finalprojectcode

Tong Su

12/11/2021

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(car)
```

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

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(patchwork)
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
thedata <- read_csv("2020_ontario_public_library_statistics_open_data.csv")
```

```
## Warning: Missing column names filled in: 'X328' [328]
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   `Library Full Name` = col_character(),
##   `Library Number` = col_character(),
```

```
##   `A1.3 Ontario Library Service (OLS) Region (English)` = col_character(),
##   `A1.4 Type of Library Service (English)` = col_character(),
##   `A1.5 Mailing Address` = col_character(),
##   `A1.9 Street Address` = col_character(),
##   `A1.10 City/Town` = col_character(),
##   `A1.11 Province` = col_character(),
##   `A1.12 Postal Code` = col_character(),
##   `A1.13 Web Site Address` = col_character(),
##   `A1.14  No. of Active Library Cardholders` = col_number(),
##   `B1.1  Net Balance brought forward from previous year` = col_number(),
##   `B2.1.1  Public Library Operating Grant (PLOG)` = col_number(),
##   `B2.1.2  Pay Equity Grant` = col_number(),
##   `B2.1.3  First Nation Salary Supplement Grant` = col_number(),
##   `B2.1.4  Total Provincial Operating Funding` = col_number(),
##   `B2.2  Local Operating Funding (e.g. Municipality or Band local operating funding)` = col_number()
##   `B2.3  Contract Revenue (funds from other municipalities, neighbouring public library boards, Local
##   `B2.4.1.03  Trillium Foundation funding` = col_number(),
##   `B2.4.1.04  Provincial student employment funding` = col_number()
##   # ... with 145 more columns
## )
## i Use `spec()` for the full column specifications.
```

```
thedata <- clean_names(thedata)
```

```
thedata <- thedata %>%
  mutate(total_funds = b4_02_1_total_funds_not_including_employee_benefits+b4_02_2_employee_benefits,
         number_of_resources=b4_01_1_general_include_all_physical_items_that_are_not_electronic_e_g_bool
         number_of_visitors=g1_5_1_w_no_of_visits_to_the_library_made_in_person+g1_5_2_w_no_of_electroni
         region = ifelse(a1_3_ontario_library_service_ols_region_english == "Southern Ontario Library Se
```

```
thedata <- thedata %>%
  filter(number_of_visitors<10000)

thedata <- thedata %>%
  filter(number_of_visitors>0)
```

```
thedata <- thedata %>%
  dplyr::select(a1_14_no_of_active_library_cardholders,
         b2_9_total_operating_revenues,
         b5_0_total_operating_expenditures,
         e7_1_in_the_space_provided_please_provide_the_total_combined_square_footage_of_all_the_facilit
         g1_1_3_w_total_circulation_of_all_library_materials,
         total_funds,
         number_of_resources,
         number_of_visitors,
         region,
         library_full_name)
```

```
thedata <-rename(thedata, number_of_cardholders=a1_14_no_of_active_library_cardholders,
         total_operating_revenues=b2_9_total_operating_revenues,
         total_operating_expenditures=b5_0_total_operating_expenditures,
         total_space_provided=e7_1_in_the_space_provided_please_provide_the_total_combined_square_footag
         total_circulation_of_materials=g1_1_3_w_total_circulation_of_all_library_materials,
         Name = library_full_name)
```

```r
nrow(thedata)
```

```
## [1] 269
```

```r
# create a 50/50 split in the data
set.seed(1)
train <- thedata[sample(1:nrow(thedata), 135, replace=F), ]
test <- thedata[which(!(thedata$Name %in% train$Name)),]
```

```r
summary(train$number_of_visitors)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    39.5   147.0  1064.1  1371.5  9599.0
```

```r
summary(train$number_of_resources)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    5994   16664   81478   53316 1037885
```

```r
summary(train$number_of_cardholders)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      10     334    1037    4866    4545   49269
```

```r
summary(train$total_funds)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   36126   90016  474003  387744 6401088
```

```r
summary(train$total_operating_revenues)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1338   54174  150829  767698  599590 9006450
```

```r
summary(train$total_operating_expenditures)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1785   54173  158952  737140  594486 9006450
```

```r
summary(train$total_space_provided)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    1119    3200   10263    9686   95085
```

```r
summary(train$total_circulation_of_materials)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1      57     281    1454    1191   18429
```

```r
summary(test$number_of_visitors)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    47.5   282.0  1201.7  1425.2  9128.0
```

```r
summary(test$number_of_resources)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    6418   23388   87129   71161  971317
```

```r
summary(test$number_of_cardholders)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##      10     445    1731    5176    6763   50181
```

```
summary(test$total_funds)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   37511  147742  563179  550580 4791843
```

```
summary(test$total_operating_revenues)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3023   71937  251698  880984  836715 8878557
```

```
summary(test$total_operating_expenditures)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3023   57548  222314  855198  829297 9527220
```

```
summary(test$total_space_provided)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    1365    4904   11188   13450   98045
```

```
summary(test$total_circulation_of_materials)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     1.00    83.25   475.50  2706.31  1830.00 93901.00
```

```
sd(train$number_of_visitors)
```

```
## [1] 1911.006
```

```
sd(train$number_of_resources)
```

```
## [1] 175534.8
```

```
sd(train$number_of_cardholders)
```

```
## [1] 8755.451
```

```
sd(train$total_funds)
```

```
## [1] 909788.8
```

```
sd(train$total_operating_revenues)
```

```
## [1] 1427611
```

```
sd(train$total_operating_expenditures)
```

```
## [1] 1381359
```

```
sd(train$total_space_provided)
```

```
## [1] 17043.68
```

```
sd(train$total_circulation_of_materials)
```

```
## [1] 3159.106
```

```
sd(test$number_of_visitors)
```

```
## [1] 1838.827
```

```
sd(test$number_of_resources)
```

```
## [1] 172337.9
```

```
sd(test$number_of_cardholders)
```

```
## [1] 8444.281
```

```
sd(test$total_funds)
```

```
## [1] 968200.7
```

```
sd(test$total_operating_revenues)
```

```
## [1] 1545803
```

```
sd(test$total_operating_expenditures)
```

```
## [1] 1536276
```

```
sd(test$total_space_provided)
```

```
## [1] 17239.24
```

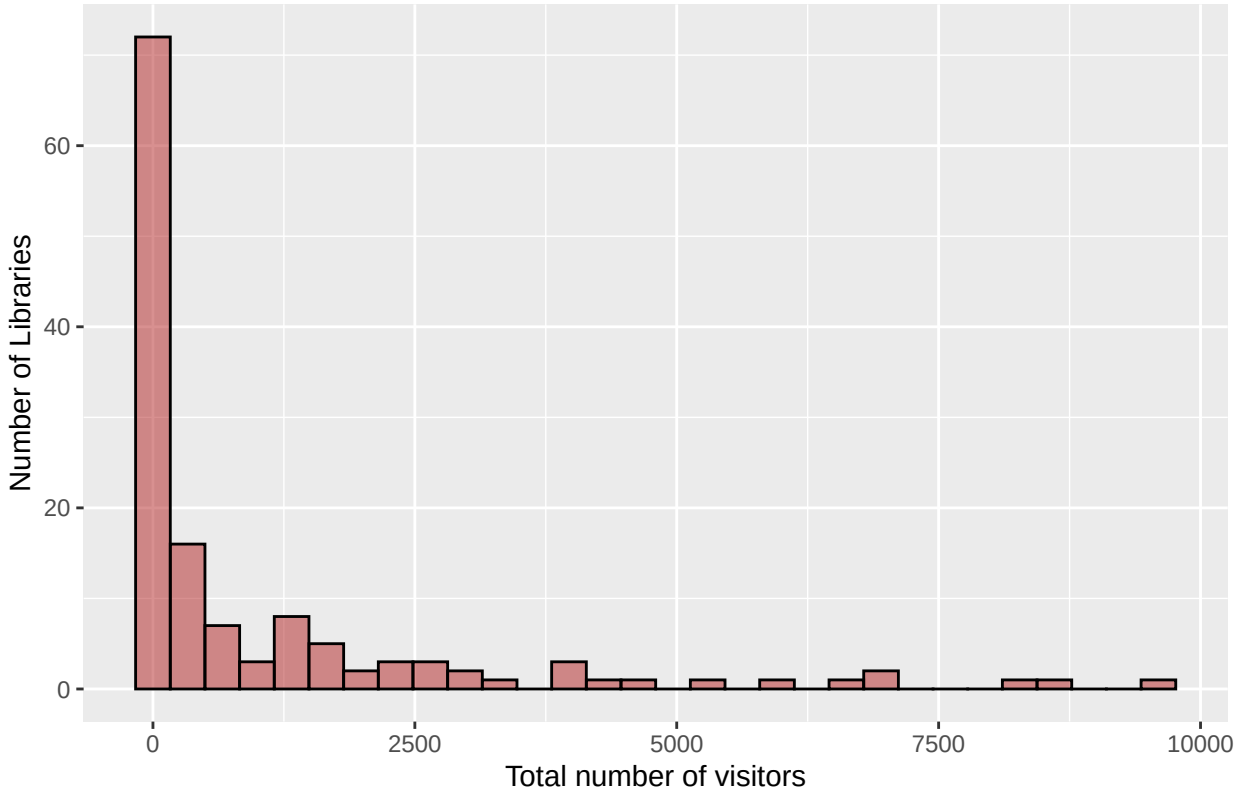
```
sd(test$total_circulation_of_materials)
```

```
## [1] 9161.654
```

```
# EDA of the response variable
train %>%
  ggplot(aes(x=number_of_visitors)) +
  geom_histogram(colour="black", fill="firebrick", alpha=0.5) +
  labs(x="Total number of visitors", y = "Number of Libraries" , title = "Fig 1. Histogram of the total
```

