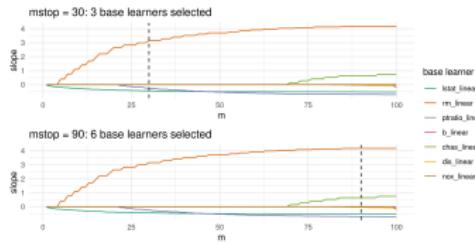


Introduction to Machine Learning

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:

$$b_j(x_j|\theta_j) = \sum_{g=1}^G \theta_{j,g} \mathbb{1}_{\{g=x_j\}} = (\mathbb{1}_{\{x_j=1\}}, \dots, \mathbb{1}_{\{x_j=G\}}) \theta_j$$



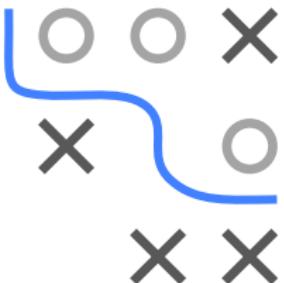
Hence, b_j incorporates a one-hot encoded feature with group means $\theta \in \mathbb{R}^G$ as estimators.

- 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



Advantages of simultaneously handling all categories in CWB:

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

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

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

HANDLING OF CATEGORICAL FEATURES

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

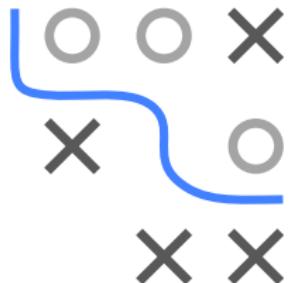
There are two options to handle the intercept in CWB. In both, the loss-optimal constant $f^{[0]}(\mathbf{x})$ is an initial model intercept.

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 $\hat{\theta}^{[1]} = (\hat{\theta}_{j1}^{[1]}, \hat{\theta}_{j2}^{[1]})$ 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 $f^{[0]}(\mathbf{x}) + \sum_{m=1}^M \hat{\theta}_{j1}^{[m]}$



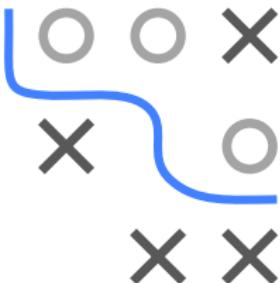
EXAMPLE: LIFE EXPECTANCY

Consider the life expectancy data set (WHO, available on
▶ [Click for source](#)) : regression task to predict life expectancy.

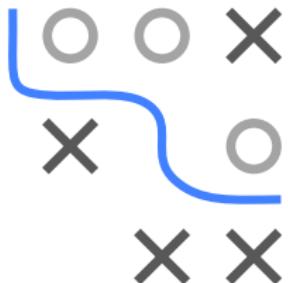
We fit a CWB model with linear BLs (with intercept)

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

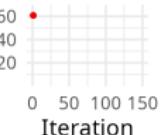


Model after 0 iterations

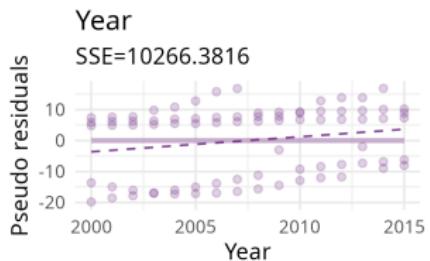
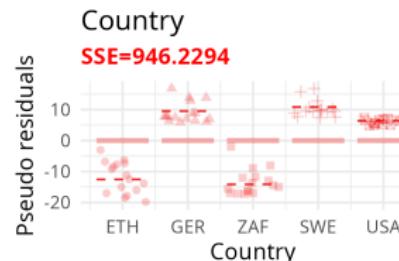
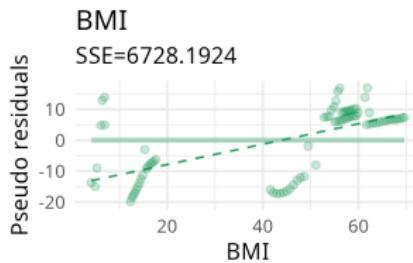
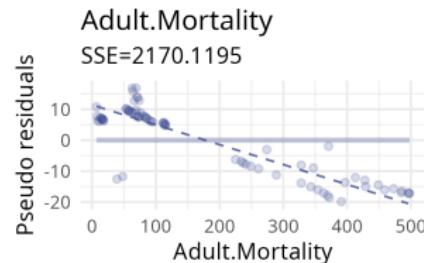
Year_linear
Country_ridge
BMI_linear
Adult.Mortality_linear



