HW8

Teryl Schmidt | [*tschmidt6@wisc.edu*](mailto:tschmidt6@wisc.edu) *| 9072604920 | Discussion 312 | Grader: Chi-Shain Dai*
12/1/2018

Problem 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?

```r
Treatment1 = c(63, 54, 69, 50, 72)
Treatment2 = c(45, 60, 40, 56)
Treatment3 = c(31, 40, 45, 25, 23, 28)
boxplot(Treatment1, Treatment2, Treatment3)

Treatmentdata = data.frame(Treatment = c(rep("Treatment1",5), rep("Treatment2", 4), rep("Treatment3", 6)),
Data = c(Treatment1, Treatment2, Treatment3), gpmeans = c(rep(61.6, 5), rep(50.25, 4), rep(32, 6)))

require(ggplot2)
```
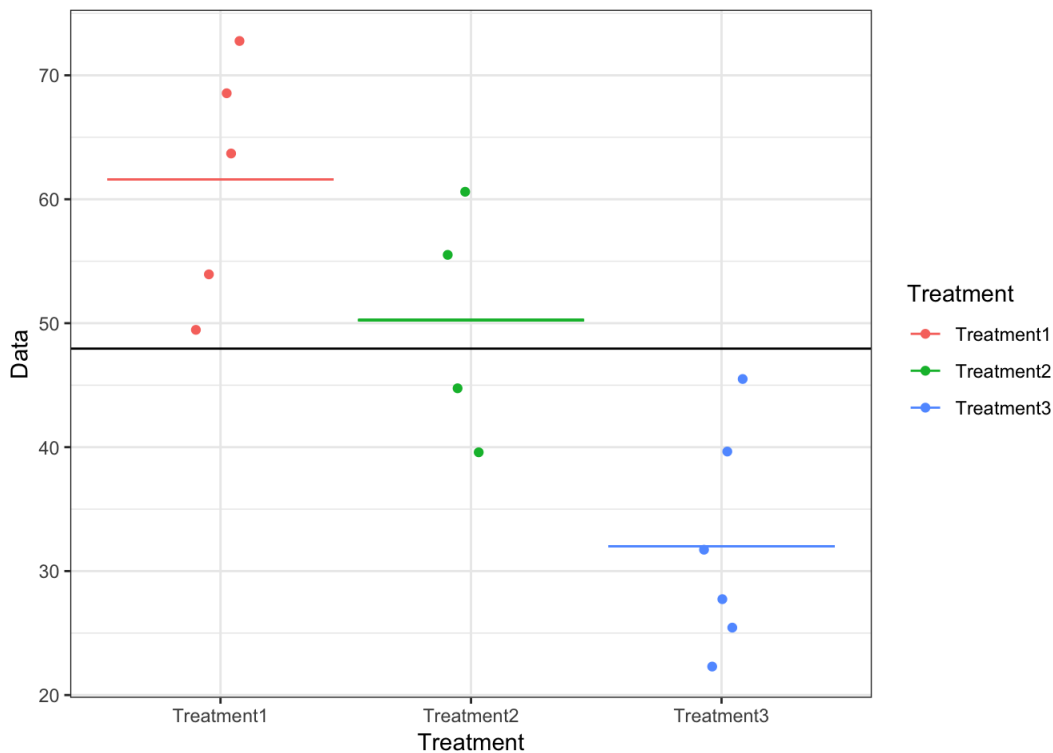
```
## Loading required package: ggplot2
```
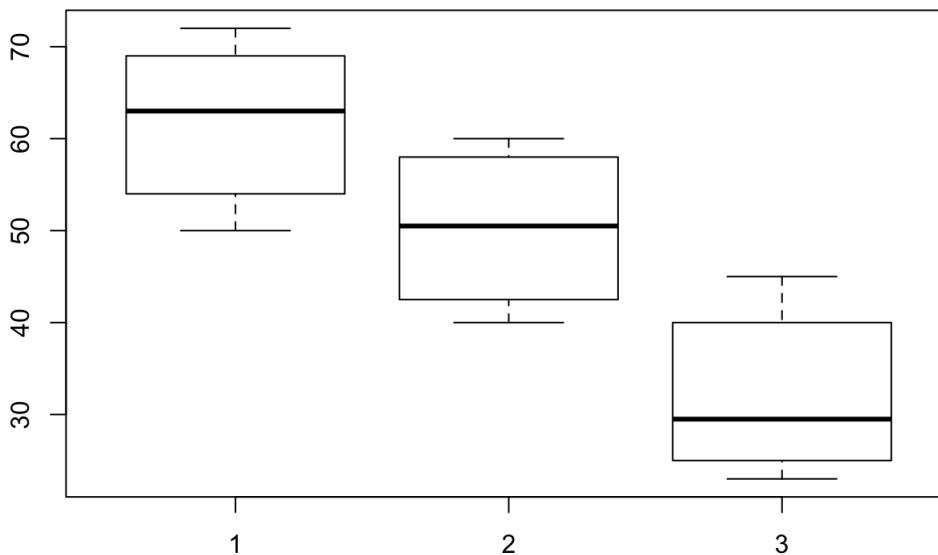


```r
ggplot(data=Treatmentdata, aes(x=Treatment, y=Data, color=Treatment))+
  geom_jitter(width=0.1)+
  theme_bw()+
  geom_errorbar(aes(ymax=gpmeans, ymin=gpmeans))+
  geom_hline(yintercept=47.95)
```

