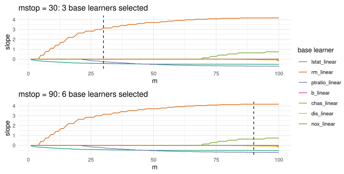
Introduction to Machine Learning

Boosting
Gradient Boosting: CWB Basics 2



Learning goals

- Handling of categorical features
- Intercept handling
- Practical example

HANDLING OF CATEGORICAL FEATURES

Feature x_j with G categories. Two options for encoding:

- One base learner to simultaneously estimate all categories:

$$b_j(x_j|\boldsymbol{\theta}_j) = \sum_{g=1}^{G} \theta_{j,g} \mathbb{1}_{\{g=x_j\}} = (\mathbb{1}_{\{x_j=1\}}, ..., \mathbb{1}_{\{x_j=G\}})\boldsymbol{\theta}_j$$

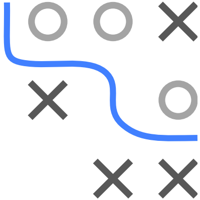
  Hence, b_j incorporates a one-hot encoded feature with group means $\boldsymbol{\theta} \in \mathbb{R}^G$ as estimators.

- One binary base learner per category:

$$b_{j,g}(x_j|\theta_{j,g}) = \theta_{j,g} \mathbb{1}_{\{g=x_j\}}$$

  Including all categories of the feature means adding G base learners $b_{j,1}, \dots, b_{j,G}$

HANDLING OF CATEGORICAL FEATURES / 2



Advantages of simultaneously handling all categories in CWB:

- Much faster estimation compared to using individual binary BLs
- Explicit solution of $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^G} \sum_{i=1}^{n} (y^{(i)} - b_j(x_j^{(i)} | \boldsymbol{\theta}))^2$:

$$\hat{\theta}_g = n_g^{-1} \sum_{i=1}^{n} y^{(i)} \mathbb{1}_{\{x_j^{(i)} = g\}}$$

- For features with many categories we usually add a ridge penalty

HANDLING OF CATEGORICAL FEATURES / 3

Advantages of including categories individually in CWB:

- Enables finer selection since non-informative categories are simply not included in the model.
- Explicit solution of $\hat{\theta}_{j,g} = \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} (y^{(i)} - b_g(x_j^{(i)} | \theta))^2$ with:

$$\hat{\theta}_{j,g} = n_g^{-1} \sum_{i=1}^{n} y^{(i)} \mathbb{1}_{\{x_j^{(i)} = g\}}$$

Disadvantage of individually handling all categories in CWB:

- Fitting CWB is slower
- Penalization and selection become difficult since base learner has exactly one degree of freedom.

INTERCEPT HANDLING

There are two options to handle the intercept in CWB. In both, the loss-optimal constant $f^{[0]}(\mathbf{x})$ is an initial model intercept.

1. Include an intercept BL:
   - Add BL $b_{\text{int}} = \theta$ as potential candidate considered in each iteration and remove intercept from all linear BLs, i.e., $b_j(\mathbf{x}) = \theta_j x_j$.
   - Final intercept is given as $f^{[0]}(\mathbf{x}) + \hat{\theta}$. Linear BLs without intercept only make sense if covariates are centered (see ▸ Hofner et al. 2014 tutorial, p. 7)
2. Include intercept in each linear BL and aggregate into global intercept post-hoc:
   - Assume linear base learners $b_j(\mathbf{x}) = \theta_{j1} + \theta_{j2} x_j$. If base learner $\hat{b}_j$ with parameter $\hat{\boldsymbol{\theta}}^{[1]} = (\hat{\theta}_{j1}^{[1]}, \hat{\theta}_{j1}^{[1]})$ is selected in first iteration, model intercept is updated to $f^{[0]}(\mathbf{x}) + \hat{\theta}_{j1}^{[1]}$.
   - During training, intercept is adjusted M times to yield $f^{[0]}(\mathbf{x}) + \sum_{m=1}^{M} \hat{\theta}_{j1}^{[m]}$

