Intro To Linear Models

Data Science
Week 5

Welcome to Linear Models!

Today:

- Intro to Linear Models Lecture (interspersed with quizzes and breaks)
- Big Break
- WWF Report Release
- Class Activity (test your skills!)

Welcome to Linear Models!

- Linear Models = Workhorse of the Data Scientist
- Can answer a million different questions
- Core: as one thing changes, how does another? And/or, what is the difference between these groups?

For Example…

1. How do pollution levels vary between different rivers?
2. Have fish populations gone down through time?
3. Does the way fish populations change through time depend on river pollution?
4. Is there a correlation between how much bacon humans eat and life expectancy?
5. Based on data about income level across age groups and countries, what is the average (and st dev) income of 20-40yr old Spanish people?
6. Can we predict the most likely leaf size of a tree based on its species, average temp of the region, and soil condition?

In other words…

- We can use Linear Models for explanation, or prediction
- We can use them with one or many predictor variables
- These predictors can be continuous or categorical

- In this class, I will use the terms ‘predictor’ and ‘response’, but you could also use ‘independent’ and ‘dependent’, or ‘x’ and ‘y’

Quiz Part 1!

1. How do pollution levels (ammonium levels) vary between different rivers?
2. Have fish populations gone down through time?
3. Does the way fish populations change through time depend on river pollution (ammonium levels)?
4. Is there a correlation between how much bacon humans eat (rashers per week) and life expectancy?
5. Based on data about income level across age groups and countries, what is the average (and st dev) income of 20-40yr old Spanish people?
6. Can we predict the most likely leaf size of a tree based on its species, average temp of the region, and soil condition?

For each, write down which variable is the 'response' variable, and which variable(s) are the predictors.

As you write each predictor, specify if it is categorical (i.e. groups) or continuous (numerical)

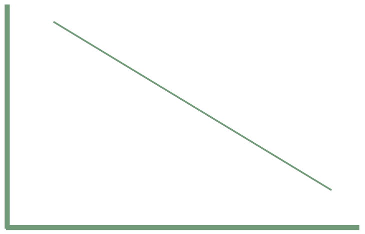
When *shouldn't* we use one?

- When you want to include many *response* variables in one model
  (many predictors is fine)

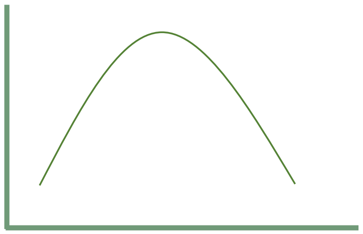
- E.g. how does river pollution vary by distance to city, where river pollution is measured by phosphate, nitrate and ammonium levels

- (note you could just do one model separately for each pollutant, but if you want to look at them simultaneously)

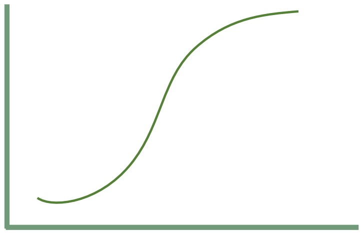
- When you don't expect simple relationships between variables



✅



✅



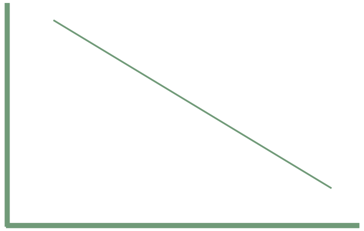
✅

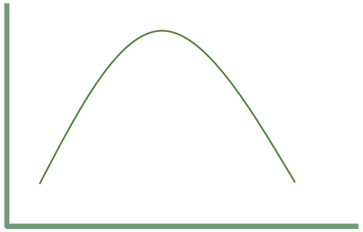
When *shouldn't* we use one?

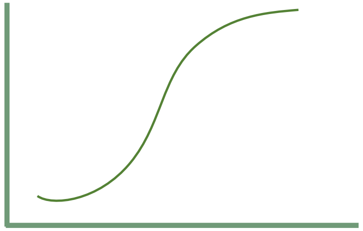
- When you want to include many *response* variables in one model
  (many predictors is fine)

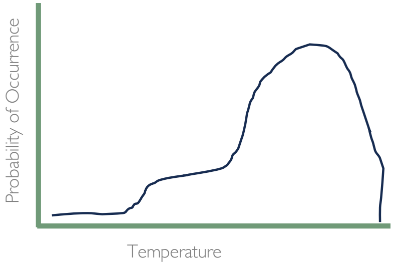
- E.g. how does river pollution vary by distance to city, where river pollution is measured by phosphate, nitrate and ammonium levels
- (note you could just do one model separately for each pollutant, but if you want to look at them simultaneously)

- When you don't expect simple relationships between variables


✅


✅


✅



✗

We will be spending 4 weeks

- *Really* getting to know Linear Models.

1. *(Week 5) – The Basics*
2. *(Week 6) – Extensions for more complex data*
3. *(Week 7) – More complex questions and improving model fit*
4. *(Week 8) – Does this model actually represent the real world?*

ALL WEEKS are relevant to the WWF report, including Week 8 (i.e. day before report is due!)

We'll allow some class time at the end of that week for you to update

Today – The Basics

- Some disclaimers…

- I am teaching this in a way that's (hopefully) more digestible. It won't always be totally accurate…

- There are different ways of doing Linear Models

Today – The Basics

- ‘Frequentist’
  - Designed out of experimental settings
  - At the base of estimates of confidence: ‘if we ran this experiment again, would we be likely to get the same results?
  - We often don’t apply to experiments now, but it’s important to know the thinking of where they started

Today – The Basics

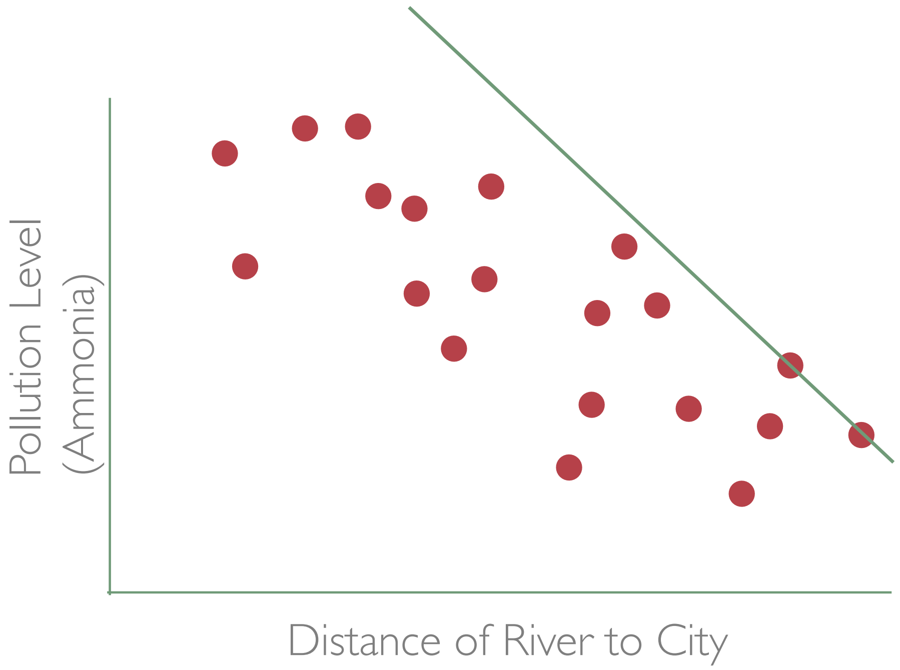
- 'Bayesian'
  - Designed to consider our knowledge about the real world
  - At the base of estimates of confidence "how likely is what we're observing to be real?"
  - Many people argue Bayesian is better… But the field is entrenched in Frequentist statistics, and it's sometimes easier to teach. So. We teach frequentist.

All about averages

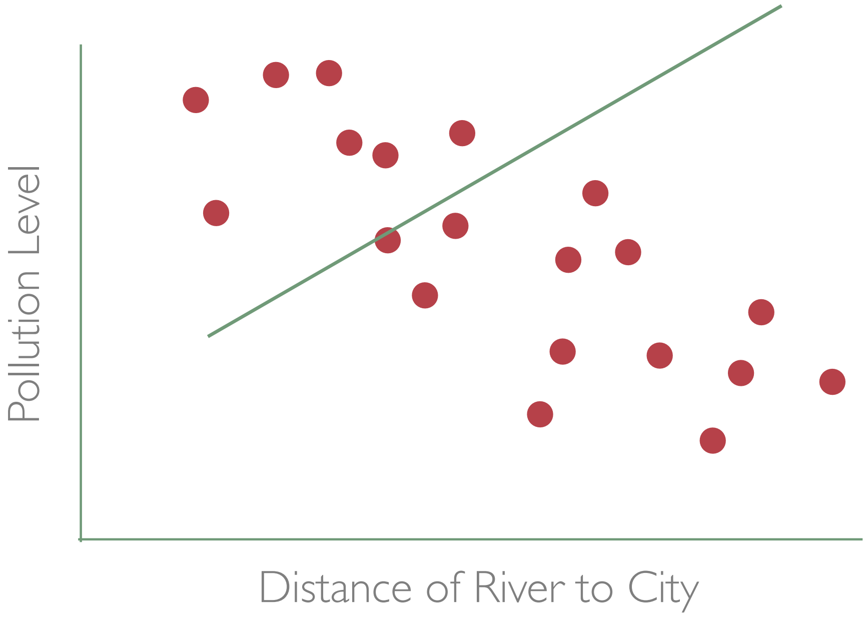
- Model gives us an average (plus a range about that average), plus confidence in how likely it is we got this estimate (i.e. p value)

- P value is the least interesting bit!!!

