Statistical Learning HW1

Yuheng Ma

3/30/2020

There is seven questions from **An Introduction to Statistical Learning with Applications in R** by G. James that we should answer by this task.

1. Is there a relationship between advertising budget and sales? Our first goal should be to determine whether the data provide evidence of an association between advertising expenditure and sales. If the evidence is weak, then one might argue that no money should be spent on advertising!
2. How strong is the relationship between advertising budget and sales? Assuming that there is a relationship between advertising and sales, we would like to know the strength of this relationship. In other words, given a certain advertising budget, can we predict sales with a high level of accuracy? This would be a strong relationship. Or is a prediction of sales based on advertising expenditure only slightly better than a random guess? This would be a weak relationship.
3. Which media contribute to sales? Do all three media—TV, radio, and newspaper—contribute to sales, or do just one or two of the media contribute? To answer this question, we must find a way to separate out the individual effects of each medium when we have spent money on all three media.
4. How accurately can we estimate the effect of each medium on sales? For every dollar spent on advertising in a particular medium, by what amount will sales increase? How accurately can we predict this amount of increase?
5. How accurately can we predict future sales? For any given level of television, radio, or newspaper advertising, what is our prediction for sales, and what is the accuracy of this prediction?
6. Is the relationship linear? If there is approximately a straight-line relationship between advertis- ing expenditure in the various media and sales, then linear regression is an appropriate tool. If not, then it may still be possible to trans- form the predictor or the response so that linear regression can be used.
7. Is there synergy among the advertising media? Perhaps spending $50,000 on television advertising and $50,000 on radio advertising results in more sales than allocating $100,000 to either television or radio individually. In marketing, this is known as a synergy effect, while in statistics it is called an interaction effect.

The data we use is **Credit** from ISLR, data package attached to the book. We shall answer these questions by a linear model.

Main Code

```
# The regression-test function filed up for further call
linreg<-function(type,training,test,training0=0,residualplot=0){
  # regress everything
  if (type=="simple"){
    lmodel=lm(Balance ~    Income+Limit+Rating+Cards+Age+Education+Gender+Married+
                Student+Ethnicity,training)
    print(summary(lmodel))
    par(mfrow=c(2,2))
    plot(lmodel)
    result<-predict(lmodel,test)
    result[result<0]=0
    # squared residue
    error=result-test$Balance
```

```
    error=error*error
    # divided by degree of freedom
    print("MSE:")
    print(sum(error)/length(error))
  }

  else if(type=="select"){
    lmodel=lm(Balance ~    Income+Limit+Rating+Cards+Age+Education+Gender+Married+
                Student+Ethnicity,training)
    #using step variable selection
    lmodel=step(lmodel)
    print(summary(lmodel))
    par(mfrow=c(2,2))
    plot(lmodel)
    result<-predict(lmodel,test)
    result[result<0]=0
    error=result-test$Balance
    error=error*error
    print("MSE:")
    print(sum(error)/length(error))
  }
  else if(type=="cutoff"){
    # now training set is actually training_1
    lmodel=lm(Balance ~    Income+Limit+Rating+Cards+Age+Education+Gender+Married+
                Student+Ethnicity,training)
    lmodel=step(lmodel,trace = 0)
    if (residualplot==0){
    print(summary(lmodel))
    par(mfrow=c(2,2))
    plot(lmodel)
    # validate how well training_0 is predicted
    validate<-predict(lmodel,training0)
    validate[validate<0]=0
    # print rate of correction
    print("Correct Rate:")
    print(length(validate[validate==0])/length(validate))
    result<-predict(lmodel,test)
    result[result<0]=0
    error=result-test$Balance
    error=error*error
    print("MSE:")
    print(sum(error)/length(error))
    }
    if (residualplot==1){
    par(mfrow=c(2,3))
    for (i in 2:7){
      plot(training[[i]],lmodel$residuals)
    }
    }
  }
  else if(type=="cutoffinter"){
    # now training set is actually training_1
    lmodel=lm(Balance ~ (Income+Limit+Rating+Cards+Age+Education+Gender+Married+
```

```
    Student+Ethnicity)*(Income+Limit+Rating+Cards+Age+Education+Gender+Married+
                          Student+Ethnicity),training)
    lmodel=step(lmodel,trace = 0)
    if (residualplot==0){
    print(summary(lmodel))
    par(mfrow=c(2,2))
    plot(lmodel)
    # validate how well training_0 is predicted
    validate<-predict(lmodel,training0)
    validate[validate<0]=0
    # print rate of correction
    print("Correct Rate:")
    print(length(validate[validate==0])/length(validate))
    result<-predict(lmodel,test)
    result[result<0]=0
    error=result-test$Balance
    error=error*error
    print("MSE:")
    print(sum(error)/length(error))
    }
    if (residualplot==1){
    par(mfrow=c(2,3))
    for (i in 2:7){
      plot(training[[i]],lmodel$residuals)
    }
    }
  }
  else if(type=="printer"){
    # now training set is actually training_1
    lmodel=lm(Balance ~    Income+Limit+Rating+Cards+Age+Education+Gender+Married+
                Student+Ethnicity,training)
    lmodel=step(lmodel,trace = 0)
  summary(lmodel)
  }
  else if(type=="confint"){
    # now training set is actually training_1
    lmodel=lm(Balance ~    Income+Limit+Rating+Cards+Age+Education+Gender+Married+
                Student+Ethnicity,training)
    lmodel=step(lmodel,trace = 0)
  confint(lmodel)
  }
}
```

