Advanced Analysis of Real Estate Valuation and Seed Germination Data

Sydney Murphy

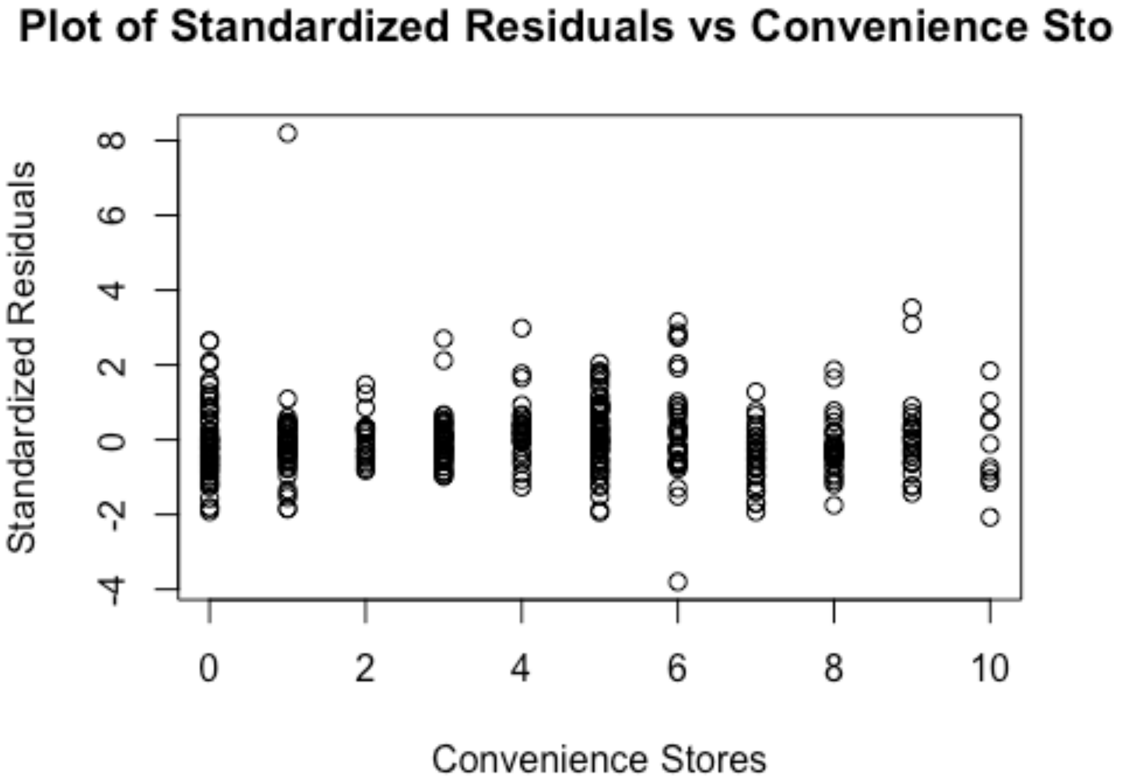
2023-12-11

The data in real-estate-valuation-data-set.csv is a subset of the dataset hosted at https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set that contains information about the unit price of houses in New Taipei City, Taiwan.

Load again the data in R. In homework 2, we noticed that the relationship between unit_price and distance appears to be exponential. This suggests that using the logarithm of distance instead of distance might help. Fit again the multiple linear regression model of homework 2, where unit_price is regressed on convenience_stores and on the logarithm of distance.

