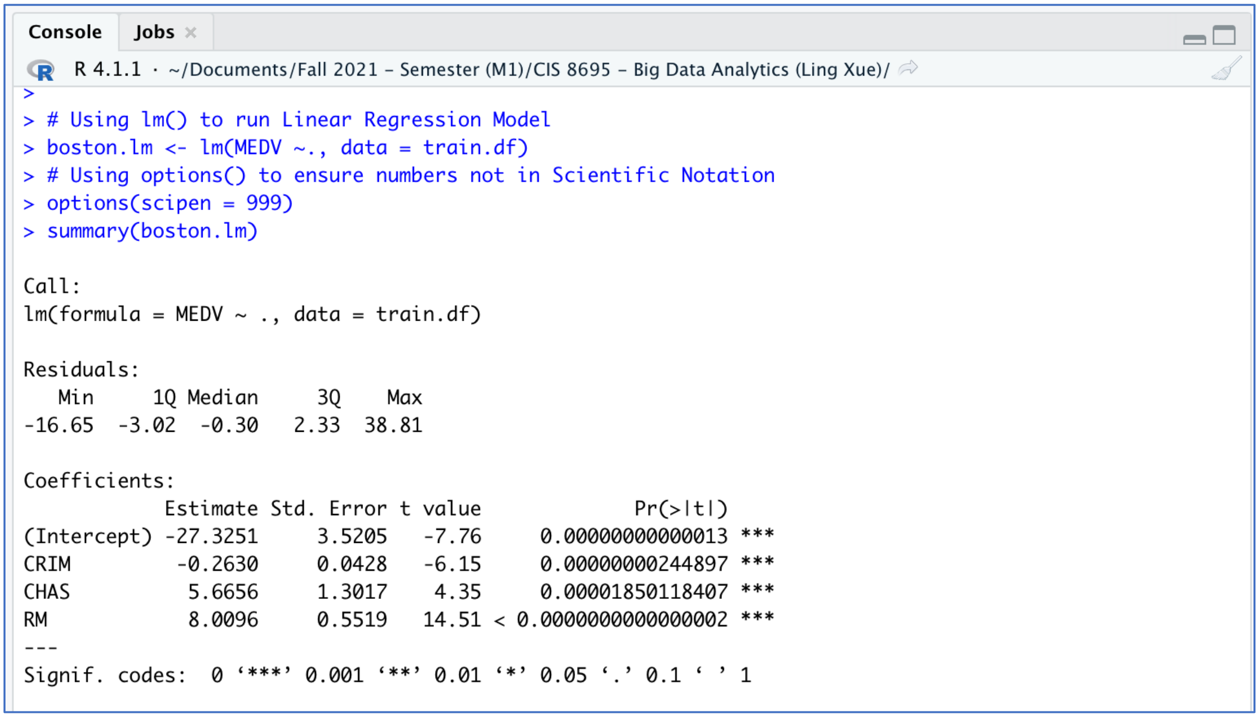
Problem-1: Linear Regression (Predicting Boston Housing Prices)

6.1 b. Fit a multiple linear regression model to the median house price (MEDV) as a function of CRIM, CHAS, and RM. Write the equation for predicting the median house price from the predictors in the model.

Ans: The equation for median house price from the predictors in the model can be described as:
MEDV = -27.3251 + (-0.2630 *CRIM) + (5.6656 *CHAS) + (8.0096 *RM)

