Estimating Medical Costs using Multiple Regression Models

2023-04-01

Prepared by:

Daksh Balkrushna Patel, 30190603

Monica Chandramurthy, 30191289

Saketh Ram Mamillapalli, 30187315

```
#library(XQuartz)
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(mctest)
library(agricolae)
library(binom)
library(dbplyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:dbplyr':
##
##     ident, sql
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(EnvStats)
```

```
##
## Attaching package: 'EnvStats'
```

```
## The following objects are masked from 'package:agricolae':
##
##     kurtosis, skewness
```

```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
## The following object is masked from 'package:base':
##
##     print.default
```

```
library(ggformula)
```

```
## Loading required package: ggstance
```

```
##
## Attaching package: 'ggstance'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh
```

```
## Loading required package: scales
```

```
##
## Attaching package: 'scales'
```

```
## The following objects are masked from 'package:psych':
##
##     alpha, rescale
```

```
## Loading required package: ggridges
```

```
##
## New to ggformula?  Try the tutorials:
##  learnr::run_tutorial("introduction", package = "ggformula")
##  learnr::run_tutorial("refining", package = "ggformula")
```

```
library(ggplot2)
library(htmltools)
library(ISLR)
library(knitr)
library(markdown)
library(mosaic)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected
by this.
```

```
##
## Attaching package: 'mosaic'
```

```
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following object is masked from 'package:scales':
##
##     rescale
```

```
## The following object is masked from 'package:EnvStats':
##
##     iqr
```

```
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
```

```
## The following objects are masked from 'package:psych':
##
##     logit, rescale
```

```
## The following object is masked from 'package:caret':
##
##     dotPlot
```

```
## The following object is masked from 'package:ggplot2':
##
##     stat
```

```
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
library(mdsr)
library(mosaicData)
library(nycflights13)
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following object is masked from 'package:mosaic':
##
##     count
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
library(purrr)
```

```
##
## Attaching package: 'purrr'
```

```
## The following object is masked from 'package:plyr':
##
##     compact
```

```
## The following object is masked from 'package:mosaic':
##
##     cross
```

```
## The following object is masked from 'package:scales':
##
##     discard
```

```
## The following object is masked from 'package:caret':
##
##     lift
```

```
library(rmarkdown)
library(rvest)
library(shiny)
library(stringi)
library(tibble)
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following objects are masked from 'package:Matrix':
##
##     expand, pack, unpack
```

```
library(tidyselect)
library(tinytex)
library(yaml)
library(shiny)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## The following objects are masked from 'package:mosaic':
##
##     deltaMethod, logit
```

```
## The following object is masked from 'package:EnvStats':
##
##     qqPlot
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
library(readr)
```

```
##
## Attaching package: 'readr'
```

```
## The following object is masked from 'package:rvest':
##
##     guess_encoding
```

```
## The following object is masked from 'package:scales':
##
##     col_factor
```

INTRODUCTION:

Medical expenses are one of the significant recurring expenses in human life. The rising cost of healthcare is an essential concern for many individuals and families. It's common knowledge that one lifestyle and various physical parameters dictate diseases or ailments one can have, and these ailments dictate medical expenses. According to multiple studies, significant factors contributing to higher personal medical care expenses include smoking, aging, and BMI. In this study, we aim to find a correlation between personal medical expenses and different factors and compare them. Then we use the prominent attributes as predictors to predict medical costs by creating linear regression models and comparing them using ANOVA. Our findings will provide insights into the factors driving medical costs and inform strategies for managing healthcare expenses.

DATASET

This dataset is in the public domain (available on https://github.com/stedy/Machine-Learning-with-R-datasets (https://github.com/stedy/Machine-Learning-with-R-datasets) or https://www.kaggle.com/mirichoi0218/insurance (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. Treatment costs depend 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 understand 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.

Variables description:

1. AGE: age of the primary beneficiary; Quantitative Data

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

```
medical_cost= read_csv("project/medical_cost.csv")
```

```
## Rows: 1338 Columns: 7
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (3): sex, smoker, region
## dbl (4): age, bmi, children, charges
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
str(medical_cost)
```

```
## spc_tbl_ [1,338 × 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ age     : num [1:1338] 19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr [1:1338] "female" "male" "male" "male" ...
##  $ bmi     : num [1:1338] 27.9 33.8 33 22.7 28.9 ...
##  $ children: num [1:1338] 0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr [1:1338] "yes" "no" "no" "no" ...
##  $ region  : chr [1:1338] "southwest" "southeast" "southeast" "northwest" ...
##  $ charges : num [1:1338] 16885 1726 4449 21984 3867 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   age = col_double(),
##   ..   sex = col_character(),
##   ..   bmi = col_double(),
##   ..   children = col_double(),
##   ..   smoker = col_character(),
##   ..   region = col_character(),
##   ..   charges = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
head(medical_cost, 4)
```

```
## # A tibble: 4 × 7
##     age sex      bmi children smoker region    charges
##   <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
## 1    19 female  27.9        0 yes    southwest  16885.
## 2    18 male    33.8        1 no     southeast   1726.
## 3    28 male    33          3 no     southeast   4449.
## 4    33 male    22.7        0 no     northwest  21984.
```

DATA CLEANING:

The Dataset is clean and does not have any missing/null values. As seen below:

```
#Checking for null/missing values
sum(is.na(medical_cost))
```

```
## [1] 0
```

DESCRIPTIVE DATA ANALYSIS:

Our response variable is ‘charges’, and our independent variables are ‘age’, ‘bmi’, ‘sex’, ‘children’, ‘smoker’ and ‘region’.

We will conduct some descriptive analysis, which includes determining the mean, median, and standard deviation of charges, and presenting the trends in a line graph. Additionally, we will examine the skewness and kurtosis of the distribution of charges.

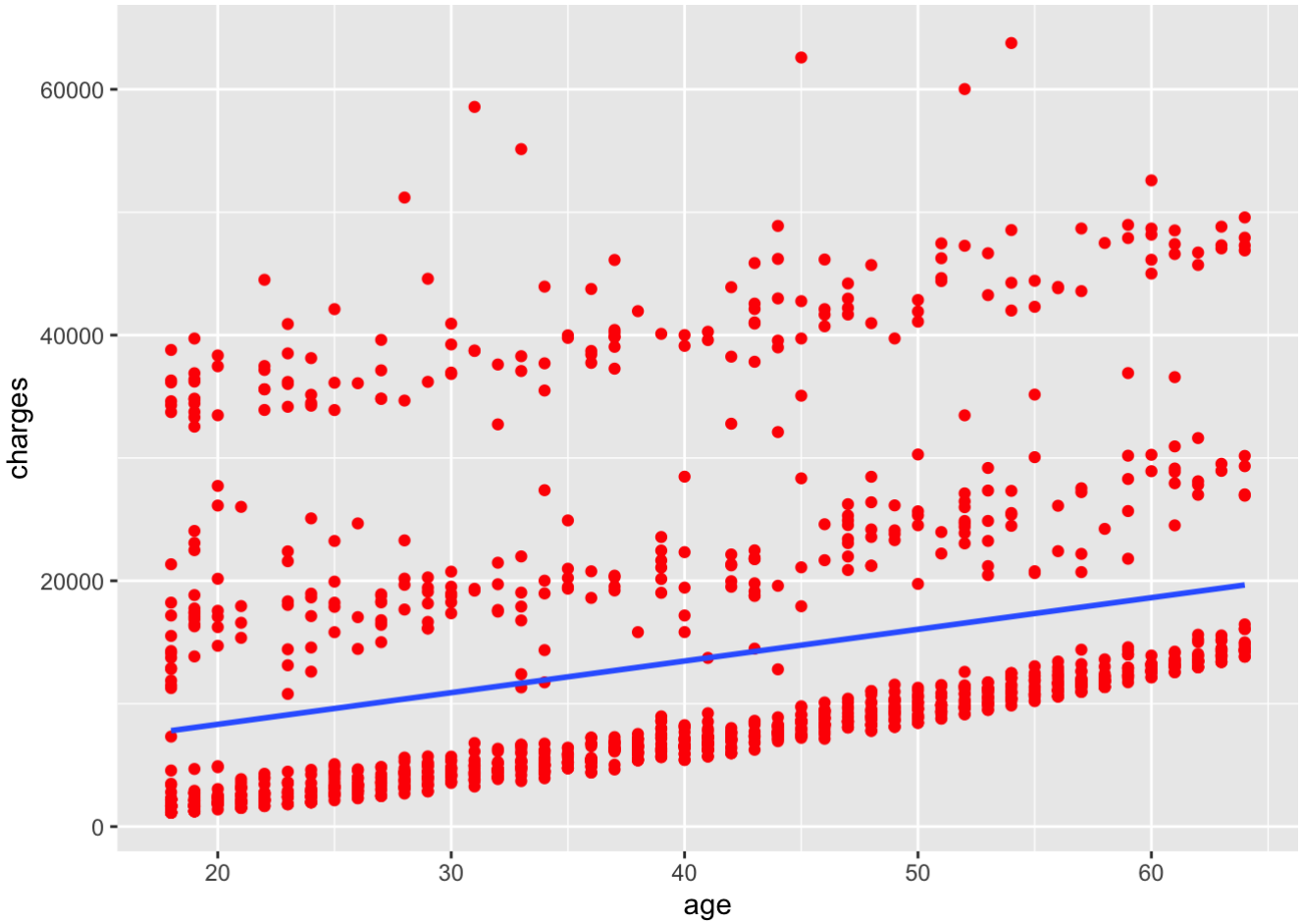
```
data_summary <- medical_cost %>%
summarise(mean_charges = mean(charges),
median_charges = median(charges),
sd_charges = sd(charges))
print(data_summary)
```

```
##   mean_charges median_charges sd_charges
## 1     13270.42       9382.033   12110.01
```

EXPLORATORY DATA ANALYSIS

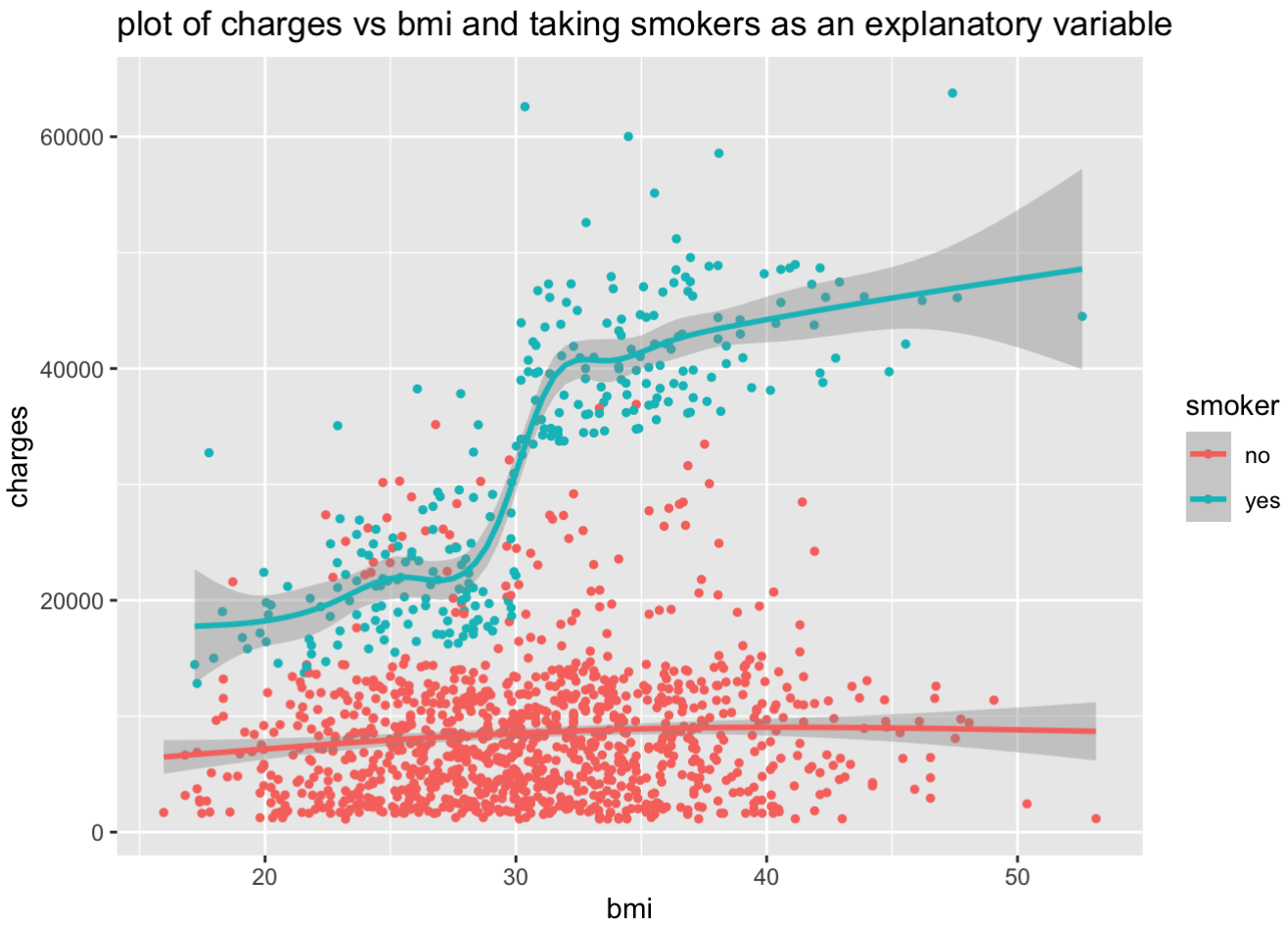
```
# Graph to check relationshp between age and charges
ggplot(data=medical_cost,mapping= aes(x=age,y=charges))+geom_point(color='red')+ geom_sm
ooth(method = "lm", se = FALSE)
```

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




```
# Graph to check relationship between bmi and charges by smoker
ggplot(data=medical_cost,aes(x=bmi,y=charges,colour=smoker))+geom_point(size=1)+geom_smo
oth()+labs(title = "plot of charges vs bmi and taking smokers as an explanatory variabl
e",x="bmi",y="charges")
```