Data Exploration

```
library(ISLR)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

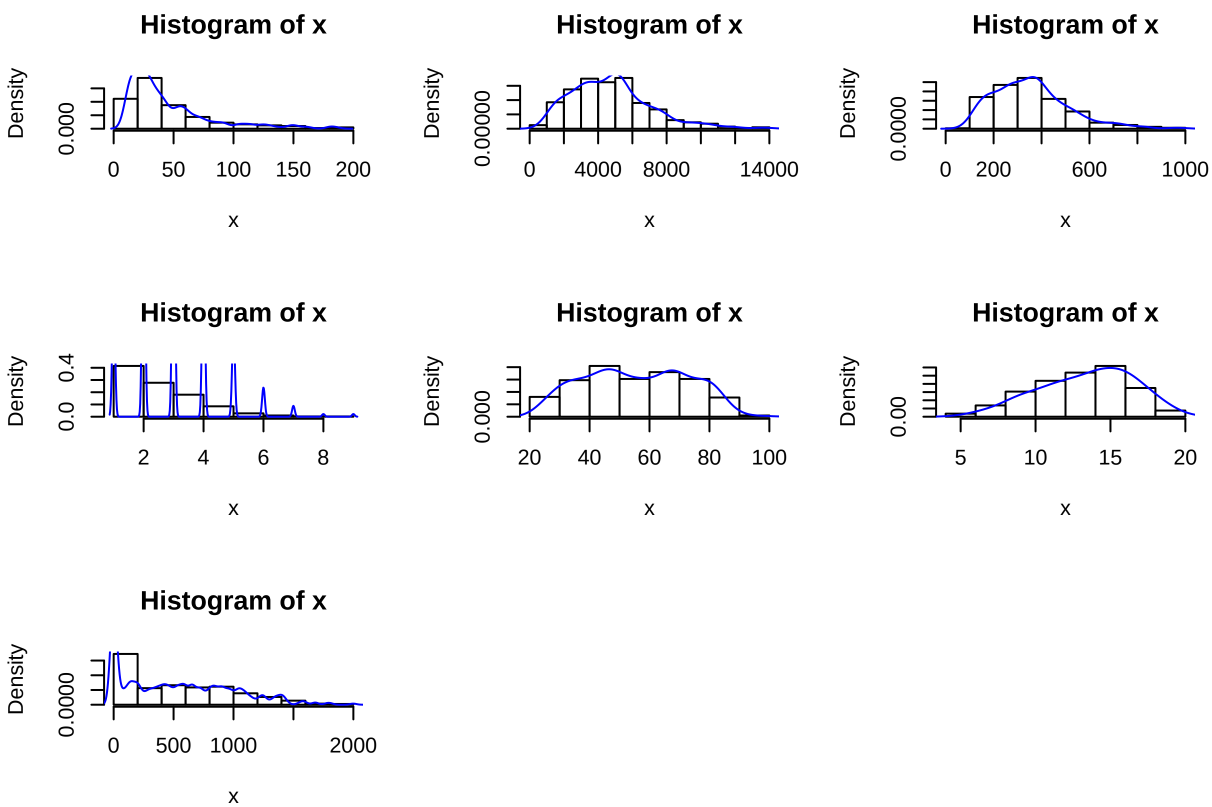
```
data("Credit")
str(Credit)
```

```
## 'data.frame':    400 obs. of  12 variables:
```

```
## $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
## $ Income   : num  14.9 106 104.6 148.9 55.9 ...
## $ Limit    : int  3606 6645 7075 9504 4897 8047 3388 7114 3300 6819 ...
## $ Rating   : int  283 483 514 681 357 569 259 512 266 491 ...
## $ Cards    : int  2 3 4 3 2 4 2 2 5 3 ...
## $ Age      : int  34 82 71 36 68 77 37 87 66 41 ...
## $ Education: int  11 15 11 11 16 10 12 9 13 19 ...
## $ Gender   : Factor w/ 2 levels " Male","Female": 1 2 1 2 1 1 2 1 2 2 ...
## $ Student  : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 2 ...
## $ Married  : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 1 1 1 2 ...
## $ Ethnicity: Factor w/ 3 levels "African American",..: 3 2 2 2 3 3 1 2 3 1 ...
## $ Balance  : int  333 903 580 964 331 1151 203 872 279 1350 ...
```

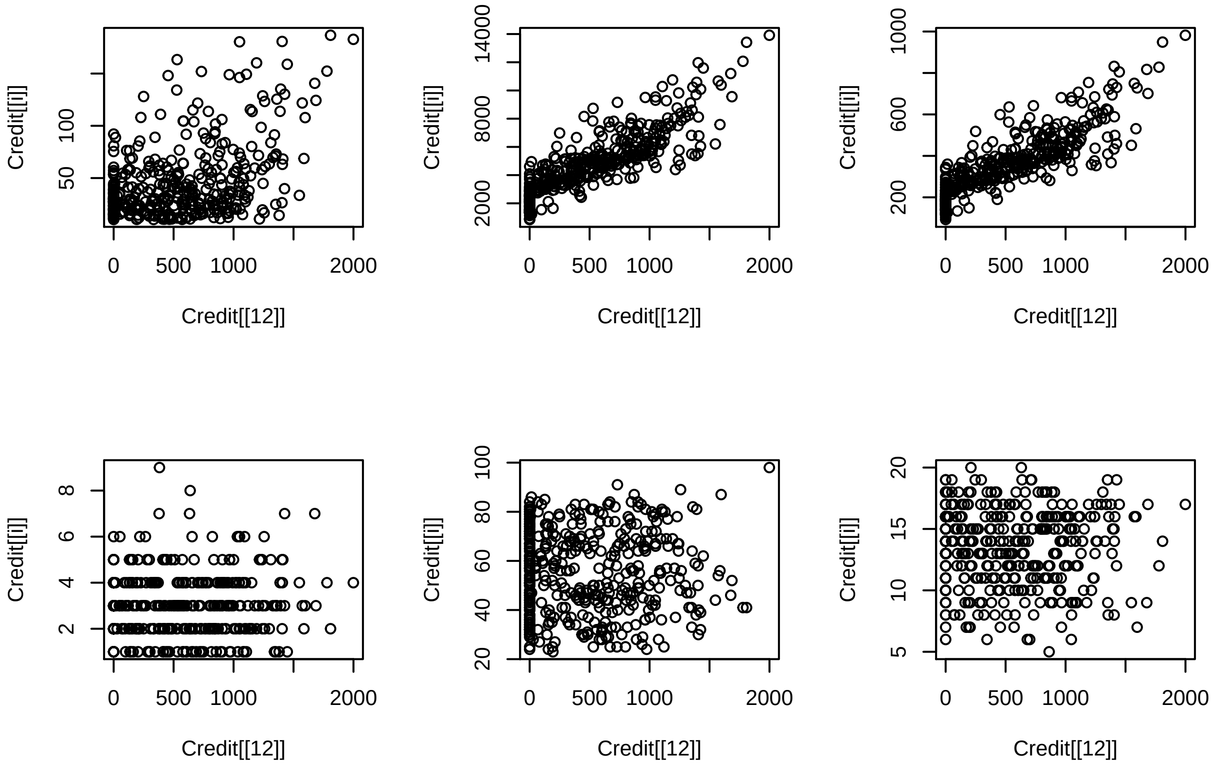
We take a look at the distribution of numerical variables.

