UNIVERSITÀ DEGLI STUDI DI PADOVA

Department of Mathematics "Tullio Levi-Civita"

Master's Degree in Data Science

INSURANCE CHARGES PREDICTION AND DATA ANALYSIS

Project in Statistical Learning

Carlo De Dominicis 2026816
Nicolò Malatesta 2026504

Academic Year 2021/2022

Contents

1 Introduction

A health insurance company can only make money if it collects more than what it spends on the medical care of its beneficiaries. On the other hand, even though some conditions are more prevalent for certain segments of the population, medical costs are difficult to predict since most money comes from rare conditions of the patients. The aim of this project is to first analyze the factors that influence medical costs by explorating the dataset and all the components in order to discover correlations between datas, and secondly try to build an adequate model that can accurately predict insurance costs bsed on the data and optimize its performance.

2 Obtaining Data

This dataset is in public domain (available on https://github.com/stedy/Machine-Learning-with-R-datasets or https://www.kaggle.com/mirichoi0218/insurance), provided from "Machine Learning with R" by Brett Lantz, this is a clean dataset, as we will see in the next paragraph.

The cost of treatment depends on many factors: diagnosis, type of clinic, city of residence, age and so on. We have no data on the diagnosis of patients. But we have other information that can help us to make a conclusion about the health of patients and practice regression analysis. Nonetheless, it is good to have an understanding of what they are. Here are some factors collected by insurance, on which we will study the influence on the cost of medical insurance premiums: We have a dataset that includes 1338 observations on 7 variables

2.1 Variables description

- AGE: age of primary beneficiary;
- SEX: insurance contractor's gender (female or male);
- BMI: body mass index which is expressed as the ratio between weight and square of an individual's height and is used as an indicator of the state of healthy weight (kg / m ^ 2). The ideal weight is excellent from 18.5 to 24.9;
- CHILDREN: Number of children covered by health insurance;
- SMOKER: Smoking/ Non-smoking
- REGION: The beneficiary's residential area in the USA (northeast, southeast, southwest, northwest);
- CHARGES: Individual medical costs billed by health insurance;

[2]:
```
data <- read.csv(file = "../input/insurance/insurance.csv")
glimpse(data)
```

```
Rows: 1,338
Columns: 7
$ age      <int> 19, 18, 28, 33, 32, 31, 46, 37, 37, 60, 25,
62, 23, 56, 27, ...
$ sex      <fct> female, male, male, male, male, female,
female, female, male...
$ bmi      <dbl> 27.900, 33.770, 33.000, 22.705, 28.880,
25.740, 33.440, 27.7...
$ children <int> 0, 1, 3, 0, 0, 0, 1, 3, 2, 0, 0, 0, 0, 0, 0,
```

```
1, 1, 0, 0, 0, ...
$ smoker   <fct> yes, no, no, no, no, no, no, no, no, no, no,
yes, no, no, ye...
$ region   <fct> southwest, southeast, southeast, northwest,
northwest, south...
$ charges  <dbl> 16884.924, 1725.552, 4449.462, 21984.471,
3866.855, 3756.622...
```

3 Clean and filter data

Let's check if there are any duplicated observations on train dataset.

[3]:
```
cat("Cheking for duplicated rows...","The dataset has",sum(duplicated(data)) ,␣
 ↪"duplicated rows.\n")
cat("Checking for NA values...", "The dataset has", sum(is.na(data)), "null␣
 ↪values\n")

cat("\nThe duplicated row is:")
data[duplicated(data),]
```

```
Cheking for duplicated rows... The dataset has 1 duplicated rows.
Checking for NA values... The dataset has 0 null values

The duplicated row is:
```

	age	sex	bmi	children	smoker	region	charges
582	19	male	30.59	0	no	northwest	1639.563

There is one. It's unlikely that two people have the same age, sex, BMI, and children from the same region, both non-smokers, and have exactly the same medical charges. We can drop this duplicated row.

[4]:
```
data <- data %>% distinct()
```

4 Exploratory data analysis

In this section we are going to explore the given dataset, trying to dig up some usefull information hidden between the variables, in order to build an effective model able to make predictions on the insurance charges. Moreover in some cases we will make use of some statistical tests to try explain the significance of some behaviours on the data, in particular:

- *F-test* to check the homogenity of two population's variances.
- *T-test* to verify possible significative differences between the means of two sets of data.
- *Bartlett's Test*, to test homoscedasticity between samples of the same populations.
- *ANOVA Test*, for the equality of two or more population's means

For all the cited tests we will take in consideration a significance level $\alpha = 0.05$

4.0.1 Descriptive statistic:

[5]:
```
summary(data)
```

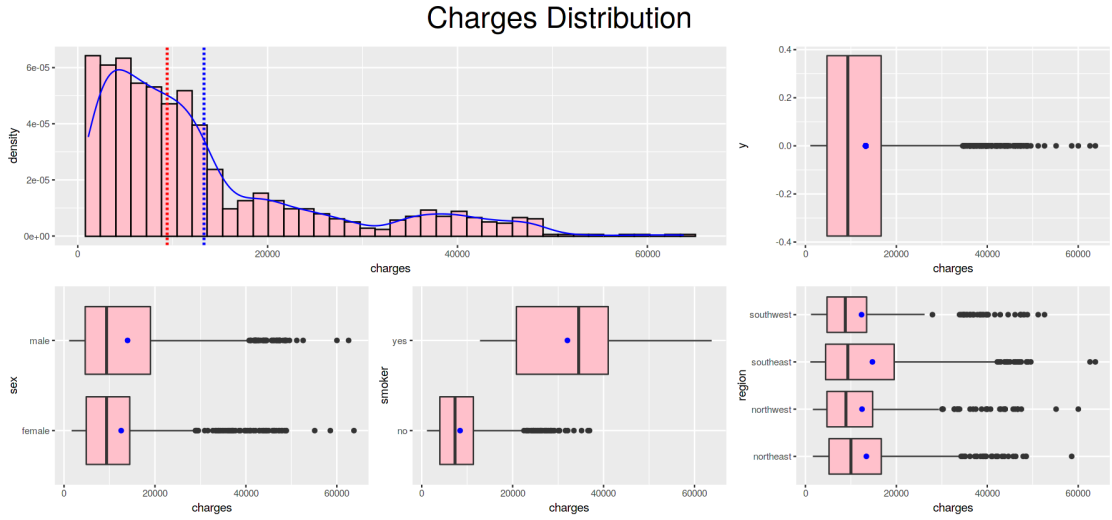
```
      age            sex            bmi           children     smoker
 Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1063
 1st Qu.:27.00   male  :675   1st Qu.:26.29   1st Qu.:0.000   yes: 274
 Median :39.00                Median :30.40   Median :1.000
 Mean   :39.22                Mean   :30.66   Mean   :1.096
 3rd Qu.:51.00                3rd Qu.:34.70   3rd Qu.:2.000
 Max.   :64.00                Max.   :53.13   Max.   :5.000
       region       charges
 northeast:324   Min.   : 1122
 northwest:324   1st Qu.: 4746
 southeast:364   Median : 9386
 southwest:325   Mean   :13279
                 3rd Qu.:16658
                 Max.   :63770
```

In terms of categorical features, the dataset has a similar number of people for each category, except for smokers. We have more non-smokers than smokers. The charges itself varies greatly from around $\$1,000$ to $\$64,000$.

4.1 Distributions on categorical varables

[6]:
```
p1 <- ggplot(data, aes(x=charges)) +
    geom_histogram(aes(y=..density..),color="black", fill="pink", bins=40) +
    geom_density(color="blue") +
    geom_vline(aes(xintercept= mean(charges)), color="blue", linetype="dashed",␣
 ↪size=1) +
    geom_vline(aes(xintercept= median(charges)), color="red", linetype="dashed",␣
 ↪size=1)

p2 <- ggplot(data, aes(x = charges)) +
      geom_boxplot(fill = "pink") +
        geom_point(aes(x= mean(charges), y=0), color="blue")

p3 <- ggplot(data, aes(x=sex, y=charges, group=sex)) +
    geom_boxplot(fill="pink") +
    stat_summary(fun=mean, geom="point", color="blue") +
    coord_flip()

p4 <- ggplot(data, aes(x=smoker, y=charges, group=smoker)) +
    geom_boxplot(fill="pink") +
    stat_summary(fun=mean, geom="point", color="blue") +
    coord_flip()

p5 <- ggplot(data, aes(x=region, y=charges, group=region)) +
```

```
    geom_boxplot(fill="pink") +
    stat_summary(fun=mean, geom="point", color="blue") +
    coord_flip()


options(repr.plot.width=15, repr.plot.height=7)
layout<- "AAB \n CDE"
p1 + p2 + p3 + p4 + p5 + plot_layout(design = layout) +
    plot_annotation(title="Charges Distribution", theme = theme(plot.title =␣
 ↪element_text(size = 28, hjust = 0.5)))
```



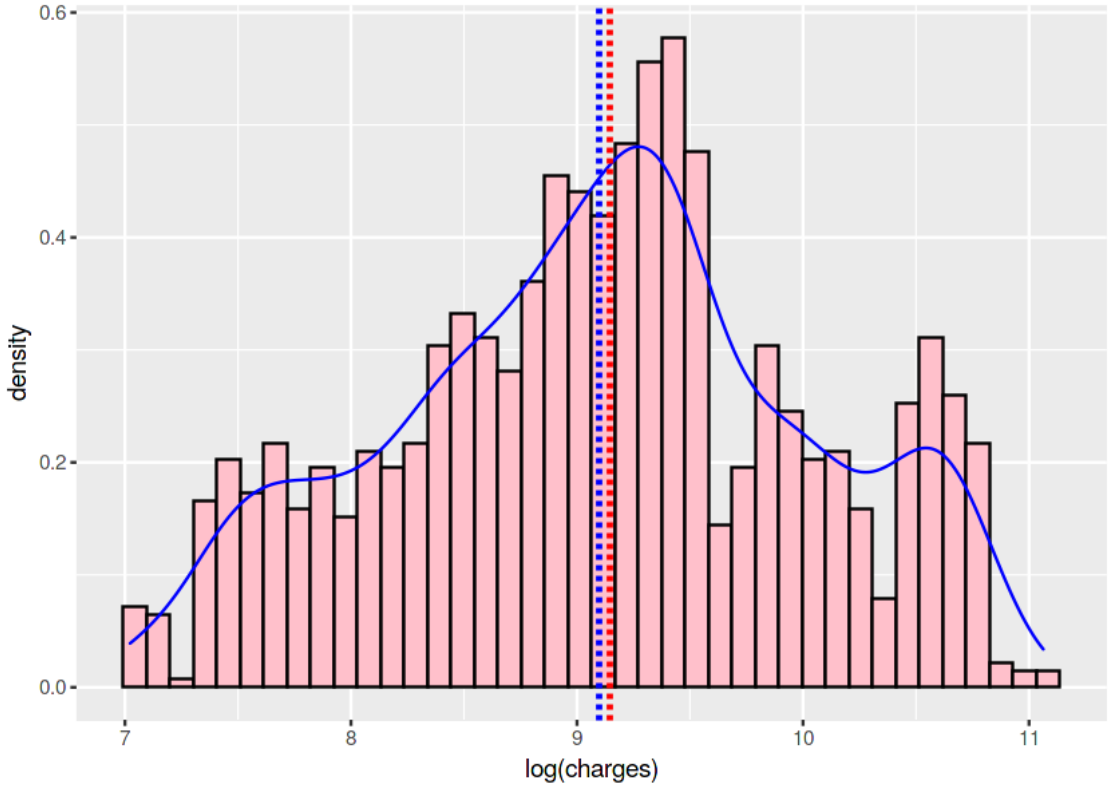
The distribution is right-skewed with a long tail to the left. There's a bump at around $40,000, perhaps another hidden distribution. To dig this up, we need categorical features that we are going to explore.

There does not seems to be a significant difference in the median charges due to variables Sex and Region if not some slightly difference in charges between males and females, and between the southest region and the others. However smoking habit is a very important factor determining charge. The median charge for smokers is more than double the median insurance charges of non smokers.

Below a plot for the log transformation of the charges: It helps have a normal distribution which could help us in a number of different ways such as outlier detection, and for our predictive modell in the next section.

