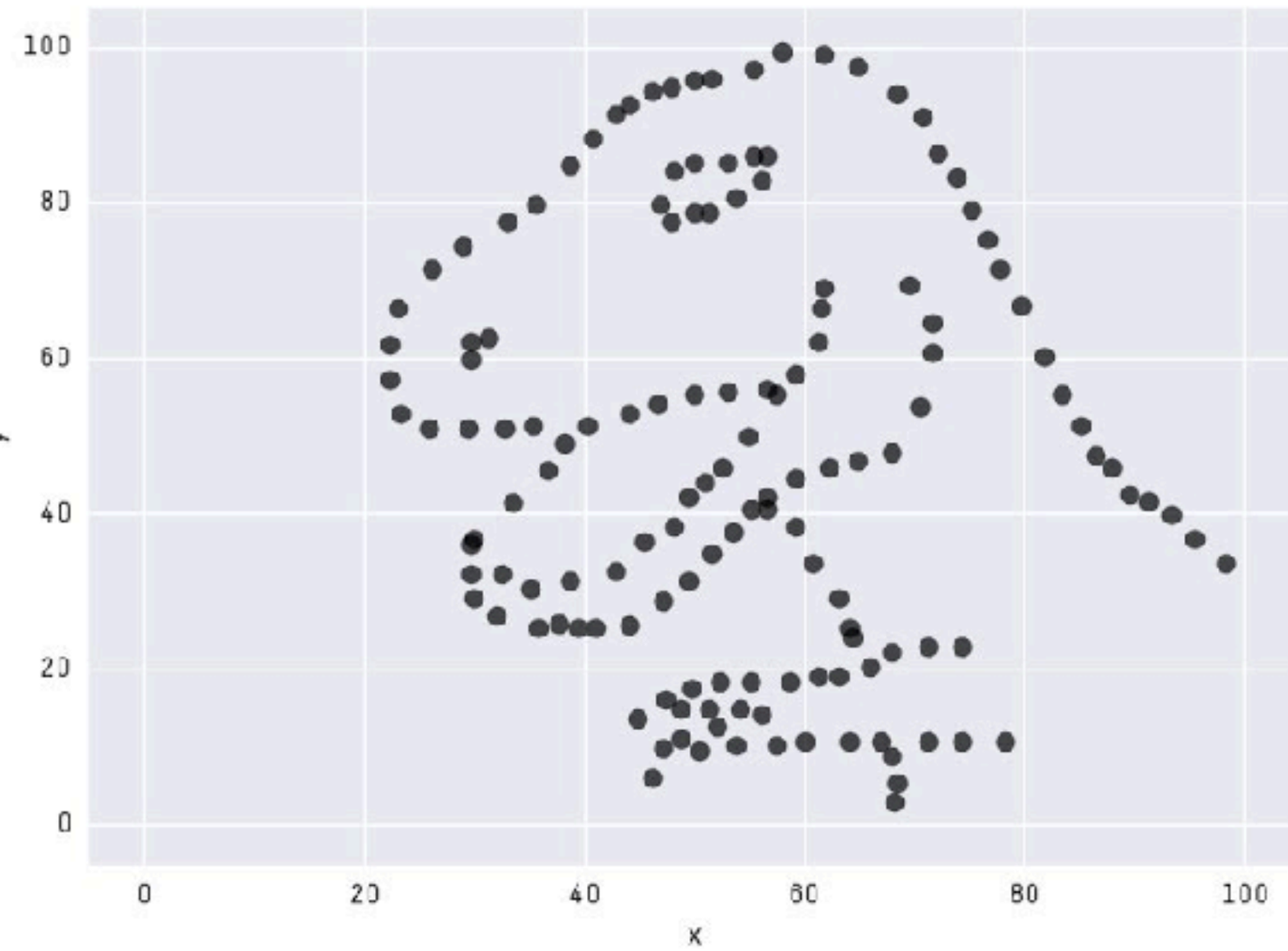
Show me your predictive model

Why and how?

Przemysław Biecek
University of Warsaw
Warsaw University of Technology

Do we really need plots to understand numbers?

