Ethics 2: Live Session 2

Zak Varty

- Summary and further exploration of the topics covered this week.

Accumulated Local Effects Plot

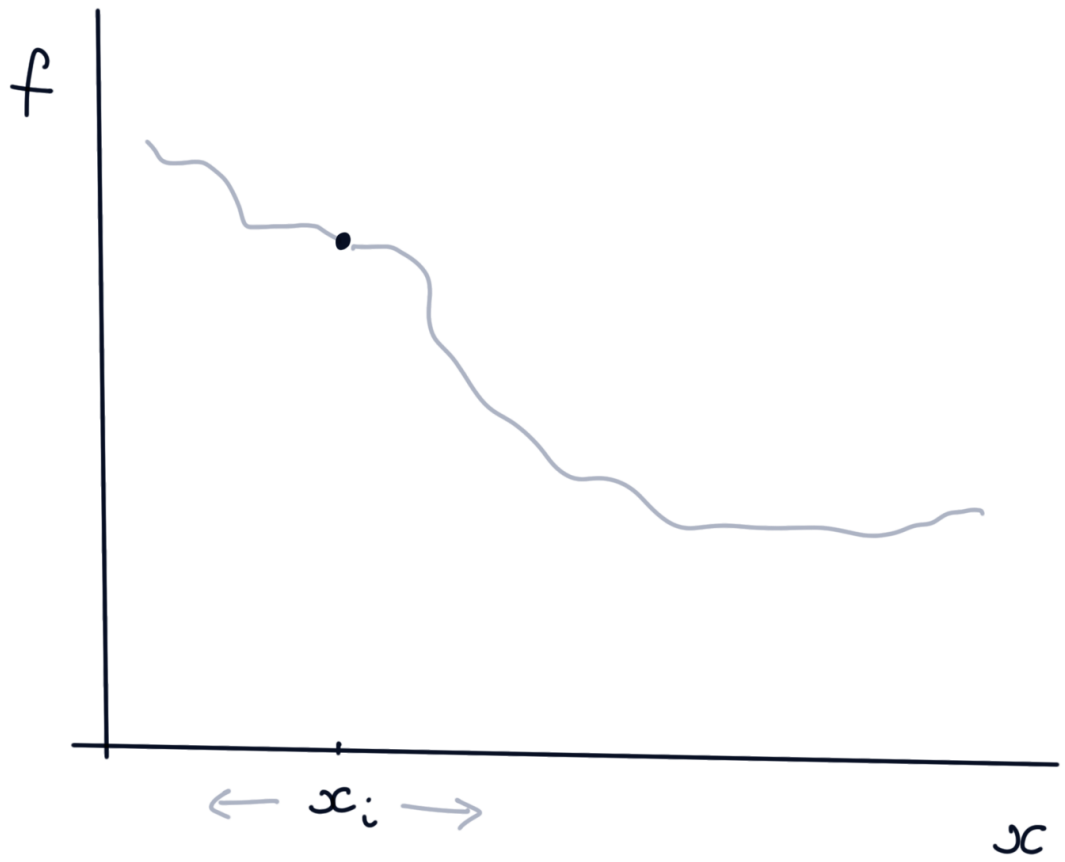
Consider a model $f(x, y, z; \theta)$ with:

- three predictors X, Y, and Z,
- fitted parameter values $\hat{\theta}$,
- estimated using on n observations: $\{(x_i, y_i, z_i), a_i\}$.

