Advanced Institute for Artificial Intelligence

Machine Learning - Regression

Raphael Cóbe

raphaelmcobe@gmail.com

Regression

Introduction

Links and References

- Book: *Artificial Intelligence: A Modern Approach*
- Book: *An Introduction to Statistical Learning*
- Book: *The Elements of Statistical Learning*
- Regression Analysis Tutorial
- *Scikit-learn* Tutorial
- Video Lecture (Andrew Ng): Univariate Linear Regression
- Video Lecture (Stanford): Linear Regression and Gradient Descent

Regression: Definition

Regression: What is it?

- Tries to **predict** numerical values directly **from attributes** of a new example.

Examples

- **Predict** tomorrow's temperature **from** atmospheric conditions.
- **Estimate** the price of a house **from** its size.

More Formally.....

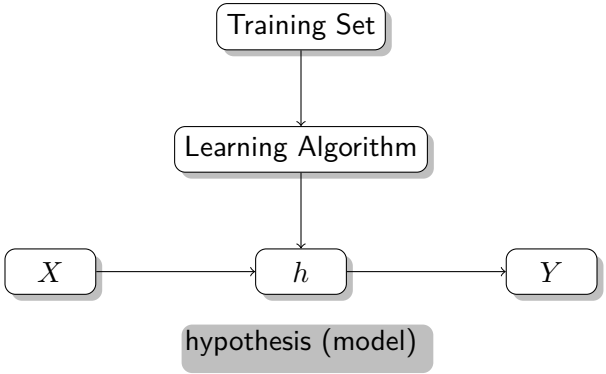
Problem definition

- Approximate a quantitative variable $Y \in \mathbb{R}$ (*response*)
- From predictor variables $X_1, \ldots, X_n \in \mathbb{R}$
  - When $n = 1$: Simple or univariate regression
  - When $n > 1$: Multivariate regression
- **Objective:** Find the function h (*hypothesis*):

$$Y \approx h(X_1, \ldots, X_n)$$

Strategy:

- Use a set of examples (*dataset*) where the correct response is known to "learn" a model.



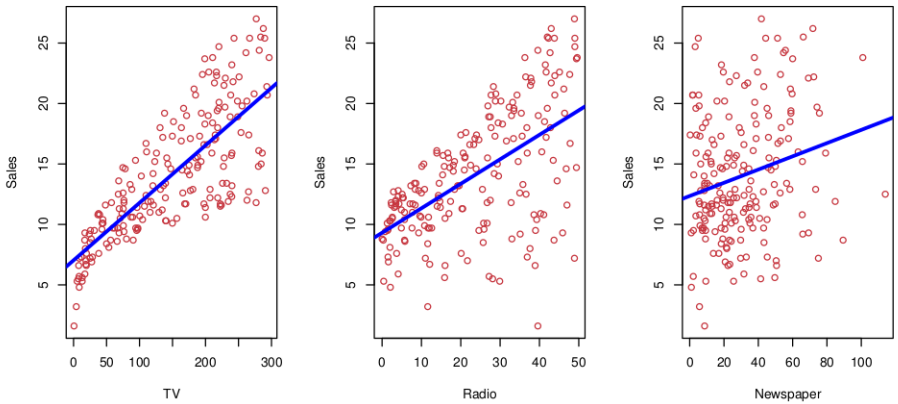
Learning the Model

Ideally, the algorithm for learning the model should:

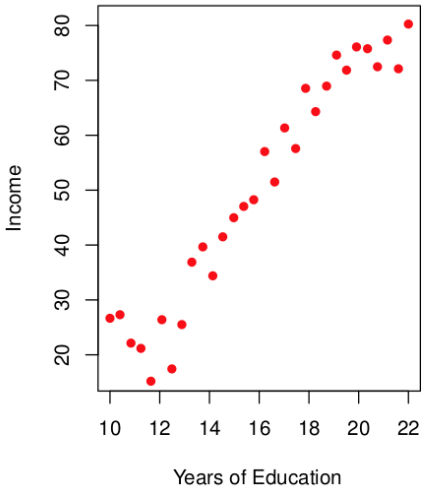
- Be able to reconstruct the modeled phenomenon as accurately as possible (except for a quantifiable error)
- Require as little data as possible for learning
- Represent the model as simply as possible (Occam's Razor)

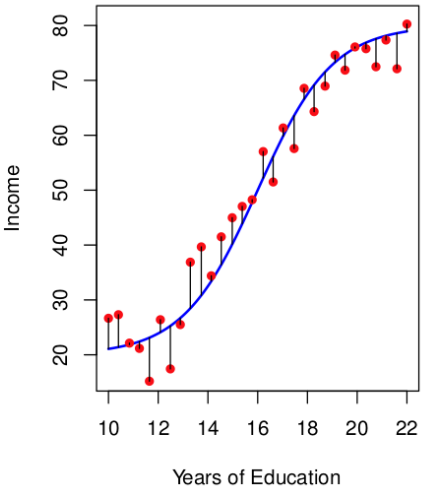
- There is no "correct answer" for all problems
- There are many types of models (linear models, trees, neural networks, etc.)

