The Price of Health

Multiple Linear Regression Analysis of Medical Charges Based on Insurance Data

Wenxin Wang, Lingzhi Pan, Xinrui Wang, Zihan Geng, Chen Yang

```
library(mctest)
library(lmtest)
```

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

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

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

```
library(ggplot2)
library(GGally)
```

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

```
library(olsrr)
```

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

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

Dataset

```
df <- read.csv('insurance.csv')
```

```
df$child <- rep(1, 1338)
df$child[df$children==0] <- 0
attach(df)
```

Model

Additive model

Firstly, we will find the additive model which the model only contain the main effects. We first fit the fullmodel and observe the analysis results.

```
model1 <- lm(charges~age+factor(sex)+bmi+factor(child)+factor(smoker)+factor(region))
stepmod=ols_step_both_p(model1,p_enter = 0.05, p_remove = 0.1, details=TRUE)
```

```
## Stepwise Selection Method
## -------------------------
##
## Candidate Terms:
##
## 1. age
## 2. factor(sex)
## 3. bmi
## 4. factor(child)
## 5. factor(smoker)
## 6. factor(region)
##
##
## Step   => 0
## Model  => charges ~ 1
## R2     => 0
##
## Initiating stepwise selection...
##
## Step      => 1
## Selected  => factor(smoker)
## Model     => charges ~ factor(smoker)
## R2        => 0.62
##
## Step      => 2
## Selected  => age
## Model     => charges ~ factor(smoker) + age
## R2        => 0.721
##
## Step      => 3
## Selected  => bmi
## Model     => charges ~ factor(smoker) + age + bmi
## R2        => 0.747
##
## Step      => 4
## Selected  => factor(child)
## Model     => charges ~ factor(smoker) + age + bmi + factor(child)
## R2        => 0.749
##
##
## No more variables to be added or removed.
```

```r
summary(stepmod$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12160.1  -2978.6   -970.7   1488.1  29556.8
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -12149.61     948.52 -12.809  < 2e-16 ***
## factor(smoker)yes  23810.76     411.70  57.835  < 2e-16 ***
## age                  257.94      11.91  21.653  < 2e-16 ***
## bmi                  321.73      27.41  11.738  < 2e-16 ***
## factor(child)1       990.74     335.96   2.949  0.00324 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6075 on 1333 degrees of freedom
## Multiple R-squared:  0.7491, Adjusted R-squared:  0.7484
## F-statistic:   995 on 4 and 1333 DF,  p-value: < 2.2e-16
```

From the output, we can see the model applied Stepwise Regression Procedure with p_enter=0.05 and p_remove=0.1 is:

$$\hat{y} = -12149.61 + 23810.76x_{smoker} + 257.94x_{age} + 321.73x_{bmi} + 990.74x_{child}$$

We will show the screening process of variables throughout the model:

```r
summary(model1)
```

```
##
## Call:
## lm(formula = charges ~ age + factor(sex) + bmi + factor(child) +
##     factor(smoker) + factor(region))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11483.5  -2894.7   -956.5   1478.1  30059.6
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -12001.01     993.99 -12.074  < 2e-16 ***
## age                       256.91      11.92  21.562  < 2e-16 ***
## factor(sex)male          -126.41     333.31  -0.379  0.70456
## bmi                       339.51      28.63  11.858  < 2e-16 ***
## factor(child)1            999.58     335.88   2.976  0.00297 **
```

```
## factor(smoker)yes        23849.65     413.63  57.660  < 2e-16 ***
## factor(region)northwest   -352.22     476.88  -0.739  0.46029
## factor(region)southeast  -1057.33     479.27  -2.206  0.02755 *
## factor(region)southwest   -944.26     478.40  -1.974  0.04861 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6069 on 1329 degrees of freedom
## Multiple R-squared:  0.7503, Adjusted R-squared:  0.7488
## F-statistic: 499.3 on 8 and 1329 DF,  p-value: < 2.2e-16
```

From this output, independent variable sex has $t_{cal} = -0.379$ and $p - value = 0.70456 > 0.05$. We can not reject $H_0 : \beta_{sex} = 0$. We can consider that sex has no influence on insurance charges. We can delete it. Dummy variable region_northwest has $t_{cal} = -0.739$ and $p - value = 0.46029 > 0.05$. We can not reject $H_0 : \beta_{region_northwest} = 0$. But other region dummy variables have $p - value < 0.05$. So we keep the region variable. We can get a reduce model:

```
model1_reduce <- lm(charges~age+bmi+factor(child)+factor(smoker)+factor(region))
summary(model1_reduce)
```

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

From this output, only dummy variable region_northwest has $t_{cal} = -0.737$ and $p - value = 0.4611 > 0.05$. Other variables have $p - value < 0.05$.
We can try to delete region variable and use ANOVA to test the model before and after deletion.

```
model1_nonregion <- lm(charges~age+bmi+factor(child)+factor(smoker))
summary(model1_nonregion)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + factor(child) + factor(smoker))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12160.1  -2978.6   -970.7   1488.1  29556.8
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -12149.61     948.52 -12.809  < 2e-16 ***
## age                  257.94      11.91  21.653  < 2e-16 ***
## bmi                  321.73      27.41  11.738  < 2e-16 ***
## factor(child)1       990.74     335.96   2.949  0.00324 **
## factor(smoker)yes  23810.76     411.70  57.835  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6075 on 1333 degrees of freedom
## Multiple R-squared:  0.7491, Adjusted R-squared:  0.7484
## F-statistic:   995 on 4 and 1333 DF,  p-value: < 2.2e-16
```

After deleting the region variable, the R2 and RSE of the model do not change significantly. We used ANOVA to compare the two models:

```
anova(model1_nonregion, model1_reduce)
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ age + bmi + factor(child) + factor(smoker)
## Model 2: charges ~ age + bmi + factor(child) + factor(smoker) + factor(region)
##   Res.Df        RSS Df Sum of Sq      F  Pr(>F)
## 1   1333 4.9192e+10
## 2   1330 4.8956e+10  3 236131249 2.1383 0.09361 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Source	SS	df	MS	F	p-value
Regression	236131249	3	78710416	2.1383	0.09361
Residual	4.8956×10^{10}	1330	36809023		
Total	4.9192×10^{10}	1333			

From the ANOVA output, $F_{cal} = 2.1383$ with $df = 3, 1330$ and $p - value = 0.09361$. We can not reject $H_0 : \beta_{region} = 0$. We can consider region has no influence on insurance charges. We can delete region variable. Therefore, we can consider that age, bmi, child and smoker have significantly influence on insurance charges.

