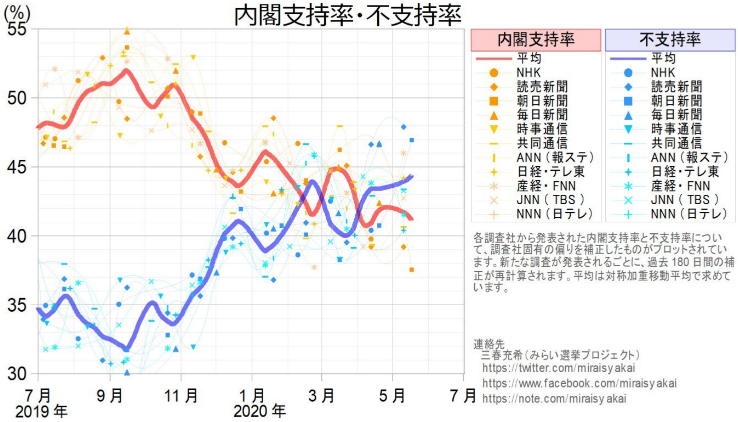
Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#331 Smoothing

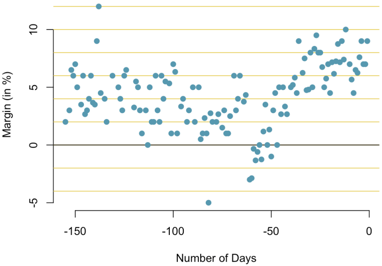
été 2020

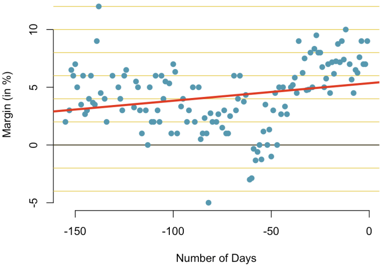
Natura non facit saltus



Natura non facit saltus

We want a continuous function... but probably not linear...
Data source: http://www.pollster.com/08USPresGEMvO-2.html
pollsters for the popular vote between Obama and McCain (2008 US presidential election), last 150 days.

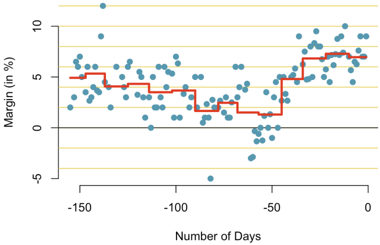


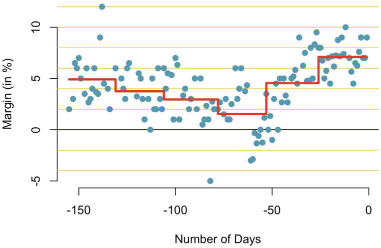


Regressogram

From Tukey (1961) Curves as parameters, and touch estimation, the regressogram is defined as

$$\hat{m}_{\boldsymbol{a}}(x) = \frac{\sum_{i=1}^{n} \mathbf{1}(x_i \in [a_j, a_{j+1}))y_i}{\sum_{i=1}^{n} \mathbf{1}(x_i \in [a_j, a_{j+1}))}$$

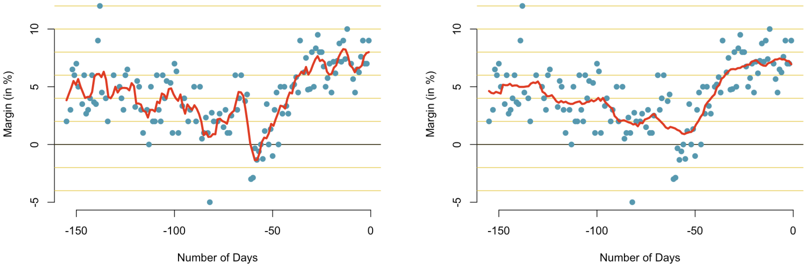




Moving Regressogram

and the moving regressogram is

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} \mathbf{1}(x_i \in [x \pm h_n]) y_i}{\sum_{i=1}^{n} \mathbf{1}(x_i \in [x \pm h_n])}$$



with bandwidth h_n (size of the neighborhood around x)

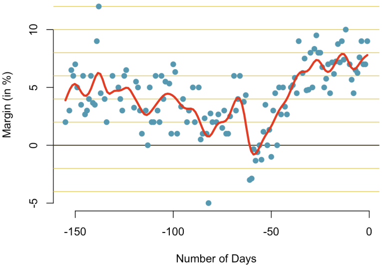
Local Regression

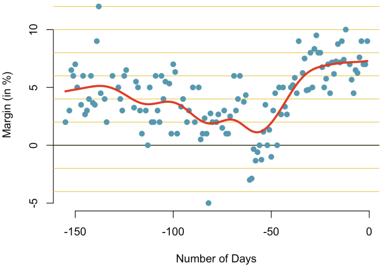
More generally, as moving from the histogram to kernel estimate

$$\tilde{m}(x) = \frac{\sum_{i=1}^{n} y_i \kappa_h(x - x_i)}{\sum_{i=1}^{n} \kappa_h(x - x_i)}$$

Observe that this regression estimator is a weighted average

$$\tilde{m}(x) = \sum_{i=1}^{n} \omega_i(x) y_i \text{ with } \omega_i(x) = \frac{\kappa_h(x - x_i)}{\sum_{i=1}^{n} \kappa_h(x - x_i)}$$





