STA 2408 : REGRESSION MODELLING II

Dennis Ochambo
0705979402
ogot.dennis@jkuat.ac.ke

Pre-requisite

STA 2312 Regression modelling I II

Purpose

To enable the student handle non-linear parametric and non-parametric regression

Objectives

By the end of this course you should be able to

i) Describe a non-linear regression model
ii) Describe the concept of mixed models
iii) Explain the concepts of non-parametric regression models.
iv) Explain the concepts of Neural Network regression models

Description

A review of linear models

Non-linear regression

*Non-linear least squares, their estimation and asymptotic properties

Mixed models

Simple non-parametric regression

* Models, estimation and asymptotic properties.

Simple Neural Networks

Estimators & their asymptotic properties

REVIEW OF LINEAR REGRESSION MODELS

- As you should know, the linear regression model is normally characterised with the following eqn

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_K X_K + \varepsilon_i$$

- We often use an estimation technique known as OLS to estimate this regression model
- OLS seeks to minimise the following eqn ;

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- SSE is sum of squared error with observed y and predicted y $(\hat{y})$ utilised in the eqn

- In an OLS regression model that includes only one explanatory variable, the slope B_1 is estimated with to the following least squares eqn

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - x)^2 / n-1}$$

- Notice that the variance of x appears in the denominator whereas the numerator is part of the formula for the $Cov(x,y)$
- Given the slope the intercept is given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Estimation is slightly more complicated in a multiple OLS regression model
- If you recall the matrix notation you may have seen this model represented as

$$Y = \underline{XB} + \varepsilon$$

- Usually the letters are bolded/capitalised to represent vectors and matrices

Y is a vector of values for the outcome variable/Dependent

X indicates a matrix of explanatory variables/Independent

β is a vector of regression coefficients including the intercept β_0 and the slopes denoted by β_i 's.

- The OLS regression coefficients may be estimated with the following equation:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

- A vector of residuals is then given by:

$$\varepsilon = Y - X\hat{\beta}$$

- You should recall that residuals play an important role in linear regression analysis.
- Assuming a sample and that the model includes an intercept, some of the properties of OLS residuals are:

Properties of OLS residuals

1. They sum to zero ie $\sum \varepsilon_i = 0$
2. They have a mean of 0, $E[\varepsilon_i] = 0$.
3. They are uncorrelated with the predicted values of the outcome variable ie $r(\varepsilon_i, \hat{y}) = 0$.

Analysts often wish to infer something about a target population from the sample, thus you may recall that the standard error of the slope is needed since in conjunction with the slope it allows estimation of the t-values and the p-values.
These provide the basis for inference in linear regression modelling.
The Standard Error (SE) of the slope in a simple OLS regression model is computed as

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / n-2}{\sum (x_i - \bar{x})^2}} = \sqrt{\frac{SSE / n-2}{SS(x)}}$$

- Assuming we have a multiple OLS regression model, the SE formula requires modification

$$SE(\hat{\beta}_i) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum(x_i - \bar{x})^2(1 - R_i^2)(n-k-1)}}$$

- The matrix formulation of SE is based on deriving the variance covariance matrix of the OLS estimator
- A simpler version of its computation is

$$\delta_\varepsilon^2 (X^T X)^{-1}$$

$$\text{with } \hat{\delta}_\varepsilon^2 = \frac{\varepsilon^T \varepsilon}{n-k} = \frac{\sum \varepsilon_i^2}{n-k}$$

NOTE :-

- The numerator in the eqn above is simply your sum of squared error.
- The equation is called residual variance or the MSE (Mean Squared Error)
- You may recognize that it provides an estimate of the variance of errors
- The estimate is biased but consistent.
- The square roots of diagonal elements of the variance-covariance matrix yield the SE of the regression coefficients
- Several of the assumptions of the OLS regression model are related to the accuracy of the SE's and thus the inferences that can be made to the target population.
- OLS results in the smallest value of SSE if some of the specific assumptions of the model are satisfied

$e \sim N(0, \sigma^2)$

linear relationship -

- If this is the case, the model is said to result in the best linear unbiased estimator (BLUE)

- It is important to note that this says Best linear and so we are concerned here with linear estimators (which means that there are also non-linear estimators)
- In any event, BLUE implies that the estimators such as slopes from an OLS regression model are unbiased, efficient and consistent
- Unbiasedness refers to whether the mean of the sampling distribution of a statistic equals the parameter it is meant to estimate in the population for example
  If the slope estimated from a sample a good estimate of an analogous slope of the population.
- Even though we rarely have more than one sample, simulation studies indicate that the mean of the sample slopes from the OLS regression model if we take many samples from a population on average equals a population slope
- Efficiency refers to how stable a statistic is from one sample to the next
  A more efficient statistic has less variability from sample to sample
  It is therefore on average more precise.
  If some of the assumptions are satisfied, OLS derived estimates are more efficient than those that might be estimated using other techniques (they have a smaller sampling variance).

with ↑ in sample size the closer the statistic is to the parameter.

- Consistency refers to whether the statistic converges to the population parameter as the sample size increases. It combines characteristics of both unbiasedness and efficiency

Assignment

The linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \quad + \beta_k x_k + \varepsilon_i$$

Consider this eqn and try to answer the following questions

What does the y_i represent, the β, the x? the x's also often include the subscript i, why?

How do we judge the size and direction of the β

How do we decide which x are important and which ones are not

What are some of the limitations in trying to make this decision?
Given this equation, what is the difference between prediction and explanat
What is this model best suited for?
What role does the mean of y play in linear regression models?
Can the model provide causal explanation of social phenomena
What are some of its limitations for studying social phenomena and causal processes?
What are some of the ways we may judge whether the model is a good fit"
How is the R^2 value computed?
Why do some prefer the adjusted-R^2?
What is the root Mean Squared error (RMSE) and why is it useful?

Stats
Numerical Methods

Exponential Model

Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ best fit $y = ae^{bx}$ to the data.

The variables a and b are the constants of the exponential model.

