Homework3

Sherine George

13 November, 2023

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. The subset of the data that we will use contains the following columns: • age: the age of the house in years • distance: the distance to the nearest Mass Rapid Transit (MRT) station from the house (in meters) • convenience_stores: the number of convenience stores near the house • unit_price: the unit price of the house, measured in 10,000 New Taiwan Dollars/Ping (where 1 Ping = 3.3 squared meters).

Question 1 - 5 points

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.

Solution:

```
data <- read_excel("D:\\Downloads\\real+estate+valuation+data+set (2)\\Real estate valuation data set.xl
head(data)
```

```
## # A tibble: 6 x 8
##      No `X1 transaction date` `X2 house age` distance convenience_stores
##   <dbl>                  <dbl>          <dbl>    <dbl>              <dbl>
## 1     1                  2013.           32       84.9                 10
## 2     2                  2013.           19.5    307.                   9
## 3     3                  2014.           13.3    562.                   5
## 4     4                  2014.           13.3    562.                   5
## 5     5                  2013.            5      391.                   5
## 6     6                  2013.            7.1   2175.                   3
## # i 3 more variables: `X5 latitude` <dbl>, `X6 longitude` <dbl>,
## #   unit_price <dbl>
```

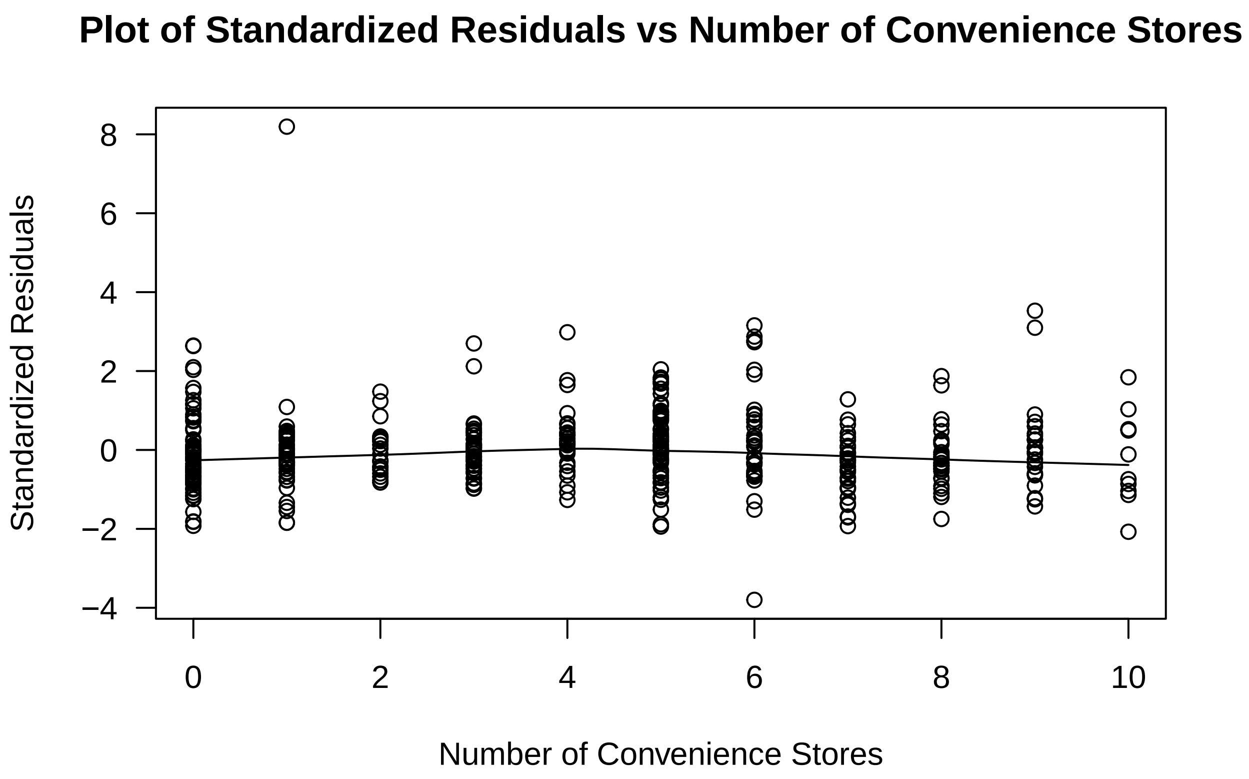
```
data$log_distance <- log(data$distance)
model <- lm(unit_price ~ convenience_stores + log_distance, data = data)
summary(model)
```

```
##
## 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
```

Question 2 - 10 points

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?

```
Diagnosti_Plot1 <- standard_res <- rstandard(model)
scatter.smooth(standard_res ~ data$convenience_stores,
las = 1,
  xlab = "Number of Convenience Stores",
  ylab = "Standardized Residuals",
  main = "Plot of Standardized Residuals vs Number of Convenience Stores")
```

