Statistical Modelling

401-3622-00L Autumn Semester 2019

Classical Linear Regression

Coefficients of determination R^2

- For OLS, $R^2 = r^2_{\hat{\boldsymbol{y}}\boldsymbol{y}}$. (For simple linear regression, $R^2 = r^2_{\boldsymbol{x}\boldsymbol{y}}$.)
- Model comparison only meaningful if every model uses the same response variable and has the same number of paramaters.
- R^2 can be arbitrarily low when model is completely correct (low signal to noise ratio), can be close to 1 when the model is totally wrong (misspecified), cannot be compared between a model with untransformed Y and one with transformed Y (can go down when the model assumptions are better fulfilled).

Statistical properties

Let $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $E(\boldsymbol{\epsilon}) = 0, \mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}, \mathrm{rank}(\boldsymbol{X}) = p$.

- (Gauss-Markov Theorem). Furthermore, let $\boldsymbol{b}$ be an arbitrary p-dimensional vector and $\hat{\boldsymbol{\beta}}$ the LSE. Then $\boldsymbol{b}^\top\hat{\boldsymbol{\beta}}$ has minimal variance amongst all linear unbiased estimators of $\boldsymbol{b}^\top\boldsymbol{\beta}$ (BLUE).
- Let furthermore $\boldsymbol{\epsilon}$ be normally distributed. Then $\boldsymbol{b}^\top\hat{\boldsymbol{\beta}}$ has minimal variance amongst all unbiased estimators of $\boldsymbol{b}^\top\boldsymbol{\beta}$ (UMVU).
- Additionally, assume $\boldsymbol{X}^\top\boldsymbol{X}/n \to \boldsymbol{V}$ as $n \to \infty$ where $\boldsymbol{V}$ is a positive definite matrix, and that $\max_j \boldsymbol{H}_{jj} \to 0$. Then $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2_{ML}$ and $\hat{\sigma}^2$ are consistent. The least squares estimator asymptotically follows a normal distribution $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \sigma^2\boldsymbol{V}^{-1})$.
- Assume normality of $\boldsymbol{\epsilon}$. $\hat{\beta}$ and $\hat{\sigma}^2$ are independent. $SSE/\sigma^2 \sim \chi^2_{n-p}$.

Exact F-test

- The least squares estimator under $H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d}$ gives $\hat{\boldsymbol{\beta}}^R = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{C}^\top(\boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})$.
- The difference in residual sum of squares $\Delta SSE = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})^\top(\boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})$.
- Assume normality of $\boldsymbol{\epsilon}$. ΔSSE and SSE are independent. Under H_0, $\Delta SSE/\sigma^2 \sim \chi^2_r$, the test statistic $F = \frac{n-p}{r}\frac{\Delta SSE}{SSE} = \frac{1}{r}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})^\top(\widehat{\mathrm{Cov}}(C\hat{\boldsymbol{\beta}}))^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}) \sim F_{r,n-p}$. A global F-test for $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$, $F = \frac{n-p}{k}\frac{R^2}{1-R^2} \sim F_{k,n-p}$.

Model specification

- Missing variables: Coefficients of model with missing variables $\tilde{\boldsymbol{\beta}}_1$ are biased, but have smaller variance, unless covariates are orthogonal.
- Irrelevant variables: Both $\hat{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_1$ are biased, but coefficients of model with irrelevant variables have larger variance, unless covariates are orthogonal.

- Prediction quality: Expected squared prediction error $SPSE = n\sigma^2 + |M|\sigma^2 + \sum_i(\mu_{iM} - \mu_i)$, where $y_{n+i} = \mu_i + \epsilon_{n+i}, i = 1, \ldots, n$ and $\mu_{iM} = E(\hat{y}_{iM})$.

