CC12 Practical Q14

Ananda Biswas

- The cost of maintenance of shipping tractors seems to increase with the age of the tractor.

(a) Fit the model $Y = \beta_0 + \beta_1 X + \epsilon$.

(b) Is the model suitable ?

Age (Years) x	6 month's cost y
4.5	619
4.5	1049
4.5	1033
4.0	495
4.0	729
4.0	681
5.0	890
5.0	1522
5.5	987
5.0	1194
0.5	163
0.5	182
6.0	764
6.0	1373
1.0	978
1.0	466
1.0	549

- **Loading the data-set and other initials**

```
tractor_maintainance_cost_data <- read.csv("D:\\data_sets\\cc12_prac_q14_data.csv")

dim(tractor_maintainance_cost_data)

## [1] 17  2

names(tractor_maintainance_cost_data)

## [1] "years" "cost"
```

```
tractor_maintainance_cost_data

##    years cost
## 1    4.5  619
## 2    4.5 1049
## 3    4.5 1033
## 4    4.0  495
## 5    4.0  729
## 6    4.0  681
## 7    5.0  890
## 8    5.0 1522
## 9    5.5  987
## 10   5.0 1194
## 11   0.5  163
## 12   0.5  182
## 13   6.0  764
## 14   6.0 1373
## 15   1.0  978
## 16   1.0  466
## 17   1.0  549
```

