TP1_MRR

Jingzhuo HUI, You ZUO

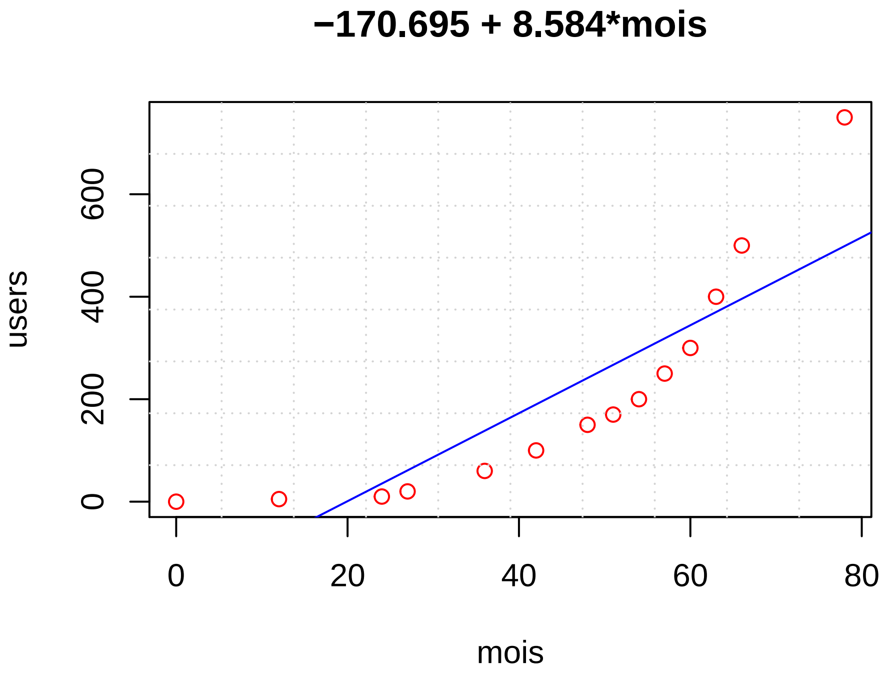
2019/9/27

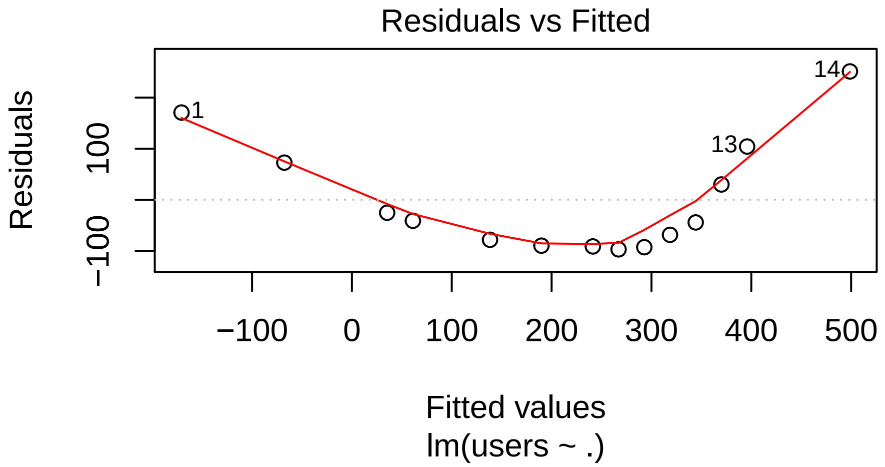
IV Facebook data set

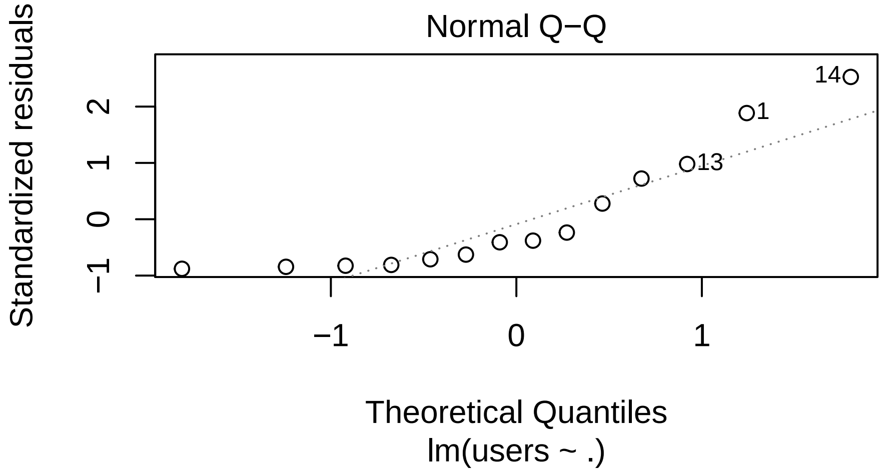
There are two variables in the data **facebook**, the number of months and users, a total of 14 sets of observation. We first make a linear regression model with the month as the independent variable and the number of users as the target, and here is what we get:

```
##
## Call:
## lm(formula = users ~ ., data = tab)
##
## Coefficients:
## (Intercept)         mois
##    -170.695        8.584
```

```
