Homework 8 Solution

Chelsey Green

11/28/2018

Completion points: 10, Accuracy Points: 40, Total points: 50

1. A study was conducted to explore the effects of ethanol on sleep time. Fifteen rats were randomized to one of three treatments. Treatment 1 got only water (control). Treatment 2 got 1g of ethanol per kg of body weight, and treatment 3 got 2g/kg. The amount of REM sleep in a 24hr period was recorded, in minutes. Data are below:

   Treatment 1: 63, 54, 69, 50, 72
   Treatment 2: 45, 60, 40, 56
   Treatment 3: 31, 40, 45, 25, 23, 28

   (a) Make a preliminary graph of the data (ok to do by hand if you prefer). Why did you choose the graph that you did and what does it tell you?

   +2 reasonable graph *ANSWER: I will make a dot plot for each group so I can assess whether the groups look to have equal variance and differences in means. There looks to be approximately equal variance and differences between the group means.*

   (b) Calculate relevant summary statistics that will be useful in an ANOVA analysis.

   ANSWER: group and overall means, size and standard deviation

mean	$\bar{y_{1i}} = 61.6$	$\bar{y_{2i}} = 50.25$	$\bar{y_{3i}} = 32$	$\bar{y_{..}} = 46.73$
sd	$s_1 = 9.45$	$s_2 = 9.32$	$s_3 = 8.72$	
replicates	$n_1 = 5$	$n_2 = 4$	$n_3 = 6$	$N = 15$

   (c) Create an ANOVA table for the data using the original or summary values above. Show your work. You may use R to check your answers.

   +8 points df, SS for all, and MS for TRT and Error $SSTreat = 5 * (61.6 - 46.73)^2 + 4 * (50.25 - 46.73)^2 + 6 * (32 - 46.75)^2 = 2460.521$
   $SSERR = (5-1) * (9.45)^2 + (4-1) * (9.32)^2 + (6-1) * (8.72)^2 = 2460.521 = 997.9892$
   $SSTot = SSTreat + SSErr = 2460.521 + 997.9892 = 3458.51$

Source	*df*	*SS*	*MS*	*F*	*p-value*
TRT	$3 - 1 = 2$	*2456.983*	*1228.492*	*14.77*	*0.0006*
Err	$15 - 3 = 12$	*997.95*	*83.16*		
Total	$15 - 1 = 14$	*3454.933*			

   (d) Evaluate the ANOVA assumptions numerically and graphically. Was ANOVA appropriate for this data?

   +4 points; 2 for each qqplot and residual plot *ANSWER: we need a residuals vs fitted values (sample means) graph and a QQ plot of residuals. These graphs show that the sample variances are fairly consistent across treatments. We also see the ratio between the smallest (8.72) and largest (9.45) sd is within a factor of 2. The residuals (shown within the code below) show some s-shape, but it is still close enough to linear to be ok. These will not be boarderline differences, so I think ANOVA will be ok.*

   (e) Based on the ANOVA table, make a conclusion in the context of the problem.

+3 points F, approx p value, reject null *ANSWER: Such a large F value and small p value gives us strong evidence against the null of no difference in means between groups. The evidence suggests at least two of the three treatment groups have different group means from the other.*

(f) Use R to obtain the relevant multiplier and then create 95% CIs for all pairwise comparisons of means using the Tukey method. Do this by hand and show your work. You may use R to check your answers. Summarize your results using letter codes. What do you conclude?

+6 points; 2 each for CI $\mu_1 - \mu_2 : 61.60 - 50.25 \pm 2.668\sqrt{83.16(1/5 + 1/4)} = (-4.971, 27.6711)$
$\mu_1 - \mu_3 : 61.60 - 32 \pm 2.668\sqrt{83.16(1/5 + 1/6)} = (14.867, 44.332)$
$\mu_2 - \mu_3 : 50.25 - 32 \pm 2.668\sqrt{83.16(1/4 + 1/6)} = (2.545, 33.95)$

Treatment	*Sample Mean*	*Letter*
3	*32.00*	*A*
2	*50.25*	*B*
1	*61.60*	*B*

There is a significant difference between μ_1 and μ_3 and μ_2 and μ_3, but not significant difference between μ_1 and μ_2 when using Tukey pairwise comparisons.

(g) Calculate the difference between the Tukey CI t multiplier and that used with a Bonferroni adjustment.

Tukey Multiplier: 2.668, vs Bonferroni: 2.779473, difference: 2.779473-2.668=0.111473

(h) How does your conclusion change, if at all, if instead you chose to use the Kruskal- Wallis followed by pairwise comparisons with bonferroni adjustment?

Overall Kruskal-Wallis chi-squared = 9.8535, df = 2, p-value = 0.00725 (less strong evidence against null- more conservative.) Pairwise comparison: 1 vs 2 p-value $= 0.1905 * 3 = 0.5715$*, 1 vs 3 p-value* $= 0.004329 * 3 = 0.012987$*, 2 vs 3 p-value* $= 0.04157 * 3 = 0.12471$*. The pairwise comparisons are also more conservative, only finding a significant difference between groups 1 and 3.*

2. A study was conducted to compare the effect of three diet types on the milk yield of cows (in lbs). The sample size, sample mean, and sample variance for each method are given below.

Diet A: $n_1 = 9$, $\bar{x}_{1\cdot} = 39.1$, $s_1^2 = 24.6$

Diet B: $n_2 = 8$, $\bar{x}_{2\cdot} = 29.9$, $s_2^2 = 16.4$

Diet C: $n_3 = 10$, $\bar{x}_{3\cdot} = 45.9$, $s_2^2 = 10.3$

(a) Construct an ANOVA table including all relevant sums of squares, mean squares, and degrees of freedom.

