Statistics

Shing Tak Lam

May 11, 2022

Contents

1 Estimation

Unless explicitly mentioned, we have data $X = (X_1, \ldots, X_n)$, where each one is iid with density $f_X(x \mid \theta)$, θ unknown parameter.

Definition 1.1 (Statistic)

A statistic is a function on the data.

Definition 1.2 (Estimator)

An estimator $\hat{\theta}$ of θ is a statistic which does not explicitly depend on θ, and can be used to estimate the true value of θ.

Definition 1.3 (Sampling distribution)

Suppose T is an estimator. Then the distribution of (the random variable) $T(X)$ is known as the sampling distribution.

Definition 1.4 (Bias)
The bias of an estimator is

$$\text{bias}(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}\right] - \theta$$

Remark 1.5. $\text{bias}(\hat{\theta})$ is a function of the true parameter θ. If $\text{bias}(\hat{\theta}) = 0$ for all θ, then we say $\hat{\theta}$ is unbiased.

Definition 1.6 (Mean squared error)
The mean squared error of an estimator is

$$\text{mse}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

Proposition 1.7 (Bias-variance decomposition).

$$\text{mse}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \left(\text{bias}(\hat{\theta})\right)^2$$

1.1 Sufficiency

Definition 1.8 (Sufficient statistic)
A statistic T is sufficient if the conditional distribution of X given $T(X)$ does not depend on θ.

Theorem 1.9 (Factorisation criterion). T is sufficient for θ if and only if $f_X(x \mid \theta) = g(T(x), \theta)h(x)$ for some g, h.

Proof. We only prove the discrete case. Suppose the density factorises, and suppose $T(x) = t$. Then

$$\begin{aligned} f(x \mid T(x) = t) &= \frac{\mathbb{P}(X = x, T(x) = t)}{\mathbb{P}(T(x) = t)} \\ &= \frac{g(T(x), \theta)h(x)}{\sum_{y \in T^{-1}\{t\}} g(T(y), \theta)h(y)} \\ &= \frac{g(t, \theta)h(x)}{\sum_{y \in T^{-1}\{t\}} g(t, \theta)h(y)} \\ &= \frac{h(x)}{\sum_{y \in T^{-1}\{t\}} h(y)} \end{aligned}$$

is independent of θ. Conversely, suppose T is sufficient.

$$f(x \mid \theta) = \mathbb{P}(X = x) = \mathbb{P}(X = x, T(x) = T(x)) = \mathbb{P}(X = x \mid T(X) = T(x))\mathbb{P}(T(X) = T(x))$$

□

Definition 1.10 (Minimal sufficient)

A sufficient statistic T is minimal sufficient if it is a function of every other sufficient statistic. In particular, if S is also a sufficient statistic, then $S(x) = S(y) \implies T(x) = T(y)$.

Theorem 1.11. Suppose $T(x) = T(y)$ if and only if $\frac{f(x|\theta)}{f(y|\theta)}$ is constant in θ. Then T is minimal sufficient.

Remark 1.12. Write $x \sim_1 y$ for $\dfrac{f(x \mid \theta)}{f(y \mid \theta)}$ is constant in θ, and $x \sim_2 y$ for $T(x) = T(y)$. Then these define equivalence relations. So we have that the equivalence classes are the same.

Proof. For any value t of the statistic, let z_t be a representative from $T^{-1}\{t\}$. Then

$$f(x \mid \theta) = f(z_{T(x)} \mid \theta)\frac{f(x \mid \theta)}{f(z_{T(x)} \mid \theta)}$$

So T is sufficient by factorisation. Now let S be any sufficient statistic. Then we have factorisation $f(x \mid \theta) = g(S(x), \theta)h(x)$ for some g, h. Suppose $S(x) = S(y)$. Then

$$\frac{f(x \mid \theta)}{f(y \mid \theta)} = \frac{g(S(x), \theta)h(x)}{g(S(y), \theta)h(y)} = \frac{h(x)}{h(y)}$$

is constant in θ. So $x \sim_1 y$, which means that $x \sim_2 y$, so $T(x) = T(y)$. □

Theorem 1.13 (Rao-Blackwell)**.** Let T be a sufficient statistic for θ, and $\tilde{\theta}$ be an estimator with finite second moment $\mathbb{E}[\tilde{\theta}^2]$. Define $\hat{\theta} = \mathbb{E}[\tilde{\theta} \mid T(X)]$. Then $\text{mse}(\hat{\theta}) \leq \text{mse}(\tilde{\theta})$, where equality holds if and only if $\tilde{\theta}$ is a function of T. Furthermore, $\text{bias}(\hat{\theta}) = \text{bias}(\tilde{\theta})$.

