Regression Analysis 2

Quantitative Analysis

Week 09

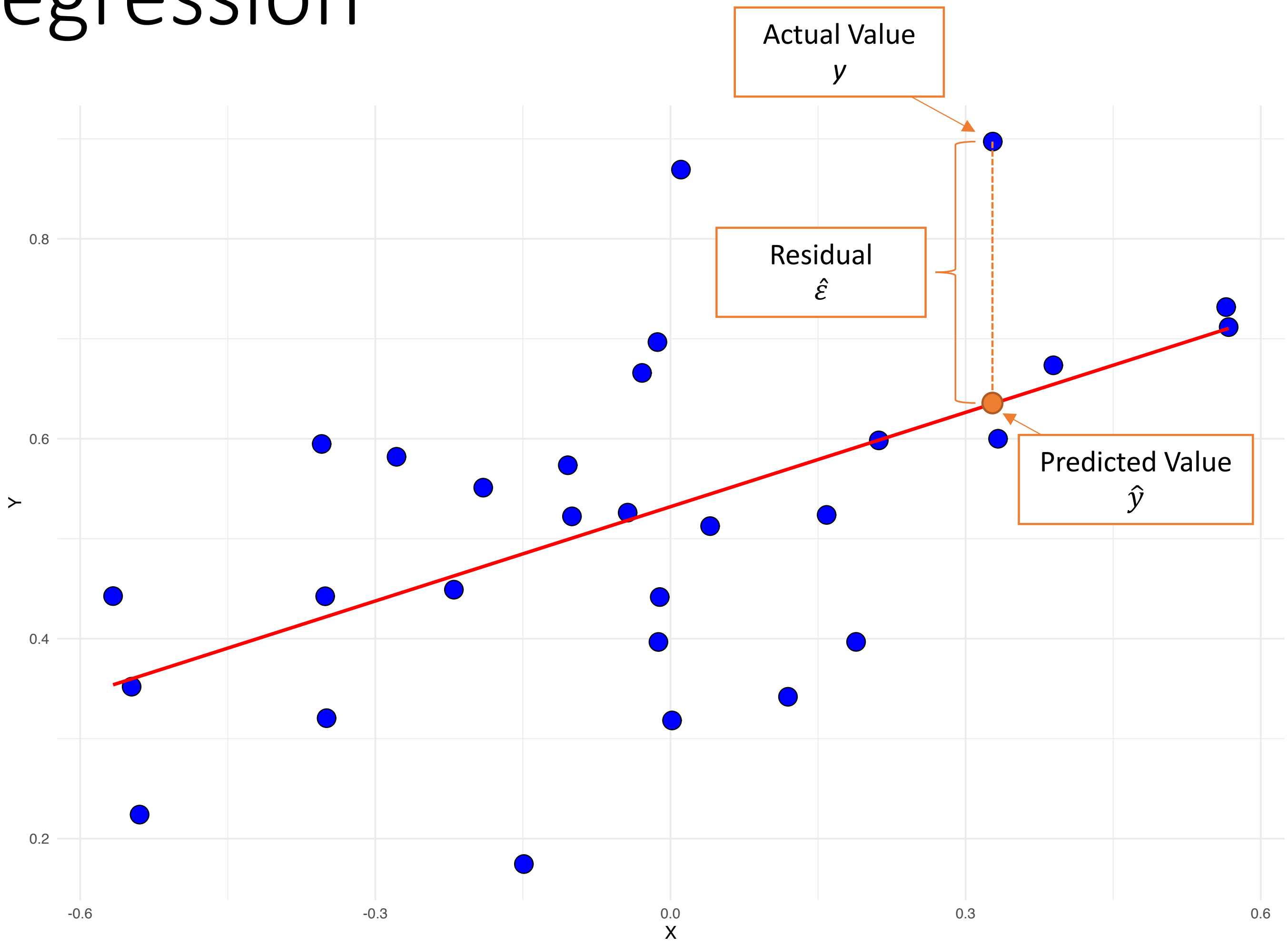
Group Project Update

- Now that teams are finalised, each group needs to decide on a **research question** which they will work on.
- When choosing a research question, try to bear in mind the following factors:
  1. Is this a question you can realistically research in a short amount of time?
  2. Is there **prior literature** about this question? (As you discuss with your teammates, you should be using Google Scholar etc. to check if papers about this topic exist.)
  3. Is there **data you can access** regarding this question? (Think of the kind of data you'd need to answer your question, and search to see if it exists. Running a small survey of your own to collect data is also an option.)
- There will be a group assignment on Moodle – you need to submit your research question, and a quick explanation of your reasons for choosing this question, and any progress you've made with data / literature (it doesn't have to be much so far). The deadline is Monday.

Review: Linear Regression

- In the last class, we introduced our first **statistical model**, namely **linear regression** – a process for finding the optimal line to describe the relationship between two variables.

- We saw that creating a linear regression line involves minimising the difference between the **actual values *y*** and the **predicted values *ŷ*.** These differences are called **residuals**.

