Session 10: Linear regression

wifi: GA-Guest, yellowpencil

GENERAL ASSEMBLY

```
cd ~/Documents/ga-ldn-ds37
git commit -am “your commit message here”
git pull
```

Today’s session plan

1800-1830	Standup & review
1830-1900	Linear regression theory
1900-1920	Break
1920-2100	Linear regression exercises
US Presidential election exercise	

At the end of the session, you will be able to ...

Understand how linear regression models are fitted to data

Appreciate pitfalls to bear in mind when interpreting the coefficients of linear regression

Assess the goodness of fit of a model

Visualise a linear model in simple cases

Data Science Part Time

What is linear regression?

Linear regression

Where are the independent and dependent variables?

What do the coefficients and constants mean?

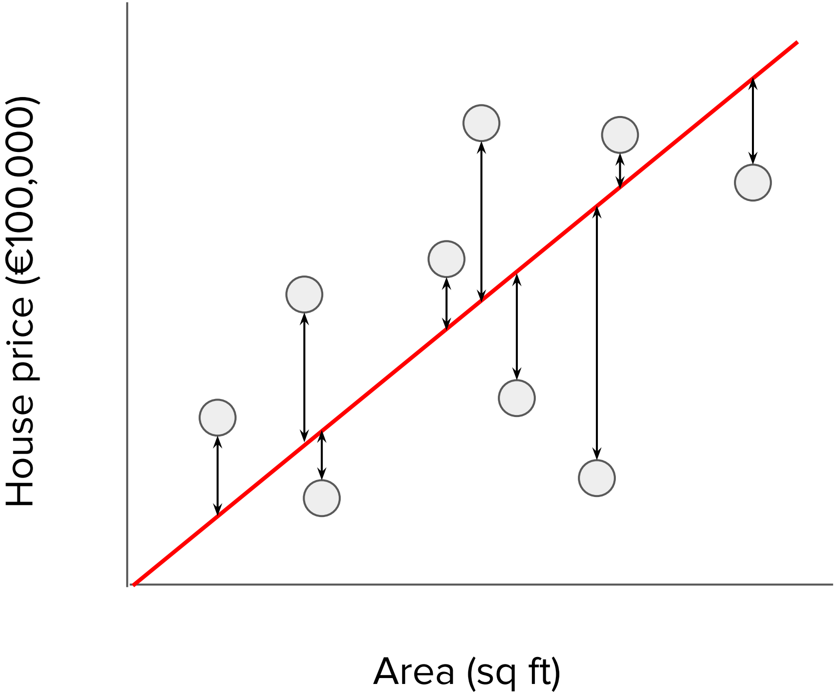
We can have one feature...

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

Or multiple features...

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

Finding the line of best fit



To fit a linear model, we minimise the sum of squared errors.

Model Prediction

$$SS_{residuals} = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Observed Result

Solo Exercise:

Sum of squared residuals

10 minutes

Let’s cement our understanding of linear regression.

We have the following data:

```
X = [1, 1, 5, 7]     y = [4, 5, 9, 12]
```

Find the sum-of-squared residuals between the data and each of these two linear models, and choose the model that best fits the data:

```
y = 2x + 2
```

```
y = x + 3
```

Solo Exercise:

Sum of squared residuals

10 minutes

$X = [1, 1, 5, 7]$ $y = [4, 5, 9, 12]$

$y = 2x + 2$:

$y_{hat} = [4, 4, 12, 16]$

$$\begin{aligned}
SSR &= (4-4)^2 + (5-4)^2 + (9-12)^2 + (12-16)^2 \\
&= 0 + 1 + 9 + 16 \\
&= 26
\end{aligned}$$

$y = x + 3$:

$y_{hat} = [4, 4, 8, 10]$

$$\begin{aligned}
SSR &= (4-4)^2 + (5-4)^2 + (9-8)^2 + (12-10)^2 \\
&= 0 + 1 + 1 + 4 \\
&= 6
\end{aligned}$$