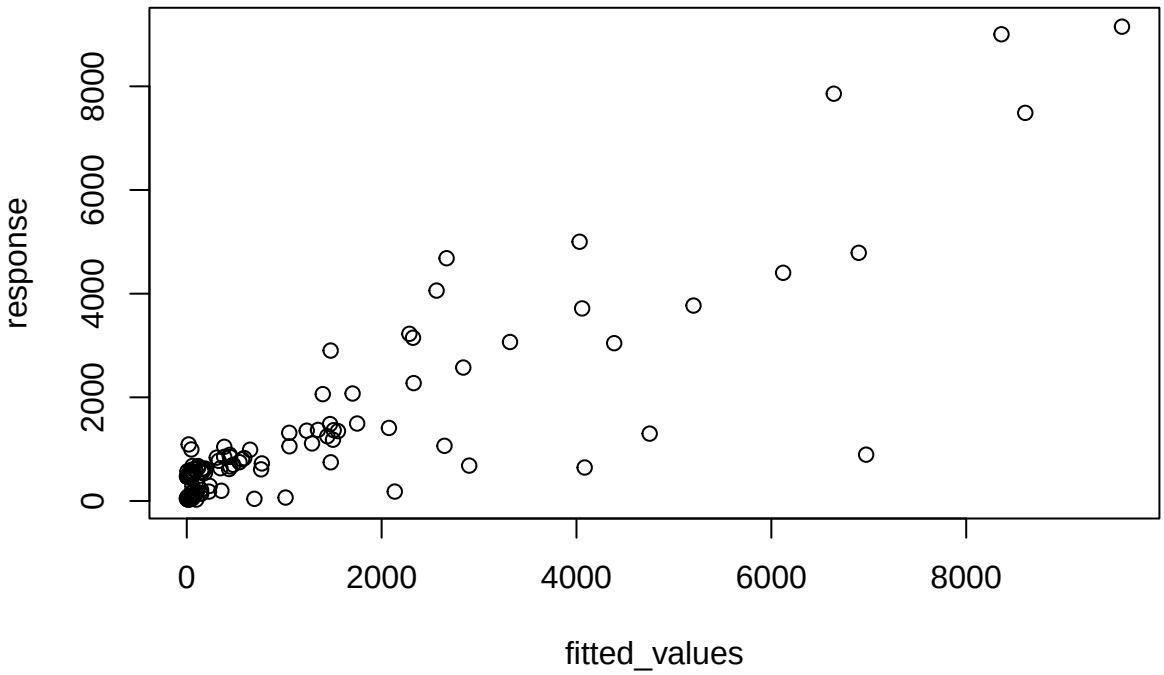
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Fig 1. Histogram of the total number of visitors

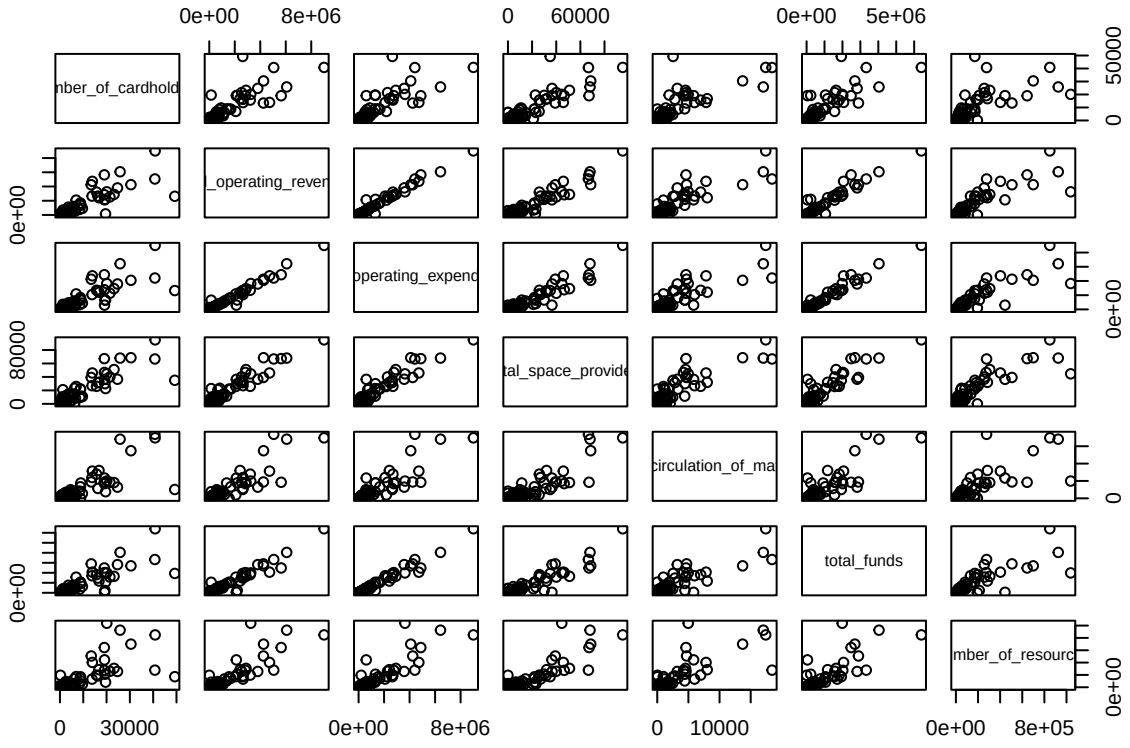



```
#primary full model
model_full <- lm(number_of_visitors ~ . -Name, data=train)
```

