Dealing with non-linearity: Categorization, Polynomial functions and Splines

Giorgio Melloni

01/20/2022

OUTLINE

INTRODUCTION

- Correct reporting of continuous risk
- How is risk reported in the literature?
- Comparison of methods

DICHOTOMIZATION/CATEGORIZATION

- Why dichotomizing is a bad idea
- Why categories are also a bad idea
- Cases where categories are acceptable

NON-LINEAR RELATIONSHIP

- Linear regression and categories
- Polynomials
- Splines

CONCLUSIONS

*The Strengthening the Reporting of Observational Studies in Epidemiology (**STROBE**) Statement: Guidelines for reporting observational studies* devised a 22-item checklist for the correct reporting of observational studies

Number 11 suggests:

"Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why"

https://www.strobe-statement.org/checklists/
Elm *et al.* 2014

Reporting estimates of a continuous outcome (e.g. Risk) on the levels of a continuous risk factor is not always straightforward.

From a 2010 survey of 58 studies reporting results for continuous risk factors:

- Categorisation occurred in 50 (86%) of them
- Of those, 42% also analyzed the variable continuously and 24% considered alternative groupings.
- Most (78%) used 3 to 5 groups.
- Categorical risk estimates were most commonly (66%) presented as pairwise comparisons to a reference group, usually the highest or lowest (79%).

Turner *et al.* , Epidemiol Perspect Innov , 2010

Transformation function

$$y = f(x)$$

Outcome *Observation*

If linearity holds

- Dichotomize/Categorize
- Linear Regression

Beyond linearity

- Polynomial Regression
- Splines

$$\mathrm{E}(y|x) = \beta_0 + \beta_1 f(x)$$
$$\mathrm{Logit}(\mathrm{P}(y = 1|x)) = \beta_0 + \beta_1 f(x)$$
$$\log(Hazard|x) = h_0 + \beta_1 f(x)$$

Valid in every model that expresses risk continuously

Linear

Logistic

Cox

While easy to explain, **dichotomizing** a continuous variable is **almost always a very bad idea**

- Dichotomization leads to a significant loss of power
  - Many simulation studies demonstrated a loss of power equivalent to losing 1/3 of the sample size
  - Efficiency is reduced to 65% for normally distributed predictors and 48% for exponential/skewed distributions (Lagakos SW. , Statistics in Medicine , 1988)

Royston *et al.,* Statistics in Medicine, 2006

Dichotomize can be an **arbitrary decision**

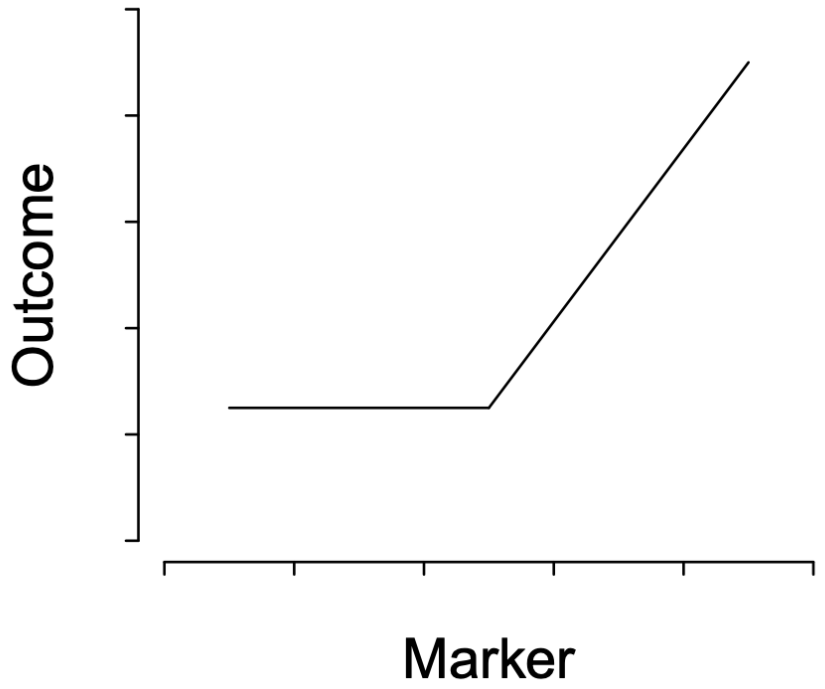
- We could use an **established** cutpoint (e.g. BMI > 30)
  - It's not always possible
  - If we adjust for other variables, the model might make the cutpoint invalid

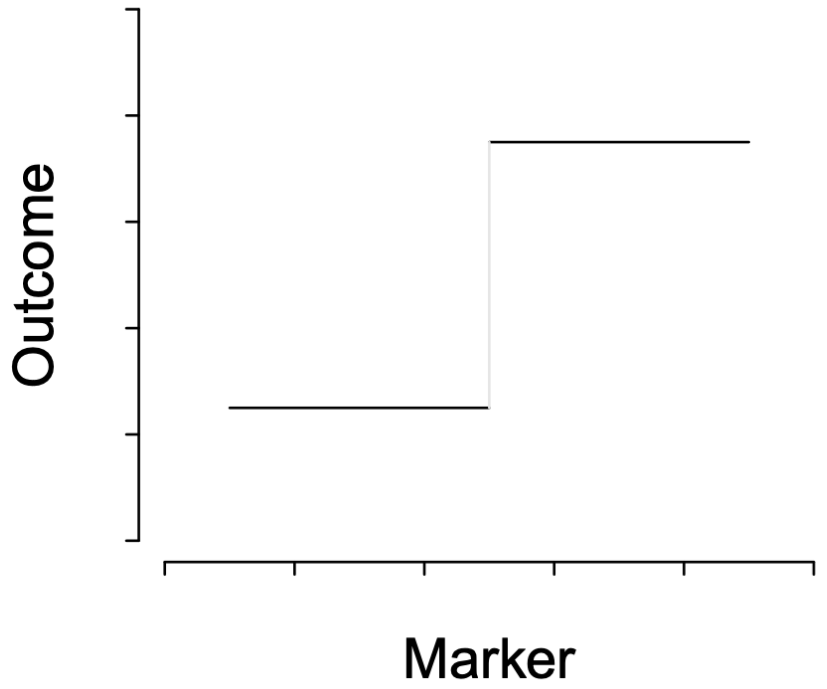
- We can use a **data-driven** cutpoint (e.g. median split)
  - Median split is the best choice for a cutoff just because it balances the sample size (good for t-test and chi-square)
  - heavily data dependent
  - impossible to reproduce in a different dataset

- We can use an "**optimal**" cutpoint by choosing the lowest p-value obtained through different cutpoints
  - Increase in type I error if we don't consider multiple testing hypothesis
  - impossible to reproduce in a different dataset

Royston *et al.,* Statistics in Medicine, 2006

DICHOTOMIZATION/CATEGORIZATION





Real cutpoint. Mathematically, it is a discontinuity in the first derivative (slope change) and there are specific mathematical tools to find it.

Artificial Cutpoint. Assume that the risk is homogeneous on either sides of the cutpoint, which never happens

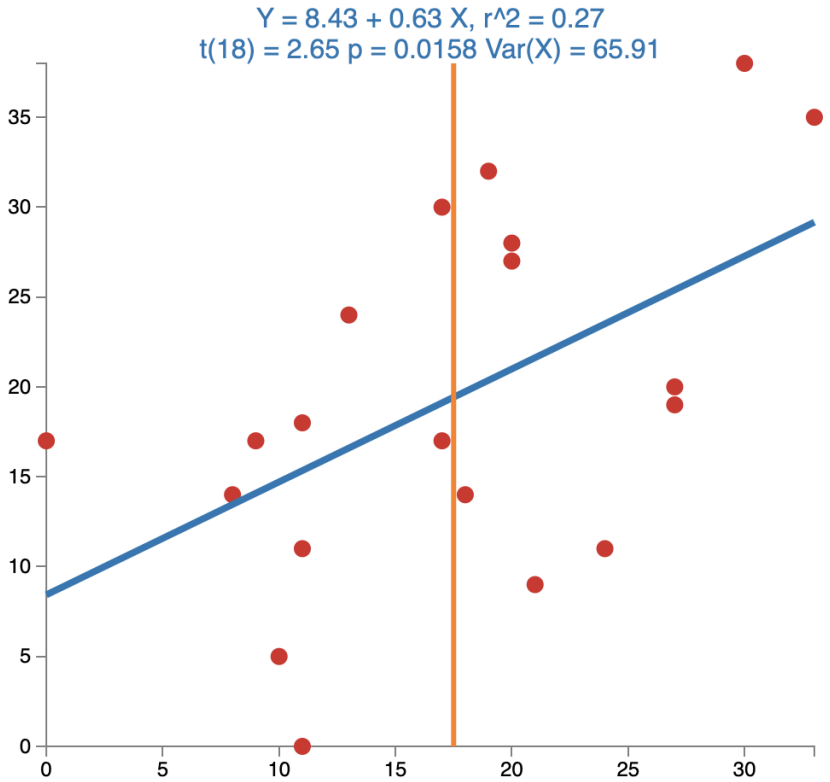
FE Harrell & JC Slaughter, Biostatistics for Biomedical Research, updated on Feb 2021

DICHOTOMIZATION/CATEGORIZATION

What does happen when you dichotomize?