Proof.

$$\text{mse}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \mathbb{E}\left[\left(\mathbb{E}\left[\tilde{\theta} \mid T\right] - \theta\right)^2\right] = \mathbb{E}\left[\mathbb{E}[\tilde{\theta} - \theta \mid T]^2\right] \leq \mathbb{E}\left[\mathbb{E}\left[(\tilde{\theta} - \theta)^2 \mid T\right]\right] = \text{mse}(\tilde{\theta})$$

By the tower law, $\mathbb{E}\left[\hat{\theta}\right] = \mathbb{E}\left[\mathbb{E}\left[\tilde{\theta} \mid T(X)\right]\right] = \mathbb{E}\left[\tilde{\theta}\right]$, so $\text{bias}(\hat{\theta}) = \text{bias}(\tilde{\theta})$. □

1.2 Maximum likelihood estimators

Definition 1.14 (Likelihood)

The likelihood function is

$$L(\theta) = f_X(x \mid \theta) = \prod_i f_{X_i}(x_i \mid \theta)$$

Definition 1.15 (Log-likelihood)

The log-likelihood function is

$$\ell(\theta) = \log(L(\theta)) = \sum_i \log(f_{X_i}(x_i \mid \theta))$$

Definition 1.16 (Maximum likelihood estimator)

A maxmimum likelihood estimator for θ is one that maxmimises L or ℓ over all θ.

Proposition 1.17. If T is a sufficient statistic for θ, then the MLE is a function of T.

Proof. Since T is sufficient, we have that $L(\theta) = g(T(x), \theta)h(x)$, so it only depends on $T(x)$. □

Proposition 1.18. If $\phi = H(\theta)$, where H is a bijection, and $\hat{\theta}$ is the MLE for θ, then $H(\hat{\theta})$ is the MLE for ϕ.

Proposition 1.19 (Asymptotic normality).

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \approx N(0, \Sigma) \quad \text{as} \quad n \to \infty$$

1.3 Confidence intervals

Definition 1.20 (Confidence interval)
For $\gamma \in (0, 1)$, a $100\gamma\%$ confidence interval for the parameter γ is a random interval $[A(X), B(X)]$ such that

$$\mathbb{P}(A(X) \leq \theta \leq B(X)) = \gamma$$

for all θ fixed.

Definition 1.21 (Pivot)
A pivot $R(X, \theta)$ is a random variable whose distribution is independent of θ.

Remark 1.22. Finding a $100\gamma\%$ CI for R gives us a way to find a $100\gamma\%$ CI for θ.

Definition 1.23 (Confidence set)
A random set $A(X)$ is a $100\gamma\%$ confidence set if

$$\mathbb{P}(\theta \in A(X)) = \gamma$$

2 Bayesian inference

Definition 2.1 (Prior)
The parameter θ is considered to be a random variable. We call the distribution of θ, $\pi(\theta)$ the prior.

Definition 2.2 (Posterior)
The posterior distribution of θ, is

$$\pi(\theta \mid x) = \frac{\pi(\theta) f(x \mid \theta)}{f(x)}$$

where $f(x \mid \theta)$ is the likelihood.

Definition 2.3 (Bayes estimator)
For a parameter θ, the Bayes estimator is the one which minimises

$$h(\delta) = \int L(\theta, \delta)\pi(\theta \mid x)\mathrm{d}\theta$$

where $L(\theta, \delta)$ is a loss function.

Proposition 2.4. With $L(\theta, \delta) = (\theta - \delta)^2$, the Bayes estimator is the expectation.

Proposition 2.5. With $L(\theta, \delta) = |\theta - \delta|$, the Bayes estimator is the median.

