Diagnostics

2025-12-05

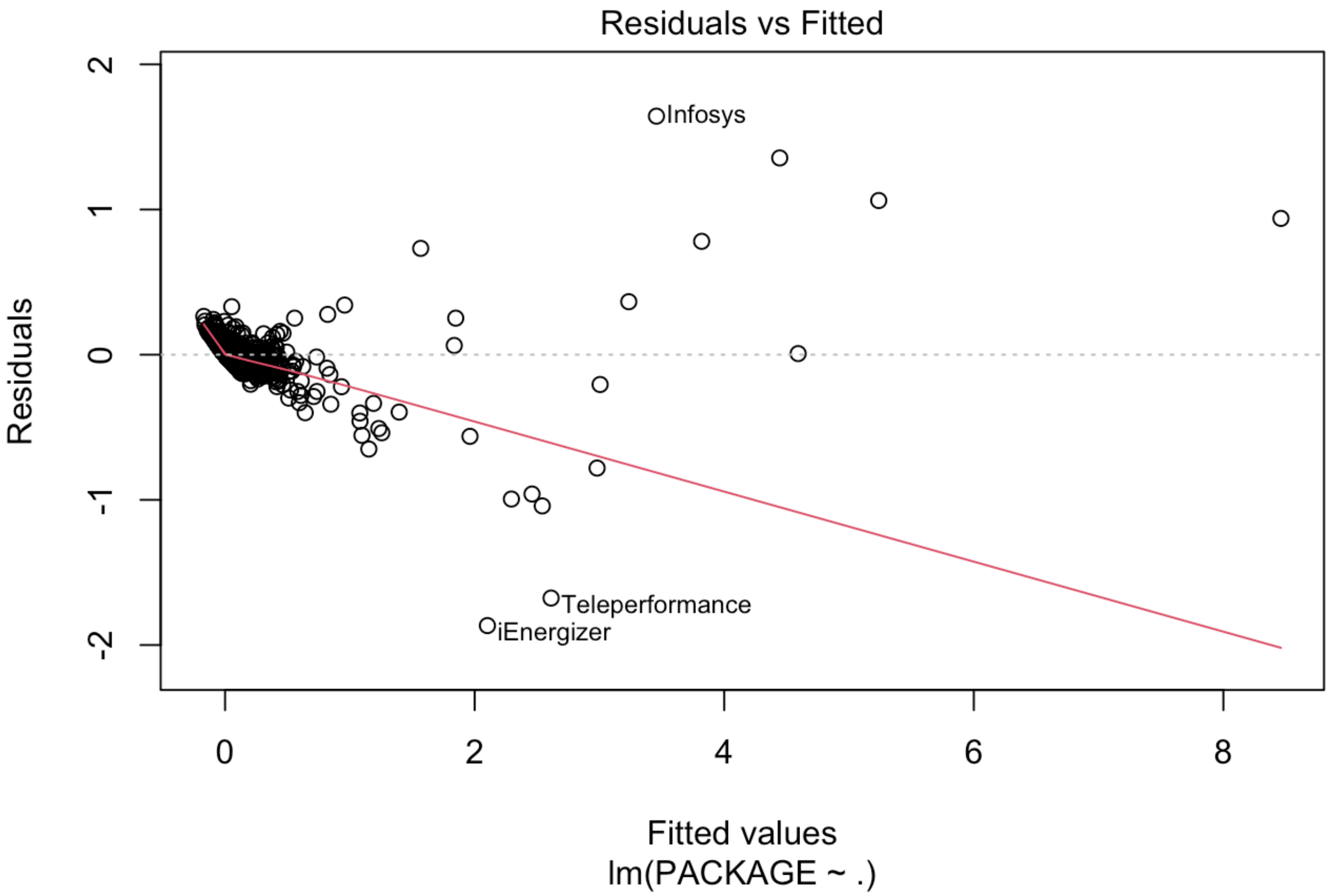
Loading the data:

```
##           YEARS.OLD                 INDUSTRY              INDIA.HQ
## TCS              56 IT Services & Consulting Bangalore / Bengaluru
## Accenture        35 IT Services & Consulting Bangalore / Bengaluru
## Wipro            30 IT Services & Consulting Bangalore / Bengaluru
## Cognizant        79 IT Services & Consulting                 Other
## Capgemini        57 IT Services & Consulting Bangalore / Bengaluru
## HDFC Bank       30                     Other                Mumbai
##           TOTAL_EMPLOYEES BRANCHES RATING REVIEWS PACKAGE
## TCS              11.91839      430    3.4  110000     9.4
## Accenture        11.91839      245    3.7   67900     6.3
## Wipro            11.91839      367    3.7   60900     4.6
## Cognizant        11.91839      224    3.7   57800     5.8
## Capgemini        11.91839      180    3.7   49500     4.6
## HDFC Bank        11.91839     1778    3.8   47900     1.5
```