ANSWER: Either compute SSTrt $= 9\times(39.1-38.9)^2+8\times(29.9-38.9)^2+10\times(45.9-38.9)^2 = 1138.36$ *or SSErr* $= (9-1)\times 24.6 + (8-1)\times 16.4 + (10-1)\times 10.3 = 404.30$*. The ANOVA table therefore is*

Source	df	SS	MS
Diet type	2	1138.36	569.18
Error	24	404.30	16.84
Total	26	1542.66	

(b) Perform an overall F test to determine whether the population means of milk yield are the same or not among the three diet types.

ANSWER: To test H_0 : all the population means are the same vs. H_A : Atleast two population means are different. The observed F test statistic is $F_{obs} = \frac{569.18}{16.84}$ on (2, 24) df. The p-value is $P(F_{2,24} > F_{obs})$ which is < 0.01 and so we would reject H_0 at level 0.05.

3. Suppose we are interested in exploring the relationship between city air particulate and rates of childhood asthma (data not from actual tests). We sample 15 cities for particulate measured in parts-per-million (ppm) of large particulate matter and for the rate of childhood asthma measured in percents. The data are as follows: (Do some of this by hand from sample summary statistics, check calculations in R - I won't ask you to do anything too crazy, but you will have to do some computations by hand.). You can also use the fact that: $\sum_{i=1}^{15}(y_i - \bar{y})(x_i - \bar{x}) = 79.134$

particulate (x)	11.6	15.9	15.7	7.9	6.3	13.7	13.1	10.8	6.0	7.6	14.8	7.4	16.2	13.1	11.2
asthma % (y)	14.5	16.6	16.5	12.6	12.0	15.8	15.1	14.2	12.2	13.1	16.0	13.5	16.4	15.4	14.4

variable:	size	mean	variance
particulate	15	11.42	13.05
asthma	15	14.55	2.52

(a) Plot the data as you see fit and summarize the pattern's shape, direction, and strength in the context of the problem.

ANSWER: There appears to be a strong, positive, linear pattern between level of particulate and percent asthma for the range of values observed in this data set.

(b) **+3 value** Calculate the correlation coefficient by hand and confirm your findings in R. *ANSWER:* $r = \frac{79.134}{\sqrt{14*13.05*14*2.52}} = 0.9857$. *This value is very close to the exact value calculated in R: 0.9854*

(c) Build a linear regression model for the data (in R or by hand using the summary statistics) with the estimated values for the slope and intercept. Interpret the intercept and slope in the context of the question.

+4 linear model (no points for interpretation) *ANSWER:* $\hat{asthma\%} = 0.433 * (particulate) + 9.6$ *estimated intercept: 9.6 (this is the estimated average rate of asthma in cities with 0 particulate - this value is outside the range of our data), estimated slope: 0.433. This suggests that for each unit increase in particulate, measured cities will tend to exhibit an increase of 0.43 percent particulate in the rate of childhood asthma.*

(d) Graphically assess whether the assumptions for linear regression are met.

Given the small sample size, there are no clear deviations from constant variance or clear curvature of the residuals. Thee is 1 larger residual for value 12 that makes it look like without that, equal variance would would be ok.

(e) Conduct a formal test to determine whether or not the slope is significantly different from 0 (In R or by hand). If the slope is significant, can we, using only this statistical information, conclude that increased particulate concentration causes an increase in childhood asthma?

ANSWER: $H_o : \beta_1 = 0$ *vs* $H_o : \beta_1 \neq 0$, *is* $p = 2.21 * 10^{-11}$ *which is much smaller than* $\alpha = 0.05$. *Therefore the slope is highly significant. However, we cannot conclude that increased particulate*

causes increased asthma. This is a plausible conclusion if we include knowledge of biology and other simular studies, but these observational data by themselves are not sufficient for such a conclusion.

(f) Calculate and interpret the R^2 value in the context of the problem.

$R^2 = 0.971$ *so the linear regression on paarticulate explains about* 97.1% *of the variation in the asthma rates.*

(g) Suppose we sample a new city whose particulate is 13 ppm. If reasonable, create a 95% confidence interval for the predicted rate of childhood asthma in this city. If not reasonable, explain why.

+4 total, +2 pt estimate,+2 correct SE *ANSWER: We use the above summary of the model to estimate the standard deviation of the residuals as* $s = 0.28$. *Additionally, the average particulate matter is* $\bar{x} = 11.42$ *and the variance of the partculate matter is* $s_x^2 = 13.05$. *Therefore, we estimate the standard error of* $\hat{y}$ *conditional on* x^* *as* $SE = 0.28\sqrt{1 + \frac{1}{15} + \frac{(13-11.42)^2}{(15-1)*13.05}} = 0.29$. *So the CI is given by* $(9.61 + 0.43 * 13) \pm 2.16 * 0.29 = 15.2 \pm 0.63 = (14.57, 15.83)$.

(h) If reasonable, create a different 95% confidence interval for the average rate of childhood asthma among cities with 10 ppm of large particulate. If not reasonable, explain why.

+4 total, +2 pt estimate,+2 correct SE *ANSWER: We use the above summary of the model to estimate the standard deviation of the residuals as* $s = 0.28$. *Additionally, the average particulate matter is* $\bar{x} = 11.42$ *and the variance of the partculate matter is* $s_x^2 = 13.05$. *Therefore, we estimate the standard error of* $\hat{y}$ *conditional on* x^* *as* $SE = 0.28\sqrt{\frac{1}{15} + \frac{(10-11.42)^2}{(15-1)*13.05}} = 0.078$. *So the CI is given by* $(9.61 + 0.43 * 10) \pm 2.16 * 0.078 = 13.91 \pm 0.17 = (13.74, 14.08)$.

(i) If reasonable, create a different 95% confidence interval for the predicted rate of childhood asthma in a city with 3 ppm of large particulate. If not reasonable, explain why.

+2 pts - cant calculate *Since 3 ppm is far outside of the values of x that were used to build the model, it would be extrapolation and therefore we should not be confident in the predictions we would get.*

(j) Will a 95% confidence interval for the average rate of childhood asthma constructed at x=12 be wider/narrower/same width as one constructed at x=7? Explain your answer.

