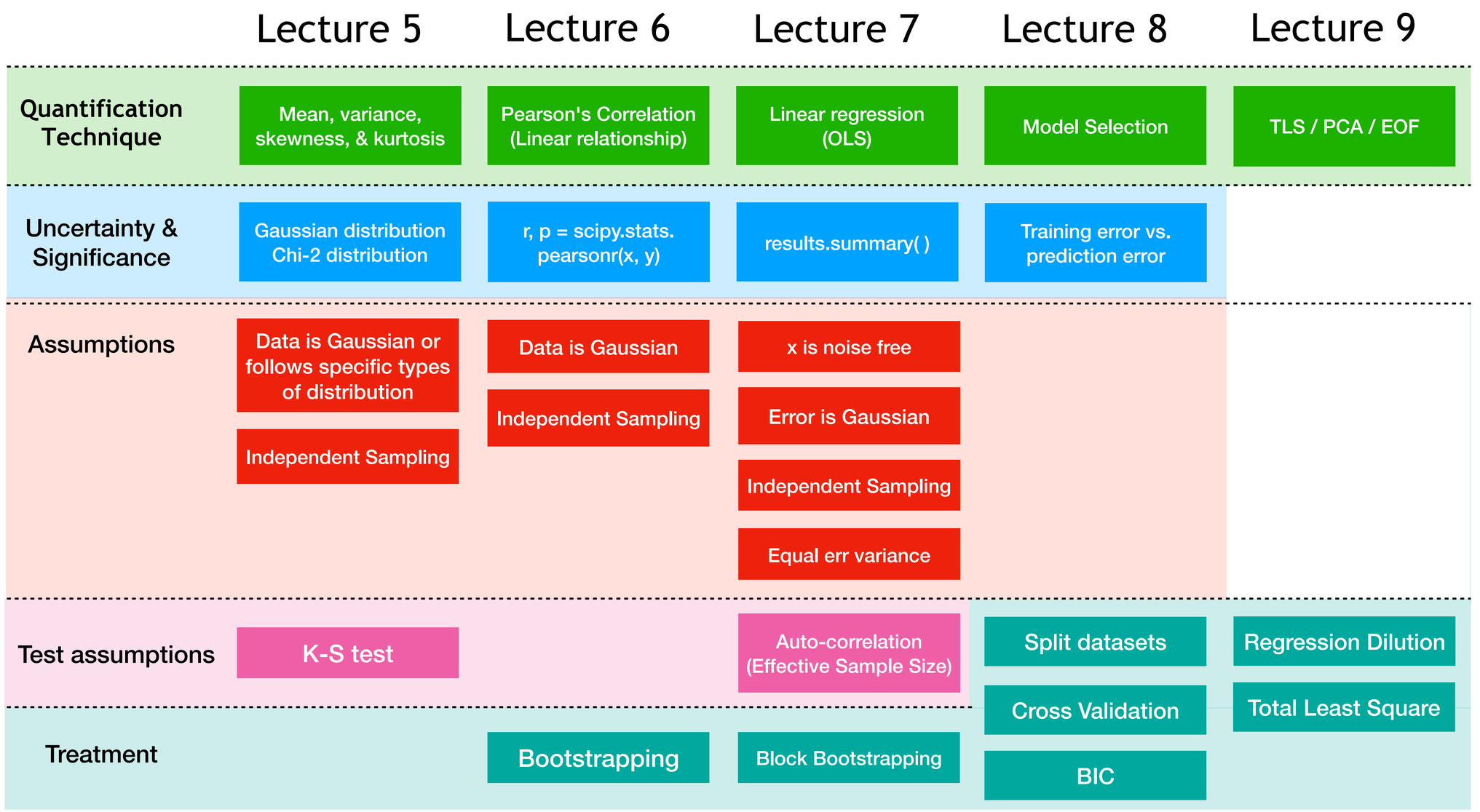
Lecture 10:

Review & Advanced Regression Techniques

Essential to pass the course

Bonus to get higher scores

Road Map of the Statistics Part



Explore a single dataset - a crucial building block for later analysis (Problem 1)

Remove mean / seasonal cycle...

