Statistical Modelling

Introduction

Models

- Ordinary (Normal) regression used for a continuous response
- Logistic regression used for a binary response
- Poisson regression used for count data

Assumptions

1. The ith observation's response, Yi, comes from a normal distribution.
2. The mean of the response, μi, is a linear combination of the explanatory terms.
3. The variance, σ2, is the same for all observations.
4. Each observation's response is independent of all others.

Procedure

- Explore data
  - Indicator variables uses dummy variables to index different levels, N.B. the baseline level is meaningless unless it's compared to other levels to measure the relative effect
  - Interaction of indicators variables to measure different combinations, N.B. the formula can be written out using I_x (0,1) depending on the level is absent or present
- Fit the model
  - A probability distribution for the response.
  - A linear combination of the explanatory variables.
  - A mathematical function (called the link function, essentially a transformation of y with respect to X) that relates the expected value of the response to the linear combination of the explanatory variables.
- Conduct diagnostic checks
  - Data issues
  - Model assumptions
- Perform inference using the fitted mode
  - Inference about the model coefficients.
  - Inference concerning the value of the response given values of the explanatory variables.
  - Inferences concerning which variables are "needed" in the model.

Analysis of variance (ANOVA) - sequential tests, so ordering matters!

- The F-statistic for LM is defined in terms of the residual sums of squares (RSS) for the two models and the number of explanatory variables k in each model:

$$f_0 = \frac{(RSS_S - RSS_F)/(k_F - k_S)}{RSS_F/(n - k_F - 1)}$$

p-value=Pr($F \geq f_0$) where F $\sim F_{k_F-k_S,n-k_F-1}$

- Chi-square test for GLM's a submodel can be tested against a full model using the change in deviance between the two models as the test statistic, i.e. $d_o = Dev_{sub} - Dev_{full}$ and p-value=Pr($D \geq d_0$) can be calculated from do $\sim \chi^2_{k_F-k_S}$ where v is the number of additional regressors in the full model

Matrix Notation

- We will define Y as the vector that contains random variables Y1, . . . Yn.

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

- Further we will define the mean vector μ as:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} \mathrm{E}(Y_1) \\ \mathrm{E}(Y_2) \\ \vdots \\ \mathrm{E}(Y_n) \end{bmatrix}$$

- Finally, we define the covariance matrix of Y as:

$$\boldsymbol{\Sigma}_\mathbf{Y} = \begin{bmatrix} \mathrm{var}(Y_1) & \mathrm{cov}(Y_1, Y_2) & \cdots & \mathrm{cov}(Y_1, Y_n) \\ \mathrm{cov}(Y_2, Y_1) & \mathrm{var}(Y_2) & \cdots & \mathrm{cov}(Y_2, Y_n) \\ \vdots & \vdots & & \vdots \\ \mathrm{cov}(Y_n, Y_1) & \mathrm{cov}(Y_n, Y_2) & \cdots & \mathrm{var}(Y_n) \end{bmatrix}$$

If the Yi's are independent and all have variance σ2:

$$\boldsymbol{\Sigma}_\mathbf{Y} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \end{bmatrix} = \sigma^2 \mathbf{I}_n$$

$$\begin{bmatrix} 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

where In is the n × n identity matrix

- Then, we can extend it to explanatory variables and coefficients

$$\underbrace{\begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}}_{\boldsymbol{\beta}} = \mathbf{X}\boldsymbol{\beta}$$

- And there we go,
  - Normal regression:

$$\mathbf{Y} \sim \text{Normal}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$$
$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}.$$

  - Poisson regression (count data):

$$\mathbf{Y} \sim \text{Poisson}(\boldsymbol{\mu}),$$
$$\log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}.$$

  - Logistic regression (binary data):

$$\mathbf{Y} \sim \text{Binomial}(\mathbf{n}, \mathbf{p}),$$
$$\text{logit}(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta}.$$

where **n** and **p** are vectors of the n_i's and p_i's respectively.

Estimate Coefficients

1. Write down the (log) likelihood function – the maximizing the (log) likelihood gives the same parameter estimates and it's easier to use.
2. Take the partial derivatives with respect to the parameters.
3. Set partial derivatives to zero and solve.
4. For most GLM's the equations produced are not linear. Thus, solutions can only be found numerically using IRLS (iteratively re-weighted least squares) -> sampling distribution in this case is approximations (central limit theorem as the theoretical basis)

(General) Linear/Regression Model

Definition

1. $$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$
$$\epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$$

Where we specify the model as random error, εi , about some relationship, β0 + β1X1i + β2X2i + . . ., for the ith observation

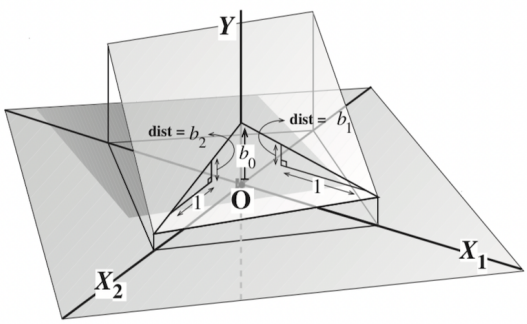
- Yi is the value of the response variable
- X1i is the value of explanatory variable X1.
- X2i is the value of explanatory variable X2.
- εi is the error (the difference between Yi and its expected value).

2. $$Y_i \sim \text{Normal}(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots$$

where we assume the ith observation's response, Yi, comes from a normal distribution with mean β0 + β1X1i + β2X2i + . . . and variance σ2 (we attribute the randomness directly to the response variable.)

