Chapter 11: Linear models with a single factor explanatory variable having three or more levels (One-way analysis of variance)

STATS 201/8

University of Auckland

Learning Outcomes

In this chapter you will learn about:

- Fitting a model with a single explanatory factor variable with multiple levels, a.k.a., one-way analysis of variance (ANOVA)[1]
- Interpretting the fitted model
- Multiple pairwise comparisons of means
- Using `emmeans` to solve the multiple comparisons problem
- Relevant `R`-code
- Alternative parameterizations of the one-way ANOVA model[2]

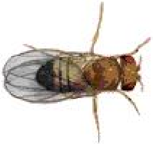
[1] **NOTE:** When people use the term ANOVA (Analysis of Variance), they are referring to a linear regression model in which all the explanatory variable are factors.

[2] Optional section.

Section 11.1
Example with a 5-level explanatory factor variable

Example – Fruit flies

In this case study we look at how the male fruit-fly's longevity is related to his reproductive activity.



Fruit flies are a very commonly used animal for laboratory experiments because they are easy to maintain and breed. Their short lifespan allows several generations to be observed within a few months. They also have a genome that is very close to that of humans with many genes discovered in humans also found in fruit flies.[3]

Previous studies have shown that the longevity (life span) of female fruit flies decreases with an increase in reproduction, and this leads to a similar question related to males.

[3] See https://www.yourgenome.org/facts/why-use-the-fly-in-research for more background on research with fruit flies.

Fruit fly

The experiment compared the lifespan of males that were divided into 5 treatment groups that varied according to the presence or absence and number of uninterested or interested females.[4]

How does one define “interest” in female fruit flies? Here is this study’s definition:

> Newly inseminated females will not usually mate again for at least two days. So, the males in the uninterested females treatment groups were always living with newly inseminated females.

The **primary focus of this Example** is the following: Due to the explanatory factor variable (treatment group) having several levels, we will need to apply an adjustment to relevant *P*-values and confidence intervals when we are making inference about differences in the expected lifespans between pairs of treatment groups.

[4] Had there been only two treatment groups then we could have used the two sample two-sample *t*-test discussed in Chapter 5

Fruit fly. . .

The response variable measured was `days`, the number of days the male fly lived.

The explanatory factor variable was `group`, with five levels:

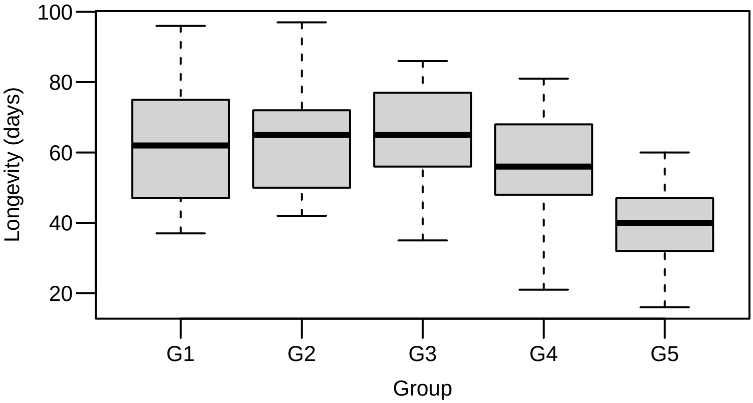
- `G1` males living alone,
- `G2` males living with one interested female,
- `G3` males living with eight interested females,
- `G4` males living with one uninterested female, and
- `G5` males living with eight uninterested females.

There were 25 male flies in each group, for a total sample size of 125.

Fruit fly. . .

Let us take a look at the data:

```
> Fruitfly.df = read.csv("Data/Fruitfly.csv", header=T)
> Fruitfly.df$group=factor(Fruitfly.df$group)
> boxplot(days ~ group, data = Fruitfly.df, ylab = "Longevity (days)")
```



It looks like male fruit flies do not live as long when in the presence of ‘uninterested’ females (`G5`), especially when there are several of them.

Fruit fly. . .

Linear model with multi-level (> 2) explanatory factor

As seen in previous chapters that involved categorical explanatory variables, our model specification uses indicator variables. In this case:

$$\texttt{days} = \beta_0 + \beta_1 \times \texttt{D2} + \beta_2 \times \texttt{D3} + \beta_3 \times \texttt{D4} + \beta_4 \times \texttt{D5} + \epsilon$$

where, as usual $\epsilon \overset{iid}{\sim} N(0, \sigma^2)$, and

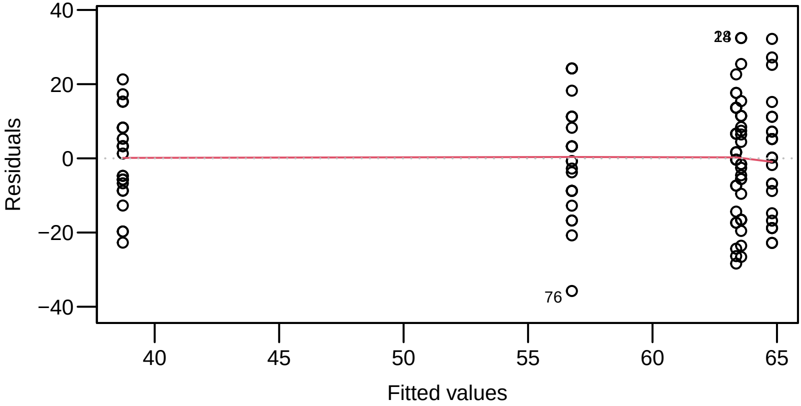
- D2 is an indicator variable whereby D2=1 if the fruit fly is in group 2, otherwise it is 0.
- D3 is an indicator variable whereby D3=1 if the fruit fly is in group 3, otherwise it is 0.
- ... and so on.

For example, β_1 and β_2 represent the differences in expected longevity (days) when we compare groups 2 and 3 to group 1 (the baseline).

Fruit fly. . .

Assumption checks

```
> Fruitfly.fit = lm(days ~ group, data = Fruitfly.df)
> plot(Fruitfly.fit, which=1)
```

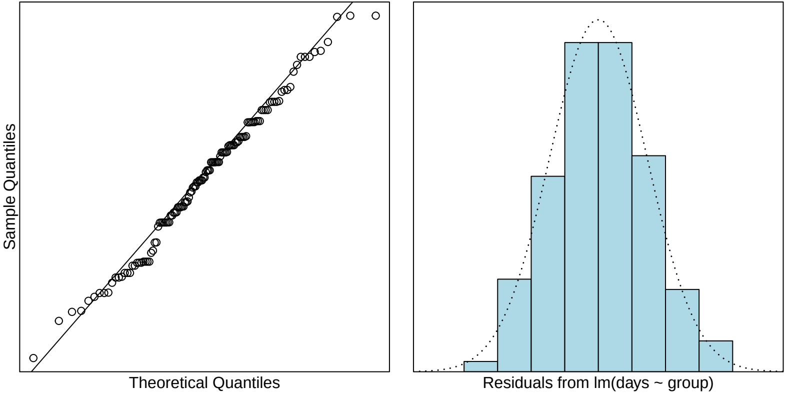


The **EOV** assumption seem to be okay.

Fruit fly. . .

Assumption checks. . .

```
> normcheck(Fruitfly.fit)
```

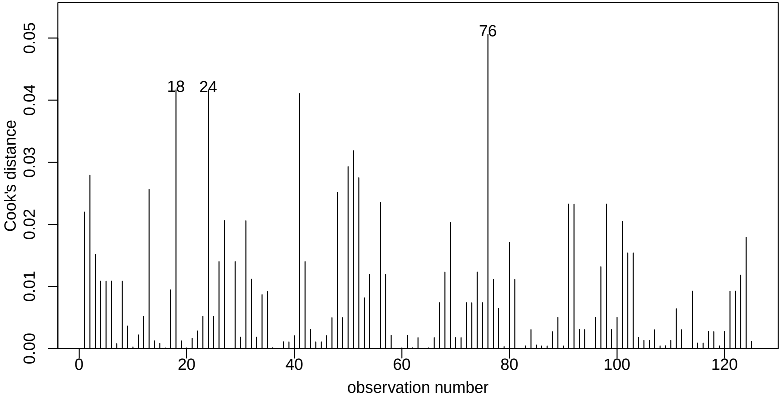


The normality assumption seem to be okay.

Fruit fly. . .

Assumption checks. . .

```
> cooks20x(Fruitfly.fit)
```



No unduly influential data points.

Fruit fly. . .

R^2 and ANOVA table

We can trust the fitted model. What can we conclude?[5]

```
> anova(Fruitfly.fit)
Analysis of Variance Table

Response: days
           Df Sum Sq Mean Sq F value    Pr(>F)    
group       4  11939 2984.82  13.612 3.516e-09 ***
Residuals 120  26314  219.28                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This allows us to say that there is very strong evidence of a difference in expected longevity between the five groups, which was fairly obvious from the boxplot.

A significant result means we should now investigate how the groups differ from one another – there is more work to be done.

[5] Recall from Chapter 9 that we have to use the anova function to check the significance of a factor variable with more than two levels.

Section 11.2
Interpreting the output

Fruit fly. . .

Interpretation

Now we know that the variable `group` helps explain longevity, what can we say about these groups? Let us investigate.

```
> summary(Fruitfly.fit)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   63.560      2.962  21.461  < 2e-16 ***
groupG2        1.240      4.188   0.296    0.768
groupG3       -0.200      4.188  -0.048    0.962
groupG4       -6.800      4.188  -1.624    0.107
groupG5      -24.840      4.188  -5.931 2.98e-08 ***
---
Residual standard error: 14.81 on 120 degrees of freedom
Multiple R-squared:  0.3121,Adjusted R-squared:  0.2892
F-statistic: 13.61 on 4 and 120 DF,  p-value: 3.516e-09
```

Our fitted model only gives the pairwise difference between the baseline group `G1` and the other four groups, and so is only providing partial information. What else can we do?

Fruit fly. . .

Interpretation of grand and group means

Some researchers like to examine the group means and their deviations from the overall (or so-called "grand") mean.[6] These deviations are commonly called group "effects".

The estimated grand mean is simply the sample mean over all 125 male flies:

```
> grand.mean=mean(Fruitfly.df$days)
> grand.mean
[1] 57.44
```

The estimated group means are just the sample means within each group. We can quickly obtain these using the incredibly useful[7] dplyr package:

```
> library(dplyr())
> Df=Fruitfly.df %>% group_by(group) %>% summarize(group.mean=mean(days)) %>%
+   data.frame()
```

[6] See the optional final Section of this Chapter for more on this topic.

[7] dplyr and its associated packages are widely used for "data wrangling" (the process of cleaning and re-arranging data sets for easy access and analysis).

Fruit fly. . .

Interpretation of grand and group means

The above code groups the data by `group` and then applies the `mean` function to the `days` values within each group.

The estimated group means are

```
> Df$group.mean
[1] 63.56 64.80 63.36 56.76 38.72
```

and the estimated group effects are therefore

```
> Df$group.mean-grand.mean
[1]   6.12   7.36   5.92  -0.68 -18.72
```

We have seen that the overall average longevity of the 125 male flies in the study is about 57.4 days.

We also see that group `G5` has markedly lower longevity (18.72 fewer days) compared to the overall mean.

We could test null hypotheses and calculate confidence intervals for the above conclusions, but our focus on this course will be making inference about the differences in group means.

Fruit fly. . .

Pairwise comparisons

We'd really like to get the pairwise comparisons between every possible pair of groups. However, we've seen that the fitted model is restricted to examining how the groups `G2-G5` differ from the baseline group `G1`.

If we wish to see how the other groups differed from group `G2`, say, then we could achieve this by changing the baseline group to group `G2`. Recall that this can be done using the `relevel` function:

```
> Fruitfly.df$newgroup = relevel(Fruitfly.df$group, ref="G2")
```

But to get all pairwise comparisons (i.e., `G3 vs G4`, `G4 vs G5`, `...`) we have to do this re-leveling for `G2`, `G3` and `G4`, and refit the model each time. This is too tedious.

We can get `R` to do the 'heavy lifting' for us by using the `emmeans` function from the `R` package of the same name. Moreover, `emmeans` solves the multiple comparisons problem that is discussed below.

Section 11.3
The multiple comparisons problem

Fruit fly. . .

Multiple comparisons

Note that when we are looking at all pair-wise comparisons of 5 groups, we have a total of 10 different possibilities:

`G1 vs G2`, `G1 vs G3`, `G1 vs G4`, `G1 vs G5`, (4 comparisons)
`G2 vs G3`, `G2 vs G4`, `G2 vs G5`, (3 comparisons)
`G3 vs G4`, `G3 vs G5`, (2 comparisons)
`G4 vs G5`, (1 comparisons).

In general, if there are m groups then there are ${}^{m}C_{2}$ possible pairwise comparisons.[8]

Each comparison requires a hypothesis test for a significant difference and an accompanying confidence interval. The multiple comparisons problem arises because, of all null hypotheses that are true, 5% are falsely rejected (Type 1 error). Equivalently, of all 95% confidence intervals, 5% of them do not contain the true parameter value.

[8] In `R` this is given by `choose(m,2)` and is the number of ways of choosing 2 objects from m objects. E.g., `choose(5,2)=10`.

Erroneous evidence of an effect from multiple testing

Multiple comparisons...

The following R code fits a simple linear regression model to iid (independent and identically distributed) normal data.

NOTE: The null hypothesis H_0 : slope $= 0$ is **true**.

```
> x = 1:30 ## Our explanatory variable
> x
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
[26] 26 27 28 29 30
> y = rnorm(30) ## y has NO relationship with x
> summary(lm(y˜x))$coef ## Print only the coefficient table
              Estimate Std. Error   t value   Pr(>|t|)
(Intercept)  0.4945317  0.3884378  1.273130 0.21344114
x           -0.0413633  0.0218802 -1.890444 0.06908972
```

If this code is run many times over, then approximately 5% of the time the slope will have P-value < 0.05.[9]

That is, there will be erroneous evidence of an effect of x (i.e., evidence for a non-zero slope) about 1 time in 20!

[9] In fact, it can be shown that the P-value is uniformly distributed between 0 and 1 when H_0 is true.

Erroneous evidence of an effect from multiple testing. . .

Multiple comparisons. . .

When we do multiple tests (i.e., the 10 paired comparisons in this example) then we greatly increase the probability of obtaining at least one erroneous conclusion[10].

This is known as the multiple comparison problem. It essentially says that if you look at enough things you will find something ‘happening’, even when there’s nothing going on.

Remember, data always have variability, and if we are not careful we can ‘discover’ false structure that is not really there.

So, when we look at these 10 comparisons we need to adjust so that the overall error rate (the probability of any spurious significance) over all 10 comparison is no more the 5%. This can be done using a Tukey adjustment.

[10] Assuming independent comparisons, if we do 10 95% CIs we have an overall error rate of $1 - (1 - .05)^{10} = 40\%$, which is much higher than our original 5% error rate per comparison.

Example—Fruit fly

Tukey simultaneous confidence intervals

Let's get *simultaneous* 95% confidence intervals for all 10 comparisons via the pairs and emmeans functions of the emmeans package.[11]

```
> library(emmeans)
> Fruitfly.pairs = pairs(emmeans(Fruitfly.fit, ~group, infer=T))
> Fruitfly.pairs
 contrast estimate   SE  df t.ratio p.value
 G1 - G2     -1.24 4.19 120  -0.296  0.9983
 G1 - G3      0.20 4.19 120   0.048  1.0000
 G1 - G4      6.80 4.19 120   1.624  0.4854
 G1 - G5     24.84 4.19 120   5.931  <.0001
 G2 - G3      1.44 4.19 120   0.344  0.9970
 G2 - G4      8.04 4.19 120   1.920  0.3127
 G2 - G5     26.08 4.19 120   6.227  <.0001
 G3 - G4      6.60 4.19 120   1.576  0.5158
 G3 - G5     24.64 4.19 120   5.883  <.0001
 G4 - G5     18.04 4.19 120   4.307  0.0003

P value adjustment: tukey method for comparing a family of 5 estimates
```

[11] These confidence intervals are called "simultaneous" since we can be 95% confident that **they all** contain the true group difference simultaneously.

Fruit fly

Tukey simultaneous confidence intervals...

We see that the majority of these pairwise comparisons are not significantly different. Let's extract only the CIs where the Tukey adjusted *P*-values are less than 0.05.

```
>   Fruitfly.pairs=data.frame(Fruitfly.pairs)
>   ## Which pairwise comparisons have a P-value less than 0.05?
>   mc.signif = Fruitfly.pairs[,"p.value"] < 0.05
>   mc.signif
 [1] FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE
>   ## Print comparisons which have a P-value less than 0.05
>   print(Fruitfly.pairs[mc.signif, ], digits = 4)
   contrast estimate    SE  df t.ratio   p.value
4   G1 - G5    24.84 4.188 120   5.931 2.958e-07
7   G2 - G5    26.08 4.188 120   6.227 7.232e-08
9   G3 - G5    24.64 4.188 120   5.883 3.701e-07
10  G4 - G5    18.04 4.188 120   4.307 3.240e-04
```

Note the use of the `data.frame` function in the above code. We needed to convert `Fruitfly.pairs` to a dataframe before we could take the subset of rows, otherwise `emmeans` gets confused and thinks we are doing a smaller number of pairwise comparisons.

Fruit fly. . .

Some conclusions:

- Our model explains 31% of variability in fruit fly longevity.
- We see that group 5 (males with 8 uninterested females) is different from all the others.

On average, group 5 males live fewer days than:

- Group 1 (males living alone) by 13 to 36 fewer days.
- Group 2 (males living with one interested female) by 14 to 38 fewer days.
- Group 3 (males living with eight interested females) by 13 to 36 fewer days.
- Group 4 (males living with one uninterested female) by 6 to 30 fewer days.

Fruit fly. . .

On a lighter note there is little evidence of a difference in longevity if no females or only one uninterested female is about, or if females are there and 'interested' in them — but in the presence of multiple uninterested females they die earlier (they 'drop like flies').

Recall also that in the other studies it was seen that females did not live as long if they reproduced, which can be attributed to the physical demands of producing and laying eggs. With males, perhaps it is sexual frustration that is killing them!

For more on this topic see the research article written by Branco et al. (2017, Reproductive activity triggers accelerated male mortality and decreases lifespan: genetic and gene expression determinants in Drosophila. Heredity 118, 221-228 https://doi.org/10.1038/hdy.2016.89) at `https://www.nature.com/articles/hdy201689`.

Section 11.4
Closing remarks and relevant R-code

Understanding the anova function output

```
> anova(Fruitfly.fit)
Analysis of Variance Table

Response: days
           Df Sum Sq Mean Sq F value    Pr(>F)
group       4  11939 2984.82  13.612 3.516e-09 ***
Residuals 120  26314  219.28
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the above output we see that the variability we observe in our longevity data can be broken down into two components `group` and `residual`.

The amount of variability that the variable `group` (as shown in the `Sum Sq` column) explains is 11939. The residual variability (left over) is 26314. The total variability is $11939 + 26314 = 38253$. The % of variability explained by `group` is therefore

$$100 \times \left(\frac{11939}{11939 + 26314} \right) = 100 \times \left(1 - \frac{26314}{11939 + 26314} \right) = 31\%.$$

Note that we have just calculated the R^2 – the proportion of the variability in the response variable that is explained by the explanatory variables, 0.31.

Most of the R-code you need for this chapter

Use box plots to inspect the data for each level of the factor.

```
> boxplot(days ~ group, data = Fruitfly.df)
```

You do not need to create indicator variables - R does that for you. The baseline can be changed if you wish by using the `relevel` function.

```
> Fruitfly.df$newgroup = relevel(Fruitfly.df$group, ref="G2")
```

Fit the model and use the ANOVA table to see if any of the means differ from one another (regardless of the baseline chosen).

```
> anova(Fruitfly.fit)
```

Adjust confidence intervals for multiple pairwise comparisons by using the Tukey adjustment to obtain simultaneous intervals CIs:

```
> Fruitfly.pairs = pairs(emmeans(Fruitfly.fit, ~group, infer=T))
```

Section 11.5
Alternative parameterizations of the 1-way ANOVA model

(This is an optional Section:
- your lecturer will advise whether it is examinable)

Alternative parameterizations of the 1-way ANOVA model

The reference cell model

Recall the linear model[12] we used to represent the longevity, in days, of a male fruitfly, i.e.

$$\texttt{days} = \beta_0 + \beta_1 \times \texttt{D2} + \beta_2 \times \texttt{D3} + \beta_3 \times \texttt{D4} + \beta_4 \times \texttt{D5} + \epsilon$$

The *parameters* $\beta_0, \beta_1, \ldots, \beta_4$ denote the true values of some attribute (e.g. longevity) of the population of male fruitflies. Here, β_0 represents the mean longevity of male fruitflies in group G1. The parameters $\beta_1, \ldots, \beta_4$ represent the deviations in mean longevity of males in groups G2,...,G5, respectively, from the mean longevity of males in group G1.

The values in the Estimate column of the regression summary table[13] result in the following equation for predicted longevity:

$$\widehat{\texttt{days}} = 63.56 + 1.24 \times \texttt{D2} + (-0.20) \times \texttt{D3} + (-6.80) \times \texttt{D4} + (-24.84) \times \texttt{D5}$$

[12] See slide 8.

[13] See slide 14; Coefficients rounded to 2 decimal places.

Alternative parameterizations of the linear model

The reference cell model

Each cell within a column in the table below corresponds to a level of the Group factor. One way to 'parametrise' these cells is to use means, i.e. $\mu_1, \mu_2, \ldots, \mu_5$. Another is to select one of the cells as a reference cell (here Group G1) and the remaining cells are then parametrised the deviations of the current row's group mean from the reference cell's group mean.

		parameterization			
Group	Data	Means	Estimate[14]	Reference cell	Estimate[15]
G1	$40, 37, \ldots, 44$	μ_1	63.56	$\beta_0 = \mu_1$	63.56
G2	$46, 42, \ldots, 92$	μ_2	64.80	$\beta_1 = \mu_2 - \mu_1$	1.24
G3	$35, 37, \ldots, 77$	μ_3	63.36	$\beta_2 = \mu_3 - \mu_1$	-0.20
G4	$21, 40, \ldots, 68$	μ_4	56.76	$\beta_3 = \mu_4 - \mu_1$	-6.80
G5	$16, 19, \ldots, 44$	μ_5	38.72	$\beta_4 = \mu_5 - \mu_1$	-24.84

The parameterization of the model shown on the previous slide is therefore known as the *reference cell* model.

[14] See estimates of Group means on slide 16

[15] See regression coefficients table on slide14

Alternative parameterizations of the linear model

The means model

From the above table we can see that there is an alternative, but equivalent, *means* model parameterization, i.e. linear model for the longevity of the jth ($j = 1, 2, \ldots, 25$) male fruitfly in `Group` i ($i = 1, 2, \ldots, 5$) may be written as

$$days_{ij} = \mu_i + \epsilon_{ij}$$

where μ_i denotes the mean longevity, in days, of a male in `Group` i and, as usual, $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.

Alternative parameterizations of the linear model

The effects model

Another parameterization is to set the overall mean longevity, μ, as the reference and then define the *effect*, τ_i, on longevity due to being in Group i as the difference between the Group i mean and the overall mean, i.e. $\tau_i = \mu_i - \mu$.

		parameterization			
Group	Data	Means	Estimate	Effects	Estimate[16]
G1	$40, 37, \ldots, 44$	μ_1	63.56	$\tau_1 = \mu_1 - \mu$	6.12
G2	$46, 42, \ldots, 92$	μ_2	64.80	$\tau_2 = \mu_2 - \mu$	7.36
G3	$35, 37, \ldots, 77$	μ_3	63.36	$\tau_3 = \mu_3 - \mu$	5.92
G4	$21, 40, \ldots, 68$	μ_4	56.76	$\tau_4 = \mu_4 - \mu$	-0.68
G5	$16, 19, \ldots, 44$	μ_5	38.72	$\tau_5 = \mu_5 - \mu$	-18.72

The linear *effects* model for the longevity of the jth ($j = 1, 2, \ldots, 25$) male fruitfly in Group i ($i = 1, 2, \ldots, 5$) may therefore be written as

$$days_{ij} = \mu + \tau_i + \epsilon_{ij}$$

where, again, $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.

[16] See overall mean (57.44 days) and deviations of group means from overall means on slide 16.