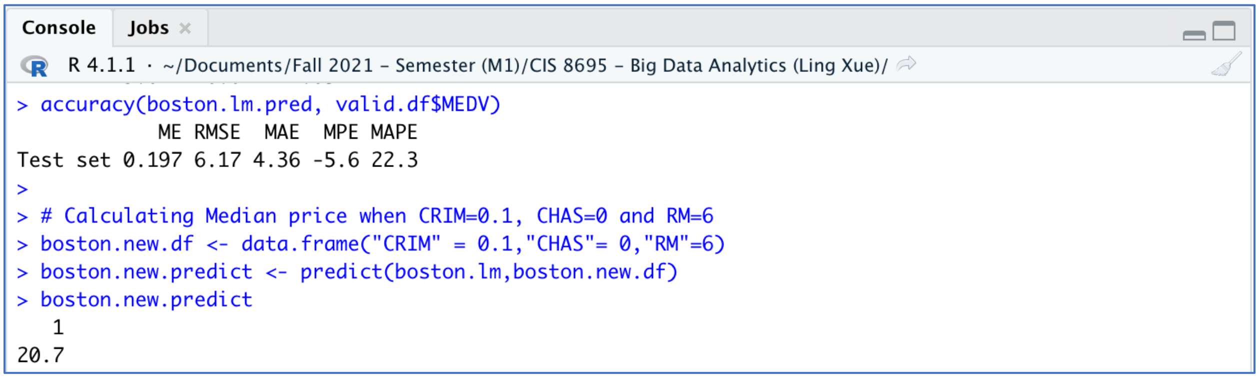

Console
Jobs
R 4.1.1 · ~/Documents/Fall 2021 – Semester (M1)/CIS 8695 – Big Data Analytics (Ling Xue)/

```
>
> # Using lm() to run Linear Regression Model
> boston.lm <- lm(MEDV ~., data = train.df)
> # Using options() to ensure numbers not in Scientific Notation
> options(scipen = 999)
> summary(boston.lm)

Call:
lm(formula = MEDV ~ ., data = train.df)

Residuals:
   Min     1Q Median     3Q    Max
-16.65  -3.02  -0.30   2.33  38.81

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) -27.3251     3.5205   -7.76     0.00000000000013 ***
CRIM         -0.2630     0.0428   -6.15     0.00000000244897 ***
CHAS          5.6656     1.3017    4.35     0.00001850118407 ***
RM            8.0096     0.5519   14.51 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

c. Using the estimated regression model, what median house price is predicted for a tract in the Boston area that does not bound the Charles River, has a crime rate of 0.1, and where the average number of rooms per house is 6? What is the prediction error?

Ans: The median house price for tract in Boston area satisfying the above conditions (CRM=0.1,RM=6.CHAS=0) is: MEDV = -27.3251 + (-0.2630 *0.1) + (5.6656 *0) + (8.0096 *6) = 20.7

The prediction error is 6.17 and can be calculated by looking at the Root Mean Square Error (RMSE)


Console
Jobs
R 4.1.1 · ~/Documents/Fall 2021 – Semester (M1)/CIS 8695 – Big Data Analytics (Ling Xue)/

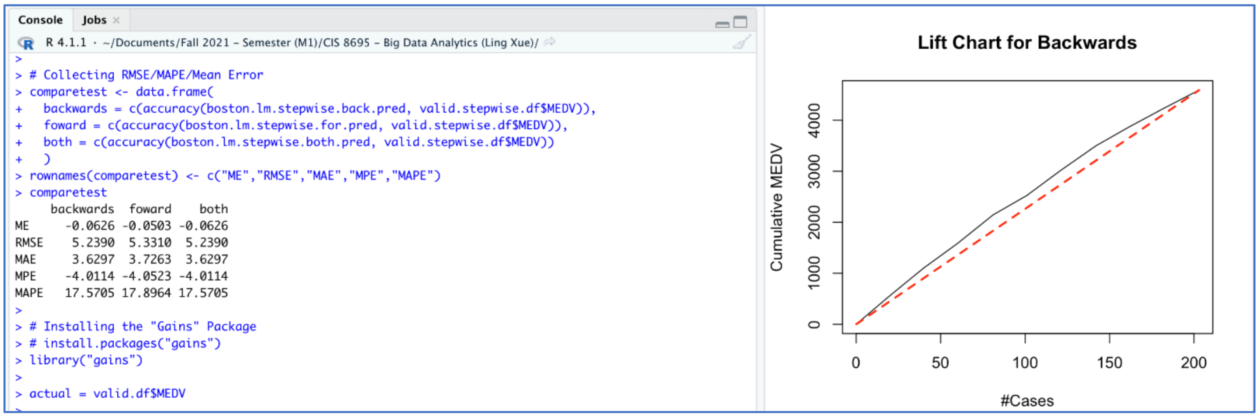
```
> accuracy(boston.lm.pred, valid.df$MEDV)
            ME RMSE  MAE  MPE MAPE
Test set 0.197 6.17 4.36 -5.6 22.3
>
> # Calculating Median price when CRIM=0.1, CHAS=0 and RM=6
> boston.new.df <- data.frame("CRIM" = 0.1,"CHAS"= 0,"RM"=6)
> boston.new.predict <- predict(boston.lm,boston.new.df)
> boston.new.predict
   1
20.7
```

d. iii) Use stepwise regression with the three options (backward, forward, both) to reduce the remaining predictors as follows: Run stepwise on the training set. Choose the top model from each stepwise run. Then use each of these models separately to predict the validation set. *Compare RMSE, MAPE, and mean error, as well as lift charts. Finally, describe the best model*.

Ans: Comparing RMSE/MAPE/Mean Error for the three options, we observe nearly identical values:

Step (BOTH): ME = -0.0626, RMSE = 5.24, MAPE = 17.6
Step (Forward): ME = -0.0503, RMSE = 5.33, MAPE = 17.9
Step (Backward): ME = -0.0626, RMSE = 5.24, MAPE = 17.6

The best performing model will ideally have lowest residual mean square, that would internally maximize the multiple correlation value R^2. Another statistic, in determining the best model in multiple linear regression is Cp criterion.


Console
Jobs
R 4.1.1 · ~/Documents/Fall 2021 – Semester (M1)/CIS 8695 – Big Data Analytics (Ling Xue)/
Lift Chart for Backwards
Cumulative MEDV
0
1000
2000
3000
4000
0
50
100
150
200
#Cases

