HW9 - Multiple Regression

Madhu Peduri

July 11, 2021

0.0.1 Problem 1

```
# Read the data
senic <- read.delim("SENIC.txt", header = FALSE, sep = "")
names(senic) <- c("id", "stay", "age", "inf_risk", "cul_r", "xray_r",
    "beds", "med", "region", "daily_avg", "nurses", "services")
head(senic)
```

```
##   id  stay  age inf_risk cul_r xray_r beds med region daily_avg nurses services
## 1  1  7.13 55.7      4.1   9.0   39.6  279   2      4       207    241       60
## 2  2  8.82 58.2      1.6   3.8   51.7   80   2      2        51     52       40
## 3  3  8.34 56.9      2.7   8.1   74.0  107   2      3        82     54       20
## 4  4  8.95 53.7      5.6  18.9  122.8  147   2      4        53    148       40
## 5  5 11.20 56.5      5.7  34.5   88.9  180   2      1       134    151       40
## 6  6  9.76 50.9      5.1  21.9   97.0  150   2      2       147    106       40
```

```
summary(senic)
```

```
##        id           stay            age          inf_risk        cul_r
##  Min.   :  1   Min.   : 6.70   Min.   :38.8   Min.   :1.30   Min.   : 1.6
##  1st Qu.: 29   1st Qu.: 8.34   1st Qu.:50.9   1st Qu.:3.70   1st Qu.: 8.4
##  Median : 57   Median : 9.42   Median :53.2   Median :4.40   Median :14.1
##  Mean   : 57   Mean   : 9.65   Mean   :53.2   Mean   :4.35   Mean   :15.8
##  3rd Qu.: 85   3rd Qu.:10.47   3rd Qu.:56.2   3rd Qu.:5.20   3rd Qu.:20.3
##  Max.   :113   Max.   :19.56   Max.   :65.9   Max.   :7.80   Max.   :60.5
##      xray_r           beds          med           region       daily_avg
##  Min.   : 39.6   Min.   : 29   Min.   :1.00   Min.   :1.00   Min.   : 20
##  1st Qu.: 69.5   1st Qu.:106   1st Qu.:2.00   1st Qu.:2.00   1st Qu.: 68
##  Median : 82.3   Median :186   Median :2.00   Median :2.00   Median :143
##  Mean   : 81.6   Mean   :252   Mean   :1.85   Mean   :2.36   Mean   :191
##  3rd Qu.: 94.1   3rd Qu.:312   3rd Qu.:2.00   3rd Qu.:3.00   3rd Qu.:252
##  Max.   :133.5   Max.   :835   Max.   :2.00   Max.   :4.00   Max.   :791
##      nurses       services
##  Min.   : 14   Min.   : 5.7
##  1st Qu.: 66   1st Qu.:31.4
##  Median :132   Median :42.9
##  Mean   :173   Mean   :43.2
##  3rd Qu.:218   3rd Qu.:54.3
##  Max.   :656   Max.   :80.0
```

```
# lm for infection risk
lm1 <- lm(stay ~ inf_risk, data = senic)
summary(lm1)
```

```
##
## Call:
## lm(formula = stay ~ inf_risk, data = senic)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.059 -0.778 -0.149  0.716  8.280
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.337      0.521   12.16  < 2e-16 ***
## inf_risk       0.760      0.114    6.64  1.2e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.62 on 111 degrees of freedom
## Multiple R-squared:  0.285,  Adjusted R-squared:  0.278
## F-statistic: 44.1 on 1 and 111 DF,  p-value: 1.18e-09
```

```
# lm for available facilities
lm2 <- lm(stay ~ services, data = senic)
summary(lm2)
```

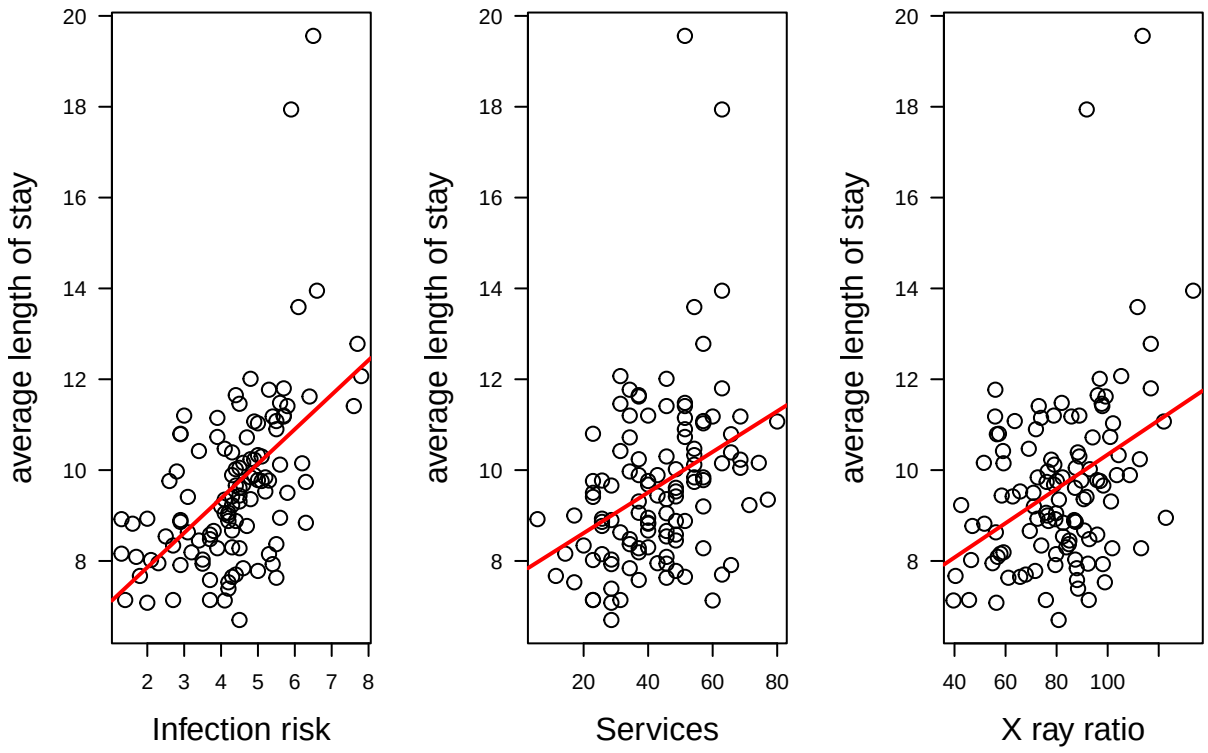
```
##
## Call:
## lm(formula = stay ~ services, data = senic)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.271 -1.072 -0.282  0.758  9.543
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.7188     0.5102   15.13  < 2e-16 ***
## services      0.0447     0.0112    4.01  0.00011 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.79 on 111 degrees of freedom
## Multiple R-squared:  0.126,  Adjusted R-squared:  0.119
## F-statistic: 16.1 on 1 and 111 DF,  p-value: 0.000111
```

```
# lm for xray ratio
lm3 <- lm(stay ~ xray_r, data = senic)
summary(lm3)
```

```
##
## Call:
## lm(formula = stay ~ xray_r, data = senic)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.923 -1.081 -0.271  0.820  8.701
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  6.56637    0.72609    9.04  5.7e-15 ***
## xray_r       0.03776    0.00866    4.36  2.9e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.77 on 111 degrees of freedom
## Multiple R-squared:  0.146,  Adjusted R-squared:  0.139
## F-statistic:   19 on 1 and 111 DF,  p-value: 2.91e-05
```

```r
# lm plots
par(mfrow = c(1, 3))
plot(senic$stay ~ senic$inf_risk, cex = 1.5, cex.lab = 1.5, las = 1,
    cex.main = 1.5, xlab = "Infection risk", ylab = "average length of stay")
abline(lm1, lwd = 2, col = "red")
plot(senic$stay ~ senic$services, cex = 1.5, cex.lab = 1.5, las = 1,
    cex.main = 1.5, xlab = "Services", ylab = "average length of stay")
abline(lm2, lwd = 2, col = "red")
plot(senic$stay ~ senic$xray_r, cex = 1.5, cex.lab = 1.5, las = 1,
    cex.main = 1.5, xlab = "X ray ratio", ylab = "average length of stay")
abline(lm3, lwd = 2, col = "red")
```



0.0.1.1 Observation

- If we observe the p-values, we have values < 0.05 for three predictors. This says that the null hypothesis of having coefficients $= 0$ is rejected and linear relation between predictor and response holds good. This suggest the linearity with confidence (1- pvalue)%.
- Plots of regression functions also shows the linear relationship provide a good fit for all the three

predictors.

0.0.2 Problem 2

```
# Mean squared error
mse1 <- mean(lm1$residuals^2)
mse2 <- mean(lm2$residuals^2)
mse3 <- mean(lm3$residuals^2)
mse <- c(mse1, mse2, mse3)
r2 <- c(summary(lm1)$r.squared, summary(lm2)$r.squared, summary(lm3)$r.squared)
df <- data.frame(MSE = mse, R2 = r2)
row.names(df) <- c("Infection risk", "Facilities and Services",
    "Chest X-ray ratio")
df
```

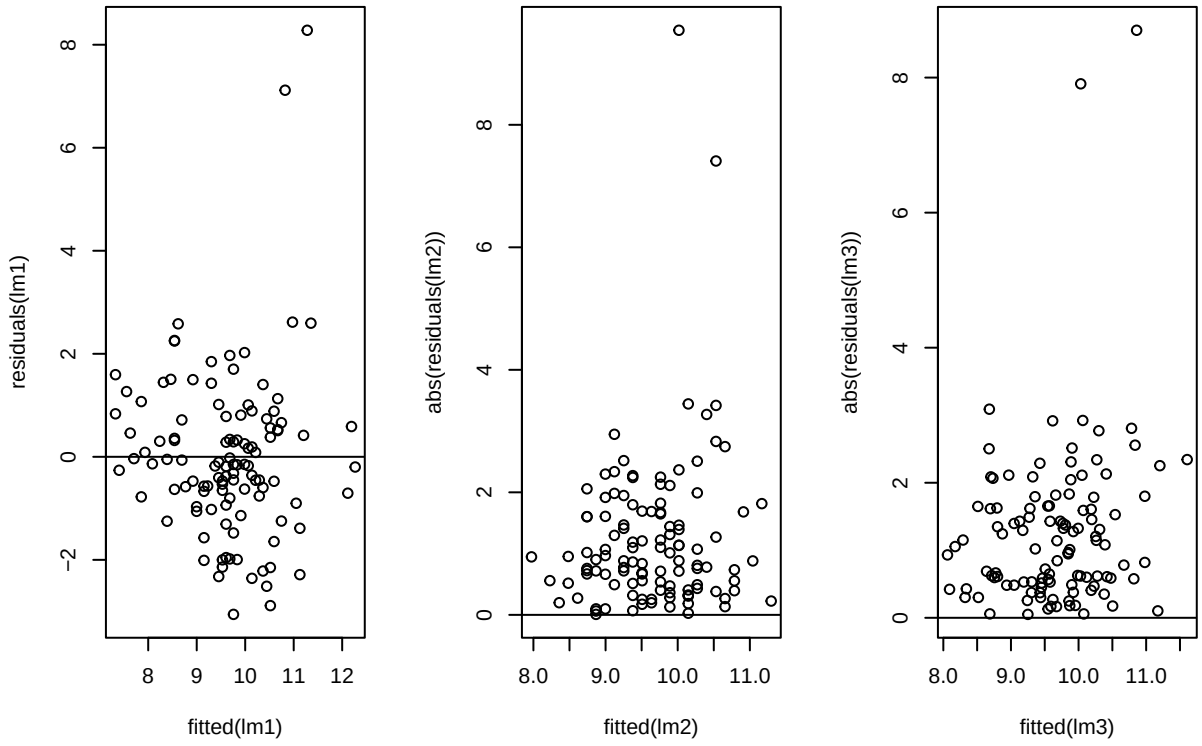
```
##                           MSE    R2
## Infection risk          2.59 0.285
## Facilities and Services 3.16 0.126
## Chest X-ray ratio       3.09 0.146
```

0.0.2.1 Observation

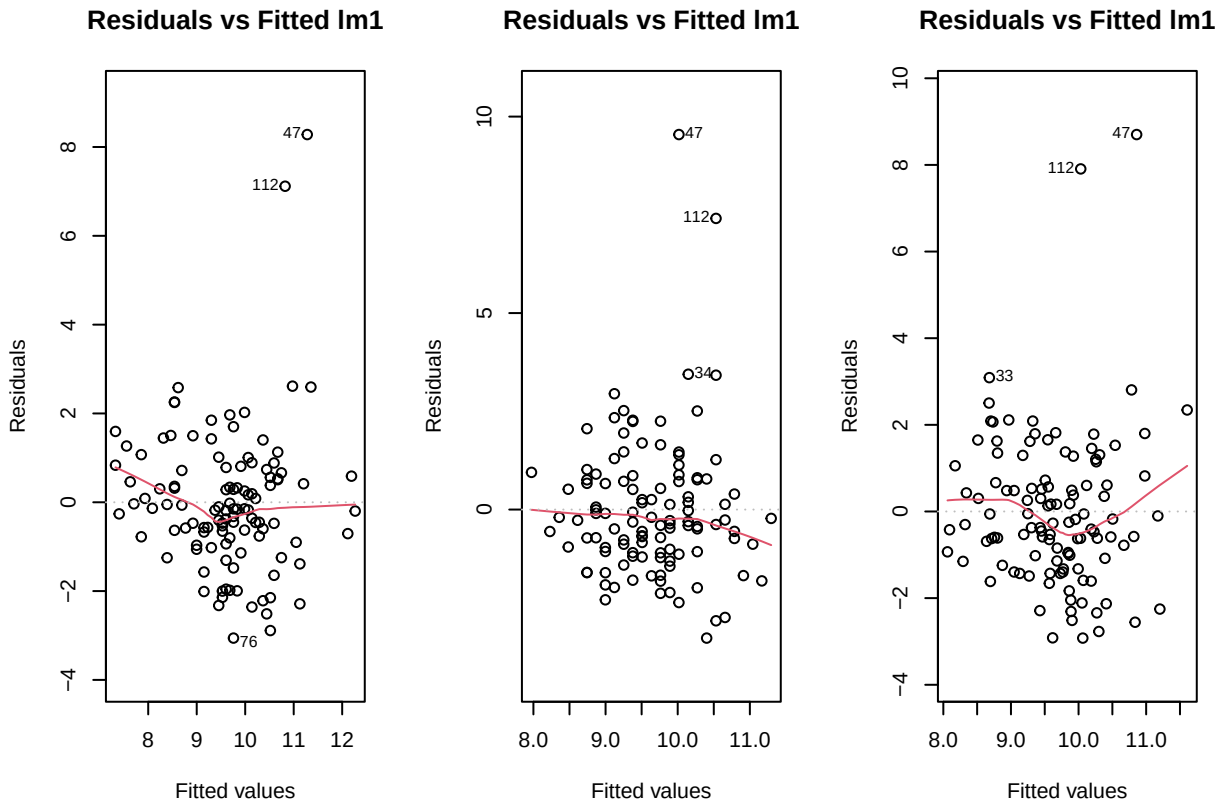
- R-square values are under 0.3 for all the three predictors. This shows that less than 30% of variance in dependent variable is explained by these univariate models.
- For the 'Infection risk' predictor, we have less mean square value and hightest Rsquare, so we can say out of three predictors, 'Infection risk' has the largest reduction in variablity of the average length of stay.

0.0.3 Problem 3

```
# Residual plots
par(mfrow = c(1, 3))
plot(fitted(lm1), residuals(lm1))
abline(h = 0)
plot(fitted(lm2), abs(residuals(lm2)))
abline(h = 0)
plot(fitted(lm3), abs(residuals(lm3)))
abline(h = 0)
```




```
par(mfrow = c(1, 3))
plot(lm1, which = c(1), main = "Residuals vs Fitted lm1", caption = "")
plot(lm2, which = c(1), main = "Residuals vs Fitted lm1", caption = "")
plot(lm3, which = c(1), main = "Residuals vs Fitted lm1", caption = "")
```



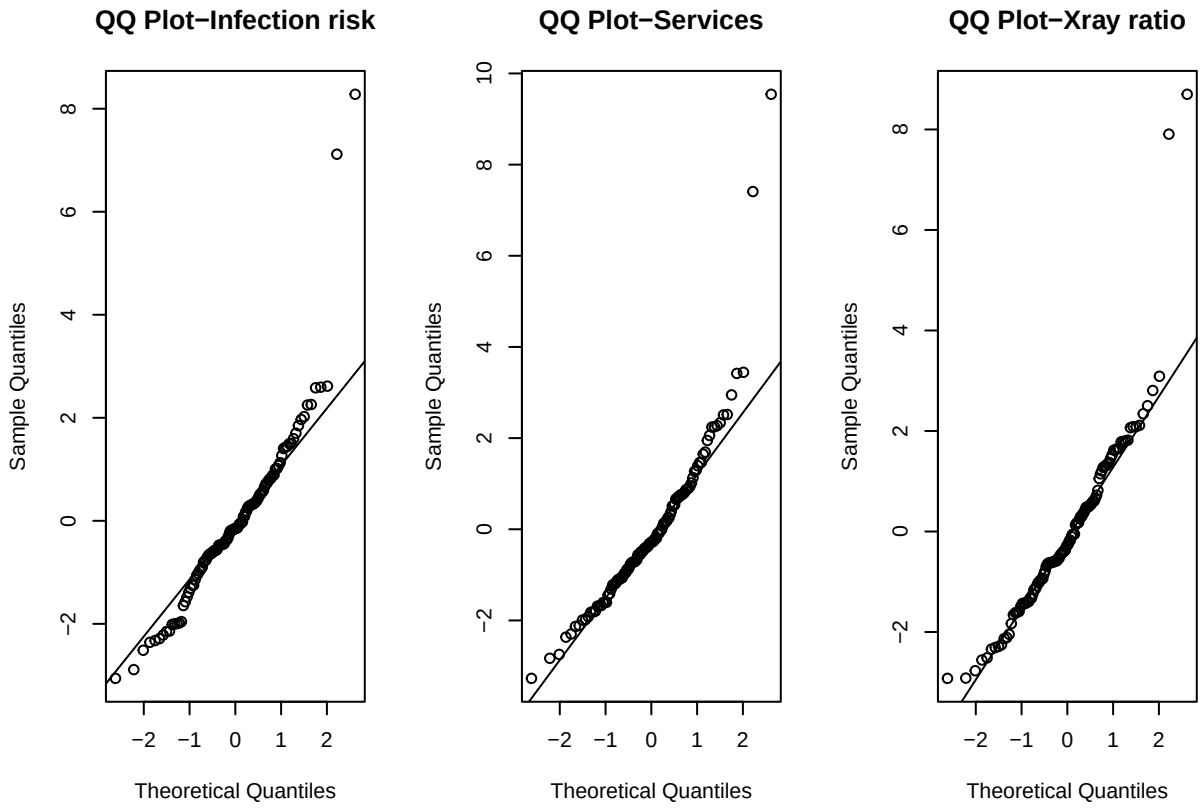

```
# Linearity of residuals
par(mfrow = c(1, 3))
qqnorm(residuals(lm1), main = "QQ Plot-Infection risk")
qqline(residuals(lm1))
shapiro.test(residuals(lm1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm1)
## W = 0.9, p-value = 2e-08
```

```
qqnorm(residuals(lm2), main = "QQ Plot-Services")
qqline(residuals(lm2))
shapiro.test(residuals(lm2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm2)
## W = 0.9, p-value = 7e-09
```

```
qqnorm(residuals(lm3), main = "QQ Plot-Xray ratio")
qqline(residuals(lm3))
```




```
shapiro.test(residuals(lm3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm3)
## W = 0.9, p-value = 3e-08
```

```
# Homoscedasticity
ncvTest(lm1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 23.4, Df = 1, p = 1e-06
```

```
ncvTest(lm2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 10.7, Df = 1, p = 0.001
```

```
ncvTest(lm3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 17.1, Df = 1, p = 4e-05
```

```
# Correlation
dwtest(lm1)
```

```
##
##  Durbin-Watson test
##
## data:  lm1
## DW = 2, p-value = 0.7
## alternative hypothesis: true autocorrelation is greater than 0
```

```
dwtest(lm2)
```

```
##
##  Durbin-Watson test
##
## data:  lm2
## DW = 2, p-value = 0.6
## alternative hypothesis: true autocorrelation is greater than 0
```

```
dwtest(lm3)
```

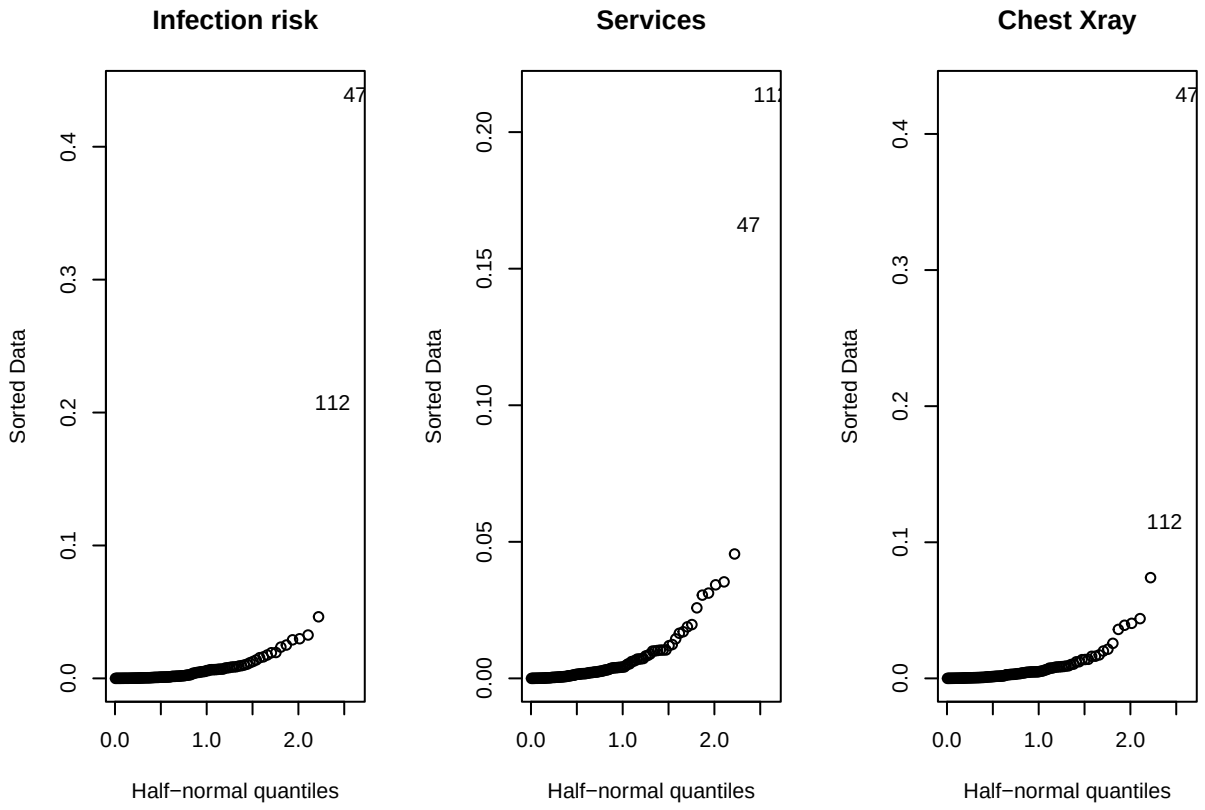
```
##
##  Durbin-Watson test
##
## data:  lm3
## DW = 2, p-value = 0.3
## alternative hypothesis: true autocorrelation is greater than 0
```

0.0.3.1 Observation

- If we see the residual plots, model with infection risk is better compared to other two.
- Normality: The p-values of shapiro test are less than the significant value < 0.05. Low p-values reject the null hypothesis of normality and that suggests that residuals are not following normal distribution. However, if we see the Qplots, we can see the normality, but presence of outliers could be the reason behind the poor p-values.
- Homoscedasticity: ncvtest has null hypothesis of constant variance for residuals. We have p-values less than the significant value 0.05. This shows the variance is changing basing on the change in the fitted valus.
- Correlation: Dwtest has the null hypothesis that the residuals from a linear regression are uncorrelated. Our test has p-values higher than 0.05 supports the null hypothesis suggesting that dependent and independent variables are not autocorrelated. DW=2 also suggests the no autocorrelation.

0.0.4 Problem 4

```
# Halfnormal quantiles to see influencing points
par(mfrow = c(1, 3))
halfnorm(cooks.distance(lm1), main = "Infection risk")
halfnorm(cooks.distance(lm2), main = "Services")
halfnorm(cooks.distance(lm3), main = "Chest Xray")
```




```
# Influencing points
lm1I <- influence.measures(lm1)
which(apply(lm1I$is.inf, 1, any))
```

```
##   2  13  40  47  49  53  54  85  93 107 108 112
##   2  13  40  47  49  53  54  85  93 107 108 112
```

```
# Refit the model without the case 47 and 112 senic1 <-
# senic[!(senic$id %in% c(47,112)),]
lm1_r <- lm(stay ~ inf_risk, data = senic, subset = -c(47, 112))
str(summary(lm1_r))
```

```
## List of 11
##  $ call         : language lm(formula = stay ~ inf_risk, data = senic, subset = -c(47, 112))
##  $ terms        :Classes 'terms', 'formula'  language stay ~ inf_risk
##   .. ..- attr(*, "variables")= language list(stay, inf_risk)
##   .. ..- attr(*, "factors")= int [1:2, 1] 0 1
##   .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. ..$ : chr [1:2] "stay" "inf_risk"
##   .. .. .. ..$ : chr "inf_risk"
##   .. ..- attr(*, "term.labels")= chr "inf_risk"
##   .. ..- attr(*, "order")= int 1
##   .. ..- attr(*, "intercept")= int 1
##   .. ..- attr(*, "response")= int 1
##   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. ..- attr(*, "predvars")= language list(stay, inf_risk)
##   .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
```

```
##   .. .. ..- attr(*, "names")= chr [1:2] "stay" "inf_risk"
##  $ residuals    : Named num [1:111] -2.219 0.995 -0.156 -1.314 0.875 ...
##   ..- attr(*, "names")= chr [1:111] "1" "2" "3" "4" ...
##  $ coefficients : num [1:2, 1:4] 6.8492 0.6097 0.4014 0.0888 17.0645 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "(Intercept)" "inf_risk"
##   .. ..$ : chr [1:4] "Estimate" "Std. Error" "t value" "Pr(>|t|)"
##  $ aliased      : Named logi [1:2] FALSE FALSE
##   ..- attr(*, "names")= chr [1:2] "(Intercept)" "inf_risk"
##  $ sigma        : num 1.24
##  $ df           : int [1:3] 2 109 2
##  $ r.squared    : num 0.302
##  $ adj.r.squared: num 0.296
##  $ fstatistic   : Named num [1:3] 47.1 1 109
##   ..- attr(*, "names")= chr [1:3] "value" "numdf" "dendf"
##  $ cov.unscaled : num [1:2, 1:2] 0.10515 -0.02225 -0.02225 0.00515
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "(Intercept)" "inf_risk"
##   .. ..$ : chr [1:2] "(Intercept)" "inf_risk"
##  - attr(*, "class")= chr "summary.lm"
```

```
# Prediction interval for the old model
senic[c("inf_risk", "stay")][c(47, 112), ]
```

```
##     inf_risk stay
## 47       6.5 19.6
## 112      5.9 17.9
```

```
fitted.values(lm1)[c(47, 112)]
```

```
##   47  112
## 11.3 10.8
```

```
newdata <- data.frame(inf_risk = c(6.5, 5.9))
pio <- predict(lm1, newdata = newdata, interval = "prediction")
row.names(pio) <- c("47", "112")
pio
```

```
##      fit  lwr  upr
## 47  11.3 8.01 14.5
## 112 10.8 7.57 14.1
```

```
# Prediction interval for the new model
pi <- predict(lm1_r, newdata = newdata, interval = "prediction")
row.names(pi) <- c("47", "112")
pi
```

```
##      fit  lwr  upr
## 47  10.8 8.32 13.3
## 112 10.4 7.97 12.9
```

```
# Assumptions of refit model
shapiro.test(residuals(lm1_r))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm1_r)
```

```
## W = 1, p-value = 0.7
```

```
ncvTest(lm1_r)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2.27, Df = 1, p = 0.1
```

```
dwtest(lm1_r)
```

```
##
##  Durbin-Watson test
##
## data:  lm1_r
## DW = 2, p-value = 0.8
## alternative hypothesis: true autocorrelation is greater than 0
```

0.0.4.1 Observation

- The R2 value has improved from 2.8 to 3 for the refit model. The pr value of refit model is less than the original but not significantly less.
- Original 'Length of stay' values are Y=(19.56, 17.94) if their 'Risk of infection' values are X = (6.5,5.9). Using the refit model, we get the prediction intervals as (8.32, 13.3), (7.97, 12.9) for X respectively.
- We can say that, original values do not fall under the bounds of the prediction intervals.
- If we can compare the prediction intervals of original and refit model for 'Infection risk', we can say original model's interval are better than refit model. Removing the points (47,112) did not perform better.
- However, for refit mode, the assumptions are better than the original model.
- Normality : Refit model has p-value > 0.05 which satifies the null hypothesis of Normal distribution for model's residuals.
- Homoscedasticity : Refit model has p-value > 0.05 which satisfies the null hypothesis of constant variance. Refit model has constant variance between dependent and independent variables.
- Collinearity : We have p-value > 0.05 which satisfies the null hypothesis of uncorrelation.

0.0.5 Problem 5

```
y <- "stay"

# regsubsets
x1 <- c("id", "log(age)", "inf_risk", "log(cul_r)", "log(xray_r)",
    "log(beds)", "med", "log(region)", "log(daily_avg)", "log(nurses)",
    "log(services)")

fm <- as.formula(paste(y, paste(x1, collapse = "+"), sep = "~"))
b <- regsubsets(fm, force.in = 1, data = senic)
rs <- summary(b)
kable(with(rs, cbind(which, rss, adjr2, cp, bic)), digits = 4)
```

	(Intercept)	id	log(age)	inf_risk	log(cul_r)	log(xray_r)	log(beds)	med	log(region)	log(daily_avg)	log(nurses)	log(services)	rss	adjr2	cp	bic
2	1	1	0	1	0	0	0	0	0	0	0	0	289	0.2807	6.2	-25.0
3	1	1	0	1	0	0	0	0	1	0	0	0	225	0.4343	7.8	-48.4
4	1	1	0	1	0	0	0	0	1	1	0	0	202	0.4882	5.0	-56.1
5	1	1	0	1	0	0	0	0	1	1	1	0	186	0.5241	6.8	-60.7

	(Intercept)	id	log(age)	inf_risk	log(cul_r)	log(xray_r)	log(beds)	med	log(region)	log(daily_avg)	log(nurses)	log(services)	rss	adjr2	cp	bic
6	1	1	1	1	0	0	0	0	1	1	1	0	172	0.555	10.27	-64.6
7	1	1	1	1	0	1	0	0	1	1	1	0	165	0.569	7.70	-64.7
8	1	1	1	1	0	1	0	0	1	1	1	1	163	0.572	8.14	-61.6

```
# Model with subset1
x <- c("log(id)", "log(age)", "inf_risk", "log(xray_r)", "log(region)",
    "log(daily_avg)", "log(nurses)", "log(services)")
#'id','log(cul_r)','log(beds)','med',

fm <- as.formula(paste(y, paste(x, collapse = "+"), sep = "~"))
mod <- lm(fm, data = senic, subset = -c(47, 112, 43, 80, 52,
    78, 26, 81, 54))
xtable(summary(mod))
```

```
## % latex table generated in R 4.1.0 by xtable 1.8-4 package
## % Sun Jul 11 01:02:04 2021
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\
##   \hline
## (Intercept) & -17.7294 & 5.0944 & -3.48 & 0.0008 \\
##   log(id) & 0.0835 & 0.0932 & 0.90 & 0.3727 \\
##   log(age) & 5.0726 & 1.1179 & 4.54 & 0.0000 \\
##   inf\_risk & 0.3330 & 0.0932 & 3.57 & 0.0006 \\
##   log(xray\_r) & 1.0155 & 0.4157 & 2.44 & 0.0164 \\
##   log(region) & -1.0972 & 0.1887 & -5.81 & 0.0000 \\
##   log(daily\_avg) & 1.0468 & 0.2761 & 3.79 & 0.0003 \\
##   log(nurses) & 0.0019 & 0.3339 & 0.01 & 0.9955 \\
##   log(services) & -0.9454 & 0.3791 & -2.49 & 0.0144 \\
##    \hline
## \end{tabular}
## \end{table}
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = fm, data = senic, subset = -c(47, 112, 43, 80, 52,
##     78, 26, 81, 54))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7599 -0.6164  0.0797  0.5049  2.0714
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -17.7294     5.0944   -3.48  0.00076 ***
## log(id)          0.0835     0.0932    0.90  0.37273
## log(age)         5.0726     1.1179    4.54  1.7e-05 ***
```

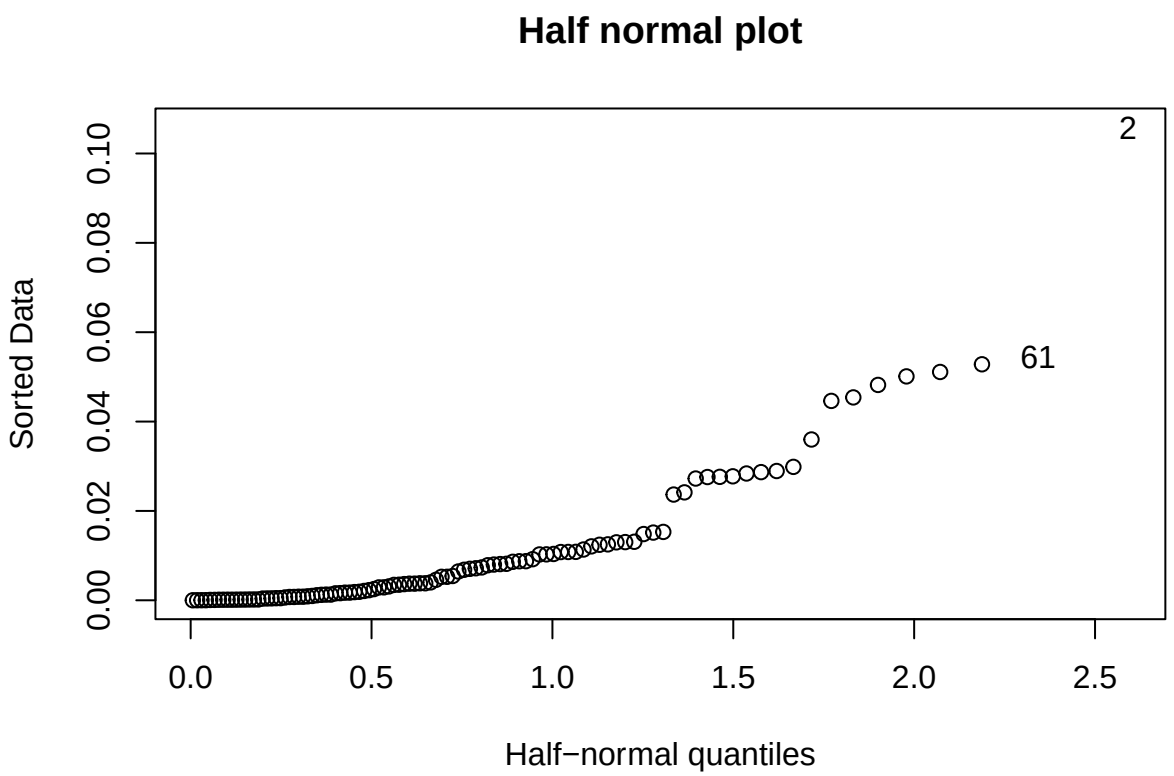
```
## inf_risk         0.3330     0.0932    3.57  0.00056 ***
## log(xray_r)      1.0155     0.4157    2.44  0.01641 *
## log(region)     -1.0972     0.1887   -5.81  8.2e-08 ***
## log(daily_avg)   1.0468     0.2761    3.79  0.00026 ***
## log(nurses)      0.0019     0.3338    0.01  0.99547
## log(services)   -0.9454     0.3791   -2.49  0.01436 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.891 on 95 degrees of freedom
## Multiple R-squared:  0.664,  Adjusted R-squared:  0.636
## F-statistic: 23.5 on 8 and 95 DF,  p-value: <2e-16
```

0.0.5.1 Observations

- We can observe from the regsubsets, we have lowest BIC, Mallow's Cp and high adjusted R2 when features 'cultures performed', 'number of beds' and 'region' are omitted.
- After we remove the features suggested by regsubsets, we use the log transformation to normalize the features that have high values.
- We remove the influencing points and perform the regression model on the selected set of features.

