Submitted for the Summer 2022 QEMP 8000

by Examinee J116

University of Georgia

May 17, 2023

Contents

List of Figures

List of Tables

Summary

In the medicare system, the healthcare cost and insurance cost are closely related for both insurance companies and insured individuals. Knowing the possible healthcare expense of individuals in advance will help the insurers and stack holders for better business plan and healthcare resource allocation. It also benefits patient since they can choose a better insurance plan according to their possible high or low medical expenses. Therefore, accurate prediction of the medical cost with related variables is important. In this report, the medical charge of individuals in the provided data is the response variable while other 6 variables: age, bmi, number of children, smoker (yes/no), sex (female/male), and residential region (NW,NE,SW,SE), are the explanatory variables. The data was divided into training and testing data in different ways, multiple linear regression models were built using the training data and compared. Model performance was evaluated based on the model R^2 values and goodness of fitting with the testing data. The age, bmi, smoker, children, and region variables were found to significantly affect the medical charges at $\alpha = 0.05$. The final model contains some main and 2 factor interaction effects as shown below. It has high prediction accuracy ($R^2 = 0.814$, 81.4% explained sample variance). Its prediction accuracy is constant through out the whole sample set as confirmed by the small range of square root of mean square residual (SMSR) for varying portions of data. The final model can be used to accurately predict medical cost, therefore, for insurers to determine the insurance cost of individuals. And the smoker variable should be considered as the the most important variable for the prediction, while the effects of age, children, bmi, region, and smoker*bmi interaction were also significant.

$$\hat{charges} = -12194 + 228 \times age + 747 \times bmi + 546 \times children + 10384 \times I_sn + 898 \times I_rne \quad (1)$$

$$-121 \times I_rnw - 327 \times I_rsw - 723 \times I_sn \times bmi$$

Where I_sn is not smoking, I_rne, I_rnw, and I_rsw are the indicators that the individual is from the Northeast,Northwest, or Southwest of the US.

1 Introduction

In health care system, accurately predicting medical cost of individuals can not only help insurers determine insurance cost but also guide patients for choosing appropriate insurance plans knowing their predicted future medicare expense. It is highly useful for properly advocating limited healthcare resources. Therefore, the objective of this study is to build accurate prediction models for healthcare cost of individuals with provided related variables. The effects of the six variables (age,sex, bmi, children, smoker, and region) on the medical cost were analyzed. Different multiple linear regression models were built and compared, the model accuracy was considered. Important variables for the prediction were also identified based on the estimates. It is found the smoker (yes/no) variable is the most important variable for predicting medicare cost, therefore influencing insurance cost. Other variables including bmi, age, children and region also significantly affect the medical cost at 95% confidence level. The interaction effect between smoker and bmi is significant at 95% confidence level. The final model has high prediction accuracy with $R^2 = 81.4\%$. The good fitting and consistency of model is confirmed by the training-testing process.

2 Exploratory data analysis

2.1 Data summary

The provided data includes 6 explanatory variables, 1 response variables, and 1338 observations. 3 of the explanatory variables are continuous, i.e., insured individual's age (age), body mass index (bmi), number of dependents (children), and medical cost ($) during the year billed to health insurance (charges). The rest 3 explanatory variables are categorical, i.e., whether the person smokes (smoker), the person's US residential area (region), and the gender (sex). In detail, the smoker variable has two levels: yes or no; region variable has shown 5 levels: Northeast (NE), Northwest (NW), Southeast (SE), Southwest (SW), and #1!a. the sex variable has two levels: female and male.

Table 1 shows the statistical summary for the continuous variables. The sample age ranges from 18-64 years old, average 18 years old. The sample bmi is in the range of 15.96-53.13 with mean at 30.66. It is known the ideal range of bmi is 18.5-24.9. The charges range for the individuals is 1122-63770 dollars with mean of 13270 dollars. In terms of children number, the sampled data shows the range of -2 to 5 children, which is unreasonable. 31 of the observations have negative children numbers, which are likely due to collection mistake. Therefore these numbers are changed to the corresponding positive numbers before visualizing the data.

Table 2 lists the counts of each level for the categorical variables. It is seen that there are much more smoker (1064) than non-smoker (274) samples in the data. The sample numbers are quite even at different levels of region and sex. There are only 24 observations showing region #1!a, it is unlikely to have a region called or labeled as #1!a. It is possible that this level is labeled mistakenly. Therefore, the region of these samples is remarked as missing data NA before plotting. More information about the distribution of sample age, bmi, charges, and children is provided in the appendix (Figures 3-6).

2.2 Data visualization and cleaning

After exploring plots with different variables from the data, some interesting patterns were found amongst the charges, age, bmi, and smoker variables. Figure 1 (left) shows the scatter plot of charges with age and the combined variable bmic from smoker and bmi. The bmic

Table 1: Summary for continuous variables

variable	age	bmi	children	charge
n	1184	1338	1338	1338
mean	18	30.66	1	13270
median	39	30.40	1	9382
min	18	15.96	-2	1122
max	64	53.13	5	63770
SD	14	6.10	1	12110

Table 2: Frequency count for categorical variables.

smoker	region	sex
1064 (Yes)	317(NE)	659 (female)
274 (No)	317 (NW)	666 (male)
-	349 (SE)	-
-	320 (SW)	-
-	24 (#!a)	-

variable is coded with fours levels: bmi in the ideal range & not smoke (inno), bmi in the ideal range & smokes (inyes), bmi out of the ideal range & not smoke (outno), bmi out of the ideal range & smokes (outyes). It is seen age tends to increase individual's medical cost in three linear trends. The insured individuals who smoke and have bmi out of the ideal range (black dots) tend to have the highest medical bills (the highest band), most sampled patients who do not smoke have relatively low medical bills (green and red dots in two lower bands). Both smoker and bmi variables may contribute to the charges leading to the median range of charges for people with ideal bmi and smoking habit (blue dots). Also other factors such as region and children may affect charges, which may contribute to the formation of the middle scatter band. From the Figure 1 (right), some interaction between smoker and bmi variables may be spotted for predicting charges. In non-smoking samples (red dots), there is no obvious trend with increasing bmi. While in smoking samples. A linear increasing trend can be seen for the charges with increasing bmi. More plots of charges with explanatory variables and plots between explanatory variables can be found in the appendix (Figures 7-10).

Before analyzing the data, data cleaning needs to be completed. As shown in Table 3, there are 11.5%, 1% and 2.6% of the data missing for the age, sex, and region variables respectively. More information about the numbers and distribution of missing data in the three variables can be found in the appendix (Figures 11-12) . To maximize the application of provided information, deleting data rows with missing entries is not recommended. Therefore, here, the age missing values were imputed using predictive mean matching (*pmm*) function in R MICE package, which imputes the missing data points with non-missing data points that has the nearest predicted age value to the predicted missing one using age and bmi as predictors with linear regression. The missing data for sex was filled using bmi with logistic regression since the sex variable shows a significant effect on the bmi value at 90% level (p-value=0.09). For imputation, female was coded as 1 and male was coded as 0. The parameter estimates for the logistic model are found to be 0.454 and -0.015. The effect of bmi is significant (p-value=0.09, AIC=1813) at $\alpha = 0.1$ The model is shown in equation (2). Then probability values for the missing sex were predicted using the model, entries with predicted probability ≥ 0.5 were assigned as "female" and entries with predicted probability < 0.5 were assigned as "male". Missing data in region was imputed with randomly generated data based on the sample probabilities. The sample probabilities/proportions for non-missing region data are 0.243, 0.243, 0.248, and 0.246 for

Table 3: Estimates for the locations .

Nemuro	46.98	Akita	16.35	Hachijojima	2.36	Sumoto	-1.54	Izuhara
Kushiro	43.41	Miyako	16.18	Fukui	1.68	Osaka	-1.69	Tokyo
Wakkanai	41.25	Yamagata	12.81	Mito	1.65	Hiroshima	-1.80	Shizuoka
Abashiri	37.15	Takayama	12.34	Utsunomiya	1.36	Hamada	-1.91	Saga
Kutchan	36.91	Sakata	12.01	Maizuru	1.18	Tokushima	-1.96	Nagasaki
Mombetsu	36.32	Shirakawa	11.94	Tsuruga	0.94	Gifu	-2.43	Fukuoka
Rumoi	36.30	Nagano	10.59	Toyooka	0.81	Tanegashima	-2.50	Miyazaki
Urakawa	35.87	Aikawa	10.07	Matsue	0.36	Matsuyama	-2.58	Uwajima
Muroran	34.30	Matsumoto	8.81	Choshi	0.12	Oita	-2.65	Kumamoto
Hiroo	33.63	Sendai	8.61	Yonago	-0.22	Shionomisaki	-2.73	Nobeoka
Asahikawa	32.23	Niigata	8.01	Tottori	-0.33	Nagoya	-2.74	Kochi
Iwamizawa	31.55	Wajima	6.64	Tateyama	-0.96	Hamamatsu	-2.92	Miyakojima
Obihiro	31.16	Takada	6.23	Okayama	-0.99	Oshima	-3.17	Ishigakijima
Sapporo	30.30	Fukushima	5.72	Kumagaya	-1.03	Fukue	-3.26	Naha
Esashi	29.11	Onahama	5.10	Tsu	-1.13	Miyakejima	-3.31	Minamidaitojim
Hakodate	28.62	Toyama	3.61	Kyoto	-1.26	Kagoshima	-3.38	Kumejima
Aomori	22.42	Iida	3.40	Kobe	-1.38	Yokohama	-3.41	Naze
Hachinohe	20.27	Saigo	3.10	Shimonoseki	-1.38	Kofu	-3.45	Funchatoge
Shinjo	19.12	Hikone	2.48	Nara	-1.40	Owase	-3.57	
Morioka	18.72	Kanazawa	2.38	Takamatsu	-1.41	Wakayama	-3.62	

Northeast, Northwest, Southeast, and Southwest, respectively.

$$log(\frac{p}{1-p}) = 0.454 - 0.015 * bmi \tag{2}$$

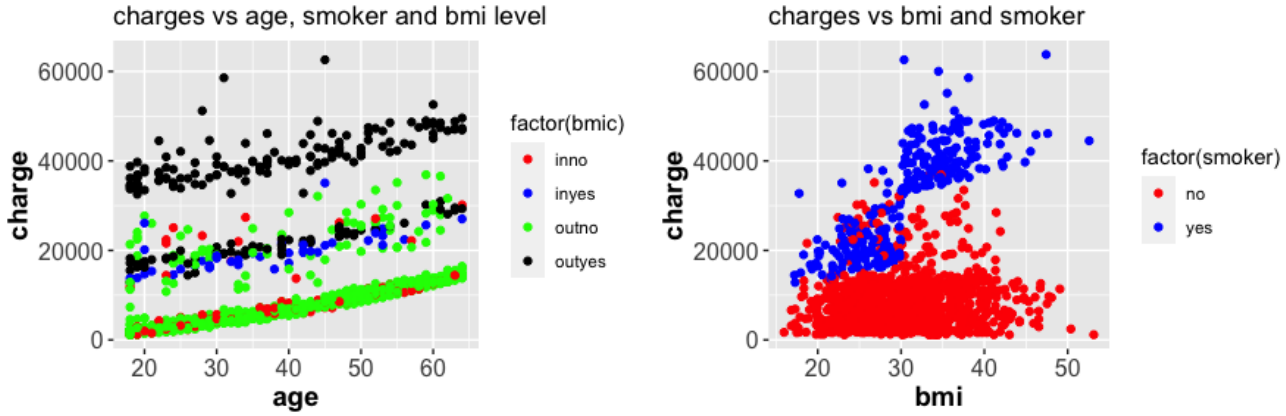

Figure 1: Plots of charges with age, bmi,and smoker.

3 Statistical methods

To build an accurate prediction model for the charges with provided variables, multiple regression methods are used in this analysis with both continuous and categorical variables. Main and interaction effects are considered. The data was modeled in three ways. First, the complete data was randomly divided into 80% training data to build the model and 20% testing data to evaluate the model. Model accuracy was evaluated based on the R^2 values and calculated

Table 4: Missing data percentages.

age	sex	bmi	children	smoker	region	charges
11.5	1.0	0	0	0	2.6	0

square root of mean square residual (SMSR). Residual plots were used to check the model assumptions. Second, the data was randomly divided into five folds for training (4 folds) and testing (1 fold) in turn. This specific train-testing application can confirm the model performance. The third model was built with only with main effects from all variables to study the importance/significance of the variables for predicting medical cost and influencing insurance cost. R studio software was used to analyze the data and generate plots.

