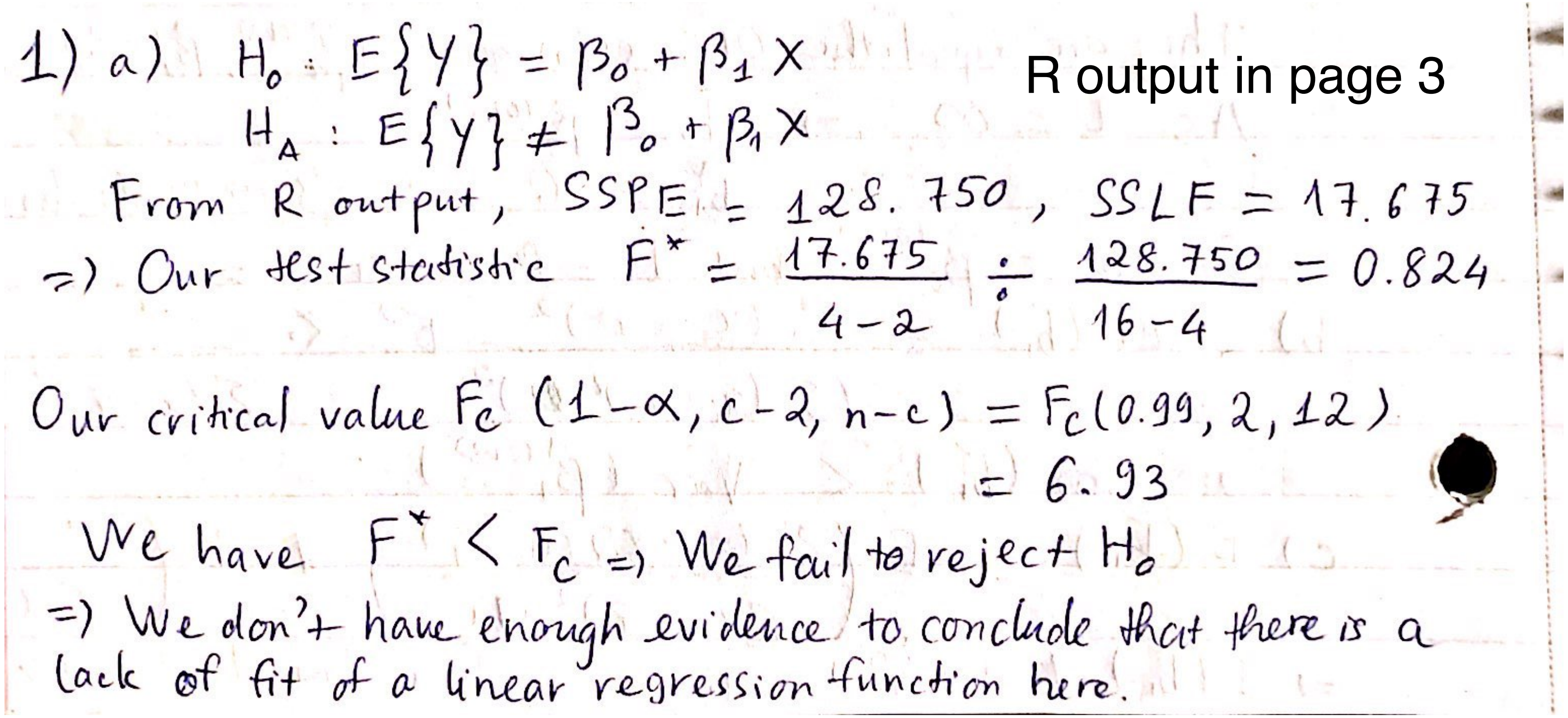
Giang Vu - 4445745
STAT 2131
Sep 3, 2020

Homework 3



R output in page 3

1) a) $H_0 : E\{Y\} = \beta_0 + \beta_1 X$

$H_A : E\{Y\} \neq \beta_0 + \beta_1 X$

From R output, $SSPE = 128.750$, $SSLF = 17.675$

$\Rightarrow$ Our test statistic $F^* = \frac{17.675}{4-2} \div \frac{128.750}{16-4} = 0.824$

Our critical value $F_c(1-\alpha, c-2, n-c) = F_c(0.99, 2, 12)$
$= 6.93$

We have $F^* < F_c \Rightarrow$ We fail to reject H_0

$\Rightarrow$ We don't have enough evidence to conclude that there is a lack of fit of a linear regression function here.

