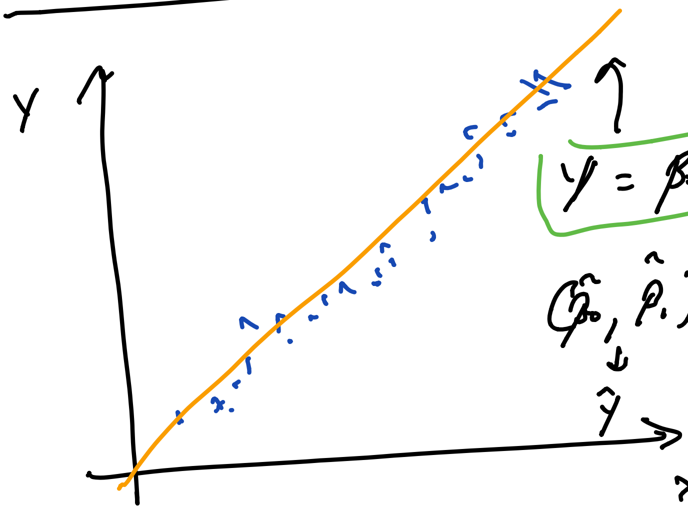
Regression Trees



$$\boxed{y = \beta_0 + \beta_1 x + \varepsilon}$$

$\downarrow$ fit using OLS or MLE

$(\hat\beta_0, \hat\beta_1)$

$\downarrow$

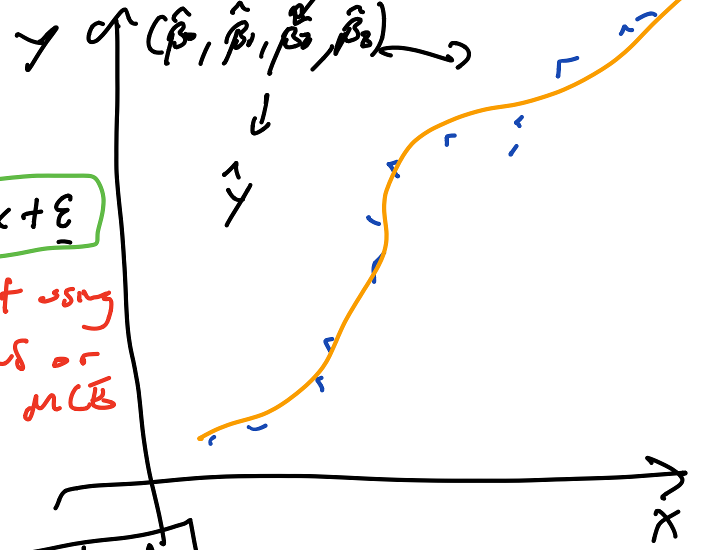
$\hat y$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

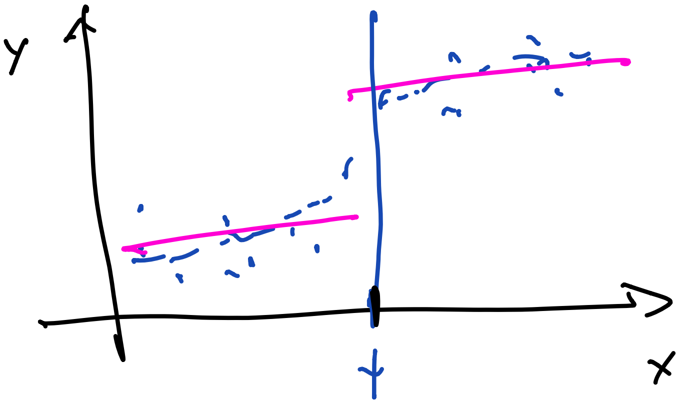
$(\hat\beta_0, \hat\beta_1, \hat\beta_2, \hat\beta_3)$