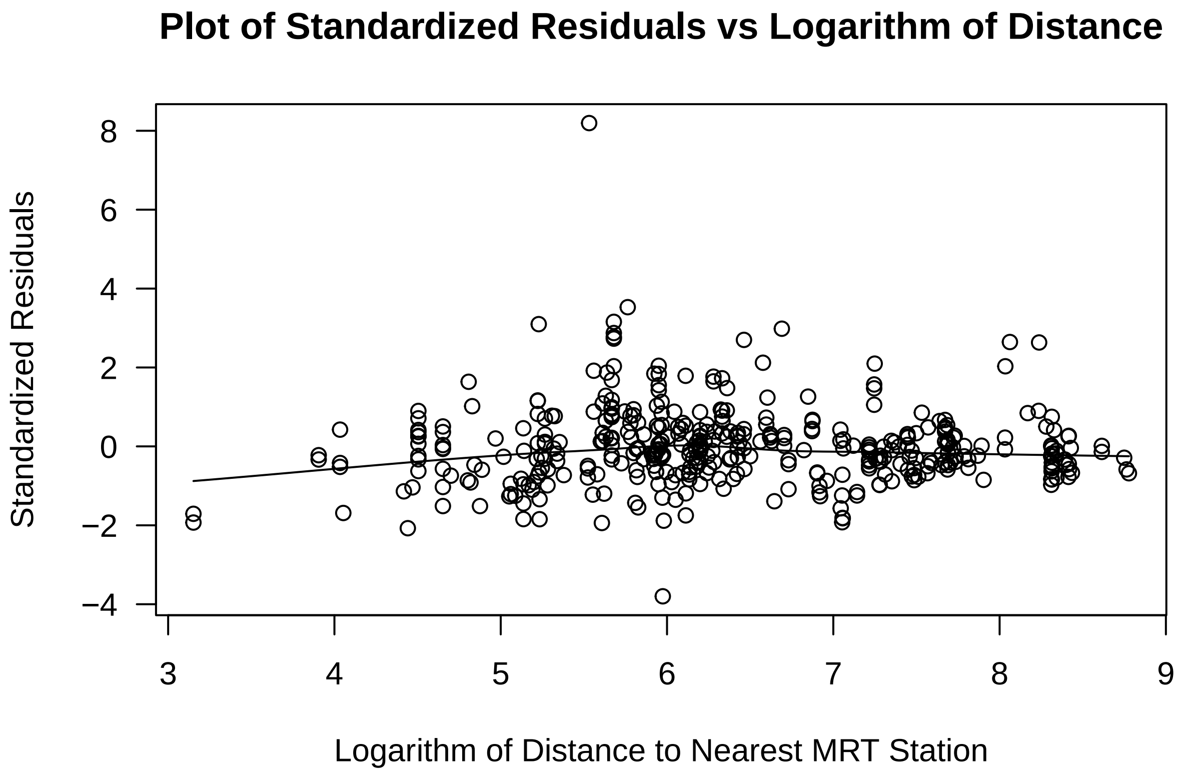


The plot of standardized residuals against the number of convenience stores reveals no discernible trends or curvature. This lack of clear patterns implies that the model adequately captures the linear relationship between the predictor variable (convenience_stores) and the response variable (unit_price). Consequently, concerns about the model's ability to represent this relationship in terms of linearity are alleviated.

However, the plot does draw attention to outliers in specific scenarios. For instances where only one store is nearby, the model tends to inaccurately represent higher-than-predicted house prices. Conversely, when there are six stores in proximity, the model struggles to accurately capture lower-than-average house prices. These outliers suggest areas where the model may benefit from further refinement to better account for variations in house prices under these particular conditions. **Question 3 - 10 points**

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?

```
Diagnosti_Plot2<-scatter.smooth(
standard_res ~ data$log_distance,
las = 1,
  xlab = "Logarithm of Distance to Nearest MRT Station",
  ylab = "Standardized Residuals",
  main = "Plot of Standardized Residuals vs Logarithm of Distance")
```



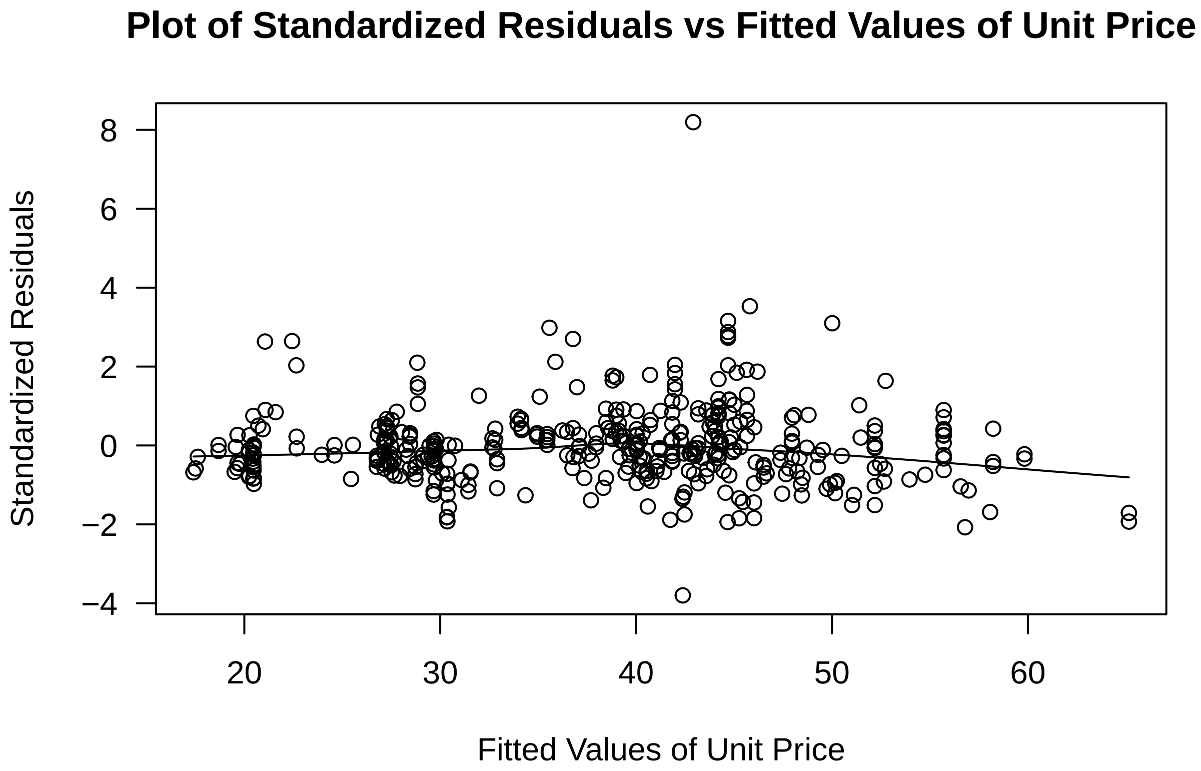
In the scatter plot investigating the predictor 'distance to nearest MRT station,' a lack of noticeable curvature is observed, and data points generally cluster around the mean. However, the plot reveals heteroskedasticity. This is discernible as the residuals exhibit smaller variances and are more tightly concentrated around the mean for lower distance values. Conversely, for median distances, the residuals are more dispersed, indicating higher variances and suggesting non-constant variances in the residuals.

It is apparent that, for shorter distances, the model tends to overestimate real estate prices, causing a slight deviation from the mean trend. Notably, outliers within the distance range of 244 to 403 meters from the houses are particularly conspicuous, signifying specific instances where the model's accuracy diminishes in capturing the variability in real estate prices.

Question 4 - 10 points

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?

```r
fitted_values <- fitted(model)
Diagnosti_Plot3 <-scatter.smooth( standard_res ~ fitted_values,
las = 1,
     xlab = "Fitted Values of Unit Price",
     ylab = "Standardized Residuals",
     main = "Plot of Standardized Residuals vs Fitted Values of Unit Price")
```