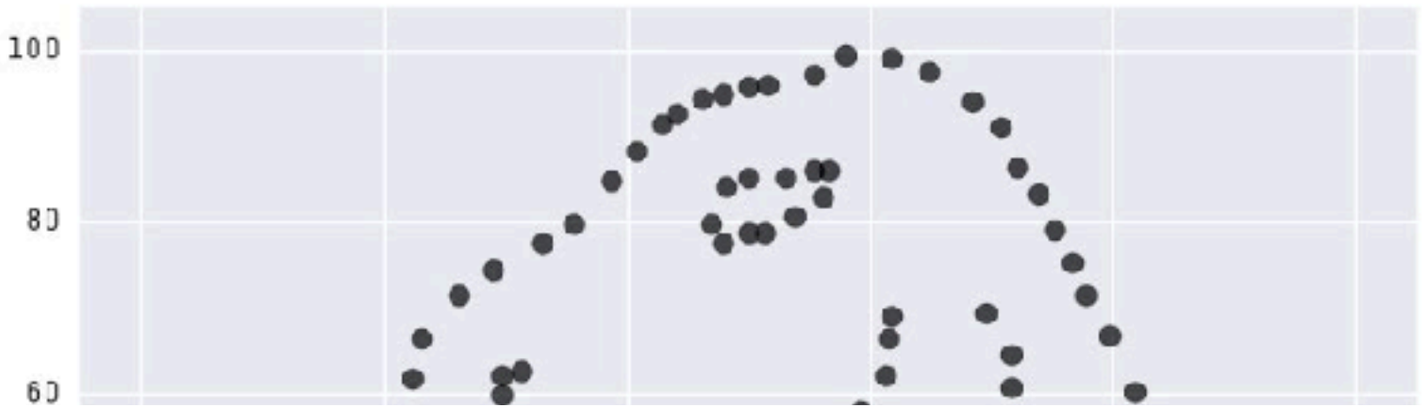
Do we really need plots to understand numbers?



X Mean: 54.2632025
Y Mean: 47.8315781
X SD : 16.7650109
Y SD : 26.9353144
Corr. : -0.0645195

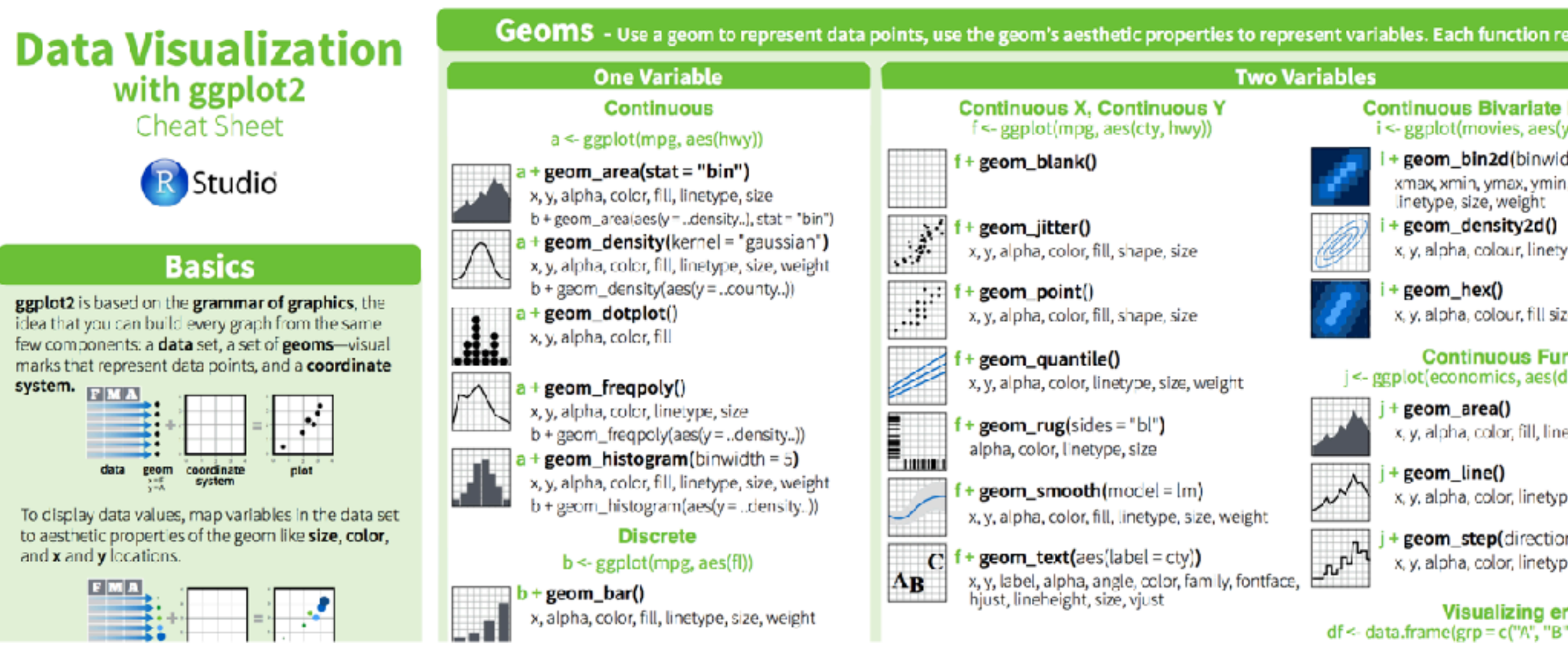
Package datasauRus
https://www.autodeskresearch.com/publications/samestats

Do we really need plots to understand numbers?



X Mean: 54.26320 25

Y Mean: 47.83157 81



Do we need plots to understand models?

Do we need plots to understand models?

