STAT 540: Companion to Lecture 7: Linear Models

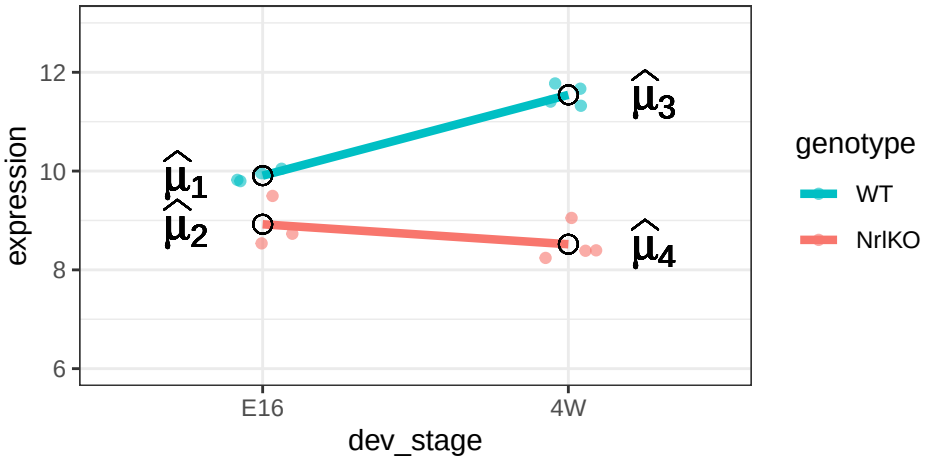
Keegan Korthauer

Note that the source Rmd for this document can be found here

Two-way ANOVA or a linear model with interaction

Which group means are we comparing in a model with 2 factors?

For simplicity, we first consider only two levels of `dev_stage`: E16 and 4W



$$\mu_1 = E[Y_{(WT,E16)}]$$
$$\mu_2 = E[Y_{(NrlKO,E16)}]$$
$$\mu_3 = E[Y_{(WT,4W)}]$$
$$\mu_4 = E[Y_{(NrlKO,4W)}]$$

Reference-treatment effect parametrization

By default, `lm` assumes a **reference-treatment effect** parametrization. We just need *more* indicator variables!!

Mathematically (a bit more difficult...)

$$Y_{ijk} = \theta + \tau_{KO} x_{KO,ijk} + \tau_{4W} x_{4W,ijk} + \tau_{KO:4W} x_{KO,ijk} x_{4W,ijk} + \varepsilon_{ijk}$$

Subscripts: i indexes samples per group, $j = \{WT, NrlKO\}$, $k = \{E16, 4W\}$

The names of these parameters and variables look overwhelming but think of them as just names for:

> $x_{KO,ijk}$: a indicator variable with value 1 for NrlKO genotype samples (any sample with j=NrlKO), and 0 otherwise. I call this variable x_{KO}
>
> $x_{4W,ijk}$: a different indicator variable with value 1 for 4W samples (any sample with k=4W), and 0 otherwise. I call this variable x_{4W}

τ_{KO}, τ_{4W}, and $\tau_{KO:4W}$: parameters to model the *simple* effects of genotype (NrlKO), development (4W), and their interaction

Note: in this "simple" version with 2 levels per factor we need only one indicator variable per factor: x_{KO} and x_{4W}. But this model can be extended to multiple factors with multiple levels. You just need to add more indicator variables!

Reference: WT & E16

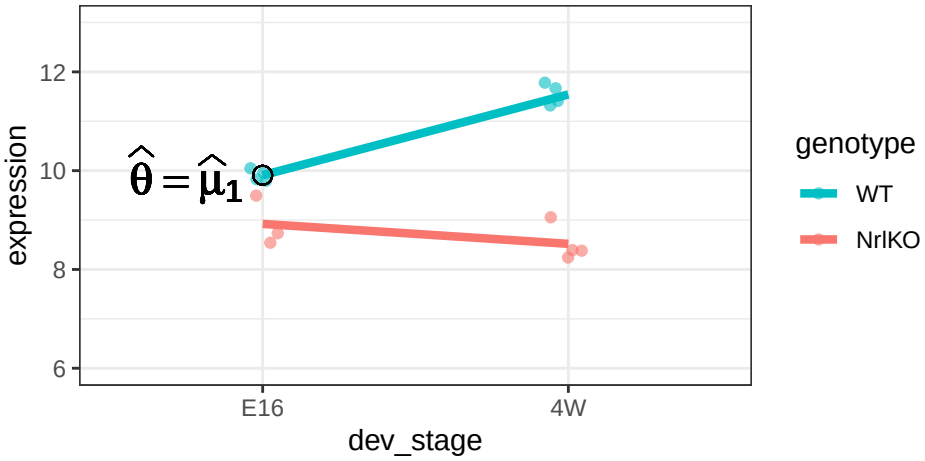
As before, comparisons are relative to a reference but now we have reference levels in both factors: **E16** and **WT**

$$Y_{ijk} = \theta + \tau_{KO}x_{KO,ijk} + \tau_{4W}x_{4W,ijk} + \tau_{KO:4W}x_{KO,ijk}x_{4W,ijk} + \varepsilon_{ijk}$$

For any sample i in the reference group: $j = WT$ and $k = E16$, then $x_{KO} = 0$ and $x_{4W} = 0$ (I'm omiting subscripts for clarity). Then only θ remains and we get:

$$E[Y_{WT,E16}] = \theta$$

as before θ is the mean of the reference group



Here is the `lm` output (the `coeff`) table for the two factor fit.

```
twoFactFit <- lm(expression ~ genotype * dev_stage, oneGene)
summary(twoFactFit)$coeff
```

```
##                              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                 9.9069542  0.1574053 62.939133 2.017456e-15
## genotypeNrlKO              -0.9844049  0.2404406 -4.094171 1.776894e-03
## dev_stage4W                 1.6366093  0.2226047  7.352087 1.444463e-05
## genotypeNrlKO:dev_stage4W -2.0403721  0.3276653 -6.227001 6.465669e-05
```

Notice that the **`lm` estimate**, $\hat{\theta}$, is the sample mean of the reference group (WT E16).

```
(means.2Fact <- group_by(oneGene, dev_stage, genotype) %>%
          summarize(cellMeans = mean(expression)) %>%
          ungroup() %>%
          mutate(txEffects = cellMeans - cellMeans[1],
                 lmEst = as.vector(summary(twoFactFit)$coeff[,1])))
```

```
## `summarise()` has grouped output by 'dev_stage'. You can override using the `.groups` argument.
```

```
## # A tibble: 4 x 5
##   dev_stage genotype cellMeans txEffects  lmEst
##   <fct>     <fct>        <dbl>     <dbl>  <dbl>
## 1 E16       WT            9.91     0      9.91
## 2 E16       NrlKO         8.92    -0.984 -0.984
## 3 4W        WT           11.5      1.64   1.64
## 4 4W        NrlKO         8.52    -1.39  -2.04
```

To show this explicitly, we pull out the `lm` estimate for the reference group (WT E16):

```
means.2Fact %>% filter(dev_stage == "E16" & genotype == "WT") %>%
  pull(lmEst)
```

```
## [1] 9.906954
```

And now the sample mean of the reference group (WT E16):

```
means.2Fact %>% filter(dev_stage == "E16" & genotype == "WT") %>%
  pull(cellMeans)
```

```
## [1] 9.906954
```

***Simple* genotype effect: WT *vs* NrlKO at E16**

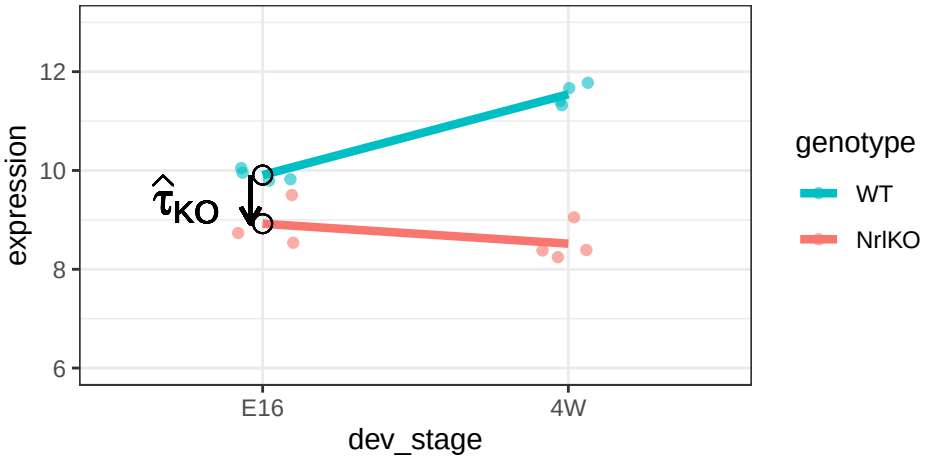
$$Y_{ijk} = \theta + \tau_{KO}x_{KO,ijk} + \tau_{4W}x_{4W,ijk} + \tau_{KO:4W}x_{KO,ijk}x_{4W,ijk} + \varepsilon_{ijk}$$

For any WT sample at E16: $x_{KO} = 0$ and $x_{4W} = 0$. Then $E[Y_{WT,E16}] = \theta$

For any KO sample at E16: $x_{KO} = 1$ and $x_{4W} = 0$. Then $E[Y_{NrlKO,E16}] = \theta + \tau_{KO}$

Substracting these expectations we get τ_{KO}, the *conditional* genotype effect at E16 :

$$\tau_{KO} = E[Y_{NrlKO,E16}] - E[Y_{WT,E16}]$$



And its **`lm` estimate**, $\hat{\tau}_{KO}$, is the *difference* of sample respective means.

To show this explicitly, we pull out the `lm` estimate for the KO effect (diff between E16:NrlKO and E16:WT):

```
means.2Fact %>% filter(dev_stage == "E16" & genotype == "NrlKO") %>%
  pull(lmEst)
```

```
## [1] -0.9844049
```

And now the differences in sample means between the E16:NrlKO group and the reference group (WT E16):

```
means.2Fact %>% filter(dev_stage == "E16" & genotype == "NrlKO") %>%
  pull(txEffects)
```

```
## [1] -0.9844049
```

***Simple* developmental effect: E16 *vs* 4W in WT**

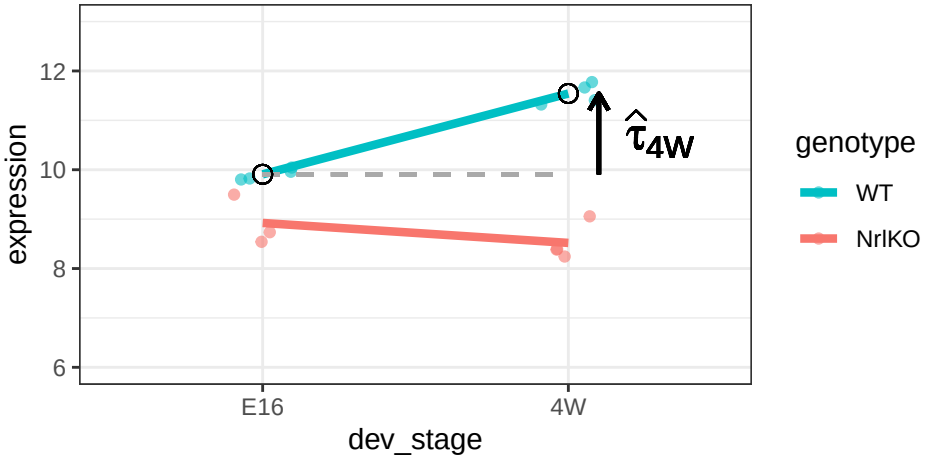
$$Y_{ijk} = \theta + \tau_{KO} x_{KO,ijk} + \tau_{4W} x_{4W,ijk} + \tau_{KO:4W} x_{KO,ijk} x_{4W,ijk} + \varepsilon_{ijk}$$

For any WT sample at E16: $x_{KO} = 0$ and $x_{4W} = 0$. Then $E[Y_{WT,E16}] = \theta$

For any WT sample at 4W: $x_{KO} = 0$ and $x_{4W} = 1$. Then $E[Y_{WT,4W}] = \theta + \tau_{4W}$

Substracting these expectations we get τ_{4W}, the *conditional* development effect in WT :

$$\tau_{4W} = E[Y_{WT,4W}] - E[Y_{WT,E16}]$$



And its **`lm` estimate**, $\hat{\tau}_{4W}$, is the *difference* of respective sample means.

To show this explicitly, we pull out the `lm` estimate for the 4W effect (diff between 4W:WT and E16:WT):

```
means.2Fact %>% filter(dev_stage == "4W" & genotype == "WT") %>%
  pull(lmEst)
```

```
## [1] 1.636609
```

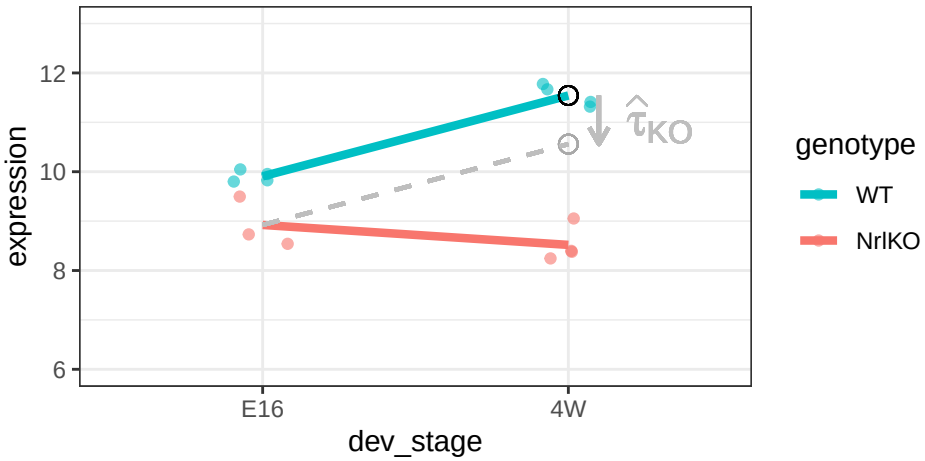
And now the differences in sample means between the E16:NrlKO group and the reference group (WT E16):

```
means.2Fact %>% filter(dev_stage == "4W" & genotype == "WT") %>%
  pull(txEffects)
```

```
## [1] 1.636609
```

Interaction effect

Can we simply add up the simple effect of genotype NrlKO, and the simple effect of developmental stage 4W, to get the effect at 4W in NrlKO?? If so, we'd expect the 4W:NrlKO group to have a mean predicted by the dotted grey line (i.e.**that the effect of KO is the same at E16 as it is at 4W**:



We see that this does not seem to be the case. This is where the **interaction** effect comes in. Let's see what it means mathematically.

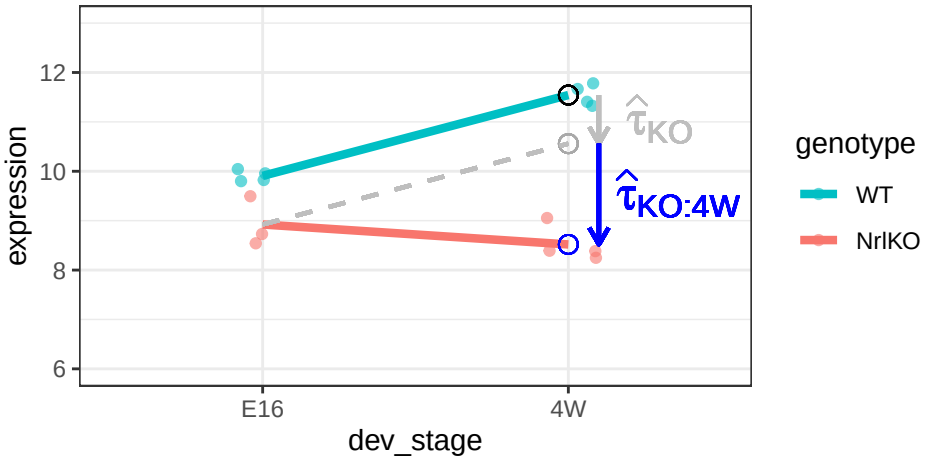
$$Y_{ijk} = \theta + \tau_{KO}x_{KO,ijk} + \tau_{4W}x_{4W,ijk} + \tau_{KO:4W}x_{KO,ijk}x_{4W,ijk} + \varepsilon_{ijk}$$

For any KO sample at 4W: $x_{KO} = 1$ and $x_{4W} = 1$. Then $E[Y_{NrlKO,4W}] = \theta + \tau_{KO} + \tau_{4W} + \tau_{KO:4W}$

Using the expectations from above, you can show that:

$$\tau_{KO:4W} = (E[Y_{NrlKO,4W}] - E[Y_{WT,4W}]) - (E[Y_{NrlKO,E16}] - E[Y_{WT,E16}])$$

This term is represented by the blue arrow:



And its **`lm` estimate**, $\hat{\tau}_{KO:4W}$, is the *difference of the differences* between NrlKO and WT at each developmental stage.

To show this explicitly, we pull out the `lm` estimate for the interaction effect:

```
means.2Fact %>% filter(dev_stage == "4W" & genotype == "NrlKO") %>%
  pull(lmEst)
```

```
## [1] -2.040372
```

And now the differences in sample means between the E16:NrlKO group and the reference group (WT E16):

```r
mean.E16.WT <- means.2Fact %>% filter(dev_stage == "E16" & genotype == "WT") %>% pull(cellMeans)
mean.E16.KO <- means.2Fact %>% filter(dev_stage == "E16" & genotype == "NrlKO") %>% pull(cellMeans)
mean.4W.WT <- means.2Fact %>% filter(dev_stage == "4W" & genotype == "WT") %>% pull(cellMeans)
mean.4W.KO <- means.2Fact %>% filter(dev_stage == "4W" & genotype == "NrlKO") %>% pull(cellMeans)

(mean.4W.KO - mean.4W.WT) - (mean.E16.KO - mean.E16.WT)
```

```
## [1] -2.040372
```

Two-way ANOVA without interaction: additive models

The interpretation of the coefficients changed when we drop the interaction terms

Mathematically

$$Y_{ijk} = \theta + \tau_{KO}x_{KO,ijk} + \tau_{4W}x_{4W,ijk} + \varepsilon_{ijk}$$

Note that this model is simpler and it has fewer parameters! but what do these mean?

As before, let's look at the expectations of each group:

- for any sample i with $j = WT$ and $k = E16$: $x_{KO} = 0$ and $x_{4W} = 0$ (only θ remains):

$$E[Y_{WT,E16}] = \theta$$

- for any sample i with $j = WT$ and $k = 4W$: $x_{KO} = 0$ and $x_{4W} = 1$:

$$E[Y_{WT,4W}] = \theta + \tau_{4W}$$

- for any sample i with $j = NrlKO$ and $k = E16$: $x_{KO} = 1$ and $x_{4W} = 0$:

$$E[Y_{KO,E16}] = \theta + \tau_{KO}$$

- for any sample i with $j = NrlKO$ and $k = 4W$: $x_{KO} = 1$ and $x_{4W} = 1$:

$$E[Y_{KO,4W}] = \theta + \tau_{KO} + \tau_{4W}$$

After some simple algebra, you get:

$$(E[Y_{WT,4W}] - E[Y_{WT,E16}]) + (E[Y_{KO,4W}] - E[Y_{KO,E16}]) = 2\tau_{4W}$$

Then,

$$\tau_{4W} = (\text{Eff}_{4W|WT} + \text{Eff}_{4W|KO})/2$$

is the average effect of 4W over the levels of `genotype`!!

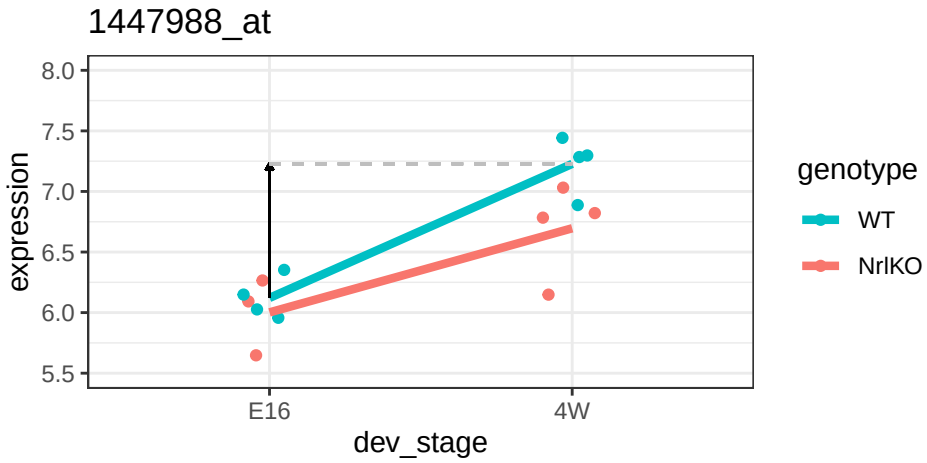
Similar for the other *main effect*.

The intercept parameter is now $\bar{Y} - \bar{x}_{ij,KO}\hat{\tau}_{KO} - \bar{x}_{ij,4W}\hat{\tau}_{4W}$

Some additional examples

Example 4: development in WT is statistically significant

Here is an example gene which has only the effect of developmental stage significant (in WT). The other two terms are not significant (effect of genotype at E16 and interaction).




```
##                               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                  6.1213470  0.1429312 42.8272394 1.370697e-13
## genotypeNrlKO               -0.1194771  0.2183310 -0.5472294 5.951490e-01
## dev_stage4W                  1.1065598  0.2021352  5.4743551 1.935791e-04
## genotypeNrlKO:dev_stage4W -0.4122782  0.2975349 -1.3856463 1.933054e-01
```

Again, the interaction effect is not significant, so there may be a development effect *regardless* of the genotype. Or likewise a genotype effect *regardless* of developmental stage. We need to test those hypotheses (main effects) using `anova`

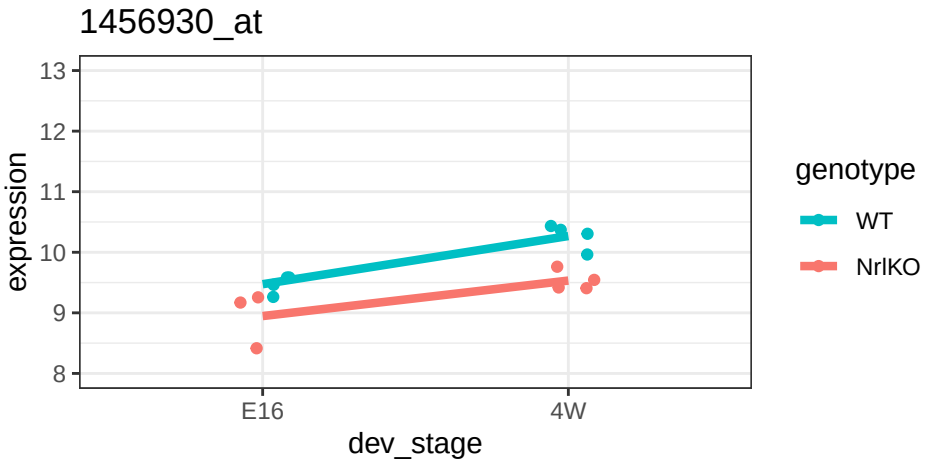
```
anova(multFit)
```

```
## Analysis of Variance Table
##
## Response: expression
##                     Df  Sum Sq Mean Sq F value    Pr(>F)
## genotype             1 0.28444 0.28444  3.4808   0.08896 .
## dev_stage            1 3.11838 3.11838 38.1606 6.933e-05 ***
## genotype:dev_stage   1 0.15690 0.15690  1.9200   0.19331
## Residuals           11 0.89889 0.08172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that there is indeed a significant main effect of developmental stage. But the main effect of genotype is not significant.

Example 5: both simple development and genotype are statistically significant

but not the interaction... note the almost parallel pattern




```
##                               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                  9.4731686  0.1269533 74.6192979 3.114098e-16
## genotypeNrlKO               -0.5261161  0.1939244 -2.7129955 2.018255e-02
## dev_stage4W                  0.7950743  0.1795391  4.4284178 1.014273e-03
## genotypeNrlKO:dev_stage4W -0.2091718  0.2642744 -0.7914946 4.453874e-01
```

Note that the main effects for both are also significant (but not the interaction).

```
## Analysis of Variance Table
##
## Response: expression
##                     Df  Sum Sq Mean Sq F value    Pr(>F)
## genotype             1 1.29452 1.29452 20.0799 0.0009298 ***
## dev_stage            1 1.81238 1.81238 28.1126 0.0002516 ***
## genotype:dev_stage   1 0.04039 0.04039  0.6265 0.4453874
## Residuals           11 0.70915 0.06447
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```