STA305/1004 Class Notes - Week 6

Nathan Taback

February 12, 2016

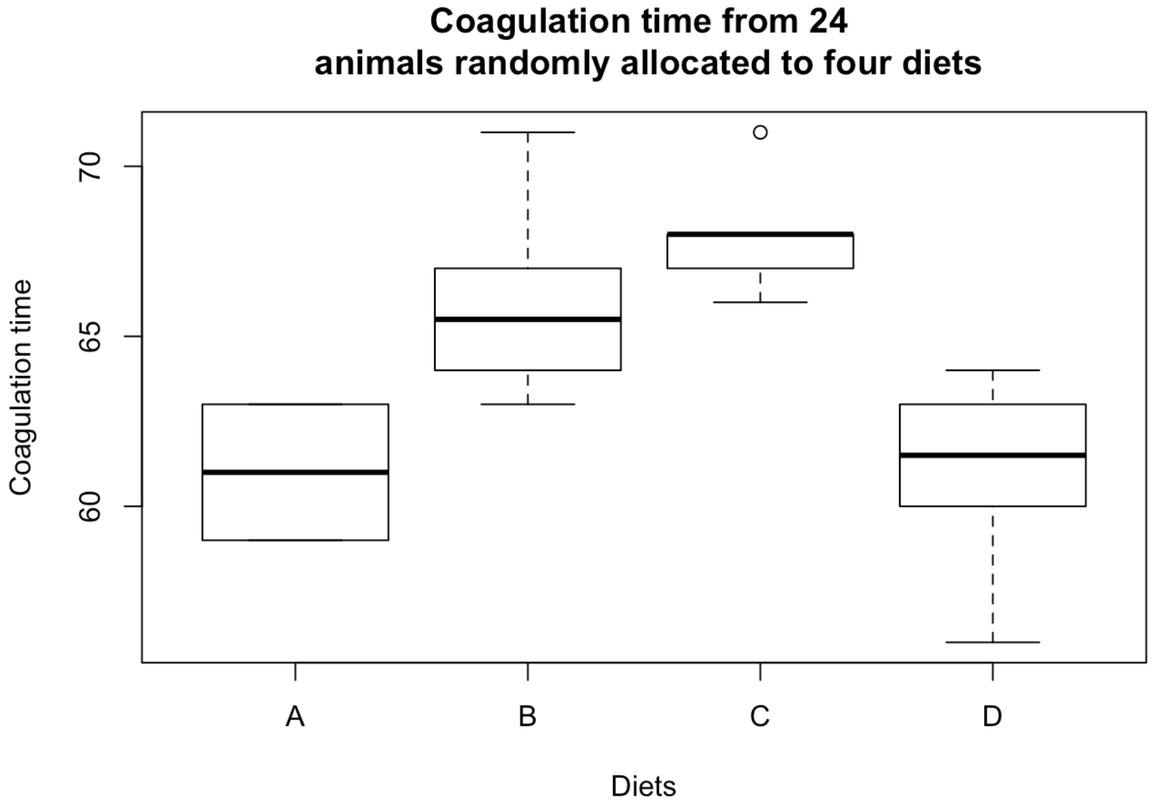
ANOVA - Comparing more than two groups

The following example is taken from Chapter 4 of Box, Hunter, and Hunter (2005). The table below gives coagulation times for blood samples drawn from 24 animals receiving four different diets A, B, C, and D.

	A	B	C	D
	60	65	71	62
	63	66	66	60
	59	67	68	61
	63	63	68	64
	62	64	67	63
	59	71	68	56
Treatment Average	61	66	68	61
Grand Average	64	64	64	64
Difference	-3	2	4	-3

Boxplots of the data are shown below.

```
boxplot(y~diets,data=tab0401,xlab="Diets",ylab="Coagulation time",main="Coagulation time from 2
4 \n animals randomly allocated to four diets")
```



Question: Is there evidence to indicate a difference in mean coagulation times for the four different diets?

An idea due to Fisher is to compare the variation in mean coagulation times *between* the diets to the variation of coagulation times *within* a diet. These two measures of variation are often summarized in an analysis of variance (ANOVA) table.

Analysis of Variance (ANOVA) table

The between treatments variation and within treatment variation are two components of the total variation in the response.

