Simple Linear Regression

Lecture 07 - CPSC392

Simple Linear Regression

Simple Linear Regression

- Model the relationship between 2 variables

Simple Linear Regression

- Model the relationship between 2 variables
  - The 2 variables are a dependent variable (denoted by y) and an independent variable (denoted by x)

Simple Linear Regression

- Model the relationship between 2 variables
    - The 2 variables are a dependent variable (denoted by y) and an independent variable (denoted by x)
- Linear regression is in fact a comparison of 2 models

The Data

- You have randomly selected 5 tests from a pool of tests and want to see if you can predict the test score using only this data.

The Data

- You have randomly selected 5 tests from a pool of tests and want to see if you can predict the test score using only this data.
- We only have 1 variable here, but we can still make a model!

The Data

- You have randomly selected 5 tests from a pool of tests and want to see if you can predict the test score using only this data.
- We only have 1 variable here, but we can still make a model!
- Let’s plot the data

Model 1

- What can we say about this data?
- What is the score for test 6 or 7 going to be?

Model 1

- What can we say about this data?
- What is the score for test 6 or 7 going to be?

- Our best estimate, in this case, is the mean

Model 1

- What can we say about this data?
- What is the score for test 6 or 7 going to be?

- Our best estimate, in this case, is the mean
- So for every future test, we can predict that the score is going to be 7

Model 1

- What can we say about this data?
- What is the score for test 6 or 7 going to be?

- Our best estimate, in this case, is the mean
- So for every future test, we can predict that the score is going to be 7

“ with only one variable, and no other information, the best prediction for the next measurement is the mean of the sample”

Goodness of Fit

Goodness of Fit

- How good a line fits the y-values

Goodness of Fit

- How good a line fits the y-values
- This is very similar to the concept of standard deviation

Residuals/Errors

- Distance between the best fit line to the observed values

Sum of Squared Errors (SSE)

Sum of Squared Errors (SSE)

- Measure of discrepancy between the data and the estimated model

Sum of Squared Errors (SSE)

- Measure of discrepancy between the data and the estimated model
- Calculated by squaring all errors and summing them up

Goals

- The goal of a simple linear regression is to create a linear model that minimizes the SSE

Goals

- The goal of a simple linear regression is to create a linear model that minimizes the SSE
- If the regression model is “significant”, it will take away a large chunk of the SSE

Goals

- The goal of a simple linear regression is to create a linear model that minimizes the SSE
- If the regression model is “significant”, it will take away a large chunk of the SSE
- The model should “fit” the data better and minimize the residuals once we introduce an independent variable

The SSE of the model with just test scores is 20. Let’s introduce a new independent variable, total hours of study, and see if we can create a linear regression model using this attribute

Lines

Lines

- **$y = mx + b$**

- **m** = slope (rise/run)
- **b** = y-intercept (point where $x = 0$)

Lines

- $y = mx + b$
- $y = \beta_0+\beta_1x+\varepsilon$

- β_1 = slope parameter
- β_0 = y-intercept parameter
- ε = error term (unexplained variation in y)

Lines

- $y = mx + b$
- **$y = \beta_0+\beta_1x+\varepsilon$ (for population data)**

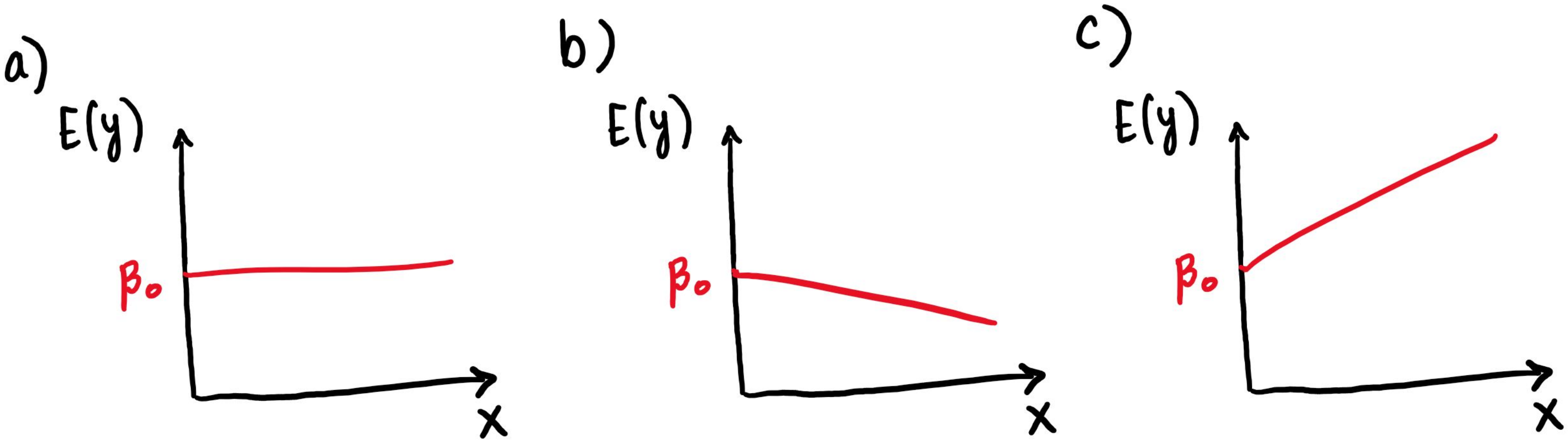
- $\boldsymbol{\beta_1}$ = slope parameter
- $\boldsymbol{\beta_0}$ = y-intercept parameter
- $\boldsymbol{\varepsilon}$ **=** error term (unexplained variation in y)