Definition 2.6 (Credible interval)
For a fixed x, a $100\gamma\%$ credible interval for the parameter θ is an interval $[A(x), B(x)]$ such that $\pi(A(x) \leq \theta \leq B(x)) = \gamma$

3 Hypothesis testing

Definition 3.1 (Hypothesis)
A hypothesis is an assumption about the distribution of the data X.

3.1 Simple hypotheses

Definition 3.2 (Simple hypothesis)
A simple hypothesis is one which fully specifies the distribution of the data X. If a hypothesis isn't simple, then we say that it is composite.

Definition 3.3 (Critical region)
A test of H_0 is defined by a critical region C, where we reject H_0 if $X \in C$, and accept H_0 if $X \notin C$.

Definition 3.4 (Type I, II error)
Type I error is H_0 is true and we reject it (false positive). Type II error is H_0 is false and we accept it (false negative).

Definition 3.5 (Size, power)
For simple hypothese H_0, H_1 define

$$\alpha = \mathbb{P}_{H_0}(X \in C) \quad \text{and} \quad \beta = \mathbb{P}_{H_1}(X \notin C)$$

The size of the test C is α, and the power is $1-\beta$.

Definition 3.6 (Likelihood ratio test)

Suppose H_0, H_1 are simple hypotheses, and we have distributions f_0, f_1 respetcively. Define the likelihood ratio

$$\Lambda = \frac{f_1(x)}{f_0(x)}$$

The likelihood ratio test is

$$C = \{x : \Lambda > k\}$$

Theorem 3.7 (Neyman-Pearson). Suppose $\text{Supp}(f_0) = \text{Supp}(f_1)$, and there exist $k > 0$ which gives a likelihood ratio test C of size α. Then among all tests with size $\leq \alpha$, the LRT has the smallest β.

Proof. Let $\overline{C} = X \setminus C$. Then

$$\alpha = \int_C f_0(x)\mathrm{d}x \quad \text{and} \quad \beta = \int_{\overline{C}} f_1(x)\mathrm{d}x$$

Let C^* be any other test, with size $\alpha^* \leq \alpha$, and power $1-\beta^*$. Then

$$\begin{aligned}
\beta - \beta^* &= \int_{\overline{C}} f_1(x)\mathrm{d}x - \int_{\overline{C^*}} f_1(x)\mathrm{d}x \\
&= \int_{\overline{C}\cap C^*} f_1(x)\mathrm{d}x - \int_{C\cap\overline{C^*}} f_1(x)\mathrm{d}x \\
&= \int_{\overline{C}\cap C^*} \Lambda f_0(x)\mathrm{d}x - \int_{C\cap\overline{C^*}} \Lambda f_0(x)\mathrm{d}x \\
&\leq k\left(\int_{\overline{C}\cap C^*} f_0(x)\mathrm{d}x - \int_{C\cap\overline{C^*}} f_0(x)\mathrm{d}x\right) \\
&= k\left(\int_{C^*} f_0(x)\mathrm{d}x - \int_C f_0(x)\mathrm{d}x\right) \\
&= k(\alpha^* - \alpha) \\
&\leq 0
\end{aligned}$$

□

Definition 3.8 (p-value)

For a test statistic T, and a critical region $\{x : T(x) > k\}$, then the p value of an observation x^* is

$$p = \mathbb{P}_{H_0}(T(X) \geq T(x^*))$$

3.2 Compostite hypotheses

Suppose we have hypotheses

$$H_0 : \theta \in \Theta_0 \quad \text{and} \quad H_1 : \theta \in \Theta_1$$

Definition 3.9 (Power)
The power function for a test C is

$$W(\theta) = \mathbb{P}_\theta(X \in C)$$

Definition 3.10 (Size)
The size of a test C is

$$\alpha = \sup_{\theta \in \Theta_0} W(\theta)$$

Definition 3.11 (Uniformly most powerful)
A test C with power W and size α is UMP if for any other test C^* with size $\alpha^* \leq \alpha$, we have $W(\theta) \geq W^*(\theta)$ for all $\theta \in \Theta_1$.

