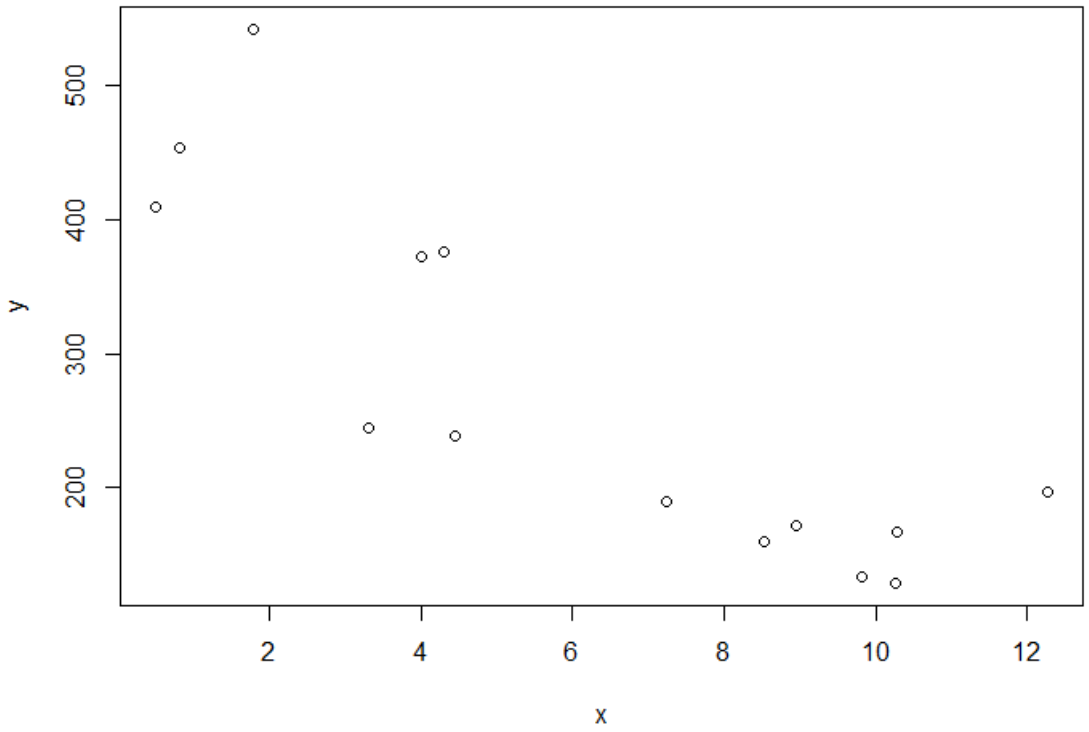
1

100 points

(a) R code and its result for the scatter plot are as follows.

```
x<-c(7.23, 8.53, 9.82, 10.26, 8.96, 12.27, 10.28, 4.45, 1.78, 4.0, 3.3, 4.3, 0.8, 0.5)
y<-c(190, 160, 134, 129, 172, 197, 167, 239, 542, 372, 245, 376, 454, 410)
plot(x,y)
```



We can see that x and y are negatively correlated.

(b) (Total **(+24 points)**) The sample correlation coefficient is

$$r = \hat{\varphi} = \frac{-5830.04}{\sqrt{198.2883}\sqrt{233081.5}} = -0.8576. \quad \textbf{(+4 points)}$$

Thus, we can know that x and y are negatively correlated **(+2 points)** and are strongly linearly correlated. **(+2 points)**

For the test, we have

$$|t| = \left| \frac{(-0.8576)\sqrt{14-2}}{\sqrt{1-(-0.8576)^2}} \right| = 5.7762, \quad \textbf{(+4 points)}$$

which is larger than $t_{0.025,12} = 2.1788$, **(+2 points)** so that we must reject H_0 at $\alpha = 0.05$. **(+2 points)**

R codes and their results are as follows.

```
> cor(x,y)
[1] -0.8575691
> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = -5.7754, df = 12, p-value = 8.805e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9540484 -0.5999903
sample estimates:
       cor
-0.8575691
```

Since the p-value is smaller than 0.05, the results are same as above. **(+(4+4) points)**

(c)

$$\hat{\beta} = B = \frac{S_{xy}}{S_{xx}} = \frac{-5830.04}{198.2883} = -29.4018$$

$$\hat{\alpha} = A = \bar{Y} - B\bar{x} = \frac{3787}{14} - (-29.4018)\frac{86.48}{14} = 452.1194$$

$$\hat{Y} = 452.1194 - 29.4018x$$

R code and its result are as follows.

```
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      452.1        -29.4
```

The result is same as above.

