Theory of Inference - Reviewed Notes

Dom Hutchinson

March 18, 2020

TODO - JAGS & RJAGS

Contents

1 General

1.1 Approaches to Inference

Definition 1.1 - *Statistical Inference*
Statistical Inference is the process of taking some data and infering a property of the world from it. This is done by theorising a *Statistical Model* which may have generated the data and then calculating parameters for it from the data.

Definition 1.2 - *Statistical Model*
Statistical Models are a, simplified, mathematical description for how a set of data could have been generated. In particular, a *Statistical Model* describes the random variability in the data generating process.

Definition 1.3 - *Frequentist Inference*
The *Frequentist Approach* to *Statistical Inference* treats model unknowns (paramters or functions) as fixed states of nature whose values we want to estimate.
There is no modelling of random variability and thus any that occurs during data collection will be inherited by the model.

Remark 1.1 - *Frequentist Inference*
Often in *Frequentist Inference* we use *asymptotic results* which only become exact as the sample size tends to infty. This has practical drawbacks.

Definition 1.4 - *Bayesian Inference*
The *Bayesian Approach* to *Statistical Inference* treats unknown model parameters as random variables. We define our initial uncertainty about parameter values (the *Prior Distribution*, $\mathbb{P}(\Theta)$), observed data is used to update these distributions in order to reach a *Posterior Distribution*, $\mathbb{P}(\Theta|X)$.
N.B. This is done by using *Bayes' Theorem*.

Remark 1.2 - *Bayesian Inference*
Often in *Bayesian Inference* we use *simulation methods*, which only become exact as the sample size tends to infty. Again, there are practical drawbacks to this.

Remark 1.3 - *Statistical Design*
When trying to infer a model from data there are a few common questions we ask

i) What range of parameter values are consistent with the data?

ii) Which of several alternative models could most plausibly have generated the data?

iii) Could our model have generate the data at all?

iv) How coudl we better arrange the data gatehering process to improve the ansers to the preceeding questions?

1.2 Models

Definition 1.5 - *Nested Models*
Let $\mathbf{X}_1 \sim f_1(\cdot;\boldsymbol{\theta}_1),\ \mathbf{X}_2 \sim f_2(\cdot;\boldsymbol{\theta}_2)$ for $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_1$.
If $\boldsymbol{\theta}_1 \subset \boldsymbol{\theta}_2$ then $\mathbf{X}_1$ is *Nested* in $\mathbf{X}_2$.

Definition 1.6 - *Predictor Variables*
Predictor Variables are the dependent variables of a system, whose values we observe.
N.B. Typically denoted $\mathbf{x}$ or $\mathbf{X}$.

Definition 1.7 - *Metric*
Metrics are *Predictor Variables* which measure an explict quantity.

Definition 1.8 - *Factor*
Factors are *Predictor Variables* which act as labels to whether an observation belongs in a particular class due a property which cannot be explicitly quantified. (*e.g.* Male or Female).

Definition 1.9 - *Response Variables*
Response Variables are the independent variables of a system, whose value we observe.
N.B. Typically denoted y or $\mathbf{y}$.

Definition 1.10 - *Fitted Values, $\hat{y}$*
Fitted Values are our estimated values for the *Response Variable*.

$$\hat{y}_i := f(\mathbf{x}_i)$$

Definition 1.11 - *Regular Models*
Let $\mathbf{X} \sim f(\cdot;\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ be a *Statistical Model*.
A *Statistical Model* is deemed *Regular* if it fulfils all the following:

i) Densities for distinct $\boldsymbol{\theta}$ are distinct.
N.B. If not, parameters won't be identifiable & thus no guaranteed consistency.

ii) $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$.
N.B. Otherwise we cannot approximat *Log-Likelihood* by *Taylor Expansion* in the region of $\boldsymbol{\theta}^*$.

iii) Within some neighbourhood of $\boldsymbol{\theta}^*$

- The first three derivatives of the *Log-Likelihood* exist & are bounded.
- $\mathcal{I} := \mathbb{E}\left(\left.\frac{\partial \ell}{\partial \boldsymbol{\theta}}\right|_{\theta^*} \left.\frac{\partial \ell}{\partial \boldsymbol{\theta}^T}\right|_{\theta^*}\right) \equiv -\mathbb{E}\left(\left.\frac{\partial^2 \ell}{\partial \boldsymbol{\theta}\boldsymbol{\theta}^T}\right|_{\theta^*}\right)$ is satisfied.
- The *Fisher Information Matrix*, $\mathcal{I}$, is positive-definite & finite.

N.B. These requirements are required for certain results.

1.3 Graphical Models

Definition 1.12 - *Directed Acyclic Graphs*
A *Directed Acyclic Graph*, DAG, is a graph which consists of directed edges and no cycles.
The direction of edges show dependence with the target element being dependent on the origin element.
Proposition 1.1 - *Graphical Model*
A *Model* can be represented graphically using a *Directed Acyclic Graph* with variables for nodes & dependence from the edges.
The distribution of a variable is completely known if you know the values of all its parent nodes.
The nodes in a *Graphical Model* take three types

i) *Stocastic Nodes* are variables with a distribution which depends stochastically on otehr nodes.
These can be observed or unobserved.

ii) *Deterministic Nodes* are nodes that are deterministic functions of other nodes. They cannot be observed.

iii) *Constant Nodes* are fixed numbers and have no parents.

The edges in a *Graphical Model* show one of two relationship types

i) *Deterministic Relationship* between nodes (usually a dashed arrow).

ii) *Stochastic Relationship* between nodes (usually a solid arrow).

Proposition 1.2 - *Distributions from Graphical Model*
Let x_i represent the variable associated with the i^{th} node in the graph.
We have joint distribution of the non-constant nodes

$$f(\mathbf{x}) = \prod_i f(x_i|\text{Parent}(x_i))$$

1.4 Inference by Mathmetical Manipulation

Remark 1.4 - *Inference by Mathematical Manipulation*
Bayesian and *Frequentist Inference* use mathetmatical computaiton to make inferences about parameter valus & their uncertainty. An alternative approach is to use mathematical manipulation, rather than computation.

Definition 1.13 - *Bootstrapping*
Let X be a set of observed data.
Bootstrapping is a *simulation* of the data gathering process.
In *Bootstrapping* we uniformly sample values from X with replacement until we reach a desired sample size (often $|X|$).

Proposition 1.3 - *Inference by Resampling*
Once we have used *Bootstrapping* to generate a set of new data-sets we can use *Bayesian* & *Frequentist Inference* techniques in order to estimate parameter values.

Remark 1.5 - *Bootstrap Interval*
An *Interval* generated from *Bootstrapping* datasets are generally narrower than those produced by *Bayesian* or *Frequentist Approaches*.
N.B. This discrepancy is reduced as sample size increases.

Proposition 1.4 - *Bootstrap Percentiles*
When wishing to create an interval for θ using *Bootstrapped* data, we treat the $\hat{\theta}$ values as if they came from $\mathbb{P}(\mu|X)$.

1.5 Causality

Definition 1.14 - *Causality*
Causality is a relationship between two events where one of the events caused the other.

Definition 1.15 - *Correlation*
Causality is a relationship between two events where the events are likely to occur together. This does not mean that one event has caused the other, but they may have been caused by the same variable (which may be hidden).

Remark 1.6 - *When two variables are highly correlated it is hard to distinguish their effects*

Definition 1.16 - *Counfounding Variable*
A *Confounding Variable* is a variable which influences both the *Predictor* & *Response Variables* in a system.
Suppose x, h are highly-correlated and $y = \beta_0 + \beta_1 x$. The model $y = \beta_2 + \beta_3 h$ would appear statistically good even though h had no part in geneating y. Here x is the *Confounding Variable*.
N.B. AKA *Hidden Variables*

1.5.1 Controlled Experiments

Definition 1.17 - *Randomisation*
Randomisatiation is a technique used when designing experiments to break relationships between *Confounder Variables* and our *Response Variable*.
Randomisation involves taking a set of subjects and randomly assigning them to different "treatments".
Provided these treaments only vary the *Predictor Variable* we wish to test, this breaks association between other *Prectior Variables* & our *Response Variable*. Making these *Predictor Variables* now part of the ranom variability of the model, ε.

Remark 1.7 - *Randomisation is the Gold Standard for Inferring Causation*

Proposition 1.5 - *Mathematical Justification*
Consider model matrix $(\mathbf{X}, \mathbf{H})$ where $\mathbf{X}$ is formed from observed *Predictor Variables* & $\mathbf{H}$ is from *Confounding Variables* (*N.B.* we would not know its value in practice).
Assume the oolumns of $\mathbf{H}$ are centred on 0.
We now have *Least Squares Estimate* for the parameters of

$$\begin{pmatrix} \tilde{\boldsymbol{\beta}}_X \\ \tilde{\boldsymbol{\beta}}_H \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{H} \\ \mathbf{H}^T\mathbf{X} & \mathbf{H}^T\mathbf{H} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}^T \\ \mathbf{H}^T \end{pmatrix} \mathbf{y}$$

If $\mathbf{X}, \mathbf{H}$ are dependent on each other then $\mathbf{X}^T\mathbf{H} \neq \mathbf{0} \implies \tilde{\boldsymbol{\beta}}_X \neq (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \hat{\boldsymbol{\beta}}_X$.
If $\mathbf{X}, \mathbf{H}$ are <u>independent</u> of each other then $\mathbf{X}^T\mathbf{H} = \mathbf{0}$, for a large sample size,
$\implies \tilde{\boldsymbol{\beta}}_X = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \hat{\boldsymbol{\beta}}_X$.
N.B. Randomisation ensures $\mathbf{X}$ and $\mathbf{H}$ are independent.

1.5.2 Instrumental Variables

Remark 1.8 - *Motivation*
In some scenarios it is impractical or unethical to perform randomised experiments.

Definition 1.18 - *Instrumental Variables*
Let $(\mathbf{X}, \mathbf{H})$ be a *Model Matrix* with $\mathbf{X}$ observed & $\mathbf{H}$ confounding for our *Response Variable*, $\mathbf{y}$.
A variable is an *Instrumental Variable* if

- It is <u>not</u> part of the true model of the **Response Variable**;
- It <u>is</u> correlated with the *Predictor Variables* in $\mathbf{X}$;

And, It is <u>not</u> correlated with the *Confounding Variables* in $\mathbf{H}$.

Remark 1.9 - *Finding Instrumental Variables is Hard, very!*

Proof 1.1 - *Confounding Variables affect Least Squares Estimate in Linear Models*
Let $(\mathbf{X}, \mathbf{H})$ be a *Model Matrix* where $\mathbf{X}$ comes from some observed variables & $\mathbf{H}$ is from *Confounding Variables*.
This means the true model is of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_X + \mathbf{H}\boldsymbol{\beta}_H + \boldsymbol{\varepsilon}$$

Since we only know $\mathbf{X}$ we try to fit $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_X + \mathbf{e}$.
Effectively $\mathbf{e} = \mathbf{H}\boldsymbol{\beta}_H + \boldsymbol{\varepsilon}$, which is unlikey to fulfil the assumptions of indepdence & constant variance.
This is a problem for the model. Further

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}_X) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}, \mathbf{H})\boldsymbol{\beta} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_X + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{H}\boldsymbol{\beta}_H \\
&= \boldsymbol{\beta}_X + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{H}\boldsymbol{\beta}_H \\
&\neq \boldsymbol{\beta}_X
\end{aligned}$$

N.B. The space spanned by the columns of $\mathbf{X}$ is not orthogonal to $\mathbf{e}$.

Proposition 1.6 - *Instrumental Variables for Linear Models*
Let $(\mathbf{X}, \mathbf{H})$ be a *Model Matrix* where $\mathbf{X}$ comes from some observed variables & $\mathbf{H}$ is from *Confounding Variables*.
Let $\mathbf{Z}$ be the *Model Matrix* for some *Instrumental Variables*.
We assume $\text{Rank}(\mathbf{Z}) \geq \text{Rank}(\mathbf{X})$.
Perform the projection of $\mathbf{X}$ onto the column space of $\mathbf{Z}$. Giving

$$\mathbf{X}_Z := \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{X} = \mathbf{A}_Z\mathbf{X}$$

This gives us the least squares estimate of the model parameters

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

Since $\mathbf{Z}$ is independent of $\mathbf{H}$. $\mathbf{Z}^T\mathbf{H} \simeq \mathbf{0} \implies \mathbf{A}_Z\mathbf{H} \simeq \mathbf{0}$. Thus

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}_X) &= (\mathbf{X}^T\mathbf{A}_Z\mathbf{X})^{-1}\mathbf{X}^T A_Z(\mathbf{X}, \mathbf{H})\boldsymbol{\beta} \\
&= (\mathbf{X}\mathbf{A}_Z\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}_Z\mathbf{X}\boldsymbol{\beta}_X + \underbrace{(\mathbf{X}\mathbf{A}_Z\mathbf{X})^{-1}\mathbf{X}^T \underbrace{\mathbf{A}_Z}_{\simeq 0}\mathbf{H}\boldsymbol{\beta}_H}_{\simeq 0} \\
&= (\mathbf{X}\mathbf{A}_Z\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}_Z\mathbf{X}\boldsymbol{\beta}_X \\
&= \boldsymbol{\beta}_X
\end{aligned}$$

This shows the estimate produced by using *Instrumental Variables* produces a much more accurate estimate of the model parameters, than when they are not used & *Convolution Variables* exist.