[7]:
```
options(repr.plot.width=7, repr.plot.height=5)
ggplot(data, aes(x=log(charges))) +
    geom_histogram(aes(y=..density..),color="black", fill="pink", bins=40) +
    geom_density(color="blue") +
    geom_vline(aes(xintercept= mean(log(charges))), color="blue",␣
 ↪linetype="dashed", size=1) +
```

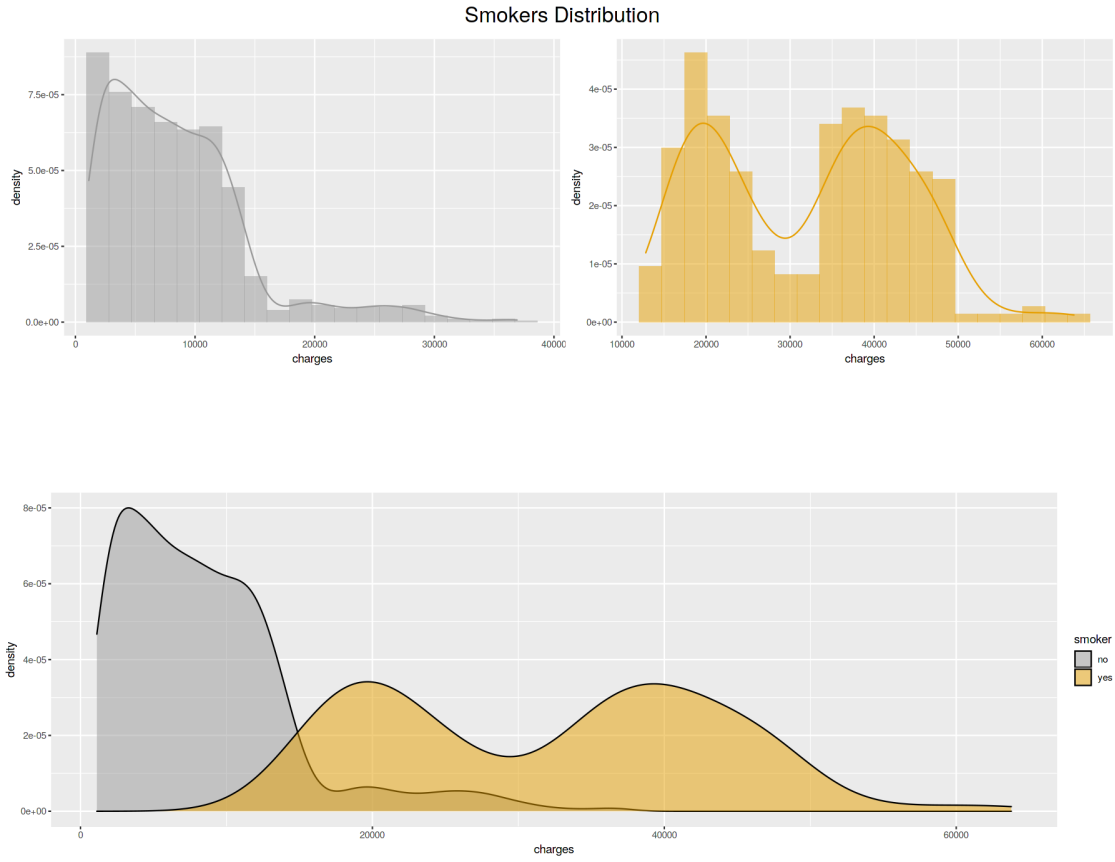
```
    geom_vline(aes(xintercept= median(log(charges))), color="red",␣
↪linetype="dashed", size=1)
```



Let's draw again the distribution of charges, now categorizing them into smoker.

[8]:
```
options(repr.plot.width=15, repr.plot.height=5)

smokers.color <- c("#999999", "#E69F00")

p1 <- ggplot(data[data$smoker == "no",], aes(x=charges, y= ..density..))+
    geom_histogram(fill=smokers.color[1],bins=20, alpha=.5)+
    geom_density(color=smokers.color[1])

p2 <- ggplot(data[data$smoker == "yes",], aes(x=charges, y= ..density..))+
    geom_histogram(fill=smokers.color[2],bins=20, alpha=.5)+
    geom_density(color=smokers.color[2])

p3 <- ggplot(data, aes(x = charges, fill = smoker)) +
      geom_density(alpha = 0.5) +
     scale_fill_manual(values=smokers.color)

p1 + p2 +
```

```
    plot_annotation(title="Smokers Distribution", theme = theme(plot.title =␣
 ↪element_text(size = 20, hjust = 0.5)))
p3
```



Smokers definitely have more charges than non-smokers. That is a clear explanation of the bump in charges we have seen before.

About the sligtly differences of charges in gender and region mentioned above, below we are going to make some hypothesis testing to see if these are relevant factors.

4.1.1 Hypothesis test 1:

We want to test the equality in the means of male and female smokers performing a statistical t-test. Since the salmples is enough large we assume the normality of the distribution. Before checking for the means, this test requires the two populations to be equal in the variance, so, a F-statistic will be perfomed taking in consideration: σ_m the standard deviation of charge costs for male smokers, and σ_f the standard deviation of charger costs for female smokers and let us define the two hypothesis:

- $H_0 : \sigma_m = \sigma_f$
- $H_1 : \sigma_m \neq \sigma_f$

[9]:
```
male_smokers <- data %>%
        group_by(sex) %>%
        filter(sex == "male", smoker == "yes")

female_smokers <- data %>%
        group_by(sex) %>%
        filter(sex == "female", smoker == "yes")

males <- male_smokers$charges
females <- female_smokers$charges

var.test(males, females, alternative="two.sided")
```

```
        F test to compare two variances

data:  males and females
F = 0.88511, num df = 158, denom df = 114, p-value = 0.4762
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6256613 1.2402158
sample estimates:
ratio of variances
         0.8851142
```

The p-value for the F-test is higher than the p-value, so we can accept the null hypothesis for which we have equal variance and perform a t-test on the means.
Let us define μ_m the mean of charge costs for smoker males, and μ_f the mean of charger costs for smoker females

- $H_0 : \mu_m \leq \mu_f$
- $H_1 : \mu_m > \mu_f$

[10]:
```
t.test(males, females, alternative = "greater", var.equal = FALSE)
```

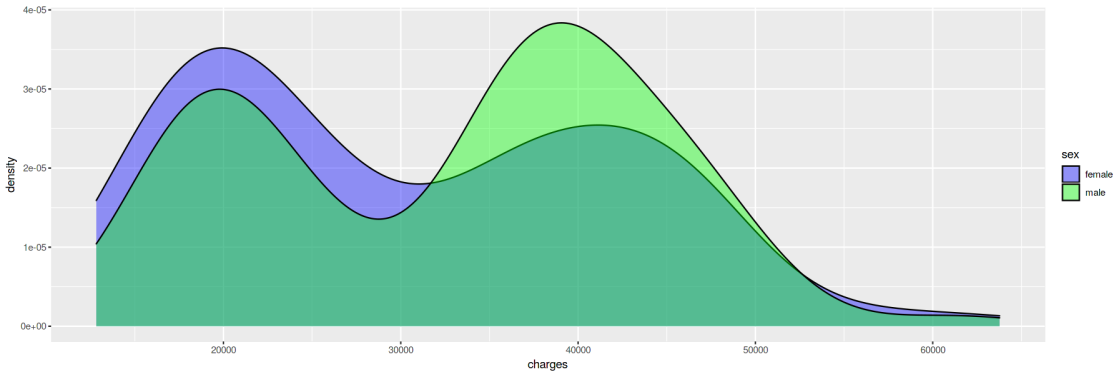
```
        Welch Two Sample t-test

data:  males and females
t = 1.6617, df = 236.69, p-value = 0.04895
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 14.73576      Inf
sample estimates:
mean of x mean of y
 33042.01  30679.00
```

The **p-value** of the test is a little less than the significance level $\alpha = 0.05$. So we can conclude that

men smoker's average in charges is significantly greater than the women smoker's one.

Then H_0 is rejected.

[11]:
```
ggplot(data[data$smoker == "yes",], aes(x = charges, fill = sex)) +
      geom_density(alpha = 0.4) +
      scale_fill_manual(values=c("blue", "green"))
```



And there we can see a bump for males compared to female smokers where the difference is pretty clear.

4.1.2 Hypothesis test 2

Let us perform a Bartlett Test to check the homogenity of the variance of charges between the regions before going to test the equality on the mean:

- $H_0 : \sigma_{ne} = \sigma_{nw} = \sigma_{se} = \sigma_{sw}$

[12]:
```
bartlett.test(charges ~ region, data = data)
```

```
        Bartlett test of homogeneity of variances

data:  charges by region
Bartlett's K-squared = 25.86, df = 3, p-value = 1.02e-05
```

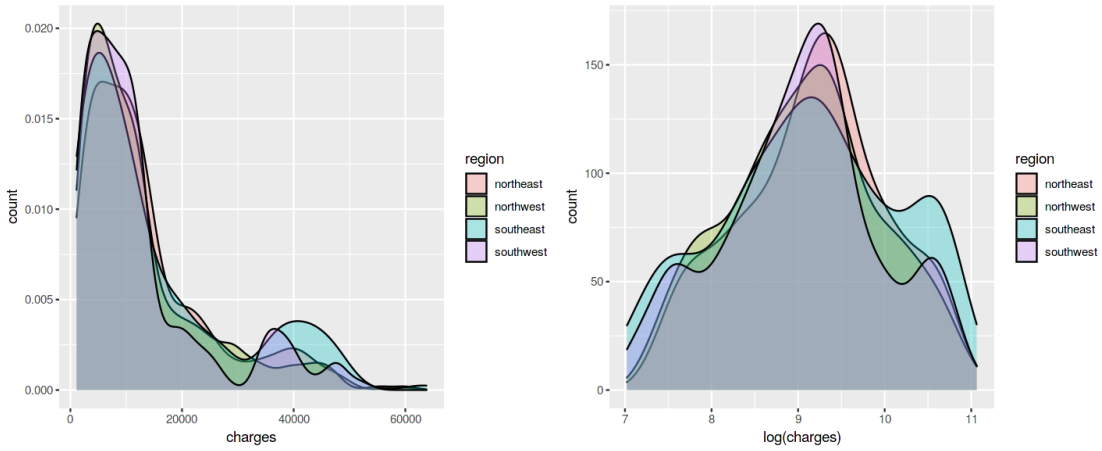
[13]:
```
bartlett.test(log(charges) ~ region, data = data)
```

```
        Bartlett test of homogeneity of variances

data:  log(charges) by region
Bartlett's K-squared = 17.705, df = 3, p-value = 0.000506
```

It seems that even applying a log transformation the charges, the P-value remains under the significance level, meaning that we have to reject the (null) hypothesis of equal variance, that is needed for the ANOVA test

Below the plots showing the shapes of charges ~ region and log(charges) ~ region

[14]:
```
options(repr.plot.width=12, repr.plot.height=5)
p1 <- ggplot(data, aes(x=charges, y=..count.., fill=region))+
geom_density(alpha=0.3)
p2 <- ggplot(data, aes(x=log(charges), y=..count.., fill=region))+
geom_density(alpha=0.3)

p1 + p2
```