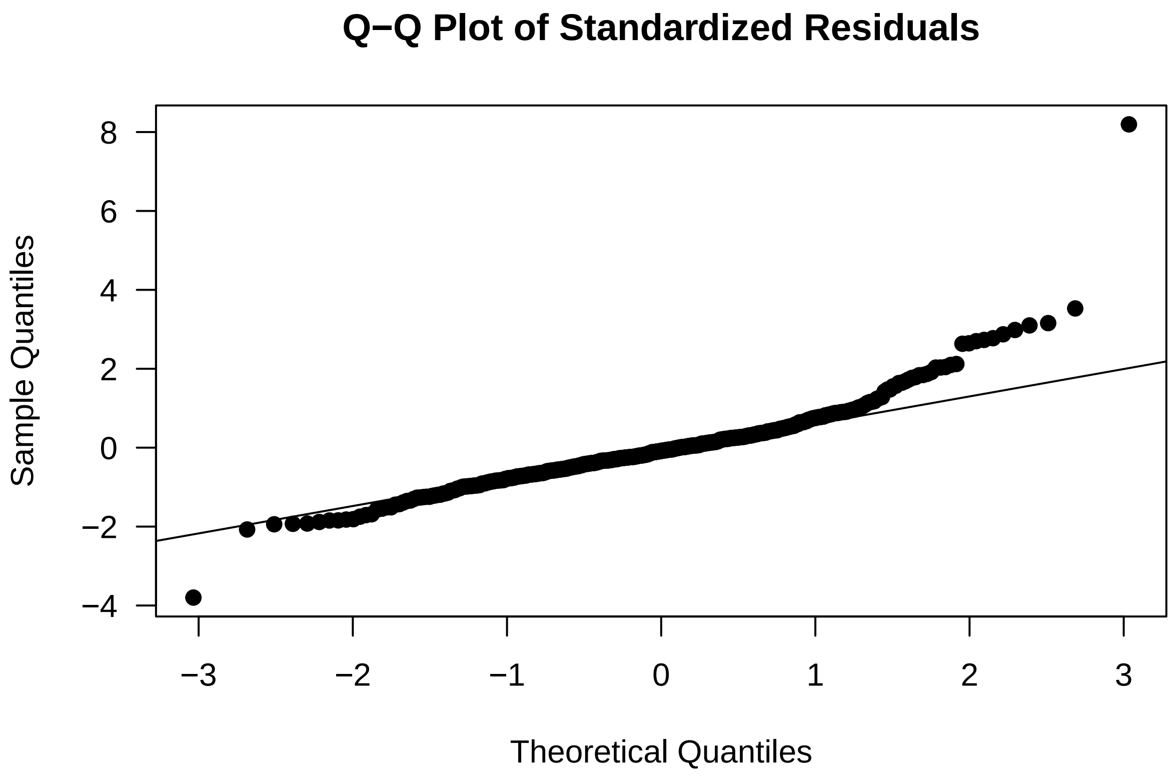


The scatter plot of fitted values shows an absence of apparent non-linear trends or specific patterns, with data points seemingly evenly distributed around the mean. However, a more detailed examination reveals the presence of heteroskedasticity. This becomes particularly evident at the extreme ends, notably on the right side, where variances are smaller compared to those in the central quantile values. Additionally, there are conspicuous outliers scattered significantly away from the mean. This observation suggests that employing a power transformation might be advantageous in normalizing the distribution of these values.

Question 5 - 10 points

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?

```r
#Q-Q plot for standardized residuals
Diagnosti_Plot4 <- qqnorm(standard_res,las = 1, pch = 19, main = "Q-Q Plot of Standardized Residuals")
qqline(standard_res, col = "black")
```



The Q-Q plot of standardized residuals primarily displays a clustering of points along the diagonal line, suggesting an approximately normal distribution. However, there is a noticeable skewness at the right extreme of the normal curve, indicating the presence of heavier tails in the distribution. A minor deviation from the diagonal is also observed at the left extreme. These observations suggest an opportunity to improve the model's fit, particularly in achieving a more normalized distribution of residuals.

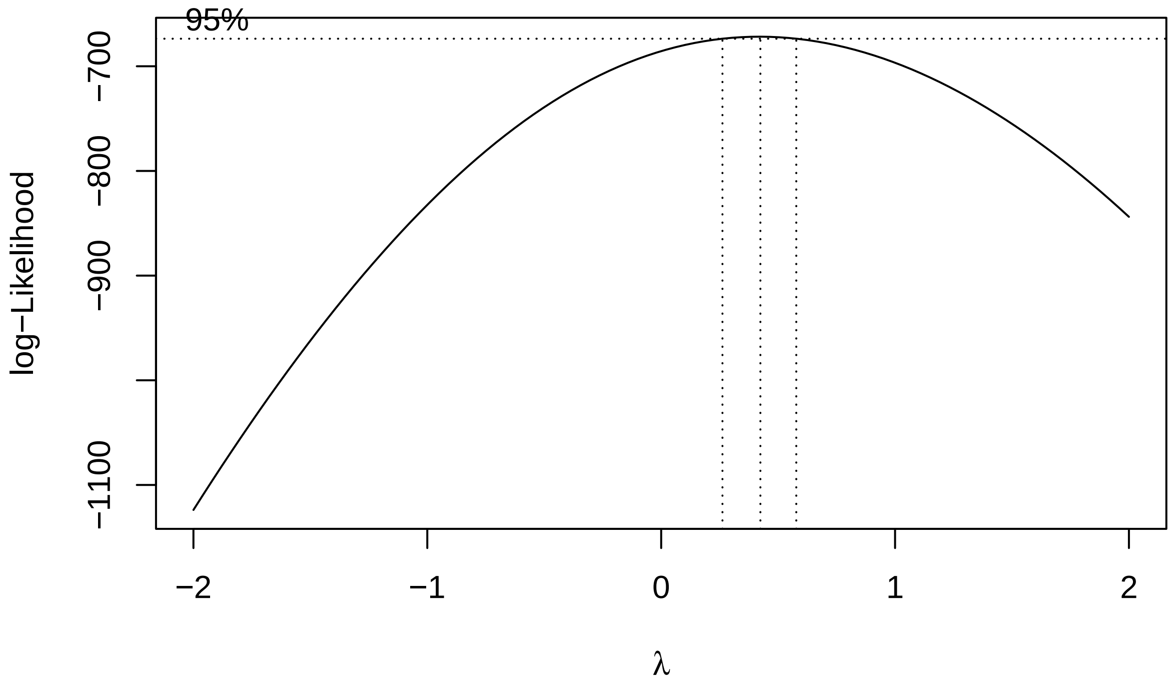
Question 6 - 10 points

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

```r
library(MASS)

power_law <- Vectorize (
  function(t,lambda) {
    if (t < 0) stop("t must be strictly positive")
    if (lambda != 0) {
      return((t^lambda - 1) / lambda)
    }
      return(log(lambda))
  },
"t"
```

```
)
box_cox <- boxcox(unit_price ~ power_law(log_distance, 0.1) + power_law(convenience_stores,0.1),data=da
```



```
lambda <- box_cox$x[which.max(box_cox$y)]
#Best lambda value is:
lambda
```

```
## [1] 0.4242424
```