R_{emp}

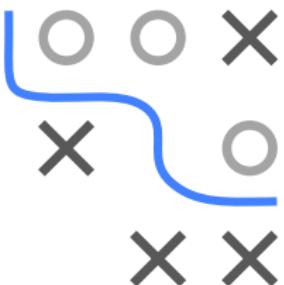


Iteration



— Partial feature effect ---- Base learner fit to pseudo residuals

EXAMPLE: LIFE EXPECTANCY



Model after 2 iterations

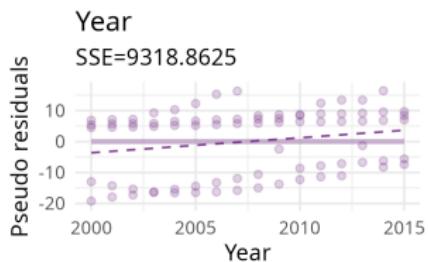
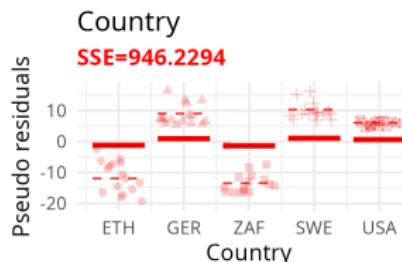
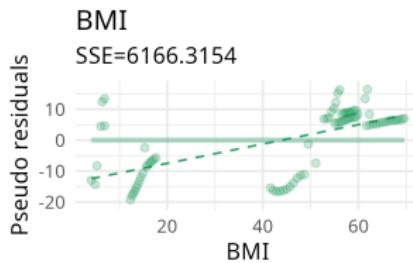
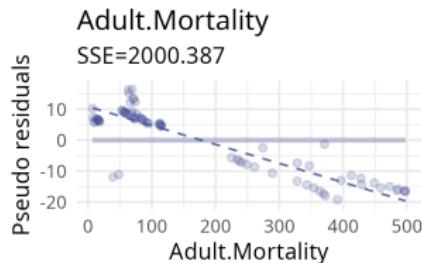
Year_linear
Country_ridge
BMI_linear
Adult.Mortality_linear



R_{emp}



Iteration



— Partial feature effect --- Base learner fit to pseudo residuals

EXAMPLE: LIFE EXPECTANCY

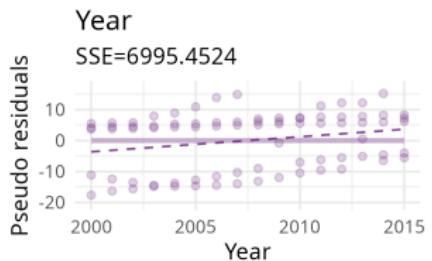
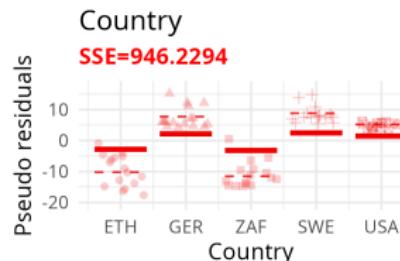
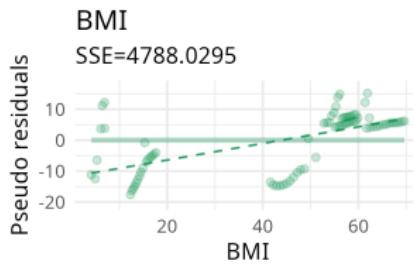
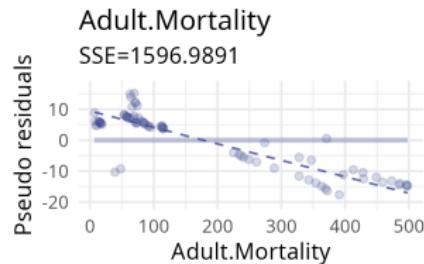
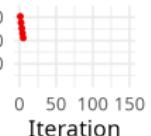


