HW4

2023-09-18

2.25 (a) (b) (d), 2.45, 3.25
Additional instructions:

- For 2.25 (a), follow the ANOVA table format given in Lecture 3 Slide #8. Provide all elements present in the former.
- For 2.25 (b), follow the five steps listed in the lecture slides.
- For 2.25 (d), "r" is the Pearson's correlation coefficient between Y and X. In R you can use **cor(X, Y)** to calculate the result.
- For 3.25, for the required "normal probability plot" create a Q-Q plot in R. See demo code in this module to create the required plots.

(3.25 in p.6)

2.25. Refer to Airfreight breakage Problem 1.21.

1.21. Airfreight breakage. A substance used in biological and medical research is shipped by airfreight to users in cartons of 1,000 ampules. The data below, involving 10 shipments, were collected on the number of times the carton was transferred from one aircraft to another over the shipment route (X) and the number of ampules found to be broken upon arrival (Y). Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	4	5	6	7	8	9	10
X_i:	1	0	2	0	3	1	0	1	2	0
Y_i:	16	9	17	12	22	13	8	15	19	11

a. Set up the ANOVA table. Which elements are additive?

	SS	DF	MS
Regression	SSR = Σ (Y.hat – mean)^2 = **160**	DE.R = p - 1 = **1**	MSR = **160**
Error	SSE = Σ (Y-Y.hat)^2 = **17.6**	DF.E = n – p = **8**	MSE = **2.2**
Total	SSTO = Σ (Y-mean)^2 = **177.6**	DF.TO =n - 1 = **9**	MSTO =**19.73333**

Sum of squares and degree of freedom are **additive: SSTO** = SSR + SSE and **DF.TO** = DF.R + DR.E
Yet, MS is not additive.

b. Conduct an F test to decide whether or not there is a linear association between the number of times a carton is transferred and the number of broken ampules; control the a risk at .05. State the alternatives, decision rule, and conclusion.

Step 1: Check Assumptions

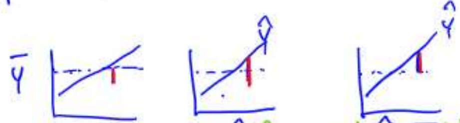
$\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

Step 2: Make Hypothesis

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$ (linear association)

Step 3: Calculate test-statistic.



$$\Sigma(Y_i - \bar{Y})^2 = \Sigma(Y_i - \hat{Y})^2 + \Sigma(\hat{Y} - \bar{Y})^2$$

SSTO = SSE + SSR

(total error) (Std. err) (regression error)

D.F $\quad n-1 = n-p + p-1$

$$\frac{SSTO}{n-1} \neq \frac{SSE}{n-p} + \frac{SSR}{n-1}$$

(MSTO) (MSE) (MSR)

$F_{obs} = \frac{MSR}{MSE}$ ↑ larger ↓ smaller F_{obs} ↑, P ↓, reject H_0.

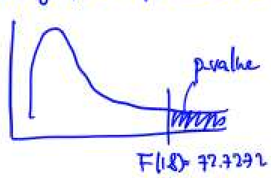
$F(p-1, n-p) = F(1,8) = 72.7272$

Step 4: p-value

$= 2.749e-05$

1-Pt(q, p-1, n-p)

Pt(q, p-1, np, lower.tail = F)



Step 5: Conclusion.

Because p-value = 2.749e-05 < α=0.05, we reject H_0 and conclude $\beta_1 \neq 0$, suggesting there is a linear association between Y and X.

Broken ampule No. of transfer.

d. Calculate R^2 and r. What proportion of the variation in Y is accounted for by introducing -X into the regression model?

```
R^2 = SSR/SST 0.9009
r = 0.949158
```

- There's a significant linear relationship between the number of times a carton is transferred and the number of broken ampules.
- About **90.09%** of the variation in broken ampules can be explained by the number of transfers.
- The **positive correlation** (0.9492) suggests that as the number of transfers increases, the number of broken ampules tends to increase as well.

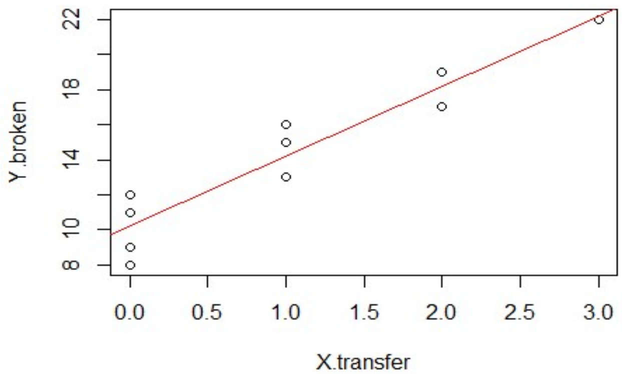
R code:

Import data

```
# Set the following to YOUR OWN folder
setwd("C:/Users/jyang/OneDrive - Arizona State University/10
Classes_OneDrive/2023_STP530_Regression")

# Import the dataset. The txt data file needs to exist in the folder above.
mydata <- read.table("CH01PR21.txt")
head(mydata)
##   V1 V2
## 1 16  1
## 2  9  0
## 3 17  2
## 4 12  0
## 5 22  3
## 6 13  1
# Rename the columns
colnames(mydata) <- c("Y.broken", "X.transfer") # Rename the columns
head(mydata)
##   Y.broken X.transfer
## 1       16          1
## 2        9          0
## 3       17          2
## 4       12          0
## 5       22          3
## 6       13          1
```

Plot data and find the linear regression model

```
# Fit the linear regression model
m <- lm(Y.broken ~ X.transfer, data=mydata)
plot(Y.broken ~ X.transfer, data=mydata)
abline(coef(m), col="red")
```




```
summary(m)
##
## Call:
## lm(formula = Y.broken ~ X.transfer, data = mydata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.2   -1.2    0.3    0.8    1.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.2000     0.6633  15.377 3.18e-07 ***
## X.transfer    4.0000     0.4690   8.528 2.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.483 on 8 degrees of freedom
## Multiple R-squared:  0.9009, Adjusted R-squared:  0.8885
## F-statistic: 72.73 on 1 and 8 DF,  p-value: 2.749e-05
```

ANOVA table- mannual calculation

```
# Manually calculate the elements in the ANOVA table
Y <- mydata$Y.broken
Y.bar <- mean(Y)
Y.hat <- predict(m)
cbind(Y, Y.bar, Y.hat)
##     Y Y.bar Y.hat
## 1  16  14.2  14.2
## 2   9  14.2  10.2
## 3  17  14.2  18.2
## 4  12  14.2  10.2
## 5  22  14.2  22.2
## 6  13  14.2  14.2
## 7   8  14.2  10.2
## 8  15  14.2  14.2
## 9  19  14.2  18.2
## 10 11  14.2  10.2
SSTO <- sum((Y - Y.bar) ^ 2)
SSR <- sum((Y.hat - Y.bar) ^ 2)
SSE <- sum((Y - Y.hat) ^ 2)
```

```
n <- nrow(mydata)
p <- 2

df.TO <- n - 1
df.R <- p - 1
df.E <- n - p

MSTO <- SSTO / df.TO
MSR <- SSR / df.R
MSE <- SSE / df.E

SSTO; SSR; SSE
## [1] 177.6
## [1] 160
## [1] 17.6
df.TO; df.R; df.E
## [1] 9
## [1] 1
## [1] 8
MSTO; MSR; MSE
## [1] 19.73333
## [1] 160
## [1] 2.2
var(Y) # equal to MSTO
## [1] 19.73333
```

ANOVA

```
anova(m)
## Analysis of Variance Table
##
## Response: Y.broken
##             Df Sum Sq Mean Sq F value    Pr(>F)
## X.transfer   1  160.0   160.0  72.727 2.749e-05 ***
## Residuals    8   17.6     2.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-test and p-value

F(p-1, n-p) F=MSR/MSE