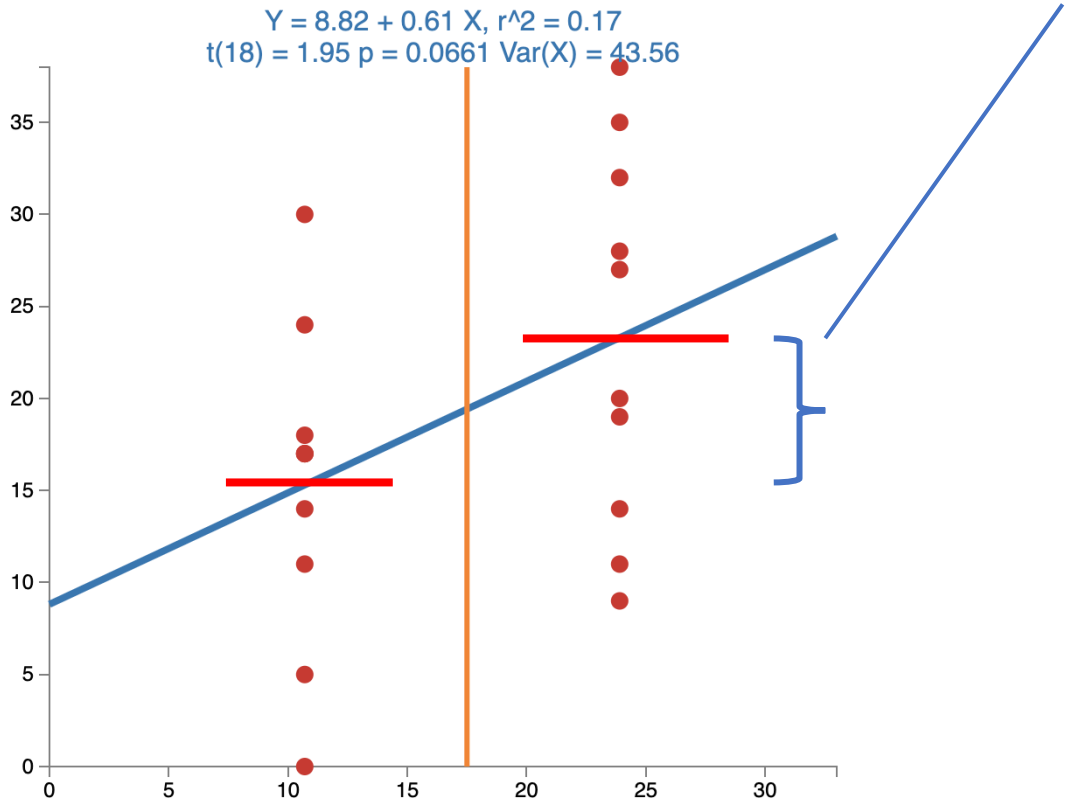
- In most cases, power is affected, leading to **higher p-values**
- Regression line doesn't change much but R2 decreases
- R2 decreases because the **variability decreases**. Loss of information = Loss of power

Equivalent to a **t-test** on the means of the left hand side and right hand side



Original Data

We split the predictor at the median. Instead of assign 0/1 values, we assign the mean of each group



Median Split

McClelland, G. , Journal of Consumer Psychology, 2015.
http://psych.colorado.edu/~mcclella/MedianSplit/

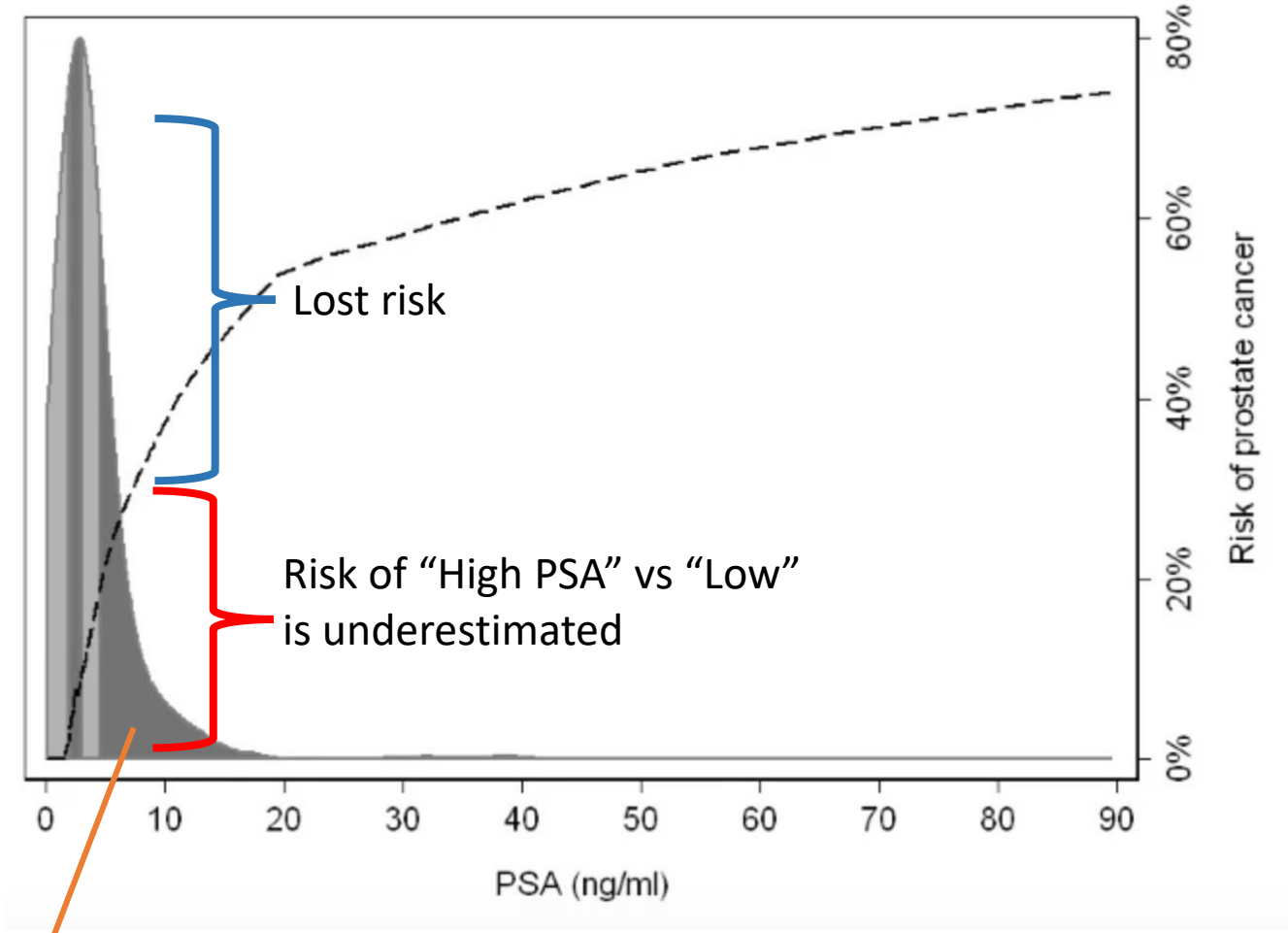
DICHOTOMIZATION/CATEGORIZATION

Dividing in several groups based on quantiles is not different.

