Agricultural Land Rent Project

Variable Selection Analysis

Brady Fisher

Land Rent Problem:

It is thought that rent for land planted to alfalfa relative to rent for other agricultural purposes would be higher in areas with a high density of dairy cows and rents would be lower in counties where liming is required, since that would mean additional expense. Explore this data with regard to understanding rent structure. Summarize your results.

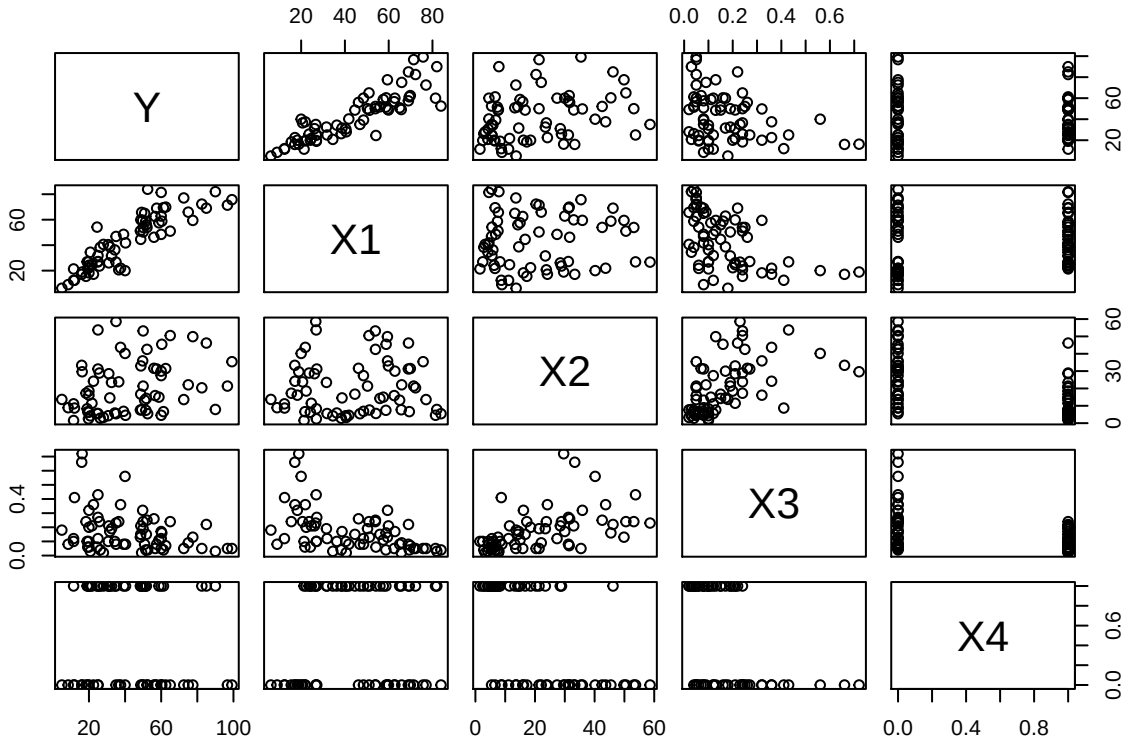
To see how the rent for land planted to alfalfa relative to rent for other agricultural purposes changes with high dairy cow density and the liming requirement, I will use Y as the response variable, where Y is the average land rent per acre planted to alfalfa. I will use the given explanatory variables X1(average rent paid for all tillable land), X2 (density of dairy cows), X3 (proportion of farmland used as pasture), and X4 (1 if liming is required, 0 otherwise).

```
data = data.frame(Y=landrent$Y,X1=landrent$X1,X2=landrent$X2,X3=landrent$X3,X4=landrent$X4)
head(data)
```

```
##       Y    X1    X2   X3 X4
## 1 18.38 15.50 17.25 0.24  0
## 2 20.00 22.29 18.51 0.20  1
## 3 11.50 12.36 11.13 0.12  0
## 4 25.00 31.84  5.54 0.12  1
## 5 52.50 83.90  5.44 0.04  0
## 6 82.50 72.25 20.37 0.05  1
```

Now I will create a scatter plot matrix to see how each explanatory variable relates to the response variable average rent per acre Y planted to alfalfa.

```
pairs(data)
```



From the scatter plot and problem description I can see that the variable X4 should be treated as a factor of 2 levels. I can also see from the scatter plot that the other predictors are relatively related to the response variable. Some of the relationships appear to be somewhat curved, specifically Y's relationship with X3 so I will test to see if transformations on any of the variables will help produce a better model.