Planar Response Surfaces

- For 2 regressors, the linear model can be depicted as a plane in R3



- For k regressors we can imagine a hyper-plane of dimension k in Rk+1

Estimated Coefficients

- Least squares (LS) - normal regression

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$$

- Minimizing $\sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2$ to obtain βˆ.
- x′βˆ approximates the conditional mean of y given x.

- Least absolute deviation (LDA) - quantile regression

$$Q(\beta_q) = q \underbrace{\sum_{i:y_i \geq \mathbf{x}_i'\beta_q}^{n} |y_i - \mathbf{x}_i'\beta_q|}_{\text{underprediction}} + (1-q) \underbrace{\sum_{i:y_i < \mathbf{x}_i'\beta_q}^{n} |y_i - \mathbf{x}_i'\beta_q|}_{\text{overprediction}}$$

- Minimizing $Q(\beta_q)$ to obtain βˆ.
- x′βˆ approximates the conditional median of y given x.

Inferences

$$\hat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta},\ \sigma^2 (\mathbf{X}^t\mathbf{X})^{-1})$$

- The diagonal elements of σ2(XtX)−1 are the variances of the βˆi′s and the off-diagonal elements are covariances between pairs of βˆi′s.
- To get a (1 − α)100% confidence interval for βi : βˆi ±tn−k−1(1−α/2)×se(βˆi)
- To test H0: βi = constant, calculate

$$\text{t-stat} = \frac{\hat{\beta}_i - \text{constant}}{\text{se}(\hat{\beta}_i)}$$

$$\text{p-value} = 2 \times \Pr(t_{n-k-1} \geq |\text{t-stat}|)$$

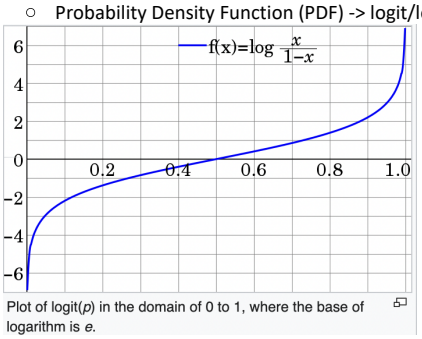
Generalized Linear Model (GLM)

Definition

- A probability distribution (which comes from the exponential family) for the response.
- A linear combination of the explanatory variables.
- A mathematical function (called the link function) that relates the expected value of the response to the linear combination of the explanatory variables.

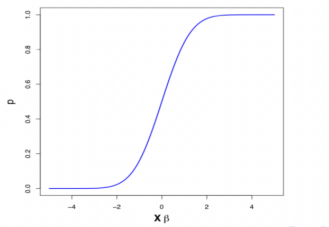
Logistic regression

- Link function
  - Probability Density Function (PDF) -> logit/log-odds (default)



Plot of logit(*p*) in the domain of 0 to 1, where the base of logarithm is *e*.

  - Cumulative Density Function (CDF) -> Probit Regression (link=probit), better suited for where we're estimating probabilities near 0 or 1



- Predictions on a assumption of normal distribution
  - Predict() to get responses (…type="response"…), then create a CI
  - Predict() to get estimated y' (…type="link"…), then create a CI using inversed link function exp(y')/(1+exp(y')) to get responses

Poisson Regression

- Link function is log(…) and the inverse link function is exp(…)
- Overdispersion: the observed variance is greater than expected
  - Quasipoisson model to adjust the underestimated variability by dispersion parameter
  - Negative Binomial Regression, note that the changing result as theta is evaluated every time the model is fitted
- Offset: by including an offset, it means that the other regressors are now modelling log μ_x for the population of the size x, in the form of … + offset(log(p/x)), or …, offset = log(p/x)
- Zero-inflated/Hurdle Poisson regression is used to model count data that has an excess of zero counts.

Others

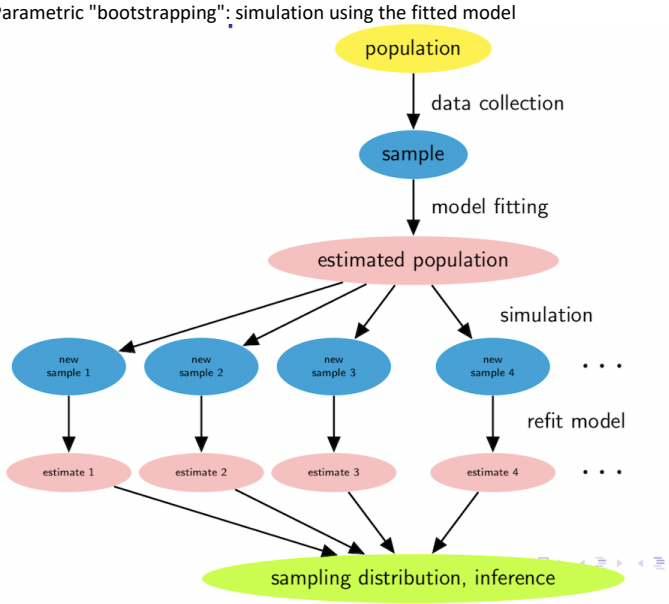
- Gamma and inverse-Gaussian, for continuous responses on the interval [0, ∞).
- Beta, for continuous responses on the interval [0, 1].

