data_exploration

Lissy Denkers

1/4/2022

Data import

Background information

This is the data of the whitefly development bioassay on grafts of MM and LA1840. The nymphs in all developmental stages (first to fourth instar) were counted every other day. After the fourth instar stage, whiteflies develop into adults, leaving behind the larval skin called exuviae. Nymphs in the last phase of fourth instar stage and exuviae were removed from the leaf after each count to prevent a whitefly outbreak in the greanhouse.

See metadata file for more information.

The raw data

```
flies <- read.csv("raw_data.csv",
                          header = T,
                          sep = ",",
                          dec = ".",
                          check.names = FALSE)
# The first 10 rows of the data
knitr::kable(flies[1:10,1:8])
```

genotype	place	date	day	stage	number	eggs_start	hatched
LA1840	1	26/11/2021	2	0_egg	60	60	40
LA1840	1	26/11/2021	2	1_first_instar	0	60	40
LA1840	1	26/11/2021	2	2_second_instar	0	60	40
LA1840	1	26/11/2021	2	3_third_instar	0	60	40
LA1840	1	26/11/2021	2	4_early_fourth_instar	0	60	40
LA1840	1	26/11/2021	2	5_late_fourth_instar	0	60	40
LA1840	1	26/11/2021	2	6_exuviea	0	60	40
MM	2	26/11/2021	2	0_egg	66	66	46
MM	2	26/11/2021	2	1_first_instar	0	66	46
MM	2	26/11/2021	2	2_second_instar	0	66	46

Preparing the data for analysis

A few improvements to the data should be made before analysis.

First, three plants must be excluded from analysis, based on observations during counting (see metadata file).

```
flies <- flies %>%
  dplyr::filter(place != 4) %>%
  dplyr::filter(place < 13)
```

Next, I make data wide for access to separate life stages.

```
flies <- flies %>%
  mutate(stage = as.factor(stage)) %>%
  pivot_wider(names_from = stage, values_from = number)
```

Because the fourth instars are the end point and taken of after counting, combine fourth instars and exuviae and make their numbers cumulative over time.

```
flies <- flies %>%
  mutate(fourth_exuviae = `5_late_fourth_instar` + `6_exuviea`) %>%
  group_by(place) %>%
  mutate(`5_completed_lifecycle` = cumsum(fourth_exuviae)) %>%
  mutate(`6_total_fourth_instars` = `4_early_fourth_instar` + `5_completed_lifecycle`) %>%
  ungroup()
```

Remove separate and non-cumulative late fourth instar and exuviae columns.

```
flies <- flies %>%
  dplyr::select(-`5_late_fourth_instar`, -`6_exuviea`, -fourth_exuviae)
```

Now the data can be made long again.

```
flies <- flies %>%
  pivot_longer(cols = `0_egg`:`6_total_fourth_instars`,
               names_to = "stage", values_to = "number")
```

Add the relative number of nymphs as percentage from the number of eggs and as percentage from the number of hatched eggs

```
flies <- flies %>%
  mutate(percentage_from_eggs = number / eggs_start * 100) %>%
  mutate(percentage_from_hatched = number / hatched * 100)
```

Add a new dataframe for the total number of nymphs per day without percentage, based on “flies”

```
flies_wide <- flies  %>%
  dplyr::select(-percentage_from_eggs) %>%
  dplyr::select(-percentage_from_hatched)
```

and make the data wide.

```
flies_wide <- flies_wide %>%
  mutate(stage = as.factor(stage)) %>%
  pivot_wider(names_from = stage, values_from = number)
```

Add the total number of nymphs per sample per day and the relative total as percentage from the number of eggs and as percentage from the number of hatched eggs

```
flies_wide <- flies_wide %>%
  mutate(total = rowSums(flies_wide[,8:12], na.rm = TRUE)) %>%
  mutate(percentage_from_eggs = total / eggs_start * 100) %>%
  mutate(percentage_from_hatched = total / hatched * 100)
```

Different developmental stages, different effects

In the original dataset of Arjen with ungrafted plants, you can see different effects of LA1840 on the different whitefly lifestages. The number of eggs was higher on LA1840 than on MM, while the number of fourth instar nymphs was drasticly decreased on LA1840. While the number of eggs is probably the result of the effect the plant has on adult flies, the nymph development is more likely to be influenced by the diet of the nymphs themselves. Another effect the plants could have, is affecting the hatching of the eggs. This can however not be analysed from Arjens data.

Because I want to have a clear view of those different effects, I will go through each part separately.

Eggs

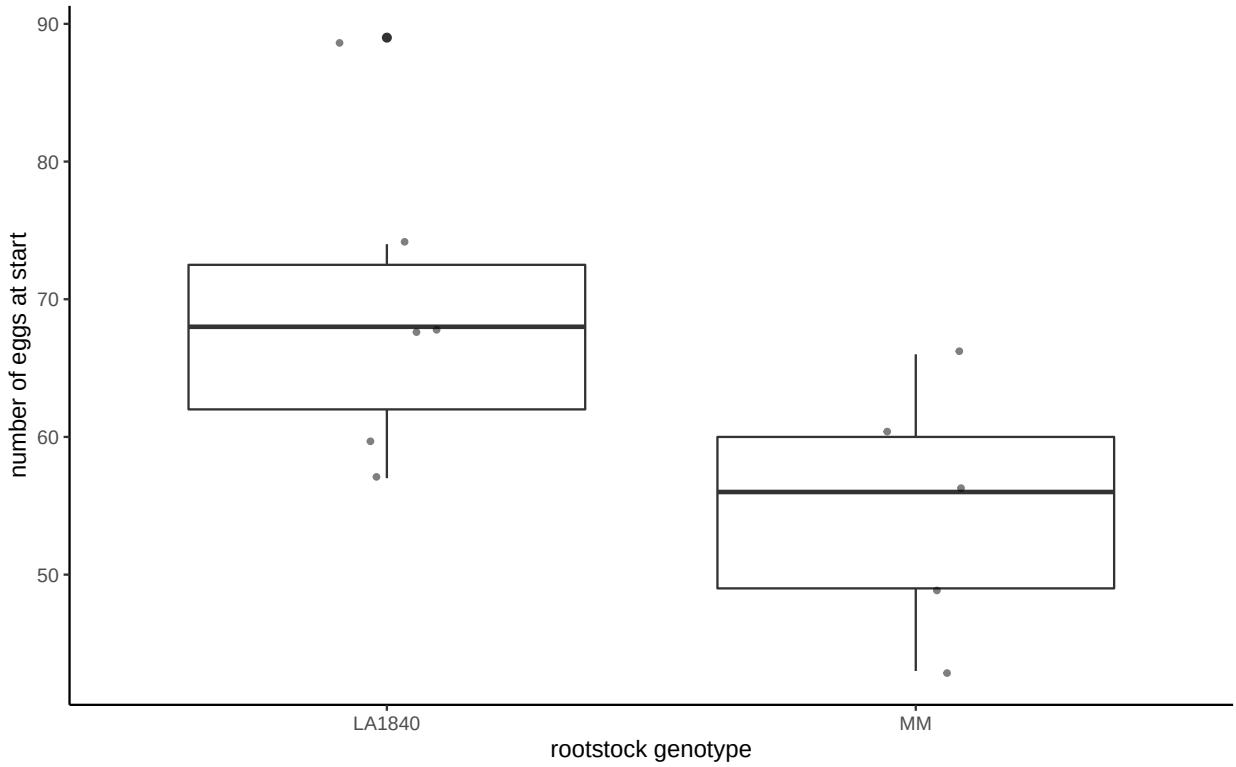
On normal LA1840 plants, the number of eggs is increased compared to MM plants. Can we see the same effect when an LA1840 rootstock is used with a MM scion?

First a quick overview of the data:

```
flies_wide %>%
  group_by(genotype) %>%
  get_summary_stats(`0_egg`, type = "mean_sd")
```

```
## # A tibble: 2 x 5
##   genotype variable     n  mean    sd
##   <chr>    <chr>    <dbl> <dbl> <dbl>
## 1 LA1840   0_egg        6  69.3 11.4
## 2 MM       0_egg        5  54.8  9.04
```

And now visualized in a boxplot:



It seems like the number of eggs might be higher on the grafts with an LA1840 rootstock.

Before testing this, check the assumptions: - normality - homogeneity of variance - no significant outliers

```
shapiro.test(resid(aov(flies_wide$`0_egg`~flies_wide$genotype)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(aov(flies_wide$`0_egg` ~ flies_wide$genotype))
## W = 0.95145, p-value = 0.6625
```

The p-value of the Shapiro-Wilk test is >0.05, so the data is normally distributed

```
leveneTest(`0_egg`~genotype, data = flies_wide)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.0447 0.8372
##        9
```

The Levene's test also has p>0.05, so the variance is equal between the two groups.