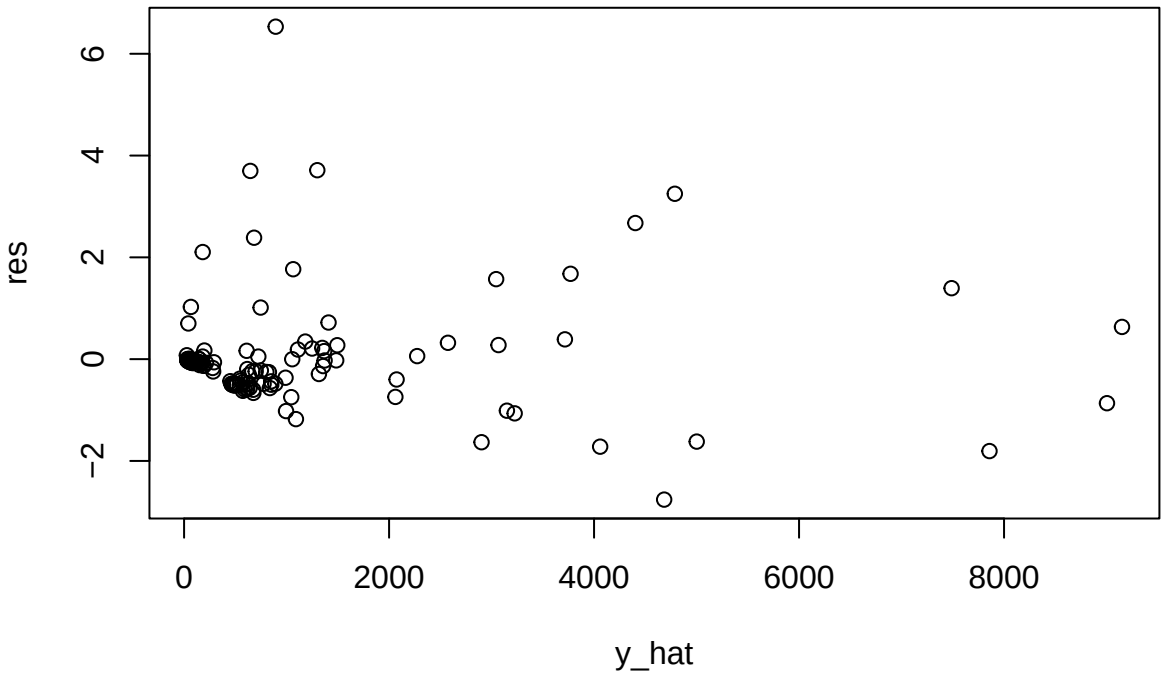
```
response <- fitted(model_full)
fitted_values <- train$number_of_visitors

plot(fitted_values, response)
```



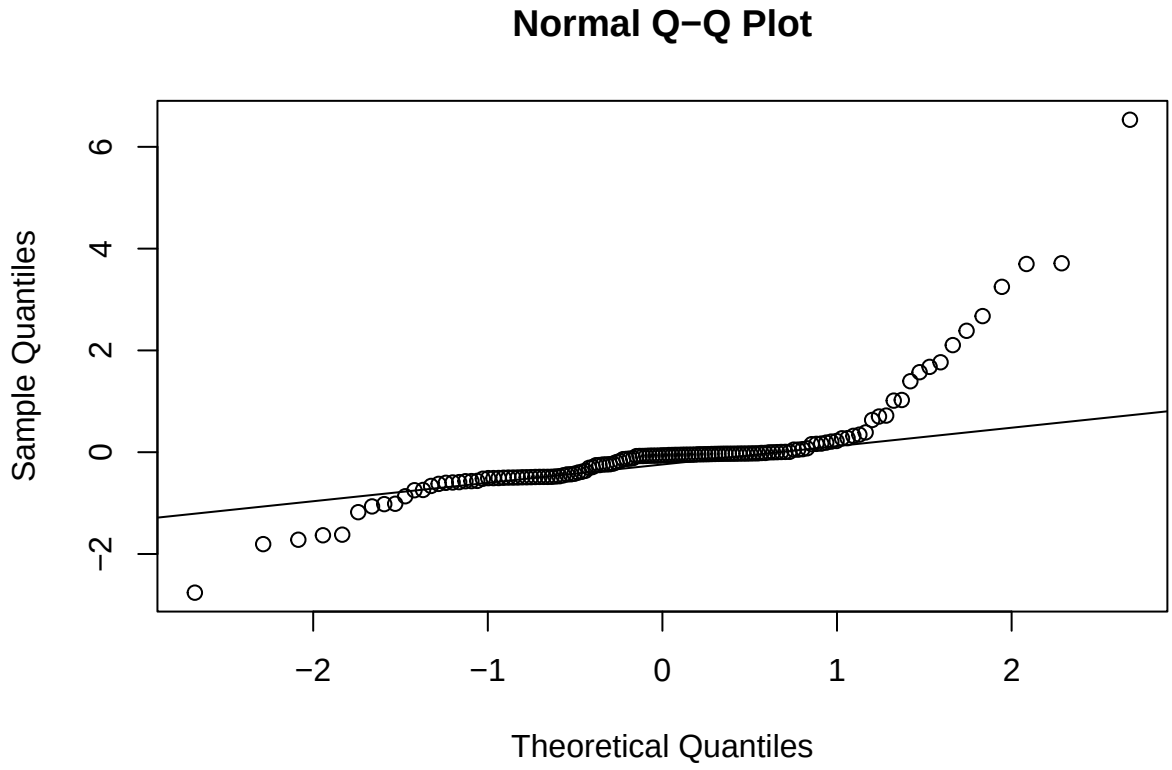

```
pairs(~number_of_cardholders+total_operating_revenues+total_operating_expenditures+total_space_provided+
```



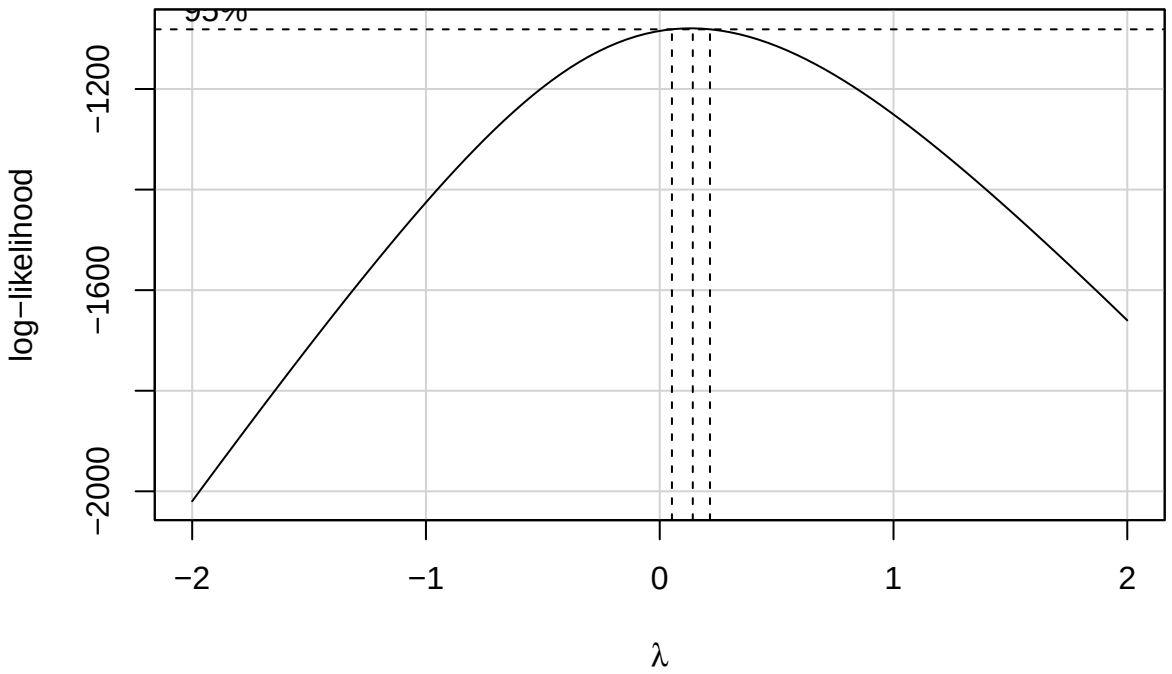

```
res<-rstandard(model_full)
y_hat <- fitted(model_full)

plot(y_hat, res)
```




```
qqnorm(res)
qqline(res)
```




```
train <- train %>%
  mutate(total_space_provided = total_space_provided+0.00000000000001,
         total_funds = total_funds+0.00000000000001,
         number_of_resources = number_of_resources+0.00000000000001)
boxCox(model_full)
```




```
summary(powerTransform(cbind(train$number_of_cardholders,
                             train$total_operating_revenues,
                             train$total_operating_expenditures,
                             train$total_space_provided,
                             train$total_circulation_of_materials,
                             train$total_funds,
                             train$number_of_resources)))
```

