Regression Analysis for Data Scientists (ZZSC5806)

Atefeh Zamani
University of New South Wales
School of Mathematics and Statistics

H2- 2023

Notations

- **Random Variables**: $Y_1, Y_2, \ldots, Y_n$ (not bold, CAPITAL LETTERS)
- **Realisations**: $y_1, y_2, \ldots, y_n$ (not bold, small letters)
- **Parameters**: α, β, etc. (Greek Letters)
- **Estimators**: $\hat{\alpha}$, $\hat{\beta}$, etc.
- **Vectors**: $\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$, $\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ (not italic/italic, small letters, **bold**)
- **Matrices**: $\mathbf{X} = \begin{bmatrix} X_{11} & \ldots & X_{1p} \\ X_{21} & \ldots & X_{2p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \ldots & X_{np} \end{bmatrix}$ (not italic/italic, CAPITAL LETTERS, **bold**)
- **Transpose**: $\mathbf{y}^T$, $\boldsymbol{y}^T$, $\mathbf{X}^T$.

Week 1: Estimation and Inference for Regression Models

Introduction to regression modelling and key statistical concepts that we will use throughout the course. This includes:

- estimation methods,
- maximum likelihood estimation,
- least squares estimation,
- exponential family of distributions,
- prediction,
- inference.

1.1 Introduction to regression analysis

- **What is regression Analysis?**
  Regression Analysis investigates the functional **relationship** between **statistical variables**. Data are usually **multiple observations** of a random vector $(Y, \mathbf{x})$.
  - $\mathbf{x} = (X_1, \ldots, X_p)^\top$ is a p-vector of variables (explanatory variables, regressors, predictors, input variables or independent variables).
  - Y may be continuous ($\in \mathbb{R}$), discrete ($\in \{1, \ldots, K\}$) or ordinal (ordered discrete)(response variable, target variable, output variable, outcome variable or dependent variable).

  Response variables are usually treated as **random variables**, while **predictors** are treated as **fixed observations**.

1.1 Introduction to regression analysis

- **Response variable**
  - **nominal**: categories
  - **ordinal**: classes
  - **continuous**: continuous scale, at least in theory.

binary 2 category / multinomial 2 < category

coin

eye color

$\in \mathbb{R}$

Nominal and ordinal data are discrete variables and can be qualitative or quantitative (e.g. counts). Continuous data are quantitative.

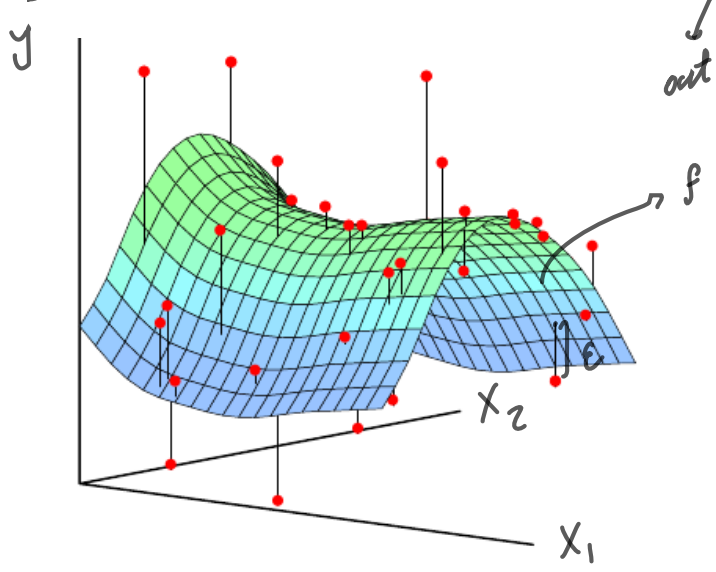
- **Explanatory variables**
  - quantitative (Covariate) or qualitative (Factor).

1.1 Introduction to regression analysis

Response	Explanatory	Method
Continuous	Binary	t-test
	Nominal, > 2categories	ANOVA → P=2 ≡ t-test
	Ordinal	ANOVA
	Continuous →	Multiple regression
Binary	Categorical	~~Contingency tables~~
	Continuous →	Logistic or probit regression ✓
Nominal, > 2categories	Nominal	~~Contingency tables~~
	Categorical & Continuous	Nominal logistic regression
Ordinal	Categorical & Continuous	Ordinal logistic regression ✗
Counts	Categorical	Log-linear models ✓
	Categorical & Continuous	Poisson regression ✓

1.1 Introduction to regression analysis

- **Regression**
  Aim: Find a "good" functional relationship of the form $Y = f(\mathbf{x}) + \varepsilon$.



to be estimated
output
predictors
$\varepsilon \geq 0$
$\varepsilon < 0$
$E(\varepsilon) = 0$

Figure: Regression of Y (vertical, continuous) on (X_1, X_2) (horizontal).

1.1 Introduction to regression analysis

- **General framework of statistical learning**
  - **Statistical learning**: A vast set of tools for understanding data.
    - supervised
    - unsupervised
  - **Regression**: statistical model for predicting or estimating an output based on one or more inputs (Supervised).

  In contrast, unsupervised methods cover situations where there are inputs but no supervising output. In these type of analysis we learn about relationships and structure of data. Example of unsupervised analysis is cluster analysis.

1.2 Scope of the Course

- **Week 1**: Introduction to regression modelling and key statistical concepts.
- **Week 2**: Dive in the world of regression modelling with focus on the popular linear models.
- **Week 3**: Move towards a more flexible regression model: the Generalised Linear Model or GLM. This allows to tailor models to specific problems and types of data.
- **Week 4**: Look into how to select the best model amongst a pool of models and estimate the error that comes with the selected model based on new (test) data.
- **Week 5**: Learn about nonlinear Regression (more complex in terms of interpretation and inference with higher predictive power).
- **Week 6**: Discover Generalised Additive Models (GAMs).

1.3 Fundamental definitions

- **The purpose of regression analysis**: Prediction and Inference
  - **Prediction**
    In many situations, **x** is available but the output Y cannot be easily obtained.
    Since $E(\varepsilon) = 0$,

$$\hat{Y} = \hat{f}(X), \tag{1.3.1}$$

    where
    - $\hat{f}$: estimate for f
    - $\hat{Y}$: the resulting prediction for Y.

    The estimate $\hat{f}$ is characterised by a **reducible** error and by an **irreducible** error.

$$\mathbb{E}[Y - \hat{Y}]^2 = \mathbb{E}[f(X) + \varepsilon - \hat{f}(X)]^2 = [f(X) - \hat{f}(X)]^2 + \mathbb{V}ar(\varepsilon) \tag{1.3.2}$$

1.3 Fundamental definitions

- **The purpose of regression analysis**
  - **Inference**
    Goal: Understanding the relationship between **x** and Y. How Y changes as a function of $X_1, X_2, \ldots, X_p$.
    - Which predictors are associated with the response?
    - What is the relationship between the response and each predictor?
    - Is the relationship between Y and each predictor linear or more complex?

  Depending on the purpose of the analysis, various methods of estimating f may be more appropriate.
  - **Prediction**: Non-linear models (in some cases overfitting causes problems)
  - **Inference**: linear models

1.3 Fundamental definitions

- **Method of estimation**
  - **Parametric**: Assumption on the functional form of f

$$f(\mathbf{x}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \tag{1.3.3}$$

    Method to fit the model (estimating $\beta_0, \beta_1, \ldots, \beta_p$)
  - ~~**Nonparametric**~~: no explicit assumptions about the functional form of f at the price of not having a small and fixed amount of parameters (i.e. larger sample sizes are needed)

1.3 Fundamental definitions

- **Maximum likelihood estimation (MLE)**
  - $\mathbf{y} = [Y_1, Y_2, \ldots, Y_n]^\top$: a random vector.
  - $f(\boldsymbol{y}; \boldsymbol{\theta})$: **joint** probability density function of Y_i, depends on $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]^\top$.