Here's an example using PSA values and risk of prostate cancer

The last quartile takes a larger range of PSA values and the risk estimate is attenuated

The risk keeps increasing in a non-linear fashion that cannot be captured by the division in quartiles

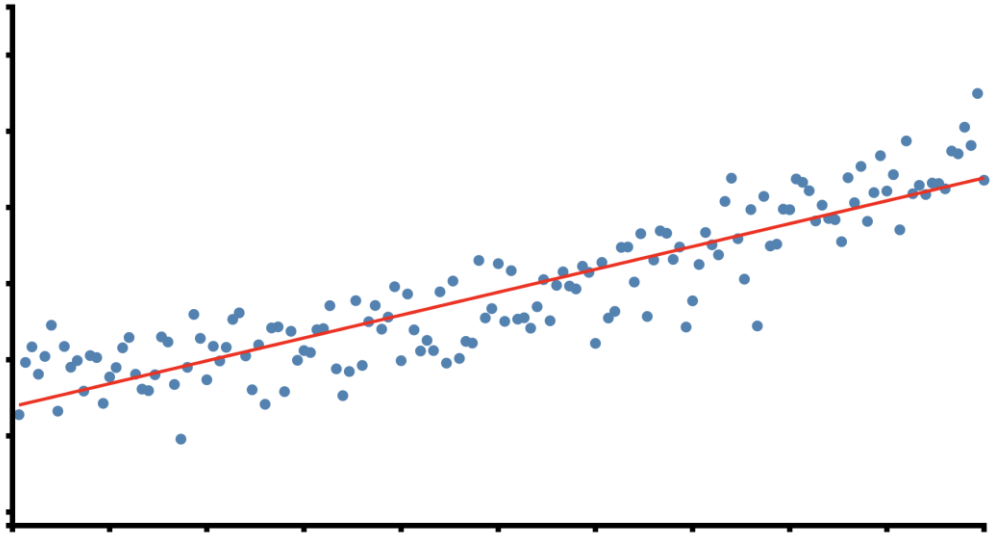


The highest quartile has the largest range of PSA value

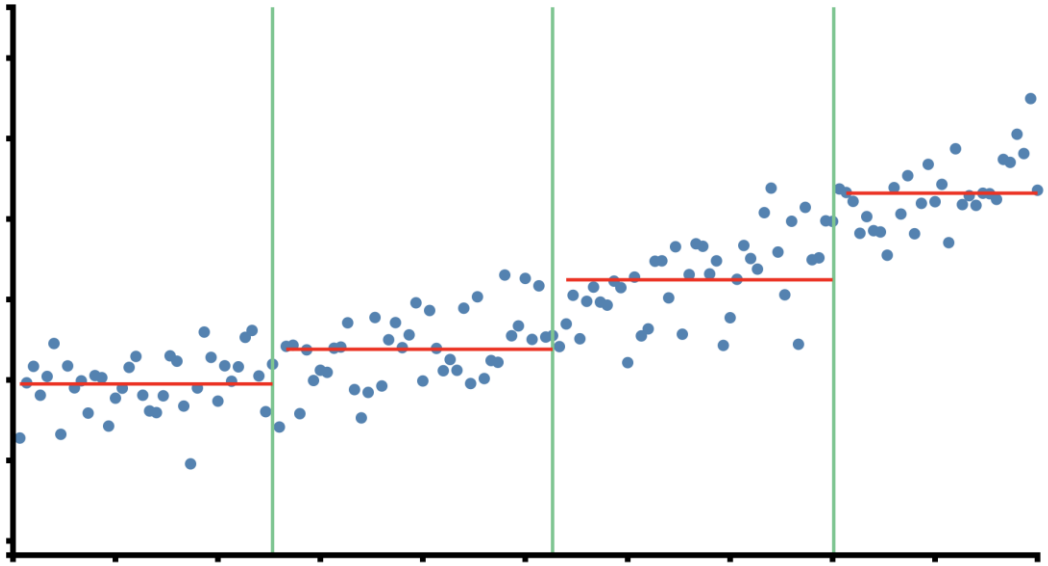
A simple log transformation could lead to reliable linear results

Bennette *et al.* , BMC Med Res Met, 2012

DICHOTOMIZATION/CATEGORIZATION



Linear Regression
R-squared = 0.803
N. of parameters = 2



Quartile step function
R-squared = 0.788
N. of parameters = 4

If the underlying relationship is truly linear, the loss in terms of overall model fit is not substantial

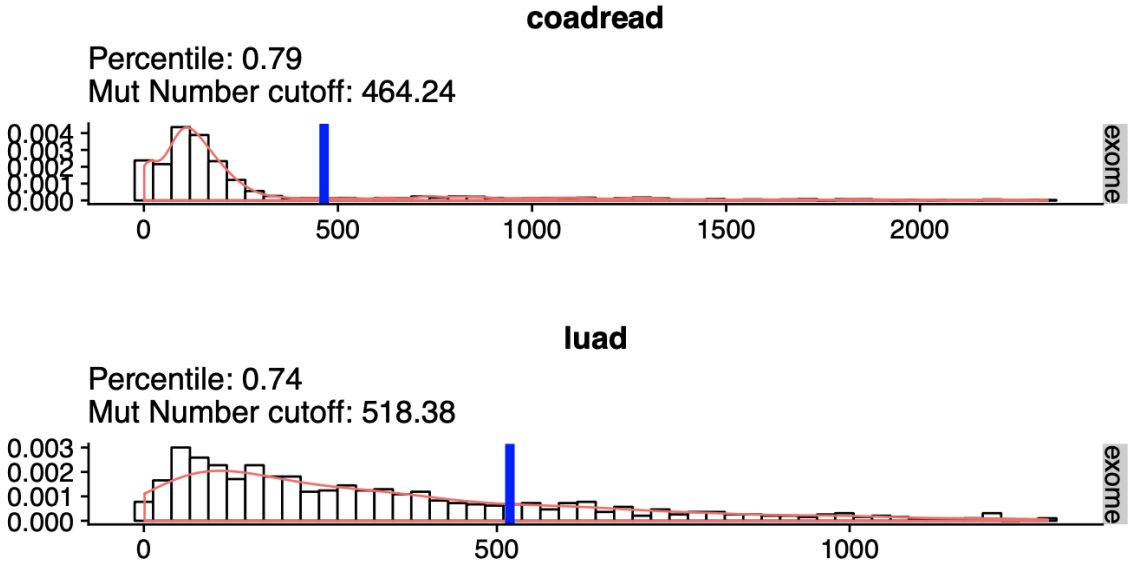
… but we are using an excessive number of parameters (aka degrees of freedom) to obtain the same result of a linear regression.

4 estimates = 4 dummies = 4 p-values which leads to a level of accepted significance of 0.05/4 = 0.0125

Simulations carried out using https://pclambert.net/interactivegraphs

When is categorization acceptable?

- Categorize Y (risk) is better than categorize X (predictor) because it makes the decision making about the outcome (a posteriori) rather than about the cause (a priori)
- Time can be dichotomized (before / after comparison). KM curves are a good of example of naturally occurring step functions in time
- When the population under exam is truly non-homogeneous for the predictor values



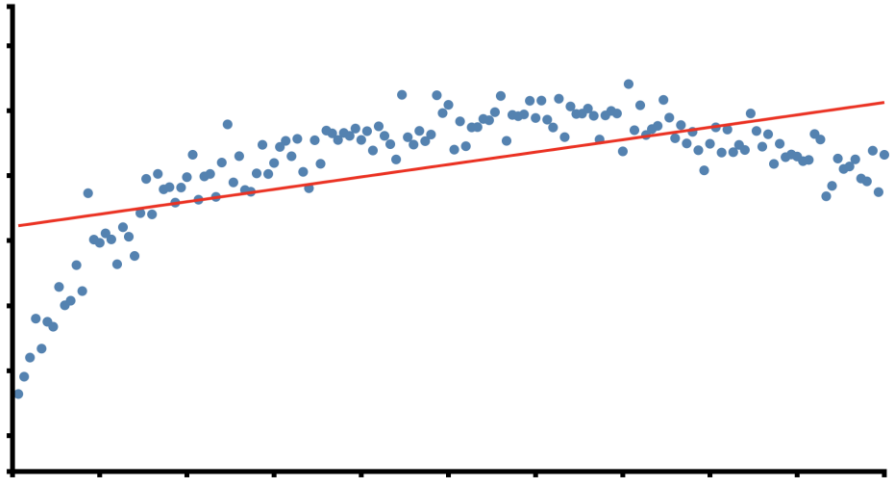
2 distributions: normal mutant and hypermutant phenotypes