Model selection

- Expected squared prediction error (SPSE)
  - Approach 1: Estimate the expected squared prediction error using independent data, $\widehat{SPSE} = \sum_{j=1}^m(y_{n+j} - \hat{y}_{jM})^2$.
  - Approach 2: Use all data to estimate $\hat{\boldsymbol{\beta}}_M$ and correct sum of squared residuals, $\widehat{SPSE} = \sum_{i=1}^n(y_i - \hat{y}_{iM})^2 + 2|M|\hat{\sigma}^2$.
- Mallow's complexity parameter $C_p = \frac{\sum_{i=1}^n(y_i - \hat{y}_{iM})^2}{\hat{\sigma}^2} - n + 2|M|$. Can be understood as estimate of $SMSE/\sigma^2$.
- Akaike information criterion $AIC = -2l(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2_{ML}) + 2(|M| + 1)$. In a linear model with Gaussian errors, $AIC = n\log(\hat{\sigma}^2_{ML}) + 2(|M| + 1)$ (ignoring constants).
- Bayesian information criterion $BIC = -2l(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2_{ML}) + \log(n)(|M| + 1)$. In a linear model with Gaussian errors, $BIC = n\log(\hat{\sigma}^2_{ML}) + \log(n)(|M| + 1)$ (ignoring constants).
- Adjusted coefficient of determination $\bar{R}^2 = 1 - \frac{n-1}{n-p}(1 - R^2)$.
- Leave-one-out cross validation $LOOCV = \frac{1}{n}\sum_{i=1}^n(y_i - \hat{y}^{-i}_{iM})^2$. Special case for LSE, $LOOCV = \frac{1}{n}\sum_{i=1}^n\left(\frac{y_i - \hat{y}_{iM}}{1 - \boldsymbol{H}_{iiM}}\right)^2$.

Model diagnostics

- Residual plots: (standardized or studentized) residuals against $\hat{y}_i$, (standardized or studentized) residuals against each predictor $\boldsymbol{x}_j$, "scale-location" plot ($\sqrt{|\text{standardised residual}_i|}$ against $\hat{y}_i$).
- Transformation: power family $\psi(u, \lambda) = \begin{cases} u^\lambda, & \text{for } \lambda \neq 0 \\ \log(u), & \text{for } \lambda = 0 \end{cases}$ (for strictly positive u), scaled power family $\psi_S(x, \lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \text{for } \lambda \neq 0 \\ \log(x), & \text{for } \lambda = 0 \end{cases}$ (for strictly positive x, preserves direction of association), variance stabilising transformation (for positive y):
  - $\sqrt{y}$: Relatively "mild" and most appropriate when the response follows a Poisson distribution
  - $\log(y)$: Most commonly used transformation, appropriate when the error standard deviation is a percent of the response (rather than absolute units), "empirical log rule"
  - $1/y$: Often applied when the response is a time until an event; Can be appropriate when responses are mostly close to 0, but occasional large values occur
- Autoccorelation detection: correlograms, Durbin-Watson test for $H_0 : \rho = 0$ in $AR(1)$ model, $d = \frac{\sum_{i=2}^n(\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=1}^n\hat{\epsilon}_i^2} \approx 2(1 - \hat{\rho})$.
- Colinearity analysis: variance inflation factor $VIF_j = \frac{1}{1-R_j^2}$, "serious collinearity problem" exists when $VIF_j > 10$.
- Outlier detection: studentised "leave-one-out" residuals $r_i^* = \frac{\hat{\epsilon}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + \boldsymbol{x}_i^\top(\boldsymbol{X}_{(i)}^\top\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i}} = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - \boldsymbol{H}_{ii}}} \sim t_{n-p-1}$.
- Influence analysis: leverage $1/n \leq \boldsymbol{H}_{ii} \leq 1$, should pay attention to observations with $\boldsymbol{H}_{ii} > 2p/n$ (twice the average), but large values do not necessarily lead to problems, Cook's distance $D_i = \frac{\|\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}}\|_2^2}{p\hat{\sigma}^2} = \frac{1}{p}\frac{\hat{\sigma}_i^2}{\hat{\sigma}^2(1 - \boldsymbol{H}_{ii})}\frac{\boldsymbol{H}_{ii}}{1 - \boldsymbol{H}_{ii}}$, plot of the standardized residuals against $\boldsymbol{H}_{ii}$ is often used to detect influential observations
- Linearity analysis: residual Prediction (RP) tests, scaled residuals $\tilde{\boldsymbol{r}} = \frac{\hat{\boldsymbol{\epsilon}}}{\|\hat{\boldsymbol{\epsilon}}\|_2^2}$ has the same distribution of $\frac{(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{z}}{\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{z}\|_2^2}$ where $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I})$

