Model Selection and Evaluation

Dr Rafael de Andrade Moral
Associate Professor of Statistics, Maynooth University

rafael.deandrademoral@mu.ie
https://rafamoral.github.io

Outline

- Measuring model fit
- Nested model comparisons
- Out-of-sample predictive performance
- Cross-validation and information criteria
- Model averaging
- Variable selection
- Bayesian model comparison methods

Measuring Model Fit

Measuring Model Fit

- In any statistical analysis, we assume our data is drawn from some probability distribution
- This is sometimes known as the *probabilistic generative model*, and in fact is exactly what we mean by the statistical model
- This is a model of the statistical population, which could also be described as the true generative model (although we will never know the *true* process)
- We aim to find a good (or good enough) model of the population

Measuring Model Fit

- One general way we can evaluate a model is by asking if the data is compatible with the model
- One way to look at this is to calculate the probability of the data according to the model
- If the probability of observing the data is relatively high in one model than in another, then the data is more compatible with the former than the latter model
- We often refer to the probability of the data according to the model as the model's *likelihood*

A Very Small Example

- Imagine we are studying the height of human populations
- Let's sample four heights (in cm) at random from this group:

180	191	162	171

- We now assume a model for the population – say we assume the normal model:

$$Y_i \sim \mathsf{N}(\mu, \sigma^2), \quad i = 1, 2, 3, 4$$

- We don't know the values of μ and σ^2, but we can estimate them based on the sampled data

A Very Small Example

- Say we use $\hat{\mu} = \bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$ and $\hat{\sigma} = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}}$ as our estimates
- We would then be assuming that our population model is

$$Y_i \sim \mathsf{N}(\mu = 176, \sigma^2 = 12.4^2)$$

- What is the *likelihood* that the random sample came from the model above?
- We write is as a function of the data given the model:

$$L(y_1, y_2, \ldots, y_n | \mu, \sigma^2)$$

A Very Small Example

- If we assume samples are independent, then we can write

$$L(y_1, y_2, \ldots, y_n|\mu, \sigma^2) = L(y_1|\mu, \sigma^2) \times L(y_2|\mu, \sigma^2) \times \cdots \times L(y_n|\mu, \sigma^2)$$

- Recall, however, that $L(\cdot)$ is a probability, and therefore is bounded between 0 and 1
- When we multiply many probabilities together we get very small values
- To avoid numerical problems representing such small quantities in a computer, it's easier to work with logarithms
- The *log-likelihood* is $l(\cdot) = \log L(\cdot)$, and therefore we have

$$l(y_1, y_2, \ldots, y_n|\mu, \sigma^2) = l(y_1|\mu, \sigma^2) + l(y_2|\mu, \sigma^2) + \cdots + l(y_n|\mu, \sigma^2)$$

- Another example: the `cars` dataset

Regression Models

- Often, for each observed value of the variable being modelled we have observed values of other variables (known as covariates, predictor variables, independent variables)
- For the `cars` dataset, we may use the `speed` variable to predict `dist`

$$\begin{aligned} Y_i &\sim \mathsf{N}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_i \end{aligned}$$

- In other words, we are modelling `dist` as normally distributed around a mean that is a linear function of `speed`, and with a fixed variance σ^2
- This is exactly a simple linear regression model
- We don't know the values of the parameters β_0, β_1, and σ^2; we must estimate those based on the data

Regression Models

- Given values for β_0, β_1, and σ^2, what is the probability of the observed `dist` values $y_1, y_2, \ldots, y_n$ given the `speed` predictor values $x_1, x_2, \ldots, x_n$?
- This is the likelihood

$$L(y_1, \ldots, y_n | x_1, \ldots, x_n, \beta_0, \beta_1, \sigma^2) = \prod_{i-1}^{n} L(y_i | x_i, \beta_0, \beta_1, \sigma^2)$$

- The log-likelihood of the model is

$$l(y_1, \ldots, y_n | x_1, \ldots, x_n, \beta_0, \beta_1, \sigma^2) = \sum_{i-1}^{n} \log L(y_i | x_i, \beta_0, \beta_1, \sigma^2)$$

