:

$$P(Y = 1) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 X)}}$$

What are the parameters of this model?

Logistic Regression

The coefficients β_0 and β_1 now control the shape of this *S*-shaped curve.



Sigmoid Animation

Interpretation of β's

With a little bit of algebraic work, the logistic model can be rewritten as:

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X$$

$\frac{P(Y=1)}{1-P(Y=1)}$: ***odds***

Logistic regression is said to model the ***log-odds*** with a linear function of the predictors or features, X.

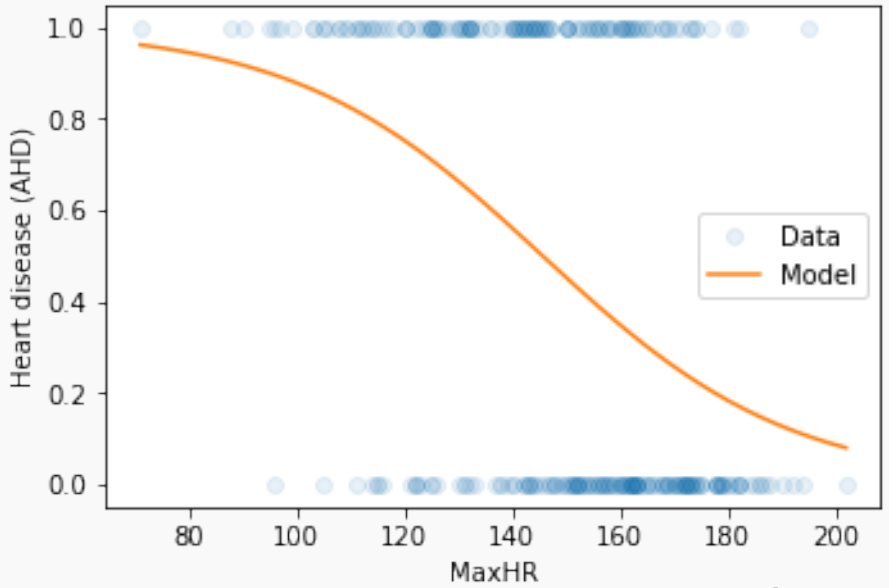
A one-unit change in X is associated with a β_1 change in the log-odds of $\mathrm{P}(Y=1)$; or better yet, a one-unit change in X is associated with an e^{β_1} change in the odds that $Y=1$.

A First Logistic Regression Model in sklearn

Here is a logistic regression output to predict Y = AHD from X = MaxHR:

```
logreg = LogisticRegression(penalty='none')
logreg.fit(df_heart[['MaxHR']], df_heart['AHD'])

print('Estimated beta1: \n', logreg.coef_)
print('Estimated beta0: \n', logreg.intercept_)
```

```
Estimated beta1:
 [[-0.04341112]]
Estimated beta0:
 [6.3249492]
```

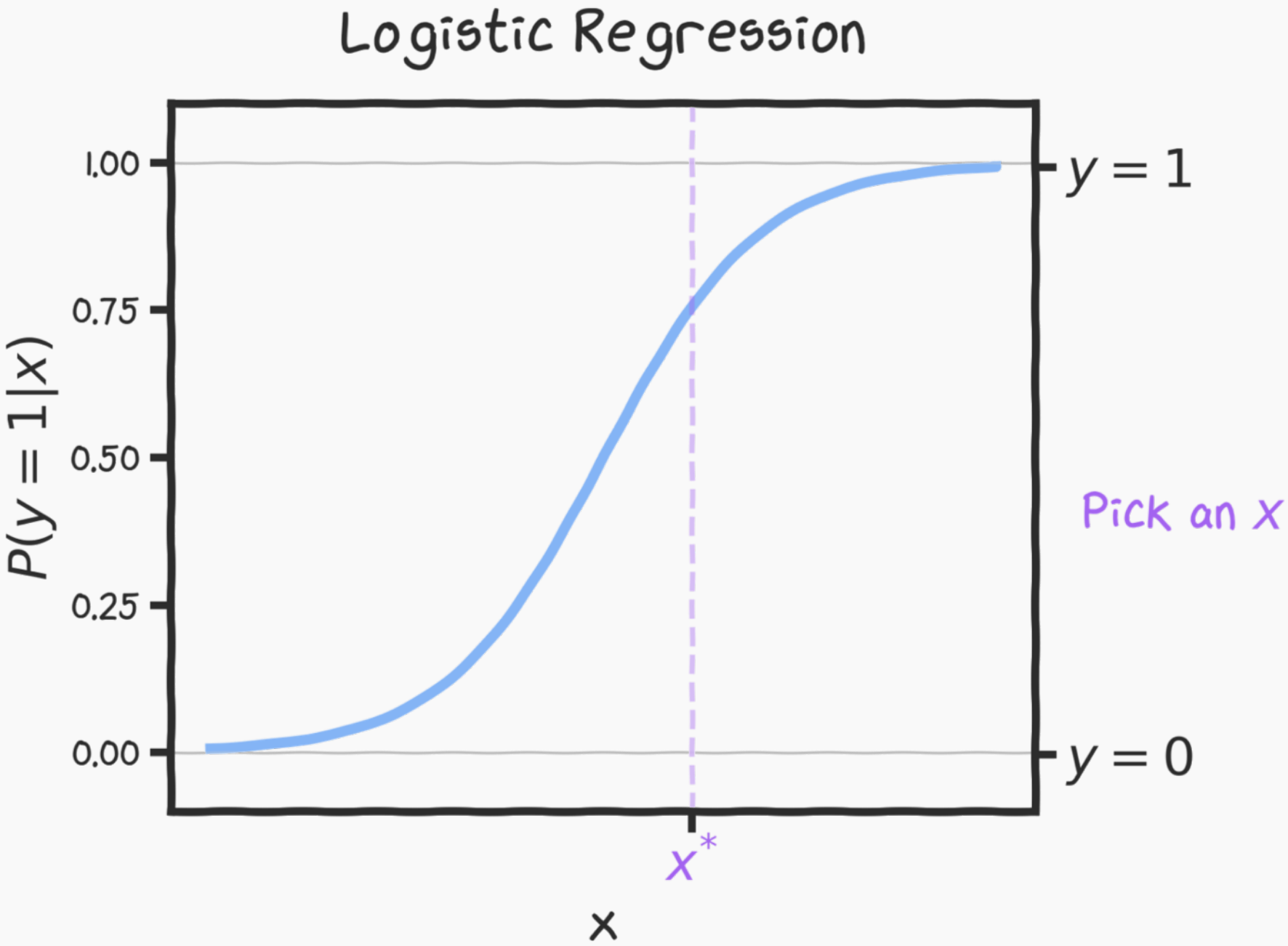


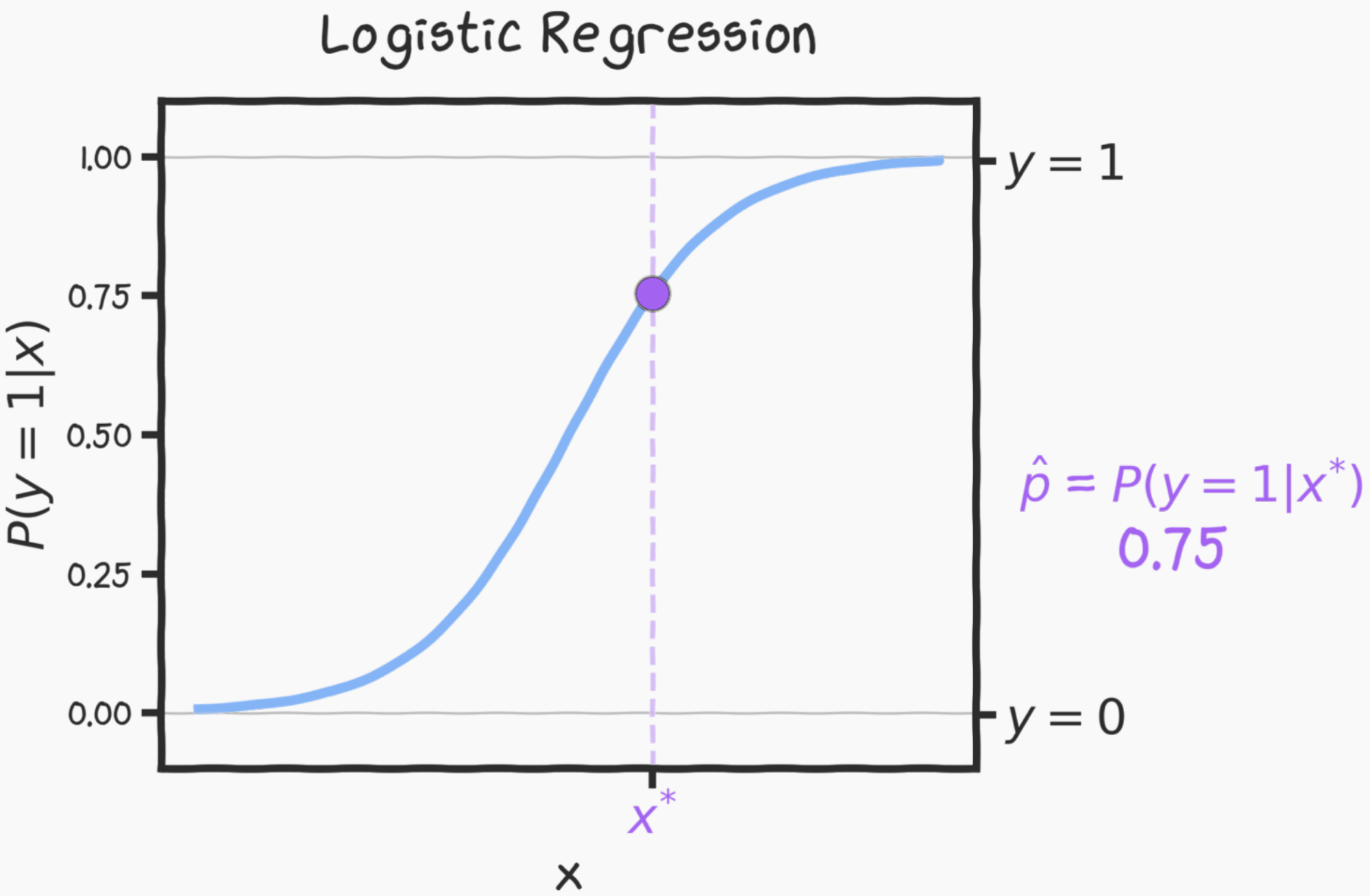
What is the estimated model? What are the interpretations of the $\hat{\beta}$s?

$$\ln\left(\frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)}\right) = 6.325 - 0.0434(MaxHR)$$

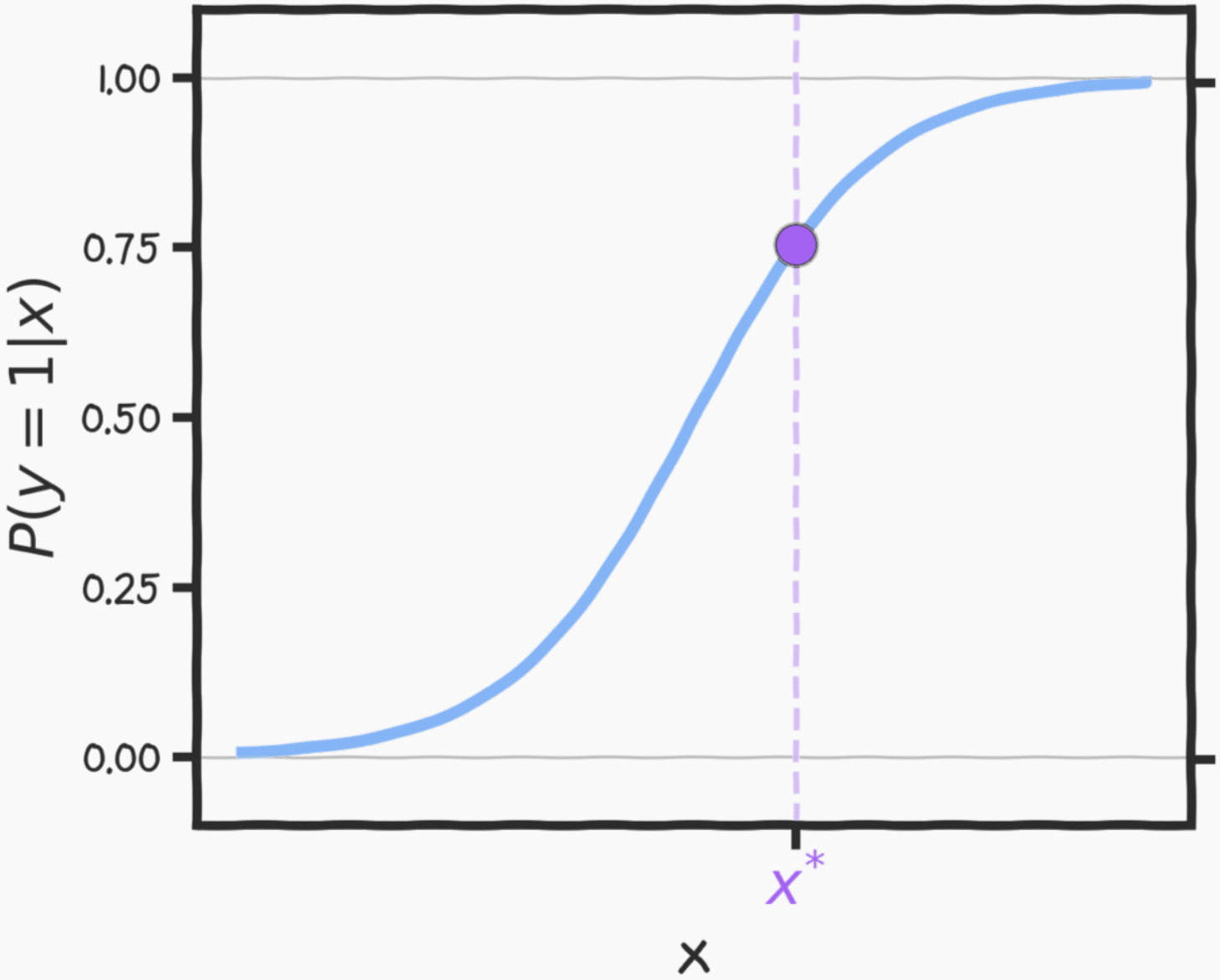
Estimating the Simple Logistic Model


Logistic Regression
1.00
0.75
0.50
0.25
0.00
$P(y = 1|x)$
$y = 1$
$y = 0$
x


Logistic Regression
1.00
0.75
0.50
0.25
0.00
$P(y=1|x)$
$y=1$
$y=0$
Pick an x
x^*
x