The interval at x=12 will be narrower because it is closer to $\bar{x} = 11.42$ *than 7. This will make a smaller standard error and thuse a narrower CI.*

4. A chemist is calibrating a spectrophotometer that will be used to measure the concentration of carbon monoxide (CO) in atmospheric samples. To check the calibration, 11 samples of known concentration are measured. The summary measures for the true concentrations (x) and the measured concentrations (y), in parts per million, are given in the following table.

$\bar{x}$	50
$\bar{y}$	47.91
sd_x	33.17
sd_y	31.25
$\sum_{i=1}^{11}(x_i-\bar{x})^2$	11000
$\sum_{i=1}^{11}(y_i-\bar{y})^2$	9768.91
$\sum_{i=1}^{11}(y_i-\bar{y})(x_i-\bar{x})$	10360
$\sum_{i=1}^{11}(y_i-\hat{y}_i)^2$	11.67

To check the calibration, the linear model $y = \beta_0 + \beta_1 x + \epsilon$ is fit. Ideally, if the machine is properly calibrated, β_0 should be 0 and β_1 should be 1.

(a) Compute the least square estimates of β_0 and β_1 by hand.

ANSWER: $\hat{\beta}_1 = \frac{\sum_{i=1}^{11}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{11}(x_i-\bar{x})^2} = \frac{10{,}360}{11{,}000} = 0.9418$, *and* $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 47.91 - (.9418) * 50 = 0.819$

ANSWER: 1st plot: scatter plot of x: True, y: Measured to ensure a linear model is appropriate. 2nd plot: scatter plot of x: fitted value from the model and y: residuals from model to ensure residual variance is relatively constant across the different value of x. 3rd plot: qqnorm plot of residuals to ensure it is reasonable that our sample of errors came from normal population

(b) Perform the following hypothesis test at significance level $\alpha = .01$:

$H_0 : \beta_1 = 1$
$H_1 : \beta_1 \neq 1.$

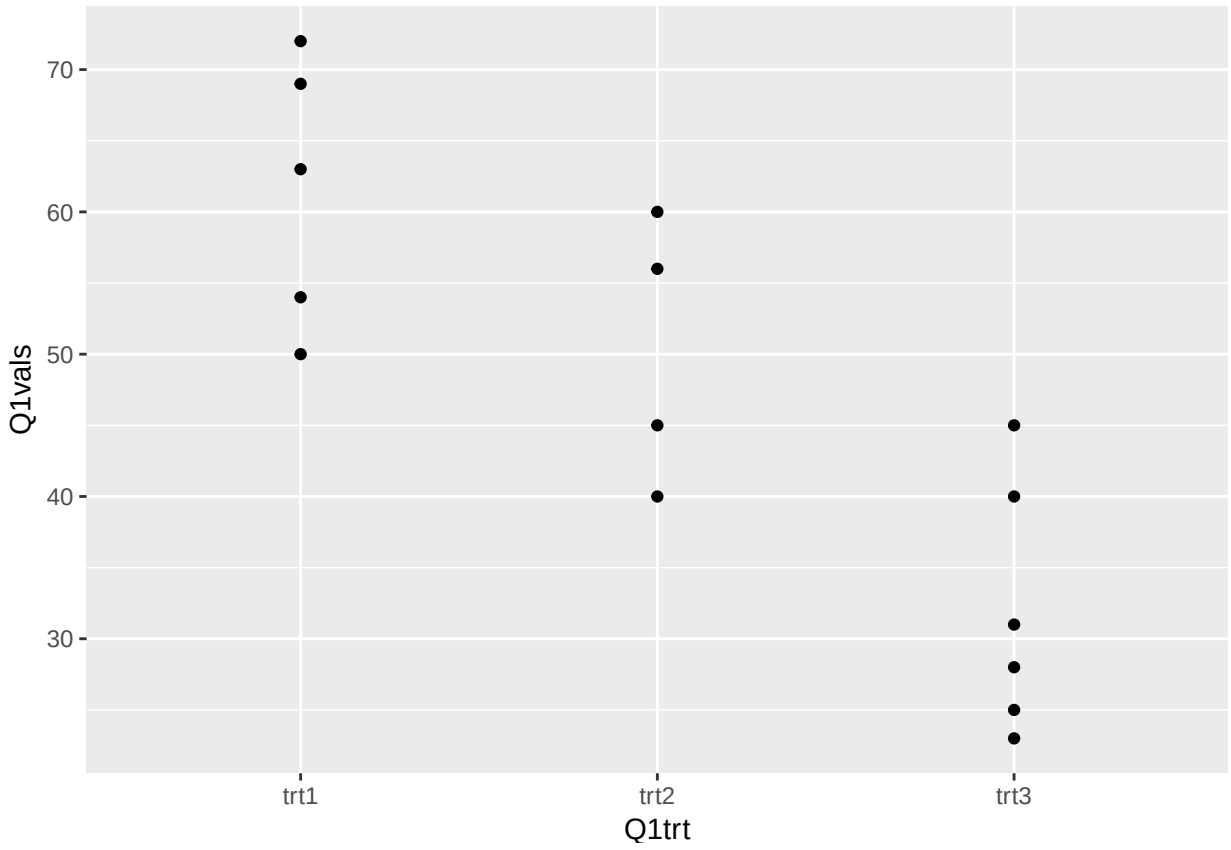
ANSWER: To test $H_0 : \beta_1 = 1$ *vs* $H_1 : \beta_1 \neq 1$, $SE(\hat{\beta}_1) = \frac{\sqrt{MSErr}}{\sqrt{\sum_{i=1}^{11}(x_i-\bar{x})^2}} = \frac{\sqrt{\frac{SSerr}{n-2}}}{\sqrt{\sum_{i=1}^{11}(x_i-\bar{x})^2}} = \frac{\sqrt{\frac{11.67}{9}}}{\sqrt{11{,}000}} = 0.0108572$. $T_{obs} = \frac{\hat{\beta}_1 - 1}{SE(\hat{\beta}_1)} = \frac{.9418-1}{0.0108572} = -5.36$. *pvalue*$= 2 * P(t_9 > 5.36) < 0.001$. *This means we can reject* $H_o : \beta_1 = 1$ *at level 0.01.*