The box plots would show if there was any extreme values in the data, there were no extreme values. Also boxplots show the average differences as all the box are at different level. Treatment 3 is slightly skewed data.

b. Calculate relevant summary statistics that will be useful in an ANOVA analysis.
Null Hypothesis: The mean sleep times for the three treatments are equal:
treatment 1 = treatment2 = treatment3

Atternative Hypothesis: The mean sleep times for the three treatments are not equal:
treatment 1 =/= treatment2 =/= treatment3

```
              Count   |   Sum   |   Mean   |   Variance   |
Treatment 1:    5     |   308   |   61.6   |   9.449868   |
Treatment 2:    4     |   201   |   50.25  |   9.322911   |
Treatment 3:    6     |   192   |   32     |   8.717798   |
```

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.

```
Source         |   SS    |        df        |         MS          |              F               |
 P  |
Treat (between)| 2456.983 | (t-1) = (3-1) = 2  | 2456.983 / 2 = 1228.492 | 1228.492 / 83.1625 = 14.77219 |0
.001|
Error (within) | 997.95  | (N-t) = (15-3) = 12 | 997.95 / 12 = 83.1625   |                              |
    |
Total          | 3454.933 | (N-1) = (15-1) = 14 |                         |                              |
    |
```

(y1 * n1) + (y2 * n2) + (y3 * n3) / (n1 + n2 + n3)
((61.6 * 5) + (50.25 * 4) + (32 * 6)) / (5 + 4 + 6) = 46.7333
(5 * (61.6 - 46.7333)^2) + (4 * (50.25 - 46.7333)^2) + (6 * (32 - 46.7333)^2) = 2456.983
4 * 9.449868^2 + 3 * 9.322911^2 + 5 * 8.717798^2 = 997.95
SSTrt + SSE = 2456.983 + 997.95 = 3454.933
$P(F_{(14.77, 2, 12)} > 12.97) = 0.001$

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

The data are independent within and between treatments:
Here groups are independent (ind. Between treatments)
The dose of ethanol (ind. Within treatments)
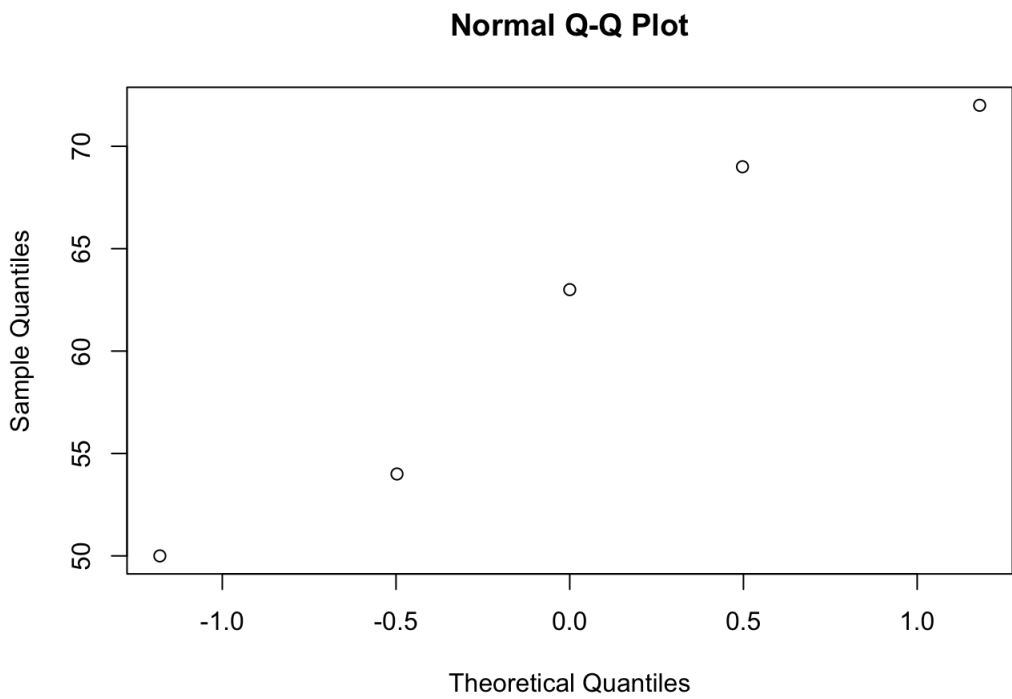
The variances are the same for all treatments:
if SD is between 0.5 - 2
9.449868 / 8.717798 = 1.083974
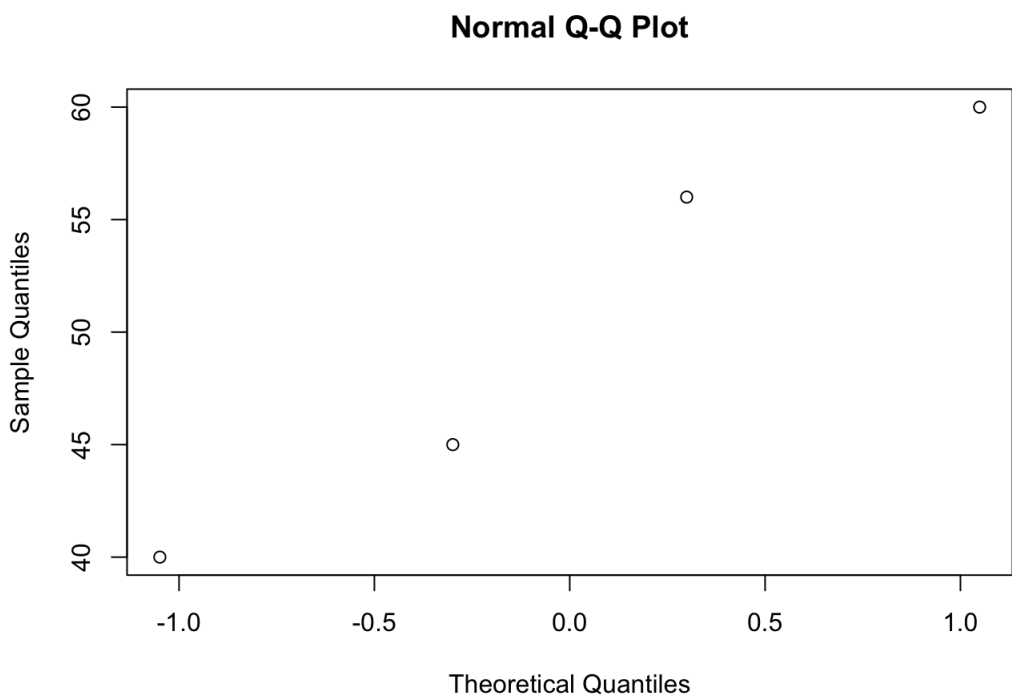So we can assume equal variance

```
qqnorm(Treatment1)
```