4.2 Distributions on Numerical variables

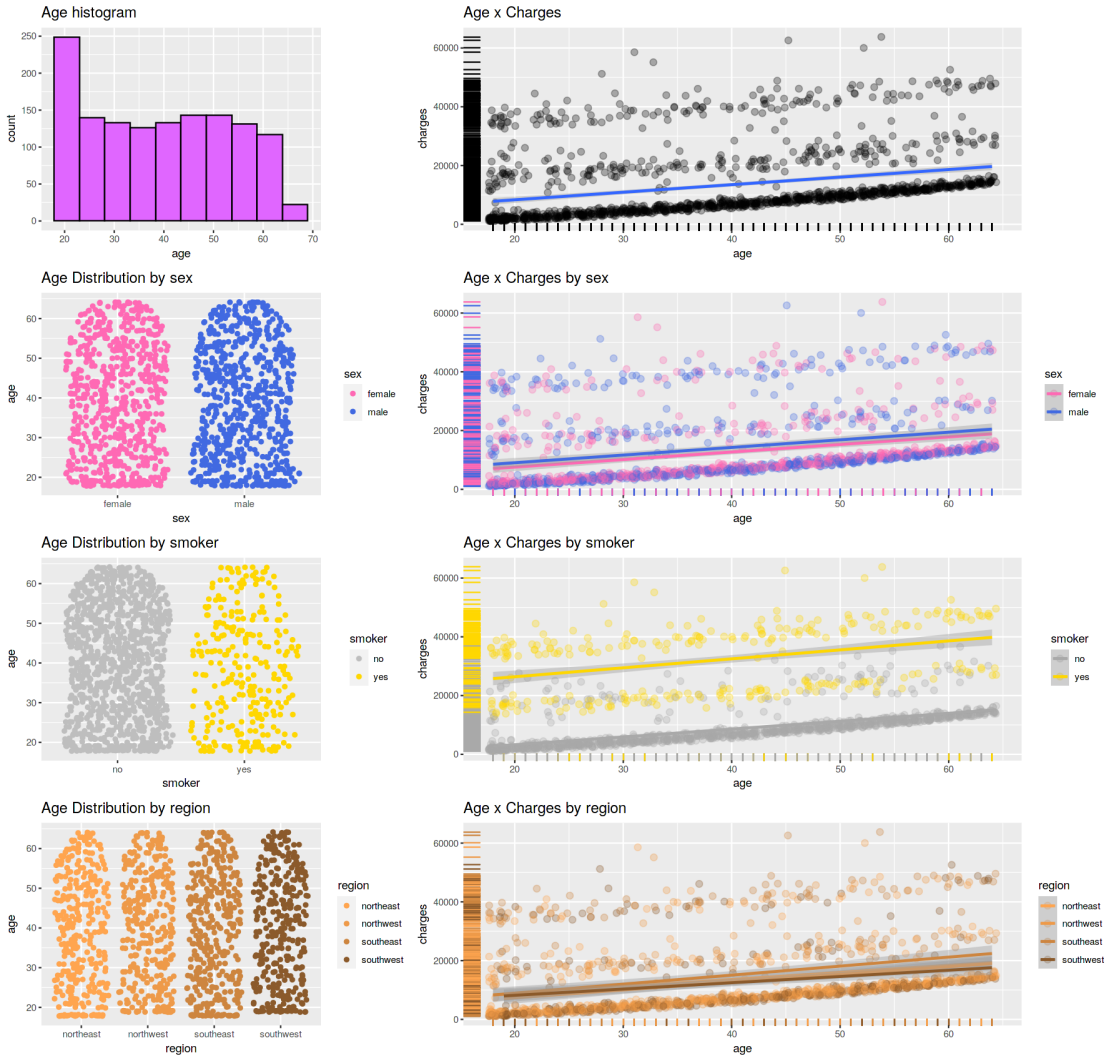
Let us now inspect the distribution of the numerical variables and the relations with the categorical ones

4.2.1 Age distributions

[15]:
```
p1 <- ggplot(data, aes(x=age)) +
    geom_histogram(color="black", fill="mediumorchid1", bins=10) +
    labs(title="Age histogram") +
    theme(plot.title = element_text(size=14))

p2 <- ggplot(data, aes(x=age, y=charges)) +
    geom_jitter(alpha=0.3, size=2.5) +
    scale_color_manual("pink") +
    geom_rug() +
    geom_smooth(method=lm, formula=y~x) +
    labs(title="Age x Charges") +
```

```
    theme(plot.title = element_text(size=14))

### age, charges, sex
p3 <- ggplot(data, aes(x=sex, y=age, color=sex)) +
    geom_sina() +
    scale_color_manual(values=c('hotpink', "royalblue")) +
    labs(title="Age Distribution by sex") +
    theme(plot.title = element_text(size=14))

p4 <- ggplot(data, aes(x=age, y=charges, color= sex)) +
    geom_jitter(alpha=0.3, size=2.5) +
    scale_color_manual(values=c('hotpink', "royalblue")) +
    geom_rug() +
    geom_smooth(method=lm, formula=y~x) +
    labs(title="Age x Charges by sex") +
    theme(plot.title = element_text(size=14))

### age, charges, smoker
p5 <- ggplot(data, aes(x=smoker, y=age, color=smoker)) +
    geom_sina() +
    scale_color_manual(values=c('grey', "gold")) +
    labs(title="Age Distribution by smoker") +
    theme(plot.title = element_text(size=14))

p6 <- ggplot(data, aes(x=age, y=charges, color= smoker)) +
    geom_jitter(alpha=0.3, size=2.5) +
    scale_color_manual(values=c('darkgrey', "gold")) +
    geom_rug() +
    geom_smooth(method=lm, formula=y~x) +
    labs(title="Age x Charges by smoker") +
    theme(plot.title = element_text(size=14))

### age, charges, region
p7 <- ggplot(data, aes(x=region, y=age, color=region)) +
    geom_sina() +
    scale_color_manual(values=c('tan1', "tan2", 'tan3', "tan4")) +
    labs(title="Age Distribution by region") +
    theme(plot.title = element_text(size=14))

p8 <- ggplot(data, aes(x=age, y=charges, color= region)) +
    geom_jitter(alpha=0.3, size=2.5) +
    scale_color_manual(values=c('tan1', "tan2", 'tan3', "tan4")) +
    geom_rug() +
    geom_smooth(method=lm, formula=y~x) +
    labs(title="Age x Charges by region") +
    theme(plot.title = element_text(size=14))
```

```
options(repr.plot.width=15, repr.plot.height=17)
layout <- "ABB \n CDD \n EFF \n GHH \n IJJ"
p1 + p2 + p3 + p4+ p5 + p6+ p7+ p8+ plot_layout(design = layout)
```



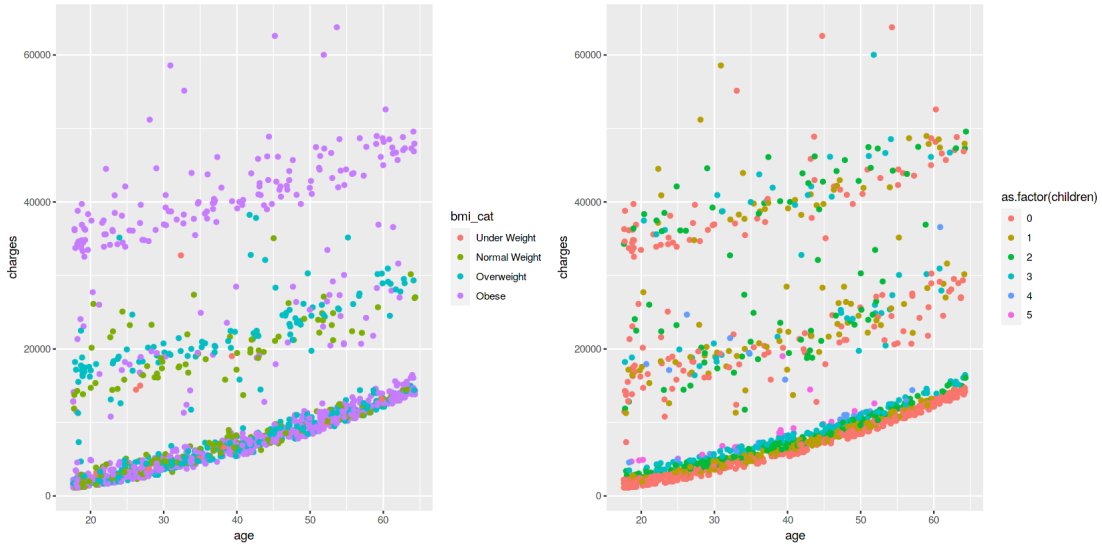
The first histogram shows us an skewed distribution characterized form an higher number of person about 20s.

Anyway the plots did not show particular details if not that the older the costumer the higher the charges and that the smoker status affect heavly the charges indiscriminately at any age. Moreover

it is possible to notice a presence of a middle level of charges that will be better analyzed.

Let us explore now the distribution of sample's ages and see if we can find someting more interesting

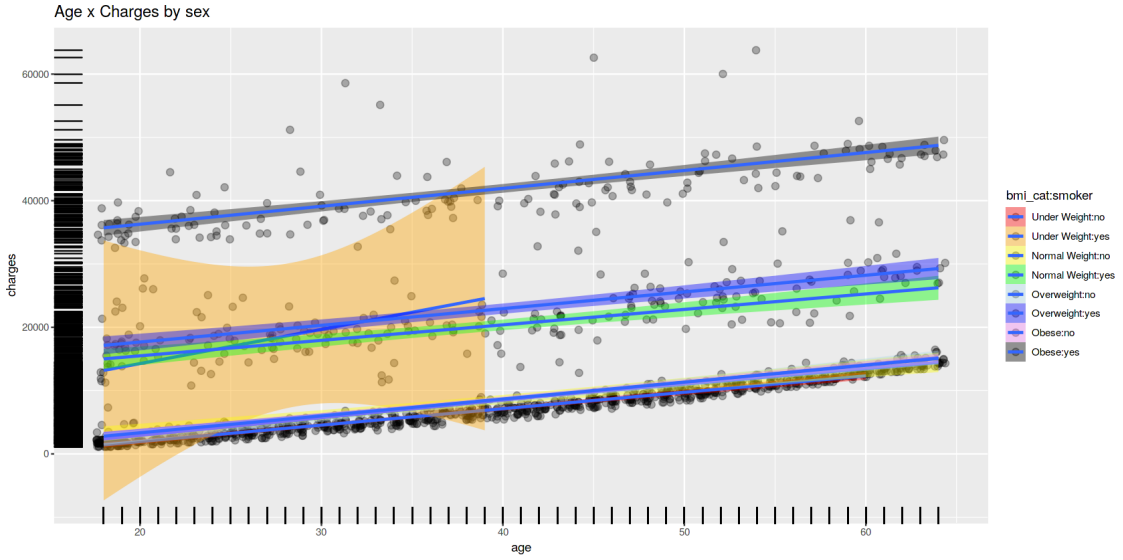
[16]:
```
options(repr.plot.width=14, repr.plot.height=7)
data <- (data %>% mutate(bmi_cat = cut(data$bmi, breaks = c(0, 18.5, 25, 30,␣
 ↪60), labels = c("Under Weight", "Normal Weight", "Overweight", "Obese"))))

p1 <- ggplot(data, aes(x=age, y=charges, color=bmi_cat))+
    geom_jitter()

p2 <- ggplot(data, aes(x=age, y=charges, color=as.factor(children)))+
    geom_jitter()

p1 + p2
```



The number of children seems not to affect this behaviour, on the contrary, by discretizing the BMI values, we can notice that in the middle level of charges ~ age there is majority of **Overwheighted** to **Normalweighted** people that can explain this behaviour, let us dig a little more about this.

[17]:
```
ggplot(data, aes(x=age, y=charges, fill= bmi_cat:smoker)) +
    geom_jitter(alpha=0.3, size=2.5)+␣
 ↪scale_fill_manual(values=c("red","orange","yellow","green","lightblue","blue",
    "violet","black")) +
    geom_rug() +
    geom_smooth(method=lm, formula=y~x) +
    labs(title="Age x Charges by sex") +
    theme(plot.title = element_text(size=14))
```

```
data$bmi_cat <- NULL
```