Q1 Analysis

```
trt1<-c(63, 54, 69, 50, 72)
trt2<-c(45, 60, 40, 56)
trt3<-c(31, 40, 45, 25, 23, 28)

Q1vals=c(trt1, trt2, trt3)
Q1trt=c(rep("trt1", times=5), rep("trt2", times=4), rep("trt3", times=6))
Q1data=data.frame(Q1trt, Q1vals)

require(ggplot2)
```

```
## Loading required package: ggplot2
```

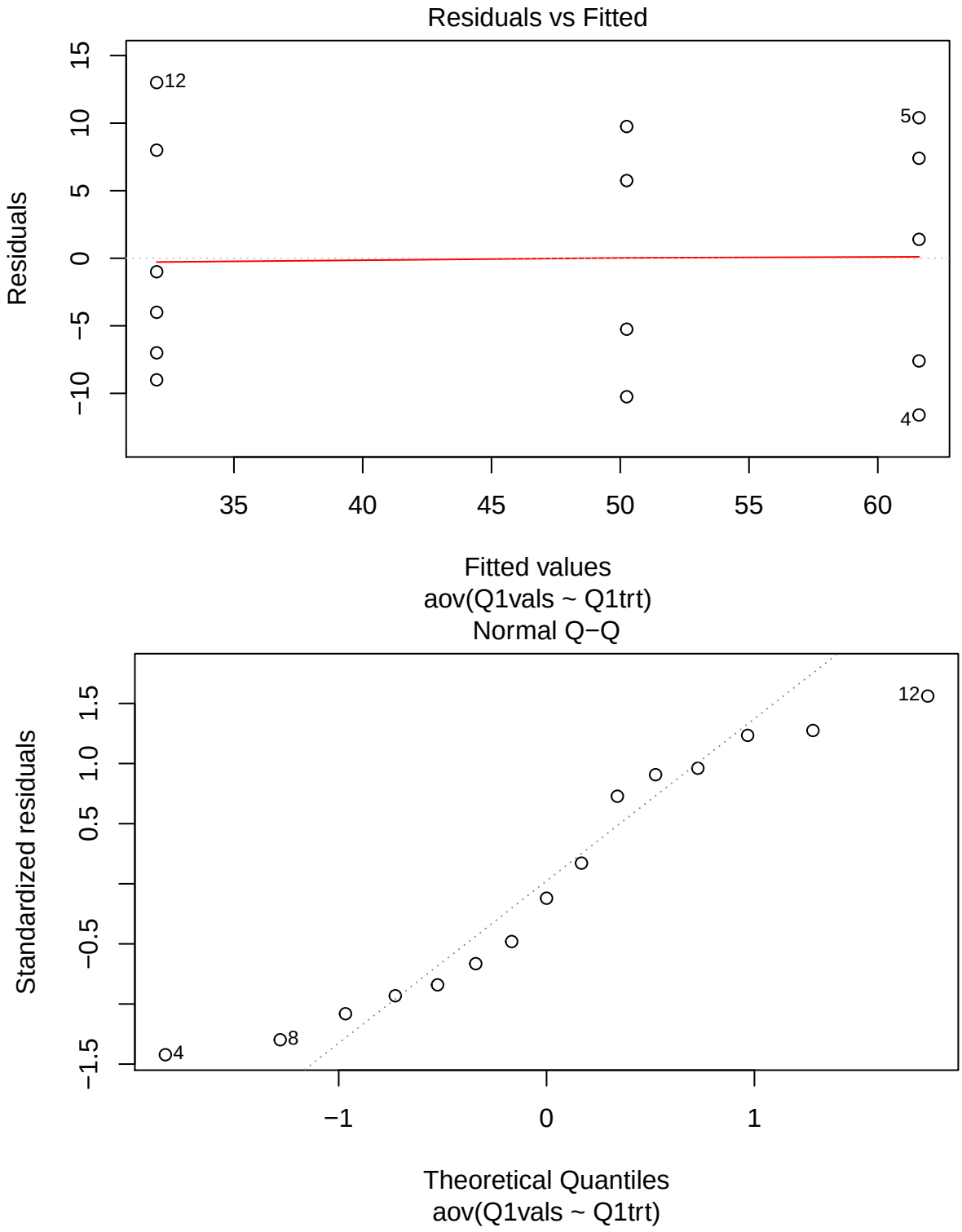
```
ggplot(data=Q1data, aes(x=Q1trt, Q1vals))+
  geom_point()
```

