Regression

Topics covered so far

1. Regression: Linear Regression
    a. General Statistical Framework
    b. Maximum Likelihood and Bayesian Estimators
    c. Linear Regression
    d. Performance Assessment - Estimating parameter means and confidence intervals for prediction

2. Regression: Model Evaluation
    a. Prediction vs Modeling
    b. Assumptions behind Regression
    c. Bias-variance tradeoff
    d. Overfitting and Regularization
    e. Cross-Validation
    f. Bootstrapping

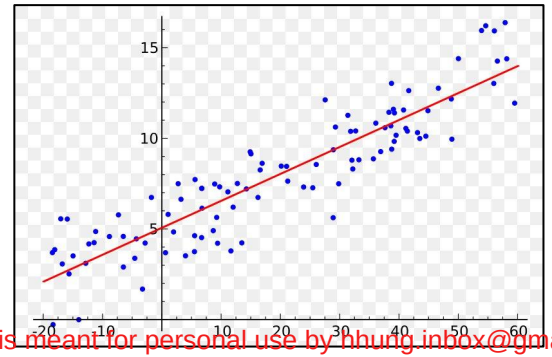
Discussion questions

1. What is linear regression and how does it work?
2. What is multiple linear regression? What are some examples where it could be used?
3. How do you measure the performance of a linear regression model?

Linear Regression

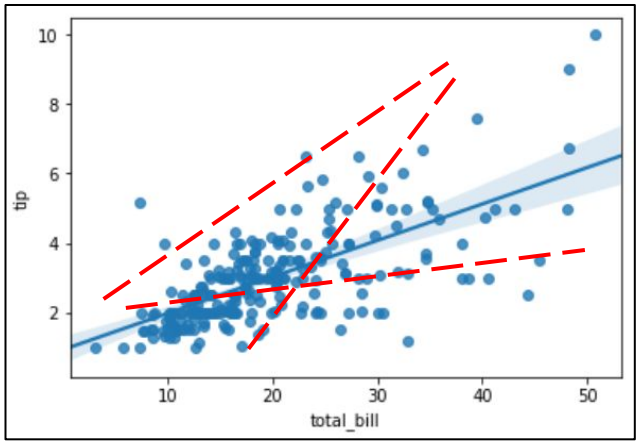
- Linear regression is a way to identify relationship between the independent variable(s) and the dependent variable.
- We can use these relationships to predict values for one variable for given value(s) of other variable(s).
- It assumes the relationship between variables can be modeled through linear equation or an equation of the line.
- The variable, which is used in prediction is termed as independent/explanatory/regressor, whereas the predicted variable is termed as dependent/target/response variable.
- In case of linear regression with a single explanatory variable, the linear combination can be expressed as:

response = intercept + constant * explanatory variable



Best fit line in the linear regression model

- Learning from the data, the model generates a line that fits the data.
- Our aim is to find a regression line that best fits the data.
- By best fit, it means that the line will be such that the cumulative distance of all the points from the line is minimized.
- Mathematically, the line that minimizes the sum of squared error of residuals is called the Regression Line or the Line of Best Fit.



- In the example here, you can see a scatter plot between the total_*tip* amount and the *total_bill* amount.
- We can see that there is a positive correlation between these variables. As the bill amount increases, the tip increases.
- The blue line is the 'best fit' line and those in red are some examples of other lines that are not the 'best fit'.

What is Multiple Linear Regression?

- This is just an extension of the concept of simple linear regression with one variable, to multiple variables.
- In the real world, any phenomenon or outcome could be driven by many different independent variables.
- Therefore there is a need to have a mathematical model that can capture this relationship.

  - **Ex:** Predicting the price of a house, we need to consider various attributes such as area, number of rooms, number of kitchens, etc. Such a regression problem is an example of multiple linear regression.
  - The equation for multiple linear regression can be represented by:

target = intercept + constant 1*feature 1 + constant 2*feature 2 + constant 3*feature 3 + ...

- The model aims to find the constants and intercept such that this line is the best fit.