Sales volume as a function of advertising budget in different media



Income as a function of education

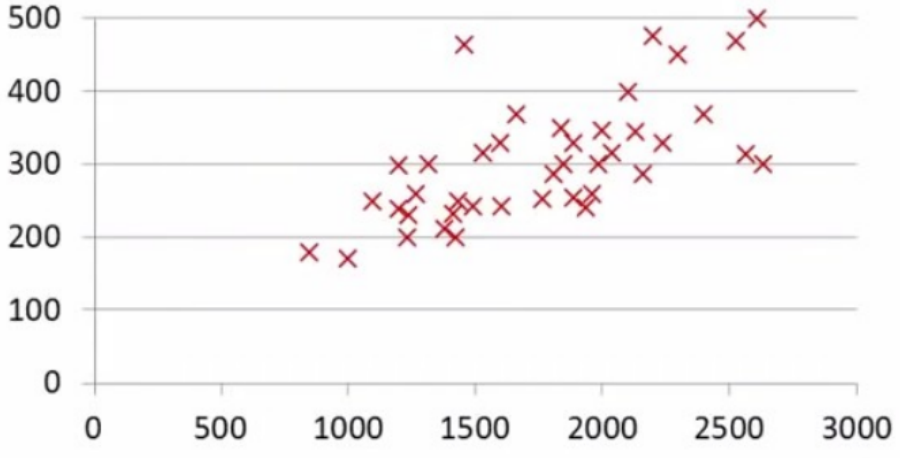




The Training Set can be visualized as a table

Size in sq ft^2	Price ($) in 1000's
2104	460
1416	232
1534	315
852	178
...	...

Table: Housing price by size in Portland (OR)


500
400
300
200
100
0
0
500
1000
1500
2000
2500
3000

K-Nearest Neighbors Regression

KNN Regression (*K-Nearest Neighbours*)

Hypothesis

- Y can be estimated by basing on the k closest examples in the training base.

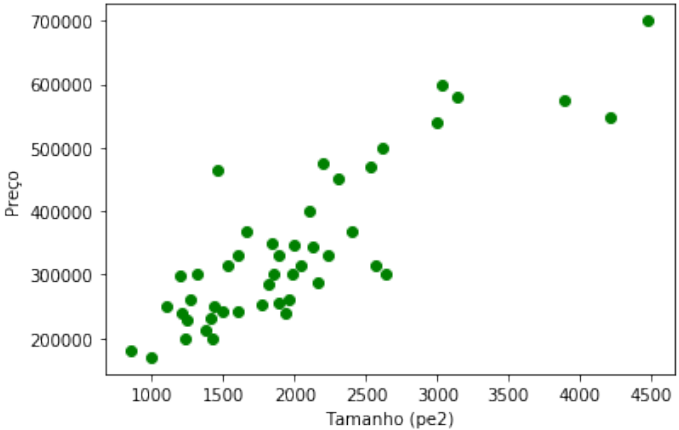
$$h_K(\boldsymbol{x}) = \sum_{j \in \mathcal{N}_K} w_j(\boldsymbol{x}, \boldsymbol{x}^j) y^j$$

- Where K is the number of neighbors, $\mathcal{N}_K$ is the set of samples present in the K-neighborhood and w_j is the weight of $\boldsymbol{x}$ relative to $\boldsymbol{x}^j$

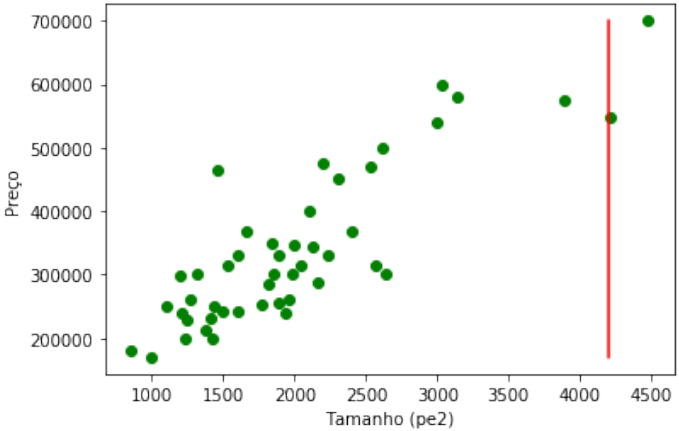
- Assuming all samples have the same weight:

$$h_K(\boldsymbol{x}) = \frac{1}{K} \sum_{j \in \mathcal{N}_K} y^j$$

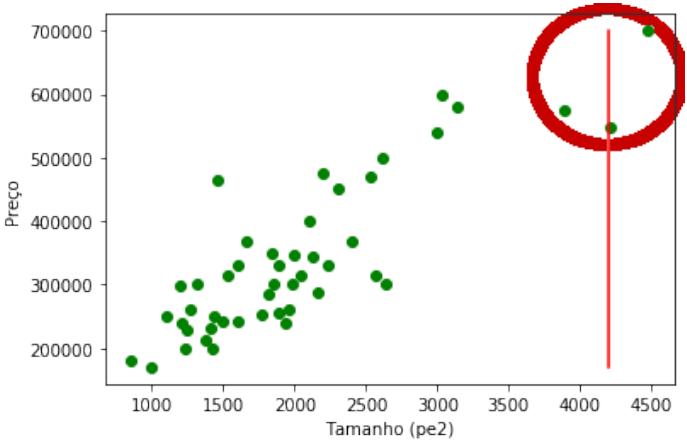
- No training process is required


700000
600000
500000
400000
300000
200000
Preço
1000
1500
2000
2500
3000
3500
4000
4500
Tamanho (pe2)

Imagine we want to predict the price of a house with size $= 4200$

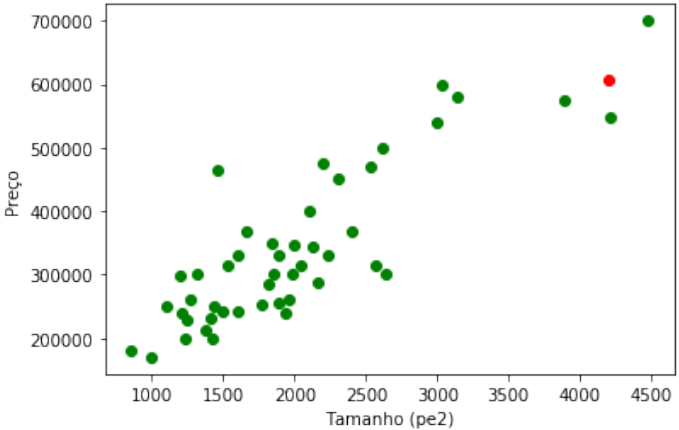


If $K = 3$, we must find the 3 examples closest to the desired value



The predicted value is defined as the average of these values

$$Y = \frac{1}{3}(699900 + 573900 + 549000) = 607600$$


700000
600000
500000
400000
300000
200000
Preço
1000
1500
2000
2500
3000
3500
4000
4500
Tamanho (pe2)