EXAMPLE: LIFE EXPECTANCY

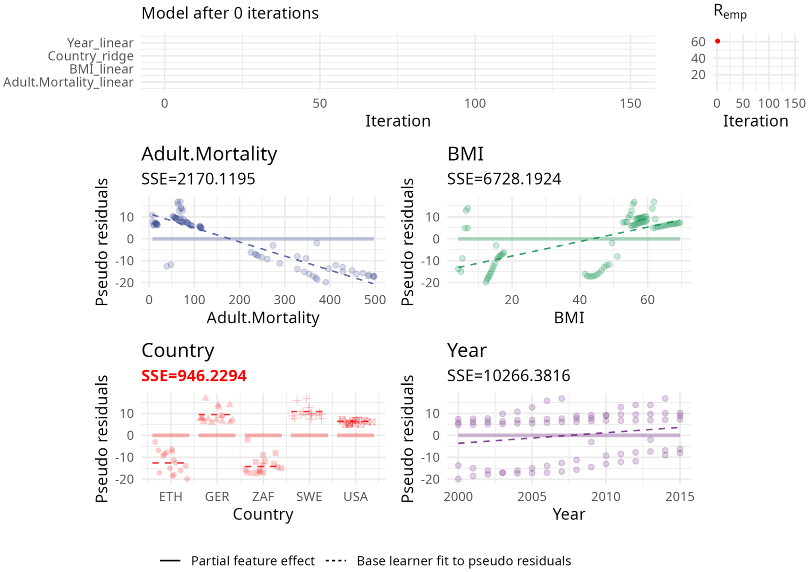
Consider the `life expectancy` data set (WHO, available on Kumar 2019) : regression task to predict life expectancy.

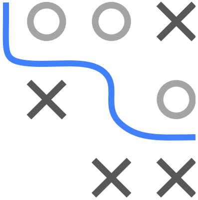
We fit a CWB model with linear BLs (with intercept)

variable	**description**
`Life.expectancy`	Life expectancy in years
`Country`	The country (just a selection GER, USE, SWE, ZAF, and ETH)
`Year`	The recorded year
`BMI`	Average BMI $= \frac{\text{body weight in kg}}{(\text{Height in m})^2}$ in a year and country
`Adult.Mortality`	Adult mortality rates per 1000 population

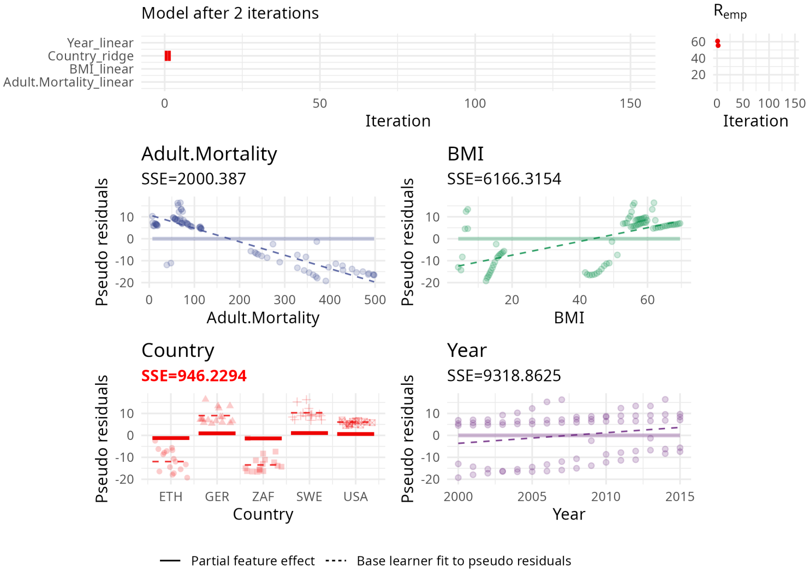
Using `compboost` with $M = 150$ iterations, we can visualize which BL was selected when and how the estimated feature effects evolve over time.