1 highly skewed distribution

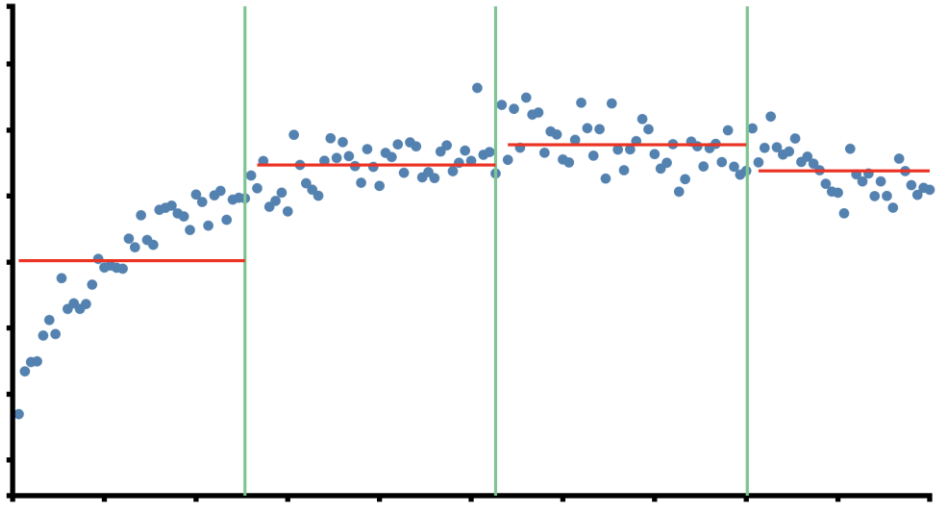
Gaussian Mixed Models can be used to identify subpopulations of interest. It acts independently from Y

Cook J, Melloni G *et al.* , Nat Comm , 2021

What if the relationship is not linear?

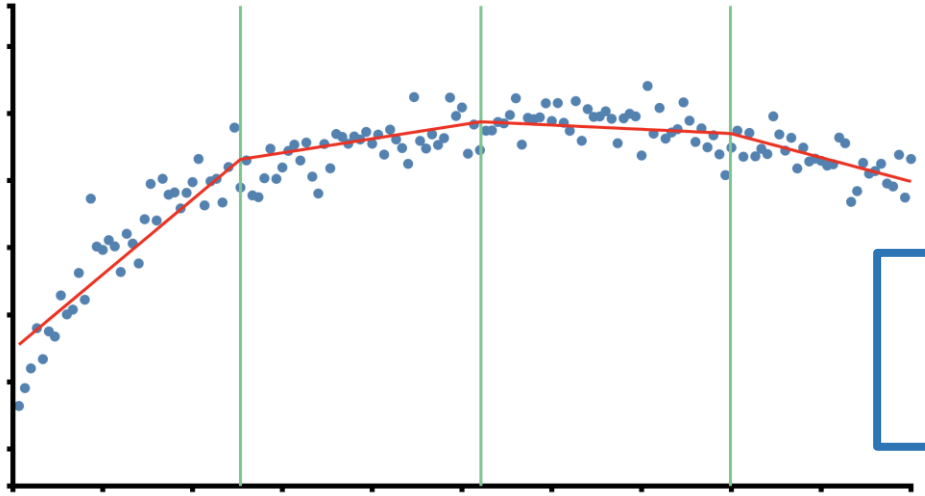


Linear regression
R-squared = 0.365
N. of parameters = 2



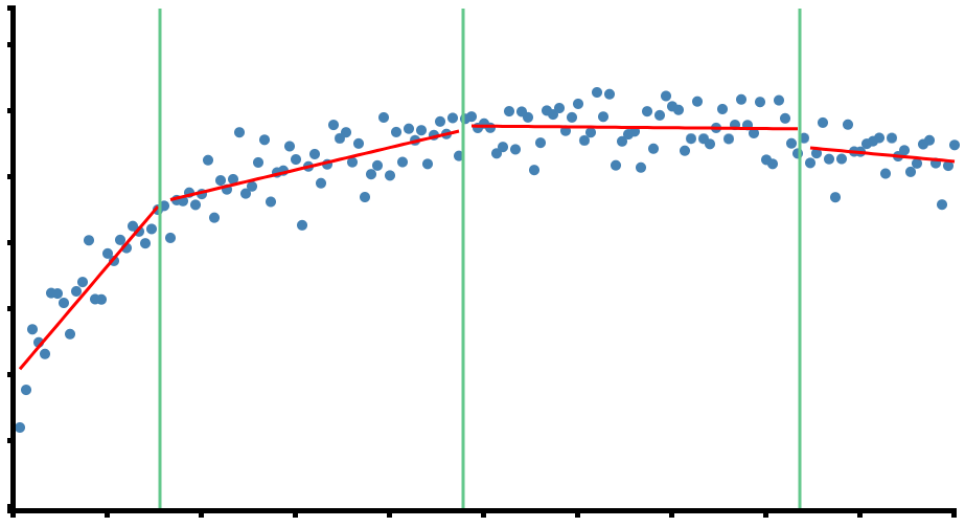
Quartile step function
R-squared = 0.597
N. of parameters = 4

COMBINE at the cost of 1 DF



Linear spline
R-squared = 0.882
N. of parameters = 5

NON-LINEAR RELATIONSHIP

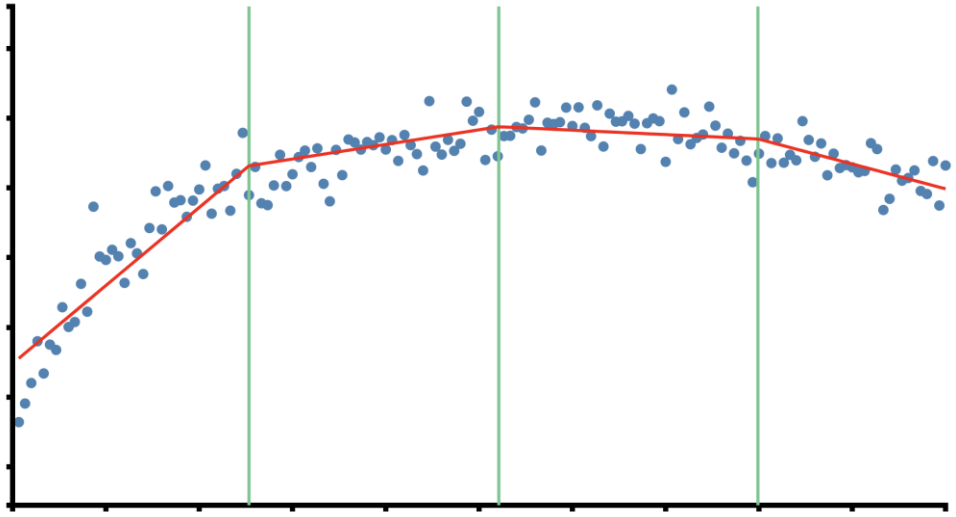


Linear spline without continuity
R-squared = 0.883
N. of parameters = 8

We have defined a better fit for our data at a minimal cost in terms of parameters, but…