(d)

$$SS_R = S_{yy} - BS_{xy} = 233081.5 - (-29.4018)(-5830.04) = 61667.83$$

$$\hat{\sigma} = \sqrt{\frac{SS_R}{n-2}} = \sqrt{\frac{61667.83}{12}} = 71.6867$$

R code and its result are as follows.

```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-110.09  -38.34  -19.07   34.47  142.22

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  452.119     36.824  12.278 3.75e-08 ***
x            -29.402      5.091  -5.775 8.81e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 71.69 on 12 degrees of freedom
Multiple R-squared:  0.7354,    Adjusted R-squared:  0.7134
F-statistic: 33.36 on 1 and 12 DF,  p-value: 8.805e-05
```

The result is same as above.

(e) (Total **(+16 points)**) $H_0 : B = 0$, $H_1 : B \neq 0$

$$|t| = \left| \frac{-29.4018}{71.6847/\sqrt{198.2883}} \right| = 5.7754 \quad \textbf{(+4 points)}$$

is larger than $t_{0.025,12} = 2.1788$, **(+2 points)** so we must reject H_0 at $\alpha = 0.05$. **(+2 points)** This test is equivalent to the test in (b). **(+4 points)**

R code and its result are as follows.

```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-110.09  -38.34  -19.07   34.47  142.22

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  452.119     36.824  12.278 3.75e-08 ***
x            -29.402      5.091  -5.775 8.81e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 71.69 on 12 degrees of freedom
Multiple R-squared:  0.7354,    Adjusted R-squared:  0.7134
F-statistic: 33.36 on 1 and 12 DF,  p-value: 8.805e-05
```

Since the p-value is smaller than 0.05, the result is same as above. **(+4 points)**

(f) (Total **(+16 points)**)

$$R^2 = 1 - \frac{SS_R}{S_{yy}} = 1 - \frac{61667.83}{233081.5} = 0.7354 \quad \textbf{(+4 points)}$$

Thus, 73.54% of the total variation in Y is explained by the linear regression model. **(+4 points)** This value is square of the correlation coefficient. **(+4 points)**

R code and its result are as follows.

```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-110.09  -38.34  -19.07   34.47  142.22

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  452.119     36.824  12.278 3.75e-08 ***
x            -29.402      5.091  -5.775 8.81e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 71.69 on 12 degrees of freedom
Multiple R-squared:  0.7354,    Adjusted R-squared:  0.7134
F-statistic: 33.36 on 1 and 12 DF,  p-value: 8.805e-05
```

The result is same as above. **(+4 points)**

(g) (Total **(+16 points)**)

$$\hat{Y} = 452.1194 - 29.4018 \times 5 = 305.1104 \quad \textbf{(+4 points)}$$

95% confidence interval for $E(Y)$ is

$$305.1104 \pm 2.179 \times 71.6867 \sqrt{\frac{1}{14} + \frac{\left(5 - \frac{86.48}{14}\right)^2}{198.2883}} = (261.3683, 348.8525) \quad \textbf{(+8 points)}$$