Definition 3.12 (Generalised likelihood ratio)
The GLR is

$$\Lambda = \frac{\sup_{\theta \in \Theta_1} f(x \mid \theta)}{\sup_{\theta \in \Theta_0} f(x \mid \theta)}$$

Definition 3.13 (Dimension)
The dimension of a hypothesis $\theta \in \Theta$ is the number of free parameters in Θ.

Theorem 3.14 (Wilks' theorem). Suppose $\Theta_0 \subseteq \Theta_1$, and $\dim(\Theta_1) - \dim(\Theta_0) = p$. $X_1, \ldots, X_n$ are iid under $f(\cdot \mid \theta)$, and $\theta \in \text{Int}(\theta_0)$. Then as $n \to \infty$, $2\log(\Lambda) \to \chi^2_p$.

3.3 Goodness of fit tests

Suppose we have $N_1, \ldots, N_k \sim \text{Multi}(n; p_1, \ldots, p_k)$. A goodness of fit test is a test of the hypotheses

$$H_0 : p = \tilde{p} \quad \text{vs} \quad H_1 : p \text{ arbitrary probability vector}$$

Then $L(p) \propto p_1^{N_1} \cdots p_k^{N_k}$, and

$$2\log(\Lambda) = 2\sum N_i \log\left(\frac{N_i}{np_i}\right) = 2\sum O_i \log\left(\frac{O_i}{E_i}\right)$$

where $O_i = N_i$ is the observed number of samples of type i, and $E_i = np_i$ is the expected number of samples of type i.

Definition 3.15 (Pearson's χ^2 statistic)
The Pearson's χ^2 statistic is

$$\sum \frac{(O_i - E_i)^2}{E_i}$$

Proposition 3.16.

$$2\log(\Lambda) \approx \sum \frac{(O_i - E_i)^2}{E_i}$$

3.4 Independence tests

Suppose we have $(X_1, Y_1), \dots, (X_n, Y_n)$ iid variables in $\{1, \dots, r\} \times \{1, \dots, c\}$. We wish to test whether X and Y are independent. Define

$$N_{ij} = \#\{l : (X_l, Y_l) = (i, j)\}$$

Then $N_{ij} \sim \text{Multi}(n; P)$. An independence test is a test of the hypotheses

$$H_0 : P_{ij} = P_{i+}P_{+j} \quad \text{vs} \quad H_1 : P \text{ arbitrary}$$

Then

$$2\log(\Lambda) = 2\sum_{i,j} N_{ij} \log\left(\frac{\hat{p}_{ij}}{\hat{p}_{i+}\hat{p}_{+j}}\right)$$

3.5 Homoegeneity tests

Suppose we have $(N_{i1}, \dots, N_{ic}) \sim \text{Multi}(n_{i+}; p_{i1}, \dots, p_{ic})$ independently for $i \in \{1, \dots, r\}$. Then we wish to test

$$H_0 : p_{1j} = \cdots = p_{rj} \text{ for all } j \quad \text{vs} \quad H_1 : p \text{ arbitrary}$$

Then

$$2\log(\Lambda) = \sum_{i,j} N_{ij} \log\left(\frac{N_{ij}}{\frac{n_{i+}N_{+j}}{n_{++}}}\right)$$

which is the same as the test for independence.

3.6 Confidence intervals

Definition 3.17 (Acceptance region)
The acceptance region of a test C is $A = X \setminus C$

Theorem 3.18. Suppose for all $\theta_0 \in \Theta$ we have a test of size α for the hypothesis $\theta = \theta_0$ and acceptance region $A(\theta)$. Then

$$I(X) = \{\theta \in \Theta : X \in A(\theta)\}$$

is a $100(1 - \alpha)\%$ confidence set.

Theorem 3.19. Suppose we have $I(X)$ which is a $100(1 - \alpha)\%$ confidence set. Then the set

$$A(\theta_0) = \{x : \theta_0 \in I(X)\}$$

is the acceptance region of a test with size α for the hypothesis $\theta = \theta_0$.

4 Multivariate normal models

Definition 4.1 (Multivariate normal)
X is multivariate normal if for all $t \in \mathbb{R}^n$, $t^\top X$ is normal.

Proposition 4.2. A MVN vector is fully specified by it's mean and variance.