Fitting Linear model

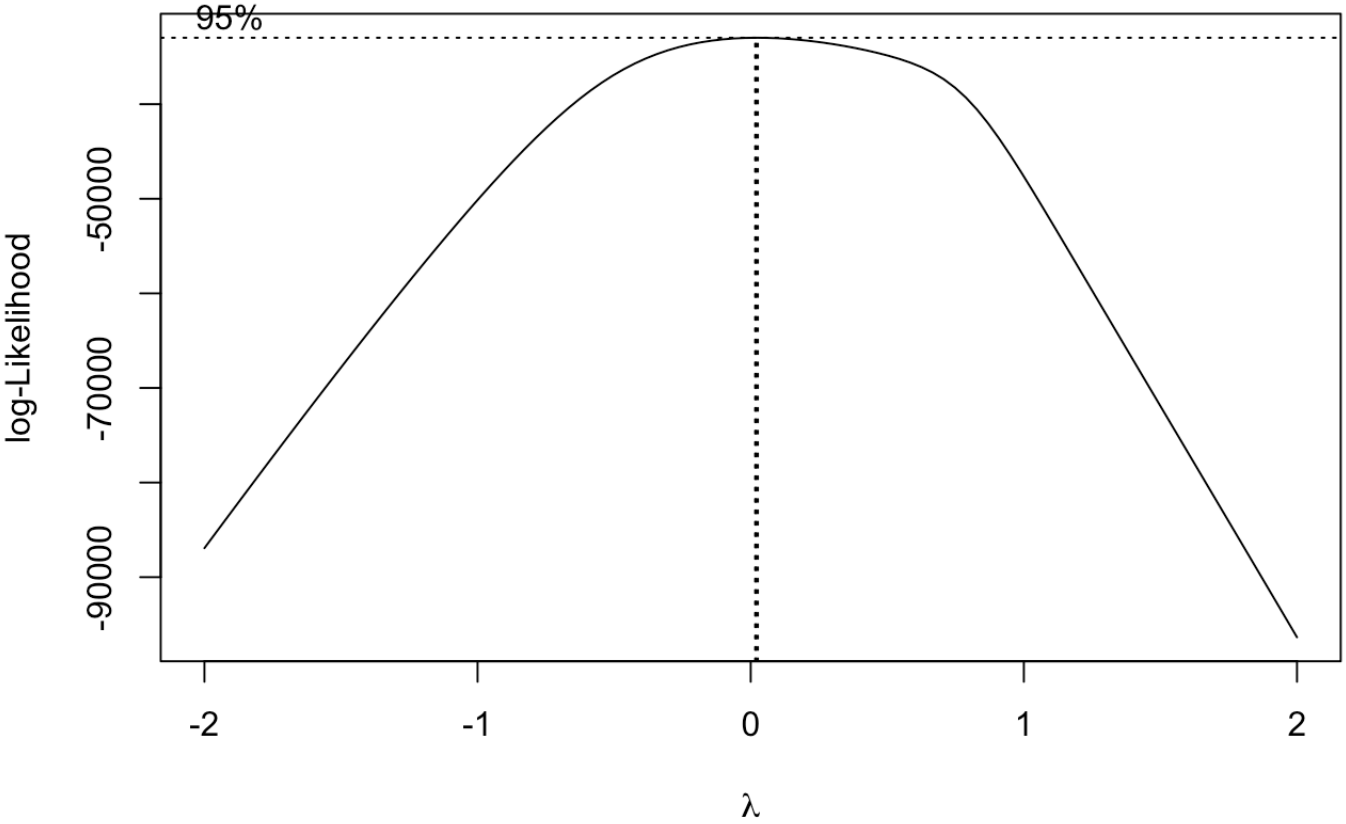
```
##
## Call:
## lm(formula = PACKAGE ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86612 -0.00896 -0.00184  0.00745  1.64315
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         8.288e-03  8.065e-03   1.028  0.30414
## YEARS.OLD                           3.633e-05  2.229e-05   1.630  0.10320
## INDUSTRYEducation & Training        9.984e-03  5.050e-03   1.977  0.04807 *
## INDUSTRYEngineering & Construction  7.992e-03  4.299e-03   1.859  0.06309 .
## INDUSTRYFinancial Services          2.672e-02  4.930e-03   5.420 6.15e-08 ***
## INDUSTRYHealthcare                  1.354e-03  4.877e-03   0.278  0.78134
## INDUSTRYIndustrial Machinery        7.762e-03  4.466e-03   1.738  0.08228 .
## INDUSTRYInternet                    2.248e-03  4.970e-03   0.452  0.65107
## INDUSTRYIT Services & Consulting    1.013e-02  3.643e-03   2.781  0.00543 **
## INDUSTRYOther                       1.085e-02  3.291e-03   3.297  0.00098 ***
## INDUSTRYPharma                      1.477e-03  4.601e-03   0.321  0.74830
## INDUSTRYReal Estate                -8.350e-04  5.427e-03  -0.154  0.87773
## INDUSTRYSoftware Product            5.931e-03  4.831e-03   1.228  0.21960
## INDIA.HQChennai                    -2.223e-03  2.924e-03  -0.760  0.44725
## INDIA.HQMumbai                     -9.491e-04  2.442e-03  -0.389  0.69754
## INDIA.HQNew Delhi                   3.823e-03  2.868e-03   1.333  0.18254
## INDIA.HQOther                      -5.139e-03  2.006e-03  -2.561  0.01044 *
## INDIA.HQPune                       -6.356e-03  2.961e-03  -2.147  0.03183 *
## TOTAL_EMPLOYEES                    -1.582e-03  4.853e-04  -3.260  0.00112 **
## BRANCHES                           -7.188e-04  1.030e-05 -69.792  < 2e-16 ***
## RATING                              4.607e-04  1.733e-03   0.266  0.79041
## REVIEWS                             7.970e-05  3.042e-07 262.002  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05885 on 7779 degrees of freedom
##   (1648 observations deleted due to missingness)
## Multiple R-squared:  0.9099, Adjusted R-squared:  0.9096
## F-statistic:  3740 on 21 and 7779 DF,  p-value: < 2.2e-16
```

Residuals vs Fitted Values



The residuals vs fitted plot indicates clear heteroscedasticity, as the spread of residuals increases for larger fitted PACKAGE values. The pattern is not centered tightly around zero, and a downward trend is visible, suggesting model misspecification or missing nonlinear terms. A few companies (e.g., Infosys, Teleperformance, iEnergizer) display extreme deviations, indicating potential outliers and influential observations