General linear models

Weighted least squares

- (Gauss-Markov Theorem). Under the assumptions of the general linear model, the WLS estimator has minimal variance among all linear and unbiased estimators.
- Assuming $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\boldsymbol{W}^{-1})$, can show that WLS estimator coincides with ML estimator for $\boldsymbol{\beta}$, and $\hat{\sigma}^2 = \frac{1}{n-p}\hat{\boldsymbol{\epsilon}}^\top\boldsymbol{W}\hat{\boldsymbol{\epsilon}}$ is unbiased.
- For grouped data with response as arithmetic mean $\bar{y}_i$, $\boldsymbol{W} = \mathrm{diag}(n_1, \ldots, n_G)$, for sum $n_i\bar{y}_i$, $\boldsymbol{W} = \mathrm{diag}(1/n_1, \ldots, 1/n_G)$.
- Two-stage least squares: Obtain estimates $\hat{\boldsymbol{\alpha}}$ from an unweighted regression between $\log(\hat{\epsilon}_i^2)$ and the variance explanatory variables $\boldsymbol{z}_i$. Then fit a linear model using the weights $\hat{\boldsymbol{w}}_i = 1/\exp(\boldsymbol{z}_i^\top\hat{\boldsymbol{\alpha}})$.
- White estimator: $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathrm{diag}(\hat{\epsilon}_1^2, \ldots, \hat{\epsilon}_n^2)\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ is consistent under general conditions. No assumption about type of heteroscedasticity but larger variances.
- Two-stage estimation (Prais-Winsten estimator) for $AR(1)$ model: $\hat{\rho} = \frac{\sum_{i=2}^n\hat{\epsilon}_i\hat{\epsilon}_{i-1}}{\sqrt{\sum_{i=2}^n\hat{\epsilon}_i^2}\sqrt{\sum_{i=2}^n\hat{\epsilon}_{i-1}^2}}$. Then insert $\hat{\rho}$ to get $\hat{\boldsymbol{W}}$.

Instrumental variables

- An instrumental variable z has to satify, relevance $\mathrm{Cov}(x, z) \neq 0$ and exogeneity $\epsilon \perp z$.
- Moment-based estimator $\hat{\beta}^{IV} = \frac{\widehat{\mathrm{Cov}}(z,y)}{\widehat{\mathrm{Cov}}(z,x)}$ is consistent.
- Two-stage least squares (2SLS): Regress x on z, construct estimate of x without influence of unobserved confounder: $\tilde{x}$, regress y on $\tilde{x}$ to obtain $\hat{\beta}^{IV-2SLS}$.
- A weak instrument z can lead to large estimation variance.

Penalised regression

Ridge regression

- Ridge regression shrinks directions in the column space of $\boldsymbol{X}$ according to how $\boldsymbol{X}$ varies along those directions
- Effective degrees of freedom $df(\lambda) = \mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top) = \sum_{j=1}^{k} \frac{d_j^2}{d_j^2+\lambda}$.
- Ridge estimator is biased but has smaller variance than OLS.

Lasso regression

- Small coefficients will be more strongly shrunken towards zero while larger coefficients will be less affected by the penalty.
- Lasso estimator is biased but has smaller variance than OLS.

Explicit solutions for design matrix with orthonormal columns, i.e., $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}$. Let $\hat{\beta}_j$ be the OLS solution, then

Estimator	Formula
Best subset	$\hat{\beta}_j 1(\hat{\beta}_j > \sqrt{\lambda})$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\mathrm{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda/2)_+$

Robust regression

- In case of heavy-tailed error distributions robust estimators have smaller variance than OLS.
- Finite sample breakdown point of an estimator: Fraction of data that can be given arbitrary values without making the estimator arbitrarily bad. 0 for mean, $(n-1)/2n$ for median.
- Huber's loss function $\rho_{H_c}(\epsilon) = \begin{cases} \frac{1}{2}\epsilon^2, & \text{if } |\epsilon| \le c \\ c(|\epsilon| - \frac{c}{2})|, & \text{if } |\epsilon| > c \end{cases}$, using $c = 1.345S \approx 2MAD$ yields 95% efficiency relative to the sample mean when the distribution is normal + resistance to outliers.
- Biweight loss function $\rho_{BW_c}(\epsilon) = \begin{cases} \frac{c^2}{6}\left\{1 - \left[1 - \left(\frac{\epsilon}{c}\right)^2\right]^3\right\}, & \text{if } |\epsilon| \le c \\ \frac{c^2}{6}, & \text{if } |\epsilon| > c \end{cases}$, using $c = 4.685S \approx 7MAD$ yields 95% efficiency relative to the sample mean when the distribution is normal.
- Using the weight function $w(\epsilon) = \psi(\epsilon)/\epsilon$.
- Bounded-influence regression: Least-trimmed-squares (LTS) regression selects regression coefficients to minimize smaller half of the squared residuals, least median of squares.