Normal Q-Q Plot


```
qqnorm(Treatment2)
```

Normal Q-Q Plot
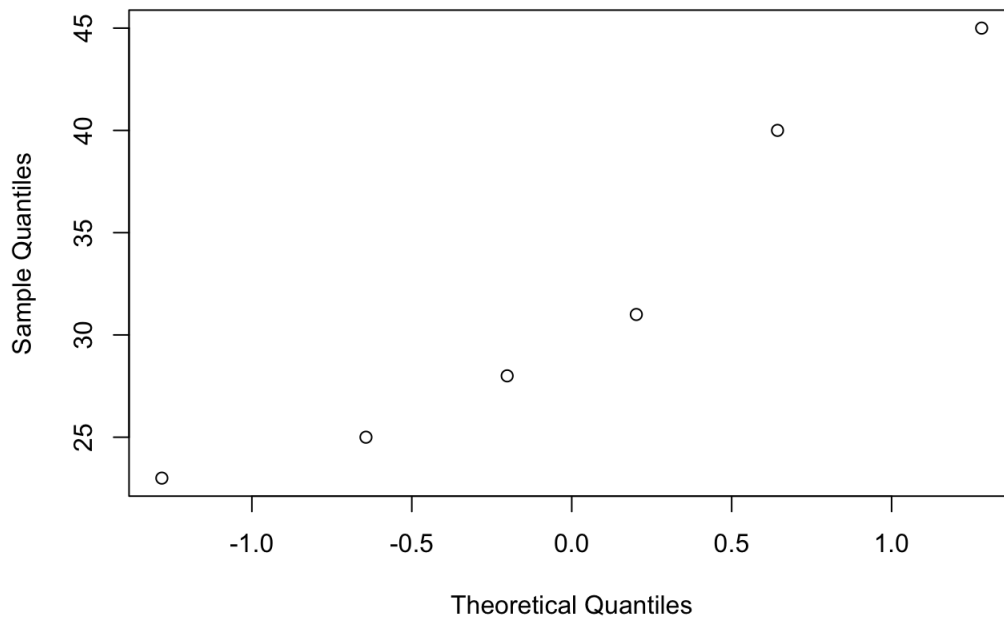


```
qqnorm(Treatment3)
```

Normal Q-Q Plot

the cases are independent, the variances are equal, and the distributions are relativley normal, thus the ANOVA test will work here.