Parameters Required for KNN

- $K \rightarrow$ number of neighbors
- A distance metric to find the "closest" neighbors
- A way to define the weight for each example

Examples of models learned with KNN

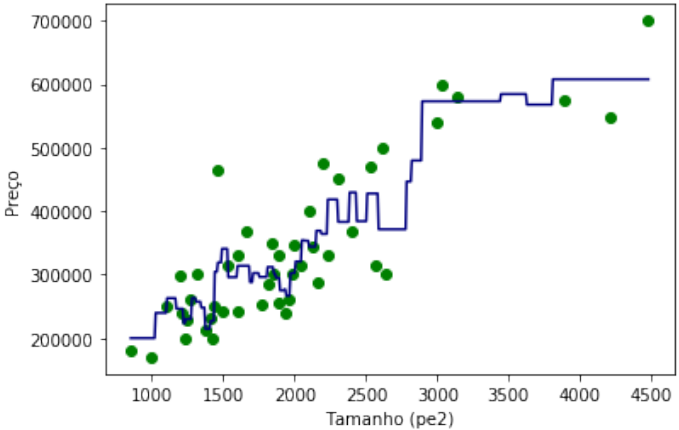


Figure: $K = 3$

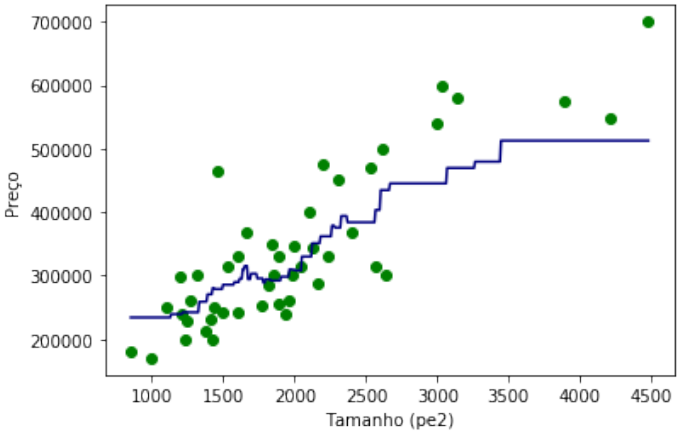

700000
600000
500000
400000
300000
200000
Preço
1000
1500
2000
2500
3000
3500
4000
4500
Tamanho (pe2)

Figure: $K = 10$

Advantages

- The learned model does not need to be linear
- No training phase
- Few parameters to be defined

Disadvantages

- Very costly inference process
- Sensitivity to noise and scale

Univariate Linear Regression

Case: Only one feature

Hypothesis

- Response variable y has a linear relationship with the attributes.

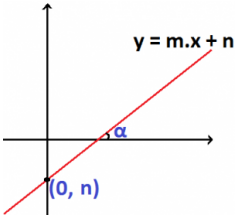
$$Y = \theta_0 + \theta_1 X$$

h is represented as a line:

$$h(\theta) = \theta_0 + \theta_1 X$$

Line Equation:

$$y = mx + n$$



- m = slope coefficient (indicates line inclination)
- n = linear coefficient (*intercept*)

$$y = mx + n$$

$$h(\theta; X) = \theta_0 + \theta_1 X^1$$

Objective: Find the best line (θ) according to the training data

And what would be the best line?

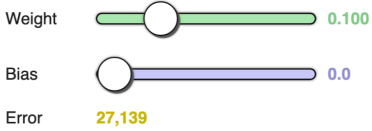
- Find the line h that passes as close as possible to all points

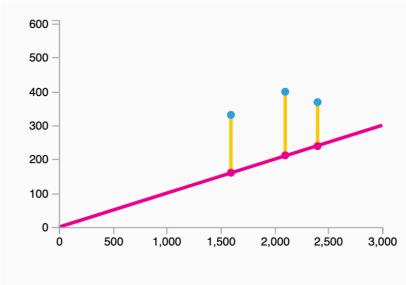
Residual

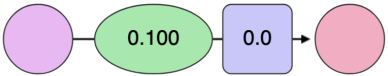
- Difference between the real Y value and the estimate $\widehat{Y} = h(\Theta; X)$