The residual at each data point x_i is

$$E_i = y_i - ae^{bx_i}$$

The sum of squares of residuals is

$$S_r = \sum_{i=1}^{n} E_i^2$$

$$= \sum_{i=1}^{n} \left(y_i - ae^{bx_i}\right)^2$$

To find the constants a and b of the exponential model, we minimize S_r by differentiating with respect to a and b and equating the resulting equations to zero.

$$\frac{\partial S_r}{\partial a} = \sum_{i=1}^{n} 2\left(y_i - ae^{bx_i}\right)\left(-e^{bx_i}\right) = 0$$

$$\frac{\partial S_r}{\partial b} = \sum_{i=1}^{n} 2\left(y_i - ae^{bx_i}\right)\left(-ax_ie^{bx_i}\right) = 0 \qquad 4(a,b)$$

or

$$-\sum_{i=1}^{n} y_i e^{bx_i} + a\sum_{i=1}^{n} e^{2bx_i} = 0$$

$$\sum_{i=1}^{n} y_i x_i e^{bx_i} - a\sum_{i=1}^{n} x_i e^{2bx_i} = 0 \qquad 5(a,b)$$

Equations (5a) and (5b) are nonlinear in a and b and thus not in a closed form to be solved as was the case for linear regression.

In general, iterative ~~theo~~, method ie Gauss-Newton iteration method method of steepest descent, direct search etc must be used to find values of a and b.

- Equation (5a), 'a' can be written explicitly in terms of 'b' as,

$$a = \sum_{i=1}^{n} \frac{y_i e^{bx_i}}{\sum_{i=1}^{n} e^{2bx_i}} \quad \ldots \quad (6)$$

Substituting in eqn (5b) we have

$$\sum_{i=1}^{n} y_i x_i e^{bx_i} - \frac{\sum_{i=1}^{n} y_i e^{bx_i}}{\sum_{i=1}^{n} e^{2bx_i}} \sum_{i=1}^{n} x_i e^{2bx_i} = 0. \quad \ldots\ldots (7)$$

This equation is still a non-linear equation in b and can be solved best by numerical methods such as the bisection method or secant method.

Example

Below is given the relative intensity of radiation as a function of time

t (hrs)	0	1	3	5	7	9
γ	1.000	0.891	0.708	0.562	0.447	0.355

If the level of the relative intensity of radiation is related to time via an exponential formula $\gamma = Ae^{\lambda t}$, find:

a) The value of the regression constants A and λ.

b) The half life of Technium-99m and

c) The radiation intensity after 24 hrs.

Solution

a) The value of λ is given by solving the non-linear equation ~~(4)~~ (7)

$$f(\lambda) = \sum_{i=1}^{n} \gamma_i t_i e^{\lambda t_i} - \frac{\sum_{i=1}^{n} \gamma_i e^{\lambda t_i}}{\sum_{i=1}^{n} e^{2\lambda t_i}} \sum_{i=1}^{n} t_i e^{2\lambda t_i} = 0. \quad \ldots\ldots \boxed{8}$$

and then the value of A from equation 6

$$A = \frac{\sum_{i=1}^{n} \gamma_i \cancel{t_i} e^{\lambda t_i}}{\sum_{i=1}^{n} e^{2\lambda t_i}} \quad \ldots\ldots \boxed{9}$$

Eqn 8 can be solved for λ using bisection method. To estimate the initial guesses, we assume $\lambda = -0.120$ and $\lambda = -0.110$.
We need to check whether these values first bracket the root of $f(\lambda) = 0$.

At $\lambda = -0.120$, the table below shows evaluation of $f(-0.120)$

Summation value for calculation of constants of model

i	t_i	y_i	$y_i t_i e^{\lambda t_i}$	$y_i e^{\lambda t_i}$	$e^{2\lambda t_i}$	$t_i e^{2\lambda t_i}$
1	0	1	0.0000	1.0000	1.0000	0.0000
2	1	0.891	0.79205	0.79205	0.78663	0.78663
3	3	0.708	1.4819	0.49395	0.48675	1.4603
4	5	0.562	1.5422	0.30843	0.30119	1.5060
5	7	0.447	1.3508	0.19297	0.18637	1.3046
6	9	0.355	1.0850	0.12056	0.11533	1.0379
$\sum_{i=1}^{6}$			6.259	2.9062	2.8763	6.0954

$n = 6$, $\sum_{i=1}^{6} y_i t_i e^{-0.120 t_i} = 6.2501$ $\qquad \sum_{i=1}^{6} e^{2(-.120)t_i} = \overset{2.8763}{2.9882}$

$\sum_{i=1}^{6} y_i e^{-0.120 t_i} = 2.9062$ $\qquad \sum_{i=1}^{6} t_i e^{2(-.120)t_i} = 6.0954$

$$f(-0.120) = 6.2601 - \frac{2.9062}{2.8763}(6.0954) = 0.091367$$

Similarly

$f(-0.110) = -0.10099$.

Since

$f(-120) \times f(-1.10) < 0$, the value of λ falls in the bracket of $[-0.120, -0.110]$. The next guess of the root then is

$$\lambda = \frac{-0.120 + (-0.110)}{2}$$

$= -0.115$

Continuing with the bisection method, the root of $f(\lambda) = 0$ is found as $\lambda = -0.11508$. This value of the root was obtained after 20 iterations with an absolute relative approximate error of less that 0.000005%.

From eqn (9), A can be calculated as

$$A = \frac{\sum_{i=1}^{6} Y_i e^{\lambda t_i}}{\sum_{i=1}^{6} e^{2\lambda t_i}}$$

$$= \frac{\left[1\left(e^{-0.11508(0)}\right) + 0.891\left(e^{-0.11508(1)}\right) + 0.708\left(e^{-0.11508(3)}\right) + \ldots + 0.562\left(e^{-0.11508(5)}\right) + 0.447\left(e^{-0.11508(7)}\right) + \cancel{e^{2(-0.11508)(9)}}\ 0.355\left(e^{-0.11509(9)}\right)\right]}{\left[e^{2(-0.11508)(0)} + e^{2(-0.11508)(1)} + e^{2(-0.11508)(3)} + e^{2(-0.11508)(5)} + e^{2(-0.11508)(7)} + e^{2(-0.11508)(9)}\right]}$$

$$= \frac{2.9373}{2.9378}$$

$$= 0.99968$$

The regresson formula is hence given by

$$Y = 0.99983\, e^{-0.11508t}$$

The half life of Technetium-99m is when $Y = \frac{1}{2} Y \Big|_{t=0}$

$$0.99983 \times e^{-0.11508t} = \frac{1}{2}(0.99983)\, e^{-0.11508(0)}$$

$$e^{-0.11508t} = 0.5$$

$$-0.11508t = \ln(0.5)$$

$$t = 6.0232 \text{ hours.}$$

The relative Intensity of the radiation after 24 hours is

$$Y = 0.99983 \times e^{-0.11508(24)}$$

$$= 6.3160 \times 10^{-2}$$

This implies that only $\frac{6.3160 \times 10^{-2}}{0.99983} \times 100 = 6.3171\%$ of the initial radioactive intensity is left after 24 hrs