Regression Model Evaluation Metrics

R-squared	Adjusted R-squared	Mean Absolute Error	Root Mean Square Error
• A measure of the % of the variance in the target variable explained by the model • Generally, the first metric to look at for linear regression model performance • Higher the better	• Conceptually, very similar to R-squared but penalizes for the addition of too many variables • Generally, used when you have too many variables as adding more variables always increases R-squared but not Adjusted R-squared • Higher the better	• Simplest metric to check prediction quality • Same unit as the dependent variable • Not sensitive to outliers, i.e. the metric is not affected too much if there are outliers • Difficult to optimize from a mathematical point of view (pure maths logic) • Lower the better	• Another metric to measure the quality of predictions • Same unit as the dependent variable • Sensitive to outliers - errors will be magnified due to the square function • But has other mathematical advantages • Lower the better

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - yhat_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - yhat_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - yhat_i)^2}$$

Discussion questions

1. What are the underlying assumptions in the linear regression model?
2. What is the Bias-Variance tradeoff?
3. What is Regularization? What are its different types?
4. Why do we use Cross-Validation? How does it work?
5. What is the concept of bootstrapping and why do we need it?

Assumptions of Linear Regression

Assumption	How to test	How to fix
There should be a linear relationship between dependent and independent variables	Pair plot / Correlation of each independent variable with dependent variables	Transform variables that appear non-linear (log, square root, etc.)
No multicollinearity in independent variables	Heatmaps of correlations or VIF (Variance Inflation Factor)	Remove correlated variables or merge them
No Heteroskedasticity - residuals should have constant variance	Plot residuals vs. fitted values and check the plot	Non-linear transformation of the dependent variable or adding other important variables
Residuals must be normally distributed	Plot residuals or use a Q-Q plot	Non-linear transformation of independent or dependent variables

Bias-Variance Tradeoff: Underfitting and Overfitting

Bias: Bias is the difference between the prediction of our model and the correct value that we are trying to predict. Models with high bias give less attention to the training data and overgeneralize the model which leads to a high error in the training and the test datasets.

Variance: Models with high variance pay a lot of attention to the training data, including the noise, and do not generalize on the test data. Therefore, such models perform very well on training data but have a high error on the test data.

In supervised learning, **underfitting** happens when a model is not able to capture the underlying pattern of the data. These models usually have high bias and low variance, whereas, **overfitting** happens when our model captures the noise along with the underlying pattern in training data. These models usually have low bias and high variance.

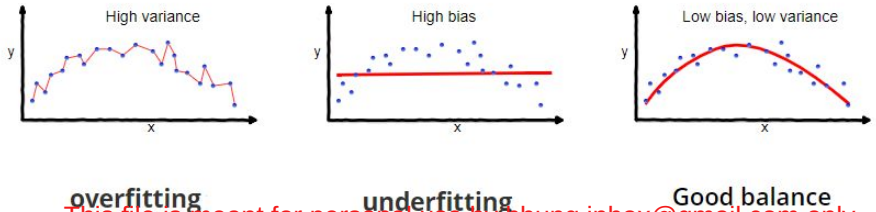


Image Source

Bias-Variance Tradeoff

If our model is too simple and has very few parameters, then it may have high bias and low variance. On the other hand, if our model has a large number of parameters, then it's going to have high variance and low bias. So, we need to find the right/good balance between overfitting and underfitting the data.

An optimal balance of bias and variance would neither overfit nor underfit the model.

