Design of “RNA-seq” Experiments

06/29/2015

Albert Lee

RabadanLab

Motivation

Motivation

- RNAseq is getting cheaper (~$600 per sample) and more sophisticated analyses are being/will be called for.

Motivation

- RNAseq is getting cheaper (~$600 per sample) and more sophisticated analyses are being/will be called for.
  - Multifactorial analysis (multiple treatments, multiple tissues , multiple time points)

Motivation

- RNAseq is getting cheaper (~$600 per sample) and more sophisticated analyses are being/will be called for.
  - Multifactorial analysis (multiple treatments, multiple tissues , multiple time points)
  - Repeated measures (more than one samples from the same individual)

Motivation

- RNAseq is getting cheaper (~$600 per sample) and more sophisticated analyses are being/will be called for.
  - Multifactorial analysis (multiple treatments, multiple tissues , multiple time points)
  - Repeated measures (more than one samples from the same individual)

Motivation

- RNAseq is getting cheaper (~$600 per sample) and more sophisticated analyses are being/will be called for.
  - Multifactorial analysis (multiple treatments, multiple tissues , multiple time points)
  - Repeated measures (more than one samples from the same individual)
- Currently, the majority of RNAseq tools heavily focus on comparing two classes

Motivation

- RNAseq is getting cheaper (~$600 per sample) and more sophisticated analyses are being/will be called for.
  - Multifactorial analysis (multiple treatments, multiple tissues , multiple time points)
  - Repeated measures (more than one samples from the same individual)
- Currently, the majority of RNAseq tools heavily focus on comparing two classes
  - edgeR & DESeq : Modified Fisher Exact test for NB distribution

Motivation

- RNAseq is getting cheaper (~$600 per sample) and more sophisticated analyses are being/will be called for.
  - Multifactorial analysis (multiple treatments, multiple tissues , multiple time points)
  - Repeated measures (more than one samples from the same individual)
- Currently, the majority of RNAseq tools heavily focus on comparing two classes
  - edgeR & DESeq : Modified Fisher Exact test for NB distribution
  - Cuffdiff : z-score based on the log-transformed ratio of expression divided by the variance of the transformed ratio

Motivation

- RNAseq is getting cheaper (~$600 per sample) and more sophisticated analyses are being/will be called for.
  - Multifactorial analysis (multiple treatments, multiple tissues , multiple time points)
  - Repeated measures (more than one samples from the same individual)
- Currently, the majority of RNAseq tools heavily focus on comparing two classes
  - edgeR & DESeq : Modified Fisher Exact test for NB distribution
  - Cuffdiff : z-score based on the log-transformed ratio of expression divided by the variance of the transformed ratio

Motivation

- RNAseq is getting cheaper (~$600 per sample) and more sophisticated analyses are being/will be called for.
  - Multifactorial analysis (multiple treatments, multiple tissues , multiple time points)
  - Repeated measures (more than one samples from the same individual)
- Currently, the majority of RNAseq tools heavily focus on comparing two classes
  - edgeR & DESeq : Modified Fisher Exact test for NB distribution
  - Cuffdiff : z-score based on the log-transformed ratio of expression divided by the variance of the transformed ratio
- One-factor-at-a-time method is ineffective.
-

Motivation

- RNAseq is getting cheaper (~$600 per sample) and more sophisticated analyses are being/will be called for.
  - Multifactorial analysis (multiple treatments, multiple tissues , multiple time points)
  - Repeated measures (more than one samples from the same individual)
- Currently, the majority of RNAseq tools heavily focus on comparing two classes
  - edgeR & DESeq : Modified Fisher Exact test for NB distribution
  - Cuffdiff : z-score based on the log-transformed ratio of expression divided by the variance of the transformed ratio
- One-factor-at-a-time method is ineffective.
-

Motivation

- RNAseq is getting cheaper (~$600 per sample) and more sophisticated analyses are being/will be called for.
  - Multifactorial analysis (multiple treatments, multiple tissues , multiple time points)
  - Repeated measures (more than one samples from the same individual)
- Currently, the majority of RNAseq tools heavily focus on comparing two classes
  - edgeR & DESeq : Modified Fisher Exact test for NB distribution
  - Cuffdiff : z-score based on the log-transformed ratio of expression divided by the variance of the transformed ratio
- One-factor-at-a-time method is ineffective.
- Multifactorial options in edgeR/DESeq/Limma-Voom may be difficult to understand in the beginning

Aim

- Goal is to identify a subset of genes ranked by *interestingness* while accounting for the structure of the experimental design

Aim

- Goal is to identify a subset of genes ranked by *interestingness* while accounting for the structure of the experimental design

Example questions

Aim

- Goal is to identify a subset of genes ranked by *interestingness* while accounting for the structure of the experimental design

Example questions

- (Level 1) What are the genes that are differentially expressed in tumor vs normal?

Aim

- Goal is to identify a subset of genes ranked by *interestingness* while accounting for the structure of the experimental design

Example questions

- (Level 1) What are the genes that are differentially expressed in tumor vs normal?
- (Level 2) What are the genes that are differentially expressed in tumor vs normal while controlling for batch effects?

Aim

- Goal is to identify a subset of genes ranked by *interestingness* while accounting for the structure of the experimental design

Example questions

- (Level 1) What are the genes that are differentially expressed in tumor vs normal?
- (Level 2) What are the genes that are differentially expressed in tumor vs normal while controlling for batch effects?
- (Level 3) What are the genes that are differentially expressed only in testis and not in other tissues?

Aim

- Goal is to identify a subset of genes ranked by *interestingness* while accounting for the structure of the experimental design

Example questions

- (Level 1) What are the genes that are differentially expressed in tumor vs normal?
- (Level 2) What are the genes that are differentially expressed in tumor vs normal while controlling for batch effects?
- (Level 3) What are the genes that are differentially expressed only in testis and not in other tissues?
- (Level 3) What are the genes that are differentially expressed between time 3 and time2 in drugged samples while controlling for vehicle effects?

There are many problems in RNAseq, but…

There are many problems in RNAseq, but…

Let’s assume that:

There are many problems in RNAseq, but…

Let’s assume that:

1. All heavy biological & computational work(sequencing, preprocessing, alignment, and counting) have been done.

There are many problems in RNAseq, but…

Let’s assume that:

1. All heavy biological & computational work(sequencing, preprocessing, alignment, and counting) have been done.

data generation problem

There are many problems in RNAseq, but…

Let’s assume that:

1. All heavy biological & computational work(sequencing, preprocessing, alignment, and counting) have been done.
2. Expression measurement follows normal distribution.

data generation problem

There are many problems in RNAseq, but…

Let’s assume that:

1. All heavy biological & computational work(sequencing, preprocessing, alignment, and counting) have been done.
2. Expression measurement follows normal distribution.

data generation problem

non-normality problem

There are many problems in RNAseq, but…

Let’s assume that:

1. All heavy biological & computational work(sequencing, preprocessing, alignment, and counting) have been done.
2. Expression measurement follows normal distribution.
3. We have a large size of experimental units(+100 samples).

data generation problem

non-normality problem

There are many problems in RNAseq, but…

Let’s assume that:

1. All heavy biological & computational work(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem

There are many problems in RNAseq, but…

Let’s assume that:

1. All heavy biological & computational work(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem
4. Normalization is taken care of.

There are many problems in RNAseq, but…

Let’s assume that:

1. All heavy biological & computational work(sequencing, preprocessing, alignment, and counting) have been done.
2. Expression measurement follows normal distribution.
3. We have a large size of experimental units(+100 samples).
4. Normalization is taken care of.

data generation problem

non-normality problem

small replicate size problem

normalization problem

There are many problems in RNAseq, but…

Let’s assume that:

1. All heavy biological & computational work(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem
4. Normalization is taken care of. — normalization problem
5. Expression measurements are collected at gene level.

There are many problems in RNAseq, but…

Let’s assume that:

1. All heavy biological & computational work(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem
4. Normalization is taken care of. — normalization problem
5. Expression measurements are collected at gene level. — Isoform problem

Level 1

Level 1

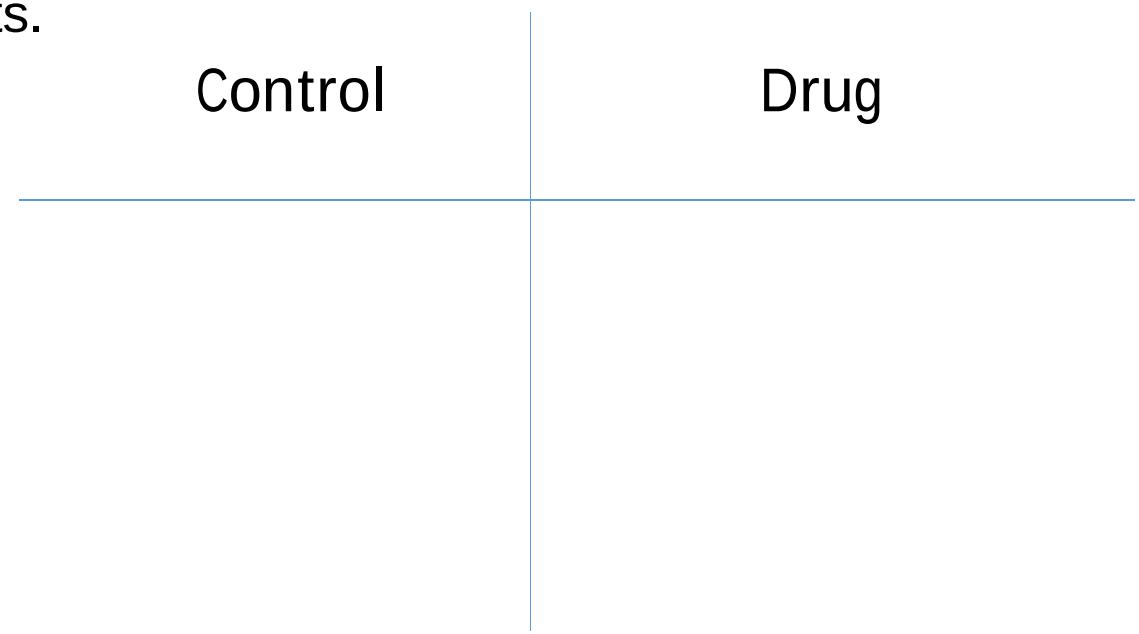
“What are the genes that are differentially expressed in tumor vs normal?”

1. Completely randomized design

- Completely randomized design:
  - the simplest experimental design.
  - With this design, experimental units are randomly assigned to treatments.

1. Completely randomized design

- Completely randomized design:
  - the simplest experimental design.
  - With this design, experimental units are randomly assigned to treatments.



1. Completely randomized design

- Completely randomized design:
  - the simplest experimental design.
  - With this design, experimental units are randomly assigned to treatments.

Control	Drug
	

1. Completely randomized design

- Completely randomized design:
  - the simplest experimental design.
  - With this design, experimental units are randomly assigned to treatments.

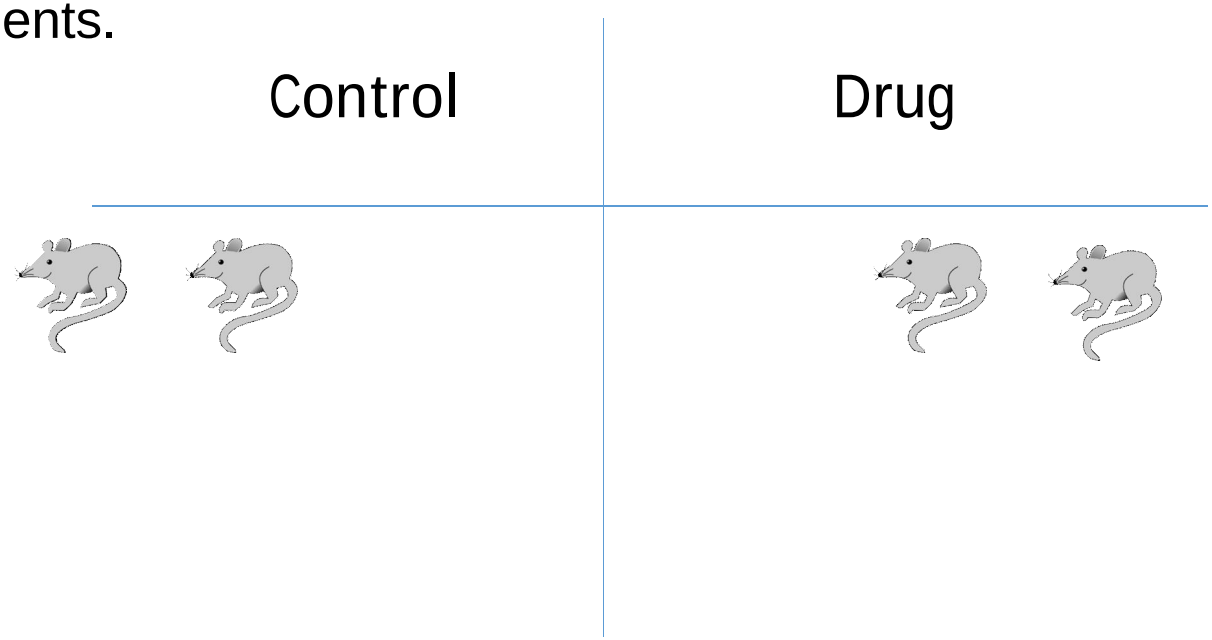
Control	Drug
	

1. Completely randomized design

- Completely randomized design:
  - the simplest experimental design.
  - With this design, experimental units are randomly assigned to treatments.

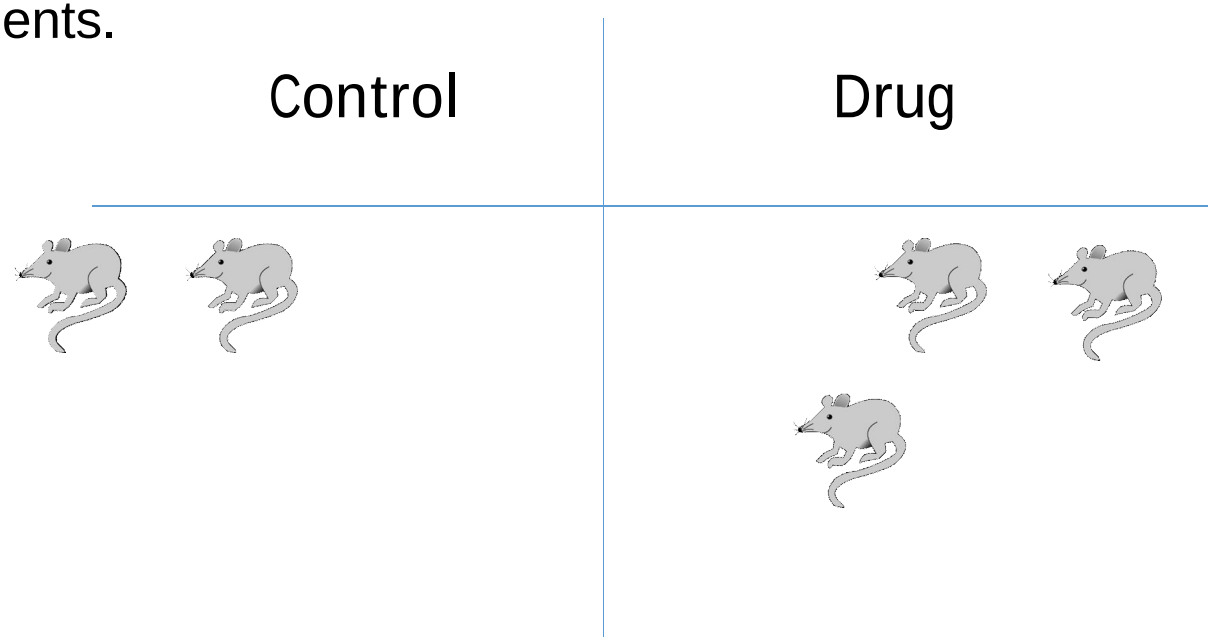
1. Completely randomized design

- Completely randomized design:
  - the simplest experimental design.
  - With this design, experimental units are randomly assigned to treatments.



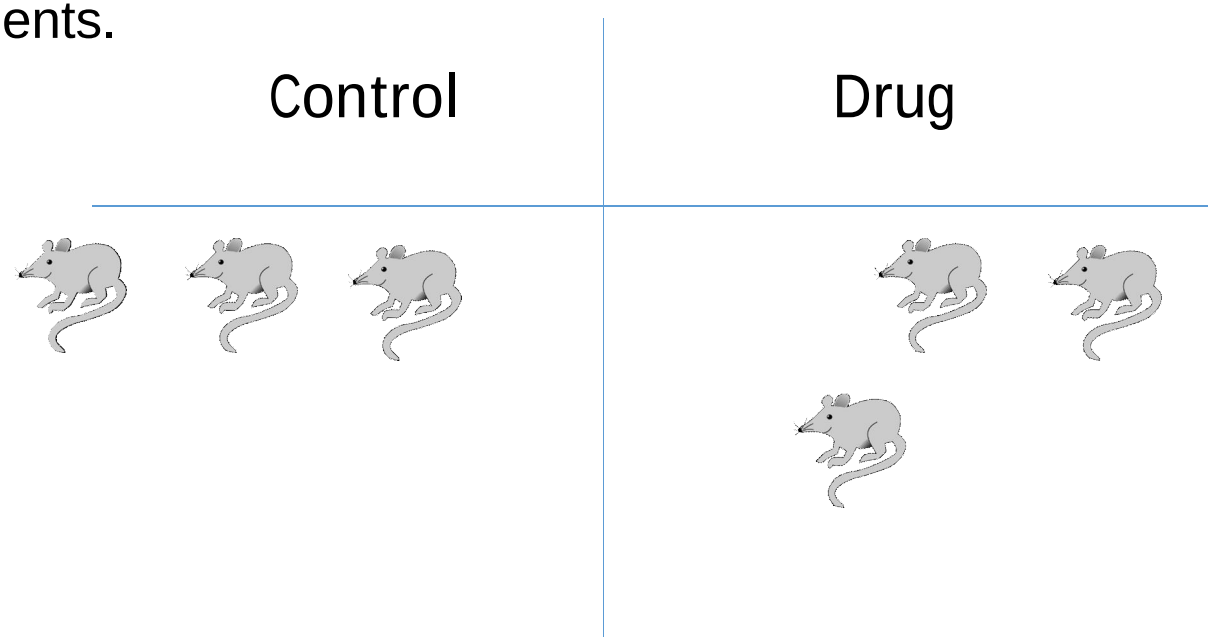
1. Completely randomized design

- Completely randomized design:
  - the simplest experimental design.
  - With this design, experimental units are randomly assigned to treatments.



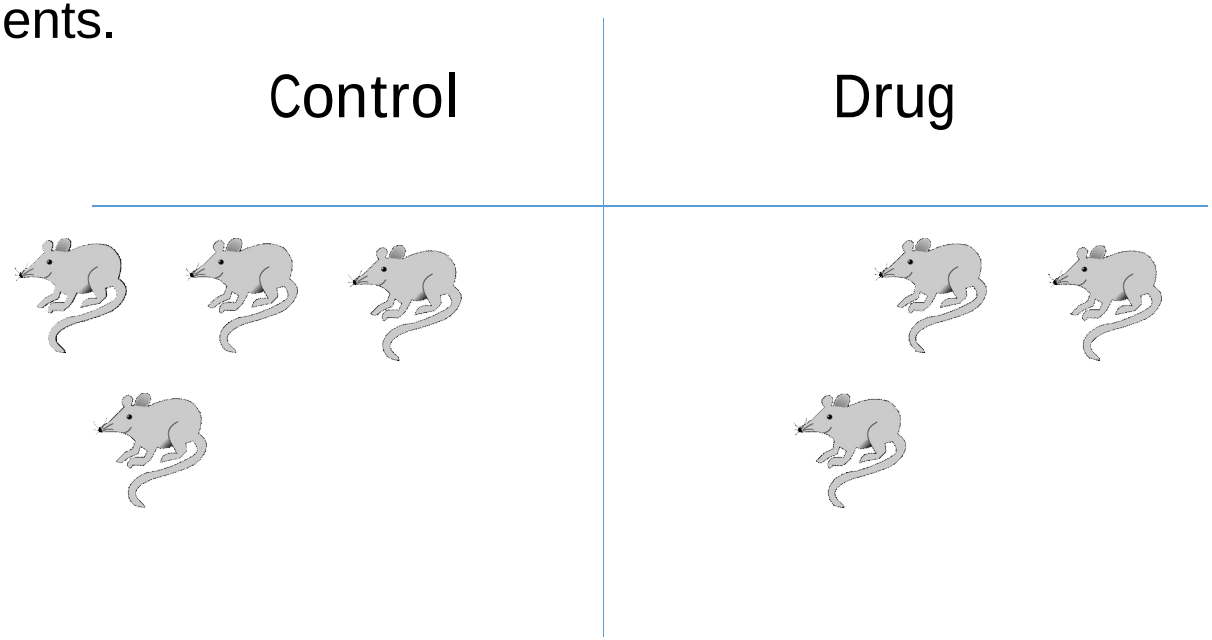
1. Completely randomized design

- Completely randomized design:
  - the simplest experimental design.
  - With this design, experimental units are randomly assigned to treatments.



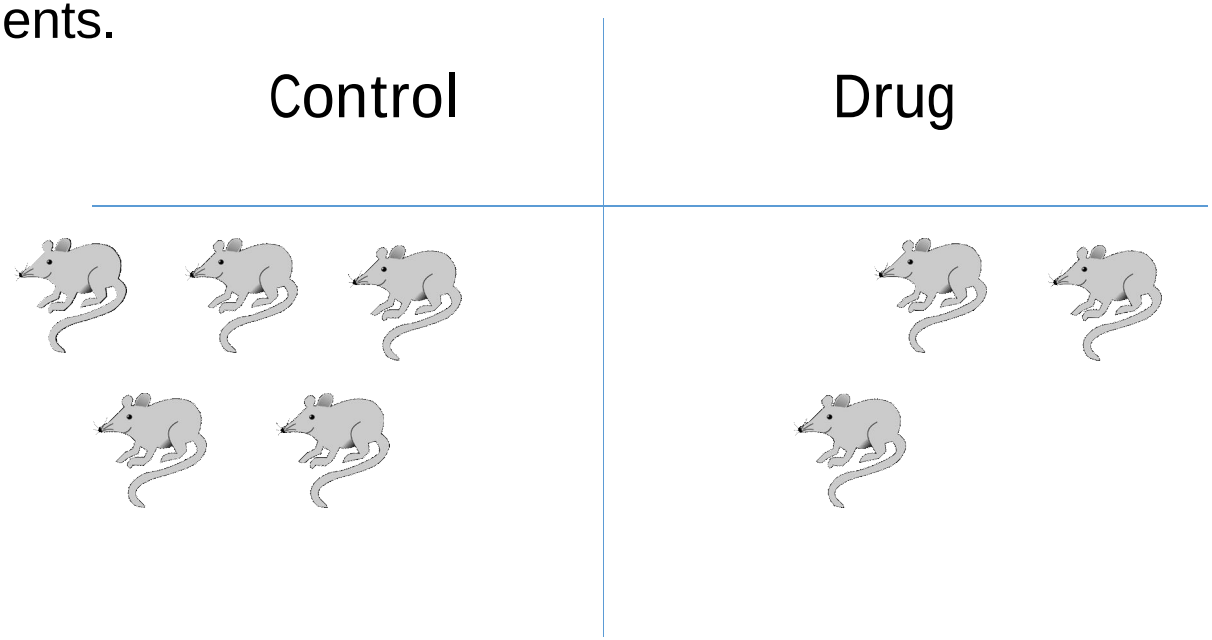
1. Completely randomized design

- Completely randomized design:
  - the simplest experimental design.
  - With this design, experimental units are randomly assigned to treatments.



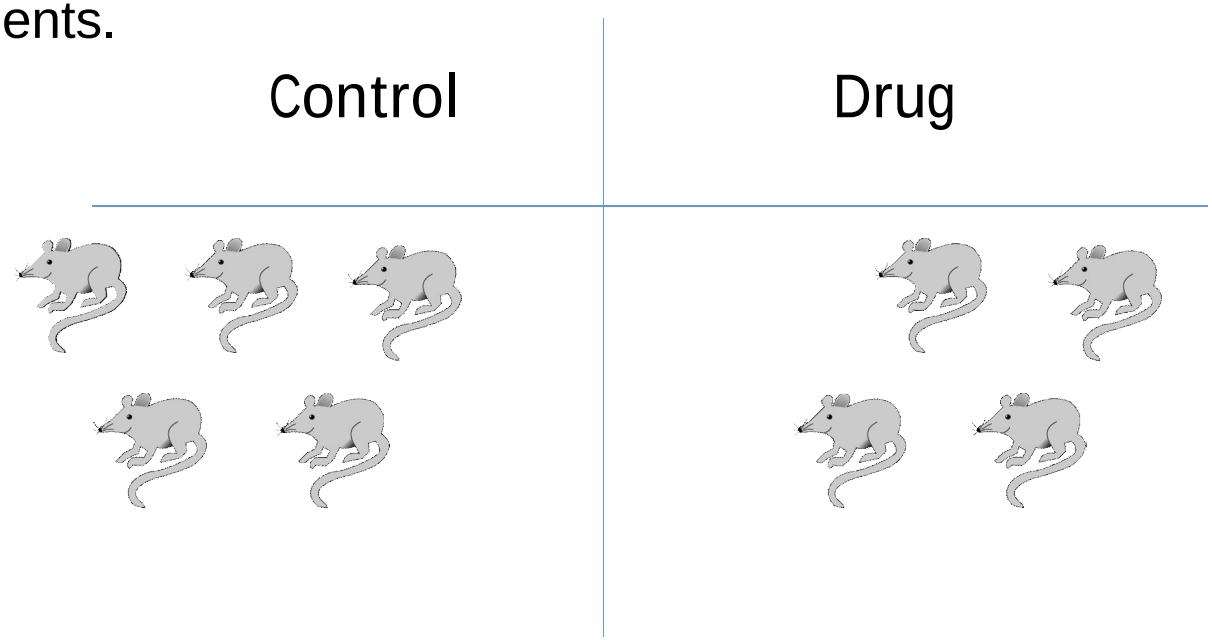
1. Completely randomized design

- Completely randomized design:
  - the simplest experimental design.
  - With this design, experimental units are randomly assigned to treatments.



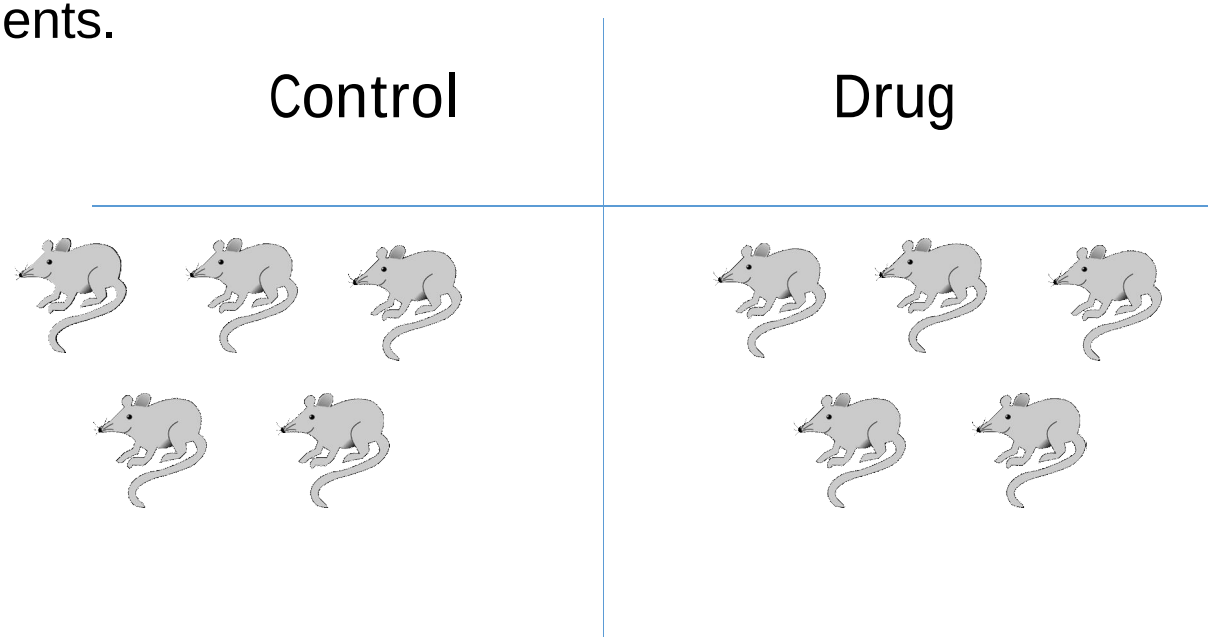
1. Completely randomized design

- Completely randomized design:
  - the simplest experimental design.
  - With this design, experimental units are randomly assigned to treatments.



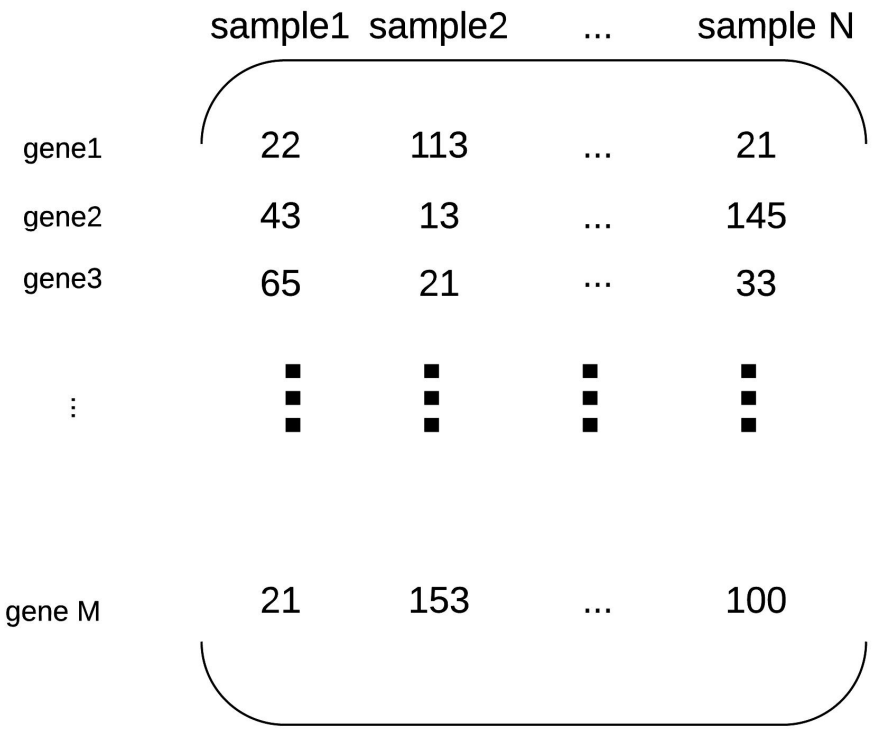
1. Completely randomized design

- Completely randomized design:
  - the simplest experimental design.
  - With this design, experimental units are randomly assigned to treatments.

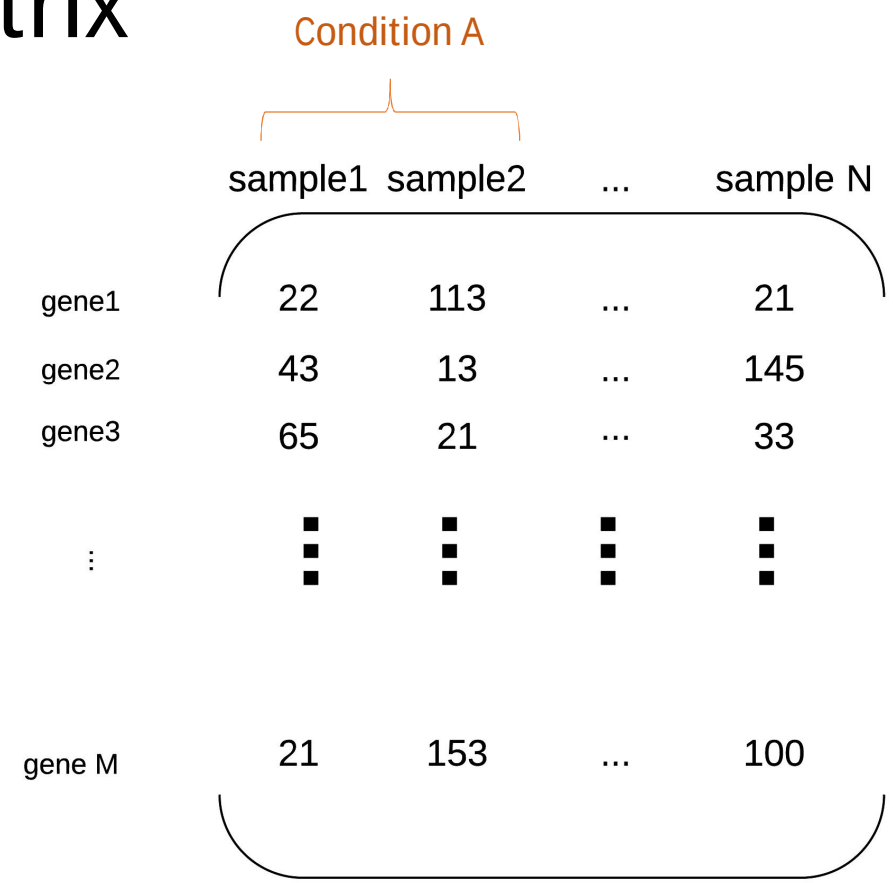


The typical matrix
you will see

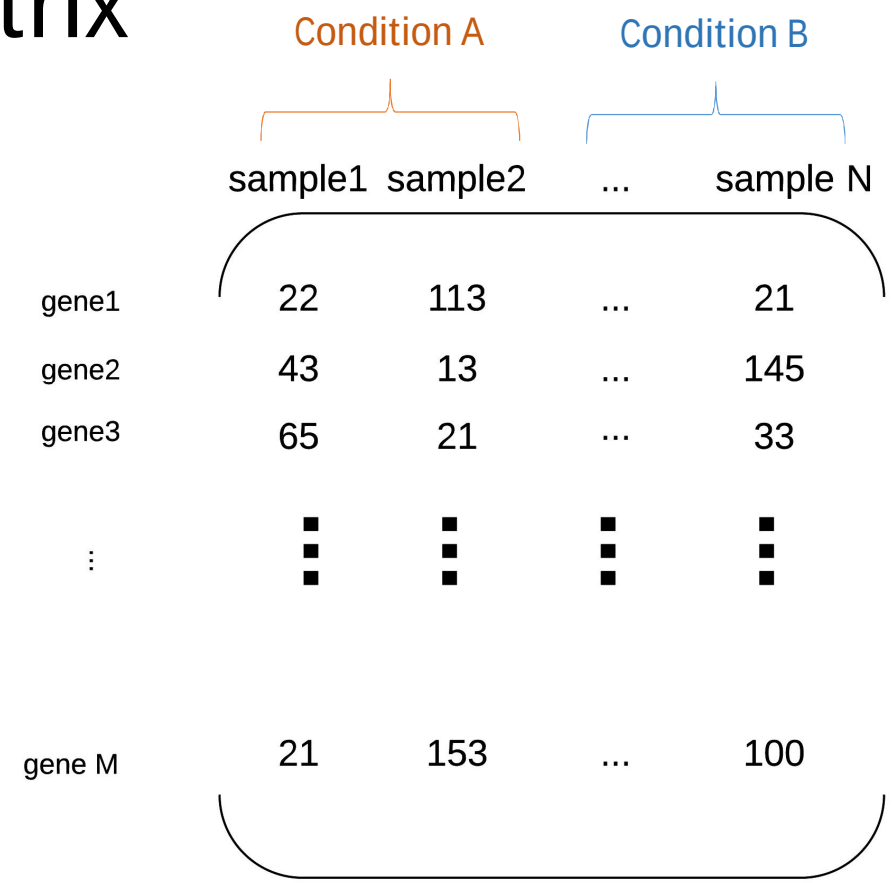
The typical matrix you will see



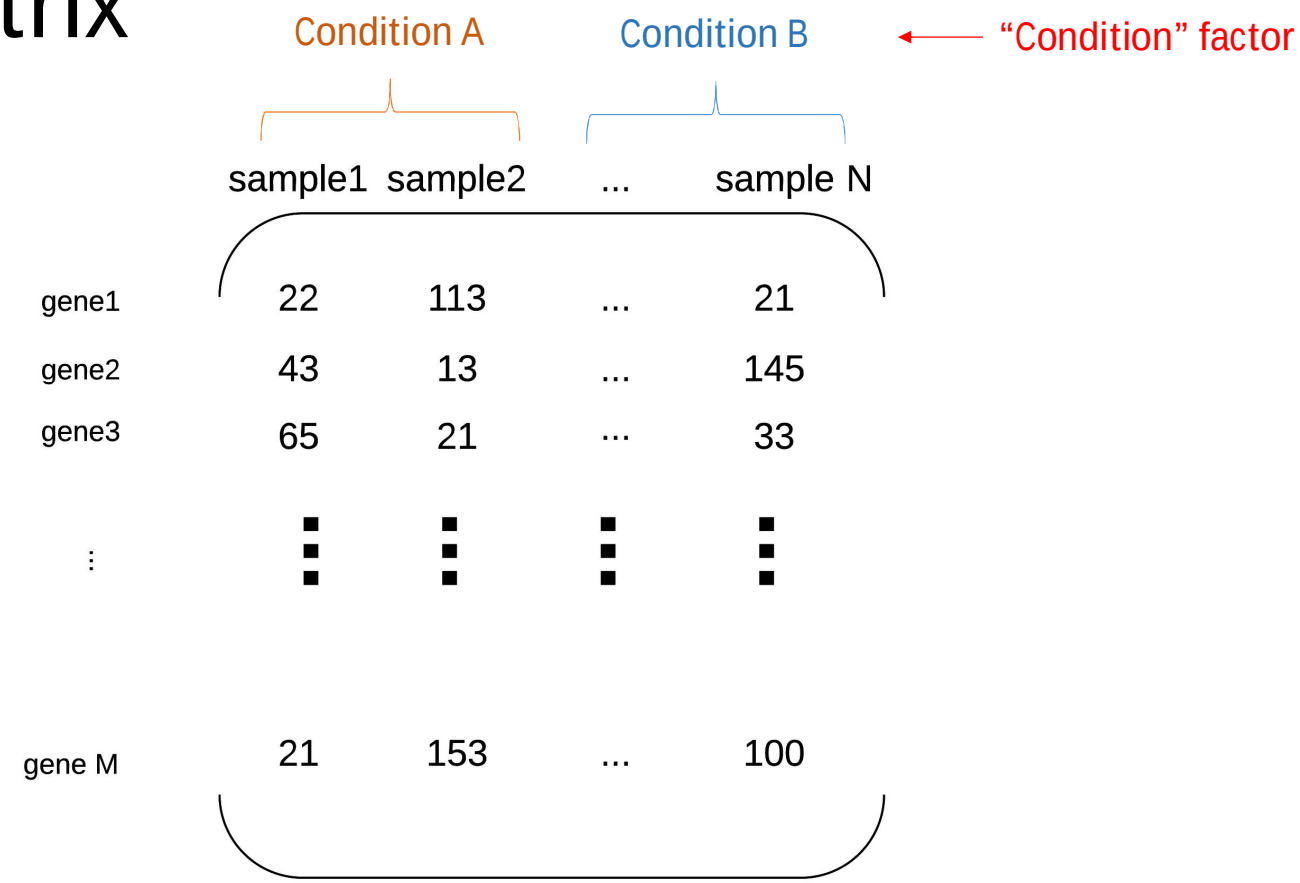
The typical matrix you will see



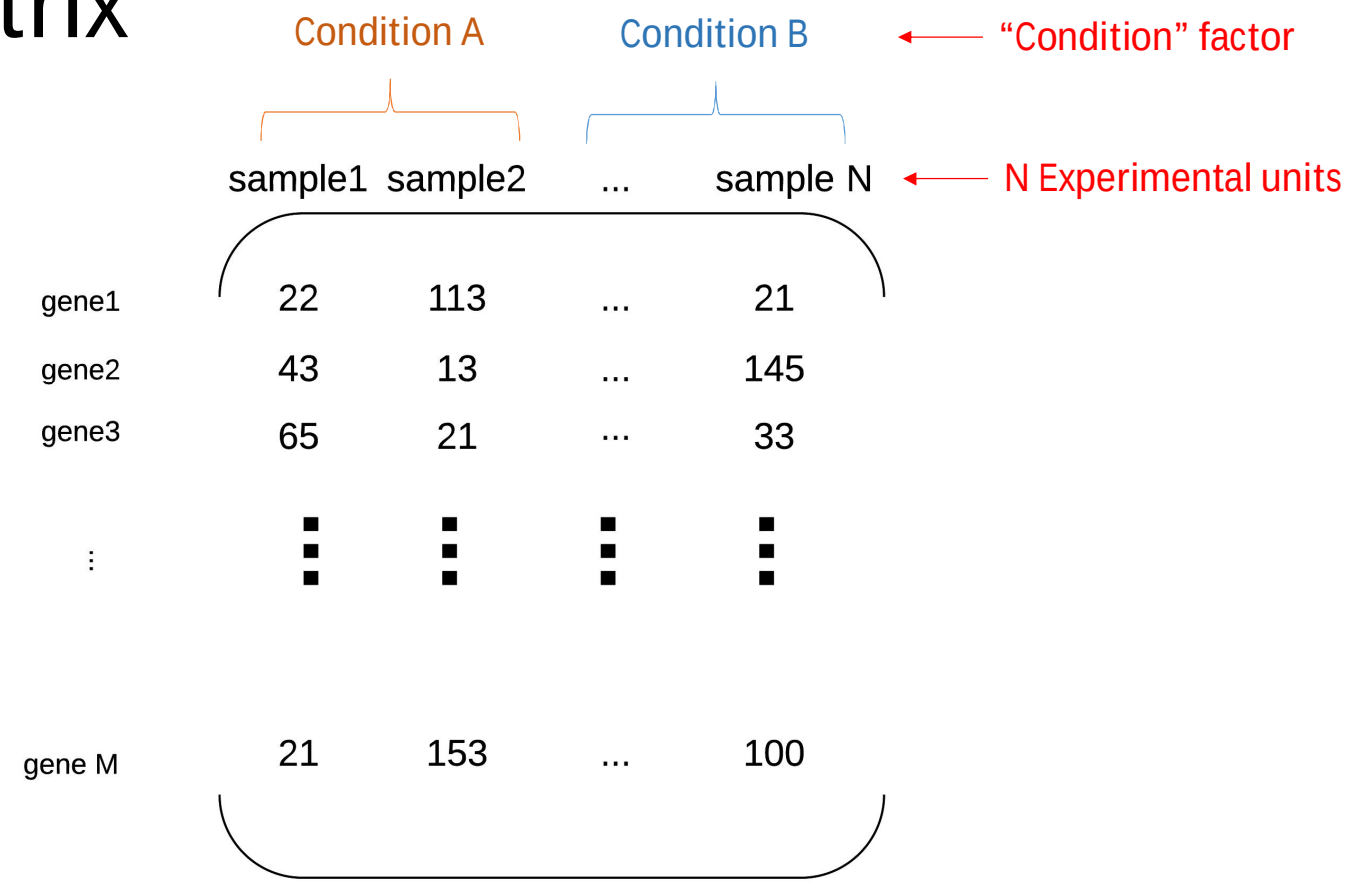
The typical matrix you will see



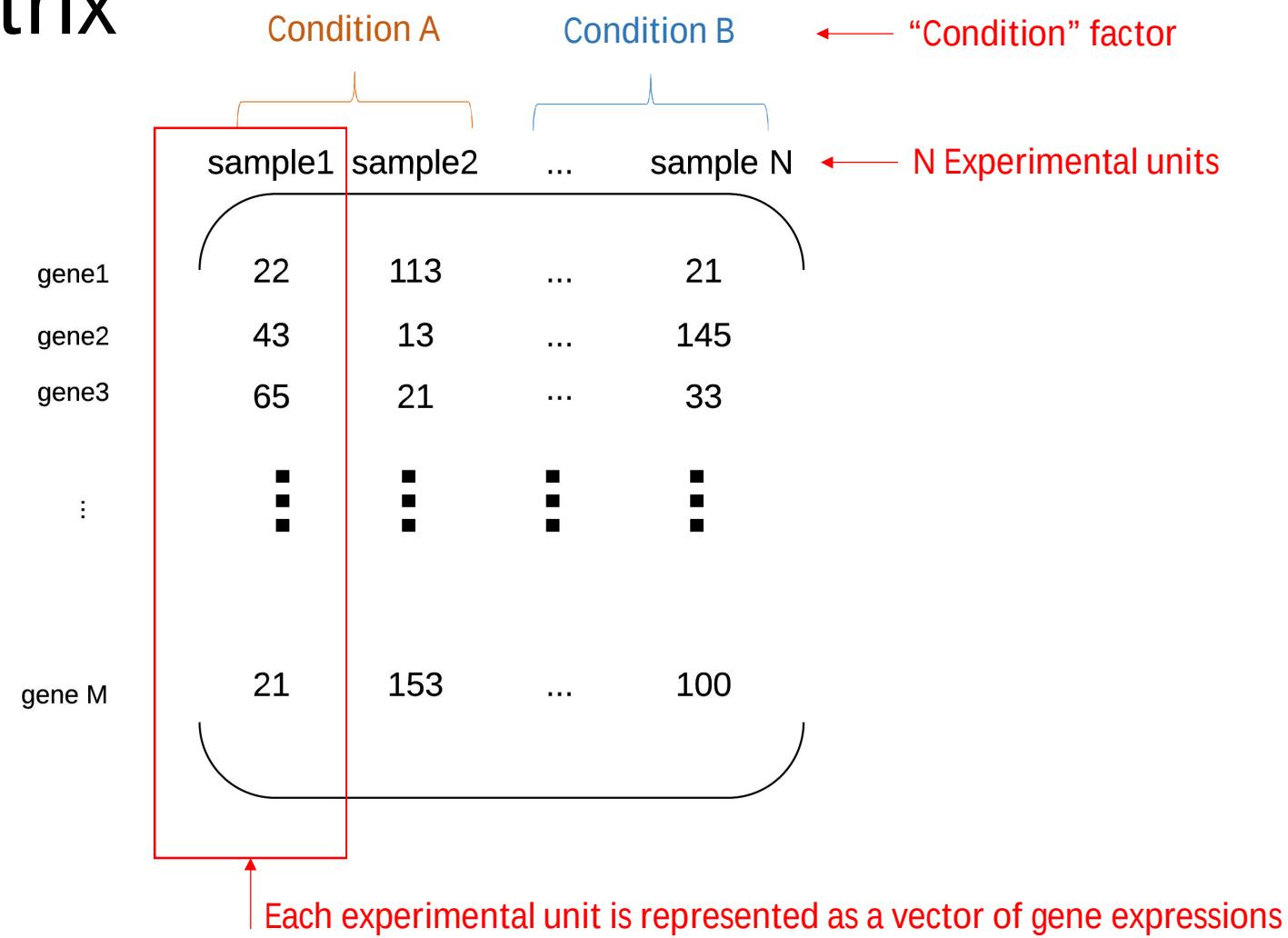
The typical matrix you will see



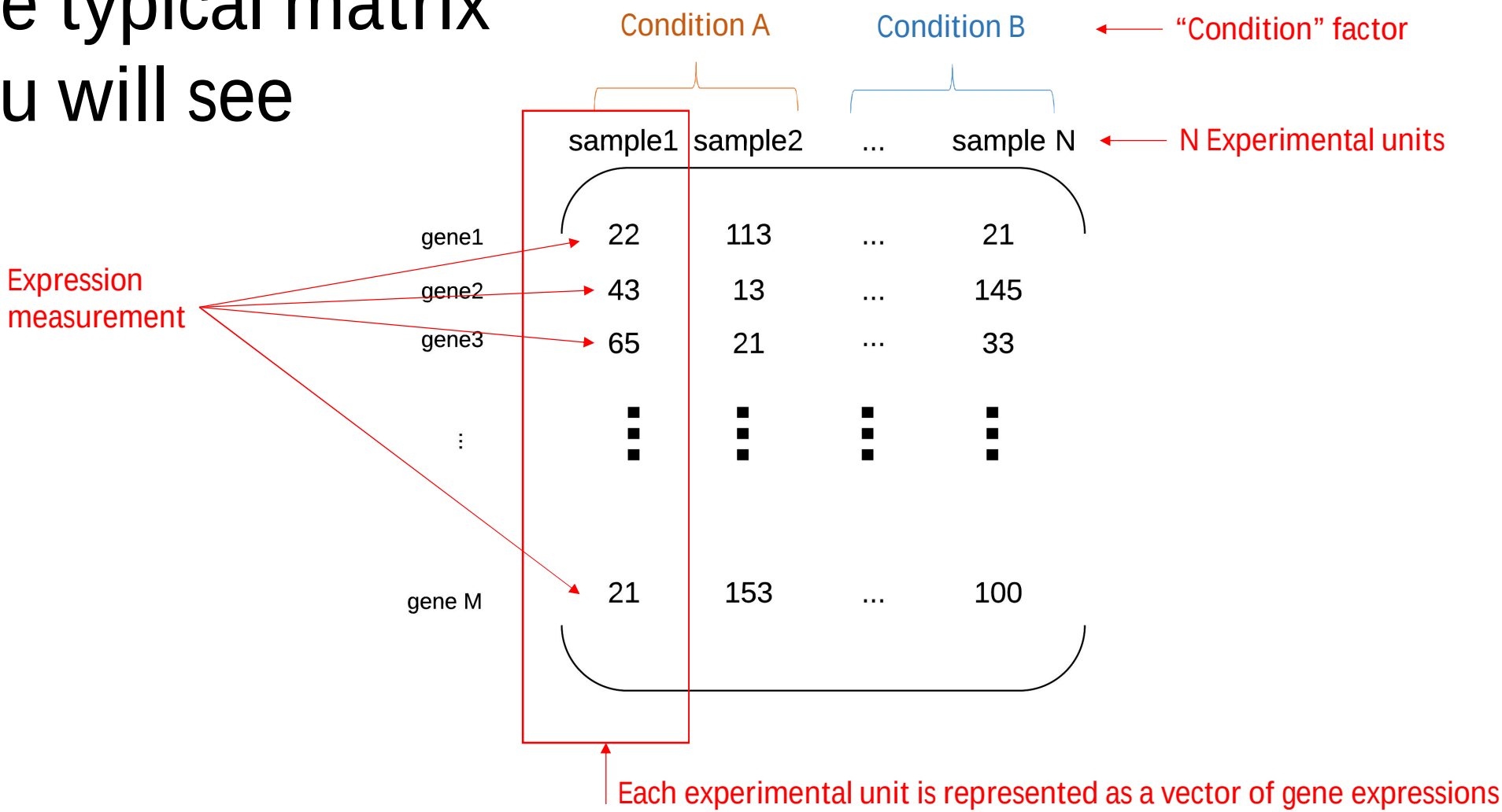
The typical matrix you will see



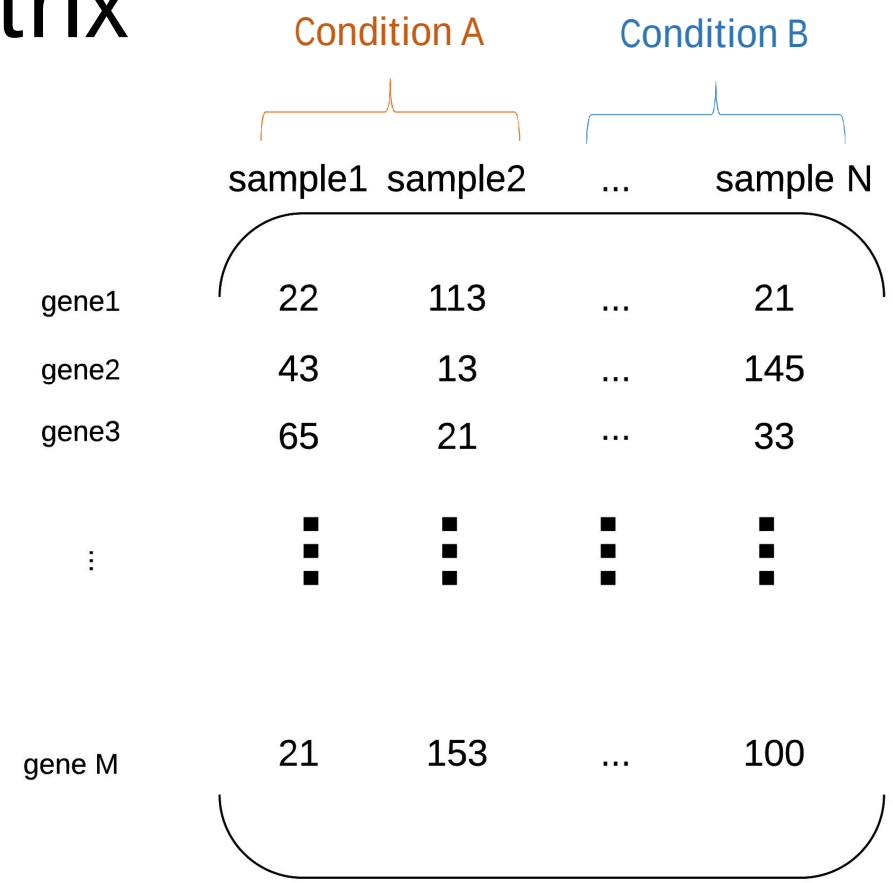
The typical matrix you will see



The typical matrix you will see



The typical matrix you will see



One thing to keep in mind

One thing to keep in mind

One thing to keep in mind

“In designing a microarray experiment,

One thing to keep in mind

or RNA-seq

“In designing a microarray experiment,

One thing to keep in mind

or RNA-seq

"In designing a microarray experiment,

we should concentrate on getting it right for *one gene*.

One thing to keep in mind

or RNA-seq

"In designing a microarray experiment,

we should concentrate on getting it right for *one gene*.

As the other 53, 999 data points are measured on subsamples of the experimental unit(!),

One thing to keep in mind

or RNA-seq

"In designing a microarray experiment,

we should concentrate on getting it right for *one gene*.

As the other 53, 999 data points are measured on subsamples of the experimental unit(!),

they have no bearing on constructing a good design."

One thing to keep in mind

or RNA-seq

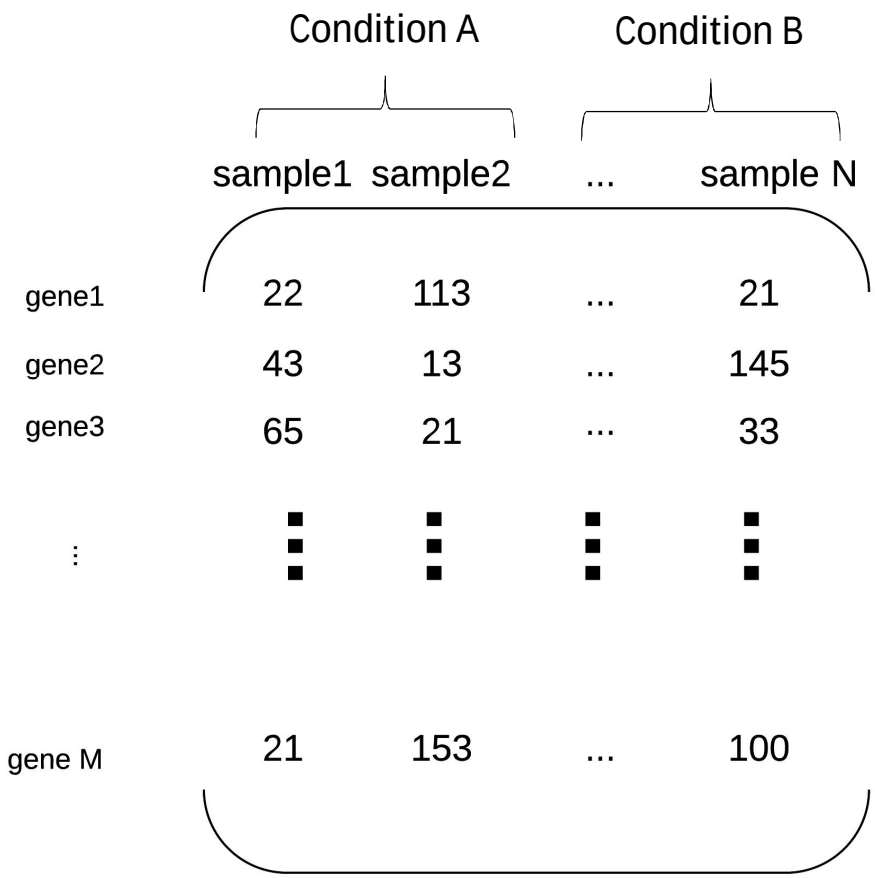
"In designing a microarray experiment,

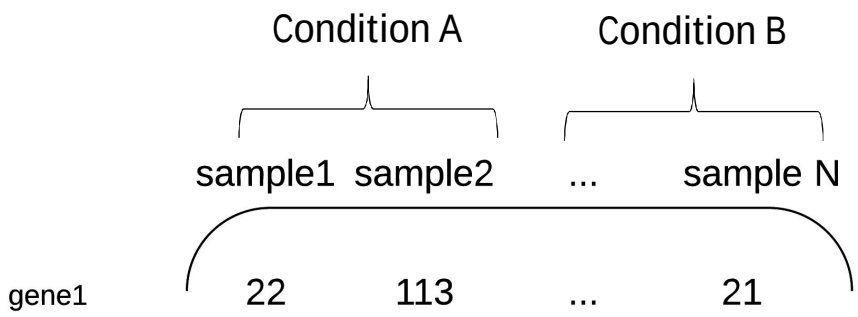
we should concentrate on getting it right for *one gene*.

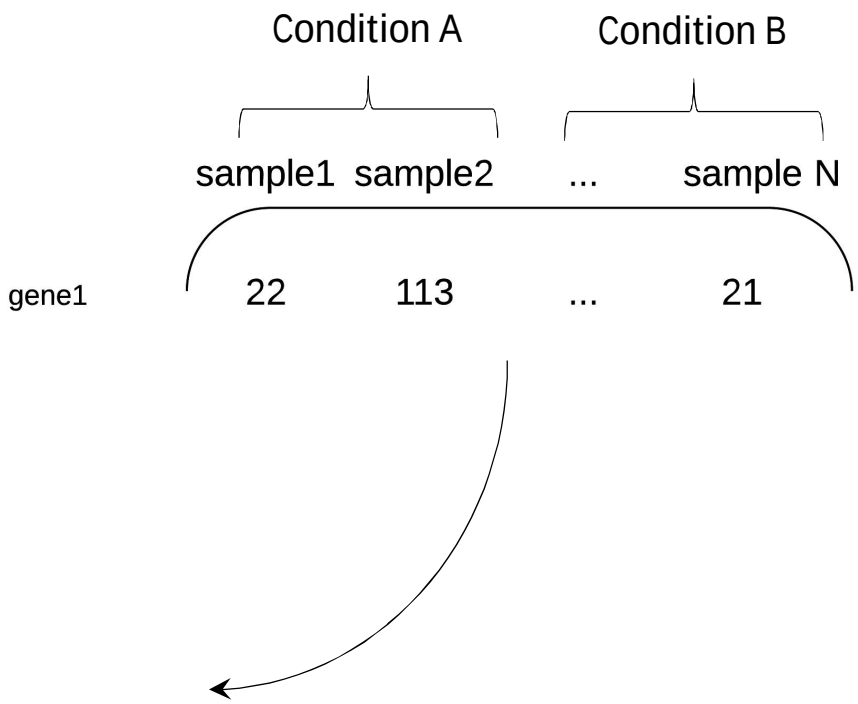
As the other 53, 999 data points are measured on subsamples of the experimental unit(!),

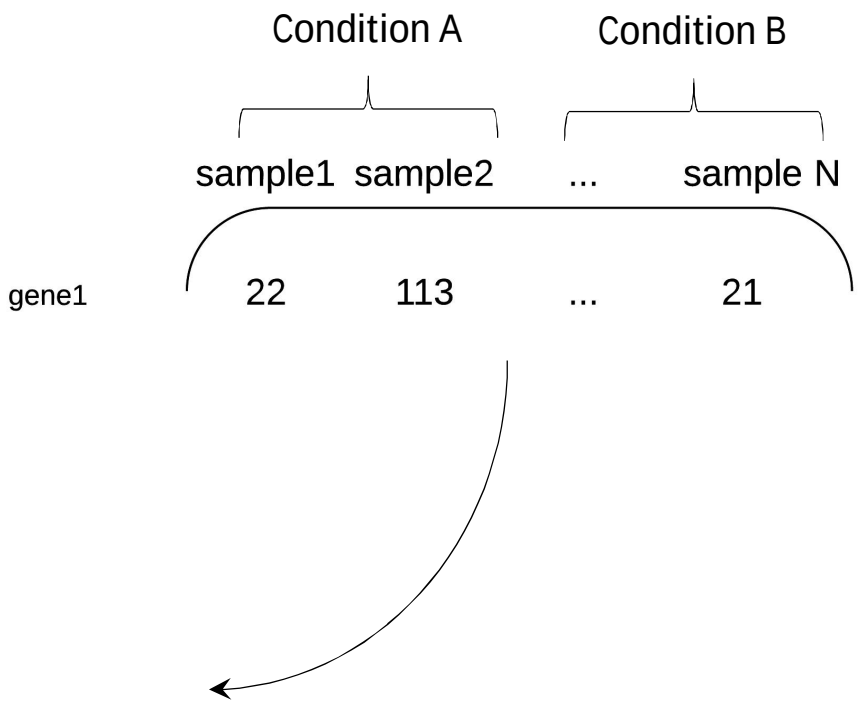
they have no bearing on constructing a good design."

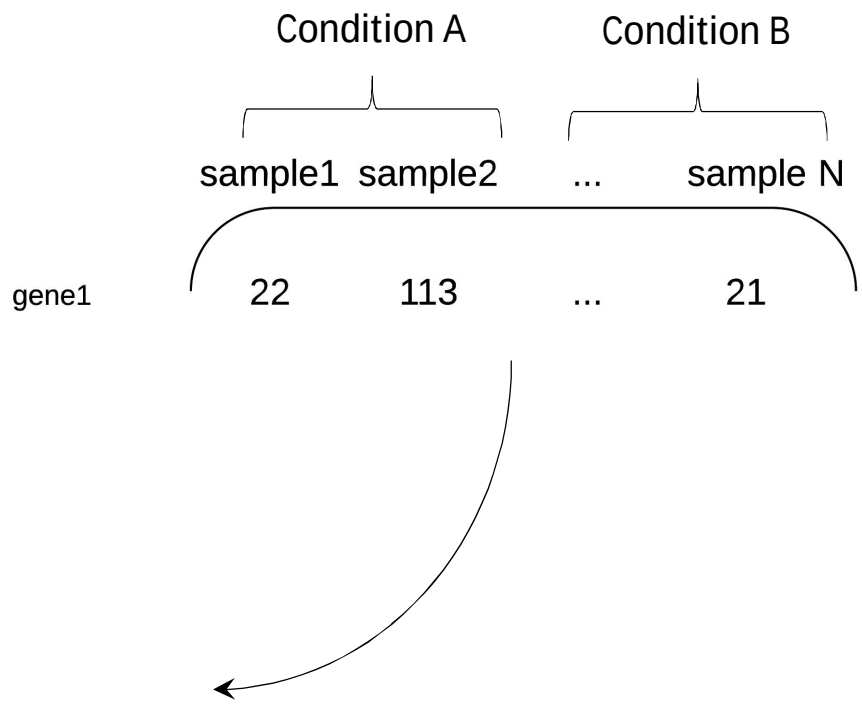
-- George Casella, "Statistical Design", Springer 2008


Condition A
Condition B
sample1
sample2
...
sample N
gene1
22
113
...
21
gene2
43
13
...
145
gene3
65
21
...
33
gene M
21
153
...
100

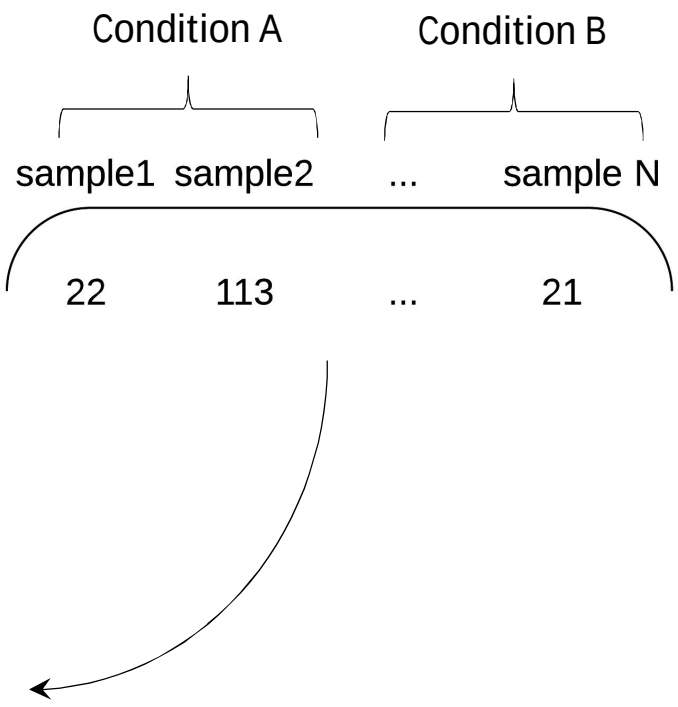

Condition A
Condition B
sample1
sample2
...
sample N
gene1
22
113
...
21


Condition A
Condition B
sample1
sample2
...
sample N
gene1
22
113
...
21


Condition A
Condition B
sample1
sample2
...
sample N
gene1
22
113
...
21

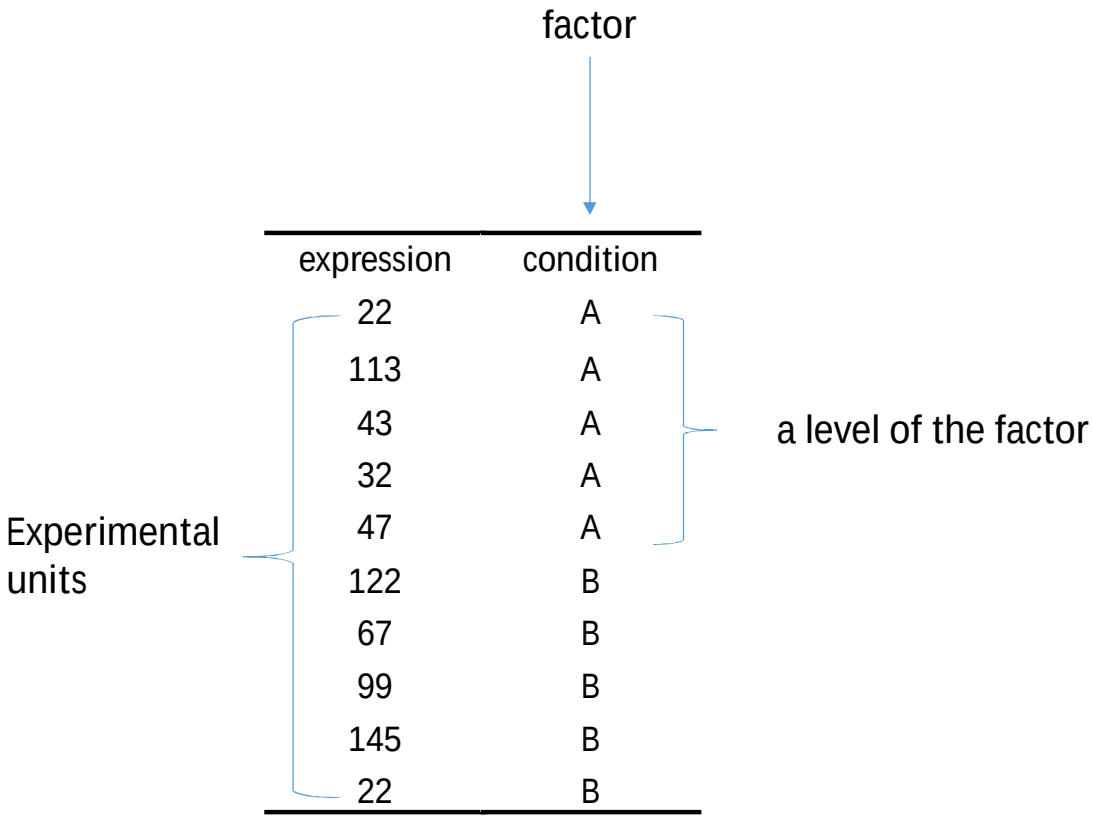


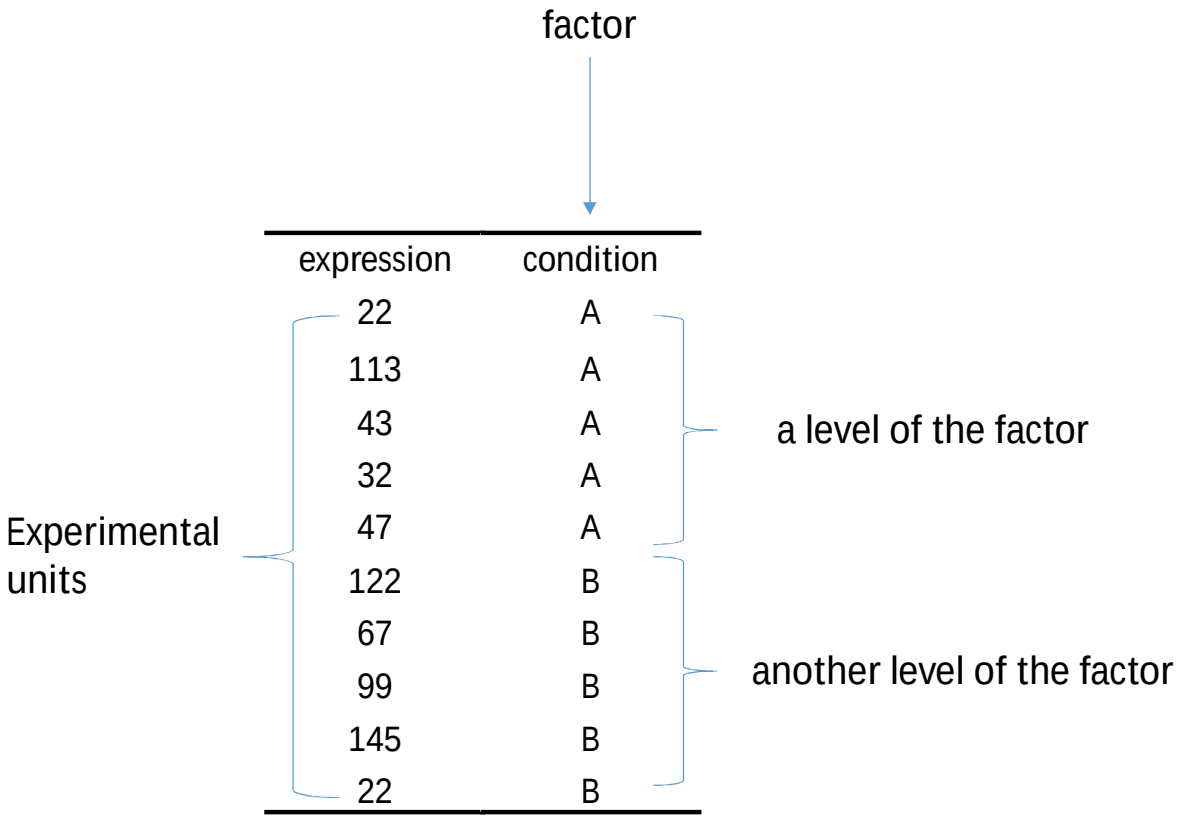
expression
22
113
43
32
47
122
67
99
145
22



expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B


factor
expression
condition
22
A
113
A
43
A
32
A
47
A
122
B
67
B
99
B
145
B
22
B
Experimental units


factor
expression
condition
22
A
113
A
43
A
32
A
47
A
122
B
67
B
99
B
145
B
22
B
a level of the factor
Experimental units


factor
expression
condition
22
A
113
A
43
A
32
A
47
A
122
B
67
B
99
B
145
B
22
B
Experimental units
a level of the factor
another level of the factor

You can compute
sample means and sample standard deviations for two groups

expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

You can compute
sample means and sample standard deviations for two groups

expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

You can compute
sample means and sample standard deviations for two groups

expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

$\bar{X}_A, S_A$

You can compute
sample means and sample standard deviations for two groups

expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

$\bar{X}_A, S_A$

You can compute
sample means and sample standard deviations for two groups

expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

$\bar{X}_A, S_A$

$\bar{X}_B, S_B$

You can compute
sample means and sample standard deviations for two groups

expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

$\bar{X}_A, S_A$

$\bar{X}_B, S_B$

Null Hypothesis : $H_o : \mu_A = \mu_B$

You can compute
sample means and sample standard deviations for two groups

expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

$\bar{X}_A, S_A$

$\bar{X}_B, S_B$

Null Hypothesis : $H_o : \mu_A = \mu_B$

Alternative Hypothesis : $H_A : \mu_A \neq \mu_B$

You can compute
sample means and sample standard deviations for two groups

expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

$\bar{X}_A, S_A$

$\bar{X}_B, S_B$

Null Hypothesis : $H_o : \mu_A = \mu_B$

Alternative Hypothesis : $H_A : \mu_A \neq \mu_B$

Compute test-statistic : $\dfrac{\bar{X}_A - \bar{X}_B}{\sqrt{\dfrac{S_A}{n_A} + \dfrac{S_B}{n_B}}} = t$

You can compute
sample means and sample standard deviations for two groups

expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

$\bar{X}_A, S_A$

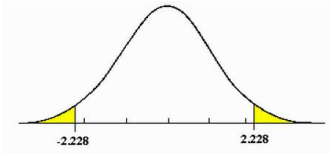
$\bar{X}_B, S_B$

Null Hypothesis : $H_o : \mu_A = \mu_B$

Alternative Hypothesis : $H_A : \mu_A \neq \mu_B$

Compute test-statistic : $$\frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{S_A}{n_A} + \frac{S_B}{n_B}}} = t$$

Compare with critical value & compute p value



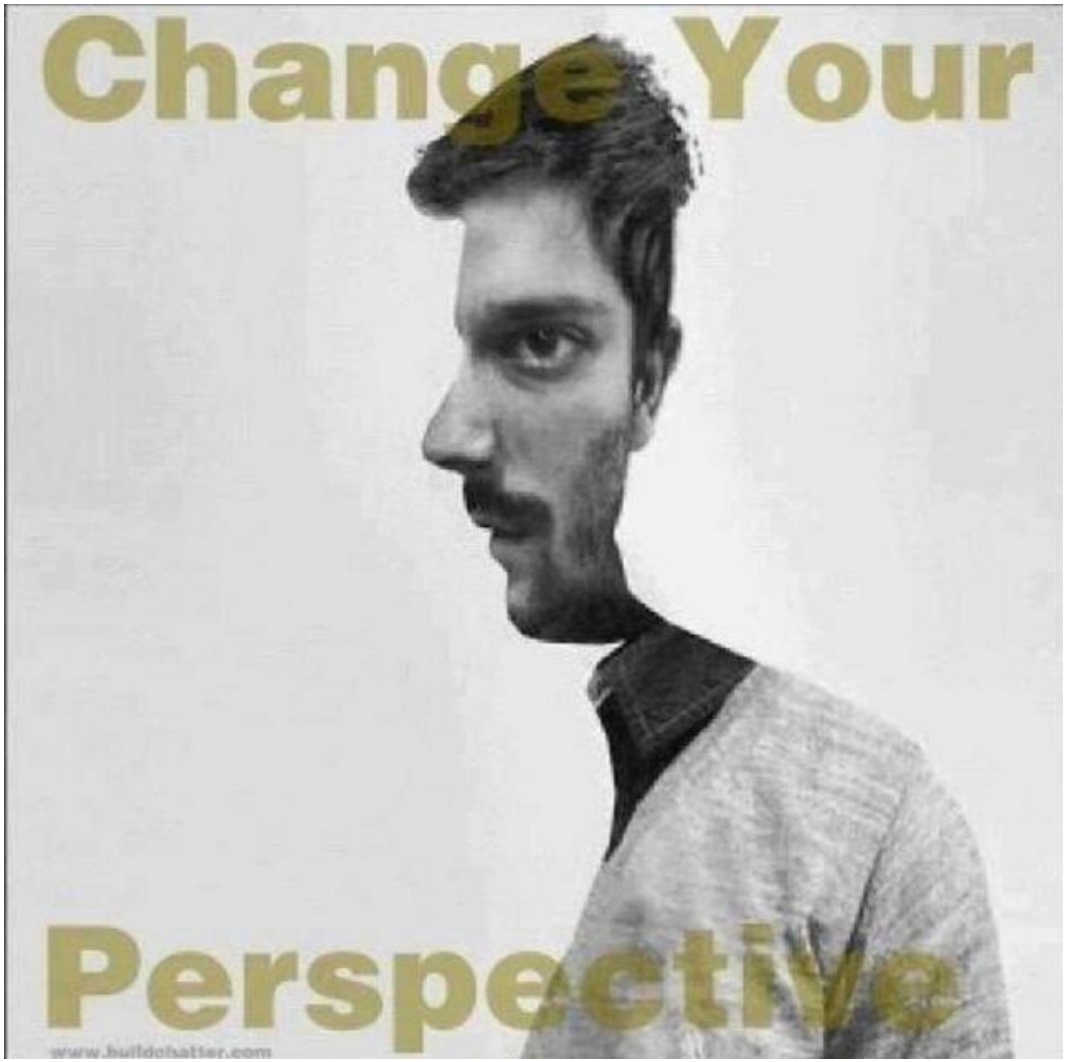


Image source: http://www.beingencouraged.com/

You can describe the same thing with linear modeling

Expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

You can describe the same thing with linear modeling

Expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

You can describe the same thing with linear modeling

Expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

You can describe the same thing with linear modeling

Expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

Instead of focusing comparison of two groups,

You can describe the same thing with linear modeling

Expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

You can describe the same thing with linear modeling

Expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

You can describe the same thing with linear modeling

Expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

You can describe the same thing with linear modeling

Expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

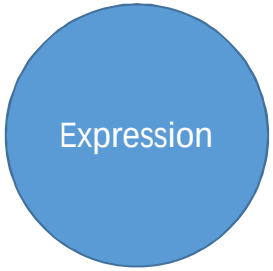
Focus on modeling the expression as a function of factors

You can describe the same thing with linear modeling

Expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B

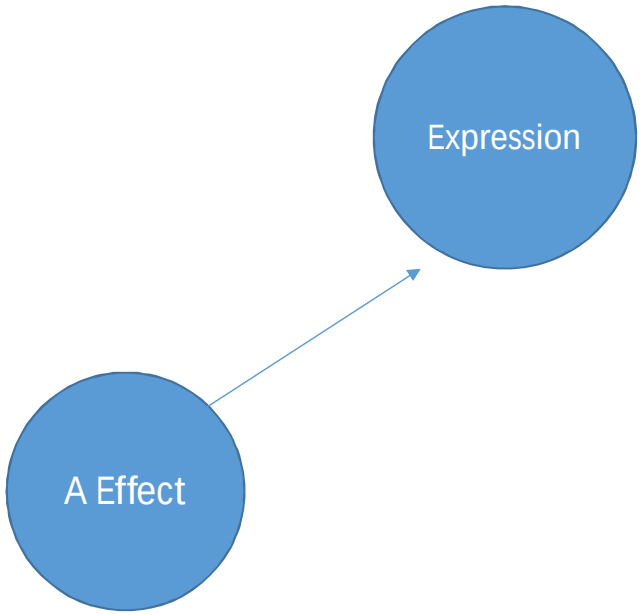
You can describe the same thing with linear modeling

Expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B



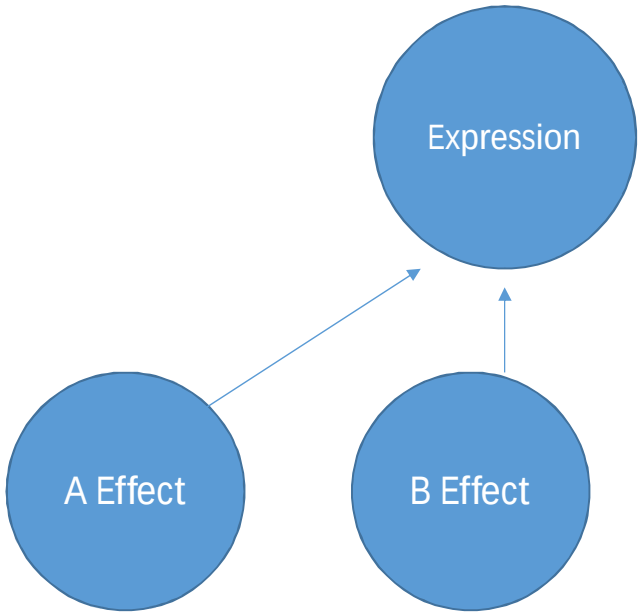
You can describe the same thing with linear modeling

Expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B



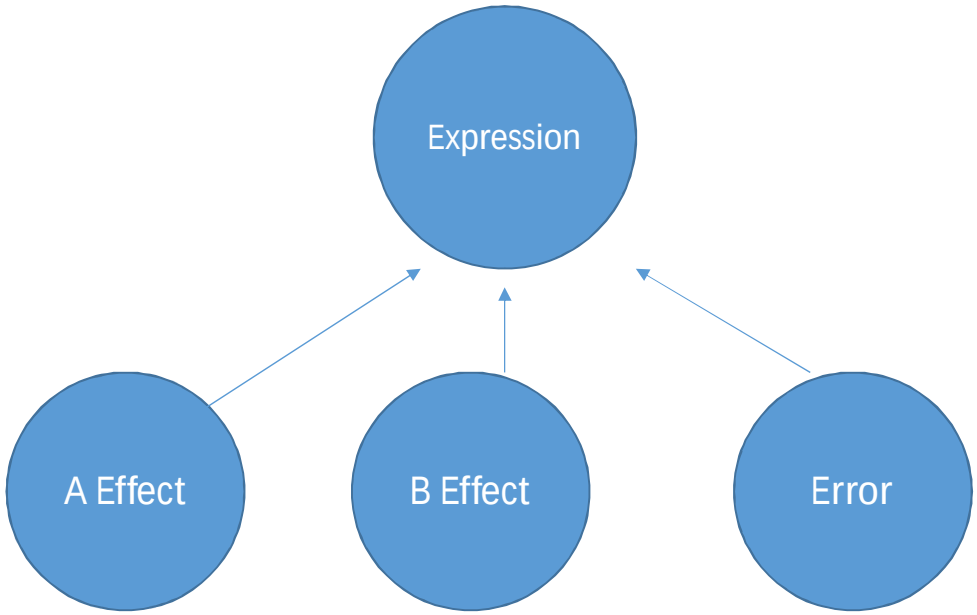
You can describe the same thing with linear modeling

Expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B



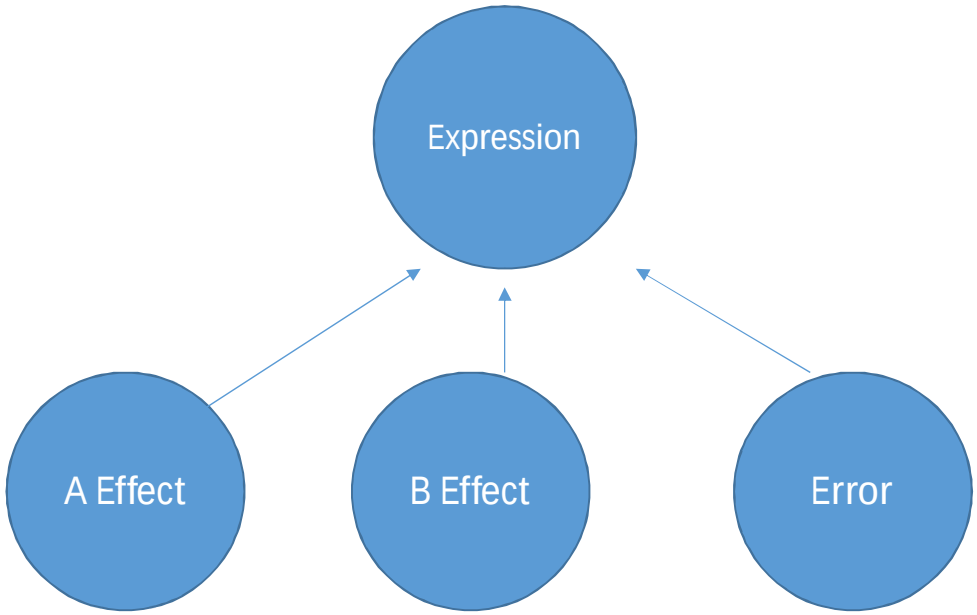
You can describe the same thing with linear modeling

Expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B



You can describe the same thing with linear modeling

Expression	condition
22	A
113	A
43	A
32	A
47	A
122	B
67	B
99	B
145	B
22	B



Expression = default effect + B Effect + Error

Use dummy variables to indicate a membership

Expression(Y)	Default(A)	Effect of B
22	1	0
113	1	0
43	1	0
32	1	0
47	1	0
122	1	1
67	1	1
99	1	1
145	1	1
22	1	1

Use dummy variables to indicate a membership

Expression(Y)	Default(A)	Effect of B
22	1	0
113	1	0
43	1	0
32	1	0
47	1	0
122	1	1
67	1	1
99	1	1
145	1	1
22	1	1

Expression = default effect + B Effect + Error

Use dummy variables to indicate a membership

Expression(Y)	Default(A)	Effect of B
22	1	0
113	1	0
43	1	0
32	1	0
47	1	0
122	1	1
67	1	1
99	1	1
145	1	1
22	1	1

Expression = default effect + B Effect + Error

Assume

$$error \sim Normal(0, \sigma^2)$$

$$E[Y|x] = \beta_A + \beta_B x$$

Use dummy variables to indicate a membership

Expression(Y)	Default(A)	Effect of B
22	1	0
113	1	0
43	1	0
32	1	0
47	1	0
122	1	1
67	1	1
99	1	1
145	1	1
22	1	1

Expression = default effect + B Effect + Error

Assume

$$error \sim Normal(0, \sigma^2)$$

$$E[Y|x] = \beta_A + \beta_B x$$

Estimate parameters

$$\sigma^2, \beta_A, \beta_B$$

Use dummy variables to indicate a membership

Expression(Y)	Default(A)	Effect of B
22	1	0
113	1	0
43	1	0
32	1	0
47	1	0
122	1	1
67	1	1
99	1	1
145	1	1
22	1	1

Expression = default effect + B Effect + Error

Assume

$$error \sim Normal(0, \sigma^2)$$

$$E[Y|x] = \beta_A + \beta_B x$$

Estimate parameters

$$\sigma^2, \beta_A, \beta_B$$

In R: lm (expression ~ condition, data=data)

“Response variable”

“Design matrix”

Expression		Default(A)	Effect of B
22		1	0
113		1	0
43		1	0
32		1	0
47	=	1	0
122		1	1
67		1	1
99		1	1
145		1	1
22		1	1

“Response variable”

“Design matrix”

Expression		Default(A)	Effect of B
22		1	0
113		1	0
43		1	0
32		1	0
47	=	1	0
122		1	1
67		1	1
99		1	1
145		1	1
22		1	1

“Response variable”

“Design matrix”

Expression		Default(A)	Effect of B
22		1	0
113		1	0
43		1	0
32		1	0
47	=	1	0
122		1	1
67		1	1
99		1	1
145		1	1
22		1	1

“Response variable”

“Design matrix”

Expression
22
113
43
32
47
122
67
99
145
22

=

Default(A)	Effect of B
1	0
1	0
1	0
1	0
1	0
1	1
1	1
1	1
1	1

"Response variable"

Expression
22
113
43
32
47
122
67
99
145
22

=

"Design matrix"

Default(A)	Effect of B
1	0
1	0
1	0
1	0
1	0
1	1
1	1
1	1
1	1

"Coefficients"

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

“Response variable” “Design matrix” “Coefficients”

Expression
22
113
43
32
47
122
67
99
145
22

$$=\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

“Response variable” “Design matrix” “Coefficients”

$$\begin{bmatrix} 22 \\ 113 \\ 43 \\ 32 \\ 47 \\ 122 \\ 67 \\ 99 \\ 145 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

“Response variable”

“Design matrix”

“Coefficients”

$$\boldsymbol{Y} = \begin{bmatrix} 22 \\ 113 \\ 43 \\ 32 \\ 47 \\ 122 \\ 67 \\ 99 \\ 145 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

“Response variable” “Design matrix” “Coefficients”

$$\boldsymbol{Y} = \begin{bmatrix} 22 \\ 113 \\ 43 \\ 32 \\ 47 \\ 122 \\ 67 \\ 99 \\ 145 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

"Response variable" "Design matrix" "Coefficients"

$$\boldsymbol{Y} = \boldsymbol{X}$$

$$\begin{bmatrix} 22 \\ 113 \\ 43 \\ 32 \\ 47 \\ 122 \\ 67 \\ 99 \\ 145 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

"Response variable" "Design matrix" "Coefficients"

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$$

$$\begin{bmatrix} 22 \\ 113 \\ 43 \\ 32 \\ 47 \\ 122 \\ 67 \\ 99 \\ 145 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

"Response variable"

"Design matrix"

"Coefficients"

$$\boldsymbol{Y} = \boldsymbol{X} \boldsymbol{\beta}$$

$$\begin{bmatrix} 22 \\ 113 \\ 43 \\ 32 \\ 47 \\ 122 \\ 67 \\ 99 \\ 145 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Solve for $\boldsymbol{\beta}$

$$\boldsymbol{Y} = \boldsymbol{X}\ \boldsymbol{\beta}$$

Solve for $\boldsymbol{\beta}$

$$\boldsymbol{Y} = \boldsymbol{X}\ \boldsymbol{\beta}$$

Solve for $\boldsymbol{\beta}$

$$Y = X\beta$$

$$X^T Y = X^T X \beta$$

Solve for β

$$Y = X\beta$$

$$X^T Y = X^T X \beta$$

$$(X^T X)^{-1} X^T Y = (X^T X)^{-1}(X^T X)\beta$$

Solve for β

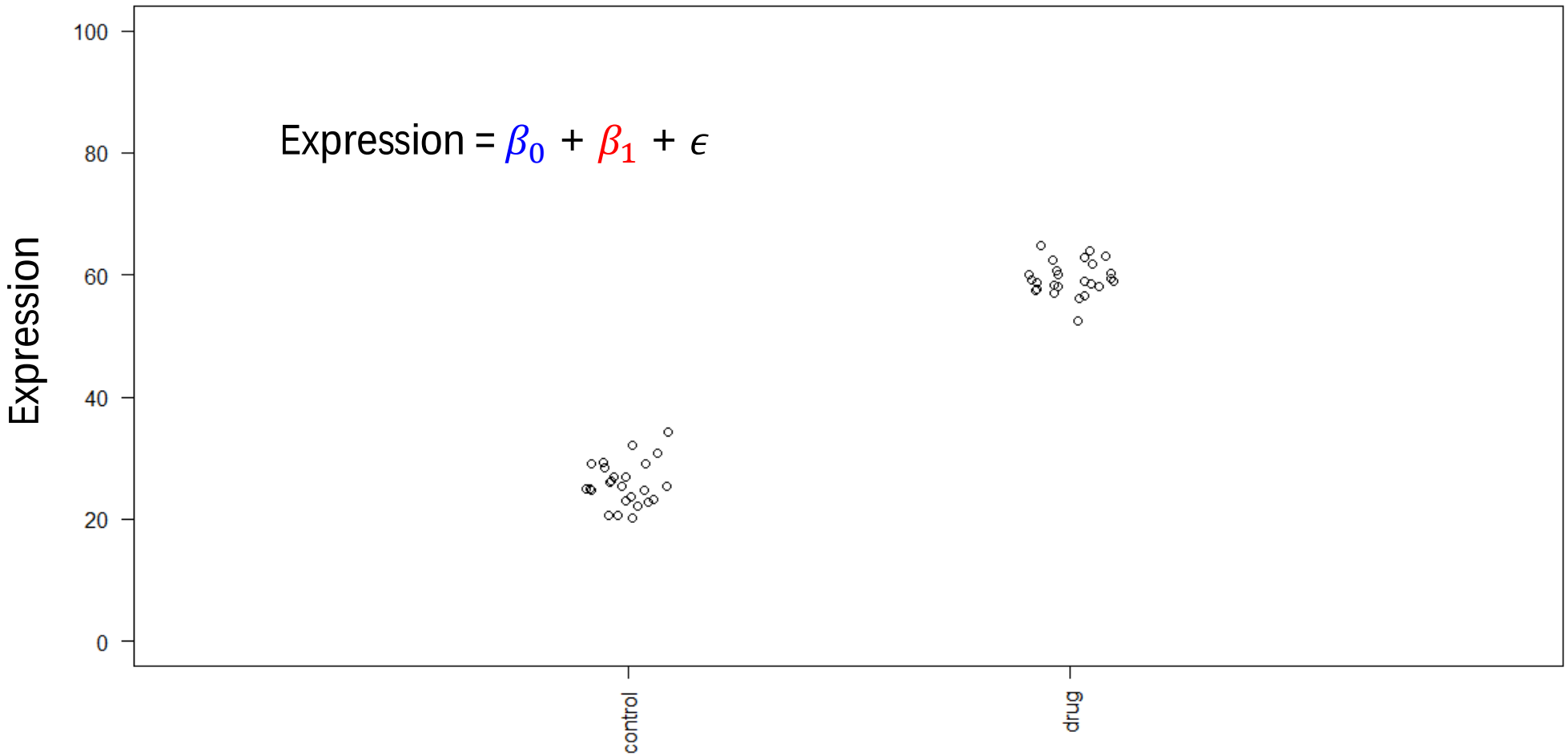
$$Y = X\beta$$

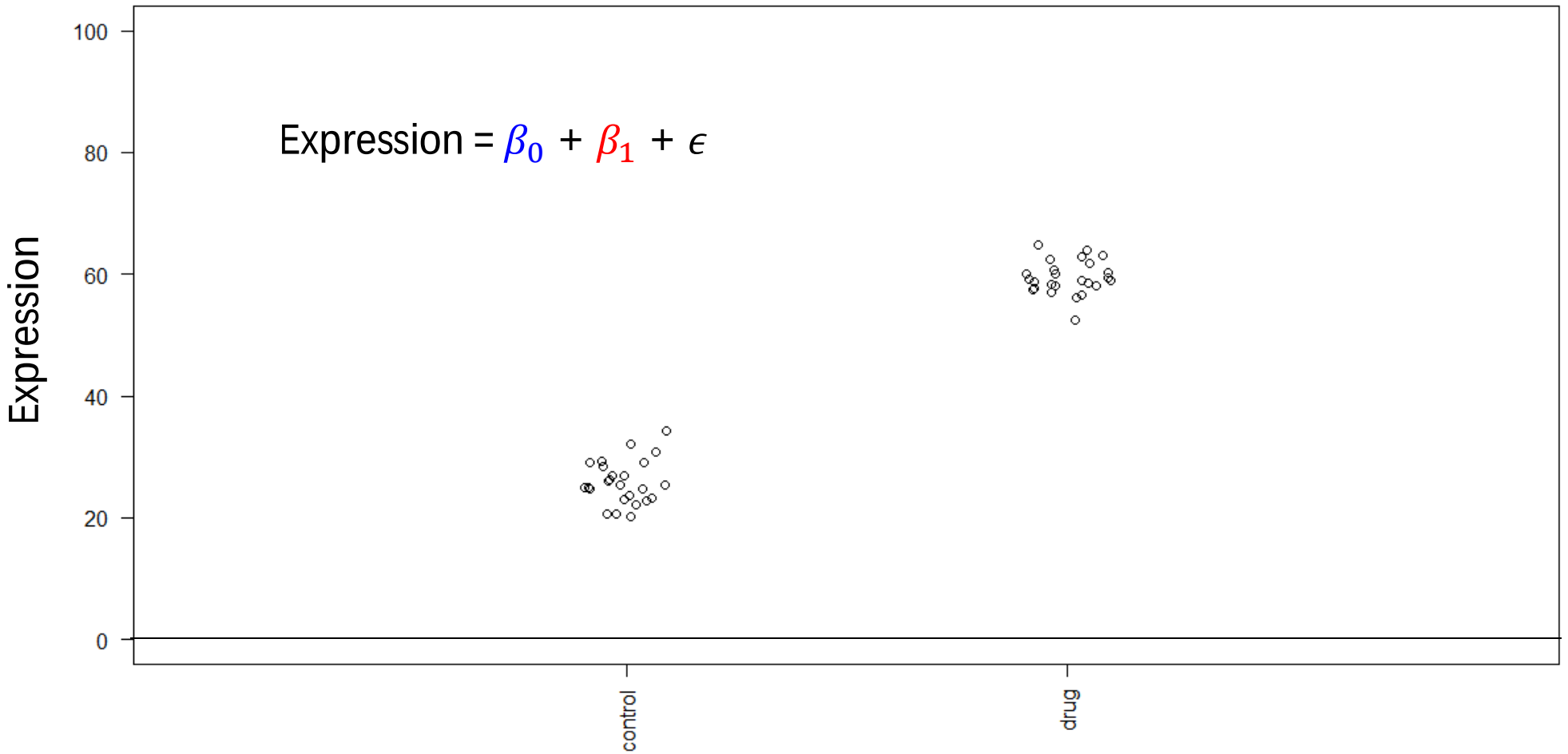
$$X^T Y = X^T X \beta$$

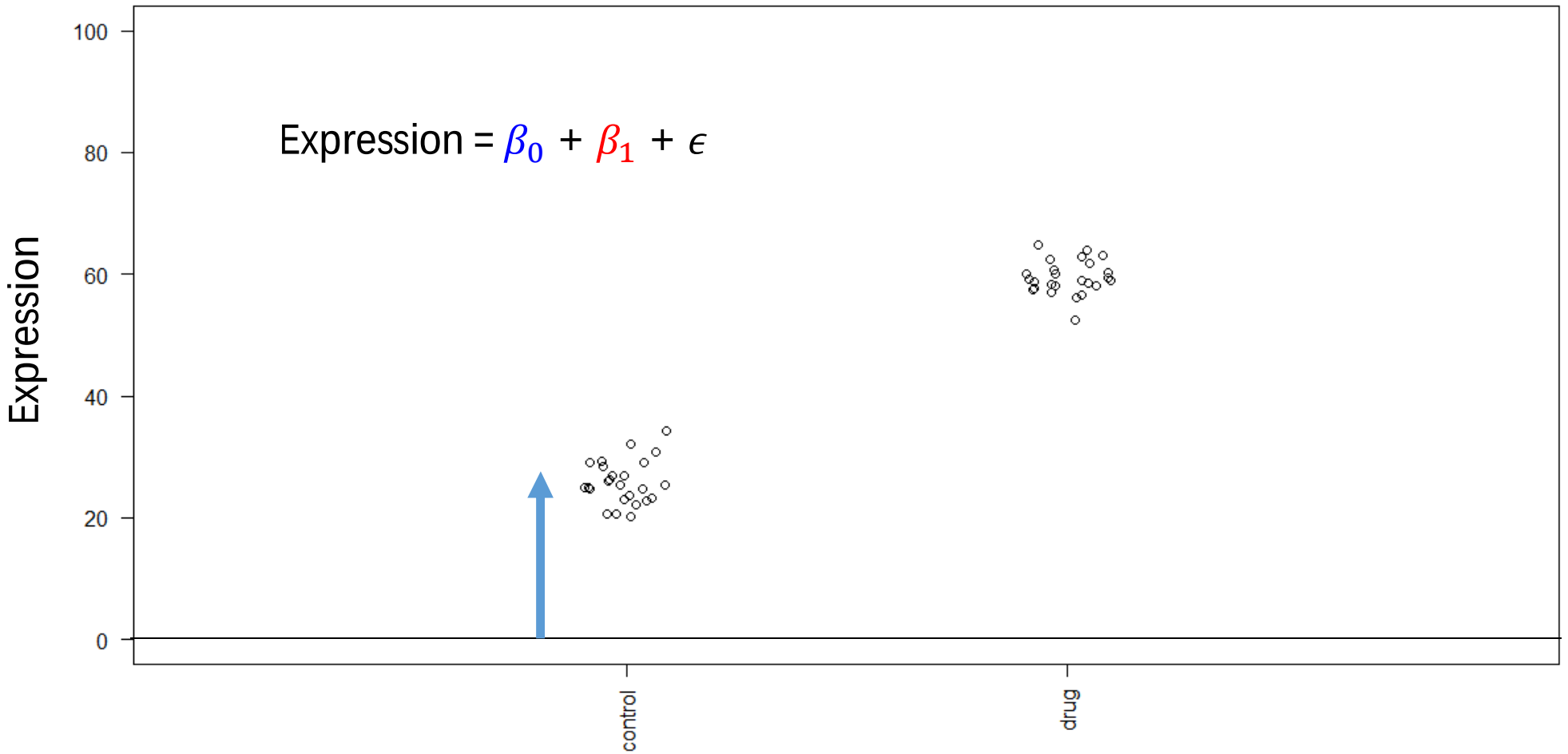
$$(X^T X)^{-1} X^T Y = (X^T X)^{-1}(X^T X)\beta$$

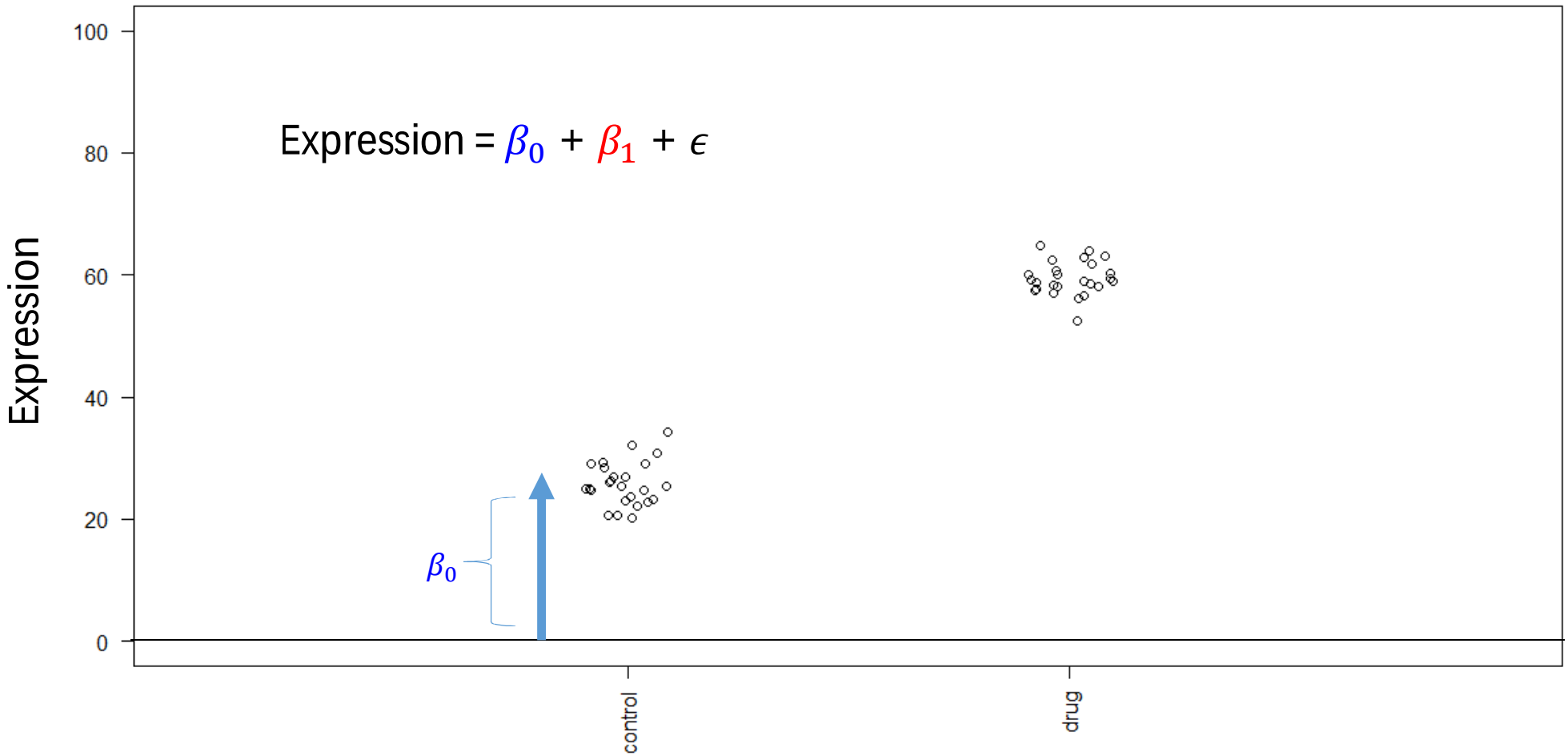
$$(X^T X)^{-1} X^T Y = \beta$$

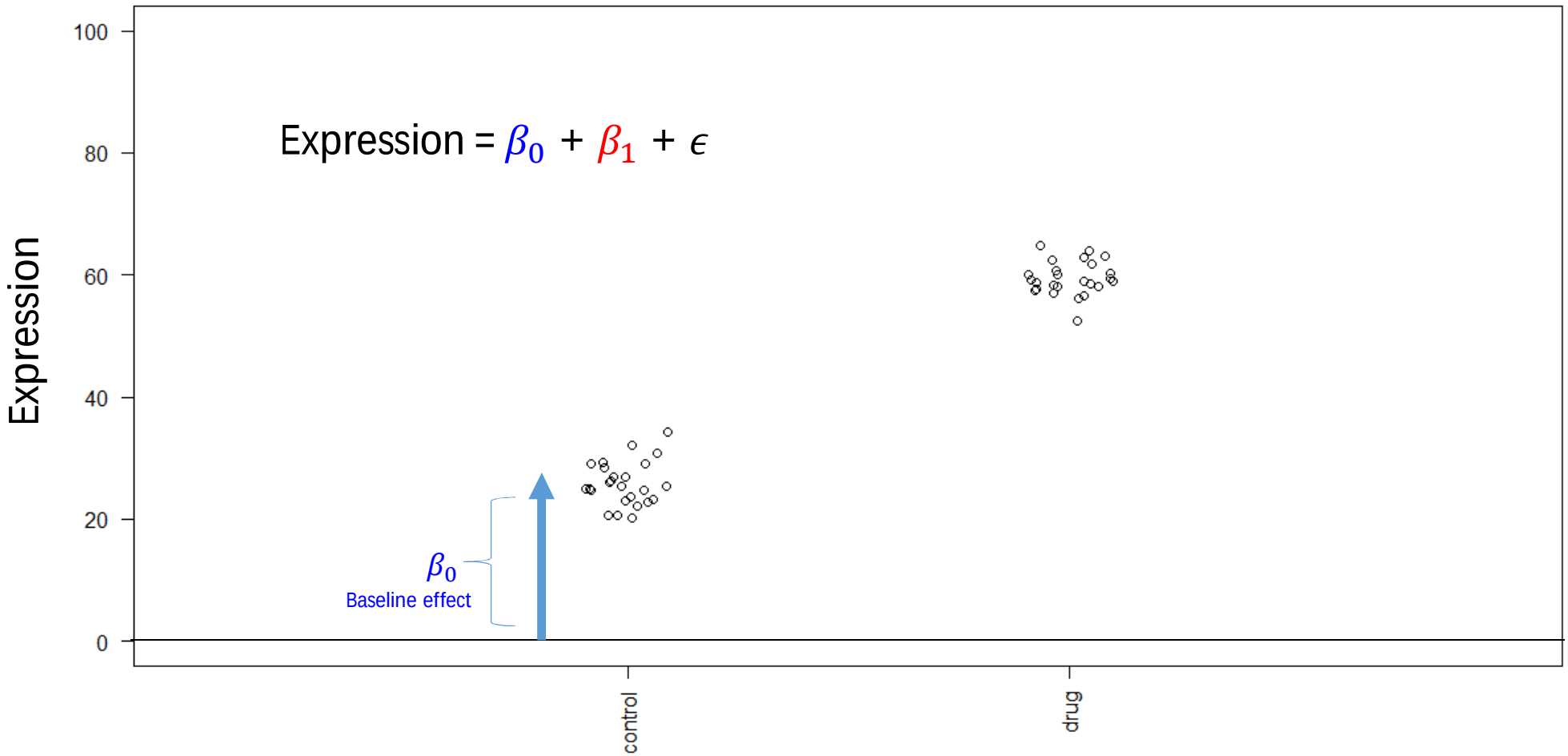
Solve for β

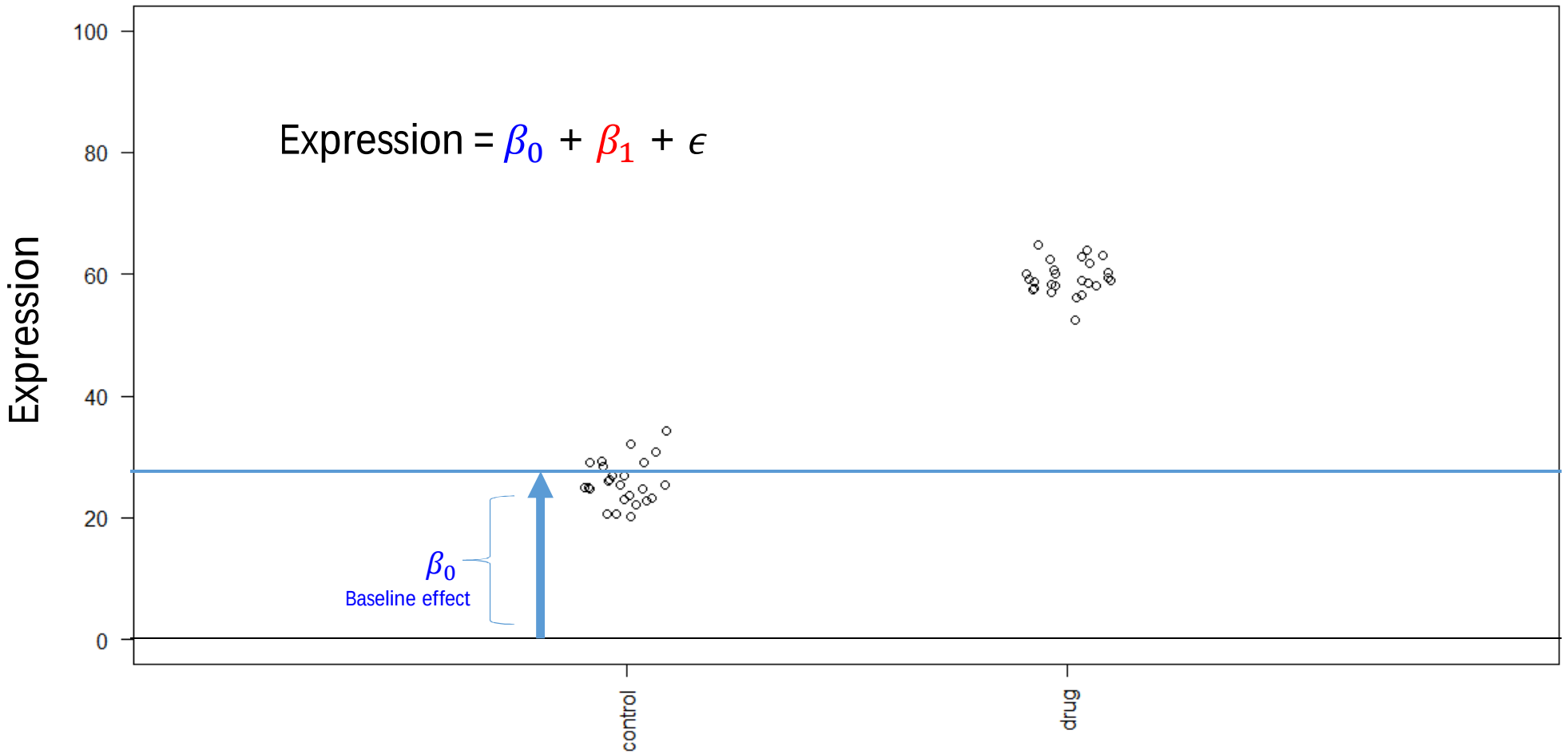

Expression = β_0 + β_1 + ϵ
Expression
100
80
60
40
20
0
control
drug

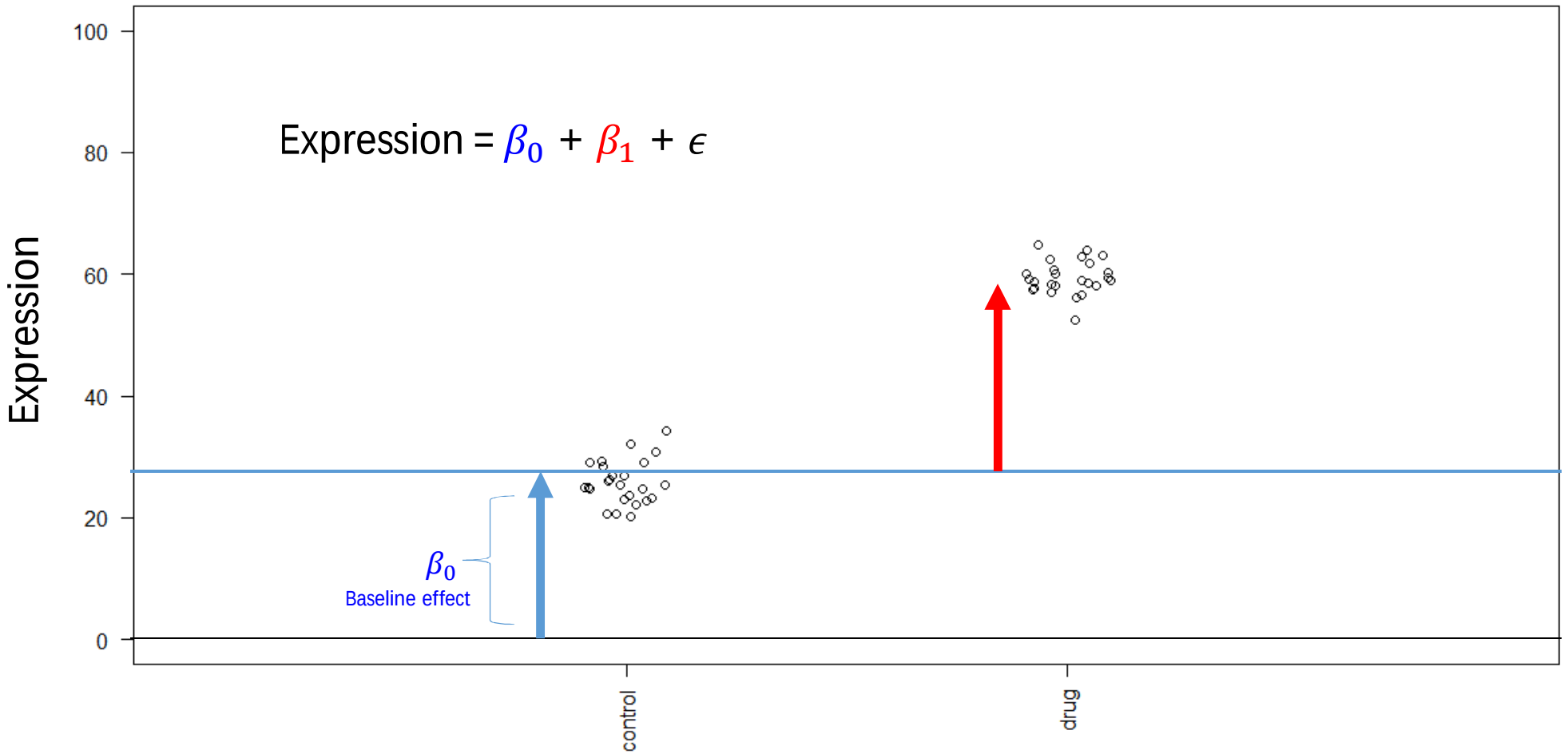

Expression = β_0 + β_1 + ϵ
Expression
100
80
60
40
20
0
control
drug

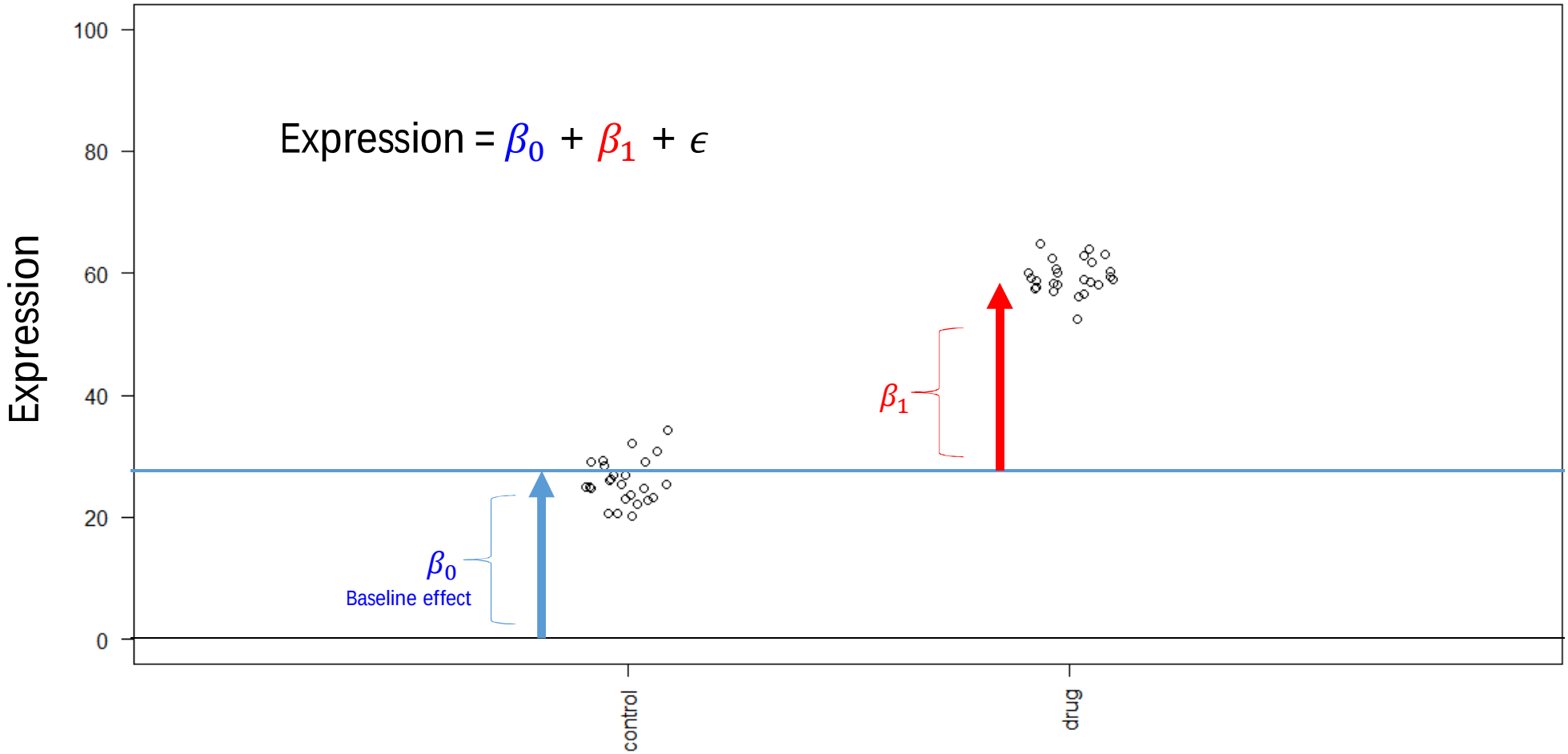

Expression = β_0 + β_1 + ϵ
Expression
100
80
60
40
20
0
control
drug

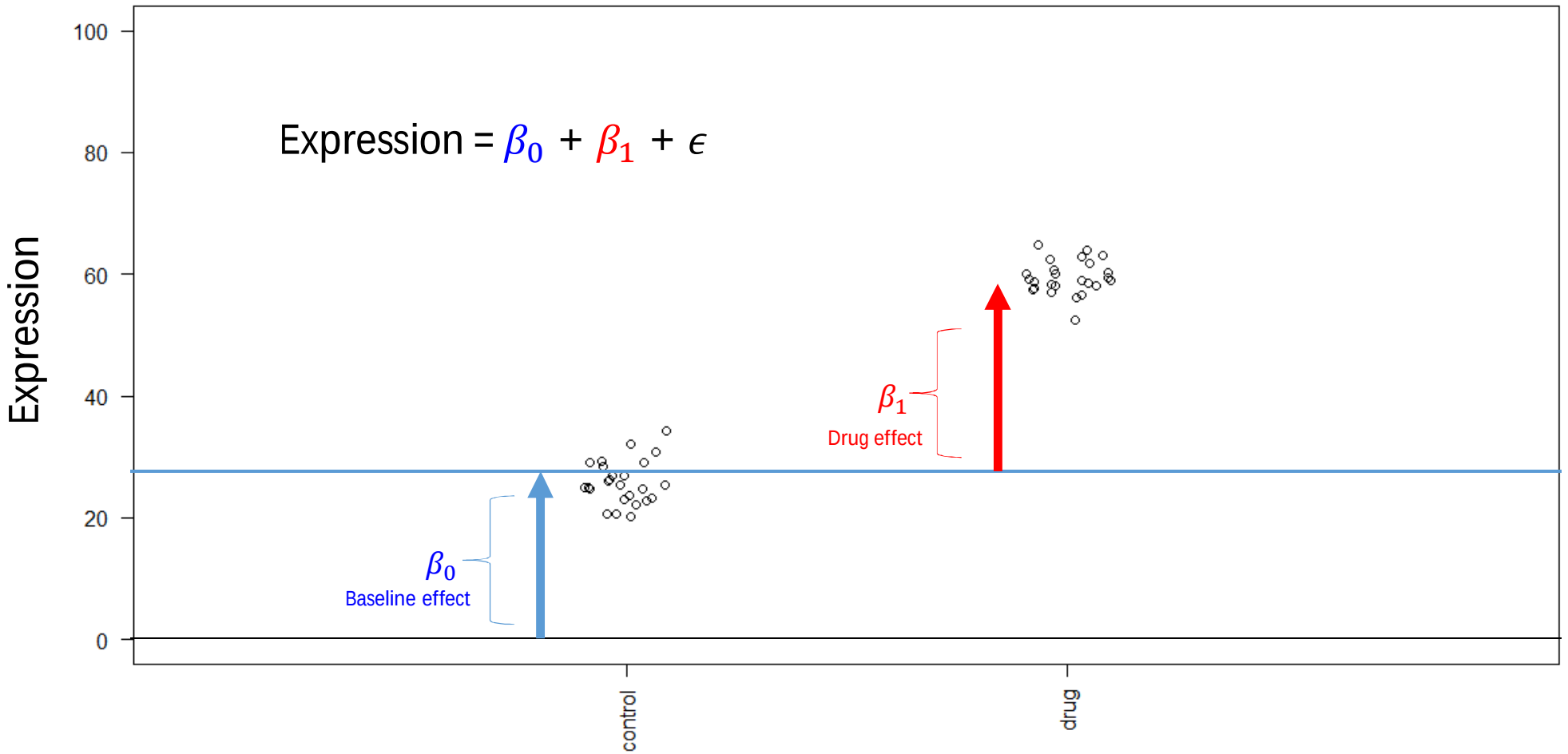

Expression = β_0 + β_1 + ϵ
Expression
100
80
60
40
20
0
β_0
control
drug


Expression = β_0 + β_1 + ϵ
Expression
100
80
60
40
20
0
β_0
Baseline effect
control
drug

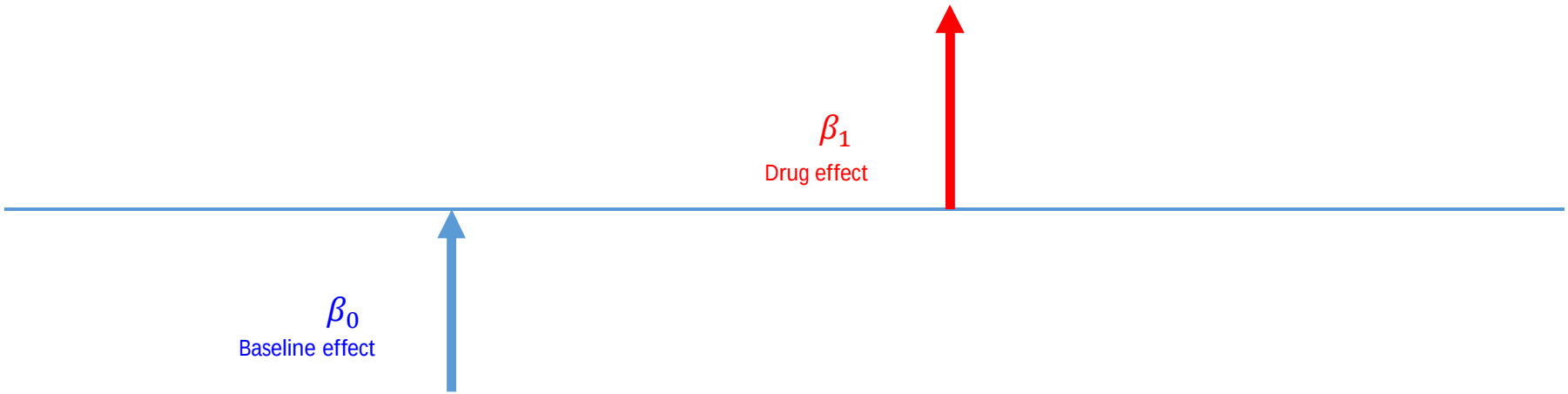

Expression = β_0 + β_1 + ϵ
Expression
100
80
60
40
20
0
β_0
Baseline effect
control
drug


Expression = β_0 + β_1 + ϵ
Expression
100
80
60
40
20
0
β_0
Baseline effect
control
drug

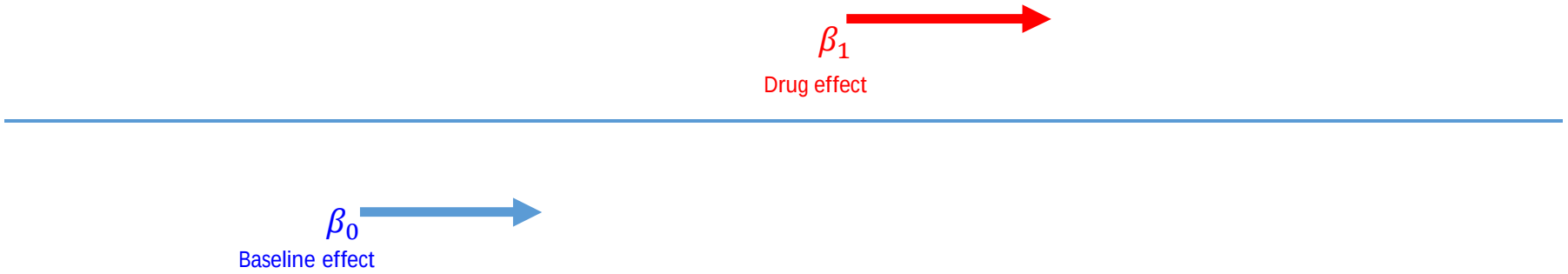

Expression = β_0 + β_1 + ϵ
Expression
100
80
60
40
20
0
β_1
β_0
Baseline effect
control
drug


Expression = β_0 + β_1 + ϵ
Expression
100
80
60
40
20
0
β_1
Drug effect
β_0
Baseline effect
control
drug

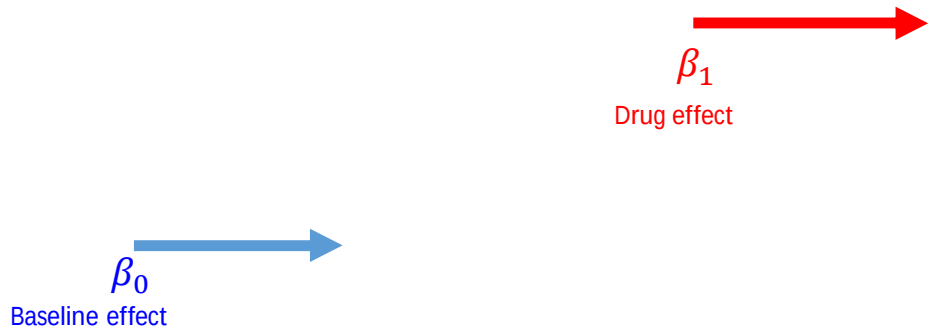
Expression = β_0 + β_1 + ϵ



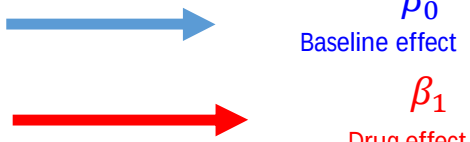
Expression = β_0 + β_1 + ϵ



Expression = β_0 + β_1 + ϵ



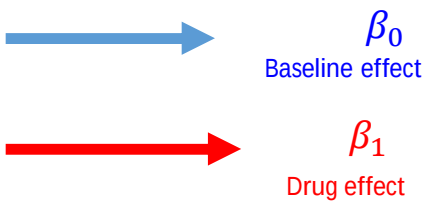
Expression = β_0 + β_1 + ϵ



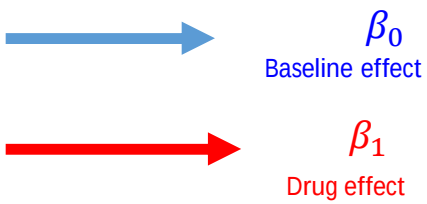
β_0
Baseline effect

β_1
Drug effect

Expression = β_0 + β_1 + ϵ

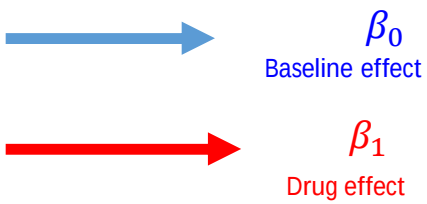


Expression = β_0 + β_1 + ϵ



Here, we *partitioned* the variance of expression into three parts: Baseline, drug effect, and residuals

Expression = β_0 + β_1 + ϵ

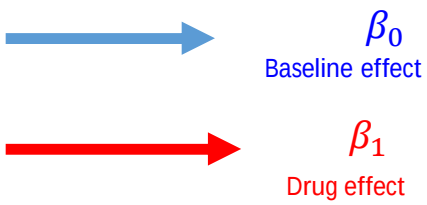


Here, we *partitioned* the variance of expression into three parts: Baseline, drug effect, and residuals

Now we can do analyze each term separately

For example, we can check if drug effect is significant

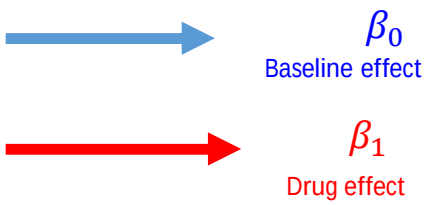
Expression = β_0 + β_1 + ϵ



Here, we *partitioned* the variance of expression into three parts: Baseline, drug effect, and residuals

Now we can do analyze each term separately

Expression = β_0 + β_1 + ϵ



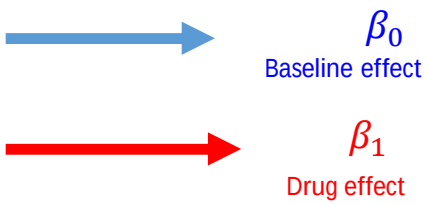
For example, we can check if drug effect is significant

$$H_0: \beta_1 = 0$$

Here, we *partitioned* the variance of expression into three parts: Baseline, drug effect, and residuals

Now we can do analyze each term separately

Expression = $\beta_0 + \beta_1 + \epsilon$



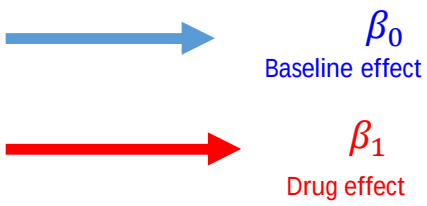
For example, we can check if drug effect is significant

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

Here, we *partitioned* the variance of expression into three parts: Baseline, drug effect, and residuals

Now we can do analyze each term separately

Expression = β_0 + β_1 + ϵ



Here, we *partitioned* the variance of expression into three parts: Baseline, drug effect, and residuals

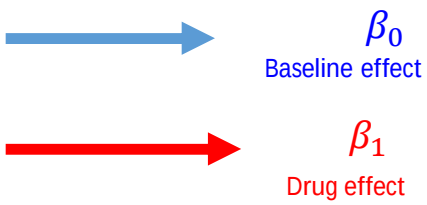
Now we can do analyze each term separately

For example, we can check if drug effect is significant

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

$$\frac{\beta_1}{SE(\beta_1)} \sim t$$

Expression = β_0 + β_1 + ϵ



Here, we *partitioned* the variance of expression into three parts: Baseline, drug effect, and residuals

Now we can do analyze each term separately

For example, we can check if drug effect is significant

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

$$\frac{\beta_1}{SE(\beta_1)} \sim t$$

$$t \rightarrow p\ value$$

$$\beta = (X^T X)^{-1} X^T Y$$

Least square

$$\beta = (X^T X)^{-1} X^T Y$$

Least square

$$Var(\beta) = Var((X^T X)^{-1} X^T Y)$$

$$\beta = (X^T X)^{-1} X^T Y$$

Least square

$$Var(\beta) = Var\left((X^T X)^{-1} X^T Y \right)$$

$$= (X^T X)^{-1} X^T Var(Y) \left((X^T X)^{-1} X^T \right)^T$$

Since $Var(AX) = AVar(X)A^T$

$$\beta = (X^T X)^{-1} X^T Y$$ Least square

$$Var(\beta) = Var\left((X^T X)^{-1} X^T Y \right)$$

$$= (X^T X)^{-1} X^T Var(Y) \left((X^T X)^{-1} X^T \right)^T$$ Since $Var(AX) = AVar(X)A^T$

$$= (X^T X)^{-1} X^T \sigma^2 \left((X^T X)^{-1} X^T \right)^T$$ Since Var(Y) = σ^2

$$\beta = (X^T X)^{-1} X^T Y$$

Least square

$$Var(\beta) = Var\left((X^T X)^{-1} X^T Y\right)$$

$$= (X^T X)^{-1} X^T Var(Y) \left((X^T X)^{-1} X^T\right)^T$$

Since $Var(AX) = AVar(X)A^T$

$$= (X^T X)^{-1} X^T \sigma^2 \left((X^T X)^{-1} X^T\right)^T$$

Since Var(Y) = σ^2

$$= (X^T X)^{-1} X^T \sigma^2 X (X^T X)^{-1}$$

Since $(AX)^T = X^T A^T$ and $A^{-1^T} = A^{T^{-1}}$

$$\beta = (X^T X)^{-1} X^T Y$$ Least square

$$Var(\beta) = Var\left((X^T X)^{-1} X^T Y \right)$$

$$= (X^T X)^{-1} X^T Var(Y) \left((X^T X)^{-1} X^T \right)^T$$ Since $Var(AX) = AVar(X)A^T$

$$= (X^T X)^{-1} X^T \sigma^2 \left((X^T X)^{-1} X^T \right)^T$$ Since Var(Y) = σ^2

$$= (X^T X)^{-1} X^T \sigma^2 X (X^T X)^{-1}$$ Since $(AX)^T = X^T A^T$ and $A^{-1^T} = A^{T^{-1}}$

$$= (X^T X)^{-1} \sigma^2 X^T X (X^T X)^{-1}$$ Since $AA^{-1} = I$

$$\beta = (X^T X)^{-1} X^T Y$$ Least square

$$Var(\beta) = Var\left((X^T X)^{-1} X^T Y \right)$$

$$= (X^T X)^{-1} X^T Var(Y) \left((X^T X)^{-1} X^T \right)^T$$ Since $Var(AX) = AVar(X)A^T$

$$= (X^T X)^{-1} X^T \sigma^2 \left((X^T X)^{-1} X^T \right)^T$$ Since Var(Y) = σ^2

$$= (X^T X)^{-1} X^T \sigma^2 X (X^T X)^{-1}$$ Since $(AX)^T = X^T A^T$ and $A^{-1^T} = A^{T^{-1}}$

$$= (X^T X)^{-1} \sigma^2 X^T X (X^T X)^{-1}$$ Since $AA^{-1} = I$

$$= (X^T X)^{-1} \sigma^2$$

$$\beta = (X^T X)^{-1} X^T Y$$ Least square

$$Var(\beta) = Var\left((X^T X)^{-1} X^T Y \right)$$

$$= (X^T X)^{-1} X^T Var(Y) \left((X^T X)^{-1} X^T \right)^T$$ Since $Var(AX) = AVar(X)A^T$

$$= (X^T X)^{-1} X^T \sigma^2 \left((X^T X)^{-1} X^T \right)^T$$ Since Var(Y) = σ^2

$$= (X^T X)^{-1} X^T \sigma^2 X (X^T X)^{-1}$$ Since $(AX)^T = X^T A^T$ and $A^{-1^T} = A^{T^{-1}}$

$$= (X^T X)^{-1} \sigma^2 X^T X (X^T X)^{-1}$$ Since $AA^{-1} = I$

$$= (X^T X)^{-1} \sigma^2$$

$$SE(\beta) = \sqrt{Var(\beta)} = \sqrt{(X^T X)^{-1} \sigma^2}$$

“Are the two means different?”

“Are the two means different?”

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{S_A}{n_A} + \frac{S_B}{n_B}}}$$

“Are the two means different?”

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{S_A}{n_A} + \frac{S_B}{n_B}}}$$

“Are the two means different?”

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{S_A}{n_A} + \frac{S_B}{n_B}}}$$

“Is there a *additional* effect of B (if we have A as a baseline)?”

“Are the two means different?”

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{S_A}{n_A} + \frac{S_B}{n_B}}}$$

“Is there a *additional* effect of B (if we have A as a baseline)?”

$$Y = \beta_A + \beta_B$$

Limitations of CRD

- If experimental units are heterogeneous, there might be high false positives or false negatives

Level 2

“What are the genes that are differentially expressed in tumor vs normal while controlling for batch effects?”

2. Randomized block design

	Control	Drug
Batch 1		
Batch 2		
Batch 3		

2. Randomized block design

	Control	Drug
Batch 1		
Batch 2		
Batch 3		

2. Randomized block design

	Control	Drug
Batch 1		
Batch 2		
Batch 3		

2. Randomized block design

	Control	Drug
Batch 1		
Batch 2		
Batch 3		

2. Randomized block design

	Control	Drug
Batch 1		
Batch 2		
Batch 3		

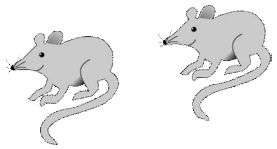
2. Randomized block design

	Control	Drug
Batch 1		
Batch 2		
Batch 3		

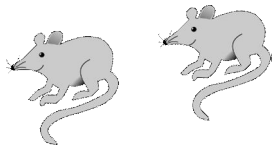
2. Randomized block design

	Control	Drug
Batch 1		
Batch 2		
Batch 3		

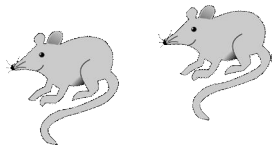
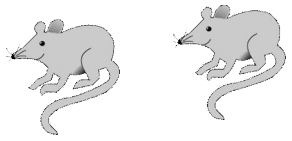
2. Randomized block design

	Control	Drug
Batch 1		
Batch 2		
Batch 3		

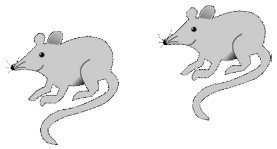
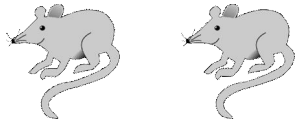
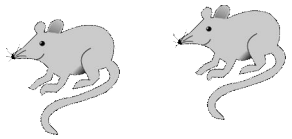
2. Randomized block design

	Control	Drug
Batch 1		
Batch 2		
Batch 3		

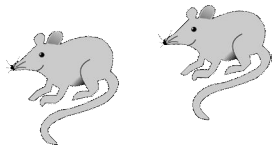
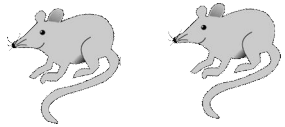
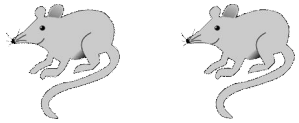
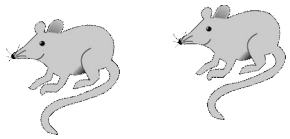
2. Randomized block design

	Control	Drug
Batch 1		
Batch 2		
Batch 3		

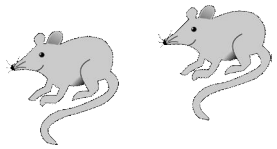
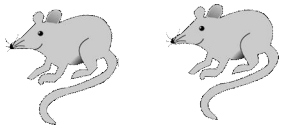
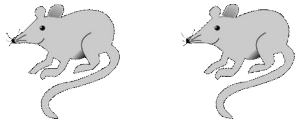
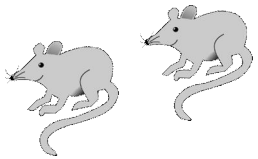
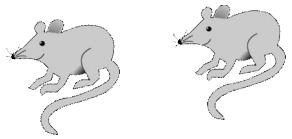
2. Randomized block design

	Control	Drug
Batch 1		
Batch 2		
Batch 3		

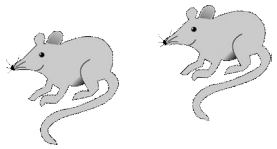
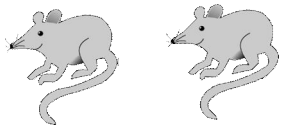
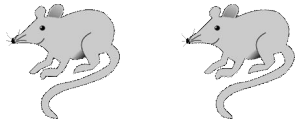
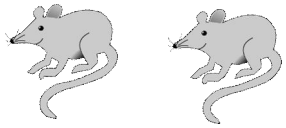
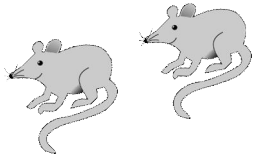
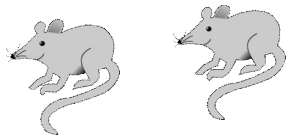
2. Randomized block design

	Control	Drug
Batch 1		
Batch 2		
Batch 3		

2. Randomized block design

	Control	Drug
Batch 1		
Batch 2		
Batch 3		

2. Randomized block design

	Control	Drug
Batch 1		
Batch 2		
Batch 3		

2. Randomized block design

	1st factor ↓	2nd factor ↓
expression	condition	batch
22	A	1
113	A	1
43	A	2
32	A	2
47	A	3
122	A	3
67	B	1
99	B	1
145	B	2
22	B	2
32	B	3
21	B	3

2. Randomized block design

	1st factor ↓	2nd factor ↓
expression	condition	batch
22	A	1
113	A	1
43	A	2
32	A	2
47	A	3
122	A	3
67	B	1
99	B	1
145	B	2
22	B	2
32	B	3
21	B	3

2. Randomized block design

	1st factor ↓	2nd factor ↓
expression	condition	batch
22	A	1
113	A	1
43	A	2
32	A	2
47	A	3
122	A	3
67	B	1
99	B	1
145	B	2
22	B	2
32	B	3
21	B	3

$$Expression \sim condition + batch$$

2. Randomized block design

	1st factor ↓	2nd factor ↓
expression	condition	batch
22	A	1
113	A	1
43	A	2
32	A	2
47	A	3
122	A	3
67	B	1
99	B	1
145	B	2
22	B	2
32	B	3
21	B	3

$$Expression \sim condition + batch$$

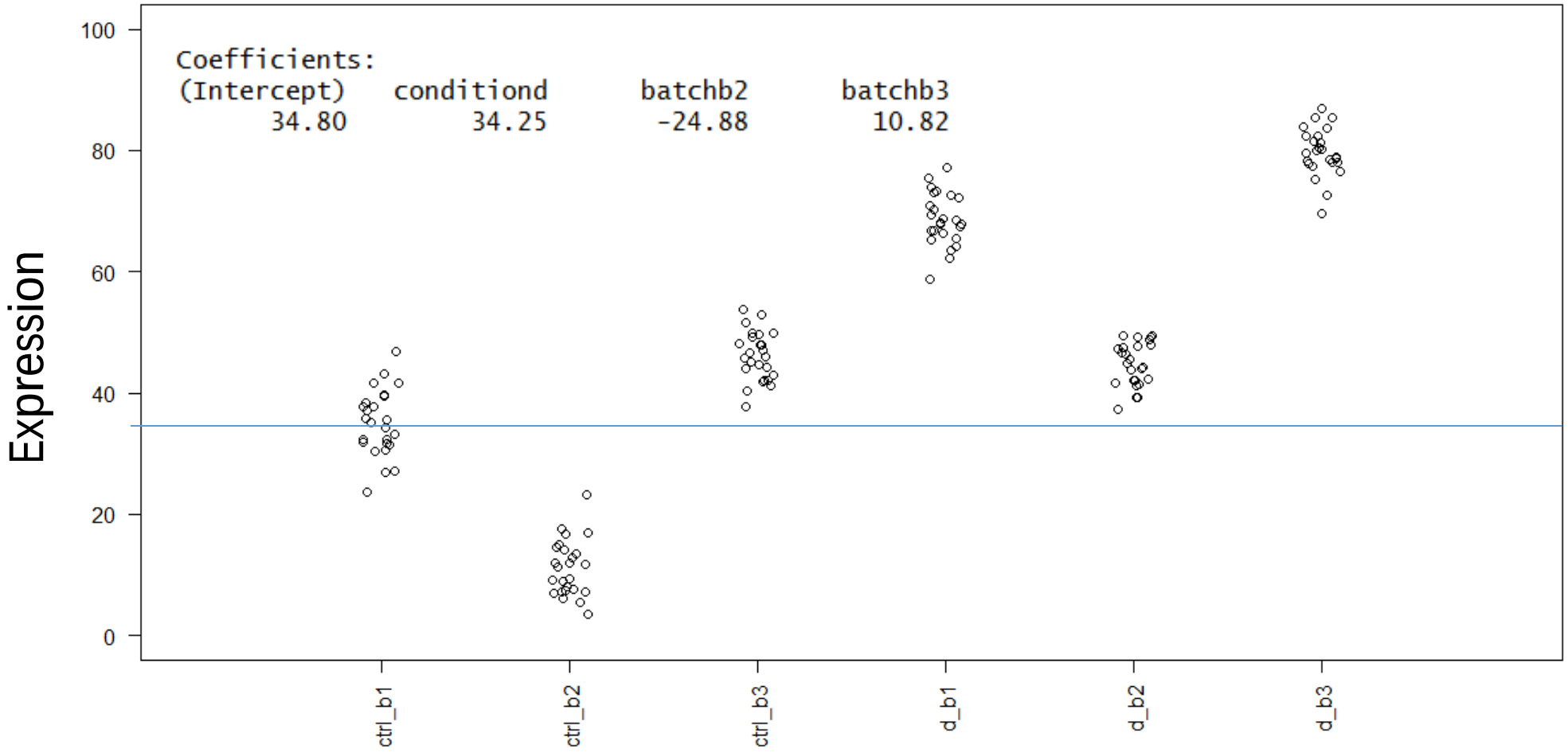
```
> (m2 <- lm(expression ~ condition + batch, data))

Call:
lm(formula = expression ~ condition + batch, data = data)

Coefficients:
(Intercept)   conditiond      batchb2      batchb3
      34.80        34.25       -24.88        10.82
```

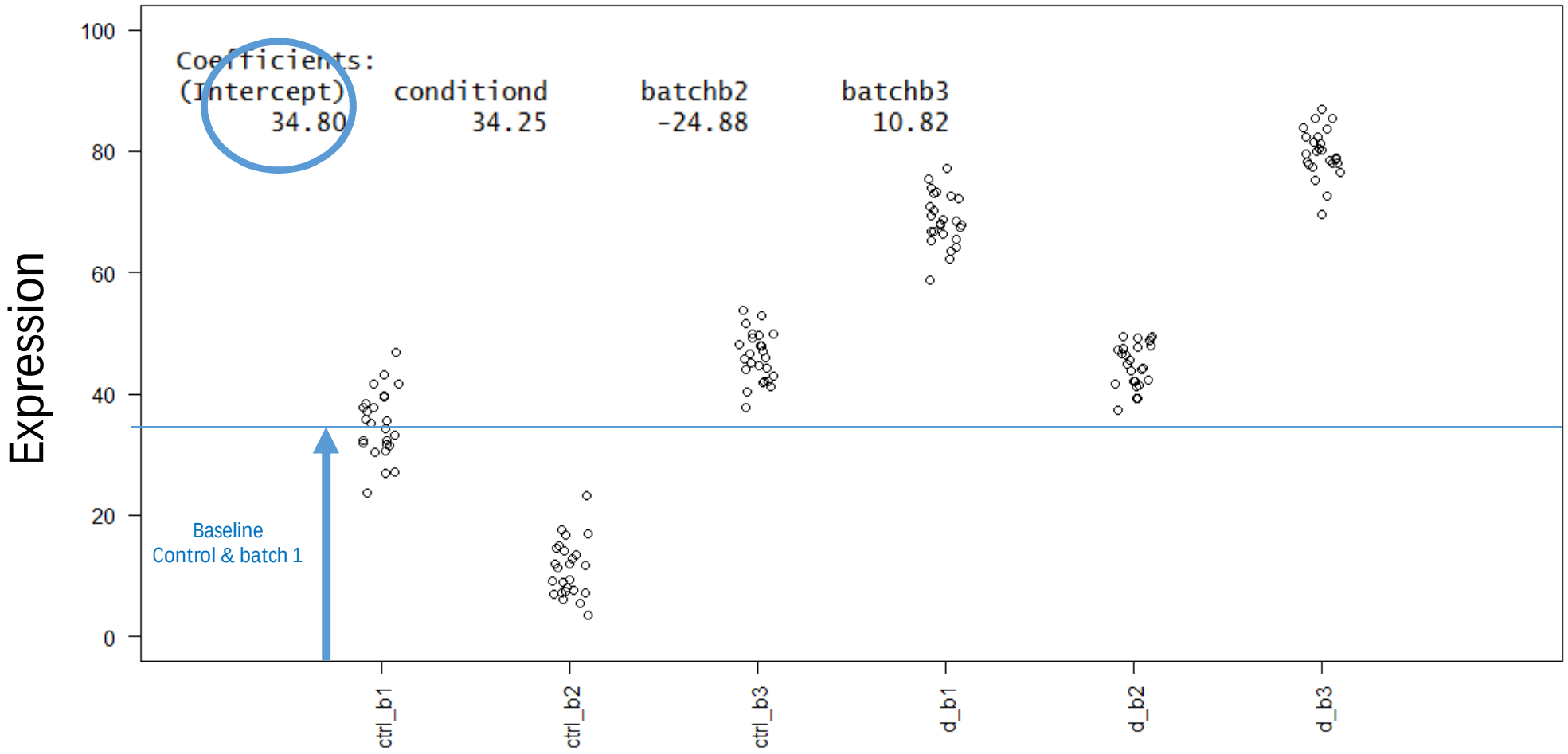
2. Randomized block design

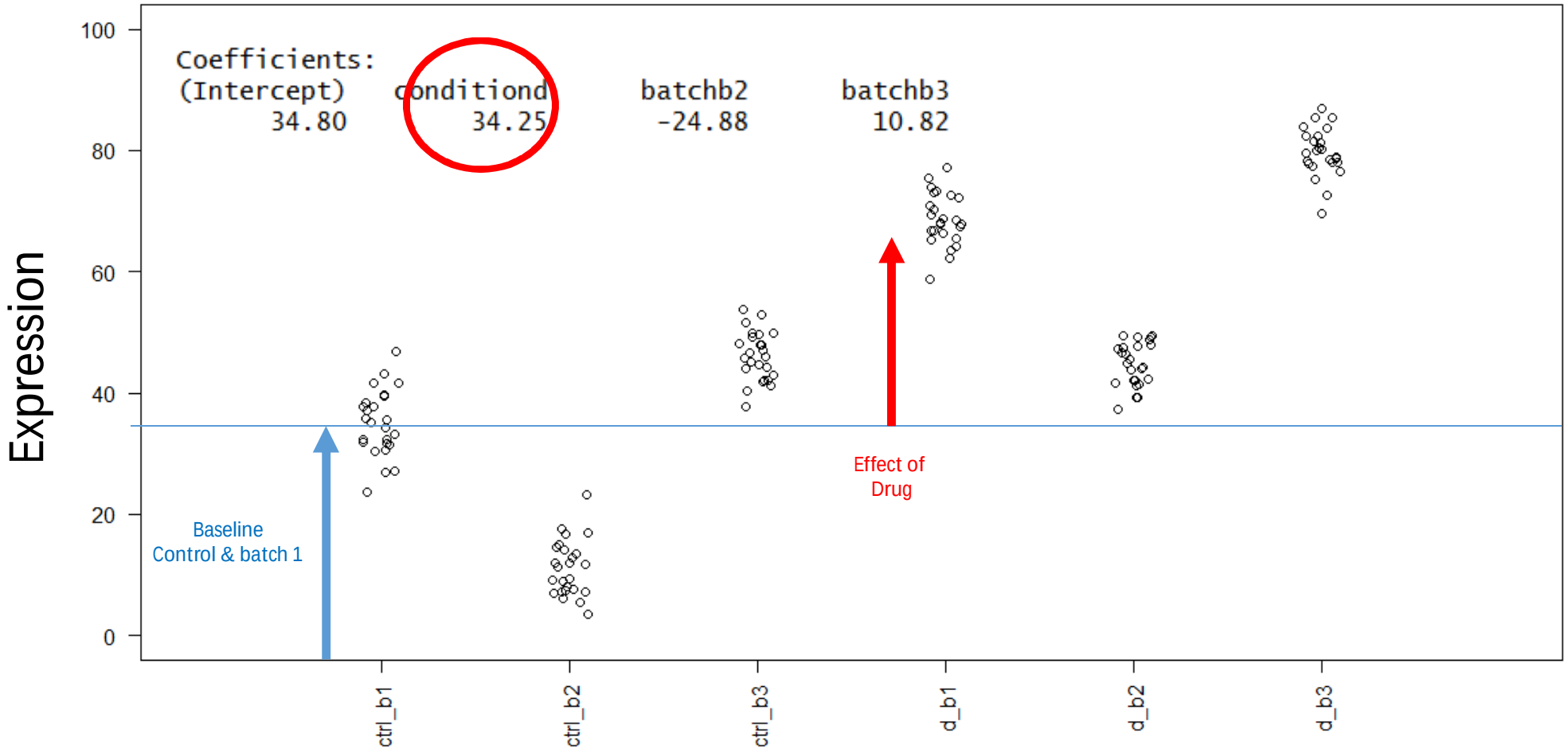
```
Coefficients:
(Intercept)   conditiond      batchb2      batchb3
      34.80        34.25       -24.88        10.82
```

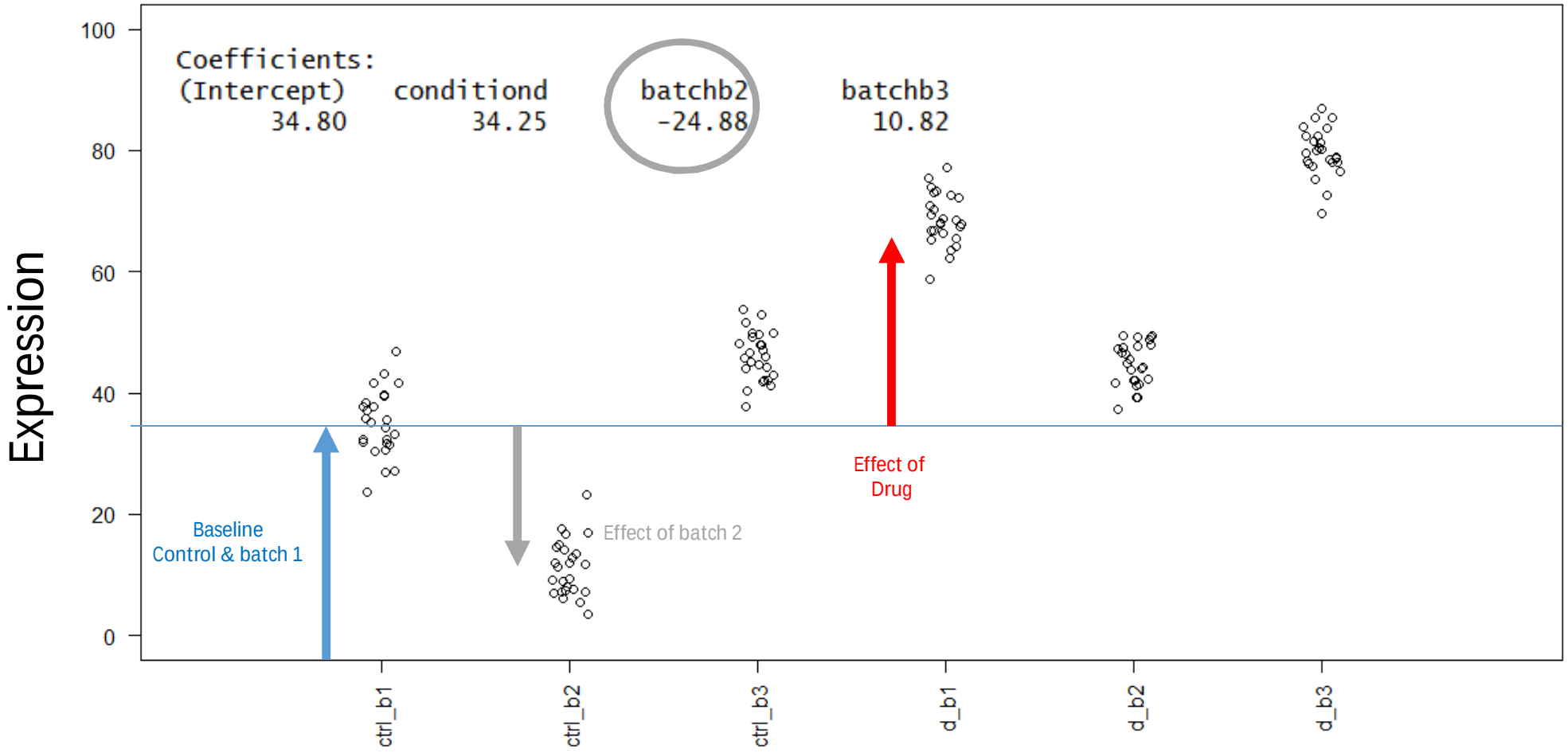
```
Coefficients:
(Intercept)   conditiond      batchb2      batchb3
      34.80        34.25       -24.88        10.82
```

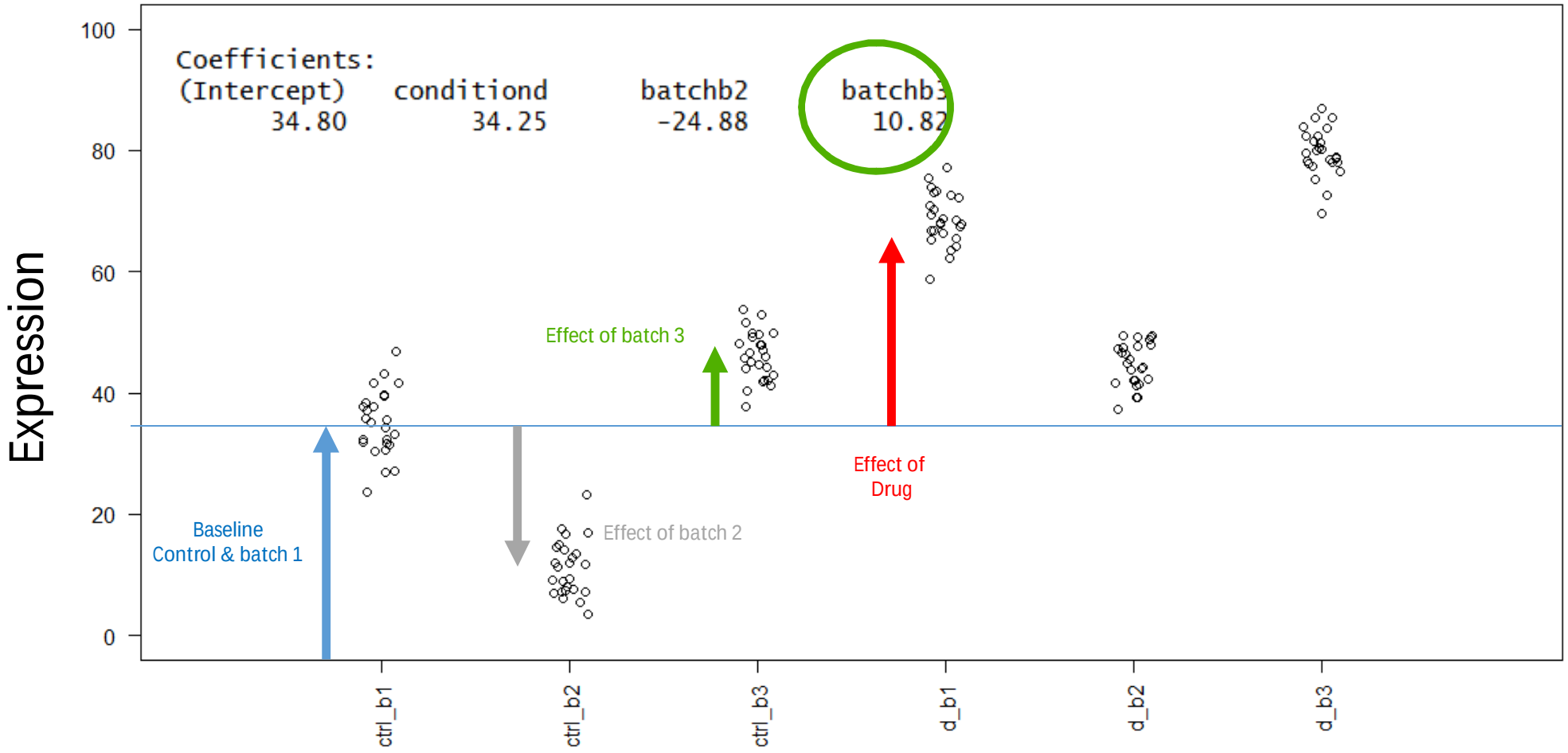
```
Coefficients:
(Intercept)   conditiond      batchb2      batchb3
      34.80        34.25       -24.88        10.82
```

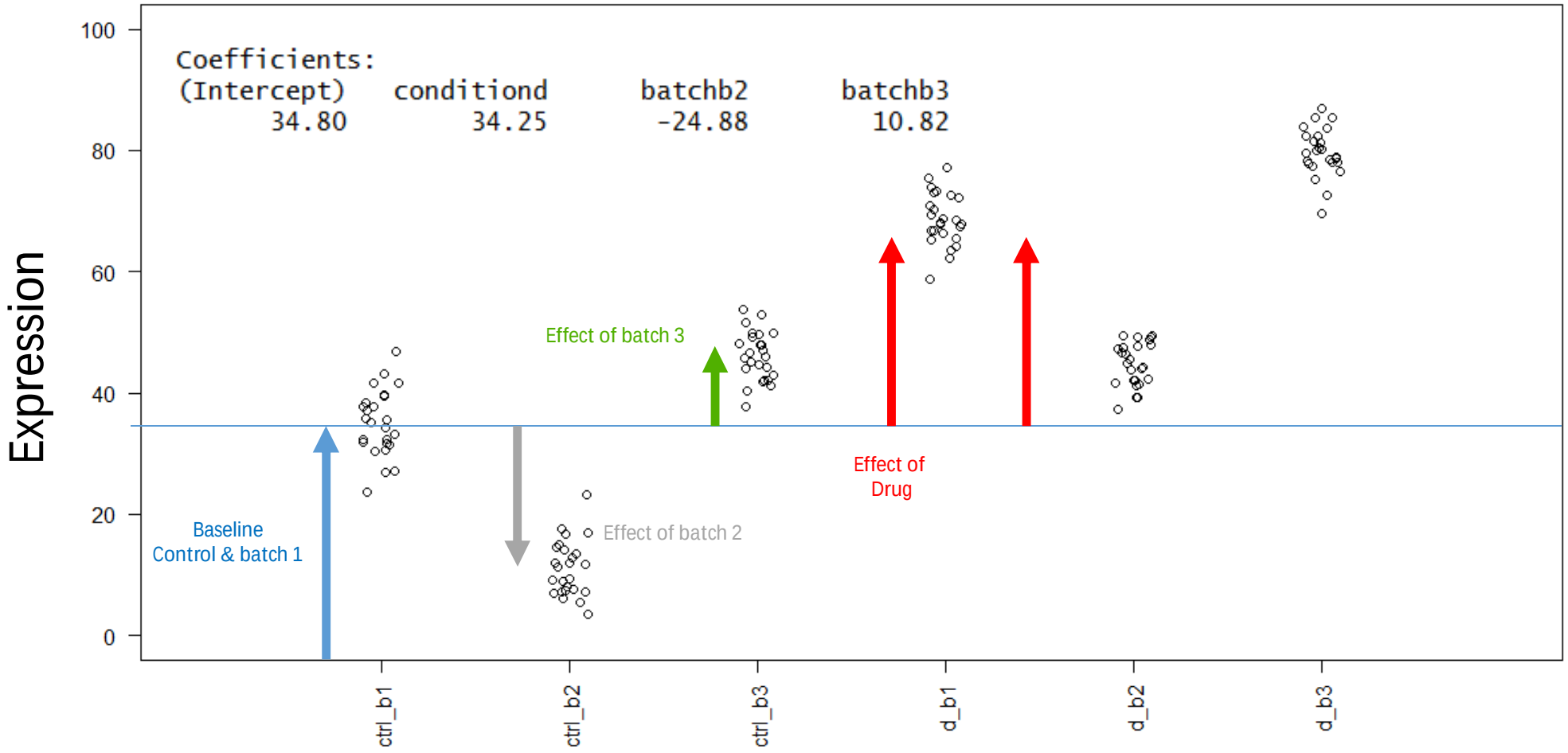

Coefficients:
(Intercept) conditiond batchb2 batchb3
34.80 34.25 -24.88 10.82
Expression
100
80
60
40
20
0
ctrl_b1
ctrl_b2
ctrl_b3
d_b1
d_b2
d_b3

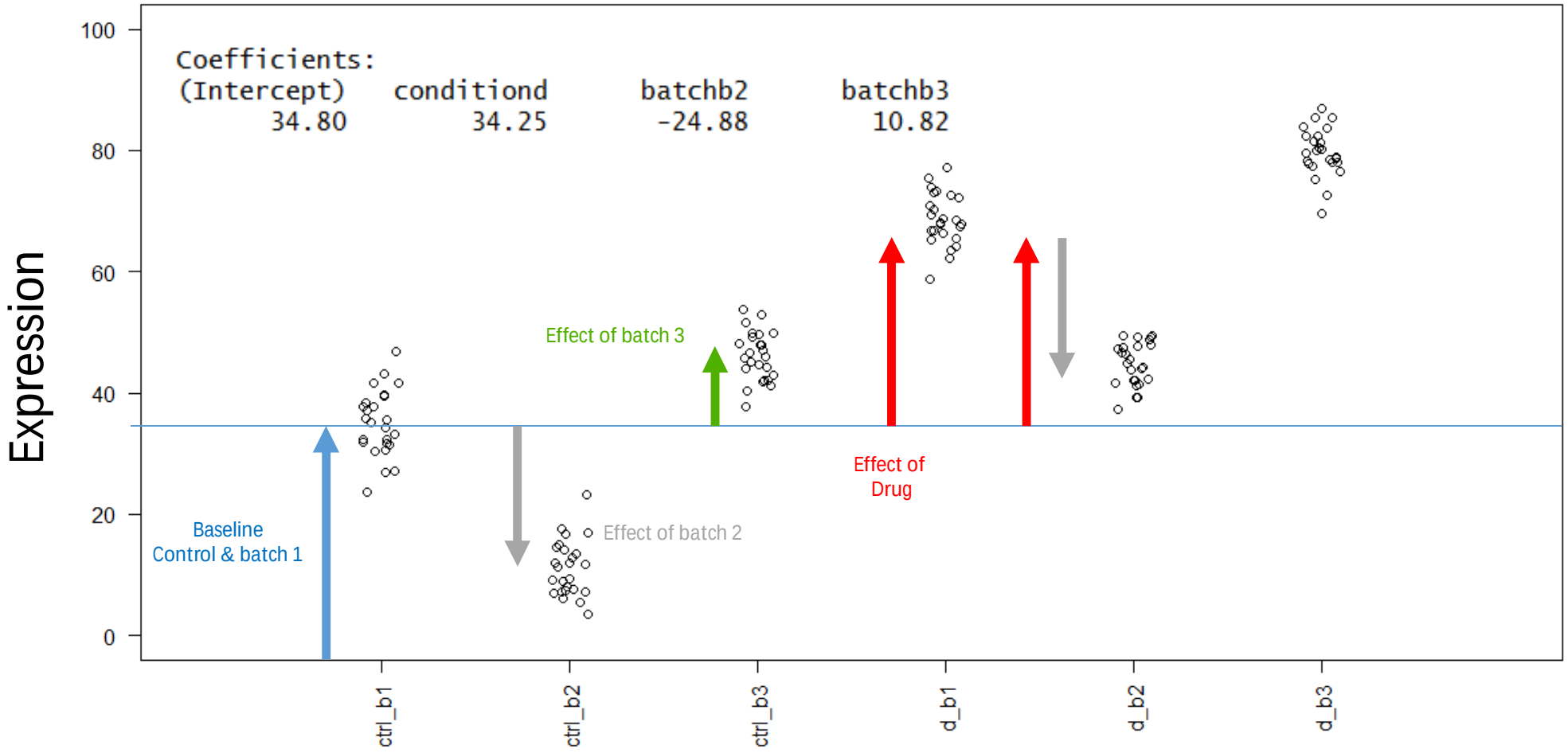

Coefficients:
(Intercept) conditiond batchb2 batchb3
34.80 34.25 -24.88 10.82
Expression
100
80
60
40
20
0
Baseline
Control & batch 1
ctrl_b1
ctrl_b2
ctrl_b3
d_b1
d_b2
d_b3

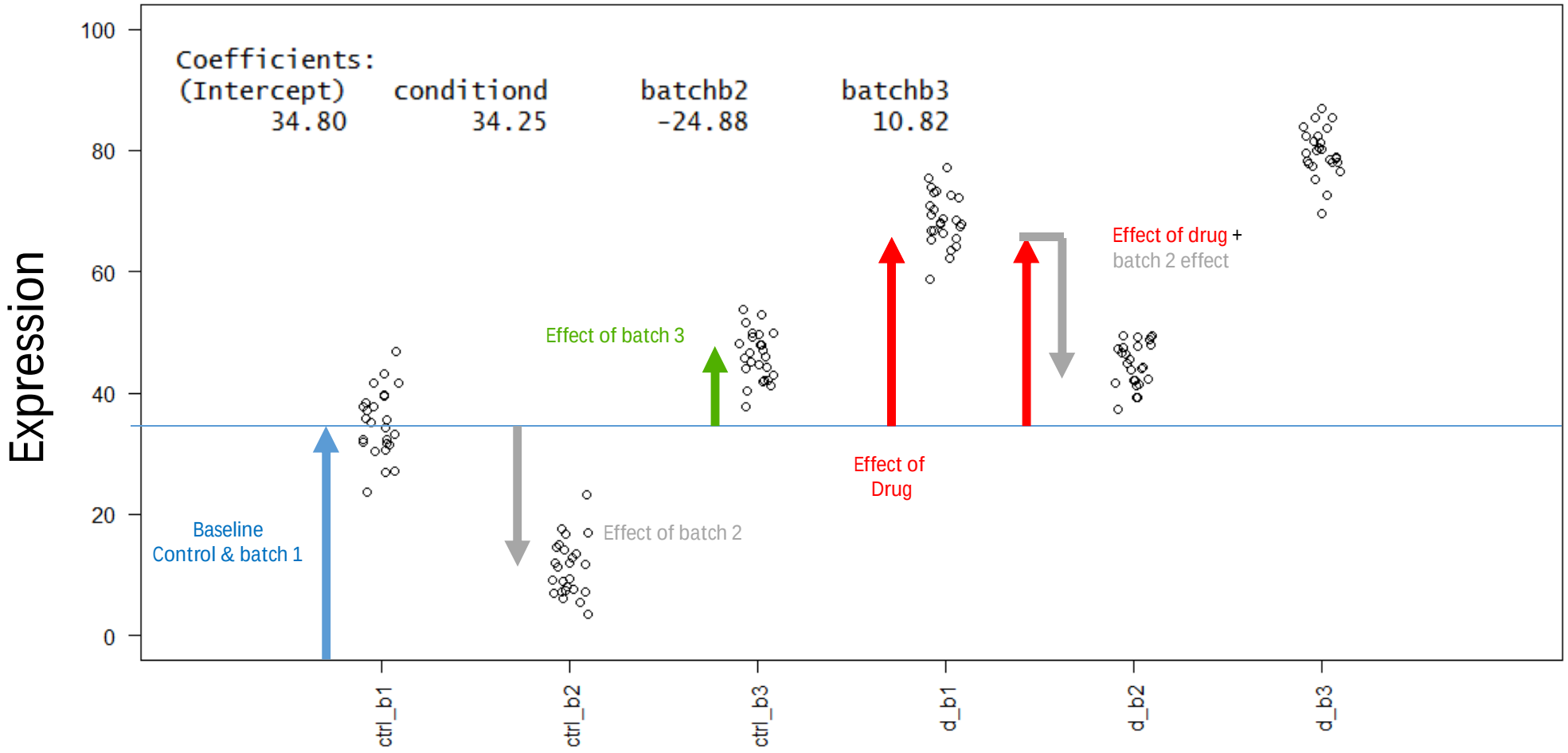

Coefficients:
(Intercept) conditiond batchb2 batchb3
34.80 34.25 -24.88 10.82
Expression
100
80
60
40
20
0
Baseline
Control & batch 1
Effect of
Drug
ctrl_b1
ctrl_b2
ctrl_b3
d_b1
d_b2
d_b3

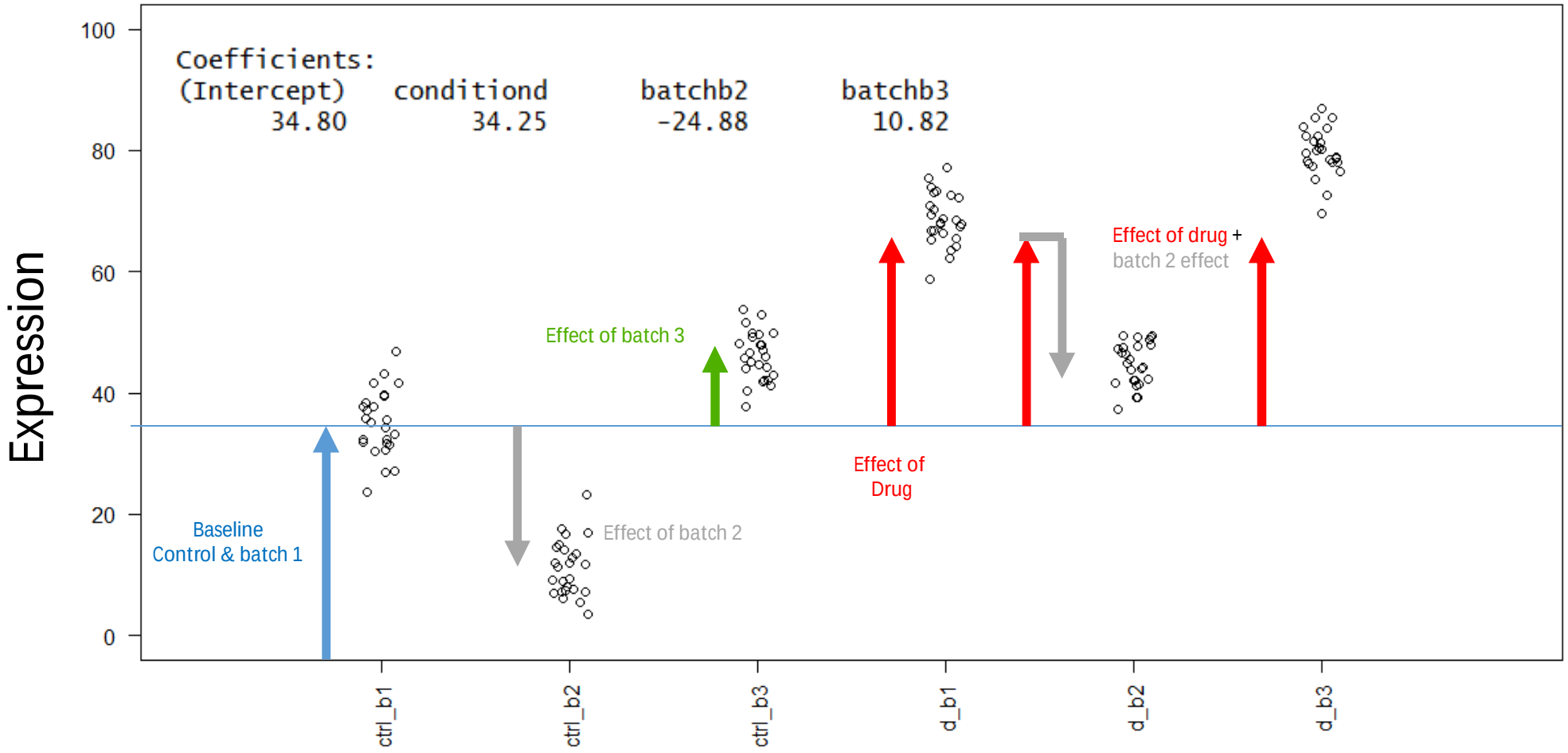

Coefficients:
(Intercept) conditiond batchb2 batchb3
34.80 34.25 -24.88 10.82
Expression
100
80
60
40
20
0
Baseline
Control & batch 1
Effect of batch 2
Effect of
Drug
ctrl_b1
ctrl_b2
ctrl_b3
d_b1
d_b2
d_b3

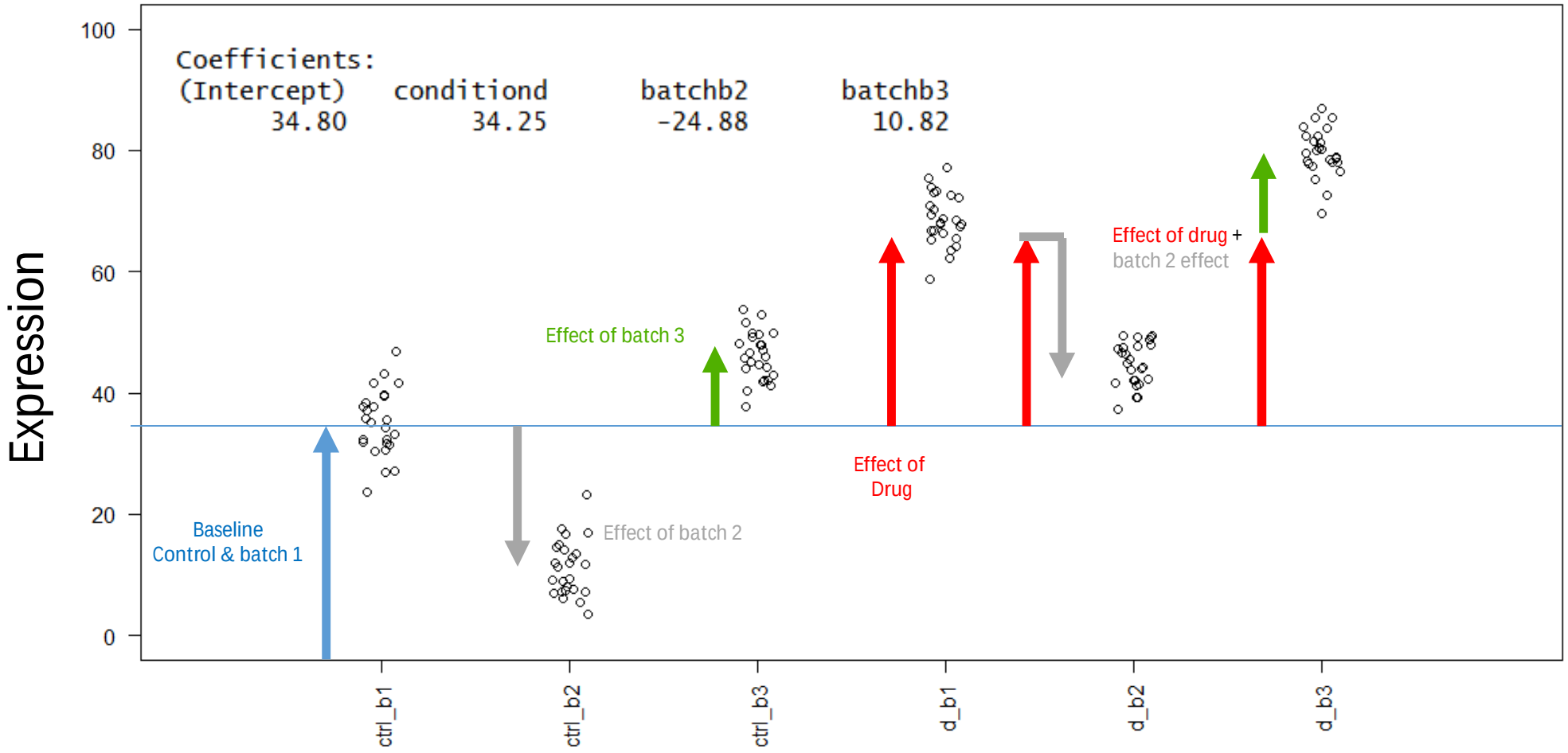

Coefficients:
(Intercept) conditiond batchb2 batchb3
34.80 34.25 -24.88 10.82
Expression
100
80
60
40
20
0
Effect of batch 3
Effect of Drug
Effect of batch 2
Baseline
Control & batch 1
ctrl_b1
ctrl_b2
ctrl_b3
d_b1
d_b2
d_b3

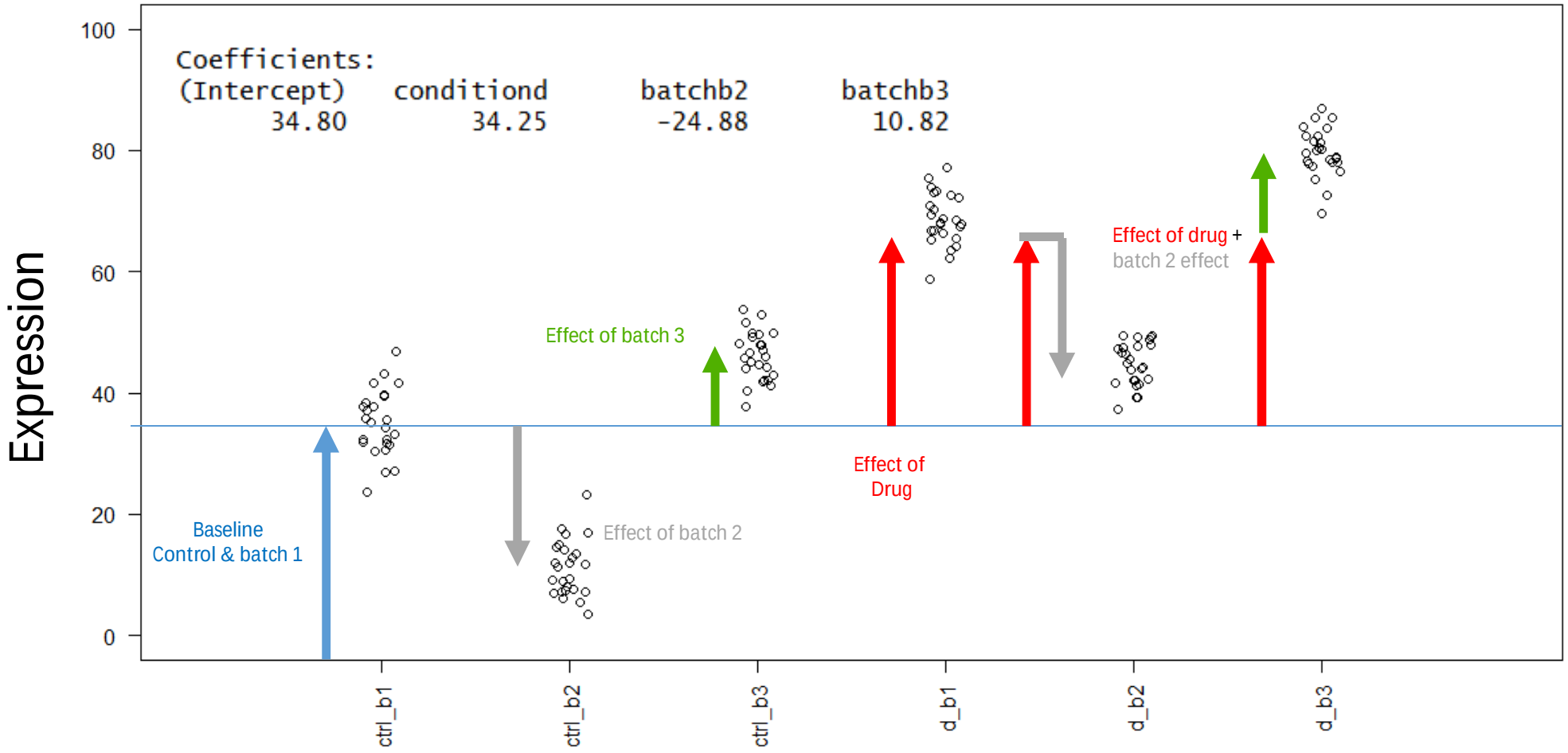

Coefficients:
(Intercept) conditiond batchb2 batchb3
34.80 34.25 -24.88 10.82
Expression
100
80
60
40
20
0
Baseline
Control & batch 1
Effect of batch 2
Effect of batch 3
Effect of
Drug
ctrl_b1
ctrl_b2
ctrl_b3
d_b1
d_b2
d_b3

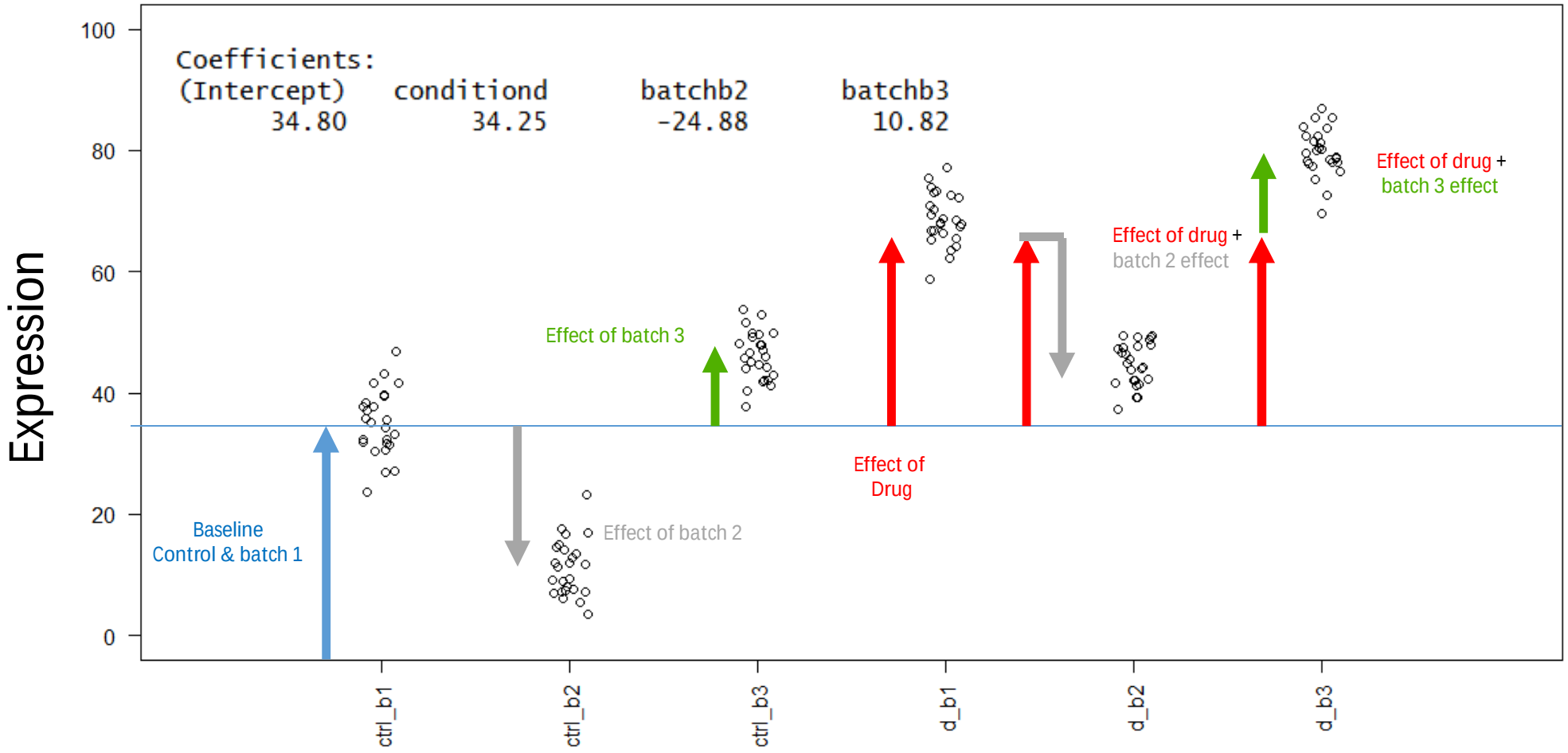

Coefficients:
(Intercept) conditiond batchb2 batchb3
34.80 34.25 -24.88 10.82
Expression
100
80
60
40
20
0
Baseline
Control & batch 1
Effect of batch 2
Effect of batch 3
Effect of
Drug
ctrl_b1
ctrl_b2
ctrl_b3
d_b1
d_b2
d_b3

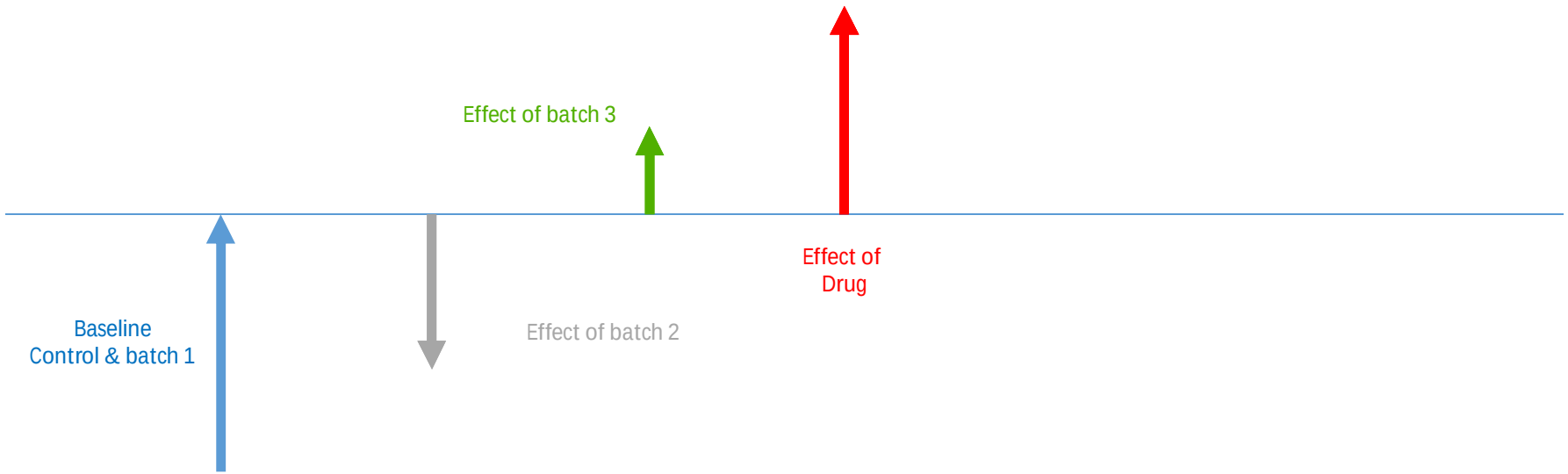

Coefficients:
(Intercept) conditiond batchb2 batchb3
34.80 34.25 -24.88 10.82
Expression
100
80
60
40
20
0
Baseline
Control & batch 1
Effect of batch 2
Effect of batch 3
Effect of
Drug
Effect of drug +
batch 2 effect
ctrl_b1
ctrl_b2
ctrl_b3
d_b1
d_b2
d_b3

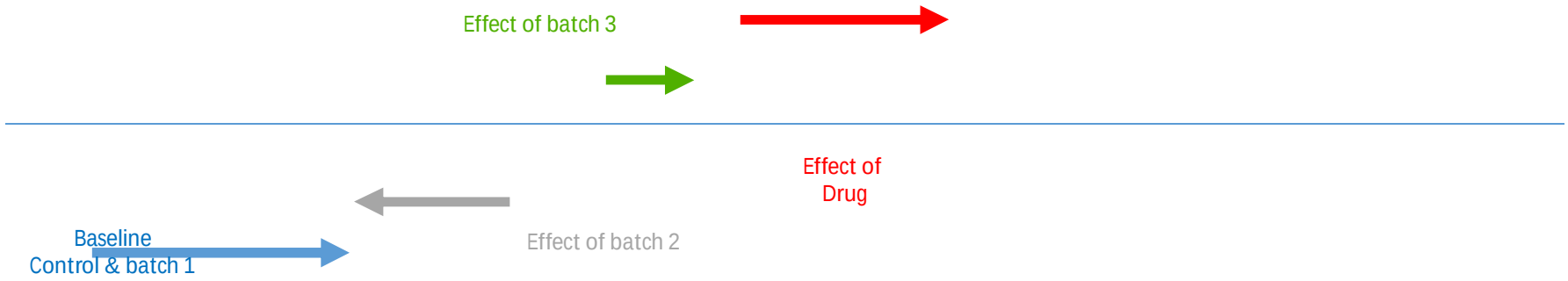

Coefficients:
(Intercept) conditiond batchb2 batchb3
34.80 34.25 -24.88 10.82
Expression
100
80
60
40
20
0
Baseline
Control & batch 1
Effect of batch 2
Effect of batch 3
Effect of
Drug
Effect of drug +
batch 2 effect
ctrl_b1
ctrl_b2
ctrl_b3
d_b1
d_b2
d_b3

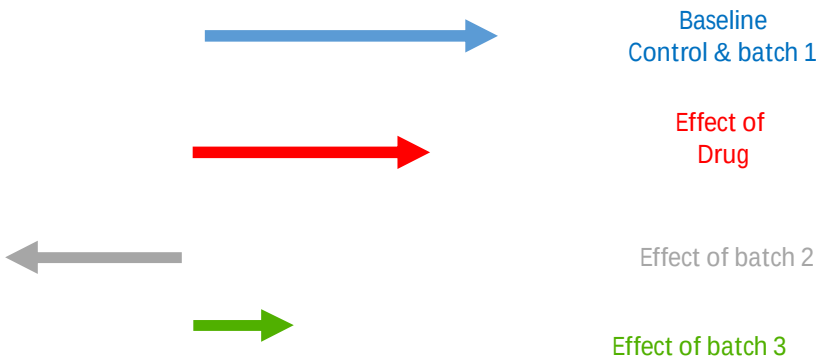

Coefficients:
(Intercept) conditiond batchb2 batchb3
34.80 34.25 -24.88 10.82
Expression
100
80
60
40
20
0
Baseline
Control & batch 1
Effect of batch 2
Effect of batch 3
Effect of
Drug
Effect of drug +
batch 2 effect
ctrl_b1
ctrl_b2
ctrl_b3
d_b1
d_b2
d_b3

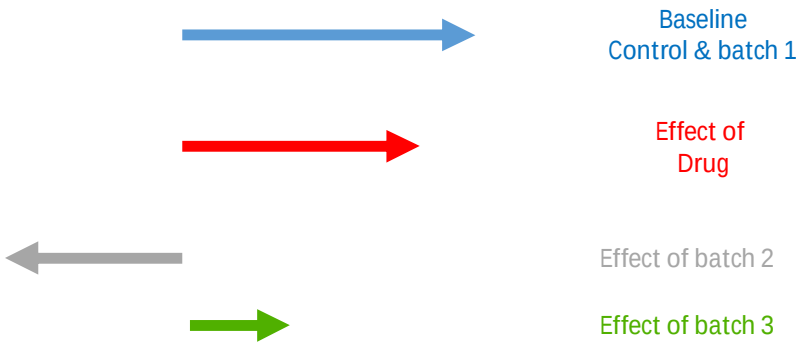

Coefficients:
(Intercept) conditiond batchb2 batchb3
34.80 34.25 -24.88 10.82
Expression
100
80
60
40
20
0
Baseline
Control & batch 1
Effect of batch 2
Effect of batch 3
Effect of
Drug
Effect of drug +
batch 2 effect
ctrl_b1
ctrl_b2
ctrl_b3
d_b1
d_b2
d_b3

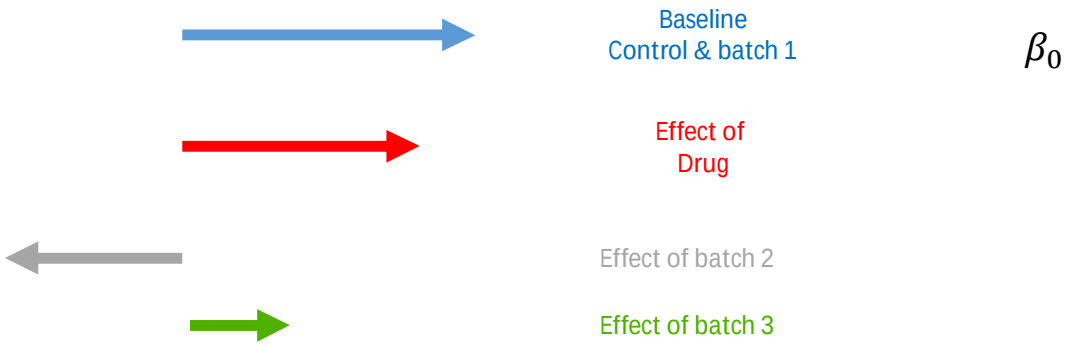

Coefficients:
(Intercept) conditiond batchb2 batchb3
34.80 34.25 -24.88 10.82
Expression
100
80
60
40
20
0
Baseline
Control & batch 1
Effect of batch 2
Effect of batch 3
Effect of
Drug
Effect of drug +
batch 2 effect
Effect of drug +
batch 3 effect
ctrl_b1
ctrl_b2
ctrl_b3
d_b1
d_b2
d_b3


Effect of batch 3
Effect of
Drug
Baseline
Control & batch 1
Effect of batch 2

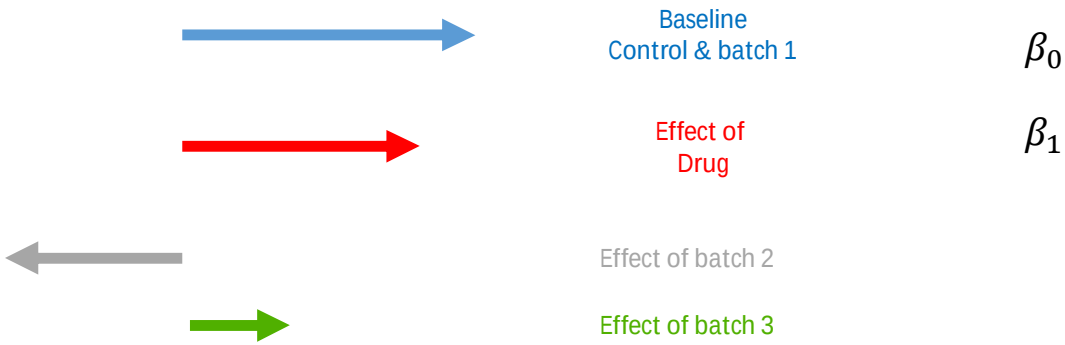

Effect of batch 3
Effect of
Drug
Effect of batch 2
Baseline
Control & batch 1

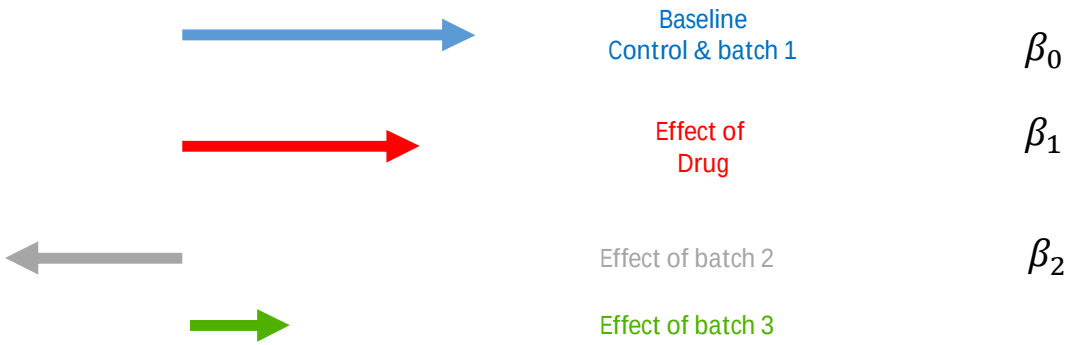

Baseline
Control & batch 1
Effect of
Drug
Effect of batch 2
Effect of batch 3

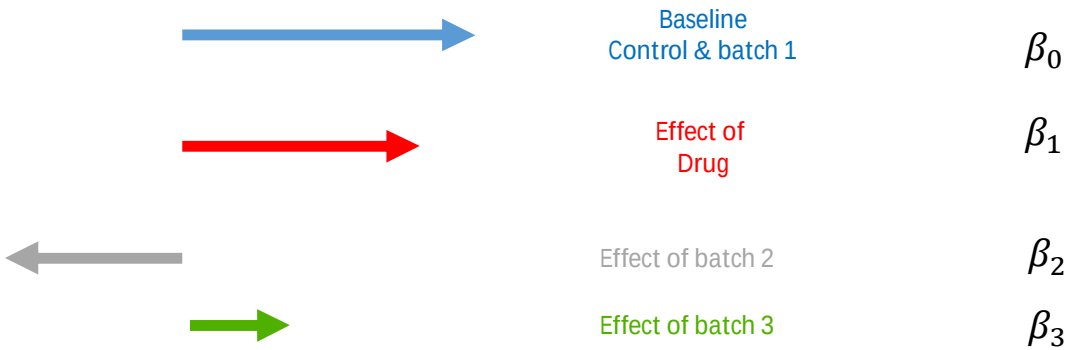

Baseline
Control & batch 1
Effect of
Drug
Effect of batch 2
Effect of batch 3

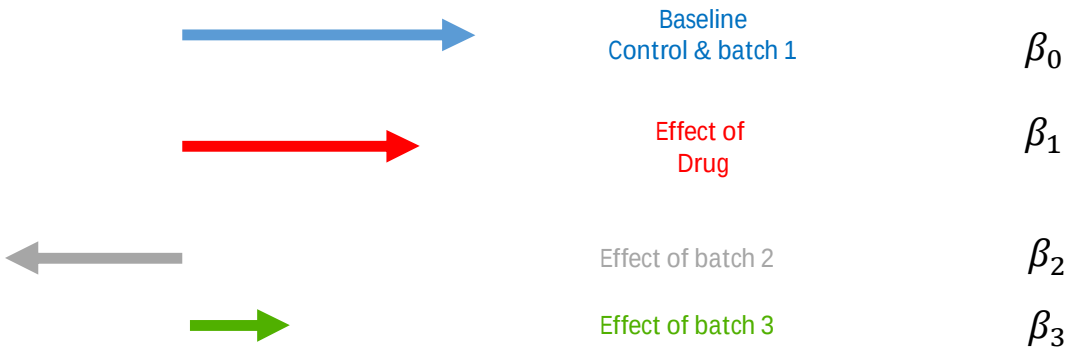

Baseline
Control & batch 1
Effect of
Drug
Effect of batch 2
Effect of batch 3

β_0

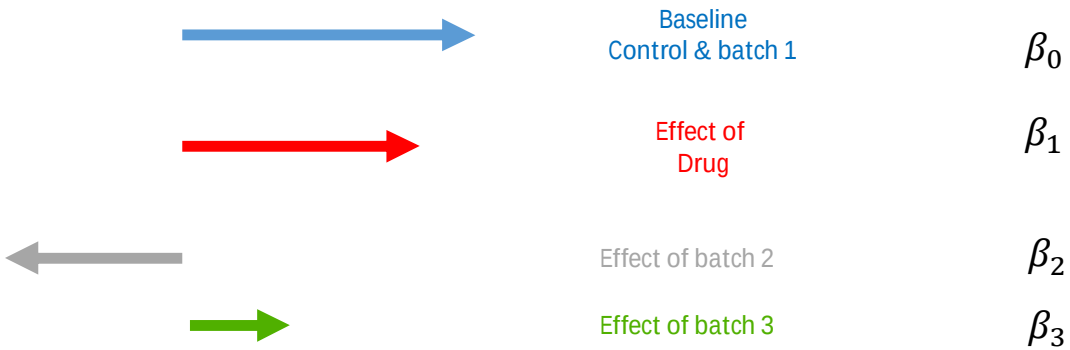

Baseline
Control & batch 1
β_0
Effect of
Drug
β_1
Effect of batch 2
Effect of batch 3


Baseline
Control & batch 1
β_0
Effect of
Drug
β_1
Effect of batch 2
β_2
Effect of batch 3


Baseline
Control & batch 1
β_0
Effect of
Drug
β_1
Effect of batch 2
β_2
Effect of batch 3
β_3

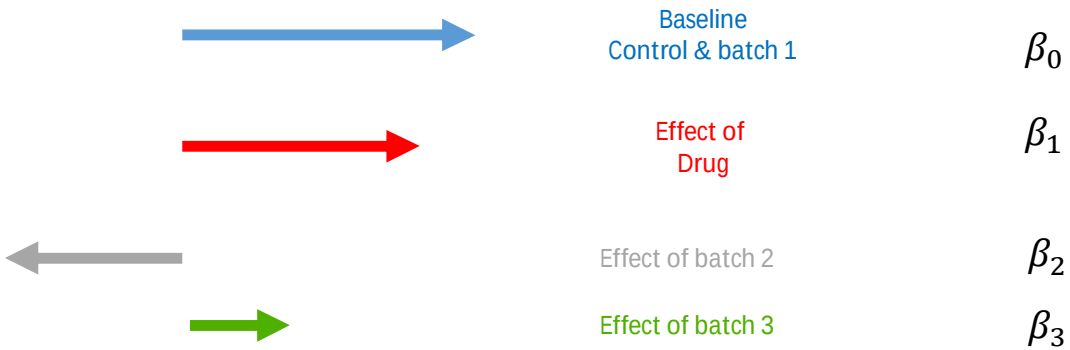


We effectively have partitioned the variance.



We effectively have partitioned the variance.

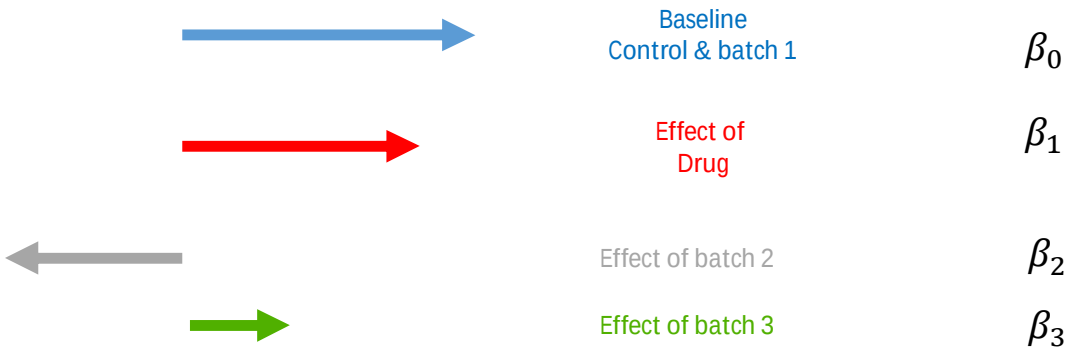
That is, we separated batch effects from treatment



We effectively have partitioned the variance.

That is, we separated batch effects from treatment

Now we can test if there is a effect of drug and if it's significant

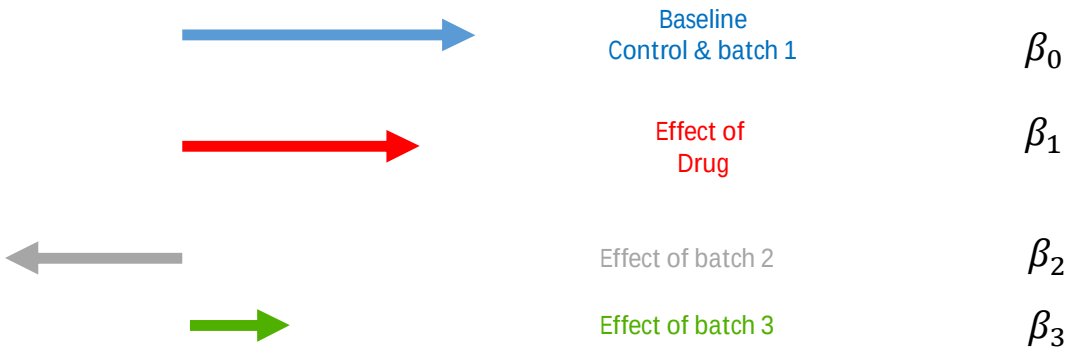


We effectively have partitioned the variance.

That is, we separated batch effects from treatment

Now we can test if there is a effect of drug and if it's significant

Null hypothesis : H_0: $\beta_1 = 0$



We effectively have partitioned the variance.

That is, we separated batch effects from treatment

Now we can test if there is a effect of drug and if it's significant

Null hypothesis : H_0: $\beta_1 = 0$ Test statistics: $\frac{\beta_1}{SE(\beta_1)} \sim t_{df}$

Effect of Drug

β_1

We effectively have partitioned the variance.

That is, we separated batch effects from treatment

Now we can test if there is a effect of drug and if it's significant

Null hypothesis : H_0: $\beta_1 = 0$ Test statistics: $\frac{\beta_1}{SE(\beta_1)} \sim t_{df}$

Effect of Drug

β_1

33.6132

We effectively have partitioned the variance.

That is, we separated batch effects from treatment

Now we can test if there is a effect of drug and if it's significant

Null hypothesis : H_0: $\beta_1 = 0$ Test statistics: $\frac{\beta_1}{SE(\beta_1)} \sim t_{df}$

Effect of Drug

β_1	$SE(\beta_1)$
33.6132	0.7102

We effectively have partitioned the variance.

That is, we separated batch effects from treatment

Now we can test if there is a effect of drug and if it's significant

Null hypothesis : H_0: $\beta_1 = 0$ Test statistics: $\frac{\beta_1}{SE(\beta_1)} \sim t_{df}$

Effect of Drug

β_1	$SE(\beta_1)$	t-statistic
33.6132	0.7102	47.33

We effectively have partitioned the variance.

That is, we separated batch effects from treatment

Now we can test if there is a effect of drug and if it's significant

Null hypothesis : H_0: $\beta_1 = 0$ Test statistics: $\frac{\beta_1}{SE(\beta_1)} \sim t_{df}$

Effect of Drug

β_1	$SE(\beta_1)$	t-statistic	p-value
33.6132	0.7102	47.33	<2e-16

We effectively have partitioned the variance.

That is, we separated batch effects from treatment

Now we can test if there is a effect of drug and if it's significant

Null hypothesis : H_0: $\beta_1 = 0$ Test statistics: $\frac{\beta_1}{SE(\beta_1)} \sim t_{df}$

Level 3

“What are the genes that are differentially expressed between drugged mice in time 2 and in time1 while controlling for vehicle effects?”

Research Question

- Setting: Mice are randomly assigned to drug and vehicle groups and their expression measures are obtained at time 1, time 2, and time3 (not repeated measure design)

3. Time series

	Control	Drug
Time 1		
Time 2		
Time 3		

3. Time series

	Control	Drug
Time 1		
Time 2		
Time 3		

3. Time series

	Control	Drug
Time 1		
Time 2		
Time 3		

3. Time series

	Control	Drug
Time 1		
Time 2		
Time 3		

3. Time series

	Control	Drug
Time 1		
Time 2		
Time 3		

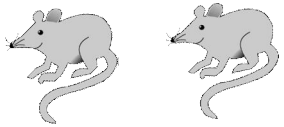
3. Time series

	Control	Drug
Time 1		
Time 2		
Time 3		 

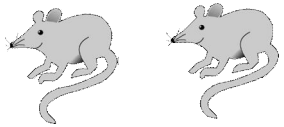
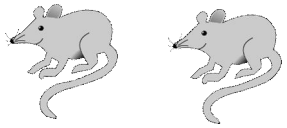
3. Time series

	Control	Drug
Time 1		
Time 2		
Time 3		 

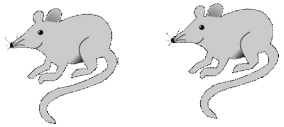
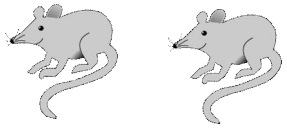
3. Time series

	Control	Drug
Time 1		
Time 2		
Time 3		

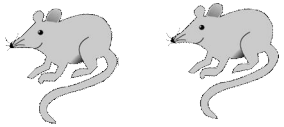
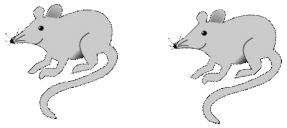
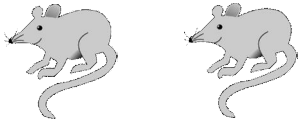
3. Time series

	Control	Drug
Time 1		
Time 2		
Time 3		

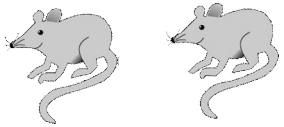
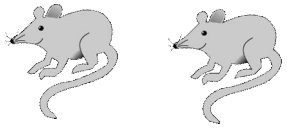
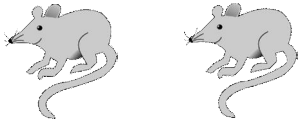
3. Time series

	Control	Drug
Time 1		
Time 2		
Time 3		

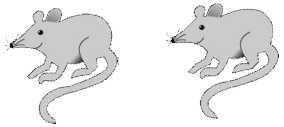
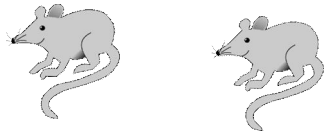
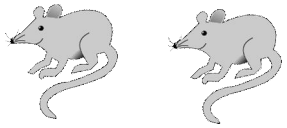
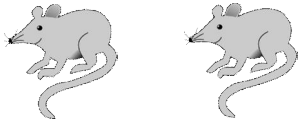
3. Time series

	Control	Drug
Time 1		
Time 2		
Time 3		

3. Time series

	Control	Drug
Time 1		
Time 2		
Time 3		

3. Time series

	Control	Drug
Time 1		
Time 2		
Time 3		

Data and model

Data and model

Data and model

class	time	expression
control	t1	16.861675
control	t1	16.419413
control	t1	11.655116
control	t2	17.914727
control	t2	12.319448
control	t2	19.408569
control	t3	29.051403
control	t3	32.112288
control	t3	31.293219
drug	t1	6.850348
drug	t1	4.649544
drug	t1	4.786212
drug	t2	22.730259
drug	t2	24.259060
drug	t2	21.780425
drug	t3	34.377084
drug	t3	29.345437
drug	t3	25.977785

Data and model

class	time	expression
control	t1	16.861675
control	t1	16.419413
control	t1	11.655116
control	t2	17.914727
control	t2	12.319448
control	t2	19.408569
control	t3	29.051403
control	t3	32.112288
control	t3	31.293219
drug	t1	6.850348
drug	t1	4.649544
drug	t1	4.786212
drug	t2	22.730259
drug	t2	24.259060
drug	t2	21.780425
drug	t3	34.377084
drug	t3	29.345437
drug	t3	25.977785

Model :

Data and model

class	time	expression
control	t1	16.861675
control	t1	16.419413
control	t1	11.655116
control	t2	17.914727
control	t2	12.319448
control	t2	19.408569
control	t3	29.051403
control	t3	32.112288
control	t3	31.293219
drug	t1	6.850348
drug	t1	4.649544
drug	t1	4.786212
drug	t2	22.730259
drug	t2	24.259060
drug	t2	21.780425
drug	t3	34.377084
drug	t3	29.345437
drug	t3	25.977785

Model :

$$expression \sim class + time$$

Data and model

class	time	expression
control	t1	16.861675
control	t1	16.419413
control	t1	11.655116
control	t2	17.914727
control	t2	12.319448
control	t2	19.408569
control	t3	29.051403
control	t3	32.112288
control	t3	31.293219
drug	t1	6.850348
drug	t1	4.649544
drug	t1	4.786212
drug	t2	22.730259
drug	t2	24.259060
drug	t2	21.780425
drug	t3	34.377084
drug	t3	29.345437
drug	t3	25.977785

Model :

$$expression \sim class + time + time{:}class$$

Data and model

class	time	expression
control	t1	16.861675
control	t1	16.419413
control	t1	11.655116
control	t2	17.914727
control	t2	12.319448
control	t2	19.408569
control	t3	29.051403
control	t3	32.112288
control	t3	31.293219
drug	t1	6.850348
drug	t1	4.649544
drug	t1	4.786212
drug	t2	22.730259
drug	t2	24.259060
drug	t2	21.780425
drug	t3	34.377084
drug	t3	29.345437
drug	t3	25.977785

Model :

$$expression \sim class + time + time{:}\,class$$

Interaction term

Data and model

class	time	expression
control	t1	16.861675
control	t1	16.419413
control	t1	11.655116
control	t2	17.914727
control	t2	12.319448
control	t2	19.408569
control	t3	29.051403
control	t3	32.112288
control	t3	31.293219
drug	t1	6.850348
drug	t1	4.649544
drug	t1	4.786212
drug	t2	22.730259
drug	t2	24.259060
drug	t2	21.780425
drug	t3	34.377084
drug	t3	29.345437
drug	t3	25.977785

Model :

$$expression \sim class + time + time{:}class$$

Interaction term

Interaction : the effect of a factor on the response variable is different depending on the level of another factor

Data and model

class	time	expression
control	t1	16.861675
control	t1	16.419413
control	t1	11.655116
control	t2	17.914727
control	t2	12.319448
control	t2	19.408569
control	t3	29.051403
control	t3	32.112288
control	t3	31.293219
drug	t1	6.850348
drug	t1	4.649544
drug	t1	4.786212
drug	t2	22.730259
drug	t2	24.259060
drug	t2	21.780425
drug	t3	34.377084
drug	t3	29.345437
drug	t3	25.977785

Model :

$$expression \sim class + time + time{:}class$$

Interaction term

Interaction : the effect of a factor on the response variable is different depending on the level of another factor

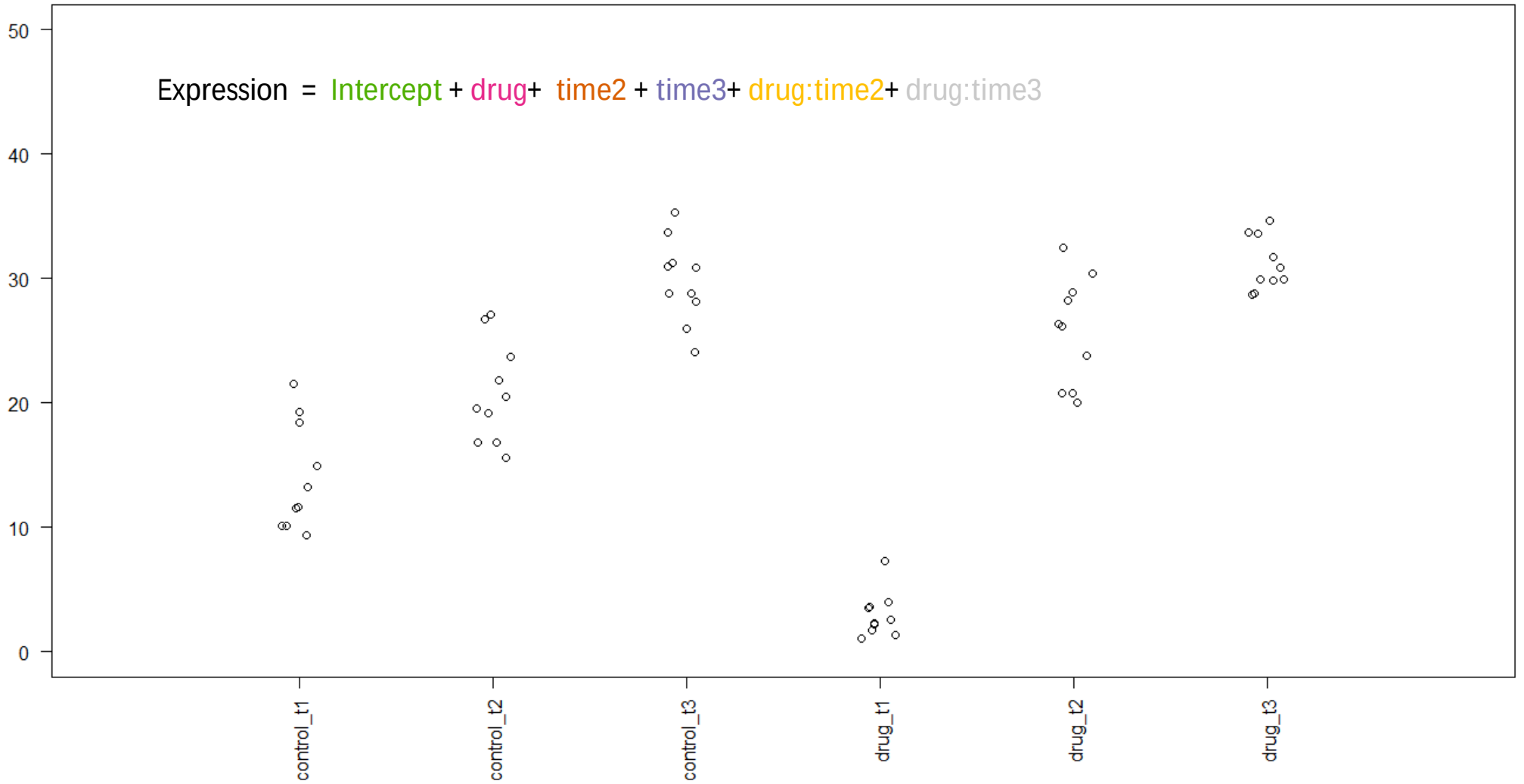
Ex) drug will act differently in t2 from control in t2

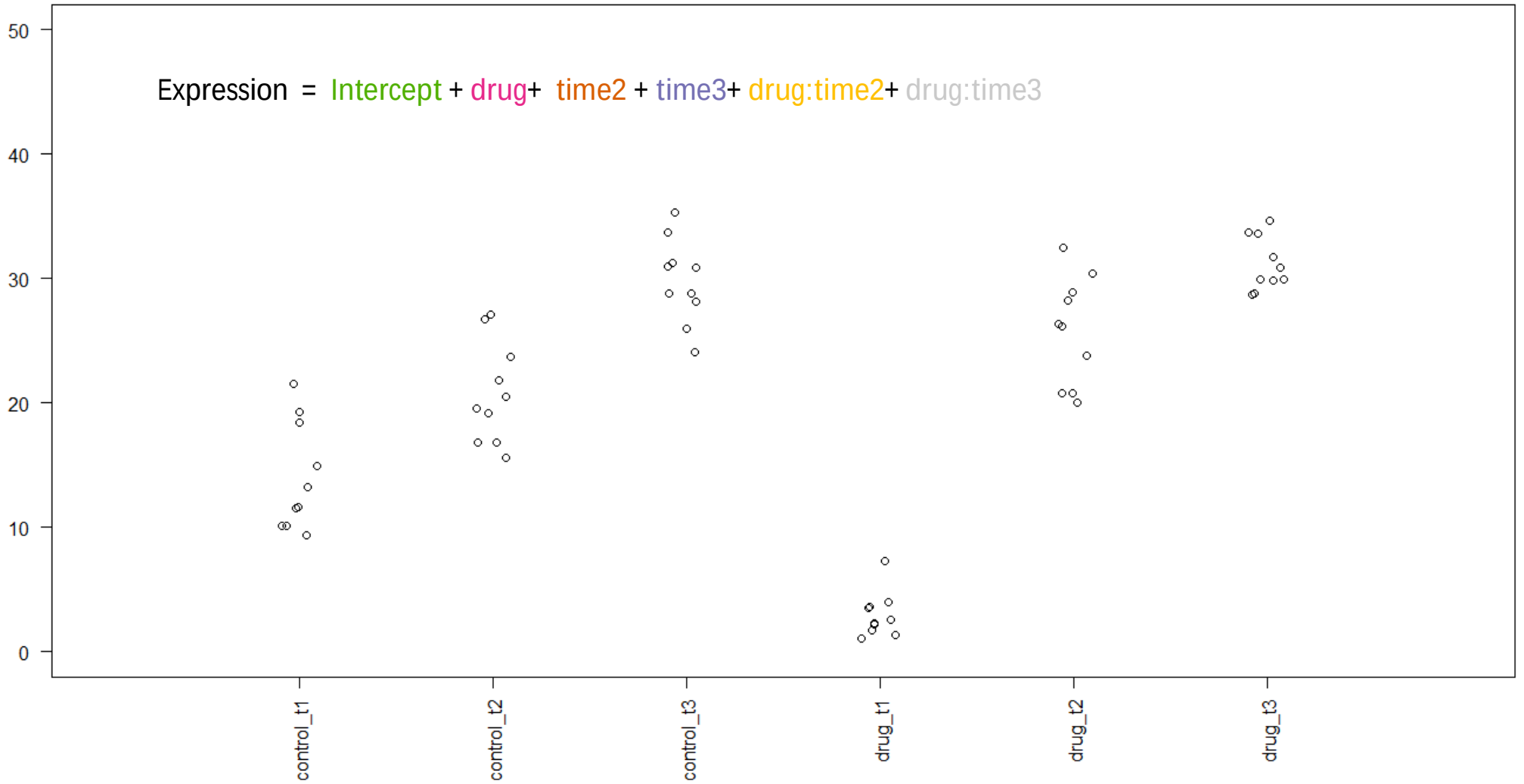
In R..

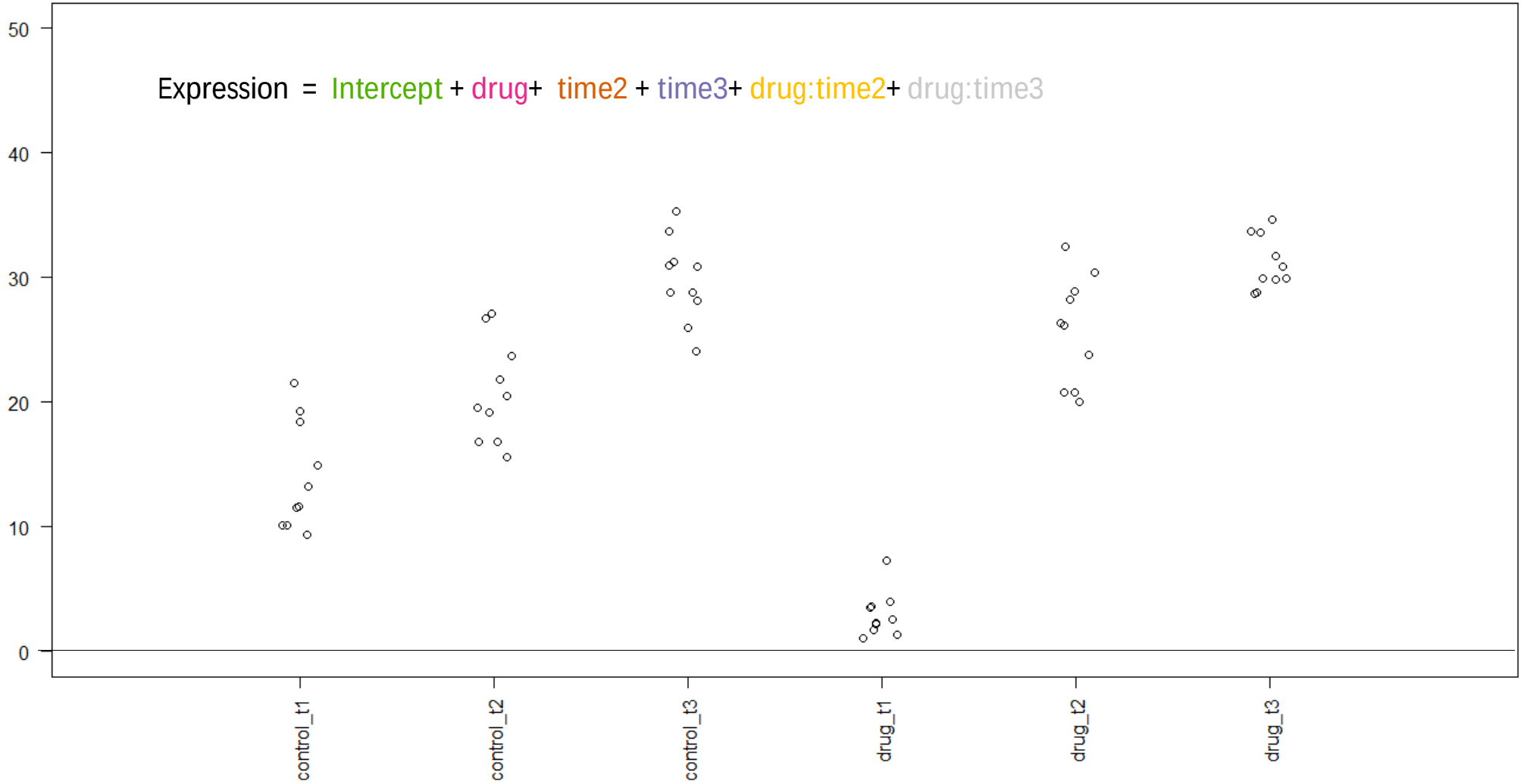
```
> m3 <- lm(expression ~ class + time + time:class , data=data)
> summary(m3)
```

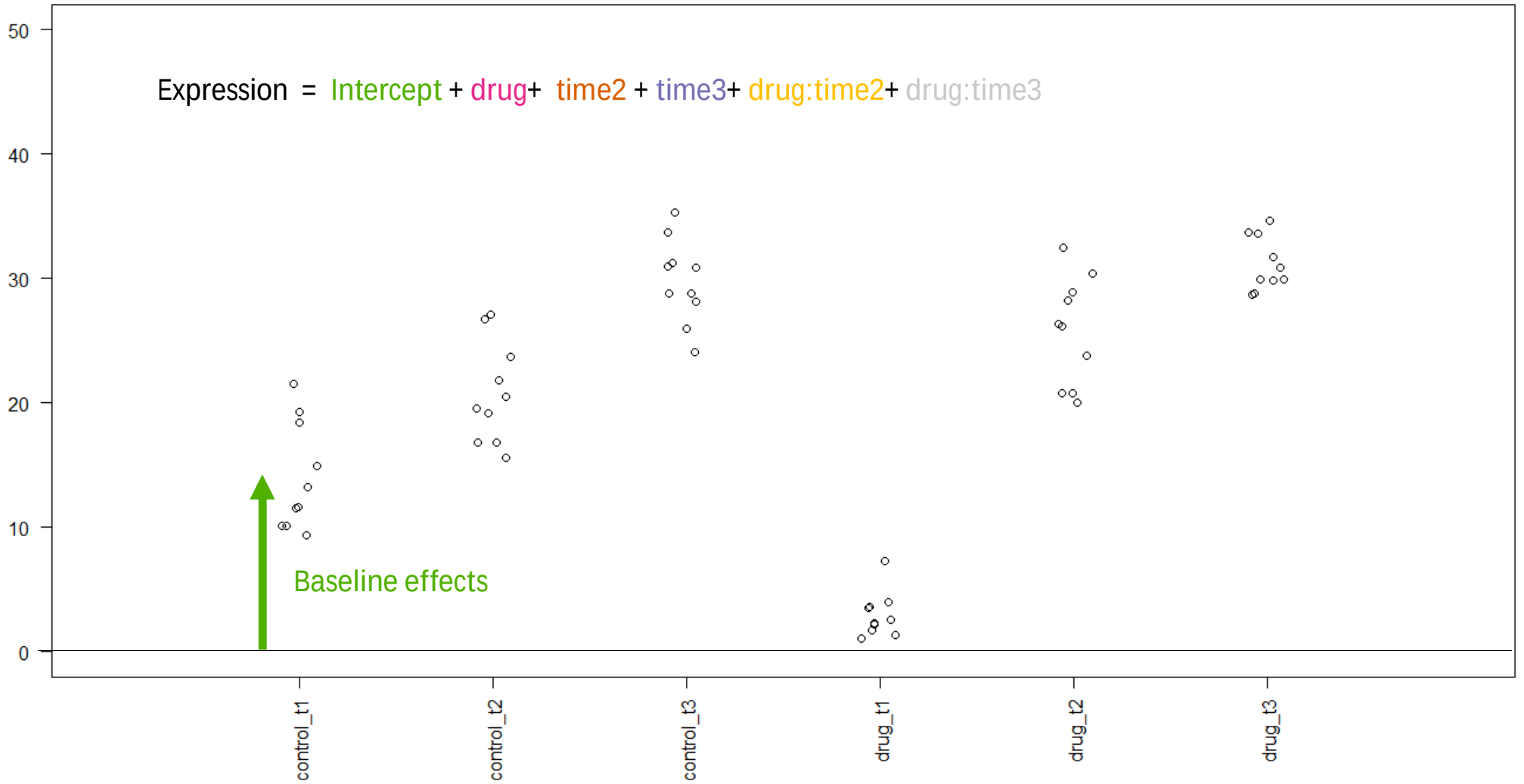
In R..

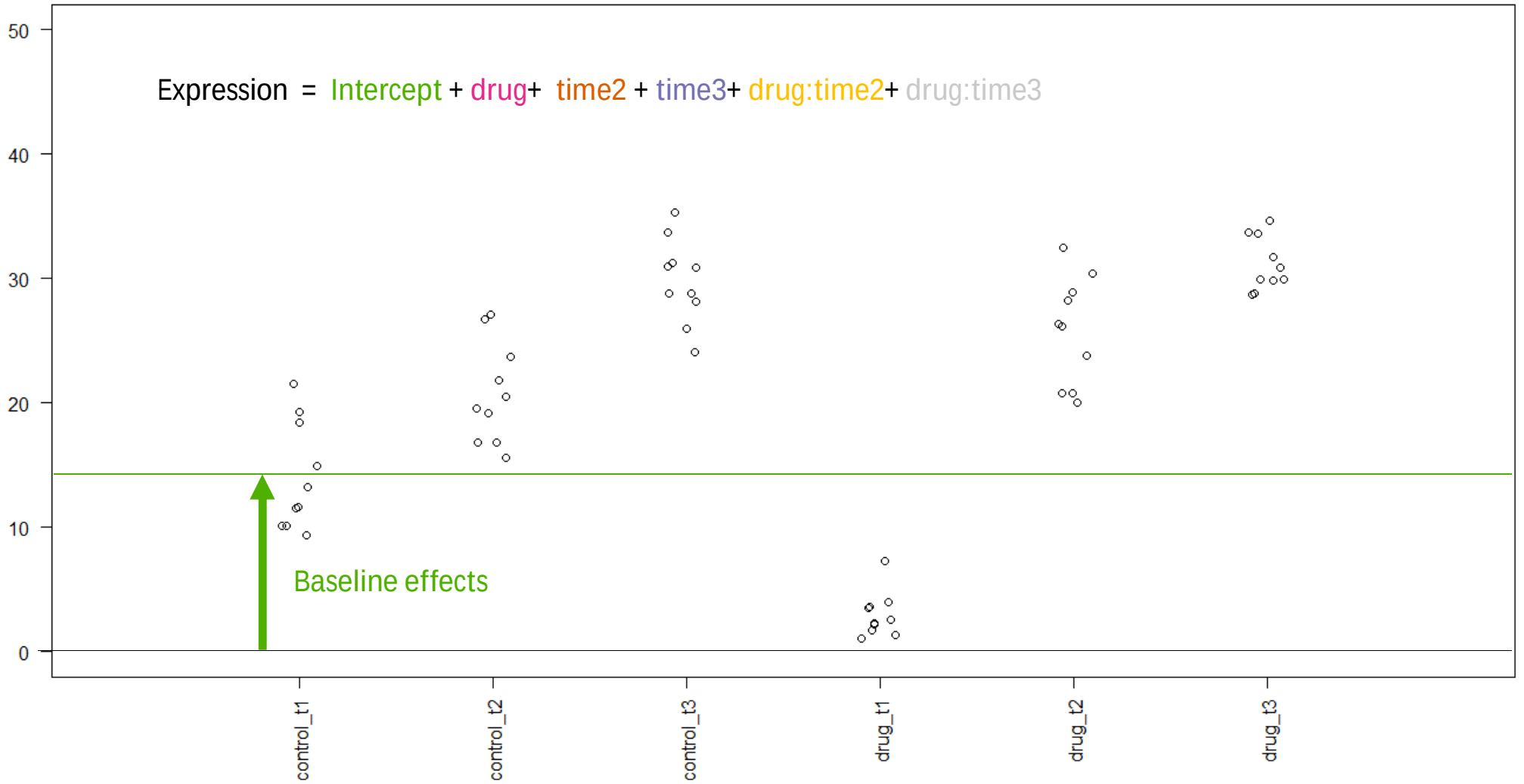
```
> m3 <- lm(expression ~ class + time + time:class , data=data)
> summary(m3)

Call:
lm(formula = expression ~ class + time + time:class, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7107 -1.9776 -0.3892  1.8724  7.3487

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       13.0269     0.9788  13.309  < 2e-16 ***
classdrug         -8.0131     1.3843  -5.789 3.71e-07 ***
timet2             7.0032     1.3843   5.059 5.20e-06 ***
timet3            17.1055     1.3843  12.357  < 2e-16 ***
classdrug:timet2  11.7185     1.9576   5.986 1.80e-07 ***
classdrug:timet3   7.6516     1.9576   3.909 0.000261 ***
```

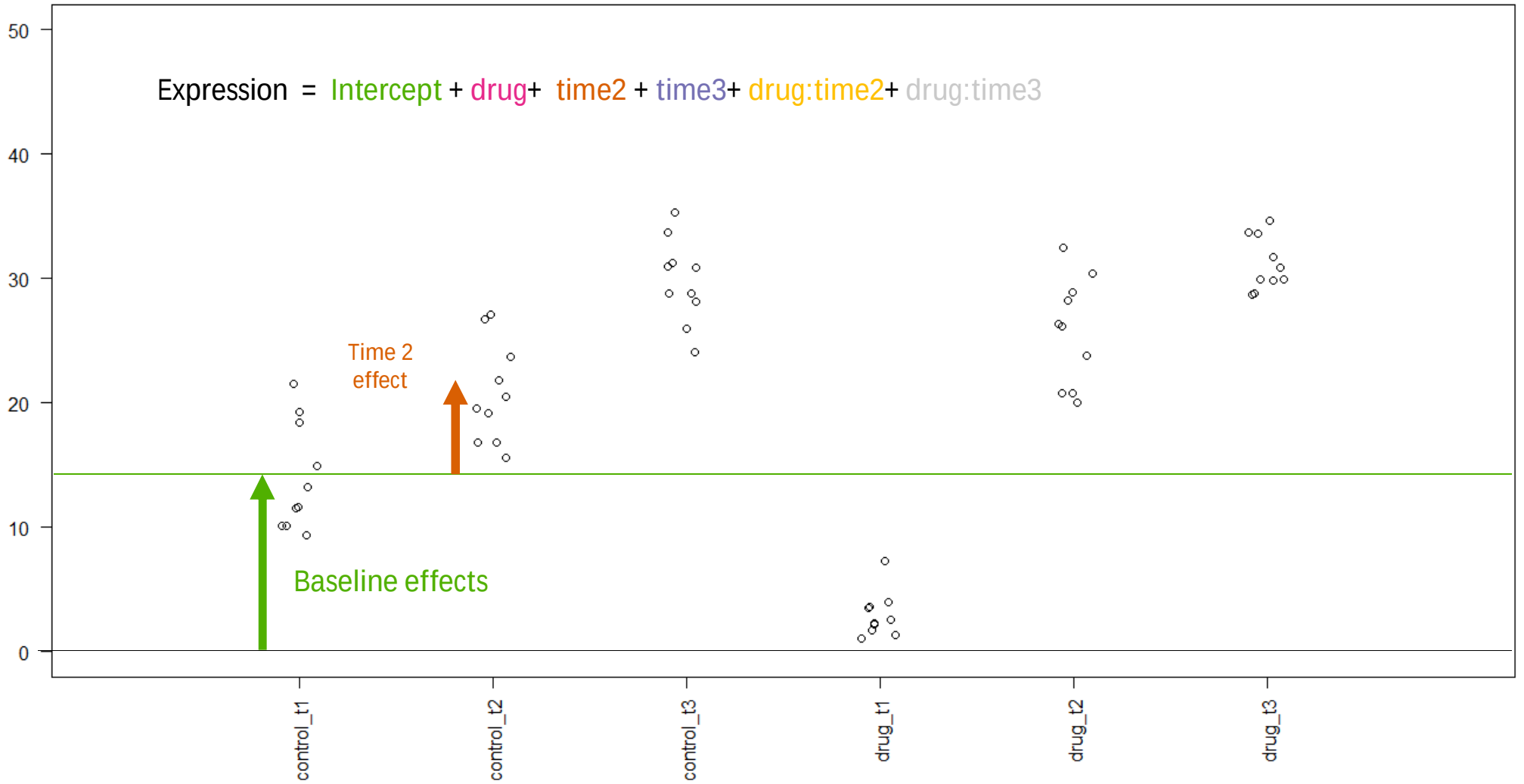

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

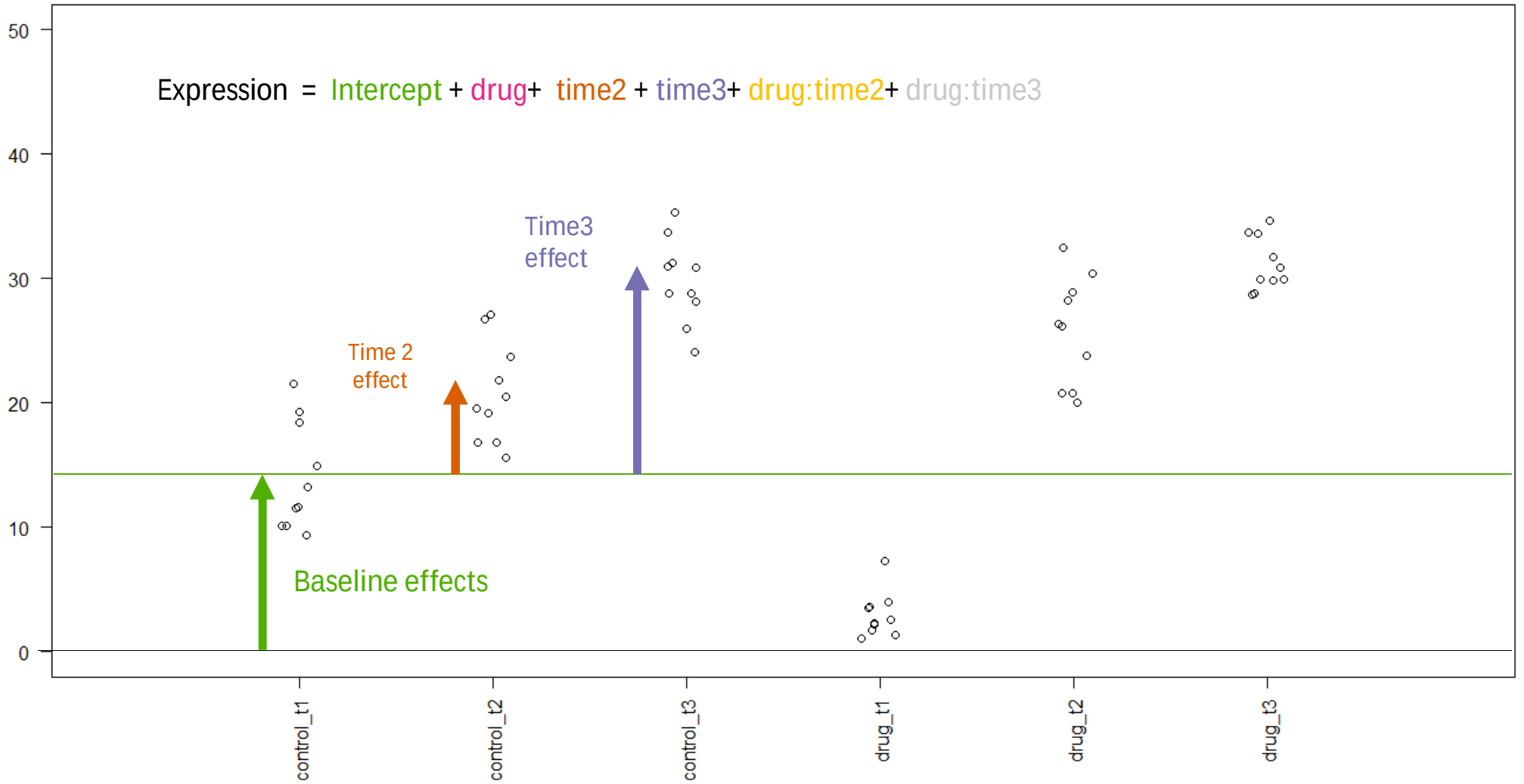

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

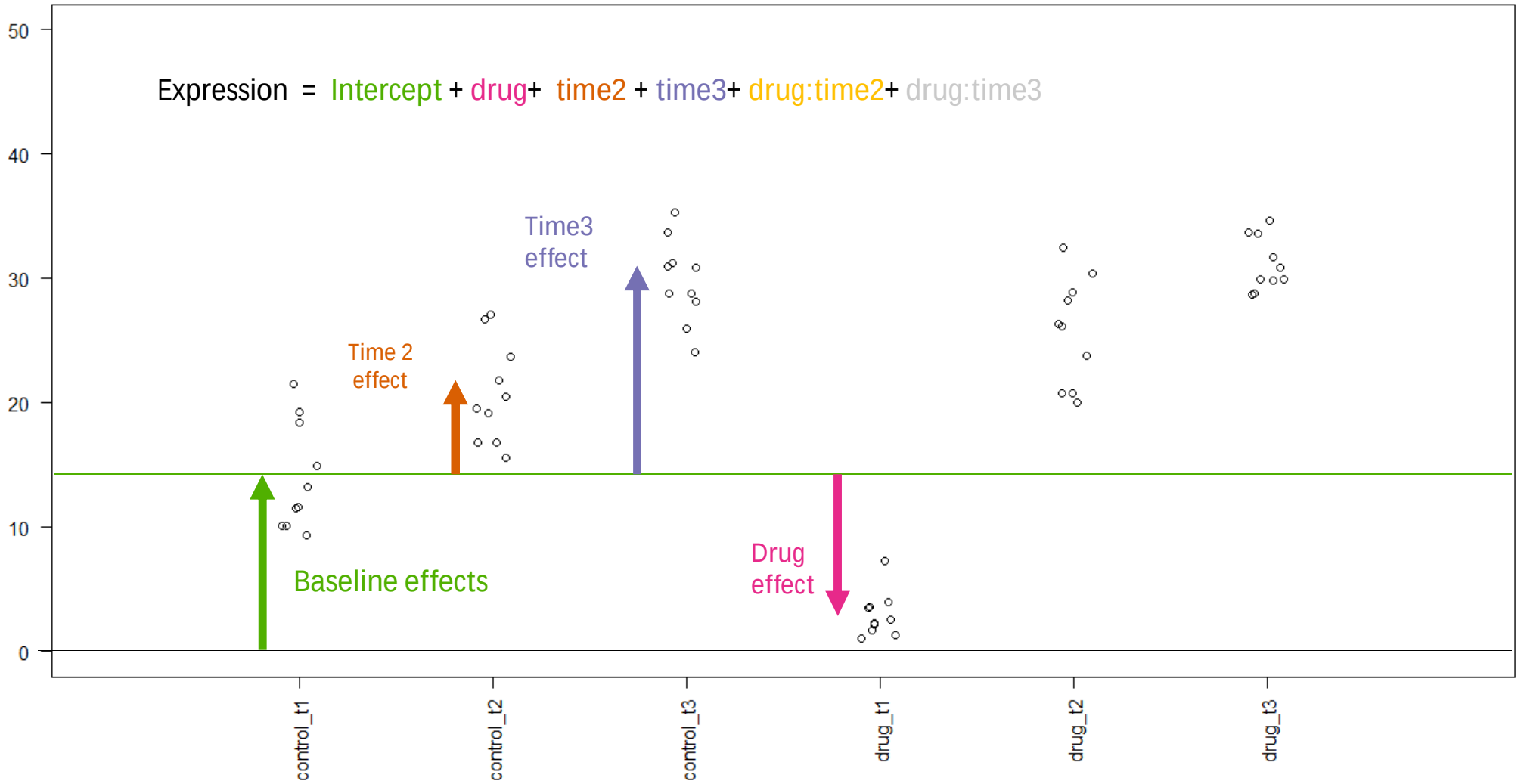

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

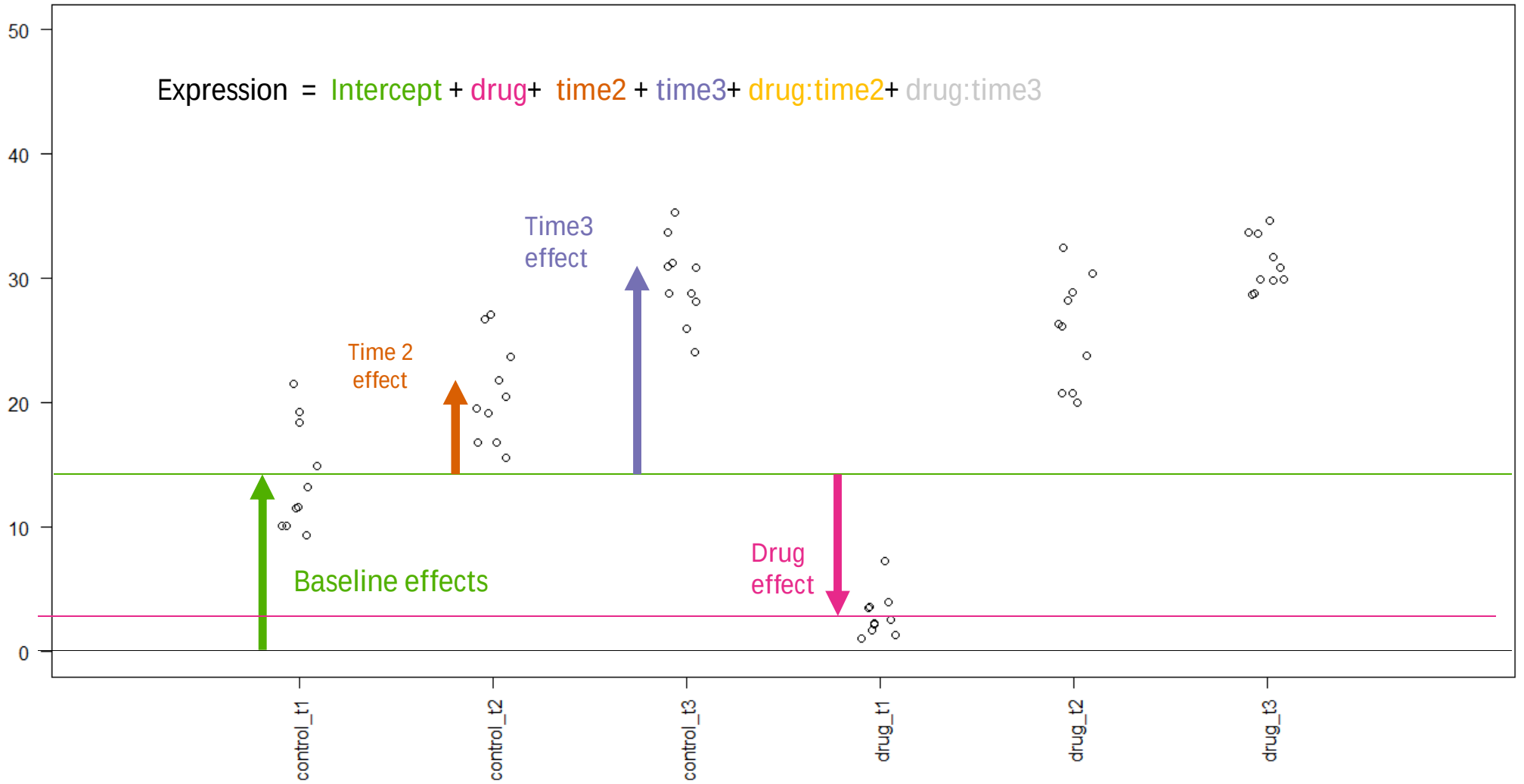

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
50
40
30
20
10
0
Baseline effects
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

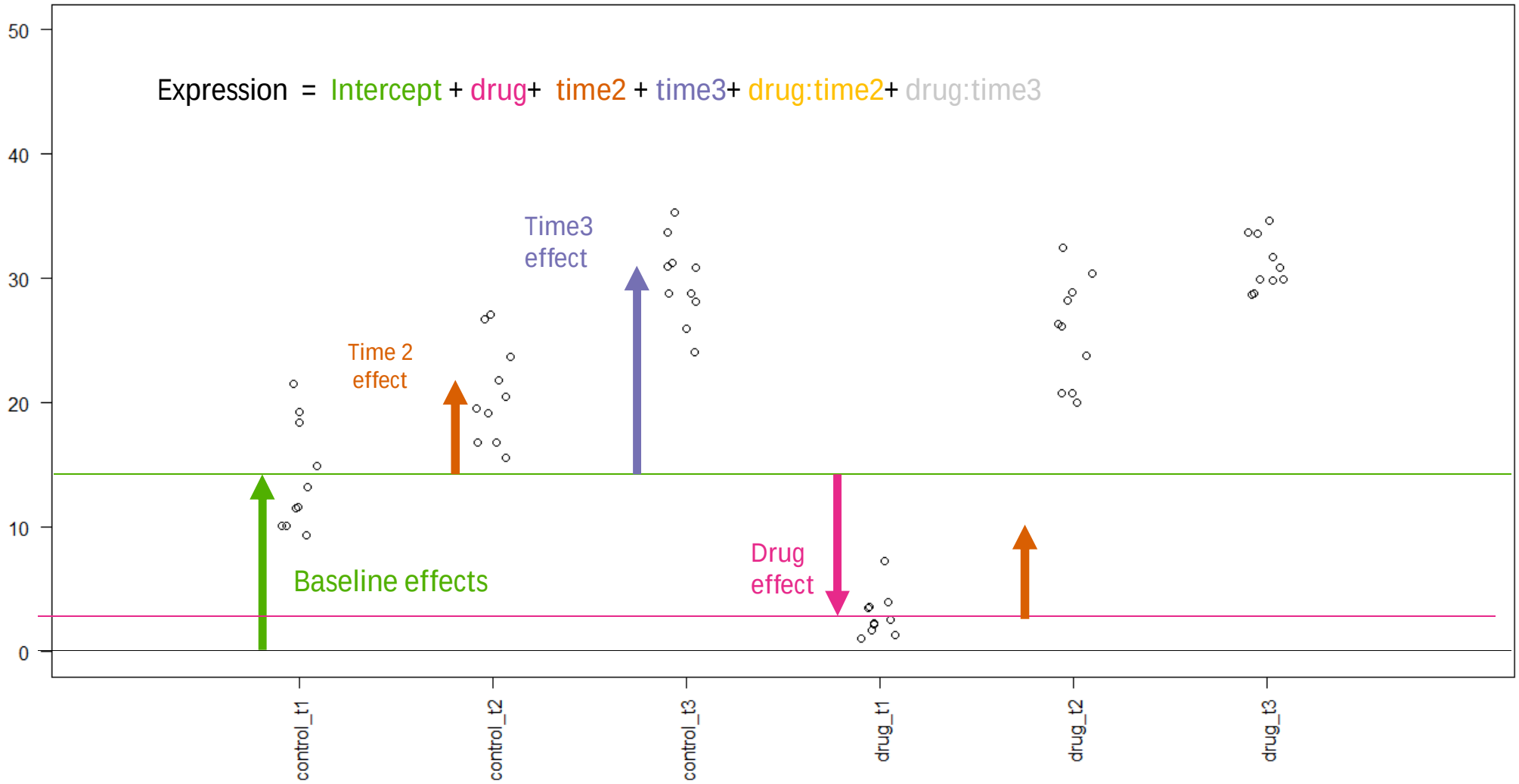

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
50
40
30
20
10
0
Baseline effects
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

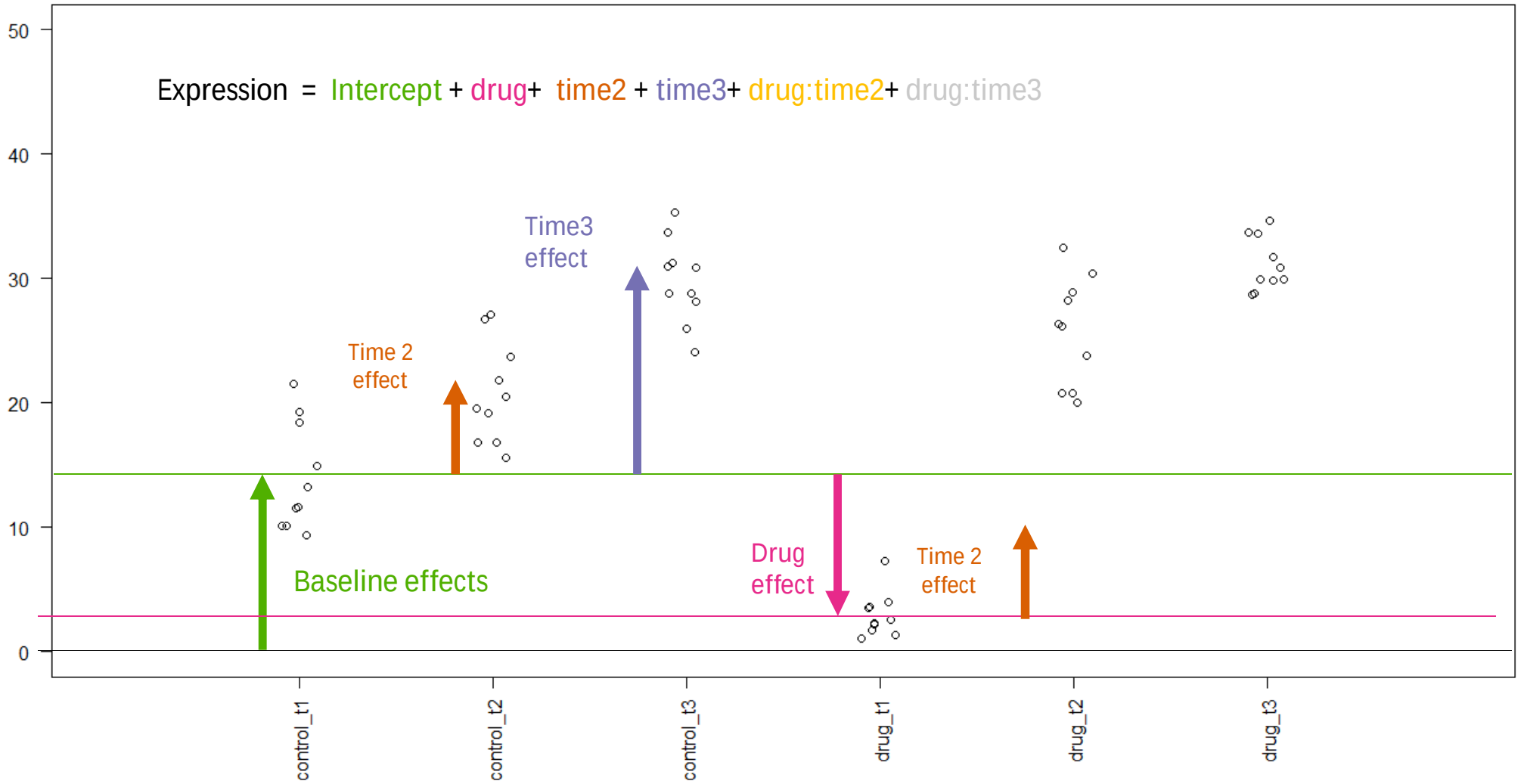

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time 2 effect
Baseline effects
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

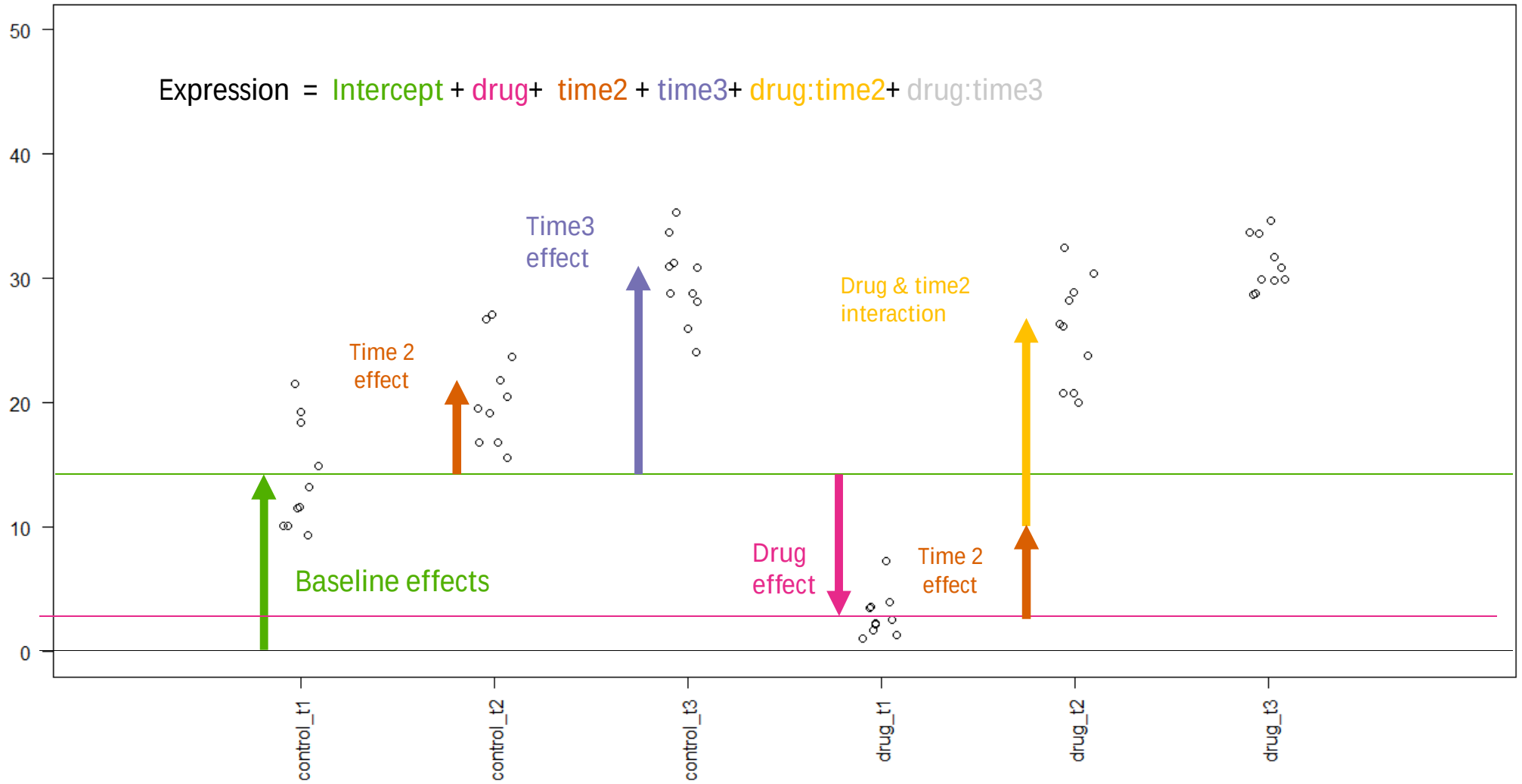

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time3 effect
Time 2 effect
Baseline effects
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

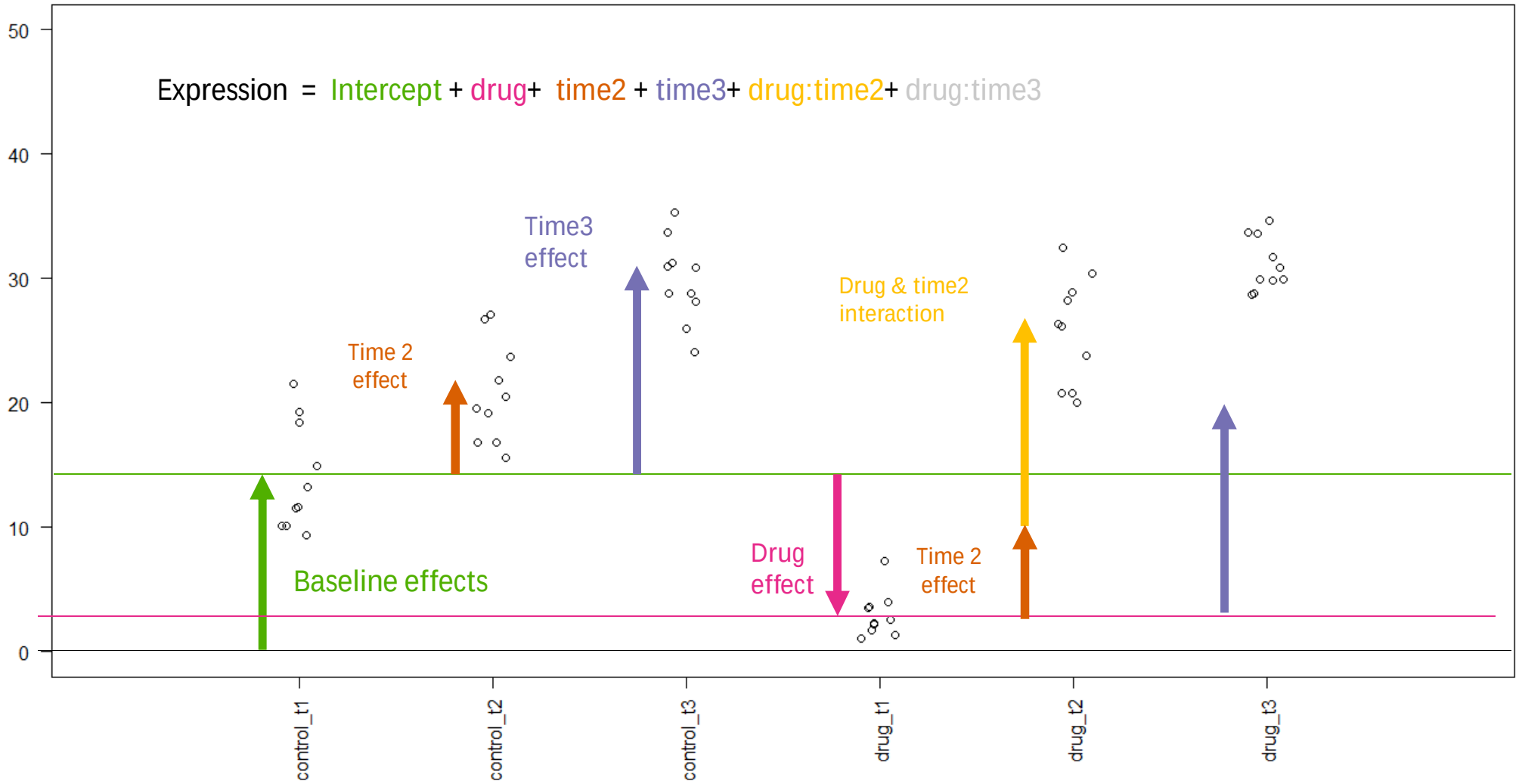

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time3 effect
Time 2 effect
Baseline effects
Drug effect
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

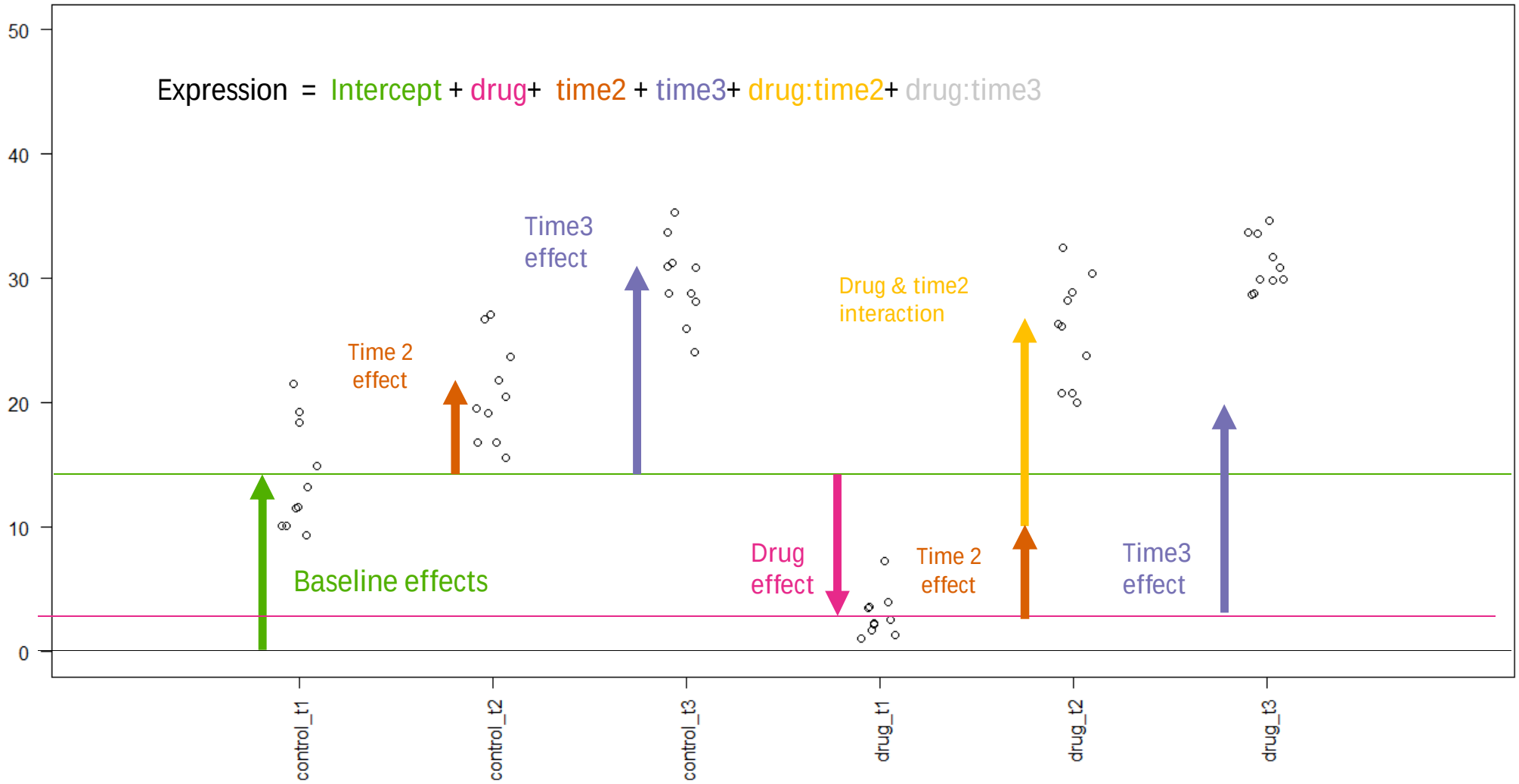

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time3 effect
Time 2 effect
Drug effect
Baseline effects
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

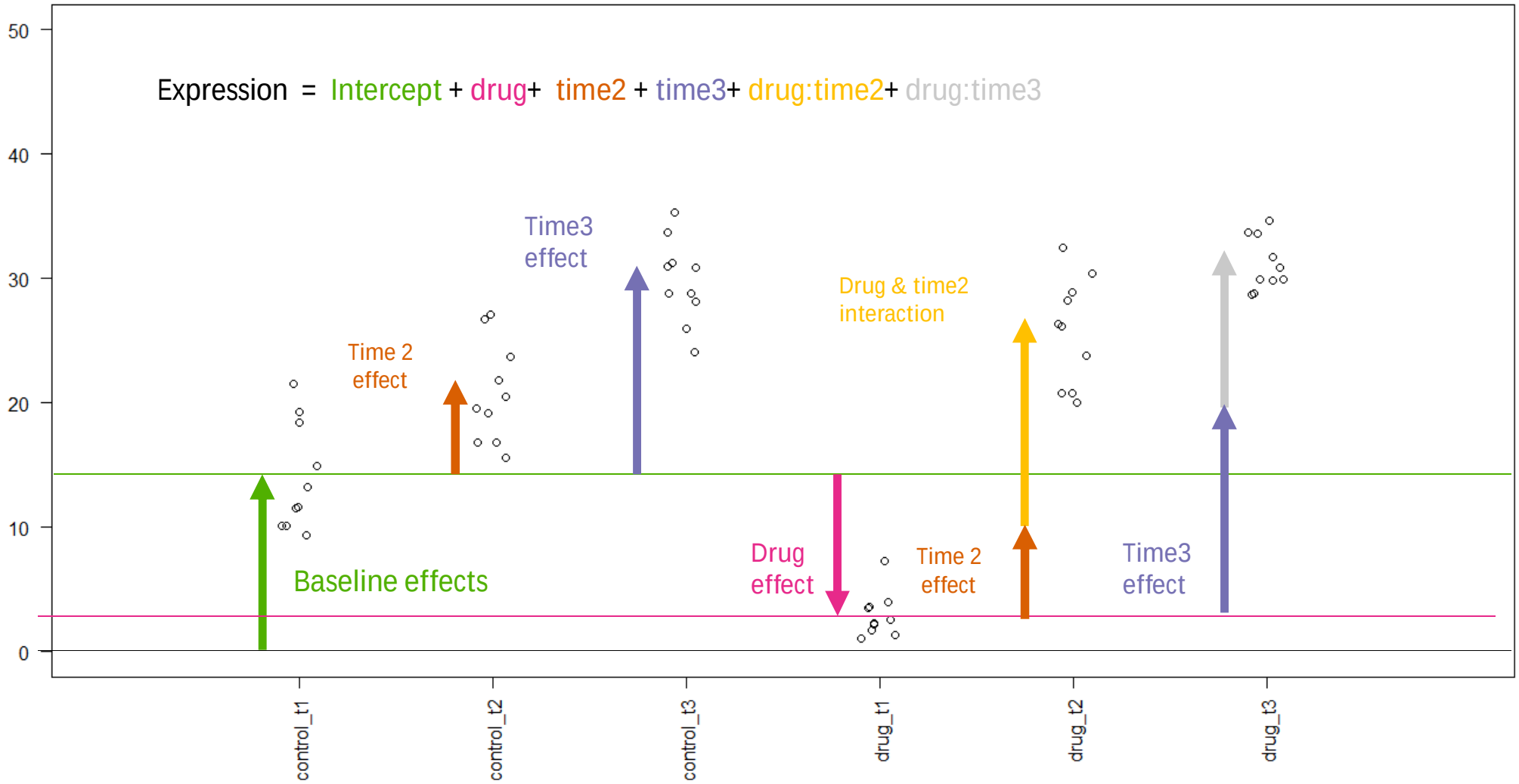

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time3 effect
Time 2 effect
Baseline effects
Drug effect
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

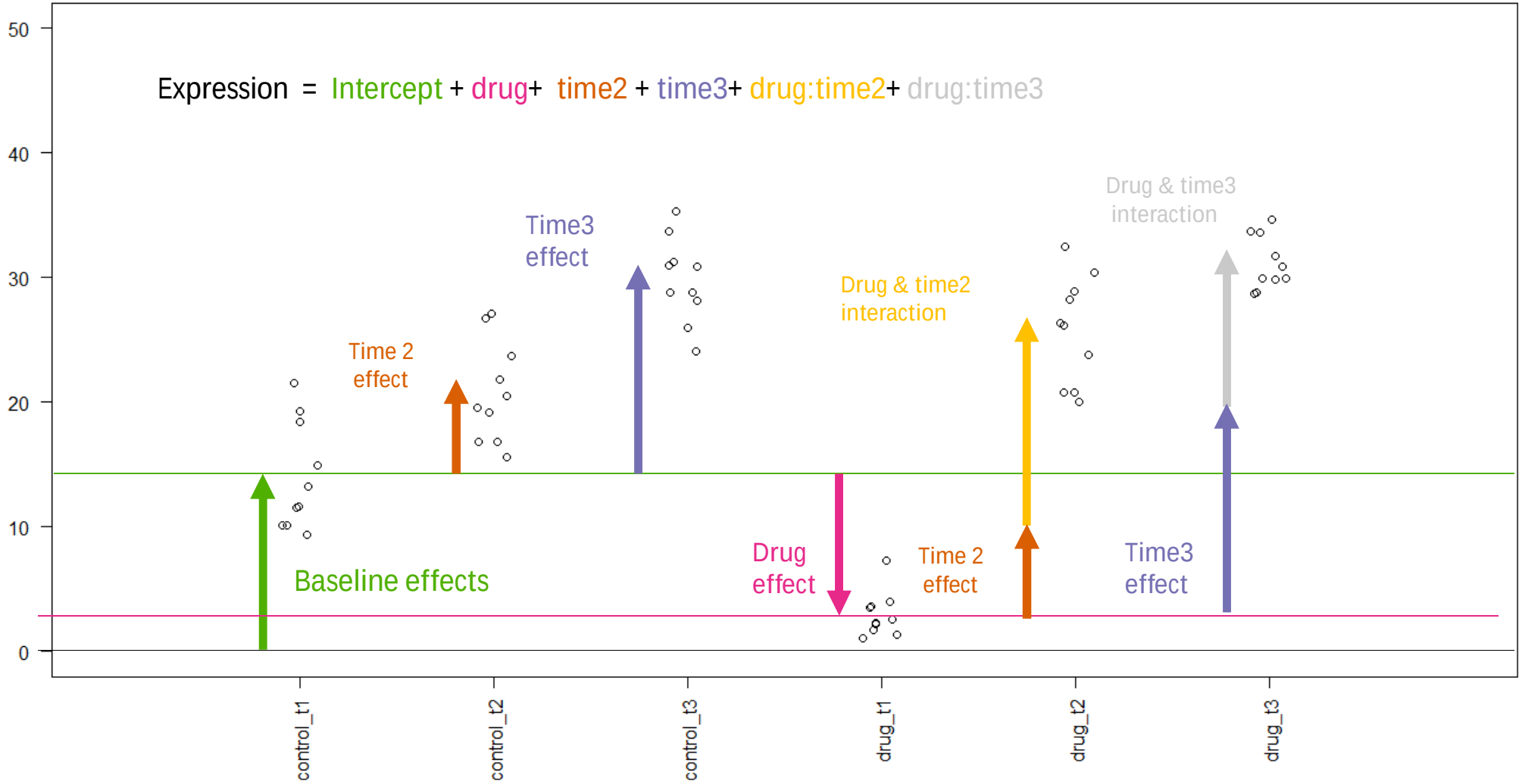

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
50
40
30
20
10
0
Time3
effect
Time 2
effect
Baseline effects
Drug
effect
Time 2
effect
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

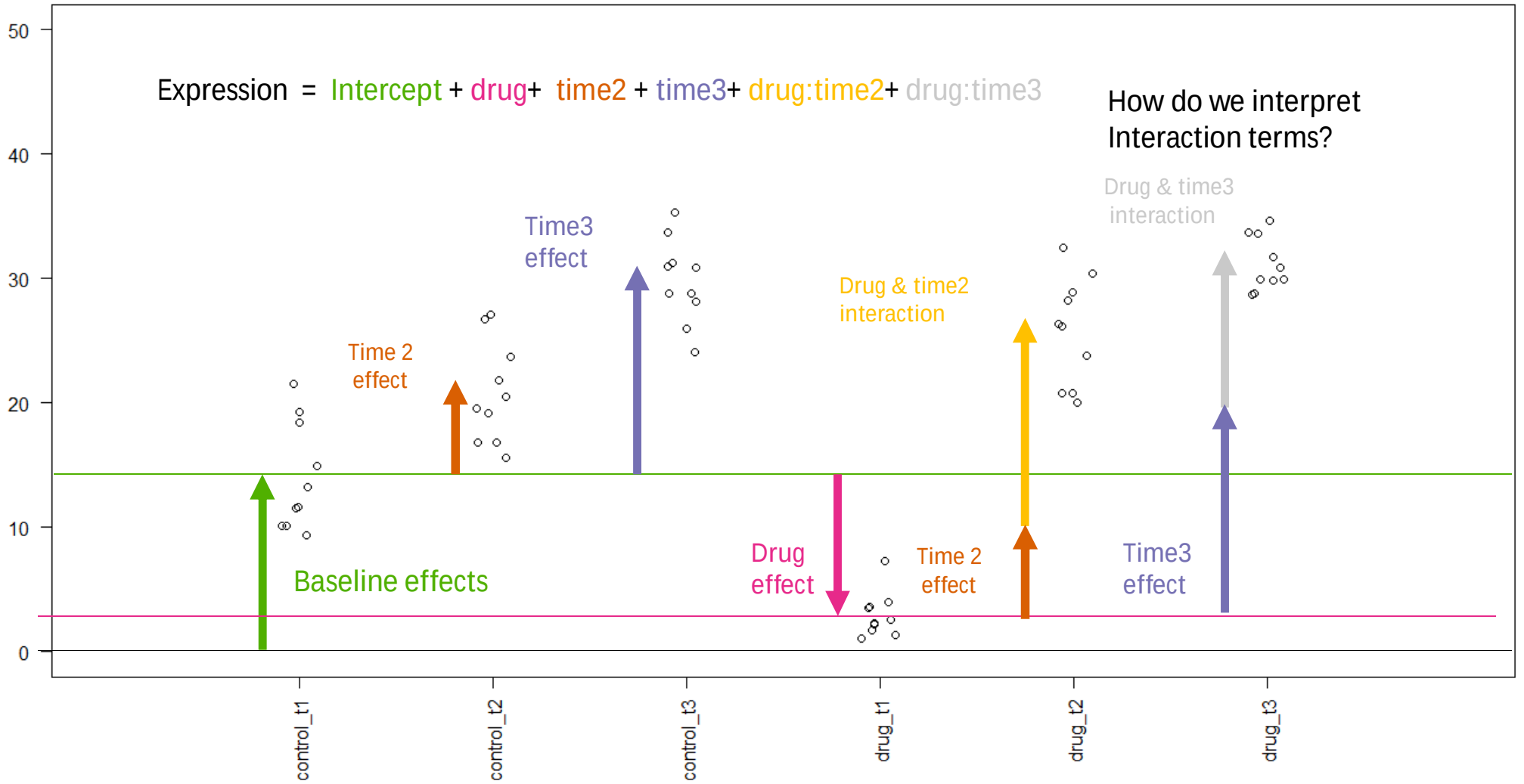

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time3 effect
Time 2 effect
Drug & time2 interaction
Baseline effects
Drug effect
Time 2 effect
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

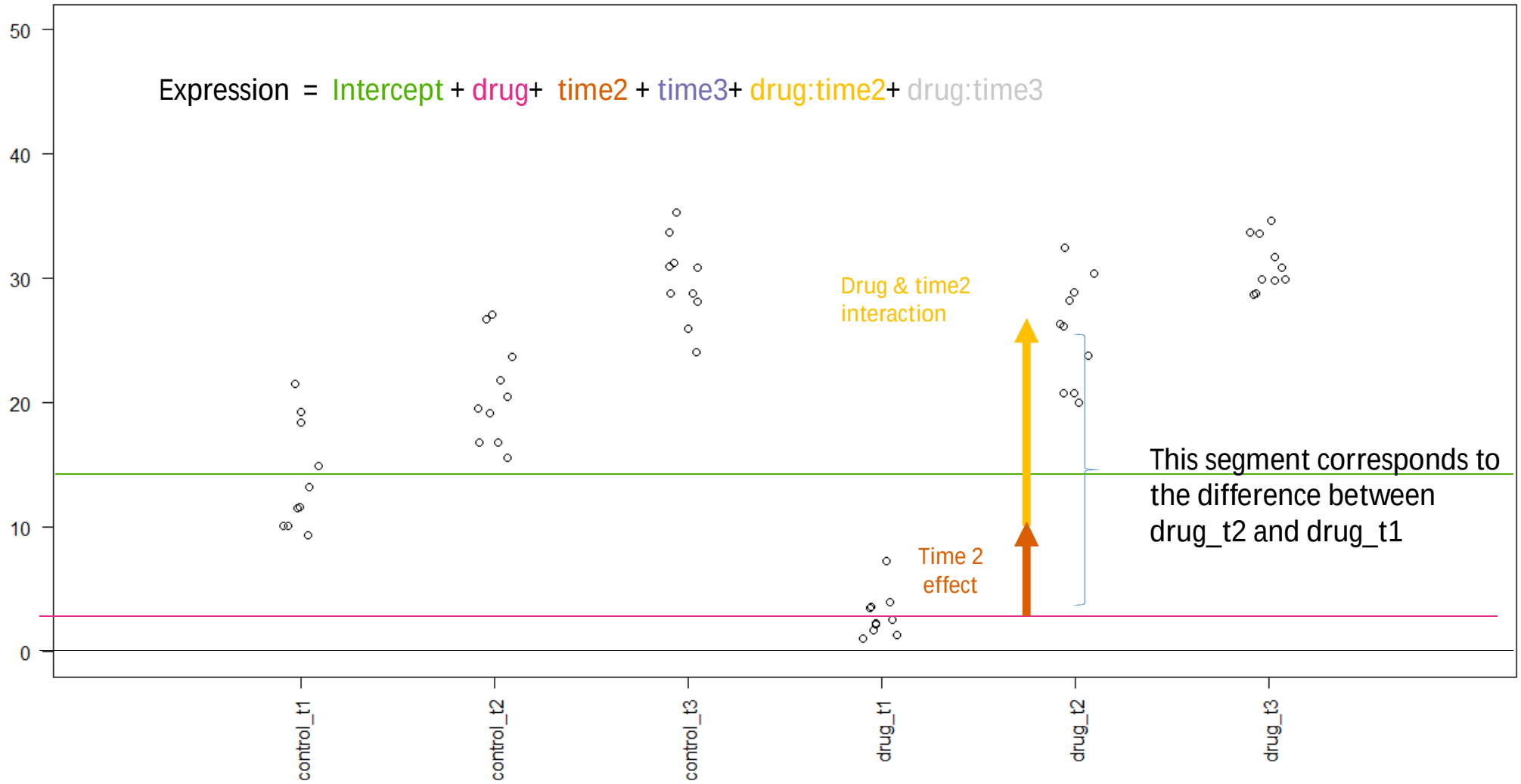

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time3 effect
Time 2 effect
Drug & time2 interaction
Baseline effects
Drug effect
Time 2 effect
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

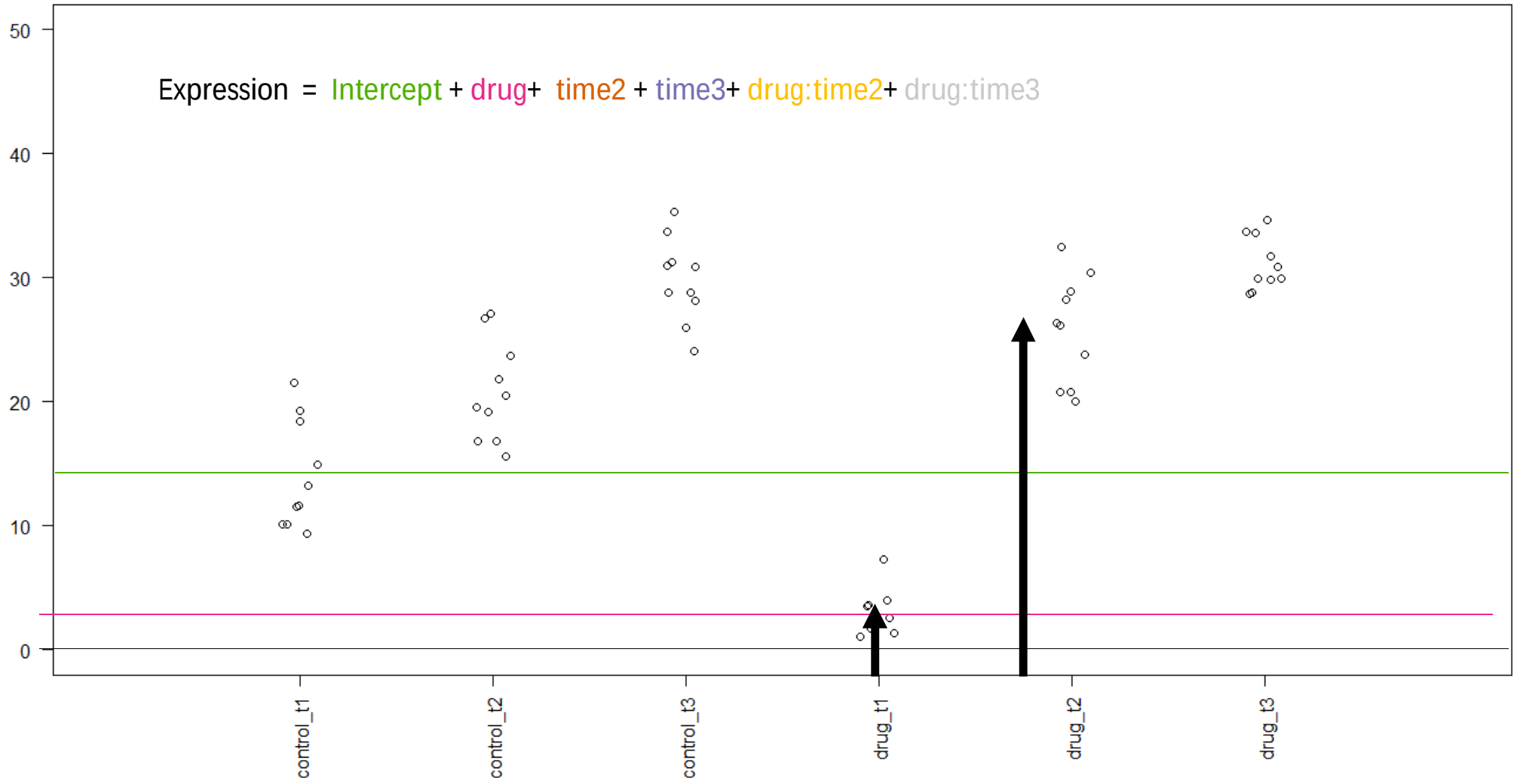

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
50
40
30
20
10
0
Time3
effect
Time 2
effect
Drug & time2
interaction
Baseline effects
Drug
effect
Time 2
effect
Time3
effect
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

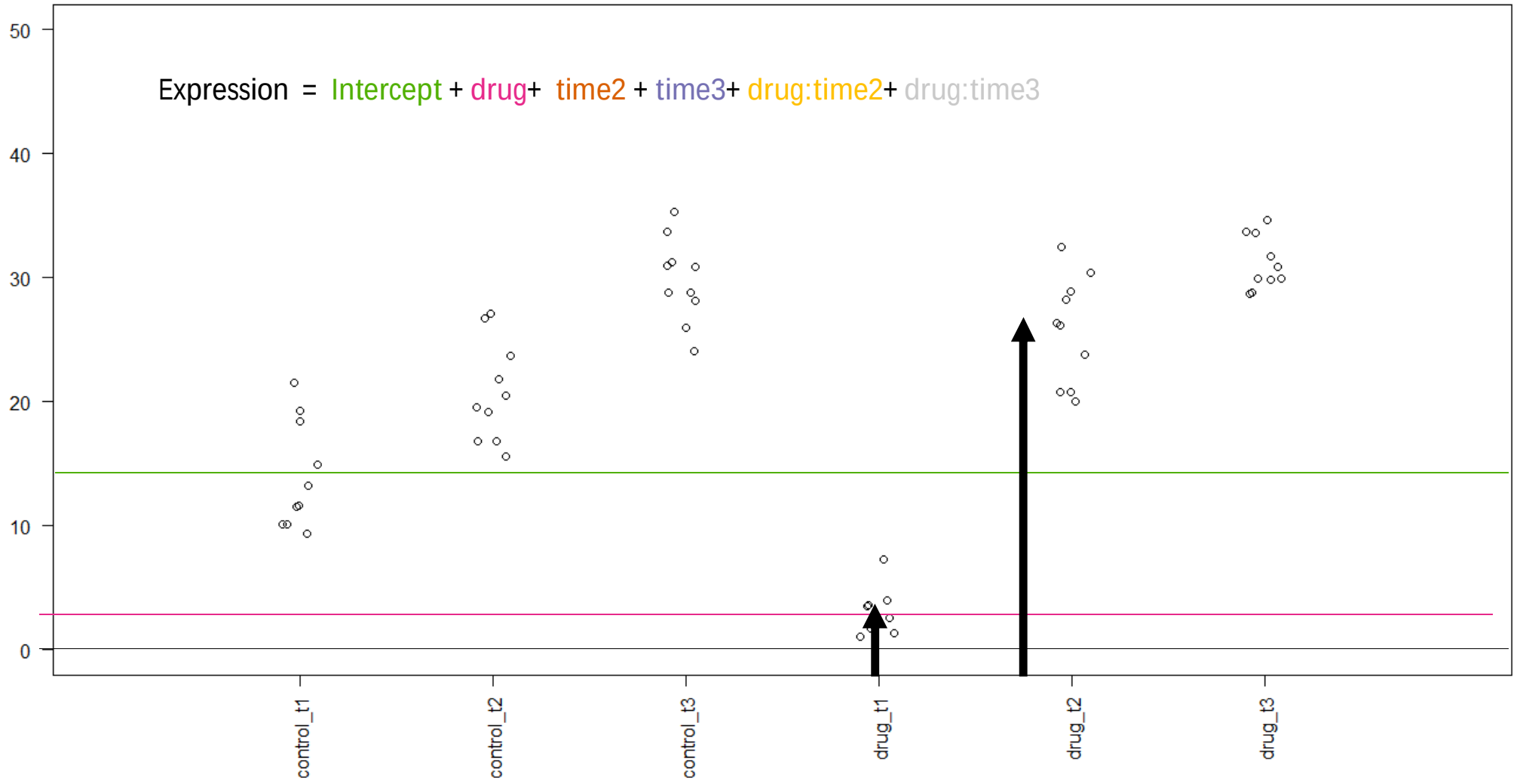

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
50
40
30
20
10
0
Time3
effect
Time 2
effect
Drug & time2
interaction
Baseline effects
Drug
effect
Time 2
effect
Time3
effect
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

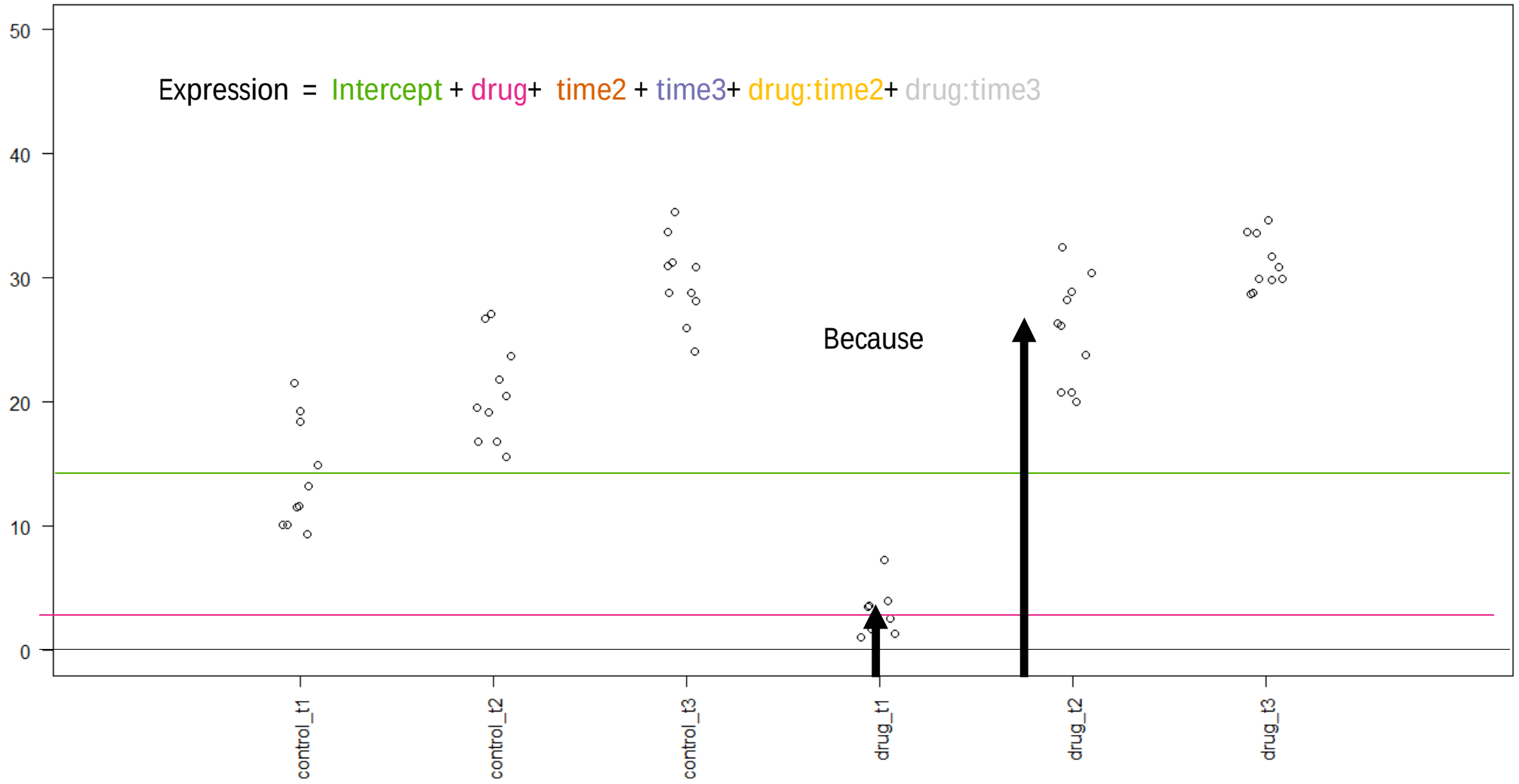

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
50
40
30
20
10
0
Baseline effects
Time 2 effect
Time3 effect
Drug effect
Drug & time2 interaction
Time 2 effect
Drug & time3 interaction
Time3 effect
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

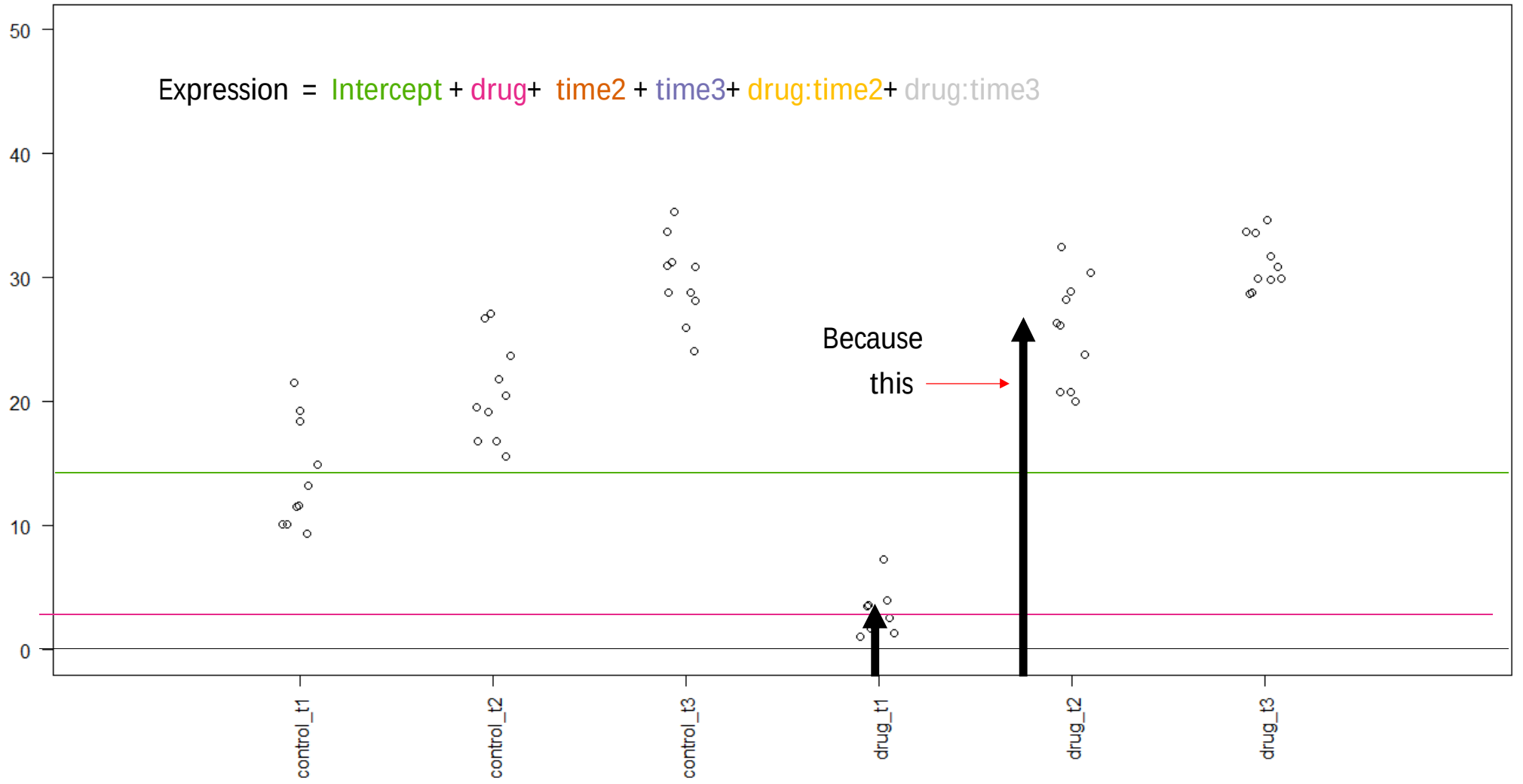

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
How do we interpret Interaction terms?
Drug & time3 interaction
Time3 effect
Drug & time2 interaction
Time 2 effect
Baseline effects
Drug effect
Time 2 effect
Time3 effect
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

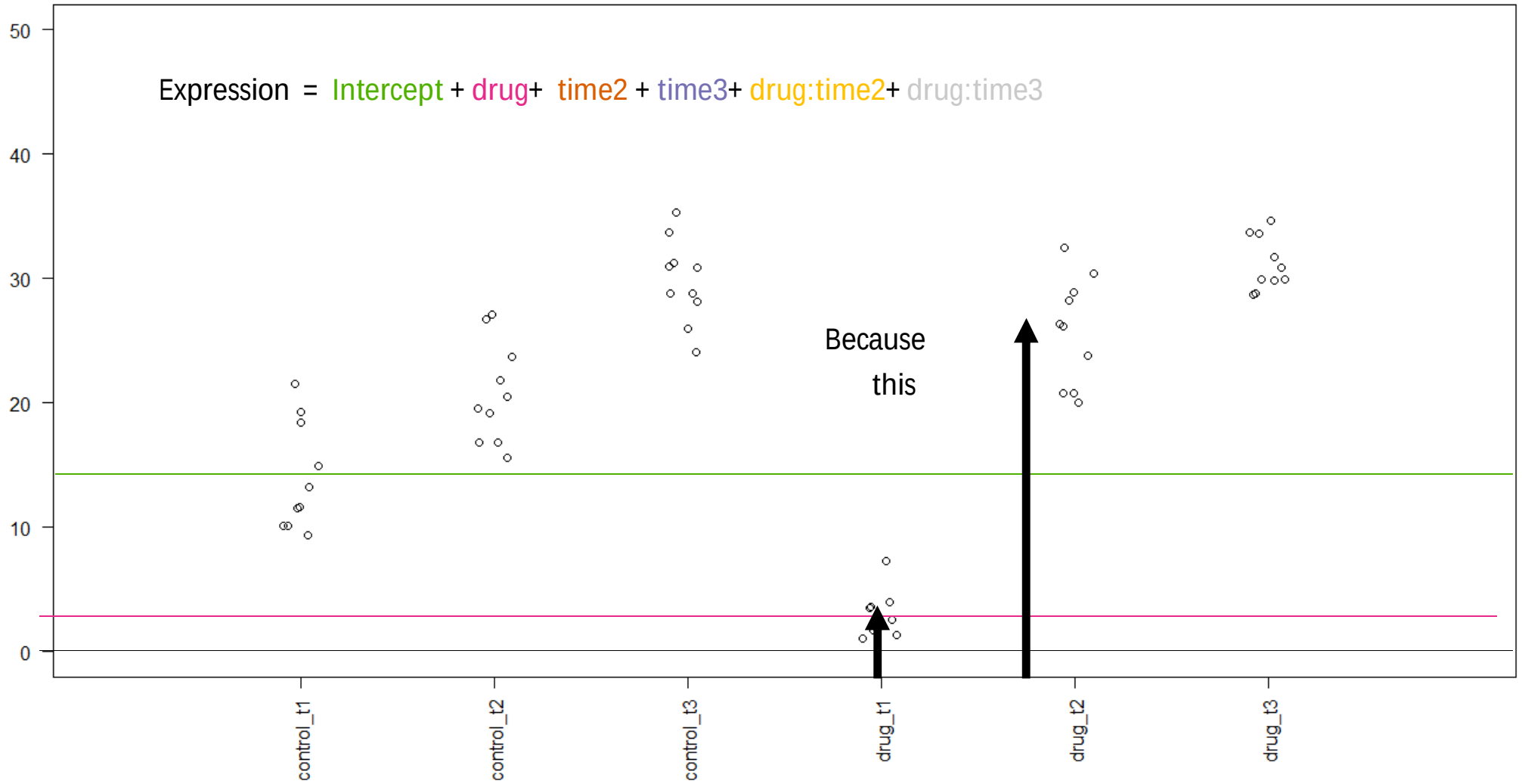

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Drug & time2 interaction
Time 2 effect
This segment corresponds to the difference between drug_t2 and drug_t1
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

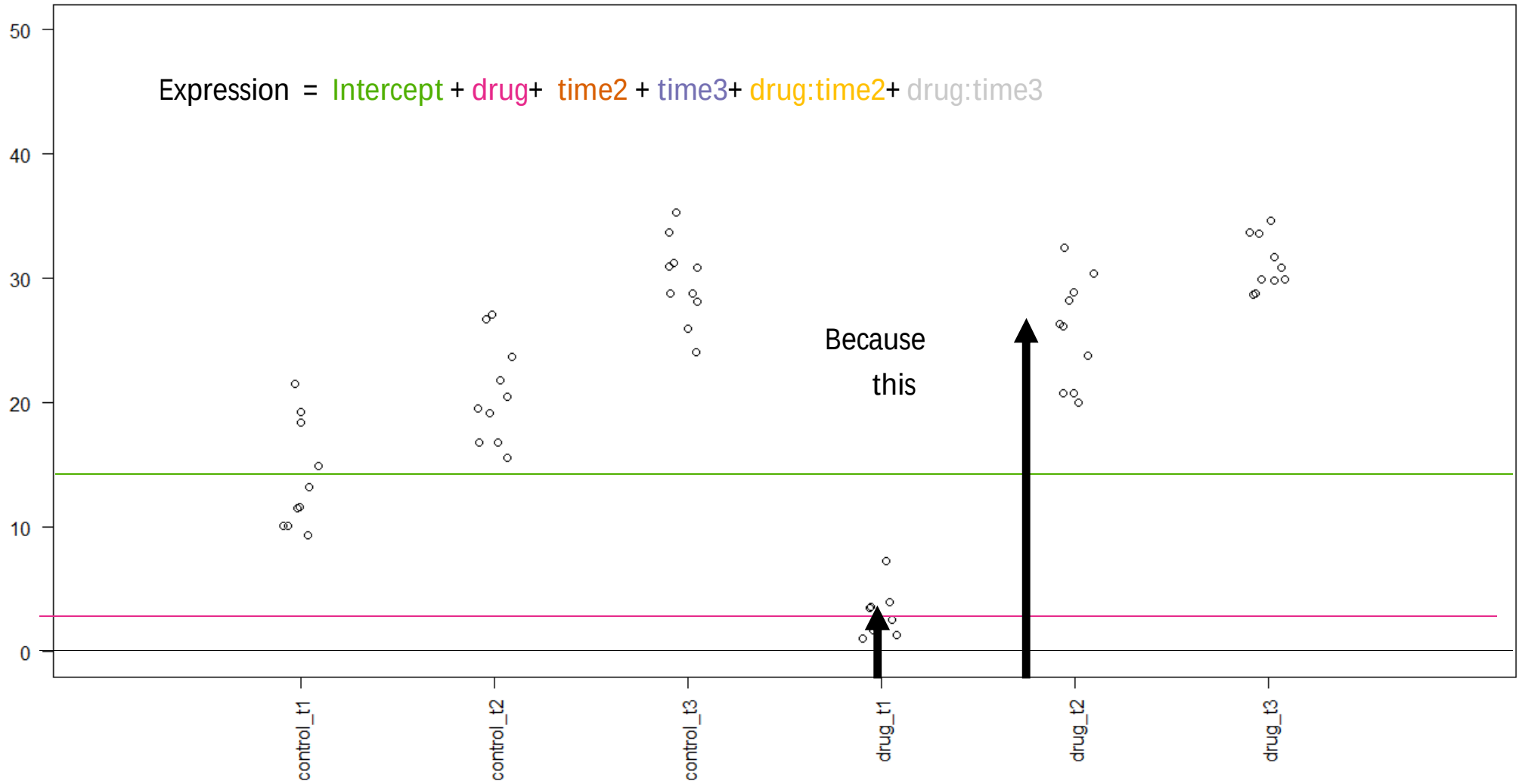

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

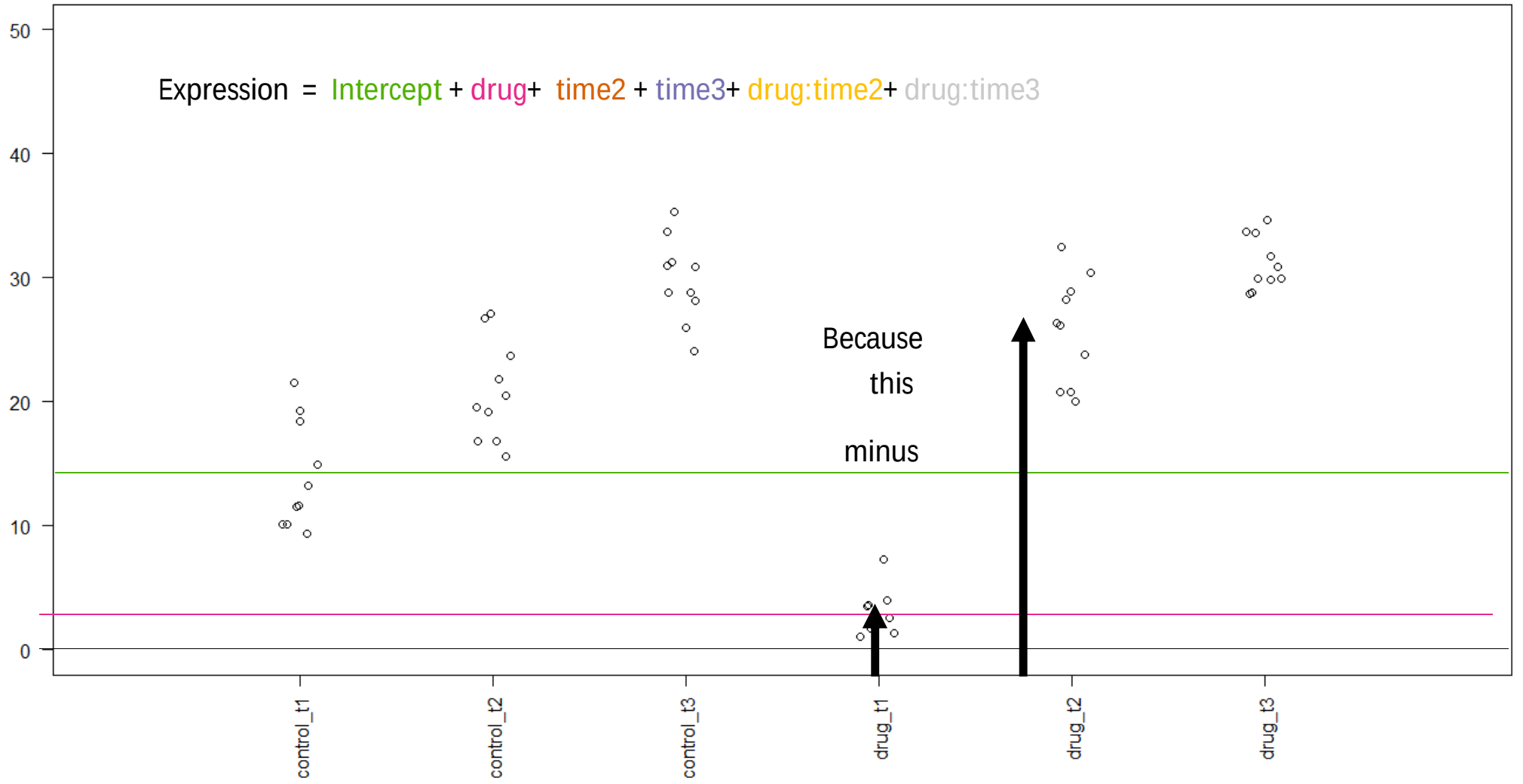

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

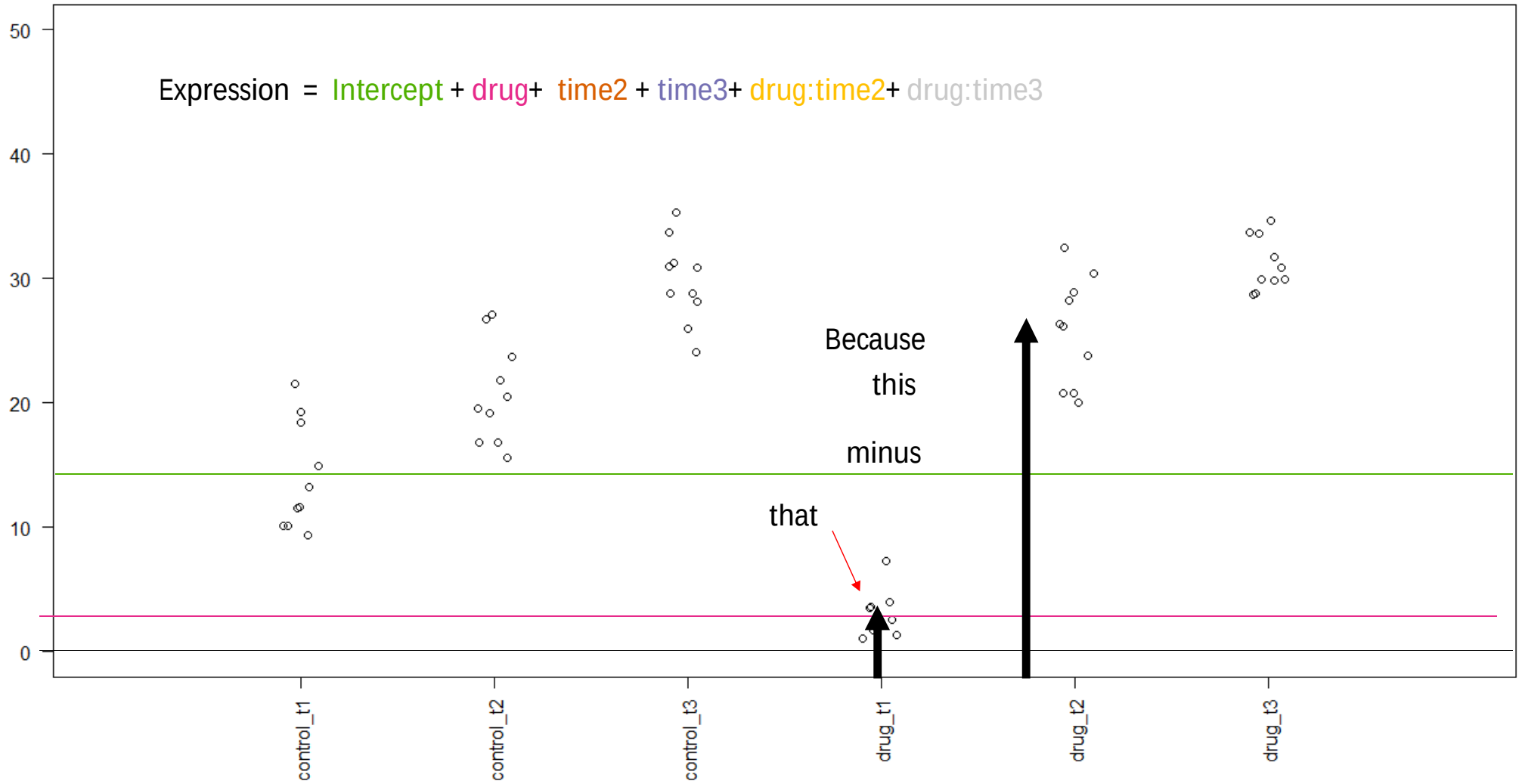

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Because
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

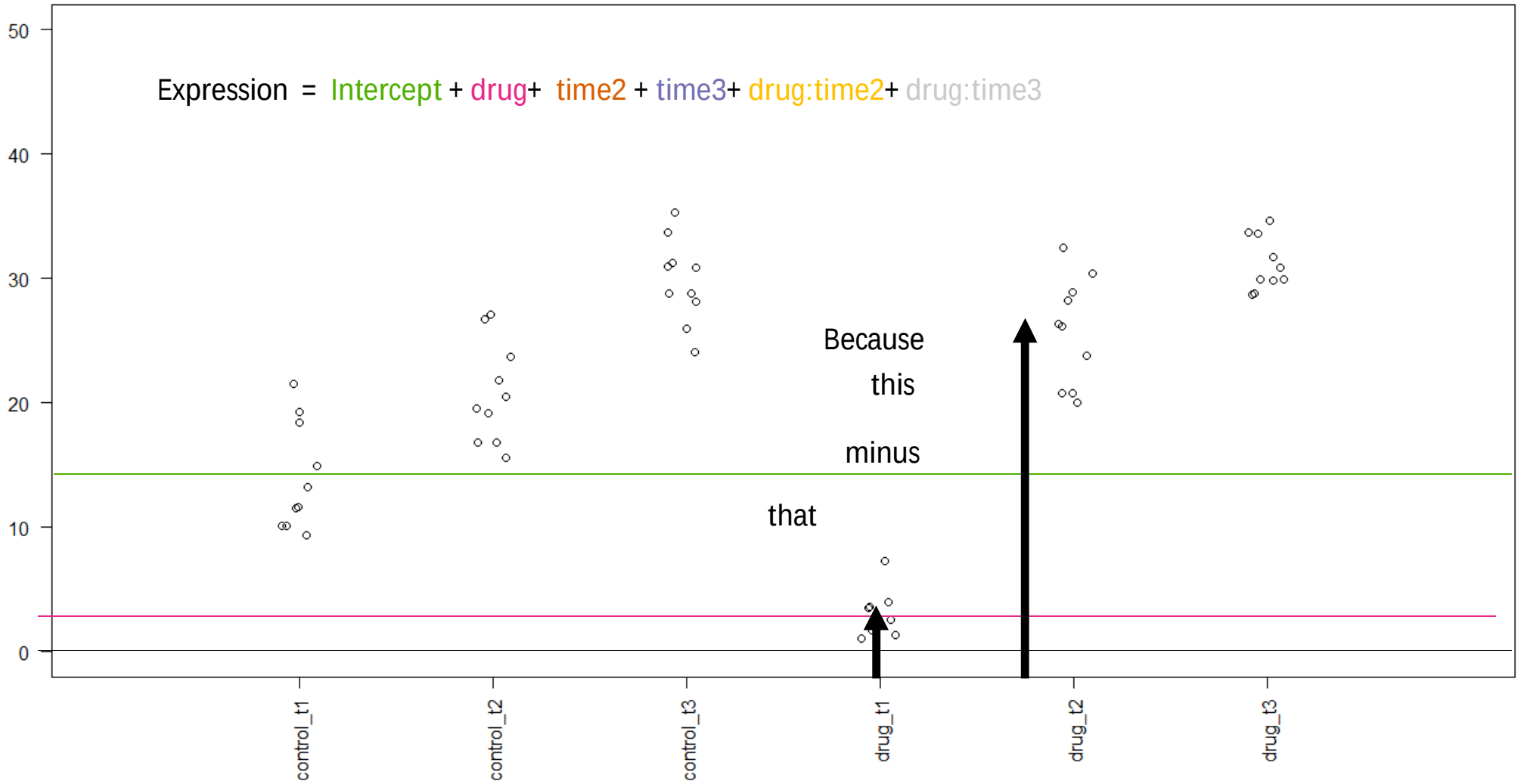

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Because this
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

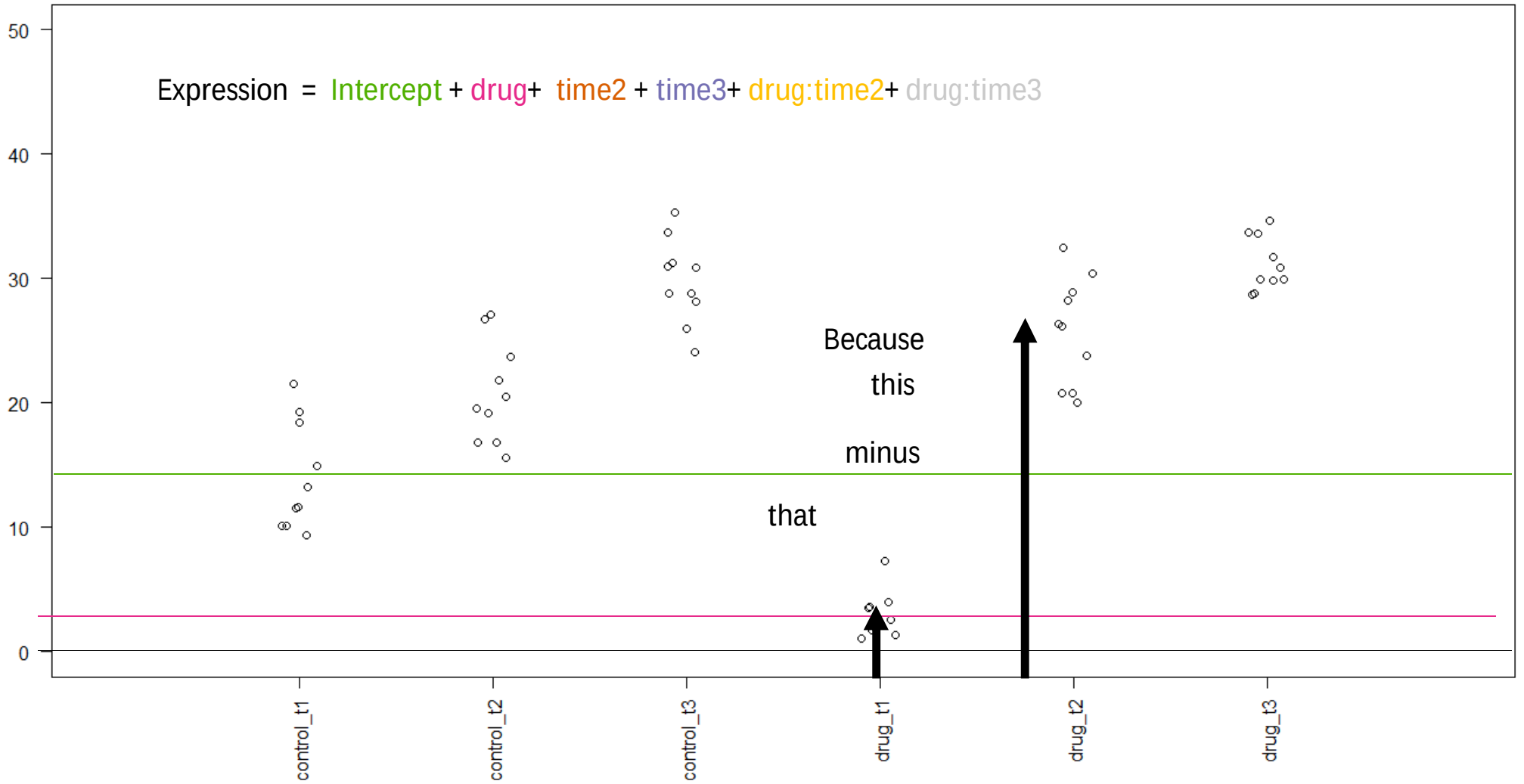

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Because this
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

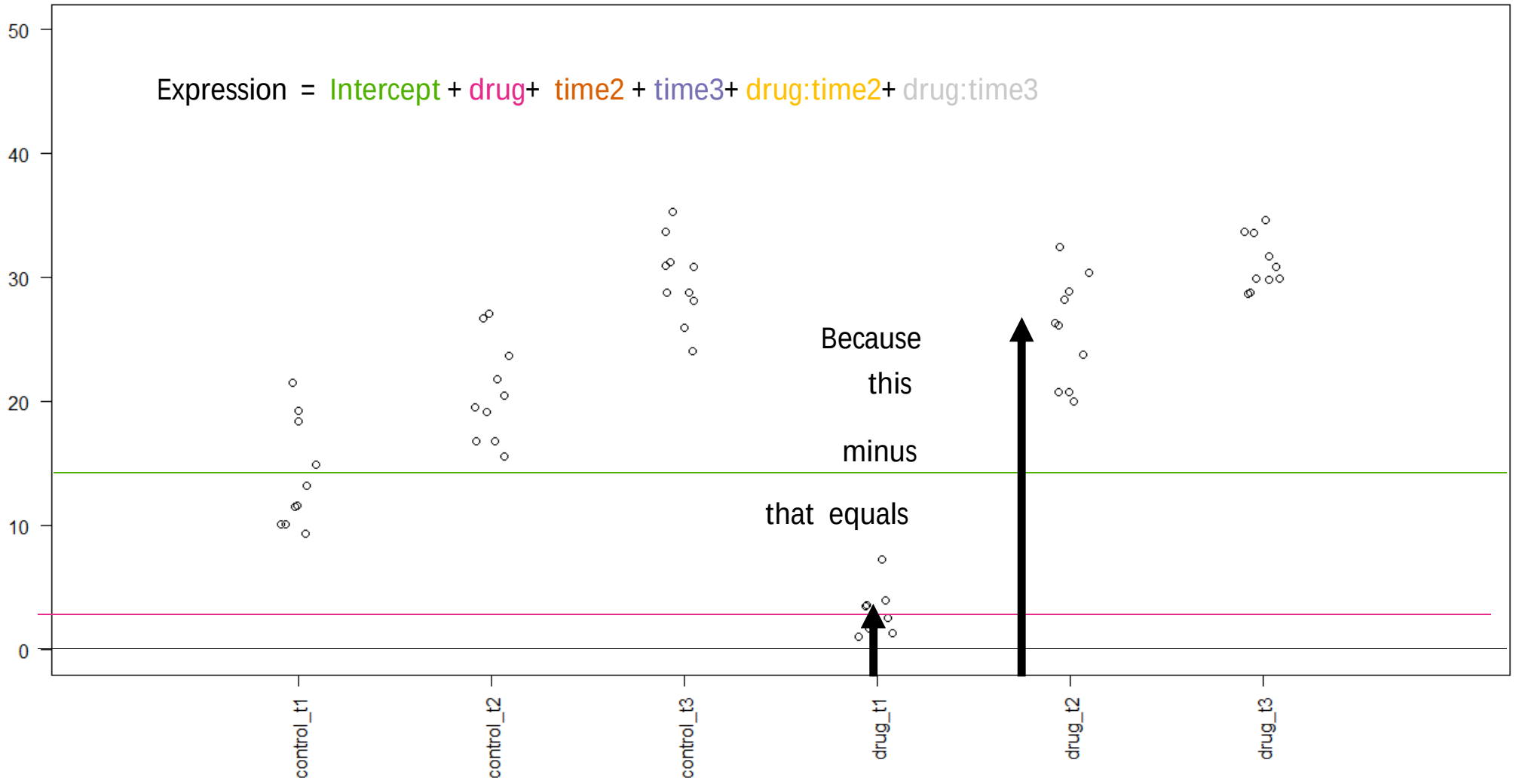

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Because this
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

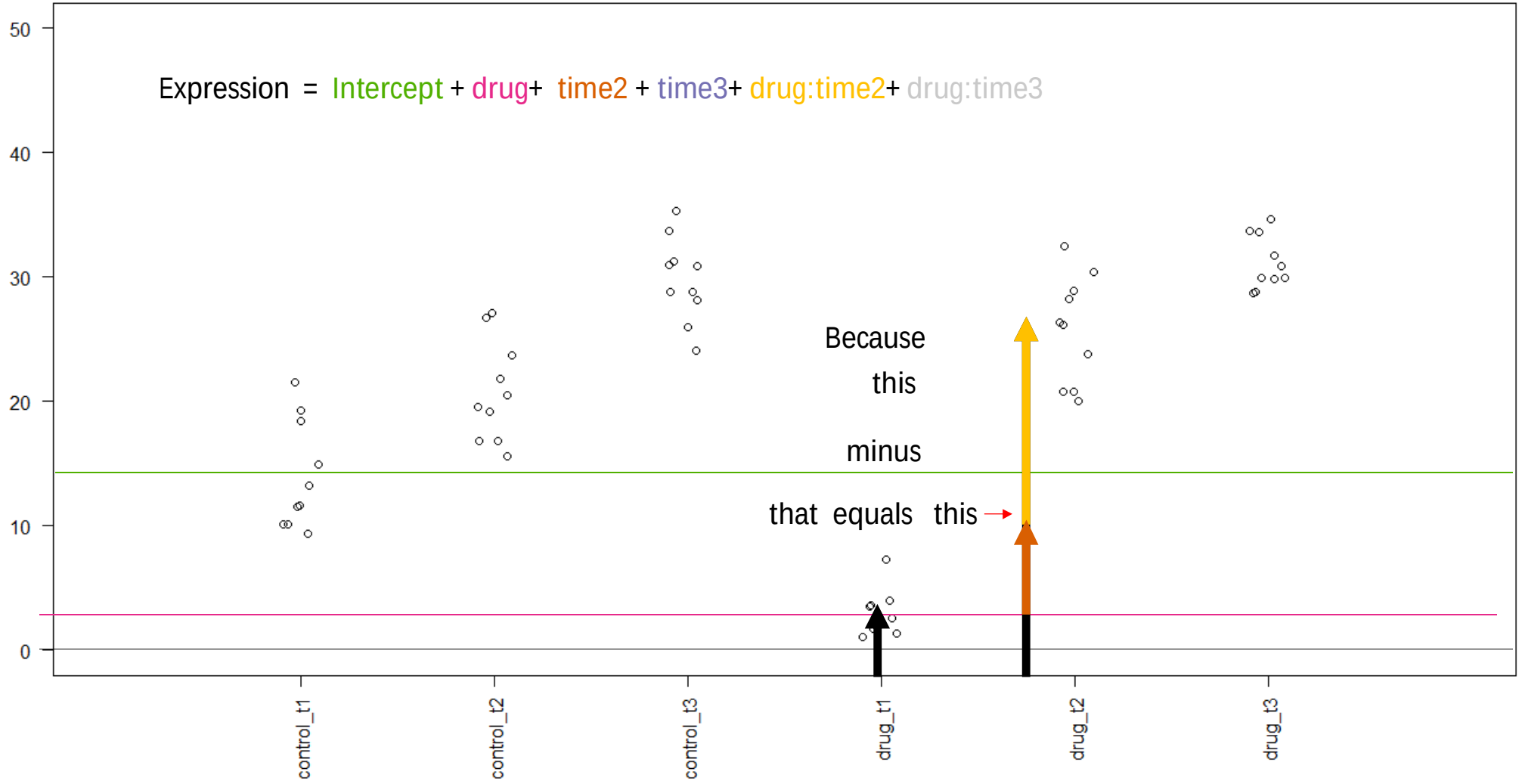

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Because this minus
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

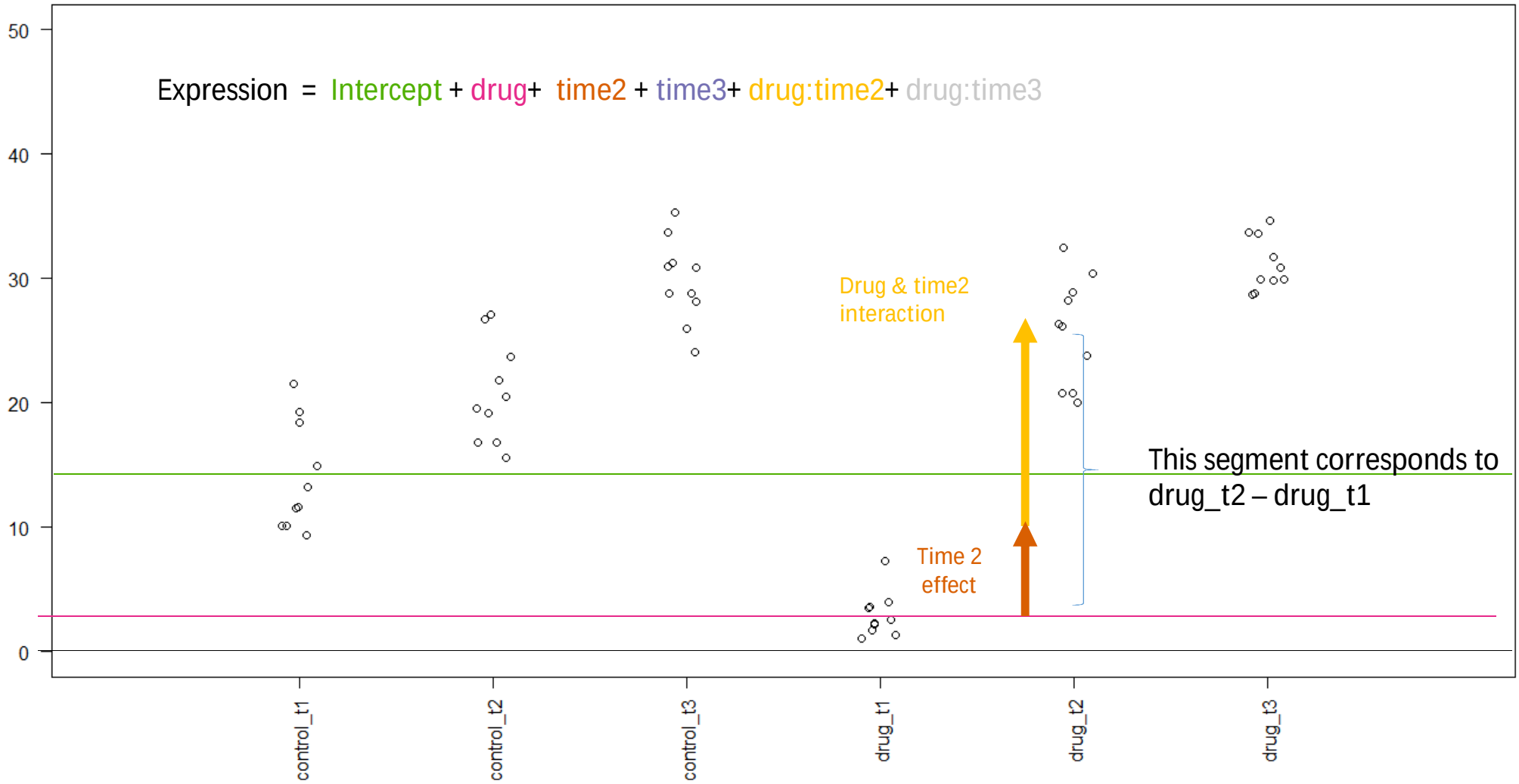

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Because this minus that
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

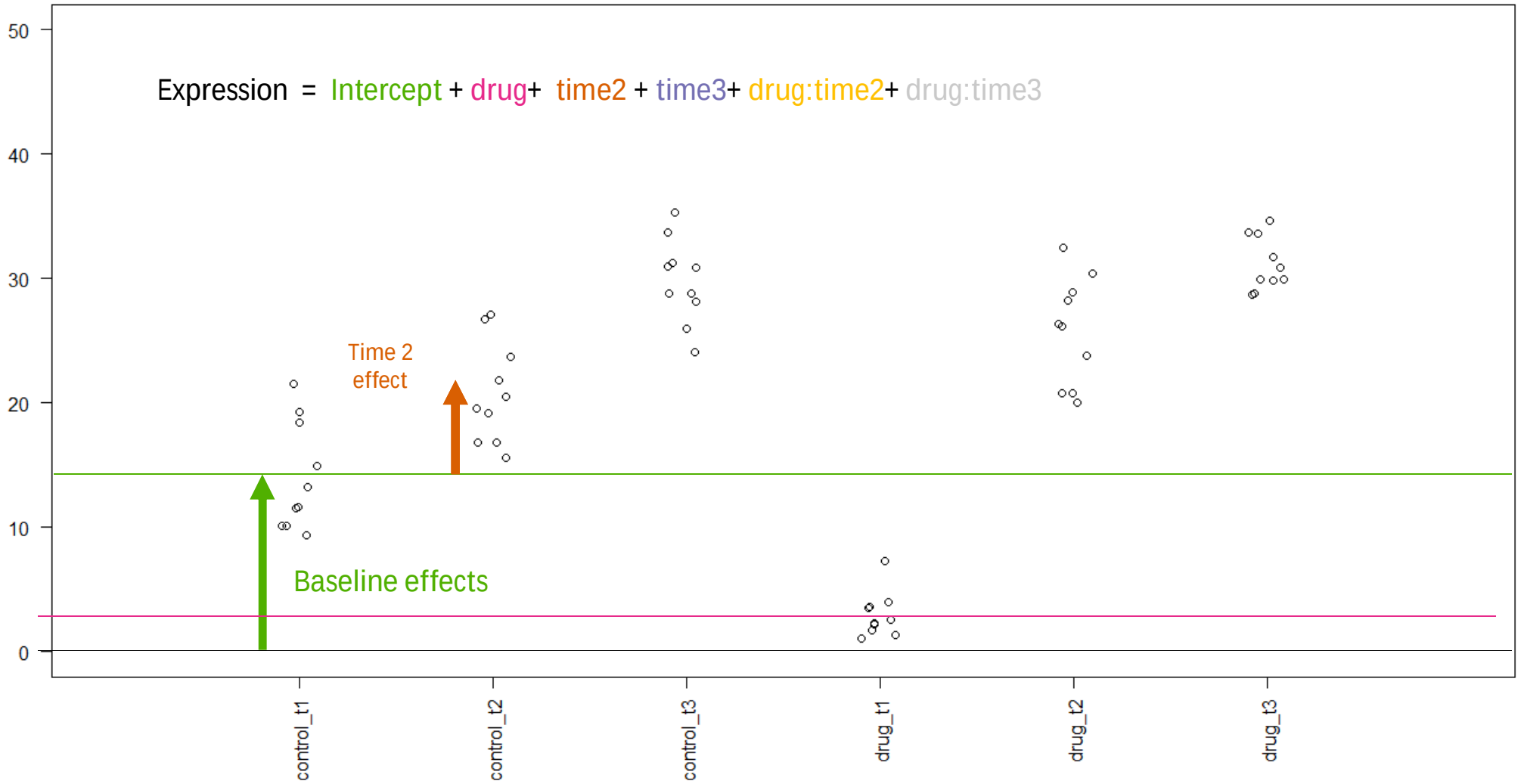

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Because this minus that
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

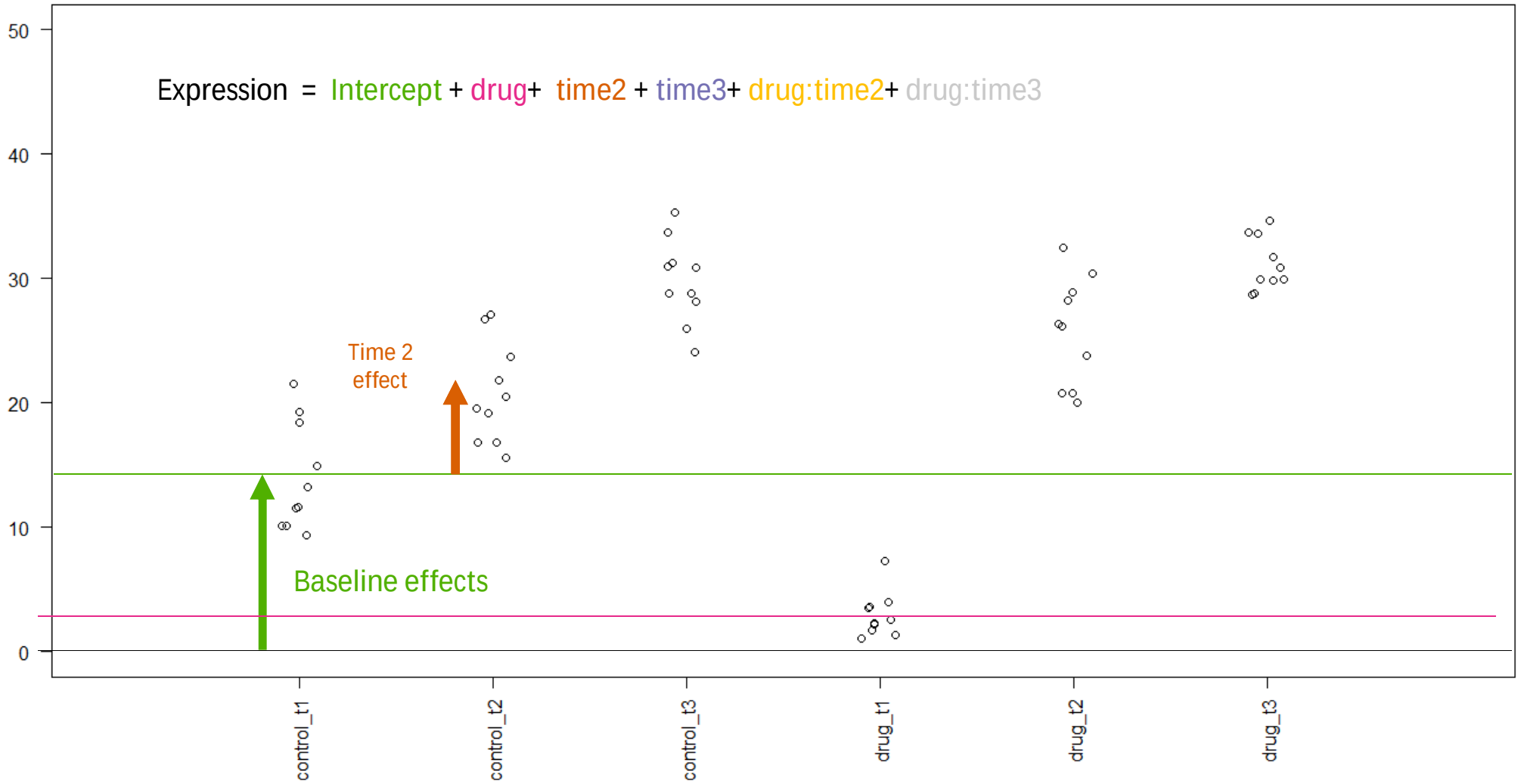

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Because this minus that
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

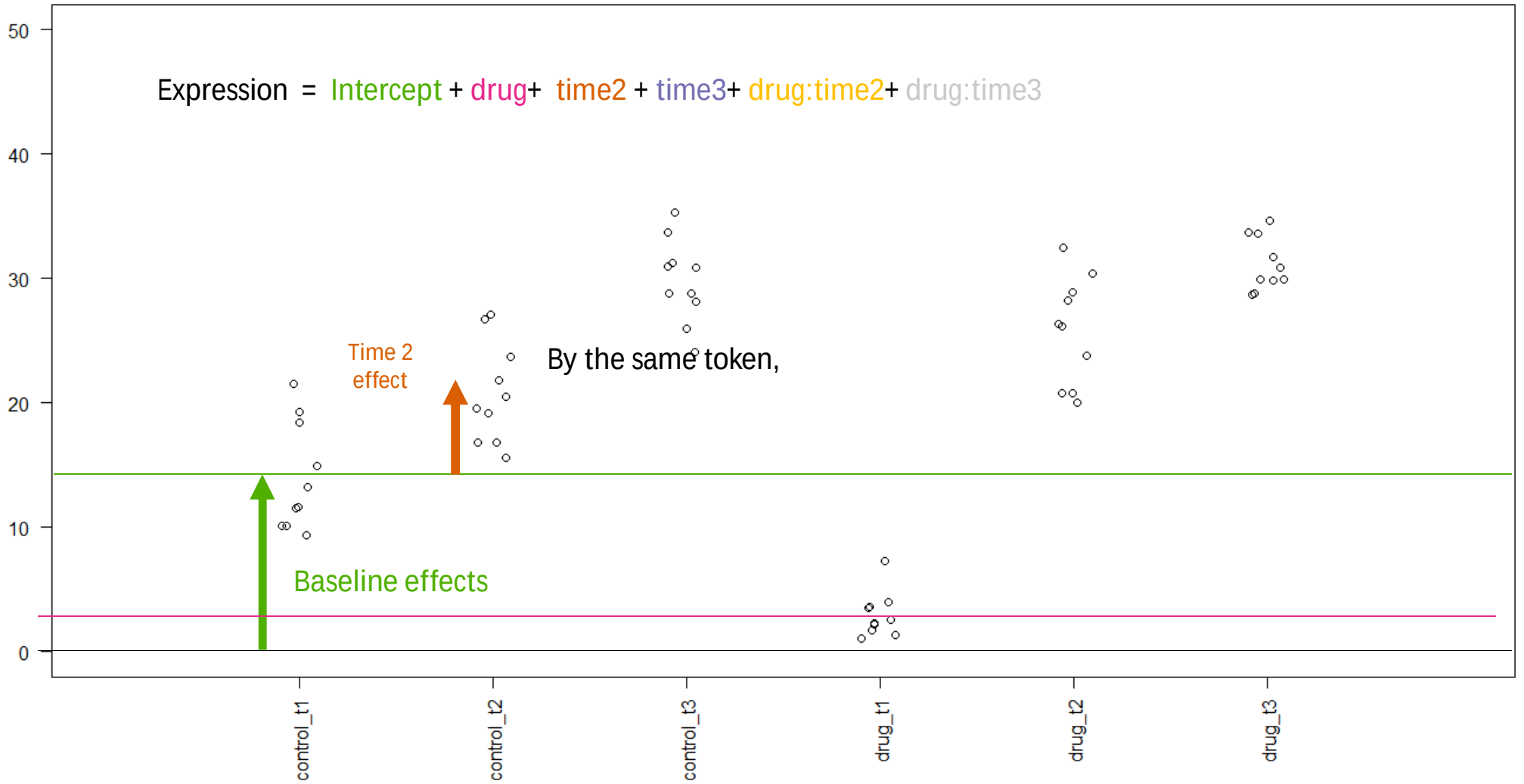

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Because this minus that equals
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

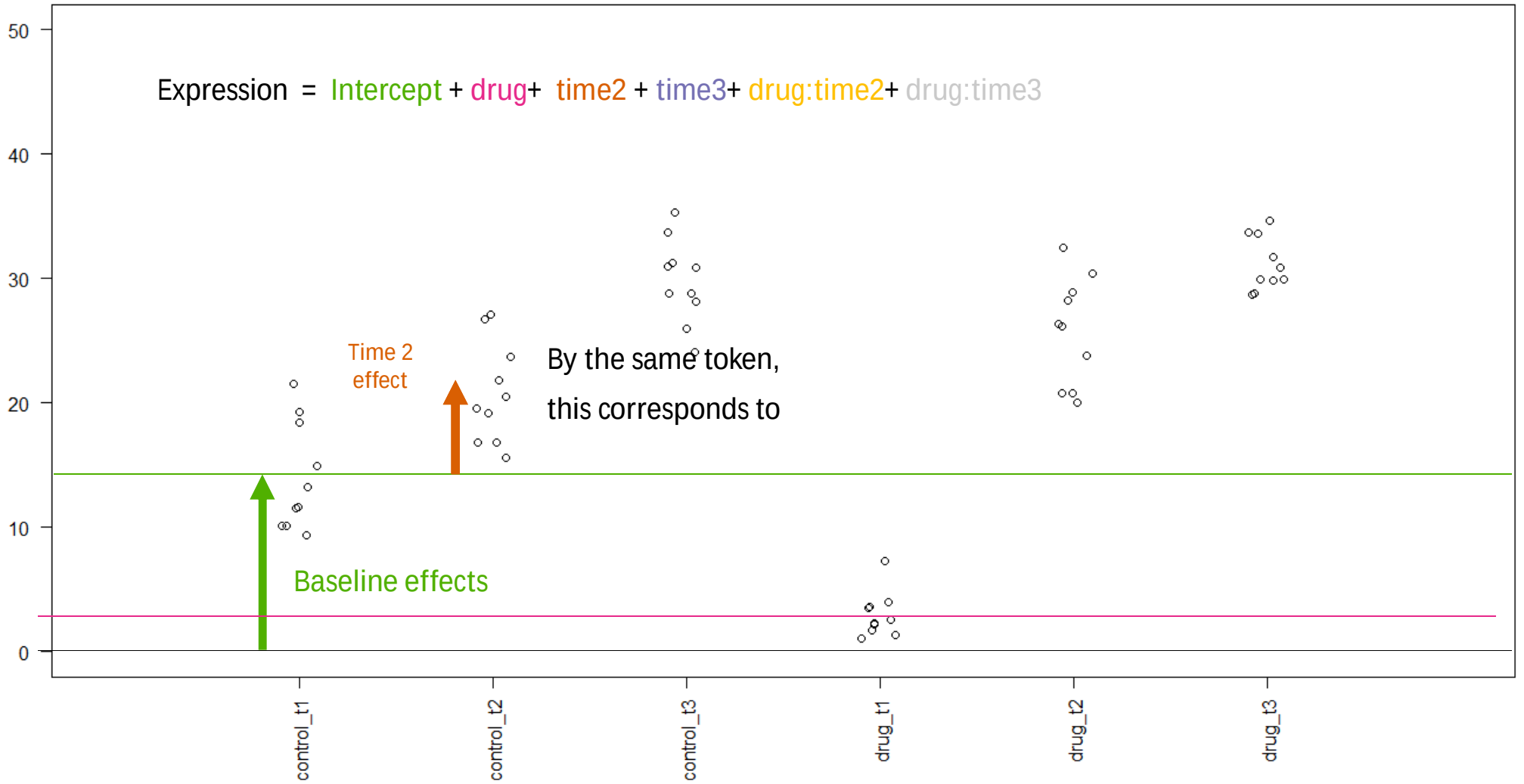

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Because this minus that equals this
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

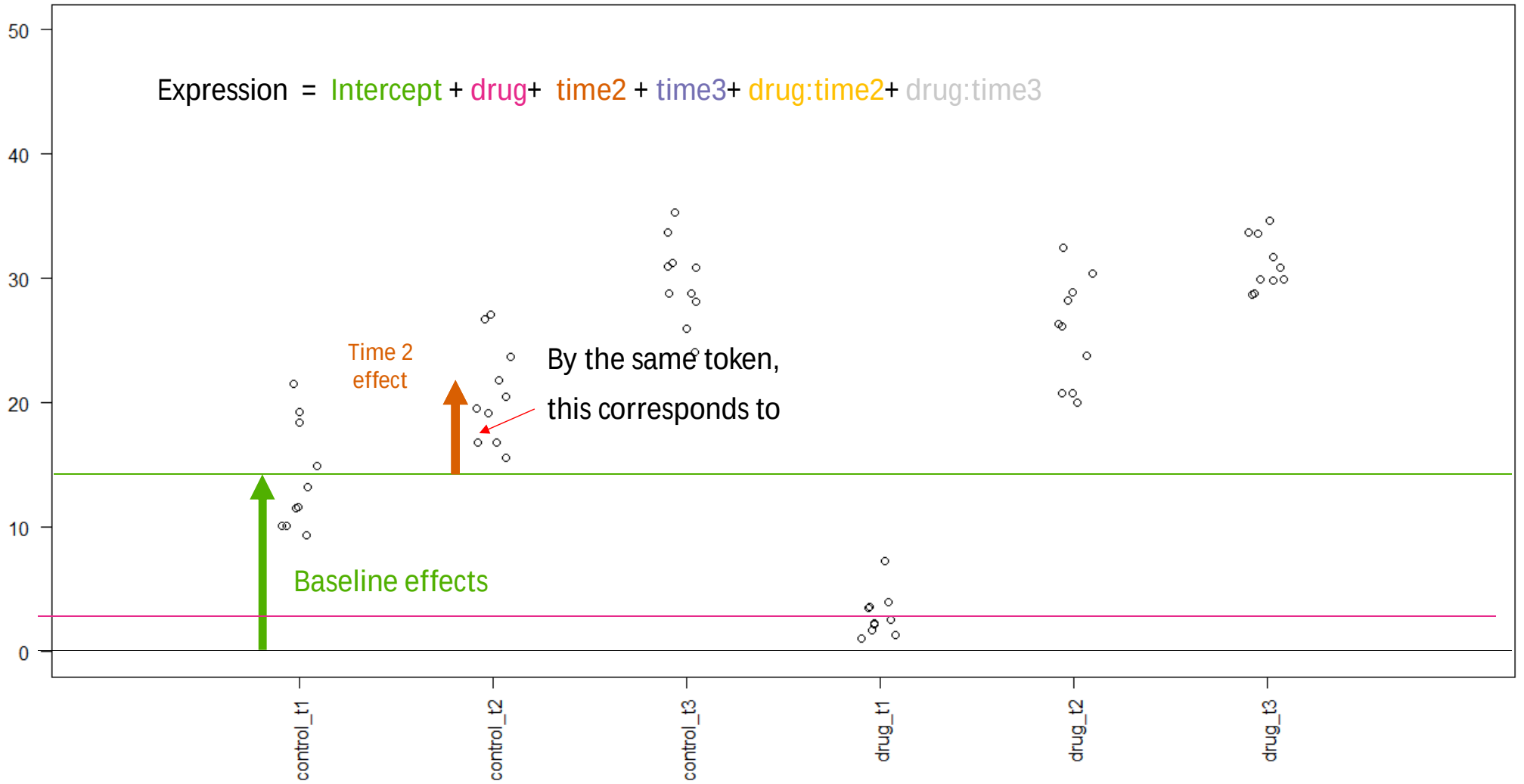

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Drug & time2 interaction
Time 2 effect
This segment corresponds to drug_t2 – drug_t1
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

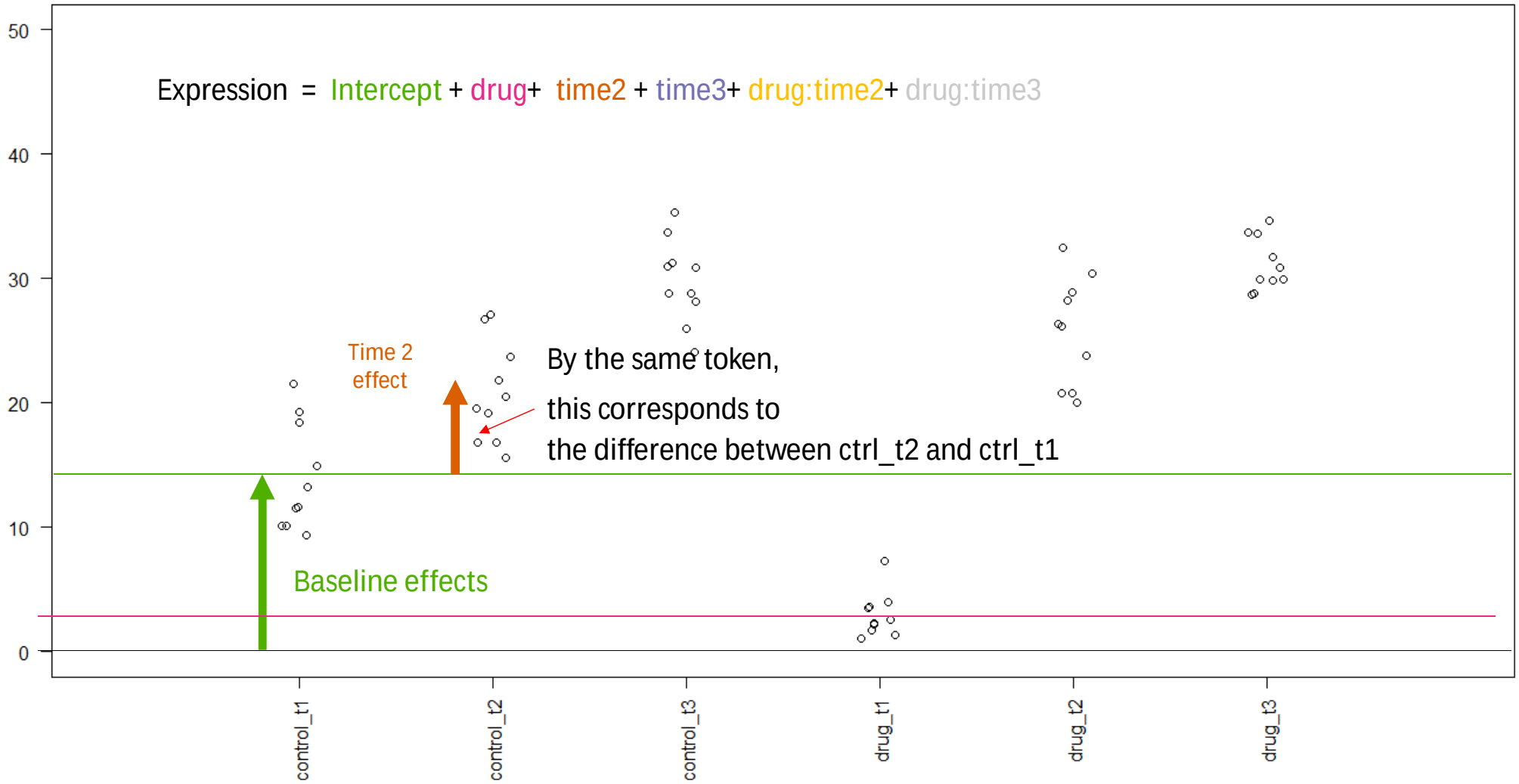

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time 2 effect
Baseline effects
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

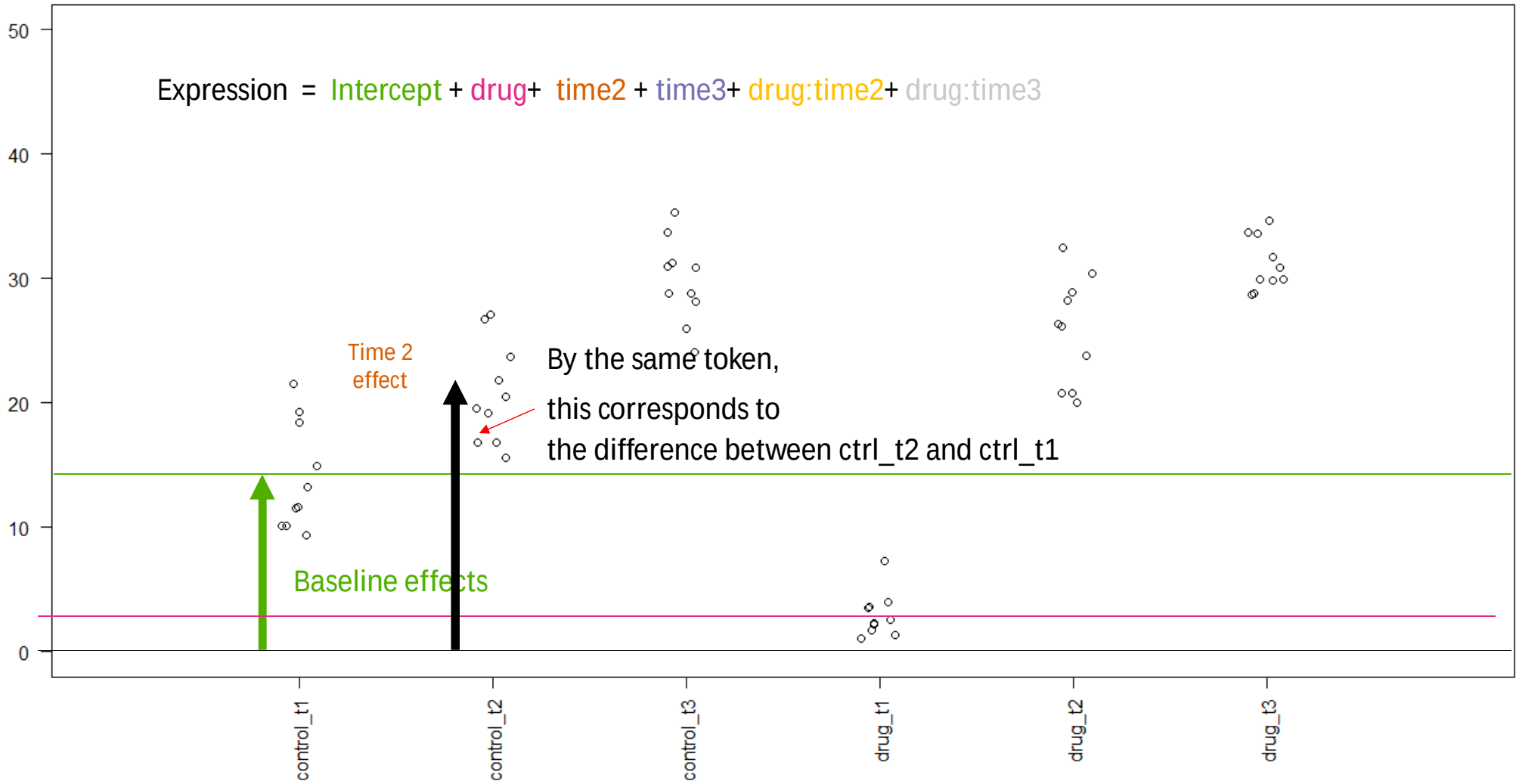

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time 2 effect
Baseline effects
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

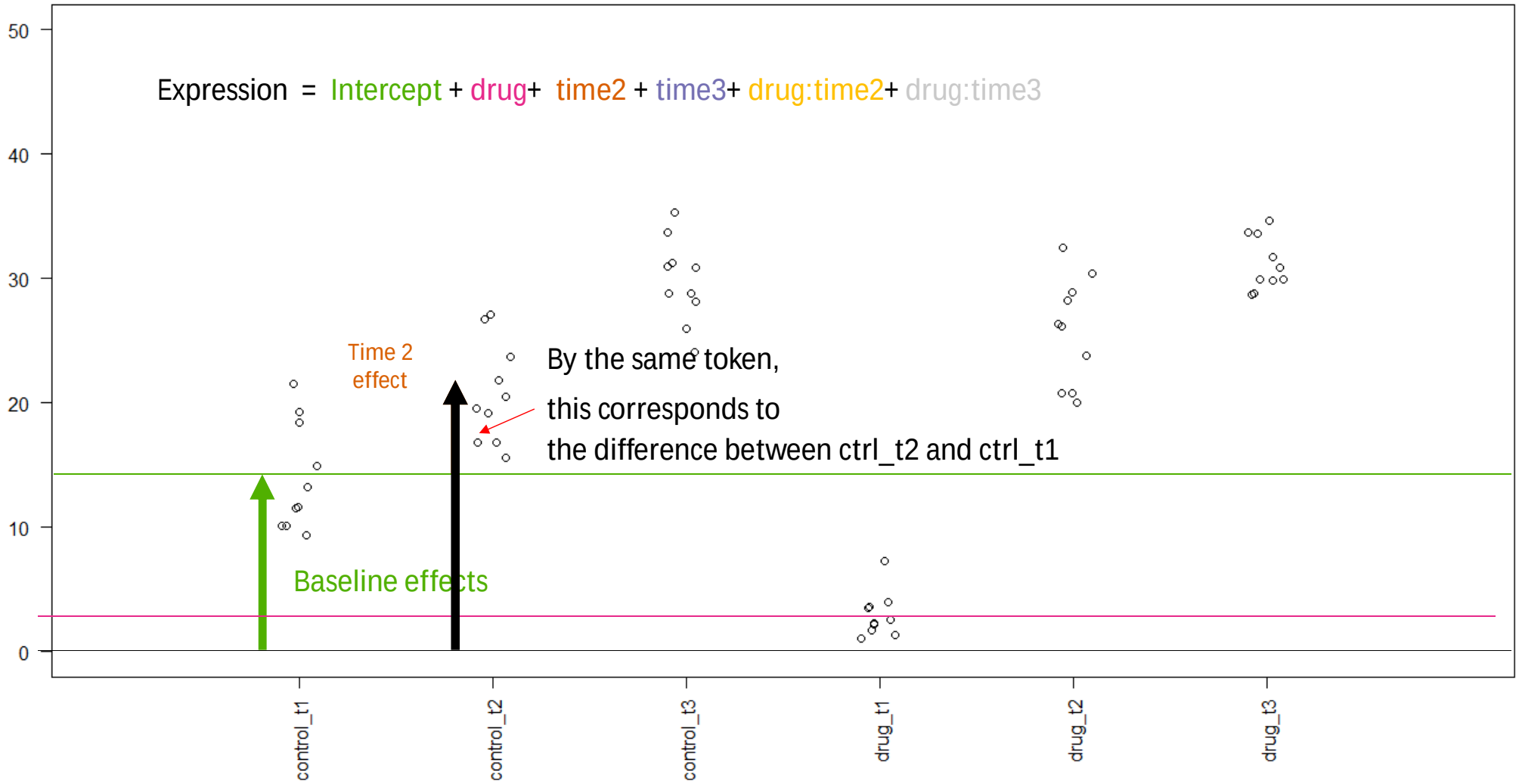

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time 2 effect
By the same token,
Baseline effects
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

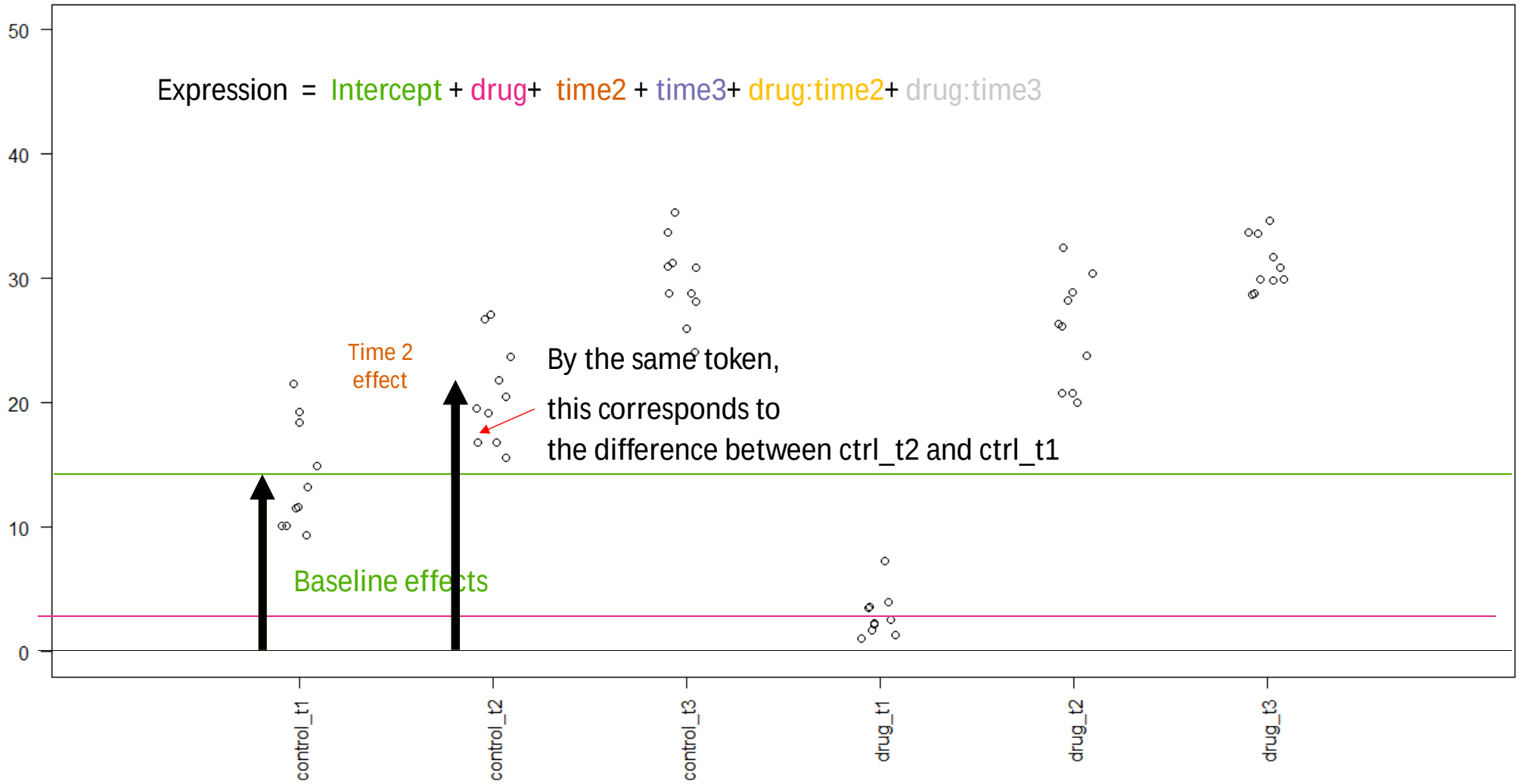

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time 2 effect
By the same token,
this corresponds to
Baseline effects
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

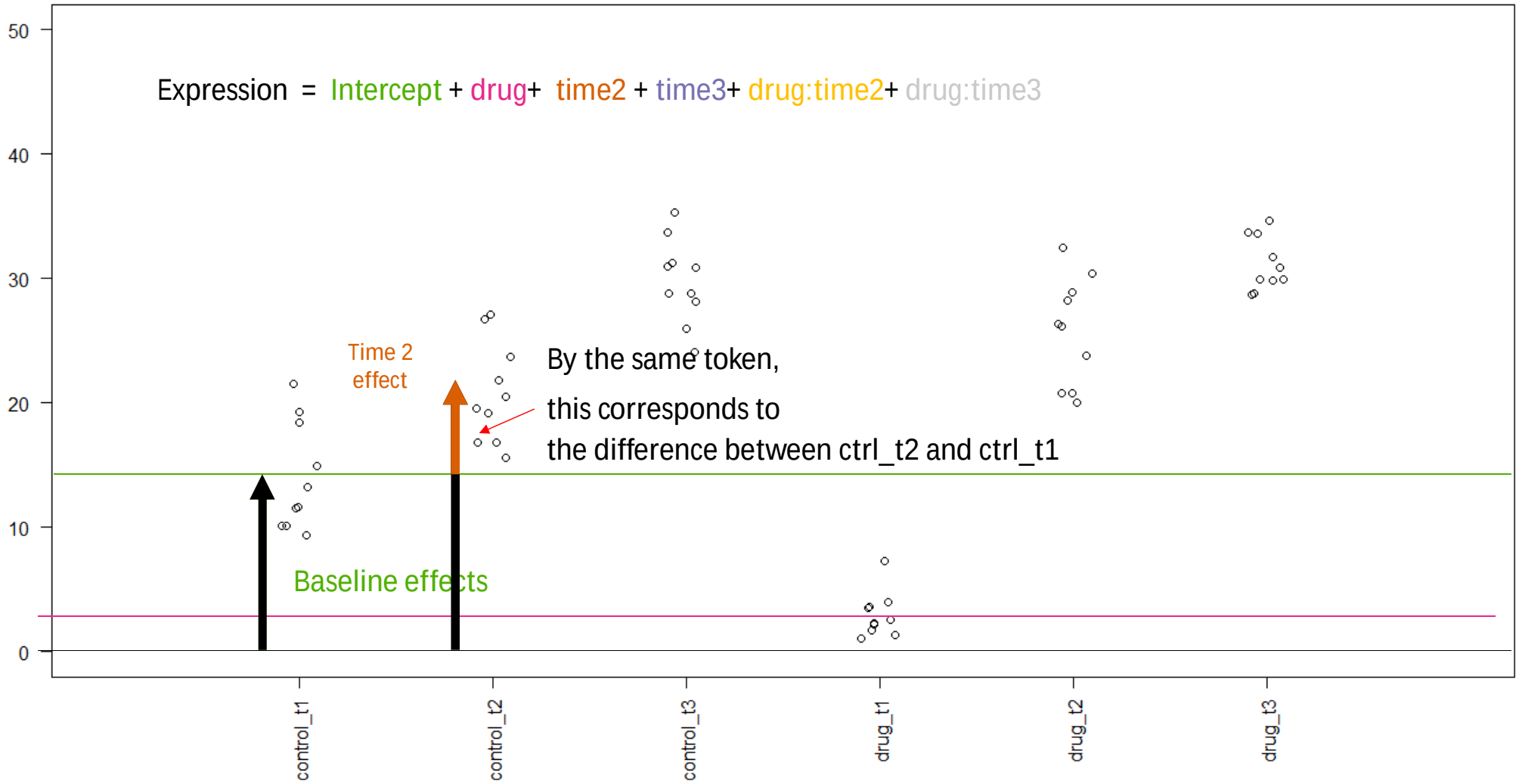

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time 2 effect
By the same token,
this corresponds to
Baseline effects
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

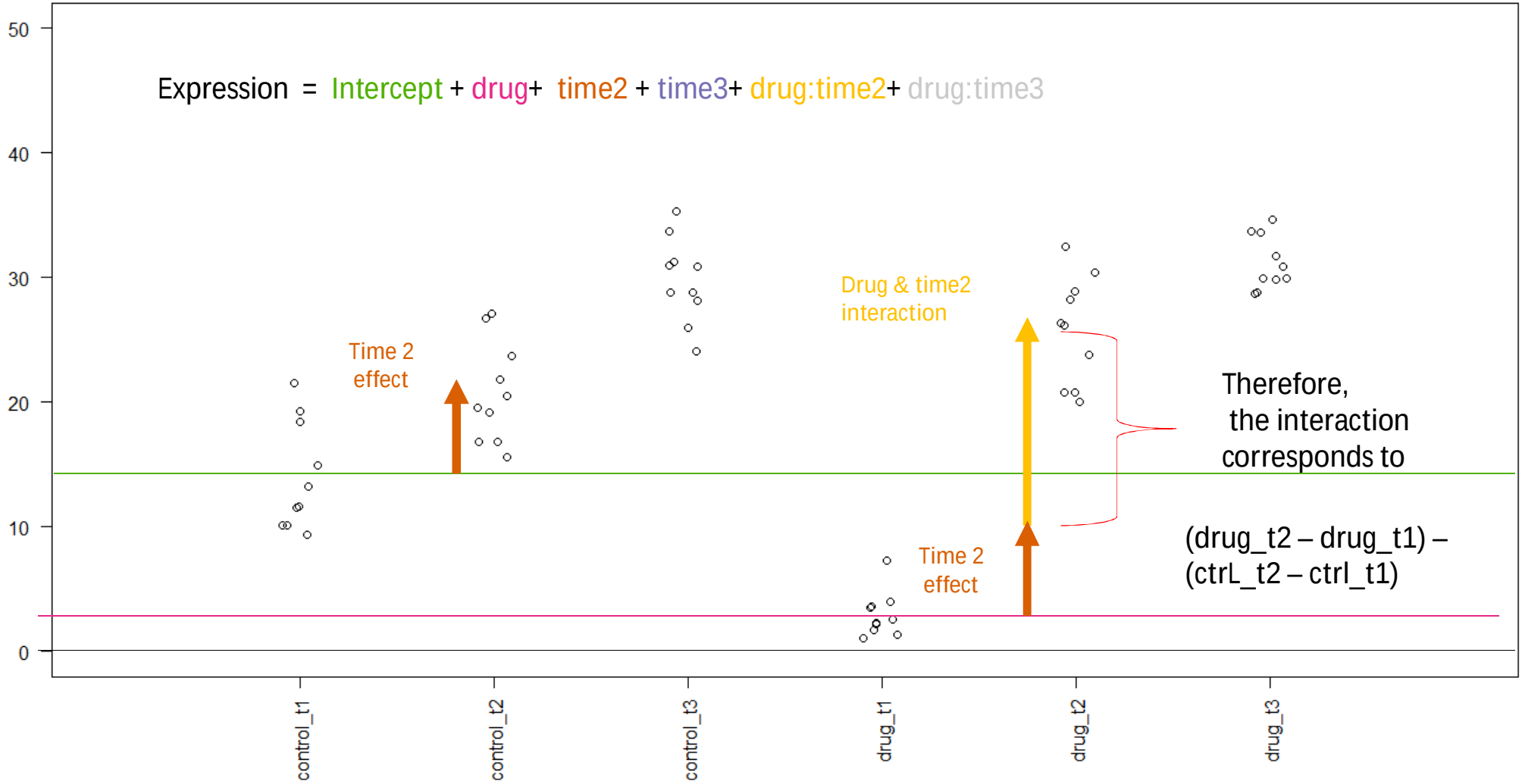

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time 2 effect
By the same token,
this corresponds to
the difference between ctrl_t2 and ctrl_t1
Baseline effects
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3


Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time 2 effect
By the same token,
this corresponds to
the difference between ctrl_t2 and ctrl_t1
Baseline effects
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3


Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time 2 effect
By the same token,
this corresponds to
the difference between ctrl_t2 and ctrl_t1
Baseline effects
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

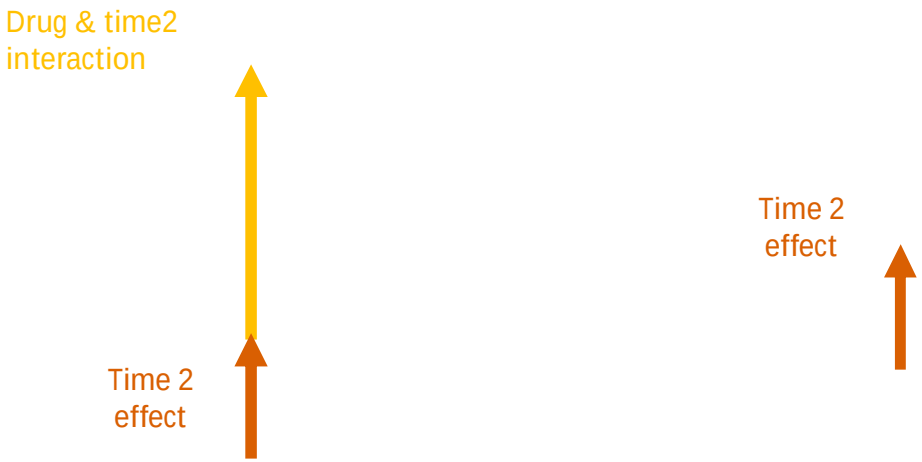

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time 2 effect
By the same token,
this corresponds to
the difference between ctrl_t2 and ctrl_t1
Baseline effects
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3


Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time 2 effect
By the same token,
this corresponds to
the difference between ctrl_t2 and ctrl_t1
Baseline effects
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

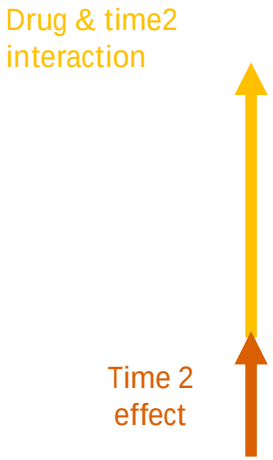

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
50
40
30
20
10
0
Time 2 effect
Drug & time2 interaction
Time 2 effect
Therefore, the interaction corresponds to
(drug_t2 – drug_t1) – (ctrL_t2 – ctrl_t1)
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

Interaction: Difference of Differences

Interaction: Difference of Differences



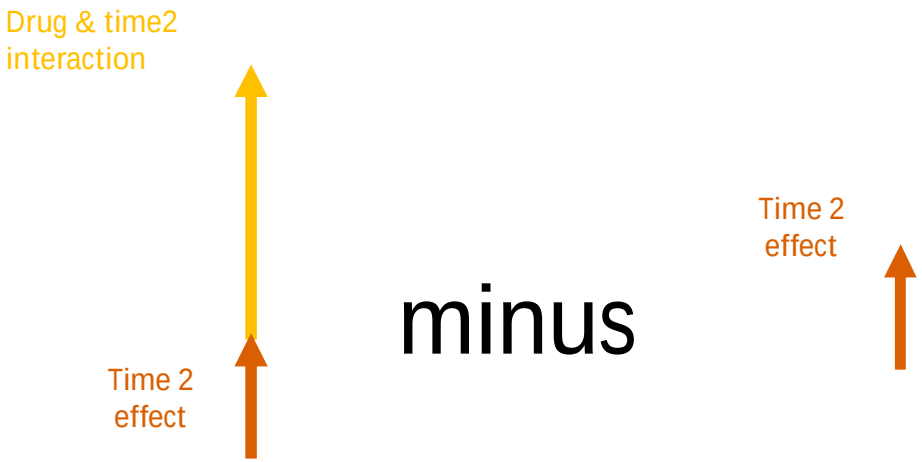
Interaction: Difference of Differences



(drug_t2 – drug_t1)

(ctrL_t2 – ctrl_t1)

Interaction: Difference of Differences

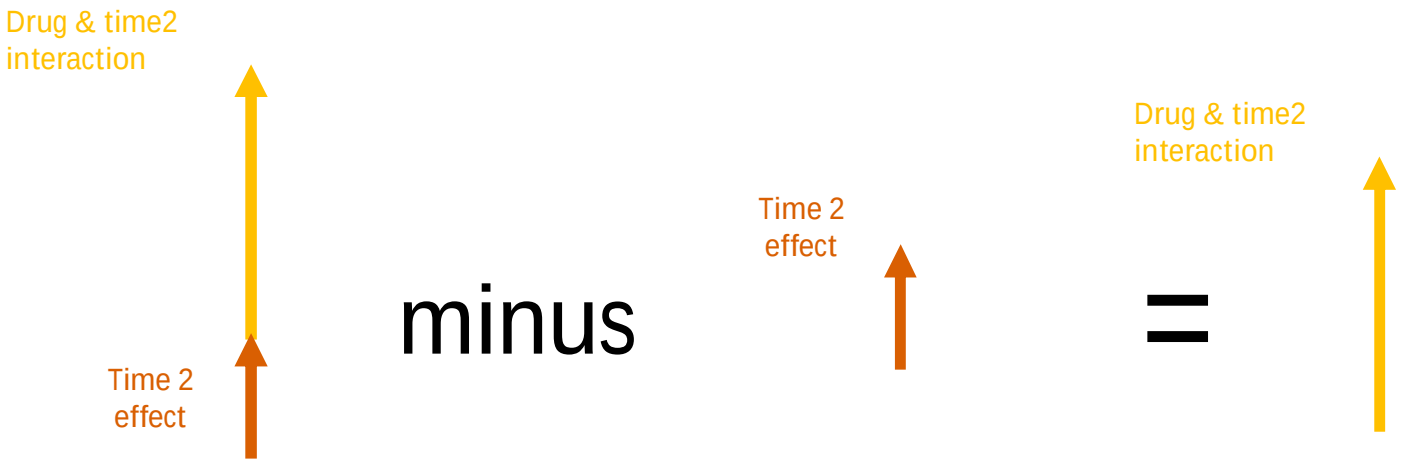


(drug_t2 – drug_t1) - (ctrL_t2 – ctrl_t1)

Interaction: Difference of Differences

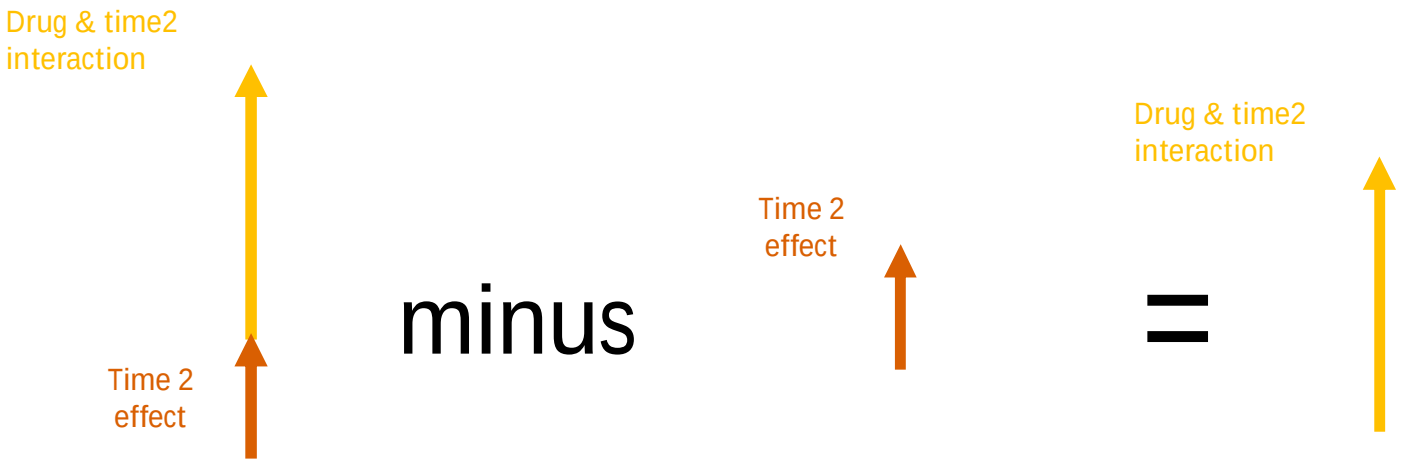
(drug_t2 – drug_t1) - (ctrL_t2 – ctrl_t1)

Interaction: Difference of Differences



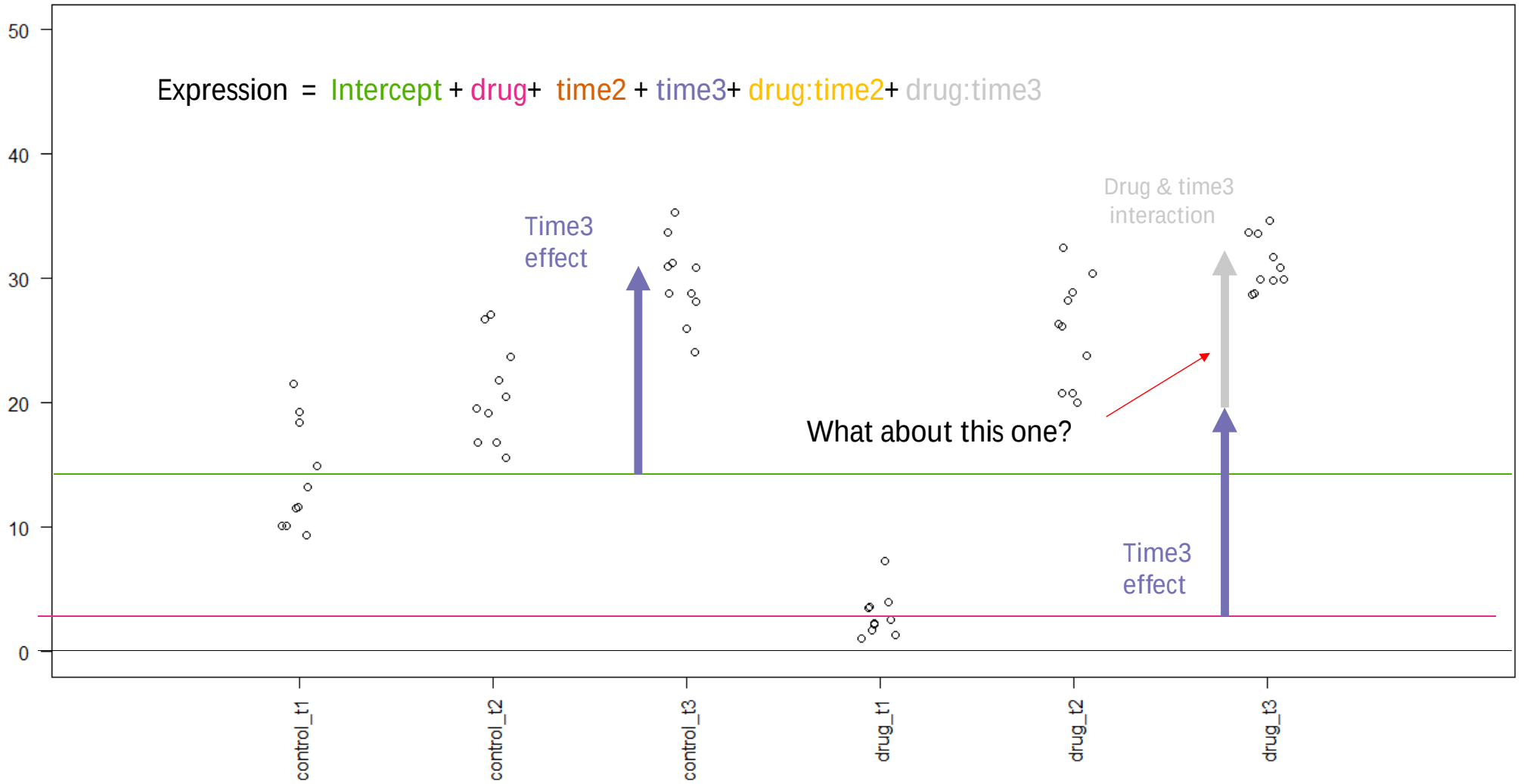
(drug_t2 – drug_t1) - (ctrL_t2 – ctrl_t1)

Interaction: Difference of Differences



(drug_t2 – drug_t1) - (ctrL_t2 – ctrl_t1)

Does this gene respond differently to drug at time t2 vs time1 than to placebo at time2 vs time1?


Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3
Time3 effect
Drug & time3 interaction
What about this one?
Time3 effect
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

What if we want to compute

(drug_t3 – drug_t2) - (ctrl_t3 – ctrl_t2) ?

What if we want to compute

(drug_t3 – drug_t2) - (ctrl_t3 – ctrl_t2) ?

“What are the genes that are differentially expressed between drugged mice in time3 and in time2 while controlling for vehicle effects?”

What if we want to compute

(drug_t3 – drug_t2) - (ctrl_t3 – ctrl_t2) ?

“What are the genes that are differentially expressed between drugged mice in time3 and in time2 while controlling for vehicle effects?”

We know that drug:time2 corresponds to?

What if we want to compute

(drug_t3 – drug_t2) - (ctrl_t3 – ctrl_t2) ?

“What are the genes that are differentially expressed between drugged mice in time3 and in time2 while controlling for vehicle effects?”

We know that drug:time2 corresponds to?

drug:time2:

What if we want to compute

(drug_t3 – drug_t2) - (ctrl_t3 – ctrl_t2) ?

“What are the genes that are differentially expressed between drugged mice in time3 and in time2 while controlling for vehicle effects?”

We know that drug:time2 corresponds to?

drug:time2: (drug_t2 – drug_t1) - (ctrl_t2 – ctrl_t1)

What if we want to compute

(drug_t3 – drug_t2) - (ctrl_t3 – ctrl_t2) ?

“What are the genes that are differentially expressed between drugged mice in time3 and in time2 while controlling for vehicle effects?”

We know that drug:time2 corresponds to?

drug:time2: (drug_t2 – drug_t1) - (ctrl_t2 – ctrl_t1)

drug:time3

What if we want to compute

(drug_t3 – drug_t2) - (ctrl_t3 – ctrl_t2) ?

“What are the genes that are differentially expressed between drugged mice in time3 and in time2 while controlling for vehicle effects?”

We know that drug:time2 corresponds to?

drug:time2: (drug_t2 – drug_t1) - (ctrl_t2 – ctrl_t1)

drug:time3 (drug_t3 – drug_t1) - (ctrl_t3 – ctrl_t1)

What if we want to compute

(drug_t3 – drug_t2) - (ctrl_t3 – ctrl_t2) ?

“What are the genes that are differentially expressed between drugged mice in time3 and in time2 while controlling for vehicle effects?”

We know that drug:time2 corresponds to?

drug:time2: (drug_t2 – drug_t1) - (ctrl_t2 – ctrl_t1)
drug:time3 (drug_t3 – drug_t1) - (ctrl_t3 – ctrl_t1)

So,

What if we want to compute

(drug_t3 – drug_t2) - (ctrl_t3 – ctrl_t2) ?

"What are the genes that are differentially expressed between drugged mice in time3 and in time2 while controlling for vehicle effects?"

We know that drug:time2 corresponds to?

drug:time2: (drug_t2 – drug_t1) - (ctrl_t2 – ctrl_t1)
drug:time3 (drug_t3 – drug_t1) - (ctrl_t3 – ctrl_t1)

So,
drug:time3 - drug:time2:

What if we want to compute

(drug_t3 – drug_t2) - (ctrl_t3 – ctrl_t2) ?

"What are the genes that are differentially expressed between drugged mice in time3 and in time2 while controlling for vehicle effects?"

We know that drug:time2 corresponds to?

drug:time2: (drug_t2 – drug_t1) - (ctrl_t2 – ctrl_t1)
drug:time3 (drug_t3 – drug_t1) - (ctrl_t3 – ctrl_t1)

So,
drug:time3 - drug:time2:
(drug_t3 – drug_t2) - (ctrl_t3 – ctrl_t2)

Use contrast matrix for arbitrary comparison

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3

Use contrast matrix for arbitrary comparison

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3

Use contrast matrix for arbitrary comparison

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3

coefficients

$$\begin{bmatrix} \text{Intercept} \\ \text{Drug} \\ \text{Time2} \\ \text{Time3} \\ \text{Drug:time2} \\ \text{Drug:time3} \end{bmatrix}$$

Use contrast matrix for arbitrary comparison

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3

contrast

[0, 0, 0, 0, -1, 1]

coefficients

Intercept
Drug
Time2
Time3
Drug:time2
Drug:time3

Use contrast matrix for arbitrary comparison

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3

contrast

[0, 0, 0, 0, -1, 1]

coefficients

Intercept
Drug
Time2
Time3
Drug:time2
Drug:time3

=

Use contrast matrix for arbitrary comparison

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3

contrast

[0, 0, 0, 0, -1, 1]

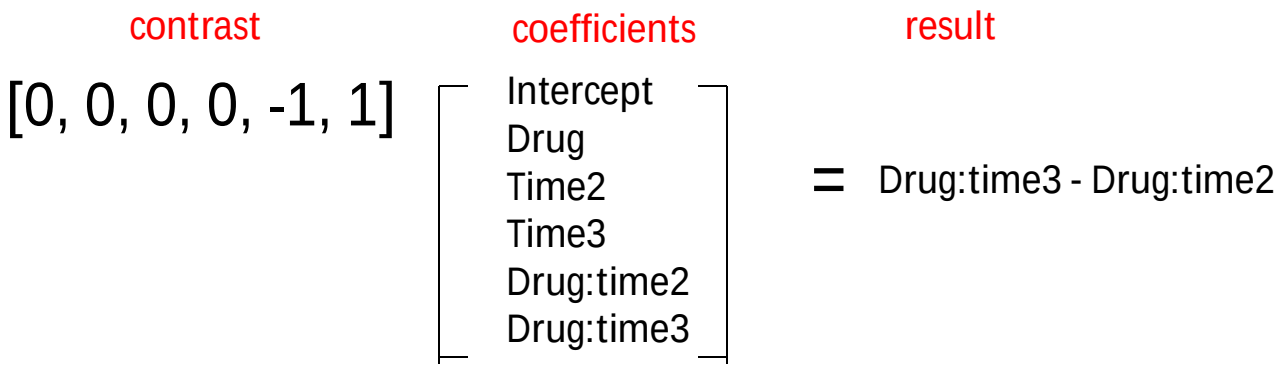
coefficients

Intercept
Drug
Time2
Time3
Drug:time2
Drug:time3

= Drug:time3 - Drug:time2

Use contrast matrix for arbitrary comparison

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3



Use contrast matrix for arbitrary comparison

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3

Use contrast matrix for arbitrary comparison

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3

```
> require(multcomp)
```

Use contrast matrix for arbitrary comparison

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3

```
> require(multcomp)
> m3 <- lm(expression ~ class + time + time:class , data=data)
```

Use contrast matrix for arbitrary comparison

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3

```
> require(multcomp)
> m3 <- lm(expression ~ class + time + time:class , data=data)
> C <- matrix(c(0,0,0,0,-1,1), 1) # Define contrast vector
```

Use contrast matrix for arbitrary comparison

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3

```
> require(multcomp)
> m3 <- lm(expression ~ class + time + time:class , data=data)
> C <- matrix(c(0,0,0,0,-1,1), 1) # Define contrast vector
> summary(glht(m3, linfct=C))
```

Use contrast matrix for arbitrary comparison

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3

```
> require(multcomp)
> m3 <- lm(expression ~ class + time + time:class , data=data)
> C <- matrix(c(0,0,0,0,-1,1), 1) # Define contrast vector
> summary(glht(m3, linfct=C))
```

general linear hypothesis test

Use contrast matrix for arbitrary comparison

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3

```
> require(multcomp)
> m3 <- lm(expression ~ class + time + time:class , data=data)
> C <- matrix(c(0,0,0,0,-1,1), 1) # Define contrast vector
> summary(glht(m3, linfct=C))

             Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = expression ~ class + time + time:class, data = data)

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)
1 == 0   -4.067      1.958  -2.077   0.0425 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

general linear hypothesis test

Use contrast matrix for arbitrary comparison

Expression = Intercept + drug+ time2 + time3+ drug:time2+ drug:time3

```
> require(multcomp)
> m3 <- lm(expression ~ class + time + time:class , data=data)
> C <- matrix(c(0,0,0,0,-1,1), 1) # Define contrast vector
> summary(glht(m3, linfct=C))

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = expression ~ class + time + time:class, data = data)

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)
1 == 0   -4.067      1.958  -2.077   0.0425 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

If using interaction is confusing…

class	time	expression
control	t1	10.100046
control	t1	16.419413
control	t1	14.077125
control	t2	17.380658
control	t2	17.914727
control	t2	25.256204
control	t3	28.275683
control	t3	26.393841
control	t3	31.831851
drug	t1	8.899719
drug	t1	6.202112
drug	t1	4.599608
drug	t2	22.730259
drug	t2	26.322069
drug	t2	31.084232
drug	t3	25.977785
drug	t3	24.361100
drug	t3	30.258291

If using interaction is confusing…

class	time	expression
control	t1	10.100046
control	t1	16.419413
control	t1	14.077125
control	t2	17.380658
control	t2	17.914727
control	t2	25.256204
control	t3	28.275683
control	t3	26.393841
control	t3	31.831851
drug	t1	8.899719
drug	t1	6.202112
drug	t1	4.599608
drug	t2	22.730259
drug	t2	26.322069
drug	t2	31.084232
drug	t3	25.977785
drug	t3	24.361100
drug	t3	30.258291

Make combined factor

If using interaction is confusing…

class	time	expression
control	t1	10.100046
control	t1	16.419413
control	t1	14.077125
control	t2	17.380658
control	t2	17.914727
control	t2	25.256204
control	t3	28.275683
control	t3	26.393841
control	t3	31.831851
drug	t1	8.899719
drug	t1	6.202112
drug	t1	4.599608
drug	t2	22.730259
drug	t2	26.322069
drug	t2	31.084232
drug	t3	25.977785
drug	t3	24.361100
drug	t3	30.258291

Make combined factor

class	time	expression	classtime
control	t1	10.100046	control_t1
control	t1	16.419413	control_t1
control	t1	14.077125	control_t1
control	t2	17.380658	control_t2
control	t2	17.914727	control_t2
control	t2	25.256204	control_t2
control	t3	28.275683	control_t3
control	t3	26.393841	control_t3
control	t3	31.831851	control_t3
drug	t1	8.899719	drug_t1
drug	t1	6.202112	drug_t1
drug	t1	4.599608	drug_t1
drug	t2	22.730259	drug_t2
drug	t2	26.322069	drug_t2
drug	t2	31.084232	drug_t2
drug	t3	25.977785	drug_t3
drug	t3	24.361100	drug_t3
drug	t3	30.258291	drug_t3

If the interaction is confusing…

class	time	expression	classtime
control	t1	10.100046	control_t1
control	t1	16.419413	control_t1
control	t1	14.077125	control_t1
control	t2	17.380658	control_t2
control	t2	17.914727	control_t2
control	t2	25.256204	control_t2
control	t3	28.275683	control_t3
control	t3	26.393841	control_t3
control	t3	31.831851	control_t3
drug	t1	8.899719	drug_t1
drug	t1	6.202112	drug_t1
drug	t1	4.599608	drug_t1
drug	t2	22.730259	drug_t2
drug	t2	26.322069	drug_t2
drug	t2	31.084232	drug_t2
drug	t3	25.977785	drug_t3
drug	t3	24.361100	drug_t3
drug	t3	30.258291	drug_t3

Fit classtime instead, and use contrast

class	time	expression	classtime
control	t1	10.100046	control_t1
control	t1	16.419413	control_t1
control	t1	14.077125	control_t1
control	t2	17.380658	control_t2
control	t2	17.914727	control_t2
control	t2	25.256204	control_t2
control	t3	28.275683	control_t3
control	t3	26.393841	control_t3
control	t3	31.831851	control_t3
drug	t1	8.899719	drug_t1
drug	t1	6.202112	drug_t1
drug	t1	4.599608	drug_t1
drug	t2	22.730259	drug_t2
drug	t2	26.322069	drug_t2
drug	t2	31.084232	drug_t2
drug	t3	25.977785	drug_t3
drug	t3	24.361100	drug_t3
drug	t3	30.258291	drug_t3

Fit classtime instead, and use contrast

class	time	expression	classtime
control	t1	10.100046	control_t1
control	t1	16.419413	control_t1
control	t1	14.077125	control_t1
control	t2	17.380658	control_t2
control	t2	17.914727	control_t2
control	t2	25.256204	control_t2
control	t3	28.275683	control_t3
control	t3	26.393841	control_t3
control	t3	31.831851	control_t3
drug	t1	8.899719	drug_t1
drug	t1	6.202112	drug_t1
drug	t1	4.599608	drug_t1
drug	t2	22.730259	drug_t2
drug	t2	26.322069	drug_t2
drug	t2	31.084232	drug_t2
drug	t3	25.977785	drug_t3
drug	t3	24.361100	drug_t3
drug	t3	30.258291	drug_t3

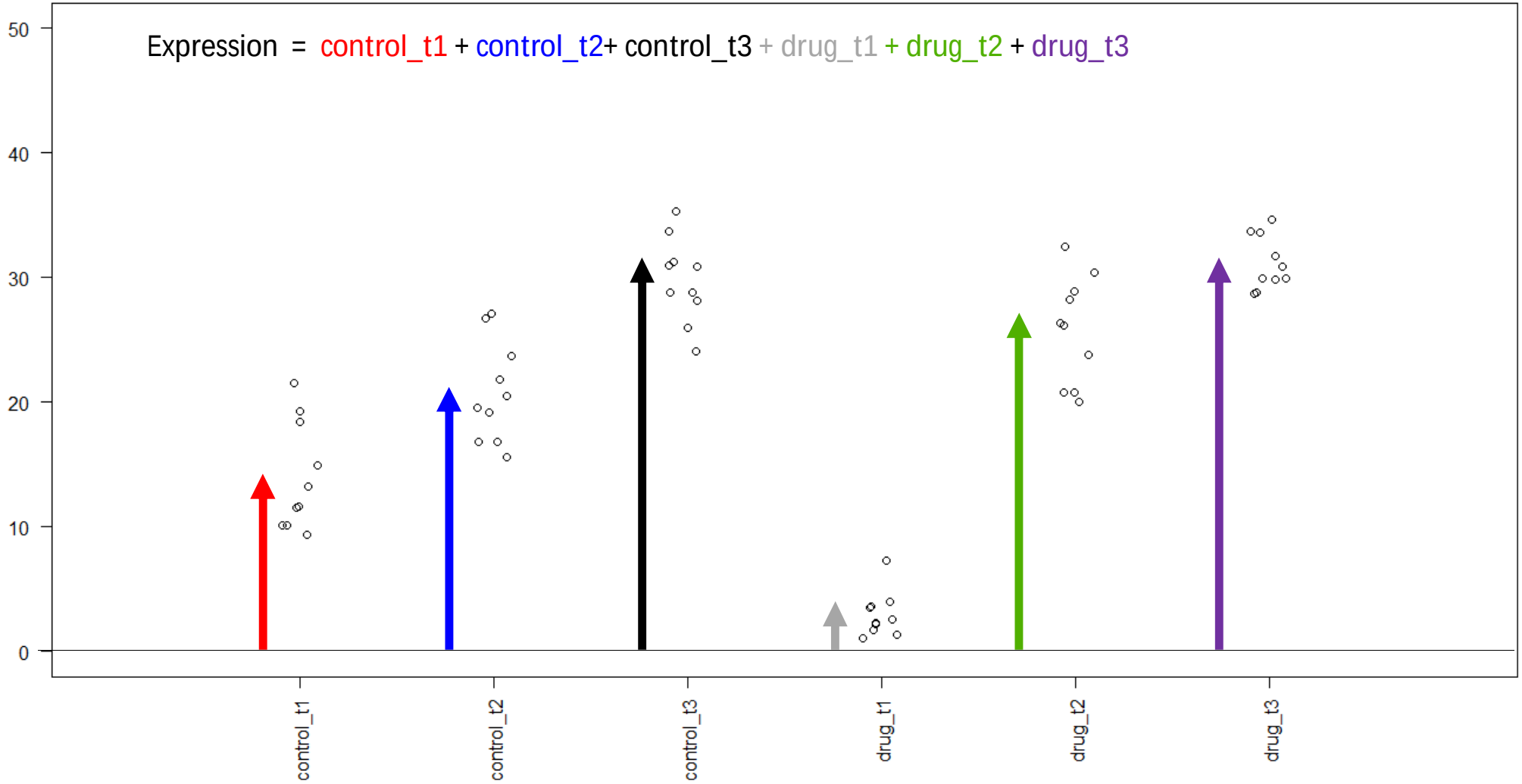
lm (expression ~ classtime ,data = data)

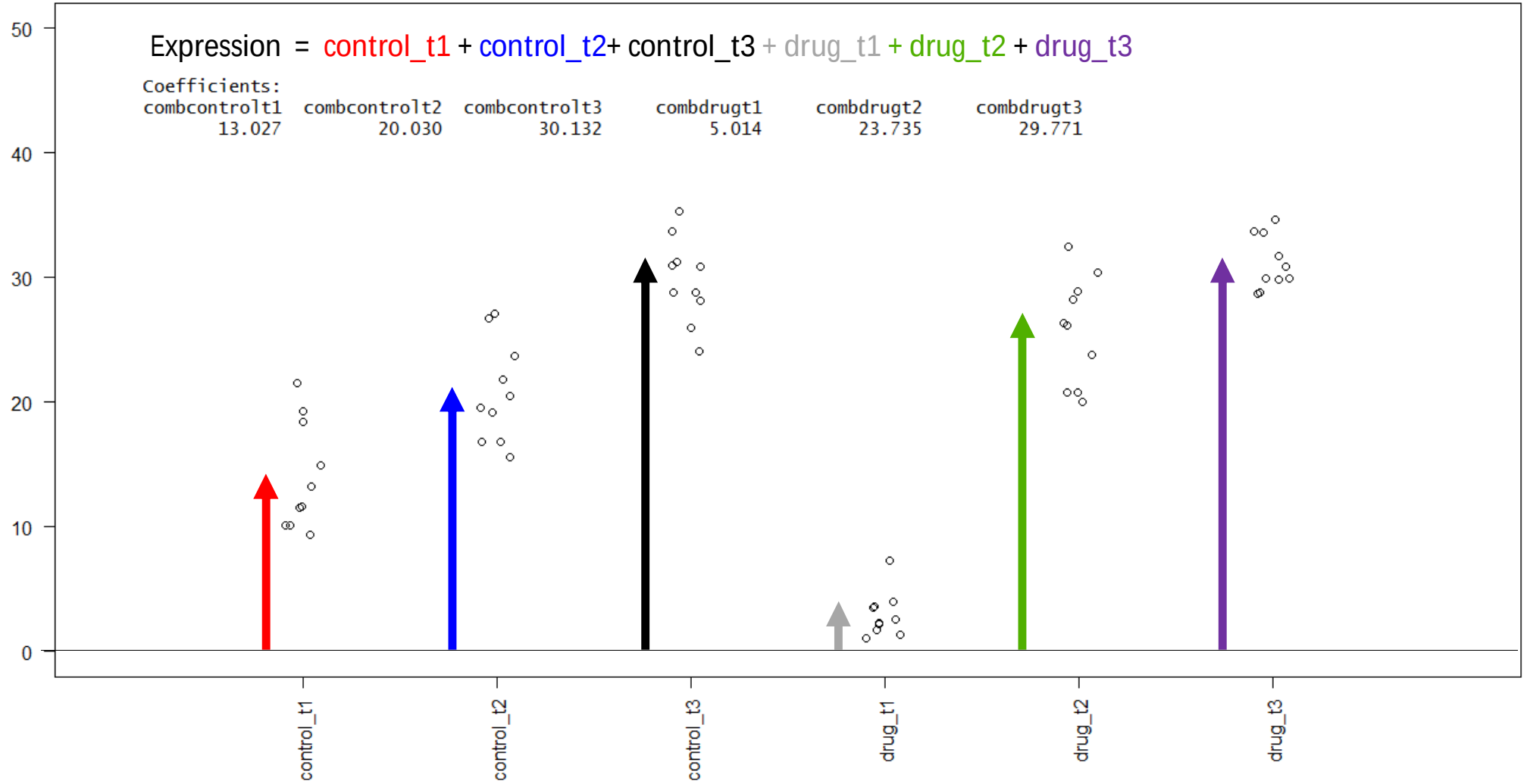
Fit classtime instead, and use contrast

	class	time	expression	classtime
control	control	t1	10.100046	control_t1
control	control	t1	16.419413	control_t1
control	control	t1	14.077125	control_t1
control	control	t2	17.380658	control_t2
control	control	t2	17.914727	control_t2
control	control	t2	25.256204	control_t2
control	control	t3	28.275683	control_t3
control	control	t3	26.393841	control_t3
control	control	t3	31.831851	control_t3
drug	drug	t1	8.899719	drug_t1
drug	drug	t1	6.202112	drug_t1
drug	drug	t1	4.599608	drug_t1
drug	drug	t2	22.730259	drug_t2
drug	drug	t2	26.322069	drug_t2
drug	drug	t2	31.084232	drug_t2
drug	drug	t3	25.977785	drug_t3
drug	drug	t3	24.361100	drug_t3
drug	drug	t3	30.258291	drug_t3

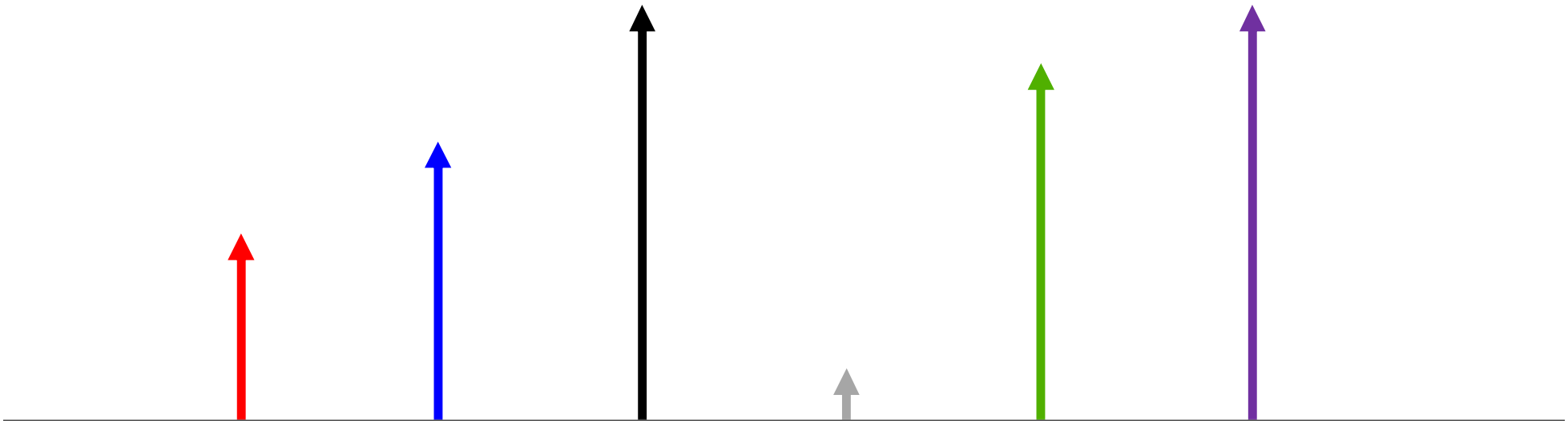
lm (expression ~ classtime ,data = data)

```
Coefficients:
combcontrolt1  combcontrolt2  combcontrolt3     combdrugt1     combdrugt2     combdrugt3
       13.027         20.030         30.132          5.014         23.735         29.771
```

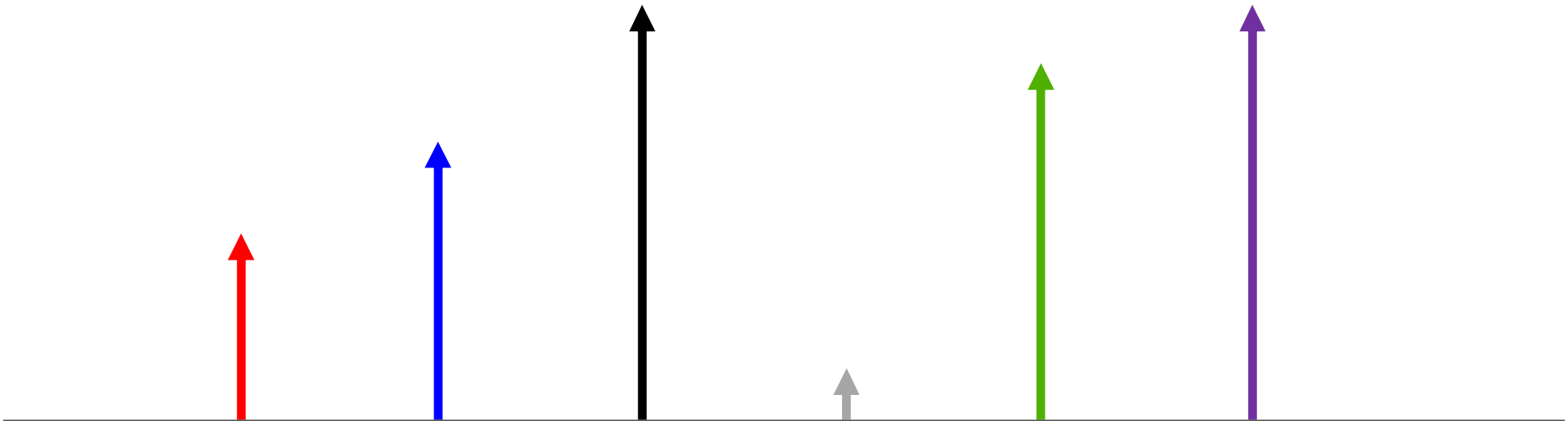

Expression = control_t1 + control_t2+ control_t3 + drug_t1 + drug_t2 + drug_t3
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3


Expression = control_t1 + control_t2+ control_t3 + drug_t1 + drug_t2 + drug_t3
Coefficients:
combcontrolt1 combcontrolt2 combcontrolt3 combdrugt1 combdrugt2 combdrugt3
13.027 20.030 30.132 5.014 23.735 29.771
50
40
30
20
10
0
control_t1
control_t2
control_t3
drug_t1
drug_t2
drug_t3

Expression = control_t1 + control_t2+ control_t3 + drug_t1 + drug_t2 + drug_t3

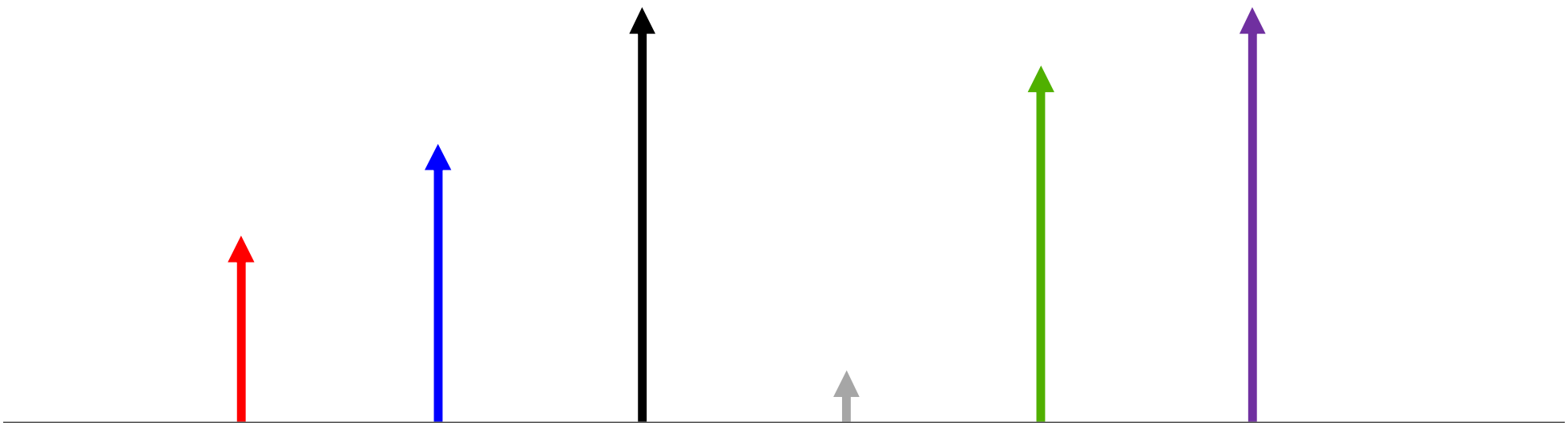


Expression = control_t1 + control_t2+ control_t3 + drug_t1 + drug_t2 + drug_t3



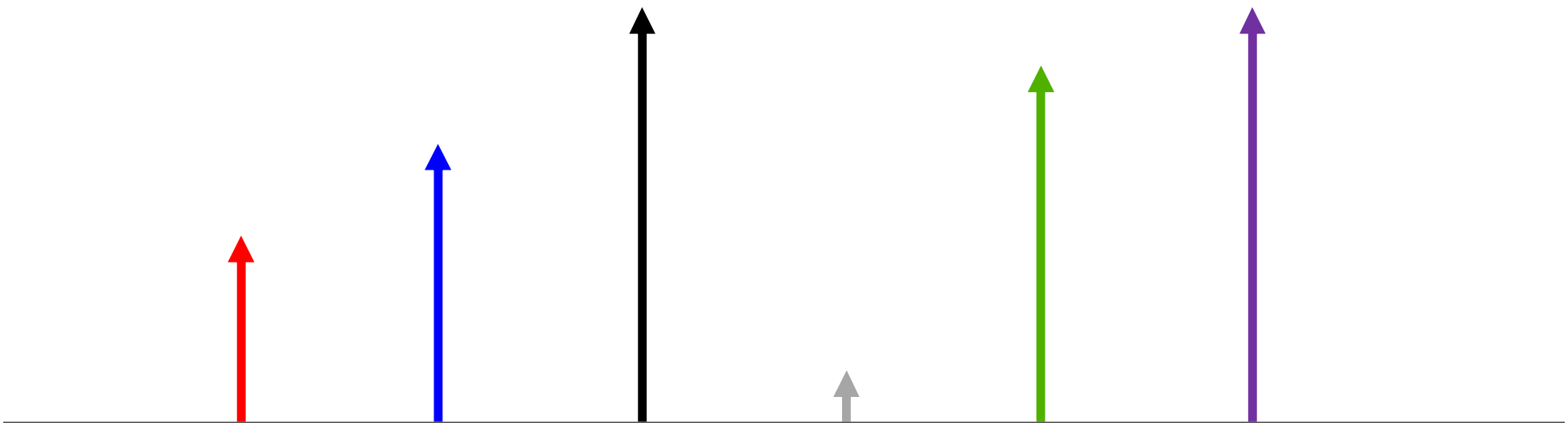
To compute , (drug_t3- drug_t2) – (control_t3 – control_t2)

Expression = control_t1 + control_t2+ control_t3 + drug_t1 + drug_t2 + drug_t3



To compute , (drug_t3- drug_t2) – (control_t3 – control_t2)

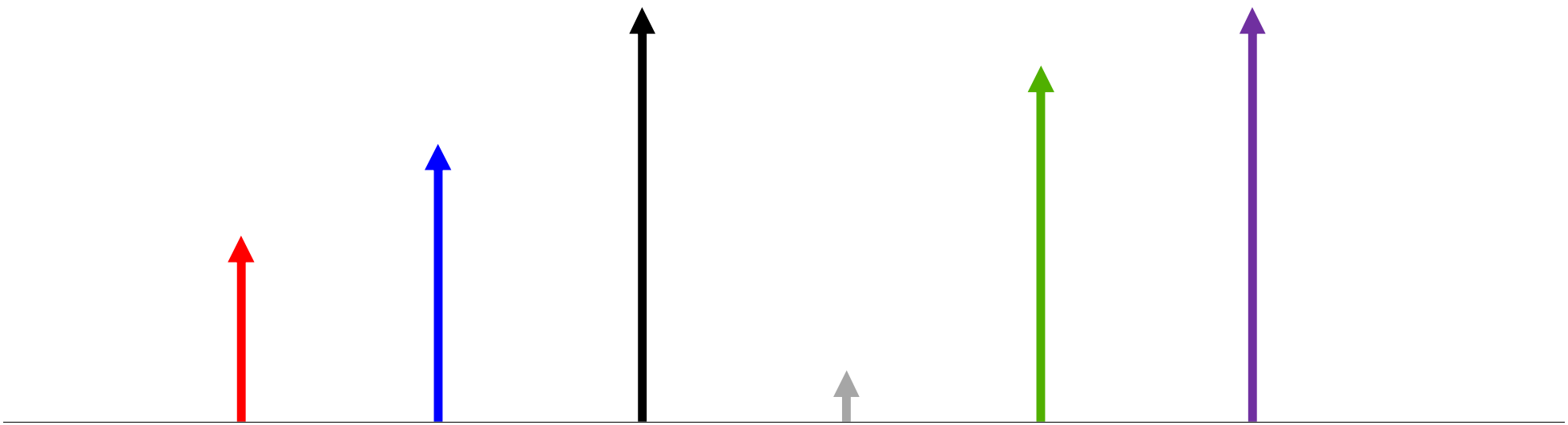
Expression = control_t1 + control_t2+ control_t3 + drug_t1 + drug_t2 + drug_t3



Use contrast:

To compute , (drug_t3- drug_t2) – (control_t3 – control_t2)

Expression = control_t1 + control_t2+ control_t3 + drug_t1 + drug_t2 + drug_t3



Use contrast: [0, 1, -1 , 0, -1, 1]

To compute , (drug_t3- drug_t2) – (control_t3 – control_t2)

Use contrast: [0, 1, -1 , 0, -1, 1]

To compute , (drug_t3- drug_t2) – (control_t3 – control_t2)

Use contrast: [0, 1, -1 , 0, -1, 1]

To compute , (drug_t3- drug_t2) – (control_t3 – control_t2)

Use contrast: [0, 1, -1 , 0, -1, 1]

To compute , (drug_t3- drug_t2) – (control_t3 – control_t2)

Use contrast: [0, 1, -1 , 0, -1, 1]

```
m4 <- lm(expression ~ 0 + comb , dat3_df_tidy %>% mutate(comb=paste0(class,time)))
C <- matrix(c(0,1,-1,0,-1,1), 1)
diff_diff <- glht(m4, linfct=C)
summary(diff_diff)
```

To compute , (drug_t3- drug_t2) – (control_t3 – control_t2)

Use contrast: [0, 1, -1 , 0, -1, 1]

```
m4 <- lm(expression ~ 0 + comb , dat3_df_tidy %>% mutate(comb=paste0(class,time)))
C <- matrix(c(0,1,-1,0,-1,1), 1)
diff_diff <- glht(m4, linfct=C)
summary(diff_diff)
```

```
         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = expression ~ 0 + comb, data = dat3_df_tidy %>% mutate(comb = paste0(class,
    time)))

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)
1 == 0   -4.067      1.958  -2.077   0.0425 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

To compute , (drug_t3- drug_t2) – (control_t3 – control_t2)

Use contrast: [0, 1, -1 , 0, -1, 1]

```
m4 <- lm(expression ~ 0 + comb , dat3_df_tidy %>% mutate(comb=paste0(class,time)))
C <- matrix(c(0,1,-1,0,-1,1), 1)
diff_diff <- glht(m4, linfct=C)
summary(diff_diff)
```

```
         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = expression ~ 0 + comb, data = dat3_df_tidy %>% mutate(comb = paste0(class,
    time)))

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)
1 == 0   -4.067      1.958  -2.077   0.0425 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

To compute , (drug_t3- drug_t2) – (control_t3 – control_t2)

Use contrast: [0, 1, -1 , 0, -1, 1]

```
m4 <- lm(expression ~ 0 + comb , dat3_df_tidy %>% mutate(comb=paste0(class,time)))
C <- matrix(c(0,1,-1,0,-1,1), 1)
diff_diff <- glht(m4, linfc
summary(diff_diff)


         Simultaneous Tests for Gener

Fit: lm(formula = expression ~ 0 + cc
    time)))

Linear Hypotheses:
       Estimate Std. Error t value Pr
1 == 0   -4.067      1.958  -2.077
---
Signif. codes:  0 '***' 0.001 '**' 0.
(Adjusted p values reported -- single
```

Previous method

```
> require(multcomp)
> m3 <- lm(expression ~ class + time + time:class , data=data)
> C <- matrix(c(0,0,0,0,-1,1), 1) # Define contrast vector
> summary(glht(m3, linfct=C))

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = expression ~ class + time + time:class, data = data)

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)
1 == 0   -4.067      1.958  -2.077   0.0425 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

To compute , (drug_t3- drug_t2) – (control_t3 – control_t2)

Use contrast: [0, 1, -1 , 0, -1, 1]

```
m4 <- lm(expression ~ 0 + comb , dat3_df_tidy %>% mutate(comb=paste0(class,time)))
C <- matrix(c(0,1,-1,0,-1,1), 1)
diff_diff <- glht(m4, linfc
summary(diff_diff)
```

```
        Simultaneous Tests for Gener

Fit: lm(formula = expression ~ 0 + cc
    time)))

Linear Hypotheses:
       Estimate Std. Error t value Pr
1 == 0   -4.067      1.958  -2.077
---
Signif. codes:  0 '***' 0.001 '**' 0.
(Adjusted p values reported -- single
```

Previous method

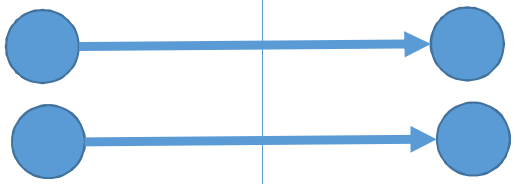
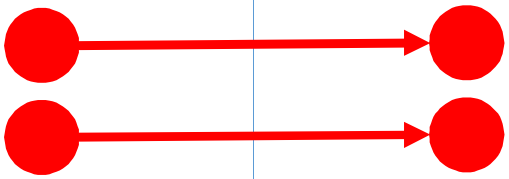
```
> require(multcomp)
> m3 <- lm(expression ~ class + time + time:class , data=data)
> C <- matrix(c(0,0,0,0,-1,1), 1) # Define contrast vector
> summary(glht(m3, linfct=C))

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = expression ~ class + time + time:class, data = data)

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)
1 == 0   -4.067      1.958  -2.077   0.0425 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Level 4

“Pre-Post-Control” Design

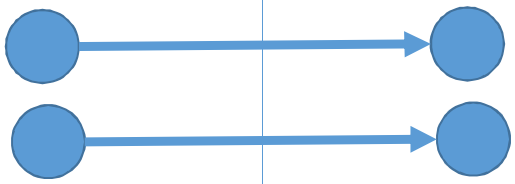
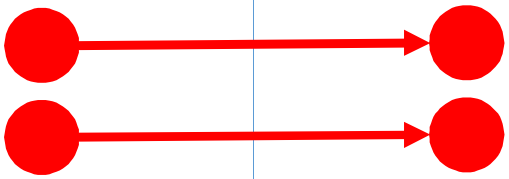
Study design

- Pre-post-control study design

	pre treatment	post treatment
control		
Drug		

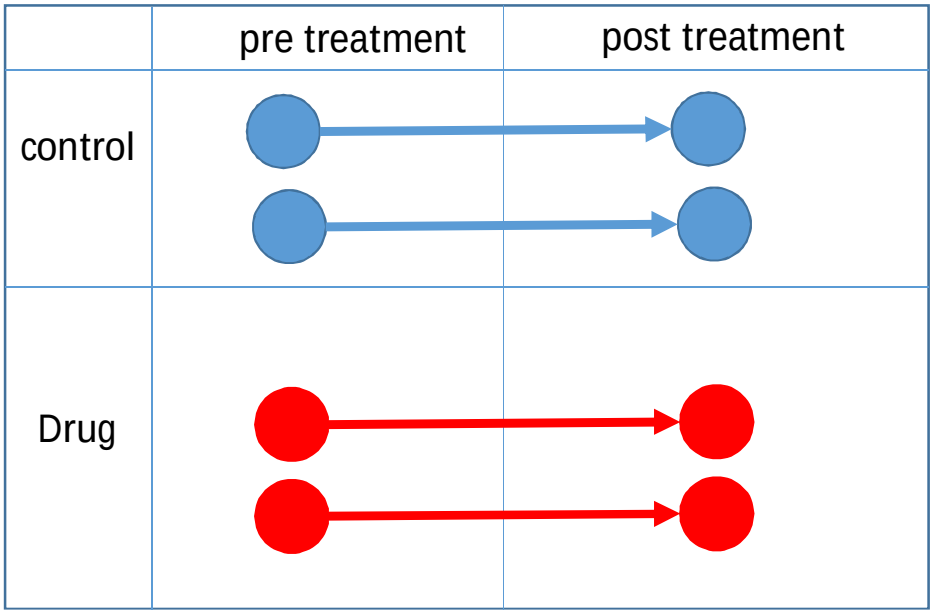
Study design

- Pre-post-control study design

	pre treatment	post treatment
control		
Drug		

Study design

- Pre-post-control study design



expression	patient	time	treatment
27.13863	patient1	pre	control
28.83620	patient1	post	control
30.76516	patient2	pre	control
32.46562	patient2	post	control
32.53838	patient3	pre	control
35.62194	patient3	post	control
31.10912	patient4	pre	control
31.21611	patient4	post	control
30.47516	patient5	pre	control
30.89179	patient5	post	control
29.57050	patient6	pre	drug
30.64656	patient6	post	drug
34.91339	patient7	pre	drug
31.69578	patient7	post	drug
26.97193	patient8	pre	drug
27.11820	patient8	post	drug
28.50800	patient9	pre	drug
26.26490	patient9	post	drug
28.76724	patient10	pre	drug
34.15198	patient10	post	drug

TIMTOADY
(There’s more than one way to do it)

TIMTOADY
(There’s more than one way to do it)

TIMTOADY
(There’s more than one way to do it)

- Using Interaction

TIMTOADY
(There’s more than one way to do it)

- Using Interaction
  - Use expression as dependent variable and use treatment, Patient and treatment interaction, and time and treatment interaction as independent variables

TIMTOADY
(There's more than one way to do it)

- $\sim TTrreeaattmmeennttTTrreeaattmmeennttPPaattiieennttTTrreeaattmmeennttTTiimmee$
- Using Interaction
  - Use expression as dependent variable and use treatment, Patient and treatment interaction, and time and treatment interaction as independent variables

    $$Expression \sim Treatment + Treatment{:}\,Patient + Treatment{:}\,Time$$

TIMTOADY
(There's more than one way to do it)

- $\sim TTrreeaattmmeennttTTrreeaattmmeennttPPaattiieennttTTrreeaattmmeennttTTiimmee$
- Using Interaction
  - Use expression as dependent variable and use treatment, Patient and treatment interaction, and time a ndent variables
    $eatment: Patient + Treatment: Time$

expression	patient	time	treatment
27.13863	patient1	pre	control
28.83620	patient1	post	control
30.76516	patient2	pre	control
32.46562	patient2	post	control
32.53838	patient3	pre	control
35.62194	patient3	post	control
31.10912	patient4	pre	control
31.21611	patient4	post	control
30.47516	patient5	pre	control
30.89179	patient5	post	control
29.57050	patient6	pre	drug
30.64656	patient6	post	drug
34.91339	patient7	pre	drug
31.69578	patient7	post	drug
26.97193	patient8	pre	drug
27.11820	patient8	post	drug
28.50800	patient9	pre	drug
26.26490	patient9	post	drug
28.76724	patient10	pre	drug
34.15198	patient10	post	drug

TIMTOADY
(There's more than one way to do it)

- $\sim TTrreeaattmmeennttt+TTrreeaattmmeennttt{:}PPaattiieennttt+TTrreeaattmmeennttt{:}TTiimmee$
- Using Interaction
  - Use expression as dependent variable and use treatment, Patient and treatment interaction, and time as independent variables
    $eatment: Patient \cdot$

expression	patient	time	treatment
27.13863	patient1	pre	control
28.83620	patient1	post	control
30.76516	patient2	pre	control
32.46562	patient2	post	control
32.53838	patient3	pre	control
35.62194	patient3	post	control
31.10912	patient4	pre	control
31.21611	patient4	post	control
30.47516	patient5	pre	control
30.89179	patient5	post	control
29.57050	patient6	pre	drug
30.64656	patient6	post	drug
34.91339	patient7	pre	drug
31.69578	patient7	post	drug
26.97193	patient8	pre	drug
27.11820	patient8	post	drug
28.50800	patient9	pre	drug
26.26490	patient9	post	drug
28.76724	patient10	pre	drug
34.15198	patient10	post	drug

expression	patient	time	treatment
27.13863	patient1	pre	control
28.83620	patient1	post	control
30.76516	patient2	pre	control
32.46562	patient2	post	control
32.53838	patient3	pre	control
35.62194	patient3	post	control
31.10912	patient4	pre	control
31.21611	patient4	post	control
30.47516	patient5	pre	control
30.89179	patient5	post	control
29.57050	patient1	pre	drug
30.64656	patient1	post	drug
34.91339	patient2	pre	drug
31.69578	patient2	post	drug
26.97193	patient3	pre	drug
27.11820	patient3	post	drug
28.50800	patient4	pre	drug
26.26490	patient4	post	drug
28.76724	patient5	pre	drug
34.15198	patient5	post	drug

TIMTOADY
(There's more than one way to do it)

- $\sim TTrreeaattmmeennttt{+}TTrreeaattmmeennttt{:}PPaattiieennttt{+}TTrreeaattmmeennttt{:}TTiimmee$
- Using Interaction
  - Use expression as dependent variable and use treatment, Patient and treatment interaction, and time and treatment interaction as independent variables

    $$Expression \sim Treatment + Treatment{:}\,Patient + Treatment{:}\,Time$$

TIMTOADY
(There’s more than one way to do it)

- $\sim TTrreeaattmmeennttt{+}TTrreeaattmmeennttt{:}PPaattiieennttt{+}TTrreeaattmmeennttt{:}TTiimmee$
- Using Interaction
  - Use expression as dependent variable and use treatment, Patient and treatment interaction, and time and treatment interaction as independent variables
- Using Mixed model

TIMTOADY
(There’s more than one way to do it)

- $\sim TTrreeaattmmeennttTTrreeaattmmeennttPPaattiieennttTTrreeaattmmeennttTTiimmee$
- Using Interaction
  - Use expression as dependent variable and use treatment, Patient and treatment interaction, and time and treatment interaction as independent variables
- Using Mixed model
  - Use expression as dependent variable and use treatment and treatment and time interaction as fixed effects and patients as random effects.

TIMTOADY
(There's more than one way to do it)

- $\sim TTrreeaattmmeennttTTrreeaattmmeennttTTiimmee + 1treatment\ 1ttrreeaattmmeennttt\ 1treatment\ PPaattiieenntt)$
- $\sim TTrreeaattmmeennttTTrreeaattmmeennttPPaattiieenntttTTrreeaattmmeennttTTiimmee$
- Using Interaction
  - Use expression as dependent variable and use treatment, Patient and treatment interaction, and time and treatment interaction as independent variables
- Using Mixed model
  - Use expression as dependent variable and use treatment and treatment and time interaction as fixed effects and patients as random effects.

    $Expression \sim Treatment + Treatment{:}Time + (1 + treatment\,|Patient)$

TIMTOADY
(There's more than one way to do it)

- $\sim TTrreeaattmmeennttt+TTrreeaattmmeennttt:TTiimmee + 1+treatment\ 1+ttrreeaattmmeennttt\ 1+treatment\ PPaattiieennttt)$
- $\sim TTrreeaattmmeennttt+TTrreeaattmmeennttt:PPaattiieennttt+TTrreeaattmmeennttt:TTiimmee$
- Using Interaction
  - Use expression as dependent variable and use treatment, Patient and treatment interaction, and time and treatment interaction as independent variables
- Using Mixed model
  - Use exp able and use treatment and treatment and time interaction as fixed effects a cts.

    $Expr$ $itment:Time\ + (1 + treatment\ |Patient)$

expression	patient	time	treatment
27.13863	patient1	pre	control
28.83620	patient1	post	control
30.76516	patient2	pre	control
32.46562	patient2	post	control
32.53838	patient3	pre	control
35.62194	patient3	post	control
31.10912	patient4	pre	control
31.21611	patient4	post	control
30.47516	patient5	pre	control
30.89179	patient5	post	control
29.57050	patient6	pre	drug
30.64656	patient6	post	drug
34.91339	patient7	pre	drug
31.69578	patient7	post	drug
26.97193	patient8	pre	drug
27.11820	patient8	post	drug
28.50800	patient9	pre	drug
26.26490	patient9	post	drug
28.76724	patient10	pre	drug
34.15198	patient10	post	drug

TIMTOADY
(There's more than one way to do it)

- $\sim TTrreeaattmmeennttt{+}TTrreeaattmmeennttt{:}TTiimmee + 1{+}treatment\ 1{+}ttrreeaattmmeennttt\ 1{+}treatment\ PPaattiieennttt)$
- $\sim TTrreeaattmmeennttt{+}TTrreeaattmmeennttt{:}PPaattiieennttt{+}TTrreeaattmmeennttt{:}TTiimmee$
- Using Interaction
  - Use expression as dependent variable and use treatment, Patient and treatment interaction, and time and treatment interaction as independent variables
- Using Mixed model
  - Use exp able and use treatment and treatment and time interaction as fixed effects a cts.

    $Expr\ itment{:}Time\ + (1 + treatment\ |Patient)$

expression	patient	time	treatment
27.13863	patient1	pre	control
28.83620	patient1	post	control
30.76516	patient2	pre	control
32.46562	patient2	post	control
32.53838	patient3	pre	control
35.62194	patient3	post	control
31.10912	patient4	pre	control
31.21611	patient4	post	control
30.47516	patient5	pre	control
30.89179	patient5	post	control
29.57050	patient6	pre	drug
30.64656	patient6	post	drug
34.91339	patient7	pre	drug
31.69578	patient7	post	drug
26.97193	patient8	pre	drug
27.11820	patient8	post	drug
28.50800	patient9	pre	drug
26.26490	patient9	post	drug
28.76724	patient10	pre	drug
34.15198	patient10	post	drug

Assume Patients are random samples (from the hypothetical population)

TIMTOADY
(There’s more than one way to do it)

$$iff\ Expression \sim Treatment$$

$$ost\ Expression \sim Treatment + Pre\ Expression\ Covariate$$

TIMTOADY
(There’s more than one way to do it)

$$iff\ Expression \sim Treatment$$

$$ost\ Expression \sim Treatment + Pre\ Expression\ Covariate$$

TIMTOADY
(There’s more than one way to do it)

- Using Gain of score

$$iff\ Expression \sim Treatment$$

$$ost\ Expression \sim Treatment + Pre\ Expression\ Covariate$$

TIMTOADY
(There's more than one way to do it)

- Using Gain of score
  - Compute the difference between post and pre expression as a response variable (Diff Expression) and use treatment as a independent variable

$$iff\ Expression \sim Treatment$$

$$ost\ Expression \sim Treatment + Pre\ Expression\ Covariate$$

TIMTOADY
(There's more than one way to do it)

- $EExxppprreessssiioonn \sim TTrreeaattmmeennttt$
- Using Gain of score
  - Compute the difference between post and pre expression as a response variable (Diff Expression) and use treatment as a independent variable

$$Diff\ Expression \sim Treatment$$

$$ost\ Expression \sim Treatment + Pre\ Expression\ Covariate$$

TIMTOADY
(There's more than one way to do it)

- $EExxppprreessssiioonn \sim TTrreeaattmmeennttt$
- Using Gain of score
  - Compute the difference between post and pre expression as a response variable (Diff Expression), and use treatment as a independent variable

expression	patient	time	treatment
27.13863	patient1	pre	control
28.83620	patient1	post	control
30.76516	patient2	pre	control
32.46562	patient2	post	control
32.53838	patient3	pre	control
35.62194	patient3	post	control
31.10912	patient4	pre	control
31.21611	patient4	post	control
30.47516	patient5	pre	control
30.89179	patient5	post	control
29.57050	patient6	pre	drug
30.64656	patient6	post	drug
34.91339	patient7	pre	drug
31.69578	patient7	post	drug
26.97193	patient8	pre	drug
27.11820	patient8	post	drug
28.50800	patient9	pre	drug
26.26490	patient9	post	drug
28.76724	patient10	pre	drug
34.15198	patient10	post	drug

$pression \sim Treatment$

$eatment + Pre\ Expression\ Covariate$

TIMTOADY
(There's more than one way to do it)

- $EExxppprreessssiioonn \sim TTrreeaattmmeennttt$
- Using Gain of score
  - Compute the difference between post and pre expression as a response variable (Diff Expression), and use treatment as a independent variable

$pression \sim Treatment$

expression	patient	time	treatment
27.13863	patient1	pre	control
28.83620	patient1	post	control
30.76516	patient2	pre	control
32.46562	patient2	post	control
32.53838	patient3	pre	control
35.62194	patient3	post	control
31.10912	patient4	pre	control
31.21611	patient4	post	control
30.47516	patient5	pre	control
30.89179	patient5	post	control
29.57050	patient6	pre	drug
30.64656	patient6	post	drug
34.91339	patient7	pre	drug
31.69578	patient7	post	drug
26.97193	patient8	pre	drug
27.11820	patient8	post	drug
28.50800	patient9	pre	drug
26.26490	patient9	post	drug
28.76724	patient10	pre	drug
34.15198	patient10	post	drug

$eatment + Pre Ex$

patient	treatment	pre	post	diff
patient1	control	27.13863	28.83620	1.6975655
patient10	drug	28.76724	34.15198	5.3847320
patient2	control	30.76516	32.46562	1.7004617
patient3	control	32.53838	35.62194	3.0835634
patient4	control	31.10912	31.21611	0.1069878
patient5	control	30.47516	30.89179	0.4166333
patient6	drug	29.57050	30.64656	1.0760649
patient7	drug	34.91339	31.69578	-3.2176176
patient8	drug	26.97193	27.11820	0.1462705
patient9	drug	28.50800	26.26490	-2.2431038

TIMTOADY
(There's more than one way to do it)

- $EExxppprreessssiioonn \sim TTrreeaattmmeennttt$
- Using Gain of score
  - Compute the difference between post and pre expression as a response variable (Diff Expression), and use treatment as a independent variable

$pression \sim Treatment$

expression	patient	time	treatment
27.13863	patient1	pre	control
28.83620	patient1	post	control
30.76516	patient2	pre	control
32.46562	patient2	post	control
32.53838	patient3	pre	control
35.62194	patient3	post	control
31.10912	patient4	pre	control
31.21611	patient4	post	control
30.47516	patient5	pre	control
30.89179	patient5	post	control
29.57050	patient6	pre	drug
30.64656	patient6	post	drug
34.91339	patient7	pre	drug
31.69578	patient7	post	drug
26.97193	patient8	pre	drug
27.11820	patient8	post	drug
28.50800	patient9	pre	drug
26.26490	patient9	post	drug
28.76724	patient10	pre	drug
34.15198	patient10	post	drug

$eatment + Pre Ex$

patient	treatment	pre	post	diff
patient1	control	27.13863	28.83620	1.6975655
patient10	drug	28.76724	34.15198	5.3847320
patient2	control	30.76516	32.46562	1.7004617
patient3	control	32.53838	35.62194	3.0835634
patient4	control	31.10912	31.21611	0.1069878
patient5	control	30.47516	30.89179	0.4166333
patient6	drug	29.57050	30.64656	1.0760649
patient7	drug	34.91339	31.69578	-3.2176176
patient8	drug	26.97193	27.11820	0.1462705
patient9	drug	28.50800	26.26490	-2.2431038

TIMTOADY
(There's more than one way to do it)

- $EExxppprreessssiioonn \sim TTrreeaattmmeennttt$
- Using Gain of score
  - Compute the difference between post and pre expression as a response variable (Diff Expression), and use treatment as a independent variable

$pression \sim Treatment$

expression	patient	time	treatment
27.13863	patient1	pre	control
28.83620	patient1	post	control
30.76516	patient2	pre	control
32.46562	patient2	post	control
32.53838	patient3	pre	control
35.62194	patient3	post	control
31.10912	patient4	pre	control
31.21611	patient4	post	control
30.47516	patient5	pre	control
30.89179	patient5	post	control
29.57050	patient6	pre	drug
30.64656	patient6	post	drug
34.91339	patient7	pre	drug
31.69578	patient7	post	drug
26.97193	patient8	pre	drug
27.11820	patient8	post	drug
28.50800	patient9	pre	drug
26.26490	patient9	post	drug
28.76724	patient10	pre	drug
34.15198	patient10	post	drug

$eatment + Pre Ex$

patient	treatment	pre	post	diff
patient1	control	27.13863	28.83620	1.6975655
patient10	drug	28.76724	34.15198	5.3847320
patient2	control	30.76516	32.46562	1.7004617
patient3	control	32.53838	35.62194	3.0835634
patient4	control	31.10912	31.21611	0.1069878
patient5	control	30.47516	30.89179	0.4166333
patient6	drug	29.57050	30.64656	1.0760649
patient7	drug	34.91339	31.69578	-3.2176176
patient8	drug	26.97193	27.11820	0.1462705
patient9	drug	28.50800	26.26490	-2.2431038

TIMTOADY
(There's more than one way to do it)

- $EExxppprreessssiioonn \sim TTrreeaattmmeennttt$
- Using Gain of score
  - Compute the difference between post and pre expression as a response variable (Diff Expression) and use treatment as a independent variable

$$Diff\ Expression \sim Treatment$$

$$ost\ Expression \sim Treatment + Pre\ Expression\ Covariate$$

TIMTOADY
(There's more than one way to do it)

- $EExxppprreessssiioonn \sim TTrreeaattmmeennttt$
- Using Gain of score
  - Compute the difference between post and pre expression as a response variable (Diff Expression) and use treatment as a independent variable

$$Diff\ Expression \sim Treatment$$

$$ost\ Expression \sim Treatment + Pre\ Expression\ Covariate$$

TIMTOADY
(There's more than one way to do it)

- $EExxppprreessssiioonn \sim TTrreeaattmmeennttt$
- Using Gain of score
  - Compute the difference between post and pre expression as a response variable (Diff Expression) and use treatment as a independent variable
- Using ANCOVA

$$ost\ Expression \sim Treatment + Pre\ Expression\ Covariate$$

TIMTOADY
(There's more than one way to do it)

- $EExxppprreessssiioonn \sim TTrreeaattmmeennttt$
- Using Gain of score
  - Compute the difference between post and pre expression as a response variable (Diff Expression) and use treatment as a independent variable
- Using ANCOVA
  - Use post as a response variable (Y) and fit the linear/general linear model using factors and also pre as a covariate

$$ost\ Expression \sim Treatment + Pre\ Expression\ Covariate$$

TIMTOADY
(There's more than one way to do it)

- $EExxppprreessssiioonn \sim TTrreeaattmmeennttt+PPrree\ EExxppprreessssiioonn\ CCoovvaarriiaattee$
- $EExxppprreessssiioonn \sim TTrreeaattmmeennttt$
- Using Gain of score
  - Compute the difference between post and pre expression as a response variable (Diff Expression) and use treatment as a independent variable
- Using ANCOVA
  - Use post as a response variable (Y) and fit the linear/general linear model using factors and also pre as a covariate

$$Post\ Expression \sim Treatment + Pre\ Expression\ Covariate$$

$$ost\ Expression \sim Treatment + Pre\ Expression\ Covariate$$

TIMTOADY
(There's more than one way to do it)

- $EExxppprreessssiioonn \sim TTrreeaattmmeennttt+PPrree\ EExxppprreessssiioonn\ CCoovvaarriiaattee$
- $EExxppprreessssiioonn \sim TTrreeaattmmeennttt$
- Using Gain of score
  - Compute the difference between post and pre expression as a response variable (Diff Expression) and use treatment as a independent variable
- Using ANCOVA
  - Use post as a response variable (Y) and fit the linear/general linear model using factors and a

$Post\ Expression \sim ession\ Covariate$

$ost\ Expression \sim ession\ Covariate$

patient	treatment	pre	post
patient1	control	27.13863	28.83620
patient10	drug	28.76724	34.15198
patient2	control	30.76516	32.46562
patient3	control	32.53838	35.62194
patient4	control	31.10912	31.21611
patient5	control	30.47516	30.89179
patient6	drug	29.57050	30.64656
patient7	drug	34.91339	31.69578
patient8	drug	26.97193	27.11820
patient9	drug	28.50800	26.26490

TIMTOADY
(There's more than one way to do it)

- $EExxppprreessssiioonn \sim TTrreeaattmmeennttt+PPrree\ EExxppprreessssiioonn\ CCoovvaarriiaattee$
- $EExxppprreessssiioonn \sim TTrreeaattmmeennttt$
- Using Gain of score
  - Compute the difference between post and pre expression as a response variable (Diff Expression) and use treatment as a independent variable
- Using ANCOVA
  - Use post as a response variable (Y) and fit the linear/general linear model using factors and a

$Post\ Expression \sim$ [illegible] $ession\ Covariate$

$ost\ Expression \sim$ [illegible] $ession\ Covariate$

patient	treatment	pre	post
patient1	control	27.13863	28.83620
patient10	drug	28.76724	34.15198
patient2	control	30.76516	32.46562
patient3	control	32.53838	35.62194
patient4	control	31.10912	31.21611
patient5	control	30.47516	30.89179
patient6	drug	29.57050	30.64656
patient7	drug	34.91339	31.69578
patient8	drug	26.97193	27.11820
patient9	drug	28.50800	26.26490

In the real world...

We have assumed that

We have assumed that

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done.

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done.

data generation problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done.
2. Expression measurement follows normal distribution.

data generation problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done.
2. Expression measurement follows normal distribution.
3. We have a large size of experimental units(+100 samples).

data generation problem

non-normality problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem
4. Normalization is taken care of.

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem
4. Normalization is taken care of. — normalization problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem
4. Normalization is taken care of. — normalization problem
5. Expression measurements are collected at gene level.

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem
4. Normalization is taken care of. — normalization problem
5. Expression measurements are collected at gene level. — Isoform problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done.

data generation problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done.

data generation problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done.

   ➔ Gee. Tophat takes hours and hours

data generation problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done.

data generation problem

➔ Gee. Tophat takes hours and hours

Remedy:

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done.

data generation problem

➔ Gee. Tophat takes hours and hours

Remedy:

1) Use STAR aligner instead. It should take 4 to 8 minutes.

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done.

data generation problem

➔ Gee. Tophat takes hours and hours

Remedy:

1) Use STAR aligner instead. It should take 4 to 8 minutes.

2) Utilize multithreads.

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done.

data generation problem

➔ Gee. Tophat takes hours and hours

Remedy:

1) Use STAR aligner instead. It should take 4 to 8 minutes.

2) Utilize multithreads.

Ex) What could takes 30 minutes can be done in 30 seconds

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done.
2. Expression measurement follows normal distribution.

data generation problem

non-normality problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done.
2. Expression measurement follows normal distribution.

data generation problem

non-normality problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem

➔ “Not really. RNAseq is usually measured in counts”

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem

➔ “Not really. RNAseq is usually measured in counts”

Remedy:

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem

➔ “Not really. RNAseq is usually measured in counts”

Remedy:

1) Assume negative binomial and fit generalized linear model

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem

➔ “Not really. RNAseq is usually measured in counts”

Remedy:

1) Assume negative binomial and fit generalized linear model

2) Transform your counts to log and fit linear model

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem

   ➔ Not really. RNAseq sample size is usually 2~ 3 per group

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem

   ➔ Not really. RNAseq sample size is usually 2~ 3 per group

   Remedy:

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem

➔ Not really. RNAseq sample size is usually 2~ 3 per group

Remedy:

Use some sort of shrinkage method to stabilize variance (borrowing information of all genes)

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem

➔ Not really. RNAseq sample size is usually 2~ 3 per group

Remedy:

Use some sort of shrinkage method to stabilize variance (borrowing information of all genes)

Use edgeR, DESeq, Voom

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem
4. Normalization is taken care of. — normalization problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem
4. Normalization is taken care of. — normalization problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem
4. Normalization is taken care of. — normalization problem

Remedy:

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem
4. Normalization is taken care of. — normalization problem

Remedy:

There are many normalization schemes. Choose one.

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem
4. Normalization is taken care of. — normalization problem
5. Expression measurements are collected at gene level. — Isoform problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem
4. Normalization is taken care of. — normalization problem
5. Expression measurements are collected at gene level. — Isoform problem

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem
4. Normalization is taken care of. — normalization problem
5. Expression measurements are collected at gene level. — Isoform problem

Remedy:

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem
4. Normalization is taken care of. — normalization problem
5. Expression measurements are collected at gene level. — Isoform problem

Remedy:

1) Use Cufflinks, RSEM, Stringtie, etc.

We have assumed that

1. All heavy biological & computational works(sequencing, preprocessing, alignment, and counting) have been done. — data generation problem
2. Expression measurement follows normal distribution. — non-normality problem
3. We have a large size of experimental units(+100 samples). — small replicate size problem
4. Normalization is taken care of. — normalization problem
5. Expression measurements are collected at gene level. — Isoform problem

Remedy:

1) Use Cufflinks, RSEM, Stringtie, etc.
2) Exon level counting

Big Picture

Prof. Friedman’s Talk

Prof. Peter Sims’ Talks

Alexander Lachmann’s Talk

Albert Lee’s Talk

Prof. Friedman’s Talk

Prof. Peter Sims’ Talks

Alexander Lachmann’s Talk

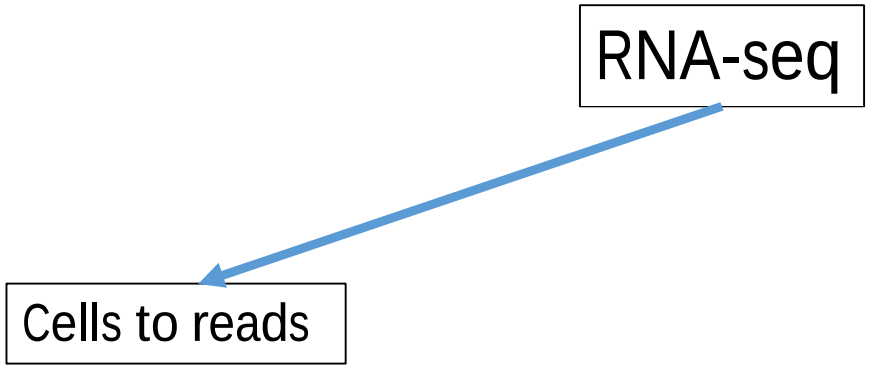
Albert Lee’s Talk

RNA-seq

Prof. Peter Sims' Talks

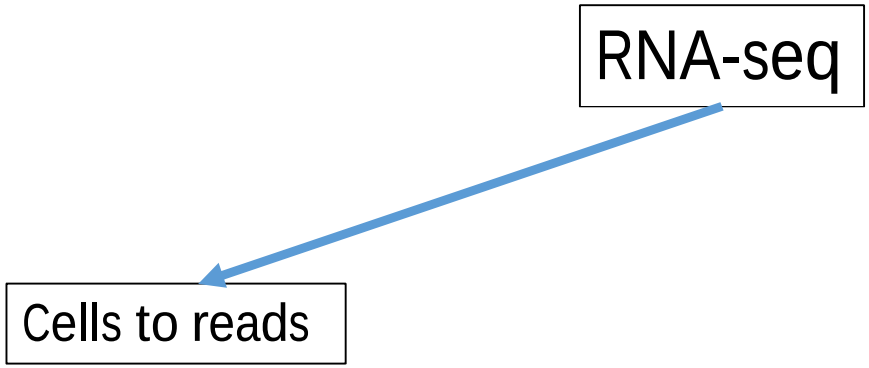
Alexander Lachmann's Talk

Albert Lee's Talk



Alexander Lachmann's Talk

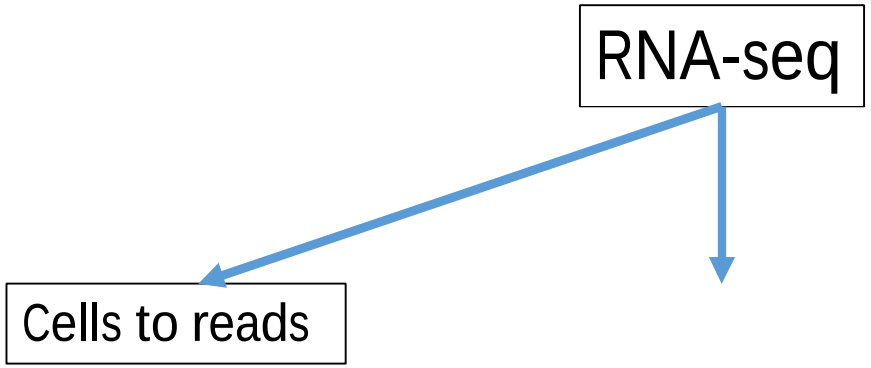
Albert Lee's Talk



Alexander Lachmann's Talk

"How do you efficiently capture the high quality biological data with little cost yet high resolution?"

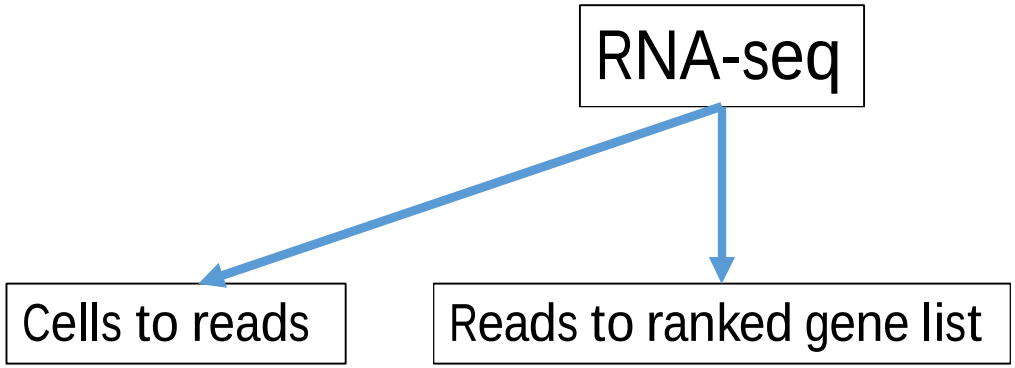
Albert Lee's Talk



Alexander Lachmann's Talk

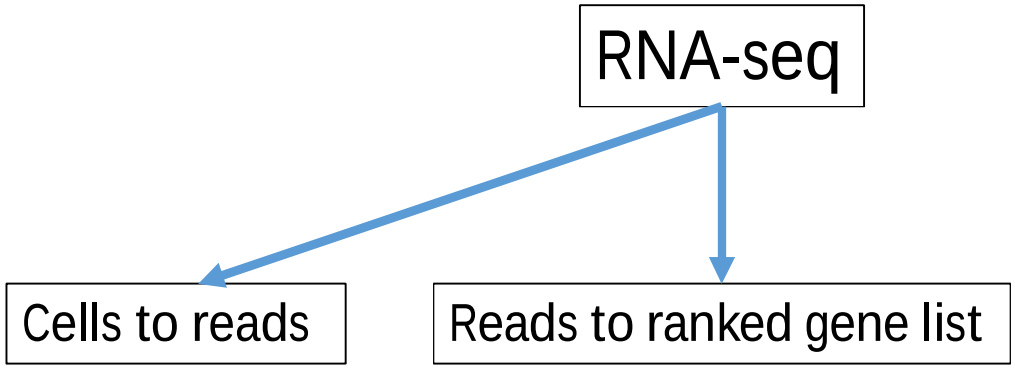
"How do you efficiently capture the high quality biological data with little cost yet high resolution?"

Albert Lee's Talk



Alexander Lachmann's Talk

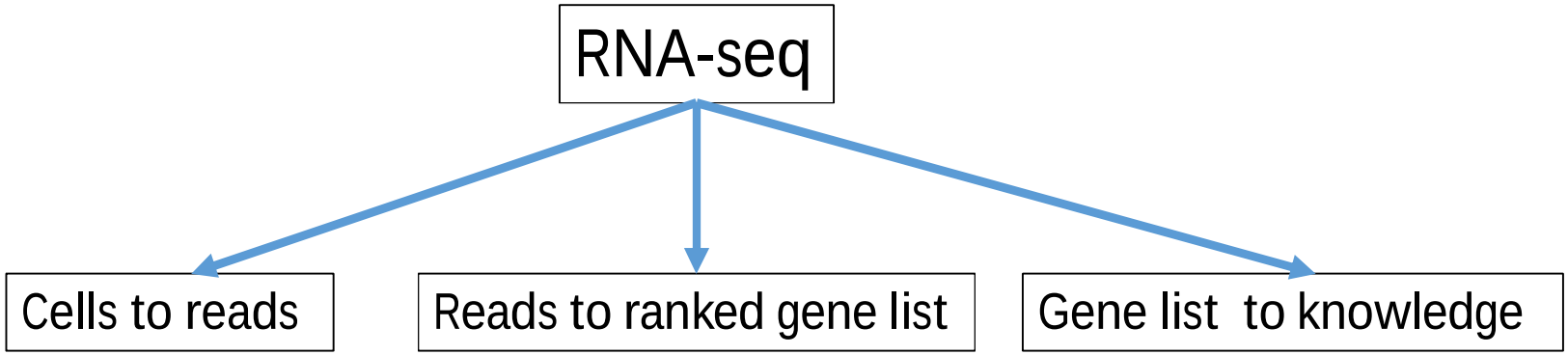
"How do you efficiently capture the high quality biological data with little cost yet high resolution?"



Alexander Lachmann's Talk

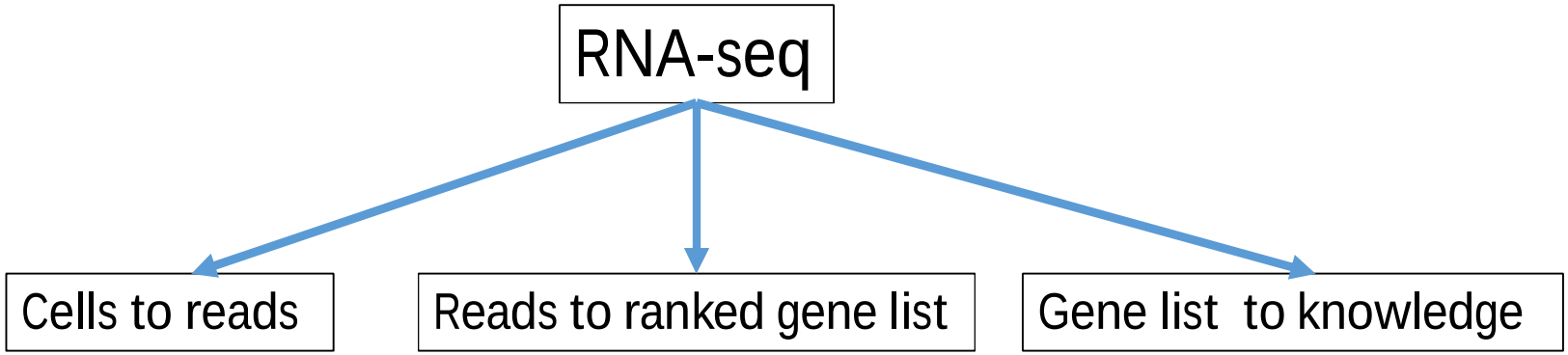
"How do you efficiently capture the high quality biological data with little cost yet high resolution?"

"What is the most interesting subset of genes within our experiment design?"



"How do you efficiently capture the high quality biological data with little cost yet high resolution?"

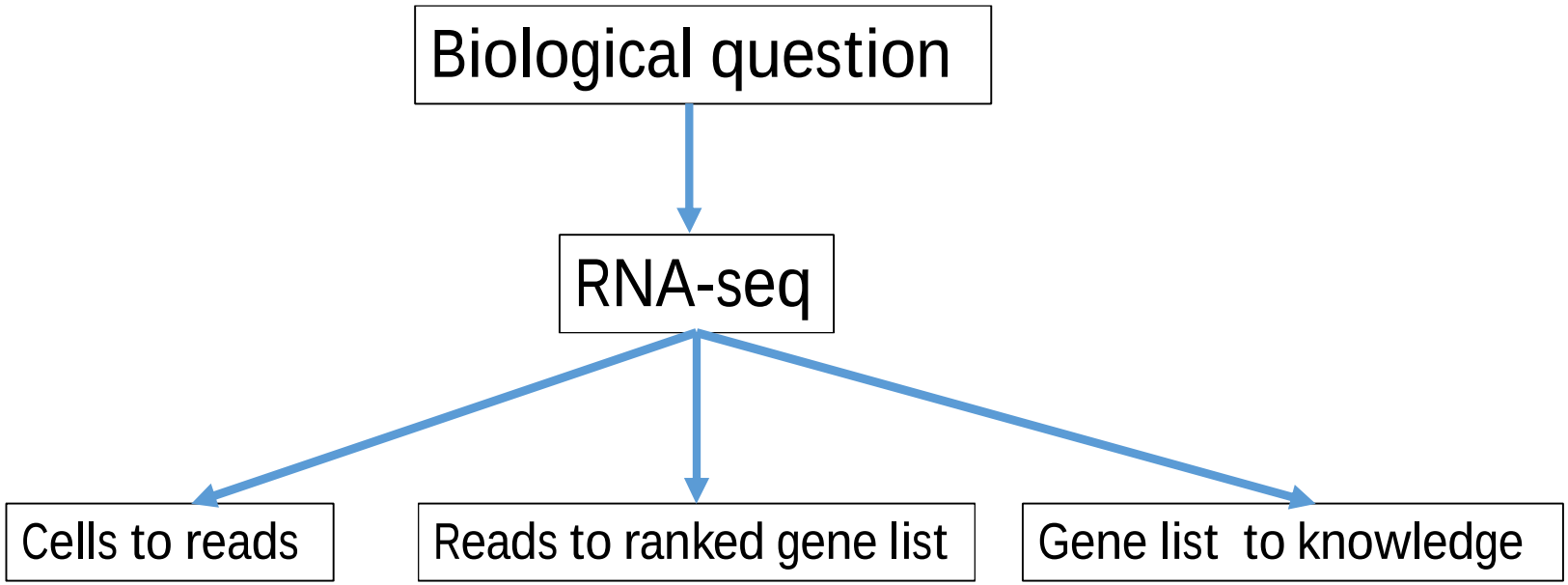
"What is the most interesting subset of genes within our experiment design?"

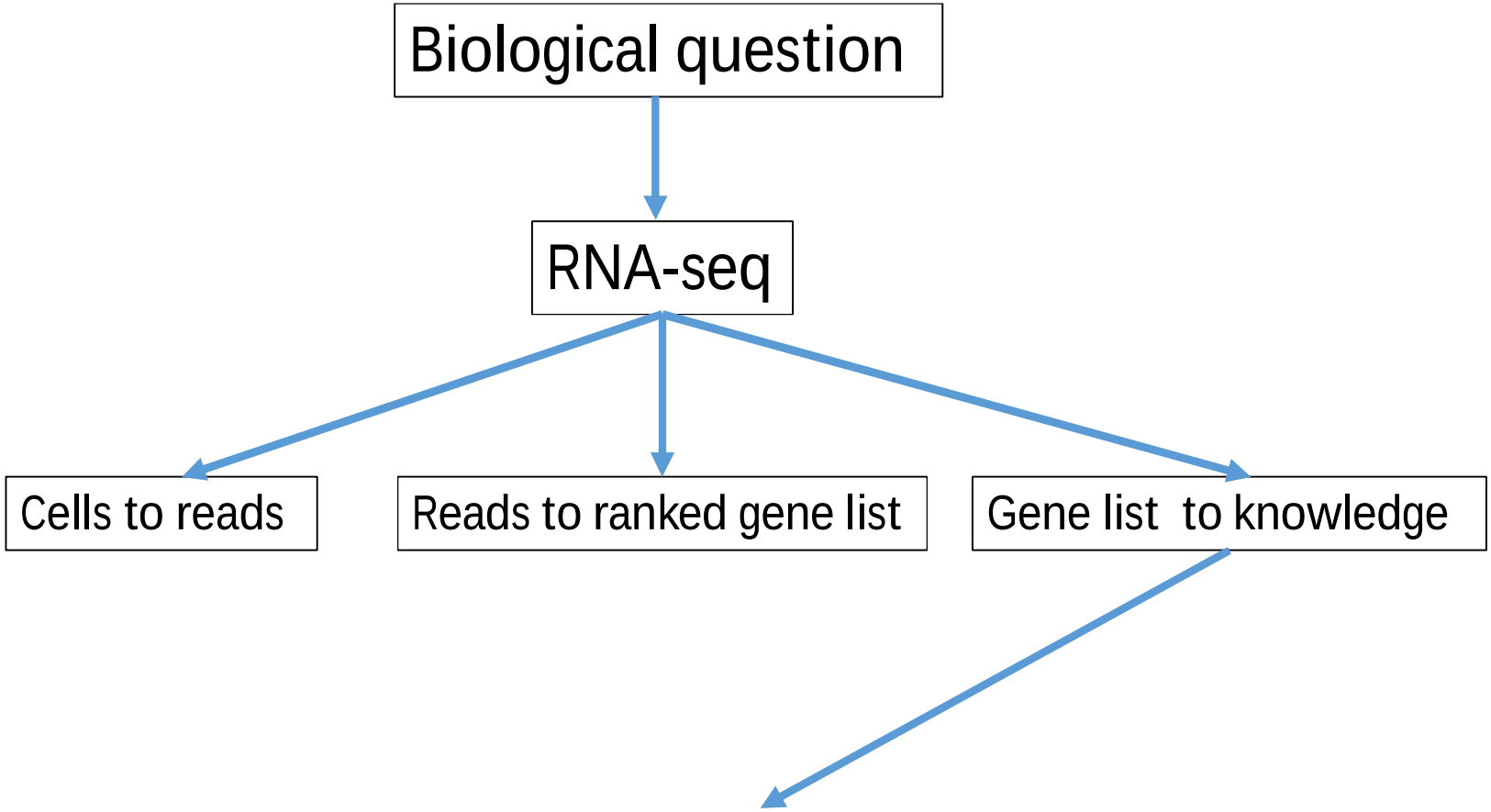


"How do you efficiently capture the high quality biological data with little cost yet high resolution?"

"What is the most interesting subset of genes within our experiment design?"

"What's the biological meaning of those genes that are affected?"


Biological question
RNA-seq
Cells to reads
Reads to ranked gene list
Gene list to knowledge


Biological question
RNA-seq
Cells to reads
Reads to ranked gene list
Gene list to knowledge

Biological question

RNA-seq

Cells to reads

Reads to ranked gene list

Gene list to knowledge

Cell

nature

NEJM

On that note….

On that note….

Sir Ronald Aylmer Fisher
Father of “Design of Experiments”

On that note….

Sir Ronald Aylmer Fisher
Father of “Design of Experiments”

“The statistician cannot evade the responsibility for understanding the process he applies or recommends”

THE END