Data Analytics I: Predictive Econometrics
Assignment I

Lauritz Storch
Sebastian Maser
Ivan Medvedev
Jan Gutjahr

2022-11-25

Data Preparation and Package Installation

Download the data set insurance-all.Rdata[1] from Canvas. Load the data into R. Install and load the package ggplot.

```
## Set working directory to load data
setwd("~/Downloads")


## Packages and Library
# Packages function
install_if_missing <- function(p) {
  if (p %in% rownames(installed.packages()) == F) {
    install.packages(p, dependencies = T)
  } else {
    cat(paste("Skipping already installed package:", p, "\n"))
  }
}

# Define packages
packages = c("ggplot2")

# Install
invisible(sapply(packages, install_if_missing))
```

```
## Skipping already installed package: ggplot2
```

```
# read library
for(pkg in packages){
  library(pkg, character.only = TRUE)
}


## Read in data
load("insurance-all.Rdata")
```

[1] Please make sure your device is connected to the internet and you are logged into Canvas to download the dataset.

Task 1

Q: How many observations are in the whole dataset? How many covariates were collected? (0.5 points)

```
# Number of observations
cat("The number of obersations in the dataset: ", nrow(data))
```

```
## The number of obersations in the dataset:  1204
```

```
# Number of covariates (charges is not a covariate)
cat("The number of covariates in the dataset: ", ncol(data)-1)
```

```
## The number of covariates in the dataset:  6
```

Task 2