```r
par(mfrow=c(3,3))
for (i in c(2:7,12)){
  x=Credit[[i]]
hist(x,probability=T)
d<-density(x, bw = "sj")
lines(d,col="blue")
}
```



It seems apparent that while other data is of some usual distribution, such as poisson or normal, the data of balance, which has a mode at balance=0, is clearly ill-posed. The reason is that, balance can't be negative. People with great tendency not to borrow from bank will have the same balance, zero, with people have less but positive tendency. Thus, this is a cut-off data. This can also be seen if we make paired plot of the variables.

```r
par(mfrow=c(2,3))
for (i in c(2:7)){
  plot(Credit[[12]],Credit[[i]])
}
```



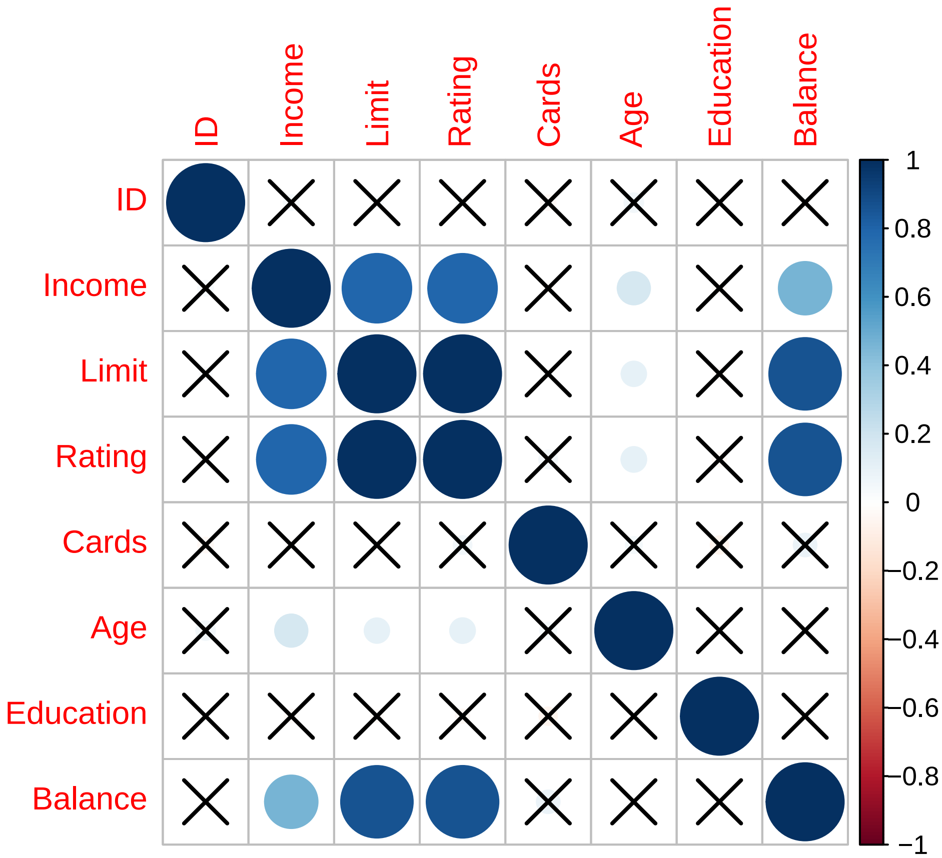
Note that there is a clear cut-off at Balance=0. Knowing this, the problem becomes a multi-region linear regression. Thus we design the learning process as

1. Separate the data set into training set, X (250 data) and test set X_T (150 data). Separate training set into $X_0 = \{x|x \in X, x\$balance = 0\}$ and $X_1 = \{x|x \in X, x\$balance > 0\}$.

2. Do linear regression on X_1 and get a 95% correct rate on X_0.

Also, we might want to check the relationship between variables.