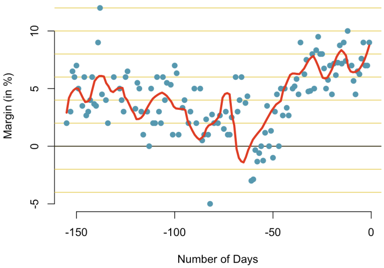
k-Nearest Neighbors

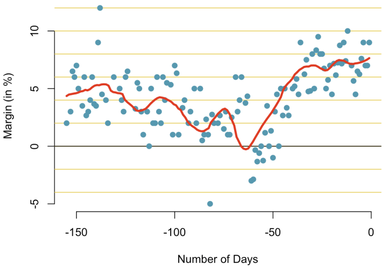
An alternative is to consider

$$\tilde{m}_k(x) = \frac{1}{n}\sum_{i=1}^{n} \omega_{i,k}(x) y_i$$

where $\omega_{i,k}(x) = \dfrac{n}{k}$ if $i \in \mathcal{I}_x^k$ with

$$\mathcal{I}_x^k = \{i : x_i \text{ one of the } k \text{ nearest observations to } x\}$$

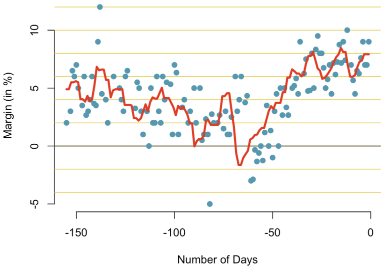


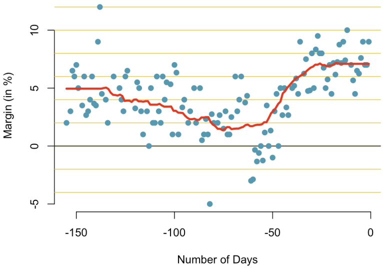


LOESS (locally weighted polynomial) Regression

Solve

$$\tilde{m}(x) = \text{argmin}\left\{\sum_{i=1}^{n} \omega_i(x)\big(y_i - \alpha - \beta x_i\big)^2\right\},\ \omega_i(x) = \frac{\kappa_h\left(x - x_i\right)}{\sum_{i=1}^{n} \kappa_h\left(x - x_i\right)}$$



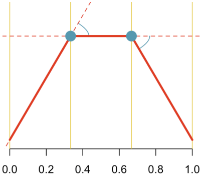


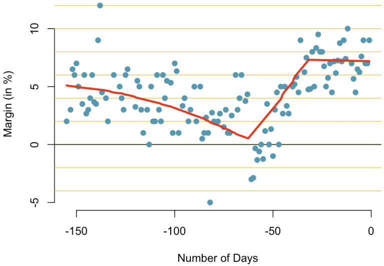
(Linear) Spline Regression

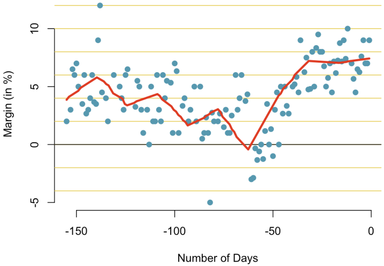
Select some knots $\{s_1, \cdots, s_k\}$, then with $s_0 = 0$

$$\tilde{m}(x) = \alpha + \sum_{j=0}^{k} \beta_j (x - s_k)_+$$

where $(x - s)_+ = (x - s)$ if $x > s$, 0 otherwise





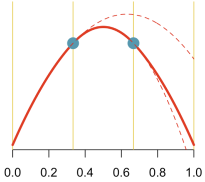


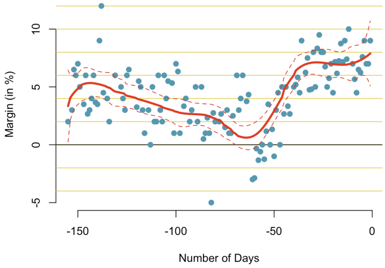
(Quadratic) Spline Regression

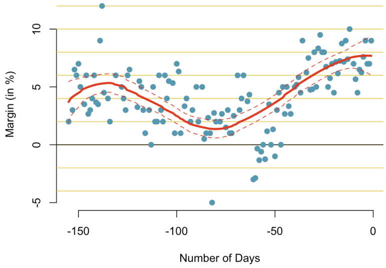
Select some knots $\{s_1, \cdots, s_k\}$, then with $s_0 = 0$

$$\tilde{m}(x) = \alpha + \gamma x + \sum_{j=0}^{k} \beta_j (x - s_k)_+^2$$

where $(x - s)_+^2 = (x - s)^2$ if $x > s$, 0 otherwise

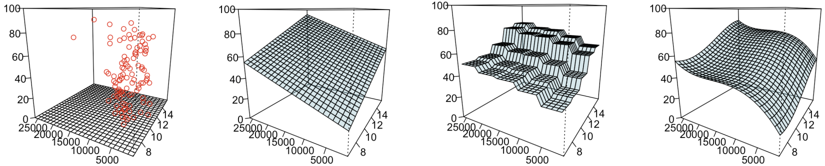






Bivariate Smoothing

Can be extended in higher dimension



from the Prestige.txt dataset