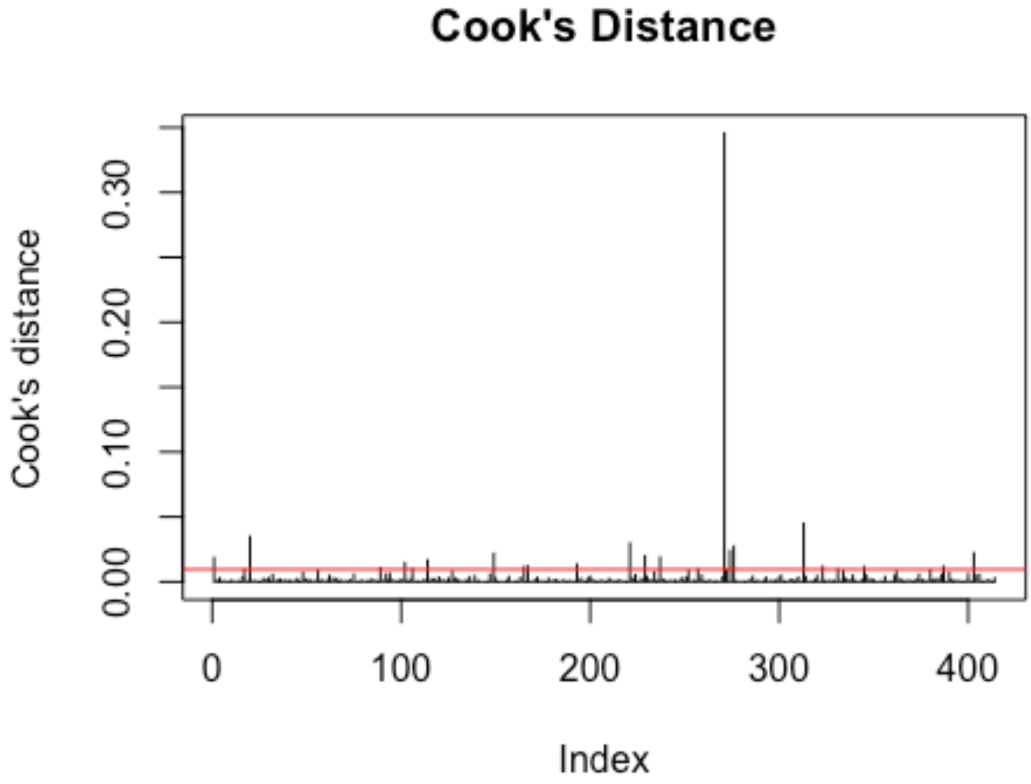
```
## [1] "age"                "distance"           "convenience_stores"
## [4] "unit_price"

## 
## Call:
## lm(formula = unit_price ~ convenience_stores + log_distance, 
##     data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -34.783  -5.106  -0.756   3.462  74.582 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         85.8141     4.2006   20.43  < 2e-16 ***
## convenience_stores   0.5891     0.2104    2.80  0.00536 ** 
## log_distance        -7.8611     0.5536  -14.20  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 9.171 on 411 degrees of freedom
## Multiple R-squared:  0.5479, Adjusted R-squared:  0.5457 
## F-statistic:   249 on 2 and 411 DF,  p-value: < 2.2e-16
```

Plot the standardized residuals of this model against the predictor convenience_stores. Comment on the diagnostic plot. Do you see anything suspicious that might indicate problems with the model?



```
## Overall, the plot does not show obvious signs of model violations
##      such as heteroscedasticity or non-linearity. The residuals do not
##      show a clear pattern or trend with respect to the number of
convenience
##      stores. The spread of residuals seems consistent across the different
##      values of convenience stores, which indicates homoscedasticity. There
##      seems to be no increase or decrease in the spread of residuals as the
##      number of convenience stores changes. Though there are a few points
##      with higher residuals (above 2 and below -2), they aren't
significant.
##      However, there are noticeable residuals that are quite far from the
##      center, particularly the point above 6. I consider this point as an
##      outlier. Most of the residuals are distributed around the zero line,
##      which aligns with a well-fitted model.
```

```
## I am interested in loking more into the influence of the outliers on
##     the model. I am thinking this could be done by checking their leverage
##     and influence measures, such as 'Cook's distance'. I believe that
would
##     a fomula like this: 4/(n–k–1), where n is the number of observations
and
##     k is the number of predictors.
```

