Lecture 7 – Linear Models

STAT/BIOF/GSAT 540

Keegan Korthauer

2020/01/27

Slides by: Gabriela Cohen Freue with contributions from Jenny Bryan and Keegan Korthauer

Recall from last class...

- show how to compare means of different groups (2 or more) using a linear regression model

- dummy variables to model the levels of a qualitative explanatory variable

- write a linear model using matrix notation

- understand which matrix is built by R

- distinguish between single and multiple hypotheses

- t-tests vs F-tests

Quick review: from t-test to linear regression

HOW??

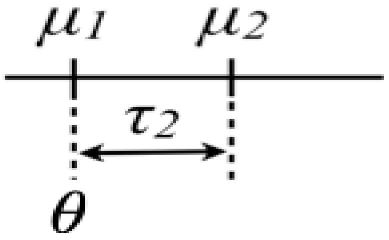
Changing the parametrization and using dummy variables

$$Y \sim G;\ E[Y] = \mu_Y;\ Z \sim G;\ E[Z] = \mu_Z$$

$$\downarrow$$

$$Y_{ij} = \theta + \tau_2 \times x_{ij2} + \varepsilon_{ij};\ i = 1, \ldots, n;\ j = 1, 2$$

$$\downarrow$$



$$\begin{aligned} E[Y_{i1}] &= \theta = \mu_1 \\ E[Y_{i2}] &= \theta + \tau_2 = \mu_1\ + (\mu_2 - \mu_1) = \mu_2 \end{aligned}$$

Using matrix notation ...

$$Y_{ij} = \theta + \tau_2 \times x_{ij2} + \varepsilon_{ij}$$

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{n_1 1} \\ Y_{12} \\ \vdots \\ Y_{n_2 2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \theta \\ \tau_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{n_1 1} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{n_2 2} \end{bmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X\alpha} + \boldsymbol{\varepsilon}$$

... and similarly beyond 2 groups comparisons (ANOVA)

WHY??

$$Y = \boxed{X}\alpha + \varepsilon$$

This gives us a VERY FLEXIBLE framework!!

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 1.22 \\ 1 & 2.02 \\ 1 & 1.42 \\ \vdots & \vdots \\ 1 & 1.89 \\ 1 & 2.01 \\ \vdots & \vdots \\ 1 & 1.56 \\ 1 & 2.17 \\ 1 & 1.51 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 1.22 & 0 \\ 1 & 0 & 2.02 & 0 \\ 1 & 0 & 1.42 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1.89 & 0 \\ 1 & 1 & 2.01 & 2.01 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1.56 & 1.56 \\ 1 & 1 & 2.17 & 2.17 \\ 1 & 1 & 1.51 & 1.51 \end{bmatrix}$$

Parametrizations

Different ways of writing this [design matrix, parameter vector] pair correspond to different parametrizations of the model

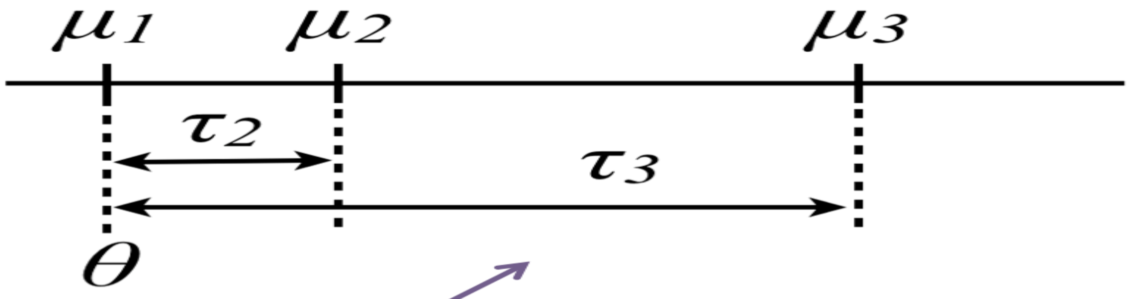
$$Y = [X\alpha] + \varepsilon$$

Understanding these concepts makes it easier ...

- to interpret fitted models
- to fit models such that comparisons you care most about are directly addressed in the inferential "report"

For example: comparisons of mean expression levels between groups!

$$
\begin{aligned}
\theta &= \mu_1 \\
\tau_2 &= \mu_2 - \mu_1 \\
\tau_3 &= \mu_3 - \mu_1
\end{aligned}
$$



These are **NOT population means**

These are **differences of means**

By default, `lm` estimates mean differences (with respect to a reference group):

```
summary(lm(gExp~devStage,subset(devDat,gene=="theHit")))$coeff
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 5.5408571  0.1021381 54.248698 1.307554e-34
## devStageP2  0.3040179  0.1398583  2.173756 3.678022e-02
## devStageP6  0.2433929  0.1398583  1.740282 9.085489e-02
## devStageP10 0.8342679  0.1398583  5.965093 9.559065e-07
## devStage4W  3.6325179  0.1398583 25.972843 5.266481e-24
```

Today... more complex models

- more than one factor with multiple levels

- how to model many categorical variables and their interaction

- distinguish between simple and main effects

- `lm` vs `anova` tests

- nested models

- t-tests vs F-tests

- continuous explanatory variables

- the regression line

Increasing the complexity of the linear model ...