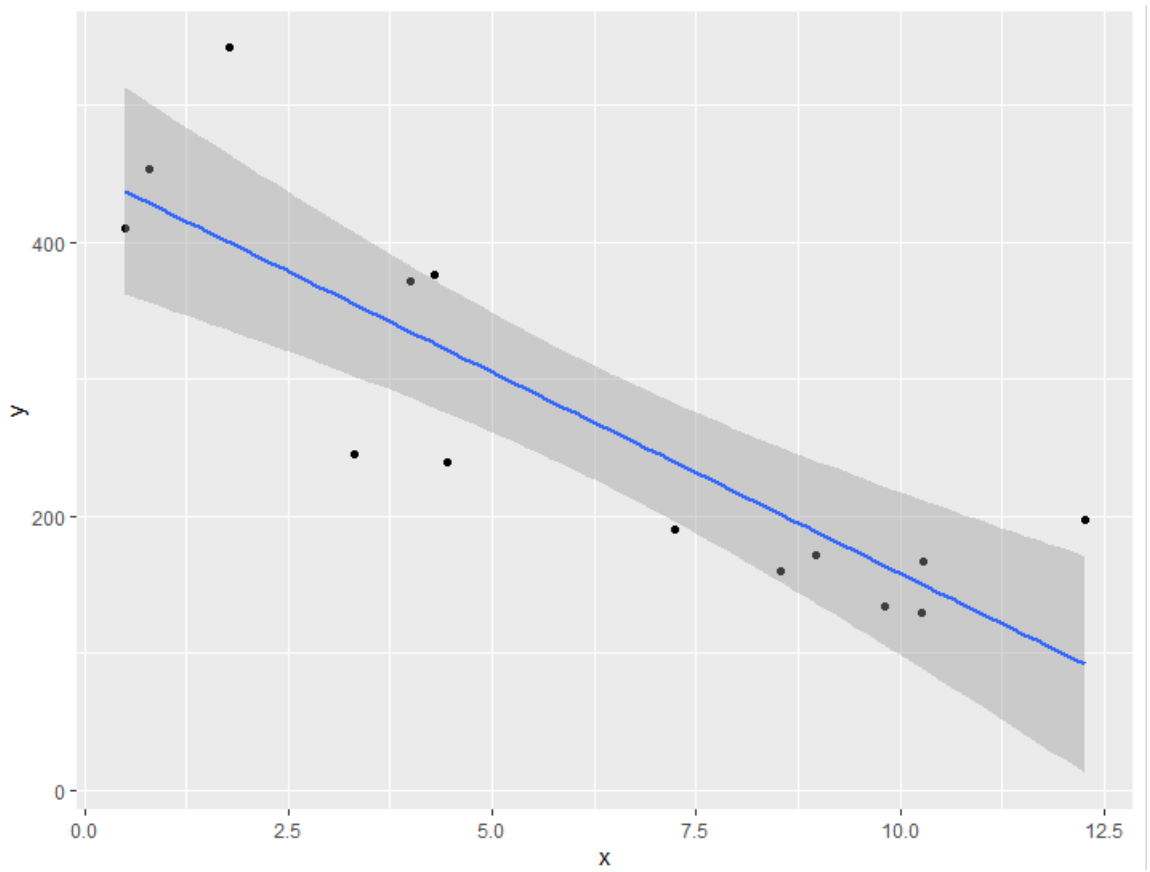
R code and its result are as follows.

```
> predict(lm(y~x),data.frame(x=5),interval='confidence')
       fit      lwr      upr
1 305.1102 261.3718 348.8485
```

The result is same as above. **(+4 points)**

(h) R code and its result are as follows.

```
library(ggplot2)
ggplot(data=data.frame(x,y),aes(x,y))+geom_point()+geom_smooth(method=lm,formula=y~x)
```



(i) (Total **(+12 points)**) 95% prediction interval for Y_{n+1} is

$$305.1104 \pm 2.179 \times \sqrt{\left(\frac{15}{14} + \frac{\left(5 - \frac{86.48}{14}\right)^2}{198.2883}\right) \times \frac{61667.83}{12}} = (142.8961, 467.3247) \quad \textbf{(+8 points)}$$