EXAMPLE: LIFE EXPECTANCY

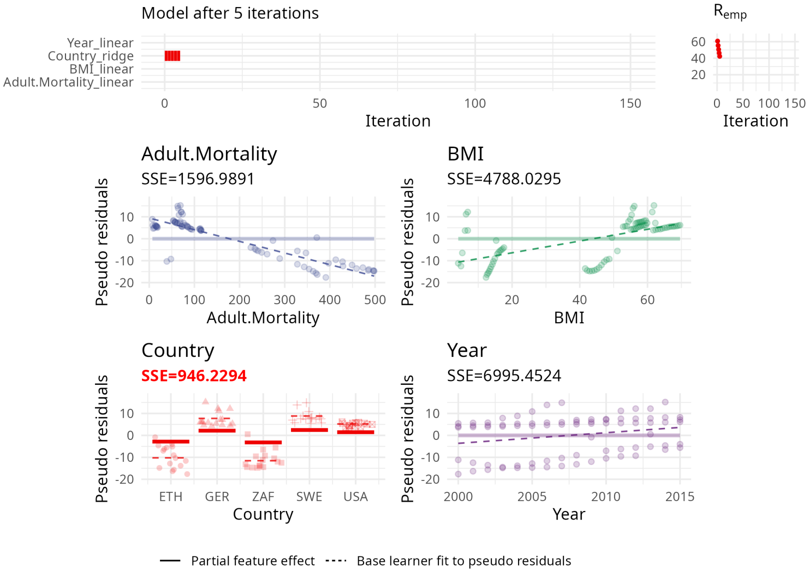


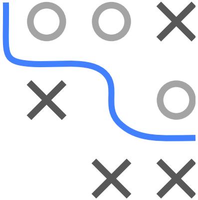


EXAMPLE: LIFE EXPECTANCY

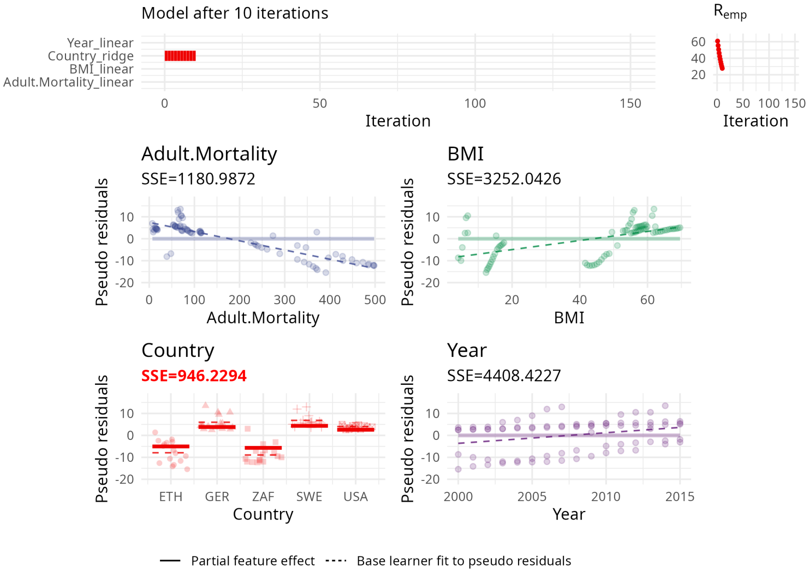


EXAMPLE: LIFE EXPECTANCY

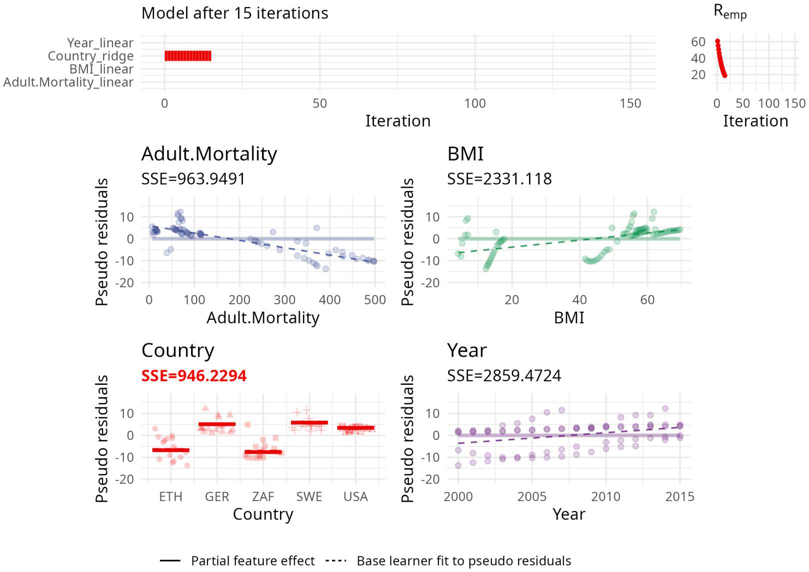