  The **likelihood function** $L(\boldsymbol{\theta}; \boldsymbol{y})$ is the same as $f(\boldsymbol{y}; \boldsymbol{\theta})$ but the emphasis is on $\boldsymbol{\theta}$ while $\boldsymbol{y}$ stays fixed.

Note: L is a random variable!

The maximum likelihood estimator of θ

MLE of $\boldsymbol{\theta}$ is the value $\hat{\boldsymbol{\theta}}$ which maximizes the likelihood function:

$$L(\hat{\boldsymbol{\theta}}, \boldsymbol{y}) \geq L(\boldsymbol{\theta}, \boldsymbol{y}) \quad \text{for all} \quad \boldsymbol{\theta} \in \Theta \quad \text{(parameter space)}. \tag{1.3.4}$$

Equivalently, $\hat{\boldsymbol{\theta}}$ is the value that maximizes the log-likelihood function $l(\boldsymbol{\theta}, \boldsymbol{y}) = \ln L(\boldsymbol{\theta}, \boldsymbol{y})$.

1.3 Fundamental definitions

- **Cont. Maximum likelihood estimation (MLE)**
  - **Method**
    1. Differentiate $l(\boldsymbol{\theta}, \boldsymbol{y})$ with respect to each element θ_j of $\boldsymbol{\theta}$
    2. Solving the simultaneous equations

$$\frac{\partial l(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \theta_j} = 0 \qquad j = 1, \ldots, p \tag{1.3.6}$$

    3. If the matrix of second derivatives

$$\frac{\partial^2 l(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \theta_j \partial \theta_k} \tag{1.3.7}$$

evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ is negative definite, then $\hat{\boldsymbol{\theta}}$ maximizes $l(\boldsymbol{\theta}, \boldsymbol{y})$ in the **interior** of Θ.

negative eigenvalues

Note: Check if there are any values of $\boldsymbol{\theta}$ at the **edges** of Θ that give local maxima of $l(\boldsymbol{\theta}, \mathbf{y})$. When all local maxima have been identified, the value of $\hat{\boldsymbol{\theta}}$ corresponding to the largest one is the MLE.

1.3 Fundamental definitions

- **Maximum likelihood estimation**:
  - **Example** (Poisson distribution)
    Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with Poisson distribution

$$f(y_i, \theta) = \frac{\theta^{y_i} e^{-\theta}}{y_i!},$$

    $y_i = 0, 1, \ldots,$ with the same parameter θ. Find the MLE of θ.

1.3 Fundamental definitions

$\hat{\theta}$ as MLE of θ $\Rightarrow$ $g(\hat{\theta})$ MLE of $g(\theta)$

- **Property of MLE**
  - **Invariance**: If $g(\theta)$ is any function of the parameters θ, then the maximum likelihood estimator of $g(\theta)$ is $g(\hat{\theta})$.
  - **Consistency**: Let $\hat{\theta}$ be the MLE of θ, obtained based on n i.i.d observations. Then, $\hat{\theta}$ is consistent for θ; i.e.,

$$\Pr\left(|\hat{\theta}_n - \theta| \geq \varepsilon\right) \to 0, \text{ as } n \to \infty \quad \equiv \quad \hat{\theta}_n \xrightarrow{P} \theta \ \checkmark$$

  - **Asymptotic normality**: Let $\hat{\theta}$ be the MLE of θ, obtained based on n i.i.d observations. Then,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right)$$

consistent → asymp normal

  - **Asymptotic efficiency**: A CAN estimator $\hat{\theta}_n$ of θ is said to be asymptotically efficient if the asymptotic variance of $\hat{\theta}_n$ equals $(I(\theta))^{-1}$.
  - **Sufficiency**

1.3 Fundamental definitions

- **Least squares estimation**
  $Y_1, \ldots, Y_n$: independent r.v. with $\mathbb{E}(Y_i) = \mu_i(\boldsymbol{\beta})$, where $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^\top$, $p < n$, is the parameter vector.
  We want to **estimate** $\boldsymbol{\beta}$.
  Least squares estimator: finding the estimator $\hat{\boldsymbol{\beta}}$ that **minimizes the sum of squares (SS)** of the differences between Y_i's and their expected values

$$SS = \sum_{i=1}^{n} [Y_i - \mu_i(\boldsymbol{\beta})]^2. \tag{1.3.8}$$

  and $\hat{\boldsymbol{\beta}}$ is obtained by differentiating SS with respect to the elements β_j:

$$\frac{dSS}{d\beta_j} = 0 \qquad j = 1, \ldots, p \tag{1.3.9}$$

1.3 Fundamental definitions

- **Weighted least squares**
  If Y_i's have variances σ_i^2 (not necessarily equal), **minimize the weighted sum of squared (WSS)** differences

$$WSS = \sum_{i=1}^{n} w_i [Y_i - \mu_i(\beta)]^2. \tag{1.3.10}$$

  where $w_i = (\sigma_i^2)^{-1}$.
  Idea: less reliable observations have less influence on the estimates.
  General form: $\mathbf{y} = [Y_1, \ldots, Y_n]^\top$ is a random vector with mean vector $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^\top$ and variance-covariance matrix $\mathbf{V}$, then

$$WSS = (\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}). \tag{1.3.11}$$

1.3 Fundamental definitions

- **Comments**
  - Least squares estimator: No assumptions on the distribution of the response variables (In MLE, the distribution is needed).
  - For many situations, MLE and least squares estimates are identical.
  - In many cases, numerical methods are used for parameter estimation.

1.3 Fundamental definitions

- **Model fitting**

Step 1. **Model specification**: an equation linking the response and the explanatory variables and a probability distribution for the response variable

Step 2. **Estimation of the parameter of the model**

Step 3. **Checking the adequacy of the model**

Step 4. **Inference**: calculating confidence interval, testing hypotheses

1.3 Fundamental definitions

- **Some Examples**
  - **Example**: Australian longitudinal study on women's health, Lee et al. (2005)
  - **Example**: Relating income to years of education

1.3 Fundamental definitions

- **What have we learnt from these examples?**
  1. What is the scale of measurement?
  2. What is a reasonable distribution to model the data?
  3. What is the relationship with other variables?

$$\mathbb{E}[Y] = \alpha + \beta x$$
$$\log[\mathbb{E}(Y)] = \alpha + \beta \sin(\gamma x)$$

  4. What is the best parameter estimation process? MLE, Least Squares, Bayesian methods
  5. Why choosing a restrictive (parametric method) instead of a very flexible (nonparametric) approach?
  6. Model checking: residual checking, plots, checking of the assumptions

1.3 Fundamental definitions

- **Don’t under-evaluate exploratory statistics!**

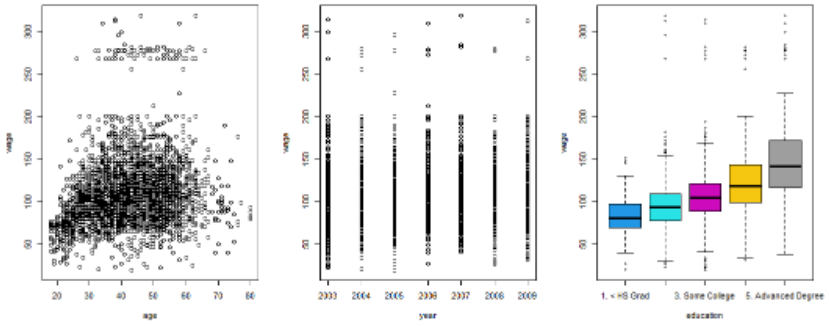


Figure: Representations of wage as a function of age, year and education

- Many statistical learning methods are relevant and useful in a wide range of disciplines.
- Statistical learning should not be viewed as a series of black boxes.
- While it is important to know what job is performed by each tool, it is not necessary to have the skills to construct the machine inside the box.

