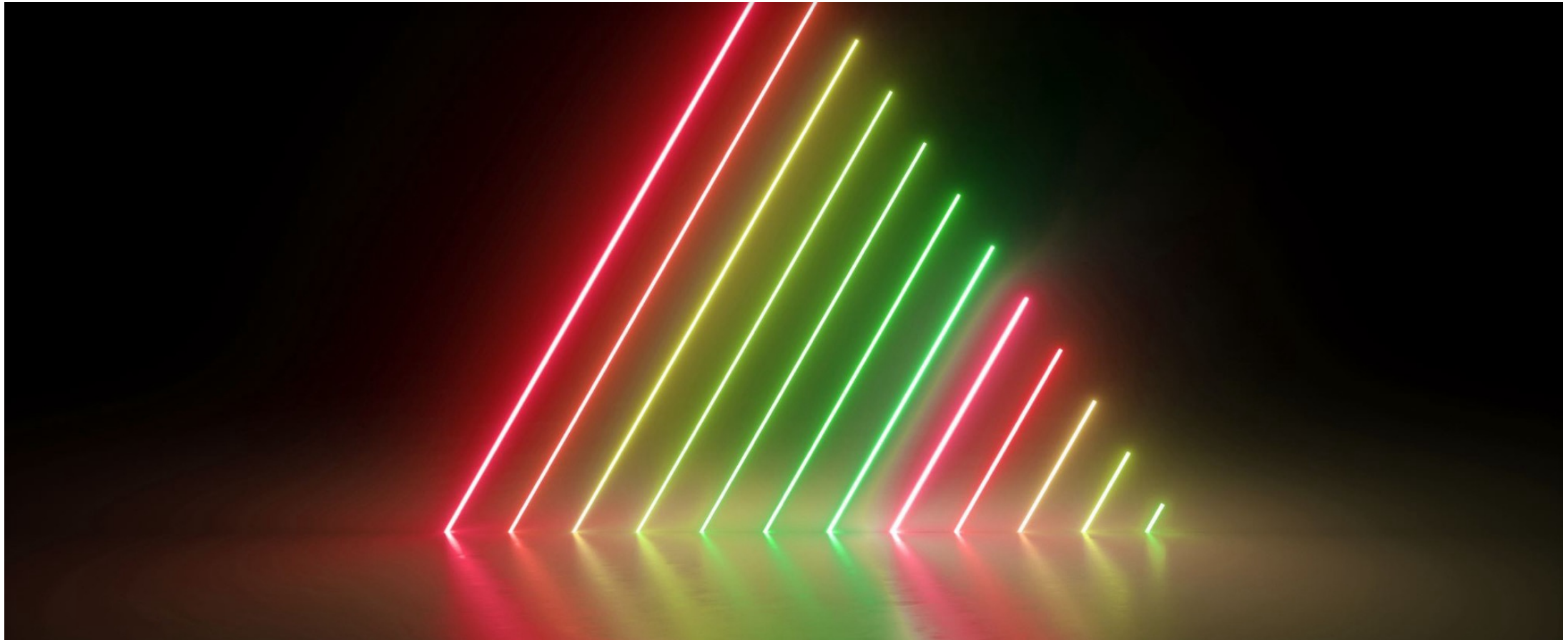
Advanced Regression Topics

Quantitative Analysis

Week 9



Review

In the past two classes, we have looked at all the techniques required to do basic regression analysis.

We learned how to model the relationship between variables by fitting a straight line to the data in such a way as to minimise the **residuals** / **errors**, using a technique called Ordinary Least Squares (OLS).

We then saw how additional variables can be added to this model, turning it into **multiple regression**, and learned how the coefficients in a multiple regression must be interpreted **ceteris paribus** – i.e., with all other variables being held equal.

Review (2)

We learned how to include categorical variables in regression models by creating **dummy variables**, which contrast each category against a chosen default category.

When you have two categories, you create a single binary (1/0) variable; for multiple categories, you create *n-1* dummies, for every category except the default.

Finally, we learned how to model **interactions**, or **moderation effects** – which occur when the effect of one predictor on the outcome is influenced by the value of another predictor, even though the two predictors are not correlated.