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

Since the p-value < 0.001 which is less the alpha = 0.01, hence we reject the null hypothesis and conclude that their is significant difference and The mean sleep times for the three treatments are not equal:
treatment 1 =/= treatment2 =/= treatment3

    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?

df_{Err} = 12
MSE = 83.1625
α = 0.05
t multiplier = $t_{(0.025, 12)}$ = 2.179
μ_1 - μ_2: 61.6 - 50.25 +- 2.179 * sqrt(83.1625 * (1/5 + 1/4)) = (-1.979911, 24.67991) * 0 included
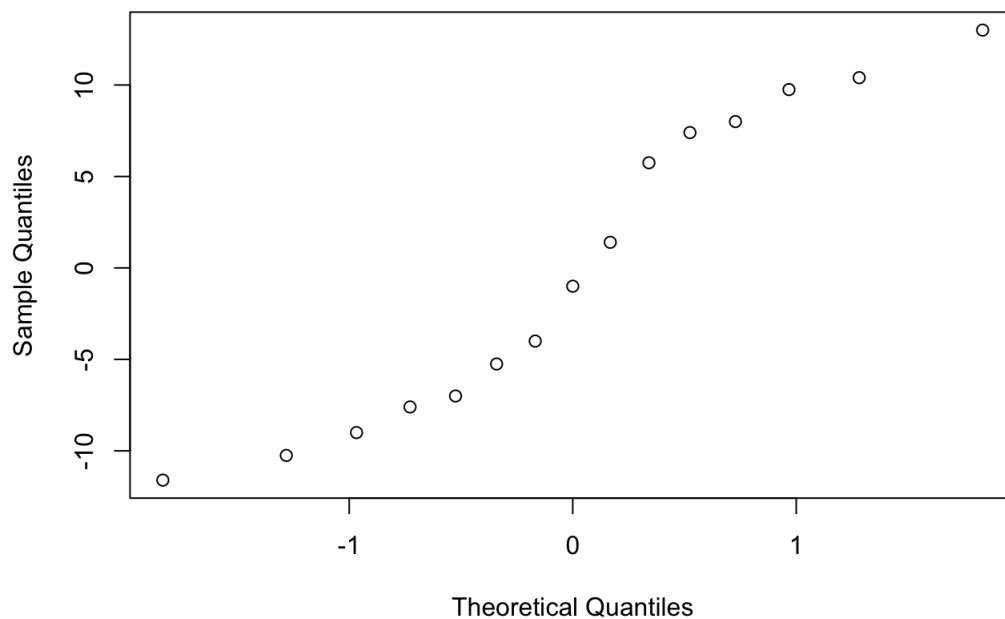μ_1 - μ_3: 61.6 - 32 +- 2.179 * sqrt(83.1625 * (1/5 + 1/6)) = (17.56748, 41.63252) * 0 not included
μ_2 - μ_3: 50.25 - 32 +- 2.179 * sqrt(83.1625 * (1/4 + 1/6)) = (5.423287, 31.07671) * 0 not included
Doing the ANOVA in R

```
Treatmentanova<-aov(Data~Treatment, data=Treatmentdata)
qqnorm(Treatmentanova$residuals)
```

Normal Q-Q Plot


```
Treatmentlm<-lm(Data~Treatment, data=Treatmentdata)
anova(Treatmentlm)
```

```
## Analysis of Variance Table
##
## Response: Data
##            Df  Sum Sq Mean Sq F value     Pr(>F)
## Treatment   2 2456.98 1228.49  14.772 0.0005808 ***
## Residuals  12  997.95   83.16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Fisher LSD CIs:
ft<-qt(p=0.05/2, df=12, lower.tail=FALSE)
#Sample sizes of 5 and 4
invisible((F5_4ME=ft*sqrt(83.1625*(1/5+1/4))))
invisible((diff1_2<-61.6-50.25))
diff1_2+F5_4ME; diff1_2-F5_4ME
```

```
## [1] 24.67877
```

```
## [1] -1.978766
```

```
#Sample sizes of 5 and 6
invisible((F5_6ME=ft*sqrt(83.1625*(1/5+1/6))))
invisible((diff1_3<-61.6-32))
diff1_3+F5_6ME; diff1_3-F5_6ME
```

```
## [1] 41.63149
```

```
## [1] 17.56851
```

```
#Sample sizes of 4 and 6
invisible((F4_6ME=ft*sqrt(83.1625*(1/4+1/6))))
invisible((diff2_3<-50.25-32))
diff2_3+F4_6ME; diff2_3-F4_6ME
```

```
## [1] 31.07561
```

```
## [1] 5.424389
```

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

```
#Bonferroni Multiple Comparison Adjustment
qt(p=0.0005808, df=12, lower.tail=FALSE)
```

```
## [1] 4.233094
```

```
#Tukey Multiple Comparison Adjustment
TukeyHSD(Treatmentanova)
```

```
##    Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = Data ~ Treatment, data = Treatmentdata)
##
## $Treatment
##                        diff       lwr        upr     p adj
## Treatment2-Treatment1 -11.35 -27.67051   4.970507 0.1939650
## Treatment3-Treatment1 -29.60 -44.33205 -14.867952 0.0004632
## Treatment3-Treatment2 -18.25 -33.95442  -2.545584 0.0231619
```

```
qtukey(p=(1-0.05), nmeans=3, df=(15-3)) / sqrt(2)
```