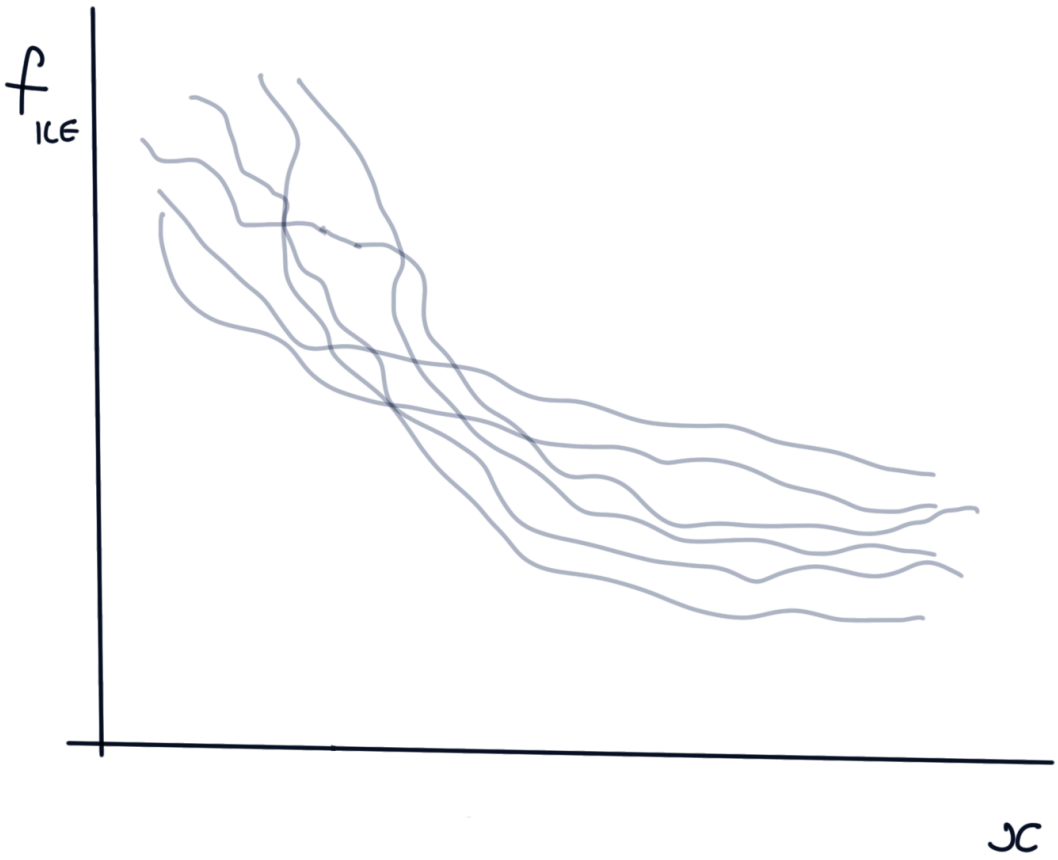
ICE plots

Individual conditional expectation plots show the counter-factual prediction of the outcome as one predictor is varied for each individual.

$$f_i(x; \hat{\theta}) = f(x, y_i, z_i; \hat{\theta}).$$

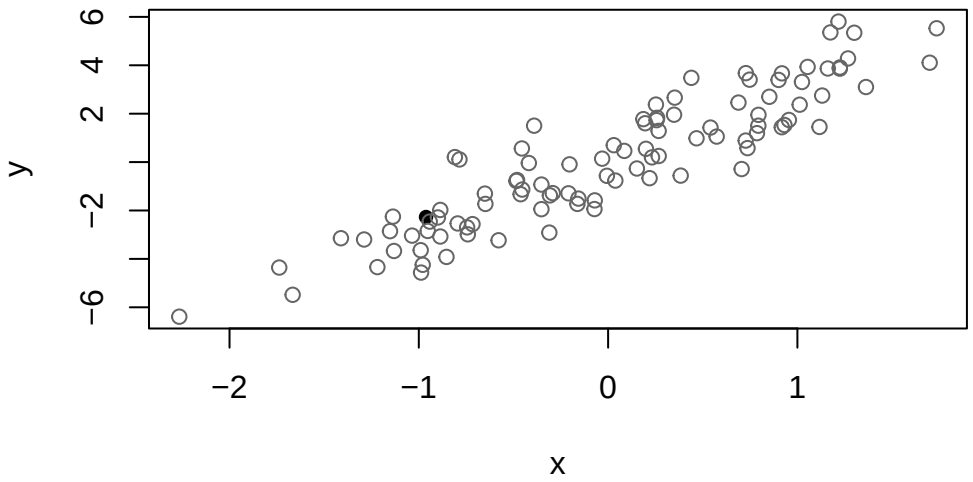


We can construct 1 such ICE plot per observation or person within our dataset.



We have to be careful about extrapolating to unreasonable combinations (x, y, z). Just because we can make a prediction doesn't mean we should.

In the example below, increasing x to 1 while keeping y fixed does not seem reasonable.



These sorts of checks rapidly get harder in higher dimensions.