Logistic Regression
1.00
0.75
0.50
0.25
0.00
$P(y=1|x)$
$y=1$
$y=0$
x^*
x
$\hat{p} = P(y=1|x^*)$
0.75

Logistic Regression

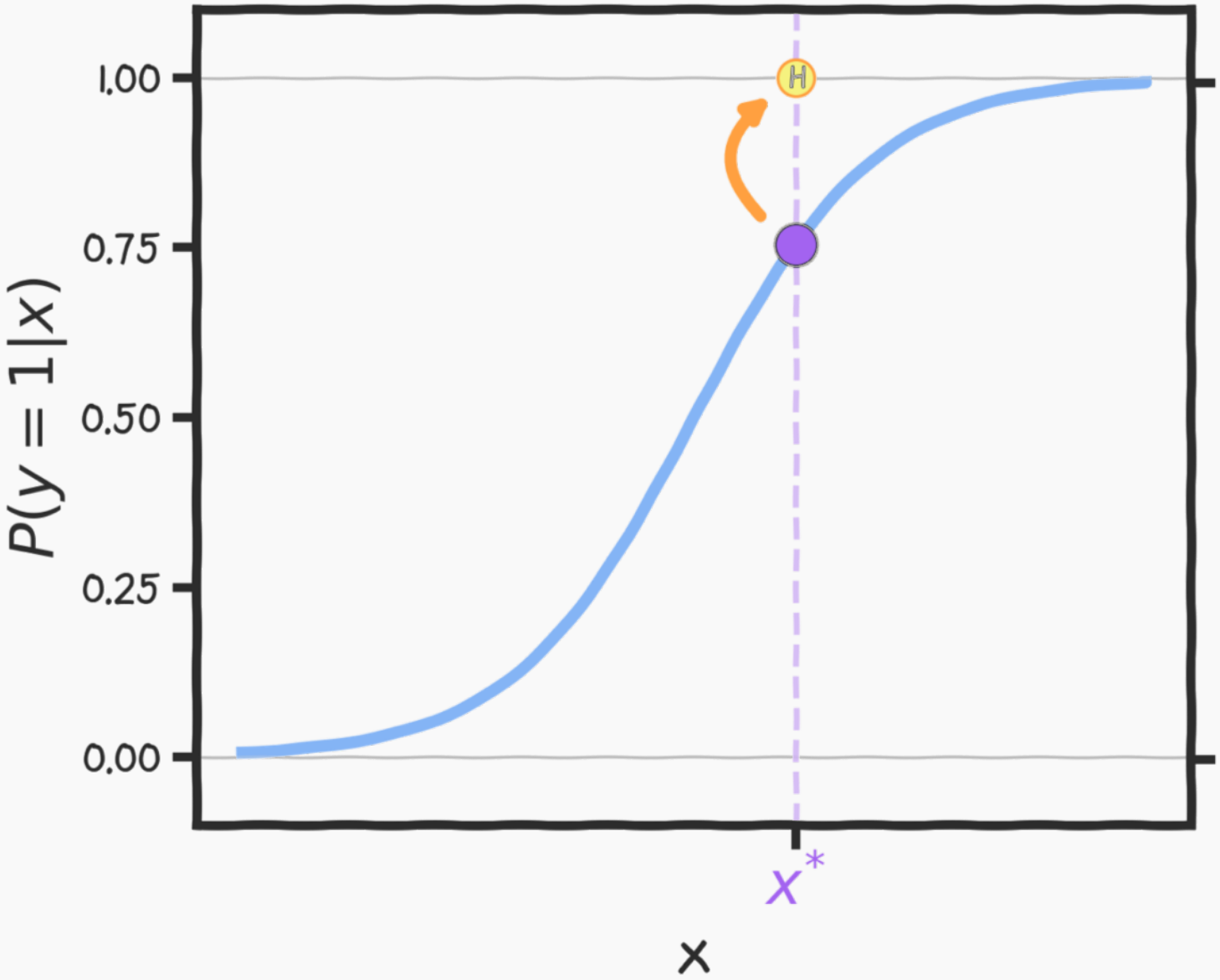


Flip a Coin with a
$\hat{P}$ Probability of Heads

$\hat{p} = P(y=1|x^*)$

0.75

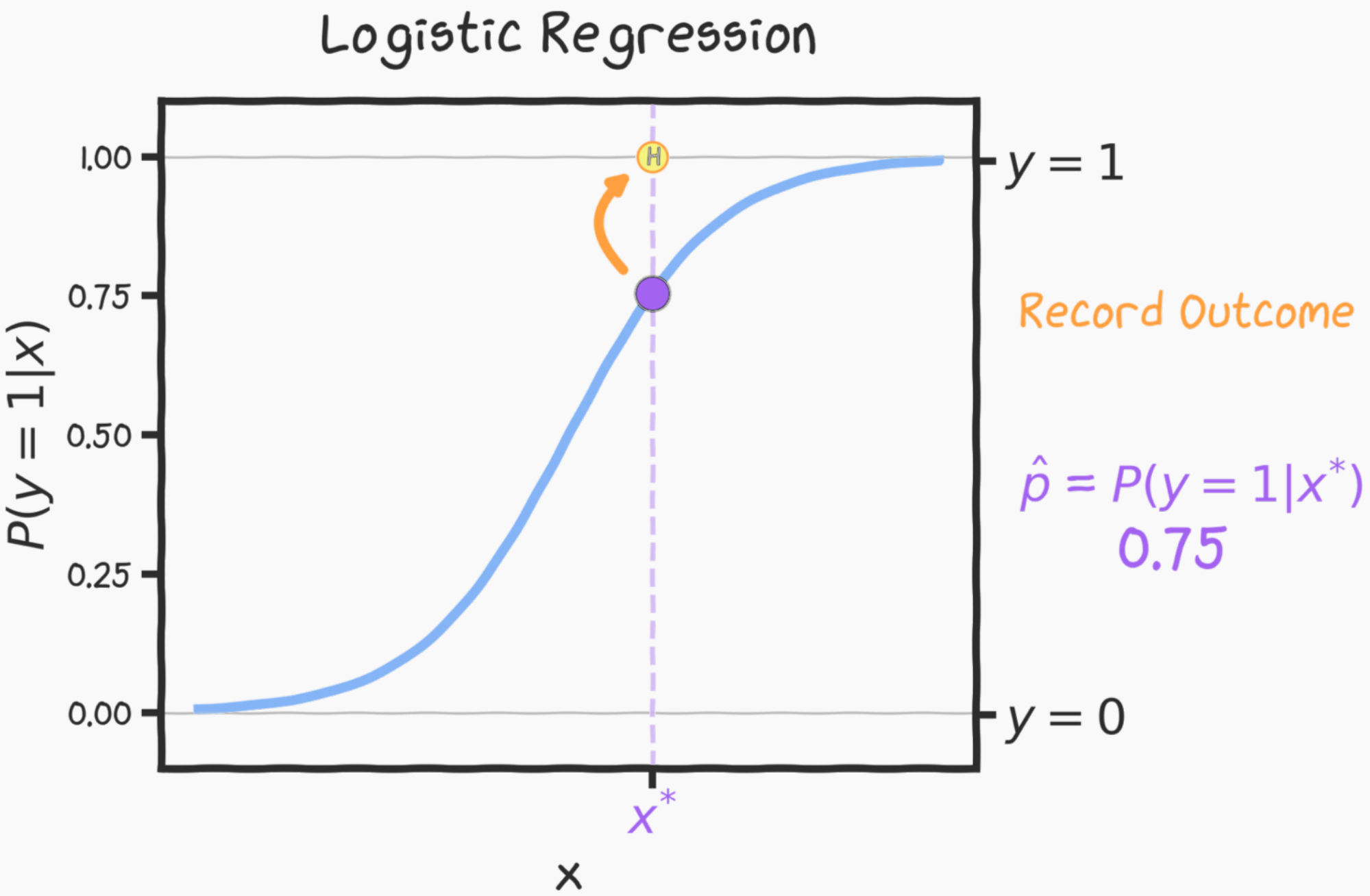
Logistic Regression

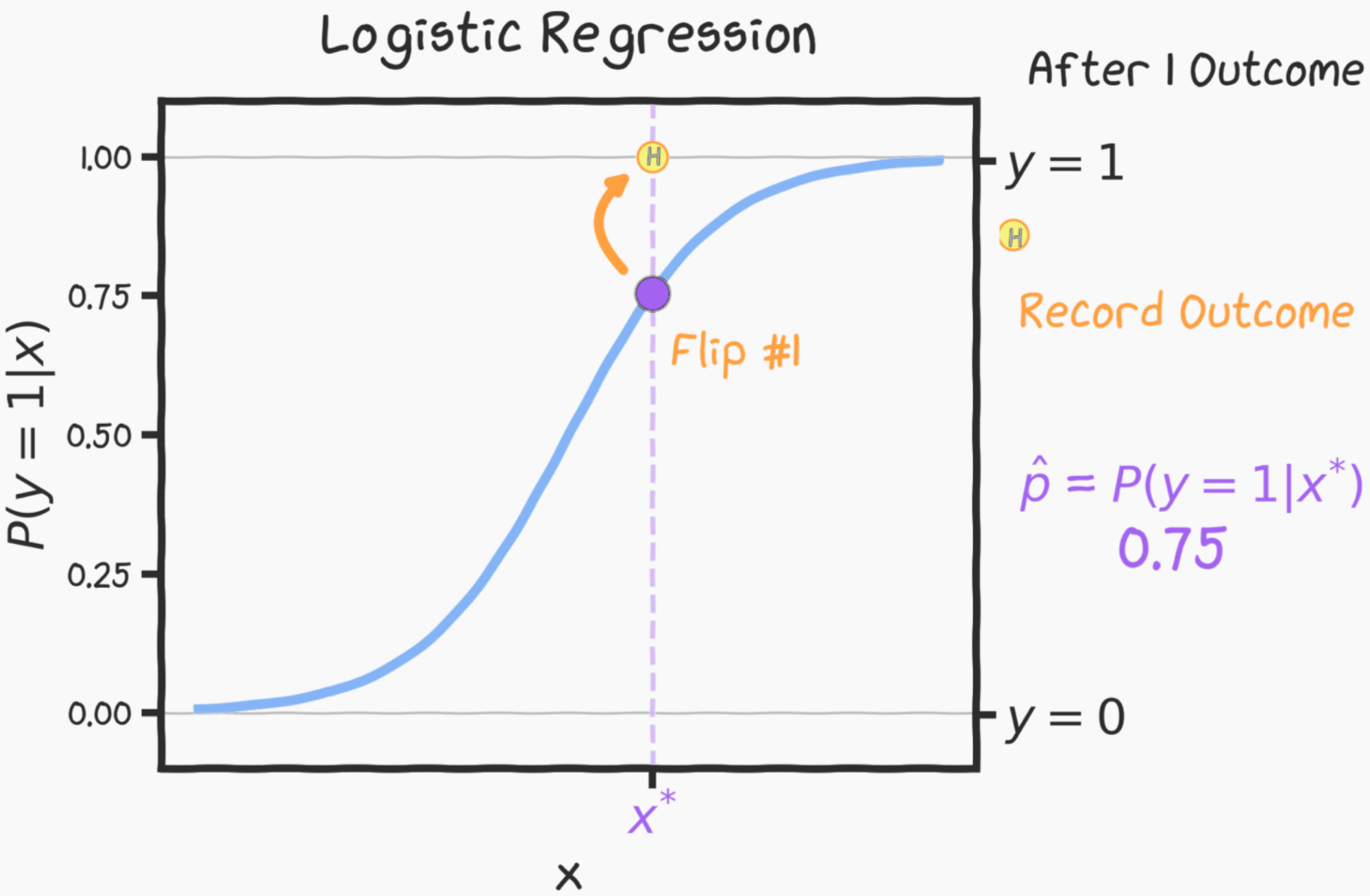


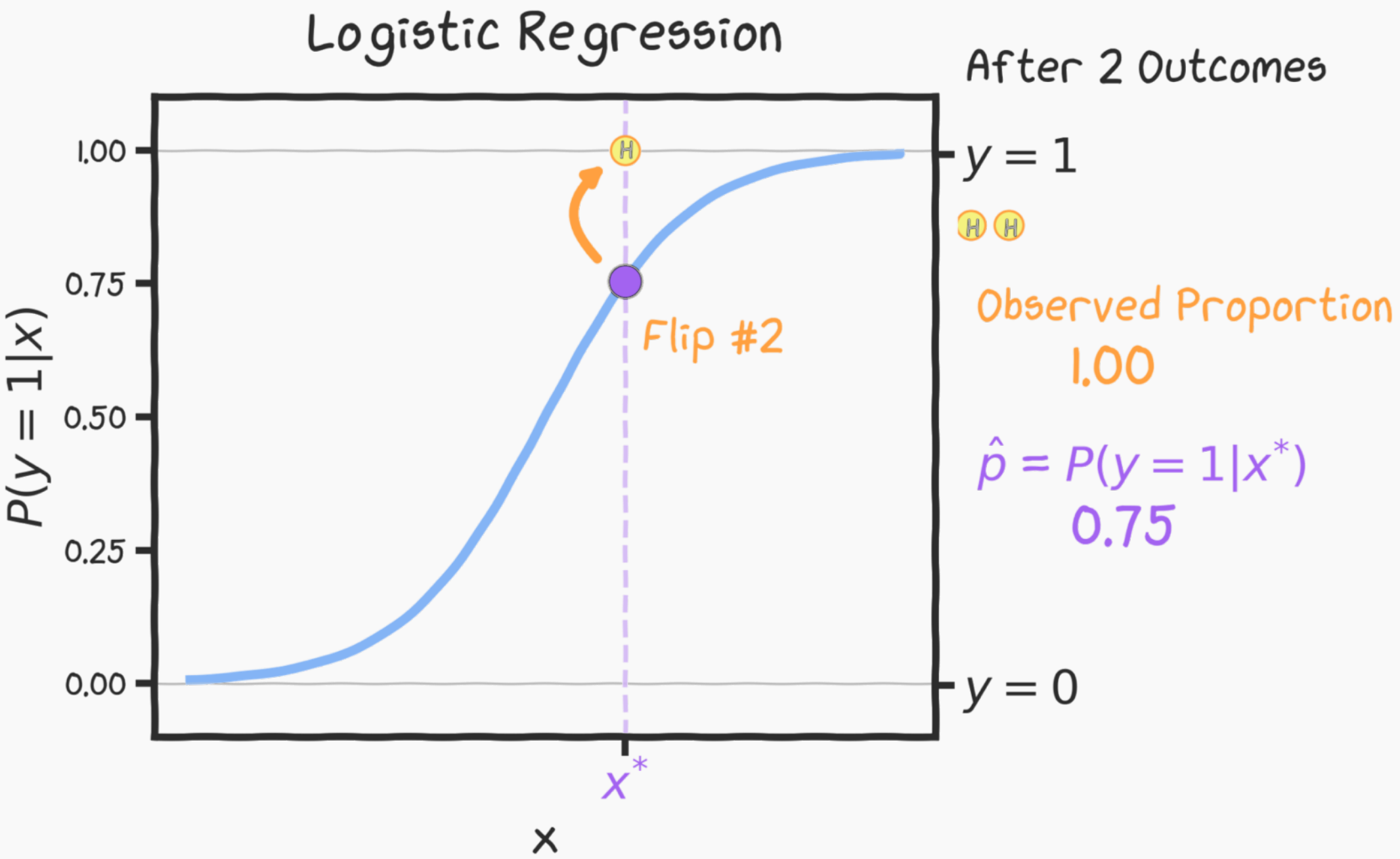
Flip a Coin with a
$\hat{P}$ Probability of Heads