```
F.statistic <- MSR/MSE
F.statistic
## [1] 72.72727
1 - pf(q=72.7272, df1=(p-1) , df2=(n-p))
## [1] 2.748679e-05
pf(q=72.7272, df1=(p-1) , df2=(n-p), lower.tail=F)
## [1] 2.748679e-05
```

Pearson's correlation r

```
cor(y=mydata$Y.broken,, x=mydata$X.transfer)
## [1] 0.949158
```

==

3.25. Refer to the CDI data set in Appendix C.2 and Project 1.43. For each of the three fitted regression models, obtain the residuals and prepare a residual plot against X and a normal probability plot. Summarize your conclusions. Is linear regression model (2.1) more appropriate in one case than in the others?

1.43. Refer to the CDI data set in Appendix C.2. The number of active physicians in a CDI (Y) is expected to be related to total population, number of hospital beds, and total personal income. Assume that first-order regression model (1.1) is appropriate for **each of the three predictor** variables.

a. Regress the number of active physicians in turn on each of the three predictor variables. State the estimated regression functions.
b. Plot the three estimated regression functions and data on separate graphs. Does a linear regression relation appear to provide a good fit for each of the three predictor variables?
c. Calculate MSE for each of the three predictor variables. Which predictor variable leads to the smallest variability around the fitted regression line?

Data Set C.2 CDI

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992. The 17 variables are:

Variable Number	Variable Name	Description
1	Identification number	1–440
2	County	County name
3	State	Two-letter state abbreviation
4	Land area	Land area (square miles)
5	Total population	Estimated 1990 population
6	Percent of population aged 18–34	Percent of 1990 CDI population aged 18–34
7	Percent of population 65 or older	Percent of 1990 CDI population aged 65 years old or older
8	Number of active physicians	Number of professionally active nonfederal physicians during 1990
9	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	Total serious crimes	Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	Percent bachelor's degrees	Percent of adult population (persons 25 years old or older) with bachelor's degree
13	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI labor force that is unemployed
15	Per capita income	Per capita income of 1990 CDI population (dollars)
16	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
17	Geographic region	Geographic region classification is that used by the U.S. Bureau of the Census, where: 1 = NE, 2 = NC, 3 = S, 4 = W

The three regression models:

1. Model 1 (The number of active physicians vs. Population):
   - R^2: 0.819, indicating that 81.9% of the variance in the number of physicians can be explained by the population.
   - Coefficient for Population: 0.0029
2. Model 2 (The number of active physicians vs. bed.no):
   - R^2: 0.792, indicating that 79.2% of the variance in the number of physicians can be explained by the number of hospital beds.
   - Coefficient for bed.no: 1.3214
3. Model 3 (The number of active physicians vs. TO.income.m):

- R^2: 0.841, indicating that <u>84.1% of the variance in the number of physicians can be explained by total personal income in millions.</u>
- Coefficient for TO.income.m: 0.1340

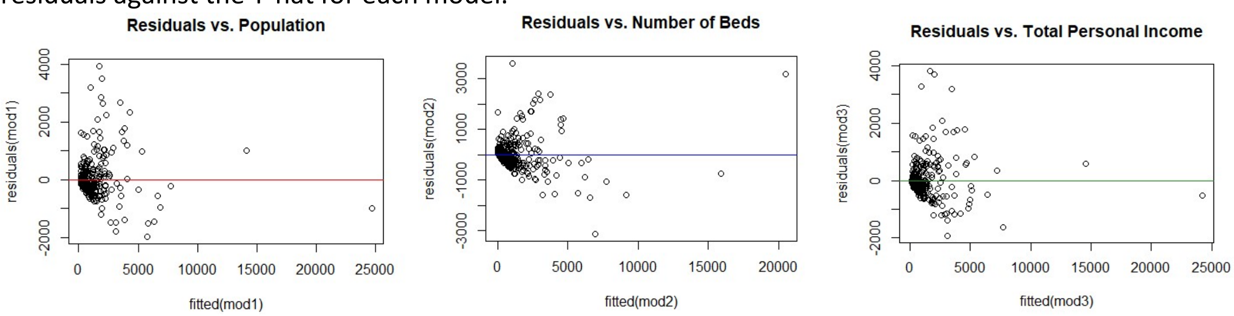
Problem:

(1) Plot the residuals against the predictor (or Y-hat).

(a) Whether there seems to be a linear or non-linear relationship?
(b) Is the residuals heteroskedastic? variance (vertical spread) across the board

<u>Residual Plots against Predictors:</u>
Note: A horizontal line around zero in the residual plot would indicate a good fit. Non-linearity or patterns in the residuals may suggest that the model isn't capturing some aspect of the data's structure.

To check if there's a linear relationship between the predictors and the response variable, I plotted the residuals against the Y-hat for each model.



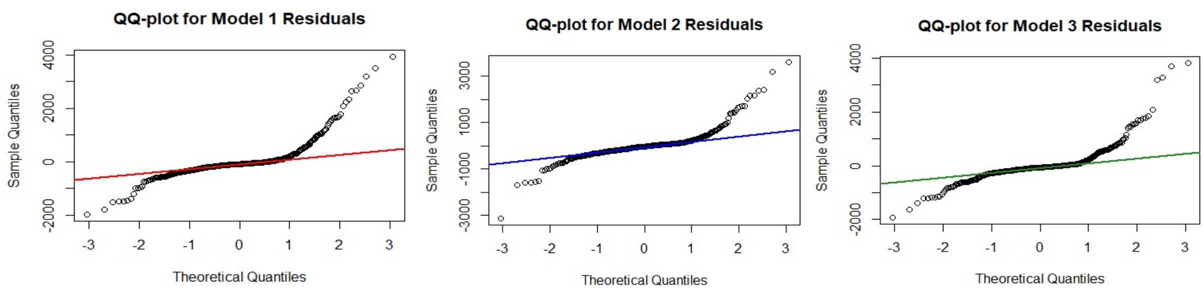
- If residuals scatter randomly around the horizontal line (y = 0), it indicates a linear relationship.
  ➔ The analysis shows that **patterns might suggest non-linearity**.
- If the residuals fan out (or show a funnel shape), it might indicate heteroskedasticity.
  ➔ All three models show evidence of **heteroskedasticity** (funnel shape opening to the right), where the spread of the residuals is not consistent across the range of fitted values.

Problem

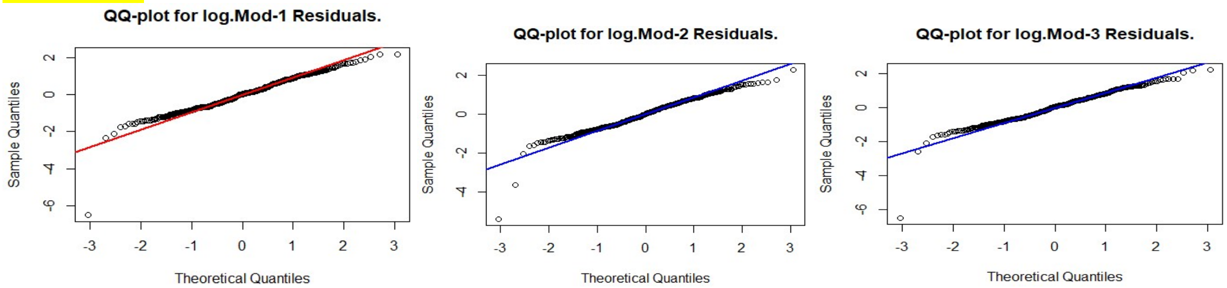
(2) Whether the residuals are normally distributed

Note: The QQ-plots can determine if the residuals are approximately normally distributed. Points closely following the **<u>straight line</u>** suggest the residuals are approximately **normally distributed**.

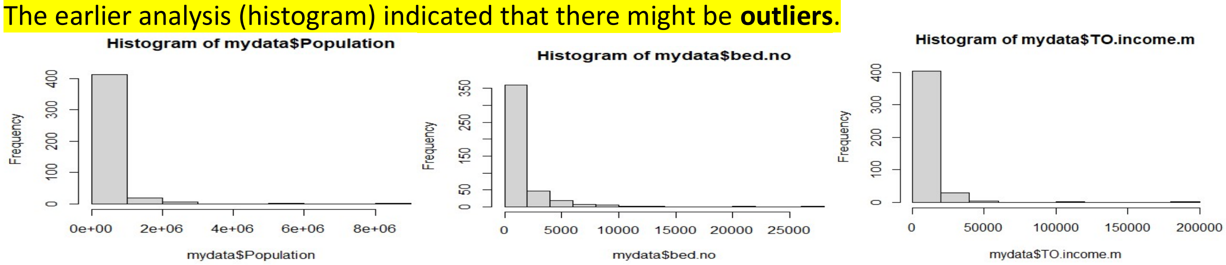
For all three models, the residuals appear to **deviate from the line** (see below), especially in the tails, suggesting they **may not be perfectly normally distributed**.



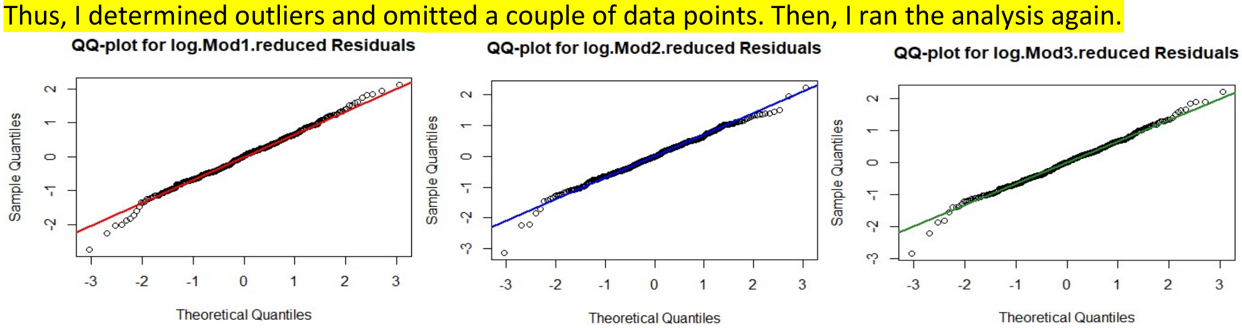
The three residual plots show funnel shape. Given the observations of heteroskedasticity in the residual plots, **log-transforming the models** could be a suitable corrective action. I perform log-transformation of the models.



The earlier analysis (histogram) indicated that there might be **outliers**.



Thus, I determined outliers and omitted a couple of data points. Then, I ran the analysis again.



After removing these outliers and re-running the analysis, the fit appears improved.

➔ The log-transformed models result in a better fit, it suggests that the original relationship between the variables might be exponential.
➔ Model 3 (Physician.no vs. TO.income.m) seems to be the closest to satisfying the assumptions of linear regression, though it still has potential issues.

➔ Given the patterns in the residuals and the deviations in the QQ-plots, transformations (like log-transformation) or more complex models might be considered for a better fit.

R code:

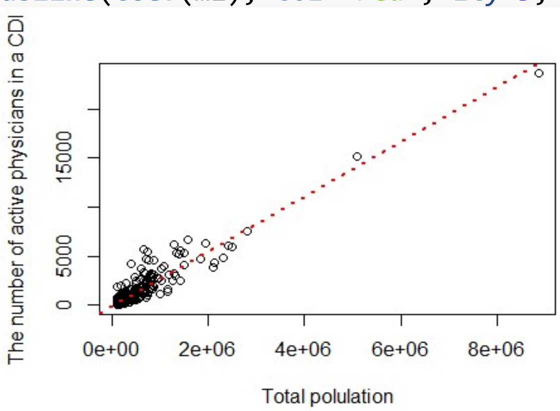
Import data

