Introduction to Regression

Part C – Linear Models

CS109A Introduction to Data Science

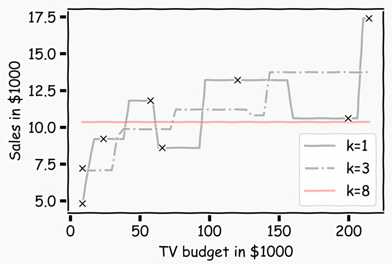
Pavlos Protopapas, Kevin Rader and Chris Tanner


I finally remember what Zoom meetings remind me of.

Lecture Outline

- Linear models
- Estimate of the regression coefficients
- Model evaluation
- Interpretation

Linear Models



Note that in building our kNN model for prediction, we did not compute a closed form for $\hat{f}$.

What if we ask the question:

"how much more sales do we expect if we double the TV advertising budget?"

Alternatively, we can build a model by first assuming a simple form of f:

$$\boxed{f(x) = \beta_0 + \beta_1 X}$$

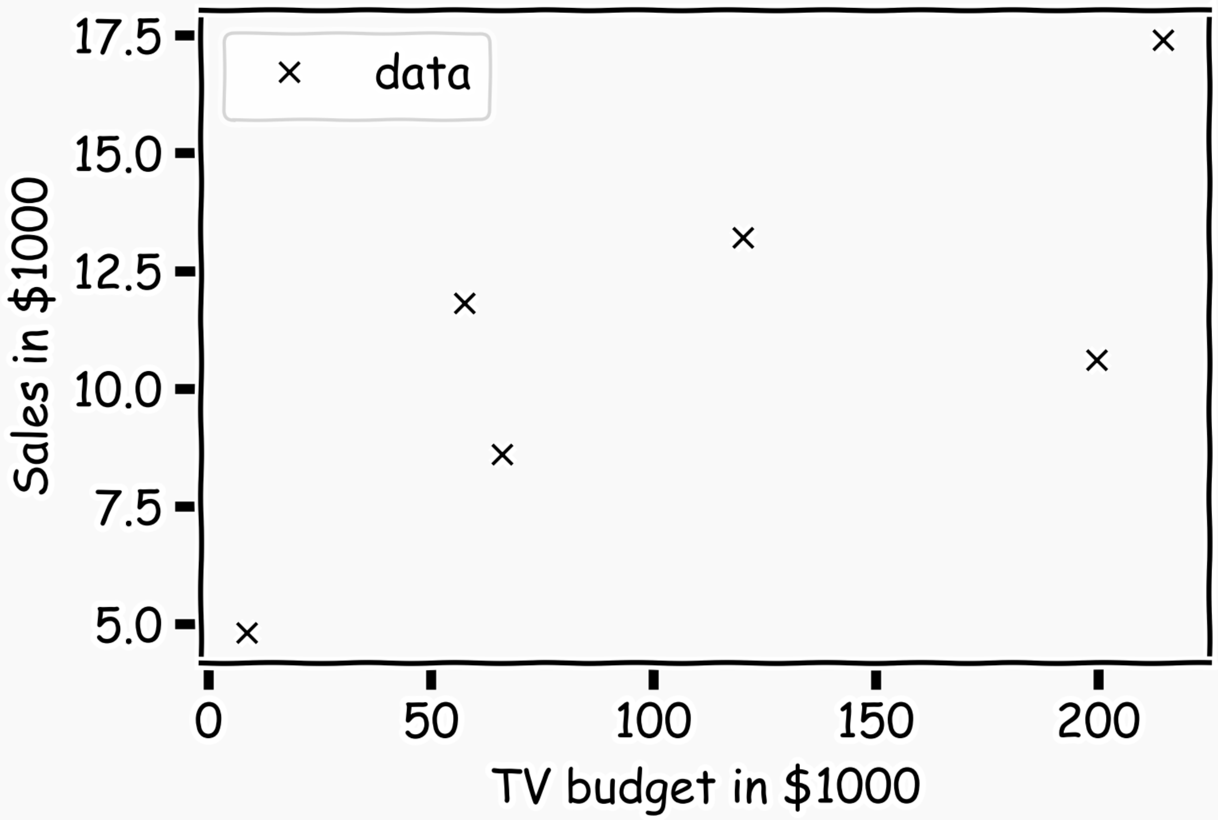
… then it follows that our estimate is:

$$\widehat{Y} = \widehat{f}(X) = \widehat{\beta_1} X + \widehat{\beta_0}$$

where $\hat{\beta}_1$ and $\hat{\beta}_0$ are **estimates** of β_1 and β_0 respectively, that we compute using observations.