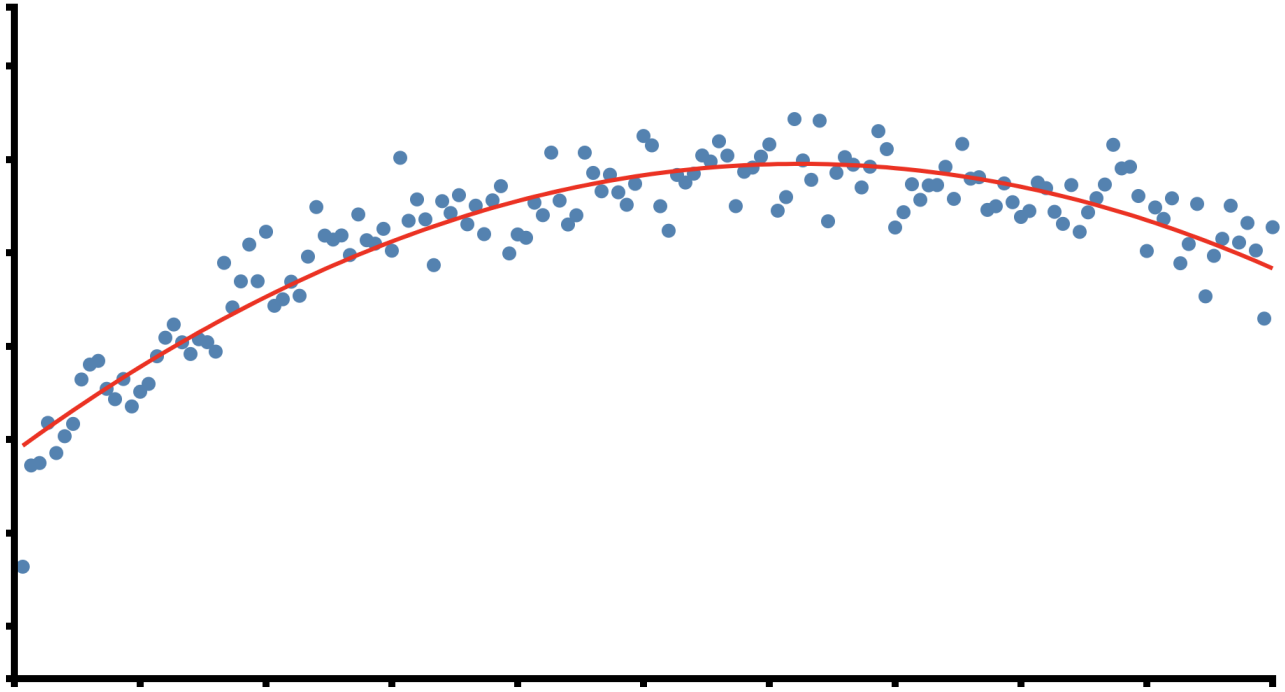
We are still heavily dependent on cutpoints

We also have to introduce a new constraint of **continuity at cutpoints** to avoid another step function (lines join at cutpoints)



Linear spline with continuity
R-squared = 0.882
N. of parameters = 5

NON-LINEAR RELATIONSHIP



Quadratic Regression
R-squared = 0.865
N. of parameters = 3

Instead of acting on cutpoints, we can add a quadratic term to our regression:

$$C(Y|X_1) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2.$$

$$H_0 : \beta_2 = 0.$$

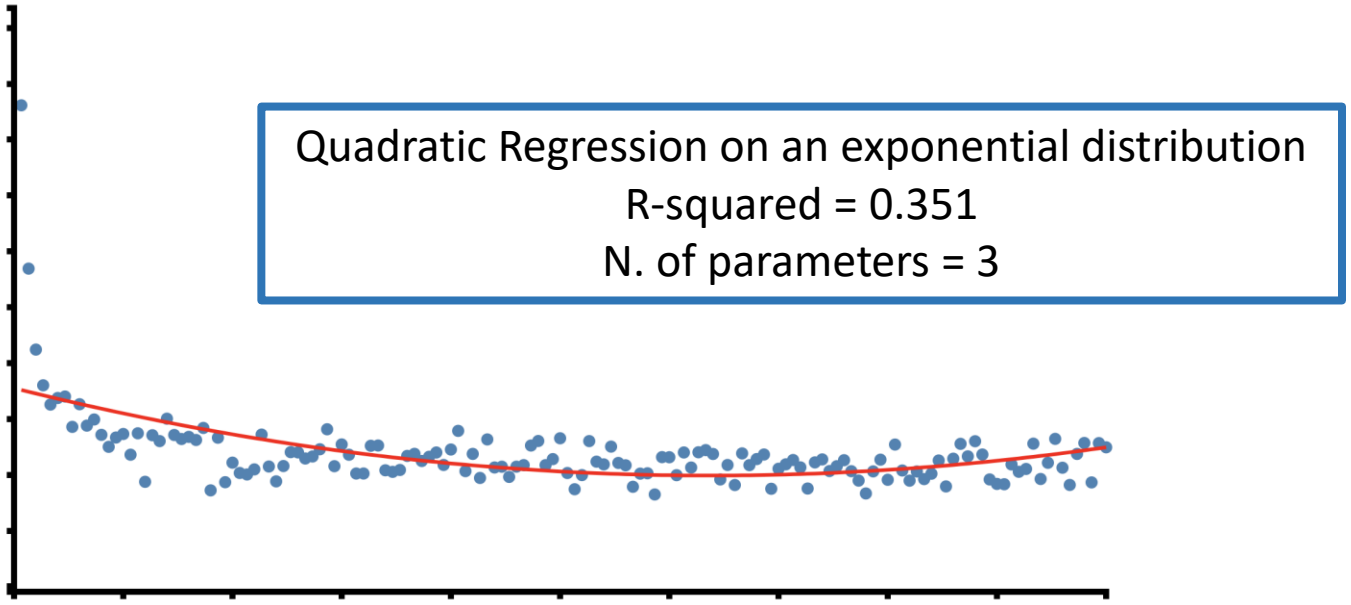
PROS

- Less parameters
- High R-squared
- We can easily test if the model is quadratic in X

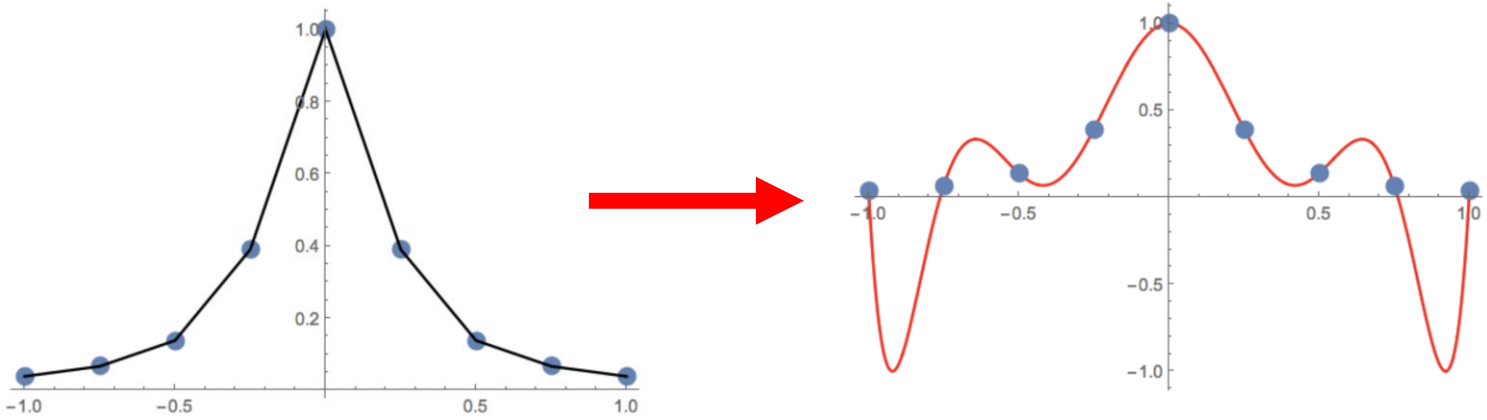
CONS

- Quadratic or cubic terms are hard to justify
- Polynomials are flexibles but they behave strangely at the extremes
- Polynomials can't account for exponential/logarithmic behaviors

NON-LINEAR RELATIONSHIP



Polynomials can't account for exponential/logarithmic behaviors

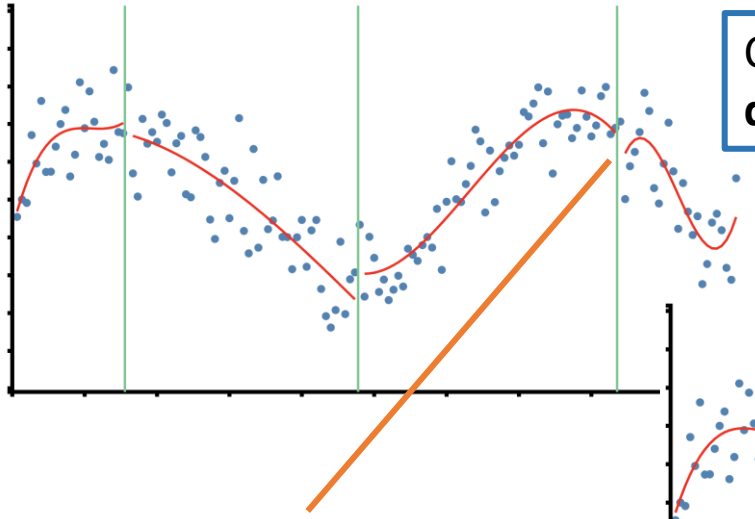


To force a single polynomial to pass by all the points we are left with a lot of unwanted "wiggling" that leads to poor accuracy of estimates

