CORE Skills

Delivering Data Science
In Resources & Energy

Data Analysis II: Simple Predictions - Regression and statistical model building

DAY 5

15-Day Data Science Springboard

Dr Jeremy Mitchell
Data Mettle

Dr Ying Yap
Data Mettle

Program partners

_dm

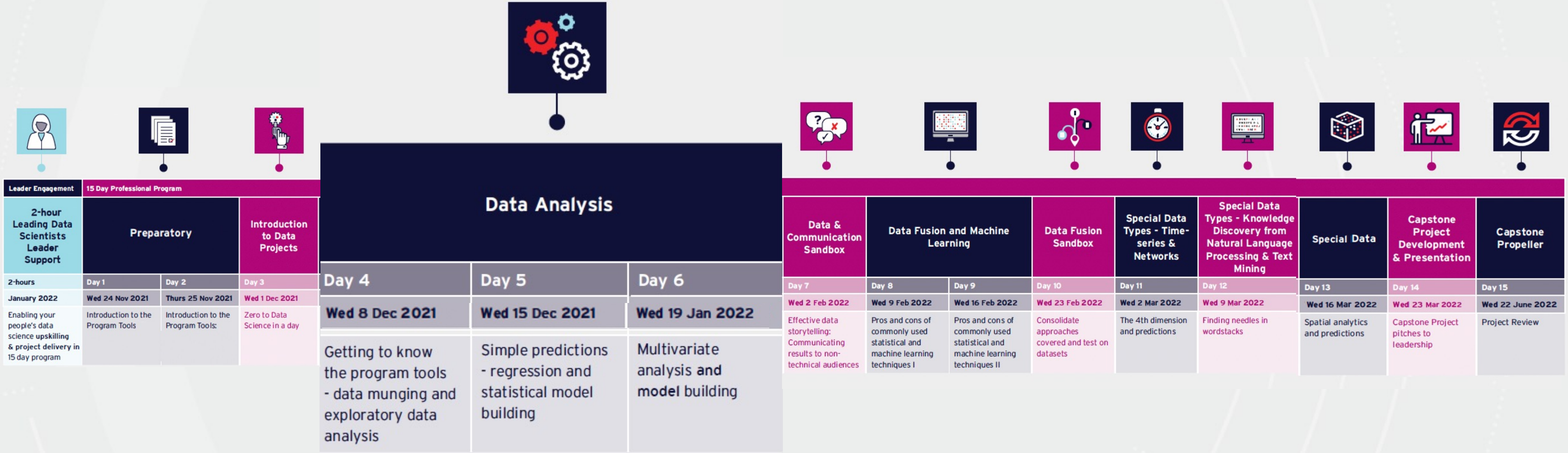
THE UNIVERSITY OF WESTERN AUSTRALIA

Curtin University

A CORE partnership

Program Timeline

DAY 4, 5 & 6: Data Analysis



Leader Engagement	15 Day Professional Program														
2-hour Leading Data Scientists Leader Support	Preparatory		Introduction to Data Projects	Data Analysis			Data & Communication Sandbox	Data Fusion and Machine Learning		Data Fusion Sandbox	Special Data Types - Time-series & Networks	Special Data Types - Knowledge Discovery from Natural Language Processing & Text Mining	Special Data	Capstone Project Development & Presentation	Capstone Propeller
2-hours	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14	Day 15
January 2022	Wed 24 Nov 2021	Thurs 25 Nov 2021	Wed 1 Dec 2021	Wed 8 Dec 2021	Wed 15 Dec 2021	Wed 19 Jan 2022	Wed 2 Feb 2022	Wed 9 Feb 2022	Wed 16 Feb 2022	Wed 23 Feb 2022	Wed 2 Mar 2022	Wed 9 Mar 2022	Wed 16 Mar 2022	Wed 23 Mar 2022	Wed 22 June 2022
Enabling your people's data science upskilling & project delivery in 15 day program	Introduction to the Program Tools	Introduction to the Program Tools:	Zero to Data Science in a day	Getting to know the program tools - data munging and exploratory data analysis	Simple predictions - regression and statistical model building	Multivariate analysis and model building	Effective data storytelling: Communicating results to non-technical audiences	Pros and cons of commonly used statistical and machine learning techniques I	Pros and cons of commonly used statistical and machine learning techniques II	Consolidate approaches covered and test on datasets	The 4th dimension and predictions	Finding needles in wordstacks	Spatial analytics and predictions	Capstone Project pitches to leadership	Project Review

Q&A, Issues & Announcements

Before we Get Started

- Resources & Tasks on Github.
- We'll start talking about **projects** a bit this afternoon, and help you start setting up yours
- Make notes about any ideas, perspectives or issues you encounter throughout the day.
- If you have aspects you'd like to go over throughout the day, feel free to post them to the general channel and we'll try to address the straightforward ones as we break.
- We'll come together to discuss before we close out this afternoon.

