1. Introduction (1-3 paragraphs)

According to a report presenting statistics on health insurance coverage in the United States by Current Population Survey Annual Social and Economic Supplements (CPS ASEC) and the American Community Survey (ACS), 91.2% of American have health insurance in 2017. One of the most frequently asked questions that people have is that how much they will be charged if they want to have a health insurance. Thus, we would like to use the technique of linear regression to deal with this issue. To be more specific, we would like to apply different linear regression skills with a health insurance dataset to build a regression model. From this model, we will be able to discover the relationships between different factors and charge.
We got the dataset from this link: https://www.kaggle.com/mirichoi0218/insurance

2. Method (2 pages)

(a) Brief Description of Data
This is a health insurance dataset containing totally 1338 rows with 6 predictors and 1 response variable. The following is the information of the variables.
Response variable:
charges: individual medical costs billed by health insurance; numerical
Input variables:
age: age of primary beneficiary; numerical
sex: insurance contractor gender, may be female or male; categorical
bmi: body mass index; numerical
children: number of children covered by health insurance; numerical
smoker: whether smoke or not, may be yes or no; categorical
region: the beneficiary's residential area in the US, may be northeast, southeast, southwest, northwest; categorical

(b) Description of Preliminary Exploratory Analysis

Firstly, we drew a histogram to show the distribution of response variable charges. We realized that most of the values are between 0 to 20000. To make it more spread out, we decided to do a logarithm transformation on the charges. Then, we drew multiple scatter plots between log(charges) and predictors. From the plots, we may find out several relationships between predictors and response variable. Firstly, log(charges) increase as the age grows. Also, it seems that log(charges) increase as bmi gets higher, though the relationship is pretty weak. Also, log(charges) and children have positive correlation. As for categorical variables, we can tell from box plots that female and male have similar medians with male has larger variability. Similarly, the four regions also have similar medians. However, smoker with value yes have much higher charges than those with no.

(c,d) Models Buliding Process

Firstly, we built a main effect model *m1* using all six predictor variables and log(charge) as response variable. The model has the adjusted R-squared value of 0.7666, which means that this model can explain only about 77% of the response variable. However, since the p-value is less than 2.2e-16, we can tell that most of the predictors in the model are significant.

Next, we tried to build a model *m2* with full interaction, which contains all possible interaction between predictor variables. Now the adjusted R-squared has increased to 0.8316 with smaller AIC and a small p-value less than 2.2e-16. Though this model seems ideal, it contains many statistically insignificant predictors. Furthermore, it's too complex to interpret. Therefore, we would like to use this model to build a more precise model.

We then used stepwise variable selection method with both direction to build model *m3*. We used the main effect model *m1* as our base step and full interaction model *m2* as an upper limit, hoping to get significant interactions. It turns out that *m3* has similar adjusted R-squared as *m2*. However, this step was just needed for us to determine which interactions should be included in our model. Based on the result, the interaction *age:smoker, bmi:smoker, age:children, children:smoker, age:sex, age:region* are significant. This led us to think about the variable *bmi* in depth. Thus, we made an assumption that since bmi is considered 'normal' until 29.9, only bmi with higher than 30 might affect charges significantly.

To further explore the effect of *bmi*, we created a variable called *bmiOver30* which is a factor. Then, we added this variable to build the model *m4* with 6 original predictors and new interactions *age:smoker, bmiOver30:smoker, age:children, children:smoker, age:sex, age:region, bmiOver30:region, age:children:smoker*. This model has adjusted R-squared 0.8367. Besides, it has better AIC value. However, since there is still predictor that is not significant in the model, we would like to build a model *m5* based on *m4* by deleting those insignificant variables. Now all the predictors are significant and the adjusted R-squared is 0.8324.

(e) Model Checking

To make sure the final model is feasible, significance test, ANOVA test, residual analysis, normality test, error independence test, homoscedasticity test and multicollinearity test are applied. The final model passes all of above tests so that it is feasible and reasonable.

3. Results (2 pages)

(a) Model Selection

Name	**Model**	**Adjusted R^2**	**AIC**	**p-value**
m5	lm(log(charges)~age+sex+children+smoker+bmiOver30:smoker+age:children+age:smoker+smoker:children)	0.8324	1194.529	<2.2e-16
m4	lm(log(charges)~age+sex+bmi+children+smoker+age:smoker+bmiOver30:smoker+age:children+children:smoker+age:sex+age:region+bmiOver30:region+age:children:smoker)	0.8367	1168.562	<2.2e-16
m3	step(m1, scope=list(upper=m2), direction = "both")	0.8374	1169.492	<2.2e-16
m2	lm(log(charges) ~ age * sex * bmi * children * smoker * region)	0.8316	1312.767	<2.2e-16
m1	lm(log(charges) ~ age + sex + bmi + children + smoker + region)	0.7666	1637.033	<2.2e-16

(b) ANOVA table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	314.96	314.96	2222.7288	< 2.2e-16 ***
sex	1	0.31	0.31	2.2101	0.1374
children	1	21.77	21.77	153.6260	< 2.2e-16 ***
smoker	1	520.59	520.59	3673.8955	< 2.2e-16 ***
smoker:bmiOver30	2	28.78	14.39	101.5526	< 2.2e-16 ***
age:children	1	6.02	6.02	42.4722	1.015e-10 ***
age:smoker	1	45.67	45.67	322.2676	< 2.2e-16 ***
children:smoker	1	4.20	4.20	29.6514	6.154e-08 ***
Residuals	1328	188.18	0.14	-	-

(c) Parameter estimates table for final model

Parameter	Point Estimate	Standard Error	t-statistic	p-value	Confidence Interval	
					2.5 %	97.5 %
(Intercept)	6.9340643	0.042773	162.113	< 2e-16	6.850154218	7.01797433

age	0.0447368	0.000965	46.373	< 2e-16	0.042844285	0.04662934
sexmale	-0.086777	0.020725	-4.187	3.01e-05	-0.12743272	-0.0461202
children	0.2881687	0.027424	10.508	< 2e-16	0.234370046	0.34196740
smokeryes	2.5666818	0.081120	31.641	< 2e-16	2.407544836	2.72581867
smokerno:bmiOver30	-0.005278	0.023250	-0.227	0.82	-0.05088825	0.04033223
smokeryes:bmiOver30	0.6855582	0.045708	14.999	< 2e-16	0.595891332	0.77522505
age:children	-0.004069	0.000653	-6.23	6.07e-10	-0.00534855	-0.0027883
age:smokeryes	-0.032201	0.001839	-17.508	< 2e-16	-0.03580850	-0.0285925
children:smokeryes	-0.119442	0.021935	-5.445	6.15e-08	-0.16247248	-0.0764112

4. Discussion (1 page)

After running five models and concise modeling selection process, we choose m5 as our final model. Based on several types of modeling checking methods, we could reach the conclusion that our model is reasonable and feasible.