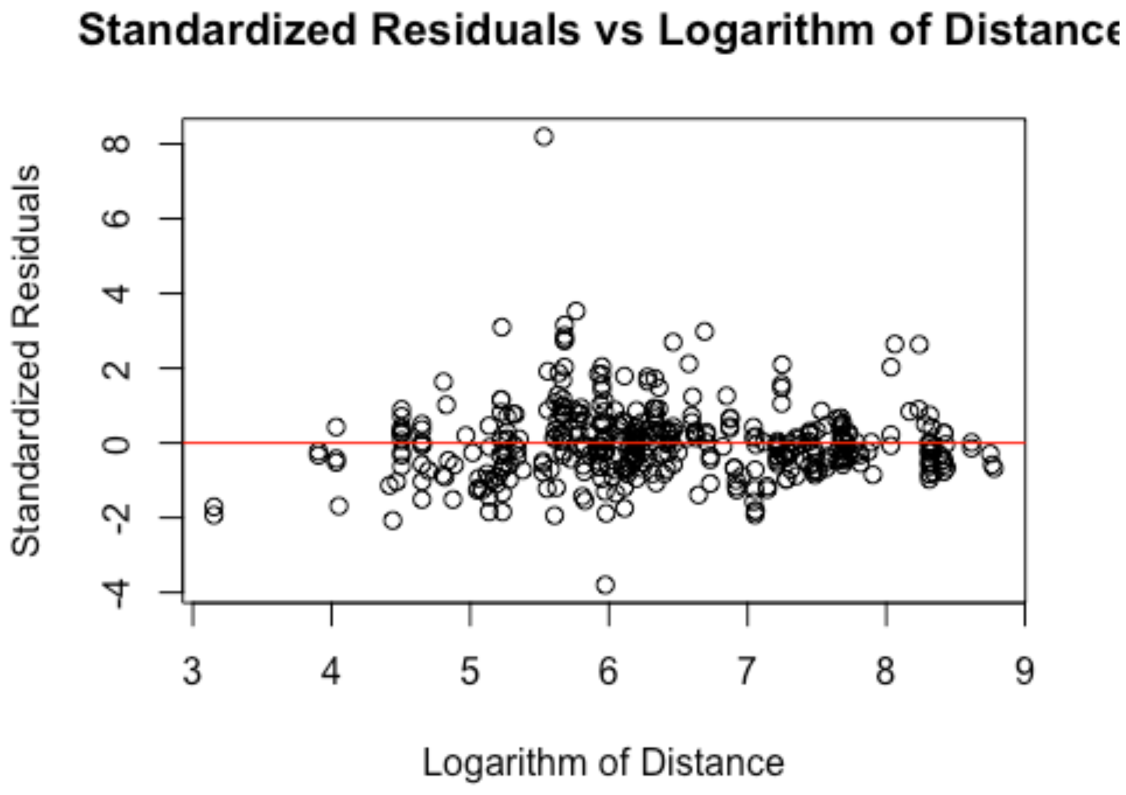


```
##   1  20  89 102 106 114 149 165 167 193 221 229 237 271 274 276 313 323
345 387
##   1  20  89 102 106 114 149 165 167 193 221 229 237 271 274 276 313 323
345 387
## 403
## 403
```

```
## The influential observations suggest that while the model fits well
## for the majority of the data, there are specific points that are not well
## explained by this model.
```

