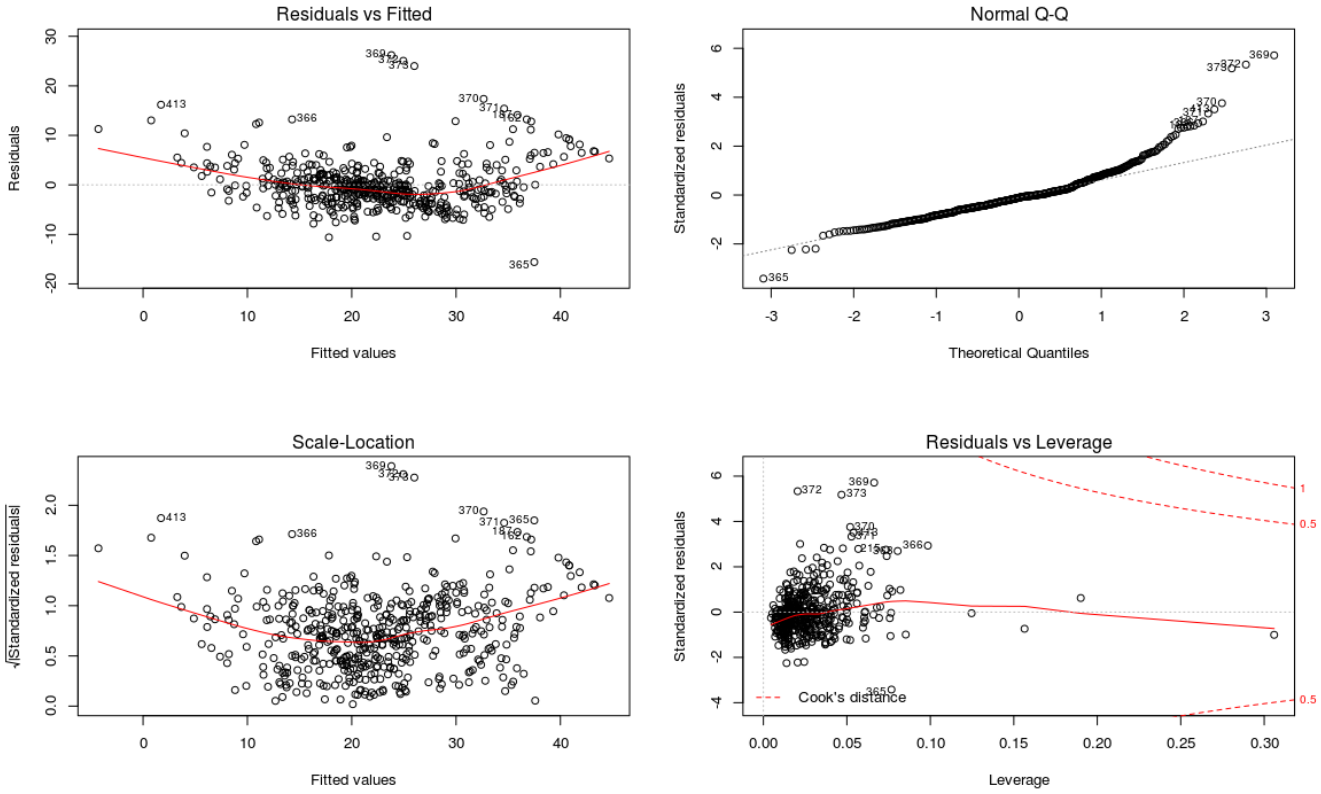
Page 1 Code fore regression and resulting model

```
library(glmnet)
library(MASS)
library(ISLR)

summary(Boston)
fit1=lm(medv~.,Boston)
summary(fit1)
par(mfrow=c(2,2))
plot(fit1, id.n = 10) # which gives 365, 369, 372, 373

# Calculate hat values
hv = hatvalues(fit1)
hv_max_index = which(hv==max(hv)) # which gives 381

# Calcualte standardized residuals
sr = rstandard(fit1)
sr_max_index = which(sr==max(sr)) # which gives 369

# Calculate cooks distance
ck = cooks.distance(fit1)
ck_max_index = which(ck==max(ck)) # which gives 369
```

Page 2 A screenshot of your diagnostic plot and a few sentences of your explanation