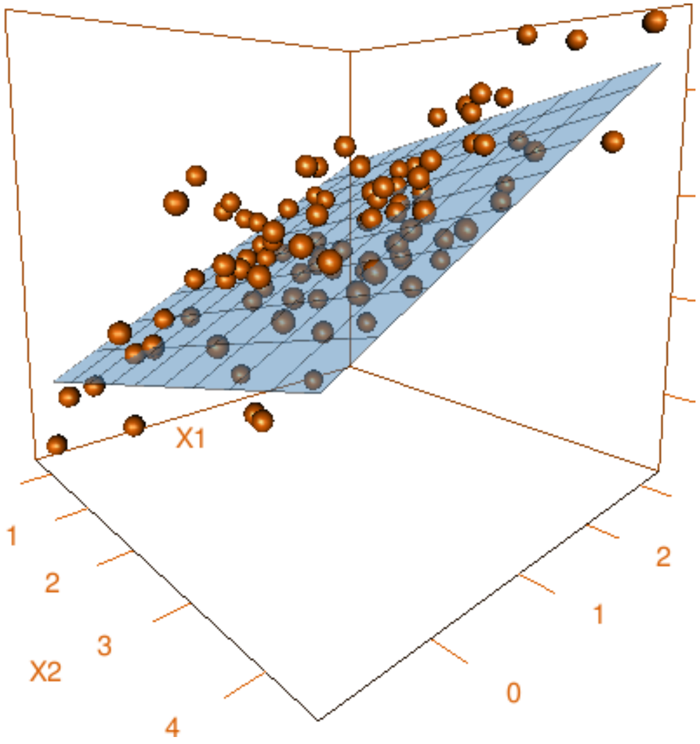


Review: Linear Regression (2)

- **Residuals** are essentially the difference in the dependent variable ***y*** which cannot be accounted for by the independent variable ***x***. In the equation for a regression model, the residuals are represented by the **error term *ε***.
- That equation is simply: $\hat{y} = \alpha + \beta x + \varepsilon$
- α is the **intercept / constant** (the value of $\hat{y}$ when $x = 0$), and β is the **slope** of the line, sometimes called the **coefficient**.
- The linear regression model also has a ***p*-value** indicating statistical significance, and an $\mathbf{R^2}$ value which shows how much of the change in ***y*** is explained by changes in ***x***.

Review: Multiple Regression

- We can also calculate a regression with **multiple independent variables** (i.e. multiple factors which we hypothesise all have an influence on the outcome variable).
- $\hat{y} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$
- This **multiple regression** involves calculating a different slope (and ***p*-value**) for each independent variable, as seen in the equation above.
- In this basic model, each slope is still a straight line, but the overall shape now has more dimensions – like the 2D plane seen in this diagram, which predicts $\boldsymbol{\hat{y}}$ based on two independent variables, $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$.



