Chapter 3

Linear Regression

STAT303-2

Northwestern | WEINBERG COLLEGE OF ARTS & SCIENCES

A Linear Regression Model

The underlying function that is impossible to know and estimated as $\hat{f}$ using the observed data.

We discussed in the last lecture that:

- The main goal of statistical learning is to estimate $\hat{f}$ in

$$\hat{Y} = \hat{f}(X),$$

- The simplest approach to this problem is to assume that f is linear.

$\hat{Y}$ is a linear combination of the predictors.

$X_1, X_2, \ldots, X_p$

$$\hat{Y} = \hat{f}(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

- A linear regression model
  - For a dataset with p predictors
- $\beta_0, \beta_1, \ldots, \beta_p$ are the **parameters** of the model
- Optimizing/training the Linear Regression model means finding the best parameters

?

Simple Linear Regression (SLR)

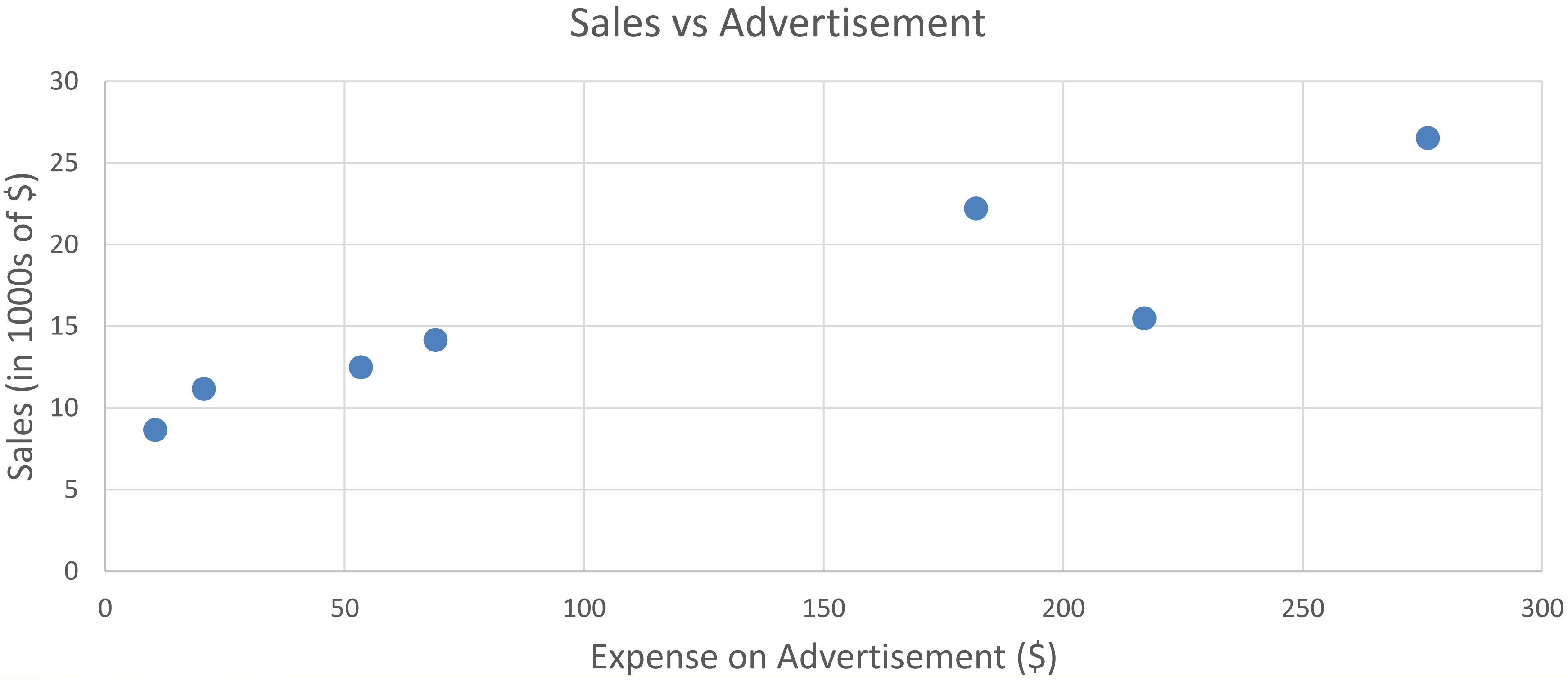
- At first, we will consider only one predictor: Simple Linear Regression
- Our assumption becomes:

$$\hat{Y} = \hat{f}(X) \quad = \beta_0 + \beta_1 X_1 \longrightarrow \text{Two parameters to estimate: } \beta_0, \beta_1$$

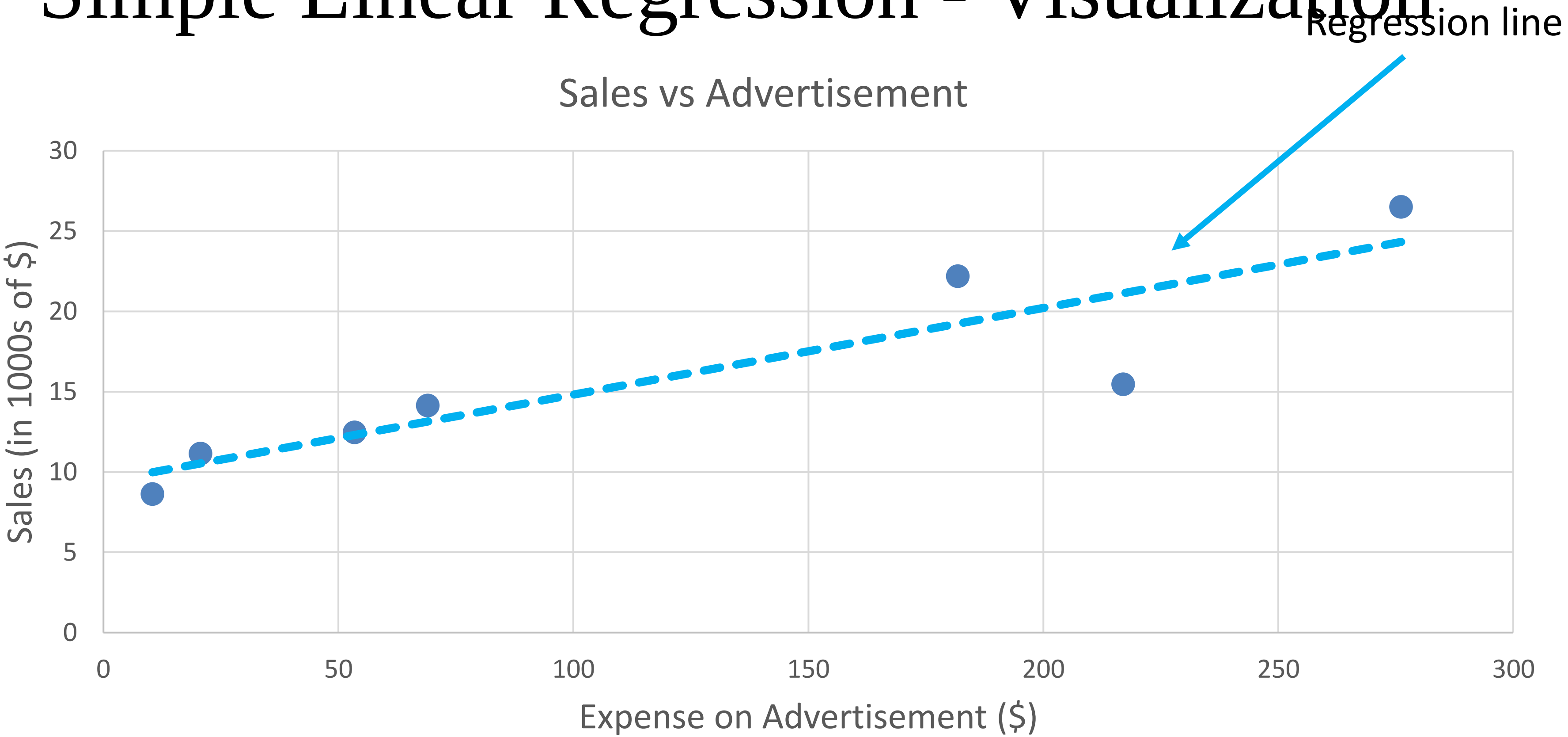
(intercept, slope)

