Chapter 11: Linear models with a factor variable with three or more levels

STATS 201/8

University of Auckland

Learning Outcomes

In this chapter you will learn about:

- Explanatory factor with multiple levels—One-way analysis of variance
- The multiple comparisons problem
- Relevant R-code.

Explanatory factor with multiple levels (One-way analysis of variance)

Example—Fruit fly

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

Data from http://www.cvgs.k12.va.us:81/digstats/Imain.html.

Fruit fly

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.

The hypothesis was that the males living alone or with uninterested females would live longer than the males living with the interested females. Since there are more than two group means to compare[1] an adjustment to how we interpret our model is used to determine if there is a significant difference between these group means.

How does one define "interest" in fruit-flies?

Here is this study's definition of the question above:

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

[1] If there were only two groups we could use a two sample two-sample *t*-test discussed in chapter 5

Fruit fly. . .

The variable measured was:

`days` the number of days the male fly lived

The variables that were controlled were:

`group` the group they were allocated to where:

`G1` are males living alone,

`G2` are males living with one interested female,

`G3` are males living with eight interested females,

`G4` are males living with one uninterested female, and

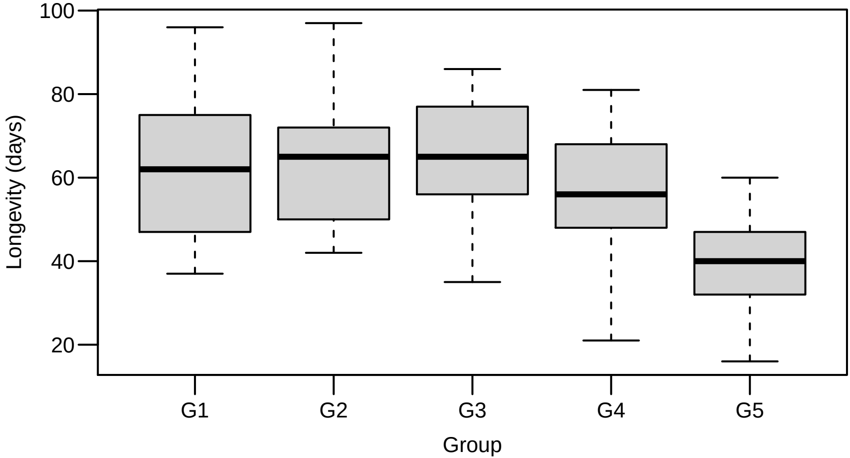
`G5` are males living with eight uninterested females.

There were 25 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)
> boxplot(days ~ group, data = Fruitfly.df)
```



Fruit fly. . .

It seems male fruit flies do not live as long when they are in the presence of ‘uninterested’ females (G5). A result we were not expecting.

For females reproduction came at a cost (shorter lifespan), whereas for males, a lack of reproduction seems to cost them. Let us see if this effect is ‘real’ (or not).

Fruit fly...

Explanatory factor with many levels (> 2)

If we believe this to be true then our underlying model is:

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

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

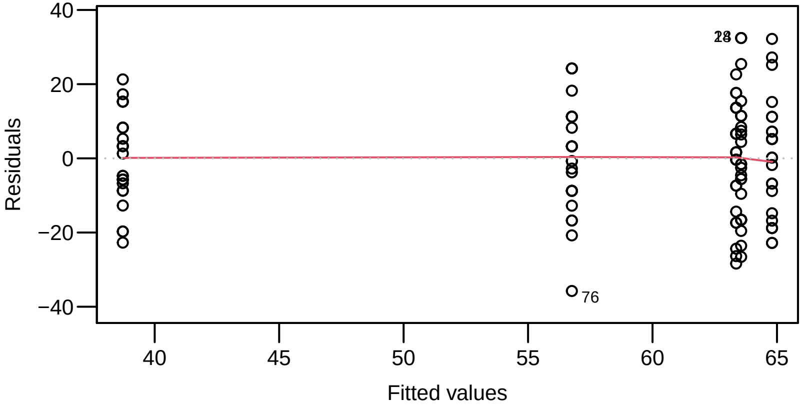
- D2 is a dummy variable whereby: D2=1 if the fruit fly is in group 2—otherwise it is 0.
- D3 is a dummy variable whereby: D3=1 if the fruit fly is in group 3—otherwise it is 0.
- ... And so on.

For example, β_1 and β_2 represent the differences in average 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)
```