1.3 Fundamental definitions

- **Measuring the quality of a fit**

  Goal: How well do the model predictions match the data??
  - In a regression setting, use the **mean squared error (MSE)**

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{1.3.15}$$

    - select the model which **minimises the MSE**.
    - Low *MSE* can hide problems of **overfitting**.

1.3 Fundamental definitions

- **Measuring the quality of a fit**

  Goal (Corrected): ~~What is the~~ accuracy of the predictions when we apply the method on unseen/new data??

  test data

Step 1. Select (training) observations $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.
Step 2. Estimate $f(\boldsymbol{x})$.
Step 3. Consider some new set of observations (x_0, y_0), called testing observations.
Step 4. Use $\hat{f}$, the obtained estimation in Step 2, to find the **test MSE**

$$\text{Ave}(\hat{f}(x_0) - y_0)^2$$

Then:

- select the model with **minimum test MSE**, if test observations are available
- select the model with minimum training MSE, if test observations are not available
- use estimation method for the test MSE, like cross-validation

1.3 Fundamental definitions

- **The bias-variance trade-off**

  The expected test MSE, for x_0, can be decomposed into **three quantities**:

$$MSE(x_0) = \mathbb{E}\left[y_0 - \hat{f}(x_0)\right]^2 = Bias^2\left(\hat{f}(x_0)\right) + \mathbb{V}ar\left(\hat{f}(x_0)\right) + \mathbb{V}ar[\varepsilon]$$

  The **test MSE** can **never** be **lower than** $\mathbb{V}ar[\varepsilon]$

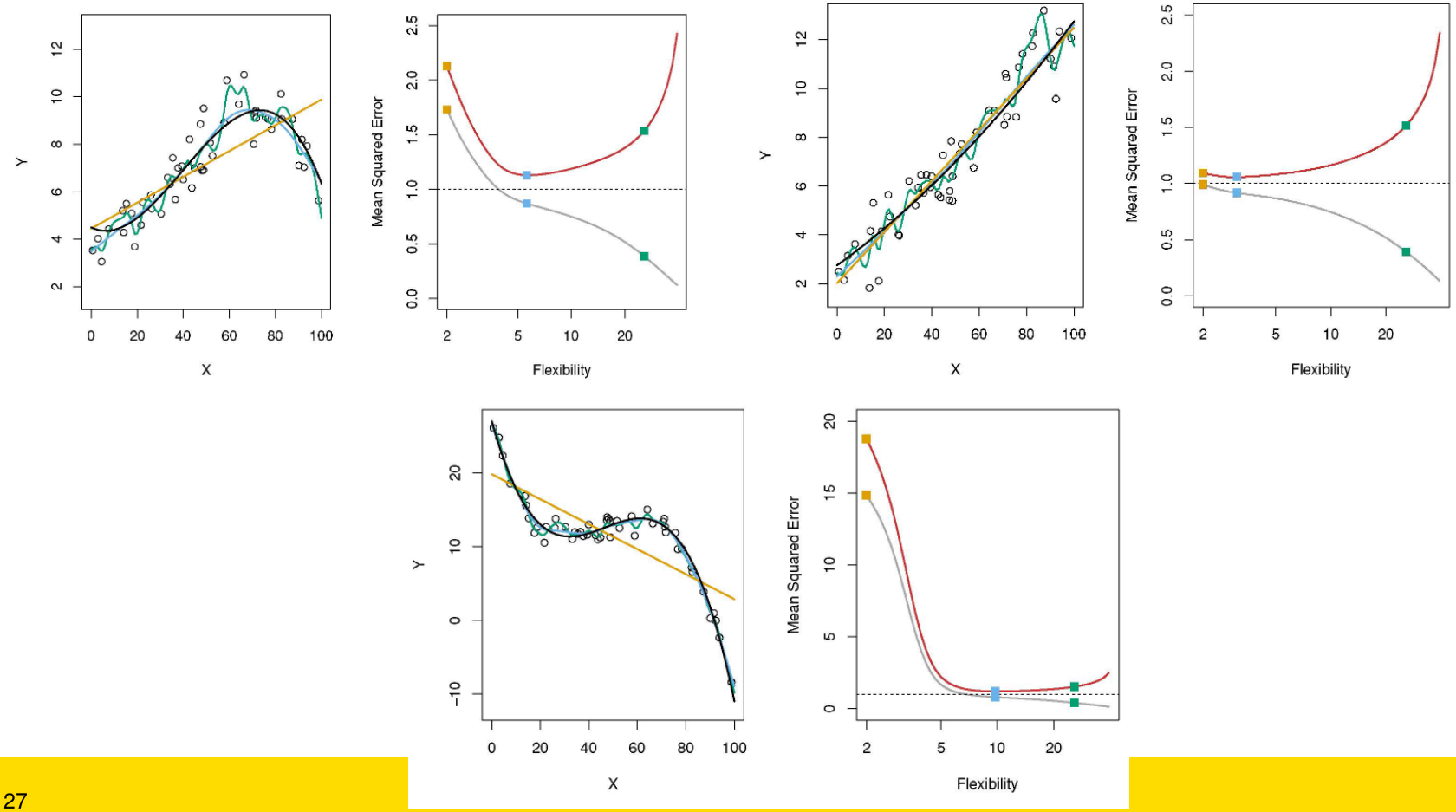
- Two competing properties of statistical learning methods
  - **Bias**: the error that is introduced by approximating a potentially complicated relationship between Y and X with a simpler model:

$$Bias\left(\hat{f}(x_0)\right) = \left[f(x_0) - \mathbb{E}(\hat{f}(x_0))\right]$$

  - **Variance**: the amount by which $\hat{f}$ would change when changing the training dataset:

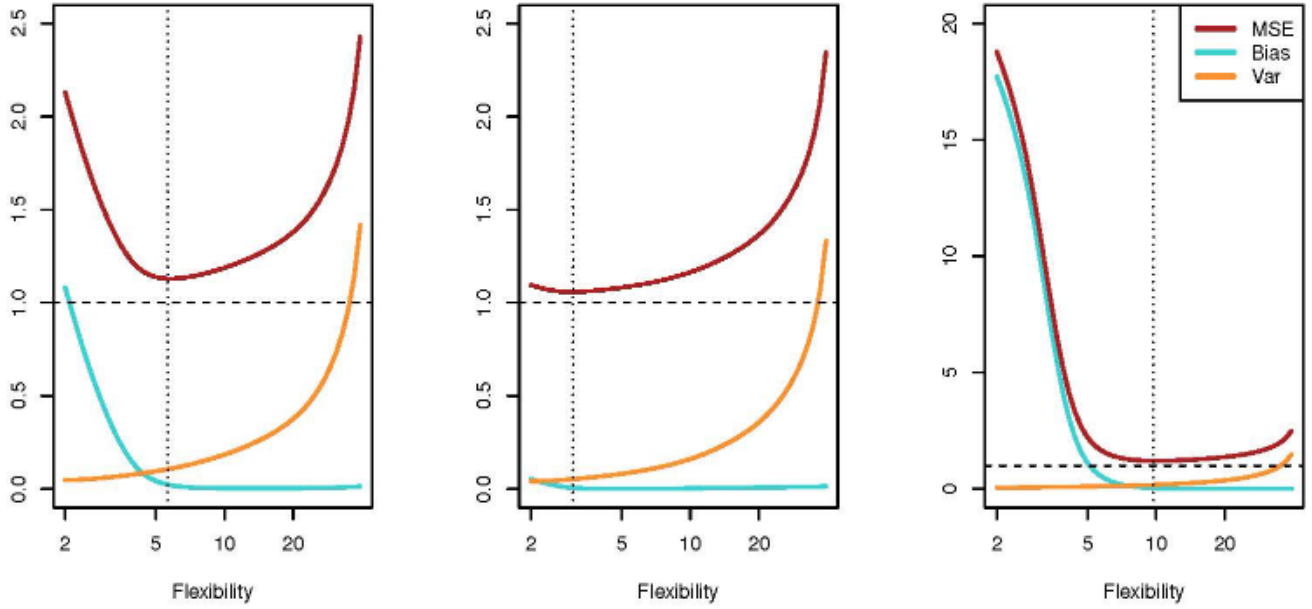
$$\mathbb{V}ar\left(\hat{f}(x_0)\right) = \mathbb{E}\left[\hat{f}(x_0) - \mathbb{E}(\hat{f}(x_0))\right]^2$$