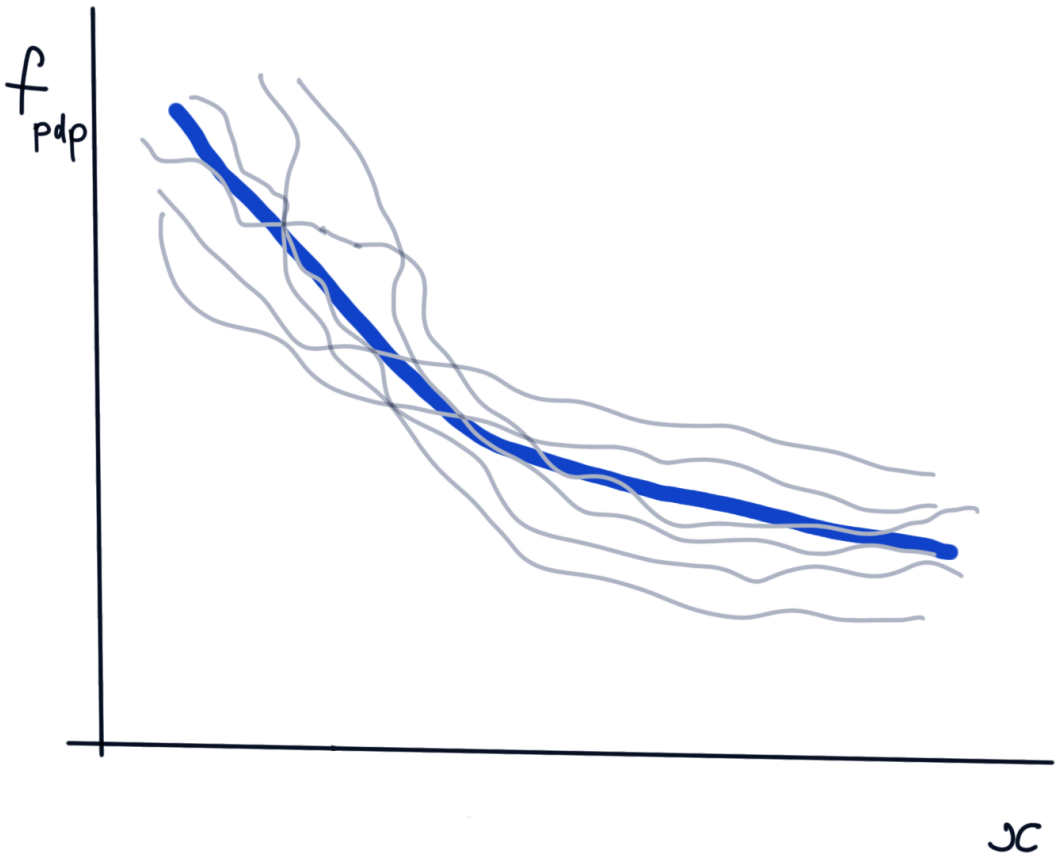
PDPs

Partial dependence plots, which you met last week show the expected response as a function of one predictor, **averaged over the values of all other covariates**.