Lines (Simple Linear Regression)

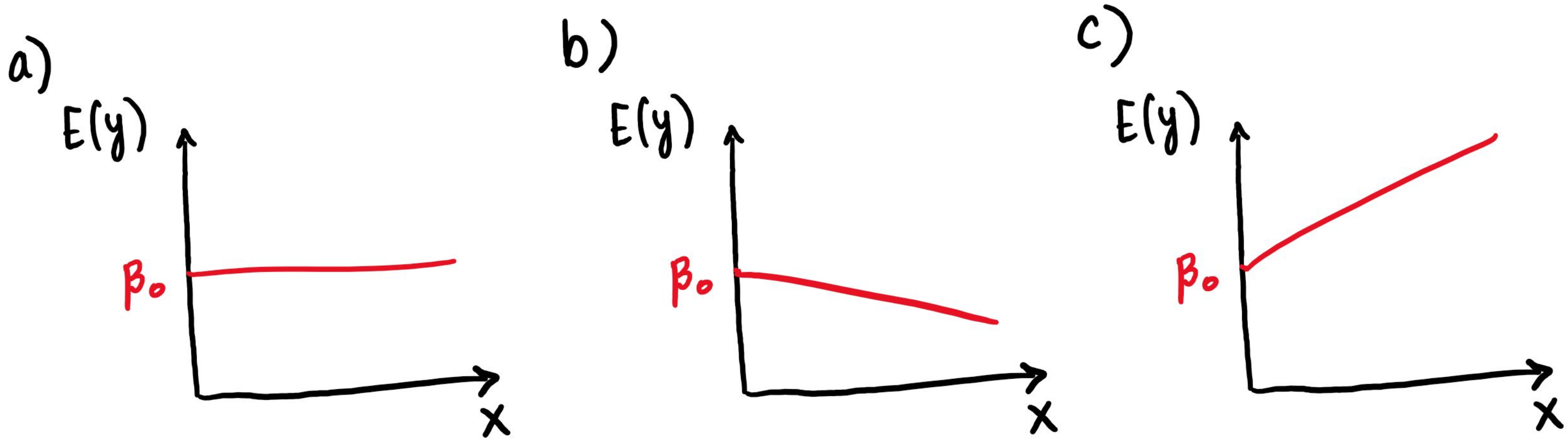
- $E(y) = \beta_0 + \beta_1 x$

- β_1 = slope parameter
- β_0 = y-intercept parameter
- **E(y)** = mean or expected value of y, given some x

What are the equations for these lines?

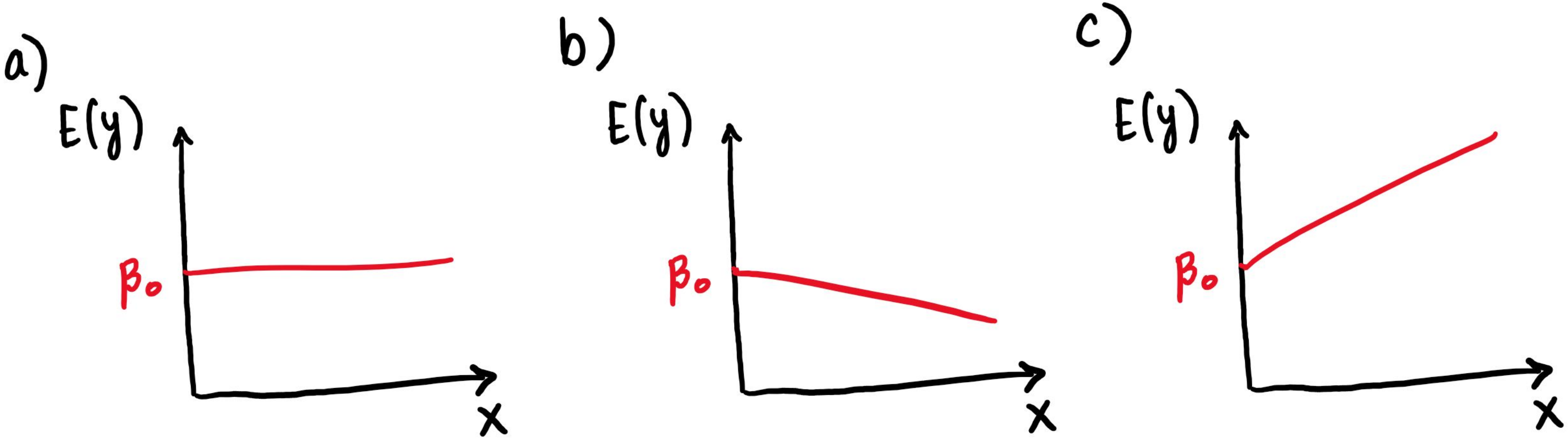


What are the equations for these lines?



$E(y) = \beta_0 + (0)\, x$

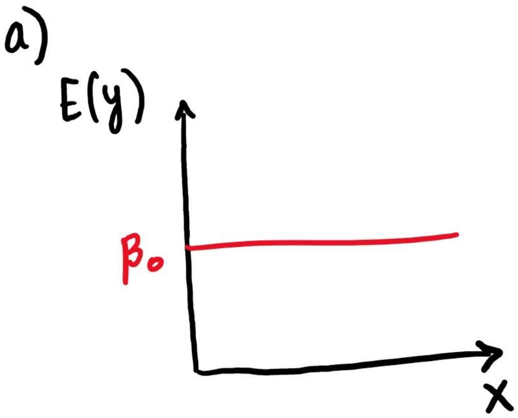
What are the equations for these lines?