Model after 5 iterations

Year_linear
Country_ridge
BMI_linear
Adult.Mortality_linear



R_{emp}



— Partial feature effect - - - Base learner fit to pseudo residuals

EXAMPLE: LIFE EXPECTANCY

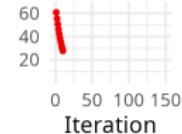


Model after 10 iterations

Year_linear
Country_ridge
BMI_linear
Adult.Mortality_linear

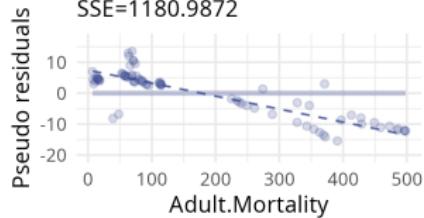


R_{emp}



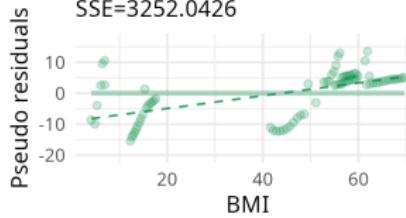
Adult.Mortality

SSE=1180.9872



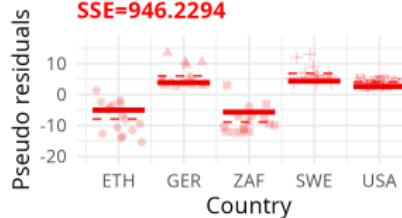
BMI

SSE=3252.0426



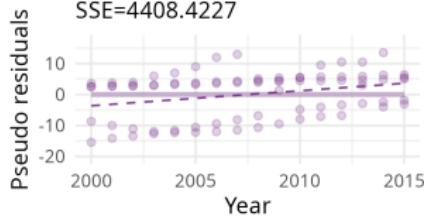
Country

SSE=946.2294



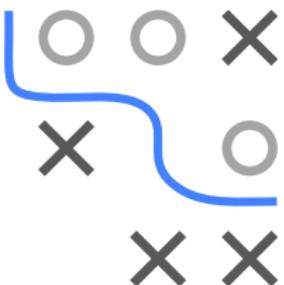
Year

SSE=4408.4227



— Partial feature effect --- Base learner fit to pseudo residuals

EXAMPLE: LIFE EXPECTANCY

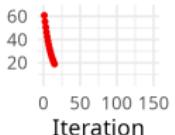


Model after 15 iterations

Year_linear
Country_ridge
BMI_linear
Adult.Mortality_linear

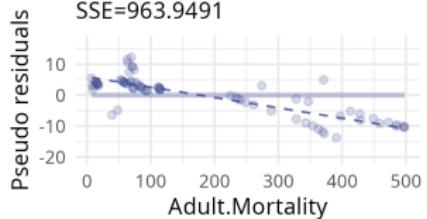


R_{emp}



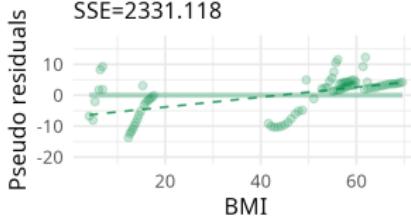
Adult.Mortality

SSE=963.9491



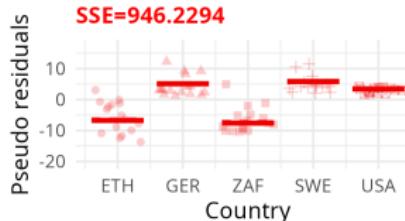
BMI

SSE=2331.118



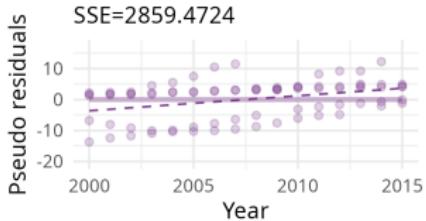
Country

SSE=946.2294



Year

SSE=2859.4724



— Partial feature effect ---- Base learner fit to pseudo residuals

EXAMPLE: LIFE EXPECTANCY

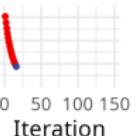


Model after 16 iterations

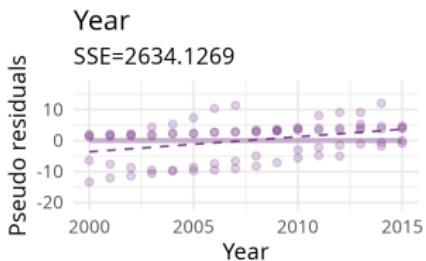
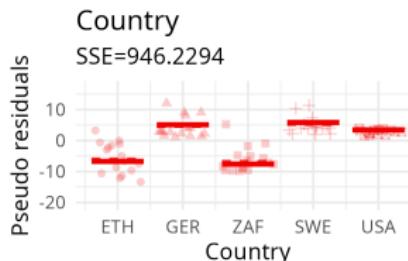
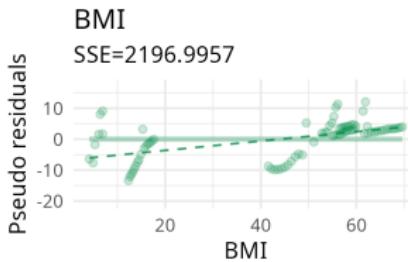
Year_linear
Country_ridge
BMI_linear
Adult.Mortality_linear