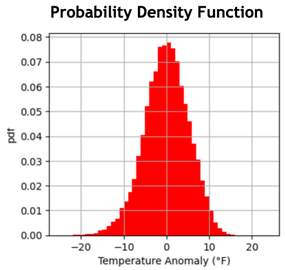
Plot and visualise your data

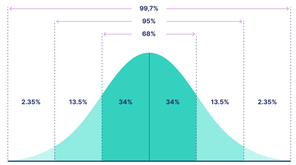
Calculate Summary Statistics

Is data Gaussian or Normally distributed?

K-S test

p-value?





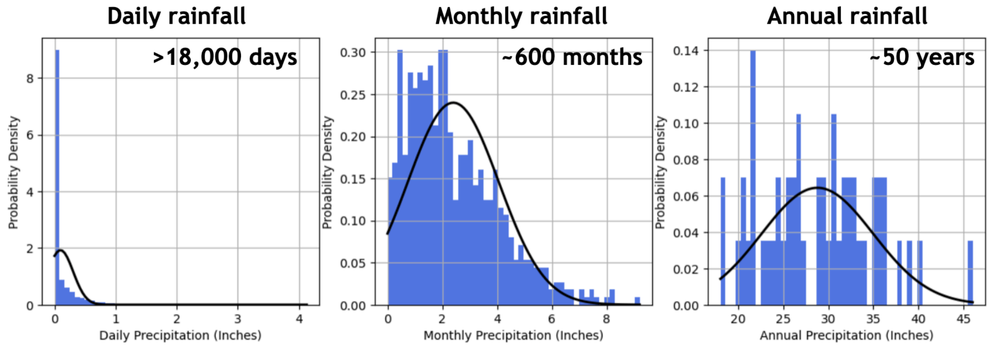
Skewness of 0

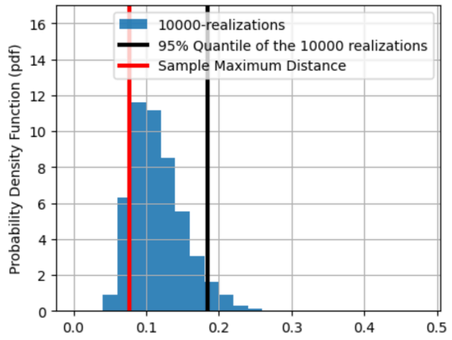
Kurtosis of 3

Lecture/Practical 5

Central Limit Theorem:

Likely Gaussian when average or sum over >30 independent samples (such as monthly or annual mean values).





Terminology	Meaning
Null hypothesis (H_0)	This is our starting assumption that the effect being studied does not exist
Alternative hypothesis (H_1)	This is what we might believe to be true if we find sufficient evidence against the null hypothesis.
Test statistics	This is a calculated value from our data that we use to test our hypothesis.
Null distribution	This represents what we would expect to see from our test statistic purely by chance if the null hypothesis were true.
significance level (α)	This is a threshold we set to decide when to reject the null hypothesis.
p-value	The probability of obtaining our data, or something more extreme, if the null hypothesis is true.

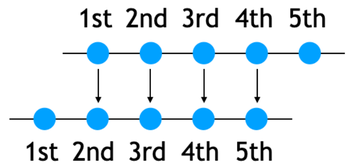
Lecture/Practical 7

Are data independent?