Bootstrapping

Bootstrapping refers to any resampling procedure (taking samples from the sample)

- The bootstrapping is useful for GLMs since sampling distributions based on theory are often only approximate even if the model assumptions are satisfied, but the bootstrapping is NOT used to get a better estimate of a statistic – the expected value of the bootstrap estimate is equal to the sample estimate, BUT to generate a reference distribution without relying on model assumptions
- It's tricky that we start with an estimated value rather than a true value to create a CI, so we need to adjust the differences between the true value and the estimated value by inverting the interval if the distribution is skewed -> 2 * estimated value - quantile on the other side

Parametric "bootstrapping": simulation using the fitted model



1. Fit the model using the original data.
2. Treat this fitted model as the true relationship and generate a new set of values for the response (note that the values of explanatory variables remain fixed).
3. Refit the model using the new set of response values.
4. Record the estimated value of the statistic of interest.
5. Repeat steps 2 through 4 a large number of times to create an empirical sampling distribution for the statistic of interest.

(Non-parametric) bootstrapping - resample from the set of observations (rows of the data frame) and each bootstrap sample contains the same number of observations as the original sample with replacement

Comparison
- Parametric bootstrapping is particularly useful when the null hypothesis needs to be true
- Non-parametric bootstrapping is particularly useful when the data do not appear to strictly satisfy the distributional assumption

Diagnostics

Data
- High leverage point - observations that have the potential to have a big impact on the fitted model - these points have unusual combinations of values for the regressors
  - The matrix $H = X\,(Xt\,X)^{-1}\,Xt$ is called the hat matrix as it converts the observed values Y into the fitted values $\hat{\mu}$, for GLM, the hat matrix diagonals are defined with respect to the final iteration of the IRLS procedure.
  - The hat matrix diagonals (HMDs) are used to detect high leverage points

$$\underbrace{\begin{bmatrix}\widehat{\mu}_1\\ \widehat{\mu}_2\\ \widehat{\mu}_3\\ \widehat{\mu}_4\\ \vdots\\ \widehat{\mu}_n\end{bmatrix}}_{\hat{\boldsymbol{\mu}}_{\mathbf{Y}}} = \underbrace{\begin{bmatrix}h_{11} & h_{12} & h_{13} & h_{14} & \cdots & h_{1n}\\ h_{21} & h_{22} & h_{23} & h_{24} & \cdots & h_{2n}\\ h_{31} & h_{32} & h_{33} & h_{34} & \cdots & h_{3n}\\ h_{41} & h_{42} & h_{43} & h_{44} & \cdots & h_{4n}\\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots\\ h_{n1} & h_{n2} & h_{n3} & h_{n4} & \cdots & h_{nn}\end{bmatrix}}_{\mathbf{H}} \underbrace{\begin{bmatrix}Y_1\\ Y_2\\ Y_3\\ Y_4\\ \vdots\\ Y_n\end{bmatrix}}_{\mathbf{Y}}$$

  - $0<h_{ii}<1$ for all diagonal elements, and $-1<h_{ij}<1$ for all off diagonal elements
  - As h_{ii} goes to 1, then h_{ij} goes to 0 for all $j \neq i$, thus large values of h_{ii} indicate points where y_i has a large influence
  - Average $h_{ii} = (k + 1)/n$, h_{ii} 's greater than $5(k + 1)/n$ or even $3(k + 1)/n$ are considered large
- Outliers - observations with large error terms where the value of the response is unusual given the values of the explanatory variables
  - For LM: $r=y-Hy=(I-H)y$, consequently, residuals do not have equal variances: $r \sim N(0, (I-H)\sigma^2)$, thus $Var(r_i)=(1-h_{ii})\sigma^2$
    - This can be refined as standardized residuals (internally studentized residuals)

$$r_i^* = \frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

- And can be taken further as studentized residuals (externally studentized residuals)

$$r_i^{\dagger} = \frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

where σˆ(i) ignores the i observation in estimating σ

- For GLM: both deviance residuals and Pearson residuals can be standardized by being divided by sqrt(1 − hii)
  - Deviance residuals measure the contribution of each observation to the residual deviance for the model

$$\text{deviance} = \sum d_i^2$$

where di2 is the difference in 2 log f (yi) between the maximalmodel and the specified model (residual is positive if yi > μi and negative otherwise)

  - Pearson residuals are the difference between the response and the fitted value divided by the standard error of the fitted value

$$P_i = \frac{y_i - \widehat{\mu}_i}{\widehat{\sigma}_i}$$

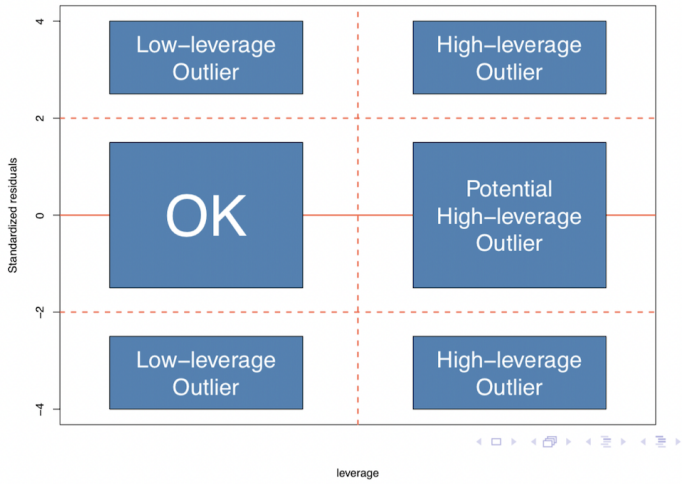
where σi is the estimated standard deviation for observation i

