Keosotra Veng
Stat 6620-02

Homework 5

Step 1: Step 1 - collecting data

- We got the data from: http://www.sci.csueastbay.edu/~esuess/classes/Statistics_6620/Presentations/ml10/insurance.csv

Step 2: Exploring and preparing the data ----

- Load the data into R

```
insurance <- read.csv("http://www.sci.csueastbay.edu/~esuess/classes/Statistics_6620/Presentations/ml10/insurance.csv", stringsAsFactors = TRUE)
str(insurance)
```

```
## 'data.frame':    1338 obs. of  7 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 25.7 33.4 27.7 29.8 25.8 ...
##  $ 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 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ expenses: num  16885 1726 4449 21984 3867 ...
```

- Summarize the expense variable with the mean of $9382

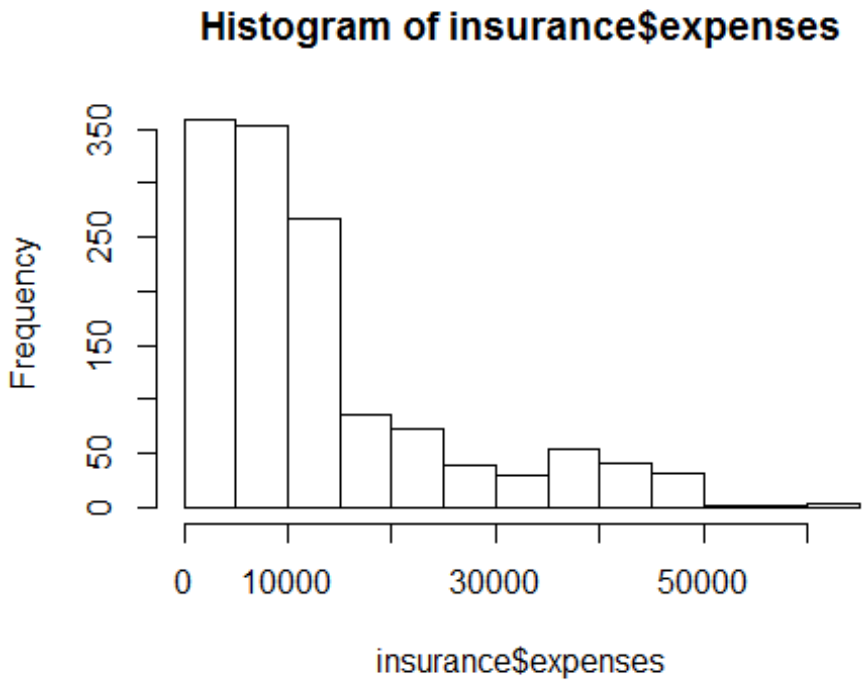
```
summary(insurance$expenses)
```

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

- Create a histogram of insurance expense. The histogram is skewed to the right.

```
hist(insurance$expenses)
```

Keosotra Veng
Stat 6620-02



- Create a table with a list of all the regions.

```
table(insurance$region)

##
## northeast northwest southeast southwest
##       324       325       364       325
```

- We look at the correlation between age, bmi, children, and exenpses using correlation matrix

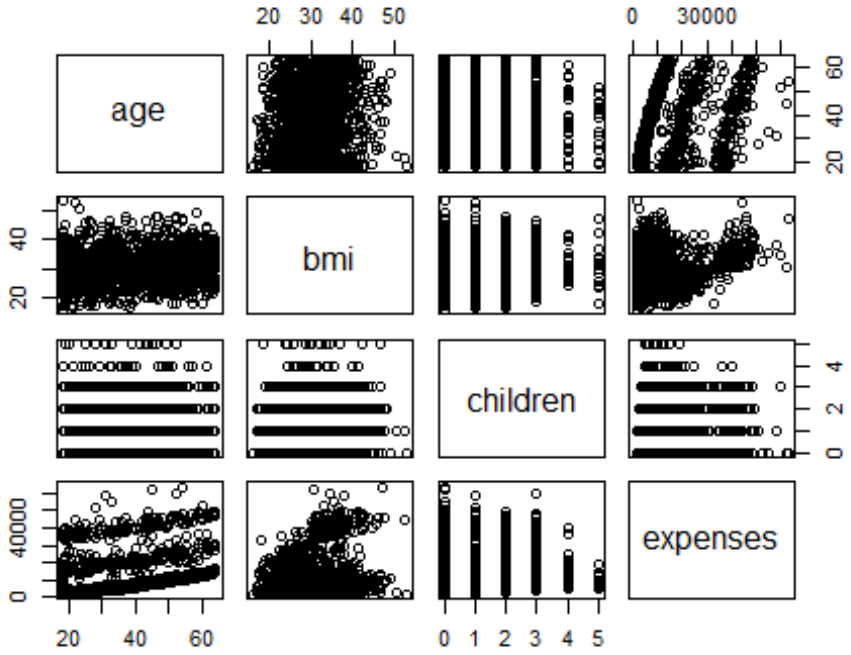
```
cor(insurance[c("age", "bmi", "children", "expenses")])

##                 age        bmi   children   expenses
## age      1.0000000 0.10934101 0.04246900 0.29900819
## bmi      0.1093410 1.00000000 0.01264471 0.19857626
## children 0.0424690 0.01264471 1.00000000 0.06799823
## expenses 0.2990082 0.19857626 0.06799823 1.00000000
```

- Visualing relationships among features: scatterplot matrix

```
pairs(insurance[c("age", "bmi", "children", "expenses")])
```

