Residual Plots

Pranav Gopalkrishna, 1940223

2021-09-23

# INITIAL SET-UP

library(ggplot2)  
setwd("/Users/pranav/Documents/Study/computerScience/programming/r/data")

# DATA SET

Not much is mentioned about this data set. But from what can be learned and assumed from the given data, the following data presents observations made on a sample of patients. The fields of observation include

* Age of various patients
* Severity (of their illness, presumably)
* Their measured / self-indicated anxiety levels (in general, presumably)
* Their measured / self-indicated satisfaction levels (in general, presumably)

In this assignment, I am going to study the linear relationship, if it exists, between satisfaction and the other variables. Satisfaction here is considered the response variable, while the others are potential factors.

dat = read.csv("patientSatisfaction.csv")  
head(dat)

## Satisfaction Age Severity Anxiety  
## 1 68 55 50 2.1  
## 2 77 46 24 2.8  
## 3 96 30 46 3.3  
## 4 80 35 48 4.5  
## 5 43 59 58 2.0  
## 6 44 61 60 5.1

y = dat$Satisfaction

# FUNCTION FOR PLOTTING THE GRAPHS FOR A MODEL

## Function for showing residual plots for each regression model

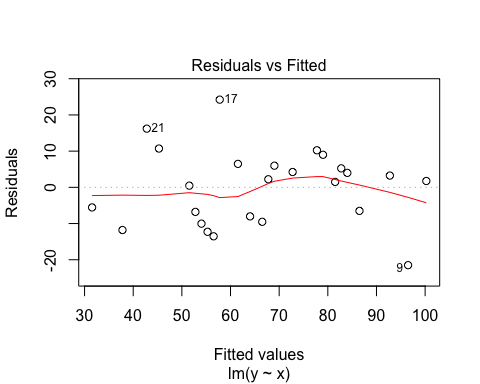
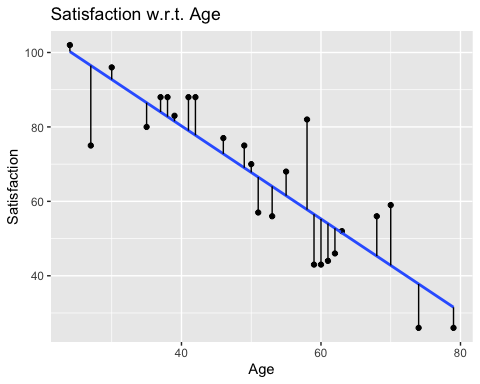
modelPlots = function(x, y, xName, yName)  
{  
 # MODEL  
 y\_x\_model = lm(y~x)  
 y\_x\_model  
 # MODEL SUMMARY  
 summary(y\_x\_model)  
 # DATA FRAME FROM MODEL  
 y\_fitted = y\_x\_model$coefficients[1] + y\_x\_model$coefficients[2]\*x  
 y\_x\_dat = data.frame(y, x, y\_fitted)  
 # PLOTS  
 # Fitted vs. actual  
 plotName = paste(yName, "w.r.t.", xName)  
 y\_x\_plot = ggplot(y\_x\_dat, aes(x, y))  
 y\_x\_plot = y\_x\_plot + geom\_point()  
 y\_x\_plot = y\_x\_plot + geom\_smooth(aes(x, y\_fitted))  
 y\_x\_plot = y\_x\_plot + geom\_segment(aes(xend = x, yend = y\_fitted))  
 y\_x\_plot = y\_x\_plot + labs(title = plotName, x = xName, y = yName)  
 print(y\_x\_plot)  
 # Residual plots  
 plot(y\_x\_model)  
}

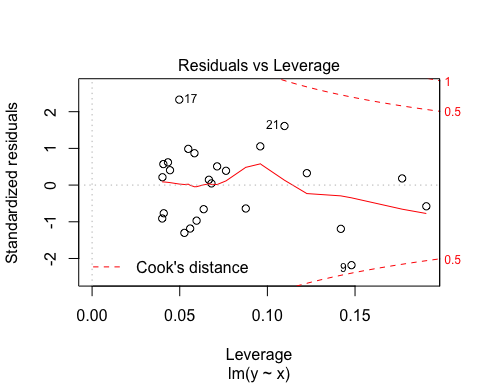
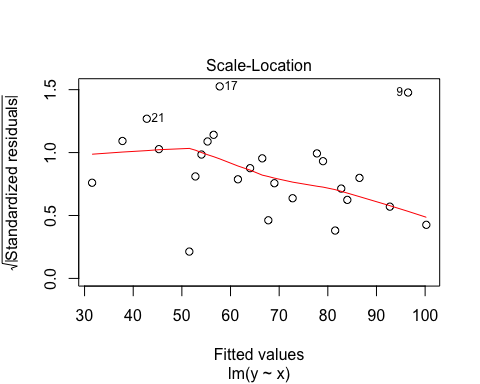
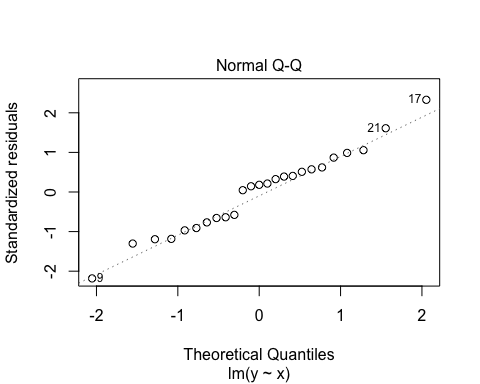
## Finding plots for every model