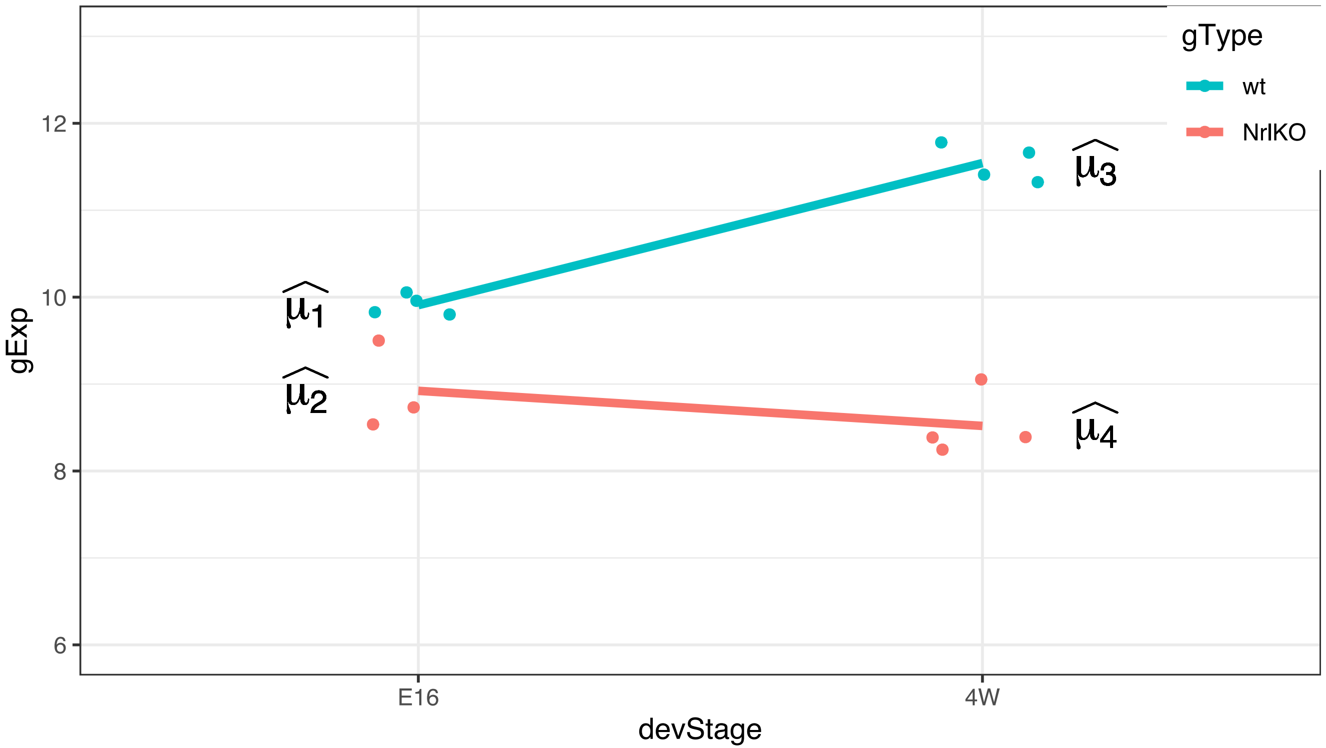
What if you have two categorical variables?

e.g., `gType` and `devStage` (for simplicity, let's consider only E16 and 4W)

- ANOVA is usually used to study models with one or more categorical variables (factors)
- Can we combine levels into 4 groups to simplify the analysis??

Two-way ANOVA or a linear model with interaction

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



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

Reference-treatment effect parametrization

By default, `lm` assumes a **reference-treatment effect** parametrization (mathematically, we need *more* dummy variables, see math handout)

```
twoFactFit <- lm(gExp ~ gType * devStage, twoDat)
```

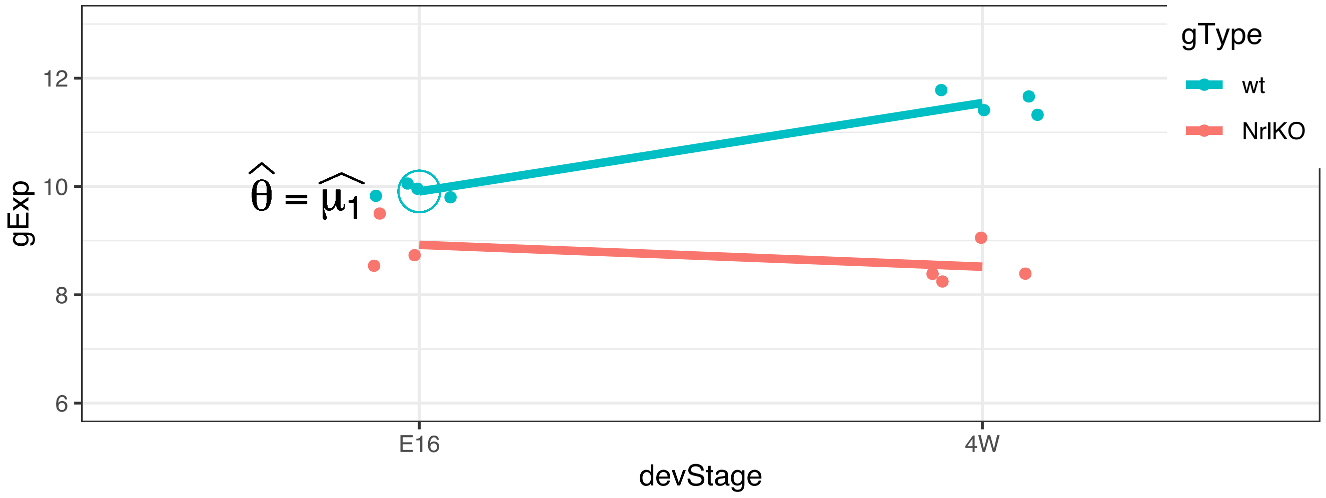
```
##                             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                9.9080000  0.1574912 62.911469 2.027211e-15
## gTypeNrlKO                -0.9856667  0.2405717 -4.097184 1.767824e-03
## devStage4W                 1.6345000  0.2227261  7.338609 1.469261e-05
## gTypeNrlKO:devStage4W -2.0380833  0.3278440 -6.216626 6.560671e-05
```

```
means.2Fact <- as.data.frame(twoDat %>%
          group_by(grp) %>% summarize(cellMeans=mean(gExp)))
(means.2Fact <-means.2Fact %>%
          mutate(txEffects=cellMeans-cellMeans[1],
          lmEst=summary(twoFactFit)$coeff[,1]))
```

```
##         grp cellMeans  txEffects      lmEst
## 1    wt.E16  9.908000  0.0000000  9.9080000
## 2 NrlKO.E16  8.922333 -0.9856667 -0.9856667
## 3     wt.4W 11.542500  1.6345000  1.6345000
## 4  NrlKO.4W  8.518750 -1.3892500 -2.0380833
```

The reference: wt & E16

As before, comparisons are relative to a reference but in this case there is a reference level in each factor: wt and E16



The reference: wt & E16

Mean of reference group: $\theta = E[Y_{wt,E16}]$

`lm` estimate: $\hat{\theta}$ is the sample mean of the group

```
##                           Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)              9.9080000  0.1574912 62.911469 2.027211e-15
## gTypeNrlKO              -0.9856667  0.2405717 -4.097184 1.767824e-03
## devStage4W               1.6345000  0.2227261  7.338609 1.469261e-05
## gTypeNrlKO:devStage4W -2.0380833  0.3278440 -6.216626 6.560671e-05
```