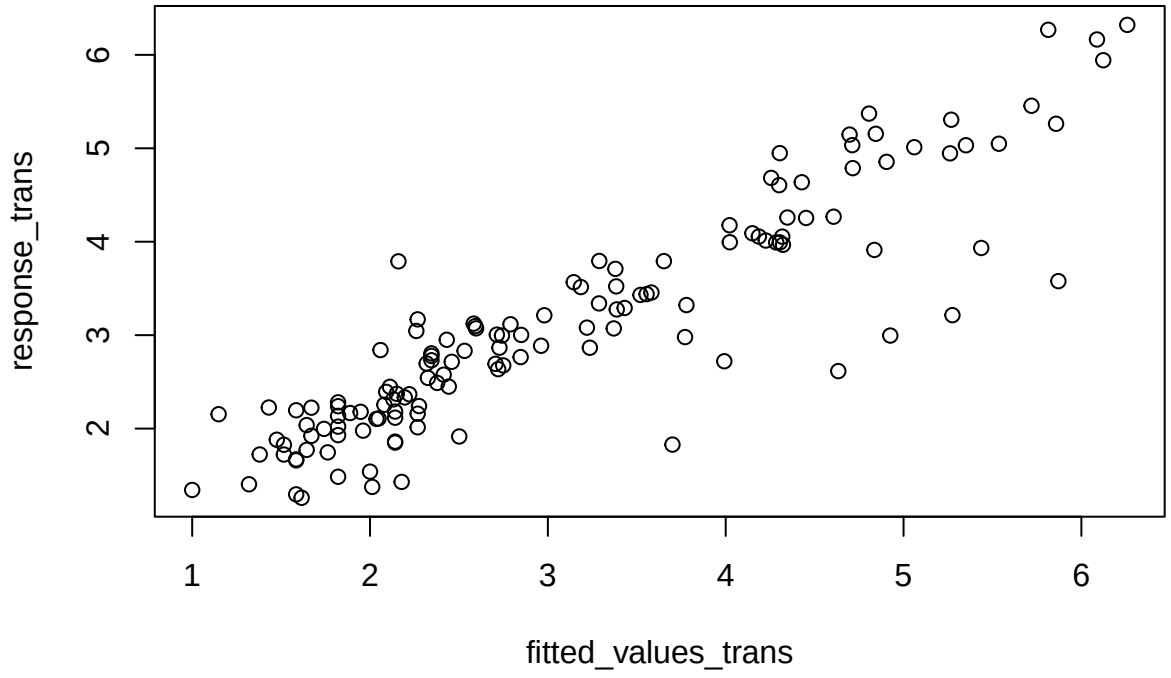
```
## bcPower Transformations to Multinormality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.1812        0.18       0.1283       0.2341
## Y2    0.2137        0.21       0.1687       0.2586
## Y3    0.2172        0.22       0.1821       0.2522
## Y4    0.1788        0.18       0.1503       0.2073
## Y5    0.2070        0.21       0.1610       0.2530
## Y6    0.2724        0.27       0.2459       0.2989
## Y7    0.2410        0.24       0.2161       0.2660
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                       LRT df       pval
## LR test, lambda = (0 0 0 0 0 0 0) 1505.603  7 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                       LRT df       pval
## LR test, lambda = (1 1 1 1 1 1 1) 4030.521  7 < 2.22e-16
```

```
train_trans <- train %>%
  mutate(number_of_cardholders=number_of_cardholders^(0.2),
                             total_operating_revenues=total_operating_revenues^(0.2),
```

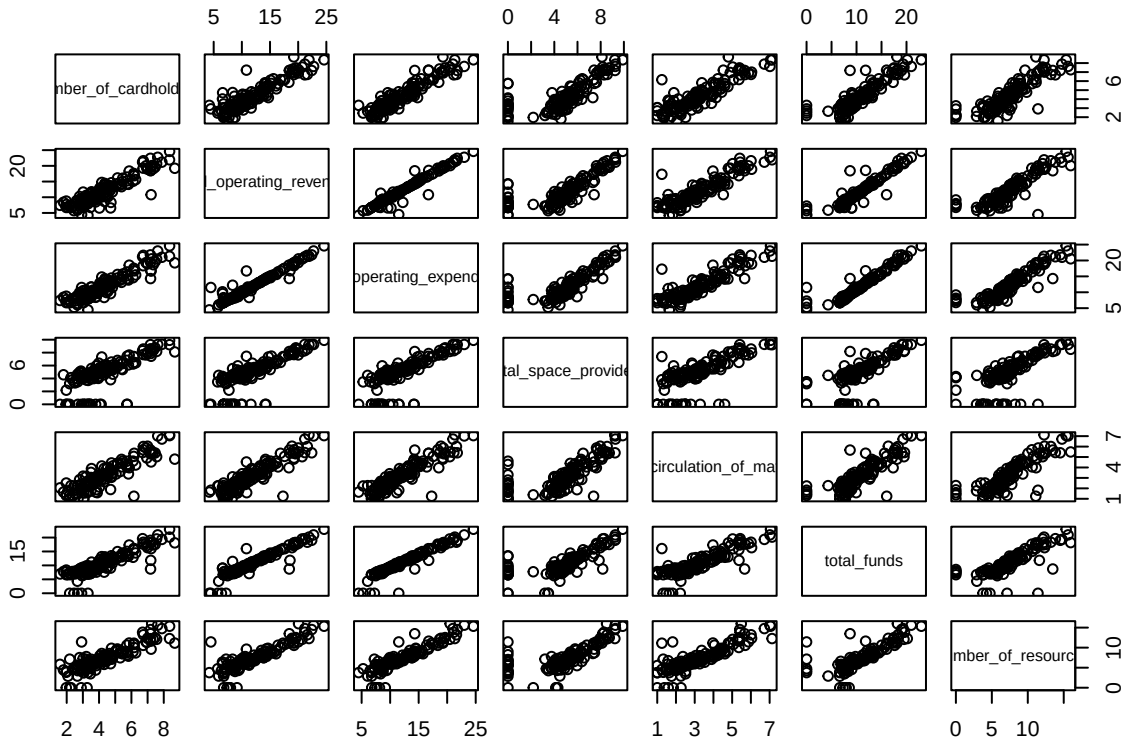
```
                            total_operating_expenditures=total_operating_expenditures^(0.2),
                            total_space_provided=total_space_provided^(0.2),
                            total_circulation_of_materials=total_circulation_of_materials^(0.2),
                            total_funds=total_funds^(0.2),
                            number_of_resources=number_of_resources^(0.2),
         number_of_visitors=number_of_visitors^(0.2))
```

```
model_trans_full <- lm(number_of_visitors ~ . -Name, data=train_trans)
```

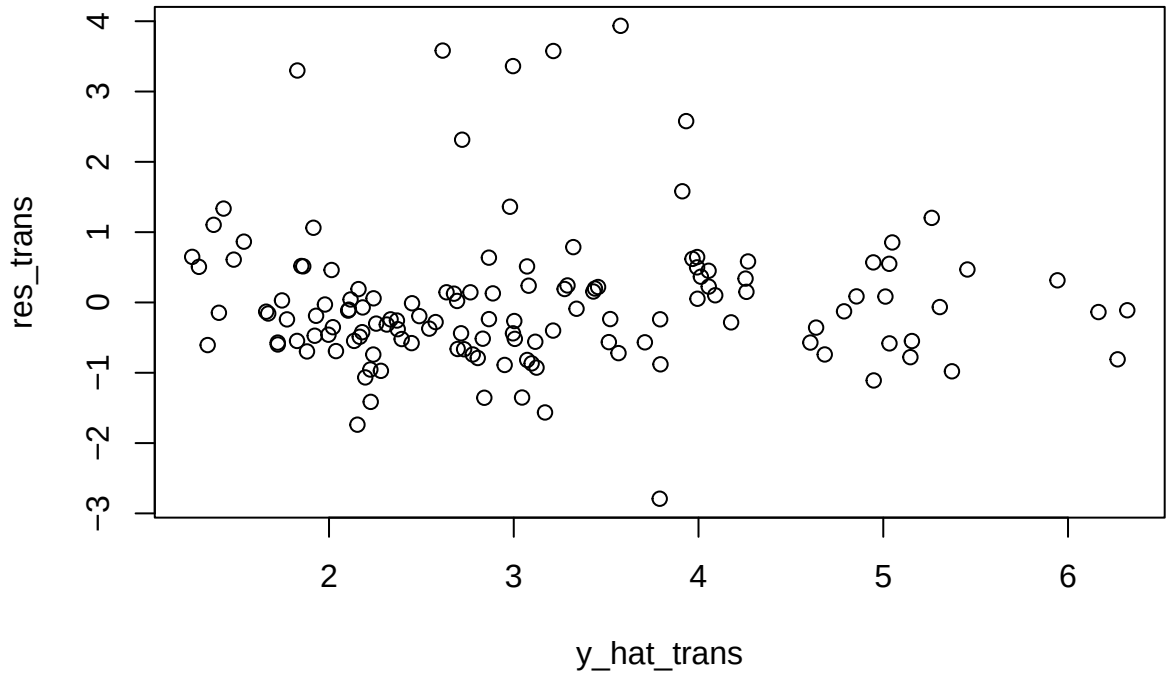
```
response_trans <- fitted(model_trans_full)
fitted_values_trans <- train_trans$number_of_visitors
plot(fitted_values_trans, response_trans)
```