Now the outliers:

```
flies_wide %>%
  group_by(genotype) %>%
  identify_outliers(`0_egg`)
```

```
## # A tibble: 1 x 18
##   genotype place date         day eggs_start hatched `0_egg` `1_first_instar`
##   <chr>    <int> <chr>      <int>      <int>   <int>   <int>            <int>
## 1 LA1840       7 26/11/2021     2         89      57      89                0
## # ... with 10 more variables: 2_second_instar <int>, 3_third_instar <int>,
## #   4_early_fourth_instar <int>, 5_completed_lifecycle <int>,
## #   6_total_fourth_instars <int>, total <dbl>, percentage_from_eggs <dbl>,
## #   percentage_from_hatched <dbl>, is.outlier <lgl>, is.extreme <lgl>
```

There is one outlier, but it is not extreme.

Because all assumptions for a t-test are met, we can do a Students t-test.

```
flies_wide %>%
  t_test(`0_egg` ~ genotype, var.equal = TRUE) %>%
  add_significance()
```

```
## # A tibble: 1 x 9
##   .y.   group1 group2    n1    n2 statistic    df      p p.signif
## * <chr> <chr>  <chr>  <int> <int>     <dbl> <dbl>  <dbl> <chr>
## 1 0_egg LA1840 MM        78    65      2.30     9 0.0468 *
```

```
flies_wide %>%
  cohens_d(`0_egg` ~ genotype, var.equal = TRUE)
```

```
## # A tibble: 1 x 7
##   .y.   group1 group2 effsize    n1    n2 magnitude
## * <chr> <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 0_egg LA1840 MM        1.39    78    65 large
```

The difference in number of eggs on MM and LA1840 grafts is significant (p=0.0468) and the effectsize of the rootstock genotype on the number of eggs is large (d=1.39).

The d-value indicates the difference between the two means in number of standard deviations. So the mean of LA1840 grafts is 1.39 standard deviations higher than that of MM grafts.

Effect of rootstock genotype on the number of eggs

Similar to the previous findings on normal LA1840 and MM plants, the number of eggs are increased on grafts with an LA1840 rootstock compared to grafts with a MM rootstock. This indicates that the signal responsible for the increased oviposition is transported from the LA1840 rootstock to the MM scion.

Hatching

Is there an effect of rootstock genotype on the hatching of the eggs? For this I will use the absolute number of hatched eggs, as well as the percentage.

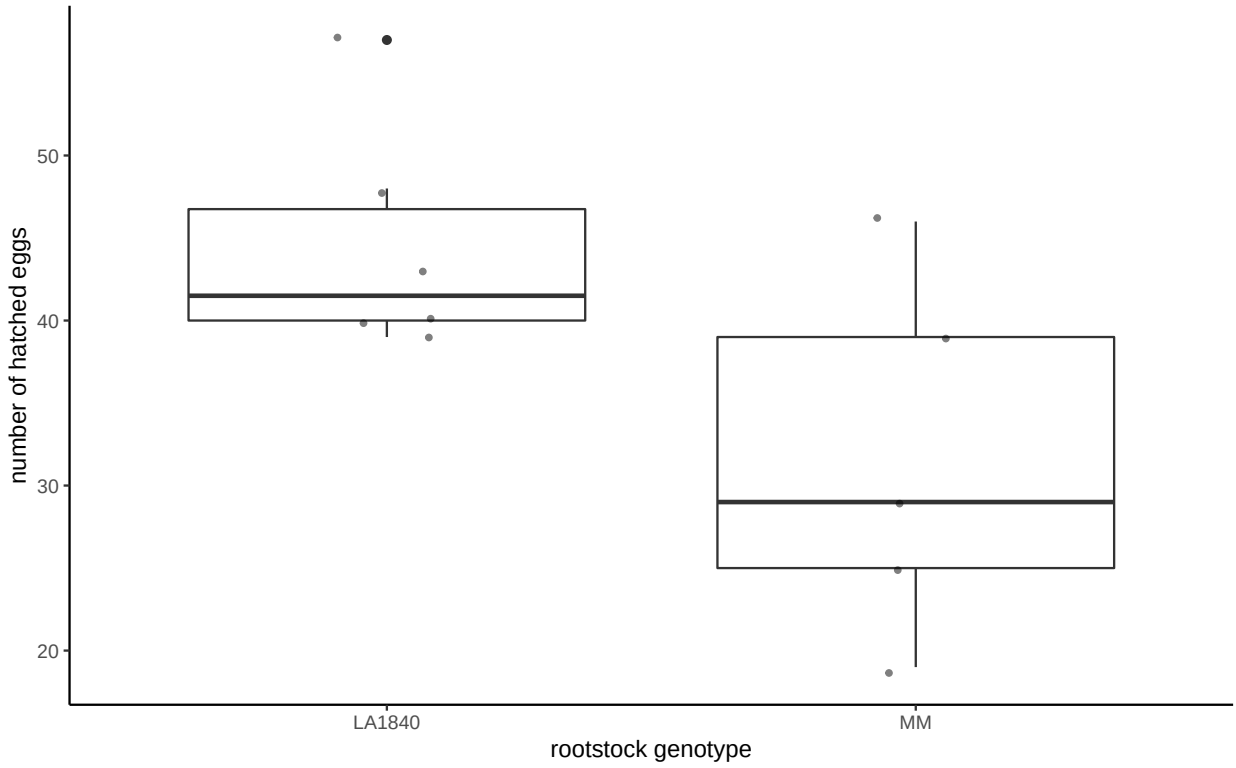
absolute number of hatched eggs

An overview of the data:

```
flies_wide %>%
  mutate(day = as.factor(day)) %>%
  group_by(genotype) %>%
  dplyr::filter(day=="2") %>%
  get_summary_stats(hatched, type = "mean_sd")
```

```
## # A tibble: 2 x 5
##   genotype variable     n  mean    sd
##   <chr>    <chr>    <dbl> <dbl> <dbl>
## 1 LA1840   hatched      6  44.5  6.95
## 2 MM       hatched      5  31.6 10.9
```

```
flies_wide %>%
  mutate(day = as.factor(day)) %>%
  dplyr::filter(day == "2") %>%
  ggplot(., aes(x = genotype, y = hatched)) +
  geom_boxplot()  +
  geom_jitter(size = 1, alpha = 0.5, width = 0.1) +
  labs(y="number of hatched eggs",
       x= "rootstock genotype") +
  theme_classic()
```



Shapiro-Wilk test:

```
flies_wide %>%
  mutate(day = as.factor(day)) %>%
```

```
  dplyr::filter(day == "2") %>%
  group_by(genotype) %>%
  shapiro_test(hatched)
```

```
## # A tibble: 2 x 4
##   genotype variable statistic      p
##   <chr>    <chr>        <dbl>  <dbl>
## 1 LA1840   hatched      0.821 0.0907
## 2 MM       hatched      0.965 0.843
```

Levene's test

```
flies_wide %>%
  mutate(day = as.factor(day)) %>%
  dplyr::filter(day == "2") %>%
  levene_test(hatched~genotype)
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     1     9     0.858 0.378
```

Identifying extreme outliers:

```
flies_wide %>%
  mutate(day = as.factor(day)) %>%
  dplyr::filter(day == "2") %>%
  group_by(genotype) %>%
  identify_outliers(hatched)
```

```
## # A tibble: 1 x 18
##   genotype place date       day   eggs_start hatched `0_egg` `1_first_instar`
##   <chr>    <int> <chr>      <fct>      <int>   <int>   <int>            <int>
## 1 LA1840       7 26/11/2021 2            89      57      89                0
## # ... with 10 more variables: 2_second_instar <int>, 3_third_instar <int>,
## #   4_early_fourth_instar <int>, 5_completed_lifecycle <int>,
## #   6_total_fourth_instars <int>, total <dbl>, percentage_from_eggs <dbl>,
## #   percentage_from_hatched <dbl>, is.outlier <lgl>, is.extreme <lgl>
```

The assumptions for a t-test are met.

```
flies_wide %>%
  mutate(day = as.factor(day)) %>%
  dplyr::filter(day == "2") %>%
  t_test(hatched ~ genotype, var.equal = TRUE) %>%
  add_significance()
```

```
## # A tibble: 1 x 9
##   .y.     group1 group2    n1    n2 statistic    df      p p.signif
## * <chr>   <chr>  <chr>  <int> <int>     <dbl> <dbl>  <dbl> <chr>
## 1 hatched LA1840 MM         6     5      2.39     9 0.0403 *
```

```
flies_wide %>%
  mutate(day = as.factor(day)) %>%
  dplyr::filter(day == "2") %>%
  cohens_d(hatched ~ genotype, var.equal = TRUE)
```

```
## # A tibble: 1 x 7
##   .y.     group1 group2 effsize    n1    n2 magnitude
## * <chr>   <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 hatched LA1840 MM        1.45     6     5 large
```

Again, there is a significant difference between LA1840 and MM grafts (p=0.04) with a large effectsize of the rootstock genotype (d=1.45). This could, however, be caused by the difference in number of eggs. Therefore, I will also check the success of hatching relative to the amount of eggs.