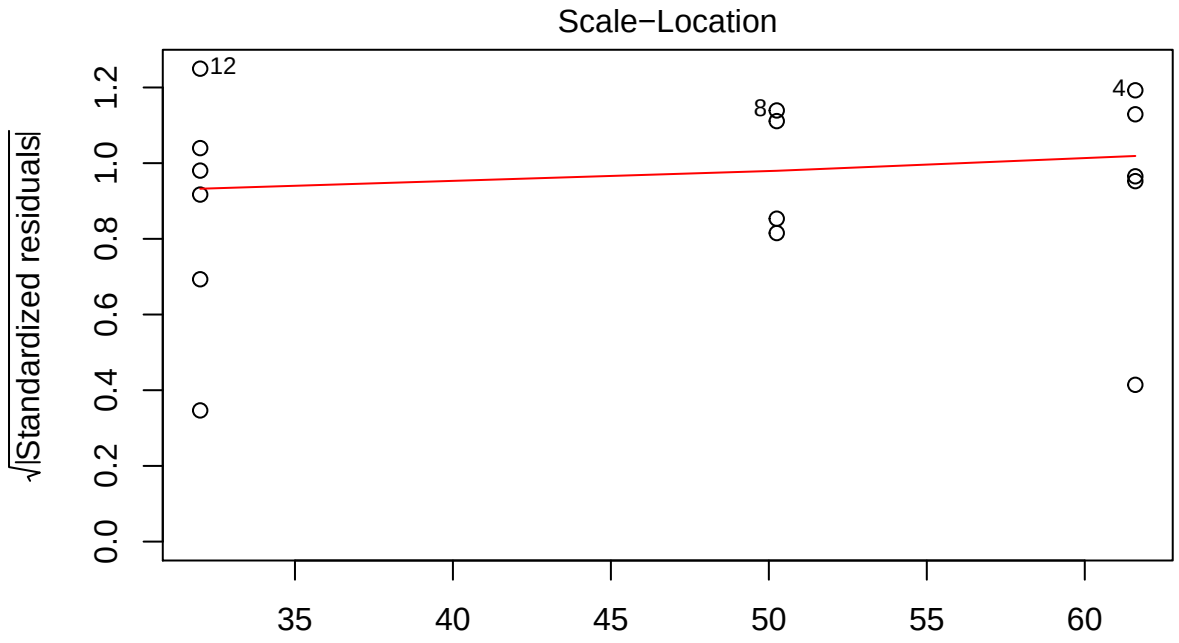


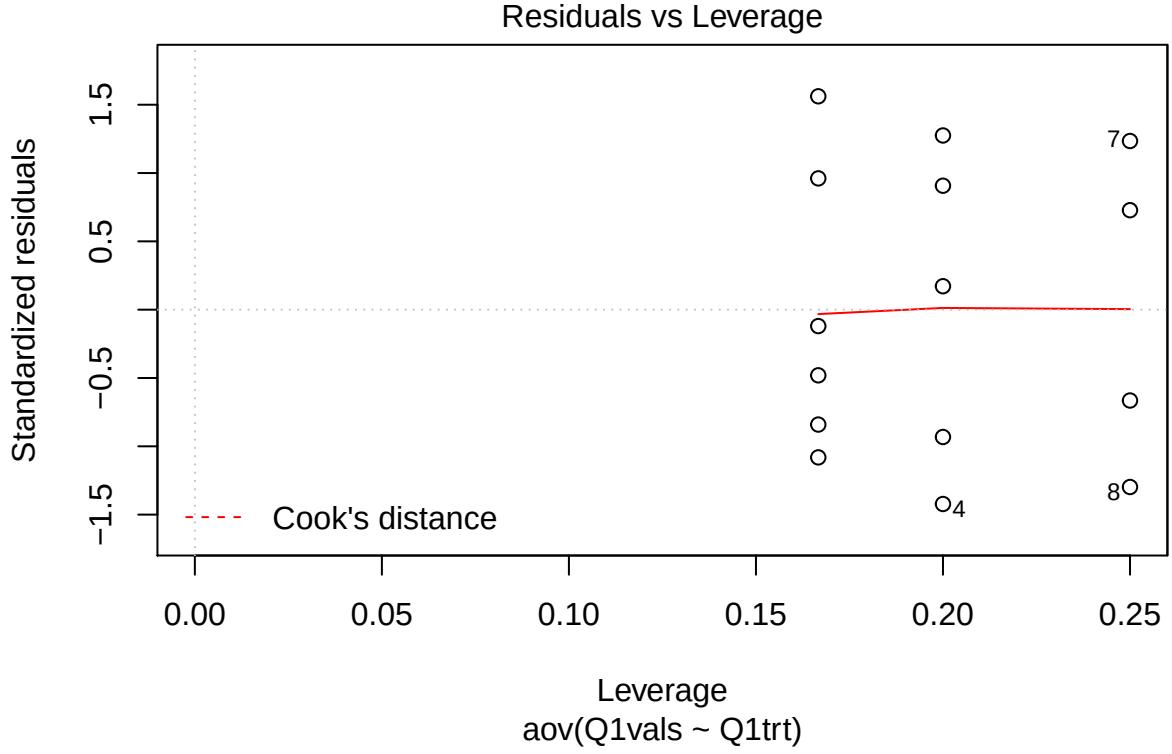

```
tr1.mean=mean(trt1) ; tr1.sd=sd(trt1)  #mean=61.6, sd=9.45
tr2.mean=mean(trt2); tr2.sd=sd(trt2)   #mean=50.25, sd=9.32
tr3.mean=mean(trt3); tr3.sd=sd(trt3)   #mean=32, sd=8.72

Q1data$Q1over_mean<-mean(Q1data$Q1vals)
Q1data$Q1gp_means<-c(rep(tr1.mean, times=5), rep(tr2.mean, times=4), rep(tr3.mean, times=6))

SSTrt<-sum((Q1data$Q1gp_means-Q1data$Q1over_mean)^2)
SSErr<-sum((Q1data$Q1vals-Q1data$Q1gp_means)^2)
SSTot<-sum((Q1data$Q1vals-Q1data$Q1over_mean)^2)

Q1mod<-aov(Q1vals~Q1trt)
summary(Q1mod)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Q1trt        2   2457  1228.5   14.77 0.000581 ***
## Residuals   12    998    83.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(Q1mod)[1:2]
```