Q: What is the highest number of children who are covered by one health insurance? (0.5 points)

```
# Highest number of children
cat("The highest number of children who are covered by one
health insurance: ", max(data[["children"]]))
```

```
## The highest number of children who are covered by one
## health insurance:  5
```

```
# ID of families
cat("Families with ", max(data[["children"]]), " children are \n",
rownames(data[c(which(data[["children"]] == 5)),]))
```

```
## Families with  5  children are
##  933 439 414 878 426 938 985 569 641 1131 167 1273 970 72
```

Task 3

Q: Look at the percentage shares of smokers and non-smokers within each region, e.g. in one region there could be 80% smokers and 20% non-smokers. Which region has the lowest share of smokers? What is the share of smokers there? (0.5 points)

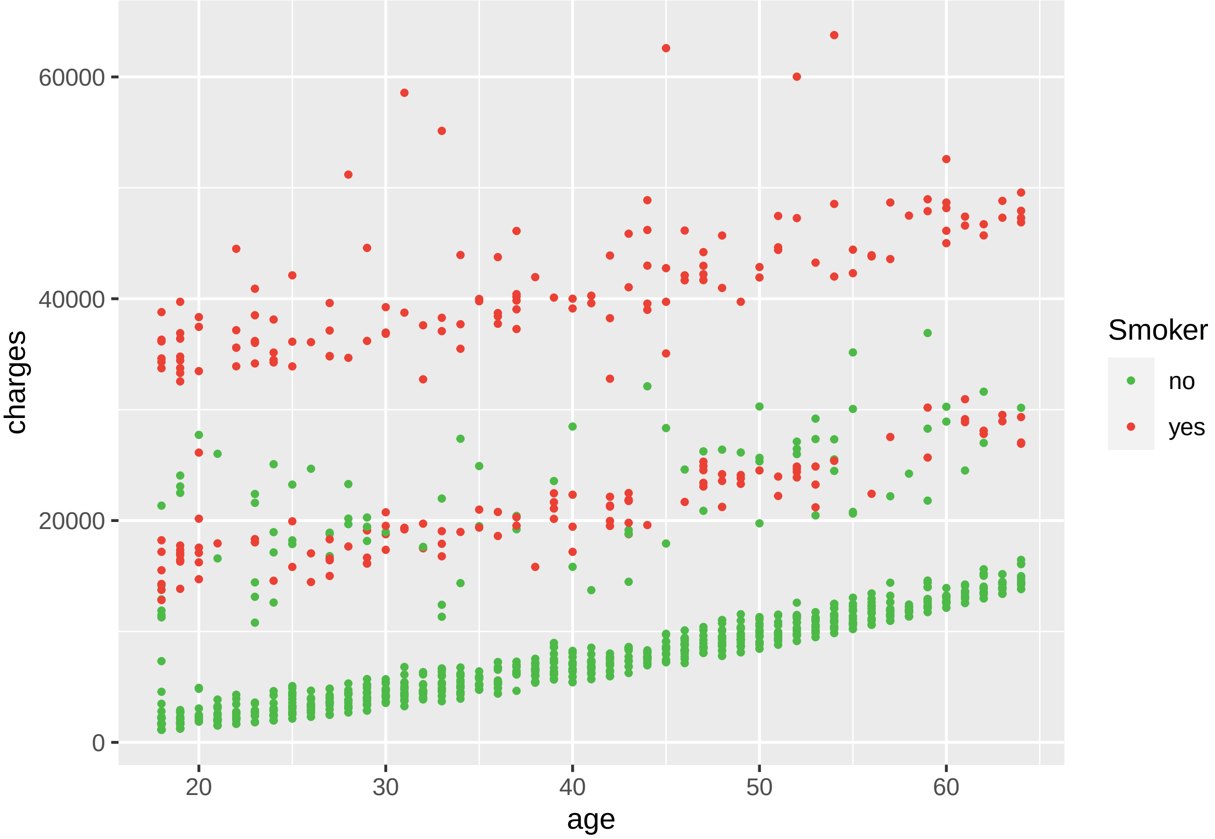
```
# Filter Regions
regions <- as.vector(unique(data$region))

# Percentage shares
share_vector <- c()
for (i in 1:length(regions)){
  sum_smoker <- length(which(data$region == regions[i] & data$smoker == "yes"))
  sum_region <- length(which(data$region == regions[i]))

  #assign(paste0("smoker_share_",regions[i]),sum_smoker/sum_region)
  share_vector <- c(share_vector, sum_smoker/sum_region)
}
cat(paste0("Region ",regions[which(share_vector == min(share_vector))],
           " has the lowest share of smokers.\n", round(min(share_vector),4),
           "% of people smoke there."))
```

```
## Region northwest has the lowest share of smokers.
## 0.169% of people smoke there.
```

