**Linear Regression**

Chart, scatter chart

Description automatically generated

Linear Regression is a linear approach for modelling the relationship between a scalar response (Y) and one or more explanatory variables (X). In the case of a single explanatory varable it is called simple linear regression, for more than one the process is called multiple linear regression.

How to find the regression line

Chart

Description automatically generated

Linear Regression minimizes the sum of squared differences between the predictions and the actual values. We want to find the global minimum of this term.

In other words, we are looking for a scalar β (which multiplies the predictor variables X) that minimizes this term by finding its global minimum.

As we can see in the graph, as β goes from 1 to 2.5, it hits a global minimum at 1.6 (important to note that the curve is differentiable everywhere and has a single minimum)

Chart, line chart, scatter chart

Description automatically generatedLet’s find the perfect fit line for this toy example.

Does this black line perfectly fit the data? No, for example, for people with an IQ of 70 we would predict an income of 40k, but the majority of datapoints with IQ 70 are above 40k.

In order to find the perfect fit line we have to use the mean line (dotted blue line) and from its center, go 1 standard deviation across x-axis and r standard deviations across y-axis, creating the red line. The red line is the best fit line that minimizes the squared residuals.

Chart, scatter chart

Description automatically generated

The black line would be the best prediction if the correlation was 1 (if there was no noise), but if there is any variability, the real regression line tilts towards the mean line (**Regression to the mean effect**)

The blue line would be the best prediction if the correlation were 0.

Chart, scatter chart

Description automatically generatedThe magnitude of the regression to the mean depends on the correlation strength.

Never stop at linear regression, there are always other confounding factors.

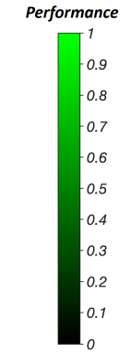
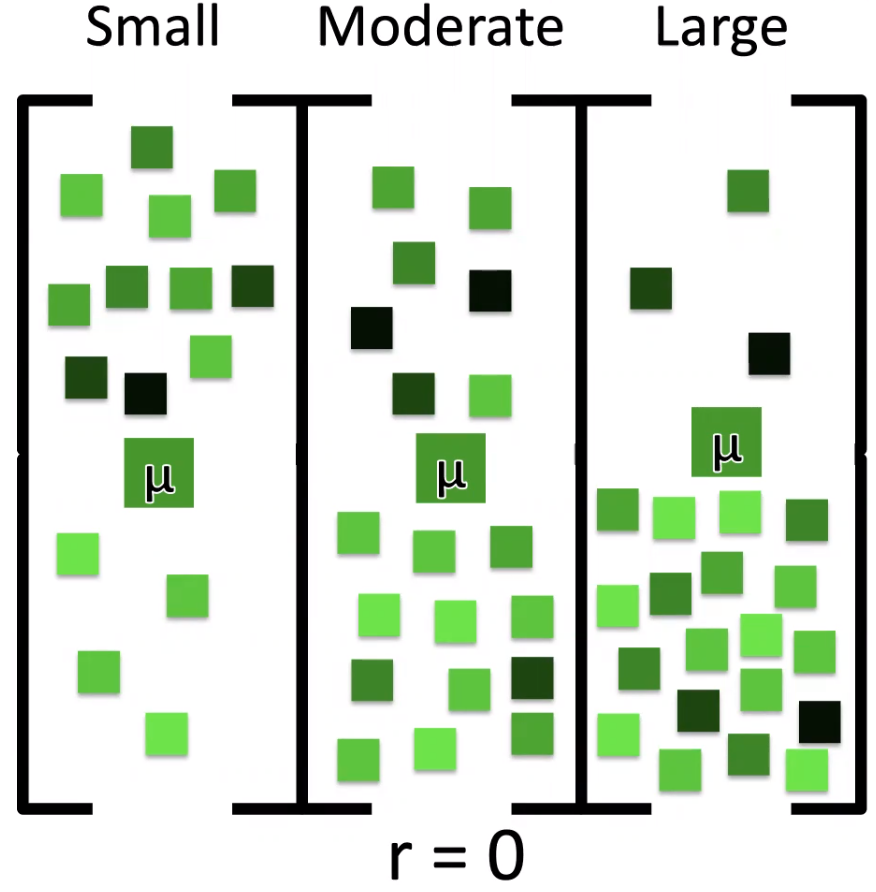
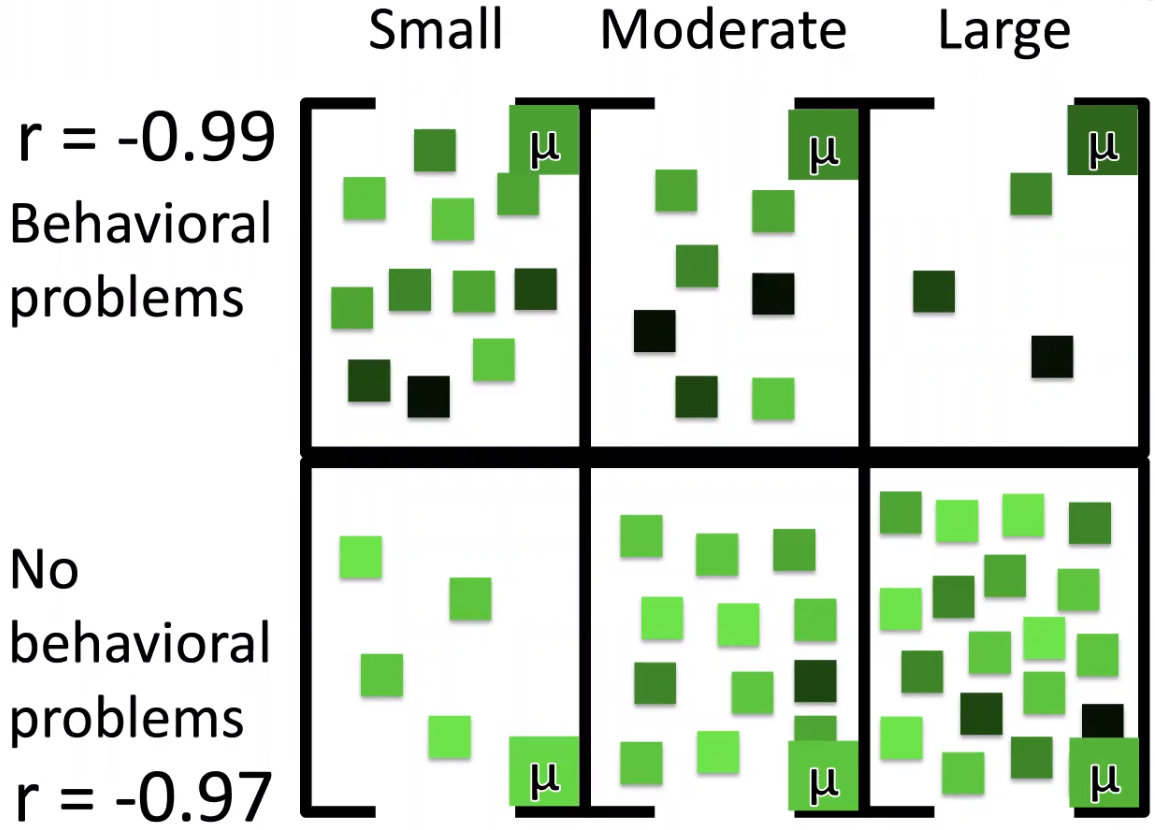
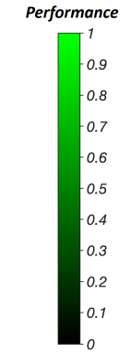
Let’s take a look at an example. We work for the Deparment of Eduaction and want to know whether classroom size is related to student learning outcomes. It is not appropiate to randomly assign students to different classroom sizes (parents are not happy with that), so let’s record data from all schools in a county of NY state and note class size and learning outcomes.