GROWTH MODEL

Growth models common in scientific fields have been developed and used successfully for specific situations.
They are used to describe how something grows with changes in the regressor variable (often the time).
Examples include growth of thin films or population with time.
Growth models include:

$$y = \frac{a}{1+be^{-cx}}$$

where a, b and c are constants. At $x=0$, $y=\frac{a}{1+b}$, and as $x \to \infty$, $y \to a$.
The residuals at each data point x_i are

$$E_i = y_i - \frac{a}{1+be^{-cx_i}}$$

The sum of the squares of residuals is

$$S_r = \sum_{i=1}^{n} E_i^2$$

$$= \sum_{i=1}^{n} \left(y_i - \frac{a}{1+be^{-cx_i}} \right)^2$$

To find the constants a, b and c we minimize S_r by differentiating w.r.t. a, b and c and equating the resulting equations to zero.

$$\frac{\partial S_r}{\partial a} = \sum_{i=1}^{n} \left(\frac{2e^{cx_i}\left[ae^{cx_i} - y_i(e^{cx_i}+b)\right]}{(e^{cx_i}+b)^2} \right) = 0$$

$$\frac{\partial S_r}{\partial b} = \sum_{i=1}^{n} \left(\frac{2ae^{cx_i}\left[by_i + e^{cx_i}(y_i-a)\right]}{(e^{cx_i}+b)^3} \right) = 0$$

$$\frac{\partial S_r}{\partial c} = \sum_{i=1}^{n} \left(\frac{-2abx_ie^{cx_i}\left[by_i + e^{cx_i}(y_i-a)\right]}{(e^{cx_i}+b)^3} \right) = 0$$

One can use the Newton-Raphson method to solve the above set of simultaneous nonlinear equations for a, b and c.

Example 2

The height of a child is measured at different ages as follows

t (years)	0	5.0	8	12	16	18
H (in m)	20	36.2	52	60	69.2	70

Estimate the height of the child as an adult of 30 yrs of age using the growth model. $H = \frac{a}{1+be^{-ct}}$

Solution

The saturation growth model of height, H vs age, t is given as

$$H = \frac{a}{1+be^{-ct}}$$

where a, b, and c are roots of the simultaneous non-linear equation system.

Initial guesses of roots: suppose we use three of the given data points such as (0, 20), (2, 60) and (18, 70) to find the initial guesses of roots, we have

$$20 = \frac{a}{1+be^{-c(0)}}$$

$$60 = \frac{a}{1+be^{-c(12)}}$$

$$70 = \frac{a}{1+be^{-c(18)}}$$

One can solve three unknowns a, b and c from the three equations as

$a = 7.5534 \times 10^{1}$

$b = 2.7767$

$c = 1.9772 \times 10^{-1}$

Applying the Newton-Raphson method for simultaneous non-linear equations, one can get the roots.

$a = 7.4321 \times 10^{1}$, $b = 2.8233$, $c = 2.1715 \times 10^{-1}$

The saturation growth model of the height of the child then is

$$H = \frac{7.4321 \times 10^{1}}{1+2.8233e^{-2.1715 \times 10^{-1} t}}$$

The height of the child as an adult of 30 years of age is

$$H = \frac{7.4321 \times 10^{1}}{1+2.8233e^{-2.1715 \times 10^{-1} \times 30}} = 74'$$

In the logistic growth model, an example of a growth model in which a measurable quantity y varies with some quantity x is

$$y = \frac{ax}{b+x}$$

for $x = 0$, $y = 0$ while as $x \to \infty$, $y \to a$; To linearize the data for this method,

$$\frac{1}{y} = \frac{b+x}{ax}$$

$$= \frac{b}{a}\frac{1}{x} + \frac{1}{a}$$

let $z = \frac{1}{y}$ $w = \frac{1}{x}$, $a_0 = \frac{1}{a}$ implying that $a = \frac{1}{a_0}$

$a_1 = b/a$ implying $b = a_1 \times a = \frac{a_1}{a_0}$

Then $z = a_0 + a_1 w$

The relationship btn z and w is linear with the coefficients a_0 and found as follows

$$a_1 = \frac{n\sum_{i=1}^{n} w_i z_i - \sum_{i=1}^{n} w_i \sum_{i=1}^{n} z_i}{n\sum_{i=1}^{n} w_i^2 - \left(\sum_{i=1}^{n} w_i\right)^2}$$

$$a_0 = \left(\frac{\sum_{i=1}^{n} z_i}{n}\right) - a_1\left(\frac{\sum_{i=1}^{n} w_i}{n}\right)$$

Finding a_0 and a_1, then give the constants of the original growth model as $a = \frac{1}{a_0}$, $b = \frac{a_1}{a_0}$

Power Functions

The power function equation describes many scientific and engineering phenomena. In chemical engineering, the rate of chemical reaction is often written in power function form as $y = ax^b$
The method of least squares is applied to the power function by first linearizing the data (the assumption is that b is not known).

If the only unknown is a, then a linear relation exists between x^n and y. The linearization of the data is as follows:

$$\ln(y) = \ln(a) + b\,\ln(x)$$

The resulting eqn shows a linear relation between $\ln(y)$ and $\ln(x)$

Let $z = \ln y$ $w = \ln(x)$ $a_0 = \ln a$ implying $a = e^{a_0}$ $a_1 = b$

We get

$$z = a_0 + a_1 w$$

$$a_1 = \frac{n\sum_{i=1}^{n} w_i z_i - \sum_{i=1}^{n} w_i \sum_{i=1}^{n} z_i}{n\sum_{i=1}^{n} w_i^2 - \left(\sum_{i=1}^{n} w_i\right)^2}$$

$$a_0 = \frac{\sum_{i=1}^{n} z_i}{n} - a_1 \frac{\sum_{i=1}^{n} w_i}{n}$$

Since a_0 and a_1 can be found, the original constant of the model are $b = a_1$, $a = e^{a_0}$

Example

The progress of a homogeneous chemical reaction is followed and it is desired to evaluate the rate constant and the order of the reaction. The rate law expression for the reaction is known to follow the power function form

~~–~~ $-r = kC^n$.