Question 7 - 5 points

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

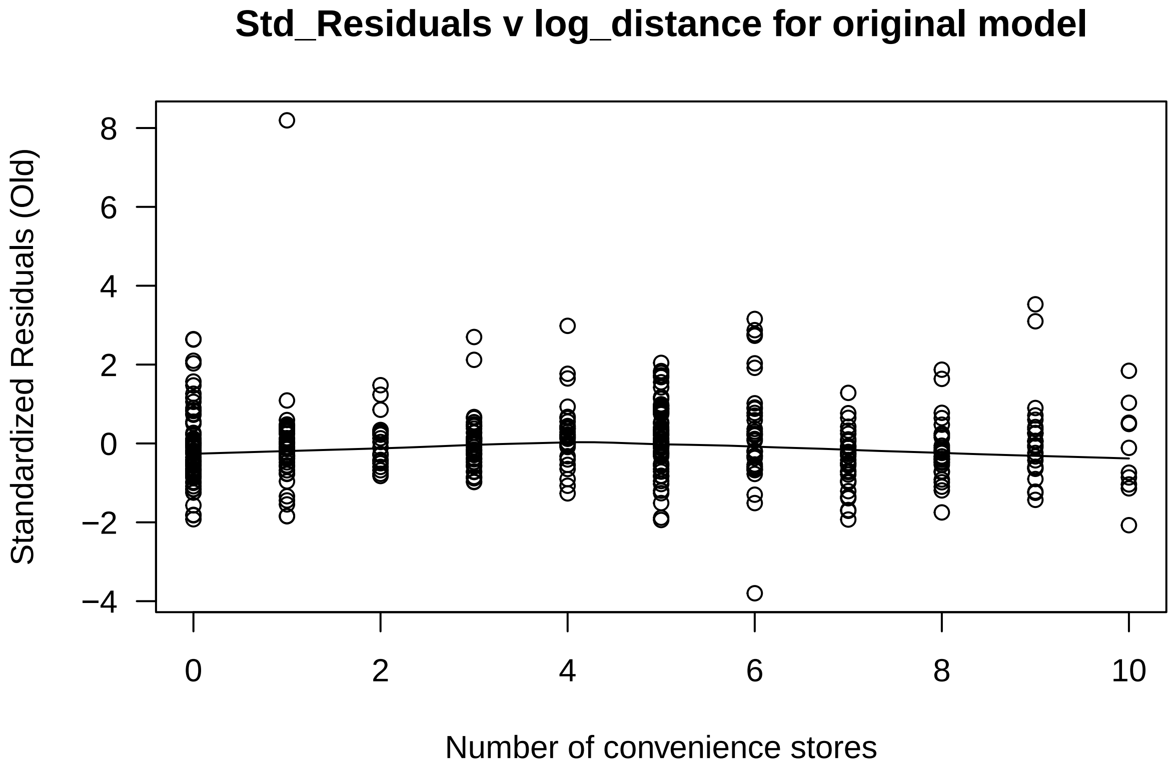
```
transformed_model <- lm(unit_price^lambda ~ log_distance + convenience_stores, data = data)
summary(transformed_model)
```

```
##
## Call:
## lm(formula = unit_price^lambda ~ log_distance + convenience_stores,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.48447 -0.26887 -0.01534  0.23991  2.68967
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.21748    0.21442  33.661  < 2e-16 ***
## log_distance        -0.43165    0.02826 -15.276  < 2e-16 ***
## convenience_stores   0.03501    0.01074   3.259  0.00121 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4681 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
```

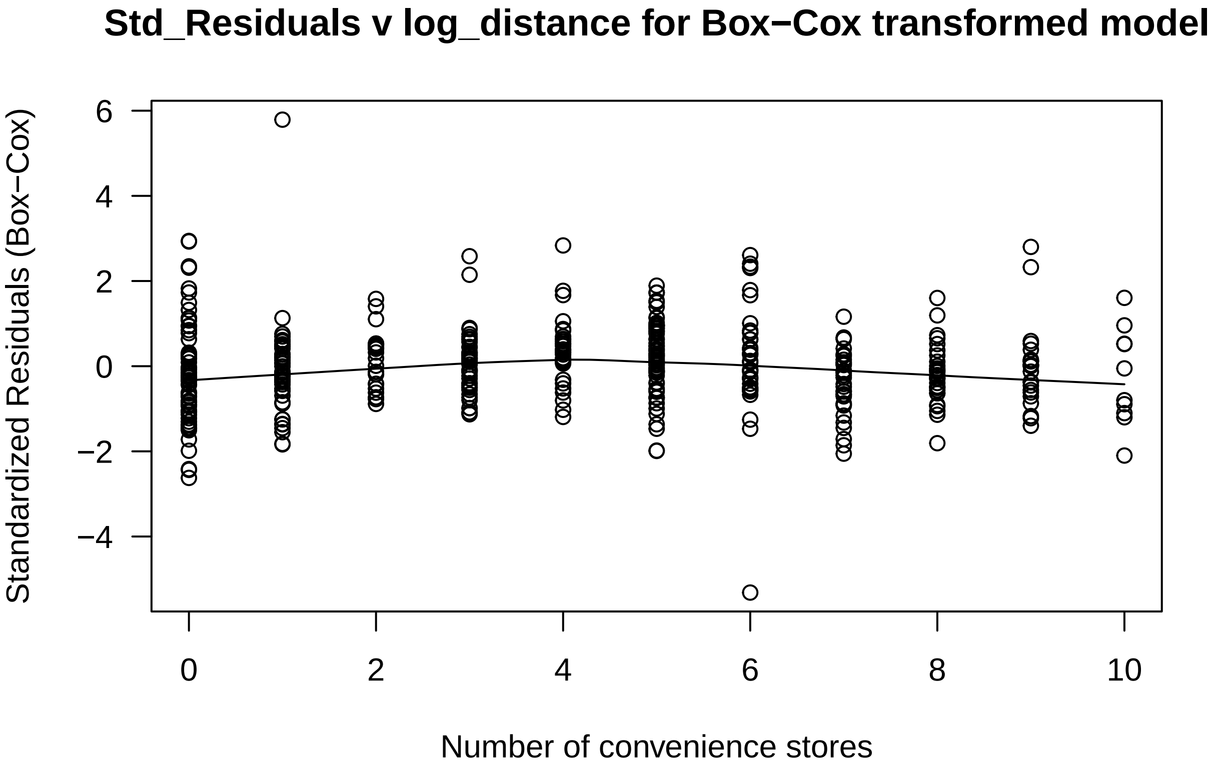
Question 8 - 10 points

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

```
standard_res_0 <- rstandard(model)
Diagnosti_Plot4 <- scatter.smooth(standard_res_0 ~ data$convenience_stores,
las = 1,
xlab = "Number of convenience stores",
ylab = "Standardized Residuals (Old)",
main = "Std_Residuals v log_distance for original model"
)
```