We perform Linear Regression and come to the conclusion that there is no relationship between class size and performance at all. This is counter-intuitive, how is going on?

Schools might be assigning their best teachers to larger classes to mitigate the effect that those large classes have on students’ performance. We should take this into account by doing Multiple Linear Regression.

If we take a look at the division on classes by size, we can see that there is a high variety of perfrmance levels in each division, which yields a similar mean for all of them. (Left Figure).

Now, if we subdivide them into students with and without behavioral problems, we can see that in bith rows there is a very strong (and negative) effect of classroom size (Right Figure)



Calendar

Description automatically generatedCalendar

Description automatically generated with medium confidenceDiagram

Description automatically generated with medium confidenceClassroom size and students’ behavioral problems are probably not the only causal factors for performance but there is never really an ending, we should just add as many as we can.

Chart, scatter chart

Description automatically generated

Everytime we subdivide by another factor we are controlling for every possible option but the data is quickly subdivied, leaving less and less data in each category. In this case, with only 4 factors, we already have 16 cubes and we might be running into the **coverage problem** – some cells might have no data in them.

Text

Description automatically generated

Chart, line chart

Description automatically generated

Why aren’t all the points on the mean? It could be noise, systematic error, etc. and this needs to be explained.

This variance from the mean to the actual point is what needs to be explained. We manage to explain all the variance from the mean to our prediction but any variance from the point to our prediction we could not explain.

Chart

Description automatically generated

A picture containing shape

Description automatically generatedNow, the squared total deviations (black squares) is equal to the sum of the squared explained deviations (blue squares) and the squared unexplained/residual deviations (red squares).

What makes a model better than another model?

A model is better if it explaines more of the variance.

COD: Coefficient of Determination (the ratio of variance that the model explains) which is the same as R2

Text

Description automatically generatedChart

Description automatically generated with medium confidence