

Machine Learning I Group Work

Adriana Ricklin, Yvonne Schaerli, Christina Sudermann, Carole Mattmann

11 June 2020

Contents

1 Packages	2
2 Import and data cleaning	2
3 Linear models -> Christina	3
3.1 Basic analysis of continuous variables	3
3.2 Basic analysis of categorical variables	5
3.3 Basic analysis of categorical and continuous variables	6
3.4 Fitting a first Linear Model	12
3.5 Final Linear Model Development	15
4 Linear models (GAM & Polynomial) -> Yvonne	25
5 GLM and cross validation -> Carole	25
5.1 Generalised Linear Models for count data	25
5.2 Generalised Linear Models for binomial data	27
5.3 Cross Validation	29
6 Decision Trees -> Adriana	32
7 Support Vector Machines -> Carole	32
7.1 two classes linear	32
7.2 two classes polynomial	36
8 Neural Networks	38
9 Cross validation and discussion	38

1 Packages

```
library(gridExtra)
library(tidyverse) # included ggplot2
library(e1071)      # for SVM
#library(ggplot2)
```

2 Import and data cleaning

```
insurance <- read.csv("../01_data/insurance.csv", header=TRUE)
str(insurance)
```

```
## 'data.frame': 1338 obs. of 7 variables:
## $ age : int 19 18 28 33 32 31 46 37 37 60 ...
## $ sex : chr "female" "male" "male" "male" ...
## $ bmi : num 27.9 33.8 33 22.7 28.9 ...
## $ children: int 0 1 3 0 0 0 1 3 2 0 ...
## $ smoker : chr "yes" "no" "no" "no" ...
## $ region : chr "southwest" "southeast" "southeast" "northwest" ...
## $ charges : num 16885 1726 4449 21984 3867 ...
```

```
# smoker = 1 / nonsmoker = 0
insurance$smoker <- as.character(insurance$smoker)
insurance$smoker[insurance$smoker == "yes"] <- "1"
insurance$smoker[insurance$smoker == "no"] <- "0"
insurance$smoker <- as.factor(insurance$smoker)

# sex
insurance$sex <- as.character(insurance$sex)
insurance$sex[insurance$sex == "yes"] <- "1"
insurance$sex[insurance$sex == "no"] <- "0"
insurance$sex <- as.factor(insurance$sex)

# region
insurance$region <- as.character(insurance$region)
insurance$region[insurance$region == "southwest"] <- "1"
insurance$region[insurance$region == "southeast"] <- "0"
insurance$region[insurance$region == "northeast"] <- "2"
insurance$region[insurance$region == "northwest"] <- "3"
insurance$region <- as.factor(insurance$region)
```

```
head(insurance)
```

```
##   age   sex   bmi children smoker region   charges
## 1  19 female 27.900         0       1      1 16884.924
## 2  18  male 33.770         1       0      0 1725.552
```

```
## 3 28 male 33.000      3      0      0 4449.462
## 4 33 male 22.705      0      0      3 21984.471
## 5 32 male 28.880      0      0      3 3866.855
## 6 31 female 25.740      0      0      0 3756.622
```

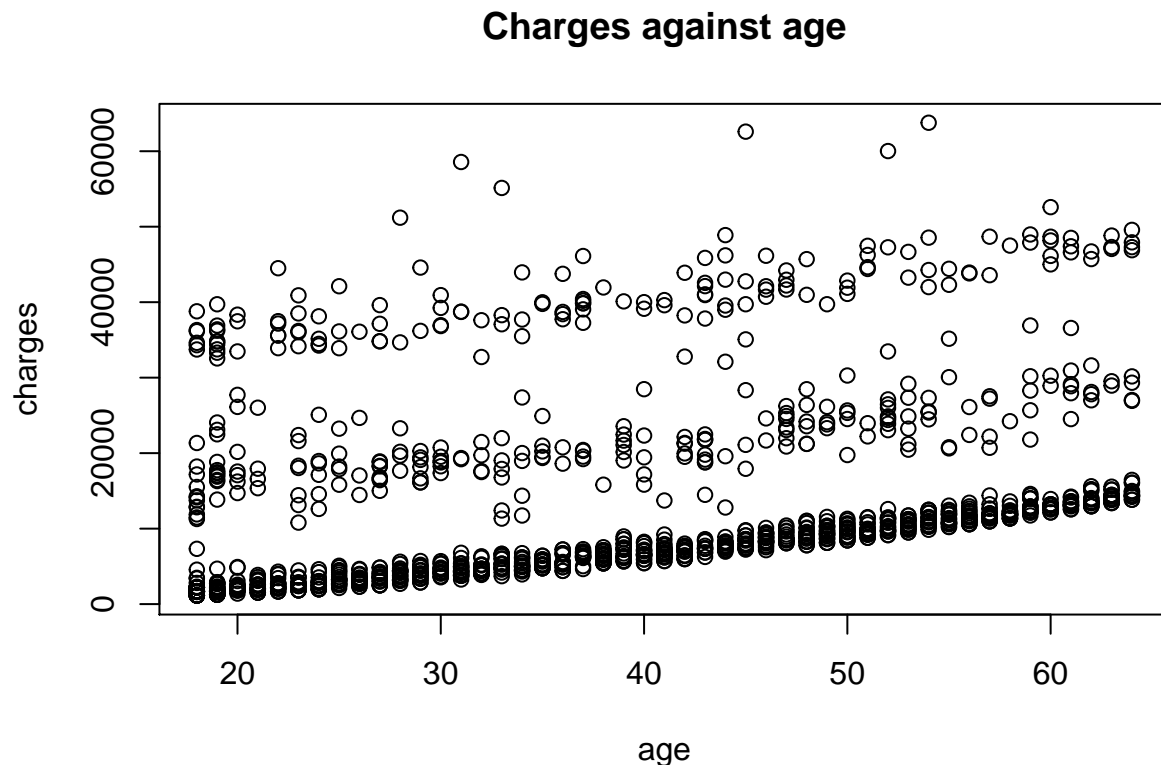
3 Linear models -> Christina

The data set used for this project contains 1338 observations and seven variables containing the age, sex, bmi, children, smoker, region and charges. Those variables are covering continuous and categorical variables. The region as an categorical variable is covering four different levels.

3.1 Basic analysis of continuous variables

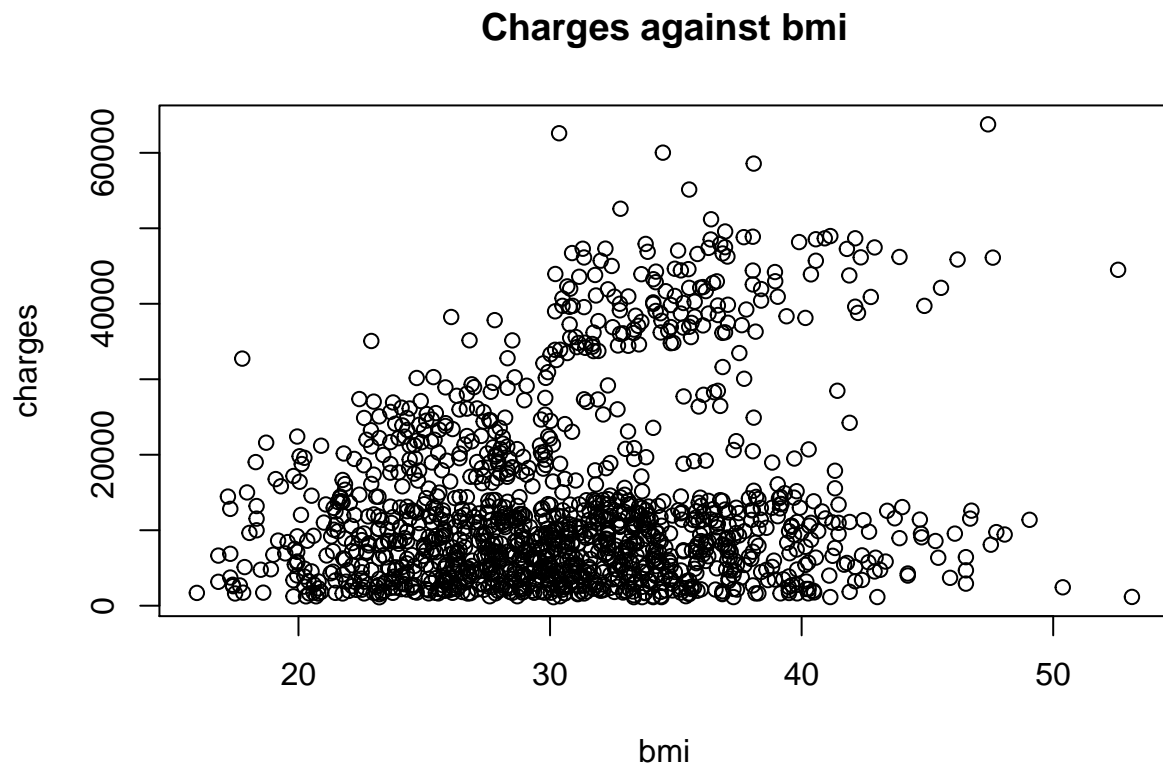
Here we will begin with a graphical analysis. Therefore we will plot the response variable against the given predictors to gather first relationships, which can be used later in the modelling process.

```
plot(charges ~ age, data = insurance, main = 'Charges against age')
```



There seems to be a positive relationship between age and charges.

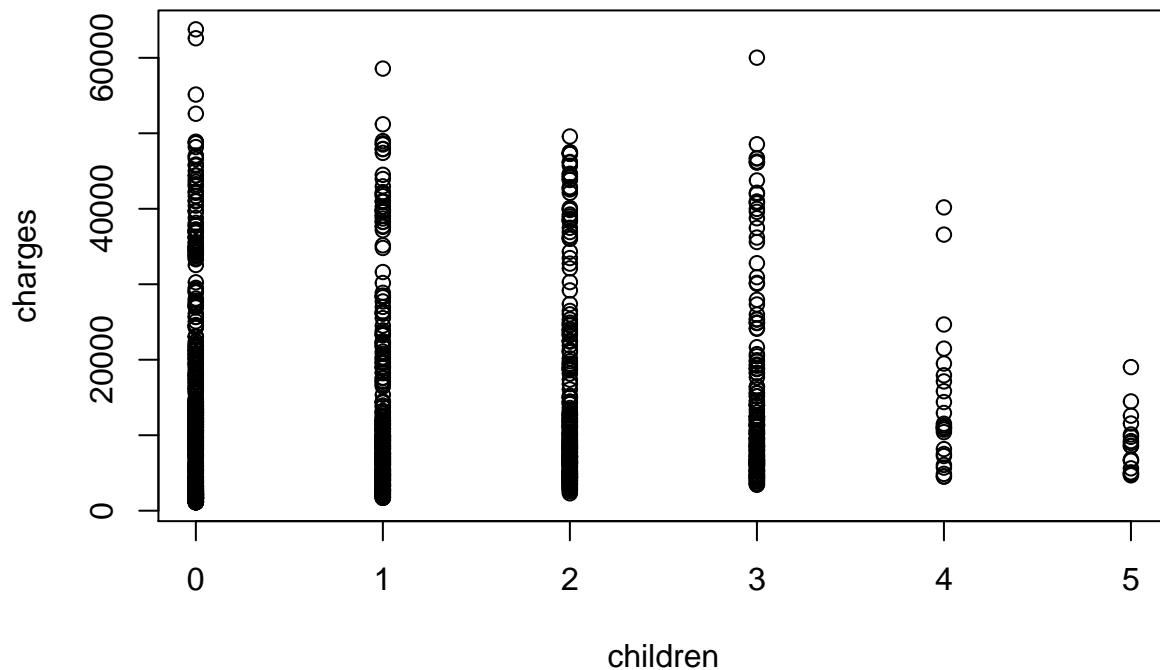
```
plot(charges ~ bmi, data = insurance, main = 'Charges against bmi')
```



The above plot is not showing a clear relationship between bmi and the corresponding charges.

```
plot(charges ~ children, data = insurance, main = 'Charges against children')
```

Charges against children



There might be an affect between the number of children and the charges. As shown above it might be possible to interpret that with a rising number of children the charges a decreasing. Still this is not a clear relationship, more a wide interpretation of the plot.

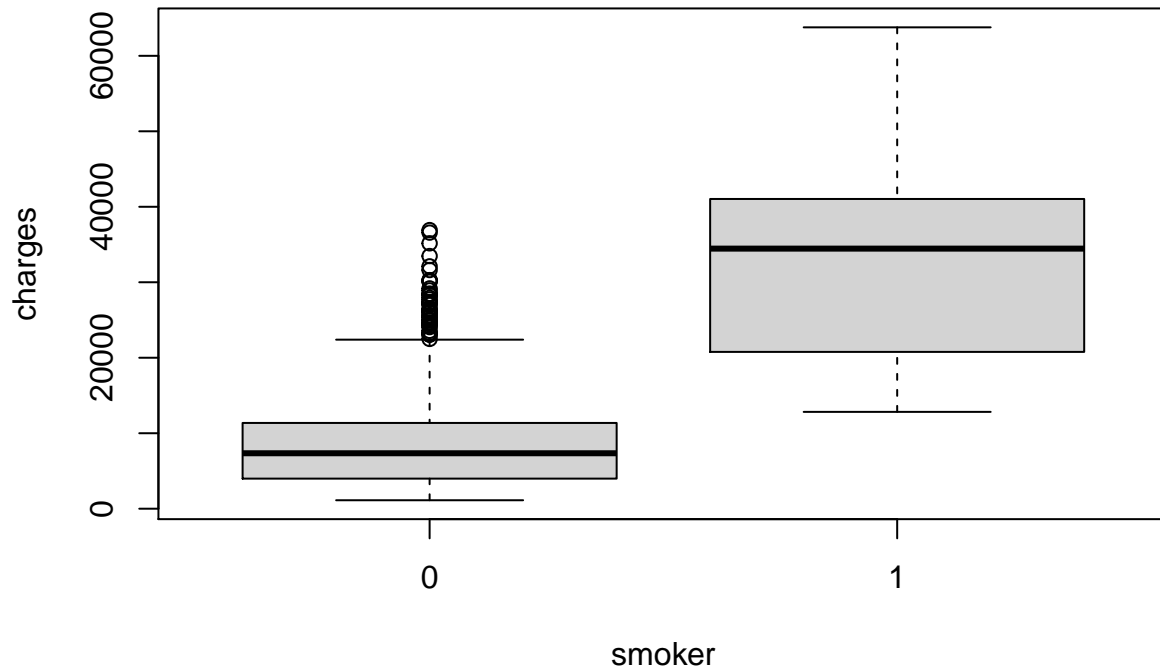
3.2 Basic analysis of categorical variables

```
##plot(charges ~ sex, data =insurance, main = 'Charges against sex')
```

As shown above there are some differences in the charges, comparing the boxplot for the two given genders. It seems that the range for 50% of the observations is bigger than the one for the female group. Also the 95% quantile is about 10'000 lower than the same quantile for the male group.

```
plot(charges ~ smoker, data =insurance, main = 'Charges against smoker')
```

Charges against smoker



This plot is showing a clear affect of smoking on the charges. As you can see persons who are not smoking having mean charges that are three times less as the one of people who are smoking. even the outliers of the smokers are still in the range where only 50% of the smokers are located.

```
##plot(charges ~ region, data =insurance, main = 'Charges against region')
```

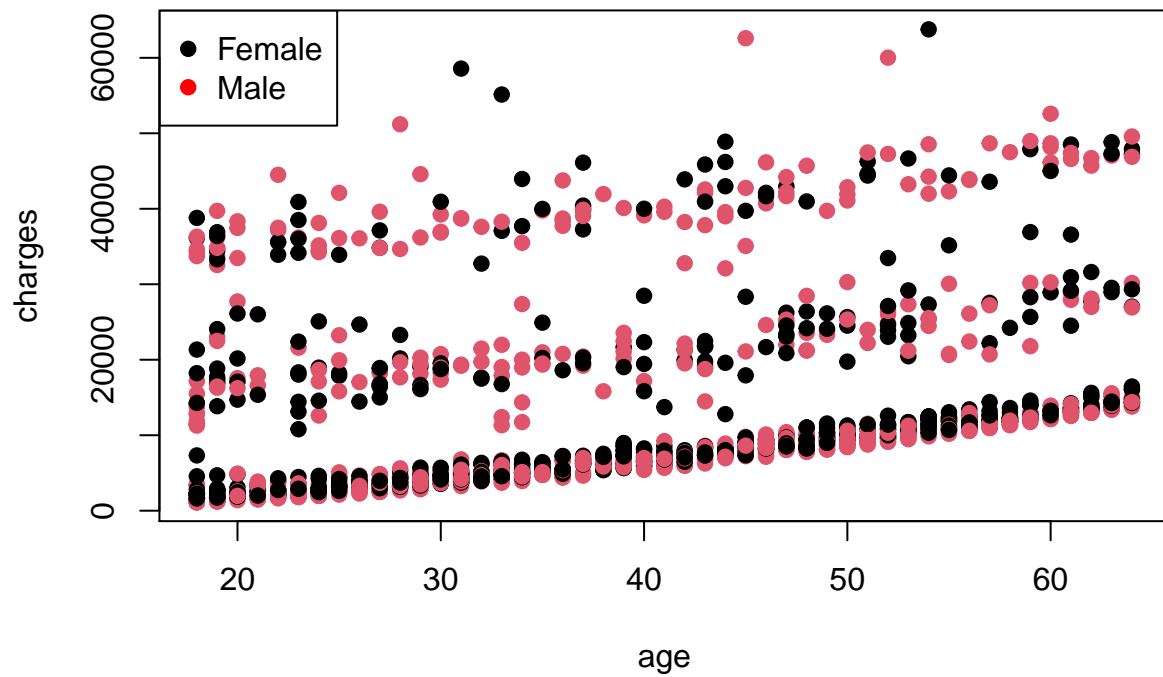
In this case there isn't any clear relationship or trend visible between the region and the corresponding charges. Comparing the four regions together seems that there are slightly small differences in the width of the box, distribution of 50% of the observation as well the setting of the 95% quantile. Those differences will be analysed later.

3.3 Basic analysis of categorical and continous variables

In this part we want to show exemplary how the relationship between charges and the age can be affected by adding additionally a categorical variable. This can be used for further data explorations and as well be adapted on all other variables. For this section we will just perform this enlargement for the case of the age to illustrate some possible relationships, which also can be helpful for the following modeling process.

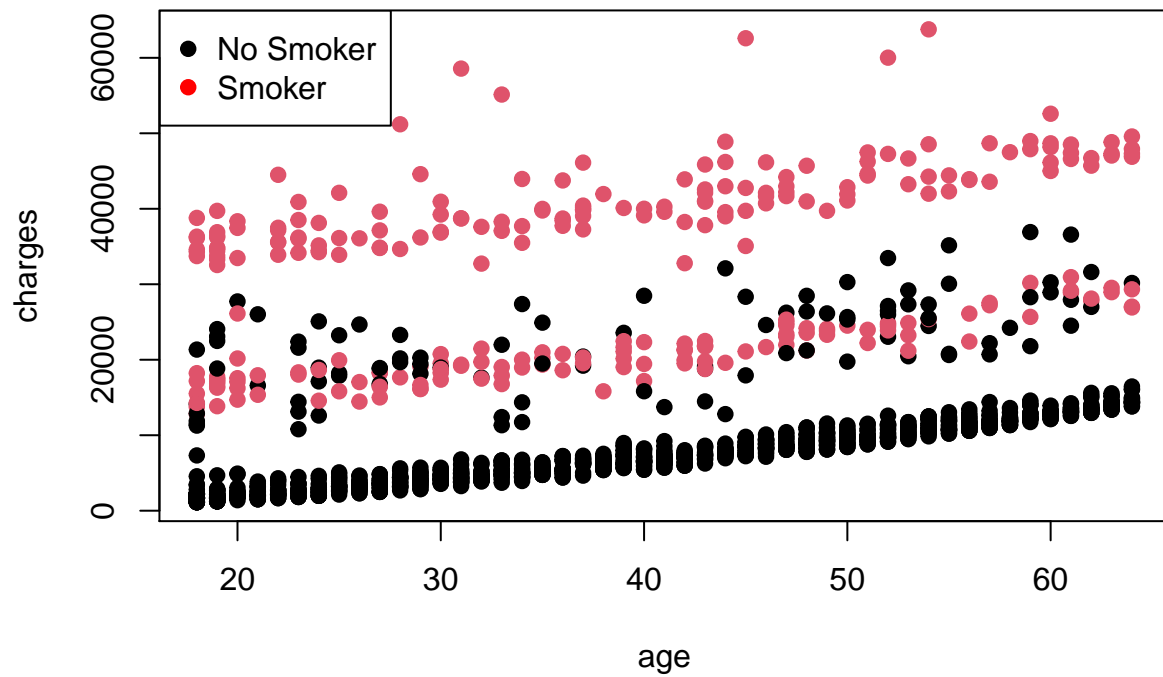
```
plot(charges ~ age, data = insurance,  
     col = sex,  
     pch = 19,  
     main = "Charges against age (sex)")  
legend("topleft",  
      pch = 19,  
      legend = c("Female", "Male"),  
      col = c("black", "red"))
```

Charges against age (sex)



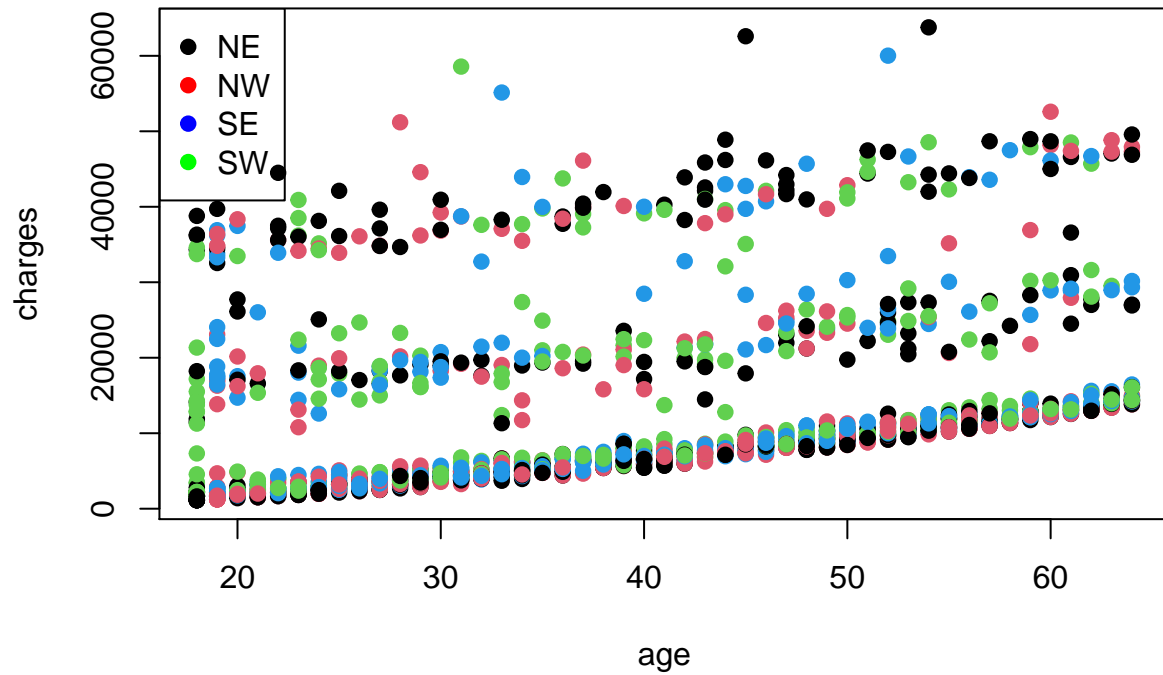
```
plot(charges ~ age, data = insurance,
     col = smoker,
     pch = 19,
     main = "Charges against age (smoker)")
legend("topleft",
     pch = 19,
     legend = c("No Smoker", "Smoker"),
     col = c("black", "red"))
```

Charges against age (smoker)



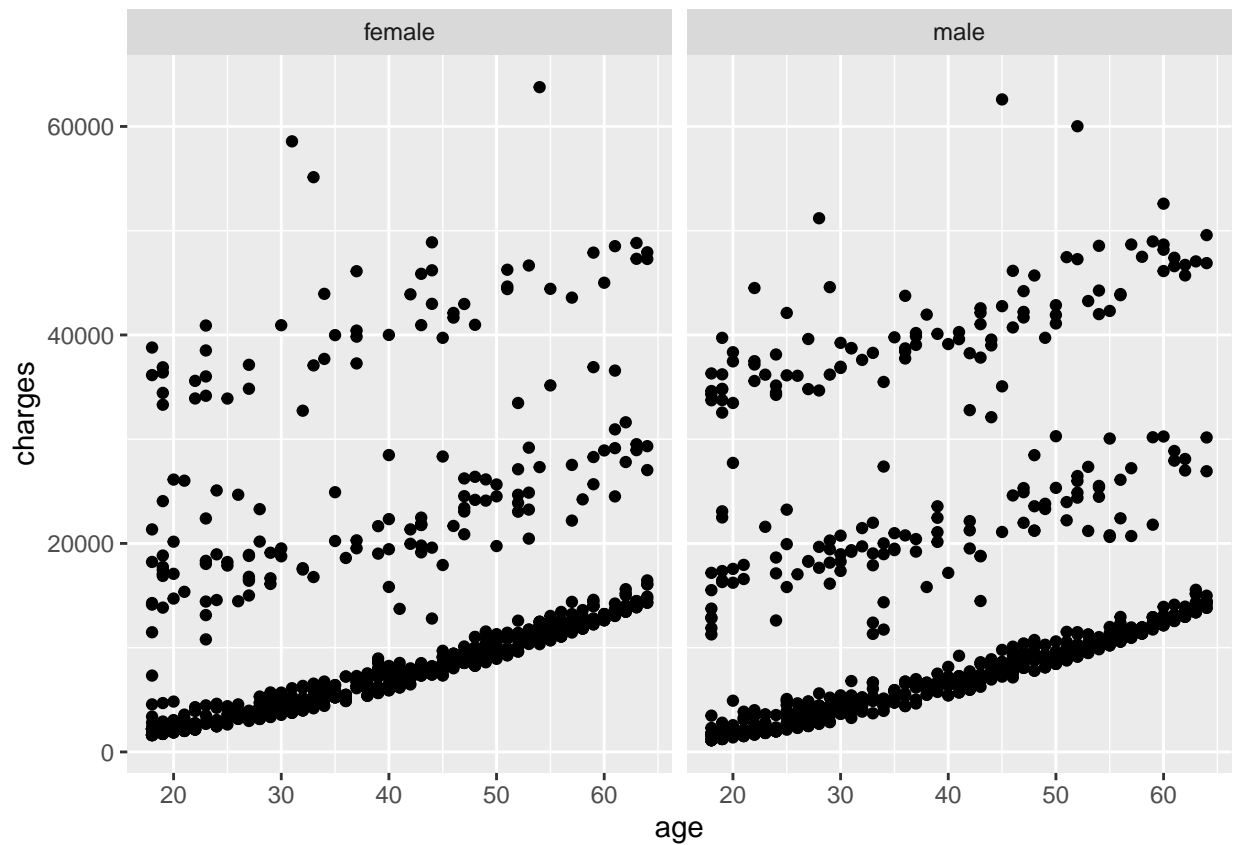
```
plot(charges ~ age, data = insurance,  
     col = region,  
     pch = 19,  
     main = "Charges against age (region)")  
legend("topleft",  
      pch = 19,  
      legend = c("NE", "NW", "SE", "SW"),  
      col = c("black", "red", "blue", "green"))
```


Charges against age (region)



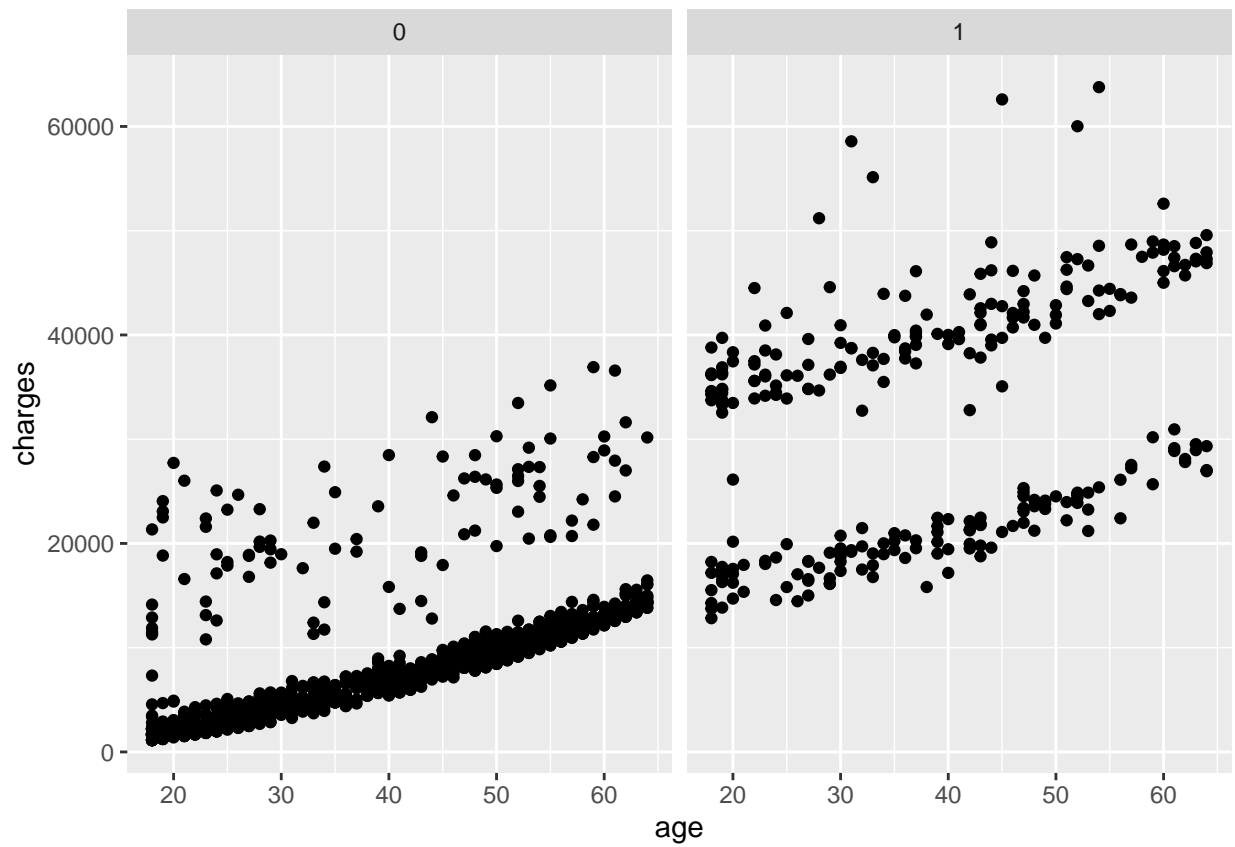
Since it is possible to have overlapping observations, we will plot the same set-up using the `qplot`. Having separate facets will avoid overlapping and therefore might serve different results as already seen above.

```
qplot(y=charges, x=age,  
      data = insurance,  
      facets = ~ sex)
```

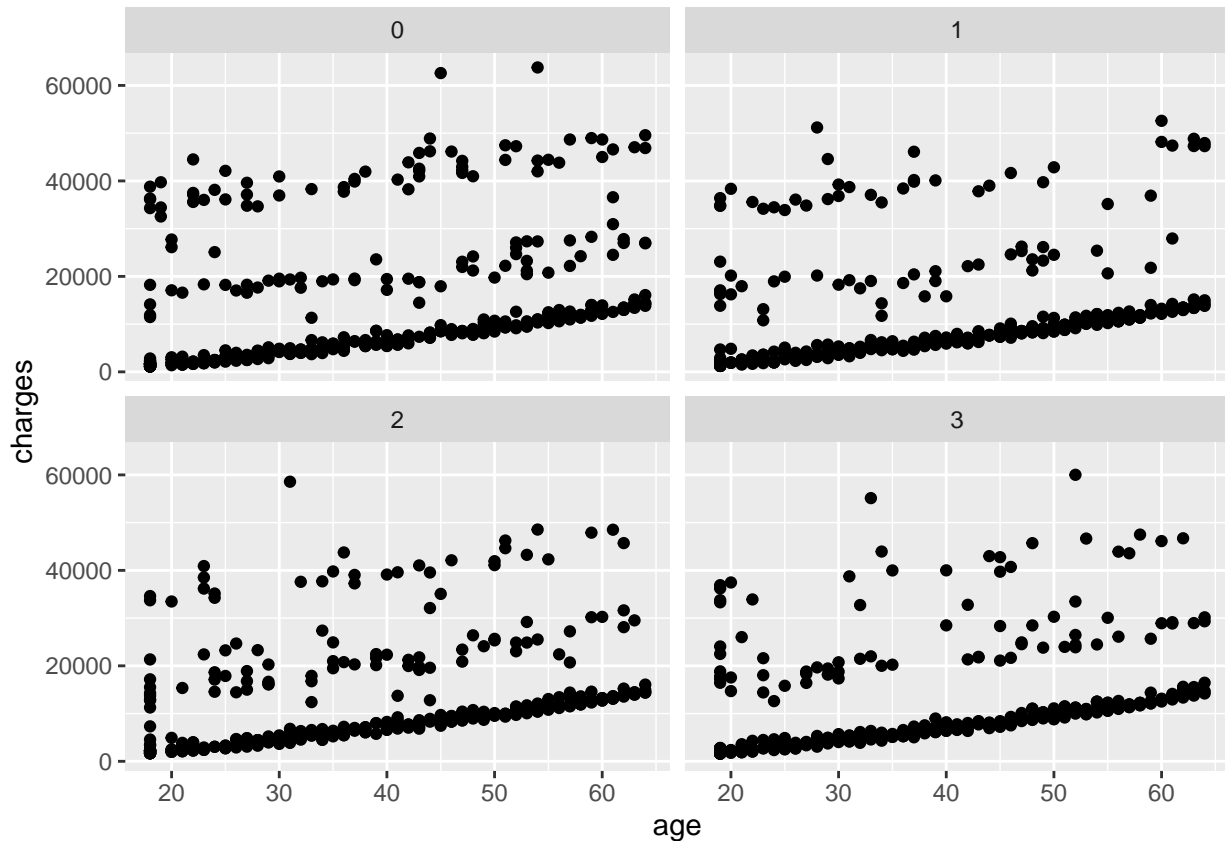


```
lm.insurance <- lm(charges ~ age, data = insurance)
```

```
qplot(y=charges, x=age,
      data = insurance,
      facets = ~ smoker)
```



```
qplot(y=charges, x=age,  
      data = insurance,  
      facets = ~ region)
```



3.4 Fitting a first Linear Model

As usually we will start by fitting a simple regression model to the insurance dataset. Therefore we will start with one variable and add additional complexity in each following subset. This step-by-step approach helps to explore the data slightly better and will simplify the final modelling.

3.4.1 Linear Model with one variable

For the first simple regression model we will use the age.

```
lm.insurance <- lm(charges ~ age, data = insurance)
summary(lm.insurance)
```

```
##
## Call:
## lm(formula = charges ~ age, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8059   -6671   -5939    5440   47829
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3165.9      937.1    3.378 0.000751 ***
```

```
## age                257.7          22.5  11.453  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11560 on 1336 degrees of freedom
## Multiple R-squared:  0.08941,    Adjusted R-squared:  0.08872
## F-statistic: 131.2 on 1 and 1336 DF,  p-value: < 2.2e-16
```

3.4.1.1 Coefficient and Interpretation As shown in the R Output a person with the age of 0 will have a base charge of 3165.885. This intercept and its interpretation seems to be nonsensical. With each year the charges are increasing by 257.7226, which is the slope for this regressionline

3.4.1.2 P-values With a value of 2e-16 the age seems to have a very strong effect on the charges. This means that the slope of the charges is not flat, not zero, so the hypothesis has to be thrown away.

3.4.1.3 Including the gender as a second variable In this subset we will consider also the sex for modelling a simple regression model.

```
lm.insurance.2 <- lm(charges ~ age + sex, data = insurance)
summary(lm.insurance.2)
```

```
##
## Call:
## lm(formula = charges ~ age + sex, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8821  -6947  -5511   5443  48203
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2343.62     994.35   2.357  0.0186 *
## age          258.87      22.47  11.523 <2e-16 ***
## sexmale      1538.83     631.08   2.438  0.0149 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11540 on 1335 degrees of freedom
## Multiple R-squared:  0.09344,    Adjusted R-squared:  0.09209
## F-statistic: 68.8 on 2 and 1335 DF,  p-value: < 2.2e-16
```

3.4.1.4 Including interaction Since a second variable was added we want to explore if there might be significant interaction between age and sex, that might have to be considered for the modelling process.

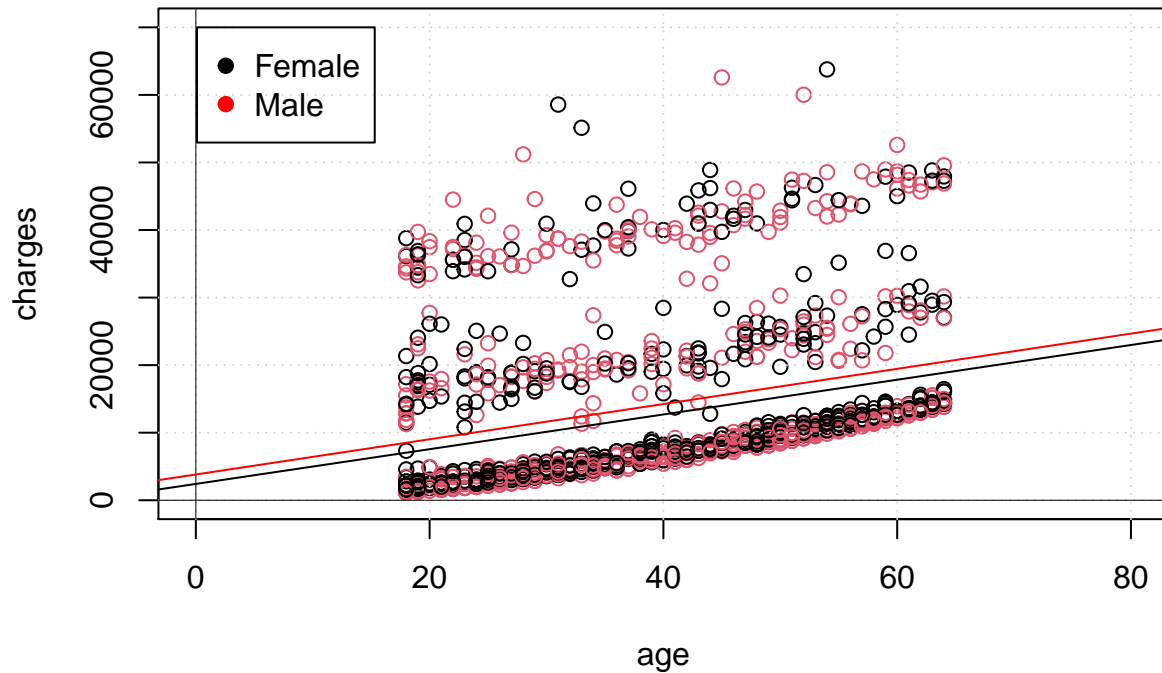
```
lm.insurance.3 <-lm(charges ~ age * sex, data =insurance)
summary(lm.insurance.3)
```

```
##
## Call:
## lm(formula = charges ~ age * sex, data = insurance)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8823  -6936  -5500   5456  48187
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2416.85    1339.35   1.804  0.0714 .
## age          257.01      31.95   8.045 1.89e-15 ***
## sexmale      1394.92    1872.30   0.745  0.4564
## age:sexmale    3.67      44.95   0.082  0.9349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11540 on 1334 degrees of freedom
## Multiple R-squared:  0.09345,    Adjusted R-squared:  0.09141
## F-statistic: 45.84 on 3 and 1334 DF,  p-value: < 2.2e-16
```

```
plot(charges ~ age, data = insurance,
     main = "Model 'lm.insurance.3'",
     xlim = c(0, 80),
     ylim = c(0, 70000),
     col = sex)
##
grid()
##
abline(h = 0, lwd = 0.5)
abline(v = 0, lwd = 0.5)
##
abline(a = coef(lm.insurance.3)[1],
       b = coef(lm.insurance.3)["age"])
abline(a = coef(lm.insurance.3)[1] + coef(lm.insurance.3)["sexmale"] ,
       b = coef(lm.insurance.3)["age"] + coef(lm.insurance.3)["age:sexmale"],
       col = "red")
legend(x = 0.1, y = 70000,
       pch = 19,
       legend = c("Female", "Male"),
       col = c("black", "red"))
```

Model 'lm.insurance.3'



```
coef(lm.insurance.3)
```

```
## (Intercept)      age      sexmale age:sexmale  
## 2416.848522  257.011490 1394.925331    3.669849
```

```
summary(lm.insurance.3)$coefficients
```

```
##              Estimate Std. Error   t value    Pr(>|t|)  
## (Intercept) 2416.848522 1339.35262  1.80449008 7.137997e-02  
## age         257.011490   31.94634  8.04509917 1.887603e-15  
## sexmale     1394.925331 1872.29663  0.74503437 4.563822e-01  
## age:sexmale    3.669849   44.95054  0.08164194 9.349437e-01
```

3.5 Final Linear Model Development

In this section we want to find an appropriate model, which accounts all relevant parameters and interactions. Afterwards we will then compare the fitted model with a base model and test its performance.

3.5.1 Linear Model with all variables

```
lm.insurance.all <- lm(charges ~ age + sex + bmi + children + smoker + region , data = insurance)  
summary(lm.insurance.all)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker +
##     region, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12973.56    1079.16  -12.022  < 2e-16 ***
## age          256.86      11.90    21.587  < 2e-16 ***
## sexmale     -131.31     332.95   -0.394  0.693348
## bmi         339.19      28.60    11.860  < 2e-16 ***
## children    475.50     137.80    3.451  0.000577 ***
## smoker1     23848.53    413.15   57.723  < 2e-16 ***
## region1      74.97     470.64    0.159  0.873460
## region2     1035.02     478.69    2.162  0.030782 *
## region3      682.06     478.96    1.424  0.154669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

As the above R Output shows not all variables seems to have a significant effect on the charges.

3.5.1.1 Testing sex before dropping Before removing those two variables we first will make a deeper analysis.

```
lm.sex <- lm(charges ~ sex, data=insurance)
summary(lm.sex)
```

```
##
## Call:
## lm(formula = charges ~ sex, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12835  -8435  -3980   3476   51201
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12569.6      470.1   26.740  <2e-16 ***
## sexmale     1387.2      661.3    2.098   0.0361 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12090 on 1336 degrees of freedom
## Multiple R-squared:  0.003282, Adjusted R-squared:  0.002536
## F-statistic: 4.4 on 1 and 1336 DF,  p-value: 0.03613
```



```
coef(lm.sex)
```

```
## (Intercept)      sexmale  
##    12569.579    1387.172
```

As seen in the output considering only the sex it is not a significant and standalone explaining variable for the charges. Comparing the p-value of the sex within the full model including all possible variables it is so high with a value of 0.693348, that it can be dropped from our final model.

```
lm.region.1 <- lm(charges ~ region, data = insurance)  
aggregate(charges ~ region, FUN = mean, data = insurance)
```

3.5.1.2 Testing region before dropping

```
##   region  charges  
## 1      0 14735.41  
## 2      1 12346.94  
## 3      2 13406.38  
## 4      3 12417.58
```

```
summary(lm.region.1)
```

```
##  
## Call:  
## lm(formula = charges ~ region, data = insurance)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -13614  -8463  -3793   3385   49035   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)  14735.4      633.3   23.266  <2e-16 ***  
## region1      -2388.5      922.2   -2.590   0.0097 **   
## region2      -1329.0      922.9   -1.440   0.1501        
## region3      -2317.8      922.2   -2.513   0.0121 *     
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 12080 on 1334 degrees of freedom  
## Multiple R-squared:  0.006634,    Adjusted R-squared:  0.0044   
## F-statistic:  2.97 on 3 and 1334 DF,  p-value: 0.03089
```

There is strong evidence that the mean charges for northeast is not equal to zero. But there isn't any evidence that all other regions differ from the reference region northeast. To have a better understanding we will make an anova test between the above shown model and a base model `lm.region.0` as shown below.

```
lm.region.0 <- lm(charges ~ 1 , data = insurance)
coef(lm.region.0)
```

```
## (Intercept)
##      13270.42
```

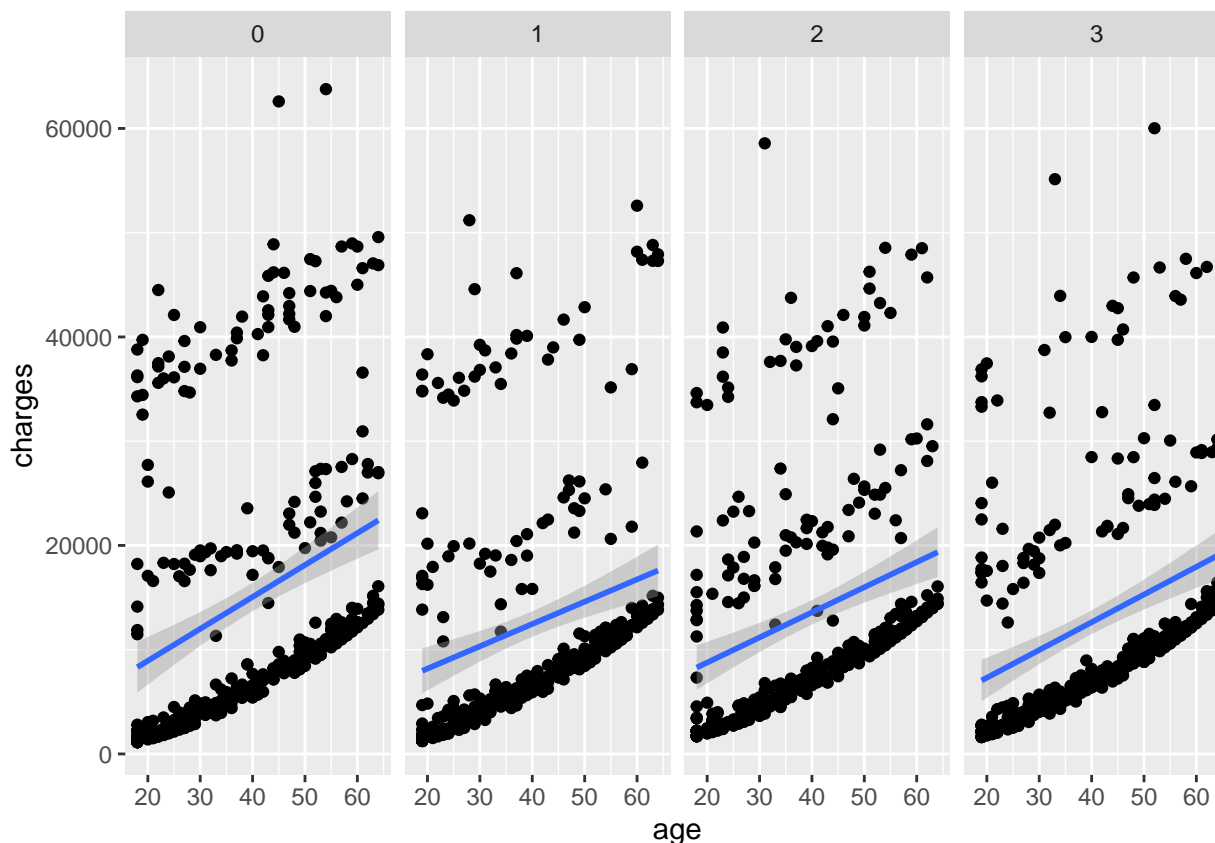
```
anova(lm.region.0, lm.region.1)
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ 1
## Model 2: charges ~ region
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1    1337 1.9607e+11
## 2    1334 1.9477e+11  3 1300759681 2.9696 0.03089 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova test shows that there is a low evidence that the model with more parameters (in this case only the region) better fits the data. The F-value seems to be very small and also the p-value is not really significant with 0.03089. Anyhow there is a drop in the RSS, by adding 3 additional parameters of the different regions. To have a better understanding we can additionally perform posthoc contrasts to decide afterwards if we will drop the region finally from our model. Before performing several posthoc test and repeating this exercise to all possible combinations we will use the ggplot to explore the data quickly upfront.

```
ggplot(data=insurance,
       mapping = aes(y = charges, x= age))+ geom_point()+geom_smooth(method = 'lm') + facet_grid(. ~reg
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



Using the visualization it seems that there seems to be no obvious difference and effect. Therefore we decide to drop also the variable region from our linear model.

3.5.2 Dropping variables from the model

```
drop1(lm.insurance.all, test="F")
```

```
## Single term deletions
##
## Model:
## charges ~ age + sex + bmi + children + smoker + region
##      Df Sum of Sq      RSS   AIC  F value    Pr(>F)
## <none>             4.8840e+10 23316
## age      1 1.7124e+10 6.5964e+10 23717  465.9837 < 2.2e-16 ***
## sex      1 5.7164e+06 4.8845e+10 23315   0.1556  0.693348
## bmi      1 5.1692e+09 5.4009e+10 23449  140.6627 < 2.2e-16 ***
## children 1 4.3755e+08 4.9277e+10 23326   11.9063  0.000577 ***
## smoker   1 1.2245e+11 1.7129e+11 24993 3331.9680 < 2.2e-16 ***
## region   3 2.3343e+08 4.9073e+10 23317   2.1173  0.096221 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lm.insurance.1 <- update(lm.insurance.all, .~. -region -sex)
formula(lm.insurance.1)
```

```
## charges ~ age + bmi + children + smoker
```

```
summary(lm.insurance.1)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11897.9  -2920.8   -986.6   1392.2  29509.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12102.77     941.98  -12.848 < 2e-16 ***
## age           257.85       11.90   21.675 < 2e-16 ***
## bmi           321.85       27.38   11.756 < 2e-16 ***
## children      473.50      137.79    3.436 0.000608 ***
## smoker1      23811.40     411.22   57.904 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6068 on 1333 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7489
## F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```

Comparing the summary output of `lm.insurance.0` and `lm.insurance.1` the latest model is only including variables having a p-value that indicates a significant factor.

3.5.3 Considering Interactions

```
drop1(lm.insurance.1, test="F")
```

```
## Single term deletions
##
## Model:
## charges ~ age + bmi + children + smoker
##      Df Sum of Sq      RSS   AIC F value    Pr(>F)
## <none>                 4.9078e+10 23315
## age      1 1.7297e+10 6.6375e+10 23717  469.789 < 2.2e-16 ***
## bmi      1 5.0884e+09 5.4167e+10 23445  138.203 < 2.2e-16 ***
## children 1 4.3477e+08 4.9513e+10 23325   11.809 0.0006077 ***
## smoker   1 1.2345e+11 1.7253e+11 24995 3352.911 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lm.insurance.2 <- update(lm.insurance.1, .~. +age:bmi + age:children + age:smoker + bmi:children + bmi:smoker)
summary(lm.insurance.2)
```

```
##
```

```
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + age:bmi +
##     age:children + age:smoker + bmi:children + bmi:smoker + children:smoker,
##     data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13996.3  -1947.6  -1331.5   -406.4   29570.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.745e+02  2.106e+03  -0.320    0.749
## age            2.055e+02  4.963e+01   4.140  3.7e-05 ***
## bmi           -6.339e+01  6.756e+01  -0.938    0.348
## children       6.957e+02  6.341e+02   1.097    0.273
## smoker1       -1.983e+04  1.861e+03 -10.651 < 2e-16 ***
## age:bmi        1.912e+00  1.561e+00   1.225    0.221
## age:children   1.201e+00  8.527e+00   0.141    0.888
## age:smoker1    -9.141e-01  2.384e+01  -0.038    0.969
## bmi:children   -5.334e+00  1.863e+01  -0.286    0.775
## bmi:smoker1     1.437e+03  5.317e+01  27.029 < 2e-16 ***
## children:smoker1 -3.858e+02  2.841e+02  -1.358    0.175
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4874 on 1327 degrees of freedom
## Multiple R-squared:  0.8392, Adjusted R-squared:  0.838
## F-statistic: 692.8 on 10 and 1327 DF, p-value: < 2.2e-16
```

As the above output shows not all of the added interactions have to be considered in the model. Upfront we also tried to generate a linear model including all possible interactions. Since the output was not satisfying we skipped this analysis at this point. Therefore the working assumption in this step was quite easy by editing only possible and simple interactions. Based on the results we will now drop the unnecessary interactions.

```
drop1(lm.insurance.2, test="F")
```

```
## Single term deletions
##
## Model:
## charges ~ age + bmi + children + smoker + age:bmi + age:children +
##     age:smoker + bmi:children + bmi:smoker + children:smoker
##              Df Sum of Sq      RSS   AIC  F value Pr(>F)
## <none>                3.1519e+10 22735
## age:bmi              1 3.5624e+07 3.1555e+10 22734   1.4998 0.2209
## age:children         1 4.7151e+05 3.1520e+10 22733   0.0199 0.8880
## age:smoker           1 3.4909e+04 3.1519e+10 22733   0.0015 0.9694
## bmi:children         1 1.9467e+06 3.1521e+10 22733   0.0820 0.7747
## bmi:smoker           1 1.7353e+10 4.8872e+10 23319 730.5705 <2e-16 ***
## children:smoker      1 4.3796e+07 3.1563e+10 22734   1.8439 0.1747
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

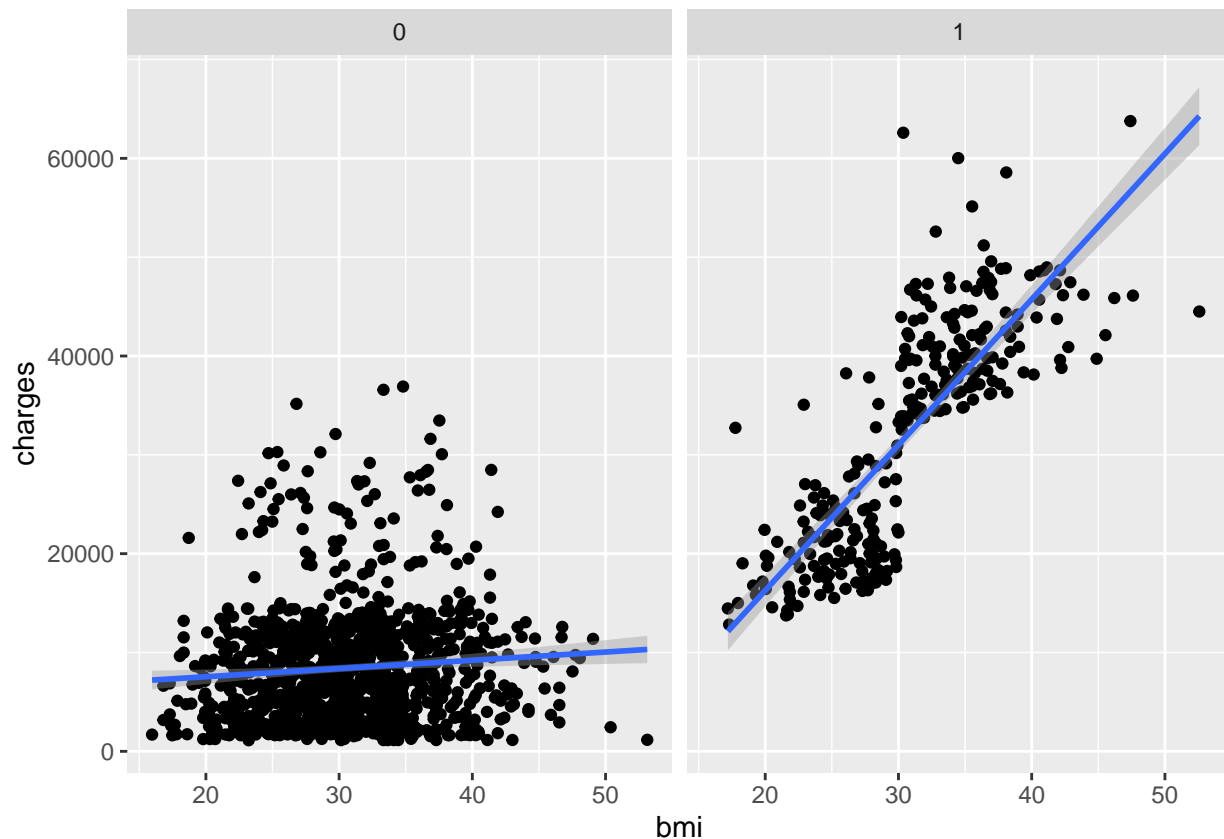
```
lm.insurance.3 <- update(lm.insurance.2, .~. -age:bmi - age:children - age:smoker - bmi:children -childr
summary(lm.insurance.3)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + bmi:smoker,
##     data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14598.6  -1924.4  -1321.4   -465.6  29892.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2729.002     831.270  -3.283  0.00105 **
## age           264.948       9.553   27.735 < 2e-16 ***
## bmi           5.656       24.873    0.227  0.82014
## children     508.924     110.615    4.601 4.61e-06 ***
## smoker1     -20194.709    1654.505  -12.206 < 2e-16 ***
## bmi:smoker1   1433.788      52.823   27.143 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4871 on 1332 degrees of freedom
## Multiple R-squared:  0.8388, Adjusted R-squared:  0.8382
## F-statistic: 1387 on 5 and 1332 DF, p-value: < 2.2e-16
```

After considering also the interaction between bmi and smoker, which seems to be significant with a small p-value of 2e-16. The variable bmi itself has now a p-value of 0.82014. Therefore lets have a look on the relationship between bmi and smoker to have a better understanding:

```
ggplot(data=insurance,
       mapping = aes(y = charges, x= bmi))+ geom_point()+ geom_smooth(method = 'lm') +facet_grid(. ~ sm

## 'geom_smooth()' using formula 'y ~ x'
```



As expected there seems to be a clear relationship between charges and bmi, considering if a person is smoking or not. Since this relationship makes even sense from the domain perspective, we will definitely keep this in our model.

3.5.4 Measure of Fit

In this section we were able to fit several models, having different levels of complexity. For the following analysis we have chosen our four models to measure the individual fit. Moreover we will compare them among each other to find the most appropriate model for the given insurance dataset.

```
formula(lm.insurance.all)
```

```
## charges ~ age + sex + bmi + children + smoker + region
```

```
summary(lm.insurance.all)$r.squared
```

```
## [1] 0.750913
```

```
summary(lm.insurance.all)$adj.r.squared
```

```
## [1] 0.7494136
```

```
formula(lm.insurance.1)
```

```
## charges ~ age + bmi + children + smoker
```

```
summary(lm.insurance.1)$r.squared
```

```
## [1] 0.7496945
```

```
summary(lm.insurance.1)$adj.r.squared
```

```
## [1] 0.7489434
```

```
formula(lm.insurance.2)
```

```
## charges ~ age + bmi + children + smoker + age:bmi + age:children +  
##      age:smoker + bmi:children + bmi:smoker + children:smoker
```

```
summary(lm.insurance.2)$r.squared
```

```
## [1] 0.8392491
```

```
summary(lm.insurance.2)$adj.r.squared
```

```
## [1] 0.8380378
```

```
formula(lm.insurance.3)
```

```
## charges ~ age + bmi + children + smoker + bmi:smoker
```

```
summary(lm.insurance.3)$r.squared
```

```
## [1] 0.8388379
```

```
summary(lm.insurance.3)$adj.r.squared
```

```
## [1] 0.8382329
```

For the comparison we used R-Squared and the adjusted R-Squared to measure the performance of our models. Since the adjusted R-squared can provide a more precise view of that correlation by also taking into account how many independent variables are added to our particular models against we will base our conclusion on this parameter.

Therefore we are happy to state that the latest model number three is able to explain the charges with a percentage of 83.82329 % based on the independent variables.

Comparing the latest model with the first one there is an increase of 8.88% in the adjusted R-squared. Even if the latest model is performing the best compared to the others, it is always a trade off between the gain in the fit and the corresponding effort.

4 Linear models (GAM & Polynomial) -> Yvonne

5 GLM and cross validation -> Carole

5.1 Generalised Linear Models for count data

5.1.1 Original data

The number of children an insured person has is analysed. We have the following data on children per person. The number of children ranges from 0 to 5 with a median of 1.

```
summary(insurance$children)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.000   0.000   1.000   1.095   2.000   5.000
```

5.1.2 Poisson model

To model count data (number of children) the poisson model is used. An analysis performed beforehand showed that only the variables “charges” and “smoker” have a significant impact on the number of children.

```
glm.children <- glm(children ~ smoker+charges,
                    data=insurance,
                    family = "poisson")
```

```
summary(glm.children)
```

```
##
## Call:
## glm(formula = children ~ smoker + charges, family = "poisson",
##      data = insurance)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.8561  -1.4318  -0.1057   0.7768   2.9717
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.706e-02  4.213e-02  -0.880   0.3790
## smoker1     -3.239e-01  1.058e-01  -3.061   0.0022 **
## charges      1.419e-05  3.365e-06   4.217 2.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 2001.6  on 1337  degrees of freedom
## Residual deviance: 1984.1  on 1335  degrees of freedom
## AIC: 3879.4
##
## Number of Fisher Scoring iterations: 5
```

To get the coefficients, the log transformation needs to be reversed:

```
exp(coef(glm.children))
```

```
## (Intercept)      smoker1      charges  
##    0.9636169    0.7233085    1.0000142
```

Smoker (factor): The model shows that for the factor smoker (yes/no), a smoker has on average 72% of the number of children a non-smoker has. The more common-sense interpretation might be the other way around, that people who have 1 or more children smoke less, but for the moment we have no proof of that.

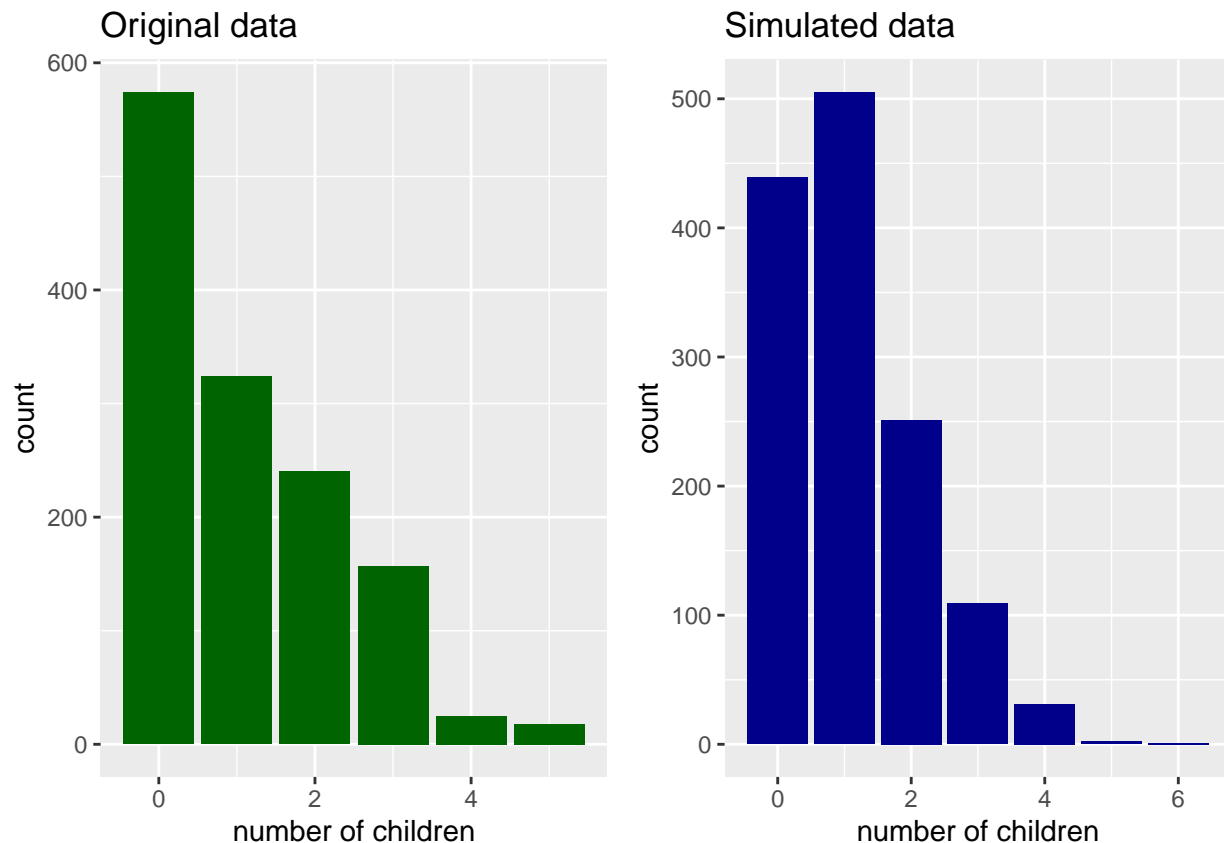
Charges: A person with higher charges will on average have more children. If charges are increased by 1000 dollars, the calculated number of children increases by 1.4%.

5.1.3 Simulation of data and comparison

With the calculated model, data is simulated:

```
##      sim_1  
## Min.    :0.000  
## 1st Qu.:0.000  
## Median :1.000  
## Mean    :1.102  
## 3rd Qu.:2.000  
## Max.    :6.000
```

The original and the simulated data are compared visually. The number of children from the simulated data (0-6) seem to be plausible. The distribution has a strong downwards trend starting at 1 like the original data. However the model does not seem to generate enough data with 0 children.



5.2 Generalised Linear Models for binomial data

A model is fitted that predicts if a person is a smoker or not. Only the significant values age, bmi and charges are used.

```
glm.smoker.2 <- glm(smoker ~ age+bmi+charges,
  data=insurance,
  family = "binomial")
```

```
summary(glm.smoker.2)
```

```
##
## Call:
## glm(formula = smoker ~ age + bmi + charges, family = "binomial",
##     data = insurance)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.09442  -0.10998  -0.04475  -0.00970   1.53727
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.311e+00  1.029e+00   5.163 2.43e-07 ***
## age         -9.875e-02  1.300e-02  -7.597 3.02e-14 ***
```

```
## bmi          -3.481e-01  4.309e-02  -8.078 6.60e-16 ***
## charges       3.822e-04  2.917e-05  13.104 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 1356.63  on 1337  degrees of freedom
## Residual deviance: 311.98  on 1334  degrees of freedom
## AIC: 319.98
##
## Number of Fisher Scoring iterations: 8
```

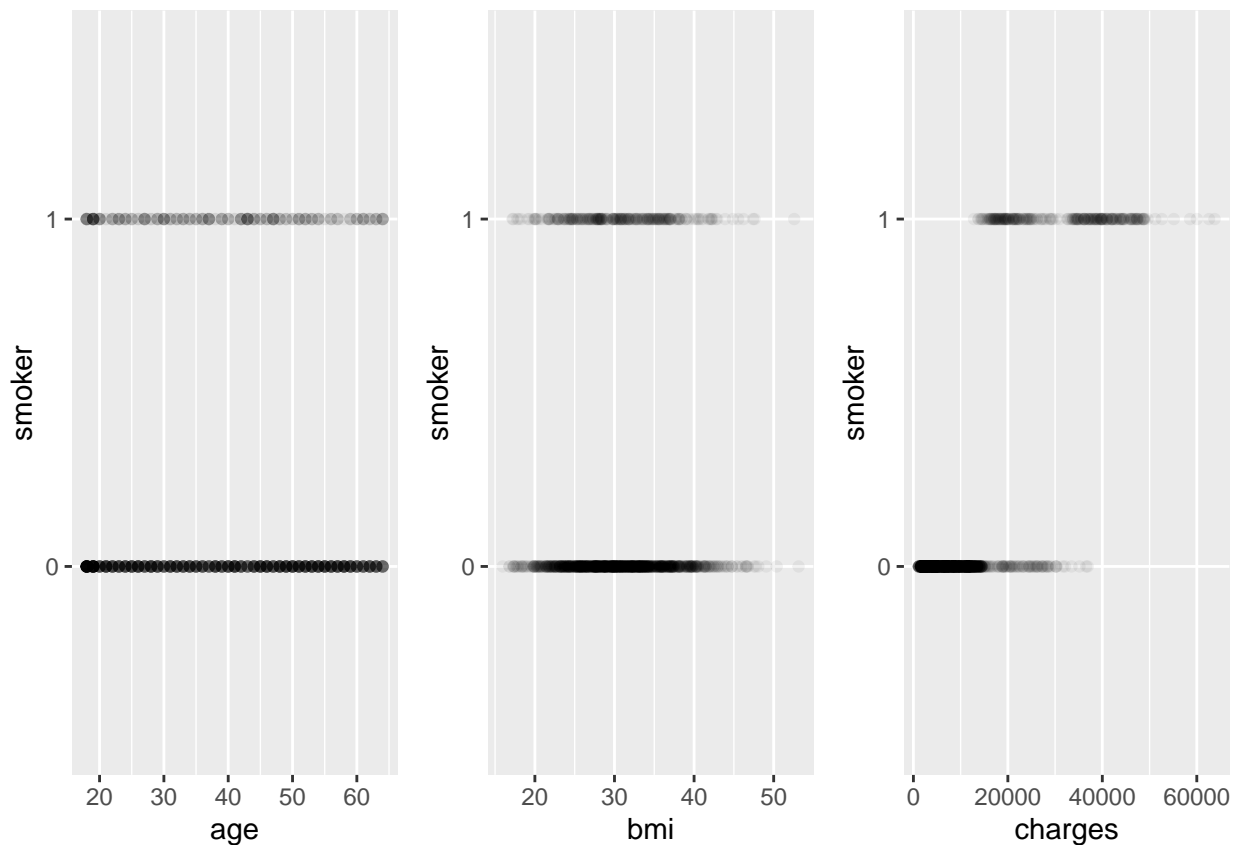
```
exp(coef(glm.smoker.2))
```

```
## (Intercept)      age      bmi      charges
## 202.5686249    0.9059677    0.7060520    1.0003823
```

Age and BMI has a negative effect on smoker. This means the higher a persons BMI and age, the lower the probability that the person is a smoker. Charges has a positive effect. This means the higher a persons charges, the higher is the possibility that the person smokes.

5.2.1 Graphical analysis

This can also be explored graphically, at least for charges it is clearly visible that smokers have higher charges.



5.2.2 Estimating the model performance

The predicted values are transformed into binary (beforehand they indicated the probability) and compared with the actual data.

```
fitted.smoker.disc <- ifelse(fitted(glm.smoker.2) < 0.5,  
                             yes = 0, no = 1)  
head(fitted.smoker.disc)
```

```
## 1 2 3 4 5 6  
## 1 0 0 1 0 0
```

```
d.obs.fit.smoker <- data.frame(obs = insurance$smoker,  
                               fitted = fitted.smoker.disc)  
head(d.obs.fit.smoker)
```

```
##   obs fitted  
## 1    1      1  
## 2    0      0  
## 3    0      0  
## 4    0      1  
## 5    0      0  
## 6    0      0
```

We observe the following fit:

```
##      fit  
## obs    0    1  
##   0 1028   36  
##   1   23  251
```

5.3 Cross Validation

Three linear models are cross validated:

```
lm.1 <- lm(data=insurance, charges~children+smoker+bmi+age+region+sex)  
lm.2 <- lm(data=insurance, charges~children+smoker)  
lm.3 <- lm(data=insurance, charges~(poly(bmi, degree=3))  
           +(poly(age, degree=2))+children+smoker)
```

The in sample performance, using R Squared as measure is the following:

```
summary(lm.1)$r.squared
```

```
## [1] 0.750913
```

```
summary(lm.2)$r.squared
```

```
## [1] 0.6236038
```

```
summary(lm.3)$r.squared
```

```
## [1] 0.754424
```

The out of sample performance is computed by using 50:50 training and test data and repeating the process 100 times.

```
set.seed(5)

r.squared.lm.1 <- c()
r.squared.lm.2 <- c()
r.squared.lm.3 <- c()

for(i in 1:100){

  # prepare data

  train.YES <- sample(x=c(TRUE,FALSE),
                     size=nrow(insurance),
                     replace = TRUE)

  table(train.YES)

  insurance.train <- insurance[train.YES, ]
  insurance.test <- insurance[!train.YES, ]

  # fit model with train data

  lm.1.train <- lm(formula = formula(lm.1),
                  data = insurance.train)

  lm.2.train <- lm(formula = formula(lm.2),
                  data = insurance.train)

  lm.3.train <- lm(formula = formula(lm.3),
                  data = insurance.train)

  # make prediction on test data
  lm.1.predict <- predict(lm.1.train,
                        newdata = insurance.test)

  lm.2.predict <- predict(lm.2.train,
                        newdata = insurance.test)

  lm.3.predict <- predict(lm.3.train,
                        newdata = insurance.test)

  # compute r.squared and save in list
```

```

r.squared.lm.1[i] <- cor(lm.1.predict,
                        insurance.test$charges)^2

r.squared.lm.2[i] <- cor(lm.2.predict,
                        insurance.test$charges)^2

r.squared.lm.3[i] <- cor(lm.3.predict,
                        insurance.test$charges)^2
}

```

The out of sample performance, using R Squared as measure is the following:

```

#lm.1
mean(r.squared.lm.1)

```

```
## [1] 0.7474233
```

```

#lm.2
mean(r.squared.lm.2)

```

```
## [1] 0.6235398
```

```

#lm.3
mean(r.squared.lm.3)

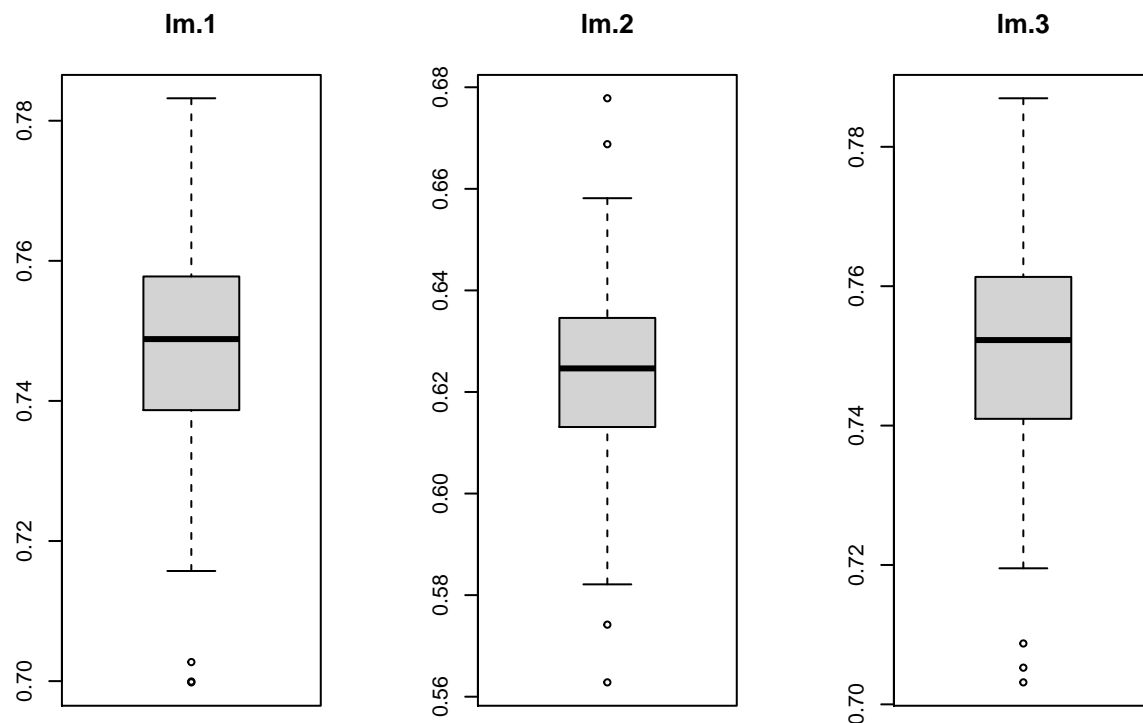
```

```
## [1] 0.7505379
```

```

par(mfrow=c(1,3))
boxplot(r.squared.lm.1, main="lm.1")
boxplot(r.squared.lm.2, main="lm.2")
boxplot(r.squared.lm.3, main="lm.3")

```



It can be observed that lm.3 performs slightly better but it is also the most complicated. lm.1 might be the better model, the performance is just slightly lower and it is simpler.

6 Decision Trees -> Adriana

7 Support Vector Machines -> Carole

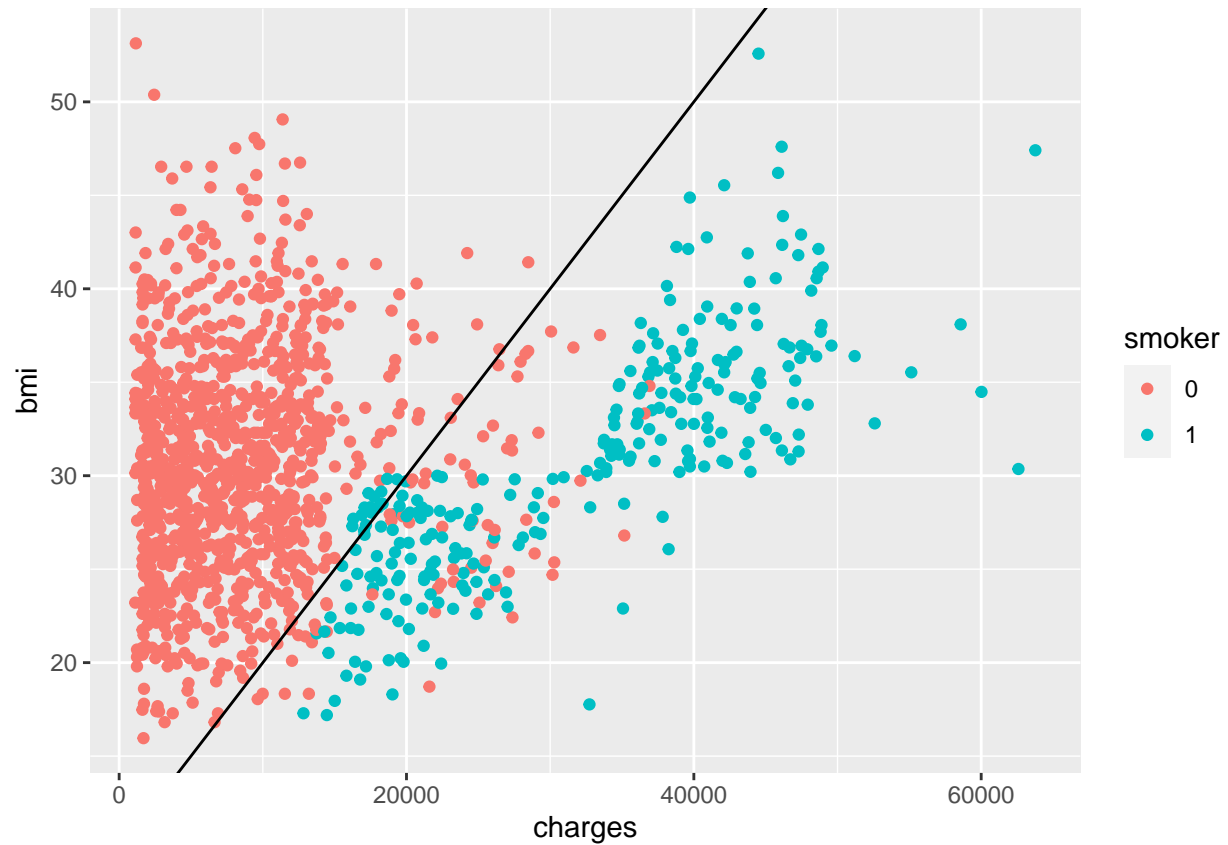
In this chapter, Support Vector Machines are applied to the insurance dataset.

7.1 two classes linear

By visual exploration, we discover that smokers and non-smokers form two groups. We fit an SVM for those two groups. The lines in the graph are the first guess of how the data could be divided.

```
# guess of hyperplane
slope <- 0.001
intercept <- 10

ggplot(data = insurance) +
  geom_point(aes(x = charges, y = bmi, color=smoker))+
  geom_abline(slope = slope, intercept = intercept)
```

A linear SVM is tuned using different cost ranges.

```
set.seed(5)
#get optimal cost
cost_range <-
  c(1e-10,
    1e-7,
    1e-5,
    0.001,
    0.0025,
    0.005,
    0.0075,
    0.01,
    0.1,
    1,
    5,
    10,
    100,
    200)

tune.out.1 <- tune(
  svm,
  smoker ~ charges+bmi,
  data = insurance.train,
  kernel = "linear",
  ranges = list(cost = cost_range)
```

```
)
```

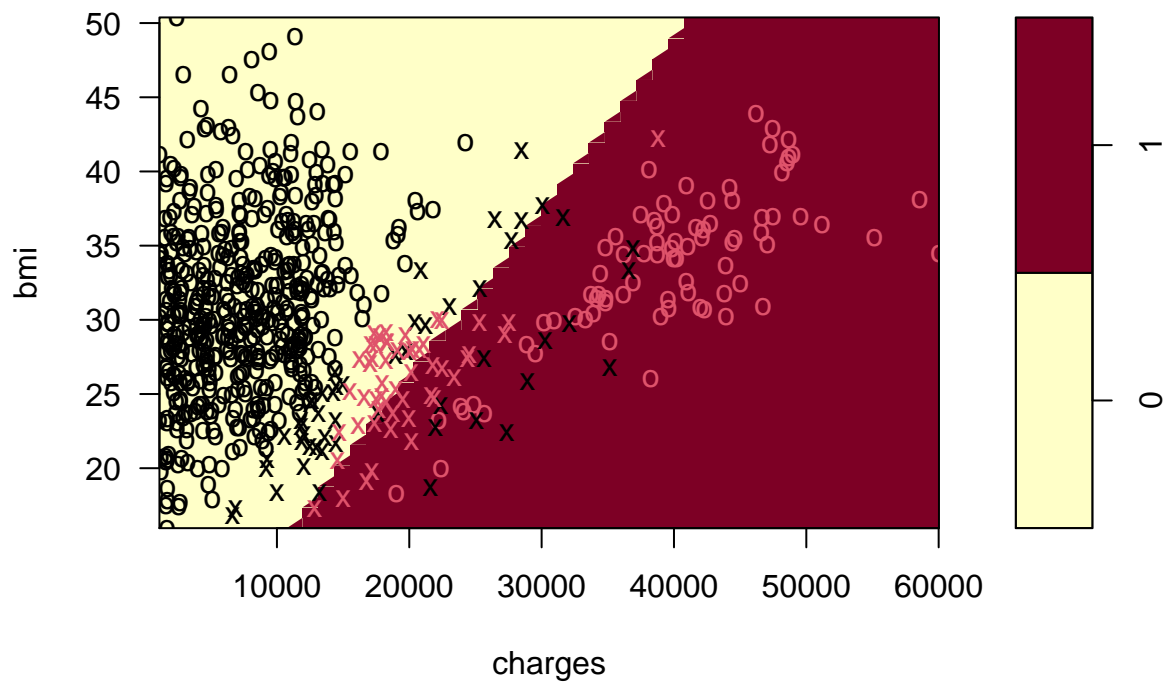
The following is the best model. A cost of 5 is used. There are 104 support vectors.

```
#best model
bestmod <- tune.out.1$best.model
summary(bestmod)

##
## Call:
## best.tune(method = svm, train.x = smoker ~ charges + bmi, data = insurance.train,
##   ranges = list(cost = cost_range), kernel = "linear")
##
##
## Parameters:
##   SVM-Type:  C-classification
##   SVM-Kernel: linear
##         cost: 5
##
## Number of Support Vectors: 104
##
## ( 52 52 )
##
##
## Number of Classes: 2
##
## Levels:
## 0 1

plot(bestmod, insurance.train, bmi~charges)
```

SVM classification plot



The confusion matrices and the classification error rate for train and test data are:

```
# confusion matrix train
confusion.lin.train <- table(predict = predict(bestmod, insurance.train),
                             truth = insurance.train$smoker)
confusion.lin.train
```

```
##          truth
## predict   0   1
##          0 514 26
##          1  18 106
```

```
#classification error rate train
(confusion.lin.train[1,2]+confusion.lin.train[2,1])/sum(confusion.lin.train[1:2,1:2])
```

```
## [1] 0.06626506
```

```
# confusion matrix linear test
confusion.lin.test <- table(predict = predict(bestmod, insurance.test),
                             truth = insurance.test$smoker)
confusion.lin.test
```

```
##          truth
## predict   0   1
##          0 509 18
##          1  23 124
```

```
#classification error rate test
(confusion.lin.test[1,2]+confusion.lin.test[2,1])/sum(confusion.lin.test[1:2,1:2])
```

```
## [1] 0.06083086
```

7.2 two classes polynomial

A polynomial model was also tested but the classification error rate was slightly higher:

```
set.seed(5)
#get optimal cost and degree for polynomial

cost_range <-
  c(1e-10,
    1e-7,
    1e-5,
    0.001,
    0.0025,
    0.005,
    0.0075,
    0.01,
    0.1,
    1,
    5,
    10,
    100)

degree_range <- 2:4

tune.out.poly <- tune(
  svm,
  smoker ~ charges+bmi,
  data = insurance.train,
  kernel = "polynomial",
  ranges = list(cost = cost_range, degree = degree_range))

#best model poly
tune.out.poly$best.parameters
```

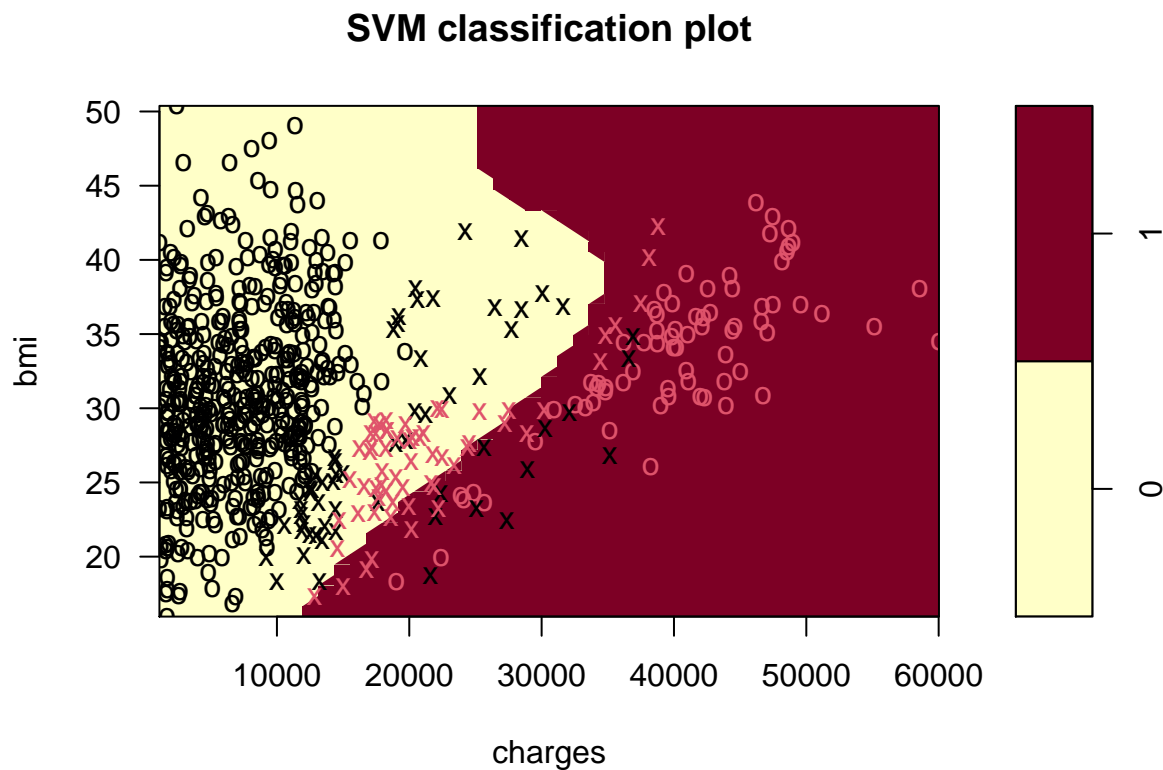
```
##      cost degree
## 26   100      3
```

```
bestmod.poly <- tune.out.poly$best.model
summary(bestmod.poly)
```

```
##
## Call:
## best.tune(method = svm, train.x = smoker ~ charges + bmi, data = insurance.train,
##      ranges = list(cost = cost_range, degree = degree_range), kernel = "polynomial")
##
##
```

```
## Parameters:
##   SVM-Type: C-classification
##   SVM-Kernel: polynomial
##     cost: 100
##     degree: 3
##     coef.0: 0
##
## Number of Support Vectors: 120
##
## ( 60 60 )
##
## Number of Classes: 2
##
## Levels:
## 0 1
```

```
plot(bestmod.poly, insurance.train, bmi~charges)
```



```
# confusion matrix
confusion.poly.train <- table(predict = predict(bestmod.poly, insurance.train),
                              truth = insurance.train$smoker)

confusion.poly.test <- table(predict = predict(bestmod.poly, insurance.test),
                              truth = insurance.test$smoker)
```

```
#classification error rate
```

```
(confusion.poly.train[1,2]+confusion.poly.train[2,1])/sum(confusion.poly.train[1:2,1:2])
```

```
## [1] 0.0753012
```

```
(confusion.poly.test[1,2]+confusion.poly.test[2,1])/sum(confusion.poly.test[1:2,1:2])
```

```
## [1] 0.06379822
```

8 Neural Networks

9 Cross validation and discussion