1.6 Hypothesis Testing

Remark 1.10 - *Hypothesis Testing is mostly a Frequentist method, not Bayesian.*

Remark 1.11 - *Hyphothesis Testing amounts to choosing the simplist model which has no obvious incosistencies with the data*

Definition 1.19 - *Simple Hypothesis*
A *Simple Hypothesis* states that a parameter takes an exact value.

i.e. $\theta = \theta_0$ for $\theta_0 \in \Theta$.

Definition 1.20 - *Composite Hypothesis*
A *Composite Hypothesis* states that a parameter takes a value from a set.
i.e. $\theta \in \Theta_0$ for $\Theta_0 \subseteq \Theta$.

Definition 1.21 - *Test Statistic*
A *Test Statistic* is a random variable whose value depends on the observed set of data.
Test Statistics are used to assess the likelihood of observing a certain data set under a given *Null Hypothesis* in *Hypothesis Testing*.

Definition 1.22 - *p-Value*
p-Value measures the goodness of fit of statistical models.
The *p-Value* is the probability of observing more extreme that the data used for fitting, under the theorised model, $\boldsymbol{\Theta}_0$.
Let $\mathbf{X} \sim f(\cdot;\boldsymbol{\theta})$ and $\mathbf{x}$ be a realisation.

$$p(\mathbf{x}) := \sup_{\theta \in \Theta_0} \mathbb{P}(\mathbf{X} \geq \mathbf{x};\boldsymbol{\theta}) = \sup_{\theta \in \Theta_0} f(\mathbf{x};\boldsymbol{\theta})$$

A higher *p-Value* suggests a better model fit.

Remark 1.12 - *p-Values are good at testing the fit of a model, but not at comparing the relative fit of two models*

Remark 1.13 - $p(\boldsymbol{x})$ *is the smallest Significance Level at which we would reject the Null Hypothesis*

Definition 1.23 - *Hypothesis Testing*
Hypothesis Testing is the porcess of determining which of two hypotheses about model parameters is more consistent with the data.
We define a *Null Hypothesis* and an *Alternative Hypothesis*. The *Null Hypothesis* acts as our default position, and we only reject it is if the observed data is too extreme (given that it is true).
N.B. Null and *Alternative Hypothesis* are mutually exclusive.

Proposition 1.7 - *Process for Hypothesis Testing*
Let $\mathbf{x}$ be a realisation of $\mathbf{X}$

i) Choose a model $f(\cdot;\theta)$ st $\mathbf{X} \sim f(\cdot;\theta)$ for $\theta \in \Theta$.

ii) Define a *Null Hypothesis*, H_0, and an *Alternative Hypothesis*, H_1.

iii) Define a *Test Statistic*, $T(\cdot)$.

iv) Choose a *Significance Level*, α, and calculate the equivalent *Critical Value*, c, for the *Test Statistic*.

v) Calculate value of the *Test Statistic* under the observed data, $t_{\text{obs}} = T(\mathbf{x})$.

vi) If $t_{\text{obs}} \geq c$ then reject H_0 in favour of H_1, otherwise accept H_0.

Definition 1.24 - *Neymann-Pearson Test Statistic*
The *Neymann-Pearson Test Statistic* is a generalisation of the *Likelihood Ratio*.
It measures the likelihood of the *Alternative Hypothesis* being correct, relative to the *Null*

Hypothesis, given the data.
Let $\mathbf{X} \sim f(\cdot;\boldsymbol{\theta})$, $\mathbf{x}$ be a realisation of $\mathbf{X}$ and $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ be our *Null Hypothesis*.

$$T_{np}(\mathbf{x}) := \frac{p(\mathbf{x};\boldsymbol{\Theta})}{p(\mathbf{x};\boldsymbol{\Theta}_0)} = \frac{\sup_{\theta\in\Theta} f(\mathbf{x};\boldsymbol{\theta})}{\sup_{\theta_0\in\Theta_0} f(\mathbf{x};\boldsymbol{\theta}_0)} = \frac{f(\mathbf{x};\hat{\boldsymbol{\theta}}_{\text{MLE}})}{\sup_{\theta_0\in\Theta_0} f(\mathbf{x};\boldsymbol{\theta}_0)} \geq 1$$

Lower values of the *Likelihood Ratio* indicate that H_0 is more likely to be true.

Definition 1.25 - *Power Function, π*
The *Power Function*, $\pi(\cdot)$, measures the probability of rejecting the *Null-Hypothesis* given that another set of parameter values is true (usually test with the *Alternative Hypothesis*).
Let $\mathbf{X} \sim f(\cdot;\theta)$, $T(\cdot)$ be a *Test Statistic* and c be the *Critical Value* of T. Then

$$\pi(\theta_1; T, c) = \mathbb{P}(T(\mathbf{X}) \geq c; \theta_1)$$

Theorem 1.1 - *Neyman-Pearson Lemma*
NON-EXAMINABLE.
Let $\mathbf{X} \sim f(\cdot;\boldsymbol{\theta})$ and $\mathbf{x}$ be a realisation of $\mathbf{x}$.
Consider testing two *Simple Hypothese* with the *Neyman Pearson Test Statistic.*
Choose some $\alpha \in [0,1]$ and find c_{NP} st $\alpha = \mathbb{P}(T_{NP} \geq c_{NP}; \boldsymbol{\theta}_0)$.
Then the test (T_{NP}, c_{NP}) is equivalent to the uniformly most powerful test.

Definition 1.26 - *Generalised Likelihood Ratio Test Statistic*
JUST NEED TO KNOW RESULT.
Let $\mathbf{X} \sim f(\cdot;\boldsymbol{\theta})$ be a model & $\mathbf{x}$ be a realisation of $\mathbf{X}$.
Consider testing two, <u>non-nested</u>, hypotheses

$$H_0 : \mathbf{R}(\boldsymbol{\theta}) = \mathbf{0} \quad \text{against} \quad H_1 : \mathbf{R}(\boldsymbol{\theta}) neq \mathbf{0}$$

where $\mathbf{R}(\cdot)$ is a vector-valued function of $\boldsymbol{\theta}$ st H_0 imposes r restrictions on $\boldsymbol{\Theta}$.
The *Generalised Likelihood Ratio Test Statistic* is defined as

$$T_{GLR} := 2[\ell(\hat{\boldsymbol{\theta}}_{\text{MLE}}) - \sup_{\mathbf{R}(\boldsymbol{\theta})=\mathbf{0}} \ell(\boldsymbol{\theta})] \sim \chi^2_r$$

1.7 Intervals

Definition 1.27 - *Random Interval*
Let $\mathbf{X} \sim f_n(\cdot;\theta^*)$ for $\theta^* \in \Theta$ and $L, U : \mathcal{X}^n \to \Theta$ st $\forall\ \mathbf{x} \in \mathcal{X}^n\ L(\mathbf{x}) < U(\mathbf{x})$.
A *Random Interval* is an *Interval* whose bounds depends on a *Random Variable.*
Here $\mathcal{I}(\mathbf{X}) := [L(\mathbf{X}), U(\mathbf{X})]$ is a *Random Interval.*
N.B. $L(\cdot)$ & $U(\cdot)$ are maps from observed data to parameter values.

Definition 1.28 - *Coverage of an Interval*
Let $\mathcal{I}(\mathbf{X}) := [L(\mathbf{X}), U(\mathbf{X})]$ be a *Random Interval* for θ with true value θ^*.
The *Coverage of an Interval* is the probability that the true value of the parameter it is estimating lies in the inverval.

$$C_{\mathcal{I}} = \mathbb{P}(\theta^* \in \mathcal{I}(\mathbf{X}); \theta^*)$$

Definition 1.29 - *Confidence Interval*
A $1-\alpha$ *Confidence Interval* for a parameter is an interval with *Coverage* at least $1-\alpha$.

$$\mathcal{I}(\mathbf{X}) := [L(\mathbf{X}), U(\mathbf{X})] \text{ is a } 1-\alpha \textit{ Confidence Interval} \text{ if } \mathbb{P}(\theta^* \in \mathcal{I}) \geq 1-\alpha$$

N.B. If $\mathbb{P}(\theta^* \in \mathcal{I}(\mathbf{X})) = 1 - \alpha$ then $\mathcal{I}$ is an *Exact Confidence Interval*.

Remark 1.14 - *Confidence Intervals are a part of Frequentist Statistics, not Bayesian*

Definition 1.30 - *Credible Interval*
Let $\mathbf{X} \sim f(\cdot;\boldsymbol{\theta})$ and $\mathbf{x}$ be a realisation of $\mathbf{X}$.
A $1-\alpha$ *Credible Interval* is as an interval (θ_1, θ_2) where the probability that the true parameter lies in the parameter is $1 - \alpha$.

$$1 - \alpha = \mathbb{P}(\theta^* \in (\theta_1, \theta_2) \iff \int_{\theta_2}^{\theta_1} p(\theta|\theta) d\theta = 1 - \alpha$$

When the model has multiple parameters we find a different intervval for each parameter.
N.B. Multiple such intervals will exist and we find these intervals by sampling.

Remark 1.15 - *Credible Intervals are a part of Bayesian Statistics, not Frequentist*

Definition 1.31 - *Test Inversion*
Test Inversion is the process of finding a *Confidence Interval/Set* by finding the range of values which would cause the *Null Hypothesis* to be accepted.
N.B. AKA *Wilk's Intervals*.

Proposition 1.8 - *Every value in a Wilk's Interval has a greater likelihood than any value outside.*

1.8 Other ways of Assessing Fit

Definition 1.32 - *Residual*
The *Residual* is the difference between the true value of the *Response Variables* & our *Fitted Values*.

$$\epsilon := |y_i - \hat{y}_i|$$

Definition 1.33 - *Residual Sum of Squares*
The *Residual Sum of Squares* is the sum of the squared value of the *Residuals* for each observation.
The *RSS* is used as a measure for how well our model fits the data

$$RSS := \sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2$$

Remark 1.16 - *More complicated models tend to have greater likelihood*
whether or not the extra complications reflect anything in the true data generating process.

Definition 1.34 - *Kullback-Leibler Divergence*
Kullback-Leibler Divergence is a measure of how similar two models are.
Let $f(\cdot)$ be the true model & $g(\cdot)$ be an approximation.

$$KL(f, g) = \int [\ln f(\mathbf{y}) - \ln g(\mathbf{y})] f(\mathbf{y}) d\mathbf{y}$$

N.B. This is a measure of information loss caused by $g(\cdot)$.

Proposition 1.9 - *Testing Models with KL Divergence*
Let f be a proposed model with true parameters values $\boldsymbol{\theta}^*$.
To compare a proposed set of parameters, $\boldsymbol{\theta}$, we want to calculate

$$KL(f^*, f_\theta) = \int [\ln f(\mathbf{y};\boldsymbol{\theta}^*) - \ln f(\mathbf{y};\boldsymbol{\theta})] f(\mathbf{y};\boldsymbol{\theta}^*) d\mathbf{y}$$

N.B. This is not tractable since f^* is unknown (otherwise we would use it).

Definition 1.35 - *Akaike's Information Criterion*
Akaike's Information Criterion, AIC, is a goodness-of-fit measure for models.

$$AIC(\boldsymbol{\theta}) = -2\ell(\boldsymbol{\theta}) + p$$

where $p := |\boldsymbol{\theta}|$ (the number of estimated parameters).
Lower AIC values indicate better fit.
N.B. The factor of 2 is to put AIC on the scale as the T_{GLR}.

Remark 1.17 - *AIC is non-consistent.*
As $n \to \infty$ the probability of choosing the correct model does not tend to 1.