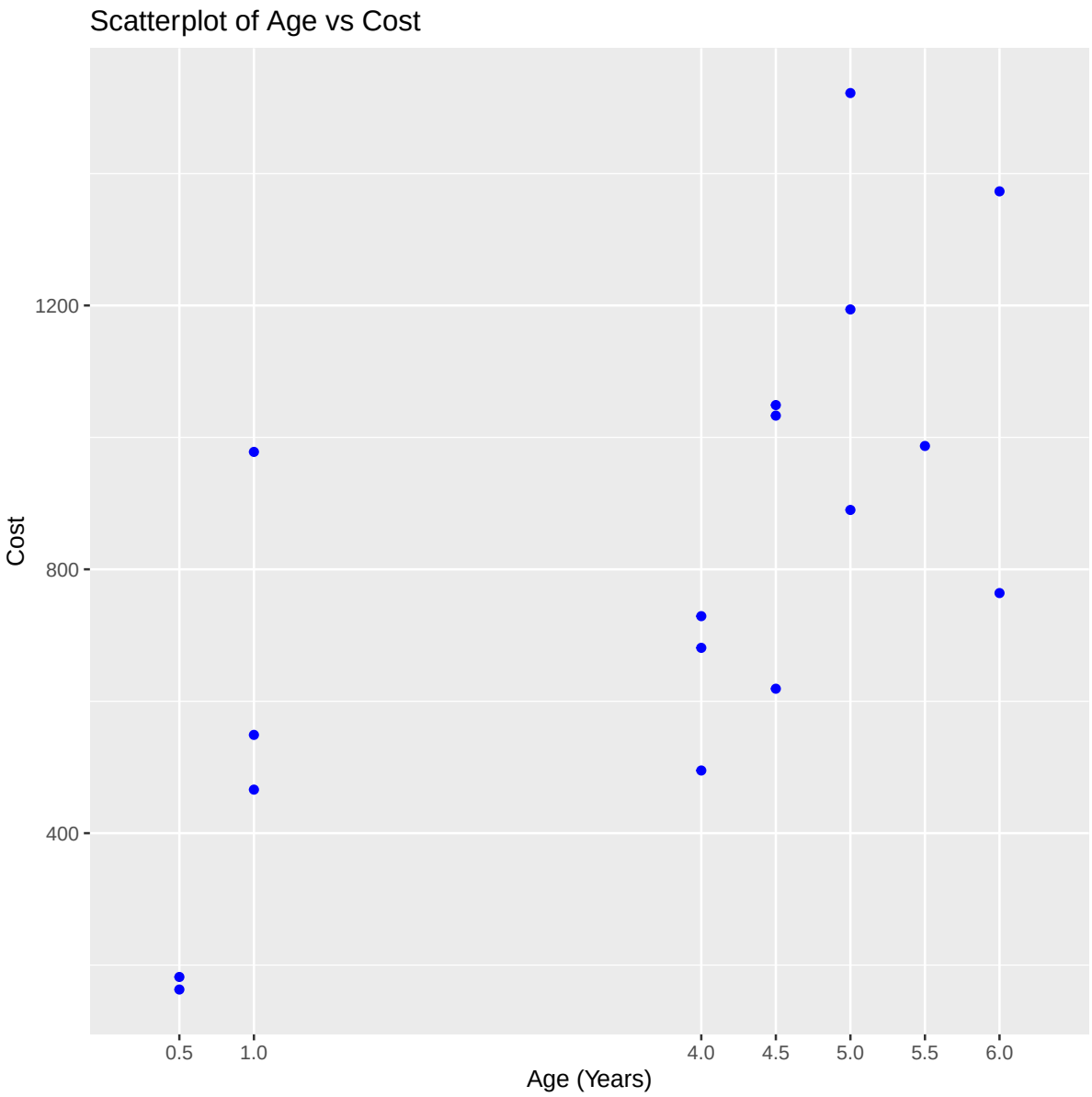
```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.3
## Warning: package 'ggplot2' was built under R version 4.2.2
## Warning: package 'tibble' was built under R version 4.2.3
## Warning: package 'tidyr' was built under R version 4.2.3
## Warning: package 'readr' was built under R version 4.2.2
## Warning: package 'purrr' was built under R version 4.2.3
## Warning: package 'dplyr' was built under R version 4.2.3
## Warning: package 'stringr' was built under R version 4.2.3
## Warning: package 'forcats' was built under R version 4.2.2
## Warning: package 'lubridate' was built under R version 4.2.2
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.1     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

- **Scatterplot of Age of Tractor vs 6 month's Maintenance Cost**

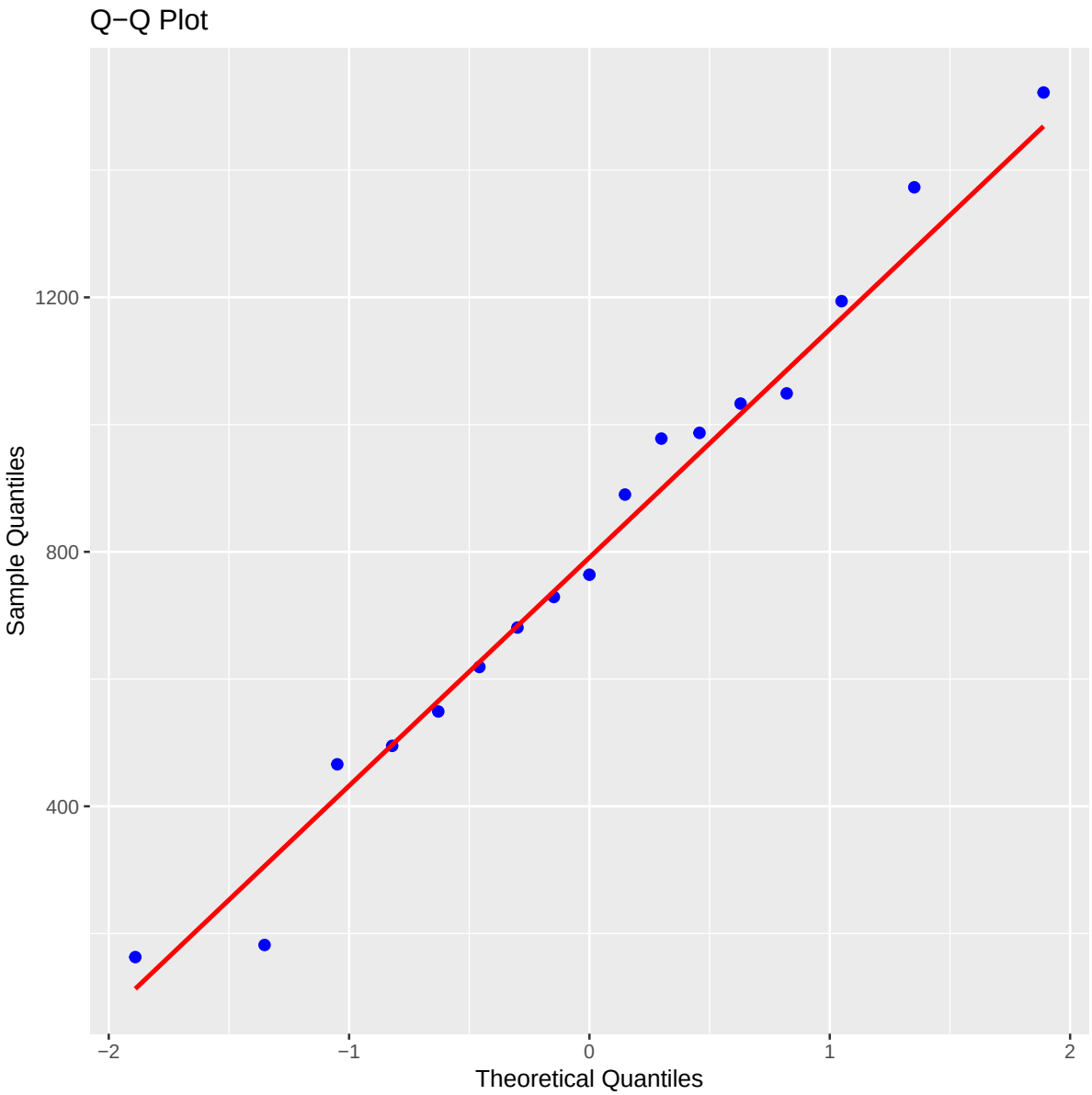
```
tractor_maintainance_cost_data %>%
    ggplot(aes(x = years, y = cost)) + geom_point(size = 1.5, col = "blue") +
    scale_x_discrete(limits = tractor_maintainance_cost_data$years) + labs(x = "Age (Years)",
    y = "Cost", title = "Scatterplot of Age vs Cost")

