Linear regression

The best gift from statistics to data scientists

Learning materials:

Some slides come from this awesome series video tutorials from **Brandon Folz**:
https://www.youtube.com/watch?v=ZkjP5RJLQF4

Concepts & examples from **OpenIntro Statistics**, chapter 7 (pages 331-371)

Learning goals

- ✖ Memorize all the mathematical process and the formulas
- Understand the main ideas and the intuition behind a linear model
- Use the SciKitLearn library to train a linear model using python, understanding the function parameters and output.
- Understand when would we use a linear model in real life.

What’s a model

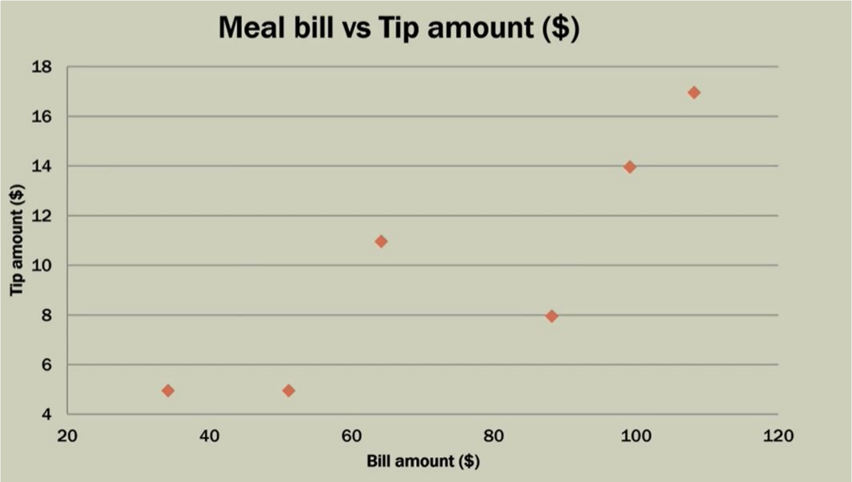
Modelling means trying to explain data with something more simple.

“All models are wrong – but some of them are useful.”

The linear model is one of the most simple models.

We want to predict the tip that customers give

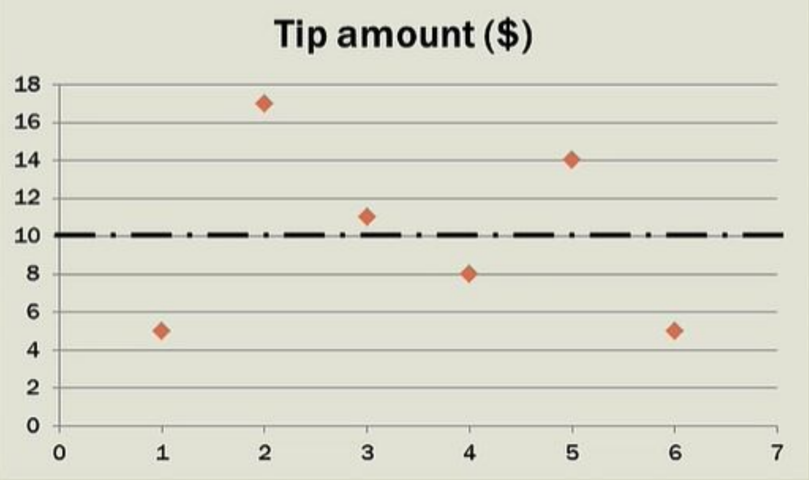
Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00
$\bar{x} = 74$	$\bar{y} = 10$



Check the correlation coefficient!

How do you make a prediction / estimate when you don’t have other variables?