According to our final model, the variable, age, is significant. Older people may have higher insurance charges. Female would pay less for their insurance. The number of children, is relevant to charges. When people have more children, they need to pay more for insurance. The charges for smokers are higher than those for non-smokers. In terms of interactions, the interactions between smokers with bmi over 30 and non-smokers with bmi over 30 are both significant for the insurance charges. Among people whose bmi is over 30, smokers tend to pay more than non-smokers.

However, the model and results given by this project have some limitations as well. On the one hand, we only collected about 1400 set of data in the dataset because of the limited resource. Though it is not a small number, it could help get a more accurate model if we could have more data. On the other hand, actually, there should be more variables which could influence the insurance charges, like the customers’ job, salaries and so on. Here, only six variables are considered. It is necessary to discuss how other variables can affect the insurance charges. A more comprehensive dataset can help solve this problem. The interactions between age and children, age and smoker, children and smoker are also significant.

5. References (1 page)

[1] https://www.kaggle.com/mirichoi0218/insurance

[2] Grumbach, Kevin, et al. "Charges for obstetric liability insurance and discontinuation of obstetric practice in New York." *Journal of Family Practice*, Jan. 1997, p. 61+. *Academic OneFile*, https://link.galegroup.com/apps/doc/A19122984/AONE?u=googlescholar&sid=AONE&xid=1a611fbd. Accessed 26 Nov. 2018.

[3] https://www.census.gov/library/publications/2018/demo/p60-264.html

6. Appendix A: Code

```r
# transform some variables into factor
insurance$sex = factor(insurance$sex)
insurance$smoker = factor(insurance$smoker)
insurance$region = factor(insurance$region)

# draw plots to see relationships
plot(age, charges)
plot(bmi, charges)
plot(children, charges)
boxplot(charges ~ sex, xlab='sex', ylab='charges')
boxplot(charges ~ smoker, xlab='smoker', ylab='charges')
boxplot(charges ~ region, xlab='region', ylab='charges')

plot(age, bmi)
plot(age, children)
plot(bmi, children)

plot(insurance)

# main effects model
m1 = lm(log(charges) ~ age + sex + bmi + children + smoker +
region, data = insurance)
summary(m1)
anova(m1)
AIC(m1)

layout(matrix(c(1,2,3,4), 2,2))
```

```r
plot(m1)

# all possible interaction
m2 = lm(log(charges) ~ age * sex * bmi * children * smoker *
region, data = insurance)
summary(m2)
anova(m2)
AIC(m2)

# Half-way model
m3 <- step(m1,scope=list(upper=m2),direction = "both")
summary(m3)
AIC(m3)

# Considered model
bmiOver30 <- ifelse(insurance$bmi >= 30, 1, 0)
m4 <-
lm(log(charges)~age+sex+bmi+children+smoker+age:smoker+bmiOver30:s
moker+age:children+children:smoker+age:sex+age:region+bmiOver30:re
gion+age:children:smoker,data=insurance)
summary(m4)
AIC(m4)

# Final Model
m5 <-
lm(log(charges)~age+sex+children+smoker+bmiOver30:smoker+age:child
ren+age:smoker+smoker:children,data=insurance)
summary(m5)
AIC(m5)
confint(m5)

#Compare m4 and m5
anova(m4,m5)

# visualize model checking
library(ggfortify)
autoplot(m5)

# significance test # comprehensive test
summary(m5)

# anova test
```

```r
anova(m5)
plot(m5)


# residual analysis:delete outliers
plot(m5,which=1:4)
qqPlot(m5,id.method='identify',simulate=T)
outlierTest(m5)

# normality test
res <- residuals(m5)
plot(density(res))
shapiro.test(res)
# => pass!

# homoscedasticity test
ncvTest(m5)
# => pass!
spreadLevelPlot(m5)

# multicolinearity test
vif(m5)
sqrt(vif(m5))>10
# => pass!

# comprehensive test
library(gvlma)
gvmodel <- gvlma(m5)
gvmodel
```