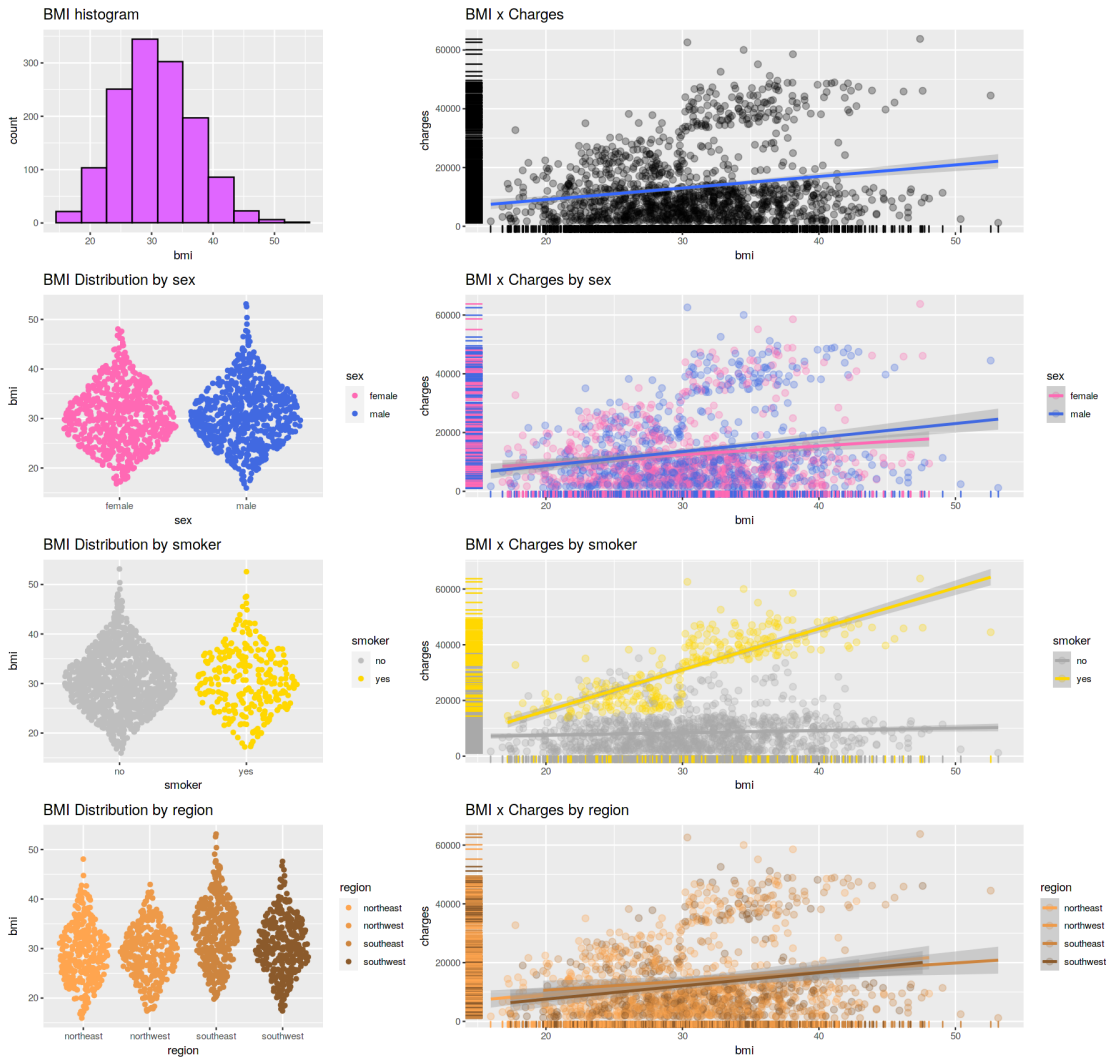


In fact we can notice that in the middle we can find a majority of Normal and Overweighted smokers, it is also interesting to notice that from 18 to almost 40 years old the Under Wheight smokers have a huge variability in the charges. By the way we are not going to investigate further on this factor.

4.2.2 BMI Distributions

[18]:
```
p1 <- ggplot(data, aes(x=bmi))+
    geom_histogram(color="black", fill="mediumorchid1", bins=10)+
    labs(title="BMI histogram" )+
    theme(plot.title = element_text(size=14))

p2 <- ggplot(data, aes(x=bmi, y=charges)) +
    geom_jitter(alpha=0.3, size=2.5) +
    scale_color_manual("mediumorchid1") +
    geom_rug() +
    geom_smooth(method=lm, formula=y~x) +
    labs(title="BMI x Charges") +
    theme(plot.title = element_text(size=14))

### bmi, charges, sex
p3 <- ggplot(data, aes(x=sex, y=bmi, color=sex)) +
    geom_sina() +
    scale_color_manual(values=c('hotpink', "royalblue")) +
    labs(title="BMI Distribution by sex") +
    theme(plot.title = element_text(size=14))
```

```
p4 <- ggplot(data, aes(x=bmi, y=charges, color= sex)) +
    geom_jitter(alpha=0.3, size=2.5) +
    scale_color_manual(values=c('hotpink', "royalblue")) +
    geom_rug() +
    geom_smooth(method=lm, formula=y~x) +
    labs(title="BMI x Charges by sex") +
    theme(plot.title = element_text(size=14))

### age, charges, smoker
p5 <- ggplot(data, aes(x=smoker, y=bmi, color=smoker)) +
    geom_sina() +
    scale_color_manual(values=c('grey', "gold")) +
    labs(title="BMI Distribution by smoker") +
    theme(plot.title = element_text(size=14))

p6 <- ggplot(data, aes(x=bmi, y=charges, color= smoker)) +
    geom_jitter(alpha=0.3, size=2.5)+
    scale_color_manual(values=c('darkgrey', "gold")) +
    geom_rug()+
    geom_smooth(method=lm, formula=y~x) +
    labs(title="BMI x Charges by smoker") +
    theme(plot.title = element_text(size=14))

### age, charges, region
p7 <- ggplot(data, aes(x=region, y=bmi, color=region)) +
    geom_sina() +
    scale_color_manual(values=c('tan1', "tan2", 'tan3', "tan4")) +
    labs(title="BMI Distribution by region") +
    theme(plot.title = element_text(size=14))

p8 <- ggplot(data, aes(x=bmi, y=charges, color= region)) +
    geom_jitter(alpha=0.3, size=2.5) +
    scale_color_manual(values=c('tan1', "tan2", 'tan3', "tan4")) +
    geom_rug() +
    geom_smooth(method=lm, formula=y~x) +
    labs(title="BMI x Charges by region") +
    theme(plot.title = element_text(size=14))


options(repr.plot.width=15, repr.plot.height=17)
layout<-"ABB \n CDD \n EFF \n GHH \n IJJ"
p1 + p2 + p3 + p4+ p5 + p6+ p7+ p8+ plot_layout(design = layout)
```



Comparing BMI~Charges with regard of the categorical values we can notice a clear increase of medical costs corresponding to the increase of BMI, with an important increase with the smoker status, and a higher BMI in the south est region, let us doing some statistical test to investigate about the significance of those last one:

4.2.3 Hypothesis test 3:

Let us perform a Bartlett Test to check the homogenity of the variance of charges between the regions before going to test the equality on the mean:

- $H_0 : \sigma_{Bne} = \sigma_{Bnw} = \sigma_{Bse} = \sigma_{Bsw}$

[19]:
```
bartlett.test(formula = bmi ~ region, data = data)
```

```
        Bartlett test of homogeneity of variances

data:  bmi by region
Bartlett's K-squared = 18.5, df = 3, p-value = 0.0003468
```

[20]:
```
bartlett.test(formula = log(bmi) ~ region, data = data)
```

```
        Bartlett test of homogeneity of variances

data:  log(bmi) by region
Bartlett's K-squared = 6.2626, df = 3, p-value = 0.09951
```

Using a log transformation on the BMI it seems that we are able to get rid of the difference between the variances in the BMIs distributions per region, let us now perform an ANOVA test on the log(bmi) to check the equality of the mean.

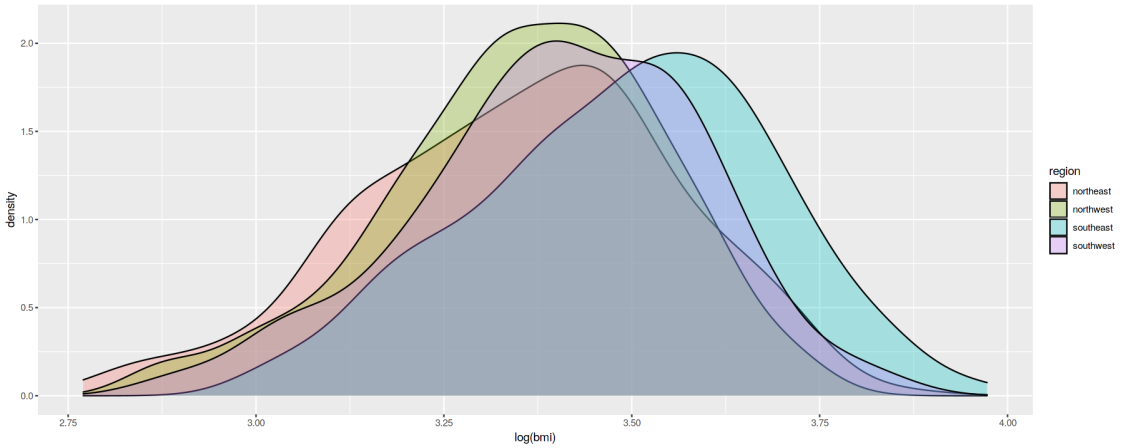
- $H_0 : \mu_{Bne} = \mu_{Bnw} = \mu_{Bse} = \mu_{Bsw}$

[21]:
```
log.bmi.aov <- aov(formula = log(bmi) ~ region, data = data)
summary(log.bmi.aov)
```

```
              Df Sum Sq Mean Sq F value Pr(>F)
region         3   4.13   1.377    36.2 <2e-16 ***
Residuals   1333  50.69   0.038
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Nonetheless, the **p-value** results to be significantly lower than the significance level, than we reject the null hypothesis, meaning that at least one of the means is significantly different from the one of the other regions

[22]:
```
options(repr.plot.width=15, repr.plot.height=6)

ggplot(data, aes(x=log(bmi), fill=region))+
    geom_density(alpha=.3)
```

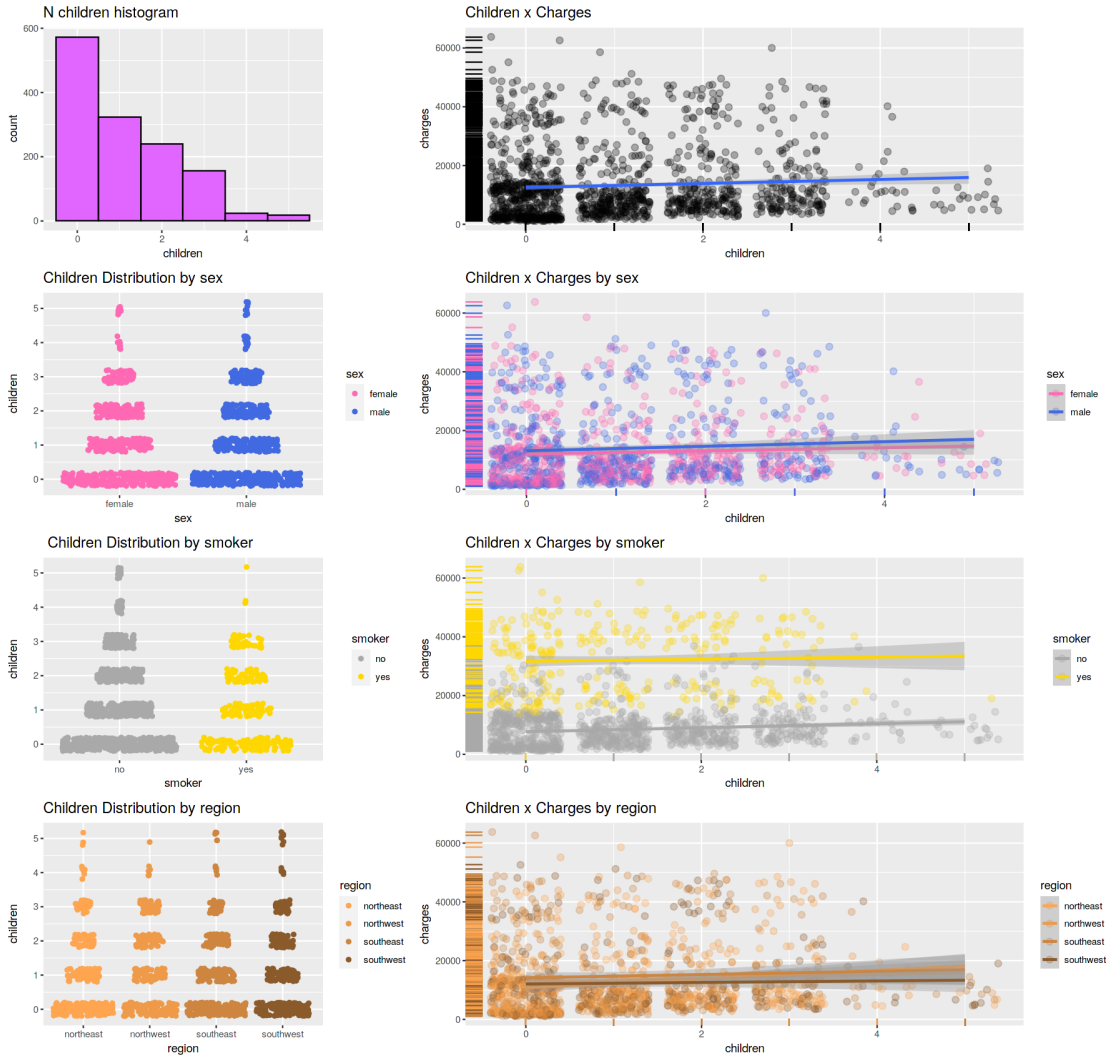


In fact seems that the southeast has a larger BMI compared to the other regions

4.2.4 Children Distributions

[23]:
```
p1<-ggplot(data, aes(x=children)) +
    geom_histogram(color="black", fill="mediumorchid1", bins=6) +
    labs(title="N children histogram") +
    theme(plot.title = element_text(size=14))

p2<-ggplot(data, aes(x=children, y=charges)) +
    geom_jitter(alpha=0.3, size=2.5) +
    scale_color_manual("pink") +
    geom_rug() +
    geom_smooth(method=lm, formula=y~x) +
    labs(title="Children x Charges") +
    theme(plot.title = element_text(size=14))

