Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

FITTING REGRESSION

EMMA BEAUXIS-AUSSALET

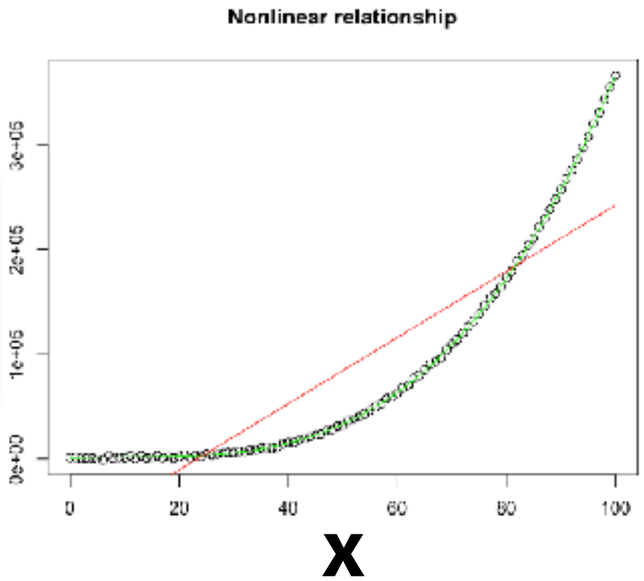
e.m.a.l.beauxis@hva.nl

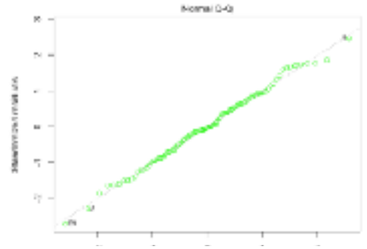
UNDER FITTING

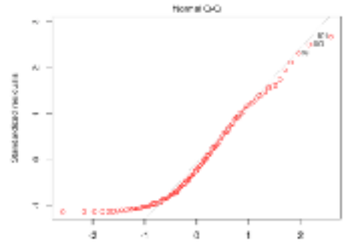
Regressions underfit the data when **essential patterns are not modelled**.
The **residuals are not random** w.r.t. the predictor or predicted variables
(e.g., not normally distributed with mean 0).

$\hat{y} = a_1\mathbf{x} + a_2\mathbf{x^2} + a_3\mathbf{x^3} + b$

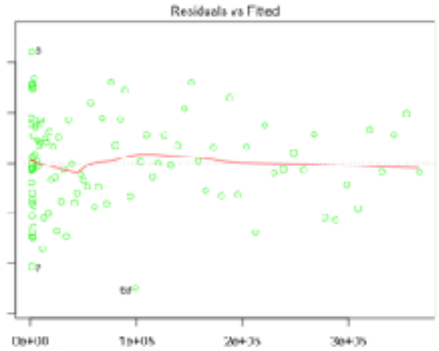
$\hat{y} = a\mathbf{x} + b$





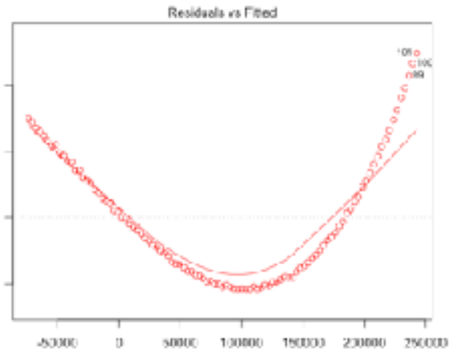


$y - \hat{y}$



$\hat{y} = a_1\mathbf{x} + a_2\mathbf{x^2} + a_3\mathbf{x^3} + b$

$y - \hat{y}$



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

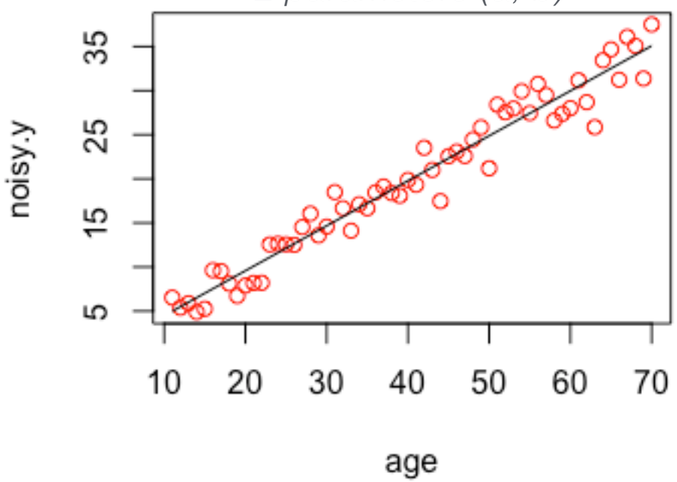
OVER FITTING

Regressions overfit the data when **sample-specific patterns are modelled**.
The **residuals may be low and random** (normally distributed with mean 0).

*More and more **complex regressions**, with more parameters to fit, and **higher risk of overfitting**…*

Simple linear (Degree 1)

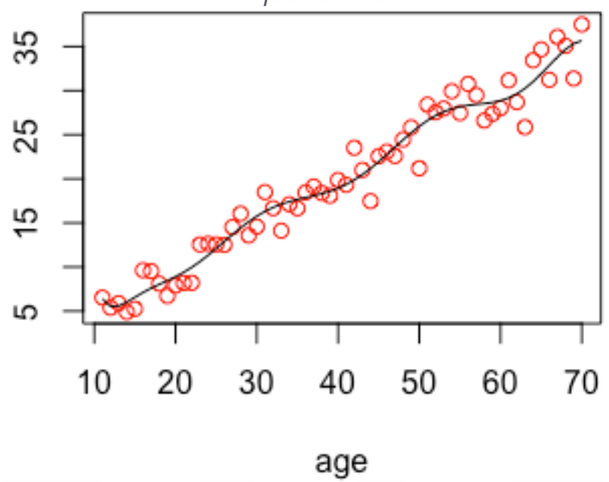
2 parameters (a, b)



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

Degree 10

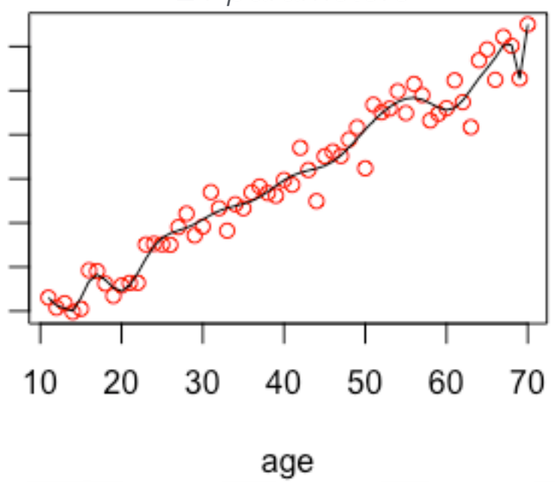
11 parameters



$$\hat{y} = a_1\mathbf{x} + a_2\mathbf{x^2} + \ldots + a_{10}\mathbf{x^{10}} + b$$

