The Jackknife

…is a more specific resampling plan based on a leave-one-out idea:

- Let $\hat{\theta}_n : \mathcal{X}_0^n \to \mathbb{R}$ be an estimator.
- For $x = (x_1, \ldots, x_n)' \in \mathcal{X}_0^n$ and $i \in \{1, \ldots, n\}$, let

$$\hat{\theta}_{(i)}(x) := \hat{\theta}_{n-1}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

  be the estimator computed without the i-th observation.
- Write $\hat{\theta}_{(\cdot)}(x) := \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{(i)}(x)$.
- The Jackknife estimate of the squared standard error (=variance) of $\hat{\theta}_n$ is given by:

$$\boxed{\hat{se}^2(x) := \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\theta}_{(i)}(x) - \hat{\theta}_{(\cdot)}(x) \right)^2 .}$$

$$\hat{se}^2 := \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)} \right)^2 .$$

This does not look quite right! Why is it not a sample variance?

Consider the case $\hat{\theta}_n(x) = \frac{1}{n} \sum_{i=1}^{n} x_i$. Then

$$\hat{\theta}_{(i)} = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} x_j = \frac{1}{n-1} \left(n \bar{X}_n - x_i \right)$$

$$\hat{\theta}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{(i)} = \frac{1}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} \left(n \bar{X}_n - x_i \right)$$

$$= \frac{1}{n-1} \left(n \bar{X}_n - \bar{X}_n \right) = \bar{X}_n$$

universität wien

$$\hat{se}^2 := \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat\theta_{(i)} - \hat\theta_{(\cdot)}\right)^2.$$

$$\hat{se}^2 = \frac{n-1}{n}\sum_{i=1}^{n}\Big(\underbrace{\frac{1}{n-1}(n\bar X_n - X_i) - \bar X_n}_{\circledast}\Big)^2$$

$$\circledast = \frac{n}{n-1}\bar X_n - \bar X_n - \frac{1}{n-1}X_i = \frac{1}{n-1}(\bar X_n - X_i)$$

$$\hat{se}^2 = \frac{\cancel{n-1}}{n}\underbrace{\frac{1}{(n-1)^2}\sum_{i=1}^{n}(X_i - \bar X_n)^2}_{\hat\sigma_n^2} = \frac{\hat\sigma_n^2}{n}$$

$$\mathbb{E}_\theta\left[\hat{se}^2\right] = se_\theta^2$$

Recall HW 2.1a: In the iid model $se_\theta^2 := \mathrm{Var}_\theta[\hat\theta_n] = \frac{\sigma^2}{n}$ and $\mathbb{E}_\theta[\hat\sigma_n^2] = \sigma^2$.

Unfortunately, however, the Jackknife does not always produce good estimates for the standard error!

e.g., sample quantiles HW

For these kinds of parametric problems, the Jackknife idea is kind of outdated.

However, ...

- We observe iid pairs $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$ from (marginal) sample space $\mathcal{X}_0 = \mathbb{R}^p \times \mathbb{R}$.
- Let (X_0, Y_0) be another independent pair with identical distribution (prediction period).
- We observe X_0 but not Y_0. Want to predict the value of Y_0.
- Use a predictor/learning algorithm $\hat{m}_n : \mathbb{R}^p \to \mathbb{R}$ to predict the value of Y_0 by $\hat{m}_n(X_0)$.
- Actually $\hat{m}_n$ depends also on the training data! So $\hat{m}_n : \mathcal{X}_0^n \times \mathbb{R}^p \to \mathbb{R}$, $\hat{m}_n(X_0) = \hat{m}_n(Z_1, \ldots, Z_n; X_0)$.
- For example:
    - $\hat{m}_n(x) = x'\hat{\beta}_n$ with $\hat{\beta}_n = (X'X + \lambda I_p)^{-1} X'Y$, $X = [X_1, \ldots, X_n]'$, $Y = (Y_1, \ldots, Y_n)'$
    - $\hat{m}_n$ is a CNN with weights obtained from SGD.