The coagulation study data we can break up each observation's deviation from the grand mean into two components: treatment deviations; and residuals within treatment deviations.

$$y_{ij} - \bar{y}_{\cdot\cdot} = (y_{i\cdot} - \bar{y}_{\cdot\cdot}) + \left(y_{ij} - \bar{y}_{i\cdot}\right)$$

Let y_{ij} be the jth observation taken under treatment $i = 1, \ldots, a$.

$$E(y_{ij}) = \mu_i = \mu + \tau_i,$$

and $Var(y_{ij}) = \sigma^2$ and the observations are mutually independent. The parameter τ_i is the ith treatment effect.

We are interested in testing if the a treatment means are equal.

$$H_0 : \mu_1 = \cdots = \mu_a \quad \text{vs.} \quad H_1 : \mu_i \neq \mu_j,\ i \neq j.$$

There will be n observations under the ith treatment.

$$y_{i\cdot} = \sum_{j=1}^{n} y_{ij}, \qquad \bar{y}_{i\cdot} = y_{i\cdot}/n,$$

$$y_{\cdot\cdot} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}, \qquad \bar{y}_{\cdot\cdot} = y_{\cdot\cdot}/N,$$

where $N = an$ is the total number of observations. The "dot" subscript notation means sum over the subscript that it replaces.

The ANOVA identity

The total sum of squares $SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2$ can be written as

$$\sum_{i=1}^{a}\sum_{j=1}^{n}\left[(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (y_{ij} - \bar{y}_{i\cdot})\right]^2$$

by adding and subtracting $\bar{y}_{i\cdot}$ to SS_T.

It can be shown that

$$\begin{aligned} SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2 &= \underbrace{n\sum_{i=1}^{a}(y_{i\cdot} - \bar{y}_{\cdot\cdot})^2}_{\text{Sum of Squares Due to Treatment}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2}_{\text{Sum of Squares Due to Error}} \\ &= SS_{Treat} + SS_E. \end{aligned}$$

This is sometimes called the analysis of variance identity. It shows how the total sum of squares can be split into two sum of squares: one part that is due to differences between treatments; and one part due to differences within treatments.

For example, the decomposition of the first observation $y_{11} = 60$ in diet A is

$$
\begin{aligned}
y_{11} - \bar{y}_{..} &= (y_{1.} - \bar{y}_{..}) + \left(y_{11} - \bar{y}_{1.}\right) \\
60 - 64 &= (61 - 64) + (60 - 61) \\
-4 &= -3 + -1
\end{aligned}
$$

Example - Blood coagulation study

The deviations from the grand average $\left(y_{ij} - \bar{y}_{..}\right)$ are in the table below:

```
  A  B C  D
-4  1 7 -2
-1  2 2 -4
-5  3 4 -3
-1 -1 4  0
-2  0 3 -1
-5  7 4 -8
```

The total sum of squares is obtained by squaring all the entries in this table and summing: $SS_T = (-4)^2 + (-1)^2 + \cdots + (-8)^2 =$ 340.

The between treatment deviations $(y_{i.} - \bar{y}_{..})$ are in the table below:

```
  A B C  D
-3 2 4 -3
-3 2 4 -3
-3 2 4 -3
-3 2 4 -3
-3 2 4 -3
-3 2 4 -3
```

The sum of squares due to treatment is obtained by squaring all the entries in this table and summing: $SS_{Treat} = (-3)^2 + (2)^2 + \cdots + (-3)^2 =$ 228.

The within treatment deviations $\left(y_{ij} - \bar{y}_{i.}\right)$ are in the table below:

```
  A  B  C  D
-1 -1  3  1
 2  0 -2 -1
-2  1  0  0
 2 -3  0  3
 1 -2 -1  2
-2  5  0 -5
```

The sum of squares due to error $\left(y_{ij} - \bar{y}_{i.}\right)$ is obtained by squaring the entries in this table and summing: $SS_E = (-1)^2 + (2)^2 + \cdots + (-5)^2 =$ 112.

$$
\underbrace{340}_{SS_T} = \underbrace{228}_{SS_{Treat}} + \underbrace{112}_{SS_E}.
$$

Which illustrates the ANOVA identity for the blood coagulation study.