Residuals vs Fitted
Residuals
15
10
5
0
−5
−10
12
5
4
35
40
45
50
55
60
Fitted values
aov(Q1vals ~ Q1trt)
Normal Q−Q
Standardized residuals
1.5
1.0
0.5
−0.5
−1.5
12
8
4
−1
0
1
Theoretical Quantiles
aov(Q1vals ~ Q1trt)






```
## NULL
```

```
qtukey(p=(1-.05), nmeans=3, df=12)/sqrt(2)
```

```
## [1] 2.667864
```

```r
qtukey(p=.05, nmeans=3, df=12, lower.tail=FALSE)/sqrt(2)  #3.772929/sqrt(2)=2.668
```

```
## [1] 2.667864
```

```r
TukeyHSD(Q1mod, conf.level=0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Q1vals ~ Q1trt)
##
## $Q1trt
##              diff       lwr        upr     p adj
## trt2-trt1 -11.35 -27.67051   4.970507 0.1939650
## trt3-trt1 -29.60 -44.33205 -14.867952 0.0004632
## trt3-trt2 -18.25 -33.95442  -2.545584 0.0231619
```

```r
qt(p=(.05/3)*1/2, df=12, lower.tail=FALSE)
```

```
## [1] 2.779473
```

```r
kruskal.test(Q1data$Q1vals~Q1data$Q1trt)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Q1data$Q1vals by Q1data$Q1trt
## Kruskal-Wallis chi-squared = 9.8535, df = 2, p-value = 0.00725
```

```r
kruskal.test(Q1vals ~ Q1trt, data=Q1data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Q1vals by Q1trt
## Kruskal-Wallis chi-squared = 9.8535, df = 2, p-value = 0.00725
```

```r
wilcox.test(x=trt1, y=trt2)  #W = 16, p-value = 0.1905
```

```
##
##  Wilcoxon rank sum test
##
## data:  trt1 and trt2
## W = 16, p-value = 0.1905
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(x=trt1, y=trt3)  #W = 30, p-value = 0.004329
```

```
##
##  Wilcoxon rank sum test
##
## data:  trt1 and trt3
## W = 30, p-value = 0.004329
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(x=trt2, y=trt3)  #W = 22, p-value = 0.04157
```

```
## Warning in wilcox.test.default(x = trt2, y = trt3): cannot compute exact p-
## value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  trt2 and trt3
## W = 22, p-value = 0.04157
## alternative hypothesis: true location shift is not equal to 0
```

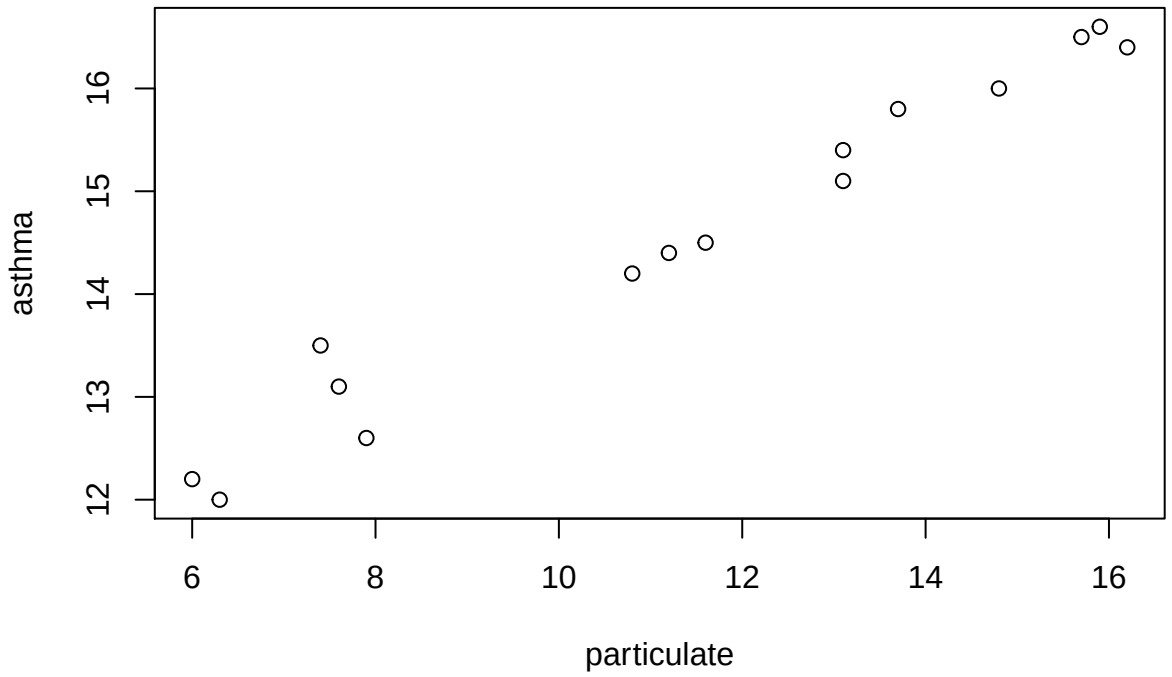
Question 2 Analysis

```
pf(569.18/16.84,2,24, lower.tail=FALSE)
```

```
## [1] 1.046816e-07
```

Question 3 Analysis

```
particulate<-c(11.6, 15.9, 15.7, 7.9, 6.3, 13.7, 13.1, 10.8, 6.0, 7.6, 14.8, 7.4, 16.2, 13.1, 11.2)
asthma<-c(14.5, 16.6, 16.5, 12.6, 12.0, 15.8, 15.1, 14.2, 12.2, 13.1, 16.0, 13.5, 16.4, 15.4, 14.4)
plot(x=particulate, y=asthma)
```




```
mpart=mean(particulate); var(particulate)
```

```
## [1] 13.05029
```

```
masth=mean(asthma); var(asthma)
```

```
## [1] 2.521238
```

```
sum((particulate-mpart)*(asthma-masth))
```

```
## [1] 79.134
```

```
r=cor(particulate, asthma)
r_hand=79.134/sqrt(14*13.05*14*2.52)

slope_hand=r*sqrt(2.52)/sqrt(13.05)
slope_hand2=79.134/(14*13.05)
```

