Statistics 324 Homework 10

Svadrut Kukunooru

*Submit your homework to Canvas by the due date and time. Email your instructor if you have extenuating circumstances and need to request an extension.

*If an exercise asks you to use R, include a copy of the code and output. Please edit your code and output to be only the relevant portions.

*If a problem does not specify how to compute the answer, you many use any appropriate method. I may ask you to use R or use manual calculations on your exams, so practice accordingly.

*You must include an explanation and/or intermediate calculations for an exercise to be complete.

*Be sure to submit the HWK 10 Auto grade Quiz which will give you ~20 of your 40 accuracy points.

*50 points total: 40 points accuracy, and 10 points completion

Exercise 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 given below.

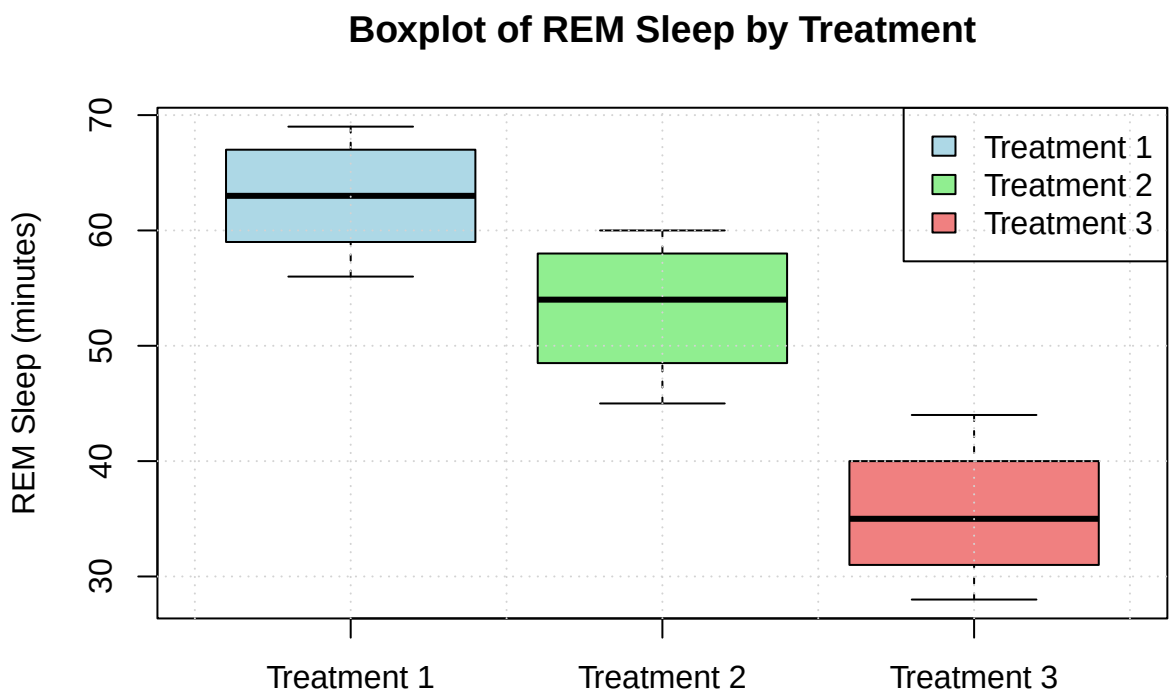
The researchers plan to perform a test to help decide between a model that says a mean amount of REM sleep for all three treatment is equal $H_o : \mu_1 = \mu_2 = \mu_3$ and a model that allows for at least one group mean to be different H_A : at least one μ_i is different.