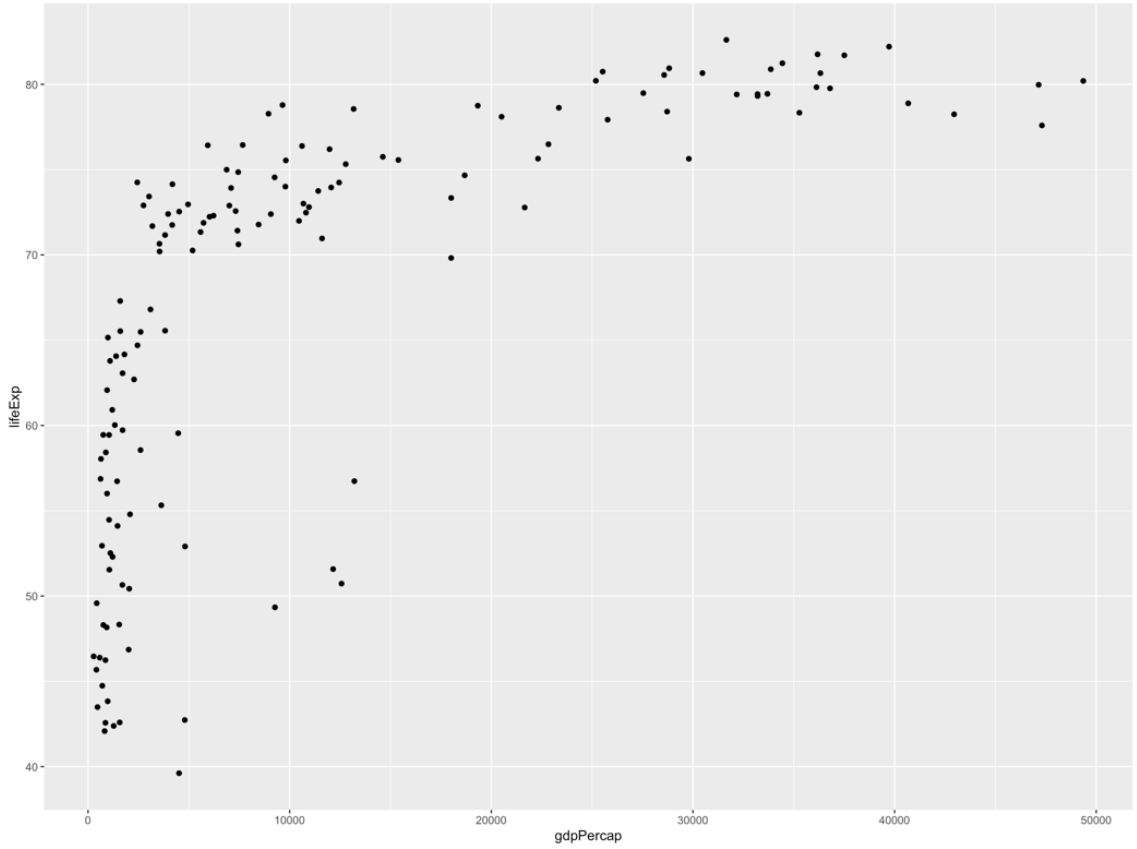
A lot of published research relies solely on these methodologies – simple and multiple regression, sometimes using categorical variables and interactions.

More complex methodology is not necessarily better methodology.

Choose methods that fit your data and questions, not methods that show off your advanced knowledge!

This Week

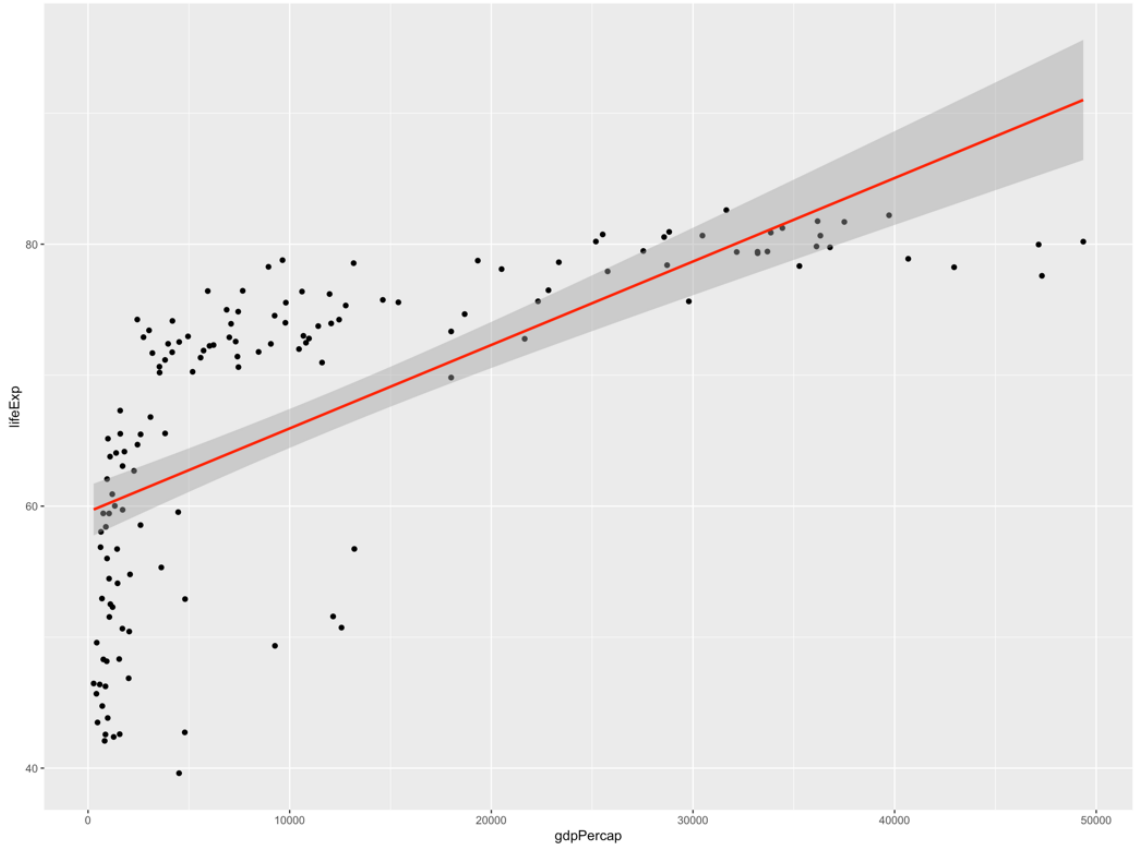
What about situations where the relationship between variables is not a straight line?



