Statistics

Klejdi Sevdari

January 2022

Optimality Theory

Sufficient Statistic: V is a sufficient statistic if $P(X = x|V = v)$ does not depend on θ.

Factorization theorem **(p)**: $V = V(X)$ is sufficient iff there exist functions $g(\theta)$ and h such that for all x and θ, $p_\theta(x) = g_\theta(V(x))h(x)$.

V is minimally sufficient iff it is a function of every other sufficient statistic.

Transformations of sufficient statistics: Let V be a sufficient statistic and suppose that $V = f(V*)$. Then $V*$ is also sufficient. If f is one to one then the converse holds.
UMVU: T is called UMVU for $g(\theta)$ if T is an unbiased estimator for $g(\theta)$ and $var_\theta T \leq var_\theta S \; \forall\theta$ and all unbiased estimators S for $g(\theta)$.

Rao-Blackwell: Let $V = V(X)$ be a sufficient statistics and let $T = T(X)$ be an arbitrary real-values estimator for $g(\theta)$. Then there exists an estimator $T^* = T^*(V)$ for $g(\theta)$ that depends only on V, such that $E_\theta T^* = E_\theta T$ and $var_\theta T^* \leq var_\theta T$ for all θ. In particular we have $MSE(T^*) \leq MSE(T)$.

A family of probability densities p_θ is called a $k-$dimensional exponential family if $p_\theta(x) = c(\theta)h(x)\exp(\sum Q_j(\theta)V_j(x))$.

Theorem for Expo family and UMVU: If the set $\{(Q_1(\theta), \ldots, Q_k(\theta)) : \theta \in \Theta\}$ has an interior point, then any unbiased estimator $T = T(V)$ for $g(\theta)$ is UMVU.

Cramer-Rao Lower Bound **(p)**: $var_\theta T \geq g'(\theta)^2/I_\theta$ for any unbiased T (if $p_\theta(x)$ is differentiable for every x).

The MLE is asymptotically UMVU. $E_\theta\hat{\theta}_n, var_\theta\hat{\theta}_n = 1/ni_\theta$.

Neyman Pearson Lemma: $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$. $L(\theta_0, \theta_1, X) = p_{\theta_1}(x)/p_{\theta_0}(x)$. Suppose that these exists a number c_{α_0} such that $P_{\theta_0}(L(\theta_1, \theta_0, X) \geq c_{\alpha_0}) = \alpha_0$. Then the test with critical region $K = \{x : L(\theta_0, \theta_1, X) \geq c_{\alpha_0}\}$ is the most powerful at level α_0 for testing the aforementioned hypothesis.

Main Regression models

Linear Regression

- Assumptions: $E[Y] = \sum_{i=1}^{p} \beta_i x_i$, $e_i \sim N(0, \sigma^2)$.
- Simple Linear Regression with intercept:
  - Model: $Y_i = \alpha + \beta x_i + e_i$, $e_i \overset{iid}{\sim} N(0, \sigma^2) \implies Y_i \overset{indep.}{\sim} N(\alpha + \beta x_i, \sigma^2)$
  - Log-Likelihood: $-\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2$
  - MLE: $\hat{\alpha} = \overline{Y} - \hat{\beta}\bar{x}$, $\hat{\beta} = \frac{s_Y}{s_x}r_x, Y$, $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$

- $SS_{tot} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, $SS_{res} = \sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$
- Residuals: $\hat{e_i} = Y_i - \hat{\alpha} - \hat{\beta}x_i, i = 1, ..., n$. Check constant variance by plotting residuals vs fitted values. Check normality with QQ-plots/histograms and tests (Shapiro-Wilk, Kolmogorov-Smirnov)
- T-Test for $H_0 : \beta = \beta_0$ vs $H_1 : \beta \neq \beta_0$: $T = \frac{\hat{\beta} - \beta_0}{\sqrt{\hat{var}\hat{\beta}}}$. Under H_0, $T \sim t_{n-2}$