Treatment 1: 63, 56, 69, 59, 67
Treatment 2: 45, 60, 52, 56
Treatment 3: 31, 40, 44, 33, 37, 28

a. Make a preliminary graph of the data that enables you to compare the centers and spreads of the three samples.

```r
# Data
treatment1 <- c(63, 56, 69, 59, 67)
treatment2 <- c(45, 60, 52, 56)
treatment3 <- c(31, 40, 44, 33, 37, 28)

# Combine data into a list
data <- list(Treatment1 = treatment1, Treatment2 = treatment2, Treatment3 = treatment3)

# Create boxplot
boxplot(data, col = c("lightblue", "lightgreen", "lightcoral"),
        main = "Boxplot of REM Sleep by Treatment",
        ylab = "REM Sleep (minutes)",
        names = c("Treatment 1", "Treatment 2", "Treatment 3"))

# Add grid for better readability
```

```r
grid()

# Add legend
legend("topright", legend = c("Treatment 1", "Treatment 2", "Treatment 3"),
       fill = c("lightblue", "lightgreen", "lightcoral"))
```



» a.1 Does this graph suggest the three samples come from populations with the same mean value (H_o : is true?) or that an H_A model is better? What about the graph makes you say that?

This graph suggests that the H_A model is better, due to the centers being different.

a.2 What does this preliminary graph tell you about an equal variance assumption in the three populations?

Based on the size of the boxes in the boxplot, there is an equal variance assumption in the three populations.

b. Compute the following summary statistics that will be useful in an ANOVA analysis. Keep values to at least 3 decimal places.

```r
# Calculate mean, sd, and n for each treatment
mean_treatment1 <- mean(treatment1)
sd_treatment1 <- sd(treatment1)
n_treatment1 <- length(treatment1)

mean_treatment2 <- mean(treatment2)
```

```r
sd_treatment2 <- sd(treatment2)
n_treatment2 <- length(treatment2)

mean_treatment3 <- mean(treatment3)
sd_treatment3 <- sd(treatment3)
n_treatment3 <- length(treatment3)

# Calculate overall mean, sd, and N
overall_mean <- mean(c(treatment1, treatment2, treatment3))
overall_sd <- sd(c(treatment1, treatment2, treatment3))
overall_n <- length(c(treatment1, treatment2, treatment3))
```

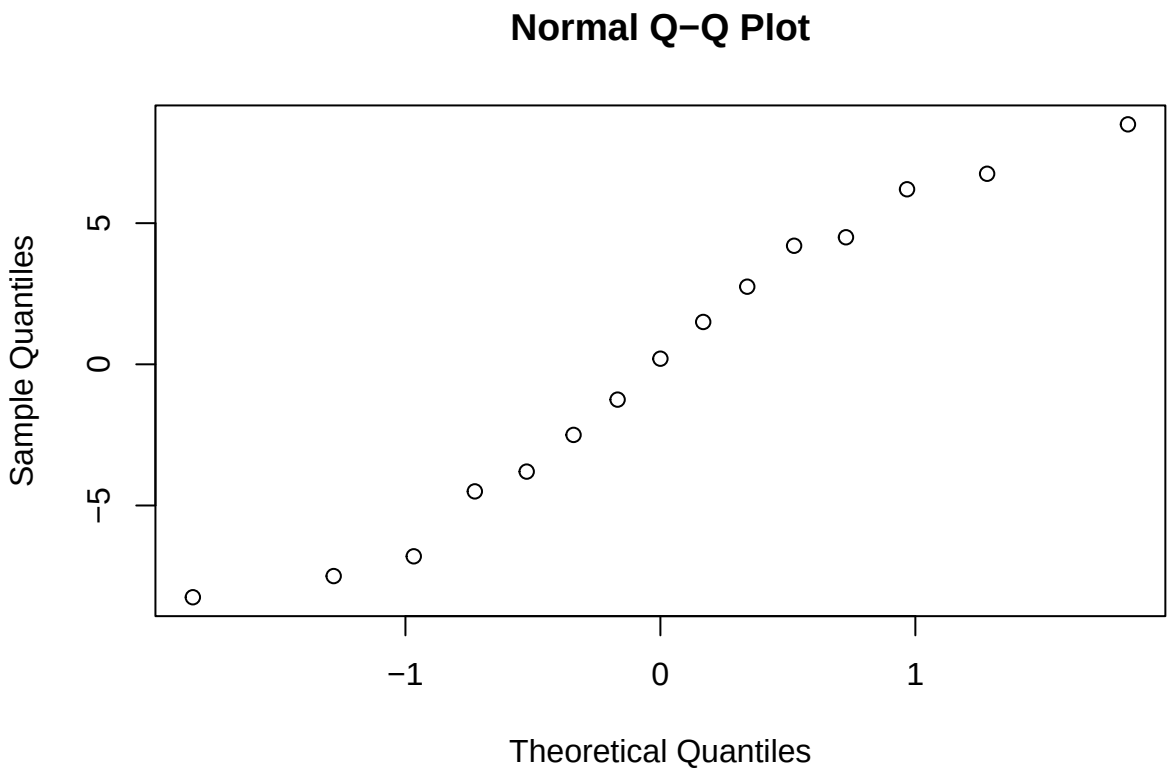
Sample	Treatment 1	Treatment 2	Treatment 3	Overall
mean	62.80	53.25	35.50	49.33
sd	5.40	6.40	5.96	13.45
n	5	4	6	15

c. Create an ANOVA table for the data using the relevant function in R.

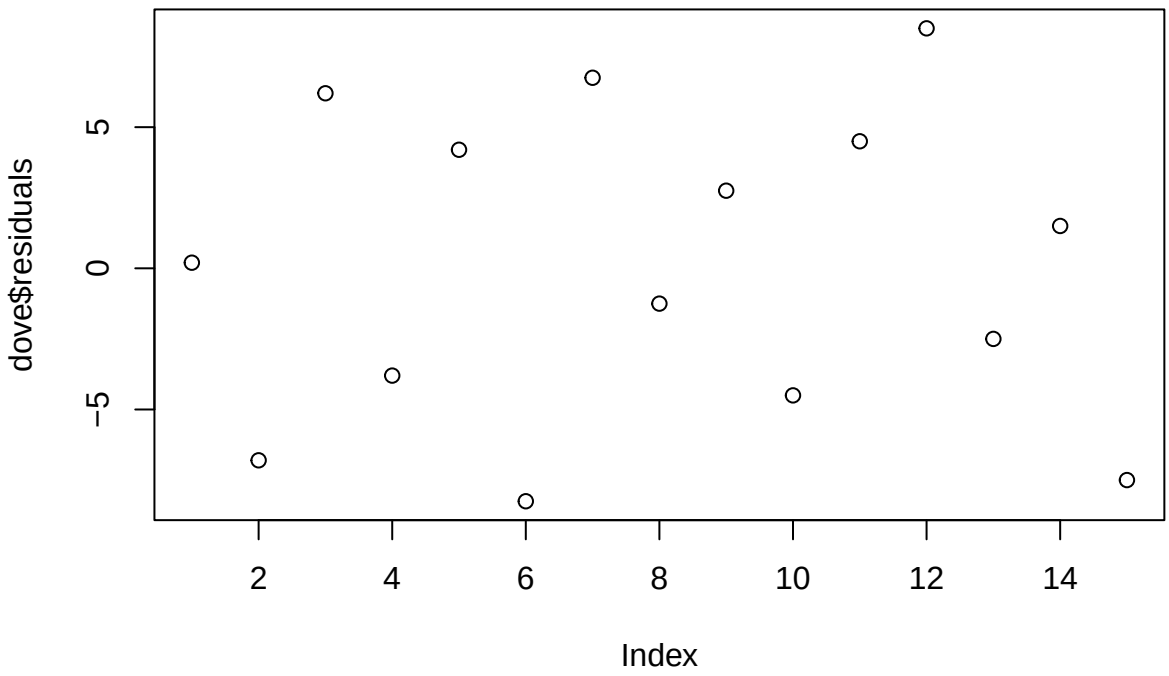
Source	DF	SS	MS	F	p-value
Treatment	2	2116.283	1058.141	30.44075	1.992433e-05
Error	12	417.128	34.76067		
Total	14	2532.635	-		

d. Create a residual plot and qqplot of your model's residuals. Use the graphs and the summary values above to explain why the assumptions for an ANOVA analysis are well met for this data.

```r
df = data.frame(REM = c(treatment1, treatment2, treatment3), TRT = c(rep("treatment1", length(treatment
dove = aov(REM ~ TRT, data = df)

qqnorm(dove$residuals)
```




```
plot(dove$residuals)
```



Based on the description of the study, we can assume that the sets of measurements are independent random samples from their respective populations. Additionally, based on the box-plots, we can assume that the variances of the t populations are equal.

e. Based on the ANOVA table, make a conclusion in the context of the problem. (Write out your hypotheses, identify your test statistic, and p value here.)

Hypotheses:

Null Hypothesis (H_0): The mean amount of REM sleep is equal for all three treatments. Alternative Hypothesis (H_a): At least one group mean is different.

Observed Test Statistic and p-value:

Test Statistic (F): 30.44075 p-value: 1.992433e-05 (very close to 0)

Conclusion in context: With a p-value much smaller than the significance level (commonly set at 0.05), we reject the null hypothesis. There is sufficient evidence to conclude that at least one group mean of REM sleep is different among the three treatments, since the F-statistic of 30.44075 indicates a significant difference in means.

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. Use a built-in R function to check your answers.

```
# Assuming 'anova_model' is the model obtained from the ANOVA analysis

# Obtain the relevant multiplier for Tukey method
```

```r
tukey_result <- TukeyHSD(dove)

# Display the Tukey multiple comparisons
tukey_result
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = REM ~ TRT, data = df)
##
## $TRT
##                           diff       lwr        upr     p adj
## treatment2-treatment1  -9.55 -20.10051   1.000507 0.0776081
## treatment3-treatment1 -27.30 -36.82364 -17.776364 0.0000165
## treatment3-treatment2 -17.75 -27.90223  -7.597770 0.0014645
```

g. Summarize your results regarding which groups are found significantly different using letter codes. What do you conclude?

Treatment	Sample Mean	Letter (according to Tukey's)
3	35.5	A
2	53.25	A
1	62.8	B

h. Now analyze the same data using an overall Kruskal-Wallis - report the test statistic and p value. Followed up with Wilcoxon Rank Sum tests with a Bonferroni adjustment. How does your conclusion change, if at all?

```r
kruskal.test(REM ~ TRT, data = df)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  REM by TRT
## Kruskal-Wallis chi-squared = 11.411, df = 2, p-value = 0.003328
```

```r
pairwise = pairwise.wilcox.test(df$REM, df$TRT, p.adjust.method = "bonferroni")
```

```
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute
## exact p-value with ties
```

Treatment	Sample Mean	Letter (according to Wilcoxon RS w/ Bonferroni's)
3	35.5	A
2	53.25	A
1	62.8	B

i. What test and conclusions would you recommend the scientist use based on your findings and your evaluation of the assumptions of the tests?

Given that the data might not meet the assumptions of normality required for ANOVA, the Kruskal-Wallis test is a suitable non-parametric alternative for comparing the medians of three or more groups. If the Kruskal-Wallis test indicates a significant difference between groups, use pairwise Wilcoxon Rank Sum tests to identify which specific groups differ from each other. The Bonferroni adjustment helps control for multiple testing.

