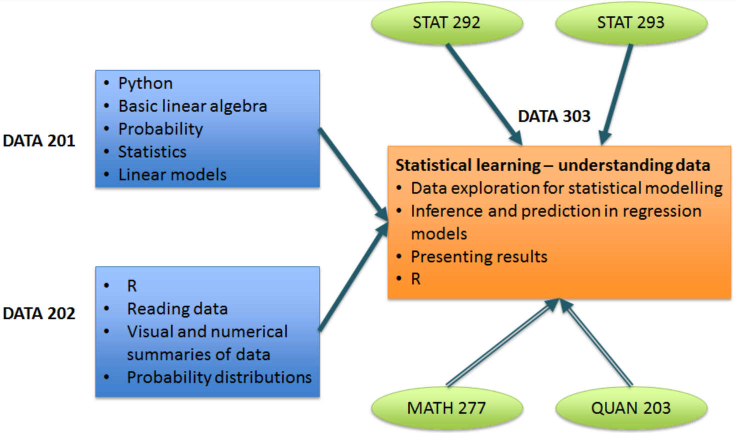
DATA 303: Statistics for Data Science

Week 1: Introduction and review of linear regression

Dr Nokuthaba Sibanda
Cotton Building 543
nokuthaba.sibanda@vuw.ac.nz

About the course

How the course connects to other courses



What is the course about?

Statistical modelling

- regression modelling framework
- inference and prediction - what is the difference?
- regression model for continuous response variables
  - estimation
  - assessing model accuracy
  - extending the linear model: interactions, polynomial regression, generalised additive models
  - model selection - subset selection, shrinkage methods
- count data regression
- regression model for binary response variables

All data manipulation and analysis will be done using R in RStudio.

Course delivery

Lectures: Mon, Wed, Fri 2:10-3pm. See https://sms.wgtn.ac.nz/Courses/DATA303_2020T1/CourseDiary for topics and timetable. Lectures recorded and videos available via Blackboard.

- Lecture Notes and slides available on course website. Annotated slides uploaded at end of week - both on **Lecture Notes** page.

Labs: Thursday 10-11:50am (MY221)

- Lab instructions available on **Labs** page
- Labs start in week1

Assessments

- 4 written assignments (20% total) - due weeks 3, 6, 9 and 12
- 2 quizzes (10% total) - weeks 4 and 10 during lectures
- 1 computer-based test (20%) - week 7 during lab time
- Final 2-hour exam (50%)

Assignments

- Submit online via DATA 303 course webpage: https://sms.wgtn.ac.nz/Courses/DATA303_2020T1/Assignments
- Must be prepared using Rmarkdown and submitted as a .Rmd file
- Make sure your .Rmd file successfully runs and produces a .pdf file before submitting it.
- Assignment 0 available for practice - use anytime to check your submission.
- Marked assignment and mark posted on course webpage.

Other info

Office Hours:

- Nokuthaba: Monday 3-4pm, Wednesday 1-2pm or email for appointment
- Yuichi and Ryan will announce theirs in due course
- or check https://sms.wgtn.ac.nz/Main/OfficeHours

Mandatory Course Requirements

- None

Course announcements

- email announcements via Blackboard
- information notices posted on course webpage

Tutor: Linda Martis

Class Rep

Video from VUWSA

https://www.vuwsa.org.nz/class-representatives/

How to sign up if elected:

https://www.youtube.com/watch?v=ofRy3oIoXD4

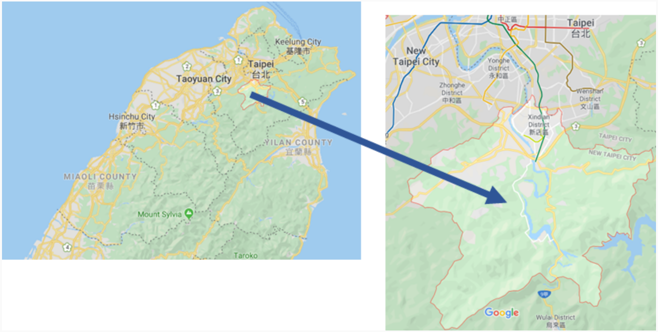
Motivating example: Predicting house prices in Taipei

Scenario

- You've been approached by a real estate agent in Taipei, Taiwan to develop an automated system they can use to estimate the selling price of a house that is to be listed for sale.
- Estimation of house price should be based on a number of characteristics or predictors.
- House price and predictor information is available in a dataset on house sales in 2012 and 2013 for 414 residential properties in Xindian District, New Taipei City, Taiwan.
- The data are from a 2018 study by Yeh and Hsu (Yeh, I. C., & Hsu, T. K. (2018). Building real estate valuation models with comparative approach through case-based reasoning. *Applied Soft Computing*, 65, 260-271.)