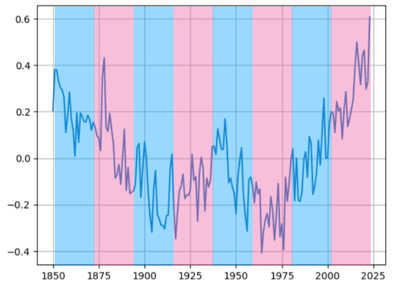
Auto-correlation test

No

Calculate Effective Sample size



$$\mathrm{ESS} = \frac{N}{1 + 2\sum_{k=1}^{\infty} \rho_k}$$

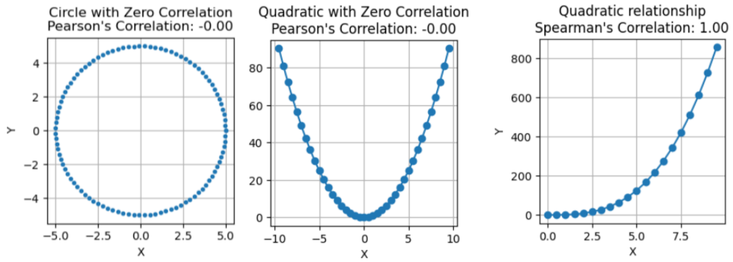


Steps for evaluating correlations between two variables (Problem 2)

Prepare

Remove mean / seasonal cycle...

Plot and visualise your data



Are we interested in a linear relationship?

No → Use other methods
e.g., Spearman's rank correlation, or regression

Yes

Analysis

Calculate Pearson's Correlation

`r, p = scipy.stats.pearsonr(x, y)`

Is data Gaussian?

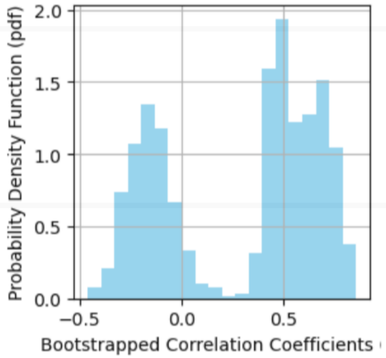
Significance

K-S test

No → Bootstrapping

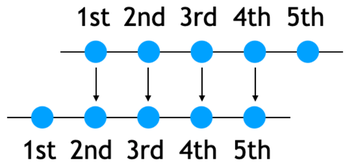
Yes

use scipy's p value to evaluate significance

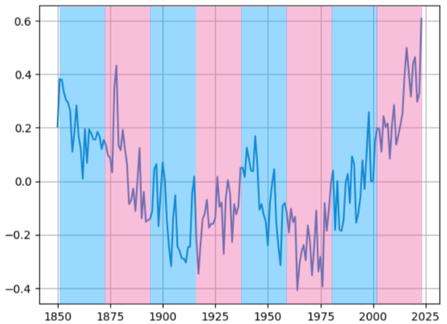


Lecture/Practical 6

Lecture/Practical 7



$$\mathrm{ESS} = \frac{N}{1 + 2\sum_{k=1}^{\infty} \rho_k}$$



Are data independent?

Auto-correlation test

No → Calculate Effective Sample size

Block Bootstrapping

Yes

Steps for evaluating correlations between two variables (Problem 3)

Prepare

Do all data have the same unit? → No → **JUST STOP HERE!!!**

Yes

Remove mean /seasonal cycle

Weigh by the square root of latitude

Reshape data: space x time

Total Least Square Only

PCA/EOF Only

Common steps

Analysis

SVD

U, D, VT = np.linalg.**svd**(A)

(1) Individual columns of **U** are **orthogonal**.

The new directions are perpendicular to each other.

(2) Individual columns of **V** are **orthogonal**.

Pearson's correlations of locations in the new coordinate is zero.

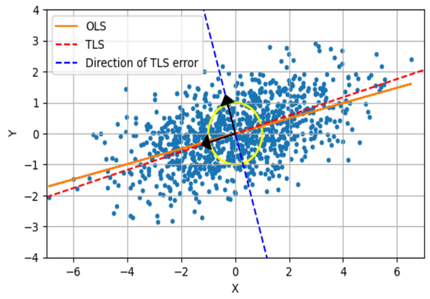
(3) D_i is ranked in a **descending order**.