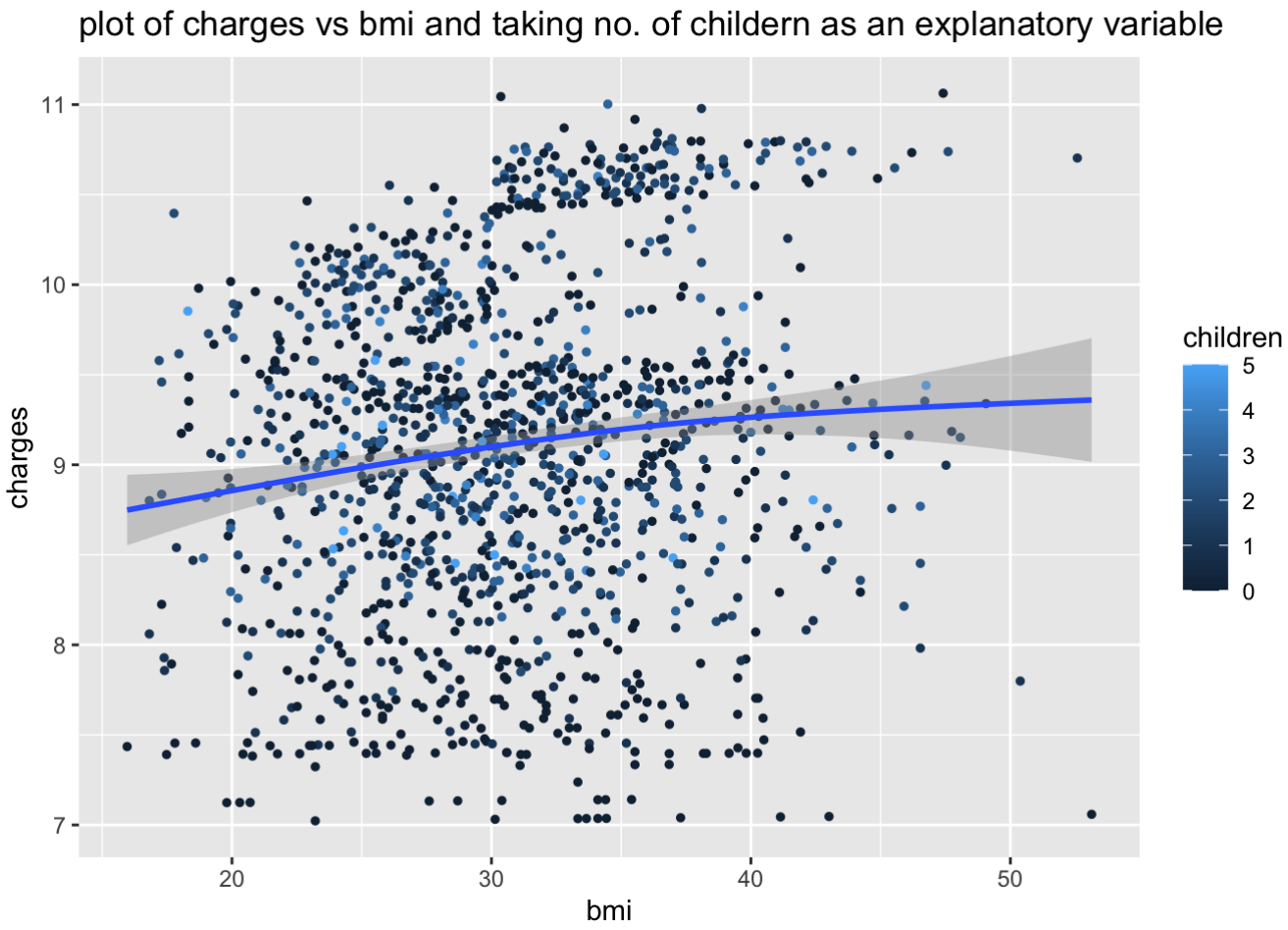
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



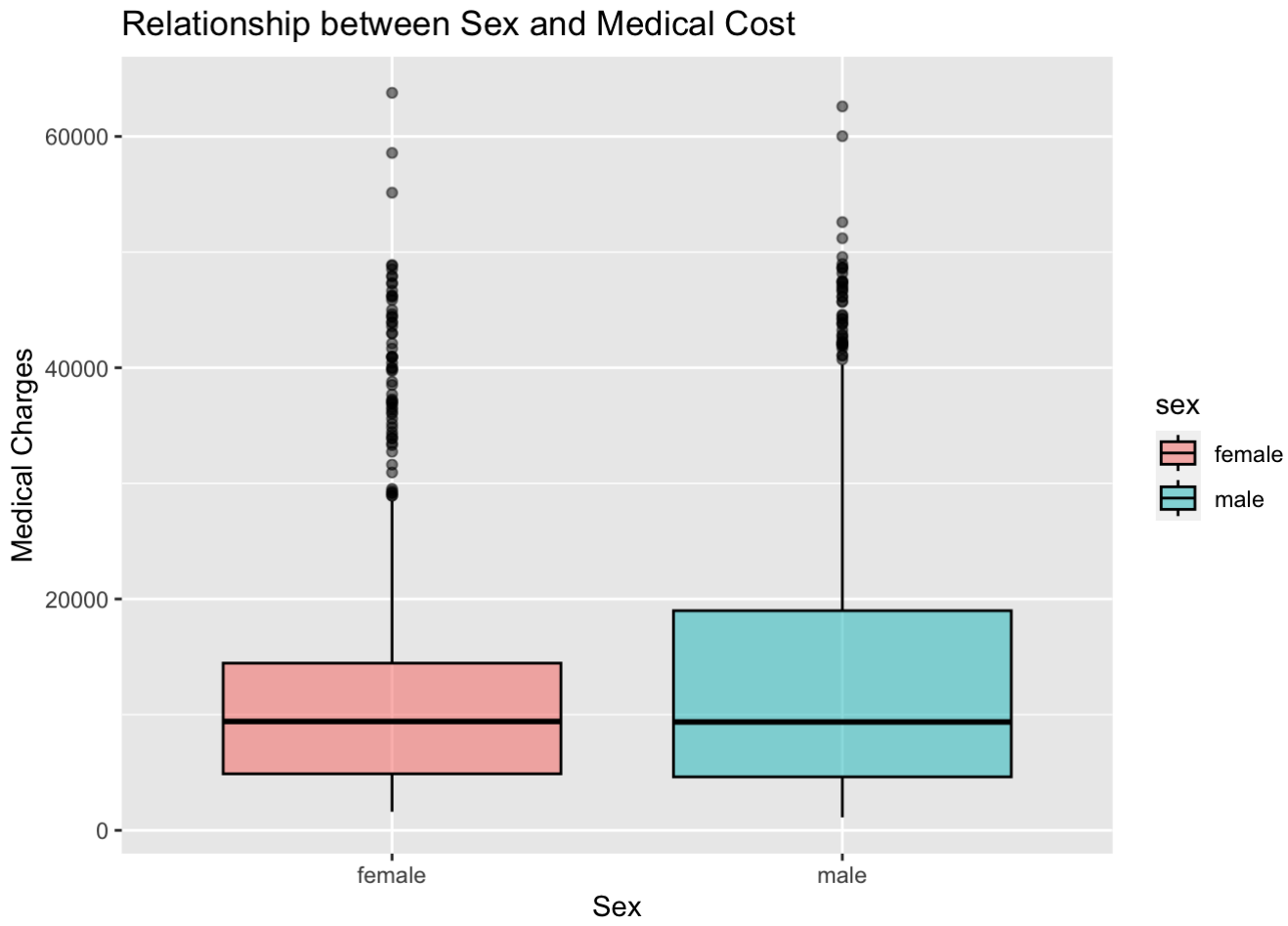

```
# Graph to check relationshp between children and charges by children
ggplot(data=medical_cost,aes(x=bmi,y=log(charges),colour=children))+geom_point(size=1)+g
eom_smooth()+labs(title = "plot of charges vs bmi and taking no. of childern as an expla
natory variable",x="bmi",y="charges")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

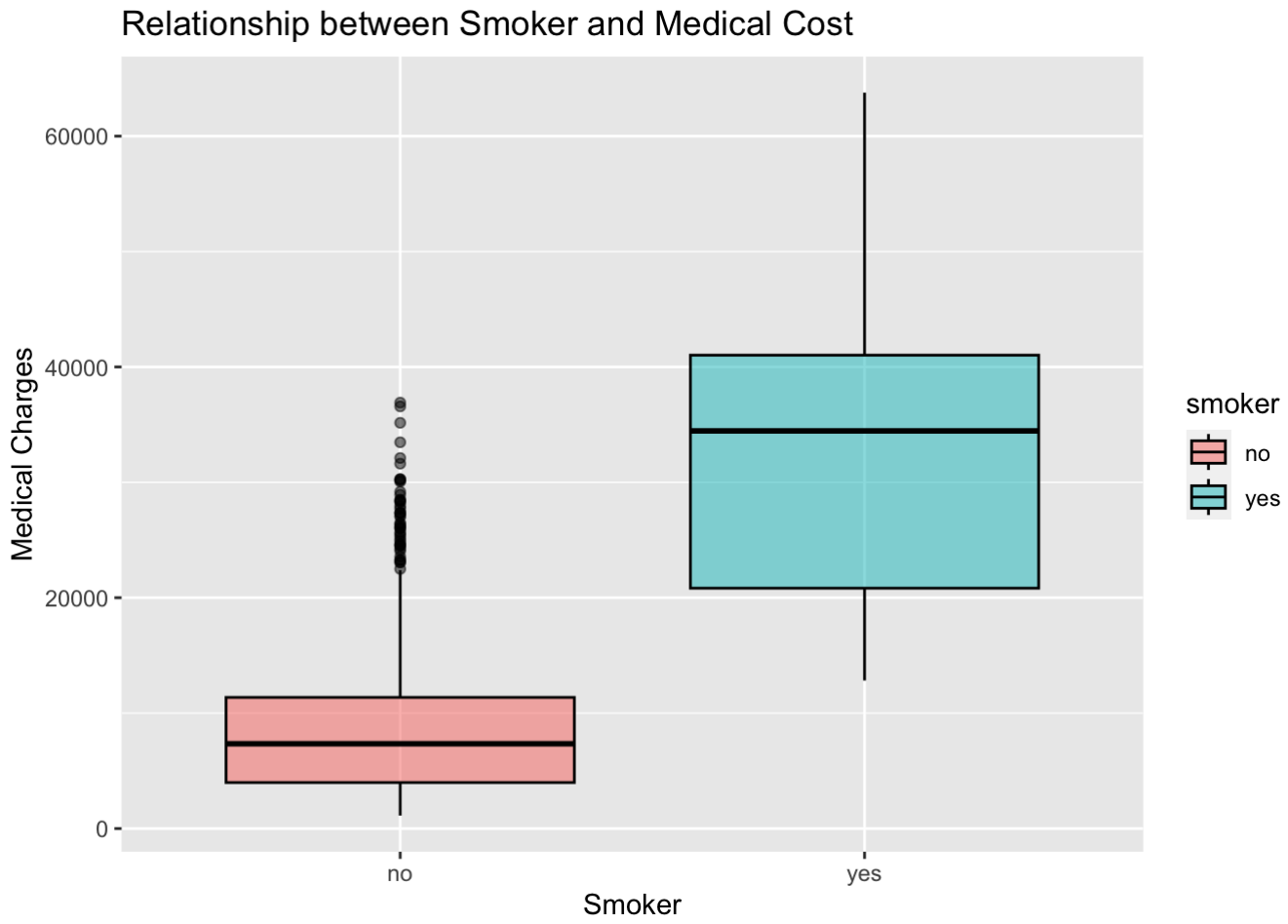
```
## Warning: The following aesthetics were dropped during statistical transformation: col
our
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



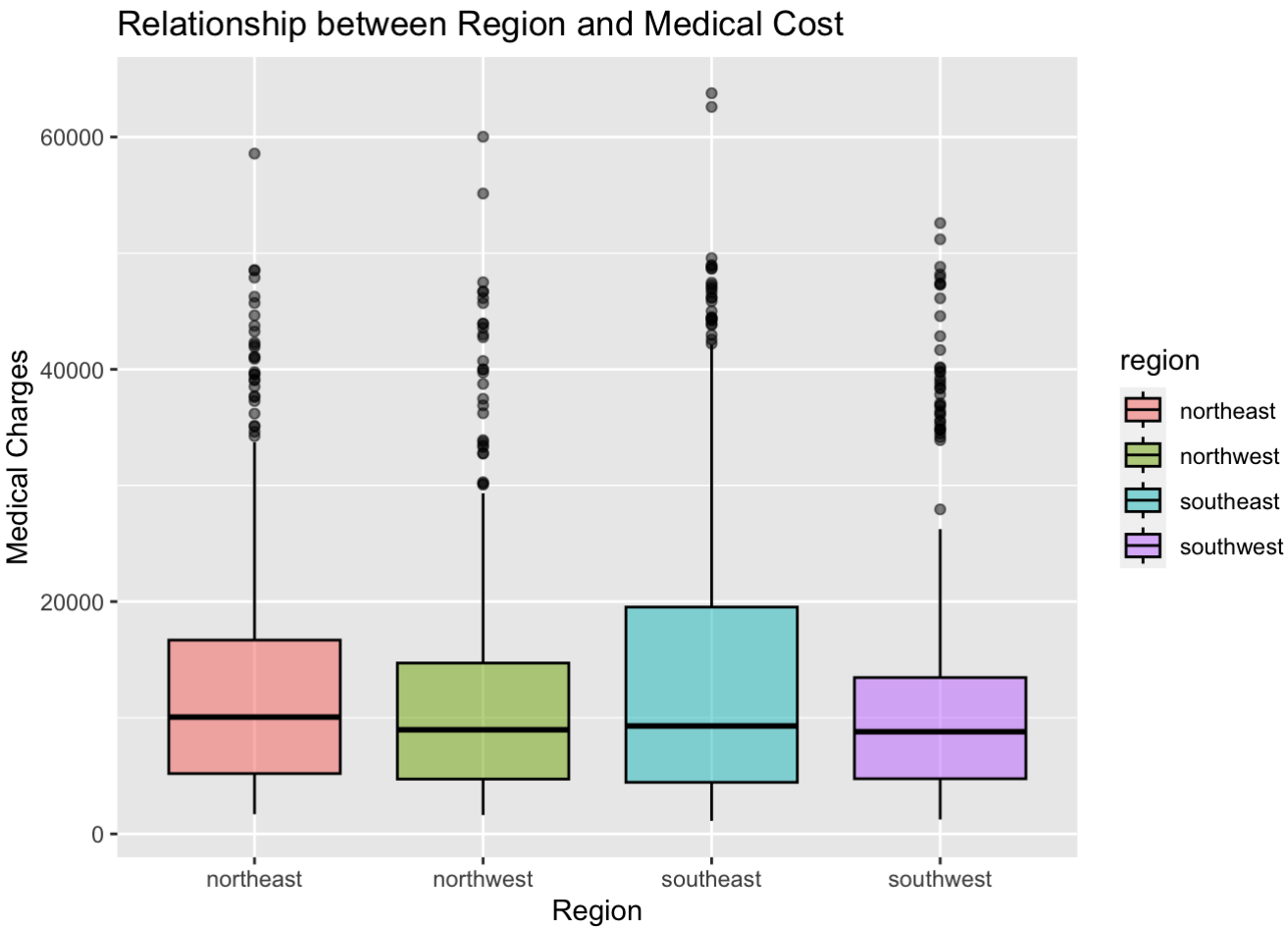

```
# Graph to check distribution of charges among male and female
ggplot(medical_cost, aes(x = sex, y = charges)) +
   geom_boxplot(aes(fill=sex), color = "black", alpha = 0.5) +
   xlab("Sex") +
   ylab("Medical Charges") +
   ggtitle("Relationship between Sex and Medical Cost")
```