4 Results and discussion

4.1 Data Modeling

After randomly dividing the data into 80% for training and 20% for testing, the training data was fitted into a model with all main effects and 2 factor interaction (2fi) effects. The hypotheses are shown below. It was found the model p-value < 0.001 indicating not all non-intercept parameters are 0 at 95% confidence level. Table 4 lists part of the Anova table showing all main factors and significant 2fi effect. It is seen that the main effects of age, bmi, and smoker, as well as the bmi*smoker 2fi effect are significant at $\alpha = 0.5$. The sex factor has the highest p-value (0.838) and lowest F value (0.04). To reduce the model, all insignificant 2fi terms and the sex factor were dropped from the next model. The model Anova result is shown in Table 5. All fitted main factors and 2fi effect are significant at $\alpha = 0.05$ with p-values < 0.001. The model sum of square error (SSE) is 2.85e+10, close to the one above (2.78e+10) obtained with much more parameters. The general form of this model is shown in equation (3). Table 6 shows the parameter estimates for this model. The R^2 for the model is 0.814 (81.4% of the sample variance explained by the model) indicating its high prediction accuracy. So, the final model can be written as equation (4). The 95% CI for the parameters can be found in Table 7. Due to the presence of bmi*smoker interaction in the model, the effects of bmi and smoker cannot be directly interpreted from their estimate CIs. However, an significant increasing of charges with higher ages, more children, or at Northeast region can be predicted at 95% confidence level from the positive CI ranges of their estimates. After fitting the data, the generated residual plots in Figure 2 are close to a random scattering pattern around 0, and an approximate straight line, respectively. This indicates the data meets the assumption of normality and randomization. Therefore, the regression model can be used for charge prediction.

Then, the performance of developed model was evaluated by fitting the testing data and predicting the charges using the explanatory variables from the testing data set. The square root of mean square residual (SMSR) between fitted charges and actual charges was calculated using equation (5). The calculated SMSR for this final model is 32239.3.

H_o: all parameters except the intercept are 0; H_a: Not all of non-intercept parameters are 0.

Table 5: Anova table for model with main and 2 factor interactions (only the significant 2fi is shown).

	Sum Sq	Df	F value	Pr(>F)
Intercept	5.56e+08	1	20.68	<0.001
age	3.44e+08	1	12.80	<0.001
sex	1.12e+06	1	0.04	0.838
bmi	1.98e+09	1	73.63	<0.001
children	6.13e+07	1	2.28	0.131
smoker (no)	2.26e+09	1	83.95	<0.001
region	5.47e+07	3	0.68	0.565
bmi:smoker	1.29e+10	1	480.98	< 0.001
...	...	...	...	...
Residuals	2.78e+10	1036	-	-

Table 6: Anova table for the final model.

	Sum Sq	Df	F value	P-value
Intercept	3.59e+09	1	133.28	<0.001
age	1.09e+10	1	402.78	<0.001
bmi	1.48e+10	1	548.89	<0.001
children	4.59e+08	1	17.05	<0.001
smoker	3.15e+09	1	116.86	<0.001
region	2.97e+08	3	3.68	0.012
bmi*smoker	1.47e+10	1	546.37	< 0.001
Residuals	2.85e+10	1060	-	-

$$charges = \beta_0 + \beta_1 \times age + \beta_2 \times bmi + \beta_3 \times children + \beta_4 \times I_sn + \beta_5 \times I_rne + \beta_6 \times I_rnw \quad (3)$$

$$+\beta_7 \times I_rsw + \beta_8 \times I_sn \times bmi + \epsilon$$

Where I_sn is not smoking,I_rne,I_rnw,and I_rsw are the indicators that the individual are from the NE,NW, SW of the US, and ϵ follows N(0,σ^2).

Table 7: Model summary for the final model.

	Estimate	Std. Error	t value	P-value
Intercept	-12194.05	1056.26	-11.55	<0.001
age	228.10	11.37	20.07	<0.001
bmi	746.62	31.87	23.43	<0.001
children	546.09	132.25	4.13	<0.001
smoker(no)	10383.89	960.60	10.81	<0.001
region(NE)	897.56	276.91	3.24	0.001
region(NW)	-120.56	281.90	-0.43	0.669
region(SW)	-327.48	283.17	-1.16	0.248
bmi*smoker(no)	-723.36	30.95	-23.37	<0.001

$$\hat{charges} = -12194 + 228 \times age + 747 \times bmi + 546 \times children + 10384 \times I_sn + 898 \times I_rne \quad (4)$$

$$-121 \times I_r nw - 327 \times I_r sw - 723 \times I_s n \times bmi$$

Where $I_s n$ is not smoking, $I_r ne$, $I_r nw$, and $I_r sw$ are the indicators that the individual are from the NE,NW, SW of the US.

Table 8: 95% CI for the model estimates.

	2.5%	97.5%
Intercept	-14266.7	-10121.4
age	205.8	250.4
bmi	684.1	809.2
children	286.6	805.6
smoker(no)	8499.0	12268.8
region1 (NE)	354.2	1441.0
region2 (NW)	-673.7	432.6
region(SW)	-883.1	228.2
bmi:smoker(no)	-784.1	-662.6

$$SMSR = \sqrt{\frac{\sum_1^n (\hat{y}_i - y_i)^2}{n}} \tag{5}$$

Where n is the testing sample size, $\hat{y}_i$ is the predicted charge, and y_i is the actual charge.

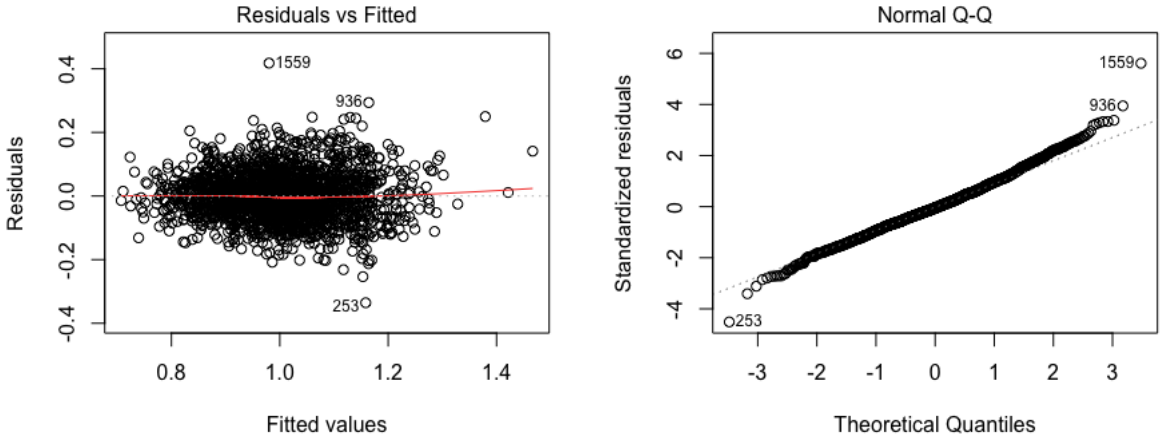

Figure 2: Residual and QQ plots.

4.2 Model accuracy

The model above shows high accuracy with high $R^2 = 0.814$ and low SMSR = 32239.3 through the training-testing process. Here, to check the model performance and consistency, the complete data was randomly divided into 5 folds, each fold of data set was used as the testing data to evaluate the model built using the rest of 4 folds of data. For all five tests, the same variables from equation (4) were found to be significant at $\alpha = 0.05$ and included in the models. The R^2 and SMSR values for the five models and model evaluation are reported in Table 8. The small ranges of both R^2 and SMSR values confirm the consistency of the model performance and the high R^2 values suggest the high model accuracy.

Table 9: Training and testing result.

test	R_2	prediction SMSR
fold1	0.802	33265.4
fold2	0.826	32118.3
fold3	0.813	31949.2
fold4	0.830	33548.7
fold5	0.834	33239.6

4.3 Compare variables

To compare the significance of the individual variables for predicting insurance cost, all main variables are fitted into the model without interaction effects. Table 9 show the model summary. It is found that smoker is the most important variable with much higher absolute estimate (11592.19) than those of other variables. The negative estimate indicates that individuals who does not smoke has much lower predicted medicare cost. In addition, the age and bmi variables are also significant with positive estimates, it means higher age or bmi leads to higher predicted medicare cost. Overall, the smoker variable should be considered as the most important factor for influencing insurance cost. The R^2 of this model is only 0.718 (< 0.814), and the SSE is 4.33e+10, higher than that of the model (4), it shows lower prediction accuracy than the final model (4).

Table 10: Model with only main variables for comparison.

	Estimate	Std. Error	t value	P-value
Intercept	-217.34	1136.83	-0.19	0.848
age	218.61	13.98	15.64	<0.001
bmi	365.59	33.71	10.84	<0.001
children	466.77	162.74	2.87	0.004
smoker(no)	-11598.29	242.65	-47.80	<0.001
region(NE)	744.55	340.67	2.19	0.029
region(NW)	46.62	346.84	0.13	0.893
region(SE)	-485.10	348.38	-1.39	0.164
sex(female)	-121.62	195.82	0.62	0.635

5 Conclusion

In this report, multiple linear regression and training-testing procedures were used to build prediction models for medical charges of insured individuals with provided variables (age, bmi, children number, smoker, sex, region). The sex variable was found to be insignificant at $\alpha = 0.05$, while all other variables and 2fi effect of smoker*bmi were found to be significant for predicting the medical charges at $\alpha = 0.05$. The final model is presented below in equation (6). It has high accuracy (R^2=0.814) and its accuracy is consistent through the whole sample. The smoker variable appears to be the most important variable for predicting medicare cost and affecting insurance costs. Individuals who smoke tend to have higher medical cost. Other variables including age, bmi, region, and children number are also significant. Individuals from Northeast with more children and at higher age tend to have increasing predicted healthcare charges.

$$\hat{charges} = -12194 + 228 \times age + 747 \times bmi + 546 \times children + 10384 \times I_sn + 898 \times I_rne \quad (6)$$

$$-121 \times I_r nw - 327 \times I_r sw - 723 \times I_s n \times bmi$$

Where $I_s n$ is not smoking, $I_r ne$, $I_r nw$, and $I_r sw$ are the indicators that the individual are from the NE,NW, SW of the US.

Reference

1. https://datascienceplus.com/imputing-missing-data-with-r-mice-package/ 2. https://statisticsbyjim.co in-regression-analysis/ 3. https://www.statology.org/variance-inflation-factor-r/

Appendix 1-Figures

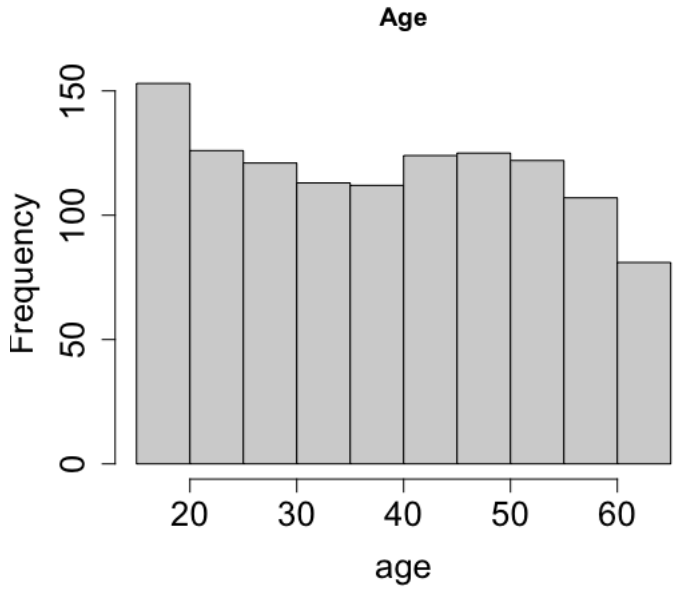

Figure 3: Distribution of age.

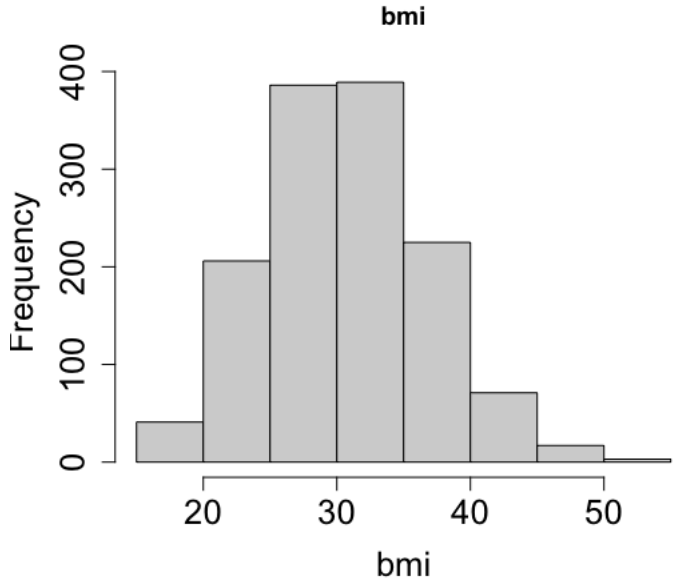

Figure 4: Distribution of bmi.

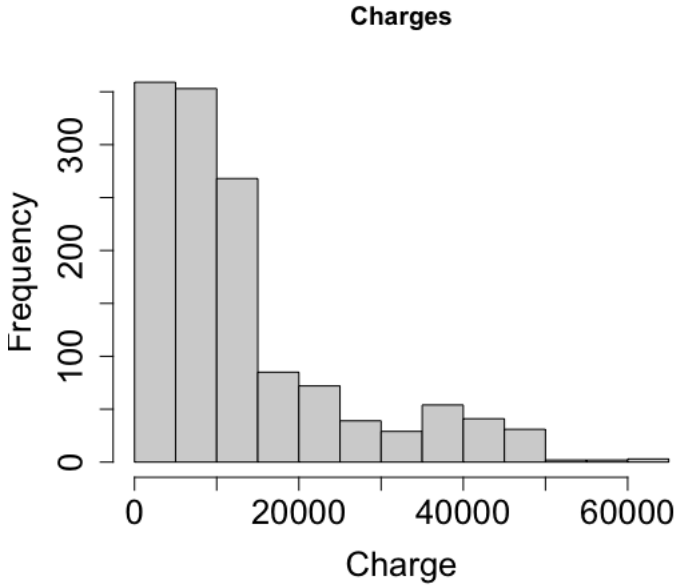

Figure 5: Distribution of charges.

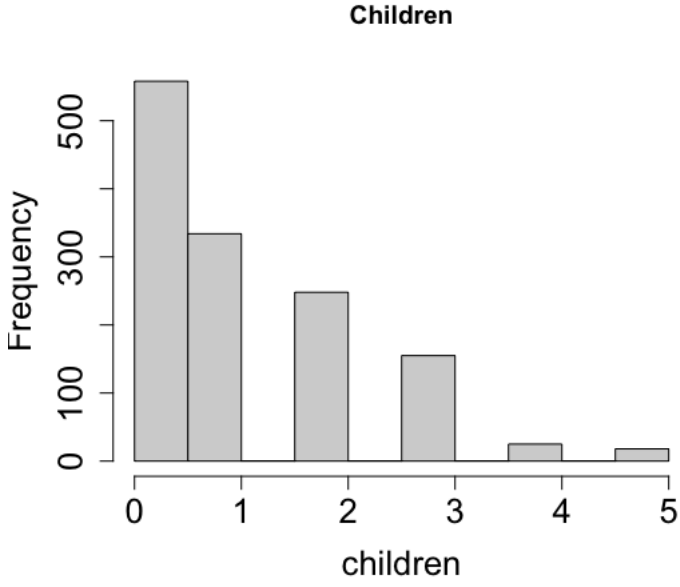

Figure 6: Distribution of children numbers.

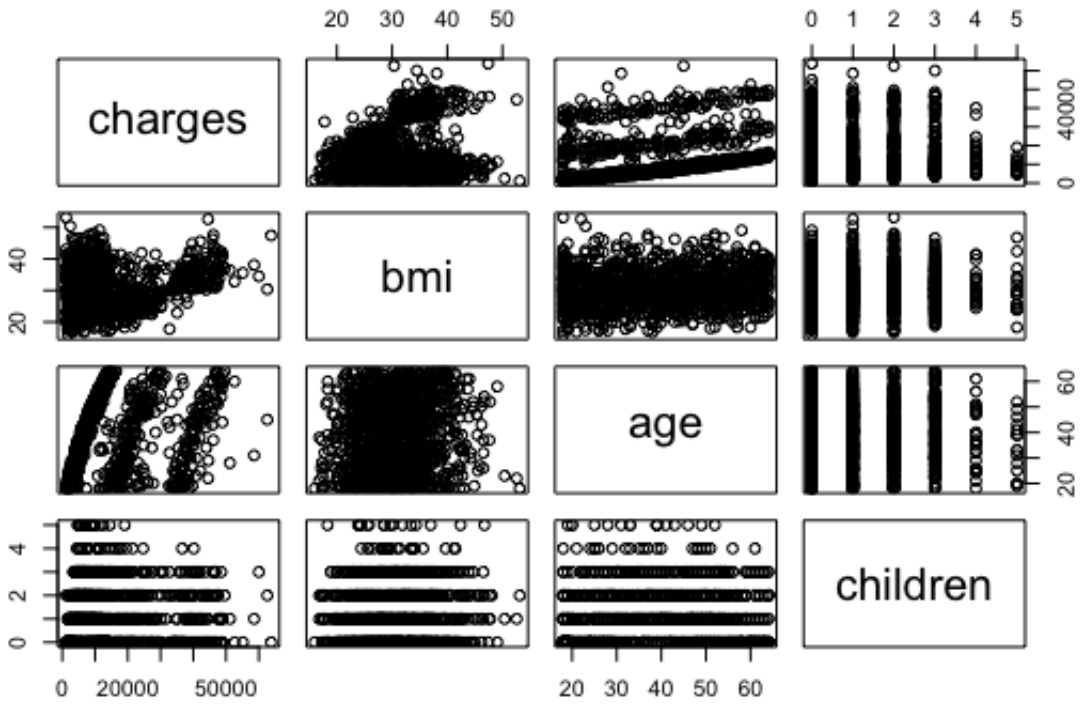

Figure 7: Matrix plots of the numerical variables.

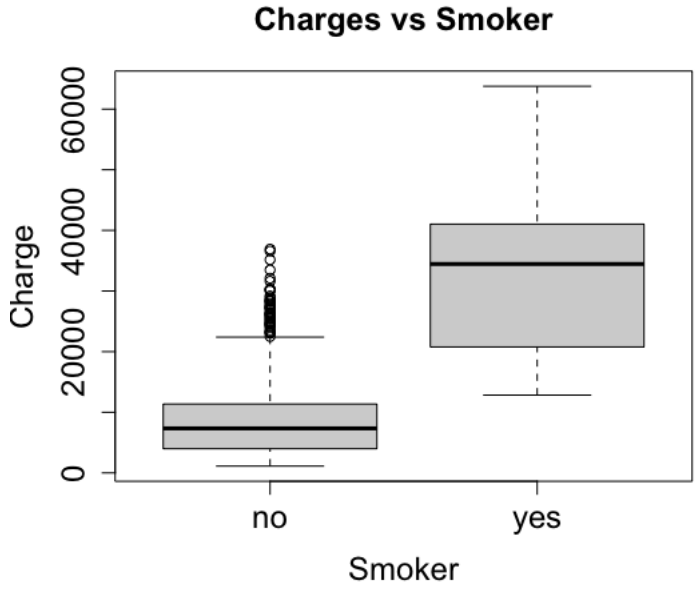

Figure 8: Plot of charges with smoker variable.

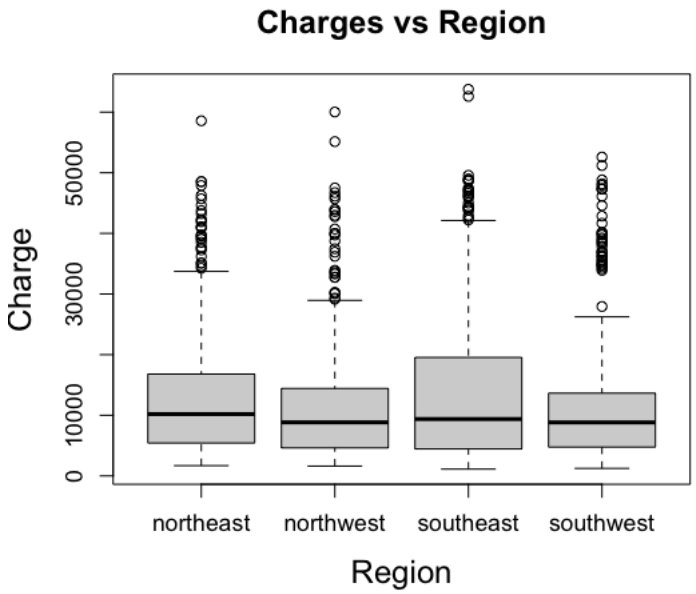

Figure 9: Plot of charges with region variable.

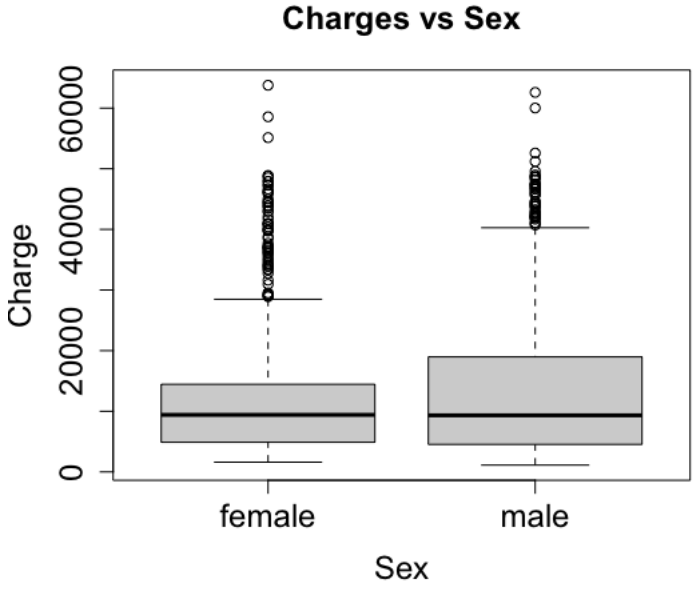

Figure 10: Plot of charges with sex variable.

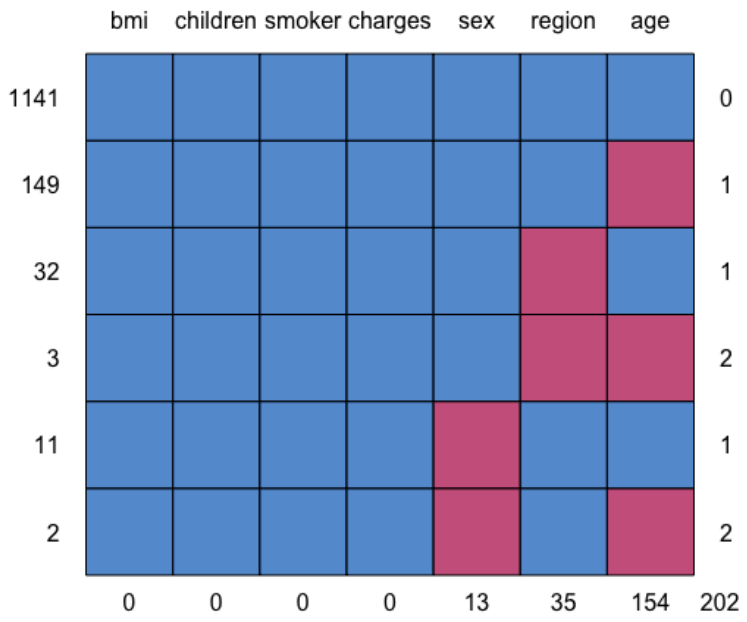

Figure 11: Numbers of missing data.

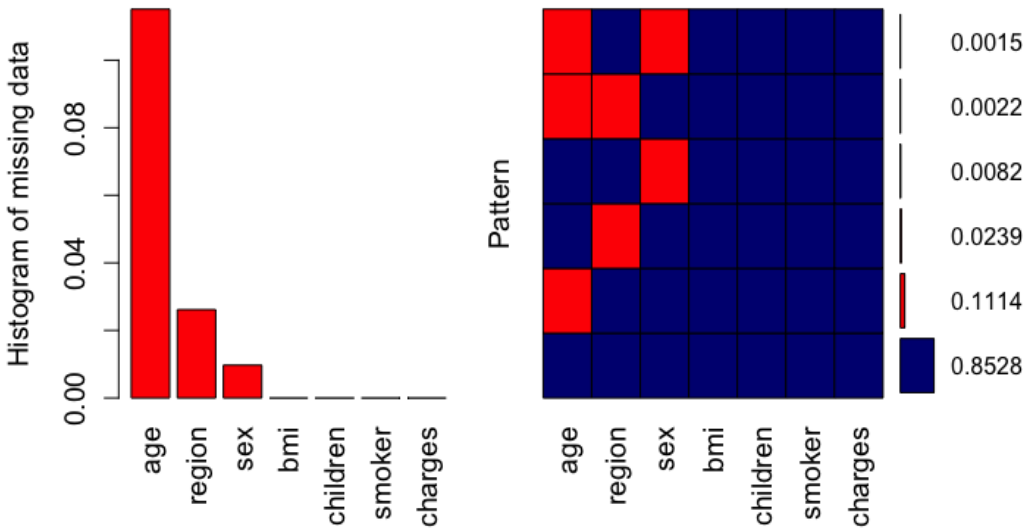

Figure 12: Percentages and distribution of the missing data

Appendix 2-R Codes

```
setwd("~/Desktop/STAT classes/STAT 6365E Programming/Working Directory")
library(tidyverse)
library(mice)
library(plyr)
insur<-read_csv("insurance.csv")
age<-insur$age
sex<-insur$sex
bmi<-insur$bmi
children<-insur$children
smoker<-insur$smoker
region<-insur$region
charges<-insur$charges

#data summary
summary(insur)
sd(insur$age,na.rm=TRUE)
sd(insur$bmi,na.rm=TRUE)
sd(insur$children,na.rm=TRUE)
sd(insur$charges,na.rm=TRUE)

table(insur$smoker)
table(insur$region)
table(insur$sex)

#children: negetive to positive
range(children)# consider negative as a mistake when putting the data;
#Change to positive, Could be predicted with a model.
length(children[is.na(children)]) # no missing
length(children[children%%1!=0])# check any non-integers, no non-integer
length(children[children<0])
length(children[children<0])/length(children)*100  ##### 2.3% data is negetive

location<-which(children<0,arr.ind=TRUE)
children[location] #show negative ages
insur$children[location]<-children[location]*(-1) ##### replace with positive values
length(insur$children[insur$children<0])

# change k#!a to NA
levels(factor(region)) #replaced with randomized level with their proportions
ct<-table(region);ct   # 24 with "k#!a"
ct[['k#!a']]/length(region)*100   #1.8% is k#!a low percent, consider NA
length(region[is.na(region)])
length(region[is.na(region)])/length(region)*100   ###### missing 0.8%
which(is.na(region),arr.ind=TRUE)
length(region[region=="k#!a"| region =="NA"]) # k#!a includes NA somehow
insur$region<-replace(insur$region, region=="k#!a",NA)
#Consider and replace k#!a as NA.
length(insur$region[is.na(insur$region)])
```

```
pMiss <- function(x){sum(is.na(x))/length(x)*100}
apply(insur,2,pMiss)  # missing %
md.pattern(insur)
library(VIM)
aggr(insur, col=c('navyblue','red'),
numbers=TRUE, sortVars=TRUE,
     labels=names(insur),
     cex.axis=1.2, gap=3,
     ylab=c("Histogram of missing data","Pattern"))
#data visualization
age<-insur$age
sex<-insur$sex
bmi<-insur$bmi
children<-insur$children
smoker<-insur$smoker
region<-insur$region
charges<-insur$charges
x<-cbind(charges,bmi,age,children)
pairs(x)
cor(x)

bmismok<-replace(bmi, (bmi>24.9 | bmi<18.5)&smoker=="yes","outyes")
bmismok2<-replace(bmismok,(bmi>24.9 | bmi<18.5)&smoker=="no","outno")
bmismok3<-replace(bmismok2,(bmi<=24.9 & bmi>=18.5)&smoker=="yes","inyes")
bmismok4<-replace(bmismok3,(bmi<=24.9 & bmi>=18.5)&smoker=="no","inno")
insurc<-mutate(insur,bmic=bmismok4)

sp<-ggplot(data=insurc) +
  geom_point(aes(x=age, y=charges, group=1,color=factor(bmic)))+
  ylab("charge")+xlab("age")+ggtitle("charges vs age, smoker and bmi level")
sp+scale_color_manual(values=c("red", "blue", "green","black"))
+theme(axis.text=element_text(size=12),
axis.title=element_text(size=14,face="bold"))

sp<-ggplot(data=insurc) +
  geom_point(aes(x=bmi, y=charges, group=1,color=factor(smoker)))+
  ylab("charge")+xlab("bmi")+ggtitle("charges vs bmi and smoker")
sp+scale_color_manual(values=c("red", "blue"))+theme(axis.text=element_text(size=12),
       axis.title=element_text(size=14,face="bold"))
###
hist(insurc$age,xlab="age",cex.axis=1.6,cex.lab=1.6,main="Age")
hist(insurc$bmi,xlab="bmi",cex.axis=1.6,cex.lab=1.6,main="bmi")
hist(insurc$children,xlab="children",cex.axis=1.6,cex.lab=1.6,main="Children")
hist(insurc$charges,xlab="Charge",cex.axis=1.6,cex.lab=1.6,main="Charges")

barplot(table(insurc$smoker))
barplot(table(insurc$region))
barplot(table(insurc$sex))
plot(insurc$charges~factor(insurc$smoker),ylab="Charge",xlab="Smoker",
```

```
     cex.axis=1.5,cex.lab=1.5,cex.main=1.5,main="Charges vs Smoker")
plot(insurc$charges~factor(insurc$region),ylab="Charge",xlab="Region",
     cex.axis=1.1,cex.lab=1.5,cex.main=1.5,main="Charges vs Region")
plot(insurc$charges~factor(insurc$sex),ylab="Charge",xlab="Sex",
     cex.axis=1.5,cex.lab=1.5,cex.main=1.5,main="Charges vs Sex")
###
#impute age by mice ppm
insur2<-insur[,-7];head(insur2)
tempData <- mice(insur2,m=5,maxit=50,meth='pmm',seed=500)
summary(tempData) #using bmi and children
tempData$imp$age
tempData$meth
completedData <- complete(tempData,2)
insur3<-cbind(completedData,insur[,7])
head(insur3)

##imput sex by bmi logistic
g=lm(insur3$bmi~factor(insur3$sex))
summary(g)
anova(g)
sex1<-replace(insur3$sex,insur3$sex=='female',1)
sex2<-as.numeric(replace(sex1,sex=='male',0))
head(sex2)

g1  =  glm(sex2~insur3$bmi,  family=binomial)
summary(g1)
x<-insur3$bmi[which(is.na(insur3$sex),arr.ind=TRUE)]
ilogit = function(x){ exp(x)/(1+exp(x)) }
predp = ilogit(g1$coef[1]+g1$coef[2]*x);predp
predp2<-replace(predp,predp>=0.5,"female")
predp3<-replace(predp2,predp<0.5,"male")
insur3$sex[which(is.na(insur3$sex),arr.ind=TRUE)]<-predp3

# replace region with randomly generated data based on sample probability
ct<-table(insur3$region);ct
cne<-ct[[1]]; cne
cnw<-ct[[2]]; cnw
cse<-ct[[3]]; cse
csw<-ct[[4]]; csw
tt<-cne+cnw+cse+csw;tt
pne<-cne/tt;pne
pnw<-cnw/tt;pnw
pse<-cse/tt;pse
psw<-csw/tt;psw
set.seed(123)
rsample<-sample(c("northeast", "northwest", "southeast","southwest"), size = 35,
replace = TRUE, prob = c(pne, pnw, pse,psw));
rsample
insur3$region[is.na(insur3$region)]<-rsample
length(insur3$region[is.na(insur3$region)])
```

```
pMiss <- function(x){sum(is.na(x))/length(x)*100}
apply(insur3,2,pMiss)
dim(insur3)
1338/10*8
train<-sample(1:1338,1070,replace=FALSE) #80%
trains<-insur3[train,] #size 1070
tests<-insur3[-train,] #size 268

age<-trains$age
sex<-trains$sex
bmi<-trains$bmi
children<-trains$children
smoker<-trains$smoker
region<-trains$region
charges<-trains$charges

g2<-lm(trains$charges~(trains$age+trains$sex
+trains$bmi+trains$children+trains$smoker
+trains$region)^2)
summary(g2)
Anova(g2,type=3)
#Final model building
g3<-lm(trains$charges~trains$age+trains$bmi
+trains$children+trains$smoker+trains$region
+trains$sex+trains$bmi*trains$smoker)
summary(g3)
confint(g3)
Anova(g3,type=3)
plot(g3, 1:2)
#To understand the effects
g4<-lm(trains$charges~trains$age+trains$bmi
+trains$children+trains$smoker+trains$region
+trains$sex)
summary(g4)
anova(g4)
confint(g4)

testsx<-insur3[-train,-7]
head(testsx)
pred<-predict(g3,testsx, se.fit = FALSE)
sqrt(sum((pred-tests[,7])^2)/268)  #32242.16

#divide into 5 folds
1338/10*2
268*4-1338

fold1<-sample(1:1338,268,replace=FALSE) #80%
fold1s<-insur3[fold1,] #size 268
fold1sx<-insur3[fold1,-7] #size 268
```

```
rest1<-insur3[-fold1,]

dim(rest1)
fold2<-sample(1:1070,268,replace=FALSE) #80%
fold2s<-rest1[fold2,] #size 268
fold2sx<-rest1[fold2,-7] #size 268
rest2<-rest1[-fold2,]
dim(rest2)
fold3<-sample(1:802,268,replace=FALSE) #80%
fold3s<-rest2[fold3,] #size 268
fold3sx<-rest2[fold3,-7] #size 268
rest3<-rest2[-fold3,]
dim(rest3)
fold4<-sample(1:534,268,replace=FALSE) #80%
fold4s<-rest3[fold4,] #size 268
fold4sx<-rest3[fold4,-7] #size 268

fold5s<-rest3[-fold4,] # size 266
fold5sx<-rest3[-fold4,-7] # size 266
dim(fold5s)

pred1<-predict(g31,fold1sx, se.fit = FALSE)
sqrt(sum((pred1-fold1s[,7])^2)/268)
trainf1<-rbind(fold5s,fold2s,fold3s,fold4s)
trainf2<-rbind(fold1s,fold5s,fold3s,fold4s)
trainf3<-rbind(fold1s,fold2s,fold5s,fold4s)
trainf4<-rbind(fold1s,fold2s,fold3s,fold5s)
trainf5<-rbind(fold1s,fold2s,fold3s,fold4s)
dim(trainf1) #1070
dim(trainf2) #1070
dim(trainf3) #1070
dim(trainf4) #1070
dim(trainf5) #1072

g21<-lm(trainf1$charges~(trainf1$age+trainf1$sex
+trainf1$bmi+trainf1$children+trainf1$smoker
+trainf1$region)^2)
summary(g21)
Anova(g21,type=3)
g31<-lm(trainf1$charges~(trainf1$age+trainf1$bmi
+trainf1$children+trainf1$smoker+trainf1$region
+trainf1$smoker*trainf1$bmi))
summary(g31)
Anova(g31,type=3)
plot(g31, 1:2)
pred<-predict(g31,fold1sx, se.fit = FALSE)
sqrt(sum((pred-fold1s[,7])^2)/268)  #33239.63

g22<-lm(trainf2$charges~(trainf2$age+trainf2$sex
+trainf2$bmi+trainf2$children+trainf2$smoker
```

```
+trainf2$region)^2)
summary(g22)
Anova(g22,type=3)
g32<-lm(trainf2$charges~(trainf2$age+trainf2$bmi
+trainf2$children+trainf2$smoker+trainf2$region
+trainf2$smoker*trainf2$bmi))
summary(g32)
Anova(g32,type=3)
plot(g32, 1:2)
pred<-predict(g32,fold2sx, se.fit = FALSE)
sqrt(sum((pred-fold2s[,7])^2)/268)  #33265.44

g23<-lm(trainf3$charges~(trainf3$age+trainf3$sex
+trainf3$bmi+trainf3$children+trainf3$smoker
+trainf3$region)^2)
summary(g22)
Anova(g22,type=3)
g33<-lm(trainf3$charges~(trainf3$age+trainf3$bmi
+trainf3$smoker+trainf3$children+trainf3$region
+trainf3$smoker*trainf3$bmi))
summary(g33)
Anova(g33,type=3)
plot(g33, 1:2)
pred<-predict(g33,fold3sx, se.fit = FALSE)
sqrt(sum((pred-fold3s[,7])^2)/268)  #32118.35

g24<-lm(trainf4$charges~(trainf4$age+trainf4$sex
+trainf4$bmi+trainf4$children+trainf4$smoker
+trainf4$region)^2)
summary(g24)
Anova(g24,type=3)
g34<-lm(trainf4$charges~(trainf4$age+trainf4$bmi
+trainf4$smoker+trainf4$children+trainf4$region
+trainf4$smoker*trainf4$bmi))
summary(g34)
Anova(g34,type=3)
plot(g34, 1:2)
pred<-predict(g34,fold4sx, se.fit = FALSE)
sqrt(sum((pred-fold4s[,7])^2)/268)  #31949.23

g25<-lm(trainf5$charges~(trainf5$age+trainf5$sex
+trainf5$bmi+trainf5$children+trainf5$smoker
+trainf5$region)^2)
summary(g25)
Anova(g25,type=3)
g35<-lm(trainf5$charges~(trainf5$age+trainf5$bmi
+trainf5$smoker+trainf5$children+trainf5$region
+trainf5$smoker*trainf5$bmi))
summary(g35)
Anova(g35,type=3)
```

```
plot(g35, 1:2)
pred<-predict(g35,fold5sx, se.fit = FALSE)
sqrt(sum((pred-fold5s[,7])^2)/266)  #33546.75


#####
plot(charges~factor(smoker))
contrasts(factor(trains$sex))
contrasts(factor(trains$smoker))
contrasts(factor(trains$region))
confint(g3)
#smoker no is 1
#female is 1
#ref:southwest". 1-3:"northeast" "northwest" "southeast"
```