Now I will take a look at the potential model without any transformations to confirm transformations will potentially be helpful.

```
m1 = lm(Y ~ X1+X2+X3+as.factor(X4), data=data)
summary(m1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + as.factor(X4), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2287  -4.8686  -0.0287   4.7547  27.7666
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.8282     4.6749  -0.605 0.547399
## X1                0.8833     0.0690  12.801  < 2e-16 ***
## X2                0.4318     0.1080   3.999 0.000172 ***
## X3              -11.3804    11.8937  -0.957 0.342359
## as.factor(X4)1   -1.0117     2.8490  -0.355 0.723706
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 9.311 on 62 degrees of freedom
## Multiple R-squared:  0.8404, Adjusted R-squared:  0.8301
## F-statistic:  81.6 on 4 and 62 DF,  p-value: < 2.2e-16
```

This output shows that the variables X1(average rent paid for all tillable land), and X2(density of dairy cows) are the only significant variables without any transformations. I can say this since the p-values less than 0.05 for these variables.

I will now find the best transformations for the non-binary explanatory variables using the powerTransform method.

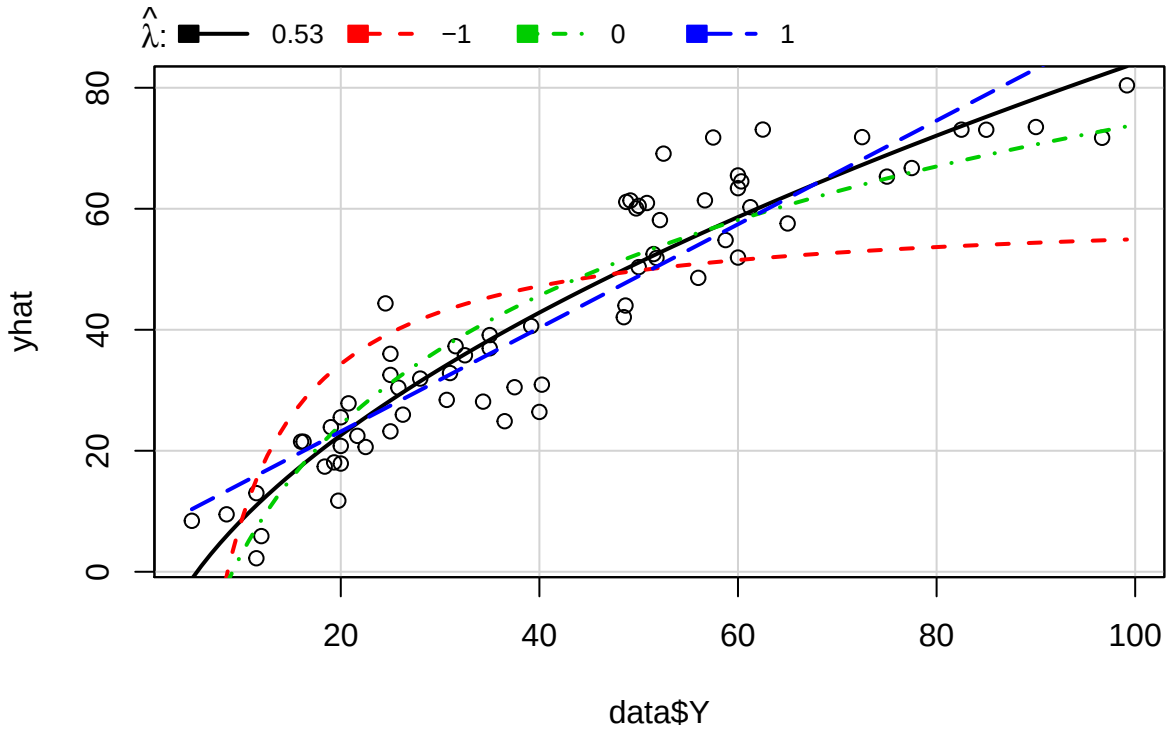
```
m2 = powerTransform(object=cbind(data$X1, data$X2, data$X3))
summary(m2)
```

```
## bcPower Transformations to Multinormality
##    Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
## Y1    0.7903   0.2030           0.3924           1.1882
## Y2    0.2371   0.1218          -0.0016           0.4759
## Y3    0.0825   0.0991          -0.1118           0.2768
##
## Likelihood ratio tests about transformation parameters
##                                 LRT df         pval
## LR test, lambda = (0 0 0)  23.155504  3 3.747847e-05
## LR test, lambda = (1 1 1) 102.374387  3 0.000000e+00
## LR test, lambda = (1 0 0)   5.253666  3 1.541374e-01
```

