hw7

Kyu Park

2021 3 1

1.

a) $b_1 = r\frac{S_y}{S_x}$ where $r = 0.67, S_y = 9.41, S_x = 10.37$.

$$b_1 = 0.67\frac{9.41}{10.37} = 0.60797$$

$y - \overline{y} = b_1(x - \overline{x})$, where $\overline{y} = 171.14, \overline{x} = 107.2$

$$y - 171.14 = 0.60797(x - 107.2)$$

$$y = 105.965616 + 0.60797x$$

where y = height and x = shoulder girth.

b) For every 1 cm increase of the shoulder girth, we expect 0.60797 cm increase in height. If someone has 0 cm shoulder girth, we expect that person to have height of 105.965616 cm. However, the meaning of the intercept is not relevant to this problem, since it is very unlikely that someone has 0 shoulder girth.

c)

$$r^2 = 0.67^2 = 0.4489$$

This means that 44.89% of the data fits into the regression model. 44.89% of the sample will fit into the linear regression that describes the correlation between shoulder girth and height.

d) $r^2 = \frac{s^2_{height} - s^2_{RES}}{s^2_{height}}$ where $s^2_{height} = 9.41^2 = 88.5481$.

$$0.4489 = \frac{88.5481 - s^2_{RES}}{88.5481}, s^2_{RES} = 48.79885791$$

Therefore, $s_{RES} = \sqrt{48.79885791} = 6.98561$

e) $y = 105.965616 + 0.60797x, x = 100, y = 166.762616$ Expected height is 166.762616 cm. Confidence interval could be calculated as $\hat{y} \pm t^* s_{\hat{y}}$ where $s_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum(x_i - \overline{x})^2}}$. $s_{\hat{y}} = 6.98561\sqrt{1 + \frac{1}{25} + \frac{(100 - 107.2)^2}{25(10.37)^2}} = 7.18969$

```
qt(0.025, 23, lower.tail = FALSE)
```

```
## [1] 2.068658
```

$CI = 166.762616 \pm (2.068658)(7.18969)$ CI = (151.88960, 181.63562)

f) $y_i - \hat{y}_i = 160 - 166.762616 = -6.762616$. Residual is -6.762616 cm. This means that the actual height is -6.762616 cm away from the expected height from the regression model.

g) No, the range of the data collection, the range of the shoulder girth is unknown. This may be extrapolation beyond the range. Therefore, it would not be appropriate to use this linear model, especially since a one year shoulder girth is highly likely to be beyond the data collection range.

2.

a)

$$t_{value}\ Bwt = 4.0341/0.2503 = 16.1171$$

```r
pt(16.1171, 142, lower = FALSE)
```

```
## [1] 3.530464e-34
```

$$Pr(>|t|)\ Bwt = 2p(t_{142} > 16.1171) = 2(3.530464e-34) = 7.060928e-34$$

$$Adjusted\ Rsquared = 1 - \frac{S^2}{SST/143}, SST = \frac{142(1.452^2)}{1-0.6466} = 847.13969$$

$$1 - \frac{1.452^2}{847.13969/143} = 1 - 0.35588 = 0.64411$$

$$F\ statistic = \frac{SSR/dfR}{SSE/dfE}, SST = SSR+SSE, SST = 847.13969, SSE = 142(1.452^2) = 299.379, SSR = 547.761$$

$$F\ statistic = \frac{547.761/1}{299.379/142} = 259.81134\ on\ 1\ and\ 142\ df$$

Note that F statistics could also be calculated by calculating the square of the t-statistic for the slope, $16.1171^2 = 259.8$.

```r
pf(259.81134, 1, 142, lower.tail = FALSE)
```

```
## [1] 6.998051e-34
```

$$p = 6.998051e-34$$

b) $y = -0.3567 + 4.0341x$ where y = heart weight (in g) and x = body weight (in kg)
c) Intercept is -0.3567, which means that if the body weight is 0 kg, the heart weight is -0.3567 g. In this context, such interpretation is meaningless.
d) The slope is 4.0341, which means that for every 1 kg increase in body weight, the heart weight increases by 4.0341 g.
e) R squared is 0.6466, which means that 64.66% of the data could be explained by the regression model above.
f) $R = \sqrt{R^2} = \sqrt{0.6466} = 0.8041$
g) Yes, the residual plot does not show clear pattern, but rather the residuals are evenly spread throughout and are approximately normally distributed. The variance of the residual is also constant. This indicates that the model assumptions are valid.

3.

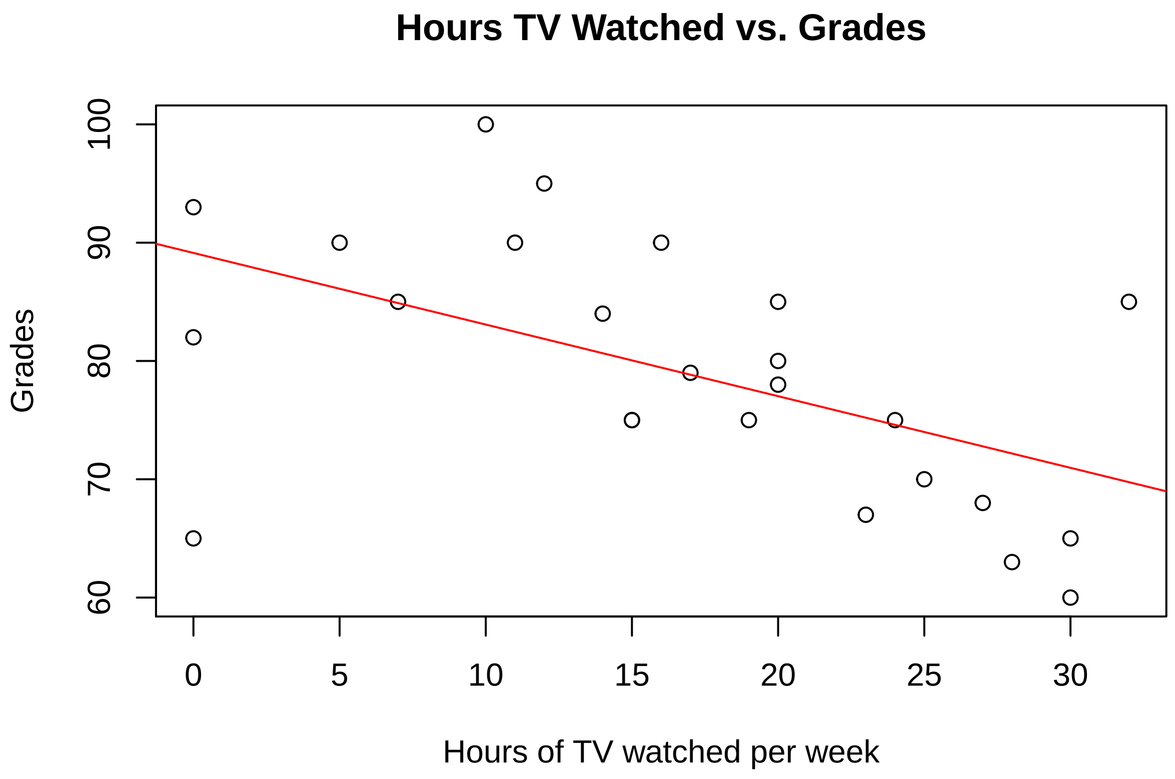
a) TV = Number of hours per week students watch TV, grades = grades students got in a statistics class (out of 100)

b)

```
lm(grades ~ tv, data)
```

```
##
## Call:
## lm(formula = grades ~ tv, data = data)
##
## Coefficients:
## (Intercept)           tv
##     89.1316      -0.6054
```