R_{emp}

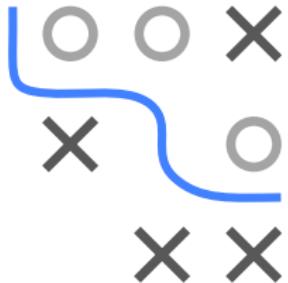


Iteration



— Partial feature effect ---- Base learner fit to pseudo residuals

EXAMPLE: LIFE EXPECTANCY

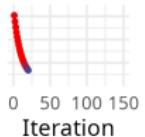


Model after 20 iterations

Year_linear
Country_ridge
BMI_linear
Adult.Mortality_linear

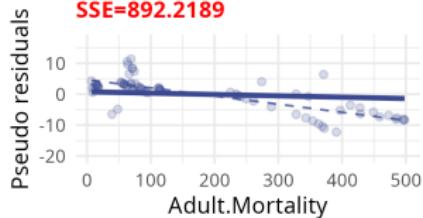


R_{emp}



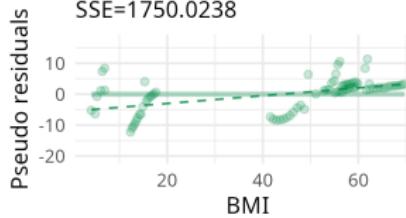
Adult.Mortality

SSE=892.2189



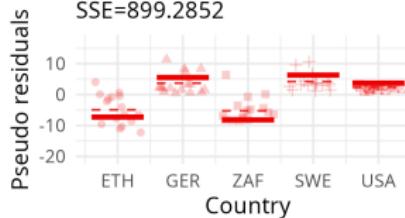
BMI

SSE=1750.0238



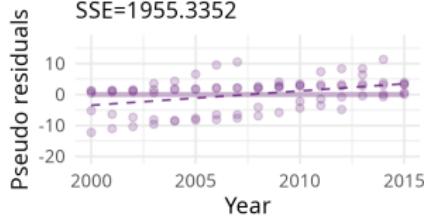
Country

SSE=899.2852



Year

SSE=1955.3352



— Partial feature effect ---- Base learner fit to pseudo residuals

EXAMPLE: LIFE EXPECTANCY

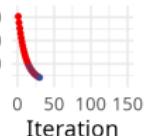


Model after 30 iterations

Year_linear
Country_ridge
BMI_linear
Adult.Mortality_linear

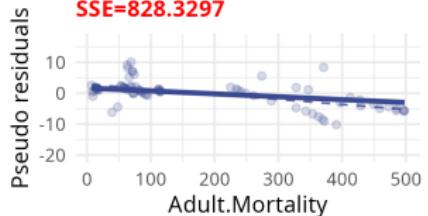


R_{emp}



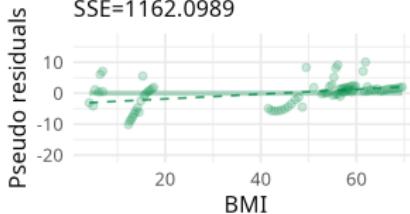
Adult.Mortality

SSE=828.3297



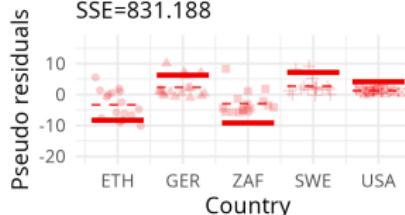
BMI

SSE=1162.0989



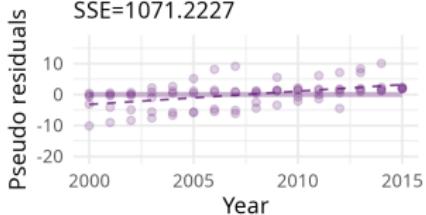
Country

SSE=831.188



Year

SSE=1071.2227



— Partial feature effect ---- Base learner fit to pseudo residuals

EXAMPLE: LIFE EXPECTANCY

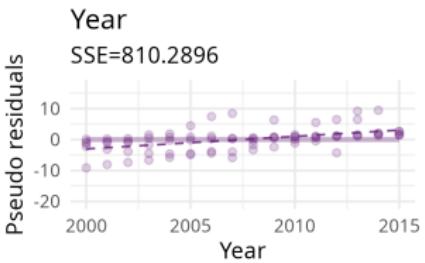
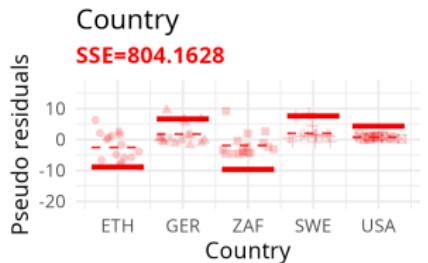
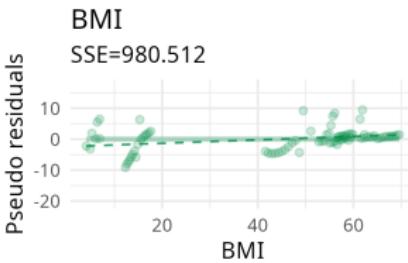
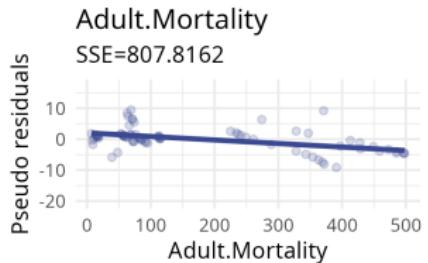
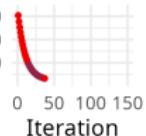