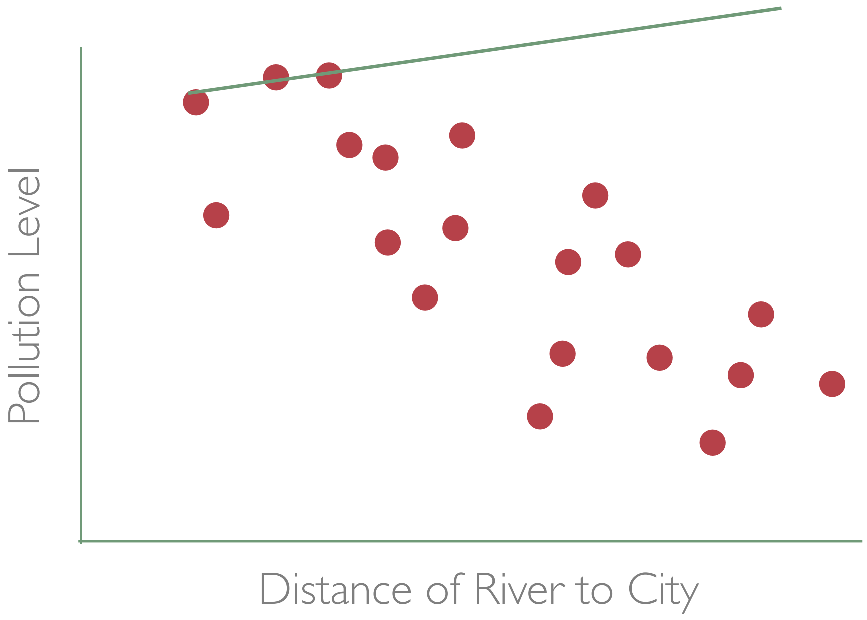
e.g. Linear Predictor



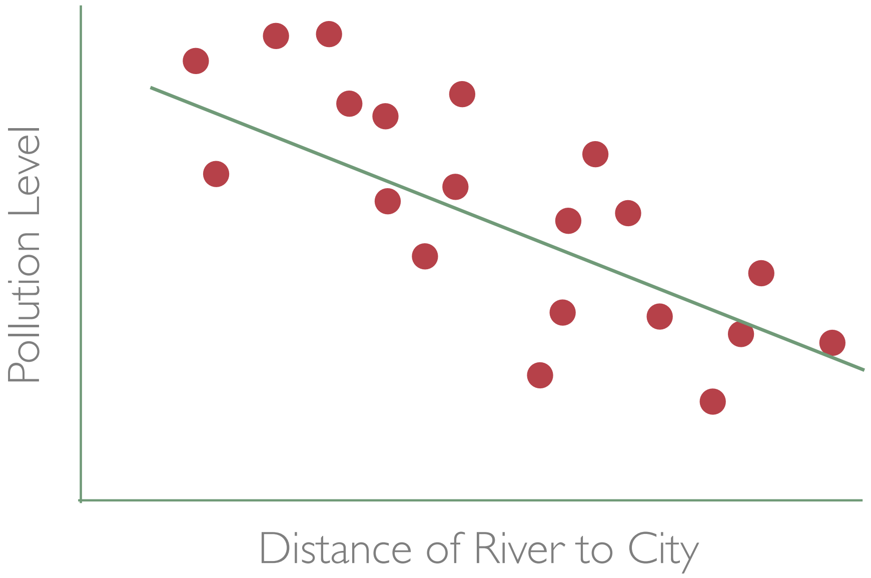
e.g. Linear Predictor



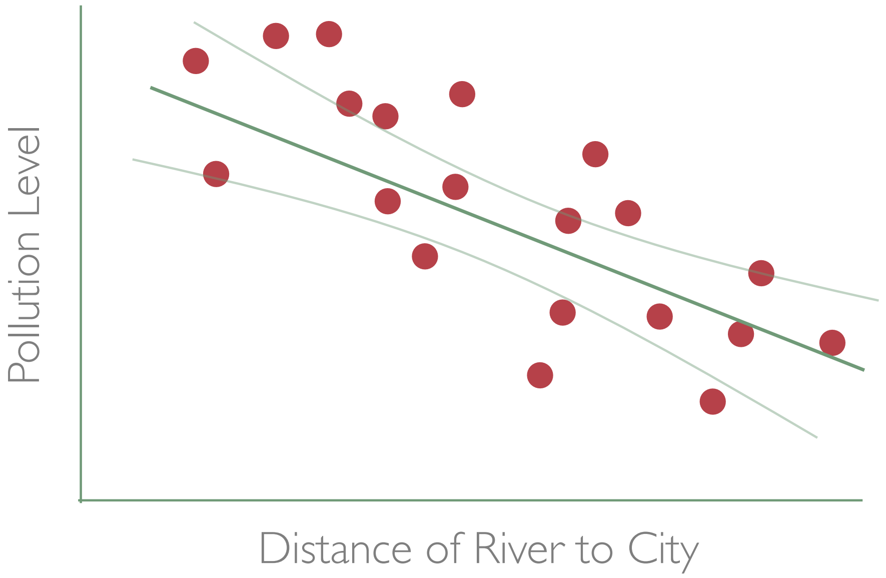
e.g. Linear Predictor



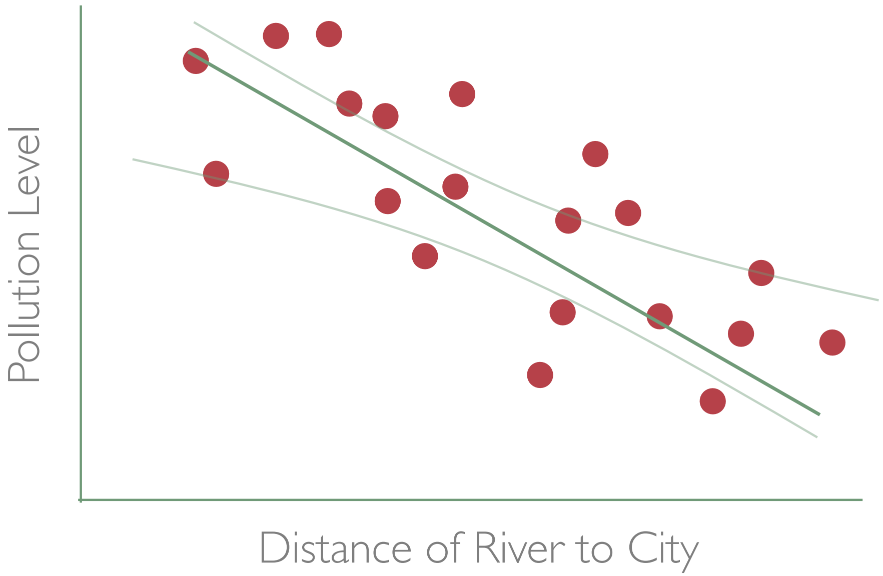
e.g. Linear Predictor



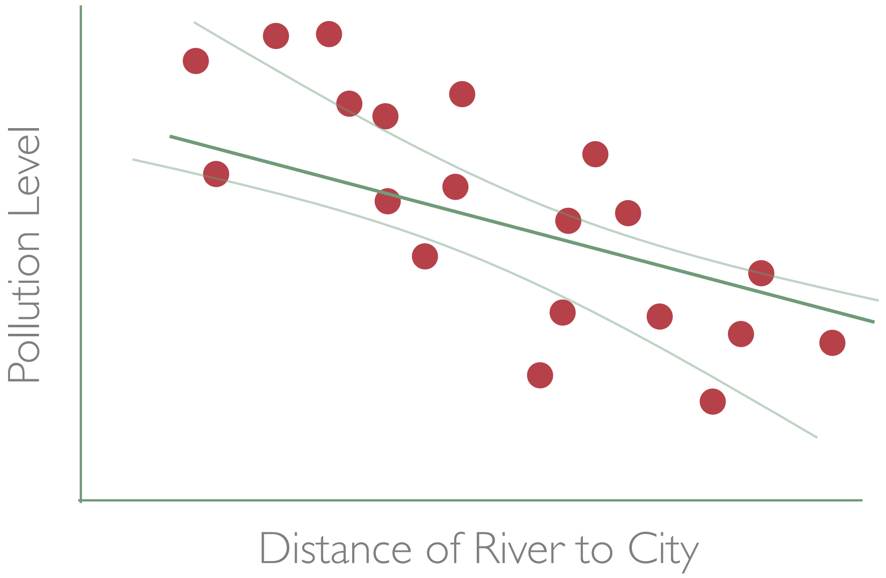
e.g. Linear Predictor



e.g. Linear Predictor

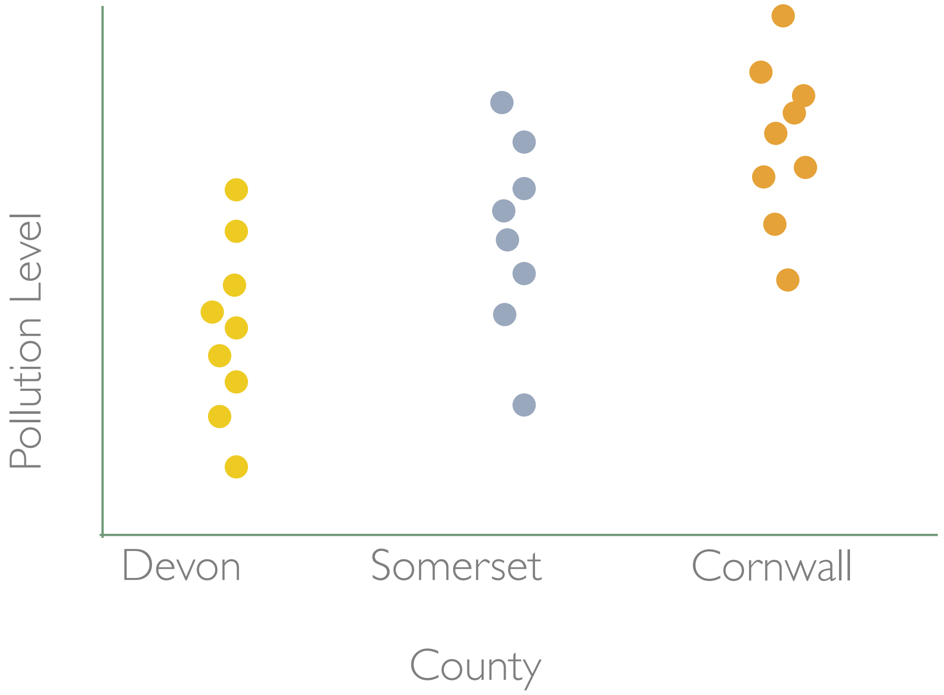


e.g. Linear Predictor

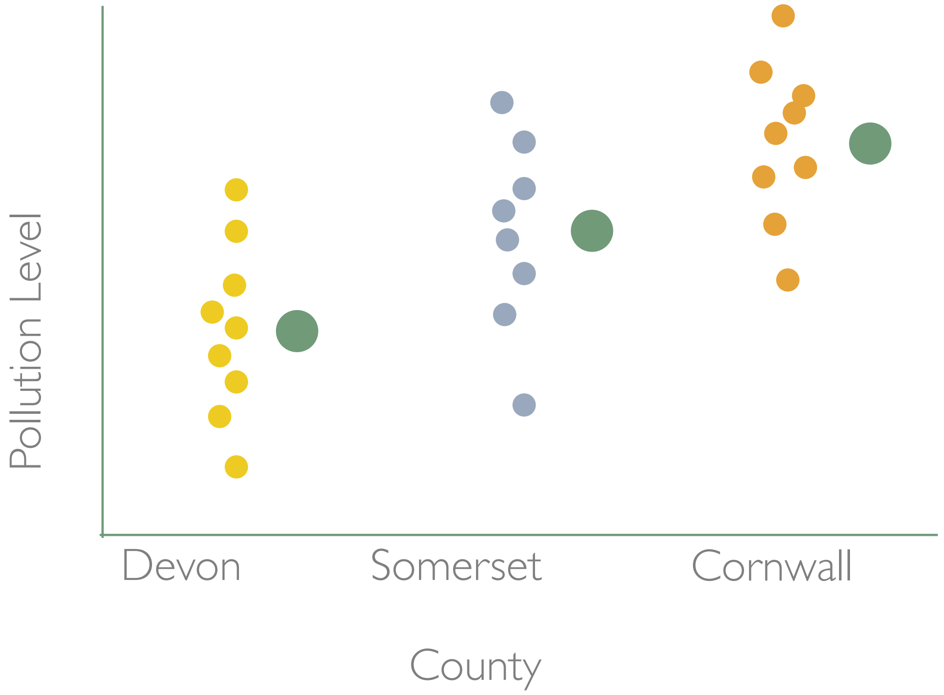


e.g. Categorical Predictor

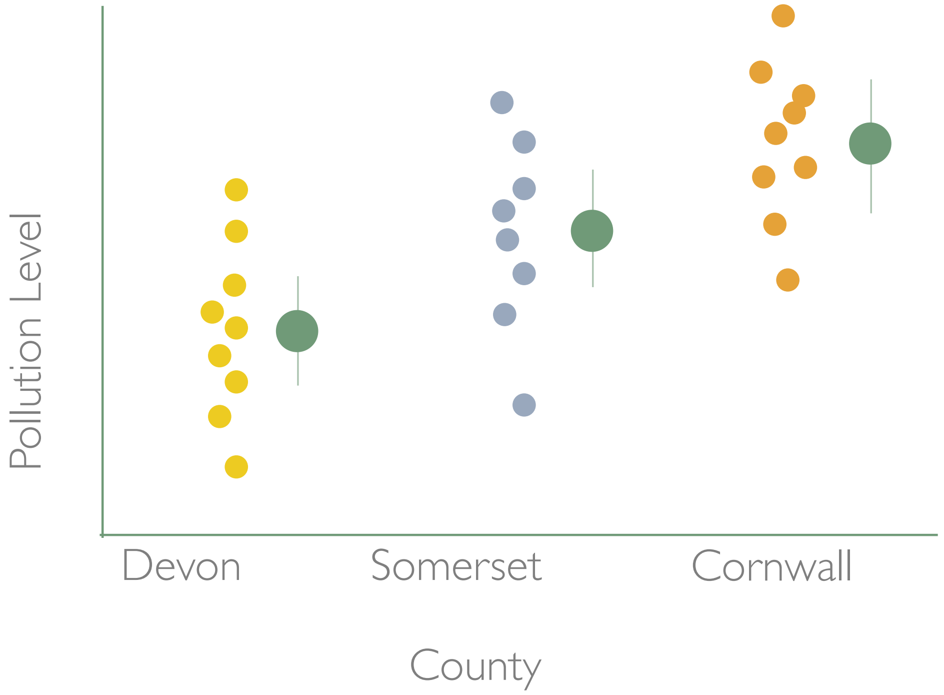
e.g. Categorical Predictor



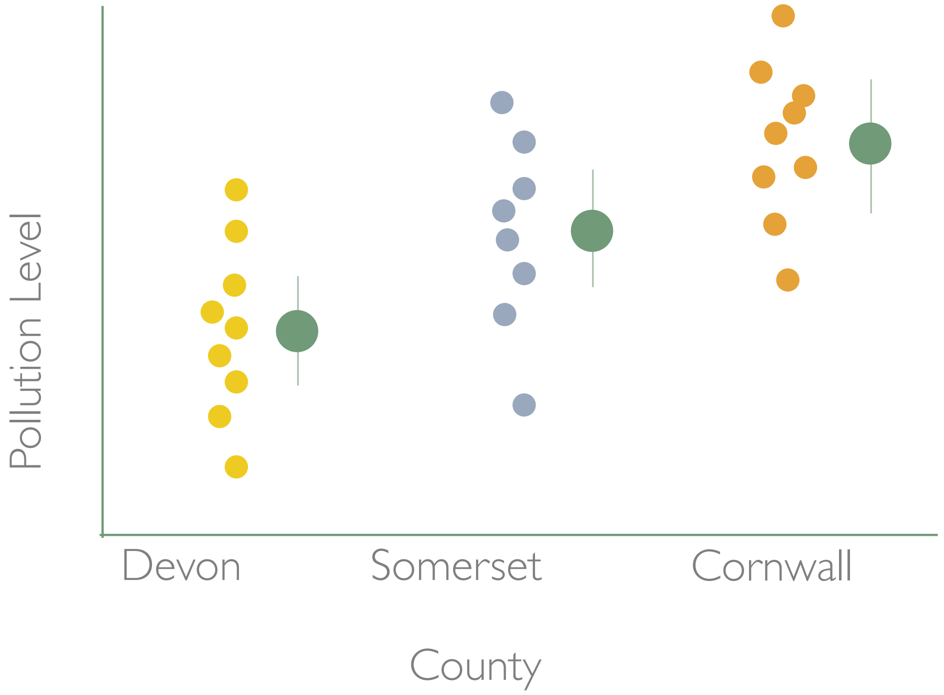
e.g. Categorical Predictor



e.g. Categorical Predictor

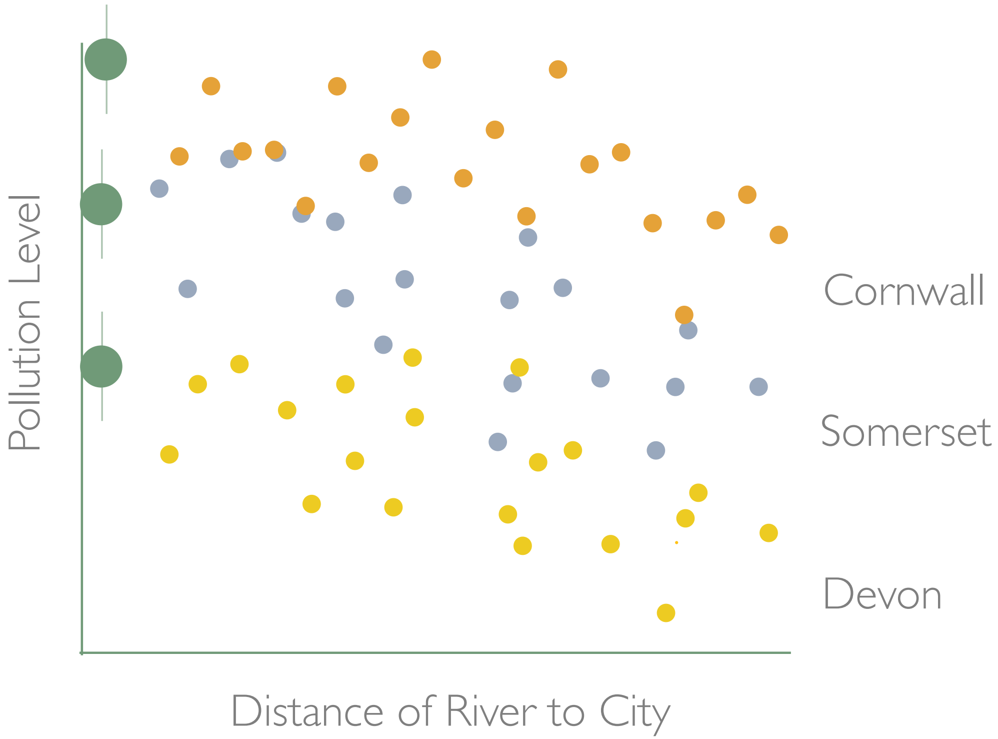


e.g. Categorical Predictor



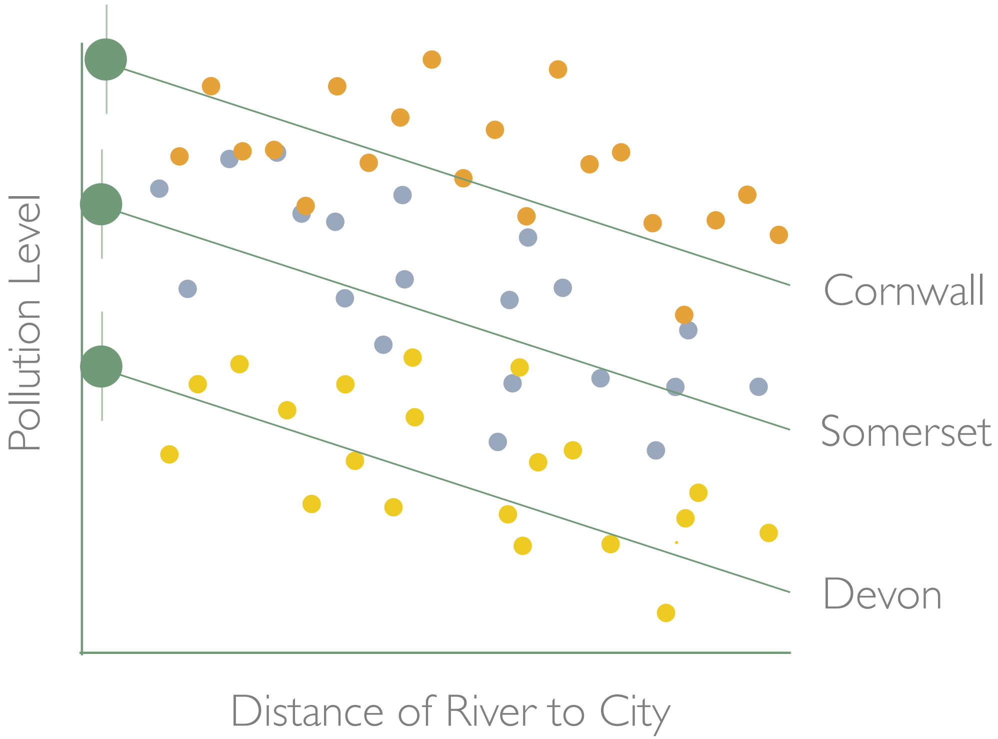
Note: A linear model with one categorical predictor is no different to an ANOVA!