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

```
wilcox.test(Treatment1,Treatment2) #Unadjusted = 0.1905; Adjusted = 3 * 0.1905 = 0.5715 diff
```

```
##
##   Wilcoxon rank sum test
##
## data:  Treatment1 and Treatment2
## W = 16, p-value = 0.1905
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(Treatment1,Treatment3) #Unadjusted = 0.004329; Adjusted = 3 * 0.004329 = 0.012987 diff
```

```
##
##   Wilcoxon rank sum test
##
## data:  Treatment1 and Treatment3
## W = 30, p-value = 0.004329
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(Treatment2,Treatment3) #Unadjusted = 0.04157; Adjusted = 3 * 0.04157 = 0.12471 diff
```

```
## Warning in wilcox.test.default(Treatment2, Treatment3): cannot compute
## exact p-value with ties
```

```
##
##   Wilcoxon rank sum test with continuity correction
##
## data:  Treatment2 and Treatment3
## W = 22, p-value = 0.04157
## alternative hypothesis: true location shift is not equal to 0
```

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?

```
#Kruskal-Wallis Test
kruskal.test(Data~Treatment, data=Treatmentdata)
```

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

Slightly more conservative overall test. KW has an overall p value of 0.00725 which is higher than that found by the F test. The pairwise comparisons were similar in the grouping - which groups were significantly different or not.

```
    Trt    |   Mean  | Unadjusted Group | Boneferonni Group |
Treatment 3:   32    |      A           |       A           |
Treatment 2:   50.25 |        B         |         B         |
Treatment 1:   61.6  |        B         |         B         |
```

Problem 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.

```
        | n_i      | x̄_i           | s²_i          | n_i * (X_i - X_grand)^2 | (n_i - 1) * s²_i |
Diet A | n_1 = 9  | x̄_1 = 39.1   | s²_1 = 24.6   |  0.387                  | 196.8            |
Diet B | n_2 = 8  | x̄_2 = 29.9   | s²_2 = 16.4   |  646.934                | 114.8            |
Diet C | n_3 = 10 | x̄_3 = 45.9   | s²_2 = 10.3   |  491.038                | 92.7             |
        Grand Mean = 38.8926                      | 1138.359                | 404.30
```

(39.1 * 9) + (29.9 * 8) + (45.9 * 10) / 27 = 38.8926

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

```
Source          | SS       |         df          |            MS               |               F               |
 P  |
Treat (between)| 1138.359 | (t-1) = (3-1) = 2    | 1138.359 / 2 = 569.1795 | 569.1795 / 16.8458 = 33.78756 |0
.001|
Error (within) | 404.3    | (N-t) = (27-3) = 24 | 404.3 / 24 = 16.84583   |                               |
    |
Total          | 1542.66  | (N-1) = (27-1) = 26 |                         |                               |
    |
```

(y1 * n1) + (y2 * n2) + (y3 * n3) / (n1 + n2 + n3)
((39.1 * 9) + (29.9 * 8) + (45.9 * 10)) / (9 + 8 + 10) = 38.8926
(9 * (39.1 - 38.8926)^2) + (8 * (29.9 - 38.8926)^2) + (10 * (45.9 - 38.8926)^2) = 1138.359
SSTrt + SSE = 1138.359 + 404.3 = 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.

$P(F_{(33.78756, 2, 24)} > 9.34) = 0.001$
cause test statistic F = 33.78756 falls in rejection region; therefore we reject null hypothesis. We have suffiicient evidence to conclude that population means of milk yield are not same among three diet types.

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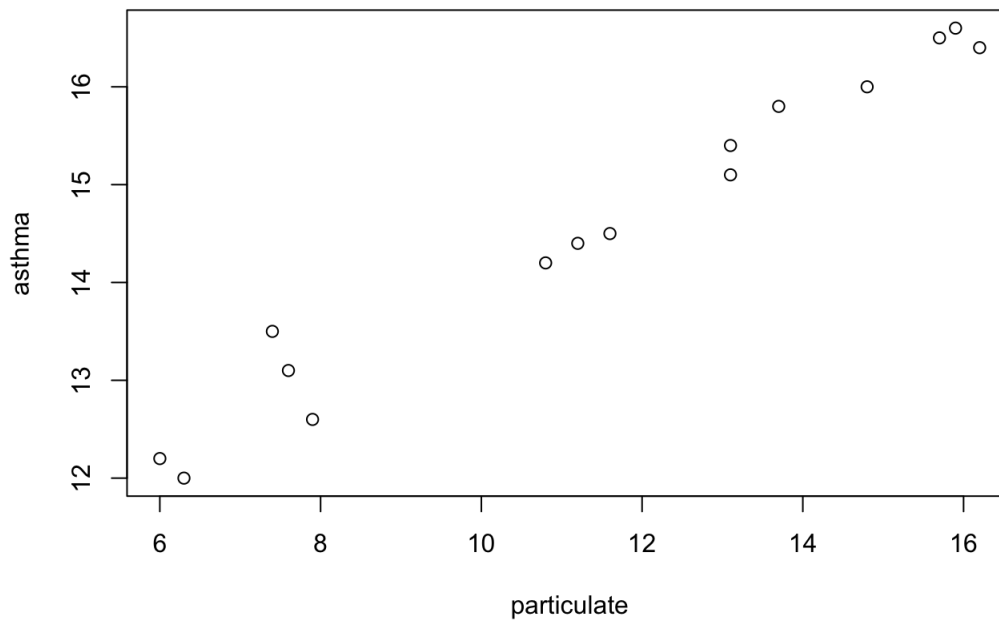
   a. Plot the data as you see fit and summarize the pattern's shape, direction, and strength in the context of the problem.

```
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)
cor(particulate, asthma)
```

```
## [1] 0.9854127
```

```
plot(particulate, asthma)
```


