Regression Analysis on Medical Cost in United States

Qichen Huang

October 3, 2022

1 Introduction

1.1 Motivation

Medical insurance is a highly relevant ongoing social issue. Exploring how other health factors might affect a person's medical expenses is essential not only for medical insurance companies to turn a profit, but also for policymakers to ensure welfare is distributed fairly.
Thus, our project would like to explore how different factors affect a person's medical expenses charge to the insurance plan in 2010s in the United States.

1.2 Source of data

Our dataset can be found on Kaggle at https://www.kaggle.com/mirichoi0218/insurance.
This data was collected by Brett Lantz in the book Machine Learning by R. Sources of data inlcude demographic statistics published by the US Census Bureau in 2010s.

1.3 Description of data

The dataset consists of 1,338 patients' health data; initial data were collected with respective to four different US regions, but we decided that demographic regions are not very relevant in our context hence we will merge them.
Variables that were obtained are charges, age, sex, BMI, children, smoker; sex and smoker are categorical variables. See Table 1.

Table 1:

Variables	Explanation	Unit
charges	annual medical expenses charged	US dollars
Age	the age of the primary beneficiary	years
Sex	the biological gender of the beneficiary	Female (baseline), Male
BMI	Body Mass Index	Kg/m^2
Children	number of children/dependents	NA
Smoker	whether the beneficiary smokes	No (baseline), Yes

1.4 Goal

Our goal is to find a predictive relationship with charges and the explanatory variables (age, sex, BMI, children, smoker).

2 Data analysis

2.1 Data preprocess

First, we process the dataset, throw away unused column, and convert categorical variables into factors.

```
data <- read.csv("insurance.csv")
data$region <- NULL
data$sex = as.factor(data$sex)
data$smoker = as.factor(data$smoker)
str(data)
```

```
'data.frame': 1338 obs. of  6 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ charges : num  16885 1726 4449 21984 3867 ...
```

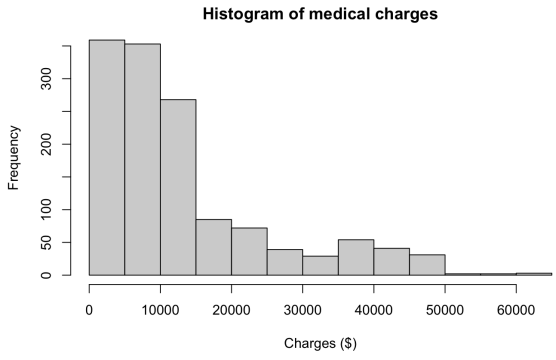
2.2 Explore the data

2.2.1 Overview

First, let's get an idea of general distribution of the variable of interest 'Charges'.

```
summary(data$charges)
hist(data$charges, main = "Histogram of medical charges", xlab="Charges ($)")
```