Degree 20

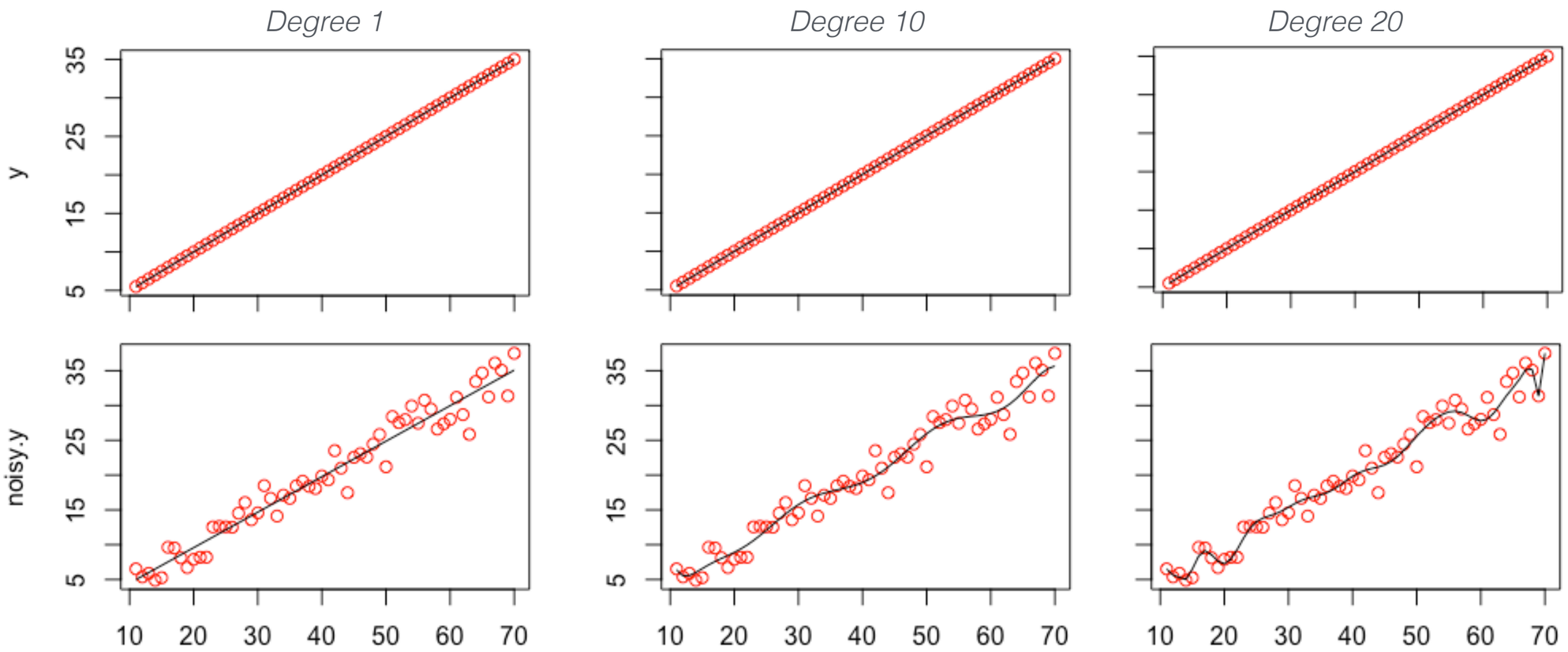
21 parameters



$$\hat{y} = a_1\mathbf{x} + a_2\mathbf{x^2} + \ldots + a_{20}\mathbf{x^{20}} + b$$

NOISE

The risk of overfitting arise from **naturally occurring random variations** (i.e., noise).



PREDICTION VERSUS DESCRIPTION

Under-fitted models have **poor predictive & descriptive properties**.

Over-fitted models have **poor predictive properties** but **better descriptive properties**.