b. Calculate the correlation coefficient by hand and confirm your findings in R.

```
X Values
∑ = 171.3
Mean = 11.42
∑(X - Mx)² = SSx = 182.704

Y Values
∑ = 218.3
Mean = 14.553
∑(Y - My)² = SSy = 35.297

X and Y Combined
N = 15
∑(X - Mx)(Y - My) = 79.134

∑X * Y = 2572.12

∑X² = 2138.95

∑Y² = 3212.29
```

R Calculation r = $\sum((X - My)(Y - Mx)) / \sqrt{((SSx)(SSy))}$

r = 79.134 / $\sqrt{((182.704)(35.297))}$ = 0.9854

```
mod<-lm(asthma~particulate)
summary(mod)
```

```
##
## 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
```

```
residuals = mod$residuals
fitted = mod$fitted.values
plot(fitted, residuals)
```
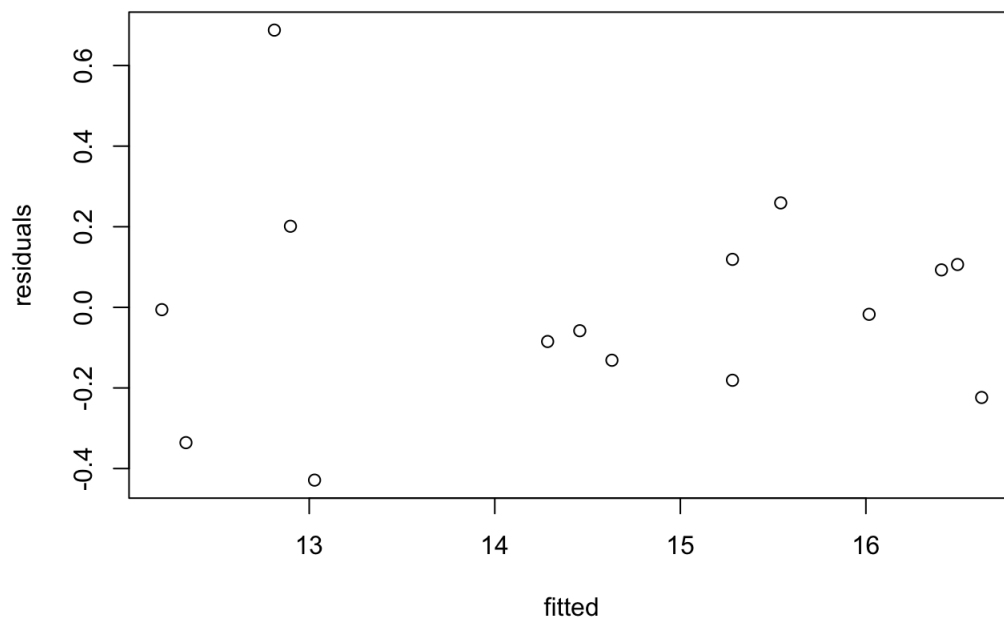


```
qqnorm(residuals)
```

Normal Q-Q Plot
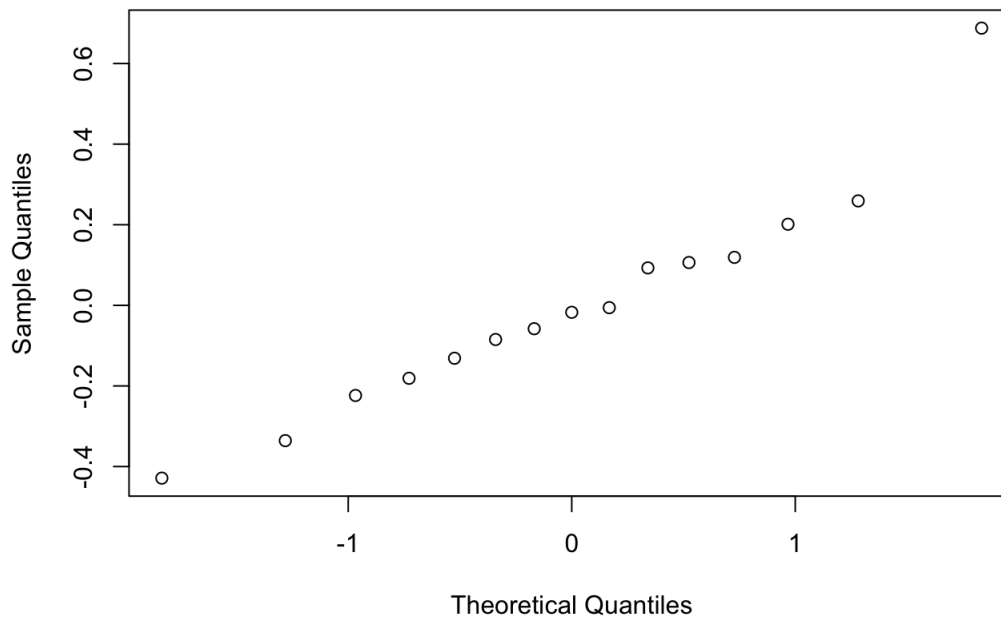

```
sum(residuals^2) / 13 # (n - 2) = (15 - 2) = 13
```

```
## [1] 0.07863667
```

```
#MSErr = 35.297 / 13 = 2.715154
# 14 * 182.704^2 = 467330.5
#SE = sqrt(2.715154) / sqrt(467330.5) = 0.002410378
# 79.134 / 182.704 = 0.4331268
# Tobs = (0.4331268 - 1) / 0.002410378 = -235.1802
pt(-235.1802, df=13, lower.tail=FALSE)*2
```