- Influential points are often the result of an observation that is both an outlier and a high leverage point
  - Cook's distance considers the overall change in β when the i observation is removed

$$D_i = \frac{r_i^{*2}}{k+1} \times \frac{h_{ii}}{1-h_{ii}}$$

where 0.5 indicates a potential HLP and 1 indicates an HLP

  - Leverage vs residual plot



- Multicollinearity is a near linear dependency between the explanatory variables (the X-values are not very well spread out), which indicates (an overlap of/confounding) information between two or more of the regressors
  - Impacts
    - Affect the stability of the fitted regression plane – applies to GLM's as well as ordinary regression
    - Inflate the standard errors in the estimated coefficients
    - May find that a number of coefficients are not significant, but if you remove any one of these variables the others become significant
    - Can be a serious problem if a goal of the analysis is to evaluate the impact of one of the affected regressors on the response – the data may not be well suited to that purpose, but Less of a problem if the goal is to find a good predictive model for the response – in this case subset selection methods may well eliminate one or more of the regressors involved.
  - Variance Inflation Factor (VIF) represents the amount that Var(βj) is inflated due to the correlation between Xj and the remaining regressors.

$$\text{VIF}_j = \frac{1}{1-R_j^2}.$$

    - values of 1 indicate a regressor that is orthogonal to all other regressors (i.e. independent)
    - values of 5 or more indicate multicollinearity.
    - values of approximately 10 or more indicate serious multicollinearity.

Model assumptions

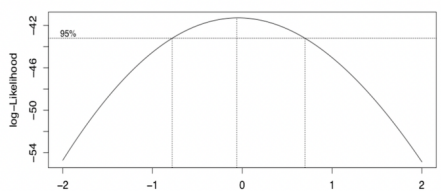
The linear regression surface is the most important assumption as it impacts the interpretation and accuracy of the fitted model. The others do not affect the validity of the fitted model but they will affect the sampling distributions for the fitted coefficients and thus statistical inference that depends on these sampling distributions.

- Error assumptions
  - Diagnostics
    - QQ-plot of residuals

    - Possible remedies
        - Ignore - inferences concerning β and μ are insensitive to the distribution of Y due to the central limit theorem – i.e. the sampling distributions of $\hat{\beta}$ and $\hat{\mu}$ will almost always be Normal to a very good approximation. The only situation where the distribution of Y is of real concern is if you want to form prediction intervals as these are predictions about an individual observation
        - Use a GLM with a more appropriate distributional assumption
        - Transform the response
- Non-Planar Regression Surface
    - Diagnostics
        - Preliminary plots used to explore the data often give the first indication of non-linear relationships between the response and the regressors
        - Plots of residuals versus fitted values or residuals versus individual regressors
        - gam plots or partial residual plots
    - Possible remedies
        - Transform the response
        - Transform one or more of the regressors
        - Add polynomial terms for one or more of the regressors,
        - Add interaction terms
- Correlated data
    - Diagnostics
        - A careful assessment of how the data was collected.
        - For data collected over time a plot of residuals versus time order or a correlogram.
    - Possible remedies
        - If the amount of correlation is small – ignore.
        - Adopt a more appropriate analysis to take into account the correlation – e.g. time series models, mixed models
- Non-constant Error Variance
    - Diagnostics
        - Plot of residuals versus fitted values to match normal distribution
        - A scale location plot for detecting unwanted patterns
    - Possible remedies
        - Transform the response
        - Use weighted least squares.
- Link function (GLM only)
    - Diagnostics
        - Preliminary plots used to explore the data where the response has been transformed using the link function
        - Plots of residuals versus fitted values or residuals versus individual regressors
        - GAM plots
    - Possible remedies
        - Use a different link function
        - Transform one or more of the regressors
        - Add polynomial terms for one or more of the regressors,
        - Add interaction terms
- Response Distribution (GLM only)
    - Diagnostics
        - The nature of the response often indicates what distribution should be used – the binomial is often used for proportions and the Poisson is often used for counts
        - Plots of studentized residuals may have unusually large values
    - Possible remedies
        - Try fitting models with alternative distributions (the quasi-binomial for proportions and the quasi-Poisson distribution or negative binomial distribution for counts)
        - Use a more suitable response distribution
- Sampling Distribution of Residual Deviance / Chi-Square Goodness of Fit Test (GLM only)
    - Reference distribution is $\chi 2_{n-k-1}$ for Poisson regression and $\chi 2_{m-k-1}$ for grouped binomial regression where m is the number of covariate patterns, note that residual is of little diagnostic value in a logistic regression modelling ungrouped data or poisson regression which have an estimated scale parameter (e.g. quasibinomial, quasipoisson, etc.) as the underlining reference distribution (X2) is not a good approximation
    - The null hypothesis is that samples come from a specified distribution

Transformations

- Transform the response



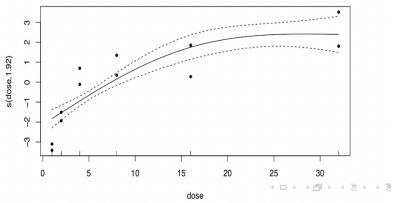
A Box-Cox plot is a convenient method of assessing whether or not transforming the response will be useful. For different values of λ fit the model and calculate the log-likelihood, then plot the values of the log-likelihood versus λ – the maximum value of the log-likelihood corresponds to the value of λ that results in the model that best fits the data.

$$Y^{\lambda} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

- Choose a value of λ that corresponds to a value of the log-likelihood that is close to the maximum value.
- λ = 0 represents a log transformation.