You may remember this graph from Week 6, when we first discussed correlation (using Pearson's R and Spearman's rho).

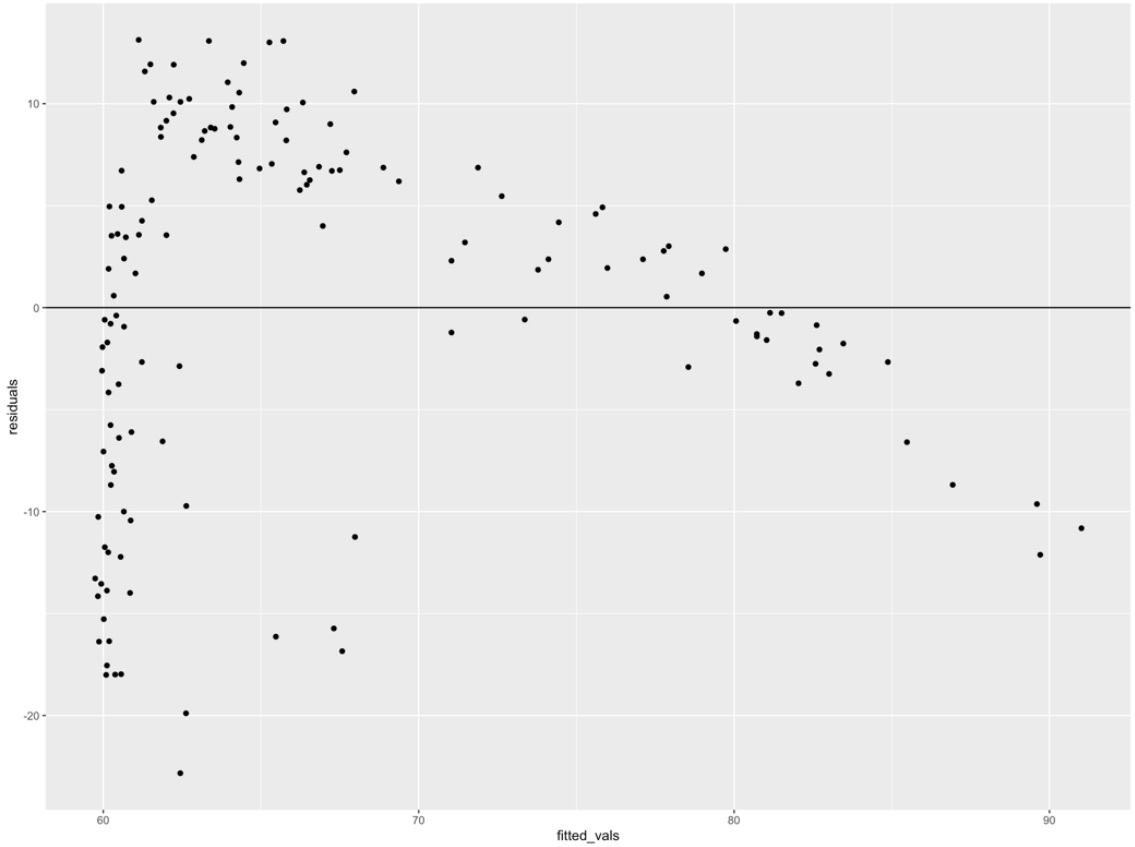
It shows the relationship between GDP per capita and life expectancy across a large sample of countries in 2007.

At the time, we noted that while the correlation tests show a positive relationship, a straight line doesn't describe that relationship very well.



You can see that clearly when we run a standard linear regression on this data. The line is positive and statistically significant – but it obviously doesn't model the actual relationship very well.