Data Science Part Time

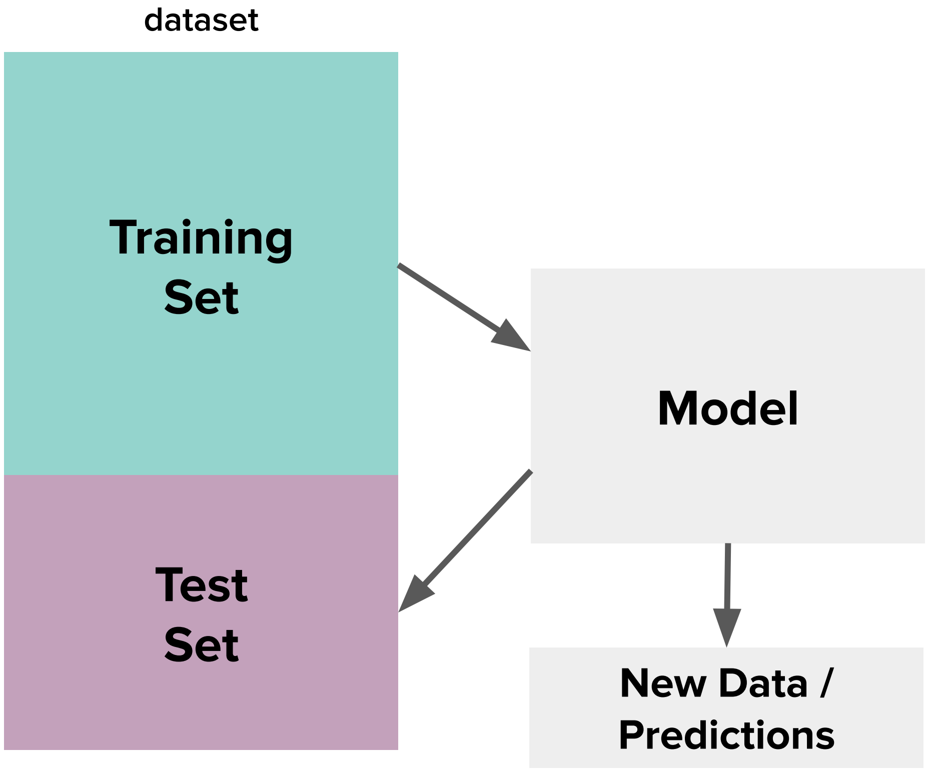
How good is your model?

Training vs testing accuracy

We will typically calculate the accuracy (or **error**) of a model using the sum of squared residuals, or the root mean squared error.

The **training** error doesn't tell us about how well our model will predict unseen data, only how well it fits our training data.

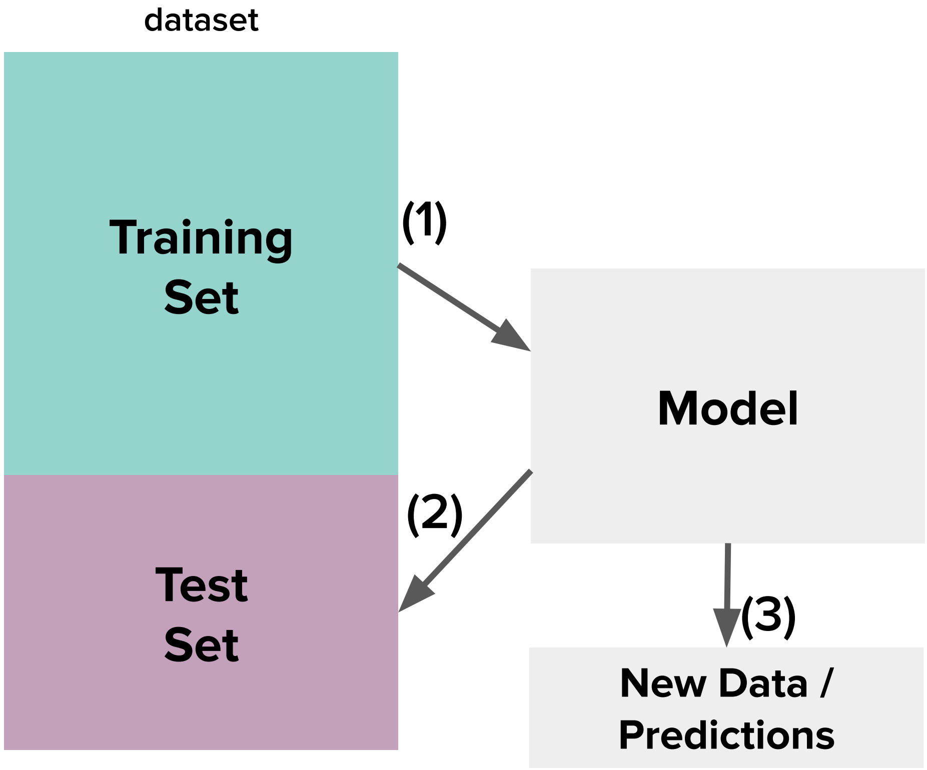
The **testing** error tells us how well our model generalises.