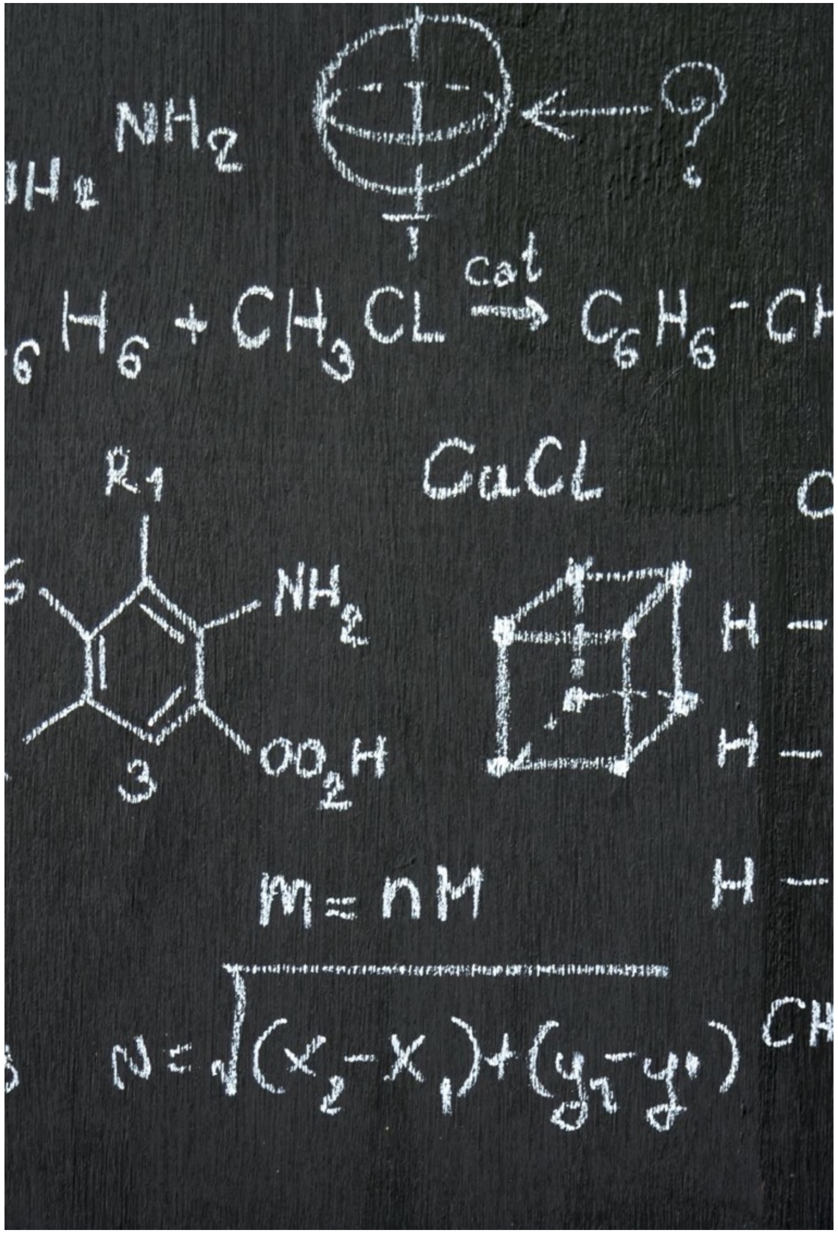
Review: R Commands

- In R, we learned how to do a regression analysis using the `lm()` command (**l**inear **m**odel).
  - We pass it a data frame containing the variables we are interested in, and tell it how they should relate to each other using a **model specification**.
  - For example, to make a regression model predicting ***y*** using ***x***, both of which are columns in the data frame ***df***:

    ```
    lm(y ~ x, data = df)
    ```

  - For a multiple regression, we can simply add extra variables to the model specification:

    ```
    lm(y ~ x1 + x2, data = df)
    ```



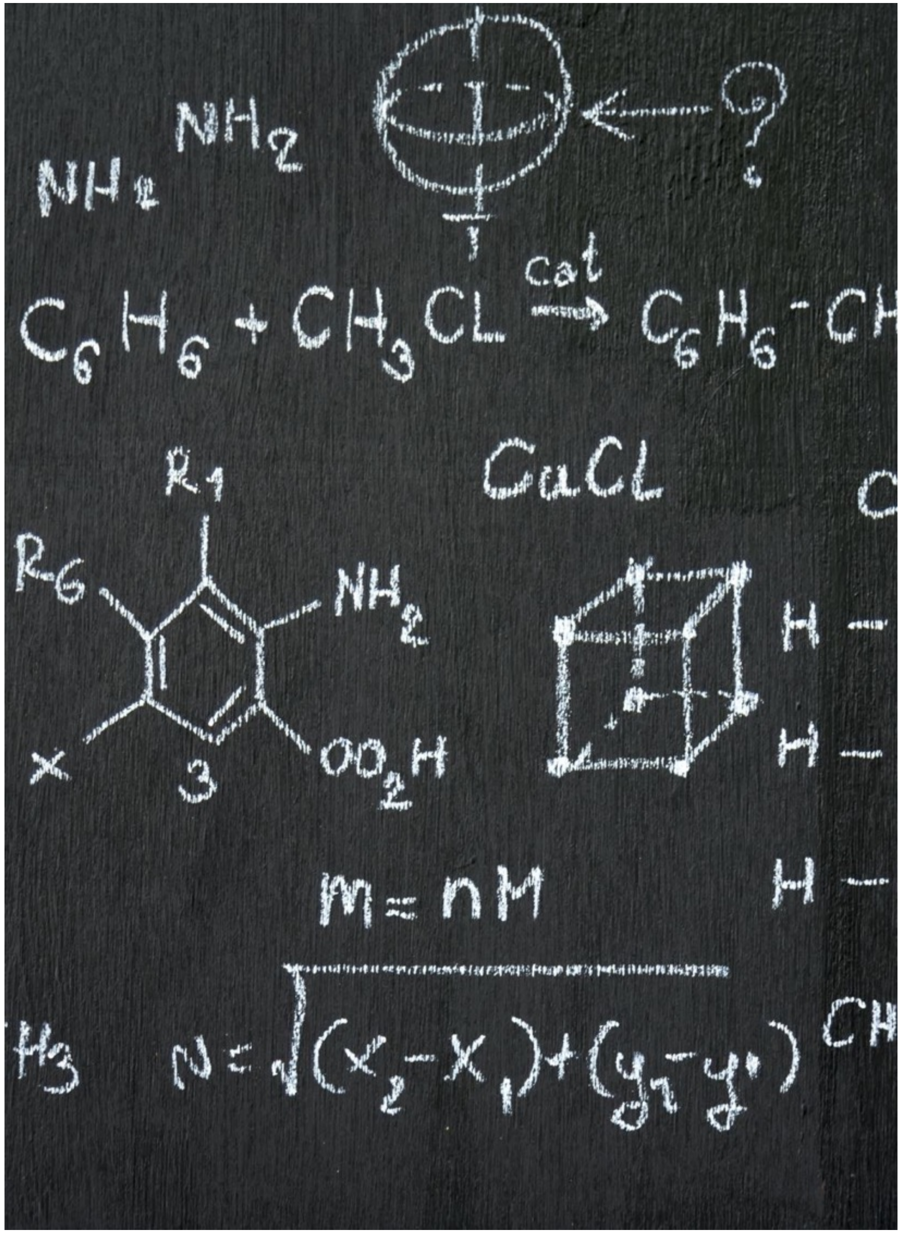
Interpreting Coefficients

- In a simple linear regression, it's easy to interpret the slope coefficient ***β*** – it is the amount by which ***y*** changes for every one-unit change in ***x***.
  - For example, if I did a regression of `salary ~ age` (with salary in Euro and age in Years) and found that the slope coefficient for age was 1200, this would mean that people earn an extra €1200 for each year older they get.