Likelihood Ratios

- We have two models of the `dist` variable
- The normal model has a log-likelihood of -232.9, and the regression model has a log-likelihood of -206.6
- Let's denote these log-likelihoods as $\log L_0$ and $\log L_1$, respectively
- The log of the ratio of likelihoods is as follows:

$$\begin{aligned} \log \frac{L_1}{L_0} &= \log L_1 - \log L_0 \\ &= -206.6 - (-232.9) \\ &= 26.3 \end{aligned}$$

- In other words,

$$\frac{L_1}{L_0} = e^{26.3} \approx 270 \text{ billion}$$

Residual Sum of Squares (RSS)

- The RSS is a measure of unexplained variability in the data (*bigger is worse*)
- The RSS is calculated as

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

- Note that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- The RSS is a measure of the model's lack of fit
- Note that, for the normal model,

$$\log L = -\frac{n}{2}\{\log(2\pi) - \log n + \log(\text{RSS}) + 1\}$$

and therefore in two normal models of the same data, the differences in likelihood are determined only by differences in RSS

Root Mean Square Error (RMSE)

- The larger the sample size, the larger the RSS
- An alternative to RSS as a measure of model fit is the square root of the mean of the squared residuals, known as the *root mean square error* (RMSE):

$$\text{RMSE} = \sqrt{\frac{\text{RSS}}{n}}$$

- This is the maximum likelihood estimator for σ, i.e. $\hat{\sigma}_{\text{MLE}} = \text{RMSE}$

Mean Absolute Error (MAE)

- Related to the RMSE is the mean absolute error (MAE), which is the mean of the absolute values of the residuals:

$$\text{MAE} = \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{n}$$

Deviance

- Deviance is used as a measure of model fit in generalized linear models
- Strictly speaking, the deviance of model M_0 is

$$2(\log L_s - \log L_0)$$

, where $\log L_0$ is the log-likelihood (at its maximum) of model M_0, and $\log L_s$ refers to a saturated model, i.e. one with as many parameters as there are data points
- When comparing two models, M_0 and M_1, the saturated model is the same, and so the difference between deviances is

$$(-2 \log L_0) - (-2 \log L_1) = D_0 - D_1$$

, and so the deviance of M_0 is usually defined simply as $-2 \log L_0$

Deviance

- Differences of deviances are equivalent to likelihood ratios:

$$\begin{aligned} D_0 - D_1 &= -2\log L_0 - (-2\log L_1) \\ &= -2(\log L_0 - \log L_1) \\ &= -2\log\frac{L_0}{L_1} \\ &= 2\log\frac{L_1}{L_0} \end{aligned}$$

- $\frac{L_1}{L_0}$ is the factor by which the likelihood of model M_1 is greater than that of model M_0
- Therefore, the difference of the deviance of models M_0 and M_1, $(D_0 - D_1)$, gives the (two times) the logarithm of the factor by the likelihood of model M_1 is greater than that of model M_0
- The larger $D_0 - D_1$, the greater the likelihood of M_1 compared to M_0

Deviance Residuals

- Deviance residuals are values such that their sum of squares is equal to the model's deviance
- For $Y_i \sim \text{Bernoulli}(p_i)$, we have the pmf $P(Y_i = y_i) = p_i^{y_i}(1-p_i)^{1-y_i}$, and therefore the log-likelihood is

$$y_i \log p_i + (1-y_i)\log(1-p_i)$$

- If we multiply it by -2, its sum gives the deviance:

$$-2\{y_i \log p_i + (1-y_i)\log(1-p_i)\}$$

- So the sum of squares of the following expression is the deviance:

$$\sqrt{-2\{y_i \log p_i + (1-y_i)\log(1-p_i)\}}$$

- Since all these values are positive, the deviance residuals are defined as

$$r_i^D = \text{sign}(y_i - \hat{y}_i)\sqrt{-2\{y_i \log p_i + (1-y_i)\log(1-p_i)\}}$$