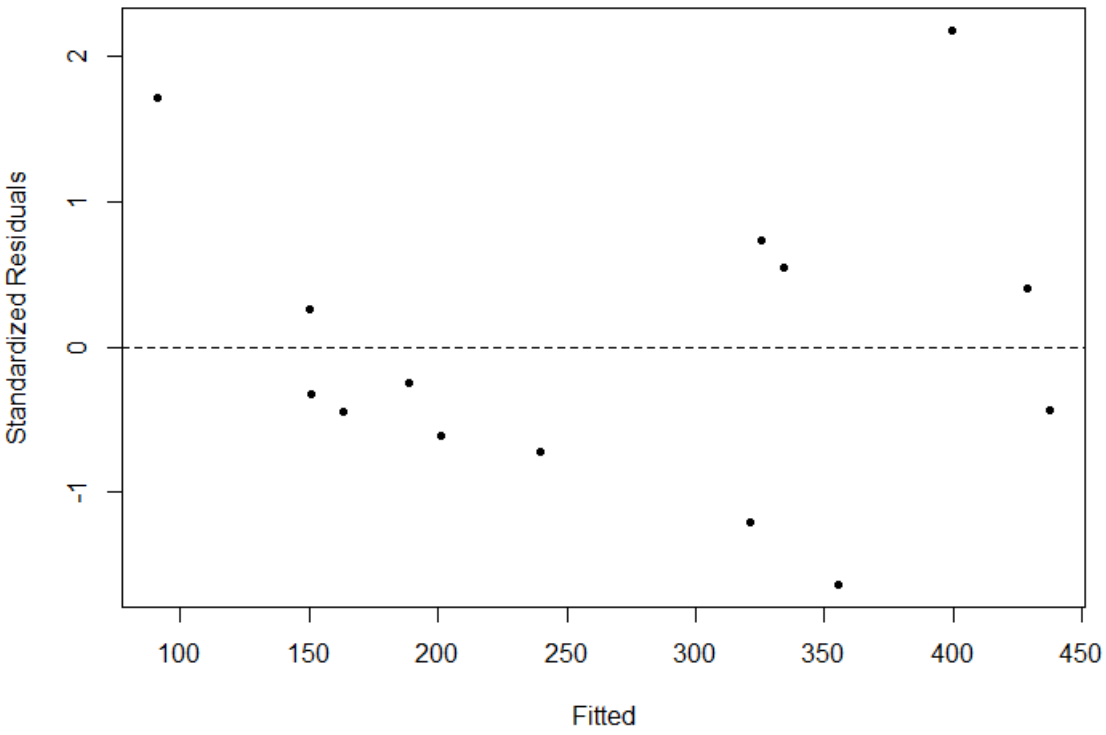
R code and its result are as follows.

```
> predict(lm(y~x),data.frame(x=5),interval='prediction')
       fit    lwr      upr
1 305.1102 142.91 467.3103
```

The result is same as above. **(+4 points)**

(j) (Total **(+16 points)**) The standardized residual plot is as follows.