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00



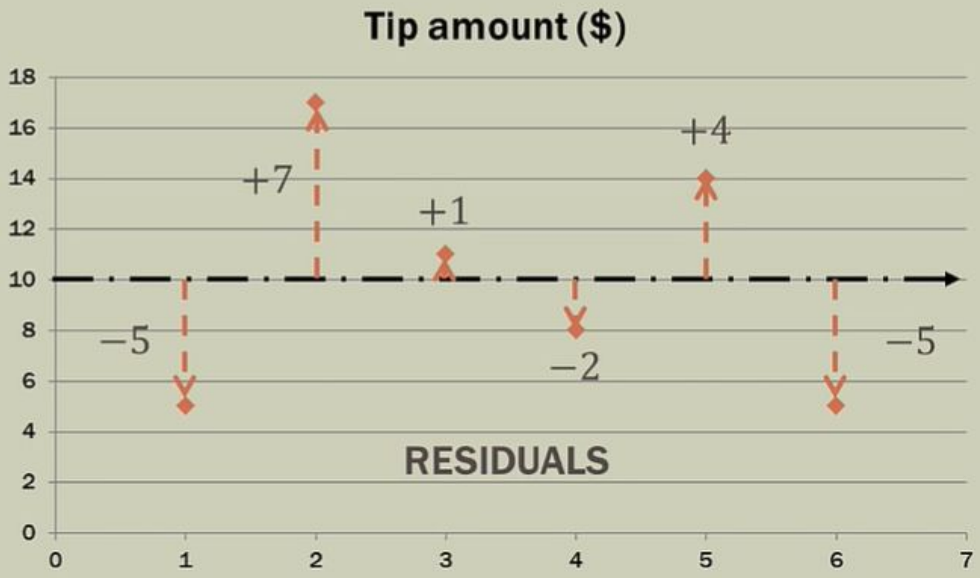
$\bar{y} = \$10$

$\bar{y} = \$10$

Each prediction comes with an error (residual)

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

$$\bar{y} = \$10$$



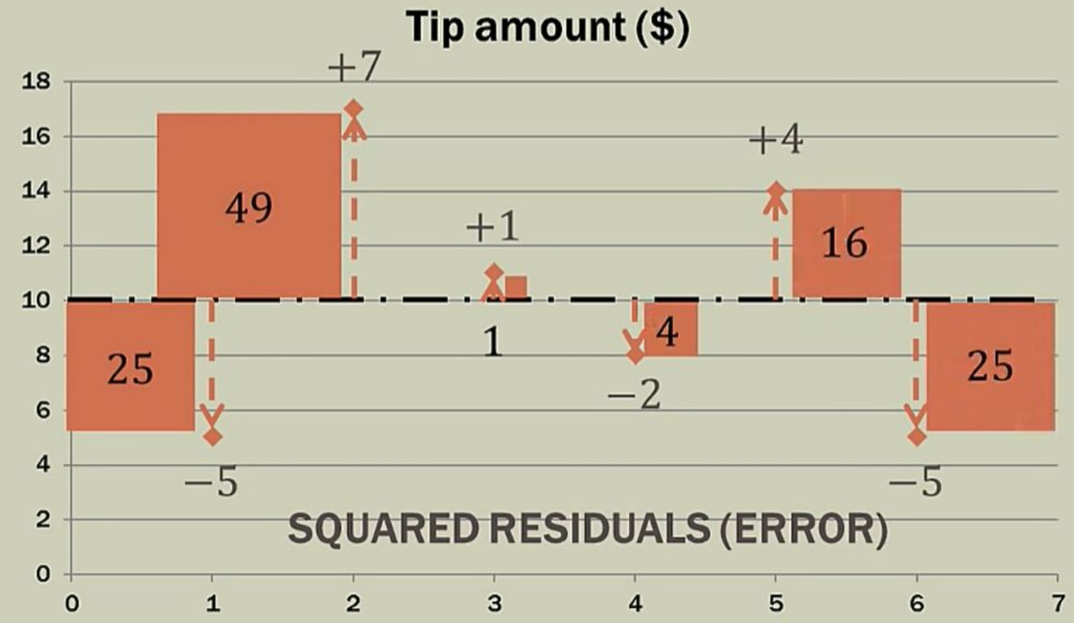
Residual: difference between observed and expected

The residual of the i^{th} observation (x_i, y_i) is the difference of the observed response (y_i) and the response we would predict based on the model fit $(\hat{y}_i)$:

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

We typically identify $\hat{y}_i$ by plugging x_i into the model.

The sum of the squared errors tells us how well does the line fit the data



Meal#	Residual	Residual^2
1	−5	25
2	+7	49
3	+1	1
4	−2	4
5	+4	16
6	−5	25

$$Sum\ of\ squared\ errors\ (SSE) = 120$$

Why square the residuals instead of using their absolute value?

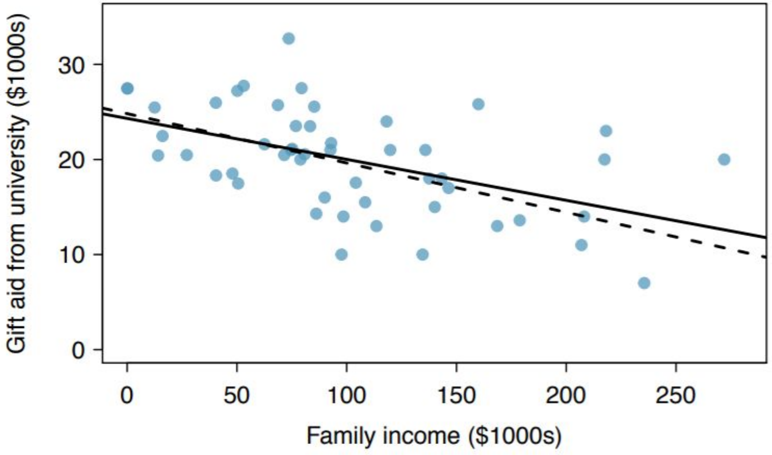


Figure 7.12: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College. Two lines are fit to the data, the solid line being the *least squares line*.

We want to have the least amount of error possible

The goal of simple linear regression is to create a linear model that minimizes the sum of squares of the residuals / error (SSE).

This is the SSE we got by just using the mean as a predictor. We're going to use it as our "baseline"

Let’s introduce the ‘explanatory variable’: the ‘bill’ amount

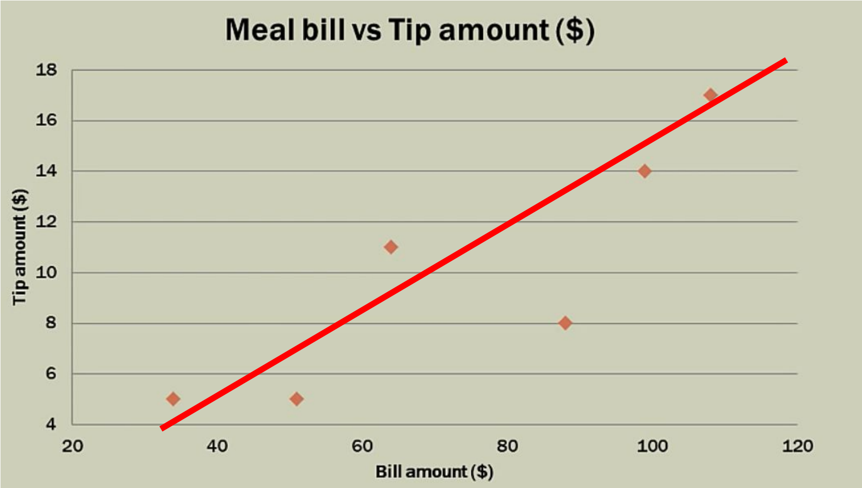
Objective: find the line that minimizes the sum of squared errors.

Process:

1. Plot the data on a scatter plot.

2. Look for a linear pattern. If there’s no linear trend, find another method.

3. Do some calculations (or… use python)



Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00

Least squares method

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

y_i = observed value of dependent variable (tip amount)

$\hat{y}_i$ = estimated(predicted)value of the dependent variable (predicted tip amount)

Plain English. The goal is to minimize the sum of the squared differences between the observed value for the dependent variable (y_i) and the estimated/predicted value of the dependent variable $(\hat{y}_i)$ that is provided by the regression line. Sum of the squared residuals.

Response or **dependent variable**: what we want to predict

Model parameters or **coefficients**. Our job is to estimate them!

$$y = \beta_0 + \beta_1 x$$

Intercept

Slope

Predictor, explanatory or **independent variable**: the information we have to predict.

Y-hat: or predictions

Estimation of parameters or **coefficients**.

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

Intercept

Slope

Explanatory variable

- Our job is to estimate β_0 and β_1
- Our estimates are called b_0 and b_1

Calculating the parameters*

Intercept

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

Slope

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

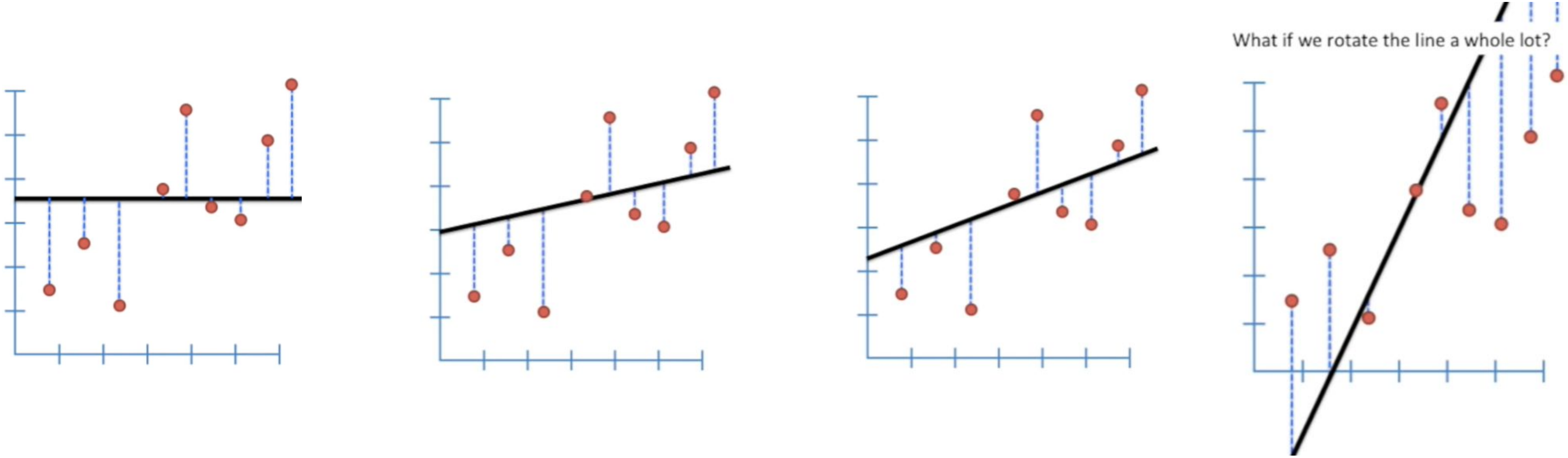
$\bar{x}$ = mean of the independent variable

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

x_i = value of independent variable

y_i = value of dependent variable

The gradient descent approach



Calculating the slope

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

1. For each data point.
2. Take the x-value and subtract the mean of x.
3. Take the y-value and subtract the mean of y.
4. Multiply Step 2 and Step 3
5. Add up all of the products.

---

1. For each data point.
2. Take the x-value and subtract the mean of x.
3. Square Step 2
4. Add up all the products.

Calculating the intercept

$$b_0 = \bar{y} - b_1\bar{x} \qquad b_1 = 0.1462$$

$$b_0 = 10 - 0.1462(74)$$

$$b_0 = 10 - 10.8188$$

$$b_0 = -0.8188$$

Total bill ($)	Tip amount ($)
x	y
34	5
108	17
64	11
88	8
99	14
51	5
$\bar{x} = 74$	$\bar{y} = 10$

Calculations step-by-step here:

https://docs.google.com/spreadsheets/d/1N046WwcIPaD5h5vFHoQ2ovntptXze7KvC-yuPP0hkaI/edit?usp=sharing

The regression line

$\hat{y}_i = b_0 + b_1 x_i$ $b_0 = -0.8188$ $b_1 = 0.1462$

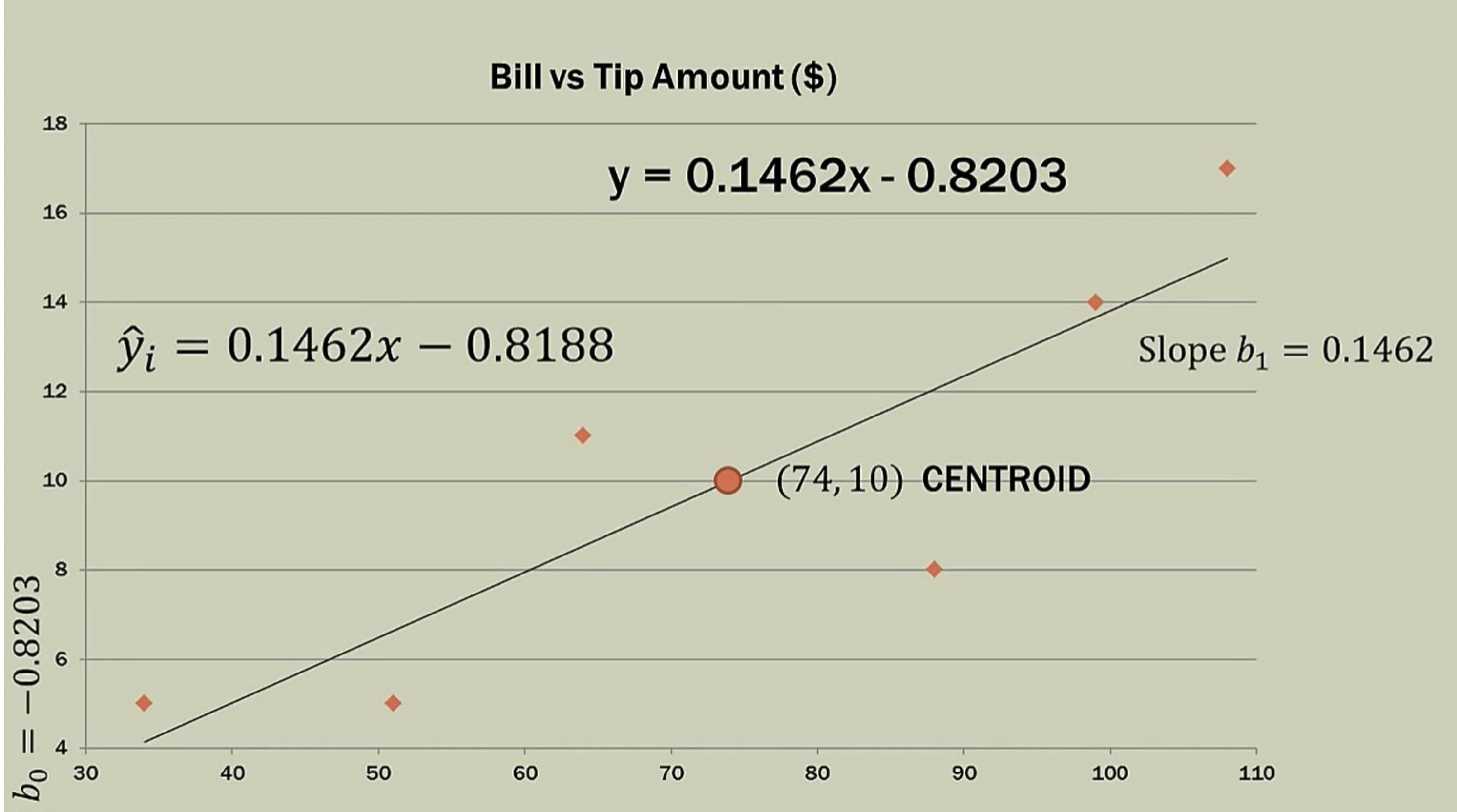
intercept slope

$$\hat{y}_i = -0.8188 + 0.1462x$$

OR

$$\hat{y}_i = 0.1462x - 0.8188$$

The regression line in the scatter plot



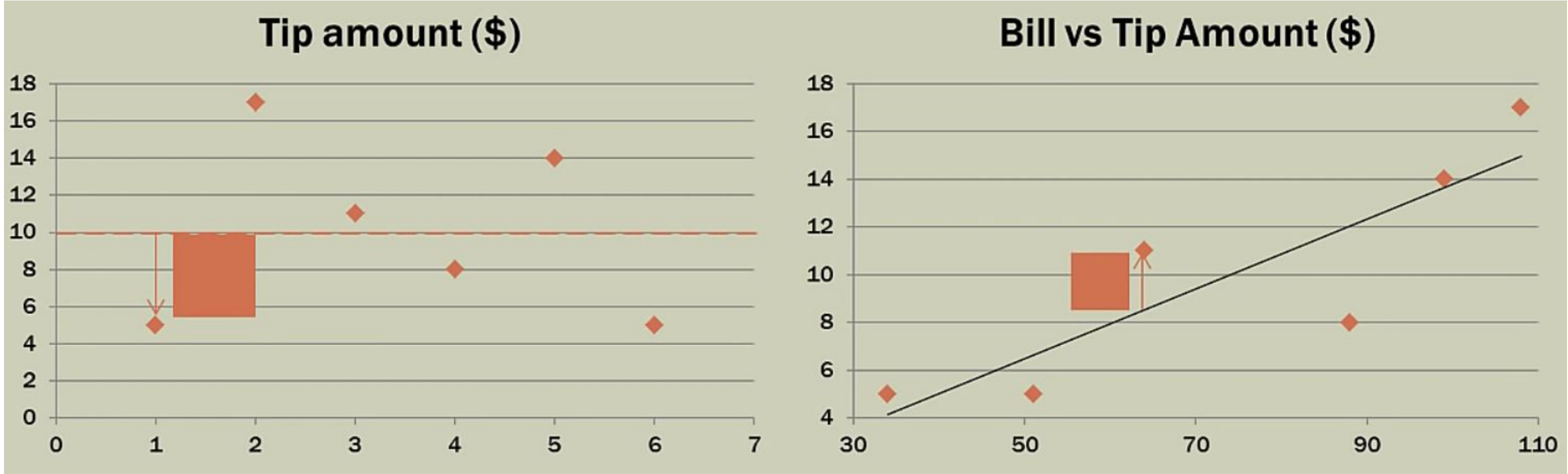
Interpretation

$$\hat{y}_i = 0.1462x - 0.8188$$

For every $1 the bill amount (x) increases, we would expect the tip amount to increase by $0.1462 or about 15-cents.

If the bill amount (x) is zero, then the expected/predicted tip amount is $-0.8188 or negative 82-cents! Does this make sense? NO. The intercept may or may not make sense in the "real world."

Evaluation: is the model better than just taking the mean?



Coefficient of determination = r^2

- How much better is the regression line compared to just using the mean of the response variable?

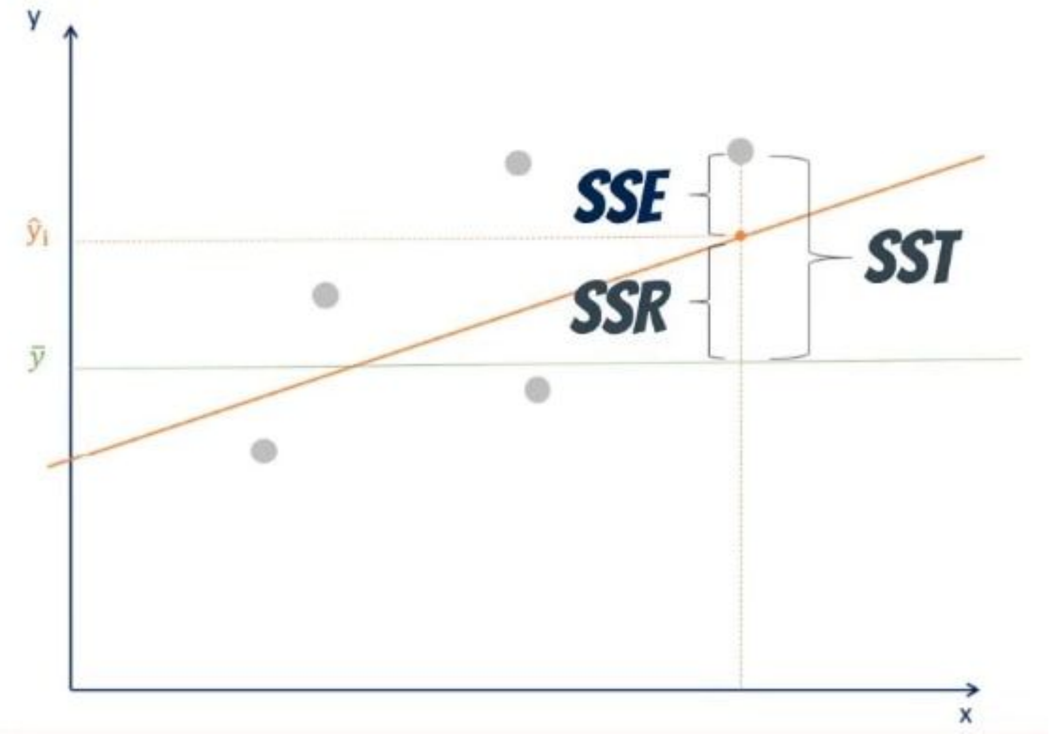
or, in other words...

