# Module 5: Solutions to Recommended Exercises

TMA4268 Statistical Learning V2023

Kenneth Aase, Emma Skarstein, Daesoo Lee, Stefanie Muff
Department of Mathematical Sciences, NTNU

February 9, 2023

## Problem 1

a) See Figure 1

| | | | | |
|---|---|---|---|---|
| validate | train | train | train | train |
| train | validate | train | train | train |
| train | train | validate | train | train |
| train | train | train | validate | train |
| train | train | train | train | validate |

b) In run 1, fold 1 is kept aside and folds 2 to $k$ are used to train the method. The error is calculated on fold 1 ($\text{MSE}_1$). We do this for the remaining $k-1$ runs. The cross-validation error is given by

$$CV_k = \frac{1}{n}\sum_{j=1}^{k} n_j \cdot \text{MSE}_j \ ,$$

where $n_j$ is the size of the $j^{\text{th}}$ fold.

For a regression problem $\text{MSE}_j = \frac{1}{n_j}\sum_{i\in V_j}(y_i - \hat{y}_i)^2$ for all $i$ in the validation set $V_j$. For a binary classification problem the error might be the average of correctly classified observations.

c) Find the optimal number of neighbors $K$ in $K$NN or the optimal polynomial degree in polynomial regression.

d) For a classification problem we can use CV to choose between QDA or LDA.

## Problem 2

Advantages and disadvantages of $k$-fold Cross-Validation

a) The validation set approach:

D: The $k$-fold cross validation is computationally more expensive.

A: The advantage is that it has less variance and less bias (the validation set approach unifies the worst of both worlds, so to say).

b) In LOOCV there is no randomness in the splits.

A: $k$-fold-CV is computationally less expensive, as we run the model $k$ times where $k \ll n$

A: $k$-fold-CV has less variance, because in LOOCV we are averaging from $n$ fitted models that are trained on nearly the same data, therefore we have positively correlated data.

D: $k$-fold-CV has more bias, as in LOOCV we use a larger data set to fit the model, which gives a less biased version of the test error.

c) We know that if $k = n =$ LOOCV the estimator of test error will have small bias but high variance and it is computationally expensive.

If $k$ is too small (for example 2), the estimator will have larger bias but lower variance.

Experimental research (simulations) has found that $k = 5$ or $k = 10$ to be good choices.

## Problem 3

No solution is provided on top of the guidelines in the exercise sheet.

## Problem 4

a) $P(\text{draw } X_i) = \frac{1}{n}$ and $P(\text{not draw } X_i) = 1 - P(\text{draw } X_i) = 1 - \frac{1}{n}$

b) $P(\text{not draw any } X_i) = (1 - \frac{1}{n})^n$ and $P(\text{draw at least one } X_i) = 1 - (1 - \frac{1}{n})^n$

c) $P(X_i \text{ in bootstrap sample}) = 1 - (1 - \frac{1}{n})^n \approx 1 - \exp(-1) = 0.632$

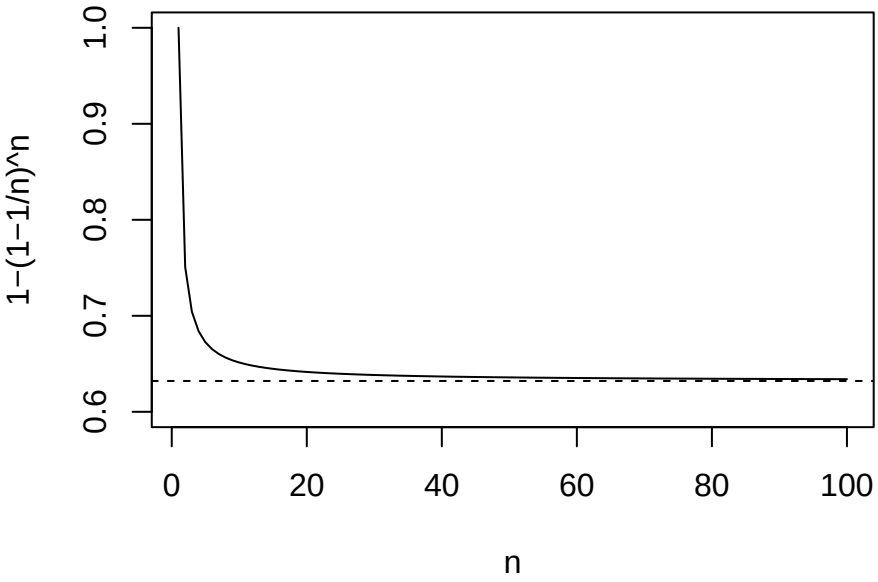
d)

```
n <- 100
B <- 10000
j <- 1
res <- rep(NA, B)
for (b in 1:B) res[b] <- (sum(sample(1:n, replace = TRUE) == j) > 0)
mean(res)
```

```
## [1] 0.6466
```

The approximation becomes quickly very good as $n$ increases, as the following graph shows:

```
curve(expr = 1 - (1 - 1/x)^x, from = 1, to = 100, ylim = c(0.6, 1), xlab = "n", ylab = "1-(1-1/n)^n")
abline(h = 1 - 1/exp(1), lty = 2)
```




## Problem 5

We repeat the following for $b = 1, \ldots, B$:

- Draw with replacement a bootstrap sample.
- Fit the model.
- Store $\hat{\beta}_b$.

Calculate $\hat{SD}(\hat{\beta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\beta}_b - \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b)^2}$.

For the 95% CI, we can calculate the 0.025 and 0.975 quantiles of the sample $\hat{\beta}_b$, $b = 1, \ldots, B$.

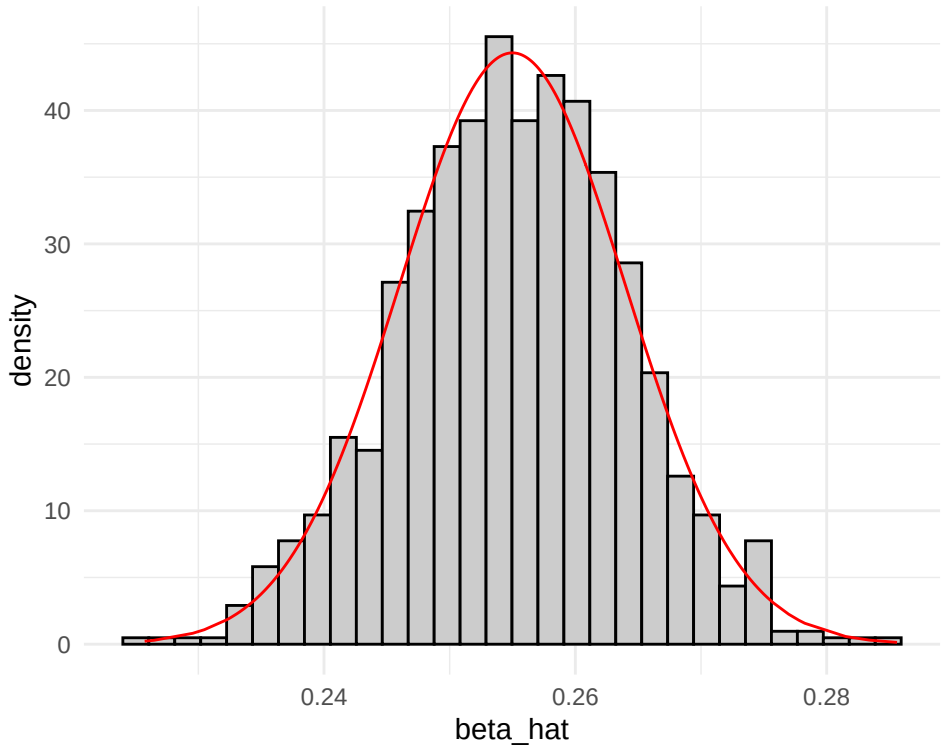
## Problem 6

```
library(carData)
boot.fn <- function(data, index) {
    return(coef(lm(wages ~ ., data = SLID, subset = index)))
}

beta_hat <- c()
B <- 1000
for (i in 1:B) {
    beta_hat[i] <- boot.fn(data = SLID, index = sample(nrow(SLID), nrow(SLID), replace = TRUE))["age"]
}
```

We can for example look at the histogram of the samples of $\hat{\beta}$ to get an idea of the distribution:

```
library(ggplot2)
data <- data.frame(beta_hat = beta_hat, norm_den = dnorm(beta_hat, mean(beta_hat),
    sd(beta_hat)))
ggplot(data) + geom_histogram(aes(x = beta_hat, y = ..density..), fill = "grey80",
    color = "black") + geom_line(aes(x = beta_hat, y = norm_den), color = "red") +
    theme_minimal()
```




The 95% CI for $\hat{\beta}_{age}$ can now be derived by either using the 2.5% and 97.5% quantiles of the samples, or by using the $\hat{\beta} \pm 1.96 \cdot \text{SD}(\hat{\beta})$ idea:

```
sd(beta_hat)
```

```
## [1] 0.009002064
```

```
quantile(beta_hat, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.2371012 0.2727316
```

```
c(mean(beta_hat) - 1.96 * sd(beta_hat), mean(beta_hat) + 1.96 * sd(beta_hat))
```

```
## [1] 0.2373471 0.2726352
```

We can compare these results to what we would obtain directly from fitting the model and calculating the confidence interval

```
SLID.lm <- lm(wages ~ ., data = SLID)
confint(SLID.lm)
```

```
##                     2.5 %     97.5 %
## (Intercept)    -9.0891576 -6.6884008
## education       0.8484610  0.9847670
## age             0.2380522  0.2722214
## sexMale         3.0452719  3.8655493
## languageFrench -0.8518577  0.8214111
## languageOther  -0.4946904  0.7798996
```

As expected, the 95% CI for `age` is essentially the same as the one we obtained from bootstrapping.

If you prefer to use the built in function `boot()`

```r
library(boot)
bo <- boot(SLID, boot.fn, 1000)
bo
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = SLID, statistic = boot.fn, R = 1000)
##
##
## Bootstrap Statistics :
##        original        bias    std. error
## t1* -7.88877921 -2.149479e-02 0.662378711
## t2*  0.91661398  9.858572e-04 0.039006093
## t3*  0.25513680  1.285178e-05 0.009080922
## t4*  3.45541061  7.024093e-03 0.211985641
## t5* -0.01522333  8.711755e-03 0.440081970
## t6*  0.14260463 -2.880760e-03 0.321504593
```

---

# Summing up

## Take home messages

- Use $k = 5$ or 10 fold cross-validation for model selection or assessment.
- Use bootstrapping to estimate the standard deviation of an estimator, and understand how it is performed before module 8 on trees.

# Further reading

- Videoes on YouTube by the authors of ISL, Chapter 5, and corresponding slides
- Solutions to exercises in the book, chapter 5