Proposition 1.10 - *Using KL Divergence*
Let f be a proposed model with true parameters $\boldsymbol{\theta}^*$.
We want to find the set of parameters, $\hat{\boldsymbol{\theta}}$, which perform closely to the true values.
i.e. $\text{argmin}_{\hat{\boldsymbol{\theta}}} KL(f_{\hat{\theta}}, f^*)$ where $f_{\hat{\theta}} := f(\cdot; \hat{\boldsymbol{\theta}})$ & $f^* := f(\cdot; \boldsymbol{\theta}^*)$.
It turns out that $\mathbb{E}(KL(f_{\hat{\theta}_{\text{MLE}}}, f^*))$ is tractable.

$$\hat{\mathbb{E}}(KL(f_{\hat{\theta}_{\text{MLE}}}, f^*)) = -\ell(\boldsymbol{\theta}_{\text{MLE}}) + p + \int \ell(\boldsymbol{\theta}^*; \mathbf{y}) f(\mathbf{y}; \boldsymbol{\theta}^*) d\mathbf{y}$$

Where $p := |\boldsymbol{\theta}|$ (the number of estimaed parameters).
Note that this involves the true parameter value, thus is minised by whichever model minimises *Akaike's Information Criterion.*

Proof 1.2 - $\hat{\mathbb{E}}(KL(f_{\hat{\theta}_{MLE}}, f^*)) = -\ell(\boldsymbol{\theta}_{MLE}) + p + \int \ell(\boldsymbol{\theta}^*; \boldsymbol{y}) f(\boldsymbol{y}; \boldsymbol{\theta}^*) d\boldsymbol{y}$
Define $\boldsymbol{\theta}_{KL} := \text{argmin}_{\boldsymbol{\theta}} KL(f_\theta, f^*)$, the parameters which minimise *KL Divergence.*
By *Taylor's Theorem*

$$\begin{aligned}
\ln f_{\hat{\theta}_{\text{MLE}}}(\mathbf{y}) &\simeq \ln f_{\theta_{KL}}(\mathbf{y}) + (\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_{KL})^T \frac{\partial \ln f_\theta}{\partial \theta}\bigg|_{\theta_{KL}} \\
&+ \frac{1}{2}(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_{KL})^T \frac{\partial^2 \ln f_\theta}{\partial \theta \partial \theta^T}\bigg|_{\theta_{KL}} (\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_{KL}) \\
\implies KL(f_{\hat{\theta}_{\text{MLE}}}, f^*) &\simeq \int \ln f^*(\mathbf{y}) - f^*(\mathbf{y}) \Bigg[\ln f_{\theta_{KL}}(\mathbf{y}) + (\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_{KL})^T \frac{\partial \ln f_\theta}{\partial \theta}\bigg|_{\theta_{KL}} \\
&+ \tfrac{1}{2}(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_{KL})^T \frac{\partial^2 \ln f_\theta}{\partial \theta \partial \theta^T} (\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_K) \Bigg] d\mathbf{y} \\
&= \int \ln f^*(\mathbf{y}) - f^*(\mathbf{y}) \ln f_{\theta_{KL}}(\mathbf{y}) + (\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_K)^T \underbrace{\frac{\partial \ln f_\theta}{\partial \theta}\bigg|_{\theta_{KL}} f^*(\mathbf{y})}_{=0 \text{ since } \theta_{KL} \text{ minimises}} \\
&+ \frac{1}{2} f^*(\mathbf{y})(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_{KL})^T \frac{\partial^2 \ln f_\theta}{\partial \theta \partial \theta^T}\bigg|_{\theta_{KL}} (\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_{KL})^T d\mathbf{y} \\
&= \int \ln f^*(\mathbf{y}) - f^*(\mathbf{y}) \ln f_{\theta_{KL}}(\mathbf{y}) + (\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_K)^T \\
&+ \frac{1}{2} f^*(\mathbf{y})(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_{KL})^T \frac{\partial^2 \ln f_\theta}{\partial \theta \partial \theta^T}\bigg|_{\theta_{KL}} (\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_{KL})^T d\mathbf{y} \\
&= KL(f_{\theta_{KL}}, f^*) + \int \frac{1}{2}(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_{KL})^T \frac{\partial^2 \ln f_\theta}{\partial \theta \partial \theta^T}\bigg|_{\theta_{KL}} (\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_{KL}) f^*(\mathbf{y}) d\mathbf{y} \\
&= KL(f_{\theta_{KL}}, f^*) + \tfrac{1}{2}(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_{KL})^T \mathcal{I}_{\theta_{KL}} (\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_{KL})
\end{aligned}$$

where $\mathcal{I}_{\theta_{KL}} := \mathbb{E}\left(\frac{\partial \ell}{\partial\theta\partial\theta^T}\Big|_{\theta=\theta^*}\right)$ (Fisher Information Matrix).
Assuming the model, f, is sufficiently regular then for large n

$$\mathbb{E}(\hat{\boldsymbol{\theta}}_{\text{MLE}}) \simeq \boldsymbol{\theta}_{KL} \;\&\; \text{Cov}(\hat{\boldsymbol{\theta}}_{MLE}) \simeq \mathcal{I}_{\theta_{KL}}$$

Since $2[\ell(\hat{\boldsymbol{\theta}}_{MLE} - \ell(\hat{\boldsymbol{\theta}}_0] \sim \chi^2_r$ we have

$$\begin{aligned} \mathbb{E}[\ell(\hat{\boldsymbol{\theta}}_{MLE} - \ell(\hat{\boldsymbol{\theta}}_0] &\simeq \mathbb{E}\left(\tfrac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{KL})\mathcal{I}_{\theta_{KL}}(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}_{KL}\right) \\ &= \tfrac{p}{2} \text{ where } p := |\boldsymbol{\theta}| \text{ by result at end of 8.6} \quad (1) \\ \implies \mathbb{E}(KL(f_{\hat{\theta}_{MLE}}, f^*)) &\simeq KL(f_{\theta_{KL}}, f^*) + \tfrac{p}{2} \quad (2) \end{aligned}$$

This still contains the true value, which is unknown.
We have

$$\begin{aligned} \mathbb{E}(-\ell(\hat{\boldsymbol{\theta}}_{MLE})) &= \mathbb{E}(-\ell(\hat{\boldsymbol{\theta}}_{MLE}) + \ell(\boldsymbol{\theta}_{KL}) - \ell(\boldsymbol{\theta}_{KL})) \\ &= \mathbb{E}(-\ell(\boldsymbol{\theta}_{KL}) - \{\ell(\hat{\boldsymbol{\theta}}_{MLE}) - \ell(\boldsymbol{\theta}_{KL})\}) \\ &= -\mathbb{E}(\ell(\boldsymbol{\theta}_{KL})) - \mathbb{E}(\ell(\hat{\boldsymbol{\theta}}_{MLE}) - \ell(\boldsymbol{\theta}_{KL})) \\ &\simeq -\underbrace{\int \ell(\boldsymbol{\theta}_{KL}) f^*(\mathbf{y}) d\mathbf{y}}_{\text{by def of } \mathbb{E}} - \underbrace{\frac{p}{2}}_{\text{by (1)}} \\ \text{Further} &= -\int \ell(\boldsymbol{\theta}_{KL}) f^*(\mathbf{y}) d\mathbf{y} - \frac{p}{2} + \int \ell(\boldsymbol{\theta}^*) f^*(\mathbf{y}) d\mathbf{y} - \int \ell(\boldsymbol{\theta}^*) f^*(\mathbf{y}) d\mathbf{y} \\ &= \int [\ell(\boldsymbol{\theta}^*) - \ell(\boldsymbol{\theta}_{KL})] f^*(\mathbf{y}) d\mathbf{y} - \int \ell(\boldsymbol{\theta}^*) f^*(\mathbf{y}) d\mathbf{y} - \frac{p}{2} \\ &= \underbrace{KL(f_{\theta_{KL}}, f^*)}_{\text{by def of KL}} - \int \ell(\boldsymbol{\theta}^*) f^*(\mathbf{y}) d\mathbf{y} - \frac{p}{2} \end{aligned}$$

By rearranging the final result

$$\begin{aligned} KL(f_{\theta_{KL}}, f^*) &= \int \ell(\boldsymbol{\theta}^*) f^*(\mathbf{y}) d\mathbf{y} + \frac{p}{2} + \mathbb{E}(-\ell(\hat{\boldsymbol{\theta}}_{MLE})) \\ &= \int \ell(\boldsymbol{\theta}^*) f^*(\mathbf{y}) d\mathbf{y} + \frac{p}{2} - \ell(\hat{\boldsymbol{\theta}}_{MLE}) \qquad (3) \end{aligned}$$

Thus

$$\mathbb{E}(KL(f_{\hat{\theta}_{MLE}}, f^*) \simeq \int \ell(\boldsymbol{\theta}^*) f^*(\mathbf{y}) d\mathbf{y} + p - \ell(\hat{\boldsymbol{\theta}}_{MLE}) \text{ by substitution (3) into (2)}$$

N.B. $f_\theta := f(\cdot;\boldsymbol{\theta})$ & $f^* := f(\cdot;\boldsymbol{\theta}^*)$ (for compactness).

2 Linear Models

Proposition 2.1 - *Implementing Factors*
Suppose a model has *Factor Variable* g_i which seperates observations into n categories.
In the function for y_i, g_i would be represented by a single term γ_{g_i} whose value depends on the value of g_i. (*i.e.* A different weight is assigned to each group).
We want to find the n values which γ_{g_i} can take.
We can express this in terms of matrices, as below, with each row on the LHS denoting which category each observation belongs to

$$\begin{pmatrix} 0 & 1 & 0 & \dots \\ 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_n \end{pmatrix}$$

N.B. Each row has a single 1 element and $n-1$ zero elements.

Remark 2.1 - *We want to find the parameters to the Precitor Variables which produce accurate values for the Response Variables.*

Definition 2.1 - *Model Matrix*
Each element in a *Model Matrix* is a function of the *Predictor Variables.*
Each row depends on a different set of observations.

Definition 2.2 - *Linear Model*
A *Linear Model* is a *Statistical Model* whose response vector, $\mathbf{y}$, is linear wrt its *Model Matrix*, $\mathbf{X}$, and some zero-mean random error, $\boldsymbol{\varepsilon}$.

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

$\mathbf{X} \in \mathbb{R}^{n\times p},\ \boldsymbol{\beta} \in \mathbb{R}^p$ and $\varepsilon \sim \text{Normal}(\mathbf{0}, \sigma^2 I)$.

2.1 Frequentist Approach

Proposition 2.2 - *Frequentist Approach to Linear Models*
In the *Frequentist Approach* to *Linear Models* we treat $\boldsymbol{\beta}$ and σ^2 as fixed (but unknown) states of nature.
Thus all random variability from the data will be inherited into the model.

2.1.1 Estimation

Proposition 2.3 - *Point Value Estimates*
We can make *Point Value Estimates* of parameter values by finding the set of parameters $\boldsymbol{\beta}$ which minimises the *Residual Sum of Squares.*

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{LSE}} &:= \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{N} (y_i - (\mathbf{X}\boldsymbol{\beta})_i)^2 \\ &= \text{argmin}_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2\end{aligned}$$

N.B. This is the *Least Squares Estimate* of $\boldsymbol{\beta}$.

Remark 2.2 - $\hat{\boldsymbol{\beta}}_{LSE} = R^{-1}Q^T\boldsymbol{y}$
where Q, R are from the decomposition of $\mathbf{X}$ st $\mathbf{X} = QR$ with Q being *Orthogonal* and R being *Upper-Triangle.*

Proposition 2.4 - *Deriving Least Squares Estimate for* $\boldsymbol{\beta}$
Let $\mathbf{X}, \mathbf{y}$ be n observed data points & $\boldsymbol{\beta} \in \mathbb{R}^p$ be a parameter vector we are fitting to our model.
We want to find $\hat{\boldsymbol{\beta}}_{\text{LSE}} = \text{argmin}_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$.
Since $\mathbf{X}$ is a real-valued matrix it can be decomposed into

$$\mathbf{X} = \mathcal{Q}\begin{pmatrix} R \\ \mathbf{0} \end{pmatrix} = QR^1$$

where $R \in \mathbb{R}^{p\times p}$ is an upper triangle matrix, $\mathcal{Q} \in \mathbb{R}^{n\times n}$ is orthogonal & $Q \in \mathbb{R}^{n\times p}$ is the first p columns of $\mathcal{Q}$.

[1] Known as *QR Decomposition* and can be performed in R using `qr.Q(qr(X), complete = TRUE)` & `qr.R(qr(X))`

Note that $\mathcal{Q}^T$ is *Orthogonal*.
Thus

$$\begin{aligned}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 &= \left\|\mathbf{y}-\mathcal{Q}\begin{pmatrix}R\\\mathbf{0}\end{pmatrix}\boldsymbol{\beta}\right\|^2\\ &= \left\|\mathcal{Q}^T\mathbf{y}-\mathcal{Q}^T\mathcal{Q}\begin{pmatrix}R\\\mathbf{0}\end{pmatrix}\boldsymbol{\beta}\right\|^2 \text{ since } \mathcal{Q}^T \text{ is orthogonal}\\ &= \left\|\mathcal{Q}^T\mathbf{y}-\begin{pmatrix}R\\\mathbf{0}\end{pmatrix}\boldsymbol{\beta}\right\|^2\end{aligned}$$

Decompose $\mathcal{Q}^T\mathbf{y}=\begin{pmatrix}\mathbf{f}\\\mathbf{r}\end{pmatrix}$ with $\mathbf{f}\in\mathbb{R}^p$ & $\mathbf{r}\in\mathbb{R}^{n-p}$.
Note that $\mathbf{f}=Q^T\mathbf{y}$.
$\mathbf{f}$ is the first p rows of $\mathcal{Q}^T\mathbf{y}$ and $\mathbf{r}$ is the last $n-p$ rows.
Thus

$$\begin{aligned}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 &= \left\|\begin{pmatrix}\mathbf{f}\\\mathbf{r}\end{pmatrix}-\begin{pmatrix}R\\\mathbf{0}\end{pmatrix}\boldsymbol{\beta}\right\|^2\\ &= \|\mathbf{f}-R\boldsymbol{\beta}\|^2+\|\mathbf{r}\|^2\end{aligned}$$

$\|\mathbf{r}\|^2$ is indepdent of $\boldsymbol{\beta}$ and thus irreducible.
This final expression is minimised when $\|\mathbf{f}-R\boldsymbol{\beta}\|^2=0$ (Meaning $\|\mathbf{r}\|^2=\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2$).
Thus

$$\hat{\boldsymbol{\beta}}_{\text{LSE}}=R^{-1}\mathbf{f}=R^{-1}Q^T\mathbf{y}$$

This requires that R is full rank, in order for its inverse to exist.
Further, $\mathbf{X}$ has to have full rank, which we can ensure by our design of the model.