```
summary(model1_nonregion)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + factor(child) + factor(smoker))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12160.1  -2978.6   -970.7   1488.1  29556.8
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -12149.61     948.52 -12.809  < 2e-16 ***
## age                  257.94      11.91  21.653  < 2e-16 ***
## bmi                  321.73      27.41  11.738  < 2e-16 ***
## factor(child)1       990.74     335.96   2.949  0.00324 **
## factor(smoker)yes  23810.76     411.70  57.835  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6075 on 1333 degrees of freedom
## Multiple R-squared:  0.7491, Adjusted R-squared:  0.7484
## F-statistic:   995 on 4 and 1333 DF,  p-value: < 2.2e-16
```

The final additive model is:

$$\hat{y} = -12149.61 + 257.94x_{age} + 321.73x_{bmi} + 990.74x_{child} + 23810.76x_{smoker}$$

where y is insurance charges;
x_{age} is the individual's age;
x_{bmi} is the body mass index;
x_{child} is a dummy variable. $x_{child} = 1$ is have children; $x_{child} = 0$ is no children;
x_{smoker} is a dummy variable. $x_{smoker} = 1$ is smoker; $x_{smoker} = 0$ is non-smoker;

After we find the final additive model, we will check the multicollinearity.

```
imcdiag(model1_nonregion, method="VIF")
```

```
##
## Call:
## imcdiag(mod = model1_nonregion, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                      VIF detection
## age                1.0149         0
## bmi                1.0122         0
## factor(child)1     1.0025         0
## factor(smoker)yes  1.0008         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
```

```
##
## ==================================
```

All VIFs are close to 1, which means that there is very weak multicollinearity, but it is not severe enough to warrant corrective measures. Therefore, for these main effects we do not need to consider the multicollinearity.

Interaction Model

After we find the additive model and main effects, we consider add some interaction terms into model. First, we add all the interaction variables and observe their significance.

```
model_in <- lm(charges~(age+bmi+factor(child)+factor(smoker))^2)
summary(model_in)
```

```
##
## Call:
## lm(formula = charges ~ (age + bmi + factor(child) + factor(smoker))^2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13690.6  -2056.0  -1248.6   -310.8  29372.1
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -8.727e+02  2.180e+03  -0.400   0.6889
## age                                2.020e+02  4.998e+01   4.042 5.61e-05 ***
## bmi                               -5.717e+01  6.976e+01  -0.820   0.4126
## factor(child)1                     1.755e+03  1.527e+03   1.149   0.2508
## factor(smoker)yes                 -1.960e+04  1.869e+03 -10.486  < 2e-16 ***
## age:bmi                            1.989e+00  1.564e+00   1.272   0.2036
## age:factor(child)1                 5.556e+00  1.950e+01   0.285   0.7757
## age:factor(smoker)yes             -6.733e-02  2.395e+01  -0.003   0.9978
## bmi:factor(child)1                -2.530e+01  4.479e+01  -0.565   0.5723
## bmi:factor(smoker)yes              1.437e+03  5.331e+01  26.953  < 2e-16 ***
## factor(child)1:factor(smoker)yes -1.174e+03  6.736e+02  -1.743   0.0816 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4886 on 1327 degrees of freedom
## Multiple R-squared:  0.8384, Adjusted R-squared:  0.8372
## F-statistic: 688.6 on 10 and 1327 DF,  p-value: < 2.2e-16
```

From the output, only the interaction term $bmi \times smoker$ has $t_{cal} = 26.953$ and $p - value < 0.05$. We can reject $H_0 : \beta_{bmi \times smoker} = 0$. We can consider the interaction term $bmi \times smoker$ has significantly influence on insurance charges. After delete other interaction terms, the reduce interaction model is:

```
model_in1 <- lm(charges ~ age+bmi+factor(child)+factor(smoker)+bmi*factor(smoker))
summary(model_in1)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + factor(child) + factor(smoker) +
##     bmi * factor(smoker))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14992.4  -2033.6  -1242.8   -372.9  29880.8
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -2753.520    838.589  -3.284 0.001052 **
## age                      265.192      9.585  27.667  < 2e-16 ***
## bmi                        6.423     24.950   0.257 0.796894
## factor(child)1           960.640    270.224   3.555 0.000391 ***
## factor(smoker)yes     -20082.183   1659.601 -12.101  < 2e-16 ***
## bmi:factor(smoker)yes   1430.143     52.987  26.991  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4886 on 1332 degrees of freedom
## Multiple R-squared:  0.8378, Adjusted R-squared:  0.8372
## F-statistic:  1376 on 5 and 1332 DF,  p-value: < 2.2e-16
```

From this output about reduce interaction model, all the interaction term is significant.We can compare the R^2_{adj} and RSE between additive and interaction model.

```
data.frame(Model = c( "additive", "interaction"),
           AdjRsq=c(summary(model1_nonregion)$adj.r.squared,summary(model_in1)$adj.r.squared),
           RSE=c(summary(model1_nonregion)$sigma,summary(model_in1)$sigma))
```

```
##         Model    AdjRsq      RSE
## 1    additive 0.7483611 6074.821
## 2 interaction 0.8372067 4886.105
```

The interaction model has a greater $R^2_{adj} = 0.8372067$, and smaller \$RSE=4886.105 \$. We prefer the interaction model. We also compare the additive and interaction models by ANOVA.

```
anova(model1_nonregion, model_in1)
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ age + bmi + factor(child) + factor(smoker)
## Model 2: charges ~ age + bmi + factor(child) + factor(smoker) + bmi *
##     factor(smoker)
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   1333 4.9192e+10
## 2   1332 3.1800e+10  1 1.7392e+10 728.49 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Source	SS	df	MS	F	p-value
Regression	1.7392×10^{10}	1	1.7392×10^{10}	728.49	0
Residual	3.18×10^{10}	1332	2.3874×10^{7}		
Total	4.9192×10^{10}	1333			

From ANOVA table, $F_{cal} = 728.49$ with $df = 1, 1332$ and $p - value < 0.05$. We can reject $H_0 : \beta_{bmi \times smoker} = 0$. Therefore, interaction model is better to fit insurance charges which is:

$$\hat{y} = -2753.520 + 265.192x_{age} + 6.423x_{bmi} + 960.640x_{child} - 20082.183x_{smoker} + 1430.143x_{bmi}x_{smoker}$$