```
# Clean up the workspace for the new analysis
rm(list=ls())

# Set the following to YOUR OWN folder
setwd("C:/Users/jyang/OneDrive - Arizona State University/10
Classes_OneDrive/2023_STP530_Regression")

# Import the dataset. The txt data file needs to exist in the folder above.
mydata <- read.table("APPENC02.txt")
head(mydata)
##   V1          V2 V3   V4      V5   V6   V7    V8    V9    V10  V11  V12
V13
## 1  1 Los_Angeles CA 4060 8863164 32.1  9.7 23677 27700 688936 70.0 22.3
11.6
## 2  2        Cook IL  946 5105067 29.2 12.4 15153 21550 436936 73.4 22.8
11.1
## 3  3      Harris TX 1729 2818199 31.3  7.1  7553 12449 253526 74.9 25.4
12.5
## 4  4   San_Diego CA 4205 2498016 33.5 10.9  5905  6179 173821 81.9 25.3
8.1
## 5  5      Orange CA  790 2410556 32.6  9.2  6062  6369 144524 81.2 27.8
5.2
## 6  6       Kings NY   71 2300664 28.3 12.4  4861  8942 680966 63.7 16.6
19.5
##   V14   V15    V16 V17
## 1 8.0 20786 184230   4
## 2 7.2 21729 110928   2
## 3 5.7 19517  55003   3
## 4 6.1 19588  48931   4
## 5 4.8 24400  58818   4
## 6 9.5 16803  38658   1
# Rename the columns
colnames(mydata) <- c("ID","Country", "State", "Area.2mile", "Population",
"Age18to34.pst","Age65over.pst", "Physician.no", "bed.no", "crime.no",
"hischool.pst","BS.pst", "Low.income.pst", "unemployment.pst",
"Per.income.dollar","TO.income.m", "Geo") # Rename the columns
head(mydata)
##   ID     Country State Area.2mile Population Age18to34.pst Age65over.pst
## 1  1 Los_Angeles    CA       4060    8863164          32.1           9.7
```

```
## 2  2      Cook   IL     946  5105067         29.2          12.4
## 3  3    Harris   TX    1729  2818199         31.3           7.1
## 4  4 San_Diego   CA    4205  2498016         33.5          10.9
## 5  5    Orange   CA     790  2410556         32.6           9.2
## 6  6     Kings   NY      71  2300664         28.3          12.4
##   Physician.no bed.no crime.no hischool.pst BS.pst Low.income.pst
## 1        23677  27700   688936         70.0   22.3           11.6
## 2        15153  21550   436936         73.4   22.8           11.1
## 3         7553  12449   253526         74.9   25.4           12.5
## 4         5905   6179   173821         81.9   25.3            8.1
## 5         6062   6369   144524         81.2   27.8            5.2
## 6         4861   8942   680966         63.7   16.6           19.5
##   unemployment.pst Per.income.dollar TO.income.m Geo
## 1              8.0             20786      184230   4
## 2              7.2             21729      110928   2
## 3              5.7             19517       55003   3
## 4              6.1             19588       48931   4
## 5              4.8             24400       58818   4
## 6              9.5             16803       38658   1
```

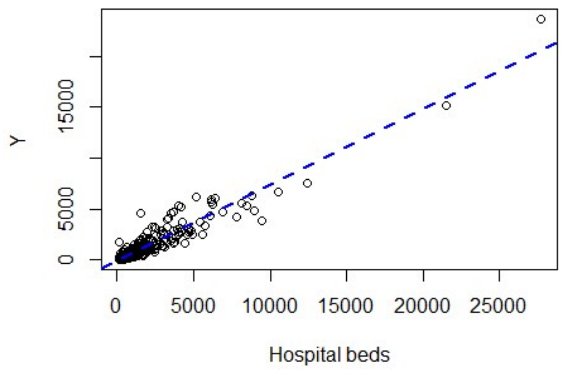
Fitted plot, linear regression model and possible outliers