Data description

Location of Xindian in relation to Taipei and New Taipei City.



Data description

The data are in the file *houseprice.csv* and include the following variables:

- `year`: transaction year (2012 or 2013)
- `house.age`: age of house (years)
- `distMRT`: distance to the nearest MRT/metro station (metres)
- `stores`: number of convenience stores within walking distance
- `latitude`: latitude of house location (degrees)
- `longitude`: longitude of house location (degrees)
- `price`: house price per unit area (10000 New Taiwan Dollars/Ping, where Ping is a local unit of area, 1 Ping $\approx$ 3.3m^2)

Planning your approach

Given such a dataset:

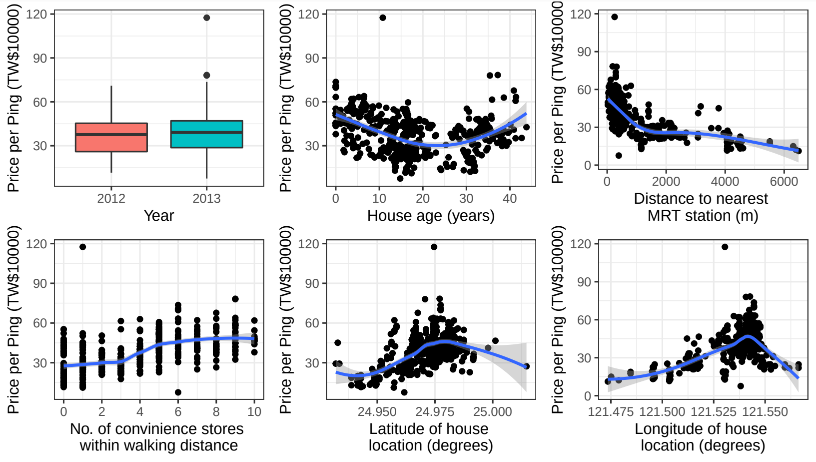
1. How would you use these data to develop the house price estimation system.

-
-
-

2. What would be the limitations of such a system?

-
-
-

The graphs below display the relationship between `price` and each of the predictors.



What do the graphs show?

- Median house prices were slightly higher in 2013 compared to 2012
-
-
-
-
-

Questions and further analyses

- Which predictors should we use to get the most accurate predictions of house price?
- Is the effect of `distMRT` on `price` the same for all values of `stores`? We may wish to consider the **interaction** between these two variables.
- Is the relationship between `price` and `distMRT` significantly non-linear?
- To get more accurate predictions, we should predict house price in a way that accounts for possible non-linear relationships between `price` and each of `house.age`, `distMRT`, `stores`, and the interaction between `distMRT` and `stores`.

Regression modelling

Regression models

Regression models are used to describe and quantify the relationship between a **response variable** and one or more **predictor variables**.

In the Taipei house price example:

- the selling price is the **response variable** - usually represented using Y
- the property characteristics are the **predictor variables** or **predictors** - usually denoted using X, with subscripts used to differentiate between predictors. For example, set X_1 to be `year`, X_2 to be `house.age`, etc.

Multiple linear regression model

Given a **quantitative** response variable Y and p different predictors, $X_1, X_2, \ldots, X_p$, we assume that there is a relationship between Y and $X = (X_1, X_2, \ldots, X_p)$ that can be written as:

$$Y = \underbrace{f(X)}_{\substack{\text{systematic} \\ \text{component}}} + \underbrace{\epsilon}_{\text{error}}.$$

For example:

Multiple linear regression model

Generally, f is unknown and we use the observed data to estimate

- its form (e.g. linear/non-linear), and
- the coefficients for the predictors.

Multiple linear regression model

The model

$$Y = f(X) + \epsilon$$

is a type of regression model, called a **multiple linear regression** model, in which

- Y is related to **multiple predictors** through f,
- Y has a **linear** or **curvi-linear** relationship with each of the predictors in the model, and
- Y is assumed to have a normal distribution with mean $E[Y|X] = f(X)$ and variance σ^2.

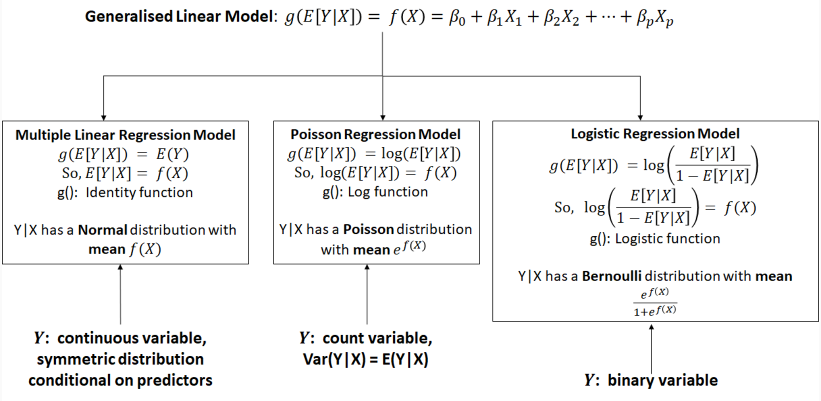
Generalised linear models

The multiple linear regression model is part of class of regression models called **generalised linear models**.

The specific type of regression model used for a particular dataset is driven largely by **the type of response variable** Y **and its distribution.**

Generalised linear models

The figure below summarises the different model types covered in DATA 303.



Generalised linear models

Note:

The link between the predictors X and the response variable Y is through the **mean** $E[Y|X]$

that is $g(E[Y|X]) = f(X)$.

So a regression model is used to determine how the **mean** of Y changes as the predictor values change.

For example, in Poisson regression we have $E[Y|X] = e^{f(X)}$ or $\log(E[Y|X]) = f(X)$.

Generalised linear models

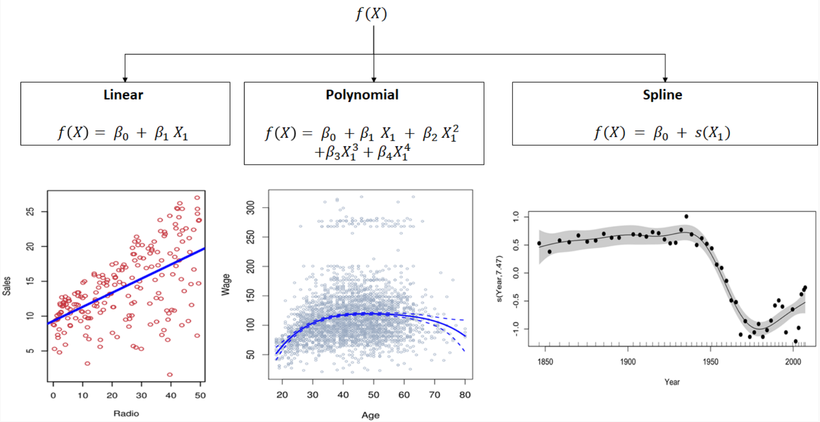
To **predict** values of Y for specific predictor values in a multiple linear regression model we use:

$$\hat{Y} = \widehat{E[Y|X]} = \widehat{f(X)} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p,$$

For accurate predictions we need accurate:

- form of $f(X)$ and
- estimates $\hat{\beta}_j$

Forms of $f(X_j)$ for a single predictor



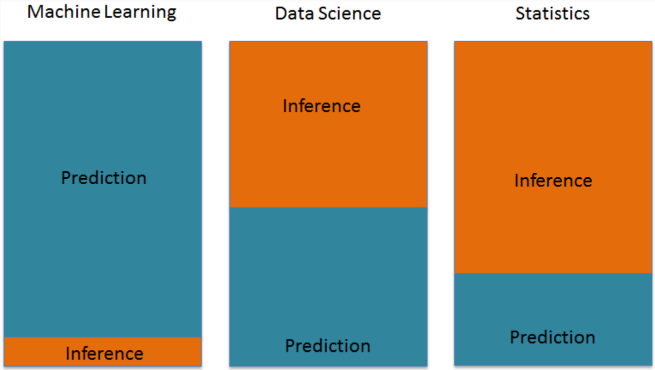
Summary of regression tasks

In summary, regression modelling is composed of the following main tasks, some of which are undertaken simultaneously:

1. Determining the most appropriate way of linking $E(Y)$ to $f(X)$. This decision is based on:
   a) the variable type for Y (eg continuous, count, categorical)
   b) the shape of the distribution (eg symmetric/non-symmetric) of Y
2. Determining the 'best' form of $f(X)$ according to criteria such as:
   a) goodness-of-fit
   b) predictive accuracy
3. Estimating the unknown parameters in the model.

Regression in the context of inference and prediction

Inference and prediction



Inference

In inference, we use the model to describe **the process that generated the data**.

The basic workflow for inference exercises is:

1. **Model hypotheses** - construct hypotheses about the potential model equations that describe the data generation process.
2. **Model fitting and validation** - Fit (estimate) the candidate models and validate them using residual analysis and goodness-of-fit tests.
3. **Conclusion** - Select and state the model that most accuractly describes the data generation process.

Model interpretability is important for inference.

Prediction

Less focus on describing the data generation process and more on finding a model that gives the most accurate predictions for Y.

Predictions of Y are calculated using

$$\hat{Y} = \widehat{E[Y|X]}.$$

When constructing a model for prediction, the most inmportant consideration is minimising prediction error $(Y - \hat{Y})$ on new data.

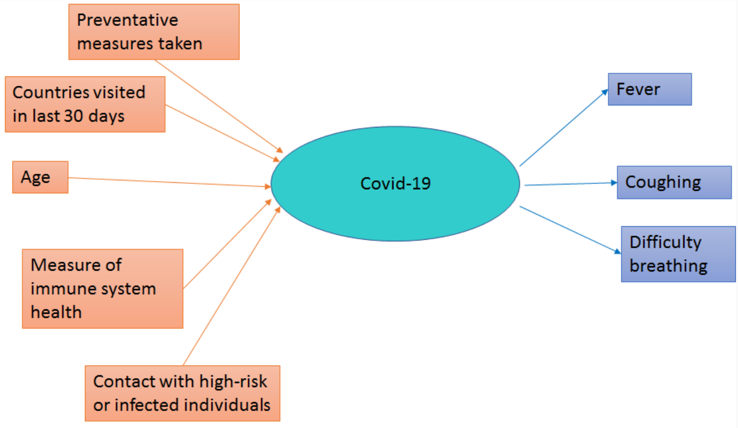
Example: Inference vs prediction

Suppose we wish to construct a model to estimate the risk that an individual has contracted the coronavirus.

Inference would focus on the (causal) factors that affect the risk of contracting the virus

Prediction could include these factors but would also include things caused by the virus eg symptoms

Example: Inference vs prediction



Multiple linear regression

The linear model

To provide a solution to the estate agency you decide to construct a regression model.

Key requirement - the model should give **accurate predictions** of house prices.

The estate agency may also wish to answer the following questions:

1. Is there a relationship between *house.age* and *price*?
2. If there is a relationship between *house.age* and *price*, how strong is it?
3. All else being equal, does the number of stores within walking distance influence house price?
4. How accurately can we predict house prices for new listings?

Therefore, an **interpretable model** that accurately estimates the data generation process is required.

The linear model

Since the response variable *price* is a continuous random variable, the first step is to use a multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon,$$

based on the underlying **assumption**

$$\epsilon \sim N(0, \sigma^2).$$

The linear model

This means that

$$Y|X_1, X_2, \ldots, X_p \sim N(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p, \sigma^2),$$

with

$$E[Y|X_1, X_2, \ldots, X_p] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

and

$$Var(Y|X_1, X_2, \ldots, X_p) = \sigma^2.$$

Model assumptions

The following four **assumptions** are made in the model:

- The error term, ϵ, is assumed to follow a normal distribution with a mean of zero.
- The error term, ϵ, has a variance σ^2 that is constant for all values of the predicor variables.
- The errors are independent of each other.
- The response variable has a linear or curvi-linear relationship with the predictor variables.

Interpreting the model assumptions

The first three assumptions imply the following about the response variable Y:

-
-

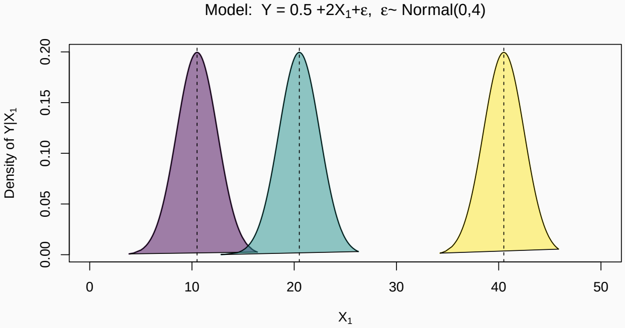
Therefore, the regression model is used to predict the **mean response** for a set of predictor variable values.

Checking if model assumptions hold

We check whether linear regression is a suitable model for a given dataset by checking whether the **errors** (estimated using residuals) do follow a normal distribution with mean zero and a constant variance (estimated by $\hat{\sigma}^2$.)

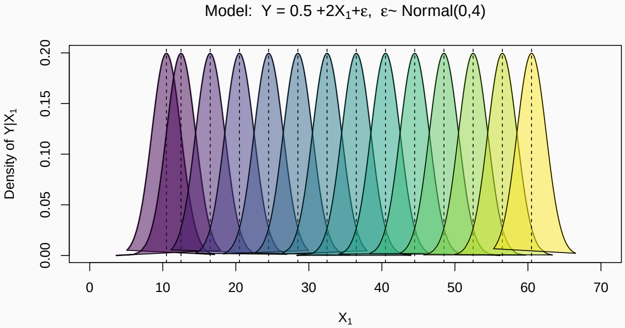
Why not check assumptions using Y?

This is more difficult since the mean of Y, $E[Y|X_1, X_2, \ldots, X_p]$, changes as the predictor values change.



Why not check assumptions using Y?

In a real dataset there are likely to be many different values of X_1, leading to:



The Y we observe is a mixture of all those Ys with different means, so we cannot expect Y to have a normal distribution, but $Y|X$ will have a normal distribution.

Coming up next week

- Estimation
- Linear regression and model assessment in R
- Interactions
- Model building guidelines