This output shows that I should transform X2 and X3 by taking their log, since their estimated powers are close to 0. I should leave X1 since its estimated power is close to 1.

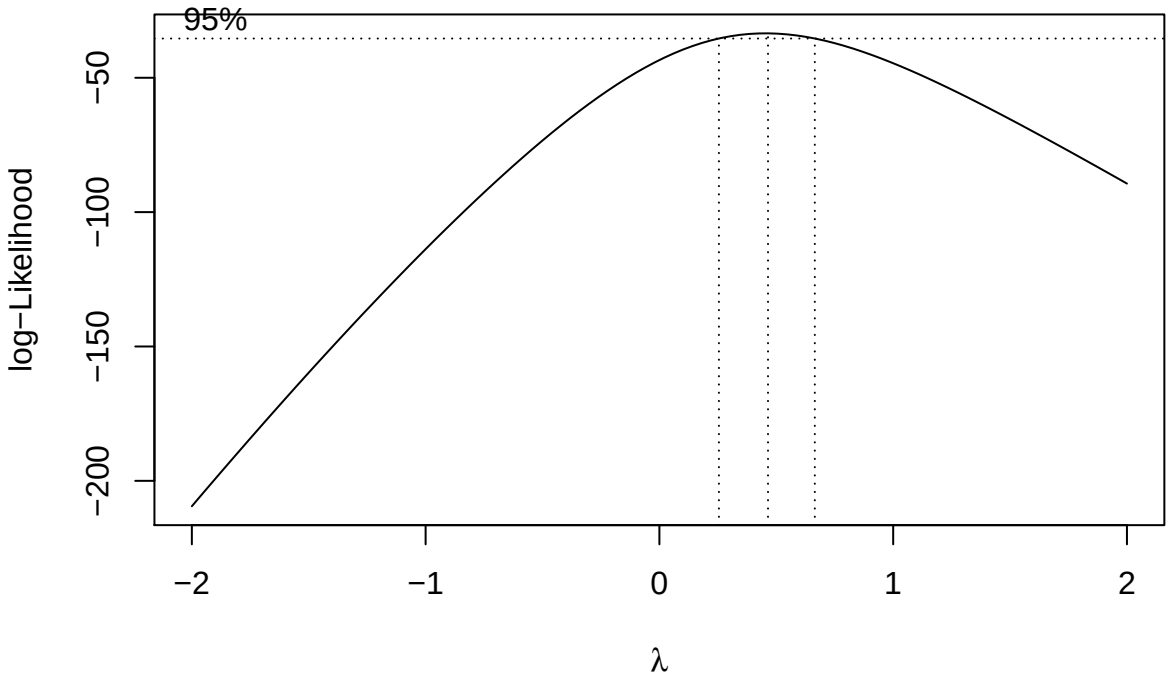
So my new potential full model will be as follows denoted m3. I will then run the inverseReponsePlot and the boxcox methods to see if the variable Y needs to be transformed.

```
m3 = lm(data$Y ~ data$X1 + log(data$X2) + log(data$X3) + as.factor(data$X4))
inverseResponsePlot(m3)
```




```
##       lambda       RSS
## 1  0.5298776  3456.775
## 2 -1.0000000 13700.624
## 3  0.0000000  4623.946
## 4  1.0000000  4138.313
```

```
boxcox(m3)
```



