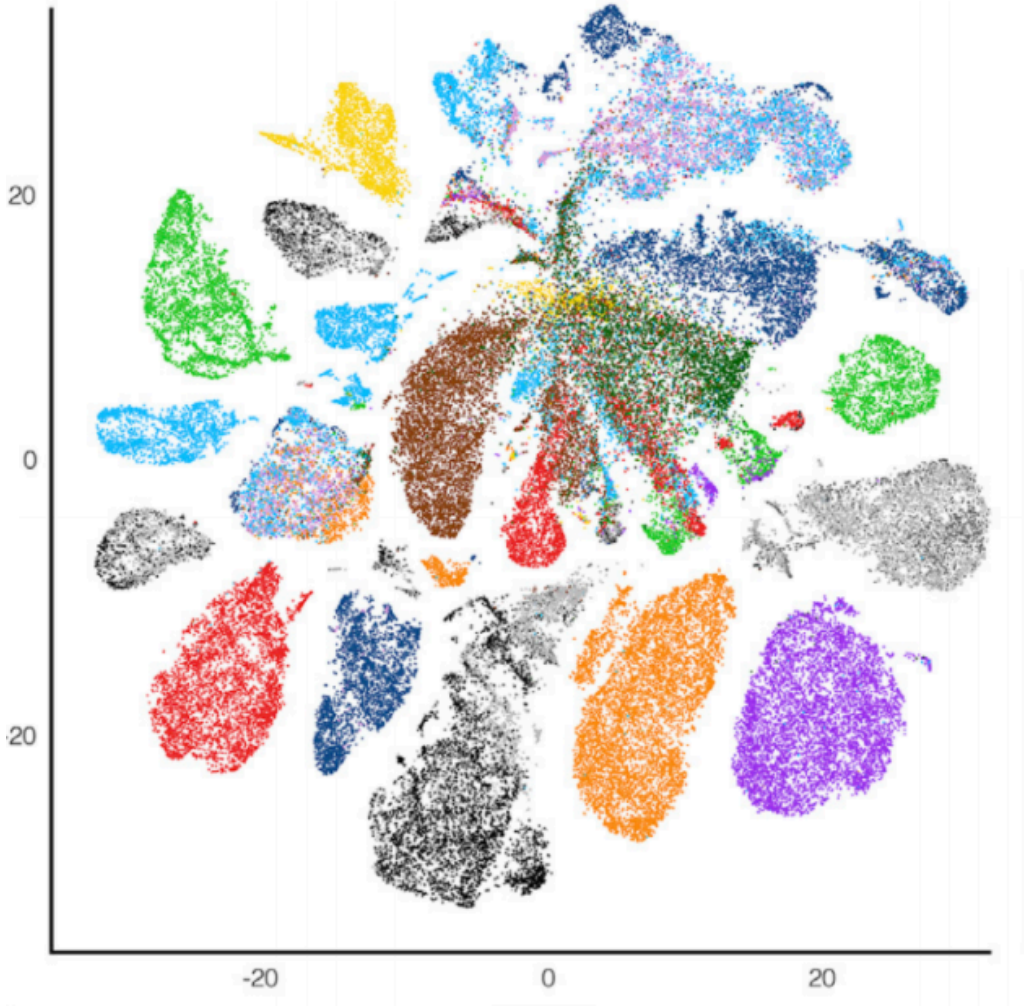
Clustering Methods in R

What is clustering? And why do it?