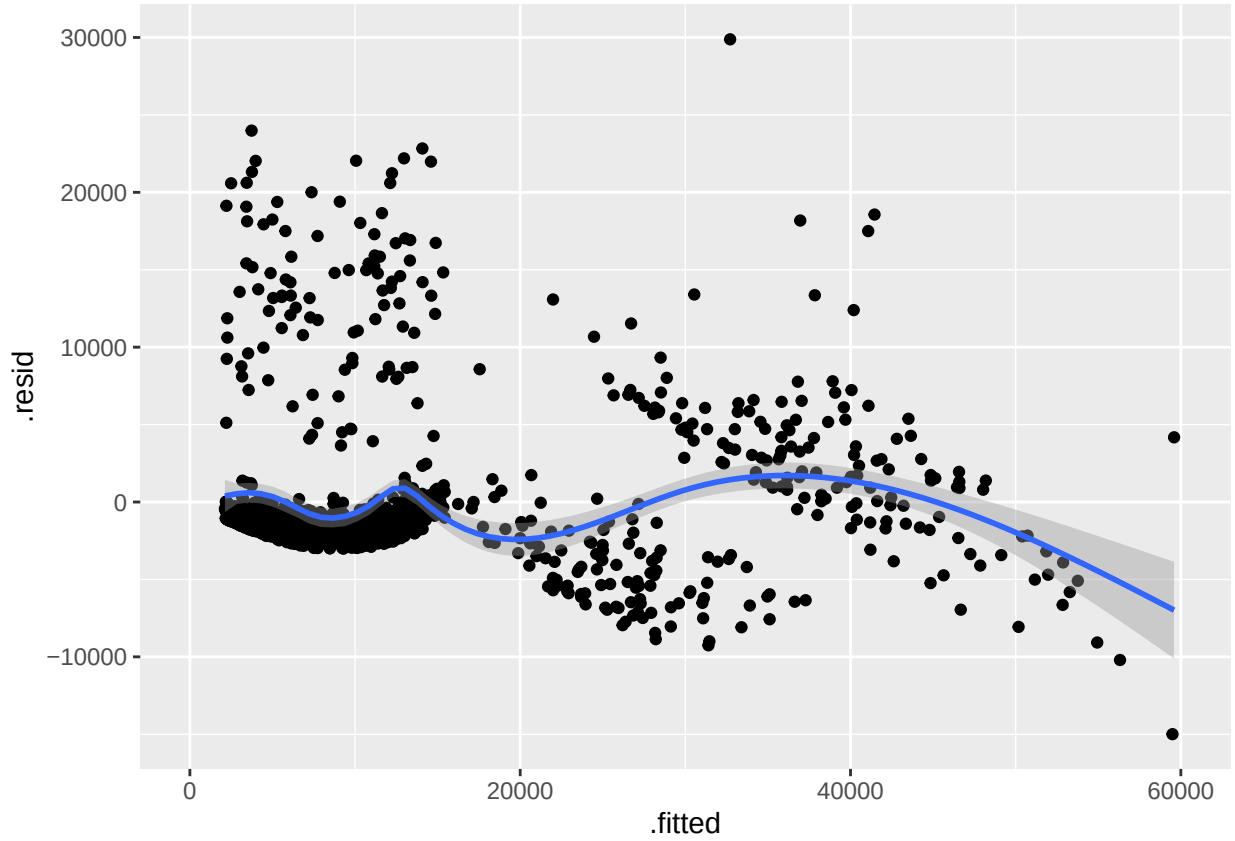
where y is insurance charges;
x_{age} is the individual's age;
x_{bmi} is the body mass index;
x_{child} is a dummy variable. $x_{child} = 1$ is have children; $x_{child} = 0$ is no children;
x_{smoker} is a dummy variable. $x_{smoker} = 1$ is smoker; $x_{smoker} = 0$ is non-smoker;

Assumption

After we find the best fit model, we need to check the assumptions for model. First, we check the linearity assumption:

```
ggplot(model_in1, aes(x=.fitted, y=.resid)) +
  geom_point() +geom_smooth()+xlim(0,60000)
```

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



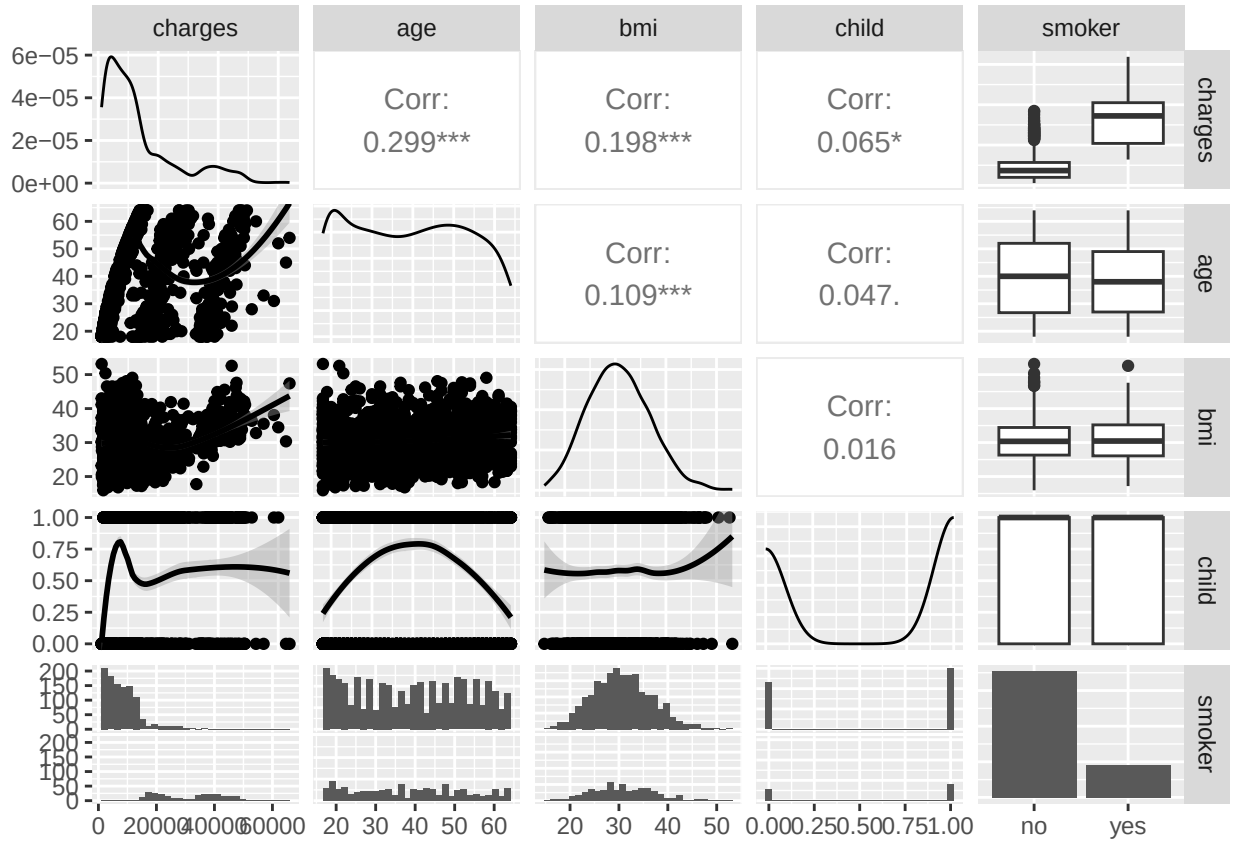

```
  geom_hline(yintercept = 0)
```

```
## mapping: yintercept = ~yintercept
## geom_hline: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

There seems to be some sort of pattern happening with our residuals. So we will use 'ggpairs' draw scatter plots and see if there are higher-order variables.