UNDER FITTING

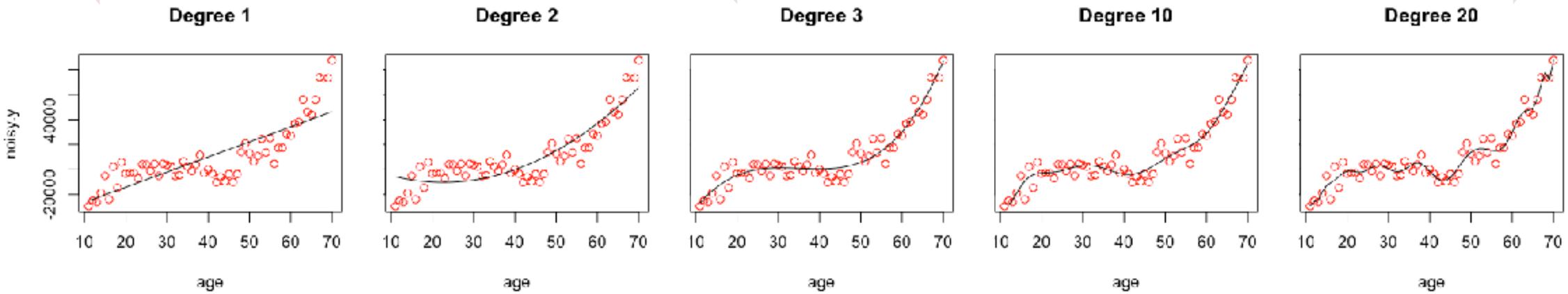
Less predictive, less descriptive…

GOOD FIT

More predictive…

OVER FITTING

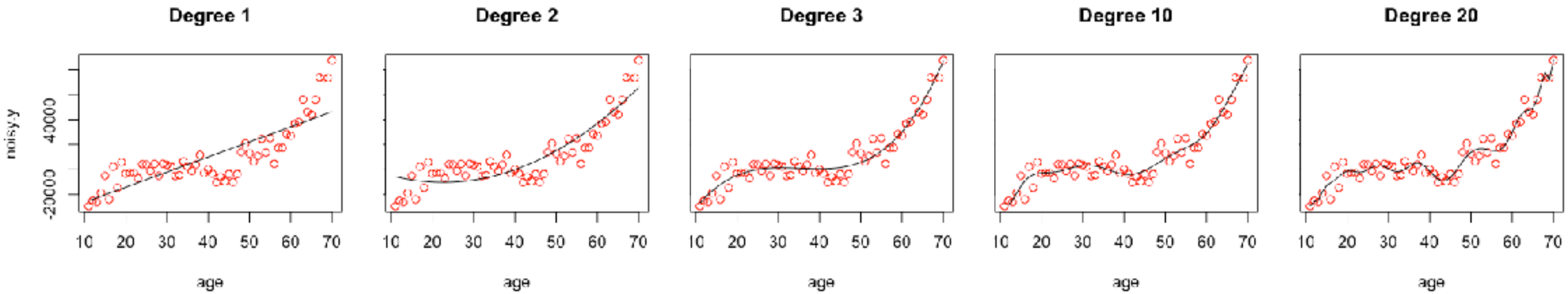
…more descriptive



R CODE

```r
###### Simple linear regression
model <- lm(y ~ age)
plot(model)
```

```r
###### Polynomial degreee 20
model <- lm(y ~ poly(age,20))
```



Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

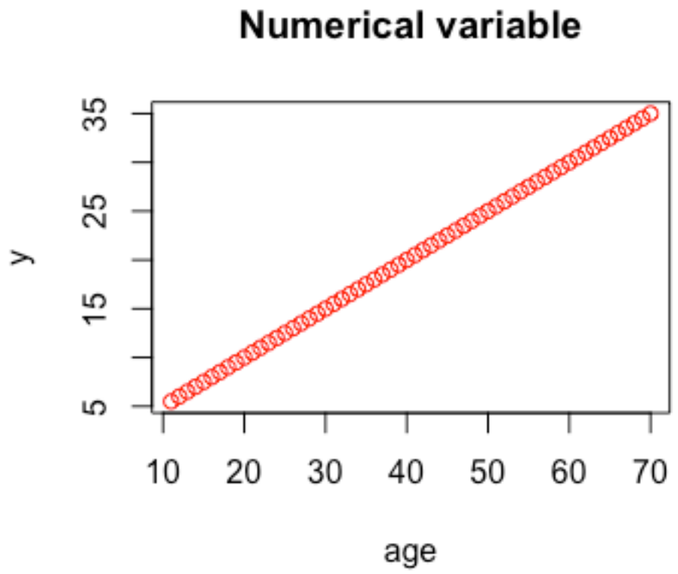
REGRESSION WITH ORDINAL DATA

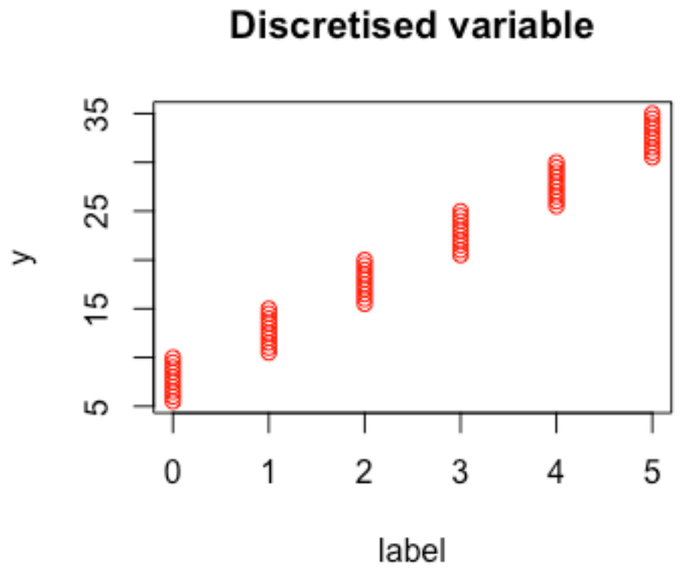
EMMA BEAUXIS-AUSSALET

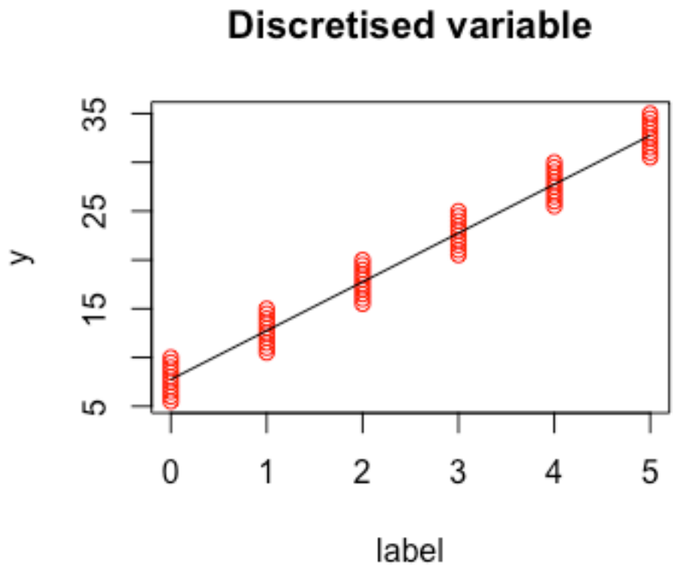
e.m.a.l.beauxis@hva.nl

ORDINAL DATA

Regressions can be fitted on ordered categorical variables by using **integer instead of categories**.





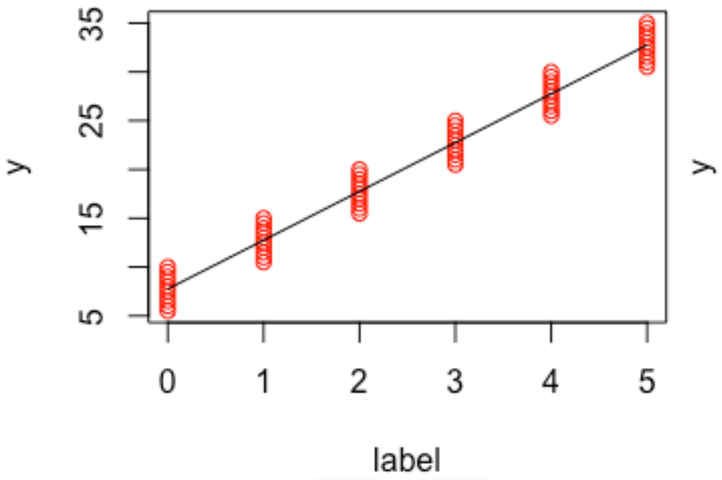



```
###### Linear regression
model <- lm(y ~ label)
```

ORDINAL DATA

Models fitted on discretised data can be reused with numerical data, assuming the patterns are consistent.

