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

- $y = mx + b$
- $y = \beta_0+\beta_1 x+\varepsilon$

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

Lines

- $y = mx + b$
- **$y = \beta_0+\beta_1 x+\varepsilon$ (for population data)**

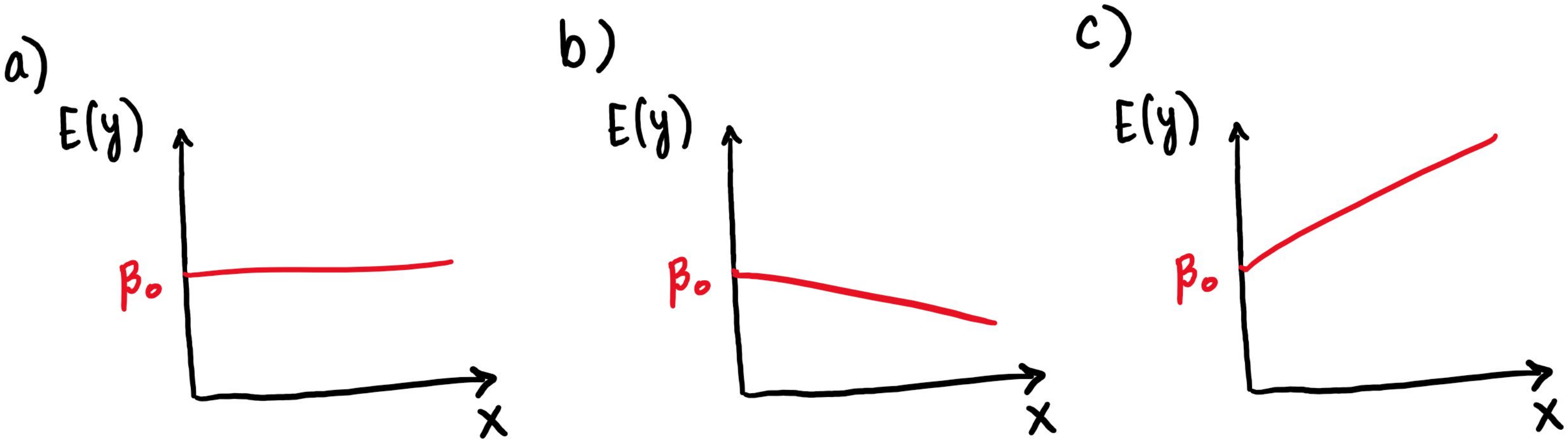
- β_1 = slope parameter
- β_0 = y-intercept parameter
- ε = error term (unexplained variation in y)

Lines (Simple Linear Regression)

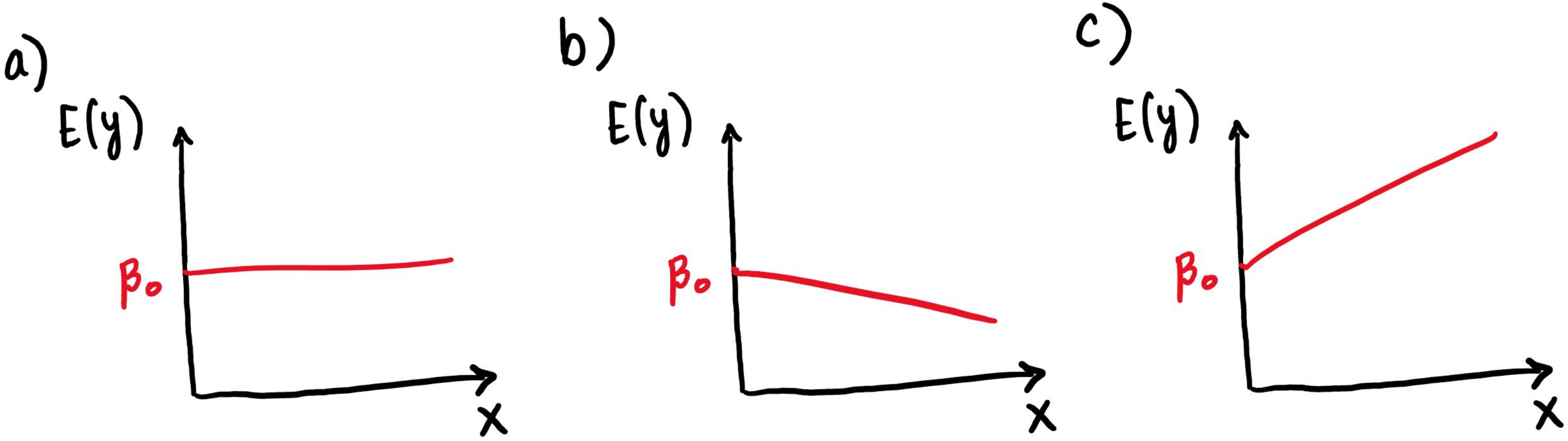
- **$E(y) = \beta_0+\beta_1x$**

- **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?

