University of St.Gallen

Model Evaluation

University of St. Gallen
School of Management, Economics, Law,
Social Sciences, International Affairs
and Computer Science

Assignment 1

Data Analytics I: Predictive Econometrics
Prof. Jana Mareckova

submitted by

Cyril Janak, 16-611-287
Jonas Husmann, 16-610-917
Niklas Kampe, 16-611-618
Robin Scherrer, 18-617-969

01.12.2021

Contents

Requirements

To solve the following tasks, the required library and the data set are loaded first. The library *ggplot2* is used for plotting various graphics.

```
library(ggplot2)
load("GHA/insurance-all.RData")
```

Exercise 1

The number of observations in the data set corresponds to the number of rows and the number of covariates collected corresponds to the number of columns minus one (dependent variable). Thus there are 1204 observations and 6 covariates.

```
(n_obs <- nrow(data))
```

```
## [1] 1204
```

```
(n_cov <- ncol(data) - 1)
```

```
## [1] 6
```

Exercise 2

The highest number of children who are covered by one health insurance is 5.

```
(max_children <- max(data$children))
```

```
## [1] 5
```

Exercise 3

The region *northwest* has the lowest share of smokers. The share of smokers in this region is 16.9 percent.

```
pct_smokers_by_region <- aggregate(data$smoker == "yes",
                                   by=list(region=data$region),
                                   FUN=function(x) sum(x)/length(x))

pct_smokers_by_region[which.min(pct_smokers_by_region$x),]
```

```
##      region         x
## 2 northwest 0.1689655
```