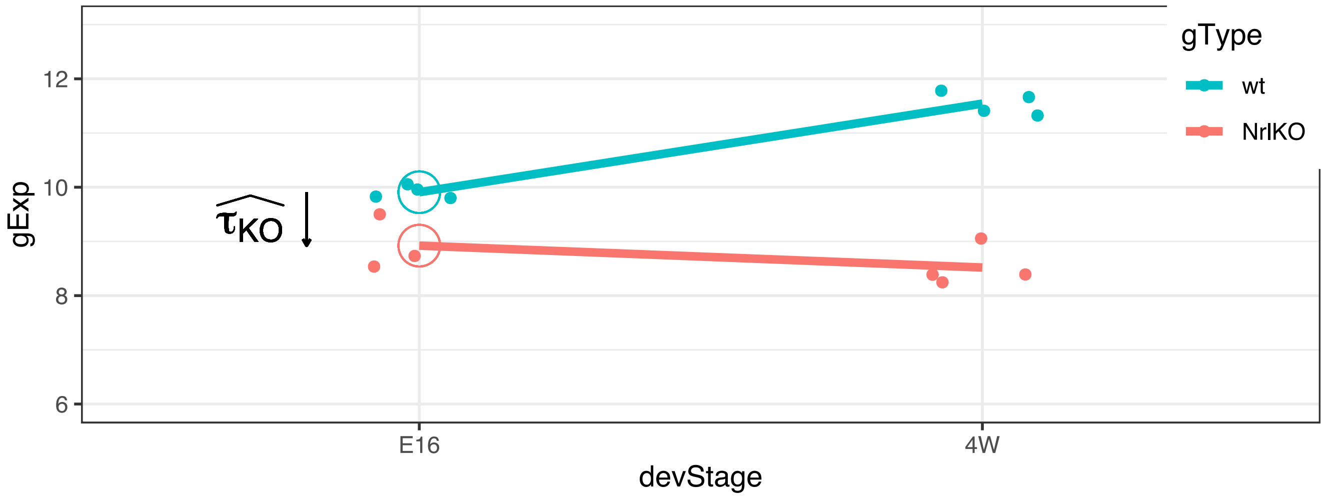
```
##         grp cellMeans  txEffects      lmEst
## 1    wt.E16  9.908000  0.0000000  9.9080000
## 2 NrlKO.E16  8.922333 -0.9856667 -0.9856667
## 3     wt.4W 11.542500  1.6345000  1.6345000
## 4  NrlKO.4W  8.518750 -1.3892500 -2.0380833
```

In general, one is not interested in: $H_0 : \theta = 0$

Simple genotype effect: wt *vs* NrlKO at E16

And now the "treatment effects"...

> **Important**: effects are not marginal but *conditional* effects (at a given level of the other factor, e.g., at E16), usually called **simple effects**



Simple genotype effect: wt *vs* NrlKO at E16

Effect of genotype at E16: $\tau_{KO} = E[Y_{NrlKO,E16}] - E[Y_{wt,E16}]$

`lm` estimate: $\hat{\tau}_{KO}$ is the *difference* of sample respective means (check below)

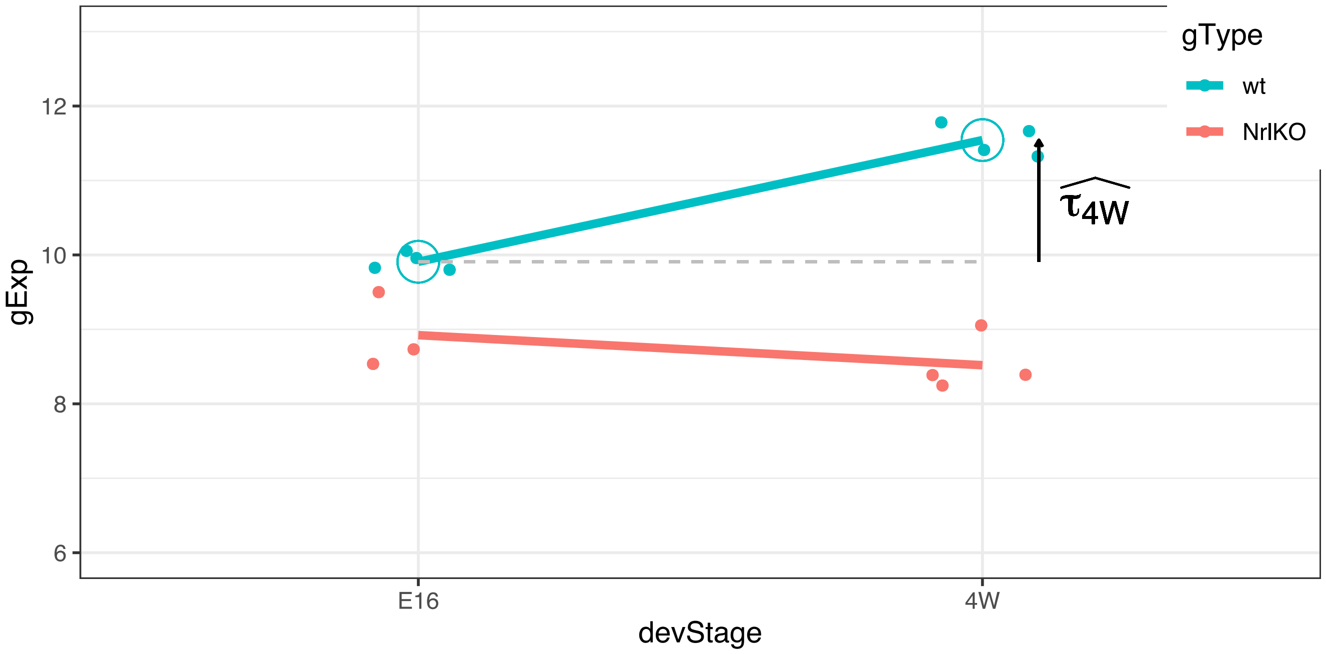
```
##                         Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)            9.9080000  0.1574912 62.911469 2.027211e-15
## gTypeNrlKO            -0.9856667  0.2405717 -4.097184 1.767824e-03
## devStage4W             1.6345000  0.2227261  7.338609 1.469261e-05
## gTypeNrlKO:devStage4W -2.0380833  0.3278440 -6.216626 6.560671e-05
```

```
##          grp cellMeans  txEffects      lmEst
## 1     wt.E16  9.908000  0.0000000  9.9080000
## 2  NrlKO.E16  8.922333 -0.9856667 -0.9856667
## 3      wt.4W 11.542500  1.6345000  1.6345000
## 4   NrlKO.4W  8.518750 -1.3892500 -2.0380833
```

But, do you want to test the *conditional* effect at E16: $H_0 : \tau_{KO} = 0$??