Generalised linear models

Logistic regression

- Fisher scoring algorithm $\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + (\boldsymbol{X}\boldsymbol{V}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \hat{\boldsymbol{\pi}}^{(t)})$, where $\boldsymbol{V}^{(t)} = \mathrm{diag}(\hat{\pi}_1^{(t)}(1-\hat{\pi}_1^{(t)}), \ldots, \hat{\pi}_n^{(t)}(1-\hat{\pi}_n^{(t)}))$.
- Asymptotic distribution under (weak) regularity conditions $\hat{\boldsymbol{\beta}} \xrightarrow{d} \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}}))$
- Under $H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d}$ with $\mathrm{rank}(\boldsymbol{C}) = r$, Wald statistic $w = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})^\top(\boldsymbol{C}\boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{C})^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}) \sim \chi_r^2$.
- Deviance $D(\hat{\boldsymbol{\pi}}) = -2l(\hat{\boldsymbol{\pi}})$ for individual data model, and $D(\hat{\boldsymbol{\pi}}) = 2(l(\tilde{\boldsymbol{\pi}}) - l(\hat{\boldsymbol{\pi}})) \xrightarrow{d} \chi^2_{G-p}$ for group data model, if there is no overdispersion, and if approximation is based on a limiting operation where G is fixed and $n_i \to \infty$.
- Deviance residual $r_{D_i} = \mathrm{sign}(\hat{y}_i - \hat{\pi}_i)\sqrt{d_i}$, $\sum_i r_{D_i}^2 = D$, Pearson residual $r_{P_i} = \frac{\hat{y}_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)/n_i}}$, based on $\chi^2 = \sum_i r_{P_i}^2$.
- This empirical variance $\bar{y}_i(1-\bar{y}_i)/n_i$ is often much larger than $\hat{\pi}_i(1-\hat{\pi}_i)$. If $\mathrm{Var}(\bar{y}_i) = \phi\pi_i(1-\pi_i)$, one estimates $\hat{\phi}_P = \frac{1}{n-p}\chi^2$ or $\hat{\phi}_D = \frac{1}{n-p}D$. Then $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = \hat{\phi}\boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$.
- Metrics: accuracy $\frac{\#TP+\#TN}{n}$, true positive rate / recall / sensitivity $TPR = \frac{\#TP}{\#TP+\#FN}$, specificity $\frac{\#TN}{\#TN+\#FP}$, false positive rate $FPR = \frac{\#FP}{\#TN+\#FP}$

General settings

- Linear predictor $\eta = \boldsymbol{x}^\top\boldsymbol{\beta}$, response function $\mu = h(\eta)$, link function $\eta = g(\mu)$.
- Distribution of response $Y|\theta$ in exponential family $f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{\phi}w + c(y, \phi, w)\right)$.
- For individual data, $w = 1$. For group data when y is group mean, $w = n_i$; when y is group total, set $w = 1/n_i$.
- $E(y) = \mu = b'(\theta)$, $\mathrm{Var}(y) = \phi b''(\theta)/w$.
- Unique canonical link function specified by $\eta = \theta$. Then log-likelihood is always concave so that the ML estimator is always unique (if it exists), and $\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{H}(\boldsymbol{\beta})$.

Cumulative model

- Let $u_i = -\boldsymbol{x}_i^\top\boldsymbol{\beta} + \epsilon_i$, where ϵ_i had cdf F. Then $Y_i = r \equiv \theta_{r-1} < u_i \le \theta_r$ for $r = 1, 2, \ldots, c+1$, where $-\infty = \theta_0 < \theta_1 < \cdots < \theta_{c+1} = \infty$.
- Logistic model is called proportional odds model because ratio of the cumulative odds for subpopulations characterized by $\boldsymbol{x}_i$ and $\tilde{\boldsymbol{x}}_i$ is $\frac{P(Y_i \le r|\boldsymbol{x}_i)/P(Y_i > r|\boldsymbol{x}_i)}{P(Y_i \le r|\tilde{\boldsymbol{x}}_i)/P(Y_i > r|\tilde{\boldsymbol{x}}_i)} = \exp((\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i)^\top\boldsymbol{\beta})$