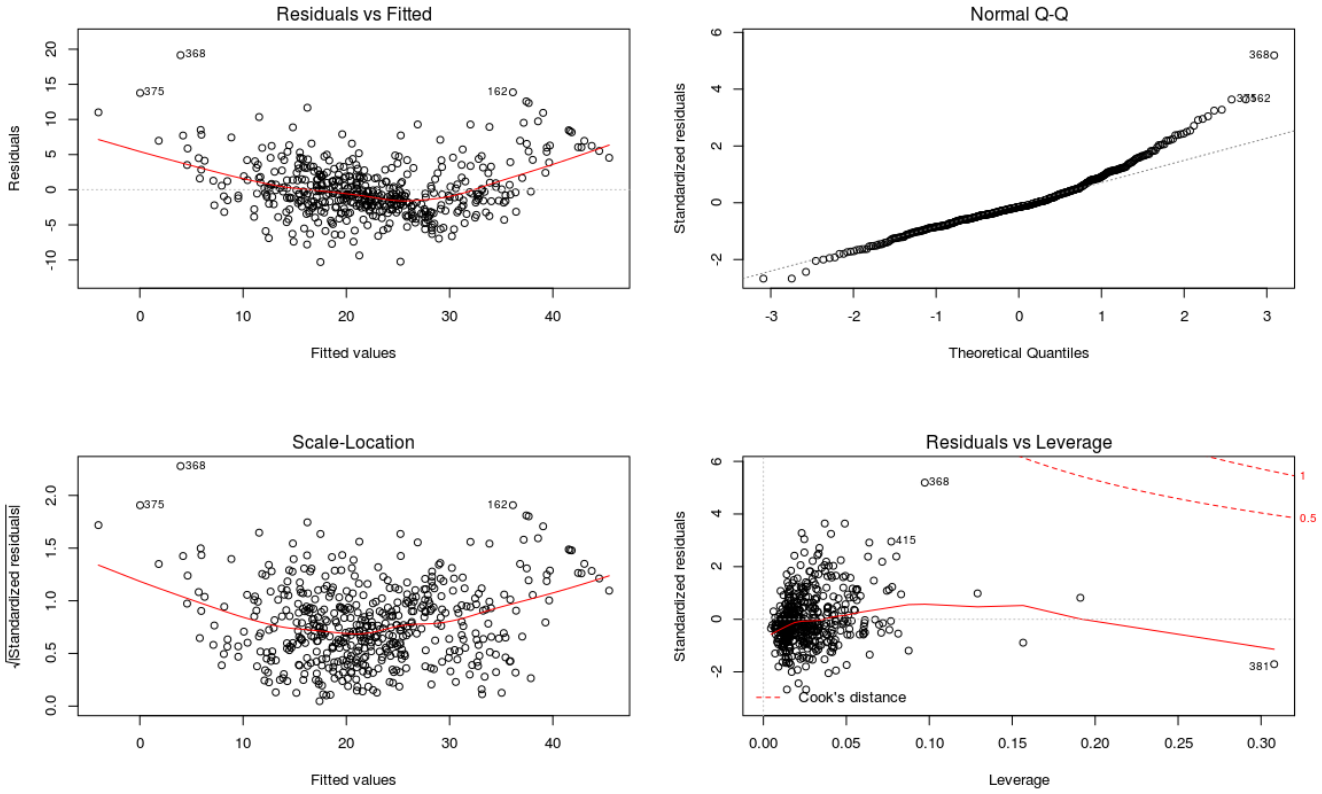
The regression model has been created for the house prices against all other features in the given data. The diagnostic plot is obtained:



As seen from the bottom right Residuals vs Leverage plot, there are several points 372, 373, 369, having large residual values which are almost 6 deviations away from the mean; Similarly, the point 370, 371, 366, 365, 413 has large residual values of 4 deviations away from the mean. These points are identified as outliers. There are several other points on the right of the same plot have high leverage, without having a large residual value. These values may or may not present problems, thus are not identified as outliers.

