Gain insight in “black box” models like Random Forest by using Partial Dependence Plots (PDP) and Individual Conditional Expectations (ICE) Plots

Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation

ALEX GOLDSTEIN*, ADAM KAPELNER†, JUSTIN BLEICH‡, AND EMIL PITKIN§

The Wharton School of the University of Pennsylvania

March 21, 2014

arXiv:1309.6392v2 [stat.AP] 20 Mar 2014

R package: ICEbox
R-Code for paper figures: https://github.com/kapelner/ICEbox

Oultine

- Explanatory or predictive modeling
- What is the research question?
- Logistic regression model approach for explanation?
  - (check goodness of fit)
  - gain insight into the model by
    - partial residual plots
    - partial dependence plots
    - individual conditional expectation plots
- Random Forest model approach for prediction?
  - brief reminder on RF
  - (check performance)
  - gain insight into the model by
    - partial dependence plots
    - individual conditional expectation plots

Project: Complications after intervention

In 207 interventions we observed a complication rate of ~50%).
Can we model the complication risk?

Binary outcome: complication.6w (0:no, 1:yes)
5 predictors: age, kps, sex, op.indication, admission.source

```
complication.6w
0: 97
1:110
```

Do we want a predictive or a descriptive or an explanatory model?

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

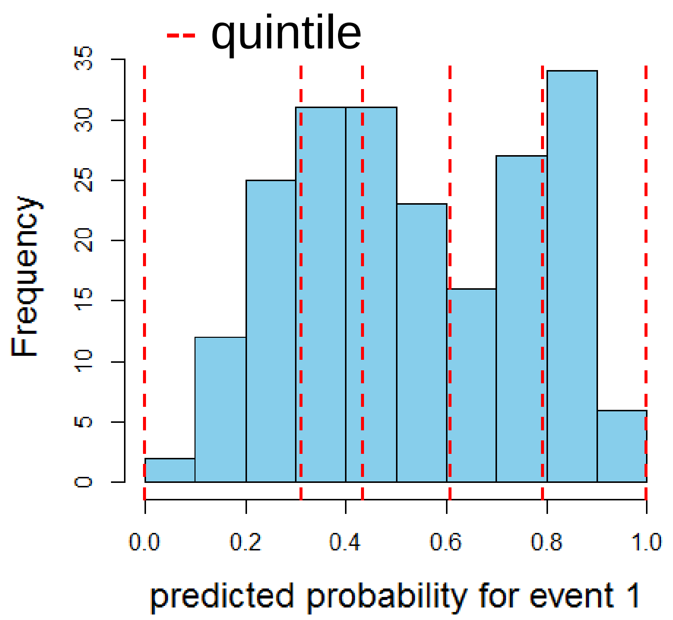
To Explain or to Predict?

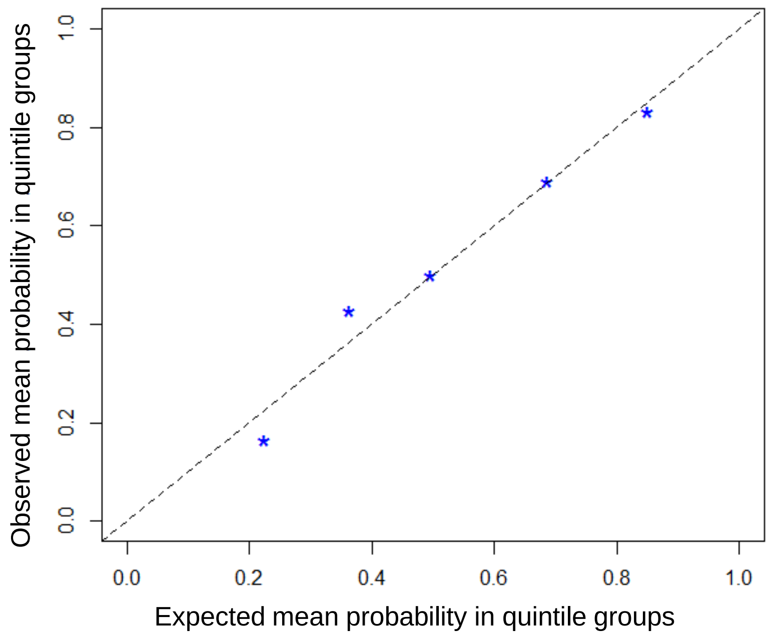
Galit Shmueli

FIG. 2. *Steps in the statistical modeling process.*

The bottom line is nicely summarized by Hagerty and Srinivasan (1991): “We note that the practice in applied research of concluding that a model with a higher predictive validity is “truer,” is not a valid inference. This paper shows that a parsimonious but less true model can have a higher predictive validity than a truer but less parsimonious model.”