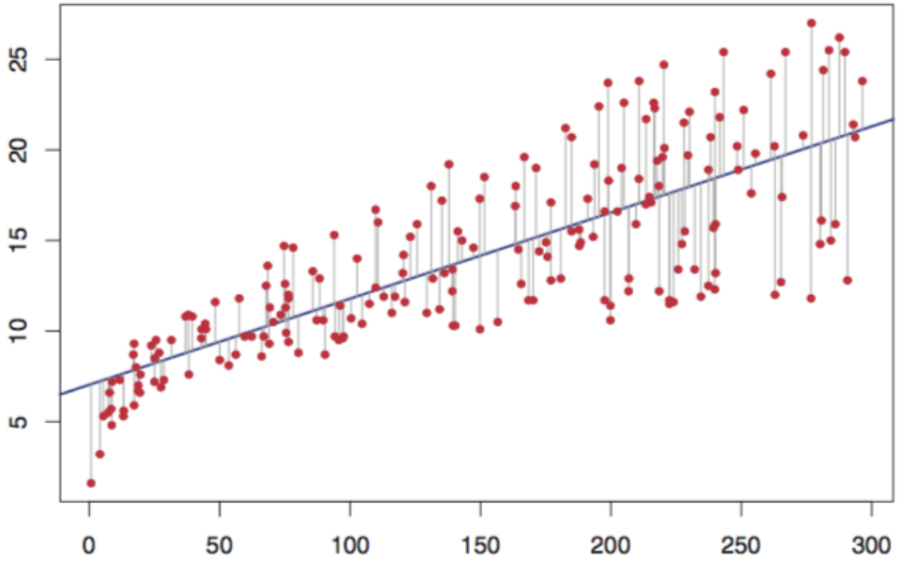
$$\epsilon_i = y_i - \widehat{y}_i$$

Finding the Weights Manually







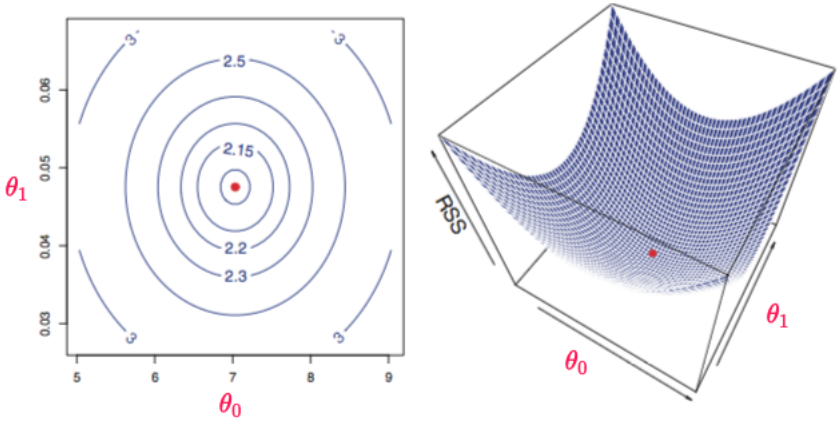

25
20
15
10
5
0
50
100
150
200
250
300

- One way to calculate θ_0 and θ_1 is to base it on the sum of squared residuals (RSS - ***R**esidual **S**um of **S**quares*)

$$J(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \epsilon_i^2$$

Cost Function Variation

Examine ϵ as a function of θ_0 and θ_1



Finding values of θ_0 and θ_1

Gradient Calculation

- The gradient of a vector is a generalization of the derivative and is represented by the vector operator ∇. This operation is used to minimize our cost function (**RSS**):
- Ordinary Least Squares method

Finding values of θ_0 and θ_1

The problem with the analytical solution

- The analytical solution in its vector representation has the form:

$$\Theta = (X^T X)^{-1} X^T Y$$

- Disadvantages of using the analytical solution:
  - $X^T X$ is not always invertible;
  - The complexity of calculating the inverse is of order $O(n^3)$:
    - If the number of features is high, it can become **computationally expensive**;
    - Very high memory consumption

Finding values of θ_0 and θ_1

Inverse calculation

Problem

Imagine a dataset containing 10^5 features and 10^6 observations, in this case $X^T X$ would have $10^5 \times 10^5$ floating points which, at 8 bytes per number would give **80 gigabytes**. The inverse calculation would then consume on the order of $O(n^3)$ (**80 kilo yottabytes!!!**);

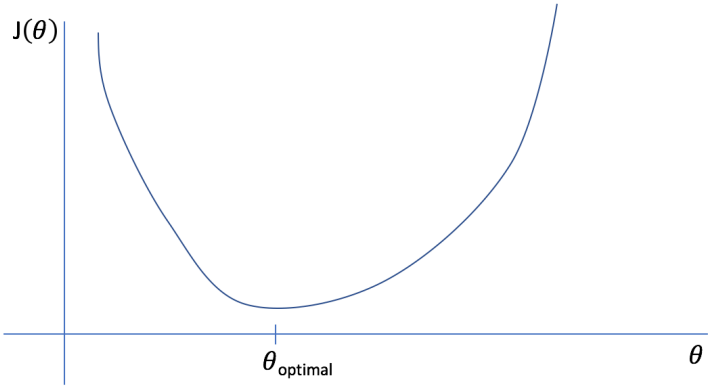
Finding values of θ_0 and θ_1

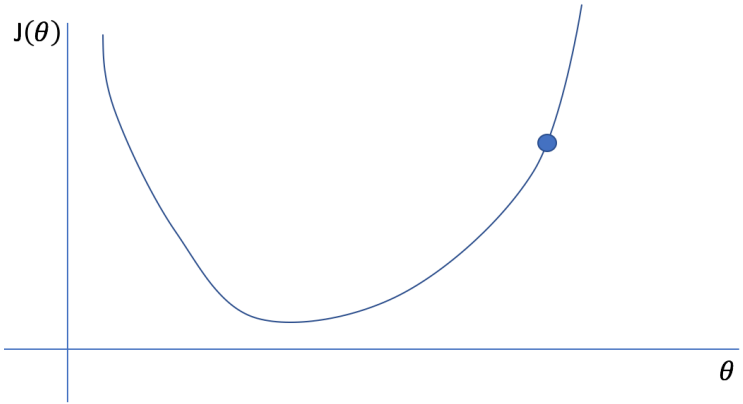
Gradient Descent Technique

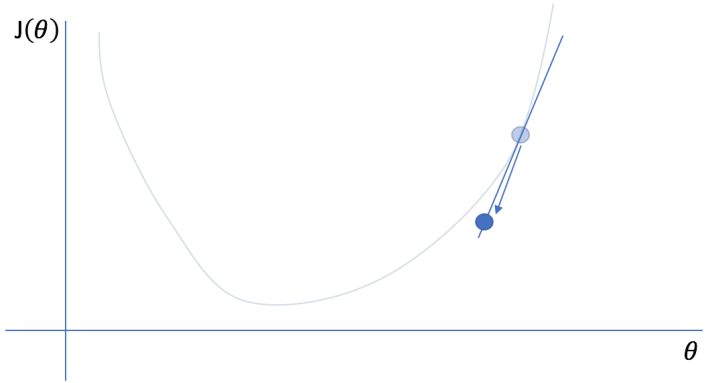
- Iterative calculation of the matrix Θ with:

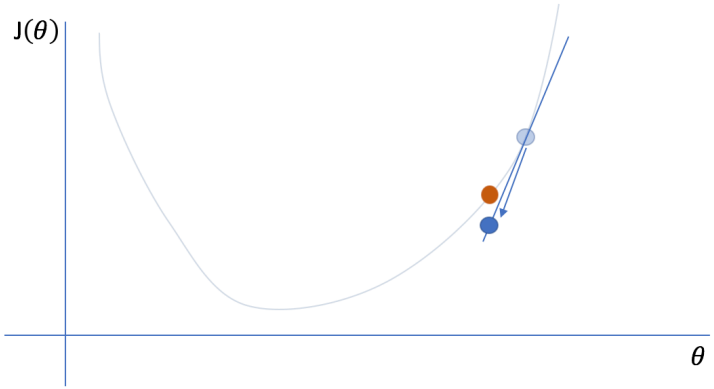
$$\begin{aligned} \theta_0^{(t+1)} &= \theta_0^{(t)} - \alpha \frac{\partial RSS}{\partial \theta_0} \\ \theta_1^{(t+1)} &= \theta_1^{(t)} - \alpha \frac{\partial RSS}{\partial \theta_1} \end{aligned}$$

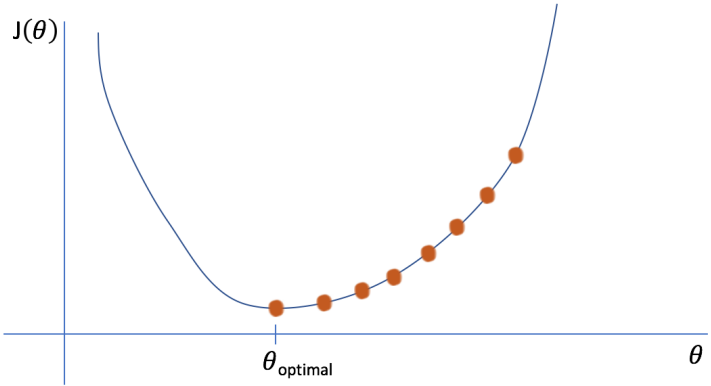
- Where α is the Learning Rate, i.e. the step size toward the minimum cost value;


J(θ)
θ optimal
θ

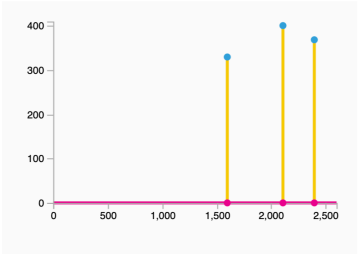

J(θ)
θ

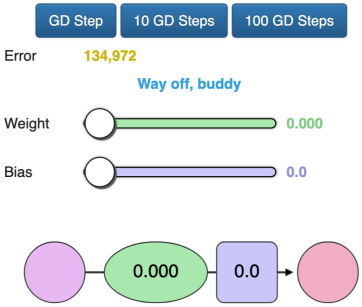

J(θ)
θ


J(θ)
θ

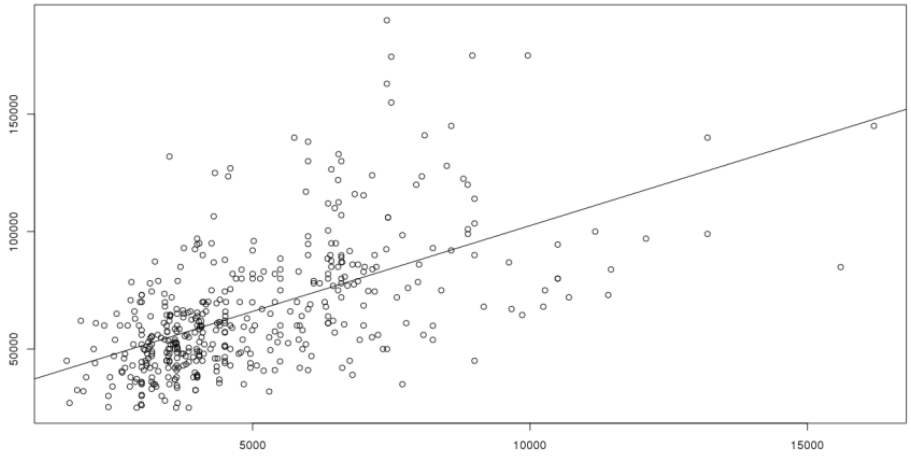

J(θ)
θ optimal
θ

Visualizing Gradient Descent in Action

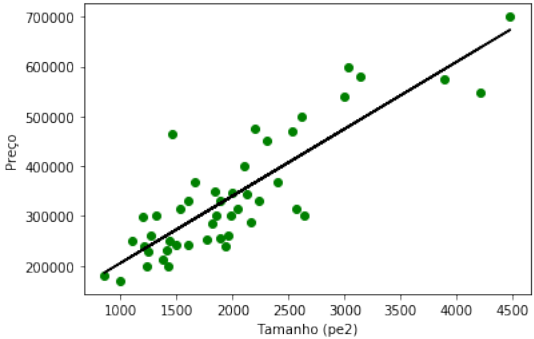




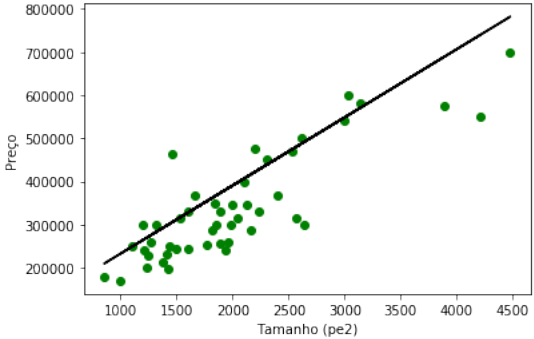
How to know if the result was good?



