ECO 602
Analysis of Environmental Data

FALL 2019 – UNIVERSITY OF MASSACHUSETTS

DR. MICHAEL NELSON

McGarigal Chapter 3

Very dense chapter, we'll cover the material in approximately 3 lectures:

1. sections 1 – 4
2. **sections 4 – 6**
3. Sections 7 - 9

Today's Agenda

1. More single variable plots
2. Associations
3. Interpreting figures in real time
4. Question set consulting time

Quantile-quantile plots and empirical distribution plots

Useful for interpreting normality.

They make more sense in the context of inference and model diagnostics.

We'll come back to these later in the course

Associations: describing relationships

What's a relationship?

What are some ways to describe relationships?

Use your model thinking to describe a relationship beyond a statistical framework.

Associations

Association is a neutral term.

Unlike correlation, it doesn't imply causality or mechanistic connections between two variables.

It doesn't imply a specific kind of relationship.

Correlation and covariance refer to **linear** relationships, association is broader.

Associations: describing relationships

We will focus on graphical and numerical descriptions of associations in this course, but it is important to remember that the concept is much broader.

Let's consider how to quantitatively describe associations:

- Covariance/correlation
- Shape of the relationship: linear, polynomial, others?

But first: sampling units and variables

- A sampling unit is an entity of interest
- A sampling unit may have **one or more** attributes of interest that we might like to measure.
- These attributes are the variables.
- Sampling units may have **many variables**.
- We'll use the term **sample** to refer to the group of sampling units that are available to us for measurement.

Variance: measure of spread

Describes a **single** variable, i.e. one attribute of a sampling unit.

Calculated from all the values in the sample.

The mean squared difference between each SU's value and the sample mean.

Always positive (because of the squared term).

Units are squared, so we don't try to interpret variance directly.

Covariance: like variance but different

A measure of spread, kind of like variance

Describes the strength of a **linear** relationship between two variables.

Because it involves the product of two variables, the units can be very weird. We don't usually interpret them directly.

Pearson's correlation coefficient, a normalized version of the covariance, is more intuitive.

Covariance: like variance but different

- There are no squared terms, so covariance can be positive or negative (or 0).
- The sign tells us whether the association is positive or negative.
- The magnitude of the absolute value gives us information about the spread, but a direct interpretation is difficult.

Covariance: a walkthrough of the calculation

It’s not necessary to memorize the calculation, but it’s helpful to have an idea of how it works.

Consider a sampling unit with two variables: x and y.

First, calculate the mean values of x and y in the whole sample.

These are usually denoted with bars: $\bar{x}$, $\bar{y}$.

Covariance: a walkthrough of the calculation

Next, for each sampling unit, calculate the difference between its value of x and the sample mean $\bar{x}$.

Do the same for the sampling unit's value of y and the $\bar{y}$.

Multiply these two differences.

Do this for each sampling unit and calculate the sum.

Covariance is the mean value of these products of differences.

You can look this up if you ever need to know the details.

Covariance: intuition

The degree to which values of x and y within sampling units are consistently:

1. Both high: positive covariance/correlation
2. One high, one low: negative covariance/correlation
3. Totally unrelated: zero covariance/correlation.

This is why the two variables of the sampling units are paired for the covariance calculations.

Correlation

Pearson's correlation:

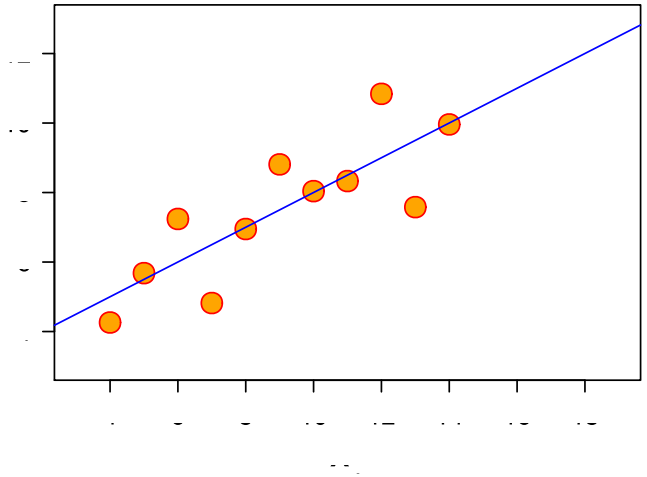
- A normalized version of covariance: it is always between -1 and 1.
- Is appropriate when there is a linear relationship between two variables.