The deviations

- SS_{Treat} is called the sum of squares due to treatments (i.e., between treatments), and SS_E is called the sum of squares due to error (i.e., within treatments).
- There are $an = N$ total observations. So SS_T has $N - 1$ degrees of freedom.
- There are a treatment levels so SS_{Treat} has $a - 1$ degrees of freedom.
- Within each treatment there are n replicates with $n - 1$ degrees of freedom. There are a treatments. So, there are $a(n-1) = an - a = N - a$ degrees of freedom for error.

$$SS_E = \sum_{i=1}^{a} \left[\sum_{j=1}^{n} \left(y_{ij} - \bar{y}_{i\cdot} \right)^2 \right]$$

If the term inside the brackets is divided by $n - 1$ then it is the sample variance for the ith treatment

$$S_i^2 = \frac{\sum_{j=1}^{n} \left(y_{ij} - \bar{y}_{i\cdot} \right)^2}{n-1}, \qquad 1 = 1, \ldots, a.$$

Combining these a variances to give a single estimate of the common population variance

$$\frac{(n-1)S_1^2 + \cdots + (n-1)S_a^2}{(n-1) + \cdots + (n-1)} = \frac{SS_E}{N-a}.$$

Thus, SS_E is a pooled estimate of the common variance σ^2 within each of the a treatments.

If there were no differences between the a treatment means $\bar{y}_{i.}$ we could use the variation of the treatment averages from the grand average to estimate σ^2.

$$\frac{SS_{Treat}}{a-1}$$

is an estimate of σ^2 of the treatment means are all equal.

The analysis of variance identity gives two estimates of σ^2. One is based on the variability within treatments and one based on the variability between treatments. If there are no differences in the treatment means then these two estimates should be similar. If these estimates are different then this could be evidence that the difference is due to differences in the treatment means.

The mean square for treatment is defined as

$$MS_{Treat} = \frac{SS_{Treat}}{a-1}$$

and the mean square for error is defined as

$$MS_E = \frac{SS_E}{N-a}.$$

SS_{Treat} and SS_E are independent and it can be shown that $SS_{Treat}/\sigma^2 \sim \chi^2_{a-1}$ and $SS_E/\sigma^2 \sim \chi^2_{N-a}$. Thus, if $H_0 : \mu_1 = \cdots = \mu_a$ is true then the ratio

$$F = \frac{MS_{Treat}}{MS_E} \sim F_{a-1,N-a}.$$

The ANOVA table for the coagulation data can be calculated in R.

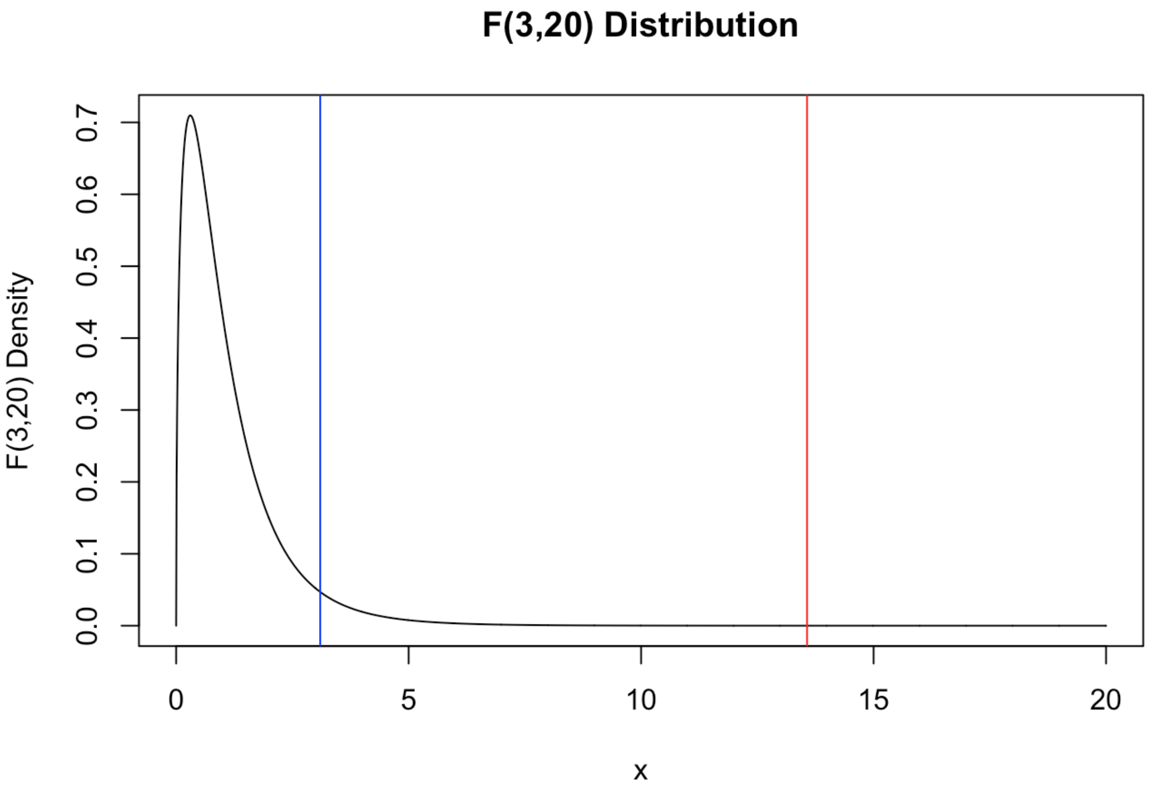
```
aov.diets <- aov(y~diets,data=tab0401)
summary(aov.diets)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
diets        3    228    76.0   13.57 4.66e-05 ***
Residuals   20    112     5.6                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this example $a - 1 = 3, N - a = 20, SS_{Treat} = 228, SS_E = 112, MS_{Treat} = 228/3 = 76.0, MS_E 112/20 = 5.6, F = 76/5.6 = 13.57.$

The observed F value of 13.57 is shown on the $F_{3,20}$ distribution. The p-value of the test is the area under the density to the right of 13.57 (red line). The 95% critical value of the $F_{3,20}$ is 3.10 (blue line). In other words, $P(F_{3,20} > 3.10) = 0.05$.