Schedule

DAY 5

AWST	AEST	Agenda	Educator
07:30	**09:30**	**Q&A, Issues & Announcements**	
07:45	09:45	Models & Regression	Jeremy
09:15	11:15	*Morning Tea*	
09:30	11:30	Linear Models I	Jeremy
11:00	13:00	*Lunch*	
11:45	13:45	Linear Models II	Jeremy
13:15	15:15	*Afternoon Tea*	
13:30	15:30	Robust Regression	Jeremy
13:50	15:50	Apply to your own Problem	
14:45	16:45	Closeout – Reflections, Takeaways & Project Update	Jeremy
14:55	16:55	Menti	Tamryn
17:00	17:00	**Close**	

Content developed between CORE Skills, Dr Morgan Williams, Dr Cericia Martinez & Dr Guillaume Rongier | CSIRO

Aims & Learning Outcomes

DAY 5

CORE Skills

Aims

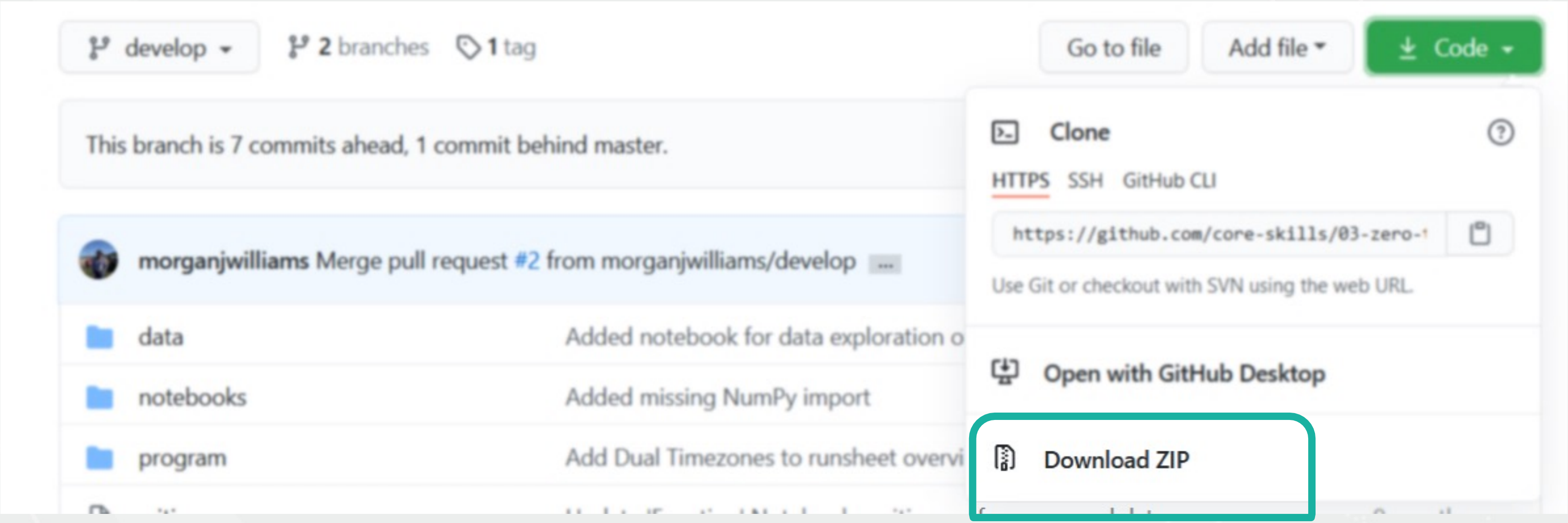
- Perform statistical model building.
- Conduct regression.
- Generate simple predictions - regression.

Learning Outcomes

- Understand regression as the basis for prediction.
- Understand how outliers and noisy data affect results.
- Understand the impact of missing data and recall practical solutions to work with incomplete data sets.
- Understand how to choose between basic statistical models and evaluate their effectiveness (e.g. linear vs polynomial).
- Have an understanding of hierarchical models as a means of modelling connections between datasets or processes.

GitHub Content for Today

[github.com / core-skills / 05-simple-predictions](https://github.com/core-skills/05-simple-predictions)



This branch is 7 commits ahead, 1 commit behind master.

morganjwilliams Merge pull request #2 from morganjwilliams/develop ...

data	Added notebook for data exploration o
notebooks	Added missing NumPy import
program	Add Dual Timezones to runsheet overvi

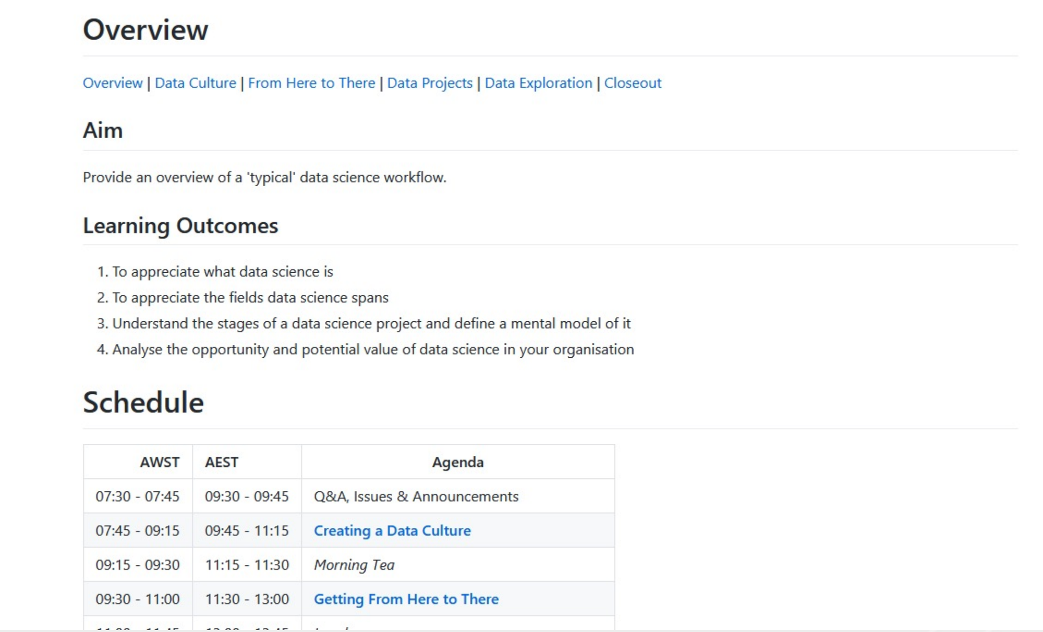
```
https://github.com/core-skills/03-zero-
```

Use Git or checkout with SVN using the web URL.