Multiple Regression Coefficients

- In a multiple regression, we have slope coefficients for each independent variable.
- We can still interpret these in the same way, but with a major caveat: the coefficient is the change in ***y*** for each unit change in ***x***, ***ceteris paribus***, meaning “all else being equal”.
- In other words, the coefficient for ***x_1*** is the change in ***y*** for each unit change in ***x_1***, **assuming that *x_2, x_3, …, x_n* all stay exactly the same**.

- Remember also that if your variables have **multicollinearity** – i.e. some independent variables are correlated with each other – the coefficients for those variables become difficult to interpret.



“All models are wrong, but some are useful”

George Box (1919-2013)

Every model is wrong…

- Every statistical model we make is an **approximation of reality** – it is not reality itself and does not capture the full complexity of reality.
- $\boldsymbol{\hat{y}}$ – the estimated values on the regression line – are simplified, incorrect versions of $\boldsymbol{y}$ – the actual values of the outcome variable.
- We know they are incorrect because they have residuals $\boldsymbol{\varepsilon}$ – these residuals, or errors, exist because our model is **wrong**, in that it is a simplified, less detailed version of reality.

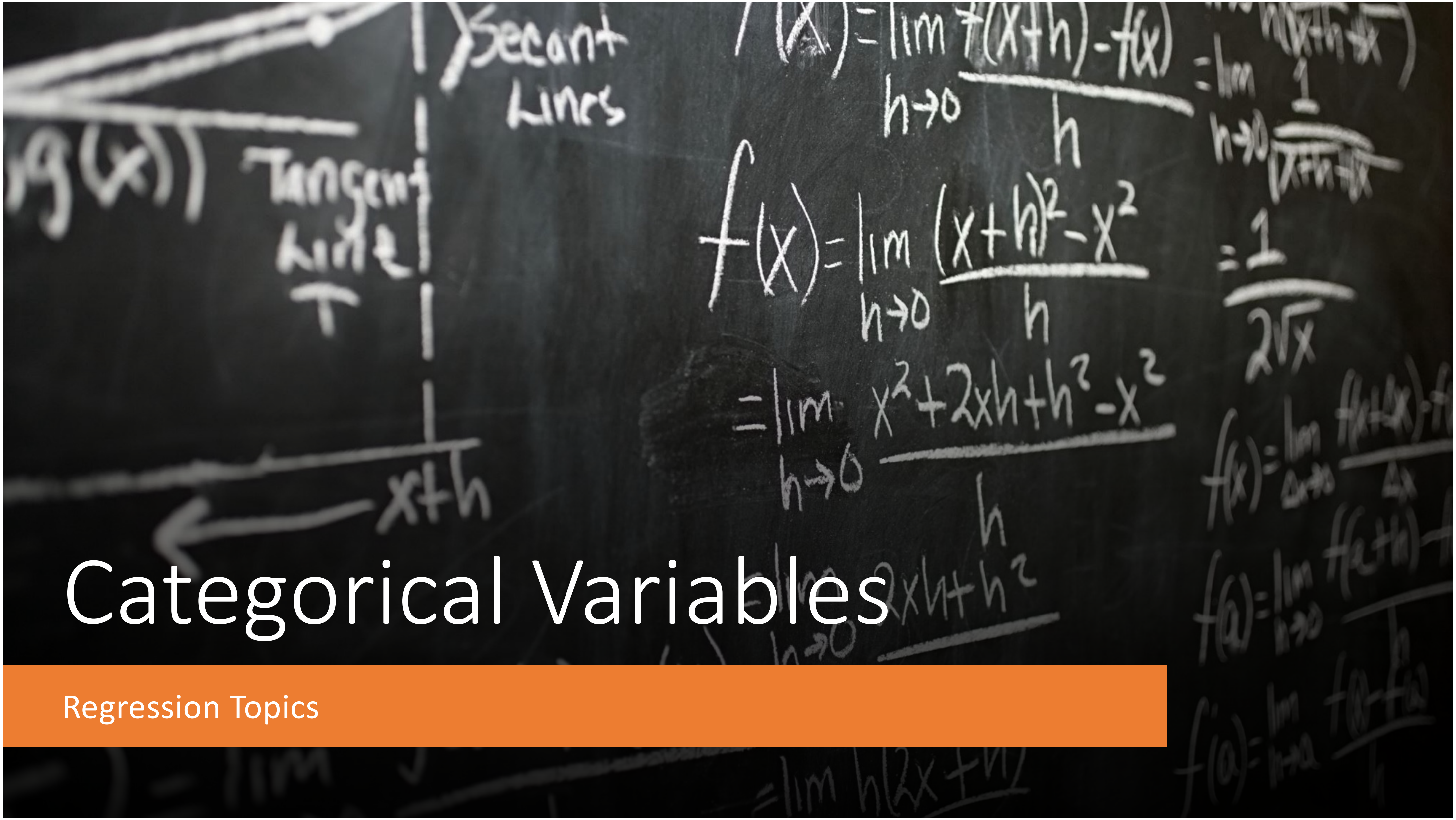
… But some are useful.