Proposition 2.5 - *Least Squares Estimate of Parameter Vector is Unbiased*

$$\mathbb{E}(\hat{\boldsymbol{\beta}})=\mathbb{E}(R^{-1}Q^T\mathbf{y})=R^{-1}Q^T\mathbb{E}(\mathbf{y})=R^{-1}Q^T\mathbf{X}\boldsymbol{\beta}=R^{-1}Q^TQR\boldsymbol{\beta}=\boldsymbol{\beta}$$

Proposition 2.6 - *Variance of Least Squares Estimate of Parameter Vector*

$$\begin{aligned}&&\operatorname{Cov}(\mathbf{y}) &= I\sigma^2\\ \implies &&\operatorname{Cov}(\mathbf{f}) &= Q^T\mathbf{y}\\ &&&= Q^TQ\sigma^2\\ &&&= I\sigma^2\\ \implies &\operatorname{Cov}(\hat{\boldsymbol{\beta}}_{\text{LSE}}) &= \operatorname{Cov}(R^{-1}\mathbf{f})\\ &&&= R^{-1}\operatorname{Cov}(\mathbf{f})R^{-T}\\ &&&= R^{-1}I\sigma^2R^{-T}\\ &&&= R^{-1}R^{-T}\sigma^2\end{aligned}$$

Remark 2.3 - *Least Squares Estimation - Geometric Interpretation*
Linear Models state that $\mathbb{E}(\mathbf{y})$ lies on the space spanned by all possible linear combinations of the columns of the *Model Matrix*.
Least Squares Estimation finds the point in the space closest to $\mathbf{y}$.
Thus *Least Sqaures Estimation* amounts to find the orthogonal projection of $\mathbf{y}$ in the linear space spanned by the columns of $\mathbf{X}$.

Definition 2.3 - *Influence Matrix*
The *Influence Matrix*, A, is the orthogonal projection of the response variables onto the linear space spanned by the columns of $\mathbf{X}$.
Since

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

the *Influence Matrix* is

$$A = QQ^T$$

N.B. The *Influence Matrix* is *Idempotent*, $AA = A$.

Proposition 2.7 - *Results in terms of Model Matrix*

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ \Sigma_{\hat{\boldsymbol{\beta}}} &= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \\ A &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \end{aligned}$$

N.B. Substituting $\mathbf{X} = QR$ gets back to previous results.

2.1.2 Checking

Remark 2.4 - *Assumptions*
We assume that each ε_i is independent & has constant variance (we also assume they are normally distributed but this generaly holds due to CLT).
We need a way to check this assumption holds in order for inferences (beyound point estimates) to be sound.

Proposition 2.8 - *Graphical Checks*
Plotting $\hat{\epsilon} = y_i - (\mathbf{X}\hat{\boldsymbol{\beta}})_i$ on a graph tends to indicate whether an assumption has been broken, and if so, how it was broken.

- Systematic patterns in the mean indicate independence assumption is broken.
- Systematic patterns in the variability indicate the constant variance assumption is broken.

2.1.3 Evaluating

Remark 2.5 - *Choice of measure to minimise?*
Was choosing to minimise *Residual Sum of Squares* a good one?
N.B. Choosing $\sum_i |\epsilon_i|$, $\sum_i \epsilon^4$,... could have worked.

Remark 2.6 - *Problem with $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$ as measure*
Suppose our data has a lot of information for estimating β_i but not mch for β_j, should we weight them equally?

Remark 2.7 - *Preferred Estimators*
We require estiamtors to be *Unbiased*, and then we shall choose the estimator with the least variance among those which are *Unbiased*.
N.B. Least variance means smallest covariance matrix (in a way which accounts for weighting individual parameters).

Theorem 2.1 - *Gauss Markov Theorem*
Let $\mathbf{X}, \mathbf{y}$ be some observed data.
Consider a model where $\boldsymbol{\mu} := \mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\Sigma_y^2 = \sigma^2 I$.
Let $\tilde{\theta} := \mathbf{c}^T\mathbf{y}$ be any *Unbiased Linear Estimaor* of $\theta = \mathbf{t}^T\boldsymbol{\beta}$ for some arbitrary vector, $\mathbf{t}$.
Then

$$\text{Var}(\tilde{\theta}) \geq \text{Var}(\hat{\theta})$$

where $\hat{\theta} = \mathbf{t}^T\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}} = R^{-1}Q^T\mathbf{y}$ where $\mathbf{X} = QR$.
Thus each element of $\hat{\boldsymbol{\beta}}$ is a *minimum variance unbiased estimator*, since $\mathbf{t}$ is arbitrary.

2.1.4 Hypothesis Testing & Intervals

Remark 2.8 - *Populat Hypothesis Test*
Often we want to test whether any $\beta_i = 0$ as this would indicate that those predictors do not affect the model accuracy.

Proposition 2.9 - *Distribution of* $\hat{\boldsymbol{\beta}}$
We assume that $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$. Thus

$$\begin{aligned} \mathbf{y} &\sim \text{Normal}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I) \\ \implies \hat{\boldsymbol{\beta}} &\sim \text{Normal}(\boldsymbol{\beta}, R^{-1}R^{-T}\sigma^2) \end{aligned}$$

Note that $\boldsymbol{\beta}$ and σ^2 are unknown.
N.B. $\mathbf{X} = QR$ where Q is orthogonal & R upper-triangle.

Proposition 2.10 - $\dfrac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} \sim t_{n-p}$

Note that we can produce a decomposition $\mathbf{X} = \mathcal{Q}\begin{pmatrix} R \\ \mathbf{0} \end{pmatrix}$ where $\mathcal{Q}$ is orthogonal & R is upper triangular.
We have

$$\text{Cov}(\mathcal{Q}^T\mathbf{y}) = \mathcal{Q}^T\text{Cov}(\mathbf{y})\mathcal{Q}^{-T} = \mathcal{Q}^T\text{Cov}(\mathbf{y})\mathcal{Q} = \mathcal{Q}^T I\sigma^2\mathcal{Q} = I\sigma^2$$

This implies that elements of $\mathcal{Q}^T\mathbf{y}$ are independent, due to their assumed normal distribution.
Note that

$$\mathbb{E}(\mathcal{Q}^T\mathbf{y}) = \mathbb{E}\left(\begin{pmatrix} \mathbf{f} \\ \mathbf{r} \end{pmatrix}\right) \quad \text{and} \quad \mathbb{E}(\mathcal{Q}^T\mathbf{y}) = \mathcal{Q}^T\mathbb{E}(\mathbf{y}) = \mathcal{Q}^T\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} R \\ \mathbf{0} \end{pmatrix}\boldsymbol{\beta}$$

Thus

$$\mathbb{E}\left(\begin{pmatrix} \mathbf{f} \\ \mathbf{r} \end{pmatrix}\right) = \begin{pmatrix} R \\ \mathbf{0} \end{pmatrix}\boldsymbol{\beta} \implies \mathbb{E}(\mathbf{f}) = R\boldsymbol{\beta} \;\&\; \mathbb{E}(\mathbf{r}) = \mathbf{0}$$

Further

$$\mathbf{f} \sim \text{Normal}(R\boldsymbol{\beta}, I_p\sigma^2) \quad \text{and} \quad \mathbf{r} \sim \text{Normal}(\mathbf{0}, I_{n-p}\sigma^2)$$

and $\mathbf{f}$ & $\mathbf{r}$ are independent.
Thus $\hat{\boldsymbol{\beta}}$ & $\hat{\sigma}^2$ are independent.
Since each $r_i \sim \text{Normal}(0, \sigma^2)$

$$\frac{\|\mathbf{r}\|^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n-p} r_i^2 \sim \chi^2_{n-p}$$

Since $\mathbb{E}(\chi^2_{n-p}) = n - p \implies \hat{\sigma}^2 := \frac{1}{n-p}\|\mathbf{r}\|^2$ is an unbiased estimator of σ^2.
$\hat{\Sigma}_{\hat{\boldsymbol{\beta}}} := \Sigma_{\hat{\boldsymbol{\beta}}}\frac{\hat{\sigma}^2}{\sigma^2} = R^{-1}R^{-T}\hat{\sigma}^2$ is an unbiased etimator of $\Sigma_{\hat{\boldsymbol{\beta}}}$.
Thus $\hat{\sigma}_{\hat{\beta}_i} := \sqrt{[\hat{\Sigma}_{\hat{\beta}}]_i} = \sigma_{\hat{\beta}_i}\frac{\hat{\sigma}}{\sigma}$.
Finally

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} = \frac{\hat{\beta}_i - \beta_i}{\sigma_{\hat{\beta}_i}\frac{\hat{\sigma}}{\sigma}} = \frac{\frac{1}{\sigma_{\hat{\beta}_i}}(\hat{\beta}_i - \beta_i)}{\sqrt{\hat{\sigma}^2/\sigma^2}} = \frac{\frac{1}{\sigma_{\hat{\beta}_i}}(\hat{\beta}_i - \beta_i)}{\sqrt{\frac{1}{\sigma^2}\frac{1}{n-p}\|\mathbf{r}\|^2}} \sim \frac{\text{Normal}(0,1)}{\sqrt{\frac{1}{n-p}\chi^2_{n-p}}} \sim t_{n-p}$$

N.B. $\|\mathbf{r}\|^2 = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$ by the results in **Proposition 2.4**.

Proposition 2.11 - *Confidence Interval for β_i*
Using the result in **Proposition 2.9** we can construct the following $1-\alpha$ confidence interval

$$\mathbb{P}\left(-t_{n-p,\frac{\alpha}{2}} < \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} < t_{n-p,\frac{\alpha}{2}}\right) = \mathbb{P}\left(\hat{\beta}_i - t_{n-p,\frac{\alpha}{2}}\sigma_{\hat{\beta}_i} < \beta_i < \hat{\beta}_i + t_{n-p,\frac{\alpha}{2}}\sigma_{\hat{\beta}_i}\right) = 1-\alpha$$

Proposition 2.12 - *Hypothesis Testing on β_i*
Suppose we want to test $H_0 : \beta_i = \beta_{i0}$ against $H_1 : \beta_i \neq \beta_{i0}$.
We use test statistic

$$T = \frac{\hat{\beta}_i - \beta_{i0}}{\hat{\sigma}_{\hat{\beta}_i}}$$

under H_0 $T \sim t_{n-p}$ where n is the number of observations & p the number of parameters.
Thus we can assess the test using $p = \mathbb{P}(|T| \geq |t_{obs}|)$.

Proposition 2.13 - *Testing Multiple Variables in a Model*
This can be expressed as the test of $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ against $H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{d}$ where $\mathbf{C} \in \mathbb{R}^{q\times p}$ & $\mathbf{d} \in \mathbb{R}^q$ with $q < p$.
Under H_0 we have $(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) \sim \text{Normal}(\mathbf{0}, \mathbf{C}\Sigma_{\hat{\boldsymbol{\beta}}}\mathbf{C}^T)$.
We can produce a *Cholesky Decomposition* $\mathbf{L}^T\mathbf{L} = \mathbf{C}\Sigma_{\hat{\beta}}\mathbf{C}^T$.
Thus

$$\begin{aligned}
& & \mathbf{L}^{-T}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) &\sim \text{Normal}(0, I) \\
\implies & (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^T(\mathbf{C}\Sigma_{\hat{\boldsymbol{\beta}}}\mathbf{C}^T)^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) &= (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^T\mathbf{L}^{-1}\mathbf{L}^{-T}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) \\
& &= \|\mathbf{L}^{-T}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})\|^2 \\
& &\sim \sum_{i=1}^{q} \text{Normal}(0,1)^2 \\
& &\sim \chi^2_q
\end{aligned}$$

Setting $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}} := \frac{\hat{\sigma}^2}{\sigma^2}\Sigma_{\hat{\boldsymbol{\beta}}}$ we can produce a test statistic

$$F := \frac{1}{q}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^T(\mathbf{C}\Sigma_{\hat{\boldsymbol{\beta}}}\mathbf{C}^T)^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})$$

Which has the distribution

$$\begin{aligned}
F &= \tfrac{1}{q}\|\mathbf{L}^{-T}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})\|^2 \\
&= \frac{\sigma^2}{q\hat{\sigma}^2}\|\mathbf{L}^{-T}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})\|^2 \\
&= \frac{\frac{1}{q}\|\mathbf{L}^{-T}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})\|^2}{\hat{\sigma}^2/\sigma^2} \\
&= \frac{\frac{1}{q}\|\mathbf{L}^{-T}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})\|^2}{\frac{1}{\sigma}^2\frac{1}{n-p}\|\mathbf{r}\|^2} \\
&\sim \frac{\frac{1}{q}\chi^2_q}{\frac{1}{n-p}\chi^2_{n-p}} \\
&\sim F_{q,n-p}
\end{aligned}$$

Proposition 2.14 - $F = \dfrac{\frac{1}{q}(RSS_0 - RSS_q)}{\frac{1}{n-p}RSS_1}$
Where RSS_0 is the residual sum of squares when H_0 is true and RSS_1 is the residual sum of squares when H_1 is true.

