Übung 2

Computational Statistics

Sommersemester 2019

April 15, 2019

J. Groos (FBMN, h_da)

Name:

A 1

```
set.seed(42)
load(file = "Donald.RData")  # importiere Datensatz

## Vergleich der Variablenverteilung zwischen Trainings- und
## Testdatensatz/Lineare Regression (Drei Mal)
for (i in 1:3) {
    smp_size <- 100
    train_ind <- sample(seq_len(nrow(Donald_1)), size = smp_size)

    train <- Donald_1[train_ind, ]
    test <- Donald_1[-train_ind, ]
    print(i)
    cat("train: \n\n")
    print(summary(train))
    cat("\n")
    cat("test: \n\n")
    print(summary(test))
    linreg <- lm(Trump ~ Alter + Geschlecht + Minderheit + Fremdenfeindlich +
        IQ, data = train)
    pred <- predict.lm(linreg, test)
    cat("\n")
    cat("MSE: \n")
    print(mean((test$Trump - pred)^2))
    cat("_______________________________ \n\n")
}
#> [1] 1
#> train:
#>
#>    Geschlecht       Alter         Minderheit   Fremdenfeindlich
#>  Min.   :0.00   Min.   : 7.00   Min.   :0.00   Min.   : 1.00
#>  1st Qu.:0.00   1st Qu.:40.00   1st Qu.:0.00   1st Qu.: 2.00
#>  Median :0.00   Median :50.50   Median :0.00   Median : 5.00
#>  Mean   :0.43   Mean   :49.79   Mean   :0.12   Mean   : 5.35
#>  3rd Qu.:1.00   3rd Qu.:58.50   3rd Qu.:0.00   3rd Qu.: 9.00
#>  Max.   :1.00   Max.   :92.00   Max.   :1.00   Max.   :11.00
#>        IQ            Trump
#>  Min.   : 66.0   Min.   :  0.00
#>  1st Qu.: 92.0   1st Qu.: 15.00
#>  Median :103.5   Median : 45.00
```

```
#>  Mean   :102.9   Mean   : 49.35
#>  3rd Qu.:112.0   3rd Qu.: 81.25
#>  Max.   :145.0   Max.   :100.00
#>
#> test:
#>
#>    Geschlecht       Alter          Minderheit   Fremdenfeindlich
#>  Min.   :0.00   Min.   : 23.00   Min.   :0.00   Min.   : 1.00
#>  1st Qu.:0.00   1st Qu.: 42.00   1st Qu.:0.00   1st Qu.: 3.00
#>  Median :1.00   Median : 53.00   Median :0.00   Median : 6.00
#>  Mean   :0.54   Mean   : 54.14   Mean   :0.08   Mean   : 5.82
#>  3rd Qu.:1.00   3rd Qu.: 66.50   3rd Qu.:0.00   3rd Qu.: 9.00
#>  Max.   :1.00   Max.   :102.00   Max.   :1.00   Max.   :11.00
#>        IQ             Trump
#>  Min.   : 48.00   Min.   :  0.00
#>  1st Qu.: 87.25   1st Qu.: 21.25
#>  Median : 95.00   Median : 70.00
#>  Mean   : 96.68   Mean   : 57.40
#>  3rd Qu.:106.00   3rd Qu.: 88.75
#>  Max.   :138.00   Max.   :100.00
#>
#> MSE:
#> [1] 35.98927
#> ------------------------------
#>
#> [1] 2
#> train:
#>
#>    Geschlecht       Alter          Minderheit  Fremdenfeindlich
#>  Min.   :0.00   Min.   :  7.00   Min.   :0.0   Min.   : 1.00
#>  1st Qu.:0.00   1st Qu.: 41.00   1st Qu.:0.0   1st Qu.: 2.00
#>  Median :0.00   Median : 52.00   Median :0.0   Median : 6.00