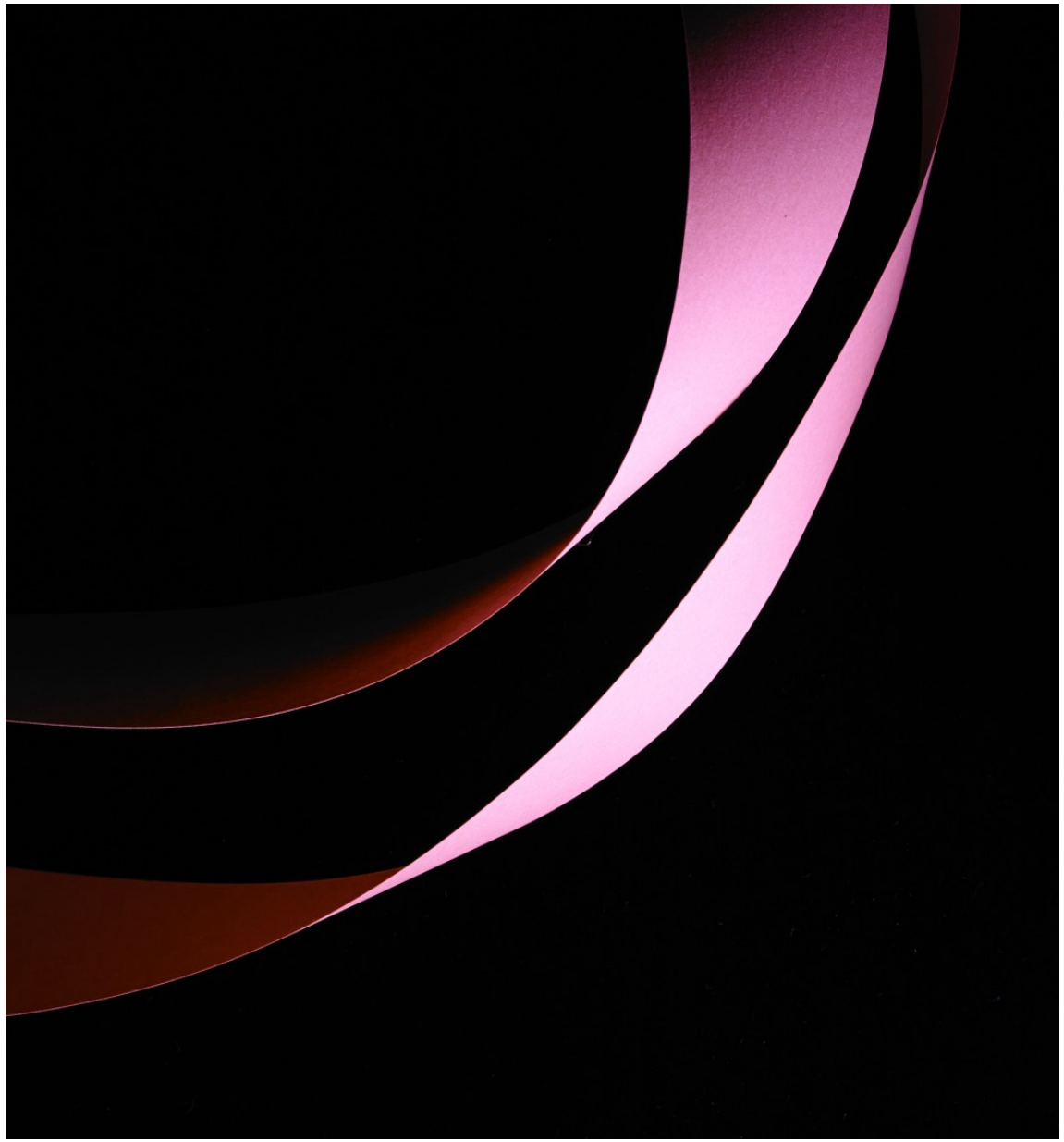
The biggest problem here is that the errors are **not random**. For some ranges of GDP, the residuals (errors) are all negative; in other ranges, they are all positive. There are **systematic errors**, not random, or **spherical**, errors.



This graph shows the residuals in the linear model.

Ideally, these should be randomly scattered in a normal distribution around the zero line (zero meaning the predicted value perfectly matched the measured value).

Here, there is an obvious pattern to the errors, so there is a problem with the model. This could mean **omitted variable bias** – but in this case, we know from visual inspection of the data that the problem is a poorly fitted regression line.



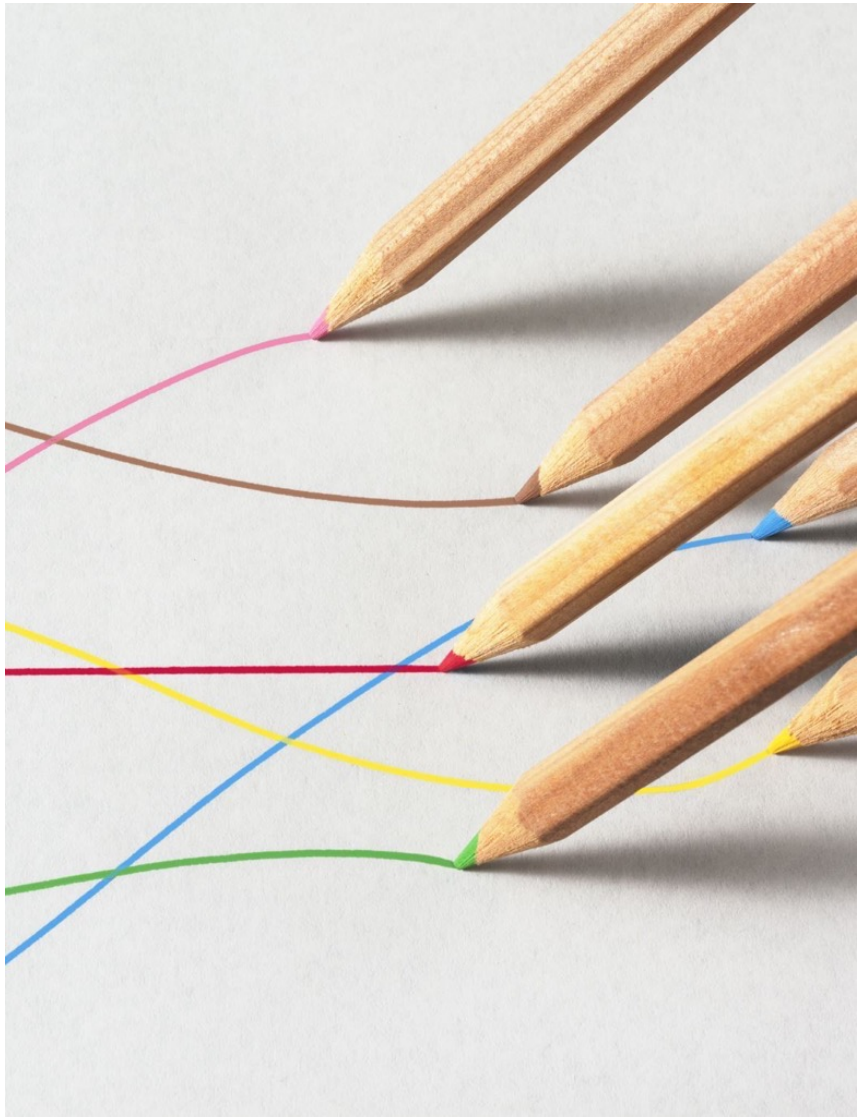
Modelling Curvilinear Relationships

- Polynomial Regression
- Variable Transformation

Core Concepts

The central idea of modelling a **curvilinear relationship** is the same as for simple linear regression; we're looking for an equation that describes the curve that best fits the data, i.e., which minimises the residuals.

We can do this in two basic ways: either by adding exponential terms to the model description (**polynomial regression**), or by transforming the variable using a function that creates a curve (**logarithmic** or **reciprocal** transformations, among others).

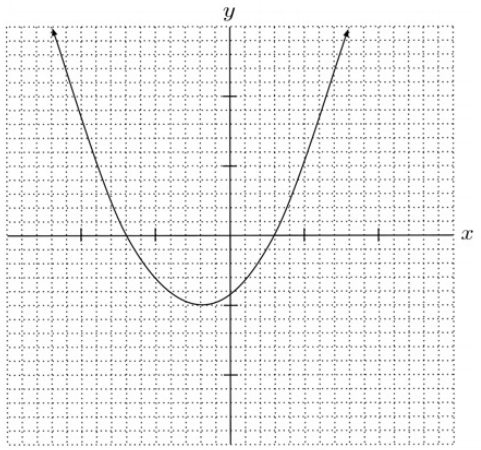


Polynomial Regression

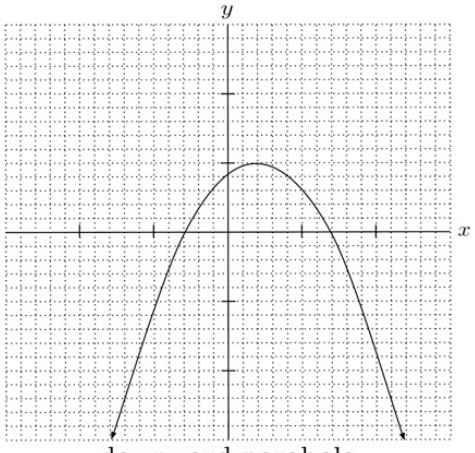
$$y \sim \beta_1 x + \beta_2 x^2 + \varepsilon$$

This is the simplest form of polynomial regression, featuring just two terms – the variable x, and its square, x^2.

x^2 is called a **quadratic term** and including it in the equation changes the line from a straight line into a U-shaped parabolic curve.



upward parabola



downward parabola

$$y \sim \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon$$

Polynomial regression simply adds more power terms to the quadratic equation.

(Technically, $x + x^2 + x^3$ is a **cubic term**; adding x^4 makes it **quartic**, x^5 is **quintic**... Collectively it's easiest to call them all **polynomials**.)

Each term you add creates a more complex and refined curve. In theory, you could add lots of terms and create a line that perfectly fitted your data, weaving through every point on the graph.

Adding too many terms is called **overfitting** – it makes your model incredibly sensitive to outliers in the data and makes it useless for prediction.

How many Polynomials?

To decide the right number of polynomials to use, we use the **p-value** to estimate the statistical significance of each polynomial term.

You usually start from quite a high number of polynomials – 10 is common – and work backwards, reducing the number until you find a model where the highest polynomial is **statistically significant**.

Above that number, adding complexity to the curve isn't adding any meaningful value to the model – it's just overfitting to the data.

Note that we keep lower-order polynomials in the model even if they are not statistically significant – leaving them out would skew the results in undesired ways.

