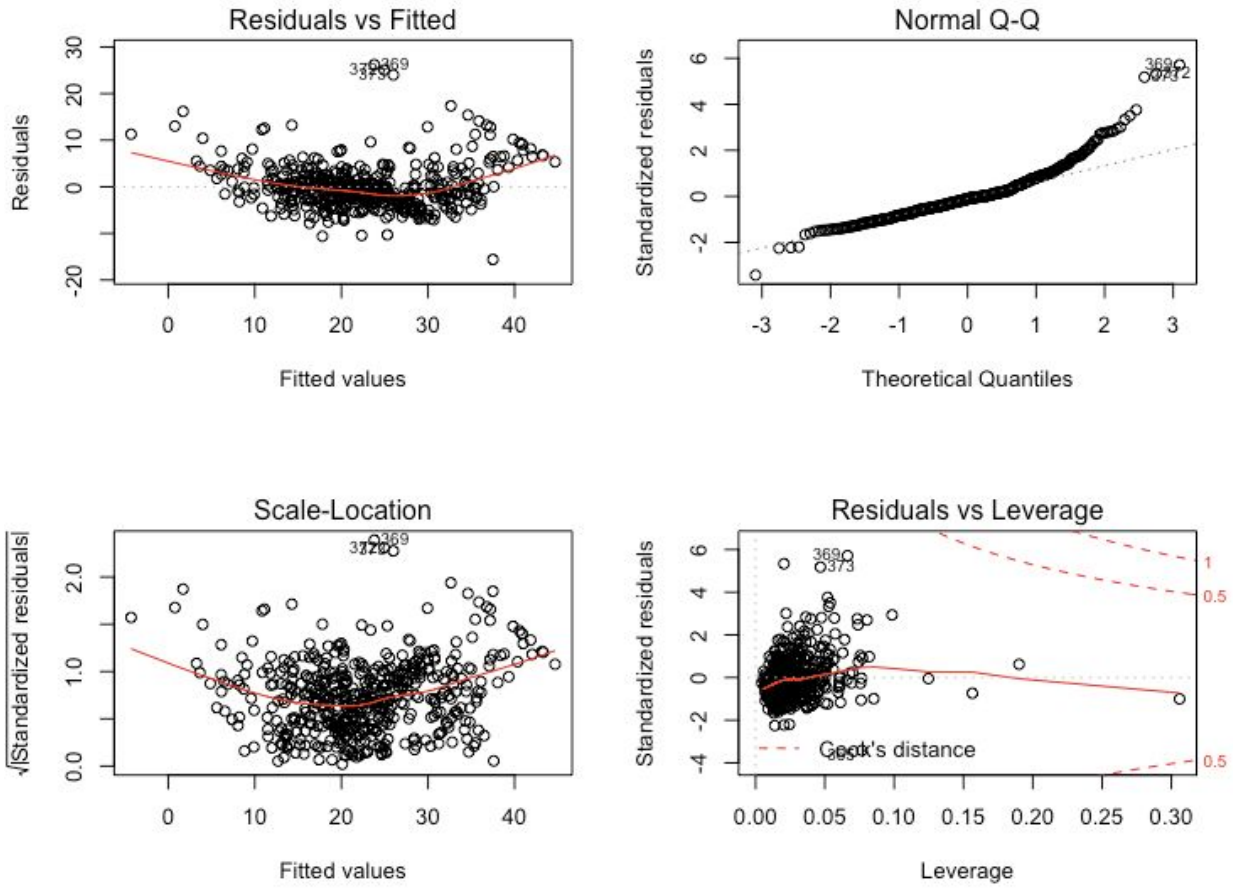
Group Members: Gassan Soukaev and Edgar Pino
CS498 Applied Machine Learning
CS498 AMO