Simple developmental effect: E16 *vs* 4W at wt

Similarly, for the other factor:



Simple developmental effect: E16 *vs* 4W at wt

Effect of development at wt: $\tau_{4W} = E[Y_{wt,4W}] - E[Y_{wt,E16}]$

`lm` estimate: $\hat{\tau}_{4W}$ is the *difference* of respective sample means (check below)

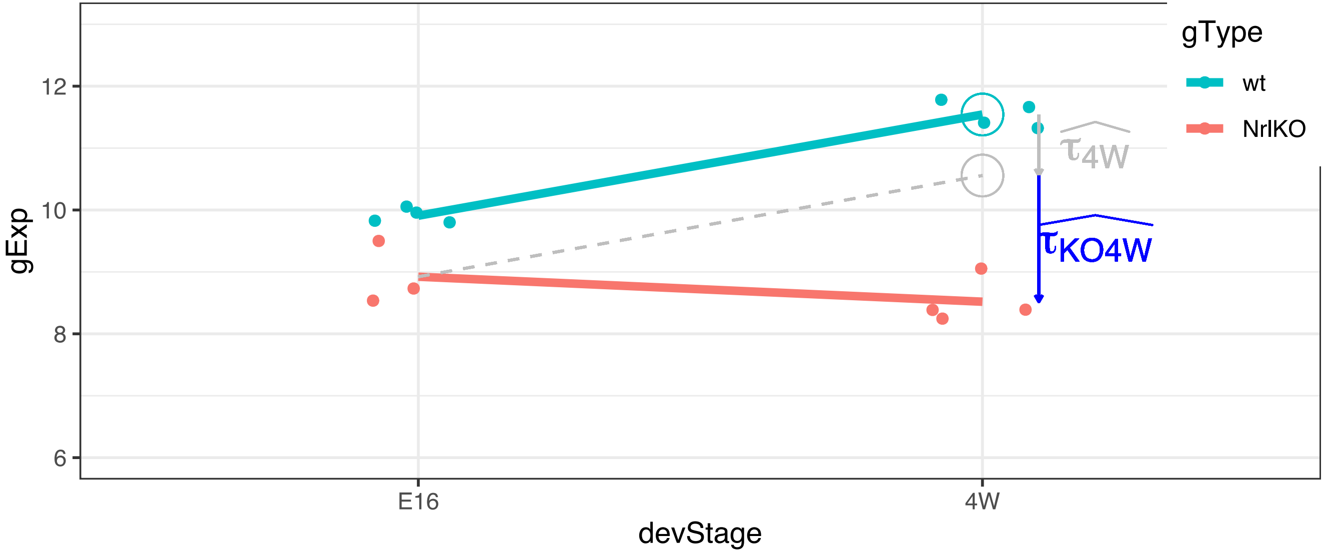
```
##                           Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)              9.9080000  0.1574912 62.911469 2.027211e-15
## gTypeNrlKO              -0.9856667  0.2405717 -4.097184 1.767824e-03
## devStage4W               1.6345000  0.2227261  7.338609 1.469261e-05
## gTypeNrlKO:devStage4W -2.0380833  0.3278440 -6.216626 6.560671e-05
```

```
##          grp cellMeans  txEffects      lmEst
## 1     wt.E16  9.908000  0.0000000  9.9080000
## 2   NrlKO.E16  8.922333 -0.9856667 -0.9856667
## 3      wt.4W 11.542500  1.6345000  1.6345000
## 4    NrlKO.4W  8.518750 -1.3892500 -2.0380833
```

Interaction effect

Is the effect of genotype the same at different developmental stages? (or does hte effect of development depend on genotype?)

Yes if, there's no interaction effect, i.e., $\tau_{KO4W} = 0$



The genotype effect at E16 is τ_{KO}. However, τ_{KO} does not seem to be the effect at 4W. The difference is the interaction effect!

Interaction effect

$$\tau_{KO4W} = (E[Y_{NrlKO,4W}] - E[Y_{wt,4W}]) - (E[Y_{NrlKO,E16}] - E[Y_{wt,E16}])$$

```
##                          Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)             9.9080000  0.1574912 62.911469 2.027211e-15
## gTypeNrlKO             -0.9856667  0.2405717 -4.097184 1.767824e-03
## devStage4W              1.6345000  0.2227261  7.338609 1.469261e-05
## gTypeNrlKO:devStage4W -2.0380833  0.3278440 -6.216626 6.560671e-05
```

```
means.2Fact
```

```
##          grp cellMeans  txEffects      lmEst
## 1     wt.E16  9.908000  0.0000000  9.9080000
## 2 NrlKO.E16  8.922333 -0.9856667 -0.9856667
## 3      wt.4W 11.542500  1.6345000  1.6345000
## 4   NrlKO.4W  8.518750 -1.3892500 -2.0380833
```

```
((means.2Fact$cellMeans[4]-means.2Fact$cellMeans[3])-
    (means.2Fact$cellMeans[2]-means.2Fact$cellMeans[1]))
```

```
## [1] -2.038083
```

NA

Summary of model parameters: with interaction

model parameter	R estimate	stats	interpretation
θ	(Intercept)	$E[Y_{wt,E16}]$	reference
τ_{KO}	gTypeNrlKO	$E[Y_{NrlKO,E16}] - E[Y_{wt,E16}]$	*conditional* effect of NrlKO *at* E16
τ_{4W}	devStage4_weeks	$E[Y_{wt,4W}] - E[Y_{wt,E16}]$	*conditional* effect of 4W *at* wt
τ_{KO4W}	gTypeNrlKO: devStage4_weeks	$E[Y_{NrlKO,4W}] - E[Y_{wt,4W}] - \tau_{KO}$	*interaction* effect of NrlKO and 4W

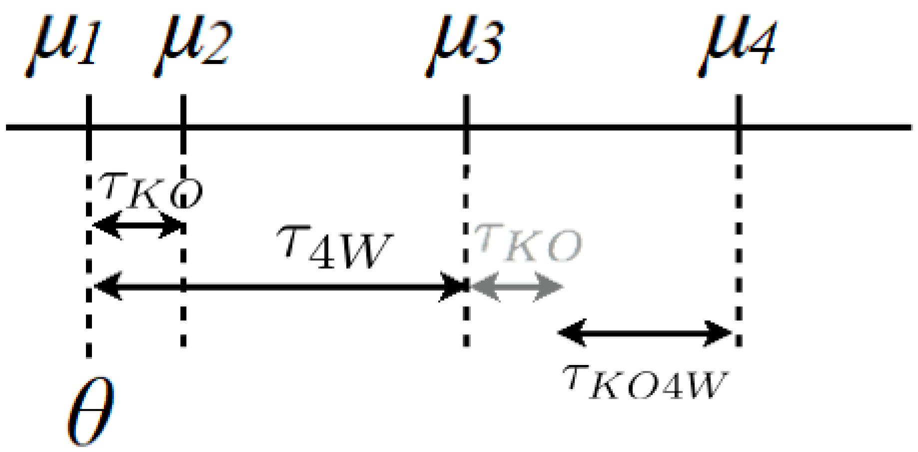
It is *important* to remember that `lm` reports *simple, not main* effects!! why?? because of the parametrization used!! (see see math handout)