0.0.6 Problem 6

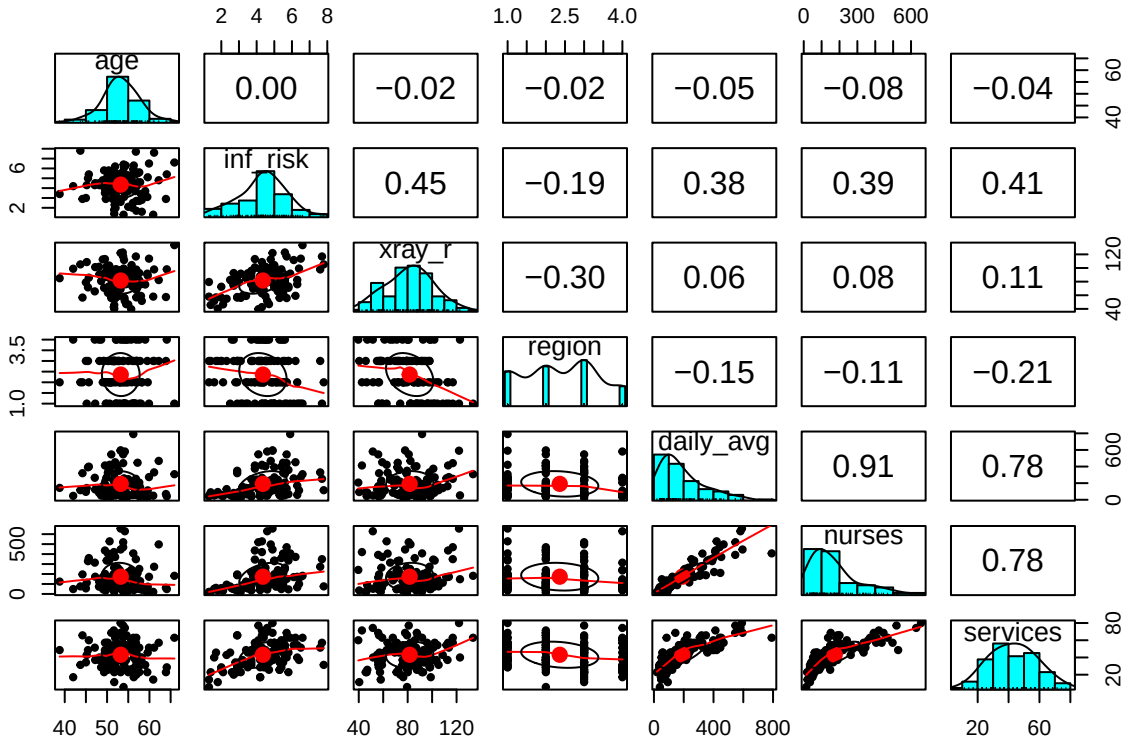
```
# Halfnormal quantiles to see influencing points
halfnorm(cooks.distance(mod), main = "Half normal plot")
```



```
# Influencing points
modI <- influence.measures(mod)
which(apply(modI$is.inf, 1, any))
```

```
##   1   2   4  21  49  93 101 107
##   1   2   4  21  46  85  93  99
```

```
# Correlation
dat0 <- senic[, c("age", "inf_risk", "xray_r", "region", "daily_avg",
    "nurses", "services")]
pairs.panels(dat0)
```




```
# Assumptions
shapiro.test(residuals(mod))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod)
## W = 1, p-value = 0.4
```

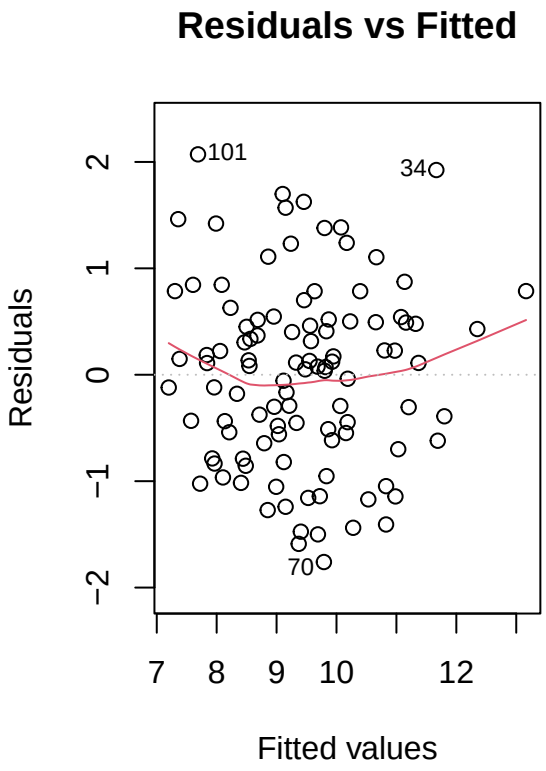
```
ncvTest(mod)
```

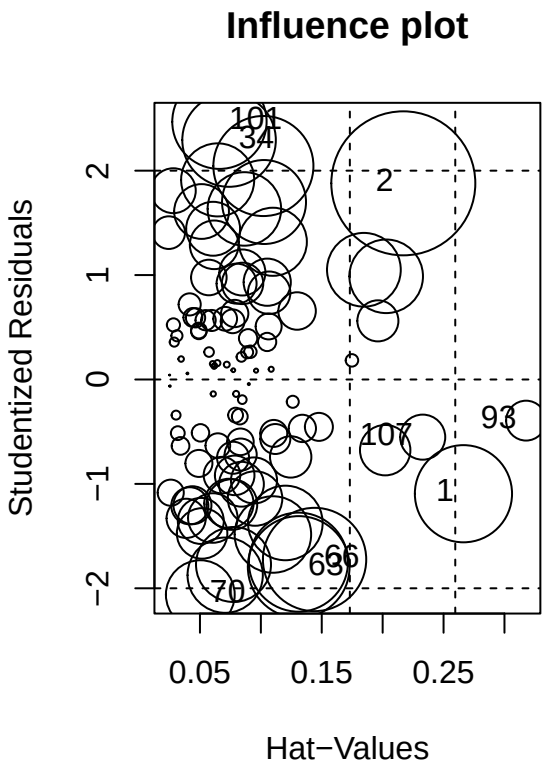
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.0113, Df = 1, p = 0.9
```