Check goodness of fit for a logistic regression





sd>0.05

Hosmer-Lemeshow goodness of fit test:

```
Hosmer and Lemeshow goodness of fit (GOF) test

data:  dat$complication.6w, fitted(f.glm)
X-squared = 207, df = 3, p-value < 2.2e-16
```

Cessie – van Houwelingen – Copas – Hosmer unweighted sum of squares test

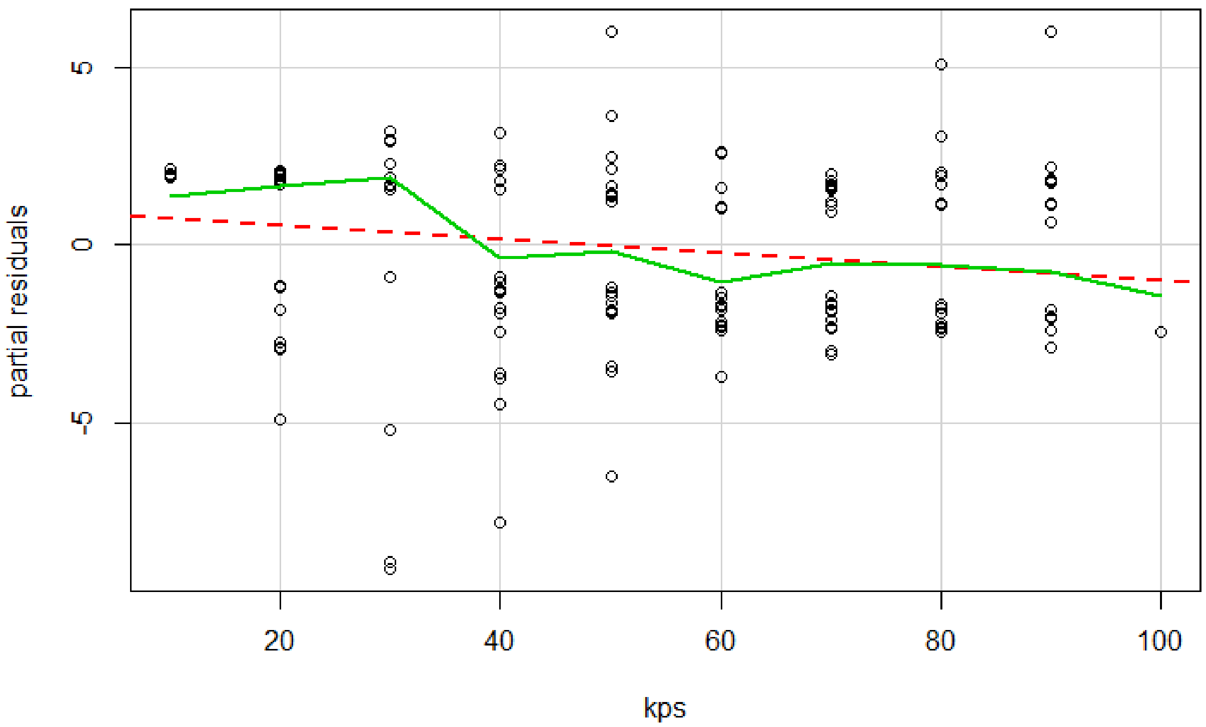
```
Sum of squared errors     Expected value|H0              SD
           40.3592793            40.6282989       0.2220258
                    Z                     P
           -1.2116591             0.2256429
```

Partial residual: marginal relation between predictor and outcome

The partial residuals give insight of the relationship between predictor x_i and "adjusted outcome", which is corrected for effect of all other predictors

The partial residuals are a matrix of working residuals of a model, where x_i was omitted from the model formula.

$$r_i^W = \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i(1-\hat{\mu}_i)}, \qquad \hat{\mu}_i = \frac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}}$$



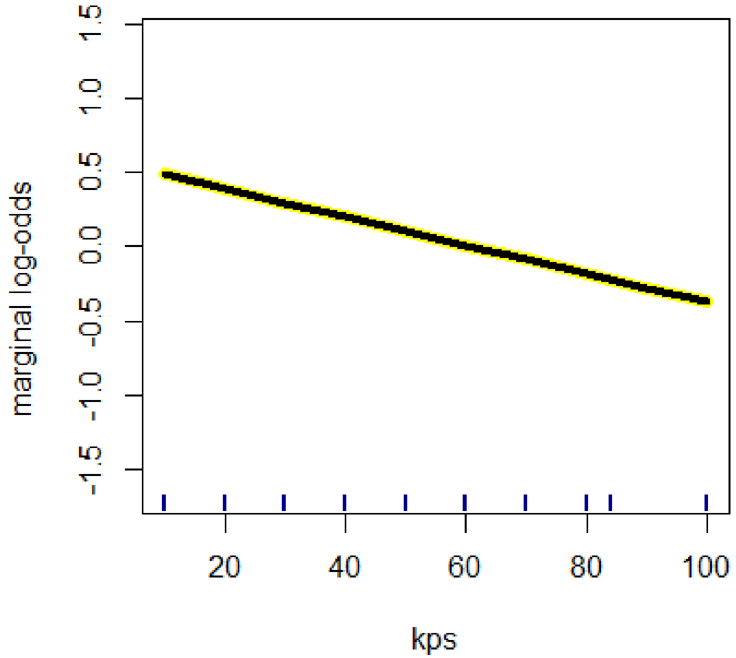
Here we see the partial residuals and a fitted linear line as well as a smoother.

Partial dependence plot & ICE plot show the fitted marginal response

Classical partial dependence plots (PDPs) plots the change in the average predicted value as the specified feature(s) vary over their marginal distribution.