1.3 Fundamental definitions



1.3 Fundamental definitions

- **The bias-variance trade-off**
  - **Comments**
    - **Variance**: in general, flexible methods have larger variance
    - **Bias**: in general, restrictive methods have a larger bias



1.3 Fundamental definitions

- **The classification setting**
  Goal: How to measure model accuracy for **categorical outputs**??
  The model accuracy is measured by the (training) **error rate**

$$\mathrm{ER} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(y_i \neq \hat{y}_i) \tag{1.3.16}$$

  where $\hat{y}_i$ is the predicted label and

$$\mathbb{I}(y_i \neq \hat{y}_i) = \begin{cases} 0 & \text{if } y_i = \hat{y}_i \text{ (correct classification)} \\ 1 & \text{if } y_i \neq \hat{y}_i \text{ (miss-classification)} \end{cases}$$

- **Test Error rate**:

$$Ave\left(\mathbb{I}(y_0 \neq \hat{y}_0)\right)$$

  where $\hat{y}_0$ is the prediction obtained by applying the classifier on input x_0.

1.3 Fundamental definitions

- **The Bayes classifier** The test error rate is minimised, on average, by the **Bayes classifier**.
  - **Rule**: Assigns each observation to the most likely class given its predictor values, i.e.,

$$\hat{y}_0 = j \quad \text{if } \Pr(Y = j|X = x_0) > \Pr(Y = i|X = x_0) \text{ for all categories} i \neq j$$

  - **Example** If there are only 2 classes, the Bayes classifier is predicts class 1 if

$$\hat{y}_0 = \begin{cases} 0 & \text{if } \Pr(Y = 0|X = x_0) > 0.5 \\ 1 & \text{if } \Pr(Y = 1|X = x_0) > 0.5 \end{cases}$$

1.3 Fundamental definitions

- The **expected prediction error** is

$$\mathrm{EPE} = \mathbb{E}\left[\mathbb{I}(y_0 \neq \hat{y}_0)\right] = \mathbb{E}_X \sum_{j=0}^{1} \left[\mathbb{I}(y_0 \neq \hat{y}_0)\right] \Pr(y_0 = j | X = x)$$

  and to minimise the expected error, you need to minimize the probability of being wrong, or

$$\hat{y}_0 = 1 \qquad \text{if} \quad \Pr(y_0 = 1 | X = x_0) = \max_{j \in \{0,1\}} \Pr(y_0 = j | X = x_0)$$

- The **expected Bayes error rate** is then

$$1 - \mathbb{E}_X[\max_{j \in \{0,1\}} \Pr(Y_0 = j | X)]$$

Note: In practice, we do not know the conditional distribution of Y given X, so the Bayes classifier is an **unattainable gold standard**.

1.3 Fundamental definitions

- **K-nearest neighbours**
  - The **K-nearest neighbours (KNN) classifier** **estimates** the conditional probability of Y given X and classifies a given observation to the class with the **highest estimated probability**.
    - simplicity
    - often close to the Bayes classifier.
  - Let K be a positive integer and x_0 be a test input observation, then:
    - Identifies the K points in the training dataset closest to x_0 - a neighbourhood of x_0, N_0
    - Computes $\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} \mathbb{I}(y_i = j)$
    - Applies the Bayes rule $\Pr(X = x_0|Y = j) = \frac{\Pr(Y=j|X=x_0)}{\Pr(Y=j)}$
    - Classifies the test observation x_0 to the class with the largest $\Pr(X = x_0|Y = j)$

What is the best choice of K??

1.3 Fundamental definitions

- **K-nearest neighbours**

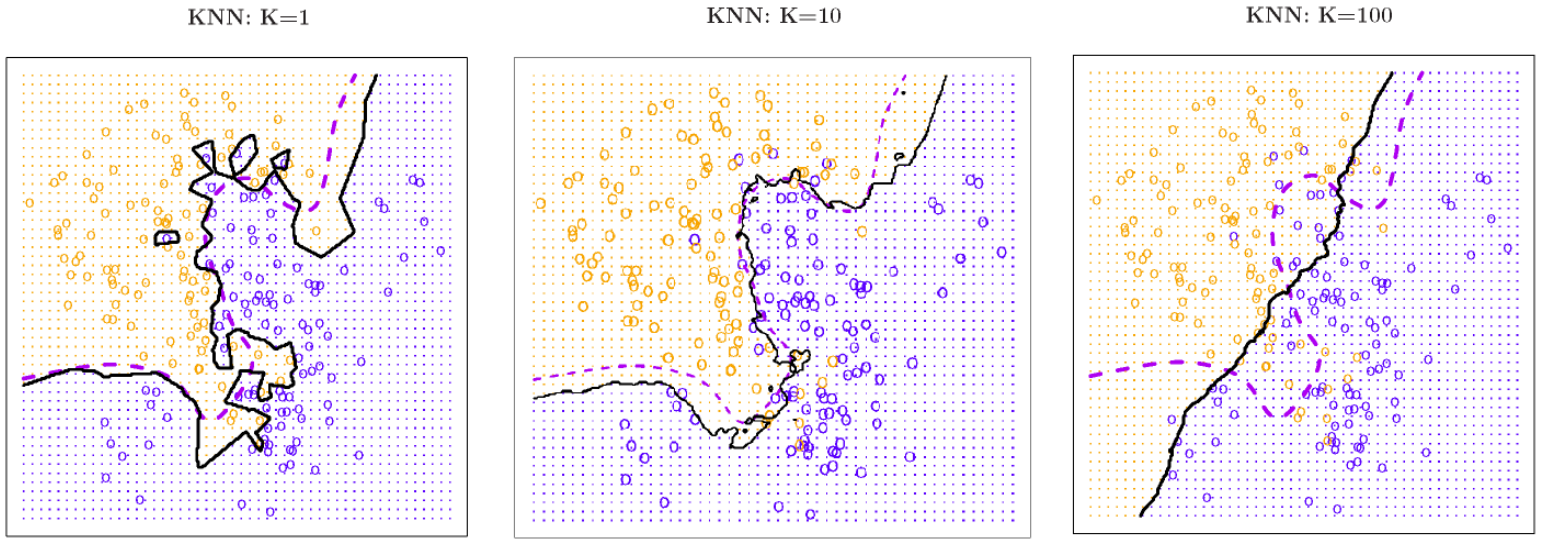


Figure: KNN with $K = 1$, $K = 10$ and $K = 100$. KNN decision boundary compared with Bayes decision boundary.