- A simplified model, even if it is wrong, can be a very effective tool for analysing and understanding aspects of human society and behaviour.
  - A map is an incredibly simplified representation of a city, but it can still help us to find our way!
- However, we must be clear about **how** the model is wrong. What was left out or smoothed over to create this model?
- A model which leaves out important factors has **omitted variable bias** which may create dangerous misunderstandings – but a model which simply removes **noise** so that we can more clearly understand the **signal** is useful.
- Studying the **residuals** – ensuring they are **i.i.d.** (independent and identically distributed), also called **spherical errors** – helps to ensure that your model isn't missing key factors.

Some more complex regression analysis problems…

How can we do regression when our variables are categorical, rather than continuous?

What about a situation where the effect of an independent variable on the outcome changes based on another independent variable?



Categorical Variables

Regression Topics

Regression with Categorical Variables

- We have previously discussed the different **types of variable** that can be used in quantitative analysis.
- So far, we've looked at regression using **continuous** variables – i.e. numeric data.
- There are two key types of **categorical** variable we might also want to use in regression:
  - **Nominal** categories have no particular rank order, such as cities (Tokyo / Beijing / Seoul etc.), genders (Male / Female etc.) or ethnicities (Asian / White / Black / Hispanic etc.).
  - **Ordinal** categories have a clear ranking order, such as education level (High School → University → Grad School) or age groups (18-24 → 25-34 → 35-44).

Nominal Categories: Binary Variables

- To use **nominal categories** (including **binary categories**, a special case with only two options, such as Yes / No) in a regression, we can represent them with **dummy variables**.
- For a **binary category variable** like gender or a yes/no answer, we just choose one value to be 0, and the other is 1.
  - For example, it's common to see analysis using a variable called "gender" where male is represented by a value of 0 and female by a value of 1.
- The regression results then indicate the effect of being in one category or the other.
  - In the above example, a positive coefficient would mean being female has a positive effect on the outcome, and vice versa.
  - Similarly, a subject being in the treatment or control group of an experiment would be represented by a 0 or 1 in a dummy variable.

Nominal Categories: Multi-Category Variables

- For a **multi-category variable** like ethnicity, location, and so on, we also must choose one category to be the default.
- We now create a separate dummy for **every other category**. The number of dummy variables is the number of categories minus one (the default category doesn't need a dummy).
  - When making dummy variables for race, we might decide that "Asian" will be our default.
  - We now create dummies like "raceWhite", "raceBlack", and "raceHispanic", which will be coded as 1 for those races, and 0 for all others.
- The results from regression with these dummy variables can be interpreted as the difference between that category and the chosen default.
  - Note that it's also a good idea to test for significant differences between your categories with ANOVA! This may also help you to decide which category to use as the default in your model.

Dummy Variable Coding

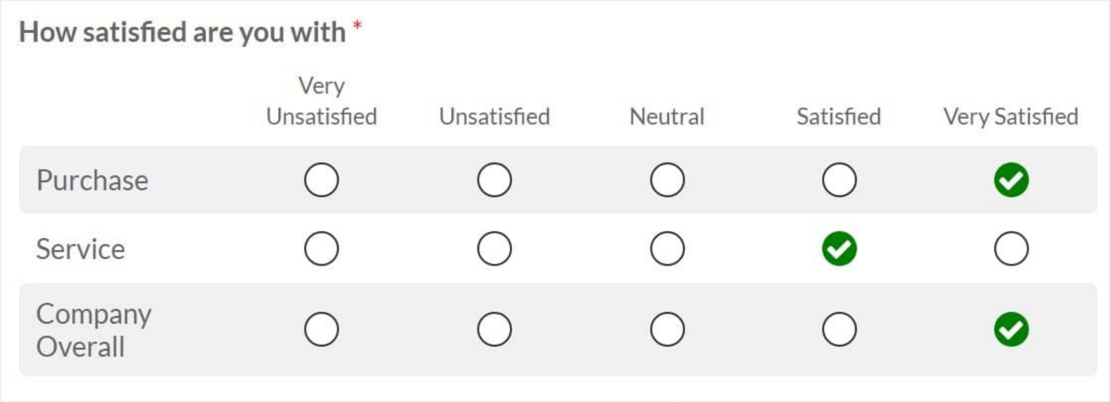
Race
Asian
Black
Asian
White
Hispanic
Black
Asian

raceBlack	raceWhite	raceHispanic
0	0	0
1	0	0
0	0	0
0	1	0
0	0	1
1	0	0
0	0	0

This kind of coding is sometimes called “**One-Hot Coding**” – each observation becomes represented by a vector of variables in which one is “hot” (1) and the others are all 0.

Ordinal Categorical Variables

- **Ordinal categories** are categories that have a strict ranking, so they can logically be ordered.
- Probably the most common example of ordinal categorical variables in social science is **Likert Scales**, which are the scales often used in survey questionnaires.