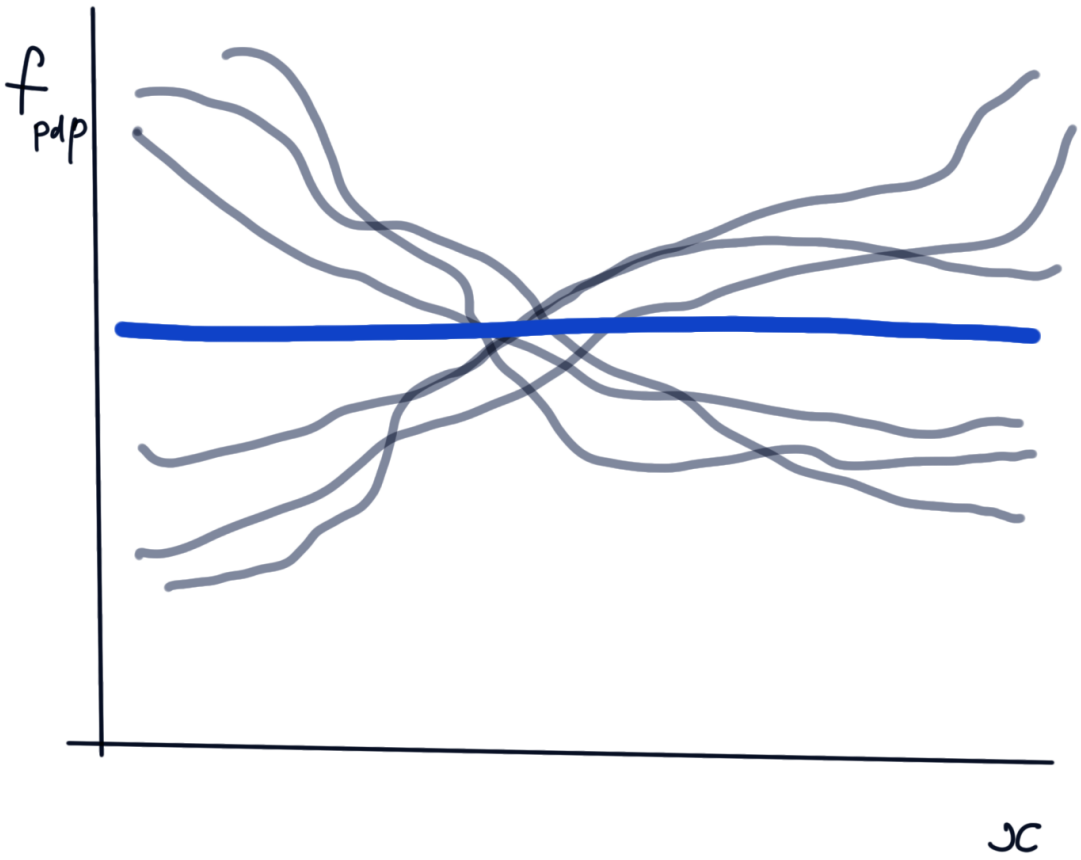
So in our example the PDP for X would be:

$$f_{pdp}(x;\hat{\theta}) = \mathbb{E}_{Y,Z}[f(x,Y,Z;\hat{\theta})] \approx \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x,y_i,z_i).$$

PDPs can be considered as the expected (or point wise mean) of the ICE curves.

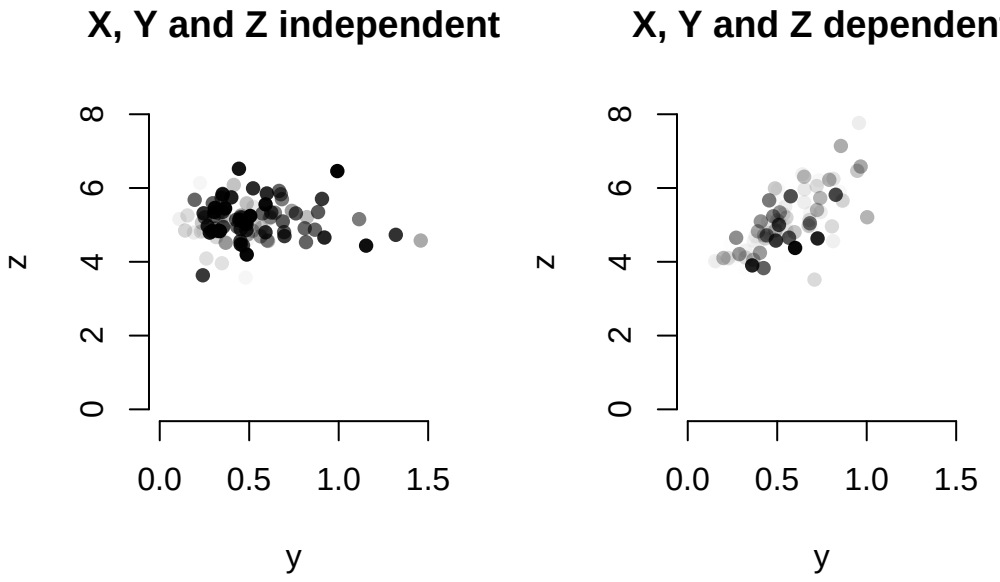


- While this describes how a response changes on average, that does not mean we expect any individual ICE cure to look anything like the PDP.



Similar to ICEs we have to be careful about extrapolating / interpolating with PDPs to low-density regions of predictor space.