##
## Call:
## lm(formula = users ~ ., data = tab)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -97.07 -86.94 -42.70  62.00 251.17
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -170.695     70.952  -2.406   0.0332 *
## mois           8.584      1.449   5.926 6.97e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 115 on 12 degrees of freedom
## Multiple R-squared:  0.7453, Adjusted R-squared:  0.7241
## F-statistic: 35.11 on 1 and 12 DF,  p-value: 6.97e-05
```







and the predictive value of users when $mois = 80$ is:

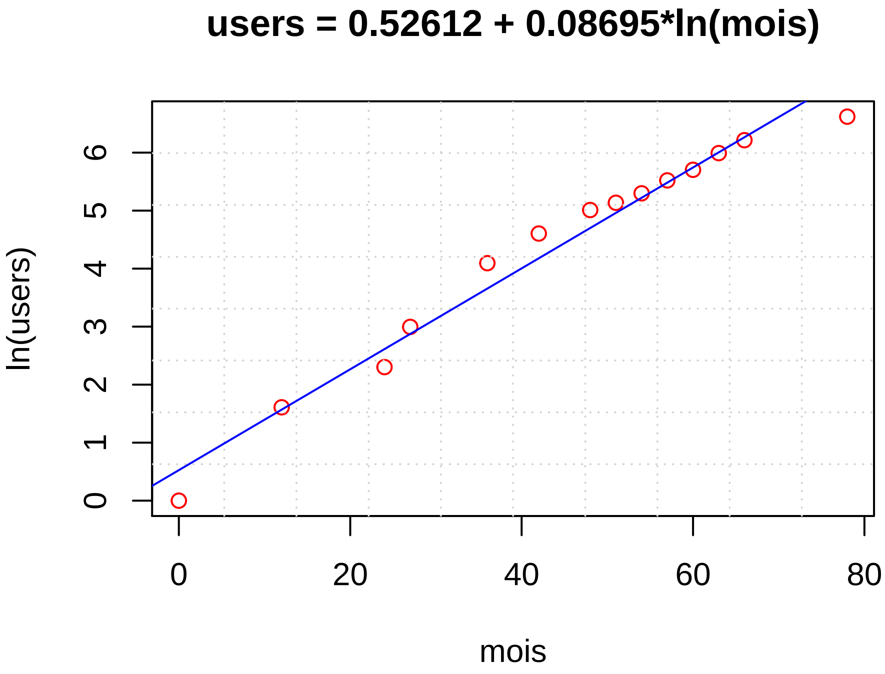
```
##         fit       lwr       upr
## 1 516.0015 232.9323 799.0707
```

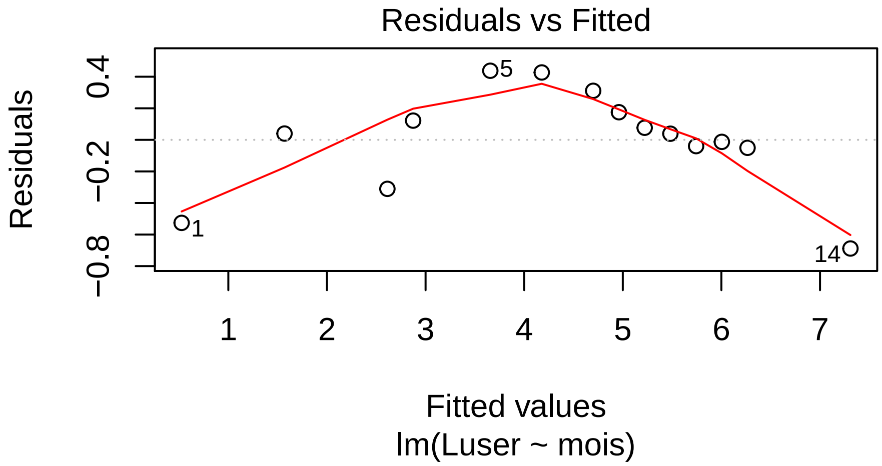
The predictive value is 516, which is even less than the users on month 78.