Proposition 2.15 - *Testing whether a Factor Variable belongs in a Model*
Factor Variables have multiple parameters associated to them in a model and thus to test

whether the *Factor Variable* should be in the model requires testing whether all of these parameters should equal 0.
This can be tested using the results in **Propostion 2.12** with $\mathbf{d} = \mathbf{0}$ and $\mathbf{C}$ is the rows of the I_p which indicate the parameters we wish to test.
In this case q is the number of parmeters we wish to test.

2.2 Bayesian Approach

Proposition 2.16 - *Prior Distributions*
We need to define *Prior Distributions* for $\boldsymbol{\beta}$ and σ^2.

$$\boldsymbol{\beta} \sim \text{Normal}(\boldsymbol{\beta}_0, \boldsymbol{\phi}^{-1}) \quad \text{and} \quad \frac{1}{\sigma^2} =: \tau \sim \Gamma(a, b)$$

where $\boldsymbol{\beta}_0, \boldsymbol{\phi}, a, b$ are given by us.
N.B. In order for results to be tractable we use conjugate priors. $\tau := \frac{1}{\sigma^2}$ is called *Precision*.

Proposition 2.17 - *Resulting Distribution*

$$\begin{aligned}
f(\mathbf{y}, \boldsymbol{\beta}, \tau) &\propto \frac{\tau^{\alpha-1+\frac{n}{2}} e^{-\frac{\tau}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2}}{e^{\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\boldsymbol{\phi}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)} e^{-b\tau}} \\
f(\tau|\boldsymbol{\beta}, \mathbf{y}) &\propto \frac{\tau^{\alpha-1\frac{n}{2}}}{e^{\frac{\tau}{2}(b+\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2}} \\
&\sim \Gamma\left(\tfrac{n}{2} + a, b + \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2\right) \\
f(\boldsymbol{\beta}|\tau, \mathbf{y}) &\propto \exp\left\{ -\tfrac{1}{2}(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\tau - 2\boldsymbol{\beta}\mathbf{X}^T\mathbf{y}\tau + \boldsymbol{\beta}^T\boldsymbol{\phi}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{\phi}\boldsymbol{\beta}_0) \right\} \\
&\sim \text{Normal}\left((\mathbf{X}^T\mathbf{X}\tau + \boldsymbol{\phi})^{-1}(\tau\mathbf{X}^T\mathbf{y} + \boldsymbol{\phi}\boldsymbol{\beta}_0),\ (\mathbf{X}^T\mathbf{X}\tau + \boldsymbol{\phi})^{-1}\right)
\end{aligned}$$

If sample size tends to infinity or the prior precision matrix tends to $\mathbf{0}$, then

$$f(\boldsymbol{\beta}|\tau, \mathbf{y}) \xrightarrow{\sim} \text{Normal}(\hat{\boldsymbol{\beta}}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

This is the same as in the *Frequentist Approach*, so the same results acan be used for intervals & hypothesis testing.
N.B. As sample size tends to infinity $\mathbf{X}^T\mathbf{X}\tau$ dominates $\boldsymbol{\phi}$.

Proposition 2.18 - *Find Posterion,* $f(\boldsymbol{\beta}, \tau|\boldsymbol{y})$1

- Iteratively find the posterior modes of $\boldsymbol{\beta}$ (given the estimated mode of τ) and the posterior mode of τ (given the estimated modes of $\boldsymbol{\beta}$), until convergence.
  Then plug this into $f(\boldsymbol{\beta}|\tau, \mathbf{y})$.

Empirical Bayes Integrate $\boldsymbol{\beta}$ outo of $f(\tau|\boldsymbol{\beta}, \mathbf{y})$ to obtain $f(\tau|\mathbf{y})$.
Maximise $f(\tau|\mathbf{y})$ to find $\hat{\tau}$.
Then plug this into $f(\boldsymbol{\beta}|\tau, \mathbf{y})$.

Gibbs Sampling Alternate simulation of $\boldsymbol{\beta}$ from $f(\boldsymbol{\beta}|\tau, \mathbf{y})$ (give simulated τ) with simulation of τ from $f(\tau|\boldsymbol{\beta}, \mathbf{y})$ (given simulated $\boldsymbol{\beta}$)

3 Maximum Likelihood Estimation

3.1 Frequentist

Proposition 3.1 - *Frequentist Approach to Linear Models*
Parameters, $\boldsymbol{\beta}$, are treataed as fixed states of nature and all uncertainty occurs in our estimation

of these parameters.

Definition 3.1 - *Likelihood*
Let $\mathbf{y}$ a set of n observations from $f(\cdot;\boldsymbol{\theta})$.
Likelihood measures the probability of observing specified outcomes, given a possible set of parameter values.

$$L_n(\boldsymbol{\theta};\mathbf{y}) := f_n(\mathbf{y};\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i;\boldsymbol{\theta})$$

Often we use the *Log-Likelihood* function as it turns products into summations and exponents into parameter.

$$\ell_n(\boldsymbol{\theta};\mathbf{y}) := \ln L(\boldsymbol{\theta};\mathbf{y}) = \sum_{i=1}^{n} f(y_i;\boldsymbol{\theta})$$

N.B. $\text{argmax}_\theta L(\theta) \equiv \text{argmax}_\theta \ell(\theta)$.

Definition 3.2 - *Maximum Likelihood Estimate*
Let $\mathbf{y}$ a set of n observations from $f(\cdot;\boldsymbol{\theta})$.
The *Maximum Likelihood Estimate*,$\hat{\boldsymbol{\theta}}_{\text{MLE}}$, for a set of parameter, $\boldsymbol{\theta}$ is the set of parameter values which maximise the *Likelihood Function*.

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \text{argmax}_\theta L(\boldsymbol{\theta};\mathbf{y}) = \text{argmax}_\theta \ell(\boldsymbol{\theta};\mathbf{y})$$

Proposition 3.2 - *Finding Maximum Likelihood Estimate*
Let $\mathbf{y}$ be a set n of observations from the model $f(\mathbf{X};\boldsymbol{\theta})$ with $|\boldsymbol{\theta}| = m$.
To find the *Maximum Likelihood Estimate* of $\boldsymbol{\beta}$

i) Define the *Log-Likelihood Function* $\ell(\boldsymbol{\theta};\mathbf{y})$.

ii) Find the gradient of $\ell(\boldsymbol{\theta};\mathbf{y})$ wrt $\boldsymbol{\theta}$

$$\nabla\ell(\boldsymbol{\theta};\mathbf{y}) := \begin{pmatrix} \frac{\partial}{\partial\theta_1}\ell(\boldsymbol{\theta};\mathbf{y}) & \dots & \frac{\partial}{\partial\theta_m}\ell(\boldsymbol{\theta};\mathbf{y}) \end{pmatrix}$$

iii) Equate the gradient to the zero-vector and solve for $\boldsymbol{\theta}$ to find extrema of ℓ

$$\nabla\ell(\boldsymbol{\theta};\mathbf{y}) = \mathbf{0}$$

iv) Calculate the *Hessian* of $\ell(\boldsymbol{\theta};\mathbf{x})$

$$\nabla^2\ell(\boldsymbol{\beta};\mathbf{y}) = \begin{pmatrix} \frac{\partial}{\partial\theta_1^2}\ell(\boldsymbol{\theta};\mathbf{y}) & \dots & \frac{\partial}{\partial\theta_1\partial\theta_m}\ell(\boldsymbol{\theta};\mathbf{y}) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial\theta_m\partial\theta_1}\ell(\boldsymbol{\theta};\mathbf{y}) & \dots & \frac{\partial}{\partial\theta_m^2}\ell(\boldsymbol{\theta};\mathbf{y}) \end{pmatrix}$$

v) Test each extremum, $\hat{\boldsymbol{\theta}}$ to see if any are maxima

$$\text{If } \det(H(\hat{\boldsymbol{\theta}})) > 0 \text{ and } \tfrac{\partial}{\partial\theta_1^2}\ell(\hat{\boldsymbol{\theta}};\mathbf{y}) < 0 \text{ then } \hat{\boldsymbol{\theta}} \text{ is a local maximum.}$$
$$\textit{i.e.} \text{ If } H(\hat{\boldsymbol{\theta}}) \text{ is negative definite.}$$

Remark 3.1 - *It is rare to find explicit expressions for MLEs. Instead we use numerical optimisation*

3.2 Performance

Remark 3.2 - *Here we look at properties of an MLE when we have a large sample size*

Definition 3.3 - *Fisher Information Matrix*
Let $\ell(\cdot) := \ln f(\mathbf{y};\boldsymbol{\theta}))$ & $\boldsymbol{\theta}^*$ be the true parameter values.
Fisher Information describes how much information the X carries about the parameters, $\boldsymbol{\theta}$.
The *Fisher Information Matrix* is defined as

$$\mathcal{I} := \mathbb{E}\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \frac{\partial \ell}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right)$$

Proposition 3.3 - *Interpretting Fisher Information Matrix*
If $\mathcal{I}$ carries lots of information if it has large magnitude eigenvalues & less information if they have small magnitude.
N.B. See **Proposition 3.5** for a different formulation of *Fisher Information Matrix*.

Remark 3.3 - *Properties of Expected Log-Likelihood*
Here are some properties of the *Expected Log-Likelihood* as *Large Sample Theory* of *MLEs* relies on them.

Theorem 3.1 - *Expect a turning point in Log-Likelihood at the true parameter values*
Let $\ell(\cdot) := \ln f(\mathbf{y};\boldsymbol{\theta}))$ & $\boldsymbol{\theta}^*$ be the true parameter values.

$$\mathbb{E}\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right) = \mathbf{0}$$

Proof

$$\mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\mathbf{y};\boldsymbol{\theta})\right) = \int \frac{1}{f(\mathbf{y};\boldsymbol{\theta})} \frac{\partial f}{\partial \boldsymbol{\theta}} f(\mathbf{y};\boldsymbol{\theta}) d\mathbf{y} = \int \frac{\partial f}{\partial \boldsymbol{\theta}} d\mathbf{y} = \frac{\partial}{\partial \boldsymbol{\theta}} \int f(\mathbf{y};\boldsymbol{\theta}) d\mathbf{y} = \frac{\partial \mathbf{1}}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

□

Theorem 3.2 - *Covariance as Expectation*
Let $\ell(\cdot := \ln f(\mathbf{y};\boldsymbol{\theta}))$ & $\boldsymbol{\theta}^*$ be the true parameter values.

$$\mathrm{Cov}\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right) = \mathbb{E}\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \frac{\partial \ell}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right)$$

Proof
Follows from **Theorem 3.1** and the defiinition of a covariance matrix, noting that $\frac{\partial \ell}{\partial \boldsymbol{\theta}}$ is a column vector and $\frac{\partial \ell}{\partial \boldsymbol{\theta}^T}$ is a row vector.

Proposition 3.4 - *Fisher Information Matrix as negative expectation*
Let $\ell(\cdot) := \ln f(\mathbf{y};\boldsymbol{\theta}))$ & $\boldsymbol{\theta}^*$ be the true parameter values.

$$\mathcal{I} := \mathbb{E}\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \frac{\partial \ell}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right) \equiv -\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right)$$

Proof

$$\begin{aligned}
& \int \frac{\partial \ell}{\partial \boldsymbol{\theta}} f(\mathbf{y};\boldsymbol{\theta}) d\mathbf{y} &&= \mathbf{0} \\
\implies & \int \frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} f(\mathbf{y};\boldsymbol{\theta}) + \frac{\partial \ell}{\partial \boldsymbol{\theta}} \frac{\partial f}{\partial \boldsymbol{\theta}^T} d\mathbf{y} &&= \mathbf{0} \\
\text{but} \quad & \frac{\partial \ell}{\partial \boldsymbol{\theta}^T} &&= \frac{1}{f} \frac{\partial f}{\partial \boldsymbol{\theta}^T} \\
\implies & \int \frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} f(\mathbf{y};\boldsymbol{\theta}) d\mathbf{y} &&= -\int \frac{\partial \ell}{\partial \boldsymbol{\theta}} \frac{\partial \ell}{\partial \boldsymbol{\theta}^T} f(\mathbf{y};\boldsymbol{\theta}) d\mathbf{y}
\end{aligned}$$

□

N.B. $\ell(\theta) = \ln f(\theta) \implies \frac{\partial}{\partial\theta}\ell(\theta) = \frac{\frac{\partial}{\partial\theta}f(\theta)}{f(\theta)}$.