Scanned with CamScanner

b) Having an equal number of replications at each of the X levels generally doesn't affect our F test much as no matter how many replications there are in each level, we are taking the sum of squared difference of their value from the level's mean anyway. The exception is when there is no replications in one level then that level won't have any effect on the calculation of our SSPE.

c) If (a) concludes that the regression function is not linear then We can use the same lack-of-fit test to apply to other non-linear models to see which is the appropriate for our data and our study.

hw3

HOMEWORK 3

1) a)

```
#read data
hw2_data <- read.csv("/Users/giangvu/Desktop/STAT 2131 - Applied Stat Methods 1/HW/hw2/hw2_data.csv",
                     header = T,sep = ",")

#linear regression model
hw2_model <- lm(Y ~ X, data = hw2_data)

#SSE - RESIDUAL SUM OF SQUARES (or SSE(R))
ssr = sum((fitted(hw2_model) - hw2_data$Y)**2)
ssr
```

```
## [1] 146.425
```

```
#SSPE or SSE(F)
hw2_data <- hw2_data %>% mutate(level=as.numeric(factor(X)))
hw2_data <- hw2_data %>% group_by(level) %>% mutate(lvl_mean=mean(Y))
hw2_data <- hw2_data %>% mutate(lvl_err_sqr=(Y-lvl_mean)**2)
sspe <- sum(hw2_data$lvl_err_sqr)
sspe
```

```
## [1] 128.75
```

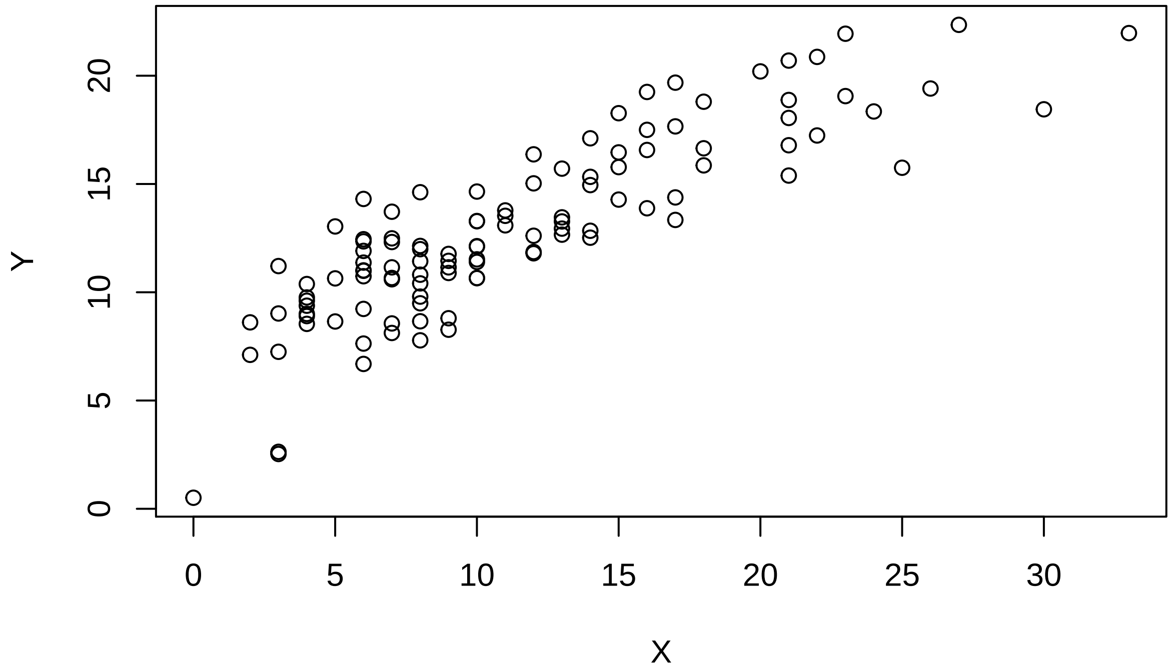
```
#SSLF = SSE - SSPE
sslf <- ssr - sspe
sslf
```

```
## [1] 17.675
```