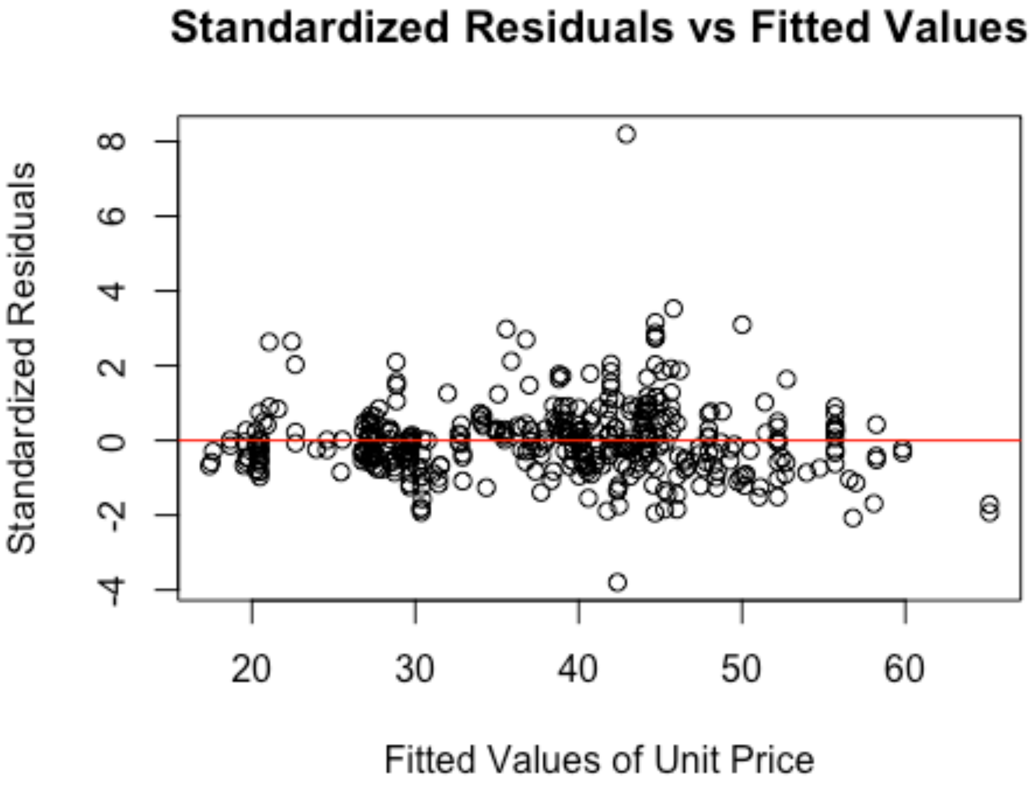
```
##      age   distance convenience_stores unit_price log_distance
## 1   32.0   84.87882                 10       37.9     4.441225
## 20   1.5   23.38284                  7       47.7     3.152002
## 89   8.9 1406.43000                  0       48.0     7.248810
## 102 12.7  170.12890                  1       32.9     5.136556
## 106  0.0  292.99780                  6       71.0     5.680165
## 114 14.8  393.26060                  6        7.6     5.974472
## 149 16.4 3780.59000                  0       45.1     8.237635
## 165  0.0  185.42960                  0       55.2     5.222675
## 167  0.0  292.99780                  6       73.6     5.680165
## 193 43.8   57.58945                  7       42.7     4.053339
## 221 37.2  186.51010                  9       78.3     5.228485
## 229 11.9 3171.32900                  0       46.6     8.061906
## 237  3.6  373.83890                 10       61.9     5.923825
## 271 10.8  252.58220                  1      117.5     5.531737
## 274 13.2  170.12890                  1       29.3     5.136556
## 276  1.5   23.38284                  7       49.7     3.152002
## 313 35.4  318.52920                  9       78.0     5.763714
## 323 12.9  187.48230                  1       33.1     5.233684
## 345 34.6 3085.17000                  0       41.2     8.034362
## 387  0.0  185.42960                  0       55.3     5.222675
## 403 12.7  187.48230                  1       28.5     5.233684
```

Plot the standardized residuals of this model against the predictor logarithm of distance. Comment on the diagnostic plot. Do you see anything suspicious that might indicate problems with the model?



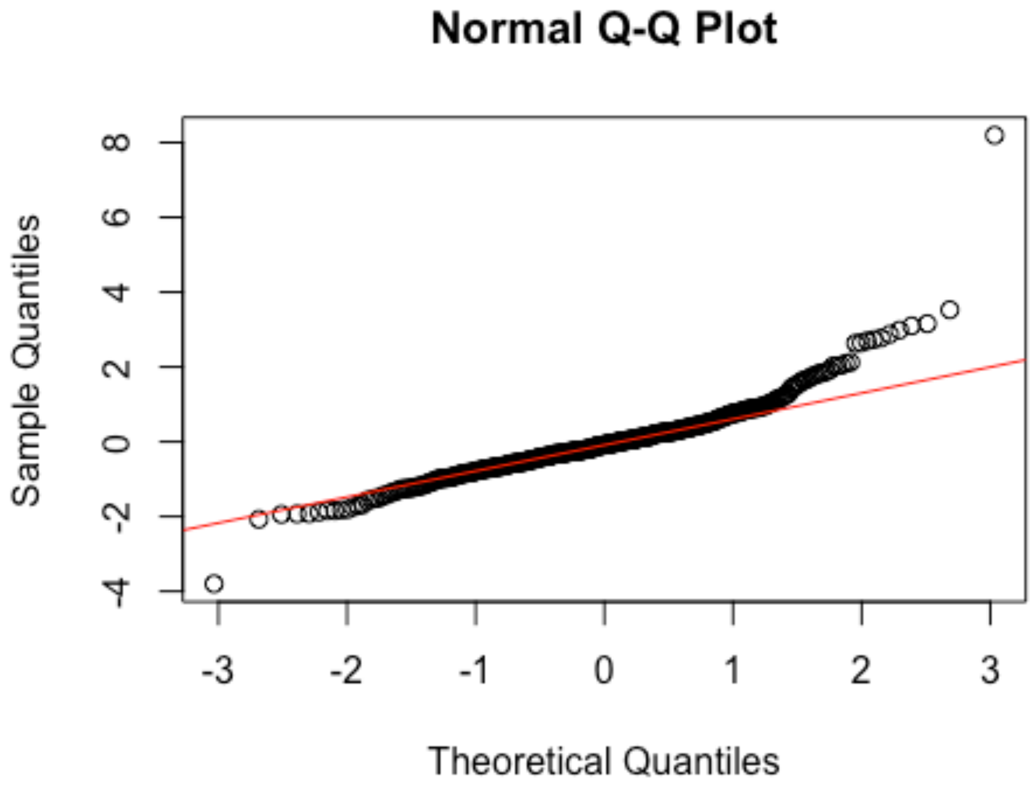
```
## Overall, the plot shows that the model fits well with the data, except
## for the potential presence of influential outliers, which I already found
after
## calculating the Cook's Distance. The residuals don't follow a pattern of
the
## logarithm of distance, which suggests that the model does not express
## non-linearity when comparing the standardized residual models vs log of
##     distance.
```