- Transform regressors

```
library(mgcv)
budworm.gam<- gam(cbind(s, n-s) ~ sex + s(dose,k=3),
                  family=binomial, data = budworm.df)
plot(budworm.gam,residuals=T,pages=1,pch=20)
```



General Additive Models (GAMs) can be used to explore the nature of the curvature in the regression surface

$$Y = g_1(x_1) + g_2(x_2) + \cdots + g_k(x_k) + \varepsilon$$

- The transformations can be set to be “smoothers”
- Plots of the smoothers suggest suitable transformations for the explanatory variables

Information Criteria

is mostly used to screen/shortlist a large amount of candidate models

Akaike Information Criterion (AIC): AIC = −2l + 2k where l is the model’s log-likelihood, and k is the number of parameters

- The first term measures model fit, and the second term penalises model complexity
- The smaller the better. If a model has a high log-likelihood (a good thing), then the first term will be small. If a model is simple (another good thing), the second term will also be small

AICc (corrected AIC) is a version of AIC that includes an additional term to correct the penalty when we do not have large samples:

$$\text{AICc} = -2\ell + \left(2k + \frac{2k^2 + 2k}{n - k - 1}\right)$$

where l is the model’s log-likelihood, and k is the number of parameters

- When sample sizes are small, we should use AICc rather than AIC.
- When sample sizes are large, then the additional term is almost zero, so AIC and AICc are virtually equivalent.

Bayesian Information Criterion (BIC): BIC = −2l + log(n)k where l is the model’s log-likelihood, and k is the number of parameters

- The penalty for each parameter is log(n), rather than 2, where n is the number of observations
- As long as n > 8, BIC penalises additional parameters more harshly than AIC or AICc, so BIC tends to favour simpler models.

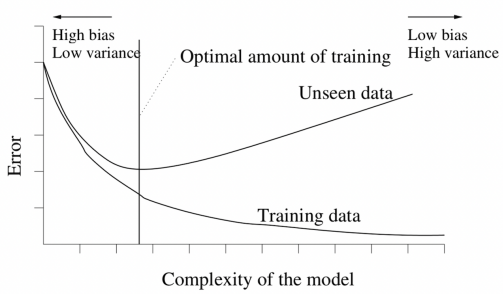
Model search

- Exhaustive search
- Stepwise search (forwards/backwards)
  1. Choose your information criterion (e.g., AICc)
  2. Fit the (null/full) model.
  3. Consider all possible models with one additional/less parameter, calculating the criterion for each.
  4. If the (null/full) model has the best criterion value, select it as your final model. Otherwise, move to the model with the best criterion value, and repeat steps 3–4

Model Selection

Datasets

- The training set is used for model selection
- The test set is used for model evaluation
- Training and test sets are independent and identically distributed (IID)
- To minimize prediction error, model complexity must be balanced against how well the model fits the current data



$$E_{test} = \underbrace{(E_{test} - E_{train})}_{variance} + \underbrace{(E_{train} - E_{best})}_{bias} + \underbrace{E_{best}}_{noise}$$

- E_{best} is the irreducible error (lowest possible error for any model), e.g. irreducible error for predicting coin flips is 0.5
- Underfitting: simple models tend to be underfitted to the training set, so $E_{train} \uparrow$ -> high bias, low variance
- Overfitting: complex models tend to be overfitted to the training set, so $E_{train} \downarrow$ -> low bias, high variance

Resampling
It's common for the small training set to be further resampled to training set and validation set to better estimate the predication error for the unseen data, note that we usually re-train on the full training set for the final model

- Jackknifing
  i. Delete d observations randomly from the data
  ii. Fit the model to the remaining data
  iii. Calculate the PE of the fitted model using the deleted observations
  iv. Repeat the above steps a number of times and compute the mean PE
- K-fold cross-validation
  i. Split the data into K (roughly) equal-sized parts
  ii. Fit the model to all the data except the kth part
  iii. Calculate the PE of the fitted model over the kth part
  iv. Repeat the above steps for k = 1,...,K and compute the mean PE

Regularisation for high dimensional data to be penalised more heavily on the model complexity - fit a model with all predictors, but with their coefficients being constrained or regularised. Note that we use normalization or standardization beforehand to keep the data on the same scale, and k-fold cross-validation to choose λ

- Lasso regression

  $$RSS + \lambda \sum_{k=1}^{n} |\beta_j|$$

  where the penalty term uses the length of B, aka, L1 norm, this is equivalent to eliminating the unnecessary coefficients
  - Tuning parameter λ ≥ 0 controls the strength of the penalty term. (0 when λ=∞)
  - The bias

    $$\mathbb{E}[\hat{\boldsymbol{\beta}}^{ridge} - \mathbb{E}[\boldsymbol{\beta}]] = ((X^T X + \lambda I)^{-1} X^T X - I) - \boldsymbol{\beta}$$

  - The variance

    $$\mathrm{Var}(\hat{\boldsymbol{\beta}}^{ridge}) = (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)$$

  - The degree of freedom

    $$df(\hat{Y}) = trace(X(X^T X + \lambda I_p)^{-1} X^T)$$

- Ridge Regression

  $$RSS + \lambda \sum_{k=1}^{n} \beta_j^2$$

  where the penalty term uses the length of B to the second power, aka, L2 norm, this is equivalent to shrinking the coefficients towards 0
  - Tuning parameter λ ≥ 0 controls the strength of the penalty term. (null model when λ=∞)
  - The bias increases as λ (amount of shrinkage) increases.
  - The variance decreases as λ (amount of shrinkage) increases.
  - The degree of freedom

    $$df(\hat{Y}) = \mathbb{E}[\text{number of nonzero coefficients in } \hat{\boldsymbol{\beta}}^{lasso}]$$