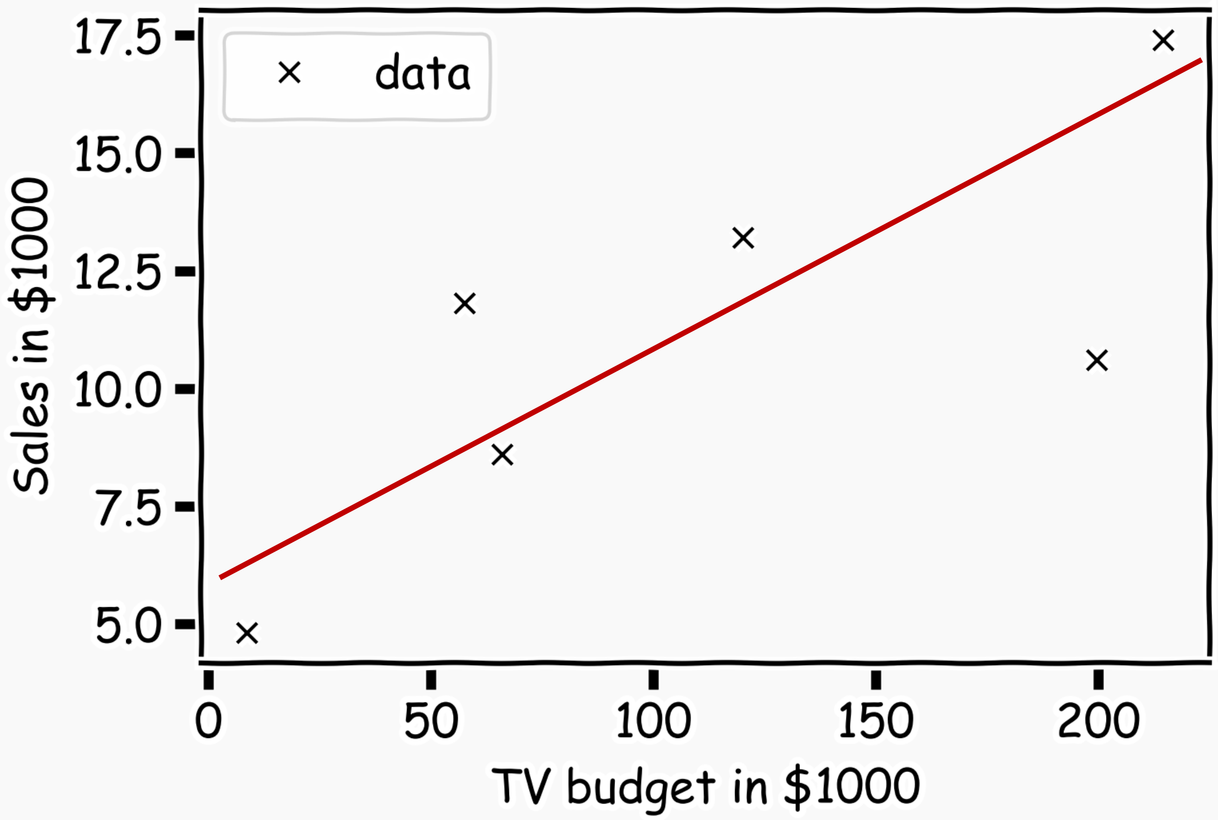
Estimate of the regression coefficients

For a given data set



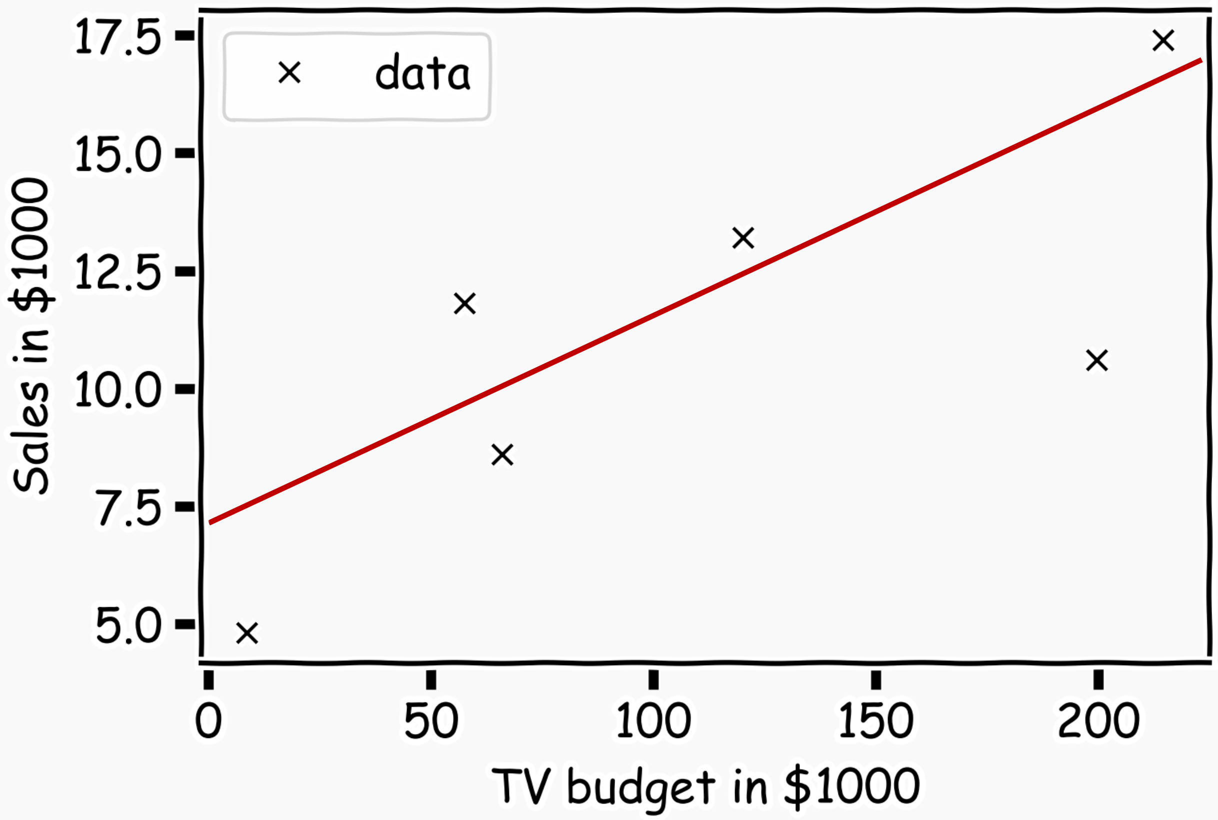
Estimate of the regression coefficients (cont)

Is this line good?