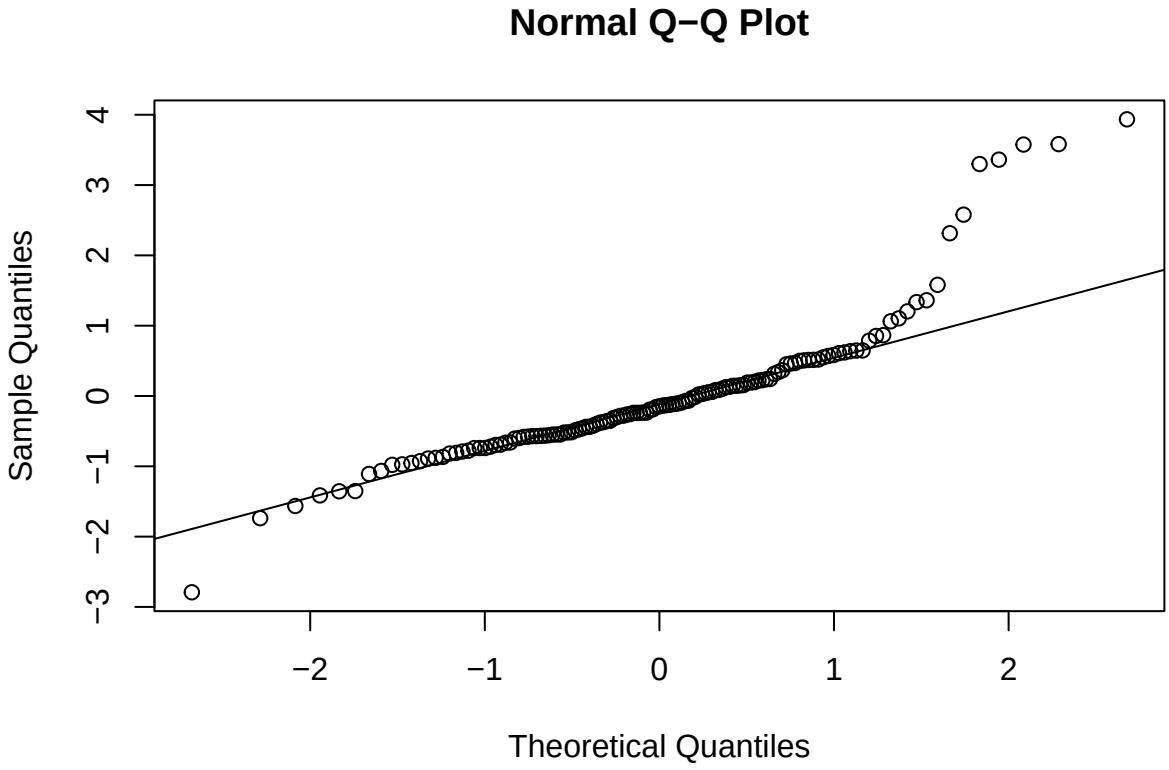

```
pairs(~number_of_cardholders+total_operating_revenues+total_operating_expenditures+total_space_provided
```




```
res_trans<-rstandard(model_trans_full)
y_hat_trans <- fitted(model_trans_full)

plot(y_hat_trans, res_trans)
```




```
qqnorm(res_trans)
qqline(res_trans)
```




```
library(MASS)
```

```
##
## Attaching package: 'MASS'
## The following object is masked from 'package:patchwork':
##
##     area
## The following object is masked from 'package:dplyr':
##
##     select
```

```
stepAIC(model_trans_full, direction="both", k=2)
```

```
## Start:  AIC=-133.53
## number_of_visitors ~ (number_of_cardholders + total_operating_revenues +
##     total_operating_expenditures + total_space_provided + total_circulation_of_materials +
##     total_funds + number_of_resources + region + Name) - Name
##
##                                  Df Sum of Sq    RSS      AIC
## - total_space_provided            1    0.0004 43.942 -135.525
## - number_of_resources             1    0.0004 43.942 -135.525
## - total_operating_expenditures    1    0.0010 43.942 -135.523
## - total_funds                     1    0.0428 43.984 -135.395
## - number_of_cardholders           1    0.2854 44.227 -134.653
## - region                          1    0.4397 44.381 -134.182
## - total_operating_revenues        1    0.4836 44.425 -134.049
## <none>                                       43.941 -133.527
## - total_circulation_of_materials  1   14.3235 58.265  -97.437
##
```

```
## Step:  AIC=-135.53
## number_of_visitors ~ number_of_cardholders + total_operating_revenues +
##     total_operating_expenditures + total_circulation_of_materials +
##     total_funds + number_of_resources + region
##
##                                  Df Sum of Sq    RSS      AIC
## - number_of_resources             1    0.0004 43.942 -137.524
## - total_operating_expenditures    1    0.0011 43.943 -137.522
## - total_funds                     1    0.0427 43.984 -137.394
## - number_of_cardholders           1    0.2851 44.227 -136.652
## - region                          1    0.4420 44.384 -136.174
## - total_operating_revenues        1    0.4966 44.438 -136.008
## <none>                                       43.942 -135.525
## + total_space_provided            1    0.0004 43.941 -133.527
## - total_circulation_of_materials  1   14.3322 58.274  -99.416
##
## Step:  AIC=-137.52
## number_of_visitors ~ number_of_cardholders + total_operating_revenues +
##     total_operating_expenditures + total_circulation_of_materials +
##     total_funds + region
##
##                                  Df Sum of Sq    RSS     AIC
## - total_operating_expenditures    1    0.0029 43.945 -139.51
## - total_funds                     1    0.0529 43.995 -139.36
## - number_of_cardholders           1    0.2859 44.228 -138.65
## - region                          1    0.4540 44.396 -138.14
## - total_operating_revenues        1    0.4971 44.439 -138.00
## <none>                                       43.942 -137.52
## + number_of_resources             1    0.0004 43.942 -135.53
## + total_space_provided            1    0.0004 43.942 -135.53
## - total_circulation_of_materials  1   14.6645 58.607 -100.65
##
## Step:  AIC=-139.52
## number_of_visitors ~ number_of_cardholders + total_operating_revenues +
##     total_circulation_of_materials + total_funds + region
##
##                                  Df Sum of Sq    RSS     AIC
## - total_funds                     1    0.0808 44.026 -141.27
## - number_of_cardholders           1    0.3140 44.259 -140.55
## - region                          1    0.4579 44.403 -140.12
## - total_operating_revenues        1    0.6556 44.601 -139.52
## <none>                                       43.945 -139.51
## + total_operating_expenditures    1    0.0029 43.942 -137.52
## + number_of_resources             1    0.0022 43.943 -137.52
## + total_space_provided            1    0.0004 43.945 -137.52
## - total_circulation_of_materials  1   14.7623 58.707 -102.42
##
## Step:  AIC=-141.27
## number_of_visitors ~ number_of_cardholders + total_operating_revenues +
##     total_circulation_of_materials + region
##
##                                  Df Sum of Sq    RSS     AIC
## - number_of_cardholders           1    0.2668 44.293 -142.45
## - region                          1    0.4810 44.507 -141.80
```

```
## <none>                                        44.026 -141.27
## - total_operating_revenues        1    0.6611 44.687 -141.25
## + total_funds                     1    0.0808 43.945 -139.51
## + total_operating_expenditures    1    0.0308 43.995 -139.36
## + number_of_resources             1    0.0010 44.025 -139.27
## + total_space_provided            1    0.0006 44.025 -139.27
## - total_circulation_of_materials  1   14.9272 58.953 -103.85
##
## Step:  AIC=-142.45
## number_of_visitors ~ total_operating_revenues + total_circulation_of_materials +
##     region
##
##                                 Df Sum of Sq    RSS      AIC
## - region                         1    0.5002 44.793 -142.936
## <none>                                       44.293 -142.452
## + number_of_cardholders          1    0.2668 44.026 -141.267
## + total_funds                    1    0.0336 44.259 -140.554
## + number_of_resources            1    0.0293 44.263 -140.541
## + total_operating_expenditures   1    0.0045 44.288 -140.465
## + total_space_provided           1    0.0002 44.292 -140.452
## - total_operating_revenues       1    1.6329 45.926 -139.564
## - total_circulation_of_materials  1   18.9229 63.216  -96.428
##
## Step:  AIC=-142.94
## number_of_visitors ~ total_operating_revenues + total_circulation_of_materials
##
##                                 Df Sum of Sq    RSS      AIC
## <none>                                       44.793 -142.936
## + region                         1    0.5002 44.293 -142.452
## + number_of_cardholders          1    0.2860 44.507 -141.800
## + total_funds                    1    0.0476 44.745 -141.079
## + number_of_resources            1    0.0093 44.784 -140.964
## + total_space_provided           1    0.0081 44.785 -140.960
## + total_operating_expenditures   1    0.0012 44.792 -140.939
## - total_operating_revenues       1    1.9281 46.721 -139.246
## - total_circulation_of_materials  1   20.0400 64.833  -95.017
##
## Call:
## lm(formula = number_of_visitors ~ total_operating_revenues +
##     total_circulation_of_materials, data = train_trans)
##
## Coefficients:
##                     (Intercept)        total_operating_revenues
##                         0.18090                         0.06453
## total_circulation_of_materials
##                         0.65747
```

```
model_AIC <- lm(number_of_visitors~ total_operating_revenues+total_circulation_of_materials, data=train
```

```
stepAIC(model_trans_full, direction="both", k=log(nrow(train_trans)))
```

```
## Start:  AIC=-107.38
## number_of_visitors ~ (number_of_cardholders + total_operating_revenues +
##     total_operating_expenditures + total_space_provided + total_circulation_of_materials +
```

```
##     total_funds + number_of_resources + region + Name) - Name
##
##                                  Df Sum of Sq    RSS      AIC
## - total_space_provided            1    0.0004 43.942 -112.283
## - number_of_resources             1    0.0004 43.942 -112.283
## - total_operating_expenditures    1    0.0010 43.942 -112.281
## - total_funds                     1    0.0428 43.984 -112.153
## - number_of_cardholders           1    0.2854 44.227 -111.410
## - region                          1    0.4397 44.381 -110.940
## - total_operating_revenues        1    0.4836 44.425 -110.807
## <none>                                       43.941 -107.379
## - total_circulation_of_materials  1   14.3235 58.265  -74.195
##
## Step:  AIC=-112.28
## number_of_visitors ~ number_of_cardholders + total_operating_revenues +
##     total_operating_expenditures + total_circulation_of_materials +
##     total_funds + number_of_resources + region
##
##                                  Df Sum of Sq    RSS      AIC
## - number_of_resources             1    0.0004 43.942 -117.187
## - total_operating_expenditures    1    0.0011 43.943 -117.185
## - total_funds                     1    0.0427 43.984 -117.057
## - number_of_cardholders           1    0.2851 44.227 -116.316
## - region                          1    0.4420 44.384 -115.837
## - total_operating_revenues        1    0.4966 44.438 -115.671
## <none>                                       43.942 -112.283
## + total_space_provided            1    0.0004 43.941 -107.379
## - total_circulation_of_materials  1   14.3322 58.274  -79.079
##
## Step:  AIC=-117.19
## number_of_visitors ~ number_of_cardholders + total_operating_revenues +
##     total_operating_expenditures + total_circulation_of_materials +
##     total_funds + region
##
##                                  Df Sum of Sq    RSS      AIC
## - total_operating_expenditures    1    0.0029 43.945 -122.084
## - total_funds                     1    0.0529 43.995 -121.930
## - number_of_cardholders           1    0.2859 44.228 -121.217
## - region                          1    0.4540 44.396 -120.705
## - total_operating_revenues        1    0.4971 44.439 -120.574
## <none>                                       43.942 -117.187
## + number_of_resources             1    0.0004 43.942 -112.283
## + total_space_provided            1    0.0004 43.942 -112.283
## - total_circulation_of_materials  1   14.6645 58.607  -83.216
##
## Step:  AIC=-122.08
## number_of_visitors ~ number_of_cardholders + total_operating_revenues +
##     total_circulation_of_materials + total_funds + region
##
##                                  Df Sum of Sq    RSS      AIC
## - total_funds                     1    0.0808 44.026 -126.741
## - number_of_cardholders           1    0.3140 44.259 -126.028
## - region                          1    0.4579 44.403 -125.590
## - total_operating_revenues        1    0.6556 44.601 -124.990
```

```
## <none>                                       43.945 -122.084
## + total_operating_expenditures    1    0.0029 43.942 -117.187
## + number_of_resources             1    0.0022 43.943 -117.185
## + total_space_provided            1    0.0004 43.945 -117.180
## - total_circulation_of_materials  1   14.7623 58.707  -87.889
##
## Step:  AIC=-126.74
## number_of_visitors ~ number_of_cardholders + total_operating_revenues +
##     total_circulation_of_materials + region
##
##                                 Df Sum of Sq    RSS      AIC
## - number_of_cardholders          1    0.2668 44.293 -130.830
## - region                         1    0.4810 44.507 -130.179
## - total_operating_revenues       1    0.6611 44.687 -129.634
## <none>                                       44.026 -126.741
## + total_funds                    1    0.0808 43.945 -122.084
## + total_operating_expenditures   1    0.0308 43.995 -121.930
## + number_of_resources            1    0.0010 44.025 -121.839
## + total_space_provided           1    0.0006 44.025 -121.837
## - total_circulation_of_materials  1   14.9272 58.953  -92.231
##
## Step:  AIC=-130.83
## number_of_visitors ~ total_operating_revenues + total_circulation_of_materials +
##     region
##
##                                 Df Sum of Sq    RSS      AIC
## - region                         1    0.5002 44.793 -134.220
## - total_operating_revenues       1    1.6329 45.926 -130.848
## <none>                                       44.293 -130.830
## + number_of_cardholders          1    0.2668 44.026 -126.741
## + total_funds                    1    0.0336 44.259 -126.028
## + number_of_resources            1    0.0293 44.263 -126.015
## + total_operating_expenditures   1    0.0045 44.288 -125.939
## + total_space_provided           1    0.0002 44.292 -125.926
## - total_circulation_of_materials  1   18.9229 63.216  -87.712
##
## Step:  AIC=-134.22
## number_of_visitors ~ total_operating_revenues + total_circulation_of_materials
##
##                                 Df Sum of Sq    RSS      AIC
## <none>                                       44.793 -134.220
## - total_operating_revenues       1    1.9281 46.721 -133.436
## + region                         1    0.5002 44.293 -130.830
## + number_of_cardholders          1    0.2860 44.507 -130.179
## + total_funds                    1    0.0476 44.745 -129.458
## + number_of_resources            1    0.0093 44.784 -129.342
## + total_space_provided           1    0.0081 44.785 -129.339
## + total_operating_expenditures   1    0.0012 44.792 -129.318
## - total_circulation_of_materials  1   20.0400 64.833  -89.207
##
## Call:
## lm(formula = number_of_visitors ~ total_operating_revenues +
##     total_circulation_of_materials, data = train_trans)
```

```
##
## Coefficients:
##                     (Intercept)        total_operating_revenues
##                         0.18090                         0.06453
## total_circulation_of_materials
##                         0.65747
```

```
summary(model_trans_full)
```

```
##
## Call:
## lm(formula = number_of_visitors ~ . - Name, data = train_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63010 -0.31866 -0.08447  0.18876  2.29236
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    0.156930   0.168538   0.931    0.354
## number_of_cardholders          0.082702   0.091424   0.905    0.367
## total_operating_revenues       0.060697   0.051547   1.178    0.241
## total_operating_expenditures  -0.004049   0.074924  -0.054    0.957
## total_space_provided          -0.001076   0.033063  -0.033    0.974
## total_circulation_of_materials  0.613188   0.095680   6.409 2.67e-09 ***
## total_funds                   -0.015113   0.043145  -0.350    0.727
## number_of_resources           -0.001634   0.046842  -0.035    0.972
## regionSouth                    0.139012   0.123797   1.123    0.264
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5905 on 126 degrees of freedom
## Multiple R-squared:  0.8128, Adjusted R-squared:  0.8009
## F-statistic: 68.39 on 8 and 126 DF,  p-value: < 2.2e-16
```

```
model_BIC <- lm(number_of_visitors~total_circulation_of_materials, data=train_trans)
```

```
test_trans <- test %>%
  mutate(number_of_cardholders=number_of_cardholders^(0.2),
                            total_operating_revenues=total_operating_revenues^(0.2),
                            total_operating_expenditures=total_operating_expenditures^(0.2),
                            total_space_provided=total_space_provided^(0.2),
                            total_circulation_of_materials=total_circulation_of_materials^(0.2),
                            total_funds=total_funds^(0.2),
                            number_of_resources=number_of_resources^(0.2),
         number_of_visitors=number_of_visitors^(0.2))
```

```
model_AIC_test <- lm(number_of_visitors~ total_operating_revenues+total_circulation_of_materials,  data=
```

```
model_BIC_test <- lm(number_of_visitors~ total_circulation_of_materials,  data=test_trans)
```

```
summary(model_AIC)
```

```
##
## Call:
## lm(formula = number_of_visitors ~ total_operating_revenues +
```

```
##     total_circulation_of_materials, data = train_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54986 -0.31061 -0.09012  0.19942  2.28737
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    0.18090    0.14654   1.234   0.2192
## total_operating_revenues       0.06453    0.02707   2.384   0.0186 *
## total_circulation_of_materials  0.65747    0.08555   7.685 3.09e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5825 on 132 degrees of freedom
## Multiple R-squared:  0.8092, Adjusted R-squared:  0.8063
## F-statistic: 279.9 on 2 and 132 DF,  p-value: < 2.2e-16
```

```
AIC(model_AIC)
```

```
## [1] 242.1778
```

```
BIC(model_AIC)
```

```
## [1] 253.7989
```

```
summary(model_AIC_test)
```

```
##
## Call:
## lm(formula = number_of_visitors ~ total_operating_revenues +
##     total_circulation_of_materials, data = test_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92554 -0.36022 -0.05314  0.30239  2.40445
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -0.006809   0.174171  -0.039 0.968877
## total_operating_revenues      0.179804   0.027438   6.553 1.18e-09 ***
## total_circulation_of_materials  0.276948   0.076291   3.630 0.000405 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6387 on 131 degrees of freedom
## Multiple R-squared:  0.7803, Adjusted R-squared:  0.7769
## F-statistic: 232.6 on 2 and 131 DF,  p-value: < 2.2e-16
```

```
AIC(model_AIC_test)
```

```
## [1] 265.071
```

```
BIC(model_AIC_test)
```

```
## [1] 276.6624
```

```
# Compare leverage test with AIC

h_AIC <- hatvalues(model_AIC)
leverage <- 2*(length(model_AIC$coefficients)/nrow(train_trans))
which(h_AIC>leverage)
```

```
##  10  30  36  55  63  66  69  75 108 110
##  10  30  36  55  63  66  69  75 108 110
```

```
h_AIC_test <- hatvalues(model_AIC_test)
leverage_test <- 2*(length(model_AIC_test$coefficients)/nrow(test_trans))
which(h_AIC_test>leverage_test)
```

```
##   2  16  40  45  65  70 128 130
##   2  16  40  45  65  70 128 130
```

```
# Compare outlier test with AIC
outlier_AIC <- rstandard(model_AIC)
which (abs(outlier_AIC)>4)
```

```
## named integer(0)
```

```
outlier_AIC_test <- rstandard(model_AIC_test)
which (abs(outlier_AIC_test)>4)
```

```
## named integer(0)
```

```
# Compare influential test with AIC
influential_AIC <- cooks.distance(model_AIC)
cutoff_AIC <- qf(0.5, length(model_AIC$coefficients),
                 nrow(train_trans)-length(model_AIC$coefficients))
which(influential_AIC> cutoff_AIC)
```

```
## named integer(0)
```

```
influential_AIC_test <- cooks.distance(model_AIC_test)
cutoff_AIC_test <- qf(0.5, length(model_AIC_test$coefficients),
                 nrow(test_trans)-length(model_AIC_test$coefficients))
which(influential_AIC_test> cutoff_AIC_test)
```

```
## named integer(0)
```

```
summary(model_BIC)
```

```
##
## Call:
## lm(formula = number_of_visitors ~ total_circulation_of_materials,
##     data = train_trans)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6141 -0.3228 -0.1071  0.2023  2.2212
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    0.35514    0.12922   2.748  0.00682 **
## total_circulation_of_materials  0.84269    0.03643  23.135  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5927 on 133 degrees of freedom
## Multiple R-squared:  0.801,  Adjusted R-squared:  0.7995
## F-statistic: 535.2 on 1 and 133 DF,  p-value: < 2.2e-16
```

```
AIC(model_BIC)
```

```
## [1] 245.8672
```

```
BIC(model_BIC)
```

```
## [1] 254.583
```

```
summary(model_BIC_test)
```

```
##
## Call:
## lm(formula = number_of_visitors ~ total_circulation_of_materials,
##     data = test_trans)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1990 -0.4160 -0.0064  0.4082  2.0889
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    0.69509    0.15766   4.409 2.14e-05 ***
## total_circulation_of_materials  0.72089    0.04027  17.900  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7331 on 132 degrees of freedom
## Multiple R-squared:  0.7082, Adjusted R-squared:  0.706
## F-statistic: 320.4 on 1 and 132 DF,  p-value: < 2.2e-16
```

```
AIC(model_BIC_test)
```

```
## [1] 301.0642
```

```
BIC(model_BIC_test)
```

```
## [1] 309.7577
```

```
# Compare leverage test with BIC
h_BIC <- hatvalues(model_BIC)
leverage_BIC <- 2*(length(model_BIC$coefficients)/nrow(train_trans))
which(h_BIC>leverage_BIC)
```

```
##   5  32  36  55  64  66  98 108
##   5  32  36  55  64  66  98 108
```

```
h_BIC_test <- hatvalues(model_BIC_test)
leverage_BIC_test <- 2*(length(model_BIC_test$coefficients)/nrow(test_trans))
which(h_BIC_test>leverage_test)
```

```
##  16  40 130
##  16  40 130
```

```r
# Compare outlier test with BIC
outlier_BIC <- rstandard(model_BIC)
which (abs(outlier_BIC)>4)
```

```
## named integer(0)
```

```r
outlier_BIC_test <- rstandard(model_BIC_test)
which (abs(outlier_BIC_test)>4)
```

```
## 130
## 130
```

