Theory of Inference - Reviewed Notes

Dom Hutchinson

March 11, 2020

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

Definition 1.7 - *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 1.8 - *Response Variables*
Response Variables are the independent variables of a system, whose value we observe.
N.B. Typically denoted y or $\mathbf{y}$.

Definition 1.9 - *Fitted Values, $\hat{y}$*
Fitted Values are our estimated values for the *Response Variable*.

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

Definition 1.10 - *Residual*
The *Residual* is the difference between the true value of the *Response Variables* & our *Fitted Values*.

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

Definition 1.11 - *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

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

1.3 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 1.12 - *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 1.1 - *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 1.2 - *Bootstrap Percentiles*
When wishing to create an interval for θ using *Bootstrapped* data, we treat the $\hat{\theta}$ values as if they came from $\mathbb{P}(\mu|X)$.

1.4 Hypothesis Testing

Definition 1.13 - *Simple Hypothesis*
A *Simple Hypothesis* states that a parameter takes an exact value.
i.e. $\theta = \theta_0$ for $\theta_0 \in \Theta$.

Definition 1.14 - *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 1.15 - *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 1.16 - *p-Value*
The *p-Value* of a *Test Statistic* is the probability of observing a value more extreme than the one produced by the observed data, $\mathbf{x}$, under the *Null Hypothesis*.

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

i.e. Under the null-hypothesis what is the probability of observing a more extreme test statistic value.

Remark 1.6 - $p(\boldsymbol{x})$ *is the smallest Significance Level at which we would reject the Null Hypothesis*

Definition 1.17 - *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 1.3 - *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 1.18 - *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

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

1.5 Intervals

Definition 1.19 - *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 1.20 - *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 1.21 - *Confidence Interval*
A $1-\alpha$ *Confidence Interval* for a parameter is an interval with *Coverage* at least $1-\alpha$.

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

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

Remark 1.7 - *Confidence Intervals are a part of Frequentist Statistics, not Bayesian*

Definition 1.22 - *Credible Interval*

Remark 1.8 - *Credible Intervals are a part of Bayesian Statistics, not Frequentist*

1.6 Causality

Definition 1.23 - *Causality*
Causality is a relationship between two events where one of the events caused the other.

Definition 1.24 - *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 1.9 - *When two variables are highly correlated it is hard to distinguish their effects*

Definition 1.25 - *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*

1.6.1 Controlled Experiments

Definition 1.26 - *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 1.10 - *Randomisation is the Gold Standard for Inferring Causation*

Proposition 1.4 - *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,
$\implies \tilde{\boldsymbol{\beta}}_X = \overline{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}} = \hat{\boldsymbol{\beta}}_X$.
N.B. Randomisation ensures $\mathbf{X}$ and $\mathbf{H}$ are independent.

1.6.2 Instrumental Variables

Remark 1.11 - *Motivation*
In some scenarios it is impractical or unethical to perform randomised experiments.

Definition 1.27 - *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 1.12 - *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 1.5 - *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.

1.7 Regularity Conditions

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

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

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

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*

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

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*

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

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

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

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

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

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.

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

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 are 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 not base don independent observations then $\frac{\partial \ell}{\partial \boldsymbol{\theta}}$ usually has a limiting normal distribution so the result holds anyways. □

3.3 Hypothesis Teting

Definition 3.4 - *Neyman-Pearson Lemma*

Definition 3.5 - *Generalise Likelihood Ratio Test Statistic*

3.4 Numerical Optimisation

Definition 3.6 - *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 minimum so we use the negative of the *Objective Function* in order to find the maximum instead.

Definition 3.7 - *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 - *Requirements for Newton's Method*
In order to guarantee that *Newton's Method* converges to the MLE, we need to ensure the followin

i) The approximating quadratic actually has a maximum (not a minimum, inflection point etc.).
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 3.8 - *Newton's Method*
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.*
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`$(\ldots,$ `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.

3.5 Intervals

0 Appendix

0.1 Definitions

Definition 0.1 - *Parametric Models*
Parameteric Models are *Statistical Models* whose only unkowns are parameters.

Definition 0.2 - *Semi-Parametric Models*
Parameteric Models are *Statistical Models* which contain unknown parameters and unknown functions.

Definition 0.3 - *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 0.4 - *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 0.5 - *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 0.6 - *Upper Triangle Matrix*
A matrix X is an *Upper Triangle Matrix* if $X_{i,j} = 0$ for $i > j$.

Definition 0.7 - *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 0.8 - *Conjugacy*
Definition 0.9 - *Fisher Information*
Definition 0.10 - *Correlation*
Definition 0.11 - *Covariance*
Definition 0.12 - *Expected Value*
Definition 0.13 - *Variance*

Definition 0.14 - *Positive Definite Matrix*
A matrix is *Positive Definite* if it is symmetric and all its eigenvalues are positive.

Definition 0.15 - *Positive Semi-Definite Matrix*
A matrix is *Postitive Semi-Definite* if all its eigenvalues are non-negative.
Definition 0.16 - *Taylor's Theorem*
Definition 0.17 - *Casual Inference*

Definition 0.18 - *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

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