Use the data provided in the table to obtain n and k.

C_A (gmol/l)	4	2.25	1.45	1.0	0.65	0.25	0.006
$-r_A$ (gmol/l.s)	0.398	0.298	0.238	0.198	0.158	0.098	0.048

Solution

Taking the ln of both sides we obtain

$$\ln(-r) = \ln(k) + n\ln(c)$$

Let

$z = \ln(-r)$

$w = \ln(c)$

$a_0 = \ln(k)$ implying that $k = e^{a_0}$

$a_1 = n$

$z = a_0 + a_1 w$

$= e^{a_0}$

we get $z = a_0 + a_1 w$

This is a linear relation between z and w, where

$$a_1 = \frac{n\sum_{i=1}^{n} w_i z_i - \sum_{i=1}^{n} w_i \sum_{i=1}^{n} z_i}{n\sum_{i=1}^{n} w_i^2 - \left(\sum_{i=1}^{n} w_i\right)^2}$$

$$a_0 = \left(\frac{\sum_{i=1}^{n} z_i}{n}\right) - a_1\left(\frac{\sum_{i=1}^{n} w_i}{n}\right)$$

i	C	-r	w	z	wz	w^2
1	4	0.398	1.3863	-0.9218	-1.2778	1.9218
2	2.25	0.298	0.8109	-1.2107	-0.9818	0.65761
3	1.45	0.238	0.3716	-1.4355	-0.5334	-0.13806
4	1	0.198	0.0000	-1.6195	0.0000	0.0000
5	0.65	0.158	-0.4308	-1.8452	0.7949	0.18557
6	0.25	0.098	-1.3863	-2.3228	3.2201	1.9218
7	0.006	0.048	-5.1160	-3.0366	15.535	26.173
$\sum_{i=1}^{7}$			-4.3643	-12.391	16.758	30.998

$n=7$, $\sum w_i = -4.3643$ $\sum z_i = -12.391$ $\sum w_i z_i = 16.758$

$\sum w_i^2 = 30.998$

$$a_1 = \frac{7\times 16.758 - (-4.3643)\times(-12.391)}{7\times 30.998 - (-4.3643)^2}$$

$= 0.31943$

$$a_0 = \frac{-12.391}{7} - (0.31943)\frac{-4.3643}{7}$$

$= -1.5711$

$k = e^{-1.5711}$

$= 0.20782$

$n = a_1 = 0.31941$

Finally the model of progress of that chemical reaction is

$-r = 0.20782 \times C^{0.31941}$

Non - Parametric Regression

Introduction

A general non-parametric regression model may be defined as

$$Y_j = m(X_j) + \varepsilon_j \qquad , j = 1, 2, \dots, N$$

$$X_j \, , Y_j \in \mathbb{R}$$

$$\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N \sim \text{iid}$$

$$E(\varepsilon_j) = 0 \, , \; Var(\varepsilon_j) = \sigma^2 < \infty$$

The problem is to estimate the function $m(X_j)$ from the sample $(X_1, Y_1), (X_2, Y_2), \dots, (X_N Y_N)$

If $X_1, X_2, \dots, X_N$ are random and independent of $\varepsilon_1, \varepsilon_2, \dots \varepsilon_N$ then

$$m(x) = E(Y_j / X_j = x)$$

If $m(X)$ is smooth, then $m(x)$ locally approximates a constant, that means, we estimate $m(x)$ by the local average

$$m(x) = \frac{1}{N_x} \sum_{j=1}^{N} \pi_{[x-h, x+h]}(X_j)$$

More generally, local weights are usually assigned to Y_i in such a way that they decrease with the distance of X_j from x. The General form of the weighted local average is sometimes given by

$$\frac{1}{N_x} \sum_{j=1}^{n} W_{N_j}(x) Y_j$$

where the weights $W_{N_j}(x)$ is large if $|x - x_j|$ is small. and they may depend on all $X_1, X_2, \dots, X_n$ simultaneously.

Kernel Estimator

A Kernel function, K is a bounded continuous function on the real number line $\mathbb{R}$ satisfying $\int K(u)\,du = 1$.

$K_h(u) = \frac{1}{h} K(u/h)$ this being a re-scaled notation of the kernel still satisfying the $\int K_h(u)\,du = 1$

If $\sup(K) = [-1,1]$ then $\sup(K_h) = [h,h]$ is called the bandwidth or smoothing parameter.

Model 1

Deterministic Equivalent Design

The model $Y_j = m(x_j) + \varepsilon_j$ is said to be deterministic if $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N$ are iid $\sim(0, \sigma^2)$ and $x_j = \frac{1}{N}$ $j = 1, 2, \dots, N$

The PC Kernel estimate of $m(x_j) : [0,1] \to \mathbb{R}$ with bandwidth $h > 0$ is given by

$$m(x,h) = \frac{1}{N}\sum_{j=1}^{N} K_h(x - x_j)\, Y_j$$

$$= \frac{1}{Nh}\sum_{j=1}^{N} K\left[\frac{x - x_j}{h}\right] Y_j, \quad x \leftarrow [0,1]$$

The corresponding weights in this case become

$$W_{Nj}(x) = K_h(x - x_j)$$

Model 2

Stochastic Design

The model $Y_j = m(X_j) + \varepsilon_j$ is said to be stochastic if $X_1, X_2, \dots, X_N$ are iid random variables with density $\underline{p(x)}$, and $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N$ are iid $\sim (0, \sigma^2)$ and independent of $X_1, X_2, \dots, X_N$.