Task 4

Q: Create a scatter plot with charges on the y-axis and age on the x-axis. Distinguish in the plot by color which data points belong to smokers and non-smokers (coded in covariate smoker). Describe the patterns in the data. (0.5 points)

```
ggplot(data, aes(x=age, y=charges)) +
  geom_point(aes(colour = smoker == "yes"),
             size=0.8) +
  scale_color_manual(name = "Smoker",
                     values = setNames(c('#eb4034','#4dba47'),c(T, F)),
                     labels = c("no","yes"))
```

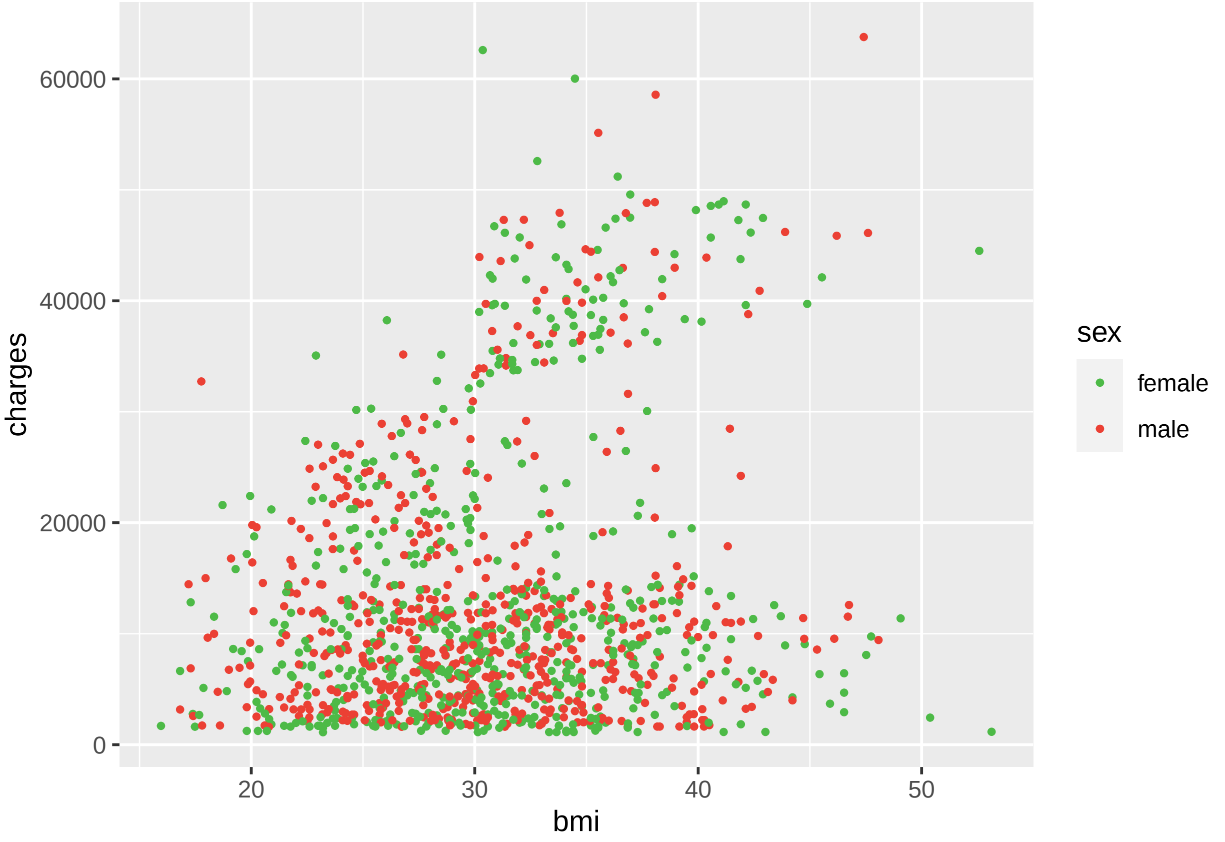


Overall, the individual medical costs billed by health insurance (charges) increase with age regardless of whether they smoke. Smokers tend to have higher charges than non-smokers.

Task 5

Q: Write a function that has a data argument and three string arguments x.variable, y.variable, color.variable which are used to generate a scatter plot of two variables against each other and use the third to color the points. Use this function to plot x.variable="bmi", y.variable="charges", color="sex" and interpret the results. (1 point)

```
# Function
scatter_fct <- function(xvar, yvar, colorvar){
  #' Make a scatter plot
  #'
  #' @param xvar Value on x-axis (string)
  #' @param yvar Value on y-axis (string)
  #' @param colorvar Binary variable used to color the points (string)
  #'
  #' @return Generates a scatter plot of two variables against each
  #' other and use the third to color the points
  #'

  unique_values <- c(as.vector(unique(data[[colorvar]])))

  ggplot(data, aes(x=get(xvar), y=get(yvar))) +
    geom_point(aes(colour = get(colorvar) == "female"),
               size=0.8) +
      scale_color_manual(name = colorvar,
                     values = setNames(c('#eb4034','#4dba47'),c(T, F)),
                     labels = unique_values) +
    xlab(xvar) + ylab(yvar)
}
# Plot
scatter_fct("bmi","charges","sex")
```