### bmi, charges, sex
p3<-ggplot(data, aes(x=sex, y=children, color=sex)) +
    geom_sina() +
    scale_color_manual(values=c('hotpink', "royalblue")) +
    labs(title="Children Distribution by sex") +
    theme(plot.title = element_text(size=14))

p4<-ggplot(data, aes(x=children, y=charges, color= sex)) +
    geom_jitter(alpha=0.3, size=2.5) +
    scale_color_manual(values=c('hotpink', "royalblue")) +
    geom_rug() +
    geom_smooth(method=lm, formula=y~x) +
    labs(title="Children x Charges by sex") +
    theme(plot.title = element_text(size=14))
```

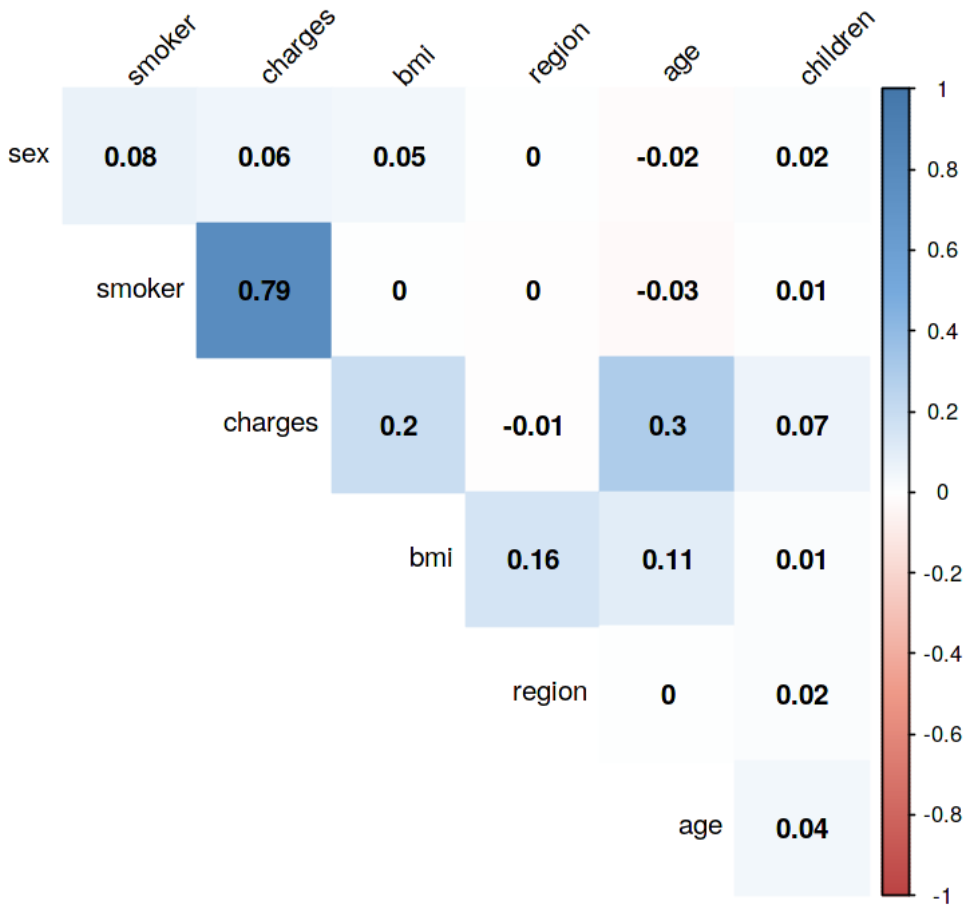
```
### age, charges, smoker
p5<-ggplot(data, aes(x=smoker, y=children, color=smoker)) +
    geom_sina() +
    scale_color_manual(values=c('darkgrey', "gold")) +
    labs(title=" Children Distribution by smoker") +
    theme(plot.title = element_text(size=14))

p6<-ggplot(data, aes(x=children, y=charges, color= smoker)) +
    geom_jitter(alpha=0.3, size=2.5) +
    scale_color_manual(values=c('darkgrey', "gold")) +
    geom_rug() +
    geom_smooth(method=lm, formula=y~x) +
    labs(title="Children x Charges by smoker") +
    theme(plot.title = element_text(size=14))

### age, charges, region
p7<-ggplot(data, aes(x=region, y=children, color=region)) +
    geom_sina() +
    scale_color_manual(values=c('tan1', "tan2", 'tan3', "tan4")) +
    labs(title="Children Distribution by region") +
    theme(plot.title = element_text(size=14))

p8<-ggplot(data, aes(x=children, y=charges, color= region)) +
    geom_jitter(alpha=0.3, size=2.5) +
    scale_color_manual(values=c('tan1', "tan2", 'tan3', "tan4")) +
    geom_rug() +
    geom_smooth(method=lm, formula=y~x) +
    labs(title="Children x Charges by region") +
    theme(plot.title = element_text(size=14))

options(repr.plot.width=15, repr.plot.height=17)
layout<-"ABB \n CDD \n EFF \n GHH \n IJJ"
p1 + p2 + p3 + p4+ p5 + p6+ p7+ p8+ plot_layout(design = layout)
```



In those last plots there not seems to be significant additional informations. We can just notice that the distribution is skewed to the right, maybe given by the high number of youngs that tends to have few or no children, charges increases with the number of children, and that people with more than 3 children tends to smoke less

4.3 Correlations

[24]:
```
options(repr.plot.width=7, repr.plot.height=7)
corr <- cor(data %>% mutate_if(is.factor, as.numeric))

col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))
corrplot(corr, method="color", col=col(200),
         type="upper", order="hclust",
         addCoef.col = "black",
         tl.col="black", tl.srt=45,
         sig.level = 0.01, insig = "blank",
         diag=FALSE)
```



From the correlation plot, it seems that the variables having the best correlation are the **smoker status** and the **charges**, there is almost no correlations between the other features

4.4 Resuming:

It can be observed that smokers have a very high insurance charge and due to this, the points in the higher regions of all the plots drawn against charges are those of smokers. It seems to be the most important variable impacting charges due to which levels are visible in all the plots.

5 Model data

After analyzing the behaviors of the dataset, we try to find the model that fits well.

5.1 Accuracy metrics

To assess the accuracy of the models we are going to implement we will make use of four metrics:

- **Adjusted** R^2: is based on R^2, which provides an indication of how well a model fits the data by adding predictors to the model. This addition decreases the effect of randomness on the R^2 value and provides more information on the issue of overfitting. We do not use R^2 because it has no meaning for non-linear models.
- **Residual Standard Error (RSE)**: measures the likelihood that the model fit deviates from the actual population. Estimates with an RSE of 25% or higher are subject to high sampling error.
- **Root Mean Squared Error (RMSE)**: is widely used in capturing large errors in data and it is sensitive to outliers it may not want to capture. It tends to increase as the complexity of the model increases (i.e. susceptible to overfitting).
- **Akaike Information Criterion (AIC)**: measures both how well the data fits the model, and how complex it is. It penalizes a model for its complexity, but rewards it for how well it fits the data.

5.2 Splitting of train data and test data

We split the data into train (80%) and test (20%) using sampling from original data.

[25]:
```
set.seed(42)

samp <- sample(1:nrow(data), ceiling(0.80*nrow(data)))
train <- data[samp,]
test <- data[-samp,]
```

5.3 Linear Regression 1

For starters, let us build the simplest model using all the available features.

[26]:
```
options(scipen = 999)
l <- lm(charges ~ age + sex + bmi + children + smoker + region, data = train)
summary(l)

l_pred <- predict(l, test)
```

```
radj <- summary(l)$adj.r.squared
rse <- sqrt(sum(residuals(l)^2) / l$df.residual )
rmse <- RMSE(l_pred, test$charges)
aic <- AIC(l)
l_reg <- cbind("Adjusted R sq"=radj, "RSE"=rse, "RMSE"=rmse, "AIC"=aic)
```

```
Call:
lm(formula = charges ~ age + sex + bmi + children + smoker +
    region, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-11365.4  -2860.1   -957.7   1499.5  30006.7

Coefficients:
                  Estimate Std. Error t value             Pr(>|t|)
(Intercept)     -11888.32    1137.13 -10.455 < 0.0000000000000002 ***
age                260.35      13.23  19.675 < 0.0000000000000002 ***
sexmale           -134.97     373.26  -0.362              0.71772
bmi                335.40      32.27  10.392 < 0.0000000000000002 ***
children           471.47     154.18   3.058              0.00228 **
smokeryes        23992.48     458.29  52.352 < 0.0000000000000002 ***
regionnorthwest   -542.48     531.59  -1.020              0.30774
regionsoutheast  -1281.62     533.19  -2.404              0.01640 *
regionsouthwest  -1149.71     535.55  -2.147              0.03204 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6067 on 1061 degrees of freedom
Multiple R-squared:  0.7539,        Adjusted R-squared:  0.7521
F-statistic: 406.4 on 8 and 1061 DF,  p-value: < 0.00000000000000022
```

Checking the p values of each explanatory variables, we see that sex is not significant, as suggested in the EDA stage. We will analyze it better in the next step.
Moreover, we have an F-statistic of 489.8 and a p-value almost equal to 0.

5.4 Linear Regression 2

Below we build the linear model without the features sex and we will analyze this.

[27]:
```
l_1 <- lm(charges ~ age + bmi + children + smoker + region, data = train)
summary(l_1)

l_1_pred <- predict(l_1, test)
radj <- summary(l_1)$adj.r.squared
rse <- sqrt(sum(residuals(l_1)^2) / l_1$df.residual )
```

```
rmse <- RMSE(l_1_pred, test$charges)
aic <- AIC(l_1)
l_1_reg <- cbind("Adjusted R sq"=radj, "RSE"=rse, "RMSE"=rmse, "AIC"=aic)
```

```
Call:
lm(formula = charges ~ age + bmi + children + smoker + region,
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-11433.0  -2834.6   -939.5   1528.7  29945.2

Coefficients:
                 Estimate Std. Error t value             Pr(>|t|)
(Intercept)     -11934.98    1129.32 -10.568 < 0.0000000000000002 ***
age                260.44      13.22  19.694 < 0.0000000000000002 ***
bmi                334.68      32.20  10.394 < 0.0000000000000002 ***
children           469.28     154.00   3.047              0.00237 **
smokeryes        23978.86     456.55  52.522 < 0.0000000000000002 ***
regionnorthwest   -538.87     531.28  -1.014              0.31068
regionsoutheast  -1277.03     532.82  -2.397              0.01671 *
regionsouthwest  -1146.39     535.25  -2.142              0.03244 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6064 on 1062 degrees of freedom
Multiple R-squared:  0.7539,        Adjusted R-squared:  0.7523
F-statistic: 464.8 on 7 and 1062 DF,  p-value: < 0.00000000000000022
```

We have five features, all of which are significant on charges except the *regionnorthwest*. From the coefficients, we know that a non-smoker zero years old who has no children and zero BMI will be charged $-\$11,934$ by health insurance. Also, since smoker has the biggest coefficient of all features, a unit change in smoker gives a bigger change in charges than a unit change in other features give, given all other features are fixed. In this case, given all other features are fixed, a non-smoker would have less charge than a smoker by $\$23,978$, which makes sense.

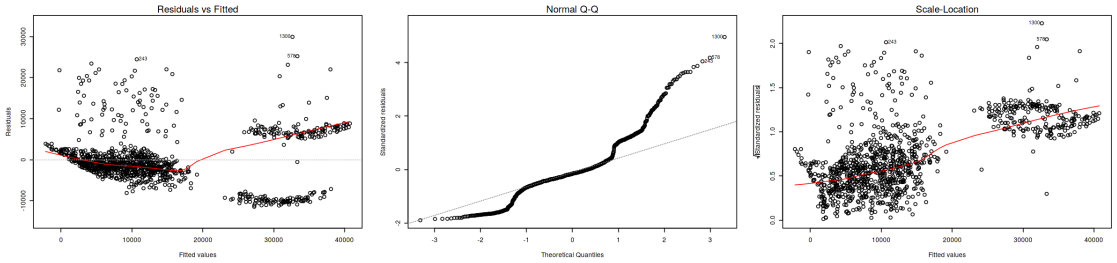
5.4.1 Model adequacy checking of Linear Regression

Let's check some assumptions of linear regression:

- Linearity of data, linear relationship between x and y: Residuals x Fitted plot
- Normality of residuals: Normal QQ plot
- Homogeneity of residuals variance (homoscedasticity), residuals with constant variance: scale location plot

[28]:
```
options(repr.plot.width=20, repr.plot.height=5)
par(mfrow=c(1,3))
```