Kernel fn - When integrated over the whole space you get 1.

a) The Rosenblatt-Parzen kernel density ~~function~~ estimate with bandwidth $h > 0$ is given as

$$\hat{p}(x,h) = \frac{1}{N}\sum_{j=1}^{n} K_h(x - X_j)$$

b) The Nadaraya-Watson (NW) kernel estimate is given by

$$\hat{m}(x,h) = \frac{1}{N}\sum_{j=1}^{n} K_h(\cancel{x}\, x - X_j)\frac{Y_j}{p(x,h)}$$

The corresponding weights

$$W_{Nj}(x) = K_h(x - X_j) / \hat{p}(x,h)$$

Types of Kernel Functions

1) Gauss - Kernel

This is given by

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

2) Epanechnikov (Bartlett Priestly) Kernel

This is given by

$$K(u) = \frac{3}{4}(1 - u^2)$$

The Nearest Neighbour estimate

Let $X_1, X_2, \ldots, X_N$ be iid R.V. with densities $P(x)$, for fixed $x \in \mathbb{R}$ let $d_1(x) \leq d_2(x) \leq \ldots \leq d_N(x)$ be ordered distances $|x - X_j|$ $j = 1, 2, \ldots, N$.

If $|x - X_K = d_K(x)|$, then X_K is called the K^{th} ~~de~~ nearest neighbour of X

Let K be a kernel function, the K nearest neighbour estimate ie K-NN of the density $P(x)$ is ~~d~~ given by.

$$\hat{P}_K(x) = \frac{1}{N d_K(x)} \sum_{j=1}^{N} K\left[\frac{x - X_j}{d_K(x)}\right]$$

$\hat{p}(x)$ is a kernel estimate with bandwidth $d_K(x)$ adapting itself to the number of data lying close to x.

It is a general smoothing estimate with weights

$$W_{Nj}(x) = \frac{1}{d_K(x)} K\left[\frac{x - X_j}{d_K(x)}\right]$$

$\pi_{(-1,+1)}$

$\pi \Rightarrow$ Indicator function

For the rectangular Kernel ie

$$K(u) = \frac{1}{2}\pi_{(-1,+1)}(u)$$

we relinquish the continuity of K and get a special case

$$\hat{p}(x) = \frac{k-1}{2N\,d_k(x)}$$

Note
$2d_k(x)$ is the length of the interval ~~x~~.
$(x - d_k(x), x + d_k(x))$ and $k-1$ is the number of observations in the interval.

* Consider Model 2 (Stochastic design) again;
Let K be a kernel, then

$$\hat{m}_k(x) = \frac{1}{N d_k(x)} \sum_{j=1}^{N} K\left[\frac{x - X_j}{d_k(x)}\right] \frac{Y_j}{\hat{p}_k(x)}$$

is called the k-NN estimate for the regression function $m(x)$.
~~$\therefore$~~ $\hat{m}_k(x)$ is a Nadaraya-Watson (NW) kernel estimate with bandwidth $d_k(x)$.
For the rectangular Kernel, we get a particular simple form reminiscent of the PC-kernel estimate as

Solely for rectangular Kernel

$$\hat{p}_k(x) = \frac{k-1}{2N d_k(x)}$$

$$\hat{m}_k(x) = \frac{1}{k-1} \sum_{j=1}^{N} \pi_{(-1,+1)}\left[\frac{x - X_j}{d_k(x)}\right] Y_j$$

$$= \frac{1}{k-1} \sum_{j=1}^{N} Y_j\, \pi_{(x-d_k(x),\, x+d_k(x))}(X_j)$$

In this case $\hat{m}_K(x)$ is just the average of Y_j corresponding to the $K-1$ nearest neighbors of X_j to X.

Example

X	11	22	33	44	60	56	67	70	78	89	90	100
Y	2337	2750	2301	2500	1700	2100	1100	1750	1000	1642	2000	1932

Use the rectangular Kernel function to find the K-NN estimate of the density $p(x)$ and regression function $m(x)$ where $x = 75$ and $K = 4$

4-NN of 75 to estimate

Solution

$$\hat{p}(x) = \frac{K-1}{2N d_K(x)} \qquad \hat{m}(x) = \frac{1}{K-1} \sum_{j=1}^{n} Y_j \, \Pi_{[x-d_K(x),\, x+d_K(x)]}(X_j)$$

$d_K(x)$ = Absolute values of differences from NN and what we are estimating

$N = 12$

$|78-75| = 3 \qquad d_1(x) = 3$

$|70-75| = 5 \qquad d_2(x) = 5$

$|67-75| = 8 \qquad d_3(x) = 8$

$|89-75| = 14 \qquad d_4(x) = \underline{14}$

$\underline{K=4}$

$$\hat{p}(x) = \frac{4-1}{2(12)14} = \frac{3}{336} = \frac{1}{112}$$

75 − 14
75 + 14

$\sum Y_j \, \Pi_{(61,89)}(X_j)$

Open Intervals Don't include the values not within

$= 1100 + 1750 + 1000 = 3850$

$\hat{m}(x) = \frac{1}{3} \times 3850$

$= 1283.33$

[5c/17]

2 $X = 49$, $k = 5\,4$

$N = 12$ 50, 44, 56, 33, 67

$d_k(x) = |67 - 49| = 18$

$$\hat{P}(x) = \frac{4}{2(12) \times 18} = \frac{4}{432} = \frac{1}{108}.$$

$$\Sigma Y_j \, \Pi_{[x - d_k(x) \, . \, x + d_k(x)]}(X_j)$$

$\Pi_{(31, 67)} = X_j$

49-18 $2301 + 2500 + 1700 + 2100 = 8601$

$\hat{m}(x) = \frac{1}{4} \times 8601 = 2150 \cdot 25$

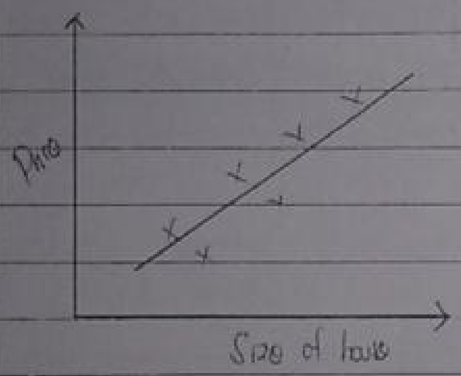
Neural Networks

A neural network is a non-linear system transforming real input variables denoted $x_1, x_2, \ldots, x_p$ into one or several output variables denoted $y_1, y_2, \ldots, y_q$ using several intermediate steps.