- Let's visualize this.

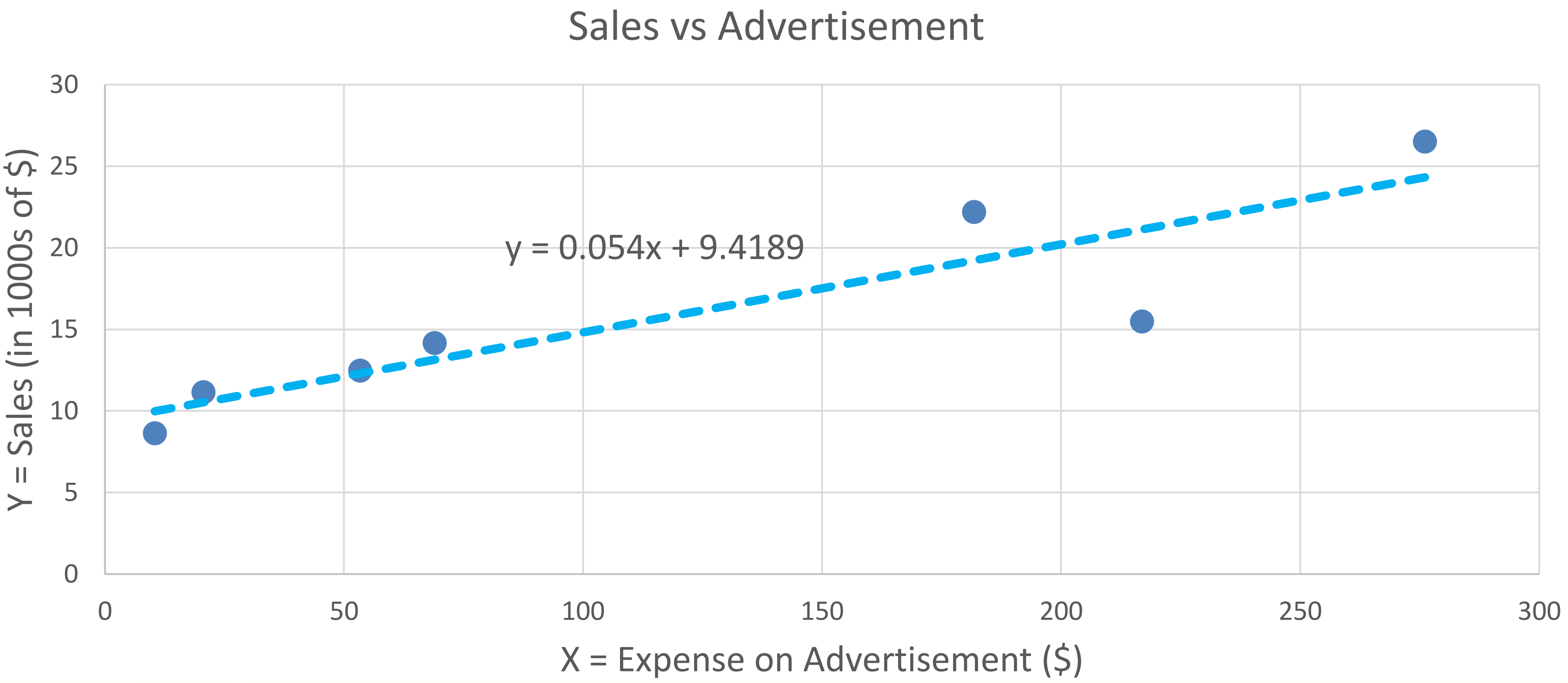
Simple Linear Regression - Visualization