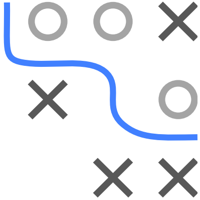


EXAMPLE: LIFE EXPECTANCY

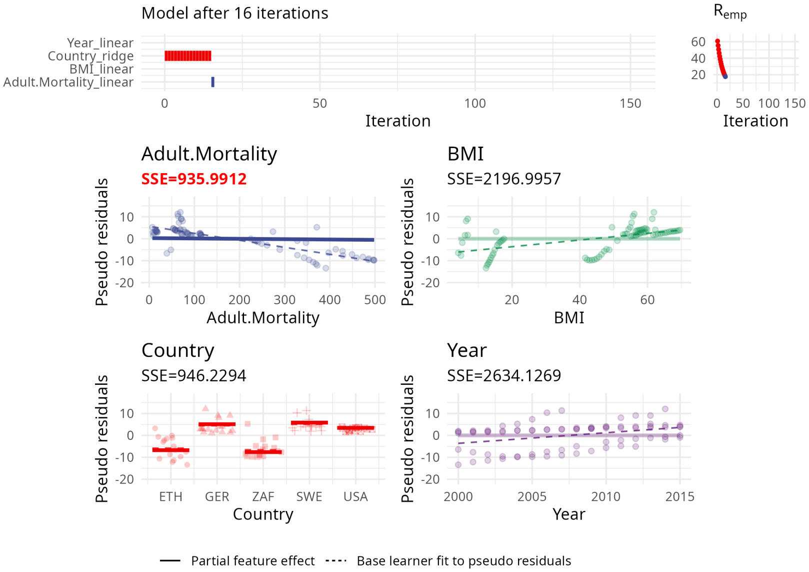


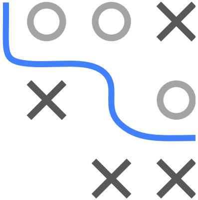
EXAMPLE: LIFE EXPECTANCY



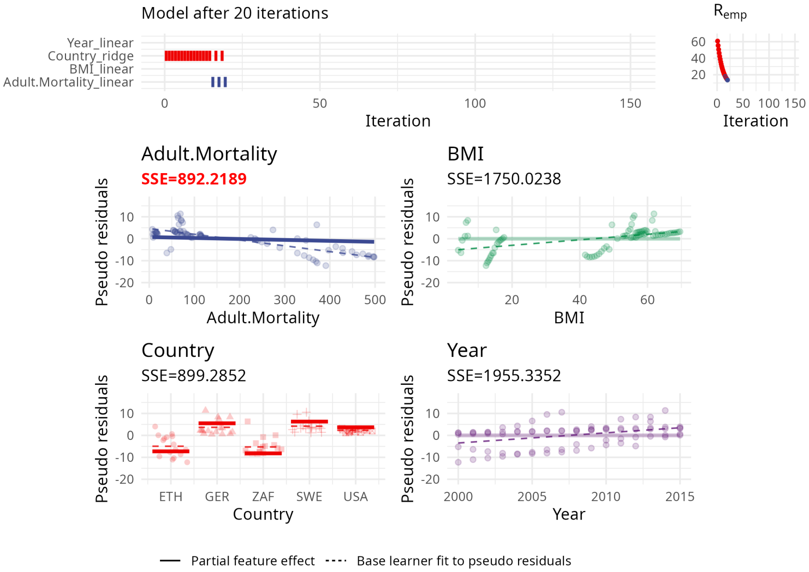


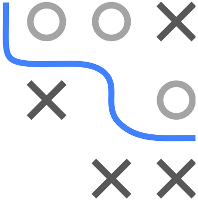
EXAMPLE: LIFE EXPECTANCY



