Tree-based Models for Regression

- Regression trees
- Regression forests
  - `randomForest` based on bagging
  - `gbm` based on boosting

Tree-based Models for Regression

- Input vector $X = (X_1, X_2, \ldots, X_p) \in \mathcal{X}$
- Response variable $Y \in \mathbb{R}$
- Trees are constructed by recursively splitting regions of $\mathcal{X}$ into two sub-regions, beginning with the whole space $\mathcal{X}$.

  For simplicity, focus on recursive binary partitions.
- `R` page: check the fitted regression tree on `BostonHousingData` based on two features `lon` and `lat`.

- Notation: node (t), child node (t_L, t_R), split (var j, value s), leaf/terminal node.
- Every leaf node (i.e. a rectangle region R_m in $\mathcal{X}$) is assigned with a constant for regression tree

$$\hat{f}(X) = \sum_m c_m I\{X \in R_m\}.$$

Advantages of Trees

- Easy to interpret
- Variable selection and interactions between variables are handled automatically
- Invariant under any monotone transformation of predictors