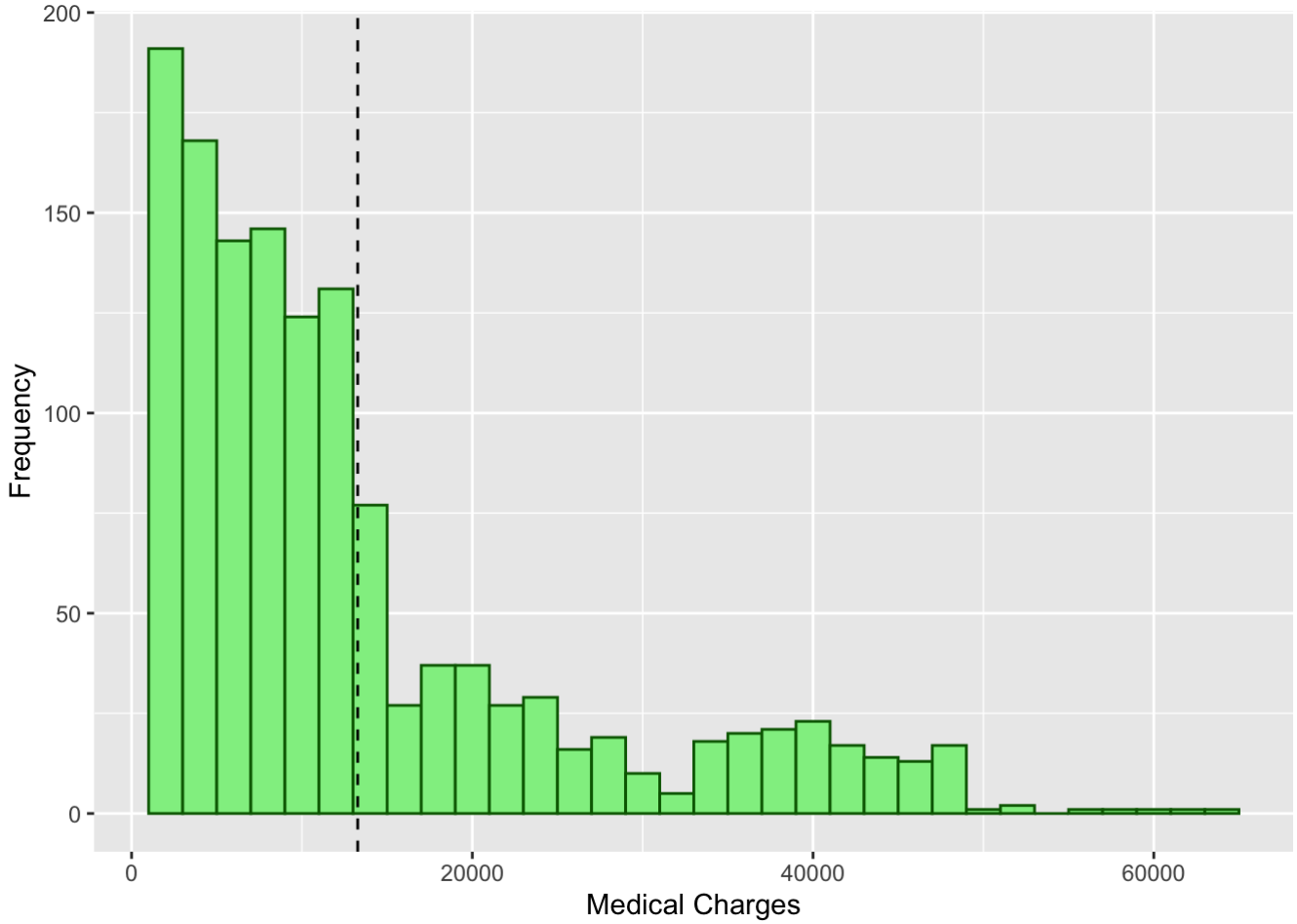

```
# Graph to check distribution of charges among smokers and non-smokers
ggplot(medical_cost, aes(x = smoker, y = charges)) +
   geom_boxplot(aes(fill=smoker), color = "black", alpha = 0.5) +
   xlab("Smoker") +
   ylab("Medical Charges") +
   ggtitle("Relationship between Smoker and Medical Cost")
```




```
# Graph to check distribution of charges across all regions
ggplot(medical_cost, aes(x = region, y = charges)) +
   geom_boxplot(aes(fill=region), color = "black", alpha = 0.5) +
   xlab("Region") +
   ylab("Medical Charges") +
   ggtitle("Relationship between Region and Medical Cost")
```




```
mean_charges = mean(medical_cost$charges)
#A histogram of the distribution of Happiness_Score
ggplot(medical_cost, aes(x = charges)) +
geom_histogram(color = "darkgreen", fill = "lightgreen", binwidth = 2000, ) +
xlab("Medical Charges") +
ylab("Frequency") +
geom_vline(xintercept = mean_charges, linetype = "dashed", color = "black")
```



We conducted several graphical analyses to explore the relationship between medical charges and different predictors in our dataset.

First, we created a scatter plot to examine the relationship between age and medical charges. The plot showed a positive linear relationship between age and charges, indicating that as age increases, medical charges also tend to increase.

Next, we examined the relationship between body mass index (BMI) and medical charges by smokers. The plot reveals a positive linear relationship between BMI and medical charges for both smokers and non-smokers. However, the charges for smokers are significantly higher than non-smokers across all BMI values. This observation highlights the importance of considering smoking habits when estimating medical costs.

To further explore the relationship between BMI and medical charges, we created a scatter plot that included the number of children as an explanatory variable. However, the plot did not show any significant pattern or relationship between BMI and medical charges with the number of children as an explanatory variable.

We also investigated the distribution of medical charges among males and females using a box plot. The plot showed that the median charges were slightly higher for males than females, but the difference was not significant.

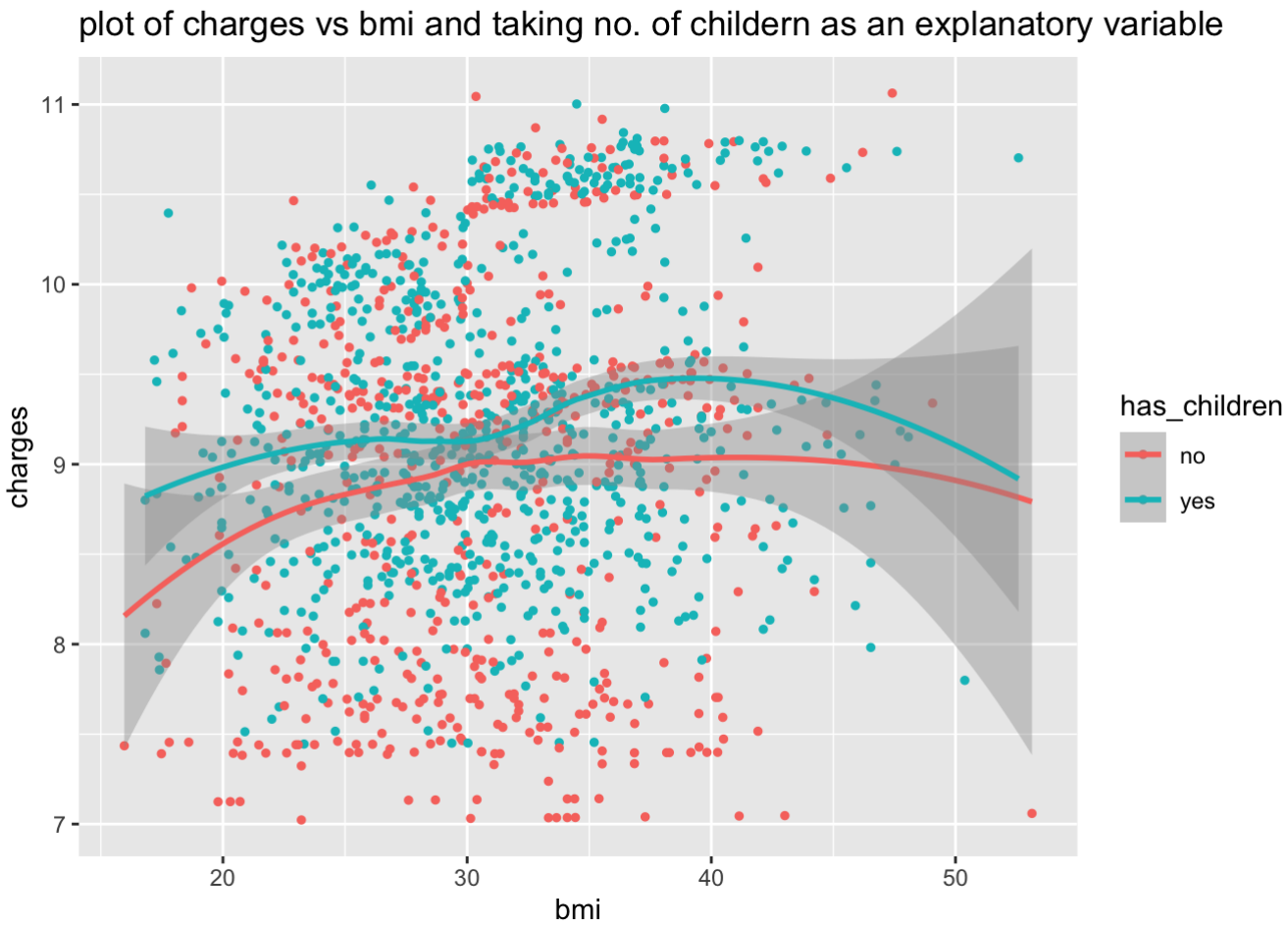
We then examined the distribution of medical charges between smokers and non-smokers using another box plot. The plot revealed a significant difference in charges between the two groups, with smokers having significantly higher charges than non-smokers.

Finally, we created a box plot to visualize the distribution of medical charges across all regions. The plot indicated that there were some differences in charges between different regions, with the southeast region having the highest median charges.

Overall, these graphical analyses provided valuable insights into the relationships between medical charges and various predictors in our dataset.

```
medical_cost <- medical_cost %>%
  mutate(has_children = ifelse(children > 0, "yes", "no"))%>%
  dplyr::select(-children)

ggplot(data=medical_cost,aes(x=bmi,y=log(charges),colour=has_children))+geom_point(size=
1)+geom_smooth()+labs(title = "plot of charges vs bmi and taking no. of childern as an e
xplanatory variable",x="bmi",y="charges")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



In an attempt to explore the relationship between BMI and medical charges with the number of children as an explanatory variable, we created a new column “has_children” in our dataset, which categorizes individuals as having children or not based on whether their “children” column value is greater than zero or not. However, when we plotted the relationship between BMI and medical charges with “has_children” as an explanatory variable, we did not observe any significant pattern or relationship between these variables. Therefore, we concluded that the number of children does not seem to have a significant impact on the relationship between BMI and medical charges.

This transformation can simplify the interpretation of the regression models by allowing us to compare the medical charges of individuals with and without children more directly.

MODELING PLAN

For this project, we will utilize the techniques acquired in Data 603. Our approach involves first conducting a linear regression analysis using all of the predictors, and performing individual t-tests on each variable at a 5% significance level. Variables that are not statistically significant will be removed. A partial F-test will then be performed to compare the full and reduced models. Once we are satisfied with the main effects, we will employ individual t-tests to examine significant higher-order terms and interactions. A subsequent F-test will be used to evaluate whether the higher-order terms and interactions are significant. If so, they will be added to the main effects to form our final model. We will then verify our model's adherence to the six assumptions listed below:

1. Linearity Assumption - Review residual plots
2. Independence Assumption
3. Normality Assumption - Using the Shapiro-Wilk normality test
4. Equal Variance Assumption (heteroscedasticity) - Using the Breusch-Pagan test
5. Multicollinearity - Using variance inflation factors (VIF)
6. Outliers - Check Cook's distance and leverage

If our model fails to satisfy any of these assumptions, we will review our methodology to see if we can make any improvements/transformations. Once our model satisfies most of the assumptions, we will use it to predict medical charges for sample data.

GUIDING QUESTIONS

Identifying the key factors that contribute to medical costs and exploring their relationships with each other.

Conducting Individual Co-efficient (t-test)

- Null Hypothesis, H0: βi = 0, i = age, factor(sex), bmi, factor(has_children), factor(smoker), factor(region)
- Alternate Hypothesis, Ha: βi != 0, i = age, factor(sex), bmi, factor(has_children), factor(smoker), factor(region)

```
medical_cost_full_model =lm(charges~age+factor(sex)+bmi+factor(has_children)+factor(smok
er)+ factor(region),data=medical_cost)
summary(medical_cost_full_model)
```

```
##
## Call:
## lm(formula = charges ~ age + factor(sex) + bmi + factor(has_children) +
##     factor(smoker) + factor(region), data = medical_cost)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11483.5  -2894.7   -956.5   1478.1  30059.6
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -12001.01     993.99 -12.074  < 2e-16 ***
## age                        256.91      11.92  21.562  < 2e-16 ***
## factor(sex)male           -126.41     333.31  -0.379  0.70456
## bmi                        339.51      28.63  11.858  < 2e-16 ***
## factor(has_children)yes    999.58     335.88   2.976  0.00297 **
## factor(smoker)yes        23849.65     413.63  57.660  < 2e-16 ***
## factor(region)northwest   -352.22     476.88  -0.739  0.46029
## factor(region)southeast  -1057.33     479.27  -2.206  0.02755 *
## factor(region)southwest   -944.26     478.40  -1.974  0.04861 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6069 on 1329 degrees of freedom
## Multiple R-squared:  0.7503, Adjusted R-squared:  0.7488
## F-statistic: 499.3 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Full Model: *Y(charges) = β0 + β1(age) + β2(factor(sex)) + β3(bmi) + β4(factor(has_children)) + β6(factor(smoker)) +β7(factor(region)) + ϵ*

From the above summary, the output shows that the factor(sex) has tcal= -0.379 with the p-value= 0.70456> 0.05,indicating that we should clearly not to reject the null hypothesis that the sex does not significantly influence on medical charges at α = 0.05.

Conducting Partial F-Test

- *Null Hypothesis, H0 : β2 = 0 in the model Y(charges) = β0 + β1(age) + β2(factor(sex)) + β3(bmi) + β4(factor(has_children)) + β6(factor(smoker)) +β7(factor(region)) + ϵ*
- *Alternate Hypothesis, Ha : β2 != 0 in the model Y(charges) = β0 + β1(age) + β2(factor(sex)) + β3(bmi) + β4(factor(has_children)) + β6(factor(smoker)) +β7(factor(region)) + ϵ*

```
#Checking our Reduced Model - First Order Model after removing sex
medical_firstordermodel = lm(charges~age+bmi+factor(has_children)+factor(smoker)+ factor
(region),data=medical_cost)
summary(medical_firstordermodel)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + factor(has_children) + factor(smoker) +
##     factor(region), data = medical_cost)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11541.7  -2874.9   -991.8   1516.5  30004.4
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -12050.69     985.01 -12.234  < 2e-16 ***
## age                         257.02      11.91  21.585  < 2e-16 ***
## bmi                         339.00      28.59  11.857  < 2e-16 ***
## factor(has_children)yes     997.67     335.73   2.972  0.00302 **
## factor(smoker)yes         23837.87     412.33  57.813  < 2e-16 ***
## factor(region)northwest    -351.46     476.72  -0.737  0.46110
## factor(region)southeast   -1056.65     479.12  -2.205  0.02760 *
## factor(region)southwest    -943.64     478.24  -1.973  0.04869 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6067 on 1330 degrees of freedom
## Multiple R-squared:  0.7503, Adjusted R-squared:  0.749
## F-statistic:   571 on 7 and 1330 DF,  p-value: < 2.2e-16
```

```
#Partial F-Test
anova(medical_cost_full_model, medical_firstordermodel)
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ age + factor(sex) + bmi + factor(has_children) + factor(smoker) +
##     factor(region)
## Model 2: charges ~ age + bmi + factor(has_children) + factor(smoker) +
##     factor(region)
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1   1329 4.8951e+10
## 2   1330 4.8956e+10 -1  -5297821 0.1438 0.7046
```

From the above data, after dropping the variable sex off the full model, the reduced output shows that Fcal = 0.1438, with df=1 and 1330 DF (p-value=0.7046 > α = 0.05), indicating that we should clearly not to reject the null hypothesis which means that we definitely drop the variable sex off the model.