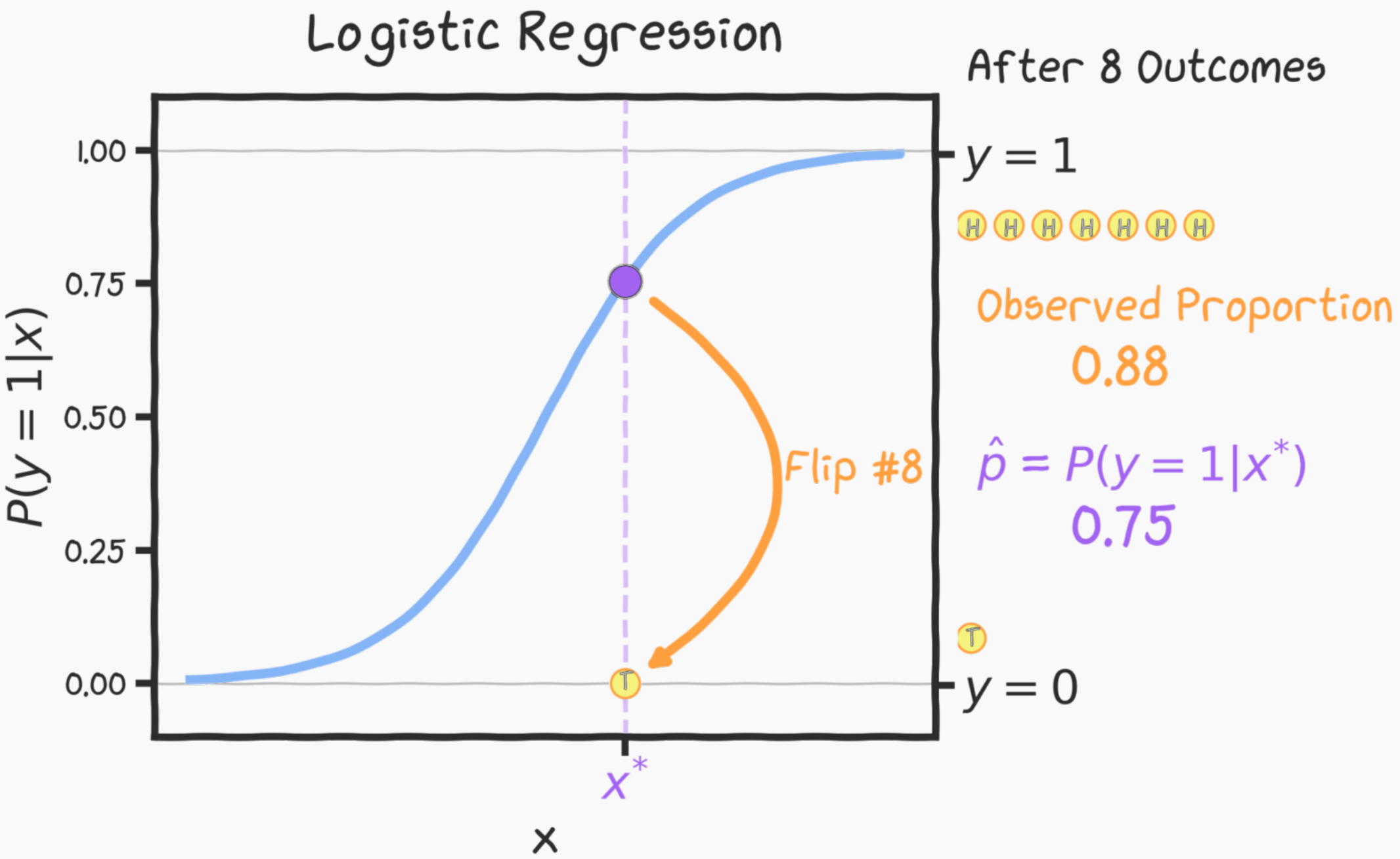
$\hat{p} = P(y=1|x^*)$

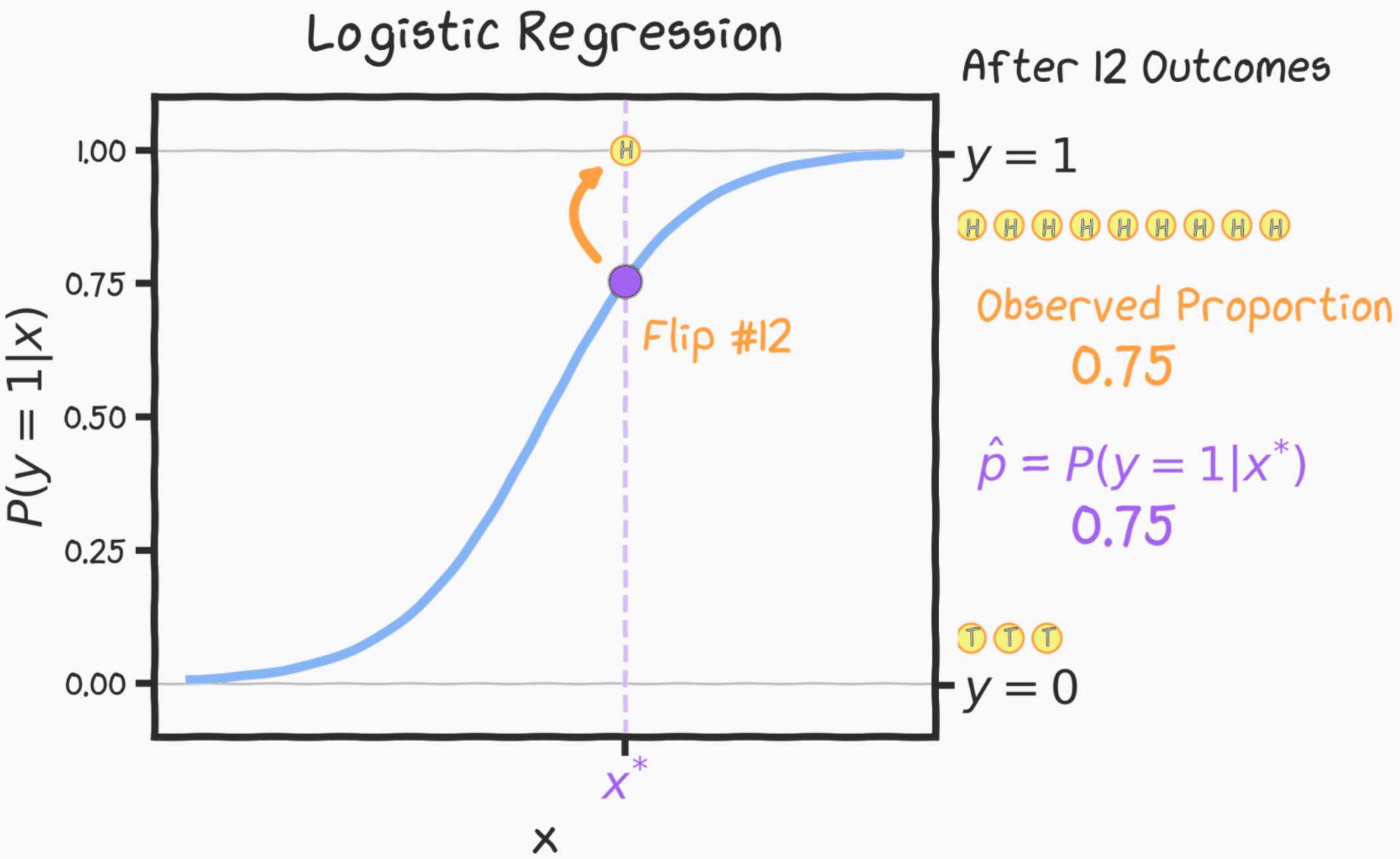
0.75

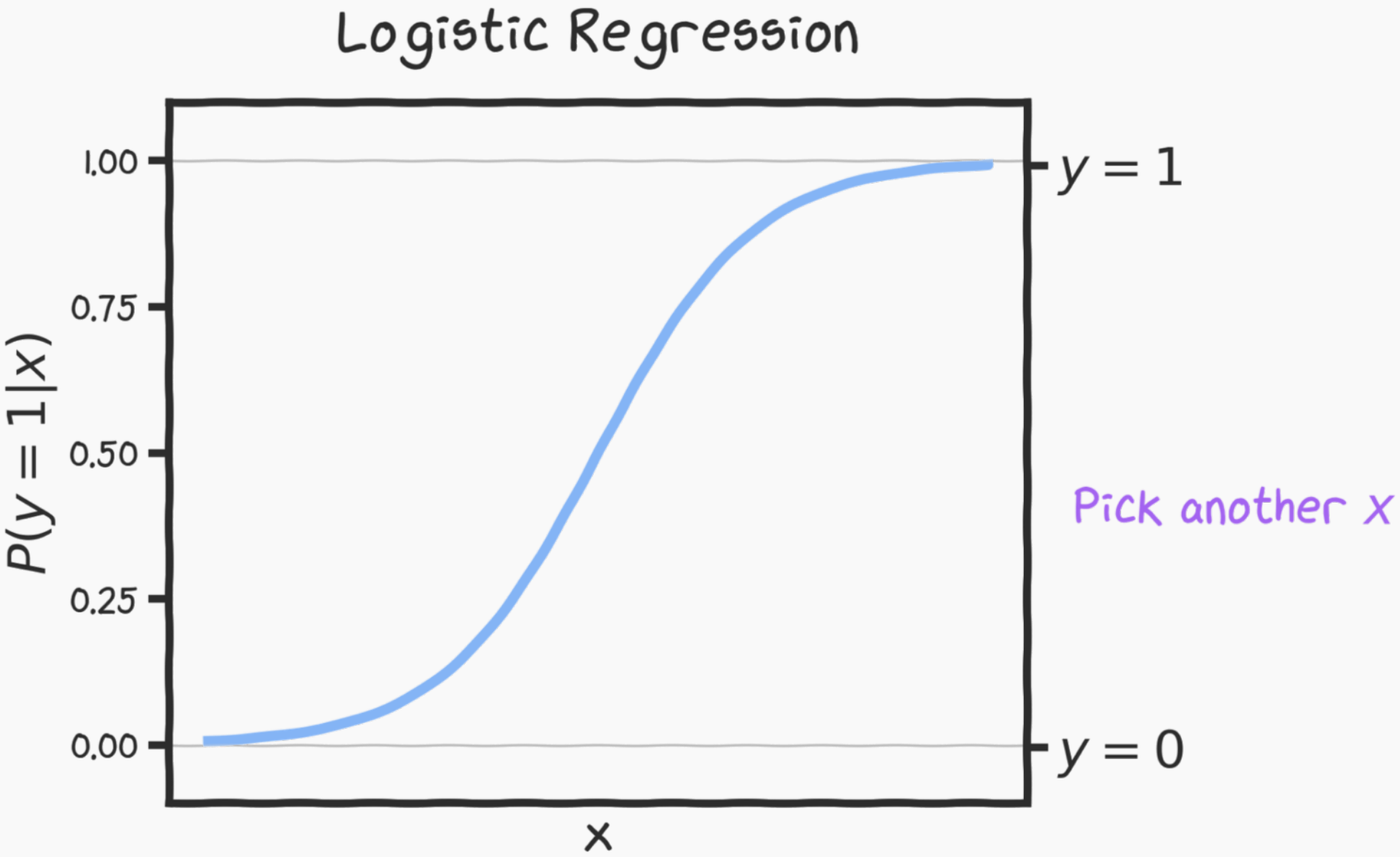

Logistic Regression
1.00
0.75
0.50
0.25
0.00
$P(y=1|x)$
$y=1$
$y=0$
H
Record Outcome
$\hat{p} = P(y=1|x^*)$
0.75
x^*
x