If X is independent of Y and Z the above expectation makes sense, it assumes that all of the observed y and z values could be observed alongside any x value in our pdp with equal probability.



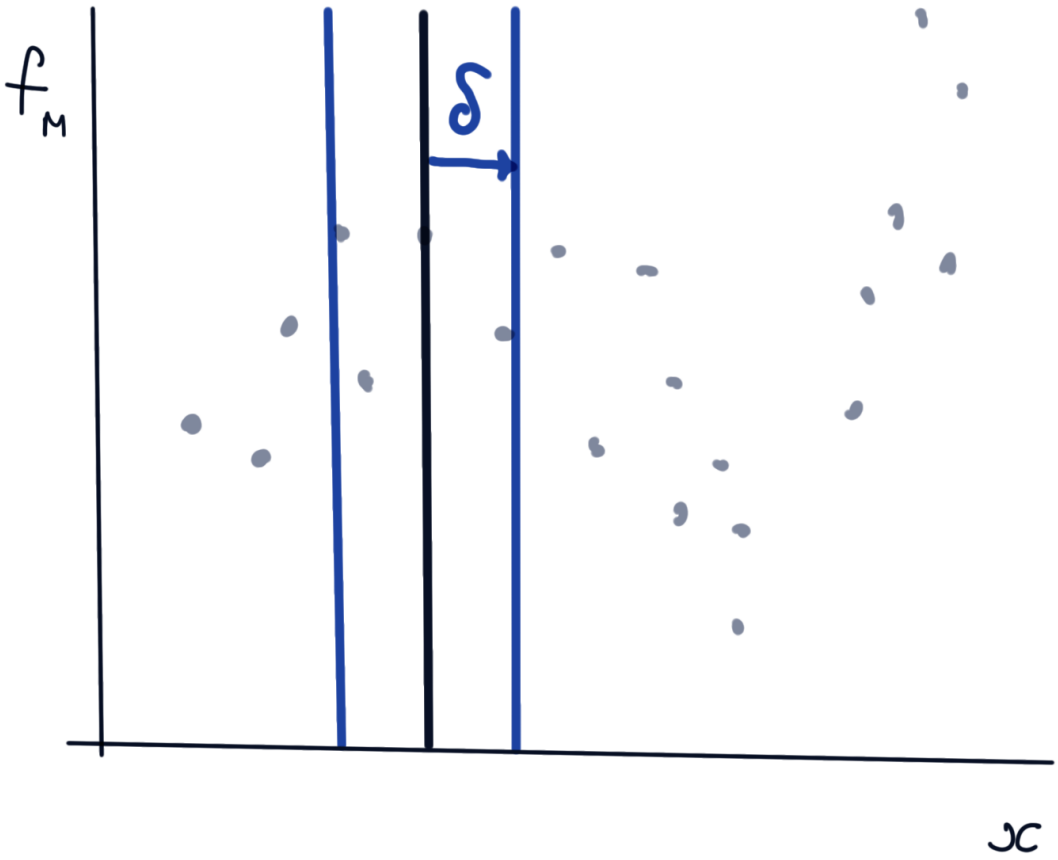
If we return to our trees example, this is clearly not the case - from context we know that a tree with very large volume is also likely to be taller.

M-plots

Accumulated Local Effect plots modify the above by taking the *conditional* expection over the remaining predictors, *given* the value of the predictor of interest.

$$f_M(x;\hat{\theta}) = \mathbb{E}_{Y,Z|X=x}[f(x,Y,Z;\hat{\theta})]$$

To evaluate this conditional expectation, we only consider observations that have X values close to each x we consider: that is those points in $B_x(\delta) = \{x' : |x' - x| < \delta\}$.

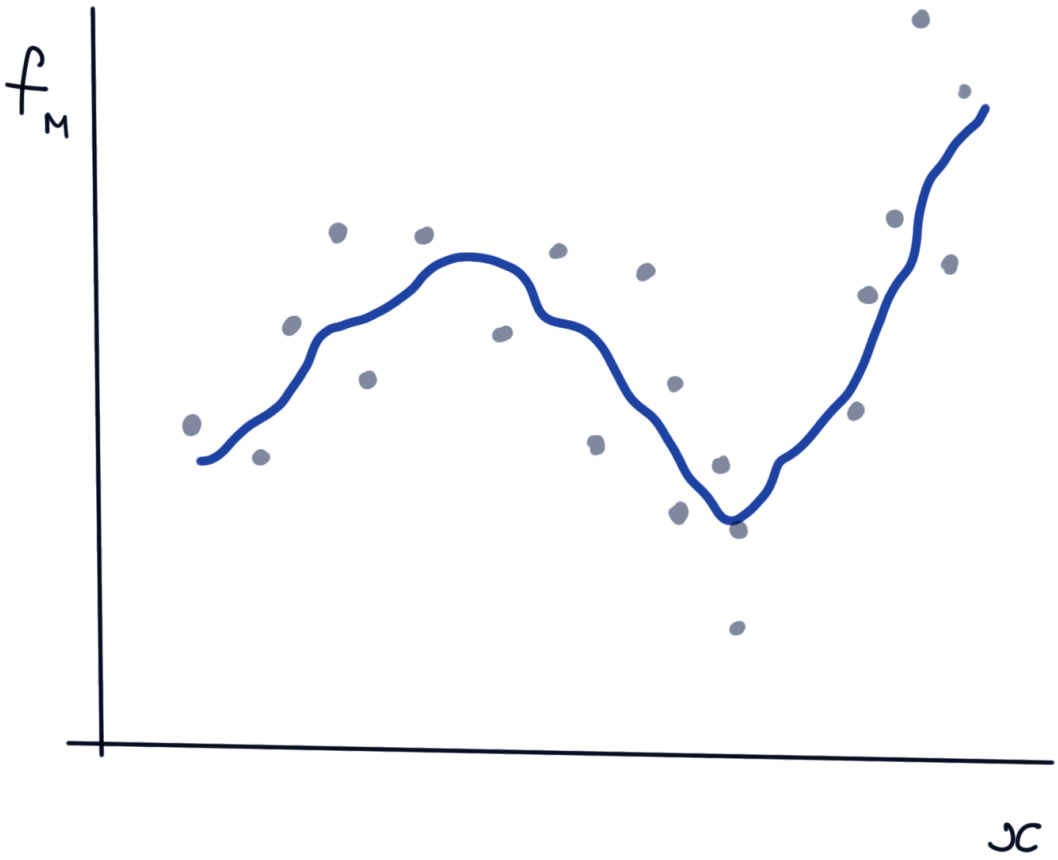


Question: What do we have to balance when picking δ?

So we approximate the conditional expectation by:

$$f_M(x;\hat{\theta}) = \mathbb{E}_{Y,Z|X=x}[f(x,Y,Z;\hat{\theta})] \approx \frac{1}{N(B_x(\delta))} \sum_{\{i:x_i \in B_x(\delta)\}} f(x,y_i,z_i;\hat{\theta}),$$

where $N(S)$ denotes the number of observations lying in subset S of the predictor space.