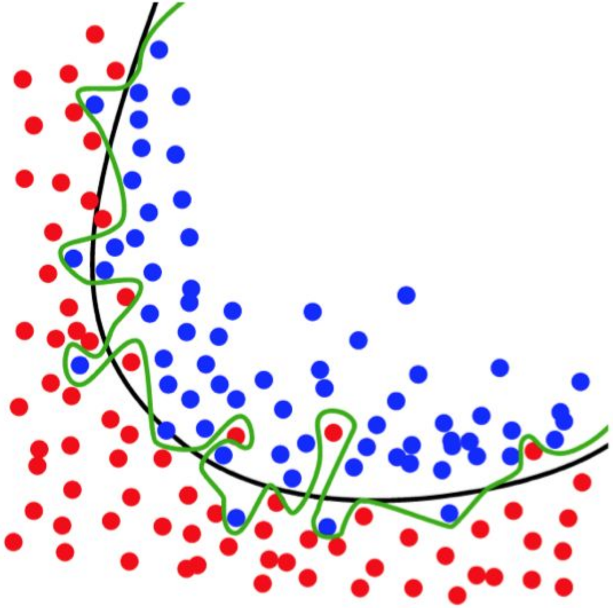
Machine Learning Validation

There are three sources of error:

1. Training error
2. Generalization error
3. Out-of-sample error



Underfitting vs Overfitting



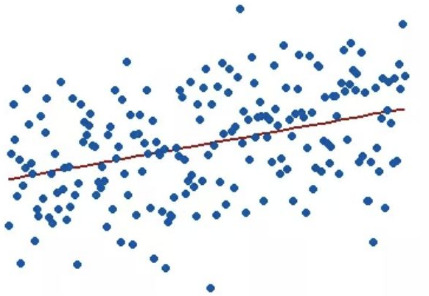
Minimizing training error does not minimize generalization error!

1. What is the training accuracy of the green line model to the left? The black line model?

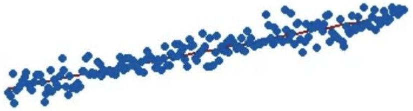
2. Is the green line a better model or the black line?

R-squared score

The R-squared score is a measure of **goodness of fit**. It evaluates the spread of the data points around the fitted regression line.



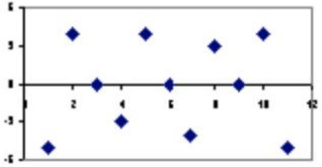
R^2 = 15%



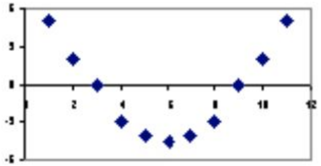
R^2 = 85%

Residual plot

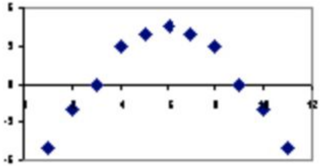
This plots the actual vs predicted values of our target variable. A residual plot that shows a largely random pattern is an indication that our linear model is a good fit for our data.



Random pattern



Non-random: U-shaped



Non-random: Inverted U

Data Science Part Time

Simple linear regression

Solo Exercise:

Simple linear regression

We’ve created some dummy data giving the % youth population and party vote share for different constituencies.

What’s the general form of the equation for the model we’re trying to fit?

Once we’ve found the coefficients, what’s the final model?

Which datasets do we use to compute:
(a) Training error
(b) Generalisation error
(c) Null/baseline model error

Data Science Part Time

Linear regression pitfalls

Multicollinearity

This occurs when the **features** in our dataset are highly correlated with each other.

A model with a high level of multicollinearity means we can't gauge how well **individual** features in our model predict the target variable.

We can spot multicollinearity by:

- Checking the correlation heatmap
- Looking for large changes in our model coefficients when our training data changes by a small amount

Intro to Python

Let’s Review

At the end of the session, you will be able to ...

Understand how linear regression models are fitted to data

Appreciate pitfalls to bear in mind when interpreting the coefficients of linear regression

Assess the goodness of fit of a model

Visualise a linear model in simple cases

Coming up next week...

- KNN classifiers
- Exploring bias and variance