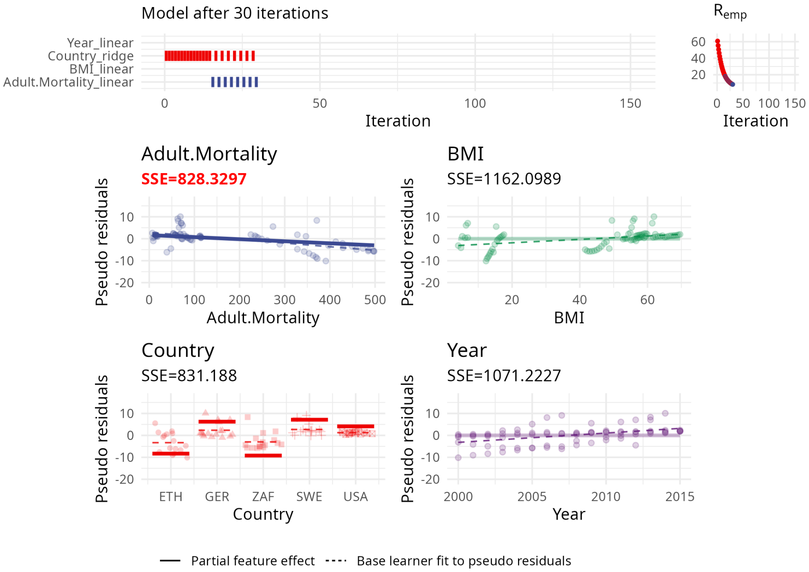


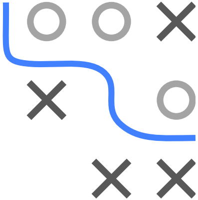
EXAMPLE: LIFE EXPECTANCY



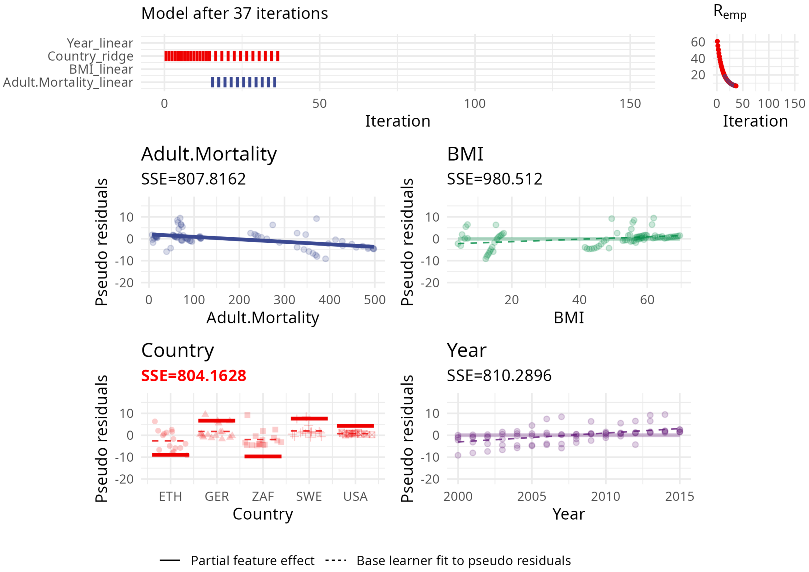


EXAMPLE: LIFE EXPECTANCY

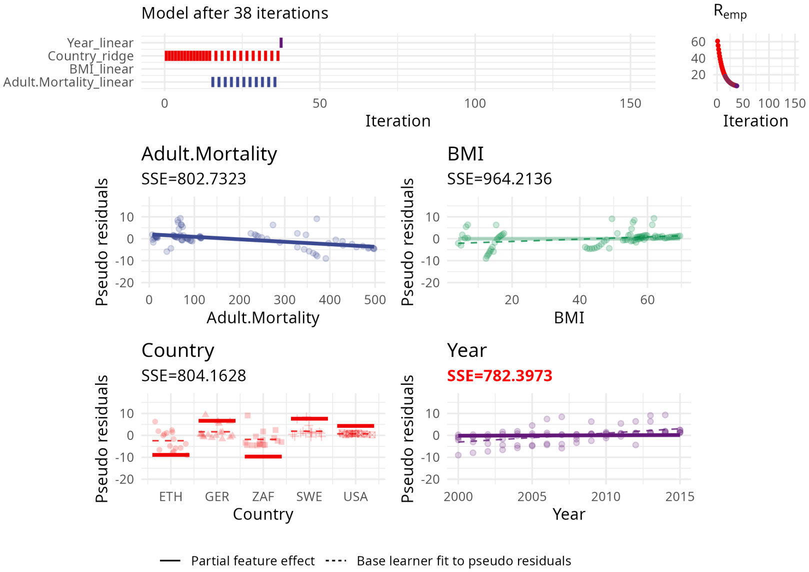


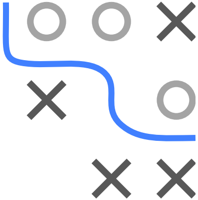