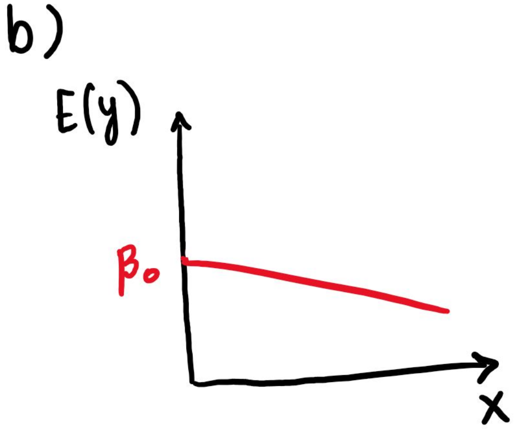
$E(y) = \beta_0 + (0)\,x$

$E(y) = \beta_0 - \beta_1\,x$

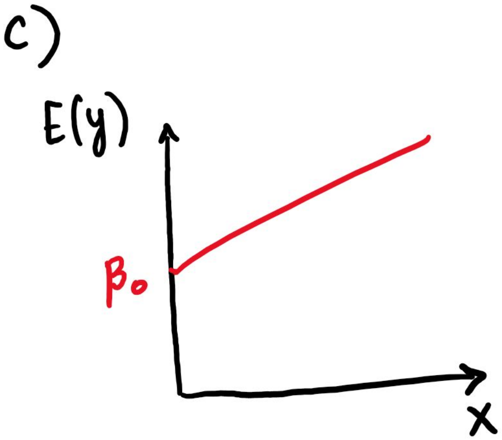
What are the equations for these lines?



$E(y) = \beta_0 + (0)\,x$



$E(y) = \beta_0 - \beta_1\,x$



$E(y) = \beta_0 + \beta_1\,x$

Linear Regression for a Sample

- $E(y) = \beta_0+\beta_1 x$
- $\hat{y} = b_0+b_1 x$

- **$\hat{y}$ (y-hat)** = estimator of E(y)

Data (with hours of study)

- Does the plot of hours of study vs test scores show some relationship?

Data (with hours of study)

- Does the plot of hours of study vs test scores show some relationship?
- If yes, then we can fit a linear regression line to predict future scores

Data (with hours of study)

- Does the plot of hours of study vs test scores show some relationship?
- If yes, then we can fit a linear regression line to predict future scores
- If no, then the linear regression model might be useless

$$\hat{y} = b_0 + b_1 x$$

$$\hat{y}_i = b_0 + b_1 x_i$$

$$b_1 = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}, \quad b_0 = \hat{y}_i - b_1 x_i$$

$\bar{x}$ = mean of independent variable

x_i = value of independent variable

$\bar{y}$ = mean of dependent variable

y_i = value of dependent variable.

Best-fit Line

$$\hat{y}_i = 3.2 + 0.95x_i$$

Best-fit Line

$$\hat{y}_i = 3.2 + 0.95x_i$$

- For every 1 hour increase in study time, you expect to get an increase in score by 0.95 points

Best-fit Line

$$\hat{y}_i = 3.2 + 0.95x_i$$

- For every 1 hour increase in study time, you expect to get an increase in score by 0.95 points
- If you don't study, x = 0, then you will end up with a score of 3.2 (practical?)

Least Square Criterion

$$\min \Sigma (y_i - \hat{y}_i)^2$$

y_i = observed value of test score

$\hat{y}_i$ = predicted value of test score

Least Square Criterion

- Goal is to minimize the sum of the squared differences between the actual value of dependent variable and the estimated (predicted) value

Least Square Criterion

- Goal is to minimize the sum of the squared differences between the actual value of dependent variable and the estimated (predicted) value
- We can find this sum and compare with the SSE of Model 1 to see how much linear regression minimizes the distance

SSR & SST

- SSE = sum of squared errors
- SST = sum of squared total
    - Equals to SSE when no independent variable is used in model
- SSR = sum of squared regression
    - SSR = SST - SSE

Coefficient of Determination (r^2)

$r^2 = SSR / SST$

- Proportion of the variance in the dependent variable that is predictable from the independent variable

Coefficient of Determination (r^2)

r^2 = SSR / SST

- Proportion of the variance in the dependent variable that is predictable from the independent variable
- For our model, r^2 = 0.90 or 90%

Coefficient of Determination (r^2)

r^2 = SSR / SST

- Proportion of the variance in the dependent variable that is predictable from the independent variable
- For our model, $r^2 = 0.90$ or 90%
- So hours of study are able to explain 90% of variation in test scores

Coefficient of Determination (r^2)

r^2 = SSR / SST

- Proportion of the variance in the dependent variable that is predictable from the independent variable
- For our model, $r^2 = 0.90$ or 90%
- So hours of study are able to explain 90% of variation in test scores
- **GOOD FIT!**

Mean Square Error (MSE)

- MSE is an estimator of the variance of error, ε

Mean Square Error (MSE)

- MSE is an estimator of the variance of error, ε
- MSE = SSE / (n-degrees of freedom)

- SSE = sum of squared errors
- n = number of observations (data points)
- Degrees of freedom = how many parameters are being used in the linear model

Mean Square Error (MSE)

- MSE is an estimator of the variance of error, ε
- MSE = SSE / (n-degrees of freedom)
- MSE = SSS / (n-2) (because of two parameters being used: b_0,b_1)

Standard Error

- Similar to standard deviation, measure of actual spread of data from the best-fit line

Standard Error

- Similar to standard deviation, measure of actual spread of data from the best-fit line
- Standard Error = sqrt (MSE)

Standard Error

- Similar to standard deviation, measure of actual spread of data from the best-fit line
- Standard Error = sqrt (MSE)
- For our model, Standard Error = 0.42
- What does it mean?

