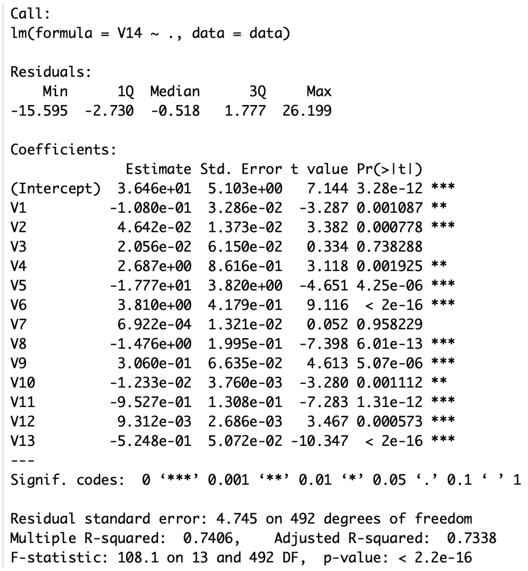
Homework 6: Outlier Detection

1. (**0 points**) **Page 1**: code for regression and resulting model.

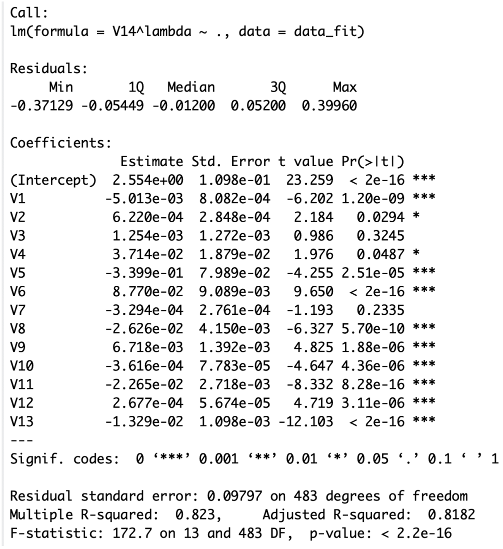
```
data = read.table("/Users/xiaoxin/Desktop/19spring/aml/6/housing.data.txt")
model = lm(V14~., data = data)
summary(model)
```

Original model:



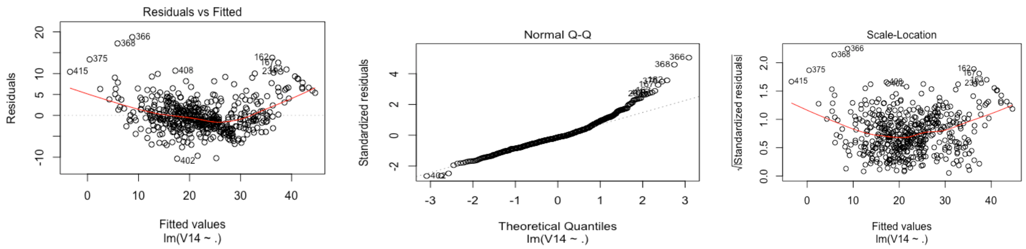
```
Call:
lm(formula = V14 ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.595  -2.730  -0.518   1.777  26.199 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
V1          -1.080e-01  3.286e-02  -3.287 0.001087 ** 
V2           4.642e-02  1.373e-02   3.382 0.000778 ***
V3           2.056e-02  6.150e-02   0.334 0.738288    
V4           2.687e+00  8.616e-01   3.118 0.001925 ** 
V5          -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
V6           3.810e+00  4.179e-01   9.116  < 2e-16 ***
V7           6.922e-04  1.321e-02   0.052 0.958229    
V8          -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
V9           3.060e-01  6.635e-02   4.613 5.07e-06 ***
V10         -1.233e-02  3.760e-03  -3.280 0.001112 ** 
V11         -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
V12          9.312e-03  2.686e-03   3.467 0.000573 ***
V13         -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338 
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

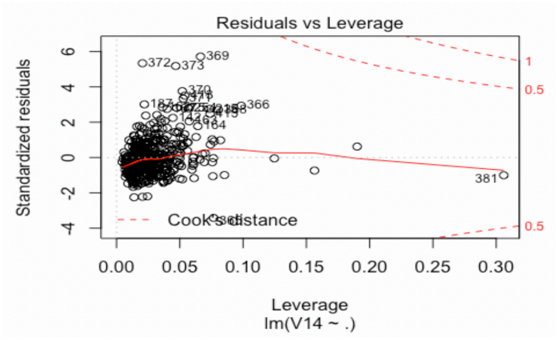
Final model after revising:



```
Call:
lm(formula = V14^lambda ~ ., data = data_fit)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.37129 -0.05449 -0.01200  0.05200  0.39960 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.554e+00  1.098e-01  23.259  < 2e-16 ***
V1          -5.013e-03  8.082e-04  -6.202 1.20e-09 ***
V2           6.220e-04  2.848e-04   2.184   0.0294 *  
V3           1.254e-03  1.272e-03   0.986   0.3245    
V4           3.714e-02  1.879e-02   1.976   0.0487 *  
V5          -3.399e-01  7.989e-02  -4.255 2.51e-05 ***
V6           8.770e-02  9.089e-03   9.650  < 2e-16 ***
V7          -3.294e-04  2.761e-04  -1.193   0.2335    
V8          -2.626e-02  4.150e-03  -6.327 5.70e-10 ***
V9           6.718e-03  1.392e-03   4.825 1.88e-06 ***
V10         -3.616e-04  7.783e-05  -4.647 4.36e-06 ***
V11         -2.265e-02  2.718e-03  -8.332 8.28e-16 ***
V12          2.677e-04  5.674e-05   4.719 3.11e-06 ***
V13         -1.329e-02  1.098e-03 -12.103  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09797 on 483 degrees of freedom
Multiple R-squared:  0.823,     Adjusted R-squared:  0.8182 
F-statistic: 172.7 on 13 and 483 DF,  p-value: < 2.2e-16
```

2. (**50 points**) **Page 2**: a screenshot of your diagnostic plot and a few sentences of your explanation.





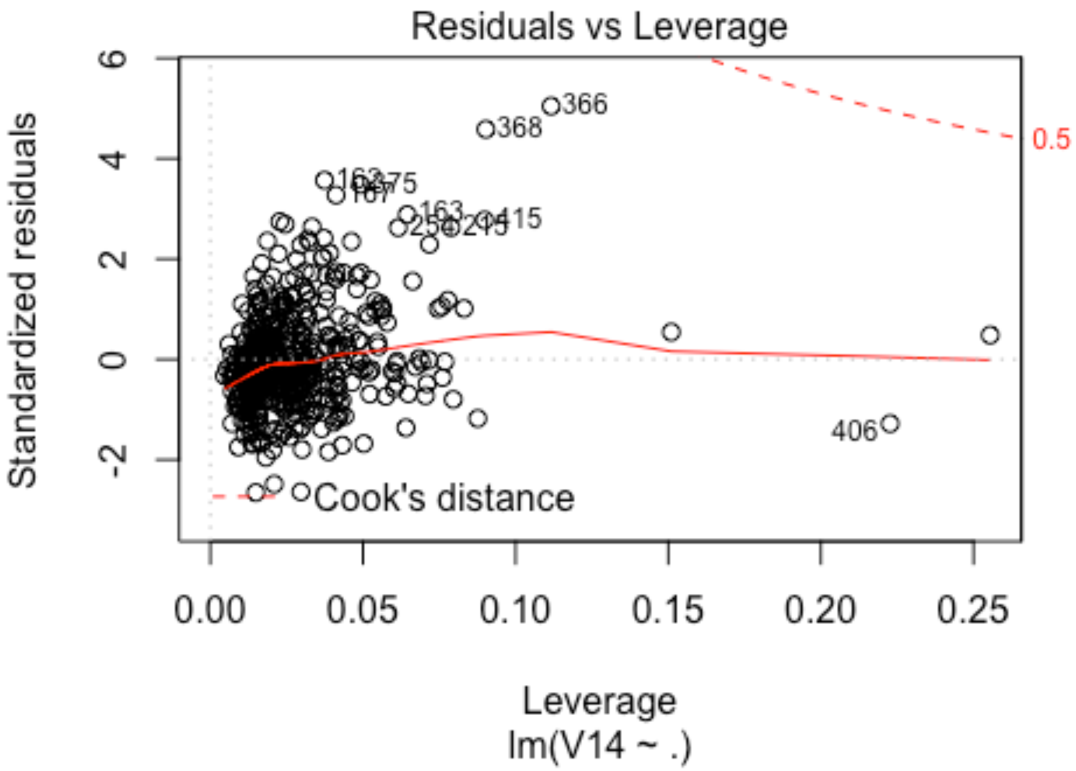
Cook’s distance, standardized residuals and leverage should be combined to detect the outliers. I choose to show 20 suspect points in the plot so that I could make a selection. To get a clearer picture, I set some thresholds and get their intersects as an reference.