If you have one of each?



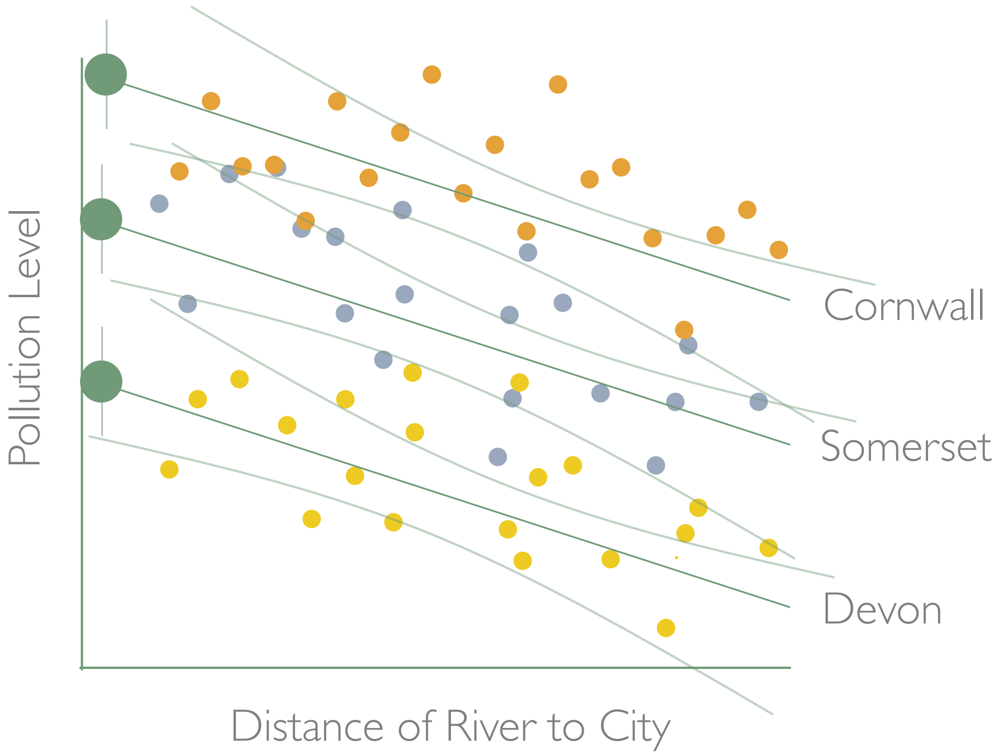
Model will find the average Y intercept for each group, then give them the same average slope

If you have one of each?



Model will find the average Y intercept for each group, then give them the same average slope

If you have one of each?



Model will find the average Y intercept for each group, then give them the same average slope

(Unless we start including interaction terms – but that’s for a future week!!)

Assumptions

This approach was designed for certain kinds of data

If your data isn’t right for the model, you can’t trust the results

There are 5 key assumptions:

Assumptions

1. Independence
2. Normal Residuals
3. Linear Relationship
4. Constant Variance
5. No Major Outliers

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Assumptions

Why does it matter?