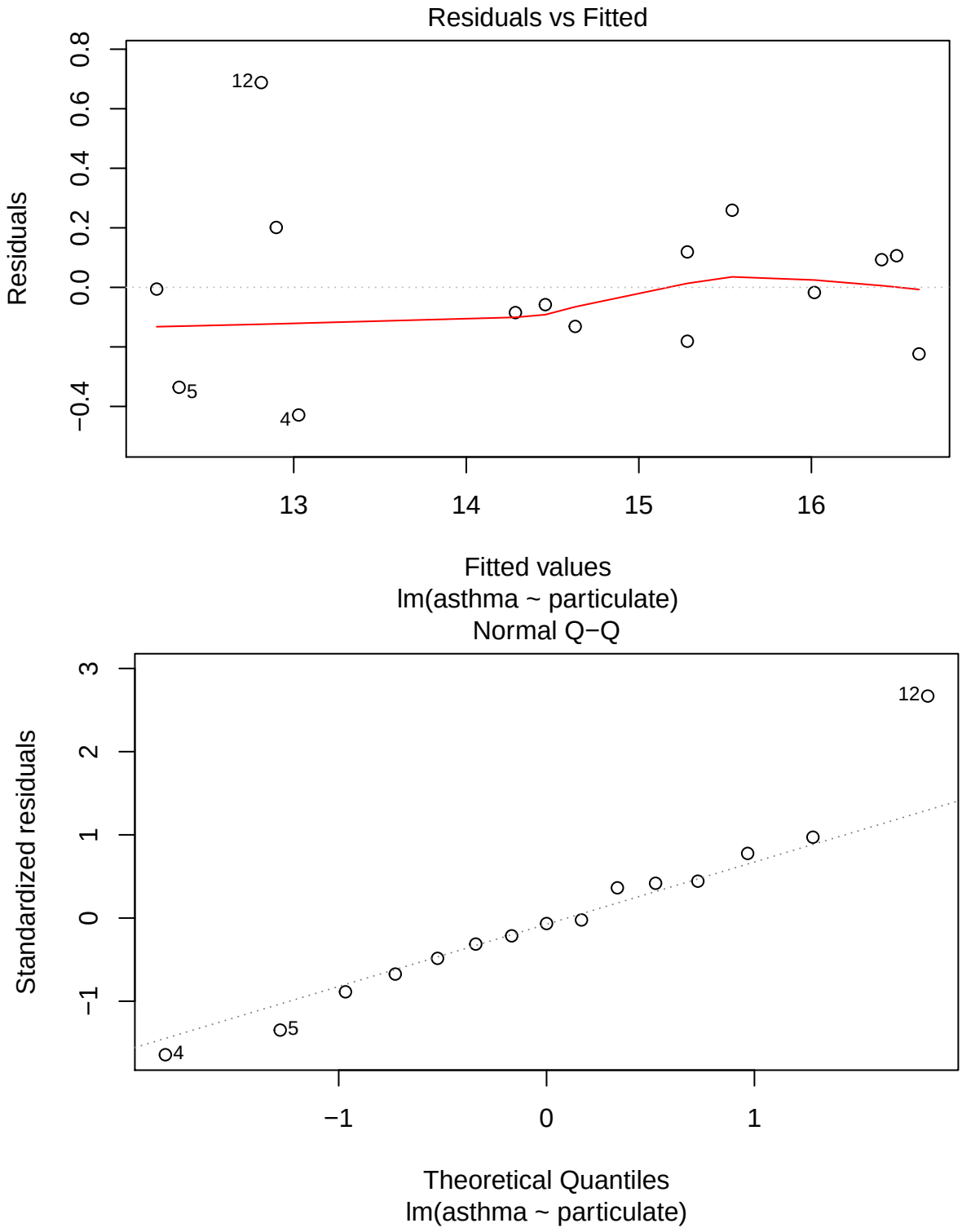
```r
Q3mod<-lm(asthma~particulate)
summary(Q3mod)
```

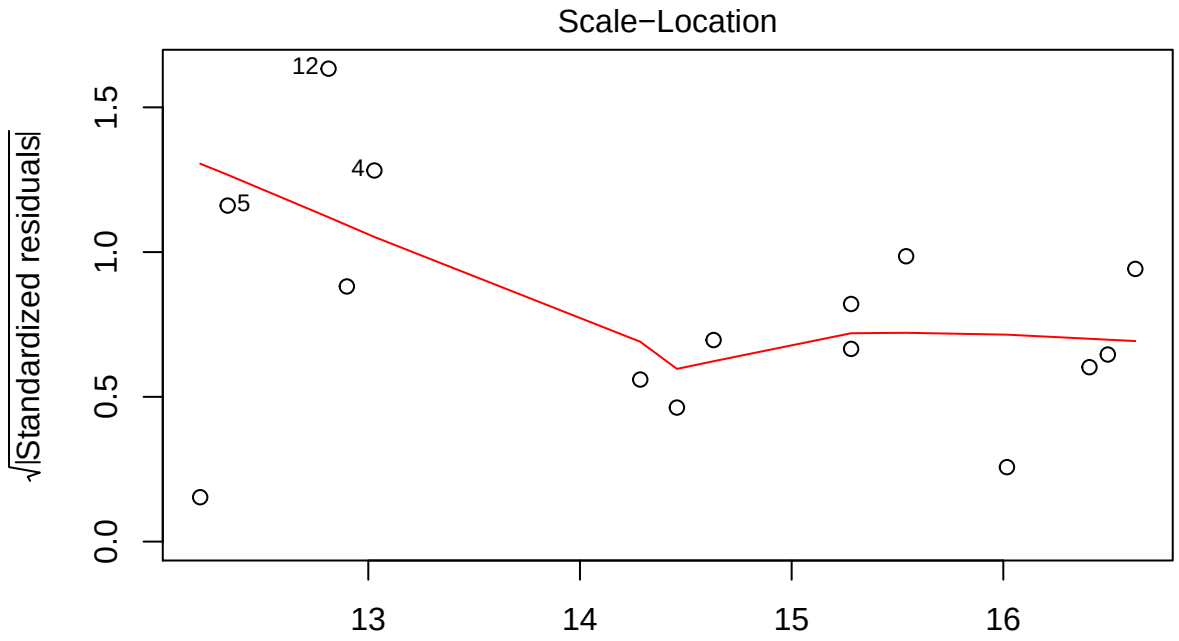
```
##
## Call:
## lm(formula = asthma ~ particulate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4287 -0.1561 -0.0173  0.1126  0.6878
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.60703    0.24774   38.78 8.00e-15 ***
## particulate  0.43313    0.02075   20.88 2.21e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2804 on 13 degrees of freedom
## Multiple R-squared:  0.971,  Adjusted R-squared:  0.9688
## F-statistic: 435.9 on 1 and 13 DF,  p-value: 2.208e-11
```

