1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
   Ans: Independent variable effect on the dependent variable.

2. Why is it important to use drop_first=True during dummy variable creation?
   Ans: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Ans: cnt vs temp, cnt vs atemp
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Ans:
   a. Check The coefficient is statistically significant. So the association is not purely by chance.
   b. We need to check if the error terms are also normally distributed

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

   - temp
   - workingday
   - windspeed

General Subjective Questions

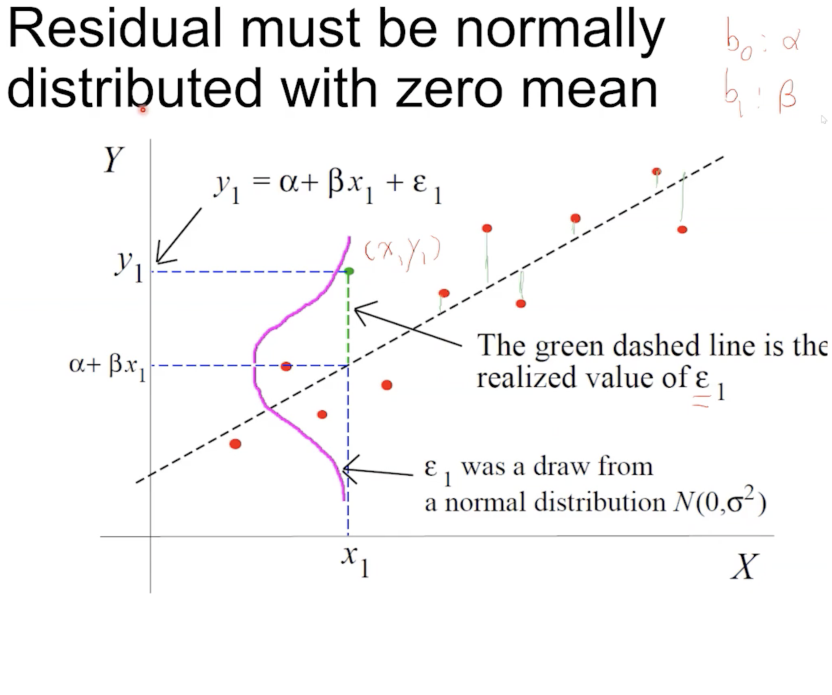
1. Explain the linear regression algorithm in detail

CORRELATION

- find correlation for check the strength and direction of a linear relationship between two variables with equation below

$$r = \frac{n\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}\sqrt{n\left(\sum y_i^2\right) - \left(\sum y_i\right)^2}}$$

- The range of the correlation coefficient is from -1 to 1
- if there is a strong positive linear relationship between the variable, the value of r will be close to. On the other hand strong negative linear relationship, the value of r will be close to -1
- If no no relationship between variables correlation will be close to 0.



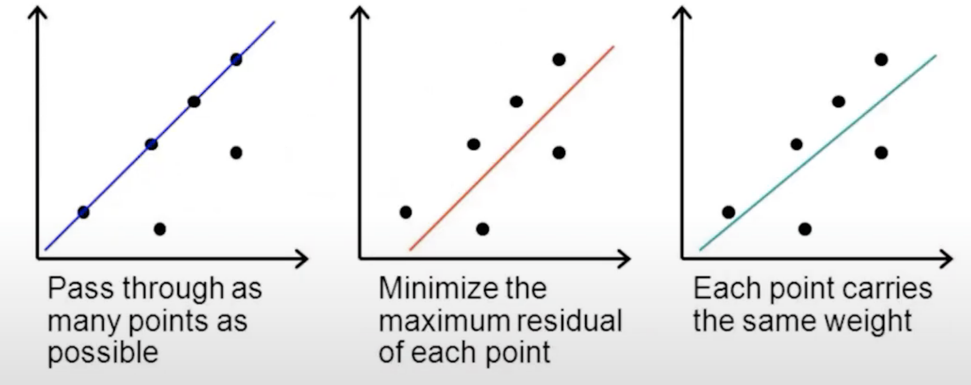
Residual must be normally distributed with zero mean

$b_0 : \alpha$
$b_1 : \beta$
Y
$y_1 = \alpha + \beta x_1 + \varepsilon_1$
y_1
(x, y_1)
$\alpha + \beta x_1$
The green dashed line is the realized value of ε_1
ε_1 was a draw from a normal distribution $N(0,\sigma^2)$
x_1
X

BEST LINE

Want to fie the “best” line to the data

- How do we define “best”?