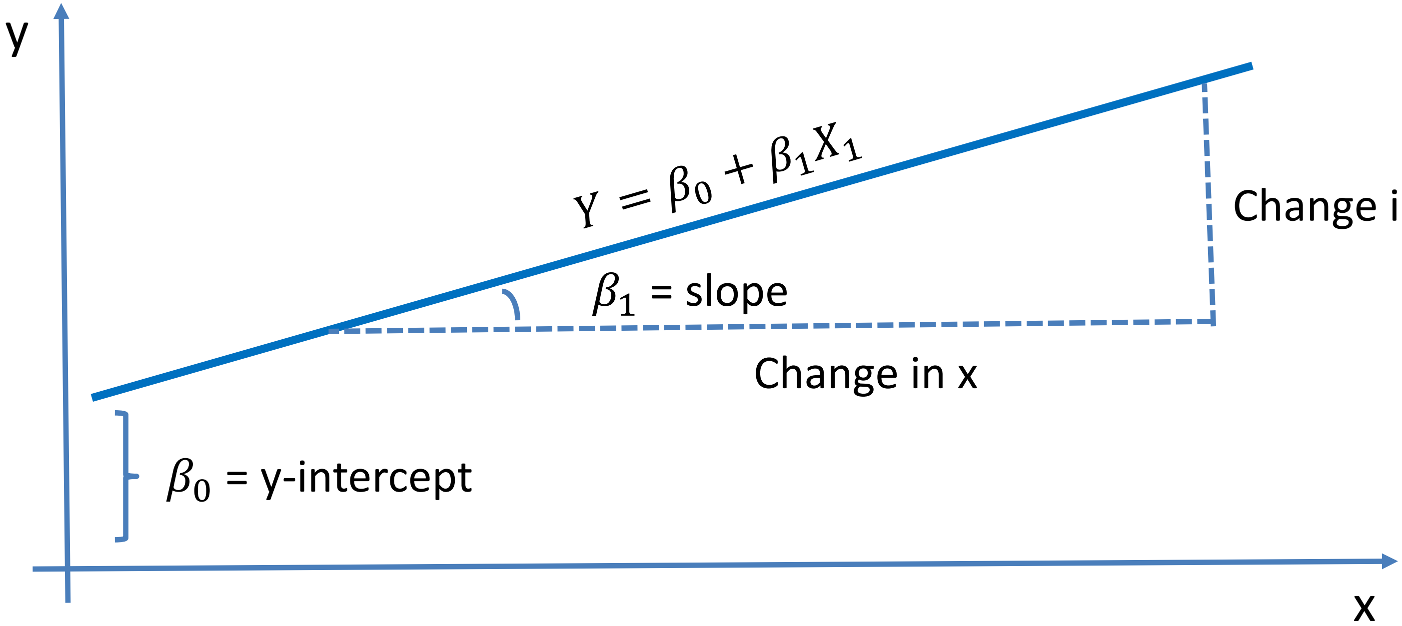
Simple Linear Regression - Visualization