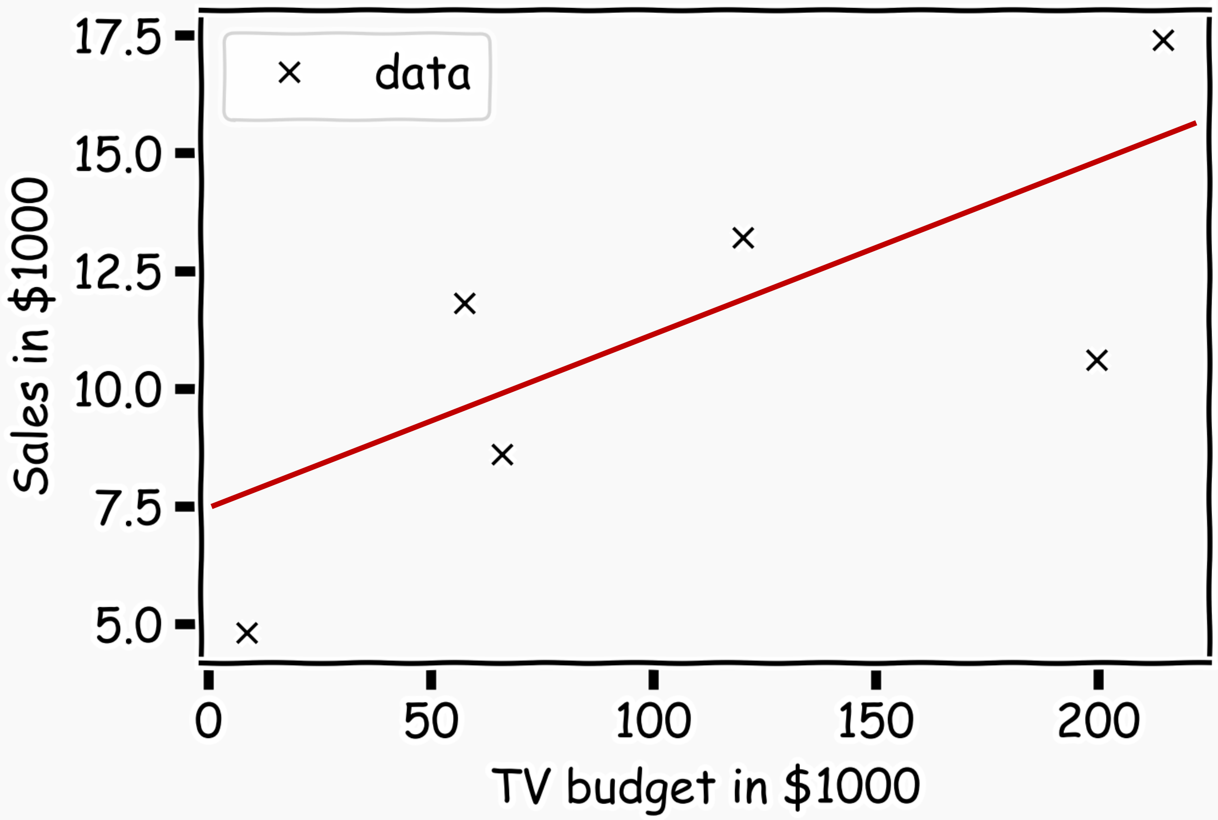
Estimate of the regression coefficients (cont)

Maybe this one?



Estimate of the regression coefficients (cont)

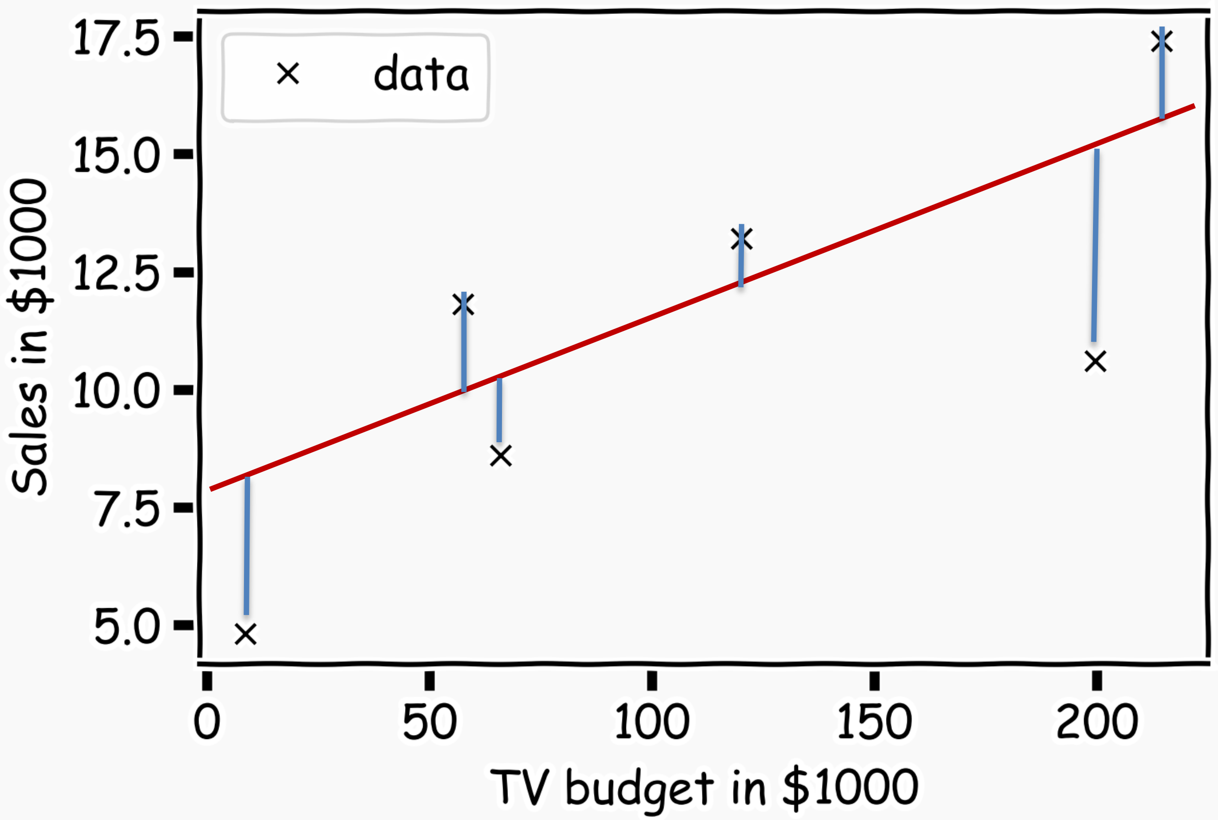
Or this one?