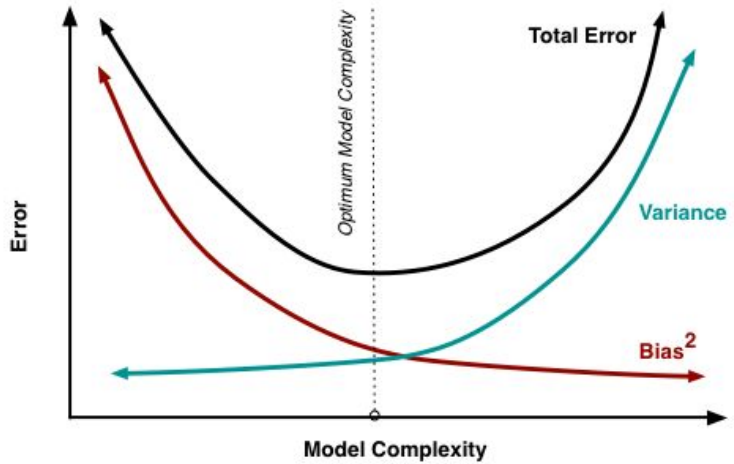


Image Source

Regularization and its types

- Regularization is the process that regularizes or shrinks the coefficients towards zero. In other words, this technique discourages learning a more complex or flexible model, to avoid the risk of overfitting.
- Regularization significantly reduces the variance of the model, without a substantial increase in its bias.
- The two most common types of regularization in regression are:
    - **Lasso Regression:** In this technique, we add $\alpha * \sum|\beta|$ as the shrinkage quantity. It only penalizes high coefficients. It has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter α is sufficiently large. This technique is also called L1 regularization.
    - **Ridge Regression:** In this technique, we modify the residual sum of squares by adding the shrinkage quantity $\alpha * \sum\beta^2$ and use α as the tuning hyperparameter that decides how much we want to penalize the flexibility of our model. This technique is also called L2 regularization.

Cross-validation and its types

Cross-validation is a technique in which we train our model using the subset of the dataset and then evaluate using the complementary subset of the dataset.

- It provides some kind of assurance that the model has got most of the pattern from the dataset correct and it is not picking up noise.
- Two most common types of cross-validation techniques are:
    1. K-Fold Cross-Validation
    2. Leave-One-Out Cross-Validation (LOOCV)

Case Study

Appendix

Statistics vs Machine Learning

The difference between machine learning and statistical learning is their purpose. Machine learning models are designed to make the most accurate predictions possible, whereas statistical models are designed for inference about the relationships between variables.
The following table highlights the major differences between statistics and the machine learning point of view:

Statistics	Machine Learning
Emphasis on deep theorems on complex models	Emphasis on the underlying algorithm
Focus on hypothesis testing and interpretability	Focus on predicting the accuracy of the model
Inference on parameter estimation, errors, and predictions	Inference on prediction
Deep understanding of simple models	The theory does not always explain the success

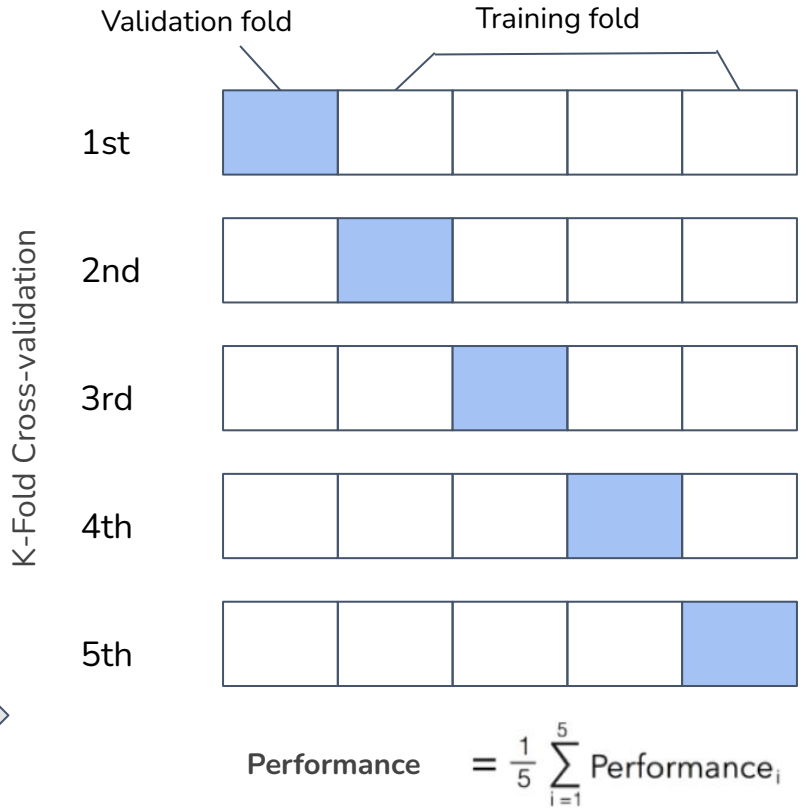
K-Fold Cross-Validation

This algorithm has a single parameter called K which refers to the number of groups that a given dataset is to be split into.