Box-Cox

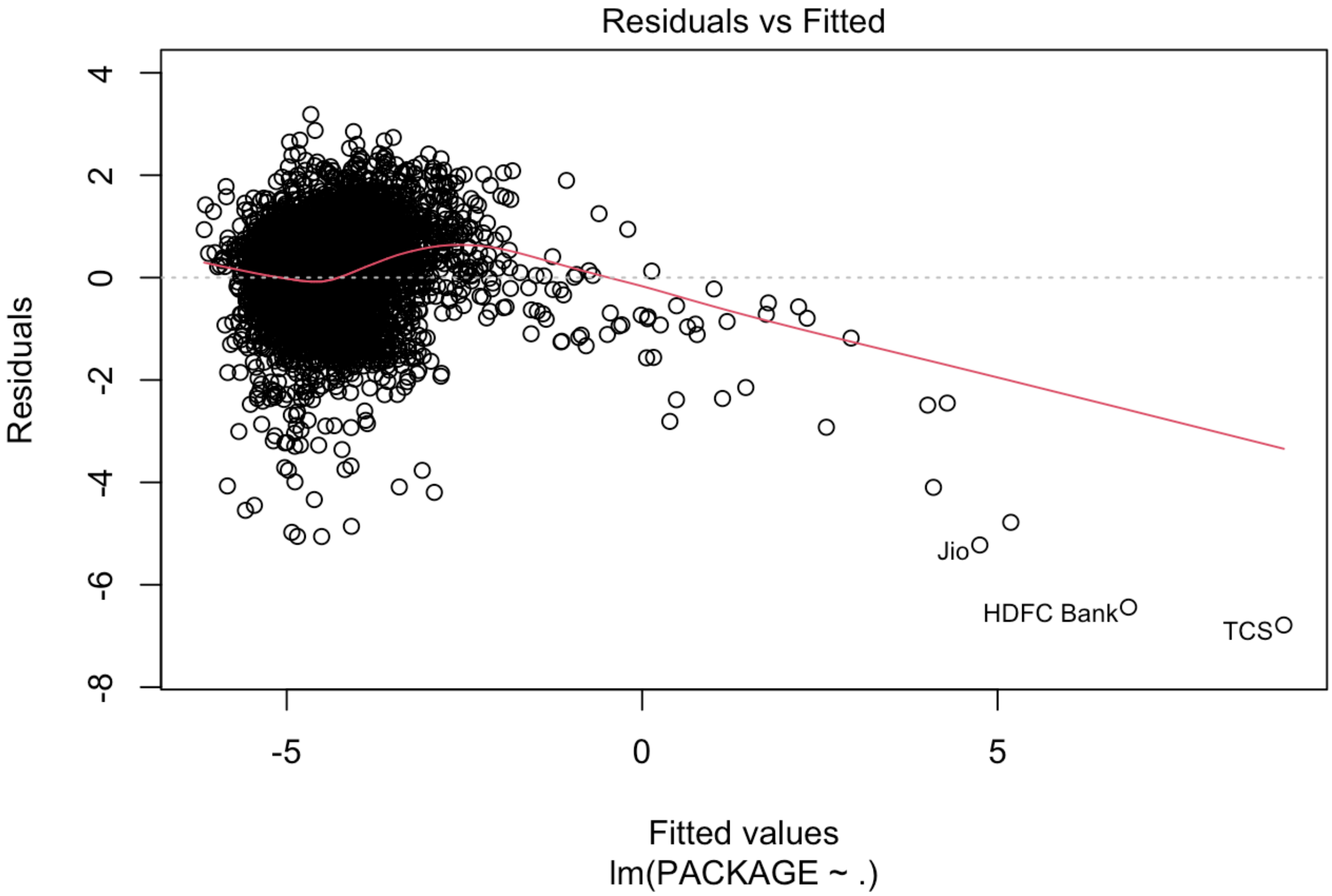

95%
log-Likelihood
-50000
-70000
-90000
-2
-1
0
1
2
λ

```
## $x
##   [1] -2.00000000 -1.95959596 -1.91919192 -1.87878788 -1.83838384 -1.79797980
##   [7] -1.75757576 -1.71717172 -1.67676768 -1.63636364 -1.59595960 -1.55555556
##  [13] -1.51515152 -1.47474747 -1.43434343 -1.39393939 -1.35353535 -1.31313131
##  [19] -1.27272727 -1.23232323 -1.19191919 -1.15151515 -1.11111111 -1.07070707
##  [25] -1.03030303 -0.98989899 -0.94949495 -0.90909091 -0.86868687 -0.82828283
##  [31] -0.78787879 -0.74747475 -0.70707071 -0.66666667 -0.62626263 -0.58585859
##  [37] -0.54545455 -0.50505051 -0.46464646 -0.42424242 -0.38383838 -0.34343434
##  [43] -0.30303030 -0.26262626 -0.22222222 -0.18181818 -0.14141414 -0.10101010
##  [49] -0.06060606 -0.02020202  0.02020202  0.06060606  0.10101010  0.14141414
##  [55]  0.18181818  0.22222222  0.26262626  0.30303030  0.34343434  0.38383838
##  [61]  0.42424242  0.46464646  0.50505051  0.54545455  0.58585859  0.62626263
##  [67]  0.66666667  0.70707071  0.74747475  0.78787879  0.82828283  0.86868687
##  [73]  0.90909091  0.94949495  0.98989899  1.03030303  1.07070707  1.11111111
##  [79]  1.15151515  1.19191919  1.23232323  1.27272727  1.31313131  1.35353535
##  [85]  1.39393939  1.43434343  1.47474747  1.51515152  1.55555556  1.59595960
##  [91]  1.63636364  1.67676768  1.71717172  1.75757576  1.79797980  1.83838384
##  [97]  1.87878788  1.91919192  1.95959596  2.00000000
## 
## $y
##   [1] -86932.21 -85359.61 -83791.28 -82227.46 -80668.41 -79114.34 -77565.51
##   [8] -76022.21 -74484.76 -72953.50 -71428.81 -69911.09 -68400.83 -66898.52
##  [15] -65404.75 -63920.15 -62445.43 -60981.43 -59529.03 -58089.30 -56663.44
##  [22] -55252.78 -53858.96 -52483.70 -51129.15 -49797.71 -48492.05 -47215.45
##  [29] -45971.34 -44763.81 -43597.27 -42476.37 -41406.20 -40391.80 -39437.90
##  [36] -38549.04 -37728.86 -36979.64 -36303.21 -35699.15 -35166.24 -34702.21
##  [43] -34303.44 -33966.11 -33685.94 -33458.54 -33279.75 -33145.74 -33052.84
##  [50] -32998.01 -32978.43 -32991.63 -33035.56 -33108.25 -33208.11 -33333.63
##  [57] -33483.60 -33657.04 -33853.28 -34072.61 -34315.64 -34585.47 -34887.06
##  [64] -35228.24 -35624.59 -36093.85 -36664.11 -37370.44 -38243.46 -39302.99
##  [71] -40564.42 -42014.13 -43621.40 -45354.48 -47179.95 -49065.53 -50990.51
##  [78] -52940.16 -54903.07 -56873.71 -58847.52 -60822.27 -62796.81 -64770.40
##  [85] -66742.99 -68714.60 -70685.45 -72655.83 -74626.03 -76596.35 -78567.07
##  [92] -80538.44 -82510.70 -84484.04 -86458.64 -88434.64 -90412.15 -92391.26
##  [99] -94372.06 -96354.65
```

```
## [1] 0.02020202
```

Since λ is close to zero, we apply log transformations.