- Multiple Linear Regression:
  - Model: $Y_i = \sum_{j=1}^{p} \beta_j x_{i,j} + e_i$, $e_i \overset{iid}{\sim} N(0, \sigma^2) \implies Y_i \overset{indep.}{\sim} N(\sum_{j=1}^{p} \beta_j x_{i,j}, \sigma^2)$
  - Matrix form: $Y = X\beta + e \implies Y \sim N(X\beta, \sigma^2 I_n)$. If we have an intercept, first column of X is made of 1's.
  - Log-Likelihood: $-\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}||Y - X\beta||^2$
  - MLE: $\hat{\beta} = (X^T X)^{-1} X^T Y$, $\hat{\sigma^2} = \frac{1}{n}||Y - X\hat{\beta}||^2$
  - $SS_{tot} = ||Y - \bar{Y_1}||, SS_{res} = ||Y - X\hat{\beta}||^2$
  - Residuals vector: $\hat{e} = Y - X\hat{\beta}$. Check above for diagnostics
  - F-Test for all predictors assuming we have an intercept (full model vs empty model): $H_0 : \beta_2 = ... = \beta_p = 0$ vs $H_1 : \exists \beta_i \neq 0, i = 2, ..., p$: Use test statistic $F = \frac{(SS_{tot} - SS_{res})/(p-1)}{SS_{res}/(n-p)}$. Under $H_0, F \sim F_{p-1,n-p}$
  - General F-test (small model vs larger model): $H_0 : \mu = X\beta \in V_0$ vs $H_1 : \mu = X\beta \notin V_0$: Use test statistic $F = \frac{||(P_V - P_{V_0})Y||^2/(p-p_0)}{||(I-P_V)Y||^2/(n-p)}$. Under $H_0, F \sim F_{p-p_0,n-p}$
  - Test one parameter (F-test/t-test): $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$: Use test statistic $T = \frac{\hat{\beta_j}}{\hat{se}\hat{\beta_j}}$. Under $H_0, T \sim t_{n-p}$

ANOVA(Two factor)

- Model: $Y_{ijk} = \mu_{ij} + e_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$ for $i = 1, ..., I, j = 1, ..., J, k = 1, ..., K_{ij}$
- Assumptions: $\sum_i \alpha_i = 0, \sum_j \beta_j = 0, \sum_{i=1}^{I} \gamma_{ij} = 0$ for $j = 1, ..., J, \sum_{j=1}^{J} \gamma_{ij} = 0$ for $i = 1, ..., I$
- The parameters satisfy: $\mu = \mu_{..}$, $\alpha_i = \mu_{i.} - \mu$, $\beta_j = \mu_{.j} - \mu$, $\gamma_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}$
- Log-Likelihood function: $-\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i,j,k}(Y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2$
- MLE's:
  - $\hat{\mu} = Y_{...}$
  - $\hat{\alpha_i} = Y_{i..} - Y_{...}$
  - $\hat{\beta_j} = Y_{.j.} - Y_{...}$
  - $\hat{\gamma_{ij}} = Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...}$
  - $\hat{\sigma^2} = \frac{1}{n}\sum_{i,j,k} = (Y_{ijk} - \hat{\mu} - \hat{\alpha_i} - \hat{\beta_j} - \hat{\gamma_{ij}})^2$

Other Regression models

Nonlinear Regression

- Model: $Y_i = f_\theta(x_i) + e_i$ for some nonlinear equation of the parameter(s) θ.
- LSE: $\theta \to \sum(Y_1 - f_\theta(x_i))^2$.
- if the form of f is not specified beforehand, then we speak of nonparametric regression.

Logistic Regression

- Model: $P(Y=1|X=x) = 1/(1+\exp(-\sum_j \beta_j x_j))$.
- Likelihood function:

$$L(\beta, Y) = \prod \left(\frac{1}{1+\exp(-\sum_j \beta_j x_{i,j})} \right)^{Y_i} \left(1 - \frac{1}{1+\exp(-\sum_j \beta_j x_{i,j})} \right)^{1-Y_i}$$

- In R, in the summary of the output, Residual Deviance is $-2log(\hat{\mathcal{L}})$. The lower this is the better.
- For testing smaller models as compared to bigger models, you use the likelihood ratio statistic. The output will tell you if the larger model is better.