```
x <-seq(0,20,by=0.01)
plot(x,df(x = x,df1 = 3,df2 = 20),type="l",ylab="F(3,20) Density",main = "F(3,20) Distributio
n")
abline(v=13.57,col="red")
abline(v=qf(p = 0.95,3,20),col="blue")
```



The p-value could also be calculated directly using the cdf of the $F_{3,20}$ distribution.

```
1-pf(q = 13.57,df1 = 3,df2 = 20)
```

```
[1] 4.66169e-05
```

The small p-value indicates that the difference between at least one pair of the treatment means is significantly different from 0.

General ANOVA

The general form of the ANOVA table is

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Between treatments	$a - 1$	SS_{Treat}	MS_{Treat}	
Within treatments	$N - a$	SS_E	MS_E	$F = \frac{MS_{Treat}}{MS_E}$

ANOVA Assumptions

The calculations that make up an ANOVA table require no assumptions. You could write 24 numbers in the ANOVA table and complete the table using the ANOVA identity and definitions of mean square and F statistic. However, using these numbers to make inferences about differences in treatment means will require certain assumptions.

Assumptions

1. Additive model.

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}.$$

The parameters τ_i are interpreted as the treatment effect of the i^{th} mean. That is, if μ_i is the mean of i^{th} group and μ is the overall mean then $\tau_i = \mu_i - \mu$.

2. The errors ϵ_{ij} are independent and identically distributed (iid) with common variance $Var(\epsilon_{ij}) = \sigma^2$, for all i, j. If the errors are iid then