$\downarrow$

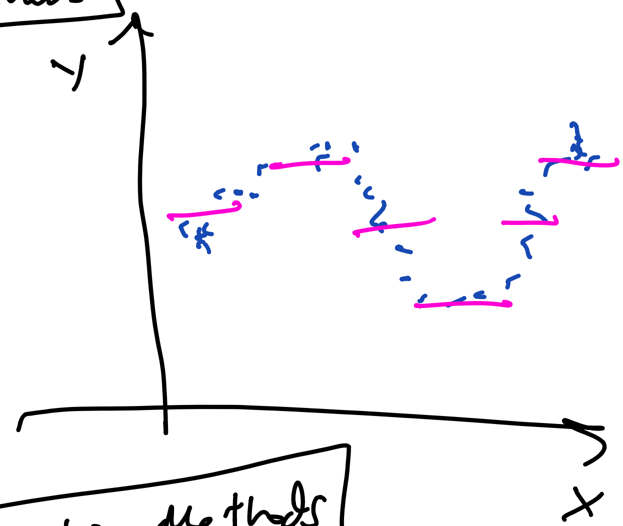
$\hat y$



$\boxed{\text{Parametric Methods}}$



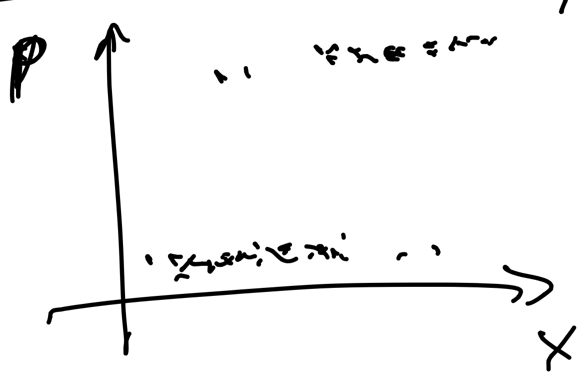
threshold



$\boxed{\text{Nonparametric Methods}}$

- We're <u>not</u> assuming a functional form a priori.

Logistic Reg.



$$\lambda = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x + \varepsilon$$

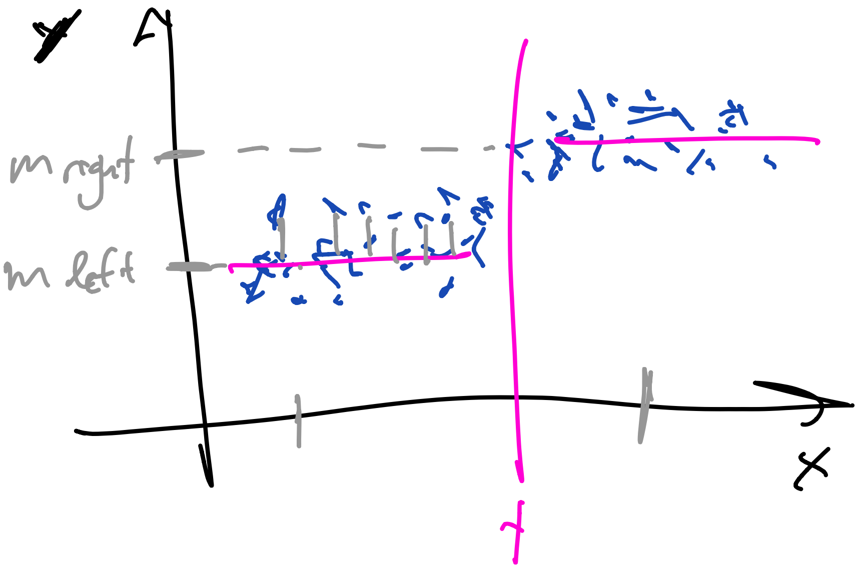
$\downarrow$

$(\hat\beta_0, \hat\beta_1)$

$\downarrow$

$\hat y$

Goal is to find a threshold s.t.
$\hat{y}_{left} = m_{left}$, $\hat{y}_{right} = m_{right}$ are
as different as possible.



At a given step k of this partitioning
algorithm

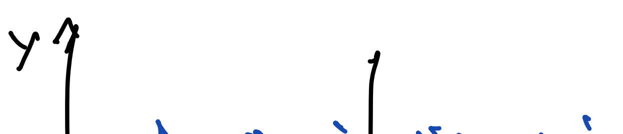
$$\underbrace{m_k}_{\text{mean of group } k} = \sum_{i \in \text{group } k} \frac{y_i}{N_k}$$

Further, $s^2_k = \sum_{i \in \text{Group } k} (y_i - m_k)^2$ (sum of squares of group k).

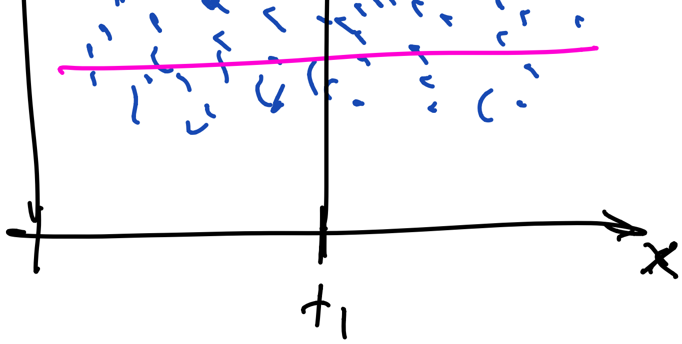
Simplest case

$$s^2_k = s^2_{k\,left} + s^2_{k\,right}$$

(only holds if x doesn't explain any variability



in y).



In general some x's can explain some variability in y. In this case (we'll prove it in the problem session).

$$\boxed{S^2_K} = \overbrace{S^2_{K\,left} + S^2_{K\,right}}^{\text{unexplained}} + \overbrace{\frac{N_{K\,left}\cdot N_{K\,right}}{N_K}\left(m_{K\,left} - m_{K\,right}\right)^2}^{\text{Explained}}$$

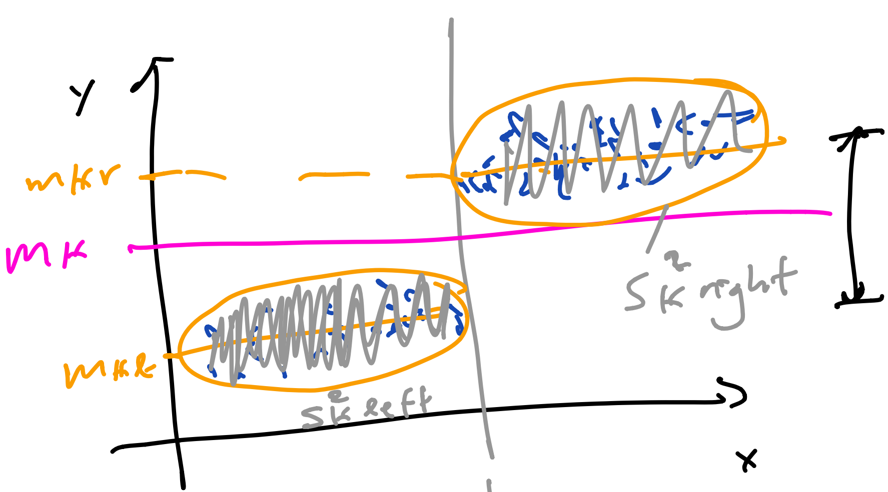
→ Our model for the data is pretty much just $m_{K\,left}$ and $m_{K\,right}$ + threshold which defines $N_{K\,left}$, $N_{K\,right}$.

Computed w.r.t m_K.

Thus, your threshold value doesn't matter.

$$S^2_{K\,left} = \sum_{i \in left}(y_i - m_{Kl})^2$$

$$S^2_{K\,right} = \sum_{i \in right}(y_i - m_{Kr})^2$$



$N_{Kl} + N_{Kr} = N_K$

$$S^2_K = S^2_{K\,left} + S^2_{K\,right} + \frac{N_{Kl}\,N_{Kr}}{N_K}\left(m_{Kl} - m_{Kr}\right)^2$$

Want to maximize $\underbrace{\frac{N_{K\ell} \cdot N_{Kr}}{N_K} (m_{K\ell} - m_{Kr})^2}$

(Maximizing Information Gain).

How do we run this regression practically?

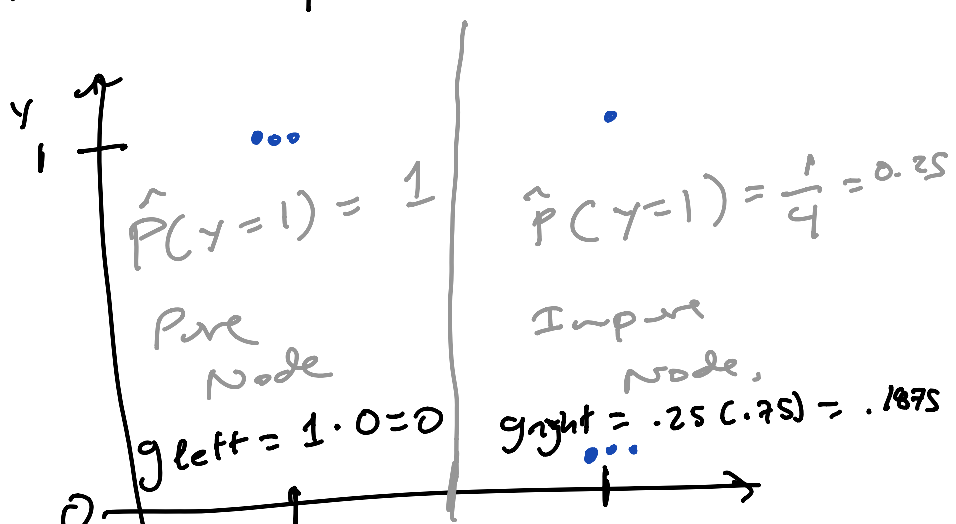
1. Pick a variable
2. Find a threshold that maxes

- Finally, let's define impurity of a node as a measure of deviation from the predicted behavior of the node.

In our regression case: $s^2_{K\ell}$ or s^2_{Kr}

• In the classification context, impurity is the probability that our prediction is incorrect.

Y	X_1
1	1
0	2
1	1
0	2
1	2
1	1
	2



✓ 0 2 1 t=1.5 2 x

we usually measure impurity by the Gini Index

$$g = \sum_{i \in K} P(y=i) \cdot (1 - P(y=i))$$

Example: