final report

Charlotte Fowler

12/6/2019

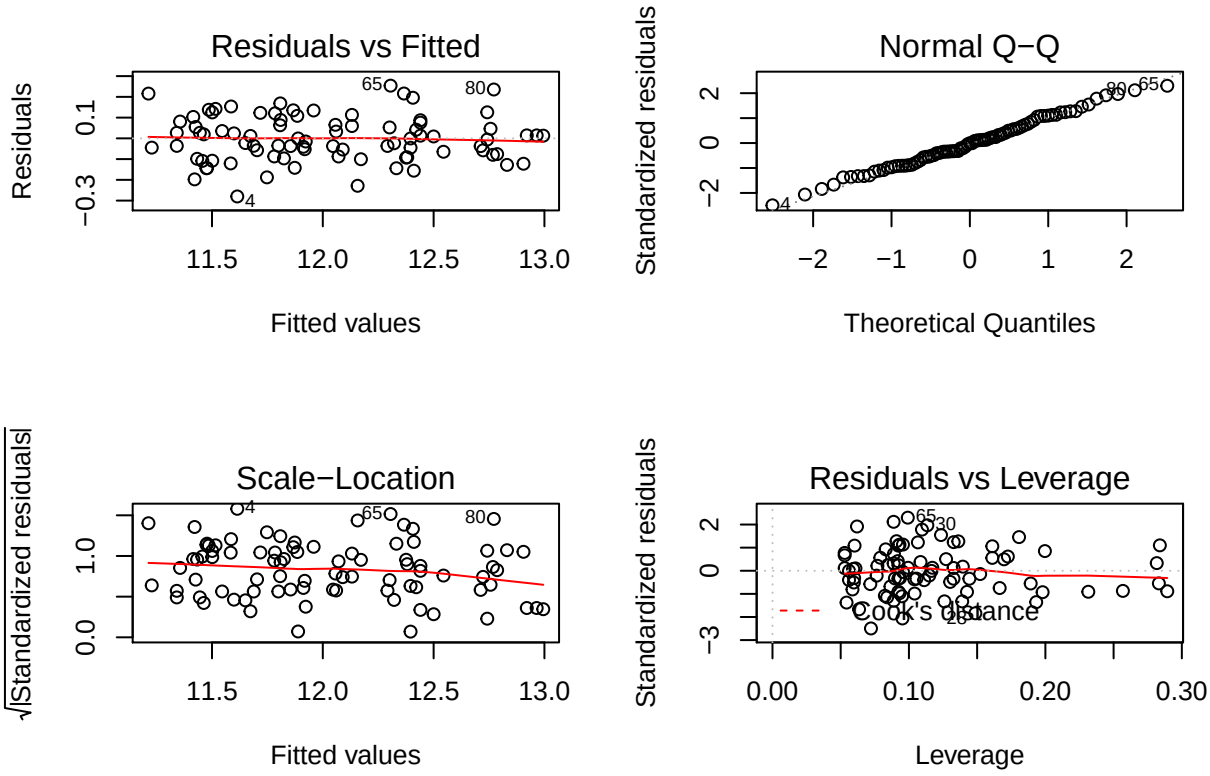
```
library(tidyverse)
library(faraway)
#library(car)
library(base)
```

```
lawsuit_full_prof =  read_csv("./Lawsuit.csv") %>%
  janitor::clean_names() %>%
  mutate(
    sal_avg = (sal94 + sal95)/2,
    ln_sal_avg = log(sal_avg),
    gender = recode(gender, "0" = "female", "1" = "male"),
    clin = recode(clin, "0" = "research", "1" = "clinical"),
    cert = recode(cert, "0" = "not certified", "1" = "certified"),
    rank = recode(rank, "1" = "assistant", "2" = "associate", "3" = "full professor"),
    dept = recode(dept, "1" = "biochemistry", "2" = "physiology", "3" = "genetics", "4" = "pediatrics",
  filter(rank == "full professor") %>%
  select(-c(sal94, sal95, rank, id, sal_avg))



lawsuit_associate =  read_csv("./Lawsuit.csv") %>%
  janitor::clean_names() %>%
  mutate(
    sal_avg = (sal94 + sal95)/2,
    ln_sal_avg = log(sal_avg),
    gender = recode(gender, "0" = "female", "1" = "male"),
    clin = recode(clin, "0" = "research", "1" = "clinical"),
    cert = recode(cert, "0" = "not certified", "1" = "certified"),
    rank = recode(rank, "1" = "assistant", "2" = "associate", "3" = "full professor"),
    dept = recode(dept, "1" = "biochemistry", "2" = "physiology", "3" = "genetics", "4" = "pediatrics",
  filter(rank == "associate") %>%
  select(-c(sal94, sal95, rank, id, sal_avg))


lawsuit_assistant =  read_csv("./Lawsuit.csv") %>%
  janitor::clean_names() %>%
  mutate(
    sal_avg = (sal94 + sal95)/2,
    ln_sal_avg = log(sal_avg),
    gender = recode(gender, "0" = "female", "1" = "male"),
    clin = recode(clin, "0" = "research", "1" = "clinical"),
    cert = recode(cert, "0" = "not certified", "1" = "certified"),
    rank = recode(rank, "1" = "assistant", "2" = "associate", "3" = "full professor"),
    dept = recode(dept, "1" = "biochemistry", "2" = "physiology", "3" = "genetics", "4" = "pediatrics",
  filter(rank == "assistant") %>%
  select(-c(sal94, sal95, rank, id, sal_avg))
```

```r
full_prof_mod = lm(ln_sal_avg ~ . - prate , lawsuit_full_prof)

summary(full_prof_mod)
```

```
##
## Call:
## lm(formula = ln_sal_avg ~ . - prate, data = lawsuit_full_prof)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.279999 -0.079443 -0.000564  0.075727  0.253899
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       11.685072   0.056901 205.359  < 2e-16 ***
## deptgenetics       0.248823   0.055205   4.507 2.38e-05 ***
## deptmedicine       0.528191   0.039209  13.471  < 2e-16 ***
## deptpediatrics     0.166793   0.066371   2.513  0.01412 *
## deptphysiology    -0.127860   0.038834  -3.292  0.00152 **
## deptsurgery        0.948635   0.050331  18.848  < 2e-16 ***
## gendermale        -0.040413   0.036217  -1.116  0.26805
## clinresearch      -0.179305   0.033412  -5.367 8.61e-07 ***
## certnot certified -0.258010   0.033989  -7.591 7.14e-11 ***
## exper              0.014871   0.002253   6.601 5.15e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1165 on 75 degrees of freedom
## Multiple R-squared:  0.9511, Adjusted R-squared:  0.9452
## F-statistic: 161.9 on 9 and 75 DF,  p-value: < 2.2e-16
```

```r
## Model diagnosis
par(mfrow=c(2,2))
plot(full_prof_mod)
```




```
shapiro.test(residuals(full_prof_mod))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(full_prof_mod)
## W = 0.99255, p-value = 0.9139
```

```
vif(full_prof_mod)
```

```
##      deptgenetics      deptmedicine    deptpediatrics    deptphysiology
##          1.443299          1.901519          1.237962          1.512136
##       deptsurgery        gendermale      clinresearch certnot certified
##          1.924784          1.256037          1.708077          1.302690
##             exper
##          1.114839
```

```
rstandard(full_prof_mod)
```

```
##           1           2           3           4           5
## -0.326441209 -0.343848228  0.727681978 -2.495992867  1.545331869
##           6           7           8           9          10
## -1.296953419 -0.368238924  0.240611881  0.240941956  0.214132121
##          11          12          13          14          15
##  0.102465249  1.228071110  0.926552858  1.087928475  0.564648247
##          16          17          18          19          20
## -1.083122030 -1.667872485 -0.507319057 -0.317117418  1.272574865
```

```
##           21           22           23           24           25
## -0.141285172  0.502710609 -0.206780106 -0.985986029 -0.005155537
##           26           27           28           29           30
## -0.971808651 -0.482727081 -1.839771755  1.127958068  1.967112458
##           31           32           33           34           35
##  1.279553349  1.086975572 -0.317402536 -0.407613846  0.317424600
##           36           37           38           39           40
##  1.456713927 -0.914134797 -1.325780731  0.178341677 -0.871546222
##           41           42           43           44           45
## -1.358305486  0.614845678  1.059823417  0.556638842  0.851648495
##           46           47           48           49           50
## -0.927243613  1.099746252 -0.887636495 -0.547133208  0.331397037
##           51           52           53           54           55
## -2.064139560  0.113454081 -0.906662801 -0.808717241 -0.211373131
##           56           57           58           59           60
## -0.004890097  1.237910900  1.915847019 -0.329878470  0.080598765
##           61           62           63           64           65
##  0.668146708 -0.551759270 -0.345190791 -0.576501131  2.296963391
##           66           67           68           69           70
##  0.771001637  0.494804526 -1.376239058 -0.398698318  0.378824548
##           71           72           73           74           75
##  1.780466163 -0.908972254 -1.323822777 -1.147378871 -0.681290220
##           76           77           78           79           80
## -1.104558014  0.417139574  0.129764061  0.119304036  2.117873129
##           81           82           83           84           85
##  1.133396899 -0.339623242  0.129856954 -0.052433531 -0.747168963
```

```
brand_hat =hatvalues(full_prof_mod)
brand_hat[brand_hat > 0.24]
```

```
##        40        47        48        49        50
## 0.2568140 0.2838955 0.2893311 0.2831800 0.2816270
```

```
# check influencial ponits

which(cooks.distance(full_prof_mod) > 0.5)
```

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

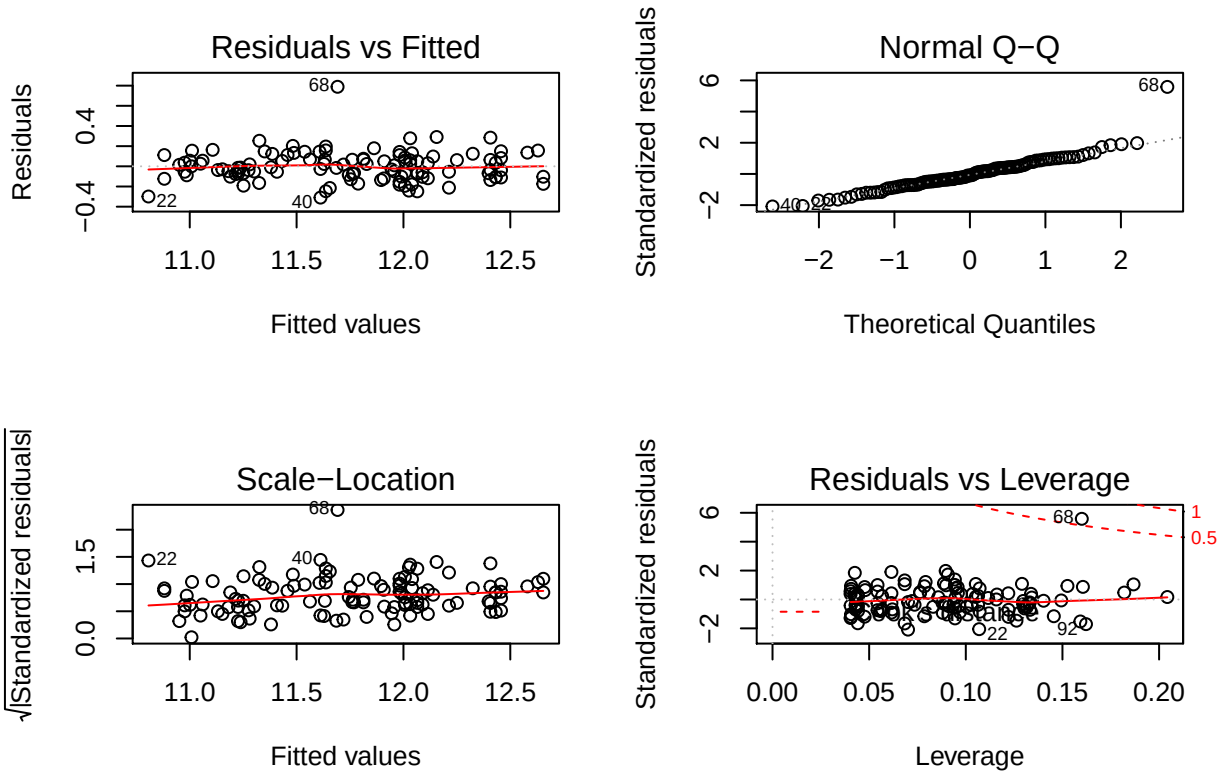
The residual has constant variance and the the residuals are normally distributed. Since the VIF is less than 5, we do not have issue with multi collinearity.

```
assistant_mod = lm(ln_sal_avg ~ . - prate , lawsuit_assistant)

summary(assistant_mod)
```

```
##
## Call:
## lm(formula = ln_sal_avg ~ . - prate, data = lawsuit_assistant)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.30995 -0.09230 -0.01370  0.07692  0.78854
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       11.257647   0.075964 148.198  < 2e-16 ***
## deptgenetics       0.143432   0.069186   2.073 0.040681 *
## deptmedicine       0.600806   0.061033   9.844  < 2e-16 ***
## deptpediatrics     0.255421   0.066715   3.829 0.000223 ***
## deptphysiology    -0.201314   0.063379  -3.176 0.001973 **
## deptsurgery        0.943082   0.070044  13.464  < 2e-16 ***
## gendermale         0.082656   0.035347   2.338 0.021316 *
## clinresearch      -0.179071   0.042166  -4.247 4.80e-05 ***
## certnot certified -0.119828   0.040890  -2.931 0.004176 **
## exper              0.024735   0.005354   4.620 1.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1541 on 102 degrees of freedom
## Multiple R-squared:  0.9126, Adjusted R-squared:  0.9049
## F-statistic: 118.3 on 9 and 102 DF,  p-value: < 2.2e-16
```

```r
## Model diagnosis
par(mfrow=c(2,2))
plot(assistant_mod)
```




```r
shapiro.test(residuals(assistant_mod))
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  residuals(assistant_mod)
## W = 0.92399, p-value = 8.107e-06
```

```
vif(assistant_mod)
```

```
##      deptgenetics      deptmedicine    deptpediatrics    deptphysiology
##          1.667507          4.031506          2.701288          1.942748
##       deptsurgery        gendermale      clinresearch certnot certified
##          3.257863          1.393034          1.828098          1.512339
##             exper
##          1.259207
```

```
rstandard(assistant_mod)
```

```
##             1             2             3             4             5
## -0.7167198938 -0.5306082101  0.0004811647 -1.3088313320 -0.2497300745
##             6             7             8             9            10
## -0.2033206579  1.0060978799  1.7248150375  0.1701821442 -0.3281385341
##            11            12            13            14            15
##  1.1114622631  0.1331774002 -0.4839636265 -0.3488720622  0.8723584032
##            16            17            18            19            20
## -0.0890942144  0.2479894645  1.0801696414  0.1769404683 -0.3705594172
##            21            22            23            24            25
##  0.7659030041 -2.0543568029 -0.6195788138  0.3822828224  0.1020856892
##            26            27            28            29            30
##  0.3899605908 -0.8485645108  1.0440544823  0.9379447511  0.4779199665
##            31            32            33            34            35
## -1.1590651223 -0.4799528367 -0.1019878246 -0.0664968952 -0.2675650572
##            36            37            38            39            40
## -0.3349679068 -0.5812431624  0.1208715019 -0.4405917685 -2.0852449469
##            41            42            43            44            45
##  1.3805772710 -0.1794008503  1.3194404376  0.4602074836  0.3636722546
##            46            47            48            49            50
## -1.6571225473  1.0597658315 -1.5212124775  0.7669420050 -0.1064126999
##            51            52            53            54            55
## -0.3707496996  0.4469504358  0.9821612399 -0.3635371353 -1.1906003870
##            56            57            58            59            60
##  0.2028331666 -0.6933293440 -0.4278309923 -0.6834605170  0.8610567986
##            61            62            63            64            65
## -1.6462776000 -0.3902000903  0.3432389370 -0.4965573806 -0.9271389767
##            66            67            68            69            70
## -0.7865320891  0.3353708537  5.5817508977 -0.2134677714  0.6467763558
##            71            72            73            74            75
##  0.4410714940  0.9458962135 -0.8005866525 -1.2054496302  0.0659797879
##            76            77            78            79            80
## -0.3335332584  0.3974453292  0.1763679480 -0.5076687923  1.2124538288
##            81            82            83            84            85
##  1.8426409586  0.1568456606  0.7353218320  0.3796609289  1.9768091891
##            86            87            88            89            90
## -0.4239957550 -1.2827617388  0.5070862573  0.5722422435 -1.1616269854
##            91            92            93            94            95
```

```
## -1.0190112855 -1.7067054146 -0.8751908272 -0.7132566375  1.9046185316
##            96            97            98            99           100
## -0.9086263222  0.4249646461 -0.7229009804 -0.2631723234  0.9988652929
##           101           102           103           104           105
##  0.9175454338 -0.7386834222  0.5493145778  1.0737354006  0.2363760922
##           106           107           108           109           110
## -1.2040136663 -1.4685938459 -0.7157251586 -0.2411194106 -0.4658657092
##           111           112
##  0.4280285854  0.8508846576
```

```
brand_hat =hatvalues(assistant_mod)
brand_hat[brand_hat > 0.18]
```

```
##         9        28        30
## 0.2042631 0.1867842 0.1819880
```

```
# check influencial ponits
which(cooks.distance(assistant_mod) > 0.5)
```

```
## 68
## 68
```

```
lawsuit_assistant1 = lawsuit_assistant [-c(68),]
assistant_mod1 = lm(ln_sal_avg ~ . - prate , lawsuit_assistant1)
summary(assistant_mod1)
```

```
##
## Call:
## lm(formula = ln_sal_avg ~ . - prate, data = lawsuit_assistant1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33395 -0.07720 -0.01032  0.08965  0.29485
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       11.328007   0.064490 175.654  < 2e-16 ***
## deptgenetics       0.143838   0.057944   2.482 0.014700 *
## deptmedicine       0.529024   0.052238  10.127  < 2e-16 ***
## deptpediatrics     0.199035   0.056512   3.522 0.000645 ***
## deptphysiology    -0.215107   0.053121  -4.049 0.000101 ***
## deptsurgery        0.893373   0.059135  15.107  < 2e-16 ***
## gendermale         0.039030   0.030318   1.287 0.200920
## clinresearch      -0.246526   0.036737  -6.711 1.14e-09 ***
## certnot certified -0.162553   0.034840  -4.666 9.46e-06 ***
## exper              0.027242   0.004499   6.054 2.42e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1291 on 101 degrees of freedom
## Multiple R-squared:  0.9381, Adjusted R-squared:  0.9326
## F-statistic:   170 on 9 and 101 DF,  p-value: < 2.2e-16
```

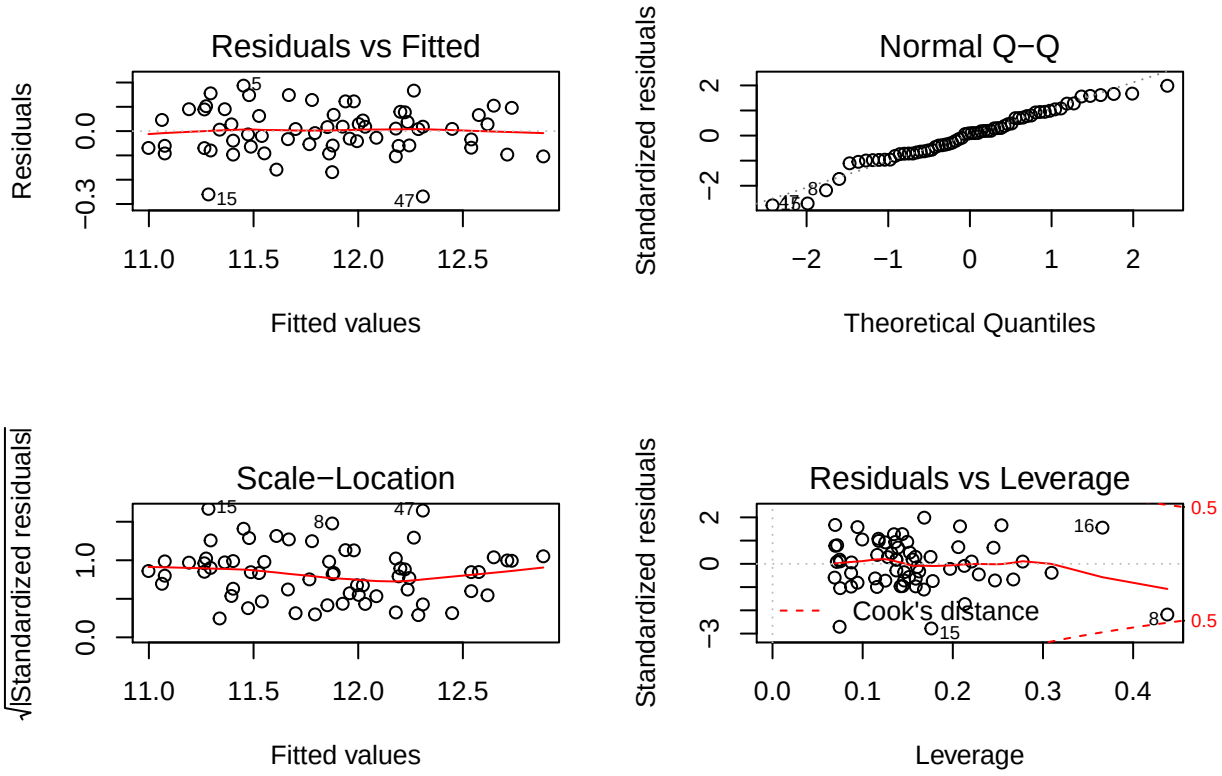
The residual has constant variance and the the residuals are normally distributed. Since the VIF is less than 5, we do not have issue with multi collinearity.