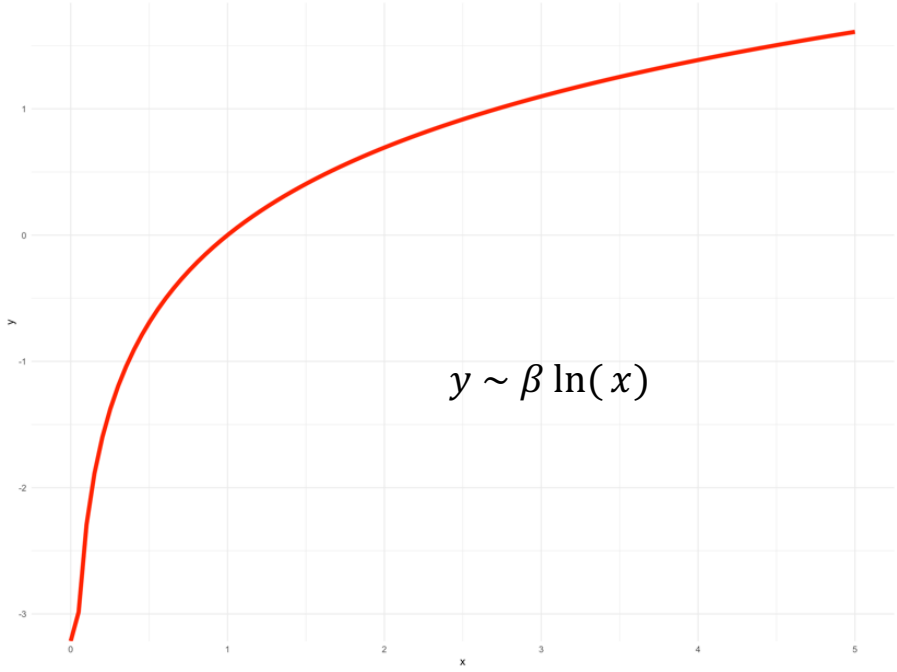
Logarithmic Transformation

Transforming the Variable

There are a number of ways to transform the variables we are using in our model to reflect curvilinear relationships.

Two common ones are **logarithmic** and **exponential** transformations.

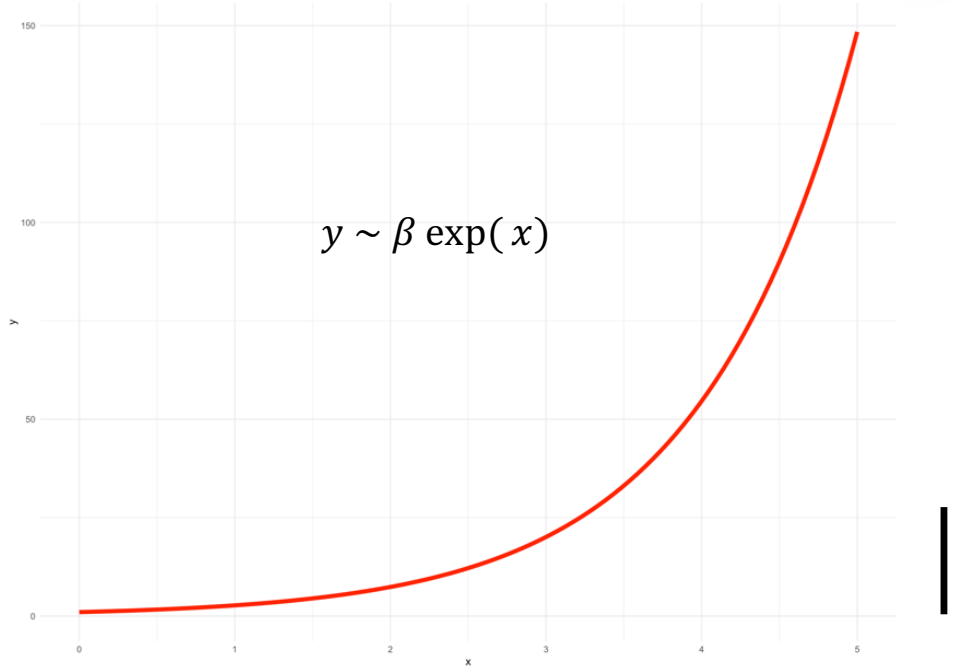
Transforming a variable using the **natural logarithm** describes a curve that grows rapidly at the start and then slows down – so it's a good fit for situations where *x* has a strong effect on *y* at low values, but the effect shrinks as the values rise.



Exponential Growth

The **exponential** transformation, on the other hand, describes almost the opposite situation – one where growth is very slow at low values, and then rises rapidly at higher values.

The inverse of this is called **exponential decay**. It creates a curve similar to the **logarithm**, but it eventually stops rising and becomes completely flat.



Reciprocal Transformation

The final type of transformation we'll look at is **reciprocal transformation**, which is mathematically very simple:

$$y \sim \beta \frac{1}{x}$$

This creates a curve that flattens out as it approaches the axes of the graph and is good for modelling curves that have a baseline value they never go below.

There are many types of transformation which express various different types of relationships – we're only discussing the most commonly used ones today.