Post-processs

Get the slope from the first column of U

slope = u_{1Y}/u_{1x}

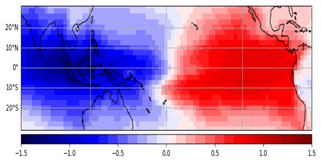
Use Bootstrapping to estimate uncertainty

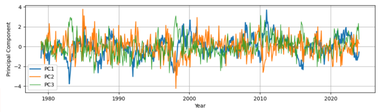


Start from the first mode

Normalise columns of V / rows of V^T

Regression to get the spatial pattern

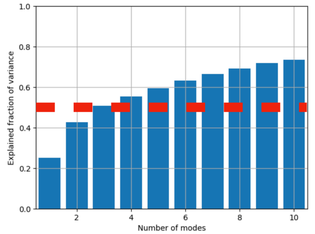




$$FC(s) = \frac{\sum_{i=1}^{s} d_i^2}{\sum_{i=1}^{n} d_i^2}$$

Stop when:

FC > 50%



Lecture/Practical 9

Steps for linear regression using ordinary least square (Problem 4)

Prepare

Remove mean / seasonal cycle... → Plot and visualise your data → Does x contain error?

Does x contain error? — Yes → Do X and Y have the same unit?

Do X and Y have the same unit? — Yes → Total Least Square

Do X and Y have the same unit? — No → Huber Regression / York Regression

Lecture/Practical 18

Does x contain error? — No → Are there apparent outliers?

Are there apparent outliers? — Yes → Huber Regression / York Regression

Are there apparent outliers? — No → **Analysis**

Analysis

```
import statsmodels.api as sm
X = ...
model = sm.OLS(Y, X)
results = model.fit()
Y_hat = results.fittedvalues
```