```
# Fit the linear regression model
Y <- mydata$Physician.no

m1 <- lm(Y~ Population, data=mydata) #related to total population
plot(Y~ Population, data=mydata,
     xlab="Total polulation",
     ylab="The number of active physicians in a CDI")
abline(coef(m1), col="red", lty=3, lwd=2)
```

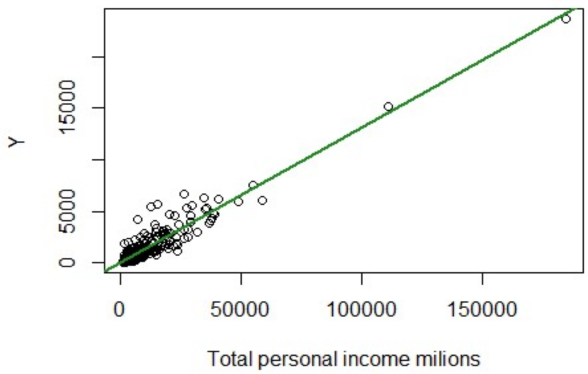



```
summary(m1)
##
## Call:
## lm(formula = Y ~ Population, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1969.4  -209.2   -88.0    27.9  3928.7
##
## Coefficients:
```

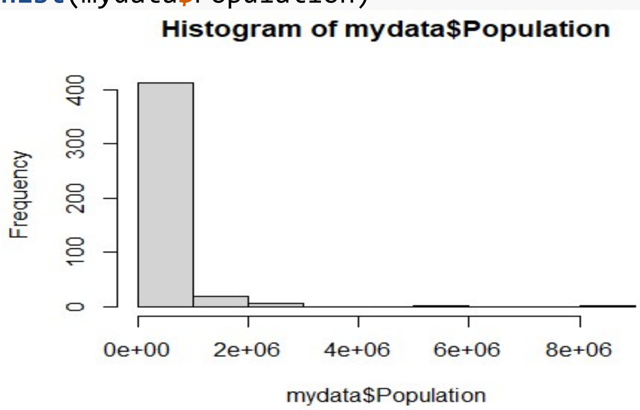
```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.106e+02  3.475e+01  -3.184  0.00156 **
## Population   2.795e-03  4.837e-05  57.793  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 610.1 on 438 degrees of freedom
## Multiple R-squared:  0.8841, Adjusted R-squared:  0.8838
## F-statistic:  3340 on 1 and 438 DF,  p-value: < 2.2e-16
m2 <- lm(Y~ bed.no, data=mydata) #related to number of hospital beds
plot(Y~ bed.no, data=mydata,
     xlab="Hospital beds")
abline(coef(m2), col="blue", lty=2, lwd=2)
```



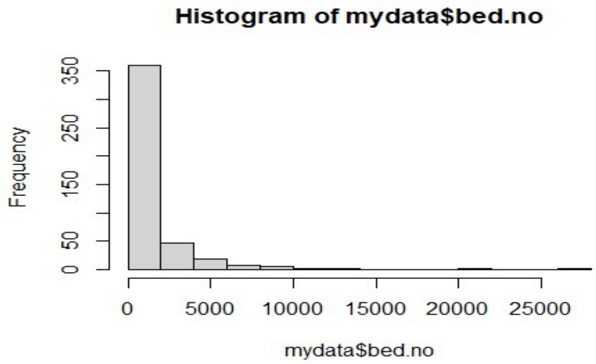

```
summary(m2)
##
## Call:
## lm(formula = Y ~ bed.no, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3133.2  -216.8   -32.0    96.2  3611.1
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept) -95.93218   31.49396  -3.046  0.00246 **
## bed.no        0.74312    0.01161  63.995  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 556.9 on 438 degrees of freedom
## Multiple R-squared:  0.9034, Adjusted R-squared:  0.9032
## F-statistic:  4095 on 1 and 438 DF,  p-value: < 2.2e-16
m3 <- lm(Y~ TO.income.m, data=mydata) #related to total personal income.
plot(Y~ TO.income.m, data=mydata,
     xlab="Total personal income milions")
abline(coef(m3), col="forestgreen",lty=1, lwd=2)
```



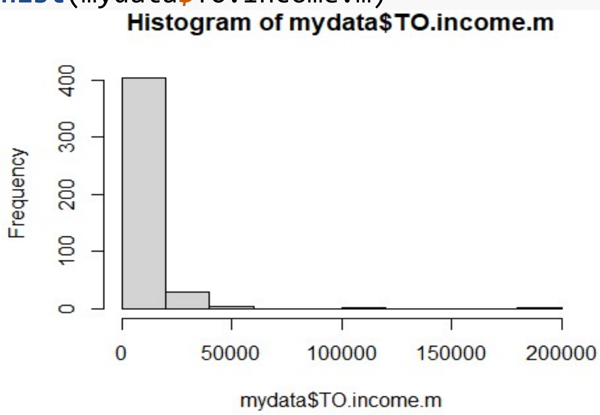

```
summary(m3)
##
## Call:
## lm(formula = Y ~ TO.income.m, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1926.6  -194.5   -66.6    44.2  3819.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -48.39485   31.83333   -1.52    0.129
## TO.income.m   0.13170    0.00211   62.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 569.7 on 438 degrees of freedom
## Multiple R-squared:  0.8989, Adjusted R-squared:  0.8987
## F-statistic:  3895 on 1 and 438 DF,  p-value: < 2.2e-16
# Possible outliners, by examining the histogram of each variable
hist(mydata$Population)
```




```
hist(mydata$bed.no)
```



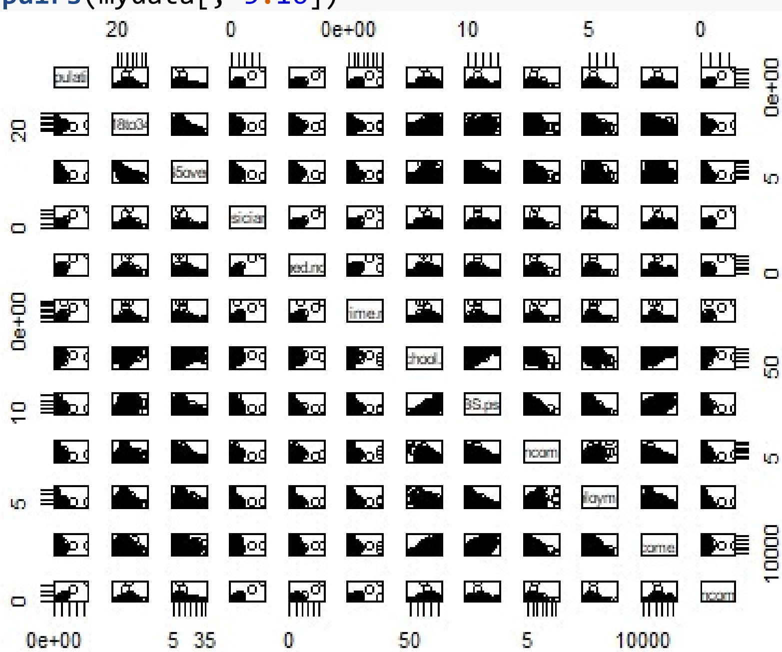

```
hist(mydata$TO.income.m)
```




```
# appears that there are a couple outliers
```

Fit a multiple regression model.