### 1. Satisfaction w.r.t. age

modelPlots(dat$Age, y, "Age", "Satisfaction")

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'





Residual vs. fitted plot shows inward closing funnel, implying that the variance of the standardized residuals is not constant for all response values.

Q-Q plot shows ideal probability distribution, hence we may conclude that

* The standardized residuals are normally distributed
* The variables are linearly associated

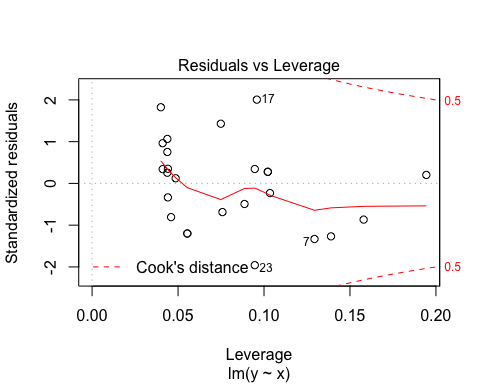
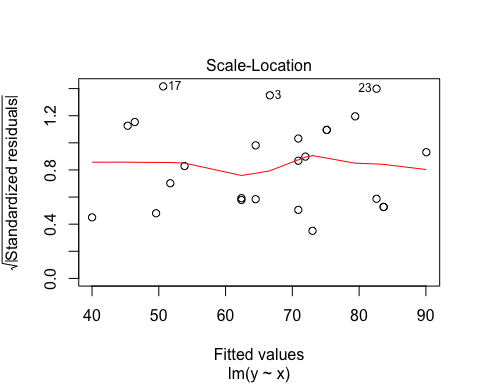
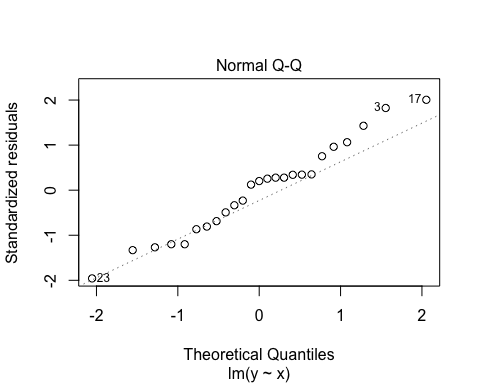
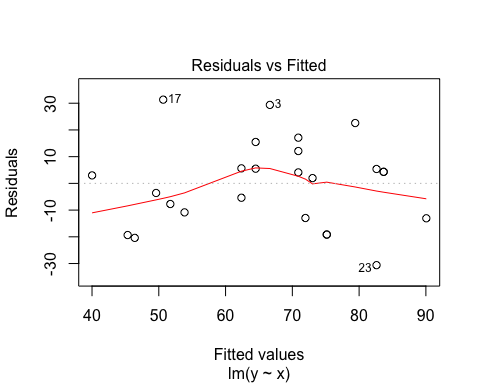
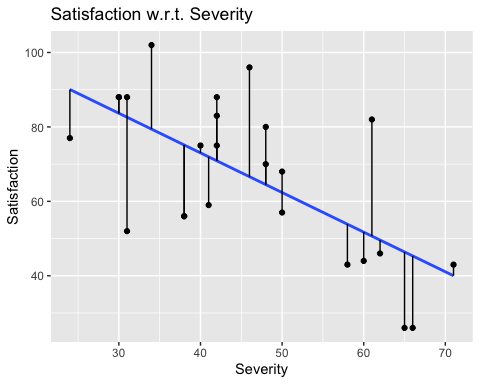
Scale-location plot shows inward closing funnel, implying that the variance of the standardized residuals is not constant for all predictor values (ages).