- Elastic Net: for highly correlated variables, we need a method between Ridge and Lasso, e.g. there exists a pair of identical covariates. l1 penalty is indifferent and the coefficients are not defined. l2 will yield two equal valued coefficients, the penalty term is defined as:

  $$\lambda \left(\frac{1}{2}(1-\alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right)$$

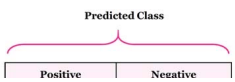
  where α ∈[0, 1] with α = 1, it reduces to the l1-norm (LASSO) and with α = 0, it becomes l2-norm (ridge)
- (Sparse) Group Lasso to penalize all the coefficients within a group for grouped or categorical variables

Measurements

- For continuous responses, the most common measure of the precision of predictions is the mean square prediction error which is defined as:

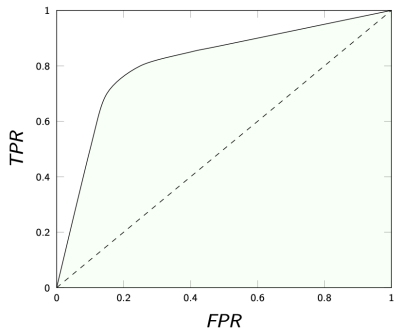
  $$\mathrm{MSPE} = \mathrm{E}\left(Y - \widehat{Y}\right)^2.$$

- For classifiers, the prediction error is the probability of a wrong classification (0's predicted as 1's, 1's predicted as 0's)
  - Confusion Matrix



Actual Class				
	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

    - Accuracy = (TP + TN)/All
    - P = Precision (Exactness) = TP/(TP + FP)
    - R = Recall (completeness) = TP/(TP + FN)
    - For multiclass classification, add up the metrics for different classes and take the average, e.g. average accuracy, [Equation]
- F-measure is the weighted measure of precision & recall
- ROC (Receiver Operating Characteristics) curves shows the trade-off between true positive rate and false positive rate & AUROC (Area Under ROC) is a measure of the accuracy of the model



- Entropy in information theory can be defined as the expected number of bits of information contained in an event, which helps to reduce the uncertainty by a reduction factor of 1/P(xi)

$$Entropy, H(X) = -\sum_{i=1}^{n} P(x_i) log_2(P(x_i))$$

    - Cross-Entropy is the average number of bits required to represent an event from one distribution, compared to another distribution

$$H(p, q) = -\sum_{i=1}^{n} p(x_i) log_2(q(x_i))$$

    - Kullback–Leibler divergence(KL divergence)

$$KL\ divergence\ =\ cross\ entropy - entropy$$

Causal Models

Causal Relationship

- A theoretical outcome based on what would have happened for a particular observation if the value of one or more of the explanatory variables had been different is called a counterfactual observation.
- A more precise (mathematical) definition of causality can be given based on counterfactual outcomes. We will do this for a situation where there is a continuous response Y and a single binary explanatory variable A.
    - The individual causal effect for observation i is

$$Y_i^{a=1} - Y_i^{a=0}.$$

        - $Y_i^{a=1} = Y_i \big| [A_i = 1]$ represents the outcome for the ith observation if the treatment is given.
        - [Equation] represents the outcome for the ith observation if the treatment is not given.
        - For observation i there is a causal effect if [Equation] and there is no causal effect if [Equation] .
    - As the individual causal effects may vary, we are really interested in the average casual effect (ACE) over the population of interest:

$$\text{ACE} = \text{E}(Y_i^{a=1} - Y_i^{a=0}).$$

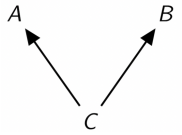
Causal Modelling

1. Designed experiments: suppose we have a representative sample of individuals from the population of interest.
    - Randomly divide the individuals into 2 groups
    - One group gets the treatment and the other does not.
    - The difference in mean value of Y between the two groups will be an unbiased estimate of the ACE.
2. Causal analysis of observational data: use causal diagrams to identify and eliminate the “contributions from other sources”, then we will be left with an unbiased estimate of the ACE and describe an underlying mechanism.
    - Correlation
        - Sampling error: there will be some correlation between A and B in our sample even if there is none in the population.

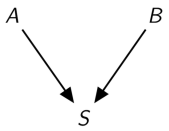
$$A \longrightarrow B$$

$$A \longleftarrow B$$

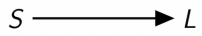
- C causes both A and B.



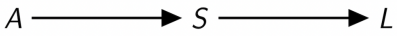
- Selection bias: A and B both affect the probability that an observation is included in the sample (S).



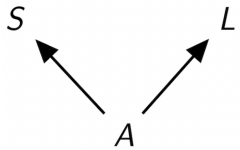
- Pathways
  - Causal paths
    - Direct effects, e.g. S on L.



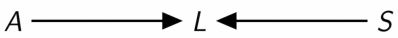
    - Indirect effects, e.g. S is included/excluded for direct/total effect of A on L.



  - Confounding paths, e.g. the cofounder A is excluded for direct/total effect of S on L.



  - Colliding paths, e.g. the collider L is excluded for direct/total effect of A on S.



Classification

Multinomial logistic regression: there are K possible outcomes (K classes) and the K -class is the reference class.