1. Choose function form
2. Choose loss function
3. Optimization

Significance

Is error Gaussian? — No → Bootstrapping

Is error Gaussian? — Yes → K-S test

K-S test → use result.summary() to evaluate significance

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.714
Model:                            OLS   Adj. R-squared:                  0.712
Method:                 Least Squares   F-statistic:                     428.7
Date:                Mon, 05 Feb 2024   Prob (F-statistic):           1.40e-48
Time:                        11:34:14   Log-Likelihood:                 31.252
No. Observations:                 174   AIC:                            -58.50
Df Residuals:                     172   BIC:                            -52.19
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0722      0.015     -4.686      0.000      -0.103      -0.042
x1             0.0064      0.000     20.706      0.000       0.006       0.007
==============================================================================
Omnibus:                        4.837   Durbin-Watson:                   0.335
Prob(Omnibus):                  0.089   Jarque-Bera (JB):                4.856
Skew:                           0.376   Prob(JB):                       0.0882
Kurtosis:                       2.679   Cond. No.                         50.2
```

Overall significance of the entire model

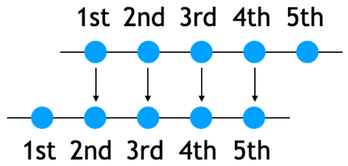
Significance of individual parameters

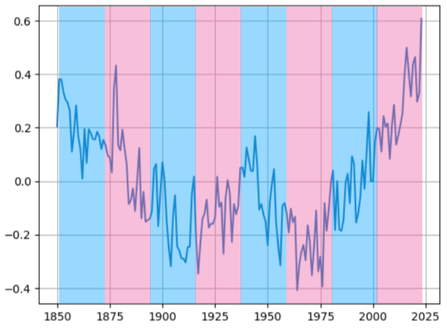
Lecture/Practical 7

Is error Gaussian? — Yes → Is error independent?

Is error independent? — Yes → Auto-correlation test

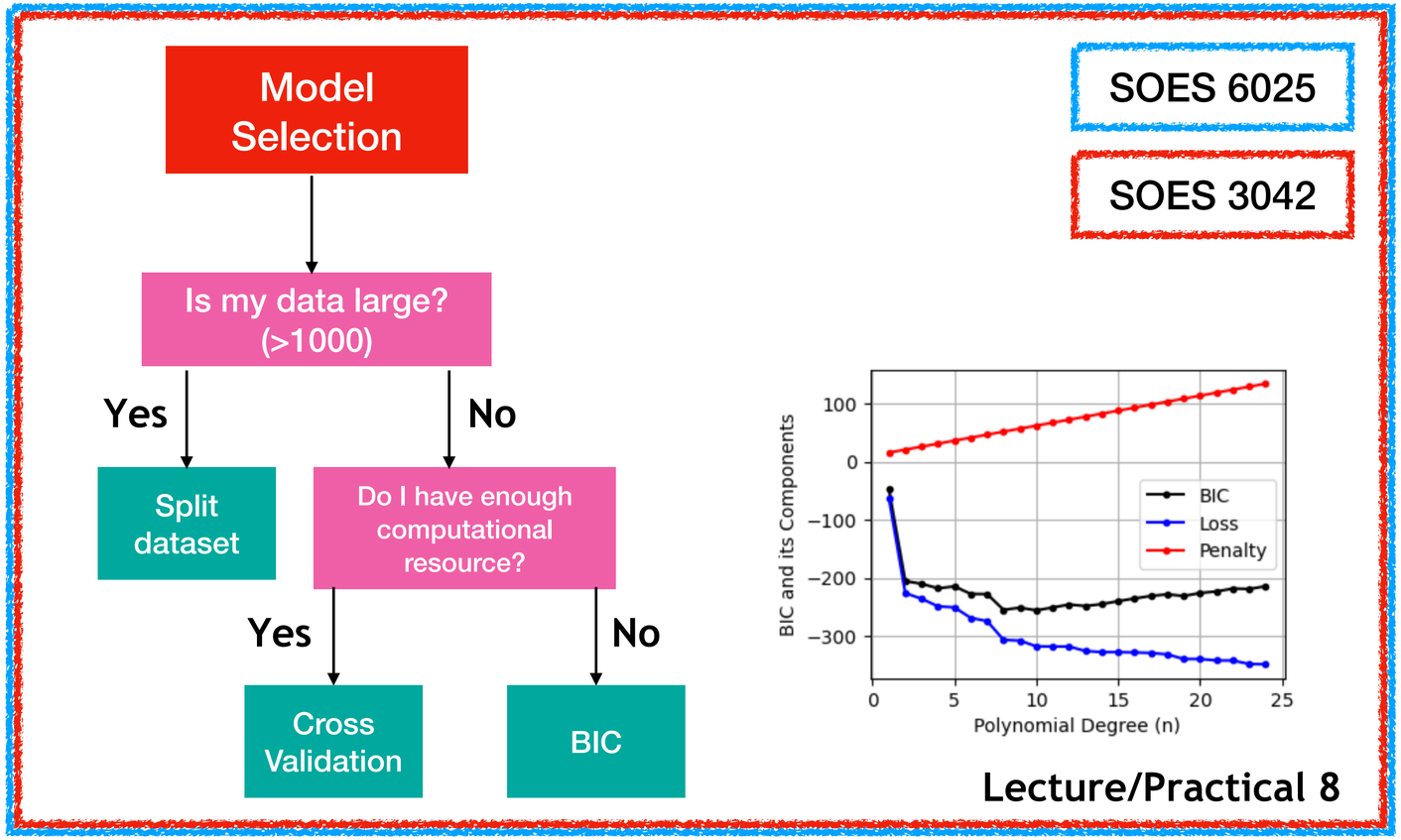
Auto-correlation test — No → Calculate Effective Sample size → Block Bootstrapping





Lecture/Practical 7