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it matter?

Data points aren't related to each other

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it matter?

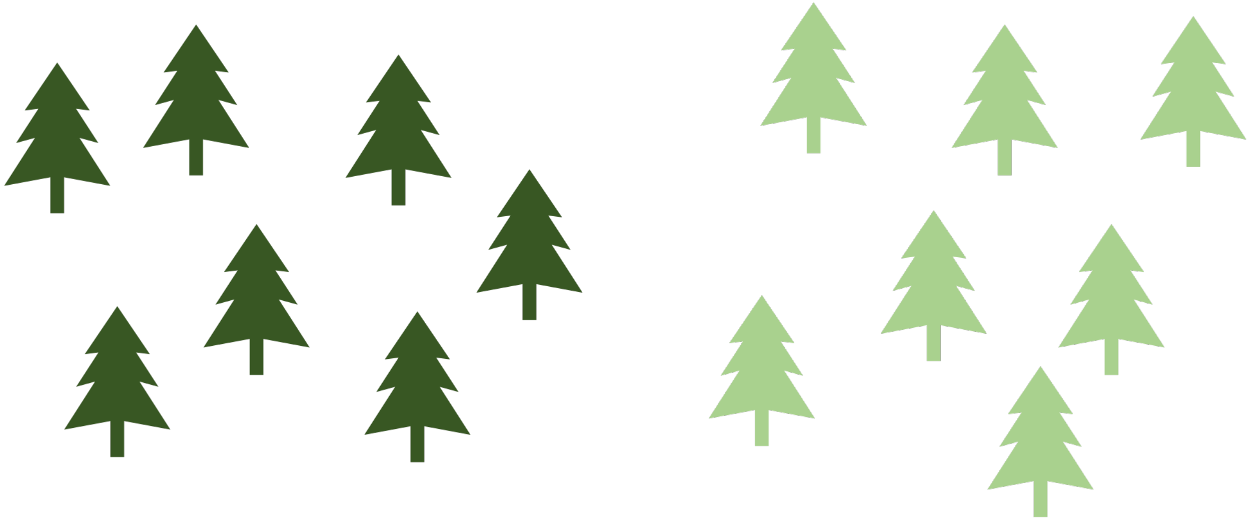
We can’t give our model an inflated sense of ego

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it matter?

e.g. Do these two tree species have different length leaves?



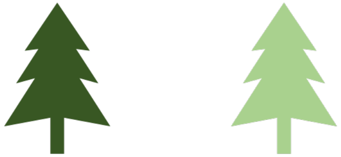
Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it matter?

e.g. Do these two tree species have different length leaves?

What about if we just sampled two trees, each 7 times?



Assumptions

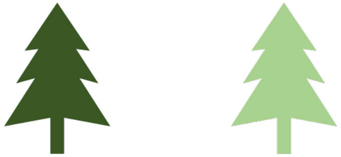
1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it matter?

e.g. Do these two tree species have different length leaves?

What about if we just sampled two trees, each 7 times?