#>  Mean   :0.42   Mean   : 51.61   Mean   :0.1   Mean   : 5.58
#>  3rd Qu.:1.00   3rd Qu.: 60.50   3rd Qu.:0.0   3rd Qu.: 9.00
#>  Max.   :1.00   Max.   :102.00   Max.   :1.0   Max.   :11.00
#>        IQ             Trump
#>  Min.   : 48.00   Min.   :  0.00
#>  1st Qu.: 89.75   1st Qu.: 20.00
#>  Median :100.00   Median : 57.50
#>  Mean   :100.26   Mean   : 53.15
#>  3rd Qu.:111.25   3rd Qu.: 85.00
#>  Max.   :145.00   Max.   :100.00
#>
#> test:
#>
#>    Geschlecht       Alter          Minderheit   Fremdenfeindlich
#>  Min.   :0.00   Min.   :16.00   Min.   :0.00   Min.   : 1.00
#>  1st Qu.:0.00   1st Qu.:41.25   1st Qu.:0.00   1st Qu.: 1.25
#>  Median :1.00   Median :49.50   Median :0.00   Median : 5.00
#>  Mean   :0.56   Mean   :50.50   Mean   :0.12   Mean   : 5.36
#>  3rd Qu.:1.00   3rd Qu.:60.75   3rd Qu.:0.00   3rd Qu.: 9.00
#>  Max.   :1.00   Max.   :85.00   Max.   :1.00   Max.   :11.00
#>        IQ           Trump
```

```
#>  Min.   : 66   Min.   :  0.00
#>  1st Qu.: 92   1st Qu.: 15.00
#>  Median :102   Median : 47.50
#>  Mean   :102   Mean   : 49.80
#>  3rd Qu.:112   3rd Qu.: 88.75
#>  Max.   :135   Max.   :100.00
#>
#> MSE:
#> [1] 40.43078
#> ------------------------------
#>
#> [1] 3
#> train:
#>
#>    Geschlecht       Alter         Minderheit   Fremdenfeindlich
#>  Min.   :0.00   Min.   : 7.00   Min.   :0.00   Min.   : 1.00
#>  1st Qu.:0.00   1st Qu.:42.00   1st Qu.:0.00   1st Qu.: 1.00
#>  Median :0.00   Median :52.50   Median :0.00   Median : 4.50
#>  Mean   :0.46   Mean   :51.76   Mean   :0.13   Mean   : 5.02
#>  3rd Qu.:1.00   3rd Qu.:62.25   3rd Qu.:0.00   3rd Qu.: 8.00
#>  Max.   :1.00   Max.   :83.00   Max.   :1.00   Max.   :11.00
#>        IQ            Trump
#>  Min.   : 48.0   Min.   :  0.0
#>  1st Qu.: 93.0   1st Qu.: 15.0
#>  Median :102.0   Median : 47.5
#>  Mean   :102.4   Mean   : 47.4
#>  3rd Qu.:113.0   3rd Qu.: 75.0
#>  Max.   :135.0   Max.   :100.0
#>
#> test:
#>
#>    Geschlecht       Alter         Minderheit   Fremdenfeindlich
#>  Min.   :0.00   Min.   : 16.0   Min.   :0.00   Min.   : 1.00
#>  1st Qu.:0.00   1st Qu.: 36.0   1st Qu.:0.00   1st Qu.: 3.25
#>  Median :0.00   Median : 48.0   Median :0.00   Median : 7.00
#>  Mean   :0.48   Mean   : 50.2   Mean   :0.06   Mean   : 6.48
#>  3rd Qu.:1.00   3rd Qu.: 57.0   3rd Qu.:0.00   3rd Qu.: 9.00
#>  Max.   :1.00   Max.   :102.0   Max.   :1.00   Max.   :11.00
#>        IQ            Trump
#>  Min.   : 66.0   Min.   :  0.00
#>  1st Qu.: 86.0   1st Qu.: 31.25
#>  Median : 94.5   Median : 70.00
#>  Mean   : 97.7   Mean   : 61.30
#>  3rd Qu.:108.8   3rd Qu.: 95.00
#>  Max.   :145.0   Max.   :100.00
#>
#> MSE:
#> [1] 43.26809
#> ------------------------------
```