Pass through as many points as possible
Minimize the maximum residual of each point
Each point carries the same weight

- In general, when we use $\hat{y}_i = b_0 + b_1 x_i$ to predict the actual response $\hat{y}_i$,we make a prediction error (or residual error) of size:

$$e_i = y_i - \hat{y}_i$$

- A line that fits the data “**best**” will be one for which **n prediction errors** - one for which the n prediction errors - one for each observed data point -- **are as small as possible in some overall sense.**

Method of least squares

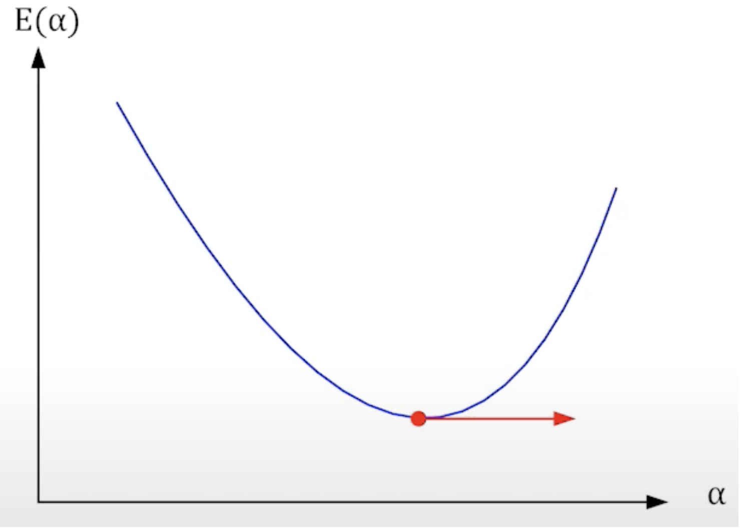
- Choose the b’s so that the sum of the squares of the errors, e_i, are minimized
- The error function is

$$S = \sum_{i=1}^{n} e_i^{2}$$

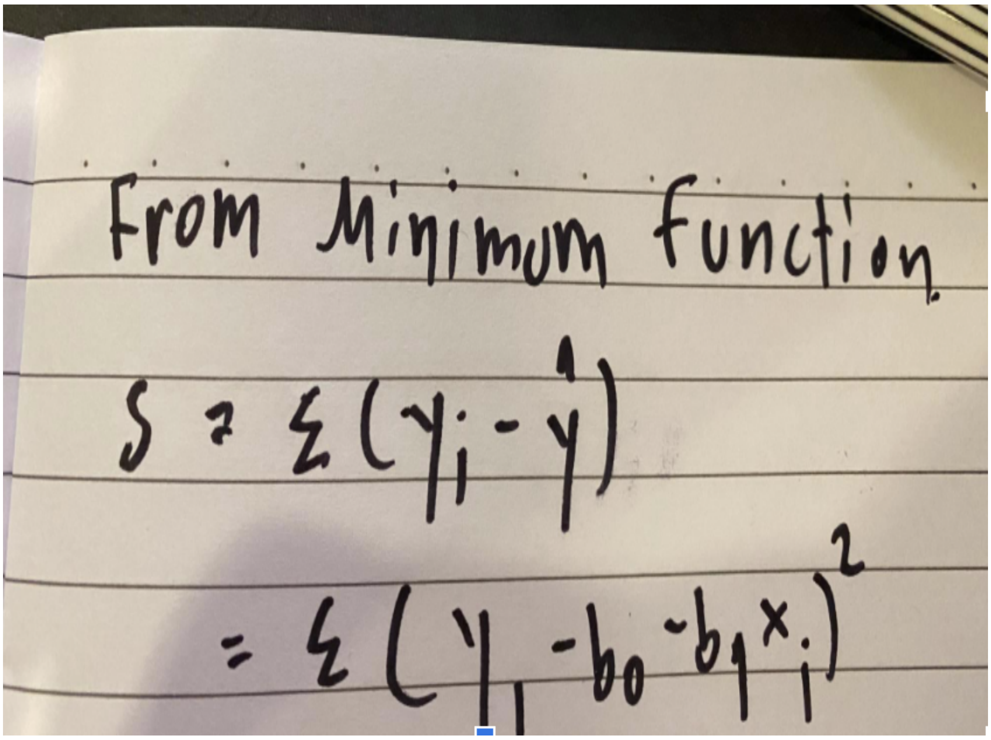
$$= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Ordinary least square solution

Minimum of a function is the point where the slope is zero



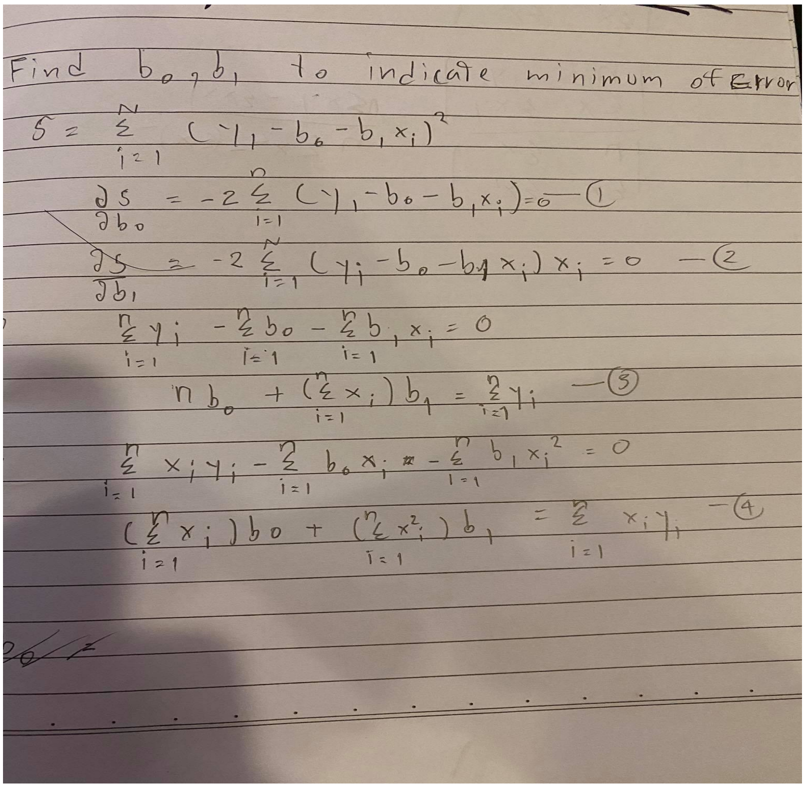
How to find sum of minimum error?