## Warning:  Continuous limits supplied to discrete scale.
## i Did you mean `limits = factor(...)` or `scale_*_continuous()`?
```



- **Test for Normality :: Q-Q Plot**

```
tractor_maintainance_cost_data %>%
    ggplot(aes(sample = cost)) + geom_qq(size = 2, col = "blue") + geom_qq_line(linewidth = 1,
    col = "red") + labs(x = "Theoretical Quantiles", y = "Sample Quantiles",
    title = "Q-Q Plot")
```



We see that it is a good fit. So normality assumption holds.

- **Test for Normality :: Shapiro-Wilk Test**

```
shapiro.test(tractor_maintainance_cost_data$cost)

##
##  Shapiro-Wilk normality test
##
## data:  tractor_maintainance_cost_data$cost
## W = 0.97911, p-value = 0.9485
```

The p-value is much higher than 0.05. So we fail to reject the null hypothesis that the data is sampled from a normal population.

```r
fit1 <- lm(cost ~ years, data = tractor_maintainance_cost_data)
```

Let us have the diagonal elements of the hat matrix.

```r
h <- hatvalues(fit1)
# lm.influence(fit1)$hat
h
```

```
##          1          2          3          4          5          6          7
## 0.07030162 0.07030162 0.07030162 0.06078886 0.06078886 0.06078886 0.08770302
##          8          9         10         11         12         13         14
## 0.08770302 0.11299304 0.08770302 0.21508121 0.21508121 0.14617169 0.14617169
##         15         16         17
## 0.16937355 0.16937355 0.16937355
```

- **Detection of Outliers**

```
standardised_residuals <- fit1$residuals/sqrt(mean(fit1$residuals^2) *
    (1 - h))

standardised_residuals

##          1          2          3          4          5          6          7
## -1.1603144  0.5154953  0.4531396 -1.3798001 -0.4724779 -0.6585953 -0.3643202
##          8          9         10         11         12         13         14
##  2.1221098 -0.2452910  0.8316841 -0.9616580 -0.8810704 -1.4247787  1.0518408
##         15         16         17
##  2.1538967  0.0428669  0.3850846
```