```r
res1 <- cor.mtest(Credit[c(1:7,12)], conf.level = 0.95)
corrplot(cor(Credit[c(1:7,12)]),sig.level = .05,p.mat = res1$p)
```



What we have done above is looking at the data mindlessly, which gives the information that some of the variables have a strong correlation.

Linear Regression

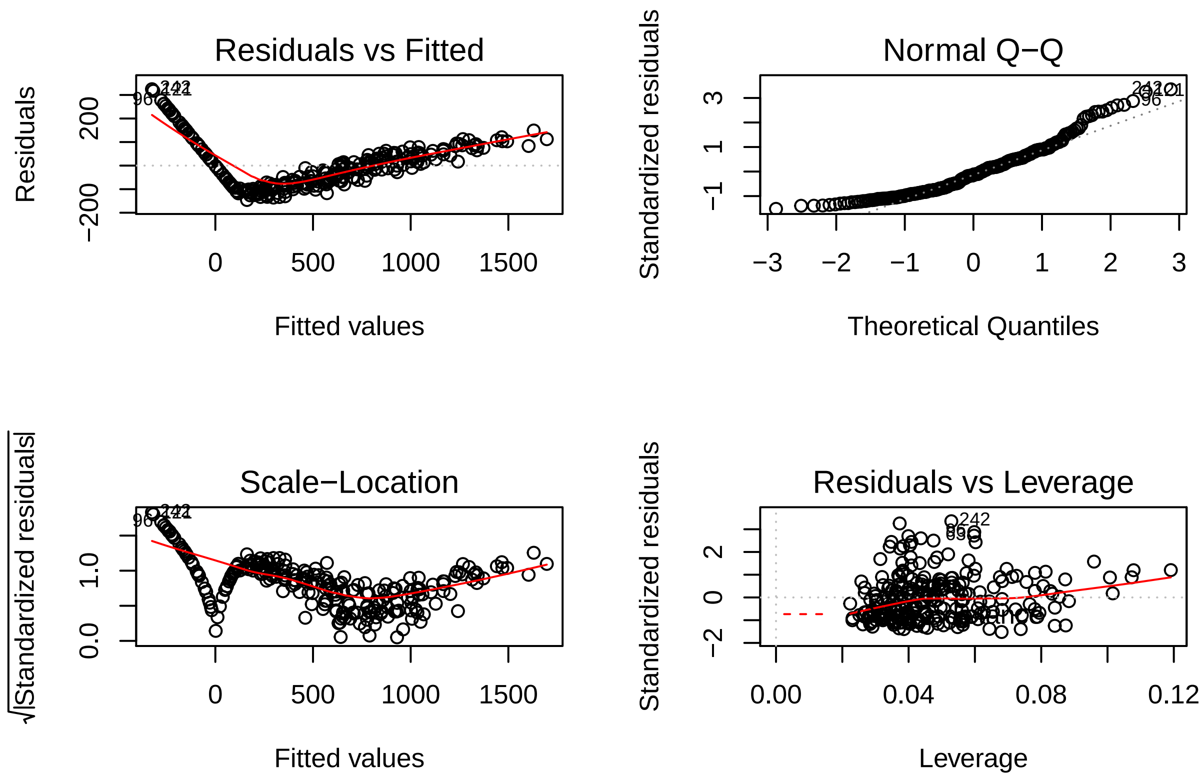
Next, we shall separate training set and test set and build predicting model step by step to see the variation of MSE. We pile up this process in a function.

```
# separating data set
training=Credit[1:250,]
training0<-training[training$Balance==0,]
training1<-training[training$Balance>0,]
test=Credit[251:400,]
```

```
# do silly regression
linreg("simple",training,test)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
##     Education + Gender + Married + Student + Ethnicity, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -145.94  -79.20  -13.11   52.63  324.36
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)        -484.34791   46.93587 -10.319  < 2e-16 ***
## Income               -7.92212    0.31036 -25.526  < 2e-16 ***
## Limit                 0.21849    0.04126   5.296 2.69e-07 ***
## Rating                0.78224    0.61634   1.269 0.205619    
## Cards                19.84145    5.72410   3.466 0.000626 ***
## Age                  -0.62144    0.37093  -1.675 0.095180 .  
## Education            -1.60303    2.12815  -0.753 0.452044    
## GenderFemale        -10.64963   12.73921  -0.836 0.404009    
## MarriedYes            5.38461   13.18787   0.408 0.683421    
## StudentYes          438.81700   20.21121  21.712  < 2e-16 ***
## EthnicityAsian        7.45505   17.23527   0.433 0.665736    
## EthnicityCaucasian    5.10920   15.33268   0.333 0.739260    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 99.41 on 238 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9555 
## F-statistic: 487.5 on 11 and 238 DF,  p-value: < 2.2e-16
```