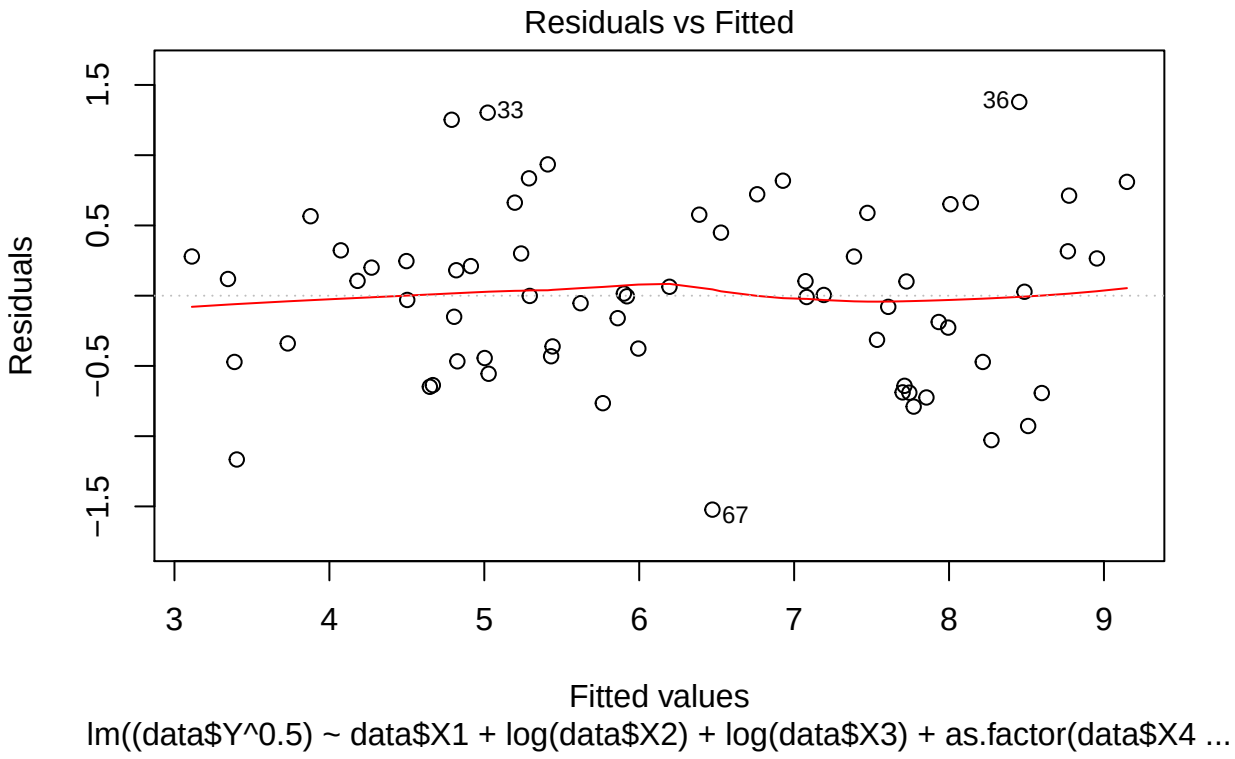
From both graphs I can see that for Y, raising it to the 1/2 would be the best transformation and may produce the better model.