- We would like to quantify the uncertainty associated with predicting the new label/response Y_0.
- Prediction interval: $PI_\alpha \subseteq \mathbb{R}$

$$\boxed{P(Y_0 \in PI_\alpha) \geq 1 - \alpha.}$$

- Would like to know the distribution of the prediction error

$$P\left(q_{\frac{\alpha}{2}} \leq Y_0 - \hat{m}_n(X_0) \leq q_{1-\frac{\alpha}{2}} \right) = 1-\alpha$$

- Could use theoretical quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ to construct

$$\boxed{PI_\alpha = [\hat{m}_n(X_0) + q_{\alpha/2}, \hat{m}_n(X_0) + q_{1-\alpha/2}].}$$

PREDICTIVE INFERENCE BY SAMPLE SPLITTING

- How to estimate/approximate the distribution of the prediction error

$$Y_0 - \hat{m}_n(Z_1, \ldots, Z_n; X_0)$$

- Traditional approach: Split the sample into $S_{train} \cup S_{val} = \{1, \ldots, n\}$, $S_{train} \cap S_{val} = \emptyset$, $n_1 = |S_{train}|$, $n_2 = |S_{val}|$, $n_1 + n_2 = n$.
- Train your algorithm on S_{train} and validate it on S_{val}, i.e., compute

n_2

$$\boxed{R_j^{ss} := Y_j - \hat{m}_{n_1}(\{Z_i : i \in S_{train}\}; X_j), \quad j \in S_{val}.}$$

- Use empirical quantiles $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$ of R_j^{ss}, $j \in S_{val}$ and compute

$$PI_\alpha = [\hat{m}_{n_1}(X_0) + \hat{q}_{\alpha/2}, \hat{m}_{n_1}(X_0) + \hat{q}_{1-\alpha/2}].$$

- Conditional on the data in S_{train}, the residuals

$$R_j^{ss} := Y_j - \hat{m}_{n_1}(\{Z_i : i \in S_{train}\}; X_j); \quad j \in S_{val}.$$

are an iid sample with distribution equal to that of

$$R^{ss} := Y_0 - \hat{m}_{n_1}(\{Z_i : i \in S_{train}\}, X_0).$$

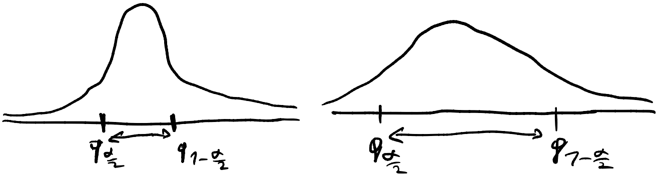
- Thus, $\hat{q}_\alpha \xrightarrow{p.} q_\alpha^{(R^{ss})}$ as $\boxed{n_2 \to \infty.}$

$$Y_0 \in PI_\alpha = [\hat{m}_{n_1}(X_0) + \hat{q}_{\alpha/2}, \hat{m}_{n_1}(X_0) + \hat{q}_{1-\alpha/2}]$$
$$\iff \hat{q}_{\alpha/2} \leq Y_0 - \hat{m}_{n_1}(X_0) \leq \hat{q}_{1-\alpha/2}$$

$$P(Y_0 \in PI_\alpha | S_{train}) = P(\hat{q}_{\alpha/2} \leq R^{ss} \leq \hat{q}_{1-\alpha/2} | S_{train}) \approx 1-\alpha$$

- Sample splitting works very well when n is large relative to p.
- Otherwise, $\hat{m}_{n_1} : \mathbb{R}^p \to \mathbb{R}$ may be much less accurate than $\hat{m}_n$.
- Recall: We need n_2 large, so $n_1 = n - n_2 \ll n$.