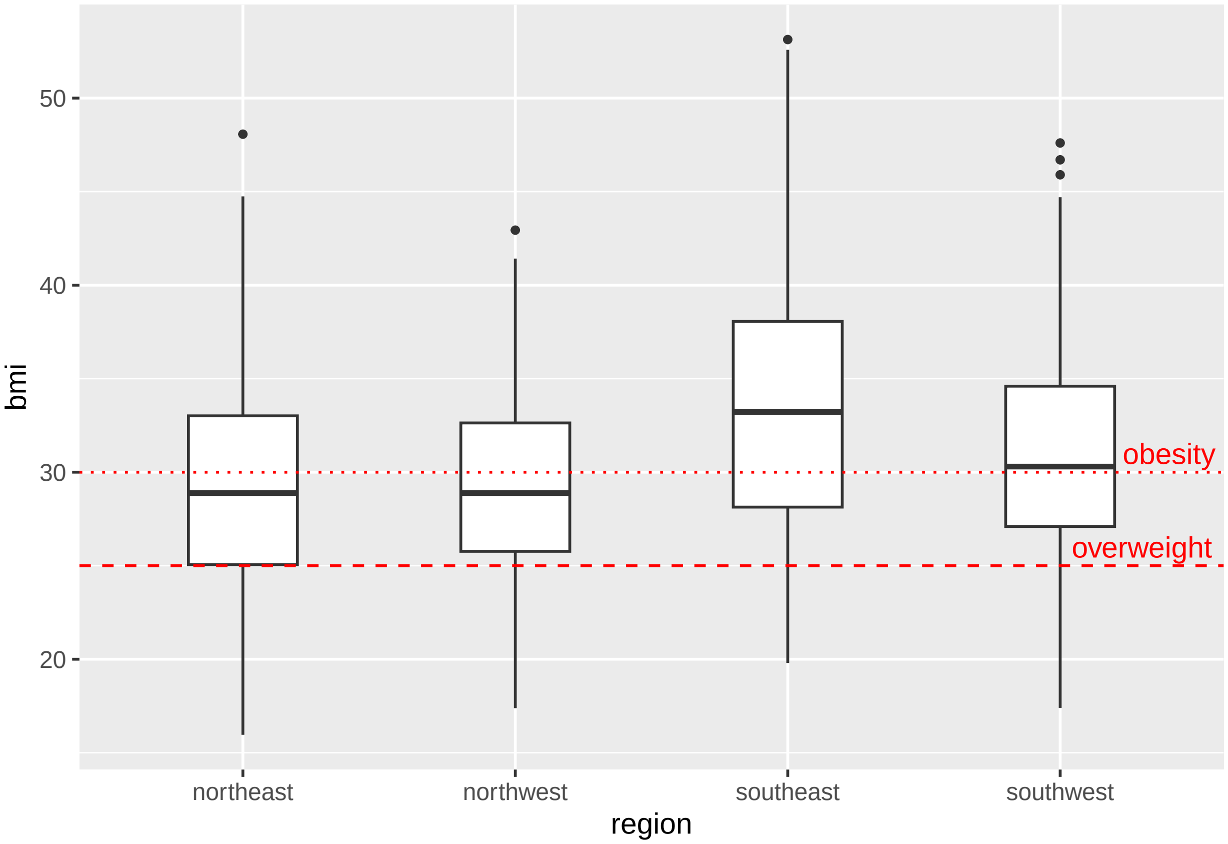


The individual medical costs billed by health insurance (charges) increase with BMI. We do not observe a deterministic relationship, as almost everybody with high charges has a higher BMI, but only some with a high BMI have high insurance costs. Nevertheless, a high BMI caused by high body fat evidently increases health risk, which comes with higher insurance costs. Men and women are equally affected, and we observe no difference between the sexes in the positive correlation between charges and BMI.

Task 6

Q: Write a function that creates a boxplot of the bmi variable split by another variable that is passed as the argument. Do you see any difference in the bmi by region? (1 point)

```
boxplot_fct <- function(xvar){
  #' Make a boxplot
  #'
  #' @param xvar Value on x-axis (string)
  #'
  #' @return creates a boxplot of the bmi variable split by another variable
  #'

  graph <- ggplot(data, aes(x=get(xvar), y=bmi)) +
    geom_boxplot(size=0.5, outlier.size=1, width=0.4) +
    xlab(xvar) + ylab("bmi")

  if (xvar == "region") {
    graph <- graph +
      geom_hline(yintercept=25, linetype="dashed", color = "red") +
      geom_hline(yintercept=30, linetype="dotted", color = "red") +
      annotate(geom="text", x=4.3, y=26, label="overweight",color="red") +
      annotate(geom="text", x=4.4, y=31, label="obesity",color="red")
  }

  return(graph)
}

boxplot_fct("region")
```



Northern regions have an equal mean slightly below 30. The percentiles differ by a margin, with the 75% percentile around 33 and the 25% at 25. Southwest has a marginally higher mean at 30. The 75% percentile is at 35, and the 25% is around 26. The southeast region has a remarkably high mean, approximately 33. The 75% percentile is about 38, and the 25% is at 28, which equals the mean of the northern regions. The southeastern region also has the highest outlier at the upper end with a value of 53, and the northeast region has the lowest outlier with a value of 16. Since all BMI values above 25 and 30 are classified as overweight and obese, respectively.[2] In the context of predicting medical costs billed by health insurance, all regions are in a critical area for health, as an increased BMI is an indicator of overweight or obesity.

[2] See also the red dashed and dotted lines in the graph.