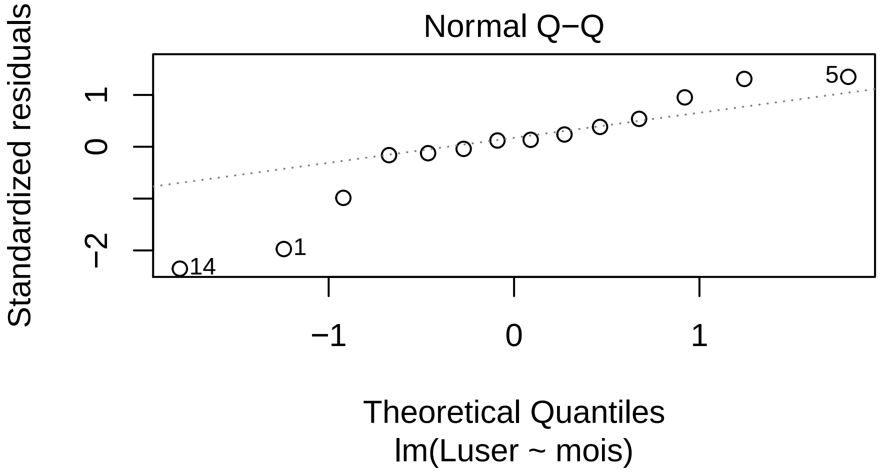
It turns out that our linear model does not fit well with the trend of the data. The residuals are not in a small range and irregularly fluctuate around zero, but at the same time we find that the data trend is similar to the exponential curve. So we decide to transform the forme of the target *users* into $ln(users)$

```
##
## Call:
## lm(formula = Luser ~ mois, data = tab)
##
## Coefficients:
## (Intercept)         mois
##     0.52612      0.08695
```

```
##
## Call:
## lm(formula = Luser ~ mois, data = tab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68848 -0.04777  0.03941  0.16172  0.43787
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.526123   0.208876   2.519    0.027 *
## mois        0.086954   0.004264  20.391 1.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3387 on 12 degrees of freedom
## Multiple R-squared:  0.9719, Adjusted R-squared:  0.9696
## F-statistic: 415.8 on 1 and 12 DF,  p-value: 1.113e-10
```







Compare the results of the new model with the results of the old model, we can see that the $p - value$ of the variable **mois** is less than that of the old model, which means that the improved linear regression model is more accurate, and the relationship between the two variables is more linear.

Using the new model to predict the number of users two months after the last observation on the list, here we get:

```
##        fit      lwr      upr
## 1 1776.603 772.1117 4087.902
```

We use the model to predict the number of users after two months, which is about 1776, also we get $[772, 4087]$ the interval of predicted values. But this time the predictive target is much larger than what we expected. From the results of the above figure, we can see that the closer the data is to the back, the more the trend falls below the straight line, which is why our predicted value is much larger than the true value.

This is actually the reason why our forecast value is much larger than the real value, because the number of facebook users shows an explosive growth at the beginning, which could be similar to exponential growth, but in the later period, as time goes on, the number of users gradually become stable, and as a consequence, new users growth rate will also become slow, which means that the points that follow will fall below the straight line of our model.

Therefore, we conclude that linear models built with long-term user data are not applicable.

V. US crime data

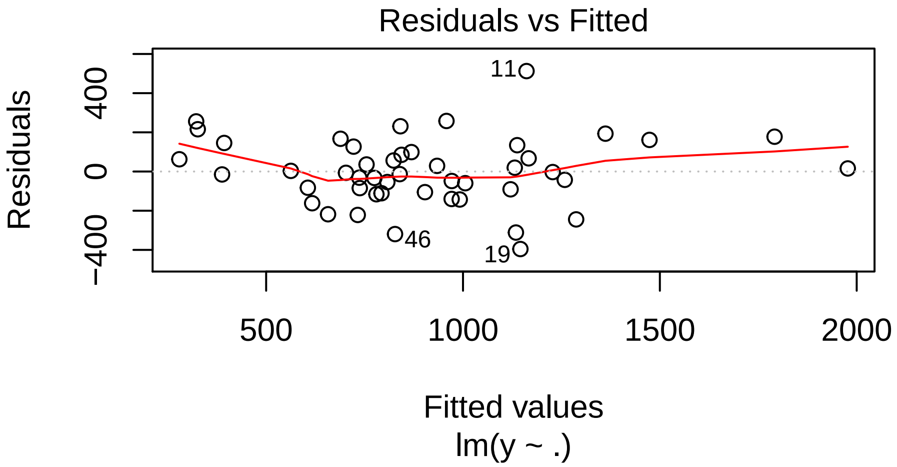
First of all, we have the data structure of data **UScrime** as below:

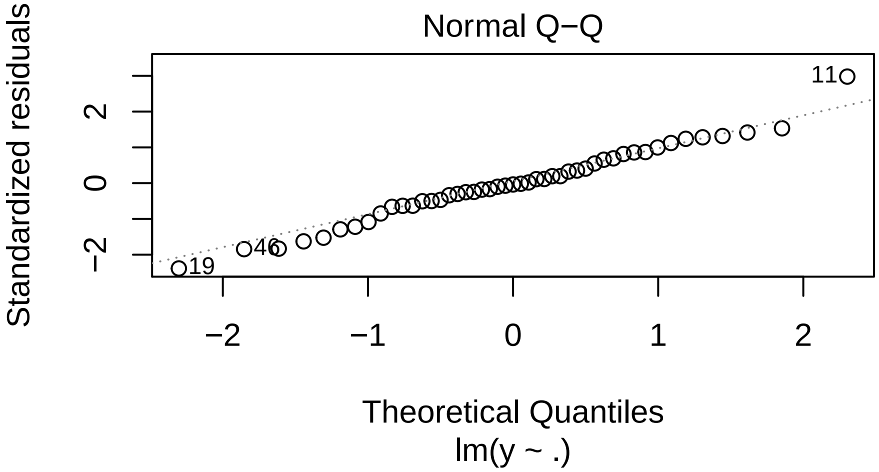
```
##     M So  Ed Po1 Po2  LF  M.F Pop  NW  U1 U2 GDP Ineq     Prob    Time
## 1 151  1  91  58  56 510  950  33 301 108 41 394  261 0.084602 26.2011
## 2 143  0 113 103  95 583 1012  13 102  96 36 557  194 0.029599 25.2999
## 3 142  1  89  45  44 533  969  18 219  94 33 318  250 0.083401 24.3006
## 4 136  0 121 149 141 577  994 157  80 102 39 673  167 0.015801 29.9012
## 5 141  0 121 109 101 591  985  18  30  91 20 578  174 0.041399 21.2998
## 6 121  0 110 118 115 547  964  25  44  84 29 689  126 0.034201 20.9995
##      y
## 1  791
## 2 1635
## 3  578
## 4 1969
## 5 1234
## 6  682
```

We establish our linear model with all the p co-variables, and we also demonstrate the plots which compare the residuals and to the fitted values and a QQ-plot of the residuals:

```
##
## Call:
## lm(formula = y ~ ., data = UScrime)
##
## Coefficients:
## (Intercept)            M           So           Ed          Po1
##  -5984.2876       8.7830      -3.8035      18.8324      19.2804
##         Po2           LF          M.F          Pop           NW
##    -10.9422      -0.6638       1.7407      -0.7330       0.4204
##          U1           U2          GDP         Ineq         Prob
##     -5.8271      16.7800       0.9617       7.0672   -4855.2658
##        Time
##     -3.4790
```