1.4 Estimation procedure

- **Introduction**

  This section is about obtaining
  - **Point estimates**
  - **Interval estimates**

  no analytical solutions $\Rightarrow$ numerical methods (e.g., Newton-Raphson algorithm).

1.4 Estimation procedure

- **The Weibull distribution [Weibull(θ, λ)]**
  - commonly used model for times to failure (or survival times) with pdf

$$f(y|\lambda,\theta) = \frac{\lambda y^{\lambda-1}}{\theta^{\lambda}} \exp\left[-\left(\frac{y}{\theta}\right)^{\lambda}\right], \qquad y > 0, \qquad \theta > 0 \qquad (1.4.1)$$

where λ is the **shape parameter**, θ is the **scale parameter**.
Equation (1.4.1) can be rewritten as

exponential family

$$f(y|\lambda,\theta) = \exp\left[\log\lambda + (\lambda-1)\log y - \lambda\log\theta - \left(\frac{y}{\theta}\right)^{\lambda}\right], \qquad y > 0, \qquad (1.4.2)$$

1.4 Estimation procedure

- **Likelihood function for Weibull(θ, λ)**
  - Let $Y_1, \ldots, Y_N \overset{\text{iid}}{\sim} \textit{Weibull}(\theta, \lambda)$. The likelihood is

$$f(y_1, \ldots, y_N | \lambda, \theta) = \prod_{i=1}^{N} \frac{\lambda y_i^{\lambda-1}}{\theta^{\lambda}} \exp\left[-\left(\frac{y_i}{\theta}\right)^{\lambda}\right] = \frac{\lambda^N (\prod y_i)^{\lambda-1}}{\theta^{N\lambda}} \times \exp -\sum \left(\frac{y_i}{\theta}\right)^{\lambda}$$

and the log-likelihood

$$\ell(\theta, \lambda; y_1, \ldots, y_N) = \sum_{i=1}^{N} \left[(\lambda - 1)\log y_i + \log\lambda - \lambda\log\theta - \left(\frac{y_i}{\theta}\right)^{\lambda}\right] \qquad (1.4.3)$$

$$= (\lambda-1)\sum \log y_i + N\log\lambda - N\lambda\log\theta - \sum\left(\frac{y_i}{\theta}\right)^{\lambda}$$

1.4 Estimation procedure

- **Estimation of the Weibull parameters**
  - First, focus on θ and consider λ as known (Exponential Family).
  - To find MLE of θ, called $\hat{\theta}$,
    - derive the derivative of the log-likelihood with respect to θ (**score function**):

$$\frac{d\ell}{d\theta} = \mathrm{U} = \sum_{i=1}^{N}\left[-\frac{\lambda}{\theta} + \frac{\lambda y_i^{\lambda}}{\theta^{\lambda+1}}\right] \tag{1.4.4}$$

    - set $\mathrm{U} = 0$

$$\sum_{i=1}^{N}\left[-\frac{\lambda}{\theta} + \frac{\lambda y_i^{\lambda}}{\theta^{\lambda+1}}\right] = 0 \iff -\frac{N\lambda}{\theta} + \frac{\lambda}{\theta^{\lambda+1}}\sum_{i=1}^{N} y_i^{\lambda} = 0 \iff \frac{-N\lambda\theta^{\lambda} + \lambda\sum_{i=1}^{N} y_i^{\lambda}}{\theta^{\lambda+1}} = 0$$

Consequently,

$$\hat{\theta} = \left(\frac{\sum_{i=1}^{N} y_i^{\lambda}}{N}\right)^{1/\lambda}$$

1.4 Estimation procedure

Example

Consider a dataset including lifetimes (times to failure in hours) of Kevlar epoxy strand pressure vessels at 70% stress level, (Andrews and Herzberg, 1985). If this data set follows the Weibull distribution, find the MLE of θ.

Figure: Histogram of times to failure (hours) for Kevlar epoxy strand pressure vessels.

1.4 Estimation procedure

- **Newton-Raphson algorithm in 1D**
  - Suppose $f : \mathbb{R} \to \mathbb{R}$ is differentiable and attains the value 0 at x_0.
  - The **Newton-Raphson algorithm** finds x_0 iteratively.
    - The slope of f at a value $x^{(m-1)}$ is given by

$$\left[\frac{df}{dx}\right]_{x=x^{(m-1)}} = f'(x^{(m-1)}) \approx \frac{f(x^{(m)}) - f(x^{(m-1)})}{x^{(m)} - x^{(m-1)}}$$

      if $x^{(m)} - x^{(m-1)}$ is small (mean value theorem).
    - If $x^{(m)} = x_0$ is the solution so that $f(x^m) = 0$ holds, then

$$x^{(m)} = x^{(m-1)} - \frac{f(x^{(m-1)})}{f'(x^{(m-1)})}$$

    - Iterating this algorithm creates a sequence $x^{(m)}$ which approaches x_0.

Visualisation on Wikipedia

1.4 Estimation procedure

- **Newton-Raphson algorithm in pD**
  - Let $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^T$ be the vector of parameters.
  - Define the **score function with respect to** θ_j as $\mathrm{U}_j := \partial \ell / \partial \theta_j$.
  - Set $\mathbf{u} := (\mathrm{U}_1, \ldots, \mathrm{U}_p)$.
  - Construct a vector sequence $\boldsymbol{\theta}^{(m)}$ converging to $\hat{\boldsymbol{\theta}}$, where $\mathbf{u}(\hat{\theta}) = 0$.
    - Write $\mathbf{u}^{(m)} := \mathbf{u}(\theta^{(m)})$.
    - By the mean value theorem, we have

$$\mathbf{u}^{(m)} - \mathbf{u}^{(m-1)} \approx H_\ell^{(m-1)} \left(\boldsymbol{\theta}^{(m)} - \boldsymbol{\theta}^{(m-1)} \right)$$

      where the Hessian H_ℓ is the $p \times p$-matrix with entry $\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k}$ at position (j, k).
    - Aiming for $\mathbf{u}^{(m)} = 0$, we arrive at a Newton-Raphson step

$$\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)} - (H_\ell^{(m-1)})^{-1} \mathbf{u}^{(m-1)}. \tag{1.4.6}$$

1.4 Estimation procedure

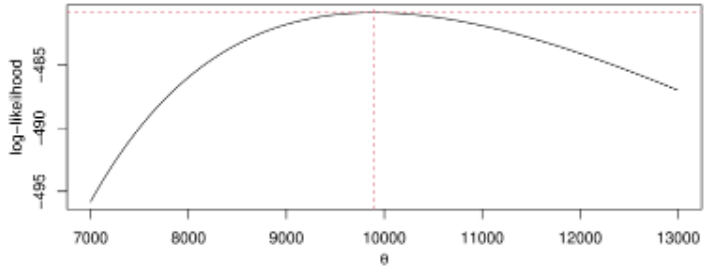
- **Method of scoring**
  - It has been shown empirically that no significant loss of accuracy is suffered when $H_\ell^{(m)}$ is replaced by minus the information matrix $-\mathcal{I} := \mathbb{E}[\mathrm{U}']$.
  - So, you may iterate:

$$\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)} + (\mathcal{I}^{(m-1)})^{-1}\mathbf{u}^{(m-1)}; \tag{1.4.7}$$

This is called the **method of scoring**.