```
# scatter plot matrix
pairs(mydata[, 5:16])
```



Problems:

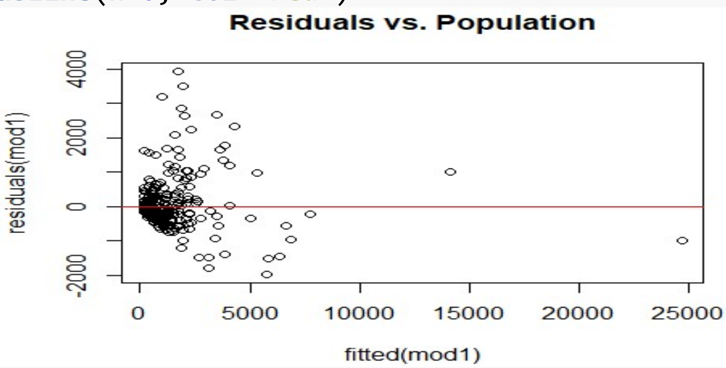
(1) Plot the residuals against the predictor (or Y-hat).

(a) Whether there seems to be a linear or non-linear relationship?

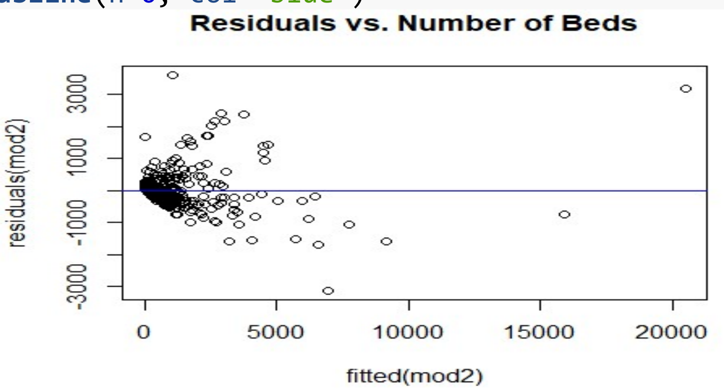
(b) Is the residuals heteroskedastic? variance (vertical spread) across the board

Funnel shape opening to the right: heteroskedasticity pattern, a Poisson distribution

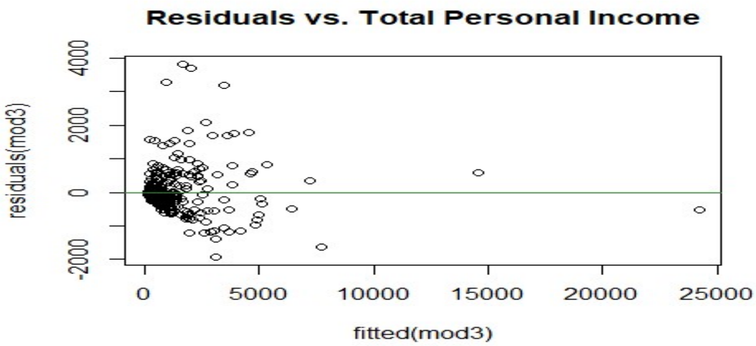
```r
# Residuals vs. Predictor for Model 1 (Population)
plot(fitted(mod1), residuals(mod1), main="Residuals vs. Population")
abline(h=0, col="red")
```



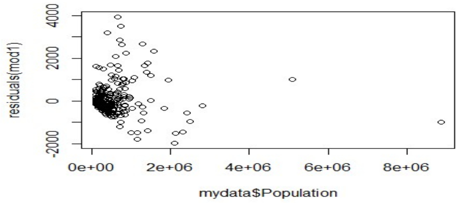

```r
# Residuals vs. Predictor for Model 2 (Number of Beds)
plot(fitted(mod2), residuals(mod2), main="Residuals vs. Number of Beds")
abline(h=0, col="blue")
```



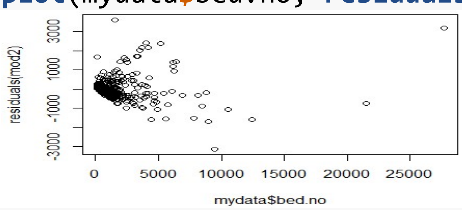

```r
# Residuals vs. Predictor for Model 3 (Total Personal Income)
plot(fitted(mod3), residuals(mod3), main="Residuals vs. Total Personal
Income")
abline(h=0, col="forestgreen")
```



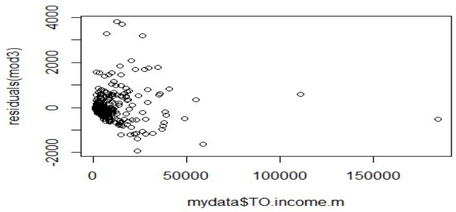

```
# Also, may plot the residuals against each X to see whether the nonlinear trend in the residuals might be potentially related to one of the Xs.
plot(mydata$Population, residuals(mod1))
```




```
plot(mydata$bed.no, residuals(mod2))
```

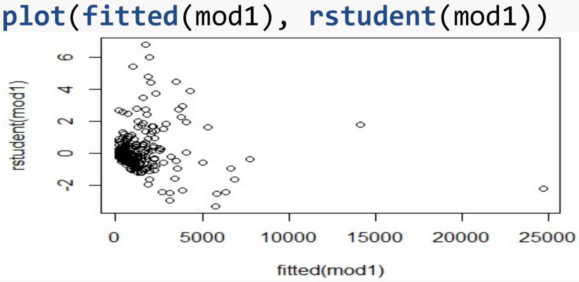



```
plot(mydata$TO.income.m, residuals(mod3))
```

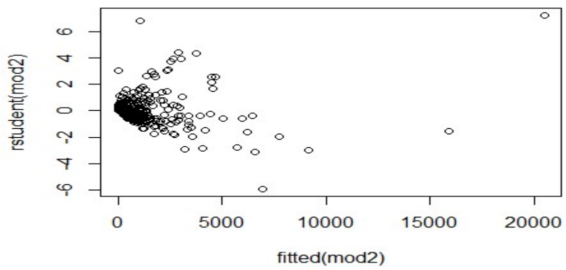


Regression outliers

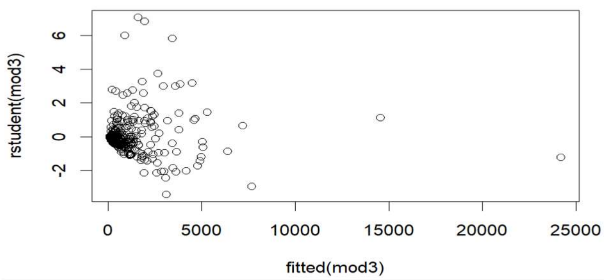
```
# Studentized residual plot
plot(fitted(mod1), rstudent(mod1))
```



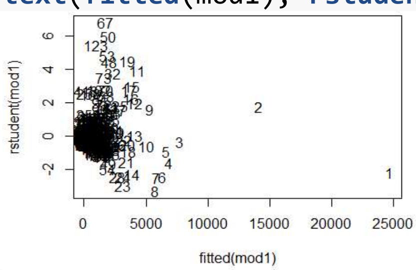

```
plot(fitted(mod2), rstudent(mod2))
```



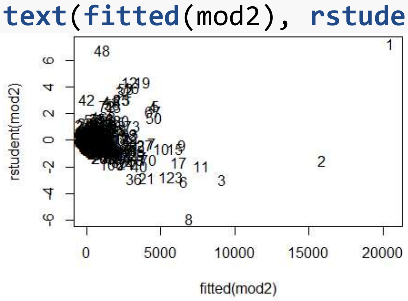

```
plot(fitted(mod2), rstudent(mod2))
```



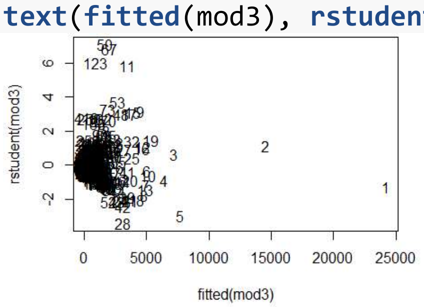