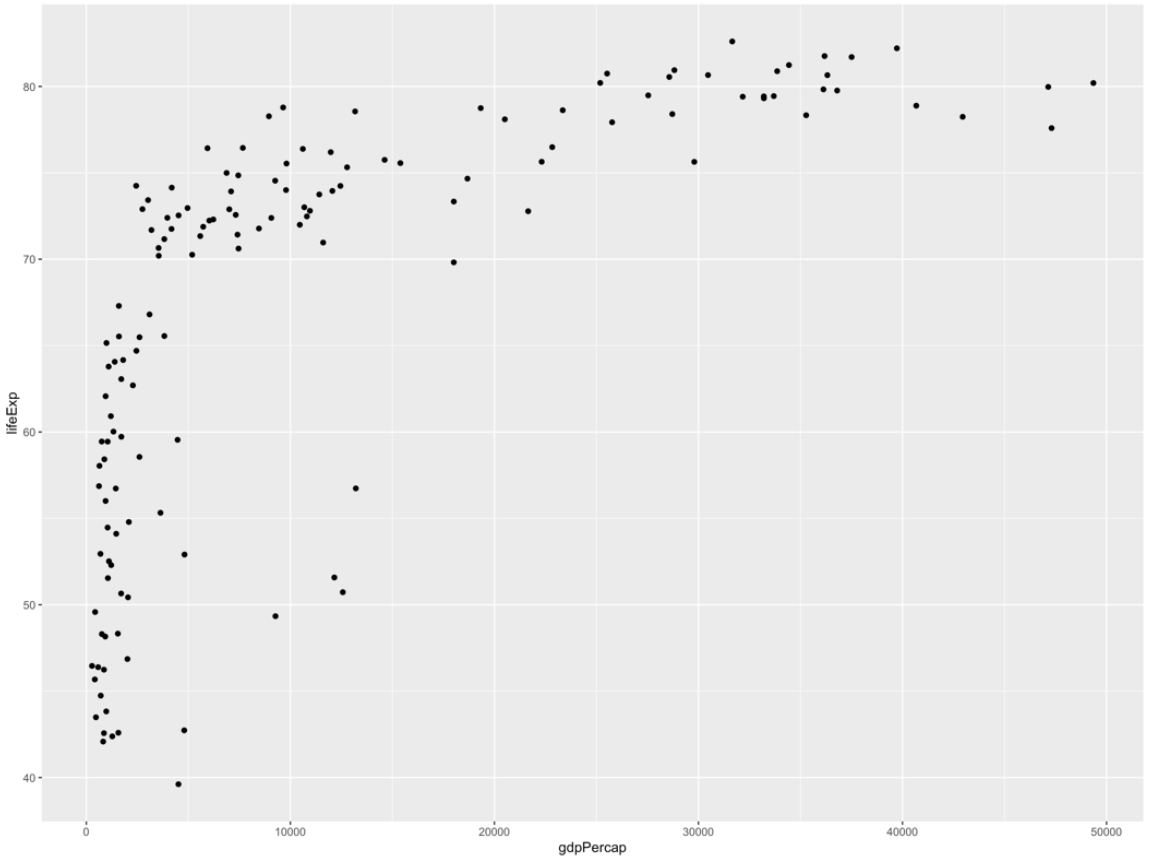
Interpreting Curvilinear Relationships

What does a curved regression line mean?

It's tempting to just find the curve that best fits our data – giving us the best R^2 values and the lowest ***p*-values** – and stop there, assuming we've found the best model. Indeed, if all we wanted was a model that gave good quality **predictions**, that would be the right approach.

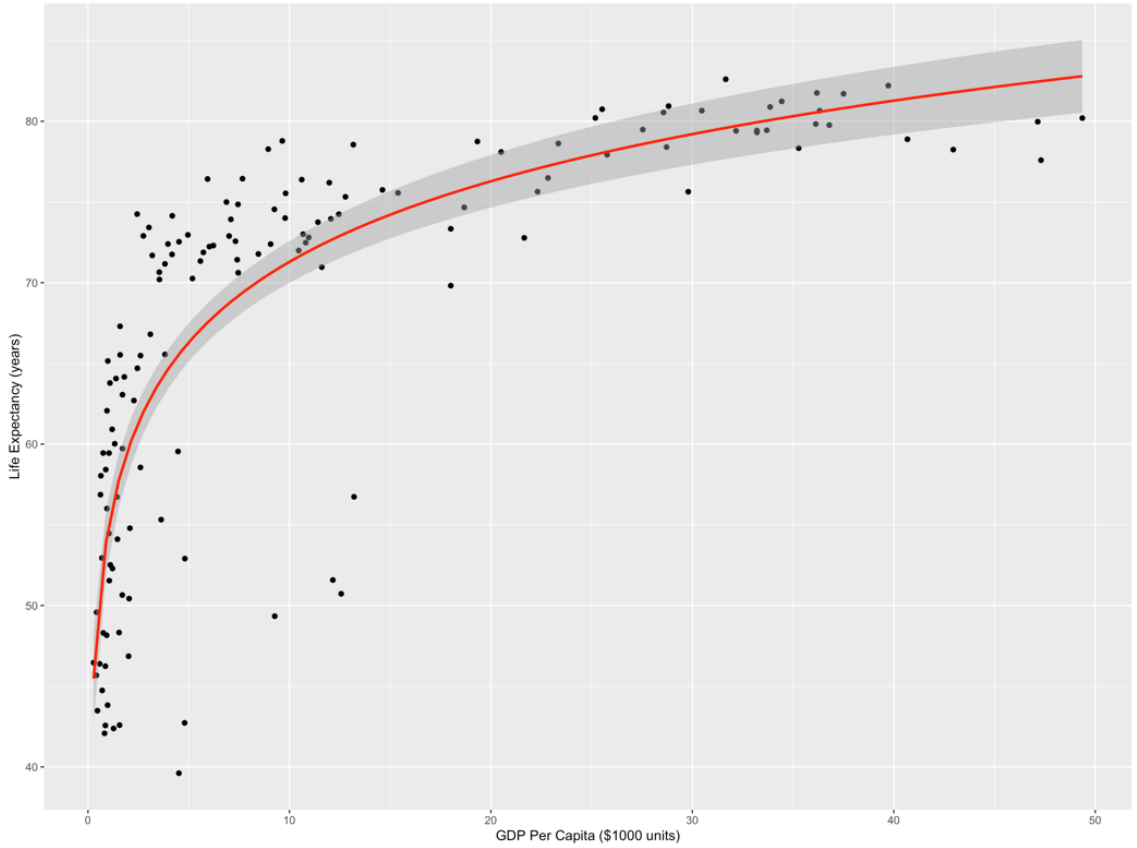
However, as researchers, we usually want to **understand** our model – to see which factors are impacting the outcome and in which ways.