RSS itself can be used



(a) Model 1: RSS $= 1.93 \times 10^{11}$



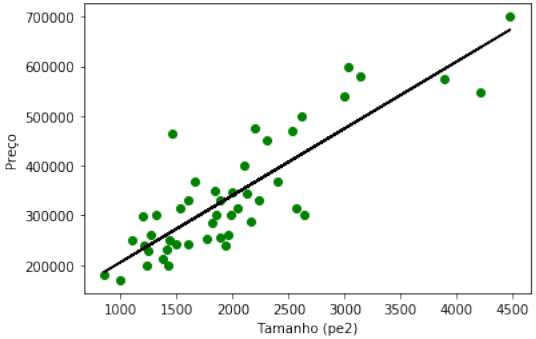
(b) Model 2: RSS $= 3.28 \times 10^{11}$

R^2 Statistic

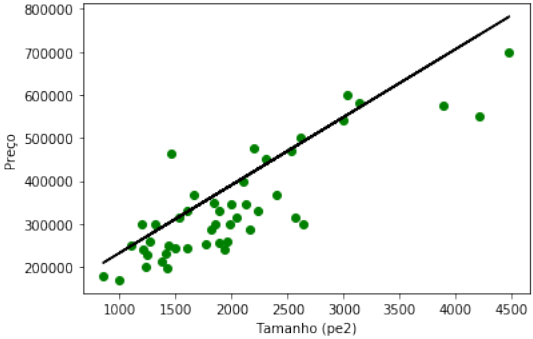
- Measures the proportion of the variability of Y that can be explained by X.

$$TSS = \sum_j (y^j - \bar{y})^2 \qquad RSS = \sum_j (y^j - \widehat{y})^2$$

$$R^2 = \frac{TSS - RSS}{TSS}$$



(c) Model 1: $R^2 = 0.63$



(d) Model 2: $R^2 = 0.54$

Linear Regression

Multivariate Regression

- In most practical problems, using only one attribute **is not enough** to estimate the response
- In this case, Linear Regression should estimate a Hyperplane as model h.

Linear Regression with Multiple Variables

Hypothesis

$$Y = h(\Theta; X) = \theta_0 + \theta_1 X_1 + \cdots + \theta_n X_n$$

Size in sq ft^2	number of bedrooms	Price ($) in 1000's
2104	3	460
1416	3	232
1534	3	315
852	2	178
...		...

- As in the univariate case, values of θ must be chosen based on the training set
- The least squares method also works for the multivariate case
- Another possibility to perform the learning is through the **Gradient Descent** method

Gradient Descent

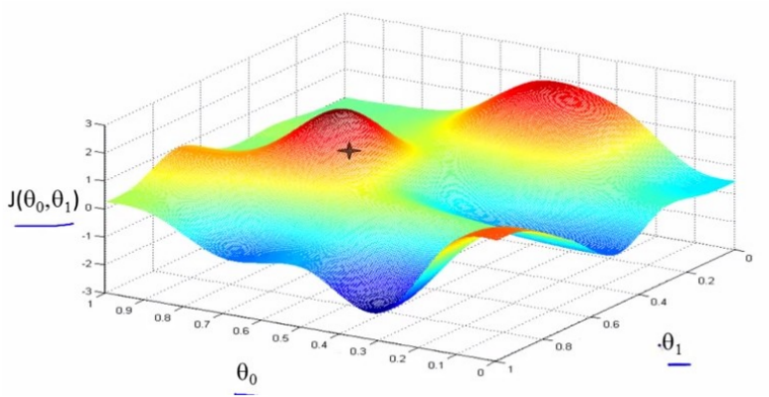
- Starting from the **Mean Squared Error** cost function

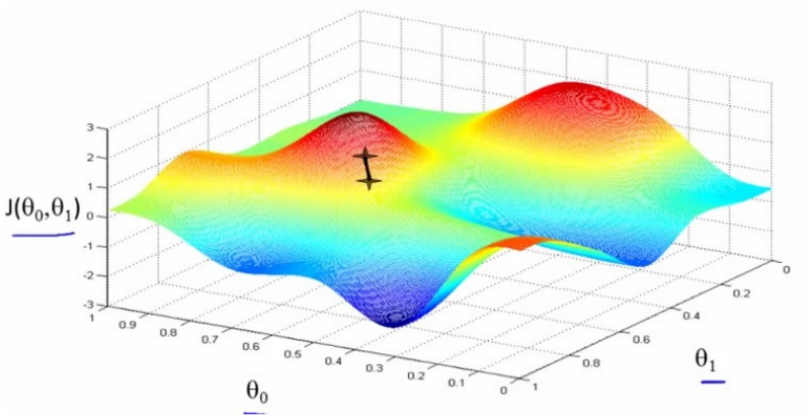
$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{n} (y_i - h(\theta; \boldsymbol{x}_i))^2$$

