Parametric Statistics

Week 11 - Advanced Regression

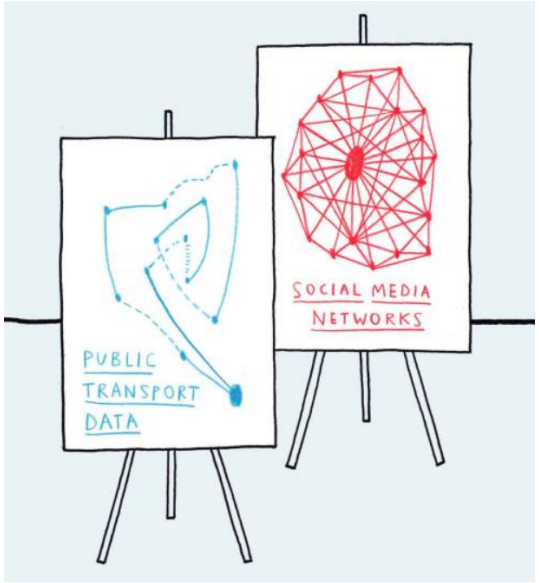
Juan Carlos Medina Serrano

Technische Universität München
Hochschule für Politik
Political Data Science

Munich, 22. January 2019

political data science

https://politicaldatascience.blogspot.de



Complex Models

What if our model is not good enough? How can we add complexity?

Interaction Effects

The first thing we can do is find out if there is some interaction between predictors.

Until now:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

In this case, the increase of one unit of X_1 signifies an increase of β_1 of Y. **Regardless of the value of** X_2

We extend the model to:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 * X_2$$

Where $X_1 * X_2$ is an interaction term. Now the change in X_2 will have an effect in the impact of X_1 on $\hat{Y}$

Interaction Effects

Check for significance and use model selection as before!
Note: If we include an interaction in our model, we should also include the single predictors (Even if they are not significant alone)

Only include interactions that make sense! You may have some theory of why the two predictors have some relationship

Example: In the headache level example, drinking whisky and vodka will increase my headache more than if I just stick to whisky or only drink vodka. Mixing alcohol can be bad for the next day! We need to add this general knowledge into our model!

$$Headache = \beta_0 + \beta_1 Vodka + \beta_2 Whisky + \beta_3 Vodka * Whisky$$

Polynomial Regression

If the relation between variables is not linear, we can add non-linear terms:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + ...$$

This is called polynomial regression. The higher the degree of the polynomial, the more non-linear the fit will be.

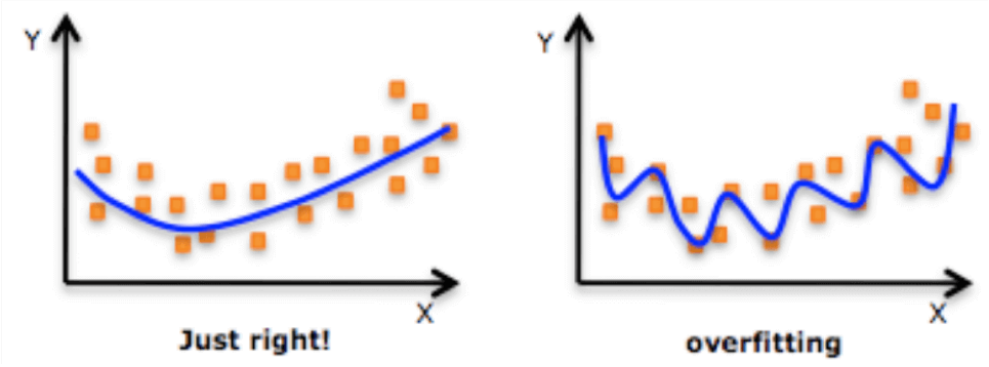
Works the same as multiple linear regression to obtain the β coefficients.

What happens when we include higher order polynomials?

Overfitting

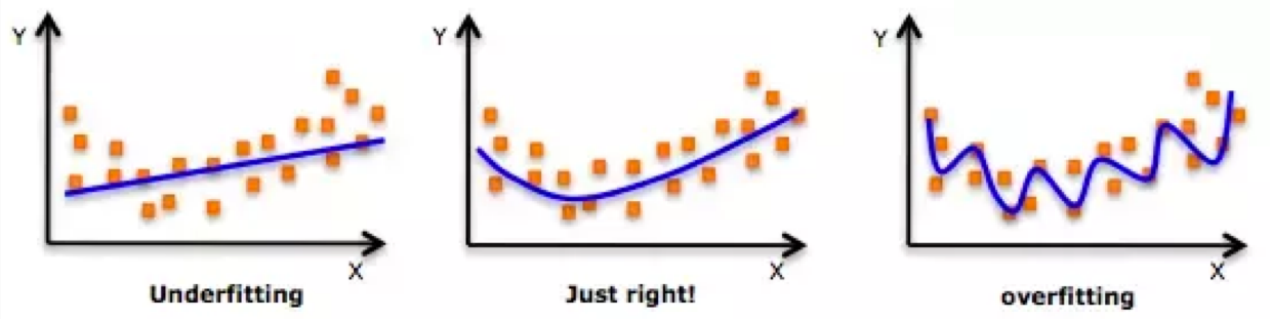
What happens when we include higher order polynomials?

It can lead to **overfitting**! High degree means an increase in variability = many curves:



Underfitting

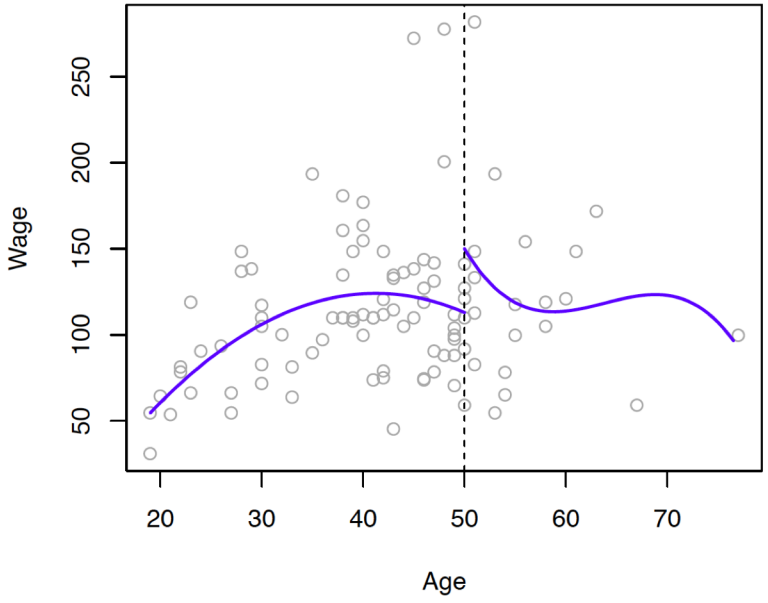
On the other extreme, if the model is too simple, we will loose important information.



Splines

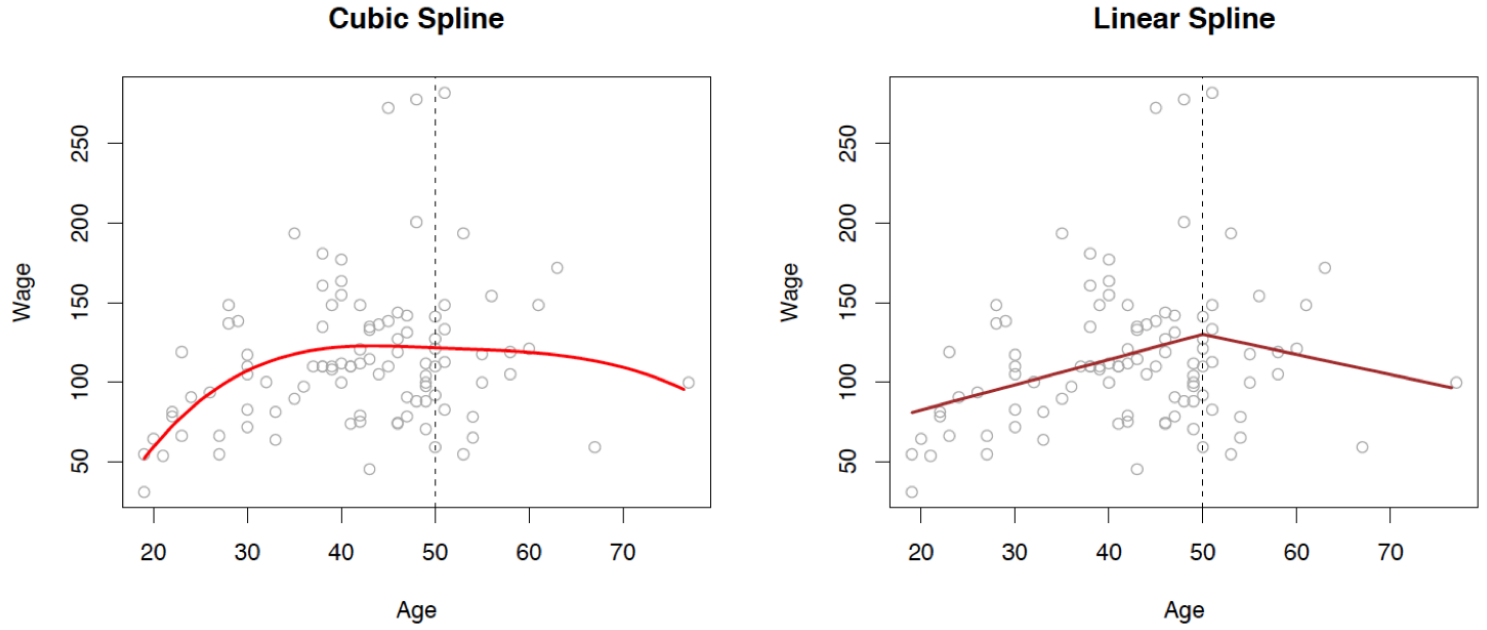
There are many more complex models (and also smarter) than simple polynomial regression.
For example, splines:

Idea: Divide the predictor X in more than one "bins". In each bin, fit a polynomial:



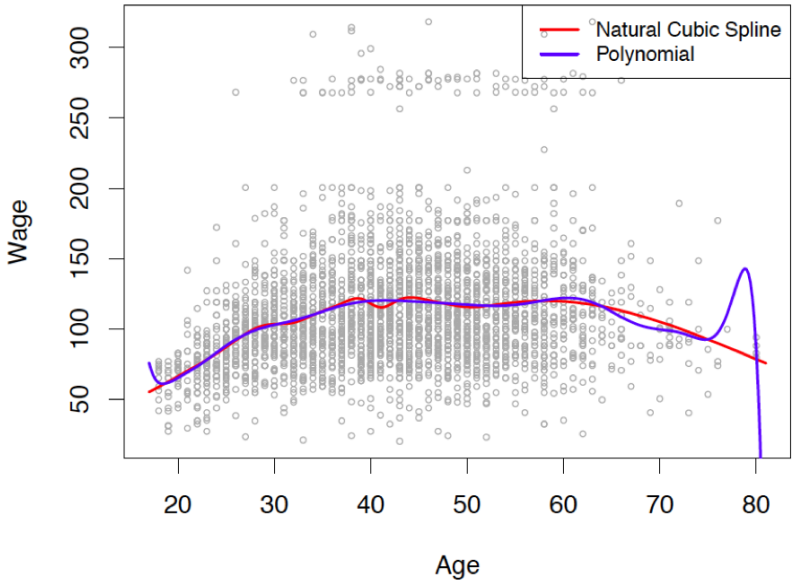
Splines

Between the "bins"there should be fixed points to have a continuous curve (knots):



Splines

Splines are often superior to polynomial regression



However: We loose interpretability! There are no β coefficients anymore to tell how much effect the predictors have on the variable.
Normal problem with such complicated models.

Literature

- Practical Statistics for Data Science: Ch 4.

For more information on complex models, I would recommend you to read: An Introduction to Statistical Learning Ch 7.