How satisfied are you with *

	Very Unsatisfied	Unsatisfied	Neutral	Satisfied	Very Satisfied
Purchase	○	○	○	○	✓
Service	○	○	○	✓	○
Company Overall	○	○	○	○	✓

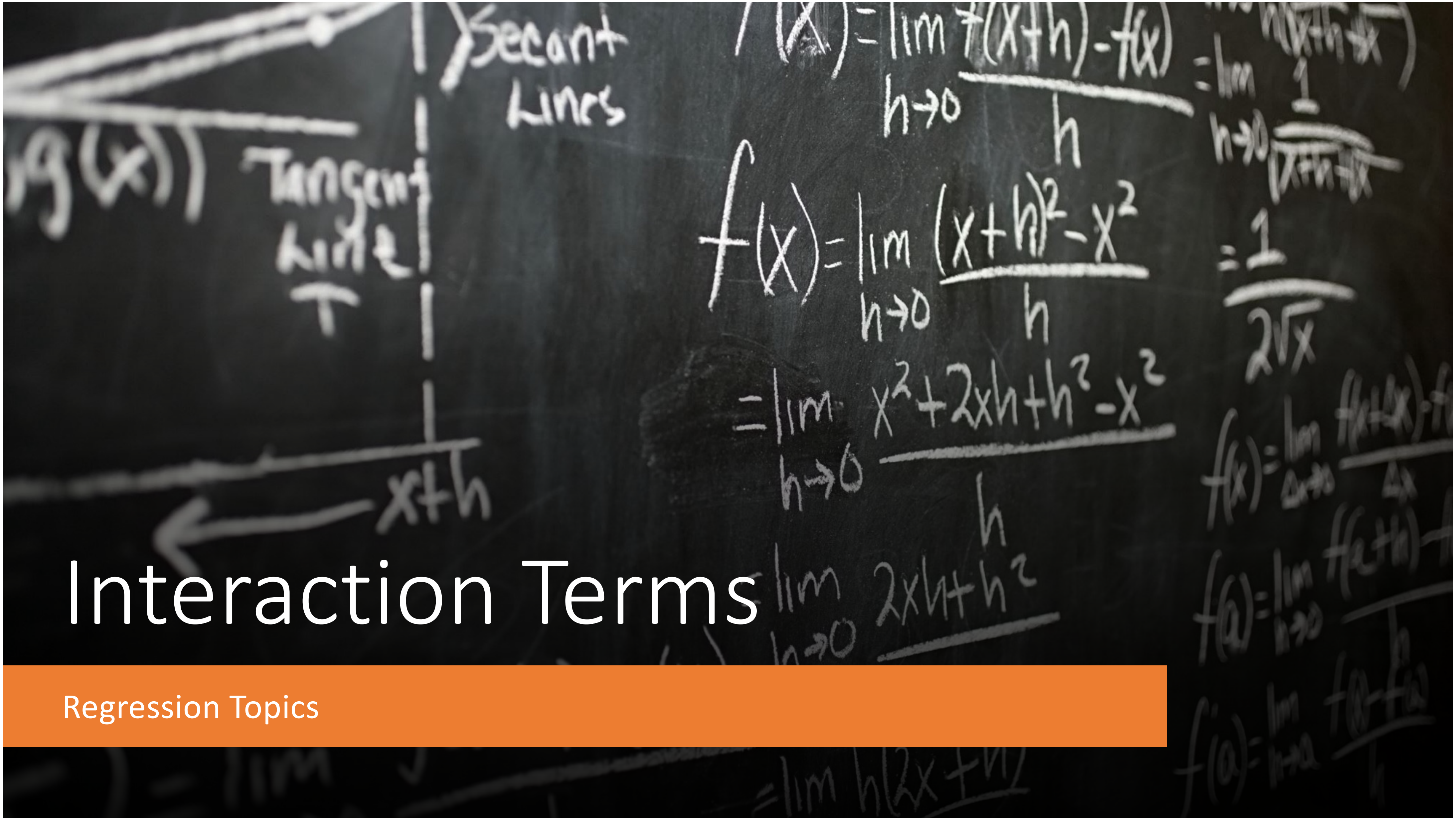
- In regression analysis, we can sometimes treat ordinal categories as continuous variables – for example, the above Likert Scale could be a variable ranging from 1 (Very Unsatisfied) to 5 (Very Satisfied).

Ordinal Categories: Warnings!

- **<u>Firstly</u>**, ordinal categories sometimes include a **“Not Applicable” option** (such as “Don’t Know” in a survey question. These need to be treated as NA (missing data) in the regression, and should not be given a number.
- **<u>Secondly</u>**, the numbers must be in **correct rank order**. This is a simple requirement but a common mistake – sometimes survey data comes with the numbers for the ranks mixed up.
- **<u>Finally</u>**, and perhaps most importantly, the ranks must represent **identical levels of difference**. In other words, the gap between rank 1 and rank 2 must be the same as the gap between rank 2 and rank 3, and so on.
  - For Likert scales, this requirement is usually satisfied, so they’re quite often treated as continuous numbers (although there is some debate over whether this is valid).
  - However, think about something like level of education (High School – University – Grad School); are the gaps between each of those ranks in some way “identical”?
  - Similarly, age or salary groupings might not cover identical ranges. 18-24 is not the same gap as 25 to 34!
- **When an ordinal category ranking does not match these requirements, it should be treated as a nominal category, and dummy variables should be created.**

Categorical Dependent Variables

- So far, we have discussed using categorical variables as independent (predictor) variables. What about the case where our dependent (outcome) variable is categorical?
- For strictly **ordinal categories**, we can use them as a continuous outcome variable – but be careful about interpretation!
- For **binary nominal categories**, we can run a **Logistic Regression** (which we'll discuss in one of the next classes).
- With **multiple nominal categories**, this becomes a type of machine learning / data science problem called **classification**.
  - Classification problems are beyond the scope of this course, but there are more advanced classes which introduce some aspects of machine learning classification algorithms.