Proposition 3.5 - *Expected Log-Likelihood has Global Maximum at True Parameter Value*
Let $\ell(\cdot) := \ln f(\mathbf{y};\boldsymbol{\theta}))$ & $\boldsymbol{\theta}^*$ be the true parameter values.

$$\forall\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \mathbb{E}[\ell(\boldsymbol{\theta})] \leq \mathbb{E}[\ell(\boldsymbol{\theta}^*)]$$

Proof Since ln is concave we can use *Jensen's Inequality*

$$\mathbb{E}\left[\ln\left(\frac{f(\mathbf{y};\boldsymbol{\theta})}{f(\mathbf{y};\boldsymbol{\theta}^*)}\right)\right] \leq \ln\left[\mathbb{E}\left(\frac{f(\mathbf{y};\boldsymbol{\theta})}{f(\mathbf{y};\boldsymbol{\theta}^*)}\right)\right] = \ln\int\frac{f(\mathbf{y};\boldsymbol{\theta})}{f(\mathbf{y};\boldsymbol{\theta}^*)}f(\mathbf{y};\boldsymbol{\theta}^*)d\mathbf{y} = \ln\int f(\mathbf{y};\boldsymbol{\theta})d\mathbf{y} = \ln 1 = 0$$

□

Theorem 3.3 - *Cramér-Rao Lower Bound*
Let $\mathbf{y} \sim f(\cdot;\boldsymbol{\theta})$ and $\mathcal{I}$ be the *Fisher Information Matrix*.
The *Cramér-Rao Lower Bound* states that

$$\mathcal{I}^{-1} \text{ is a lower bound of the variance matrix of any unbiased estimator } \tilde{\boldsymbol{\theta}}$$

In the sense that $\text{Cov}(\boldsymbol{\theta}) - \mathcal{I}^{-1}$ is positive semi-definite.
N.B. Look at proof.

Proposition 3.6 - *Consistency of MLE*
Maximum Likelihood Estimators <u>are</u> usually *Consistent*.
This is because $\frac{1}{n}\ell(\boldsymbol{\theta}) \overset{n\to\infty}{\longrightarrow} \frac{1}{n}\mathbb{E}(\ell(\boldsymbol{\theta})) \to \boldsymbol{\theta}^*$ by **Proposition 3.5** (and WLLN if log can be broken down into summations).
N.B. Consistency likeliy fails when the number of parameters increases as sample sie increases.

Proposition 3.7 - *MLE Distribution for Large Sample Size*
Let $\hat{\boldsymbol{\theta}}$ be an MLE for a set of parameters, with true value $\boldsymbol{\theta}^*$.
As the same size tends to infty

$$\hat{\boldsymbol{\theta}} \sim \text{Normal}(\boldsymbol{\theta}^*, \mathcal{I}^{-1})$$

This means that in regular situtations, for large sample sizes, MLEs are unbiased and achieve the *Cramér Rao Lower Bound*.

Proof 3.1 - *Proposition 3.7*
If the log-likelihood is based on independent observations then $\ell(\boldsymbol{\theta}) = \sum_i \ell_i(\boldsymbol{\theta})$.
$\implies \frac{\partial\ell}{\partial\boldsymbol{\theta}} = \sum_i \frac{\partial\ell_i}{\partial\boldsymbol{\theta}}$. Thus the central limit theorem applies and we get the result.
If the log-likelihood is <u>not</u> base don independent observations then $\frac{\partial\ell}{\partial\boldsymbol{\theta}}$ usually has a limiting normal distribution so the result holds anyways.

3.3 Intervals

3.4 Numerical Optimisation

Definition 3.4 - *Numerical Optimisation*
Numerical Optimisation is the process of finding the set of parameters which maximise a function by numerically evaluating the function with multiple different values. This is used when we cannot take derivatives of a function for whatever reason.
N.B. A lot of techniques focus on finding the <u>minimum</u> so we use the negative of the *Objective Function* in order to find the maximum instead.

Definition 3.5 - *Objective Function*
The *Objective Function* is the function we wish to optimise in *Numerical Optimisation.*
N.B. In *Statistical Inference* this is the *Probability Density/Mass Function.*

Proposition 3.8 - *Assumptions about Objective Function*
In order to make problems easier we assume that the *Objective Function* is:

i) Sufficiently smooth;

ii) Bounded below; and,

iii) The parameter elements are unrestricted real values.
If we want to put restrictions on $\boldsymbol{\theta}$ we need to be able to implement them as $\boldsymbol{\theta} = \mathbf{r}(\boldsymbol{\theta}_r)$ where $\mathbf{r}(\cdot)$ is a known function & $\boldsymbol{\theta}_r$ is the unrestricted parameter set.

N.B. We require f to be convex for *Newton Methods* to work but that is an assumption too far. If it is not then we may only find a local minimum, not global.

Proposition 3.9 - *Principles for Newton's Method*
EXAMINABLE.
Let $f(\cdot)$ be the function we wish to optimise.
Newton's Method for *Numerical Optimisation* is based on iteratively approximating f by a truncated Taylor expansion and seeking the minimum of the approxiamte at each step.

$$f(\boldsymbol{\theta}+\boldsymbol{\Delta}) = f(\boldsymbol{\theta}) + \nabla f(\boldsymbol{\theta})^T\boldsymbol{\Delta} + \frac{1}{2}\boldsymbol{\Delta}^T\nabla^2 f(\boldsymbol{\theta}+t\boldsymbol{\Delta})\boldsymbol{\Delta} \text{ for } t \in (0,1)$$

Thus

$$f(\boldsymbol{\theta}+\boldsymbol{\Delta}) \geq f(\boldsymbol{\theta}) \text{ for small } \boldsymbol{\Delta} \text{ is equivalent to } \nabla f(\boldsymbol{\theta}) = \mathbf{0} \text{ and } \nabla^2 f(\boldsymbol{\theta}) \text{ being positive definite.}$$
$$\textit{i.e. } \boldsymbol{\theta} \text{ is a minimum of } f.$$

We have that $\boldsymbol{\Delta} := -\mathbf{H}\nabla f(\boldsymbol{\theta})$ is a *Descent Direction* if $\mathbf{H}$ is a positive definite matrix.

Definition 3.6 - *Newton's Method for Numerical Optimisation*

i) Make an initial parameter value guess.

ii) Obtain a quadratic approximation to the *Log-Likelihood Function* which behaves similarlly to the log-likelihood function in the region of this guess.
N.B. This is done by using a *Taylor Approximation* of the *Log-Likelihood Function.*

$$f(\boldsymbol{\theta}+\boldsymbol{\Delta}) = f(\boldsymbol{\theta}) + \nabla f(\boldsymbol{\theta})^T\boldsymbol{\Delta} + \frac{1}{2}\boldsymbol{\Delta}^T\nabla^2 f(\boldsymbol{\theta}+t\boldsymbol{\Delta})\boldsymbol{\Delta}$$

Solve for $\boldsymbol{\Delta}$ which maximises.
N.B. Take derivate wrt $\boldsymbol{\Delta}$ and equate to zero

iii) Update parameter guesss to be maximiser of this quadratic approxiation.

iv) Repeat *ii)-iii)* with new guesses, until convergence.

Proposition 3.10 - *Requirements for Newton's Method*
In order to guarantee that *Newton's Method* converges to the MLE, we need to ensure the following

i) The approximating quadratic actually has a maximum (not minimum, inflection point etc.).
N.B. If it has a minimum then we use its negative value instead.

ii) The proposed change in parameter values actually increases the *Log-Likelihood* itself. If not we move the parameter back towards the previous parameter guess until the *Log-Likelihood* increases.

Definition 3.7 - *Newton's Method*
NON-EXAMINABLE.
Let $f(\boldsymbol{\theta})$ be the function we wish to optimise (this would be the *Likelihood Function*).
Newton's Method is a method for *Numerical Optimisation.*
The idea is to iteratively use a truncated *Taylor Expansion* (to the second degree) of function $f(\boldsymbol{\theta})$ and to find the minimum of this approximation at each step.

i) Make an initial input guess $\boldsymbol{\theta}^{[0]}$. Set $k = 0$.

ii) Evaluate the function & its first two derivatives

$$f(\boldsymbol{\theta}^{[k]}),\ \nabla f(\boldsymbol{\theta}^{[k]}),\ \nabla^2 f(\boldsymbol{\theta}^{[k]})$$

iii) **If $\nabla f(\boldsymbol{\theta}^{[k]}) = \mathbf{0}$ and $\nabla^2 f(\boldsymbol{\theta}^{[k]})$ is positive semi-definite:**

- $\boldsymbol{\theta}^{[k]}$ is a minimum. TERMINATE

iv) **If $\nabla^2 f(\boldsymbol{\theta}^{[k]})$ is *positive-definite*:** Set $\mathbf{H} = \nabla^2 f(\boldsymbol{\theta}^{[k]})$.
Else: Set $\mathbf{H} = U\tilde{\Lambda}U^T$ where we have decomposed $\nabla^2 f(\boldsymbol{\theta}^{[k]}) = U\Lambda U^T$ with Λ being the diagonal matrix of eigenvalues & $\tilde{\Lambda}_{ij} = |\Lambda_{ij}|$ (all values positive).
N.B. This step is to ensure that $\mathbf{H}$ is *postive definite.*

v) Solve $\Delta := -\dfrac{\nabla f(\boldsymbol{\theta}^{[k]})}{\mathbf{H}}$ where Δ is the search direction.

vi) **If not $f(\boldsymbol{\theta}^{[k]} + \Delta) < f(\boldsymbol{\theta}^{[k]})$:** repeatedly have Δ until condition is met.
N.B. Implementation of *Step Length Control.*

vii) Set $\boldsymbol{\theta}^{[k+1]} = \boldsymbol{\theta}^{[k]} + \Delta$. Set $k+ = 1$

viii) Repeat ii)-vii) until we get a termination in stage iii).

N.B. In practice we terminate in iii) if $\|\nabla f(\boldsymbol{\theta}^{[k]})\| < |\nabla f(\boldsymbol{\theta}^{[k]})|\epsilon_r + \epsilon_a$, for some small $\epsilon_a, epsilon_r$ which we set.

Remark 3.4 - *The first derivative tells us the direction, the second derivative suggests the step length*

Remark 3.5 - *$f(\boldsymbol{\theta}^{[k]})$ is only evaluated to ensure that step is an improvement.*
NON-EXAMINABLE.
If $f(\cdot)$ is not avaiable (but ∇f & $\nabla^2 f$ are) we have a few options

- Show that f is non-increasing in the direction of the step, Δ, at $\boldsymbol{\theta}+\Delta$ (*i.e.* $\nabla f(\boldsymbol{\theta}+\Delta)^T\Delta \leq 0$).

Or Replace $-\nabla^2\ell(\boldsymbol{\theta})$ with $-\mathbb{E}[\nabla^2\ell(\boldsymbol{\theta})]$ (Known as the *Fisher Scoring Matrix*).

N.B. This only affects step vi) of **Definition 3.5**.

Remark 3.6 - *Other Numerical Optimisation Techniques*

Quasi-Newton *Quasi-Newton Methods* are *Newton type Methods* in which an approximation is made for the *Hessian Matrix* ($\nabla^2 f(\cdot)$), or its inverse, by building it from the first derivative information computed at each trial.
N.B. In R this is done with `optim(...,method =' BFGS')`.

Steepest Descent Truncating the *Taylor Expansion* allows us to establish the direction to step but not length. Thus we need to implement step length control methods.

N.B. There plenty of methods not covered here.

4 Estimating Posterior Distribution

Proposition 4.1 - *Difficulty Computing Bayes' Theorem*
The *Likelihood* & *Prior Distributions* are computable in *Bayes' Theorem*, but the *Evidence Distribution*, $p(\mathbf{y})$ is not. Thus the difficult in calculating the *Posterior Distribution* lies in trying to calculate the *Evidence Distribution* for the observed data.

Remark 4.1 - *Likelihood is the probability of observing the data that we observed*

Definition 4.1 - *Markov Chain Monte Carlo Methods*
MCMC Methods are techniques for generating a sample from a probability distribution.
Here we use them to generate a sample from the posterior of a model.
They are very general & will work with almost any model, given enough iterations.

Remark 4.2 - *Convergence in MCMC Methods*
We can run MCMC Methods for as many iterations as we like, thus we need to measure whether convergence has occured so we can stop sampling at the correct time.

- This can be done by inspect plots of data & looking at the rolling values of quantiles.
  N.B. If we think the posterior is multi-modal then multiple MCMC chains should be run at once.

- Examine the *Autocorrelation Function* of the chain components.
  This considers reading the chain at different intervals & what correlations occur in doing so. Lower values indicate greater independence.

- *Autocorrelation Length* is twice the sum of the correlations (at different interval lengths) minus 1.
  We typically sum up values until the correlations appear to be 0.
  From this we can calculate the effecitve sample size (*i.e.* How many 'independent' values we have sampled from all the iterations of the chain).

$$\text{ESS} := \frac{\#\text{ chain iterations}}{\text{Autocorrelation length}}$$

From these can perform better tests sub as two-sample Kologorov-Smirnov test, or ANOVA methods for multiple chains.

Remark 4.3 - *Interval Estimation from MCMC methods*
Given a sufficiently large sample from an *MCMC Method* we can compute intervals by finding appropriate quantiles of the sample.

4.1 Markov Chains

Definition 4.2 - *Markov Chains*
A *Markov Chain* is a sequence of random vectors $\mathbf{X}_1, \mathbf{X}_2, \dots$ which satisfies the condition

$$\forall\ j,\ f(\mathbf{x}_j|\mathbf{x}_{j-1}, \dots, \mathbf{x}_1) = f(\mathbf{x}_j|\mathbf{x}_{j-1})$$

i.e. The probability of a realisation depends on the previous realisation only & non-earlier.

Definition 4.3 - *Transition Kernel*
The *Transition Kernel* of the *Markov Chain* is the probability of transition to a given state, given the current state.

$$\mathbb{P}(\mathbf{x}_j|\mathbf{x}_{j-1})$$

Definition 4.4 - *Stationary Distribution*
A *Stationary Distribution* of a *Markov Chain* is a distribution f_x which satisfies

$$f_x(\mathbf{x}_j) = \int \mathbb{P}(\mathbf{x}_j|\mathbf{x}_{j-1}) f_x(\mathbf{x}_{j-1}) d\mathbf{x}_{j-1}$$

N.B. The existence of a *Stationary Distribution* is not guaranteed & depends on $\mathbb{P}$ being irriducible (*i.e.* whenever we start the chain, there is a positive probability of visiting all possible values of $\mathbf{X}$).

Definition 4.5 - *Recurrent*
A *Markov Chain* is *Recurrent* if we can start at any value of $\mathbf{X}$ and its marginal distribution, ϕ, will eventually converge to the *Stationary Distribution*, f_x.
For a simulation of length $J \to \infty$

$$\frac{1}{J}\sum_{j=1}^{J} \phi(\mathbf{x}_j) \to \mathbb{E}_{f_x}(\phi(\mathbf{X}))$$

N.B. This is know as *Ergodicity*.

Remark 4.4 - *Reccurence is an extension of the WLLN to the correlated sequence & is why MCMCs are useful.*

Definition 4.6 - *Detailed Balanced Condition*
Let $\mathbb{P}(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j)$ be the pdf of $\boldsymbol{\theta}_i$ given $\boldsymbol{\theta}_j$ according to the *Markov Chain*.
The *Detailed Balanced Condition* is that

$$\mathbb{P}(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{j-1}) f(\boldsymbol{\theta}_{j-1}|\mathbf{y}) = \mathbb{P}(\boldsymbol{\theta}_{j-1}|\boldsymbol{\theta}_j) f(\boldsymbol{\theta}_j|\mathbf{y})$$

N.B. Also know as *Reversibility*.

Proposition 4.2 - *Using Detailed Balanced Condition*
Notice that the LHS of the *Detailed Balanced Condition* is the joint pdf of $\boldsymbol{\theta}_j$ & $\boldsymbol{\theta}_{j-1}$.
Integrating both sides wrt $\boldsymbol{\theta}_{j-1}$ gives

$$\begin{aligned}\int \mathbb{P}(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{j-1} f(\boldsymbol{\theta}_{j-1}|\mathbf{y}) d\boldsymbol{\theta}_{j-1} &= \int \mathbb{P}(\boldsymbol{\theta}_{j-1}|\boldsymbol{\theta}_j) f(\boldsymbol{\theta}_j|\mathbf{y}) d\boldsymbol{\theta}_{j-1} \\ &= f(\boldsymbol{\theta}_j|\mathbf{y})\end{aligned}$$

This means that if we start at a value, $\boldsymbol{\theta}_1$, which is not impossible (*i.e.* $f(\boldsymbol{\theta}_1|\mathbf{y}) > 0$), then the chain will generate from the target distribution.
N.B. The speed of convergence to a high-probability region of $f(\boldsymbol{\theta}|\mathbf{y})$ is a different problem.

4.2 Metropolis Hastings

Proposition 4.3 - *Metropolis Hastings Method*
The *Metropolis Hastings Method* constructs a chain with an appropriate P.

i) Propose a distribution $q(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{j-1})$ (*e.g.* Normal distribution centred around $\boldsymbol{\theta}_{j-1}$).

ii) Pick a value $\boldsymbol{\theta}_0$ & set $j = 1$.

iii) Generate $\boldsymbol{\theta}'_j$ from $q(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{j-1})$.

iv) Set $\boldsymbol{\theta}_j = \boldsymbol{\theta}'_j$ with probability

$$\alpha = \min\left\{1, \quad \frac{f(\mathbf{y}|\boldsymbol{\theta}'_j)f(\boldsymbol{\theta}'_j)q(\boldsymbol{\theta}_{j-1}|\boldsymbol{\theta}'_j)}{f(\mathbf{y}|\boldsymbol{\theta}_{j-1})f(\boldsymbol{\theta}_{j-1})q(\boldsymbol{\theta}'_j)\boldsymbol{\theta}_{j-1})}\right\}$$

v) Increment j.

vi) Repeat iii)-v) until convergence.

Remark 4.5 - *Metropolis Hastings Method - α*
Note that the q terms cancel iff q only depends on the magnitude of $(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j-1})$.
This occurs if q is a normal distribution centred on $\boldsymbol{\theta}_{j-1}$.
The priors, $f(\boldsymbol{\theta}'_j)$ & $f(\boldsymbol{\theta}_{j-1})$, cancel out if they are impropert uniform.
If both q & prior terms cancel out then

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) = \min\left\{1, \quad \frac{L(\boldsymbol{\theta}'_j)}{L(\boldsymbol{\theta}_{j-1})}\right\}$$

Remark 4.6 - *Metroplis Hastings Method - Choosing Initial Guess, $\boldsymbol{\theta}_0$*
$\boldsymbol{\theta}_0$ may be highly imporbably, meaning the chian will require many iteraions to reach a region of high-probability in $f(\boldsymbol{\theta}|\mathbf{y})$.
Usually we discard the *burn-in period* (*i.e.* first few hundred $\boldsymbol{\theta}_j$ vectors simulated where little progress is made).

Remark 4.7 - *Metropolis Hastings Method - Choosing Proposed Distribution, q*
Generally we perform several pilor runs of the *Metropolis-Hastings Sampler* in order to 'tune' the proposal distribution.

i) It is often necessary to update parameters in blocks.

ii) The perfect proposal is the posterior itself (so will never be met).

Proof 4.1 - *Metropolis Hastings Method*
Here it is shown that *Metropolis-Hastings Method* works if it satisfies *Detailed Balanced Condition*.
For notation, let $\pi(\boldsymbol{\theta}) := f(\boldsymbol{\theta})$, remembering that $f(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})$.
This means the acceptance probability from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ is

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}$$

We need to show that $\pi(\boldsymbol{\theta})P(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}')P(\boldsymbol{\theta}|\boldsymbol{\theta}')$.
This is trivial if $\boldsymbol{\theta}' = \boldsymbol{\theta}$.
Otherwise:
We know that $p(\boldsymbol{\theta}'|\boldsymbol{\theta}) = q(\boldsymbol{\theta}'|\boldsymbol{\theta})\alpha(\boldsymbol{\theta}', \boldsymbol{\theta})$.
Thus

$$\begin{aligned}
\pi(\boldsymbol{\theta})P(\boldsymbol{\theta}'|\boldsymbol{\theta}) &= \pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta}) \min\left\{1, \quad \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\} \\
&= \min\{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta}), \pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')\} \\
&= \pi(\boldsymbol{\theta}')P(\boldsymbol{\theta}|\boldsymbol{\theta}') \text{ by symmetry}
\end{aligned}$$

□

4.3 Gibbs Sampling

Definition 4.7 - *Gibbs Sampling*
Suppose the parameter row vector is partitioned into subvectors st $\boldsymbol{\theta} \equiv (\boldsymbol{\theta}^{[1]}, \dots, \boldsymbol{\theta}^{[K]})$.
Further, define

$$\tilde{\boldsymbol{\theta}}_j^{[-k]} := (\boldsymbol{\theta}_{j+1}^{[1]}, \dots, \boldsymbol{\theta}_{j+1}^{[k-1]}, \boldsymbol{\theta}_j^{[k+1]}, \dots, \boldsymbol{\theta}_j^{[K]})$$

Let $\boldsymbol{\theta}_1$ be an initial guess.
We perform J steps of the *Gibbs Sampler Process* as follows

i) For $j \in [1, J]$, $k \in [1, K]$:

(a) Simulate $\boldsymbol{\theta}_{j+1}^{[}k] \sim f(\boldsymbol{\theta}^{[k]}|\tilde{\boldsymbol{\theta}}_j^{[-k]}, \mathbf{y})$.

N.B. We are conditioning on the most recently simulated values due to definition of $\tilde{\boldsymbol{\theta}}_j^{[-k]}$.

Remark 4.8 - *Gibbs Sampling - Finding Conditional Distributions*

- Generally we will be able to recognise the conditions as some standar distribution.
  This often relies on noting that any multiplicative factors of a pdf which do not involve the argument of the pdf are part of the normalising constant & thus recognise the pdf only requires recognising the form to within a normalising constant.
- If not, we can devise some way of simulating them.
- As a last resort, *Metropolis Hastings* can be used for this step.

Remark 4.9 - *Gibbs Sampling - Limitations*

- *Gibbs Sampling* produces a slow moving chain if parameters have high posterior correlation, as sampling from these conditionals produces very small steps.
  Updating the parameters in blocks or reparameterising to reduce posterior dependence can help to improve mixing.
- If improper priors are used with *Gibbs Sampling*, then it is important to check that the posterior is actually proper .
  It is not always possible to detect impropriety for the output of the sampler.

4.4 Automatic Gibbs Sampling

Proposition 4.4 - *Gibbs Sampling uses Graphical Models (**Section 1.3**)*
Let x_i represent the variable associated with the i^{th} node in a given *Graphical Model*.
Gibbs Sampling requires simulating from the full conditionals of all unobserved *Stochastic Nodes*.
From the *Graphical Model* we can infer a distribution

$$\begin{aligned}
f(x_j|\mathbf{x}^{[-j]}) &= \frac{f(\mathbf{x})}{\int f(\mathbf{x})dx_j} \\
&= \frac{\prod_i f(x_i|\text{parents}(x_i))}{\int \prod_i f(x_i|\text{parents}(x_i))dx_j} \text{ by } \textbf{Proposition 1.2} \\
&= \frac{f(x_j|\text{parents}(x_j)) \prod_{i\in\text{children}(j)} f(x_i|\text{parents}(x_i))}{\int f(x_j|\text{parents}(x_j)) \prod_{i\in\text{children}(j)} f(x_i|\text{parents}(x_i))dx_j} \\
&\propto f(x_j|\text{parents}(x_j)) \prod_{i\in\text{children}(j)} f(x_i|\text{parent}(x_i))
\end{aligned}$$

This shows that, no matter how complicated the *Graphical Model* is, the conditional required for the *Gibbs Update* of x_j, $f(x_j|\mathbf{x}^{[-j]})$ depends only on the parent conditional densities of x_j and its children.

This is a relatively small number of terms.

Proposition 4.5 - *Conjugate Distributions*
The RHS of the final result in **Proposition 4.4** has the same structure as a *Prior* for x_j, $f(x_j|\text{parents}(x_j))$, multiplied by a *Likelihood* term for x_j.
This means we can use *conjugacy of distributions* to ascertain the density of $f(x_k|\mathbf{x}^{[-j]})$.

Proposition 4.6 - *Slice Sampling*
Slice Sampling is used when there is no convenient form of the conditional, $f(x_j|\mathbf{x}^{[-j]})$.
Slice Sampling is done as follows

i) Chose some finite $k > 0$.

ii) Sample y uniformly from $[0, kf(x)]$.

iii) Return an x which satisfies $kf(x) \geq y$.

Here we are sampling from two distributions

$$f(y|x) \sim U(0, kf(x)) \quad \text{and} \quad f(x|y) \sim U(x \in \{x : kf(x) \geq y\})$$

To identify the set $\{x : kf(x) \geq y\}$ in *Uni-Modal Distributions* requires finding a single interval, but for *Multi-Modal Distributions* we need to find serveral.

4.5 Estimating Posterior Point-Value

Definition 4.8 - *Posterior Mode*
The *Posterior Mode* is the set of parameter values with the greatest likelihood, according to the *Posterior Distribution*.

$$\hat{\boldsymbol{\theta}} := \text{argmax}_\theta f(\boldsymbol{\theta}|\mathbf{y})$$

N.B. This is the parameters which are most consistent with the data.

Proposition 4.7 - *Improper Uniform Prior*
If we specify an improper uniform prior, $f(\boldsymbol{\theta}) = k$, then $f(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})$ and the *Posterior Modes* are equivalent to the *Maximum Likelihood Estimates*.