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

Multiple R-squared: 0.971, Adjusted R-squared: 0.9688

   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.

$a = n * \sum X * Y - \sum X * \sum Y / n * \sum X^2 - (\sum X)^2$
$a = 15 * 2572.12 - 171.3 * 218.3 / 15 * 2138.95 - (171.3)^2 = 0.433$
$b = \sum Y * \sum X^2 - \sum X * \sum X * Y / n * \sum X^2 - (\sum X)^2$
$b = 218.3 * 2138.95 - 171.3 * 2572.12 / 15 * 2138.95 * (171.3)^2 = 9.607$ y = ax + b
$y = (0.433 * x) + 9.607$

```
mod<-lm(asthma~particulate)
summary(mod)
```

```
##
## 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
```

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

```
require(ggplot2)
ggplot(data=data.frame(particulate, asthma), aes(x=particulate, y=asthma))+
  geom_point()+
  geom_smooth(method="lm", se=TRUE )
```
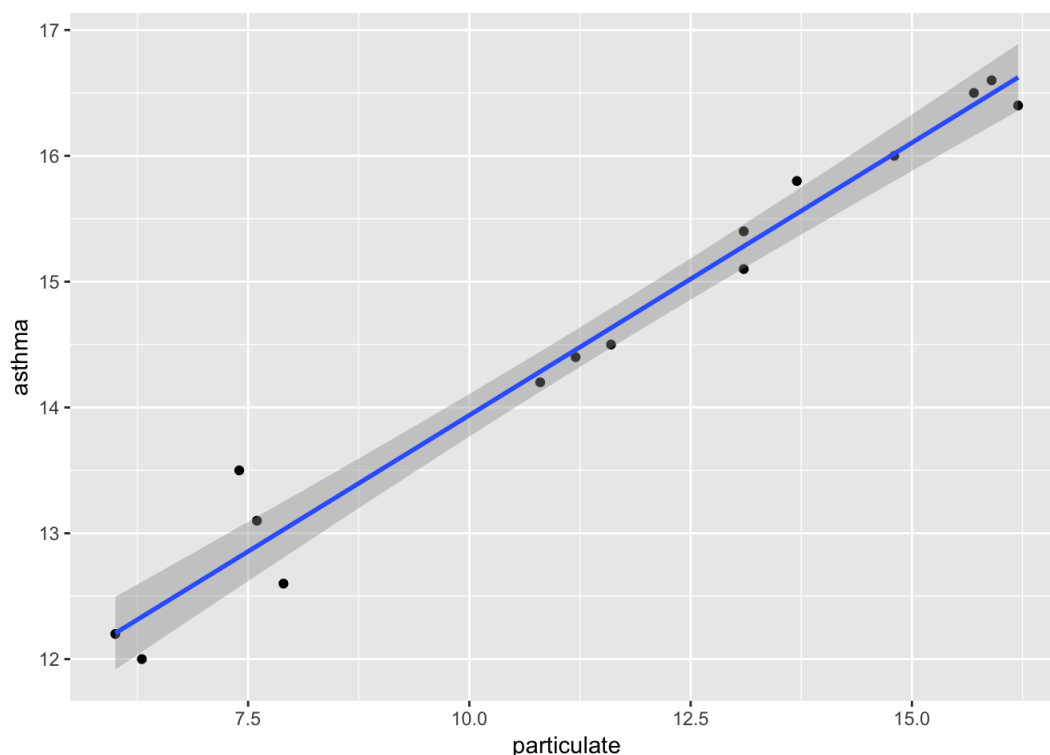

e. Conduct a formal test to determine whether or not the slope is significantly different from 0 (In R or by hand). If the slope is significant, can we, using only this statistical information, conclude that increased particulate concentration causes an increase in childhood asthma?

```
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 ***
```

H_0: The slope is zero
H_1: The slope is significantly different from zero
The p value of the particulate variable is less than 0.05:
particulate 0.43313 0.02075 20.88 2.21e-11 ***
The slope of the particulate is not zero and we can reject the null hypothesis in favor of alternate hypothesis.

