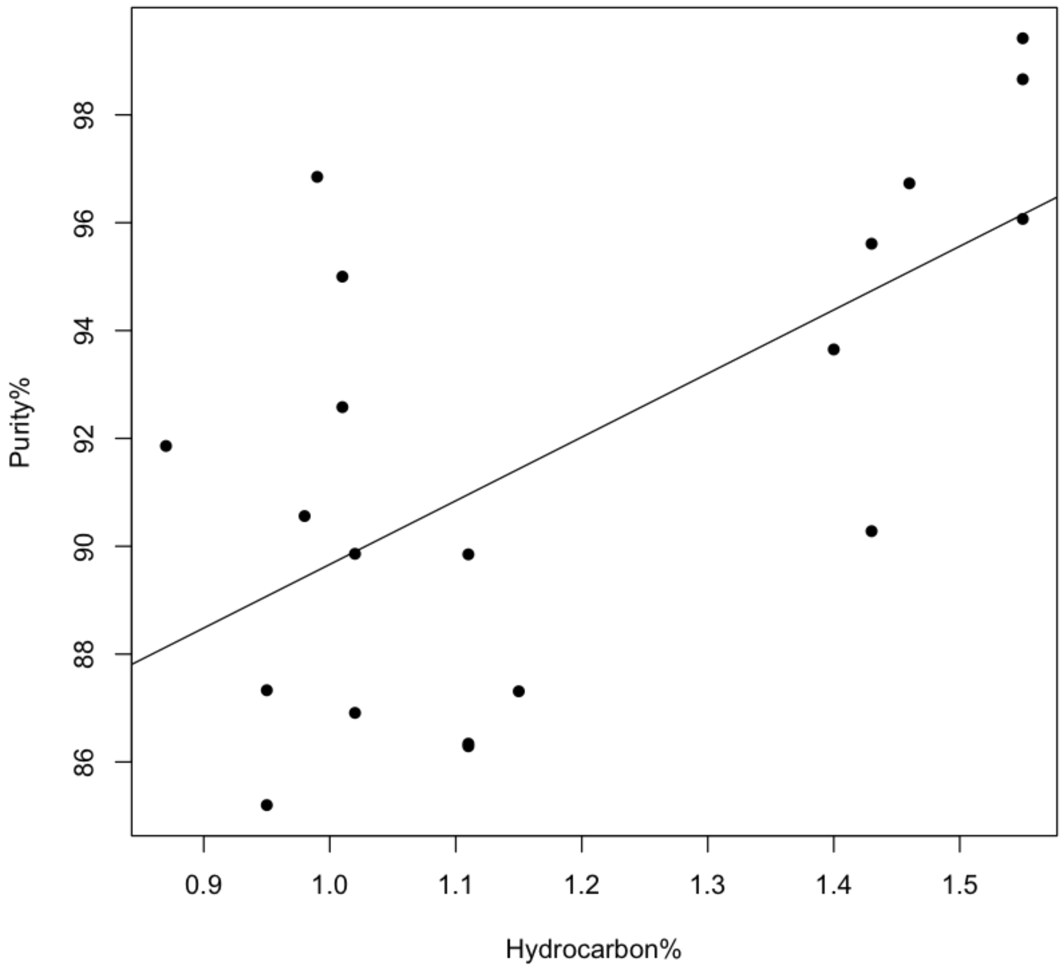
P1

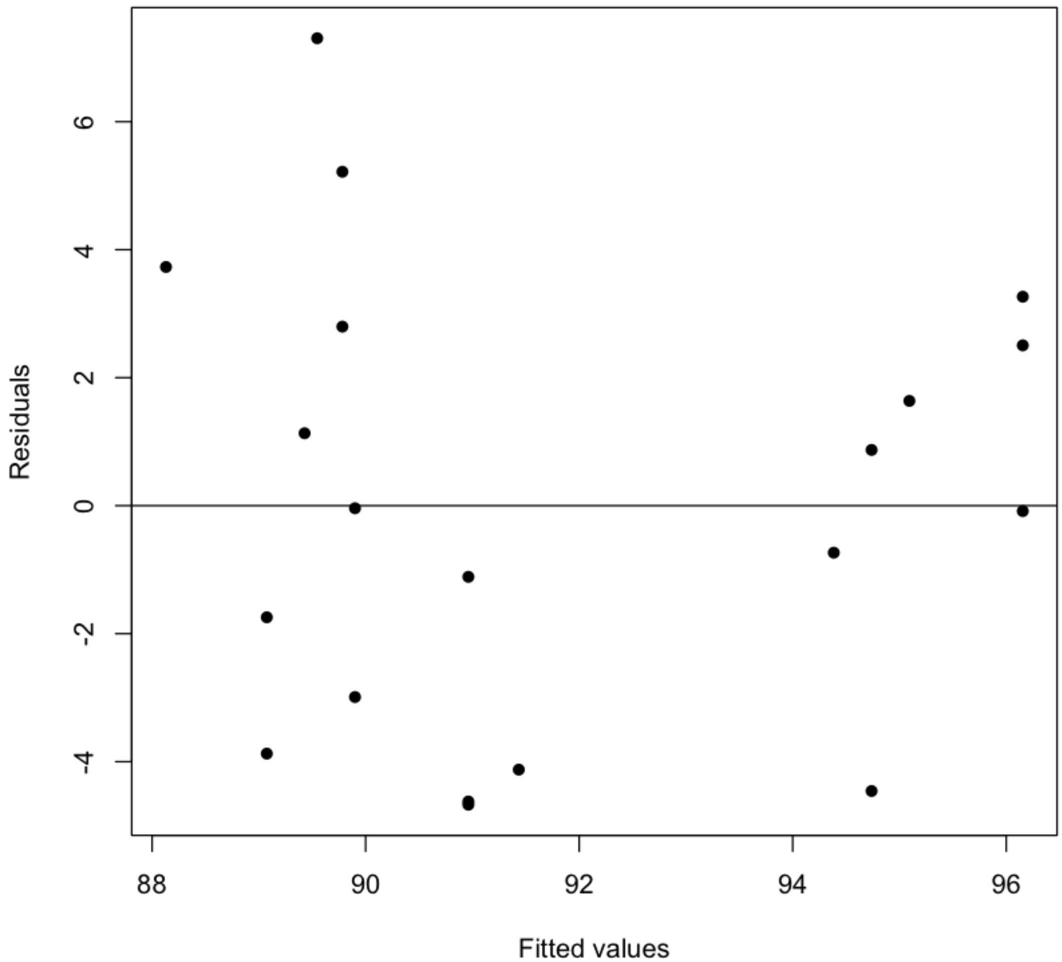
September 11, 2021

[37]:
```
df <- read.csv("problem1-oxygenpurity (1).csv")
hydrocarbon<-df$hydro
purity<-df$purity
```

[38]:
```
fit = lm(purity~hydrocarbon)
#scatterplot with regression line superimposed
plot(hydrocarbon,purity,xlab = "Hydrocarbon%",ylab = "Purity%",pch=16)
abline(fit)
#residual plot
# Residual plot
plot(fitted(fit),residuals(fit),pch=16,
xlab="Fitted values",ylab="Residuals")
abline(h=0)
#QQ plot
resid<-residuals(fit)
qqnorm(resid);qqline(resid)
#Find coefficient estimate
summary(fit)
#ANOVA table
anova(fit)
```



