46-926 Homework 2, Part I

Jingyi Guo, Pittsburgh Campus

1/28/2017

1 Exhaustive Search

```
allpreds=cbind(weight,current_coupon,time_to_maturiy,is_callable,reporting_delay,trade_size,trade_type,
               curved_based_price,received_time_diff_last1,trade_price_last1,trade_size_last1,
               trade_type_last1,curved_based_price_last1)
Xyframe=data.frame(cbind(allpreds,trade_price))
bestmod=bestglm(Xyframe,IC="AIC")
print(bestmod)
```

```
## AIC
## BICq equivalent for q in (0.855223669269258, 0.952801902678732)
## Best Model:
##                              Estimate   Std. Error   t value
## (Intercept)               1.013788e+00 5.209037e-01  1.946209
## trade_size                1.218353e-07 6.230855e-08  1.955354
## trade_type                3.388076e-01 5.604833e-02  6.044919
## curved_based_price        4.268591e-01 4.175788e-02 10.222241
## received_time_diff_last1 -1.560020e-07 5.418332e-08 -2.879152
## trade_price_last1         6.521732e-01 2.001999e-02 32.576100
## trade_type_last1         -1.350892e-01 5.780145e-02 -2.337125
## curved_based_price_last1 -9.480838e-02 4.379210e-02 -2.164965
##                              Pr(>|t|)
## (Intercept)               5.180315e-02
## trade_size                5.071384e-02
## trade_type                1.853563e-09
## curved_based_price        8.249012e-24
## received_time_diff_last1  4.040085e-03
## trade_price_last1        2.798799e-179
## trade_type_last1          1.955456e-02
## curved_based_price_last1  3.053668e-02
```

We see that the final model takes the form

$$Y_i = \beta_0+\beta_1 trade\_size+\beta_2 trade\_type+\beta_3 curved\_based\_price+\beta_4 received\_time\_diff\_last1+\beta_5 trade\_price\_last1$$

$$+\beta_6 trade\_type\_type\_last1 + \beta_7 curved\_based\_price\_last1$$

Categorical predictors trade_type, trade_type_last1 appear in the final model.

2 PRESS

First, compute the PRESS for the full model

```
fitfullmodel = lm(trade_price ~ ., data = newdata)
levs=hatvalues(fitfullmodel)
```

```
PRESSfull=sum((fitfullmodel$resid/(1-levs))^2)
print(PRESSfull)
```

```
## [1] 5536.378
```

Then, compute the PRESS for the final model in Question 1

```
fitmodel1 = lm(trade_price ~ trade_type+curved_based_price+curved_based_price+received_time_diff_last1
               +trade_price_last1+trade_type_last1+curved_based_price_last1, data = newdata)
levs1=hatvalues(fitmodel1)
PRESS1=sum((fitmodel1$resid/(1-levs1))^2)
print(PRESS1)
```

```
## [1] 5514.856
```

The PRESS value for the full model is larger than that for the AIC-optimal value found in Question 1. So according to PRESS, the model in Question 1 has higher predictive power than the full model.

3 Influential Observations

```
cookd=as.numeric(cooks.distance(fitmodel1))
sort(pf(cookd,8,1612),decreasing=TRUE)[1:5]
```

```
## [1] 2.990504e-02 7.848597e-04 2.091119e-05 1.448552e-05 1.043706e-05
```

The largest Cook's Distance is at the 3.0% of the F distribution. So there is no reason for concern from influcential observations.