EXAMPLE: LIFE EXPECTANCY

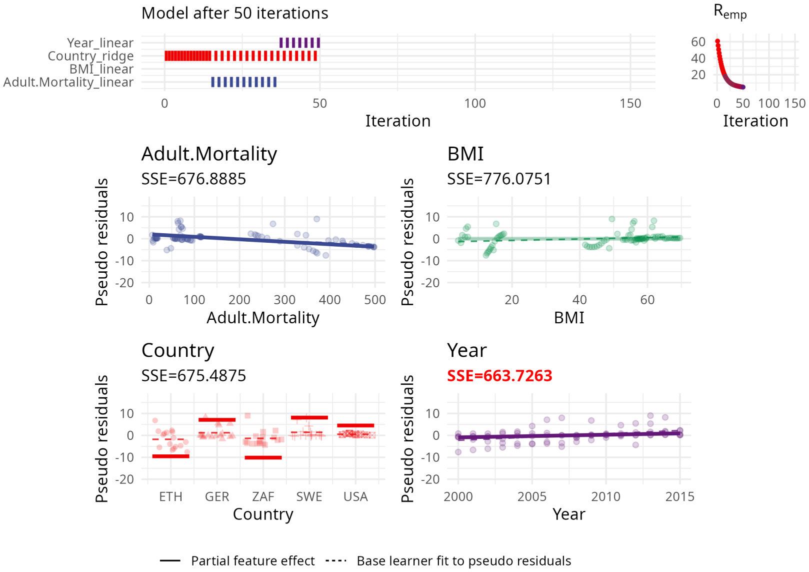


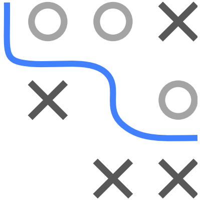
EXAMPLE: LIFE EXPECTANCY



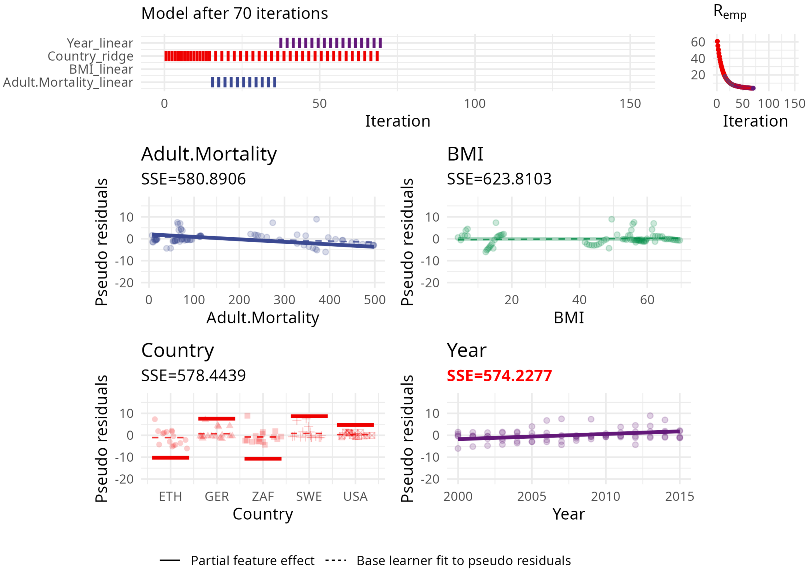


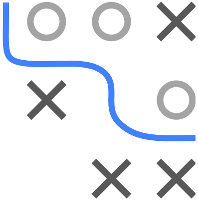
EXAMPLE: LIFE EXPECTANCY