$$E(MS_{Treat}) = \sum_{i=1}^{a} \tau_i^2 + \sigma^2, \qquad E(MS_E) = \sigma^2.$$

If there are no differences between the treatment means then $\tau_1 = \cdots = \tau_4$ then both MS_{treat} and MS_E estimate σ^2.

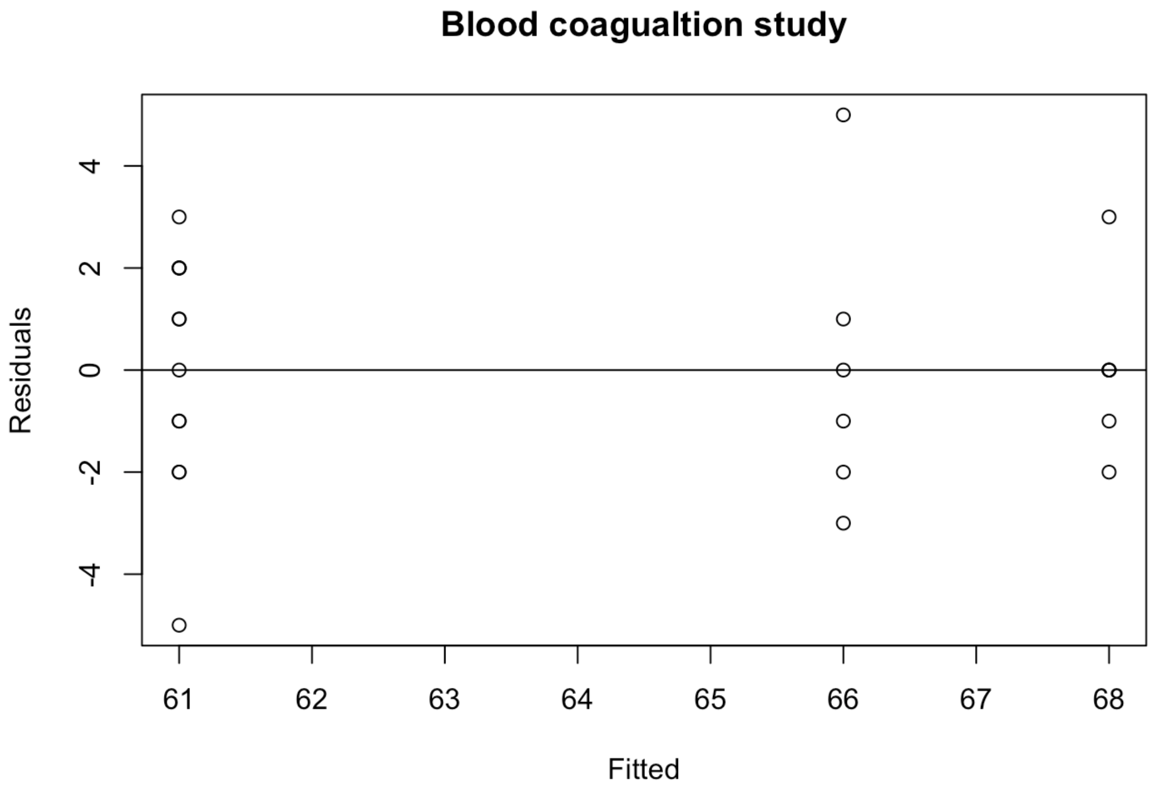
3. If $\epsilon_{ij} \sim N(0, \sigma^2)$ then MS_{Treat} and MS_E are independent. Under the null hypothesis that $\sum_{i=1}^{a} \tau_i^2 = 0$ the ratio $F = \frac{MS_{Treat}}{MS_E}$ is the ratio of two independent estimates of σ^2. Therefore, $\frac{MS_{Treat}}{MS_E} \sim F_{a-1,N-a}$.

Example - checking the assumptions in the blood coagualtion study

1. The additive model assumption seems plausible since the observations from each diet can be viewed as the sum of a common mean plus a random error term.

2. The common variance assumption can be investigated by plotting the residuals versus the fitted values of the ANOVA model. A plot of the residuals versus fitted values can be used to investigate the assumption that the residuals are randomly distributed and have constant variance. Ideally, the points should fall randomly on both sides of 0, with no recognizable patterns in the points.In the R this can be done using the following commands.

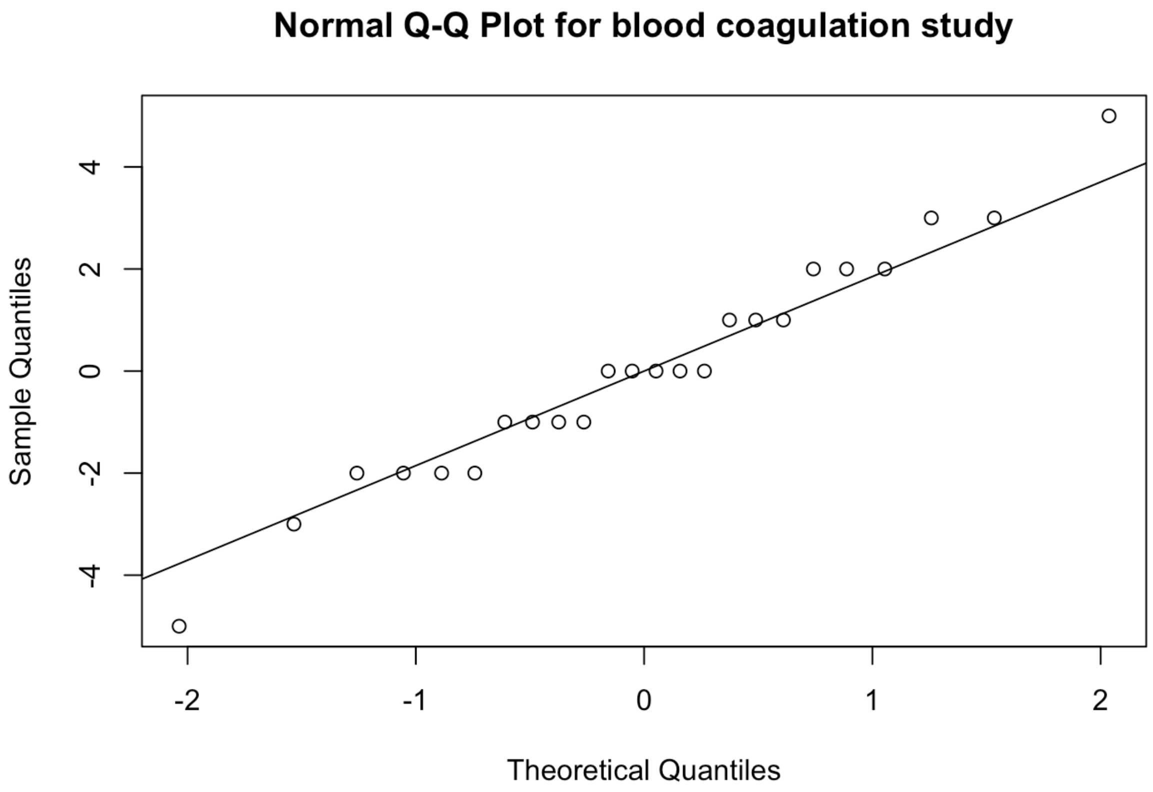
```
plot(aov.diets$fitted.values,aov.diets$residuals,ylab="Residuals",xlab="Fitted",main="Blood co
agualtion study")
abline(h=0)
```



The assumption of constant variance is satisfied for the blood coagulation study.

3. The normality of the residuals can be investigated using a normal quantile-quantile plot.