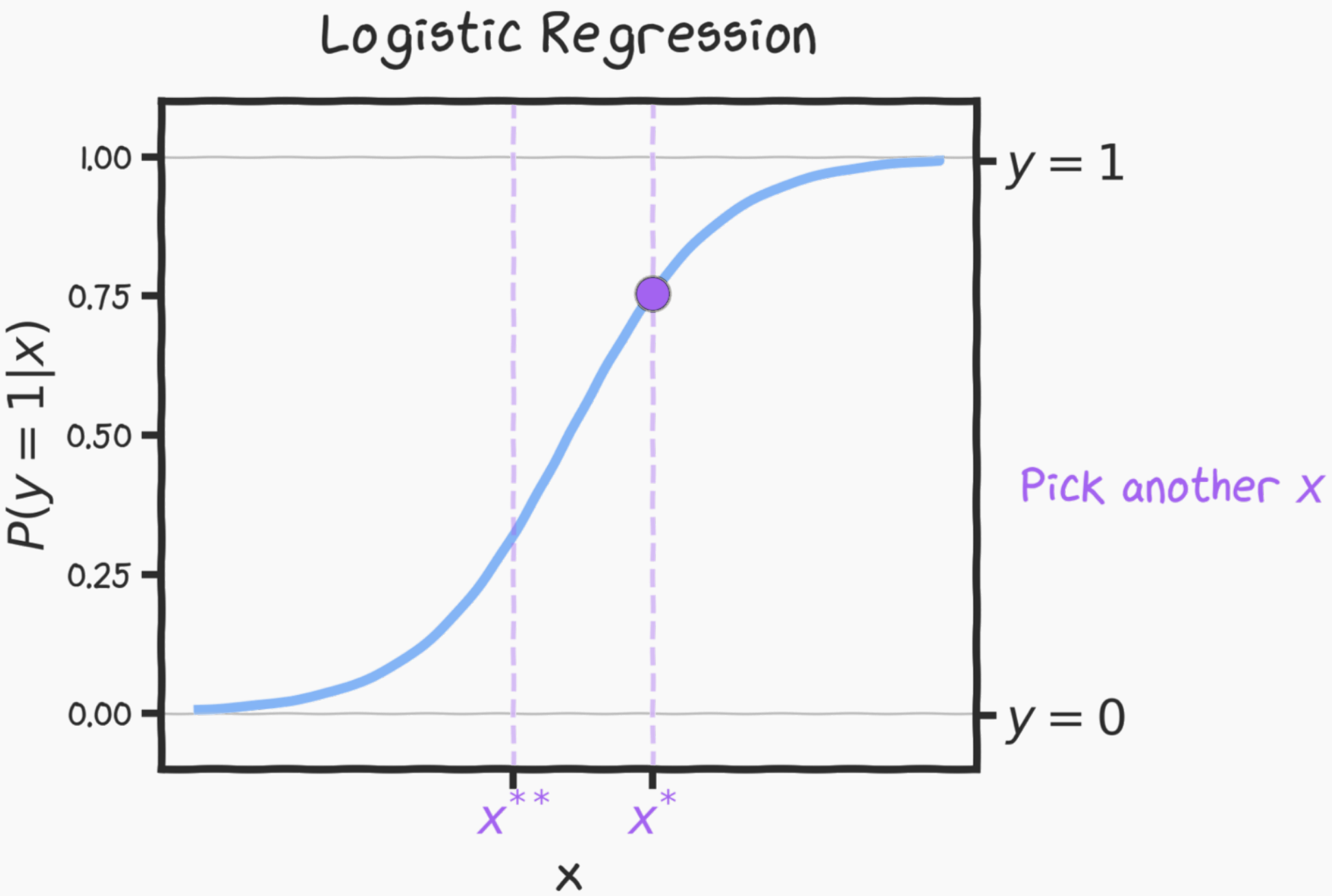

Logistic Regression
After 1 Outcome
1.00
0.75
0.50
0.25
0.00
$P(y=1|x)$
$y=1$
$y=0$
H
H
Flip #1
Record Outcome
$\hat{p} = P(y=1|x^*)$
0.75
x^*
x

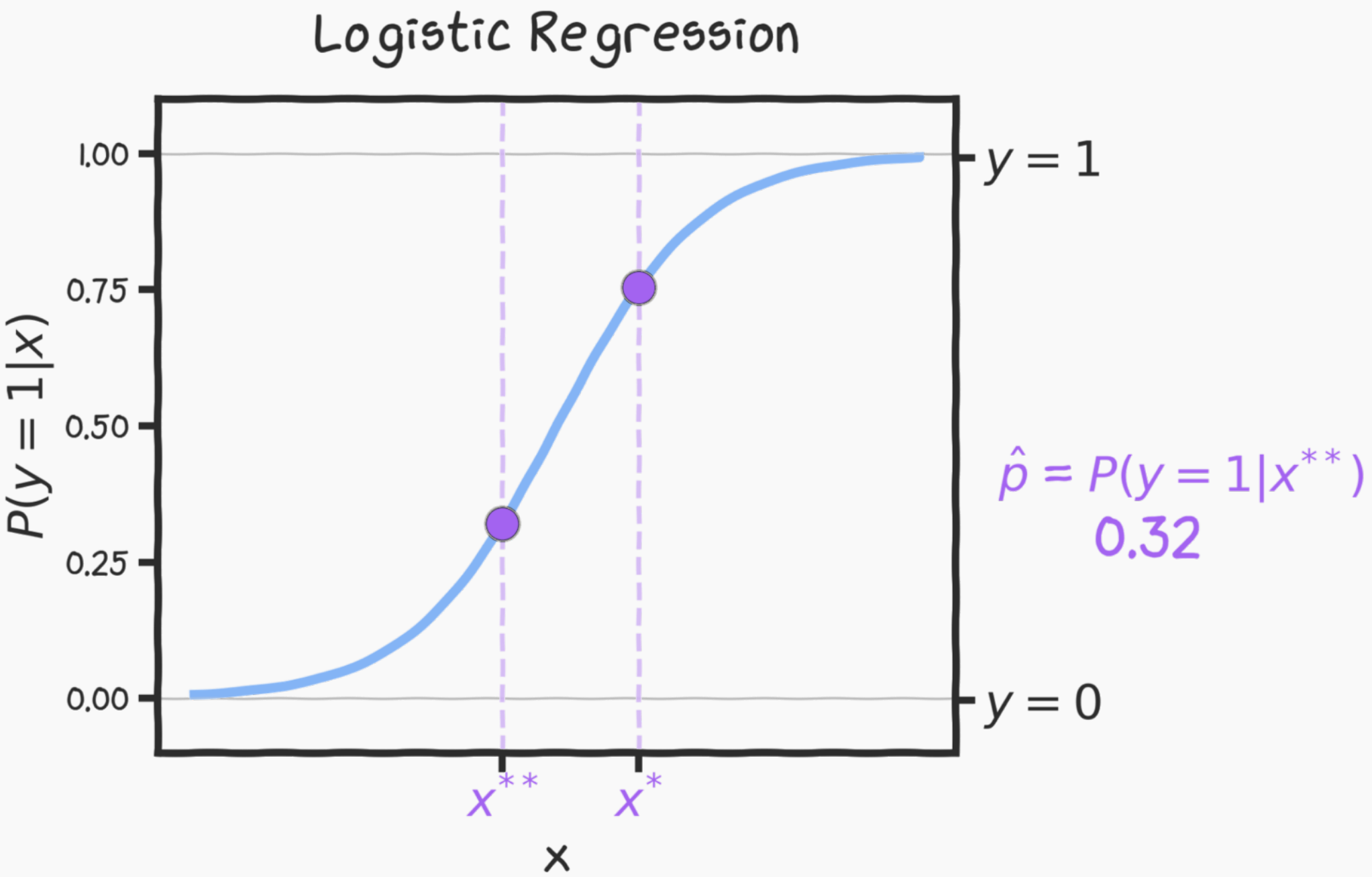

Logistic Regression
After 2 Outcomes
1.00
0.75
0.50
0.25
0.00
$P(y=1|x)$
$y=1$
$y=0$
H
H H
Flip #2
Observed Proportion
1.00
$\hat{p} = P(y=1|x^*)$
0.75
x^*
x


Logistic Regression
After 8 Outcomes
1.00
0.75
0.50
0.25
0.00
$P(y=1|x)$
$y=1$
H H H H H H H
Observed Proportion
0.88
Flip #8
$\hat{p} = P(y=1|x^*)$
0.75
T
$y=0$
x^*
x

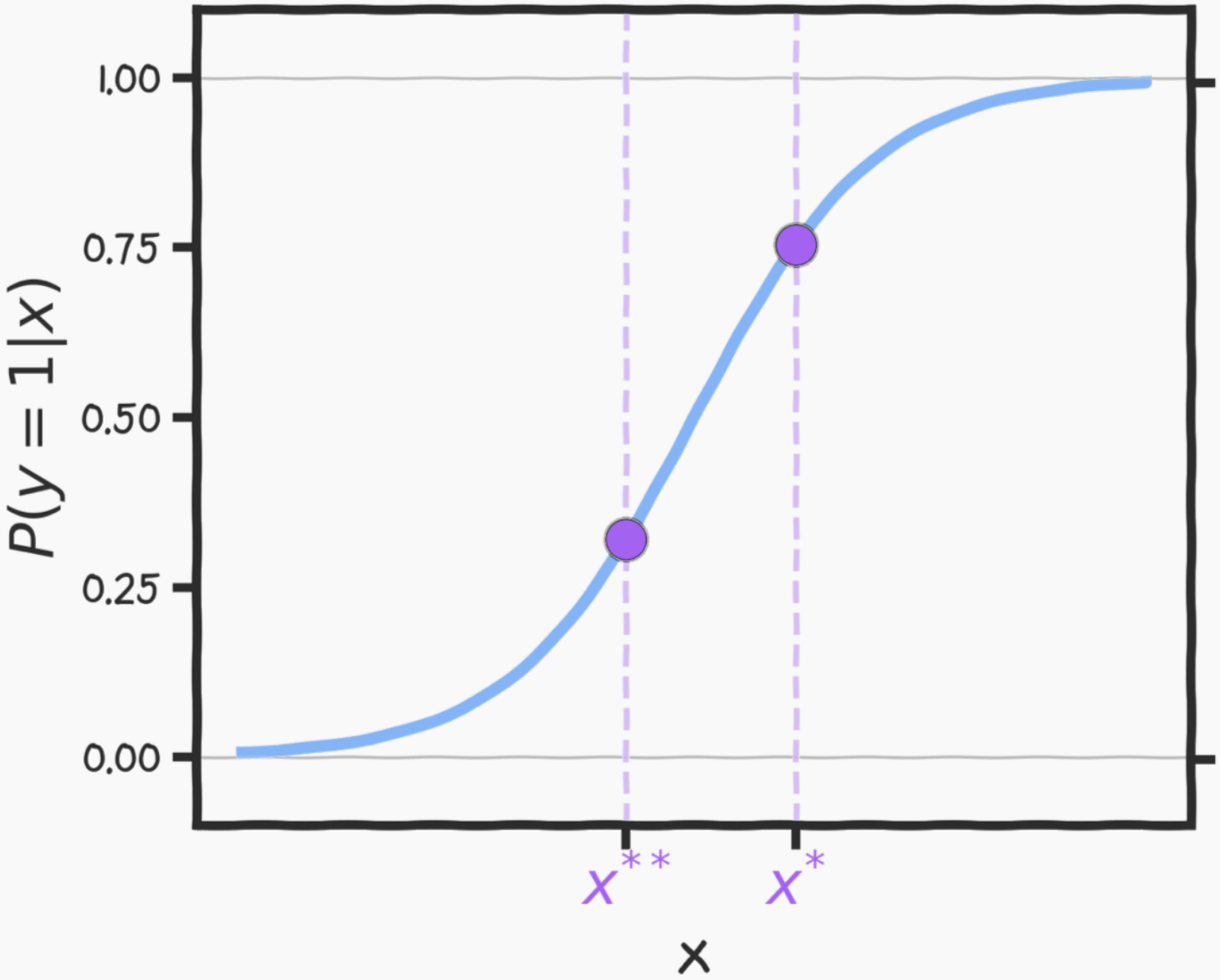

Logistic Regression
After 12 Outcomes
1.00
0.75
0.50
0.25
0.00
$P(y=1|x)$
$y=1$
Flip #12
Observed Proportion
0.75
$\hat{p} = P(y=1|x^*)$
0.75
$y=0$
x^*
x


Logistic Regression
1.00
0.75
0.50
0.25
0.00
$P(y=1|x)$
x
$y=1$
$y=0$
Pick another x


Logistic Regression
1.00
0.75
0.50
0.25
0.00
$P(y=1|x)$
$y=1$
$y=0$
Pick another x
x^{**}
x^{*}
x


Logistic Regression
1.00
0.75
0.50
0.25
0.00
$P(y=1|x)$
$y=1$
$y=0$
x^{**}
x^{*}
x
$\hat{p} = P(y=1|x^{**})$
0.32