It can also be shown that $\tau_{KO4W} = E[Y_{NrlKO,4W}] - \tau_{4W} - \tau_{KO} - \theta$

Let's examine these parameters closer and some examples

For our model, `lm` tests 4 hypotheses:

> $$H_0 : \theta = 0$$
>
> $$H_0 : \tau_{KO} = 0$$
>
> $$H_0 : \tau_{4W} = 0$$
>
> $$H_0 : \tau_{KO4W} = 0$$



We may not be interested in these hypotheses, e.g., τ_{KO} and τ_{4W} are *conditional* effects *at* a given level of a factor (*simple effects*)

Example 1: nothing is statistically significant, very flat genes *

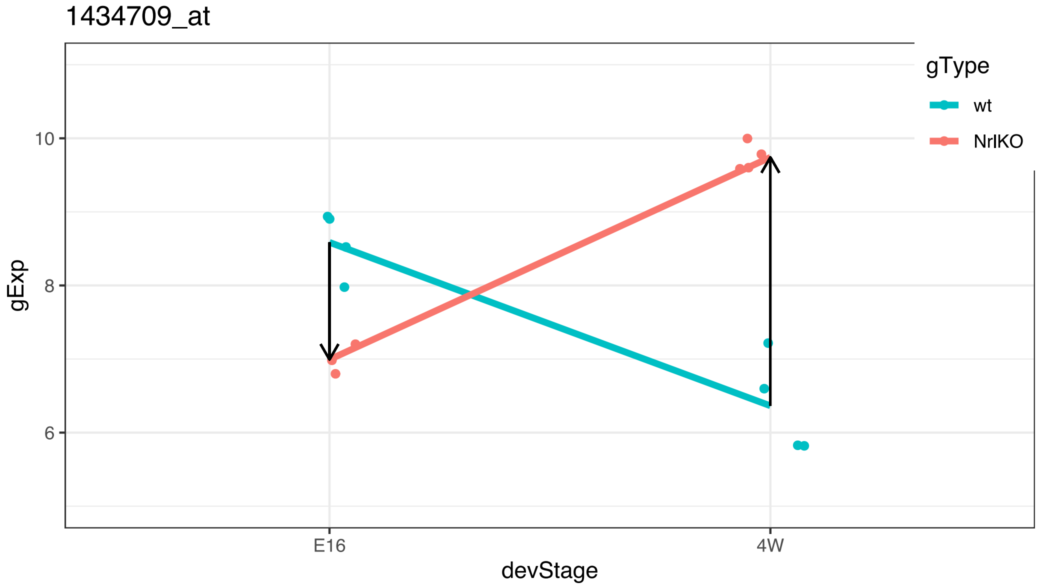
```
egDat<-prDat %>% subset(row.names(prDat) %in%
       c("1442080_at","1448243_at")) %>%
       tibble::rownames_to_column(var = "gene")%>%
       gather(sidChar, gExp,-gene) %>%
       inner_join(prDes,by="sidChar")
```

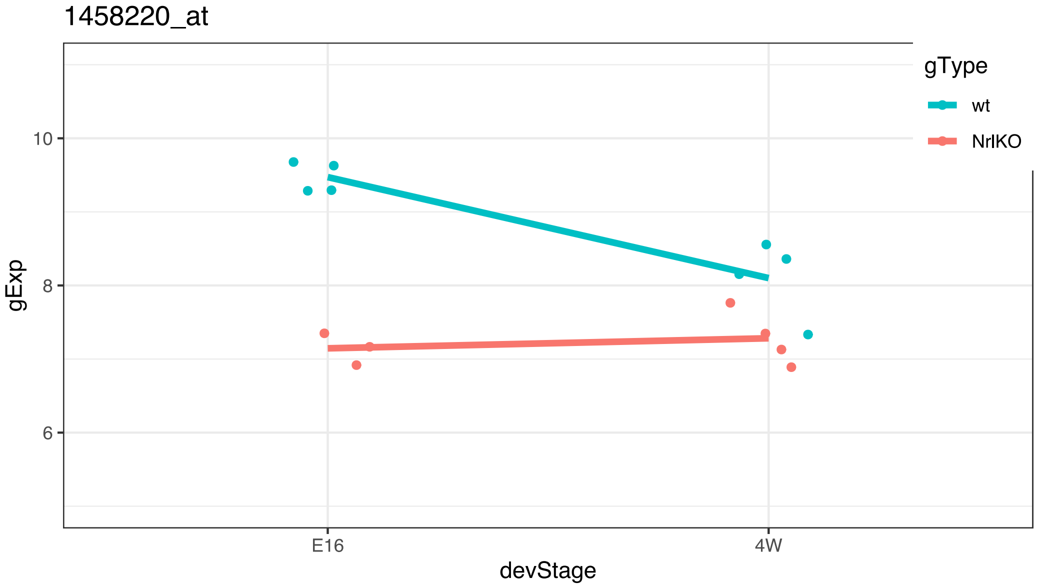
* Here and in next slides, summary of `lm` shown for the gene in the left plot

```
##                         Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)            8.5240000  0.2561305 33.2799114 2.154028e-12
## gTypeNrlKO            -0.4336667  0.3912458 -1.1084251 2.913251e-01
## devStage4W            -0.2532500  0.3622232 -0.6991545 4.989723e-01
## gTypeNrlKO:devStage4W  0.5504167  0.5331781  1.0323317 3.240804e-01
```

Example 2: statistically significant interaction effect: non-parallel

```
##                        Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)             8.58550  0.2215161 38.757897 4.083165e-13
## gTypeNrlKO             -1.59250  0.3383715 -4.706366 6.434783e-04
## devStage4W             -2.22075  0.3132711 -7.088907 2.021975e-05
## gTypeNrlKO:devStage4W   4.96975  0.4611226 10.777502 3.481389e-07
```





When the interaction effect is significant, the *simple* effects may not agree: compare the genotype effect @E16 with that @4W!

Example 3: balance & only genotype @E16 is statistically significant

To simplify future explanations, I've added a random observation in the NrlKO.E16 group (close to its mean) to have a *balanced* design

> In *unbalanced* designs the *main* effects are a *weighted* average of the simple effects, and the weights are not easy to interpret (beyond the scope of this course but worth noting the issue!)

```
egDat<-prDat %>% subset(row.names(prDat) %in%
       c("1447753_at","1431651_at")) %>%
       tibble::rownames_to_column(var = "gene")%>%
       gather(sidChar, gExp,-gene) %>%
       inner_join(prDes,by="sidChar") %>%
       mutate(grp=interaction(gType, devStage))

#duplicate sample 6 and add noise to gExp of genes
set.seed(123)
egDat <-egDat %>% subset(sidChar=="Sample_6") %>%
       mutate(sidChar="Sample_r",sidNum="r",
       gExp=gExp+rnorm(2,0,.1)) %>% rbind(egDat) %>%
       arrange(grp)
```

Example 3: only genotype @E16 is statistically significant: parallel

The interaction effect is not significant (almost parallel pattern).

Thus, there may be a genotype effect *regardless* of the developmental stage (*main* effect). However, that hypothesis is *not* tested here!!

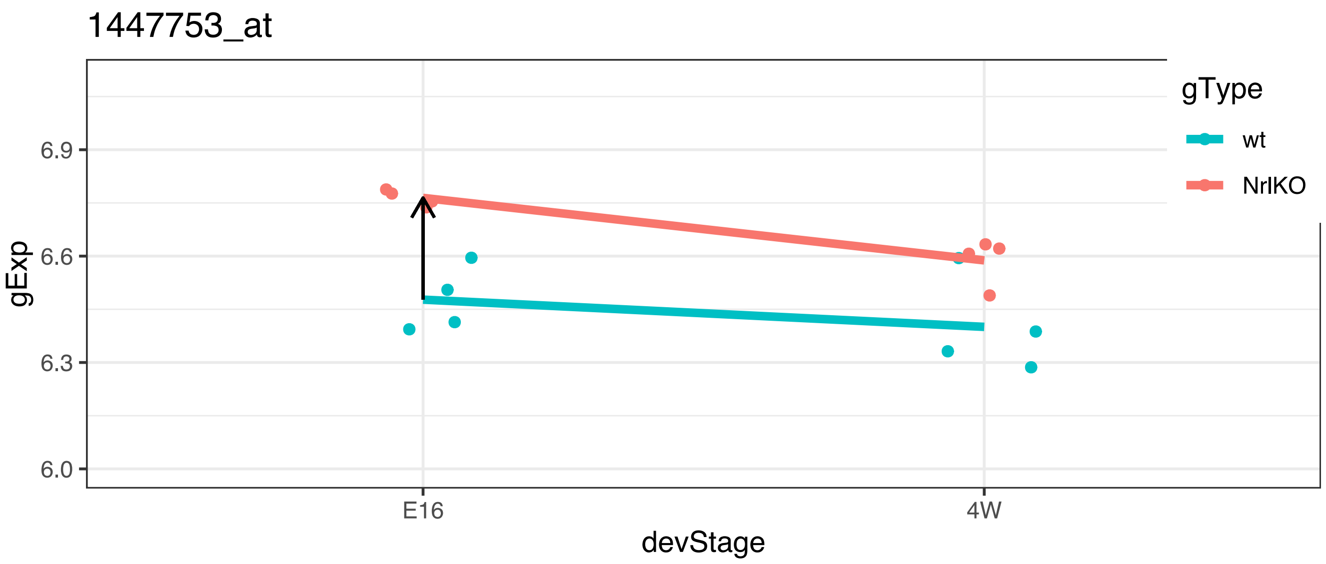
How do we test a *main effect*??

```
##                            Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)               6.47725000 0.04513333 143.513681 8.795702e-21
## gTypeNrlKO                0.28699556 0.06382816   4.496378 7.312501e-04
## devStage4W               -0.07675000 0.06382816  -1.202447 2.523805e-01
## gTypeNrlKO:devStage4W    -0.09949556 0.09026666  -1.102240 2.919732e-01
```

How do we test the *main* effects?

The main effect measures the *overall* association between the response and a factor. They are the (weighted) average of an effect over the levels of the other factor

> **Note**: a significant interaction means that the effect of a factor depends on the level of the other one. Thus, main effects may mask interesting results!



Main effects

`anova()` can be used to test the main effects:

$$H_0 : ((E[Y_{KO,E16}] - E[Y_{wt,E16}]) + (E[Y_{KO,4W}] - E[Y_{wt,4W}]))/2 = 0$$

> for unbalanced experiments $H_0 : w_1 \text{effect}_{E16} + w_2 \text{effect}_{4W} = 0$

```
tidy(anova(lm(gExp ~ gType * devStage, plot1Dat)))
```

```
## # A tibble: 4 x 6
##   term                 df   sumsq  meansq statistic   p.value
##   <chr>             <int>   <dbl>   <dbl>     <dbl>     <dbl>
## 1 gType                 1 0.225   0.225       27.6   0.000202
## 2 devStage              1 0.0640  0.0640       7.86  0.0160
## 3 gType:devStage        1 0.00990 0.00990      1.21  0.292
## 4 Residuals            12 0.0978  0.00815     NA    NA
```

As we suspected in slide #26, there is a significant genotype effect for this gene (1447753_at), i.e., its mean expression changes in NrlKO group (compared to wt), on average over develpmental stages.

* `anova` uses type II sums of squares, which follows the principle of marginality, thus order matters in unbalanced designs!

Main & interaction effects: important notes

- A **significant interaction effect** means that the effect of one factor depends on the levels of the other one.
    - e.g., the effect of genotype depends on development
- **Main effects**: are the (weighted) average of an effect over the levels of the other factor.
- A **non-significant main effect** means that, on average, there's no evidence of a factor's effect
    - e.g, no evidence of a genotype effect, on average over both developmental stages
- **Note of caution**: if the interaction is significant, it is possible that one or both simple effects are significant but the average effect (i.e., the main effect) is not. This is because the effect of a factor *depends on* the level of the other one!

Additive models

- In some applications, we need to test the interaction term
- However, additive models are easier and smaller
- If there are no statistical or theoretical grounds to include the interaction term, additive models are preferred
- Additive effects: $E[Y_{NrlKO,4W}] - E[Y_{wt,E16}] = \tau_{KO} + \tau_{4W}$

```
addFit <- summary(lm(formula = gExp ~ gType +devStage,plot1Dat))$coeff
addFit
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  6.5021239  0.0394084 164.993346 5.609641e-23
## gTypeNrlKO   0.2372478  0.0455049   5.213675 1.670701e-04
## devStage4W  -0.1264978  0.0455049  -2.779872 1.561988e-02
```

* gene=1447753_at in table

Additive models

- In an additive model, the parameters are **average effects**, over the levels of the other factor. Now, same as in anova()!!

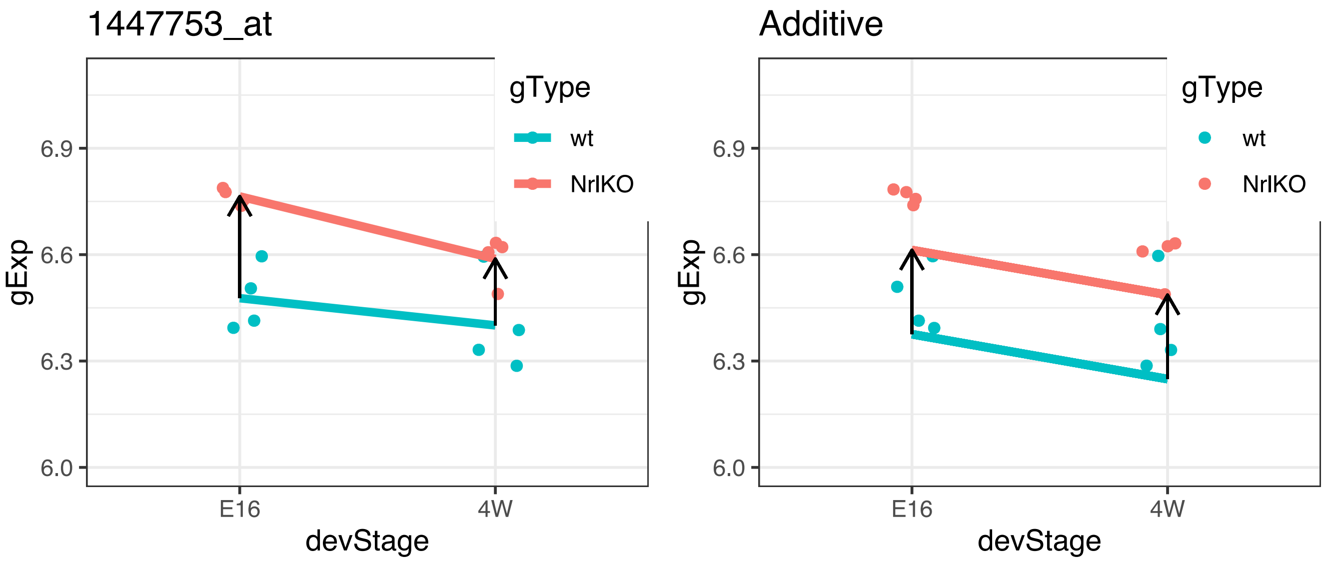
> Note the agreement!! This is gone in unbalanced designs since weights are computed differently! try!!

- TypeIII sum of squares are required for agreement in unbalanced designs (use `Anova` in `car`), beyond our scope

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  6.5021239  0.0394084 164.993346 5.609641e-23
## gTypeNrlKO   0.2372478  0.0455049   5.213675 1.670701e-04
## devStage4W  -0.1264978  0.0455049  -2.779872 1.561988e-02
```

```
tidy(anova(lm(gExp ~ gType + devStage,plot1Dat)))
```

```
## # A tibble: 3 x 6
##   term         df  sumsq  meansq statistic   p.value
##   <chr>     <int>  <dbl>   <dbl>     <dbl>     <dbl>
## 1 gType         1 0.225  0.225       27.2   0.000167
## 2 devStage      1 0.0640 0.0640       7.73  0.0156
## 3 Residuals    13 0.108  0.00828     NA    NA
```




```
multEst
```

```
##          (Intercept)            gTypeNrlKO            devStage4W
##           6.47725000            0.28699556           -0.07675000
## gTypeNrlKO:devStage4W
##          -0.09949556
```

```
addEst
```

```
## (Intercept)  gTypeNrlKO  devStage4W
##   6.5021239   0.2372478  -0.1264978
```

Factors with multiple levels

We can generalize the regression model to factors with more levels (e.g., E16, P2, P10 and 4W): we just additional dummy variables (and parameters)!!

> With interaction

```
##                           Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)             5.43325000  0.1240289 43.8063081 4.740219e-28
## gTypeNrlKO              0.25108333  0.1894573  1.3252764 1.954265e-01
## devStageP2              0.39900000  0.1754034  2.2747562 3.049627e-02
## devStageP6              0.19525000  0.1754034  1.1131483 2.747868e-01
## devStageP10             0.92000000  0.1754034  5.2450520 1.283680e-05
## devStage4W              3.96125000  0.1754034 22.5836544 5.952464e-20
## gTypeNrlKO:devStageP2  -0.22583333  0.2581868 -0.8746896 3.889296e-01
## gTypeNrlKO:devStageP6   0.06041667  0.2581868  0.2340037 8.166263e-01
## gTypeNrlKO:devStageP10 -0.20733333  0.2581868 -0.8030361 4.284868e-01
## gTypeNrlKO:devStage4W  -0.69333333  0.2581868 -2.6853939 1.185648e-02
```

Note that all the `devStage` parameters are still *simple* effects, but we now have more: one for each level compared to the reference

Factors with multiple levels (cont.)

> Without interaction: additive

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 5.52731618 0.11010606 50.1999257 9.574287e-33
## gTypeNrlKO  0.03159559 0.08783425  0.3597183 7.213497e-01
## devStageP2  0.30176103 0.14182289  2.1277315 4.091897e-02
## devStageP6  0.24113603 0.14182289  1.7002617 9.848949e-02
## devStageP10 0.83201103 0.14182289  5.8665498 1.428982e-06
## devStage4W  3.63026103 0.14182289 25.5971450 2.412597e-23
```

Parameters are now *main* effects (on average over the levels of the other factor) but we have more!

Is developmental a significant effect? We haven't tested that!!

