Insurance Data Project Report

23/10/2020

1 Introduction and Data Description

Healthcare and medical insurance are both multibillion dollar industries in the United States. Given the exorbitant cost of healthcare in the United States, coupled with the large volume of beneficiaries filing for payouts, insurance companies stand to gain valuble insight by being able to accurately predict the charges they can expect to pay for a given beneficiary's healthcare based on payouts for previous beneficiaries, especially those that are in some way similar to the beneficiary in question. To do this, we must first understand the nature of our data, summarised in the table below.

Variable	Description	Type
age	Age of beneficiary in years	continuous quantitative
sex	Sex of beneficiary (male or female)	categorical (nominal)
bmi	Body Mass Index (body mass in kilograms divided by the square of body height in meters)	continuous quantitative
children	Number of children also covered by the beneficiary's insurance policy	categorical (ordinal)
smoker	Indicates whether or not the beneficiary is a regular smoker	categorical (nominal)
region	Region of the United States in which the beneficiary lives	categorical (nominal)
charges	Medical costs billed to the insurance company	continuous quantitative (response variable)

Since the response variable is continuous, we use a regression model, and for interpretability purposes we use a linear regression model in particular. We remind ourselves that the three primary purposes of a linear model are to summarise a (linear) association, make a prediction given certain covariates, and perform causal inference. Causal inference is the most powerful of these purposes, and consequently requires the strongest assumptions to be made or met. Our data cannot be collected via a randomised controlled trial (which is essential for causal infernce) since it's unethical and or impractical to randomly assign people to smoking vs. not smoking, having a certain amount of children, having a specific BMI, etc. As such, the goal of the linear model we construct here is to make predictions about the response variable (charges) given the covariates (all the other aforementioned data). The assumptions for making predictions with a linear model are weaker than those for performing causal inference, and we will discuss how they are met in our analysis.

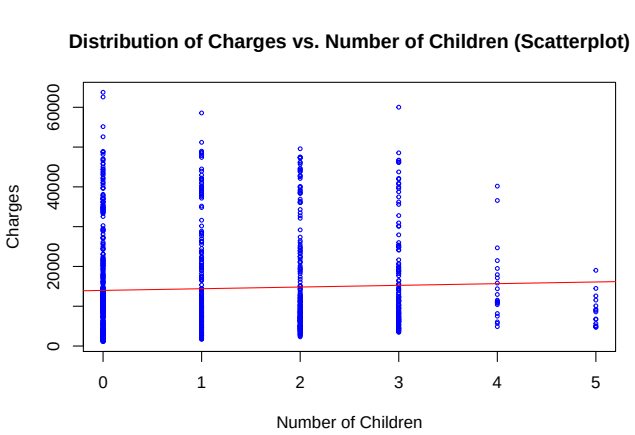
2 Exploratory Data Analysis

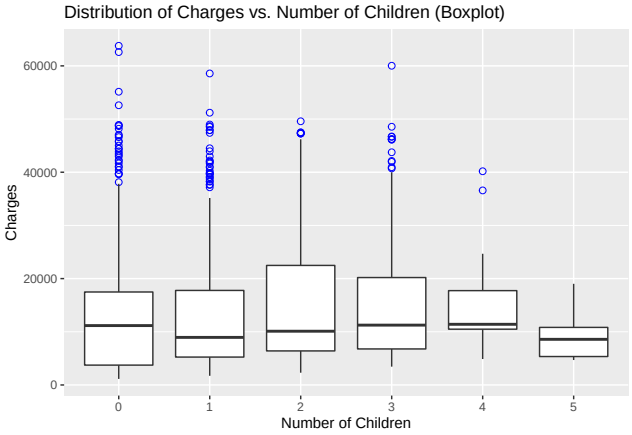
We begin by examining the data for any outliers, null/corrupted/missing values etc. While I found no `na` values in the DataFrame, I noticed that 2 observations out of the total 1100 had charges of -999. Paying a negative charge on a beneficiary's healthcare cost doesn't make sense, so presumably -999 is the designation for missing charges. Since there is no additional resource that may be able to tell us what the charges paid for these two beneficiaries were, we exclude them from our data. Additionally, I noticed that the maximum recorded BMI in the data for a beneficiary was 143.02, which is one of the highest values ever reported[1]. Unlike the previous case, this observation cannot be removed from our analysis, because it's presumably an

[1] https://en.wikipedia.org/wiki/List_of_heaviest_people

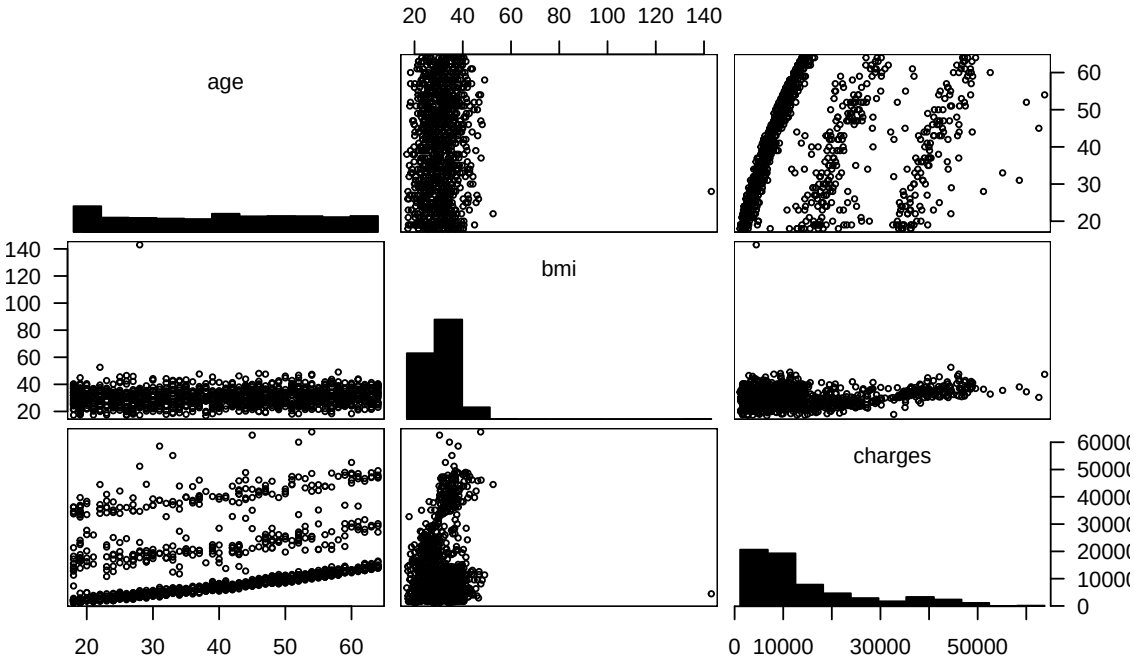
outlier instead of missing data ie. the measurement is correctly recorded and not missing, but substantially different from all our other observations. That being said, we should bear this data point in mind when constructing our model, especially since linear regression is sensitive to outliers. Having noted this, we continue our analysis with the remaining 1108 observations.

Note that while the children variable is technically discrete quantitative, since it takes on such few values we treat it as a categorical variable in our analysis instead, as mentioned in the table above. The figures below help demonstrate this. It's worth noting that based on these plots, it appears that children has little to no linear relationship with charges, at least by itself.

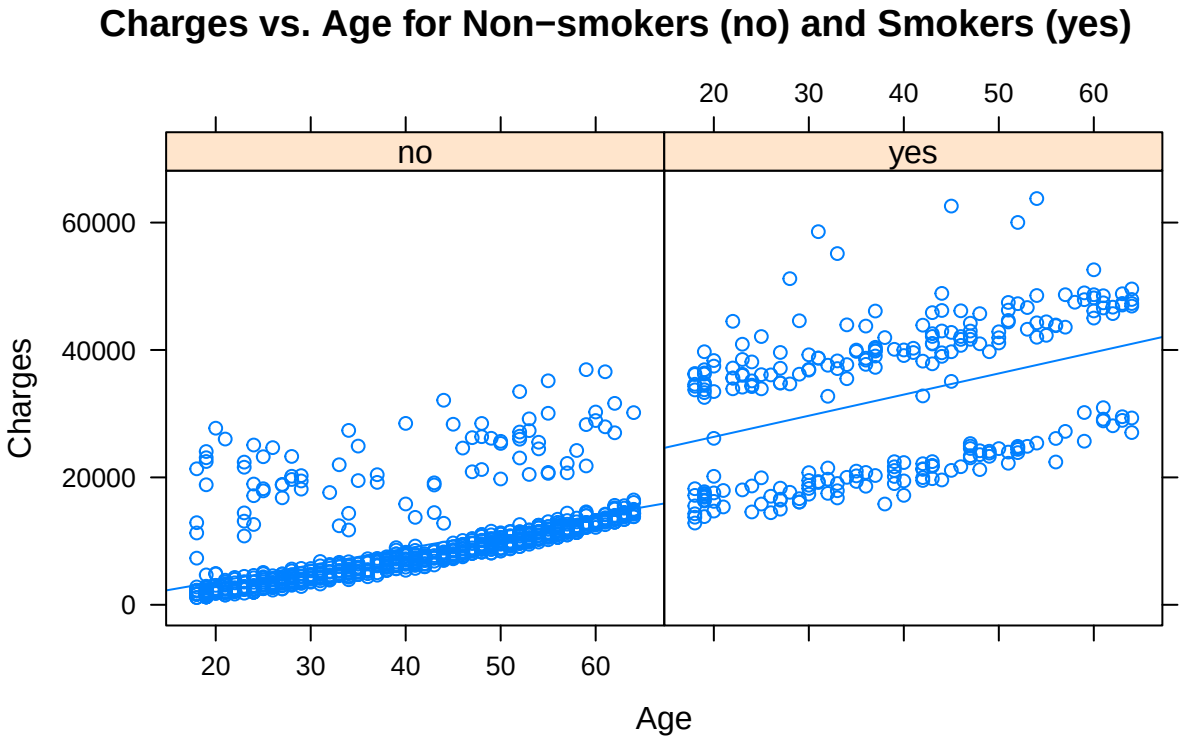




Since we treat children as a categorical variable, the only continuous quantiative variables are age, bmi, and charges. I found the pairwise correlations to be as follows: $\rho(\text{age, bmi}) = 0.1032$, $\rho(\text{age, charges}) = 0.2539$, $\rho(\text{bmi, charges}) = 0.2028$. Ideally age and bmi would have higher correlations with charges, but hopefully we'll be able to come up with an accurate model regardless. Since age and bmi are weakly correlated, we can safely include both features in our model without worrying about collinearity, which would be a problem if the two variables were highly correlated even if they were linearly independent. Note this wouldn't change the accuracy of our predictions, but would affect the interpretation of our model coefficients. Pairwise scatterplots between these variables and univariate distributions are plotted below to visually supplement the provided statistical summary.

