1495 UNIVERSITY OF ABERDEEN

CELEBRATING 525 YEARS 1495 – 2020

ABERDEEN 2040

Regression Analysis

Data Mining & Visualisation

Lecture 9

2025

Recap

- Types of Data
  - Quantitative vs Categorical Variables
- Relationships in data
  - Correlation

Today

- Supervised Learning
- Regression Analysis
- Evaluating a Regression Model

Recap: Supervised Learning

Supervised Learning

As we discussed during Lecture 2, there are a lot of overlaps between data mining and machine learning.

While the exact steps, terminology, and end-goals may differ, various aspects of today's lecture should be familiar to you.

In the next few lectures, we will discuss a few supervised learning methods that can be used for data mining and analysis.

Supervised Learning

Supervised learning is a type of machine learning where labelled training data is used to train models.

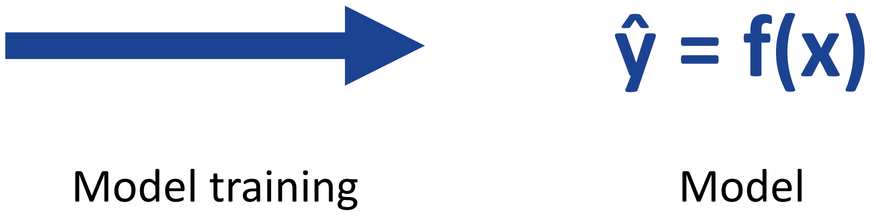
Our models learn the relationships between these variables based on the labelled data that we provide.

Supervised Learning

Example: Based on a housing dataset, we could train a model to map a house's size (in sq. ft.) to the value of that house.

Price ($)	Size (Sq. ft.)
100,000	1,000
250,000	2,500
400,000	4,000
…	…

Labelled Data



$\hat{y} = f(x)$

Model training

Model

Terminology

In supervised learning, we frequently refer to variables as either *dependent* or *independent*.

A **dependent variable** (DV) is dependent on other variables. Typically, this is what we are interested in understanding or predicting (i.e. the **output** of our model).

An **independent variable** (IV), or predictor, is what we use to understand or predict our dependent variable (i.e. the **input**).

Terminology

In the example where we compute the price from the size:

- The size of the house is our **independent variable** (IV).
- The price of the house is our **dependent variable** (DV), because our model's output (price) is dependent on the size.

However, note that for a separate model where we compute the size from the price, the price (our model's input) would be the IV, and size (our model's output) would be the DV.

Prediction Vs Interpretation

Within supervised learning, once we have trained our model, there are two common goals that we might want to use our model to achieve:

Prediction and **Interpretation**.

Prediction

In a ‘typical’ ML context, we might want to determine how well our model can be used to **predict** new (unseen) data.

We might first *train* a model, *test* it with a withheld dataset, and then *deploy* it to predict new data.

> E.g. if we have a house with 1,500 sq. ft., what should its valuation be?

Interpretation

However, when performing data analysis, we might be more interested in what we can **infer** from the model.

In other words, we might want to understand what the underlying relationships within the data can teach us.

> E.g. How does a house’s size influence its valuation?

Prediction Vs Interpretation

Note that in practice, we might want **both** prediction and interpretation. Both are valid end-goals to have.

However, there is often a trade-off between these two goals.

Certain supervised learning algorithms will typically be better for prediction and worse for interpretation (and vice versa).

Regression Vs Classification Problems

When we discuss supervised learning, there are two broad types of problems that tend to come up:

- **Regression problems**, where our DV is quantitative

- **Classification problems**, where our DV is categorical

Note that for both regression and classification problems, our IV(s) can be quantitative and/or categorical!

Regression Analysis

Regression Analysis

Regression analysis refers to a **set** of statistical methods for quantifying relationships between variables.

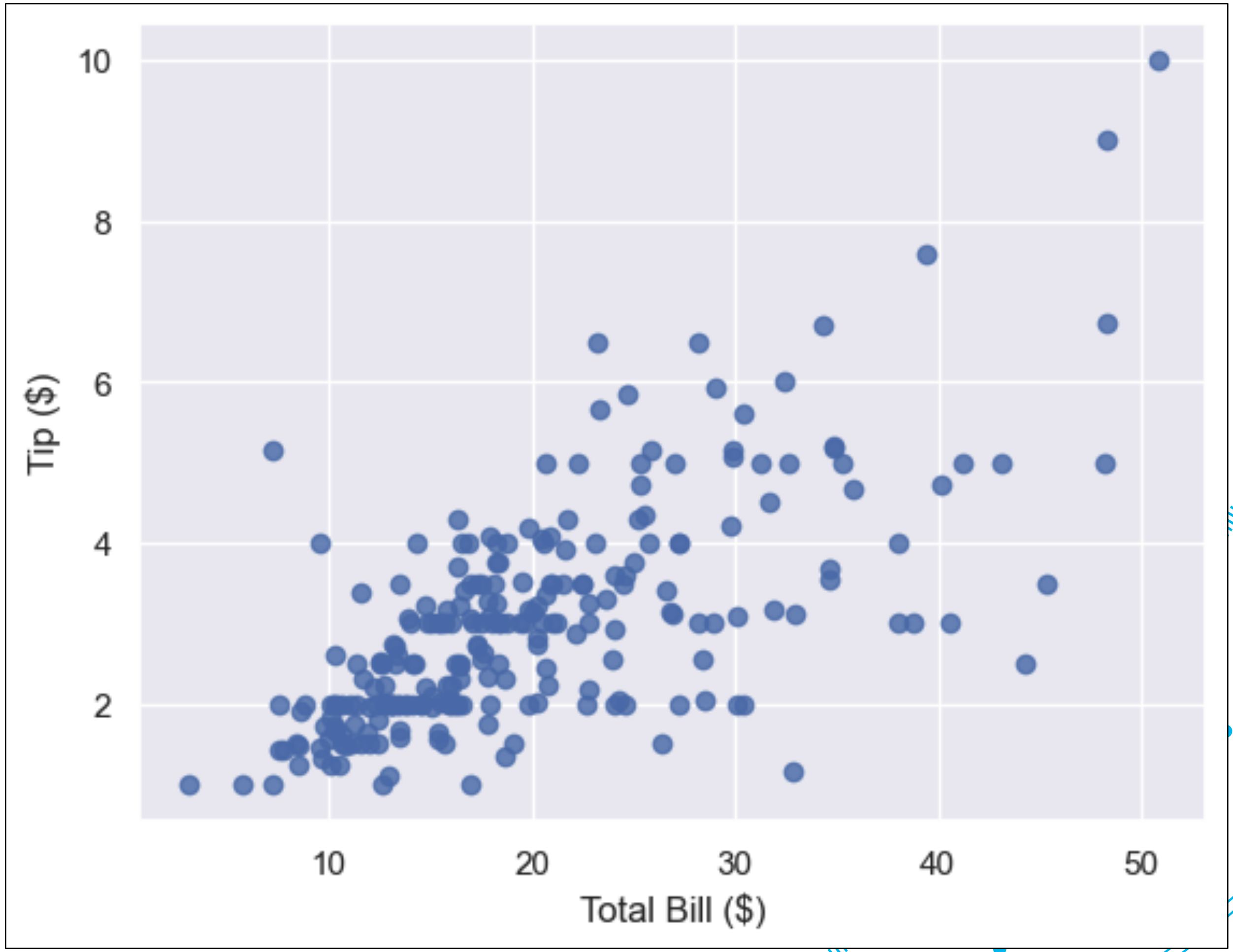
In this way, it is similar to correlation – however, regression allows us to do much more, above and beyond correlation!

They are some of the simplest (and most interpretable) techniques for supervised learning.

Regression Analysis

Let’s go back to some of our correlation examples.

Example: Do larger bills result in larger tips?

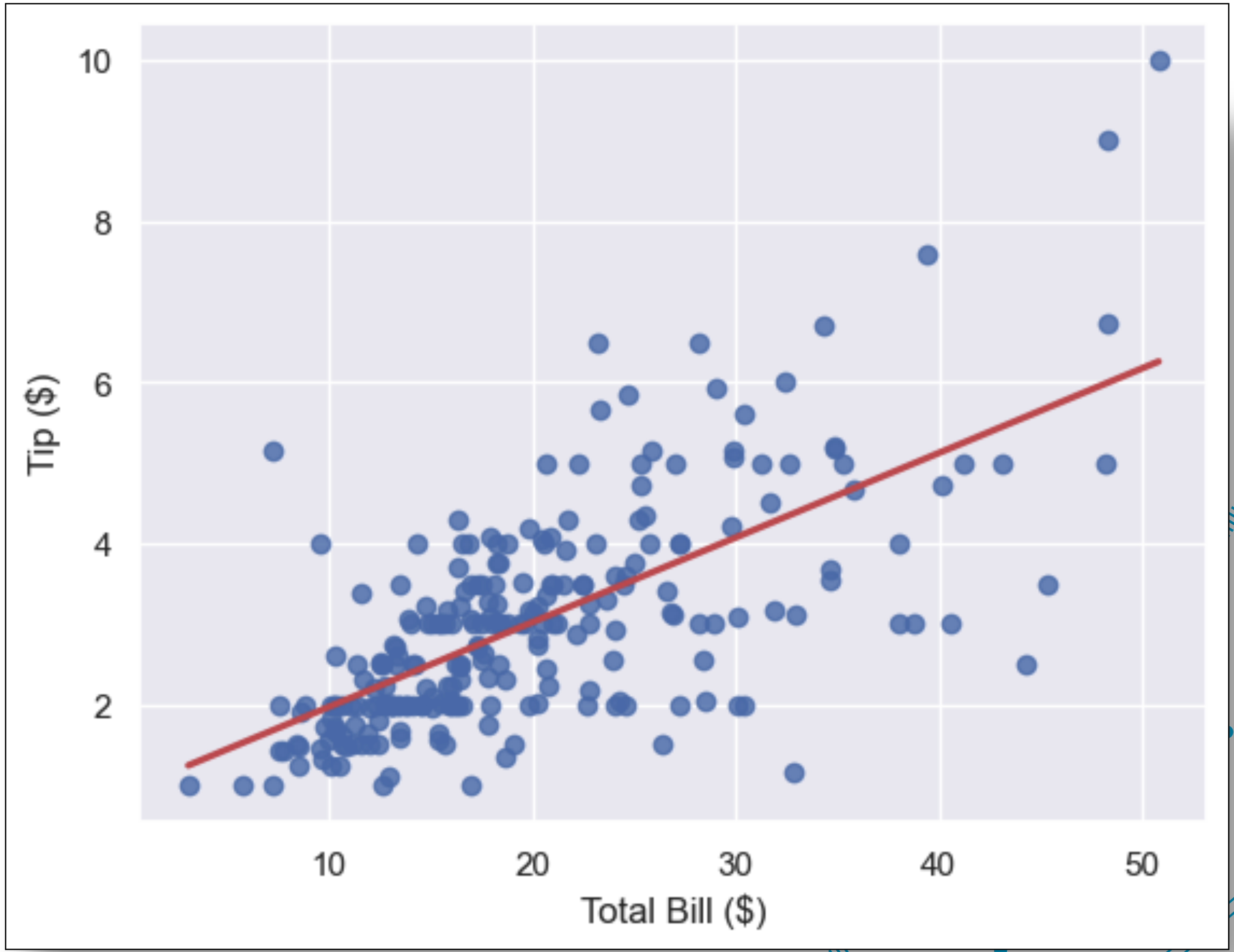


Regression Analysis

In previous examples, we have used the ‘line of best fit’ to visualise our correlations.

This is our regression model. It estimates the function that most closely fits the data.

Here, we have one predictor (IV).

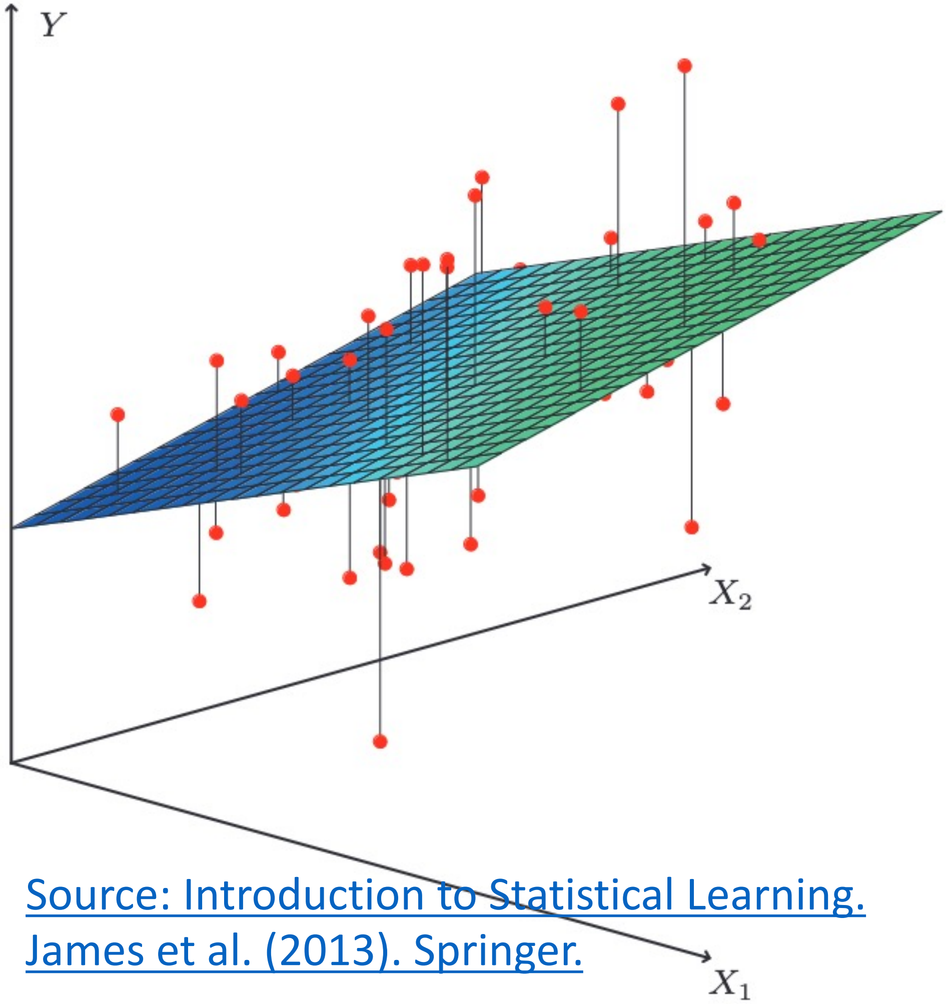


Regression Analysis

But regression allows us to have as many as we want.

Here, we now have 2 IVs (X_1 and X_2), and instead of a line, we are now using a plane.

But we could have hundreds of IVs if we wanted to.



Source: Introduction to Statistical Learning. James et al. (2013). Springer.

Regression Models: Under the Hood

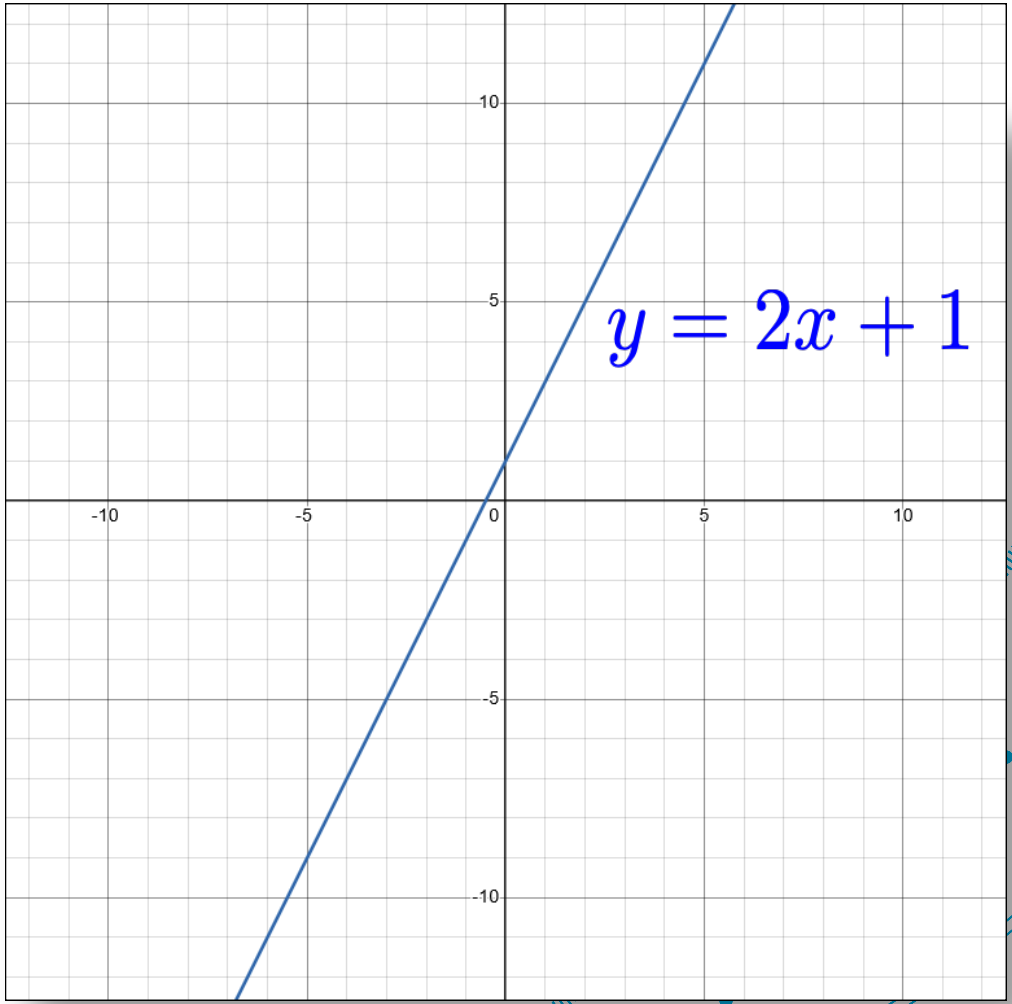
Regression Model

To understand what these regression models are, recall the equation of a line:

$$y = mx + c$$

With one IV, the formula for regression is very familiar:

$$y \approx \beta_0 + \beta_1 x_1$$

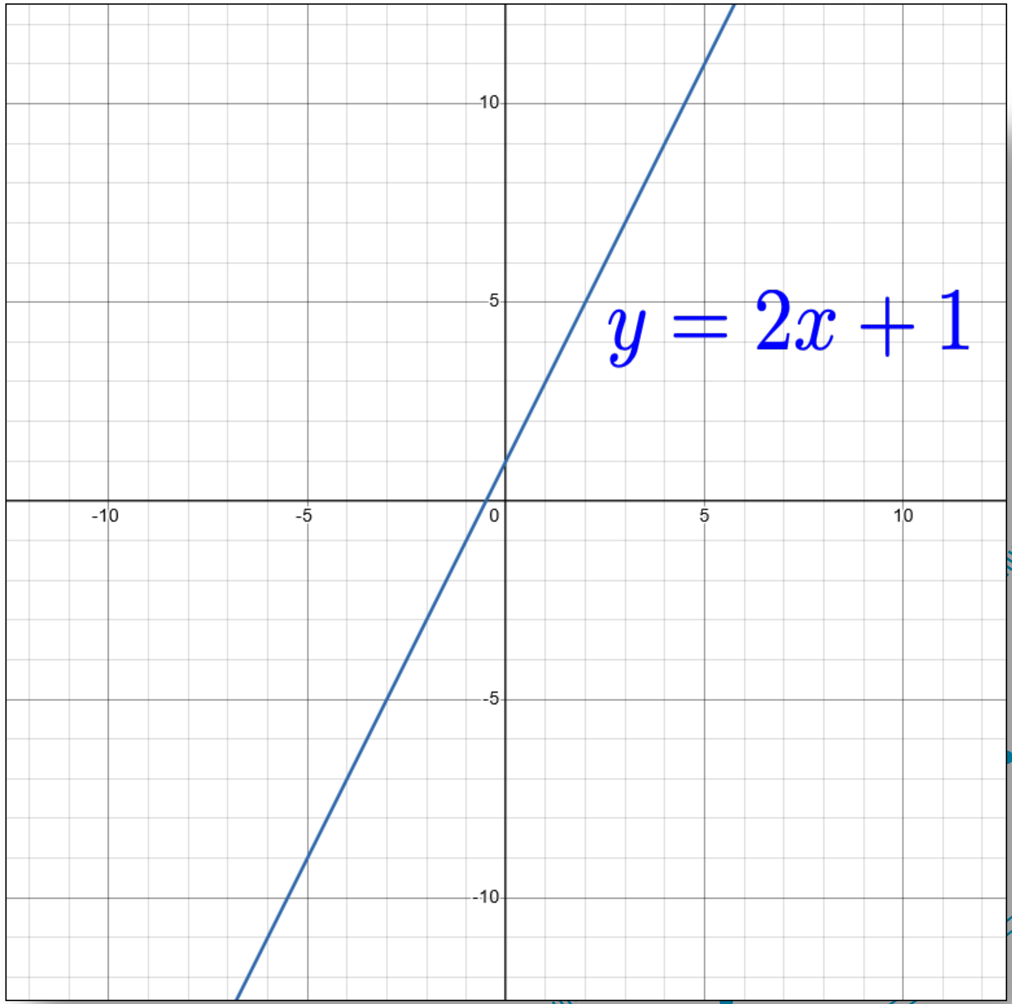


Regression Model

$$y \approx \beta_0 + \beta_1 x_1$$

Here:

- y is our DV
- x_1 is our IV
- β_0 is our y intercept
- β_1 is our gradient for x_1



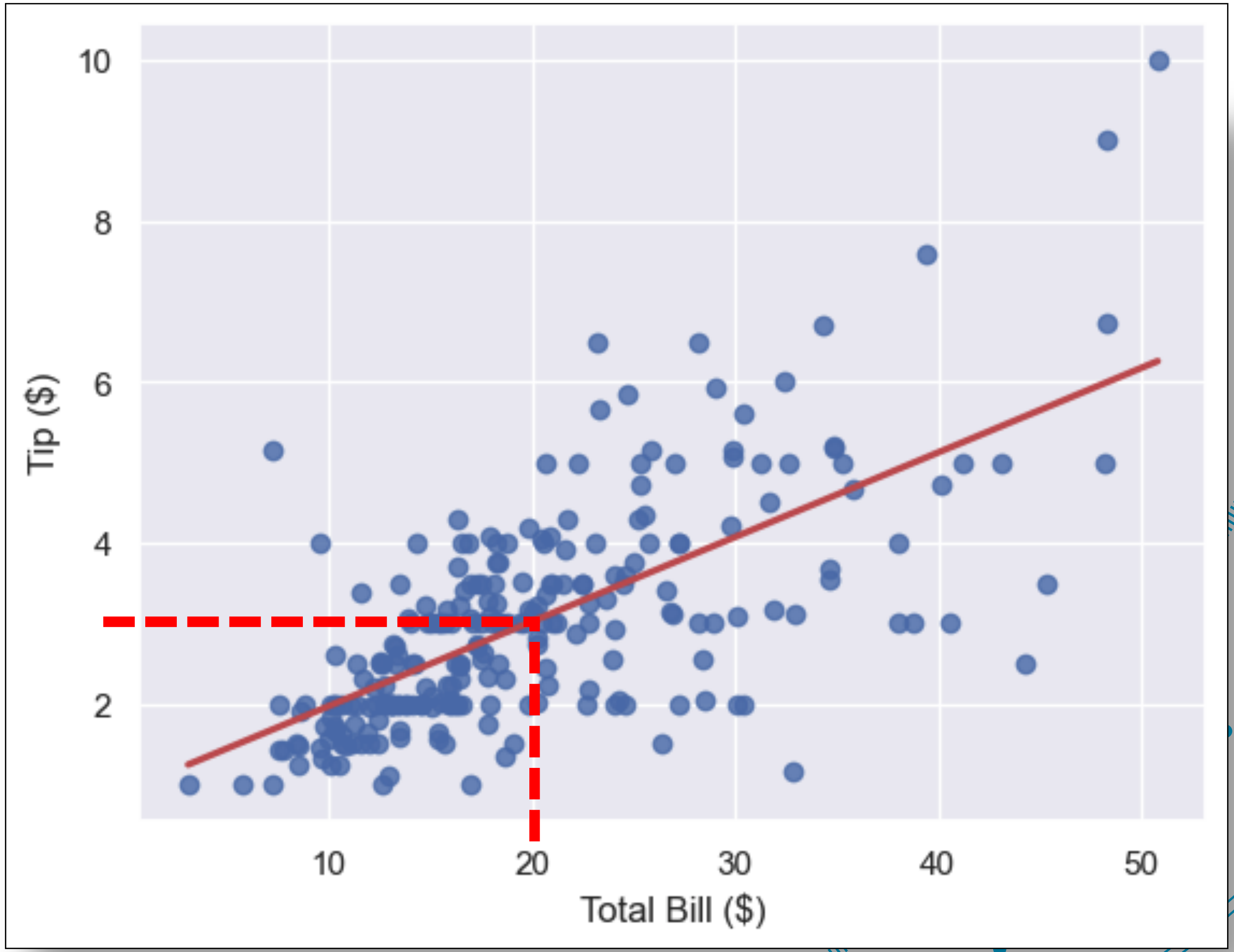
Regression Model

Intercept: $0.92

Gradient: 0.105

If we have a total bill of $20, our tip will estimate to:

> $y \approx \beta_0 + \beta_1 x_1$
>
> $y \approx 0.92 + (0.105 * 20)$
>
> $y \approx \$3.02$



Evaluating a Regression Model

Evaluating a Regression Model

We will often want to know how well our model 'fits' the data.

There are lots of ways of evaluating a regression model.

Note: You don't have to **memorise** these formulas, but you will be expected to understand the intuition of each metric, and be able to identify and differentiate between them!

Sum of Squared Errors (SSE)

The coefficients in a linear regression models will often be estimated using the 'Ordinary Least Squares' (OLS) method.

Here, the aim is to minimise the Sum of Squared Errors (SSE).

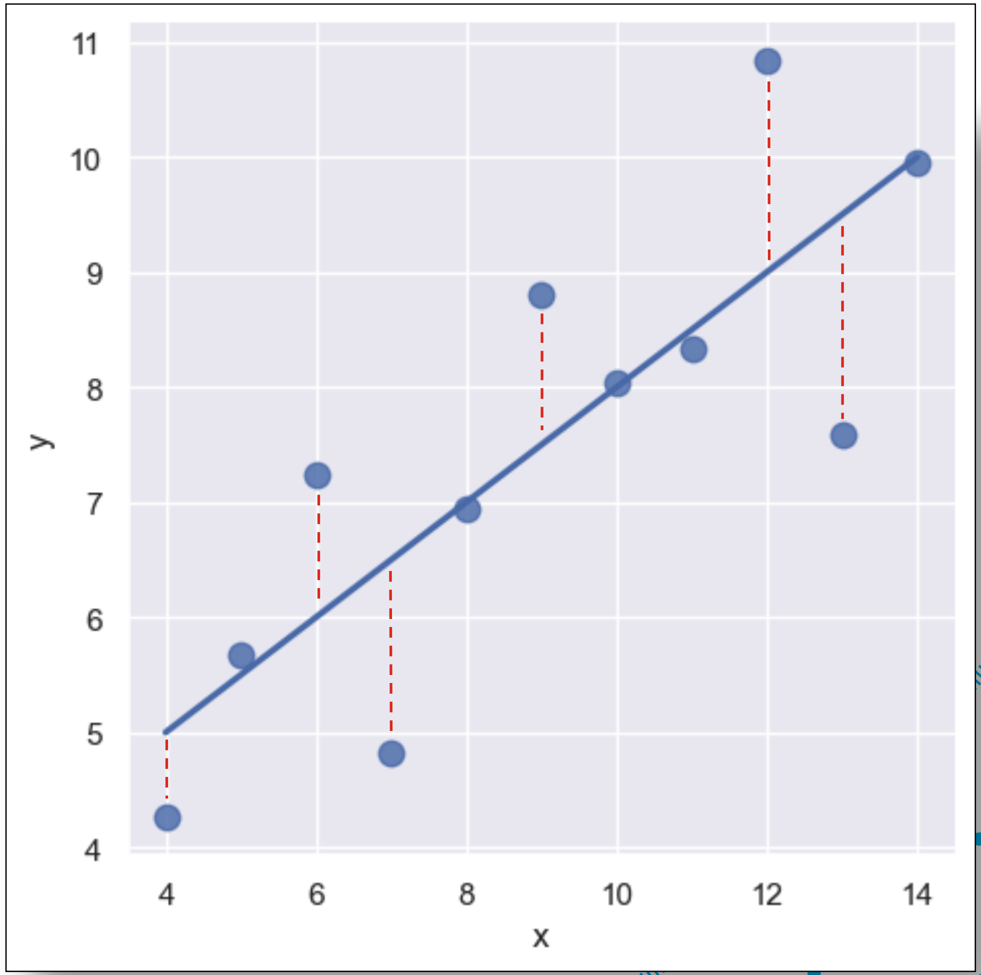
Note that SSE is sometimes referred to as Sum of Squared Residuals (SSR), and Residual Sum of Squares (RSS), but we'll stick with SSE in this course.

Sum of Squared Errors (SSE)

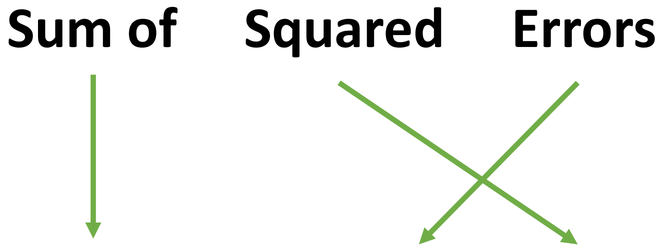
For each data point, we square the difference between the actual & estimated values.

Then, we add these up:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



Sum of Squared Errors (SSE)



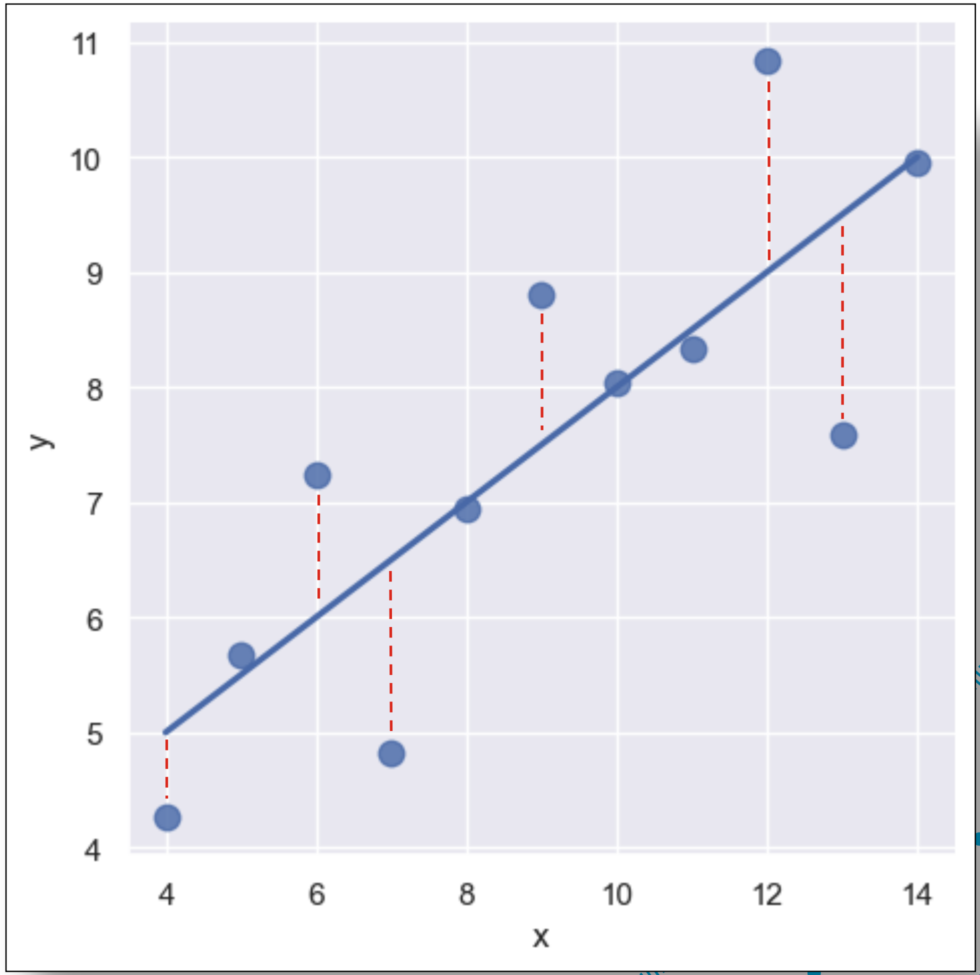
Sum of

Squared

Errors

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

A small SSE represents a close fit.



Mean Squared Error (MSE)

SSE is good for comparing different regression models against a dataset, but the SSE value will increase as *n* increases.

So instead, we can use Mean Squared Error (MSE), by multiplying SSE by 1 / (n – p – 1), where p is the number of predictors (IVs).

$$MSE = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE)

So, MSE represents the average **squared** error.

If we find the root of this, we find the average error, which we call Root Mean Squared Error (RMSE), or Standard Error (SE)

$$RMSE = \sqrt{\frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Evaluating a Regression Model

In other words, these three metrics are iterations of one another.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$MSE = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

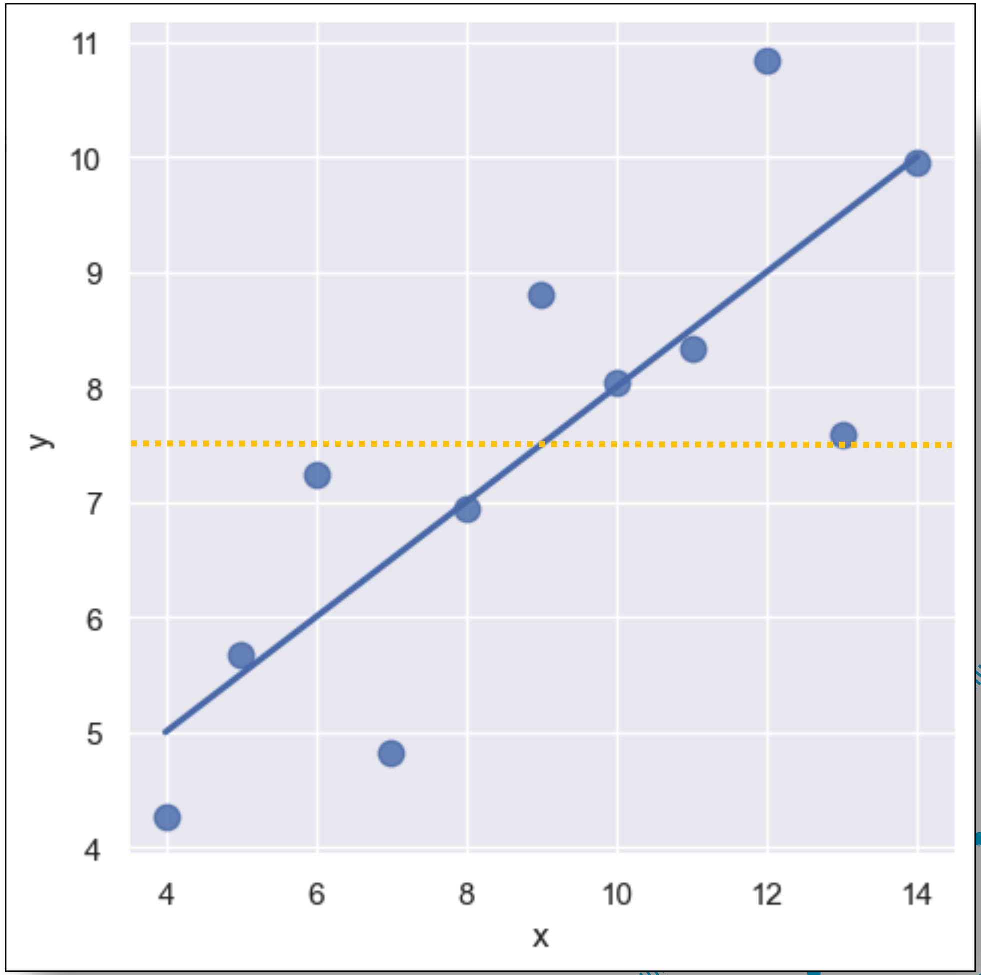
$$RMSE = \sqrt{\frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

R^2 Value

Another way to evaluate our model is by comparing it to just using the mean.

Here:

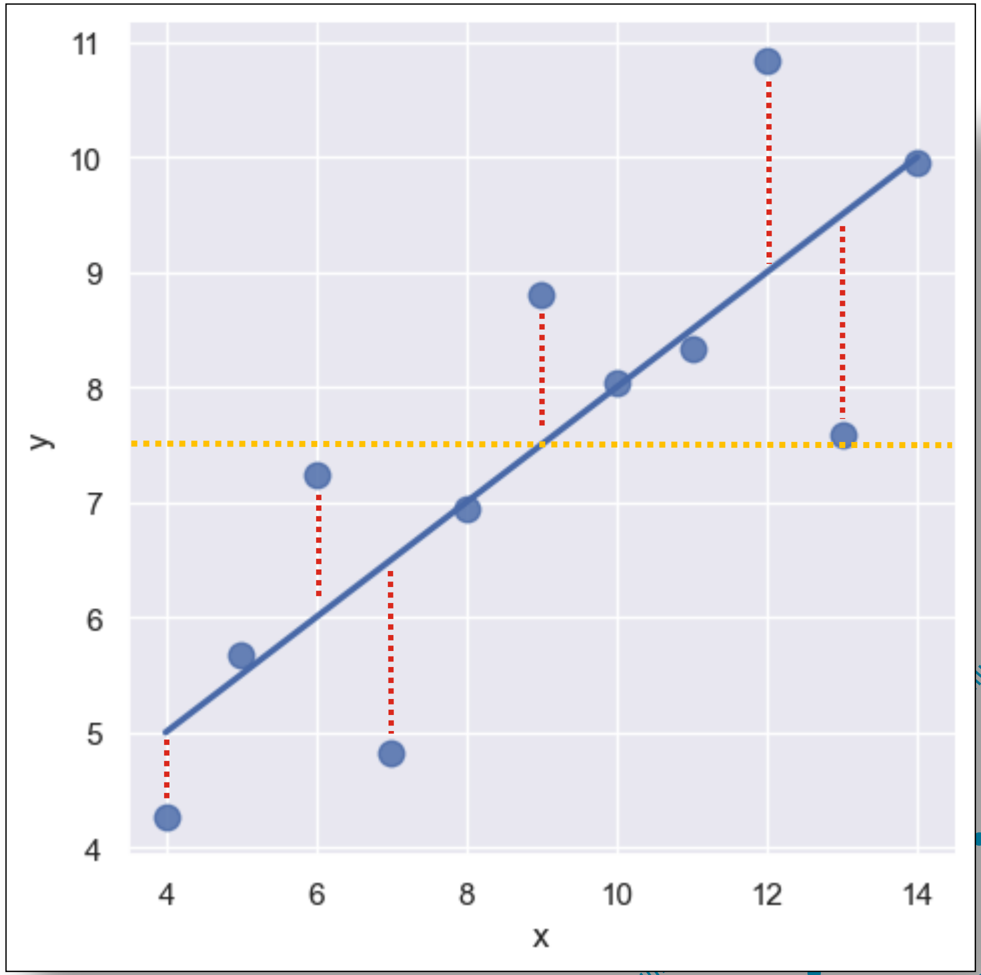
- Blue line: $y \approx \beta_0 + \beta_1 x_1$
- Yellow line: $y \approx \frac{1}{n}\sum_{i=1}^{n} x_i$



R^2 Value

Recall the SSE calculation:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



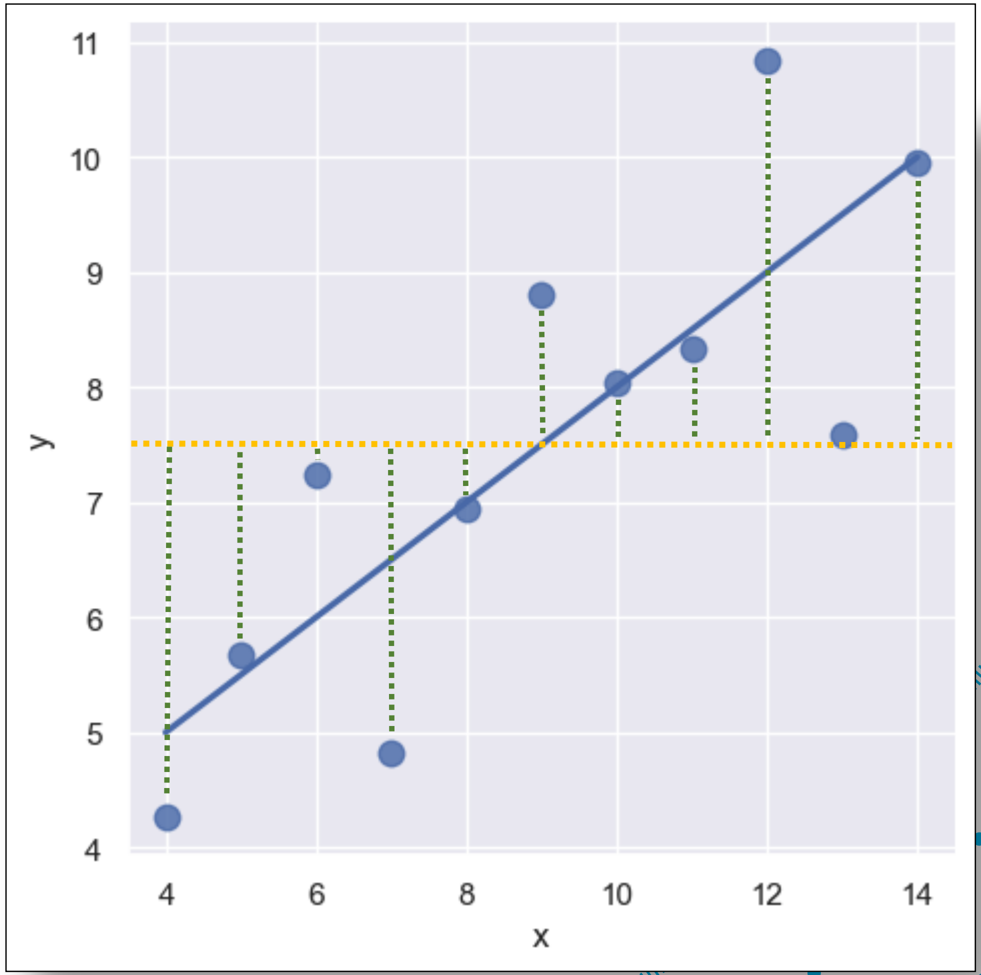
R² Value

Recall the SSE calculation:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Let's say we calculate the Sum of Squares Total (SST), using the mean instead of ŷ

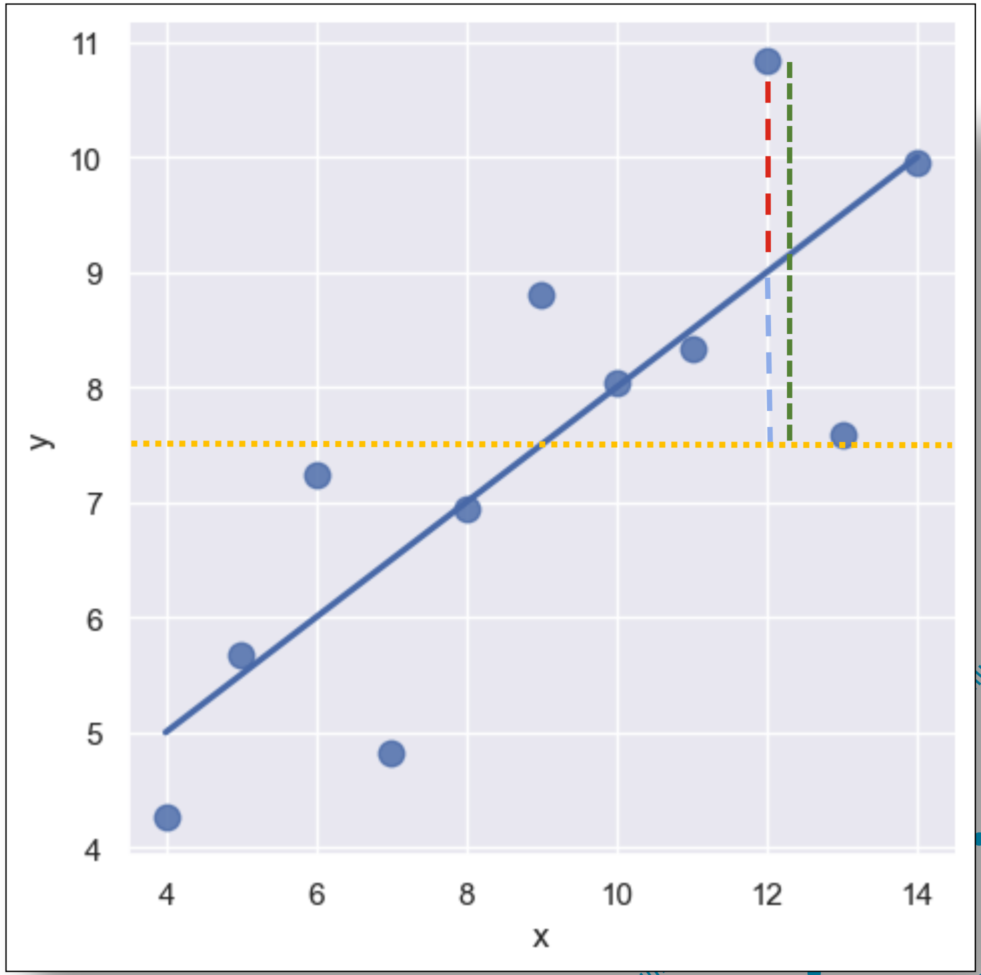
$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$



R^2 Value

One way to think of this is that we're differentiating between:

- *variance that is **explained** by our model (┈),*
- *variance that is **unexplained** by our model (┈),*
- *and the total variance (┈)*

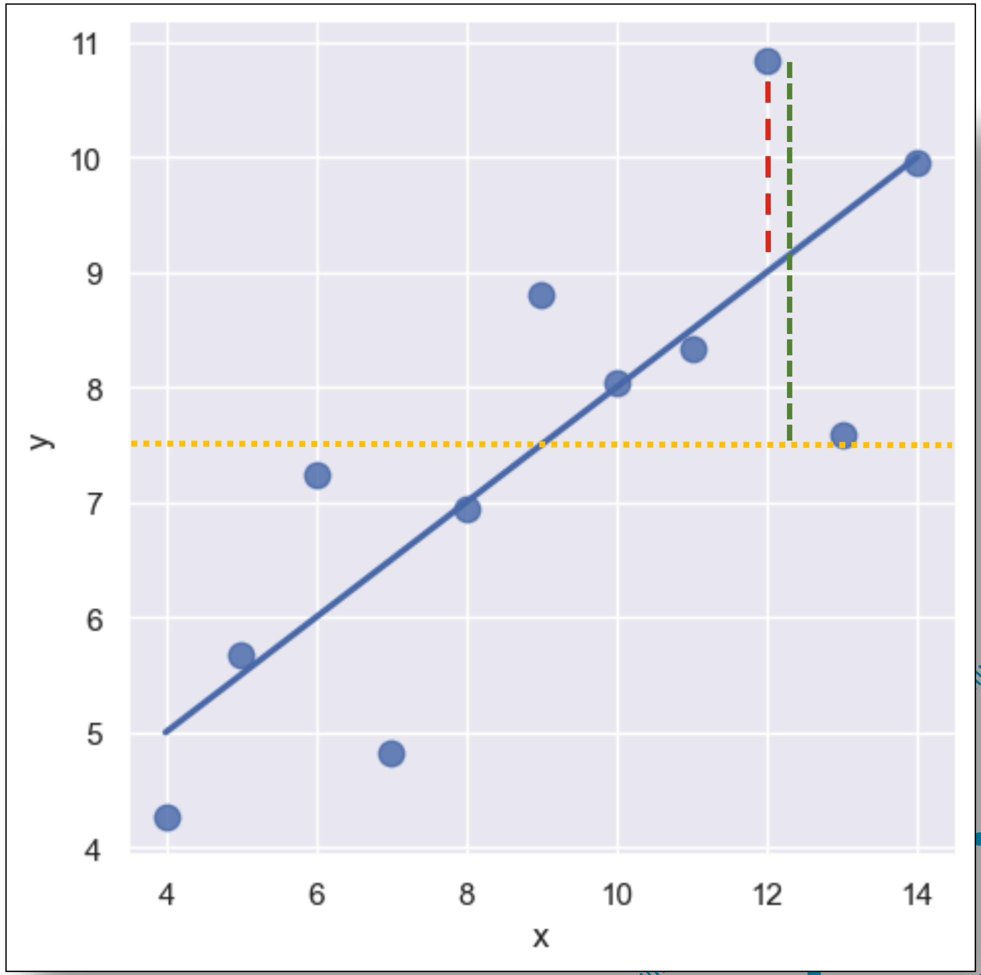


R^2 Value

- R^2 tells us what proportion of the total variance is explained by our regression model

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

$$= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

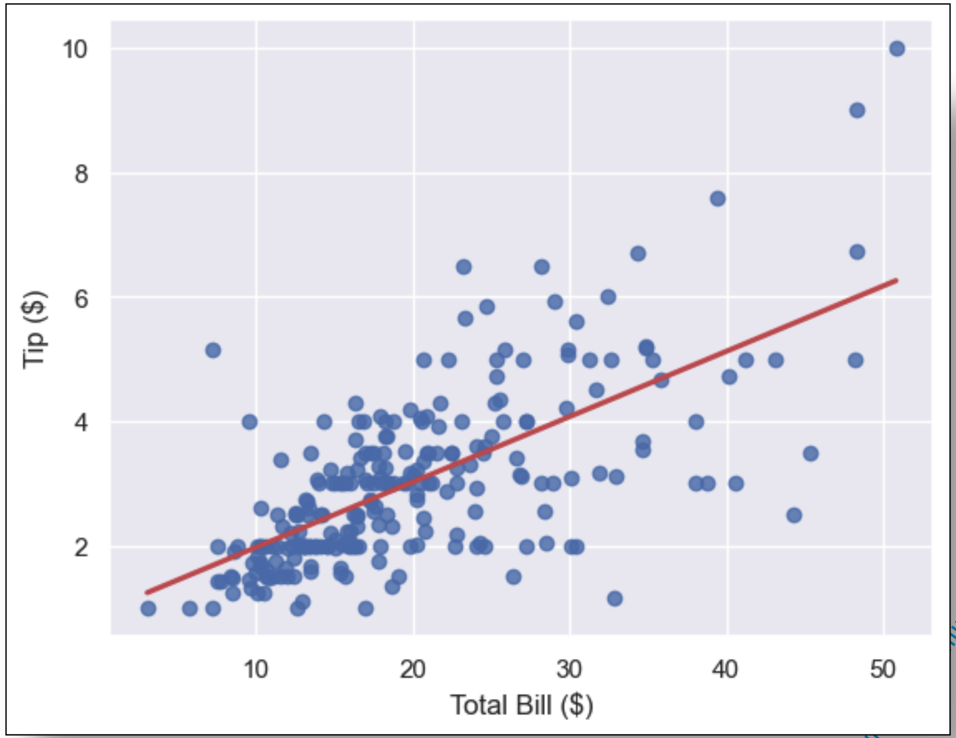


R^2 Vs RMSE

So, to clarify,

R^2 tells us **what proportion of the total variance is explained by our regression model**.

RMSE tells us **the average error between predicted and actual values** (in units of y).



Intercept:	$0.92
Gradient:	0.105
RMSE:	$0.007
R^2:	0.457

So, What's the Best Metric? R^2 or RMSE?

Both metrics tell us something useful about the model.

But keep in mind that evaluation metrics will often depend on context, and any one metric may not tell the whole story.

Therefore, <u>it is important to look at different metrics, and to visualise data (where possible) to gain a broader understanding</u>.

What Can We Do With Regression?

What Can We Do With Regression

So we’ve trained a regression model using a dataset – great!

But what can we do with this regression model?

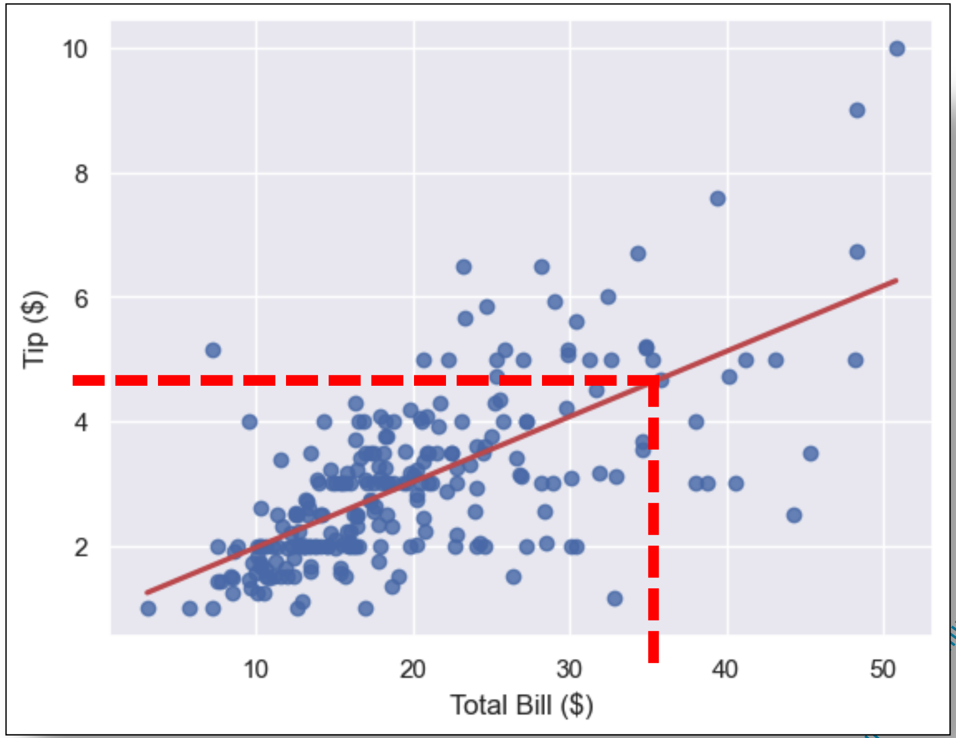
We’ve already differentiated between prediction and interpretation in supervised learning, so let’s demonstrate each of these within the context of regression.

Interpreting A Regression Model

Prediction: We can use X values to predict new Y values.

If we have a total bill of $35, our tip will estimate to:

> $y \approx \beta_0 + \beta_1 x_1$
>
> $y \approx 0.92 + (0.105 * 35)$
>
> $y \approx \$4.60$



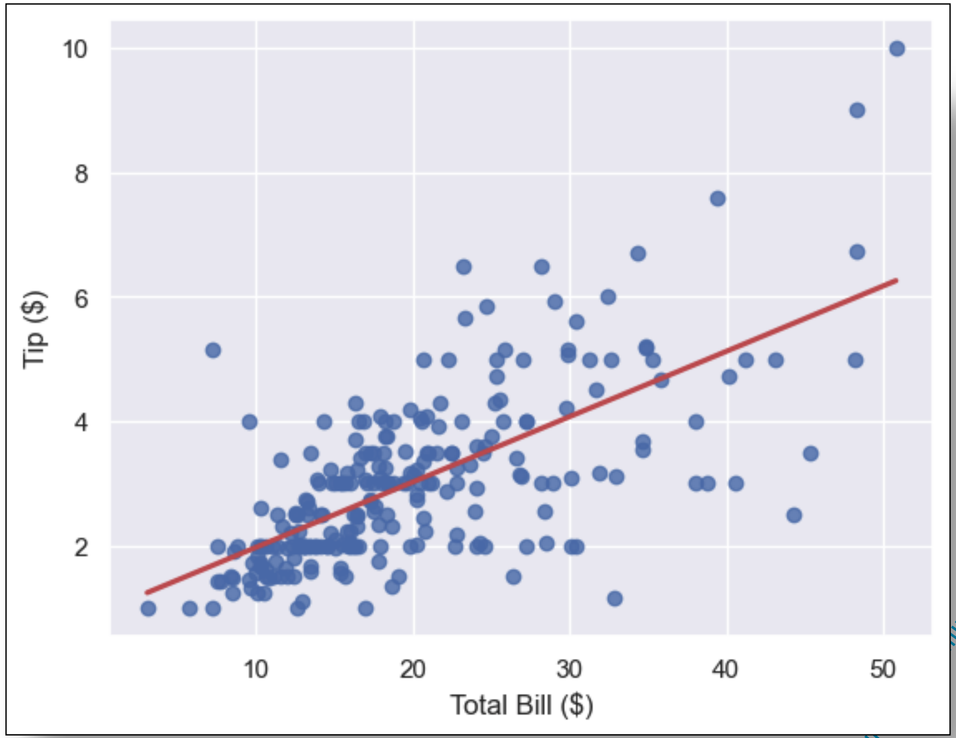
Intercept:	$0.92
Gradient:	0.105
RMSE:	$0.007
R^2:	0.457

Interpreting A Regression Model

Interpretation: We can use X values to understand how they influence the Y value.

For each $ the total bill increases, the tip tends to increase around $0.105.

This is in addition to a constant value of $0.92 (our intercept).

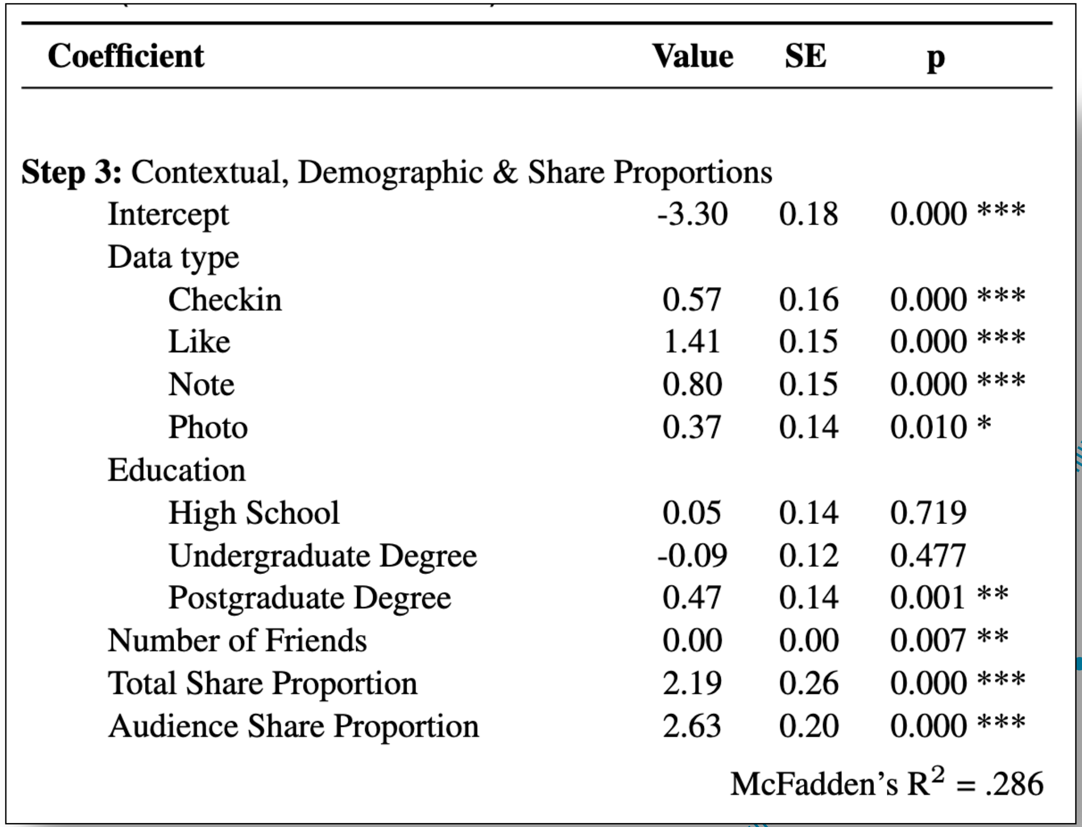


Intercept:	$0.92
Gradient:	0.105
RMSE:	$0.007
R^2:	0.457

Interpreting A Regression Model

Note that in situations where we have several variables,

interpretation becomes **very** useful for understanding how our variables relate to each other!



Coefficient	Value	SE	p
Step 3: Contextual, Demographic & Share Proportions			
Intercept	-3.30	0.18	0.000 ***
Data type			
Checkin	0.57	0.16	0.000 ***
Like	1.41	0.15	0.000 ***
Note	0.80	0.15	0.000 ***
Photo	0.37	0.14	0.010 *
Education			
High School	0.05	0.14	0.719
Undergraduate Degree	-0.09	0.12	0.477
Postgraduate Degree	0.47	0.14	0.001 **
Number of Friends	0.00	0.00	0.007 **
Total Share Proportion	2.19	0.26	0.000 ***
Audience Share Proportion	2.63	0.20	0.000 ***

McFadden’s R^2 = .286

Other Types of Regression

Multiple Linear Regression

Again, we will often be interested in understanding the effects of more than one IV (predictor) on our DV.