```
## [1] "MSE:"
## [1] 3877.899
```

We see that the MSE now is $\frac{1}{n}\sum(balance - bal\hat{a}nce)^2$=3877.899, while we directly regress everything.

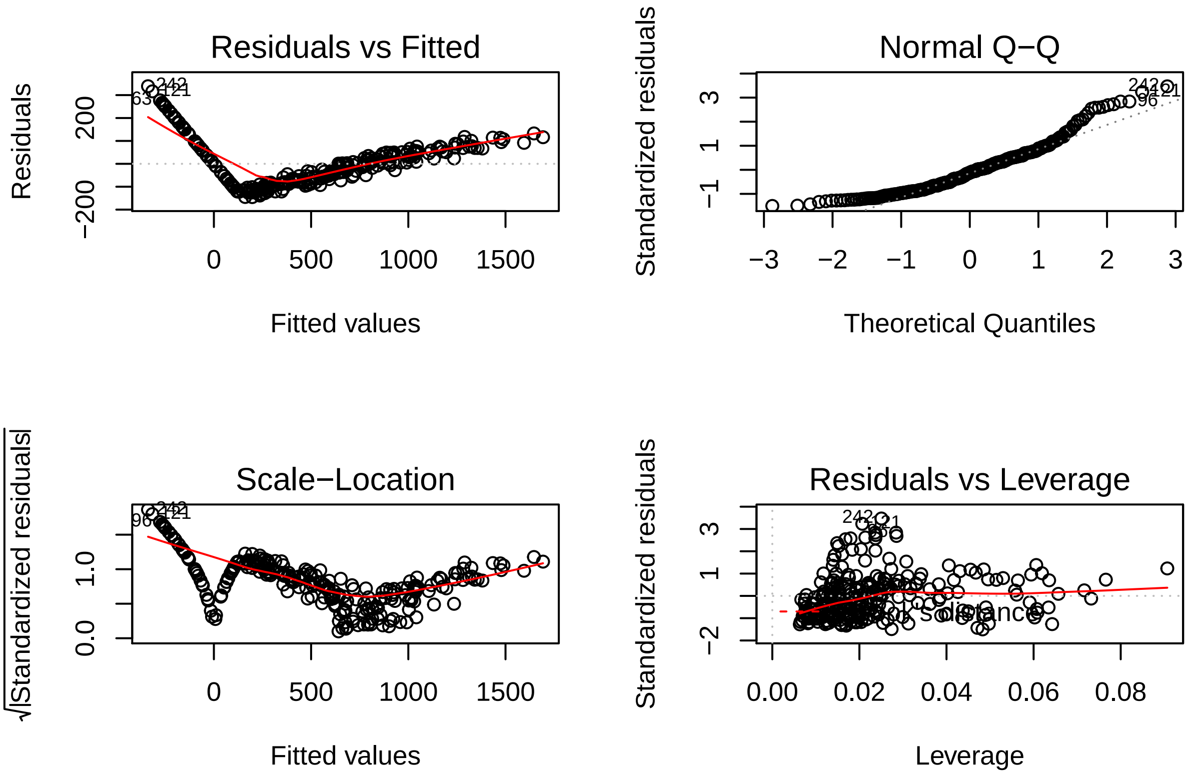
Then we plug in variable selection process.

```r
# do variable selection
linreg("select",training,test)
```

```
## Start:  AIC=2311.33
## Balance ~ Income + Limit + Rating + Cards + Age + Education + 
```

```
##     Gender + Married + Student + Ethnicity
##
##             Df Sum of Sq     RSS    AIC
## - Ethnicity  2      1967 2353992 2307.5
## - Married    1      1647 2353673 2309.5
## - Education  1      5607 2357632 2309.9
## - Gender     1      6906 2358931 2310.1
## - Rating     1     15919 2367944 2311.0
## <none>                   2352025 2311.3
## - Age        1     27738 2379763 2312.3
## - Cards      1    118740 2470765 2321.6
## - Limit      1    277189 2629214 2337.2
## - Student    1   4658514 7010539 2582.4
## - Income     1   6438944 8790969 2638.9
##
## Step:  AIC=2307.54
## Balance ~ Income + Limit + Rating + Cards + Age + Education +
##     Gender + Married + Student
##
##             Df Sum of Sq     RSS    AIC
## - Married    1      1985 2355977 2305.8
## - Education  1      6011 2360004 2306.2
## - Gender     1      7006 2360998 2306.3
## - Rating     1     15773 2369765 2307.2
## <none>                   2353992 2307.5
## - Age        1     28715 2382707 2308.6
## - Cards      1    120483 2474475 2318.0
## - Limit      1    278397 2632389 2333.5
## - Student    1   4662004 7015996 2578.6
## - Income     1   6446168 8800160 2635.2
##
## Step:  AIC=2305.75
## Balance ~ Income + Limit + Rating + Cards + Age + Education +
##     Gender + Student
##
##             Df Sum of Sq     RSS    AIC
## - Education  1      5872 2361849 2304.4
## - Gender     1      6718 2362696 2304.5
## - Rating     1     16907 2372884 2305.5
## <none>                   2355977 2305.8
## - Age        1     31736 2387713 2307.1
## - Cards      1    118814 2474791 2316.1
## - Limit      1    276414 2632391 2331.5
## - Student    1   4668273 7024250 2576.8
## - Income     1   6466218 8822196 2633.8
##
## Step:  AIC=2304.37
## Balance ~ Income + Limit + Rating + Cards + Age + Gender + Student
##
##           Df Sum of Sq     RSS    AIC
## - Gender   1      6150 2367999 2303.0
## - Rating   1     18872 2380721 2304.4
## <none>                 2361849 2304.4
## - Age      1     31448 2393297 2305.7
```