```
df1 <- df[c('charges','age','bmi','child','smoker')]
ggpairs(df1,lower = list(continuous = "smooth_loess", combo =
  "facethist", discrete = "facetbar", na = "na"))
```

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



From the scatter plots, the relationship between age and charges looks like it might be quadratic or something higher. At the same time, the relationship between bmi and charges looks like it might be quadratic or something higher. We try to add a higher-order variable of age:

```
model_age2 <- lm(charges~age+I(age^2)+bmi+factor(child)+factor(smoker)+bmi*factor(smoker))
summary(model_age2)
```

```
##
## Call:
## lm(formula = charges ~ age + I(age^2) + bmi + factor(child) +
##     factor(smoker) + bmi * factor(smoker))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15523.1  -1863.6  -1277.2   -665.9  30827.8
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.471e+03  1.387e+03   1.781   0.0751 .
## age                     -4.708e+01  6.702e+01  -0.702   0.4825
## I(age^2)                 3.935e+00  8.359e-01   4.707 2.78e-06 ***
## bmi                      3.214e+00  2.476e+01   0.130   0.8968
## factor(child)1           1.469e+03  2.891e+02   5.083 4.24e-07 ***
## factor(smoker)yes       -1.996e+04  1.647e+03 -12.123  < 2e-16 ***
## bmi:factor(smoker)yes    1.427e+03  5.258e+01  27.134  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4848 on 1331 degrees of freedom
## Multiple R-squared:  0.8405, Adjusted R-squared:  0.8398
## F-statistic:  1169 on 6 and 1331 DF,  p-value: < 2.2e-16
```

```
model_age3 <- lm(charges~age+I(age^2)+I(age^3)+bmi+factor(child)+factor(smoker)+bmi*factor(smoker))
summary(model_age3)
```

```
##
## Call:
## lm(formula = charges ~ age + I(age^2) + I(age^3) + bmi + factor(child) +
##     factor(smoker) + bmi * factor(smoker))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15519.5  -1871.5  -1282.0   -656.2  30810.8
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.046e+03  3.594e+03   0.848    0.397
## age                   -9.786e+01  3.003e+02  -0.326    0.745
## I(age^2)               5.294e+00  7.883e+00   0.672    0.502
## I(age^3)              -1.127e-02  6.496e-02  -0.173    0.862
## bmi                    3.180e+00  2.477e+01   0.128    0.898
## factor(child)1         1.471e+03  2.893e+02   5.084 4.22e-07 ***
## factor(smoker)yes     -1.996e+04  1.647e+03 -12.117  < 2e-16 ***
## bmi:factor(smoker)yes  1.427e+03  5.260e+01  27.123  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4850 on 1330 degrees of freedom
## Multiple R-squared:  0.8405, Adjusted R-squared:  0.8396
## F-statistic:  1001 on 7 and 1330 DF,  p-value: < 2.2e-16
```

```
data.frame(Model = c( "age", "age2", "age3"),
           AdjRsq=c(summary(model_in1)$adj.r.squared,summary(model_age2)$adj.r.squared,summary(model_ag
           RSE=c(summary(model_in1)$sigma,summary(model_age2)$sigma,summary(model_age3)$sigma))
```

```
##   Model    AdjRsq      RSE
## 1   age 0.8372067 4886.105
## 2  age2 0.8397518 4847.760
## 3  age3 0.8396350 4849.527
```

After we add x^2_{age} into the interaction model, the R^2_{adj} increase from 0.8372067 to 0.8397518 and the RSE decrease from 4886.105 to 4847.760 . But when we add the x^3_{age}, the R^2_{adj} of the model reduce to 0.8396350 and the RSE is increased to 4849.527. The$p-value$ of x^2_{age} is smaller than 0.05 which is significant. Therefore, we decide to add x^2_{age} into interaction model.
Then, we try to add a higher-order variable of bmi:

```r
model_bmi2 <- lm(charges~age+I(age^2)+bmi+I(bmi^2)+factor(child)+factor(smoker)+bmi*factor(smoker))
summary(model_bmi2)
```

```
##
## Call:
## lm(formula = charges ~ age + I(age^2) + bmi + I(bmi^2) + factor(child) +
##     factor(smoker) + bmi * factor(smoker))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12064.6  -1907.0  -1286.0   -409.6  30510.2
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -5.866e+03  2.820e+03  -2.080 0.037734 *
## age                   -5.184e+01  6.677e+01  -0.776 0.437697
## I(age^2)               3.971e+00  8.327e-01   4.769 2.06e-06 ***
## bmi                    5.605e+02  1.662e+02   3.373 0.000766 ***
## I(bmi^2)              -8.842e+00  2.608e+00  -3.391 0.000717 ***
## factor(child)1         1.485e+03  2.880e+02   5.156 2.91e-07 ***
## factor(smoker)yes     -2.003e+04  1.640e+03 -12.211  < 2e-16 ***
## bmi:factor(smoker)yes  1.430e+03  5.238e+01  27.295  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4829 on 1330 degrees of freedom
## Multiple R-squared:  0.8418, Adjusted R-squared:  0.841
## F-statistic:  1011 on 7 and 1330 DF,  p-value: < 2.2e-16
```

```r
model_bmi3 <- lm(charges~age+I(age^2)+bmi+I(bmi^2)+I(bmi^3)+factor(child)+factor(smoker)+bmi*factor(smo
summary(model_bmi3)
```

```
##
## Call:
## lm(formula = charges ~ age + I(age^2) + bmi + I(bmi^2) + I(bmi^3) +
##     factor(child) + factor(smoker) + bmi * factor(smoker))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8566.6 -1961.7 -1288.0  -417.5 30687.5
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.777e+04  8.693e+03   3.195 0.001433 **
## age                   -4.273e+01  6.642e+01  -0.643 0.520094
## I(age^2)               3.853e+00  8.283e-01   4.651 3.63e-06 ***
## bmi                   -2.788e+03  8.357e+02  -3.337 0.000872 ***
## I(bmi^2)               9.778e+01  2.621e+01   3.731 0.000199 ***
## I(bmi^3)              -1.093e+00  2.674e-01  -4.088 4.61e-05 ***
## factor(child)1         1.490e+03  2.863e+02   5.204 2.25e-07 ***
## factor(smoker)yes     -2.032e+04  1.632e+03 -12.450  < 2e-16 ***
## bmi:factor(smoker)yes  1.439e+03  5.212e+01  27.610  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4800 on 1329 degrees of freedom
## Multiple R-squared:  0.8438, Adjusted R-squared:  0.8429
## F-statistic: 897.4 on 8 and 1329 DF,  p-value: < 2.2e-16
```

```
model_bmi4 <- lm(charges~age+I(age^2)+bmi+I(bmi^2)+I(bmi^3)+I(bmi^4)+factor(child)+factor(smoker)+bmi*fa
summary(model_bmi4)
```