Durch die zufällige Aufteilung der Datenpunkte in zwei disjunkte Datensätze (train, test) entstehen Diskrepanzen zwischen den Verteilungen der Variablen. Diese Schwankungen wirken sich dann auch auf die Modellgüte aus. Wir wissen aus der vergangenen Übung, dass das Merkmal “Fremdenfeinlich” die Kovariate mit dem stärksten Einfluss auf die abhängige Variable ist. Vergleichen wir die Verteilung dieses Merkmals

zwischen Train- und Testdatensatz in 1 mit dem aus 3, fällt auf, dass die Diskrepanzen zwischen train und test in 3 eindeutig stärker ausfallen, als in 1. Dies spiegelt sich im jeweiligen MSE wieder: Das Modell, welches mit den Daten aus 1 trainiert und getestet wurde, weist einen deutlich niedrigeren MSE auf, als das Modell aus 3.

A 2

a)

```
mse <- c()
for (i in 1:nrow(Donald_1)) {
    lou <- lm(Trump ~ Alter + Geschlecht + Minderheit + Fremdenfeindlich + IQ,
        data = Donald_1[-i, ])
    pred <- predict.lm(lou, Donald_1[i, ])
    mse[i] <- mean((Donald_1[i, ]$Trump - pred)^2)
}

mse <- sum(mse)/nrow(Donald_1)
mse
#> [1] 38.19052
```

Der MSE für die Leave-one-out Cross-Validation liegt zwischen den berechneten MSE's aus Aufgabe 1. Dies ist zu erwarten, da wir hier untersuchen, wie gut das lineare Regressionsmodell auf "ungesehenen" Daten performt. In dem wir jeweils nur einen Datenpunkt als "Test"-Datensatz verwenden, bekommen wir so viele MSE-Werte, wie Datenpunkte im Datensatz. Diese MSE-Werte werden gemittelt, um eine "robustere" Einschätzung über die Generalisierbarkeit des Modells auf unsere Daten zu erhalten.

b)

```
packageTest("boot")

model <- glm(Trump ~ Alter + Geschlecht + Minderheit + Fremdenfeindlich + IQ,
    data = Donald_1)

cv_model <- cv.glm(Donald_1, model, K = nrow(Donald_1))

mse1 <- cv_model$delta
mse1
#> [1] 38.19052 38.18070
```

Die Implementierung des "Leave-one-out cross-validation"-Verfahrens durch cv.glm gibt zwei Modellgütemetriken zurück. Der erste Wert ist der durschnittliche MSE über die Zeilen. Dieser ist identisch mit dem Wert aus a). Man kann vermuten, dass die Implementierung des "Leave-one-out cross-validation"-Verfahrens durch cv.glm identisch mit unserer manuellen Implementierung aus a) ist. Der zweite Wert ist laut Dokumentation ein Bias-korrigierter MSE.

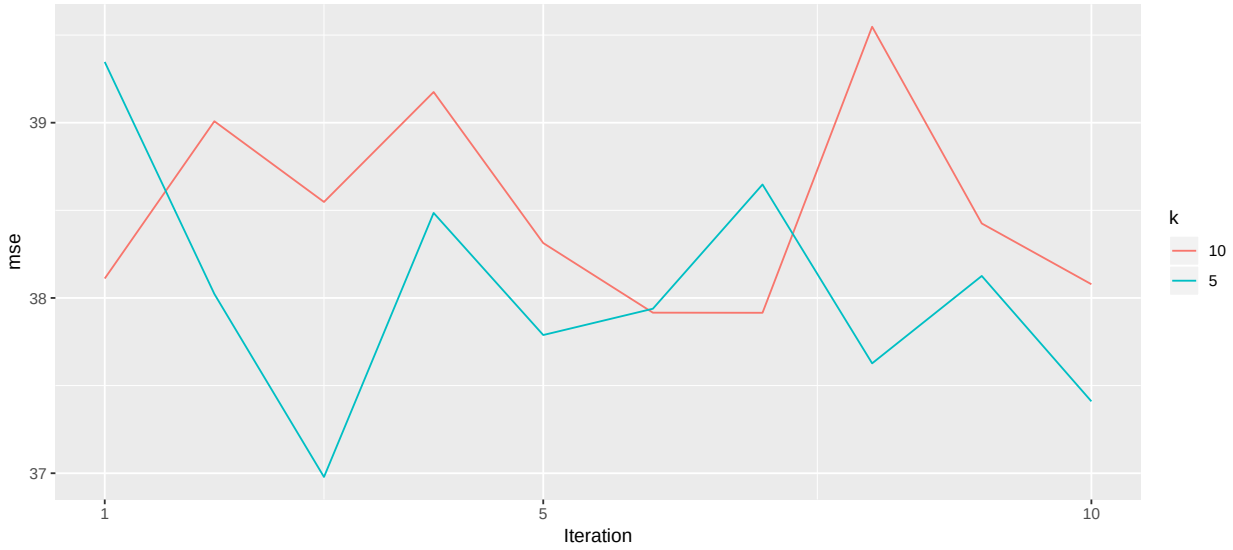
A 3

```
packageTest("ggplot2")
set.seed(42)
```

```r
mse <- matrix(NA, nrow = 20, ncol = 3)
for (k in c(5, 10)) {
    for (i in 1:10) {
        mse[ifelse(k == 5, i, i + 10), ] <- c(i, cv.glm(Donald_1, model, K = k)$delta[1],
            k)
    }
}

mse <- data.frame(mse)
names(mse) <- c("index", "mse", "k")
mse$k <- as.character(mse$k)

gg = ggplot(data = mse, mapping = aes(x = index, y = mse, color = k))
gg = gg + xlab("Iteration")
gg = gg + scale_x_continuous(breaks = c(1, 5, 10))
gg + geom_line()
```



Bei der k-fachen $k$reuzvalidierung wird die ursprüngliche Stichprobe zufällig in k gleich große Teilstichproben aufgeteilt. Von den k Teilstichproben wird eine einzige Teilstichprobe als Validierungsdaten für den Test des Modells aufbewahrt und die restlichen $k - 1$ Teilstichproben werden als Trainingsdaten verwendet. Der $k$reuzValidierungsprozess wird dann k-mal wiederholt, wobei jede der k Teilstichproben genau einmal als Validierungsdaten verwendet wird. Die k-Ergebnisse $k$önnen dann gemittelt werden, um eine einzige Schätzung zu erhalten. Der Vorteil dieser Methode gegenüber des wiederholt zufälligen Subsampling besteht darin, dass alle Beobachtungen sowohl für das Training als auch für die Validierung verwendet werden und jede Beobachtung genau einmal für die Validierung verwendet wird. Die k-fache $k$reuzvalidierung mit $k = 10$ wird häufig verwendet, aber im Allgemeinen bleibt k ein nicht fixierter Parameter.

Untereinander schwanken die MSE-Werte zufällig. Als Trend ist zu beobachten, dass für $k = 5$ die MSE-Werte niedriger sind. Alles in allem variieren die MSE-Werte um einen Wert von 38. Dies deckt sich mit den Ergebnissen aus Aufgabe 2. Die Schwankungen sind ebenfalls im Bereich der generierten MSE-Werte aus aus Aufgabe 1. Interessant ist, dass wir in Aufgabe 1 mit MSE ~ 35 einen, im Vergleich, relativ niedrigen MSE-Wert erzielt haben.