This algorithm has the following procedure:

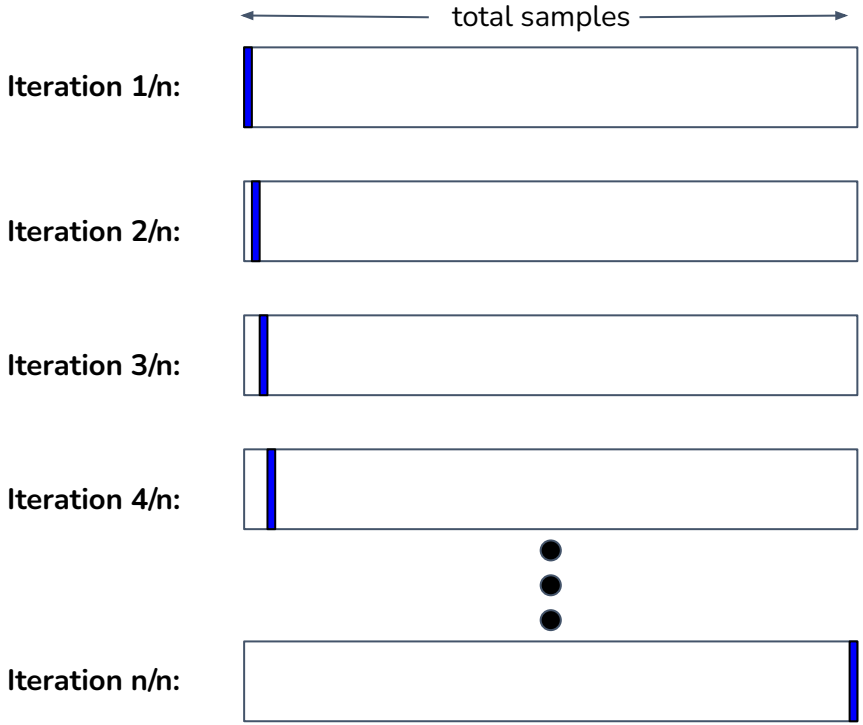
1. Shuffle the dataset randomly.
2. Split the whole dataset into K distinct groups.
3. In each iteration, take one group as a hold-out set and the remaining as the training set.
4. Repeat step 3, K times with a different group, as a validation set, in each iteration.
5. Summarize the skill of the model using the average model evaluation scores of all groups.

Here, K = 5



Leave-One-Out Cross-Validation (LOOCV)

- LOOCV is a special case of K-Fold Cross-Validation where K equals n, n being the number of data points in the dataset.
- This approach leaves 1 data point out of the training data, i.e., if there are n data points in the original dataset, then n-1 data points are used to train the model and 1 data point is used as the validation set.
- This is repeated for all combinations in which the original dataset can be separated this way, and then the error is averaged for all trials, to give an overall model performance.
- The number of possible combinations is equal to the number of data points in the original dataset, i.e., n.



Bootstrapping

Bootstrapping (also called Bootstrap sampling) is a resampling method that involves the drawing of samples from the data repeatedly with replacement to estimate a population parameter.

It involves the following steps:

1. Choose a number of bootstrap samples to perform
2. Choose a sample size n
3. For each bootstrap sample
    1. Draw a sample with replacement with the chosen size
    2. Calculate the statistic on the sample
4. Calculate the mean of the calculated sample statistics

Bootstrap sampling can be used to estimate the parameter of a population, for example, mean, standard error, etc.

Great Learning

Happy Learning !