Assignment 1

2022-10-01

Question 4

a) Using the parametrization $\mu = 0$:

```
data("iris");

Y <- iris[order(iris$Species), "Sepal.Width"];

X <- diag(3) %x% rep(1, 50);

n = 150;
I = 3;
```

Then we calculate the estimated $\hat{\beta} = (X^T X)^{-1} X^T Y$ as

```
beta = solve((t(X) %*% X)) %*% t(X) %*% Y;
```

The residual sum of squares S_Ω and S_ω of the full and reduced models respectively are

```
s1 = norm(Y - X %*% beta, type="2")^2;
s2 = norm(Y - matrix(rep(1, n),ncol=1) * mean(Y), type="2")^2;
```

The unbiased estimator of σ^2 are $\frac{S_\Omega}{n-1} = 16.962$ and $\frac{S_\omega}{n-1} = 28.3069333$.

```
unb_est = s1/(n - I);
bet_ss = s2 - s1;
bet_means = (s2 - s1)/(I);
f_val = ((s2 - s1)/(I - 1))/(s2/(n - I));

within_means = s1/(n - I);
```

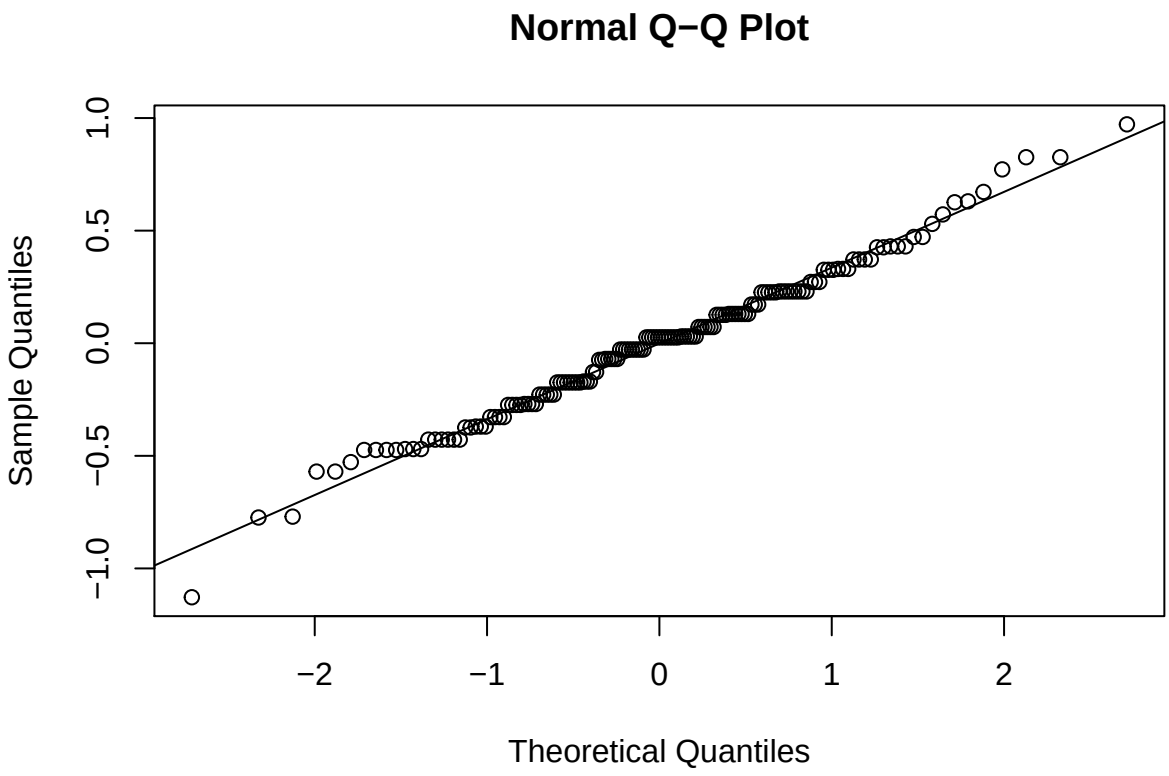
The quantities needed to complete an ANOVA table are :

- Between groups sum of square: $S_\omega - S_\Omega = 11.3449333$.
- Between groups mean square: $\frac{(S_\omega - S_\Omega)/(I-1)}{S_\Omega/(n-I)} = 3.7816444$.
- Within groups sum of square: $S_\Omega = 16.962$.
- F value $= 29.4575393$.

b) We first check for the model assumptions:

The normality of residuals with expectation zero are checked using QQ-plot, Shapiro-Wilk test, and one-sample t-test.

```
plot.new()
res <- Y - c(rep(beta, 1, each=50));
qqnorm(res)
qqline(res)
```




```
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.98948, p-value = 0.323
```

```
t.test(res, mu=0, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  res
## t = -1.2788e-14, df = 149, p-value = 0.5
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -0.04559694         Inf
## sample estimates:
##     mean of x
## -3.523043e-16
```

Since the p-values for both test are larger than 0.05, with the mean of the residuals being extremely close, we can say that the normality and zero mean assumptions hold. Next we check that $Var(e_{ij}) = \sigma^2$ using Bartlett test.

```
species <- factor(c(rep(1, 50), rep(2, 50), rep(3, 50)),
                  labels=c("setosa", "versicolor", "virginica"));
```

```r
data_iris <- data.frame(Y, species);

bartlett.test(Y ~ species, data=data_iris)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Y by species
## Bartlett's K-squared = 2.0911, df = 2, p-value = 0.3515
```

Thus, the model assumptions hold. Now we test for the mean of iris sepal width of the three species using the produced F-statistic above.

```r
pv <- pf(f_val, I - 1, n - I, lower.tail = FALSE);
```

The p-value is $1.7447978 \times 10^{-11} < 0.05$ so we can reject the null hypothesis that the means are statistically the same.

c)

```r
model <- aov(Y ~ species, data=data_iris);

summary(model)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## species       2  11.35   5.672   49.16 <2e-16 ***
## Residuals   147  16.96   0.115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen that the results from ANOVA agree with the final conclusion although some quantities are a bit off.

d)

```r
kruskal.test(Y ~ species, data=data_iris);
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Y by species
## Kruskal-Wallis chi-squared = 63.571, df = 2, p-value = 1.569e-14
```

Thus, the Kruskal-Wallis test agrees with our findings since its p-value is smaller than 0.05, and because we the normal distribution assumption holds, the location parameters are the means.