That means it's not good enough to just fit a curve on the basis that it matches the data; you must also try to understand **why** this data matches a curve better than a straight line.



Look again at our GDP / Life Expectancy data. You can probably guess what kind of curve will fit this:

- It's not U-shaped, so it won't be a **quadratic**, but a higher-order **polynomial** might work;
- It rises quickly and then slows down, so a **logarithm** model seems likely to fit.
- We might also test a **reciprocal** model, although the data doesn't seem to have a hard ceiling – it does keep rising, just very slowly.



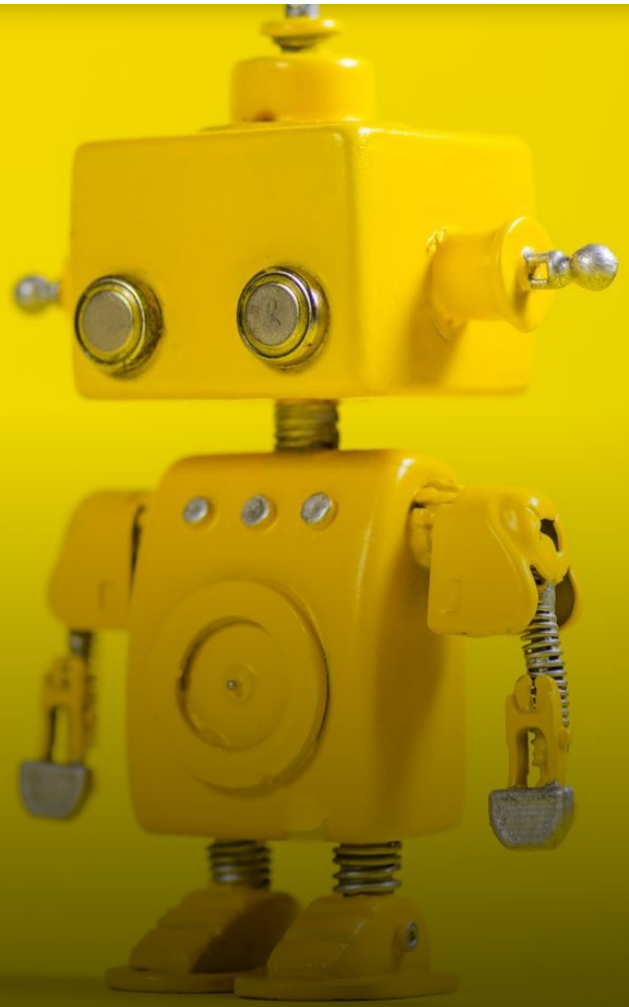
Spoiler alert: the best fit here is indeed a natural logarithmic transformation of *x*, also called a **Linear-Log Model**.

Fitting this curve is only the first step.

Now you must ask: why does this logarithmic curve fit better than a straight line? How should we interpret the shape of this curve?

What does it mean in substantive terms for GDP and Life Expectancy to have this kind of relationship?

Over to RStudio



Group Project

Each group has now submitted their Research Question assignment; you'll all receive group feedback shortly.

You have a group assignment for next week – the basic assignment is to refine your question and find the data you'll use to test your hypothesis, but each group will also get specific instructions in their feedback, so read it carefully!

There will be a further group assignment due in the last week before the winter break. Read the feedback to this assignment especially carefully – it is meant to ensure every group knows what they need to be doing in the weeks before we meet again in January.

Research Diary

Don’t forget that you should all be keeping your **individual research diary** (or lab notes) from this week onwards.

If you haven’t started already, open a new document later today, and write your first paragraph or two – explaining in your own words your group’s research question and the progress you’ve made so far. Give your thoughts on it and your personal understanding of the research process. Add a new entry each week from now on.

DO NOT SHOW YOUR RESEARCH DIARY TO ANYONE ELSE.

DO NOT COPY YOUR RESEARCH DIARY FROM ANYONE ELSE.