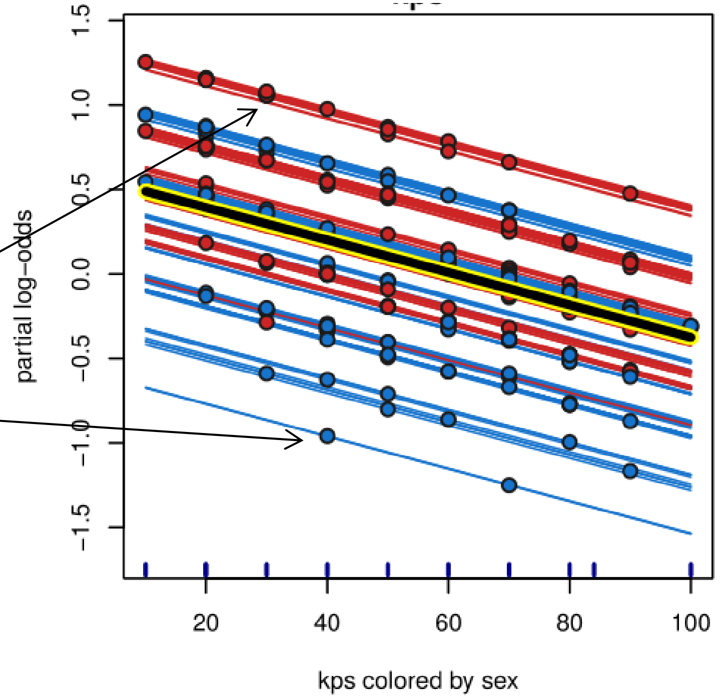
$$f_S = \mathbb{E}_{x_C}\left[f(x_S, x_C)\right] = \int f(x_S, x_C) \mathrm{dP}(x_C)$$

$$\hat{f}_S = \frac{1}{N}\sum_{i=1}^{N} \hat{f}(x_S, x_{Ci})$$



ICE plot: individual conditional expectation plots. **For each observed covariate set** we can predict the dependence of the outcome on kps:

sex	age	op.indication	revision	complication.6w	kps	admission.source
w	77	NPH	0	0	70	other care
m	73	NPH	0	0	50	other care
m	76	NPH	0	0	80	other care
m	57	NPH	1	0	40	other care
w	83	NPH	0	1	30	other care
w	48	otherH	0	1	20	other care



predict(f.glm, newdata=…)

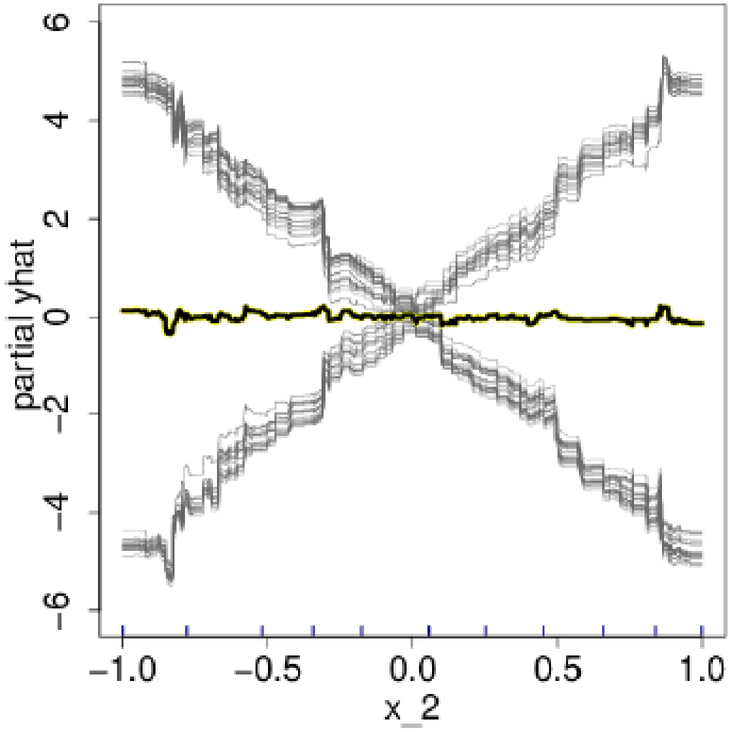
ICE plots can be generated for all models which have a predict function