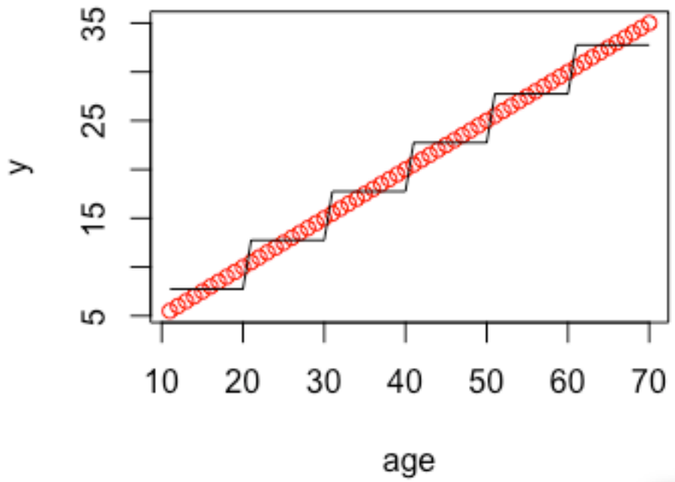
Discretised variable



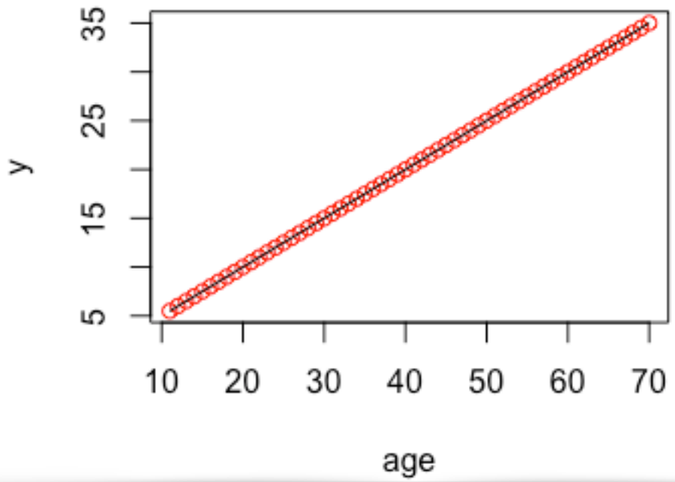

```
Call:
lm(formula = y ~ label)

Coefficients:
(Intercept)        label
       7.75         5.00
```

Numerical variable



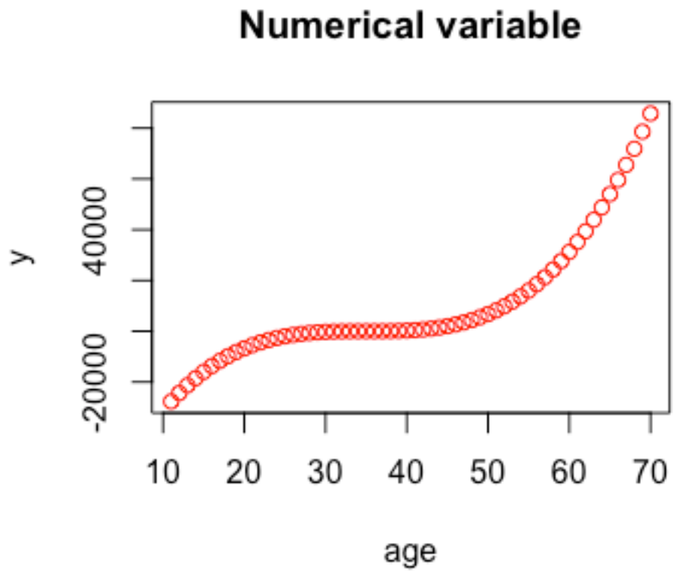
Numerical variable

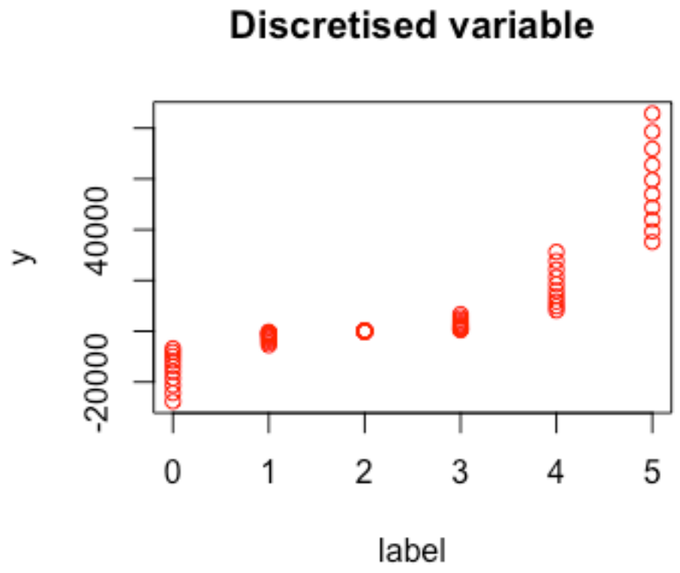


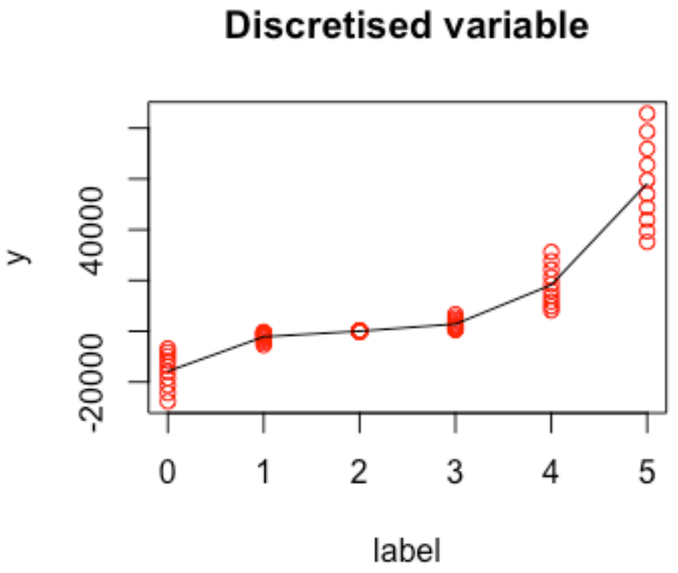

```
plot(age, y, type='p', col='red', main='Numerical variable')
y_hat = 5*(age/10-1.55) + 7.75
lines(age, y_hat)
```

ORDINAL DATA

Regressions can be fitted on ordered categorical variables (e.g., discretised data) by using **integer instead of categories**.