```
qqnorm(aov.diets$residuals, main="Normal Q-Q Plot for blood coagulation study")
qqline(aov.diets$residuals)
```



The normality assumptions is satisfied.

Estimating treatment effects using least squares

The model for diet $y_{ij} = \mu + \tau_i + \epsilon_{ij}$ can be written in terms of the dummy variables X_1, X_2, X_3 as:

$$y_{ij} = \mu + \tau_1 X_1 + \tau_2 X_2 + \tau_3 X_3 + \epsilon_{ij},$$

where,

$$X_1 = \begin{cases} 1 & \text{if diet B} \\ 0 & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if diet C} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if diet D} \\ 0 & \text{otherwise} \end{cases}$$

The least squares estimates are:

$$\begin{aligned} \hat{\mu} &= \bar{y}_{1\cdot}, \\ \hat{\tau}_1 &= \bar{y}_{2\cdot} - \bar{y}_{1\cdot}, \\ \hat{\tau}_2 &= \bar{y}_{3\cdot} - \bar{y}_{1\cdot}, \\ \hat{\tau}_3 &= \bar{y}_{3\cdot} - \bar{y}_{1\cdot}. \end{aligned}$$

Using least squares to estimate the parameters

```
lm.diets <- lm(y~diets,data=tab0401)
summary(lm.diets)
```

```


Call:
lm(formula = y ~ diets, data = tab0401)

Residuals:
   Min     1Q Median     3Q    Max 
 -5.00  -1.25   0.00   1.25   5.00 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.100e+01  9.661e-01  63.141  < 2e-16 ***
dietsB       5.000e+00  1.366e+00   3.660  0.00156 ** 
dietsC       7.000e+00  1.366e+00   5.123 5.18e-05 ***
dietsD      -9.999e-15  1.366e+00   0.000  1.00000    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.366 on 20 degrees of freedom
Multiple R-squared:  0.6706,	Adjusted R-squared:  0.6212 
F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

The averages for each of the four diets are in the table below.

Diet	A (t=1)	B (t=2)	C (t=3)	D (t=4)
Average ($y_{t.}$)	61	66	68	61

So we can verify that the least-squares estimates are differences of the treatment averages.

$$\begin{aligned}
\bar{y}_{1.} &= 61, \\
\hat{\tau}_1 &= \bar{y}_{2.} - \bar{y}_{1.} = 5 \\
\hat{\tau}_2 &= \bar{y}_{3.} - \bar{y}_{1.} = 7 \\
\hat{\tau}_3 &= \bar{y}_{3.} - \bar{y}_{1.} = -9.9 \times 10^{-15}.
\end{aligned}$$

Questions

Question A

Let $\mu_A, \mu_B, \mu_C, \mu_D$ be the mean coagulation times of diets A, B, C, and D respectively.

i. Formulate a null and alternative hypotheses to compare the mean coagulation times between the four diets.

ii. What is the test statistic and P-value of the test in part (a)?

iii. Is there a significant difference (at the 1% significance level) between at least two of the diets?

iv. What are the assumptions behind:

- The ANOVA table calculations.
- The P-value in the ANOVA table.