$$Y_0 - \hat{m}_n(X_0) \quad \text{vs.} \quad Y_0 - \hat{m}_{n_1}(X_0)$$



Predictive inference with the jackknife

- How to estimate/approximate the distribution of the prediction error

$$R := Y_0 - \boxed{\hat{m}_n}(X_0)$$

- Let $R_i^{l1o} := Y_i - \boxed{\hat{m}_{(i)}}(X_i)$, $i = 1, \ldots, n$ where

$$\hat{m}_{(i)}(X_i) = \hat{m}_{n-1}(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n; X_i)$$

  is the prediction at X_i of the learning algorithm trained on the data set with the i-th observation pair removed.
- If $\hat{m}_n \approx \hat{m}_{(i)}$, then, approximately $R_i^{l1o} \approx R$.
- The $R_1^{l1o}, \ldots, R_n^{l1o}$ are (usually) identically distributed but not independent.
- We still use empirical quantiles $\hat{q}_{\alpha/2}^{l1o}$ and $\hat{q}_{1-\alpha/2}^{l1o}$ to compute…

universität wien

- $R_i^{l1o} := Y_i - \hat{m}_{(i)}(X_i)$, $i = 1, \dots, n$
- $\hat{q}_{\alpha/2}^{l1o}$ and $\hat{q}_{1-\alpha/2}^{l1o}$ empirical quantiles.

$$\boxed{PI_\alpha^{l1o} = [\hat{m}_n(X_0) + \hat{q}_{\alpha/2}^{l1o}, \hat{m}_n(X_0) + \hat{q}_{1-\alpha/2}^{l1o}]}$$

Under some regularity assumptions, one can show

$$\mathbb{E}\left[\left|P(Y_0 \in PI_\alpha^{l1o} | Z_1, \dots, Z_n) - (1-\alpha)\right|\right] \xrightarrow[n,p\to\infty]{} 0.$$

$$P(Y_0 \in PI_\alpha^{l1o}) \approx 1-\alpha$$

$$\left| P\left(Y_0 \in PI_\alpha^{(1,0)}\right) - (1-\alpha) \right|$$

$$= \left| \mathbb{E}\left[P\left(Y_0 \in PI_\alpha^{(1,0)} \mid Z_1 \ldots Z_n \right)\right] - (1-\alpha) \right|$$

$$= \left| \mathbb{E}\left[P\left(Y_0 \in PI_\alpha^{(1,0)} \mid Z_1 \ldots Z_n \right) - (1-\alpha) \right] \right|$$

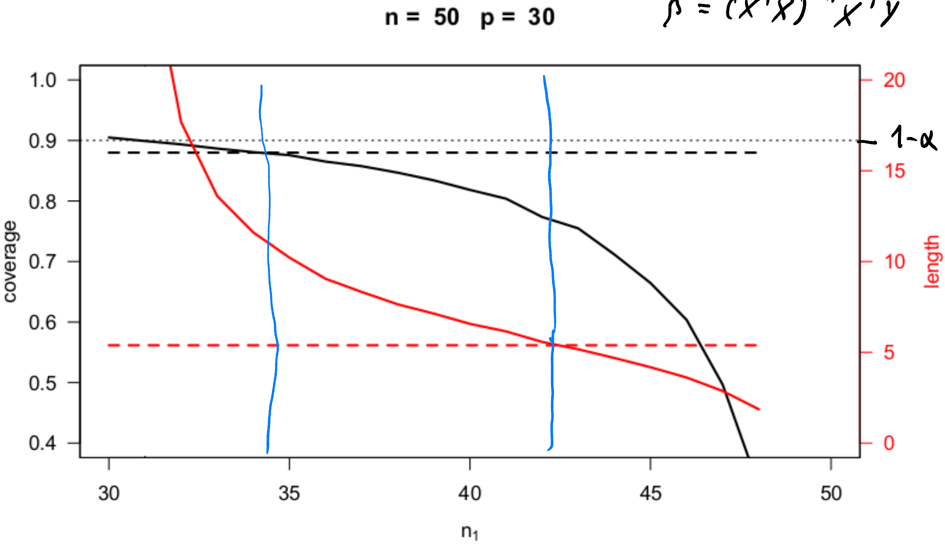
$$\leq \mathbb{E}\left[\left| P\left(Y_0 \in PI_\alpha^{(1,0)} \mid Z_1 \ldots Z_n \right) - (1-\alpha) \right| \right]$$

