Computational Statistics

Daniel Balle 2018

Multiple Linear Regression

In multiple regression we are given data $(Y_i, \mathbf{x_i})$ where we assume the response variable is a linear function of the predictors:

$$Y_i = \boldsymbol{\beta}^\mathsf{T} \mathbf{x_i} + \epsilon_i$$

or $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{Y} \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$. Errors ϵ_i are usually assumed *iid* with $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$.

Our goal is to estimate the unknown parameters $\boldsymbol{\beta} \in \mathbb{R}^p$.

Least Squares Estimator

The least squares estimator for a *linear* model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is given by minimizing the least squares error:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - X\boldsymbol{\beta}\|^2$$

Remark. This is solved theoretically by $\hat{\boldsymbol{\beta}} = (X^\intercal X)^{-1} X^\intercal \boldsymbol{Y}$. This estimator is *unbiased*, i.e. $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ and $\mathbb{E}[\hat{\boldsymbol{Y}}] = X\boldsymbol{\beta}$.

Least Squares Residuals

With the *least squares* estimator $\hat{\boldsymbol{\beta}}$ the residuals $r_i = Y_i - \hat{\boldsymbol{\beta}}^\mathsf{T} \mathbf{x}_i$ give an *unbiased* estimate of the errors ϵ_i, $\mathbb{E}[\boldsymbol{r}] = \mathbf{0}$. And

$$\hat{\sigma}^2 = (n-p)^{-1} \sum r_i^2$$

provides an *unbiased* estimate of the variance, i.e. $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$.

Least Squares Projection

Geometrically the *least squares* method performs an orthogonal projection $\boldsymbol{Y} \mapsto \hat{\boldsymbol{Y}} = X\hat{\boldsymbol{\beta}}$ with projection matrix:

$$\hat{\boldsymbol{Y}} = P\boldsymbol{Y} \implies P = X(X^\intercal X)^{-1}X^\intercal$$

Remark. The residuals can then be expressed as $\mathbf{r} = (I - P)\boldsymbol{Y}$.

Linear Regression in R

In `R` a *linear regression* model can be fitted as follows:

```
fitted <- lm(formula = LOGRUT ~ ., data = asphalt1)
```

Bias-Variance Tradeoff

The bias-variance decomposition for any *supervised* learning task of $y = f(x) + \epsilon$ is the expected generalization error:

$$\underbrace{\mathbb{E}[(f(x) - \hat{f}(x))^2]}_{\text{MSE}(x)} = (\underbrace{\mathbb{E}[\hat{f}(x)] - f(x)}_{\text{Bias}})^2 + \underbrace{\mathbb{E}[\hat{f}(x)^2] - E[\hat{f}(x)]^2}_{\text{Var}(\hat{f}(x))}$$

Remark. Optimizing this trade-off is called *regularization*, and avoids the problem of *overfitting*.

Kernel Density Estimation

Given realizations $X_i \in \mathbb{R} \sim F$, the nonparametric kernel density estimator $\hat{f}$ of the unknown density function $f = F'$ is

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

where $K(\cdot)$ is a *kernel* function, usually symmetric around 0, and the tuning parametrer h the *bandwidth*.

$$K(x) = K(-x) \quad K(x) \geq 0 \quad \int_{-\infty}^{\infty} K(x)dx = 1$$