```r
dwtest(mod)
```

```
##
##  Durbin-Watson test
##
## data:  mod
## DW = 2, p-value = 0.4
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
# Residual plots
par(mfrow = c(1, 2))
plot(mod, which = c(1), main = "Residuals vs Fitted", caption = "")
influencePlot(mod, id = list(n = 3), main = "Influence plot")
```






```
##      StudRes    Hat   CookD
## 1     -1.094 0.2663 0.04820
## 2      1.879 0.2170 0.10584
## 34     2.293 0.0739 0.04463
## 63    -1.796 0.1311 0.05282
## 66    -1.724 0.1441 0.05448
## 70    -2.061 0.0503 0.02415
## 93    -0.394 0.3177 0.00811
## 101    2.469 0.0660 0.04541
## 107   -0.552 0.2329 0.01036
```

0.0.6.1 Observations

- We can see the influencing points from influencer plot and halfnormal plot. We have removed the points for our model.
- From pair plots, we can see that feature set (daily-average, nurses), (daily-average, services), (nurses, services) has a positive correlation.
- Normality: We have p-value(=0.5) > 0.05 significant value suggesting a normal distribution for the residuals of our final model.
- Homoscedasticity: We have p-value(=0.8) > 0.05 suggesting a constant variance between dependent and independent variables.
- Collinearity: We have p-value(=0.3) > 0.05 significant value suggesting uncorrelation between dependent and independent variables.

0.0.7 Problem 7

- We have a dataset which has below sets of features and the target label that will be predicted.
- Features:
  - Id number
  - Age
  - Infection risk
  - Routine culturing ratio
  - Routine Chest x-ray ratio
  - Number of beds
  - Medical shcool affication
  - Geographic region
  - Average daily census
  - Number of nurses
  - Available facilities and services.
- Target:
  - Length of Stay
- We created a model with the features and tested for normality, Constant variance and collinearity and found our assumptions were less than significant values and tests were negative.
- For a better prediction model, our assumptions should be positive. We call difference between fitted values and actual values target variables as residuals. A better prediction model should hae residuals forllow normality, constant variance and uncorrelated.
- We used halfnormal test and determined the influencing points and built our prediction model with subset of data by omitting the influencing points.
- We observed few features are high in scale compared to others. Used log transformation to normalize the out of scale features.
- We used regsubsets method to determine the best set of variables for our model. We used parameters BIC, Mallow's Cp and adjusted R2 to determine the best set of variables out of the given set. Below are the best set of features with lowest BIC, Cp and highest adjusted R2.
- Best set of features:
  - age
  - inf_risk
  - xray_r
  - region
  - daily_avg
  - nurses
  - services
- By using above set of variables, we get the Adjusted R2 = 0.64 and assumptions normality, constant variance and Collinearity are tested positive.

0.1 Document Information.

All of the statistical analyses in this document will be performed using R version 4.1.0 (2021-05-18). R packages used will be maintained using the packrat dependency management system.

```
sessionInfo()
```

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19041)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] psych_2.1.6        leaps_3.1          faraway_1.0.7      xtable_1.8-4
##  [5] lmtest_0.9-38      zoo_1.8-9          PairedData_1.1.1   mvtnorm_1.1-2
##  [9] gld_2.6.2          ggpubr_0.4.0       car_3.0-11         carData_3.0-4
## [13] mnormt_2.0.2       vcd_1.4-8          epiDisplay_3.5.0.1 nnet_7.3-16
## [17] foreign_0.8-81     Hmisc_4.5-0        Formula_1.2-4      survival_3.2-11
## [21] lattice_0.20-44    MASS_7.3-54        ggplot2_3.3.5      rmarkdown_2.8
## [25] knitr_1.33
##
## loaded via a namespace (and not attached):
##  [1] nlme_3.1-152        RColorBrewer_1.1-2  tools_4.1.0
##  [4] backports_1.2.1     utf8_1.2.1          R6_2.5.0
##  [7] rpart_4.1-15        colorspace_2.0-1    withr_2.4.2
## [10] tidyselect_1.1.1    gridExtra_2.3       curl_4.3.1
## [13] compiler_4.1.0      formatR_1.11        htmlTable_2.2.1
## [16] scales_1.1.1        checkmate_2.0.0     proxy_0.4-26
## [19] stringr_1.4.0       digest_0.6.27       minqa_1.2.4
## [22] rio_0.5.27          base64enc_0.1-3     jpeg_0.1-8.1
## [25] pkgconfig_2.0.3     htmltools_0.5.1.1   lme4_1.1-27.1
## [28] highr_0.9           htmlwidgets_1.5.3   rlang_0.4.11
## [31] readxl_1.3.1        rstudioapi_0.13     generics_0.1.0
## [34] dplyr_1.0.7         zip_2.2.0           magrittr_2.0.1
## [37] Matrix_1.3-3        Rcpp_1.0.6          munsell_0.5.0
## [40] fansi_0.5.0         abind_1.4-5         lifecycle_1.0.0
## [43] stringi_1.6.1       yaml_2.2.1          parallel_4.1.0
## [46] forcats_0.5.1       crayon_1.4.1        lmom_2.8
## [49] haven_2.4.1         splines_4.1.0       hms_1.1.0
## [52] tmvnsim_1.0-2       pillar_1.6.1        boot_1.3-28
## [55] ggsignif_0.6.2      glue_1.4.2          evaluate_0.14
## [58] latticeExtra_0.6-29 data.table_1.14.0   nloptr_1.2.2.2
## [61] png_0.1-7           vctrs_0.3.8         cellranger_1.1.0
```

```
## [64] gtable_0.3.0        purrr_0.3.4         tidyr_1.1.3
## [67] xfun_0.23           openxlsx_4.2.4      broom_0.7.8
## [70] e1071_1.7-7         rstatix_0.7.0       class_7.3-19
## [73] tibble_3.1.2        cluster_2.1.2       ellipsis_0.3.2
```