Steps for Model Selection (Problem 4)



Accounting for Regression Dilution: York Regression

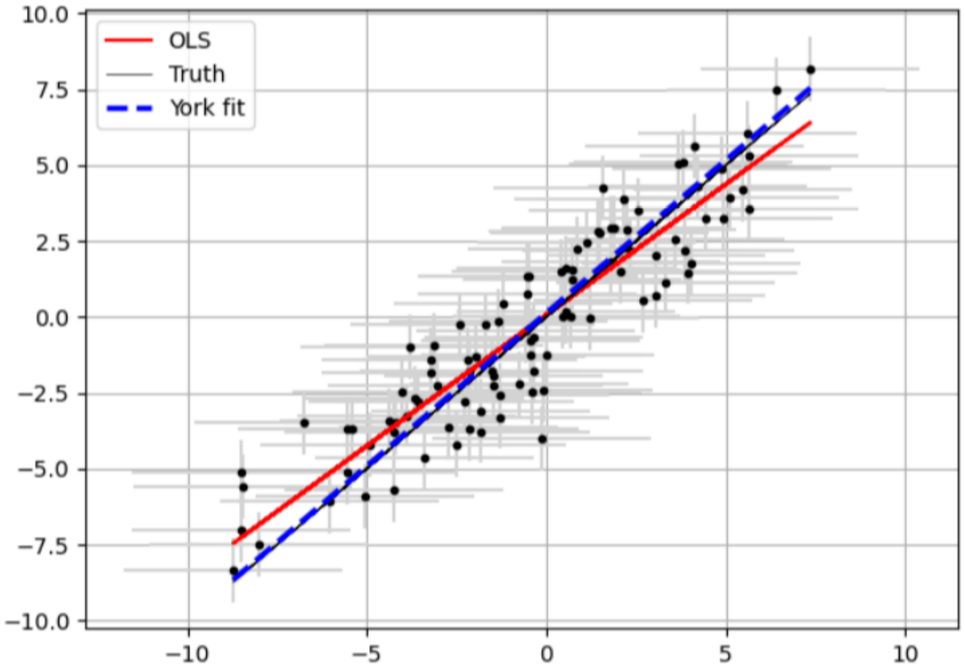
When the error of individual data points is **known**

York Regression is a way to account for regression dilution.

$$P(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp[-\frac{(x-\mu)^2}{2\sigma^2}]$$

$$\sigma^2 = \sigma_y^2 + \alpha^2\sigma_x^2 - 2\alpha\rho\sigma_x\sigma_y$$

Weight is a function of fitted slope



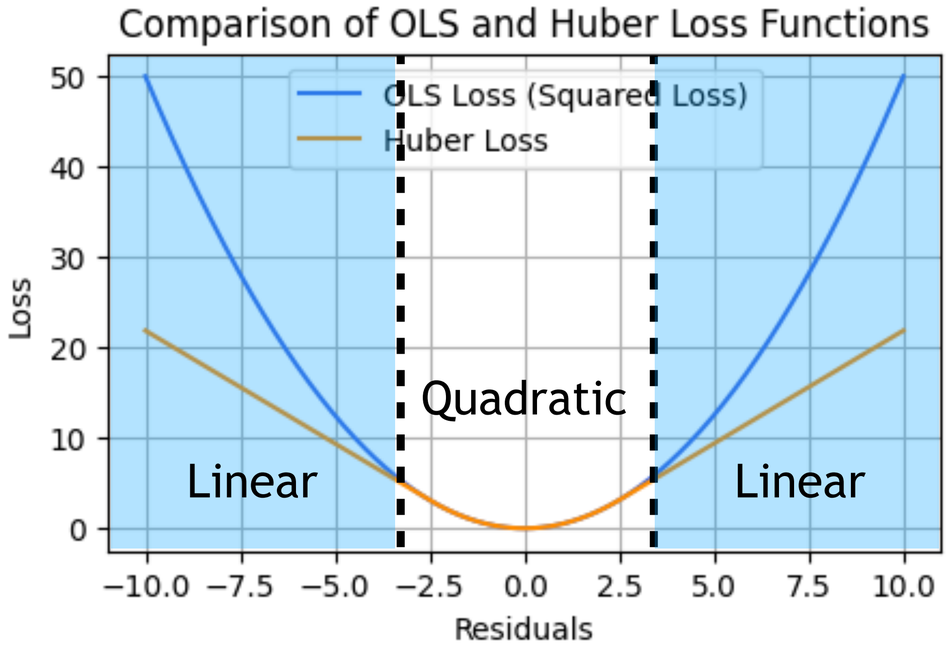
Input: x, y, σ_x, σ_y, ρ (optional)

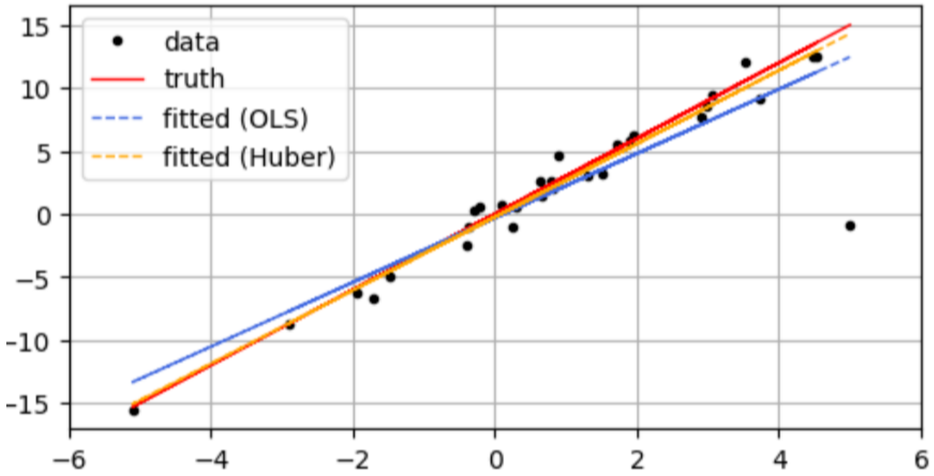
The algorithm uses an **iterative approach that converges to the solution.**

Account for Outliers: Huber Regression

When the error of individual data points is **unknown**

We need to estimate which data points are more reliable

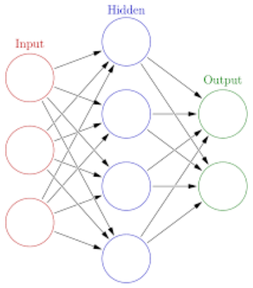




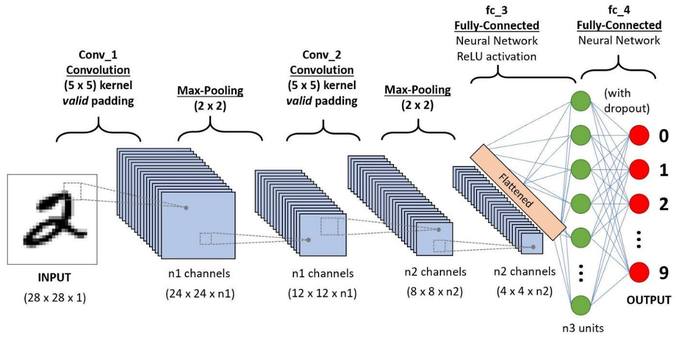