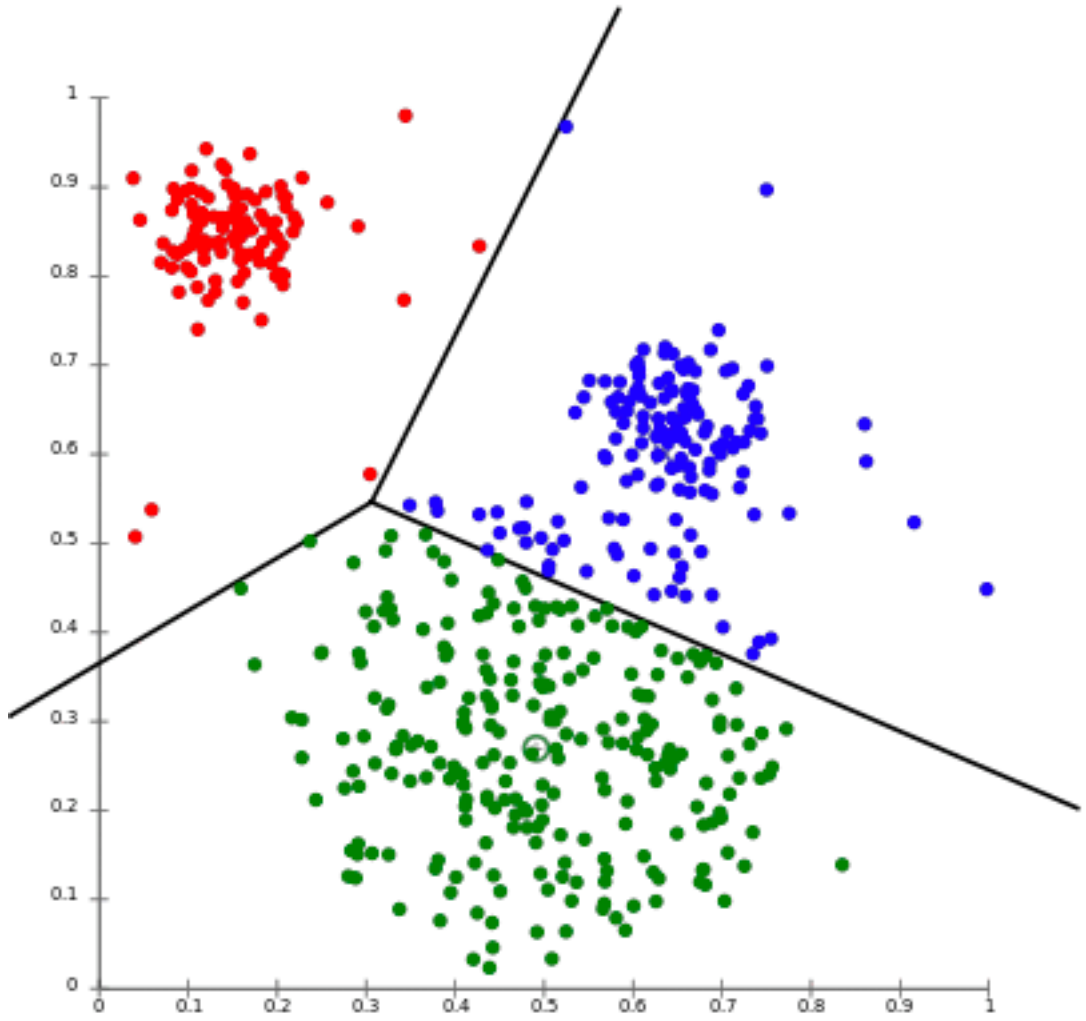
- clustering – grouping objects together into clusters by similarity
- Many, many different algorithms that try to solve this problem
- Clustering can identify patterns of variation in the data
  - Unwanted clustering like batch effect
  - Wanted clustering like by condition or to identify cell type in single cell sequencing



K-means Clustering

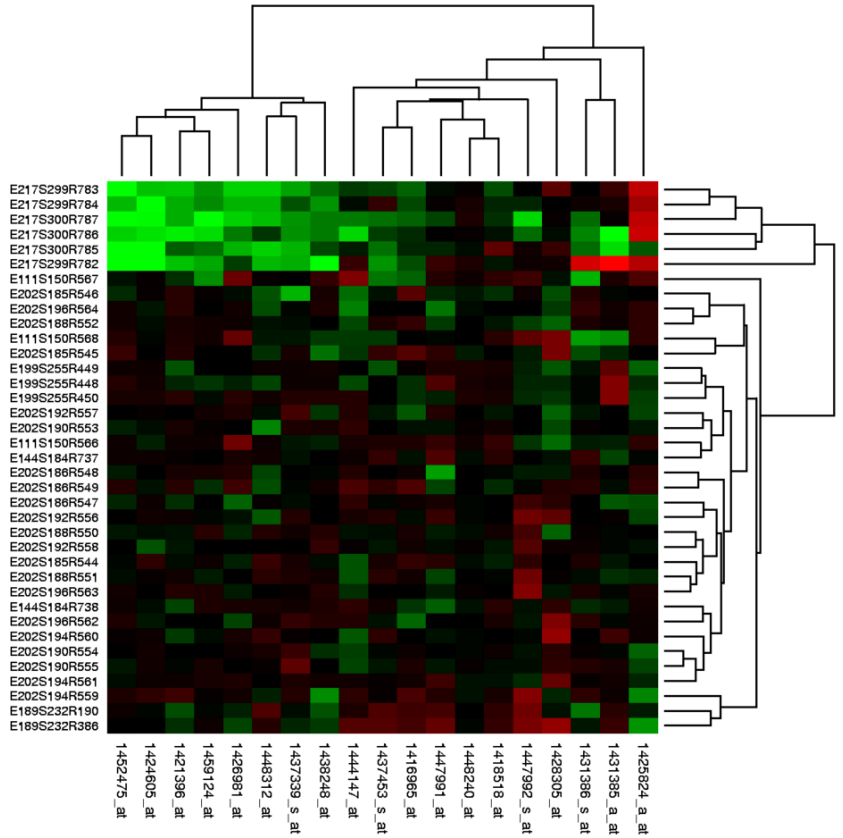
1. Pick number of clusters
2. Randomly pick a point to be the center of each cluster (centroid)
3. Assign all the datapoints to a cluster based on the nearest centroid.
4. Move the centroids
5. Repeat steps 3 and 4 until no points change clusters