Logistic Regression

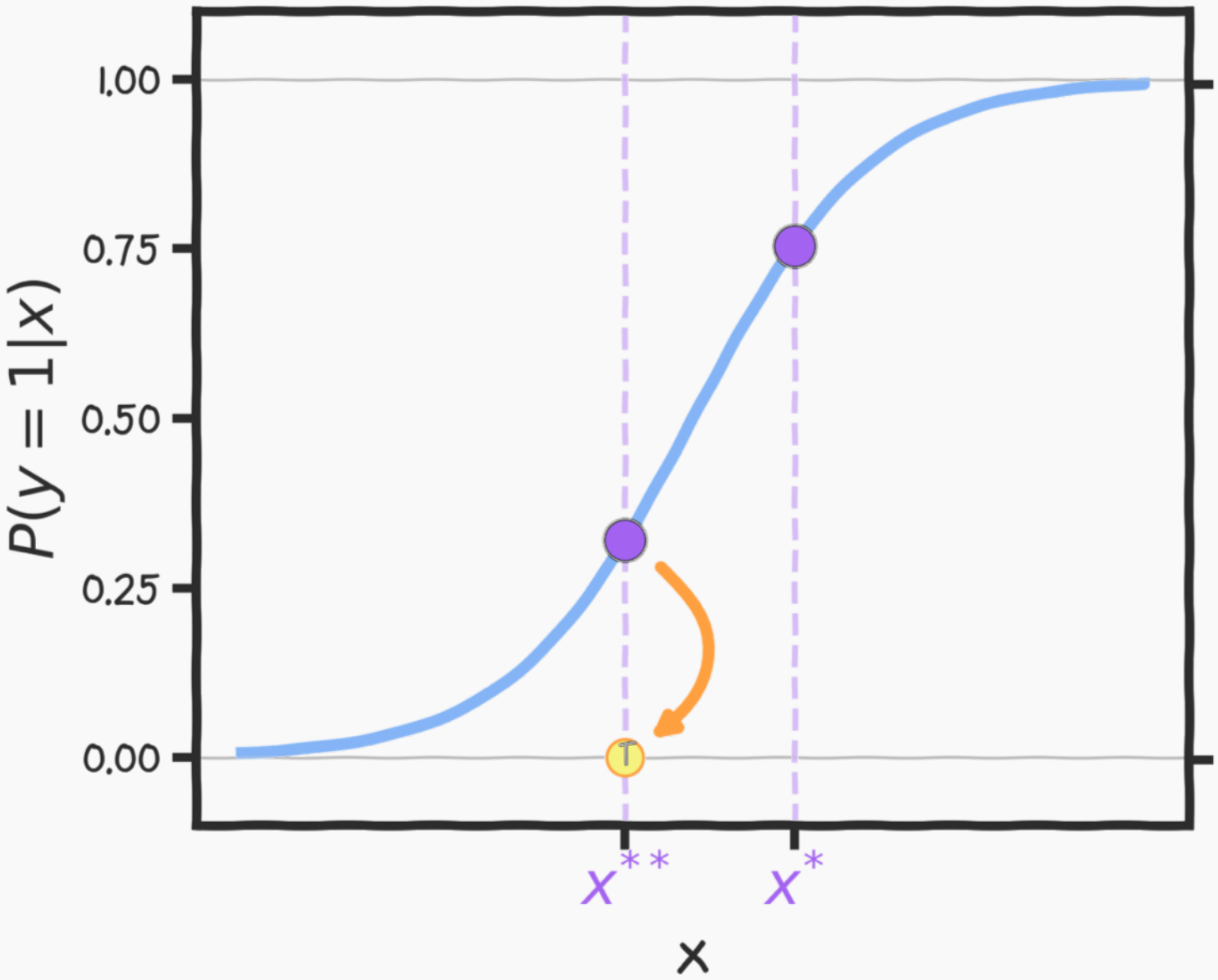


Flip a Coin with a
$\hat{P}$ Probability of Heads

$\hat{p} = P(y=1|x^{**})$

0.32

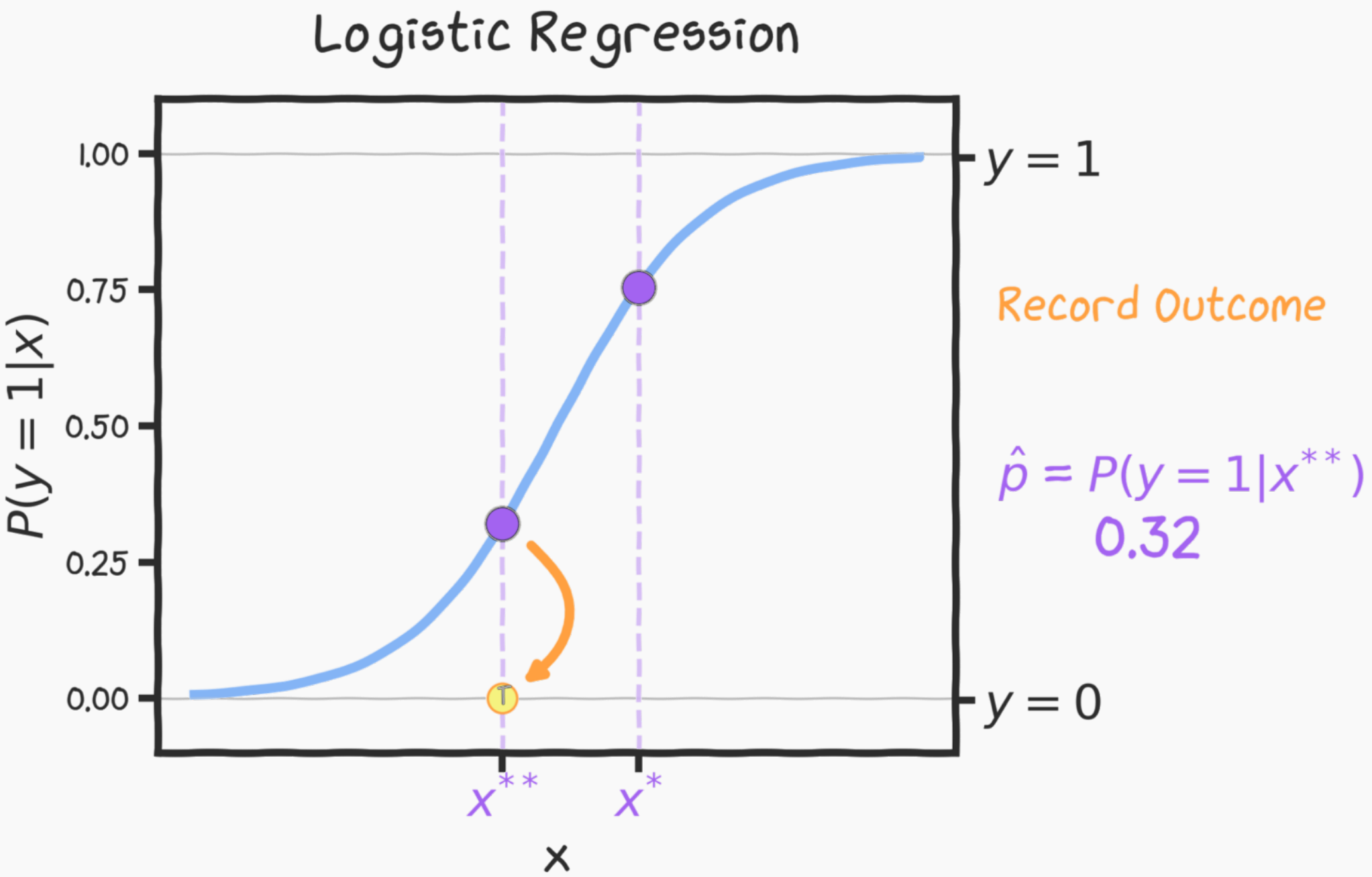
Logistic Regression

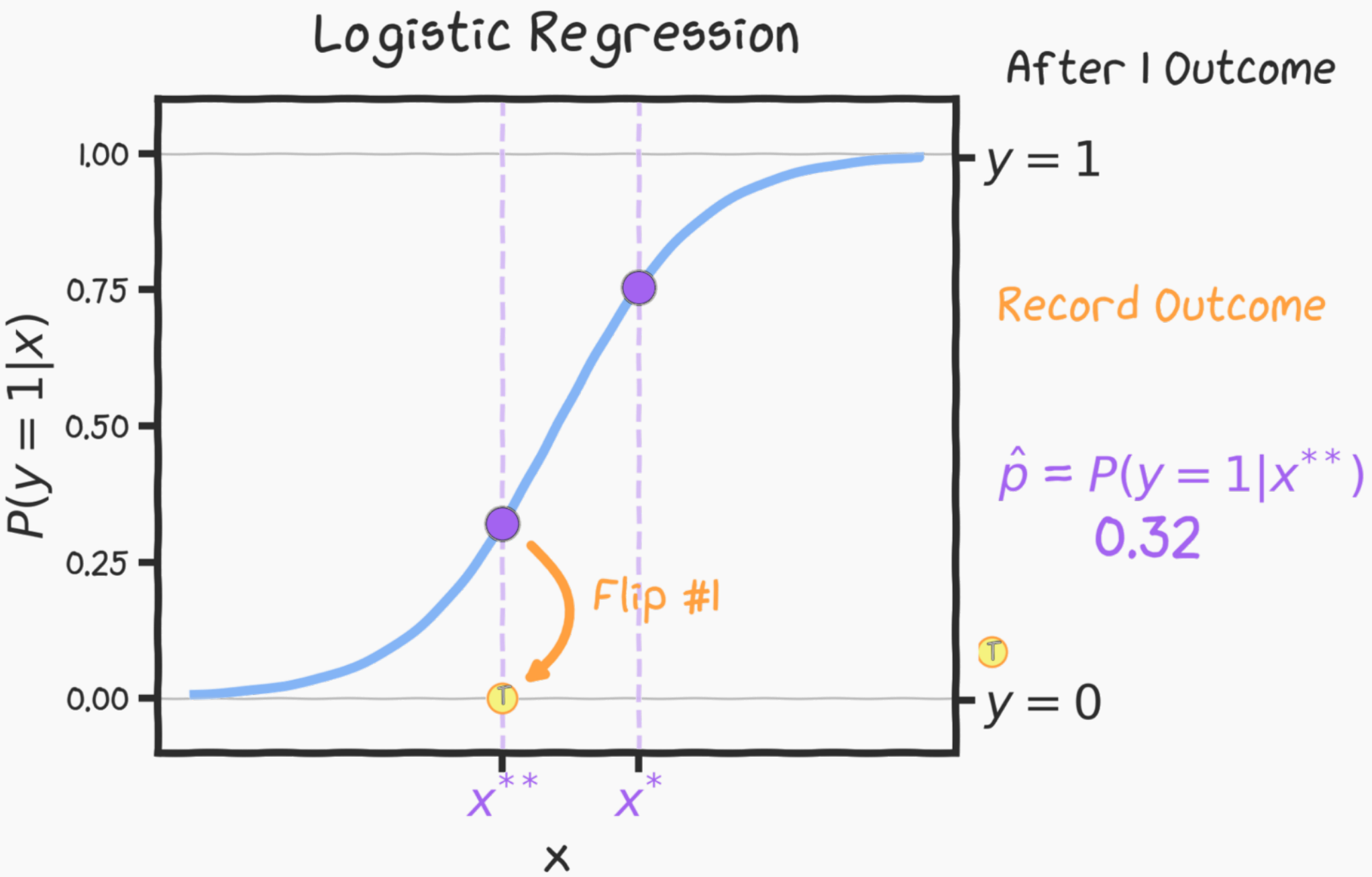


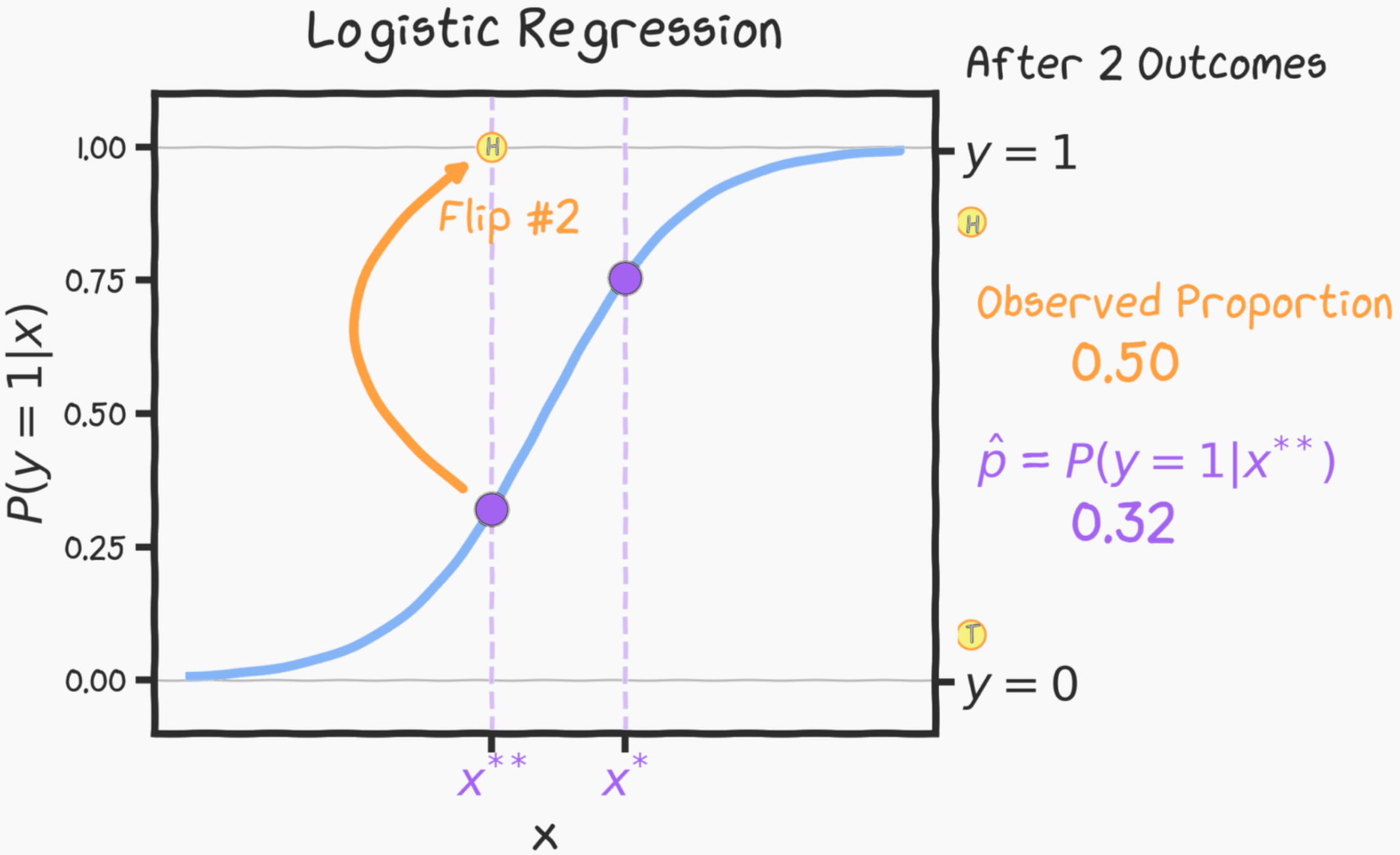
Flip a Coin with a
$\hat{P}$ Probability of Heads