Nonparametric regression

Polynomial splines

- A function $f : [a, b] \to \mathbb{R}$ is called a polynomial spline of degree $l \ge 0$ with knots $a = \kappa_1 < \cdots < \kappa_m = b$ if $f(z)$ is $(l-1)$-times continuously differentiable, and $f(z)$ is a polynomial of degree l on the intervals $[\kappa_j, \kappa_{j+1})$.
- Truncated power series uses the model $y_i = \gamma_1 + \gamma_2 z_i + \cdots + \gamma_{l+1}z_i^l + \gamma_{l+2}(z_i - \kappa_2)_+^l + \cdots + \gamma_{l+m-1}(z_i - \kappa_{m-1})_+^l + \epsilon_i$. The functions $B_1, \ldots, B_d$ are called basis functions. Numerical instabilities for covariates with large values. Basis functions are nearly collinear, especially when two knots are close to one another.
- B-splines: only positive on an interval based on $l+2$ adjacent knots, at any point $l+1$ basis functions are positive. Bounded from above, hence numerically more stable.
- P-splines: introduce an additional penalty term that prevents overfitting, minimise $PLS(\lambda) = \sum_{i=1}^n \left(y_i - \sum_{j=1}^d \gamma_j B_j(z_i)\right)^2 + \lambda\sum_{j=l+2}^d \gamma_j^2$ for TP basis.
- Smoothing splines: only assume that $f(z)$ is twice continuously differentiable, so that we can use $\sum_{i=1}^n(y_i - f(z_i))^2 + \lambda\int(f''(z))^2 dz$. Natural cubic splines with knots at the d ordered and unique covariate values $z_{(1)} < \cdots < z_{(d)}$.
- The function $f(z)$ is a natural cubic spline based on the knots $a \le \kappa_1 < \cdots < \kappa_m \le b$ if $f(z)$ is a cubic polynomial spline for the given knots, and satisfy boundry conditions $f''(a) = f''(b) = 0$, linear in the intervals $[a, \kappa_2]$ and $[\kappa_{m-1}, b]$. Need m instead of $m+2$ basis functions compared to standard cubic polynomial splines.

Local polynomial regression

- Estimation is usually based on a weighted version of the residual sum of squares $\sum_{i=0}^n \left(y_i - \sum_{j=0}^l \gamma_j(z_i - z)^j\right)^2 w_\lambda(z, z_i)$, where the weights are based on kernel functions $K\left(\frac{z - z_i}{\lambda}\right)$.
- Uniform kernel $K(u) = \frac{1}{2}1(|u| \le 1)$, Epanechnikov kernel $K(u) = \frac{3}{4}(1-u^2)1(|u| \le 1)$, Gaussian kernel $K(u) = \phi(u)$.
- Nadaraya-Watson estimator: local constant polynomial model $\hat{f}(z) = \frac{\sum_i w_\lambda(z, z_i)y_i}{\sum_i w_\lambda(z, z_i)}$.
- LOESS/LOWESS: weights $w_{\Delta(z)}(z, z_i) = K\left(\frac{|z - z_i|}{\Delta(z)}\right)$ where $\Delta(z) = \max_{i,j \in N(z)}|z_i - z_j|$ for k-nearest neighbourhood $N(z)$ of z.
- If $\hat{f}(z) = \boldsymbol{s}(z)^\top\boldsymbol{y}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$, then $\hat{f}(x) - f(z) \xrightarrow{d} \mathcal{N}(0, \sigma^2\boldsymbol{s}(z)^\top\boldsymbol{s}(z))$, assuming that $\hat{f}(z)$ is (approximately) unbiased.
- Choose smoothing parameters by leave-one-out cross validation $LOOCV = \frac{1}{n}\sum_{i=1}^n\left(\frac{y_i - \hat{f}(z)}{1 - \boldsymbol{S}_{ii}}\right)^2$, exact for classical linear model and smoothing splines.
- Assume additive structure $f(\boldsymbol{z}) = f_1(z_1) + \cdots + f_q(z_q)$ to protect against curse of dimensionality.