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1122    4740    9382   13270   16640   63770
```

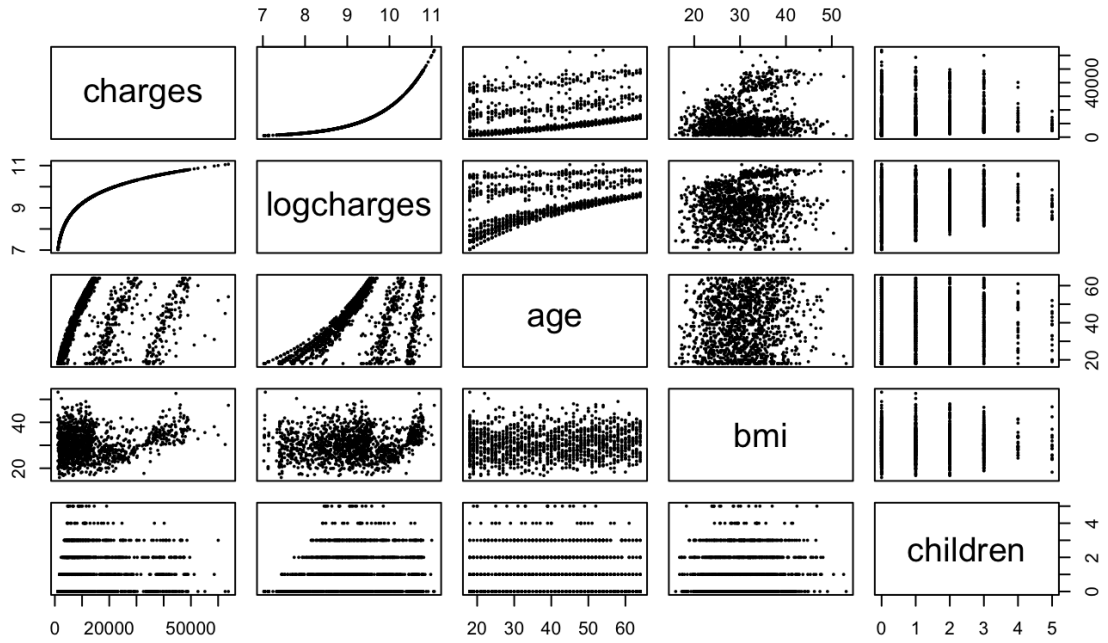


From Figure 1 we can see that most examples have charges around $0 - $15000 per year. The range is roughly $0 to $60000 with a right-skewed distribution.

2.2.2 Visualizations

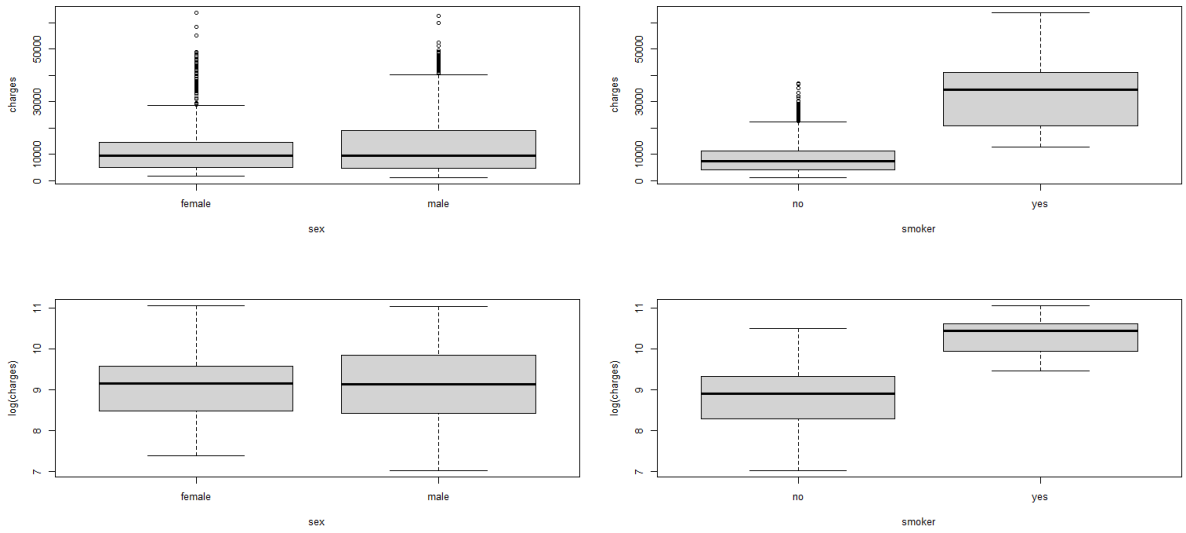
We further visualize the data by creating the scatterplot.

```
data$logcharges = log(data$charges)
pairs(data[c("charges","logcharges", "age", "bmi", "children")], panel = points, pch=16,cex=0.4)
```




```
sex<-as.factor(data$sex)
smoker<-as.factor(data$smoker)
par(mfrow=c(2,2))
plot(sex,charges, xlab="sex", ylab="charges")
plot(smoker,charges, xlab="smoker", ylab="charges")
plot(sex,log(charges), xlab="sex", ylab="log(charges)")
plot(smoker,log(charges), xlab="smoker", ylab="log(charges)")
```

Figure: Boxplots between log(charges)/charges and our categorical variables(sex/smoker).



As we transition from the scatterplots relating the explanatory variables to charges to scatterplots relating the variables to log(charges), we see that BMI becomes much more scattered,

suggesting a log transform might be appropriate. Additionally, the log(charges) vs children plot has a relatively clear positive relationship in contrast to the charges vs children plot's negative relationship.

2.2.3 Summary statistics

The frequency tables and graphs were generated using the library "epiDisplay"

```
summary(data)
sex<-data$sex
smoker<-data$smoker
tab1(sex)
tab1(smoker)
```

Univariate summary statistics:

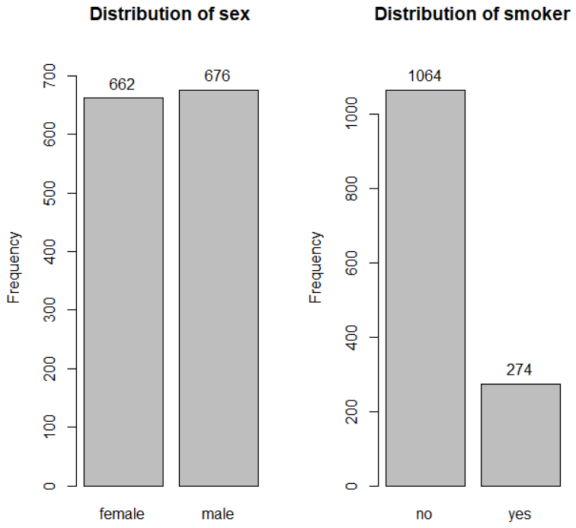
	charges	age	BMI	children
Min.	1122	18.00	15.96	0.000
1st Qu.	4740	27.00	26.30	0.000
Median	9382	39.00	30.40	1.000
Mean	13270	39.21	30.66	1.095
3rd Qu.	16640	51.00	34.69	2.000
Max.	63770	64.00	53.13	5.000

Frequency table for sex:

	Frequency	Percent	Cumulative percent
Female	662	49.5	49.5
Male	676	50.5	100.0
Total	1338	100.0	100.0

Frequency table for smokers:

	Frequency	Percent	Cumulative percent
Non-smoker	1064	79.5	79.5
Smoker	274	20.5	100.0
Total	1338	100.0	100.0



2.2.4 Correlations

We explore the correlations between numeric variables.

```
data$logcharges = log(data$charges)
cor(data[c("charges","logcharges", "age", "bmi", "children")])
```

```
             charges    logcharges   age        bmi    children
charges    1.00000000  0.8929642 0.2990082 0.1983410 0.06799823
logcharges 0.89296420  1.0000000 0.5278340 0.1326694 0.16133634
age        0.29900819  0.5278340 1.0000000 0.1092719 0.04246900
bmi        0.19834097  0.1326694 0.1092719 1.0000000 0.01275890
children   0.06799823  0.1613363 0.0424690 0.0127589 1.00000000
```

The correlation matrix shows that BMI and number of children is the least correlated with log(charges). One thing to notice is that charges and logcharges are expected to have high correlation as one is the other's logarithm transform. The numerical variables don't correlate with each other much hence we could include all in the model without worrying about collinearity.

2.2.5 Analysis Summary

From the plots, we can see that log(charges) is monotomically related to age, with a possible curvilinear relationship. The scatterplot against BMI suggest a log transform to be applied as explained in 2.2.2. Additionally, there is no clear distinct relationship with log(changes) and the covariates bmi and age, as their scatterplots show a graph with spread out points. From the correlation matrix, we can see that they are also not highly correlated with the log(Charges), with bmi being the least correlated.

From the boxplot of log(charges) by sex, there doesn't seem to be a noticeable difference in charges between the two sexes; male and female. Conversely, the boxplot of log(charges) by whether one smokes or not is noticeably different in that smokers tend to claim more money through their insurance plan.

We next fit a multiple regression model with our covariates.

2.3 Multiple Regression Model

2.3.1 Model 1: full model without interactions:

We start by fitting a full regression model with log(charges) against all other variables. The estimate and summary statistics are shown below.

```
fullWOInteraction <- lm(data = data, logcharges ~ .-charges)
summary(fullWOInteraction)
```

```
Call:
lm(formula = logcharges ~ . - charges, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.08241 -0.20315 -0.05185  0.07057  2.11173

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.0121103  0.0701685  99.932  < 2e-16 ***
age          0.0347158  0.0008781  39.536  < 2e-16 ***
sexmale     -0.0750088  0.0245899  -3.050  0.00233 **
bmi          0.0109087  0.0020225   5.394 8.16e-08 ***
children     0.1017275  0.0101688  10.004  < 2e-16 ***
smokeryes    1.5502366  0.0304293  50.946  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4477 on 1332 degrees of freedom
Multiple R-squared:  0.7638,        Adjusted R-squared:  0.7629
F-statistic: 861.5 on 5 and 1332 DF,  p-value: < 2.2e-16
```

This model provides an acceptable adjR^2 of 0.7629, with all parameter estimate having small p-values and being quite significant. Overall the model fit is decent. One thing we noticed is that some coefficient has an estimate of order of 10 to some negative power. However, since estimates like intercept and smokers are decently big, we do not want to scale the response variable (by some multiple of 10).

2.3.2 Model 2: full model with interactions:

Having the dummy variables, it is to our interest to explore how they might interact with the numeric variables. For example, it seems logical to have an interaction between gender and age, as demographic data often shows that females and males have a different lifespan. With this in mind we fit model 2 with all interaction included.