Observation 68 is an influencial point, after removing this point adj r2 increase from 90% to 93%. The coefficients of department of medicine, pediatrics, gender male and clinical research change more than 10%.

```
associate_mod = lm(ln_sal_avg ~ . - prate , lawsuit_associate)

summary(associate_mod)
```

```
##
## Call:
## lm(formula = ln_sal_avg ~ . - prate, data = lawsuit_associate)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.268895 -0.061719  0.008443  0.069568  0.186993
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       11.513803   0.057135 201.518  < 2e-16 ***
## deptgenetics       0.170588   0.055740   3.060 0.003439 **
## deptmedicine       0.507098   0.050051  10.132 4.30e-14 ***
## deptpediatrics     0.210069   0.055921   3.757 0.000424 ***
## deptphysiology    -0.189342   0.043481  -4.355 5.99e-05 ***
## deptsurgery        0.931900   0.057099  16.321  < 2e-16 ***
## gendermale        -0.013277   0.031011  -0.428 0.670252
## clinresearch      -0.220247   0.037705  -5.841 3.06e-07 ***
## certnot certified -0.200488   0.031803  -6.304 5.53e-08 ***
## exper              0.021512   0.002619   8.214 4.45e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1033 on 54 degrees of freedom
## Multiple R-squared:  0.9621, Adjusted R-squared:  0.9558
## F-statistic: 152.2 on 9 and 54 DF,  p-value: < 2.2e-16
```

```
## Model diagnosis
par(mfrow=c(2,2))
plot(associate_mod)
```




```
shapiro.test(residuals(associate_mod))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(associate_mod)
## W = 0.97494, p-value = 0.2174
```

```
vif(associate_mod)
```

```
##      deptgenetics      deptmedicine    deptpediatrics    deptphysiology
##          1.342806          2.932484          2.267822          1.495755
##       deptsurgery        gendermale      clinresearch certnot certified
##          2.364445          1.272280          2.114114          1.422567
##             exper
##          1.167069
```

```
rstandard(associate_mod)
```

```
##           1           2           3           4           5           6
## -0.81167841 -0.14321846  1.04585576  0.06007298  1.98549440 -0.97954641
##           7           8           9          10          11          12
##  1.58129451 -2.18145793 -0.39788057 -0.95872228  0.28789106 -0.72366525
##          13          14          15          16          17          18
##  0.92292654 -0.08851320 -2.78039978  1.55702444  0.93912909 -0.74035021
##          19          20          21          22          23          24
##  0.48255785 -0.21674354 -0.63935514  0.94617968 -0.97109320  1.61411614
```

```
##          25          26          27          28          29          30
## -0.71532110 -0.38535690 -0.67175672  0.09802496  1.65887191  0.18980407
##          31          32          33          34          35          36
##  0.18799276  0.69877412 -0.96301942  0.69538373 -0.56559420  0.17335605
##          37          38          39          40          41          42
##  1.28039901 -1.72977869 -0.28509314  0.79920173  0.08320143  0.77717182
##          43          44          45          46          47          48
## -0.59248122  0.10439777  1.67268186 -1.04635134 -2.70645136  0.18256564
##          49          50          51          52          53          54
##  1.27399378 -0.32748867  0.30317644  0.45449841 -0.62664958  0.38436222
##          55          56          57          58          59          60
## -0.45683103 -1.10540146 -0.71571356  0.09857326 -0.99494646  1.07847098
##          61          62          63          64
## -0.35921612  0.98679092  0.71794415  0.29790460
```

```
brand_hat =hatvalues(associate_mod)
brand_hat[brand_hat > 0.3125]
```

```
##         8        16
## 0.4379014 0.3659291
```

```
# check influencial ponits
which(cooks.distance(associate_mod) > 0.5)
```

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

The residual has constant variance and the the residuals are normally distributed. Since the VIF is less than 5, we do not have issue with multi collinearity.