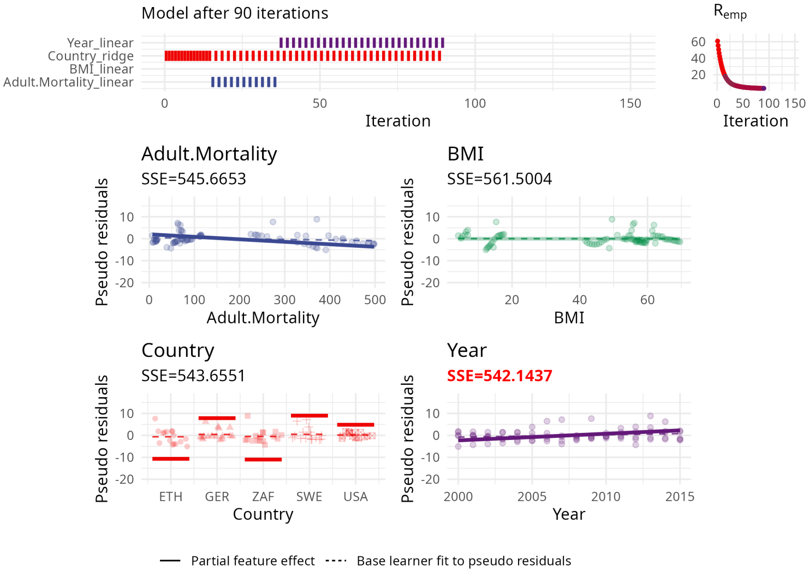


EXAMPLE: LIFE EXPECTANCY

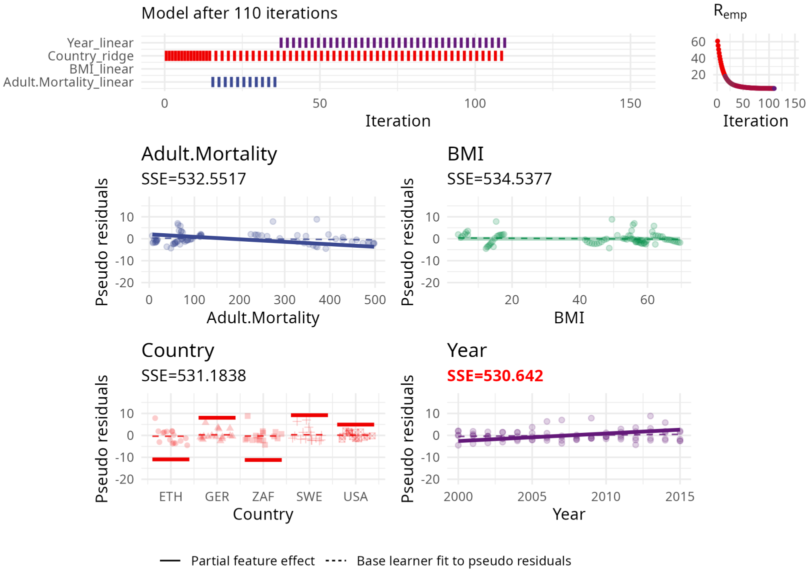


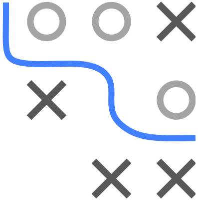


EXAMPLE: LIFE EXPECTANCY

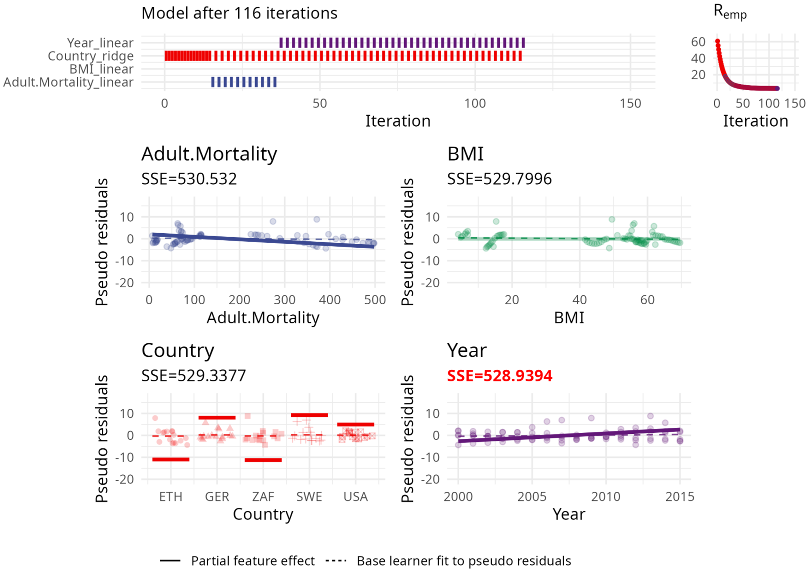


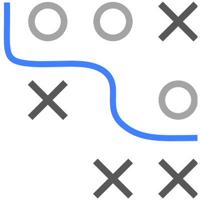