Mixed Models
A mixed model can be viewed as a regression model in which some of the parameters have been replaced by random variables, called random effects. Such a model is useful to model dependence between the observations or heterogeneities in the data.

- Linear mixed model: $Y = X\beta + Z\gamma + e$: X and Z are known $(n \times p)$ and $(n \times q)$ matrices. One can figure out the dimension of the other vectors from this information. The vector γ is not observed and it is used a device to model dependecies between the $Y_1, \ldots, Y_n$.

Model Selection

Step Down Method

- Start with the full model. Test all $H_0 : \theta_i = 0$ for all $i \in I$.
- Remove θ_i with the largest p-value (unless all parameters were significant) and repeat the procedure with the smaller model.

Step Up Method

- Start with the empty model. Test all $H_0 : \theta_i = 0$ for all $i \in I$.
- Add θ_i with the smallest p-value (unless none of the hypotheses were rejected) and repeat the procedure with the bigger model.

AIC
We observe a sample $X_1, \ldots, X_n$ from a density $p_{d,\theta}$. AIC maximizes:

$$\max d \to \log \prod p_{d,\hat{\theta}_d}(X_i) - |d|$$

The Kullback-Leibler divergence: $K(p, p_{d,\hat{\theta}_d}) = E_p \log p(X) - E_p \log p_{d,\hat{\theta}_d}(X)$.

Estimate for $-E_p \log p_{d,\hat{\theta}_d}(X)$: $-\frac{1}{n}\sum \log p_{d,\hat{\theta}_d}(X_i)$.

Oracle Inequality: $E_p K(p, p_{\hat{d},\hat{\theta}_{\hat{d}}}) \leq C \inf_d \left(K(p, \mathcal{P}_d) + |d|/n \right)$
Interpreting the Oracle inequality: The term $K(p, \mathcal{P}_d)$ can be viewed as a necessary, minimal bias when using the model $\mathcal{P}_d$. The term $|d|/n$ can be viewed as a variance term, or estimation inaccuracy, as a result of estimating $|d|$ unknown parameters when using the model $\mathcal{P}_d$. The infimum over d in the inequality shows that the model chosen by the AIC makes a smaller error than every other model (at least up to a multiplicative constant).

AIC downsides:

- The oracle inequality is not useful for a small number of low dimensional models. Models needs to have very different dimensions $|d|$, for the inequality to be useful.
- For large n, the AIC never chooses a small model.

BIC
Notation:

- Prior probabilities of the models: $p_d \ \forall d$.
- Prior density of the parameter set $\Theta_d \ \forall d: \ \theta \to \pi_d(\theta)$.
- The probability density of the data X given $(d, \theta): \ x \to p_{d,\theta}(x)$.

The conditional density of (d, θ):

$$\pi(d, \theta|X) \propto p_d \pi_d(\theta) p_{d,\theta}(X) \implies \pi(d|X) \propto \int p_d \pi_d(\theta) p_{d,\theta}(X) d\theta$$

Bayes factor for models d_1 and d_2:

$$BF(d_1, d_2) = \frac{\int \pi_{d_1}(\theta) p_{d_1,\theta}(x) d\theta}{\int \pi_{d_2}(\theta) p_{d_2,\theta}(x) d\theta}$$

Comparing AIC and BIC
For large n, the BIC penalty is significantly greater than the AIC penalty, which leads to the choice of smaller models than those given by the AIC. This higher penalty is necessary for consistent model selection. The AIC overestimates the the dimension and is not consistent. However, AIC gives good estimates for the distribution of the observations (due to minimization of the KL divergence). Consistent model selection shrinks the parameters too much for good predictions.

Cross Validation
Split the data into two data sets that have the same statistical properties and are statistically independent. Use one data set for estimation and the other for validation. Switch their roles and choose the model which performs best. Simple to implement and accurate if we have sufficient data (as data partitioning lead to a loss in efficiency). Used for finding models that predict well.

High Dimensional Statistics

High-dimensional: More parameters than observations.