Example

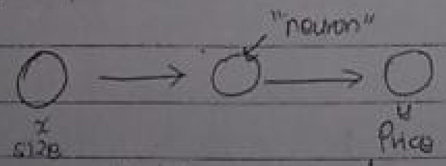
Consider an example where we have to predict the price of a house. Variables given are the size of the house in sq ft and the price of the house.

Assume we have 6 houses



A linear regression line will draw a straight line to fit the data

Using simple neural network



The neuron will take an input, apply some activation function to it and generate an output

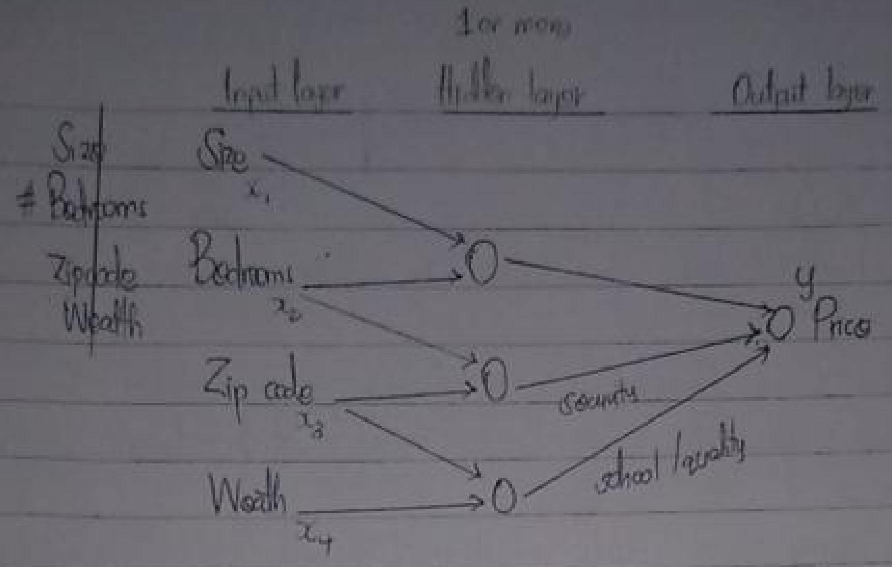
One of the most commonly used activation function is the rectified linear unit ReLU

The ReLU takes a real number as an input and returns a maximum of zero or that number.

If we pass 10, the output will be 10

-10, the output will be 0

We have seen a neural network with a single neuron but in reality we have to consider multiple features like



In the input layer, each node represents one input variable and analogously in the output layer each node represents one output variable. In between, there are one or several hidden layers, the nodes of which are neither sources nor sinks of the graph.
The network above has only one hidden layer. Additionally, it is a feed-forward network as all the edges which originate from one node end up in a node of one of the subsequent layers never in a node of the same or one of the previous layers.

Assignment (To be submitted on 25th)

1. What are linear mixed models?
2. State and explain the assumptions made by linear mixed models
3. Define what is meant by a spline and illustrate using examples. Show that splines are linear smoothers
4. Show that a linear regression estimator is a special case of a non-parametric estimator

Methods for analysing data that is non-independant, longitudinal and correlated.

Solution,

Linear Mixed models:→

These are an extension of simple linear models that allow fixed and random effects, which are especially used where there is non independence in the data.

Assumptions made by linear mixed models

The explanatory variables have a linear relationship with the response variables

The error terms have contant variance.

There is independence of error terms.

Error terms are normally distributed.

The dependant variable is quantitive.

The model

$Y = X\beta + Z\mu + \varepsilon$

Y - dependant term / vector of the respondant

X - Independant / design matrix of fixed effects.

β - Unknown vector of fixed effects parameters to be estimated.

Z - Matrix for the random effects

μ - Unknown effects of the random parameters.

ε - Random / Unobserved error

REVISION

1. Assumptions for OLS estimators to be BLUE

Linearity - The linear regression model is linear in parameters

Unbiasedness

Efficiency

2. $Y_i = \beta_1 + \beta_2 X_i + u_i$, Suppose you multiply each X_i value by a constant. Will it change the residuals and fitted values of Y_i? Explain.

3. What is the interpretation of an estimated coefficient from a simple linear regression model and that from a logistic regression model?

Simple ⇒ $Y_i = \beta_0 + \beta_1 X_i + u_i$

Response Explanatory

Holding all other terms constant

β_0 - intercept β_1 - slope ⇒ for every unit change in X, Y increases or decreases by β_1.

error

Holding all other variables constant, $Y = \beta_0$

Logistic Regression Model

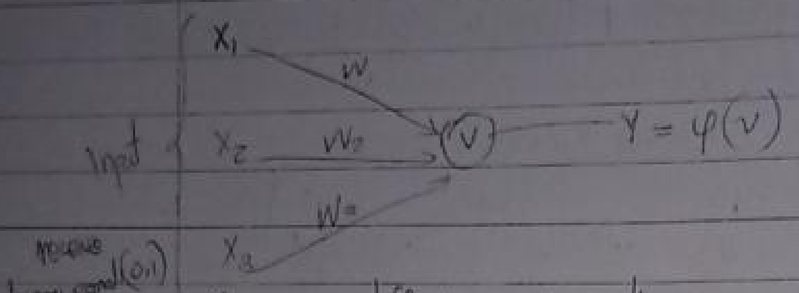
How could we judge a multiple linear regression model has a better performance in terms of statistical information

R^2

↑R^2 ⇒ better model performance

Root mean squared error

Measures difference between predicted values and actual values



How many different input patterns can this node receive?

What if the node had 4 inputs? Five? Can you give a formula that computes the number of binary input patterns for a given number of inputs?