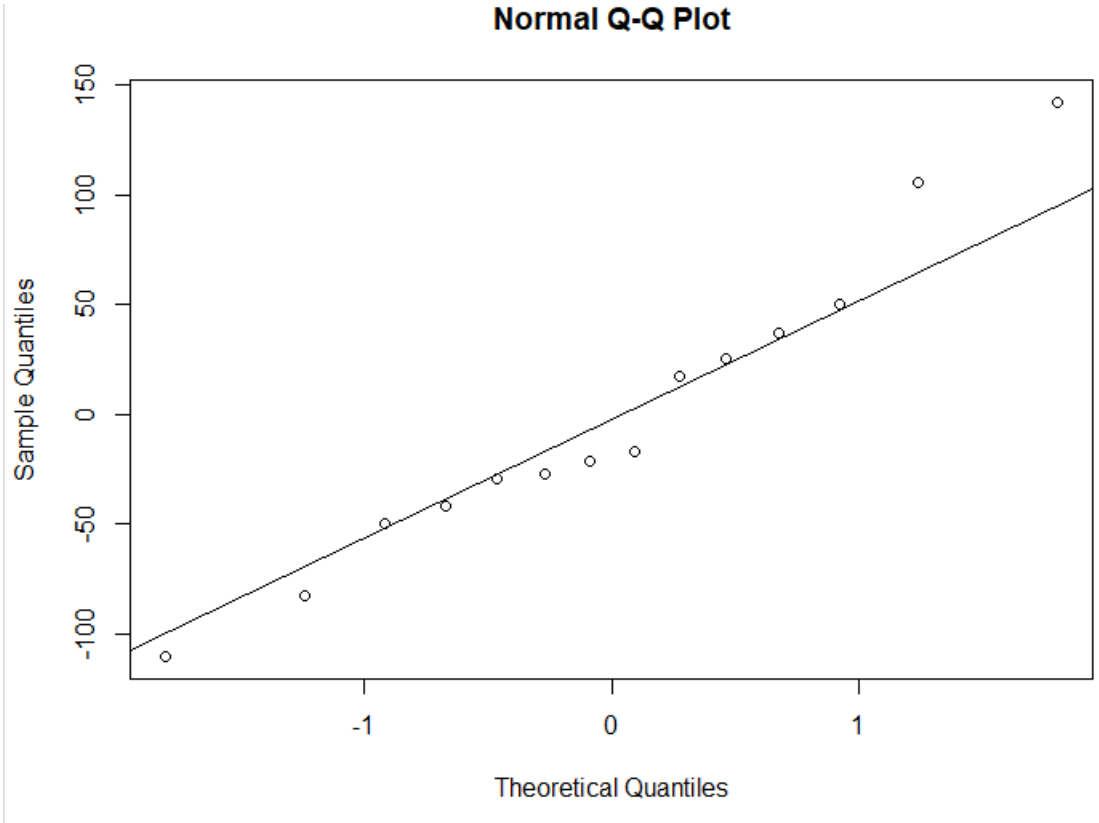
```
plot(lm(y~x)$fit,rstandard(lm(y~x)),type='p',pch=20,xlab='Fitted',ylab='Standardized Residuals')
abline(h=0,lty=2)
```



(+4 points)

The Q-Q plot is as follows.

```
qqnorm(lm(y~x)$res)
qqline(lm(y~x)$res)
```



(+4 points)

For the residual plot, there is no pattern. Points are randomly scatted around zero. There is only one point outside ± 2. For the Q-Q plot, although there is some deviation toward the right end, there seems no severe violation of normality. There are no serious violation of the assumptions: linearity, constant variance,

normality, and independence. **(+(4+4) points)** (*Note:* You will get full points if you have analyzed the plots giving appropriate reasonings, even if that analysis differs from the analysis written here.)

2 (a) Since

$$\frac{\partial}{\partial B}\sum(Y_i - Bx_i)^2 = -2\sum x_i(Y_i - Bx_i)$$

needs to be zero, we have

$$B = \frac{\sum x_i Y_i}{\sum x_i^2}.$$

(b) $E[B] = \sum x_i E[Y_i]/\sum x_i^2 = \beta$ and $Var(B) = \sum x_i^2 Var(Y_i)/(\sum x_i^2)^2 = \sigma^2/\sum x_i^2$, so that $B \sim N(\beta, \sigma^2/\sum x_i^2)$.

(c) $SS_R = \sum(Y_i - Bx_i)^2 \sim \sigma^2\chi^2_{n-1}$, since $\sum(Y_i - \beta x_i)^2 \sim \sigma^2\chi^2_n$.

(d) When $\beta = \beta_0$, $\sqrt{\sum x_i^2}(B - \beta_0)/\sigma \sim Z$. Therefore, in that case, $\sqrt{\sum x_i^2}(B - \beta_0)/\sqrt{SS_R/(n-1)} \sim T_{n-1}$. We can thus conclude that the p-value is $2P(T_{n-1} > \sqrt{\sum x_i^2}(B - \beta_0)/\sqrt{SS_R/(n-1)})$.

(e) Since $Y - Bx_0 \sim N(0, \sigma^2(1 + x_0^2/\sum x_i^2))$, the confidence interval is $Bx_0 \pm t_{\alpha/2,n-1}\sqrt{SS_R/(n-1)}\sqrt{1 + x_0^2/\sum x_i^2}$.