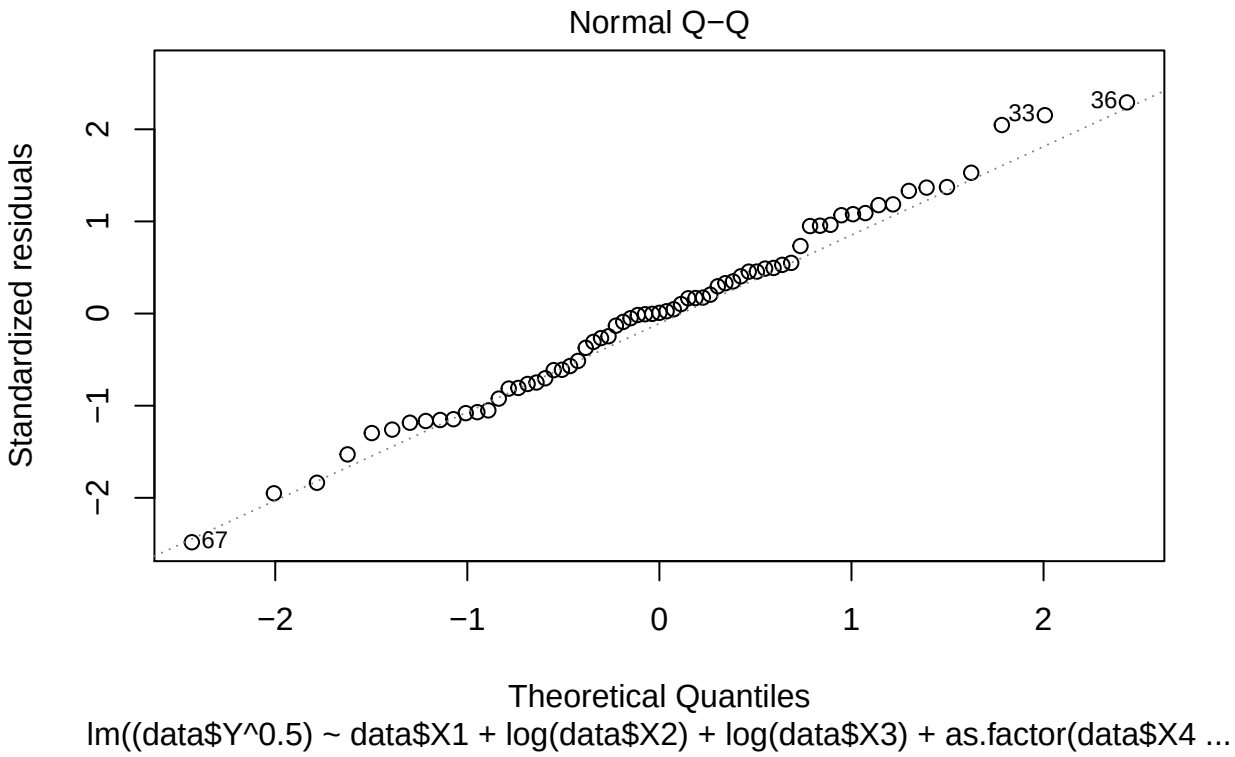
Now my new model is:

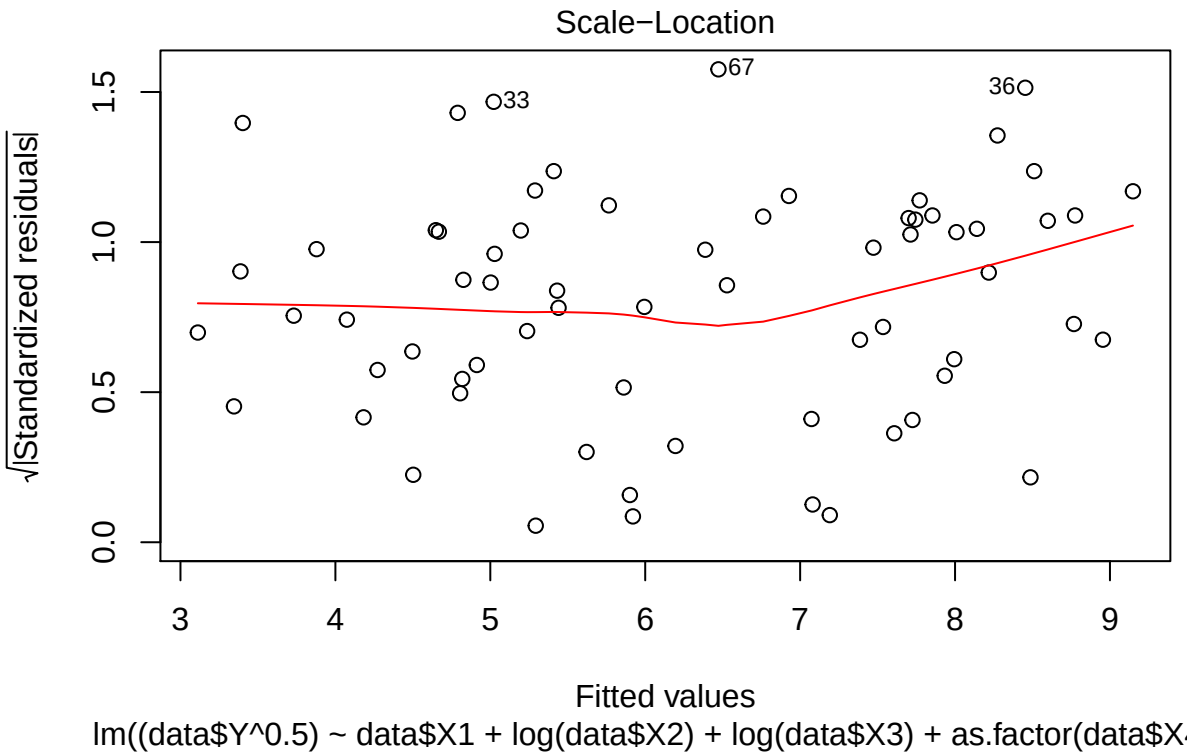
```
m4 = lm((data$Y ^.5)~  data$X1 + log(data$X2) + log(data$X3) + as.factor(data$X4))
```

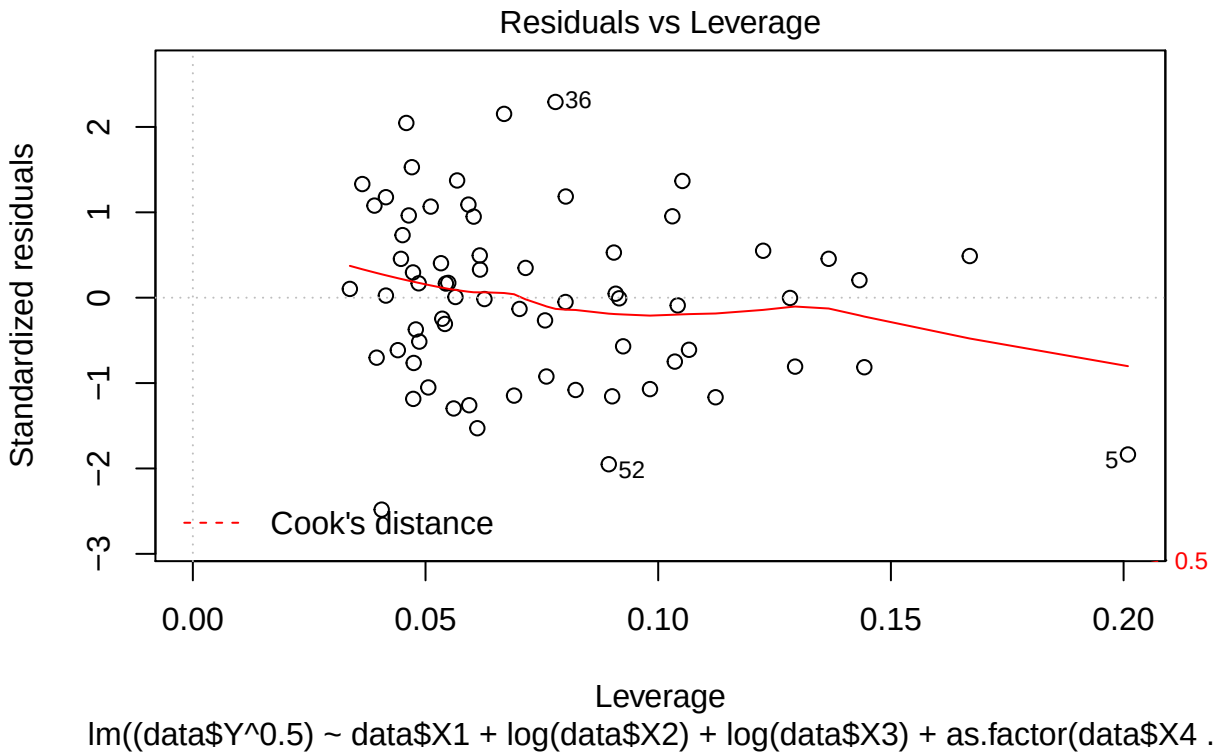
I will now check the model with a Residuals vs Fitted plot and a normal QQ Plot to check that all the appropriate assumptions are met.

```
plot(m4)
```