- The probability for a class k , k = 1, . . . , K − 1 is

$$Pr(Y = k|X) = \frac{\exp(\alpha_k + \beta_k^T X)}{1 + \sum_{l=1}^{K-1} \exp(\alpha_l + \beta_l^T X)}$$

- The probability for the reference class K is

$$Pr(Y = K|X) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\alpha_l + \beta_l^T X)}$$

- $\Sigma_{l=1}^{K} Pr(Y = l|X) = 1$
- Unknown parameters $\alpha 1, \beta 1, ..., \alpha K-1, \beta K-1$ are estimated using the MLE

Naïve Bayes for high-dimensional data mixed with quantitative and qualitative data

- A fundamental rule to classify an observation x is to use conditional probabilities:
  $\mathbb{P}(Y = j|X = x), \quad j = 1, \ldots, J,$
  which is also known as posterior probabilities
- According to Bayes' theorem, this is equivalent to:

$$\mathbb{P}(Y = j|X = x) = \frac{\pi_j f_j(x)}{\sum_{l=1}^{J} \pi_l f_l(x)}$$

  Where πj = P(Y = j) is the proportion of class j observations in the population known as the prior probability for class j, fj(x) is the density function, and [Equation] is the same for all classes
- Each fjk (xk) is a density estimate from the values of xk only.
  - For a quantitative variable, it can be a normal density estimate or a kernel density estimate (KDE)
  - For a categorical variable, it is the sample proportions for each level of xk.
- Choose the class with the largest πjfj(x)

Linear/Quadratic Discriminant Analysis (LDA/QDA) assume that the class-conditional density P(Y = k|X) to be of Gaussian and decision boundaries are [Equation]

- The discriminant function for QDA is:

$$\delta_k(x) = \log \pi_k - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) - \frac{1}{2} \log |\Sigma_k|$$

- LDA assume that all classes have a common covariance matrixΣ=Σ1 =⋯=ΣK, so the discriminant function can be reduced to:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

Non-Linear Regressions

Polynomial regression where higher-order terms of X are included in the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_d X^d + \epsilon.$$

Step functions - essentially a piece-wise constant function

- The range of X is partitioned into intervals at cut points c1,c2,...,cm:

  $$Y = \alpha_0 I(X < c_1) + \alpha_1 I(c_1 \le X < c_2) + \cdots + \alpha_{p-1} I(c_{m-1} \le X < c_m) + \alpha_m I(X \ge c_m) + \epsilon.$$

  where the indicator function is denoted by

  $$I(Z) = \begin{cases} 1, & \text{if } Z = \texttt{TRUE}; \\ 0, & \text{if } Z = \texttt{FALSE}. \end{cases}$$

- Another representation:

  $$Y = \beta_0 + \beta_1 I(X \ge c_1) + \cdots + \beta_m I(X \ge c_m) + \epsilon$$
  $$= \beta_0 + \sum_{j=1}^{m} \beta_j I(X \ge c_j) + \epsilon,$$

  where β0 = α0 and βj = αj − αj −1 (j = 1, . . . , m)

Spines is dividing the domain of X into contiguous intervals and representing f by a separate polynomial in each interval, note that the degree of freedom is sum of polynomial order and knot number.

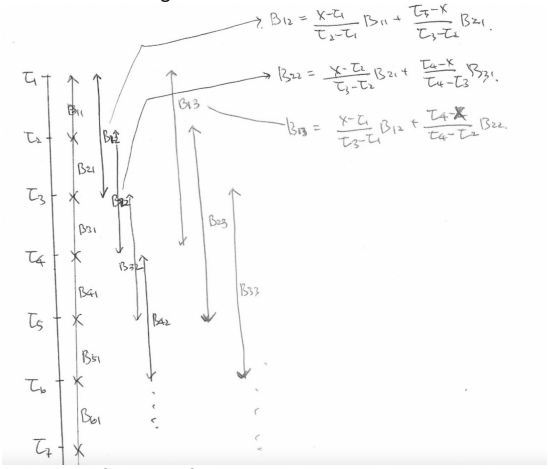
- Regression splines - essentially a piece-wise linear function with fixed knots
  - Regression spline extends the step function by replacing β0 with α0 +α1X +···+αdXd, and each knot I (X ≥ cj) with

    $$(X - c_j)_+^d = \begin{cases} (X - c_j)^d, & \text{if } X \ge c_j; \\ 0, & \text{if } X < c_j. \end{cases}$$

    and hence we can rewrite the dth-degree regression spline as

    $$Y = \sum_{j=0}^{d} \alpha_j X^d + \sum_{j=1}^{m} \beta_j (X - c_j)_+^d.$$

  - Different d's
    - if d ≥ 1, it is continuous
    - If d ≥ 2, it has a continuous (d − 1)th-order derivative and hence is smooth of order d − 1 (d = 3 is commonly used in practice)
    - If d is large, it can cause numerical problems
- Basic Spines (B Spines) for local support
  - Spline function of given degree can be expressed as a linear combination of basic function of that degree

    - Basis function of degree 1 is

      $$B_{i,1}(x) = \begin{cases} 1, & \text{if } \tau_i \le x < \tau_{i+1} \\ 0, & \text{otherwise} \end{cases}$$

      for $i = 1, \ldots, K + 2M - 1$.

    - Basis function of order 1 < m ≤ M is obtains recursively

      $$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1.m-1}(x)$$

      for $i = 1, \ldots, K + 2M - m$.