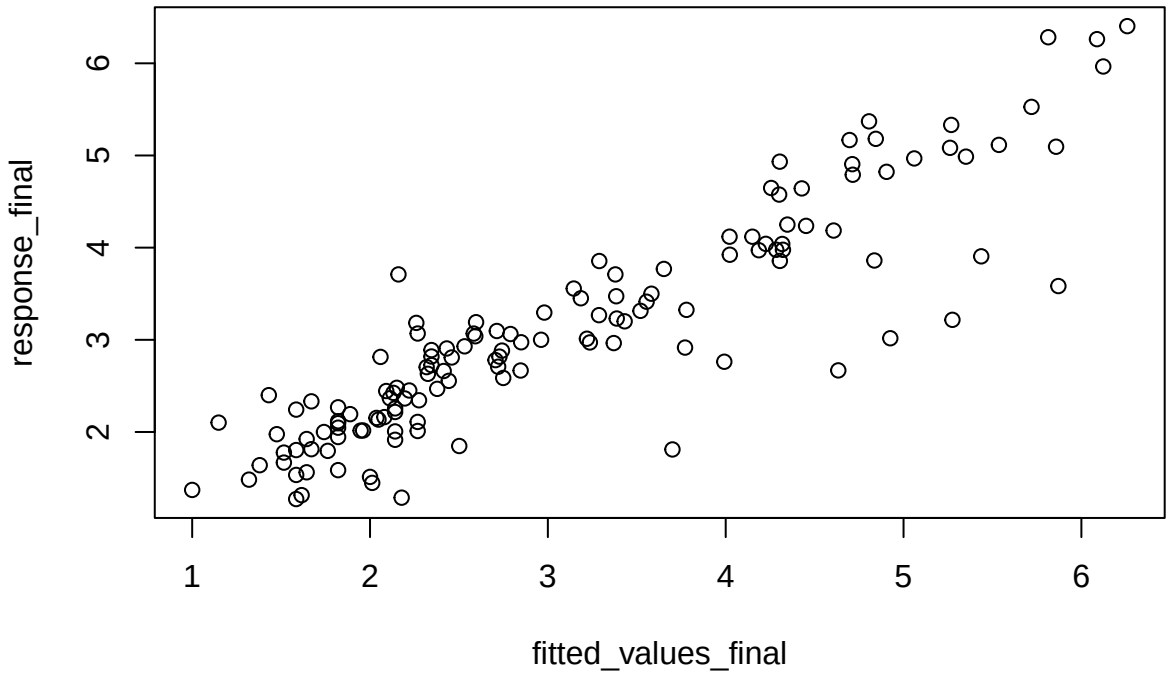
```r
# Compare influential test with BIC
influential_BIC <- cooks.distance(model_BIC)
cutoff_BIC <- qf(0.5, length(model_BIC$coefficients),
                 nrow(train_trans)-length(model_BIC$coefficients))
which(influential_BIC> cutoff_BIC)
```

```
## named integer(0)
```

