## Regression

►Look for relationships amongst variables

►Identify the relation between salary and experience, education

Terminology

Variables:

* Independent Variables/Inputs/Predictors, E.g., experience, education
* Dependent Variables/Outputs/Responses, E.g., salary of employee

Observation is a data point, row, or sample in a dataset

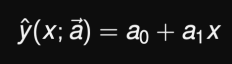
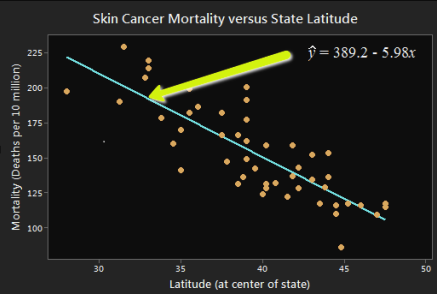
E.g., an employee's salary, experience, education, role.

When Use Regression

* To determine how variables are related
  + determine if experience or education impact salaries
* To forecast a value
  + E.g., predict electricity consumption given the outdoor temperature, time of day, and number of residents in that household

Simple Linear Regression

Regression fits a very simple equation to the data:

Best Fitting Line

Aim is that the predicted response, close to the actual response.

Calculating Parameters Intuition

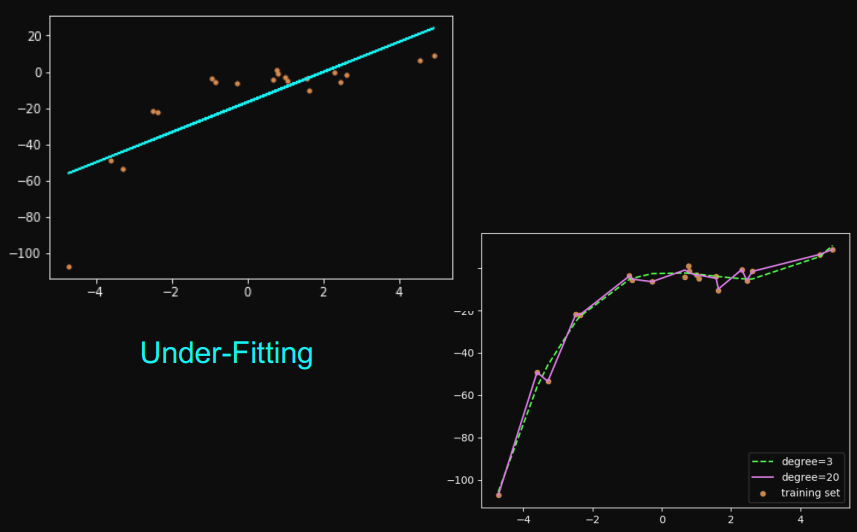
Given some data pairs, we fit a model by finding the vector that minimises the loss function:



Compare the derivatives to zero

Polynomial Regression

## Underfitting and Overfitting



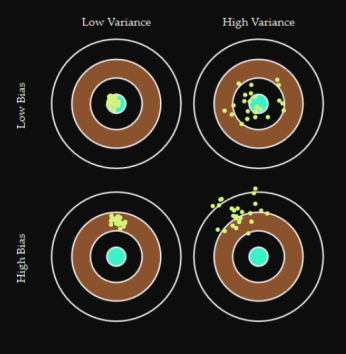
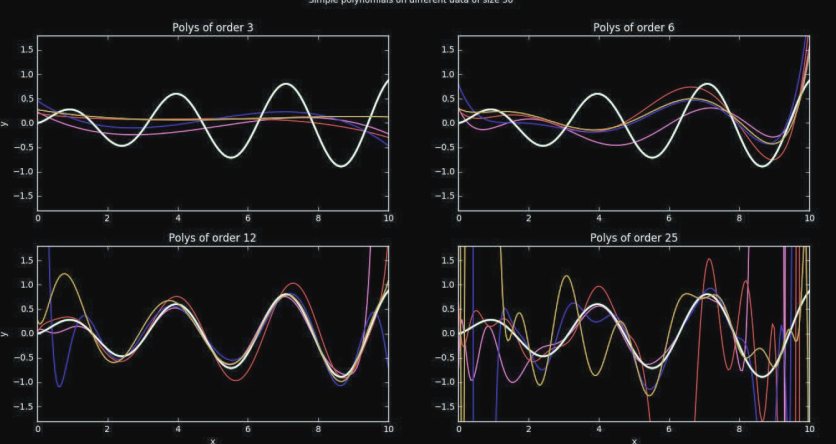
## Overfitting

The more parameters a model has, the more complicated a curve it can fit.

If we don’t have very much data and we try to fit a complicated model to it, the model will make wild predictions.

* Small polynomial; cannot fit the data well; said to have high bias
* Large polynomial; can fit the data well; fits the data too well; said to have small bias
* Poor fit due to high bias called underfitting
* Poor fit due to low bias called overfitting

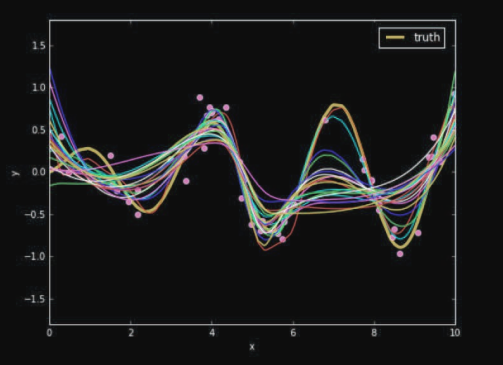
Bias and Variance

* Bias: measures how much the prediction differs from the desired regression function.
* Variance: measures how much the predictions for individual data sets vary around their average.

## Ensembles

* given only data, we do not know the truth and can only estimate what may be the “truth”
* an ensemble is a collection of possible/reasonable models
* from this we can understand the variability and range of predictions that is realistic



generating an ensemble is a whole statistical subject in itself