```
# find out which point is the outlier
plot(fitted(mod1), rstudent(mod1), type="n")
text(fitted(mod1), rstudent(mod1), names(rstudent(mod1)))
```




```
plot(fitted(mod2), rstudent(mod2), type="n")
text(fitted(mod2), rstudent(mod2), names(rstudent(mod2)))
```




```
plot(fitted(mod3), rstudent(mod3), type="n")
text(fitted(mod3), rstudent(mod3), names(rstudent(mod3)))
```

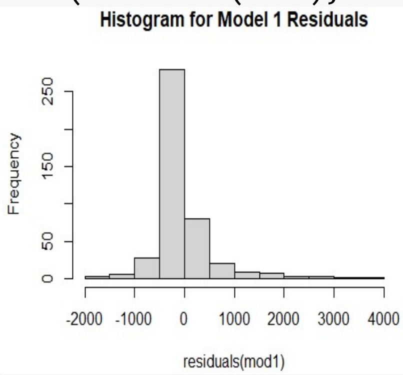


Problems:

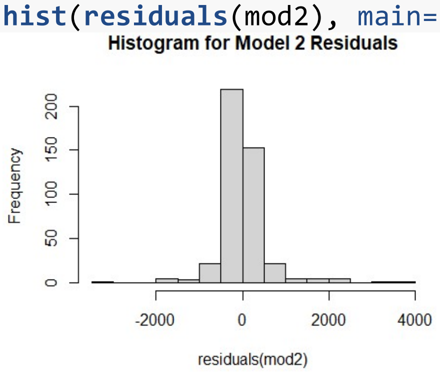
(2) Whether the residuals are normally distributed

Histogram or QQ-plot QQ-plot: The more straight line, the more normal Data do not fit well.

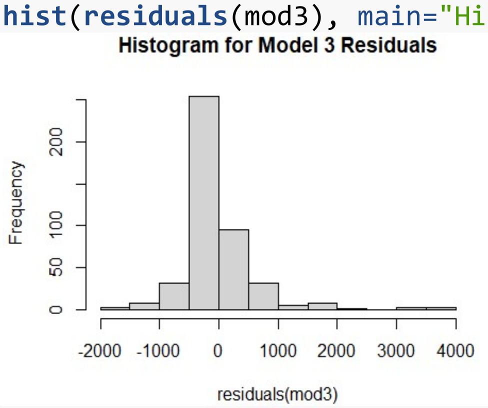
```
# Histogram
hist(residuals(mod1), main="Histogram for Model 1 Residuals")
```



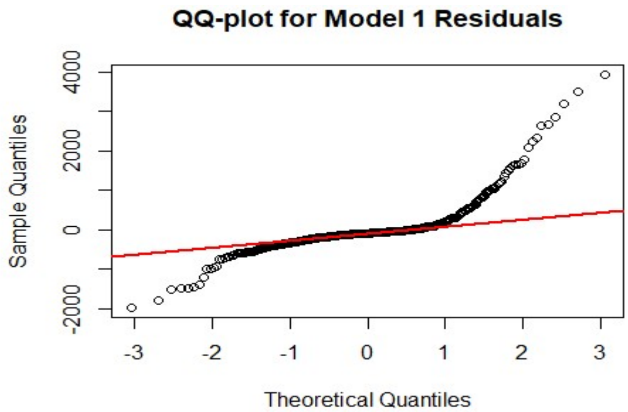

```
hist(residuals(mod2), main="Histogram for Model 2 Residuals")
```



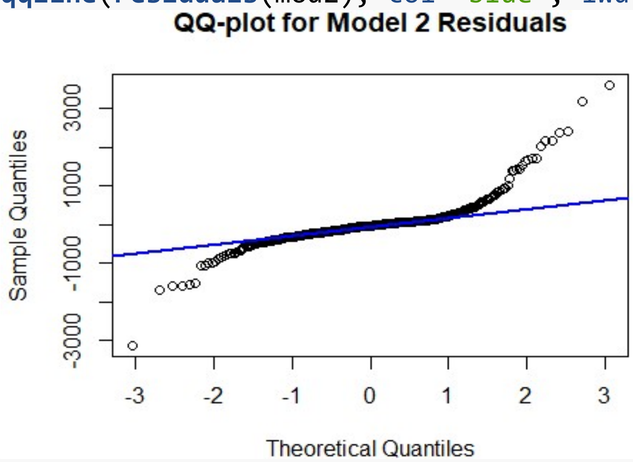

```
hist(residuals(mod3), main="Histogram for Model 3 Residuals")
```



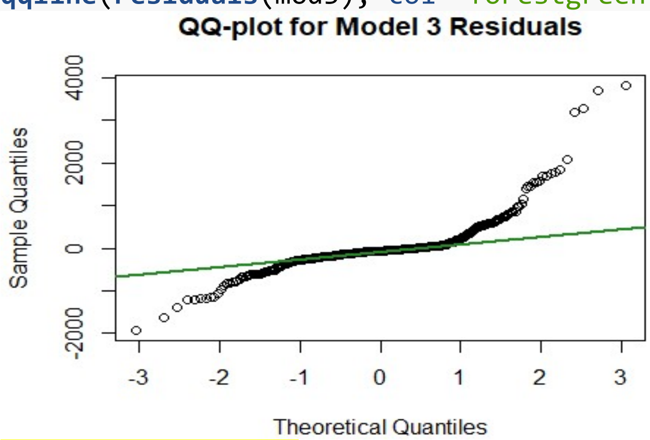

```
# Q-Q plot for model 1
qqnorm(residuals(mod1), main ="QQ-plot for Model 1 Residuals")
qqline(residuals(mod1), col="red", lwd=2)
```




```
# Q-Q plot for model 2
qqnorm(residuals(mod2), main ="QQ-plot for Model 2 Residuals")
qqline(residuals(mod2), col="blue", lwd=2)
```



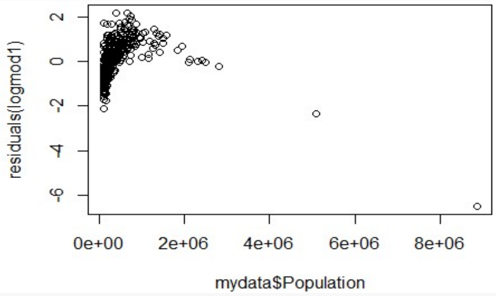

```
# Q-Q plot for model 3
qqnorm(residuals(mod3), main ="QQ-plot for Model 3 Residuals")
qqline(residuals(mod3), col="forestgreen", lwd=2)
```



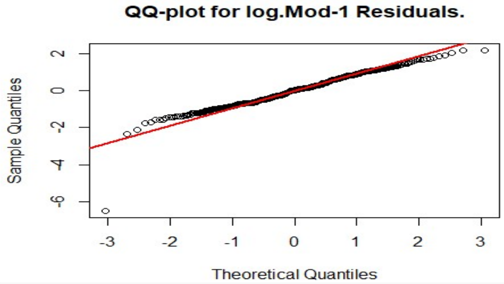
Log-regression models

```
#-----------------------------------------------
# Transform Y with log(), fit the new model, and repeat the diagnostics
mydata$log.Physician.no <- log(mydata$Physician.no)
logmod1 <- lm(log.Physician.no ~ Population, data=mydata,
```

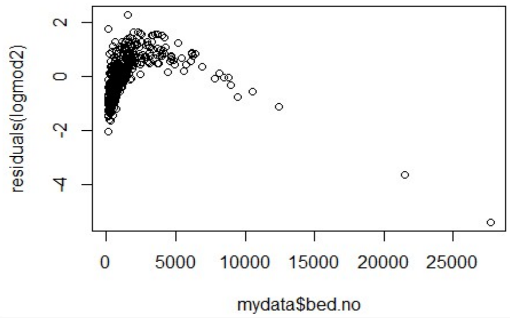
```
na.action=na.exclude)
plot(mydata$Population, residuals(logmod1))
```



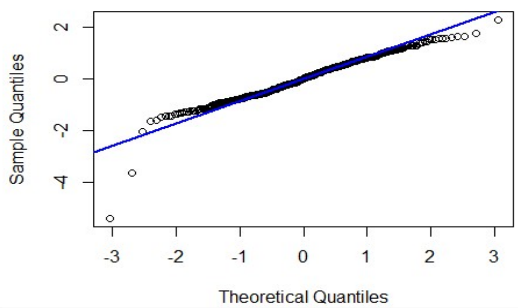