```
## - Cards    1    114114 2475964 2314.2
## - Limit    1    271795 2633644 2329.6
## - Student  1   4681704 7043553 2575.5
## - Income   1   6497310 8859159 2632.9
##
## Step:  AIC=2303.02
## Balance ~ Income + Limit + Rating + Cards + Age + Student
##
##           Df Sum of Sq     RSS    AIC
## - Rating   1     18704 2386704 2303.0
## <none>                 2367999 2303.0
## - Age      1     32933 2400932 2304.5
## - Cards    1    114271 2482270 2312.8
## - Limit    1    272488 2640487 2328.2
## - Student  1   4689851 7057850 2574.1
## - Income   1   6494279 8862278 2631.0
##
## Step:  AIC=2302.99
## Balance ~ Income + Limit + Cards + Age + Student
##
##           Df Sum of Sq      RSS    AIC
## <none>                  2386704 2303.0
## - Age      1     30512  2417215 2304.2
## - Cards    1    239342  2626045 2324.9
## - Student  1   4734882  7121586 2574.3
## - Income   1   6477032  8863735 2629.0
## - Limit    1  35722815 38109519 2993.6
##
## Call:
## lm(formula = Balance ~ Income + Limit + Cards + Age + Student,
##     data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -145.23  -80.22  -11.22   52.40  338.79
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.845e+02  2.842e+01 -17.048  < 2e-16 ***
## Income      -7.915e+00  3.076e-01 -25.733  < 2e-16 ***
## Limit        2.709e-01  4.483e-03  60.432  < 2e-16 ***
## Cards        2.343e+01  4.738e+00   4.947 1.41e-06 ***
## Age         -6.416e-01  3.633e-01  -1.766   0.0786 .
## StudentYes   4.368e+02  1.985e+01  22.001  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.9 on 244 degrees of freedom
## Multiple R-squared:  0.9569, Adjusted R-squared:  0.956
## F-statistic:  1083 on 5 and 244 DF,  p-value: < 2.2e-16
```



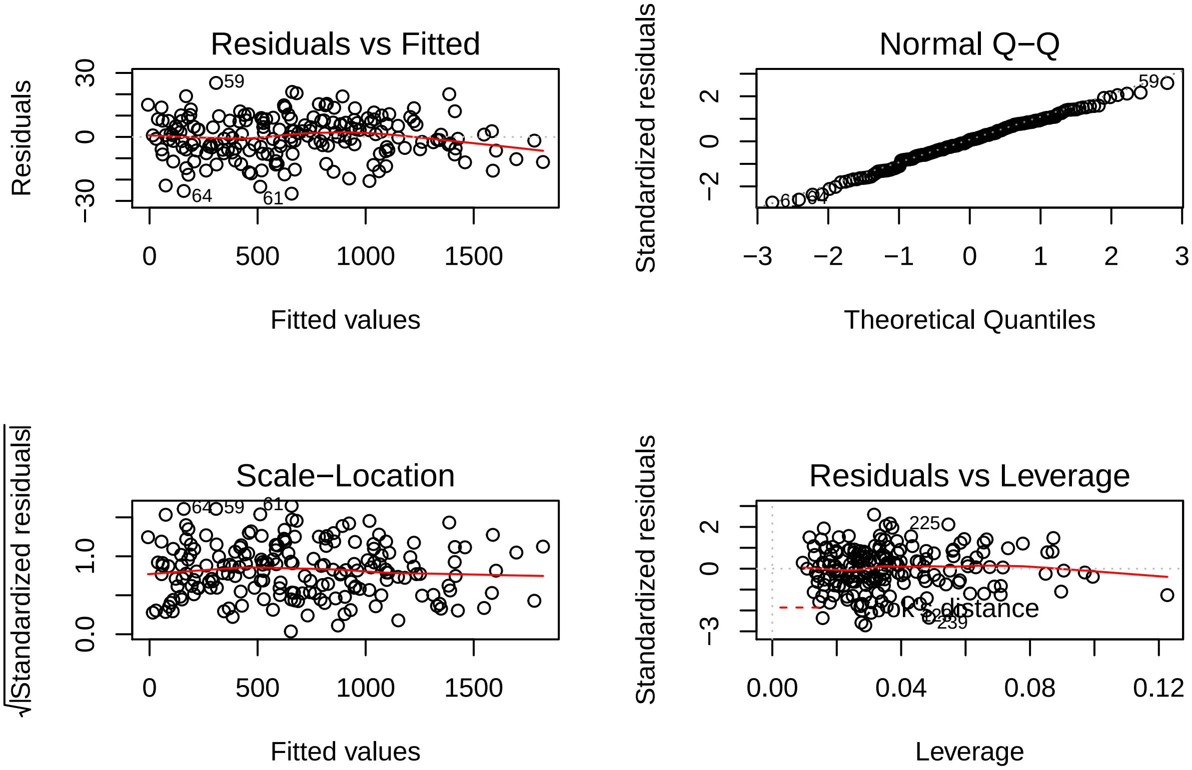

```
## [1] "MSE:"
## [1] 3704.627
```

A slight improve, but residue plot tells that we are still far from success. Now we apply procedure illustrated by cut-off data.