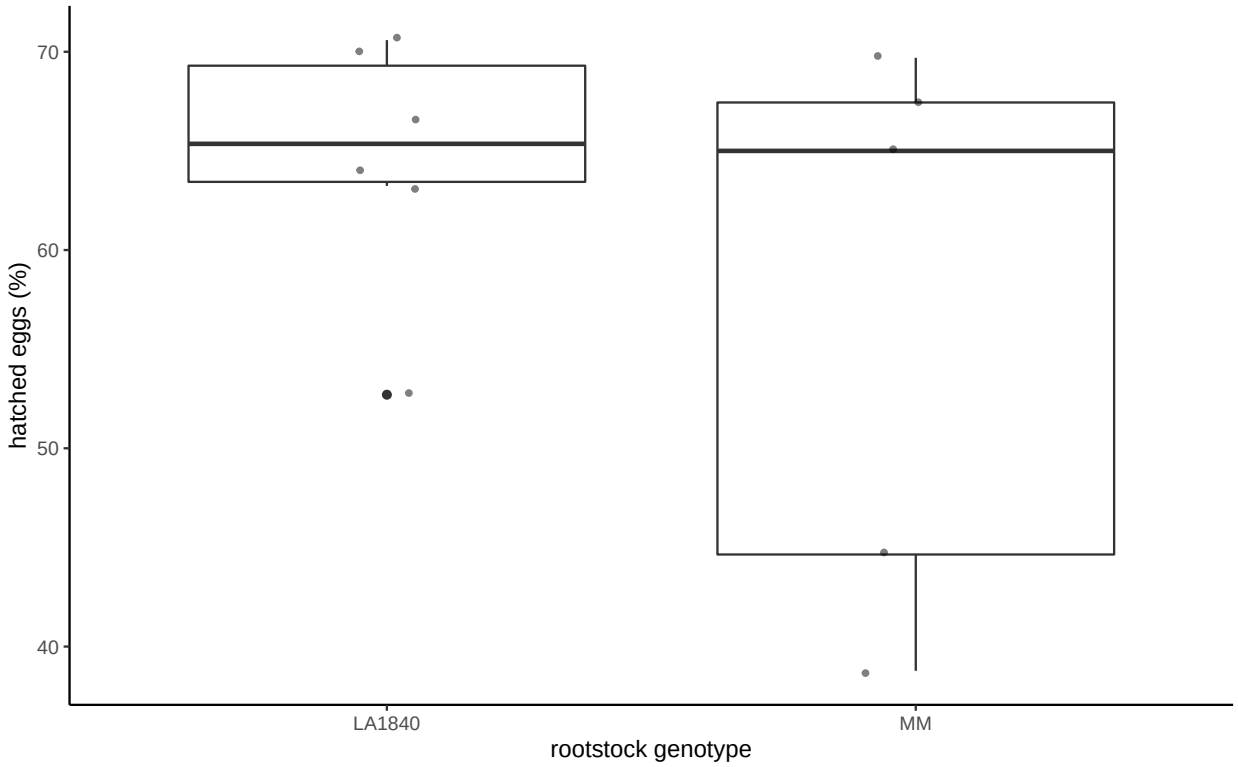
Percentage of hatching

An overview of the data:

```
flies_wide <- flies_wide %>%
  mutate(perc_hatched = hatched/eggs_start*100) %>%
  mutate(day = as.factor(day))

flies_wide %>%
  group_by(genotype) %>%
  dplyr::filter(day=="2") %>%
  get_summary_stats(perc_hatched, type = "mean_sd")
```

```
## # A tibble: 2 x 5
##   genotype variable         n  mean    sd
##   <chr>    <chr>        <dbl> <dbl> <dbl>
## 1 LA1840   perc_hatched     6  64.6  6.56
## 2 MM       perc_hatched     5  57.1 14.3
```

```
flies_wide %>%
  dplyr::filter(day == "2") %>%
  ggplot(., aes(x = genotype, y = perc_hatched)) +
  geom_boxplot()  +
  geom_jitter(size = 1, alpha = 0.5, width = 0.1) +
  labs(y="hatched eggs (%)",
       x= "rootstock genotype") +
  theme_classic()
```



Shapiro-Wilk test:

```
flies_wide %>%
  dplyr::filter(day == "2") %>%
  group_by(genotype) %>%
  shapiro_test(perc_hatched)
```

```
## # A tibble: 2 x 4
##   genotype variable     statistic     p
##   <chr>    <chr>            <dbl> <dbl>
## 1 LA1840   perc_hatched     0.871 0.229
## 2 MM       perc_hatched     0.829 0.136
```

Levene's test

```
flies_wide %>%
  dplyr::filter(day == "2") %>%
  levene_test(perc_hatched~genotype)
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     1     9      1.45 0.260
```

Identifying extreme outliers:

```
flies_wide %>%
  dplyr::filter(day == "2") %>%
  group_by(genotype) %>%
  identify_outliers(perc_hatched)
```

```
## # A tibble: 1 x 19
##   genotype place date       day   eggs_start hatched `0_egg` `1_first_instar`
##   <chr>    <int> <chr>      <fct>      <int>   <int>   <int>            <int>
## 1 LA1840       6 26/11/2021 2             74      39      74                0
## # ... with 11 more variables: 2_second_instar <int>, 3_third_instar <int>,
## #   4_early_fourth_instar <int>, 5_completed_lifecycle <int>,
## #   6_total_fourth_instars <int>, total <dbl>, percentage_from_eggs <dbl>,
## #   percentage_from_hatched <dbl>, perc_hatched <dbl>, is.outlier <lgl>,
## #   is.extreme <lgl>
```

The assumptions for a t-test are met.

```
flies_wide %>%
  dplyr::filter(day == "2") %>%
  t_test(perc_hatched ~ genotype, var.equal = TRUE) %>%
  add_significance()
```

```
## # A tibble: 1 x 9
##   .y.          group1 group2    n1    n2 statistic    df     p p.signif
## * <chr>        <chr>  <chr>  <int> <int>     <dbl> <dbl> <dbl> <chr>
## 1 perc_hatched LA1840 MM         6     5      1.15     9  0.28 ns
```

```
flies_wide %>%
  dplyr::filter(day == "2") %>%
  cohens_d(perc_hatched ~ genotype, var.equal = TRUE)
```

```
## # A tibble: 1 x 7
##   .y.          group1 group2 effsize    n1    n2 magnitude
## * <chr>        <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 perc_hatched LA1840 MM       0.696     6     5 moderate
```

There is no difference in the percentage of hatched eggs between the two rootstock genotypes. There is, however, a large variation in hatching between plants, as can be seen in the above boxplot.

```
flies_wide %>%
  dplyr::select(!(`1_first_instar`:perc_hatched)) %>%
  dplyr::filter(day == "2") %>%
  pivot_longer(cols = hatched:`0_egg`,
               names_to = "stage", values_to = "number") %>%
  mutate(place = as.factor(place)) %>%
  anova_test(number ~ stage*genotype + Error(place/stage)) %>%
  get_anova_table()
```

```
## ANOVA Table (type III tests)
##
```

```
##           Effect DFn DFd       F        p p<.05   ges
## 1       genotype   1   9   6.344 3.30e-02     * 0.378
## 2          stage   1   9 120.586 1.63e-06     * 0.650
## 3 genotype:stage   1   9   0.139 7.17e-01       0.002
```

The same result is visible in a two-way repeated measures anova. Although the number of eggs and of hatched eggs differs between the rootstock genotypes and the number of hatched eggs is lower than the initial number of eggs, there is no genotype*stage effect. This means that the rootstock genotype does not influence the hatching rate.

Effect of rootstock genotype on hatching

Although the LA1840 grafts have a higher number of hatched eggs, the grafting itself does not seem to be influenced by the rootstock, as the percentage of hatched eggs is equal on the LA1840 and MM grafts.

Nymph development

For the nymph development, I want to focus on the development from first instar to fourth instar. To cancel-out effect on oviposition and variation in hatching, take percentage of later nymphs from first instars.

First, I'll make a new dataframe for the total number of nymphs that passed a developmental stage for each plant

```
# new data set from flies_wide without unnecessary columns
flies_total <- flies_wide %>%
  dplyr::select(!c(`0_egg`, `1_first_instar`, `4_early_fourth_instar`, `5_completed_lifecycle`, total:p

# create new columns for the number of nymphs passing a developmental stage
flies_total <- flies_total %>%
  mutate(second_instars = rowSums(flies_total[,7:9])) %>%
  mutate(third_instars = rowSums(flies_total[,8:9])) %>%
  mutate(fourth_instars = `6_total_fourth_instars`)

# remove unnecessary columns
flies_total <- flies_total %>%
  dplyr::select(!(`2_second_instar`:`6_total_fourth_instars`))

# make data long
flies_total <- flies_total %>%
  pivot_longer(cols = eggs_start:fourth_instars,
               names_to = "stage", values_to = "number")

# make data wide for day
flies_total <- flies_total %>%
  dplyr::select(-date) %>%
  pivot_wider(names_from = day, values_from = number)

# make a column for the highest number per developmental stage per plant over time
flies_total <- flies_total %>%
  rowwise() %>%
  mutate(number = max(c_across(`2`:`29`)))

# remove the counts per day
```

```r
flies_total <- flies_total %>%
  dplyr::select(!c(`2`:`29`))

# order the developmental stages
flies_total$stage <- factor(flies_total$stage,
                            levels= c("eggs_start", "hatched", "second_instars",
                                      "third_instars", "fourth_instars"))

# The first 10 rows of the data
knitr::kable(flies_total[1:10,1:4])
```

genotype	place	stage	number
LA1840	1	eggs_start	60
LA1840	1	hatched	40
LA1840	1	second_instars	37
LA1840	1	third_instars	36
LA1840	1	fourth_instars	35
MM	2	eggs_start	66
MM	2	hatched	46
MM	2	second_instars	45
MM	2	third_instars	41
MM	2	fourth_instars	40

And also for the numbers relative to eggs and hatched eggs

```r
# new  wide dataframe from flies_total for numbers relative to eggs
flies_relative_eggs <- flies_total %>%
  pivot_wider(names_from = stage, values_from = number)

# make a column for each stage relative to eggs
flies_relative_eggs <- flies_relative_eggs %>%
  mutate(eggs = eggs_start/eggs_start*100) %>%
  mutate(hatched_eggs = hatched/eggs_start*100) %>%
  mutate(second = second_instars/eggs_start*100) %>%
  mutate(third = third_instars/eggs_start*100) %>%
  mutate(fourth = fourth_instars/eggs_start*100)

# remove absolute numbers and make long
flies_relative_eggs <- flies_relative_eggs %>%
  dplyr::select(!c(eggs_start:fourth_instars)) %>%
  pivot_longer(cols = eggs:fourth,
               names_to = "stage", values_to = "perc_from_eggs")

# order the developmental stages
flies_relative_eggs$stage <- factor(flies_relative_eggs$stage,
                            levels= c("eggs", "hatched_eggs", "second",
                                      "third", "fourth"))

# The first 10 rows of the data
knitr::kable(flies_relative_eggs[1:10,1:4])
```

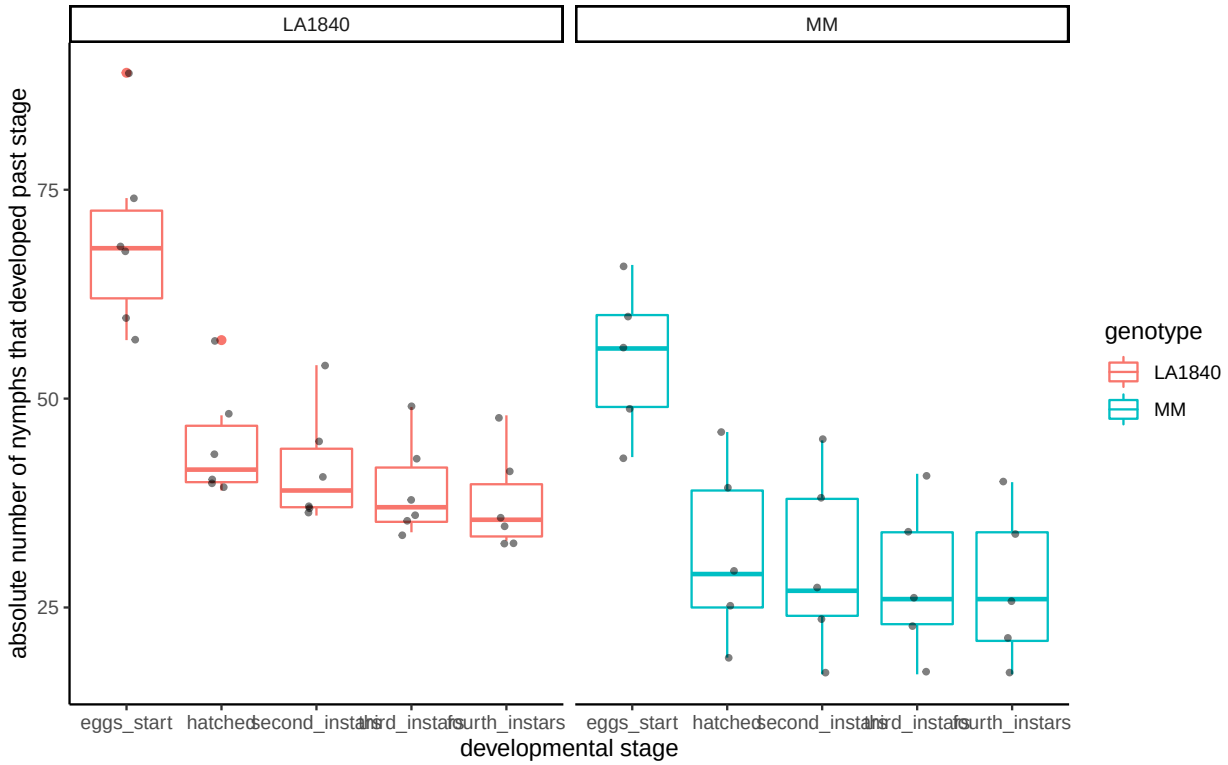
genotype	place	stage	perc_from_eggs
LA1840	1	eggs	100.00000
LA1840	1	hatched_eggs	66.66667
LA1840	1	second	61.66667
LA1840	1	third	60.00000
LA1840	1	fourth	58.33333
MM	2	eggs	100.00000
MM	2	hatched_eggs	69.69697
MM	2	second	68.18182
MM	2	third	62.12121
MM	2	fourth	60.60606

```
# new  wide dataframe from flies_total for numbers relative to hatched
flies_relative_hatched <- flies_total %>%
  pivot_wider(names_from = stage, values_from = number)

# make a column for each stage relative to hatched
flies_relative_hatched <- flies_relative_hatched %>%
  mutate(eggs = eggs_start/hatched*100) %>%
  mutate(hatched_eggs = hatched/hatched*100) %>%
  mutate(second = second_instars/hatched*100) %>%
  mutate(third = third_instars/hatched*100) %>%
  mutate(fourth = fourth_instars/hatched*100)

# remove absolute numbers and make long
flies_relative_hatched <- flies_relative_hatched %>%
  dplyr::select(!c(eggs_start:fourth_instars)) %>%
  pivot_longer(cols = eggs:fourth,
               names_to = "stage", values_to = "perc_from_hatched")

# order the developmental stages
flies_relative_hatched$stage <- factor(flies_relative_hatched$stage,
                          levels= c("eggs", "hatched_eggs", "second",
                                    "third", "fourth"))

# The first 10 rows of the data
knitr::kable(flies_relative_hatched[1:10,1:4])
```

genotype	place	stage	perc_from_hatched
LA1840	1	eggs	150.00000
LA1840	1	hatched_eggs	100.00000
LA1840	1	second	92.50000
LA1840	1	third	90.00000
LA1840	1	fourth	87.50000
MM	2	eggs	143.47826
MM	2	hatched_eggs	100.00000
MM	2	second	97.82609
MM	2	third	89.13043
MM	2	fourth	86.95652

What happens in the absolute numbers through the development?

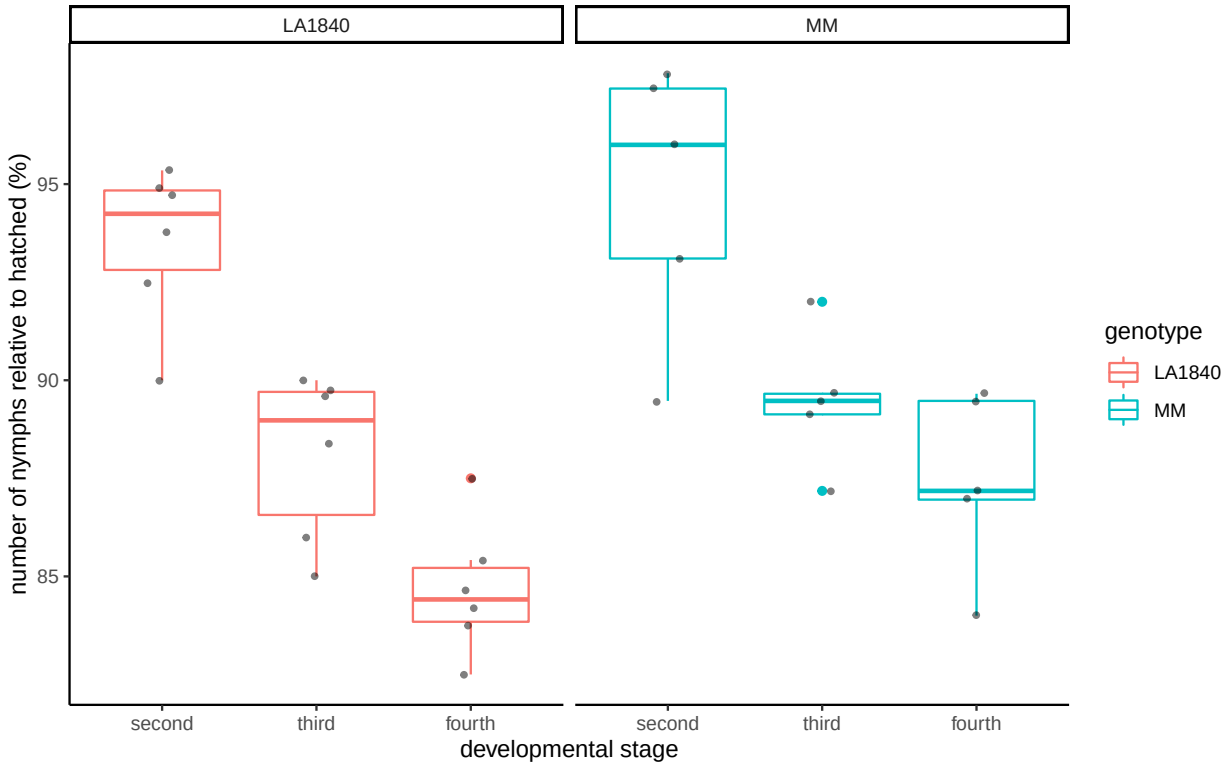
```
flies_total %>%
  mutate(number = na_if(number, 0)) %>%
  group_by(genotype, stage) %>%
  ggplot(., aes(x = stage, y = number)) +
  facet_wrap( vars(genotype), nrow = 1) +
  geom_boxplot(aes(color = genotype))   +
  geom_jitter(size = 1, alpha = 0.5, width = 0.1) +
  labs(y="absolute number of nymphs that developed past stage",
       x= "developmental stage") +
  theme_classic()
```



The number of nymphs appears to be higher on LA1840 grafts, but this could be the result of the higher number of eggs, as discussed above.

What happens if we look at the number of nymphs relative to the number of hatched eggs?

```
flies_relative_hatched %>%
  dplyr::filter(stage != "eggs") %>%
  dplyr::filter(stage != "hatched_eggs") %>%
  group_by(genotype, stage) %>%
  ggplot(., aes(x = stage, y = perc_from_hatched)) +
  facet_wrap( vars(genotype), nrow = 1) +
  geom_boxplot(aes(color = genotype))   +
  geom_jitter(size = 1, alpha = 0.5, width = 0.1) +
  labs(y="number of nymphs relative to hatched (%)",
       x= "developmental stage") +
  theme_classic()
```



Now it seems like a larger percentage of eggs which hatched on MM grafts developed into fourth instar nymph than on LA1840 grafts.

Is this a significant difference?

An overview of the data:

```
flies_relative_hatched %>%
  group_by(genotype) %>%
  dplyr::filter(stage=="fourth") %>%
  get_summary_stats(perc_from_hatched, type = "mean_sd")
```

```
## # A tibble: 2 x 5
##   genotype variable              n  mean    sd
##   <chr>    <chr>             <dbl> <dbl> <dbl>
## 1 LA1840   perc_from_hatched     6  84.7  1.70
## 2 MM       perc_from_hatched     5  87.5  2.30
```

Shapiro-Wilk test:

```
flies_relative_hatched %>%
  dplyr::filter(stage=="fourth") %>%
  group_by(genotype) %>%
  shapiro_test(perc_from_hatched)
```

```
## # A tibble: 2 x 4
##   genotype variable          statistic     p
##   <chr>    <chr>                 <dbl> <dbl>
```

```
## 1 LA1840   perc_from_hatched     0.965 0.858
## 2 MM       perc_from_hatched     0.900 0.411
```

Levene's test

```
flies_relative_hatched %>%
  dplyr::filter(stage=="fourth") %>%
  levene_test(perc_from_hatched~genotype)
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     1     9     0.343 0.572
```

Identifying extreme outliers:

```
flies_relative_hatched %>%
  dplyr::filter(stage=="fourth") %>%
  group_by(genotype) %>%
  identify_outliers(perc_from_hatched)
```

```
## # A tibble: 1 x 6
##   genotype place stage  perc_from_hatched is.outlier is.extreme
##   <chr>    <int> <fct>              <dbl> <lgl>      <lgl>
## 1 LA1840       1 fourth              87.5 TRUE       FALSE
```

The assumptions for a t-test are met.

```
flies_relative_hatched %>%
  dplyr::filter(stage=="fourth") %>%
  t_test(perc_from_hatched ~ genotype, var.equal = TRUE) %>%
  add_significance()
```

```
## # A tibble: 1 x 9
##   .y.               group1 group2    n1    n2 statistic    df      p p.signif
## * <chr>             <chr>  <chr>  <int> <int>     <dbl> <dbl>  <dbl> <chr>
## 1 perc_from_hatched LA1840 MM         6     5     -2.32     9 0.0455 *
```

```
flies_relative_hatched %>%
  dplyr::filter(stage=="fourth") %>%
  cohens_d(perc_from_hatched ~ genotype, var.equal = TRUE)
```

```
## # A tibble: 1 x 7
##   .y.               group1 group2 effsize    n1    n2 magnitude
## * <chr>             <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 perc_from_hatched LA1840 MM       -1.40     6     5 large
```

The percentage of hatched eggs that develop into fourth instar nymphs is indeed lower on LA1840 grafts than on MM grafts (p=0.046). The rootstock genotype has a large effect (d=1.40) on the development of nymphs into fourth instars after hatching.

Effect of rootstock genotype on the nymph development

In line with the previous results on normal LA1840 and MM plants, the nymph development is decreased on the LA1840 grafts compared to the MM grafts. Contrary to the previous findings, however, this is only visible in the development relative to the number of hatched eggs.

First conclusions

overall similar to results of Arjen, but diluted. Two contradicting effects of LA1840 rootstock on whiteflies: - increased oviposition - hampered nymph development

Effects found on MM scion suggests transportable nature of responsible mechanism(s).

Phenotype in detail

Now, let's have a more detailed look at the development.

Hatching (first instars)

```
# new data set from flies_wide without unnecessary columns
flies_model <- flies_wide %>%
  mutate(day = as.numeric(as.character(day))) %>%
  dplyr::select(!c(`0_egg`, `4_early_fourth_instar`, `5_completed_lifecycle`,
                   total:perc_hatched))

# create new columns for the cumulative number of nymphs per developmental stage
flies_model <- flies_model %>%
  mutate(first_instars = rowSums(flies_model[,7:10])) %>%
  mutate(second_instars = rowSums(flies_model[,8:10])) %>%
  mutate(third_instars = rowSums(flies_model[,9:10])) %>%
  mutate(fourth_instars = `6_total_fourth_instars`)

# update columns for the percentage of nymphs from hatched per stage
flies_model <- flies_model %>%
  mutate(first_instars = first_instars/hatched*100) %>%
  mutate(second_instars = second_instars/hatched*100) %>%
  mutate(third_instars = third_instars/hatched*100) %>%
  mutate(fourth_instars = fourth_instars/hatched*100)

flies_first <- flies %>%
  dplyr::filter(stage=="1_first_instar") %>%
  dplyr::select(-eggs_start, -hatched, -percentage_from_eggs)
```

I want to use the right model, so that it the first step. The drc package has a function for selection the best fitting model, but needs an initial model to start with.

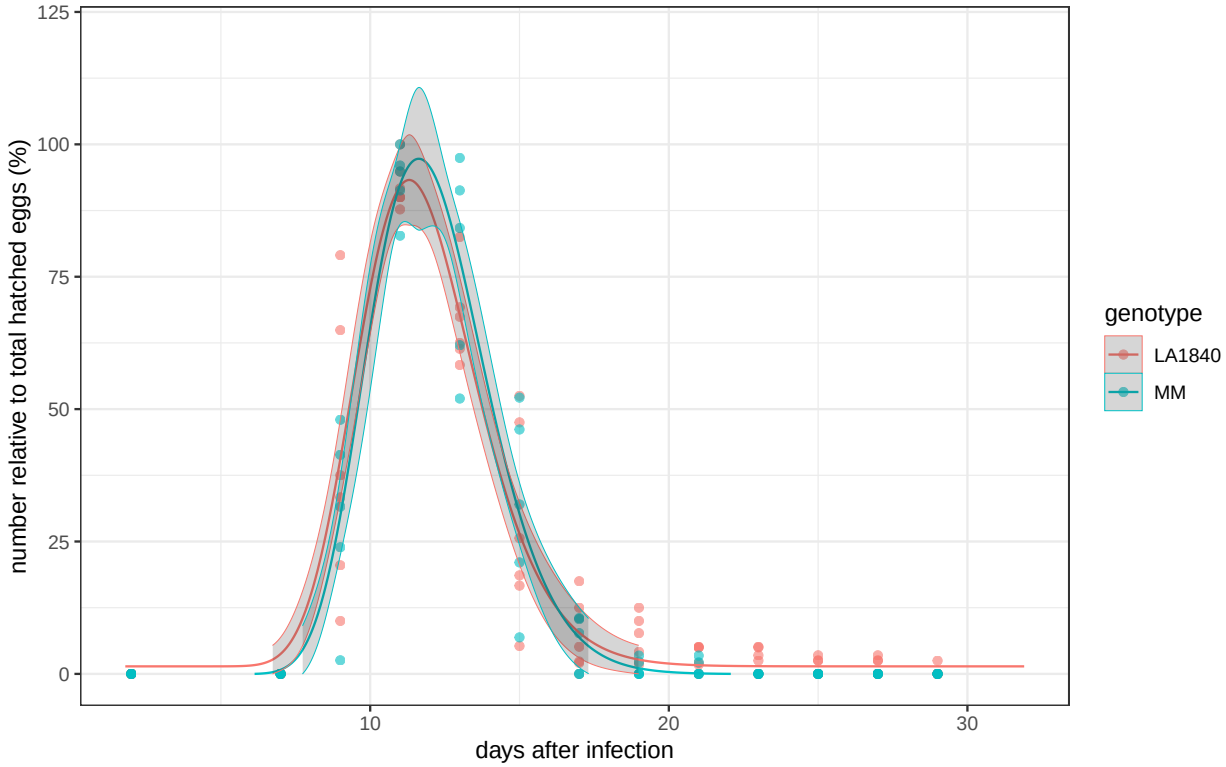
```
#set-up first model
model_test<- drm(percentage_from_hatched ~ day,
                 data = flies_first,
                 fct = lgaussian())
```

```
#select model with best fit
mselect(model_test, list(gaussian(), lgaussian()))
```

```
##              logLik       IC Lack of fit  Res var
## lgaussian -516.6169 1045.234  0.97750254 83.35643
## lgaussian -516.6169 1045.234  0.97750254 83.35643
## gaussian  -523.2087 1058.417  0.07271987 91.40669
```

The data follows a skewed bell-shaped pattern, so the best fitting model is lgaussian.

```
tidydrc_model(flies_first, day, percentage_from_hatched, model = lgaussian(),
              genotype) %>%
  tidydrc_plot(ed50 = F, color = ~genotype, confint = T) +
  labs(x = "days after infection", y = "number relative to total hatched eggs (%)") +
  ylim(0,120)
```



and cumulative:

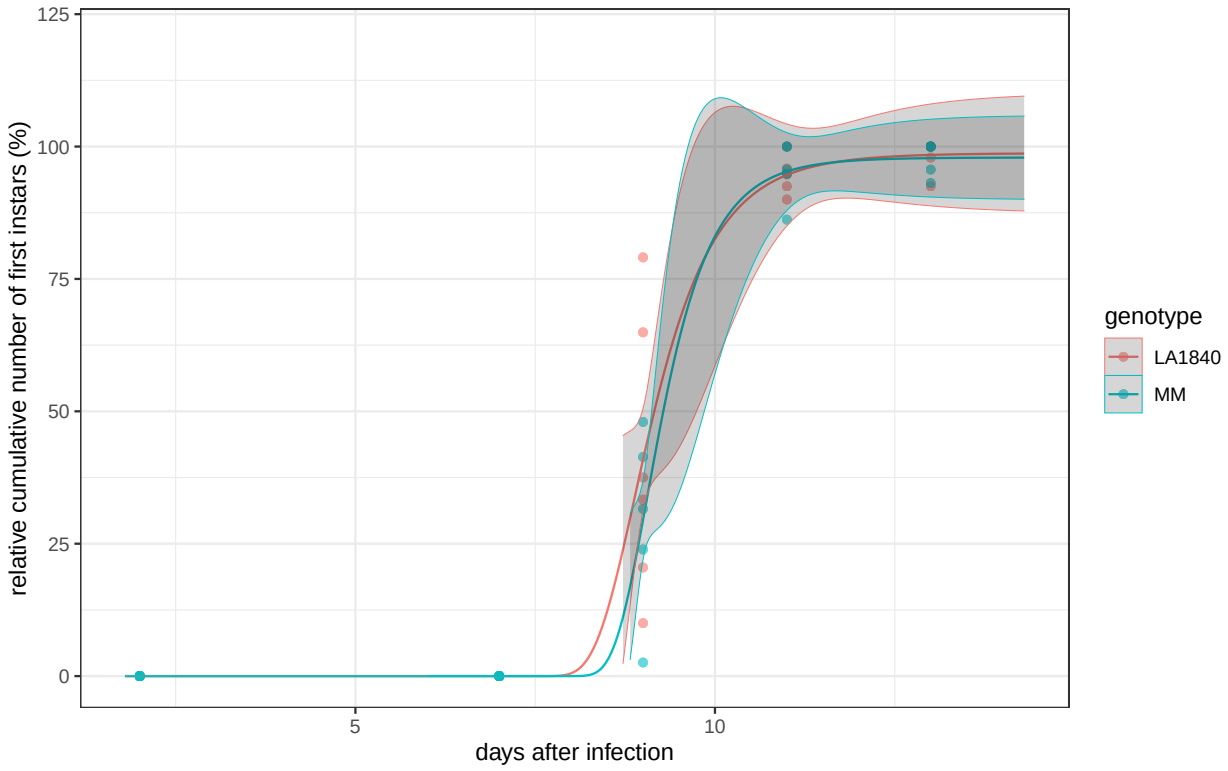
```
#set-up first model
model_test<- drm(first_instars ~ day,
                 data = filter(flies_model, day<15),
                 fct = W2.3())

#select model with best fit
mselect(model_test, list(W1.3 (),
                         W1.4 (),
```

```
                        W2.3 (),
                        W2.4 (),
                        LL.3 ()))
```

```
##         logLik       IC Lack of fit  Res var
## W1.3 -203.8932 415.7863   1.0000000 102.7716
## LL.3 -203.9020 415.8041   0.9919707 102.8048
## W2.3 -204.1723 416.3446   0.7758792 103.8201
## W2.3 -204.1723 416.3446   0.7758792 103.8201
## W1.4 -203.8932 417.7863   0.9985979 104.7868
## W2.4 -204.2108 418.4216   0.4495433 106.0041
```

W1.3 has the best fit.

```
tidydrc_model(filter(flies_model, day<15), day, first_instars,
              model = W1.3(names = c("Slope", "Upper Limit", "ED50")),
              genotype) %>%
  tidydrc_plot(ed50 = F, color = ~genotype, confint = T) +
  labs(x = "days after infection", y = "relative cumulative number of first instars (%)") +
  ylim(0,120)
```



Second instars

```
flies_second <- flies %>%
  dplyr::filter(stage=="2_second_instar") %>%
  dplyr::select(-eggs_start, -hatched, -percentage_from_eggs)
```

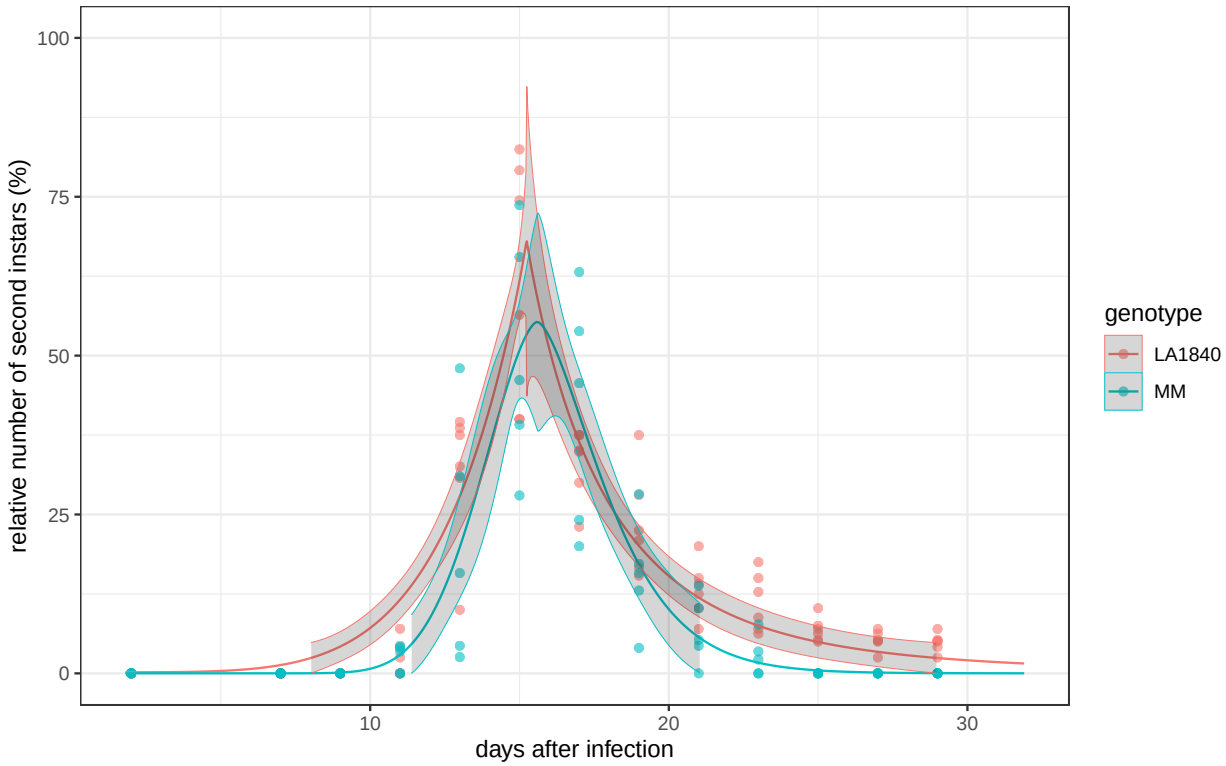
I want to use the right model, so that it the first step. The drc package has a function for selection the best fitting model, but needs an initial model to start with.

```
#set-up first model
model_test<- drm(percentage_from_hatched ~ day,
                 data = flies_second,
                 fct = lgaussian())

#select model with best fit
mselect(model_test, list(gaussian(), lgaussian()))
```

```
##              logLik       IC Lack of fit  Res var
## lgaussian -508.3260 1028.652 0.467282660 74.22996
## lgaussian -508.3260 1028.652 0.467282660 74.22996
## gaussian  -515.0255 1042.051 0.009905193 81.52155
```

The data follows a skewed bell-shaped pattern, so the best fitting model is lgaussian.

```
tidydrc_model(flies_second, day, percentage_from_hatched, model = lgaussian(),
              genotype) %>%
  tidydrc_plot(ed50 = F, color = ~genotype, confint = T) +
  labs(x = "days after infection", y = "relative number of second instars (%)") +
  ylim(0,100)
```



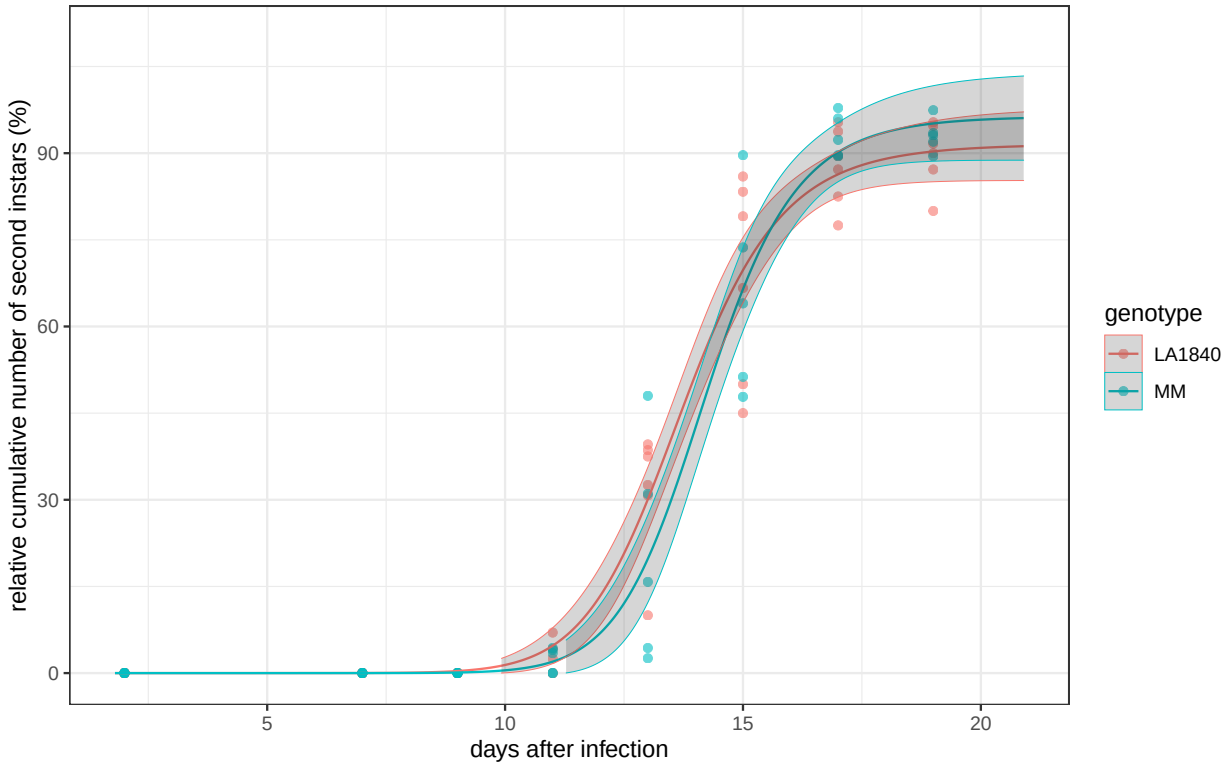
and cumulative:

```
#set-up first model
model_test<- drm(second_instars ~ day,
                 data = filter(flies_model, day<20),
                 fct = W2.3())

#select model with best fit
mselect(model_test, list(W1.3 (),
                         W1.4 (),
                         W2.3 (),
                         W2.4 (),
                         LL.3 ()))
```

```
##         logLik       IC Lack of fit  Res var
## LL.3 -309.9928 627.9856   0.9138519 69.55369
## W2.3 -310.1600 628.3199   0.8756996 69.81848
## W2.3 -310.1600 628.3199   0.8756996 69.81848
## W1.3 -310.8684 629.7367   0.6823143 70.95163
## W2.4 -309.9969 629.9938   0.8281074 70.38829
## W1.4 -310.7760 631.5520   0.5703184 71.64576
```

LL.3 has the best fit.

```
tidydrc_model(filter(flies_model, day<20), day, second_instars,
              model = LL.3(names = c("Slope", "Upper Limit", "ED50")),
              genotype) %>%
  tidydrc_plot(ed50 = F, color = ~genotype, confint = T) +
  labs(x = "days after infection", y = "relative cumulative number of second instars (%)") +
  ylim(0,110)
```



Third instars

```
flies_third <- flies %>%
  dplyr::filter(stage=="3_third_instar") %>%
  dplyr::select(-eggs_start, -hatched, -percentage_from_eggs)
```

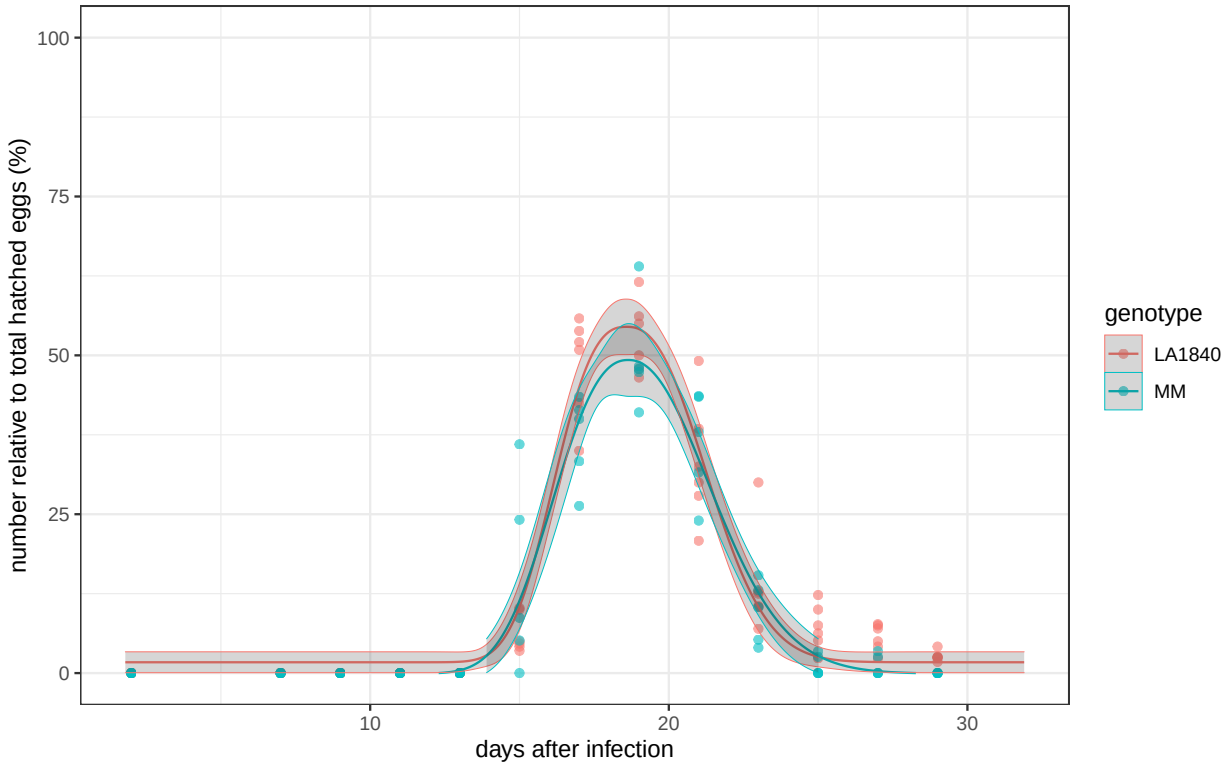
I want to use the right model, so that it the first step. The drc package has a function for selection the best fitting model, but needs an initial model to start with.

```
#set-up first model
model_test<- drm(percentage_from_hatched ~ day,
                 data = flies_third,
                 fct = lgaussian())

#select model with best fit
mselect(model_test, list(gaussian(), lgaussian()))
```

```
##              logLik       IC Lack of fit  Res var
## lgaussian -444.6945 901.3889  0.66827302 30.48445
## lgaussian -444.6945 901.3889  0.66827302 30.48445
## gaussian  -449.8571 911.7142  0.05162815 32.76699
```