Content developed between CORE Skills, Dr Morgan Williams, Dr Cericia Martinez & Dr Guillaume Rongier | CSIRO

GitHub – Program Notes

github.com / core-skills / 05-simple-predictions / program / 00_overview.md



Overview

Overview | Data Culture | From Here to There | Data Projects | Data Exploration | Closeout

Aim

Provide an overview of a 'typical' data science workflow.

Learning Outcomes

1. To appreciate what data science is
2. To appreciate the fields data science spans
3. Understand the stages of a data science project and define a mental model of it
4. Analyse the opportunity and potential value of data science in your organisation

Schedule

AWST	AEST	Agenda
07:30 - 07:45	09:30 - 09:45	Q&A, Issues & Announcements
07:45 - 09:15	09:45 - 11:15	Creating a Data Culture
09:15 - 09:30	11:15 - 11:30	*Morning Tea*
09:30 - 11:00	11:30 - 13:00	Getting From Here to There

Environment

- Open an Anaconda Prompt
- Navigate to where you have the unzipped repository material

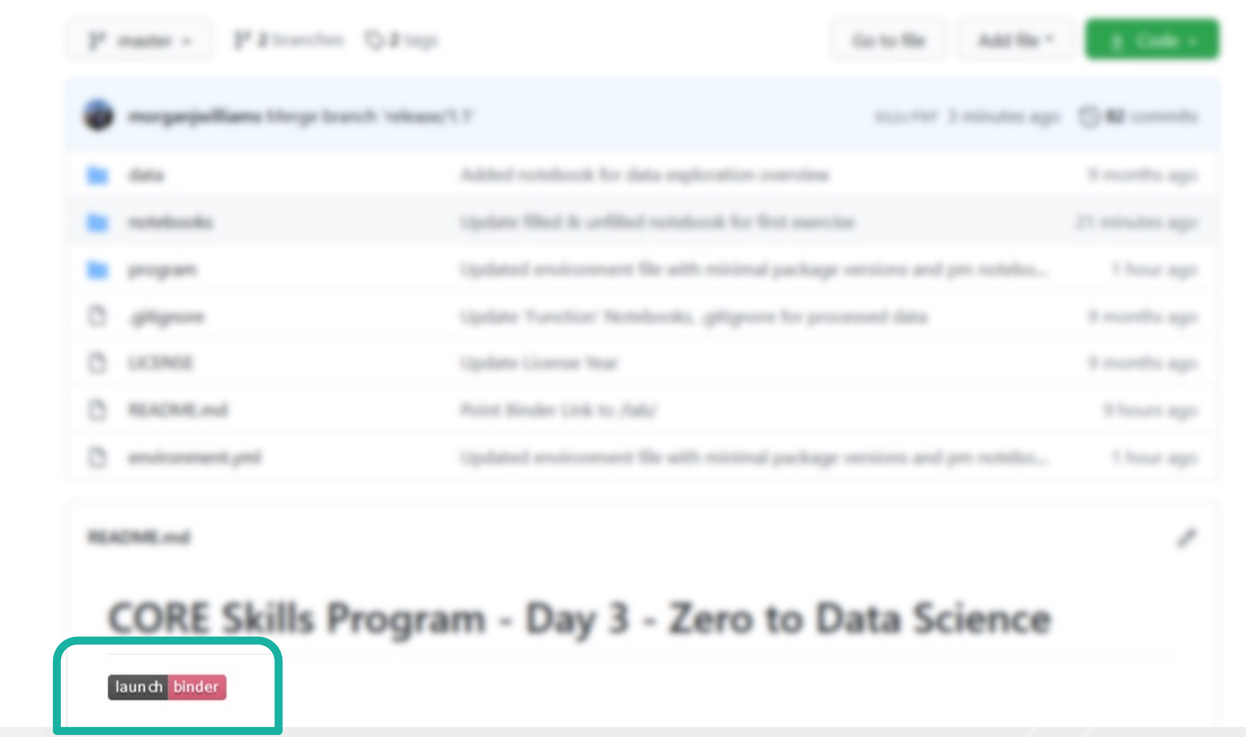
```
conda env create –f environment.yml # make new env

conda activate core05 # activate this env (Windows)

# make this available to Jupyter as a “kernel”
python -m ipykernel install --user --name=core05

jupyter lab # launch Jupyter lab
```

Binder Backup



morganjwilliams Merge branch 'release/1.1' [illegible] 3 minutes ago 82 commits

data	Added notebook for data exploration overview	9 months ago
notebooks	Update filled & unfilled notebook for first exercise	21 minutes ago
program	Updated environment file with minimal package versions and pm notebo...	1 hour ago
.gitignore	Update 'Function' Notebooks, .gitignore for processed data	9 months ago
LICENSE	Update License Year	9 months ago
README.md	Point Binder Link to /lab/	9 hours ago
environment.yml	Updated environment file with minimal package versions and pm notebo...	1 hour ago