```
# perform cut off version
linreg("cutoff",training1,test,training0 = training0)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
##     Student, data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.5978  -5.9039   0.5969   7.4329  25.3049
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -7.027e+02  4.002e+00 -175.593  < 2e-16 ***
## Income      -1.001e+01  3.645e-02 -274.605  < 2e-16 ***
## Limit        3.402e-01  4.642e-03   73.300  < 2e-16 ***
## Rating      -1.951e-01  6.863e-02   -2.842  0.00499 **
## Cards        2.569e+01  6.208e-01   41.375  < 2e-16 ***
## Age         -9.518e-01  4.198e-02  -22.673  < 2e-16 ***
## StudentYes   5.004e+02  2.104e+00  237.887  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

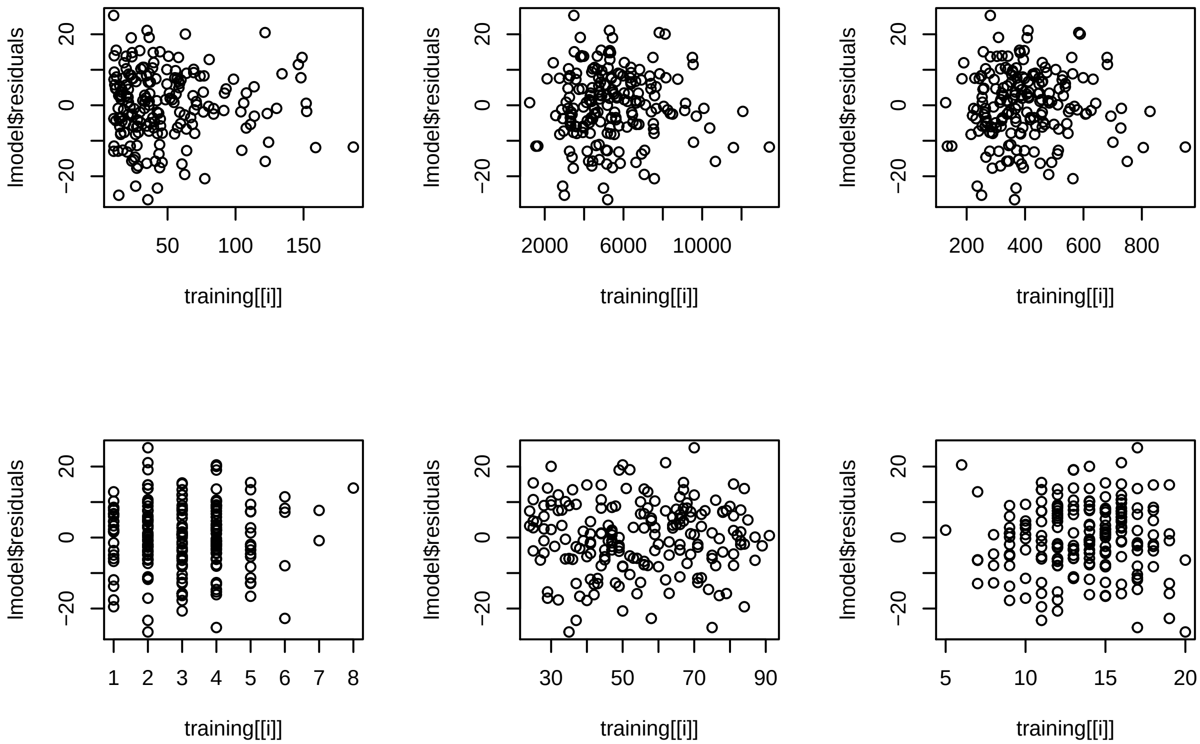
```
## Residual standard error: 9.941 on 183 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9995
## F-statistic: 5.801e+04 on 6 and 183 DF,  p-value: < 2.2e-16
```




```
## [1] "Correct Rate:"
## [1] 1
## [1] "MSE:"
## [1] 106.7317
```

The result shows significant improvement. Correct rate on X_0 is 1, which means all estimates on X_0 is less equal than 0. Diagnosis plots also give great results stating that residues are generally noninformative. R square is 0.9995, also a lot better than before. Thus we regard cut-off an effective method and do further optimization of inference on X_1. We start by plotting residues w.r.t. variables.

```r
# plot residuals
linreg("cutoff",training1,test,training0 = training0,residualplot = 1)
```