```
import statsmodels.api as sm

X = sm.add_constant(years)

model = sm.RLM(y, X,

        M=sm.robust.norms.HuberT())

results = model.fit()

GMST_hat = results.fittedvalues
```

Generalised Regression and Machine Learning

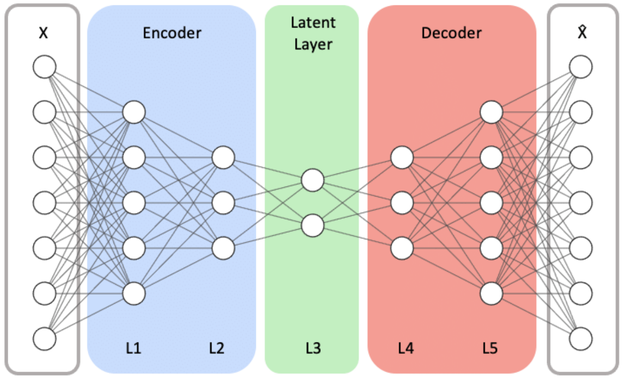
1 **Choose a functional form** → **2** **Define what it means by fit** → **3** **Find that "most"**



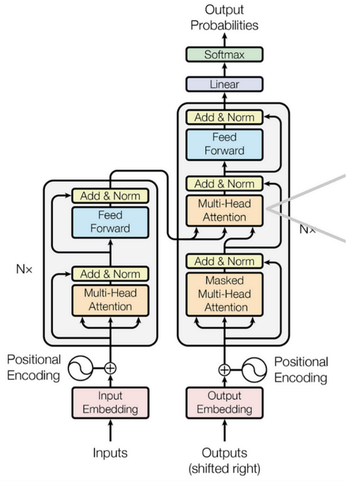
Neural Network



Convolutional Neural Network
(Image recognition; weather forecast)

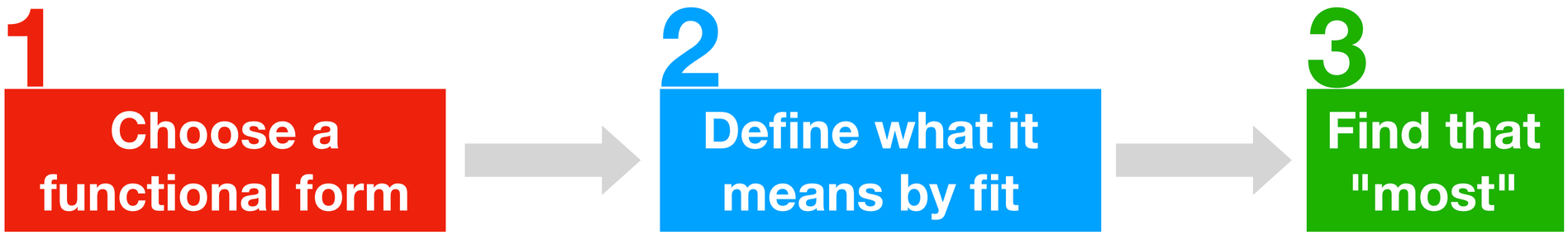


Auto encoder - decoder
(Dimension reduction; non-linear EOF)



Transformer
(GPT)

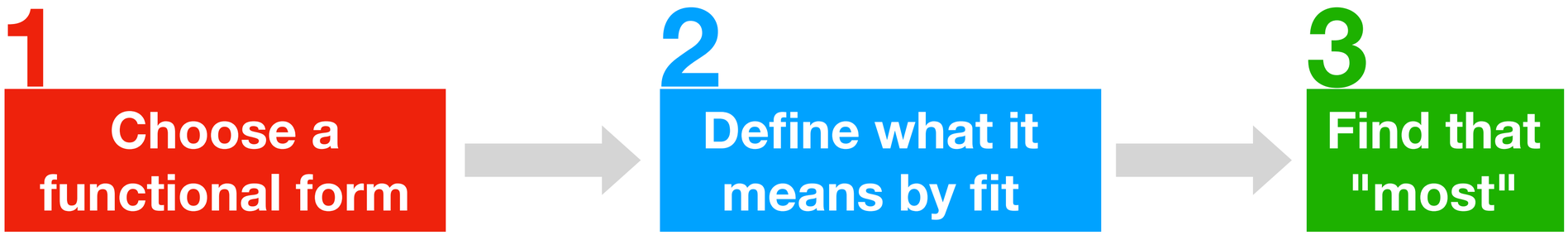
Generalised Regression and Machine Learning



Other loss functions

Absolute Loss (L1 Loss) - LASSO regression
Kullback-Leibler (KL) Divergence - For machine learning categorical data
Poisson Loss - For count data.
Binomial Loss - For binary outcomes or proportions.
Multinomial Loss - For multi-class categorical data.
Exponential Loss - For time-to-event or survival data.
Gamma Loss - For positively skewed continuous data.

Generalised Regression and Machine Learning



Optimisation

```
from scipy.optimize import minimize
result = minimize(neg_log_likelihood, initial_params,
                  args=(y, omega), bounds=bounds)
```

Classic Optimisation Methods:

Gradient Descent - Iterative gradient minimisation.
Newton's Method - Uses second derivatives.
Quasi-Newton (BFGS) - Approximates second derivatives.
Conjugate Gradient - Efficient for large-scale problems.
Simplex Method - Linear programming.
Lagrange Multipliers - Constrained optimisation.

Advanced/ML Optimisation Methods:

Stochastic Gradient Descent (SGD) - Mini-batch updates.
Adam - Adaptive learning rates.
Bayesian Optimization - Probabilistic global optimisation.
Genetic Algorithms - Evolution-inspired optimisation.
Simulated Annealing - Randomised global search.