X_1	0	1	0	0	1	0	1	1
X_2	0	0	1	0	1	1	0	1
X_3	0	0	0	1	0	1	1	1

X_1	0	1	0	0	0	1	0	0	1	0	0	1	0	1	1	1
X_2	0	0	1	0	0	1	1	0	0	1	1	1	1	1	0	1
X_3	0	0	0	1	0	0	1	1	0	0	1	1	1	0	1	1
X_4	0	0	0	0	1	0	0	1	1	1	0	0	1	1	1	1

$3 = 8 = 2^3$

$4 = 16 = 2^4$

$5 = 2^5 =$

2^n

let $p=2$ and $x_1, x_2 \in \{0,1\}$, the classification to be learned is the logical XOR $z=1$ if $x_1=1$ or $x_2=2$

$z=0$ if $x_1=0$ and $x_2=0$.

$$x^{(1)} = \begin{pmatrix}1\\1\\0\end{pmatrix} \quad x^{(2)} = \begin{pmatrix}1\\0\\1\end{pmatrix} \quad x^{(3)} = \begin{pmatrix}1\\1\\1\end{pmatrix} \quad x^{(4)} = \begin{pmatrix}1\\0\\0\end{pmatrix}$$

with correct classification $z^{(1)} = z^{(2)} = z^{(3)} = 1$, $z^{(4)} = 0$. The perceptron with weights w_0, w_1, w_2 classifies an object as 1 iff

$W^T x = w_0 x_1 + w_1 x_1 + w_2 x_2 > 0$ and as 0 elsewhere. As initial vector we use $w = (0,0,0)^T$ and we set $\eta = 1$.

The steps of the delta algorithm are now

1. $y^{(1)} = w^T x(1) = [0,0,0]\begin{pmatrix}1\\1\\0\end{pmatrix} = 0 \neq z^{(1)}$, $x^{(1)}$ is classified incorrectly

~~The~~

The weights are modified

$W_{new} = (0,0,0)^T + (1-0)(1,1,0)^T = (1,1,0)^T$

2 $y^{(2)} = W^T x^{(2)} = (1,1,0)\begin{pmatrix}1\\0\\1\end{pmatrix} = 1 = z^{(2)}$, $x^{(2)}$ is correctly classified.

3. $W^T x^{(3)} = (1,1,0)\begin{pmatrix}1\\1\\1\end{pmatrix} = 2 \equiv 1 = z^{(3)} = 1$, $x^{(3)}$ is correctly classified.

4 $W^T x^{(4)} = y^{(4)} = (1,1,0)\begin{pmatrix}1\\0\\0\end{pmatrix} = 1 \neq z^4 = 0$, $x^{(4)}$ is classified incorrectly

The weights are modified again

$W_{new} = (1,1,0)^T + (0-1)(1,0,0)^T = (0,1,0)^T$

5. $y^{(1)} = w^T x(1) = (0,1,0)\begin{pmatrix}1\\1\\0\end{pmatrix} = 1 = z^{(1)}$, $x^{(1)}$ is classified correctly

6 $y^{(2)} = W^T x^{(2)} = (0,1,0)\begin{pmatrix}1\\0\\1\end{pmatrix} = 0 \neq z^{(2)}$, $x^{(2)}$ is classified incorrectly

$\vdots$

Locally Weighted Regression
Uses data from a neighborhood around the specific location, so that the neighborhood is defined as a span, which is a function of the total points used to form the neighborhood.

Linear Regression - A linear approach to modelling the relationship between a scalar and response and one or more explanatory variables.

Non-linear regression - A form of regression analysis in which observational data are modelled by a function which has non-linear model parameters and depends on one or more independent variable

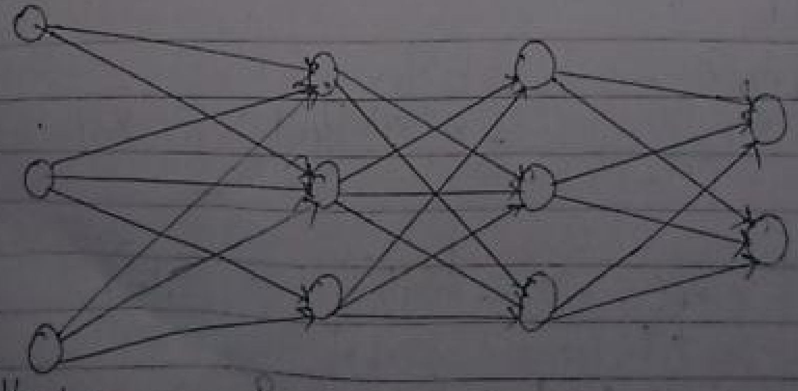
Parametric regression - Assumes some; finite set of parameters. Involves a model that ~~involves~~ assumes a ~~[illegible]~~ some finite set of parameters with certain predetermined assumptions about the data.

Non-parametric regression - Regression analysis where the predictor does not take a predetermined form but is constructed according to information derived from the data.

Neural network - A non-linear system transforming real input variables denoted by $X_1, X_2, \ldots, X_p$ into one or several output variables denoted $Y_1, Y_2, \ldots, Y_q$ using intermediate steps

Hidden layer : A layer between input and output layers, where artificial neuron's take in a set of weighted input and produce an output through an activation function.

Example :



Kernel regression - Regression that estimates the continuous dependent variable from a limited set of data points by convolving the data points to [illegible]

$\frac{1}{x-1}$ [illegible]

[illegible]

11) Suppose that a credit card company decided to deploy a new system for assessing credit worthiness of its customers.. suggest in a form of essay what should the bank have before the system can be used? Discuss the problem associated with this

The company should get hold of historical data about its customers who already took credit in the past.

This data will be used as a training set for the neural network.

It is important that the data is representative & covers as many types of customers as possible. This is ~~so~~ because the network will not be able to produce an accurate answer for a customer very different from the training set . .

We have a sample (X_i, Y_i), $i = 1, 2, \ldots, n$ from a NPR model $Y_i = m(X_i) + \varepsilon_i$ where $E[\varepsilon_i | X_1, X_2, \ldots, X_n) = 0$ and $E(\varepsilon_i^2 | X_1, \ldots, X_n) = \sigma^2 X_i$ and $X_1, \ldots, X_n$ has a pdf $f(x)$. Moreover, $E(\varepsilon_i, \varepsilon_j | X_1, \ldots, X_n) = 0$ for $i \neq j$

Define

$$C_1(h) = -\frac{2}{n} \sum_{j=1}^{n} [\varepsilon_j | \hat{m}_h(X_j) - m(X_j)], \quad C_2(h) = -\frac{2}{n} \sum_{j=1}^{n} \varepsilon_j [\hat{m}_h^{(-j)}(X_j) - m(X_j)]$$

where ~~*~~ $\hat{m}_h(X_j)$ is the N-W estimator of $m(x)$ at X_j ~~obtained~~ and $\hat{m}_h^{(-j)}(X_j)$ is the N-W estimator of $m(x)$ at X_j obtained using all the data except (X_j, Y_j)

a) Find the assymptotic expression ~~Expression~~ of $E(C_1(h) | X_1, \ldots, X_n)$