Simple Linear Regression - Visualization



Simple Linear Regression - Visualization



Simple Linear Regression - Formula

We start with the assumption: $Y = \beta_0 + \beta_1 X$

We have the dataset:
$X^1, X^2, \ldots X^N$
$Y^1, Y^2, \ldots Y^N$

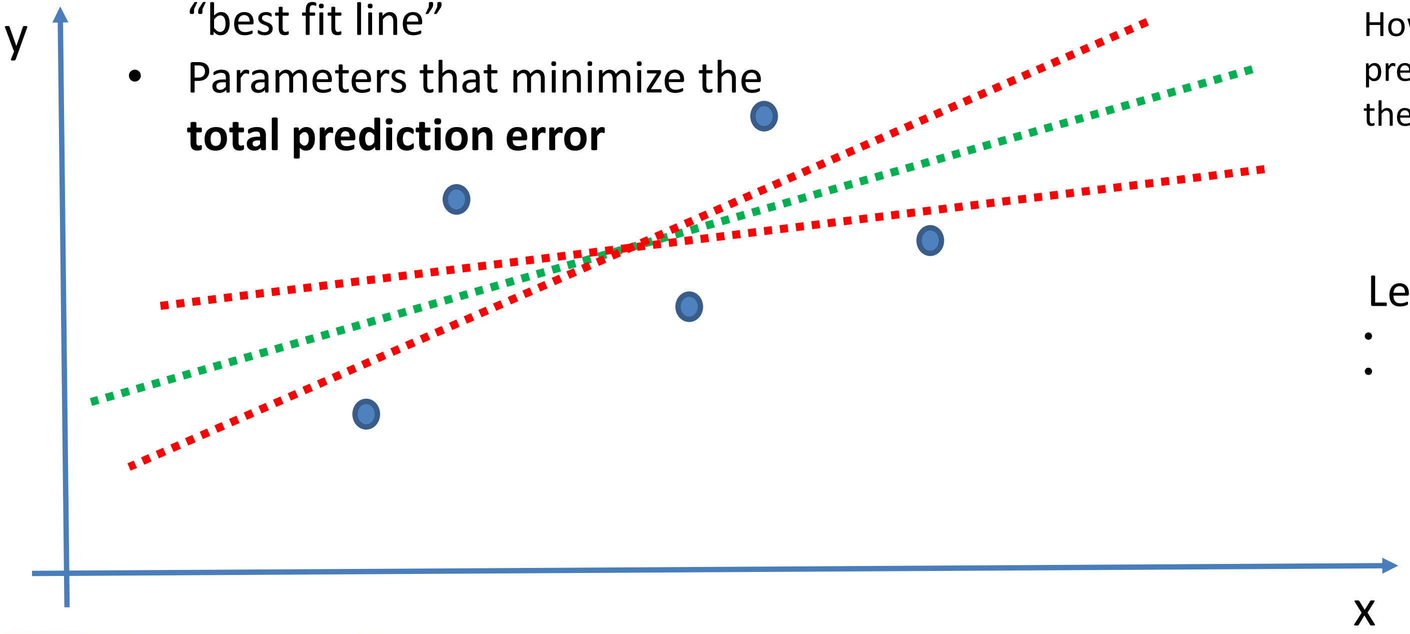
Using them, we find the optimum parameters, $\hat{\beta}_0, \hat{\beta}_1$

Two questions:
- What does optimum mean?
- How to find the optimum parameters?

Simple Linear Regression - Formula

What does optimum mean?

- Parameters that return the "best fit line"
- Parameters that minimize the **total prediction error**