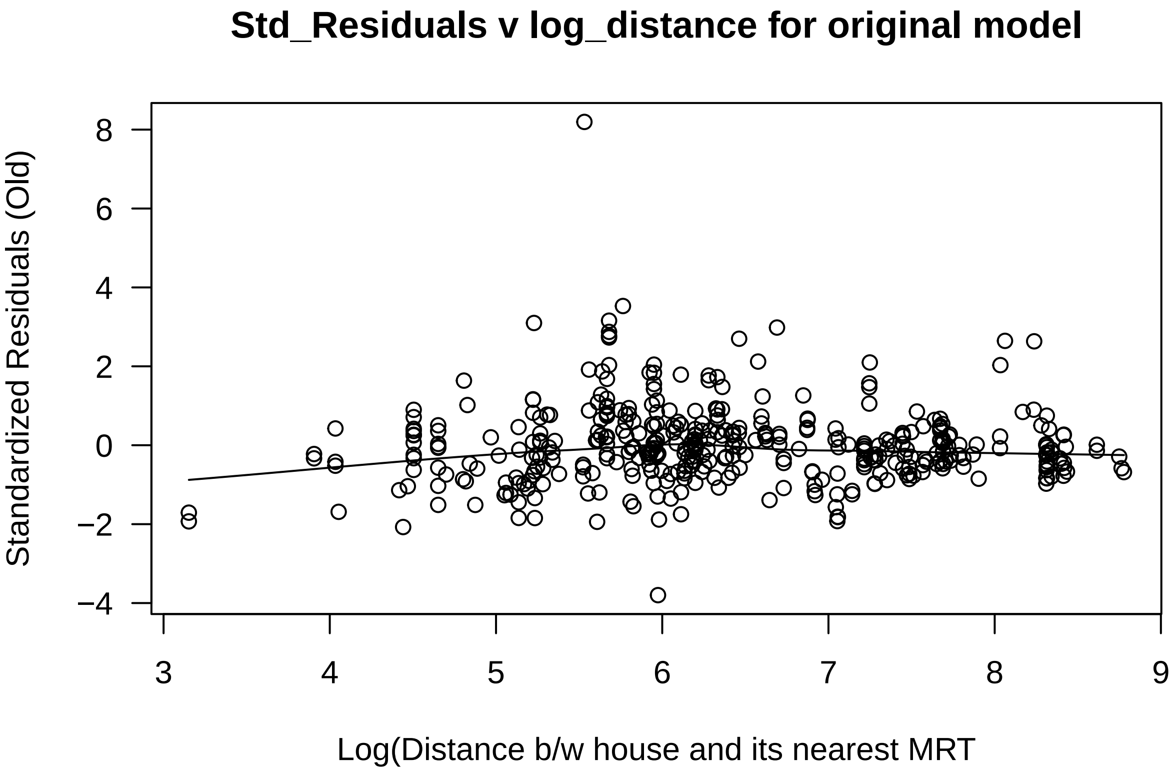


```
standard_res_1 <- rstandard(transformed_model)
Diagnosti_Plot5 <- scatter.smooth(standard_res_1 ~ data$convenience_stores,
las = 1,
xlab = "Number of convenience stores",
ylab = "Standardized Residuals (Box-Cox)",
main = "Std_Residuals v log_distance for Box-Cox transformed model"
)
```

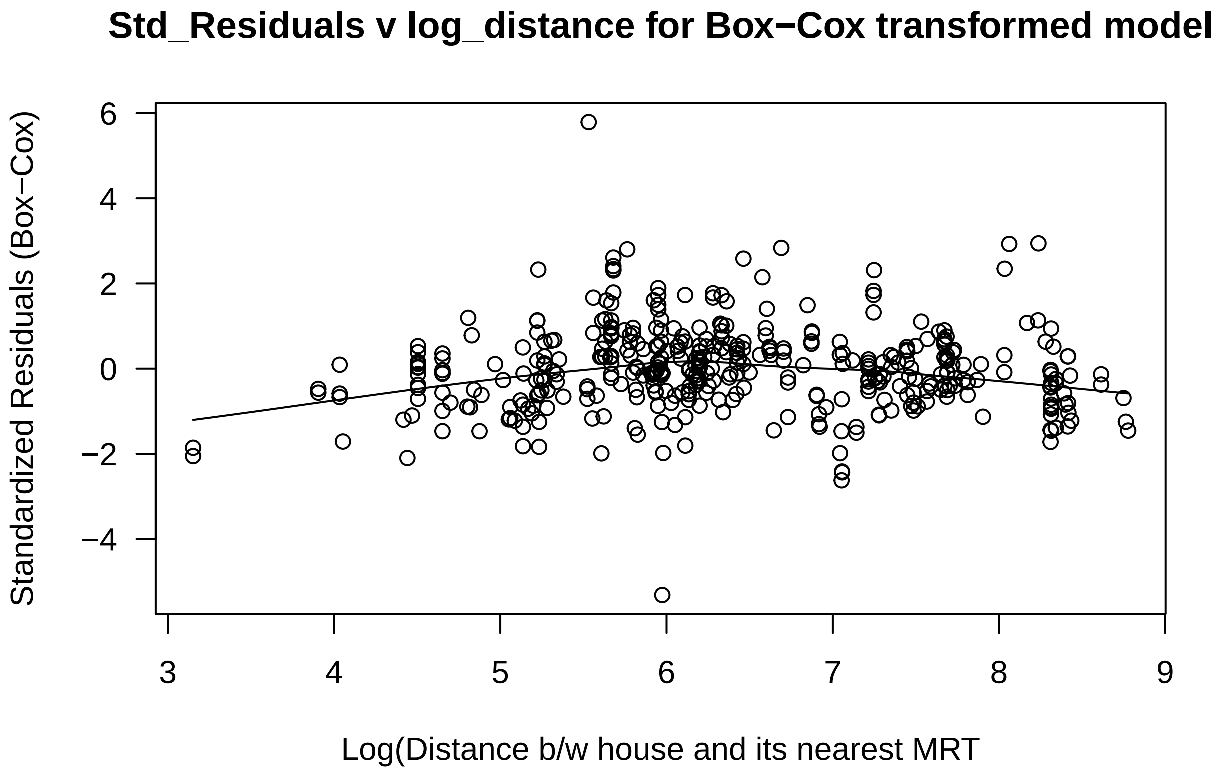


Upon comparing the plots before and after implementing the Box-Cox transformation, a discernible shift is observed in the distribution of outliers around the mean. Following the transformation, the outliers exhibit a more symmetrical distribution around the mean, indicating an enhancement towards a more normal distribution with reduced skewness. Furthermore, the transformed model demonstrates a higher R-squared value, suggesting that it captures more variance in the response variable compared to the original model. However, it is important to note that the curvature appears more visually pronounced in the transformed plot, implying that the Box-Cox transformation might not substantially contribute to addressing linearity concerns.

```
standard_res_0 <- rstandard(model)
Diagnosti_Plot6 <- scatter.smooth(standard_res_0 ~ data$log_distance,
las = 1,
ylab = "Standardized Residuals (Old)",
xlab = "Log(Distance b/w house and its nearest MRT",
main = "Std_Residuals v log_distance for original model"
)
```