This is called **Runge's Phenomenon**

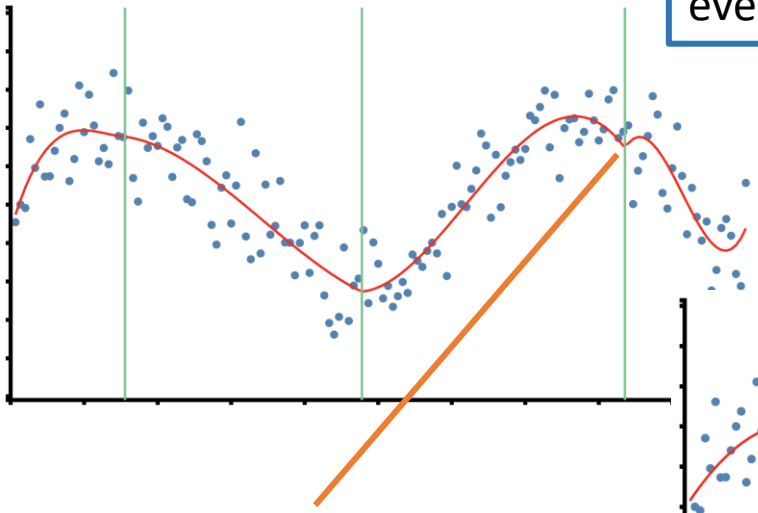
NON-LINEAR RELATIONSHIP

To obtain maximum flexibility at a minimal cost in terms of parameters, we need to combine polynomials at cutpoints (**Knots** in spline jargon) to create **splines: polynomial functions defined piecewise**



No Gaps

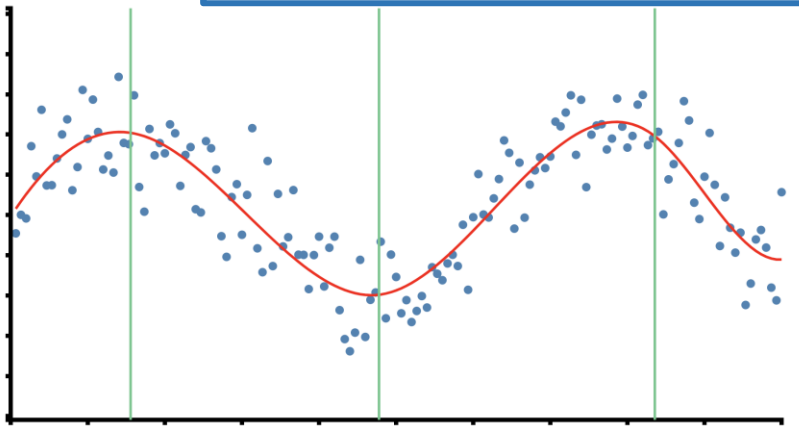
Continuity at knots' junctions is desirable, **0^{th} derivative continuity**



No sudden change in slope

The slope should be continue in every point, **1^{st} derivative continuity**

Because we generally use cubic terms, the slope of the slope should also be continue at every point, **2^{nd} derivative continuity**



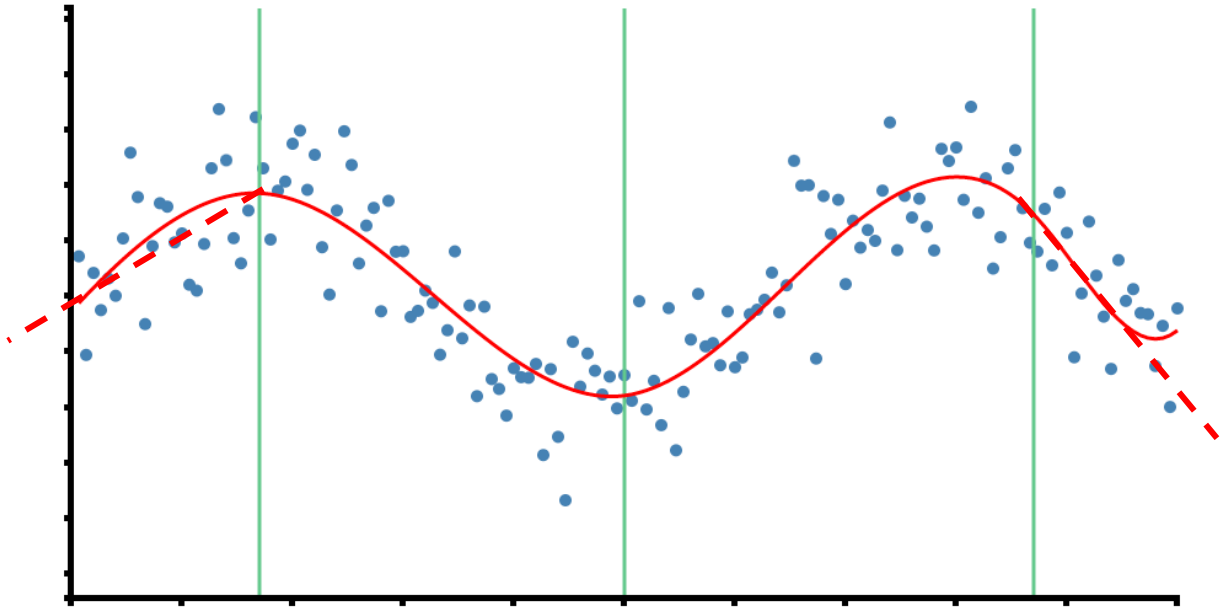
NON-LINEAR RELATIONSHIP

The most simple and widely used spline type is a **Cubic Spline**, defined by two sets of parameters (**DF**):

- A polynomial of **degree** 3 (*d* = 3) defined piecewise
  - Possible values can be 1 (linear), 2 (quadratic) or 3 (cubic). Prefer 3 if N > DF, 2 or 1 otherwise
  - Over degree 3, you almost never obtain a significantly better fit (you just add more knots instead)
  - Degree 3 with 2nd derivative continuity allows for smooth curves with no detectable change of slope

- Two boundaries **knots** and a certain number of internal **knots** (total: *k)*
  - Positioning of the internal knots has been shown to not be particularly important
  - Knots are generally equispaced [quantiles] or placed according to data density

- ***DF = knots + degree = k + d***



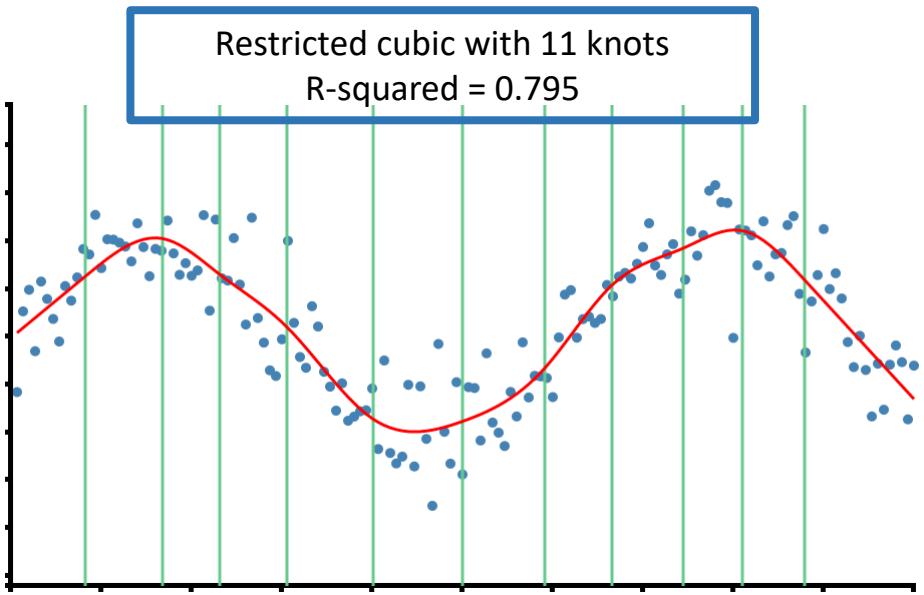
If we assume linearity after the last knot and before the first knot, we free 4 DF that can be spent in knots and avoid the erratic behavior of polynomials with no constrain. **Estimation accuracy is generally improved**