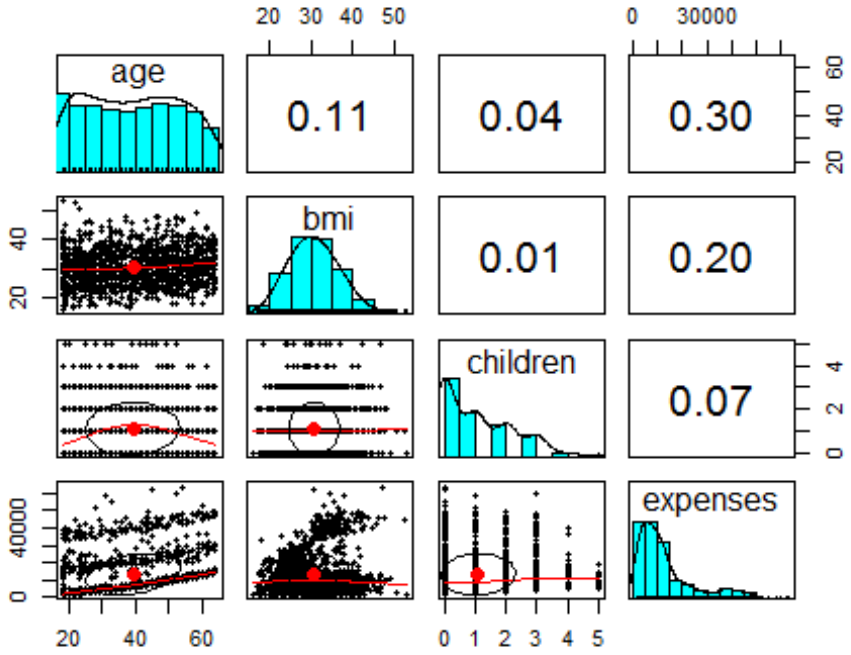
Keosotra Veng
Stat 6620-02



- more informative scatterplot matrix

```
library(psych)
pairs.panels(insurance[c("age", "bmi", "children", "expenses")])
```

Keosotra Veng
Stat 6620-02



Step 3: Training a model on the data ---- - Run linear regression with expenses as dependent varaible and age, children, bmi, sex, smoker, region as independent variables.

```
ins_model <- lm(expenses ~ age + children + bmi + sex + smoker + region,
                data = insurance)
```

- Lookat at the estimated beta coefficients

```
ins_model
```

```
##
## Call:
## lm(formula = expenses ~ age + children + bmi + sex + smoker +
##     region, data = insurance)
##
## Coefficients:
##     (Intercept)              age         children              bmi
##        -11941.6            256.8            475.7            339.3
##         sexmale        smokeryes  regionnorthwest  regionsoutheast
##          -131.4          23847.5           -352.8          -1035.6
## regionsouthwest
##          -959.3
```

Step 4: Evaluating model performance ----

- See more detail about the estimated beta coefficients. Our adjusted R-Squared is 0.7494

```
summary(ins_model)
```

```
##
## Call:
## lm(formula = expenses ~ age + children + bmi + sex + smoker +
##     region, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11302.7  -2850.9   -979.6   1383.9  29981.7
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11941.6      987.8 -12.089  < 2e-16 ***
## age                256.8       11.9  21.586  < 2e-16 ***
## children           475.7      137.8   3.452 0.000574 ***
## bmi                339.3       28.6  11.864  < 2e-16 ***
## sexmale           -131.3      332.9  -0.395 0.693255
## smokeryes        23847.5      413.1  57.723  < 2e-16 ***
## regionnorthwest   -352.8      476.3  -0.741 0.458976
## regionsoutheast  -1035.6      478.7  -2.163 0.030685 *
## regionsouthwest   -959.3      477.9  -2.007 0.044921 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.9 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Step 5: Improving model performance ----

- add a higher-order "age" term

```
insurance$age2 <- insurance$age^2
```

- add an indicator for BMI >= 30

```
insurance$bmi30 <- ifelse(insurance$bmi >= 30, 1, 0)
```

- Create final model. We can see that our R-squared has been improved with the new model to 0.8653

```
ins_model2 <- lm(expenses ~ age + age2 + children + bmi + sex +
                   bmi30*smoker + region, data = insurance)

summary(ins_model2)
```

```
##
## Call:
## lm(formula = expenses ~ age + age2 + children + bmi + sex + bmi30 *
##     smoker + region, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17297.1  -1656.0  -1262.7   -727.8  24161.6
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       139.0053  1363.1359   0.102 0.918792
## age               -32.6181    59.8250  -0.545 0.585690
## age2                3.7307     0.7463   4.999 6.54e-07 ***
## children          678.6017   105.8855   6.409 2.03e-10 ***
## bmi               119.7715    34.2796   3.494 0.000492 ***
## sexmale          -496.7690   244.3713  -2.033 0.042267 *
## bmi30            -997.9355   422.9607  -2.359 0.018449 *
## smokeryes       13404.5952   439.9591  30.468  < 2e-16 ***
## regionnorthwest  -279.1661   349.2826  -0.799 0.424285
## regionsoutheast  -828.0345   351.6484  -2.355 0.018682 *
## regionsouthwest -1222.1619   350.5314  -3.487 0.000505 ***
## bmi30:smokeryes 19810.1534   604.6769  32.762  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4445 on 1326 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8653
## F-statistic: 781.7 on 11 and 1326 DF,  p-value: < 2.2e-16
```