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

The r square value is Multiple R-squared: 0.971, which means that model is able to explain 97.1% variation in the data The value is high

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.

```
SSx = ∑x² - nx̄ = (2138.95 - 15 * 11.42²) = 182.704

SSy = ∑y² - nȳ = (3212.29 - 15 * 14.553²) = 35.44286

SSxy = ∑(X - Mx)(Y - My) = 79.134
```

$y = 9.607 + (0.433 * 13) = 15.2376$
$\sigma^2 = 1 / n - 2 * (Syy - Sxy^2 / Sxx)$
$\sigma^2 = (1 / 15) * (35.44286 - (79.134^2 / 182.704)) = 0.07785355$

$SE(\hat{y}) = sqrt(((1 + (1 / n) + (x_0 - \bar{x})^2 / Sxx) * \sigma^2))$
$SE(\hat{y}) = sqrt(((1 + (1 / 15) + ((13 - 11.42)^2 / 182.704)) * 0.07785355)) = 0.290013$
$t_{((\alpha / 2), (n - 2))} = 2.160$
$\hat{y} +- t * SE(\hat{y})$
15.238 +- 2.16 * 0.290013
(14.608, 15.867)

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.

$y = 9.607 + (0.433 * 10) = 13.937$
$\sigma^2 = 1 / n - 2 * (Syy - Sxy^2 / Sxx)$
$\sigma^2 = (1 / 15) * (35.44286 - (79.134^2 / 182.704)) = 0.07785355$

$SE(\hat{y}) = sqrt((((1 / n) + (x_0 - \bar{x})^2 / Sxx) * \sigma^2))$
$SE(\hat{y}) = sqrt((((1 / 15) + ((10 - 11.42)^2 / 182.704)) * 0.07785355)) = 0.07777829$
$t_{((\alpha / 2), (n - 2))} = 2.160$
$\hat{y} +- t * SE(\hat{y})$
13.937 +- 2.16 * 0.07777829
(13.769, 14.105)

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.

$y = 9.607 + (0.433 * 3) = 10.906$
$\sigma^2 = 1 / n - 2 * (Syy - Sxy^2 / Sxx)$
$\sigma^2 = (1 / 15) * (35.44286 - (79.134^2 / 182.704)) = 0.07785355$

$SE(\hat{y}) = sqrt((((1 / n) + (x_0 - \bar{x})^2 / Sxx) * \sigma^2))$
$SE(\hat{y}) = sqrt((((1 / 15) + ((3 - 11.42)^2 / 182.704)) * 0.07785355)) = 0.1881502$
$t_{((\alpha / 2), (n - 2))} = 2.160$
$\hat{y} +- t * SE(\hat{y})$
10.906 +- 2.16 * 0.1881502
(10.4996, 11.3124)

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 95% confidence interval at x = 12 will have a narrower width because it is closer to the mean and we are more confident.

Problem 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.

```
        x̄                      |    50
        ȳ                      |    47.91
        sdx                    |    33.17
        sdy                    |    31.25
∑¹¹ᵢ₌₁(xᵢ - x̄)²              |    11000
∑¹¹ᵢ₌₁(yᵢ - ȳ)²              |    9768.91
∑¹¹ᵢ₌₁(yᵢ - ȳ)(xᵢ - x̄)      |    10360
∑¹¹ᵢ₌₁(yᵢ - ŷᵢ)²             |    11.67
```

To check the calibration, the linear model $y = \beta_0 + \beta_1 x + \varepsilon$ 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.

$\beta_1 = \sum^{11}_{i=1}(y_i - \bar{y})(x_i - \bar{x}) / \sum^{11}_{i=1}(x_i - \bar{x})^2$
$\beta_0 = \bar{y} - (\bar{x} * \beta_1)$

$\beta_1 = 10360 / 11000 = 0.9418182$
$\beta_0 = 47.91 - (50 * 0.9418182) = 0.81909$

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

$H_0 : \beta_1 = 1$
$H_1 : \beta_1 =/= 1$

The quantity $(\hat{\beta}_1 - \beta_1) / S_{\hat{\beta}_1}$ has the t distribution with n - 2 = 9 degrees of freedom.
$r^2 = (9768.91 - 11.67) / 9768.91 = 0.9988054$
$s = sqrt((1 - r^2) * 9768.91 / 11 - 2) = 1.13871$
$S_{\hat{\beta}_1} = s / sqrt(11000) = 0.01085717$
$(\hat{\beta}_1 - \beta_1) / S_{\hat{\beta}_1} = (0.9418182 - 1) / 0.01085717 = -5.358837$
Checking the t table with 9 degrees of freedom, we find that the pvalue is less than 0.001. Therefore we reject the null hypothesis $H_0: \beta_1 = 1$