Plot the standardized residuals of this model against the fitted values of unit_price. Comment on the diagnostic plot. Do you see anything suspicious that might indicate problems with the model?



```
## The model does not show any clear signs of heteroscedasticity or
##     non-linearity. But similar to the previous plots, there are a few
points
##     with relatively high residuals. For example, the point with a residual
##     above 8. It shows that point could be an influential outlier.
```

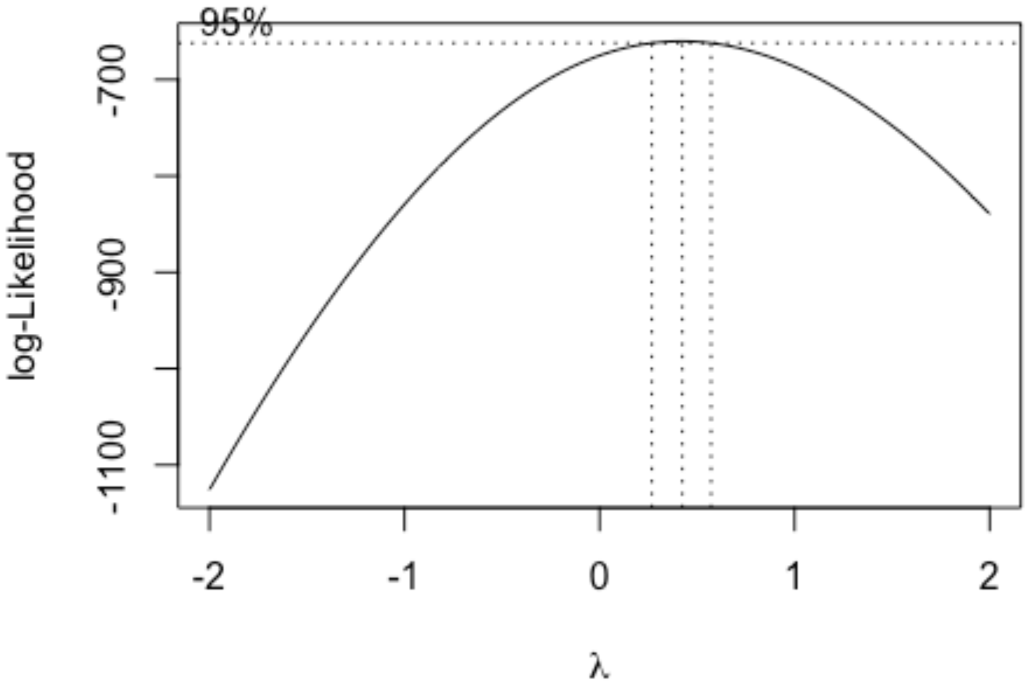
Plot the quantile-quantile plot of the standardized model residuals. Comment on the diagnostic plot. Do you see anything suspicious that might indicate problems with the model?



```
## The plot shows that the residuals of the model are mostly normally
## distributed and that there are outliers(particularly in the upper tail)
## suggesting that there may be things affecting its normality.
```

Apply the Box-Cox method to find the optimal power λ for the response variable unit_price when the predictors are convenience_stores and the logarithm of distance.