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/



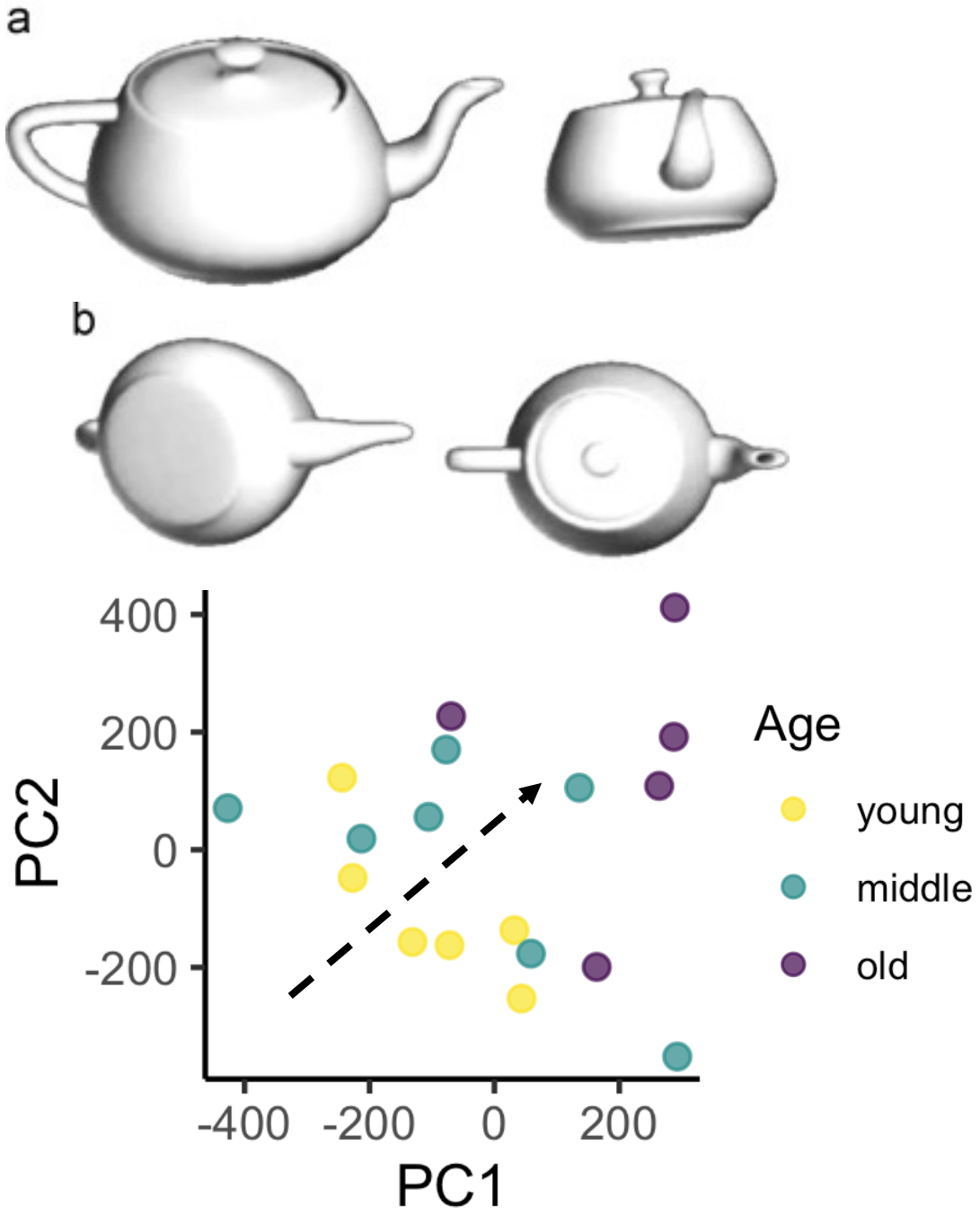
Heatmap

- A **heatmap** shows magnitude of some data as color in two dimensions (one color for lowest values transitioning to the other color for the highest values); cells are clustered by some algorithm and the dendrogram on the top and/or sides shows the relationships
- To calculate the clustering, first a measure of similarity is calculated, then a clustering algorithm is applied to the similarity scores.
- Many different algorithms can be used to calculate the clustering



PCA

- Like calculating a line of best fit, but with more than 2 dimensions
- To calculate a PCA
  - Data can be normalized by scaling and centering it (z scoring). Whether you do this affects the final outcome
  - Do some linear algebra to calculate principal components
- PC1 always explains the most variation, then PC2, then PC3 etc.
- Dimensionality reduction technique; goal is to retain most of the important information while simplifying the data