1) …there is a lot of opportunity to do visualization for machine learning. Even **many of the people working in the field don't have good intuitions for how their systems work**, and they need tools to inspect what they're doing, debug, etc…

https://eagereyes.org/blog/2017/eurovis-2017-conference-report-part-1
Robert Kosara

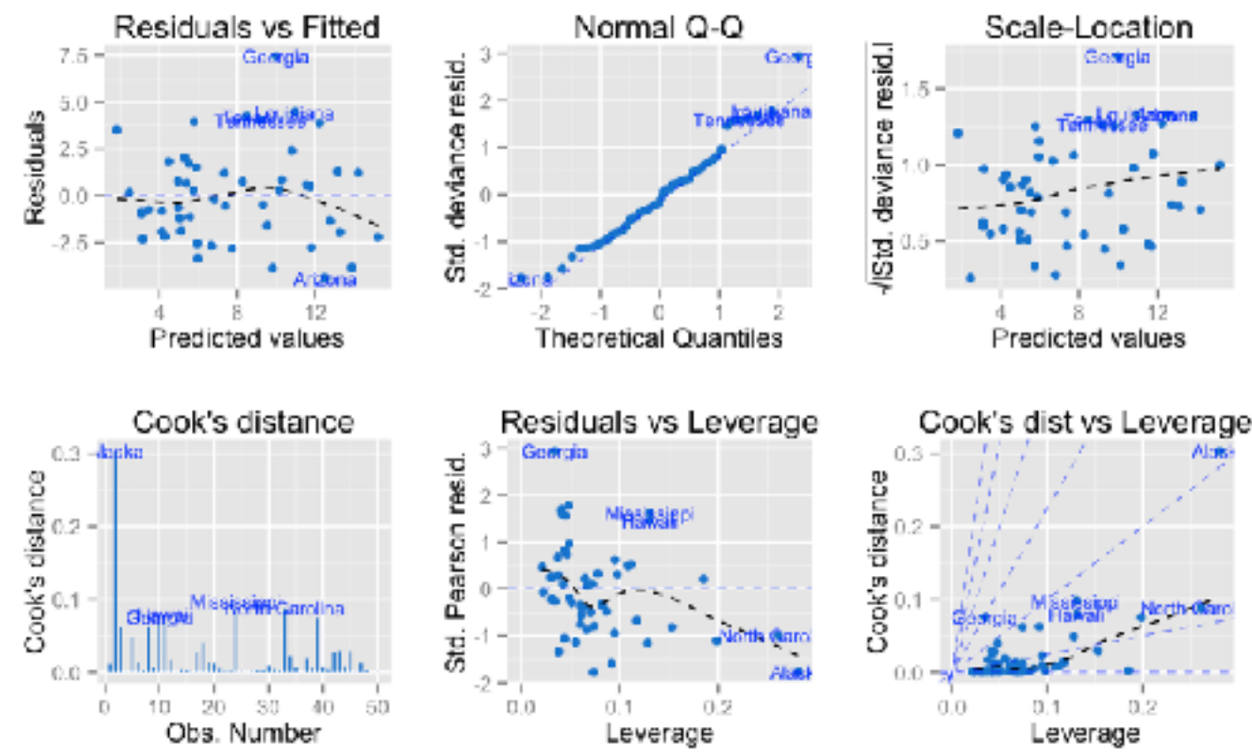
2) Understanding and trust - we need to **understand models that makes important decisions**.

"Why Should I Trust You?": Explaining the Predictions of Any Classifier
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin (2016)

3) Models have assumptions, and we need **early warnings that something is wrong** with them.