ICE plots can reveal interactions which are invisible in PDP plots

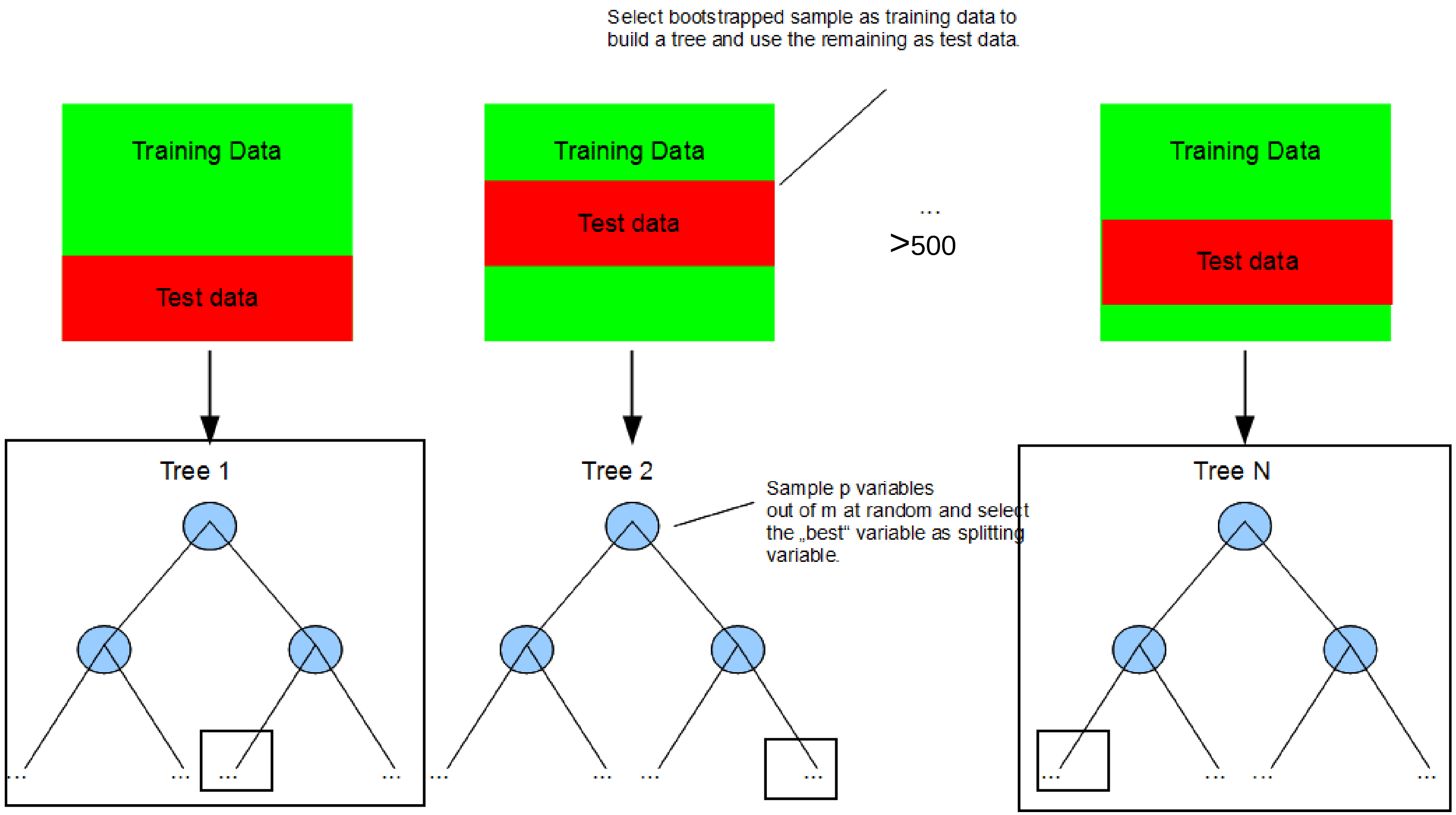
$$Y = 0.2X_1 - 5X_2 + 10X_2 \mathbb{1}_{X_3 \geq 0} + \mathcal{E},$$
$$\mathcal{E} \overset{iid}{\sim} \mathcal{N}(0, 1), \quad X_1, X_2, X_3 \overset{iid}{\sim} \mathrm{U}(-1, 1).$$

$$Y = 0.2 \cdot X_1 - 5 \cdot X_2 \quad , \text{for } \mathrm{X}_3 < 0$$
$$Y = 0.2 \cdot X_1 + 5 \cdot X_2 \quad , \text{for } \mathrm{X}_3 \geq 0$$

- Simulate 2,000 observations
- **Fit** a random forest **model which allows for interaction**.
- Check the PDP and ICE plot for X2



Random Forest



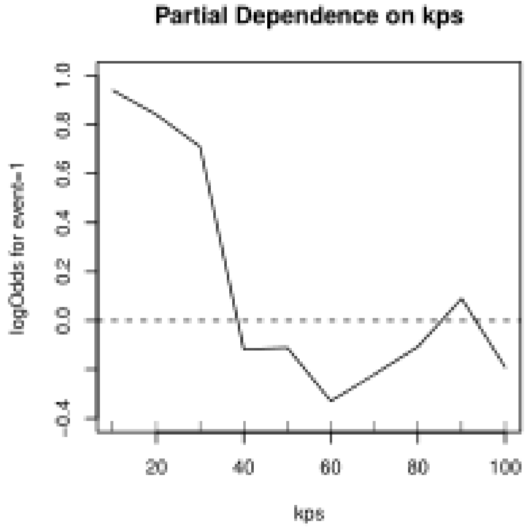
Evaluation: For each observation, construct its random forest oob-predictor by averaging only the results of those trees corresponding to bootstrap samples in which the observation was not contained.