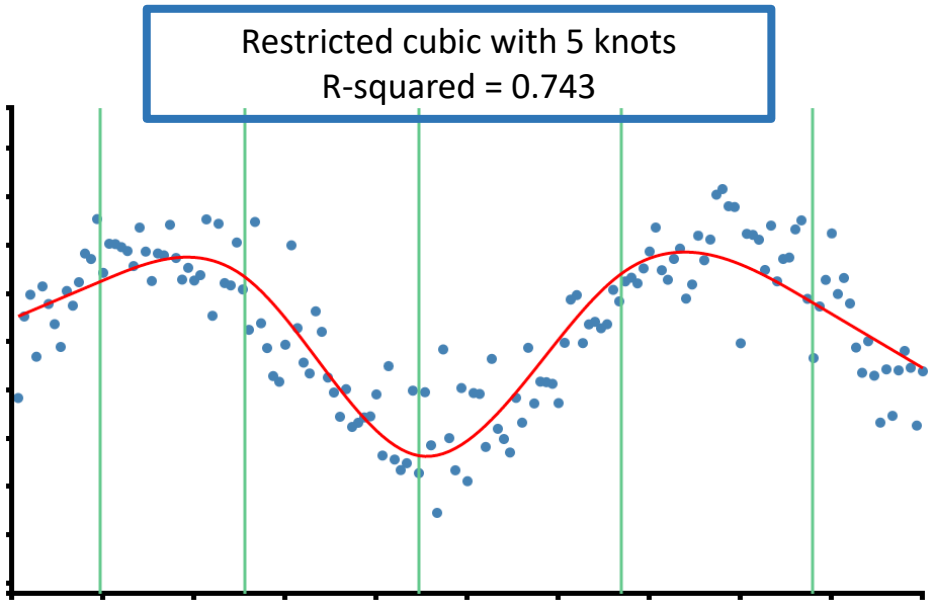
This is called a **restricted cubic spline** or **natural cubic spline**

If the degree is fixed (let's say 3), the only other decision is
how many knots should be used and where should we put them?

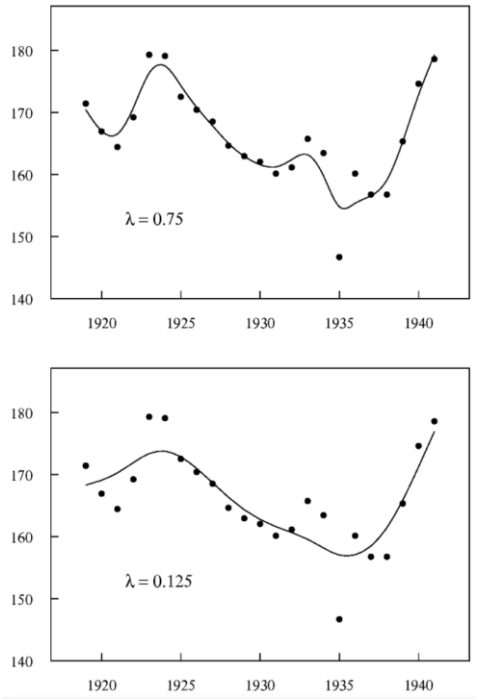
k	Quantiles						
3			.10	.5	.90		
4			.05	.35	.65	.95	
5		.05	.275	.5	.725	.95	
6	.05	.23	.41	.59	.77	.95	
7	.025	.1833	.3417	.5	.6583	.8167	.975

- **Boundaries knots** are the most important.
  - Estimation at the extreme, like we have seen for categorization, is the most erratic
  - We generally place them at 0.25/0.75 quantiles (default for Stata and R) but distribution density might be important if data is sparse
- **The number of knots is important** but after k = 5 the fit will not improve significantly (try 2,3,4,5).
  - The lower is N, the lower *k* should be. Rule of thumb *k < log(N)*
  - In most situation, *k* = 3/4 would do the trick





If the degree is fixed (let's say 3), the only other decision is
how many knots should be used and where should we put them?



- **Internal knots** can be positioned equispaced or according to data density
  - Their position doesn't matter in most situations (S. Durrleman and R. Simon, Stat Med, 1989)
  - Some procedures simply shift the problem on the **smoothness** of the curve by defining a λ parameter that controls how **wiggling** the curve will be. *k* is estimated based on that

Regular OLS on Taylor's expansion of X

Penalization on the 2nd derivative (**roughness/wigglingness**)

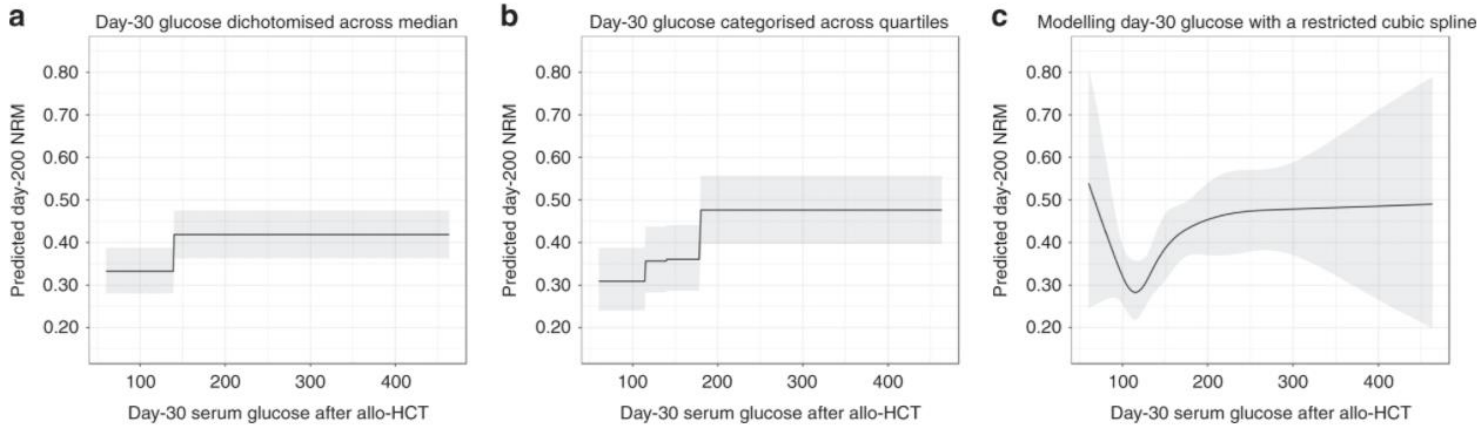
$$RSS(f,\lambda) = \sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \int [f''(t)]^2 dt$$

This is a classic penalization problem (like Ridge/Lasso regression) where we try to fit the "signal" and avoid overfitting the "noise". **λ is our noise gate**

An optimum between least square error and penalization can be found automatically using cross-validation and AIC

NON-LINEAR RELATIONSHIP

Post HTC (Hematopoietic Cell Transplantation) glucose levels and 200-day NRM risk (non-relapse mortality). A logistic model.