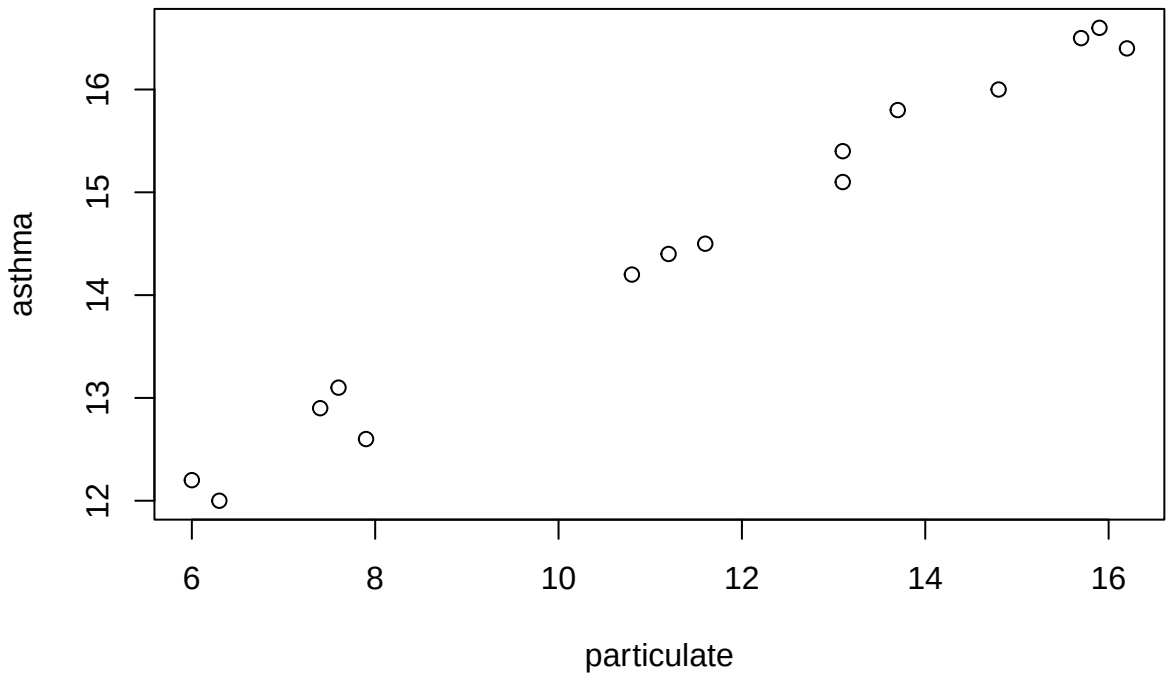
Exercise 2 Suppose we are interested in exploring the relationship between city air particulate and rates of childhood asthma. We sample 15 cities for particulate measured in parts-per-million (ppm) of large particulate matter and for the rate of childhood asthma measured in percents. The data are as follows (and are also given in asthma.csv):

variable															
ppm(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	12.9	16.4	15.4	14.4

variable:	size	mean	variance
particulate	15	11.42	13.05029
asthma	15	14.51333	2.635524

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, 12.9, 16.4, 15.4, 14.4)
plot(asthma ~ particulate)
```



b. Calculate the correlation coefficient and explain how the value corresponds to what you observed in the graph in part (a)

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

```
## [1] 0.9931873
```

This is a very high correlation coefficient and it corresponds with the graph looking linear as I observed in part A.

c. Build a linear regression model with least squares estimators for slope and y intercept for the data

(c.1) First by hand using the correlation computed in (b) and summary statistics given above.

```r
mean_particulate = mean(particulate)
mean_asthma = mean(asthma)

correlation_coefficient = cor(particulate, asthma)

slope <- correlation_coefficient * (sd(asthma) / sd(particulate))
intercept <- mean_asthma - slope * mean_particulate

# Display results
cat("Slope (b):", slope, "\n")
```

```
## Slope (b): 0.4463285
```

```r
cat("Y-intercept (a):", intercept, "\n")
```

```
## Y-intercept (a): 9.416262
```

(c.2) then check your computations using lm in R.