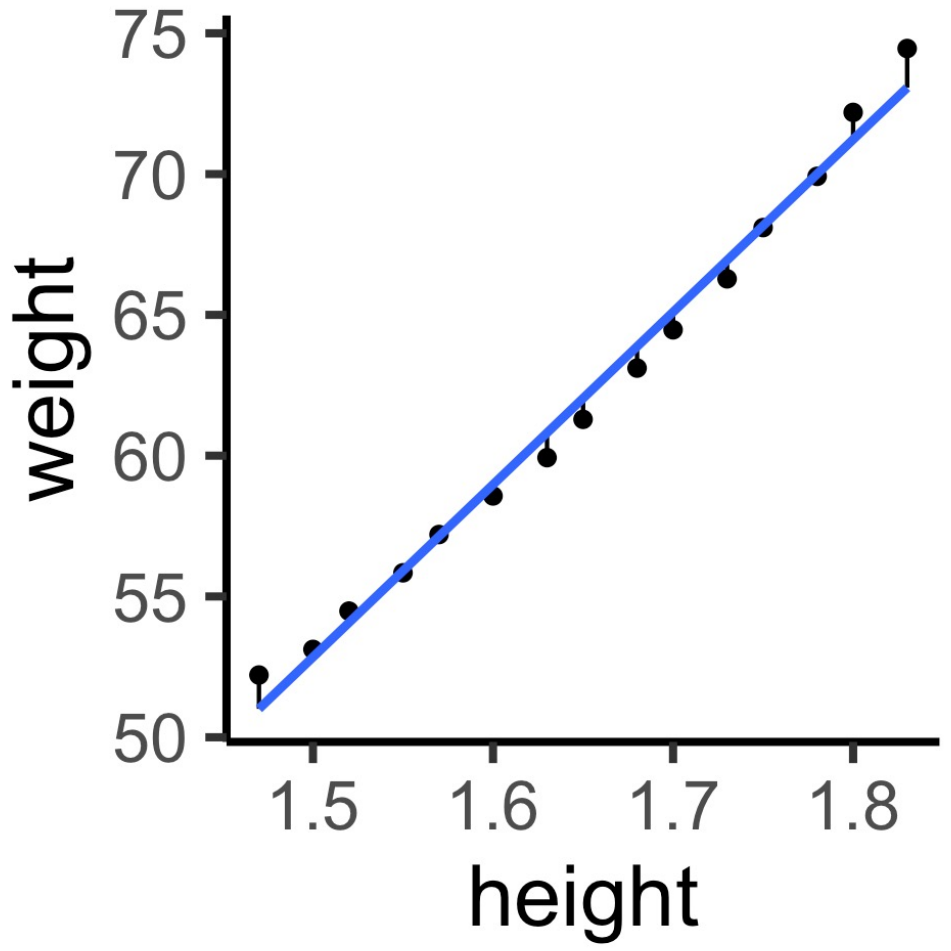


DEMO

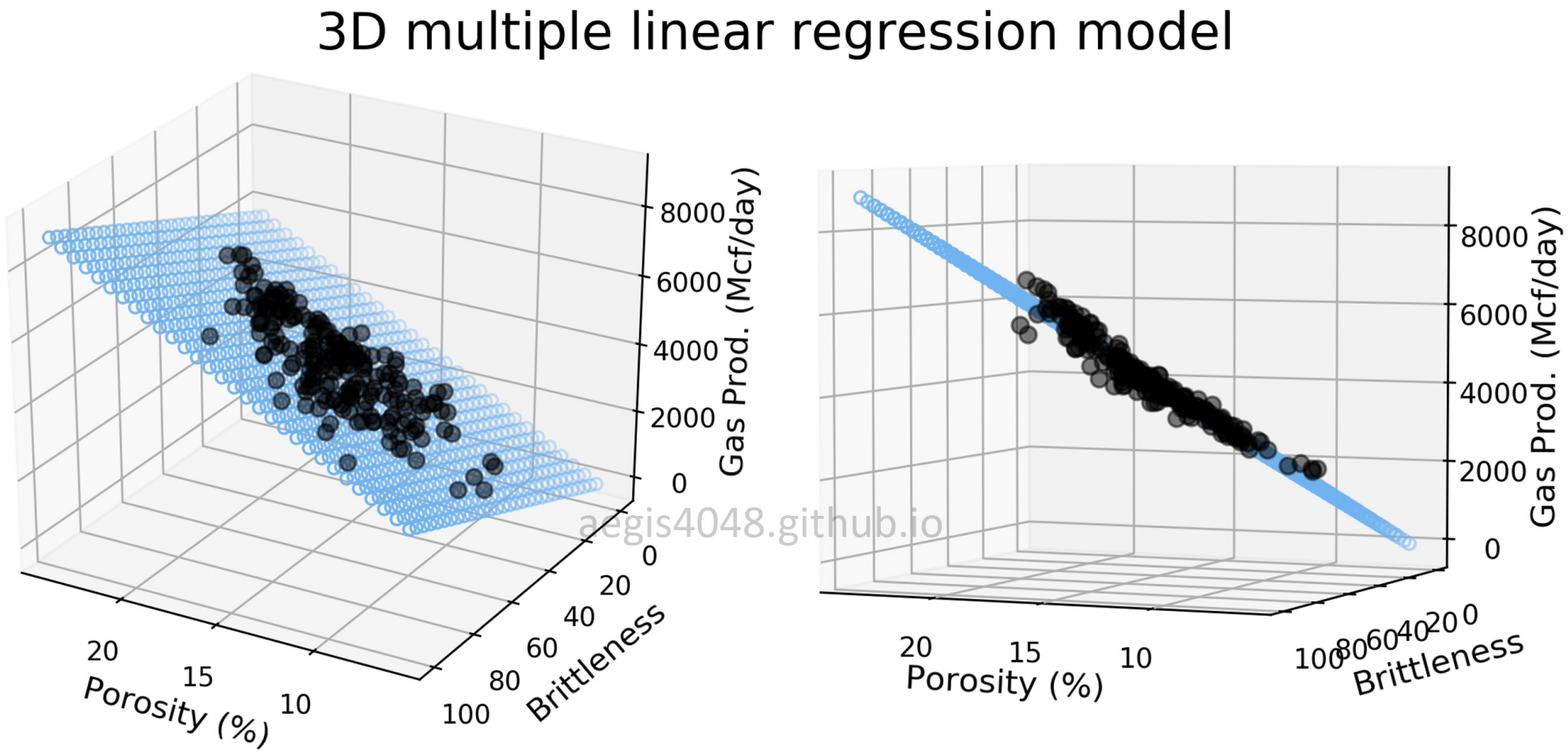
Statistics 2: More Advanced Linear Models

Calculating, Interpreting, and Extending Linear Models

- Linear models minimize the squared distance between the points of two variables
- Interpretation
  - For every 1 additional meter in height, weight increases by 61.27 kg
  - The y-intercept is often meaningless, especially in biological data. Here's it's when you weigh 0 kg, your height is -39.06 meters
- The additional value here, whether than just asking if these variables are significantly related is that you can use new measurements of X to predict what Y will be