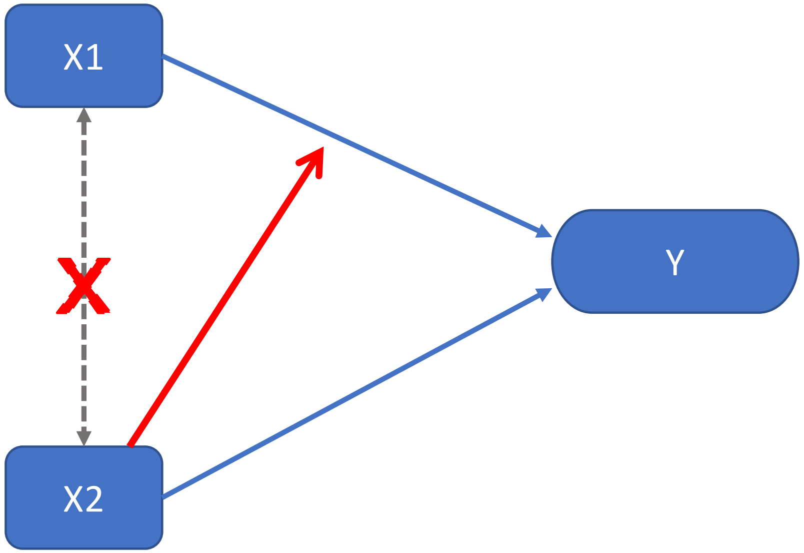


Interaction Terms

Regression Topics

Regression with Interacting Variables

- In the previous class we discussed the problem of **multicollinearity** – when independent variables are correlated with each other (as well as with the outcome variable), so it is hard to figure out the degree to which each of them is influencing the outcome.
- There are also cases when two independent variables are not correlated, but they do **interact**.
- This means that the effect of X_1 on Y is influenced by the value of X_2, even though the value of X_1 is independent from the value of X_2.



An Example of Interaction

- It’s easier to think of this with an example. Imagine a country where gender has no significant effect on the possibility of getting a university education (Japan is a good example – men and women are equally likely to attend and graduate university).
- The standard linear model, $Salary = \alpha + \beta_1 Sex + \beta_2 Univ + \varepsilon$, would let us see the effects of gender and university education on salary separately from one another – *ceteris paribus*.
- But what about the effect of being a woman with a university degree, or a man without a university degree? We might hypothesise that these variables **interact** – that employers value university education more in one gender than another, for example.

Modelling Interaction Effects

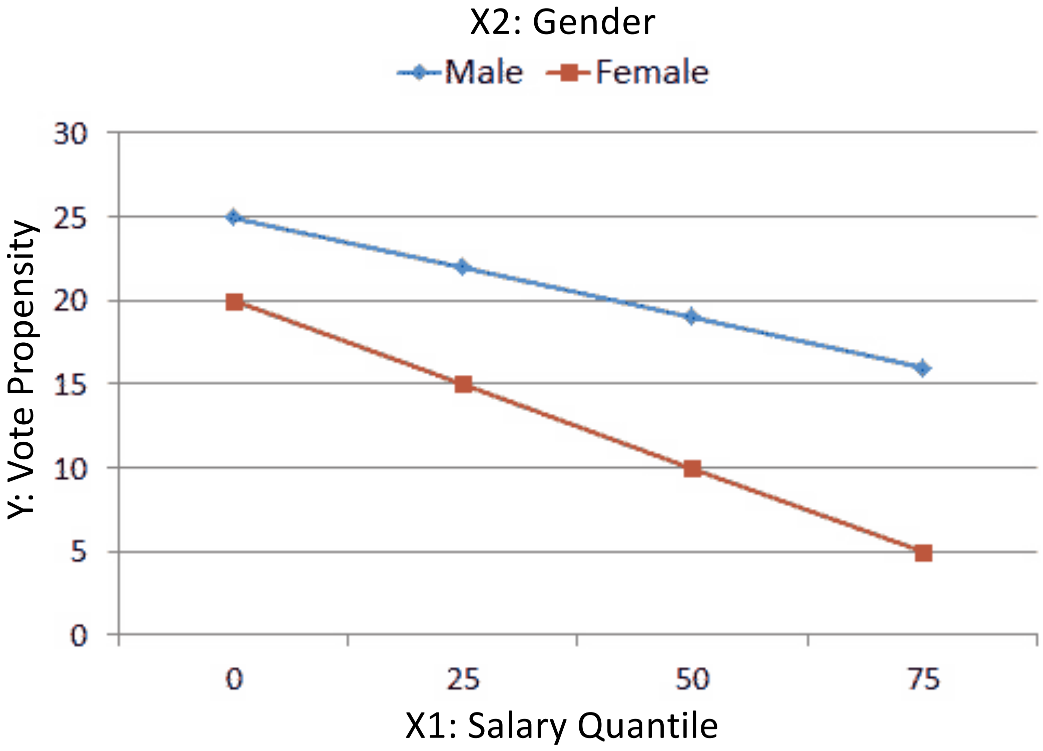
- To check the effect of this interaction, we add an **interaction term** to the regression formula.
- $Salary = \ \alpha + \beta_1 Sex + \beta_2 Univ + \beta_3 SexUniv + \varepsilon$
- In this formula, we are still looking at the effects of gender and university as separate variables, but we are also adding a new variable which combines their effects and has its own coefficient and p-value.
- The calculation for this is as simple as it looks – the interaction term is the product of the two variables we expect to interact.
  - You *could* also make interaction terms with more than two variables, but it becomes almost impossible to interpret the results, so it’s very rare to see this.

Interpreting Interaction Terms

- Generally speaking, you should show results for regression models with and without the interaction terms, to show how including the term helps to explain the phenomenon being studied.
- If an interaction term is statistically significant, and improves the Adjusted R2 for the model, this indicates that the interaction between the two variables plays a meaningful role in the outcome.
- Including interaction terms makes interpretation of coefficients difficult, so if an interaction term doesn’t meet these criteria, it shouldn’t be in your final model.
  - You should still show the model where you tested it, so readers can confirm that the interaction wasn’t relevant.

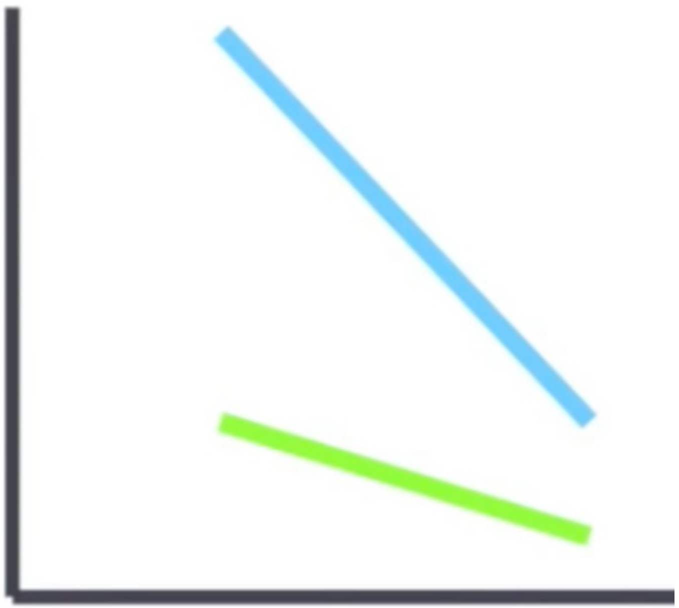
Understanding Interactions

- Because we can’t easily interpret the coefficients of a regression model with an interaction term, we need a different way to explain the interaction.
- Drawing an **interaction chart** is a useful tool. This type of chart displays the outcome variable and the two interacting independent variables.



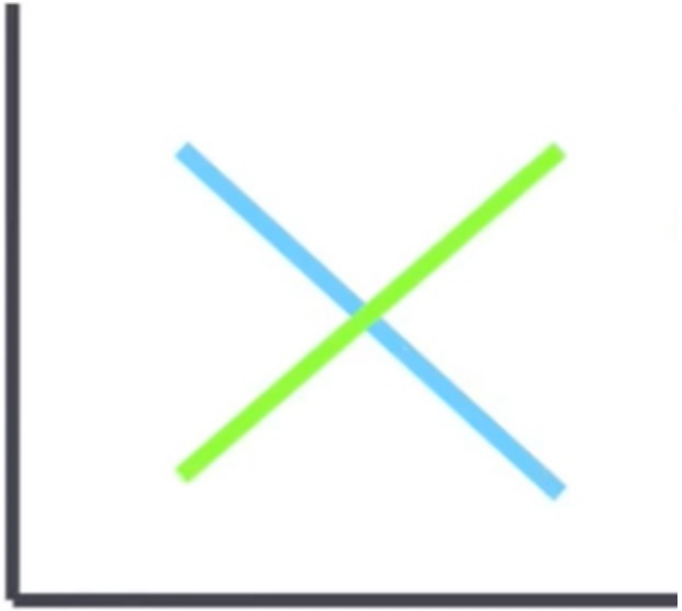
Types of Interaction

Exponential Interaction



Effects are consistent, interaction makes them stronger in some cases than others.

Antagonistic Interaction



Interaction reverses the effects in some cases compared to others.

If the lines are parallel, there is no interaction between the variables.

Note that finding non-parallel lines (either exponential or antagonistic) doesn’t mean the interaction is necessarily significant or relevant – you need to test that with a regression model!

Interactions with Continuous Variables

- So far we have been assuming that at least one of your interacting variables will be categorical (e.g. gender, race), which makes it easy to split up the effects and draw an interaction chart.
- When both variables are continuous, we can still model the interaction in the same way – but drawing and interpreting an interaction chart becomes tricky.
- In most cases it's best to **convert one continuous variable into a categorical variable**.
- This is done by splitting it up into ranges – salary ranges, age ranges, or simply "Low – Medium – High" for something like vote propensity or test scores. Then we can treat this variable as categorical and draw the interaction chart as before.

Over to R...