98
96
94
92
90
88
86
Purity%
0.9
1.0
1.1
1.2
1.3
1.4
1.5
Hydrocarbon%

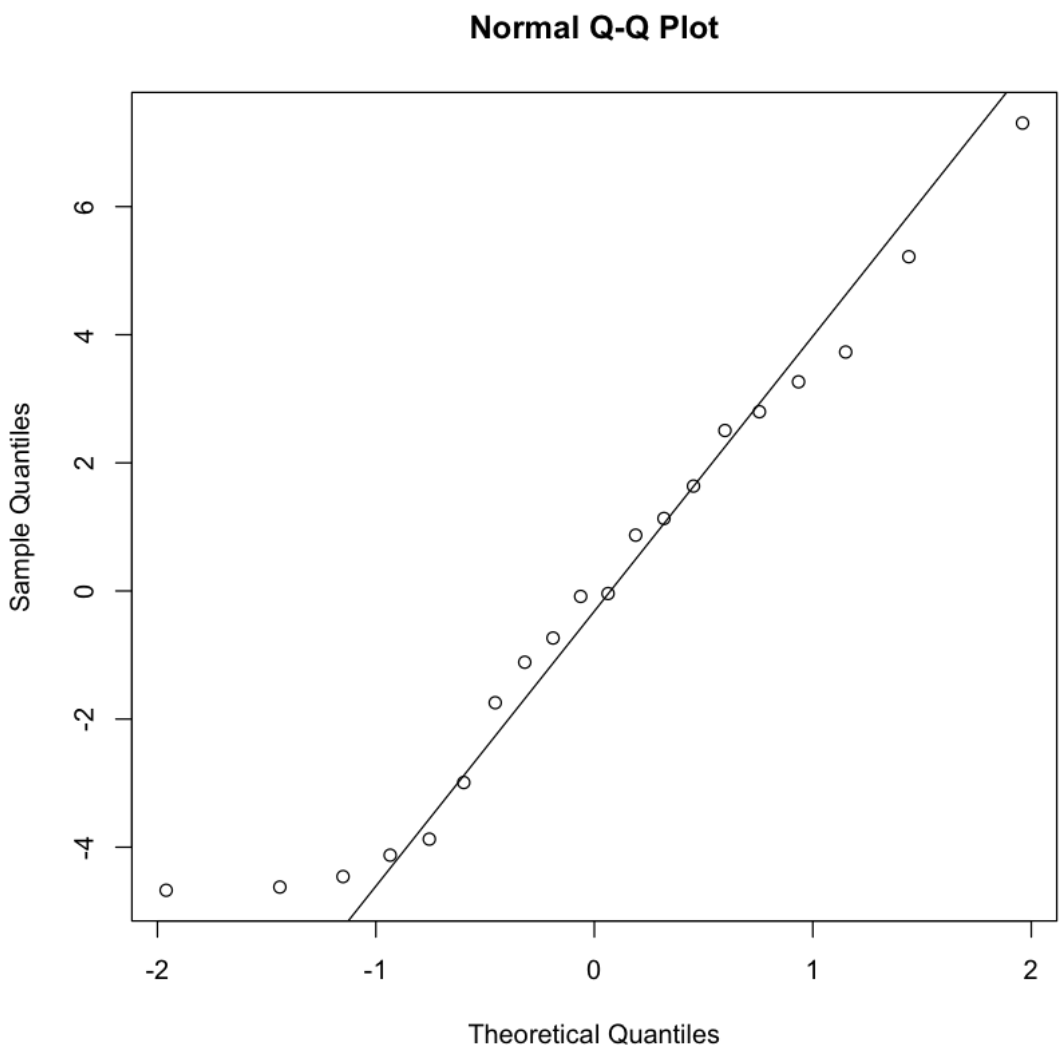



```
Call:
lm(formula = purity ~ hydrocarbon)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6724 -3.2113 -0.0626  2.5783  7.3037

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   77.863      4.199  18.544 3.54e-13 ***
hydrocarbon   11.801      3.485   3.386  0.00329 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 3.597 on 18 degrees of freedom
Multiple R-squared:  0.3891,Adjusted R-squared:  0.3552
F-statistic: 11.47 on 1 and 18 DF,  p-value: 0.003291
```

A anova: 2×5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
hydrocarbon	1	148.3130	148.31296	11.4658	0.003291122
Residuals	18	232.8344	12.93524	NA	NA



(a)

$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 381.1473$ $S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 1.0650$ $S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 12.5678$

Slope of the regression equation is

$$b_1 = \frac{S_{xy}}{S_{xx}} = 11.8010$$

and intercept of the equation will be

$$b_0 = \frac{1}{n}\left(\sum y - b_1 \sum x\right) = 77.8633$$

So the regression equation will be $y' = 77.8633 + 11.801x$

(b)

Let us find SSE first :

$$SSE = \sum y^2 - b_0 \sum y - b_1 \sum xy = 232.8344$$

So standard error of estimate will be

$$S_e = \sqrt{\frac{SSE}{n-2}} = 3.5966$$

$$s_{b_1} = \frac{S_e}{\sqrt{S_{xx}}} = 3.4851$$

T-statistics is

$$t = \frac{b_1 - 0}{s_{b_1}} = 3.386$$

Degree of freedom of test is $df = n - 2 = 20 - 2 = 18$ P-value of the test: 0.0033 Since p-value is less than 0.05 so we reject the null hypothesis.

(c)

The coeffcient of correlation is :

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.6238$$

The coeffcient of determination is: $r^2 = 0.6238 \cdot 0.6238 = 0.3891$

(d)

For $df = 18$ critical value of t for 95% confidence interval is 2.101. So confidence interval is

$$b_1 \pm t_c s b_1 = 11.801 \pm 2.101 \cdot 3.4851 = 11.801 \pm 7.322 = (4.479, 19.123)$$

(e)

```
[39]: predict(fit,data.frame(hydrocarbon=1.05),level=0.95,interval="confidence")
```

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	90.25436	88.30605	92.20268

(f)

The coeffcient of correlation is :

$$\operatorname{Cov}(Y, X) = r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.6238$$

(g)

from (b), $ t_{1} = 3.386$

p-value: $0.003291 < \alpha$

Therefore, reject the $ h_{0}$

P2

September 11, 2021

[1]:
```
#2
```

we have $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and $\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \sum yici \quad ci = \frac{x_i - \bar{x}}{s_{xx}}$

$s_{xy} = \sum y_i \left(x_i - \bar{x}\right)\ s_{xx} = \sum \left(x_i - \bar{x}\right)^2$

$$\begin{aligned}
\operatorname{Cov}\left(\bar{y}_1 b_1\right) &= E[\bar{y} - (E(\bar{y}))]\left[\beta_1 - E\left(\hat{\beta}_1\right)\right] \\
&= E\left[\hat{E}\left(\Sigma c_i y_j - \beta_1\right)\right] \\
&= \frac{1}{n}\left[\left(\Sigma \epsilon_i\right)\left(\beta_0 \Sigma c_i + \beta_1 \Sigma c_i x_i + \varepsilon c_i \varepsilon_i\right)\right. \\
&= \frac{1}{n}[0 + 0 + 0 + 0] = 0
\end{aligned}$$

$$\begin{aligned}
\operatorname{Cov}\left(\bar{y}, \hat{\beta}_1\right) &= \operatorname{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} c_i y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} c_i \operatorname{Var}\left(y_i\right) \\
&= \frac{\sigma^2}{n}\sum_{i=1}^{n} c_i = 0
\end{aligned}$$

P3

September 11, 2021

$\operatorname{Cov}^2(X,Y) \le \operatorname{Var}(X)\operatorname{Var}(Y)$

It follows from this Cauchy-Schwarz inequality that the correlation coefficient is between -1 and 1 .

$$-\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)} \le \operatorname{Cov}(X,Y) \le \sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}$$

Therefore,

$$-1 \le \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} \le 1$$

Actually the covariance is the inner product between two random variables and the standard deviation is the norm of a random variable. If we denote the inner product $\langle X,Y\rangle$ and the norm $|X|$, then the usual Cauchy-Schwarz inequality still holds: $< X,Y >^2 \le |X|^2|Y|^2$

The correlation coefficient is in fact the cosine of the angle between two variables:

$$\operatorname{Corr}(X,Y) = \frac{\langle X,Y\rangle}{|X||Y|} = \cos(\theta)$$

which is between -1 and 1.

[ ]:

P4

September 11, 2021

[1]:
```
#4
```

$$
\begin{aligned}
\mathrm{Var}[\hat{Y}_h] =& \mathrm{Var}[\hat{\beta}_0\bar{x} + \hat{\beta}_1 x_h] \\
=& \mathrm{Var}[\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_h] \\
=& \mathrm{Var}[\bar{y} + \hat{\beta}_1(x_h - \bar{x})] \\
=& \mathrm{Var}[\bar{y}] + \mathrm{Var}[\hat{\beta}_1(x_h - \bar{x})] \\
=& \frac{\sigma^2}{n} + \mathrm{Var}[\hat{\beta}_1(x_h - \bar{x})]
\end{aligned}
$$

2.9

September 11, 2021

(a) Var(Y) = (SSR + SSE)/(n − 1) = 0.0050, Cor(Y, X) is the positive square root of R^2 (positive because $\hat{\beta}_1 > 0$), which is 0.631.

(b) The estimated participation rate would be $0.203311 + 0.656040(0.45) = 0.4985$.

(c) With $\alpha = .05, n = 19$, and the information provided in Table 2.10, we may employ formula (2.38) to obtain the 95 percent confidence interval for our prediction in part (b). The result is $0.4985 \pm 2.11(0.0566)\sqrt{1 + \frac{1}{19} + \frac{(0.45-.5)^2}{18 \cdot \mathrm{Var}(X)}} = .4985 \pm .1241$.

(d) We may use the computer output and formula (2.34) to obtain $0.6560 \pm$ $2.11(0.1961) = 0.6560 \pm 0.4137$ as the 95 percent confidence interval for β_1.

(e) The critical value for the test statistic is 1.74. However, we can see that the test statistic will be negative without actually computing it; therefore, we may automatically conclude that the null hypothesis will not be rejected.

(f) R^2 would not change because $R^2 = (\mathrm{Cor}(Y, X))^2$ and $\mathrm{Cor}(Y, X) = \mathrm{Cor}(X, Y)$.