README.md

CORE Skills Program - Day 3 - Zero to Data Science

Content developed between CORE Skills, Dr Morgan Williams, Dr Cericia Martinez & Dr Guillaume Rongier | CSIRO

Schedule

DAY 5

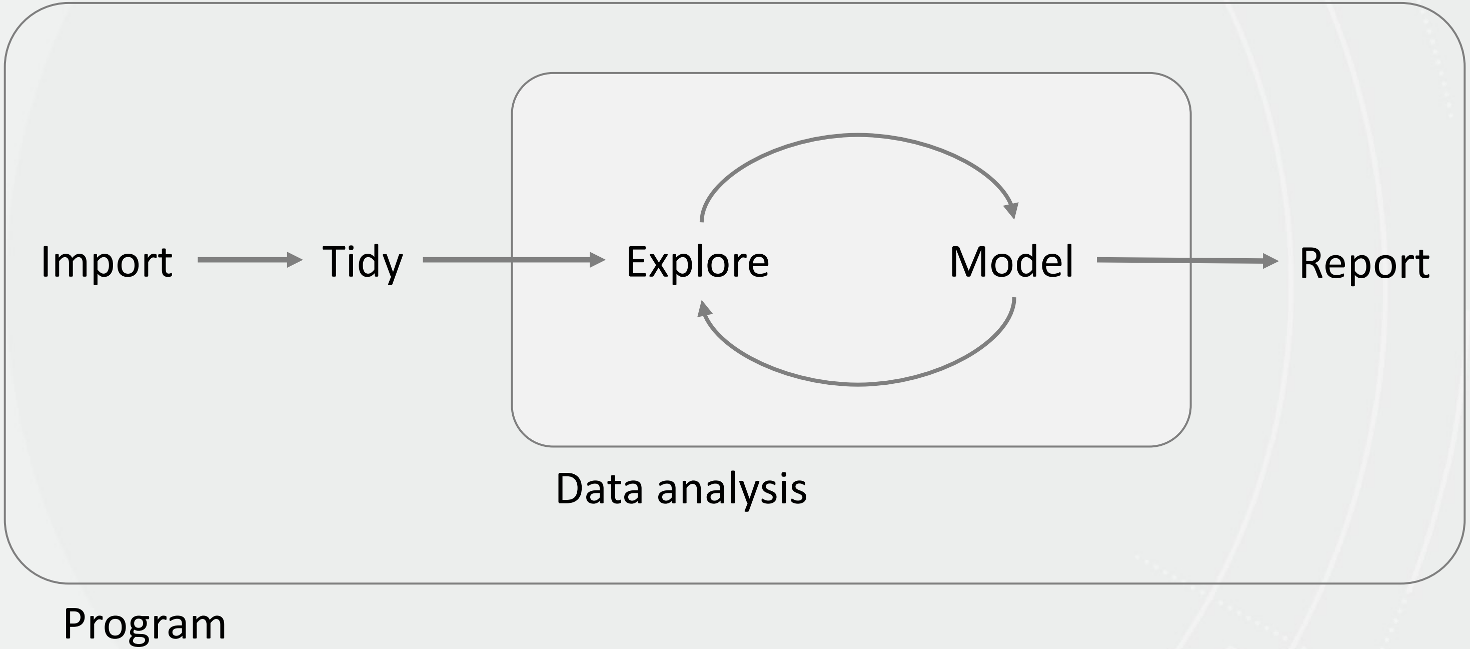
AWST	AEST	Agenda	Educator
07:30	**09:30**	**Q&A, Issues & Announcements**	
07:45	09:45	Models & Regression	Jeremy
09:15	11:15	*Morning Tea*	
09:30	11:30	Linear Models I	Jeremy
11:00	13:00	*Lunch*	
11:45	13:45	Linear Models II	Jeremy
13:15	15:15	*Afternoon Tea*	
13:30	15:30	Robust Regression	Jeremy
13:50	15:50	Apply to your own Problem	
14:45	16:45	Closeout – Reflections, Takeaways & Project Update	Jeremy
14:55	16:55	Menti	Tamryn
17:00	17:00	**Close**	

Content developed between CORE Skills, Dr Morgan Williams, Dr Cericia Martinez & Dr Guillaume Rongier | CSIRO

Models & Regression

What are the main steps of the data analysis workflow?

5 min



Import

Tidy

Explore

Model

Report

Data analysis

Program

Exercise: Explore a dataset using pandas and seaborn

Open **am1-models-and-regression.ipynb**

and go through exercise 1

Content developed between CORE Skills, Dr Morgan Williams, Dr Cericia Martinez & Dr Guillaume Rongier | CSIRO

Categories of machine learning

scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Scikit-learn basics and API

- Dataset in tidy format
- Basic API:
  - model = LinearRegression()
  - model.fit(x_train, y_train)
  - y_pred = model.predict(x_pred)

Linear Regression Model

- The linear regression model is one of the most basic statistical models used in predictive analysis
- The model proceeds by fitting a linear equation to observed paired data to attempt to model the relationship between the two variables
- One variable is commonly referred as to the explanatory variable while the other is considered the dependent variable