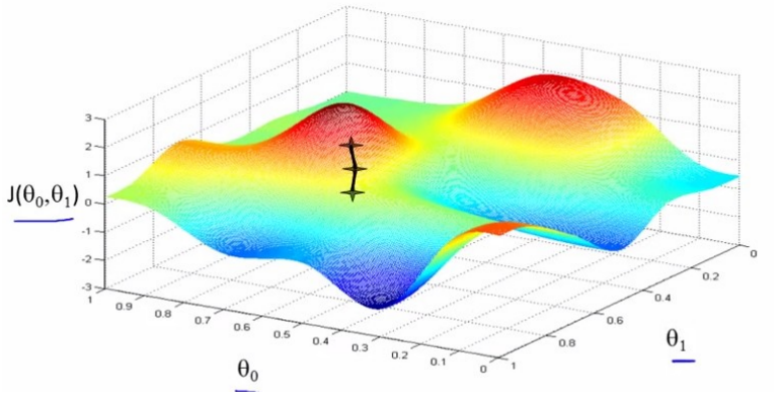
- Define parameters θ that minimize J

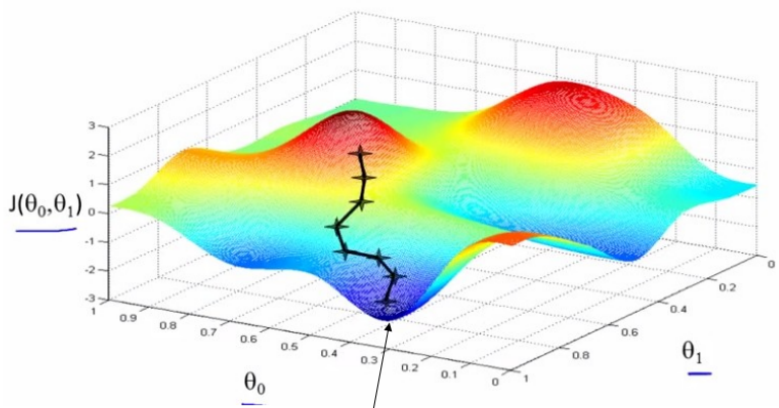
Gradient Descent Algorithm

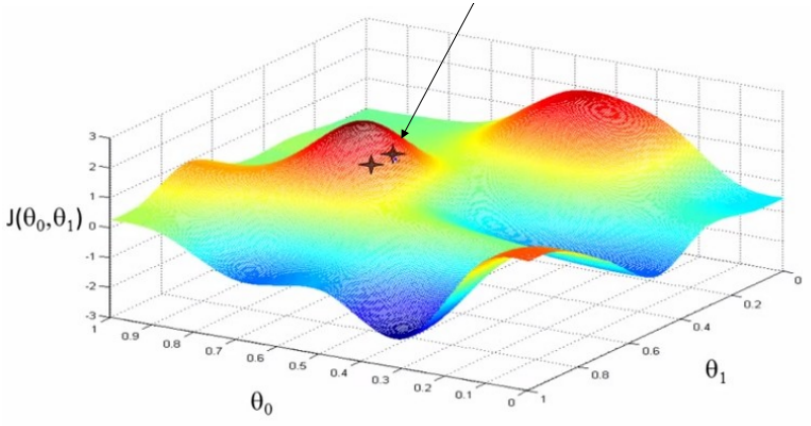
- Initialize θ randomly
- Modify values of θ (following the gradient), to reduce J until a minimum value is reached.

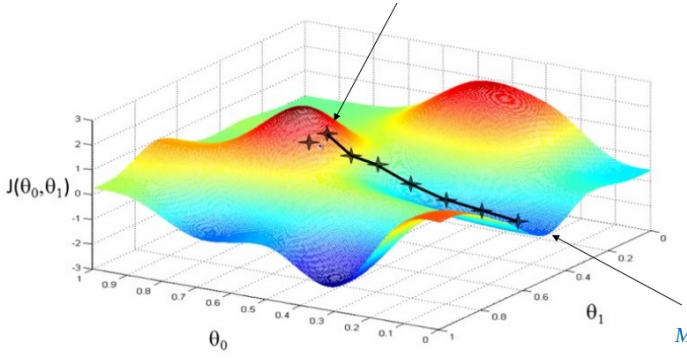

J(θ0,θ1)
θ0
θ1


J(θ0,θ1)
θ0
θ1


J(θ0,θ1)
3
2
1
0
-1
-2
-3
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
θ0
0
0.2
0.4
0.6
0.8
1
θ1


J(θ0,θ1)
θ0
θ1
Mínimo local


J(θ0,θ1)
3
2
1
0
-1
-2
-3
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
θ0
0
0.2
0.4
0.6
0.8
1
θ1



Mínimo local

Applying Gradient Descent

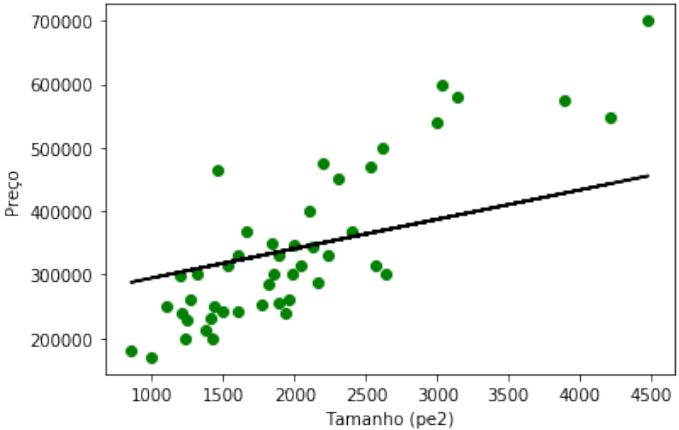


Figure: Repeating training **1** time. $RSS = 4.2 \times 10^{12}$, $R^2 = -5.76$

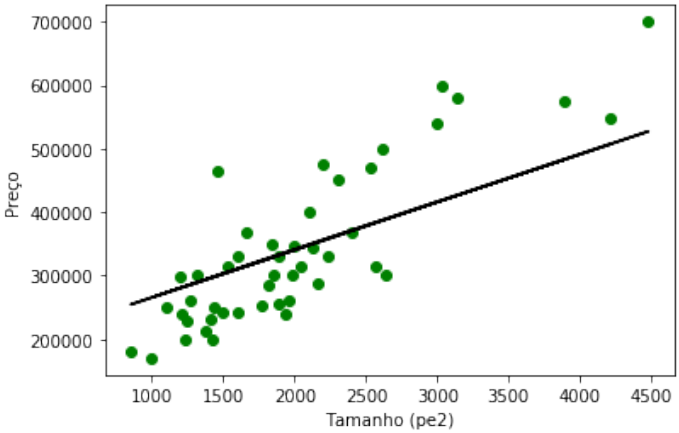


Figure: Repeating training **2** times. $RSS = 2.9 \times 10^{11}$, $R^2 = -0.79$