```
##
## Call:
## lm(formula = charges ~ age + I(age^2) + bmi + I(bmi^2) + I(bmi^3) +
##     I(bmi^4) + factor(child) + factor(smoker) + bmi * factor(smoker))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8958.9 -2029.2 -1276.1  -288.5 30525.1
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            8.021e+04  2.839e+04   2.825  0.00480 **
## age                   -3.548e+01  6.645e+01  -0.534  0.59346
## I(age^2)               3.778e+00  8.284e-01   4.561 5.57e-06 ***
## bmi                   -9.772e+03  3.695e+03  -2.644  0.00828 **
## I(bmi^2)               4.336e+02  1.751e+02   2.477  0.01339 *
## I(bmi^3)              -8.037e+00  3.590e+00  -2.239  0.02532 *
## I(bmi^4)               5.217e-02  2.689e-02   1.940  0.05260 .
## factor(child)1         1.471e+03  2.862e+02   5.140 3.16e-07 ***
## factor(smoker)yes     -2.023e+04  1.631e+03 -12.403  < 2e-16 ***
## bmi:factor(smoker)yes  1.436e+03  5.209e+01  27.570  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4796 on 1328 degrees of freedom
## Multiple R-squared:  0.8442, Adjusted R-squared:  0.8432
## F-statistic: 799.8 on 9 and 1328 DF,  p-value: < 2.2e-16
```

We try to add x_{bmi}^2, x_{bmi}^3, x_{bmi}^4 into model. The R_{adj}^2 and RSE about these model are below:

```
data.frame(Model = c( "bmi", "bmi2", "bmi3", "bmi4"),
           AdjRsq=c(summary(model_age2)$adj.r.squared,
                    summary(model_bmi2)$adj.r.squared,
                    summary(model_bmi3)$adj.r.squared,
                    summary(model_bmi4)$adj.r.squared),
           RSE=c(summary(model_age2)$sigma,summary(model_bmi2)$sigma,
                 summary(model_bmi3)$sigma,summary(model_bmi4)$sigma))
```

```
##   Model    AdjRsq      RSE
## 1   bmi 0.8397518 4847.760
## 2  bmi2 0.8410058 4828.755
## 3  bmi3 0.8428621 4800.483
## 4  bmi4 0.8431882 4795.500
```

From the table above, we can find that R^2_{adj} and RSE become smaller as the order of bmi increases, and the smallest $R^2_{adj}=0.8431882$ and $RSE = 4795.5$ are founded when we add x^4_{bmi}. But the $p-value$ of x^4_{bmi} is greater than 0.05 which is not significant. Therefore, we decide to add x^2_{bmi} and x^2_{bmi} into model. The final model after adding high-order variables is as follows:

$$\hat{y} = 27770 - 42.73x_{age} + 3.853x^2_{age} - 2788x_{bmi} + 97.78x^2_{bmi} - 1.093x^3_{bmi} + 1490x_{child} - 20320x_{smoker} + 1439x_{bmi}x_{smoker}$$

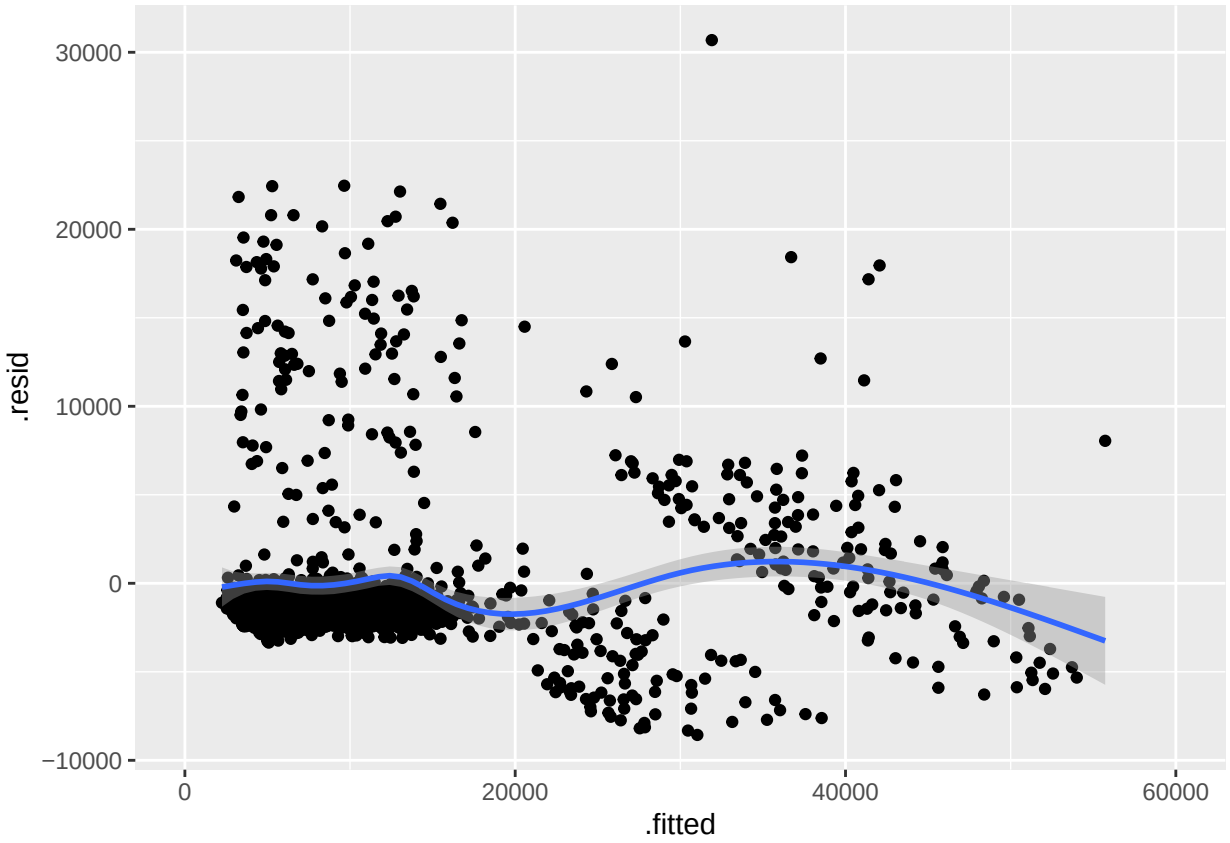
where y is insurance charges;
x_{age} is the individual's age;
x_{bmi} is the body mass index;
x_{child} is a dummy variable. $x_{child} = 1$ is have children; $x_{child} = 0$ is no children;
x_{smoker} is a dummy variable. $x_{smoker} = 1$ is smoker; $x_{smoker} = 0$ is non-smoker;

```
ggplot(model_bmi3, aes(x=.fitted, y=.resid)) +
  geom_point() + geom_smooth()+xlim(0,60000)
```

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

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



```
  geom_hline(yintercept = 0)
```

```
## mapping: yintercept = ~yintercept
## geom_hline: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