Simultaneous hypotheses again

We generally test two types of null hypotheses:

$$H_0 : \tau_j = 0$$

vs

$$H_0 : \tau_j \neq 0$$

for each *j* individually

e.g., Is gene A differencially expressed 2 days after birth?

$$H_0 : \tau_{P2} = 0$$

$$H_0 : \tau_j = 0$$

vs

$$H_0 : \tau_j \neq 0$$

for all *j* at the same time

e.g., Is gene A significantly affected by time (`devStage`)?

$$H_0 : \tau_{P2} = \tau_{P6} = \tau_{P10} = \tau_{4W} = 0$$

F-test and overall significance: a deja vu

- the *t*-test in linear regression allows us to test single hypotheses. Those are given in the summary of `lm`

$$H_0 : \tau_i = 0$$

$$H_A : \tau_j \neq 0$$

- but we often like to test multiple hypotheses *simultaneously*:

$$H_0 : \tau_{P2} = \tau_{P6} = \tau_{P10} = \tau_{4W} = 0 \text{ [AND statement]}$$

$$H_A : \tau_i \neq 0 \text{ for some i [OR statement]}$$

the *F*-test allows us to test such compound tests

Overall effects: compound tests

With interaction

> $H_0 : \tau_{KO} = 0$ (1 df) $H_0 : \tau_{P2} = \tau_{P6} = \tau_{P10} = \tau_{4W} = 0$ (at wt!, 4 df)
> $H_0 : \tau_{KOP2} = \tau_{KOP6} = \tau_{KOP10} = \tau_{KO4W} = 0$ (4 df)

```
tidy(anova(lm(gExp~gType*devStage,hitDat)))
```

```
## # A tibble: 4 x 6
##   term              df   sumsq  meansq statistic   p.value
##   <chr>          <int>   <dbl>   <dbl>     <dbl>     <dbl>
## 1 gType              1  0.0692  0.0692      1.12  2.98e- 1
## 2 devStage           4 71.0    17.8       289.    6.69e-23
## 3 gType:devStage     4  0.689   0.172       2.80  4.44e- 2
## 4 Residuals         29  1.78    0.0615     NA    NA
```

Tests of overall effects of a factor controlling for the other one

Overall effects: compound tests (cont.)

Without interaction

> $H_0 : \tau_{KO} = 0$ (1 df) $H_0 : \tau_{P2} = \tau_{P6} = \tau_{P10} = \tau_{4W} = 0$ (on average!, 4 df)

```
tidy(anova(lm(gExp~gType+devStage,hitDat)))
```

```
## # A tibble: 3 x 6
##   term         df   sumsq  meansq statistic   p.value
##   <chr>     <int>   <dbl>   <dbl>     <dbl>     <dbl>
## 1 gType         1  0.0692  0.0692     0.924  3.44e- 1
## 2 devStage      4 71.0    17.8      237.     8.40e-24
## 3 Residuals    33  2.47    0.0749    NA     NA
```

Tests of overall effects of a factor controlling for the other one

The t-test in `lm` and the F-test (1 df) in `anova` for gType are not equivalent due to unbalancedness

Nested models

These examples are just special cases of nested models

For example: does development have a significant effect on gene expression?

Compare the models with and without `devStage`!!

Model 1: gExp~gType

Model 2: gExp~gType + devStage

Mathematically:

Model 1: $Y_{ijk} = \theta + \tau_{KO} x_{KO,ijk} + \varepsilon$

Model 2: $Y_{ijk} = \theta + \tau_{KO} x_{KO,ijk} + \tau_{P2} x_{P2,ijk} + \tau_{P6} x_{P6,ijk} + \tau_{P10} x_{P10,ijk} + \tau_{4W} x_{4W,ijk} + \varepsilon$

$$H_0 : \tau_{P2} = \tau_{P6} = \tau_{P10} = \tau_{4W} = 0$$

The $x_{DD,ijk}$ are dummy variables (see math handout)

More general!

F-test: selection of nested models

$$H_0 : \beta_{k+1} = \ldots = \beta_{k+p} = 0$$

$$F = \frac{\left(SS_{reduced} - SS_{full}\right)/p}{SS_{full}/(n-\underline{p-k-1})} \sim \mathcal{F}_{p,n-p-k-1}$$

Compares:

Model 1: $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i$ (reduced: **1+k** parameters)

versus

Nested models in R

```
addReduced<- lm(gExp ~ gType, data = hitDat)
addFull<- lm(gExp ~ gType+devStage, data = hitDat)
anova(addReduced,addFull)
```

```
## Analysis of Variance Table
##
## Model 1: gExp ~ gType
## Model 2: gExp ~ gType + devStage
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     37 73.498
## 2     33  2.473  4    71.024 236.92 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
tidy(anova(addFull))
```

```
## # A tibble: 3 x 6
##   term         df   sumsq  meansq statistic   p.value
##   <chr>     <int>   <dbl>   <dbl>     <dbl>     <dbl>
## 1 gType         1  0.0692  0.0692     0.924  3.44e- 1
## 2 devStage      4 71.0    17.8      237.     8.40e-24
## 3 Residuals    33  2.47    0.0749    NA     NA
```

Another special case: goodness of fit!

Compare the full *vs* the intercept-only models (compound test)!

$$H_0 : \tau_{KO} = \tau_{P2} = \tau_{P6} = \tau_{P10} = \tau_{4W} = 0,\ (5\text{ df})$$

```
## Analysis of Variance Table
##
## Model 1: gExp ~ 1
## Model 2: gExp ~ gType + devStage
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     38 73.567
## 2     33  2.473  5    71.094 189.72 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(addFull)$fstatistic # also given in the summary of lm
```

```
##    value    numdf    dendf
## 189.7238   5.0000  33.0000
```

All is valid in the model with interaction, try it!!

Summary

- ***t*-tests** can be used to test the equality of **2** population means
- **ANOVA** can be used to test the equality of **more than 2** population means simultaneously (main effects)
- **Linear regression** provides a general framework for modelling the relationship between a response and different type of explanatory variables
  - *t*-tests are used to test the significance of *simple effects* (*individual* coefficients)
  - *F*-tests are used to test the significance of *main effects* (*simultaneously* multiple coefficients)
- *F*-tests are used to compare nested models
  - e.g., *overall* effects or *goodness of fit*

WHY??

$$Y = \boxed{X}\alpha + \varepsilon$$

This gives us a VERY FLEXIBLE framework!!

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 1.22 \\ 1 & 2.02 \\ 1 & 1.42 \\ \vdots & \vdots \\ 1 & 1.89 \\ 1 & 2.01 \\ \vdots & \vdots \\ 1 & 1.56 \\ 1 & 2.17 \\ 1 & 1.51 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 1.22 & 0 \\ 1 & 0 & 2.02 & 0 \\ 1 & 0 & 1.42 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1.89 & 0 \\ 1 & 1 & 2.01 & 2.01 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1.56 & 1.56 \\ 1 & 1 & 2.17 & 2.17 \\ 1 & 1 & 1.51 & 1.51 \end{bmatrix}$$

1 categorical covariate — **2 categorical covariates** — **1 continuous covariate** — **1 continuous 1 categorical**

AND MANY MORE .....

Tip: ?model.matrix

Next class: linear models provides a general flexible framework to study the relation of a response with many variables!