Estimate of the regression coefficients (cont)

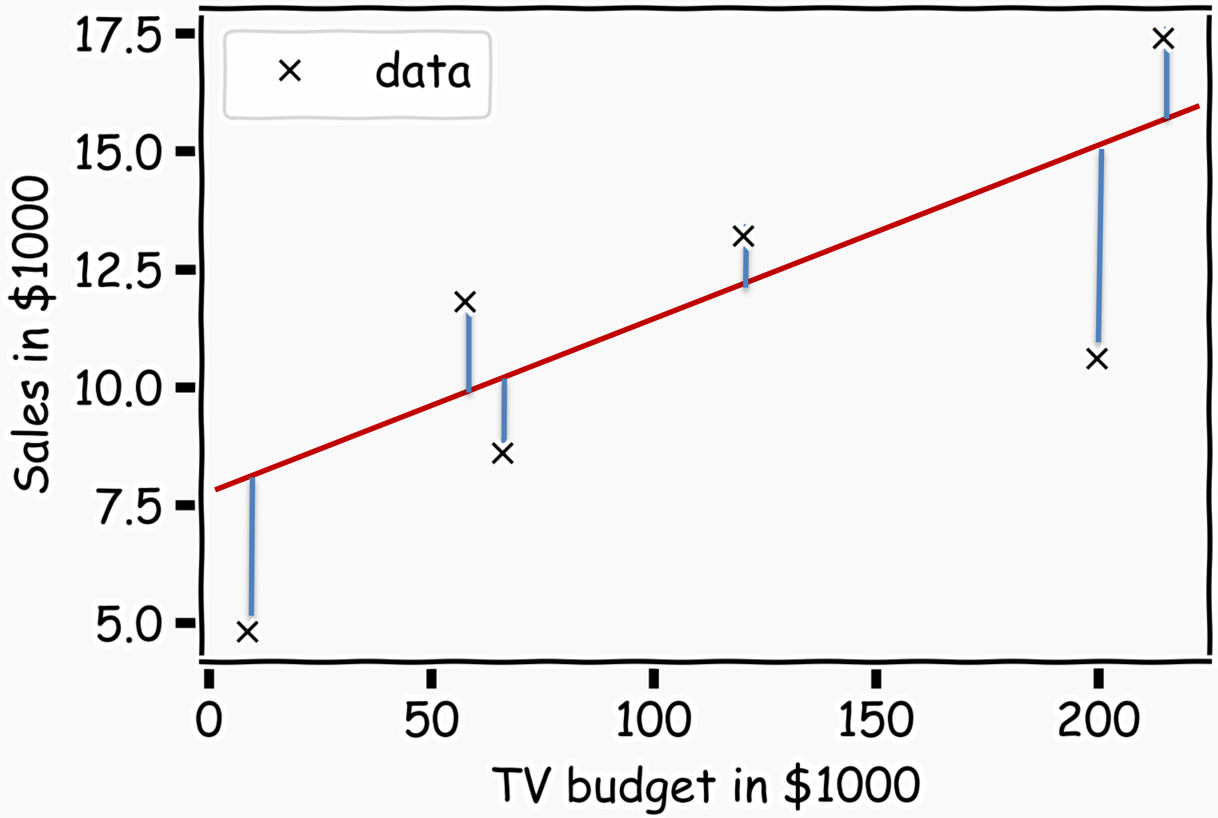
Question: Which line is the best?

For each observation (x_n, y_n), the absolute residual is calculate the residuals $r_i = |y_i - \hat{y}_i|$.



Loss Function: Aggregate Residuals

How do we aggregate residuals across the entire dateset?



1. Max Absolute Error
2. Mean Absolute Error
3. Mean Squared Error

Estimate of the regression coefficients (cont)

Again we use MSE as our **loss function**,

$$\boxed{L(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - (\beta_1 X + \beta_0)\right]^2 .}$$

We choose $\hat{\beta}_1$and $\hat{\beta}_0$ in order to minimize the predictive errors made by our model, i.e. minimize our loss function.