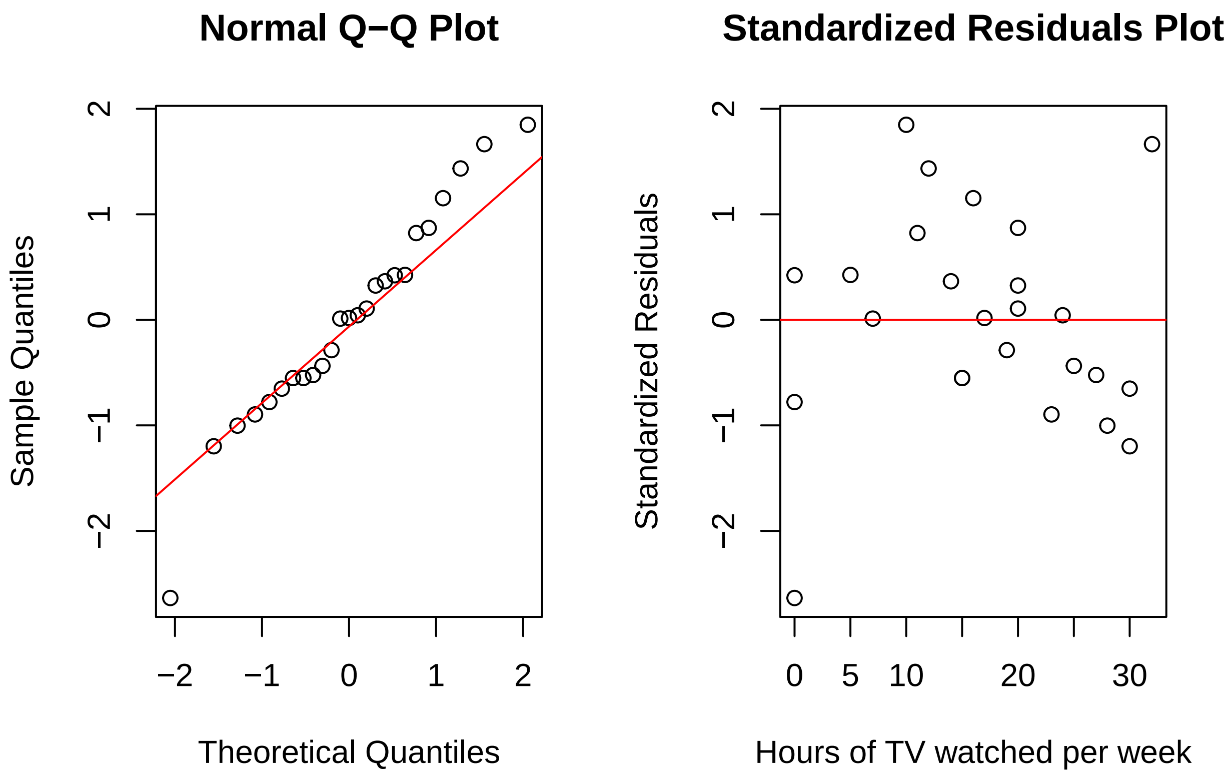
```
plot(data$tv, data$grades, xlab = "Hours of TV watched per week",
     ylab = "Grades", main = "Hours TV Watched vs. Grades")
abline(lm(grades ~ tv, data), col = "RED")
```



c)

```
par(mfrow=c(1,2))
res = (lm(grades ~ tv, data)$res - mean(lm(grades ~ tv, data)$res))/sd(lm(grades ~ tv, data)$res)
qqnorm(res)
```

```r
qqline(res, col = "RED")
plot(data$tv, res, ylab = "Standardized Residuals",
     xlab = "Hours of TV watched per week", main = "Standardized Residuals Plot")
abline(h=0, col = "RED")
```



The assumptions are valid, since the standardized residual displays the normal distribution, and the regular residual plot shows no clear pattern. The residuals also have constant variance.