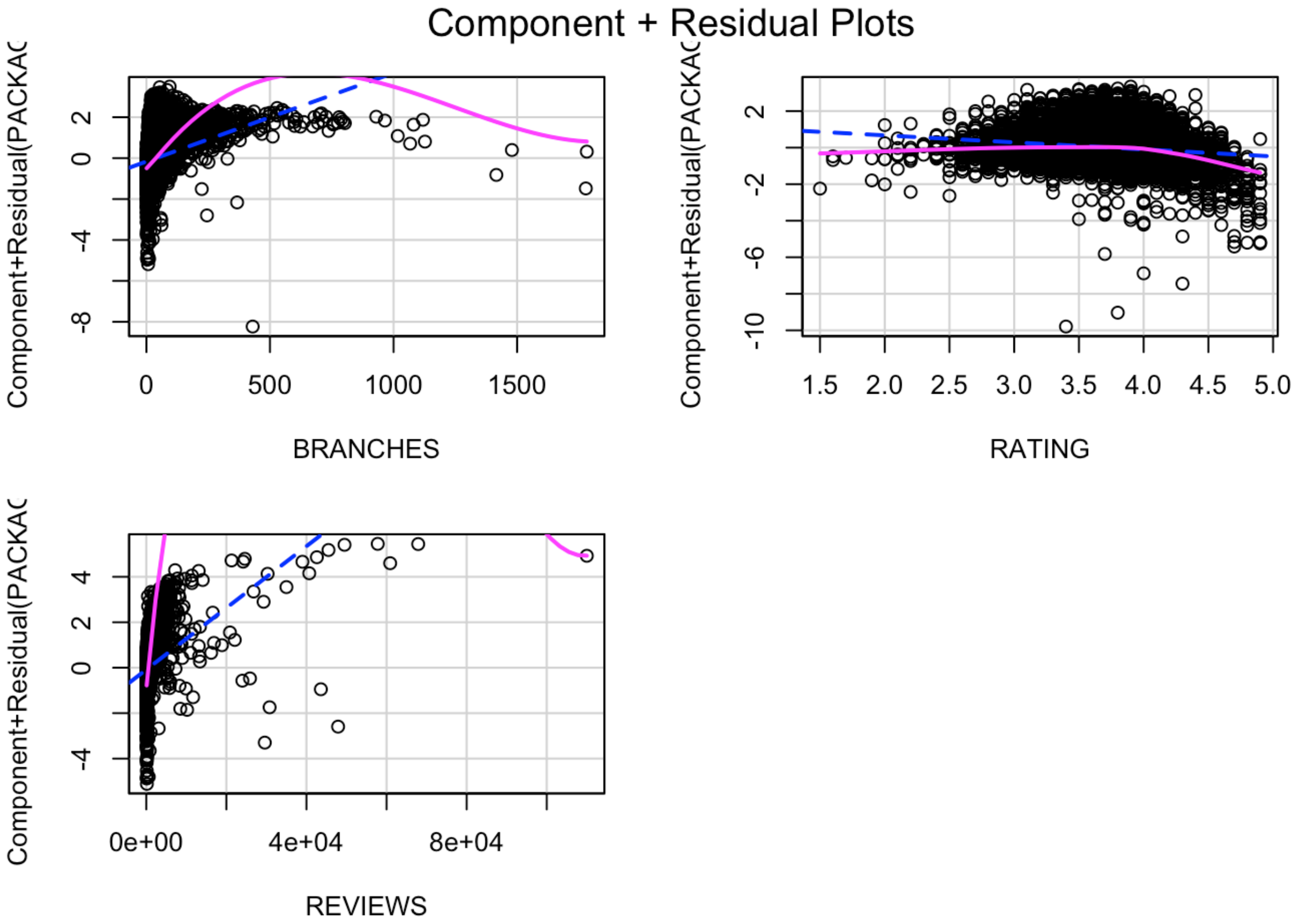
```
##
## Call:
## lm(formula = PACKAGE ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7859 -0.4371 -0.0091  0.4519  3.1871
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -4.650e+00  1.066e-01 -43.629  < 2e-16 ***
## YEARS.OLD                           5.451e-06  2.946e-04   0.019 0.985237
## INDUSTRYEducation & Training       -6.416e-01  6.673e-02  -9.615  < 2e-16 ***
## INDUSTRYEngineering & Construction -2.929e-01  5.681e-02  -5.155 2.60e-07 ***
## INDUSTRYFinancial Services         -6.095e-02  6.515e-02  -0.936 0.349498
## INDUSTRYHealthcare                 -3.471e-01  6.444e-02  -5.387 7.38e-08 ***
## INDUSTRYIndustrial Machinery       -2.035e-01  5.902e-02  -3.448 0.000568 ***
## INDUSTRYInternet                   -1.956e-01  6.568e-02  -2.978 0.002912 **
## INDUSTRYIT Services & Consulting   -3.751e-02  4.814e-02  -0.779 0.435930
## INDUSTRYOther                      -2.273e-01  4.348e-02  -5.227 1.76e-07 ***
## INDUSTRYPharma                     -1.483e-01  6.080e-02  -2.439 0.014764 *
## INDUSTRYReal Estate                -2.363e-01  7.171e-02  -3.295 0.000989 ***
## INDUSTRYSoftware Product            6.204e-02  6.383e-02   0.972 0.331117
## INDIA.HQChennai                    -1.313e-01  3.864e-02  -3.397 0.000685 ***
## INDIA.HQMumbai                     -2.782e-01  3.227e-02  -8.621  < 2e-16 ***
## INDIA.HQNew Delhi                  -4.494e-01  3.789e-02 -11.860  < 2e-16 ***
## INDIA.HQOther                      -3.184e-01  2.651e-02 -12.010  < 2e-16 ***
## INDIA.HQPune                       -7.821e-02  3.912e-02  -1.999 0.045627 *
## TOTAL_EMPLOYEES                     2.527e-01  6.413e-03  39.405  < 2e-16 ***
## BRANCHES                            3.241e-03  1.361e-04  23.811  < 2e-16 ***
## RATING                             -3.578e-01  2.290e-02 -15.622  < 2e-16 ***
## REVIEWS                             9.568e-05  4.020e-06  23.803  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7777 on 7779 degrees of freedom
##   (1648 observations deleted due to missingness)
## Multiple R-squared:  0.4814, Adjusted R-squared:   0.48
## F-statistic: 343.8 on 21 and 7779 DF,  p-value: < 2.2e-16
```



The Box–Cox (log) transformation of the response reduced heteroscedasticity, but it was not completely eliminated.

To further diagnose the issue, we examine the linearity assumption between the response and predictors. Partial residual plots can be used for Rating, Branches, and Reviews to assess nonlinear relationships.

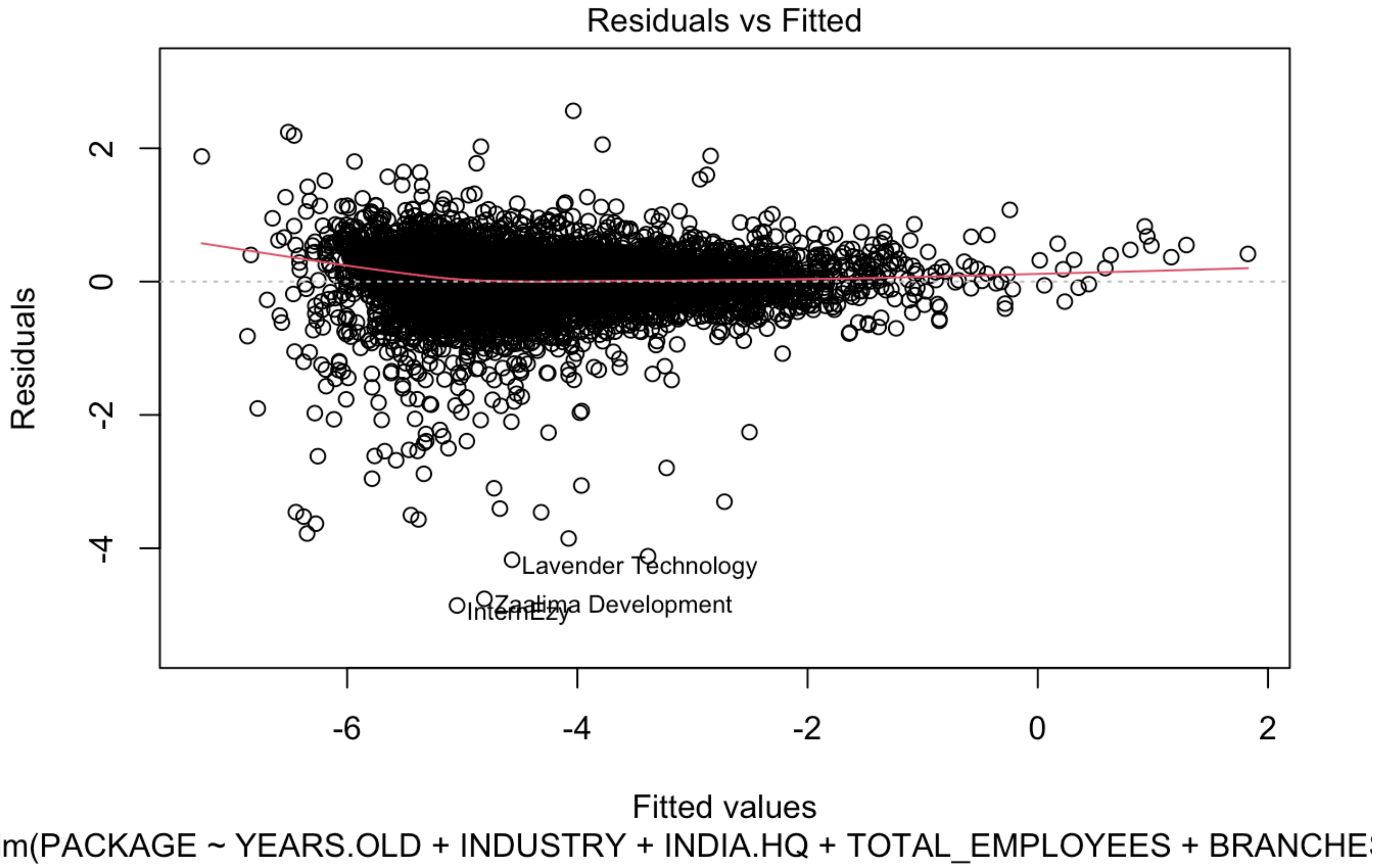
```
## Loading required package: carData
```



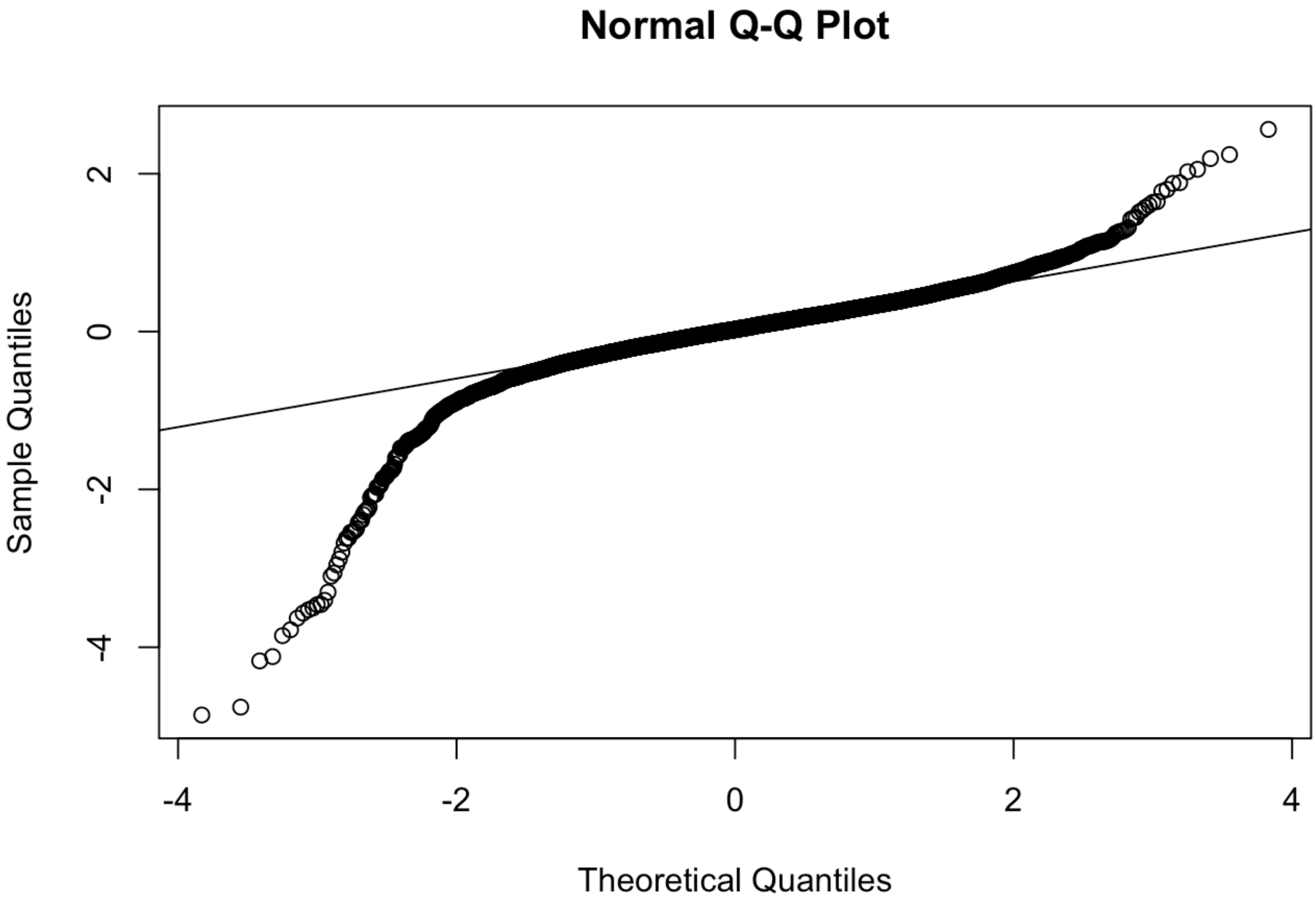
Based on these plots, additional predictor transformations were considered where nonlinearity was evident, i.e. in case of Branches and Reviews.

While Rating is almost linear, where quadratic transformation might provide a better fit.

```
## 
## Call:
## lm(formula = PACKAGE ~ YEARS.OLD + INDUSTRY + INDIA.HQ + TOTAL_EMPLOYEES + 
##     BRANCHES + RATING + I(RATING^2) + REVIEWS, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.8582 -0.1859  0.0266  0.2298  2.5614 
## 
## Coefficients:
##                                      Estimate Std. Error  t value Pr(>|t|)    
## (Intercept)                        -9.8055192  0.0427801 -229.207  < 2e-16 ***
## YEARS.OLD                          -0.0000653  0.0001644   -0.397 0.691273    
## INDUSTRYEducation & Training       -0.3008225  0.0373568   -8.053 9.30e-16 ***
## INDUSTRYEngineering & Construction -0.0810371  0.0318626   -2.543 0.010999 *  
## INDUSTRYFinancial Services          0.0708428  0.0363805    1.947 0.051538 .  
## INDUSTRYHealthcare                 -0.1321684  0.0360051   -3.671 0.000243 ***
## INDUSTRYIndustrial Machinery        0.0315252  0.0330183    0.955 0.339719    
## INDUSTRYInternet                   -0.0691997  0.0367141   -1.885 0.059491 .  
## INDUSTRYIT Services & Consulting    0.1233462  0.0269854    4.571 4.93e-06 ***
## INDUSTRYOther                      -0.0541113  0.0243447   -2.223 0.026264 *  
## INDUSTRYPharma                     -0.0581191  0.0339543   -1.712 0.086994 .  
## INDUSTRYReal Estate                 0.0235197  0.0401244    0.586 0.557778    
## INDUSTRYSoftware Product            0.1883554  0.0358515    5.254 1.53e-07 ***
## INDIA.HQChennai                    -0.0970025  0.0215788   -4.495 7.05e-06 ***
## INDIA.HQMumbai                     -0.1575785  0.0181200   -8.696  < 2e-16 ***
## INDIA.HQNew Delhi                  -0.2458446  0.0214506  -11.461  < 2e-16 ***
## INDIA.HQOther                      -0.2010602  0.0148260  -13.561  < 2e-16 ***
## INDIA.HQPune                       -0.0410601  0.0218455   -1.880 0.060204 .  
## TOTAL_EMPLOYEES                    -0.0022120  0.0041009   -0.539 0.589636    
## BRANCHES                           -0.0549002  0.0065428   -8.391  < 2e-16 ***
## RATING                             -0.5034713  0.0132970  -37.863  < 2e-16 ***
## I(RATING^2)                        -0.6278460  0.0189950  -33.053  < 2e-16 ***
## REVIEWS                             1.0140523  0.0082078  123.548  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4341 on 7778 degrees of freedom
##   (1648 observations deleted due to missingness)
## Multiple R-squared:  0.8384, Adjusted R-squared:  0.838 
## F-statistic:  1835 on 22 and 7778 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
Residuals
2
0
-2
-4
Lavender Technology
Zaalima Development
InternEzy
-6
-4
-2
0
2
Fitted values
lm(PACKAGE ~ YEARS.OLD + INDUSTRY + INDIA.HQ + TOTAL_EMPLOYEES + BRANCHE

Q-Q Plot:



The Q-Q plot shows departure from normality, with heavy left tail and mild right tail. For large samples, it might not be so critical (CLT applies). But this can be improved upon handling outliers.

Several observations (e.g., TCS, Teleperformance, iEnergizer, HDFC Bank) appear as outlying or high-leverage points in multiple plots, reinforcing influence concerns.

Leverage Points:

```
## Number of leverage points (>2p/n):  1351
```

```
## Companies with highest leverage:  Accenture Capgemini Cognizant HCLTech HDFC Bank
ICICI Bank Infosys TCS Tech Mahindra Wipro
```

A large number of observations exceed the leverage threshold 2p/n, indicating many high-leverage companies in the dataset. Most high-leverage cases belong to very large firms. Their combination of extreme workforce/branch presence makes them disproportionately influential in estimating regression coefficients.

Jackknife Residuals:

```
## Critical value:  4.407288
```

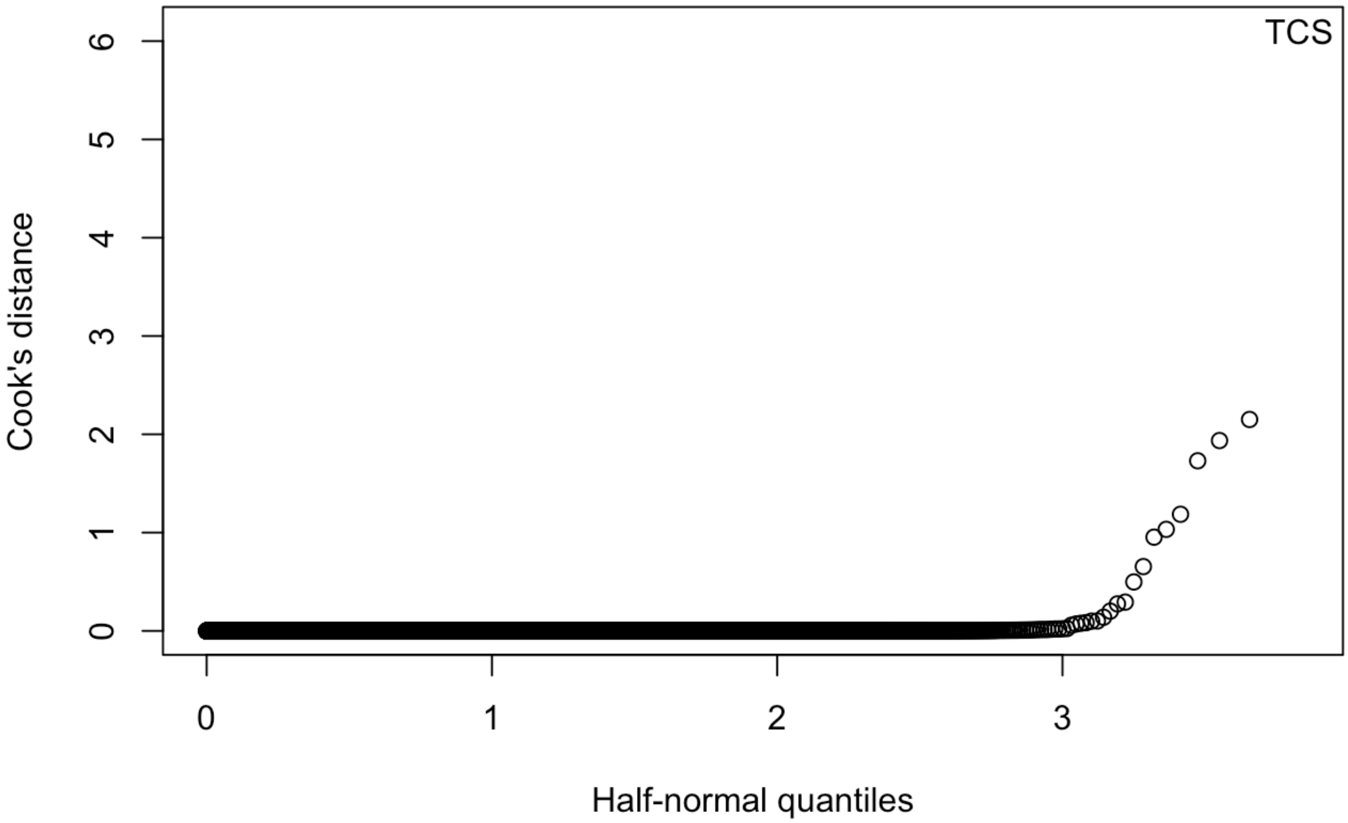
```
##                        TCS                  Accenture
##                          1                          2
##                  Cognizant                  Capgemini
##                          4                          5
##                  HDFC Bank                    Infosys
##                          6                          7
##                 ICICI Bank                    HCLTech
##                          8                          9
##                    Genpact            Teleperformance
##                         11                         12
##                  Axis Bank     Concentrix Corporation
##                         13                         14
##                        Jio                     Amazon
##                         15                         16
##                 iEnergizer            Reliance Retail
##                         17                         18
##     HDB Financial Services    Larsen & Toubro Limited
##                         21                         22
##                   Deloitte        Kotak Mahindra Bank
##                         23                         24
##              Vodafone Idea                     BYJU'S
##                         26                         27
##                        WNS                Tata Motors
##                         29                         31
##              Ernst & Young                        PwC
##                         33                         38
## Conneqt Business Solutions                    Startek
##                         44                         49
## Sutherland Global Services                        HGS
##                         56                         63
##               Ecom Express                Xyz Company
##                        138                        792
```

Using the jackknife-derived outlier threshold (crival = 4.41), we observe that many large firms have residual magnitudes far greater than the cutoff. This confirms these observations as true statistical outliers under formal studentized residual testing. Their extreme salary positions (either significantly above or below model-implied pay levels) indicate structural salary differences not captured by the current predictors.

Cook’s Distance:

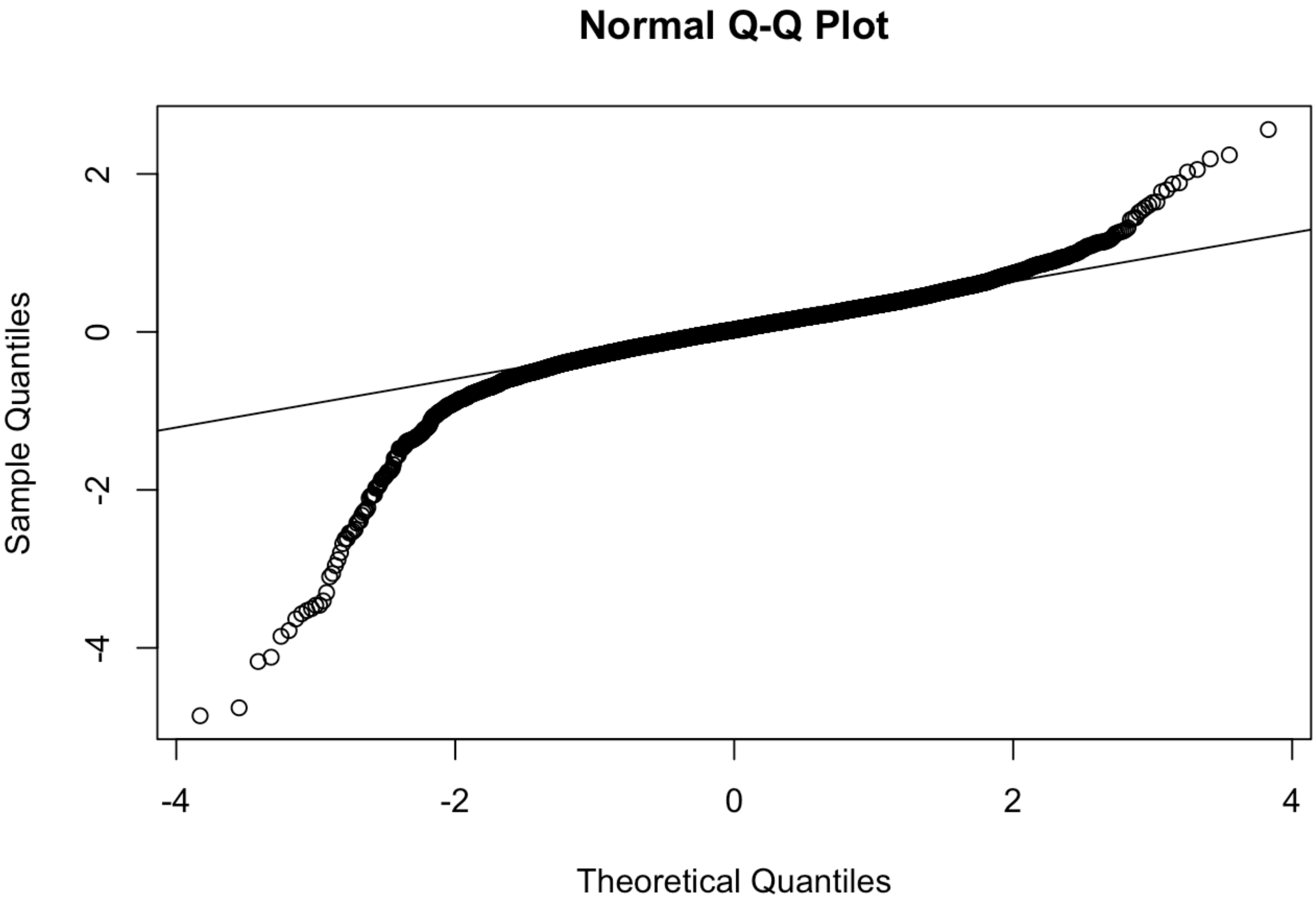
```
##
## Attaching package: 'faraway'
```

```
## The following objects are masked from 'package:car':
##
##     logit, vif
```



```
##            TCS      Accenture      Cognizant      HDFC Bank        Infosys
##              1              2              4              6              7
## Teleperformance
##             12
```

Removing Outliers that are not influential points.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
2
0
-2
-4
-4
-2
0
2
4

```
##
## Call:
## lm(formula = PACKAGE ~ YEARS.OLD + INDUSTRY + INDIA.HQ + TOTAL_EMPLOYEES +
##     BRANCHES + RATING + I(RATING^2) + REVIEWS, data = transformed_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8594 -0.1859  0.0265  0.2296  2.5620
##
## Coefficients:
##                                      Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                        -9.800e+00  4.353e-02 -225.135  < 2e-16 ***
## YEARS.OLD                          -5.603e-05  1.648e-04   -0.340 0.733963
## INDUSTRYEducation & Training       -3.014e-01  3.737e-02   -8.065 8.42e-16 ***
## INDUSTRYEngineering & Construction -8.144e-02  3.190e-02   -2.553 0.010686 *
## INDUSTRYFinancial Services          7.073e-02  3.639e-02    1.944 0.051981 .
## INDUSTRYHealthcare                 -1.325e-01  3.601e-02   -3.680 0.000235 ***
## INDUSTRYIndustrial Machinery        3.113e-02  3.303e-02    0.942 0.345986
## INDUSTRYInternet                   -7.064e-02  3.677e-02   -1.921 0.054745 .
## INDUSTRYIT Services & Consulting    1.221e-01  2.700e-02    4.522 6.22e-06 ***
## INDUSTRYOther                      -5.426e-02  2.436e-02   -2.227 0.025951 *
## INDUSTRYPharma                     -5.822e-02  3.396e-02   -1.714 0.086498 .
## INDUSTRYReal Estate                 2.298e-02  4.014e-02    0.573 0.566961
## INDUSTRYSoftware Product            1.883e-01  3.586e-02    5.251 1.56e-07 ***
## INDIA.HQChennai                    -9.712e-02  2.164e-02   -4.488 7.30e-06 ***
## INDIA.HQMumbai                     -1.572e-01  1.817e-02   -8.652  < 2e-16 ***
## INDIA.HQNew Delhi                  -2.459e-01  2.147e-02  -11.450  < 2e-16 ***
## INDIA.HQOther                      -2.009e-01  1.486e-02  -13.522  < 2e-16 ***
## INDIA.HQPune                       -4.040e-02  2.191e-02   -1.844 0.065242 .
## TOTAL_EMPLOYEES                    -2.299e-03  4.103e-03   -0.560 0.575339
## BRANCHES                           -5.467e-02  6.559e-03   -8.335  < 2e-16 ***
## RATING                             -5.029e-01  1.332e-02  -37.750  < 2e-16 ***
## I(RATING^2)                        -6.265e-01  1.902e-02  -32.946  < 2e-16 ***
## REVIEWS                             1.013e+00  8.351e-03  121.303  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4342 on 7753 degrees of freedom
##   (1647 observations deleted due to missingness)
## Multiple R-squared:  0.8305, Adjusted R-squared:   0.83
## F-statistic:  1727 on 22 and 7753 DF,  p-value: < 2.2e-16
```