Linear Regression Model

- The equation of a line is

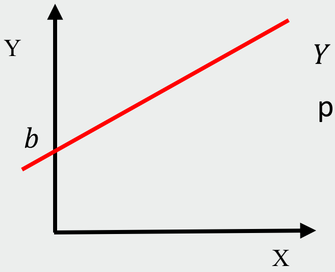
$$Y = aX + b$$

Dependent variable

Explanatory variable

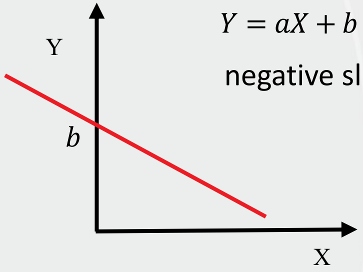
- a corresponds to the slope of the line and b to the intercept
- Parameters a and b need to be determined
- Different mathematical approaches can be used to determine the slope and the intercept. This leads to different types of regression

Linear Regression Model



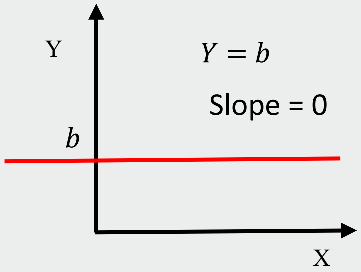
$Y = aX + b$

positive slope



$Y = aX + b$

negative slope

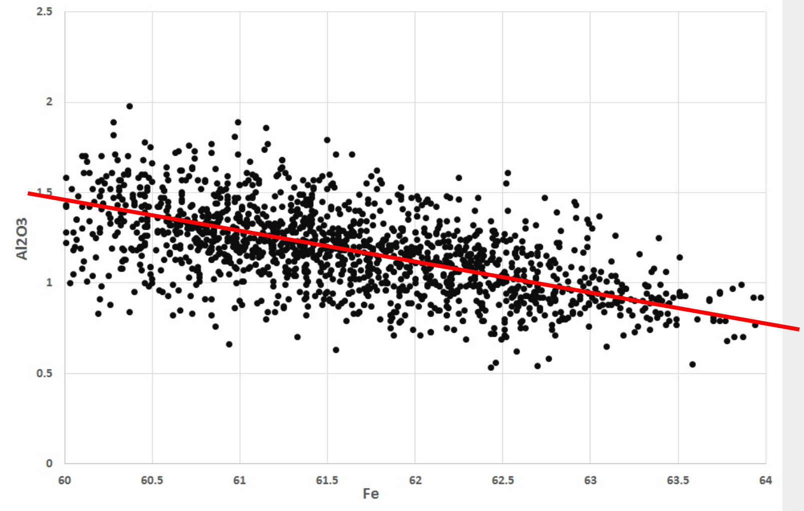


$Y = b$

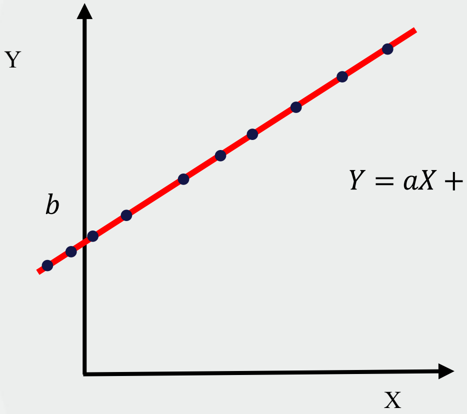
Slope = 0

Linear Regression Model

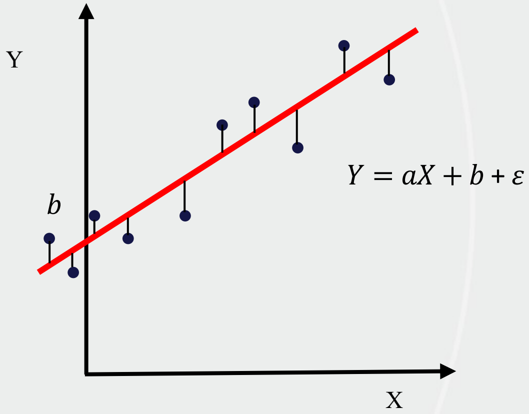
- Why do we want to fit a line?
- The scatterplot shows the paired data is scattered around a trend line
- The line is indicative of the average value of the dependent variable given the explanatory variable



Linear Regression Model



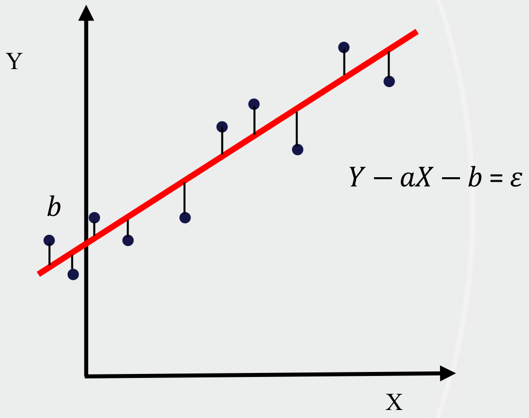
Values of X and Y are perfectly related



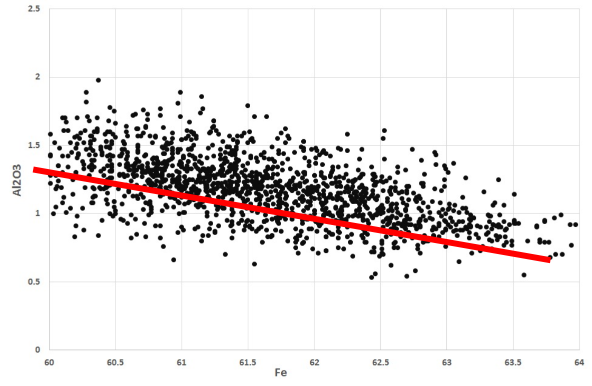
Values of X are known and it is assumed that $Y = aX + b$ plus a random term ε