Code for regression and resulting model

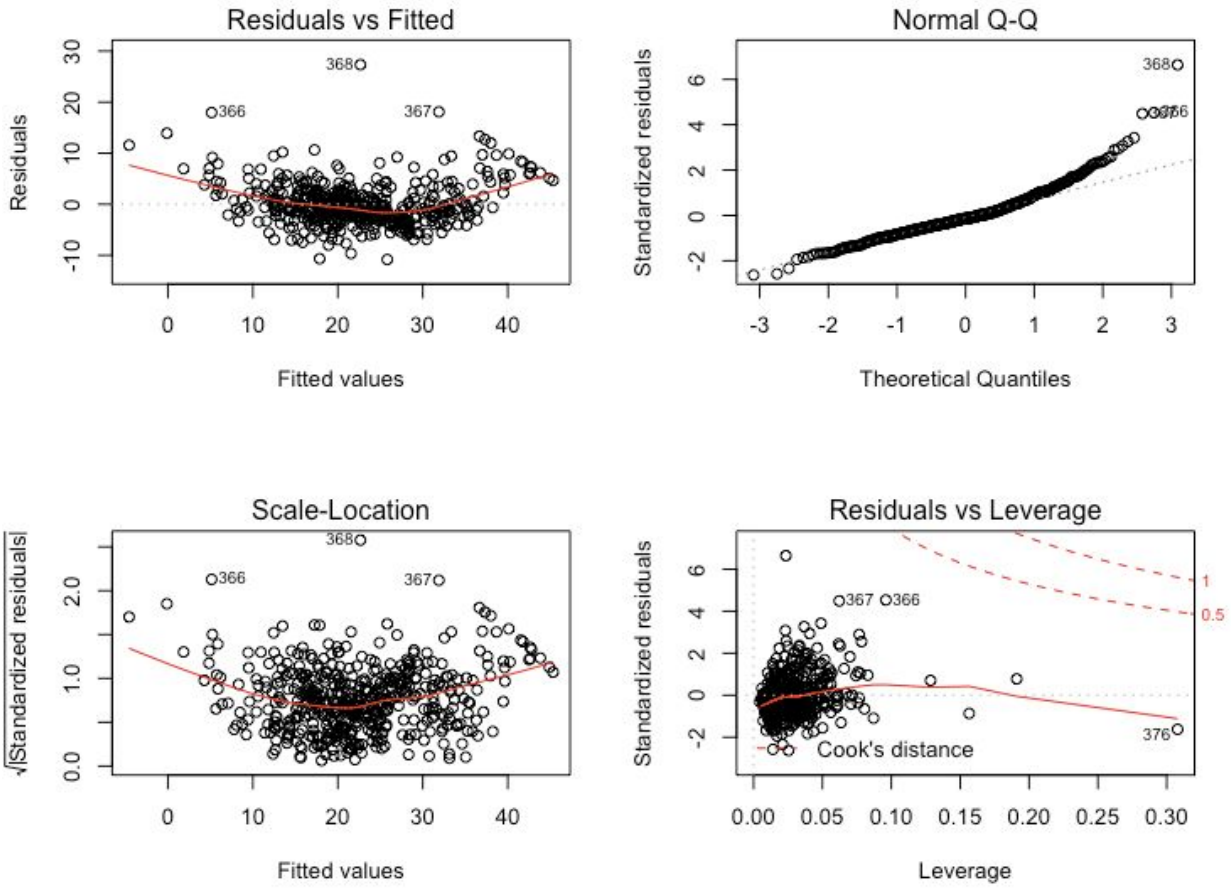
```
# Create the model
orig_model = lm(MEDV ~ ., data = housing_data)

summary(orig_model)
plot(rstandard(orig_model))

## Search for outliers manually
c_distance = cooks.distance(orig_model)
c_distance = data.frame(1:nrow(housing_data), c_distance)
colnames(c_distance) = c("idx", "value")

plot(c_distance$idx ~ c_distance$value)
text(c_distance$idx ~ c_distance$value, labels=idx, data=c_distance, cex=.75, font=2)
```

Diagnostic Plot



We decided to manually remove most of the values that were outside the range of the plots. The following values were removed: 365,366,369,373,370,413. This produced the best plot of fitted house price against true house price.