Nested Model Comparisons

Nested Model Comparisons

- Model M_1 is *nested* in model M_0 if the parameter space of M_1 is a *subset* of the parameter space of M_0
- For example, if M_0 is the following linear model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad \varepsilon_i \sim \mathsf{N}(0, \sigma^2)$$

  its parameter space is $\beta_0, \beta_1, \beta_2, \sigma^2$
- If M_1 is the following linear model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i, \quad \varepsilon_i \sim \mathsf{N}(0, \sigma^2)$$

  its parameter space is $\beta_0, \beta_1, \sigma^2$
- Any set of values of $\beta_0, \beta_1, \sigma^2$ in M_1 is a point in the parameter space $\beta_0, \beta_1, \beta_2, \sigma^2$ of M_0 if we simply set $\beta_2 = 0$
- We say that M_1 is *nested* in M_0

Nested Model Comparisons

- We can compare nested normal linear models using F tests
- We calculate RSS_0 and RSS_1, the residual sums of squares of M_0 and M_1, respectively
- Since M_1 is nested in M_0, $\text{RSS}_1 \geq \text{RSS}_0$
- Then

$$\frac{\text{RSS}_1 - \text{RSS}_0}{\text{RSS}_0}$$

represents the proportional increase in error (change in error from the simpler model to the more complex one divided by the minimal error)

Nested Normal Linear Models

- This quantity is used within an F test statistic:

$$F_{\text{obs}} = \frac{\frac{\text{RSS}_1 - \text{RSS}_0}{\text{df}_1 - \text{df}_0}}{\frac{\text{RSS}_0}{\text{df}_0}}$$

- df_j is the number of residual d.f. for model M_j
- Under the null hypothesis that the two model fits are equivalent, and given the assumptions of the normal models are correct, F_{obs} has an exact $F-$distribution with $\text{df}_1 - \text{df}_0$ and df_0 degrees of freedom

R^2

- If we have model M_1 nested in model M_0, we can calculate the proportional *decrease* in error:

$$\frac{\text{change in error (from } M_1 \text{ to } M_0)}{\text{error in } M_1}$$

- This quantity is also referred to as

$$R^2 = \frac{\text{RSS}_1 - \text{RSS}_0}{\text{RSS}_1}$$

R^2

- It can be shown that

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{ESS}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{RSS}}$$

  where TSS is the *total* sum of squares, ESS is the *explained* sum of squares and RSS is the *residual* sum of squares
- The coefficient of determination R^2 is defined as the proportion of the variation that is explained, i.e.

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

- Therefore, $0 \leq R^2 \leq 1$

Adjusted R^2

- R^2 can be used as a *goodness-of-fit* measure
- However, R^2 will always grow as the number of predictors k grows
- It can be adjusted to counteract the artificial effect of increasing numbers of predictors as follows:

$$R^2_{\text{Adj}} = R^2 \underbrace{\frac{n-1}{n-k-1}}_{\text{penalty}}$$

  where n is the sample size
- R^2_{Adj} decreases as k increases
- R^2_{Adj} is not identical to the proportion of variance explained in the *sample*, but it is an unbiased measure of the population R^2

Model Comparison with Deviance

- Again, assume that model M_1 is nested in model M_0
- Under the null hypothesis, the difference between deviances

$$\Delta_D = D_1 - D_0$$

is asymptotically distributed as χ^2 with $k_0 - k_1$ d.f., where k_j is the number of parameters in model M_j