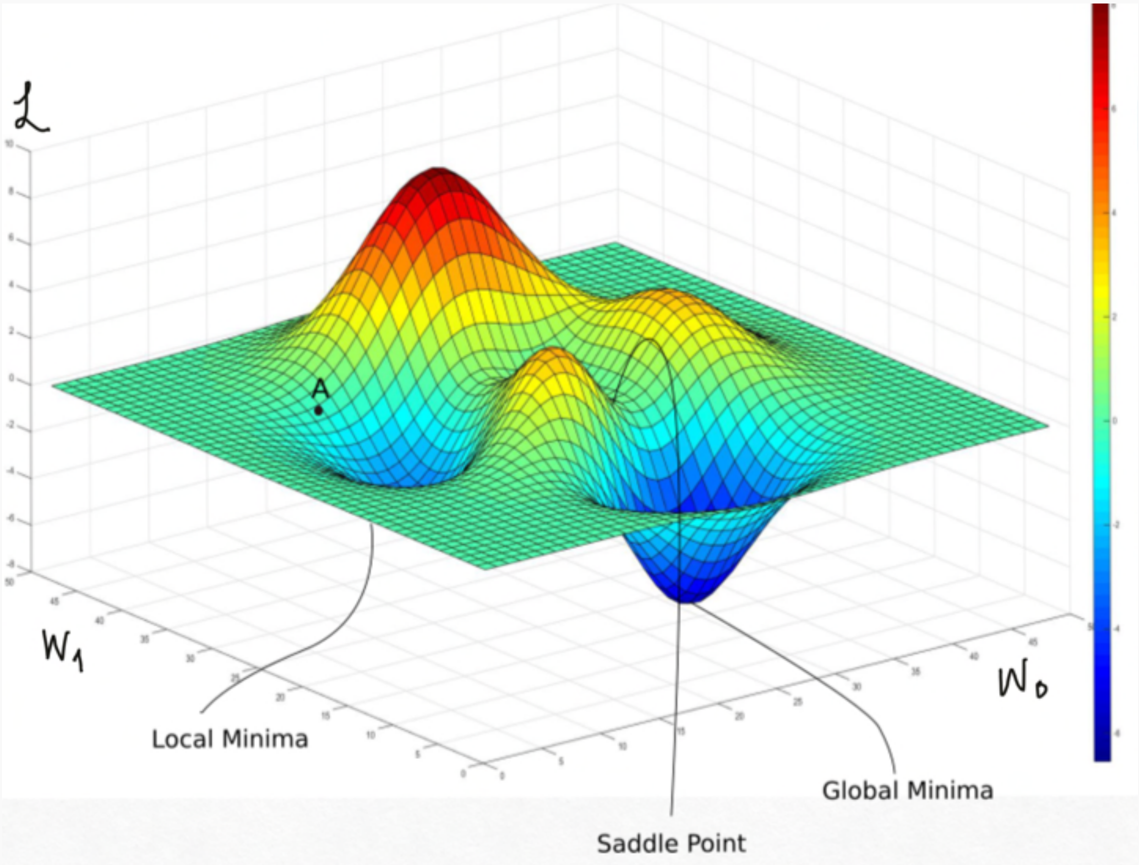
Then the optimal values for $\hat{\beta}_0$ and $\hat{\beta}_1$ should be:

$$\boxed{\widehat{\beta}_0, \widehat{\beta}_1 = \underset{\beta_0, \beta_1}{\operatorname{argmin}} L(\beta_0, \beta_1).}$$

WE CALL THIS **FITTING** OR **TRAINING** THE MODEL

Optimization

How does one minimize a loss function?



The global minima or maxima of $L(\beta_0, \beta_1)$ must occur at a point where the gradient (slope)

$$\nabla L = \left[\frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1}\right] = 0$$

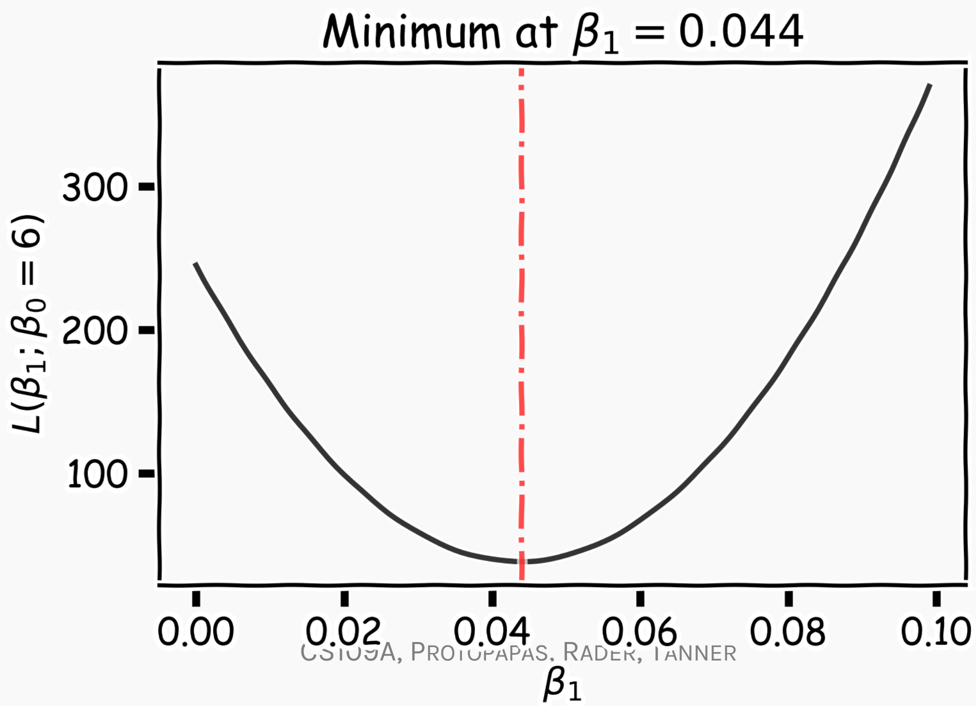
- Brute Force: Try every combination
- Exact: Solve the above equation
- Greedy Algorithm: Gradient Descent

Optimization: Estimate of the regression coefficients

Brute force

A way to estimate $\text{argmin}_{\beta_0,\beta_1} L$ is to calculate the loss function for every possible β_0 and β_1. Then select the β_0 and β_1 where the loss function is minimum.