Question B

Interpret the parameters in the additive model for ANOVA

$$y_{ti} = \mu + \tau_t + \epsilon_{ti},$$

where, y_{ti} is the i^{th} observation in the t^{th} treatment group, μ is the overall mean, and τ_t is the deviation produced by treatment t, and ϵ_{ti} is the error.

Answers to questions

Question A

Let $\mu_1, \mu_2, \mu_3, \mu_4$ be the mean coagulation times of diets A, B, C, and D respectively.

i. Formulate a null and alternative hypotheses to compare the mean coagulation times between the four diets.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

versus

$$H_1 : \mu_i \neq \mu_j, i \neq j.$$

ii. What is the test statistic and P-value of the test in part (a)?

> The F statistic is 13.571 and the p-value is 4.658e-05.

iii. Is there a significant difference (at the 1% significance level) between at least two of the diets?

> Yes, since the p-value is less than or equal to 0.01.

iv. What are the assumptions behind:

- The ANOVA table calculations.

> There are no assumptions required to carry out the calcualtions for the sums of squares , degrees of freedom, or mean squares. (see BHH pg. 138)

- The P-value in the ANOVA table.

> Additive model, errors are independent, errors are normally distributed with constant variance.
>
> Based on plots below residuals versus fitted values and the normal Q-Q of residuals the normal distribution and constant variance assumptions are fullfilled.
>
> The additive model seems plausible in this case since effects should be additive for coagualtion times. Coagulation times of rats are independent since the time for one rat does not depend on another rat.

Question B

Interpret the parameters in the additive model for ANOVA

$$y_{ti} = \mu + \tau_t + \epsilon_{ti},$$

where, y_{ti} is the i^{th} observation in the t^{th} treatment group, μ is the overall mean, and τ_t is the deviation produced by treatment t, and ϵ_{ti} is the error.

> $\hat{\mu} = \bar{y}_1$ and $\hat{\mu} + \hat{\tau}_t = \bar{y}_t, t = 2, 3, 4$, where $\bar{y}_t = \sum_{k=1}^{n_t} y_{tk}/n_t$ is the mean for for the n_t units in treatment t. So, the least squares estimates $\hat{\tau}_t = \bar{y}_t - \bar{y}_1, t = 2, 3, 4$.
>
> ϵ_{ti} has a $N(0, \sigma^2)$ distribution.

Blood coagulation - ANOVA table

```
lm.diets <- lm(y~diets,data=tab0401)
anova(lm.diets)
```