Proof. Suppose X has mean μ and variance Σ. Then consider the moment generating function

$$M_X(t) = \mathbb{E}\left[e^{t^\top X}\right] = M_{t^\top X}(1) = \exp\left(t^\top \mu + \frac{1}{2} t^\top \Sigma t\right)$$

which is a function of only μ and Σ. □

4.1 Orthogonal projection

Definition 4.3 (Orthogonal projection)
A matrix P is an orthogonal projection onto its image $\mathrm{Col}(P)$ if for all $u \in \mathrm{Col}(P)$, $Pu = u$, and for all $w \in \mathrm{Col}(P)^\perp$, $Pw = 0$.

Proposition 4.4. P is an orthogonal projection if and only if $P = P^\top$ and $P^2 = P$.

Proposition 4.5. If P is an orthogonal projection, so is $I - P$.

Proposition 4.6. If P is an orthogonal projection, let $r = \mathrm{rank}(P)$. Then we have $U \in \mathrm{Mat}_{n,r}(\mathbb{R})$ such that $P = UU^\top$ and the columns of U are an orthonormal basis of $\mathrm{Col}(P)$.

Proposition 4.7. If P is an orthogonal projection, then $\mathrm{rank}(P) = \mathrm{tr}(P)$.

Theorem 4.8. Suppose $X \sim N(0, \sigma^2 I)$, and P is an orthogonal projection. Then

$$PX \sim N(0, \sigma^2 P) \quad \text{and} \quad (I - P)X \sim N(0, \sigma^2 (I - P))$$

Furthermore, they are independent.

Proof. By computing the mean and variance of $(PX, (I - P)X)$ we get the above result. To show they are independent, let $Z_1 \sim N(0, \sigma^2 P)$ and $Z_2 \sim N(0, \sigma^2 (I - P))$ be independent. Then (Z_1, Z_2) has the same mean and variance as $(PX, (I - P)X)$, so they are independent as a MVN is characterised by its mean and variance. □

Theorem 4.9. Suppose $X \sim N(0, \sigma^2 I)$, and P is an orthogonal projection. Then

$$\frac{\|PX\|^2}{\sigma^2} \sim \chi^2_{\mathrm{rank}(P)}$$

Proof. Note $\left\|PX\right\|^2 = \left\|U^\mathsf{T}X\right\|^2$, $U^\mathsf{T}X \sim N(0, \sigma^2 I_{\mathrm{rank}(P)})$, and the result follows. □

Theorem 4.10. Suppose $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ iid. Let $\overline{X} = \frac{1}{n}\sum X_i$, and $S_{XX} = \sum (X_i - \overline{X})$. Then

$$\overline{X} \sim N(\mu, \frac{1}{n}\sigma^2) \quad \text{and} \quad \frac{S_{XX}}{\sigma^2} \sim \chi^2_{n-1}$$

Furthermore, these are independent.

Proof. Let $J = (1, \dots, 1)^\mathsf{T} \in \mathbb{R}^n$, $P = n^{-1}JJ^\mathsf{T}$ is a projection onto its image. Write $X = \mu J + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2 I)$. Then $PX = \mu J + P\varepsilon$, and each element of PX is $\overline{X}$. So $\overline{X}$ depends only on $P\varepsilon$. Furthermore,

$$S_{XX} = \sum (X_i - \overline{X})^2 = \left\|(I-P)X\right\|^2 = \left\|(I-P)\varepsilon\right\|^2$$

which gives the required results. □

4.2 Linear model

Definition 4.11 (Linear model)
Let $X \in \mathrm{Mat}_{n,p}(\mathbb{R})$, where each row is a data point in $\mathbb{R}^p$. $\beta \in \mathbb{R}^p$, then a linear model is of the form

$$Y = X\beta + \varepsilon$$

where ε is a random vector with mean 0 and variance $\sigma^2 I$.

Definition 4.12 (Moore-Penrose inverse)
The Moore-Penrose inverse of a matrix X with linearly independent columns is $(X^\mathsf{T}X)^{-1}X^\mathsf{T}$.

Definition 4.13 (Least squares estimator)
The least squares estimator for β is $\hat{\beta} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$.