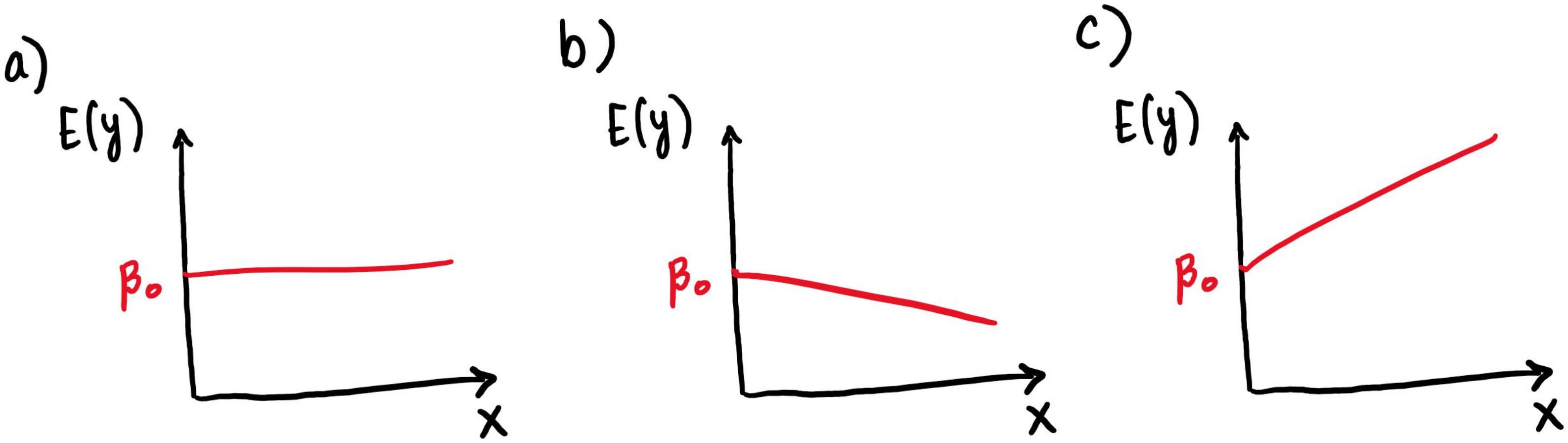


What are the equations for these lines?



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

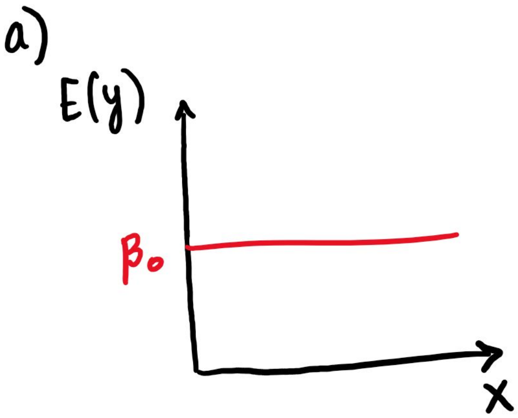
What are the equations for these lines?

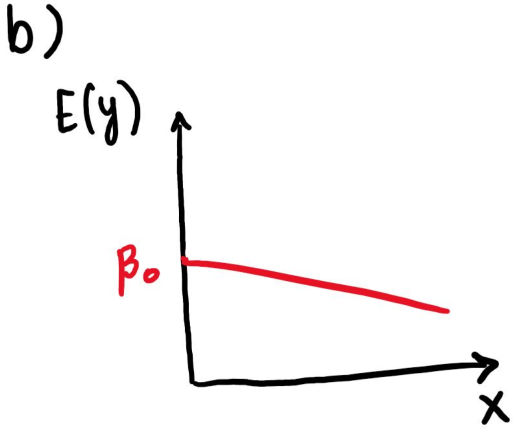


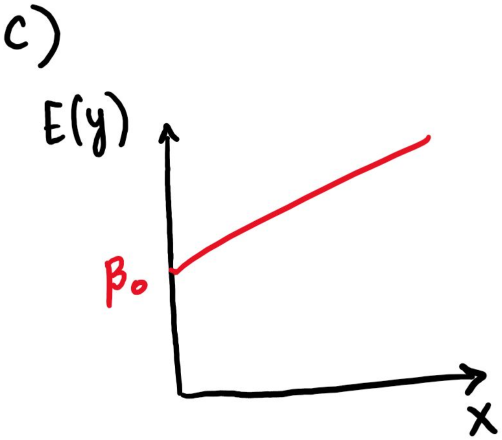
$E(y) = \beta_0 + (0)\,x$

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

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, $r^2 = 0.983$ or 98.3%

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, $r^2 = 0.983$ or 98.3%
- So hours of study are able to explain 98.3% 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, $r^2 = 0.983$ or 98.3%
- So hours of study are able to explain 98.3% of variation in test scores
- **GOOD FIT!**