Final - Akshaj Kammari

05/02/2024

```
library(readr)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(boot)
library(ISLR)
data(Carseats)
set.seed(29101)
```

1.

```
model1 <- lm(Sales ~ Advertising, data=Carseats)
summary(model1)
```

```
##
## Call:
## lm(formula = Sales ~ Advertising, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3770 -1.9634 -0.1037  1.7222  8.3208
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.7370     0.1925  35.007  < 2e-16 ***
## Advertising   0.1144     0.0205   5.583 4.38e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

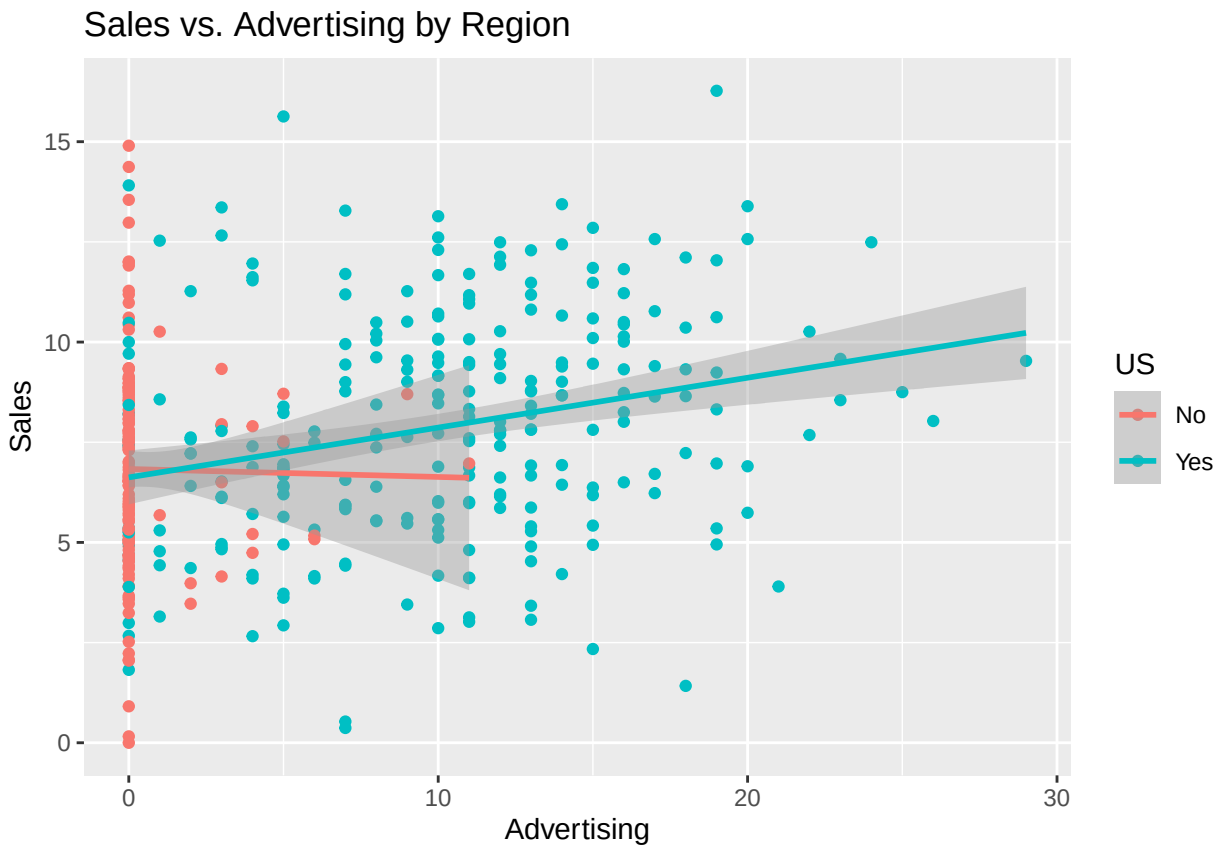
```
## Residual standard error: 2.723 on 398 degrees of freedom
## Multiple R-squared:  0.07263,    Adjusted R-squared:  0.0703
## F-statistic: 31.17 on 1 and 398 DF,  p-value: 4.378e-08
```

Null hypothesis (H0): There is no linear relationship between advertising budget and sales (i.e., coefficient of Advertising is 0). Alternative hypothesis (H1): There is a positive linear relationship (i.e., coefficient of Advertising is greater than 0). The p-value is low (commonly <0.05), reject H0, supporting the alternative hypothesis of a positive relationship.

2.

```
ggplot(Carseats, aes(x=Advertising, y=Sales, color=US)) +
  geom_point() +
  geom_smooth(method="lm") +
  labs(title="Sales vs. Advertising by Region")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
model_us <- lm(Sales ~ Advertising, data=Carseats, subset=(US=="Yes"))
model_non_us <- lm(Sales ~ Advertising, data=Carseats, subset=(US=="No"))
summary(model_us)
```

```
##
## Call:
## lm(formula = Sales ~ Advertising, data = Carseats, subset = (US ==
##     "Yes"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -7.4407 -2.0139 -0.1081  1.8053  8.3858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.62245    0.34140  19.398  < 2e-16 ***
## Advertising  0.12435    0.02938   4.232 3.22e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.787 on 256 degrees of freedom
## Multiple R-squared:  0.0654, Adjusted R-squared:  0.06175
## F-statistic: 17.91 on 1 and 256 DF,  p-value: 3.224e-05
```

```
summary(model_non_us)
```

```
##
## Call:
## lm(formula = Sales ~ Advertising, data = Carseats, subset = (US ==
##     "No"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8331 -1.7331 -0.1731  1.6894  8.0669
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.83314    0.22941  29.785   <2e-16 ***
## Advertising -0.01994    0.13371  -0.149    0.882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.612 on 140 degrees of freedom
## Multiple R-squared:  0.0001589,  Adjusted R-squared:  -0.006983
## F-statistic: 0.02224 on 1 and 140 DF,  p-value: 0.8817
```

Similar hypothesis as part 1, but separately for US and non-US stores. The advertising effect differs by region.

3. This discrepancy might be described as the "interaction effect" meaning the relationship between advertising and sales depends on the region.

4.

```
model_us_indicator <- lm(Sales ~ US, data=Carseats)
summary(model_us_indicator)
```

```
##
## Call:
## lm(formula = Sales ~ US, data = Carseats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.497 -1.929 -0.105  1.836  8.403
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.8230     0.2335   29.21  < 2e-16 ***
## USYes         1.0439     0.2908    3.59 0.000372 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.783 on 398 degrees of freedom
## Multiple R-squared:  0.03136,    Adjusted R-squared:  0.02893
## F-statistic: 12.89 on 1 and 398 DF,  p-value: 0.0003723
```

Null hypothesis (H0): No difference in sales between US and non-US stores (coefficient of US is 0). Alternative hypothesis (H1): A difference exists between US and non-US stores. (coefficient of US is not 0)

5.

```
model_multiple <- lm(Sales ~ Advertising + US, data=Carseats)
summary(model_multiple)
```

```
##
## Call:
## lm(formula = Sales ~ Advertising + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3938 -1.9653 -0.1202  1.7542  8.3564
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.76295    0.22923  29.503  < 2e-16 ***
## Advertising  0.11848    0.02815   4.209 3.18e-05 ***
## USYes       -0.08177    0.39075  -0.209    0.834
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.726 on 397 degrees of freedom
## Multiple R-squared:  0.07274,    Adjusted R-squared:  0.06806
## F-statistic: 15.57 on 2 and 397 DF,  p-value: 3.089e-07
```

The model is significant. Adding US changed the significance of the advertising effect.

6.

```
model_price <- lm(Sales ~ Price, data=Carseats)
summary(model_price)
```

```
##
## Call:
## lm(formula = Sales ~ Price, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5224 -1.8442 -0.1459  1.6503  7.5108
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.641915   0.632812  21.558   <2e-16 ***
## Price       -0.053073   0.005354  -9.912   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.532 on 398 degrees of freedom
## Multiple R-squared:  0.198,  Adjusted R-squared:  0.196
## F-statistic: 98.25 on 1 and 398 DF,  p-value: < 2.2e-16
```

Null hypothesis (H0): Price does not affect sales (i.e., coefficient of Price is 0). Alternative hypothesis (H1): Price affects sales (i.e., coefficient of Price is not 0). Check the coefficient of Price. This coefficient tells you the expected change in sales (in $1000) for each one dollar increase in price. It suggests that an increase in price leads to a decrease in sales. Generally, as the price of a product increases, the demand tends to decrease because fewer customers may be willing or able to afford the product at higher prices.

7.

```r
#permutation method

#theoretical method
t_test_result <- t.test(Sales ~ US, data=Carseats)
summary(t_test_result)
```

```
##             Length Class  Mode
## statistic   1      -none- numeric
## parameter   1      -none- numeric
## p.value     1      -none- numeric
## conf.int    2      -none- numeric
## estimate    2      -none- numeric
## null.value  1      -none- numeric
## stderr      1      -none- numeric
## alternative 1      -none- character
## method      1      -none- character
## data.name   1      -none- character
```

8. Yes, you can use the model from Q4. The regression model Sales ~ US directly provides an estimate of the difference in sales between US and non-US stores, which is exactly the hypothesis you are testing in Q7. If the coefficient for US in the model is significantly different from zero, and if the confidence interval for this coefficient does not include zero, it suggests that there is a significant difference in sales between US and non-US stores.

9.

```r
boot_diff <- function(data, indices){
  sample_data <- data[indices, ] # resampling with replacement
  diff_means <- with(sample_data, tapply(Sales, US, mean)[["Yes"]] - tapply(Sales, US, mean)[["No"]])
  return(diff_means)
}

results <- boot(data=Carseats, statistic=boot_diff, R=10000)

ci <- boot.ci(results, type="perc", conf=0.95)$percent[4:5]

cat("95% Confidence Interval:", ci, "\n")
```

```
## 95% Confidence Interval: 0.4876876 1.60992
```

Since zero is not within this interval, it suggests a significant difference at the 95% confidence level. However, just because zero is not in the confidence interval, it doesn't automatically validate the causal relationship or other underlying assumptions such as homogeneity of variances or absence of outliers.

10.

```r
anova_result <- aov(Sales ~ US, data=Carseats)
summary(anova_result)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## US            1   99.8   99.80   12.89 0.000372 ***
## Residuals   398 3082.5    7.74
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes, you can use ANOVA if the US variable is treated as a factor with two levels (US and Non-US). ANOVA tests the null hypothesis that all group means are equal, which aligns with testing whether there's a difference in sales between US and Non-US stores. The F-test in the ANOVA is significant, it suggests that there are statistically significant differences in sales between the groups, assuming the conditions for ANOVA are met (LINE test).

11.

```r
full_model <- lm(Sales ~ Advertising + Age + Price + ShelveLoc + US, data=Carseats)
summary(full_model)
```

```
##
## Call:
## lm(formula = Sales ~ Advertising + Age + Price + ShelveLoc +
##     US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3776 -1.0713 -0.0385  1.0935  4.4103
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     14.469548   0.517485  27.961   <2e-16 ***
## Advertising      0.108969   0.016394   6.647    1e-10 ***
## Age             -0.049990   0.004935 -10.129   <2e-16 ***
## Price           -0.061490   0.003380 -18.192   <2e-16 ***
## ShelveLocGood    4.817062   0.236857  20.337   <2e-16 ***
## ShelveLocMedium  1.939103   0.194757   9.957   <2e-16 ***
## USYes            0.006701   0.228126   0.029    0.977
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.585 on 393 degrees of freedom
## Multiple R-squared:  0.6896, Adjusted R-squared:  0.6848
## F-statistic: 145.5 on 6 and 393 DF,  p-value: < 2.2e-16
```

```r
best_model <- stepAIC(full_model, direction="backward")
```

```
## Start:  AIC=375.64
## Sales ~ Advertising + Age + Price + ShelveLoc + US
##
##               Df Sum of Sq     RSS    AIC
## - US           1      0.00  987.89 373.64
## <none>                      987.89 375.64
## - Advertising  1    111.05 1098.94 416.26
## - Age          1    257.91 1245.80 466.43
## - Price        1    831.88 1819.77 618.00
```

```
## - ShelveLoc    2   1049.69 2037.58 661.22
##
## Step:  AIC=373.64
## Sales ~ Advertising + Age + Price + ShelveLoc
##
##               Df Sum of Sq     RSS    AIC
## <none>                      987.89 373.64
## - Advertising  1    209.70 1197.59 448.64
## - Age          1    258.02 1245.91 464.46
## - Price        1    833.03 1820.92 616.25
## - ShelveLoc    2   1050.20 2038.09 659.32
```

```
summary(best_model)
```

```
##
## Call:
## lm(formula = Sales ~ Advertising + Age + Price + ShelveLoc, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3791 -1.0734 -0.0379  1.0946  4.4065
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     14.471205   0.513747  28.168   <2e-16 ***
## Advertising      0.109298   0.011951   9.145   <2e-16 ***
## Age             -0.049986   0.004927 -10.144   <2e-16 ***
## Price           -0.061486   0.003373 -18.227   <2e-16 ***
## ShelveLocGood    4.817154   0.236536  20.365   <2e-16 ***
## ShelveLocMedium  1.938788   0.194215   9.983   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.583 on 394 degrees of freedom
## Multiple R-squared:  0.6896, Adjusted R-squared:  0.6856
## F-statistic:   175 on 5 and 394 DF,  p-value: < 2.2e-16
```

Used backward elimination or stepwise regression to find the best model. Began with the full model and systematically remove the predictor with the highest p-value (above a chosen threshold like 0.05), one at a time, re-running the model each time. Showed intermediate results: the output from summary() for each step. Chose predictors based on statistical significance (p-values), potential multicollinearity (via VIFs), and model diagnostics (like AIC for model comparison).

12.

```
summary(best_model)
```

```
##
## Call:
## lm(formula = Sales ~ Advertising + Age + Price + ShelveLoc, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3791 -1.0734 -0.0379  1.0946  4.4065
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     14.471205   0.513747  28.168   <2e-16 ***
## Advertising      0.109298   0.011951   9.145   <2e-16 ***
## Age             -0.049986   0.004927 -10.144   <2e-16 ***
## Price           -0.061486   0.003373 -18.227   <2e-16 ***
## ShelveLocGood    4.817154   0.236536  20.365   <2e-16 ***
## ShelveLocMedium  1.938788   0.194215   9.983   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.583 on 394 degrees of freedom
## Multiple R-squared:  0.6896, Adjusted R-squared:  0.6856
## F-statistic:   175 on 5 and 394 DF,  p-value: < 2.2e-16
```

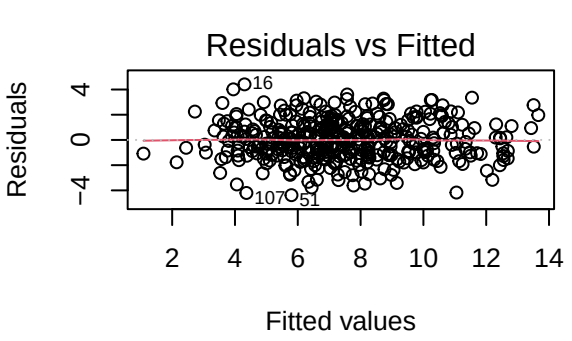
```
age_coefficient <- coef(best_model)["Age"]

cat("Estimated coefficient for Age:", age_coefficient, "\n")
```

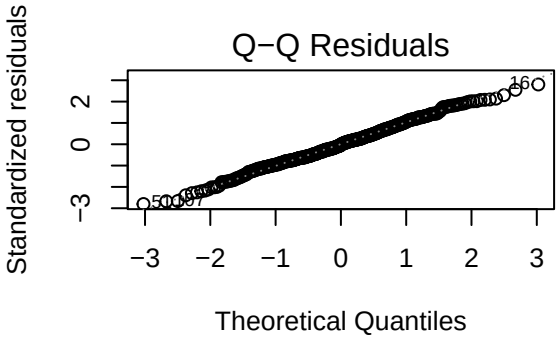
```
## Estimated coefficient for Age: -0.04998602
```

The estimated coefficient for Age from best_model explains the change in sales (in $1000s) for each one-year increase in the average age of the local population, holding all other factors constant. A negative coefficient indicates that as the average age increases, sales decrease, perhaps suggesting that younger populations are more likely buyers of car seats.

13.

```
par(mfrow=c(2,2))
plot(best_model, which=1)
plot(best_model, which=2)
hist(resid(best_model), main="Histogram of Residuals")
```





Residuals vs Fitted: Ideally, there should be a random scatter. Histogram: Should show no obvious skewness or outliers; ideally, symmetric. Q-Q Plot: Points lie roughly along the line

14.

```
table_data <- table(Carseats$ShelveLoc, Carseats$US)
chisq.test(table_data)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_data
## X-squared = 2.7397, df = 2, p-value = 0.2541
```

Null hypothesis (H0): There is no association between shelving quality and US store indicator. Alternative hypothesis (H1): There is an association.

15.

```
anova_model <- aov(Sales ~ ShelveLoc, data=Carseats)
summary(anova_model)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## ShelveLoc     2   1010   504.8   92.23 <2e-16 ***
## Residuals   397   2173     5.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

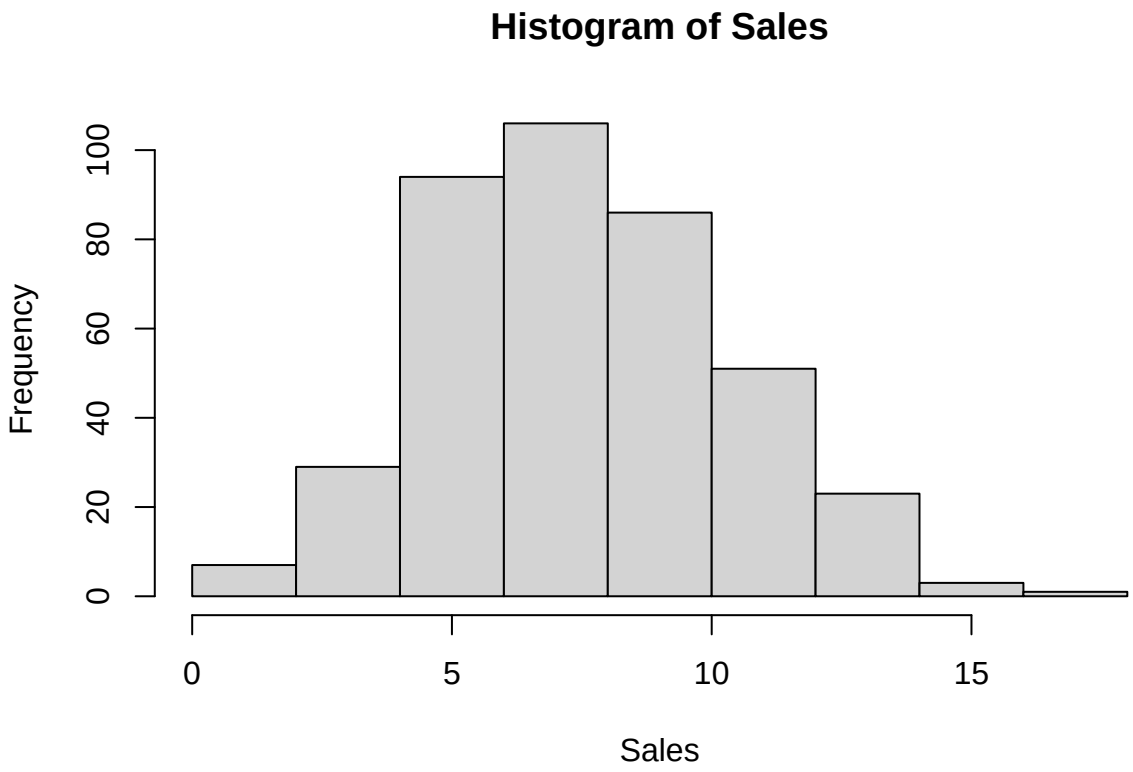
```
kruskal.test(Sales ~ ShelveLoc, data=Carseats)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Sales by ShelveLoc
## Kruskal-Wallis chi-squared = 119.39, df = 2, p-value < 2.2e-16
```

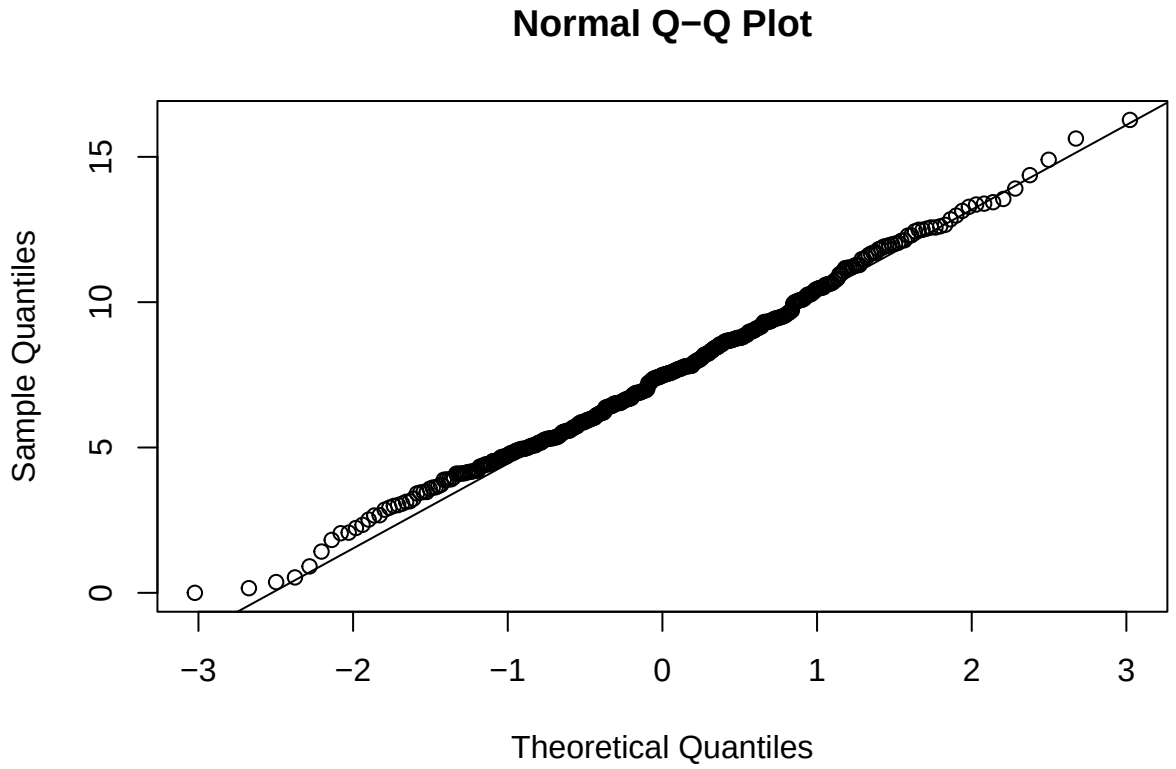
Null hypothesis (H0): No differences in average sales among the shelving quality levels. Alternative hypothesis (H1): Differences exist. Compared the results from ANOVA; consistent results across both methods strengthen the conclusions.

16.

```
hist(Carseats$Sales, main="Histogram of Sales", xlab="Sales")
```




```
qqnorm(Carseats$Sales)
qqline(Carseats$Sales)
```



Yes, the histogram appears to be symmetrical and bell-shaped.

17.

```r
sales_mean <- mean(Carseats$Sales)
sales_sd <- sd(Carseats$Sales)

within_one_sd <- mean(Carseats$Sales >= (sales_mean - sales_sd) & Carseats$Sales <= (sales_mean + sales_
more_than_one_sd_right <- mean(Carseats$Sales > (sales_mean + sales_sd))
more_than_one_sd_left <- mean(Carseats$Sales < (sales_mean - sales_sd))

cat("Proportion within one SD:", within_one_sd, "\n")
```

```
## Proportion within one SD: 0.685
```

```r
cat("Proportion more than one SD to the right:", more_than_one_sd_right, "\n")
```

```
## Proportion more than one SD to the right: 0.165
```

```r
cat("Proportion more than one SD to the left:", more_than_one_sd_left, "\n")
```

```
## Proportion more than one SD to the left: 0.15
```

34% within one standard deviation on either side of the mean. 16% more than one standard deviation on the right of the mean. 16% more than one standard deviation on the left of the mean.

PART II

18.

```r
data <- read.csv("beefbacteria.csv")

data_a <- data[data$method == "A",]

mean_function <- function(data, indices) {
  return(mean(data[indices]))
}

bootstrap_results_a <- boot(data_a$bacteria, statistic = mean_function, R = 10000)
ci_bootstrap_a <- boot.ci(bootstrap_results_a, type = "perc")$percent[4:5]

mean_a <- mean(data_a$bacteria)
sd_a <- sd(data_a$bacteria)
n_a <- length(data_a$bacteria)
error_margin_a <- qt(0.975, df = n_a - 1) * sd_a / sqrt(n_a)
ci_se_a <- c(mean_a - error_margin_a, mean_a + error_margin_a)

list(ci_bootstrap = ci_bootstrap_a, ci_se = ci_se_a)
```

```
## $ci_bootstrap
## [1] 28.73403 31.80692
##
## $ci_se
## [1] 28.70215 31.83985
```

19.

```r
data <- read.csv("beefbacteria.csv")

data_b <- data[data$method == "B",]

mean_function <- function(data, indices) {
```

```r
  return(mean(data[indices]))
}

bootstrap_results_b <- boot(data_b$bacteria, statistic = mean_function, R = 10000)
ci_bootstrap_b <- boot.ci(bootstrap_results_b, type = "perc")$percent[4:5]

mean_b <- mean(data_b$bacteria)
sd_b <- sd(data_b$bacteria)
n_b <- length(data_b$bacteria)
error_margin_b <- qt(0.975, df = n_b - 1) * sd_b / sqrt(n_b)
ci_se_b <- c(mean_b - error_margin_b, mean_b + error_margin_b)

list(ci_bootstrap = ci_bootstrap_b, ci_se = ci_se_b)
```

```
## $ci_bootstrap
## [1] 28.00603 31.32987
##
## $ci_se
## [1] 28.00672 31.36328
```

20.

```r
mean_b <- mean(data_b$bacteria)
sd_b <- sd(data_b$bacteria)
n_b <- length(data_b$bacteria)

mean_diff <- mean_a - mean_b
se_diff <- sqrt(sd_a^2/n_a + sd_b^2/n_b)
ci_diff <- mean_diff + c(-1, 1) * qt(0.975, df = min(n_a - 1, n_b - 1)) * se_diff

ci_diff
```

```
## [1] -1.711373  2.883373
```

21.

```r
t_test_results <- t.test(data_a$bacteria, data_b$bacteria, var.equal = TRUE)
t_test_results
```

```
##
##  Two Sample t-test
##
## data:  data_a$bacteria and data_b$bacteria
## t = 0.50612, df = 198, p-value = 0.6133
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.697248  2.869248
## sample estimates:
## mean of x mean of y
##    30.271    29.685
```

22. If the 95% CI for the difference in means between Method A and B does not include zero, it indicates a statistically significant difference at the 5% level.