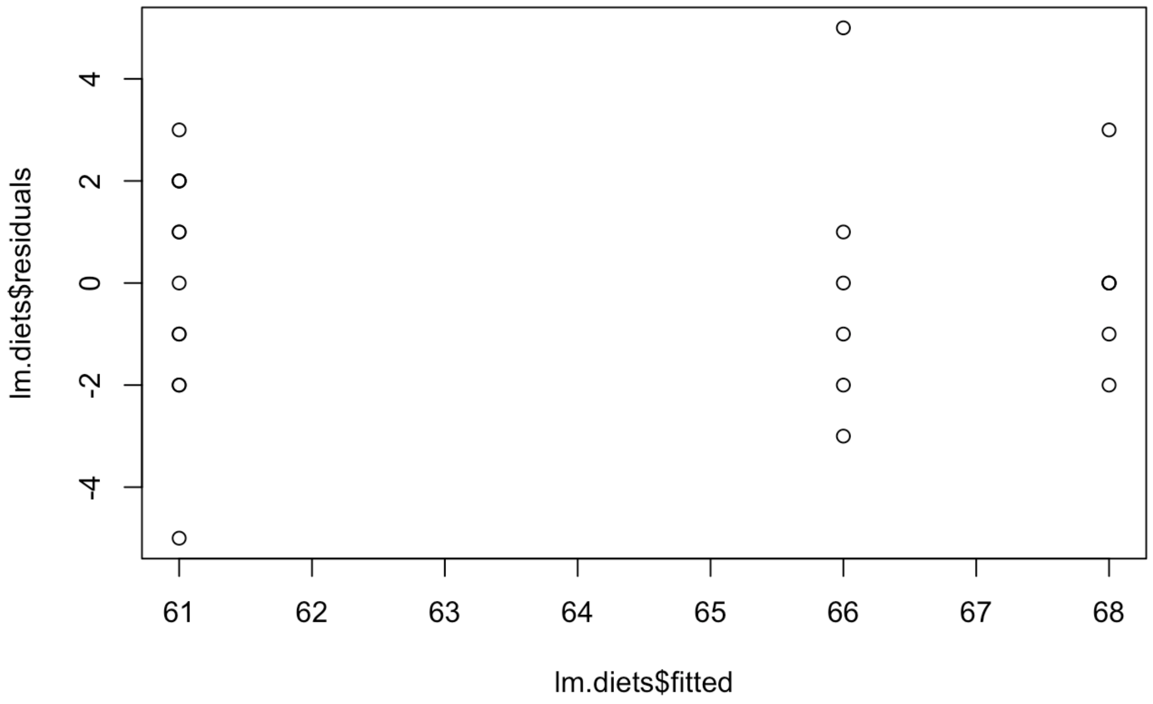
```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diets      3    228    76.0  13.571 4.658e-05 ***
## Residuals 20    112     5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Blood coagulation - linear model parameter estimates

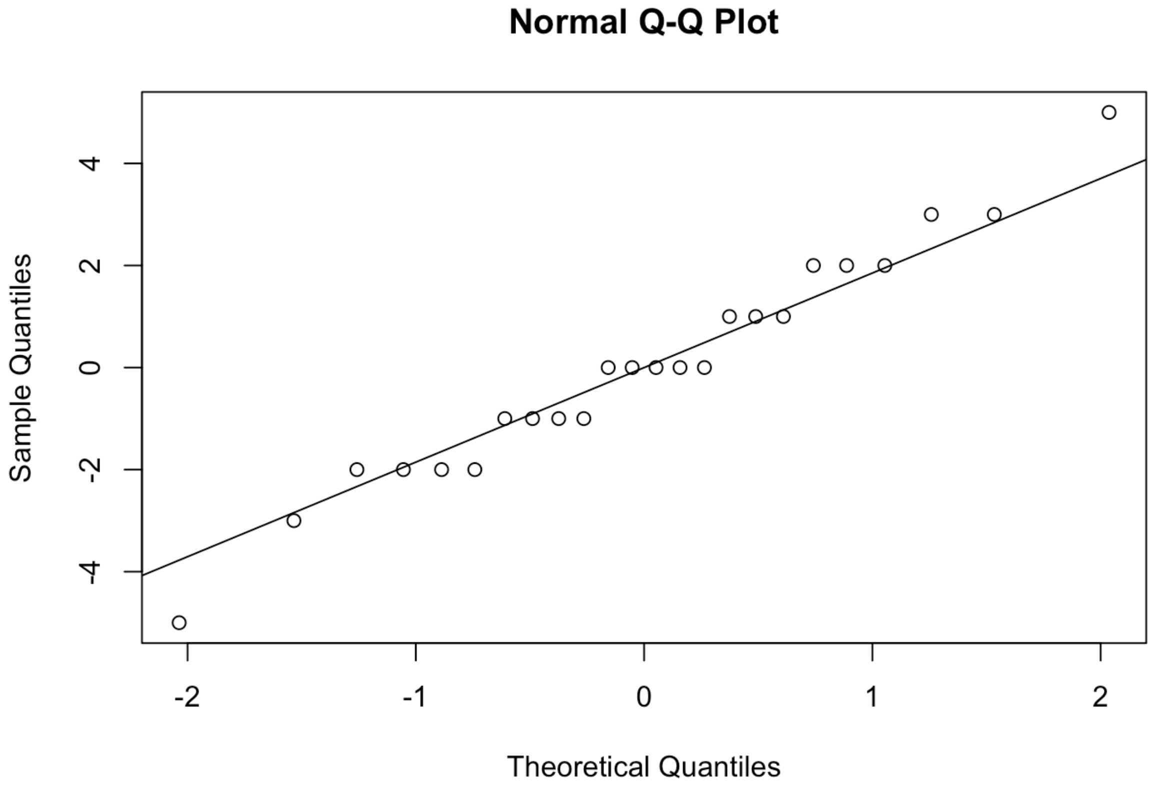
```
summary(lm.diets)
```

```
##
## Call:
## lm(formula = y ~ diets, data = tab0401)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5.00  -1.25   0.00   1.25   5.00
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.100e+01  9.661e-01  63.141  < 2e-16 ***
## dietsB       5.000e+00  1.366e+00   3.660  0.00156 **
## dietsC       7.000e+00  1.366e+00   5.123 5.18e-05 ***
## dietsD      -9.999e-15  1.366e+00   0.000  1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.366 on 20 degrees of freedom
## Multiple R-squared:  0.6706, Adjusted R-squared:  0.6212
## F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

```
plot(lm.diets$fitted,lm.diets$residuals)
```



```
qqnorm(lm.diets$residuals);qqline(lm.diets$residuals)
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles