Econometrics and Statistical Models Group Project

CRETIN Lucie – PARIS Emma – SZEPEK Maria Sofie – VINCENOT Amandine

CONTENT

- Aim of the project and data source
- Data cleaned and recoded
- Descriptive statistics for each variable
- Graph for each quantitative variable
- Graphs for each qualitative variable
- Methodology used : from classical regression (linear) to GAM
- Results and comparison of prediction accuracy
- Conclusion

Aim of the project and data source

- The aim of this project is to predict the cost of insurance using regression by leveraging personal health data.

- Our data source is composed of 1,338 rows and 7 variables : age, sex, bmi, children, smoker, region (as explaining variables) and charges (to be explained variable).

- The dataset is hosted on Kaggle:

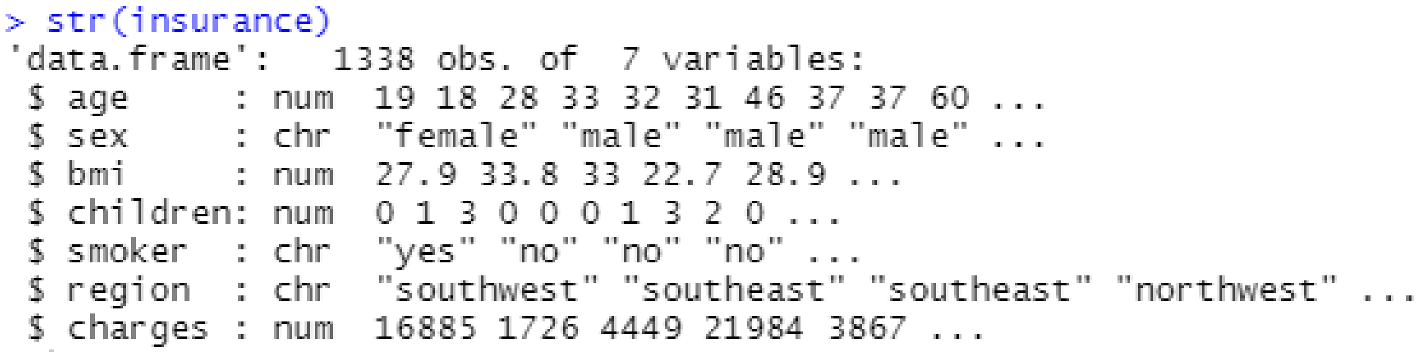
https://www.kaggle.com/mirichoi0218/insurance

Data cleaned and recoded

Step 1 : Verify data structure

This is a data frame with 1,338 rows and 7 columns :

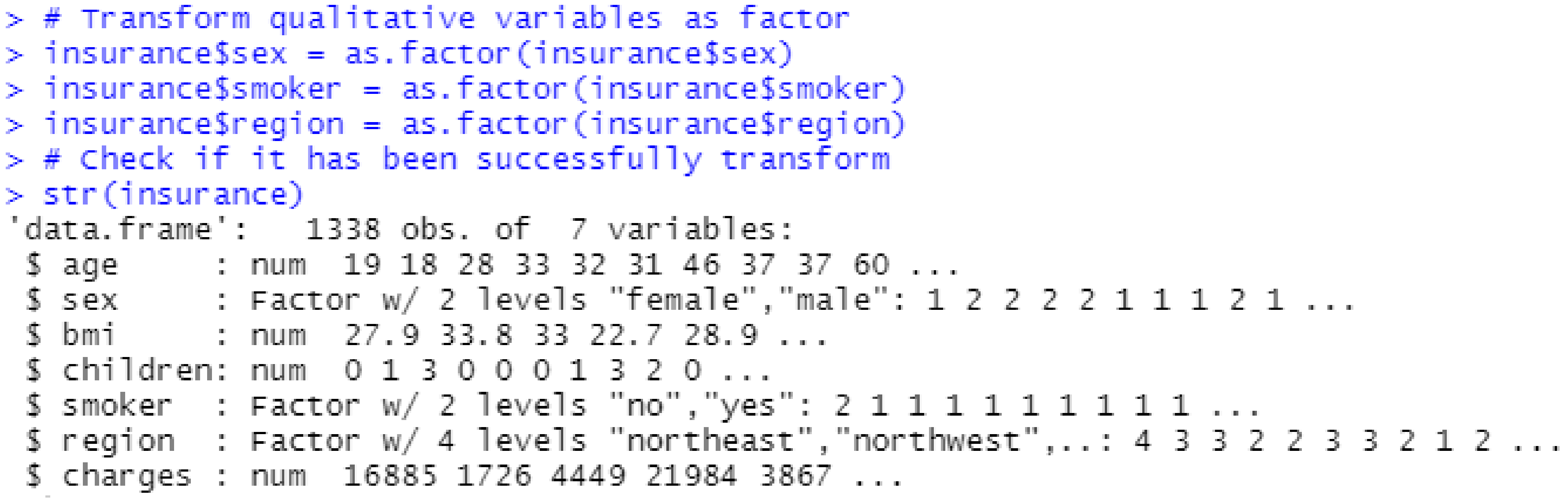
- The are 3 character strings (categorical) variables : sex, smoker, region
- There are 4 numerical variables : age, bmi, children, charges
- We consider the variable charges as dependent variable thus continuous



```
> str(insurance)
'data.frame':	1338 obs. of  7 variables:
 $ age     : num  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: num  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 ...
```

Data cleaned and recoded

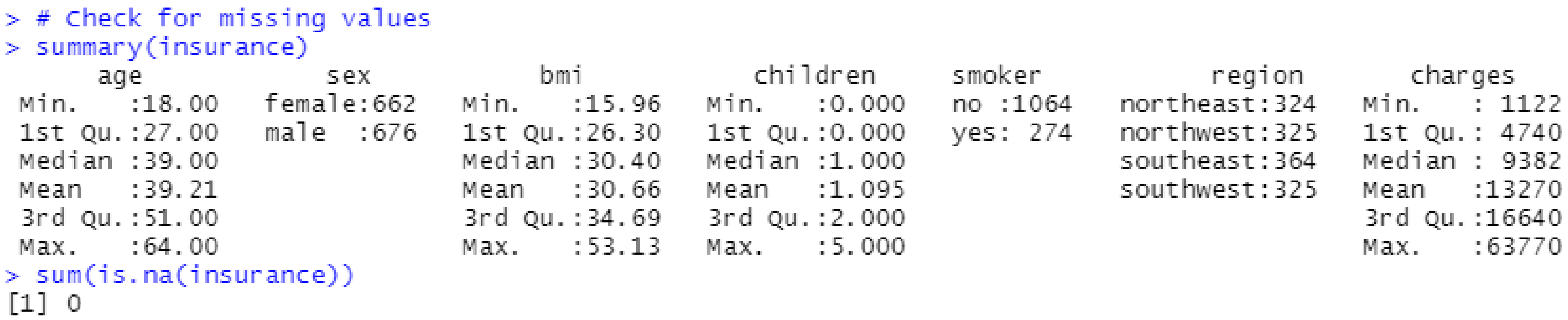
Step 2 : Recode qualitative variables by factorization



```
> # Transform qualitative variables as factor
> insurance$sex = as.factor(insurance$sex)
> insurance$smoker = as.factor(insurance$smoker)
> insurance$region = as.factor(insurance$region)
> # Check if it has been successfully transform
> str(insurance)
'data.frame':	1338 obs. of  7 variables:
 $ age     : num  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: num  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
 $ charges : num  16885 1726 4449 21984 3867 ...
```

Data cleaned and recoded

Step 3 : Check for missing values



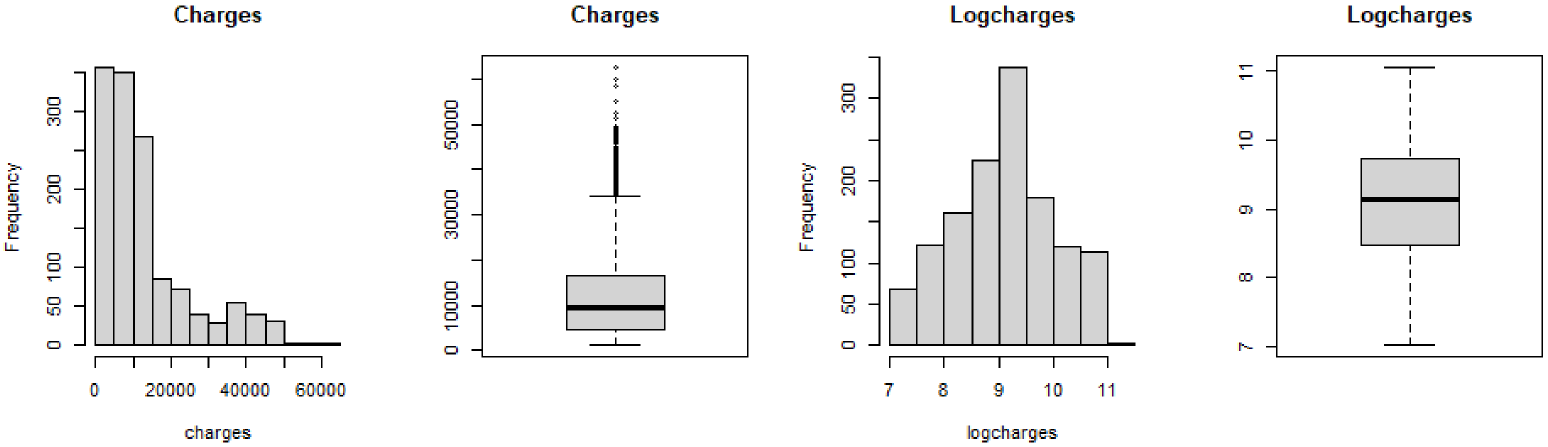
```
> # Check for missing values
> summary(insurance)
      age            sex            bmi           children       smoker          region        charges     
 Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064   northeast:324   Min.   : 1122  
 1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274   northwest:325   1st Qu.: 4740  
 Median :39.00                Median :30.40   Median :1.000              southeast:364   Median : 9382  
 Mean   :39.21                Mean   :30.66   Mean   :1.095              southwest:325   Mean   :13270  
 3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000                              3rd Qu.:16640  
 Max.   :64.00                Max.   :53.13   Max.   :5.000                              Max.   :63770  
> sum(is.na(insurance))
[1] 0
```

There is no missing values.

Data cleaned and recoded

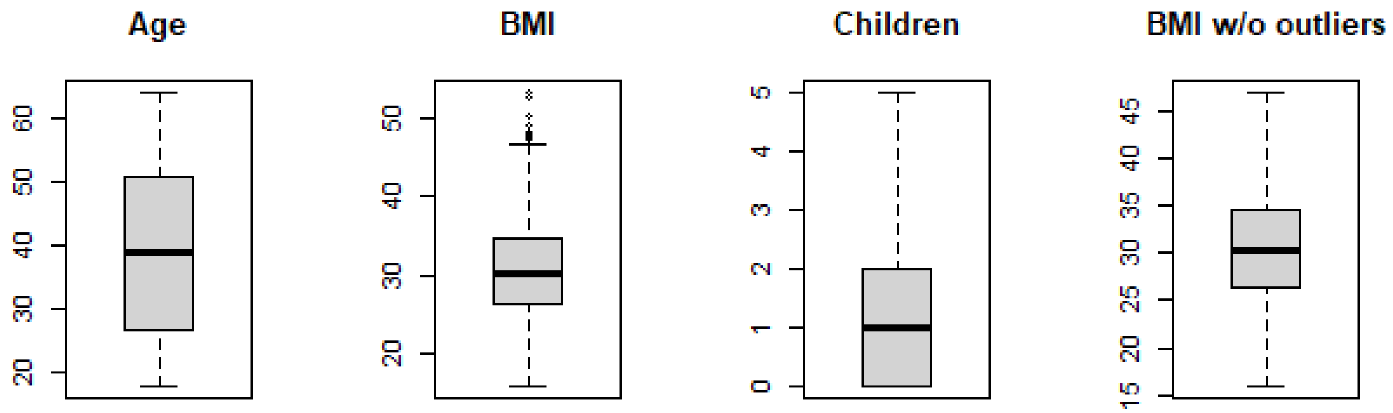
Step 3 : Check if the variable charges is normally distributed

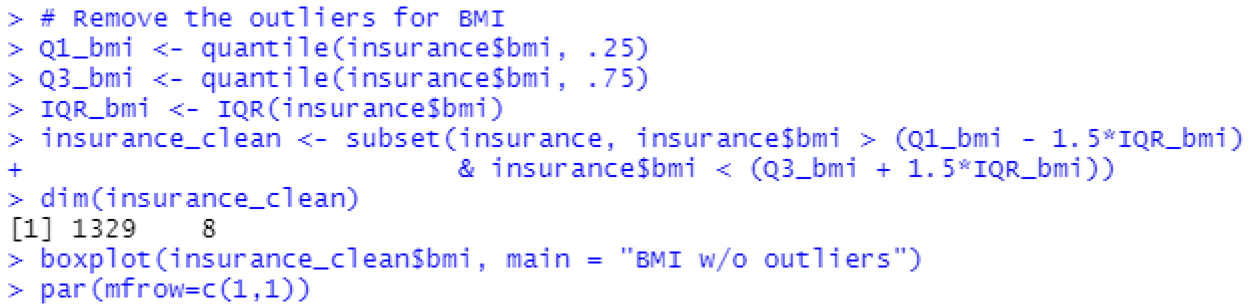
```
> # Check if the variable charges is normally distributed with histogram
> hist(charges, main = "Charges")
> boxplot(charges, main = "Charges")
> insurance$logcharges = log(insurance$charges)
> hist(logcharges, main = "Logcharges")
> boxplot(logcharges, main = "Logcharges")
> par(mfrow=c(1,1))
```



Data cleaned and recoded

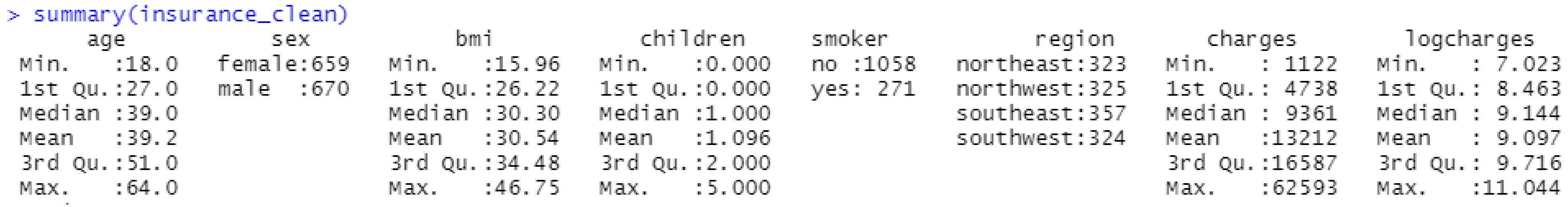
Step 4 : Remove outliers for quantitative variables





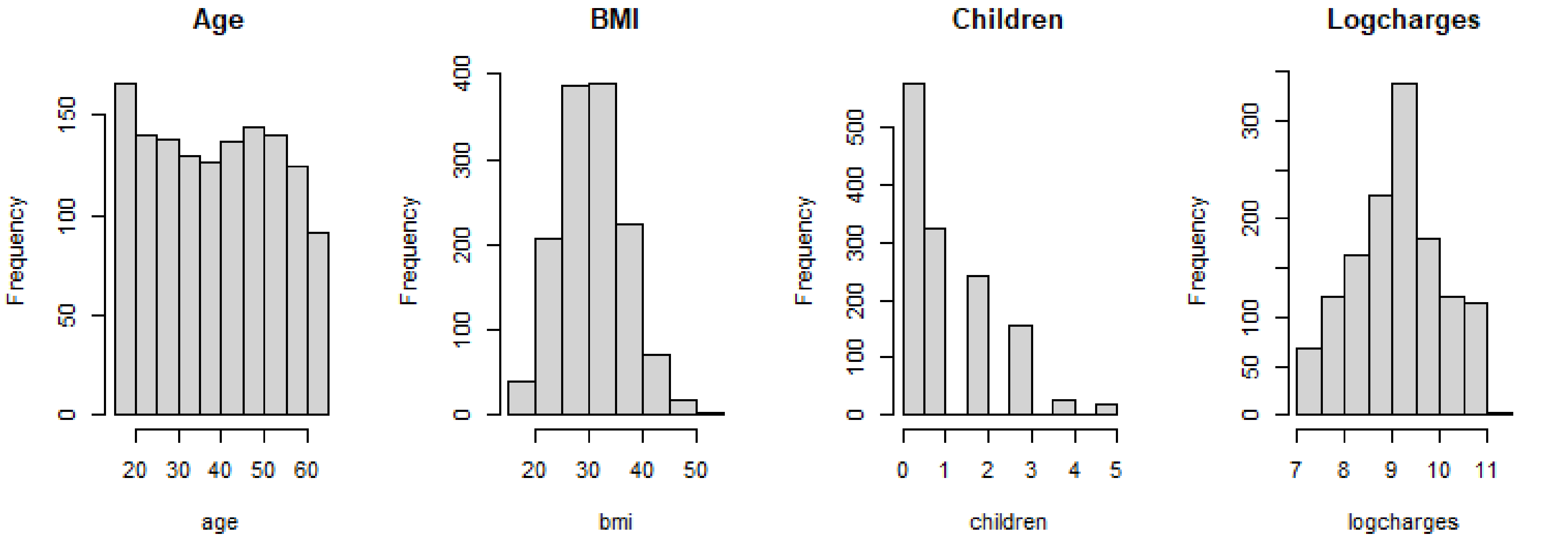
```
> # Remove the outliers for BMI
> Q1_bmi <- quantile(insurance$bmi, .25)
> Q3_bmi <- quantile(insurance$bmi, .75)
> IQR_bmi <- IQR(insurance$bmi)
> insurance_clean <- subset(insurance, insurance$bmi > (Q1_bmi - 1.5*IQR_bmi)
+                          & insurance$bmi < (Q3_bmi + 1.5*IQR_bmi))
> dim(insurance_clean)
[1] 1329    8
> boxplot(insurance_clean$bmi, main = "BMI w/o outliers")
> par(mfrow=c(1,1))
```

Descriptive statistics for each variable

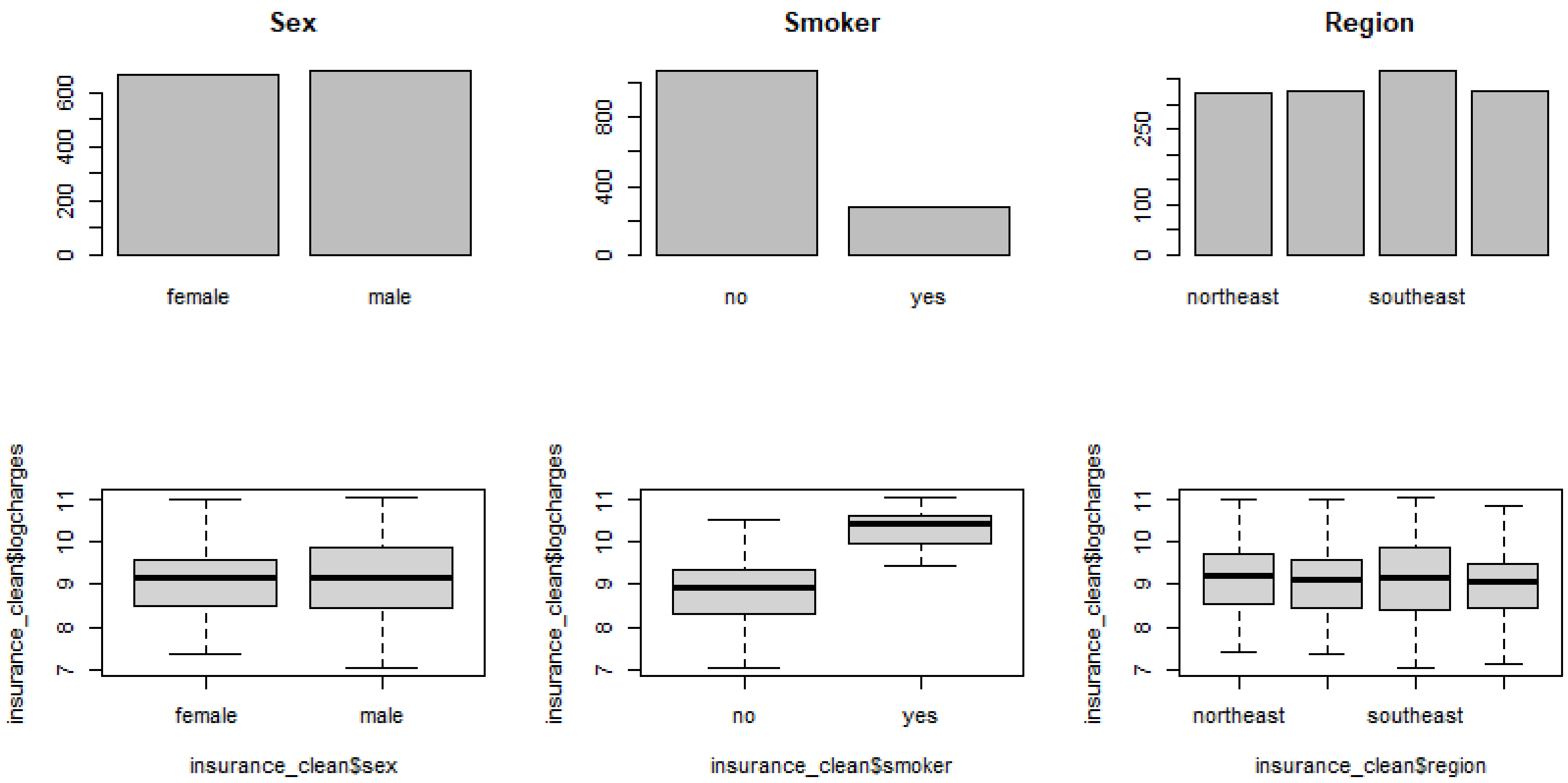


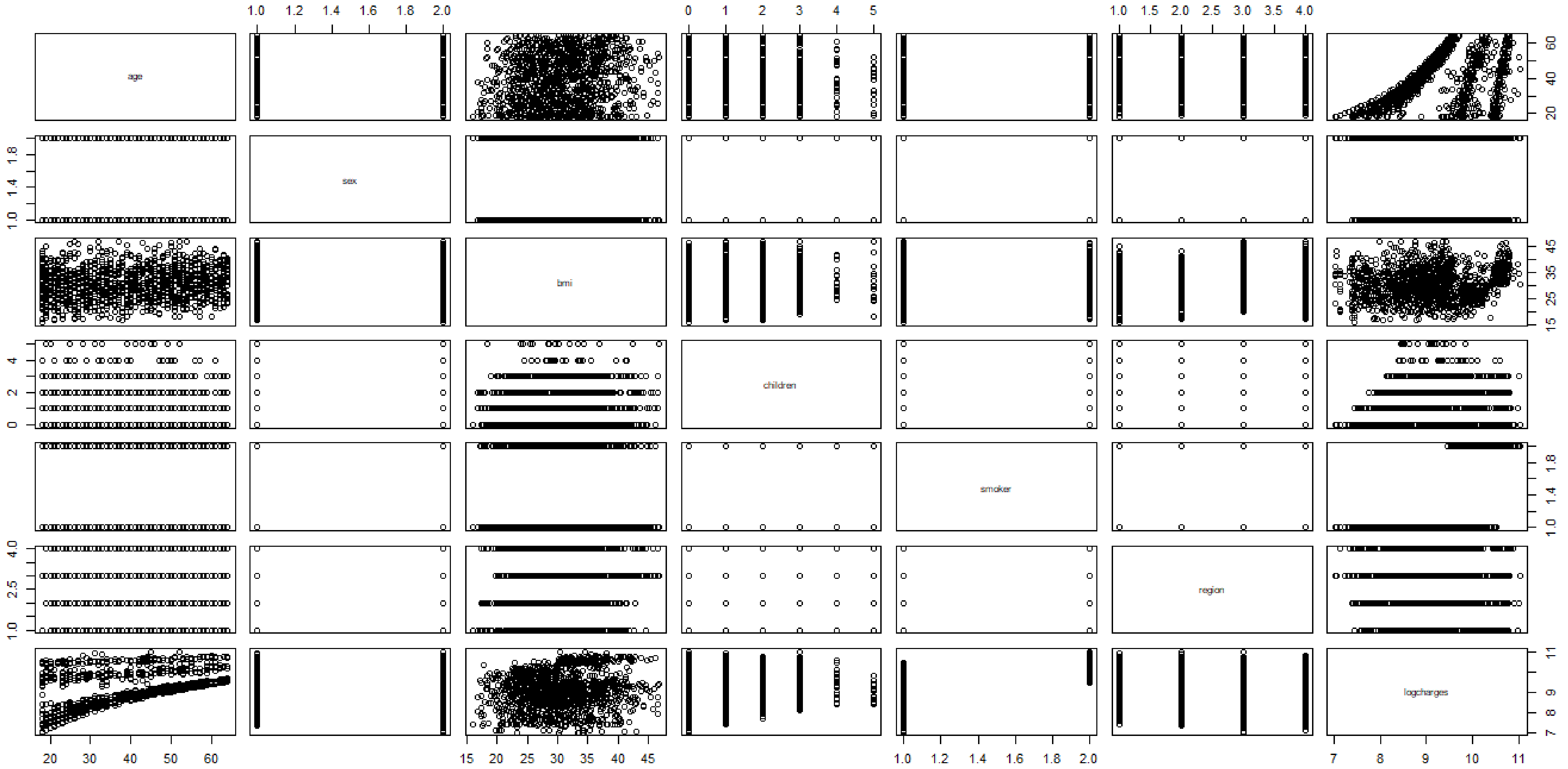
```
> summary(insurance_clean)
      age            sex           bmi           children      smoker           region       charges        logcharges    
 Min.   :18.0   female:659   Min.   :15.96   Min.   :0.000   no :1058   northeast:323   Min.   : 1122   Min.   : 7.023  
 1st Qu.:27.0   male  :670   1st Qu.:26.22   1st Qu.:0.000   yes: 271   northwest:325   1st Qu.: 4738   1st Qu.: 8.463  
 Median :39.0                Median :30.30   Median :1.000              southeast:357   Median : 9361   Median : 9.144  
 Mean   :39.2                Mean   :30.54   Mean   :1.096              southwest:324   Mean   :13212   Mean   : 9.097  
 3rd Qu.:51.0                3rd Qu.:34.48   3rd Qu.:2.000                              3rd Qu.:16587   3rd Qu.: 9.716  
 Max.   :64.0                Max.   :46.75   Max.   :5.000                              Max.   :62593   Max.   :11.044  
```

Graph for each quantitative variable



Graphs for each qualitative variable

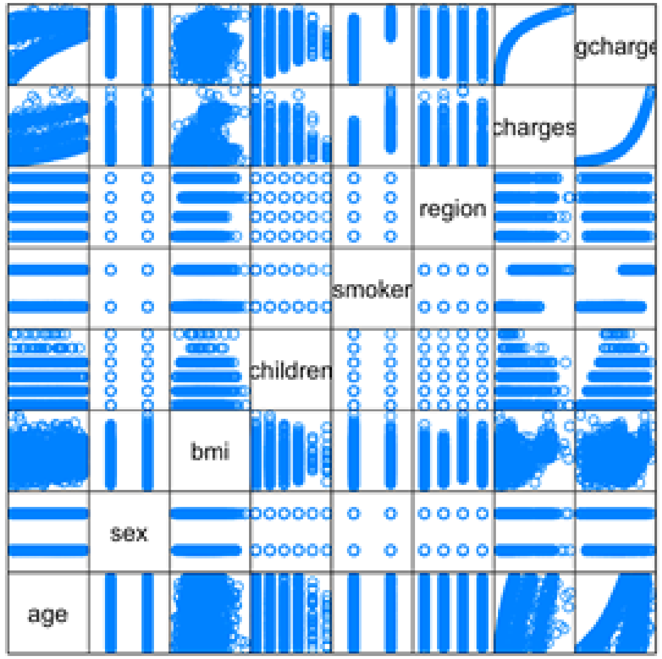




age
sex
bmi
children
smoker
region
logcharges
1.0 1.2 1.4 1.6 1.8 2.0
0 1 2 3 4 5
1.0 1.5 2.0 2.5 3.0 3.5 4.0
20 40 60
1.0 1.4 1.8
15 25 35 45
0 2 4
1.0 1.4 1.8
1.0 2.5 4.0
7 8 9 11
20 30 40 50 60
15 20 25 30 35 40 45
1.0 1.2 1.4 1.6 1.8 2.0
7 8 9 10 11

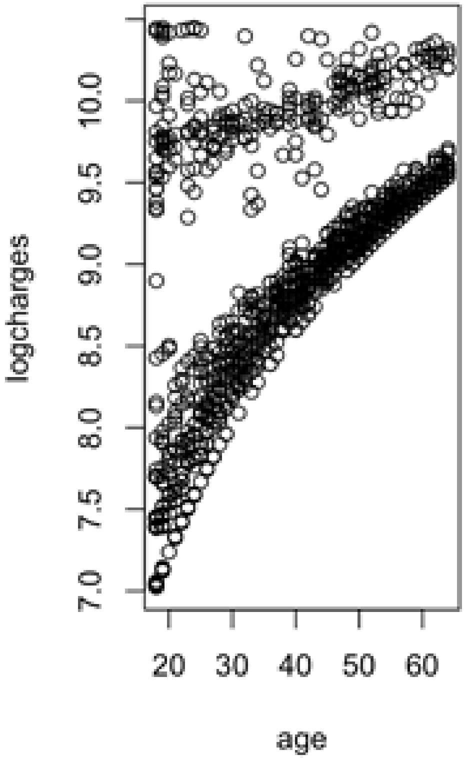
SIMPLE LINEAR MODEL REGRESSION

- SIMPLE LINEAR REGRESSION



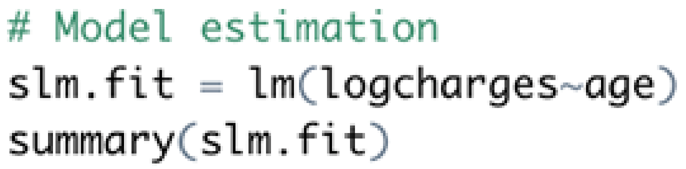
Matrice de nuages de points

By looking at the scatter-plot, we can see that the variable age seems to have a real impact on logcharges. Let's try to do a simple linear regression with these two variables

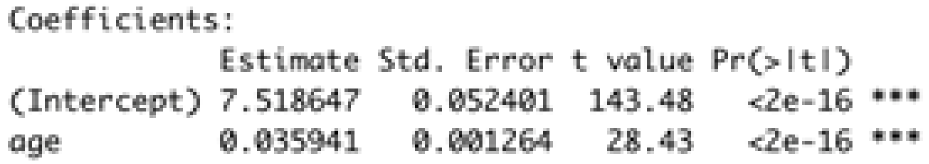


SIMPLE LINEAR MODEL REGRESSION WITH AGE

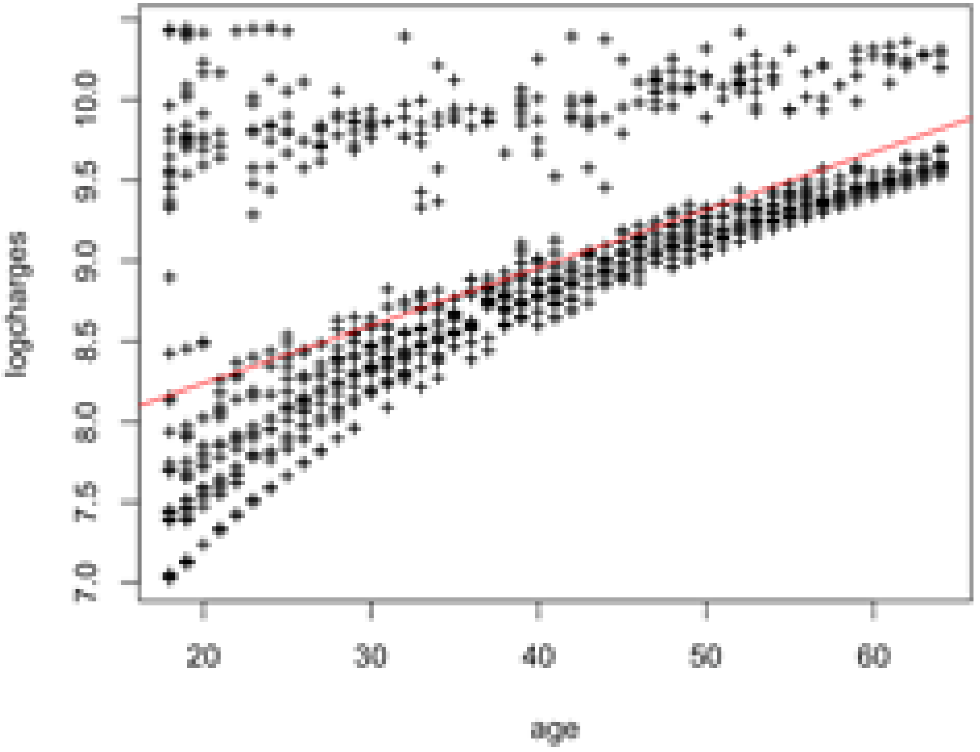
- SIMPLE LINEAR REGRESSION WITH AGE



```
# Model estimation
slm.fit = lm(logcharges~age)
summary(slm.fit)
```



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.518647   0.052401  143.48   <2e-16 ***
age         0.035941   0.001264   28.43   <2e-16 ***
```



All the p-value are less than 5%, so all the coefficient are significant.

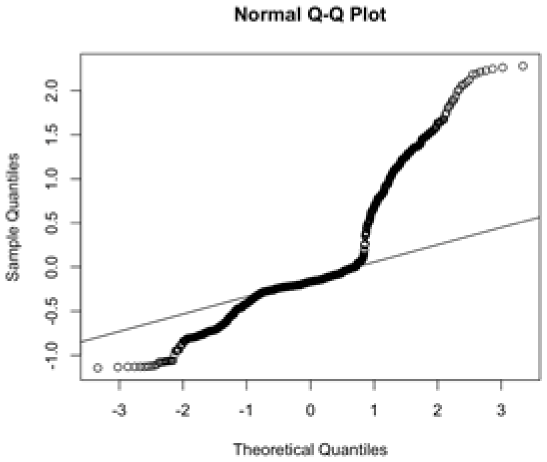
Multiple R-squared: 0.4031, Adjusted R-squared: 0.4026. This model explains 40,3% of the logcharges variations.

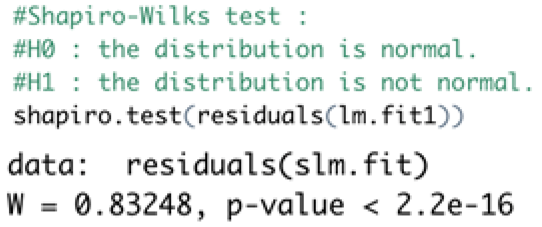
Since pvalue<5%, we can conclude that the coefficient beta_1 is significantly different from 0 and we can interpret its value.

beta_1 = 7,52 >0 so an increase by 1 unit of age impacts logcharges by an increase of 7,52.

SIMPLE LINEAR MODEL REGRESSION WITH AGE

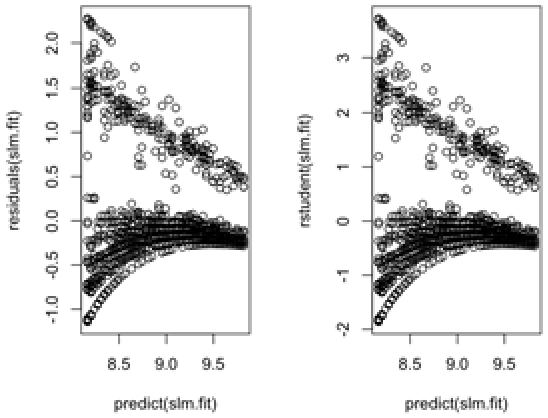
- SIMPLE LINEAR REGRESSION WITH AGE





```
#Shapiro-Wilks test :
#H0 : the distribution is normal.
#H1 : the distribution is not normal.
shapiro.test(residuals(lm.fit1))
```

```
data:  residuals(slm.fit)
W = 0.83248, p-value < 2.2e-16
```



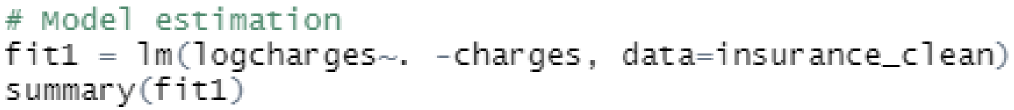
With the qqplot, we can see that the residuals are not normally distributed.

p-value < 2.2e-16 <<< 5% so we reject H0. The distribution cannot be considered as normal.

By checking at the homoscedacity, we also can see that our model can't be validate.

First estimation of the model

Multiple linear regression :



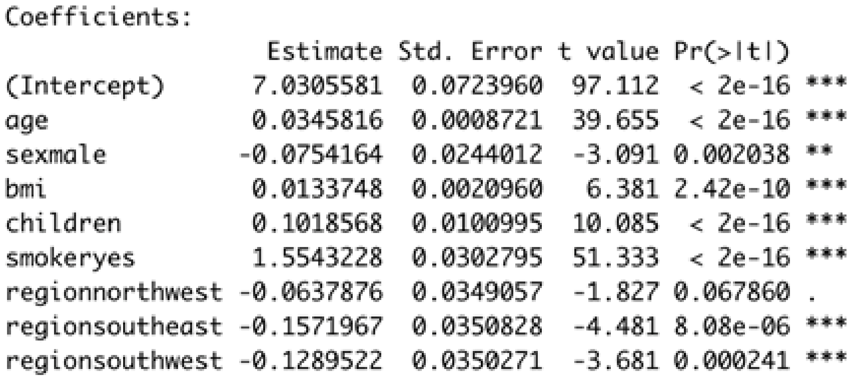
```
# Model estimation
fit1 = lm(logcharges~. -charges, data=insurance_clean)
summary(fit1)
```

We try to fit a first model with all the explanatory variables and we remove the charges variable because we use the logcharges variable.

Multiple R-squared : 0.7679

Adjusted R-squared : 0.7666

So the model is quite good.

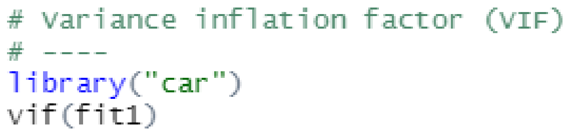


```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      7.0305581  0.0723960  97.112  < 2e-16 ***
age              0.0345816  0.0008721  39.655  < 2e-16 ***
sexmale         -0.0754164  0.0244012  -3.091 0.002038 ** 
bmi              0.0133748  0.0020960   6.381 2.42e-10 ***
children         0.1018568  0.0100995  10.085  < 2e-16 ***
smokeryes        1.5543228  0.0302795  51.333  < 2e-16 ***
regionnorthwest -0.0637876  0.0349057  -1.827 0.067860 .  
regionsoutheast -0.1571967  0.0350828  -4.481 8.08e-06 ***
regionsouthwest -0.1289522  0.0350271  -3.681 0.000241 ***
```

By looking at the pvalue (Pr(>|t|)), not all are the predictors are significant, we need to remove them. The non-significant predictor is regionnorthwest. It means that we have to remove region. We will remove it after the VIF analysis.

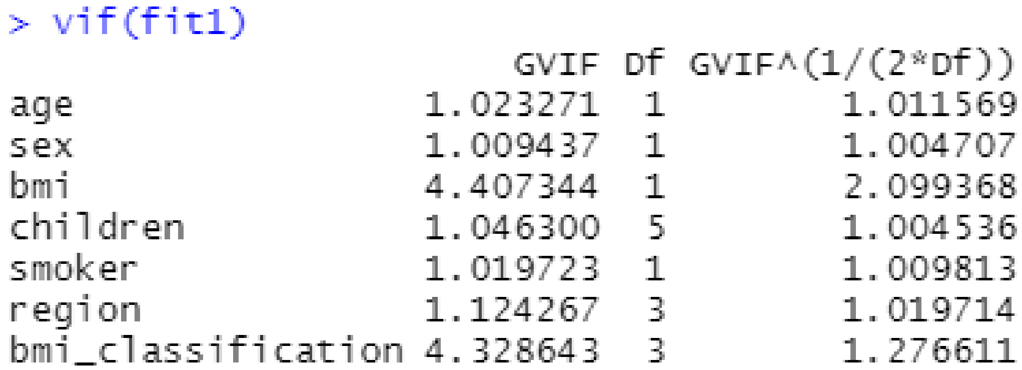
First estimation of the model

VIF analysis :



```
# variance inflation factor (VIF)
# ----
library("car")
vif(fit1)
```

The larger the VIF is, the more correlated the variables are. VIF provides the link between the variable we are considering and all the others.



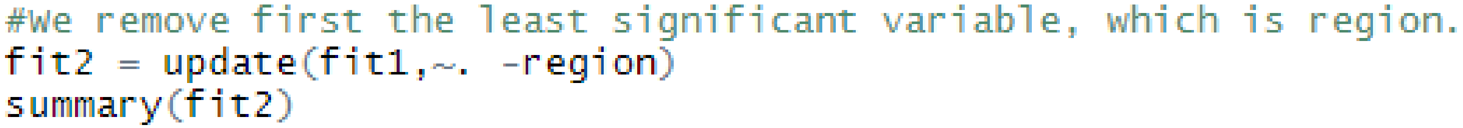
```
> vif(fit1)
                       GVIF Df GVIF^(1/(2*Df))
age                1.023271  1        1.011569
sex                1.009437  1        1.004707
bmi                4.407344  1        2.099368
children           1.046300  5        1.004536
smoker             1.019723  1        1.009813
region             1.124267  3        1.019714
bmi_classification 4.328643  3        1.276611
```

In our case, all VIF are really close to 1, so there is no problem of collinearity.

We can continue with all predictors and apply the stepwise selection.

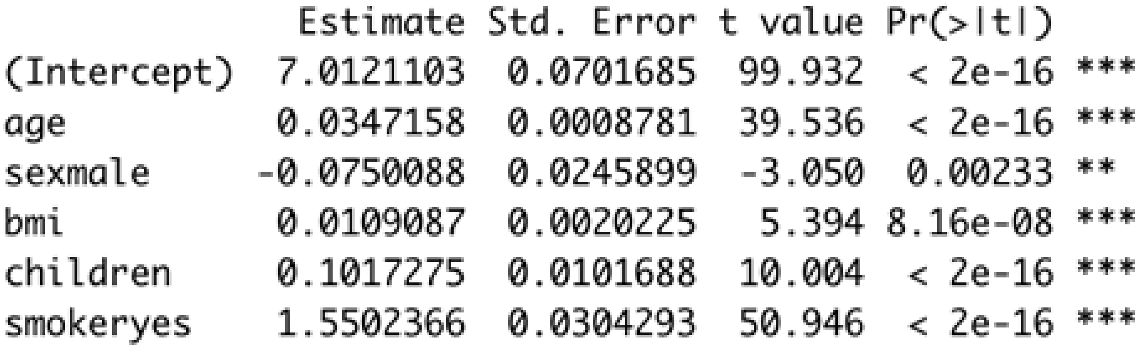
First estimation of the model

Variables stepwise selection :



```
#We remove first the least significant variable, which is region.
fit2 = update(fit1,~. -region)
summary(fit2)
```

We remove first the least significant variable, which is region.



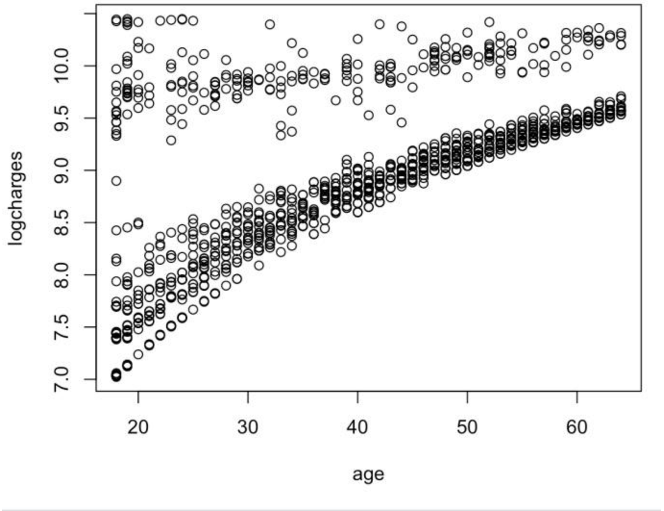
```
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.0121103  0.0701685  99.932  < 2e-16 ***
age          0.0347158  0.0008781  39.536  < 2e-16 ***
sexmale     -0.0750088  0.0245899  -3.050  0.00233 ** 
bmi          0.0109087  0.0020225   5.394 8.16e-08 ***
children     0.1017275  0.0101688  10.004  < 2e-16 ***
smokeryes    1.5502366  0.0304293  50.946  < 2e-16 ***
```

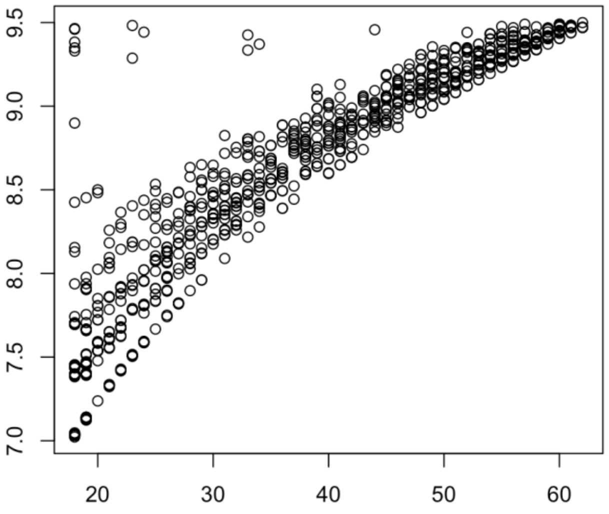
We can see that all the pvalue are less than 5%, so all the predictors are now significant.

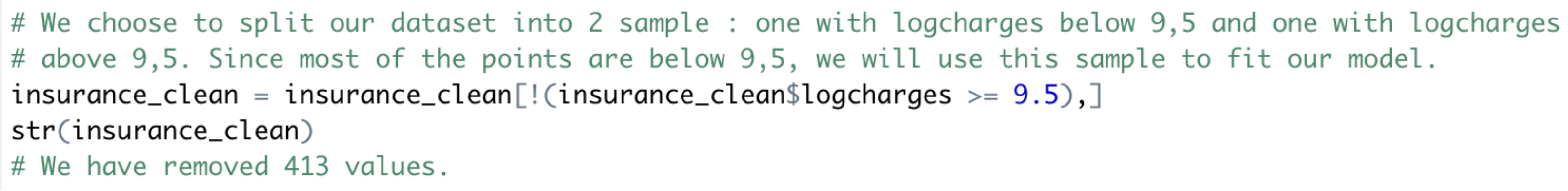
Multiple R-squared: 0.7638 and Adjusted R-squared: 0.7629, which means that the model is still quite good but not really improved.

Correction of the model

- With the plot of age and logcharges, we can see that it seems to be 2 types of populations depending on the age.



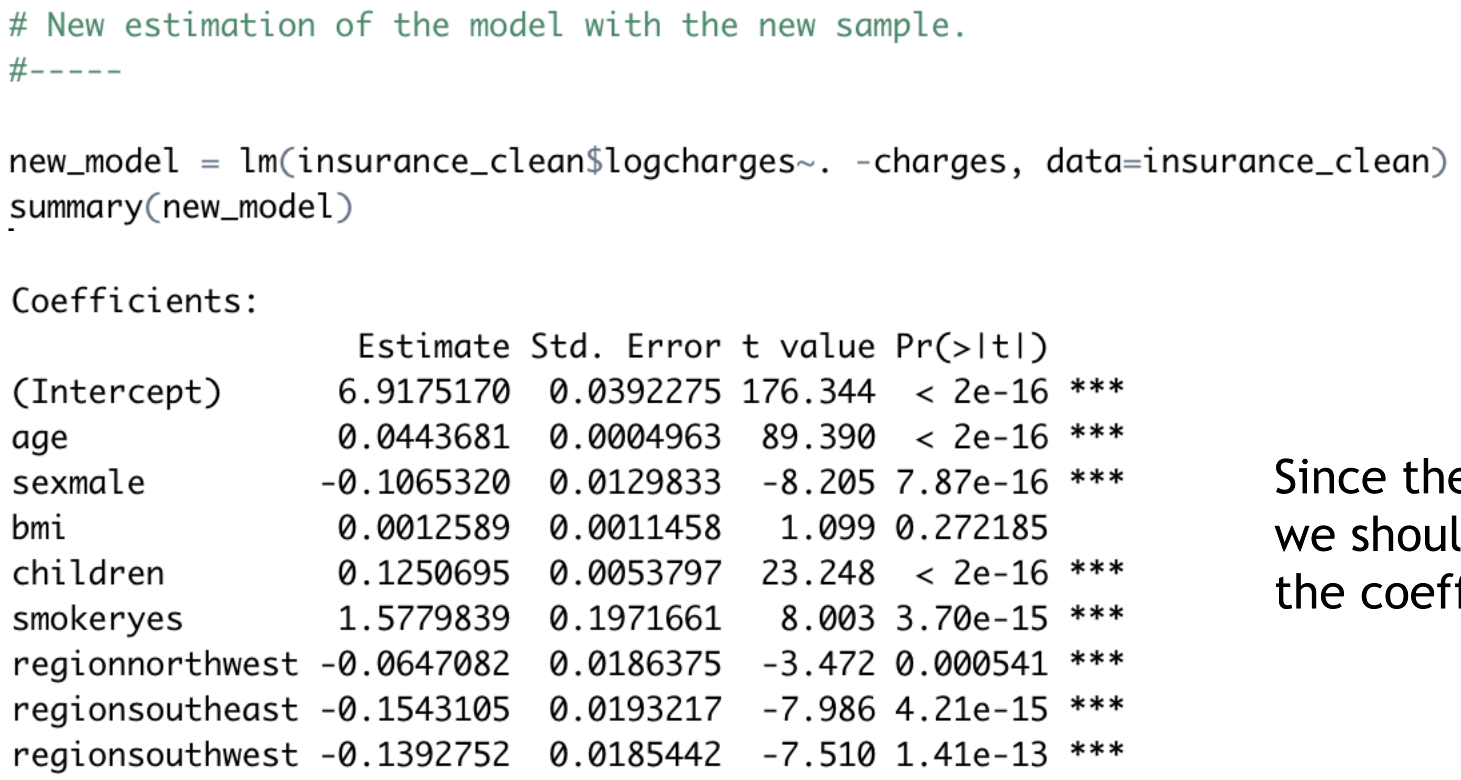




```
# We choose to split our dataset into 2 sample : one with logcharges below 9,5 and one with logcharges
# above 9,5. Since most of the points are below 9,5, we will use this sample to fit our model.
insurance_clean = insurance_clean[!(insurance_clean$logcharges >= 9.5),]
str(insurance_clean)
# We have removed 413 values.
```

Correction of the multiple linear regression

- Model fitting with the new sample.



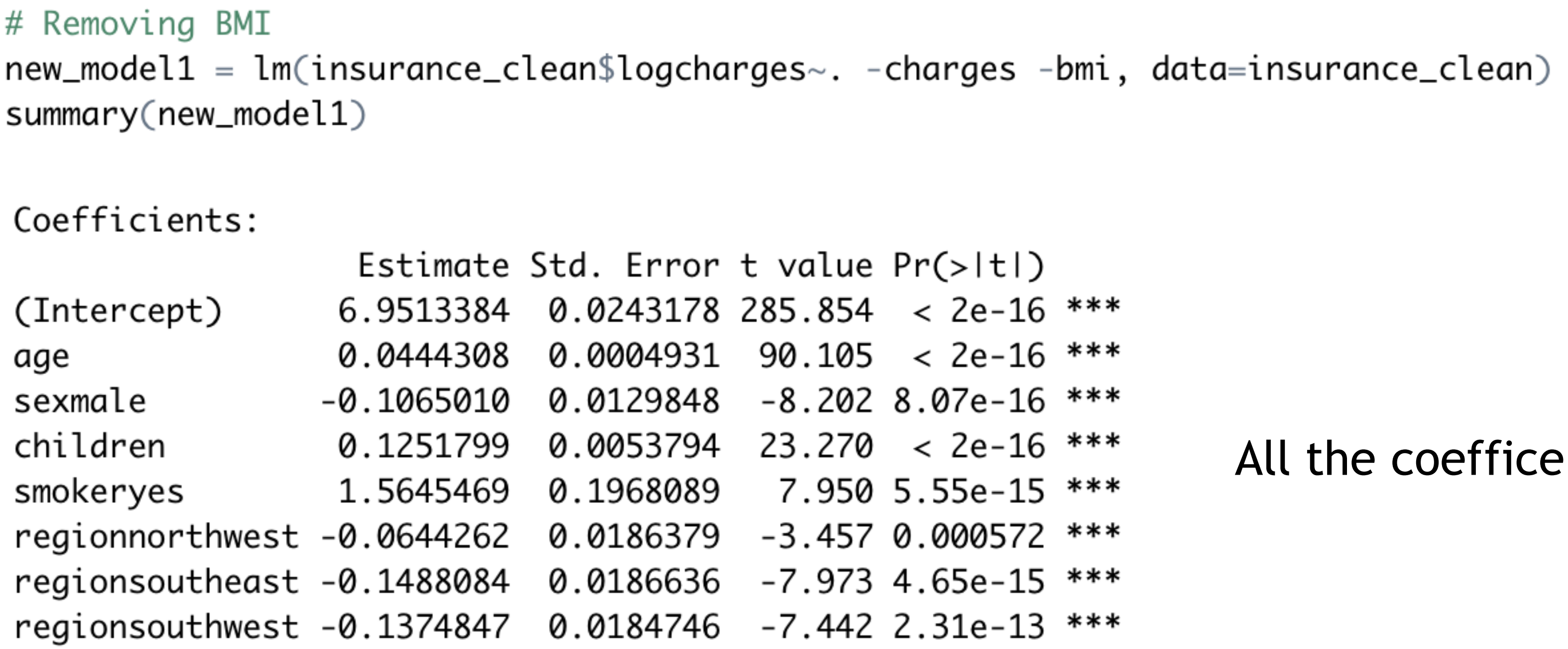
```
# New estimation of the model with the new sample.
#-----

new_model = lm(insurance_clean$logcharges~. -charges, data=insurance_clean)
summary(new_model)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      6.9175170  0.0392275 176.344  < 2e-16 ***
age              0.0443681  0.0004963  89.390  < 2e-16 ***
sexmale         -0.1065320  0.0129833  -8.205 7.87e-16 ***
bmi              0.0012589  0.0011458   1.099 0.272185    
children         0.1250695  0.0053797  23.248  < 2e-16 ***
smokeryes        1.5779839  0.1971661   8.003 3.70e-15 ***
regionnorthwest -0.0647082  0.0186375  -3.472 0.000541 ***
regionsoutheast -0.1543105  0.0193217  -7.986 4.21e-15 ***
regionsouthwest -0.1392752  0.0185442  -7.510 1.41e-13 ***
```

Since the pvalue of bmi is larger than 5%, we should remove it because it means that the coefficient is not significant.

Multiple R-squared: 0.9097, Adjusted R-squared: 0.9089
→ this model is very good and fit with the data.

Correction of the multiple linear regression

- Model fitting with the new sample.



```
# Removing BMI
new_model1 = lm(insurance_clean$logcharges~. -charges -bmi, data=insurance_clean)
summary(new_model1)
```

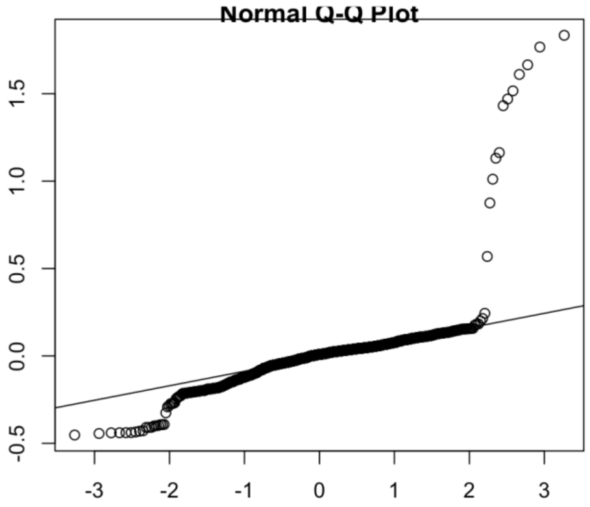
```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     6.9513384  0.0243178 285.854  < 2e-16 ***
age             0.0444308  0.0004931  90.105  < 2e-16 ***
sexmale        -0.1065010  0.0129848  -8.202 8.07e-16 ***
children        0.1251799  0.0053794  23.270  < 2e-16 ***
smokeryes       1.5645469  0.1968089   7.950 5.55e-15 ***
regionnorthwest -0.0644262  0.0186379  -3.457 0.000572 ***
regionsoutheast -0.1488084  0.0186636  -7.973 4.65e-15 ***
regionsouthwest -0.1374847  0.0184746  -7.442 2.31e-13 ***
```

All the coefficents are now significant.

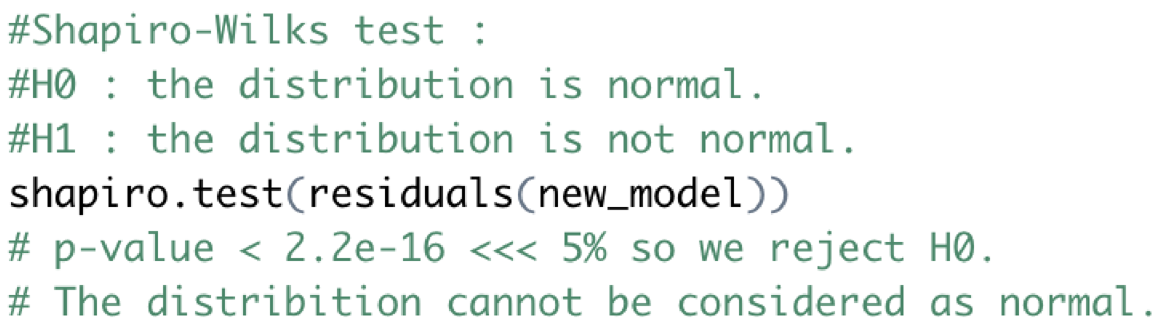
Multiple R-squared: 0.9096,   Adjusted R-squared: 0.9089
→ this model is very good and fit with the data.

Correction of the multiple linear regression

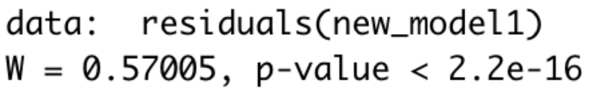
- Residual analysis for model validation



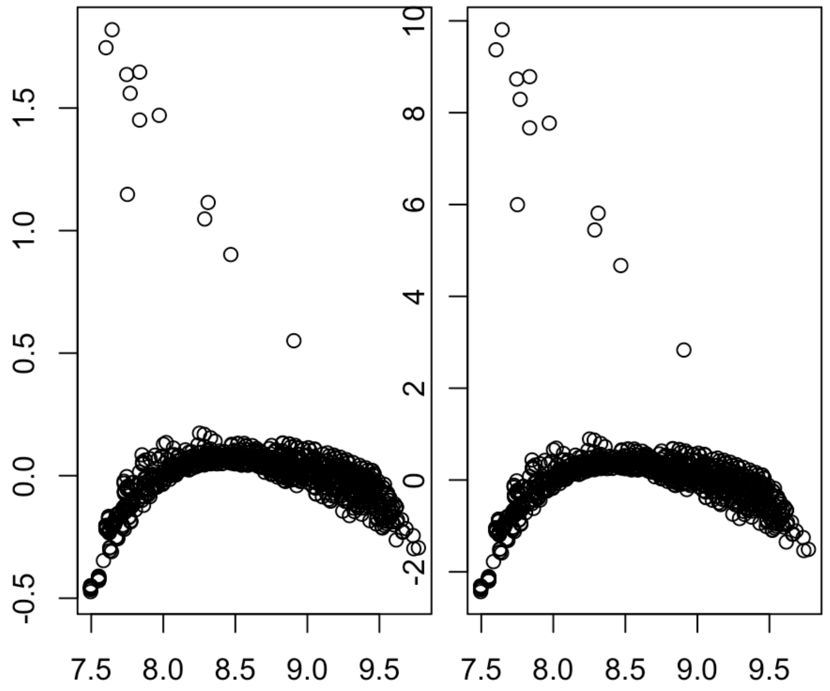
Residuals are not normally distributed.



```
#Shapiro-Wilks test :
#H0 : the distribution is normal.
#H1 : the distribution is not normal.
shapiro.test(residuals(new_model))
# p-value < 2.2e-16 <<< 5% so we reject H0.
# The distribition cannot be considered as normal.
```



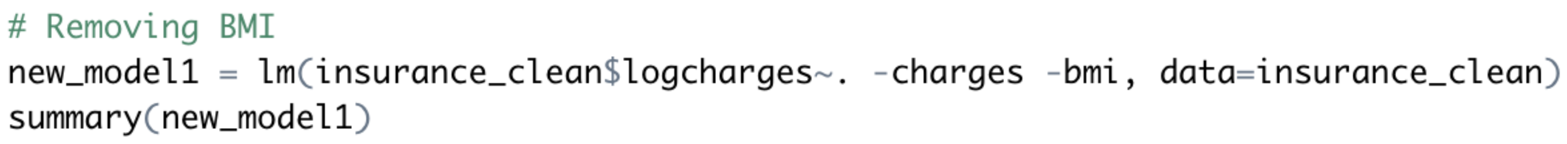
```
data:  residuals(new_model1)
W = 0.57005, p-value < 2.2e-16
```



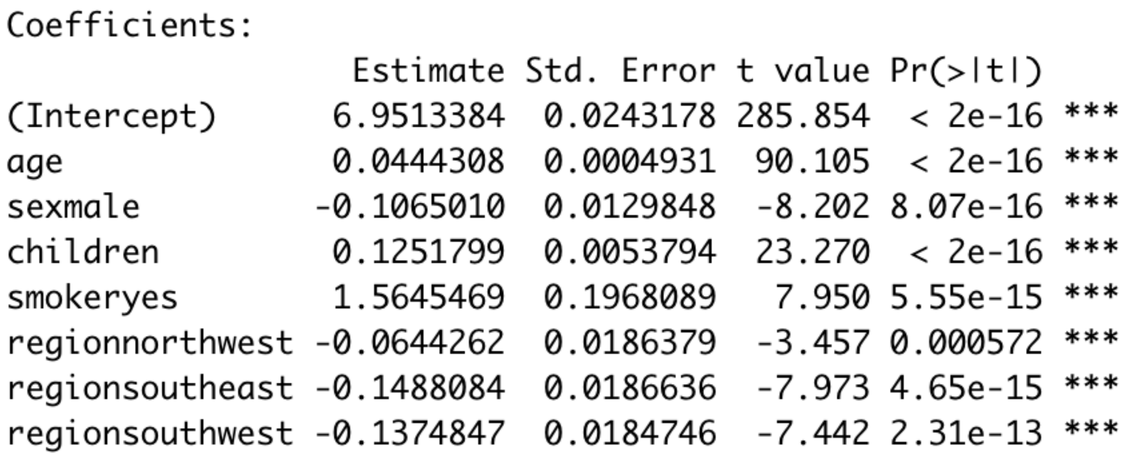
We cannot validate our model.

Correction of the multiple linear regression

- Model fitting with the new sample.



```
# Removing BMI
new_model1 = lm(insurance_clean$logcharges~. -charges -bmi, data=insurance_clean)
summary(new_model1)
```



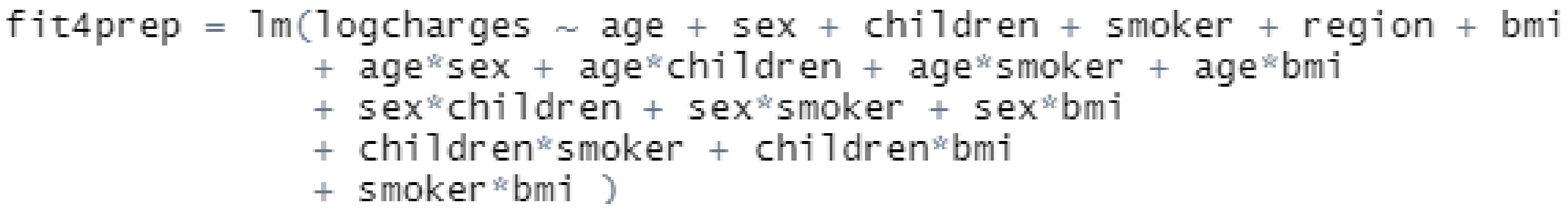
```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      6.9513384  0.0243178 285.854  < 2e-16 ***
age              0.0444308  0.0004931  90.105  < 2e-16 ***
sexmale         -0.1065010  0.0129848  -8.202 8.07e-16 ***
children         0.1251799  0.0053794  23.270  < 2e-16 ***
smokeryes        1.5645469  0.1968089   7.950 5.55e-15 ***
regionnorthwest -0.0644262  0.0186379  -3.457 0.000572 ***
regionsoutheast -0.1488084  0.0186636  -7.973 4.65e-15 ***
regionsouthwest -0.1374847  0.0184746  -7.442 2.31e-13 ***
```

All the coefficents are now significant.

Multiple R-squared: 0.9096,   Adjusted R-squared: 0.9089
→ this model is very good and fit with the data.

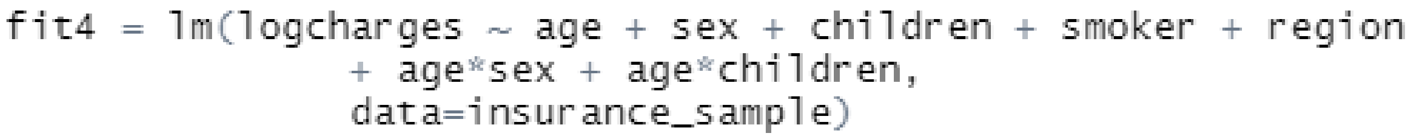
Including interaction effects

- Targetedly observe potential interaction effects

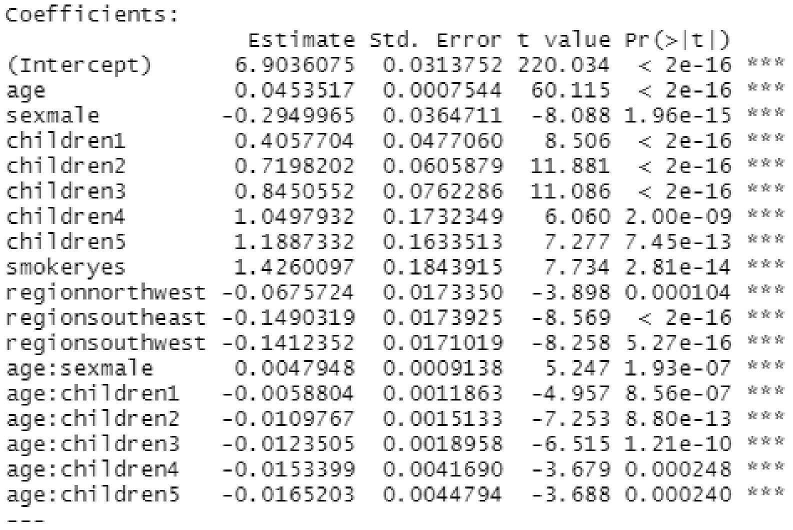


```
fit4prep = lm(logcharges ~ age + sex + children + smoker + region + bmi
                   + age*sex + age*children + age*smoker + age*bmi
                   + sex*children + sex*smoker + sex*bmi
                   + children*smoker + children*bmi
                   + smoker*bmi )
```

- Keep significant interaction terms:



```
fit4 = lm(logcharges ~ age + sex + children + smoker + region
               + age*sex + age*children,
               data=insurance_sample)
```



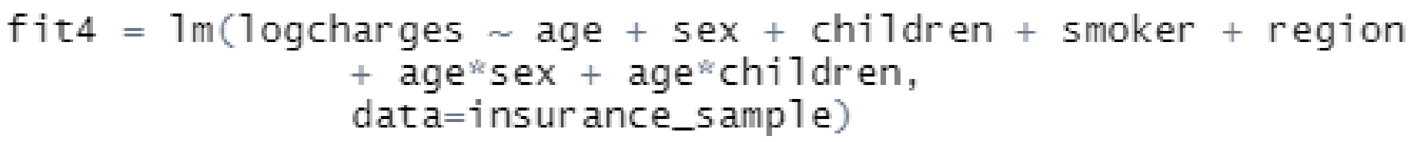
```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       6.9036075  0.0313752 220.034  < 2e-16 ***
age               0.0453517  0.0007544  60.115  < 2e-16 ***
sexmale          -0.2949965  0.0364711  -8.088 1.96e-15 ***
children1         0.4057704  0.0477060   8.506  < 2e-16 ***
children2         0.7198202  0.0605879  11.881  < 2e-16 ***
children3         0.8450552  0.0762286  11.086  < 2e-16 ***
children4         1.0497932  0.1732349   6.060 2.00e-09 ***
children5         1.1887332  0.1633513   7.277 7.45e-13 ***
smokeryes         1.4260097  0.1843915   7.734 2.81e-14 ***
regionnorthwest -0.0675724  0.0173350  -3.898 0.000104 ***
regionsoutheast -0.1490319  0.0173925  -8.569  < 2e-16 ***
regionsouthwest -0.1412352  0.0171019  -8.258 5.27e-16 ***
age:sexmale       0.0047948  0.0009138   5.247 1.93e-07 ***
age:children1    -0.0058804  0.0011863  -4.957 8.56e-07 ***
age:children2    -0.0109767  0.0015133  -7.253 8.80e-13 ***
age:children3    -0.0123505  0.0018958  -6.515 1.21e-10 ***
age:children4    -0.0153399  0.0041690  -3.679 0.000248 ***
age:children5    -0.0165203  0.0044794  -3.688 0.000240 ***
---
```

All the coefficents are now significant.

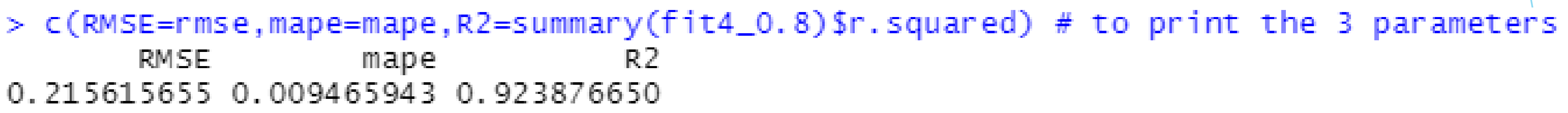
Multiple R-squared: 0.9239,   Adjusted R-squared: 0.9224
→ the inclusion of significant interaction terms improved the model fit.

Including interaction effects

- Investigate predictive power of the model fit4 using train and test set (train set = 80 %-sample):

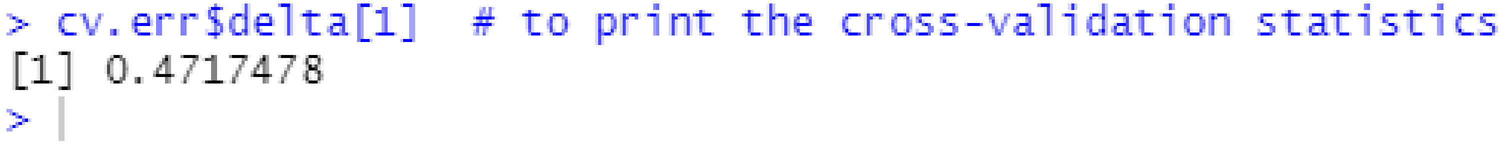


```
fit4 = lm(logcharges ~ age + sex + children + smoker + region
               + age*sex + age*children,
               data=insurance_sample)
```



```
> c(RMSE=rmse,mape=mape,R2=summary(fit4_0.8)$r.squared) # to print the 3 parameters
       RMSE        mape          R2 
0.215615655 0.009465943 0.923876650 
```

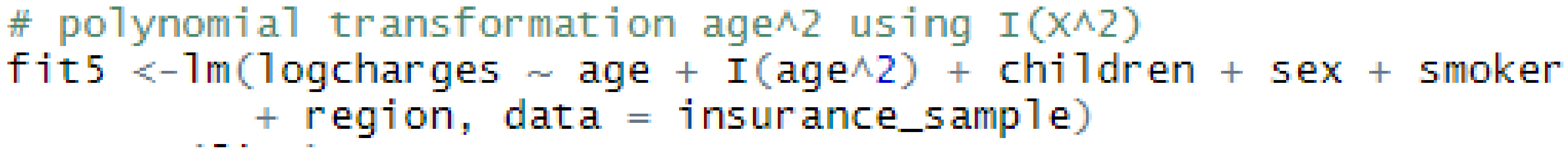
- Based on LOOCV-approach:



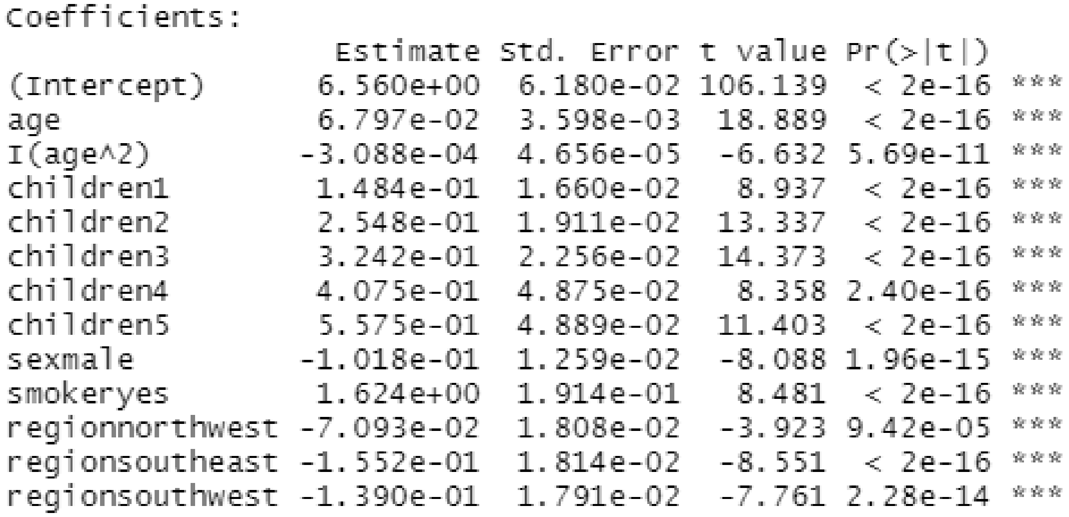
```
> cv.err$delta[1]  # to print the cross-validation statistics
[1] 0.4717478
> |
```

Polynomial transformation

- Considering age as highly explaining variable (based on preprocessing)



```
# polynomial transformation age^2 using I(X^2)
fit5 <-lm(logcharges ~ age + I(age^2) + children + sex + smoker
        + region, data = insurance_sample)
```



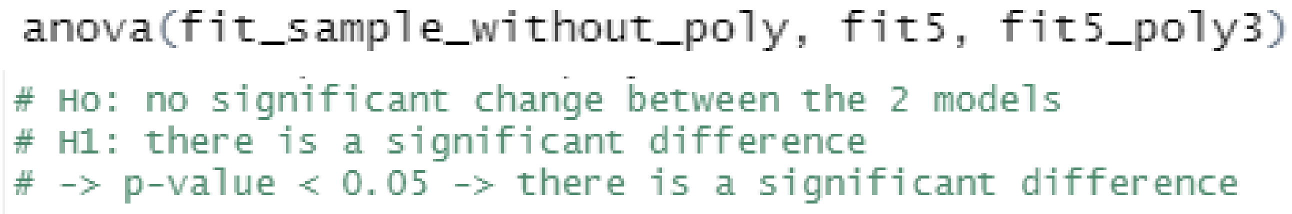
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      6.560e+00  6.180e-02 106.139  < 2e-16 ***
age              6.797e-02  3.598e-03  18.889  < 2e-16 ***
I(age^2)        -3.088e-04  4.656e-05  -6.632 5.69e-11 ***
children1        1.484e-01  1.660e-02   8.937  < 2e-16 ***
children2        2.548e-01  1.911e-02  13.337  < 2e-16 ***
children3        3.242e-01  2.256e-02  14.373  < 2e-16 ***
children4        4.075e-01  4.875e-02   8.358 2.40e-16 ***
children5        5.575e-01  4.889e-02  11.403  < 2e-16 ***
sexmale         -1.018e-01  1.259e-02  -8.088 1.96e-15 ***
smokeryes        1.624e+00  1.914e-01   8.481  < 2e-16 ***
regionnorthwest -7.093e-02  1.808e-02  -3.923 9.42e-05 ***
regionsoutheast -1.552e-01  1.814e-02  -8.551  < 2e-16 ***
regionsouthwest -1.390e-01  1.791e-02  -7.761 2.28e-14 ***
```

Polynomial term significant.
Multiple R-squared: 0.9158,	Adjusted R-squared: 0.9147

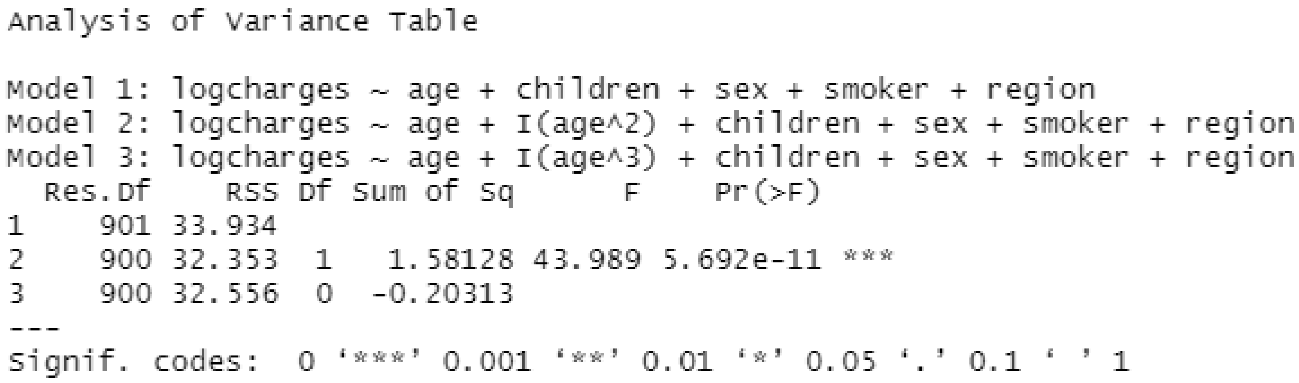
Improved fit compared to sam model without polynomial term "fit_sample_without_poly": (Multiple R-squared: 0.9117,	Adjusted R-squared: 0.9106)

Polynomial transformation

- Anova test: is the additional polynomial term significant?



```
anova(fit_sample_without_poly, fit5, fit5_poly3)
# Ho: no significant change between the 2 models
# H1: there is a significant difference
# -> p-value < 0.05 -> there is a significant difference
```

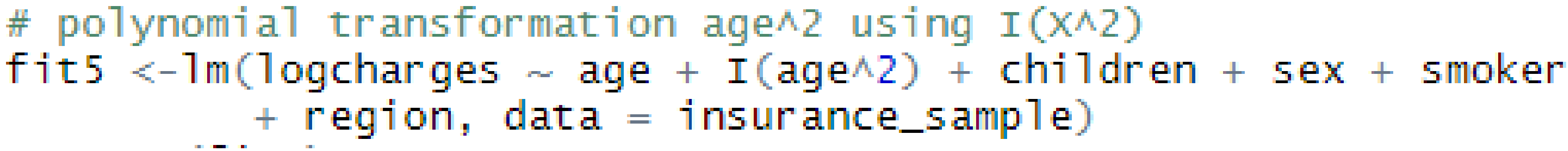


```
Analysis of Variance Table

Model 1: logcharges ~ age + children + sex + smoker + region
Model 2: logcharges ~ age + I(age^2) + children + sex + smoker + region
Model 3: logcharges ~ age + I(age^3) + children + sex + smoker + region
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1    901 33.934                                  
2    900 32.353  1   1.58128 43.989 5.692e-11 ***
3    900 32.556  0  -0.20313                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- p-value (fit5) < 0.05 -> the additional polynomial term is significant.
- The best model seems to be Model2 (fit5) since pvalue for the inclusion of I(age^3) indicates non-significance

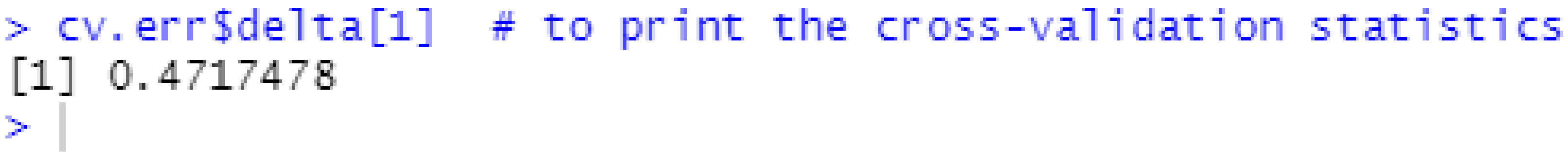
Polynomial transformation

- Investigate predictive power of the model fit5



```
# polynomial transformation age^2 using I(X^2)
fit5 <-lm(logcharges ~ age + I(age^2) + children + sex + smoker
          + region, data = insurance_sample)
```

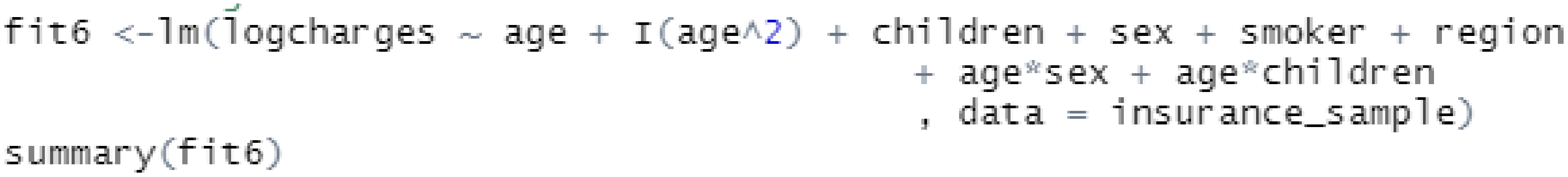
- Using LOOCV-approach:



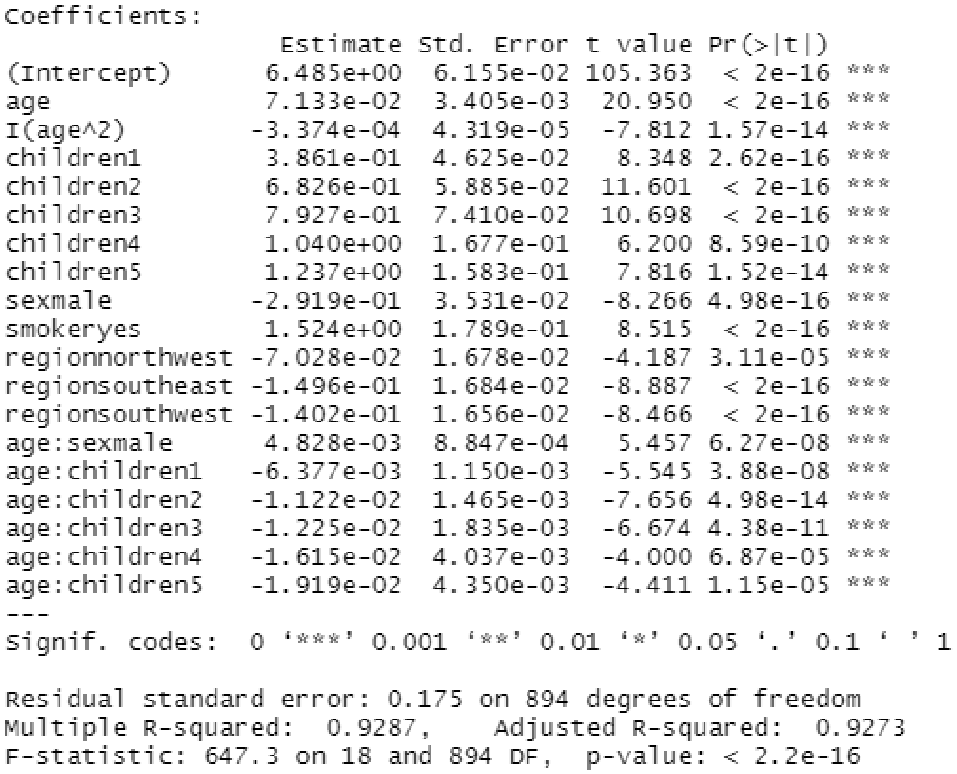
```
> cv.err$delta[1]  # to print the cross-validation statistics
[1] 0.4717478
> 
```

Polynomial transformation

- Including relevant interaction effects



```
fit6 <-lm(logcharges ~ age + I(age^2) + children + sex + smoker + region
                                      + age*sex + age*children
                                      , data = insurance_sample)

summary(fit6)
```

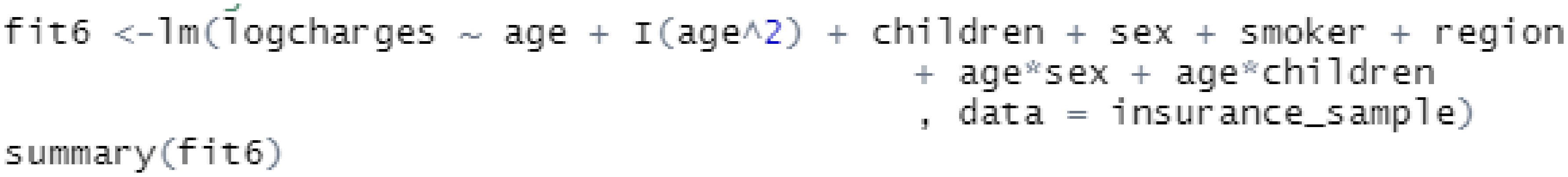


```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       6.485e+00  6.155e-02 105.363  < 2e-16 ***
age               7.133e-02  3.405e-03  20.950  < 2e-16 ***
I(age^2)         -3.374e-04  4.319e-05  -7.812 1.57e-14 ***
children1         3.861e-01  4.625e-02   8.348 2.62e-16 ***
children2         6.826e-01  5.885e-02  11.601  < 2e-16 ***
children3         7.927e-01  7.410e-02  10.698  < 2e-16 ***
children4         1.040e+00  1.677e-01   6.200 8.59e-10 ***
children5         1.237e+00  1.583e-01   7.816 1.52e-14 ***
sexmale          -2.919e-01  3.531e-02  -8.266 4.98e-16 ***
smokeryes         1.524e+00  1.789e-01   8.515  < 2e-16 ***
regionnorthwest -7.028e-02  1.678e-02  -4.187 3.11e-05 ***
regionsoutheast -1.496e-01  1.684e-02  -8.887  < 2e-16 ***
regionsouthwest -1.402e-01  1.656e-02  -8.466  < 2e-16 ***
age:sexmale       4.828e-03  8.847e-04   5.457 6.27e-08 ***
age:children1    -6.377e-03  1.150e-03  -5.545 3.88e-08 ***
age:children2    -1.122e-02  1.465e-03  -7.656 4.98e-14 ***
age:children3    -1.225e-02  1.835e-03  -6.674 4.38e-11 ***
age:children4    -1.615e-02  4.037e-03  -4.000 6.87e-05 ***
age:children5    -1.919e-02  4.350e-03  -4.411 1.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.175 on 894 degrees of freedom
Multiple R-squared:  0.9287,	Adjusted R-squared:  0.9273 
F-statistic: 647.3 on 18 and 894 DF,  p-value: < 2.2e-16
```

- Multiple R-squared: 0.9287,   Adjusted R-squared: 0.9273
- → so far best model fit

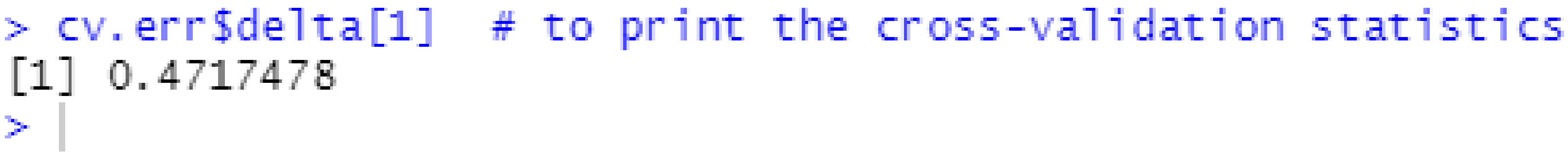
Polynomial transformation

- Investigate predictive power of the model fit5



```
fit6 <-lm(logcharges ~ age + I(age^2) + children + sex + smoker + region
                                            + age*sex + age*children
                                            , data = insurance_sample)
summary(fit6)
```

- Using LOOCV-approach:

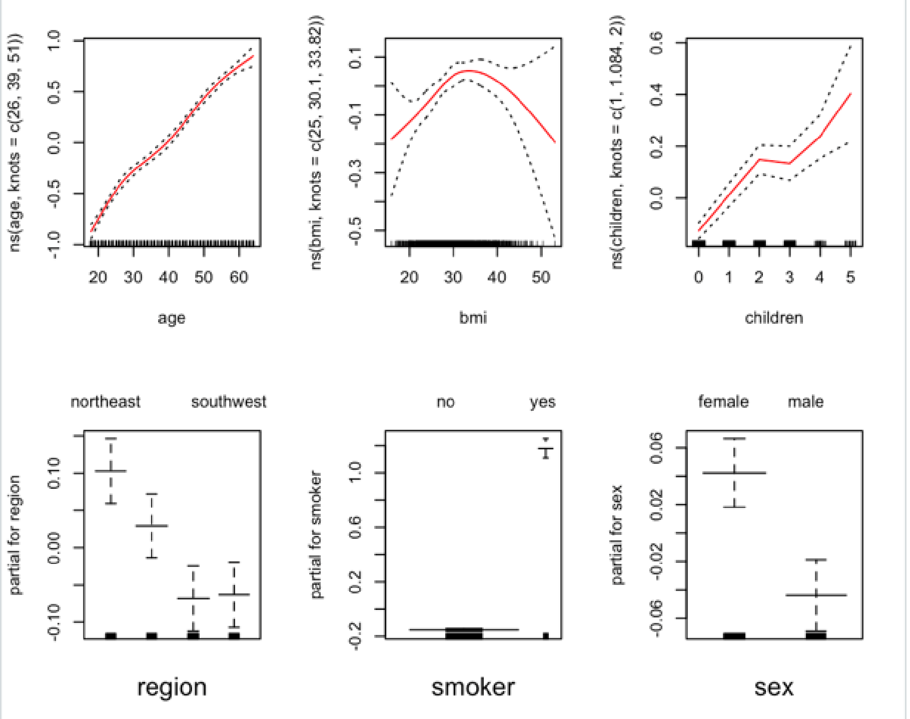


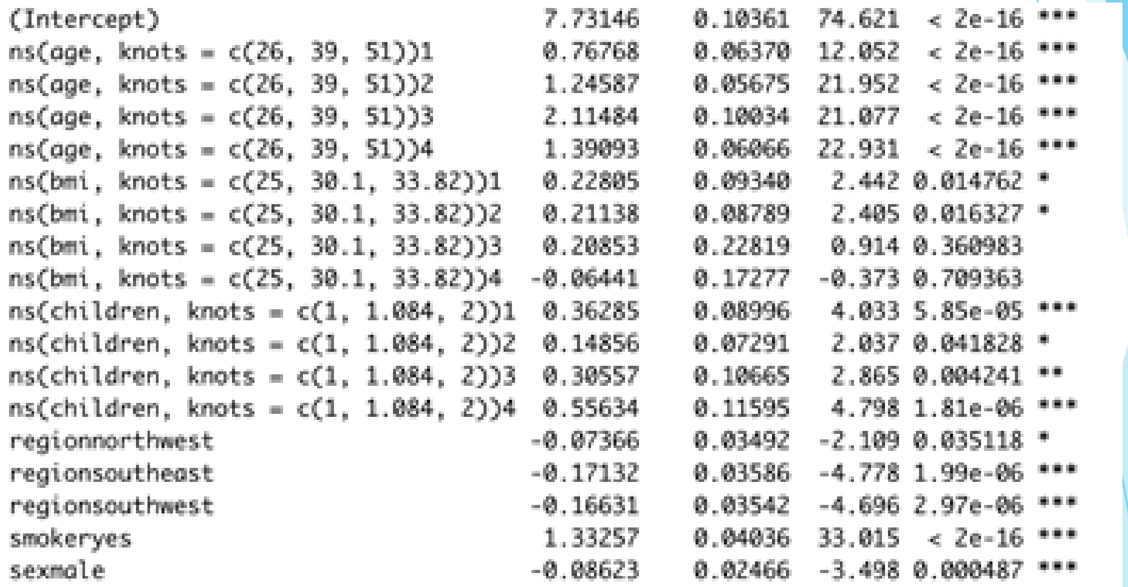
```
> cv.err$delta[1]  # to print the cross-validation statistics
[1] 0.4717478
> 
```

GAMs EXTENSIONS

- GAMs extension to multiple linear regression

```
# GAM using quartiles for the knots
gam1 = lm(logcharges~ns(age, knots=c(26, 39, 51))+ns(bmi, knots=c(25, 30.10, 33.82))
          +ns(children, knots=c(1, 1.084, 2))+
          +region+smoker+sex, data=insurance_clean)
summary(gam1)

par(mfrow=c(2,3))
plot.Gam(gam1, se=TRUE, col="red")
```





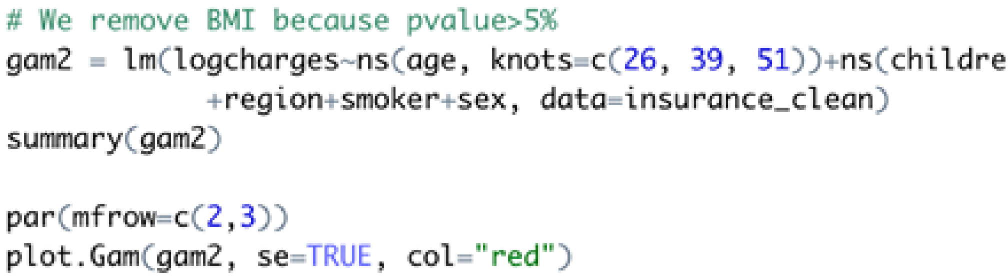
```
(Intercept)                          7.73146    0.10361  74.621  < 2e-16 ***
ns(age, knots = c(26, 39, 51))1      0.76768    0.06370  12.052  < 2e-16 ***
ns(age, knots = c(26, 39, 51))2      1.24587    0.05675  21.952  < 2e-16 ***
ns(age, knots = c(26, 39, 51))3      2.11484    0.10034  21.077  < 2e-16 ***
ns(age, knots = c(26, 39, 51))4      1.39093    0.06066  22.931  < 2e-16 ***
ns(bmi, knots = c(25, 30.1, 33.82))1  0.22805    0.09340   2.442 0.014762 *
ns(bmi, knots = c(25, 30.1, 33.82))2  0.21138    0.08789   2.405 0.016327 *
ns(bmi, knots = c(25, 30.1, 33.82))3  0.20853    0.22819   0.914 0.360983
ns(bmi, knots = c(25, 30.1, 33.82))4 -0.06441    0.17277  -0.373 0.709363
ns(children, knots = c(1, 1.084, 2))1  0.36285    0.08996   4.033 5.85e-05 ***
ns(children, knots = c(1, 1.084, 2))2  0.14856    0.07291   2.037 0.041828 *
ns(children, knots = c(1, 1.084, 2))3  0.30557    0.10665   2.865 0.004241 **
ns(children, knots = c(1, 1.084, 2))4  0.55634    0.11595   4.798 1.81e-06 ***
regionnorthwest                     -0.07366    0.03492  -2.109 0.035118 *
regionsoutheast                     -0.17132    0.03586  -4.778 1.99e-06 ***
regionsouthwest                     -0.16631    0.03542  -4.696 2.97e-06 ***
smokeryes                            1.33257    0.04036  33.015  < 2e-16 ***
sexmale                             -0.08623    0.02466  -3.498 0.000487 ***
```

Pvalue of BMI is larger than 5%, so we will remove it.

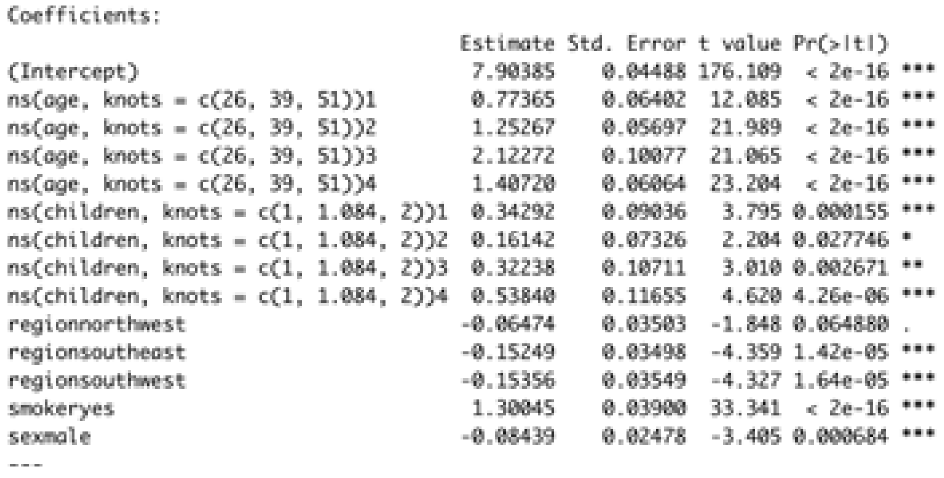
Multiple R-squared: 0.7179,   Adjusted R-squared: 0.7138

GAMs EXTENSIONS

- GAMs extension to multiple linear regression

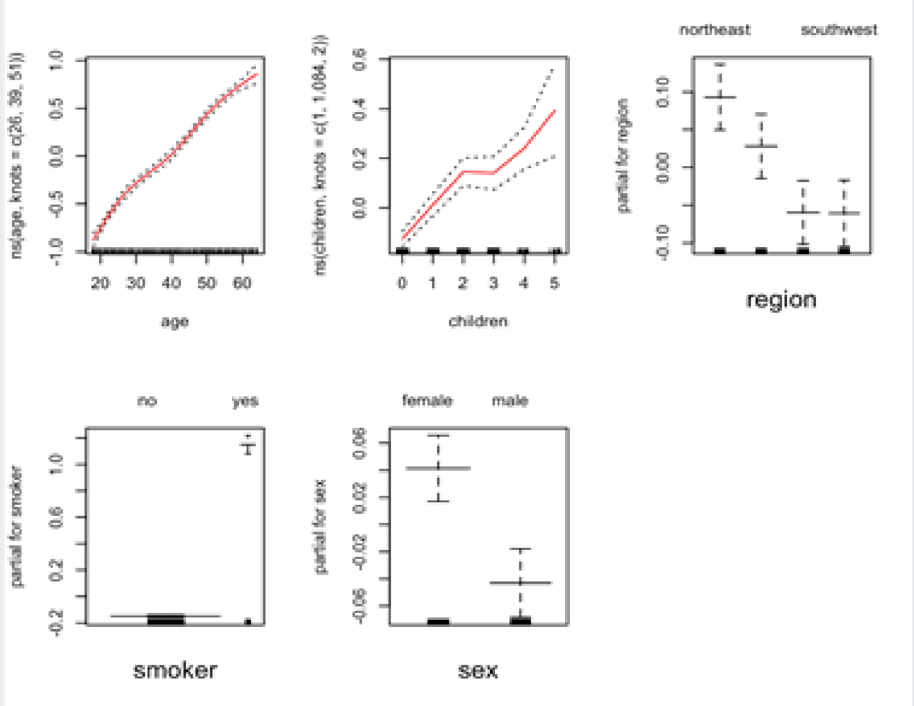


```
# We remove BMI because pvalue>5%
gam2 = lm(logcharges~ns(age, knots=c(26, 39, 51))+ns(children, knots=c(1, 1.084, 2))+
          +region+smoker+sex, data=insurance_clean)
summary(gam2)

par(mfrow=c(2,3))
plot.Gam(gam2, se=TRUE, col="red")
```



```
Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         7.90385    0.04488 176.109  < 2e-16 ***
ns(age, knots = c(26, 39, 51))1     0.77365    0.06402  12.085  < 2e-16 ***
ns(age, knots = c(26, 39, 51))2     1.25267    0.05697  21.989  < 2e-16 ***
ns(age, knots = c(26, 39, 51))3     2.12272    0.10077  21.065  < 2e-16 ***
ns(age, knots = c(26, 39, 51))4     1.40720    0.06064  23.204  < 2e-16 ***
ns(children, knots = c(1, 1.084, 2))1  0.34292    0.09036   3.795 0.000155 ***
ns(children, knots = c(1, 1.084, 2))2  0.16142    0.07326   2.204 0.027746 *  
ns(children, knots = c(1, 1.084, 2))3  0.32238    0.10711   3.010 0.002671 ** 
ns(children, knots = c(1, 1.084, 2))4  0.53840    0.11655   4.620 4.26e-06 ***
regionnorthwest                    -0.06474    0.03503  -1.848 0.064880 .  
regionsoutheast                    -0.15249    0.03498  -4.359 1.42e-05 ***
regionsouthwest                    -0.15356    0.03549  -4.327 1.64e-05 ***
smokeryes                           1.30045    0.03900  33.341  < 2e-16 ***
sexmale                            -0.08439    0.02478  -3.405 0.000684 ***
---
```

All the coefficients are now significant.

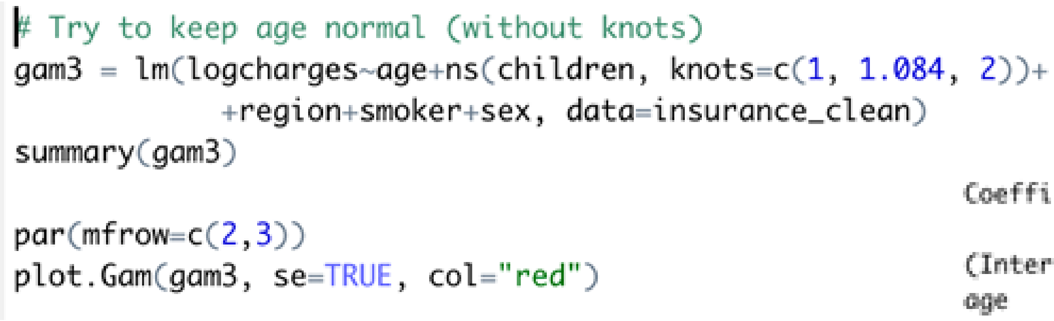


As it is already fitting well, we can try to remove knots for the variable age.

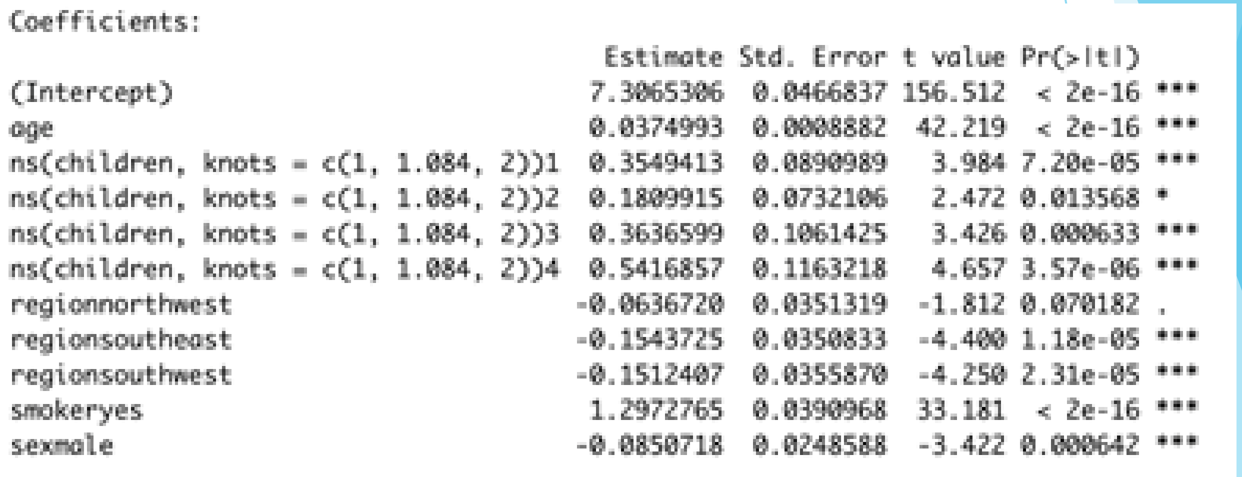
Multiple R-squared: 0.7134, Adjusted R-squared: 0.7103

GAMs EXTENSIONS

- GAMs extension to multiple linear regression

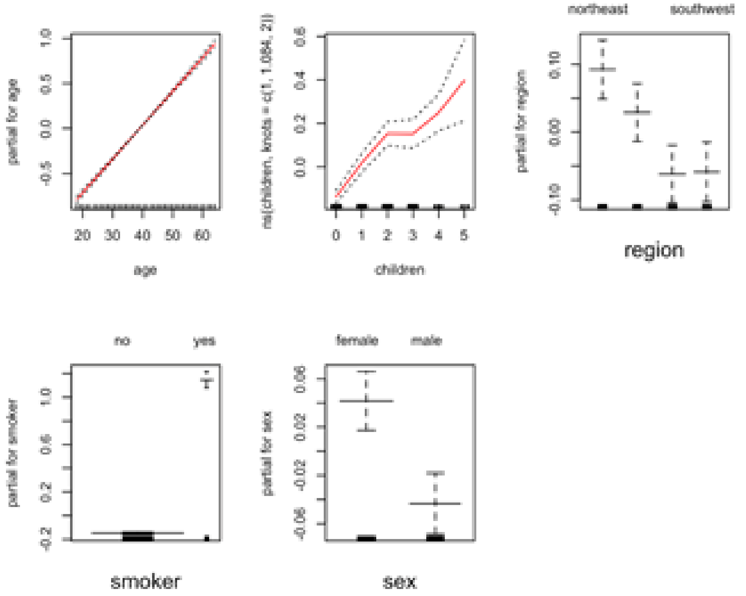


```
# Try to keep age normal (without knots)
gam3 = lm(logcharges~age+ns(children, knots=c(1, 1.084, 2))+
            +region+smoker+sex, data=insurance_clean)
summary(gam3)

par(mfrow=c(2,3))
plot.Gam(gam3, se=TRUE, col="red")
```



```
Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         7.3065306  0.0466837 156.512  < 2e-16 ***
age                                 0.0374993  0.0008882  42.219  < 2e-16 ***
ns(children, knots = c(1, 1.084, 2))1  0.3549413  0.0890989   3.984 7.20e-05 ***
ns(children, knots = c(1, 1.084, 2))2  0.1809915  0.0732106   2.472 0.013568 *  
ns(children, knots = c(1, 1.084, 2))3  0.3636599  0.1061425   3.426 0.000633 ***
ns(children, knots = c(1, 1.084, 2))4  0.5416857  0.1163218   4.657 3.57e-06 ***
regionnorthwest                    -0.0636720  0.0351319  -1.812 0.070182 .  
regionsoutheast                    -0.1543725  0.0350833  -4.400 1.18e-05 ***
regionsouthwest                    -0.1512407  0.0355870  -4.250 2.31e-05 ***
smokeryes                           1.2972765  0.0390968  33.181  < 2e-16 ***
sexmale                            -0.0850718  0.0248588  -3.422 0.000642 ***
```

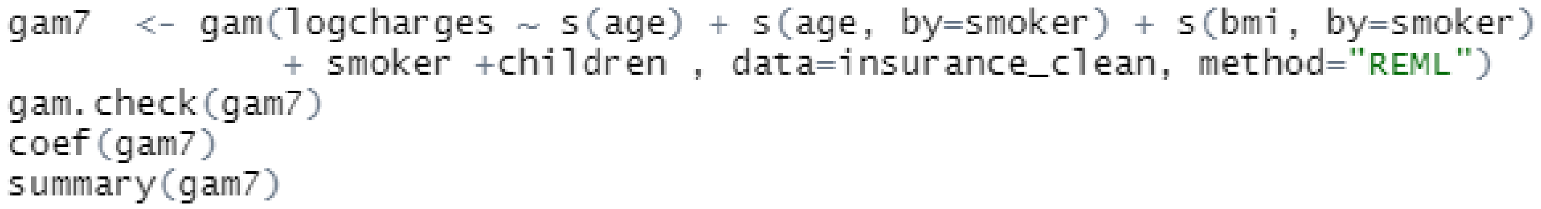
All the coefficients are still significant.



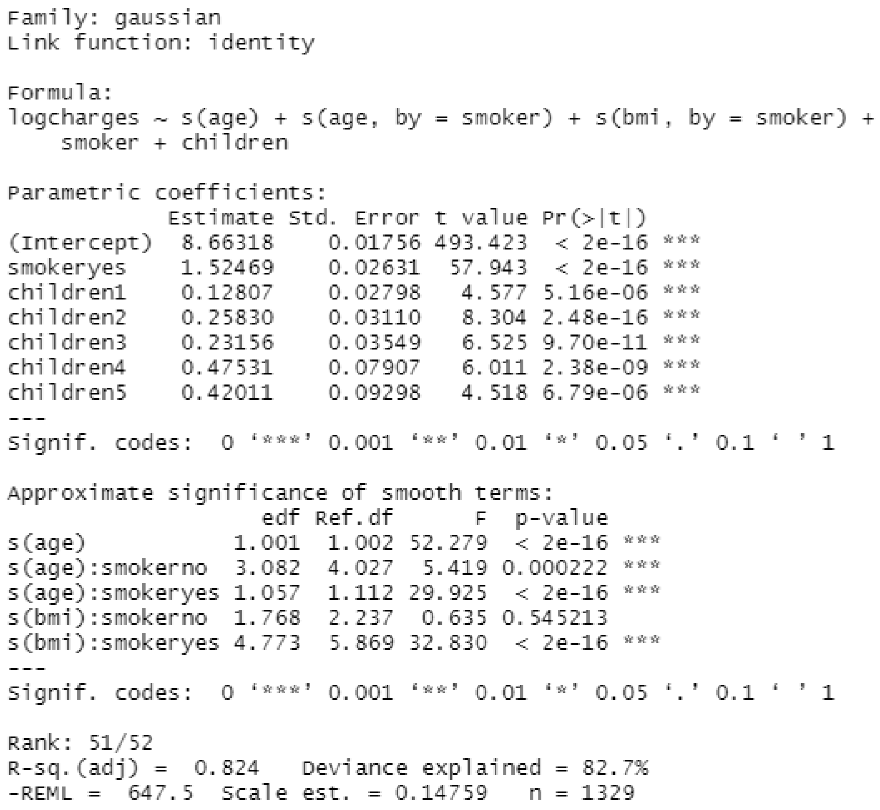
Multiple R-squared: 0.7109, Adjusted R-squared: 0.7085

GAMs EXTENSIONS

- GAM with smoothing splines: parameter determination using restricted marginal likelihood (REML)



```
gam7  <- gam(logcharges ~ s(age) + s(age, by=smoker) + s(bmi, by=smoker)
               + smoker +children , data=insurance_clean, method="REML")
gam.check(gam7)
coef(gam7)
summary(gam7)
```



```
Family: gaussian
Link function: identity

Formula:
logcharges ~ s(age) + s(age, by = smoker) + s(bmi, by = smoker) +
    smoker + children

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.66318    0.01756 493.423  < 2e-16 ***
smokeryes    1.52469    0.02631  57.943  < 2e-16 ***
children1    0.12807    0.02798   4.577 5.16e-06 ***
children2    0.25830    0.03110   8.304 2.48e-16 ***
children3    0.23156    0.03549   6.525 9.70e-11 ***
children4    0.47531    0.07907   6.011 2.38e-09 ***
children5    0.42011    0.09298   4.518 6.79e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                   edf Ref.df      F  p-value
s(age)           1.001  1.002 52.279  < 2e-16 ***
s(age):smokerno  3.082  4.027  5.419 0.000222 ***
s(age):smokeryes 1.057  1.112 29.925  < 2e-16 ***
s(bmi):smokerno  1.768  2.237  0.635 0.545213
s(bmi):smokeryes 4.773  5.869 32.830  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Rank: 51/52
R-sq.(adj) =  0.824   Deviance explained = 82.7%
-REML =  647.5  Scale est. = 0.14759   n = 1329
```

R-sq.(adj) = 0.824

Deviance explained = 82.7%

- CONCLUSION with the GAMs extension :

We can see that the R-squared of our GAMs models are not better than the multiple linear regression model we had at the beginning.
The best GAM model is the first one with a R-squared equal to 71,8%.