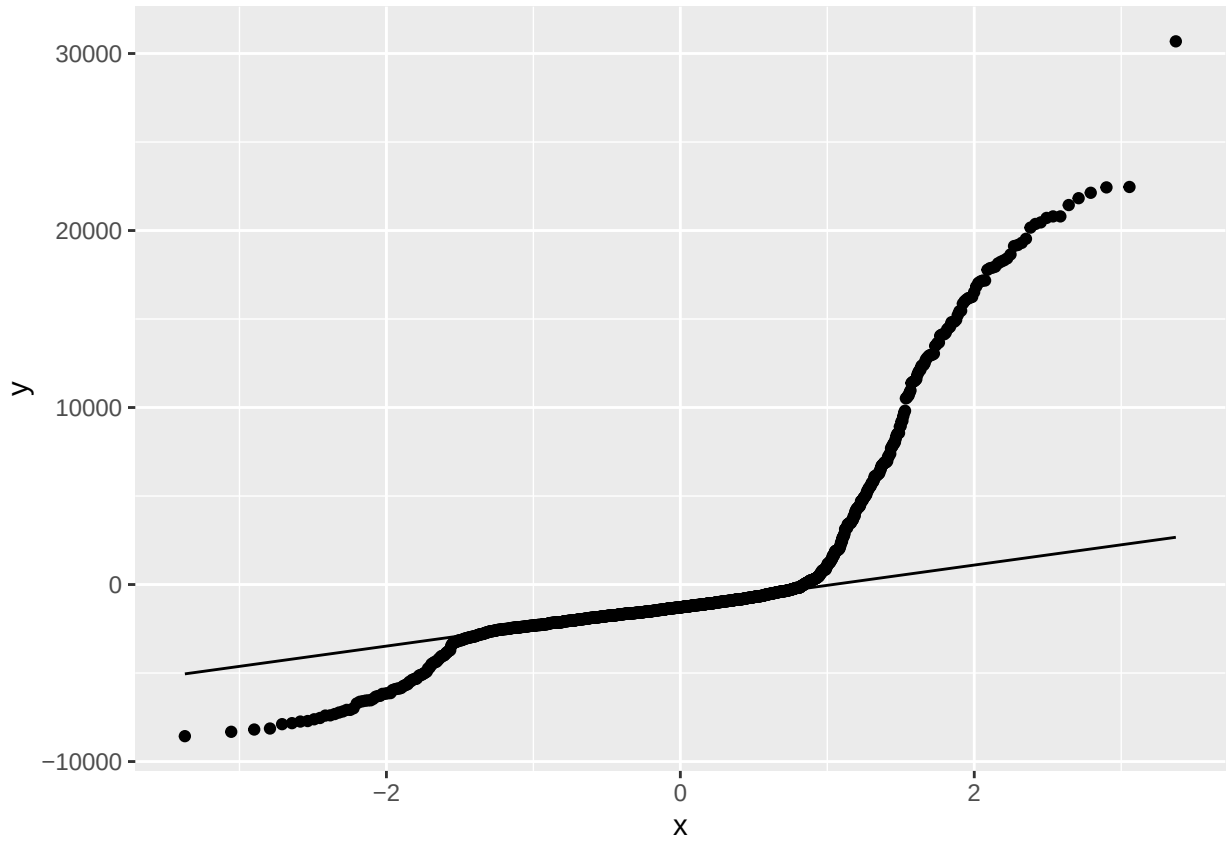
It looks a little smoother than it did before we increased the higher order.
After adding higher-order variables, we test this model for homogeneity of variance and normality assumptions. We use Breusch-Pagan test to test equal variance assumption:
H_0 :heteroscedasticity is not present
H_1 :heteroscedasticity is present

```
bptest(model_bmi3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_bmi3
## BP = 9.4923, df = 8, p-value = 0.3025
```

From the output, $p - value = 0.3025 > 0.05$, we can not reject H_0. Therefore, we can conclude that this model does meet the equal variance assumption.
Then, we draw qqplot and use Shapiro-Wilk test (S-W) to test the normality assumption: H_0 :the sample data are significantly normally distributed H_1 :the sample data are not significantly normally distributed

```
ggplot(df, aes(sample=model_bmi3$residuals)) +
  stat_qq() +
  stat_qq_line()
```

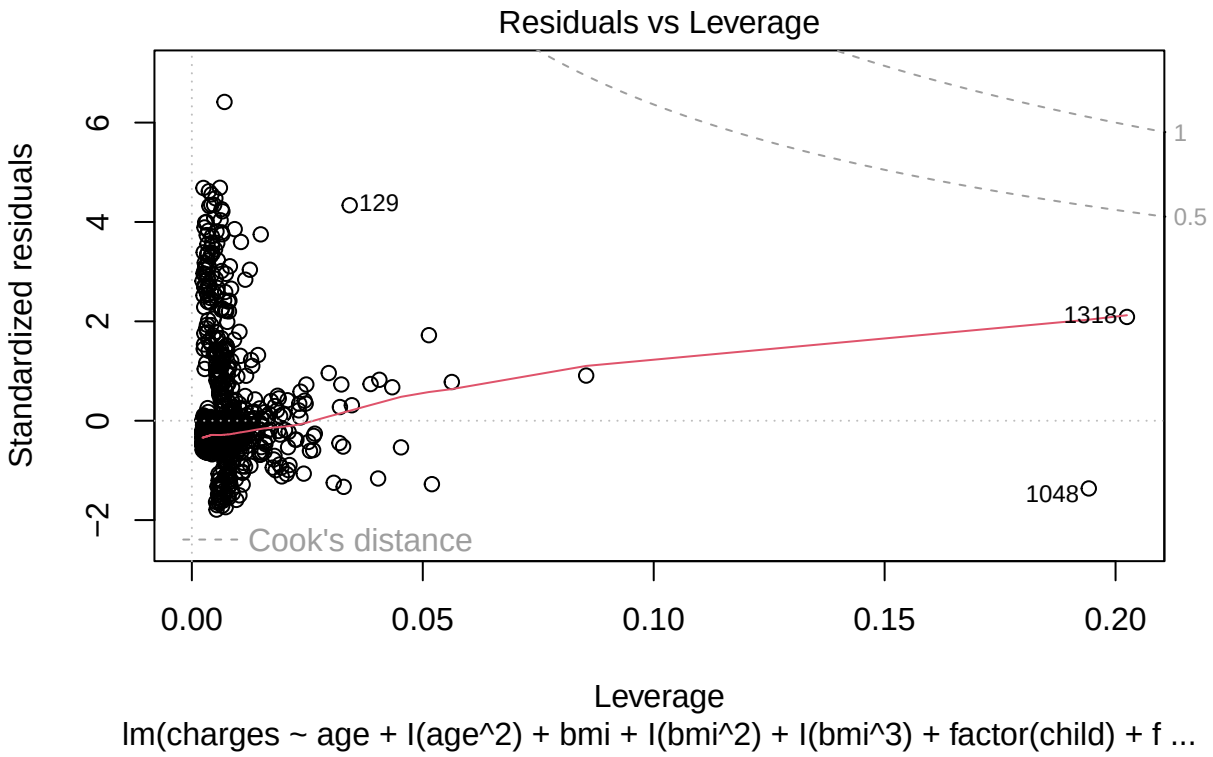



```
shapiro.test(residuals(model_bmi3))
```