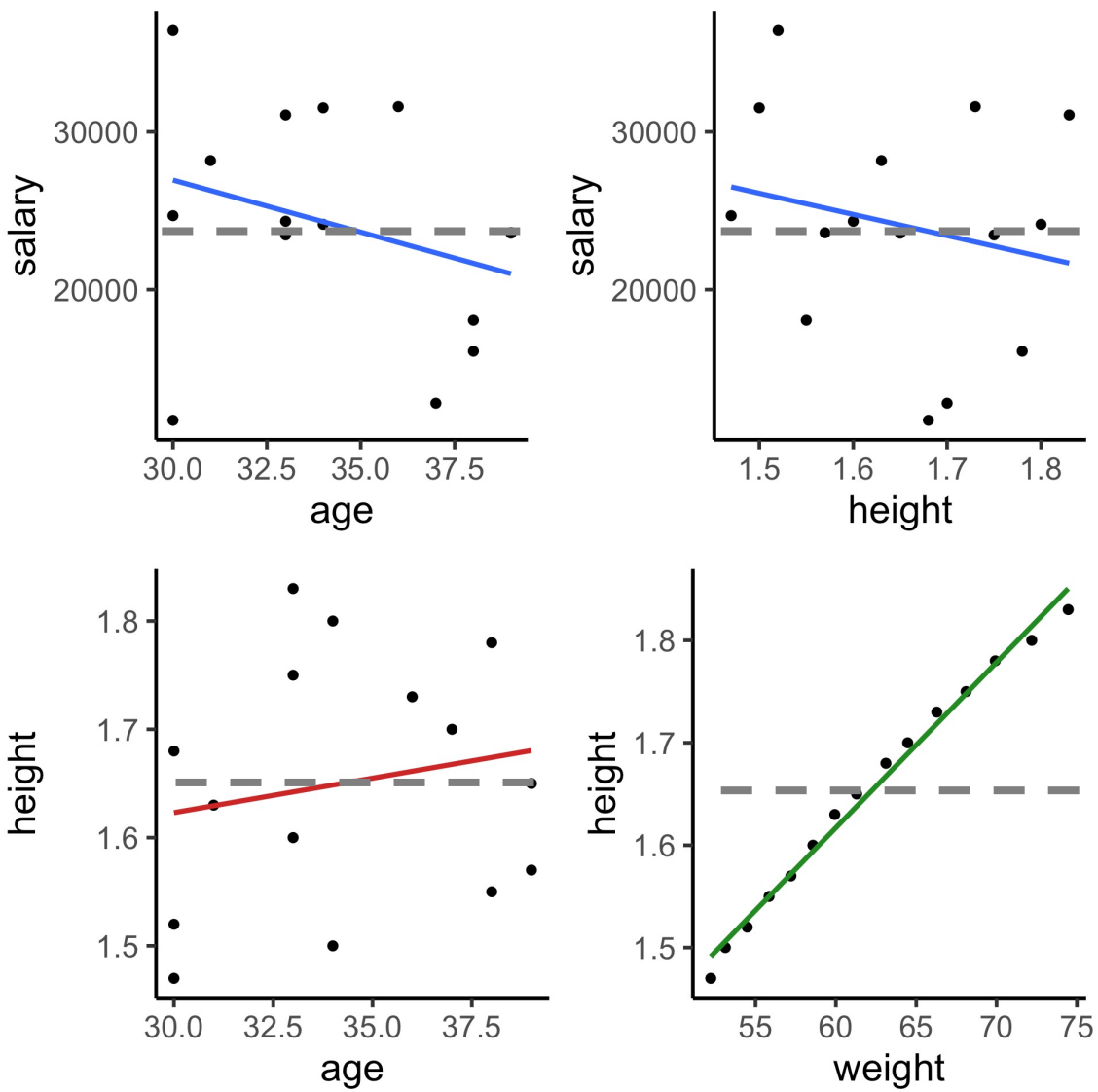


How is a linear model calculated for more than one variable?