Standard Error

- Similar to standard deviation, measure of actual spread of data from the best-fit line
- Standard Error = sqrt (MSE)
- For our model, Standard Error = 0.42
- So the average distance of the observed test scores from the fitted line is 0.42 points.

Multiple Linear Regression

- We can't just always be using one variable to predict the behaviour of another variable

Multiple Linear Regression

- We can't just always be using one variable to predict the behaviour of another variable
- For our example, test scores could be dependent upon the hours you study, your average grade in the class, and if you had breakfast that morning

Multiple Linear Regression

- Extension of Simple Linear Regression which models a one-to-one relationship

Multiple Linear Regression

- Extension of Simple Linear Regression which models a one-to-one relationship
- Multiple Linear Regression models a many-to-one linear relationship

Multiple Linear Regression

- Extension of Simple Linear Regression which models a one-to-one relationship
- Multiple Linear Regression models a many-to-one linear relationship
    - You can have multiple independent variables and a single dependent variable (x_1, x_2, etc. and y)

Things to Consider (Multiple LR)

- Having more independent variable will always increase the r^2 value

Things to Consider (Multiple LR)

- Having more independent variable will always increase the r^2 value
  - Because you are explaining more and more variation in the dependent variable using multiple independent variables

Things to Consider (Multiple LR)

- Having more independent variable will always increase the r^2 value
    - Because you are explaining more and more variation in the dependent variable using multiple independent variables
- But that does not mean that your regression is better or will predict with more accuracy!

Overfitting

- Occurs when the model is too complex

Overfitting

- Occurs when the model is too complex
  - By that we mean that the model has too many independent variables being used

Overfitting

- Occurs when the model is too complex
  - By that we mean that the model has too many independent variables being used
- Here, instead of explaining the relationship between the variables, the model starts to predict the random error in the data

Multicollinearity

- If you use too many independent variables, there is a possibility that some of those variables depend on each other too

Multicollinearity

- If you use too many independent variables, there is a possibility that some of those variables depend on each other too
    - Think in terms of derived attributes (age and age group, BMI and height etc.)

Gradient Descent

- We have so far been talking about Simple Linear Regression

Gradient Descent

- We have so far been talking about Simple Linear Regression
    - But using what is commonly known as the Ordinary Least Squares (OLS) method

Gradient Descent

- We have so far been talking about Simple Linear Regression
  - But using what is commonly known as the Ordinary Least Squares (OLS) method
    - That is, find the slope and intercept parameters by minimizing the residuals
      - $\min(\sum (y_i - \hat{y}_i)^2)$

Gradient Descent

- We have so far been talking about Simple Linear Regression
  - But using what is commonly known as the Ordinary Least Squares (OLS) method
    - That is, find the slope and intercept parameters by minimizing the residuals
      - $\min (\sum (y_i - \hat{y}_i)^2)$
- Here we say that using OLS method, we optimize the slope and the intercept

Gradient Descent

- In this method, we optimize the parameters too, but there is some learning involved!

Gradient Descent

- We first define a cost function, *J*
  - For simple linear regression, our cost function is equal to the mean squared error (MSE)
    - MSE = mean squared difference between observed and predicted values

Gradient Descent

- We first define a cost function, *J*
    - For simple linear regression, our cost function is equal to the mean squared error (MSE)
        - MSE = mean squared difference between observed and predicted values
- Next, we try to minimize this cost function

Gradient Descent

- We first define a cost function, *J*
  - For simple linear regression, our cost function is equal to the mean squared error (MSE)
    - MSE = mean squared difference between observed and predicted values
- Next, we try to minimize this cost function
- Which in turn gives us the optimal parameters for linear regression

Gradient Descent

X	Y
3	7
7	10
4	6
1	4
5	8

$$\hat{y}_i = 3.2 + 0.95\, x_i$$

$$MSE = \frac{1}{2} \Sigma \left(\hat{y}_i - y_i \right)^2$$

Gradient Descent

X	Y
3	7
7	10
4	6
1	4
5	8

$$\hat{y}_i = 3.2 + 0.95\, x_i$$

$$MSE = \frac{1}{2} \Sigma \left(\hat{y}_i - y_i \right)^2$$

- Keeping slope parameter constant (b_1), let's try to estimate the intercept parameter (b_0)

Gradient Descent

X	Y
3	7
7	10
4	6
1	4
5	8

$$\hat{y}_i = 3.2 + 0.95 x_i$$

$$MSE = \frac{1}{2} \Sigma \left(\hat{y}_i - y_i \right)^2$$

- Keeping slope parameter constant (b_1), let's try to estimate the intercept parameter (b_0)
- Start with a random value for b_0 and plug into the best-fit line equation to get predicted y values