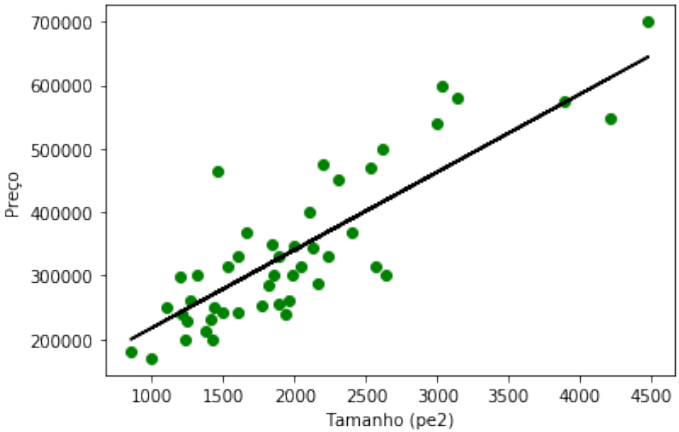


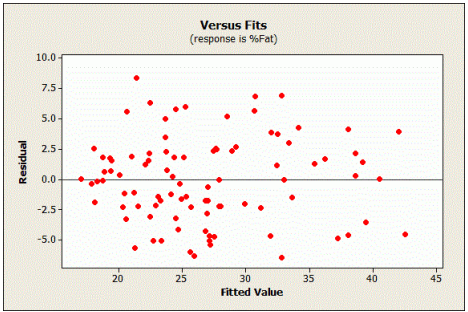
Figure: Repeating training **7** times. $RSS = 1.9 \times 10^{11}$, $R^2 = 0.55$

Residual Analysis

- The deterministic component of a regression model does such a good job of explaining the dependent variable that it leaves only the **intrinsically unexplainable** part of your study area to error;
  - Existence of non-randomness in the error: independent variables are not explaining everything they can
- Residual plots display the residual values on the y-axis and the predicted values on the x-axis;
- If they show undesirable patterns, we cannot trust the regression coefficients, i.e., the residuals are consistent with random error

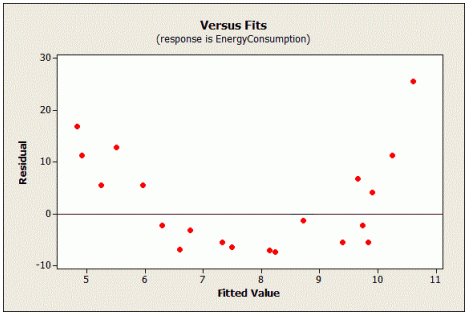
Residual Analysis

- Check if residuals are **randomly scattered around zero** for the entire range of fitted values



Residual Analysis

- When we have a pattern, unfortunately, some of the explanatory information has leaked into the supposedly random error;



F-Test

- F test:
  - The p value for each independent variable tests the null hypothesis that the variable has no correlation with the dependent variable.

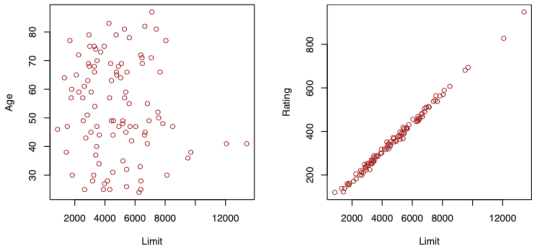
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	12833.9	4278.0	57.87	0.000
East	1	226.3	226.3	3.06	0.092
South	1	2255.1	2255.1	30.51	0.000
North	1	12330.6	12330.6	166.80	0.000
Error	25	1848.1	73.9		
Total	28	14681.9			

- Compare with significance level (1 - confidence interval)
  - Values smaller than the significance level

Multicollinearity

- Situation where two or more predictor variables are closely related to each other;
- Can represent problems in the regression context;
- Difficult to separate the individual effects of collinear variables on the response;
  - It may be difficult to determine how each one separately is associated with the response;



Multicollinearity

- A coefficient θ_i represents the average change in the dependent variable for each 1-unit change in an independent variable when **all other independent variables are held constant**.
  - You can change the value of one independent variable and **not the others**;
- When independent variables are correlated, it indicates that changes in one variable are associated with changes in another variable
  - The stronger the correlation, the more difficult it is to change one variable without changing another;

Problem

Imagine that you fit a regression model and the coefficient values, and even the signs, change drastically depending on the specific variables you include in the model.
It's a disconcerting feeling when slightly different models lead to very different conclusions.
You don't feel like you know the actual effect of each variable!

Multicollinearity

- These problems only affect the independent variables that are correlated
  - It is possible to have a model with severe multicollinearity and still some variables in the model may be completely unaffected
  - You don't always need to find a way to fix multicollinearity
- The variance inflation factor (VIF) identifies the correlation between independent variables and the strength of that correlation;
  - 1 indicates that there is no correlation between this independent variable and any others
  - Between 1 and 5 suggest that there is a moderate correlation, but it is not severe enough to warrant corrective measures
  - Values greater than 5 represent critical levels of multicollinearity

Linear Regression

Advantages

- Efficient learning
- Simple model to visualize and understand

Disadvantages

- Many real problems are not linear

Hands-on Exercises

Notebooks:

Multivariate Linear Regression