```
>
> # Collecting RMSE/MAPE/Mean Error
> comparetest <- data.frame(
+   backwards = c(accuracy(boston.lm.stepwise.back.pred, valid.stepwise.df$MEDV)),
+   foward = c(accuracy(boston.lm.stepwise.for.pred, valid.stepwise.df$MEDV)),
+   both = c(accuracy(boston.lm.stepwise.both.pred, valid.stepwise.df$MEDV))
+   )
> rownames(comparetest) <- c("ME","RMSE","MAE","MPE","MAPE")
> comparetest
     backwards  foward    both
ME     -0.0626 -0.0503 -0.0626
RMSE    5.2390  5.3310  5.2390
MAE     3.6297  3.7263  3.6297
MPE    -4.0114 -4.0523 -4.0114
MAPE   17.5705 17.8964 17.5705
>
> # Installing the "Gains" Package
> # install.packages("gains")
> library("gains")
>
> actual = valid.df$MEDV
```

Sample Code (Problem-1: Linear Regression):

P1-Linear Regression.txt

Problem-2: Logistic Regression (Financial Condition of Banks)

10.1 b. Consider a new bank whose total loans and leases/assets ratio = 0.6 and total expenses/assets ratio = 0.11. From your logistic regression model, estimate the following four quantities for this bank (use R to do all the intermediate calculations; show your final answers to four decimal places): the logit, the odds, the probability of being financially weak, and the classification of the bank (use cutoff = 0.5)

Ans: Leases/Assets Ratio = 0.6, Total Expenses/Assets Ratio = 0.11, Cutoff = 0.5

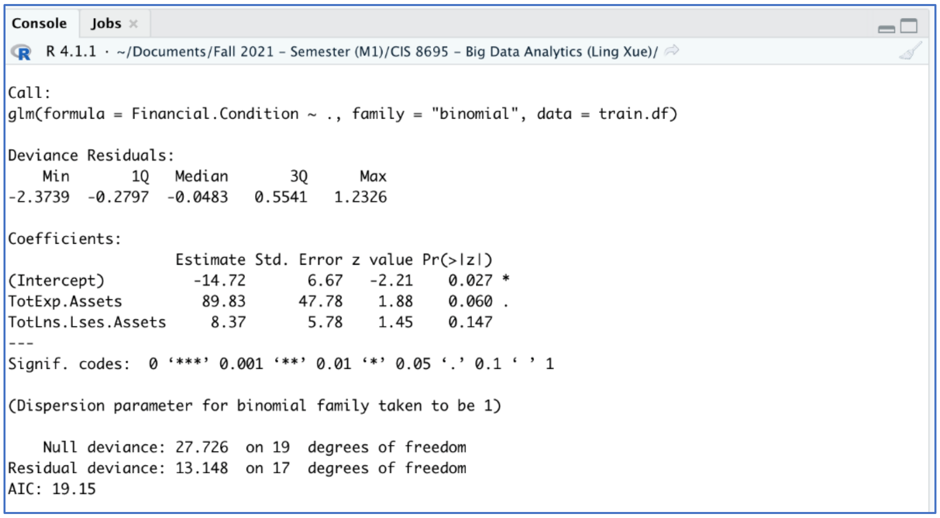
Logit Equation = -14.72 + 89.83* TotExp.Assets + 8.37 * TotLns.Lses.Assets

= (-14.72) + (89.83*0.11) + (8.37*0.6) = 0.1833

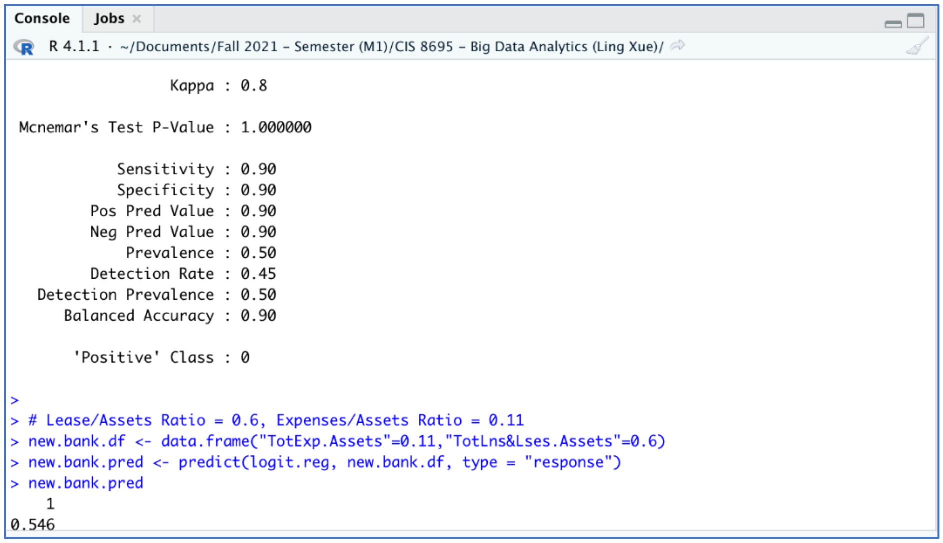
Odds: e^(0.1833) = 1.2011

Probability: Odds/(1+Odds) = = 1.2011 / (1+1.2011) = 0.5456

Classification: Financially Weak (P>0.5)


Console
Jobs
R 4.1.1 · ~/Documents/Fall 2021 – Semester (M1)/CIS 8695 – Big Data Analytics (Ling Xue)/

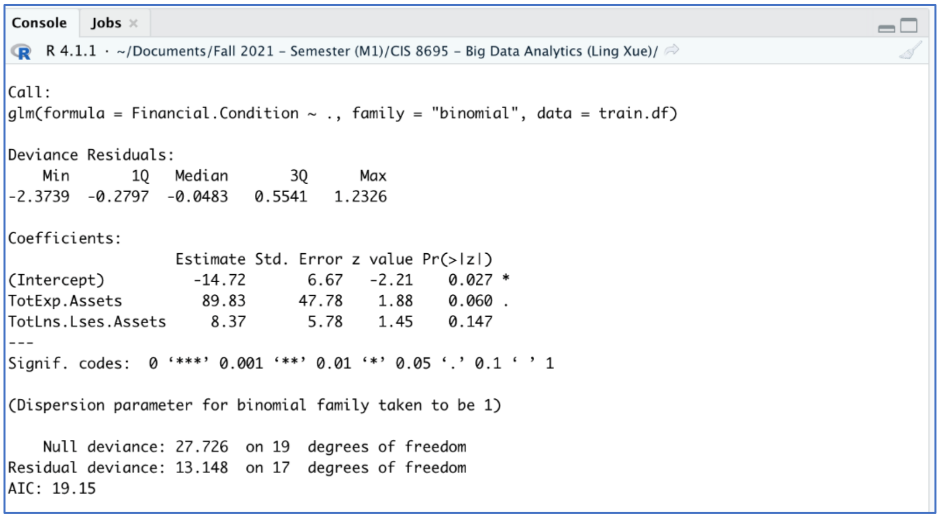
```
Call:
glm(formula = Financial.Condition ~ ., family = "binomial", data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-2.3739  -0.2797  -0.0483   0.5541   1.2326  

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)  
(Intercept)          -14.72       6.67   -2.21    0.027 *
TotExp.Assets         89.83      47.78    1.88    0.060 .
TotLns.Lses.Assets     8.37       5.78    1.45    0.147  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.726  on 19  degrees of freedom
Residual deviance: 13.148  on 17  degrees of freedom
AIC: 19.15
```


Console
Jobs
R 4.1.1 · ~/Documents/Fall 2021 – Semester (M1)/CIS 8695 – Big Data Analytics (Ling Xue)/

```
                  Kappa : 0.8             
                                          
 Mcnemar's Test P-Value : 1.000000        
                                          
            Sensitivity : 0.90            
            Specificity : 0.90            
         Pos Pred Value : 0.90            
         Neg Pred Value : 0.90            
             Prevalence : 0.50            
         Detection Rate : 0.45            
   Detection Prevalence : 0.50            
      Balanced Accuracy : 0.90            
                                          
       'Positive' Class : 0               
                                          

> 
> # Lease/Assets Ratio = 0.6, Expenses/Assets Ratio = 0.11
> new.bank.df <- data.frame("TotExp.Assets"=0.11,"TotLns&Lses.Assets"=0.6)
> new.bank.pred <- predict(logit.reg, new.bank.df, type = "response")
> new.bank.pred
    1 
0.546 
```

d. Interpret the estimated coefficient for the total loans & leases to total assets ratio (TotLns&Lses/Assets) in terms of the odds of being financially weak

Ans: The positive coefficients of TotLns&Lses/Assets translates into odds coefficient being larger than 1.



R 4.1.1 · ~/Documents/Fall 2021 – Semester (M1)/CIS 8695 – Big Data Analytics (Ling Xue)/

```
Call:
glm(formula = Financial.Condition ~ ., family = "binomial", data = train.df)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.3739  -0.2797  -0.0483   0.5541   1.2326  

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)  
(Intercept)          -14.72       6.67   -2.21    0.027 *
TotExp.Assets         89.83      47.78    1.88    0.060 .
TotLns.Lses.Assets     8.37       5.78    1.45    0.147  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.726  on 19  degrees of freedom
Residual deviance: 13.148  on 17  degrees of freedom
AIC: 19.15
```

e. When a bank that is in poor financial condition is misclassified as financially strong, the misclassification cost is much higher than when a financially strong bank is misclassified as weak. To minimize the expected cost of misclassification, should the cut off value for classification (which is currently at 0.5) be increased or decreased?

Ans: To minimize the expected cost of misclassification, the cut off values must be “decreased”, since the success class is identified as weak.

Sample Code (Problem-2: Logistic Regression):

P2-Logistic Regression.txt