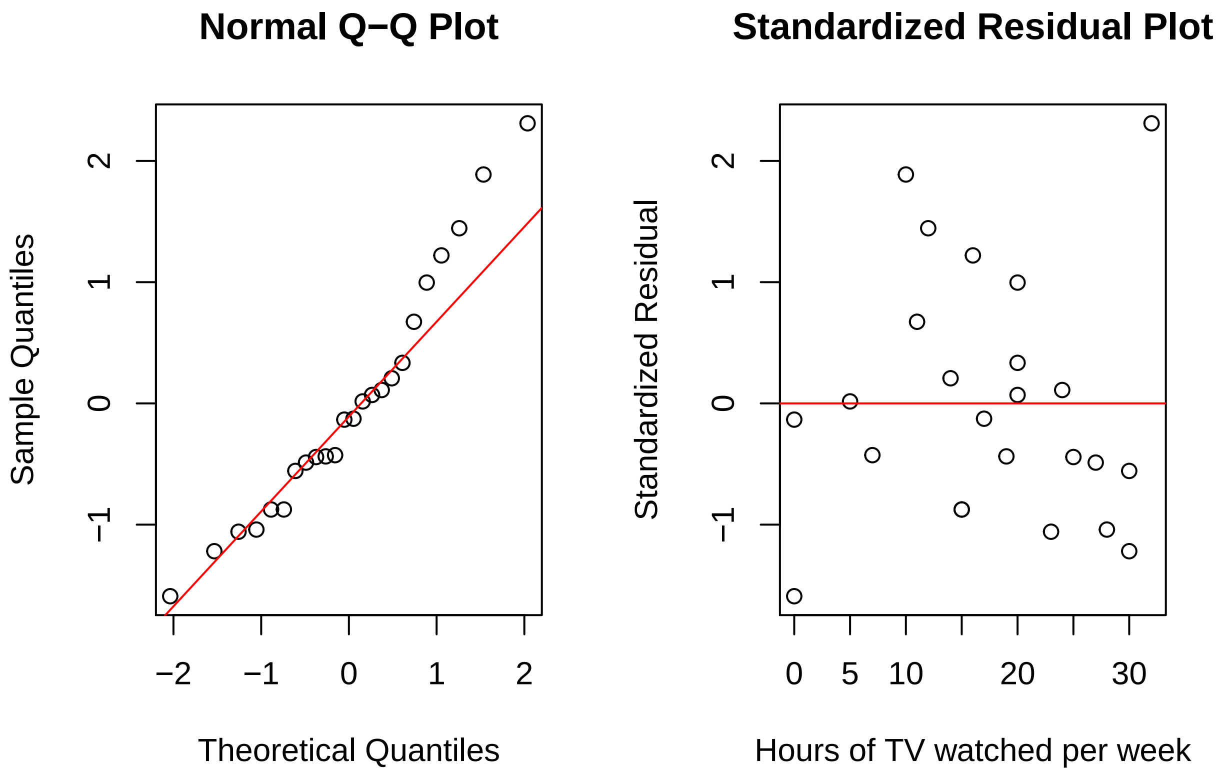
d) The very first point on the normal q-q plot seems to be the outlier. It is the point farthest from the normal qqline. This is the point where the residual shows the greatest deviation from the expected model.

e)

```r
par(mfrow=c(1,2))
lm(grades ~ tv, data[-3,])
```

```
##
## Call:
## lm(formula = grades ~ tv, data = data[-3, ])
##
## Coefficients:
## (Intercept)           tv
##     94.0104      -0.8268
```

```r
newres = (lm(grades ~ tv, data[-3, ])$res - mean(lm(grades ~ tv, data[-3, ])$res))/sd(lm(grades ~ tv, da
qqnorm(newres)
qqline(newres, col = "RED")
plot(data[-3,]$tv, newres, main = "Standardized Residual Plot",
     ylab = "Standardized Residual", xlab = "Hours of TV watched per week")
abline(h=0, col = "RED")
```



Because the outlier is omitted, the coefficients of the regression model significantly changed.

f) No, although the outlier was removed, the standardized residual plot appears approximately the same, with similar distribution. This indicates that although the outlier was removed, it doesn't necessarily improved the accuracy of the model, as it will keep produce outliers. More justification is needed to remove the outlier therefore.

4.

a) case-id number
bwt-birthweight, in ounces
gestation-length of gestation, in days
parity-binary indicator for a first pregnancy (0=first pregnancy)
age-mother's age in years
height-mother's height in inches
weight-mother's weight in pounds
smoke-binary indicator for whether the mother smokes

b)

```
lm(babies$bwt~babies$smoke)
```

```
##
## Call:
## lm(formula = babies$bwt ~ babies$smoke)
##
## Coefficients:
##  (Intercept)  babies$smoke
##      123.047        -8.938
```

If smoking had no effect on the babies' weight, then coefficient of babies$smoke would have been around 0.

c) Let H_o: predicted model is not significant and H_a: predicted model is significant

```
summary(lm(babies$bwt~babies$smoke))
```

```
##
## Call:
## lm(formula = babies$bwt ~ babies$smoke)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -68.05 -11.05   0.89  10.95  52.95
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   123.047      0.649 189.597   <2e-16 ***
## babies$smoke   -8.938      1.033  -8.653   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.68 on 1224 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.05764,    Adjusted R-squared:  0.05687
## F-statistic: 74.87 on 1 and 1224 DF,  p-value: < 2.2e-16
```

