Diagnostics

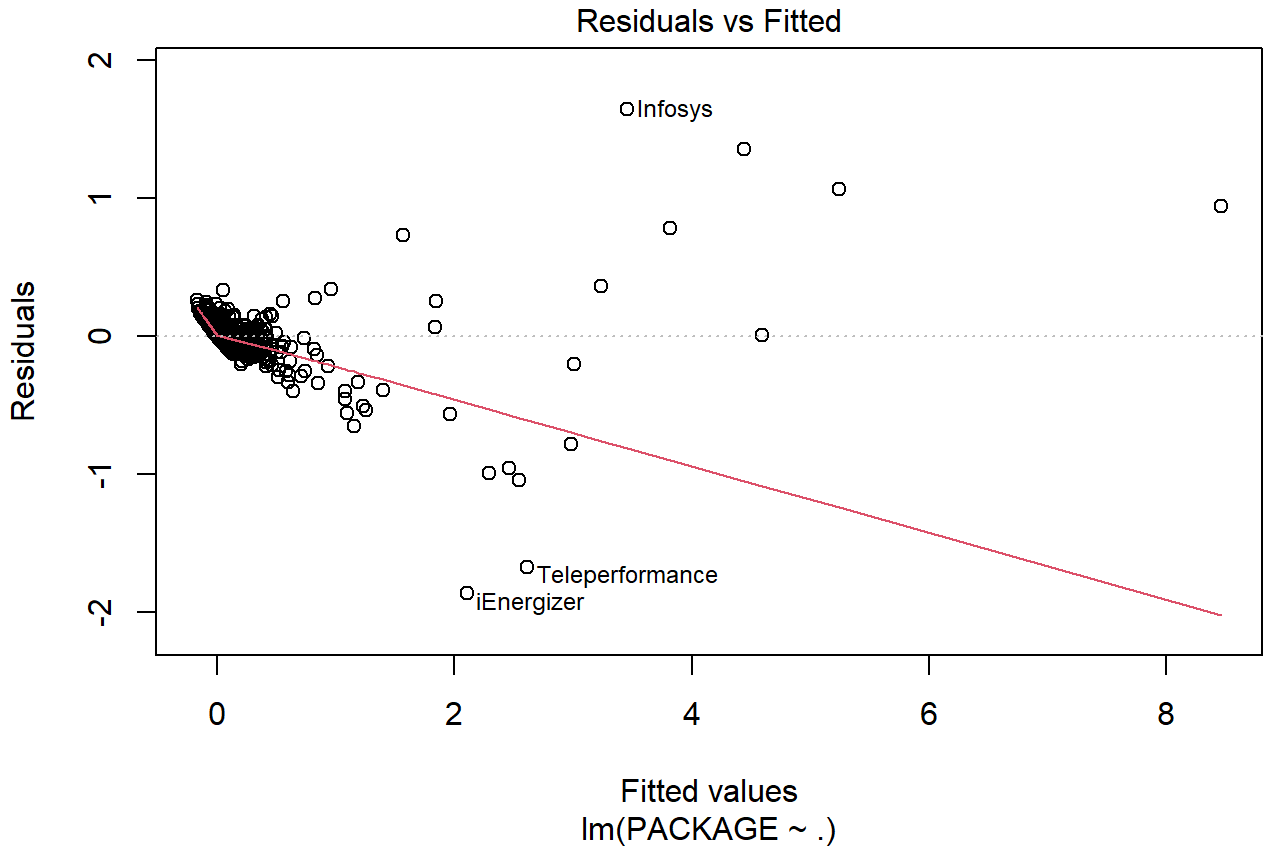
2025-12-05

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
```

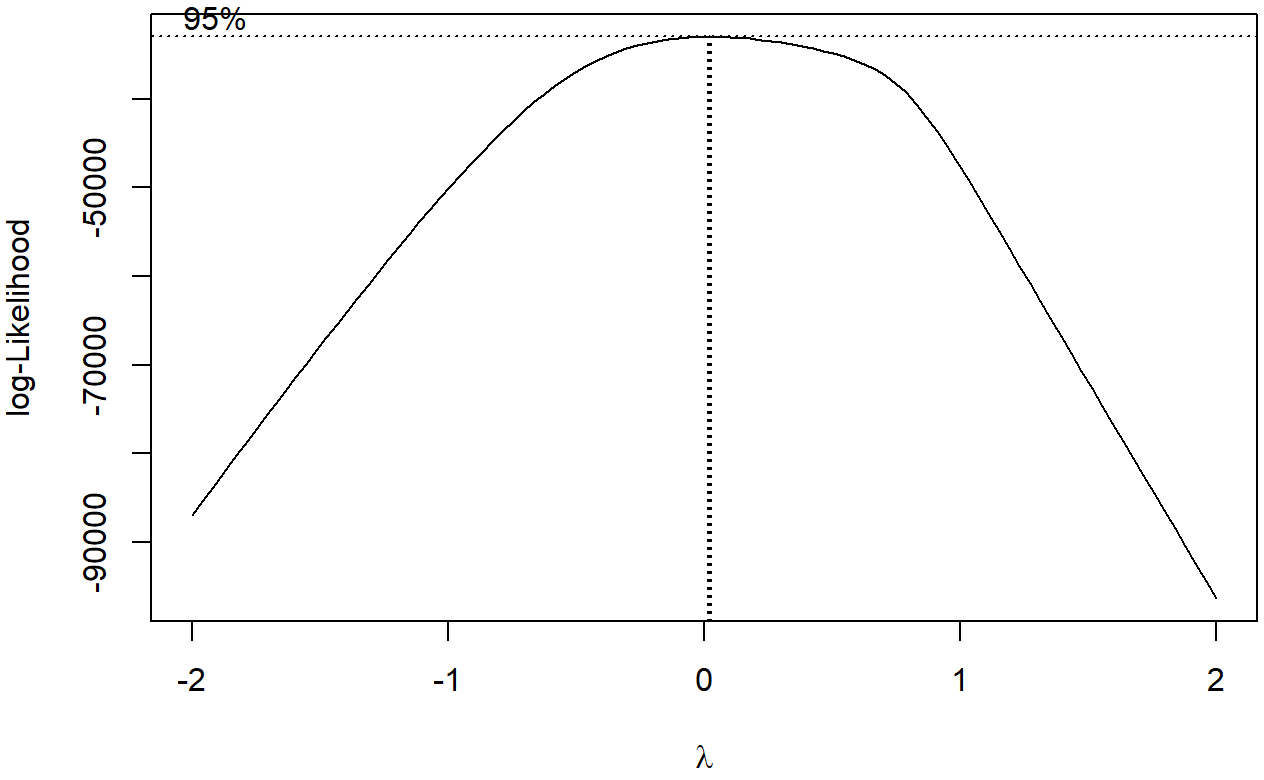
```
##
## Call:
## lm(formula = PACKAGE ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86612 -0.00896 -0.00184  0.00745  1.64315
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       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