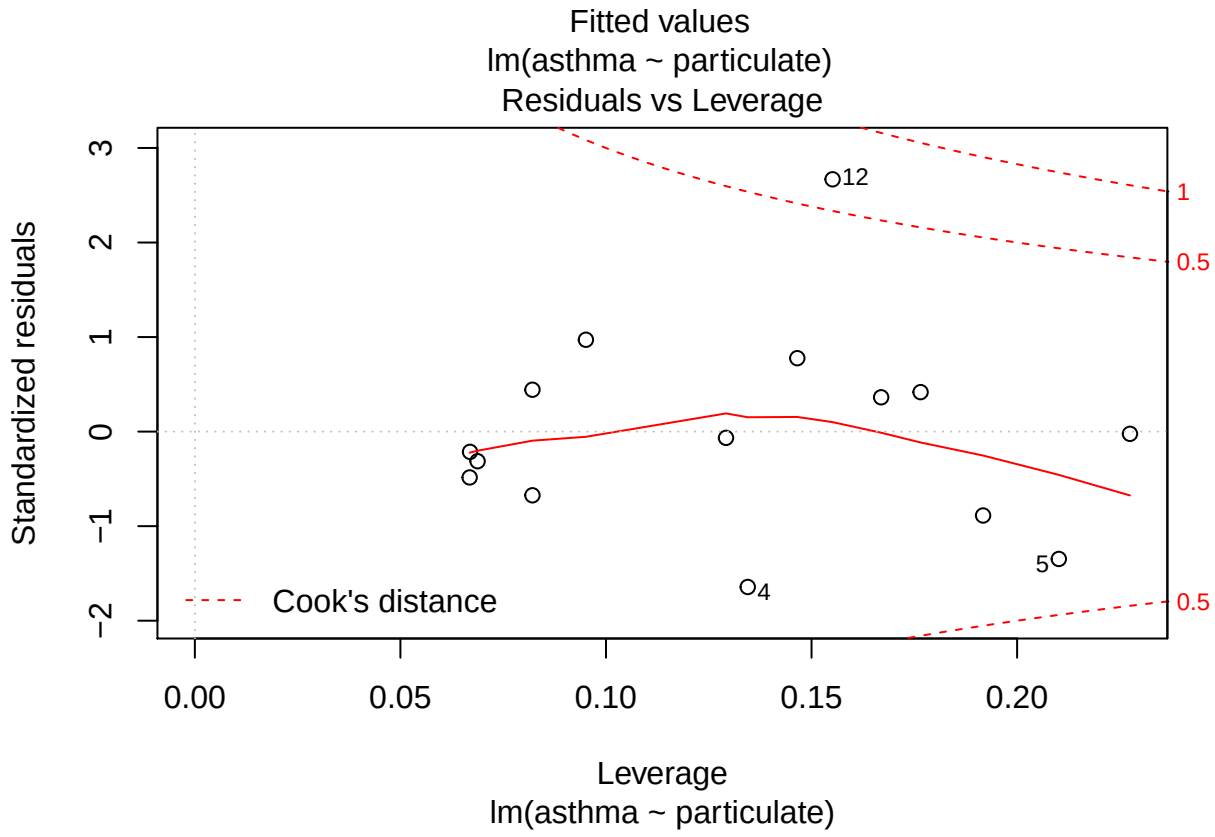
```r
anova(Q3mod)
```

```
## Analysis of Variance Table
##
## Response: asthma
##             Df Sum Sq Mean Sq F value    Pr(>F)
## particulate  1 34.275  34.275  435.87 2.208e-11 ***
## Residuals   13  1.022   0.079
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(Q3mod)
```


Residuals vs Fitted
Residuals
0.8
0.6
0.4
0.2
0.0
−0.4
12
5
4
13
14
15
16
Fitted values
lm(asthma ~ particulate)
Normal Q−Q
Standardized residuals
3
2
1
0
−1
12
5
4
−1
0
1
Theoretical Quantiles
lm(asthma ~ particulate)






```
sqrt(1.022/13)
```

```
## [1] 0.2803844
```

Question 4 Analysis

```r
x=c(0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100)
y=c(1, 11, 21, 28, 37, 48, 56, 68, 75, 86, 96)
mean(x); sd(x)
```

```
## [1] 50
```

```
## [1] 33.16625
```

```r
sum((y-mean(y))*(x-mean(x)))
```

```
## [1] 10360
```

```r
Q4mod<-lm(y~x)
summary(Q4mod)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4909 -1.1182  0.1818  0.8818  1.3455
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.81818    0.64240   1.274    0.235
## x            0.94182    0.01086  86.736 1.82e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.139 on 9 degrees of freedom
## Multiple R-squared:  0.9988, Adjusted R-squared:  0.9987
## F-statistic:  7523 on 1 and 9 DF,  p-value: 1.824e-14
```

```r
sum((Q4mod$residuals)^2)
```

```
## [1] 11.67273
```

```r
1.139/sqrt(11000)
```

```
## [1] 0.01085994
```

```r
sqrt(11.67/9)/sqrt(11000)
```

```
## [1] 0.0108572
```

```r
(.9418-1)/0.0108572
```

```
## [1] -5.360498
```

```r
2*pt(-5.36, df=9)
```

```
## [1] 0.000456314
```