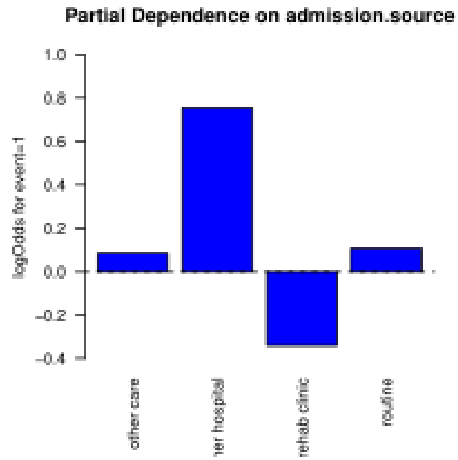
Random forest model and the resulting oob confusion matrix

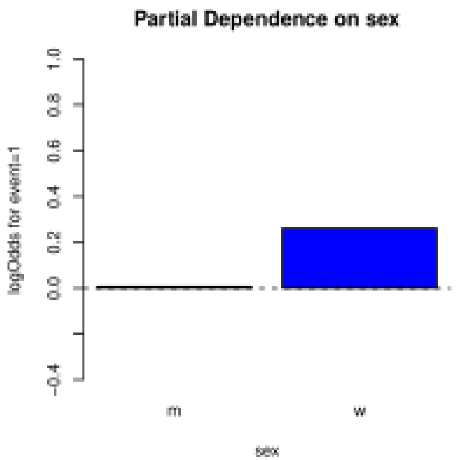
```
set.seed(99)
rf <- randomForest(complication.6w ~ sex+age+op.indication+
                                      admission.source+kps,
                   data=dat, importance=TRUE, ntree=1000 )
print(rf)
```

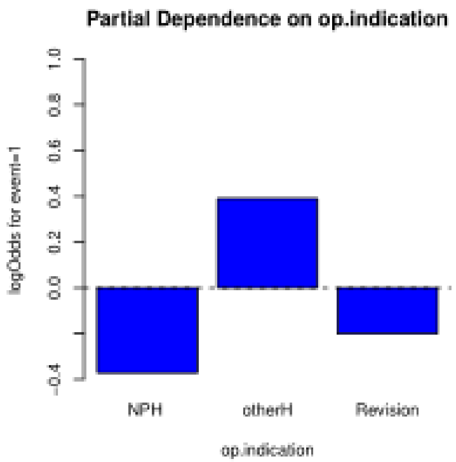
```
Call:
 randomForest(formula = complication.6w ~ sex + age + op.indication +      admis
sion.source + kps, data = dat, importance = TRUE, ntree = 1000)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 2

        OOB estimate of  error rate: 28.99%
Confusion matrix:
   0  1 class.error
0 62 35   0.3608247
1 25 85   0.2272727
```

PDP plots are provided by the randomForest package:



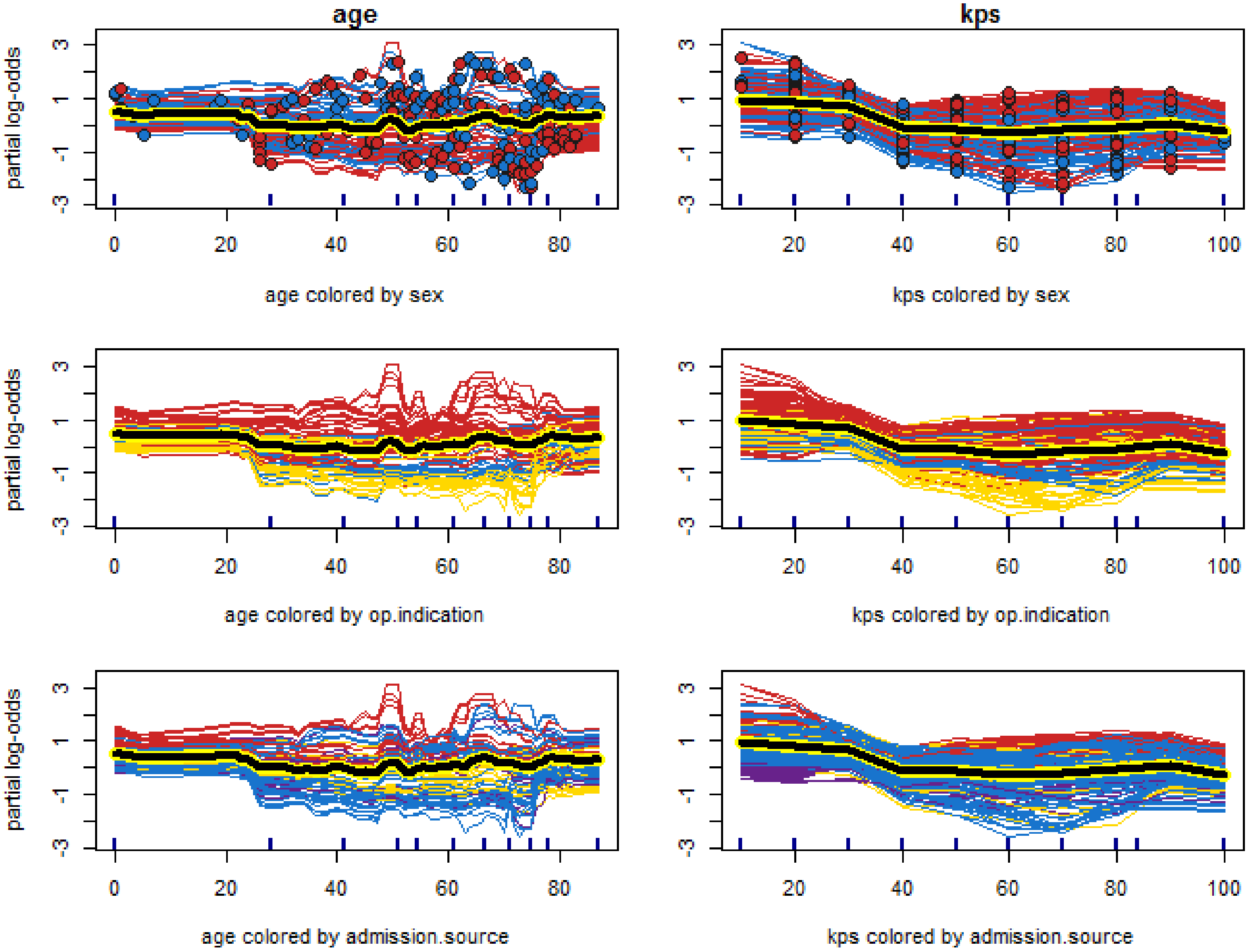







```
partialPlot(rf, dat, age,ylim=c(-0.4,1),
            which.class="1", ylab="logOdds for event=1")
abline(h=0,lty=2)
# kps:
partialPlot(rf, dat, kps,ylim=c(-0.4,1),
            which.class="1", ylab="logOdds for event=1")
abline(h=0,lty=2)
# sex:
partialPlot(rf, dat, sex,ylim=c(-0.4,1),
            which.class="1", ylab="logOdds for event=1")
abline(h=0,lty=2)
# op.indication:
partialPlot(rf, dat, op.indication, ylim=c(-0.4,1),
            which.class="1", ylab="logOdds for event=1")
abline(h=0,lty=2)
```