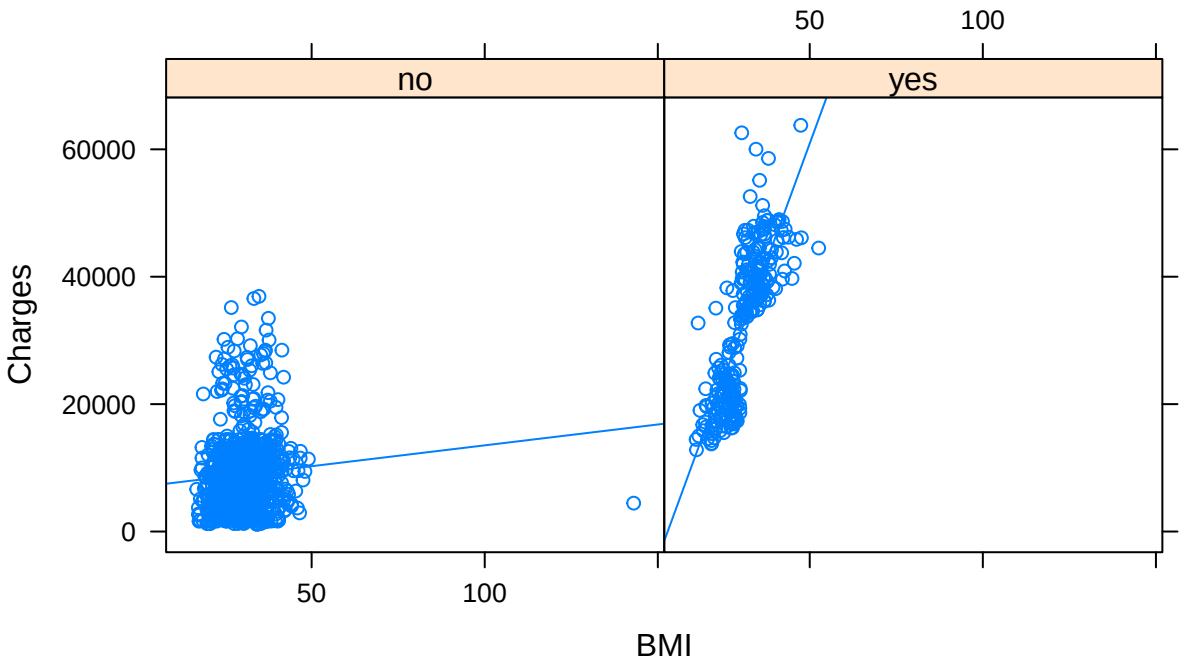


Since ρ(age, charges) $>$ ρ(bmi, charges), we further examine the relationship between age and charge first. We notice that there seem to be 3 distinct trends in the above scatterplot; if we could find some suitable categorical features or combinations thereof that correspond to these different trends, we could come up with a very accurate model. As mentioned earlier, one of the motivations behind choosing a linear model is the interpretability it offers, given that certain assumptions are met. Bearing this in mind, the approach is not simply to fit the model with the highest R^2 or even adjusted R^2 value, but rather to maximise adjusted R^2 value while still meeting the assumptions of the linear model with minimal departures; as such, we must select features carefully. Domain knowledge suggests that smoking is associated with a variety of health problems, must notably lung and throat cancer, not to mention heart disease, so we regress charges on age and smoker. The plots for each value of smoker are provided below.



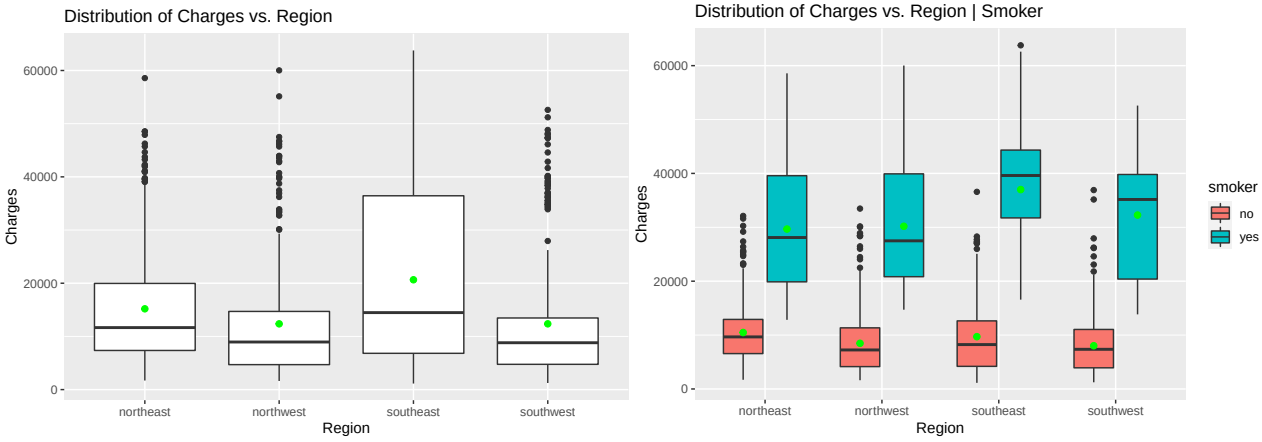
We see that in the non-smoker case, the regression model very accurately models the charges for a majority of the data points corresponding to the lower region of the earlier charges vs. age scatterplot, but still misses out on a cloud of points in the so-called middle region. In the smoker case, the regression line passes between the middle and upper region of data points, resulting in approximately homoschedastic but nontrivial residuals throughout the plot. These shortcomings notwithstanding, the model is now substantially more accurate. Visually, the slopes for both lines appears to be the same, suggesting there's no need for an interaction term, which would result in different slopes for the regression line depending on whether the beneficiary was a smoker or not. We will rigorously test this later on.

We now perform a similar analysis, where we regress charges on bmi and fit a different regression line in each case. From the earlier scatterplot, we see that the relationship between charge and bmi is not clearly linear, and transformations won't change that. Regardless, the conditional regression plot is provided below.



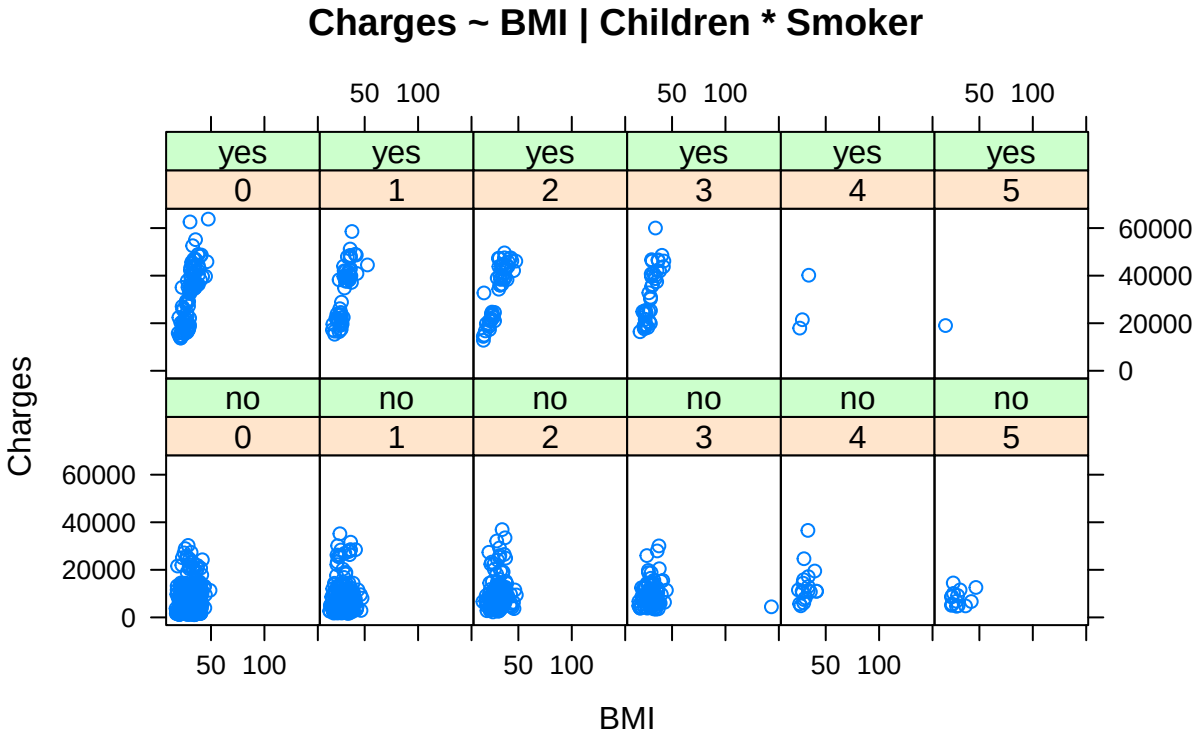
In the non-smoker case, the linear relationship between charges and bmi is still weak and strongly influenced by the single aforementioned outlier. In the smoker case, there is a much stronger linear relationship with our regression line passing through or near the bulk of the points with low residuals. Note that the slopes for btoh conditional regression lines differ significantly, which is a clear indication that our model should include an interaction term to account for this.

Since whether a beneficiary is a smoker seems to play the most important individual role in determining the charge the insurance company pays for their healthcare, we compare distributions of charges for different combinations of smoker and other categorical variables, starting with region.

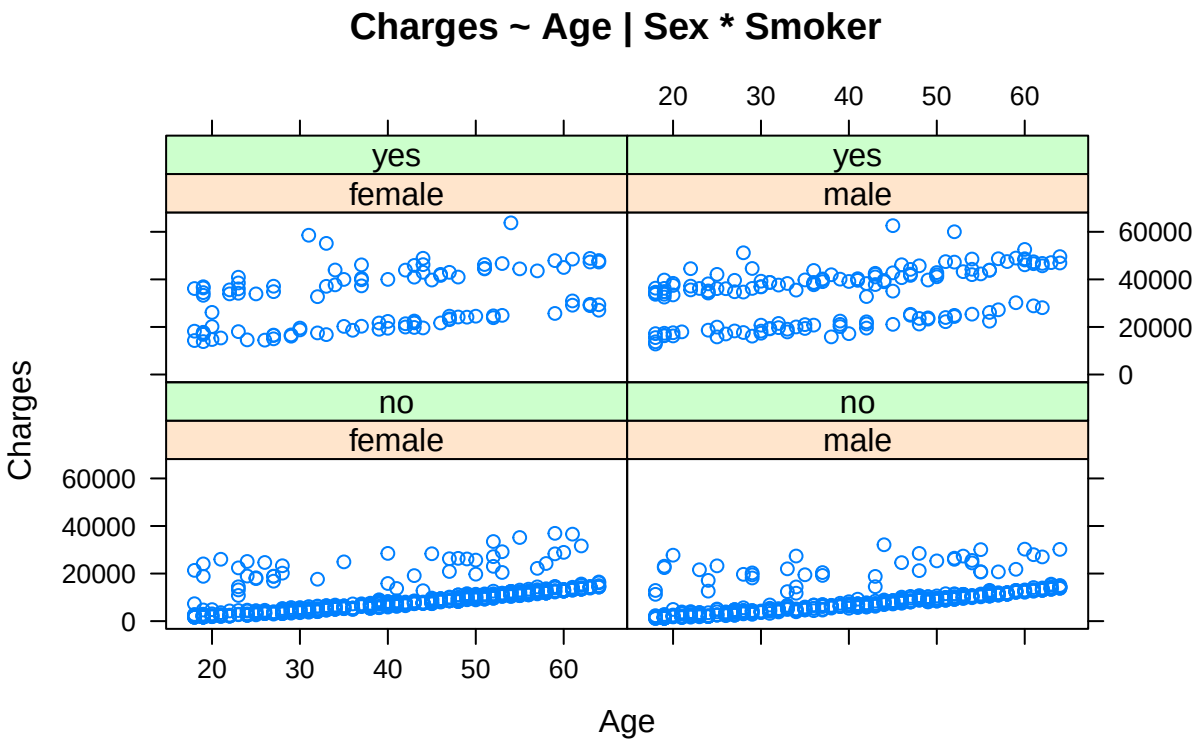


In the above plots, the group means are plotted in green. The unconditional plot already suggests that there are some differences between charges based on region. However we must also consider how region intercats with our other variables. In the second plot above, we see noticeable differences between the distribution of charges depending on whether the beneficiary is a smoker or not, as expected. We also see that per region, the difference between the mean charge for a smoker vs non-smoker is substantial. Notably in the southeast and southwest, this difference is considerably greater than in the northeast and northwest, in addition to the difference in the southeast being measurably greater than the difference in the southwest. This indicates

that not only should we include region as a factor in our model, but also that we should include relevant interaction terms with smoker, and perhaps bmi, since we already plan to include an interaction term for bmi and smoker.



The figure above reinforces our initial assumption that children has little to do with predicting charges, while the figure below suggests the same for sex. These visual assumptions are formally tested in the next section. While we could perform more EDA, such as plotting Charges ~ Age | Sex * Smoker and so on, we avoid doing so in the interest of concision.



3 Modelling

Our EDA has suggested that age, smoker, bmi, and region are variables of interest in constructing the linear model to predict charges, while sex and children do not appear so important. In this section, we quantiatively analyse these idea by examining adjusted R^2 values, experimenting with interaction terms based on statistical intuition and domain knowledge, and performing hypothesis tests regarding the values of coefficients and whether the corresponding feature should be part of our model. To have a full model to compare against, we have the following model

$$\text{model0} = \text{lm (charges regressed on ages * sex * bmi * children * smoker * region)}$$

where * refers to an interaction term between two variables and all relevant dummy variables, in accordance with the principle of marginality.

As mentioned earlier, the full model is not a practical consideration and only serves as a basis for comparison. While it has a high adjusted R^2 value (which isn't a given due to overparameterisation), 485 of the 825 estimated coefficients are not defined due to singularities. Additionally the residuals are not homoschedastic, violating the assumptions of a linear model. That being said, none of these issues affect model accuracy, and as such we compare other models to the full model. The tradeoff we're optimising for is coming up with a model with similar accuracy with greater interpretability while adhering to the assumptions of a linear model.

We iteratively build our model. Our EDA has already revealed that our model should include age, smoker, bmi, and region, and there should be an interaction term. Hence for our first nonreference model, we fit

$$\text{model1} = \text{lm (charges regressed on ages + bmi * smoker + region)}$$

. The model has a comparable adjusted R^2 value, and also has more homoschedastic residuals. Since it has only 8 parameters, of which 1 is an interaction term, it is far easier to interpret than the full model. As such, we have already found a suitable model, and now we only incrementally tweak it to see if it is overall more optimal. As mentioned earlier, the model may or not benefit from having an interaction term for age and smoker, the later of which already has an interaction term with bmi. To test this out, we fit

$$\text{model2} = \text{lm (charges regressed on ages * bmi * smoker + region)}$$

We notice almost no change in adjusted R^2 or homoschedasticity of residuals. The three added model parameters, which are all interaction terms involving age, having p-values greater than 0.05 when performing a two-sided t-test. Additionally, performing analysis of variance using the incremental sum of squares approach between model_1 and model_2 results in a p-value greater than 0.05. Since we have no reason to reject the null hypothesis that all the coefficients corresponding to the interaction terms involing age are 0, coupled with the fact that including these terms does not substantially benefit our model in terms of accuracy and or interpretability, we do not include these interaction terms in our final model.

We also earlier suspected that we might need an interaction term for region based on our earlier plots. On fitting

$$\text{model3} = \text{lm (charges regressed on ages * bmi * smoker * region)}$$

we see an increase in adjusted R^2 value while still preserving model interpretability and homoschedasticity of residuals. While earlier EDA had suggested that children was not a useful variable, I decided to include it and fit the following model:

$$\text{model4} = \text{lm (charges regressed on ages * bmi * smoker * region + children)}$$

As a reminder, in all these models, children is a categorical variable. Model4 ended up preserving interpretability (since we only added some dummy regressors) and residual homoschedasticity while also increasing adjusted R^2, so I decided to switch to using this model. I also fit

$$\text{model5} = \text{lm (charges regressed on ages * bmi * smoker * region * children)}$$

Model #	Model Formula	Adjusted R^2 Value	Total # of Estima
0	lm (charges ∼age* sex * bmi * children * smoker * region)	0.8479	283
1	lm(charges ∼age + bmi * smoker + region, filtered_data)	0.8338	8
2	lm(charges ∼age * bmi * smoker + region, filtered_data)	0.8339	11
3	lm(charges ∼age * bmi * smoker * region, filtered_data)	0.8364	19
4	lm(charges ∼age * bmi * smoker * region + children, filtered_data)	0.8386	24

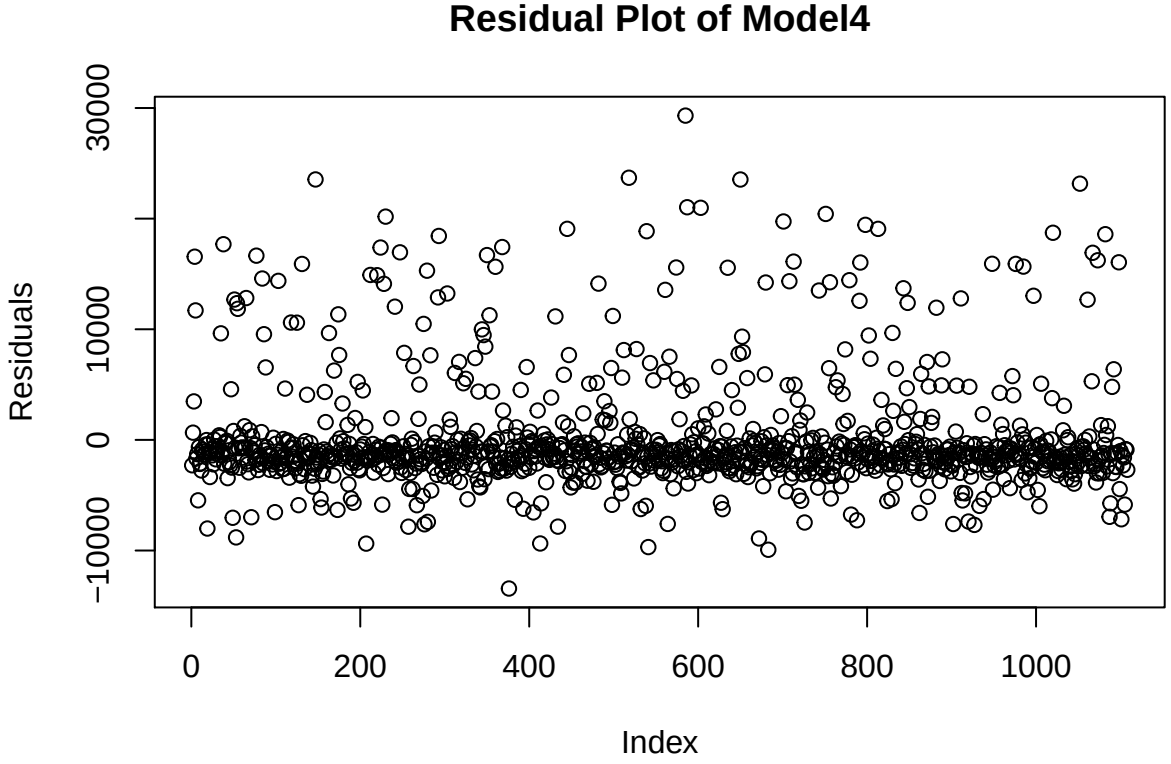
model6 = lm (charges regressed on ages * bmi * smoker * region + children + sex)

model7 = lm (charges regressed on ages * bmi * smoker * region + children * sex)

but I rejected model5 since it was no longer interpretable due to overparemeterisation, while model6 and model7 were rejected since the p-value when conducting the incremental sum of squares analysis of variance test was greater than 0.05 in both cases. The relevant results are summarised in the table below:

4 Model Interpretation and Discussion

Thus, our final model is model4. The residuals of the model are plotted below. We can see that broadly we meet the key assumptions of a linear model ie. the homoschedasticity of residuals, so we interpret our coefficients using the principle of a linear model. I'll interpret each of the 4 classes of coefficients, and provide an example of each.



1) Coefficient for the intercept: our estimate for the intercept is -4351.84, which means we expect on avaerage that when all the other covariates are set to 0, a beneficiary will cost the insurance company -$4351.84. This doesn't make practical sense and highlights that our model isn't perfect and is used best for predicting the charge when the beneficiary has covariates resembling that of other beneficiaries whose charges we already know. Fortunately we also never expect to be in the situation where all the covariates are 0, since a bmi of 0 is not biologically possible.

2) Coefficient for continuous covariate: an example of this age, for which our coefficient estimate is 255.75.

We expect that all else staying the same, an beneficiary who is a year older than another will cost the insurance company $255.75 for their healthcare.

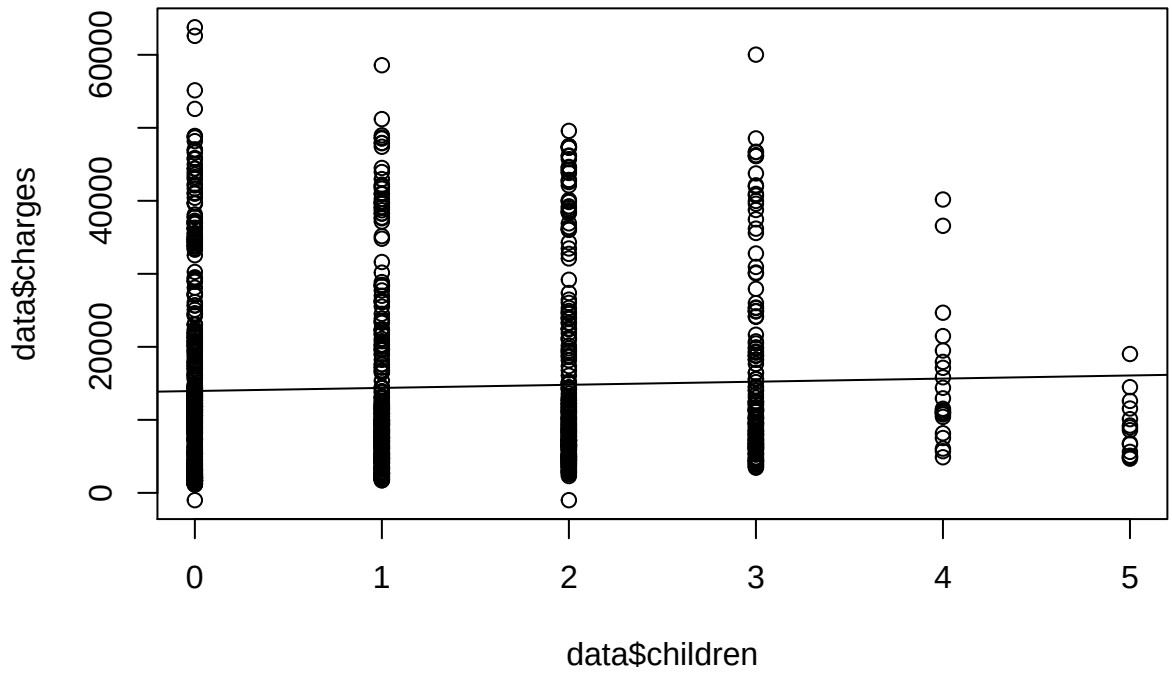
3) Coefficent for dummy variable: Note that the baseline for the dummy region variable is northeast, and the estimated coefficient for southeast is 5567.38. This means that all else staying the same, a beneficiary from the southeast will cost the insurance company $5567.38 more than if they were from the northeast. The other coefficient estimates for region are also with respect to the northeast.

4) Coefficient for interaction term: While dummy regressors can be thought of as parallel lines having different intercepts but the same slope for different values of a categorical variable, an interaction term means that the slope for the regression line is different. For example, our coefficient estimate for bmi is 113.11, but the estimate for the coefficient of the interaction term is bmi:smokeryes is 1422.19, which means that in additiond to whatever change in charge we expect to see for two individuals who differ in BMI with all else staying the same (excluding smoker), a smoker will cost on average the insurance company $1422.19 more than the non-smoker for the same change in BMI than a non-smoker.

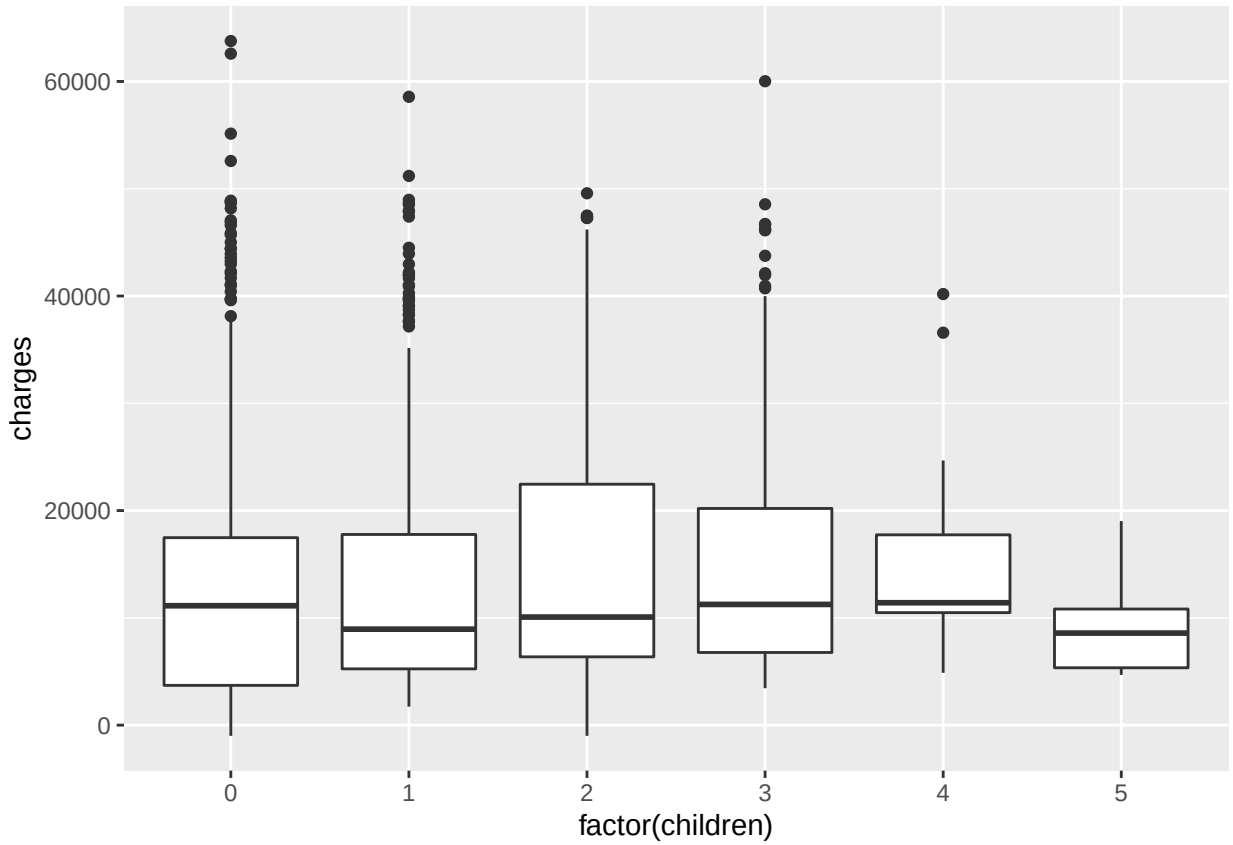
We now consider what impact the proposed wellness incentives program may have. Of the 6 covariates of interest, age, sex, region, and children can safely be expected not to be affected by this wellness program. Regarding bmi, the overall model can easily be used to predict what the expected charge to the insurance company a beneficiary would be. Note that this is not the same as causal inference because the model does not claim that a change in bmi will cause the cost to the company to decrease. Smoker is an interesting variable in this case, because even if a beneficiary stops smoking for a certain time period, there is no guarantee they will be designated non-smoker. This is indicative of a broader issue: since we don't know what time period the data are collected from, there may be a seasonal component that affects charges that could be modelled using a time series methodology that we are missing. However since we do not currently have this information, this is left as a possible future anaylsis should it become available. Returning to the point at hand, our model can easily predict the expected charge for an individual differing in bmi by applying the formula. In the case of a beneficiary stops smoking, the model may not be able to make an accurate prediction since not smoking will likely improve a beneficiary's health, but is not equivalent to never smoking. As a final note, we must consider that the interpretation of the coefficients of bmi and smoker is not as straightforward as that of intercept and age, since our model includes interaction terms for the former but not the latter. The interpretation of these coefficients is already provided above, and thus we conclude this analysis for the time being.

APPENDIX A: RELEVANT PLOTS

```
plot(data$children, data$charges)
abline(lm(charges ~ children, data = data))
```



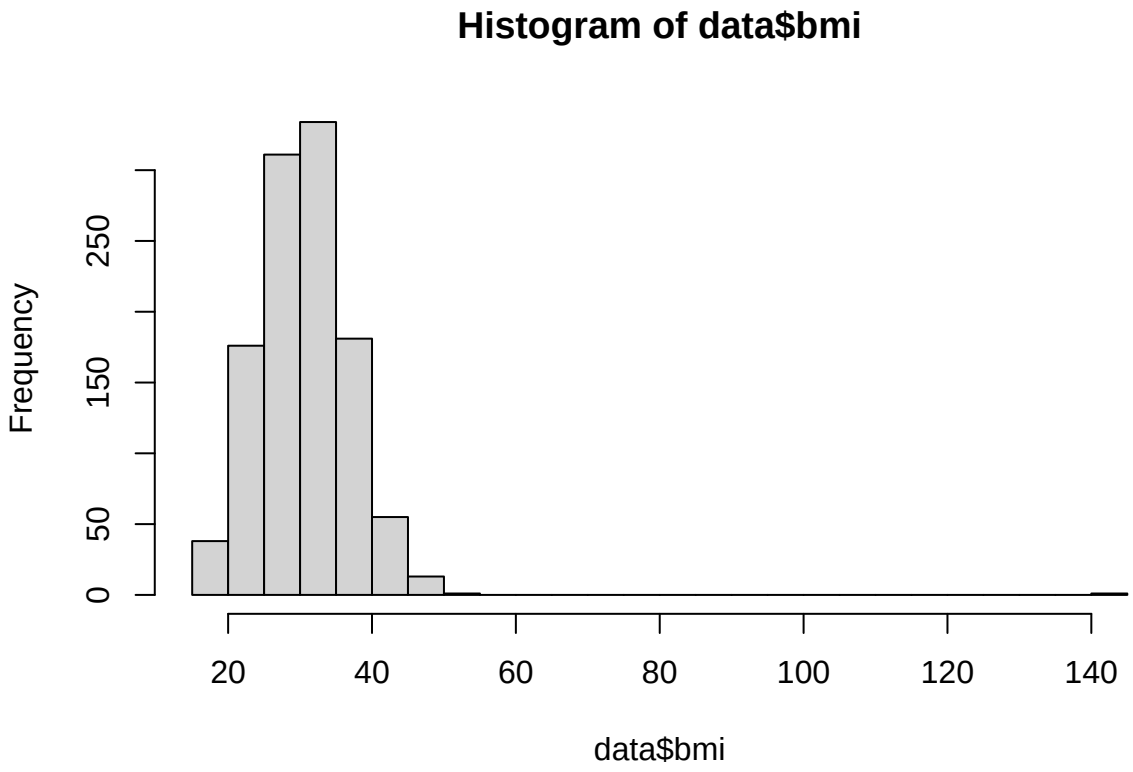

```
ggplot(data = data, aes(x = factor(children), y = charges)) + geom_boxplot()
```




```
max(data$bmi)
```

```
## [1] 143.02
```

```
hist(data$bmi, breaks = 20)
```


Histogram of data$bmi
Frequency
0
50
150
250
20
40
60
80
100
120
140
data$bmi

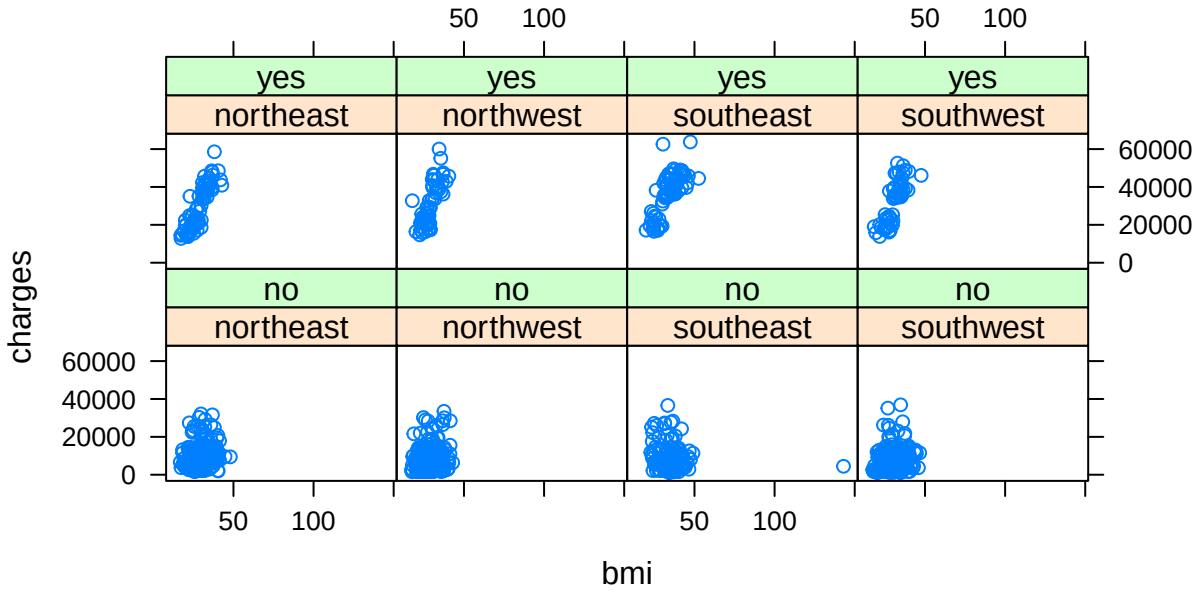
APPENDIX B: FULL R CODE

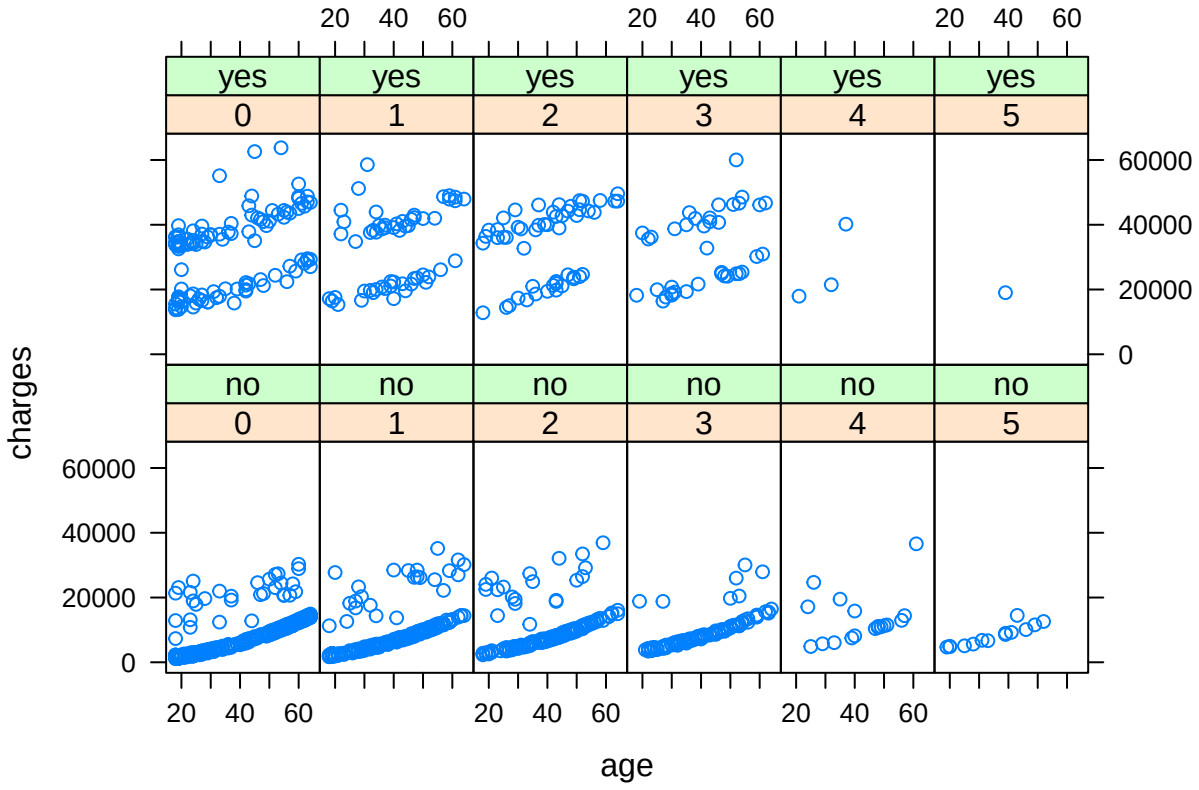
```
library(MASS)
library(gpairs)
library(ggplot2)
library(lattice)
library(scatterplot3d)
data = read.csv("MedicalInsurance.csv")
#stripchart(charges~factor(children),
#data=filtered_data,
#main="Charges for Different Number of Children on Insurance Plan",
#xlab="Number of Children",
#ylab="Charges",
#col="blue",
#vertical=TRUE,
#pch=16, cex = 0.5
#)
filtered_data = data[data$charges != -999, ]
age = filtered_data$age
sex = filtered_data$sex
bmi = filtered_data$bmi
children = filtered_data$children
smoker = filtered_data$smoker
region = filtered_data$region
charges = filtered_data$charges
high_bmi = bmi > 100
filtered_data$high_bmi = factor(high_bmi)
#unique(data$children)
plot(filtered_data$children, filtered_data$charges, col = "blue", cex = 0.5, xlab = "Number of Children"
abline(lm(charges ~ children, data = filtered_data), col = "red")
ggplot(data = filtered_data, aes(x = factor(children), y = charges)) + geom_boxplot(outlier.colour="blue
gpairs(filtered_data[, c("age", "bmi", "charges")])
xyplot(charges ~ age | smoker, type=c("p","r"), main = "Charges vs. Age for Non-smokers (no) and Smokers
without = lm(charges ~ age + smoker, filtered_data)
with = lm(charges ~ age * smoker, filtered_data)
anova(with, without)
xyplot(charges ~ bmi | smoker, type=c("p","r"), main = "Charges vs. BMI for Non-smokers (no) and Smokers
ggplot(filtered_data, aes(x = region, y = charges)) + geom_boxplot() + labs(x = "Region", y = "Charges"

ggplot(filtered_data, aes(x = region, y = charges)) + geom_boxplot(aes(fill = smoker)) + labs(x = "Regio

#ggplot(filtered_data, aes(x = smoker, y = charges)) + geom_boxplot() + labs(x = "Region", y = "Charges
xyplot(charges ~ bmi | factor(children) * smoker, filtered_data, xlab = "BMI", ylab = "Charges", main =
xyplot(charges ~ age | sex * smoker, filtered_data, xlab = "Age", ylab = "Charges", main = "Charges ~ Ag
model0 = lm (charges ~ age * sex * bmi * factor(children) * smoker * region, filtered_data)
#summary(model0)
#plot(model0$residuals)
model1 = lm(charges ~ age + bmi * smoker + region, filtered_data)
#summary(model1)
#plot(residuals(model1))
model2 = lm(charges ~ age * bmi * smoker + region, filtered_data)
#summary(model2)
#plot(residuals(model2))
```

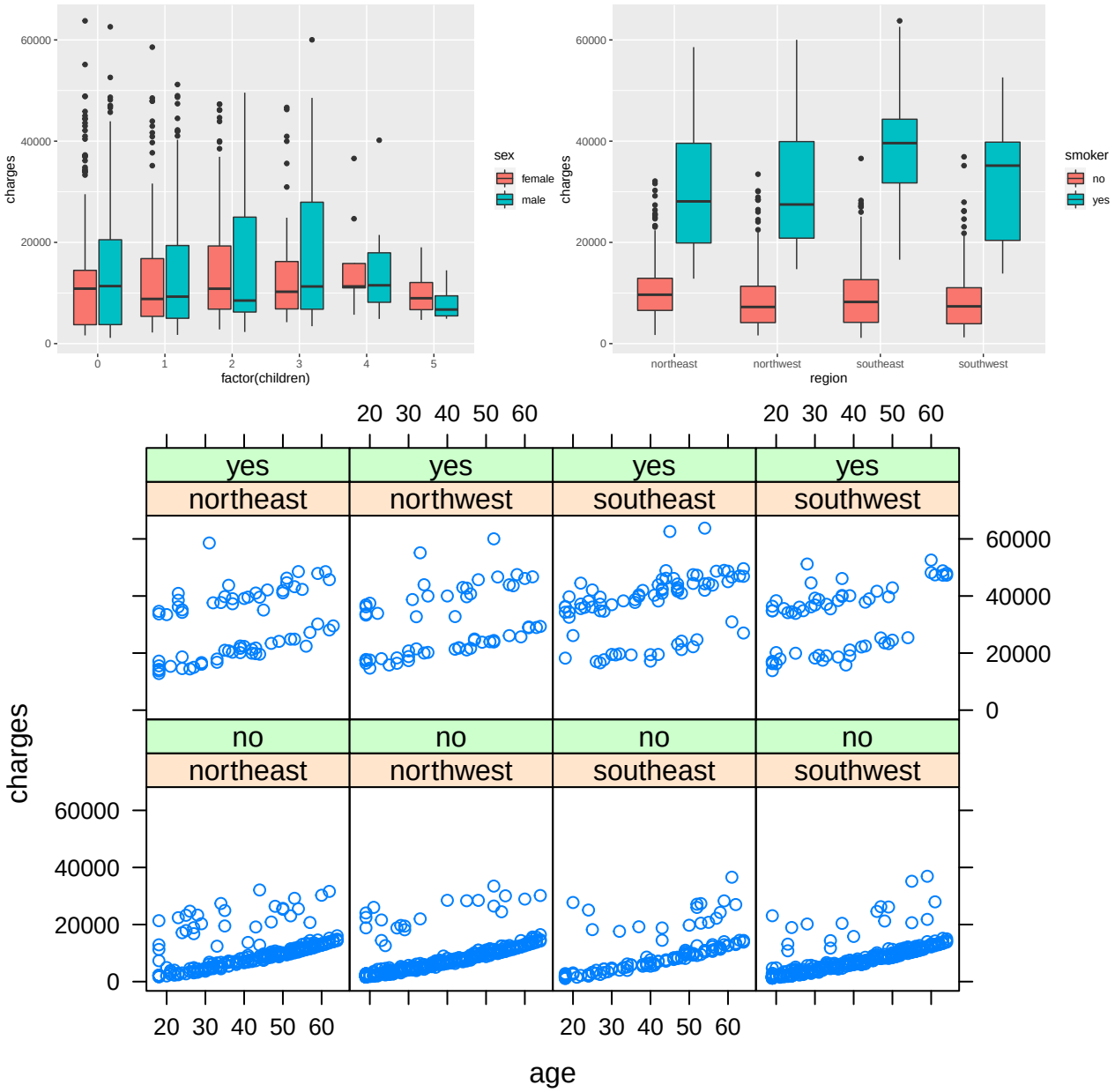
```r
#anova(model1, model2)
model3 = lm(charges ~ age + bmi * smoker * region, filtered_data)
summary(model3)
#plot(residuals(model3))
model4 = lm(charges ~ age + bmi * smoker * region + factor(children), filtered_data)
#anova(model3, model4)
#summary(model4)
model5 = lm(charges ~ age + bmi * smoker * region * factor(children), filtered_data)
#summary(model5)
model6 = lm(charges ~ age + bmi * smoker * region + factor(children) + sex, filtered_data)
model7= lm(charges ~ age + bmi * smoker * region + factor(children) * sex, filtered_data)
#anova(model4, model6)
#plot(residuals(model4))
#plot(residuals(model3))
#plot(residuals(model0))
#hist(age)
plot(residuals(model4), xlab = "Index", ylab = "Residuals", main = "Residual Plot of Model4")
plot(data$children, data$charges)
abline(lm(charges ~ children, data = data))
ggplot(data = data, aes(x = factor(children), y = charges)) + geom_boxplot()
max(data$bmi)
hist(data$bmi, breaks = 20)
labs = knitr::all_labels()
labs = labs[!labs %in% c("setup", "toc", "getlabels", "allcode")]

xyplot(charges ~ bmi | region * smoker, filtered_data)
yes_smoke = smoker == "yes"
no_smoke = filtered_data[!yes_smoke, ]
yes_smoke = filtered_data[yes_smoke, ]
yes_smoke_model = lm(charges ~ age, yes_smoke)
no_smoke_model = lm(charges ~ age, no_smoke)
par(mfrow=c(1,2))
plot(yes_smoke$age, yes_smoke$charges, xlab = "Age", ylab = "Charges", main = "Charges vs. Age for Smok
abline(yes_smoke_model)
plot(no_smoke$age, no_smoke$charges,  xlab = "Age", ylab = "Charges", main = "Charges vs. Age for Non-sm
abline(no_smoke_model)
paired_data = filtered_data
paired_data$sex = factor(sex)
paired_data$children = factor(children)
paired_data$region = factor(region)
paired_data$smoker = factor(smoker)
gpairs(paired_data[, c("age", "bmi", "charges")])
gpairs(paired_data[, c("region", "sex", "smoker", "children", "charges")])
cor(filtered_data[, c("age", "bmi", "charges")])
#out.width="50%"}
par(mfrow=c(1,2))
plot(age, charges)
xyplot(charges ~ age | factor(children) * smoker, filtered_data)
model_no_child = lm(charges ~age + smoker * region * factor(children), filtered_data)
model_child = lm(charges ~ age + smoker * region, filtered_data)
anova(model_child, model_no_child)
summary(model_no_child)
summary(model_child)
```

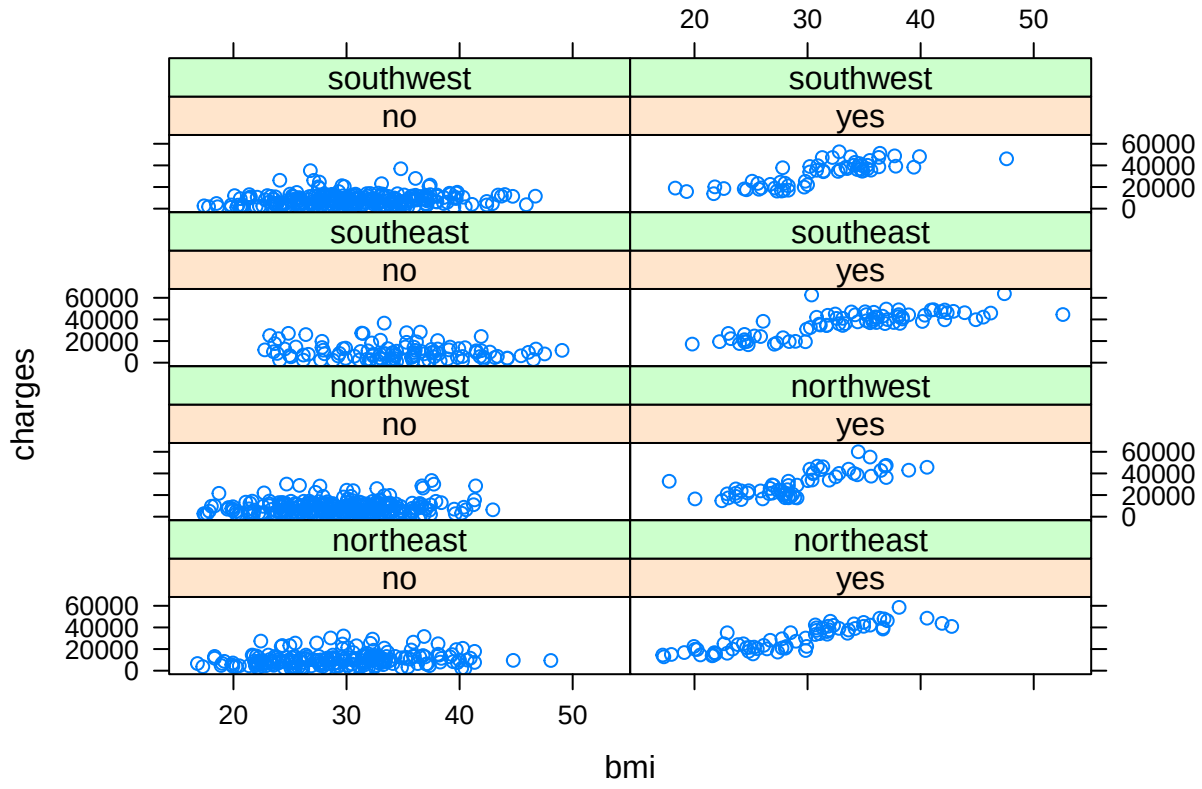
```
full_model = lm (charges ~ age* sex * bmi * children * smoker * region)
summary(full_model)
plot(full_model$residuals)
par(mar = c(4, 4, .1, .1))
ggplot(filtered_data, aes(x = factor(children), y = charges, fill = sex)) + geom_boxplot()
ggplot(filtered_data, aes(x = region, y = charges, fill = smoker)) + geom_boxplot()
xyplot(charges ~ age | region * smoker, filtered_data)
removed = filtered_data[!high_bmi, ]
xyplot(charges ~ bmi | smoker * region, removed)
scatterplot3d(x = age, y = bmi, z = charges, cex.symbols = 0.5, angle = 75)
full_model = lm(charges ~ age * bmi * factor(children) * sex * smoker * region *factor(high_bmi), filte

ggplot(filtered_data, aes(x=factor(children), y=charges)) + geom_boxplot(outlier.colour="red", outlier.

#http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization
#plot(filtered_data$children, filtered_data$charges)

model1 = lm(log(charges) ~ age * sex * bmi * children * smoker * region, filtered_data)
model2 = lm(log(charges) ~ age * sex * bmi * children * smoker + region, filtered_data)

anova(model1, model2)
paired_data = filtered_data
paired_data$sex = factor(sex)
paired_data$children = factor(children)
paired_data$region = factor(region)
paired_data$smoker = factor(smoker)
gpairs(paired_data[ , c("charges", "age", "bmi", "smoker")])

#barchart(factor(region))
yes_smoke = smoker == "yes"
no_smoke = filtered_data[!yes_smoke, ]
yes_smoke = filtered_data[yes_smoke, ]
boi = lm(charges ~ age + factor(smoker), filtered_data)


#plot(age, charges)
#lines(age, )
xyplot(charges ~ age | smoker, type=c("p","r"), main = "Charges vs. Age for Non-smokers (no) and Smokers

summary(lm(charges ~ age))
model1 = lm(charges ~ age + smoker + sex + smoker*sex, filtered_data)
model2 = lm(charges ~ age + smoker * sex, filtered_data)
model3 = lm(charges ~ age + smoker, filtered_data)
anova(model2, model1)
#summary(model1)
#summary(model2)
#summary(model3)
xyplot(charges~age|sex * factor(children), data = filtered_data)
filtered_data[filtered_data$bmi == max(bmi), ]
#https://en.wikipedia.org/wiki/Jon_Brower_Minnoch
#Something is off about max bmi
good = lm(charges ~ age + smoker * bmi + region)
good_boi = lm(charges ~ age + smoker * bmi * region + high_bmi)
```

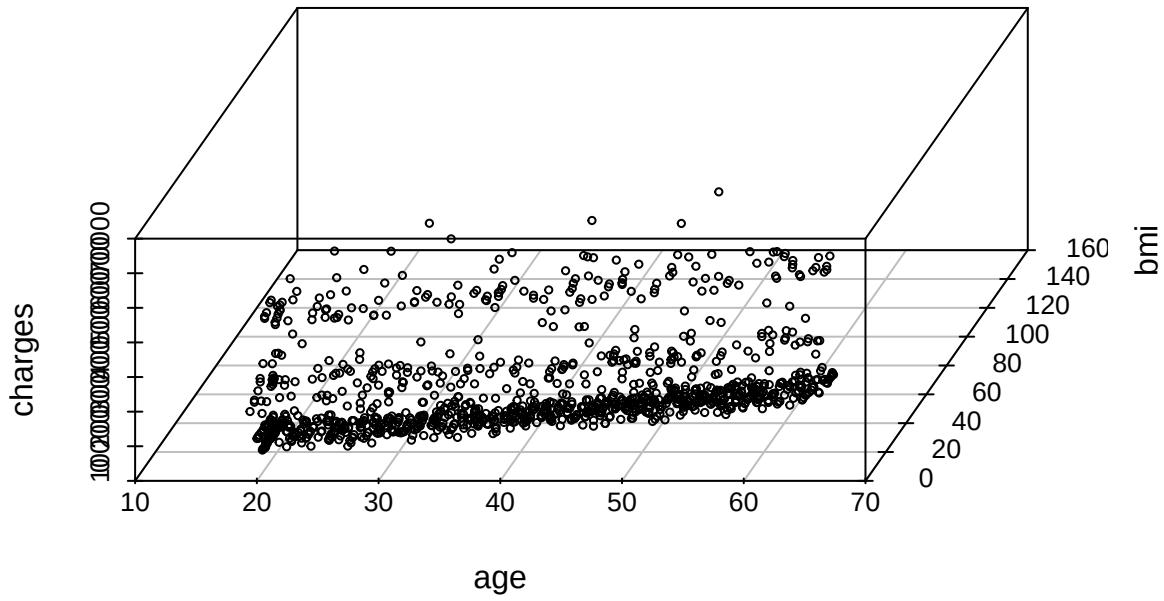
```
xyplot(charges ~ age | region * smoker, filtered_data)
xyplot(charges ~ age | region, filtered_data)
plot(age, charges)
abline(lm(charges ~ age))
current_boi = lm(charges ~ age + bmi * smoker + region)
qqnorm(current_boi$residuals, pch = 1, frame = FALSE)
qqline(current_boi$residuals, col = "steelblue", lwd = 2)
summary(model3)
```

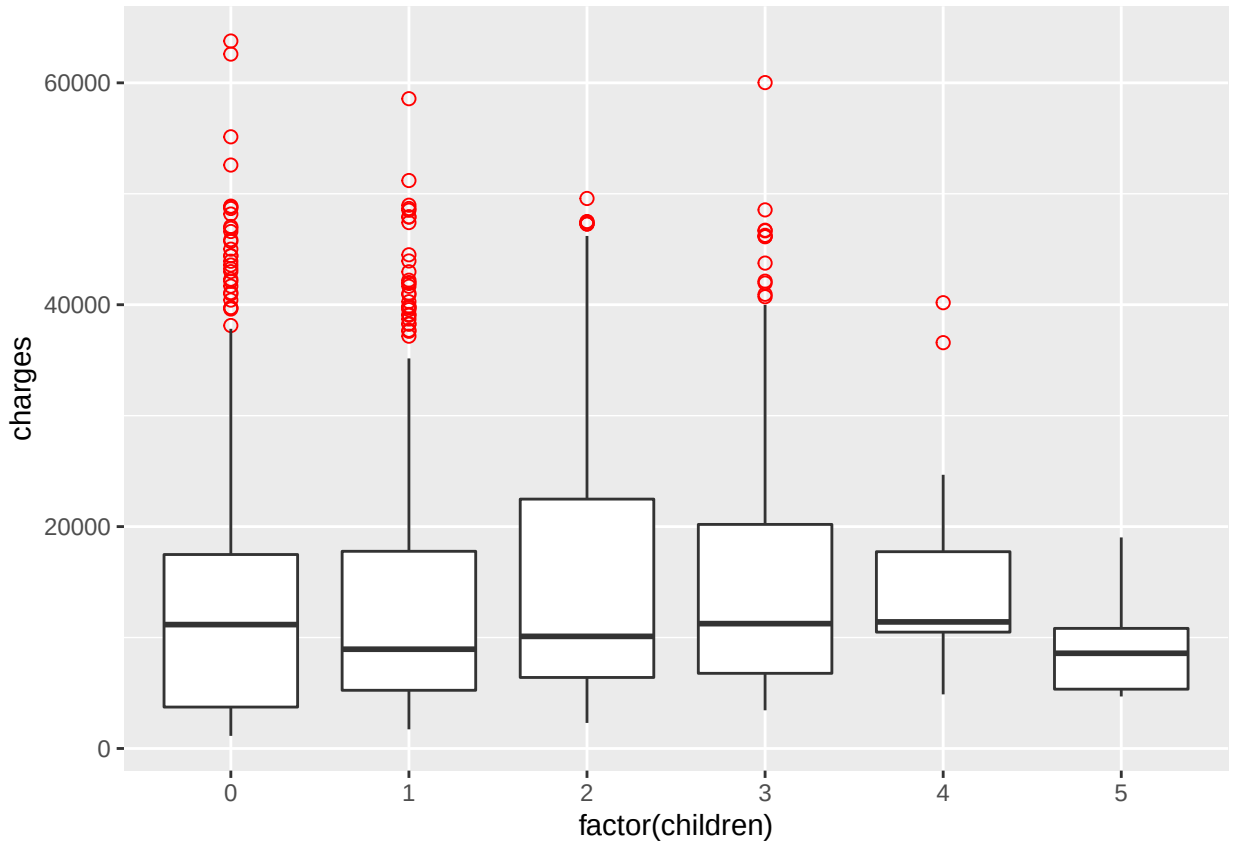




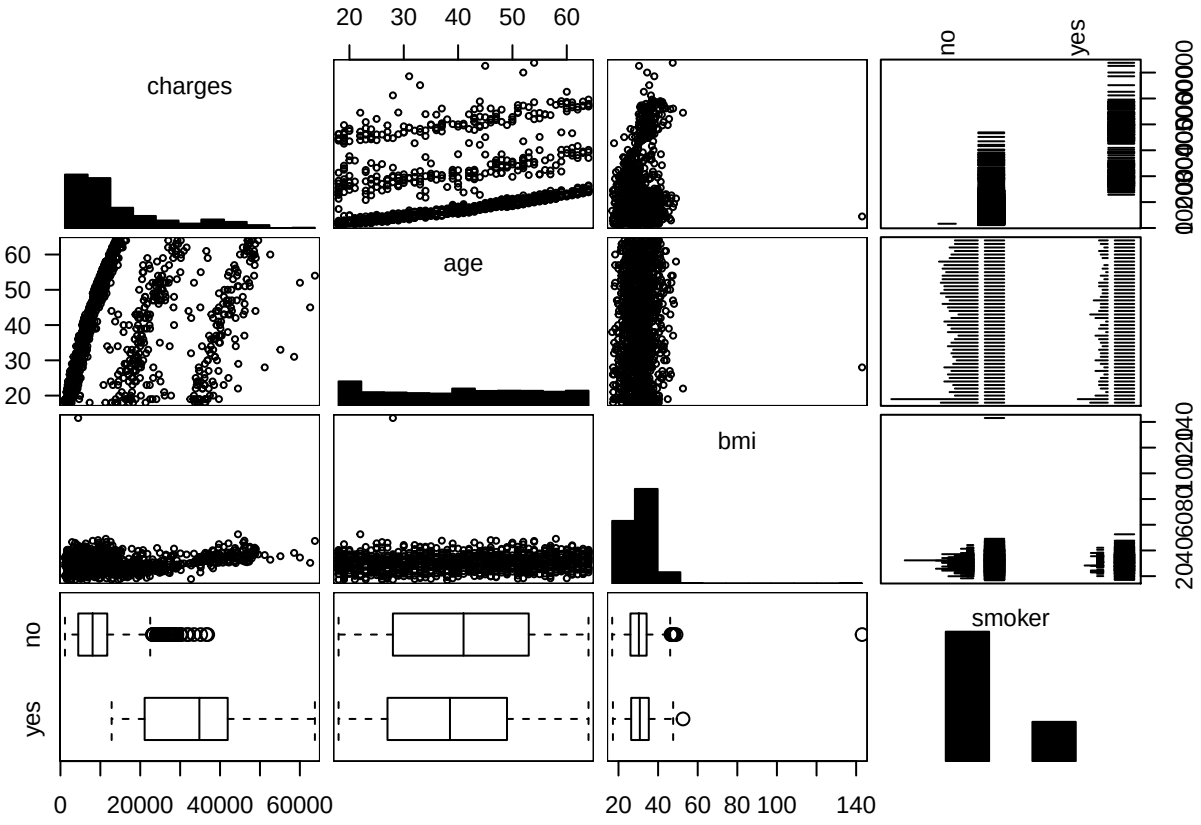

charges
60000
40000
20000
0
0
1
2
3
4
5
factor(children)
sex
female
male
charges
60000
40000
20000
0
northeast
northwest
southeast
southwest
region
smoker
no
yes
20 30 40 50 60
yes
yes
yes
yes
northeast
northwest
southeast
southwest
60000
40000
20000
0
charges
no
no
no
no
northeast
northwest
southeast
southwest
60000
40000
20000
0
20 30 40 50 60
20 30 40 50 60
age


southwest
no
southwest
yes
southeast
no
southeast
yes
northwest
no
northwest
yes
northeast
no
northeast
yes
charges
bmi
20
30
40
50
60000
40000
20000
0

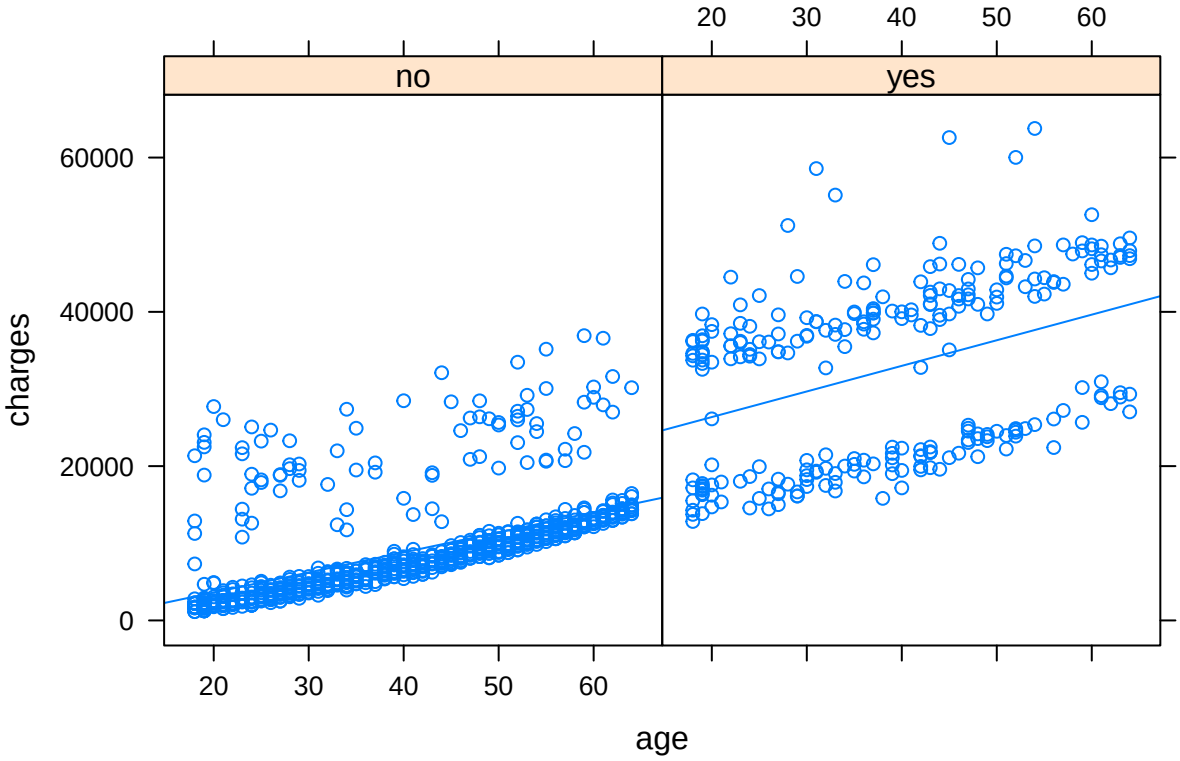

charges
age
bmi
10
20
30
40
50
60
70
0
20
40
60
80
100
120
140
160




```
## Analysis of Variance Table
##
## Model 1: log(charges) ~ age * sex * bmi * children * smoker * region
## Model 2: log(charges) ~ age * sex * bmi * children * smoker + region
##   Res.Df    RSS  Df Sum of Sq      F  Pr(>F)
## 1    980 139.18
## 2   1073 156.92 -93   -17.737 1.3429 0.02058 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



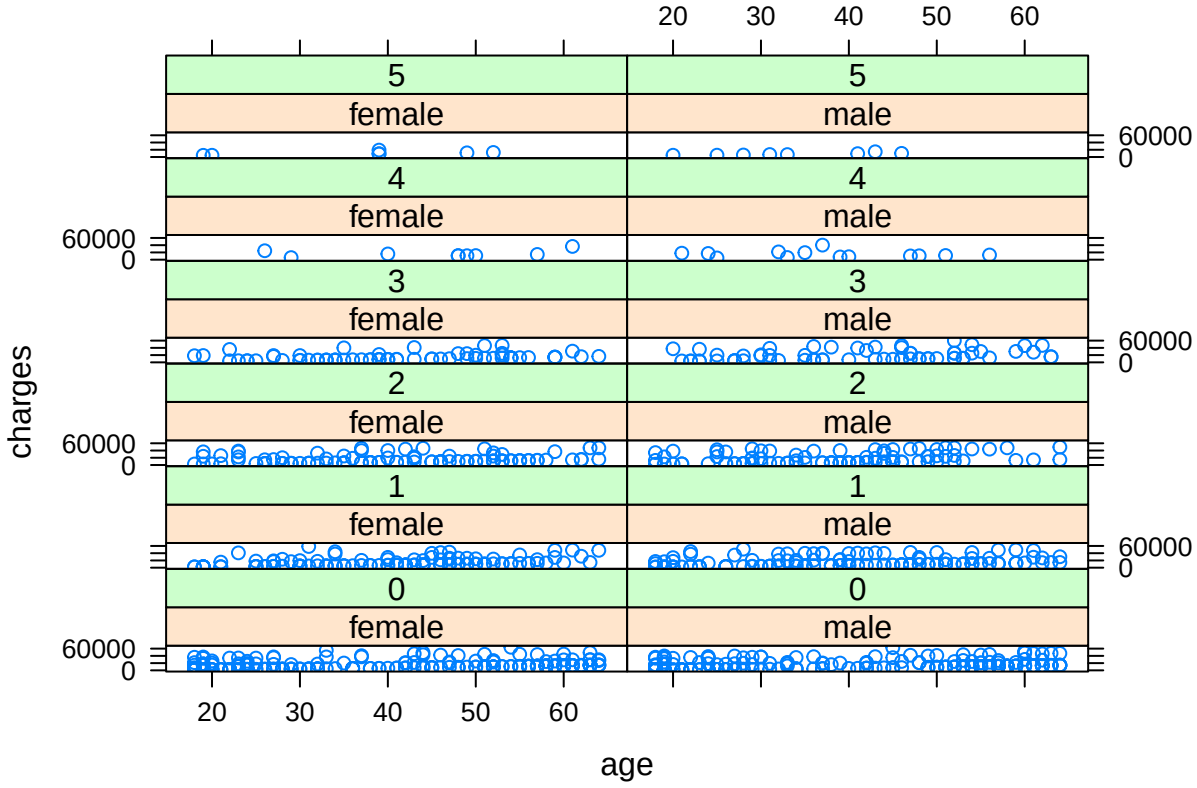
Charges vs. Age for Non−smokers (no) and Smokers (yes)



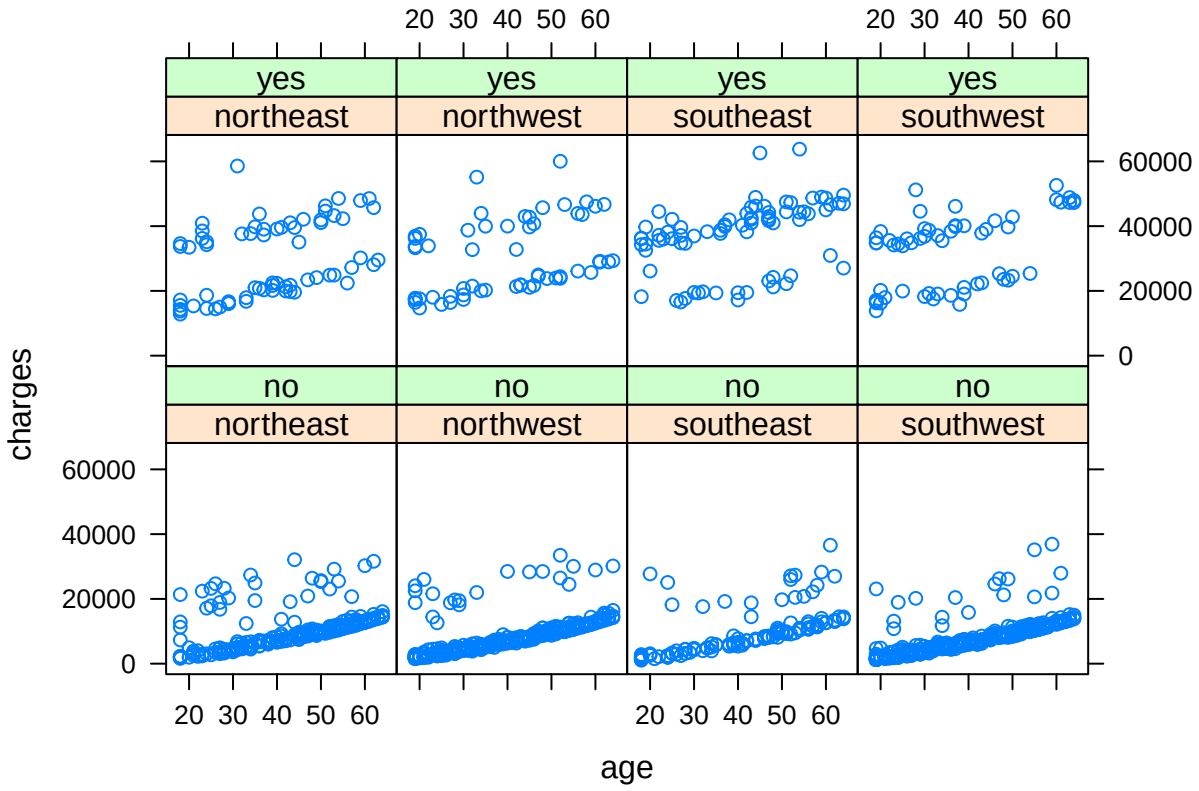
##

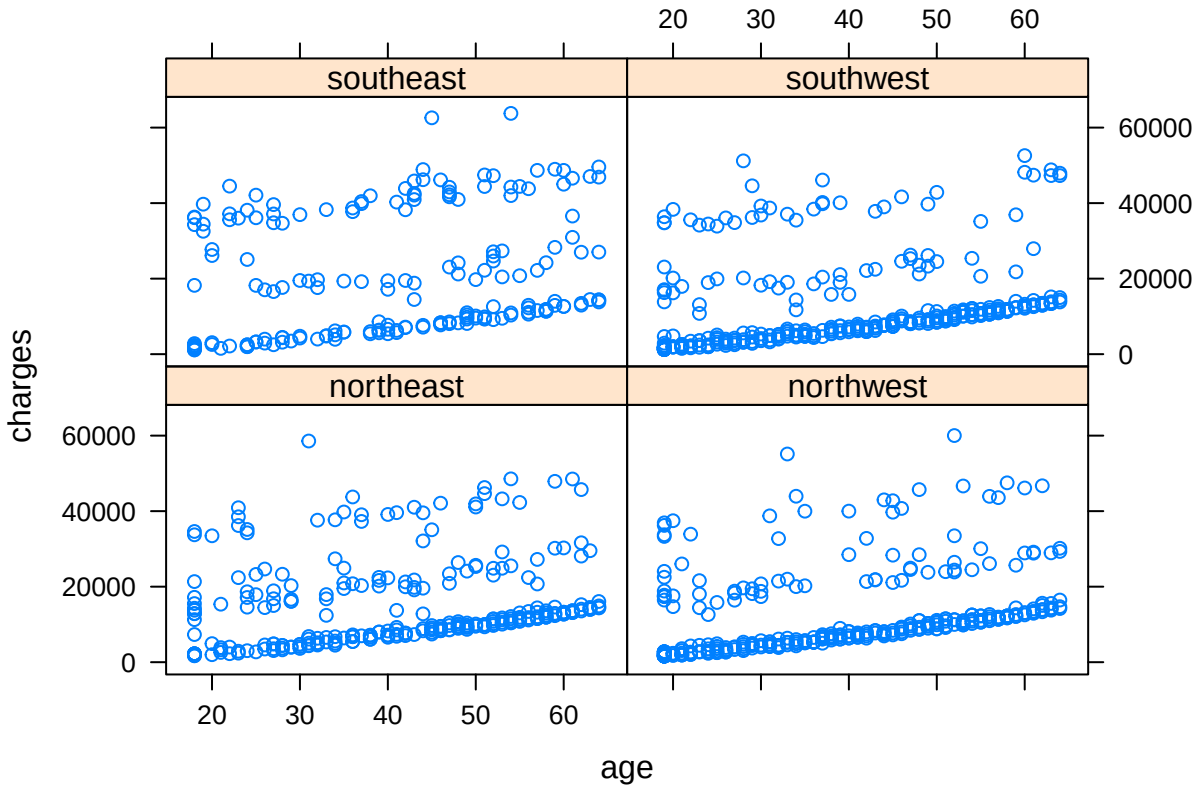
```
## Call:
## lm(formula = charges ~ age)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9115  -7646  -6590   6163  47015
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5235.75    1118.10   4.683 3.18e-06 ***
## age           229.83      26.33   8.730  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12200 on 1106 degrees of freedom
## Multiple R-squared:  0.06447,    Adjusted R-squared:  0.06362
## F-statistic: 76.21 on 1 and 1106 DF,  p-value: < 2.2e-16
```

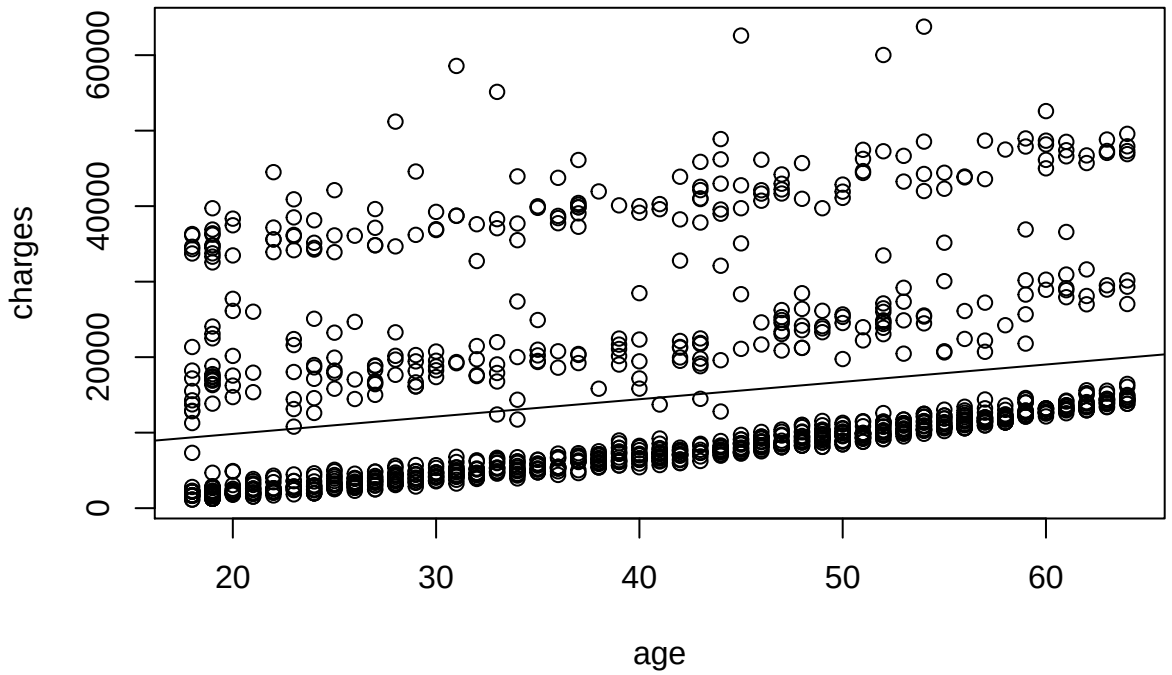
```
## Analysis of Variance Table
##
## Model 1: charges ~ age + smoker * sex
## Model 2: charges ~ age + smoker + sex + smoker * sex
##   Res.Df        RSS Df Sum of Sq F Pr(>F)
## 1   1103 4.9655e+10
## 2   1103 4.9655e+10  0         0
```

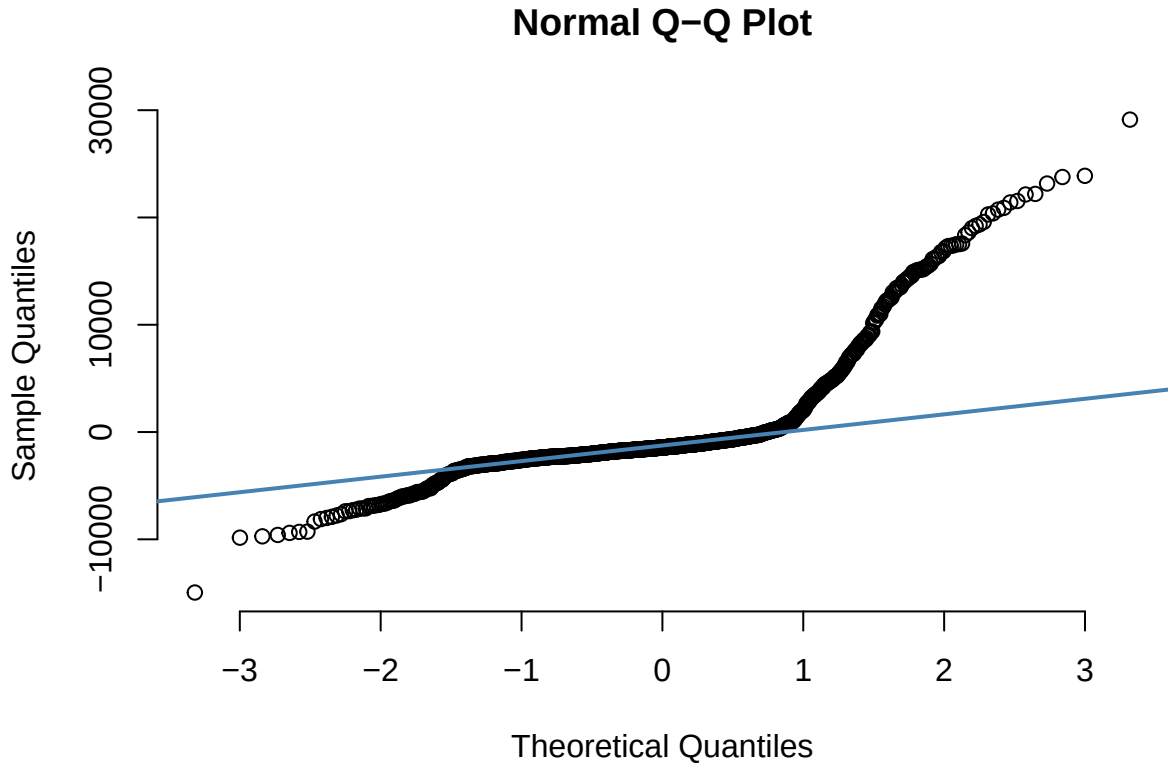



```
##     age  sex    bmi children smoker    region  charges high_bmi
## 974  28 male 143.02        3     no southeast 4454.732     TRUE
```


yes
yes
yes
yes
northeast
northwest
southeast
southwest
no
no
no
no
northeast
northwest
southeast
southwest
charges
age
0
20000
40000
60000
20
30
40
50
60


southeast
southwest
northeast
northwest
charges
age
0
20000
40000
60000
20
30
40
50
60






```
summary(model3)
```

```
##
## Call:
## lm(formula = charges ~ age + smoker, data = filtered_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -16543.0  -2236.6  -1462.8   -205.8  28292.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2258.42     634.83  -3.558  0.00039 ***
## age           276.71      14.56  19.006  < 2e-16 ***
## smokeryes   24107.20     479.64  50.261  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6734 on 1105 degrees of freedom
## Multiple R-squared:  0.7153, Adjusted R-squared:  0.7148
## F-statistic:  1388 on 2 and 1105 DF,  p-value: < 2.2e-16
```