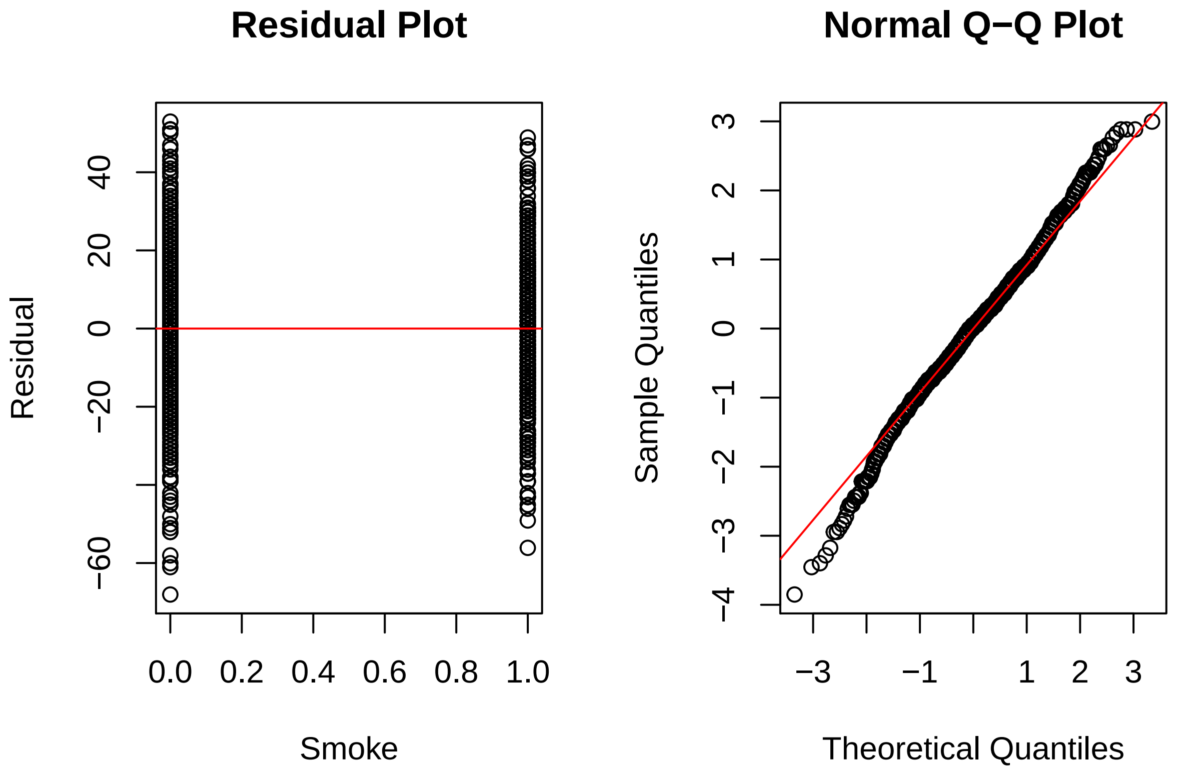
P is less than 2.2e-16, and less than 0.05. Therefore, reject H_o; the predicted model is significant, and there is a correlation between smoking and the weight of babies at 95% confidence level.

d) Assume that the residuals are independent and show no clear pattern and there is no clear outlier; the residuals have constant variance; the residuals are normally distributed.

e)

```
par(mfrow=c(1,2))
plot(na.omit(babies$smoke), lm(babies$bwt~babies$smoke)$res, main = "Residual Plot"
     , xlab = "Smoke", ylab = "Residual")
babies.lm = lm(babies$bwt~babies$smoke)
abline(h=0, col = "RED")
```

```
res=(lm(babies$bwt~babies$smoke)$res-mean(lm(babies$bwt~babies$smoke)$res))/sd(
  lm(babies$bwt~babies$smoke)$res)
qqnorm(res)
qqline(res, col = "RED")
```



The assumptions seem to hold true. The residuals seem independent and show no clear pattern and there is no clear outlier; the residuals have constant variance; the residuals are normally distributed and well follows the qqline. This indicates the linear regression model does well fit the data.