2)

a)

```
#read data
hw3_data <- read.csv("/Users/giangvu/Desktop/STAT 2131 - Applied Stat Methods 1/HW/hw3/hw3_data.csv",
                     header = T,sep = ",")
#scatter plot
plot(Y ~ X, data = hw3_data)
```



From the scatter plot generated above, there doesn't seem to be a linear relationship between our original X and Y. If we try to fit a line through the points, the line would be a non-linear curve. Therefore, some transformation of our variables would be necessary before we want to run a linear regression with the data.

b)

```
#add a column for transformed X, call it X2
hw3_data <- hw3_data %>% mutate(X2 = sqrt(X))
#linear regression of Y on X2
hw3_model <- lm(Y ~ X2, data = hw3_data)
sm3 <- summary(hw3_model)
sm3
```

```
##
## Call:
## lm(formula = Y ~ X2, data = hw3_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0008 -1.2161  0.0383  1.3367  4.1795
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.2547     0.6389   1.964   0.0521 .
## X2            3.6235     0.1895  19.124   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.99 on 109 degrees of freedom
## Multiple R-squared:  0.7704, Adjusted R-squared:  0.7683
## F-statistic: 365.7 on 1 and 109 DF,  p-value: < 2.2e-16
```

```r
anova(hw3_model)
```