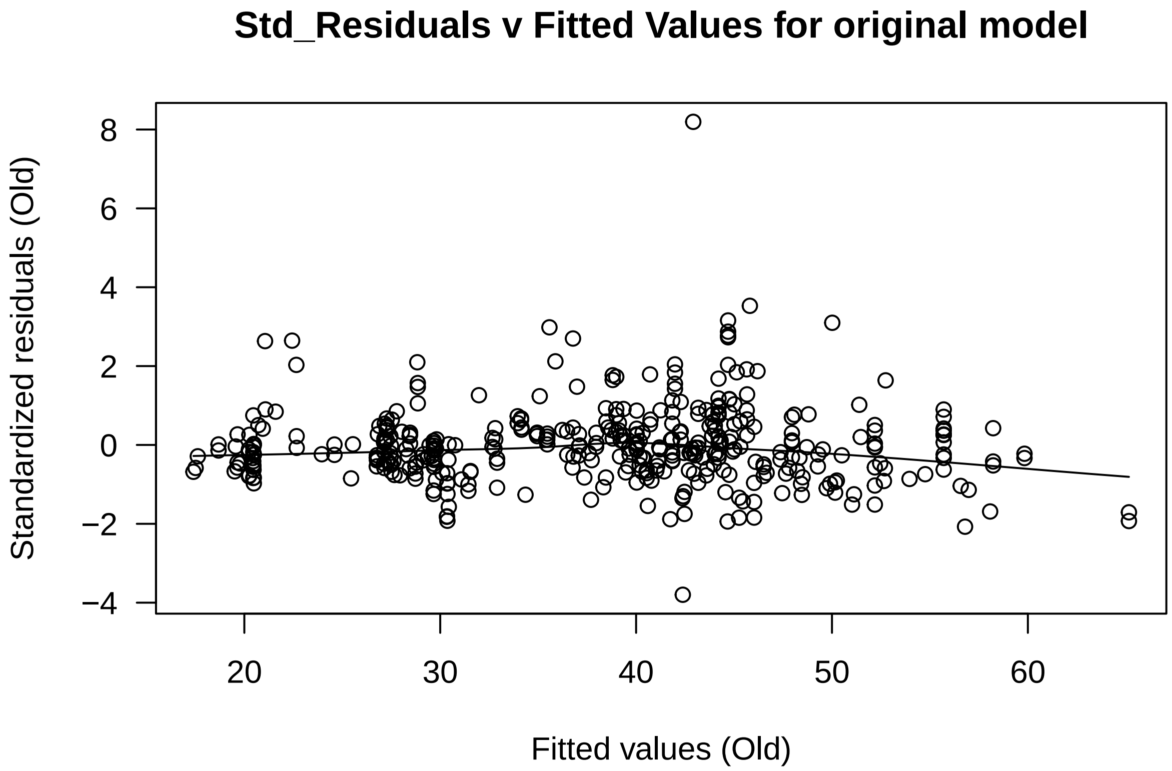
```r
standard_res_1 <- rstandard(transformed_model)
Diagnosti_Plot7 <- scatter.smooth(standard_res_1 ~ data$log_distance,
las = 1,
ylab = "Standardized Residuals (Box-Cox)",
xlab = "Log(Distance b/w house and its nearest MRT",
main = "Std_Residuals v log_distance for Box-Cox transformed model"
)
```



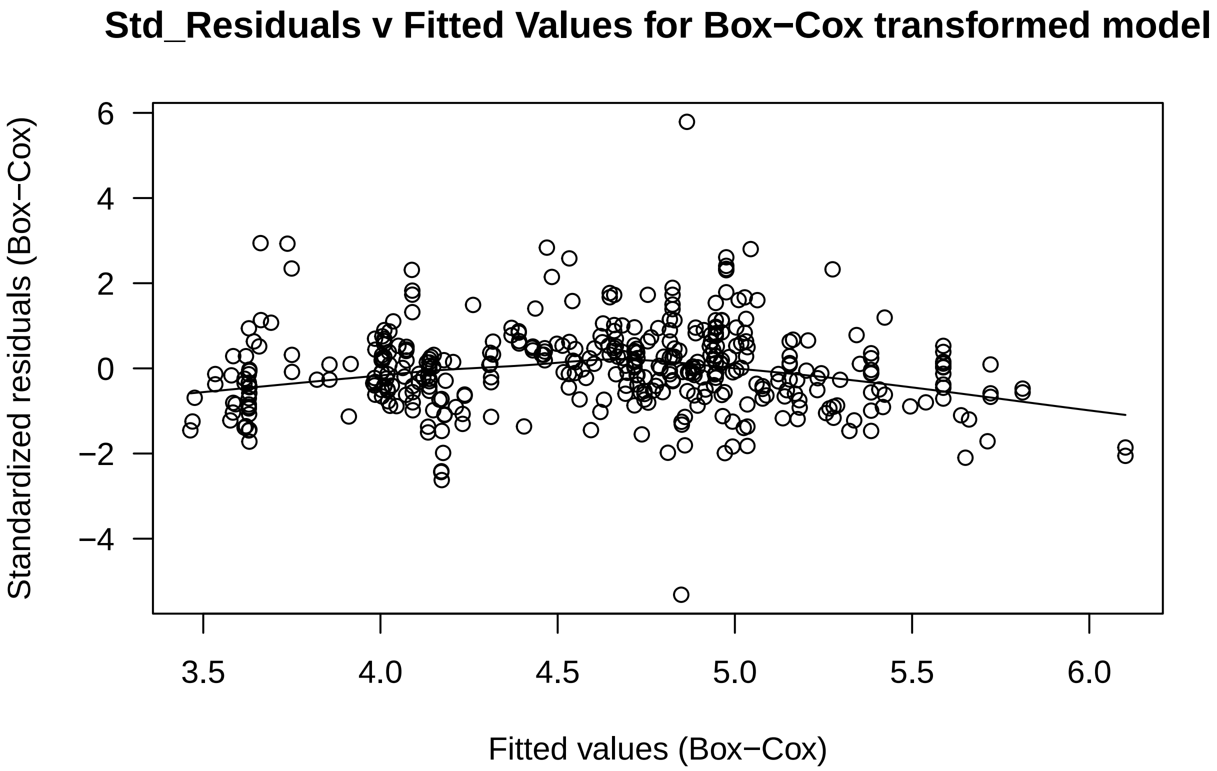
Comment on the diagnostic plot 6 and 7:

we observe a trend that mirrors the one seen in the plot of standard residuals versus convenience stores.

```
standard_res_0 <- rstandard(model)
lm_model_fitted <- fitted(model)
Diagnosti_Plot8 <-scatter.smooth(standard_res_0 ~ lm_model_fitted,
las = 1,
    ylab = "Standardized residuals (Old)",
    xlab = "Fitted values (Old)",
    main = "Std_Residuals v Fitted Values for original model"
)
```