Model after 37 iterations

Year_linear
Country_ridge
BMI_linear
Adult.Mortality_linear

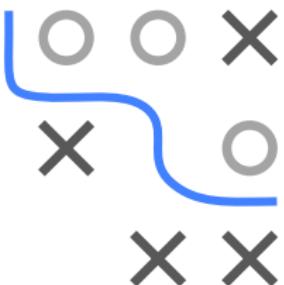


R_{emp}



— Partial feature effect ---- Base learner fit to pseudo residuals

EXAMPLE: LIFE EXPECTANCY

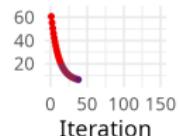


Model after 38 iterations

Year_linear
Country_ridge
BMI_linear
Adult.Mortality_linear

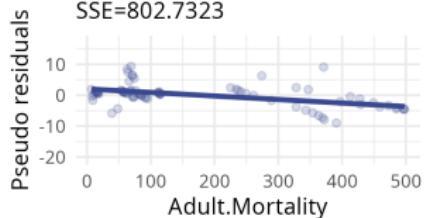


R_{emp}



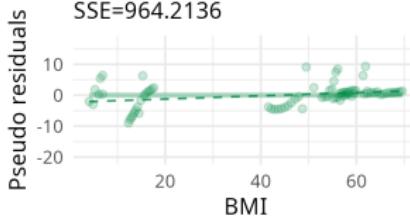
Adult.Mortality

SSE=802.7323



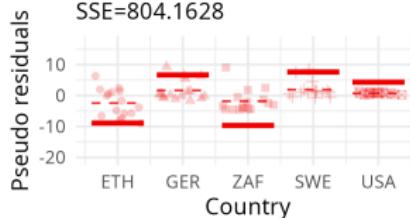
BMI

SSE=964.2136



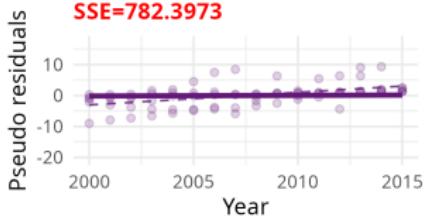
Country

SSE=804.1628



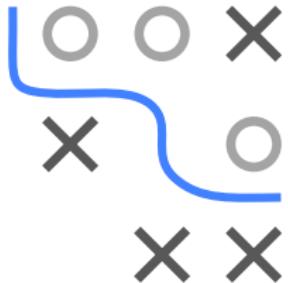
Year

SSE=782.3973



— Partial feature effect ---- Base learner fit to pseudo residuals

EXAMPLE: LIFE EXPECTANCY

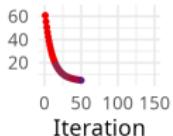


Model after 50 iterations

Year_linear
Country_ridge
BMI_linear
Adult.Mortality_linear

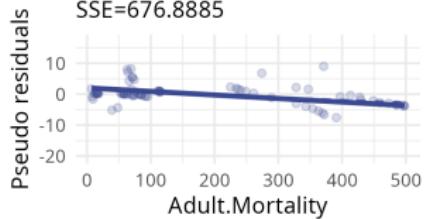


R_{emp}



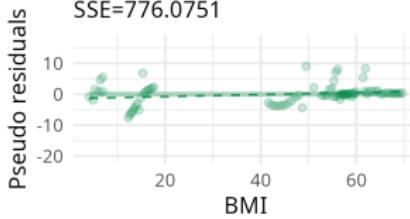
Adult.Mortality

SSE=676.8885



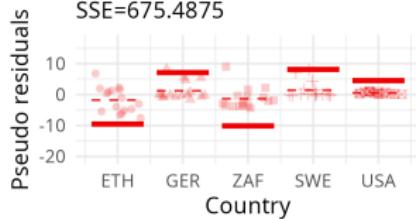
BMI

SSE=776.0751



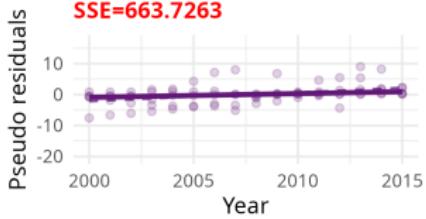
Country

SSE=675.4875



Year

SSE=663.7263



