Linear Regression in R

Manbir Singh Panesar

10/25/2021

The data file dataxy.txt on quercus contains a data set with two variables x and y. We will use this dataset to investigate heteroscedasticity (i.e. non-conatant variance). (a) Fit a simple linear regression model for y on x and obtain the residuals for the model. Make a plot of residuals versus x and a scatterplot of y versus x. Comment on your plots.

```
library(readr)
dataxy <- read.table("dataxy.txt", header = TRUE)
## Dataset ##
dataxy
```

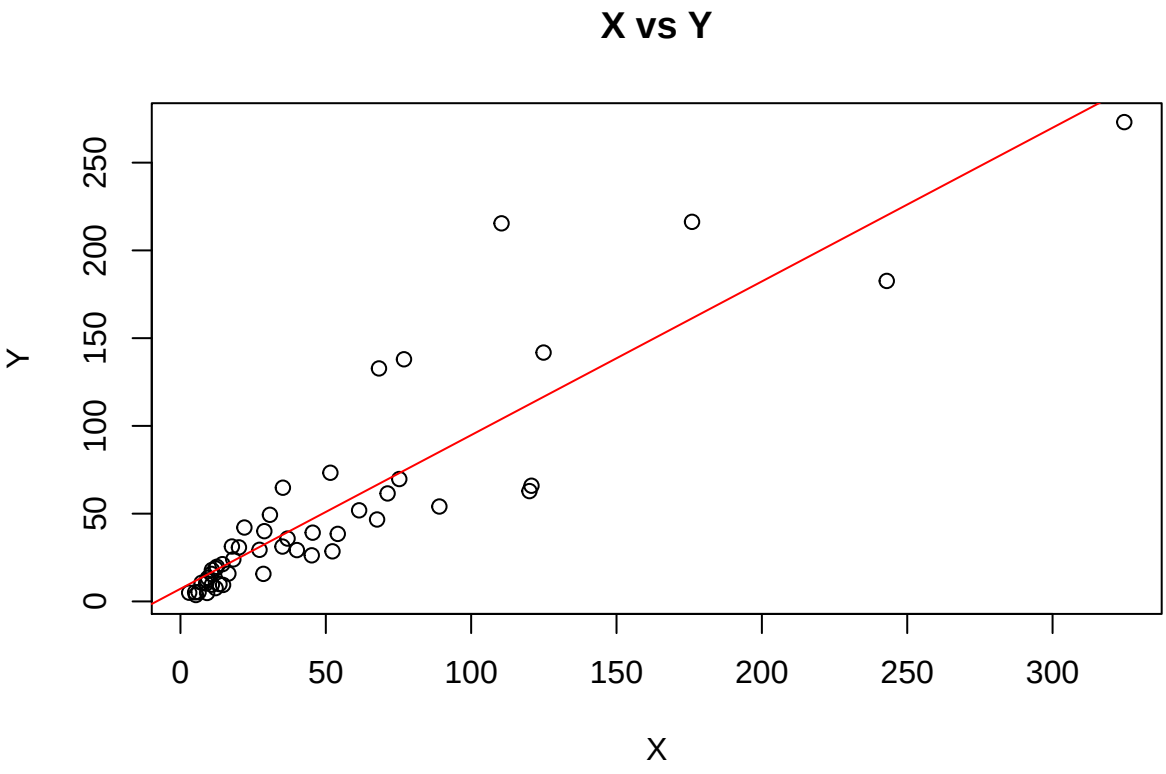
```
##             x          y
## 1  124.899403 141.794859
## 2   76.884554 137.955281
## 3   10.824256  17.907235
## 4  324.667916 273.097965
## 5  120.045005  62.860476
## 6  120.745675  65.902874
## 7  110.444777 215.424127
## 8    9.461987  13.373291
## 9   71.243568  61.510482
## 10  40.049586  29.241211
## 11  14.440574  21.226351
## 12  54.109919  38.470751
## 13  28.525079  15.696156
## 14  89.062818  54.068128
## 15  75.274861  69.720369
## 16  35.088710  31.241327
## 17  16.479621  15.827016
## 18  51.574396  73.311107
## 19   6.241330   5.287396
## 20  17.693693  31.276318
## 21  52.286436  28.527686
## 22   8.827452  10.227488
## 23  20.112870  30.781690
## 24  12.072060  18.796012
## 25   9.108619  12.007274
## 26 242.963123 182.595707
## 27  28.850766  40.021504
## 28  10.805967   9.517010
## 29  14.655687   9.379913
## 30  30.775198  49.349015
## 31  18.178231  23.948973
## 32  67.631707  46.677725
```

```
## 33  10.606957  15.670239
## 34  61.462907  51.900010
## 35   5.308919   3.644295
## 36   2.996053   4.947110
## 37   7.273225  10.619005
## 38  13.422811   9.854998
## 39  36.839086  35.799592
## 40  12.091897   7.579147
## 41  68.284269 132.749052
## 42  45.481351  39.167071
## 43  27.187009  29.393401
## 44  45.156746  26.256211
## 45  12.501935  19.729818
## 46   5.049055   5.184491
## 47   9.164551   4.806023
## 48 175.980735 216.278197
## 49  35.231971  64.788443
## 50  21.987999  42.095776
```

```r
## Plotting the data ##
x <- dataxy$x
y <- dataxy$y

plot(x, y, main = "X vs Y",
     xlab = "X", ylab = "Y")

## Creating Linear Regression Model and plotting ##
linreg <- lm(y ~ x)
abline(linreg, col = "red")
```



```r
summary(linreg)
```

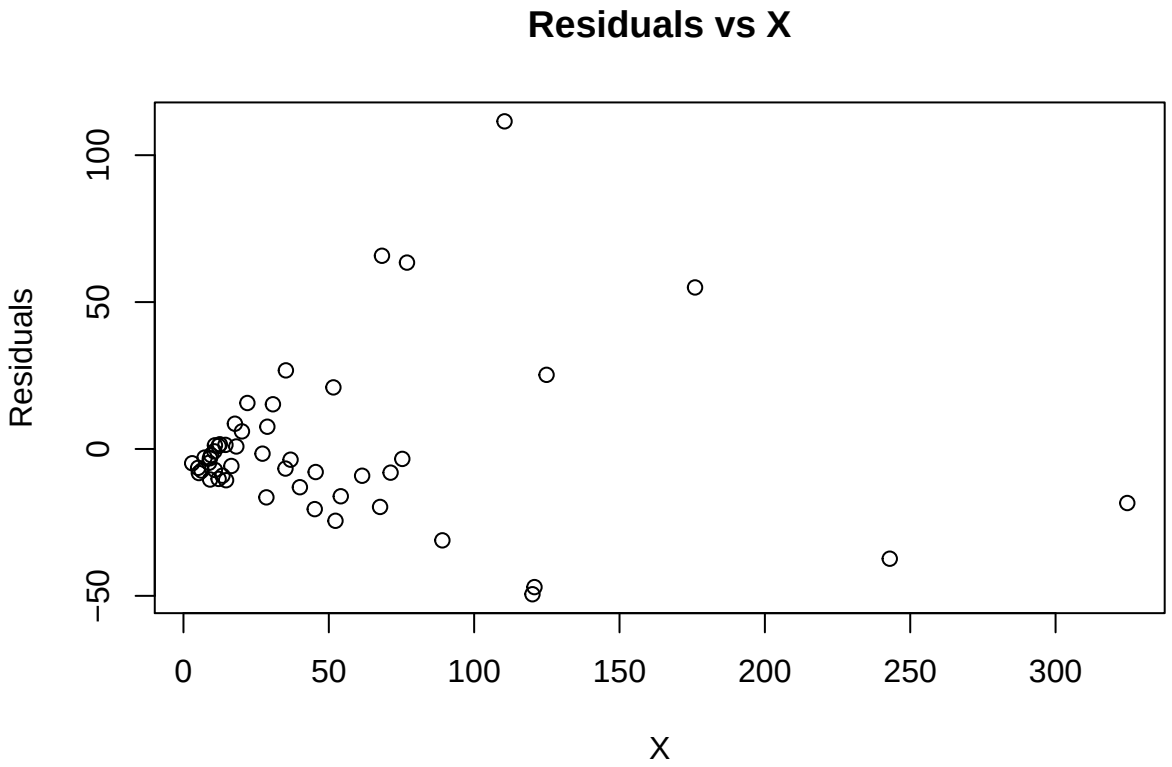
```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.451 -10.356  -4.778   1.543 111.519
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.18891    5.05175   1.423    0.161
## x            0.87569    0.06368  13.752   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.66 on 48 degrees of freedom
## Multiple R-squared:  0.7976, Adjusted R-squared:  0.7933
## F-statistic: 189.1 on 1 and 48 DF,  p-value: < 2.2e-16
```