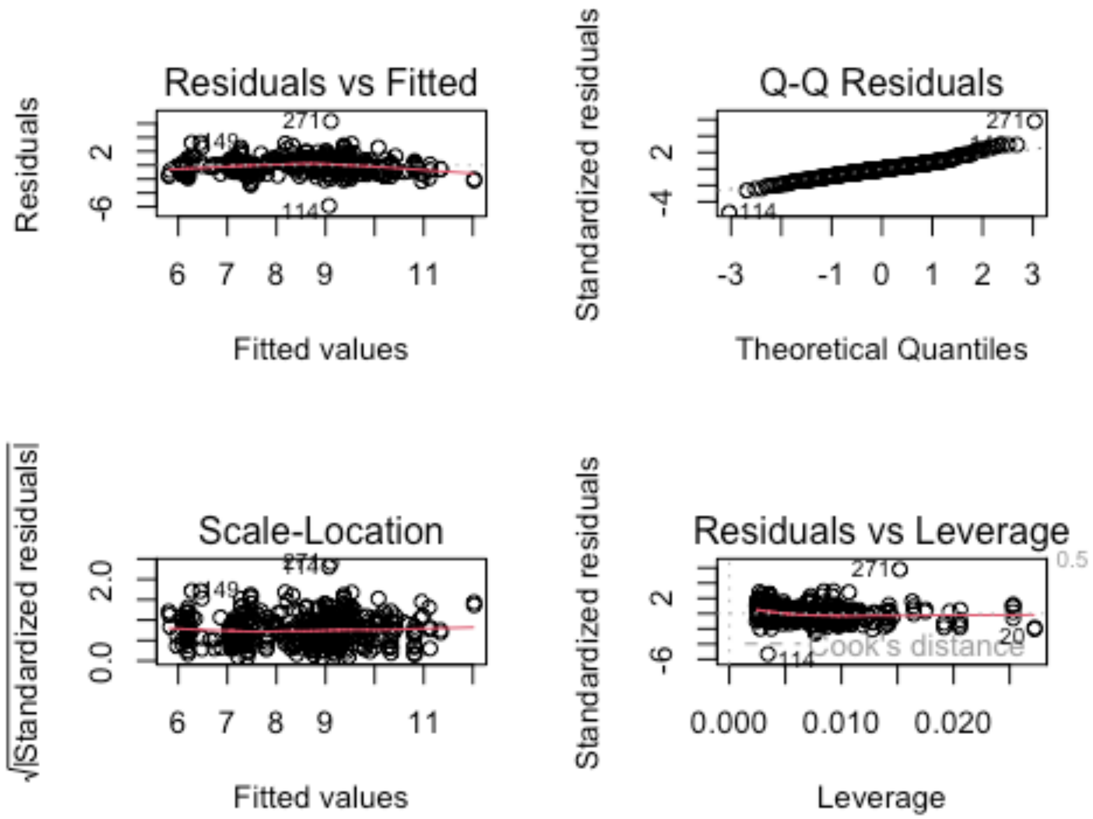
```
## Loading required package: MASS
```



Fit a multiple linear regression model of unit_priceλ (where λ is the value you found using the Box-Cox method) on convenience_stores and on the logarithm of distance.

```
## 
## Call:
## lm(formula = unit_price_transformed ~ convenience_stores + log_distance, 
##     data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.8562 -0.6338 -0.0362  0.5655  6.3399 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        14.65548    0.50541  28.997  < 2e-16 ***
## convenience_stores  0.08252    0.02532   3.259  0.00121 ** 
## log_distance       -1.01747    0.06661 -15.276  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.103 on 411 degrees of freedom
## Multiple R-squared:  0.589,  Adjusted R-squared:  0.587 
## F-statistic: 294.5 on 2 and 411 DF,  p-value: < 2.2e-16
```

Recreate the diagnostic plots for this new model and comment on them. Did the Box-Cox method produce any improvements?



```
## The diagnostic plots show that the Box-Cox transformation improved the
##     model fit. The residuals seem to be more consistent with assumptions
of
##     homoscedasticity and normality, but the outliers remain a concern.
##     The plot labeled 'Residuals vs Fitted', in particular, could be
influencing
##     the model.
```

Use the influence.measures function to compute the DFBETAs for the two predictors with respect to the model that you fitted in Question 8. Are there observations that are flagged as influential with respect to the DFBETAs scores?