```
fullInteractions <- lm(data = data, logcharges ~ (age + bmi + children) * (sex + smoker))
summary(fullInteractions)
```

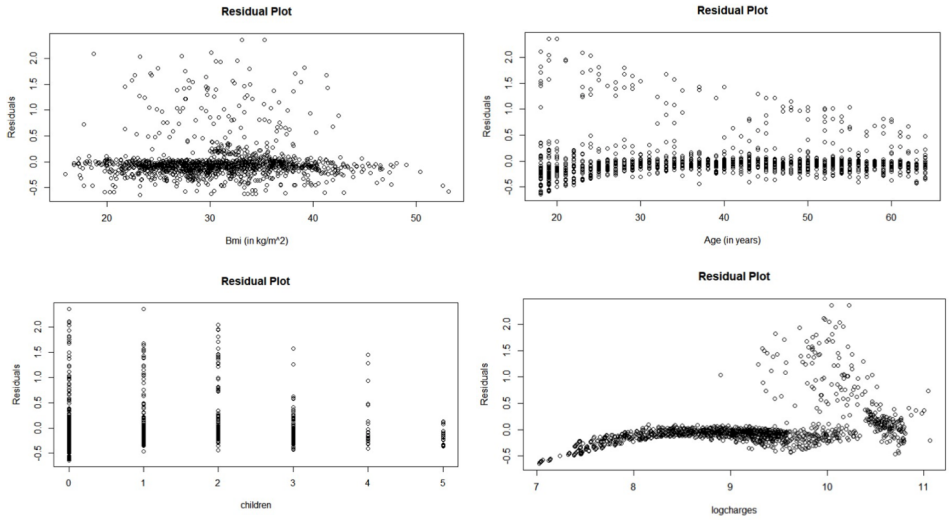
```
    Call:
lm(formula = logcharges ~ (age + bmi + children) * (sex + smoker),
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63828 -0.14944 -0.06785 -0.00122  2.35754
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         7.1906217  0.0905147  79.441  < 2e-16 ***
age                 0.0391886  0.0011253  34.827  < 2e-16 ***
bmi                -0.0012595  0.0026509  -0.475 0.634786
children            0.1217403  0.0131182   9.280  < 2e-16 ***
sexmale            -0.2998335  0.1195160  -2.509 0.012235 *
smokeryes           1.4085115  0.1472764   9.564  < 2e-16 ***
age:sexmale         0.0050445  0.0015175   3.324 0.000911 ***
age:smokeryes      -0.0331111  0.0018908 -17.512  < 2e-16 ***
bmi:sexmale         0.0001416  0.0034962   0.040 0.967702
bmi:smokeryes       0.0506777  0.0042058  12.049  < 2e-16 ***
children:sexmale    0.0148040  0.0176087   0.841 0.400656
children:smokeryes -0.1199501  0.0225450  -5.320 1.21e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3855 on 1326 degrees of freedom
Multiple R-squared:  0.8257,        Adjusted R-squared:  0.8243
F-statistic: 571.1 on 11 and 1326 DF,  p-value: < 2.2e-16
```

From the summary for model 2, we can see that the residual standard error is much smaller and adjusted R^2 is non-trivially larger than our model 1. This suggests that including interactions of the dummy variables is worth doing.
Meanwhile, We observe some coefficients appears non-significant (large p-value) compared to model 1. Therefore it is worth testing further variable selection, see section 2.4.

2.3.3 Residual plots:

Residual plots with model 2 are given below.



From the residual plots, we observe a potential curvilinear relationship with respect to age and bmi. Thus we try a quadratic model fitting in next section.

2.3.4 Model 3: Quadratic model

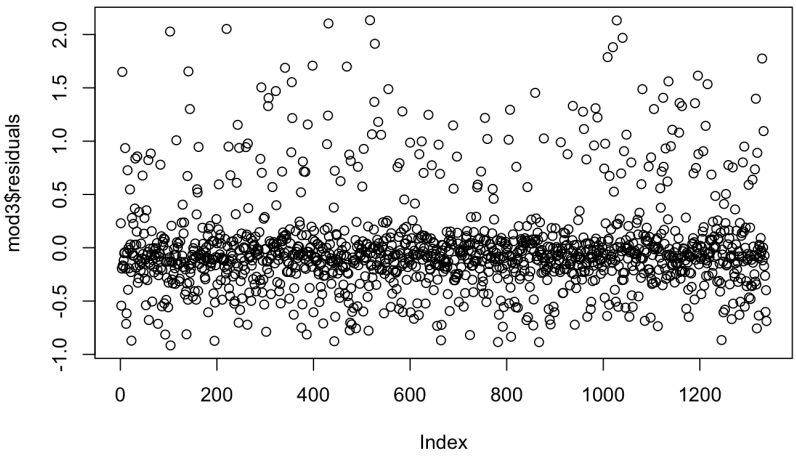
We fit a quadratic model including age2 term as follows:

```
data$age2 <- (data$age)^2
data$bmi2 <- (data$bmi)^2
mod3 <- lm(logcharges ~ . + age2 + bmi2 - charges, data=data)
summary(mod3)
```

```
Call:
lm(formula = logcharges ~ . + age2 + bmi2 - charges, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91574 -0.20527 -0.07017  0.07186  2.13437

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.0758367  0.2586364  23.492  < 2e-16 ***
age          0.0525918  0.0059804   8.794  < 2e-16 ***
sexmale     -0.0747124  0.0244533  -3.055 0.002293 **
bmi          0.0528263  0.0153022   3.452 0.000573 ***
children     0.0921280  0.0105944   8.696  < 2e-16 ***
smokeryes    1.5515463  0.0302693  51.258  < 2e-16 ***
age2        -0.0002272  0.0000746  -3.045 0.002370 **
bmi2        -0.0006607  0.0002404  -2.749 0.006061 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4452 on 1330 degrees of freedom
Multiple R-squared:  0.7668,        Adjusted R-squared:  0.7656
F-statistic: 624.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

This model gives an okay adjR^2 of 0.7656 , while age, bmi and their quadratice terms appear significant. We quickly test the residual plot for fit. It appears slightly better than previous.

Figure: Residual plot for model 3.



Thus, we will take this model as a potential candidate.

2.4 Model selection

We select the best models with different number of parameters by forward selection using regsubset():

```
s <- regsubsets(logcharges ~ (age + bmi + children) * (sex + smoker), data=D, method = "forward")
ss = summary(s)

ss$which
ss$adjr2
ss$cp
```

```
  (Intercept)   age   bmi children sexmale smokeryes age:sexmale age:smokeryes bmi:sexmale bmi:smokeryes children:sexmale children:smokeryes
1        TRUE FALSE FALSE    FALSE   FALSE     FALSE       FALSE         FALSE       FALSE         TRUE           FALSE             FALSE
2        TRUE  TRUE FALSE    FALSE   FALSE     FALSE       FALSE         FALSE       FALSE         TRUE           FALSE             FALSE
3        TRUE  TRUE FALSE    FALSE   FALSE     FALSE       FALSE          TRUE       FALSE         TRUE           FALSE             FALSE
4        TRUE  TRUE FALSE     TRUE   FALSE     FALSE       FALSE          TRUE       FALSE         TRUE           FALSE             FALSE
5        TRUE  TRUE FALSE     TRUE   FALSE      TRUE       FALSE          TRUE       FALSE         TRUE           FALSE             FALSE
6        TRUE  TRUE FALSE     TRUE   FALSE      TRUE       FALSE          TRUE       FALSE         TRUE           FALSE              TRUE
7        TRUE  TRUE FALSE     TRUE    TRUE      TRUE       FALSE          TRUE       FALSE         TRUE           FALSE              TRUE
8        TRUE  TRUE FALSE     TRUE    TRUE      TRUE        TRUE          TRUE       FALSE         TRUE           FALSE              TRUE
```

Summary statistics:

	1	2	3	4	5	6	7	8
adjr2	0.47	0.76	0.76	0.80	0.82	0.82	0.82	0.83
cp	2724	504	295	154	55	29	16	7

Based on "Forward" model selection method, adjustedR^2 statistics and Cp statistics, the best model is the last one with 8 variables, which are:
Age, children, sex, smoker, age:sex, age:smoker, bmi:smoker, children:smoker.