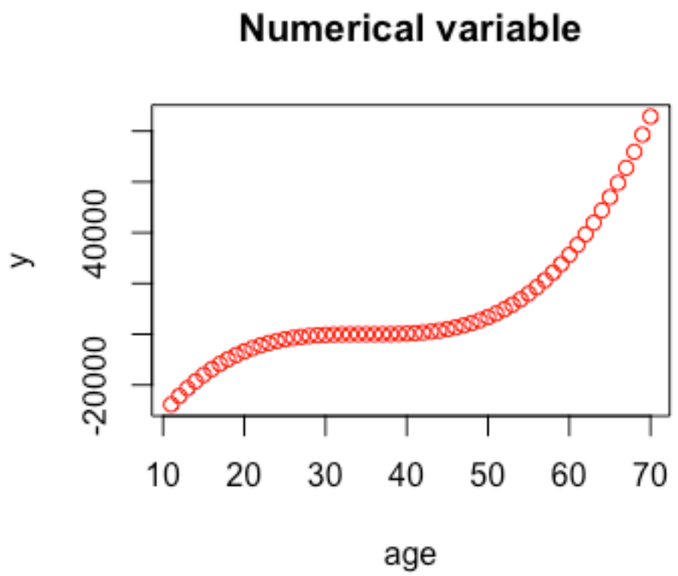


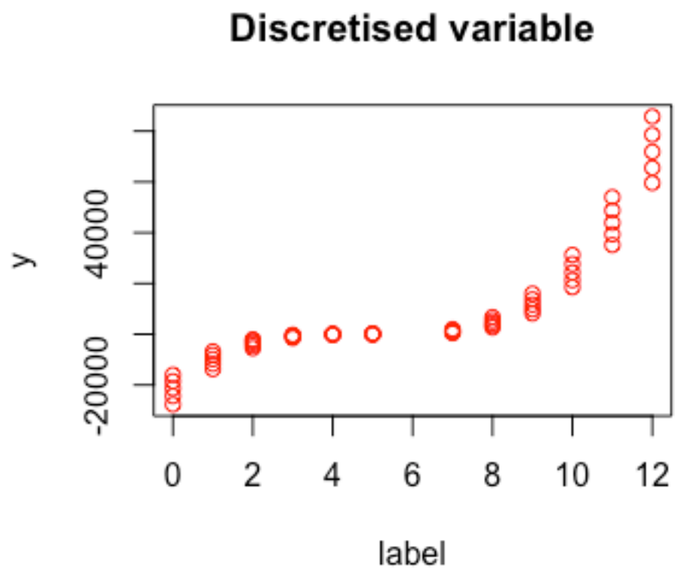


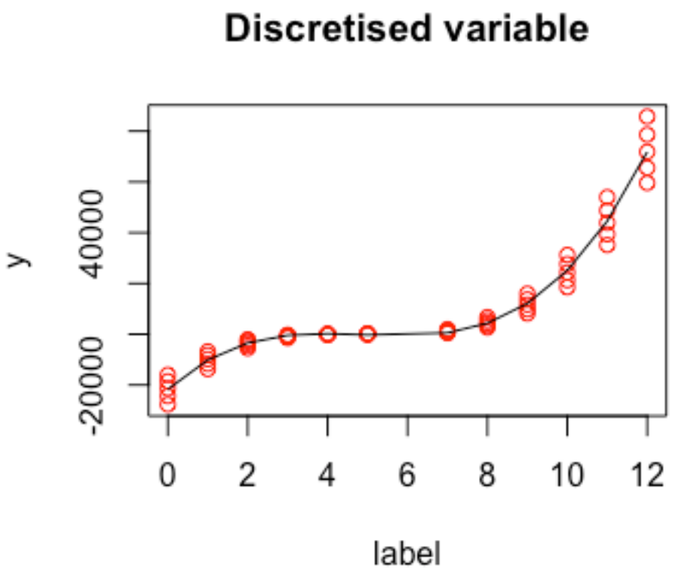

```
###### Polynomial regression - Degree 3
model <- lm(y ~ poly(label,3))
```

ORDINAL DATA

The granularity of discretisation (i.e., bin width) impacts the uncertainty.



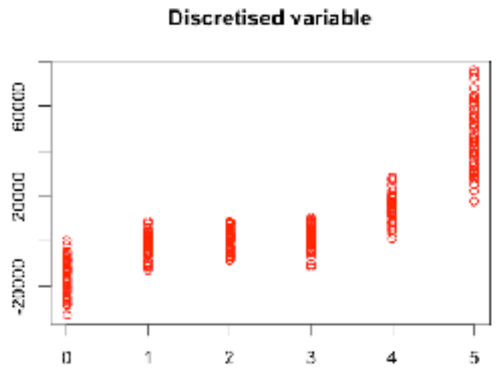


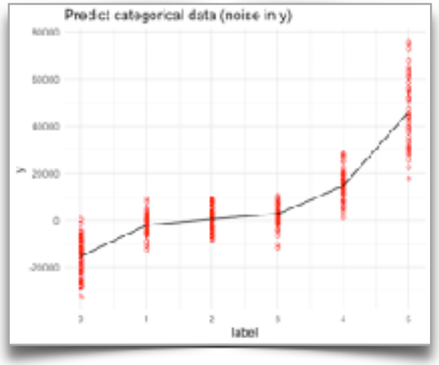


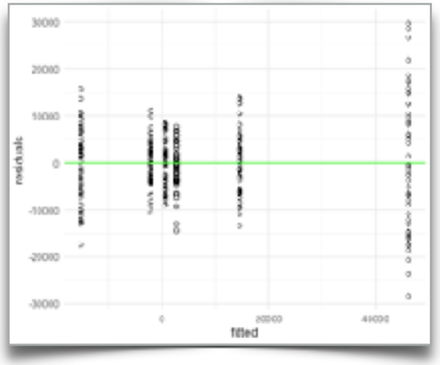

```
###### Polynomial regression - Degree 3
model <- lm(y ~ poly(label,3))
```

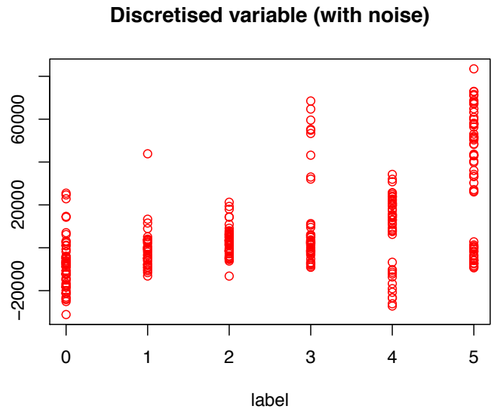
ORDINAL DATA

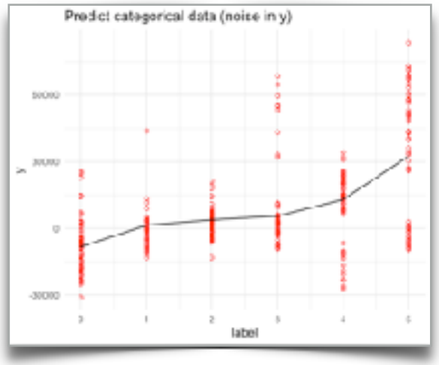
Residuals are also discretised, and subject to noise and heteroscedasticity.