Statistical assumptions

- A number of assumptions are made on the random error term ε
- First assumption is that on average the error is equal to zero
- This ensures the model has no bias

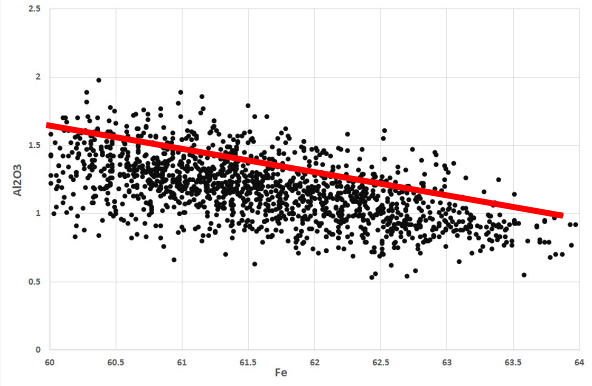


Statistical assumptions



Positive bias

Negative bias

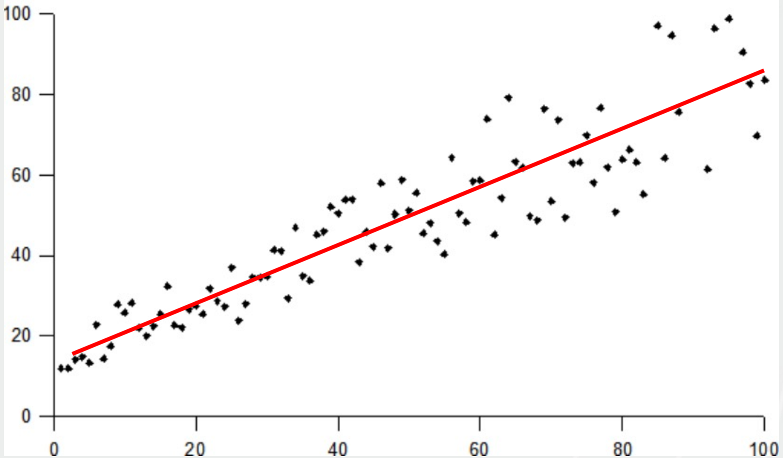


Statistical assumptions

- Second assumption is that the variance of the random error term is constant

- This is know in statistics as homoscedasticity asummption

- Heteroscedasticity refers to the violation of this assumption

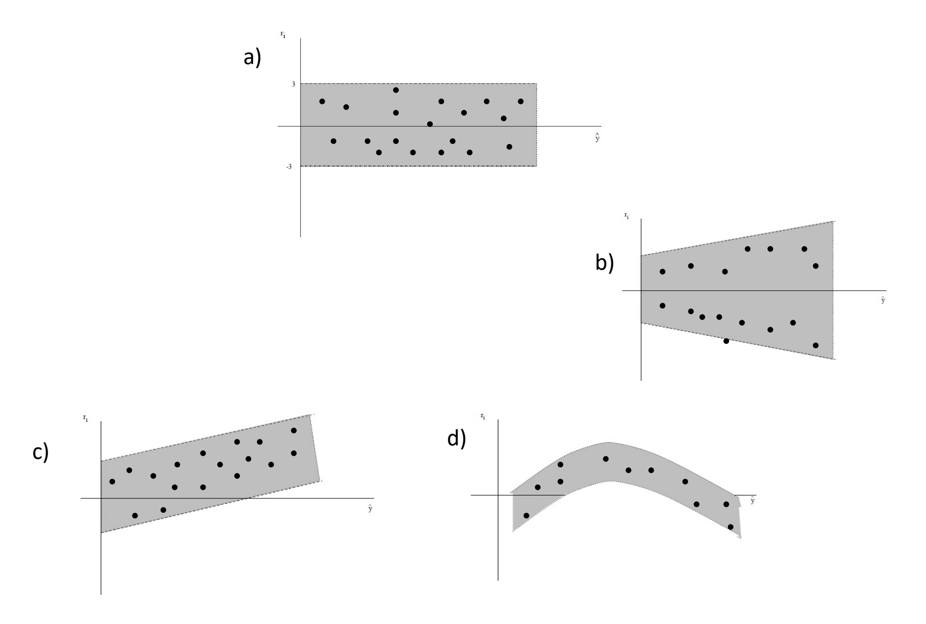


Statistical assumptions

- Third assumption is that random error term is independent of both the explanatory and dependent variable
- Extremely important assumption and provides a way to assess the goodness of the linear model
- Scatterplot of the error with either the explanatory or the dependent variable should not show any clear pattern

Statistical assumptions

- a) Appropriate
- b) homoscedasticity violated
- c) and d) are indicative that the linear regression model is not adequate



Statistical assumptions

- Fourth assumption corresponds to the statistical distribution of the error component
- It is assumed that the distribution of the random error is Gaussian with mean equal to zero and variance σ^2

$$\varepsilon \sim N(0, \sigma^2)$$

- The Gaussian assumption allows to derive all the theoretical properties of the linear regression model