Residuals vs Fitted
Residuals
1.5
0.5
−0.5
−1.5
33
36
67
3
4
5
6
7
8
9
Fitted values
lm((data$Y^0.5) ~ data$X1 + log(data$X2) + log(data$X3) + as.factor(data$X4 ...


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm((data$Y^0.5) ~ data$X1 + log(data$X2) + log(data$X3) + as.factor(data$X4 ...
67
33
36
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
3
4
5
6
7
8
9
33
67
36
Fitted values
lm((data$Y^0.5) ~ data$X1 + log(data$X2) + log(data$X3) + as.factor(data$X4 ...



From the Residuals vs. Fitted plot I can see that the model is approximately linear as the points appear to be randomly scattered around residuals of 0 with no clear pattern.

Also I can see that there is approximately equal variance as the spread of the points do not appear to change too much. There does appear to be one point that sticks out more than the rest, that being point 67, but from the Residuals vs Leverage Plot I can see that it falls below the 0.5 line, meaning it is not too influential.

From the QQ Plot I can see that the data is not exactly normal, but that should not matter too much because the data has a large sample size making the model resistant in the Normality assumption.

Lastly I will assume the data was taken so that they are independent of each other, thus satisfying the independence assumption.

Finally, I will view the summary of the model to see which explanatory variables appear to be significant in predicting the response.

```
summary(m4)
```

```
##
## Call:
## lm(formula = (data$Y^0.5) ~ data$X1 + log(data$X2) + log(data$X3) +
##     as.factor(data$X4))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.52289 -0.45559  0.00498  0.31923  1.37890
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)          0.647748   0.626563   1.034    0.305
## data$X1              0.068926   0.005411  12.738  < 2e-16 ***
## log(data$X2)         0.786132   0.156736   5.016 4.71e-06 ***
## log(data$X3)        -0.158992   0.172536  -0.922    0.360
## as.factor(data$X4)1  0.293746   0.195758   1.501    0.139
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6264 on 62 degrees of freedom
## Multiple R-squared:  0.8837, Adjusted R-squared:  0.8761
## F-statistic: 117.7 on 4 and 62 DF,  p-value: < 2.2e-16
```

From the summary I can see that the X1(average rent paid for all tillable land) and X2 (the density of dairy cows) variables are significant with pvalues less than 0.05. This means that I have evidence to believe the density of dairy cows and average rent paid for all tillable land is significant in predicting the rent for land planted to alfalfa relative to rent for other agricultural purposes. Also this means I have no evidence to believe whether or not liming is required nor the amount of pasture has any effect on the land rent.