ICE plots are provided by the ICEbox package:



sex:

Blue: male

Red: female

op.indication:

Gold: NPH

Blue: revision

Red: otherH

admission.source:

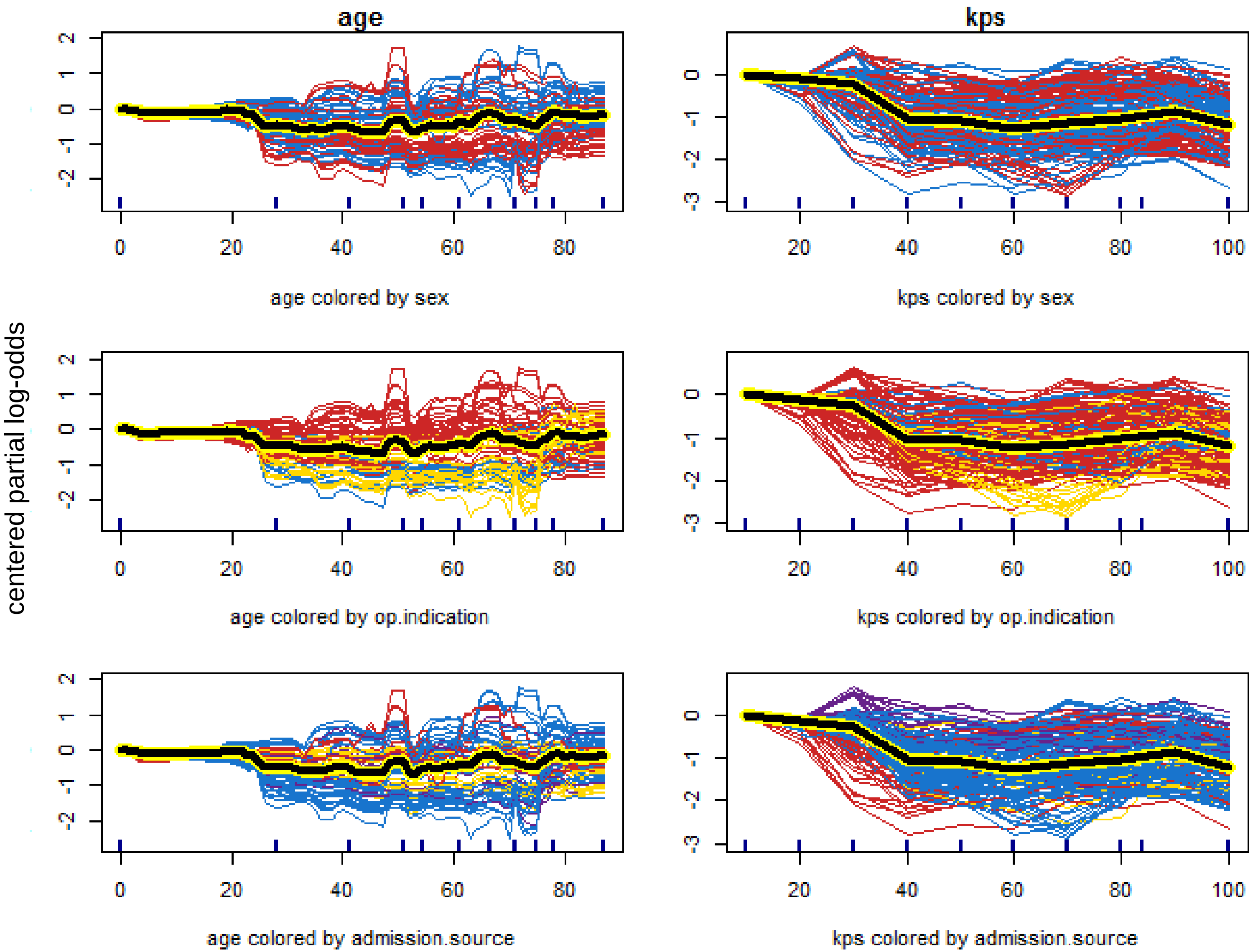
Violot: other care

Gold: reha

Blue: routine

Red: other hospital

Centered ICE reveal departure from average tendency

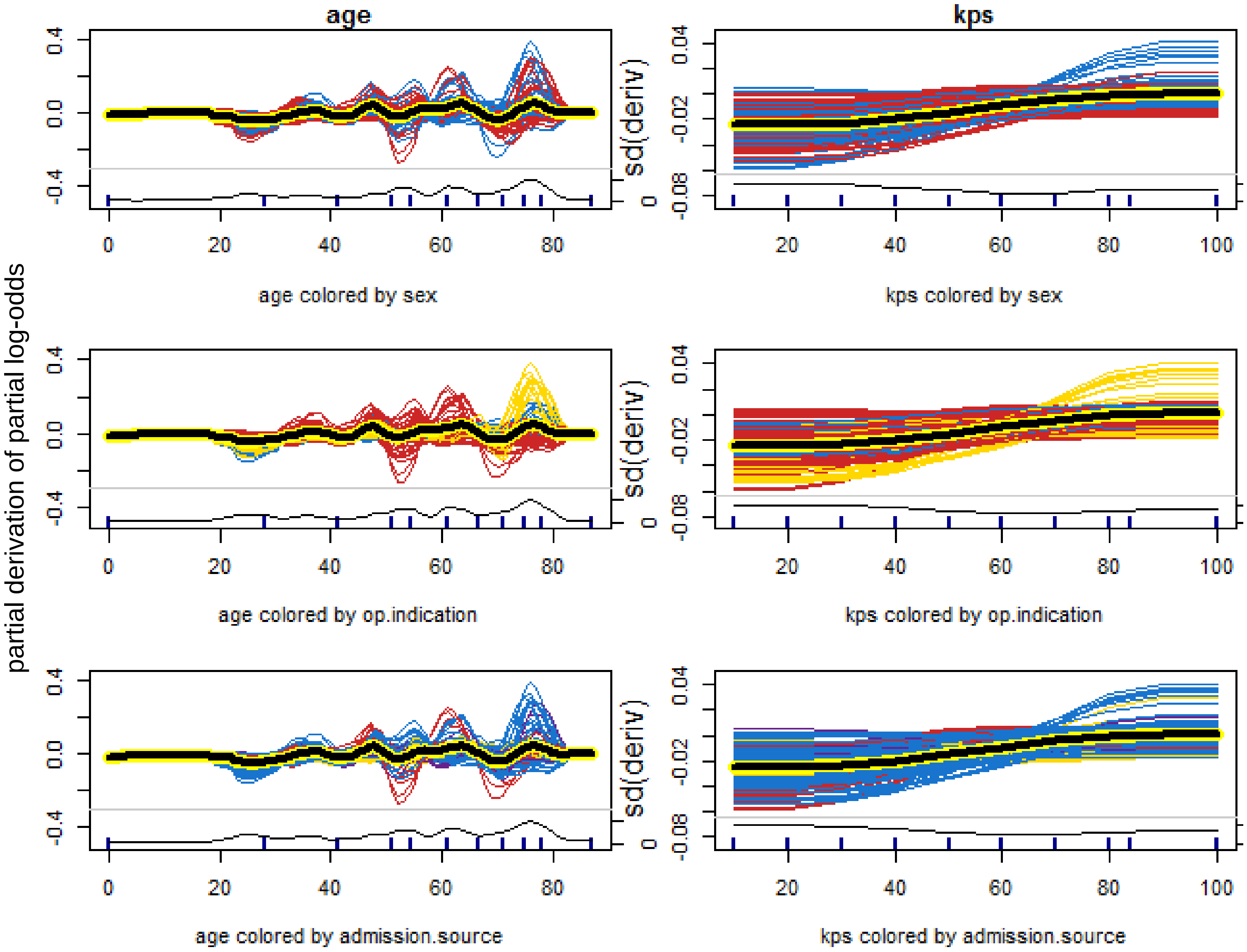


Blue: male
Red: female

Gold: NPH
Blue: revision
Red: otherH

Violot: other care