The data follows a skewed bell-shaped pattern, so the best fitting model is lgaussian.

```
tidydrc_model(flies_third, day, percentage_from_hatched, model = lgaussian(),
              genotype) %>%
  tidydrc_plot(ed50 = F, color = ~genotype, confint = T) +
  labs(x = "days after infection", y = "number relative to total hatched eggs (%)") +
  ylim(0,100)
```



and cumulative:

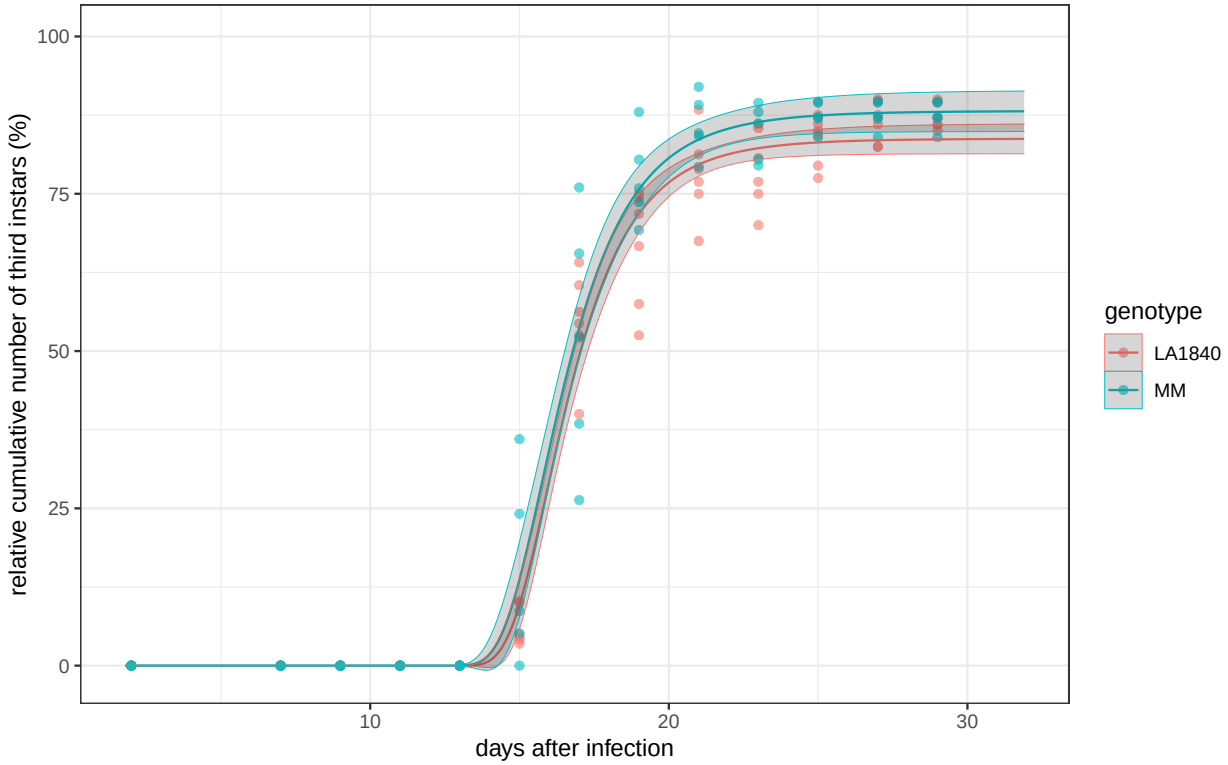
```
#set-up first model
model_test<- drm(third_instars ~ day,
                 data = flies_model,
                 fct = W2.3())

#select model with best fit
mselect(model_test, list(W1.3 (),
                         W1.4 (),
                         W2.3 (),
                         W2.4 (),
                         LL.3 ()))
```

```
##         logLik       IC  Lack of fit  Res var
## W1.3 -463.3684 934.7369 8.073609e-01 39.01720
## W1.4 -463.3592 936.7184 7.345281e-01 39.29283
## LL.3 -469.4868 946.9735 6.300209e-02 42.50296
## W2.3 -481.9737 971.9473 1.164073e-05 50.61335
## W2.3 -481.9737 971.9473 1.164073e-05 50.61335
## W2.4 -481.2323 972.4646 9.779274e-06 50.45163
```

W1.3 has the best fit.

```
tidydrc_model(flies_model, day, third_instars,
              model = W1.3(names = c("Slope", "Upper Limit", "ED50")),
              genotype) %>%
```

```
  tidydrc_plot(ed50 = F, color = ~genotype, confint = T) +
  labs(x = "days after infection", y = "relative cumulative number of third instars (%)") +
  ylim(-1,100)
```



Fourth instars

```
flies_fourth <- flies %>%
  dplyr::filter(stage=="6_total_fourth_instars") %>%
  dplyr::select(-eggs_start, -hatched, -percentage_from_eggs)
```

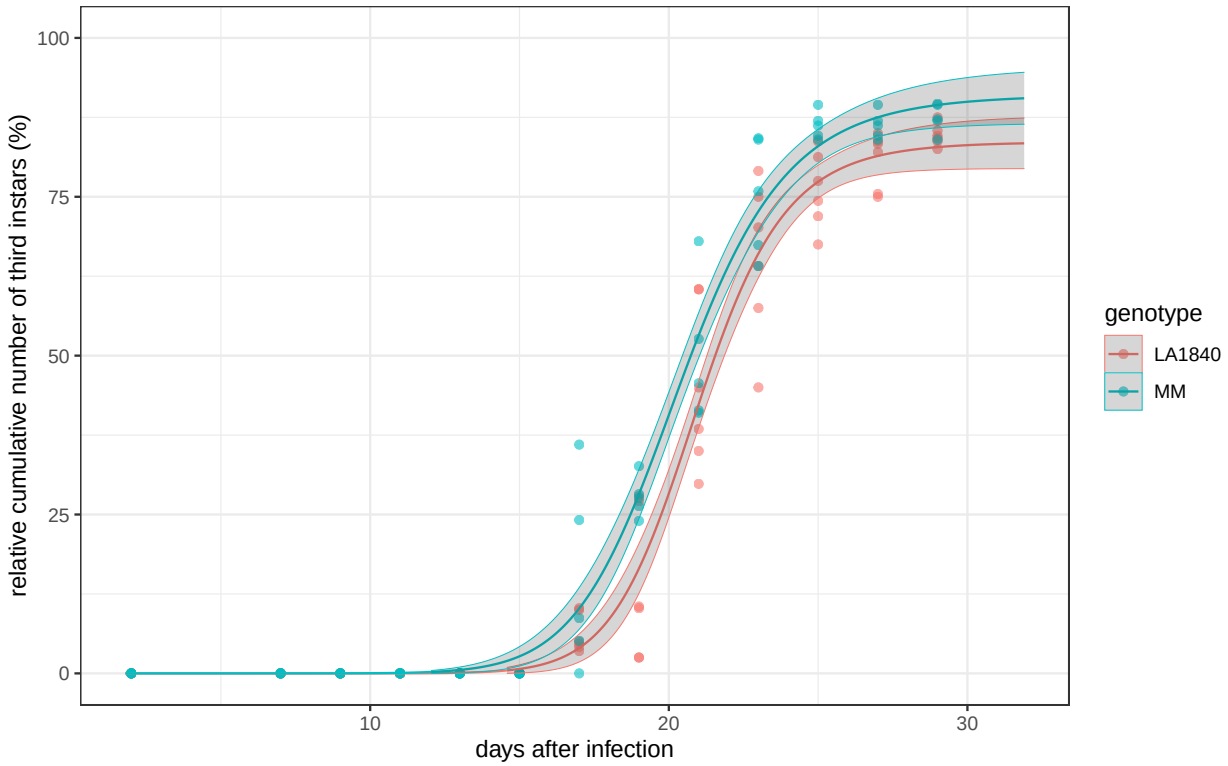
I want to use the right model, so that it the first step. The drc package has a function for selection the best fitting model, but needs an initial model to start with.

```
#set-up first model
model_test<- drm(percentage_from_hatched ~ day,
                 data = flies_fourth,
                 fct = W2.3())

#select model with best fit
mselect(model_test, list(W1.3 (),
                         W1.4 (),
                         W2.3 (),
                         W2.4 (),
                         LL.3 ()))
```

```
##         logLik       IC Lack of fit  Res var
## LL.3 -471.9820 951.9640  0.88743640 44.01243
## W2.3 -472.5888 953.1775  0.79886517 44.38751
## W2.3 -472.5888 953.1775  0.79886517 44.38751
## W2.4 -472.4306 954.8611  0.75244292 44.60804
## W1.3 -478.2705 964.5410  0.07812853 48.05870
## W1.4 -477.9226 965.8452  0.06435632 48.16950
```

The best fitting model is LL.3.

```
tidydrc_model(flies_fourth, day, percentage_from_hatched,
              model = LL.3(names = c("Slope", "Upper Limit", "ED50")),
              genotype) %>%
  tidydrc_plot(ed50 = F, color = ~genotype, confint = T) +
  labs(x = "days after infection", y = "relative cumulative number of third instars (%)") +
  ylim(0,100)
```



Trying to figure out how to get the statistics: Nest time, use summary(model_test)