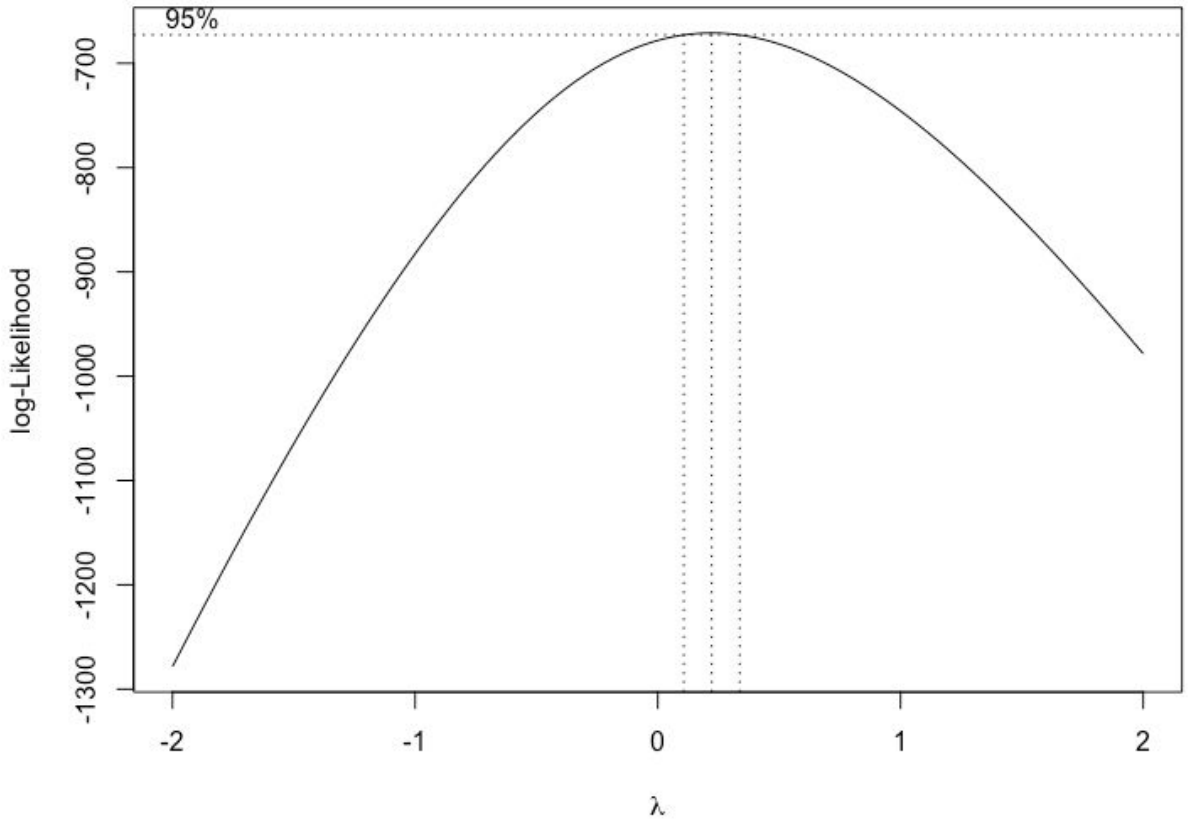
New diagnostic plot



Code for subproblem 2

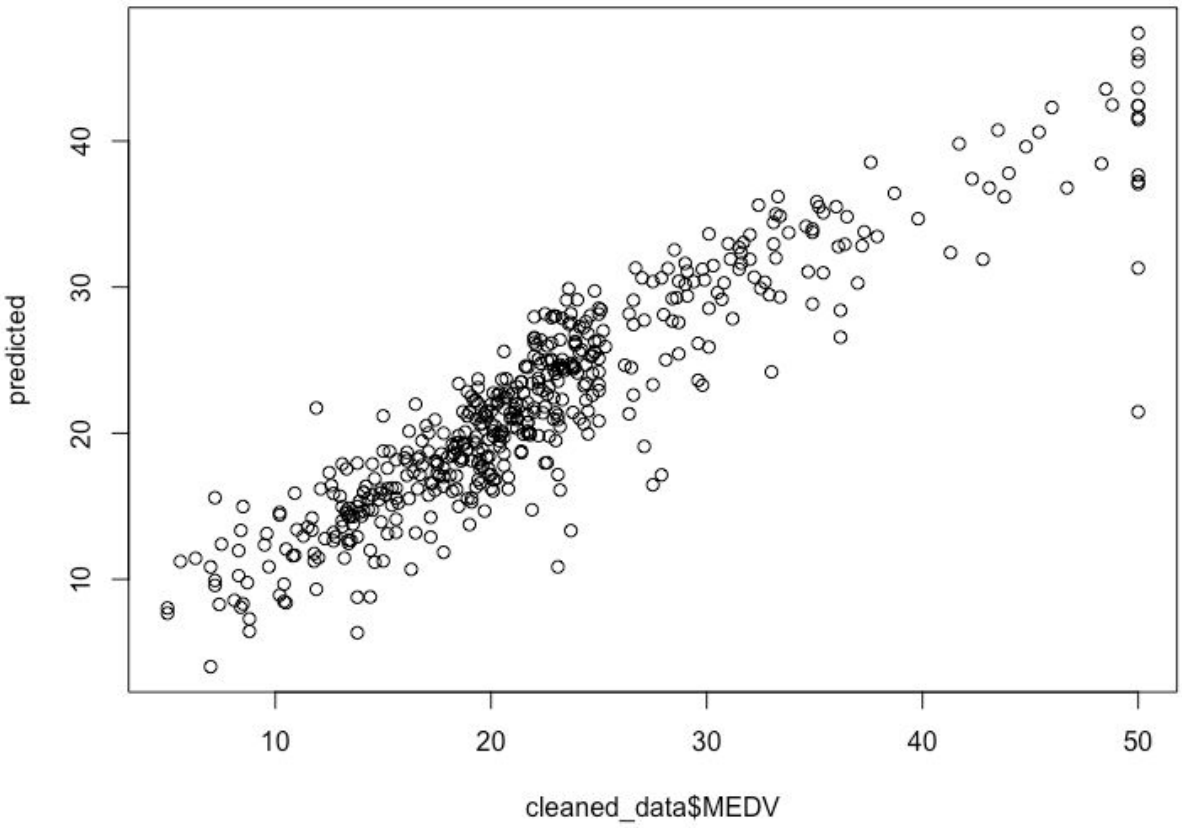
```
bc = boxcox(MEDV ~ .,  data=cleaned_data)
lambda = bc$x[which.max(bc$y)]
```

Box-Cox transformation plot



Best Value: 0.2222222

Results after Box-Cox



Code for subproblems 3 and 4

```
q4_data = cleaned_data
q4_data$MEDV = q4_data$MEDV ** lambda
new_fit = lm(MEDV ~ ., data=q4_data)
plot(rstandard(new_fit))
predicted = fitted(new_fit) ** (1/lambda)
plot(predicted ~ cleaned_data$MEDV)
```