From the perspective of Cook’s distance(i.e. influence), all the points are within the dash line of 0.5, which means they are kind of acceptable. Furthermore, I use 4/n as a threshold, but get too many points, so I would just remove those points that also have a large standardized residual.(more than 3)

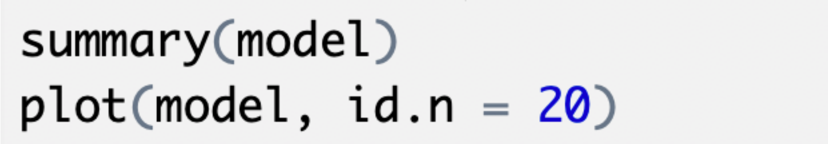
From the perspective of leverage, obviously, there are some points with high leverage. Point 381 with leverage more than 0.3 means the model’s value at that point is predicted largely by the observed value at that point instead of the other training data, which is not good for predicting, so I choose to remove it. For other points with leverage less than 0.2 but larger than 0.1, though their leverage are relatively large, their standard deviation is less than 2, which means they are acceptable.

From the perspective of standard deviation, about 99% of the sampled values of a standard normal random variable are in the range [-3,3], so I choose abs=3 as a threshold and get 8 suspects points, which are also shown on the plot. Their leverage are all above threshold, so I choose to remove them.

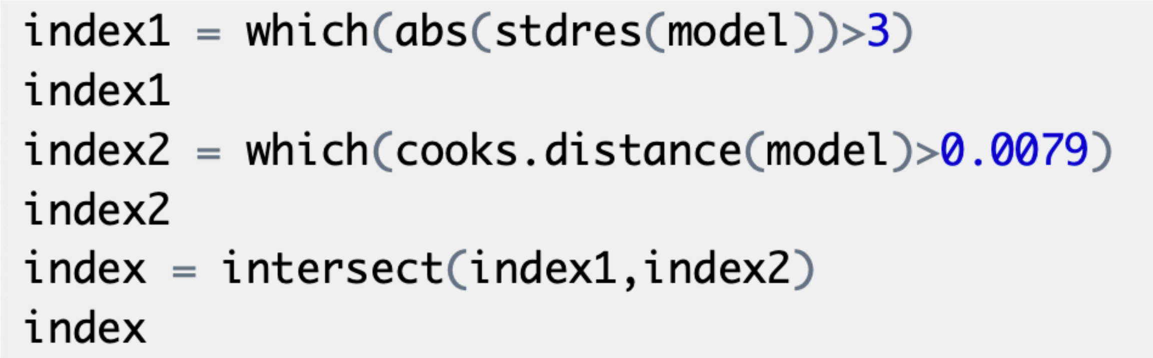
3. (**20 points**) **Page 3**: a screenshot of your new diagnostic plot.



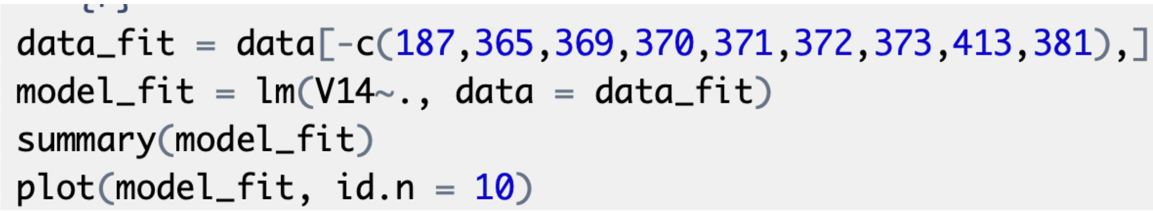
4. (**10 points**) **Page 4**: a screenshot of your code for subproblem 2.



