Bootstrap

- R live coding and debugging
- Data resampling
- Bias and variance estimation for estimators
- L08: More coding; parametric bootstrap; randomised tests

Cross validation and Bootstrap

- Cross validation splits the data in K parts and performs model estimation on $(K-1)$ parts and validation assessment on the Kth part, all for each of the parts.
- Bootstrap resamples *with replacement* to obtain a random sample of the same size as the original sample.

Basic Bootstrap resampling

Let $Y = \{(y_i, x_i), i = 1, \ldots, N\}$ be a data collection with response values y_i and predictors/covariates x_i.

- Define a *Bootstrap sample* $Y^{(j)}$ by drawing N pairs (y_i, x_i) from Y with equal probability, and with replacement.
- Repeat this procedure for $j = 1, \ldots, J$, with $J \gg 1$.

The resampling procedure draws a random sample from the *empirical distribution* for the data collection.

The Bootstrap principle

- Each boostrap sample $Y^{(j)}$ can be used to apply some model estimation procedure, each generating a parameter estimate $\widehat{\theta}^{(j)}$.
- We want to use these bootstrapped estimates to say something about the properties of the estimator $\widehat{\theta}$ which is based on the original data Y.
- Idea: The parameter estimate as a deterministic function of the data, the *empirical parameter value* for the observed sample Y: $\widehat{\theta} = \theta(Y)$ and $\widehat{\theta}^{(j)} = \theta(Y^{(j)})$.

The Bootstrap principle

According to the *Bootstrap principle*, the errors of the bootstrapped estimates have the same distribution as the error of $\widehat{\theta}$. In particular, if the true parameter is θ_{true}, then

$$\begin{aligned} \mathsf{E}(\widehat{\theta} - \theta_{\text{true}}) &= \mathsf{E}(\widehat{\theta}^{(j)} - \widehat{\theta}), \\ \mathsf{Var}(\widehat{\theta} - \theta_{\text{true}}) &= \mathsf{Var}(\widehat{\theta}^{(j)} - \widehat{\theta}). \end{aligned}$$

Bootstrap estimation

- The usual expectation and variance estimators can be used:

$$\widehat{\mathsf{E}}(\widehat{\theta} - \theta_{\text{true}}) = \frac{1}{J}\sum_{j=1}^{J}(\widehat{\theta}^{(j)} - \widehat{\theta}) = \overline{\widehat{\theta}^{(\cdot)}} - \widehat{\theta}, \quad \widehat{\mathsf{Var}}(\widehat{\theta} - \theta_{\text{true}}) = \frac{1}{J-1}\sum_{j=1}^{J}\left(\widehat{\theta}^{(j)} - \overline{\widehat{\theta}^{(\cdot)}}\right)^2.$$

- Bias adjusted etimator $\widehat{\theta} - \left(\overline{\widehat{\theta}^{(\cdot)}} - \widehat{\theta}\right)$. Properties? Need *double bootstrap*!
- Confidence intervals for θ_{true}: Consider the quantiles of the error distribution.

Find a&b s.t. $P(a < \hat{\theta}^{(j)} - \hat{\theta} < b) = 1-\alpha$ (Empirical quantiles)

Bootstrap principle: $P(a < \hat{\theta} - \theta_{\text{true}} < b) = 1-\alpha$

$$\Rightarrow P(\hat{\theta} - b < \theta_{\text{true}} < \hat{\theta} - a) = 1-\alpha$$

$$CI_{\text{boot}} = (\hat{\theta} - b, \hat{\theta} - a)$$