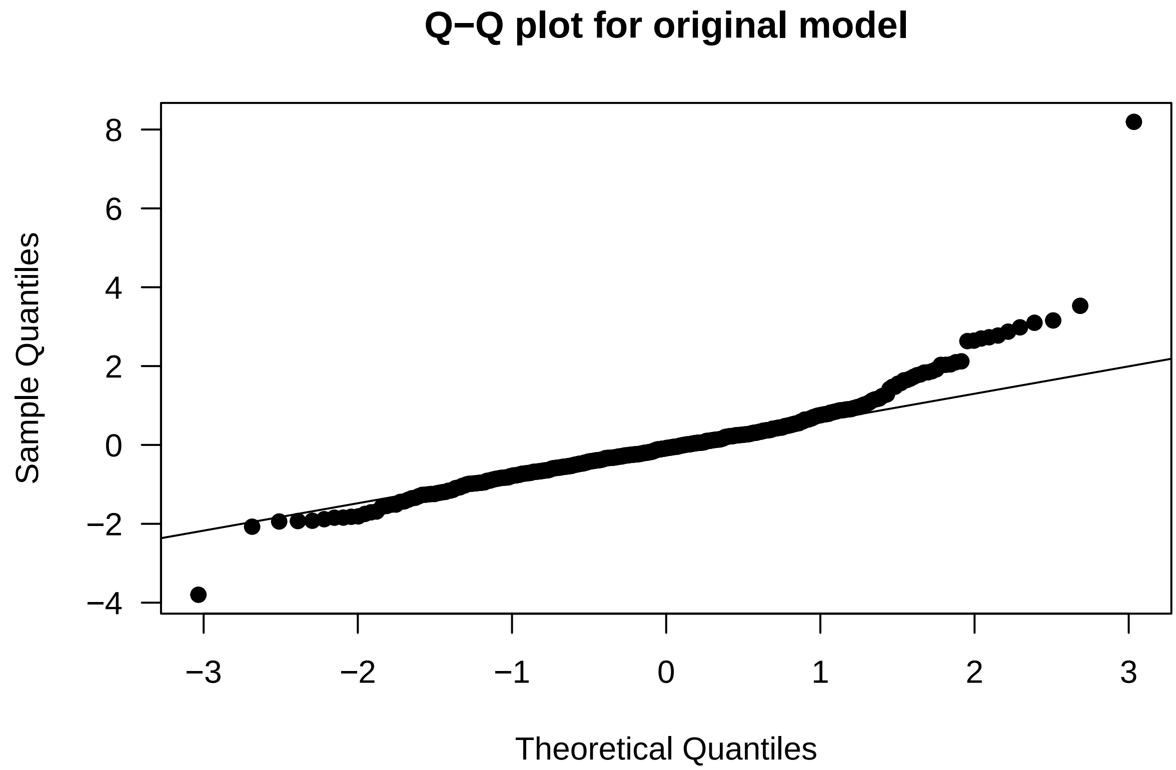


```
standard_res_1 <- rstandard(transformed_model)
lm_model_fitted_new <- fitted(transformed_model)
Diagnosti_Plot9 <-scatter.smooth(standard_res_1 ~ lm_model_fitted_new,
las = 1,
    ylab = "Standardized residuals (Box-Cox)",
    xlab = "Fitted values (Box-Cox)",
    main = "Std_Residuals v Fitted Values for Box-Cox transformed model"
)
```

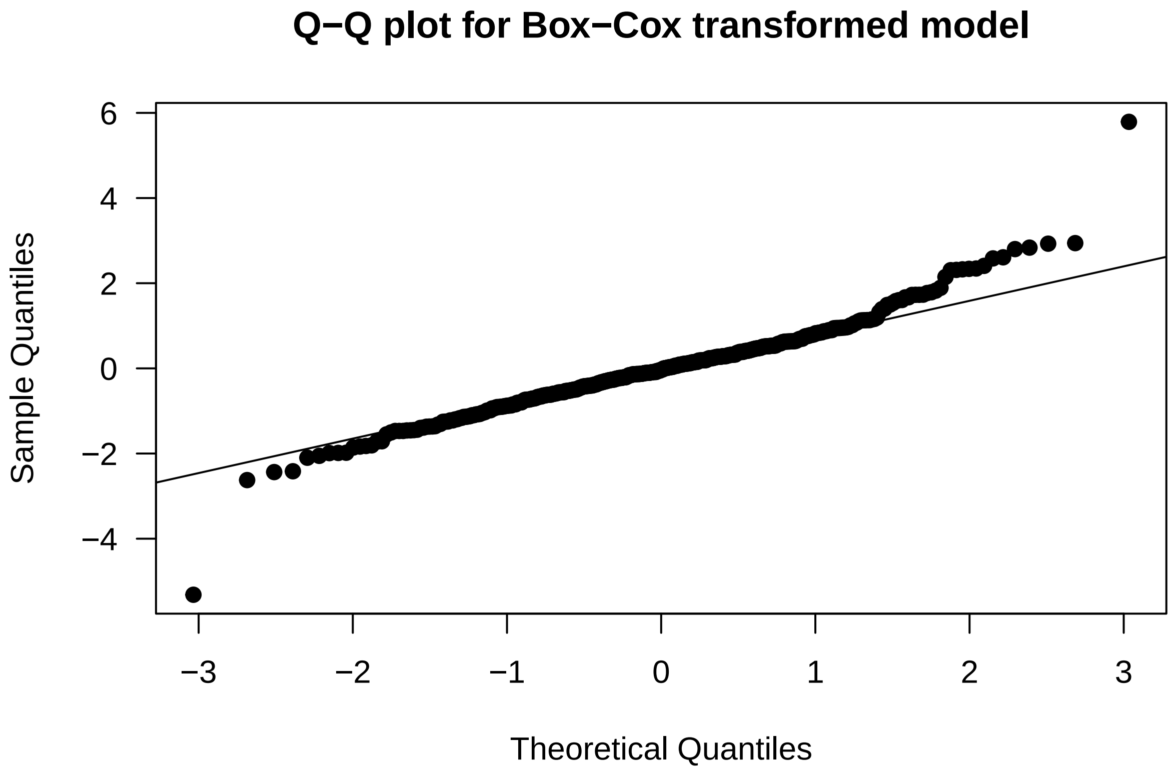


Similar trend as above.

```
standard_res_0 <- rstandard(model)
Diagnosti_Plot10 <- qqnorm(standard_res_0, las = 1, pch = 19,main = "Q-Q plot for original model")
qqline(standard_res_0)
```



```
standard_res_1 <- rstandard(transformed_model)
Diagnosti_Plot11 <- qqnorm(standard_res_1, las = 1, pch = 19,main = "Q-Q plot for Box-Cox transformed mo
qqline(standard_res_1)
```



Regarding the Q-Q Plots, it's clear that the standardized residuals in the transformed model exhibit a closer alignment with the diagonal compared to the original model. The original model's skewness has diminished, as evidenced by the tighter clustering of residuals along the diagonal and a more symmetric distribution of outliers, with both extremes now within 6 standard deviations of the mean. This adjustment contributes to an improved adherence of the model to normality. Furthermore, the elevated R-squared value indicates a more precise fit between unit_price and the response variables in the transformed model, particularly when considering the logarithm of the distance variable.

