Linear Regression

Lecture 1
Section 1
June 1st

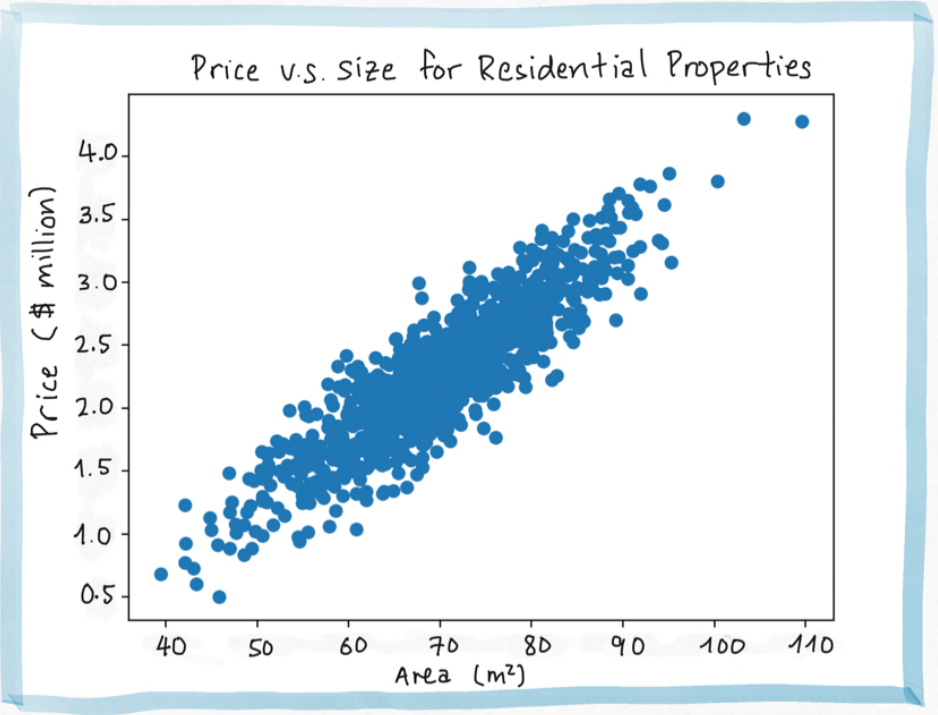
INSTITUTE FOR APPLIED
COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY

UNIVERSITY of
RWANDA

LINEAR REGRESSION

MOTIVATION: PREDICTING HOUSING PRICES

Build a model to predict selling price based on size.



This is a regression problem.

The target, y, is price.

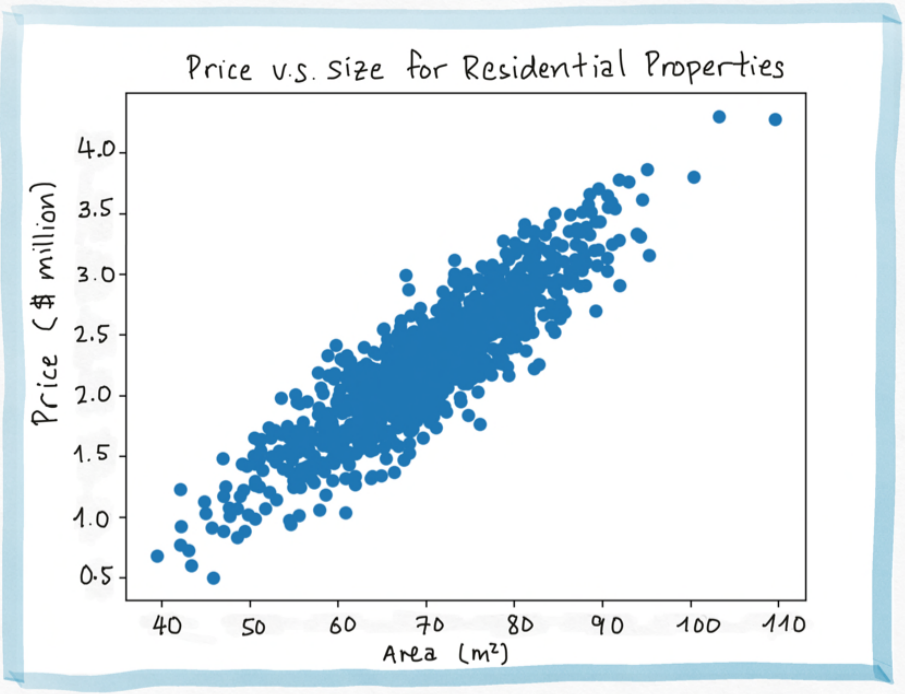
The predictor, x, is size.