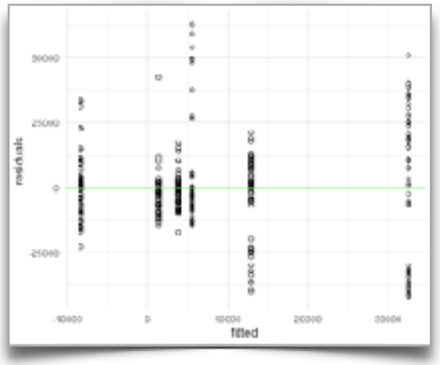












Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

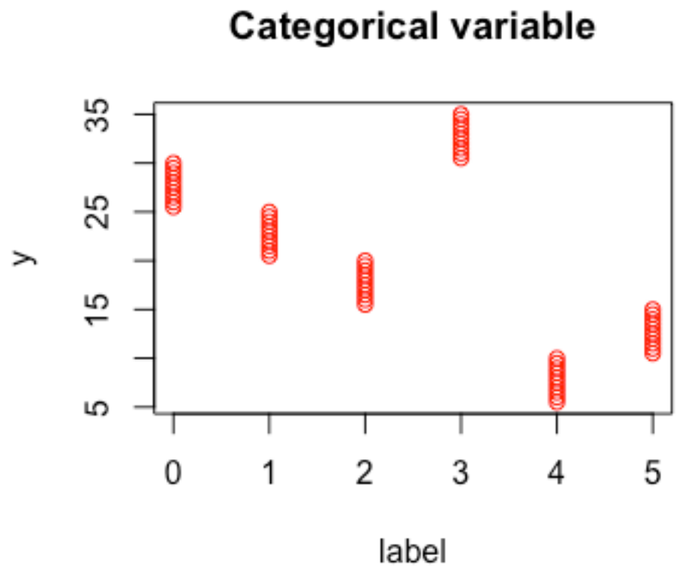
REGRESSION WITH CATEGORICAL DATA

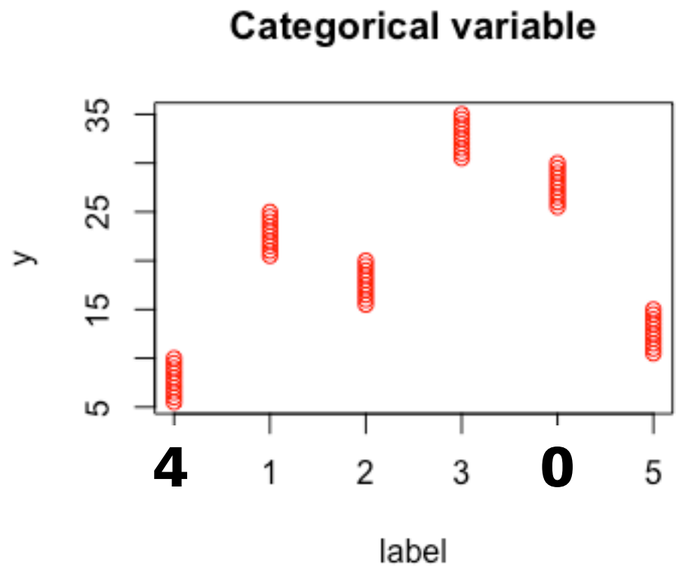
EMMA BEAUXIS-AUSSALET

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

UNORDERED CATEGORIES

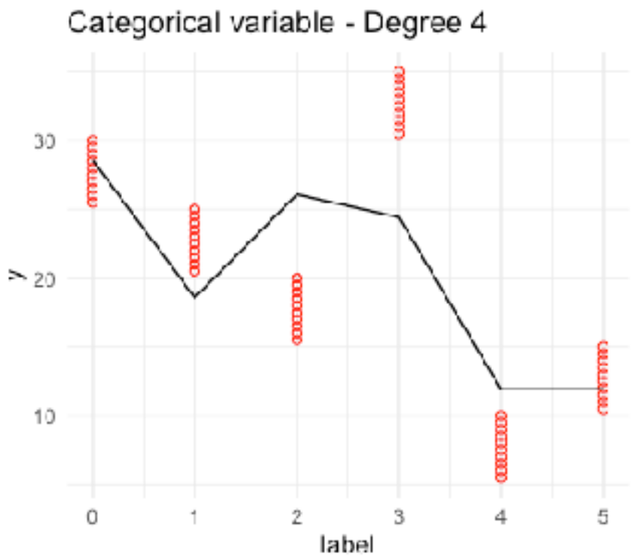
When transforming multiple categories into numerical index, the **patterns are arbitrary**.

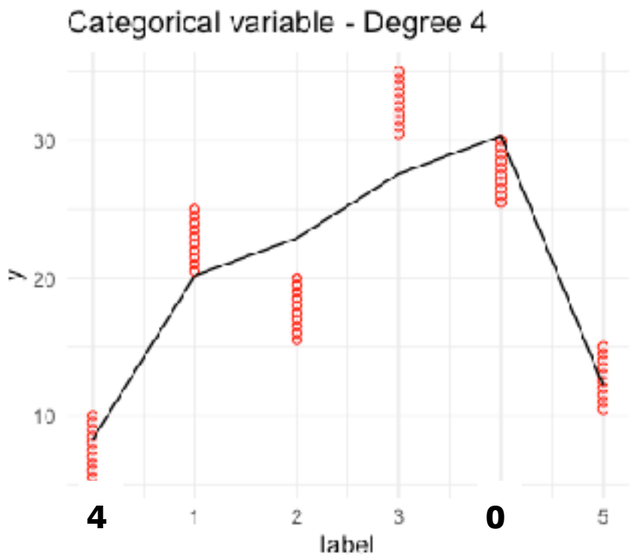




UNORDERED CATEGORIES

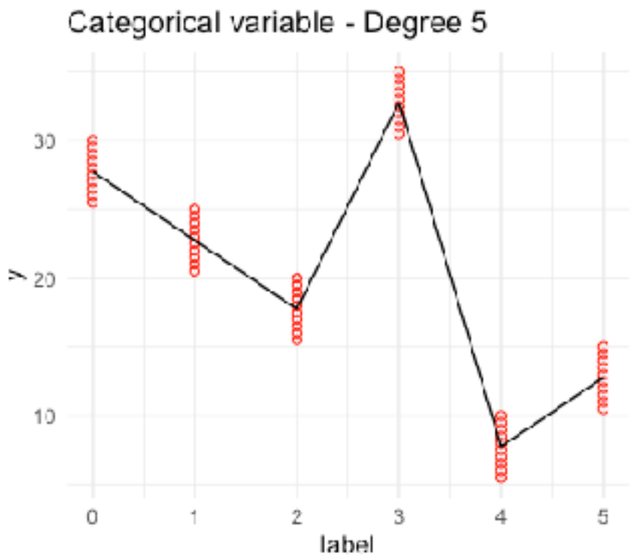
When transforming multiple categories into numerical index, the **patterns are arbitrary**. **Models are biased**.

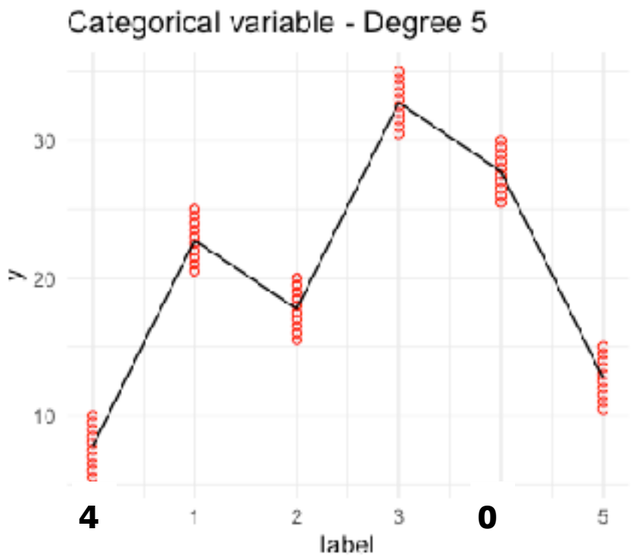




UNORDERED CATEGORIES

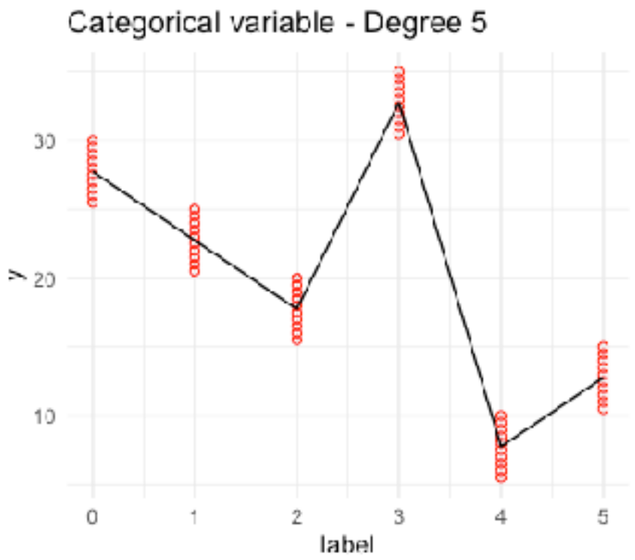
When transforming multiple categories into numerical index, the **patterns are arbitrary**. **Models are biased**… unless completely overfitting!

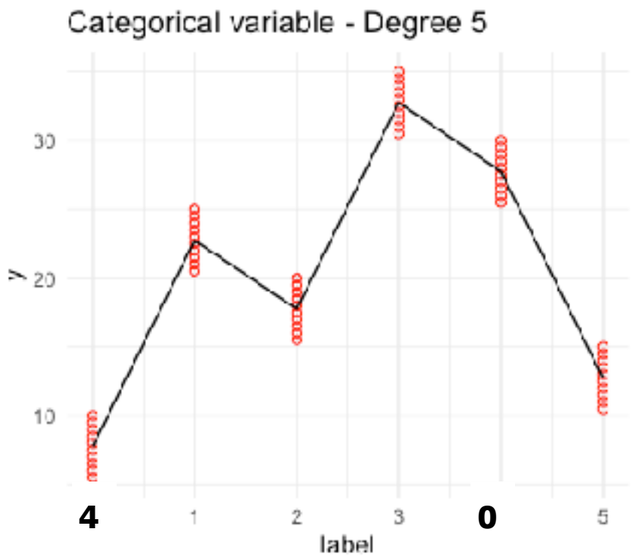




UNORDERED CATEGORIES

When transforming multiple categories into numerical index, the **patterns are arbitrary**.
Models are biased… unless completely overfitting!
And they remain **uninterpretable**, and an issue for multivariate regressions.





DUMMY CODING

Categorical data must be **transformed into "*wide*" data** with a column per category. For each data point, the **category is indicated with 1 or 0** (in category, or not).

```
> data_narrow[1:15,]
   label    y
1      A  5.5
2      A  6.0
3      A  6.5
4      A  7.0
5      A  7.5
6      A  8.0
7      A  8.5
8      A  9.0
9      A  9.5
10     A 10.0
11     F 10.5
12     F 11.0
13     F 11.5
14     F 12.0
15     F 12.5
```

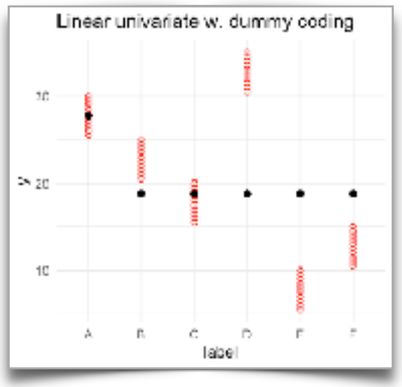
```
> data_narrow_prep[1:15,]
   id label    y dummy
1   1     A  5.5     1
2   2     A  6.0     1
3   3     A  6.5     1
4   4     A  7.0     1
5   5     A  7.5     1
6   6     A  8.0     1
7   7     A  8.5     1
8   8     A  9.0     1
9   9     A  9.5     1
10 10     A 10.0     1
11 11     F 10.5     1
12 12     F 11.0     1
13 13     F 11.5     1
14 14     F 12.0     1
15 15     F 12.5     1
```