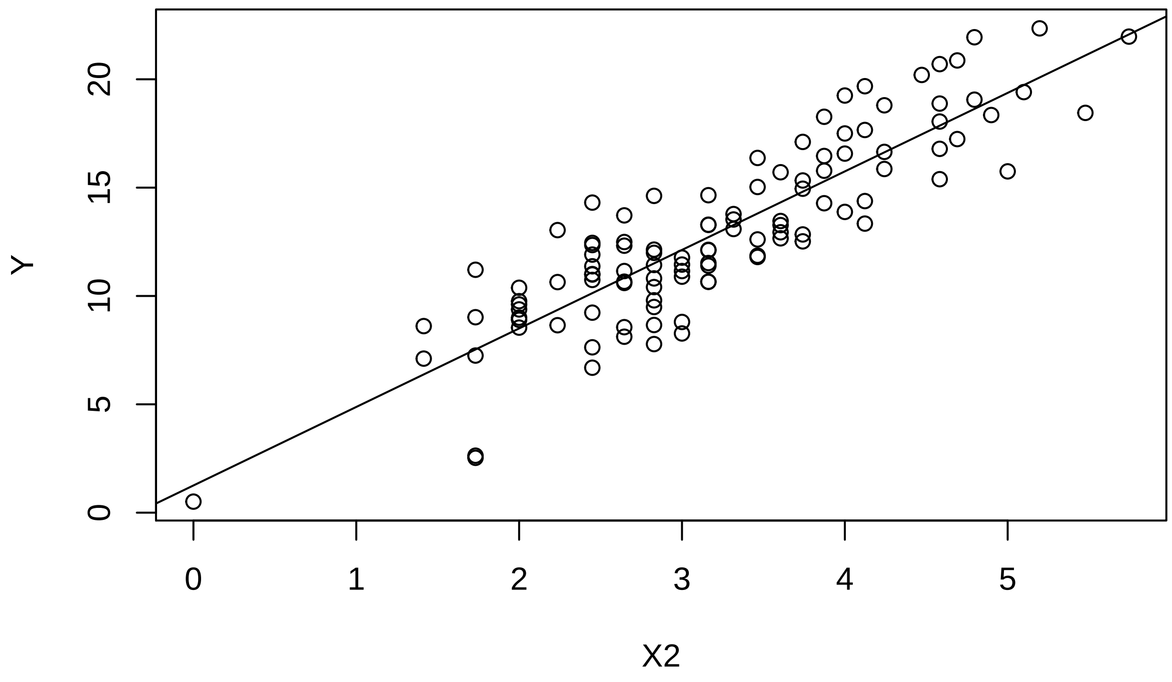
```
## Analysis of Variance Table
##
## Response: Y
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## X2          1 1448.33 1448.33  365.72 < 2.2e-16 ***
## Residuals 109  431.67    3.96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated linear regression function for the transformed data is

$$\hat{Y} = 1.2547 + 3.6235X^{'}$$

c)

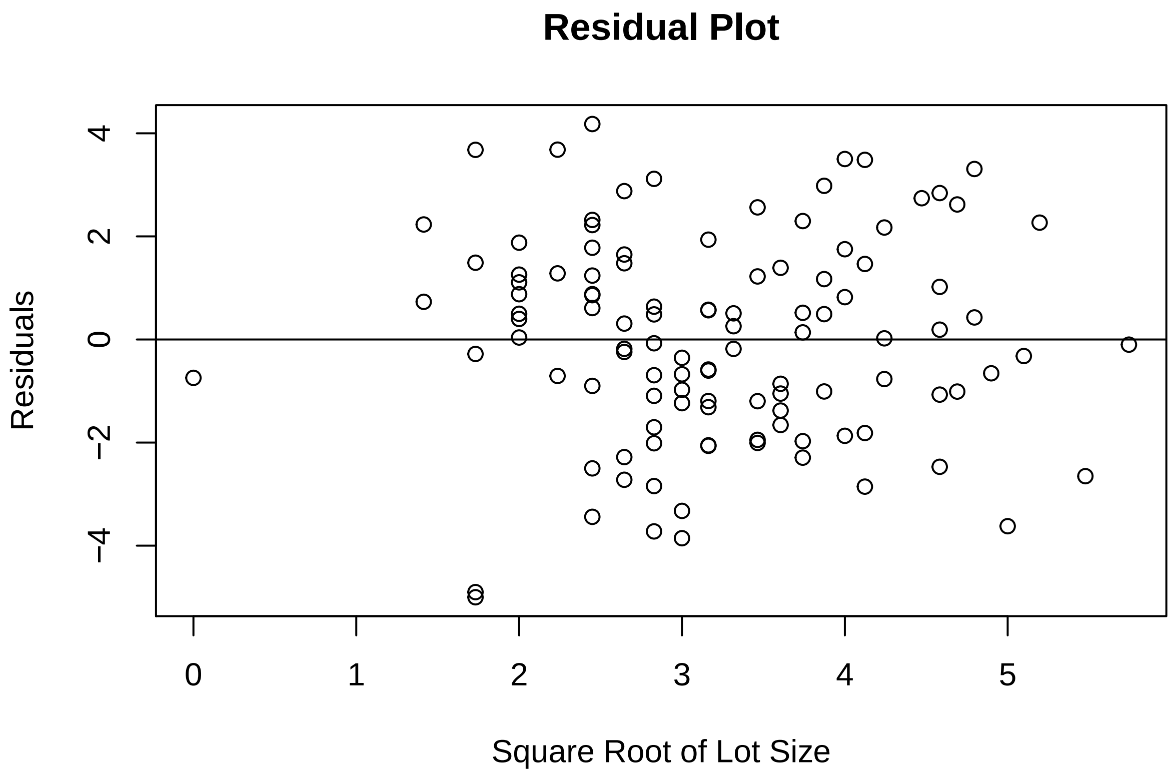
```r
#scatter plot
plot(Y ~ X2, data = hw3_data)
#regression line
abline(hw3_model)
```



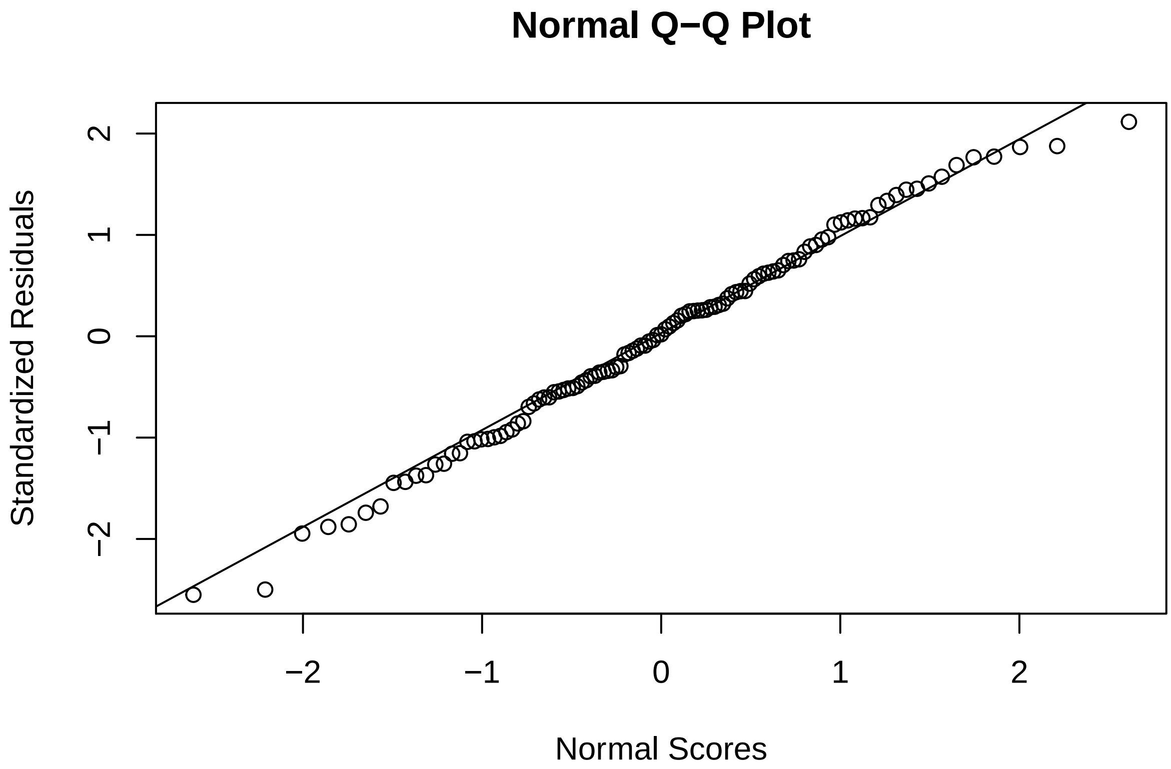
The regression line does appear to be a good fit to the transformed data.

d)

```
#Plot residuals against fitted values
plot(hw3_data$X2, resid(hw3_model),
     ylab  = "Residuals", xlab = "Square Root of Lot Size", main ="Residual Plot")
abline(0,0)
```



```
#Normal probability plot
qqnorm(rstandard(hw3_model),
       ylab="Standardized Residuals",
       xlab="Normal Scores")
qqline(rstandard(hw3_model))
```



From the residual plot, we can conclude that there are no clear correlation between the residuals and the transformed Xi's. Therefore, the regression with transformed X satisfies the assumption that all the residuals are independent of the Xi's.
From the normal probability plot, we can see that the points are close to the q-q line, that means that our sample (observed) quantiles are close to the theoretical normal quantiles. We can then conclude that the residuals from our regression are normally distributed, which satisfies another linear model assumption.

e) The estimated regression function in the original units is

$$\hat{Y} = 1.2547 + 3.6235\sqrt{X}$$