Linear models

Differential Expression

Kasper D. Hansen

Fall 2023

Linear models

Linear models is a big subject in statistics and it can take a while to become familiar with them.

In genomics, we tend to use relatively simple designs.

"Linear" refers to the fact that the model is linear in its parameters.

Modeling the mean

To include covariates Z, we can propose a linear model like

$$E(X_g) = \beta Z$$

where Z is a matrix of covariates and β is a vector of unknown parameters (fold-changes).

This is a universal "language" for modeling the effect of covariates and it is useful to learn.

NOTE: a constant source of confusion is that different choices of Z (different parameterizations) can result in the same model. There is more than one way to encode your question.

A two group comparison

Let us say we have 2 groups with means

$$\mu_1 = 9, \quad \mu_2 = 5$$

There are really 3 key numbers here

1. The expression in group 1 ($\mu_2 = 9$)
2. The expression in group 2 ($\mu_2 = 5$)
3. The difference between the two groups (log fold-change): $\mu_1 - \mu_2 = 4$.

If you know 2 out of these 3 numbers you can always get the 3rd one.

Scientifically, we're usually interested in the difference.

A two group comparison, parametrizations

Parametrization 1 - group 1 has mean μ_1 - group 2 has mean μ_2 - To get the difference we form the **contrast** $\mu_1 - \mu_2$

Parametrization 2 - group 1 has mean θ_1 - group 2 has mean $\theta_1 + \theta_2$

In parametrization 2, we recognize that $\theta_1 = \mu_1$ and $\theta_2 = \mu_2 - \mu_1$. In other words, one of the two main parameters, directly represents the parameter of interest (the difference).

In code

```
groups = c("group1", "group1", group1", "group2, "group2", "group2")
model.matrix( ~ 0 + groups)  # Parametrization 1
model,matrix( ~ groups - 1)   # Parametrization 1
model.matrix( ~ groups)       # Parametrization 2
```

For the first two approaches, you would need a contrast matrix constructed like

```
contrasts = limma::makeContrasts(groupsgroup2 - groupsgroup1, levels = design2)
```

For the second approach, in practice, it can be hard to know whether the estimated number is `group1 - group2` or `group2 - group1`. This depends on the first level of the `groups` variable represented as a factor.

More variables

The real power comes when you're modeling more than 1 factor.

Example: treatment and genotype.

```
model.matrix( ~ genotype + treatment)
model.matrix( ~ genotype * treatment) # interaction
```