Residuals vs. leverage plot shows that all standardized residuals are within Cook's distance, implying that there are no excessively influential data values.

### **2. Satisfaction w.r.t severity**

modelPlots(dat$Severity, y, "Severity", "Satisfaction")

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



Residual vs. fitted plot shows randomly distributed points, implying that the variance of the standardized residuals is constant for all response values.

Q-Q plot shows mostly ideal probability distribution, with a temporarily increased slope at the top extreme, hence we may conclude that

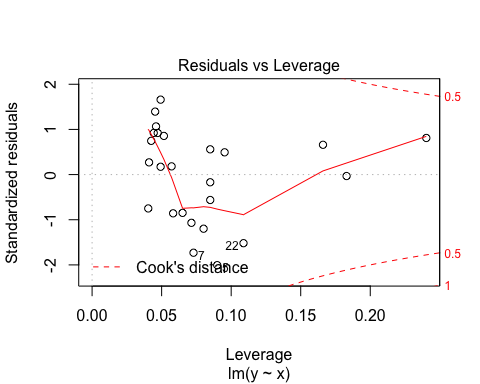
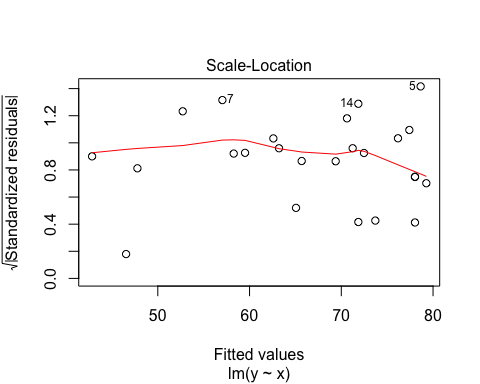
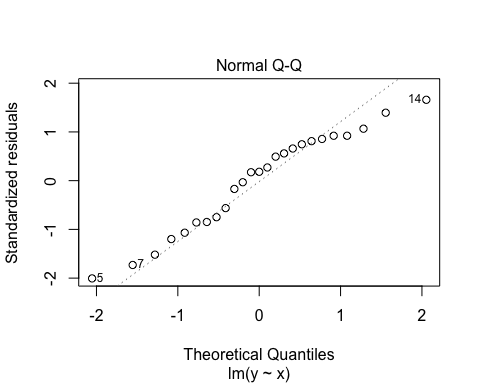
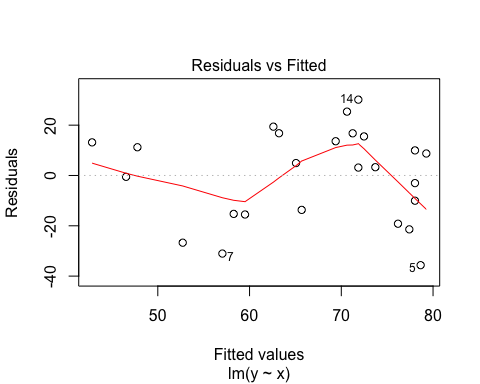
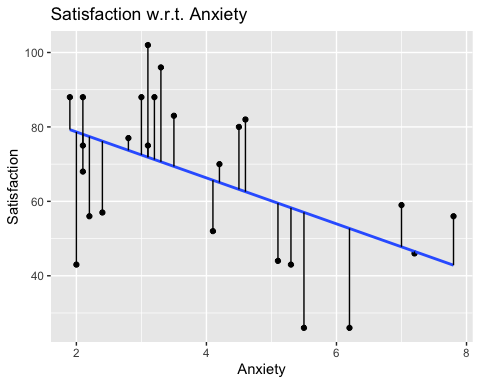
* The standardized residuals are mostly normally distributed, with some possible positive skewness
* The variables are generally linearly associated

Scale-location plot shows randomly distributed points, implying that the variance of the standardized residuals is equal for every response variable for all predictor values (severity levels).

Residuals vs. leverage plot shows that all standardized residuals are within Cook's distance, implying that there are no excessively influential data values.

modelPlots(dat$Anxiety, y, "Anxiety", "Satisfaction")

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



Residual vs. fitted plot shows randomly distributed points, implying that the variance of the standardized residuals is constant for all response values.

Q-Q plot shows mostly ideal probability distribution, with a temporarily increased slope at the top extreme, hence we may conclude that

* The standardized residuals are mostly normally distributed, with some possible positive skewness
* The variables are generally linearly associated

Scale-location plot shows randomly distributed points, implying that the variance of the standardized residuals is equal for every response variable for all predictor values (anxiety levels).

Residuals vs. leverage plot shows that all standardized residuals are within Cook's distance, implying that there are no excessively influential data values.