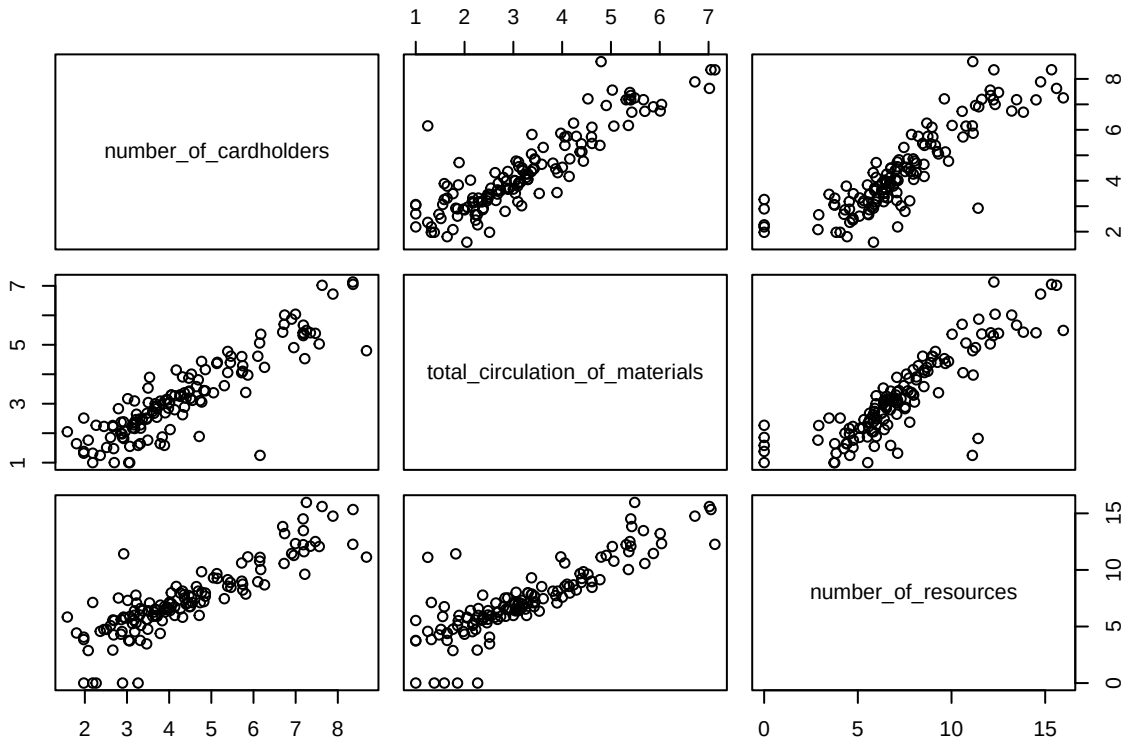
```r
influential_BIC_test <- cooks.distance(model_BIC_test)
cutoff_BIC_test <- qf(0.5, length(model_BIC_test$coefficients),
                 nrow(test_trans)-length(model_BIC_test$coefficients))
which(influential_BIC_test> cutoff_BIC_test)
```

```
##  40 130
##  40 130
```