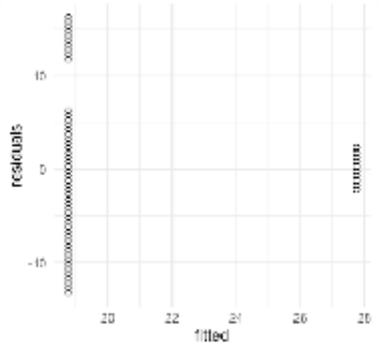
```
> data_wide <- data_narrow_prep %>% spread(label, dummy, fill=0)
> data_wide[1:15,]
   id    y A B C D E F
1   1  5.5 1 0 0 0 0 0
2   2  6.0 1 0 0 0 0 0
3   3  6.5 1 0 0 0 0 0
4   4  7.0 1 0 0 0 0 0
5   5  7.5 1 0 0 0 0 0
6   6  8.0 1 0 0 0 0 0
7   7  8.5 1 0 0 0 0 0
8   8  9.0 1 0 0 0 0 0
9   9  9.5 1 0 0 0 0 0
10 10 10.0 1 0 0 0 0 0
11 11 10.5 0 0 0 0 0 1
12 12 11.0 0 0 0 0 0 1
13 13 11.5 0 0 0 0 0 1
14 14 12.0 0 0 0 0 0 1
15 15 12.5 0 0 0 0 0 1
```

DUMMY CODING

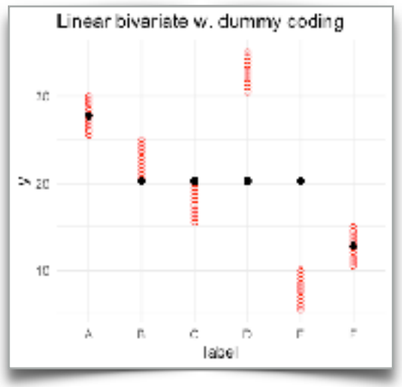
Multivariate linear regressions can accurately fit dummy coded variables.

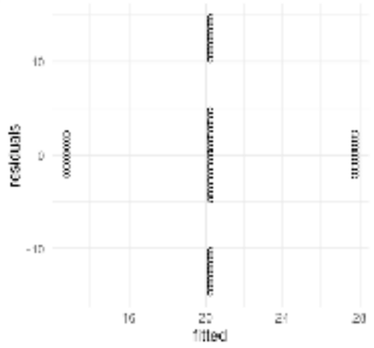
```
########## Univariate regression
model <- lm(y ~ A, data=data_wide)
```



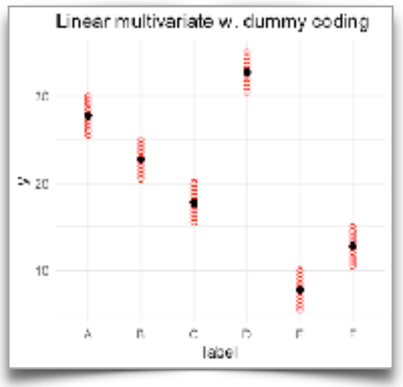


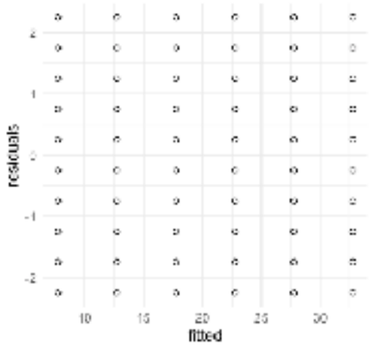

```
########## Bivariate regression
model <- lm(y ~ A+F, data=data_wide)
```






```
########## Multivariate regression
model <- lm(y ~ A+B+C+D+E+F, data_wide)
```

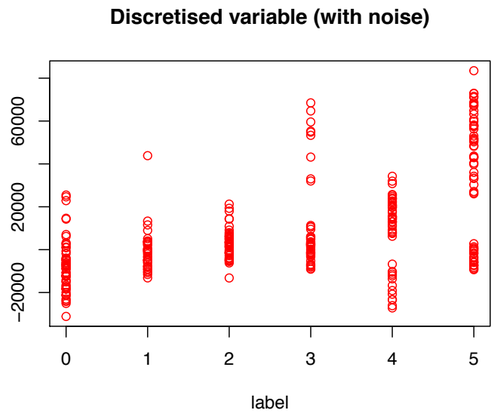


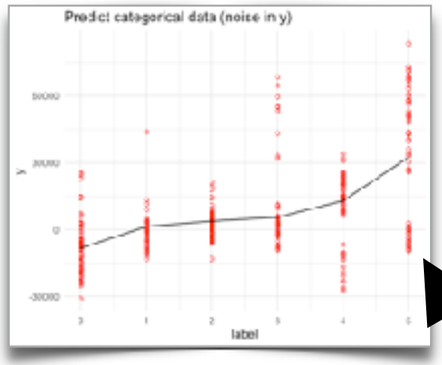


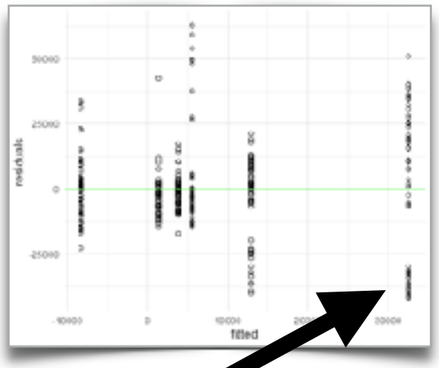
	id	y	A	B	C	D	E	F
1	1	5.5	1	0	0	0	0	0
2	2	6.0	1	0	0	0	0	0
3	3	6.5	1	0	0	0	0	0
4	4	7.0	1	0	0	0	0	0
5	5	7.5	1	0	0	0	0	0
6	6	8.0	1	0	0	0	0	0
7	7	8.5	1	0	0	0	0	0
8	8	9.0	1	0	0	0	0	0
9	9	9.5	1	0	0	0	0	0
10	10	10.0	1	0	0	0	0	0
11	11	10.5	0	0	0	0	0	1
12	12	11.0	0	0	0	0	0	1
13	13	11.5	0	0	0	0	0	1
14	14	12.0	0	0	0	0	0	1
15	15	12.5	0	0	0	0	0	1

ORDINAL DATA

Dummy coding is also recommended for ordinal data and multivariate regression.







The variable interactions creating these patterns would be better modelled with dummy coding.

Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

QUESTIONS?

EMMA BEAUXIS-AUSSALET

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