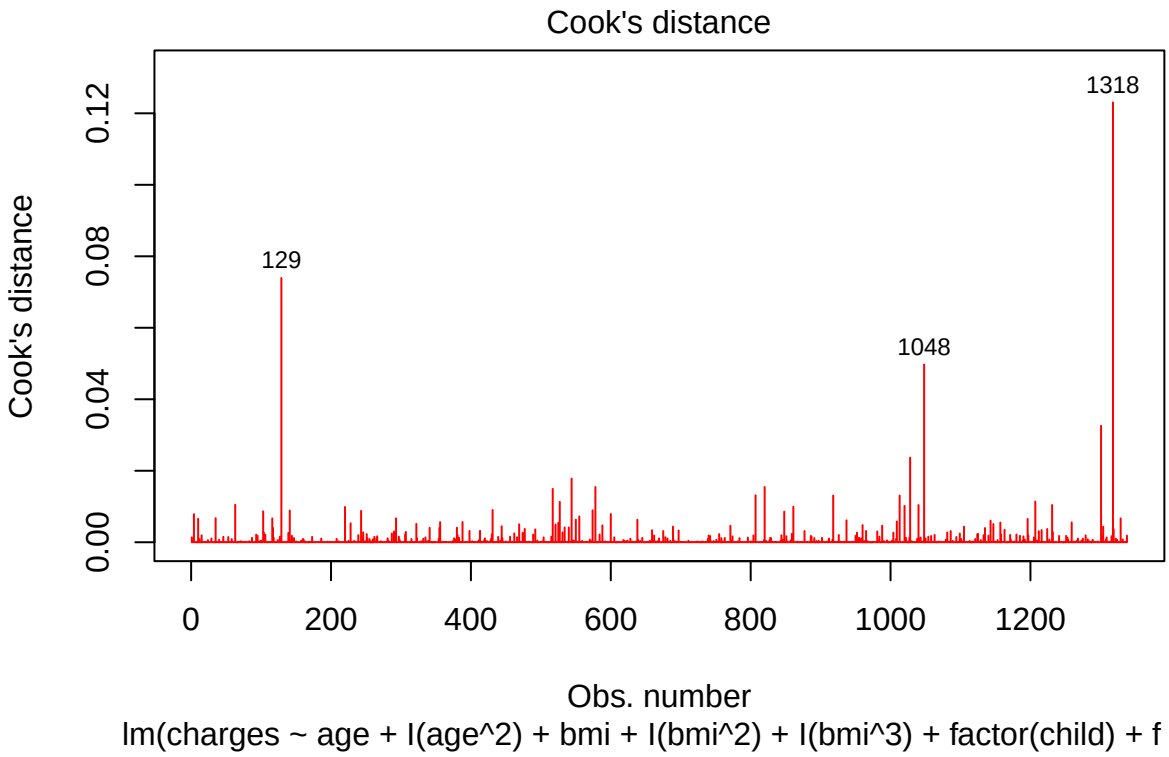
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model_bmi3)
## W = 0.66184, p-value < 2.2e-16
```

From the output, the residual data do not have normal distribution. $p - value = 0.05$. We can reject H_0. We do not have normality.
We can find the outlier points by Cooks distance:

```
plot(model_bmi3,which=5)
```




```
plot(model_bmi3,pch=18,col="red",which=c(4))
```



From the plots above, we can see that data point 129, 1048 and 1318 have a larger cook's distance than other points but smaller than 0.5 and all cases are well inside of the Cook's distance lines. Therefore, there is no influential case. Then, we can find the Leverage points

```
lev=hatvalues(model_bmi3)
p = length(coef(model_bmi3))
n = nrow(df)
outlier2p = lev[lev>(2*p/n)]
outlier3p = lev[lev>(3*p/n)]
print("h_I>3p/n, outliers are")
```

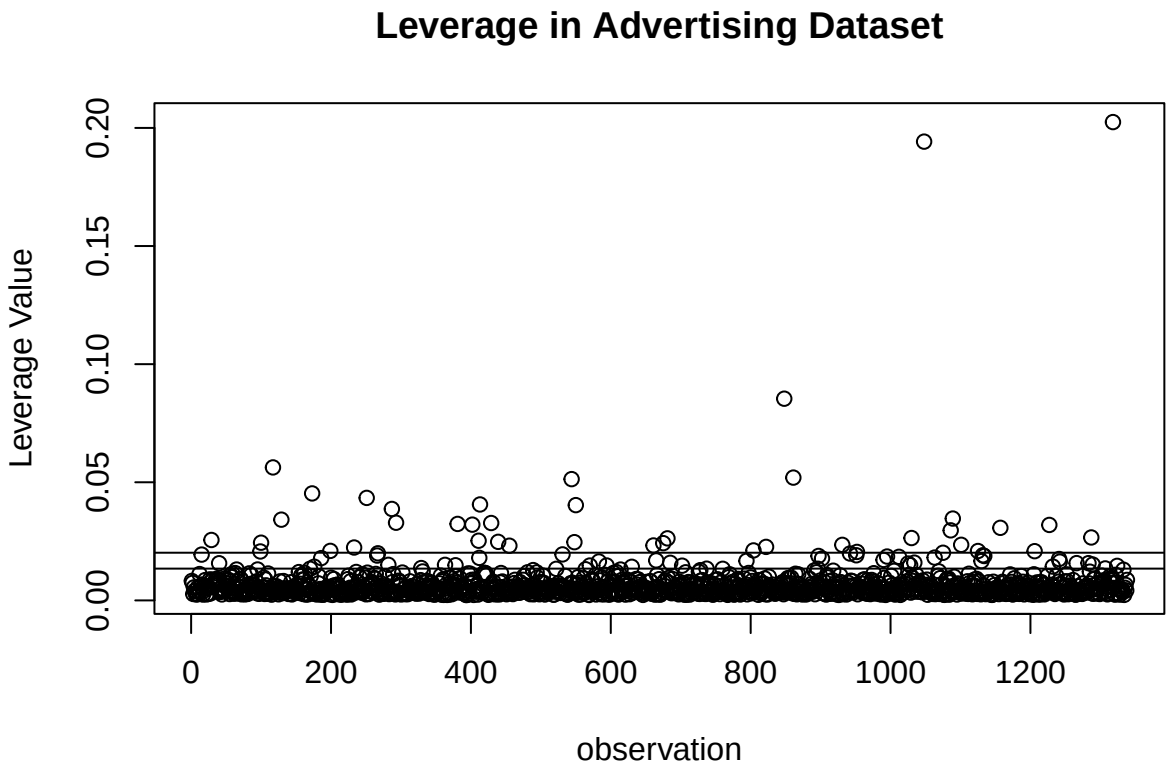
```
## [1] "h_I>3p/n, outliers are"
```