Page 3 A screenshot of your new diagnostic plot

After removing the 8 outliers identified in Problem 1, the new regression is computed and the new diagnostic plot is drawn:

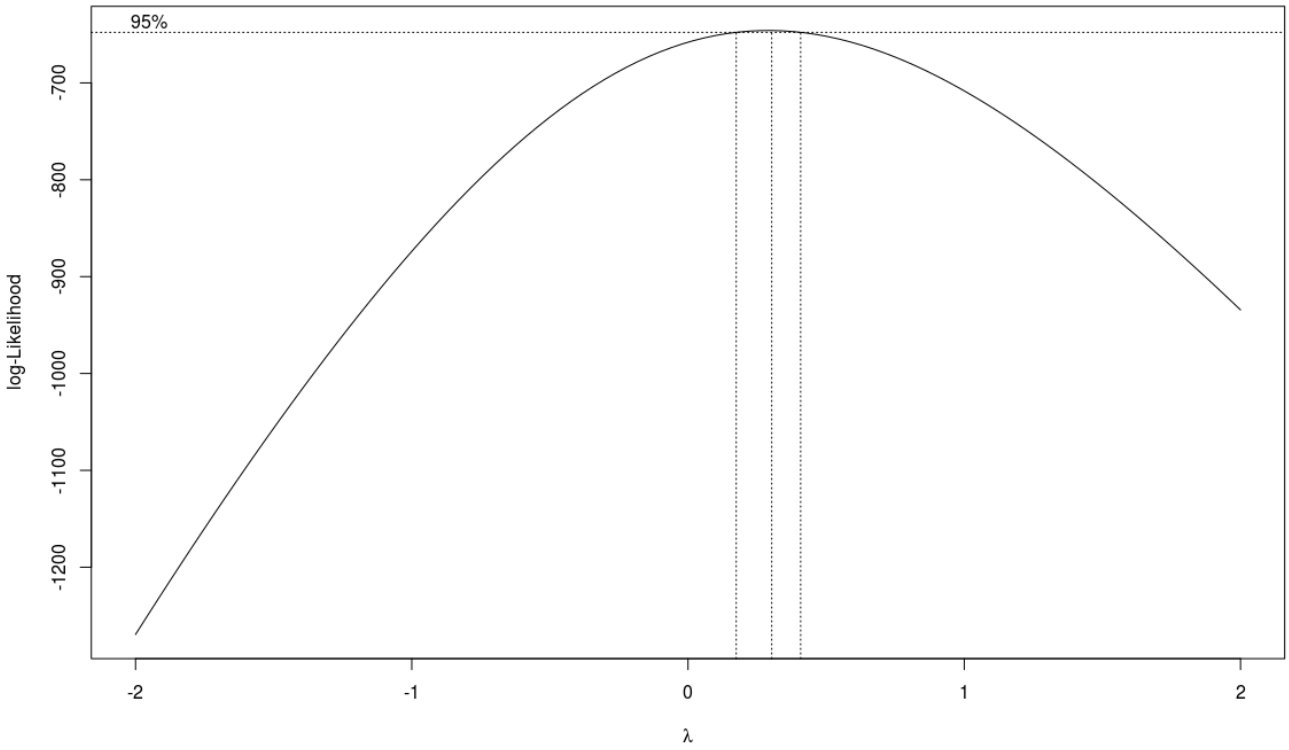


Page 4 A screenshot of your code for subproblem 2.

```
# Problem 2
# Remove the 3 outliers 365, 370, 373, 369 observed from the plot
Boston1 <- Boston[-c(365, 369, 370, 371, 372,373, 366, 413), ]
fit2=lm(medv~.,Boston1)
summary(fit2)
par(mfrow=c(2,2))
plot(fit2)
plot(fitted(fit2), residuals(fit2));
title("Residual vs Fit. value with outliers removed");
```

Page 5 A screenshot of Box-Cox transformation plot and the best value you chose

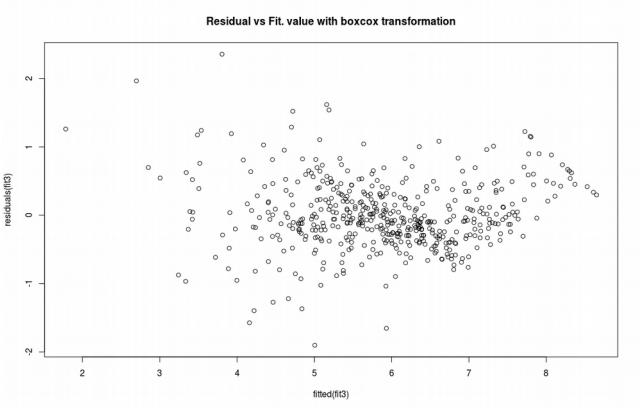
After applying the box-cox transformation to the model with outliers removed, the log plot is obtained:



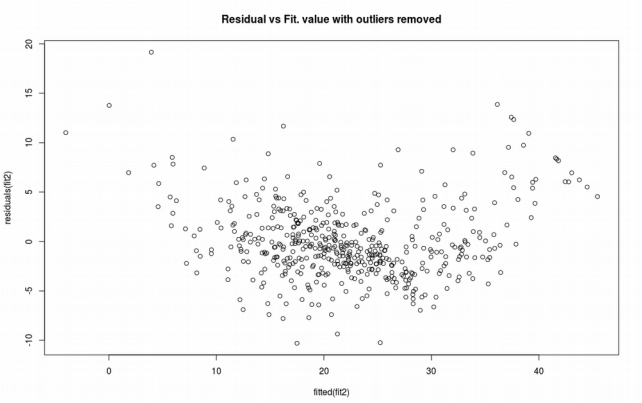
As seem from the plot and sorted in R, the lambda value with the highest log-likelihood is 0.3.

Page 6

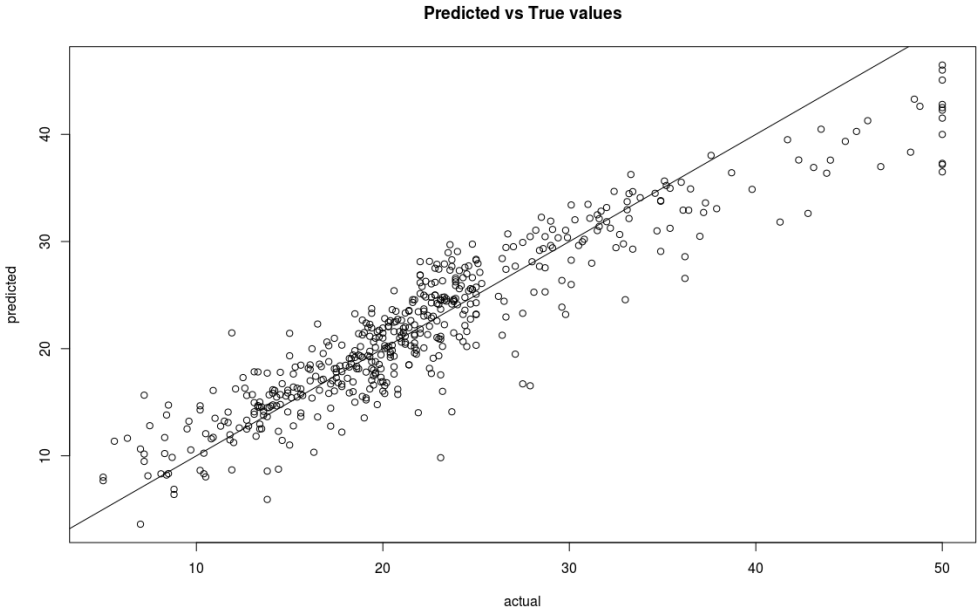
Using the lambda 0.3 value to transform the dependent values, the new linear regression model is obtained, and the standardized residual vs Fitted values plot is shown:



Compare to the previous one without box-cox transformation applied:



The “banana” shape has gone. Thus this suggests the transformation is helpful. The fitted house price against the true house price plot is shown:



Page 7 Code for subproblems 3 and 4

```
# Problem 3
# boxcox
bc <- boxcox(fit2)
lambda <- bc$x
lik <- bc$y
combined <- cbind(lambda, lik)
combined[order(-lik), ]
#lambda is  0.30303030

# regression with transformation
fit3=lm((medv^(1/3) - 1)/0.3~.,Boston1)
summary(fit3)
plot(fit3)
plot(fitted(fit3), residuals(fit3));
title("Residual vs Fit. value with boxcox transformation")

# Plot predicted against true values
plot(Boston1$medv, (predict(fit3) * 0.3 + 1 )^ 3, xlab="actual",ylab="predicted")
abline(a=0,b=1)
title("Predicted vs True values")
```