```
plot(l_1, which=c(1,2,3))
```



The distribution of error terms looks normal seeing the boxplot, skewed to the left and with many outliers.
At Residuals x Fitted plot, the non horizontal line may indicate a non-linear relationship.
At Normal QQ plot, we see that the residuals are not exactly on the straight line, indicating that they are not normally distributed.
At Scale-Location plot, the non straight line indicates heteroscedasticity.

It is an over-dispersed data, that is, it has an increased number of outliers (i.e. the distribution has fatter tails than a normal distribution).

5.5 Log Linear Regression

We want to correct the model inadequacies transforming the response variable to stabilize the variance. We are going to use the log transformation presented in paragraph 4.

[29]:
```
train$log_charges <- log(train$charges)
l_2 <- lm(log_charges ~ age + bmi + smoker + children + smoker + region, data =␣
 ↪train)
summary(l_2)

l_2_pred <- predict(l_2, test)
radj <- summary(l_2)$adj.r.squared
rse <- sqrt(sum(residuals(l_2)^2) / l_2$df.residual )
rmse <- RMSE(l_2_pred, test$charges)
aic <- AIC(l_2)
l_2_reg <- cbind("Adjusted R sq"=radj, "RSE"=rse, "RMSE"=rmse, "AIC"=aic)
```

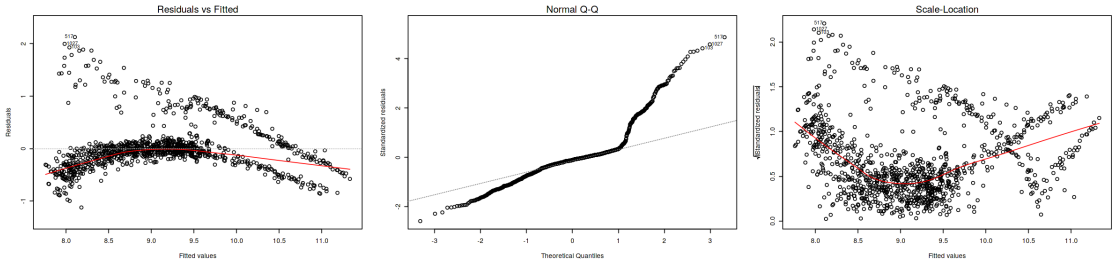
```
Call:
lm(formula = log_charges ~ age + bmi + smoker + children + smoker +
    region, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.12125 -0.19660 -0.04780  0.07315  2.12608
```

```
Coefficients:
                  Estimate Std. Error t value             Pr(>|t|)
(Intercept)      7.0127754  0.0815004  86.046 < 0.0000000000000002 ***
age              0.0343515  0.0009544  35.993 < 0.0000000000000002 ***
bmi              0.0135751  0.0023238   5.842        0.00000000686 ***
smokeryes        1.5494728  0.0329482  47.028 < 0.0000000000000002 ***
children         0.0967786  0.0111136   8.708 < 0.0000000000000002 ***
regionnorthwest -0.0730917  0.0383414  -1.906              0.056876 .
regionsoutheast -0.1719384  0.0384527  -4.471        0.00000860497 ***
regionsouthwest -0.1464416  0.0386279  -3.791              0.000158 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4376 on 1062 degrees of freedom
Multiple R-squared:  0.7716,        Adjusted R-squared:  0.7701
F-statistic: 512.6 on 7 and 1062 DF,  p-value: < 0.00000000000000022
```

In this model the significance level of *regionnorthwest* has grown the other ones. Adjusted R^2 for the model, after transforming the response variable charges, is higher than that of linear regression.

5.5.1 Model adequacy checking of Log Linear Regression

[30]:
```
par(mfrow=c(1,3))
plot(l_2, which=c(1,2,3))
```



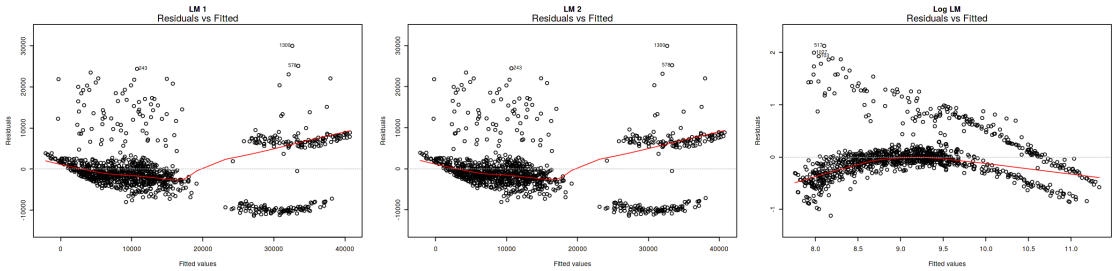
The distribution of error terms looks normal seeing the boxplot, skewed to the left and with many outliers.
At Residuals x Fitted plot, the non horizontal line may indicate a non-linear relationship.
At Normal QQ plot, we see that the residuals are not exactly on the straight line, indicating that they are not normally distributed.
At Scale-Location plot, the non straight line indicates heteroscedasticity.
It is an over-dispersed data, that is, it has an increased number of outliers (i.e. the distribution has fatter tails than a normal distribution).
The increasing pattern is still observed but the evaluation has improved. Furthermore we could

occur in overfitting.

5.6 Polynomial regression

In previous models we have used the features that, in most cases, have a low correlation. The residual plots exhibit parabolic trend, which provides a strong indication of non-linearity in the data.

[31]:
```
par(mfrow=c(1,3))

plot(l, c(1), main = 'LM 1')
plot(l_1, c(1), main = 'LM 2')
plot(l_2, c(1), main = 'Log LM')
```



We want to increase the correlation of the other features and we try to improve our model using the Polynomial regression. It is sensitive to outliers so the presence of one or two outliers can also badly affect the performance. The idea is to generate a new feature matrix consisting of all polynomial combinations of the features with degree 2.

We don't want charges to be included in the process of generating the polynomial combinations, so we take out charges from train and test and save them as y_train and y_test, respectively.

[32]:
```
y_train <- train$charges
y_test <- test$charges
```

From EDA we know that *sex* has no correlation with *charges*. We can drop it. Also, since polynomial combinations don't make sense to categorical features, we mutate *smoker* and *region* as numeric.

[33]:
```
X_train <- train %>%
  select(-c(charges, sex, log_charges)) %>%
  mutate(smoker = as.numeric(smoker), region = as.numeric(region))
X_test <- test %>%
  select(-c(charges, sex)) %>%
  mutate(smoker = as.numeric(smoker), region = as.numeric(region))
```

We use the formula below to apply polynomial combinations.

[34]:
```
formula <- as.formula(
  paste(' ~ .^2 + ', paste('poly(', colnames(X_train), ', 2, raw=TRUE)[, 2]',␣
 ↪collapse = ' + '))
)

formula
```

```
~.^2 + poly(age, 2, raw = TRUE)[, 2] + poly(bmi, 2, raw = TRUE)[,
    2] + poly(children, 2, raw = TRUE)[, 2] + poly(smoker, 2,
    raw = TRUE)[, 2] + poly(region, 2, raw = TRUE)[, 2]
```

Then, insert y_train and y_test back to the new datasets.

[35]:
```
train_poly <- as.data.frame(model.matrix(formula, data = X_train))
test_poly <- as.data.frame(model.matrix(formula, data = X_test))
train_poly$charges <- y_train
test_poly$charges <- y_test
colnames(train_poly)
```

```
'(Intercept)' 'age' 'bmi' 'children' 'smoker' 'region' 'poly(age, 2, raw = TRUE)[,␣
 ↪2]' 'poly(bmi, 2, raw = TRUE)[, 2]' 'poly(children, 2, raw = TRUE)[, 2]'␣
 ↪'poly(smoker, 2, raw = TRUE)[, 2]' 'poly(region, 2, raw = TRUE)[, 2]' 'age:bmi'␣
 ↪'age:children' 'age:smoker' 'age:region' 'bmi:children' 'bmi:smoker' 'bmi:
 ↪region' 'children:smoker' 'children:region' 'smoker:region' 'charges'
```

We can see that our new datasets train_poly and test_poly now have 22 columns:

1. **(Intercept)** is a column consists of constant 1, this is the constant term in the polynomial.
2. **age, bmi, children, smoker, region** are the original features.
3. **age², bmi², children², smoker², region²** are the square of the original features.
4. **age x bmi, age x children, age x smoker, age x region, bmi x children, bmi x smoker, children x smoker, children x region, smoker x region** are nine interactions between pairs of five features.
5. **charges** is the target feature.

Now, we are ready to make the model. As usual, we start with all features and use the backward elimination using the function *step*.

[36]:
```
temp <- lm(formula = charges ~ ., data = train_poly)
step(temp)
```

```
Start:  AIC=18127.63
charges ~ `(Intercept)` + age + bmi + children + smoker + region +
    `poly(age, 2, raw = TRUE)[, 2]` + `poly(bmi, 2, raw = TRUE)[, 2]` +
    `poly(children, 2, raw = TRUE)[, 2]` + `poly(smoker, 2, raw = TRUE)[, 2]` +
    `poly(region, 2, raw = TRUE)[, 2]` + `age:bmi` + `age:children` +
    `age:smoker` + `age:region` + `bmi:children` + `bmi:smoker` +
    `bmi:region` + `children:smoker` + `children:region` + `smoker:region`
```

```
Step:  AIC=18127.63
charges ~ `(Intercept)` + age + bmi + children + smoker + region +
    `poly(age, 2, raw = TRUE)[, 2]` + `poly(bmi, 2, raw = TRUE)[, 2]` +
    `poly(children, 2, raw = TRUE)[, 2]` + `poly(region, 2, raw = TRUE)[, 2]` +
    `age:bmi` + `age:children` + `age:smoker` + `age:region` +
    `bmi:children` + `bmi:smoker` + `bmi:region` + `children:smoker` +
    `children:region` + `smoker:region`


Step:  AIC=18127.63
charges ~ age + bmi + children + smoker + region + `poly(age, 2, raw = TRUE)[,
2]` +
    `poly(bmi, 2, raw = TRUE)[, 2]` + `poly(children, 2, raw = TRUE)[, 2]` +
    `poly(region, 2, raw = TRUE)[, 2]` + `age:bmi` + `age:children` +
    `age:smoker` + `age:region` + `bmi:children` + `bmi:smoker` +
    `bmi:region` + `children:smoker` + `children:region` + `smoker:region`

                                    Df   Sum of Sq         RSS   AIC
- `age:children`                     1        5373 23487663172 18126
- `bmi:children`                     1      934642 23488592441 18126
- `age:bmi`                          1     1321050 23488978849 18126
- `smoker:region`                    1     4815050 23492472849 18126
- `age:smoker`                       1     8677042 23496334841 18126
- region                             1    15288359 23502946158 18126
- `poly(children, 2, raw = TRUE)[, 2]`  1    16308269 23503966067 18126
- `children:region`                  1    24925975 23512583774 18127
- `children:smoker`                  1    33682542 23521340341 18127
- `poly(region, 2, raw = TRUE)[, 2]`    1    35982398 23523640196 18127
<none>                                           23487657799 18128
- age                                1    48794856 23536452655 18128
- children                           1    63099493 23550757292 18128
- `bmi:region`                       1    63299745 23550957544 18128
- `age:region`                       1    87141816 23574799615 18130
- bmi                                1   237134652 23724792451 18136
- `poly(bmi, 2, raw = TRUE)[, 2]`       1   303436138 23791093937 18139
- `poly(age, 2, raw = TRUE)[, 2]`       1   534056497 24021714296 18150
- smoker                             1  2453454932 25941112731 18232
- `bmi:smoker`                       1 13928095653 37415753452 18624

Step:  AIC=18125.63
charges ~ age + bmi + children + smoker + region + `poly(age, 2, raw = TRUE)[,
2]` +
    `poly(bmi, 2, raw = TRUE)[, 2]` + `poly(children, 2, raw = TRUE)[, 2]` +
    `poly(region, 2, raw = TRUE)[, 2]` + `age:bmi` + `age:smoker` +
    `age:region` + `bmi:children` + `bmi:smoker` + `bmi:region` +
    `children:smoker` + `children:region` + `smoker:region`
```

```
                                    Df   Sum of Sq         RSS   AIC
- `bmi:children`                     1      929728 23488592899 18124
- `age:bmi`                          1     1324396 23488987568 18124
- `smoker:region`                    1     4818492 23492481664 18124
- `age:smoker`                       1     8673841 23496337012 18124
- region                             1    15298393 23502961564 18124
- `poly(children, 2, raw = TRUE)[, 2]`  1    16366522 23504029693 18124
- `children:region`                  1    24930186 23512593358 18125
- `children:smoker`                  1    33678760 23521341932 18125
- `poly(region, 2, raw = TRUE)[, 2]`    1    36007613 23523670785 18125
<none>                                           23487663172 18126
- age                                1    48830048 23536493220 18126
- `bmi:region`                       1    63314295 23550977467 18126
- children                           1    72743918 23560407090 18127
- `age:region`                       1    87376405 23575039577 18128
- bmi                                1   237182326 23724845497 18134
- `poly(bmi, 2, raw = TRUE)[, 2]`       1   303468361 23791131533 18137
- `poly(age, 2, raw = TRUE)[, 2]`       1   534683519 24022346691 18148
- smoker                             1  2453885400 25941548571 18230
- `bmi:smoker`                       1 13929671115 37417334287 18622

Step:  AIC=18123.67
charges ~ age + bmi + children + smoker + region + `poly(age, 2, raw = TRUE)[,
2]` +
    `poly(bmi, 2, raw = TRUE)[, 2]` + `poly(children, 2, raw = TRUE)[, 2]` +
    `poly(region, 2, raw = TRUE)[, 2]` + `age:bmi` + `age:smoker` +
    `age:region` + `bmi:smoker` + `bmi:region` + `children:smoker` +
    `children:region` + `smoker:region`

                                    Df   Sum of Sq         RSS   AIC
- `age:bmi`                          1     1285569 23489878469 18122
- `smoker:region`                    1     4726078 23493318977 18122
- `age:smoker`                       1     8738572 23497331471 18122
- region                             1    15407692 23504000591 18122
- `poly(children, 2, raw = TRUE)[, 2]`  1    16965486 23505558385 18122
- `children:region`                  1    24093799 23512686698 18123
- `children:smoker`                  1    34886559 23523479458 18123
- `poly(region, 2, raw = TRUE)[, 2]`    1    35947139 23524540039 18123
<none>                                           23488592899 18124
- age                                1    49131663 23537724563 18124
- `bmi:region`                       1    62863924 23551456823 18124
- `age:region`                       1    86990553 23575583452 18126
- children                           1   194971332 23683564232 18130
- bmi                                1   236253193 23724846093 18132
- `poly(bmi, 2, raw = TRUE)[, 2]`       1   303285878 23791878777 18135
- `poly(age, 2, raw = TRUE)[, 2]`       1   535558259 24024151158 18146
- smoker                             1  2458990285 25947583184 18228
- `bmi:smoker`                       1 13984521760 37473114659 18622
```

```
Step:  AIC=18121.73
charges ~ age + bmi + children + smoker + region + `poly(age, 2, raw = TRUE)[,
2]` +
    `poly(bmi, 2, raw = TRUE)[, 2]` + `poly(children, 2, raw = TRUE)[, 2]` +
    `poly(region, 2, raw = TRUE)[, 2]` + `age:smoker` + `age:region` +
    `bmi:smoker` + `bmi:region` + `children:smoker` + `children:region` +
    `smoker:region`

                                       Df   Sum of Sq         RSS   AIC
- `smoker:region`                       1     4802062 23494680530 18120
- `age:smoker`                          1     8743947 23498622416 18120
- region                                1    15052462 23504930930 18120
- `poly(children, 2, raw = TRUE)[, 2]`  1    17061637 23506940105 18120
- `children:region`                     1    23838442 23513716911 18121
- `children:smoker`                     1    34972970 23524851439 18121
- `poly(region, 2, raw = TRUE)[, 2]`    1    35373756 23525252224 18121
<none>                                             23489878469 18122
- `bmi:region`                          1    62225557 23552104026 18123
- age                                   1    74221667 23564100135 18123
- `age:region`                          1    85780636 23575659104 18124
- children                              1   194807286 23684685755 18129
- bmi                                   1   277147175 23767025644 18132
- `poly(bmi, 2, raw = TRUE)[, 2]`       1   304188405 23794066873 18134
- `poly(age, 2, raw = TRUE)[, 2]`       1   534492866 24024371335 18144
- smoker                                1  2462483627 25952362096 18226
- `bmi:smoker`                          1 13998895652 37488774121 18620

Step:  AIC=18119.95
charges ~ age + bmi + children + smoker + region + `poly(age, 2, raw = TRUE)[,
2]` +
    `poly(bmi, 2, raw = TRUE)[, 2]` + `poly(children, 2, raw = TRUE)[, 2]` +
    `poly(region, 2, raw = TRUE)[, 2]` + `age:smoker` + `age:region` +
    `bmi:smoker` + `bmi:region` + `children:smoker` + `children:region`

                                       Df   Sum of Sq         RSS   AIC
- `age:smoker`                          1     8348559 23503029089 18118
- region                                1    10781204 23505461734 18118
- `poly(children, 2, raw = TRUE)[, 2]`  1    16919196 23511599726 18119
- `children:region`                     1    23865043 23518545573 18119
- `children:smoker`                     1    33287030 23527967560 18120
- `poly(region, 2, raw = TRUE)[, 2]`    1    34701319 23529381849 18120
<none>                                             23494680530 18120
- `bmi:region`                          1    62771024 23557451554 18121
- age                                   1    73737444 23568417974 18121
- `age:region`                          1    83677703 23578358233 18122
- children                              1   192422886 23687103417 18127
- bmi                                   1   282622632 23777303162 18131
```

```
- `poly(bmi, 2, raw = TRUE)[, 2]`       1   303828300 23798508830 18132
- `poly(age, 2, raw = TRUE)[, 2]`       1   536434288 24031114818 18142
- smoker                                1  2574495535 26069176065 18229
- `bmi:smoker`                          1 14505302518 37999983048 18632

Step:  AIC=18118.33
charges ~ age + bmi + children + smoker + region + `poly(age, 2, raw = TRUE)[,
2]` +
    `poly(bmi, 2, raw = TRUE)[, 2]` + `poly(children, 2, raw = TRUE)[, 2]` +
    `poly(region, 2, raw = TRUE)[, 2]` + `age:region` + `bmi:smoker` +
    `bmi:region` + `children:smoker` + `children:region`

                                      Df   Sum of Sq         RSS   AIC
- region                               1    10043221 23513072309 18117
- `poly(children, 2, raw = TRUE)[, 2]`  1    16803839 23519832928 18117
- `children:region`                    1    24123625 23527152714 18117
- `children:smoker`                    1    31342014 23534371103 18118
- `poly(region, 2, raw = TRUE)[, 2]`    1    33081298 23536110387 18118
<none>                                            23503029089 18118
- `bmi:region`                         1    64202541 23567231629 18119
- age                                  1    65454628 23568483717 18119
- `age:region`                         1    86048469 23589077558 18120
- children                             1   189969341 23692998430 18125
- bmi                                  1   283584181 23786613270 18129
- `poly(bmi, 2, raw = TRUE)[, 2]`       1   303244360 23806273449 18130
- `poly(age, 2, raw = TRUE)[, 2]`       1   536326061 24039355150 18140
- smoker                               1  3056059674 26559088763 18247
- `bmi:smoker`                         1 14529344591 38032373680 18631

Step:  AIC=18116.79
charges ~ age + bmi + children + smoker + `poly(age, 2, raw = TRUE)[, 2]` +
    `poly(bmi, 2, raw = TRUE)[, 2]` + `poly(children, 2, raw = TRUE)[, 2]` +
    `poly(region, 2, raw = TRUE)[, 2]` + `age:region` + `bmi:smoker` +
    `bmi:region` + `children:smoker` + `children:region`

                                      Df   Sum of Sq         RSS   AIC
- `poly(children, 2, raw = TRUE)[, 2]`  1    16283040 23529355350 18116
- `poly(region, 2, raw = TRUE)[, 2]`    1    23995854 23537068163 18116
- `children:region`                    1    26117902 23539190211 18116
- `children:smoker`                    1    32014275 23545086584 18116
<none>                                            23513072309 18117
- age                                  1    60362553 23573434862 18118
- `age:region`                         1    76514766 23589587076 18118
- `bmi:region`                         1   163985496 23677057805 18122
- children                             1   192711582 23705783891 18124
- bmi                                  1   277205351 23790277660 18127
- `poly(bmi, 2, raw = TRUE)[, 2]`       1   294854827 23807927136 18128
- `poly(age, 2, raw = TRUE)[, 2]`       1   532130089 24045202398 18139
```

```
- smoker                                1  3064259583 26577331892 18246
- `bmi:smoker`                          1 14547358532 38060430841 18630

Step:  AIC=18115.53
charges ~ age + bmi + children + smoker + `poly(age, 2, raw = TRUE)[, 2]` +
    `poly(bmi, 2, raw = TRUE)[, 2]` + `poly(region, 2, raw = TRUE)[, 2]` +
    `age:region` + `bmi:smoker` + `bmi:region` + `children:smoker` +
    `children:region`

                                     Df   Sum of Sq         RSS   AIC
- `poly(region, 2, raw = TRUE)[, 2]`  1    22461644 23551816994 18114
- `children:smoker`                   1    26655458 23556010807 18115
- `children:region`                   1    33942341 23563297690 18115
<none>                                              23529355350 18116
- age                                 1    50994949 23580350298 18116
- `age:region`                        1    77498904 23606854254 18117
- `bmi:region`                        1   158913814 23688269164 18121
- children                            1   196124861 23725480211 18122
- bmi                                 1   278906739 23808262089 18126
- `poly(bmi, 2, raw = TRUE)[, 2]`     1   297786637 23827141986 18127
- `poly(age, 2, raw = TRUE)[, 2]`     1   516175023 24045530373 18137
- smoker                              1  3115608871 26644964221 18247
- `bmi:smoker`                        1 14661799726 38191155076 18632

Step:  AIC=18114.55
charges ~ age + bmi + children + smoker + `poly(age, 2, raw = TRUE)[, 2]` +
    `poly(bmi, 2, raw = TRUE)[, 2]` + `age:region` + `bmi:smoker` +
    `bmi:region` + `children:smoker` + `children:region`

                                   Df   Sum of Sq         RSS   AIC
- `children:smoker`                 1    25849064 23577666057 18114
- `children:region`                 1    26630324 23578447317 18114
<none>                                             23551816994 18114
- age                               1    54754105 23606571099 18115
- `age:region`                      1   106084100 23657901094 18117
- `bmi:region`                      1   182206866 23734023860 18121
- children                          1   184465609 23736282603 18121
- bmi                               1   284291864 23836108858 18125
- `poly(bmi, 2, raw = TRUE)[, 2]`   1   333806262 23885623256 18128
- `poly(age, 2, raw = TRUE)[, 2]`   1   511629494 24063446488 18136
- smoker                            1  3117566284 26669383277 18246
- `bmi:smoker`                      1 14652662705 38204479698 18630

Step:  AIC=18113.72
charges ~ age + bmi + children + smoker + `poly(age, 2, raw = TRUE)[, 2]` +
    `poly(bmi, 2, raw = TRUE)[, 2]` + `age:region` + `bmi:smoker` +
    `bmi:region` + `children:region`
```

```
                                  Df   Sum of Sq         RSS   AIC
- `children:region`                1    28550656 23606216713 18113
<none>                                          23577666057 18114
- age                              1    54944040 23632610097 18114
- `age:region`                     1   105832047 23683498104 18116
- `bmi:region`                     1   181189167 23758855224 18120
- children                         1   224913541 23802579598 18122
- bmi                              1   277848401 23855514459 18124
- `poly(bmi, 2, raw = TRUE)[, 2]`  1   341759786 23919425844 18127
- `poly(age, 2, raw = TRUE)[, 2]`  1   511377570 24089043627 18135
- smoker                           1  3303920664 26881586722 18252
- `bmi:smoker`                     1 14640664628 38218330685 18628

Step:  AIC=18113.02
charges ~ age + bmi + children + smoker + `poly(age, 2, raw = TRUE)[, 2]` +
    `poly(bmi, 2, raw = TRUE)[, 2]` + `age:region` + `bmi:smoker` +
    `bmi:region`

                                  Df   Sum of Sq         RSS   AIC
<none>                                          23606216713 18113
- age                              1    50549656 23656766369 18113
- `age:region`                     1   100469074 23706685788 18116
- `bmi:region`                     1   231387464 23837604177 18122
- bmi                              1   277941179 23884157892 18124
- `poly(bmi, 2, raw = TRUE)[, 2]`  1   335495450 23941712163 18126
- `poly(age, 2, raw = TRUE)[, 2]`  1   501655696 24107872409 18134
- children                         1   587228297 24193445010 18137
- smoker                           1  3320525502 26926742215 18252
- `bmi:smoker`                     1 14674459256 38280675970 18628


Call:
lm(formula = charges ~ age + bmi + children + smoker + `poly(age, 2, raw = TRUE)[,␣
 ↪2]` +
    `poly(bmi, 2, raw = TRUE)[, 2]` + `age:region` + `bmi:smoker` +
    `bmi:region`, data = train_poly)

Coefficients:
                     (Intercept)                               age
                       14359.291                          -112.535
                             bmi                          children
                        -689.685                           643.828
                          smoker  `poly(age, 2, raw = TRUE)[, 2]`
                      -22562.223                             4.203
`poly(bmi, 2, raw = TRUE)[, 2]`                       `age:region`
                         -11.305                            18.792
                    `bmi:smoker`                       `bmi:region`
                        1512.345                           -39.363
```

Save the best model as l_p, then predict. After that, calculate the metrics.

[37]:
```
l_p <- lm(formula = charges ~ age + bmi + children + smoker + `poly(age, 2, raw␣
 ↪= TRUE)[, 2]` + `poly(bmi, 2, raw = TRUE)[, 2]` + `age:region` +
          `bmi:smoker` + `bmi:region`, data = train_poly)
summary(l_p)

l_p_pred <- predict(l_p, test_poly)
radj <- summary(l_p)$adj.r.squared
rse <- sqrt(sum(residuals(l_p)^2) / l_p$df.residual )
rmse <- RMSE(l_p_pred, test$charges)
aic <- AIC(l_p)
l_p_reg <- cbind("Adjusted R sq"=radj, "RSE"=rse, "RMSE"=rmse, "AIC"=aic)
```

```
Call:
lm(formula = charges ~ age + bmi + children + smoker + `poly(age, 2, raw = TRUE)[,␣
 ↪2]` +
    `poly(bmi, 2, raw = TRUE)[, 2]` + `age:region` + `bmi:smoker` +
    `bmi:region`, data = train_poly)

Residuals:
     Min       1Q   Median       3Q      Max
-11487.1  -1881.7  -1309.9   -369.8  30489.5

Coefficients:
                                  Estimate  Std. Error t value
(Intercept)                     14359.2907   3796.2160   3.783
age                              -112.5352     74.6947  -1.507
bmi                              -689.6853    195.2247  -3.533
children                          643.8281    125.3795   5.135
smoker                          -22562.2234   1847.7310 -12.211
`poly(age, 2, raw = TRUE)[, 2]`      4.2026      0.8855   4.746
`poly(bmi, 2, raw = TRUE)[, 2]`    -11.3049      2.9126  -3.881
`age:region`                        18.7916      8.8472   2.124
`bmi:smoker`                      1512.3454     58.9156  25.670
`bmi:region`                       -39.3631     12.2118  -3.223
                                         Pr(>|t|)
(Intercept)                              0.000164 ***
age                                      0.132210
bmi                                      0.000429 ***
children                                 0.000000336 ***
smoker                                 < 0.0000000000000002 ***
`poly(age, 2, raw = TRUE)[, 2]`          0.000002357 ***
`poly(bmi, 2, raw = TRUE)[, 2]`          0.000110 ***
`age:region`                             0.033901 *
```

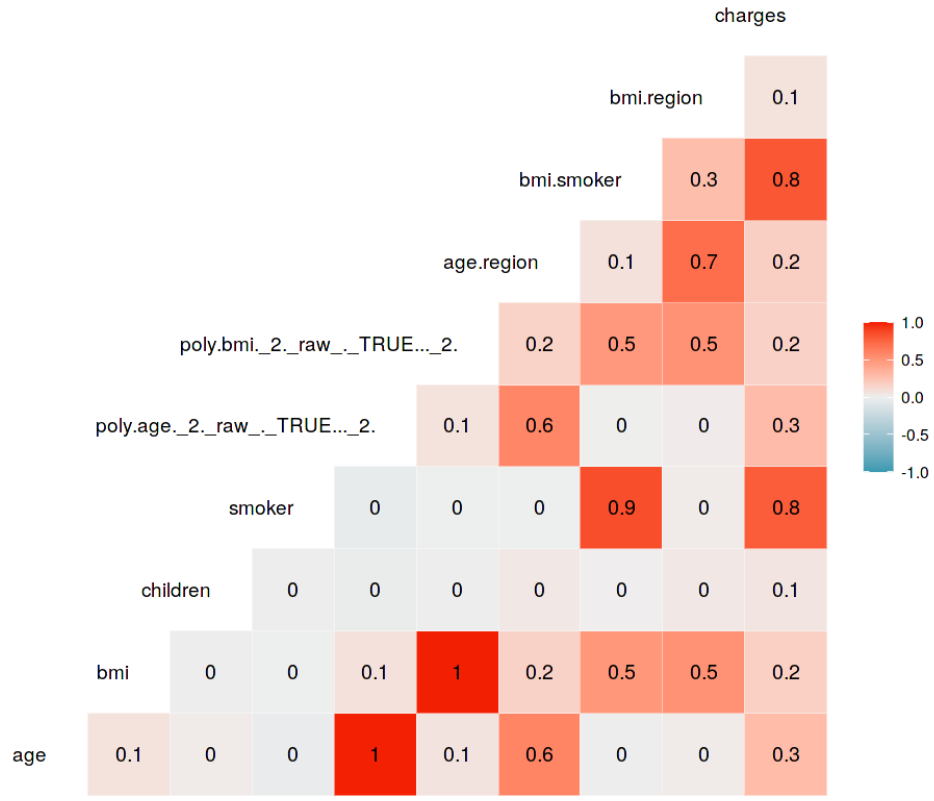
```
`bmi:smoker`                    < 0.0000000000000002 ***
`bmi:region`                                0.001306 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4719 on 1060 degrees of freedom
Multiple R-squared:  0.8513,        Adjusted R-squared:   0.85
F-statistic:   674 on 9 and 1060 DF,  p-value: < 0.00000000000000022
```

We have nine features, all of which are significant on charges, except for *age*. From the coefficients, we know that a non-smoker zero years old who has no children and zero BMI will be charged \$14,359 by health insurance (which we know this scenario is impossible). Also, since smoker has the biggest coefficient of all features, a unit change in smoker gives a bigger change in charges than a unit change in other features give, given all other features are fixed. In this case, given all other features are fixed, a non-smoker would have more charge than a smoker by -\$22358.

5.6.1 Model adequacy checking of Polynomial Regression

We need to make sure that there is a linear relationship between predictors and target variable. This can be done by visually looking at the correlation between each pair of predictor and target variable.

[38]:
```
options(repr.plot.width=9, repr.plot.height=7)

cols <- c('age', 'bmi', 'children', 'smoker', 'poly(age, 2, raw = TRUE)[, 2]',␣
 ↪'poly(bmi, 2, raw = TRUE)[, 2]', 'age:region', 'bmi:smoker', 'bmi:region')
ggcorr(train_poly %>% select(c(all_of(cols), 'charges')), hjust = 1, layout.exp␣
 ↪= 2, label = T)
```



Another way is to use hypothesis testing with Pearson's product-moment correlation.

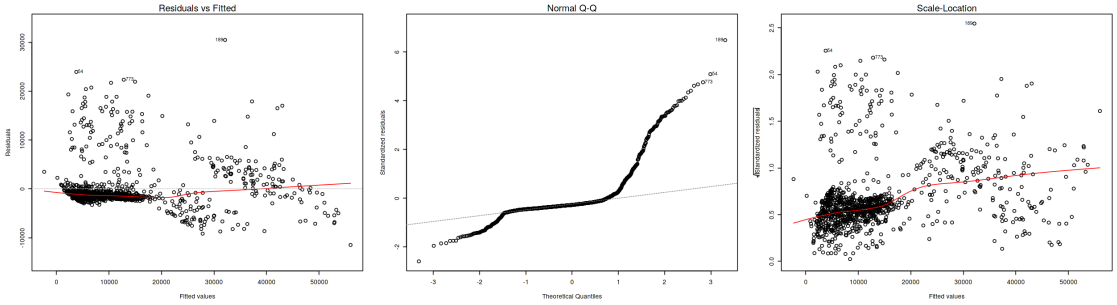
- H_0: the predictor does not correlate with *charge*
- H_1: the predictor correlates with *charge*

[39]:
```
for (col in cols) {
  cor <- cor.test(train_poly[, col], train_poly$charges)
  print(round(cor$p.value, 4))
}
```

```
[1] 0
[1] 0
[1] 0.0718
[1] 0
[1] 0
[1] 0
[1] 0
[1] 0
[1] 0.02
```

Since the p-value for each predictor-target pair is below alpha (0.05) except for *children*, reject H0. We can safely say that the predictors correlate with target variable.

[40]:
```
options(repr.plot.width=21, repr.plot.height=6)
par(mfrow=c(1,3))
plot(l_p, which=c(1,2,3))
```



The distribution of error terms looks normal seeing the boxplot, skewed to the left and with many outliers.
At Normal QQ plot, we see that the residuals are not exactly on the straight line, indicating that they are not normally distributed.
At Scale-Location plot, the non straight line indicates heteroscedasticity.
It is an over-dispersed data, that is, it has an increased number of outliers (i.e. the distribution has fatter tails than a normal distribution).

6 Models Evaluation

Now we are going to compare the metrics between al the implemented models.

[41]:
```
result <- rbind(l_reg, l_1_reg, l_2_reg, l_p_reg)
rownames(result) <- c("Linear Regression 1", "Linear Regression 2", "Log Linear␣
↪Regression", "Polynomial Regression")
result
```

	Adjusted R sq	RSE	RMSE	AIC
Linear Regression 1	0.7520835	6066.6897379	6062.979	21688.107
Linear Regression 2	0.7522864	6064.2064376	6063.061	21686.239
Log Linear Regression	0.7701225	0.4376392	17071.383	1278.087
Polynomial Regression	0.8499892	4719.1117562	5086.870	21151.546

Linear Regression 1 and Linear Regression 2
Adjusted R squared of these model is about 0.752, which means that 75.20% of the variation in charges could be explained by the independent variables we took in consideration. They might be good models, but we should also consider the other metrics used to compare their complexity with how well models fits the data.

Log Linear Regression
The metrics indicate that it is the best model, unless RMSE is used. The latter shows us that the model is very susceptible to overfitting.

Polynomial Regression
This model has a great variation of the charges indicated by the Adjusted R^2. Compared to the first two models, polynomial regression should be more complex but fits the data very well. To conclude the RMSE indicates that the model shouldn't overfitting.

7 Conclusion

The models we have built can be used for inference of how the different predictors influence the outcome but it is far from perfect. But some feature engineering such as polynomial regression plays an important role to improve the model. There is still presence of non-linearity and non-constant variance of errors. Moreover, the outliers points should be analyzed to find a better model. To obtain even more precision in its predictions, this insurance company should collect more data about its customers in order to explain the behavior of some individuals that we have noticed in the exploratory phase.

We can be sure that smoking could affect your wallet.