Gradient Descent

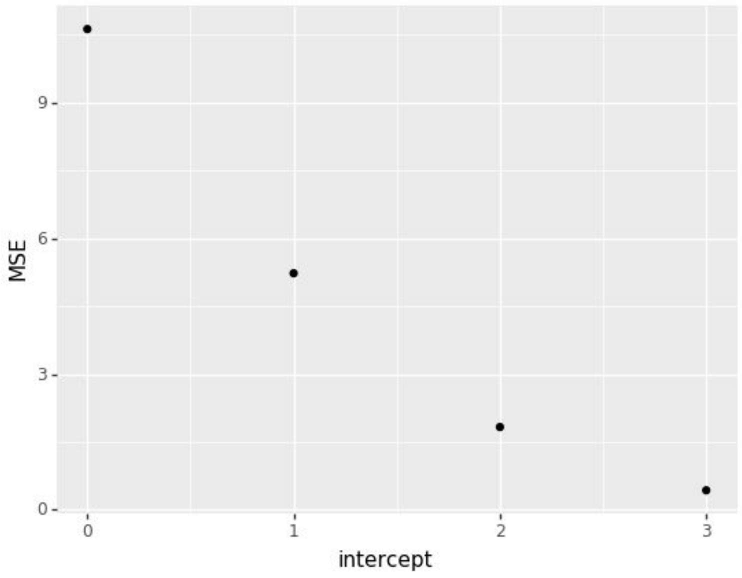
X	Y
3	7
7	10
4	6
1	4
5	8

$$\hat{y}_i = 3.2 + 0.95 x_i$$

$$MSE = \frac{1}{2}\Sigma(\hat{y}_i - y_i)^2$$

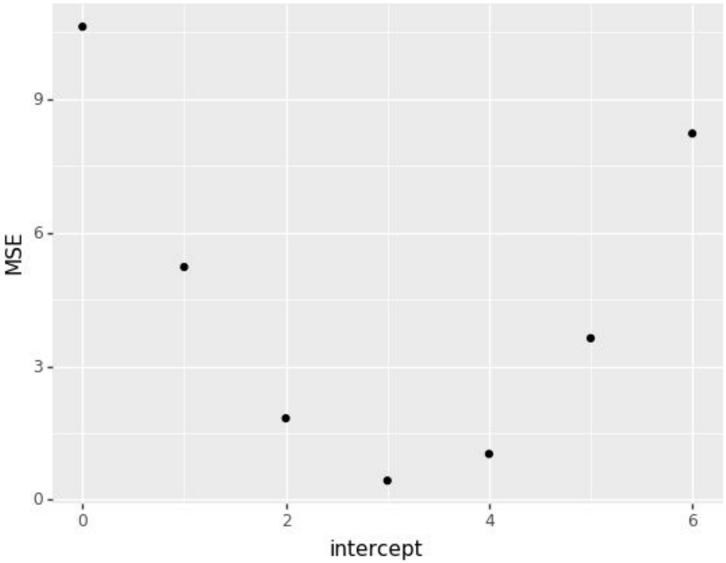
- Keeping slope parameter constant (b_1), let's try to estimate the intercept parameter (b_0)
- Start with a random value for b_0 and plug into the best-fit line equation to get predicted y values
- Calculate MSE and plot