Proposition 4.14. $\hat{\beta}$ minimises

$$S(\beta) = \left\|Y - X\beta\right\|^2$$

Proposition 4.15.

$$\mathbb{E}\left[\hat{\beta}\right] = \beta \quad \text{and} \quad \mathrm{Var}(\hat{\beta}) = \sigma^2(X^\mathsf{T}X)^{-1}$$

Theorem 4.16 (Gauss-Markov)**.** Let $\beta^* = CY$ be any other linear unbiased estimator. Then $\mathrm{Var}(\hat{\beta}) \leq \mathrm{Var}(\beta^*)$.

Remark 4.17. We define $A \leq B$ if $B - A$ is positive semidefinite.

Proof. Let $A = C - (X^\mathsf{T}X)^{-1}X^\mathsf{T}$. Then $\mathbb{E}[AY] = AX\beta = 0$, which means that $AX = 0$. Then

$$\operatorname{Var}(\beta^*) = \operatorname{Var}((A + (X^\mathsf{T}X)^{-1}X^\mathsf{T})Y) = (A + (X^\mathsf{T}X)^{-1}X^\mathsf{T})\operatorname{Var}(Y)(A + (X^\mathsf{T}X)^{-1}X^\mathsf{T})^\mathsf{T} = \sigma^2 AA^\mathsf{T} + \operatorname{Var}(\hat{\beta}) \geq \operatorname{Var}(\hat{\beta})$$

□

Definition 4.18 (Fitted values)
The fitted values are

$$\hat{Y} = X\hat{\beta} = PY$$

where $P = X(X^\mathsf{T}X)^{-1}X^\mathsf{T}$.

Definition 4.19 (Residuals)
The residuals are $Y - \hat{Y} = (I - P)Y$.

Proposition 4.20. P is an orthogonal projection onto its image.

4.3 Normal linear model

From now on assume $\varepsilon \sim N(0, \sigma^2 I)$. The log likelihood is

$$\ell(\beta, \sigma^2) = \text{const} - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\left\|Y - X\beta\right\|^2$$

Proposition 4.21. The MLE for β is the least squares estimator.

Proposition 4.22. The MLE for σ^2 is given by

$$\hat{\sigma}^2 = \frac{\left\|Y - X\beta\right\|^2}{n} = \frac{\left\|\hat{Y} - Y\right\|^2}{n} = \frac{\left\|(I - P)Y\right\|^2}{n}$$

Theorem 4.23.

$$\hat{\beta} \sim N(\beta, \sigma^2(X^\mathsf{T}X)^{-1}) \quad \text{and} \quad \frac{\hat{\sigma}^2 n}{\sigma} \sim \chi^2_{n-p}$$

Proof. $\hat{\beta}$ is a linear function of Y, so it is MVN. We have already computed the mean and variance previously.

Note that

$$n\hat{\sigma}^2 = \left\|(I - P)Y\right\|^2 = \left\|(I - P)(X\beta + \varepsilon)\right\|^2 = \left\|(I - P)\varepsilon\right\|^2$$

since $I - P$ is a projection onto $\operatorname{Col}(X)^\perp$, $(I - P)X = 0$. $I - P$ has rank $n - p$, which gives the required result.

For independence, $\hat{\sigma}^2$ is a function of $(I - P)\varepsilon$, and $\hat{\beta}$ is a function of $P\varepsilon$. □

4.4 Hypothesis testing

Definition 4.24 (Student t distribution)
Suppose $U \sim N(0,1)$, $V \sim \chi^2_n$ independent. Then

$$T = \frac{U}{\sqrt{V/N}} \sim t_n$$

Definition 4.25 (F distribution)
Suppose $V \sim \chi^2_n$, $W \sim \chi^2_m$ independent. Then

$$F = \frac{V/n}{W/m} \sim F_{n,m}$$

Proposition 4.26.

$$\frac{\frac{\beta_i - \hat{\beta}_i}{\sqrt{(X^\top X)^{-1}_{ii}}}}{\sqrt{\hat{\sigma}^2 \frac{n}{n-p}}} \sim t_{n-p}$$

Proof. Multiply numerator and denominator by σ. □

Proposition 4.27.

$$I = \hat{\beta}_i \pm t_{n-p}\left(\frac{\alpha}{2}\right)\sqrt{\frac{(X^\top X)^{-1}_{ii}\hat{\sigma}^2}{(n-p)/n}}$$

is a $100(1-\alpha)\%$ CI for β_i.

Proposition 4.28.

$$\frac{\left\|X(\hat{\beta} - \beta')\right\|^2 / p}{\hat{\sigma}^2 n/(n-p)} \leq F_{p,n-p}(\alpha)$$

Defines a $100(1-\alpha)\%$ confidence set for β.

Proof. Multiply numerator and denominator by σ. □

F test

We wish to test

$$H_0 : \beta_1 = \cdots = \beta_{p_0} = 0 \quad \text{vs} \quad H_1 : \beta \text{ arbitrary}$$

Write $X = \begin{pmatrix} X_0 & X_1 \end{pmatrix}$, where $X_0 \in \mathrm{Mat}_{n,p_0}(\mathbb{R})$ and $X_1 \in \mathrm{Mat}_{n,p-p_0}(\mathbb{R})$, and $\beta = \begin{pmatrix} \beta^{(0)} & \beta^{(1)} \end{pmatrix}^\top$, where $\beta_0 \in \mathbb{R}^p_0$ and $\beta_1 \in \mathbb{R}^{p-p_0}$.

Then under H_0, $\beta^{(0)} = 0$, so $Y = X\beta + \varepsilon = X_1\beta^{(1)} + \varepsilon$. Define $P = X(X^\top X)^{-1}X^\top$ and $P_1 = X_1(X_1^\top X_1)^{-1}X_1^\top$.

Proposition 4.29.

$$\mathrm{rank}(P_1) = p - p_0$$

Lemma 4.30.

$$(I - P)(P - P_1) = 0$$

Lemma 4.31. $P - P_1$ is an orthogonal projection with rank p_0.

The log-GLR is

$$2\log(\Lambda) = n\left(\log\left(\frac{\left\|(I-P_1)Y\right\|^2}{n}\right) - \log\left(\frac{\left\|(I-P)Y\right\|^2}{n}\right)\right)$$

which is monotone in $\dfrac{\left\|(I-P_1)Y\right\|^2}{\left\|(I-P)Y\right\|^2} = \dfrac{\left\|(I-P)Y\right\|^2 + \left\|(P-P_1)Y\right\|^2}{\left\|(I-P)Y\right\|^2}$.

Theorem 4.32.

$$\frac{\left\|(P-P_1)Y\right\|^2}{\left\|(I-P)Y\right\|^2} \sim F_{p_0, n-p}$$

Proposition 4.33. When $p_0 = 1$ we recover the t-test.

4.5 Applications

Categorical predictors

Suppose $X_{i,j} \in \{0, 1\}$ for all i, j, and we wish to test the linear model with an intercept term. That is, $X_{i,1} = 1$ for all i. Then the design matrix X does not have full rank, so the Moore-Penrose inverse $(X^\mathsf{T}X)^{-1}X^\mathsf{T}$ is not invertible. One way of fixing this is by removing one of the columns, or equivalently, setting the coefficient for one category to 0. The result will be the same as the matrix will have the same column space, so P and the fitted values will be the same.

ANOVA

Suppose we have $Y_{ij} = \alpha + \mu_j + \varepsilon_{ij}$, $i = 1, \dots, N$ and $j = 1, \dots, J$. Concatenating them along the i-axis, and using the corner point constraint, we have a linear model. Then the F test is

$$F = \frac{\frac{1}{J-1}\sum_{j=1}^{J} N(\overline{Y_j} - \overline{Y})^2}{\frac{1}{JN-J}\sum_{j=1}^{J}\sum_{i=1}^{N}(Y_{ij} - \overline{Y_j})^2}$$

which is known as the ANalysis Of VAriances test, as it is

$$F = \frac{\text{Variance between groups}}{\text{Variance within each group}}$$

Two sample testing

Suppose $Y_1, \dots, Y_n \sim N(\mu_1, \sigma^2)$ iid, and $Z_1, \dots, Z_m \sim N(\mu_2, \sigma^2)$ iid. Furthermore, Y and Z are independent. We can test the hypotheses

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2$$

using the F-test, by concatenating, and using an appropriate design matrix.