|  |
| --- |
| Course: Data Science [Z15811] Class: Modeling & Evaluation [05]  Subject: Discussion topics |

1. In R, just using lm() will train a model for you, but how does R calculate the different parameters of the linear model that minimize the RMSE? Do some research online.

RMSE = mean((observeds - predicteds)^2) %>% sqrt()

1. The tools of linear regression can be used even when the relationship between the input and output variables in non-linear. Explain this statement and show with an example.

linear regression can be significant if the true relationship is non-linear. A linear regression finds a line of best fit through your data and simply tests, whether the slope is significantly different from 0.

If it makes sense to assume that there is a non-linear relationship as for example between car speed and braking distance, then you can add squared terms (or other transformations) of your independent variable.

1. Do some research on ggplot’s geom\_smooth() function and explain how it is connected with regression.

If you are using the same x and y values that you supplied in the ggplot() call and need to plot the linear regression line then you don't need to use the formula inside geom\_smooth(), just supply the method="lm".

ggplot(data,aes(x.plot, y.plot)) +

stat\_summary(fun.data= mean\_cl\_normal) +

geom\_smooth(method='lm')

1. Why do you need to split up the data into a training set and a test set? When do you need a validation set?

There will not be enough data in the training dataset for the model to learn an effective mapping of inputs to outputs.

 the process is like this:

Train/Test split

Choose final model by cross-validation on the training data. This includes tuning. And retraining on the full training data.

Evaluate on the test data to get an impression of true performance (and bias) of your model.

1. Explain the bias variance trade-off.
   1. Draw a dataset and a fitted model with high bias.
   2. Draw a dataset and a fitted model with high variance.
   3. How can you strike the trade-off between bias and variance?

Trade-off is tension between the error introduced by the bias and the variance.

1. Explain how you can use the residuals of a model for data exploration.

Je kan het verschil tussen geobserveerde en voorspelde data weergeven

1. Come up with two examples of selection bias.

* Fouilleren van een enkele huidskleur
* Image recognition werkt beter bij blanke mannen

1. How are classification and regression different? Is RMSE a good metric to evaluate classification models? Why (not)?

Classificatie is de taak van het voorspellen van een discreet klassenlabel. Regressie is de taak van het voorspellen van een continue grootheid.

Nee, RMSE kan beter worden gebruikt om de prestaties tussen verschillende regressiemodellen te vergelijken. Bij Classificatie zal elke gekwadrateerde fout effectief een 1 zijn.

1. Find some examples on Simpson’s paradox. Is this an occurrence of bias? Why (not)?

Covid-19 tussen Italië en China,