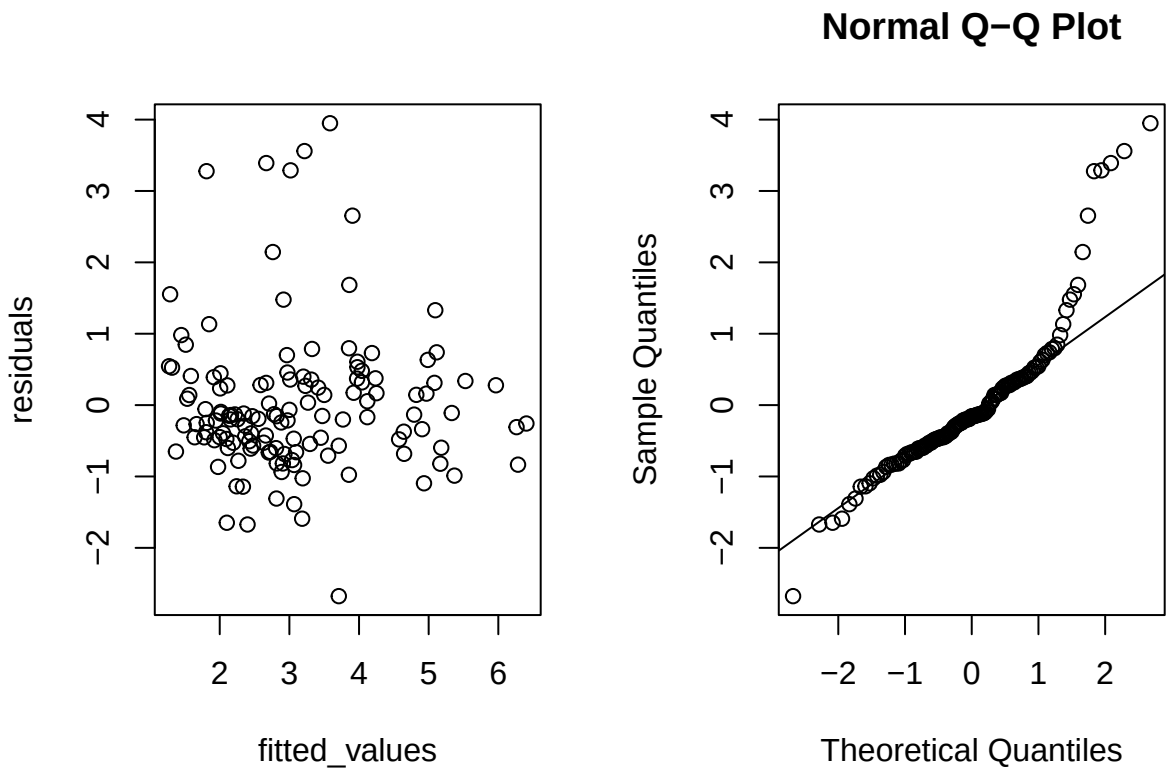
```r
#Recheck condition1
response_final <- fitted(model_AIC)
fitted_values_final <- train_trans$number_of_visitors
plot(fitted_values_final, response_final)
```




```r
#Recheck condition2
pairs(~number_of_cardholders+total_circulation_of_materials+number_of_resources, data=train_trans)
```




```
residuals<-rstandard(model_AIC)
fitted_values <- fitted(model_AIC)
par(mfrow=c(1,2))
plot(fitted_values, residuals)
qqnorm(residuals)
qqline(residuals)
```



```
vif(model_AIC)
```

```
##       total_operating_revenues total_circulation_of_materials
##                       5.710922                       5.710922
```

```
anova(model_AIC, model_trans_full)
```

```
## Analysis of Variance Table
##
## Model 1: number_of_visitors ~ total_operating_revenues + total_circulation_of_materials
## Model 2: number_of_visitors ~ (number_of_cardholders + total_operating_revenues +
##     total_operating_expenditures + total_space_provided + total_circulation_of_materials +
##     total_funds + number_of_resources + region + Name) - Name
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    132 44.793
## 2    126 43.941  6    0.8515 0.4069 0.8733
```