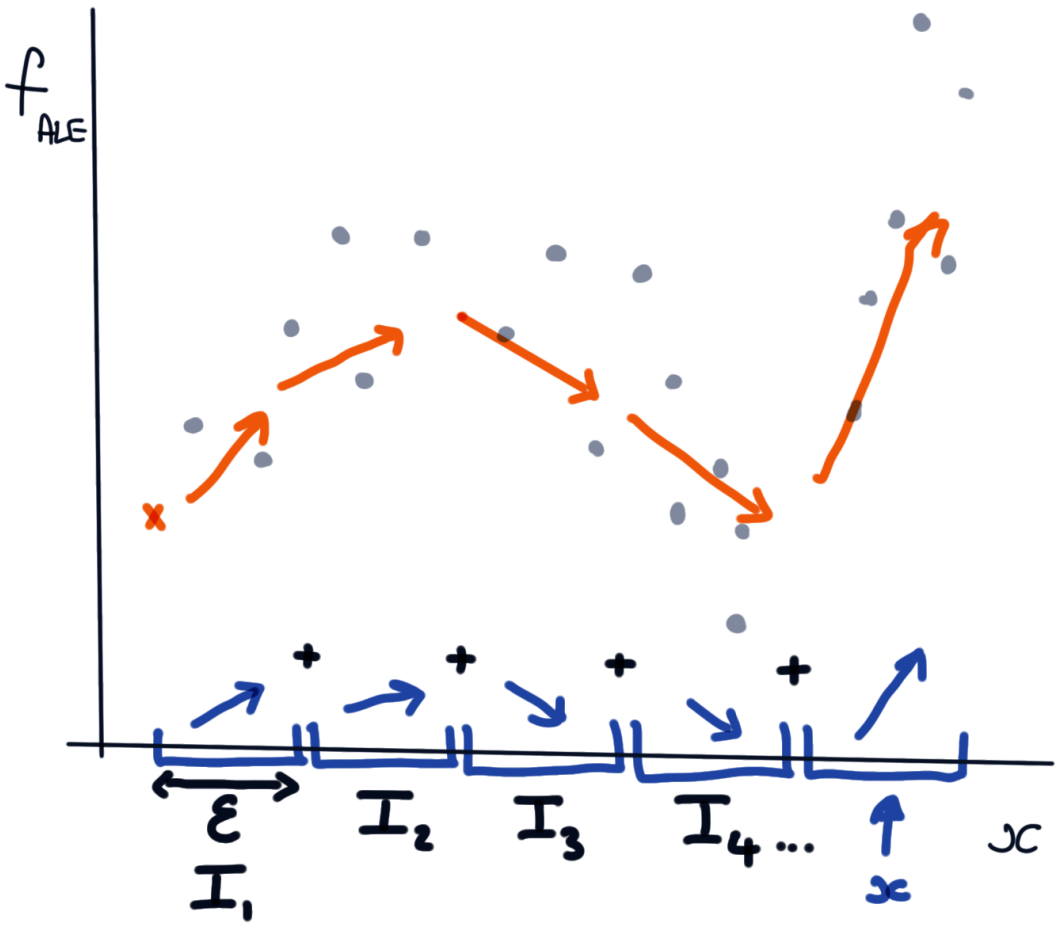
ALE plots

Rather than only basing our estimate on the small number of points that have X values close to x, we instead use all observations that have values less than x.

To do that we:

- Find $x_0 \leq \min x_i$
- Split (x_0, x) into m intervals $I_1, \ldots, I_m$ of length $\epsilon_1, \ldots, \epsilon_m$

  - standard case has all intervals of equal length
- Find the average value of $\partial f / \partial x$ using the data points in each interval
  - this gives us an estimate of how increasing x over that range changes our response
- Multiply each estimate by the interval length and add them up
  - this gives us an estimate of the cumulated effect of a particular x value on our response.



Mathematically we can express this as $J = \lceil (x - x_0)/\epsilon \rceil$ be the index of the first interval endpoint that is above x

$$f_{ALE}(x) = \sum_{j=0}^{J} \underbrace{\left(\frac{1}{N(I_{j+1})} \sum_{\{i: x_i \in I_{j+1}\}} \frac{\hat{f}(x_0 + (j+1)\epsilon, y_i, z_i) - \hat{f}(x_0 + j\epsilon, y_i, z_i)}{\epsilon} \right)}_{\text{average } \partial f/\partial x \text{ for observations in } I_{j+1}}.$$

This leads us to many similar considerations about the width of the intervals.

- If we have dense observations of x, then lots of small intervals will better approximate $\partial f/\partial x$ and so better construct the ALE plot.
- If we have few observations in one interval then we have a very noisy estimate of $\partial f/\partial x$, which effects all the ALE at all higher values.
- If we have no observations in one interval, we cannot estimate the gradient there at all!
- **Simple solution:** adaptive interval widths to maintain equal sample sizes for gradient estimation.

Predictive vs Prognostic covariates

A **prognostic** biomarker provides information about the patient's overall cancer outcome, regardless of therapy.

A **predictive** biomarker gives information about the effect of a therapeutic intervention.

Consider the following model: `outcome ~ age * sex * treatment`

Question: Which of the following terms are prognostic and which are predictive?

- `1`
- `age`
- `sex`
- `treatment`
- `age:sex`
- `age:treatment`
- `sex:treatment`
- `age:sex:treatment`

Interactions in decision trees

In general a decision tree with depth k will include k-way interaction terms.

Question: Why is this?

SHAP Quiz

Question: Can anyone explain a SHAP value?

Question: Can anyone explain how SHAP interactions extend this?

Question: Why is this difficult?