Rexample1

Lingxiao Zhou

Copier example

Let X be the number of copiers serviced and Y be the time spent (in minutes) by the technician for a known manufacturer.

```
copier <- read.csv("https://raw.githubusercontent.com/lingxiaozhou/STA4210Rmaterial/main/data/copiers.cs
    skip = 1, header = TRUE)

head(copier)  # print the first six rows of the dataframe
```

```
##   Time Copiers
## 1   20       2
## 2   60       4
## 3   46       3
## 4   41       2
## 5   12       1
## 6  137      10
```

```
nrow(copier)  # check the sample size
```

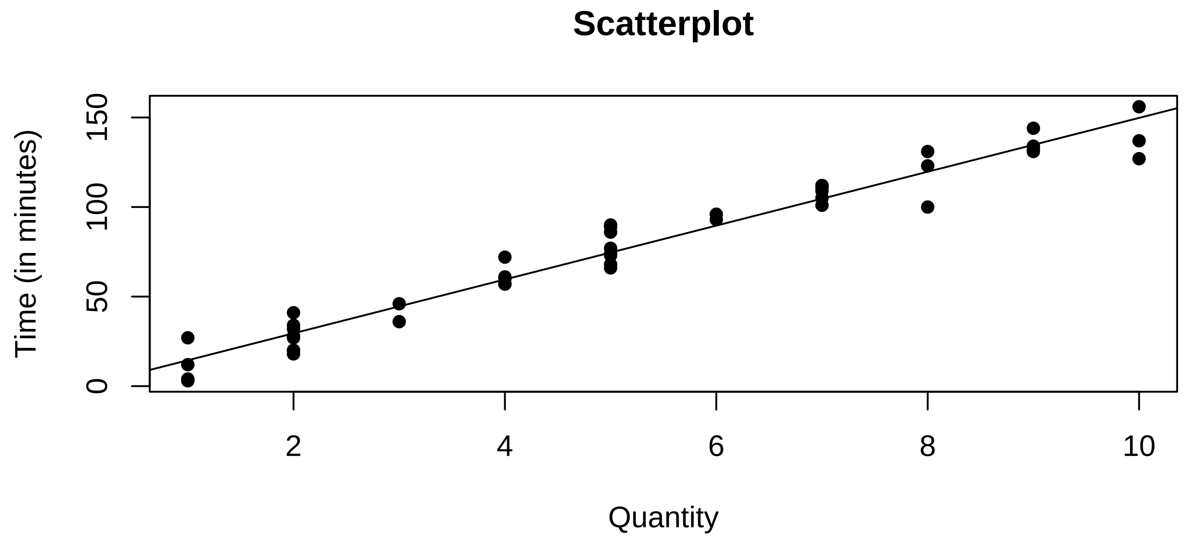
```
## [1] 45
```

```
# Scatterplot
plot(copier$Copiers, copier$Time, pch = 16, xlab = "Quantity",
    ylab = "Time (in minutes)", main = "Scatterplot")


# Fit the model
reg <- lm(Time ~ Copiers, data = copier)
summary(reg)  # get the summary of the model
```

```
##
## Call:
## lm(formula = Time ~ Copiers, data = copier)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207    0.837
## Copiers      15.0352     0.4831  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

```
abline(reg)  # add the fitted line to the scatterplot
```



The estimated equation is

$$\hat{Y} = -0.5802 + 15.0352X, \text{ for } X \in \text{approximately } [0, 10]$$

Recall that for simple linear regression model, we have

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2),$$

which implies

$$E[Y_i] = \beta_0 + \beta_1 X_i$$

- We note that the slope $b_1 = 15.0352$ implies that for each unit increase in copier quantity, the mean service time increases by 15.0352 minutes (for quantity values between 1 and 10).
- The estimated σ is s = 8.914 from the output.
- If we wish to estimate the expected time needed for a service call for 5 copiers that would be $-0.5802 + 15.0352(5) = 74.5958$ minutes.

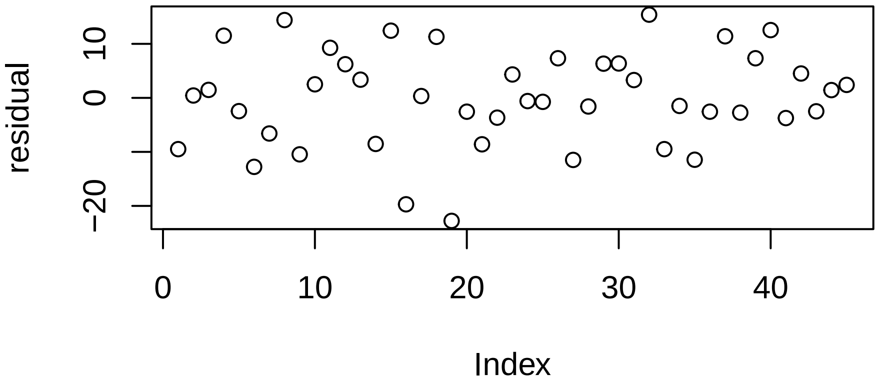
```
predict(reg, newdata = data.frame(Copiers = 5))
```

```
##        1
## 74.59608
```

```r
# print the residuals
cat("first 5 residuals:", reg$residuals[1:5], "\n")
```

```
## first 5 residuals: -9.490339 0.4391645 1.474413 11.50966 -2.455091
```

```r
plot(reg$residuals, main = "", ylab = "residual")
```



- The first residual implies that the actual observed time was 9.4903 minutes smaller than the model estimates.
- Residual magnitude in minute may not be easy to interpret whether large or small in the context of the problem.
- standardized residual can be used to compare the residuals from different models: $\frac{e_i}{\mathrm{sd}(e_i)}$
  - standardized residuals are centered and scaled to make it easier to interpret residuals from different models.
  - An observation with a standardized residual that is larger than 3 (in absolute value) is considered big.

```r
cat("first 5 standardized residuals:", rstandard(reg)[1:5])
```

```
## first 5 standardized residuals: -1.092749 0.04991894 0.1684136 1.325261 -0.2858998
```

```r
copier$Time <- copier$Time/60  # convert the unit of time to hours
```

```r
# refit the model
reg2 <- lm(Time ~ Copiers, data = copier)

# print the residuals
cat("first 5 residuals:", reg2$residuals[1:5], "\n")
```

```
## first 5 residuals: -0.1581723 0.007319408 0.02457354 0.1918277 -0.04091819
```

```r
cat("first 5 standardized residuals:", rstandard(reg2)[1:5])
```

```
## first 5 standardized residuals: -1.092749 0.04991894 0.1684136 1.325261 -0.2858998
```