From Minimum function

$$S = \sum (y_i - \hat{y})$$

$$= \sum (y_i - b_0 - b_1 x_i)^2$$



Find b_0, b_1 to indicate minimum of Error

$$S = \sum_{i=1}^{N} (y_1 - b_0 - b_1 x_i)^2$$

$$\frac{\partial S}{\partial b_0} = -2 \sum_{i=1}^{n} (y_1 - b_0 - b_1 x_i) = 0 \quad — ①$$

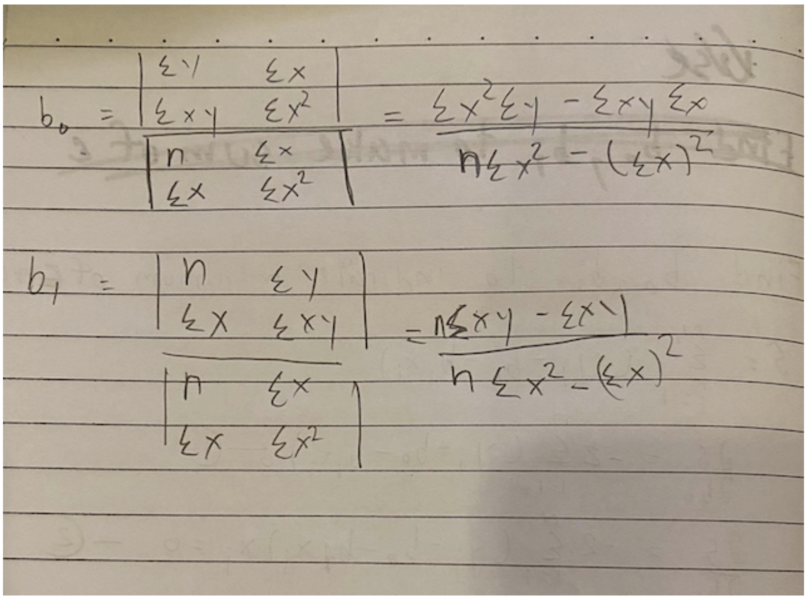
$$\frac{\partial S}{\partial b_1} = -2 \sum_{i=1}^{N} (y_i - b_0 - b_1 x_i) x_i = 0 \quad — ②$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} b_0 - \sum_{i=1}^{n} b_1 x_i = 0$$

$$n b_0 + \left(\sum_{i=1}^{n} x_i\right) b_1 = \sum_{i=1}^{n} y_i \quad — ③$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} b_0 x_i - \sum_{i=1}^{n} b_1 x_i^2 = 0$$

$$\left(\sum_{i=1}^{n} x_i\right) b_0 + \left(\sum_{i=1}^{n} x_i^2\right) b_1 = \sum_{i=1}^{n} x_i y_i \quad — ④$$



$$b_0 = \frac{\begin{vmatrix} \sum y & \sum x \\ \sum xy & \sum x^2 \end{vmatrix}}{\begin{vmatrix} n & \sum x \\ \sum x & \sum x^2 \end{vmatrix}} = \frac{\sum x^2 \sum y - \sum xy \sum x}{n \sum x^2 - (\sum x)^2}$$

$$b_1 = \frac{\begin{vmatrix} n & \sum y \\ \sum x & \sum xy \end{vmatrix}}{\begin{vmatrix} n & \sum x \\ \sum x & \sum x^2 \end{vmatrix}} = \frac{n \sum xy - \sum x y}{n \sum x^2 - (\sum x)^2}$$

Now we get b0, and b1 of linear equation

Coefficient of determination

$$R^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{SS_{yy}}{SS_{yy}} - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

- $\mathrm{SS}_{\mathrm{yy}}$ measures the deviation of the observations from their mean:

  $$SS_{yy} = \sum_i \left(y_i - \overline{y}\right)^2$$

- SSE measures the deviation of observations from their predicted values

  $$SSE = \sum_i \left(y_i - Y_i\right)^2$$

- The higher the R^2, the more useful mode
- R^2 take on values max 1