Statistical assumptions

Some of the theoretical properties are:

- The dependent variable has Gaussian distribution

  This allows to construct confidence intervals for the estimated values

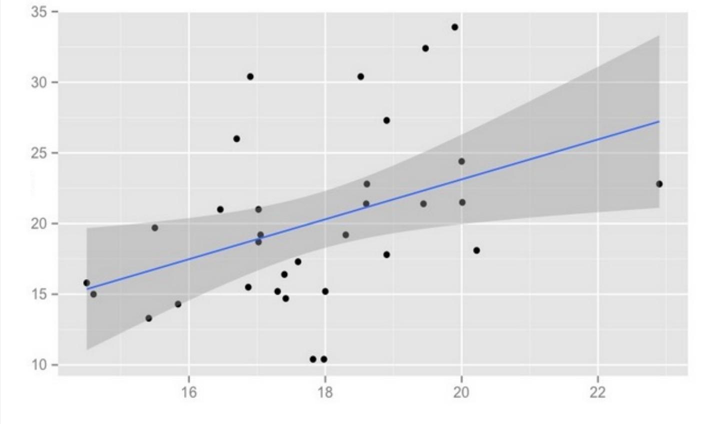
- The slope a has Gaussian distribution

  This allows to use statistical tests to assess the “importance” of the explanatory variable

Statistical tests on the parameters are more important in multivariate problems

Statistical assumptions

- Example of a linear regression model with confidence intervals
- Can you anticipate the behaviour of the residuals?
- Would you say the linear model is reasonable?



Exercise: Linear regression with scikit-learn

Return to **am1-models-and-regression.ipynb**

and go through exercise 2

Content developed between CORE Skills, Dr Morgan Williams, Dr Cericia Martinez & Dr Guillaume Rongier | CSIRO

Exercise: Perturbing perfect linear data

Return to **am1-models-and-regression.ipynb**

and go through exercise 3

A look at R^2

- $\boldsymbol{R^2}$ is also known as coefficient of determination

- $\boldsymbol{R^2}$ is used as a goodness of fit measure for linear regression models, i.e. to determine how well the regression model fits the data

$$\boldsymbol{R^2} = 1 - \frac{\sum \varepsilon_i^2}{\sum(y_i - \bar{y})^2}$$

- Proportion of the variance in the dependent variable that the explanatory variable explains

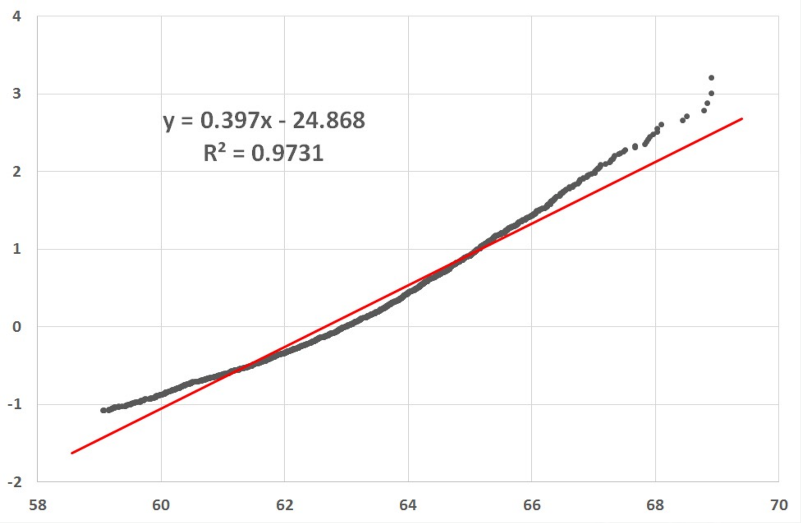
A look at R^2

- R^2 has limitations and therefore should not be the only criteria used to assess the goodness of the linear regression model

- Some of its limitations are:
  - Does not account for the number of paired data used
  - Does not indicate if the explanatory variable used is appropriate
  - Does not indicate if the regression used is appropriate
  - Does not indicate if the model is biased

A look at R^2

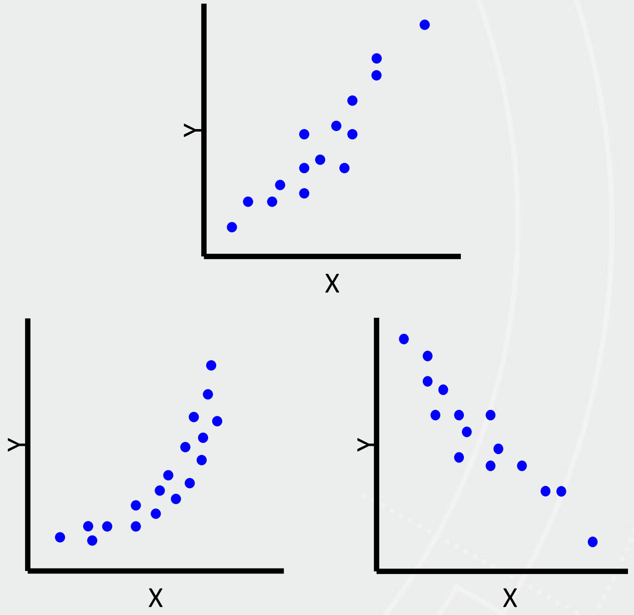
A biased model can have a high R^2 value!