Package ggfortify

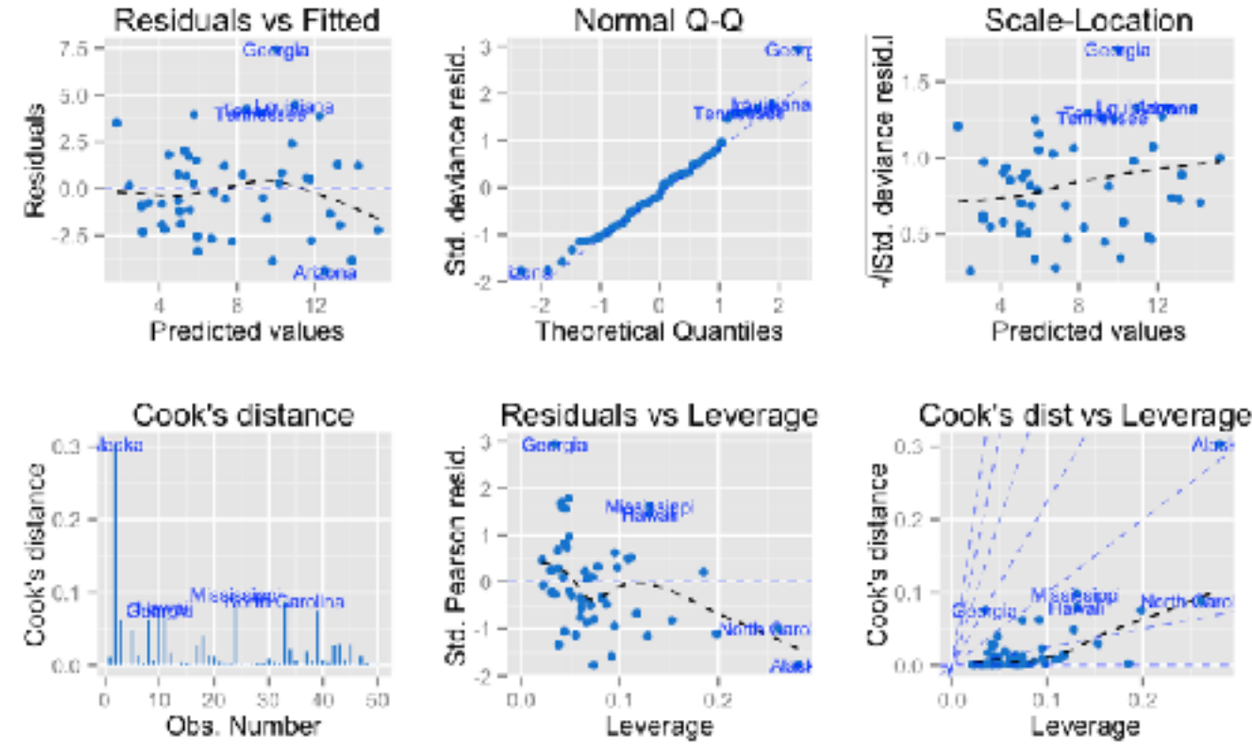
lm



Yuan Tang, Masaaki Horikoshi, and Wenxuan Li.
"ggfortify: Unified Interface to Visualize Statistical
Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.