```
summary(model)
plot(model, id.n = 20)
```

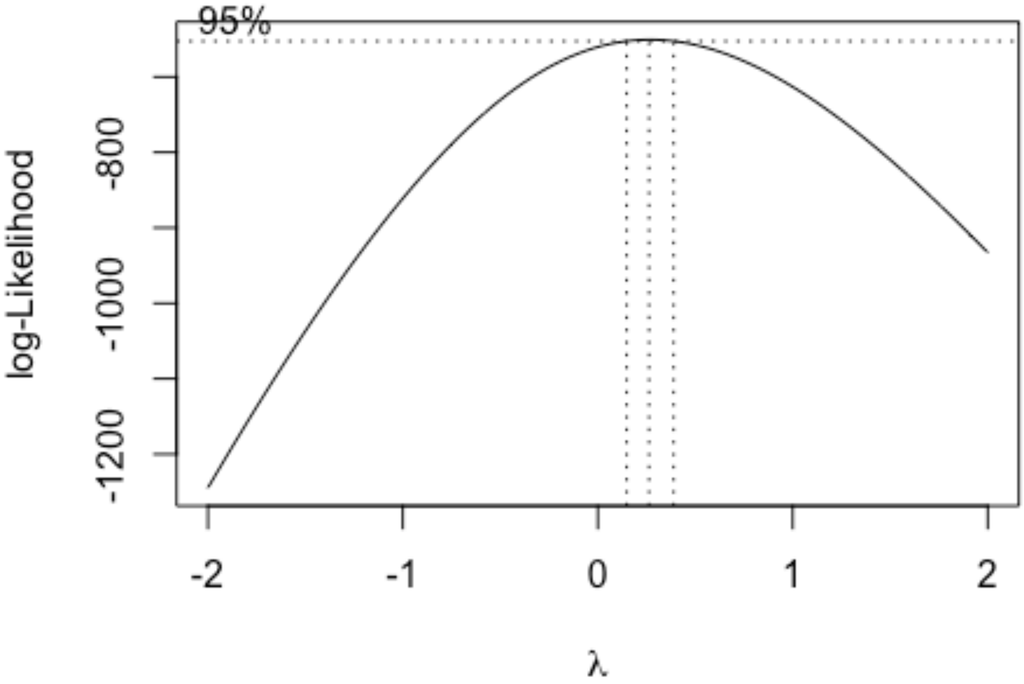