Solution

$$E(\varepsilon_j \{\hat{m}_h(X_j) - m(X_j)\} | X_1, \ldots, X_n)$$

$$= E\left(\varepsilon_j \left\{ \frac{\sum_{i=1}^{n} K_h(X_i - X_j) Y_i}{\sum_{i=1}^{n} K_h(X_i - X_j)} - m(X_j) \right\} \Big| X_1, \ldots, X_n\right)$$

$$= E\left(\varepsilon_j \left\{ \frac{\sum_{i=1}^{n} K_h(X_i - X_j)(Y_i - m(X_j))}{\sum_{i=1}^{n} K_h(X_i - X_j)} \right\} \Big| X_1, \ldots, X_n\right)$$

$$= \frac{\sum_{i=1}^{n} K_h(X_i - X_j) E((Y_i - m(X_j)) \varepsilon_j | X_1, \ldots, X_n)}{\sum_{i=1}^{n} K_h(X_i - X_j)}$$

$$= \frac{K_h(0)\sigma^2(X_j)}{\sum_{i=1}^{n} K_h(X_i - X_j)} = \frac{K(0)\sigma^2(X_j)}{nh\hat{f}(X_j)}$$

We have a sample (X_i, Y_i), $i = 1, 2, \ldots, n$ generated from the NPR model $Y_i = m(X_i) + \varepsilon_i$, $i = 1, 2, \ldots, n$. where $m(x)$ is an unknown smooth function and $X_1 < X_2 < \ldots < X_n$. Let $Y = [Y_1, \ldots, Y_n]^T$ be the response vector and $\hat{Y} = [\hat{Y}_1, \ldots, \hat{Y}_n]^T$ be estimator $\hat{m}(x)$. When $\hat{m}(x)$ is a linear smoother, we have $\hat{Y} = AY$ where A is known as the associated smoother matrix and df = trace (A) is known the associated d.f, meaning how complex the fitting model

a) First assume that $\hat{m}(x)$ is the usual regression spline smoother constructed based on the pth order truncated power basis $\Phi(x)$ using K distinct knots $T_1 < T_2 < \ldots < T_K$. When $n > k + p$ and p fixed show that df will increase as K increases.

Proof.

Let $X = [\Phi(X_1), \ldots, \Phi(X_n)]^T$, we have $\hat{Y} = X(X^TX)^{-1}X^TY$ and hence $A = X(X^TX)^{-1}X^T$. It follows that df = trace(A) = $p + 1 + K$ which increases as K increases.

b) Assume now that $\hat{m}(x)$ is the N-W estimator using a bandwidth $h > 0$ and a symmetric kernel $K(\cdot)$ which is a pdf. Show that when K is fixed and n is sufficiently large, df will decrease as h increases

Proof.

Notice that $Y_i = \hat{m}(X_i) = \sum_{j=1}^{n} \frac{K_h(X_j - X_i)Y_j}{n\hat{f}(X_i)}$

we have df = trace (A) $= \sum_{i=1}^{n} K_h(0) / (n\hat{f}(X_i)) = \frac{K(0)(b-a)}{h}[1 + o(1)]$

when n is sufficiently large. It follows that df will decrease as h increases.

$$= \frac{K_h(0)\sigma^2(X_j)}{\sum_{i=1}^{n} K_h(X_i - X_j)} = \frac{K(0)\sigma^2(X_j)}{nh\hat{f}(X_j)}$$

We have a sample (X_i, Y_i), $i = 1, 2, \ldots, n$ generated from the NPR model $Y_i = m(X_i) + \varepsilon_i$, $i = 1, 2, \ldots, n$. where $m(x)$ is an unknown smooth function and $X_1 < X_2 < \ldots < X_n$. Let $Y = [Y_1, \ldots, Y_n]^T$ be the response vector and $\hat{Y} = [\hat{Y}_1, \ldots, \hat{Y}_n]^T$ be estimator $\hat{m}(x)$. When $\hat{m}(x)$ is a linear smoother, we have $\hat{Y} = AY$ where A is known as the associated smoother matrix and $df = \text{trace}(A)$ is known the associated d.f, measuring how complex the fitting model

a) First assume that $\hat{m}(x)$ is the usual regression spline smoother constructed based on the pth order truncated power basis $\Phi(x)$ using K distinct knots $\tau_1 < \tau_2 < \ldots < \tau_K$. When $n > k + p$ and p fixed Show that df will increase as K increases.

Proof.

Let $X = [\Phi(X_1), \ldots, \Phi(X_n)]^T$, we have $\hat{Y} = X(X^TX)^{-1}X^TY$ and hence $A = X(X^TX)^{-1}X^T$. It follows that $df = \text{trace}(A) = p + 1 + K$ which increases as K increases.

b) Assume now that $\hat{m}(x)$ is the N-W estimator using a bandwidth $h > 0$ and a symmetric kernel $K(\cdot)$ which is a pdf Show that when K is fixed and n is sufficiently large, df will decrease as h increases

Proof.

Notice that $\hat{Y}_i = \hat{m}(X_i) = \sum_{j=1}^{n} \frac{K_h(X_j - X_i)Y_j}{n\hat{f}(X_i)}$

we have $df = \text{trace}(A) = \sum_{i=1}^{n} K_h(0) / (n\hat{f}(X_i)) = \frac{K(0)(b-a)}{h}[1 + o(1)$

when n is sufficiently large It follows that df will decrease as h increases.

c Assume now that $\hat{m}(x)$ is the cubic smoothing spline smoother with a smoothing parameter λ Show that df will decrease as λ increases

Prof

Since the design th time points $x_1, x_2, \ldots, x_n$ are distinct, we have $\hat{Y} = (I_n + \lambda G)^{-1} Y$ where G is the associated roughness matrix, which is non-negative and has a singular value decomposition as $G = UDU^T$ where U is an orthonomal matrix, containing all eigenvectors of G and D is a diagonal matrix, with diagonal entries as the eigen values of G. It follows that

$$df = \text{trace}(A) = \text{trace}\left((I_n + \lambda G)^{-1}\right) = \text{trace}\left((I_n + \lambda D)^{-1}_i\right) = \sum_{i=1}^{n} (1 + \lambda dt)^{-1}$$