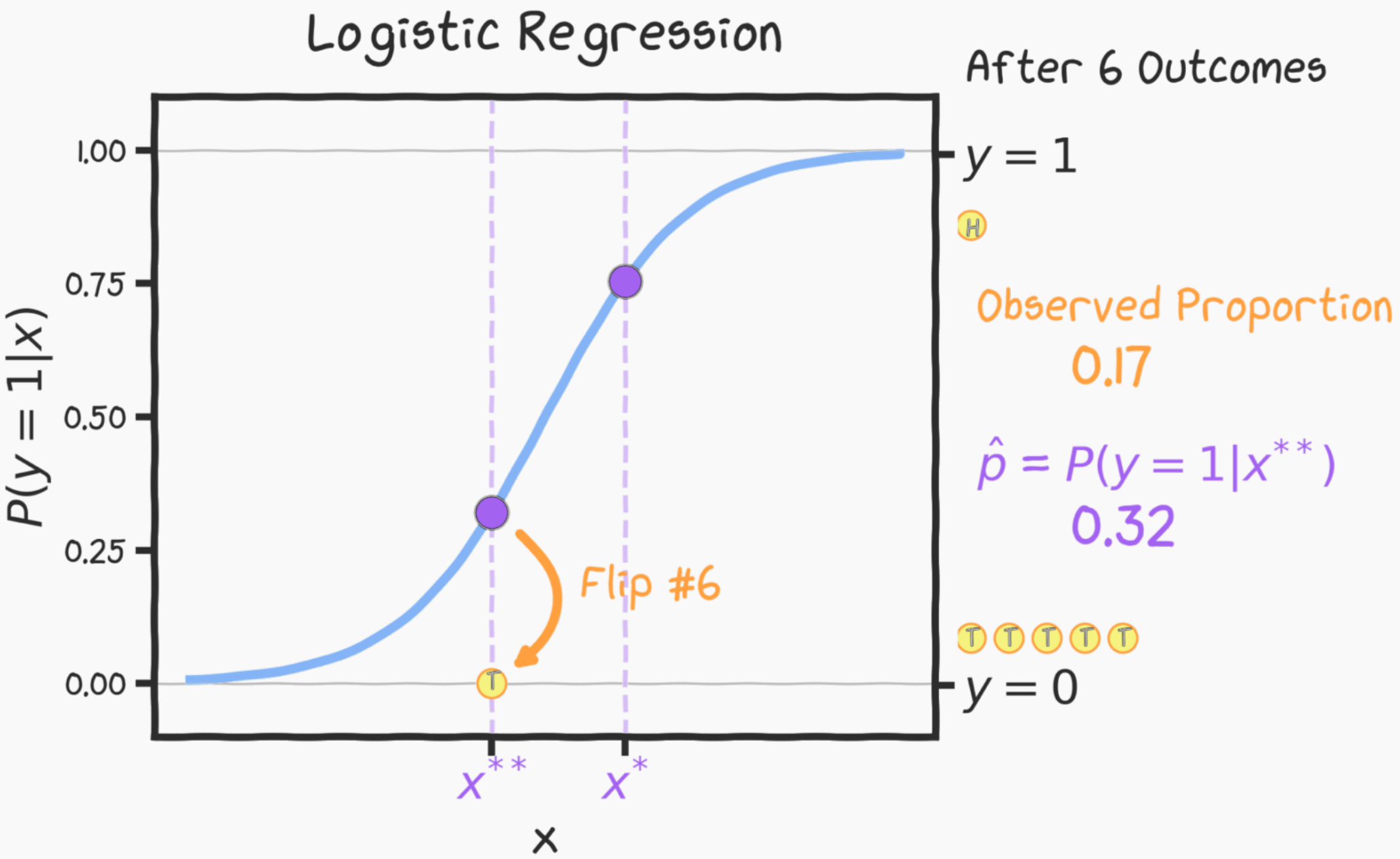
$\hat{p} = P(y=1|x^{**})$

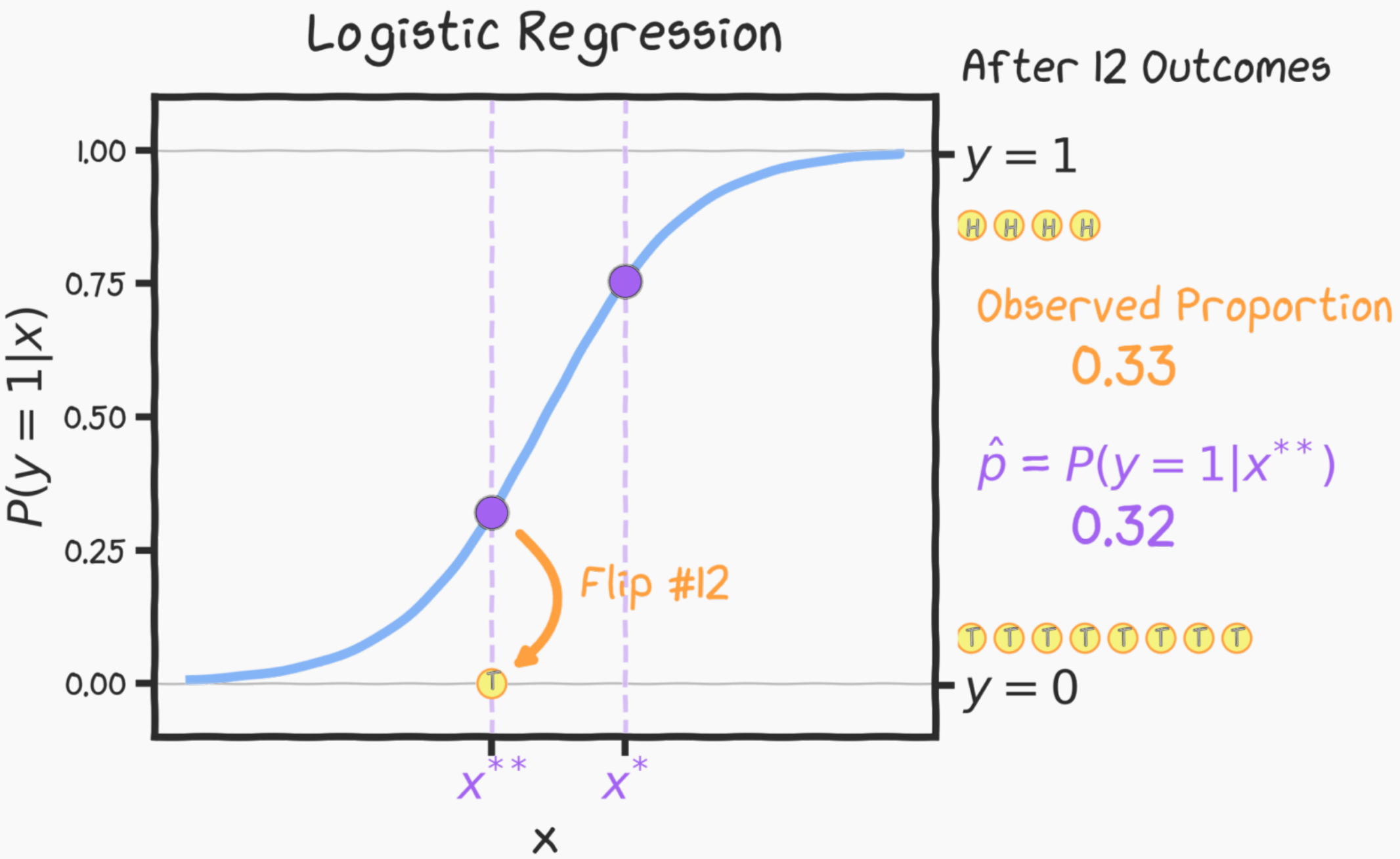
0.32


Logistic Regression
1.00
0.75
0.50
0.25
0.00
$P(y=1|x)$
$y=1$
$y=0$
Record Outcome
$\hat{p} = P(y=1|x^{**})$
0.32
x^{**}
x^{*}
x


Logistic Regression
After 1 Outcome
1.00
0.75
0.50
0.25
0.00
$P(y=1|x)$
$y=1$
$y=0$
Record Outcome
$\hat{p} = P(y=1|x^{**})$
0.32
Flip #1
T
T
x^{**}
x^{*}
x

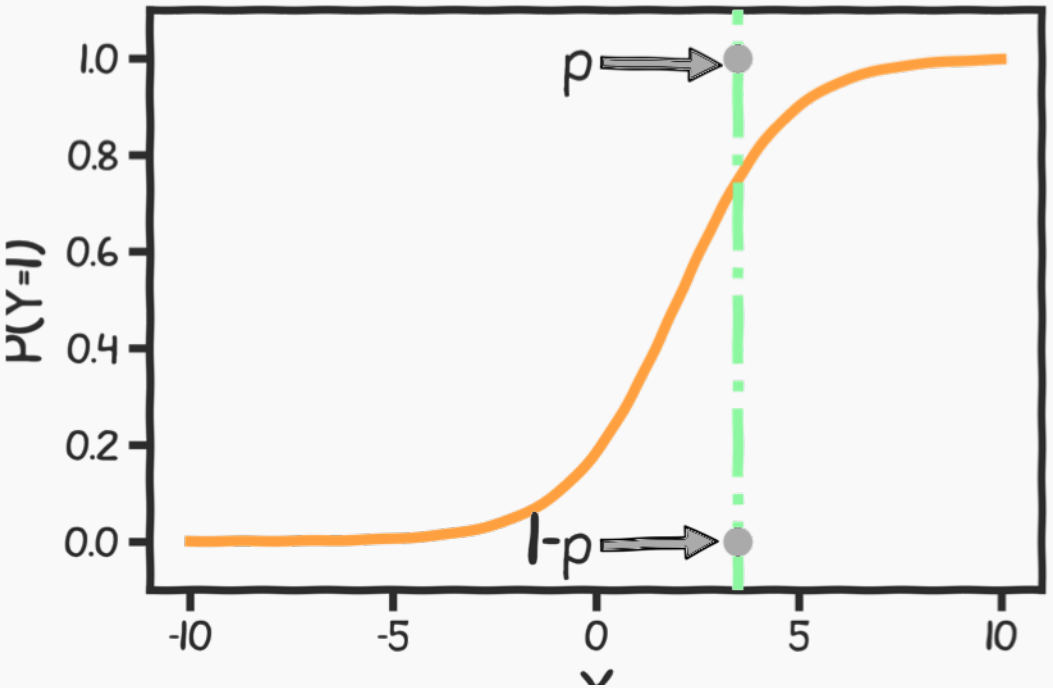

Logistic Regression
After 2 Outcomes
1.00
0.75
0.50
0.25
0.00
$P(y=1|x)$
H
Flip #2
$y = 1$
H
Observed Proportion
0.50
$\hat{p} = P(y=1|x^{**})$
0.32
T
$y = 0$
x^{**}
x^{*}
x


Logistic Regression
After 6 Outcomes
1.00
0.75
0.50
0.25
0.00
$P(y=1|x)$
$y=1$
H
Observed Proportion
0.17
$\hat{p} = P(y=1|x^{**})$
0.32
Flip #6
T T T T T
T
$y=0$
x^{**}
x^{*}
x


Logistic Regression
After 12 Outcomes
1.00
0.75
0.50
0.25
0.00
$P(y=1|x)$
$y=1$
H H H H
Observed Proportion
0.33
$\hat{p} = P(y=1|x^{**})$
0.32
Flip #12
T T T T T T T T
$y=0$
x^{**}
x^{*}
x

Estimating the Simple Logistic Model

For any X, the probability of getting heads or tails (1 or 0) is given by the logistic function shown in the plot as an orange line.



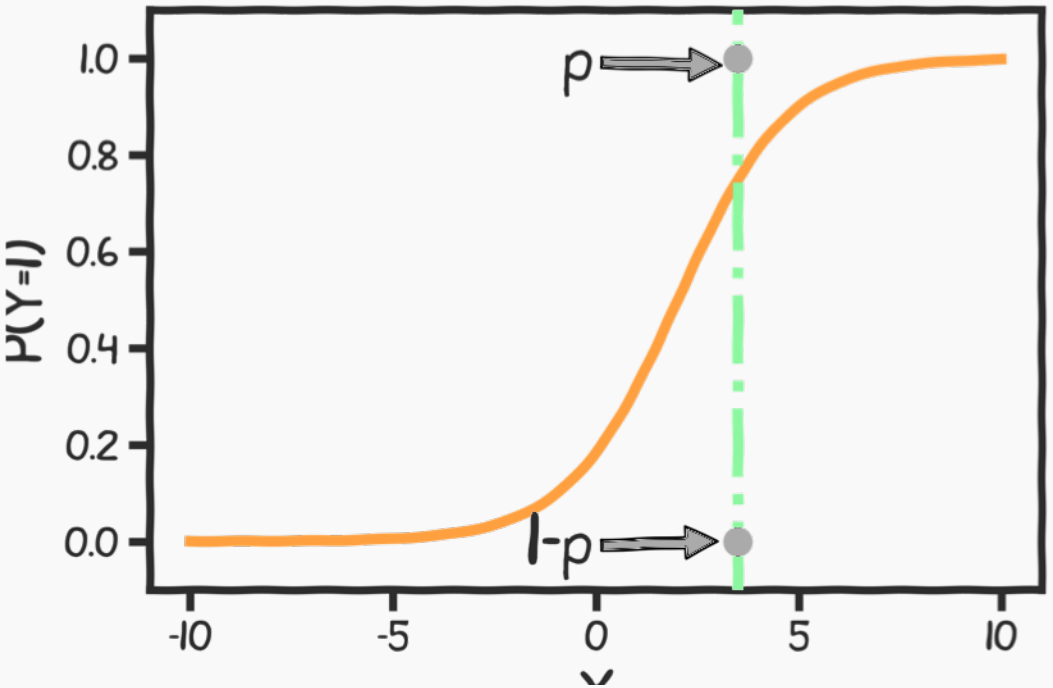
Prob. $Y = 1$: $P(Y = 1) = p$
Prob. $Y = 0$: $P(Y = 0) = 1 - p$

$$P(Y = y) = p^y (1 - p)^{(1-y)}$$

where $p = P(Y = 1|X = x)$ and therefore p depends on X. Thus, not every p is the same for each individual measurement.

Estimating the Simple Logistic Model

For any X, the probability of getting heads or tails (1 or 0) is given by the logistic function shown in the plot as an orange line.



Prob. $Y = 1$: $P(Y = 1) = p$
Prob. $Y = 0$: $P(Y = 0) = 1 - p$

$$P(Y = y) = p^y (1 - p)^{(1-y)}$$

where $p = P(Y = 1|X = x)$ and therefore p depends on X. Thus, not every p is the same for each individual measurement.

Estimating the Simple Logistic Model

The model takes the form: $P(Y = y) = p^y(1-p)^{(1-y)}$

The loss function that we use for Logistic Regression is not MSE, but Binary Cross Entropy.