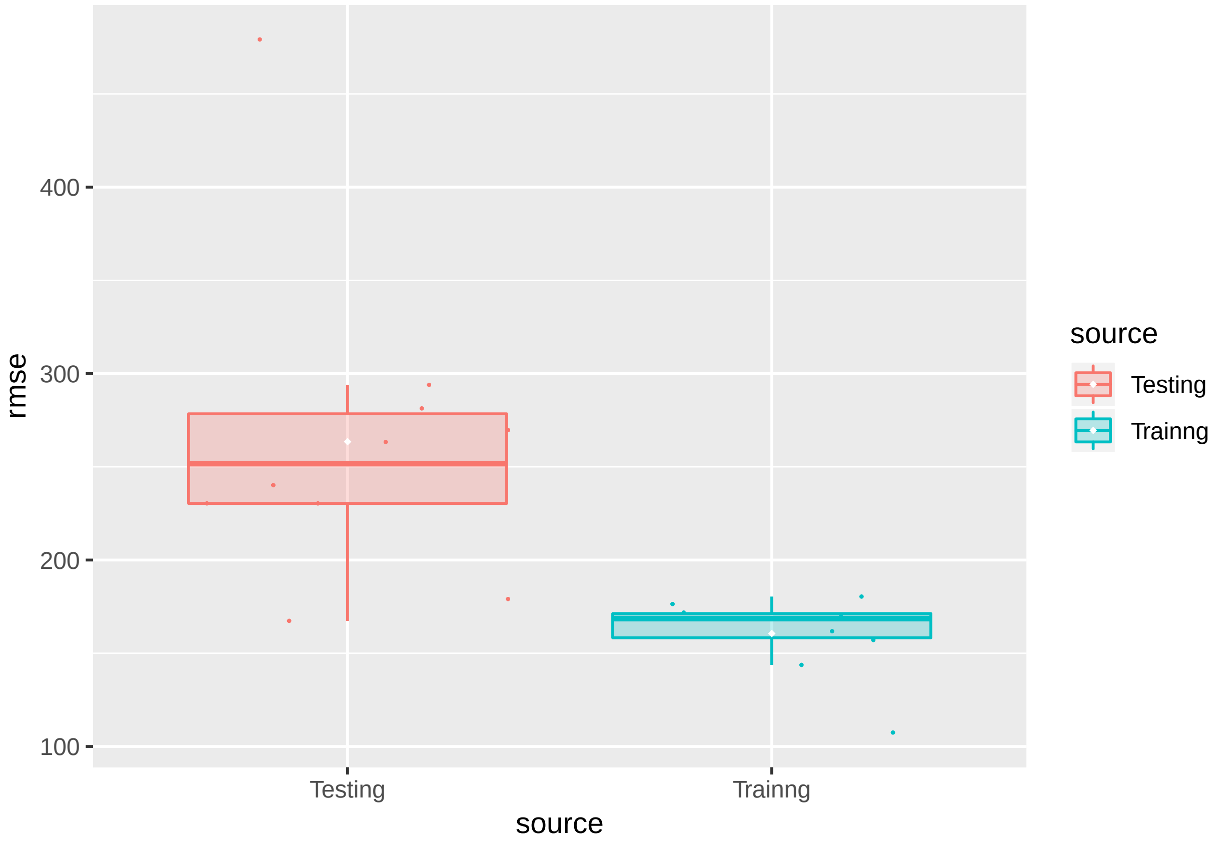
```
##
## Call:
## lm(formula = y ~ ., data = UScrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -395.74  -98.09   -6.69  112.99  512.67
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5984.2876  1628.3184  -3.675 0.000893 ***
## M               8.7830     4.1714   2.106 0.043443 *
## So             -3.8035   148.7551  -0.026 0.979765
## Ed             18.8324     6.2088   3.033 0.004861 **
## Po1            19.2804    10.6110   1.817 0.078892 .
## Po2           -10.9422    11.7478  -0.931 0.358830
## LF             -0.6638     1.4697  -0.452 0.654654
## M.F             1.7407     2.0354   0.855 0.398995
## Pop            -0.7330     1.2896  -0.568 0.573845
## NW              0.4204     0.6481   0.649 0.521279
## U1             -5.8271     4.2103  -1.384 0.176238
## U2             16.7800     8.2336   2.038 0.050161 .
## GDP             0.9617     1.0367   0.928 0.360754
## Ineq            7.0672     2.2717   3.111 0.003983 **
## Prob        -4855.2658  2272.3746  -2.137 0.040627 *
## Time           -3.4790     7.1653  -0.486 0.630708
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 209.1 on 31 degrees of freedom
## Multiple R-squared:  0.8031, Adjusted R-squared:  0.7078
## F-statistic: 8.429 on 15 and 31 DF,  p-value: 3.539e-07
```





According to the results above, we have noticed that there exist some variables which do not have a respectively high level of significativity, but we can see from the Q-Q plot that the residuals follow roughly a normal distribution. So in general our model containing all the p variables is somewhat reasonable in this level.

After that, we want to test the predictive ability of our model. We seperate our data set into two parts, which are the training set $\mathcal{D}_{train}$ and the testing set $\mathcal{D}_{test}$, and they account respectively for 75% and 25% of the total data. Next, we will compute the $RMSE$ of the $\mathcal{D}_{test}$ with our linear regression model established on the $\mathcal{D}_{train}$.



From the results above we can see that the distrubution of $RMSE$ for our different $\mathcal{D}_{test}$ is respectively more varied and the values are very high in some kind, from which the $RMSE$ of the test set is even three to four times the error of the training set, so we can think that the predicted and true values of our model have very large errors. Consequently, from the predictive point of view, the ability of our model still needs to be improved.