ID5059 L04 - loss functions and model fitting

C. Donovan

Today

- Model fitting
- Loss functions
- Continuous response example
- Minimising these - analytic and algorithmic search

My favourite equation

A model structure that will recur throughout this course (and almost every statistics course) is the apparently simple:

$$\mathbf{y} = f(\mathbf{X}, \boldsymbol{\theta}) + \mathbf{e}$$

how do we find (the best) $\boldsymbol{\theta}$?.

This is *model fitting*: given f and some data, what are the best $\boldsymbol{\theta}$?

Loss functions - continuous response

Mean Squared Error loss MSE (like Ordinary Least Squares - OLS) - the i-th *error* $= y_i - \hat{y}_i$, so:

$$\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$$

NB $\hat{y}$ comes from our fitted function i.e. $\hat{f}$ - so depends on/varies with $\boldsymbol{\theta}$, so equally ($\mathbf{x}_i$ being the i-th row of our $\mathbf{X}$ matrix):

$$\frac{1}{n}\sum_i (y_i - \hat{f}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}))^2$$

Roughly speaking, How close on average are our model predictions to the observed response?

Loss functions - continuous response

closely related, Mean Absolute Error loss MAE - we just don't square the error.

$$\frac{1}{n}\sum_i |y_i - \hat{y}_i|$$

Squaring errors has nice mathematical/theoretical properties, but makes large individual errors relatively larger.

Linear models

If f has this form (i.e. can define $\mathbf{X}$):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

then there is an analytical solution (multiple linear regression, analysis of variance, analysis of covariance, t-tests, polynomial regression).

Linear models

At the i-th observation we have ((again $\mathbf{x}_i$ being the i-th row of our $\mathbf{X}$)

$$\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$$

In the well-known case of Residual Sum of Squares we choose

$$RSS(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Differentiation gives

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

with unique solution (if $\mathbf{X}^T\mathbf{X}$ is nonsingular)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Parameter searches

Analytic solutions don't exist for all models i.e. a formula for the "best" $\boldsymbol{\theta}$, nor are they necessarily the fastest approach.

More generally we search the *parameter space*

- Grid search - inefficient
- Gradient search - varying computational approaches

Gradient search

- Gradient descent (use all the data to choose the direction to jump)
- Stochastic gradient descent (use some sample of the data to choose the direction to jump)

<drawing ensues: 1D & 2D>

Gradient search

- Initial starting point
- Learning rates/step sizes
- Convexity, local minima and global minima

For well-behaved Loss functions (like MSE or likelihoods for a linear or generalised linear model) these are easy. Others can be *veeery* computationally difficult/intense e.g. neural networks

Keywords and Reading

- **training error, loss functions, linear model, model fitting, gradient descent, stochastic gradient descent**
- James *et al*: Section 3.1/3.1.1
- HT&F: page 12, Section 3.2
- Geron: page 112-127