The linear model appears to have a moderately strong relation with b1 being 0.875. It has a positive direction indicating positive relation and is of linear form.

```r
## plotting the residual ##
dataxy$residuals <- residuals.lm(linreg)
dataxy$residuals
```

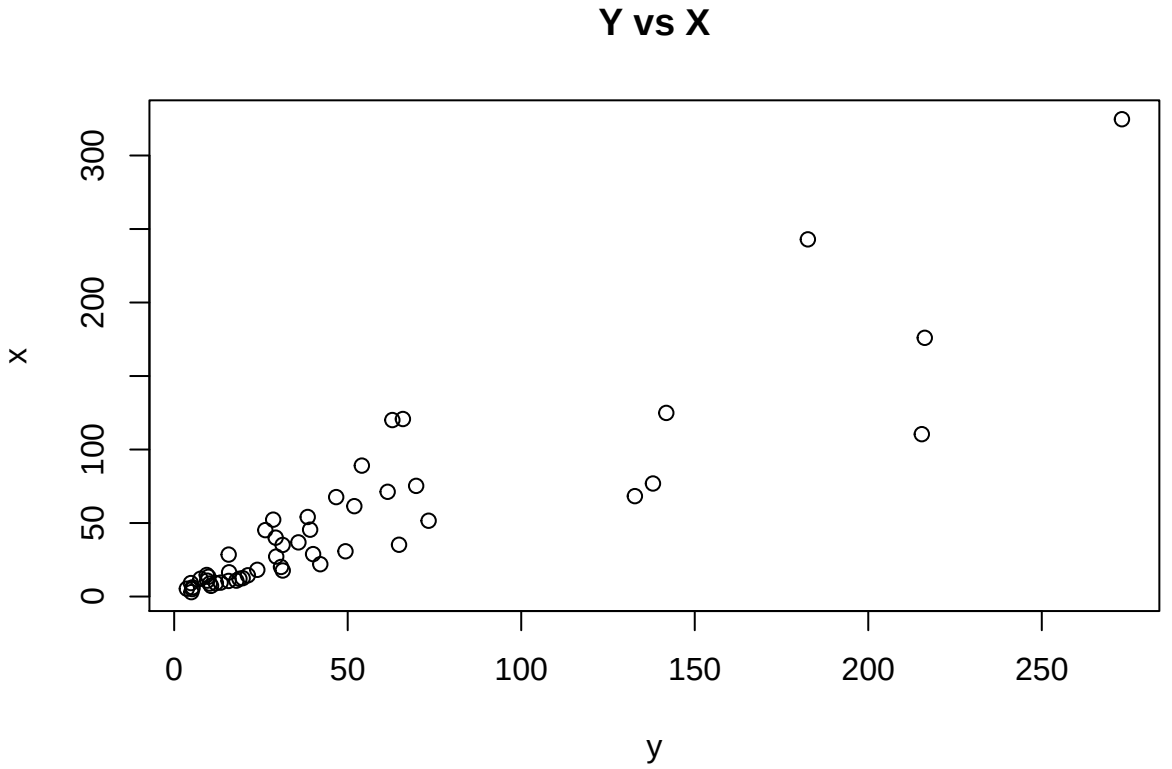
```
##  [1]  25.2322161  63.4389827   1.2395839 -18.4008888 -49.4511968 -47.0223717
##  [7] 111.5193222  -2.1014285  -8.0660352 -13.0189042   1.3919097 -16.1019221
## [13] -16.4720105 -31.1126105  -3.3863298  -6.6745758  -5.7930078  20.9587777
## [19]  -7.3670115   8.5931379 -24.4481728  -4.6915726   5.9800496   1.0356656
## [25]  -3.1580031 -37.3546996   7.5681351  -7.1346256 -10.6429016  15.2104314
## [31]   0.8414856 -19.7359050  -0.8071246  -9.1116351  -8.1936052  -4.8654259
## [37]  -2.9390274  -9.0881938  -3.6491057 -10.1985705  65.7639770  -7.8496116
## [43]  -1.6030248 -20.4762167   1.5930324  -6.4258477 -10.4082334  54.9839068
## [49]  26.7470873  15.6520952
```

```r
## residual vs x ##
plot(x, dataxy$residuals, main = "Residuals vs X", xlab = "X", ylab = "Residuals")
```



The residuals against x have most of their points around Residual = 0 indicating that the variance of the residuals is less.

```
## Scatterplot of y vs x ##
plot(y, x, main = 'Y vs X', xlab = "y", ylab = "x")
```



Likewise the if a line were to be fitted for y against x we would see strong strength given the close clusters of data

points.