The model won't know, it'll think that this…



Assumptions

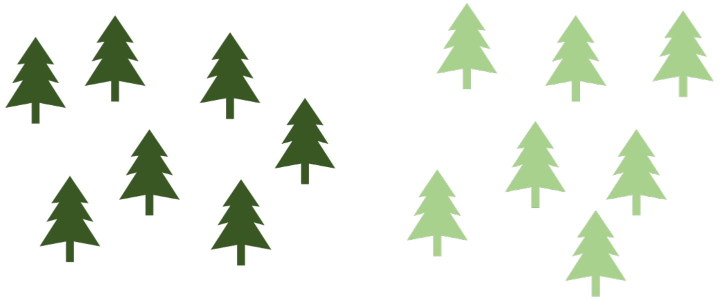
1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it matter?

e.g. Do these two tree species have different length leaves?

What about if we just sampled two trees, each 7 times?

Is actually this…



Assumptions

How to Check?

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

How to Check?

We can’t test this one with code, or any fancy analysis

It just relies on good ol intuition and your knowledge about your data

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

What to do if you violate?

Mixed effects models - See next week!

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Assumptions

Why does it Matter?

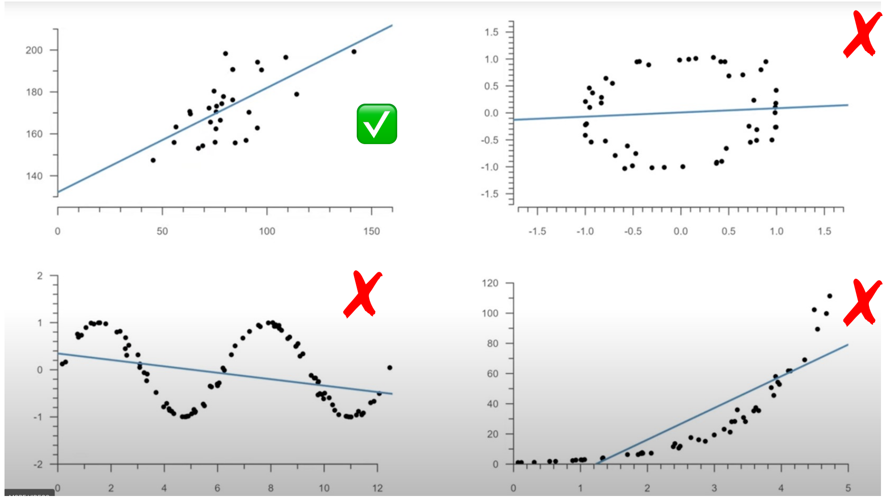
1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it Matter?

This one's easy…



Thankyou to Frans Rodenburg for visuals!! https://fransrodenburg.github.io/Applied-Statistics/simplelinearregression.html

Assumptions

1. Independence

2. Linear Relationship

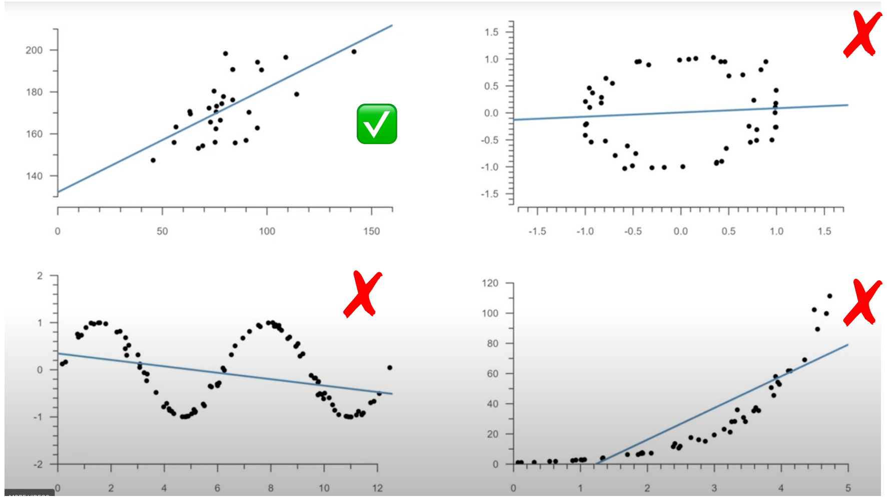
3. Normal Residuals

4. Constant Variance

5. No Major Outliers

Why does it Matter?

This one's easy…



How do you predict an average relationship with these?! There isn't one… It appears to depend on something else we're not accounting for. Any average estimate won't be representative.

Thankyou to Frans Rodenburg for visuals!! https://fransrodenburg.github.io/Applied-Statistics/simplelinearregression.html

Assumptions

How to Check?

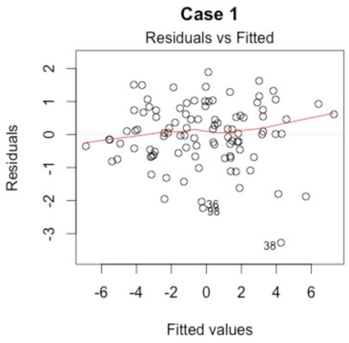
1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

How to Check?

R produces 'diagnostic plots' (see activity). The first one covers this.



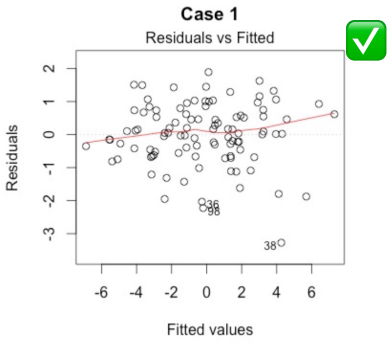
Thankyou to Frans Rodenburg for visuals!! https://fransrodenburg.github.io/Applied-Statistics/simplelinearregression.html

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

How to Check?

R produces 'diagnostic plots' (see activity). The first one covers this.



✅

Red line is fairly straight, points are fairly scattered

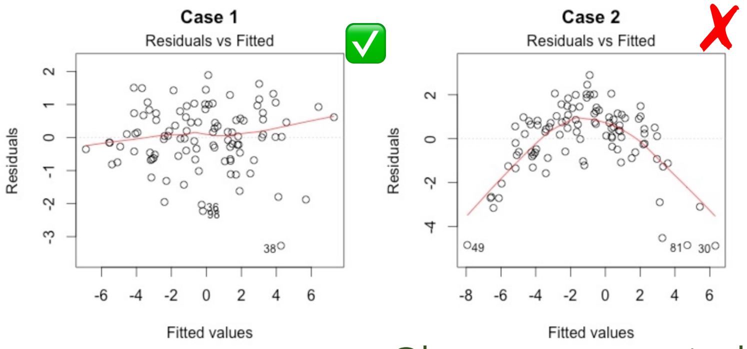
Thankyou to Frans Rodenburg for visuals!! https://fransrodenburg.github.io/Applied-Statistics/simplelinearregression.html

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

How to Check?

R produces ‘diagnostic plots’ (see activity). The first one covers this.



Clear patterns in line and data

Thankyou to Frans Rodenburg for visuals!! https://fransrodenburg.github.io/Applied-Statistics/simplelinearregression.html

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

What to do if you violate?

Sometimes it means there's a predictor variable you're missing

Sometimes you need a different kind of model

Sometimes you need a GLM – see next week!

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Assumptions

Why does it Matter?

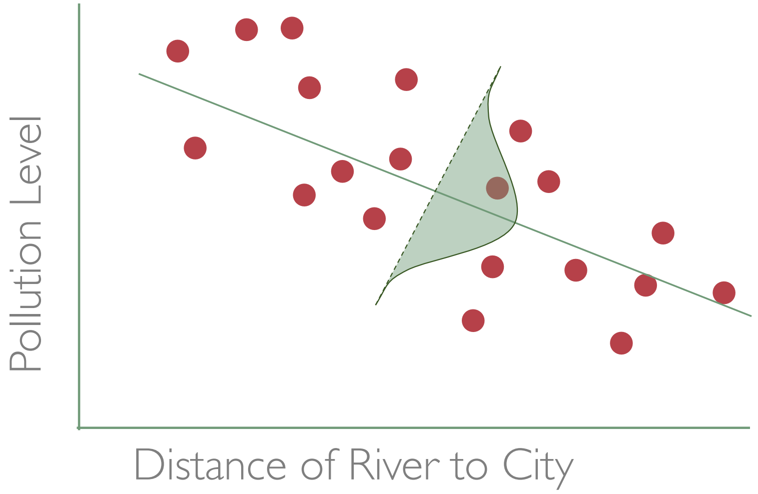
1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it Matter?

I.e. points are distributed normally above and below the trend line



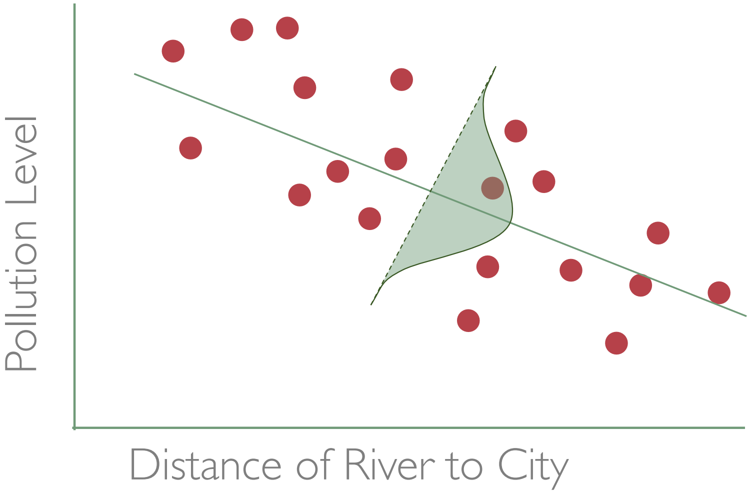
Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it Matter?

I.e. points are distributed normally above and below the trend line

All the maths that determines the output relies on the idea that the trend line can sit in the 'middle' of the points.



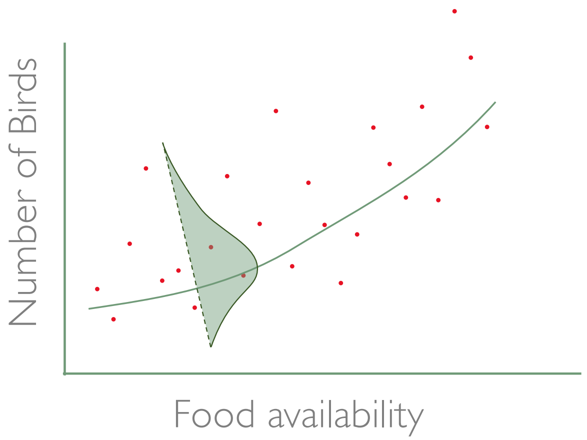
Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it Matter?

I.e. points are distributed normally above and below the trend line

E.g. this distribution is NOT normal



Assumptions

How to Check?

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

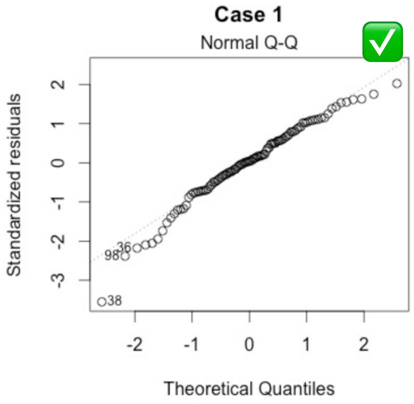
Thankyou to U Virginina for visuals!! https://library.virginia.edu/data/articles/diagnostic-plots

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

How to Check?

The second of the diagnostic plots checks for this



Points follow the diagonal pretty evenly

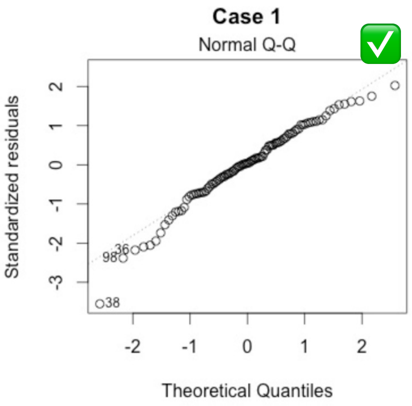
Thankyou to U Virginina for visuals!! https://library.virginia.edu/data/articles/diagnostic-plots

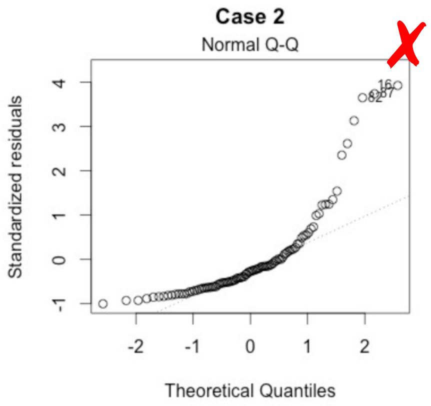
Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

How to Check?

The second of the diagnostic plots checks for this





Points are wibbly, stray very far from diagonal

Thankyou to U Virginina for visuals!! https://library.virginia.edu/data/articles/diagnostic-plots

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

What to do if you violate?

You need a GLM – see next week!

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Aka Equal Variance

Aka Homoscedasticity but I don’t call it that cos it sounds scary

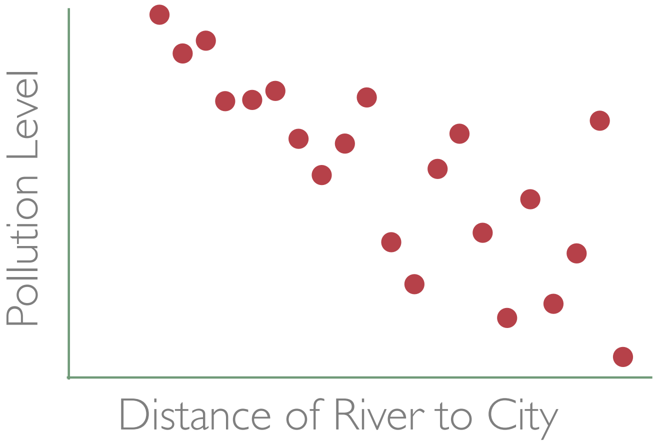
Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it Matter?

The model gives one number to estimate error about the central average. How does it get one number if the variance is different?

e.g.



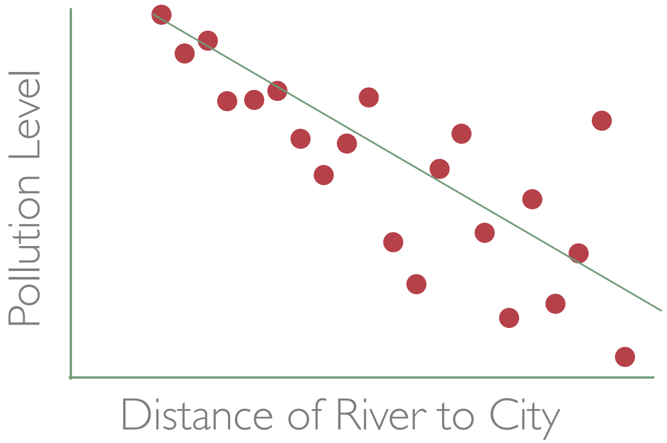
Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it Matter?

The model gives one number to estimate error about the central average. How does it get one number if the variance is different?

How do we estimate error?



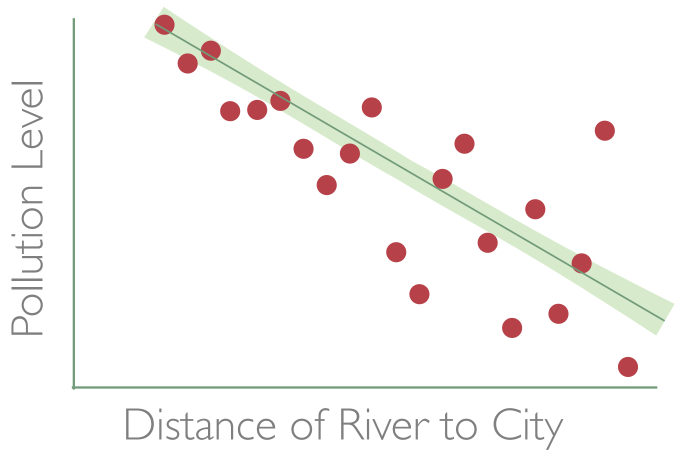
Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it Matter?

The model gives one number to estimate error about the central average. How does it get one number if the variance is different?

Is it this?



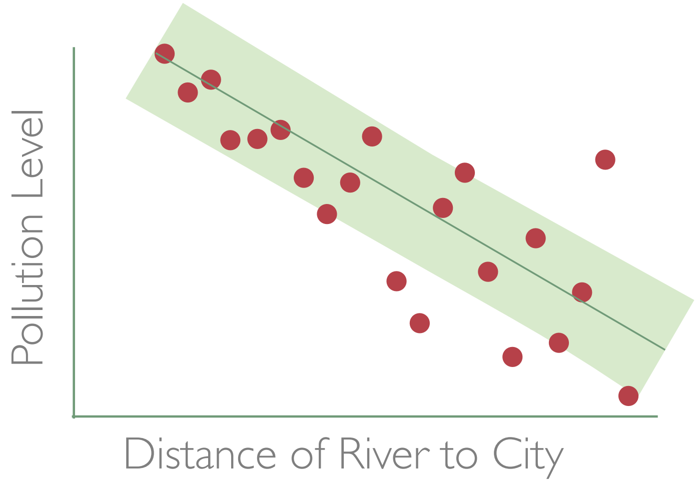
Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it Matter?

The model gives one number to estimate error about the central average. How does it get one number if the variance is different?

Or this?



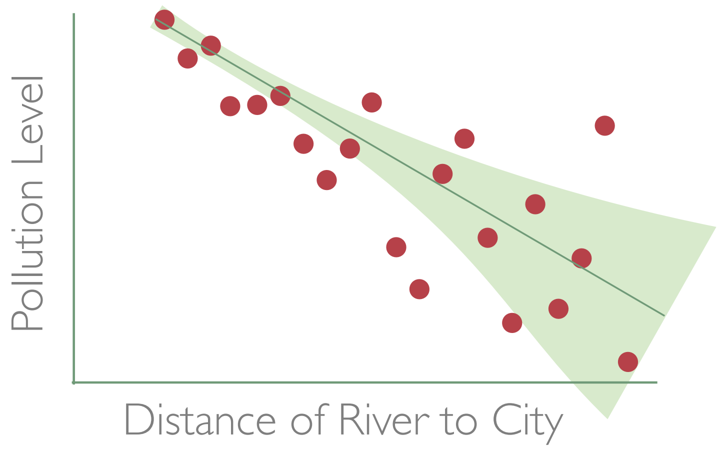
Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it Matter?

The model gives one number to estimate a plus mins error about the central average. How does it get one number if the variance is different?

Really its more this, but we can't have a plus minus value that varies



Assumptions

How to Check?

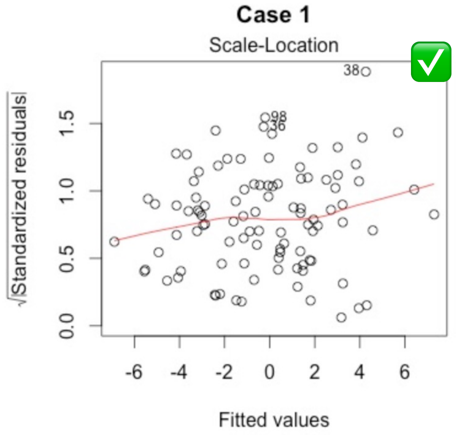
1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

How to Check?

The third of the diagnostic plots checks for this



Points are scattered evenly, line is roughly horizontal

Thankyou to U Virginina for visuals!! https://library.virginia.edu/data/articles/diagnostic-plots

Assumptions

1. Independence

2. Linear Relationship

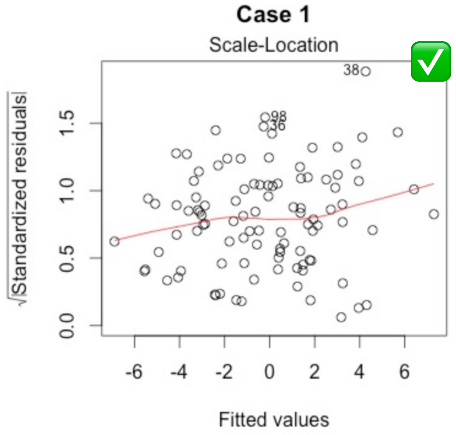
3. Normal Residuals

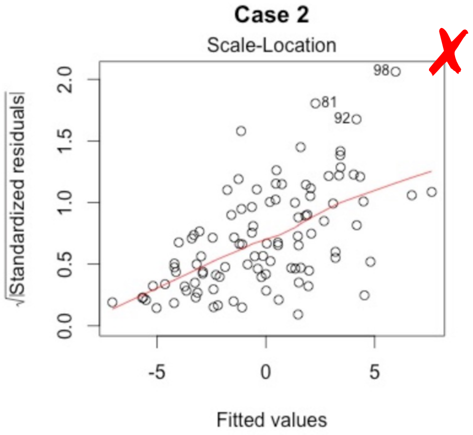
4. Constant Variance

5. No Major Outliers

How to Check?

The third of the diagnostic plots checks for this





There is a clear direction to the points, line has a trend (isn't flat)

Thankyou to U Virginina for visuals!! https://library.virginia.edu/data/articles/diagnostic-plots

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

What to do if you violate?

Violating this is less mission critical – your average estimate will still be okay (if you satisfy other assumptions).

But you can't trust your error estimates much.

(There are other things you can do, that we don't teach, but if you come across this problem in future and are referring back to these slides: try googling bootstrapping or robust standard errors)

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it matter?

What is the average of these numbers?

2 4 6 8 10

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it matter?

What is the average of these numbers?

2 4 6 8 10

It's 6.

What about:

2 4 6 8 10 24059

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it matter?

What is the average of these numbers?

2 4 6 8 10

It's 6.

What about:

2 4 6 8 10 24059

It's 4015. But is that representative?

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

Why does it matter?

Same logic, outliers can have a disproportionate impact on the result, which might not represent most points.

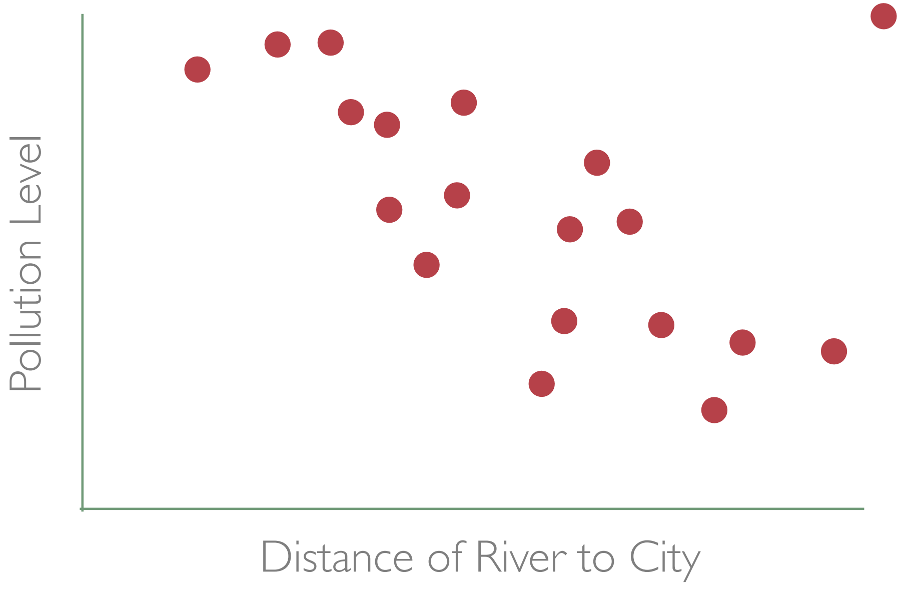
Thankyou to U Virginina for visuals!! https://library.virginia.edu/data/articles/diagnostic-plots

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

How to Check?

You can just look at your data

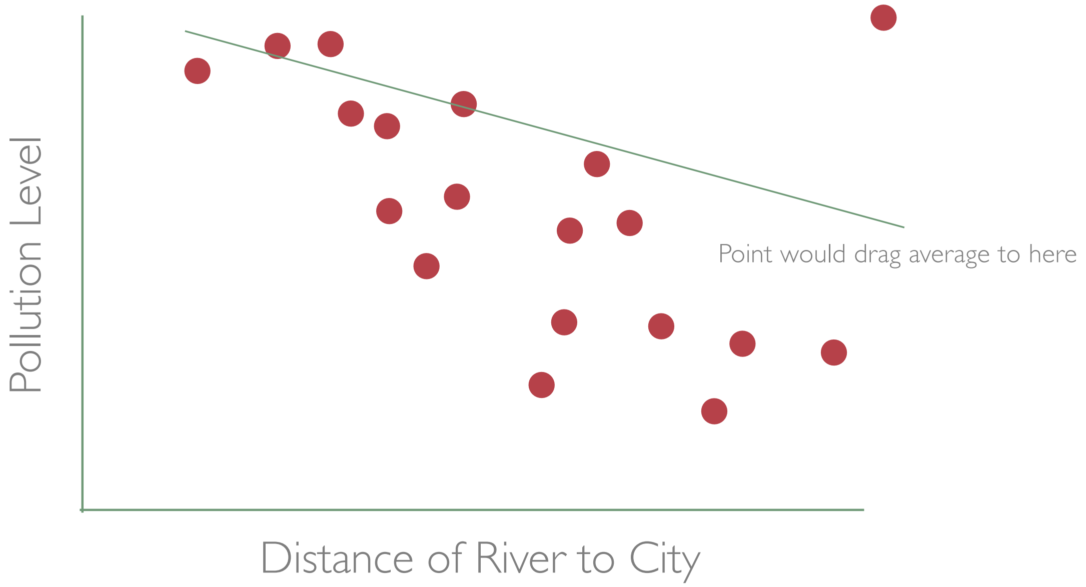


Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

How to Check?

You can just look at your data

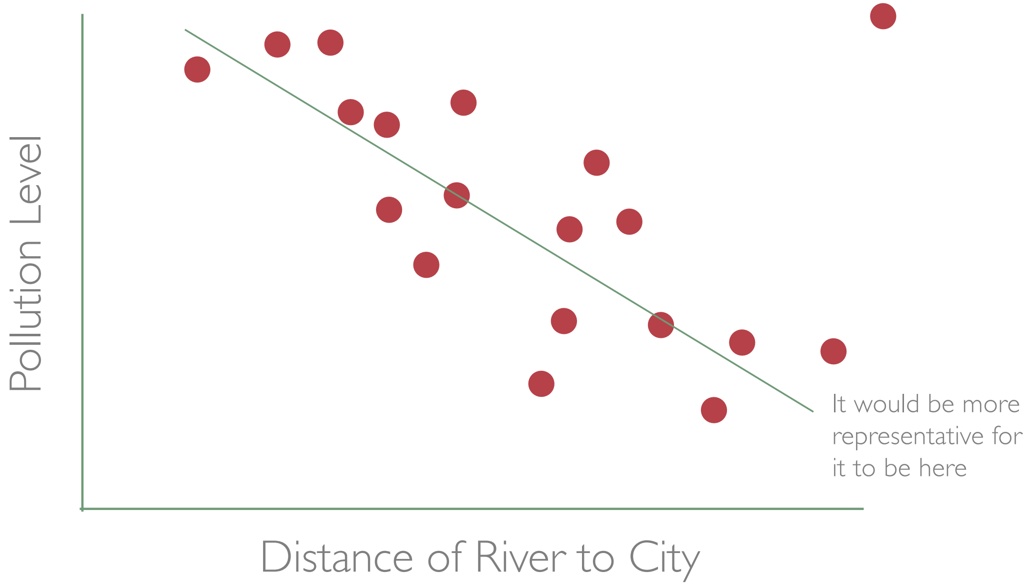


Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

How to Check?

You can just look at your data

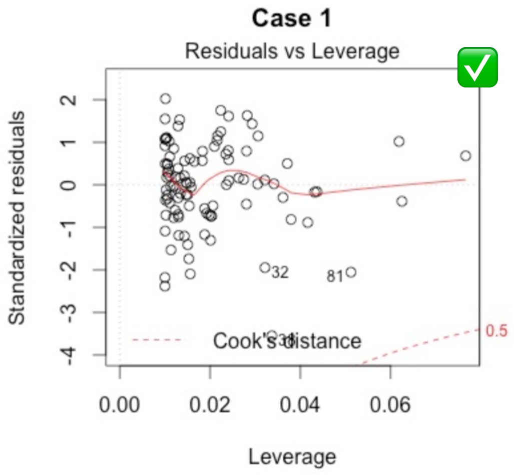


Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

How to Check?

OR The fourth of the diagnostic plots checks for this



Points are close-ish to each other, nothing is on the other side of a red dotted line

(sometimes you can't even see the line!)

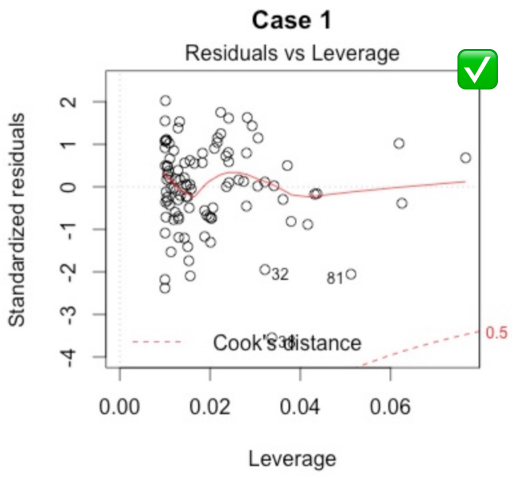
Thankyou to U Virginina for visuals!! https://library.virginia.edu/data/articles/diagnostic-plots

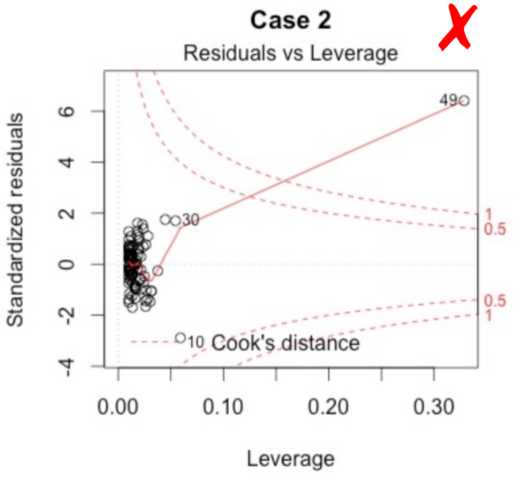
Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

How to Check?

The fourth of the diagnostic plots checks for this





One point waaay out, on the other side of the red dashes.

Thankyou to U Virginina for visuals!! https://library.virginia.edu/data/articles/diagnostic-plots

Assumptions

1. Independence
2. Linear Relationship
3. Normal Residuals
4. Constant Variance
5. No Major Outliers

What to do if you violate?

Do NOT just automatically remove the outlier! First check – is there some measurement error (or you typed something wrong in your spreadsheet or something). If so, fix.

But if the point seems genuine – consider what information its giving you!

Does it make sense to get an average without it? Is that fair?

Have A Breather…

Quiz 2

Divide into pairs.

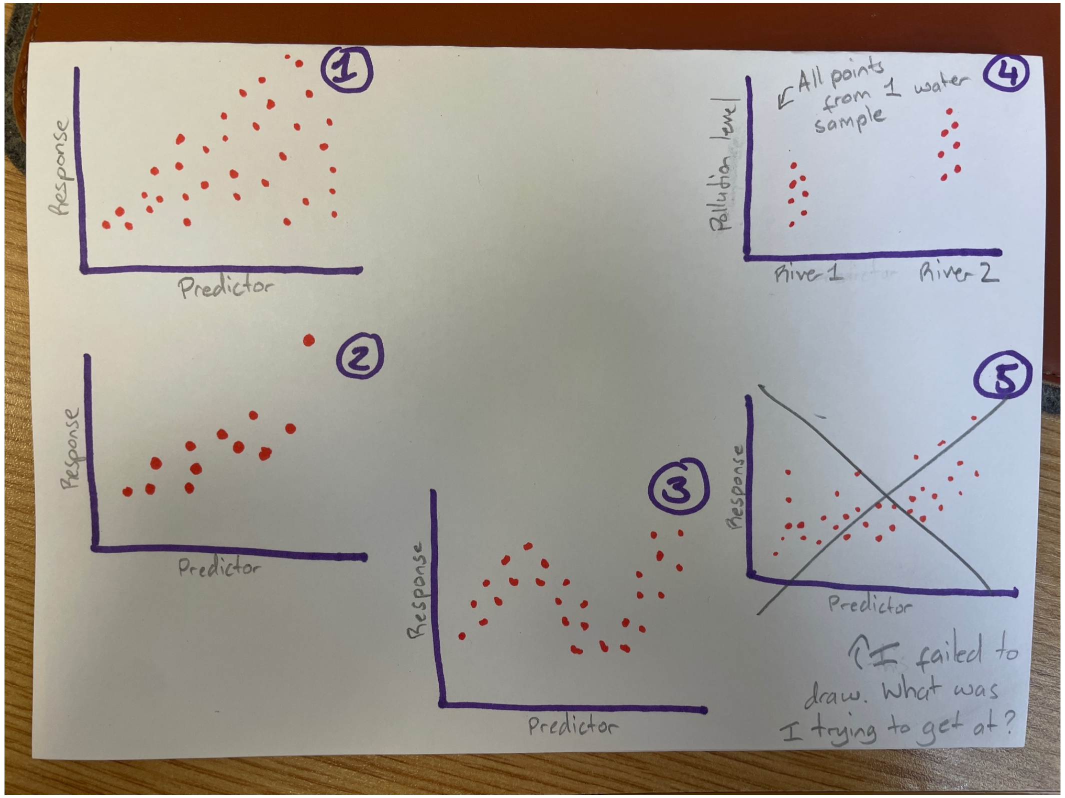
One person takes 1 & 2, the other 3 & 4. For each graph:

1. Copy it onto your paper, and try and add a trend line (or average estimate to it)
2. Figure out what assumption you think would be violated if we made a linear model of the data.
3. Explain to your partner which one it is AND why it would affect how we understand the model output.

Each assumption will affect our confidence in one of:

- How confident the model is in its representation of reality
- How accurate the overall maths of the model is
- How accurate the average estimate is
- How accurate the error estimate is
- How well the model represents the majority of data points

Which assumption isn't displayed here? Discuss it's meaning.



WWF Report (40% Grade)

You are a statistical consultant, and we (WWF) have hired you to put together a report from the Living Planet Database

We're interested in how the population trend of a species changes through time.

Take data from the Living Planet Database, that you've worked with the past few weeks

WWF Report - CRITICAL

When you click the link, it takes you to YOUR OWN ASSIGNMENT REPO. Do NOT fork it!! Just work directly off that.

REMEMBER to push your work to the repo from your R studio. Check the online repo before the deadline to check its updated!

You will work in the repo, and we will mark this, but we will also mark a PDF of your report. You MUST submit this to Learn by the deadline, in addition to having the github repo ready! Otherwise you WILL GET LATE PENALTIES, and we're not allowed to make exceptions!

WWF Report - Details

Everything is in the repo. Full details there, too long to go through in class. READ CAREFULLY.

In brief:

1. Pick a species from the database, and bags it in the assessment issue
2. Create a linear model of how your species in changing through time
3. Write an 800 report to WWF describing your model, its output, and what this means for the species
4. Plus, include a 200 word statement about your AI usage
5. The report should be in ‘R Markdown’ format within the repo. (See ‘Reporting with R Markdown’ Data Camp!)
6. Ensure the repo is well organized, including the code you used (well commented etc!)

WWF Report - Details

We teach content for this over the next four weeks. Do NOT start modelling this week, you need things from next week's class!

Do start picking your species, and playing around with wrangling the data to get familiar and get it ready.

Weeks 6 and 7 will cover everything you need to know to run a good model.

Week 8 will cover some things you should discuss in your report.

WWF Report - Marking

Model Construction (25%)

- You built and justified an appropriate model for your data
- (This includes 1) Generalised or not, 2) Mixed effects or not 3) Which predictors you chose, 4) Whether you included interaction terms)
- You tested assumptions and discuss them well

Model Interpretation (25%)

- You interpreted the model output for all predictors (sign, strength, standard error, significance) accurately and well
- You consider how likely it is that your model represents reality based on the data

WWF Report - Marking

Presentation (25%)

- You have figures and tables showing data and model output
- Report is well written
- Report is well and creatively formatted (e.g. pics of your species)

Reproducibility (25%)

- Repo is clearly organized
- Code runs without errors when we check it
- Data is cited in the report
- AI statement critically reflects on your usage

A bit on the AI Statement

You can use AI in this assignment.

USE IT WISELY. Think of Matúš talk last week!

We want 200 words reflecting on your process. How did you critique it? How did you ensure it wasn't giving your bogus advice?

What are your thoughts on how it affected your learning experience? Did it help or hinder?

Pleaaaaaase don't use it fully to write all your text. It makes text so boring to read, and often less accurate. Have a voice!

Due Friday 7th Nov, 12pm!

Everything explained in full in the repo!

Remember: Don’t start modelling this week!

But do pick a species and start having a play with the data!

Class Activity!