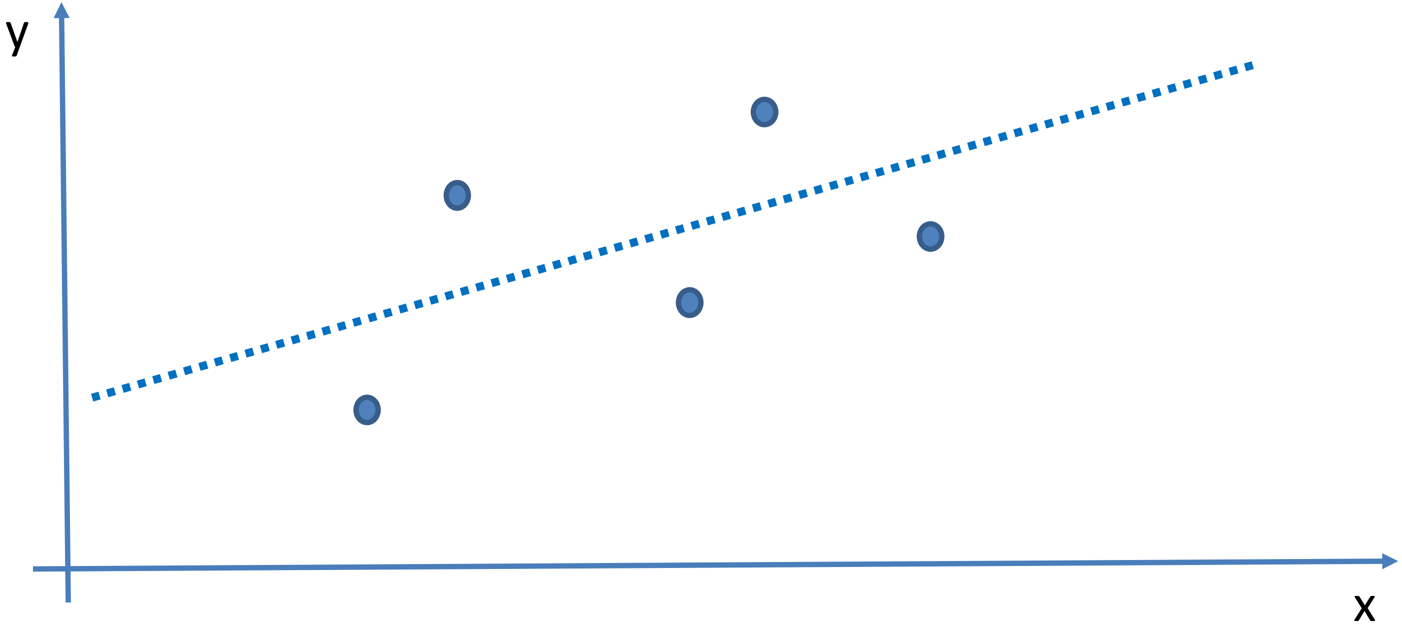
How to minimize the total prediction error? (and find the optimum parameters?)

↓

Least Squares Method

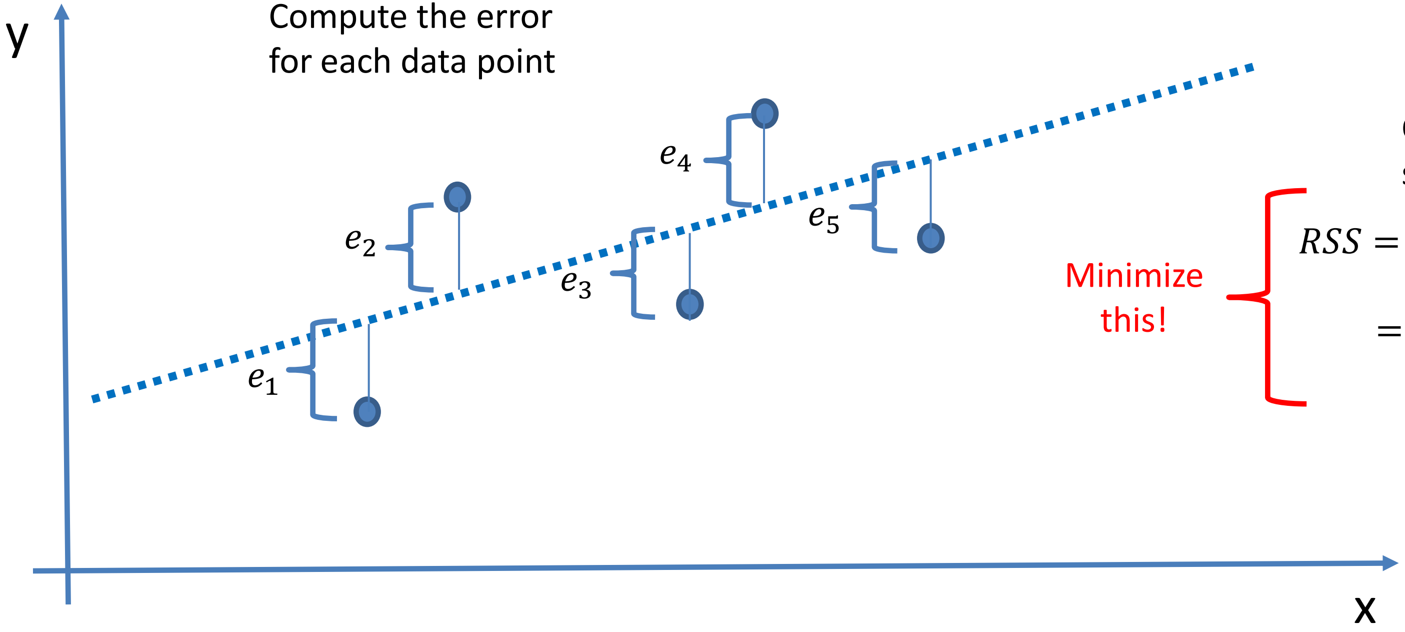
- Visualization first
- Calculation in hand-written notes

Least Squares - Visualization



Consider a line

Least Squares - Visualization



Compute the residual sum of squares (RSS)

Minimize this!

$$RSS = e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$$
$$= \sum_{i=1}^{5} e_i^2$$

Compute the error corresponding to each data point

Least Squares - Calculation

Note that the general formula for RSS is:

$$RSS = e_1^2 + e_2^2 + \cdots + e_N^2$$

where N is the total number of observations.

We can write it as:

$$RSS = \left(Y^1 - \hat{Y}^1\right)^2 + \left(Y^2 - \hat{Y}^2\right)^2 + \cdots + \left(Y^N - \hat{Y}^N\right)^2$$

And then:

$$RSS = \left(Y^1 - \hat{\beta}_0 - \hat{\beta}_1 X^1\right)^2 + \left(Y^2 - \hat{\beta}_0 - \hat{\beta}_1 X^2\right)^2 + \cdots + \left(Y^N - \hat{\beta}_0 - \hat{\beta}_1 X^N\right)^2$$

Check the hand-written notes

Least Squares - Calculation

$$RSS = \left(Y^1 - \hat{\beta}_0 - \hat{\beta}_1 X^1\right)^2 + \left(Y^2 - \hat{\beta}_0 - \hat{\beta}_1 X^2\right)^2 + \cdots + \left(Y^N - \hat{\beta}_0 - \hat{\beta}_1 X^N\right)^2$$

Minimize RSS with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$.

- The only variables in the equation
- Everything else is known from the data at hand.

The final expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ are entirely in terms of the data: X’s and Y’s

Prediction

Given a new observation, plug it in the optimum line formula: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

Obtain the prediction.

- The data at hand that is used for finding the optimum parameters is called **the training dataset.**
  - In a training dataset, both the predictor(s) and the response must be known.
- The “new” data (data that the model has not seen during training) is called **the test dataset.**
  - The point of it is to assess the trained model on a dataset it has not seen before.
  - In the test dataset, the response should be known if the test performance (RMSE or MAE) is to be obtained and compared with the training performance. (RMSE or MAE)

- If the model returns an RMSE/MAE higher than expected/desired both for training and test:
  - It means the model is not enough – a more complex model is necessary.
  - This is called **underfitting**.
  - For basic models such as Linear Regression, underfitting is usually the problem.
- If the model returns a low RMSE/MAE for training and a high RMSE/MAE for test:
  - It means the model parameters are optimized only for the training data – not good for previously unseen data
  - This is called **overfitting –** usually the problem of more complex models (STAT 303-3)

We have covered:
- Linear Regression as a concept
- Simple Linear Regression
- Optimizing the parameters – training
- Prediction
- Python implementation

Next lecture: (along with the GitHub demo)
- Inference and Uncertainty
    - Confidence intervals
    - Prediction interval
    - Coefficient of Determination (R2)
- Multiple Linear Regression

Reference

Source for slides: https://www.statlearning.com/resources-second-edition