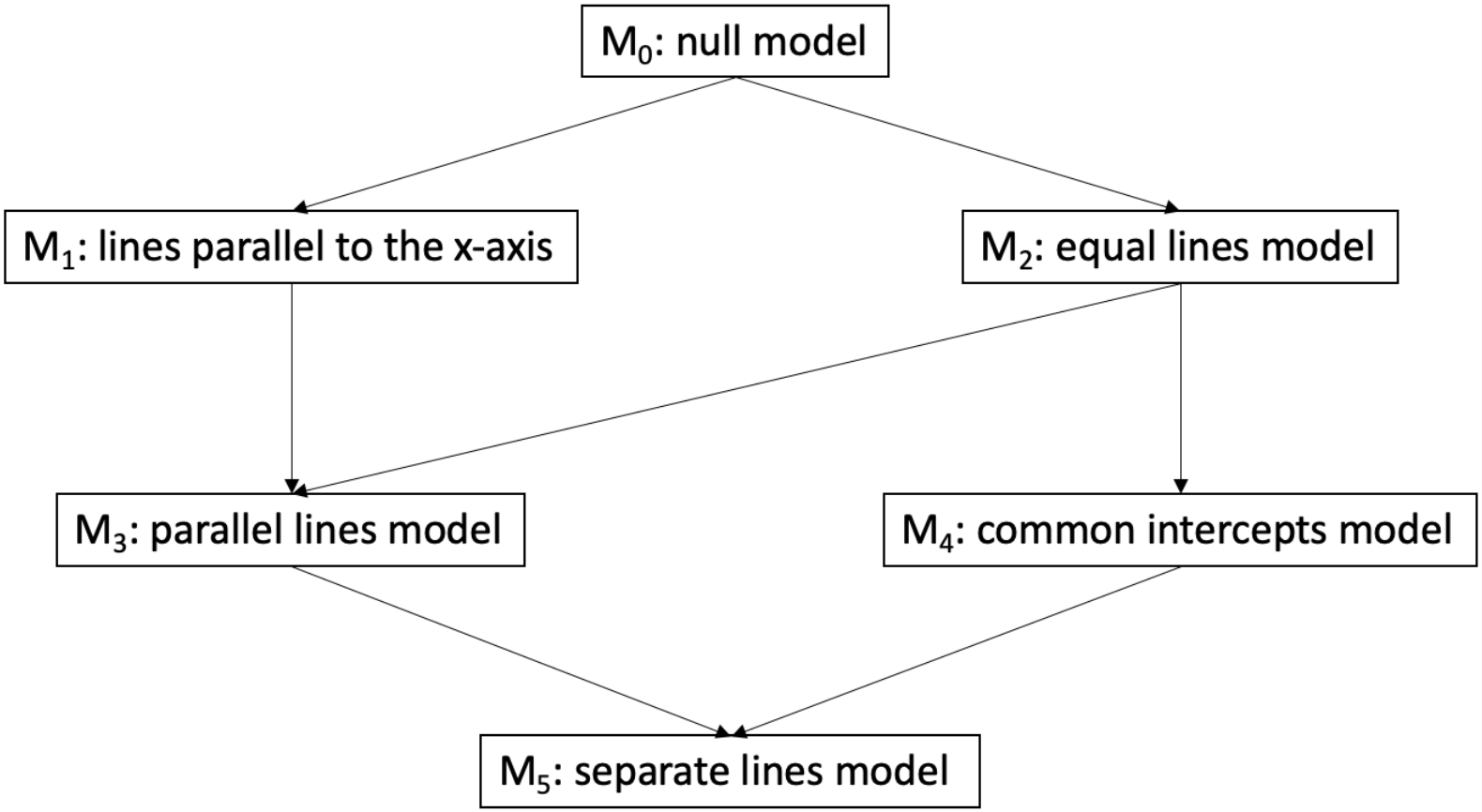
Continuous and Categorical Predictors

- When we have a continuous and a categorical predictor, we can build separate regression models for each category
- Example: in Autumn, small winged fruit called samara fall off maple trees, spinning as they go
  - A forest scientist studied the relationship between how fast they fell and their "disk loading" (a quantity based on their size and weight)
  - The data give the loadings and fall velocities for fruit from three trees.

Continuous and Categorical Predictors

- Let $Y_{ij} \sim \mathsf{N}(\mu_{ij}, \sigma^2)$ be the `Velocity` for `Tree` i and `Load` j
- We have `Load` as a continuous predictor and `Tree` as a categorical with 3 levels
- There are many different models we may consider; these include
  - $M_0 : \mu_{ij} = \beta_0$, the null model
  - $M_1 : \mu_{ij} = \beta_{0i}$, lines parallel to the x-axis
  - $M_2 : \mu_{ij} = \beta_0 + \beta_1 x_{ij}$, equal lines model
  - $M_3 : \mu_{ij} = \beta_{0i} + \beta_1 x_{ij}$, parallel lines model
  - $M_4 : \mu_{ij} = \beta_0 + \beta_{1i} x_{ij}$, common intercepts model
  - $M_5 : \mu_{ij} = \beta_{0i} + \beta_{1i} x_{ij}$, separate lines model
- There are different nesting relationships between these models

Continuous and Categorical Predictors

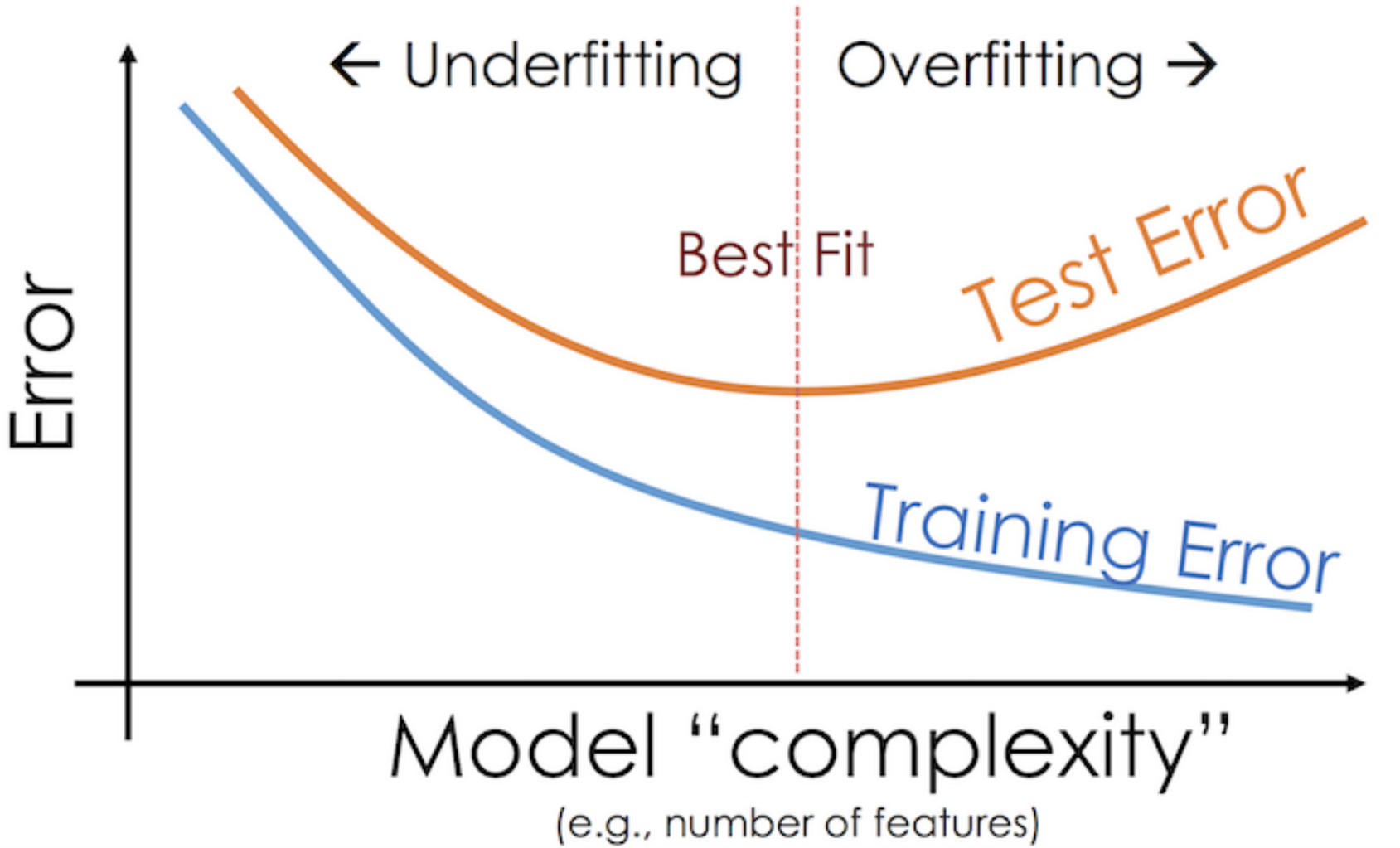


Out-of-Sample Predictive Performance

Out-of-Sample Predictive Performance

- How well can our predictions generalise to unseen samples?
- In Statistical Machine Learning jargon, we say that we have *trained* a model
- If we calculate a measure of goodness-of-fit such as the RMSE, it will reflect the error for the *training set*
- However, to understand the actual predictive power of our model, we must test it with unseen data, or a *validation/test* set
- This is because the more complex we make our model, the better the predictions will be for the training data
- Take, for instance, a saturated model: it will reproduce the training data exactly, therefore RMSE $= 0$; but this doesn't mean it will have good predictive power for unseen data

Out-of-Sample Predictive Performance



Overfitting

Out-of-Sample Predictive Performance

- More formally, if we have a model M_1 with parameters $\boldsymbol{\theta}_1$, whose MLE is $\hat{\boldsymbol{\theta}}_1$, the likelihood for the observed data y^{obs} is

$$L(y^{\text{obs}}|M_1, \hat{\boldsymbol{\theta}}_1)$$

- We will never know whether M_1 is the *true* generating model M_{true}
- We can assume, however, that y^{obs} is a sample from M_{true}
- How well does M_1 with $\hat{\boldsymbol{\theta}}_1$ predict y^{new} from M_{true}?
- The out-of-sample predictive performance is

$$\int L(y^{\text{new}}|M_1, \hat{\boldsymbol{\theta}}_1) L(y^{\text{new}}|M_{\text{true}}) dy^{\text{new}}$$

Out-of-Sample Predictive Performance

- A quantity of interest in this context is the *expected log predictive density* (elpd)
- If we have n independent samples of data from M_{true}, we can calculate the elpd as

$$\text{elpd} = \sum_{i=1}^{n} \log L(y_i^{\text{new}} | M_1, \hat{\boldsymbol{\theta}}_1)$$

Cross-Validation

- Rather than waiting for new data to be collected, a simple solution is to remove some data from the data that is used for model fitting, fit the model with the remaining data, and then test how well the fitted model predicts the reserved data
- This is known as *cross-validation*
- One common approach to cross-validation is known as *K-fold* cross-validation
  - The original data set is divided randomly into K subsets.
  - One of these subsets is randomly selected to be reserved for testing.
  - The remaining $K-1$ are used for fitting and the generalization to the reserved data set is evaluated
  - This process is repeated for all K subsets, and overall cross validation performance is the average of the K repetitions

Leave-One-Out Cross-Validation

- One extreme version of K-fold cross-validation is where $K = n$, where n is the size of the data-set
- This is called *Leave-one-out cross validation* (LOOCV)
- We divide our data into n sets
- In each set, we remove one of the observations only (for set 1 we remove y_1, for set 2 we remove y_2, and so on)
- We fit the model to each set and calculate the prediction for the observation that is missing
- We can then obtain

$$\text{elpd} = \sum_{i=1}^{n} \log L(y_i|\hat{\boldsymbol{\theta}}^{-i})$$

where $\hat{\boldsymbol{\theta}}^{-i}$ is the MLE for the model fitted to the set where observation i is missing

ELPD and AIC

- We can also look at the elpd in the "deviance scale", i.e. $-2 \times$ elpd
- Since we must re-fit the model n times, calculation of the elpd for large datasets can be computationally demanding (especially decades ago when computing power was much lower)
- In the early 70's, statistician Hirotsugu Akaike presented an approximation to LOOCV that could be easily computed
- The *Akaike Information Criterion* (AIC) is defined as

$$\begin{aligned} \text{AIC} &= 2k - 2\log L(\mathbf{y}|\hat{\boldsymbol{\theta}}) \\ &= 2k + \text{Deviance} \end{aligned}$$

- k is the number of parameters in the model

AIC

- AIC can be seen as a penalised goodness-of-fit measure to favour parsimony when two model fits are essentially equivalent
- A model's AIC value is of little value in itself, and so we only interpret differences in AIC between models
- Conventional standards (see, for example, Burnham and Anderson 2002, chap. 2):
  - AIC differences greater than 4 indicate clear superiority of the predictive power of the model with the lower AIC
  - differences of 10 or more indicate that the model with the higher value has essentially no predictive power relative to the model with the lower value
  - however, it is important to take *uncertainty* into account when comparing AIC values
- A correction for small sample sizes (typically when $\frac{n}{k} < 40$) is

$$\text{AICc} = AIC + \frac{2k(k+1)}{n-k-1}$$

Model Averaging

Akaike Weights

- We can use *Akaike weights* for model averaging
- Let

$$\Delta_j = \text{AIC}_j - \text{AIC}_{\text{min}}$$

  where AIC_{min} is the minimum AIC value in the set
- The Akaike weights w_j for a group of J models are defined as

$$w_j = \frac{\exp\left(-\frac{1}{2}\Delta_j\right)}{\sum_{j=1}^{J}\exp\left(-\frac{1}{2}\Delta_j\right)}$$

- Each w_k can be interpreted as the probability that model j has the best out-of-sample generalisation

Variable Selection

Stepwise Regression

- Iterative procedure that adds / removes one variable at a time that yields to the best AIC (or another measure of goodness-of-fit
- Different versions:
  - forward
  - backward
  - both
- Two main problems
  - Stepwise is a *greedy* algorithm
  - Chance of finding spurious associations, which leads to overfitting
- Example: student dataset

All Subsets Regression

- In this case we can simply fit all possible models using all combinations of predictors
- The number of models to fit increases exponentially with the number of predictors (2^k)
- Example: Swiss fertility data

Regularisation

- A more elegant way of performing variable selection is through the introduction of *penalty* terms in our objective functions so as to *regularise* the parameter estimates
- Regularisation is also referred to as *shrinkage*
- For example, the assumption of a normal distribution for the random effects in a linear mixed model effectively regularises, or shrinks their estimates towards zero, when compared to a fixed effects model

Ridge Regression

- Ridge regression is a method to reduce variance in estimators of regression coefficients
- It penalizes large coefficients and shrinks the towards zero
- In linear regression, it estimates the coefficients by minimising the penalized sum of squared residuals:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{k=0}^{K} \beta_k^2$$

where $\lambda > 0$ is called a *regularisation parameter*

LASSO

- The "Least Absolute Shrinkage and Selection Operator" (LASSO) is a method similar to ridge regression, but uses a penalty based on the sum of the absolute values of coefficients:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{k=0}^{K} |\beta_k|$$

- LASSO effectively drops predictors from the model, while ridge only shrinks them
- Since both penalty types rely on the β coefficient values, sometimes it is a good idea to scale the numerical predictors to have mean zero and variance one

Elastic Net

- Elastic net combines the ridge and LASSO penalties using a weighted averave:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \left\{ \alpha \sum_{k=0}^{K} |\beta_k| + (1-\alpha) \sum_{k=0}^{K} \beta_k^2 \right\}$$

  where $\alpha \in [0, 1]$
- When $\alpha = 1$, this is pure LASSO regression
- When $\alpha = 0$, this is pure ridge regression

Bayesian Regularisation

- There are specific priors in the Bayesian setting that are equivalent to ridge- or LASSO-type regularisation
- There are also other types, such as the *horseshoe*
- Examples