The **Binary Cross Entropy** is expressed as follows:

$$L_{BCE} = -\sum_i y_i \log p_i + (1 - y_i) \log(1 - p_i)$$

Outline

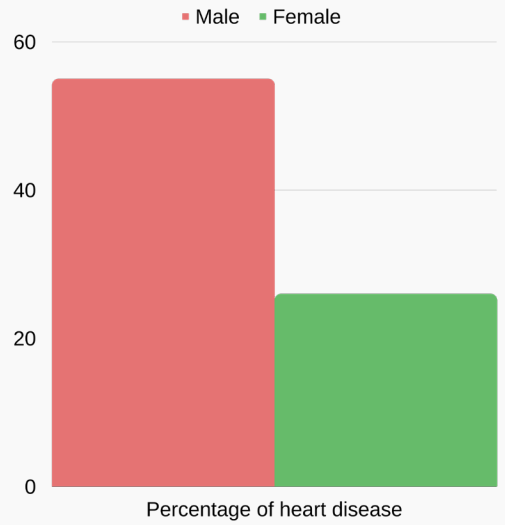
- What is Classification?
- Why not Linear Regression?
- Estimating the Simple Logistic Model
- **Inference in Logistic Regression**
- Multiple Logistic Regression
- Classification Decision Boundaries

Statistical Inference in Logistic Regression

Just like in linear regression, when the predictor, X, is binary, the interpretation of the model simplifies.

In this case, what are the interpretations of $\hat{\beta}_0$ and $\hat{\beta}_1$?

Consider the heart disease dataset represented here:



If we predict heart disease based on biological sex, how would you calculate and interpret the coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$?

A Second Logistic Regression Model in sklearn

Here is a logistic regression output to predict Y = AHD from X = Female:

```
logreg = LogisticRegression(penalty='none')
logreg.fit(df_heart[['Female']], df_heart['AHD'])

print('Estimated beta1: \n', logreg.coef_)
print('Estimated beta0: \n', logreg.intercept_)
```

```
Estimated beta1:
 [[-1.27219988]]
Estimated beta0:
 [0.21440982]
```

```
df_heart['Female'] = 1*(df_heart['Sex'] == 0)
pd.crosstab(df_heart['Female'],df_heart['AHD'])
```

AHD	No	Yes
Female		
0	92	114
1	72	25

What is the estimated model? What are the interpretations of the $\hat{\beta}$s?

$$\ln\left(\frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)}\right) = 0.2144 - 1.272(Female)$$

What is the estimated log-odd of AHD for Females and for Males? What about estimated probabilities? How does this agree with the table?

Outline

- What is Classification?
- Why not Linear Regression?
- Estimating the Simple Logistic Model
- Inference in Logistic Regression
- **Multiple Logistic Regression**
- Classification Decision Boundaries

Multiple Logistic Regression

It is simple to illustrate examples in logistic regression when there is just one predictors variable.

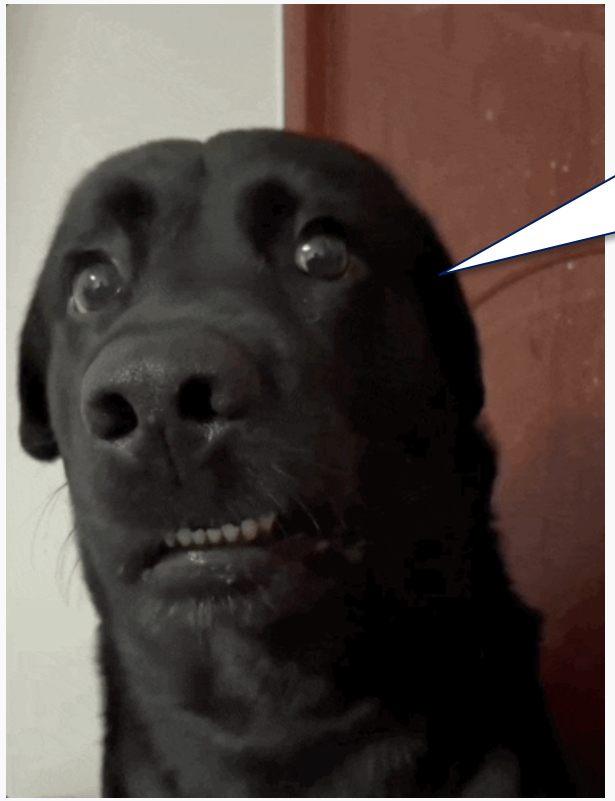
But the approach ‘easily’ generalizes to the situation where there are multiple predictors.

A lot of the same details as linear regression apply to logistic regression. Interactions can be considered; **collinearity** is a concern and so is **overfitting**.

Multiple Logistic Regression

collinearity

overfitting

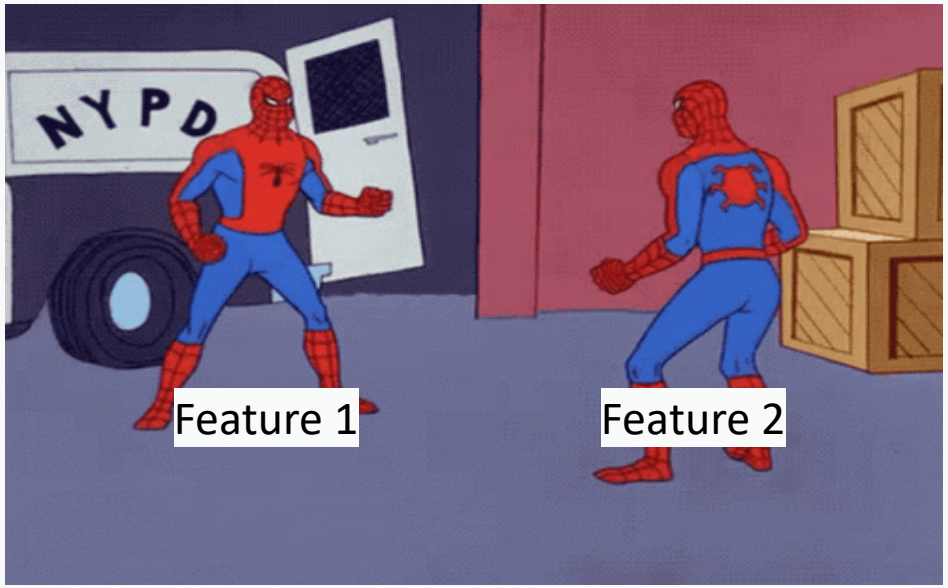


What are these?
I'm scared!

Don't fret! Let's look at each of them.

Collinearity

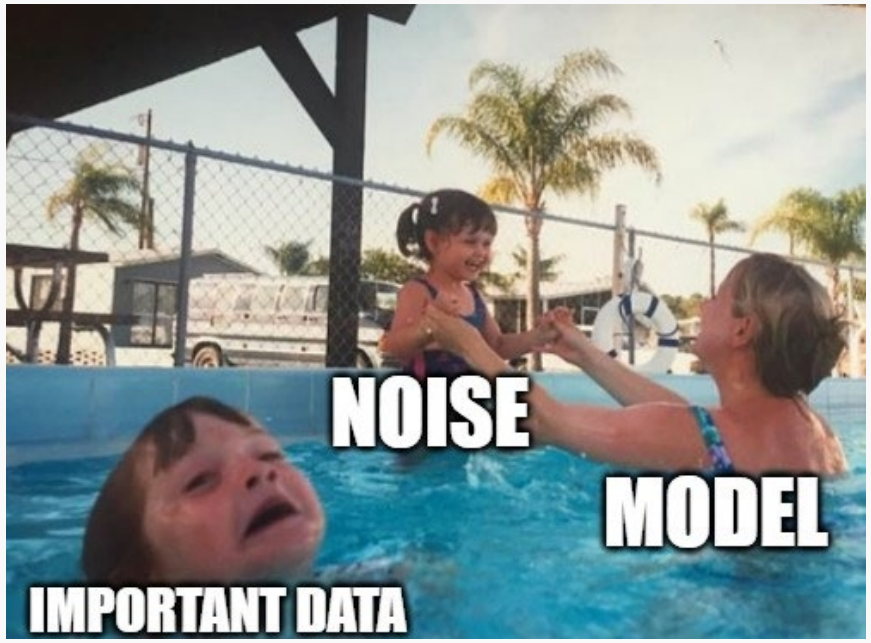
Collinearity refers to a situation where two or more predictors in a model are highly correlated with each other.



Collinearity affects our confidence in the estimated coefficients, making it challenging to assess the importance of individual predictors.

Overfitting

Overfitting is a common problem where a model learns the training data too well, capturing the noise and random fluctuations along with relevant patterns.



Overfitting indicates that the model has become too complex, tailoring itself too closely to the specifics of the training data.

Multiple Logistic Regression

So, that is what overfitting and collinearity is!

So how do we correct for such problems?

Regularization and checking though train and cross-validation!

We will get into the details of this, along with other extensions of logistic regression, in the next lecture.

Multiple Logistic Regression

Earlier we saw the general form of *simple* logistic regression, meaning when there is just one predictor used in the model:

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X$$

Multiple logistic regression is a generalization to multiple predictors. More specifically we can define a multiple logistic regression model to predict $P(Y=1)$ as such:

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_p X_p$$

Fitting Multiple Logistic Regression

The actual fitting of a Multiple Logistic Regression is easy using software (of course there's a python package for that) as the mathematical aspect of it has already been hard coded. So, we don't need to worry about it!

In the sklearn.linear_model package, you just have to create your multidimensional design matrix X to be used as predictors in the LogisticRegression function.

Outline

- What is Classification?
- Why not Linear Regression?
- Estimating the Simple Logistic Model
- Inference in Logistic Regression
- Multiple Logistic Regression
- **Classification Decision Boundaries**

Using Logistic Regression for Classification

How can we use logistic regression to perform classification?

That is, how can we predict when $Y = 1$ vs. when $Y = 0$?

We can classify all observations for which:

- Classify all observations with $\hat{P}(Y = 1) \geq 0.5$ to be in the group associated with $Y = 1$.
- Classify all observations with $\hat{P}(Y = 1) < 0.5$ to be in the group associated with $Y = 0$.

How would this extend if Y has 3+ classes?

Decision Boundaries for Classification

Recall that we could attempt to purely classify each observation based on whether the estimated $P(Y = 1)$ from the model is at least 0.5:

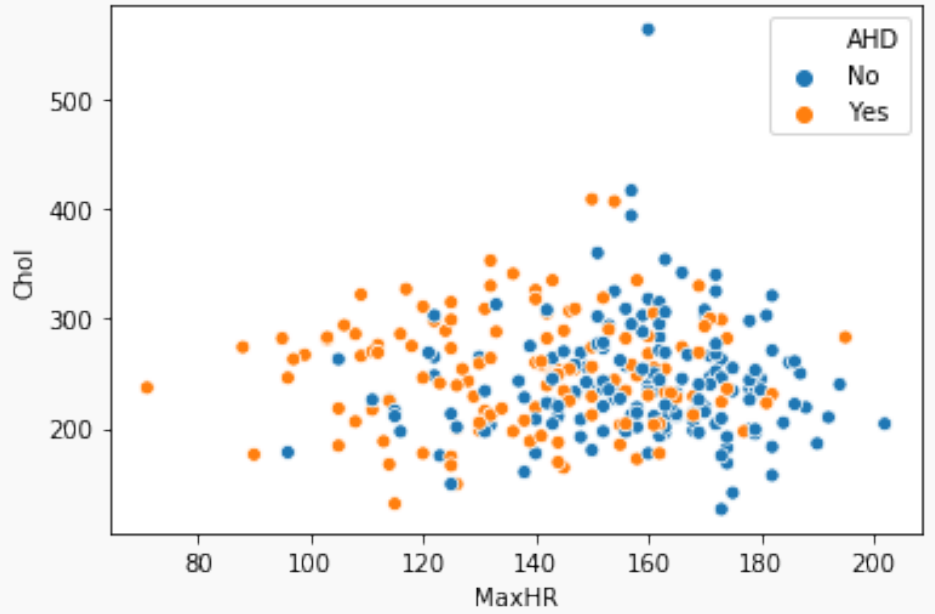
$$\hat{P}(Y = 1) \geq 0.5$$

This results in a **decision boundary**: a surface (line, curve, etc. in 2D) that separates the predicted classes into *sets.*

Decision Boundaries for Classification

Here's a 2-D plot from our Heart Data Set:

How do you expect logistic regression to draw the decision boundary?



Decision Boundaries Example

Here is the output from a logistic regression model with 2 predictors:

What is the estimated model?

```
data_x = df_heart[['MaxHR','Chol']]
data_y = df_heart['AHD']

logreg = LogisticRegression(penalty='none', fit_intercept=True)
logreg.fit(data_x, data_y);


print('Estimated beta1, beta2: \n', logreg.coef_)
print('Estimated beta0: \n', logreg.intercept_)
```

```
Estimated beta1, beta2:
 [[-0.04388093  0.00391746]]
Estimated beta0:
 [5.42271131]
```

$$\ln\left(\frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)}\right) = 5.423 - 0.0439(MaxHR) + 0.0039(Chol)$$

What will the decision boundary look like? Key: if $\hat{P}(Y=1) = 0.5$, then what are the estimated odds? What are the estimated log-odds?

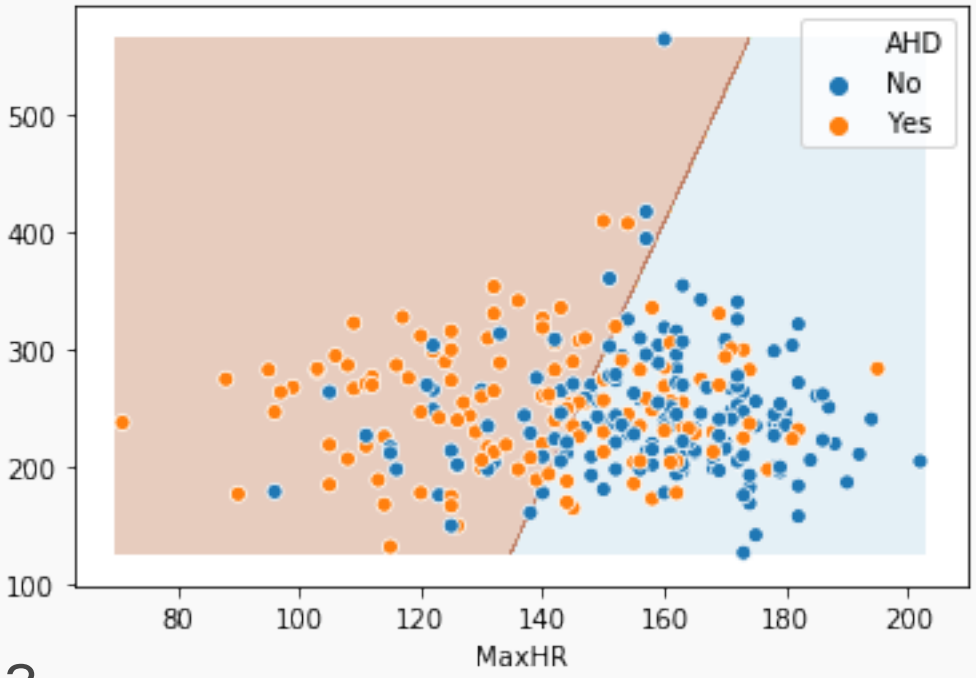
In logistic regression, the decision boundary is defined when $X\beta = 0$.