Gradient Descent



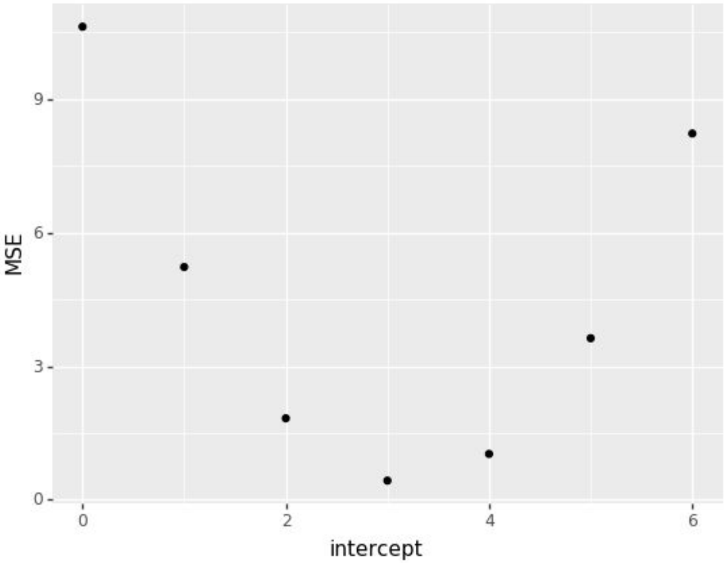
- MSE for b_0 = 0,1,2,3

Gradient Descent



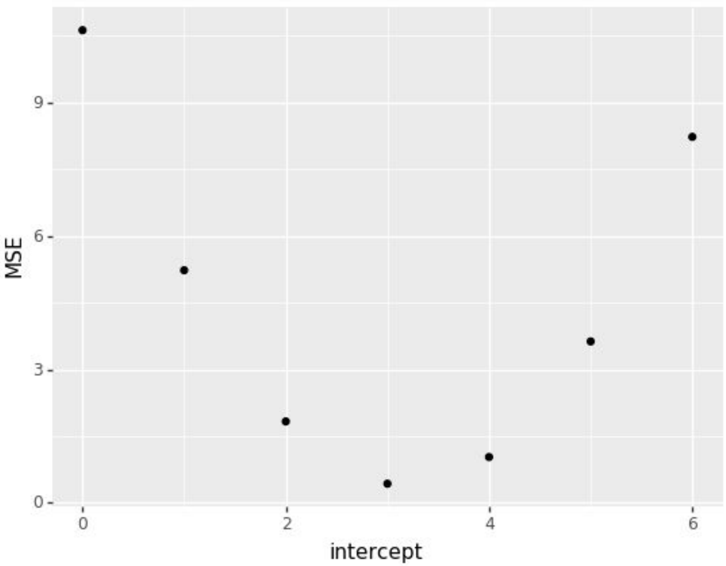
- MSE with b_0 = 0,1,2,3,4,5,6

Gradient Descent



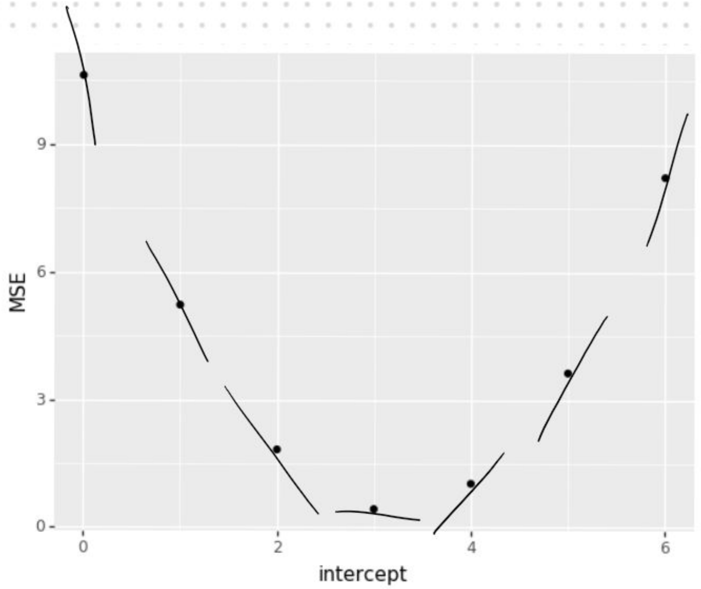
- MSE with b_0 = 0,1,2,3,4,5,6
- This graph is showing that we have moved beyond the lowest value of MSE

Gradient Descent



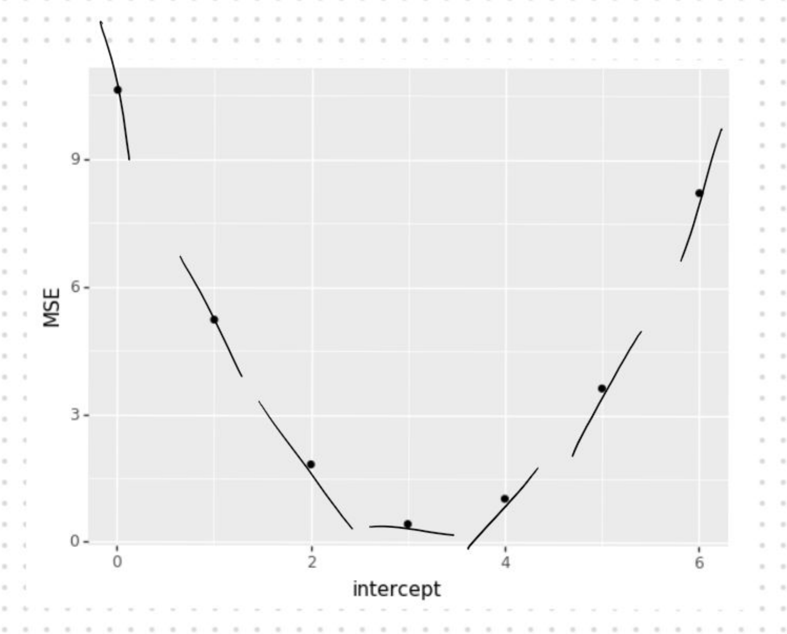
- MSE with b_0 = 0,1,2,3,4,5,6
- This graph is showing that we have moved beyond the lowest value of MSE
- So there needs to be a stopping point!

Gradient Descent



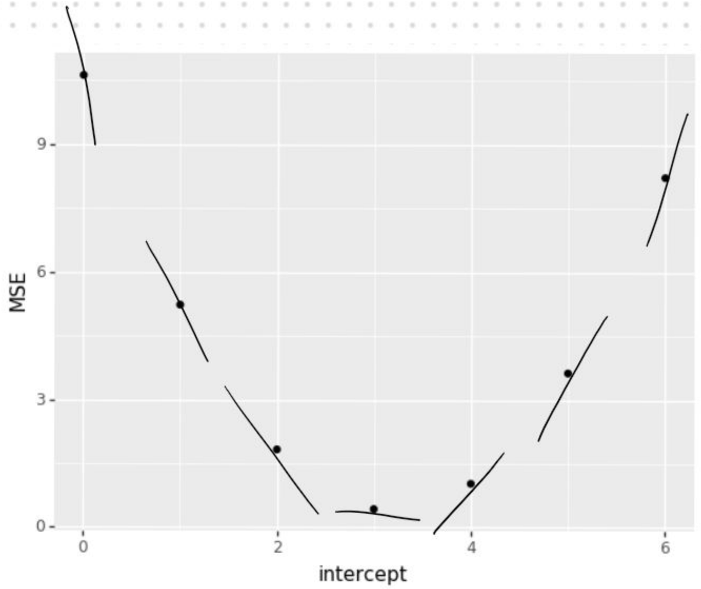
When to stop?