Correlation and Independence

- Traditional correlation is a measure of the linear relationship between two variables

- A correlation value equal to zero does not mean the variables are not related



Correlation and Independence

- The concept of statistical independence is given by the factorisation of the joint probability distribution as the product of the marginal distributions

- This means that knowing the value of one of the variables does not tell anything about the value of the other variable

- Independence implies no correlation but the reverse is not true

Schedule

DAY 5

AWST	AEST	Agenda	Educator
07:30	**09:30**	**Q&A, Issues & Announcements**	
07:45	09:45	**Models & Regression**	Jeremy
09:15	11:15	*Morning Tea*	
09:30	11:30	**Linear Models I**	Jeremy
11:00	13:00	*Lunch*	
11:45	13:45	**Linear Models II**	Jeremy
13:15	15:15	*Afternoon Tea*	
13:30	15:30	**Robust Regression**	Jeremy
13:50	15:50	Apply to your own Problem	
14:45	16:45	**Closeout** – Reflections, Takeaways & Project Update	Jeremy
14:55	16:55	**Menti**	Tamryn
17:00	17:00	**Close**	

Content developed between CORE Skills, Dr Morgan Williams, Dr Cericia Martinez & Dr Guillaume Rongier | CSIRO

Apply to your own Problem

If you don't have a dataset

- UCI Machine Learning Repository:

  archive.ics.uci.edu/ml/index.php

- Kaggle datasets:

  www.kaggle.com/datasets

Content developed between CORE Skills, Dr Morgan Williams, Dr Cericia Martinez & Dr Guillaume Rongier | CSIRO

Schedule

DAY 5

AWST	AEST	Agenda	Educator
07:30	**09:30**	**Q&A, Issues & Announcements**	
07:45	09:45	**Models & Regression**	Jeremy
09:15	11:15	*Morning Tea*	
09:30	11:30	**Linear Models I**	Jeremy
11:00	13:00	*Lunch*	
11:45	13:45	**Linear Models II**	Jeremy
13:15	15:15	*Afternoon Tea*	
13:30	15:30	**Robust Regression**	Jeremy
13:50	15:50	Apply to your own Problem	
14:45	16:45	**Closeout** – Reflections, Takeaways & Project Update	Jeremy
14:55	16:55	**Menti**	Tamryn
17:00	17:00	**Close**	

Content developed between CORE Skills, Dr Morgan Williams, Dr Cericia Martinez & Dr Guillaume Rongier | CSIRO

Takeaways & Closeout

High-level Takeaways From Today

- Understand regression as the basis for prediction.
- Understand how outliers and noisy data affect results.
- Understand the impact of missing data and recall practical solutions to work with incomplete data sets.
- Understand how to choose between basic statistical models and evaluate their effectiveness (e.g. linear vs polynomial).
- Have an understanding of hierarchical models as a means of modelling connections between datasets or processes.

Content developed between CORE Skills, Dr Morgan Williams, Dr Cericia Martinez & Dr Guillaume Rongier | CSIRO

High-level Takeaways From Today

Your thoughts?

Content developed between CORE Skills, Dr Morgan Williams, Dr Cericia Martinez & Dr Guillaume Rongier | CSIRO

Capstone Projects

Capstone Projects

Update

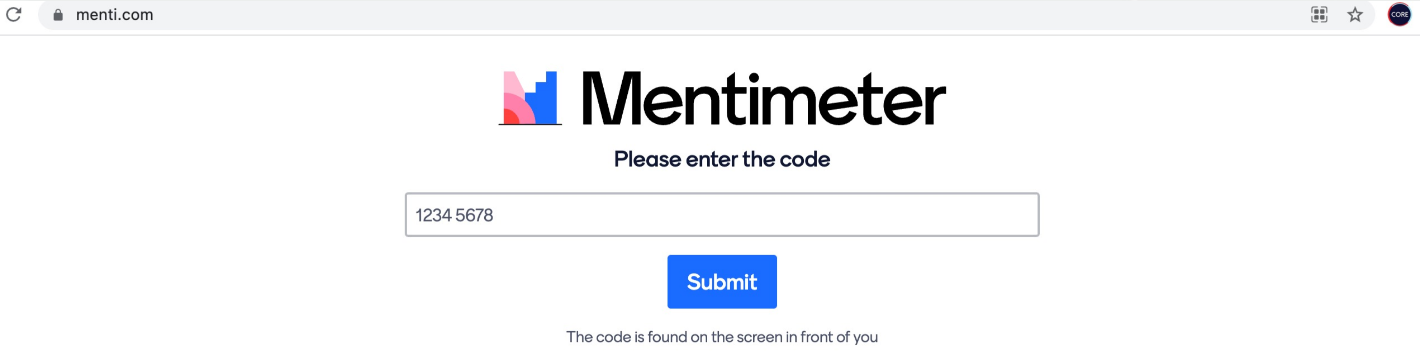
- How you are shaping up your Project

Action

- Milestone #1 → L0 in Momentum
- If not done already, update your Working Project Title & Short Project Statement (Problem/Solution/Plan) here.
- Update your Leader

Daily Feedback

Menti



Mentimeter

Please enter the code

1234 5678

The code is found on the screen in front of you

- What was one thing you like about today?
- What would you like to see more of?

CORE Skills

A CORE partnership

COREHUB.COM.AU/SKILLS