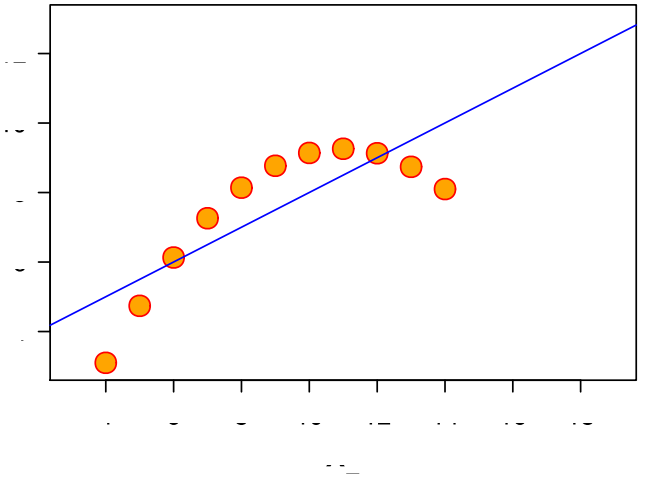
Spearman's rank correlation:

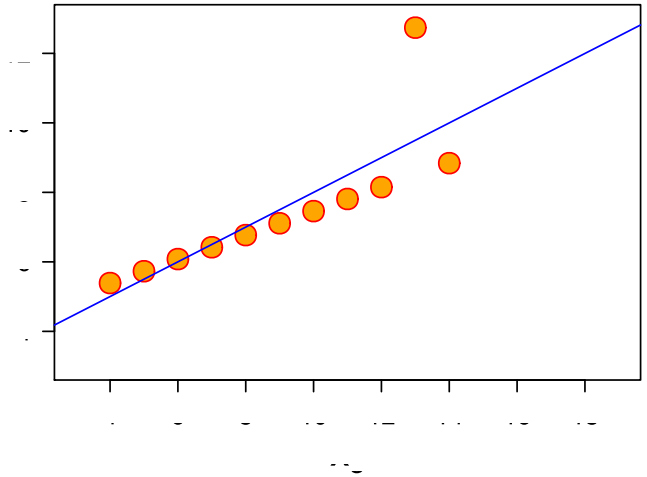
- Analogous to Pearson's correlation, but is appropriate when the relationship is monotonic (but not linear)

