



# 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**

<b>Requirements</b>	<b>1</b>
<b>Exercise 1</b>	<b>1</b>
<b>Exercise 2</b>	<b>1</b>
<b>Exercise 3</b>	<b>1</b>
<b>Exercise 4</b>	<b>2</b>
<b>Exercise 5</b>	<b>3</b>

## 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))
```

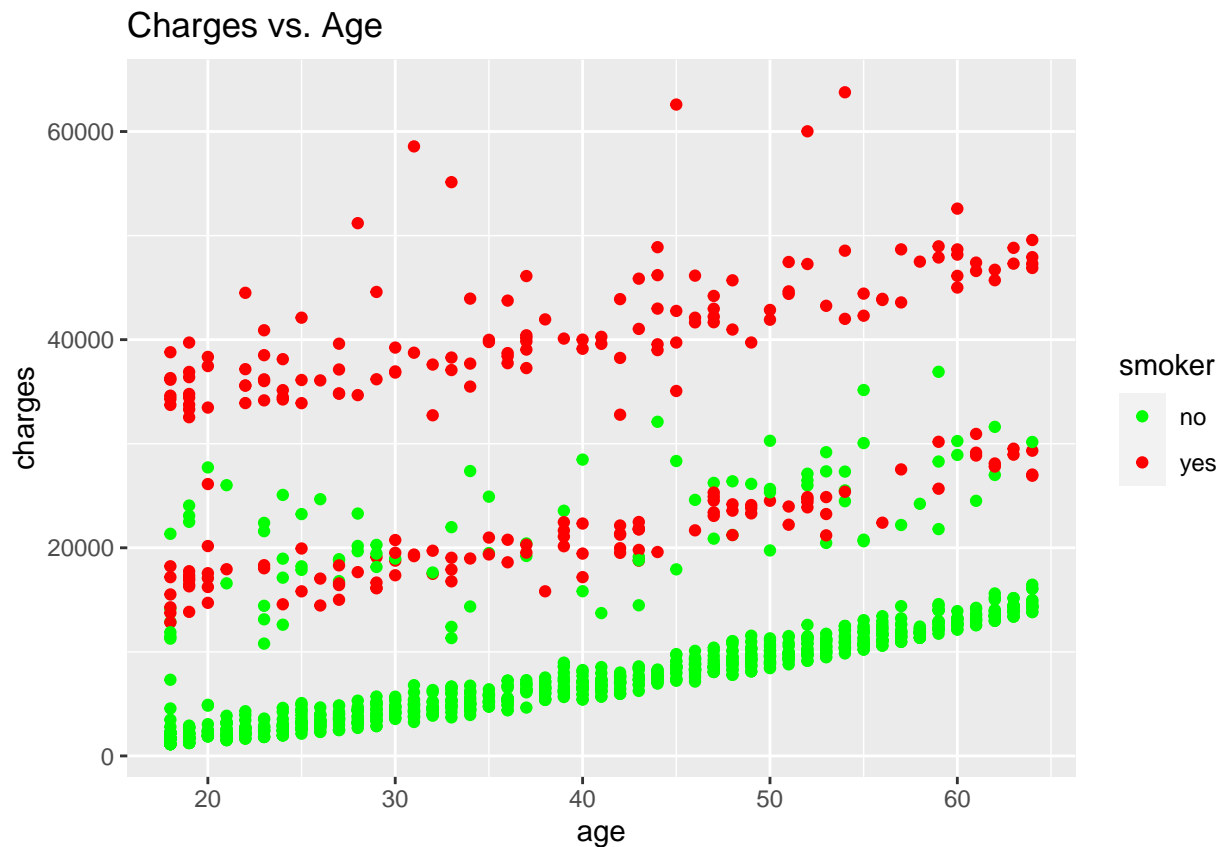
```
pct_smokers_by_region[which.min(pct_smokers_by_region$x),]
```

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

## Exercise 4

According to the plot, two patterns can be identified in the data. It can be seen that the older the primary beneficiary, the higher the individual medical costs billed by health insurance. This makes sense because the likelihood of health problems increases with age. Furthermore, it can be seen that the average medical costs billed by health insurance are higher for the group of smokers than for the group of non-smokers. This also makes sense, as smoking increases the likelihood of health problems. It can be assumed that the area where the average medical costs billed by health insurance of smokers and non-smokers overlap can be explained by other covariates.

```
ggplot2::ggplot(data = data, mapping = ggplot2::aes(x=age, y=charges, color=smoker)) +  
  ggplot2::ggtitle(label = "Charges vs. Age") +  
  ggplot2::geom_point() +  
  ggplot2::scale_color_manual(values = c("green", "red"))
```



## Exercise 5

The plot shows that the body mass index and the medical costs billed by health insurance are positively related. This makes sense, as being overweight increases the risk of health problems. One might have expected the relationship to be even stronger. It can be assumed that this can be explained by other covariates.

```
dynamic_plot <- function(data, x.variable, y.variable, color.variable) {  
  ggplot2::ggplot(data = data,  
    mapping = ggplot2::aes_string(x=x.variable,  
                                   y=y.variable,  
                                   color=color.variable)) +  
  ggplot2::geom_point()  
}  
  
dynamic_plot(data = data,  
  x.variable = "bmi",  
  y.variable = "charges",  
  color.variable = "sex")
```