Best model selected is fitted as follows:

```
mod4 <- lm(data = data, logcharges ~ age + children + sex + smoker + age:sex + age:smoker + bmi:smoker + children:smoker)
summary(mod4)
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.64709 -0.14908 -0.06703  0.00070  2.35748

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.183947   0.074270  96.728  < 2e-16 ***
age                0.039132   0.001121  34.897  < 2e-16 ***
children           0.129133   0.009716  13.290  < 2e-16 ***
sexmale           -0.281853   0.062752  -4.492 7.68e-06 ***
smokeryes          1.401956   0.146991   9.538  < 2e-16 ***
age:sexmale        0.005106   0.001506   3.391 0.000717 ***
age:smokeryes     -0.033090   0.001888 -17.529  < 2e-16 ***
smokerno:bmi      -0.001229   0.001972  -0.623 0.533217
smokeryes:bmi      0.049606   0.003708  13.379  < 2e-16 ***
children:smokeryes -0.118494   0.022449  -5.278 1.52e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3853 on 1328 degrees of freedom
```

```
Multiple R-squared:  0.8256,        Adjusted R-squared:  0.8244
F-statistic: 698.6 on 9 and 1328 DF,  p-value: < 2.2e-16
```

We've also check residual plots of model 4 and it appear not worse than residual plots in 2.3.4.

2.5 Cross Validation

We select our candidates from previous sections, including (1) model 1 a full model without interactions; (2) model 2 a full model including all interactions; (3) model 3 the quadratic model in 2.3.4;(4) the best model selected from Cp statistics (see section 2.4). We then further validate by comparing training/holdout RMSE.

2.5.1 Methodology

To perform cross-validation, we randomly split the dataset into 20:80 testing and training set. set.seed() is used for easier reproduction.

```
library(caret)
set.seed(100)
training.samples <- createDataPartition(data$logcharges, p = 0.8, list = FALSE)
train1.data  <- data[training.samples, ]
test1.data <- data[-training.samples, ]

set.seed(123)
training.samples <- createDataPartition(data$logcharges, p = 0.8, list = FALSE)
train2.data  <- data[training.samples, ]
test2.data <- data[-training.samples, ]

set.seed(666)
training.samples <- createDataPartition(data$logcharges, p = 0.8, list = FALSE)
train3.data  <- data[training.samples, ]
test3.data <- data[-training.samples, ]
modCV3 <- lm(data = train3.data, logcharges ~ .-charges)
```

Then we train and test the models respectively. Below is the example for model 1. (Model 3 needs some extra caution since we need to add quadratic terms from train.data everytime)

```
modCV11 <- lm(data = train1.data, logcharges ~ .-charges)
predictions <- predict(modCV11, test1.data)
RMSE(predictions, test1.data$logcharges)
modCV12 <- lm(data = train2.data, logcharges ~ .-charges)
predictions <- predict(modCV12, test2.data)
RMSE(predictions, test2.data$logcharges)
modCV13 <- lm(data = train3.data, logcharges ~ .-charges)
predictions <- predict(modCV12, test3.data)
RMSE(predictions, test3.data$logcharges)
```

2.5.2 Result

We take Table 5.4 from the course notes as a reference and create the following table using our code for CV.

	model 1	model 2	model 3	model 4
adjustedR^2	0.7629	0.8243	0.7656	0.8244
residual SD	0.4477	0.3855	0.4452	0.3853
rmsepred(train/holdout) 1	0.4164085	0.3387911	0.4156608	0.3386095
rmsepred(train/holdout) 2	0.4502805	0.3860948	0.4474212	0.3858685
rmsepred(train/holdout) 3	0.4749948	0.4077659	0.4692503	0.4072026

We observe that model 2 and model 4 appear to have the highest adjustedR^2 and prediction performance. We do not see the quadratic model perform better with respect to any of the criterion statistics. Among model 2 and 4, which is the two models including interaction terms, their prediction performance and statistics are similarly good. If we check their model summary

back in section 2.3 and 2.4, model 4 has its parameter estimated appear more significant.

3 Conclusion

In conclusion, we have found a good regression model (model 4), and from that model we can understand how terms interact with one another as well as which of those interactions correlate most to log(charges).

One takeaway from our interaction model is that smoking, and its interaction with other relevant covariates consistently have a very small p-value, indicating that there is a very strong chance there is a positive correlation between smoking and medical expenses. Interactions of variables including BMI have a large p-value, which means we can confidently say it that BMI may not be highly correlated to medical expenses, contrary to what one might think.

However, one limitation is that we are not entirely satisfied about the residual plots. It shows a slight tendency to underestimate the medical cost. With more data being collected and other variables (population density in residence region, social economic status etc.) to explore in the future, we might be able to obtain a more homoscedastic variation.