```
qqnorm(residuals(logmod1), main="QQ-plot for log.Mod-1 Residuals.")
qqline(residuals(logmod1), col="red", lwd=2)
```



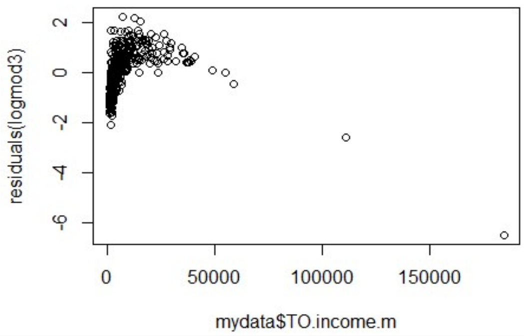

```
summary(logmod1)
##
## Call:
## lm(formula = log.Physician.no ~ Population, data = mydata, na.action =
na.exclude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5066 -0.6455  0.0228  0.6149  2.1839
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.668e+00  4.970e-02  114.05   <2e-16 ***
## Population  1.231e-06  6.918e-08   17.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8726 on 438 degrees of freedom
## Multiple R-squared:  0.4196, Adjusted R-squared:  0.4183
## F-statistic: 316.6 on 1 and 438 DF,  p-value: < 2.2e-16

logmod2 <- lm(log.Physician.no ~ bed.no, data=mydata, na.action=na.exclude)
plot(mydata$bed.no, residuals(logmod2))
```



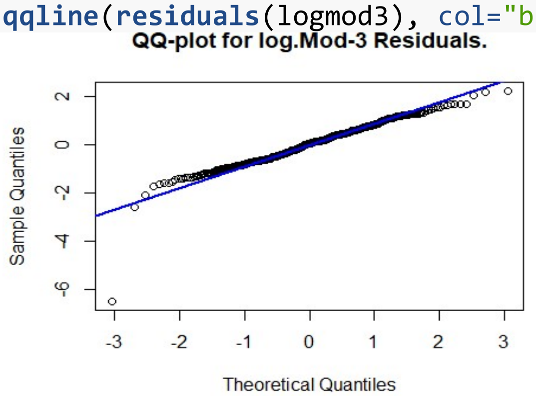

```
qqnorm(residuals(logmod2), main="QQ-plot for log.Mod-2 Residuals.")
qqline(residuals(logmod2), col="blue", lwd=2)
```




```
summary(logmod2)
##
## Call:
## lm(formula = log.Physician.no ~ bed.no, data = mydata, na.action =
na.exclude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3832 -0.5856  0.0171  0.5717  2.2725
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.635e+00  4.565e-02  123.44   <2e-16 ***
## bed.no      3.545e-04  1.683e-05   21.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8072 on 438 degrees of freedom
## Multiple R-squared:  0.5033, Adjusted R-squared:  0.5021
## F-statistic: 443.7 on 1 and 438 DF,  p-value: < 2.2e-16
logmod3 <- lm(log.Physician.no ~ TO.income.m, data=mydata,
na.action=na.exclude)
plot(mydata$TO.income.m, residuals(logmod3))
```



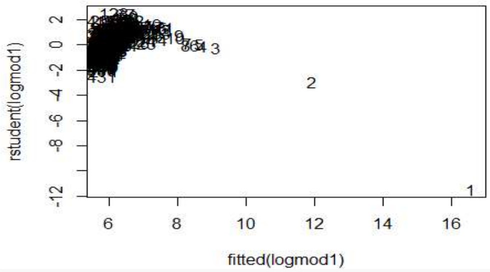

```
qqnorm(residuals(logmod3), main="QQ-plot for log.Mod-3 Residuals.")
qqline(residuals(logmod3), col="blue", lwd=2)
```



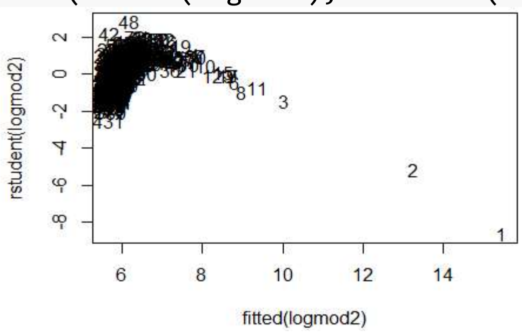

```
summary(logmod3)
##
## Call:
## lm(formula = log.Physician.no ~ TO.income.m, data = mydata, na.action =
na.exclude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5125 -0.6221  0.0349  0.5803  2.2289
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.686e+00  4.773e-02   119.1   <2e-16 ***
## TO.income.m 5.916e-05  3.164e-06    18.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8541 on 438 degrees of freedom
## Multiple R-squared:  0.4439, Adjusted R-squared:  0.4426
## F-statistic: 349.6 on 1 and 438 DF,  p-value: < 2.2e-16
```

Identify outliers

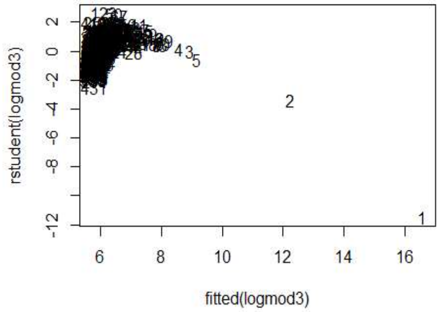
```
# identify the outlier
# logmod1
plot(fitted(logmod1), rstudent(logmod1), type='n')
text(fitted(logmod1), rstudent(logmod1), names(rstudent(logmod1)))
```




```
rstudent(logmod1)[1] # outlier check
##         1
## -11.49522
outliers_gt_2 <- rstudent(logmod1)[abs(rstudent(logmod1)) > 2] # outliers >2
print(outliers_gt_2)
##          1          2          48          50          53          67
73
## -11.495218  -2.902440   2.122138   2.339800   2.200644   2.491654
2.066237
##        123        271        431
##   2.520957  -2.036103  -2.459259
# logmod2
plot(fitted(logmod2), rstudent(logmod2), type='n')
text(fitted(logmod2), rstudent(logmod2), names(rstudent(logmod2)))
```



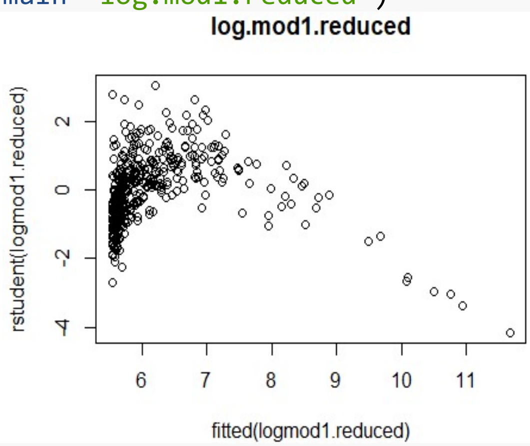