EXAMPLE: LIFE EXPECTANCY



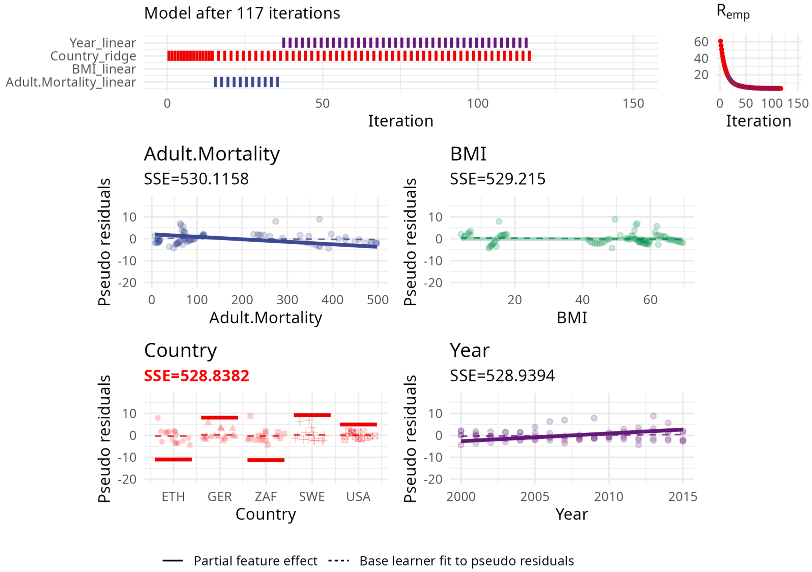


EXAMPLE: LIFE EXPECTANCY

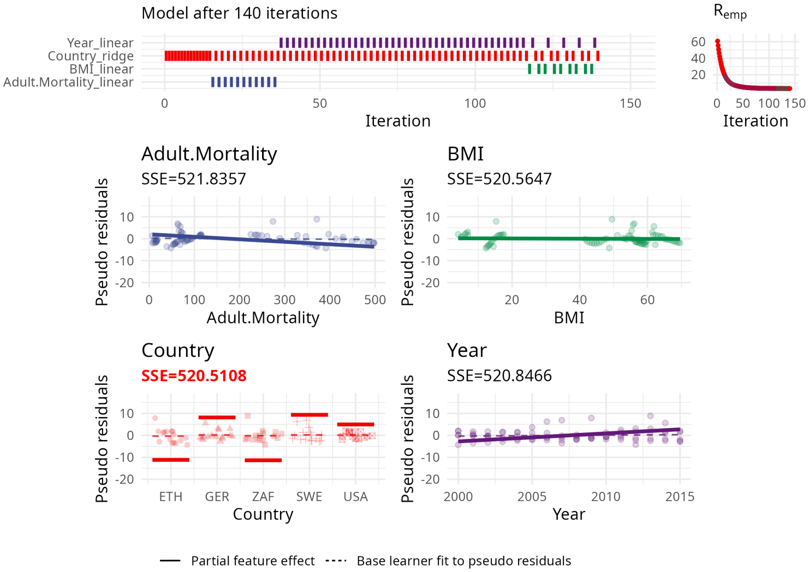


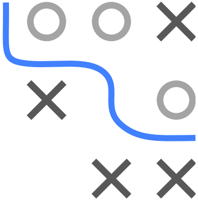


EXAMPLE: LIFE EXPECTANCY



EXAMPLE: LIFE EXPECTANCY





EXAMPLE: LIFE EXPECTANCY