Proposition 4.8 - *Large Sample Size*
For a large sample size the *Posterior Mode* is the *Maximum Likelihood Estimate*

$$\boldsymbol{\theta}|\mathbf{y} \sim \text{Normal}(\hat{\boldsymbol{\theta}}_{\text{MLE}}, \mathcal{I}^{-1})$$

5 Bayesian Inference

Definition 5.1 - *Loss Function*
A *Loss Function* quantifies the loss associated with a particular set of parameter predictions, $\hat{\boldsymbol{\theta}}$. We wish to find the set of parameters which minimise a *Loss Function*.

Definition 5.2 - *Marginal Likelihood*
The *Marginal Likelihood* of a model, M, is the marginal density of observing the outcomes

$$f(\mathbf{y}|M) = \int f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

where $\boldsymbol{\theta}$ are the parameters of M.
N.B. This is hard to calculate.

Remark 5.1 - *The value of the Marginal Likelihood is sensitive to the chosen Prior*

5.1 Bayes Factors

Definition 5.3 - *Bayes Factor*
Let M_0 & M_1 be two models and $\mathbf{y}$ some observed data.
Bayes Factor summarises the evidence for/against two alernative models.

$$B_{10} := \frac{f(\mathbf{y}|M_1)}{f(\mathbf{y}|M_0)} \equiv \frac{\mathbb{P}(M_1|\mathbf{y})\mathbb{P}(M_0)}{\mathbb{P}(M_0|\mathbf{y})\mathbb{P}(M_1)}$$

We interpret semantic meaning of *Bayes Factor* using the value of $2\ln B_{10}$

$2\ln B_{10}$	Evidence against B_{10}
$0-2$	Barely worth mentioning
$2-6$	Positive
$6-10$	Strong
>10	Very Strong

N.B. This is similar to the *p-Value* in the *Frequentist Approach.*

Remark 5.2 - *Bayes Factor v Likelihood Ratio Statistic*
Bayes Factor uses *Marginal Distributions* which requires integrating over all possible parameter values. This is the major difference between *Bayes Factor* & the *Likelihood Ratio Statistic.*

Remark 5.3 - *Due to sensitivity of Marginal Likelihood on the Prior, we usually cannot justify using Bayes Factor*
We can only justify it when the priors that differ between alternative models are really precise & are meaningful representation of prior knowledge.
Even in this case it is hard to compute the marginal likelihood.

Proposition 5.1 - *Partial Bayes Factor*
Let $\mathbf{y}$ be some observered data & $(\mathbf{x}, \mathbf{z})$ a partitioning of $\mathbf{y}$.
Using $f(\boldsymbol{\theta}|\mathbf{x})$ as the prior gives *Marginal Likelihood* of $\mathbf{z}$ under model M

$$f(\mathbf{z}|M, \mathbf{x}) = \int f(\mathbf{z}|\boldsymbol{\theta}, \mathbf{x}) f(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}$$

We know that $f(\boldsymbol{\theta}|\mathbf{x}) = \dfrac{f(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathbf{x})} = \dfrac{f(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int f(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}}$.
Thus we get the *Partial Bayes Fator*

$$\begin{aligned} f(\mathbf{z}|M, \mathbf{x}) &= \frac{\int f(\mathbf{z}|\boldsymbol{\theta}, \mathbf{x}) f(\mathbf{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int f(\mathbf{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}} \text{ by substitution} \\ &= \frac{\int f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int f(\mathbf{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}} \end{aligned}$$

Since the same prior is used on the top & bottom, the sensitivity to the prior is reduced.

Definition 5.4 - *Intrinsic Bayes Factor*
An *Intrinsic Bayes Factor* is an extension of *Partial Bayes Factors.*

When partitioning the observed data, $\mathbf{y}$, into $(\mathbf{x}, \mathbf{z})$ we ensure $\mathbf{x}$ is just large enough that $f(\boldsymbol{\theta}|\mathbf{x})$ is proper and then average the resulting *Partial Bayes Factors* over all such $\mathbf{x}$.
This removes the arbitrariness of any particular choice of $\mathbf{x}$.

Definition 5.5 - *Fractional Bayes Factor*
There is a result that $f(\mathbf{x}|\boldsymbol{\theta}) \approx f(\mathbf{y}|\boldsymbol{\theta})^b$ where $b := \dfrac{\dim(\mathbf{x})}{\dim(\mathbf{y})}$.
A *Fractional Bayes Factor* combines this result with the result of *Partial Bayes Factor* to get

$$f(\mathbf{z}|M, \mathbf{x}) \approx \frac{\int f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int f(\mathbf{y}|\boldsymbol{\theta})^b f(\boldsymbol{\theta}) d\boldsymbol{\theta}} \text{ where } b := \frac{\dim(\mathbf{x})}{\dim(\mathbf{y})}$$

For this to select the right model in large sample size, then $b \to 0$ as $n \to \infty$.

5.2 Information Criterion

Proposition 5.2 - *Information Criterion*
Information Criterion are an older method for comparing two models & they avoid the difficulties around the sensitivity wrt priors.

Definition 5.6 - *Bayesian Information Criterion*
Let $\mathbf{y}$ be observed data of dimension n & $\boldsymbol{\theta}$ be the model parameters with dimension p.
The *Bayesian Information Criterion* of a model is defined as

$$BIC := -2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}}_{MLE}) + p \ln n$$

Proposition 5.3 - *Comapring BIC*
The difference in *Bayesian Information Criterion* for two models is an approximation of twice the *log Bayes Factor*.
We should select the model with the lower *BIC*.

Proposition 5.4 - *Derivining BIC*
Let $\mathbf{y}$ be observed data of dimension n & $\boldsymbol{\theta}$ be the model parameters with dimension p.
Define P to be the marginal likelihood of observing $\mathbf{y}$.

$$P := \int f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

Define $\tilde{\boldsymbol{\theta}} := \text{argmax}_\theta f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta})$.
By a *Taylor Expansion* we have

$$\begin{aligned} \ln[f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})] &\simeq \ln[f(\mathbf{y}|\tilde{\boldsymbol{\theta}})f(\tilde{\boldsymbol{\theta}})] - \tfrac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T \left(-\frac{\partial^2 \ln[f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right) (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \\ \implies \quad f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) &\simeq f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) \exp\left\{ -\tfrac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T \left(-\frac{\partial^2 \ln[f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right) (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \right\} \end{aligned}$$

The term in the curly braces is from a multivaraite normal pdf. Thus

$$P \simeq f(\mathbf{y}|\tilde{\boldsymbol{\theta}}) f(\tilde{\boldsymbol{\theta}}) (2\pi)^{p/2} \left| -\frac{\partial^2 \ln[f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \partial \boldsymbol{\theta}^T \right|^{-1/2}$$

As $n \to \infty$ we have $-\dfrac{\partial^2 \ln[f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \partial \boldsymbol{\theta}^T = n\mathcal{I}_0$ where $\mathcal{I}_0$ is the (fixed) information matrix for a single obseration. Thus

$$\left| -\frac{\partial^2 \ln[f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \partial \boldsymbol{\theta}^T \right| = n^p |\mathcal{I}_0|$$

Thus

$$\ln P \simeq \ln f(\mathbf{y}|\tilde{\boldsymbol{\theta}}) + \ln f(\tilde{\boldsymbol{\theta}}) + \frac{p}{2}\ln 2\pi - \frac{p}{2}\ln n - \frac{1}{2}\ln|\mathcal{I}_0|$$

As $n \to \infty$, $\tilde{\boldsymbol{\theta}} \to \hat{\boldsymbol{\theta}}_{MLE}$ while ther terms that do not depend on n become negligible.
Thus we arrive at the *Bayesian Information Criterion* by removing those terms & substituting this result.

Remark 5.4 - *In Complex Bayesian Models it is not always clear how to count the number of free parameters in the model.*
Since priors restrict the freedom of a parameter to vary.

Definition 5.7 - *Deviance*
Deviance is a measure of the number of 'effective' degrees of freedom in a *Bayesian Model*.

$$D(\boldsymbol{\theta}) := -2\ln f(\mathbf{y}|\boldsymbol{\theta}) + c$$

where c is a neglectable constant depending only on the $\mathbf{y}$.

Remark 5.5 - *In the large sample domain* $D(\boldsymbol{\theta}) - D(\mathbb{E}(\boldsymbol{\theta})) \sim \chi^2_r$

Definition 5.8 - *Effective Degrees of Freedom*
Effective Degrees of Freedom is a measure of how many parameters are free in a *Bayesian Model*.

$$p_D := \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$$

Remark 5.6 -
Effective Degrees of Freedom, p_D, is a direct estimate of $\mathbb{E}[D(\boldsymbol{\theta}) - D(\mathbb{E}(\boldsymbol{\theta}))]$.
Noting that $\mathbb{E}(\chi^2_r) =$.

Definition 5.9 - *Deviance Information Criterion*
The *Deviance Information Criterion* is defined as

$$DIC := D(\bar{\boldsymbol{\theta}}) + 2p_D$$

N.B. This is similar to the *AIC*.

0 Appendix

0.1 Definitions

Definition 0.1 - *Proper*

Definition 0.2 - *Parametric Models*
Parameteric Models are *Statistical Models* whose only unkowns are parameters.

Definition 0.3 - *Semi-Parametric Models*
Parameteric Models are *Statistical Models* which contain unknown parameters and unknown functions.

Definition 0.4 - *Non-Parametric Models*
Non-Parametric Models make *few* prior assumptions about how data was generated and instead depend mainly on the observed data.
We cannot simulate data from *Non-Parameteric Models*.

Definition 0.5 - *Orthongonal Matrix*
A matrix $\mathbf{X}$ is *Orthogonal* if

$$\mathbf{X}^T\mathbf{X} = \mathbf{X}\mathbf{X}^T = I \implies \mathbf{X}^T = \mathbf{X}^{-1}$$

Orthogonal Matrices rotate & reflect vectors without changing their magnitude.
N.B. $\mathbf{X}^T$ is *Orthogonal*.

Definition 0.6 - *Full Rank Matrix*
Let $\mathbf{X} \in \mathbb{R}^{m\times n}$.
If $m > n$ then $\mathbf{X}$ has *Full Rank* iff all its columns are linearly independent.
If $n > m$ then $\mathbf{X}$ has *Full Rank* iff all its rows are linearly independent.
N.B. In statistics the number of $m > n$ always as we should have more observations than fields.

Definition 0.7 - *Upper Triangle Matrix*
A matrix X is an *Upper Triangle Matrix* if $X_{i,j} = 0$ for $i > j$.

Definition 0.8 - *Unbiased Estimator*
An *Estimator* of a parameter, $\hat{\theta}$, is unbiased if its expected value is the true value of the parameter for all possible parameter values

$$\mathbb{E}(\hat{\theta}; \theta = \theta^*) = \theta^*$$

Definition 0.9 - *Conjugacy*
Definition 0.10 - *Fisher Information*
Definition 0.11 - *Correlation*
Definition 0.12 - *Covariance*
Definition 0.13 - *Expected Value*
Definition 0.14 - *Variance*

Definition 0.15 - *Positive Definite Matrix*
A matrix is *Positive Definite* if it is symmetric and all its eigenvalues are positive.

Definition 0.16 - *Positive Semi-Definite Matrix*
A matrix is *Postitive Semi-Definite* if all its eigenvalues are non-negative.

Definition 0.17 - *Taylor's Theorem*
Definition 0.18 - *Casual Inference*

Definition 0.19 - *Consistent Estimator*
A *Parameter Estimator*, $\hat{\boldsymbol{\theta}}_n$, is *Consistent* if its value tends to the true parameter value, $\boldsymbol{\theta}^*$, as sample size tends to infinity

$$\hat{\boldsymbol{\theta}}_n \overset{n\to\infty}{\longrightarrow} \boldsymbol{\theta}^*$$

0.2 Theorems

Theorem 0.1 - *Bayes' Theorem*
Suppose $X \sim f(\cdot;\Theta)$. Then

$$\underbrace{\mathbb{P}(\Theta|X)}_{\text{Posterior}} = \frac{\overbrace{\mathbb{P}(X|\Theta)}^{\text{Likelihood}}\overbrace{\mathbb{P}(\Theta)}^{\text{Prior}}}{\underbrace{\mathbb{P}(X)}_{\text{Evidence}}}$$

Theorem 0.2 - *Euclidean Distance Identities*

$$\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} = \sum_{i=1}^{n} x_i^2$$

Theorem 0.3 -
If $\mathbf{X}$ and $\mathbf{Y}$ are independent. Then

$$\mathbf{XY} \simeq \mathbf{0}$$

Theorem 0.4 - *Jensen's Inequality*
For any random variable X and concave function f

$$f[\mathbb{E}(X)] \geq \mathbb{E}[f(X)]$$

0.3 Remarks

Remark 0.1 - *Conditional from Joint*
If we are given a joint distribution $f(x, y)$ and factor out its normalising constant, we can read off the conditional distribution of a variable ($f(x|y)$ or $f(y|x)$) by just using the terms which include the variable.
The distribution should then be recognisable & the normalising constant findable.