```
which(abs(standardised_residuals) > 3)

## named integer(0)
```

None of the absolute values of the standardized residuals is more than 3. So we drop of the suspicion of presence of any outlier.

- **Detection of High Leverage Observations**

```
which(h > (3 * 1)/17)

## 11 12
## 11 12
```

So 11th and 12th value of the covariate(here years) are high leverage observations.

```
tractor_maintainance_cost_data[which(h > (3 * 1)/17), 1]

## [1] 0.5 0.5
```

- **Detection of Influential Observations**

```
cooks_distance <- cooks.distance(fit1)

cooks_distance
```

```
##            1            2            3            4            5            6
## 0.0449145449 0.0088651274 0.0068501419 0.0543632563 0.0063743541 0.0123854097
##            7            8            9           10           11           12
## 0.0056293342 0.1909969103 0.0033814278 0.0293364438 0.1117973181 0.0938450259
##           13           14           15           16           17
## 0.1533203953 0.0835612786 0.4173514264 0.0001653089 0.0133402667
```

```
which(cooks_distance > 4/17)
```

```
## 15
## 15
```

So 15th observation is an influential observation.

```
tractor_maintainance_cost_data[which(cooks_distance > 4/17), ]
```

```
##    years cost
## 15     1  978
```

- **Test for Homoscedasticity :: Goldfeld-Quandt Test**

```
library(lmtest)
```

```
## Warning:  package 'lmtest' was built under R version 4.2.3
## Loading required package:  zoo
##
## Attaching package:  'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
gqtest(fit1, order.by = ~years, data = tractor_maintainance_cost_data)
```

```
##
##  Goldfeld-Quandt test
##
## data:  fit1
## GQ = 1.2262, df1 = 7, df2 = 6, p-value = 0.4098
## alternative hypothesis: variance increases from segment 1 to 2
```

p-value is much higher than 0.05. So we fail to reject the null hypothesis that homoscedasticity is present.

- **Test for Homoscedasticity :: Breusch-Pagan Test**

```
library(lmtest)
```

```
bptest(fit1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit1
## BP = 0.0030494, df = 1, p-value = 0.956
```

p-value is much higher than 0.05. So we fail to reject the null hypothesis that homoscedasticity is present.

- **Test for Auto-correlation :: Durbin-Watson Test**

```
library(car)

## Loading required package:  carData
##
## Attaching package:  'car'
## The following object is masked from 'package:dplyr':
##
##     recode
## The following object is masked from 'package:purrr':
##
##     some
```

```
durbinWatsonTest(fit1)

##  lag Autocorrelation D-W Statistic p-value
##    1      0.03418094      1.850764   0.642
##  Alternative hypothesis: rho != 0
```

p-value is much higher than 0.05. So we fail to reject the null hypothesis that the residuals are not auto-correlated.

```r
summary(fit1)
```

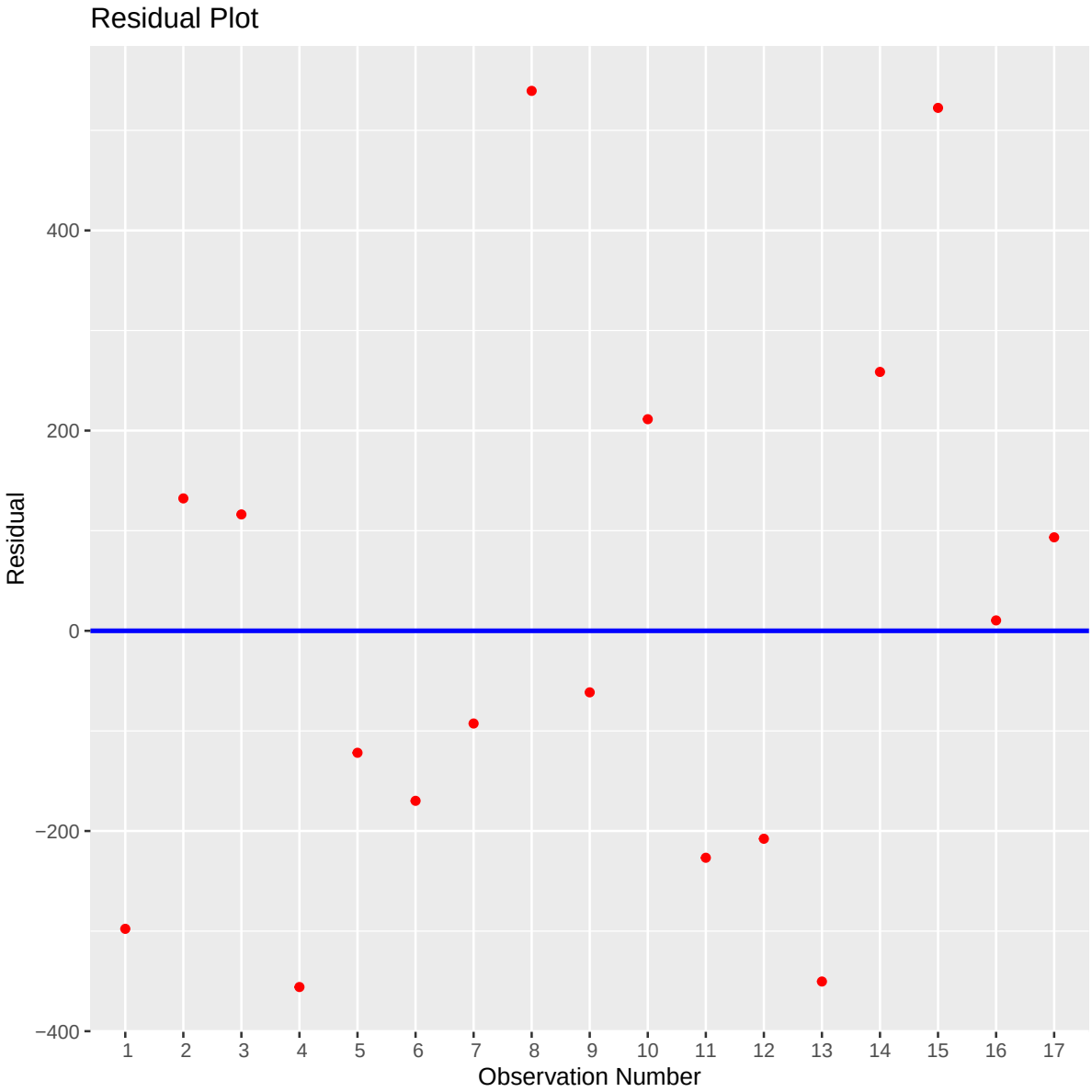
```
##
## Call:
## lm(formula = cost ~ years, data = tractor_maintainance_cost_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -355.85 -207.73  -61.48  132.27  539.40
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   323.85     146.85   2.205  0.04345 *
## years         131.75      35.59   3.702  0.00213 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 283.3 on 15 degrees of freedom
## Multiple R-squared:  0.4775,Adjusted R-squared:  0.4427
## F-statistic: 13.71 on 1 and 15 DF,  p-value: 0.002129
```

- **Residual Plot**

```
df1 <- data.frame(sample_no = 1:length(fit1$residuals), residuals = fit1$residuals)
```

```
df1 %>%
    ggplot(aes(x = sample_no, y = residuals)) + geom_point(size = 1.5,
    col = "red") + geom_hline(yintercept = 0, linewidth = 1, col = "blue") +
    scale_x_discrete(limits = 1:length(fit1$residuals)) + labs(x = "Observation Number",
    y = "Residual", title = "Residual Plot")
```

```
## Warning:  Continuous limits supplied to discrete scale.
## i Did you mean `limits = factor(...)` or `scale_*_continuous()`?
```



- **Model Checking**

Let us have a look at R-squared for the model.

```
summary(fit1)$r.squared

## [1] 0.4774928
```

Let us have a look at Adjusted R-squared for the model.

```
summary(fit1)$adj.r.squared

## [1] 0.4426589
```