Gradient Descent (derivative)



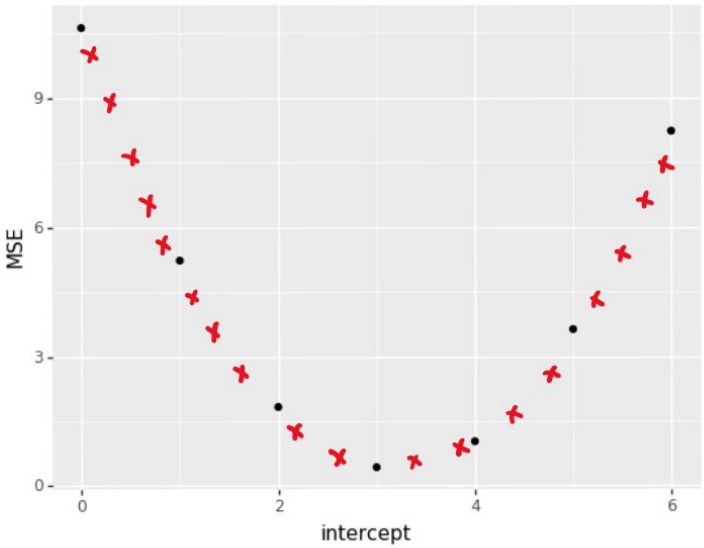
- Take the derivative of the MSE function at each step and check if closer to 0

Gradient Descent (derivative)



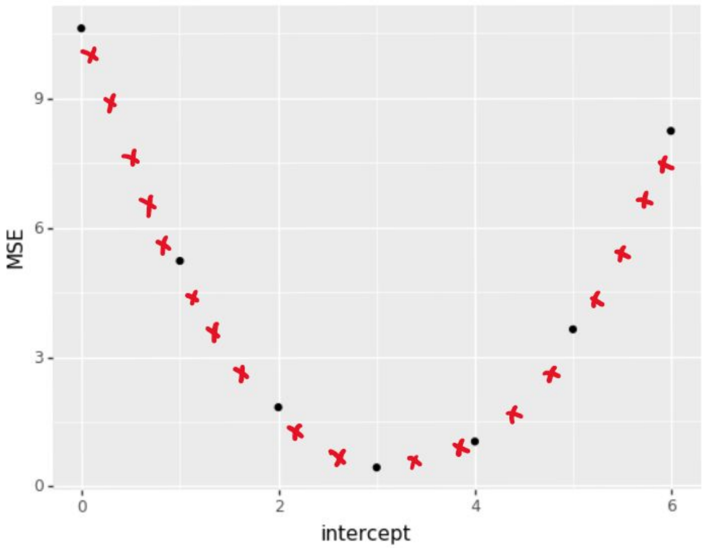
- Take the derivative of the MSE function at each step and check if closer to 0
- Stop plugging in values once closest to 0

Gradient Descent



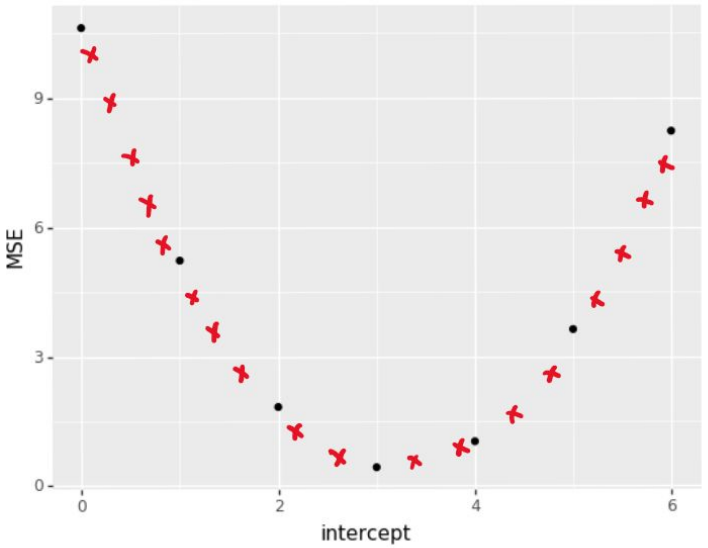
How many steps to take?

Gradient Descent (learning rate)



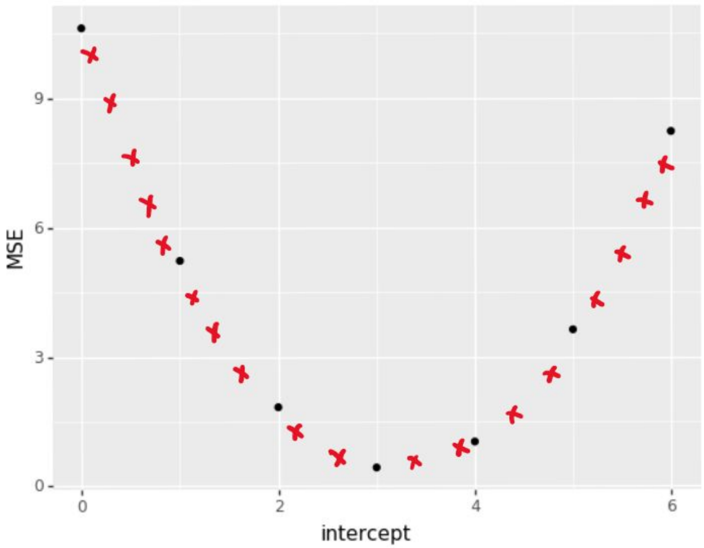
- If large value of MSE derivative, take longer steps