This is often referred to as 'multiple linear regression’, and mathematically, the idea is the same:

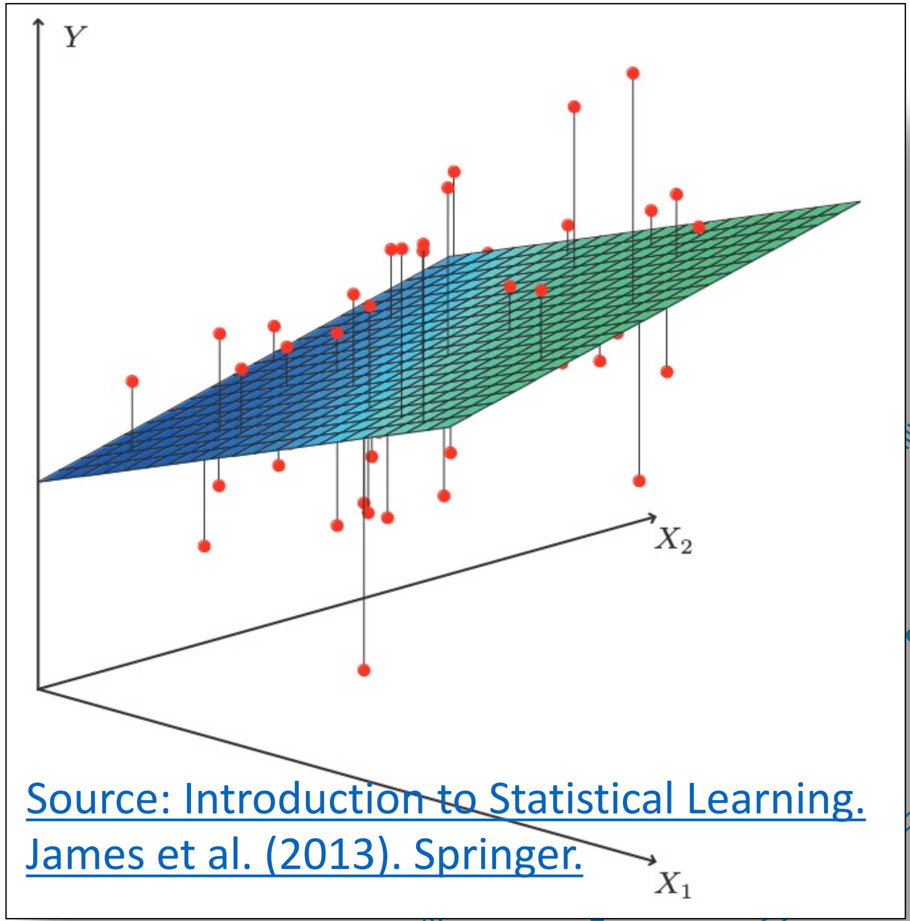
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Intercept IV 1 IV 2 And so on…

Multiple Linear Regression

As the number of IVs increase, it becomes a little more complicated to conceptualise visually.

But for 2 IVs (X_1 & X_2), instead of fitting a line, we can just think of fitting a 'plane of best fit'.



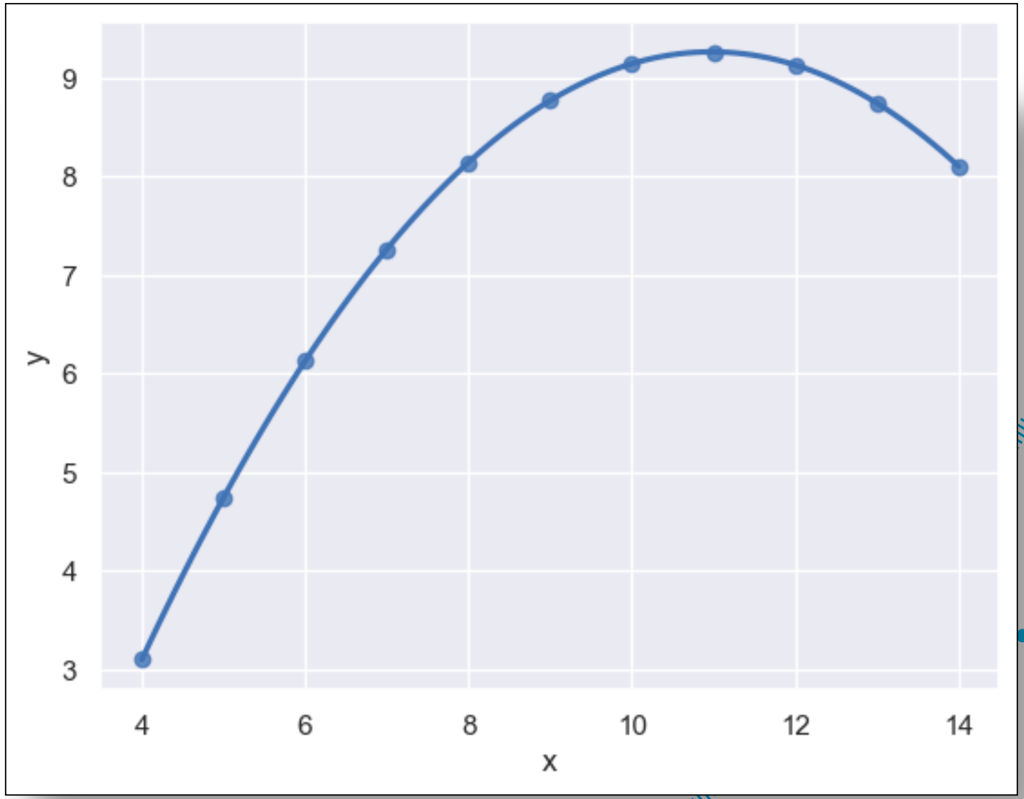
Source: Introduction to Statistical Learning. James et al. (2013). Springer.

Non-Linear Regression

We can also build regression models with non-linear functions.

This can result in a better 'fit' when relationships are non-linear.

However, interpretability becomes more of a challenge.

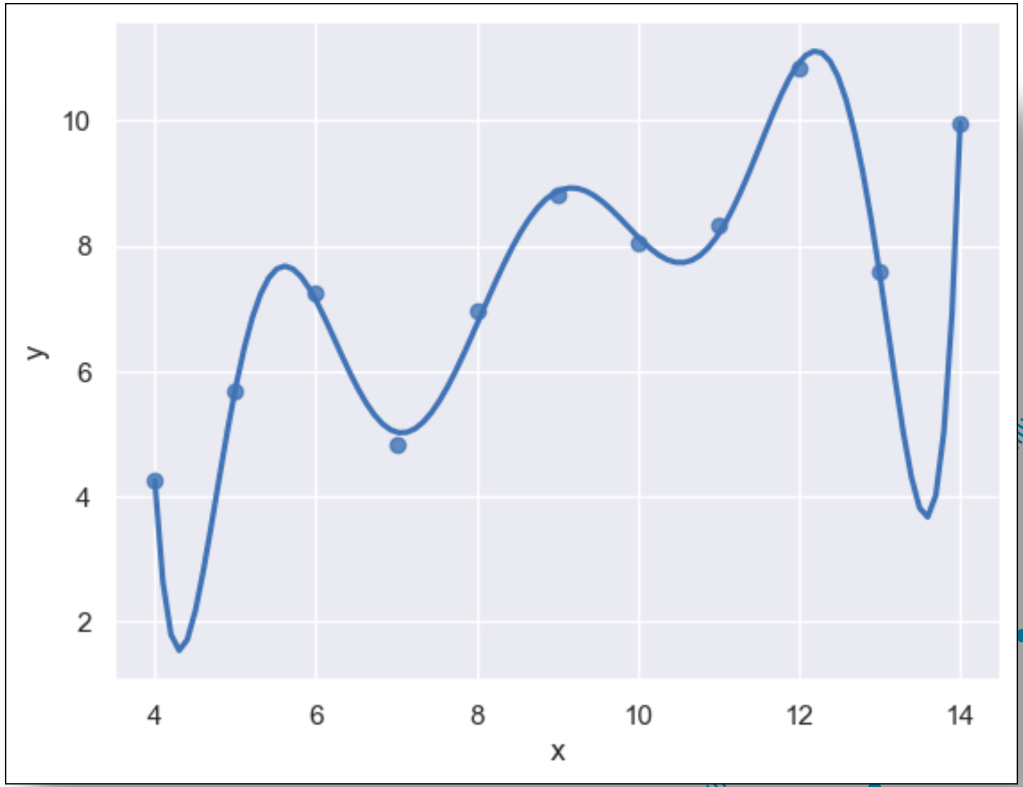


Non-Linear Regression

It also risks overfitting to the data.

This is especially the case when curves are made too complex.

Again, Python won’t stop you, so it’s important to be aware of this!



Limitations of ‘Ordinary’ Linear Regression

In short, regression is a very simple, interpretable, and useful method for analysing quantitative DVs.

But what if we want to predict a *categorical* variable?

Next up, we’ll talk a little bit about Logistic Regression, which will allow us to do this!