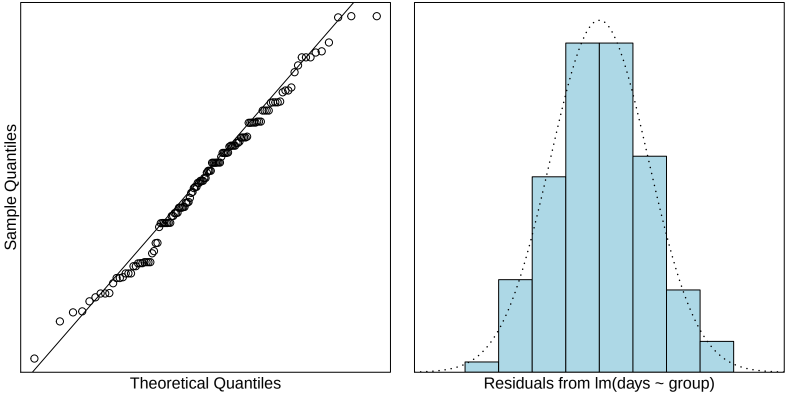


The **EOV** assumption seem to be okay.

Fruit fly. . .

Assumption checks. . .

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

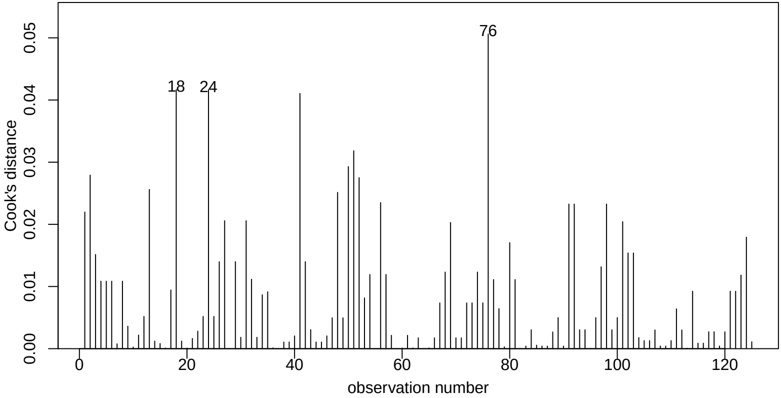


The normality assumption seem to be okay.

Fruit fly. . .

Assumption checks. . .

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



No unduly influential data points.

Fruit fly. . .

R^2 and ANOVA table

We can trust this output. What is it telling us?

```
> 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 simply allows us to say that there is a clear difference in expected longevity between the five groups, which was fairly obvious from the first plot. The next step is to see where the differences lie.

Note that a significant result means you should investigate which groups are different from one another; there is more work to be done.

Now we can investigate whether female ‘lack of interest’ is ‘killing’ these wretched male fruit flies.

Fruit fly. . .

R^2 and ANOVA table. . .

In the above output we see that the variability we observe for all our 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

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

This is another way we can think of the % of variation explained or R^2.

It is the % of the data that is not residual or ‘noise’.

Fruit fly. . .

One-way analysis of variance

Now we know that the variable `group` helps explain longevity, what can we say about these groups? Let us investigate. The `s20x` function

`summary1way` provides the group means and their deviations from the overall (or 'grand') mean.

Fruit fly. . .

One-way analysis of variance. . .

```
> summary1way(Fruitfly.fit, draw.plot = FALSE)
ANOVA Table:
                Df  Sum Squares  Mean Square  F-statistic  p-value
Between Groups  4   11939.28     2984.82      13.61195     0
Within Groups   120 26313.52     219.27933
Total           124 38252.8

Numeric Summary:
          Sample size     Mean  Median  Std Dev  Midspread
All Data          125    57.44      58 17.56389         24
G1                 25    63.56      62 16.45215         28
G2                 25    64.80      65 15.65248         22
G3                 25    63.36      65 14.53983         21
G4                 25    56.76      56 14.92838         20
G5                 25    38.72      40 12.10207         15

Table of Effects: (GrandMean and deviations from GM)
typ.val      G1      G2      G3      G4     G5
  57.44    6.12    7.36    5.92   -0.68 -18.72
```

Fruit fly. . .

Interpreting the output

We see from above 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 (18.75 fewer days) compared to the overall mean.

Note that if `group` does not explain any true underlying variation in longevity, then we expect all these group means to differ at most only moderately from the overall mean. This can be hard to judge informally, since we have to take into account the standard error of each group mean and how many groups there are.

That is why we have to rely on the *P*-value from the `anova` table.

Fruit fly. . .

Interpreting the output. . .

It is natural to ask which of the groups are different.

Here is our familiar `summary` output:

```
> 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
```

Here we see that we have evidence to believe that β_4 the parameter for group 5 is different from from zero.

We estimate that males with 8 uninterested females die, on average, 25 days earlier than males who are by themselves (our baseline group is `G1`).

Fruit fly. . .

Interpreting the output. . .

In the output above we are restricted to seeing how each of the groups, `G2-G5`, differs from the baseline group `G1`.

If we wish to see how the other groups differed from group `G4`, for example, then we could achieve this by changing the baseline group to group `G4` by reordering the levels of the group factor variable.

This is very tedious, but here is how we make `G4` the baseline level in `R`:

```
> Fruitfly.df$newgroup = factor(Fruitfly.df$group,
+                               levels = c("G4", "G1", "G2", "G3", "G5"))
```

However, what if we wish to look at all pair-wise comparisons (i.e., `G1 vs G2`, `G2 vs G3`, `...`)? Do we really have to do this re-ordering a bunch of times in order to find these out?

The answer is no: We can get `R` to do this 'heavy lifting' for us.

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).

The number 10 comes from $4 + 3 + 2 + 1 = 10$ or in fact ${}^{5}C_2 = 10$ ways of choosing 2 objects from 5 (in no particular order).

Here we are asking 10 questions (comparisons) about our data, as we are looking to test 10 null hypotheses. Of all null hypotheses that are true, 5% are falsely rejected. Of all 95% confidence intervals, 5% of them do not contain the true parameter value.

The multiple comparisons problem

Erroneous evidence of an effect from multiple testing

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) -0.107068614 0.35454802 -0.3019862 0.7648943
x           -0.004679368 0.01997124 -0.2343054 0.8164528
```

If this code is run many times over, then approximately 5% of the time the slope will have P-value < 0.05.[2]

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!

[2] 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. . .

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[3].

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.

[3] 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 `s20x` function `multipleComp`.

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

```
> multipleComp(Fruitfly.fit)
          Estimate  Tukey.L Tukey.U Tukey.p
G1  -  G2    -1.24 -12.8405 10.3605  0.9983
G1  -  G3     0.20 -11.4005 11.8005  1.0000
G1  -  G4     6.80  -4.8005 18.4005  0.4855
G1  -  G5    24.84  13.2395 36.4405  0.0000
G2  -  G3     1.44 -10.1605 13.0405  0.9970
G2  -  G4     8.04  -3.5605 19.6405  0.3127
G2  -  G5    26.08  14.4795 37.6805  0.0000
G3  -  G4     6.60  -5.0005 18.2005  0.5158
G3  -  G5    24.64  13.0395 36.2405  0.0000
G4  -  G5    18.04   6.4395 29.6405  0.0003
```

Here we see that most of these comparisons are not significantly different.

Fruit fly

Tukey simultaneous confidence intervals. . .

Lets extract the CIs where the Tukey adjusted *P*-value are less than 0.05.

```
> mc.fruitfly = multipleComp(Fruitfly.fit)
> ## Which entries have a P-value less than 0.05?
> mc.fruitfly[, "Tukey.p"] < 0.05
G1  -  G2 G1  -  G3 G1  -  G4 G1  -  G5 G2  -  G3 G2  -  G4 G2  -  G5 G3  -  G4
    FALSE     FALSE     FALSE      TRUE     FALSE     FALSE      TRUE     FALSE
G3  -  G5 G4  -  G5
     TRUE      TRUE
> ## Print entries which have a P-value less than 0.05
> mc.fruitfly[mc.fruitfly[, "Tukey.p"] < 0.05, ]
          Estimate Tukey.L Tukey.U Tukey.p
G1  -  G5    24.84 13.2395 36.4405   0e+00
G2  -  G5    26.08 14.4795 37.6805   0e+00
G3  -  G5    24.64 13.0395 36.2405   0e+00
G4  -  G5    18.04  6.4395 29.6405   3e-04
```

Fruit fly. . .

Some conclusions:

- Our model explains 31% of variability in fruit fly longevity.
- We see that the effect of group 5 (males with 8 uninterested females) seems 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 no more than 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 original studies it was seen that females did not live as long if they reproduced.

It is tempting to make similar inference about the human species but that may be going too far!

Relevant R-code

Most of the R-code you need for this chapter

Note that this code comes with the usual code/checks discussed in chapters 1 and 2.

You do not need to create dummy variables - R does that for you. The baseline can be changed if you wish rather than having R choose it for you — see relevant R-code from chapter 9.

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

```
> ## Create the pairs plot of the five numeric variables
> boxplot(days ~ group, data = Fruitfly.df)
```

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)
```

In order to see measure pair-wise differences between mean levels to adjust for multiple comparisons:

```
> multipleComp(Fruitfly.fit)
```