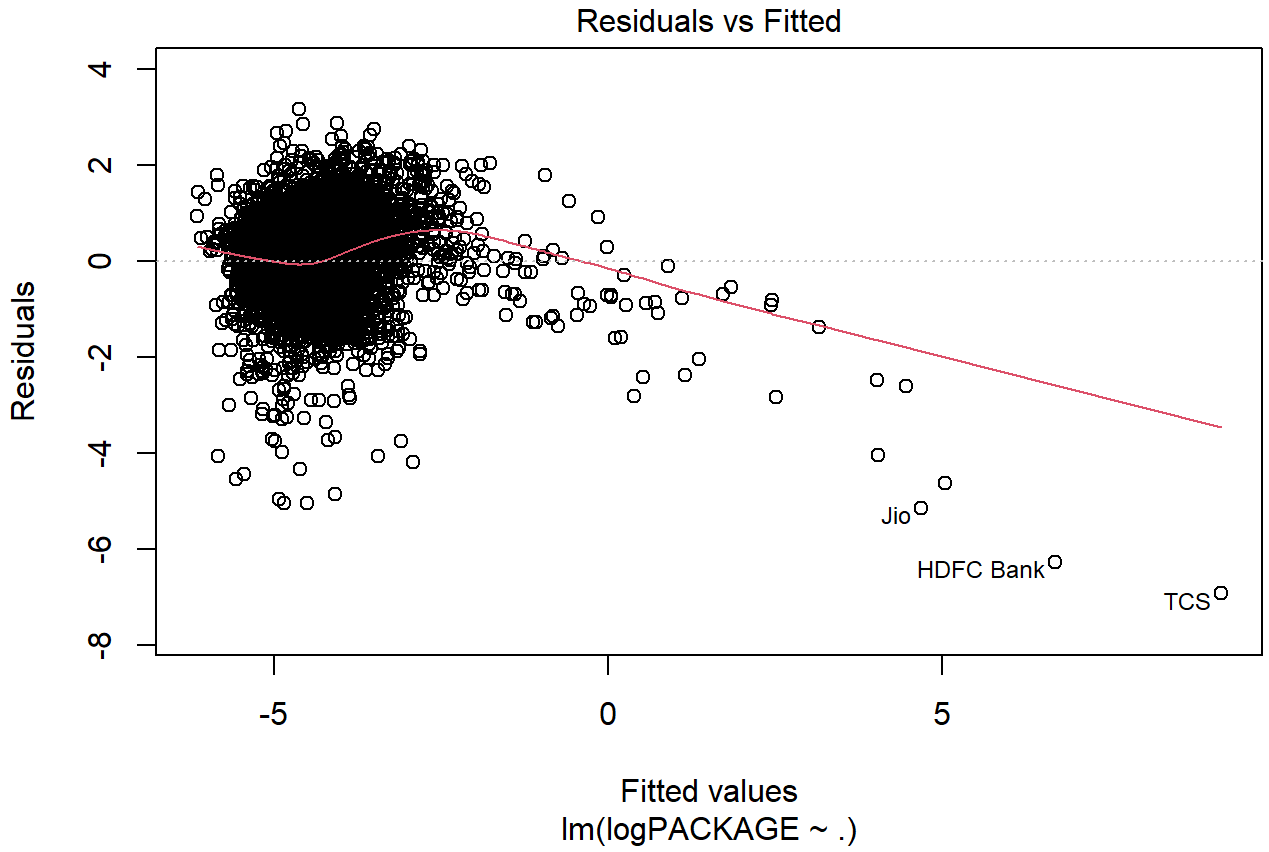

95%
log-Likelihood
-50000
-70000
-90000
-2
-1
0
1
2
λ

```
## $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
```

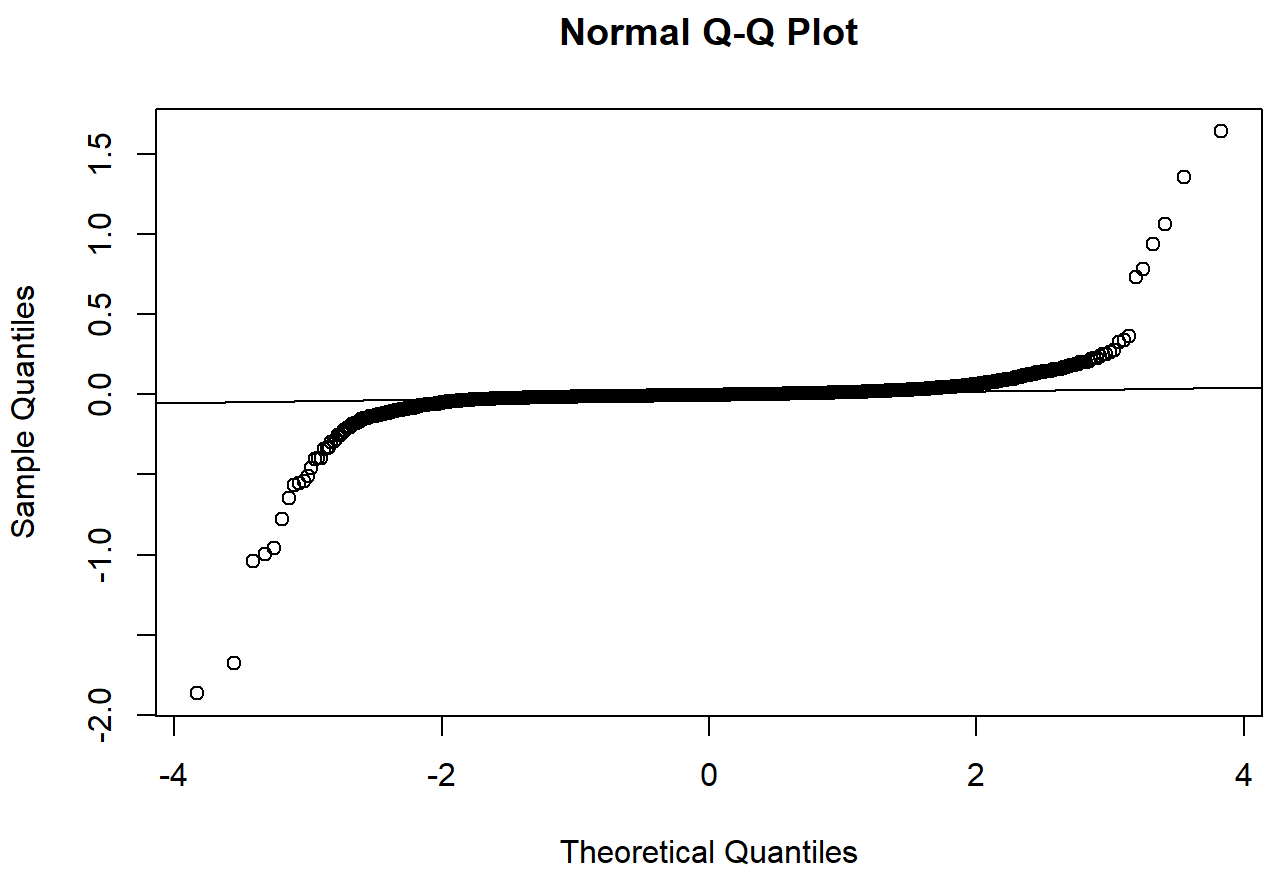
Since λ is close to zero, we apply log transformations.

```
##
## Call:
## lm(formula = logPACKAGE ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9295 -0.4381 -0.0094  0.4502  3.1559
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -4.651e+00  1.066e-01 -43.639  < 2e-16 ***
## YEARS.OLD                         -1.073e-07  2.946e-04   0.000 0.999710
## INDUSTRYEducation & Training      -6.431e-01  6.675e-02  -9.635  < 2e-16 ***
## INDUSTRYEngineering & Construction -2.941e-01  5.683e-02  -5.176 2.33e-07 ***
## INDUSTRYFinancial Services        -6.504e-02  6.527e-02  -0.996 0.319043
## INDUSTRYHealthcare                -3.473e-01  6.444e-02  -5.390 7.25e-08 ***
## INDUSTRYIndustrial Machinery      -2.047e-01  5.903e-02  -3.467 0.000528 ***
## INDUSTRYInternet                  -1.959e-01  6.568e-02  -2.983 0.002863 **
## INDUSTRYIT Services & Consulting  -3.906e-02  4.817e-02  -0.811 0.417432
## INDUSTRYOther                     -2.290e-01  4.351e-02  -5.262 1.46e-07 ***
## INDUSTRYPharma                    -1.485e-01  6.080e-02  -2.442 0.014613 *
## INDUSTRYReal Estate               -2.362e-01  7.171e-02  -3.293 0.000995 ***
## INDUSTRYSoftware Product           6.113e-02  6.384e-02   0.958 0.338283
## INDIA.HQChennai                   -1.309e-01  3.864e-02  -3.388 0.000707 ***
## INDIA.HQMumbai                    -2.780e-01  3.227e-02  -8.616  < 2e-16 ***
## INDIA.HQNew Delhi                 -4.500e-01  3.790e-02 -11.874  < 2e-16 ***
## INDIA.HQOther                     -3.176e-01  2.652e-02 -11.976  < 2e-16 ***
## INDIA.HQPune                      -7.724e-02  3.913e-02  -1.974 0.048450 *
## TOTAL_EMPLOYEES                    2.529e-01  6.417e-03  39.416  < 2e-16 ***
## BRANCHES                           3.351e-03  1.735e-04  19.306  < 2e-16 ***
## RATING                            -3.579e-01  2.290e-02 -15.625  < 2e-16 ***
## REVIEWS                            8.349e-05  1.260e-05   6.626 3.67e-11 ***
## PACKAGE                            1.530e-01  1.498e-01   1.021 0.307218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7777 on 7778 degrees of freedom
##   (1648 observations deleted due to missingness)
## Multiple R-squared:  0.4814, Adjusted R-squared:   0.48
## F-statistic: 328.2 on 22 and 7778 DF,  p-value: < 2.2e-16
```



Heteroscedasticity is still not eliminated.

Q-Q Plot:

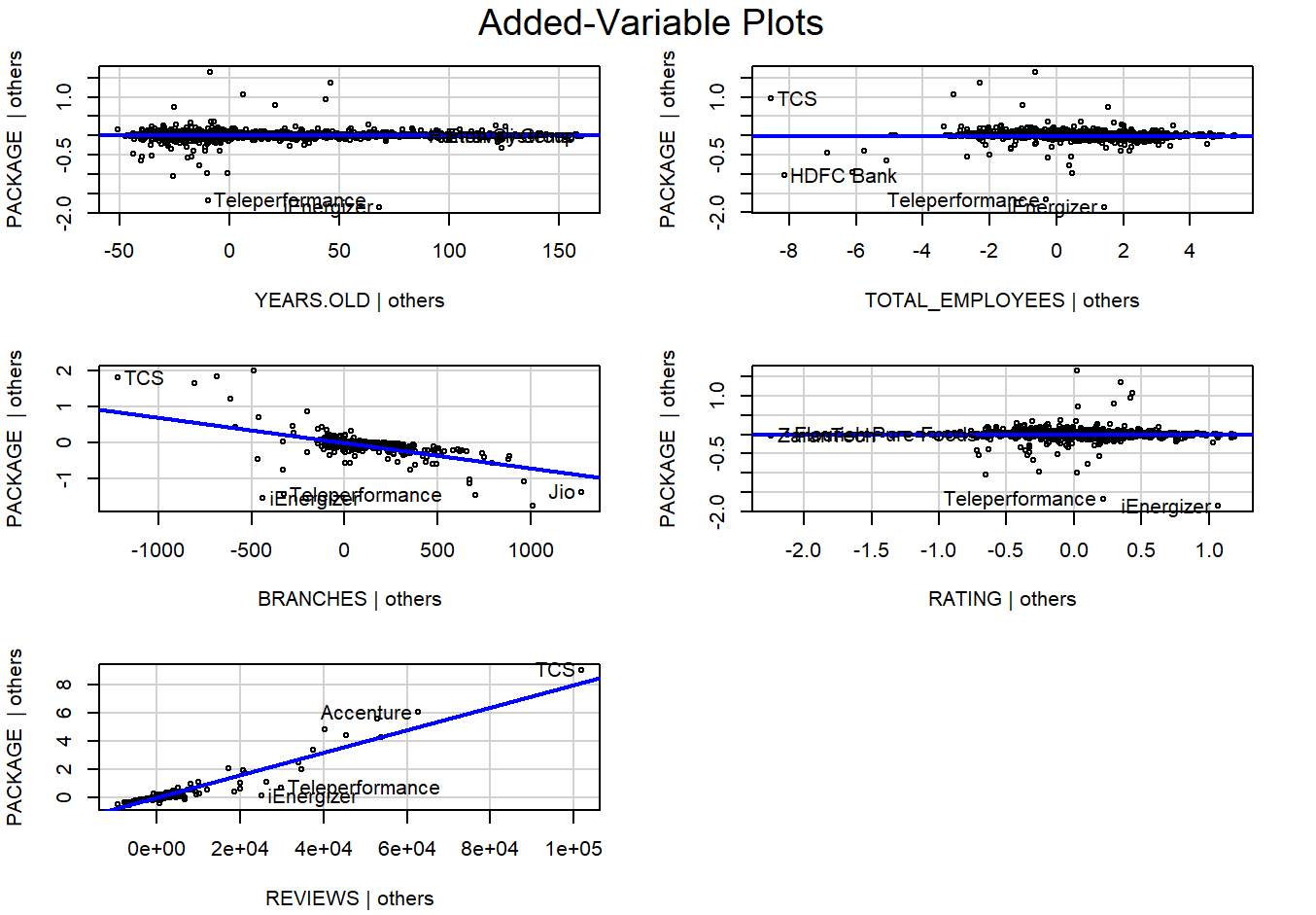


The Q-Q plot shows strong departure from normality, with heavy tails on both ends. Extreme upper-tail and lower-tail values indicate that the error distribution is heavy-tailed, violating the normality assumption. This suggests either influential companies or that certain predictors have long-tailed behavior impacting salaries. Transformation or robust regression could be considered.

Shapiro Test can’t be performed because there are more than 5000 rows in the dataset.

Residual Plots:

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



BRANCHES has a strong negative partial relationship with salary. REVIEWS shows a strong positive partial relationship with salary, increasing sharply for high review counts. YEARS.OLD, TOTAL_EMPLOYEES, and RATING demonstrate weak or flat partial effects, indicating limited independent contribution once other predictors are controlled. Several observations (e.g., TCS, Teleperformance, iEnergizer, HDFC Bank) appear as outlying or high-leverage points in multiple plots, reinforcing influence concerns.

Leverage Points:

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 or very small firms (e.g., TCS, Reliance Retail, Infosys, Axis Bank). Their combination of extreme workforce/branch presence makes them disproportionately influential in estimating regression coefficients.

Jackknife Residuals:

```
## [1] 4.407288
```

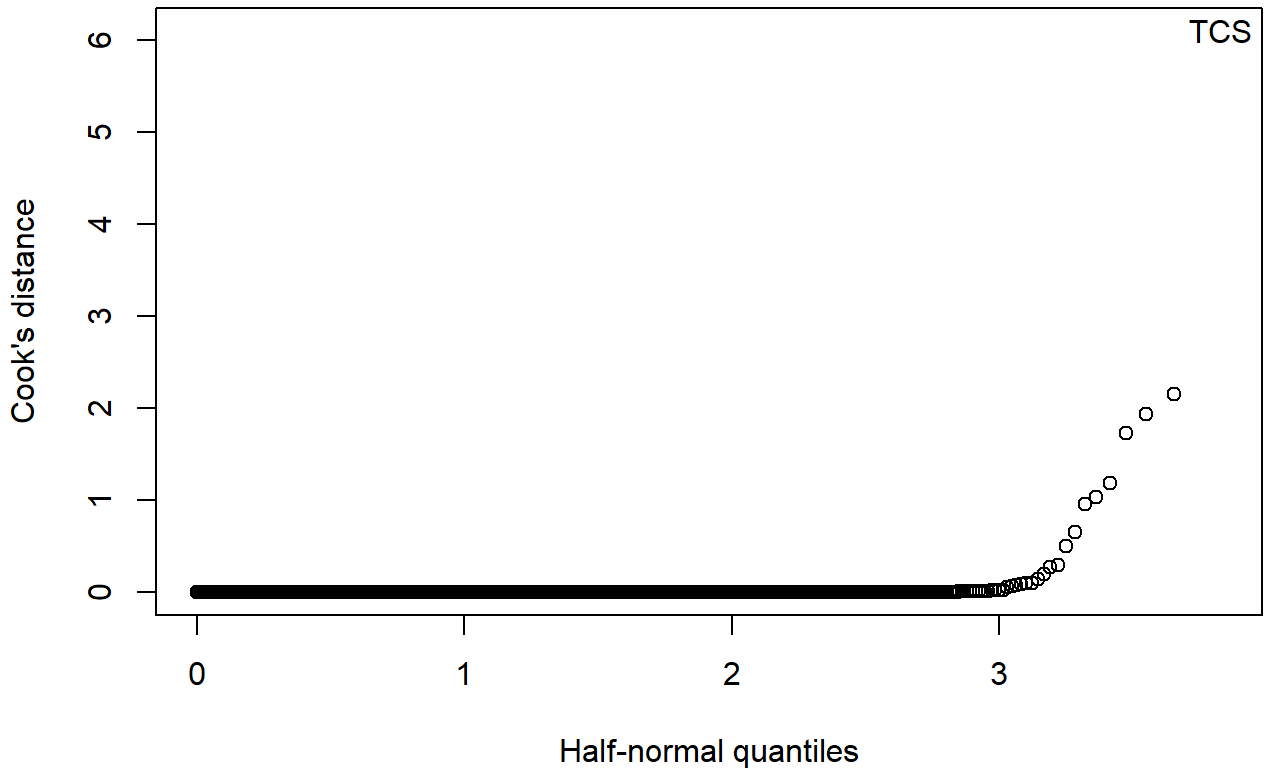
```
##                       TCS                 Accenture
##                         1                         2
##                 Cognizant                 Capgemini
##                         4                         5
##                 HDFC Bank                   Infosys
##                         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 iEnergizer, Teleperformance, Infosys, Cognizant, Accenture, and TCS have residual magnitudes far greater than the cutoff. This confirms these observations as true statistical outliers under formal studentized residual testing, rather than heuristic cutoffs. 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
```



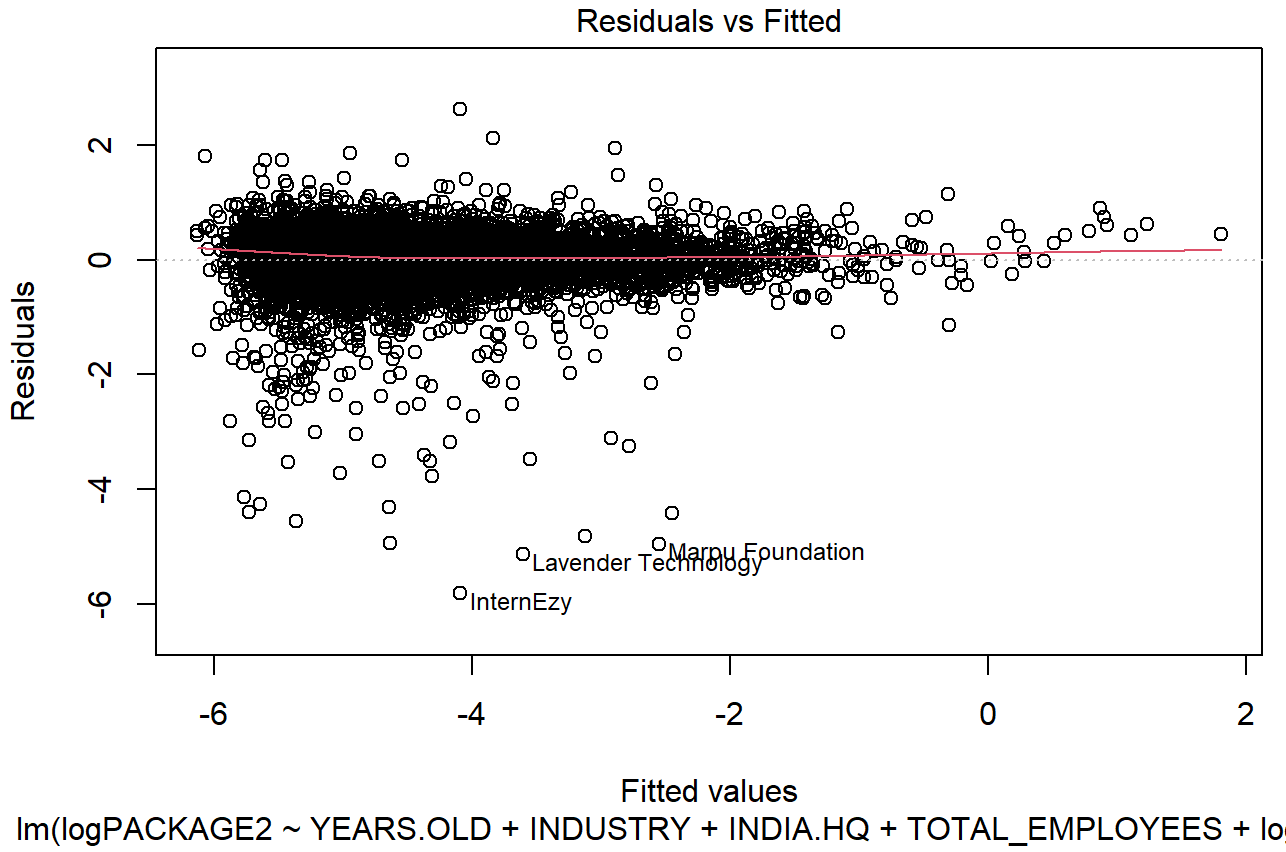

```
##             TCS       Accenture       Cognizant       HDFC Bank         Infosys
##               1               2               4               6               7
## Teleperformance
##              12
```

Log transformation since heteroscedasticity wasn’t completely eliminated:

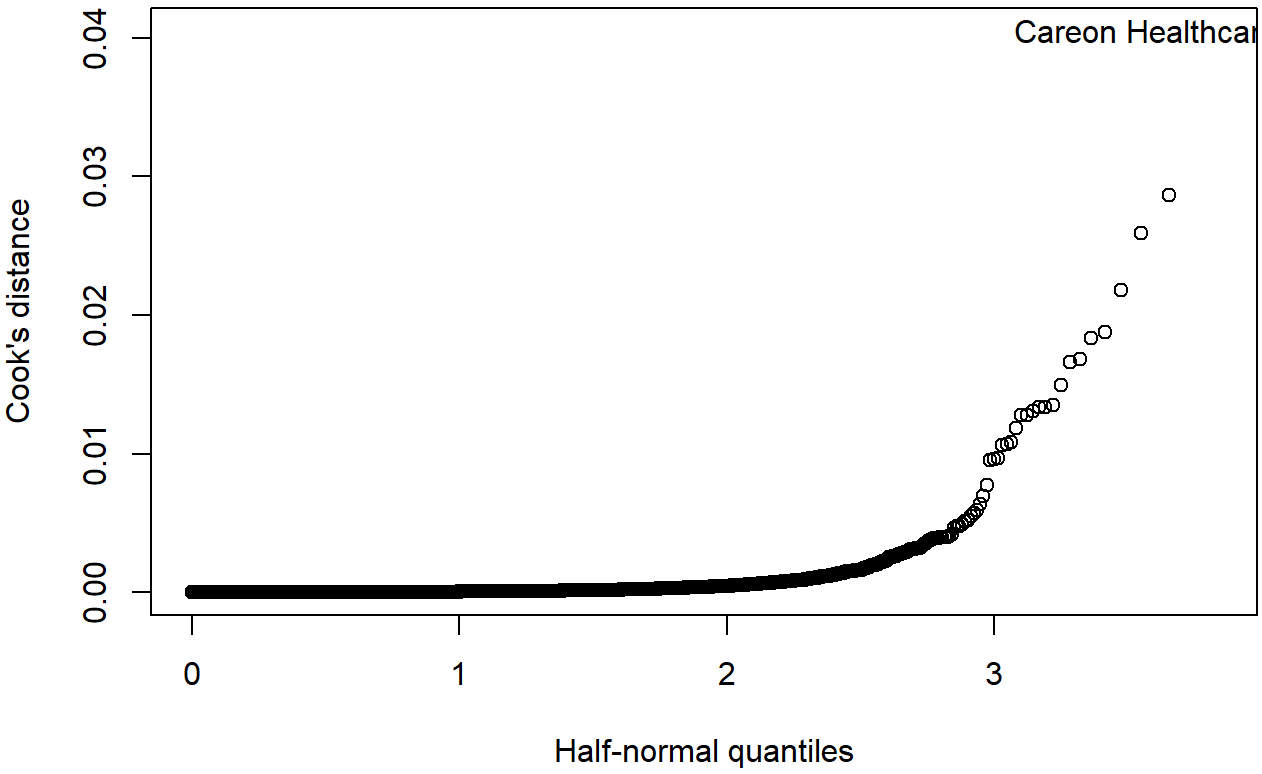
```
##
## Call:
## lm(formula = logPACKAGE2 ~ YEARS.OLD + INDUSTRY + INDIA.HQ +
##     TOTAL_EMPLOYEES + logBRANCHES + logRATING + logREVIEWS, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8133 -0.1702  0.0444  0.2432  2.6172
##
## Coefficients:
##                                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                       -8.306e+00  7.864e-02 -105.627  < 2e-16 ***
## YEARS.OLD                         -3.786e-05  1.775e-04   -0.213 0.831075
## INDUSTRYEducation & Training      -3.605e-01  4.028e-02   -8.951  < 2e-16 ***
## INDUSTRYEngineering & Construction -1.314e-01  3.435e-02   -3.826 0.000131 ***
## INDUSTRYFinancial Services         4.774e-02  3.926e-02    1.216 0.224026
## INDUSTRYHealthcare                -1.689e-01  3.884e-02   -4.347 1.40e-05 ***
## INDUSTRYIndustrial Machinery       2.704e-02  3.564e-02    0.759 0.447996
## INDUSTRYInternet                  -1.287e-01  3.958e-02   -3.252 0.001149 **
## INDUSTRYIT Services & Consulting   8.192e-02  2.909e-02    2.816 0.004878 **
## INDUSTRYOther                     -9.345e-02  2.625e-02   -3.560 0.000373 ***
## INDUSTRYPharma                    -7.052e-02  3.664e-02   -1.924 0.054339 .
## INDUSTRYReal Estate               -3.122e-02  4.327e-02   -0.721 0.470685
## INDUSTRYSoftware Product           1.324e-01  3.865e-02    3.424 0.000619 ***
## INDIA.HQChennai                   -8.190e-02  2.328e-02   -3.518 0.000438 ***
## INDIA.HQMumbai                    -1.637e-01  1.956e-02   -8.373  < 2e-16 ***
## INDIA.HQNew Delhi                 -2.745e-01  2.313e-02  -11.865  < 2e-16 ***
## INDIA.HQOther                     -2.051e-01  1.600e-02  -12.818  < 2e-16 ***
## INDIA.HQPune                      -4.346e-02  2.358e-02   -1.843 0.065297 .
## TOTAL_EMPLOYEES                   -2.333e-03  4.426e-03   -0.527 0.598058
## logBRANCHES                       -3.812e-02  7.041e-03   -5.414 6.35e-08 ***
## logRATING                         -1.231e+00  4.942e-02  -24.906  < 2e-16 ***
## logREVIEWS                         1.016e+00  8.858e-03  114.700  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4685 on 7779 degrees of freedom
##   (1648 observations deleted due to missingness)
## Multiple R-squared:  0.8118, Adjusted R-squared:  0.8113
## F-statistic:  1598 on 21 and 7779 DF,  p-value: < 2.2e-16
```

Residuals vs Fitted after transformations

```
plot(lmod_log2, 1)
```



Cook’s distance after transformations:


Careon Healthcar
Cook's distance
0.00
0.01
0.02
0.03
0.04
0
1
2
3
Half-normal quantiles