Median Split

Quartiles

RCS with 4 knots at quartiles

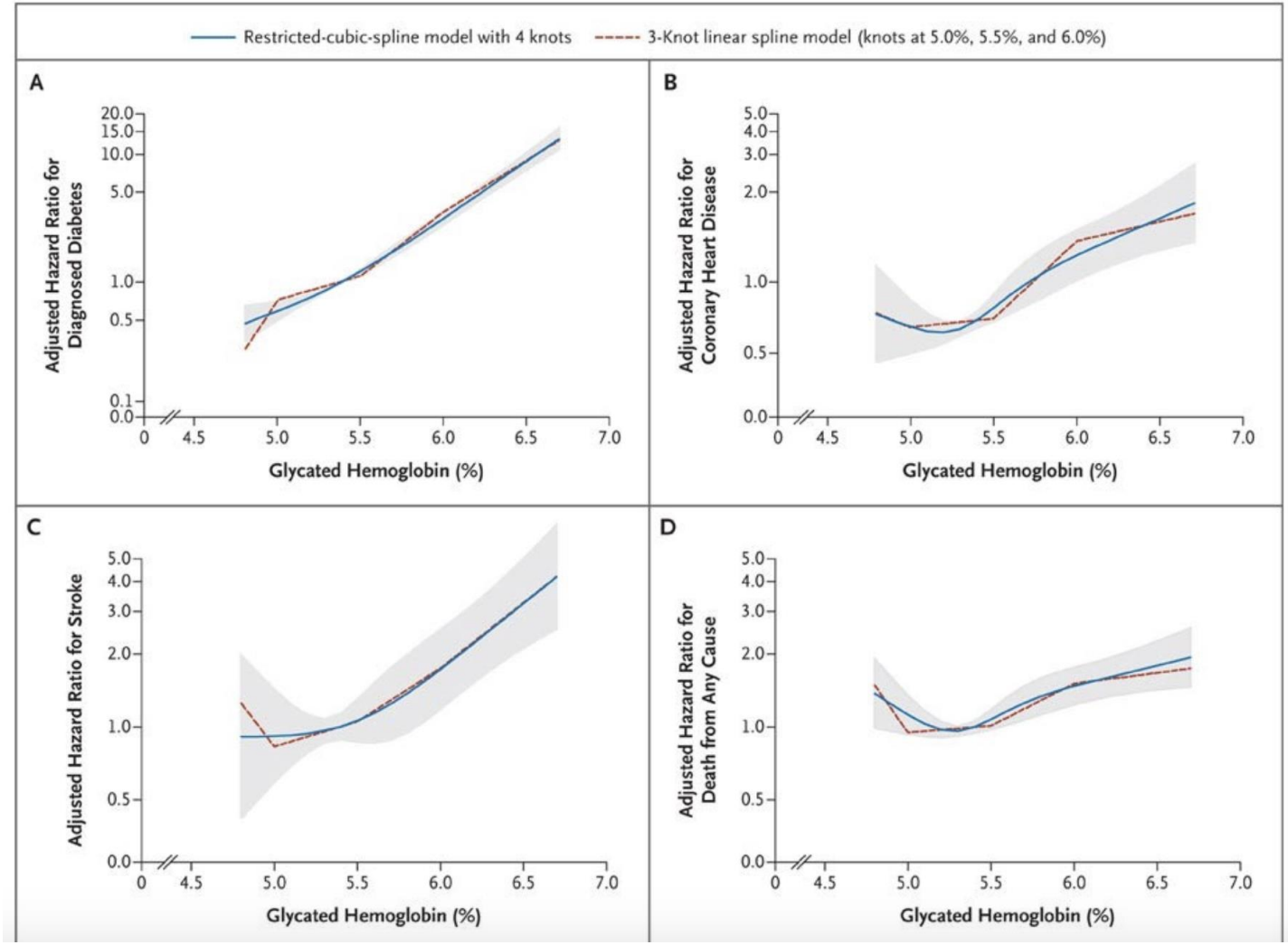
Use of splines to identify non-linear effects

Modelling glucose with the restricted cubic spline suggests a strongly non-linear relationship between glucose and NRM and allows potentially unique predictions for any glucose concentration.

As expected, extreme values (both very low and very high) are associated with an increase in the risk of NRM, while a lower risk is predicted for intermediate values

Gauthier *et al.* , Bone Marrow Transplantation, 2020

Glycated Hemoglobin compared to fasting glucose as a marker for diabetes



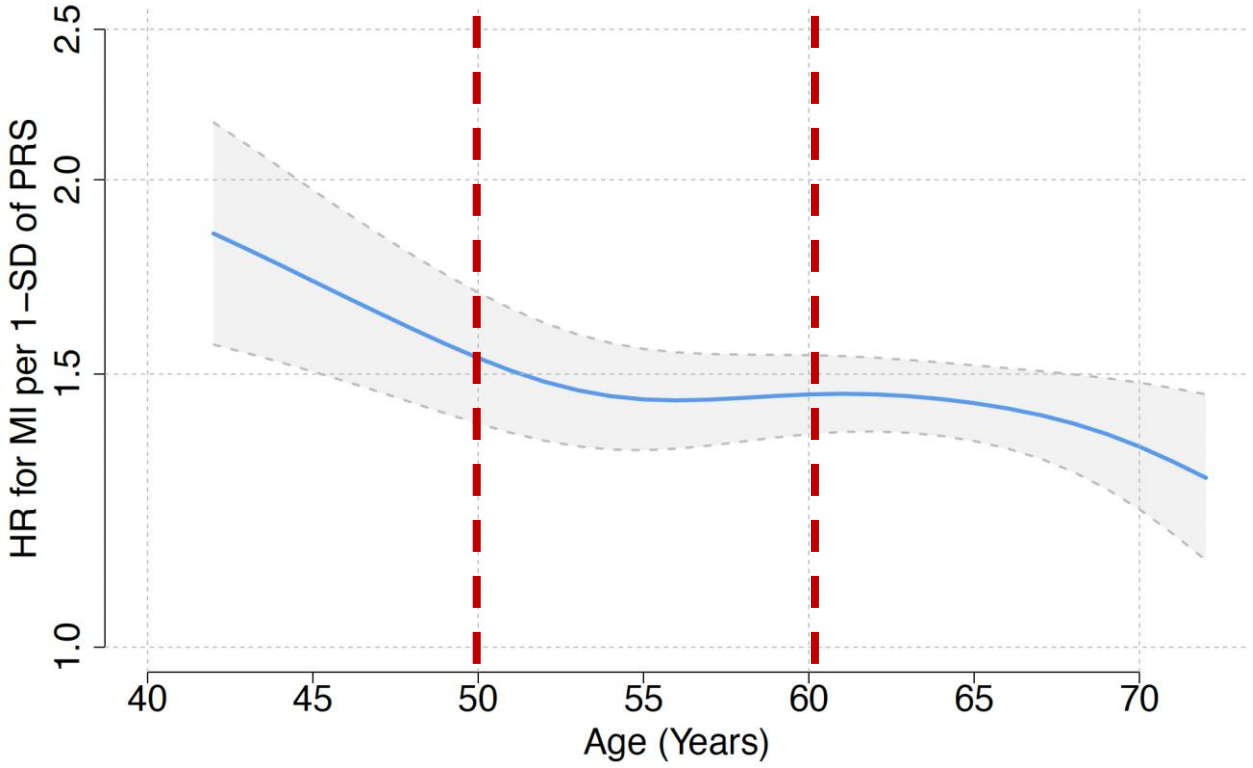
Use of splines to define possible threshold values: an example

There was no evidence of a threshold value of glycated hemoglobin for diagnosed diabetes, but there was evidence for a possible threshold for the risk of coronary heart disease

For death from any cause, we observed a J-shaped association. Participants with glycated hemoglobin values in the lowest category (<5.0%) had a significantly higher risk of death from any cause as compared with those with glycated hemoglobin levels of 5.0 to less than 5.5%

Selvin *et al.* , NEJM, 2010

Genetic risk of MI by age



Use of splines to define possible threshold values: example 2

We created a spline of HR values by 1-SD increase of PRS as a function of age

Risk looks flat after around 50-55 years of age (change of slope) and decreases again after ~60yr

We used Age <50, Age 50-60 and Age > 60 for subsequent sensitivity analyses

This plot was generated using the R package **interactionHR**, created by Andrea and me here at TIMI. It's a tool for accurate estimates of HR in spline regression with interaction terms!

Try it out at
https://github.com/gmelloni/interactionHR

Marston *et al.* , under revision at Circ Gen

When we don’t know the distribution of risk according to a risk predictor, **categorization** is often the first choice for its simplicity

The cost of simplification is high:

- Loss of power and efficiency
- Arbitrary cutpoints decision
- Hard to reproduce in a different setting or different dataset
- Optimal cutpoints don’t account for multiple hypothesis testing
- Unreasonable assumption of discontinuity (*Natura non facit saltus*)
  - E.g. what is the real difference between a 29.9 BMI and a 30.1 BMI? If we categorize, a lot, in reality, very little

CONCLUSION 2

When we don't know the distribution of risk according to a risk predictor, **using a spline regression** is a better choice

PROS

- Increased power to detect non-linear trends
- Estimates possible on any point of the distribution and for any comparison
- Graphically more appealing and faithful to the data
- Useful to find actionable cutpoints

CONS

- The model is defined piecewise (no simple explanation)
- The model requires quadratic and cubic terms that are hard to justify
- The functional form is not easy to transport to a different setting
  - E.g. create a score based on multiple parameters using splines is not trivial
- There is no formal test for non-linearity. Reporting the p-value of the 2nd and 3rd degree order of X between internal knots is the most common way to describe non-linear effects.

AKNOWLEDGEMENTS

Special thanks to

TIMI Genetics team
(Fred, Nick, Christian)

TIMI Stats team
(Sabina, Kelly, Julia, Erica, Jeong-Gun, Michael, Andrea, Jianping)

THANK YOU ALL FOR THE ATTENTION