```
print(outlier3p)
```

```
##         29         99        100        117        129        173        199
## 0.02557650 0.02071208 0.02442002 0.05628873 0.03417414 0.04528710 0.02095753
##        233        251        287        293        381        402        411
## 0.02238290 0.04338883 0.03869951 0.03287624 0.03236663 0.03206159 0.02525940
##        413        429        439        455        544        548        550
## 0.04060996 0.03275058 0.02478133 0.02314994 0.05133024 0.02464643 0.04032560
##        661        675        681        804        822        848        861
## 0.02325163 0.02422694 0.02626306 0.02114614 0.02264806 0.08537652 0.05197711
##        931        952       1030       1048       1086       1089       1101
## 0.02351193 0.02052329 0.02639497 0.19419925 0.02964513 0.03464202 0.02353288
```

```
##       1125       1157       1206       1227       1287       1318
## 0.02089923 0.03072183 0.02084025 0.03194576 0.02663214 0.20246511
```

```
plot(rownames(df),lev, main = "Leverage in Advertising Dataset", xlab="observation",
    ylab = "Leverage Value")
abline(h = 2 *p/n, lty = 1)
abline(h = 3 *p/n, lty = 1)
```



There are 41 leverage values.

Final model

Our final model is:

```
summary(model_bmi3)
```

```
##
## Call:
## lm(formula = charges ~ age + I(age^2) + bmi + I(bmi^2) + I(bmi^3) +
##     factor(child) + factor(smoker) + bmi * factor(smoker))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8566.6 -1961.7 -1288.0  -417.5 30687.5
```

```
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.777e+04  8.693e+03   3.195 0.001433 **
## age                   -4.273e+01  6.642e+01  -0.643 0.520094
## I(age^2)               3.853e+00  8.283e-01   4.651 3.63e-06 ***
## bmi                   -2.788e+03  8.357e+02  -3.337 0.000872 ***
## I(bmi^2)               9.778e+01  2.621e+01   3.731 0.000199 ***
## I(bmi^3)              -1.093e+00  2.674e-01  -4.088 4.61e-05 ***
## factor(child)1         1.490e+03  2.863e+02   5.204 2.25e-07 ***
## factor(smoker)yes     -2.032e+04  1.632e+03 -12.450  < 2e-16 ***
## bmi:factor(smoker)yes  1.439e+03  5.212e+01  27.610  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4800 on 1329 degrees of freedom
## Multiple R-squared:  0.8438, Adjusted R-squared:  0.8429
## F-statistic: 897.4 on 8 and 1329 DF,  p-value: < 2.2e-16
```

$$\hat{y} = 27770 - 42.73x_{age} + 3.853x_{age}^2 - 2788x_{bmi} + 97.78x_{bmi}^2 - 1.093x_{bmi}^3 + 1490x_{child} - 20320x_{smoker} + 1439x_{bmi}x_{smoker}$$

where y is insurance charges;
x_{age} is the individual's age;
x_{bmi} is the body mass index;
x_{child} is a dummy variable. $x_{child} = 1$ is have children; $x_{child} = 0$ is no children;
x_{smoker} is a dummy variable. $x_{smoker} = 1$ is smoker; $x_{smoker} = 0$ is non-smoker;

Interpretation

For families non-smokers the model is:

$$\hat{y} = 27770 - 42.73x_{age} + 3.853x_{age}^2 - 2788x_{bmi} + 97.78x_{bmi}^2 - 1.093x_{bmi}^3 + 1490x_{child}$$

which is

$$\hat{y} = \begin{cases} 29260 - 42.73x_{age} + 3.853x_{age}^2 - 2788x_{bmi} + 97.78x_{bmi}^2 - 1.093x_{bmi}^3 & \text{,with child} \\ 27770 - 42.73x_{age} + 3.853x_{age}^2 - 2788x_{bmi} + 97.78x_{bmi}^2 - 1.093x_{bmi}^3 & \text{,without child} \end{cases}$$

The insurance charges (y) vary with age (x_{age}). Both linear (x_{age}) and quadratic (x_{age}^2) terms are included in the model, capturing a nonlinear relationship. Initially, the charges decrease slightly with linear coefficient $\beta_{age} = -42.73$, but they begin to increase more steeply for older individuals as the quadratic coefficient $\beta_{age^2} = 3.853$, reflecting the higher medical risks associated with aging.

The insurance charges (y) change with BMI (x_{bmi}) in a nonlinear pattern, as captured by the linear (x_{bmi}), quadratic (x_{bmi}^2), and cubic (x_{bmi}^3) terms in the model. The negative linear coefficient $\beta_{bmi} = -2788$ suggests that charges decrease slightly at lower BMI levels. However, the positive quadratic coefficient $\beta_{bmi^2} = 97.78$) reflects an upward trend in charges as BMI increases. The small negative cubic coefficient ($\beta_{bmi^3} = -1.093$) further moderates this increase at extremely high BMI levels, reflecting the complex relationship between BMI and medical costs. This pattern aligns with the understanding that both underweight and overweight individuals are at greater health risks, leading to higher insurance charges.

The insurance charges (y) vary about without or with children (x_{child}). The average insurance charge is 1490 higher with children than without children, when other variables are constant. This indicates that having children has a positive effect on insurance charges.

For families with smokers the model is:

$$\hat{y} = 27770 - 20320 - 42.73x_{age} + 3.853x_{age}^2 + (1439 - 2788)x_{bmi} + 97.78x_{bmi}^2 - 1.093x_{bmi}^3 + 1490x_{child}$$

which is

$$\hat{y} = \begin{cases} 8940 + 1490 - 42.73x_{age} + 3.853x_{age}^2 - 1349x_{bmi} + 97.78x_{bmi}^2 - 1.093x_{bmi}^3 & \text{,with child} \\ 7450 - 42.73x_{age} + 3.853x_{age}^2 - 1349x_{bmi} + 97.78x_{bmi}^2 - 1.093x_{bmi}^3 & \text{,without child} \end{cases}$$

The relationship and coefficient of age variable x_{age} between smoking and non-smoking do not change. Both linear (x_{age}) and quadratic (x_{age}^2) terms are included in the model, capturing a nonlinear relationship. the charges decrease slightly with linear coefficient $\beta_{age} = -42.7318$, but they begin to increase more steeply for older individuals as the quadratic coefficient $\beta_{age^2} = 3.8525$.

The relationship of bmi variable x_{bmi} between smoking and non-smoking do not change. But part of the coefficient has changed. The negative linear coefficient increase to$\beta_{bmi} = -1349$ suggests that charges decrease more slightly at lower BMI levels than without smoker. However, the positive quadratic coefficient $\beta_{bmi^2} = 97.78$) reflects an upward trend in charges as BMI increases. The small negative cubic coefficient ($\beta_{bmi^3} = -1.093$) further moderates this increase at extremely high BMI levels.

The insurance charges (y) vary about without or with children (x_{child}). The average insurance charge is still 1490 higher with children than without children, when other variables are constant.

Smoking and non-smoking do not change the relationship and coefficients between age x_{age}, child x_{child} variables and insurance charges. But it will change the intercept and the linear coefficient of bmi. The intercept is reduced and the coefficient on bmi is increased.