— Partial feature effect --- Base learner fit to pseudo residuals

EXAMPLE: LIFE EXPECTANCY



Model after 70 iterations

Year_linear
Country_ridge
BMI_linear
Adult.Mortality_linear

Iteration

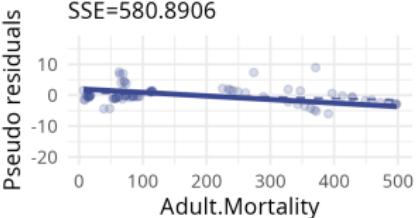
R_{emp}

60
40
20

Iteration

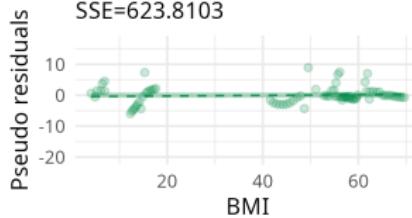
Adult.Mortality

SSE=580.8906



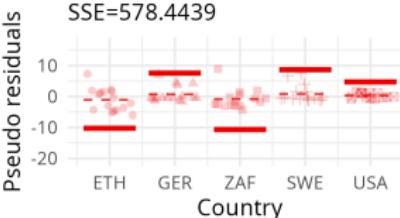
BMI

SSE=623.8103



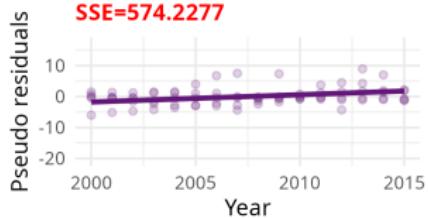
Country

SSE=578.4439



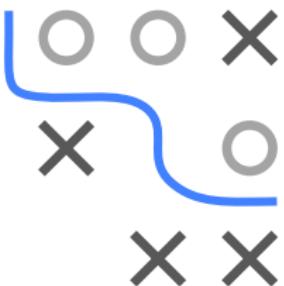
Year

SSE=574.2277

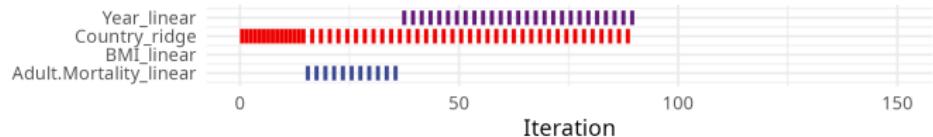


— Partial feature effect - - - Base learner fit to pseudo residuals

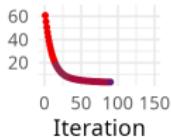
EXAMPLE: LIFE EXPECTANCY



Model after 90 iterations

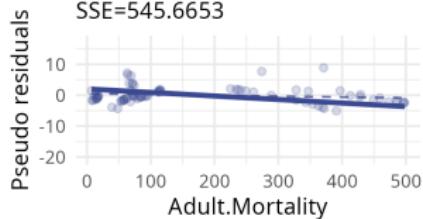


R_{emp}



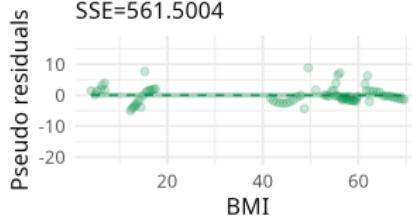
Adult.Mortality

SSE=545.6653



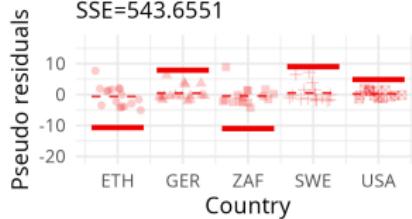
BMI

SSE=561.5004



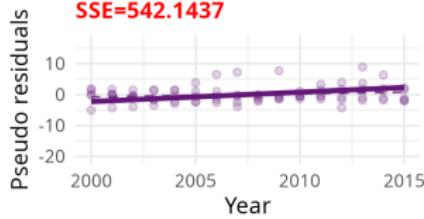
Country

SSE=543.6551



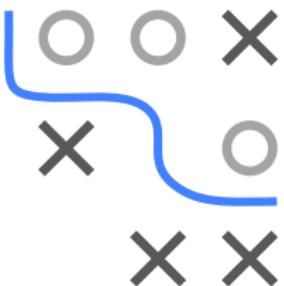
Year

SSE=542.1437

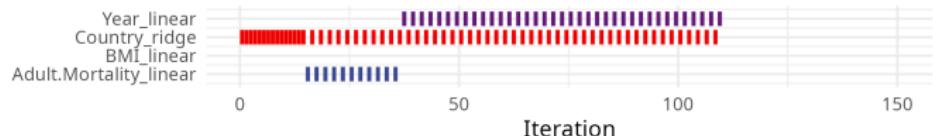


— Partial feature effect ---- Base learner fit to pseudo residuals

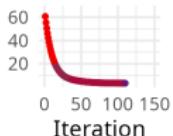
EXAMPLE: LIFE EXPECTANCY



Model after 110 iterations

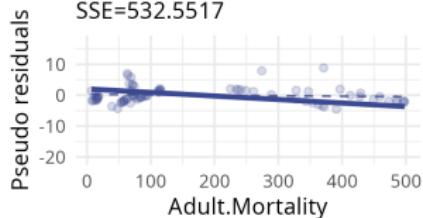


R_{emp}



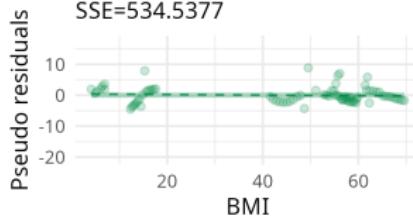
Adult.Mortality

SSE=532.5517



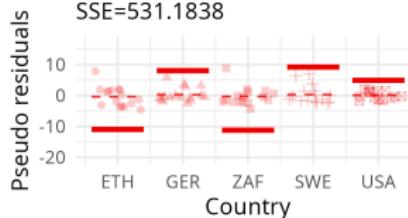
BMI

SSE=534.5377



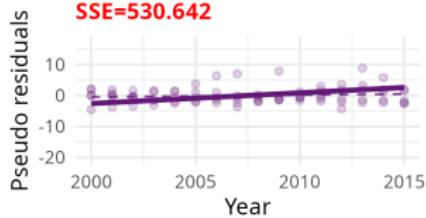
Country

SSE=531.1838



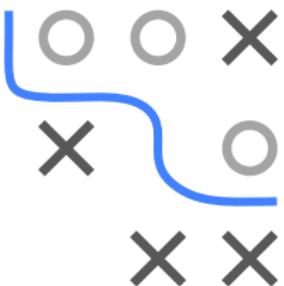
Year

SSE=530.642

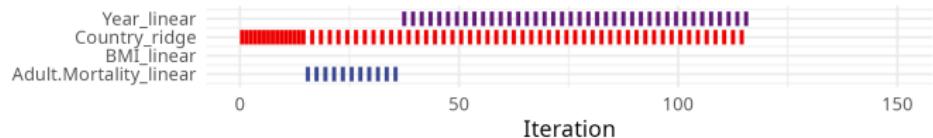


— Partial feature effect --- Base learner fit to pseudo residuals

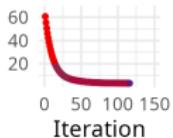
EXAMPLE: LIFE EXPECTANCY



Model after 116 iterations

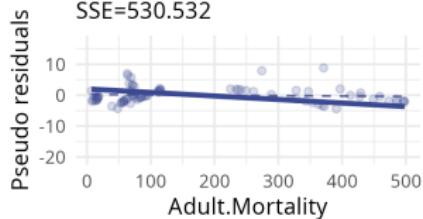


R_{emp}



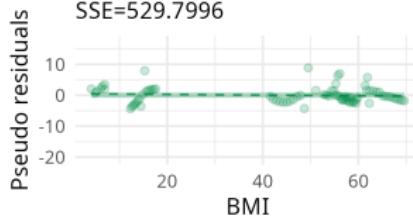
Adult.Mortality

SSE=530.532



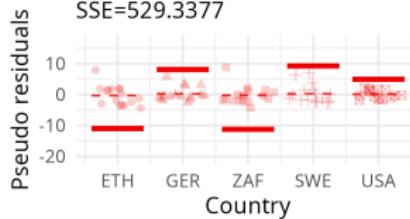
BMI

SSE=529.7996



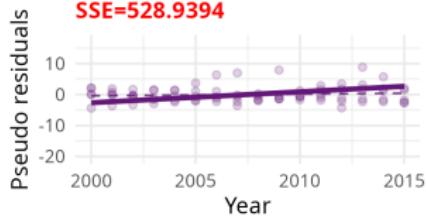
Country

SSE=529.3377



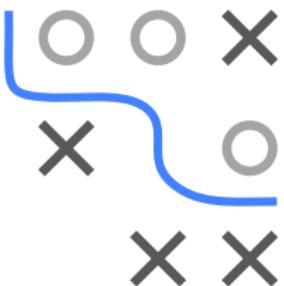
Year

SSE=528.9394



— Partial feature effect - - - Base learner fit to pseudo residuals

EXAMPLE: LIFE EXPECTANCY

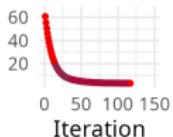


Model after 117 iterations

Year_linear
Country_ridge
BMI_linear
Adult.Mortality_linear

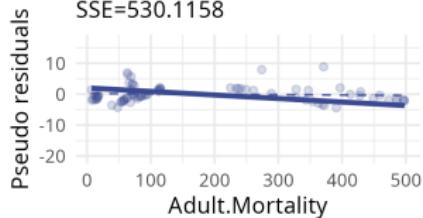


R_{emp}