- How much of the 'total error' does the regression 'solve'?

or, in other words...

- What percentage of the response variable variance does the regression model explain?

r-squared = SSR / SST



Calculating r-squared

$$Coefficient\ of\ Determination = r^2 = \frac{SSR}{SST}$$

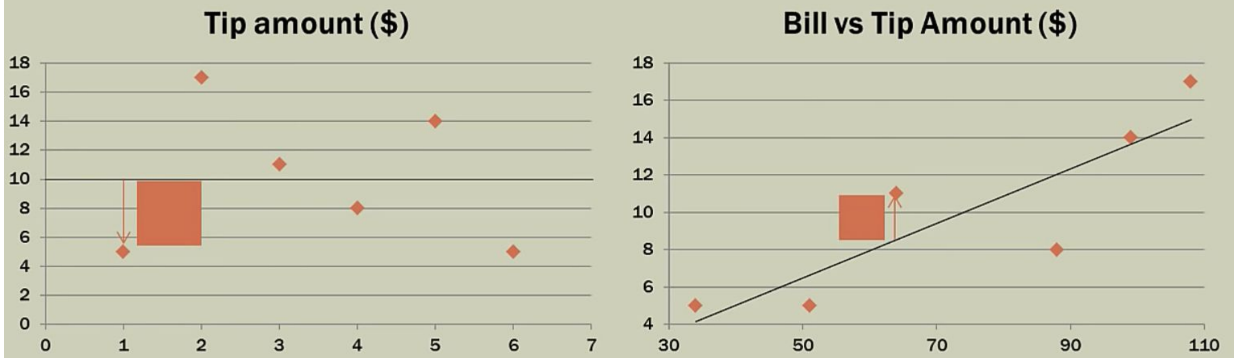
$$Coefficient\ of\ Determination = r^2 = \frac{89.925}{120}$$

$$Coefficient\ of\ Determination = r^2 = 0.7493\ or\ 74.93\%$$

Sum of Squares Total (SST): squared differences between the **observed** dependent variable and its mean.

Sum of Squares Regression (SSR): Sum of the differences between the **predicted** dependent variable and its mean

How much of the 'total error' does the regression 'solve'?*



$SSE = 120$ $SST = 120$ $SSE = 30.075$

$SSE = SST$ $120 - 30.075 = SSR$

$SST = 120$ $120 - 30.075 = 89.925$

In other words: what percentage of the variance does the model explain?