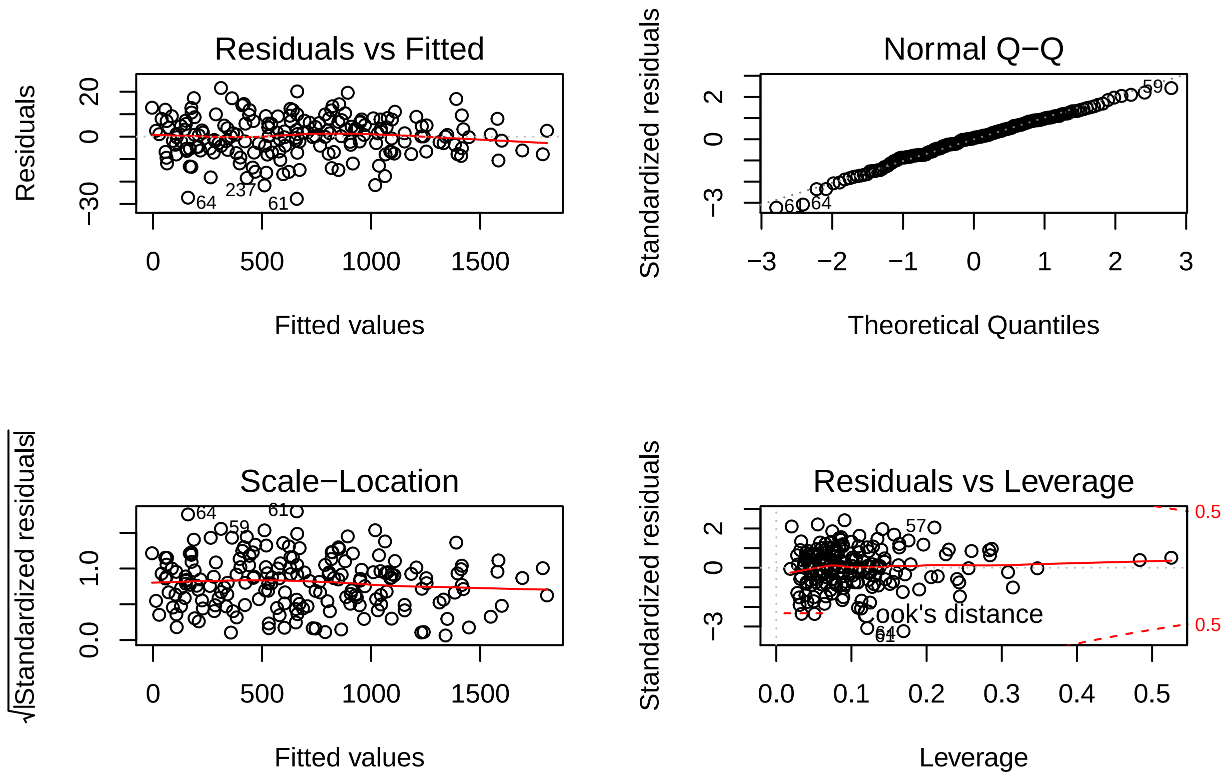


Speechless. This is too good to add anything else. Note that these plot exclude the posibility of quadratic terms. We shall consider issure of interaction terms. Though we have no clue of doing so, but after regression on them,

```r
# interaction terms
linreg("cutoffinter",training1,test,training0 = training0)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
##     Education + Married + Student + Income:Limit + Income:Education +
##     Limit:Age + Limit:Married + Rating:Cards + Rating:Age + Rating:Married +
##     Cards:Education + Age:Education + Age:Married + Married:Student,
##     data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.7065  -6.0809   0.5396   6.3082  21.6623
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -7.182e+02  1.909e+01 -37.614  < 2e-16 ***
## Income             -9.691e+00  1.089e-01 -89.018  < 2e-16 ***
## Limit               3.099e-01  1.614e-02  19.201  < 2e-16 ***
## Rating              2.305e-01  2.377e-01   0.970  0.33354
## Cards               2.764e+01  2.766e+00   9.994  < 2e-16 ***
## Age                -1.199e+00  2.798e-01  -4.286 3.04e-05 ***
## Education           1.138e+00  9.538e-01   1.193  0.23460
## MarriedYes         -3.941e+00  7.120e+00  -0.554  0.58063
## StudentYes          4.962e+02  2.999e+00 165.444  < 2e-16 ***
## Income:Limit       -1.422e-05  7.789e-06  -1.826  0.06966 .
```

```
## Income:Education      -1.668e-02  6.986e-03  -2.388  0.01803 *
## Limit:Age              3.365e-04  2.438e-04   1.380  0.16931
## Limit:MarriedYes       1.903e-02  8.146e-03   2.336  0.02063 *
## Rating:Cards           1.182e-02  3.820e-03   3.094  0.00231 **
## Rating:Age            -5.013e-03  3.609e-03  -1.389  0.16659
## Rating:MarriedYes     -2.693e-01  1.210e-01  -2.226  0.02730 *
## Cards:Education       -4.967e-01  1.673e-01  -2.969  0.00342 **
## Age:Education          2.628e-02  1.470e-02   1.788  0.07556 .
## Age:MarriedYes         1.181e-01  8.265e-02   1.429  0.15483
## MarriedYes:StudentYes  6.102e+00  3.987e+00   1.531  0.12774
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.393 on 170 degrees of freedom
## Multiple R-squared:  0.9996, Adjusted R-squared:  0.9995
## F-statistic: 2.052e+04 on 19 and 170 DF,  p-value: < 2.2e-16
```




```
## [1] "Correct Rate:"
## [1] 1
## [1] "MSE:"
## [1] 128.9456
```

OK, nothing good happened.

The Questions

1. Obviously yes.
2. For income, limit, rating, cards number, age and whetheris student, it is significantly relevant.
3. As in 2.
4. 95% confident intervals are as follows. Note that relative errors are all small enough that the prediction

is effective.

```r
linreg("confint",training1,test,training0 = training0)
```

```
##                   2.5 %        97.5 %
## (Intercept) -710.5791175 -694.78803348
## Income       -10.0802040   -9.93638645
## Limit          0.3310917    0.34940880
## Rating        -0.3304733   -0.05965973
## Cards         24.4607752   26.91045952
## Age           -1.0345934   -0.86894472
## StudentYes   496.2664318  504.56724819
```

5. MSE is 106.7317, which is that (if normal) a 95% confidence interval is of length about 400 dollar.
6. It is piecewise linear.
7. Not there is not.

Conclusion

We conclude that the linear model we specify is as follows:

```r
# a printer of model
linreg("printer",training1,test,training0 = training0)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
##     Student, data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.5978  -5.9039   0.5969   7.4329  25.3049
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -7.027e+02  4.002e+00 -175.593  < 2e-16 ***
## Income      -1.001e+01  3.645e-02 -274.605  < 2e-16 ***
## Limit        3.402e-01  4.642e-03   73.300  < 2e-16 ***
## Rating      -1.951e-01  6.863e-02   -2.842  0.00499 **
## Cards        2.569e+01  6.208e-01   41.375  < 2e-16 ***
## Age         -9.518e-01  4.198e-02  -22.673  < 2e-16 ***
## StudentYes   5.004e+02  2.104e+00  237.887  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.941 on 183 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9995
## F-statistic: 5.801e+04 on 6 and 183 DF,  p-value: < 2.2e-16
```

while estimates less equal than 0 are 0. The optimized MSE is 106.7317.