Gradient Descent (learning rate)



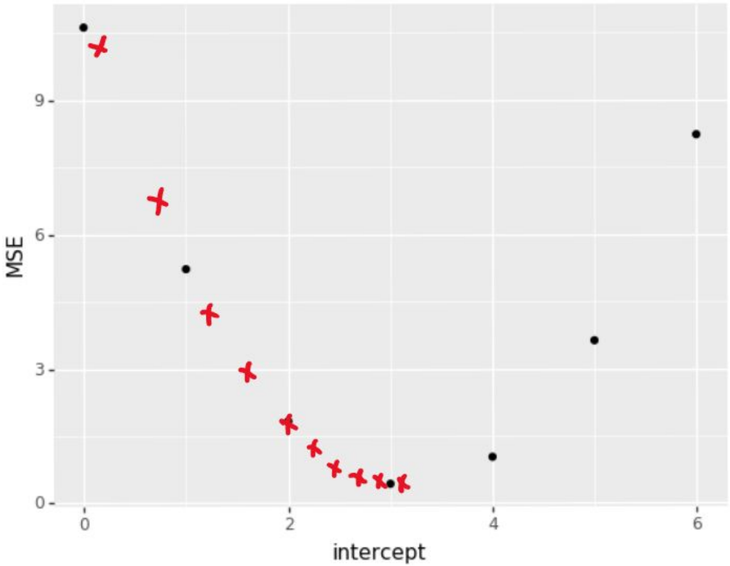
- If large value of MSE derivative, take longer steps
- As derivative gets closer to 0, start taking small steps

Gradient Descent (learning rate)



- If large value of MSE derivative, take longer steps
- As derivative gets closer to 0, start taking small steps
- Learning Rate allows to regulate the step, as a multiplier

Gradient Descent



- “Descending” into the minimum value
- “Gradient” as in derivative

Gradient Descent

$$MSE = \frac{1}{2} \Sigma (\underbrace{\hat{y}_i - y_i}_{\text{residuals}})^2$$

$$J_{(b_0, b_1)} = \frac{1}{2} \Sigma ((b_0 + b_1 x_i) - y_i)^2$$

Gradient Descent

$$J(b_0,b_1) = \frac{1}{2}\Sigma\left(\left(b_0 + b_1 x_i\right) - y_i\right)^2$$

$$J(b_0,b_1) = \frac{1}{2}\,\text{ERROR}^2. \qquad \left[\text{ERROR} = b_0 + b_1 x_i - y_i\right]$$

$$\frac{\partial J}{\partial b_1} = \frac{1}{\cancel{2}} \times \cancel{2} \times \text{ERROR} \times \frac{\partial\,\text{ERROR}}{\partial b_1} \qquad [\text{chain rule}]$$

$$= \text{ERROR} \times (0 + x_i - 0)$$

$$= \text{ERROR} \times (x_i)$$

$$\frac{\partial J}{\partial b_0} = \frac{1}{\cancel{2}} \times \cancel{2} \times \text{ERROR} \times \frac{\partial\,\text{ERROR}}{\partial b_0}$$

$$= \text{ERROR} \times (1 + 0 - 0)$$

$$= \text{ERROR}$$

Gradient Descent

$$b0_n = b_0 - ERROR \times \text{Learning rate}$$

$$b1_n = b_1 - ERROR \times x_i \times \text{Learning rate}$$

Gradient Descent (Steps)

1. Take the derivative of the cost function for each parameter
2. Pick random values for the parameters
3. Plug the values into the derivatives
4. Calculate the step size
5. Calculate the new parameters
6. Repeat 3. Until step size = 0

Summary

- Linear Regression has 4 key assumptions:
  - Linear relationship (matrix correlation plot)
  - Multivariate normality (histogram)
  - No or little multicollinearity
  - Homoscedasticity (residuals are equal across the best fit line)

Summary

- Linear Regression has 4 key assumptions:
  - Linear relationship (matrix correlation plot)
  - Multivariate normality (histogram)
  - No or little multicollinearity
  - Homoscedasticity (residuals are equal across the best fit line)
- Preprocess data first
  - Remove or impute outliers and missing values
  - Transform, standardize data if needed

Summary

- Simple Linear Regression can be done using the OLS and the Gradient Descent method
  - Each method tries to optimize the slope and intercept parameters
  - OLS tries to minimize the SSE
  - GD minimizes the cost function

Summary

- Simple Linear Regression can be done using the OLS and the Gradient Descent method
  - Each method tries to optimize the slope and intercept parameters
  - OLS tries to minimize the SSE
  - GD minimizes the cost function
- Linear Regression is applied using continuous data, to predict continuous data

Summary

- Simple Linear Regression can be done using the OLS and the Gradient Descent method
  - Each method tries to optimize the slope and intercept parameters
  - OLS tries to minimize the SSE
  - GD minimizes the cost function
- Linear Regression is applied using continuous data, to predict continuous data
- You can compare R-squared values, split the data into training and testing sets, and conduct k-fold validation to assess the model performance