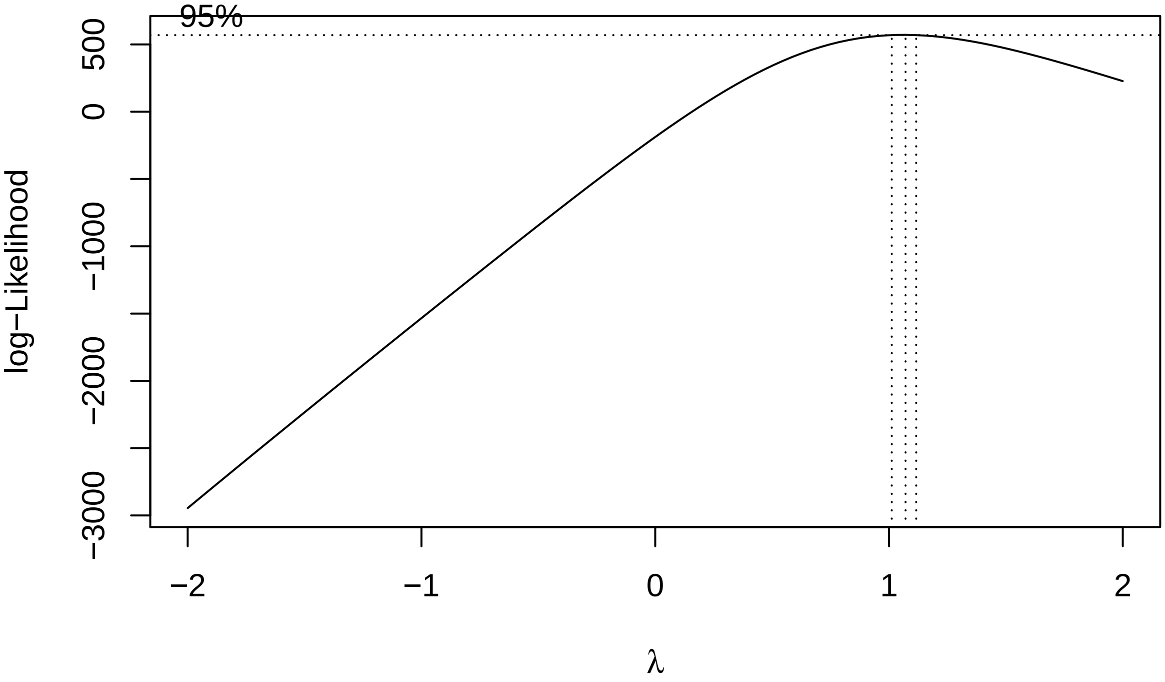
4 Robust Regression Method

```
holdrlm1 = rlm(trade_price ~ trade_type+curved_based_price+curved_based_price+received_time_diff_last1
               +trade_price_last1+trade_type_last1+curved_based_price_last1, data = newdata)
summary(holdrlm1)
```

```
##
## Call: rlm(formula = trade_price ~ trade_type + curved_based_price +
##     curved_based_price + received_time_diff_last1 + trade_price_last1 +
##     trade_type_last1 + curved_based_price_last1, data = newdata)
## Residuals:
##        Min         1Q     Median         3Q        Max
## -34.811374  -0.606539  -0.005745   0.631047   9.503415
##
## Coefficients:
##                          Value   Std. Error t value
## (Intercept)               0.1436  0.3304     0.4346
## trade_type                0.2717  0.0353     7.7068
## curved_based_price        0.4269  0.0265    16.1251
## received_time_diff_last1  0.0000  0.0000    -5.0733
## trade_price_last1         0.7554  0.0127    59.4975
## trade_type_last1         -0.0943  0.0366    -2.5734
## curved_based_price_last1 -0.1885  0.0278    -6.7904
##
## Residual standard error: 0.9114 on 1613 degrees of freedom
```

5 Box-Cox Procedure

```
boxcox(trade_price ~ .,data = newdata)
```



From the plot, 1 is very close to being in the 95% Confidence Interval of lambda. So we could approximate lambda to be 1, which means a transformation of response is not necessary.

However, to be more precise, we see that the optimal choice is

$$\lambda \approx 1.1.$$

And this can be applied to the data as follows

```
transrealizedvol=bcPower(trade_price,1.1)
```