- Smoothing splines
  - Smoothing spline takes an approach to penalise RSS by minimising

    $$\mathrm{RSS}^{\mathrm{sspline}}(f; \lambda) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int [f''(x)]^2 \, \mathrm{d}x$$

    and the solution by theory, aka, natural cubic spine is
    - a piecewise cubic polynomial with knots at the unique values of x1, . . . , xn inside the region between the two extrema (minimum and maximum) of the xi 's.
    - linear outside of this region
    - has continuous first and second derivatives at the knots (and everywhere)
  - Memory-based: all unique values of x1, . . . , xn are taken as knots
  - Smoothness: the smoothness of the spline is controlled by the value of λ, which corresponds uniquely to a value of the effective degree of freedom dfλ
    - Large λ -> smooth
    - Small λ -> wiggly

Local regression

- Local regression computes the fit at each target point x0, by assigning different weights to observations to minimise

$$\sum_{i=1}^{n} w(x_i; x_0)(y_i - \beta_0 - \beta_1 x_i)^2$$

  - Higher weights for those close to x0
  - Lower or 0 weights for those far away
- Memory-based method: all observations must be remembered for prediction
- Extension: one can further replace the simple linear regression model with a (higher-order) polynomial to improve prediction accuracy

Generalised additive models (GAM)

- A GAM is an extension to multiple linear regression and can be written as

$$Y = \beta_0 + f_1(X_1) + \cdots + f_p(X_p) + \epsilon,$$

  where fj can be a nonlinear function of Xj, e.g., a smoothing spline
- Backfitting: transforming a variable does not need to be conducted beforehand and is now just part of the model-fitting process
- Extension: Nonlinear interactions can be included

Gaussian process

is a stochastic process such that for any finite set of elements x1 , ..., xn ∈X, the associated finite set of random variables f(x1),...,f(xn) have a Gaussian distribution. Similar examples are Bernoulli process, random walk, Wiener process, Poisson process, etc.

- Random variables F = {f(x1),...,f(xn)} has a multivariate Gaussian distribution, N(M, K):

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim N\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix} \right)$$

  where real valued function m and a function k for a valid covariance matrix are defined as:

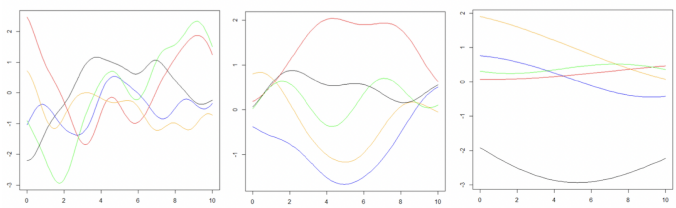
$$m(x) = \mathbb{E}[f(x)]$$

$$k(x_i, x_j) = \mathbb{E}[(f(x_i) - m(x_i))(f(x_j) - m(x_j))], \qquad x_i, x_j \in \mathcal{X}.$$

- The squared exponential kernel is used for k

$$k_{SE}(x_i, x_j) = \exp\left(-\tau \|x_i - x_j\|^2\right), \quad \tau > 0$$

  - $\|x_i - x_j\| \to 0$ then $\exp\left(-\tau \|x_i - x_j\|^2\right) \to 1$
  - $\|x_i - x_j\| \to \infty$ then $\exp\left(-\tau \|x_i - x_j\|^2\right) \to 0$
  - As the bandwidth parameter τ decreases, then points will have higher correlations hence the sampled functions tend to be smoother overall



Samples from a zero-mean Gaussian process with k_{SE} using $\tau = 0.5$(left), $\tau = 0.1$ (middle), $\tau = 0.01$ (right).

Gaussian process can be used as a prior probability distribution for non-parametric models where assumptions are minimized in exchange of flexibility, note that it scales badly with the number of observations

- A Gaussian process GP(0,K) is used as a prior for F. Then the posterior for F becomes

$$\begin{aligned} p(F|Y,X) &\propto p(Y|F,X)p(F|X) \\ &= \frac{\exp(-\frac{1}{2\sigma^2}(Y-F)^T(Y-F))}{(2\pi\sigma)^{n/2}} \frac{\exp(\frac{-1}{2}F^T K^{-1} F)}{\sqrt{(2\pi)^n |det(K)|}} \end{aligned}$$

  where the maximum posterior estimation F^ (the most probable value for F given X,Y) is the value maximizing log p(F|Y, X):

$$\hat{F} = K(K + \sigma^2 I)^{-1} Y$$

- Regression - predict [Equation] given a test data [Equation]
  - Known that f has a zero-mean Gaussian distribution with covariance function k, the marginal distribution is also a Gaussian distribution.

$$\begin{bmatrix} F \\ F^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X,X) & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{bmatrix}\right)$$

    [Equation]
  - Assuming iid noise,

$$\begin{bmatrix} E \\ E^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma^2 I_{n \times n} & 0 \\ 0 & \sigma^2 I_{n^* \times n^*} \end{bmatrix}\right)$$

- Then the summation of Gaussian random variables has a Gaussian distribution,

$$\begin{bmatrix} Y \\ Y^* \end{bmatrix} = \begin{bmatrix} F \\ F^* \end{bmatrix} + \begin{bmatrix} E \\ E^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X,X) + \sigma^2 I_{n\times n} & K(X,X^*) \\ K(K^*,X) & K(K^*,X^*) + \sigma^2 I_{n^*\times n^*} \end{bmatrix}\right)$$

- Lastly, Y* has a Gaussian distribution due to conditional density

$$Y^* \sim N(\mu^*, \Sigma^*)$$

where

$$\mu^* = K(X^*,X)(K(X,X) + \sigma^2 I_{n\times n})^{-1} Y$$

$$\Sigma^* = K(X^*,X^*) + \sigma I_{n^*\times n^*} - K(X^*,X)(K(X,X) + \sigma^2 I_{n\times n})^{-1} K(X,X^*)$$

- Classification - predict the posterior probability for y = m:

$$p(y = m|x) = \int g(f(x))N(f(x); 0, K)df$$

where y is a categorical variable, y ∈{1, ..., M} and the class probability is modelled by a function g