E.g. the loss function for different β_1 when β_0 is fixed to be 6:



Very **computationally expensive** with many coefficients

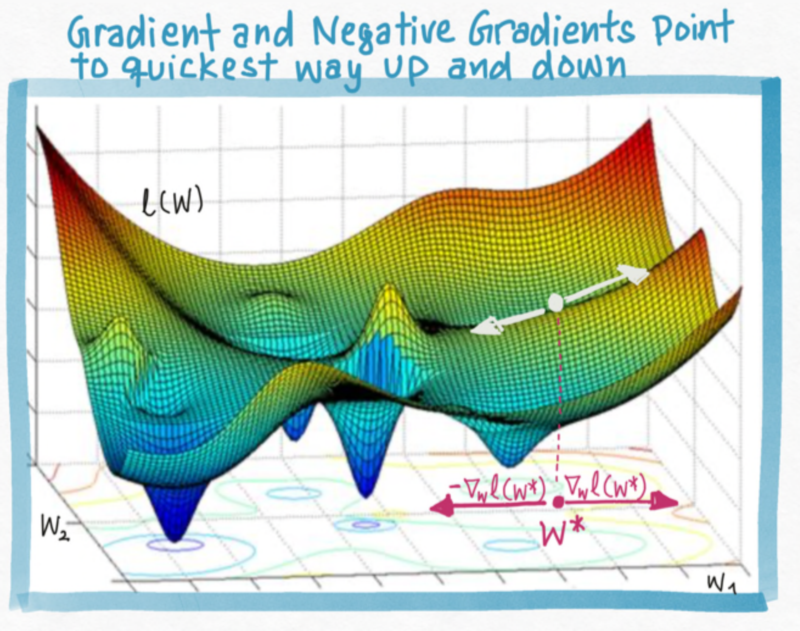
Gradient Descent

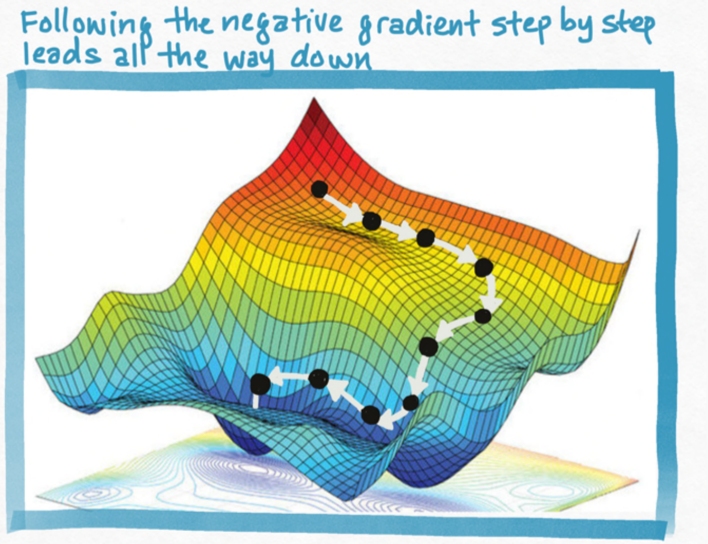
When we can't analytically solve for the stationary points of the gradient, we can still exploit the information in the gradient.

The gradient ∇L at any point is the direction of the steepest increase. The negative gradient is the direction of steepest decrease.

By following the –ve gradient, we can eventually find the lowest point.

This method is called Gradient Descent





Estimate of the regression coefficients: analytical solution

Take the gradient of the loss function and find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ where the gradient is zero: $\nabla L = \left[\frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1}\right]$=0

This does not usually yield to a close form solution. However for linear regression this procedure gives us explicit formulae for $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\boxed{\begin{gathered}\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}\end{gathered}}$$

where $\bar{y}$ and $\bar{x}$ are sample means.

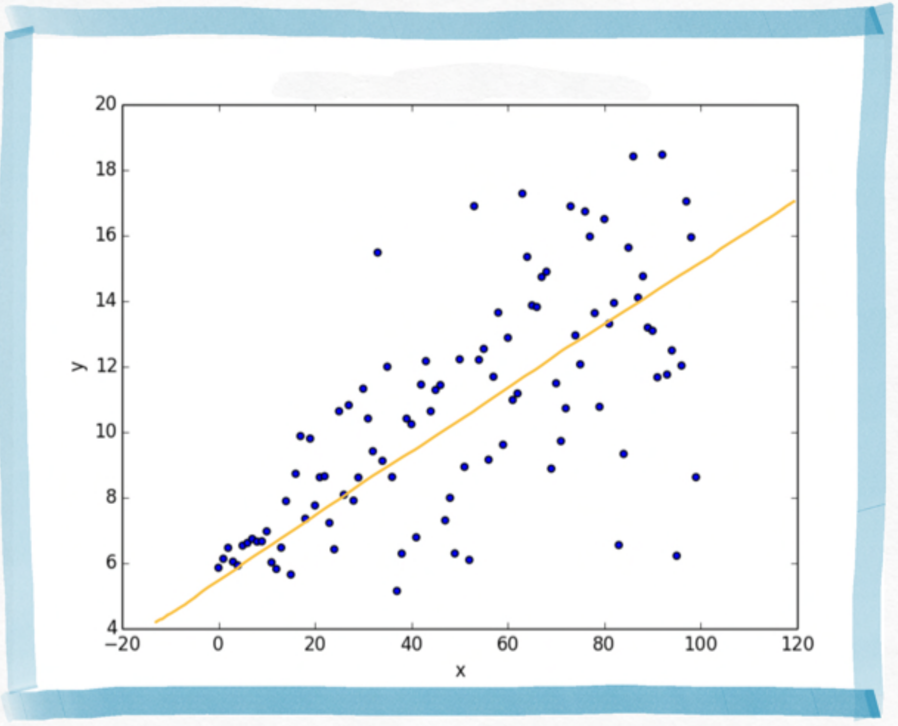
The line:

$$\boxed{\widehat{Y} = \widehat{\beta}_1 X + \widehat{\beta}_0}$$

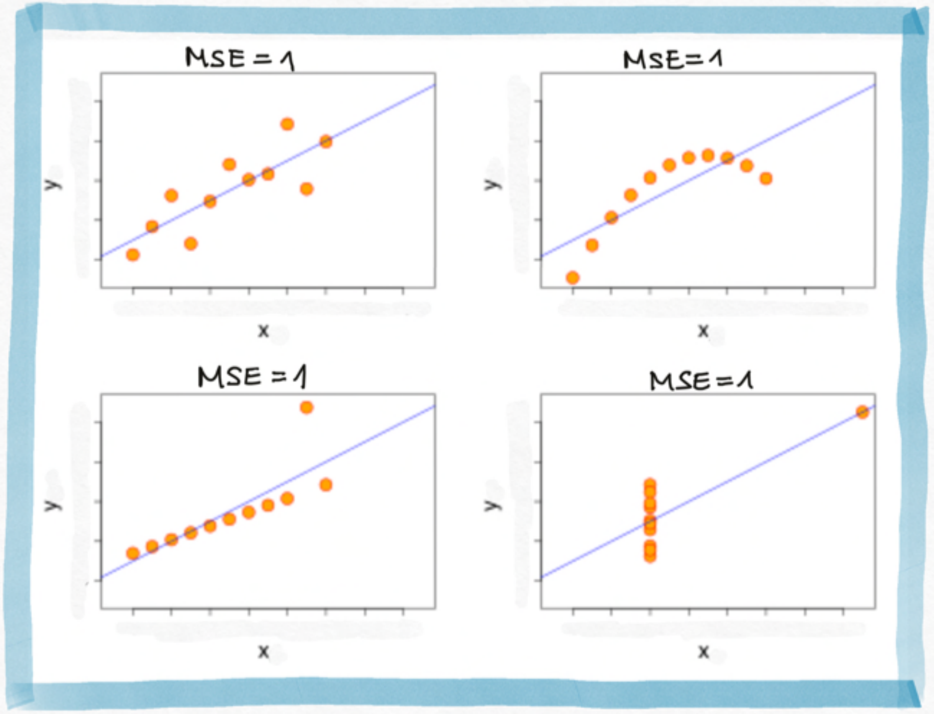
is called the **regression line**.

Evaluation: Training Error

Just because we found the model that minimizes the squared error it doesn't mean that it's a good model. We investigate the R2 but also:



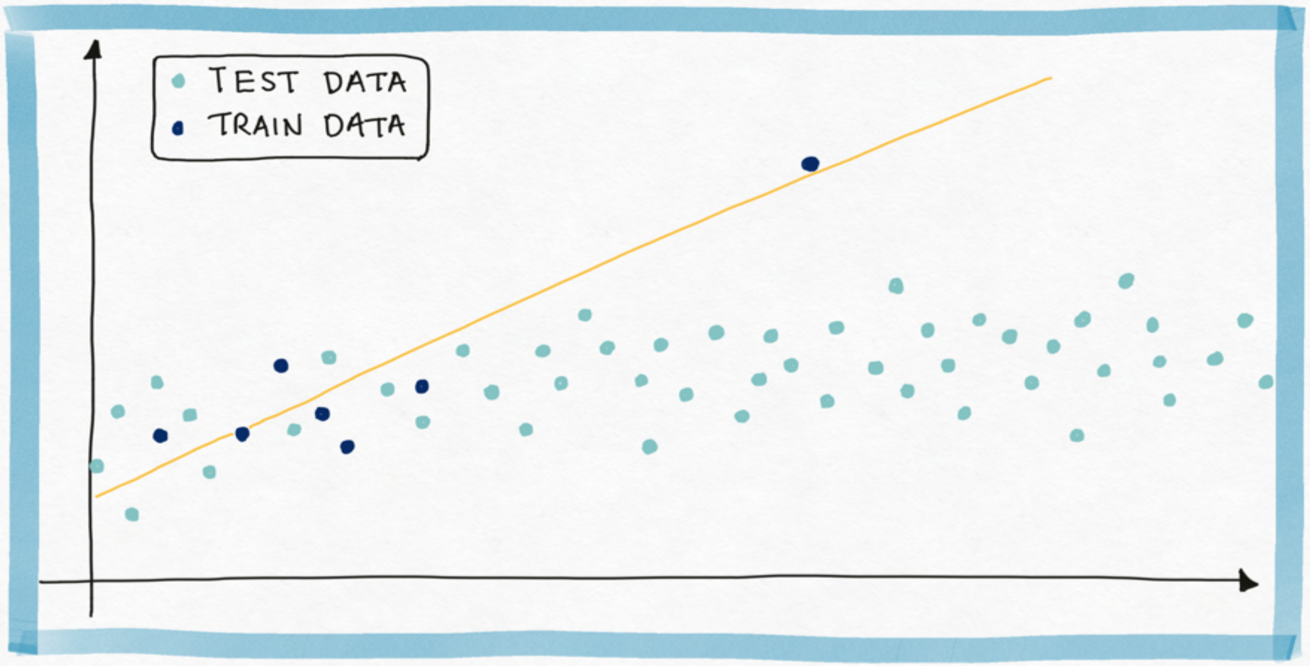
The MSE is high due to noise in the data.



The MSE is high in all four models but the models are not equal.

Evaluation: Test Error

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



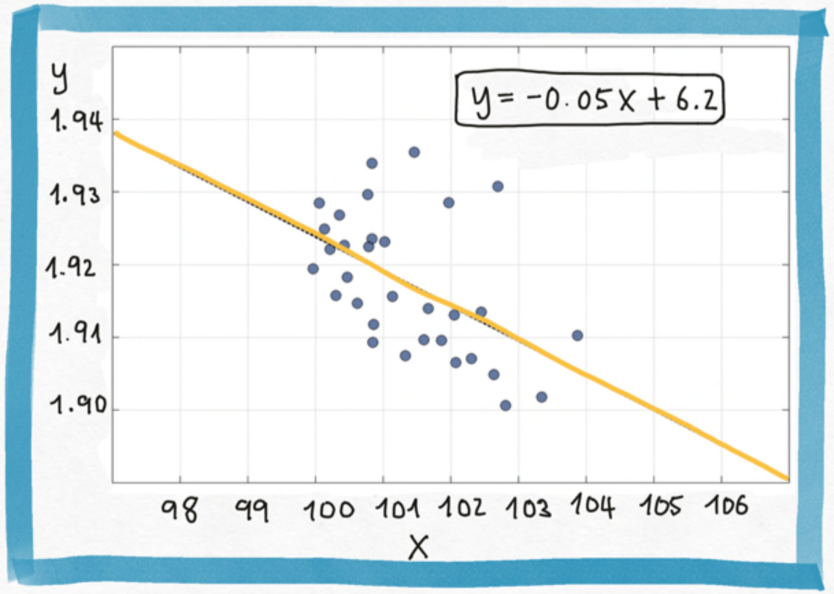
The training MSE here is 2.0 where the test MSE is 12.3.

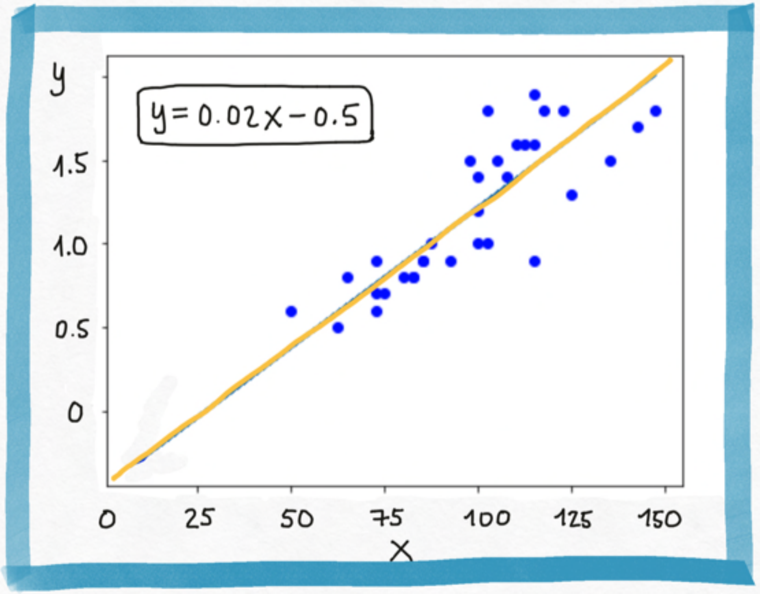
The training data contains a strange point – an outlier – which confuses the model.

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

Evaluation: Model Interpretation

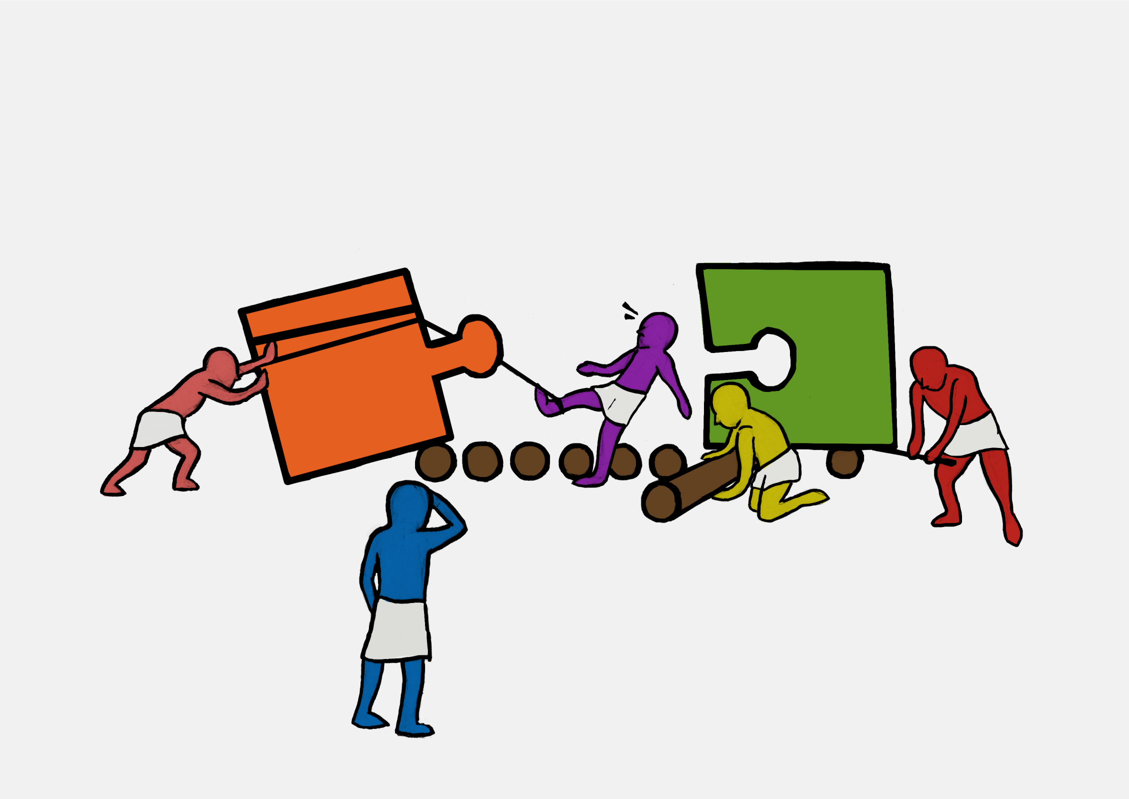
For linear models it's important to interpret the parameters





The MSE of this model is very small. But the slope is -0.05. That means the larger the budget the less the sales.

The MSE is very small but the intercept is -0.5 which means that for very small budget we will have negative sales.



Ex C.1, C.2, [C.3]