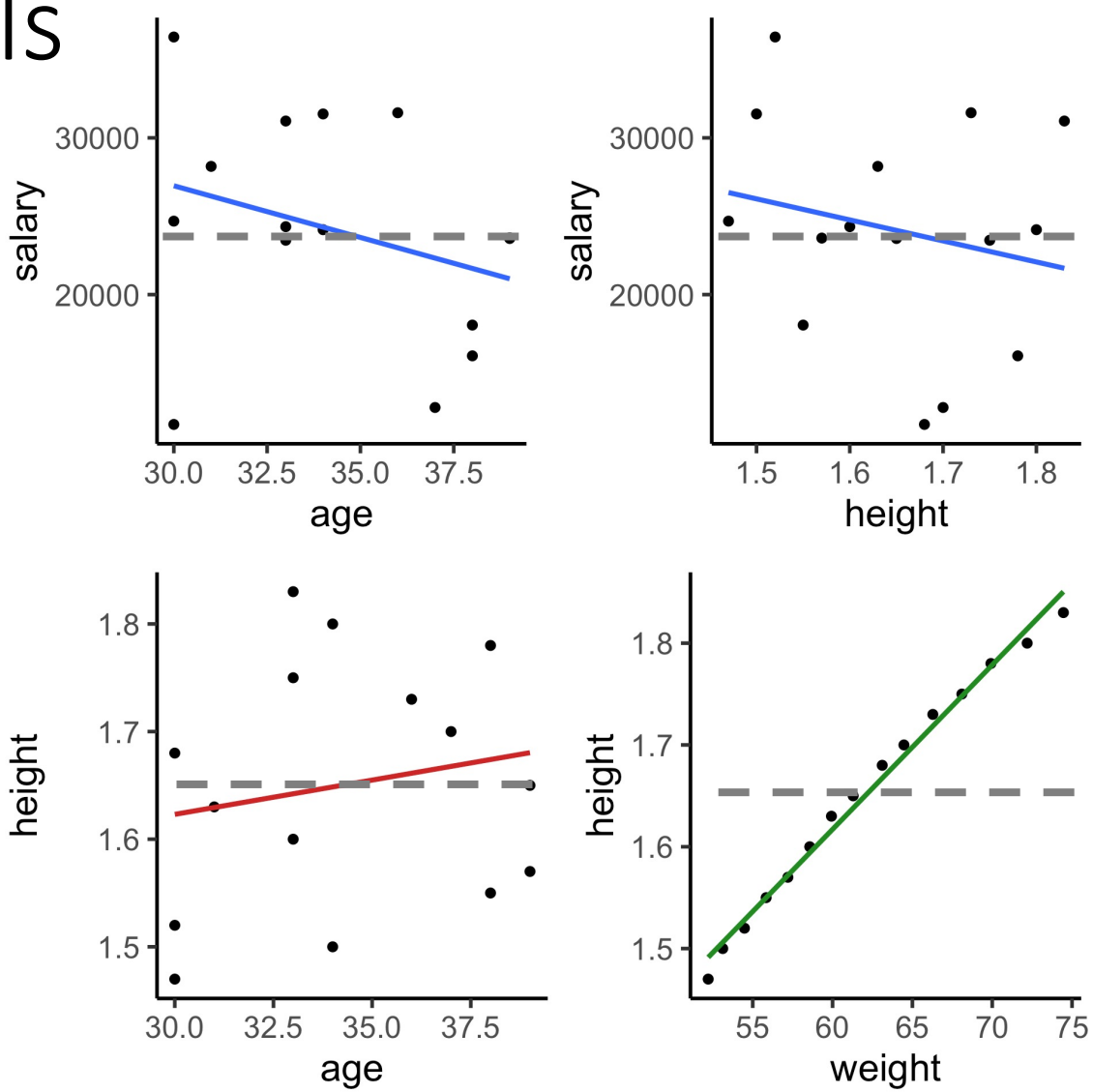
Additive Linear Models

- These are for when you have no relationship between your variables/when your variables are totally independent
- By using multiple independent variables that each explain some variation in Y, you can get a better prediction
- This is the most common type of multiple regression and what you should use by default



Interaction Linear Models

- You would use an interaction model when the values of your X variables depend on each other
- For example, height and weight. Tall people must weigh more, and short people must weigh less
- You may not know whether your variables are independent, so you will sometimes need to test both additive and interaction models



How do you pick the best model?

R^2

- R^2 explains the goodness-of-fit of a model
- It represents the percent variation the X variable explains in the Y variable
- Good models are close to 1 or -1, while bad models are close to 0
- Importantly it increases with each additional variable
- For example, for height vs weight, R^2 = 0.98, so weight explains 98% of the variation in height

Akaike Information Criterion (AIC)

- Also quantifies the goodness-of-fit of a model
- AIC selects for the model that explains the greatest amount of variation with the fewest variables (attempts to control for adding in additional variables)
- The lower the AIC the better
- For example, AIC for height vs weight is -84.47 but for salary vs age it's 312.22
- If AICs for models are similar, take the model with fewer variables

DEMO

Exploratory Data Analysis

Exploratory Data Analysis