At this point, from the initial estimated regression model is

Y(charges) = -12001.01 + 256.91(age) - 126.41(factor(sex)) + 339.51(bmi) + 999.58(factor(has_children=yes)) + 23849.65(factor(smoker=yes)) -352.22(factor(region=northwest)) -1057.33(factor(region=southeast)) -944.26(factor(region=southwest)) + ϵ

After checking individual coefficients test, the final regression model is

Y(charges) = β0 + β1(age) + β3(bmi) + β4(factor(has_children)) + β6(factor(smoker)) +β7(factor(region)) + ϵ

Y(charges) = -12050.69 + 257.02(age) + 339.00(bmi) + 997.67(factor(has_children=yes)) + 23837.87(factor(smoker=yes)) -351.46(factor(region=northwest)) -1032.43(factor(region=southeast)) -943.64(factor(region=southwest)) + ϵ

The model has an Adjusted R-squared value of 0.749, indicating that 74.9% of the variation in medical charges can be explained by the predictors included in the model. The F-statistic is 571 with a p-value < 2.2e-16, indicating that the model as a whole is statistically significant. Overall, the results suggest that age, BMI, number of children, smoker status, and region are important predictors of medical charges, and the model has a good fit.

```
#Performance metrics
l_pred <- predict(medical_cost_full_model, medical_cost)
radj <- summary(medical_cost_full_model)$adj.r.squared
rse <- sqrt(sum(residuals(medical_cost_full_model)^2) / medical_cost_full_model$df.resid
ual )
rmse <- RMSE(l_pred, medical_cost$charges)
aic <- AIC(medical_cost_full_model)
l_reg <- cbind("Adjusted R sq"=radj, "RSE"=rse, "RMSE"=rmse, "AIC"=aic)
cat("\n Performace metrics for the full model: \n")
```

```
##
##  Performace metrics for the full model:
```

```
l_reg
```

```
##      Adjusted R sq      RSE     RMSE      AIC
## [1,]     0.7488424 6069.008 6048.562 27118.55
```

```
#Performance metrics
l2_pred <- predict(medical_firstordermodel, medical_cost)
radj2 <- summary(medical_firstordermodel)$adj.r.squared
rse2 <- sqrt(sum(residuals(medical_firstordermodel)^2) / medical_firstordermodel$df.resi
dual )
rmse2 <- RMSE(l2_pred, medical_cost$charges)
aic2 <- AIC(medical_firstordermodel)
cp2 <- ols_mallows_cp(medical_cost_full_model, medical_firstordermodel)
l2_reg <- cbind("Adjusted R sq"=radj2, "RSE"=rse2, "RMSE"=rmse2, "AIC"=aic2, "CP-Criteri
on"=cp2)
cat("\n Performace metrics for the reduced model: \n")
```

```
##
##  Performace metrics for the reduced model:
```

```
l2_reg
```

```
##      Adjusted R sq      RSE     RMSE     AIC CP-Criterion
## [1,]     0.7490041 6067.054 6048.889 27116.7     5.856073
```

We considered five model performance metrics: adjusted R-squared, residual standard error (RSE), root mean squared error (RMSE), Akaike Information Criterion (AIC), and the Cp-criterion.

The performance metrics for the full model are an adjusted R-squared of 0.7488, a residual standard error (RSE) of 6069.008, a root-mean-square error (RMSE) of 6048.562, an Akaike information criterion (AIC) of 27118.55.

The performance metrics for the reduced model are an adjusted R-squared of 0.7490, an RSE of 6067.054, an RMSE of 6048.889, an AIC of 27116.7, and a Cp-criterion value of 5.8560.

Both models have similar performance metrics, but the reduced model has a slightly better adjusted R-squared and AIC value, suggesting that it is a slightly better fit for the data. However, it is important to consider the specific research question and the variables included in each model when choosing which model to use.

Checking for Interaction Model with Quantitative Predictors

- Null Hypothesis, H0: $\beta i = 0$, (i=1,2,…,p)
- Alternate Hypothesis, Ha: $\beta i \neq 0$, (i=1,2,…,p)

```
medical_interaction = lm(charges ~ (age+ bmi+ factor(has_children) +factor(smoker)+ fact
or(region))^2,data=medical_cost)
summary(medical_interaction)
```

```
##
## Call:
## lm(formula = charges ~ (age + bmi + factor(has_children) + factor(smoker) +
##     factor(region))^2, data = medical_cost)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11645.6  -2114.7  -1169.2   -104.9  29760.9
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                     -2408.615   2517.055  -0.957
## age                                               195.286     52.432   3.725
## bmi                                                53.385     82.485   0.647
## factor(has_children)yes                          1740.255   1584.465   1.098
## factor(smoker)yes                              -20683.849   1928.915 -10.723
## factor(region)northwest                         -1754.247   2261.815  -0.776
## factor(region)southeast                          3417.836   2171.190   1.574
## factor(region)southwest                          -105.028   2167.932  -0.048
## age:bmi                                             1.225      1.629   0.752
## age:factor(has_children)yes                        -1.843     19.445  -0.095
## age:factor(smoker)yes                              -1.647     23.988  -0.069
## age:factor(region)northwest                        23.549     27.413   0.859
## age:factor(region)southeast                        54.321     27.445   1.979
## age:factor(region)southwest                        49.382     27.950   1.767
## bmi:factor(has_children)yes                        -1.979     46.671  -0.042
## bmi:factor(smoker)yes                            1478.896     55.940  26.437
## bmi:factor(region)northwest                        -1.251     70.144  -0.018
## bmi:factor(region)southeast                      -192.000     60.696  -3.163
## bmi:factor(region)southwest                       -89.025     67.070  -1.327
## factor(has_children)yes:factor(smoker)yes        -1228.204    673.232  -1.824
## factor(has_children)yes:factor(region)northwest    427.415    771.144   0.554
## factor(has_children)yes:factor(region)southeast   -873.862    775.359  -1.127
## factor(has_children)yes:factor(region)southwest  -1148.593    772.207  -1.487
## factor(smoker)yes:factor(region)northwest            1.089    969.537   0.001
## factor(smoker)yes:factor(region)southeast        -1016.225    924.104  -1.100
## factor(smoker)yes:factor(region)southwest         1001.701    981.869   1.020
##                                                 Pr(>|t|)
## (Intercept)                                     0.338785
## age                                             0.000204 ***
## bmi                                             0.517604
## factor(has_children)yes                         0.272265
## factor(smoker)yes                                < 2e-16 ***
## factor(region)northwest                         0.438129
## factor(region)southeast                         0.115688
## factor(region)southwest                         0.961368
## age:bmi                                         0.452173
## age:factor(has_children)yes                     0.924491
## age:factor(smoker)yes                           0.945268
## age:factor(region)northwest                     0.390475
## age:factor(region)southeast                     0.047999 *
## age:factor(region)southwest                     0.077498 .
```

```
## bmi:factor(has_children)yes                       0.966176
## bmi:factor(smoker)yes                              < 2e-16 ***
## bmi:factor(region)northwest                       0.985779
## bmi:factor(region)southeast                       0.001596 **
## bmi:factor(region)southwest                       0.184630
## factor(has_children)yes:factor(smoker)yes         0.068328 .
## factor(has_children)yes:factor(region)northwest 0.579495
## factor(has_children)yes:factor(region)southeast 0.259931
## factor(has_children)yes:factor(region)southwest 0.137145
## factor(smoker)yes:factor(region)northwest         0.999104
## factor(smoker)yes:factor(region)southeast         0.271670
## factor(smoker)yes:factor(region)southwest         0.307823
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4840 on 1312 degrees of freedom
## Multiple R-squared:  0.8432, Adjusted R-squared:  0.8402
## F-statistic: 282.3 on 25 and 1312 DF,  p-value: < 2.2e-16
```

The interaction model shows the output shows that Fcal = 282.3, with df=25 and 1312 DF (p-value= 2.2e-16 < α = 0.05), indicating that there's clearly interaction between the terms. However from the above data, the interaction model, only the interaction between bmi:factor(smoker) is significant compared to the others with p-value < 0.05 (2e-16). Thus we will only consider only : bmi:factor(smoker)

The adjusted R-squared value is 0.8402, indicating that the model explains around 84.02% of the variance in medical charges. The F-statistic has a high value of 135.6, indicating that the model is statistically significant.

In conclusion, the multiple linear regression model with interaction terms provides insights into the impact of age, number of children, smoking status, and region on medical charges. The model is statistically significant and explains a considerable proportion of the variance in medical charges.

Conducting F-test for first order model and interaction model

- Null Hypothesis (H0): The reduced model without the interaction term is a better fit for the data.
- Alternative Hypothesis (HA): The full model with the interaction term is a better fit for the data.

```
medical_interaction_2 = lm(charges ~ age+ bmi+ factor(has_children) +factor(smoker)+ fac
tor(region) + bmi:factor(smoker) ,data=medical_cost)
summary(medical_interaction_2)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + factor(has_children) + factor(smoker) +
##     factor(region) + bmi:factor(smoker), data = medical_cost)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15044.7  -1984.1  -1282.1   -342.2  30345.1
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -2496.564    864.809  -2.887 0.003954 **
## age                         264.243      9.555  27.655  < 2e-16 ***
## bmi                          23.904     25.699   0.930 0.352452
## factor(has_children)yes     973.150    269.299   3.614 0.000313 ***
## factor(smoker)yes         -20191.130   1653.935 -12.208  < 2e-16 ***
## factor(region)northwest     -575.309    382.476  -1.504 0.132775
## factor(region)southeast    -1228.753    384.359  -3.197 0.001422 **
## factor(region)southwest    -1207.005    383.729  -3.145 0.001695 **
## bmi:factor(smoker)yes       1434.371     52.793  27.170  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4866 on 1329 degrees of freedom
## Multiple R-squared:  0.8395, Adjusted R-squared:  0.8385
## F-statistic: 868.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

```
anova(medical_firstordermodel, medical_interaction_2)
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ age + bmi + factor(has_children) + factor(smoker) +
##     factor(region)
## Model 2: charges ~ age + bmi + factor(has_children) + factor(smoker) +
##     factor(region) + bmi:factor(smoker)
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   1330 4.8956e+10
## 2   1329 3.1474e+10  1 1.7482e+10 738.18 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the output above, we can see that the F-statistic is 738.18 with 1 and 1329 degrees of freedom and a p-value less than 2.2e-16. This indicates that we can reject the null hypothesis and conclude that the full model with the interaction term is a better fit for the data than the reduced model without the interaction term. Therefore, we should include the interaction term in our final model.

The model has an adjusted R-squared value of 0.8385, indicating that it explains around 83.85% of the variance in medical charges. The F-statistic has a high value of 585.7, indicating that the model is statistically significant.

In conclusion, this model provides insights into the impact of age, number of children, smoking status, and the interaction between BMI and smoking status on medical charges. These predictors are significant in explaining the variance in medical charges. However, further analyses may be necessary to validate the assumptions of the model.

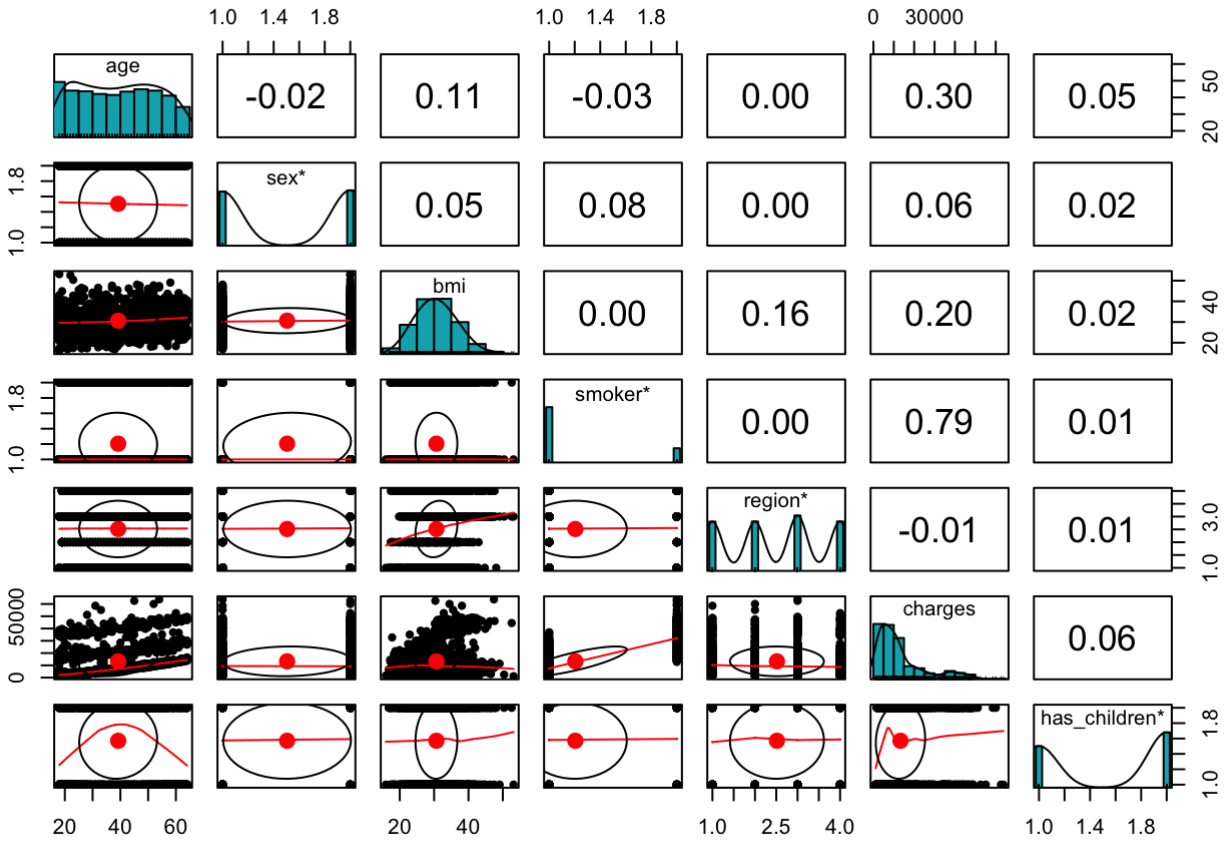
Our Regression model upto this point is:

Y(charges) = β0 + β1(age) + β2(bmi) + β3(factor(has_children)) + β4(factor(smoker)) +β5(factor(region)) + β6(bmi:factor(smoker))+ ϵ

Checking for non-linear pattern:
The “pairs.panels” function is used to create a matrix of scatterplots and correlation coefficients between all pairs of variables in the “medical_cost” dataset. The function uses Pearson’s correlation method to compute the correlation coefficients and displays histograms and density plots for each variable.

To visualize the relationships between variables and identify non-linear patterns, which may suggest the need for quadratic or higher-order terms.

```
#Checking for Quadratic terms:
pairs.panels(medical_cost,
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
             density = TRUE,  # show density plots
             ellipses = TRUE # show correlation ellipses
             )
```



The pairs plot indicates that the relationship between the dependent variable (charges) and the independent variable (bmi) may be non-linear. Therefore, a quadratic term (bmi^2) was added to the model.

- Null Hypothesis, H(0) : βp−q+1 = βp−q+2 = ... = βp = 0 : Higher order terms are not significant
- Alternate Hypothesis, H(A) : at least one βp 6 = 0 : At least one higher order term is significant

```
medical_quad_model_Reduced = lm(charges ~ age + bmi + I(bmi^2) + factor(has_children) +
factor(smoker)+ factor(region) + (bmi:factor(smoker)) ,data=medical_cost)
summary(medical_quad_model_Reduced)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + I(bmi^2) + factor(has_children) +
##     factor(smoker) + factor(region) + (bmi:factor(smoker)), data = medical_cost)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11647.6  -2076.6  -1247.3   -182.9  29964.1
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -10714.746   2621.763  -4.087 4.63e-05 ***
## age                         262.505      9.533  27.535  < 2e-16 ***
## bmi                         573.180    167.467   3.423 0.000639 ***
## I(bmi^2)                     -8.742      2.634  -3.319 0.000928 ***
## factor(has_children)yes     986.040    268.318   3.675 0.000247 ***
## factor(smoker)yes        -20275.787   1647.935 -12.304  < 2e-16 ***
## factor(region)northwest    -654.393    381.787  -1.714 0.086757 .
## factor(region)southeast   -1186.458    383.131  -3.097 0.001998 **
## factor(region)southwest   -1271.236    382.781  -3.321 0.000921 ***
## bmi:factor(smoker)yes      1437.700     52.605  27.330  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4848 on 1328 degrees of freedom
## Multiple R-squared:  0.8408, Adjusted R-squared:  0.8397
## F-statistic: 779.3 on 9 and 1328 DF,  p-value: < 2.2e-16
```

Since the p-value is less than the significant value 0.05 (2.2e-16). We reject our null hypothesis and accept our alternate hypothesis At least one higher order term is significant.

The resulting medical_quad_model shows that age, bmi, bmi^2, factor(has_children), factor(smoker), factor(region), and bmi:factor(smoker) are all significant predictors of medical charges. The adjusted R-squared value of 0.8397 suggests that the model explains a good proportion of the variance in the data.

Comparing Interaction model and Quadratic Model:

- Null hypothesis (H0): The interaction model and the quadratic model have the same performance in predicting medical charges.
- Alternative hypothesis (Ha): The interaction model performs better than the quadratic model in predicting medical charges.

```
medical_interact_model = lm(charges ~ age + bmi + factor(has_children) + factor(smoker)
+ factor(region) + (bmi:factor(smoker)), data = medical_cost)

medical_quad_model = lm(charges ~ age + bmi + I(bmi^2) + factor(has_children) + factor(s
moker) + factor(region) + (bmi:factor(smoker)), data = medical_cost)

anova(medical_quad_model, medical_interact_model)
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ age + bmi + I(bmi^2) + factor(has_children) + factor(smoker) +
##     factor(region) + (bmi:factor(smoker))
## Model 2: charges ~ age + bmi + factor(has_children) + factor(smoker) +
##     factor(region) + (bmi:factor(smoker))
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   1328 3.1215e+10
## 2   1329 3.1474e+10 -1 -258919034 11.015 0.0009281 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(medical_quad_model)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + I(bmi^2) + factor(has_children) +
##     factor(smoker) + factor(region) + (bmi:factor(smoker)), data = medical_cost)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11647.6  -2076.6  -1247.3   -182.9  29964.1
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -10714.746   2621.763  -4.087 4.63e-05 ***
## age                         262.505      9.533  27.535  < 2e-16 ***
## bmi                         573.180    167.467   3.423 0.000639 ***
## I(bmi^2)                     -8.742      2.634  -3.319 0.000928 ***
## factor(has_children)yes     986.040    268.318   3.675 0.000247 ***
## factor(smoker)yes        -20275.787   1647.935 -12.304  < 2e-16 ***
## factor(region)northwest    -654.393    381.787  -1.714 0.086757 .
## factor(region)southeast   -1186.458    383.131  -3.097 0.001998 **
## factor(region)southwest   -1271.236    382.781  -3.321 0.000921 ***
## bmi:factor(smoker)yes      1437.700     52.605  27.330  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4848 on 1328 degrees of freedom
## Multiple R-squared:  0.8408, Adjusted R-squared:  0.8397
## F-statistic: 779.3 on 9 and 1328 DF,  p-value: < 2.2e-16
```

The partial F-test resulted in a p-value of 0.0009281, which is less than the significance level of 0.05. Thus we reject the null hypothesis and conclude that the full model with both interaction and quadratic terms provides a better fit to the data than the reduced model with only main effects.

Final model is the model with both interaction and quadratic terms:

- The final model is the interaction model with the following equation:
- charges = -10714.746 + 262.505(age) + 573.180(bmi) - 8.742(bmi^2) + 986.040(factor(has_children)=yes) - 20275.787(factor(smoker)=yes) - 654.393(factor(region)=northwest) - 1186.458(factor(region)=southeast) - 1271.236(factor(region)=southwest) + 1437.700(bmi:factor(smoker)=yes) + ϵ
- Y(charges) = β0 + β1(age) + β3(bmi) + β4(bmi^2) + β5(factor(has_children)) + β6(factor(smoker)) +β7(factor(region)) + β8(bmi:factor(smoker))+ ϵ

Multiple Regression Assumptions

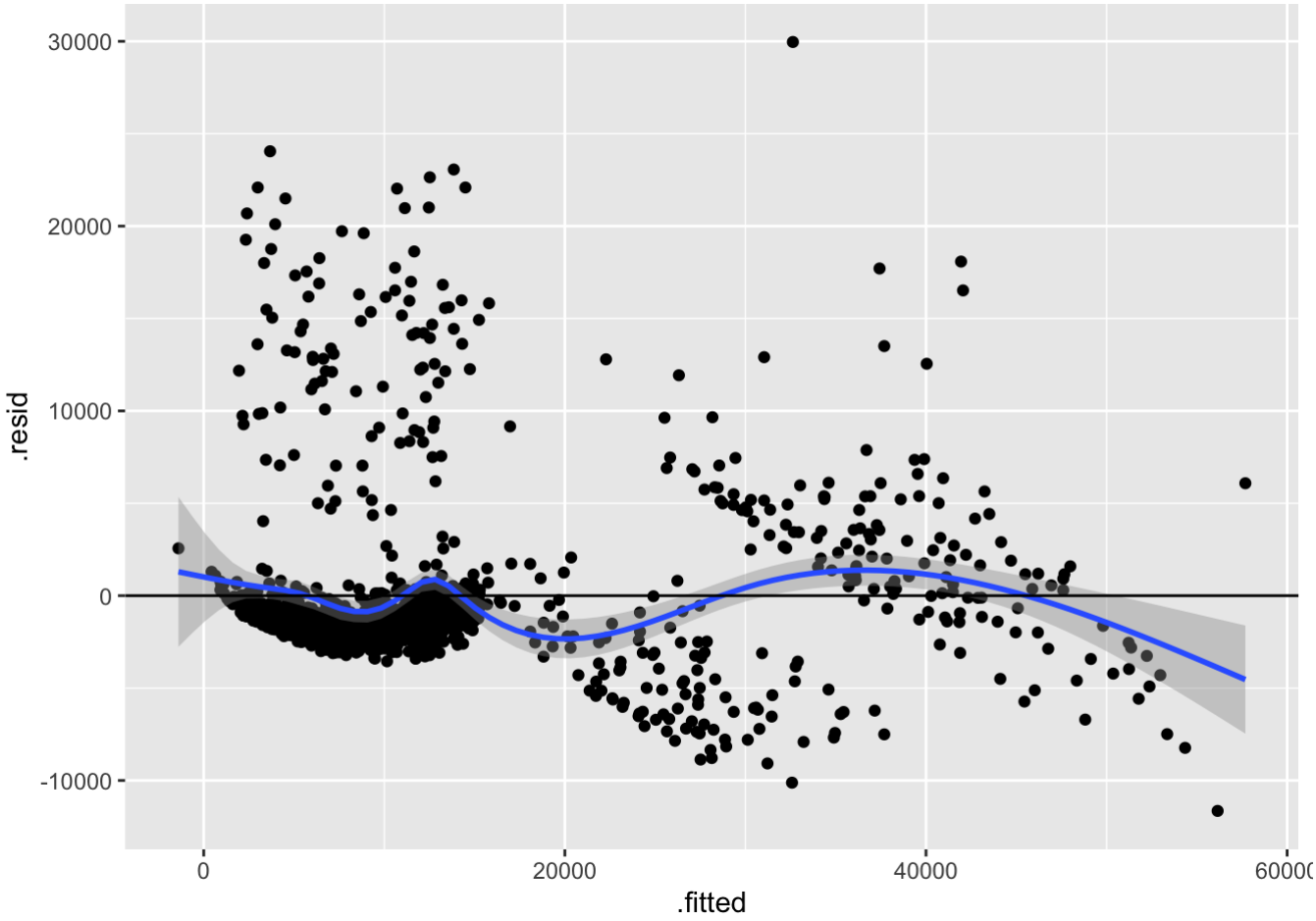
To ensure the reliability of our model results, we performed tests to verify its adherence to several assumptions associated with multiple regression. The following sections detail the methods used to test for these assumptions.

1. **Linearity Assumption:**

We can check the linearity assumption by plotting the residuals against the fitted values. The plot should show no pattern, and the residuals should be evenly scattered around zero.

```
medical_cost_best_model = lm(charges ~ age + bmi + I(bmi^2) + factor(has_children) + fac
tor(smoker)+ factor(region) + (bmi:factor(smoker)) ,data=medical_cost)
# Linearity Assumption - Review residual plots
ggplot(medical_cost_best_model, aes(x=.fitted, y=.resid)) +
  geom_point() + geom_smooth()+
  geom_hline(yintercept = 0)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



2. **Independence Assumption:**

The independence assumption states that the errors (residuals) should not be correlated with each other. In the case of the medical cost data, we do not have any indication that the residuals are dependent, so the assumption of independence is not violated.

3. **Normality Assumption:**

To ensure the validity of our multiple regression analysis, it is necessary for the residuals to have a normal distribution. This can be tested by examining the histogram and qqplot of the residuals.

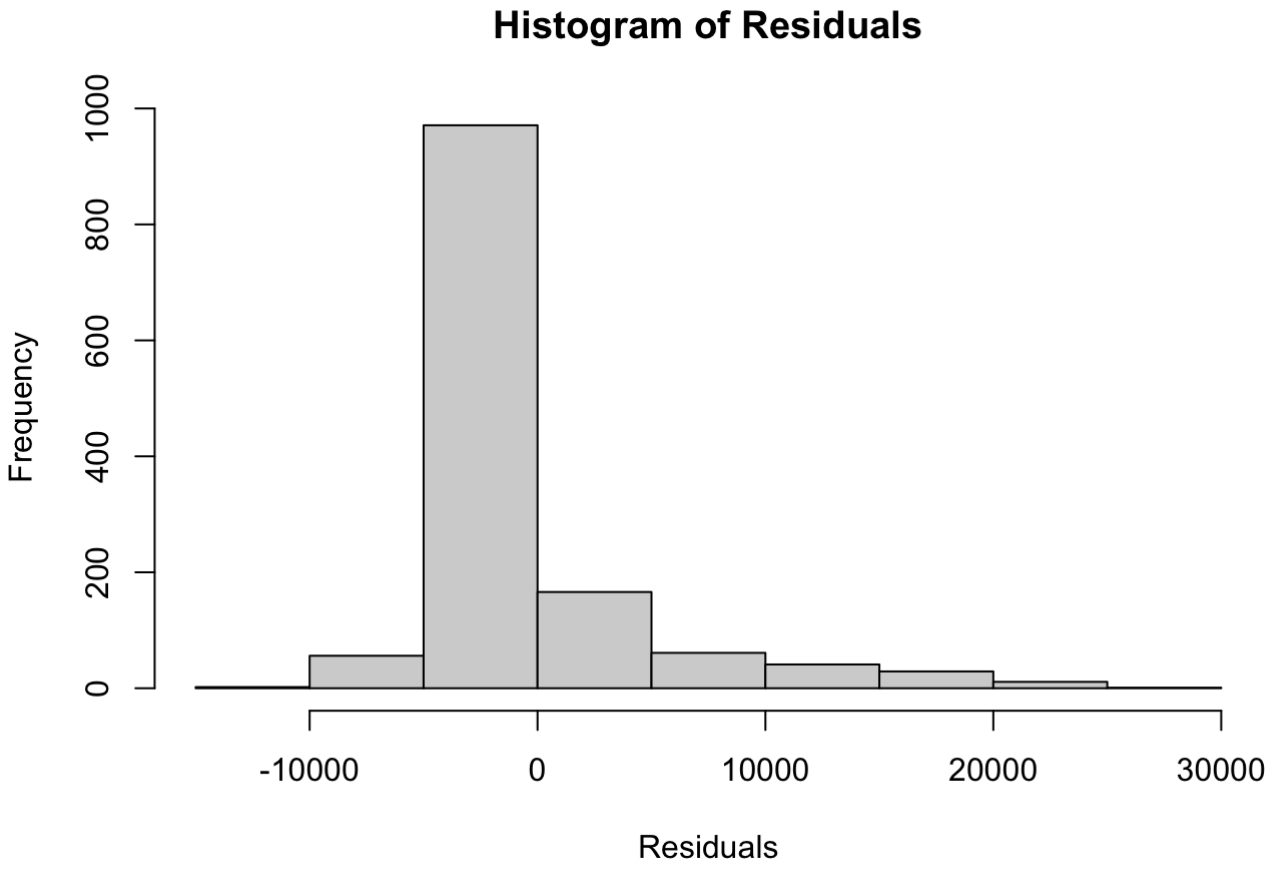
However, we observe that the data points do not conform to a normal distribution, as there are a few points deviating from the normal line towards the tails, suggesting the possibility of outliers.

We can use the Shapiro-Wilk normality test to test this assumption:

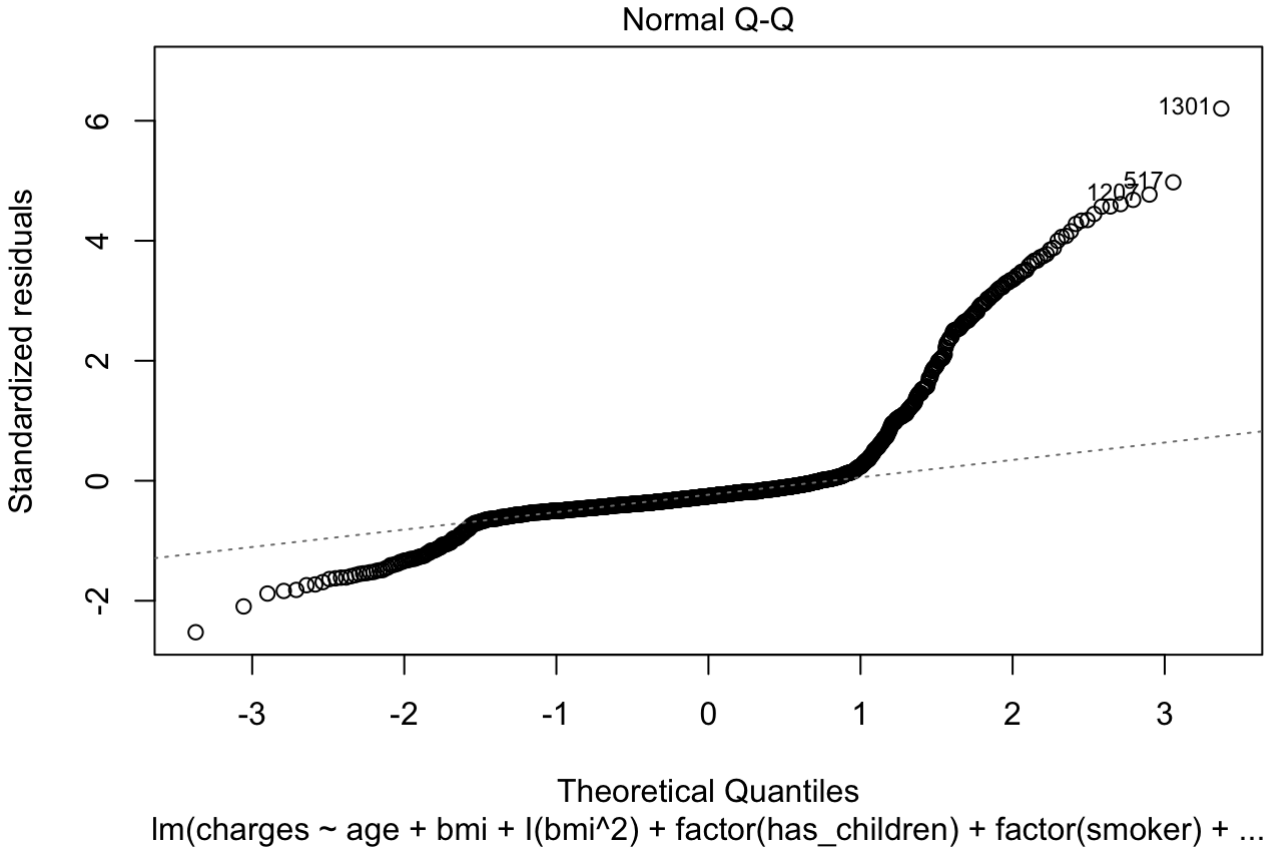
Null Hypothesis H(0): The sample data is normally distributed

Alternate Hypothesis H(A): The sample data is not normally distributed

```
# Histogram of Residuals
hist(resid(medical_cost_best_model),
     main = "Histogram of Residuals",
     xlab = "Residuals")
```




```
# QQ Plot of Residuals
plot(medical_cost_best_model, which=2)
```




```
shapiro.test(resid(medical_cost_best_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(medical_cost_best_model)
## W = 0.68877, p-value < 2.2e-16
```

The test results in a p-value of 2.2e-16, which is less than 0.05, which means that we reject our null hypothesis, that the normality assumption is violated

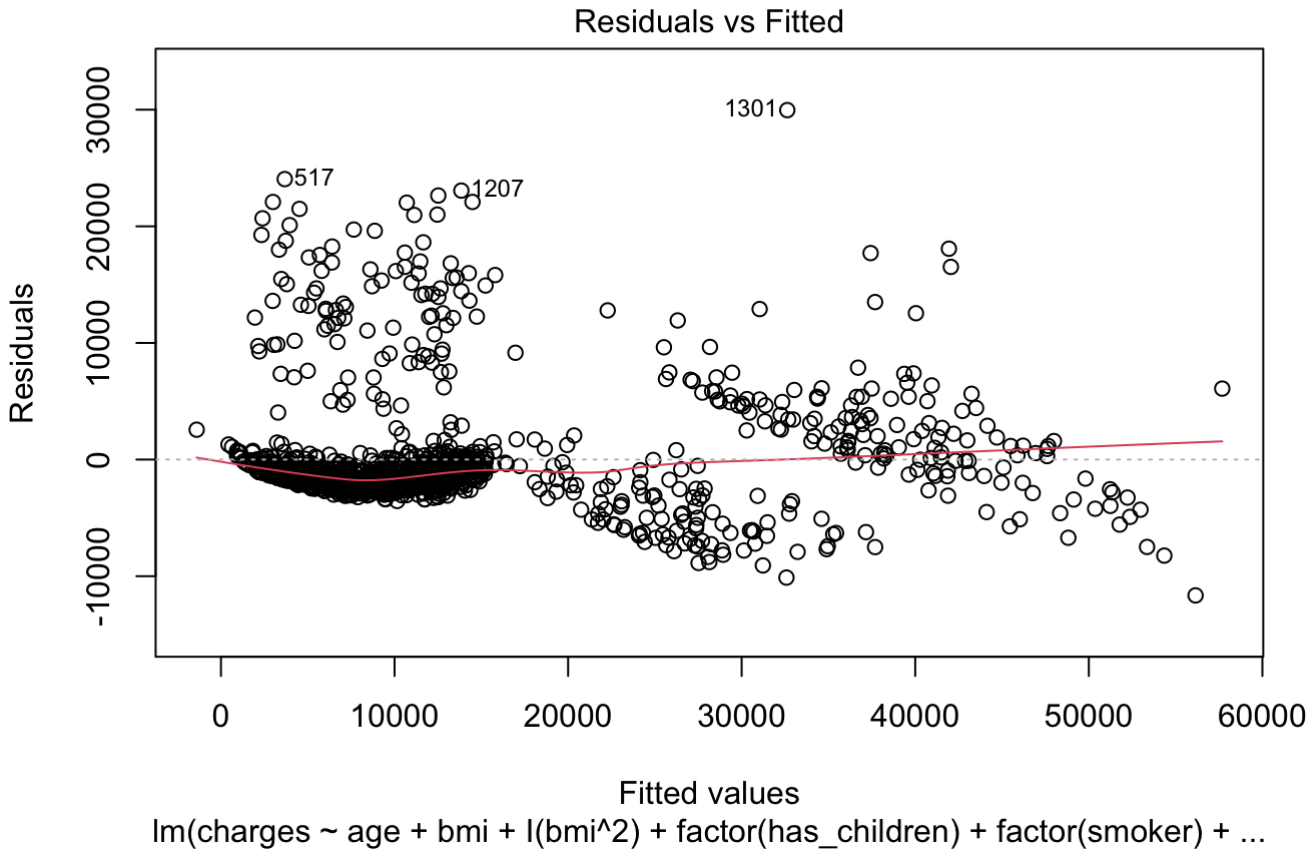
4. **Equal Variance Assumption (heteroscedasticity) :**

To test for heteroscedasticity, we used the studentized Breusch-Pagan test with the following hypotheses:

Null Hypothesis H(0): Heteroscedasticity is not present

Alternative Hypothesis H(A): Heteroscedasticity is present

```
plot(medical_cost_best_model, which=1)
```




```
bptest(medical_cost_best_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  medical_cost_best_model
## BP = 14.181, df = 9, p-value = 0.116
```

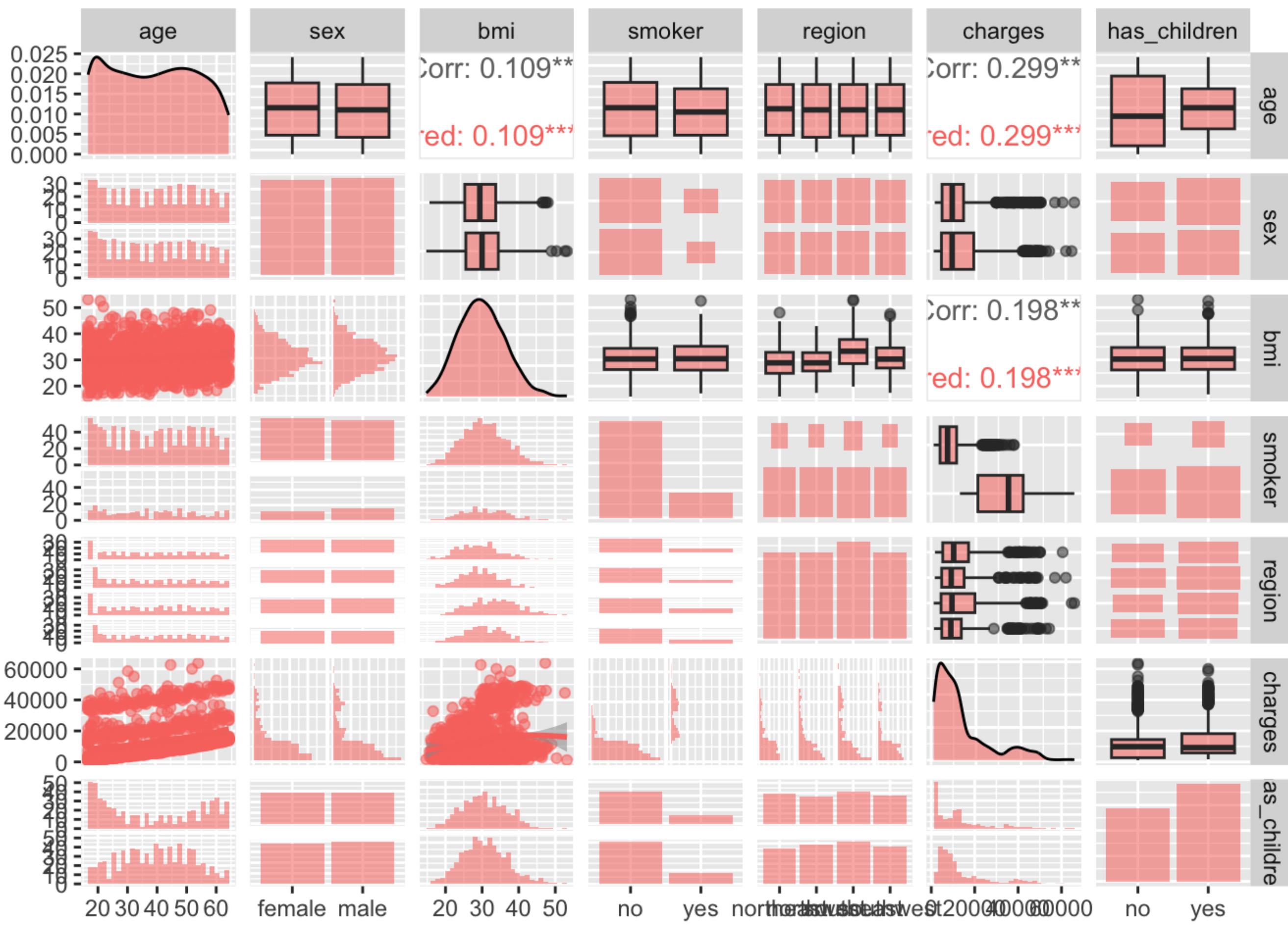
We also examined the plot of fits to residuals, which revealed an no specific pattern against fitted values, indicating equal variance. Furthermore, the results of the Breusch-Pagan test (BP = 14.181, df = 9, p-value = 0.116) failed to reject the null hypothesis, suggesting that our model is homoscedastic.

5. **Multicollinearity Assumption:**

The VIF method we used did not detect multi-collinearity in our model. To ensure that there were no highly correlated variables, we also used a ggpairs function to visualize the correlations between variables. From this, we concluded that the variables included in our first order model (age, children, bmi, smoker, and region) did not have extremely high correlations with each other ($r > 0.80$). Therefore, we can assume that multicollinearity is not a major issue in our model.

```
ggpairs(medical_cost,aes(color = "red", alpha = 0.5), lower = list(continuous = "smooth_
loess", combo = "facethist", discrete = "facetbar", na = "na"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
imcdiag(medical_firstordermodel, method="VIF")
```

```
##
## Call:
## imcdiag(mod = medical_firstordermodel, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                             VIF detection
## age                      1.0166         0
## bmi                      1.1041         0
## factor(has_children)yes 1.0037         0
## factor(smoker)yes       1.0064         0
## factor(region)northwest 1.5192         0
## factor(region)southeast 1.6525         0
## factor(region)southwest 1.5289         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

6. **Influential Points and Outliers**

To ensure the validity of our model, we assessed the presence of influential points and outliers by analyzing leverage and Cook's distance. We identified several outliers in the data using a leverage cutoff value of 3p/n. The leverage and Cook's distance plots further confirmed the presence of these outliers. By removing these observations, we aimed to improve the model's overall performance and reliability.

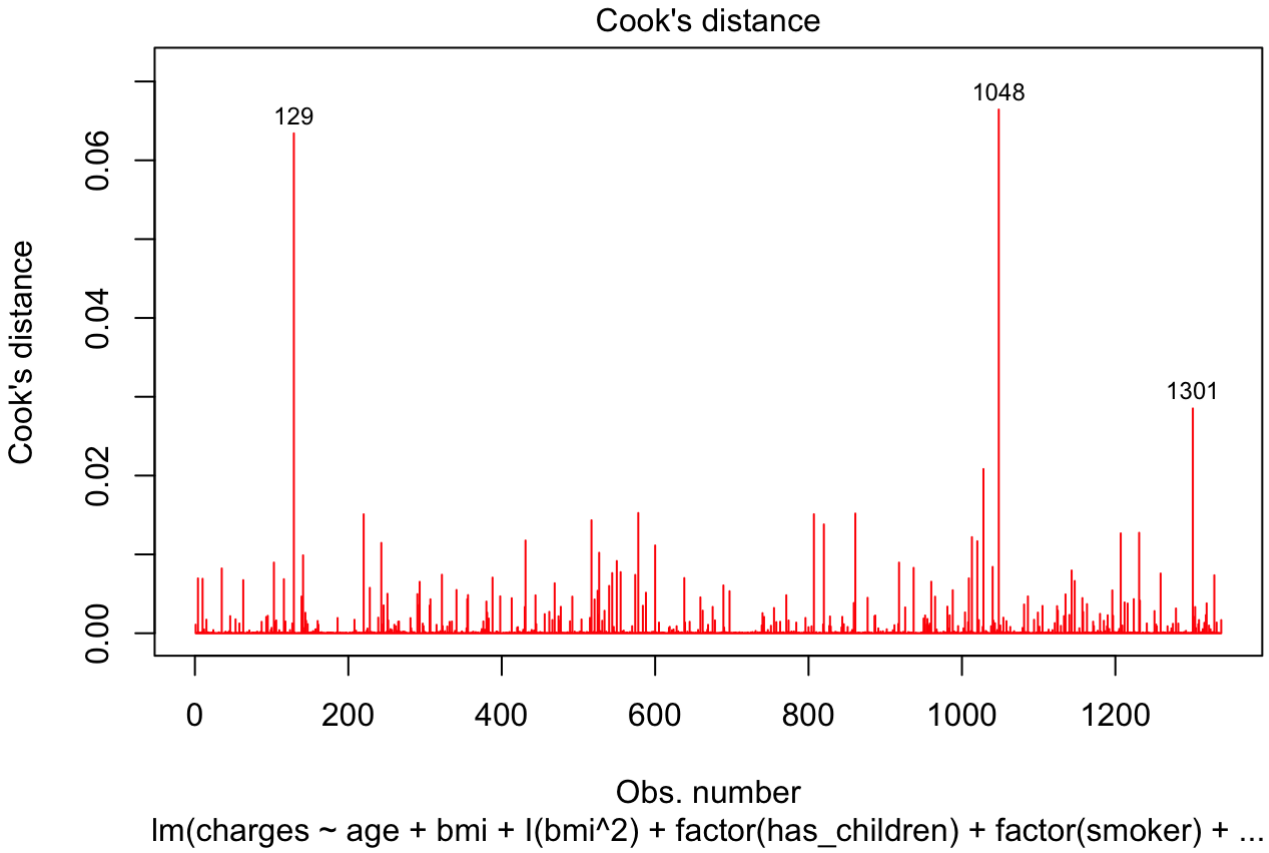
```
# Number of observations
n <- nrow(medical_cost)
# Number of predictor variables (including the intercept)
p <- length(coef(medical_cost_best_model))

# Calculate cutoff value for leverage
cutoff <- 3 * p / n
cutoff
```

```
## [1] 0.02242152
```

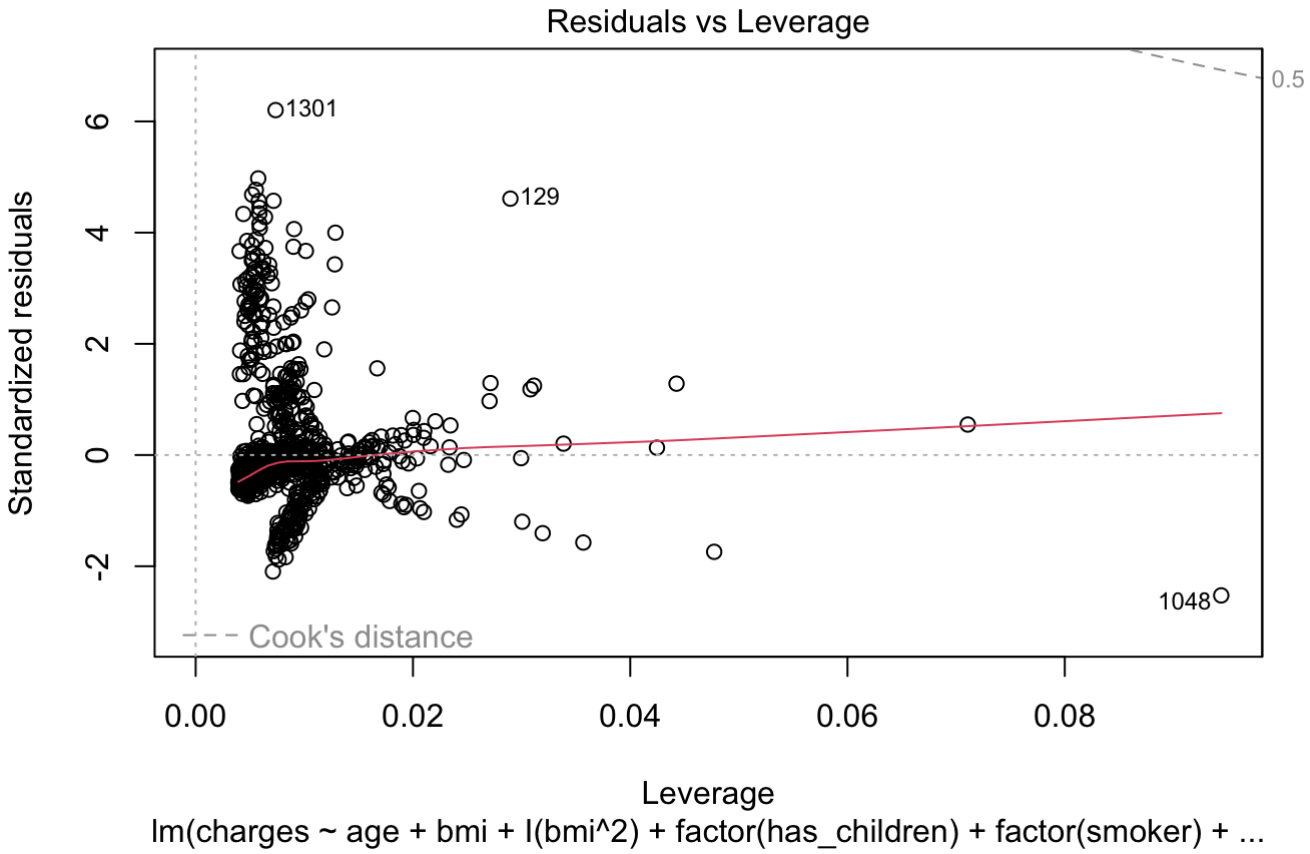
```
# Calculate leverage values
leverage_values <- hatvalues(medical_cost_best_model)
# Identify observations with high leverage values
outliers <- which(leverage_values > cutoff)

outliers
```

```
##  100  117  129  251  287  293  381  402  413  544  548  550  675  848  861 1048
##  100  117  129  251  287  293  381  402  413  544  548  550  675  848  861 1048
## 1086 1089 1125 1157 1318
## 1086 1089 1125 1157 1318
```

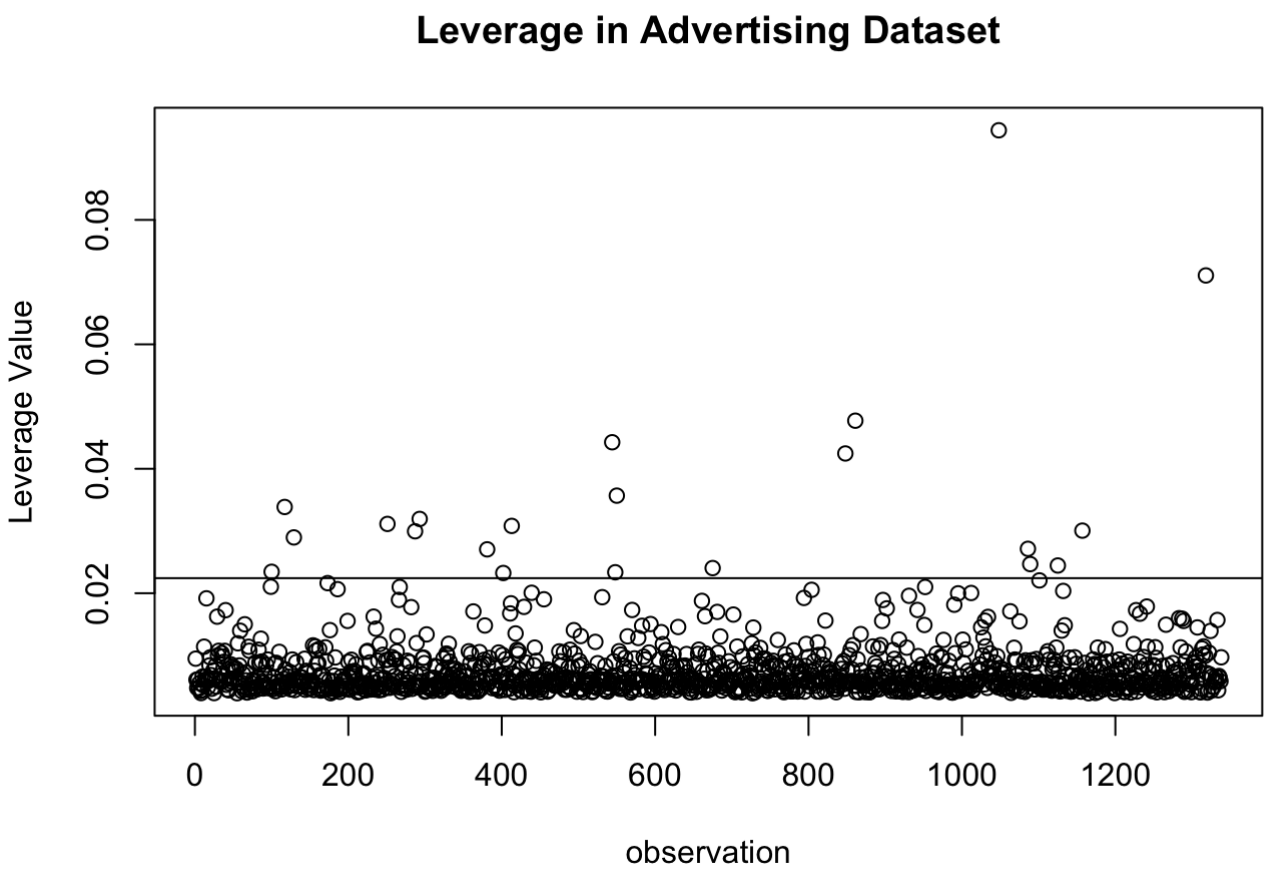
```
#Cooks Distance
plot(medical_cost_best_model, pch=18,col="red",which=c(4))
```




```
# Residual vs Leverage plot
plot(medical_cost_best_model, which = 5)
```




```
plot(rownames(medical_cost),leverage_values, main = "Leverage in Advertising Dataset", x
lab="observation",
ylab = "Leverage Value")
abline(h = 3 *p/n, lty = 1)
```




```
if (length(outliers) > 0) {
  medical_no_outliers <- medical_cost[-outliers, ]
} else {
  cat("No outliers detected.")
}
```

```
# linear regression model after removing outliers
medical_cost_best_model <- lm(charges ~ age + bmi + I(bmi^2) + factor(has_children) + fa
ctor(smoker) + factor(region) + (bmi:factor(smoker)), data = medical_no_outliers)
summary(medical_cost_best_model)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + I(bmi^2) + factor(has_children) +
##     factor(smoker) + factor(region) + (bmi:factor(smoker)), data = medical_no_outlier
s)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9979.9 -2054.7 -1248.7  -312.1 30015.8
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -9460.151   3016.597  -3.136 0.001751 **
## age                         261.524      9.470  27.615  < 2e-16 ***
## bmi                         496.050    196.692   2.522 0.011788 *
## I(bmi^2)                     -7.530      3.152  -2.389 0.017047 *
## factor(has_children)yes     960.953    266.829   3.601 0.000328 ***
## factor(smoker)yes        -25505.415   1845.442 -13.821  < 2e-16 ***
## factor(region)northwest    -699.189    378.467  -1.847 0.064912 .
## factor(region)southeast   -1191.259    380.851  -3.128 0.001800 **
## factor(region)southwest   -1304.699    380.463  -3.429 0.000624 ***
## bmi:factor(smoker)yes      1608.891     59.471  27.054  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4785 on 1307 degrees of freedom
## Multiple R-squared:  0.8393, Adjusted R-squared:  0.8382
## F-statistic: 758.5 on 9 and 1307 DF,  p-value: < 2.2e-16
```

The linear regression model after removing the outliers is as follows:

charges ~ age + bmi + I(bmi^2) + factor(has_children) + factor(smoker) + factor(region) + (bmi:factor(smoker))

The model summary indicates a strong model performance with an adjusted R-squared value of 0.8382, suggesting that approximately 83.82% of the variability in medical charges can be explained by the model. Most of the predictor variables are statistically significant at the 5% level or lower, except for the factor(region)northwest variable, which has a p-value slightly above the 5% threshold (0.0649). The model's residual standard error is 4785, and the F-statistic has a p-value less than 2.2e-16, indicating that the model is statistically significant overall.

Exploring Data Transformations for Model Improvement

Since our model failed Normality assumption we would be trying some transformation :

```
# linear regression model
medical_cost_best_model <- lm(charges ~ age + bmi + I(bmi^2) + factor(has_children) + fa
ctor(smoker) + factor(region) + (bmi:factor(smoker)), data = medical_no_outliers)

# select numerical variables to scale
num_vars <- c("charges", "age", "bmi")

# scale numerical variables
medical_cost_scaled <- medical_no_outliers
medical_cost_scaled[, num_vars] <- scale(medical_cost_scaled[, num_vars])

# fit the linear regression model using scaled data
scaled_medical_cost_best_model <- lm(charges ~ age + bmi + I(bmi^2) + factor(has_childre
n) + factor(smoker) + factor(region) + (bmi:factor(smoker)), data = medical_cost_scaled)

#checking homoscedasticity
bptest(scaled_medical_cost_best_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  scaled_medical_cost_best_model
## BP = 13.287, df = 9, p-value = 0.15
```

```
#Testing for Normality
shapiro.test(residuals(scaled_medical_cost_best_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(scaled_medical_cost_best_model)
## W = 0.66888, p-value < 2.2e-16
```

```
# Box-Cox transformation
library(MASS)
```

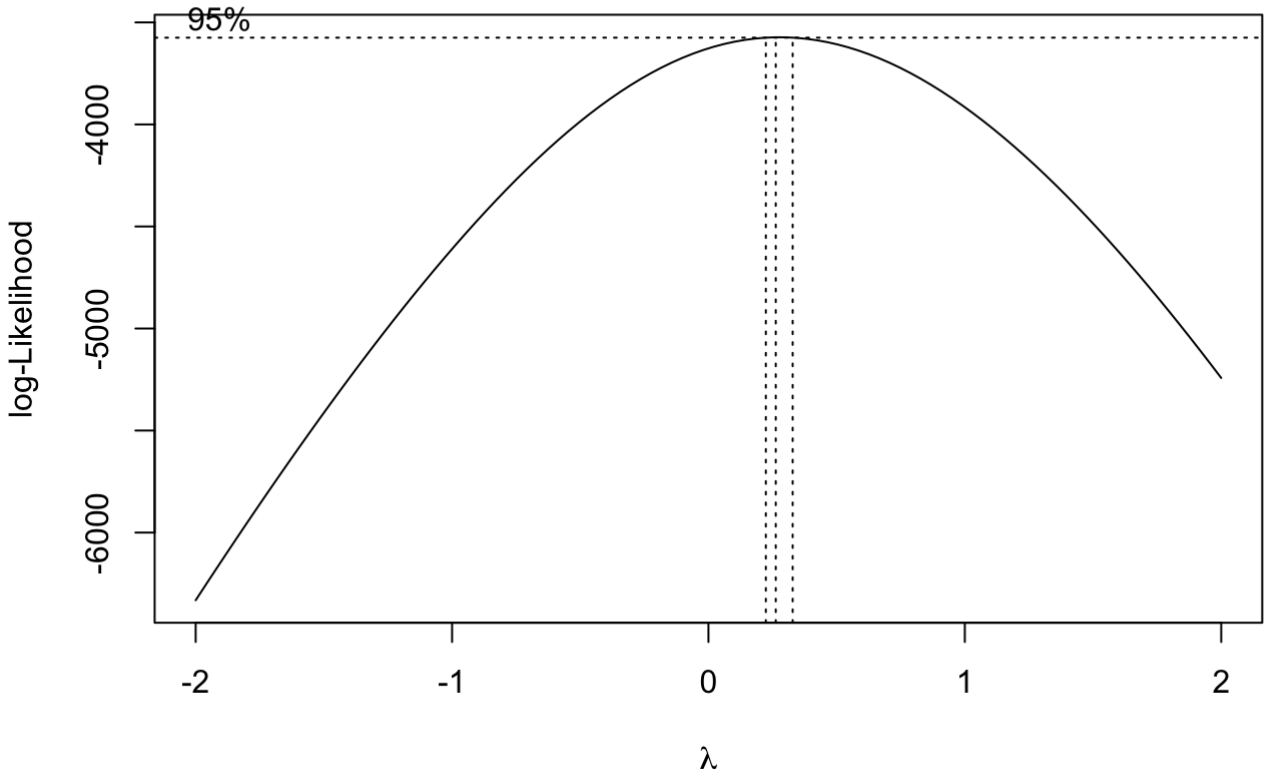
```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:EnvStats':
##
##     boxcox
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## The following object is masked from 'package:olsrr':
##
##     cement
```

```
bc <- boxcox(medical_cost_best_model, lambda = seq(-2, 2, by = 0.1))
```




```
bestlambda <- bc$x[which(bc$y == max(bc$y))]
bestlambda
```

```
## [1] 0.2626263
```

```
medical_cost_boxcox=lm((((charges^0.2626263)-1)/(0.2626263)) ~ age + bmi + I(bmi^2) + fa
ctor(has_children) + smoker+ factor(region) + (bmi:factor(smoker)), data= medical_no_out
liers)

#checking homoscedasticity
bptest(medical_cost_boxcox)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  medical_cost_boxcox
## BP = 40.488, df = 9, p-value = 6.198e-06
```

```
#Testing for Normality
shapiro.test(residuals(medical_cost_boxcox))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(medical_cost_boxcox)
## W = 0.75861, p-value < 2.2e-16
```

```
# Square root transformation
medical_cost_sqrt <- lm(sqrt(charges) ~ age + bmi + I(bmi^2) + factor(has_children) + fa
ctor(smoker)+ factor(region) + (bmi:factor(smoker)), data=medical_no_outliers)
#checking homoscedasticity
bptest(medical_cost_sqrt)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  medical_cost_sqrt
## BP = 22.953, df = 9, p-value = 0.006302
```

```
#Testing for Normality
shapiro.test(residuals(medical_cost_sqrt))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(medical_cost_sqrt)
## W = 0.69115, p-value < 2.2e-16
```

```
# Exponential transformation
medical_cost_exp <- lm(log(charges) ~ age + bmi + I(bmi^2) + factor(has_children) + fact
or(smoker)+ factor(region) + (bmi:factor(smoker)), data=medical_no_outliers)
#checking homoscedasticity
bptest(medical_cost_exp)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  medical_cost_exp
## BP = 78.295, df = 9, p-value = 3.522e-13
```

```
#Testing for Normality
shapiro.test(residuals(medical_cost_exp))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(medical_cost_exp)
## W = 0.84131, p-value < 2.2e-16
```

```
# Inverse transformation
medical_cost_inv <- lm(1/charges ~ age + bmi + I(bmi^2) + factor(has_children) + factor
(smoker)+ factor(region) + (bmi:factor(smoker)), data=medical_no_outliers)
#checking homoscedasticity
bptest(medical_cost_inv)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  medical_cost_inv
## BP = 261.21, df = 9, p-value < 2.2e-16
```

```
#Testing for Normality
shapiro.test(residuals(medical_cost_inv))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(medical_cost_inv)
## W = 0.91435, p-value < 2.2e-16
```

Since our initial model failed the normality assumption, we attempted various data transformations to address this issue, including scaling the numerical variables, applying the Box-Cox, square root, exponential, and inverse transformations. However, none of them were successful in meeting the normality assumption. It might be appropriate to consider non-parametric methods or generalized linear models (GLM) to address this issue. Collecting additional data might also be an option, if possible. It is crucial to acknowledge that violating the normality assumption may affect the model's prediction accuracy and should be cautiously considered when interpreting the findings.

```
# Our final model:
medical_cost_best_model <- lm(charges ~ age + bmi + I(bmi^2) + factor(has_children) + fa
ctor(smoker) + factor(region) + (bmi:factor(smoker)), data = medical_no_outliers)
summary(medical_cost_best_model)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + I(bmi^2) + factor(has_children) +
##     factor(smoker) + factor(region) + (bmi:factor(smoker)), data = medical_no_outlier
s)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9979.9 -2054.7 -1248.7  -312.1 30015.8
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -9460.151   3016.597  -3.136 0.001751 **
## age                         261.524      9.470  27.615  < 2e-16 ***
## bmi                         496.050    196.692   2.522 0.011788 *
## I(bmi^2)                     -7.530      3.152  -2.389 0.017047 *
## factor(has_children)yes     960.953    266.829   3.601 0.000328 ***
## factor(smoker)yes        -25505.415   1845.442 -13.821  < 2e-16 ***
## factor(region)northwest    -699.189    378.467  -1.847 0.064912 .
## factor(region)southeast   -1191.259    380.851  -3.128 0.001800 **
## factor(region)southwest   -1304.699    380.463  -3.429 0.000624 ***
## bmi:factor(smoker)yes      1608.891     59.471  27.054  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4785 on 1307 degrees of freedom
## Multiple R-squared:  0.8393, Adjusted R-squared:  0.8382
## F-statistic: 758.5 on 9 and 1307 DF,  p-value: < 2.2e-16
```

```
residuals <- residuals(medical_cost_best_model)
mse <- mean(residuals^2)
rmse <- sqrt(mse)
rmse
```

```
## [1] 4766.874
```

The final model is:

Y(charges) = β0 + β1(age) + β3(bmi) + β4(bmi^2) + β5(factor(has_children)) + β6(factor(smoker)) +β7(factor(region)) + β8(bmi:factor(smoker))+ ϵ

charges = -9460.151 + 261.524 * age + 496.050 * bmi - 7.530 * (bmi^2) + 960.953 * factor(has_children=yes) - 25505.415 * factor(smoker=yes) - 699.189 * factor(region=northwest) - 1191.259 * factor(region=southeast) - 1304.699 * factor(region=southwest) + 1608.891 * bmi:factor(smoker=yes)

Interpretation of coefficients:

Interpretation of coefficients:

β0 (Intercept): The baseline medical charge, -9460.151, is the expected charge for a non-smoking individual with no children, a BMI of 0, and residing in the northeast region. This value is not meaningful in practice, as it is not possible to have a negative medical charge or a BMI of 0.

β1 (age): A one-year increase in age is associated with an increase of 261.524 in medical charges, holding all other factors constant.

β3 (bmi) and β4 (bmi^2): The relationship between BMI and medical charges is quadratic, with an increase of 496.050 in charges per unit increase in BMI and a decrease of 7.530 in charges per unit increase in BMI squared. This suggests that the effect of BMI on charges is not linear but follows a curve.

β5 (factor(has_children=yes)): Having children is associated with an increase of 960.953 in medical charges compared to not having children, holding all other factors constant.

β6 (factor(smoker=yes)): Being a smoker is associated with a decrease of 25505.415 in medical charges compared to being a non-smoker, holding all other factors constant. However, this should be interpreted with caution due to the interaction term.

β7 (factor(region)): The coefficients for region indicate the differences in medical charges relative to the northeast region (the reference category). The northwest, southeast, and southwest regions have lower medical charges by 699.189, 1191.259, and 1304.699, respectively, compared to the northeast region, holding all other factors constant.

β8 (bmi:factor(smoker=yes)): The interaction term between BMI and smoking status indicates that for a one-unit increase in BMI, a smoker's medical charges increase by 1608.891 more than a non-smoker's, holding all other factors constant. This means that the effect of BMI on medical charges is different for smokers compared to non-smokers.

We considered the following model performance metrics: adjusted R-squared, residual standard error (RSE), and root mean squared error (RMSE).

Adjusted R-squared measures the proportion of variance in the dependent variable that is explained by the independent variables in the model, adjusted for the number of independent variables in the model. In this case, the adjusted R-squared value is 0.8382, indicating that the model explains 83.82% of the variance in the dependent variable (charges), after adjusting for the number of independent variables.

RSE and RMSE are measures of the variability of the residuals (the difference between the predicted and actual values of the dependent variable) in the model. RSE is the standard deviation of the residuals, while RMSE is the square root of the mean squared error of the residuals. In this case, the RSE is 4785. To calculate the RMSE, you can use the formula RMSE = 4766.874, where residuals are the differences between the predicted and actual values of the dependent variable.

Overall, the model seems to have a good fit with the data, as indicated by the high adjusted R-squared value, and the low RSE. However, it is important to note that the model does not satisfy the normality assumption,

Projected Charges

Based on our final linear regression model, we can predict the charges for new patients based on their age, bmi, number of children, smoking status, and region. The model equation is:

```
# linear regression model
medical_cost_best_model <- lm(charges ~ age + bmi + I(bmi^2) + factor(has_children) + fa
ctor(smoker) + factor(region) + (bmi:factor(smoker)), data = medical_no_outliers)
```

To predict the charges for a new patient, we can input their values for each predictor variable into the equation and solve for charges. We can use R to predict the charges for new patients based on our final model.

```
# Predicting charges for new data
newdata = data.frame(age=60, sex="female", bmi=25.84, has_children="yes", smoker="no", r
egion="northwest")
newdata2 = data.frame(age=19, sex="female", bmi=27.9, has_children="no", smoker="yes", r
egion="southwest")

predicted_charges_nonsmoker <- predict(medical_cost_best_model, newdata = newdata)
predicted_charges_smoker <- predict(medical_cost_best_model, newdata = newdata2)
predict(medical_cost_best_model,newdata,interval="predict")
```

```
##       fit      lwr      upr
## 1 14283.1 4868.095 23698.11
```

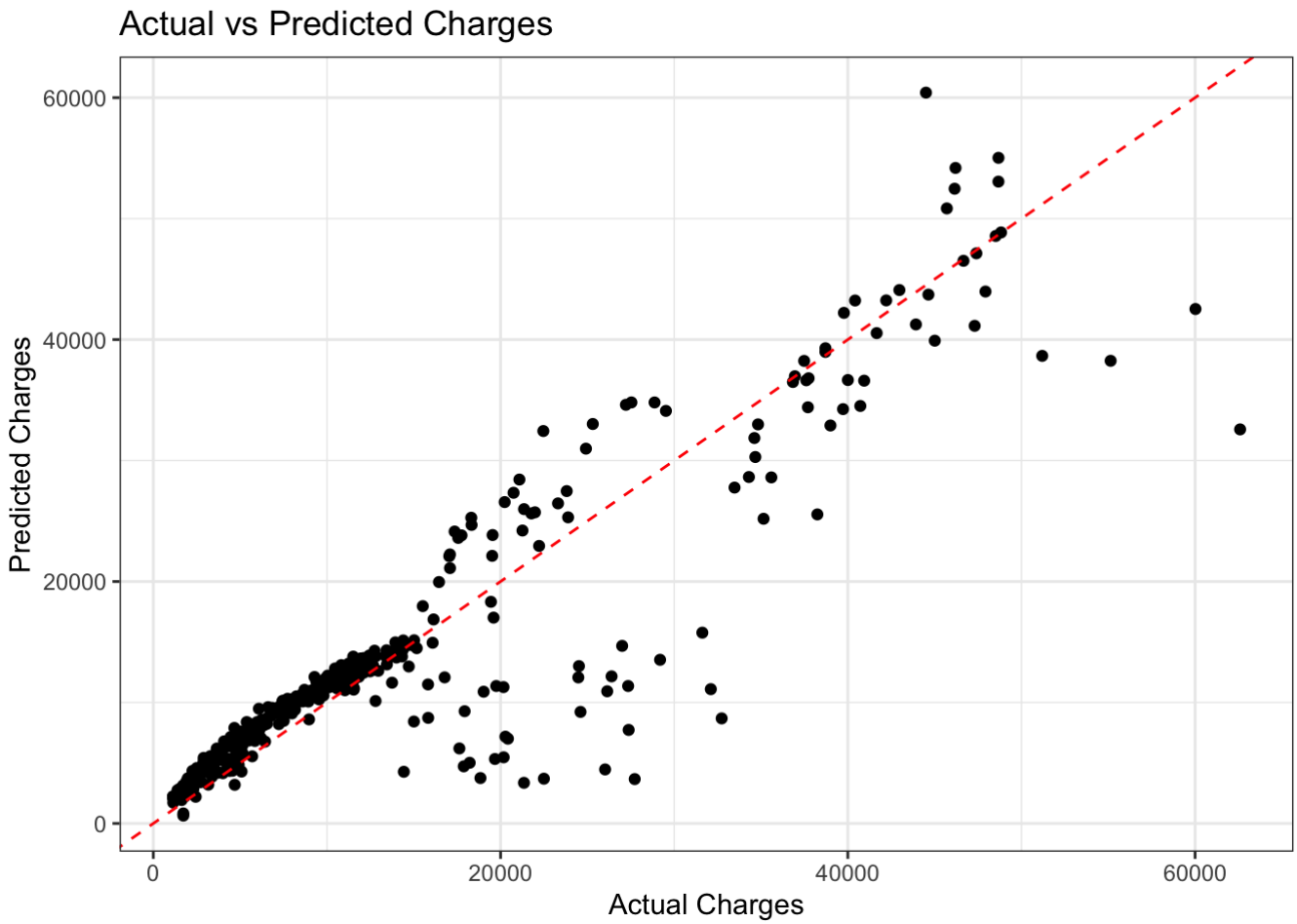
```
predict(medical_cost_best_model,newdata2,interval="predict")
```

```
##        fit     lwr      upr
## 1 21565.04 12131.4 30998.68
```

For the first data point (newdata), the predicted medical charge is $14,283.1. For this data point, the individual is a 60-year-old female with a BMI of 25.84, with children, a non-smoker, and lives in the northwest region.

For the second data point (newdata2), the predicted medical charge is $21,565.04. For this data point, the individual is a 19-year-old female with a BMI of 27.9, no children, a smoker, and lives in the southwest region.

```
# Predicting charges for 30% of the actual data
newdata <- medical_cost[sample(nrow(medical_no_outliers), nrow(medical_no_outliers)*0.
3), ]
predicted_charges <- predict(medical_cost_best_model, newdata = newdata)

# Create a data frame with actual and predicted charges
plot_data <- data.frame(actual_charges = newdata$charges,
                        predicted_charges = predicted_charges)
# Create scatter plot
ggplot(plot_data, aes(x = actual_charges, y = predicted_charges)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color='red') +
  labs(x = "Actual Charges", y = "Predicted Charges",
       title = "Actual vs Predicted Charges") +
  theme_bw()
```



We plotted predicting medical charges for 30% of the actual data: The plot visualizes the relationship between actual charges and predicted charges by the model. Each point represents an observation, with the x-axis showing the actual charges and the y-axis displaying the predicted charges. The red dashed line in the plot represents a perfect prediction, where the actual charges are equal to the predicted charges. If the model were perfect, all points would lie exactly on this line. By visually examining the scatter plot, it appears that the model's predictions are reasonably accurate for most observations, as many points are clustered around the red dashed line.

The plot also reveals that the model's performance may vary across different ranges of charges. For instance, it might perform better for lower charges and less accurately for higher charges. In the future, it would be beneficial to investigate potential limitations and areas for improvement in the model's performance. By examining the model's behavior across different ranges of charges, refining feature selection, and exploring alternative modeling techniques, we can work towards enhancing the accuracy and overall effectiveness of the predictive model.

CONCLUSION AND FUTURE RECOMMENDATIONS

In conclusion, the final model we developed has a good fit for the data with an adjusted R-squared value of 0.8382. The model includes variables such as age, BMI, BMI squared, factor(children), factor(smoker), factor(region), and an interaction term between BMI and factor(smoker). The model equation shows that charges increase with age, BMI, and the number of children. Charges are higher for smokers, and there are differences in charges based on region.

However, it is important to note that the normality assumption was violated in the model, indicating that the errors are not normally distributed. To address this issue, we tried several transformations, including Scaling the data, the Box-Cox transformation, square root transformation, exponential transformation, and inverse transformation but none were successful in satisfying the normality assumption. This suggests that there may be other factors that are not captured in the model that influence medical charges.

Future recommendations could include exploring more advanced techniques such as non-parametric regression or ensemble models (like GLM) to improve the model's performance. Additionally, collecting more data and including additional variables could also improve the model's accuracy.

Overall, our model provides a useful tool for estimating medical charges based on several important predictor variables.

References

1. Dataset: Choi, M. (2018, February 21). Medical Cost Personal Datasets. Kaggle. Retrieved March 15, 2023, from https://www.kaggle.com/datasets/mirichoi0218/insurance (https://www.kaggle.com/datasets/mirichoi0218/insurance)
2. Moran, J. L., Solomon, P. J., Peisach, A. R., & Martin, J. (2007). New models for old questions: generalized linear models for cost prediction. Journal of evaluation in clinical practice, 13(3), 381-389.
3. Dalpiaz, D. (n.d.). Applied Statistics with R. Chapter 14 Transformations. Retrieved April 4, 2023, from https://book.stat420.org/transformations.html (https://book.stat420.org/transformations.html)