```r
lm(asthma ~ particulate)
```

```
##
## Call:
## lm(formula = asthma ~ particulate)
##
## Coefficients:
## (Intercept)  particulate
##      9.4163       0.4463
```

(c.3) Interpret the estimated intercept and slope in the context of the question.

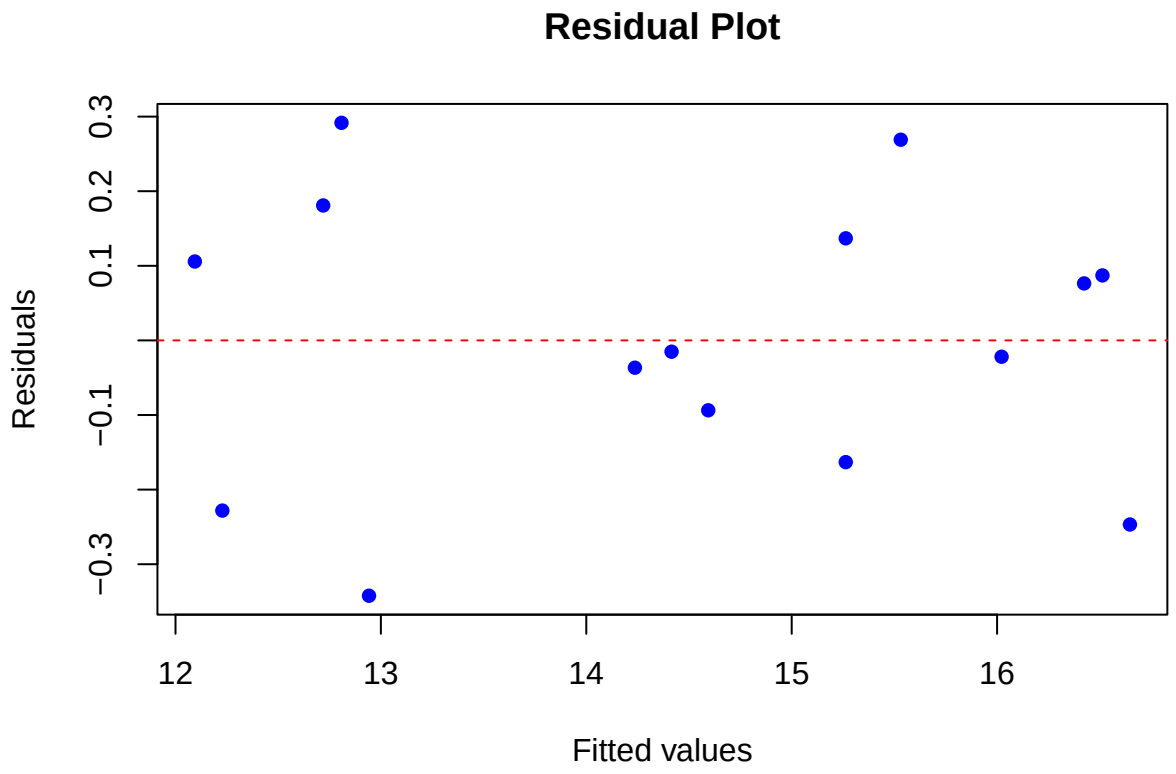
Intercept 9.4163

Slope 0.4463

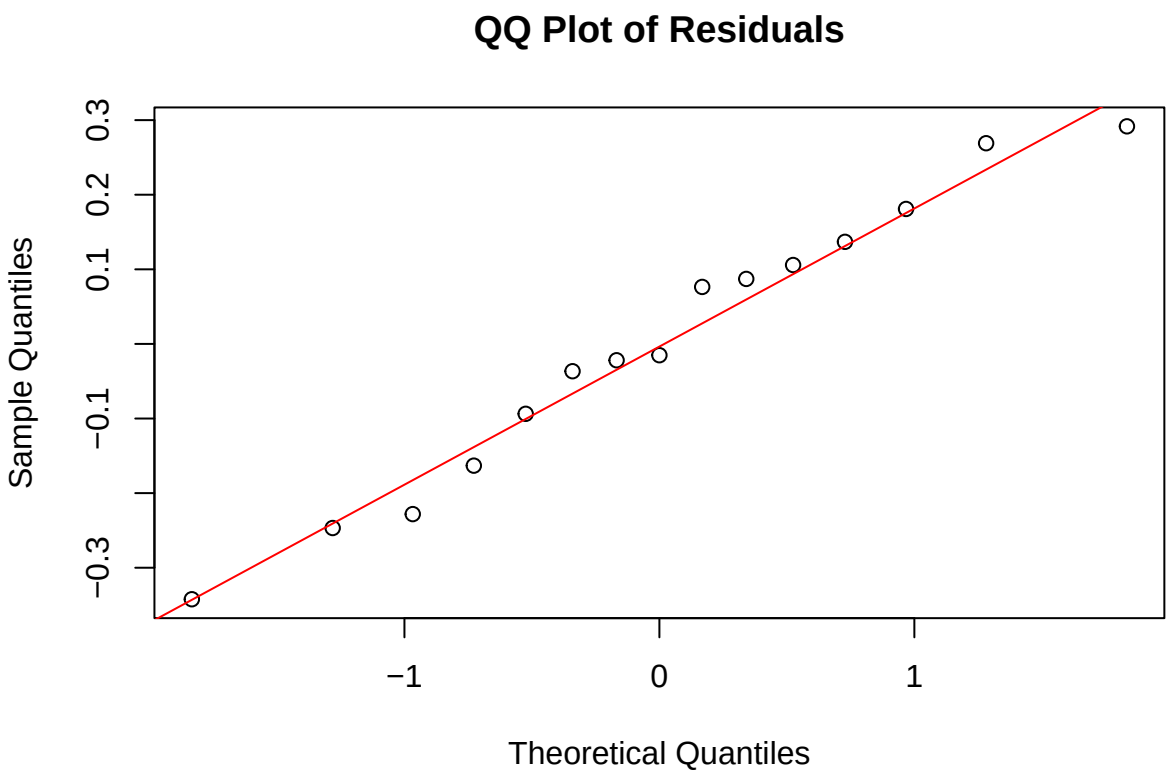
For each increase in ppm, there is a 0.446% increase in asthma. When there is no ppm, there is still 9.4% asthma.

d. Construct a residual plot of fitted y values on the x axis and residuals on the y. Also, create a qqnorm plot of the residuals. Assess whether the correct model, constant variance, and normality of errors assumptions are reasonably met.

```r
# Extract residuals from the model
linear_model <- lm(asthma ~ particulate)
fitted_values <- fitted(linear_model)
residuals <- residuals(linear_model)

# Construct a residual plot
plot(fitted(lm(asthma ~ particulate)), residuals, xlab = "Fitted values", ylab = "Residuals",
     main = "Residual Plot", pch = 16, col = "blue")
abline(h = 0, col = "red", lty = 2)  # Add a horizontal line at y = 0
```




```r
# Create a QQ plot of the residuals
qqnorm(residuals, main = "QQ Plot of Residuals")
qqline(residuals, col = "red")
```



e. Identify what (particulate, asthma) point results in the residual with the largest magnitude. Is that point above or below the fitted regression line? Show how the residual is calculated. (Makes sure that you can also identify that point on the residual plot.)

```
# Find the index of the observation with the largest residual magnitude
index_max_residual <- which.max(abs(residuals))

# Extract the corresponding observation
observation_max_residual <- c(particulate[index_max_residual], asthma[index_max_residual])

# Display the observation with the largest residual magnitude
print(observation_max_residual)
```

```
## [1]  7.9 12.6
```

```
# Calculate the residual for the observation with the largest magnitude
observed_value_max_residual <- asthma[index_max_residual]
fitted_value_max_residual <- fitted_values[index_max_residual]
residual_max <- observed_value_max_residual - fitted_value_max_residual

# Display the calculated residual
print(residual_max)
```

```
##          4
## -0.3422571
```