7. Appendix B: Output

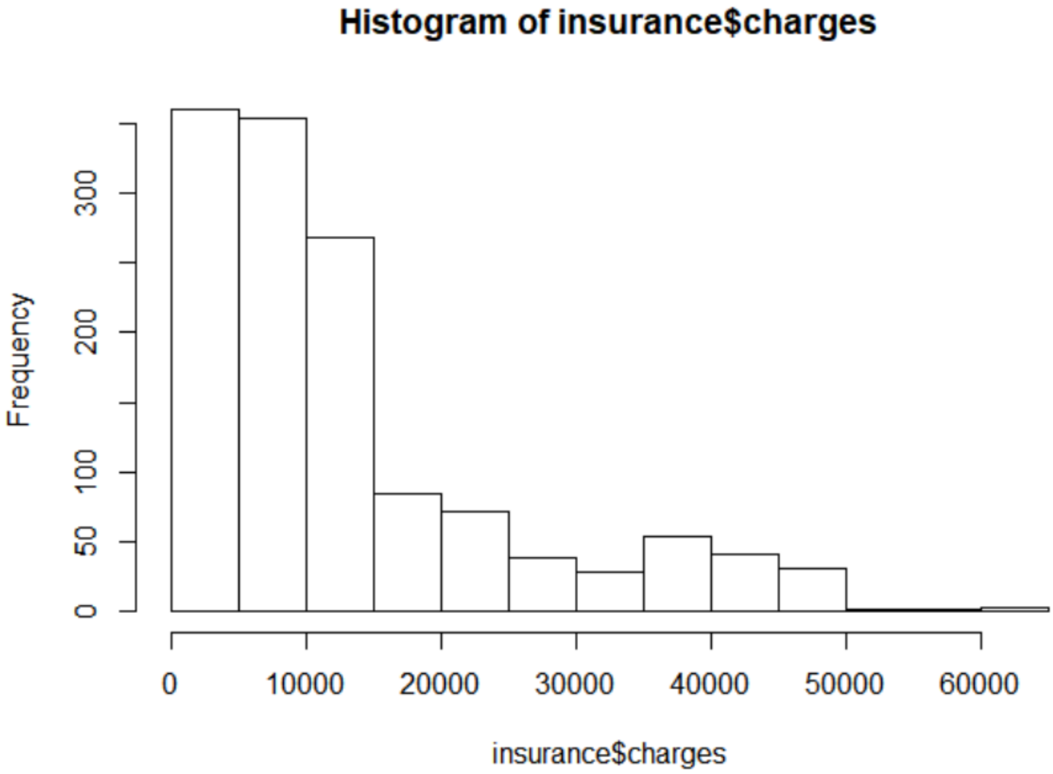

Fig. 7.1 Histogram of charges

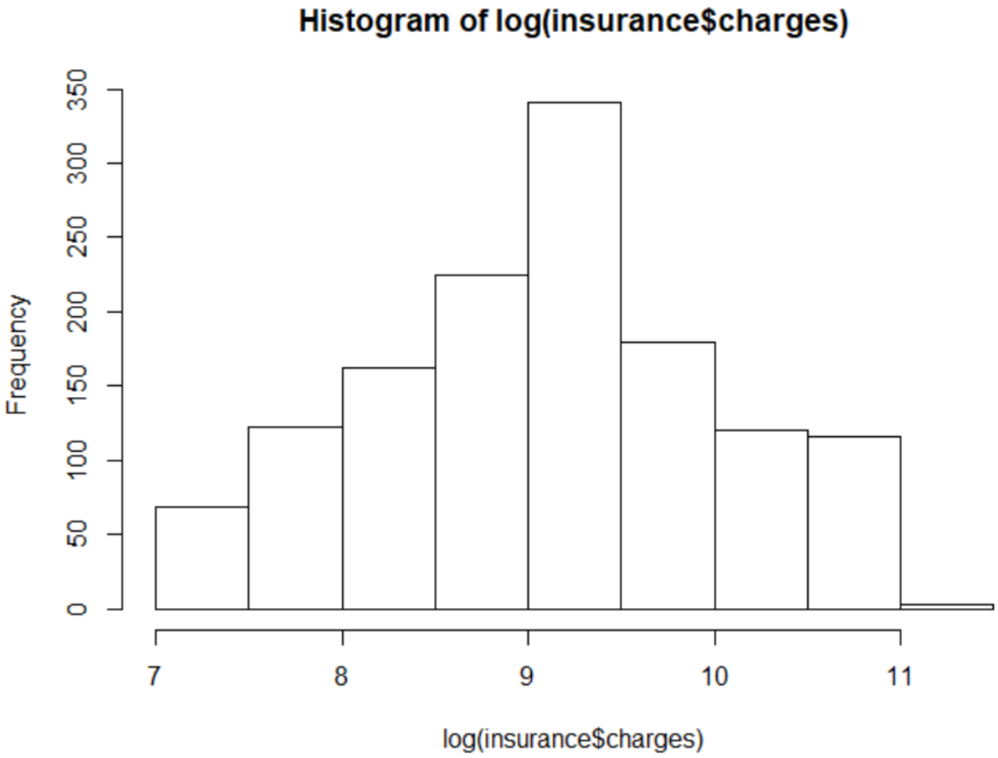

Fig. 7.2 Histogram of log(charges)

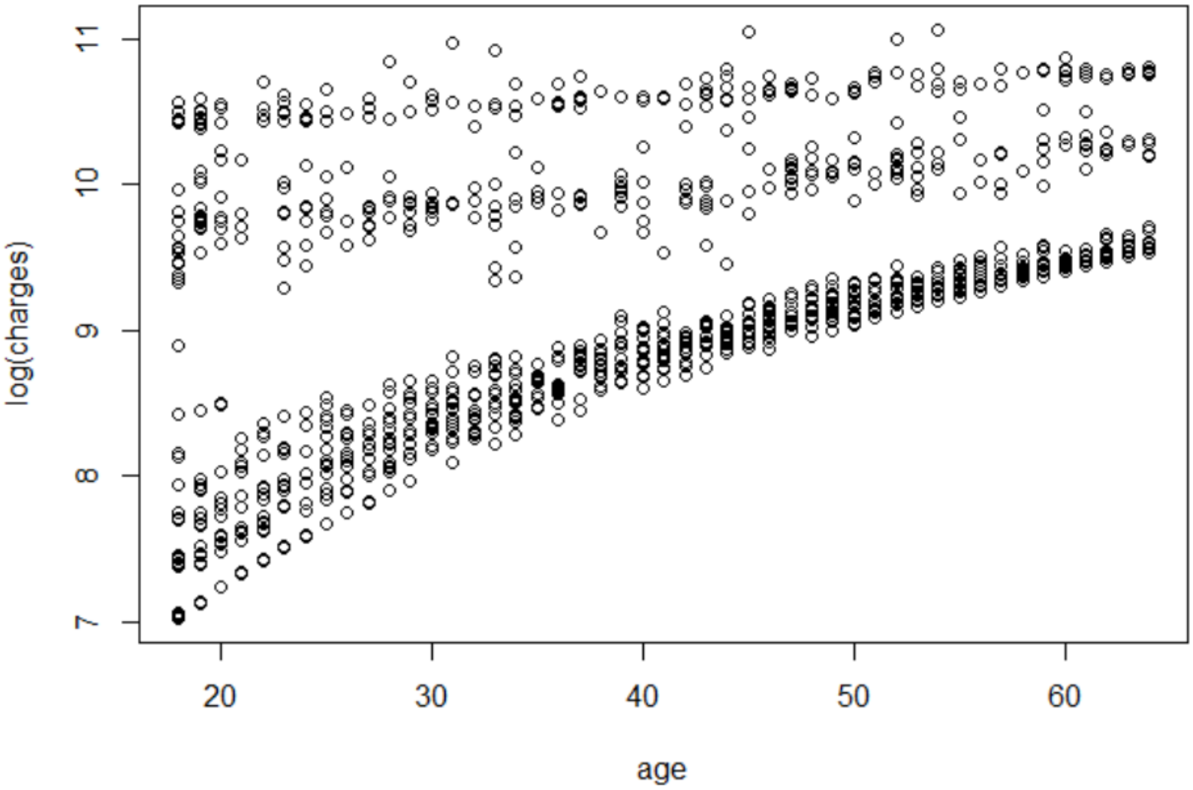

Fig. 7.3 Scatter plot of log(charges) ~ age

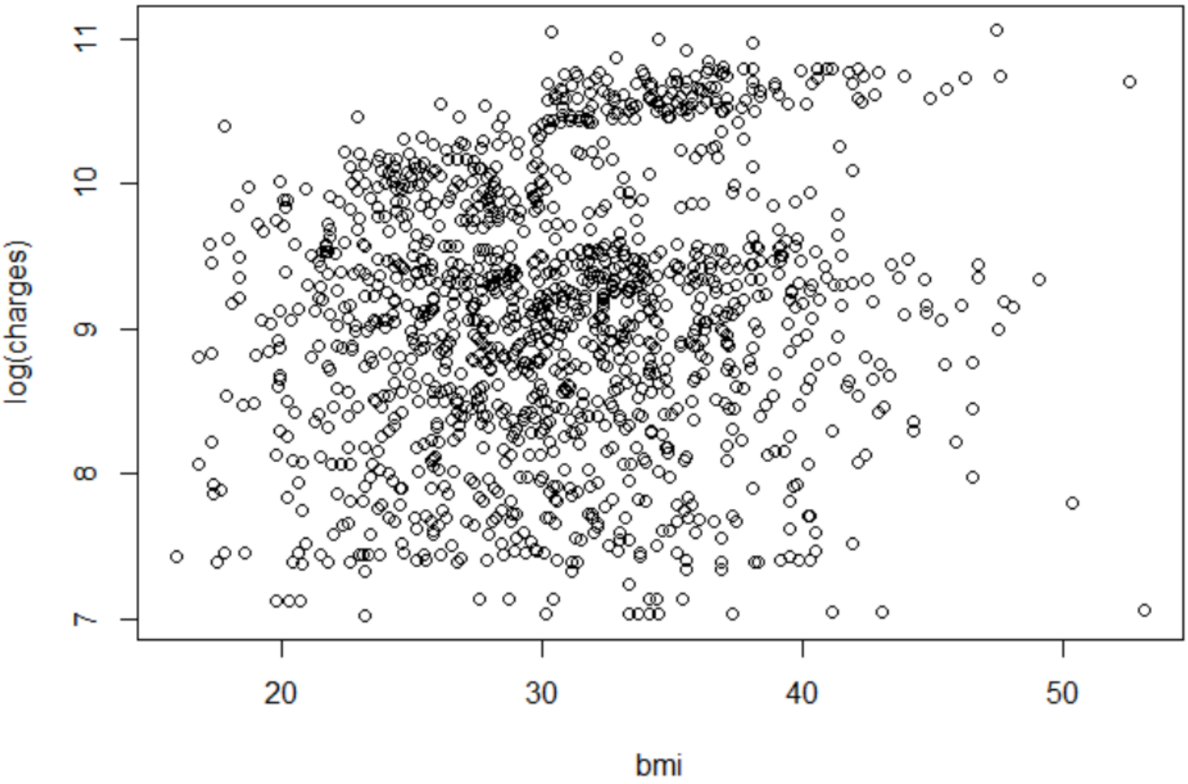

Fig. 7.4 Scatter plot of log(charges) ~ bmi

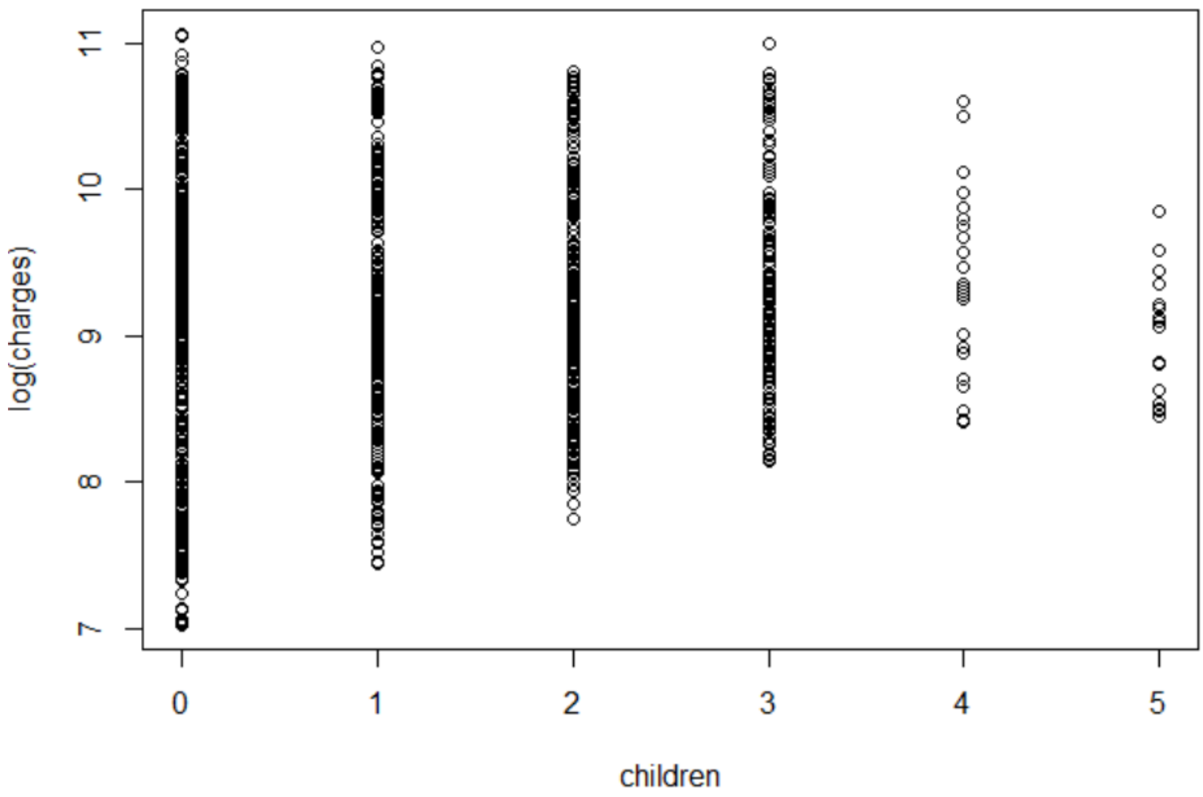

Fig. 7.5 Scatter plot of log(charges) versus children

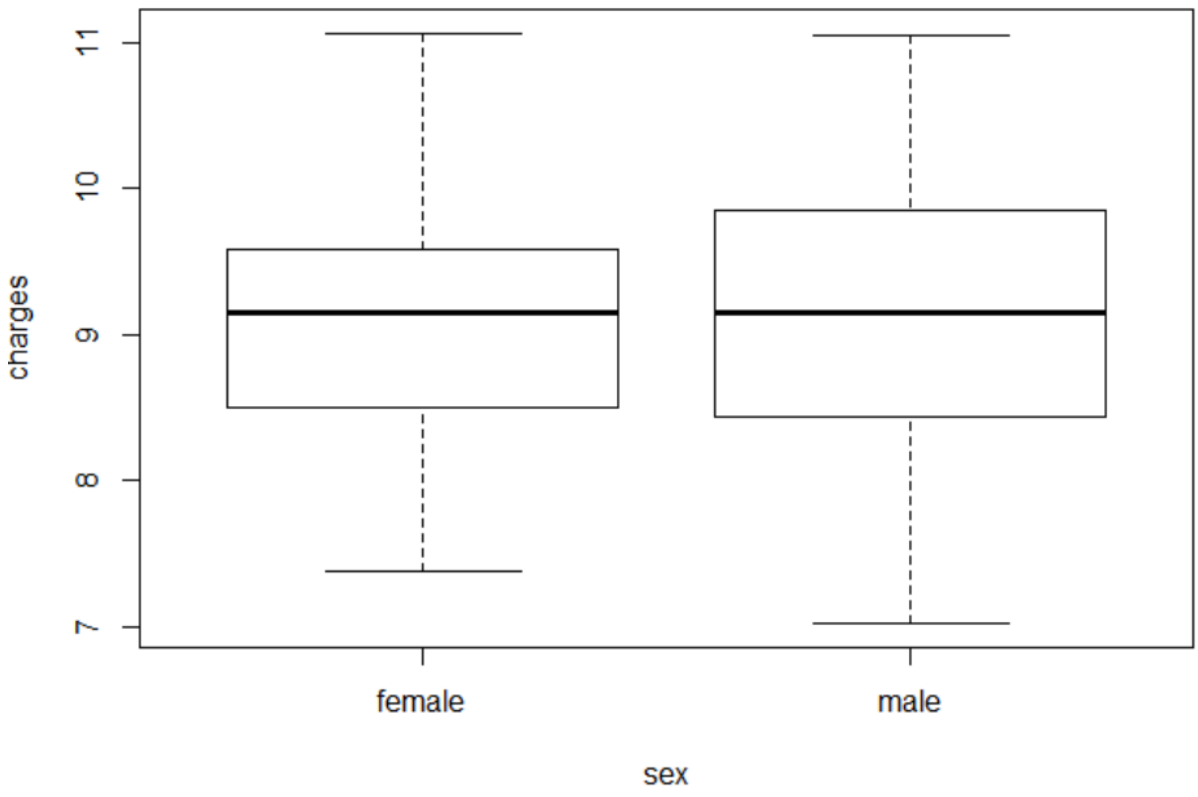

Fig. 7.6 Box plot of log(charges) versus sex

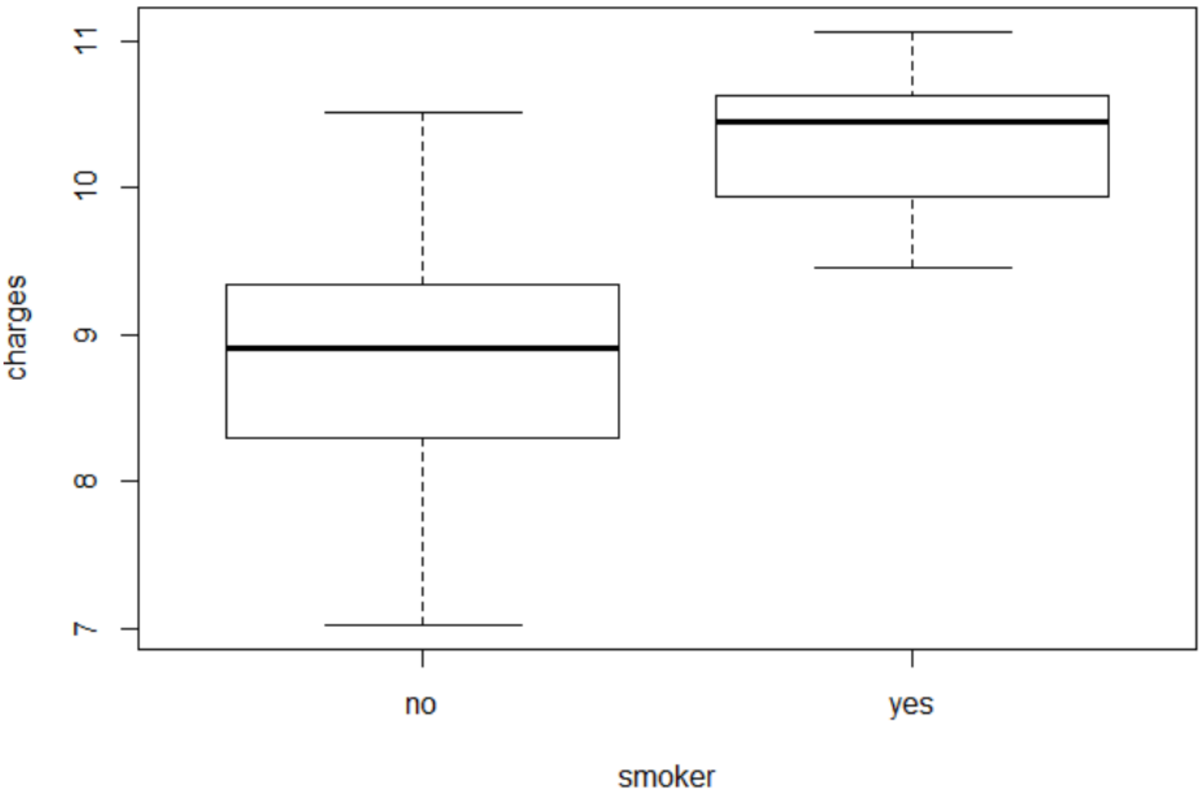

Fig. 7.7 Box plot of log(charges) versus smoker

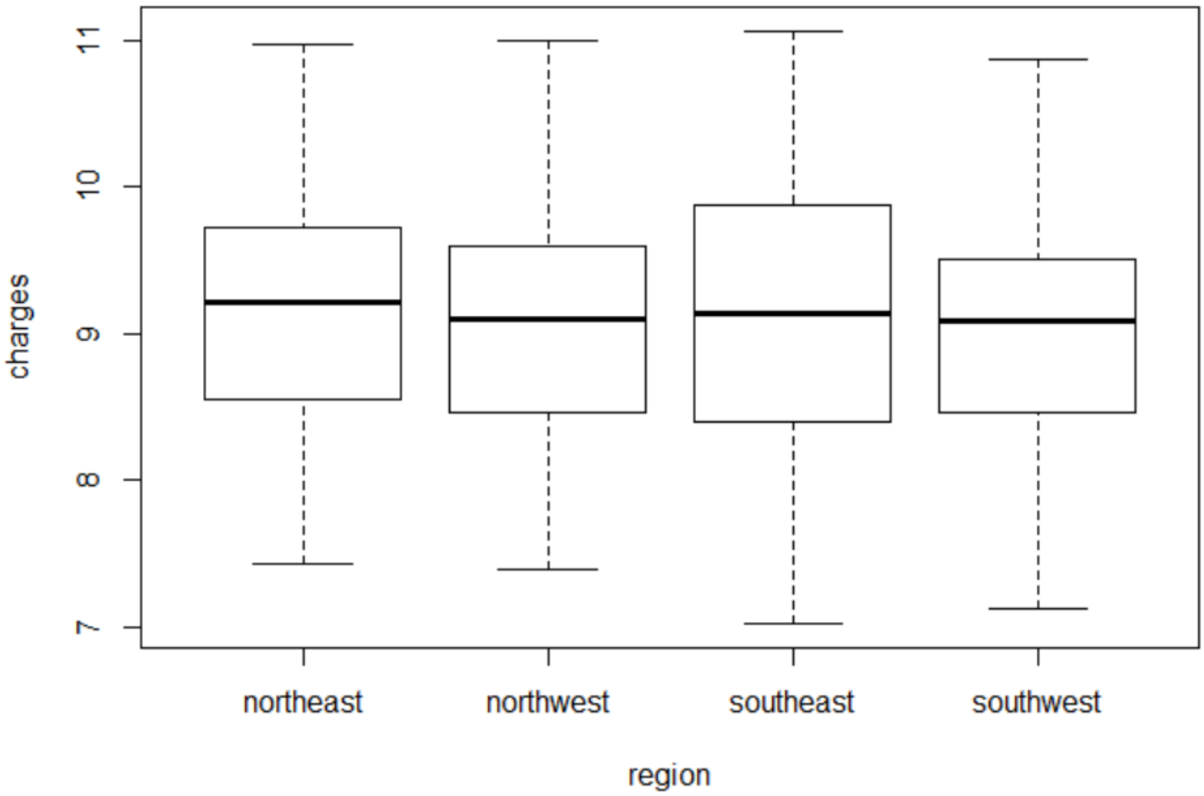

Fig. 7.8 Box plot of log(charges) versus region

```
# main effects model
```

```
> summary(m1)

Call:
lm(formula = log(charges) ~ age + sex + bmi + children + smoker + 
    region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.07186 -0.19835 -0.04917  0.06598  2.16636 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      7.0305581  0.0723960  97.112  < 2e-16 ***
age              0.0345816  0.0008721  39.655  < 2e-16 ***
sexmale         -0.0754164  0.0244012  -3.091 0.002038 ** 
bmi              0.0133748  0.0020960   6.381 2.42e-10 ***
children         0.1018568  0.0100995  10.085  < 2e-16 ***
smokeryes        1.5543228  0.0302795  51.333  < 2e-16 ***
regionnorthwest -0.0637876  0.0349057  -1.827 0.067860 .  
regionsoutheast -0.1571967  0.0350828  -4.481 8.08e-06 ***
regionsouthwest -0.1289522  0.0350271  -3.681 0.000241 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4443 on 1329 degrees of freedom
Multiple R-squared:  0.7679,	Adjusted R-squared:  0.7666 
F-statistic: 549.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

```
> AIC(m1)
[1] 1637.033
```

```
# all possible interaction
```

```
> m2 = lm(log(charges) ~ age * sex * bmi * children * smoker * region,
data = insurance)
Call:
lm(formula = log(charges) ~ age * sex * bmi * children * smoker * 
    region, data = insurance)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.68444 -0.15029 -0.06529  0.00265  2.46927 

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3774 on 1210 degrees of freedom
Multiple R-squared:  0.8476,	Adjusted R-squared:  0.8316 
F-statistic: 52.98 on 127 and 1210 DF,  p-value: < 2.2e-16
```

```
> AIC(m2)
[1] 1312.767
```

```
> anova(m1,m2)
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   1329 262.33                                   
2   1210 172.32 119    90.006 5.3109 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> # Half-way model
> summary(m3)

Call:
lm(formula = log(charges) ~ age + sex + bmi + children + smoker + 
    region + age:smoker + bmi:smoker + age:children + children:smoker + 
    age:sex + age:region + bmi:region + sex:smoker + age:bmi + 
    smoker:region + age:children:smoker, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.52439 -0.15504 -0.07881 -0.01138  2.55417 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                6.7938672  0.1871676  36.298  < 2e-16
age                        0.0442536  0.0040504  10.926  < 2e-16
sexmale                   -0.2956321  0.0616672  -4.794 1.82e-06
bmi                        0.0169406  0.0060567   2.797 0.005234
```

```
children                   0.3243387  0.0298037  10.882  < 2e-16
smokeryes                  1.4536886  0.1585072   9.171  < 2e-16
regionnorthwest           -0.0066814  0.1711631  -0.039 0.968868
regionsoutheast           -0.0027048  0.1634594  -0.017 0.986800
regionsouthwest           -0.2640395  0.1645770  -1.604 0.108877
age:smokeryes             -0.0368050  0.0022844 -16.111  < 2e-16
bmi:smokeryes              0.0497069  0.0043119  11.528  < 2e-16
age:children              -0.0049765  0.0007179  -6.932 6.50e-12
children:smokeryes        -0.3044677  0.0674578  -4.513 6.95e-06
age:sexmale                0.0048209  0.0014553   3.313 0.000950
age:regionnorthwest        0.0026388  0.0020957   1.259 0.208211
age:regionsoutheast        0.0080687  0.0021003   3.842 0.000128
age:regionsouthwest        0.0078965  0.0021344   3.700 0.000225
bmi:regionnorthwest       -0.0056689  0.0053729  -1.055 0.291579
bmi:regionsoutheast       -0.0149898  0.0046520  -3.222 0.001303
bmi:regionsouthwest       -0.0078903  0.0051422  -1.534 0.125170
sexmale:smokeryes          0.0960914  0.0516605   1.860 0.063101
age:bmi                   -0.0001964  0.0001249  -1.573 0.116030
smokeryes:regionnorthwest  0.0504945  0.0743255   0.679 0.497023
smokeryes:regionsoutheast  0.0798718  0.0707557   1.129 0.259173
smokeryes:regionsouthwest  0.1828824  0.0751375   2.434 0.015067
age:children:smokeryes     0.0046933  0.0016548   2.836 0.004636
---
(Intercept)               ***
age                       ***
sexmale                   ***
bmi                       **
children                  ***
smokeryes                 ***
regionnorthwest
regionsoutheast
regionsouthwest
age:smokeryes             ***
bmi:smokeryes             ***
age:children              ***
children:smokeryes        ***
age:sexmale               ***
age:regionnorthwest
age:regionsoutheast       ***
age:regionsouthwest       ***
bmi:regionnorthwest
bmi:regionsoutheast       **
bmi:regionsouthwest
sexmale:smokeryes         .
age:bmi
smokeryes:regionnorthwest
```

```
smokeryes:regionsoutheast
smokeryes:regionsouthwest *
age:children:smokeryes    **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3707 on 1312 degrees of freedom
Multiple R-squared:  0.8405,	Adjusted R-squared:  0.8374 
F-statistic: 276.5 on 25 and 1312 DF,  p-value: < 2.2e-16
```

```
> AIC(m3)
[1] 1169.492
```

```
> # Considered model
> summary(m4)

Call:
lm(formula = log(charges) ~ age + sex + bmi + children + smoker + 
    age:smoker + bmiOver30:smoker + age:children + children:smoker + 
    age:sex + age:region + bmiOver30:region + age:children:smoker, 
    data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.58866 -0.14054 -0.07513 -0.00679  2.47407 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                6.8558714  0.0883058  77.638  < 2e-16 ***
age                        0.0436244  0.0013549  32.197  < 2e-16 ***
sexmale                   -0.2418968  0.0606713  -3.987 7.06e-05 ***
bmi                        0.0052866  0.0028828   1.834  0.06690 .  
children                   0.3216612  0.0297813  10.801  < 2e-16 ***
smokeryes                  2.7243129  0.0955219  28.520  < 2e-16 ***
age:smokeryes             -0.0363313  0.0022811 -15.927  < 2e-16 ***
smokerno:bmiOver30         0.0380279  0.0509148   0.747  0.45526    
smokeryes:bmiOver30        0.7293499  0.0643959  11.326  < 2e-16 ***
age:children              -0.0048794  0.0007176  -6.799 1.59e-11 ***
children:smokeryes        -0.2994366  0.0674406  -4.440 9.75e-06 ***
age:sexmale                0.0040394  0.0014558   2.775  0.00560 ** 
age:regionnorthwest       -0.0001738  0.0009301  -0.187  0.85182    
age:regionsoutheast       -0.0003806  0.0010014  -0.380  0.70393    
age:regionsouthwest       -0.0010771  0.0009761  -1.103  0.27001    
bmiOver30:regionnorthwest -0.0624980  0.0581025  -1.076  0.28228    
```

```
bmiOver30:regionsoutheast -0.1361595  0.0564043  -2.414  0.01591 *
bmiOver30:regionsouthwest -0.1393545  0.0584029  -2.386  0.01717 *
age:children:smokeryes     0.0045806  0.0016527   2.772  0.00566 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3716 on 1319 degrees of freedom
Multiple R-squared:  0.8389,	Adjusted R-squared:  0.8367
F-statistic: 381.6 on 18 and 1319 DF,  p-value: < 2.2e-16
```

```
> AIC(m4)
[1] 1168.562
```

```
> # Final Model
> summary(m5)

Call:
lm(formula = log(charges) ~ age + sex + children + smoker + 
bmiOver30:smoker + 
    age:children + age:smoker + smoker:children, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.62979 -0.13713 -0.07356 -0.00603  2.35484 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          6.9340643  0.0427730 162.113  < 2e-16 ***
age                  0.0447368  0.0009647  46.373  < 2e-16 ***
sexmale             -0.0867765  0.0207245  -4.187 3.01e-05 ***
children             0.2881687  0.0274238  10.508  < 2e-16 ***
smokeryes            2.5666818  0.0811198  31.641  < 2e-16 ***
smokerno:bmiOver30  -0.0052780  0.0232497  -0.227     0.82    
smokeryes:bmiOver30  0.6855582  0.0457075  14.999  < 2e-16 ***
age:children        -0.0040684  0.0006525  -6.235 6.07e-10 ***
age:smokeryes       -0.0322005  0.0018392 -17.508  < 2e-16 ***
children:smokeryes  -0.1194418  0.0219348  -5.445 6.15e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3764 on 1328 degrees of freedom
Multiple R-squared:  0.8335,	Adjusted R-squared:  0.8324 
F-statistic: 738.9 on 9 and 1328 DF,  p-value: < 2.2e-16
```

```
> AIC(m5)
[1] 1194.529
```

```
> #Compare m4 and m5
> anova(m4,m5)
Analysis of Variance Table

Model 1: log(charges) ~ age + sex + bmi + children + smoker + age:smoker 
+ 
    bmiOver30:smoker + age:children + children:smoker + age:sex + 
    age:region + bmiOver30:region + age:children:smoker
Model 2: log(charges) ~ age + sex + children + smoker + bmiOver30:smoker 
+ 
    age:children + age:smoker + smoker:children
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1   1319 182.09                                  
2   1328 188.18 -9    -6.083 4.8958 1.777e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> confint(m5)
                            2.5 %       97.5 %
(Intercept)           6.850154218  7.017974330
age                   0.042844285  0.046629335
sexmale              -0.127432720 -0.046120240
children              0.234370046  0.341967399
smokeryes             2.407544836  2.725818673
smokerno:bmiOver30   -0.050888250  0.040332226
smokeryes:bmiOver30   0.595891332  0.775225046
age:children         -0.005348547 -0.002788304
age:smokeryes        -0.035808503 -0.028592502
children:smokeryes   -0.162472480 -0.076411186
```

```
> # anova test
> anova(m5)
Analysis of Variance Table

Response: log(charges)
                   Df Sum Sq Mean Sq  F value    Pr(>F)    
age                 1 314.96  314.96 2222.7288 < 2.2e-16 ***
sex                 1   0.31    0.31    2.2101    0.1374    
children            1  21.77   21.77  153.6260 < 2.2e-16 ***
smoker              1 520.59  520.59 3673.8955 < 2.2e-16 ***
smoker:bmiOver30    2  28.78   14.39  101.5526 < 2.2e-16 ***
```

```
age:children        1    6.02    6.02   42.4722 1.015e-10 ***
age:smoker          1   45.67   45.67  322.2676 < 2.2e-16 ***
children:smoker     1    4.20    4.20   29.6514 6.154e-08 ***
Residuals        1328 188.18    0.14
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Modeling Checking for Final Model
```

```
# residual analysis
```

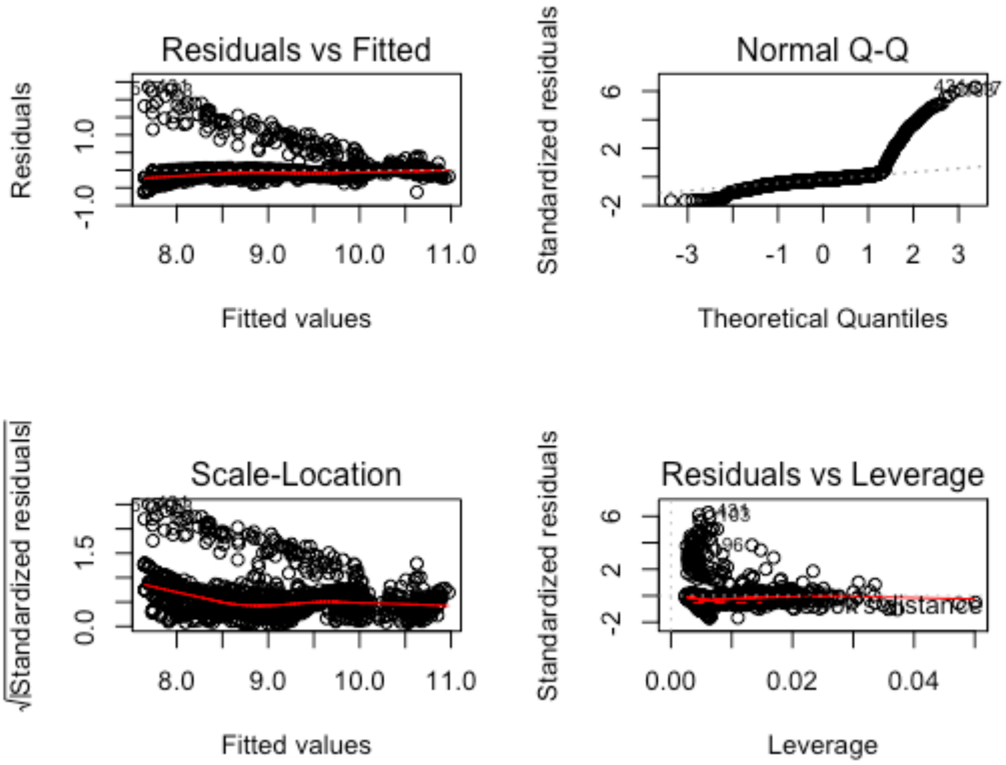

Figure 7.9 Residual Plot

```
> outlierTest(m5)
     rstudent unadjusted p-value Bonferonni p
431  6.367722         2.6381e-10   3.5297e-07
517  6.172712         8.9032e-10   1.1912e-06
103  6.033966         2.0721e-09   2.7724e-06
220  5.717868         1.3310e-08   1.7808e-05
1028 5.668935         1.7608e-08   2.3560e-05
398  5.204898         2.2470e-07   3.0065e-04
```

```
1040 5.117985          3.5424e-07   4.7397e-04
1020 5.092837          4.0359e-07   5.4000e-04
527  5.080943          4.2918e-07   5.7424e-04
341  4.948271          8.4471e-07   1.1302e-03
```

```
# normality test
> shapiro.test(res)

	Shapiro-Wilk normality test

data:  res
W = 0.59439, p-value < 2.2e-16

# => pass!
```

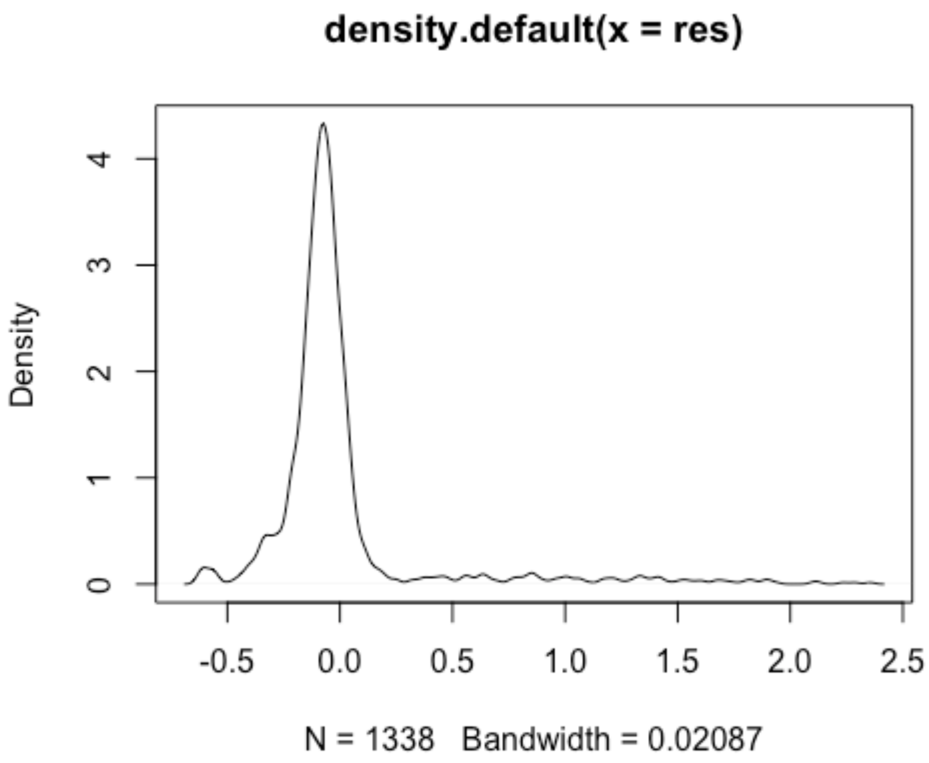

Figure 7.10 Normality Test

```
> # homoscedasticity test
> ncvTest(m5)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 422.9763, Df = 1, p = < 2.22e-16
```

```
=> pass!
```

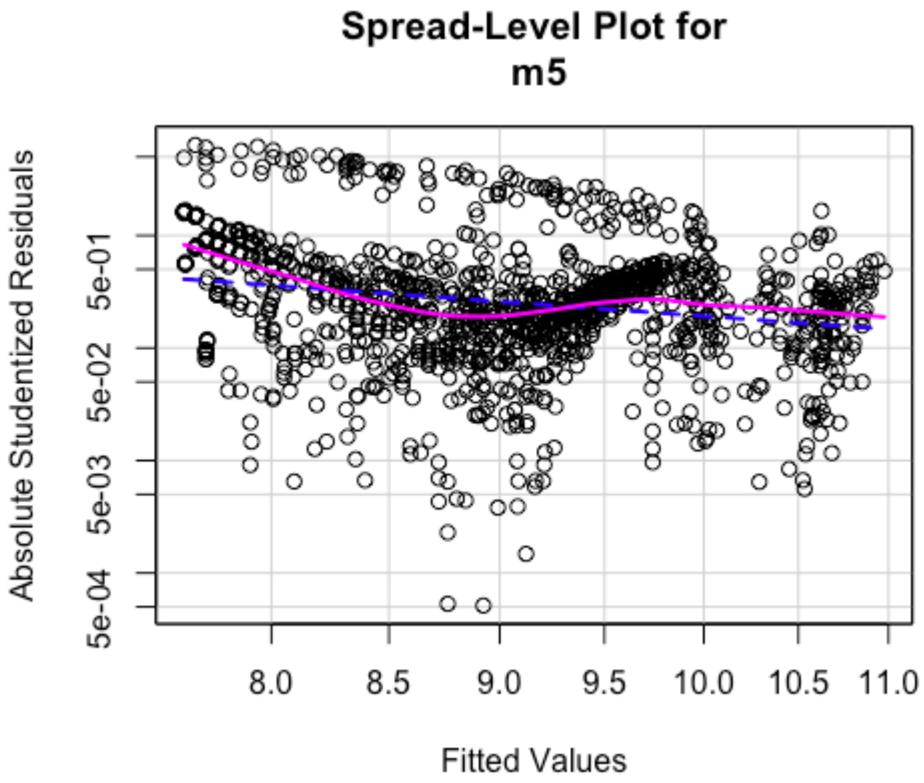

Figure 7.11 Homoscedasticity Test

```
> # multicolinearity test
> vif(m5)
                      GVIF Df GVIF^(1/(2*Df))
age               1.733436  1        1.316600
sex               1.013785  1        1.006869
children         10.312093  1        3.211245
smoker           10.118613  1        3.180977
smoker:bmiOver30  2.161474  2        1.212516
age:children     10.680614  1        3.268121
age:smoker        8.978807  1        2.996466
children:smoker   2.157729  1        1.468921
> sqrt(vif(m5))>10
                  GVIF    Df GVIF^(1/(2*Df))
age              FALSE FALSE           FALSE
sex              FALSE FALSE           FALSE
children         FALSE FALSE           FALSE
smoker           FALSE FALSE           FALSE
smoker:bmiOver30 FALSE FALSE           FALSE
```

```
age:children     FALSE FALSE           FALSE
age:smoker       FALSE FALSE           FALSE
children:smoker  FALSE FALSE           FALSE

=> pass!
```