The linear model sometimes sucks

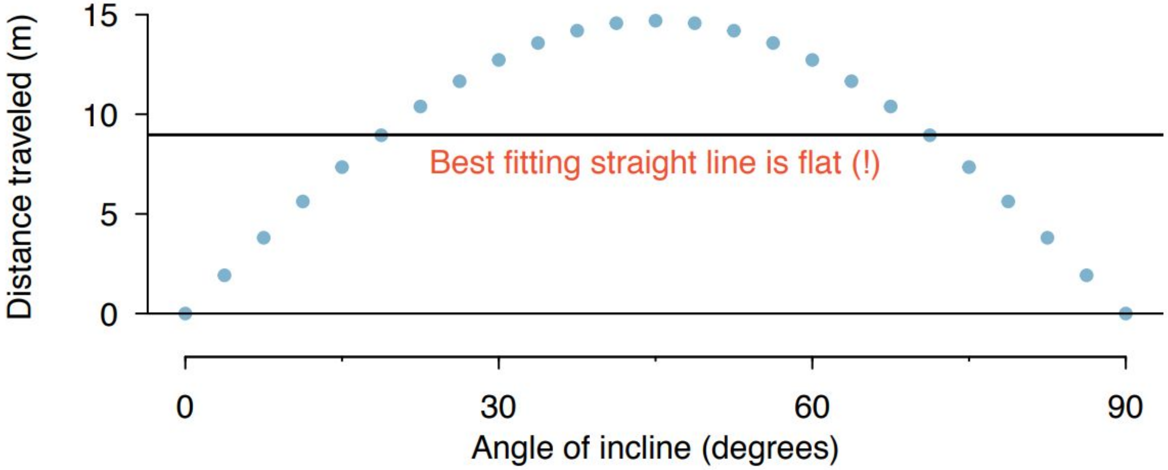
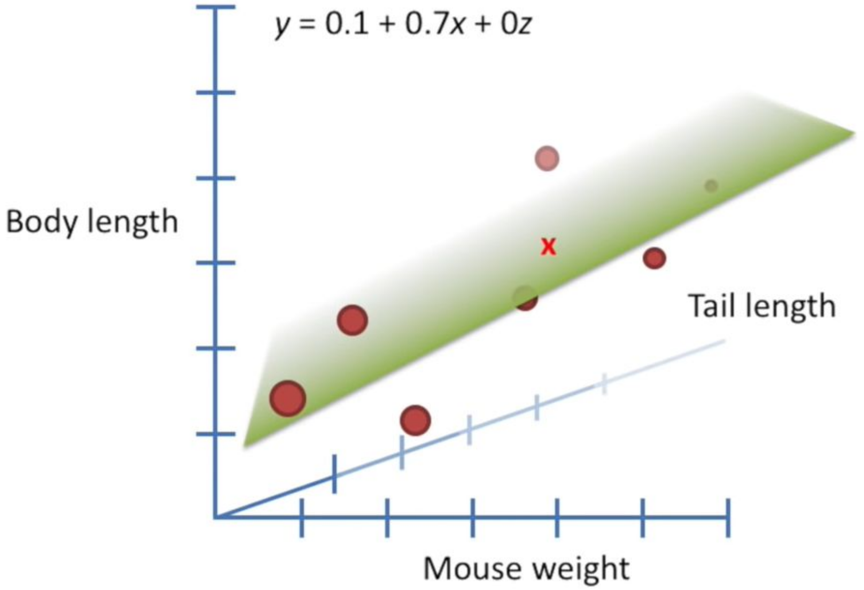
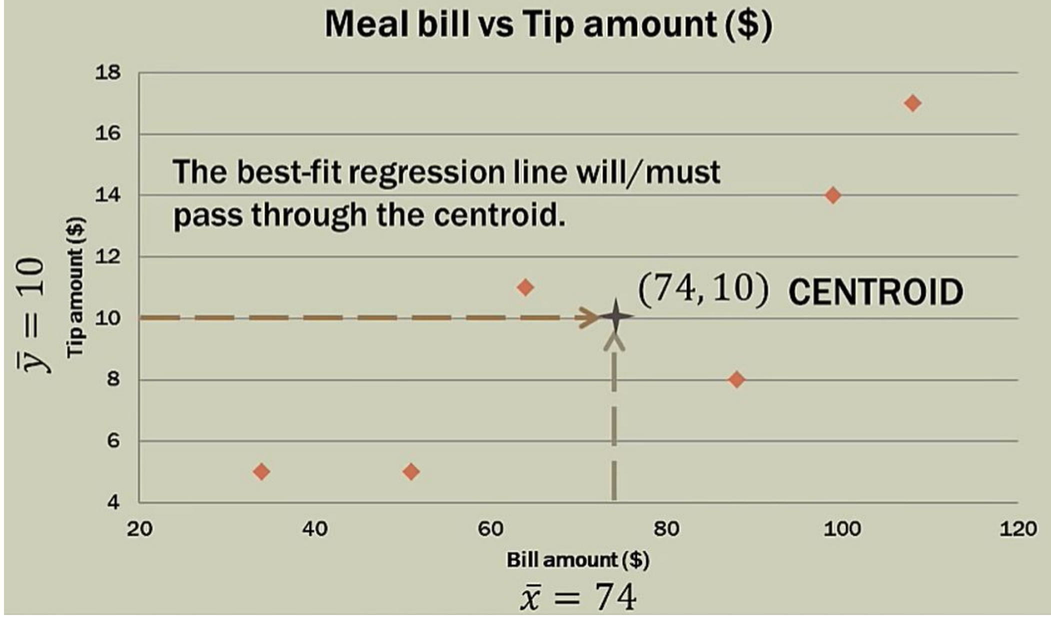


Figure 7.3: A linear model is not useful in this nonlinear case. These data are from an introductory physics experiment.

Multiple regression: you can add more variables!



The centre of the data



Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00
$\bar{x} = 74$	$\bar{y} = 10$

Why would we ever use a linear model instead of a more advanced model?

- Simplicity
- Interpretability
- Generalization
- Baseline for oher models

The expected value of y is the mean of a distribution of possible values