**Regularization**

Let’s start with an example, we want to predict the scores of an exam based on some fetaures of the students:

Chart, scatter chart

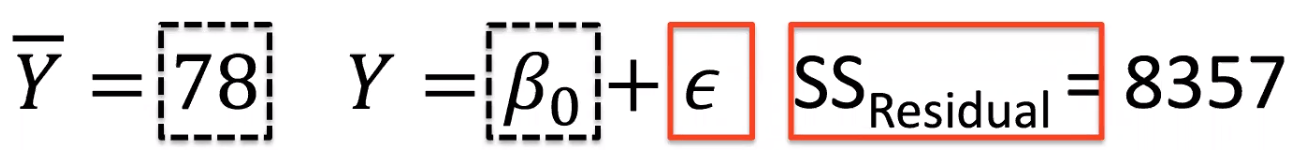
Description automatically generated

Chart, scatter chart

Description automatically generatedIf we only know the student number of each student, what is the best guess for any student’s exam? The mean itself! There is no apparent correlation betweem studet number and exam performance (obviously) so we could build a model that is just the mean line.

A picture containing text

Description automatically generated



Chart, scatter chart

Description automatically generatedIs the mean an accurate prediction? No, there are a lot of residuals. Is the noise around the mean due to pure randomness or could we find other factors that could explain that variance? Yes, let’s add hours studied.

Shape

Description automatically generated with medium confidenceDiagram

Description automatically generated with low confidence

Text

Description automatically generated with low confidence

Chart, scatter chart

Description automatically generatedIs this the best we can do? What else matters when it comes to predicting a student’s grade?

Text

Description automatically generated with medium confidenceA picture containing diagram

Description automatically generatedA picture containing text

Description automatically generatedIcon

Description automatically generatedA picture containing text, clipart

Description automatically generated

Notice that even though we have two predictors where each explains 40% and 70% of the variance, their combination explains 72.9%. Which means that they overlap, because they are not independent!

We can keep adding variables such as X3 being lab attendance with rLP=0.6 and SSResidual=2991 and R2=0.64, but how can we visualize it?

Timeline

Description automatically generated with medium confidenceChart, scatter chart

Description automatically generated

A picture containing graphical user interface

Description automatically generated

The more predictors we put in, the lesser the SSResiduals decreases (diminishing returns).

Chart

Description automatically generatedChart

Description automatically generated with medium confidenceChart, line chart, histogram

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, histogram

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Chart, histogram

Description automatically generated

We can keep adding predictors until we reach an R2 of 1 but that is overfitting the model.

There are considerations other than maximizing variance accounted for when building a model (which motivates regularization):

* Multi-collinearity
* Coverage / sparsity
* Overfitting
* …

Multi-collinearity

Nulti-collinearity arises when predictors itself are correlated. If we have two predictors that are **each** correlated with the outcome at 0.9 (R) they have to be correlated, because that means that each explains 81% (R2) of the outcome’s variance and therefore they do not account for unique parts of the variance, as this would add up to well over 100%, implying that they are at least somewhat correlated.

Note that if two variables are not correlated they are not necessarily independent.

Why is this important? Highly correlated predictors can “stump” the regression models, making **parameters estimates unreliable (β)**. The model won’t know how to distribute the β among these predictors and therefore there won’t be a unique solution and everytime this model is run, you might get a different solution (which is unacceptable). How do we fix this? Using fewer predictors to reduce correlation or use regularization.

Sparsity (the issue of sparsity)

Complex models (with may predictors) quickly need extraordinary amounts of data to estimate the parameters. There simply might not be enough data available to cover the high-dimensional space eveny and densely enogh to estimate regression parameters reliably

Overfitting

A complex model with many parameters can fit any dataset – every single data point – perfectly. The problem lays in the fact that each data point consists of true vakue plus error (“noise”) and therefore we are also fitting that noise. Therefore it won’t generalize because the noise is completely random.

Bias/variance tradeoffs

If the function we are trying to fit is too simple to capture the true relationship between predictor and outcome, we are underfitting the data. This inability to reflect the true relationship is called **bias**.

Diagram

Description automatically generatedIf the function we are trying to fit the data is too flexible, we are overfitting the data. If that is the case, while the fit to one dataset might be great, it will not generalize to other datasets, where the fit will be worse. This is called **variance**.

Diagram

Description automatically generated

Text

Description automatically generatedRegularization

The most common regularization techniques for linear regression are Ridge and Lasso regularization.

They both consist on adding a penalty term to the loss function, which simplifies the model (counteracting the increasing complexness of adding more predictors)

Text

Description automatically generated

How do we find the optimal λ

A picture containing text, newspaper, screenshot

Description automatically generated

Chart

Description automatically generated

As λ increases, the βs decrease towards 0 (simplifying the model) and where the RMSE of the test set hits a minimum we can find our optimal value.

Note that the intercept β0 is not included in the regularization term.

Text, letter

Description automatically generatedText

Description automatically generated