Package ggfortify

lm



PCA

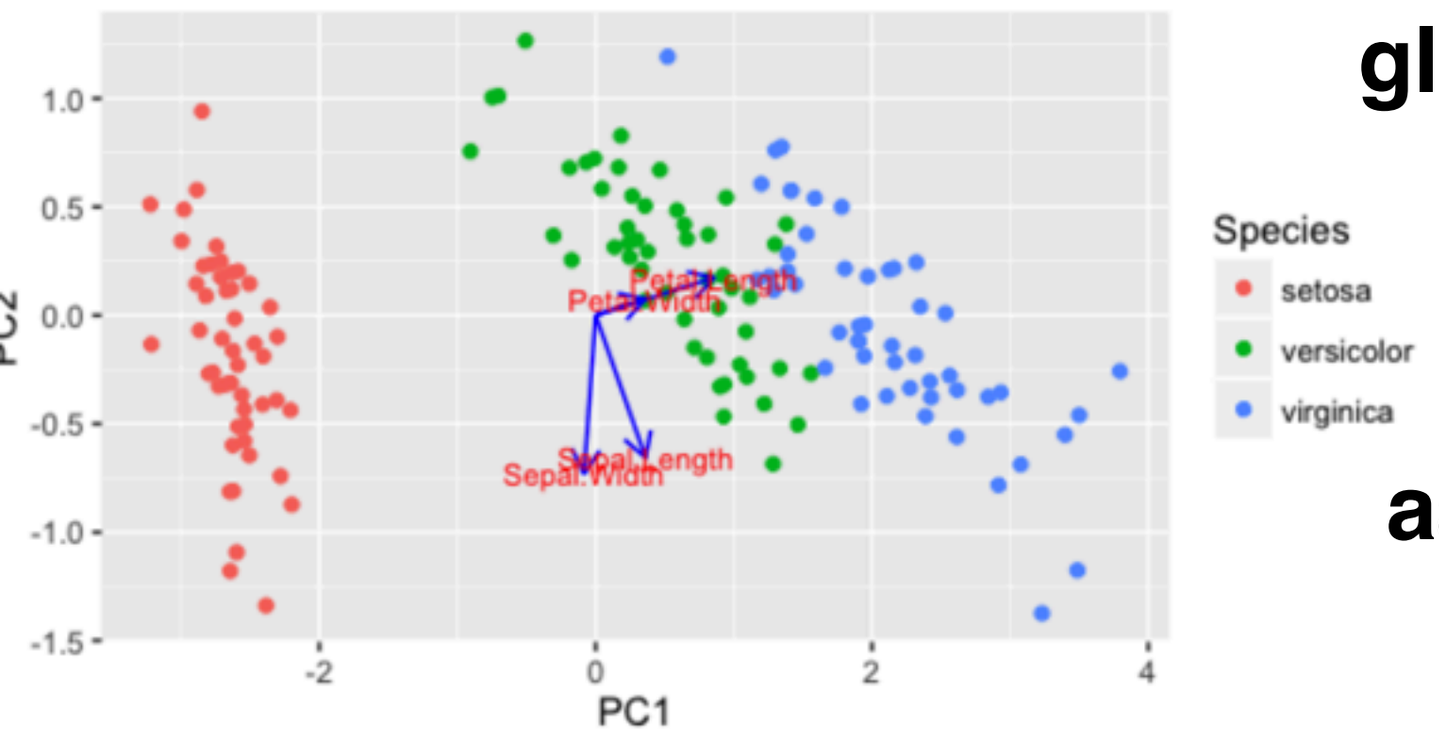
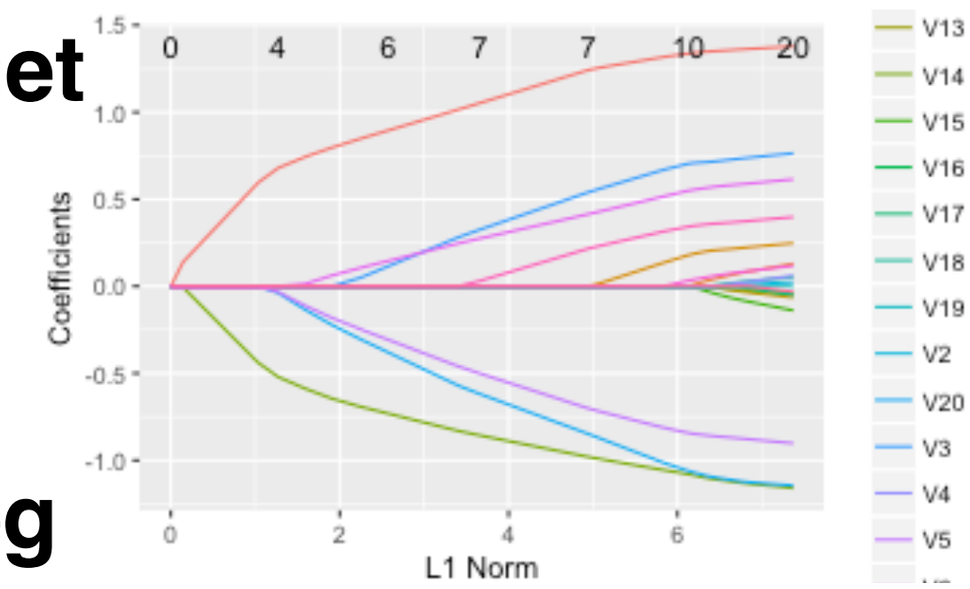
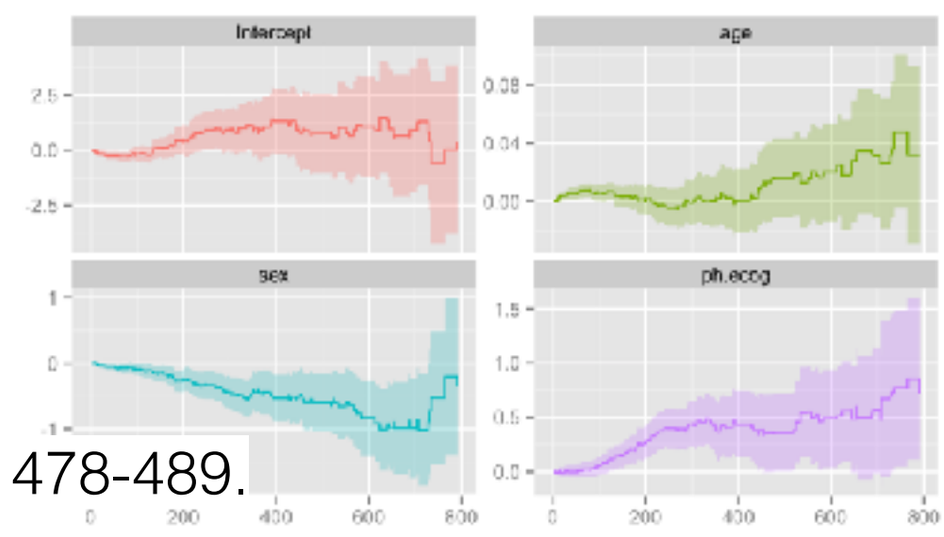


Figure 3: PCA with eigen-vectors and labels.

glmnet



aareg



Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.

Package ggfortify

lm

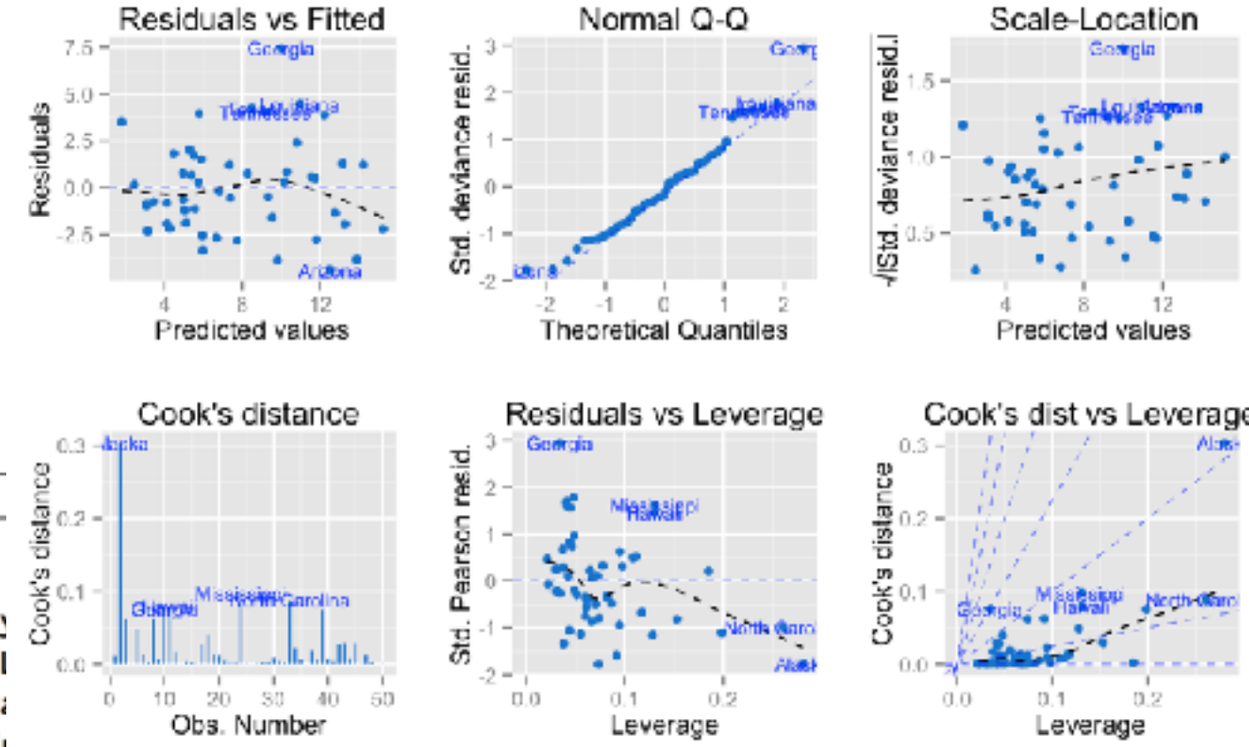
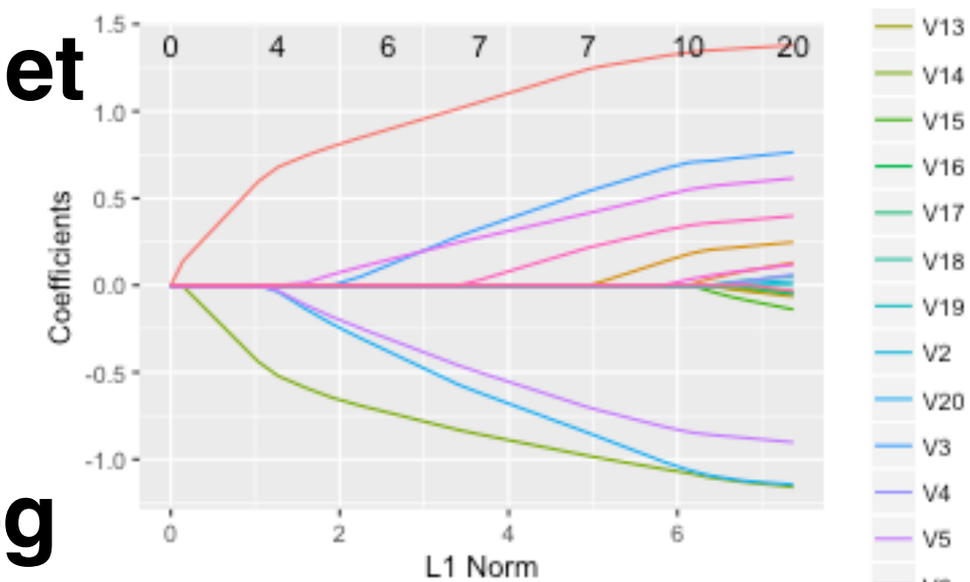


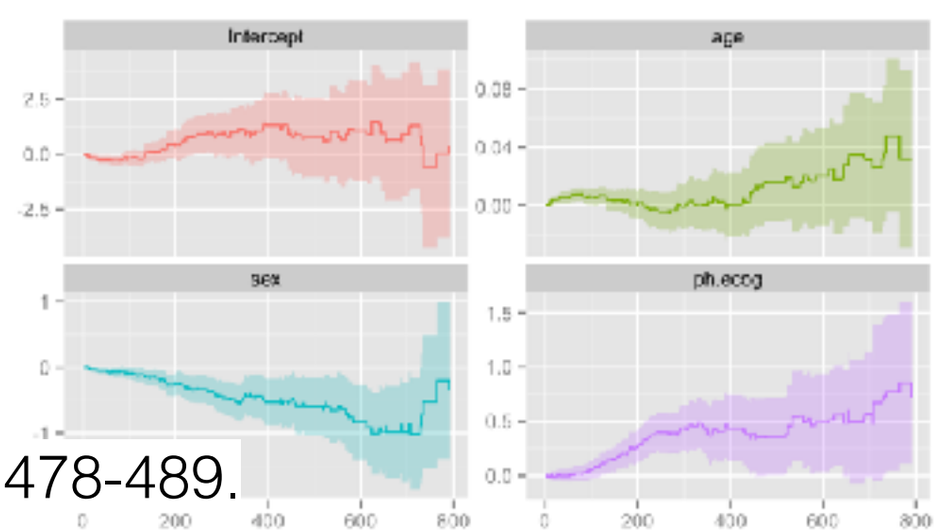
Table 1: Supported packages

package	supported types	package	supported types
base	"matrix", "table"	**sp**	"SpatialPoints", "SpatialPolygons", "Line", "Lines", "Poly "Polygons", "SpatialL "SpatialLinesDataFr "SpatialPointsDataF "SpatialPolygonsDataFrame"
cluster	"clara", "fanny", "pam"	**stats**	"HoltWinters", "lm", "acf", "ar", "Arima", "stepfun", "stl", "ts", "cmdscale", "decomposed.ts", "density", "factanal", "glm", "kmeans", "princomp", "spec"
changepoint	"cpt"	**survival**	"survfit", "survvfit.cox"
dlm	"dlmFilter", "dlmSmooth"	**strucchange**	"breakpoints", "breakpointsfull"
fGarch	"fGARCH"	**timeSeries**	"timeSeries"
forecast	"bats", "forecast", "ets", "nnetar"	**tseries**	"irts"
fracdiff	"fracdiff"	**vars**	"varprd"
glmnet	"cv.glmnet", "glmnet"	**xts**	"xts"
KFAS	"KFS", "signal"	**zoo**	"zooreg"
lfda	"lfda", "klfda", "self"	**MASS**	"isoMDS", "sammon"
maps	"map"		