2D Classification in Logistic Regression: Example #1

A logistic regression model was fit to predict Y = AHD from X_1 = MaxHR and X_2 = Chol. Results shown below:

```
data_x = df_heart[['MaxHR','Chol']]
data_y = df_heart['AHD']

logreg = LogisticRegression(penalty='none', fit_intercept=True)
logreg.fit(data_x, data_y);


print('Estimated beta1, beta2: \n', logreg.coef_)
print('Estimated beta0: \n', logreg.intercept_)
```

```
Estimated beta1, beta2:
 [[-0.04388093  0.00391746]]
Estimated beta0:
 [5.42271131]
```



What will the decision boundary look like?

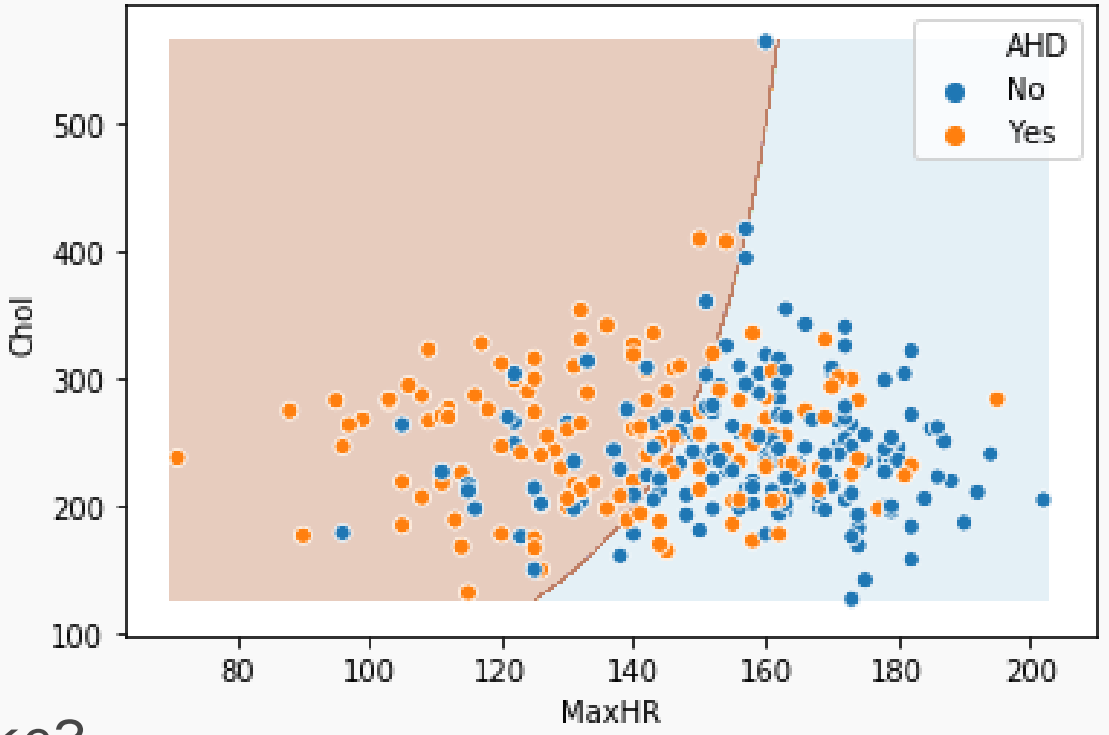
In logistic regression, decision boundaries are structured to be linear!

2D Classification in Logistic Regression: Example #2

A logistic regression model was fit to predict Y = AHD from X_1 = MaxHR and X_2 = Chol, and X_3 = their interaction. Results are shown below:

```
df_heart['Interaction'] = df_heart.MaxHR * df_heart.Chol

data_x = df_heart[['MaxHR','Chol', 'Interaction']]
data_y = df_heart['AHD']

logreg = LogisticRegression(penalty='none', fit_intercept=True)
logreg.fit(data_x, data_y);


print('Estimated beta1, beta2, beta3: \n', logreg.coef_)
print('Estimated beta0: \n', logreg.intercept_)
```

```
Estimated beta1, beta2, beta3:
 [[-0.00785835  0.02682656 -0.00015188]]
Estimated beta0:
 [5.70800455e-05]
```



What will the decision boundary look like?

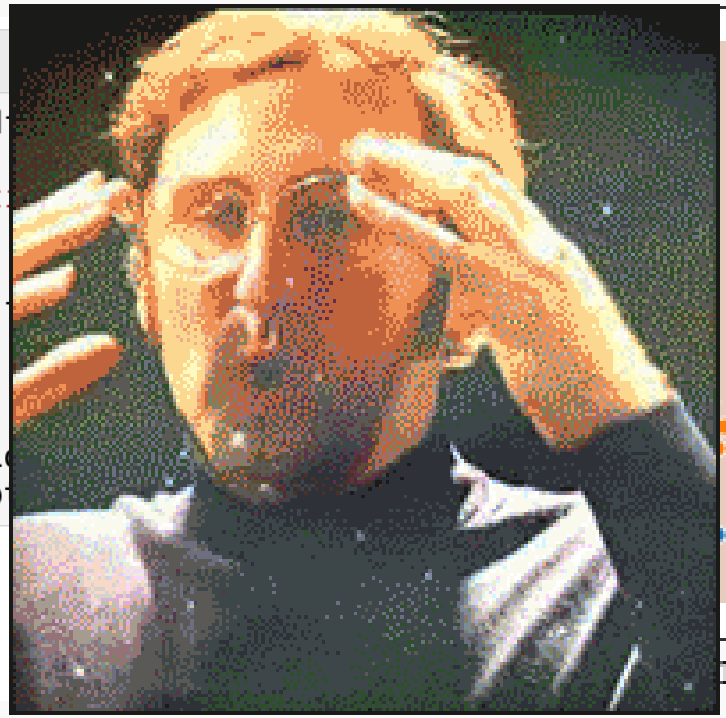
In logistic regression, decision boundaries are not always structured to be linear!

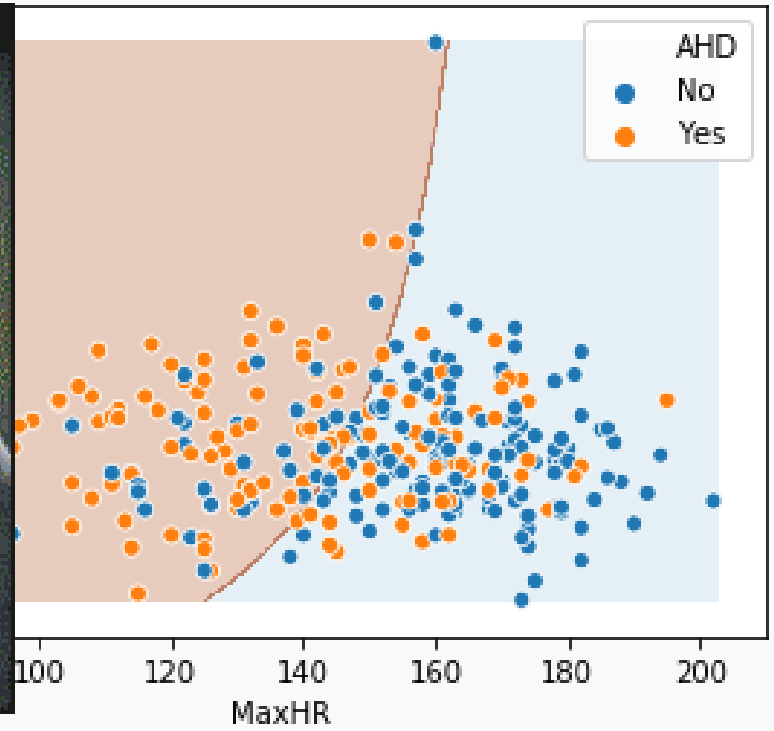
2D Classification in Logistic Regression: Example #2

A logistic regression model was fit to predict Y = AHD from X_1 = MaxHR and X_2 = Chol, and X_3 = their interaction. Results are shown below:

```
df_heart['Interaction'] = df_heart.MaxHR * df

data_x = df_heart[['MaxHR','Chol', 'Interact
data_y = df_heart['AHD']

logreg = LogisticRegression(penalty='none',
logreg.fit(data_x, data_y);


print('Estimated beta1, beta2, beta3: \n', l
print('Estimated beta0: \n', logreg.intercep
```

```
Estimated beta1, beta2, beta3:
 [[-0.00785835  0.02682656 -0.00015188]]
Estimated beta0:
 [5.70800455e-05]
```



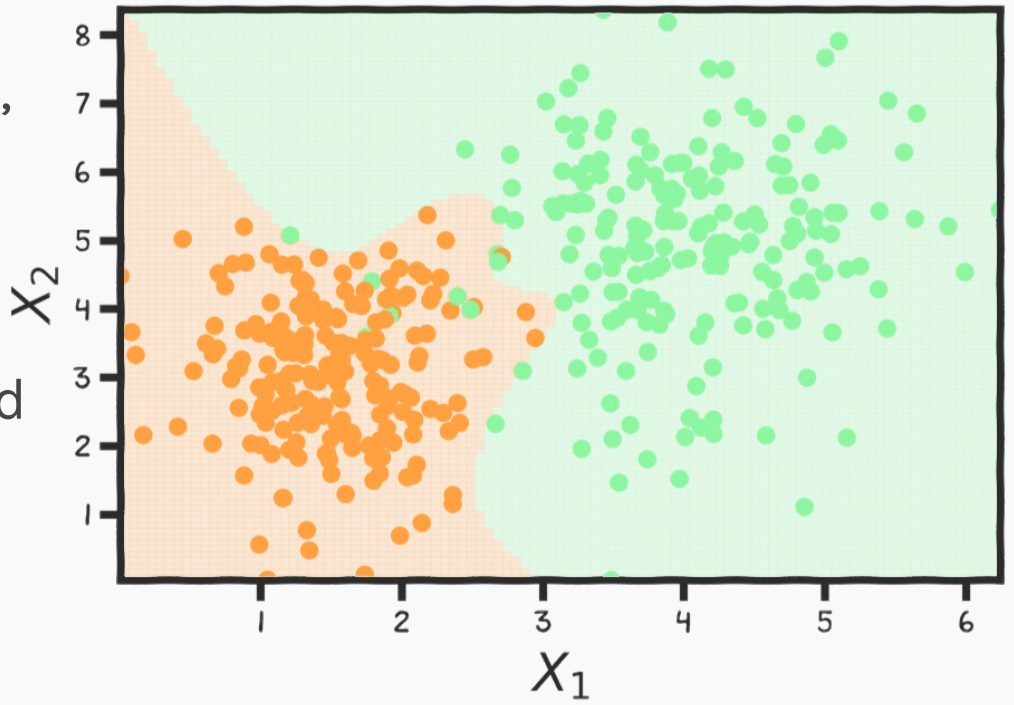


What will the decision boundary look like?

In logistic regression, decision boundaries are not always structured to be linear!

Polynomial Logistic Regression

We saw a 2-D plot last time which had two predictors, X_1, X_2. A similar one is shown here but the decision boundary is again not linear.



We can extend multiple Logistic Regression as we did with polynomial regression:

We transform the data by adding new predictors:

$$\tilde{x} = [1, \tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_M]$$

where $\tilde{x}_k = x^k$.

The polynomial Logistic Regression can be expressed as:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \tilde{X}\beta$$

Geometry of Decision Boundaries (Logistic Regression)

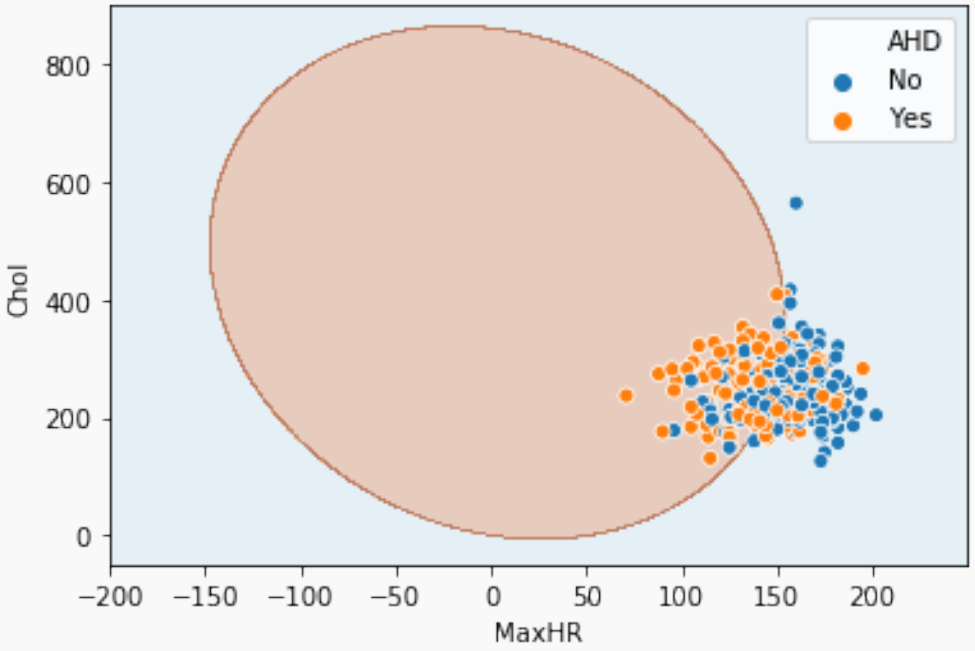
$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \tilde{X}\beta$$

Thus, we can define our logistic regression model to achieve a desired geometry.

For example, what set for $X = f(X_1, X_2)$ should we choose if we want a *circular* decision boundary?

$$X = \{X_1, X_1^2, X_2, X_2^2, X_1X_2\}$$

What could be an alternative modeling approach (think outside of logistic regression)?



Thank you!