```
index1 = which(abs(stdres(model))>3)
index1
index2 = which(cooks.distance(model)>0.0079)
index2
index = intersect(index1,index2)
index
```



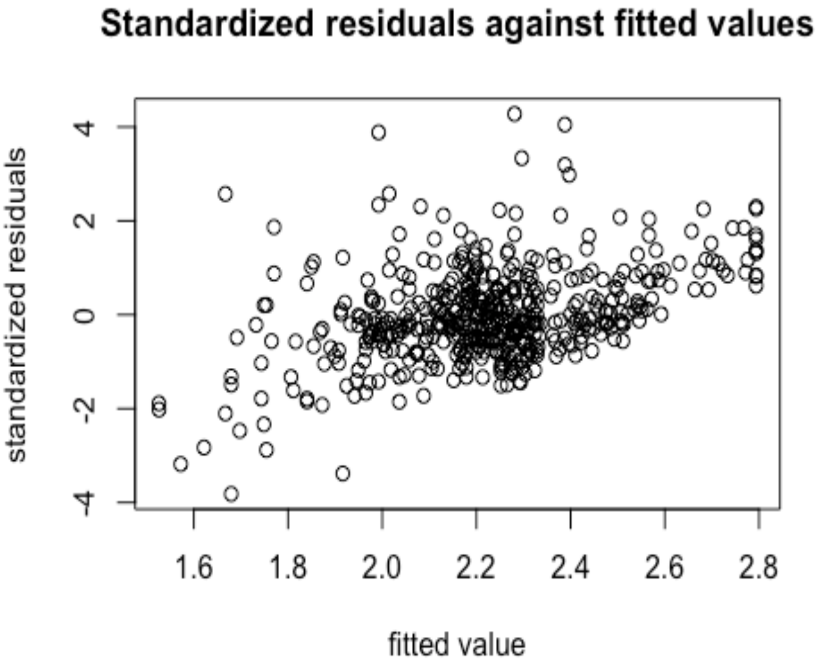
```
data_fit = data[-c(187,365,369,370,371,372,373,413,381),]
model_fit = lm(V14~., data = data_fit)
summary(model_fit)
plot(model_fit, id.n = 10)
```

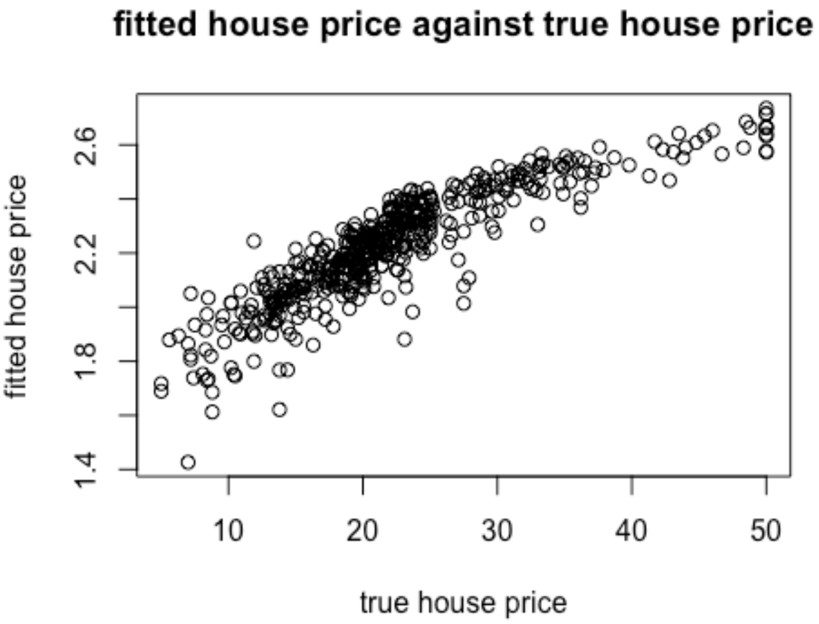
5. (**10 points**) **Page 5**: a screenshot of Box-Cox transformation plot and the best value you chose.



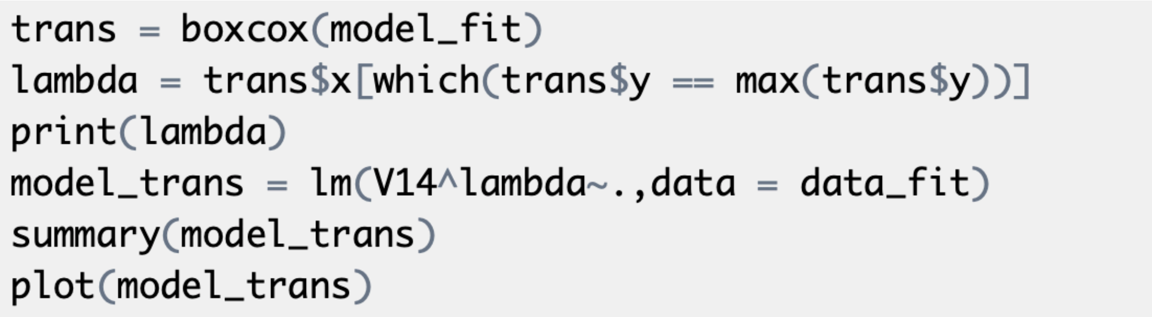
Lamda is 0.2626263.

6. (**10 points**) **Page 6**: result of the standardized residuals of the regression after Box-Cox transformation and a plot of fitted house price against true house price.

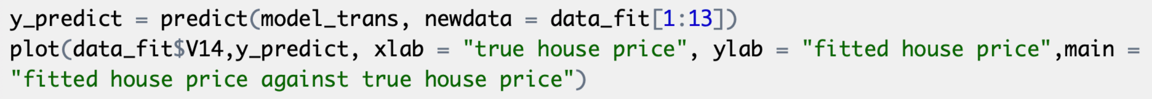




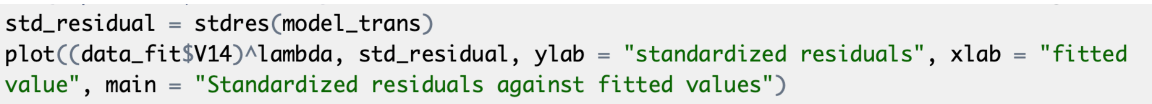
7. (**0 points**) **Page 7**: code for subproblems 3 and 4.



```
trans = boxcox(model_fit)
lambda = trans$x[which(trans$y == max(trans$y))]
print(lambda)
model_trans = lm(V14^lambda~.,data = data_fit)
summary(model_trans)
plot(model_trans)
```



```
y_predict = predict(model_trans, newdata = data_fit[1:13])
plot(data_fit$V14,y_predict, xlab = "true house price", ylab = "fitted house price",main =
"fitted house price against true house price")
```



```
std_residual = stdres(model_trans)
plot((data_fit$V14)^lambda, std_residual, ylab = "standardized residuals", xlab = "fitted
value", main = "Standardized residuals against fitted values")
```