Active set of $\beta \in \mathbb{R}^p$ is $S_\beta = \{j : \beta_j \neq 0\}$. Sparsity index is $s_\beta = |S_\beta| = \text{card}(S_\beta)$
Impossible to determine β, so we introduce an assumption: β is sparse: $s_\beta < n$. We thus introduce penalty terms that should induce sparse estimators. Penalty terms:

- l_0 penalty: penalize model size. $\hat{\beta} \in \text{argmin}_\beta \|Y - X\beta\|_2^2 + \lambda s_\beta$.
- Ridge. Computational advantage, not sparse: $\hat{\beta} \in \text{argmin}_\beta \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$
- LASSO, sparse: $\hat{\beta} \in \text{argmin}_\beta \|Y - X\beta\|_2^2 + \lambda |\beta|_1$

Sparse Gaussian Graphical models
We observe $X_1, \ldots, X_n \sim N_p(0, \Sigma)$ with $\Sigma \in \mathbb{R}^{p \times p}$ invertible. Define $\Omega = \Sigma^{-1}$.

Sparse Structure:

- $\Sigma_{i,j} = 0 \implies X_i$ independent from X_j.
- $\Omega_{i,j} = 0 \implies X_i$ independent from X_j given all the other X_k's.

Likelihood function: $p(X_1, \ldots, X_n) = \Pi \frac{|\Omega|^{1/2}}{(2\pi)^{p/2}} \exp(-\frac{1}{2} X^T \Omega X)$

Nonparametrics Statistics

Density Estimation
Model: $X_1, \dots, X_n$(i.i.d) $\sim f$, for some unkown $f \in \mathcal{F}$.
General idea: Relative frequency of observations within a small interval corresponds to the height of the density in this interval, hence:

$$f(x_0) \approx \frac{1}{|U|} \int_U f(u)du \approx \frac{\#\{i : X_i \in U\}}{n|U|}$$

Histogram estimation
For $a \in \mathbb{R}$, binwidth $h > 0$ let the number of observations in $I_{ahk} = (a + kh, a + (k + 1)h]$ be $N_{ahk} = \#\{i : X_i \in I_{ahk}\}$ for $k \in \mathbb{Z}$.

Estimator: $\hat{f}^{\text{hist}}_{\text{ahn}} = \frac{1}{nh} \sum_k N_{ahk} \mathbb{1}(\cdot \in (a + kh, a + (k + 1)h])$

Kernel Density Estimation
A kernel is a Lebesgue integrable function $K : \mathbb{R} \to \mathbb{R}$ satisfying $\int_{-\infty}^{\infty} K(u)du = 1$. Then the KDE, for bandwidth $h > 0$ is:

$$\hat{f}_n(x) = \frac{1}{nh} \sum_i K\left(\frac{X_i - x}{h}\right)$$

Some important kernels:

- Rectangular kernel: $K(u) = \frac{1}{2} 1(\cdot \in (-1, 1])$
- Triangular kernel (used in histogram integral): $K(u) = (1 - |u|)_+$
- Gaussian kernel: $K(u) = (2\pi)^{-\frac{1}{2}} e^{-\frac{u^2}{2}}$

The MSE is bounded in the following way:

$$\text{MSE}(\hat{f}_n(x)) = \text{Bias}^2 \hat{f}_n(x) + var(\hat{f}_n(x)) \leq \left(h^\beta \frac{|f|_{\mathcal{C}^\beta}}{l!} \int_{-\infty}^{\infty} |K(v)v^\beta| dv\right)^2 + \frac{\|f\|_{L^\infty} \|K\|^2_{L^2}}{nh} \leq C_1 h^{2\beta} + C_2 \frac{1}{nh}$$

Hölder Spaces
Holder seminorm:

$$|f|_{\mathcal{C}^\beta} = \sup_{x \neq y,\ x,y \in D} \frac{\left|f^{\lfloor \beta \rfloor}(x) - f^{\lfloor \beta \rfloor}(y)\right|}{|x - y|^{\beta - \lfloor \beta \rfloor}}$$