Question 9 - 10 points

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?

```
df1 <- influence.measures(transformed_model)
summary(df1)
```

```
## Potentially influential observations of
##   lm(formula = unit_price^lambda ~ log_distance + convenience_stores,      data = data) :
##
##     dfb.1_ dfb.lg_d dfb.cnv_ dffit   cov.r   cook.d hat
## 17   0.03  -0.03     0.03     0.14    0.97_*  0.01   0.00
## 20  -0.30   0.31     0.14    -0.35_*  1.00    0.04   0.03_*
## 48   0.05  -0.03    -0.06     0.14    0.96_*  0.01   0.00
## 56  -0.10   0.06     0.16    -0.22    0.97_*  0.02   0.01
## 89   0.07  -0.03    -0.14     0.20    0.98_*  0.01   0.01
## 106  0.03  -0.03     0.03     0.15    0.97_*  0.01   0.00
## 114  0.02 -0.03    -0.15    -0.32_*  0.82_*  0.03   0.00
## 124  0.03 -0.03    -0.03     0.03    1.03_*  0.00   0.03_*
## 127 -0.03  0.05     0.03     0.15    0.95_*  0.01   0.00
## 147  0.13 -0.12    -0.12     0.14    1.03_*  0.01   0.03_*
## 149 -0.09  0.14    -0.05     0.29_*  0.95_*  0.03   0.01
## 165  0.17 -0.15    -0.16     0.18    1.02_*  0.01   0.03_*
## 167  0.03 -0.03     0.04     0.16    0.96_*  0.01   0.00
## 207  0.00  0.00    -0.01    -0.01    1.02_*  0.00   0.02
## 221 -0.03  0.02     0.15     0.23    0.98_*  0.02   0.01
## 223 -0.04  0.04     0.06     0.07    1.02_*  0.00   0.02
## 229 -0.06  0.11    -0.07     0.27_*  0.95_*  0.02   0.01
## 252 -0.10  0.06     0.16    -0.21    0.97_*  0.02   0.01
## 271  0.68 -0.61    -0.65     0.75_*  0.79_*  0.17   0.02
## 274 -0.25  0.23     0.23    -0.26_*  1.00    0.02   0.02
## 276 -0.27  0.28     0.13    -0.31_*  1.01    0.03   0.03_*
## 286  0.03 -0.01    -0.04     0.11    0.98_*  0.00   0.00
## 307  0.07 -0.06    -0.06     0.07    1.03_*  0.00   0.02
## 313 -0.13  0.11     0.25     0.29_*  0.96_*  0.03   0.01
## 319 -0.04  0.03     0.06     0.07    1.02_*  0.00   0.02
## 331 -0.10  0.06     0.18    -0.23    0.96_*  0.02   0.01
## 341  0.03 -0.03    -0.04     0.04    1.03_*  0.00   0.02
## 345 -0.04  0.08    -0.06     0.22    0.98_*  0.02   0.01
## 346 -0.09  0.09     0.09    -0.10    1.03_*  0.00   0.03_*
## 357  0.03 -0.02    -0.02     0.03    1.02_*  0.00   0.02
## 380  0.03 -0.03     0.03     0.14    0.97_*  0.01   0.00
## 387  0.17 -0.16    -0.16     0.18    1.02_*  0.01   0.03_*
## 400 -0.12  0.11     0.11    -0.12    1.02_*  0.01   0.02
## 403 -0.24  0.22     0.22    -0.26_*  1.00    0.02   0.02
```

```r
df1['dfb.1_']
```

```
## $<NA>
## NULL
```

The above summary captures the key influential points in the multilinear regression model. Focusing on DFBETAs, none of them exceed an absolute value of 1. Consequently, this indicates that there's no need to mark any observation as particularly influential based on their DFBETA scores.

Question 10 - 5 points

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. The dataset contains the following columns: Germ: the number of seeds that germinated in a particular experiment Total: the number of seeds planted in a particular experiment Extract: the extract type (Bean or Cucumber) for the experiment Seeds: the type of seed (0A75 or 0A73) for the experiment.

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.

```r
library(GLMsData)
data("germ", package = "GLMsData")
germs <- as.data.frame(germ)
germs$Seeds <- factor(germs$Seeds)
germs$Extract <- factor(germs$Extract)
germs
```

```
##    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
## 7    53    74 Cucumber  OA75
## 8    55    72 Cucumber  OA75
## 9    32    51 Cucumber  OA75
## 10   46    79 Cucumber  OA75
## 11   10    13 Cucumber  OA75
## 12    8    16     Bean  OA73
## 13   10    30     Bean  OA73
## 14    8    28     Bean  OA73
## 15   23    45     Bean  OA73
## 16    0     4     Bean  OA73
## 17    3    12 Cucumber  OA73
## 18   22    41 Cucumber  OA73
## 19   15    30 Cucumber  OA73
## 20   32    51 Cucumber  OA73
## 21    3     7 Cucumber  OA73
```

```
logistic_reg_model <- glm(Germ / Total ~ Extract + Seeds, weights = Total, data = germs, family ="binom
summary(logistic_reg_model)
```

```
##
## Call:
## glm(formula = Germ/Total ~ Extract + Seeds, family = "binomial",
##     data = germs, weights = Total)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.7005     0.1507  -4.648 3.36e-06 ***
## ExtractCucumber   1.0647     0.1442   7.383 1.55e-13 ***
## SeedsOA75         0.2705     0.1547   1.748   0.0804 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 39.686  on 18  degrees of freedom
## AIC: 122.28
##
## Number of Fisher Scoring iterations: 4
```

Question 11 - 15 points

Analyze the model summary and answer the following questions. 1. What is the baseline category of each categorical predictor in the model?

Solution:

The log odds for Cucumber are expressed concerning the Extract category, with Bean serving as the reference category for the Extract predictor. Similarly, for the Seeds Extract predictor, OA73 is set as the baseline category.

2. What are the odds of germination for the baseline combination of Extract and Seeds according to the model?

Solution:

```
beta0 = -0.7005
betas = 0.2705
betae = 1.0647
odds_baseline = exp(beta0)
odds_baseline
```

```
## [1] 0.4963371
```

For the base combination of ‘Bean’ in the Extract category and ‘OA73’ in the Seeds category, the estimated odds of germination are calculated to be 49.63%.

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?

Solution:

```
#cucumber indicates Cucumber,0A73
#Baseline indicates baseline
odds_cucumber_Baseline = exp(betae)
odds_cucumber_Baseline
```

```
## [1] 2.899969
```

For Cucumber extracts and 0A73 type seeds, the germination odds exceed those of the baseline Extract and Seed combination by approximately 289.99%.

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?

Solution:

```
#Bean indicates Bean, 0A75
#Baseline indicates baseline
odds_Bean_Baseline = exp(betas)
odds_Bean_Baseline
```

```
## [1] 1.31062
```

When employing Bean extracts with 0A75 seeds, the odds of germination exceed the baseline germination odds for the standard Extract and Seed combination by approximately 131%.

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?

Solution:

```r
#Bean indicates Bean, 0A75
#Baseline indicates baseline
new_odds_Bean_Baseline = exp(betas + betae)
new_odds_Bean_Baseline
```

```
## [1] 3.800756
```

The odds of germination for Bean extracts when the seed is 0A75 are higher than the odds of germination for the baseline combination of Extract and Seeds by approximately 380%.