```
rstudent(logmod2)[1] # outlier check
##         1
## -8.621855
outliers_gt_2 <- rstudent(logmod2)[abs(rstudent(logmod2)) > 2] # outliers >2
print(outliers_gt_2)
##         1         2        41        42        48        72       380
431
## -8.621855 -5.126931  2.037006  2.183433  2.840967  2.007360 -2.033400 -
2.531219
# logmod3
plot(fitted(logmod3), rstudent(logmod3), type='n')
text(fitted(logmod3), rstudent(logmod3), names(rstudent(logmod3)))
```



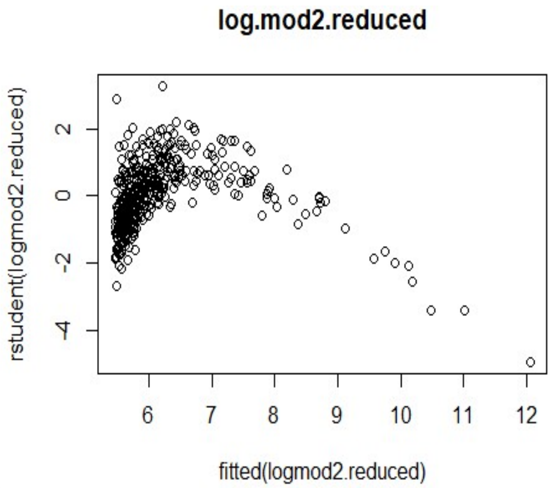

```
rstudent(logmod3)[1] # outlier check
##         1
## -11.50474
outliers_gt_2 <- rstudent(logmod3)[abs(rstudent(logmod3)) > 2] # outliers >2
print(outliers_gt_2)
##          1          2          50          67         123         271
431
## -11.504736  -3.365263   2.553202   2.414853   2.630120  -2.043067  -
2.501032
# Remove the outliers discovered and call the new data as XX.reduced
mydata.reduced <- mydata[-c(1, 2),] #remove two outliers
```

Log-regression models on the 'reduced' data

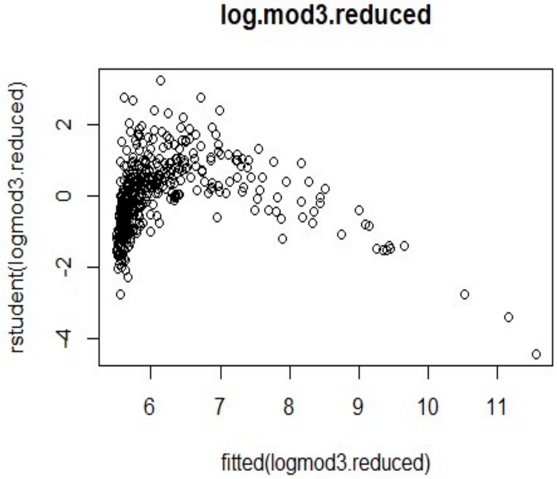
```
# Fit the log-transformed model on the reduced data
logmod1.reduced <- lm(log(Physician.no) ~ Population, data=mydata.reduced,
na.action=na.exclude)
logmod2.reduced <- lm(log(Physician.no) ~ bed.no, data=mydata.reduced,
na.action=na.exclude)
logmod3.reduced <- lm(log(Physician.no) ~ TO.income.m, data=mydata.reduced,
na.action=na.exclude)

# Residual plot of the log-transformed model
plot(fitted(logmod1.reduced), rstudent(logmod1.reduced),
main="log.mod1.reduced")
```




```
plot(fitted(logmod2.reduced), rstudent(logmod2.reduced),
main="log.mod2.reduced")
```



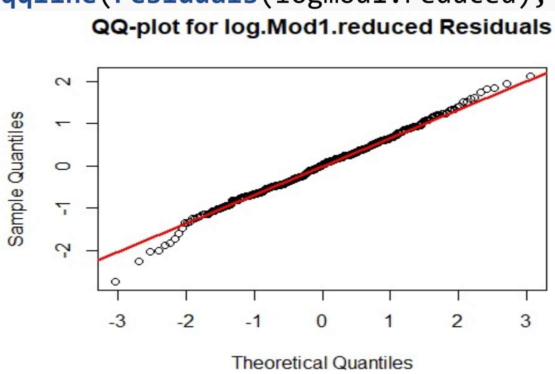

```
plot(fitted(logmod3.reduced), rstudent(logmod3.reduced),
main="log.mod3.reduced")
```



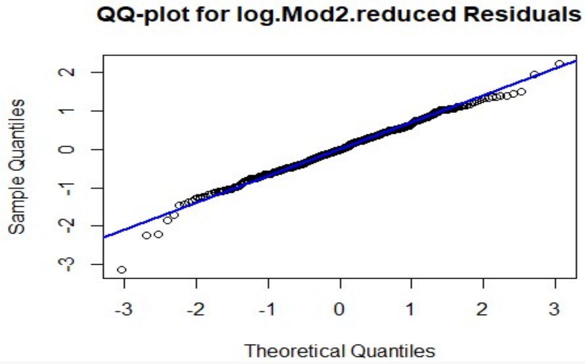
Q-Q plot of the log-transformed model

It looks better now.

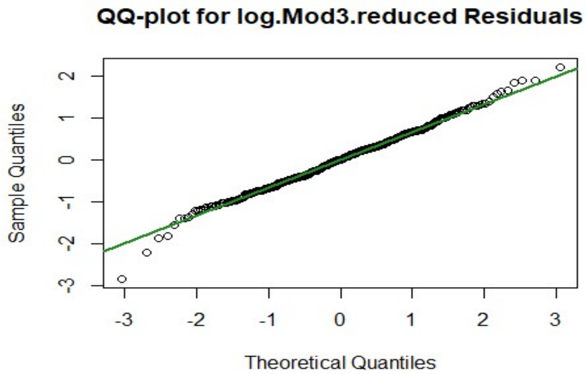
```
# Q-Q plot of the log-transformed model
qqnorm(residuals(logmod1.reduced), main="QQ-plot for log.Mod1.reduced
Residuals")
qqline(residuals(logmod1.reduced), col="red", lwd=2)
```



```
qqnorm(residuals(logmod2.reduced), main="QQ-plot for log.Mod2.reduced
Residuals")
qqline(residuals(logmod2.reduced), col="blue", lwd=2)
```




```
qqnorm(residuals(logmod3.reduced), main="QQ-plot for log.Mod3.reduced
Residuals")
qqline(residuals(logmod3.reduced), col="forestgreen", lwd=2)
```



summary table

```
summary(logmod1.reduced)
##
## Call:
## lm(formula = log(Physician.no) ~ Population, data = mydata.reduced,
##     na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74350 -0.47202  0.02536  0.43964  2.12917
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.316e+00  4.642e-02  114.53   <2e-16 ***
## Population  2.256e-06  8.780e-08   25.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7065 on 436 degrees of freedom
```

```
## Multiple R-squared:  0.6022, Adjusted R-squared:  0.6013
## F-statistic: 660.1 on 1 and 436 DF,  p-value: < 2.2e-16

summary(logmod2.reduced)
##
## Call:
## lm(formula = log(Physician.no) ~ bed.no, data = mydata.reduced,
##     na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.12684 -0.46736 -0.00468  0.47276  2.22425
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.413e+00  4.222e-02  128.20   <2e-16 ***
## bed.no      5.337e-04  1.974e-05   27.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6846 on 436 degrees of freedom
## Multiple R-squared:  0.6265, Adjusted R-squared:  0.6256
## F-statistic: 731.2 on 1 and 436 DF,  p-value: < 2.2e-16

summary(logmod3.reduced)
##
## Call:
## lm(formula = log(Physician.no) ~ TO.income.m, data = mydata.reduced,
##     na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.84951 -0.45494 -0.00187  0.44242  2.21022
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.374e+00  4.316e-02  124.51   <2e-16 ***
## TO.income.m 1.052e-04  3.892e-06   27.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6849 on 436 degrees of freedom
## Multiple R-squared:  0.6261, Adjusted R-squared:  0.6252
## F-statistic:   730 on 1 and 436 DF,  p-value: < 2.2e-16
```