1.4 Estimation procedure

- **Likelihood maximisation**
  - The curvature of the function ℓ in the vicinity of the maximum gives information about how reliable the MLE is.
  - The curvature of ℓ is defined by the rate of change of U, i.e. U$'$ or $\mathbb{E}$[U$'$]:
    - if U$'$ is small, then ℓ is flat and U is small for a wide interval of values for $\theta \Rightarrow \hat{\theta}$ is not well determind and its s.e. is large.
    - if U$'$ is large, then ℓ is concentrated around $\hat{\theta}$
  - The variance of $\hat{\theta}$ is inversely related to $\mathcal{I} = \mathbb{E}[-\text{U}']$, i.e. *s.e.*$(\hat{\theta}) = \sqrt{\frac{1}{\mathcal{I}}}$



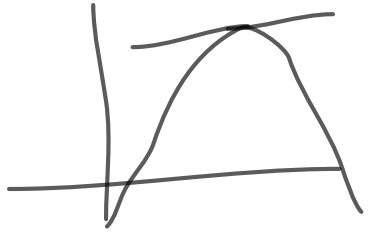


Figure: Log-likelihood function of pressure data as a function of θ.

1.5 Exponential family of distributions

Definition

Let Y be a random variable with values in $\mathbb{R}$ (or $\mathbb{N}_0$ or $\mathbb{Z}$), with pdf $f(y;\theta)$. Then, the distribution belongs to the **exponential family** if

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)}, \quad \text{for some } a, b, s, t$$

or equivalentley

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)] \tag{1.5.1}$$

where $s(y) = e^{d(y)}$ and $t(\theta) = e^{c(\theta)}$.

- $a(y) = y \Rightarrow$ the distribution of Y is called **canonical**.
- The distribution is in its canonical form, $\Rightarrow b(\theta)$ is the **natural parameter**.
- If there are other parameters, they are called **nuisance parameters**

1.5 Exponential family of distributions

- **Binomial distribution**
  - Suppose $Y \sim Bin(n, p)$. Y represents ...
  - The probability mass function of Y is:

$$f(y|p) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad y \in \{0, 1, \ldots, n\}.$$

  - Let n be known and p be the parameter of interest, which indicates ... .
  - **Does the Binomial belong to the Exponential family of distributions?**
    - Rewritten $f(y|p)$ as

$$f(y|p) = \exp\left[y \log p - y \log(1-p) + n \log(1-p) + \log \binom{n}{y} \right]$$

    - Compare with (1.5.1) :

$$a(y) = y,\ b(p) = \log \frac{p}{1-p},\ c(p) = n \log(1-p),\ d(y) = \log \binom{n}{y}. \qquad (1.5.2)$$

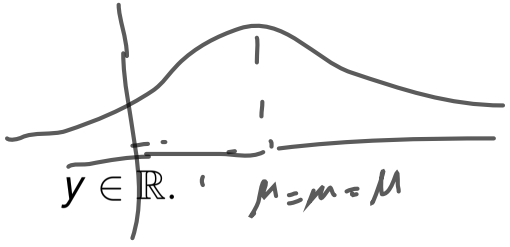
Code

1.5 Exponential family of distributions

- **Normal distribution**
  - Suppose $Y \sim N(\mu, \sigma^2)$ with the pdf:

$$f(y|\mu) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right), \quad y \in \mathbb{R}.$$

assume σ^2 is known



  - **Does the normal belong to the Exponential family of distributions?**
    - Rewritten $f(y|\mu)$ as

$$f(y|\mu) = \exp\left[-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right]$$

    - Compare with (1.5.1) :

$$a(y) = y,\ b(\mu) = \mu/\sigma^2,\ c(\mu) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2),\ d(y) = -\frac{y^2}{2\sigma^2} \qquad (1.5.3)$$

  - **Why is normal distribution important?**

Code

1.5 Exponential family of distributions

- **Poisson distribution**
  - Suppose $Y \sim \text{Pois}(\lambda)$, with the pdf:

$$f(y|\lambda) = f(y|\lambda) = \frac{\lambda^y}{e^\lambda y!}, \qquad y \in 0, 1, 2, \ldots, \ \lambda > 0$$

  - **Does the Poisson belong to the Exponential family of distributions?**
    - Rewritten $f(y|\lambda)$ as

$$f(y|\lambda) = \exp(y \log \lambda - \lambda - \log y!)$$

    - Compare with (1.5.1) :

$$a(y) = y, \ b(\lambda) = \log \lambda, \ c(\lambda) = -\lambda, \ d(y) = -\log y!. \qquad (1.5.4)$$

  - The Poisson distribution expresses the probability of a given number of events occurring in a **fixed interval of time and/or space** if these events occur with a **known average rate** and **independently** of the time since the last event.

Code

1.5 Exponential family of distributions

- **Properties of distributions in the exponential family**
  - The expected value and variance of $a(\mathrm{Y})$ are given by

$$\mathbb{E}[a(\mathrm{Y})] = -\frac{c'(\theta)}{b'(\theta)} \tag{1.5.5}$$

and

$$\mathbb{V}ar[a(\mathrm{Y})] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{b'(\theta)^3} \tag{1.5.6}$$

Proof

1.5 Exponential family of distributions

$f_Y(y) = \exp(a(y)b(\theta) + c(\theta) + d(y))$

- **Score and information**
  - The log-likelihood function for the exponential family is:

$$\ell(\theta; y) = a(y)b(\theta) + c(\theta) + d(y).$$

  - The score statistic is

$$\mathrm{U}(\theta; y) = \frac{d\ell(\theta; y)}{d\theta} = a(y)b'(\theta) + c'(\theta)$$

    which depends on y and can be interpreted as a random variable:

$$\mathrm{U} := \mathrm{U}(\theta; Y) = a(\mathrm{Y})b'(\theta) + c'(\theta).$$

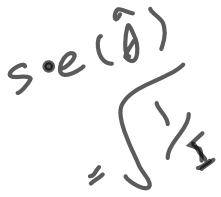
  - By (1.5.5), we have

$$\mathbb{E}(\mathrm{U}) = b'(\theta)\mathbb{E}[a(\mathrm{Y})] + c'(\theta) = 0 \qquad (1.5.7)$$

  - The variance of U or the information, $\mathcal{I}$, is:

$$\mathcal{I} = \mathbb{V}ar(\mathrm{U}) = b'(\theta)^2\mathbb{V}ar[a(\mathrm{Y})] = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta). \qquad (1.5.8)$$

1.5 Exponential family of distributions



- **Cont. Score and information**
  - Another property of the score function U is

$$\mathbb{E}(U^2) = \mathbb{V}\text{ar}(U) = -\mathbb{E}\left(\frac{dU}{d\theta}\right) = I$$

- The first follows from $\mathbb{V}\text{ar}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$, and $\mathbb{E}(U) = 0$.
- For the second equality, note that

$$\mathbb{E}\left(\frac{dU}{d\theta}\right) = \mathbb{E}\left(a(Y)b''(\theta) + c''(\theta)\right) = b''(\theta)\mathbb{E}[a(Y)] + c''(\theta)$$

$$= b''(\theta)\left[-\frac{c'(\theta)}{b'(\theta)}\right] + c''(\theta) = -\mathbb{V}\text{ar}(U) = -\mathcal{I}$$

1.5 Exponential family of distributions

- **Weibull distribution**
  - Suppose $Y \sim \text{Weibull}(\theta, \lambda)$, with the pdf:

$$f(y;\theta,\lambda) = \frac{\lambda y^{\lambda-1}}{\theta^{\lambda}} \exp\left[-\left(\frac{y}{\theta}\right)^{\lambda}\right]$$

  - **Does the Weibull belong to the Exponential family of distributions?**
    - Rewritten $f(y|\theta,\lambda)$ as

$$f(y;\theta) = \exp\left[\log\lambda + (\lambda-1)\log y - \lambda\log\theta - \left(\frac{y}{\theta}\right)^{\lambda}\right]$$

it is not in canonical form except $\lambda = 1$

    - Compare with (1.5.1) :

$$a(y) = y^{\lambda},\ b(\theta) = -\theta^{-\lambda},\ c(\theta) = \log\lambda - \lambda\log\theta,\ d(y) = (\lambda-1)\log y$$

    λ is a nuisance parameter.

1.6 Inference

The two main tools to make statistical conclusions are

- **confidence intervals**: the width of a confidence interval provides a measure of the precision of the **point estimates**.
- **hypothesis testing**: compares how well two related models fit the data.
  The logic for hypothesis testing:
  - specify a model M_0 corresponding to H_0 and a more general model M_1 corresponding to H_1
  - fit model M_0 and compute measure of goodness of fit, G_0; repeat for M_1 to obtain G_1
  - calculate the improvement in fit
  - test the null hypothesis $G_0 = G_1$
  - if $G_0 = G_1$ is not rejected, then H_0 is not rejected and M_0 is the preferred model

1.6 Inference

For **confidence intervals** and **hypothesis testing**, **sampling distributions** are required.

- **normally distributed** r.v. $\Rightarrow$ **exact**.
- **other distributions** $\Rightarrow$ large-sample asymptotic results based on the **CLT**. Under appropriate conditions (i.i.d and S being a sum), the statistic S is

$$\frac{S - \mathbb{E}(S)}{\sqrt{\mathbb{V}\text{ar}(S)}} \dot{\sim} \mathcal{N}(0,1) \tag{1.6.1}$$

or equivalently

$$\frac{[S - \mathbb{E}(S)]^2}{\mathbb{V}\text{ar}(S)} \dot{\sim} \chi^2_1 \tag{1.6.2}$$

and, in case, of p-multivariate statistics

$$[\mathbf{S} - \mathbb{E}(\mathbf{S})]^\top \mathbf{V}^{-1} [\mathbf{S} - \mathbb{E}(\mathbf{S})] \dot{\sim} \chi^2_p \tag{1.6.3}$$

Note: $\dot{\sim}$ symbol means "approximately distributed as".

1.6 Inference

- **Sampling distribution for score statistics**
  Suppose $Y_1, \ldots, Y_N$ are independent random variables with a distribution from the exponential family, with parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.
  - The **score statistics** are such that

$$E[U_j] = 0, \quad \text{for all } j = 1, \ldots, p. \tag{1.6.4}$$

  - The **The variance-covariance matrix of the score statistics** is the information matrix $\mathcal{I}$, with elements

$$\mathcal{I}_{jk} = \mathbb{E}[U_j U_k] \tag{1.6.5}$$

  - If $p = 1$,

$$\frac{U}{\sqrt{\mathcal{I}}} \dot{\sim} \mathcal{N}(0,1), \qquad \equiv \qquad \frac{U^2}{\mathcal{I}} \dot{\sim} \chi^2_1$$

  - If $p > 1$,

$$\mathbf{U} \dot{\sim} MVN(\mathbf{0}, \mathcal{I}) \qquad \equiv \qquad \mathbf{u}^\top \mathcal{I}^{-1} \mathbf{u} \dot{\sim} \chi^2_p$$

1.6 Inference

$\frac{\partial \log p}{\partial p} = \frac{1}{p}$

$E(Y) = np \quad Var(Y) = np(1-p)$

- **Example: Binomial distribution** Let $Y \sim Bin(n, p)$, the log-likelihood function and the score statistic are, respectively,

$$\ell(p) = y \log p + (n-y) \log(1-p) + \log \binom{n}{y}, \text{ and } U = \frac{d\ell}{dp} = \frac{Y}{p} - \frac{n-Y}{1-p} = \frac{Y-np}{p(1-p)}$$

Therefore,

$$\mathbb{E}(Y) = np \Rightarrow \mathbb{E}(U) = 0$$

$E(U) = E\left(\frac{Y-np}{p(1-p)}\right) = \frac{1}{p(1-p)}\left[E(Y)-np\right] = 0$

and

$$\mathbb{V}\text{ar}(Y) = np(1-p) \Rightarrow \mathcal{I} = \mathbb{V}\text{ar}(U) = \frac{1}{p^2(1-p)^2}\mathbb{V}\text{ar}(Y) = \frac{n}{p(1-p)}$$

and, hence,

$$\frac{U}{\sqrt{\mathcal{I}}} = \frac{Y-np}{\sqrt{np(1-p)}} \dot{\sim} \mathcal{N}(0,1)$$

This is known as the normal approximation to the binomial distribution.

1.6 Inference

- **Taylor approximation**
  Taylor approximations for generic functions f in a neighbourhood of t:

$$f(x) = f(t) + (x-t)\left[\frac{df}{dx}\right]_{x=t} + \frac{1}{2}(x-t)^2\left[\frac{d^2f}{dx^2}\right]_{x=t} + \dots$$

  The first three terms of the Taylor approximation for
  - log-likelihood with $p = 1$:

log likelihood function

$$\ell(\theta) = \ell(\hat{\theta}) + (\theta - \hat{\theta})\mathrm{U}(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^2\mathcal{I}(\hat{\theta}), \quad (\hat{\theta} \text{ is MLE of } \theta): \qquad (1.6.6)$$

  - log-likelihood with p-dimensional vector: $\boldsymbol{\theta}$

$$\ell(\boldsymbol{\theta}) = \ell(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathbf{u}(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathcal{I}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \qquad (1.6.7)$$

1.6 Inference

- **Taylor approximation**
  The first two terms of the Taylor approximation for
  - the score function parameter θ:

$$\mathsf{U}(\theta) \approx \mathsf{U}(\hat{\theta}) + (\theta - \hat{\theta})\mathsf{U}'(\hat{\theta}) = \mathsf{U}(\hat{\theta}) - (\theta - \hat{\theta})\mathcal{I}(\hat{\theta}) \tag{1.6.8}$$

  - the score function of a p-dimensional parameter $\boldsymbol{\theta}$:

$$\mathbf{u}(\boldsymbol{\theta}) \approx \mathbf{u}(\hat{\boldsymbol{\theta}}) - \mathcal{I}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \tag{1.6.9}$$

1.6 Inference

- **Sampling distribution of $\hat{\theta}$ (MLE of θ)**
  - The MLE is the estimator which maximises $\ell(\hat{\theta})$, i.e. $\mathbf{u}(\hat{\theta}) = \mathbf{0}$. Therefore,

$$\mathbf{u}(\boldsymbol{\theta}) \approx -\mathcal{I}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

$$-\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{u}(\boldsymbol{\theta}) \approx (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

$\mathbb{E}(U^2) = I$

$\mathbb{E}(UU^T) = I_{p \times p}$

- **Properties**
  - consistency

$$\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$$

  - variance-covariance matrix

$$\mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top\right] = \mathbb{E}[\mathcal{I}^{-1}\mathbf{u}\mathbf{u}^\top\mathcal{I}^{-1}]$$

$$= \mathcal{I}^{-1}\mathbb{E}[\mathbf{u}\mathbf{u}^\top]\mathcal{I}^{-1} = \mathcal{I}^{-1}$$

  - asymptotic sampling distribution

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top\mathcal{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \dot{\sim} \chi^2(p)$$

1.6 Inference

- **Remarks**
  - $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \mathcal{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is also known as **Wald statistics**.
  - for $p = 1$,
$$\hat{\theta} \dot{\sim} \mathcal{N}(\theta, \mathcal{I}^{-1}).$$
  - if the response variable is normally distributed, the results are exact; for other GLM, the results are asymptotic

1.7 Deviance

To assessing the adequacy of a model, compare it with a more general model with the **maximum number of parameters** that can be estimated, called **saturated (maximal/full) model**.

- For observations $Y_i, i = 1, \ldots, N$, a saturated model can be specified with N parameters.
  - In general, the maximum number of parameters m that can be estimated is smaller than N, e.g., if observations are repeated.
- Let $\boldsymbol{\theta}_{\max}$ be the parameter vector of the saturated model with the MLE $\hat{\boldsymbol{\theta}}_{\max}$.
- The likelihood for the saturated model $L(\hat{\boldsymbol{\theta}}_{\max}; \mathbf{y})$ will be larger than any other likelihood function for these observations, because it provides the most complete description of the data.

1.7 Deviance

$L(\hat{\theta}; \mathbf{y})$: maximum value of the likelihood function for the **model of interest**.

- To assess the goodness of fit for the model, use the **likelihood ratio**

$$\lambda = \frac{L(\hat{\theta}_{max}; \mathbf{y})}{L(\hat{\theta}; \mathbf{y})}$$

- In practice, use

$$\log \lambda = \ell(\hat{\theta}_{max}; \mathbf{y}) - \ell(\hat{\theta}; \mathbf{y})$$

- Large values of $\log \lambda$ suggest poor performance of the model of interest relative to the saturated model.

1.7 Deviance

The **Deviance** or **log likelihood ratio statistic** is defined as

$$D = 2[\ell(\hat{\theta}_{max}; \mathbf{y}) - \ell(\hat{\theta}; \mathbf{y})].$$

We know that

$$\ell(\boldsymbol{\theta}; \mathbf{y}) - \ell(\hat{\boldsymbol{\theta}}; \mathbf{y}) = -\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathcal{I}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \equiv 2[\ell(\hat{\boldsymbol{\theta}}; \mathbf{y}) - \ell(\boldsymbol{\theta}; \mathbf{y})] = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathcal{I}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

and consequently,

$$2[\ell(\hat{\boldsymbol{\theta}}; \mathbf{y}) - \ell(\boldsymbol{\theta}; \mathbf{y})] \dot{\sim} \chi^2_p$$

1.7 Deviance

We now have:

$$\begin{aligned} D &= 2[\ell(\hat{\boldsymbol{\theta}}_{max}; \mathbf{y}) - \ell(\hat{\boldsymbol{\theta}}; \mathbf{y})] \\ &= 2[\ell(\hat{\boldsymbol{\theta}}_{max}; \mathbf{y}) \pm \ell(\boldsymbol{\theta}_{max}; \mathbf{y}) \pm \ell(\boldsymbol{\theta}; \mathbf{y}) - \ell(\hat{\boldsymbol{\theta}}; \mathbf{y})] \\ &= 2[\ell(\hat{\boldsymbol{\theta}}_{max}; \mathbf{y}) - \ell(\boldsymbol{\theta}_{max}; \mathbf{y})] - 2[\ell(\hat{\boldsymbol{\theta}}; \mathbf{y}) - \ell(\boldsymbol{\theta}; \mathbf{y})] \\ &\quad + 2[\ell(\boldsymbol{\theta}_{max}; \mathbf{y}) - \ell(\boldsymbol{\theta}; \mathbf{y})] \end{aligned}$$

- $2[\ell(\hat{\boldsymbol{\theta}}_{max}; \mathbf{y}) - \ell(\boldsymbol{\theta}_{max}; \mathbf{y})] \sim \chi^2_m$, where m is the number of parameters in the saturated model
- $2[\ell(\hat{\boldsymbol{\theta}}; \mathbf{y}) - \ell(\boldsymbol{\theta}; \mathbf{y})] \sim \chi^2_p$, where p is the number of parameters in the model of interest
- $2[\ell(\boldsymbol{\theta}_{max}; \mathbf{y}) - \ell(\boldsymbol{\theta}; \mathbf{y})] \geq 0$, which is zero if the model of interest has a fit which is as good as the saturated model

1.7 Deviance

Therefore, the sampling distribution of the deviance is

$$D \dot{\sim} \chi^2_{m-p}(\nu) \tag{1.7.1}$$

where $\nu = 2[\ell(\boldsymbol{\theta}_{\max}; \mathbf{y}) - \ell(\boldsymbol{\theta}; \mathbf{y})]$ is a non-centrality parameter.

- **Remark**
  - The distribution is exact if the response variable is normally distributed
  - For some other distributions, D can be calculated and used directly as a goodness of fit statistic

1.7 Deviance

- **Example: Binomial distribution**
  For the independent response variables $Y_1, \ldots, Y_N$ with $Y_i \sim \text{Bin}(n_i, p_i)$,

$$\ell(\mathbf{p}; \mathbf{y}) = \sum_{i=1}^{N} \left[Y_i \log p_i - Y_i \log(1 - p_i) + n_i \log(1 - p_i) + \log \binom{n_i}{Y_i} \right]$$

  - For a saturated model, the p_i's are all different with $\hat{p}_i = \frac{Y_i}{n_i}$. So,

$$\ell(\hat{\boldsymbol{p}}_{max}; \mathbf{y}) = \sum_{i=1}^{N} \left[Y_i \log \left(\frac{Y_i}{n_i} \right) - Y_i \log \left(\frac{n_i - Y_i}{n_i} \right) + n_i \log \left(\frac{n_i - Y_i}{n_i} \right) + \log \binom{n_i}{Y_i} \right]$$

  - For any other model, $p < N$; let's $\hat{p}_i^*$ be the MLE for a non-saturated model and $\hat{Y}_i = n_i \hat{p}_i^*$ the fitted values; then

$$\ell(\hat{\boldsymbol{p}}^*; \mathbf{y}) = \sum \left[Y_i \log \left(\frac{\hat{Y}_i}{n_i} \right) - Y_i \log \left(\frac{n_i - \hat{Y}_i}{n_i} \right) + n_i \log \left(\frac{n_i - \hat{Y}_i}{n_i} \right) + \log \binom{n_i}{Y_i} \right]$$

1.7 Deviance

- **Cont. Example: Binomial distribution**
  Therefore, the deviance for binomial distribution is:

$$D = 2\sum_{i=1}^{N}\left[\mathrm{Y}_i \log\left(\frac{\mathrm{Y}_i}{\hat{\mathrm{Y}}_i}\right) + (n_i - \mathrm{Y}_i)\log\left(\frac{n_i - \mathrm{Y}_i}{n_i - \hat{\mathrm{Y}}_i}\right)\right].$$

1.7 Deviance

- **Nested model**
  We say that model M_0 is nested in model M_1 if M_0 results as a special case of M_1.

Example

If we partition θ as

$$\theta^\top = \left(\theta^{(1)^\top}, \theta^{(2)^\top}\right)$$

where θ has length p and $\theta^{(1)}$ has length q, then model M_1 could assume unrestricted θ, whereas M_0 restricts, e.g, $\theta^{(2)} = \mathbf{0}$.

1.7 Deviance

- **Cont. Nested model**
  Let M_0 (with q parameters) be nested in M_1 (with p Parameters) [$q < p < N$]. Then

$$\begin{aligned}\Delta D = D_0 - D_1 &= 2[\ell(\hat{\boldsymbol{\theta}}_{\max}; \mathbf{y}) - \ell(\hat{\boldsymbol{\theta}}_0; \mathbf{y})] - 2[\ell(\hat{\boldsymbol{\theta}}_{\max}; \mathbf{y}) - \ell(\hat{\boldsymbol{\theta}}_1; \mathbf{y})] \\ &= 2[\ell(\hat{\boldsymbol{\theta}}_1; \mathbf{y}) - \ell(\hat{\boldsymbol{\theta}}_0; \mathbf{y})]\end{aligned}$$

  Since $D_0 \sim \chi^2(N - q)$ and $D_1 \sim \chi^2(N - p)$, it can be concluded that, for large N,

$$\Delta D \sim \chi^2(p - q)$$

  If the values of ΔD is in the critical region, then model M_1 provides a significantly better description of the data.