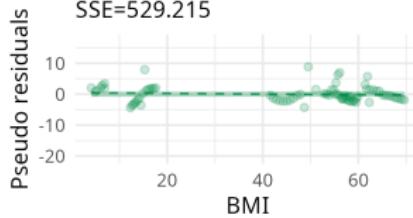
Adult.Mortality

SSE=530.1158



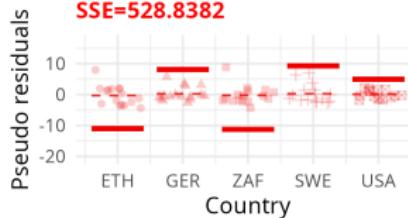
BMI

SSE=529.215



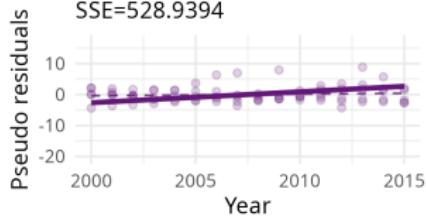
Country

SSE=528.8382



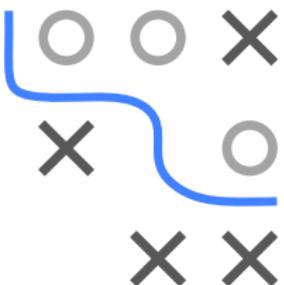
Year

SSE=528.9394

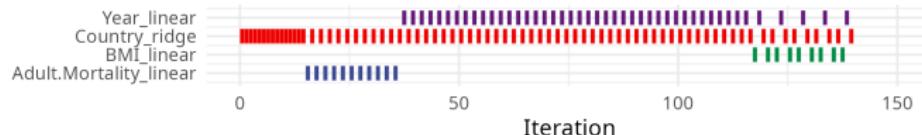


— Partial feature effect ---- Base learner fit to pseudo residuals

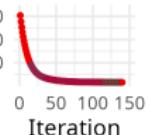
EXAMPLE: LIFE EXPECTANCY



Model after 140 iterations

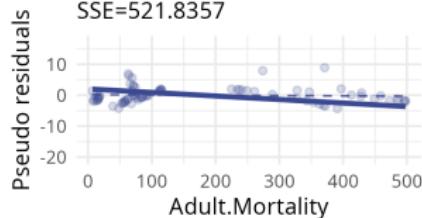


R_{emp}



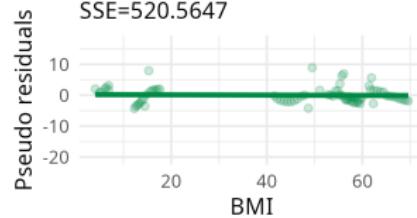
Adult.Mortality

SSE=521.8357



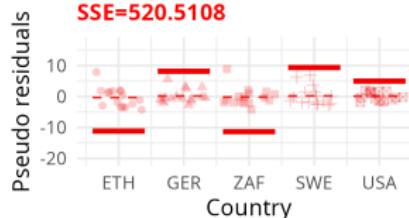
BMI

SSE=520.5647



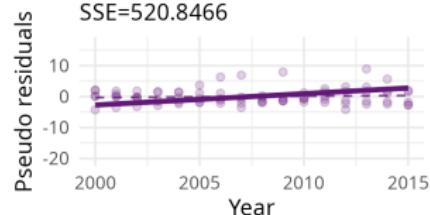
Country

SSE=520.5108



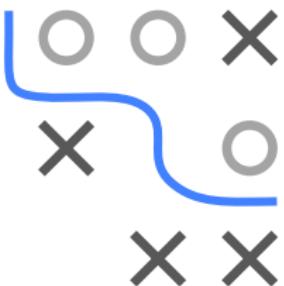
Year

SSE=520.8466

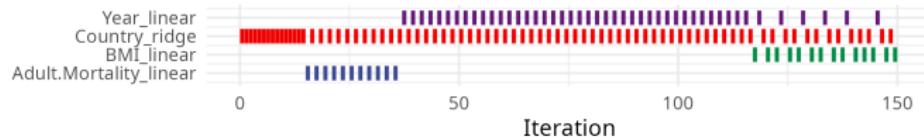


— Partial feature effect ---- Base learner fit to pseudo residuals

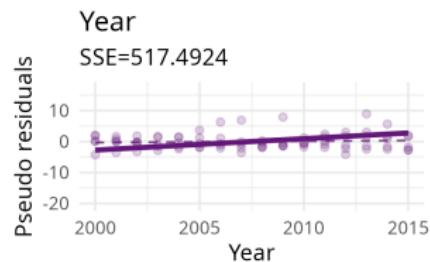
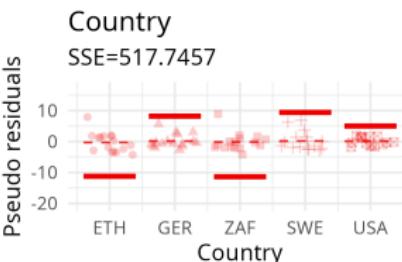
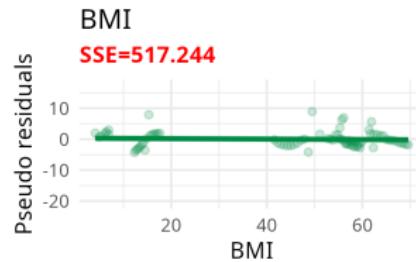
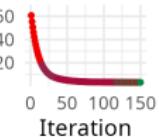
EXAMPLE: LIFE EXPECTANCY



Model after 150 iterations



R_{emp}



— Partial feature effect ---- Base learner fit to pseudo residuals