The relationship between x and y appears linear:

$$y = w_1 x + w_0$$

LINEAR REGRESSION

Hypothesis: $y = w_1 x + w_0$, where x is size and y is price



But which line fits the data best?

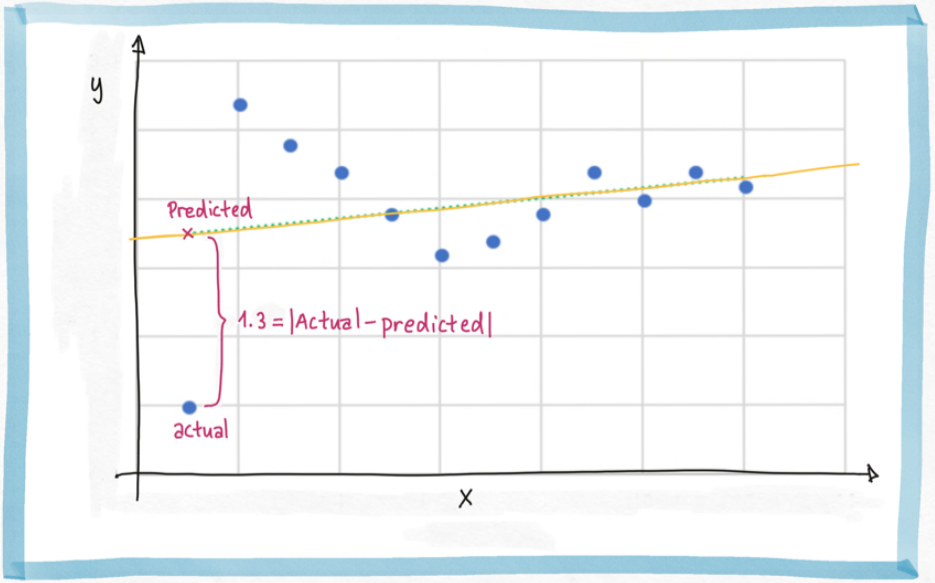
How do we choose w_1, w_0?

RESIDUALS

For each observation (x_n, y_n), the absolute residual is:

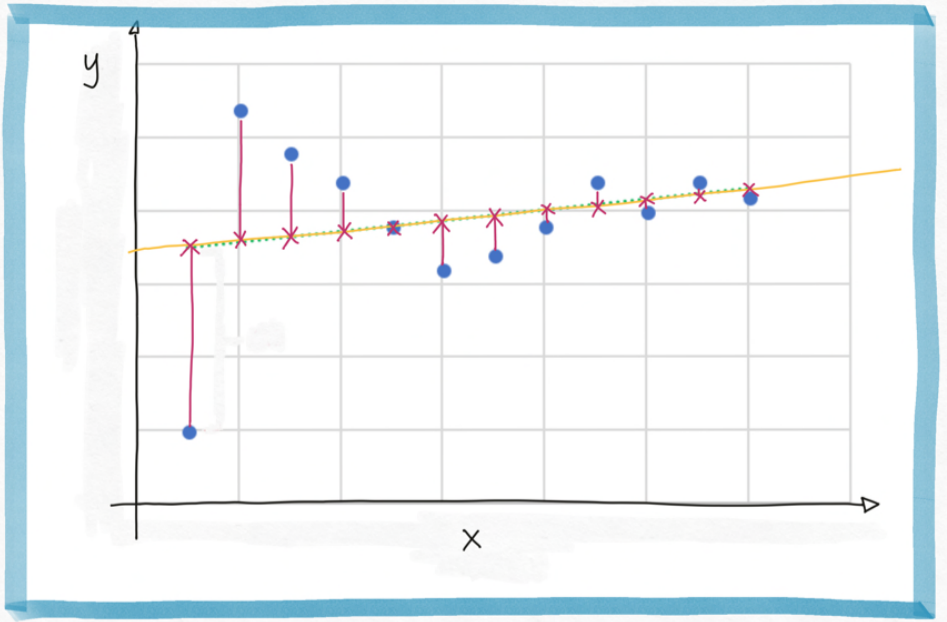
$$R_n = |y_n - \hat{y}_n|$$

Where $\hat{y}_n$ is predicted by the model, $\hat{y}_n = W_1 x_n + W_0$



LOSS FUNCTIONS: AGGREGATNG RESIDUALS

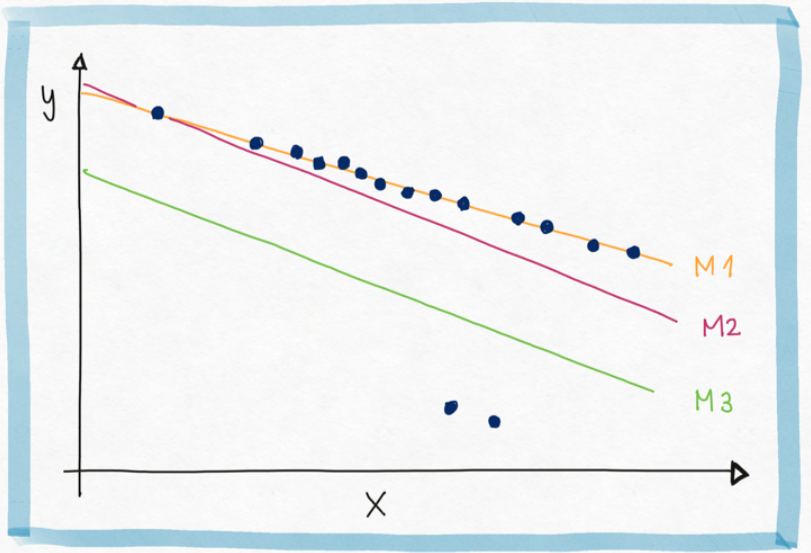
How do we aggregate residuals across the entire dataset?



1. (Max Absolute Error)
   Count only the biggest error
   $$\max_n |y_n - \hat{y}_n|$$
2. (Mean Absolute Error)
   count the average error
   $$\frac{1}{N} \sum_{n=1}^{N} |y_n - \hat{y}_n|$$
3. (Mean Squared Error)
   count the average squared error
   $$\frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2$$

COMPARING LOSS FUNCTIONS:

What is the difference between the three notions of error?



Model 1 has the lowest mean absolute error

$$\frac{1}{N}\sum_{n=1}^{N}|y_n-\hat{y}_n|$$

Model 2 has the lowest mean squared error

$$\frac{1}{N}\sum_{n=1}^{N}(y_n-\hat{y}_n)^2$$

Model 3 has the lowest max absolute error

$$\max_n |y_n-\hat{y}_n|$$

MODEL TRAINING: LINEAR REGRESSION

Given a dataset $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, we find the optimal linear model $y = w_1 x + w_0$:

1. Choose a notion of overall error: mean squared error

2. Define a loss function:

$$\mathcal{L}(w_1, w_0) = \frac{1}{N} \sum_{n=1}^{N} (y_n - (w_1 x_n + w_0))^2$$

3. Find parameters w_1, w_0 that minimizes the loss function:

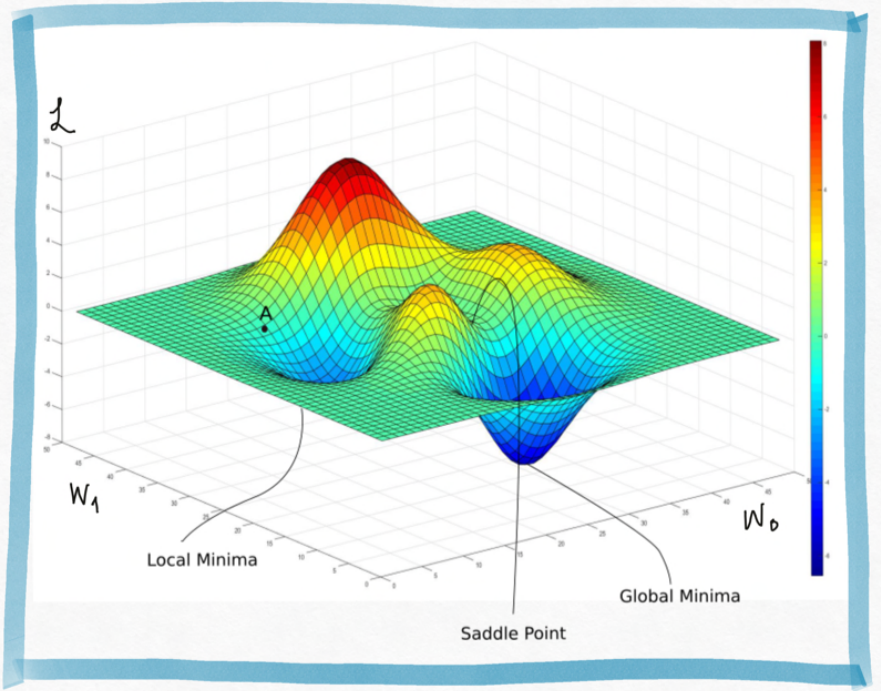
$$w_0^*, w_1^* = \underset{w_1, w_0}{\operatorname{argmin}} \mathcal{L}(w_1, w_0)$$

This is called fitting or training the model.

OPTIMIZATION

OPTIMIZATION:

How does one minimize a loss function?



The global maxima or global minima of $\mathcal{L}(W_1, W_0)$ must occur at a point where the gradient

$$\nabla \mathcal{L} = \left[\frac{\partial \mathcal{L}}{\partial W_1}, \frac{\partial \mathcal{L}}{\partial W_0}\right]$$

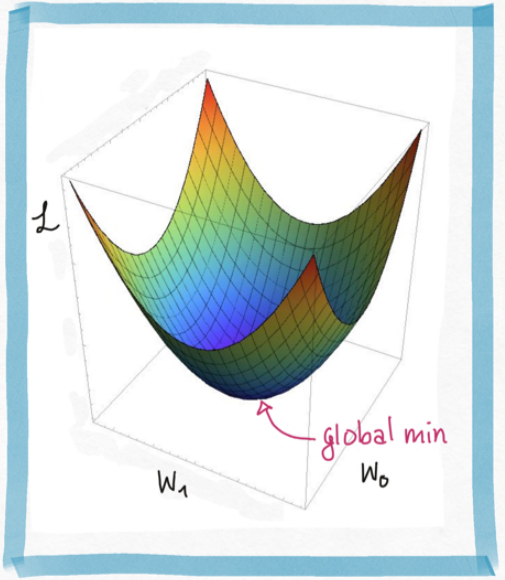
is zero.

At a stationary point where $\nabla \mathcal{L} = 0$, the loss function is neither increasing or decreasing

But not every stationary point is a global minima.

OPTIMIZING CONVEX FUNCTIONS:

For convex functions (bowl shaped, facing up), the global minimum happens at the stationary point.



To check that a function $\mathcal{L}$ is convex, we check that its Hessian, the matrix of second partial derivatives, satisfies a special property called positive semi-definiteness.

MODEL TRAINING: LINEAR REGRESSION

Given a dataset $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, we find the optimal linear model $y = w_1 x + w_0$:

Find $w_1^*, w_0^* = \underset{w_1, w_0}{\operatorname{argmin}} \mathcal{L}(w_1, w_0) = \underset{w_1, w_0}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} (y_n - (w_1 x_n + w_0))^2$

1. Compute the gradient $\nabla \mathcal{L} = \left[\frac{\partial \mathcal{L}}{\partial w_1}, \frac{\partial \mathcal{L}}{\partial w_0}\right]$

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w_1} = \frac{1}{N} \sum_{n=1}^{N} 2(y_n - (w_1 x_n + w_0))(-x_n) \\ \frac{\partial \mathcal{L}}{\partial w_0} = \frac{1}{N} \sum_{n=1}^{N} 2(y_n - (w_1 x_n + w_0))(-1) \end{cases}$$

2. Solve for stationary points: $\nabla \mathcal{L} = 0$

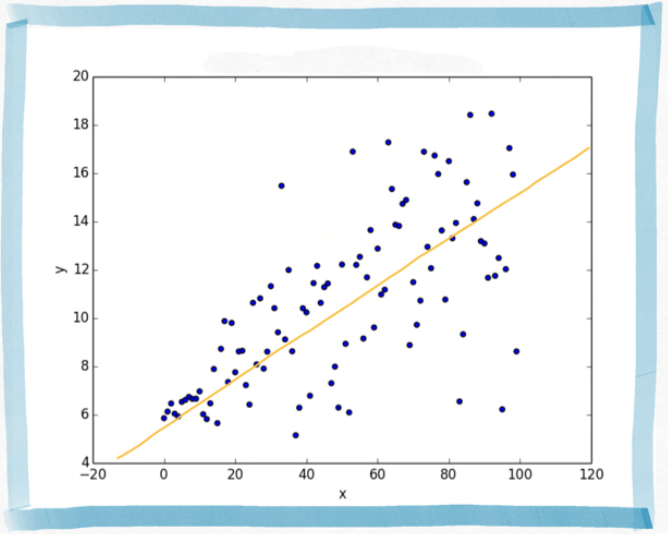
$$\begin{cases} 0 = \frac{1}{N} \sum_{n=1}^{N} 2(y_n - (w_1 x_n + w_0))(-x_n) \\ 0 = \frac{1}{N} \sum_{n=1}^{N} 2(y_n - (w_1 x_n + w_0))(-1) \end{cases} \quad \begin{cases} w_1^* = \frac{N \sum_n x_n y_n - \sum_n x_n \sum_n y_n}{N \sum_n x_n^2 - (\sum_n x_n)^2} \\ w_0^* = \frac{\sum_n y_n \sum_n x_n^2 - \sum_n x_n \sum_n x_n y_n}{N \sum_n x_n^2 - (\sum_n x_n)^2} \end{cases}$$

3. The mean squared error loss $\mathcal{L}$ is convex so the stationary point is the global minimum!

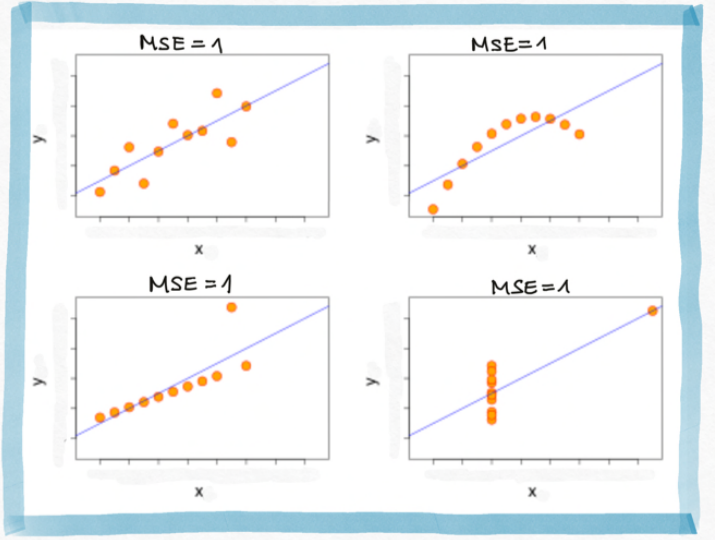
EVALUATION

EVALUATION: TRAINING ERROR

Just because we found the model $y = w_1^* x + w_0^*$ that minimizes the mean squared error it doesn't mean that it's a good model.



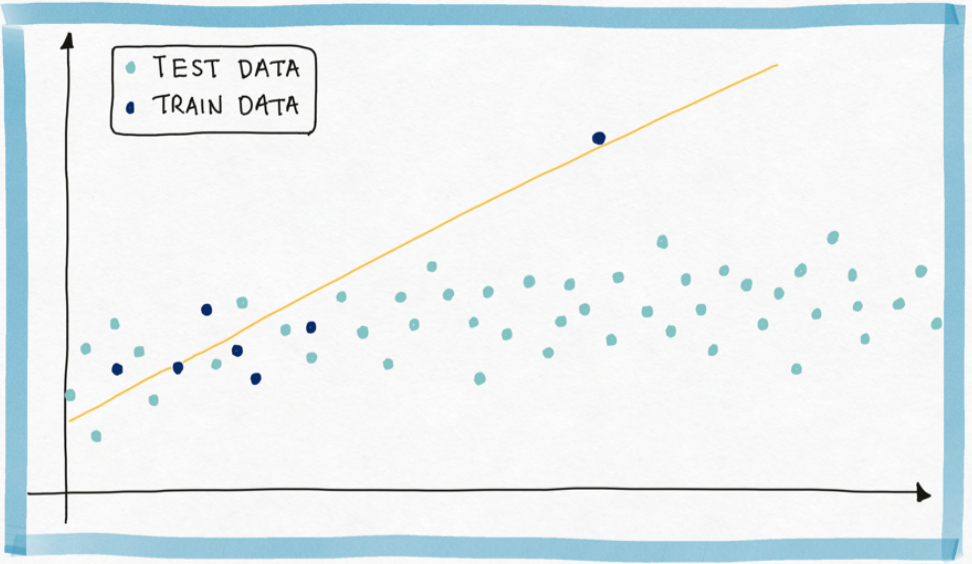
The MSE of our trained model is high due to noise in the data.



The MSE of all four models are the same. But the models are not equally good.

EVALUATION: TEST ERROR

We need to evaluate the fitted model on new data, data that the model did not train on. This new data is called test data.



The training MSE is 1.5

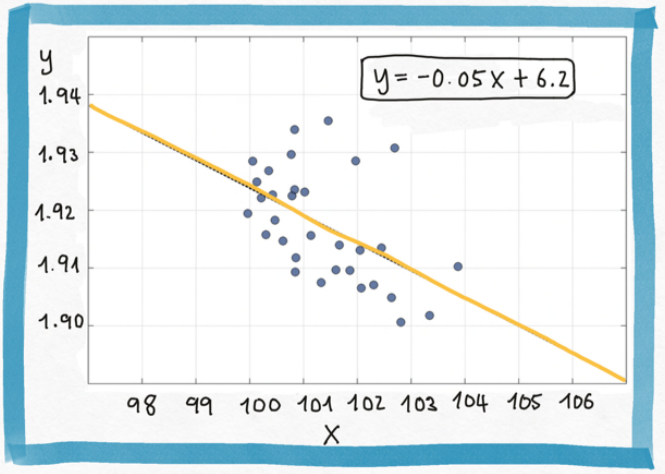
The test MSE is 10.87

The training data contains an outlier, which skewed our model.

Fitting to the meaningless patterns in the training data is called overfitting.

EVALUATION: MODEL INTERPRETATION

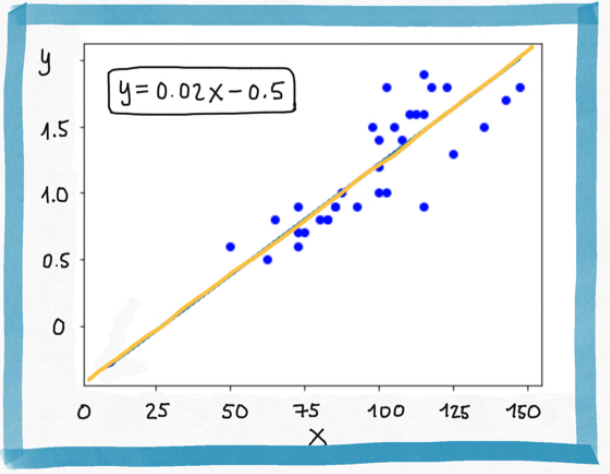
For linear models it's important to interpret the parameters.



The MSE of this model is very small.

But the slope is −0.05 million/m².

This mean that the larger the house the cheaper the price.



The MSE of this model is very small.

But the y-intercept is −0.5 million.

This mean that a very small (10m²) house will have a negative price (−0.3 million).