Gold: reha
Blue: routine
Red: other hospital

Derivate ICE plots reveal differences in shapes

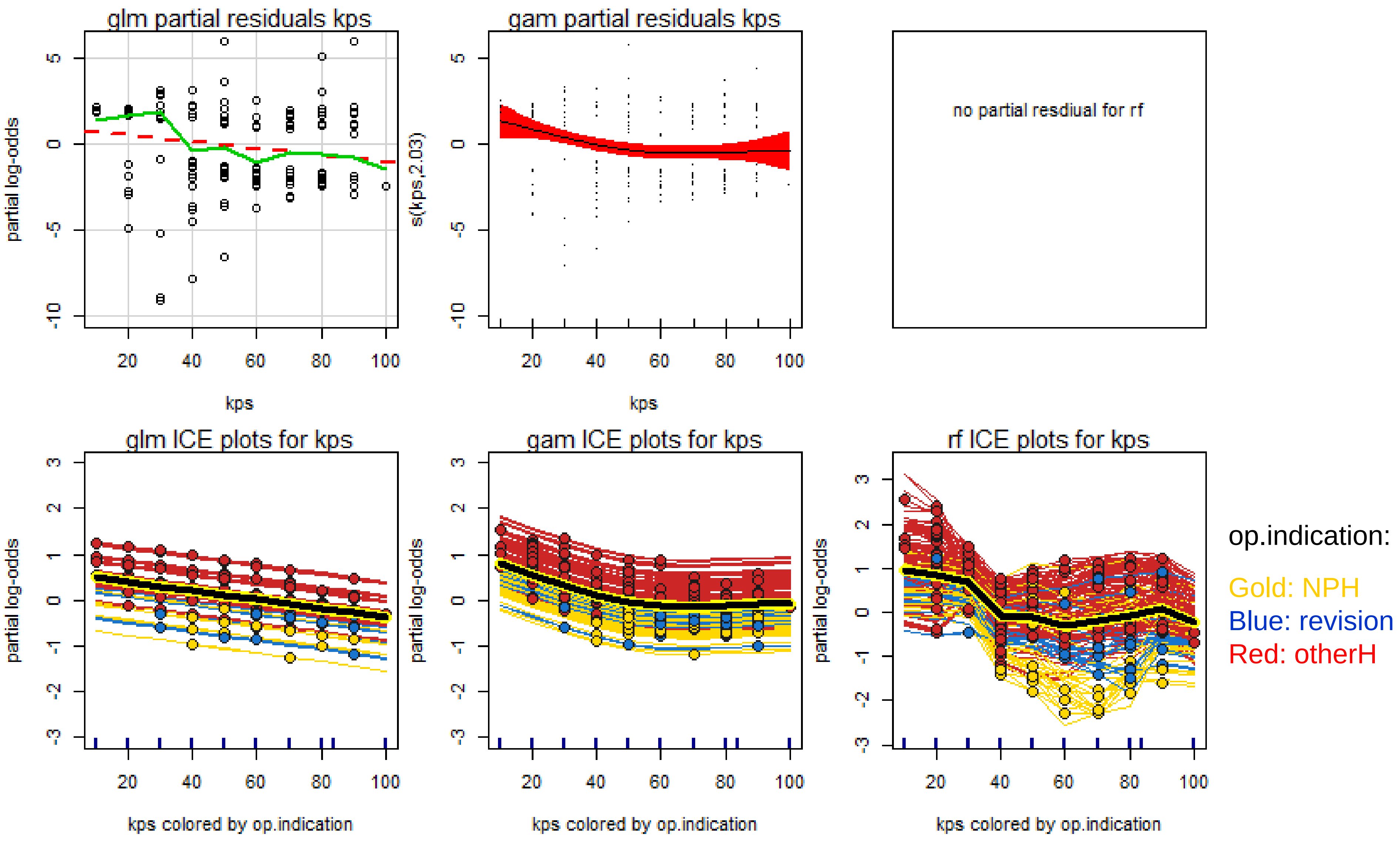


Blue: male
Red: female

Gold: NPH
Blue: revision
Red: otherH

Violot: other care
Gold: reha
Blue: routine
Red: other hospital

Comparison of 3 model types based on partial dependency on kps

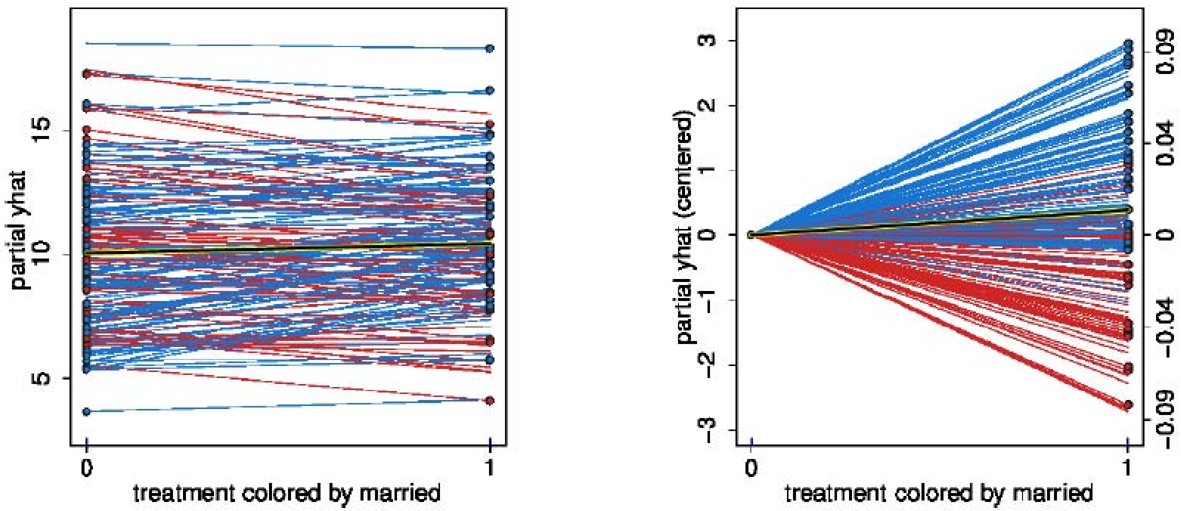


Example from the ICEbox paper

Depression clinical trial (DeRubeis et al., 2014).
The response variable is the Hamilton Depression Rating Scale.

The goal of the analysis in DeRubeis et al. (2014) is to understand how different subjects respond to different treatments (here only two treatments: 0,1), conditional on their personal covariates.

The response was modeled best as a function of the 37 covariates as well as treatment using the black-box algorithm BART.



(a) ICE (b) c-ICE