Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

BEFORE WE START…

EMMA BEAUXIS-AUSSALET

e.m.a.l.beauxis@hva.nl

MISSING VALUES

- **Delete** data points with missing values.
- **Replace** missing values (imputation) with default values (mean, median…).
- Include missing values in the regression model, and **learn** from them.

  Data point with missing value may be a special case, a special kind of phenomena.

  (e.g., make 2 categories, for data points with or without for missing values)

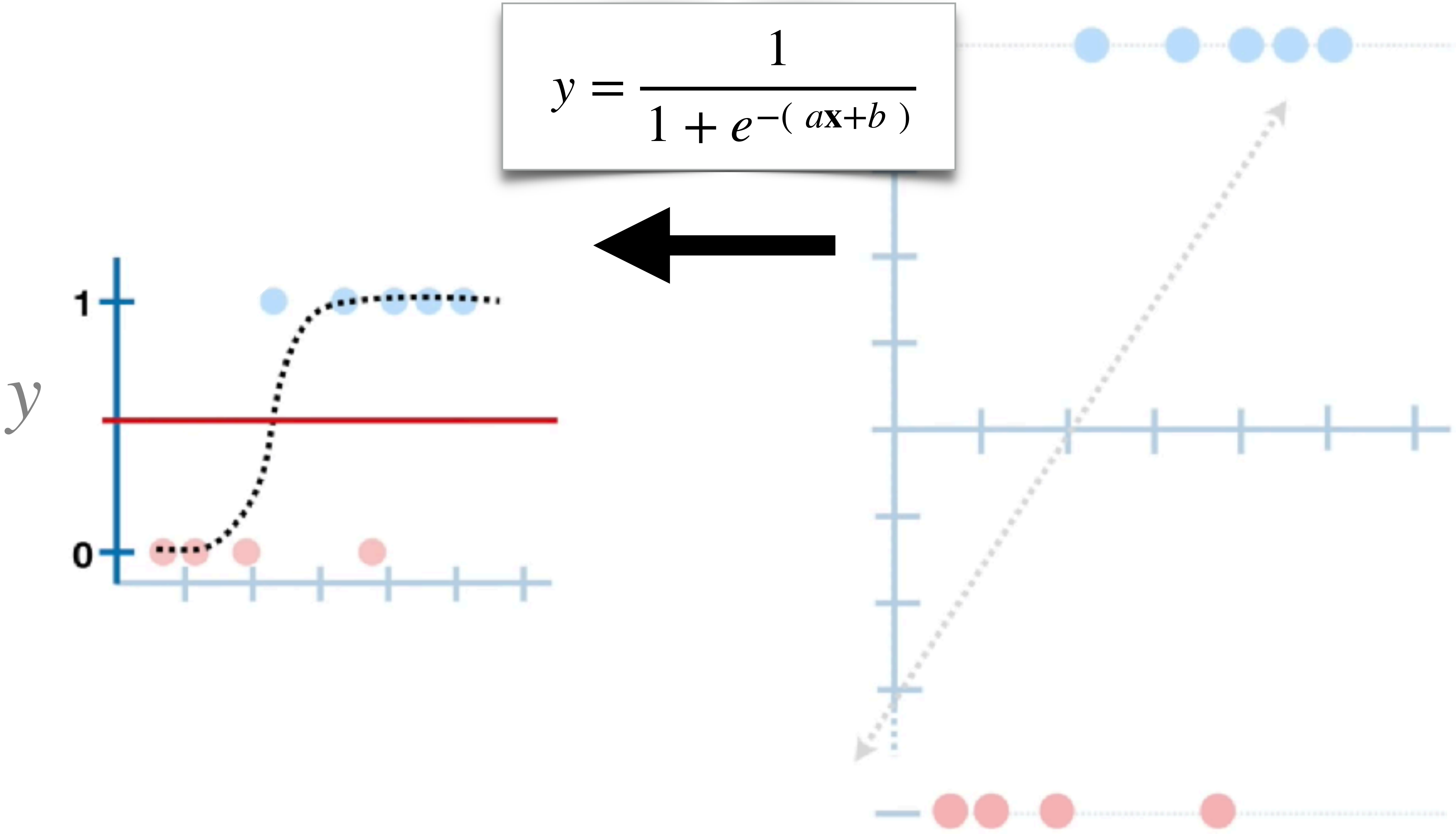
Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

LOGISTIC REGRESSION, A RECAP

EMMA BEAUXIS-AUSSALET

e.m.a.l.beauxis@hva.nl

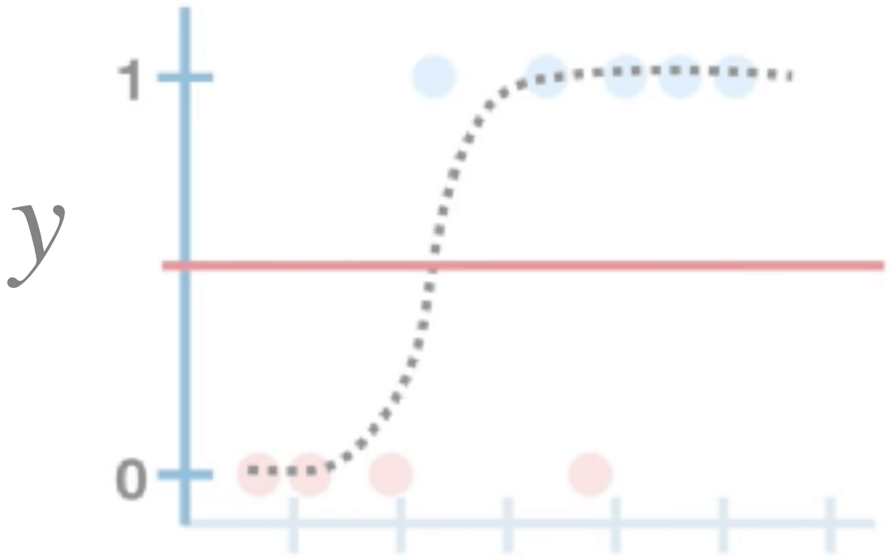
LINEAR TO SIGMOID

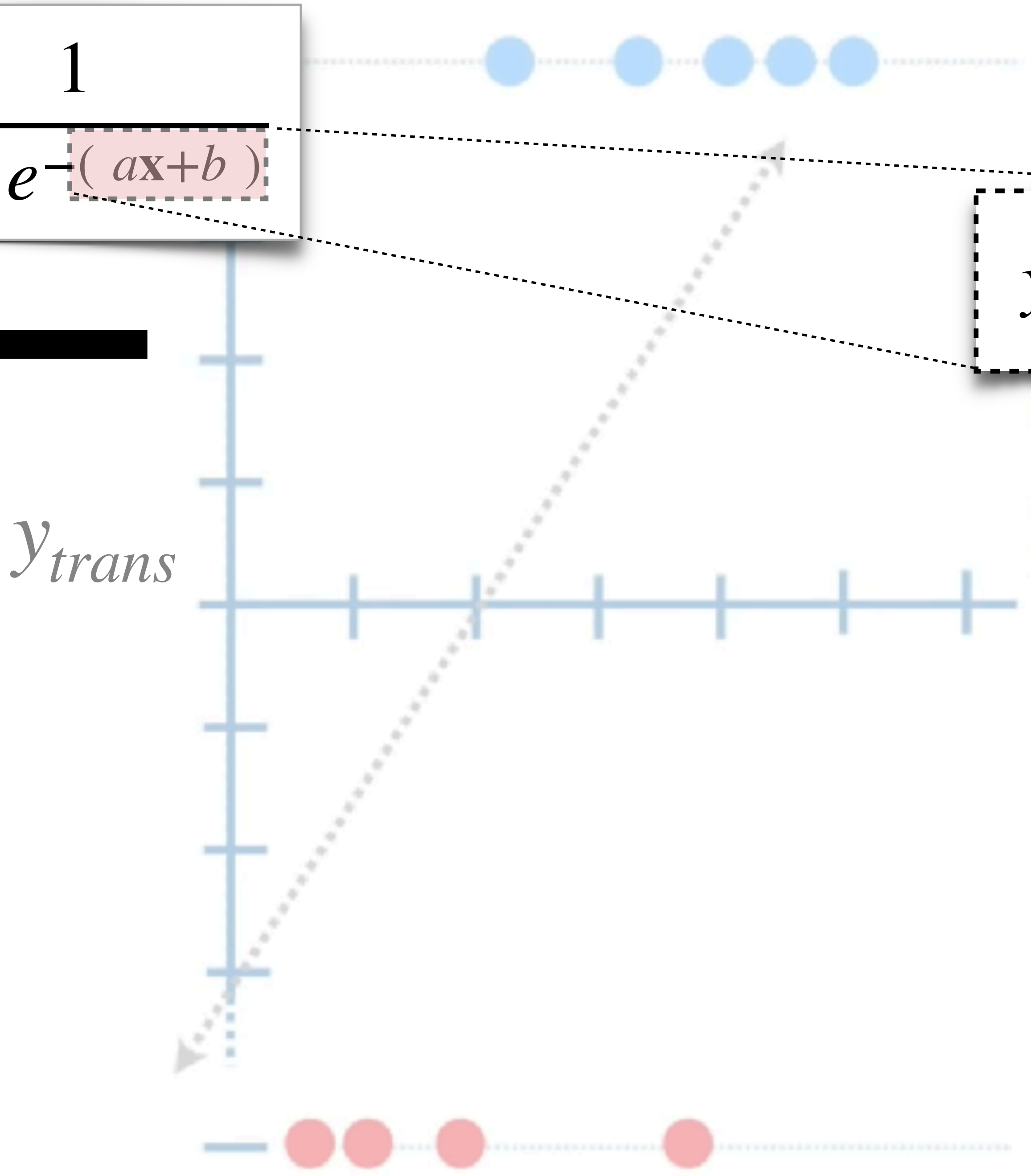


$$y = \frac{1}{1 + e^{-(\,a\mathbf{x}+b\,)}}$$

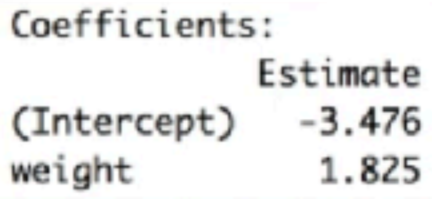
LINEAR TO SIGMOID

$$y_{trans} = \frac{1}{1 + e^{-\boxed{(a\mathbf{x}+b)}}}$$





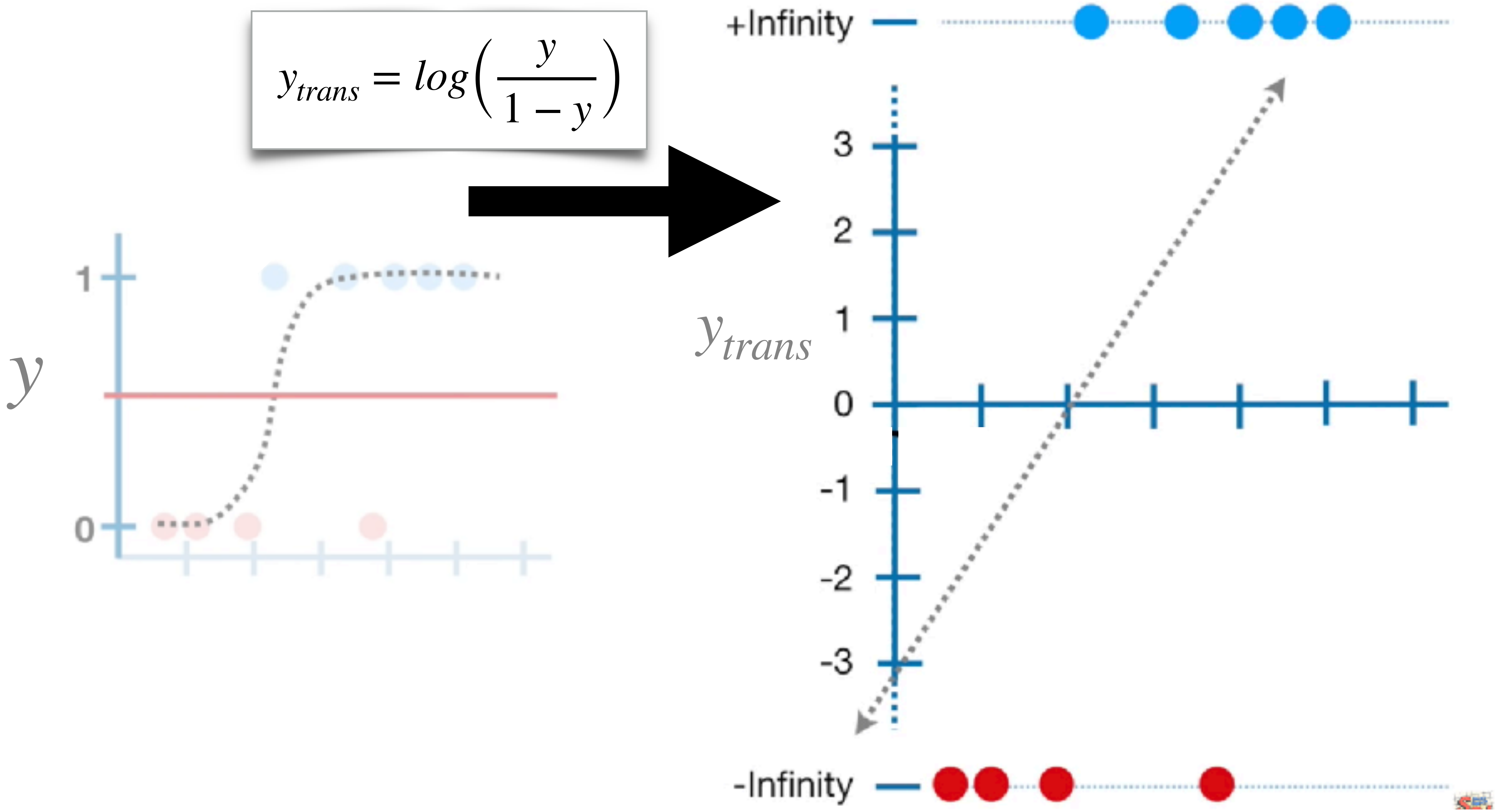
$$\boxed{y = a\mathbf{x} + b}$$



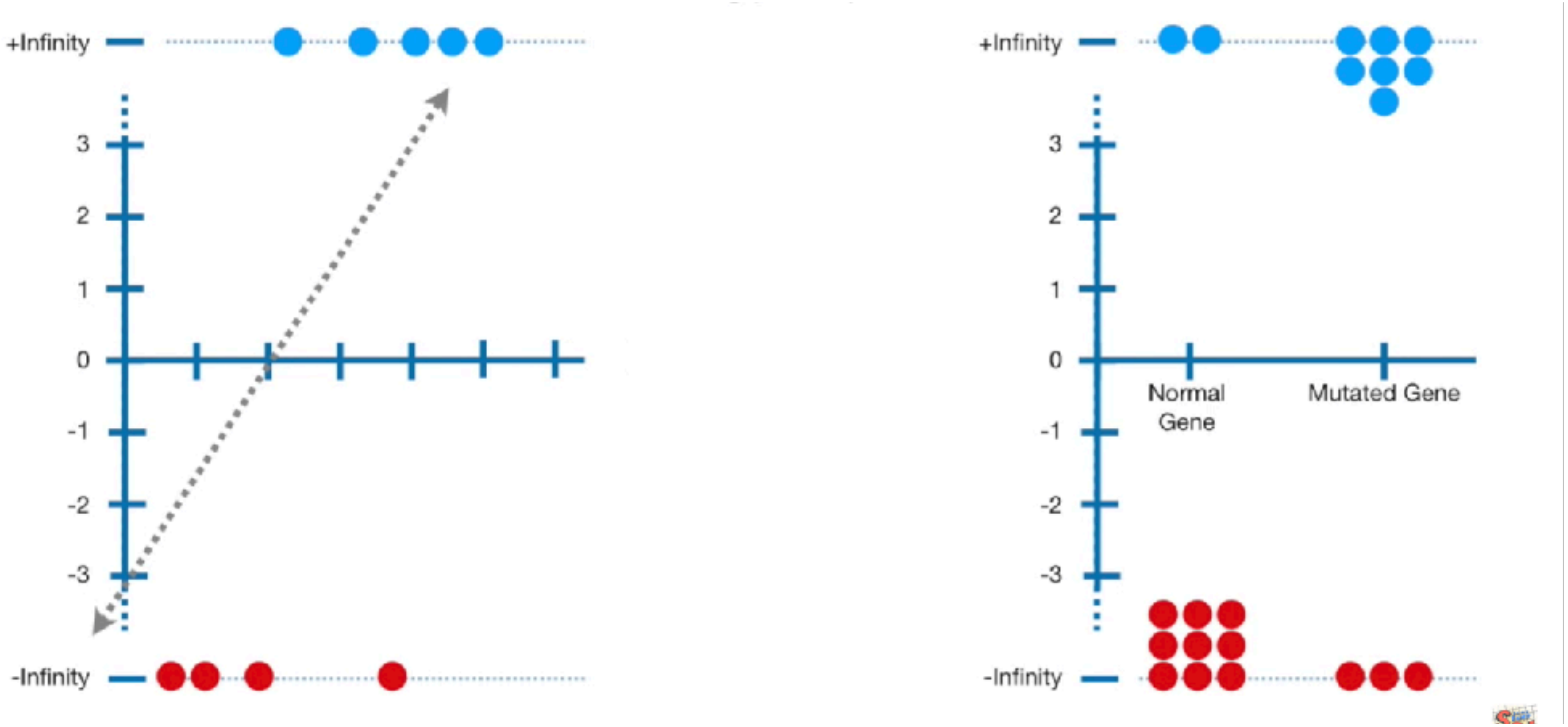
```
Coefficients:
            Estimate
(Intercept)   -3.476
weight         1.825
```

SIGMOID TO LINEAR

$$y_{trans} = log\left(\frac{y}{1-y}\right)$$

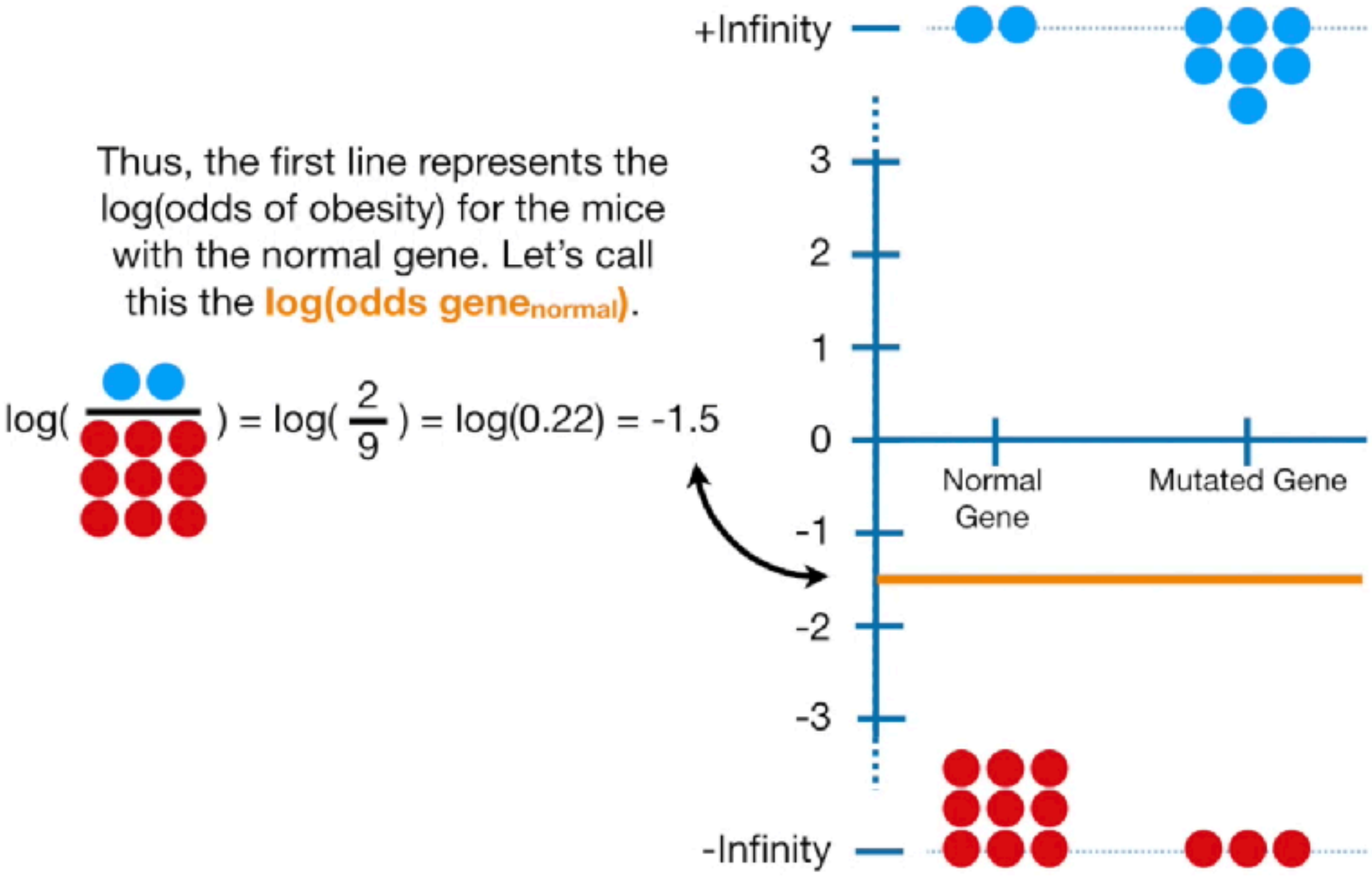


...AND CATEGORICAL DATA ?



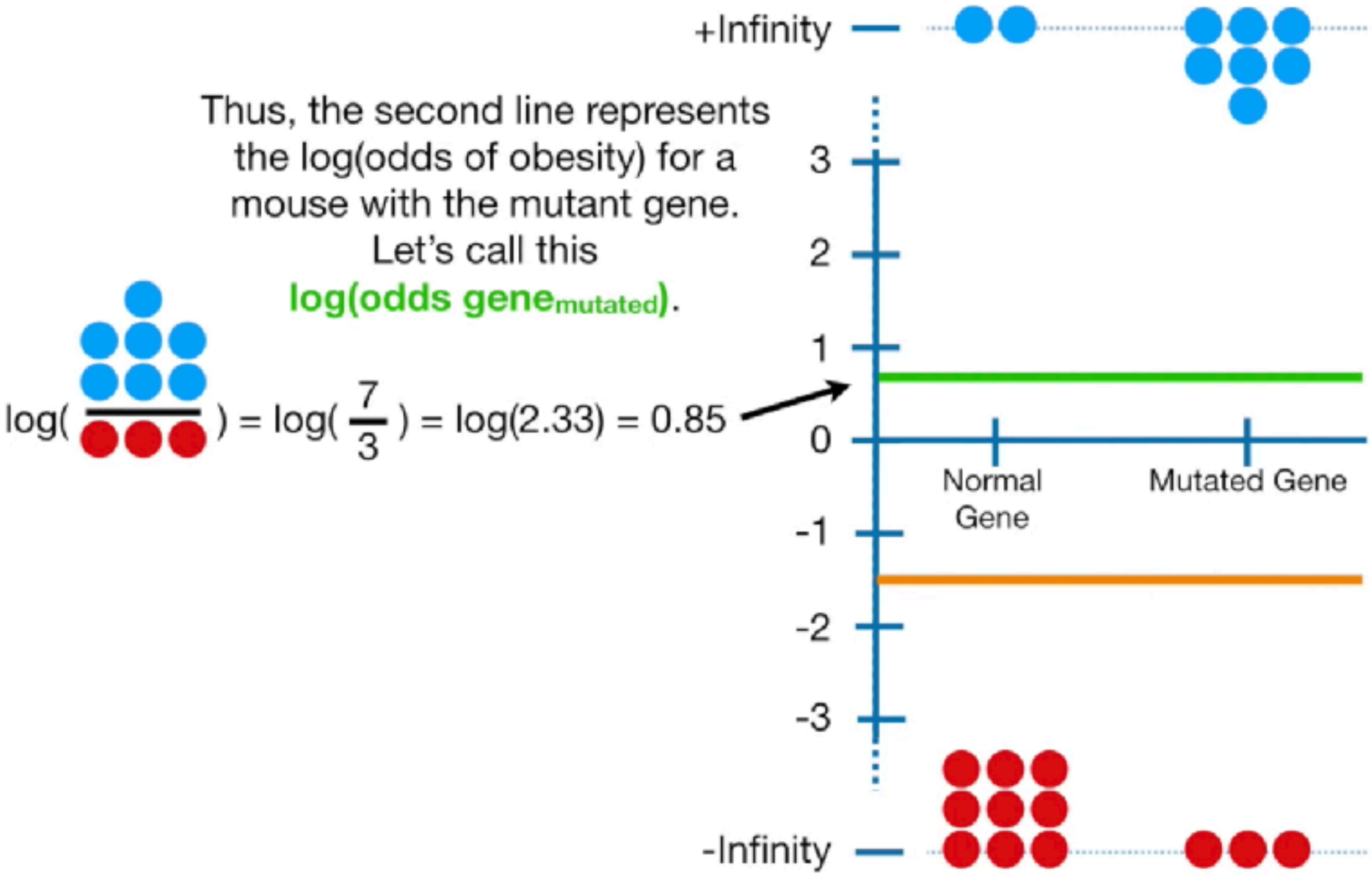
https://www.youtube.com/watch?v=vN5cNN2-HWE

…AND CATEGORICAL DATA ?



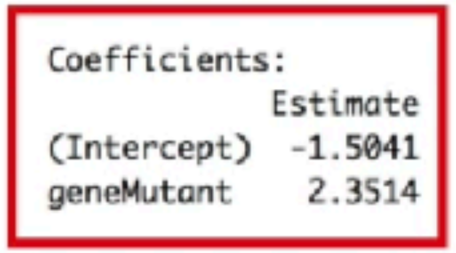
Thus, the first line represents the log(odds of obesity) for the mice with the normal gene. Let’s call this the **log(odds gene$_{normal}$)**.

…AND CATEGORICAL DATA ?

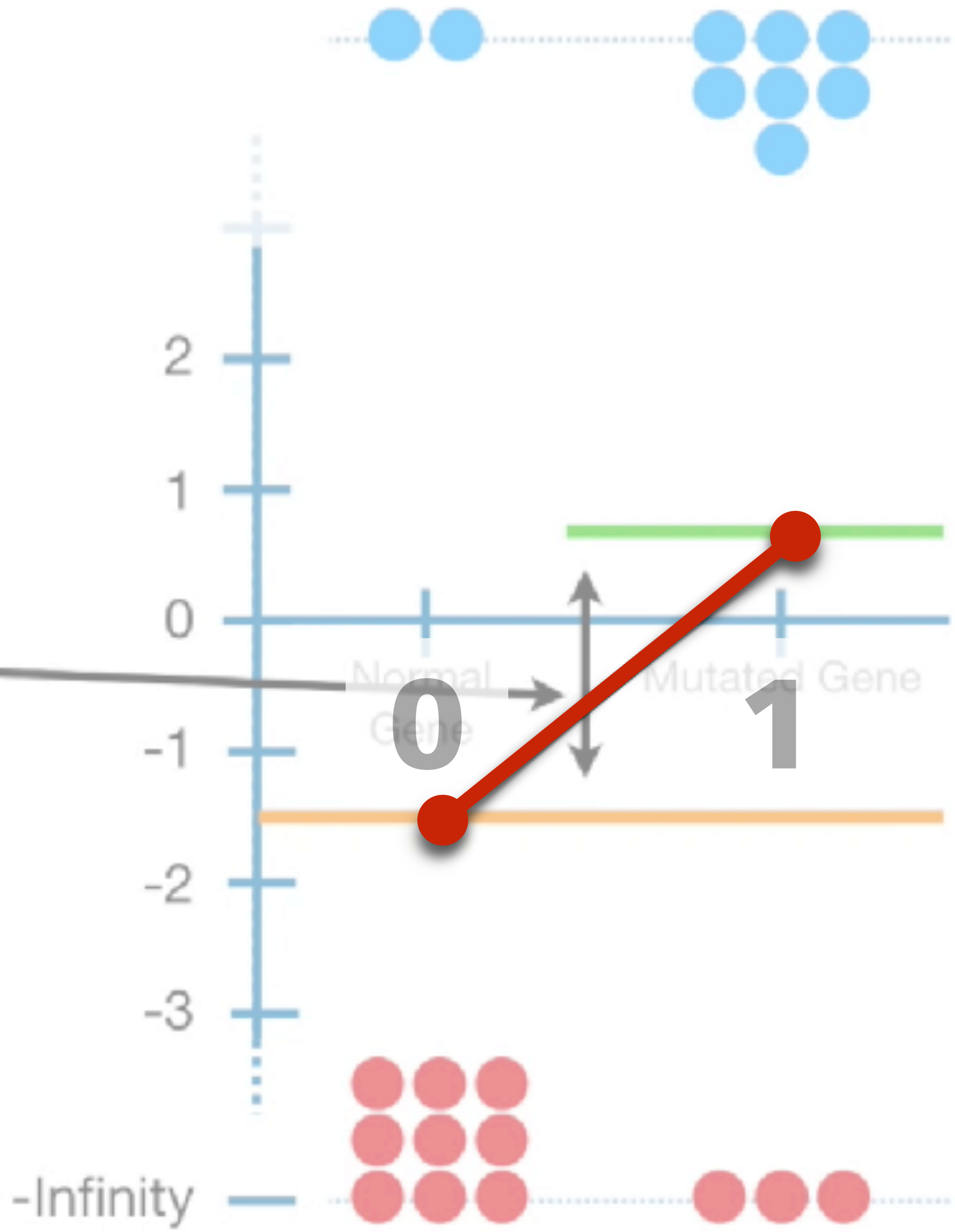

+Infinity
Thus, the second line represents the log(odds of obesity) for a mouse with the mutant gene. Let's call this log(odds gene_mutated).
log(●●●●●●● / ●●●) = log(7/3) = log(2.33) = 0.85
3
2
1
0
-1
-2
-3
Normal Gene
Mutated Gene
-Infinity

...AND CATEGORICAL DATA ?

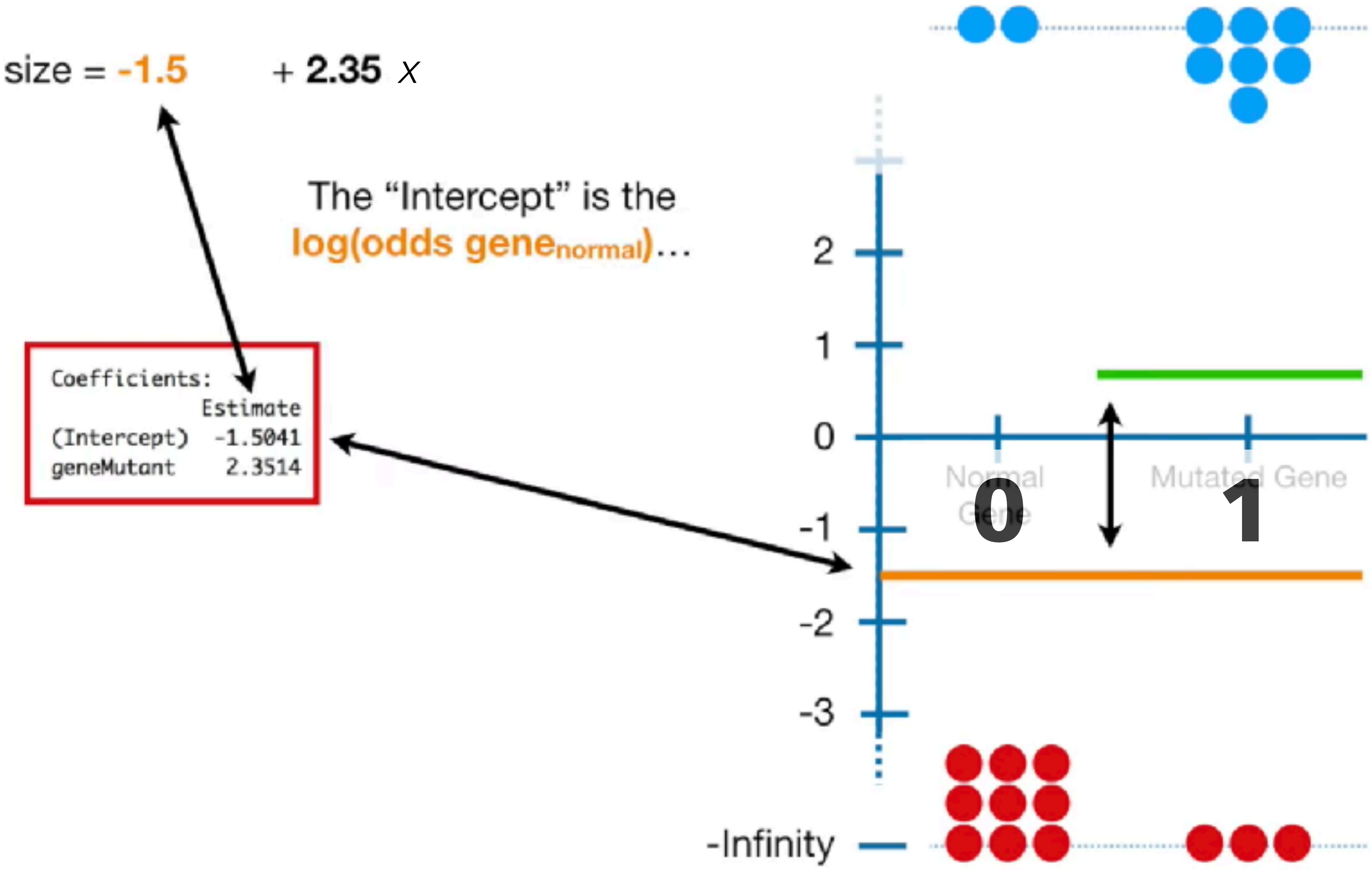
size = -1.5 $+ 2.35\ x$



```
Coefficients:
            Estimate
(Intercept)  -1.5041
geneMutant    2.3514
```



...AND CATEGORICAL DATA ?



size = -1.5 + 2.35 x

The “Intercept” is the log(odds $\text{gene}_{\text{normal}}$)...

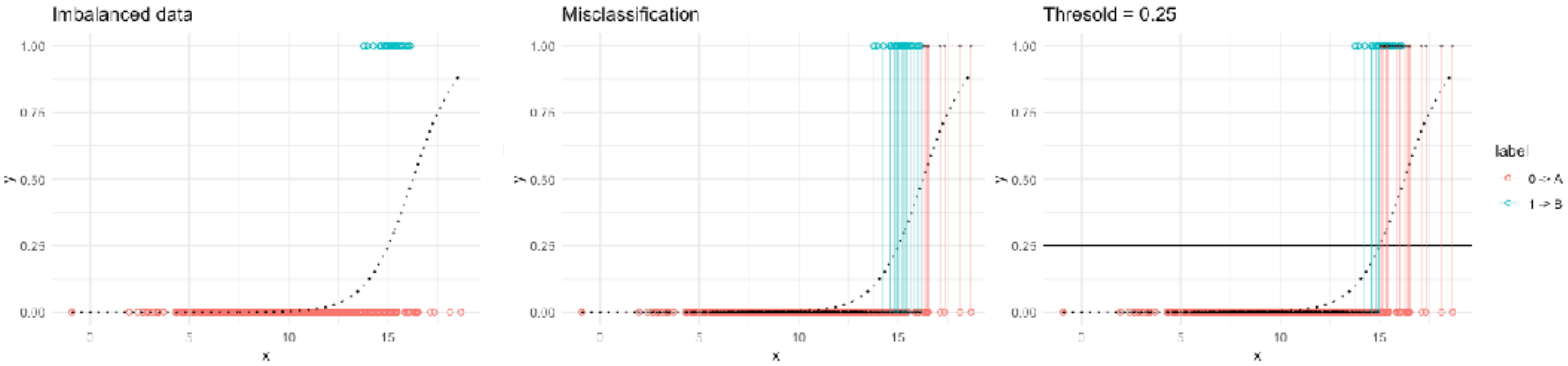
```
Coefficients:
                Estimate
(Intercept)     -1.5041
geneMutant       2.3514
```

UNBALANCED DATA

Highly imbalanced classes (more data point in one class) may yield bias.

Large classes outweigh small classes (e.g., they “drag” the regression line towards them).
In extreme cases, the small class is never predicted.

➡ Change the threshold (not always handy).

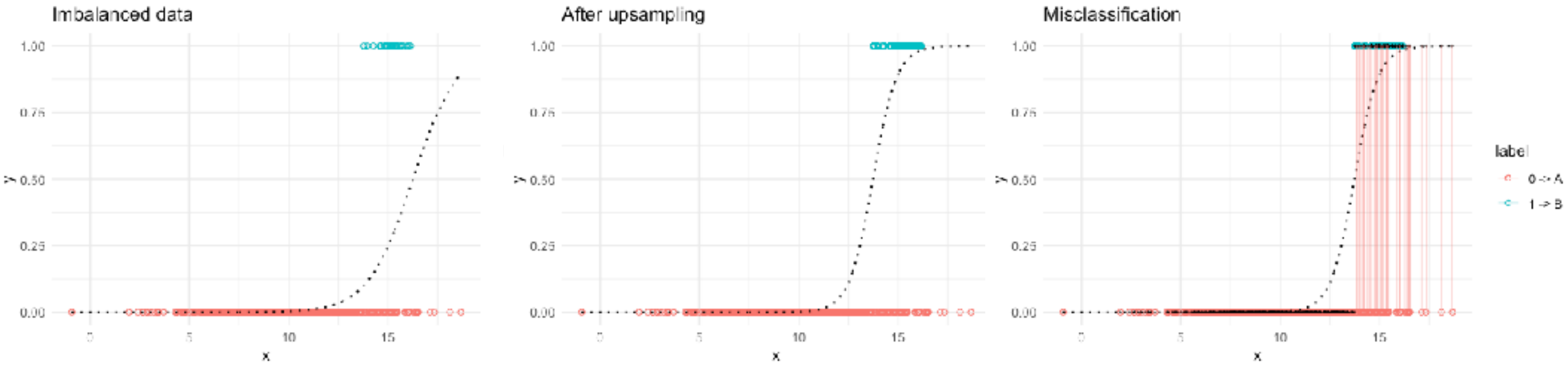


UNBALANCED DATA

Highly imbalanced classes (more data point in one class) may yield bias.

Large classes outweigh small classes (e.g., they “drag” the regression line towards them). In extreme cases, the small class is never predicted.

- Change the threshold (not always handy).
- Do upsampling (will fit the default 0.5 threshold).



UNBALANCED DATA

Highly imbalanced classes (more data point in one class) may yield bias.

Large classes outweigh small classes (e.g., they “drag” the regression line towards them). In extreme cases, the small class is never predicted.

- ➡ Change the threshold (not always handy).
- ➡ Do upsampling (will fit the default 0.5 threshold).

  Use in extreme cases.

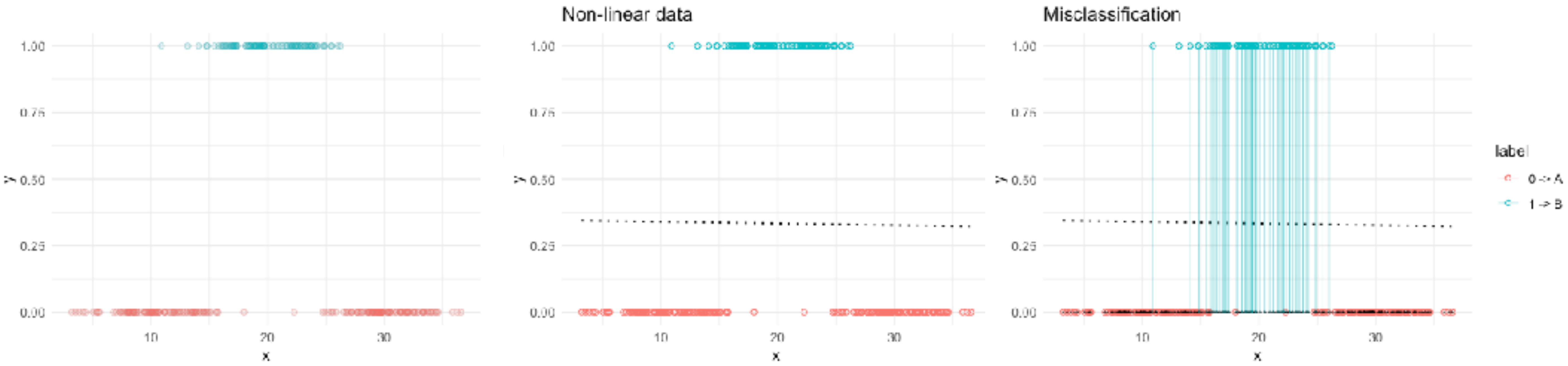
  Use downsampling if you really have large data.

  Be careful with multiclass: upsample differently for each class-specific model. The one-vs-all approach means that the positive class (label 1) may be smaller than all other classes put together.

NONLINEAR PATTERNS

Explanatory variables (predictors) may not have linear relationship with the category to predict.

➡ Transform the data (not always handy or possible).

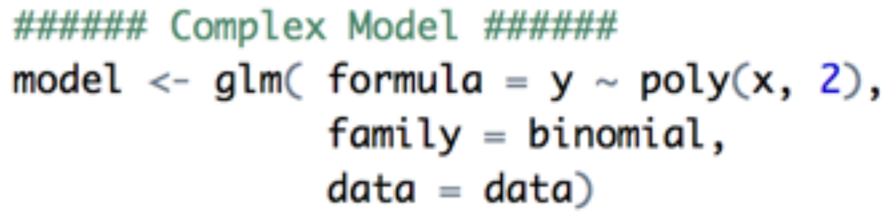

Non-linear data

Misclassification

NONLINEAR PATTERNS

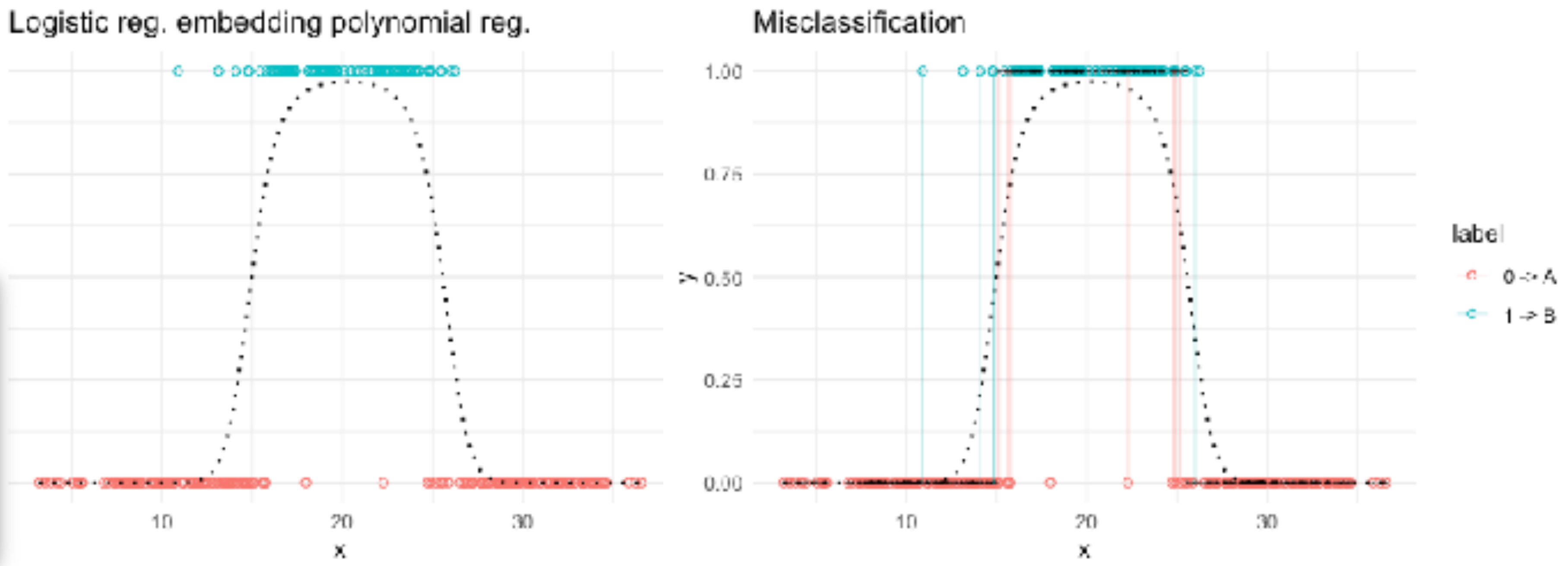
Explanatory variables (predictors) may not have linear relationship with the category to predict.

- ➡ Transform the data (not always handy or possible).
- ➡ Embed polynomial formula into logistic regression (although unusual).



```
###### Complex Model ######
model <- glm( formula = y ~ poly(x, 2),
              family = binomial,
              data = data)
```

$$y = \frac{1}{1 + e^{-(\ a_1\mathbf{x} + a_2\mathbf{x}^2 + b\ )}}$$



NONLINEAR PATTERNS

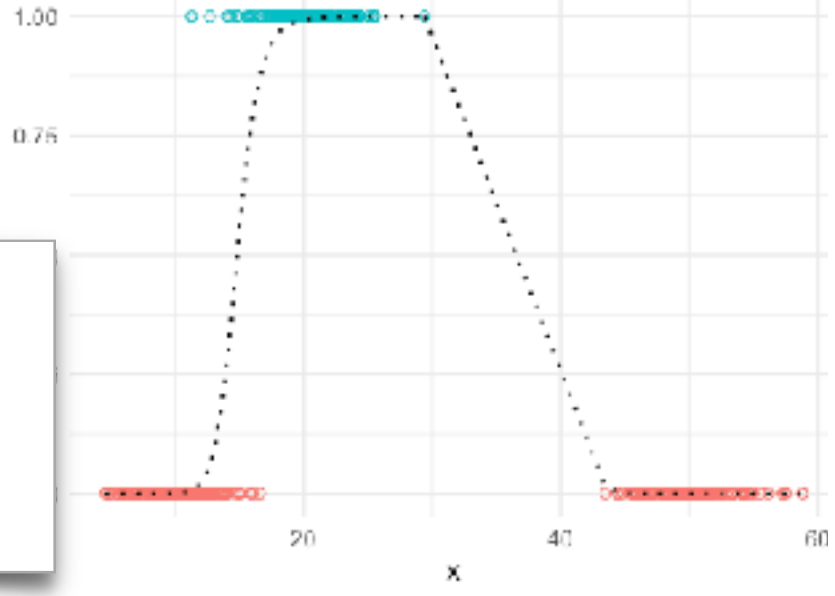
Explanatory variables (predictors) may not have linear relationship with the category to predict.

- ➡ Transform the data (not always handy or possible).
- ➡ Embed polynomial formula into logistic regression (although unusual). Be careful, do not fit too complex polynomials.

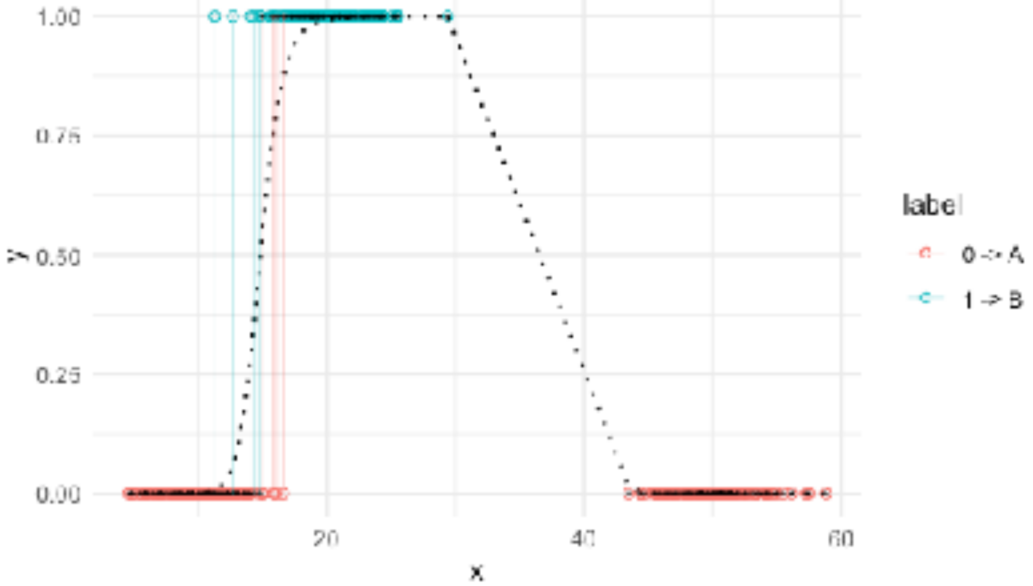
```
###### Complex Model ######
model <- glm( formula = y ~ poly(x, 2),
              family = binomial,
              data = data)
```

$$y = \frac{1}{1 + e^{-(\ a_1\mathbf{x} + a_2\mathbf{x}^2 + b\ )}}$$

Logistic reg. embedding polynomial reg.



Misclassification



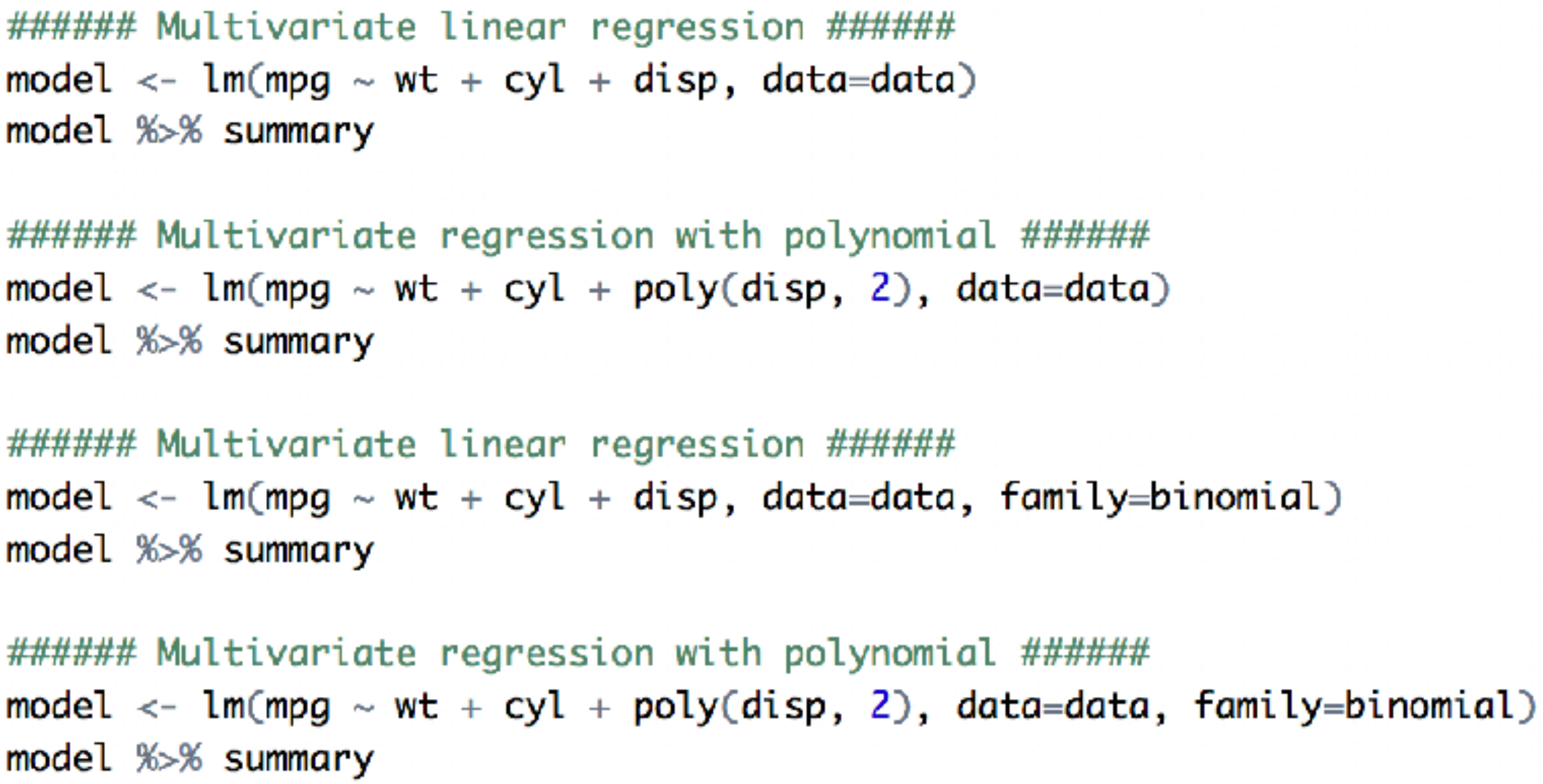
Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

MULTIVARIATE REGRESSION

EMMA BEAUXIS-AUSSALET

e.m.a.l.beauxis@hva.nl

R CODE



```
###### Multivariate linear regression ######
model <- lm(mpg ~ wt + cyl + disp, data=data)
model %>% summary

###### Multivariate regression with polynomial ######
model <- lm(mpg ~ wt + cyl + poly(disp, 2), data=data)
model %>% summary

###### Multivariate linear regression ######
model <- lm(mpg ~ wt + cyl + disp, data=data, family=binomial)
model %>% summary

###### Multivariate regression with polynomial ######
model <- lm(mpg ~ wt + cyl + poly(disp, 2), data=data, family=binomial)
model %>% summary
```

https://www.youtube.com/watch?v=q1RD5ECsSB0

https://www.youtube.com/watch?v=eTZ4VUZHzxw

MORE VARIABLES, WHAT CAN HAPPEN?

Variables with different magnitude (range of values) yield bias, and make fitted parameters & residuals incomparable.

Variables of higher magnitudes outweigh those of lower magnitudes.
(e.g., they "drag" the regression boundary towards them)

➡ Do normalisation.

MORE VARIABLES, WHAT CAN HAPPEN?

Be careful as more variables can yield more noise and overfitting.

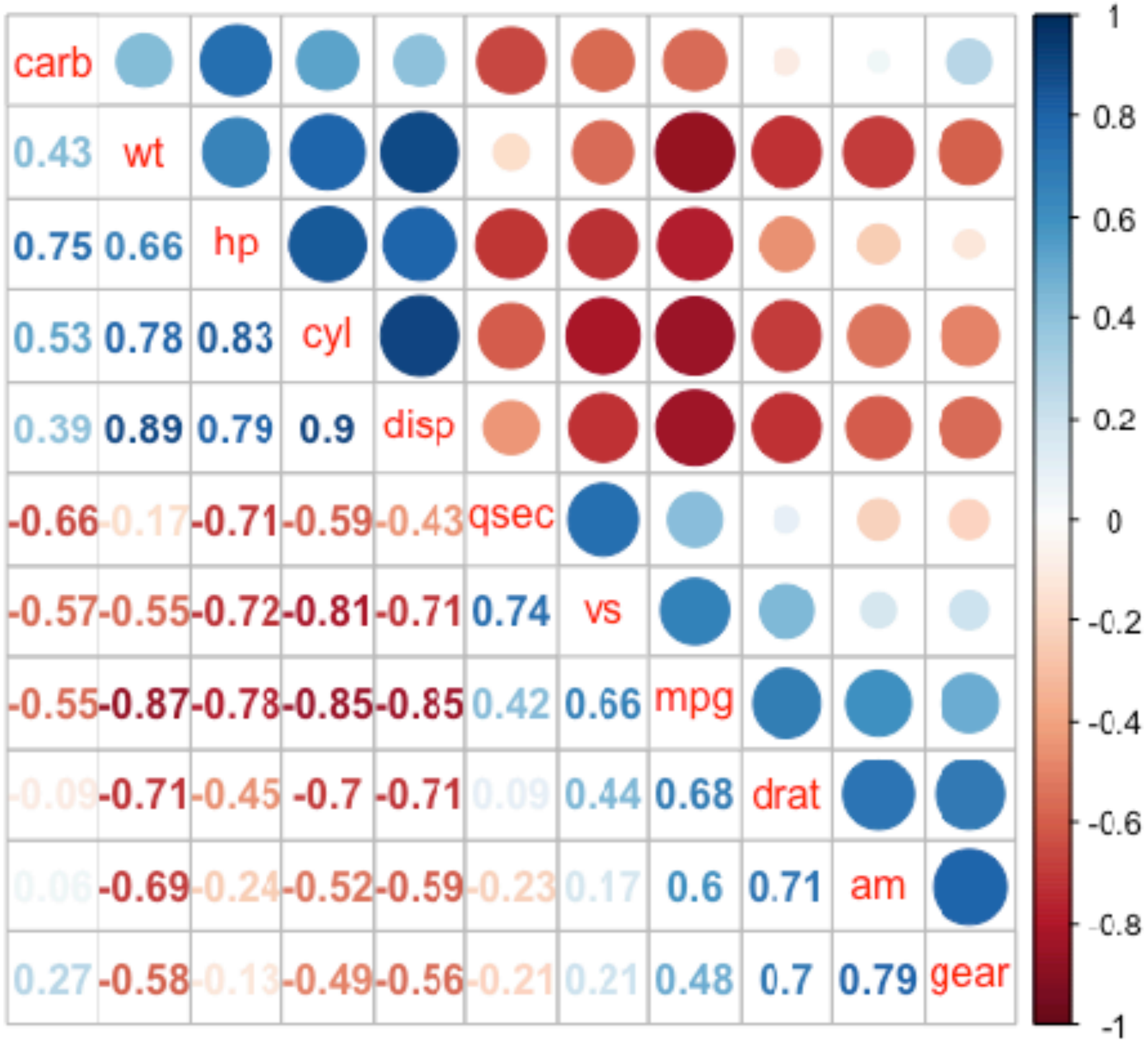
- Do feature selection (select the variables to use as predictors).

Be careful with variables that are correlated with each other (**collinearity**).

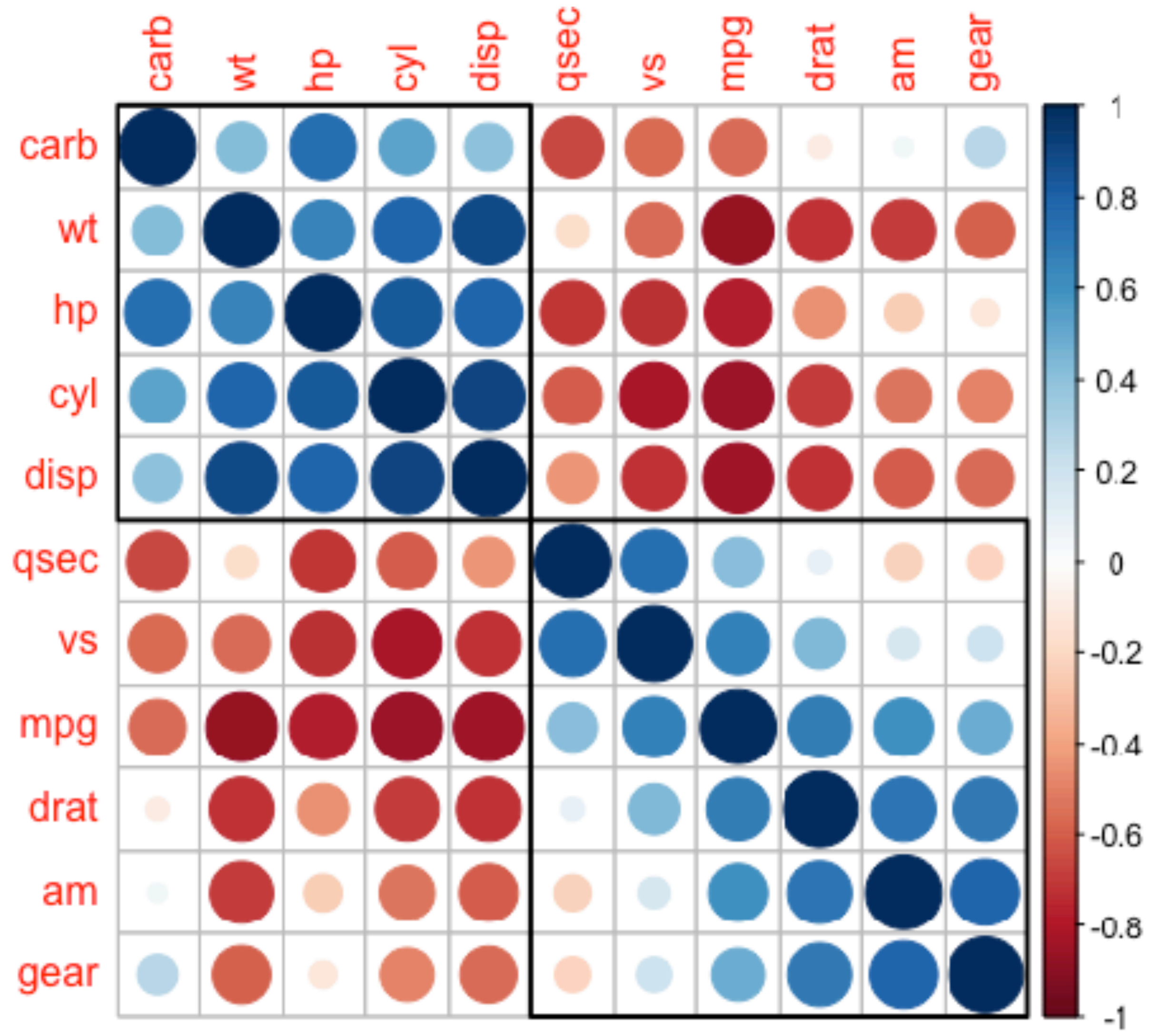
- Quick check of correlations.
- Inspect models fitted with or without a candidate variable, to differentiate confounders & collinear variables.
- Keep only one of the collinear variables.
- Keep the confounders.

CHECK CORRELATIONS

```
data %>% cor %>% corrplot.mixed()
```



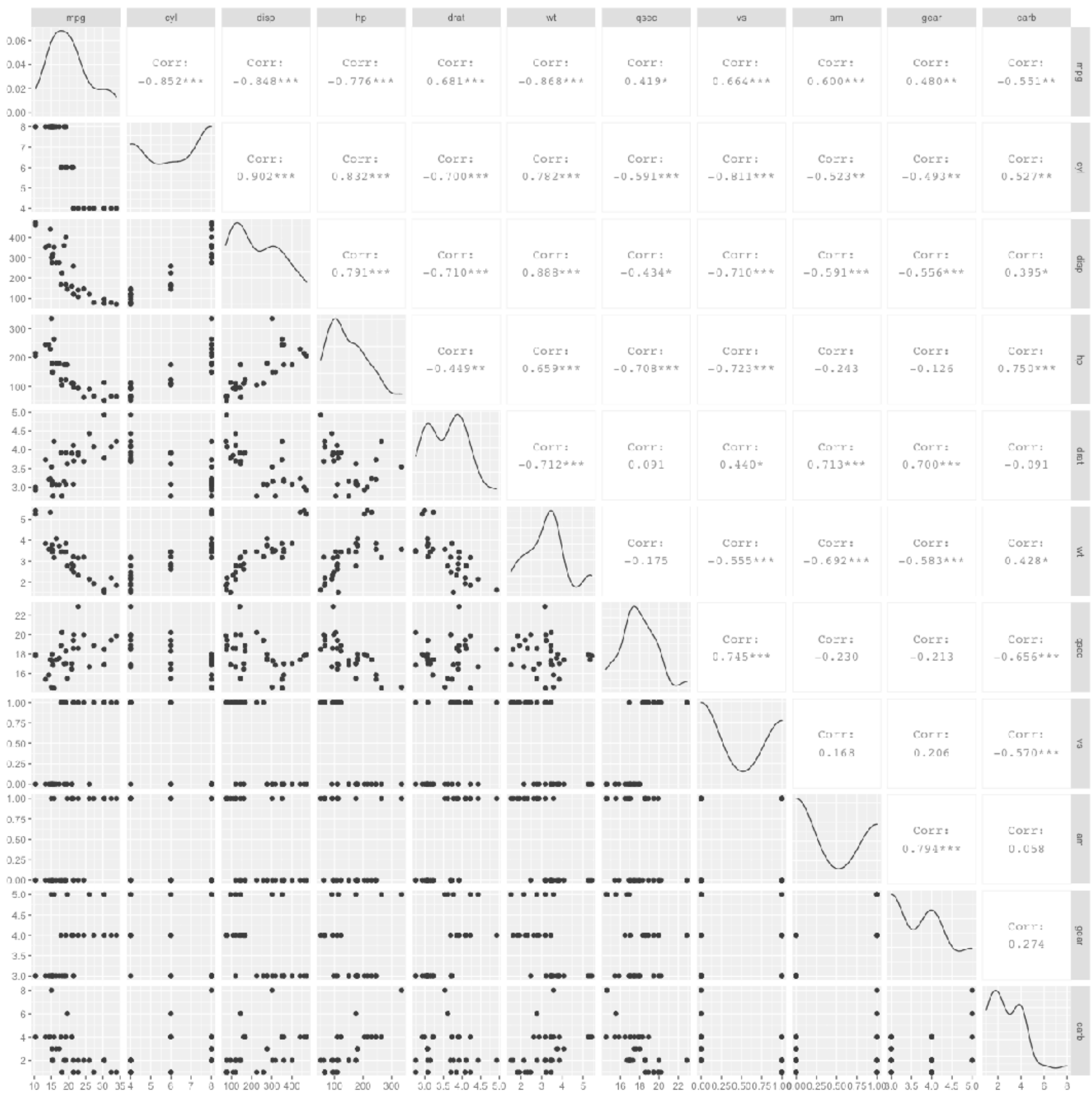
```
data %>% cor %>% corrplot(order = "hclust", addrect = 2)
```

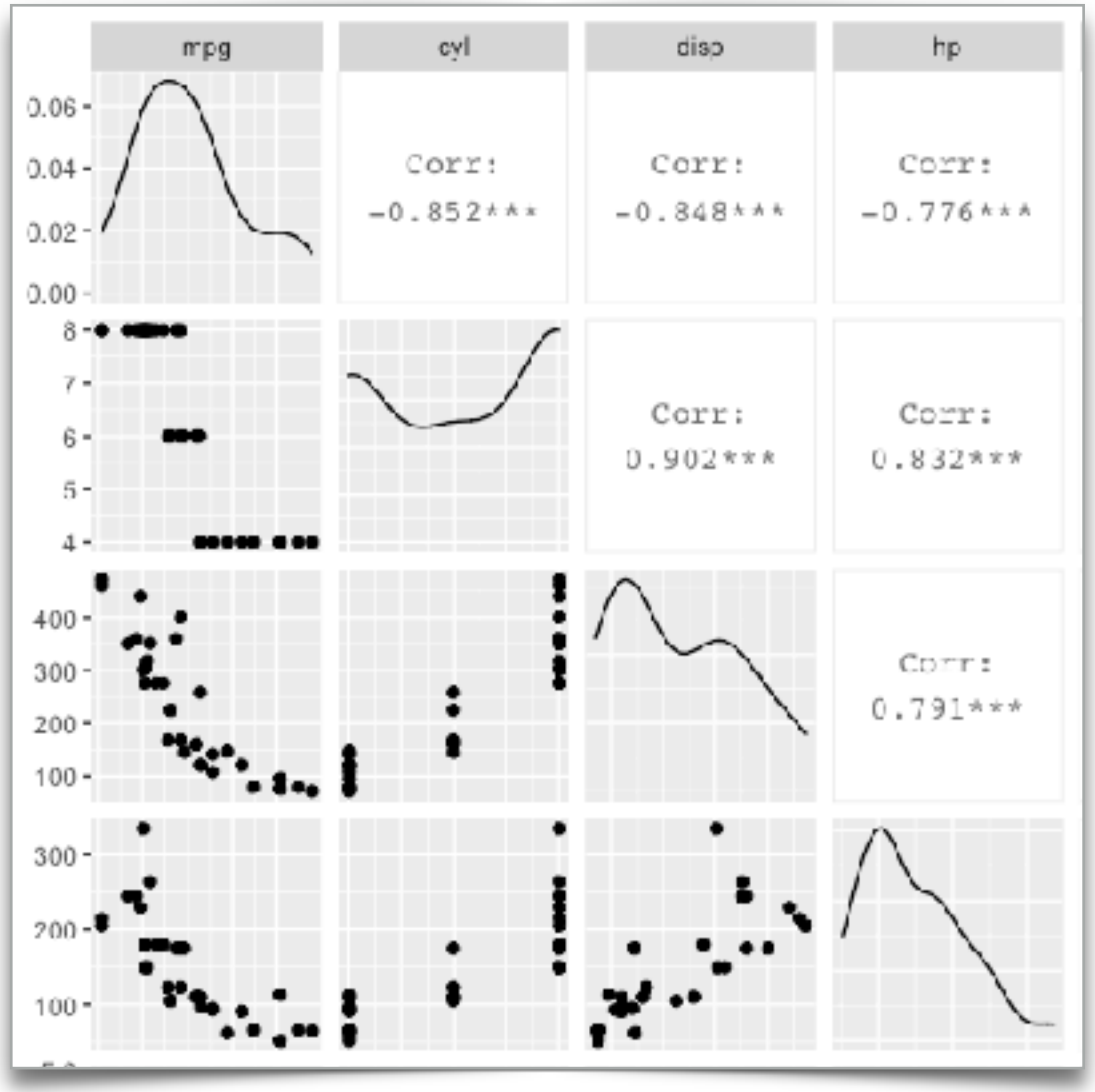


https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

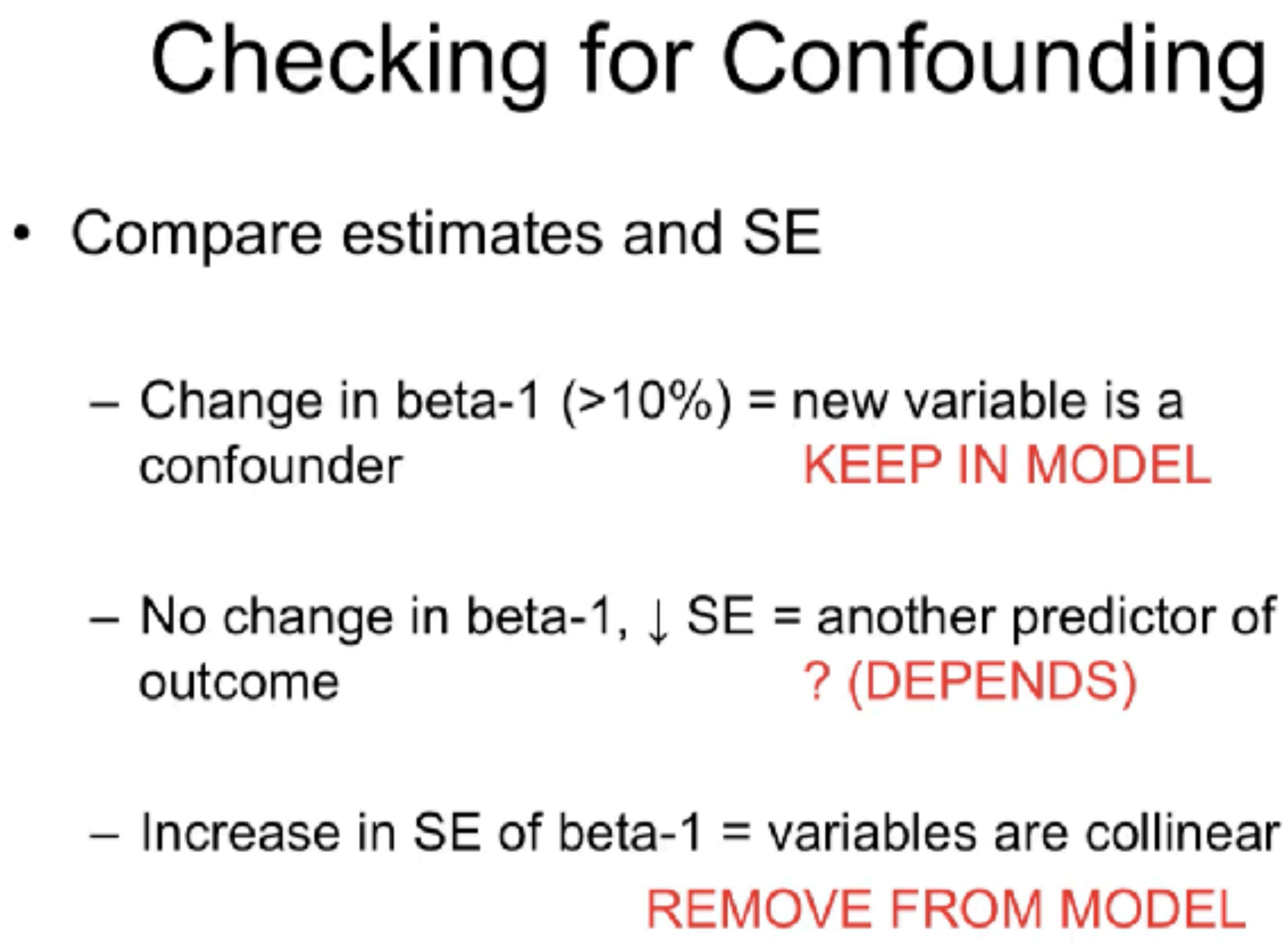
CHECK DISTRIBUTIONS

```
data %>% ggpairs()
```





CHECK IMPACT ON MODEL



Checking for Confounding

- Compare estimates and SE
  - Change in beta-1 (>10%) = new variable is a confounder KEEP IN MODEL
  - No change in beta-1, ↓ SE = another predictor of outcome ? (DEPENDS)
  - Increase in SE of beta-1 = variables are collinear REMOVE FROM MODEL

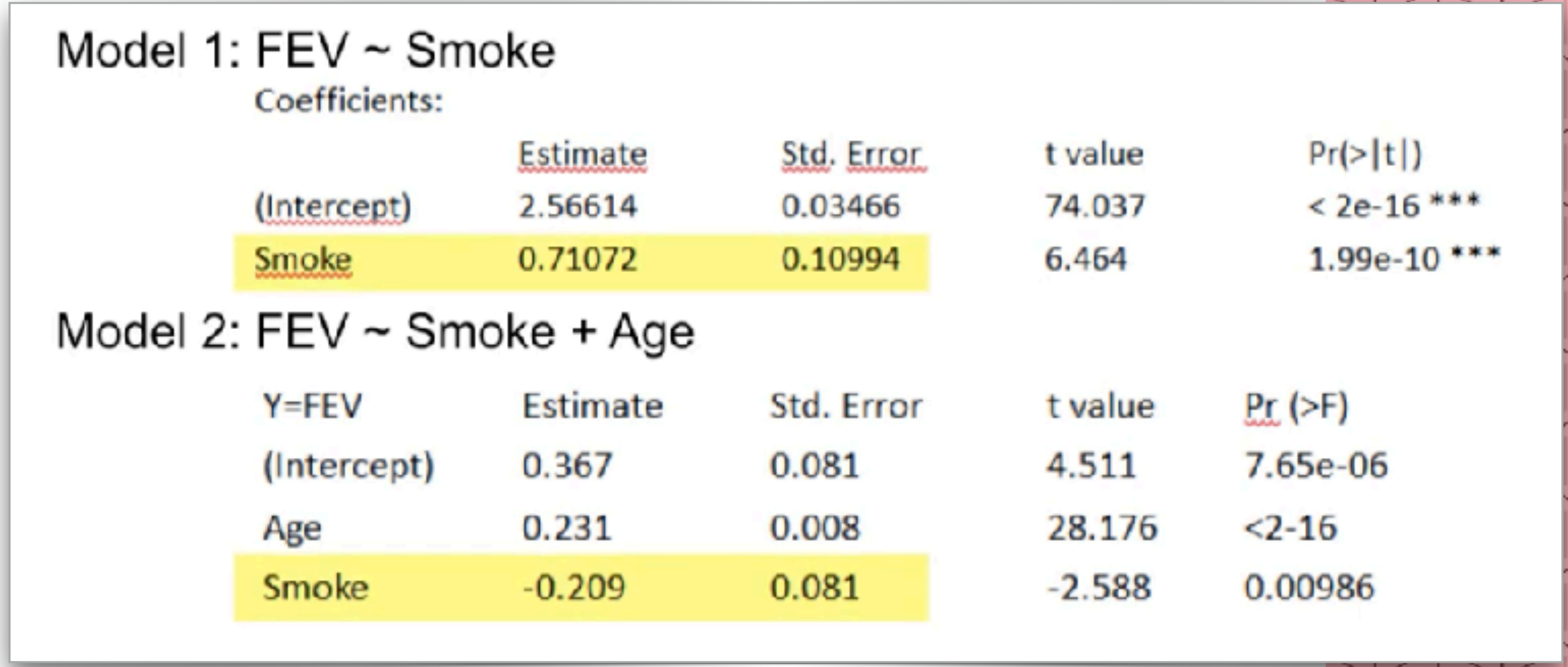
beta-1 *is the slope*

SE *is the Standard Error*
of the slope
(to make confidence intervals)

https://www.youtube.com/watch?v=Nwdp3wVxEBM

CHECK IMPACT ON MODEL

- Change in beta-1 (>10%) = new variable is a confounder KEEP IN MODEL
- No change in beta-1, ↓ SE = another predictor of outcome ? (DEPENDS)
- Increase in SE of beta-1 = variables are collinear REMOVE FROM MODEL



Model 1: FEV ~ Smoke

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.56614	0.03466	74.037	< 2e-16 ***
Smoke	0.71072	0.10994	6.464	1.99e-10 ***

Model 2: FEV ~ Smoke + Age

Y=FEV	Estimate	Std. Error	t value	Pr (>F)
(Intercept)	0.367	0.081	4.511	7.65e-06
Age	0.231	0.008	28.176	<2-16
Smoke	-0.209	0.081	-2.588	0.00986

beta-1 *is the slope*

SE *is the Standard Error*
of the slope
(to make confidence intervals)

Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

LET'S PRACTICE...

EMMA BEAUXIS-AUSSALET

e.m.a.l.beauxis@hva.nl