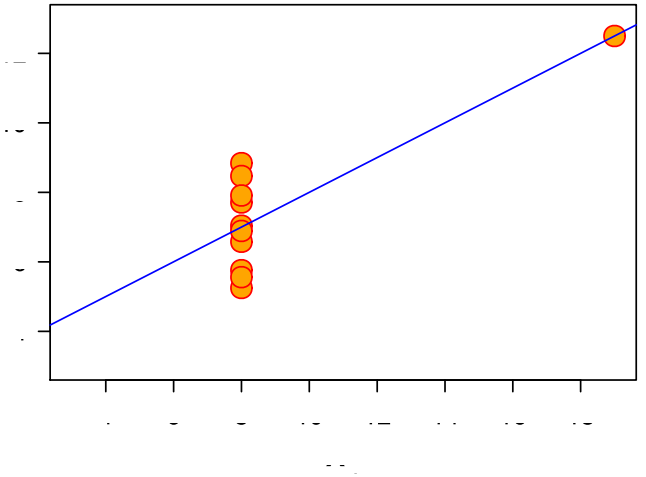
Anscombe's Quartet

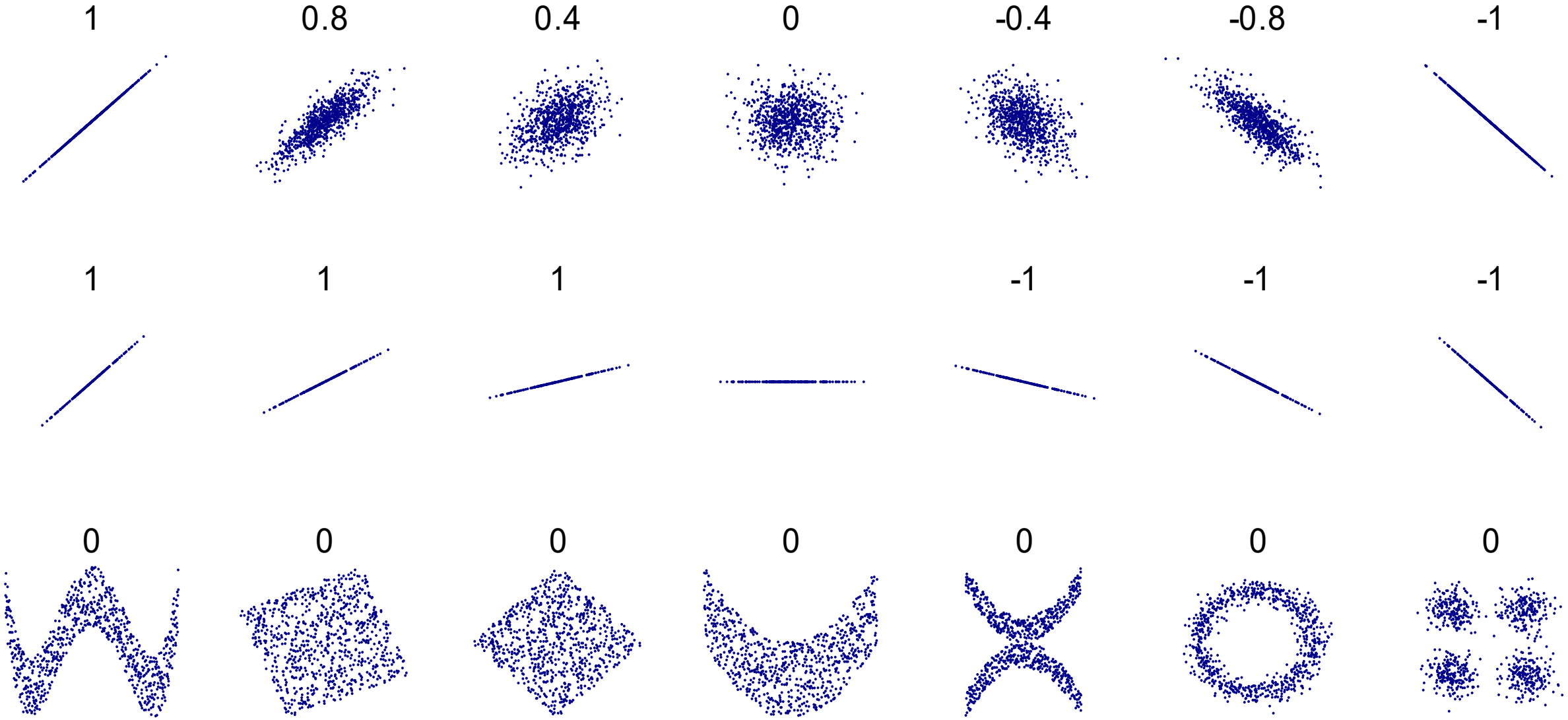
Data Set	1	2	3	4
Mean of x	9.000	9.000	9.000	9.000
Mean of y	7.500	7.500	7.500	7.500
Sample SD of x	3.317	3.317	3.317	3.317
Sample SD of y	2.031	2.031	2.030	2.030
Cor(x, y)	0.816	0.816	0.816	0.816











Public domain image: https://en.wikipedia.org/wiki/Correlation_and_dependence#/media/File:Correlation_examples2.svg

Associations: correlation and causation

A nonrandom association between two variables is not evidence for a causal relationship, but it can lend support to a causal mechanistic model.

Lurking variables can jointly influence predictor and response.

Evidence for causality is usually much harder to support than evidence of an association.

Controlled experiments can help

Scatterplots

Usually straightforward to understand

One or both axes may be on a log scale

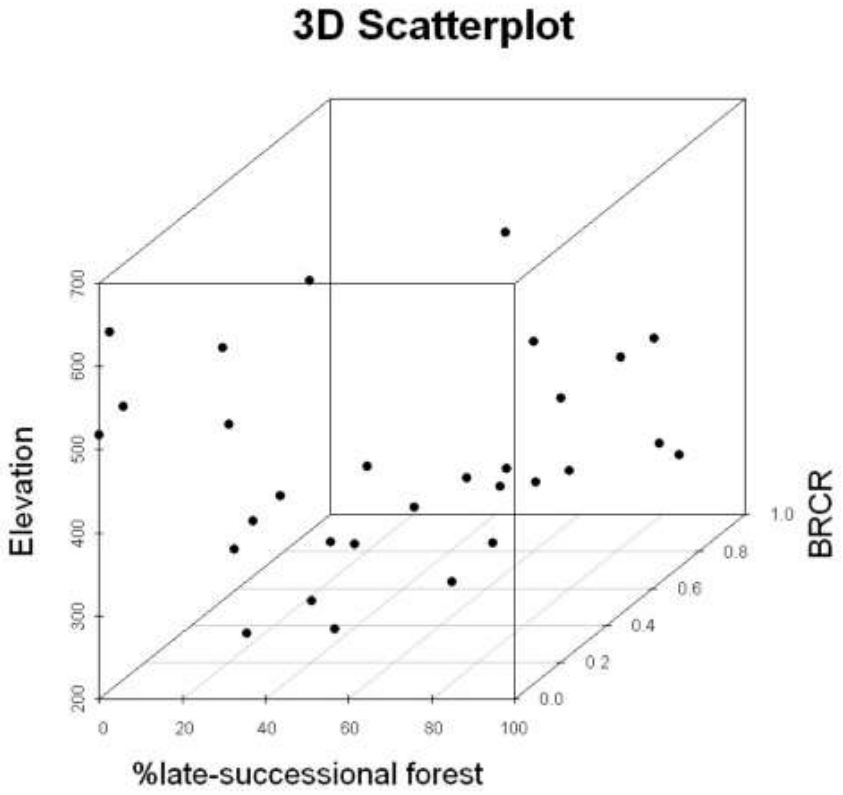
- This usually happens with very skewed data that cover a very large range of magnitudes.
- More when we discuss transformations

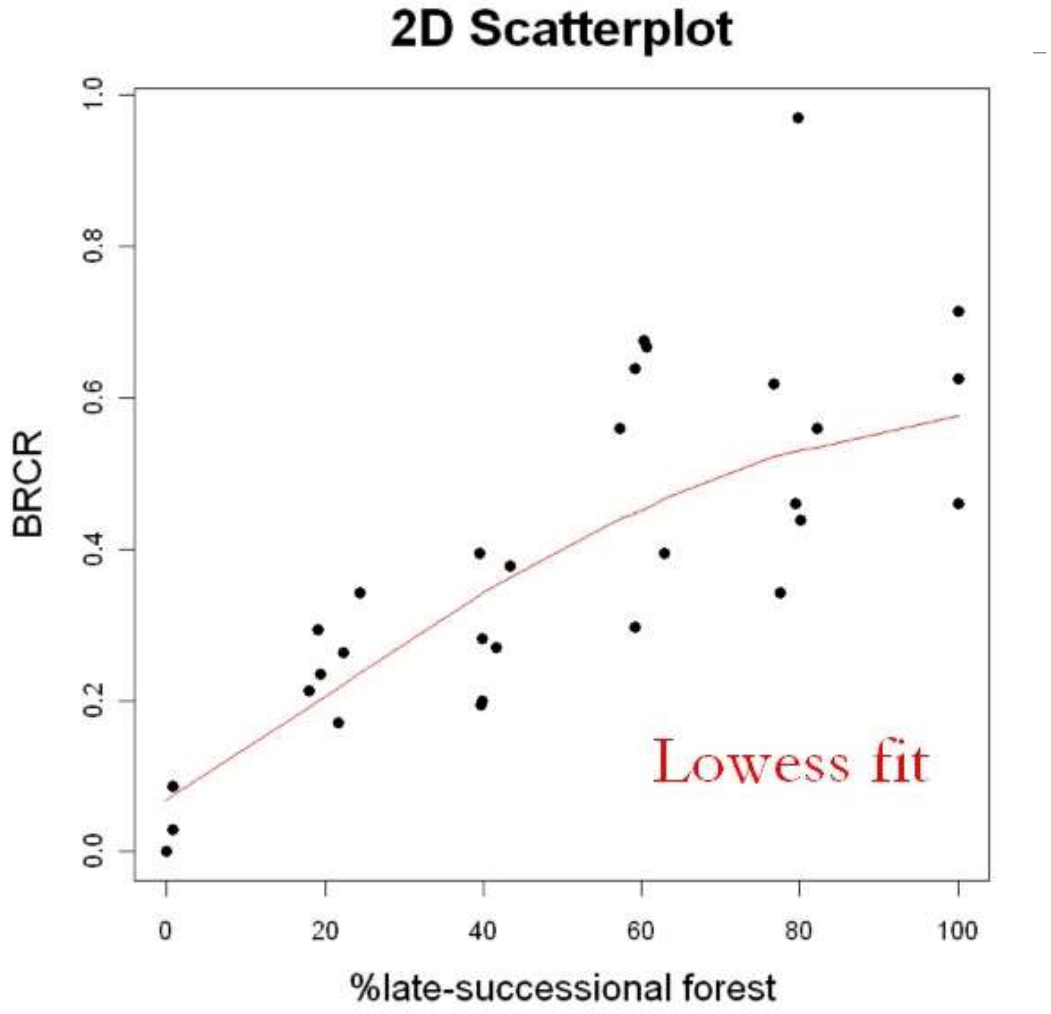
By convention:

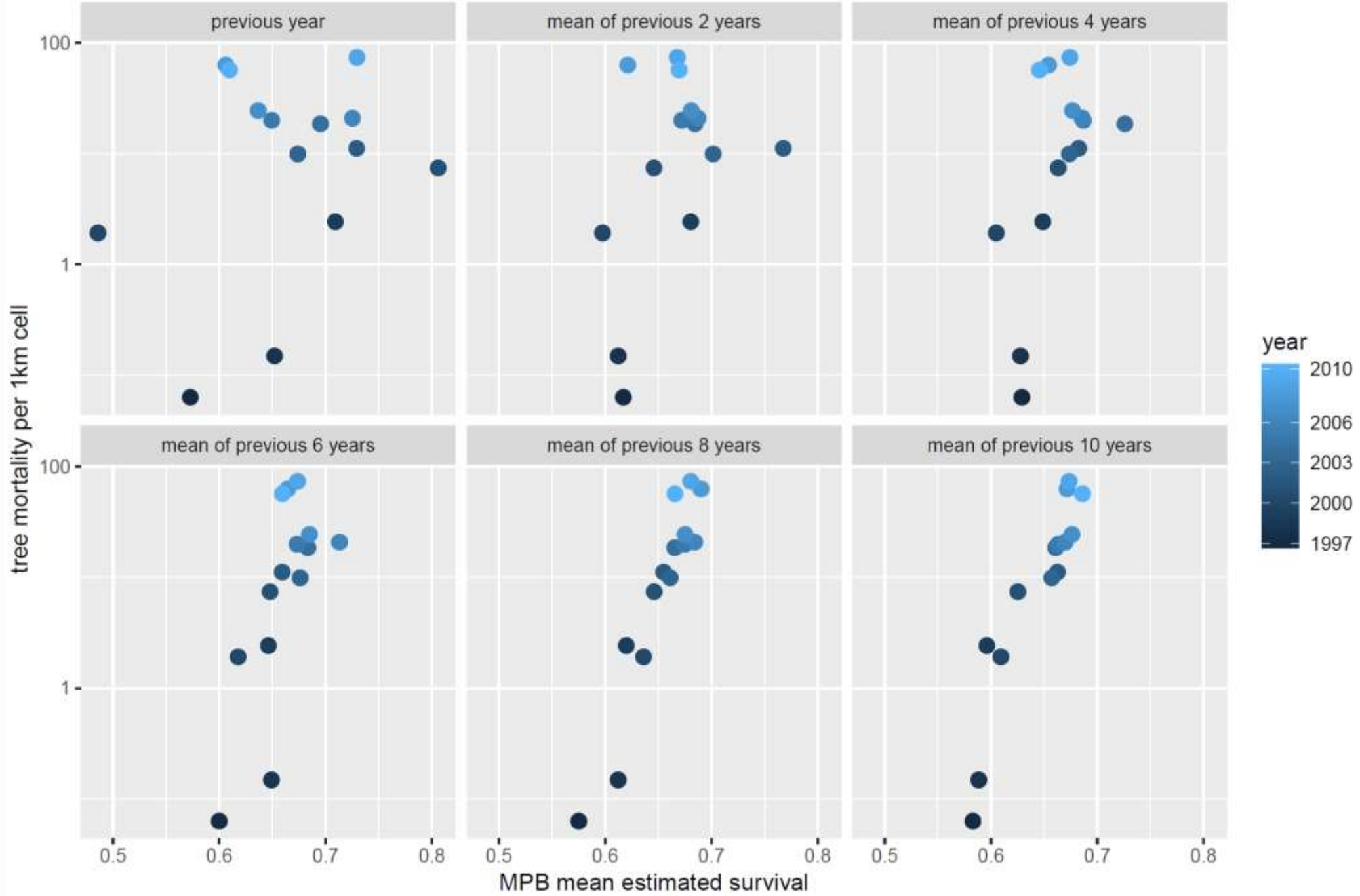
- X axis is the predictor/independent variable
- Y axis is the response/dependent variable

Scatterplot

- Graphical display of two (or more) variables






previous year
mean of previous 2 years
mean of previous 4 years
mean of previous 6 years
mean of previous 8 years
mean of previous 10 years
tree mortality per 1km cell
100
1
0.5
0.6
0.7
0.8
MPB mean estimated survival
year
2010
2006
2003
2000
1997

Interpreting plots in real time

It’s a great feeling when you produce a nice-looking, easy to read plot of your data.

You have a dataset that you have spent a lot of time with, maybe you even collected the data yourself.

You’ve done lots of exploratory data analysis, graphical and numerical, and you have a good intuitive feel for each of the variables.

Interpreting plots in real time

You build a wonderful plot to highlight some aspect of your data, perhaps an association.

You're super excited that you get to show your figure during at your upcoming lab group meeting.

Interpreting plots in real time

You advance to the slide with your figure, and you see lots of confused, blank stares.

Maybe even some frowny faces.

Not the reaction you had hoped for…

Interpreting plots in real time

Consider the audience's point of view:

Quickly making sense of a figure without knowing the system well is difficult.

It's something you will have to do often.

You are learning skills to help.

Let's experience it now, together!

Figures exercise: individual

You'll have 5 minutes to scan the paper individually:

1. Read the paper's abstract
2. Look at each figure and read the caption.
3. Try to get the general idea of the figure's meaning.
4. It's ok if you don't understand it at all, move on to the next one.

Keep in mind: uncertainty is everywhere

“I don’t know” is a perfectly valid, and honest answer.

But… as a scientist you must be able to articulate:

Why you didn’t know?

How could you find an answer?

Figures exercise:

For the paper: context, questions, null/alternative hypotheses, conclusions

For each figure:

- What are the variables, data types, axes, scales?
- What is the general story that the figure tells?
- How might the figure support or refute a null hypothesis?

Single variable plots summary

Histogram: skewness and symmetry, assessing normality, sensitive to bin size

Boxplot: spread, extreme values, not a great depiction of normality

Empirical probability plots: comparing data to a probability distribution

QQ plots: assessing normality, useful for model diagnostics (more on this later)

For next time:

Keep working through McGarigal chapter 3

Don’t get stuck on correlation matrices, scatterplot matrices, or coplots. The text is especially dense on these, and we’ll cover them next lecture.

Please send me drafts of your assignment 1 if you want feedback!

Additional office hour: Mondays 1:00 – 2:00