The data in the germ dataset of the GLMsData library contains information ex- periments where the number of seed germinations were recorded for two extracts: beans and cucumbers.

```
##  num [1:414, 1:3] -0.03172 0.01575 -0.00756 -0.01284 0.00363 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:414] "1" "2" "3" "4" ...
##   ..$ : chr [1:3] "(Intercept)" "convenience_stores" "log_distance"
## [1] "(Intercept)"        "convenience_stores" "log_distance"
```

```
##   1  20  32  56  75  89  92  94 102 114 139 147 165 221 224 230 234 237 
252 257 
##   1  20  32  56  75  89  92  94 102 114 139 147 165 221 224 230 234 237 
252 257 
## 271 272 274 276 313 323 331 374 387 400 403 
## 271 272 274 276 313 323 331 374 387 400 403
```

```
##  20  50 102 117 147 149 165 193 229 237 257 271 274 276 301 313 323 334 
348 374 
##  20  50 102 117 147 149 165 193 229 237 257 271 274 276 301 313 323 334 
348 374 
## 387 400 403 
## 387 400 403
```

Load the germ data in R and fit a logistic regression model for the proportion of seeds that germinated Germ / Total onto the predictors Extract and Seeds.

```
##   Germ Total  Extract Seeds
## 1   10    39     Bean  OA75
## 2   23    62     Bean  OA75
## 3   23    81     Bean  OA75
## 4   26    51     Bean  OA75
## 5   17    39     Bean  OA75
## 6    5     6 Cucumber  OA75
```

```
## 
## Call:
## glm(formula = GerminationRate ~ Extract + Seeds, family = quasibinomial, 
##     data = germ)
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)   
## (Intercept)       -0.9653     0.2642  -3.654  0.00181 **
## ExtractCucumber    1.0404     0.2886   3.604  0.00203 **
## SeedsOA75          0.6493     0.2884   2.251  0.03710 * 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for quasibinomial family taken to be 0.09901371)
## 
##     Null deviance: 3.9112  on 20  degrees of freedom
## Residual deviance: 2.0280  on 18  degrees of freedom
## AIC: NA
## 
## Number of Fisher Scoring iterations: 4
```

Question 11

Analyze the model summary and answer the following questions. 1. What is the baseline category of each categorical predictor in the model? 2. What are the odds of germination for the baseline combination of Extract and Seeds according to the model? 3. According to the model, by how much are the odds of germination for extracts of type Cucumber and seed of type 0A73 larger/smaller than the odds of germination for the baseline combination of Extract and Seeds? 4. According to the model, by how much are the odds of germination larger/smaller for extracts of type Beans when the seed is 0A75 compared to the odds of the baseline combination of Extract and Seed? 5. Finally, by how much are the odds of germination larger/smaller for ex- tracts of type Cucumber when the seed is 0A75 compared to the odds of the baseline combination of Extract and Seed?

```
## 1. For Extract: Since the coefficient in the output is for
##     'ExtractCucumber' the baseline category for Extract is 'Bean'.
##     The coefficient for 'ExtractCucumber' represents the change in
##     log odds of germination rate when using 'Cucumber' extract vs
##     'Bean' extract. For Seeds: Since 'SeedsOA75' appears in the output,
##     it means that 'OA73' is the baseline category.

## 2. The intercept is -0.9653, representing the log odds of germination
##     when the extract is 'Bean' and the seed type is 'OA73'. I converted
the
##     log odds to odds by exponentiating the intercept and found that the
odds
##     of a seed germinating are about: [Odds] to 1

## [1] 0.3808689

## 3. The coefficient for ExtractCucumber is 1.0404. By calculating the
change
##     in odds, I concluded that using 'Cucumber' extract, as opposed to
'Bean',
##     increases odds of a seed germination by:

## [1] 2.830349

## 4.1 I calculated the change in odds for 'Bean' extracts with '0A75' seeds
##     compared to the baseline ('Bean' and '0A73'), using the coefficient
for
##     SeedsOA75, which is 0.6493. The change in odds for SeedsOA75 is:

## [1] 1.9142

## 4.2 Therefore, the odds of germination for extracts of type 'Beans' with
##     seed '0A75' are larger than the odds of germination for the baseline
##     combo of 'Bean' extract and '0A73' seeds.

## The coefficient for ExtractCucumber is 1.0404.

## The coefficient for SeedsOA75 is 0.6493
## 5.1 I see an increase in the odds of germination due to the combined
effect
##     of using 'Cucumber' extract and '0A75' seeds,as the combined effect
is:

## [1] 1.6897

## 5.2 And the change in odds is:

## [1] 5.417855

## 5.3 This is the multiplicative change in odds due to using 'Cucumber'
## extract and '0A75' seeds together.
```