$$\xrightarrow[n,p \to \infty]{} 0$$

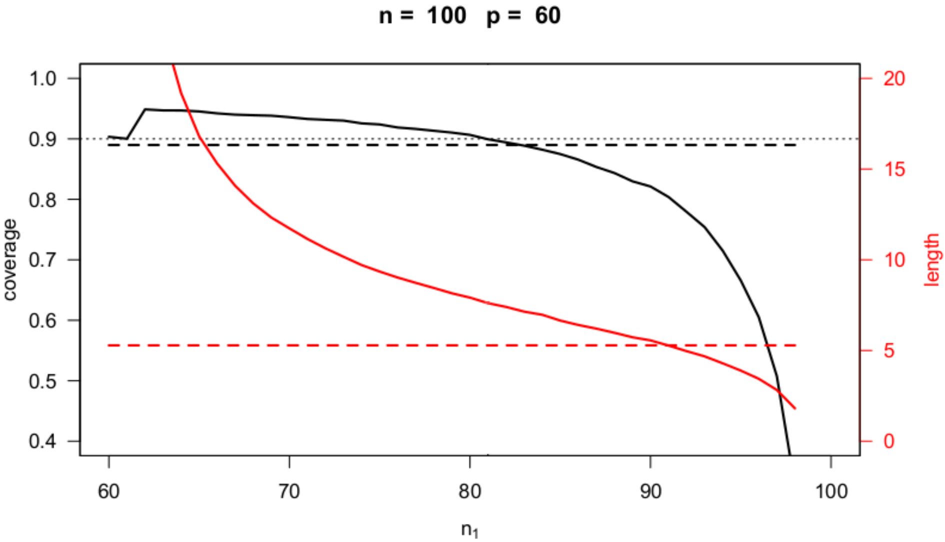
PREDICTIVE INFERENCE: SAMPLE SPLITTING VS. JACKKNIFE

$\hat{m}_1(x) = x^T \hat{\beta}$

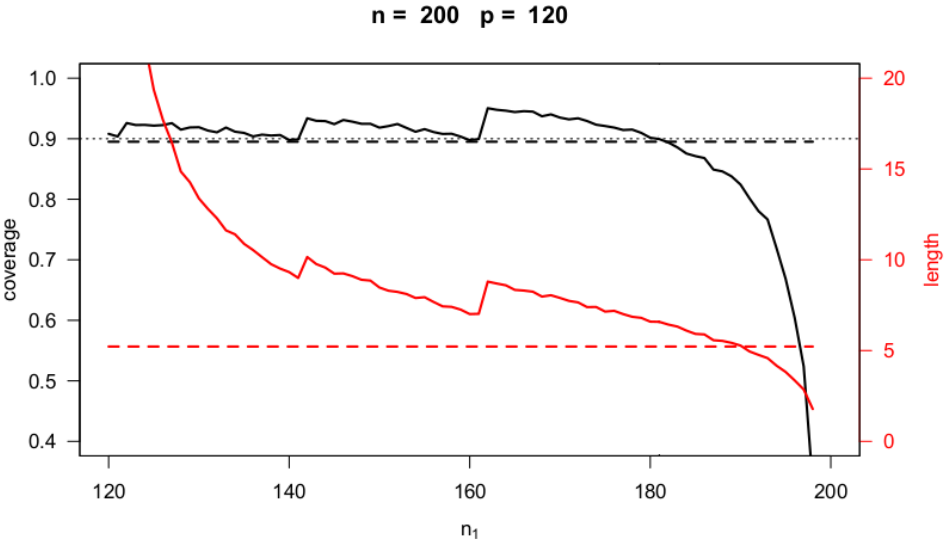
$\hat{\beta} = (X^T X)^{-1} X^T y$



PREDICTIVE INFERENCE: SAMPLE SPLITTING VS. JACKKNIFE



Predictive inference: sample splitting vs. jackknife



Predictive inference with the jackknife

Why use leave-one-out residuals

$$Y_i - \hat{m}_{(i)}(X_i)$$

and not simply

$$Y_i - \hat{m}_n(X_i)?$$

$\overset{?}{\approx}$

z_i

Does it make a big difference?

HW

Would be computationally much cheaper!!!