glmnet

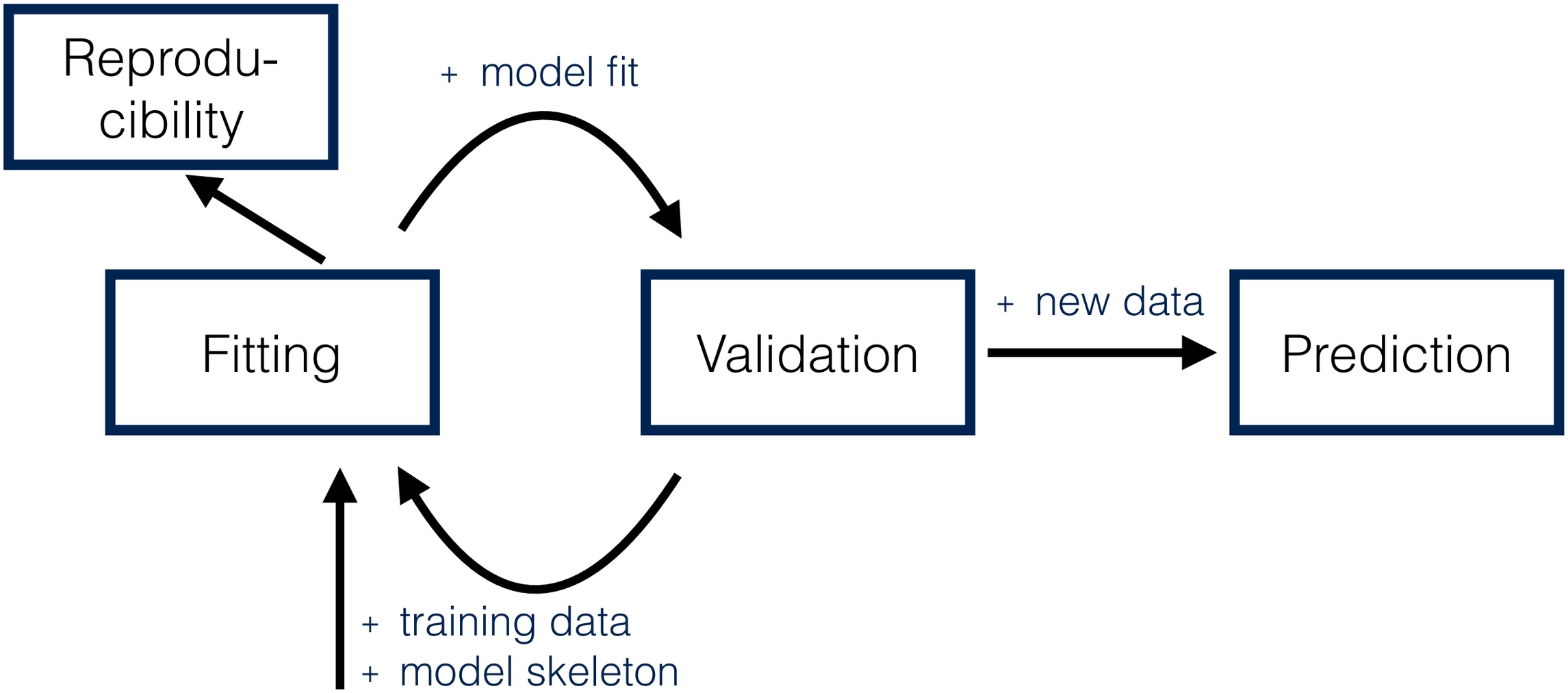


aareg

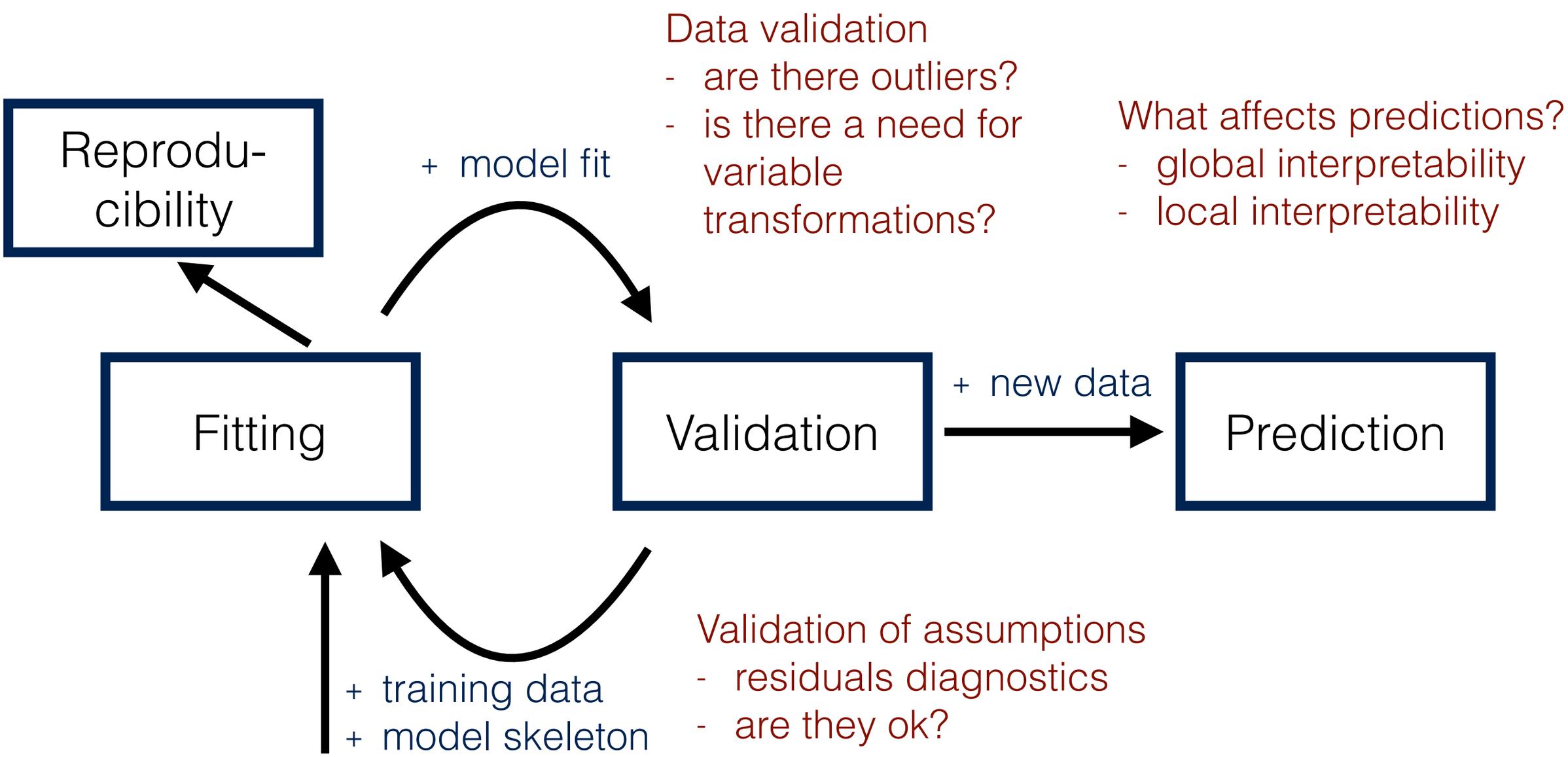


Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.

Life-cycle of a typical prognostic model



Life-cycle of a typical prognostic model



What are estimates
for model parameters?

Are convergence
criteria satisfied?

Performance charts,
is it a good model?
- models comparisons
- what is the predictive performance?

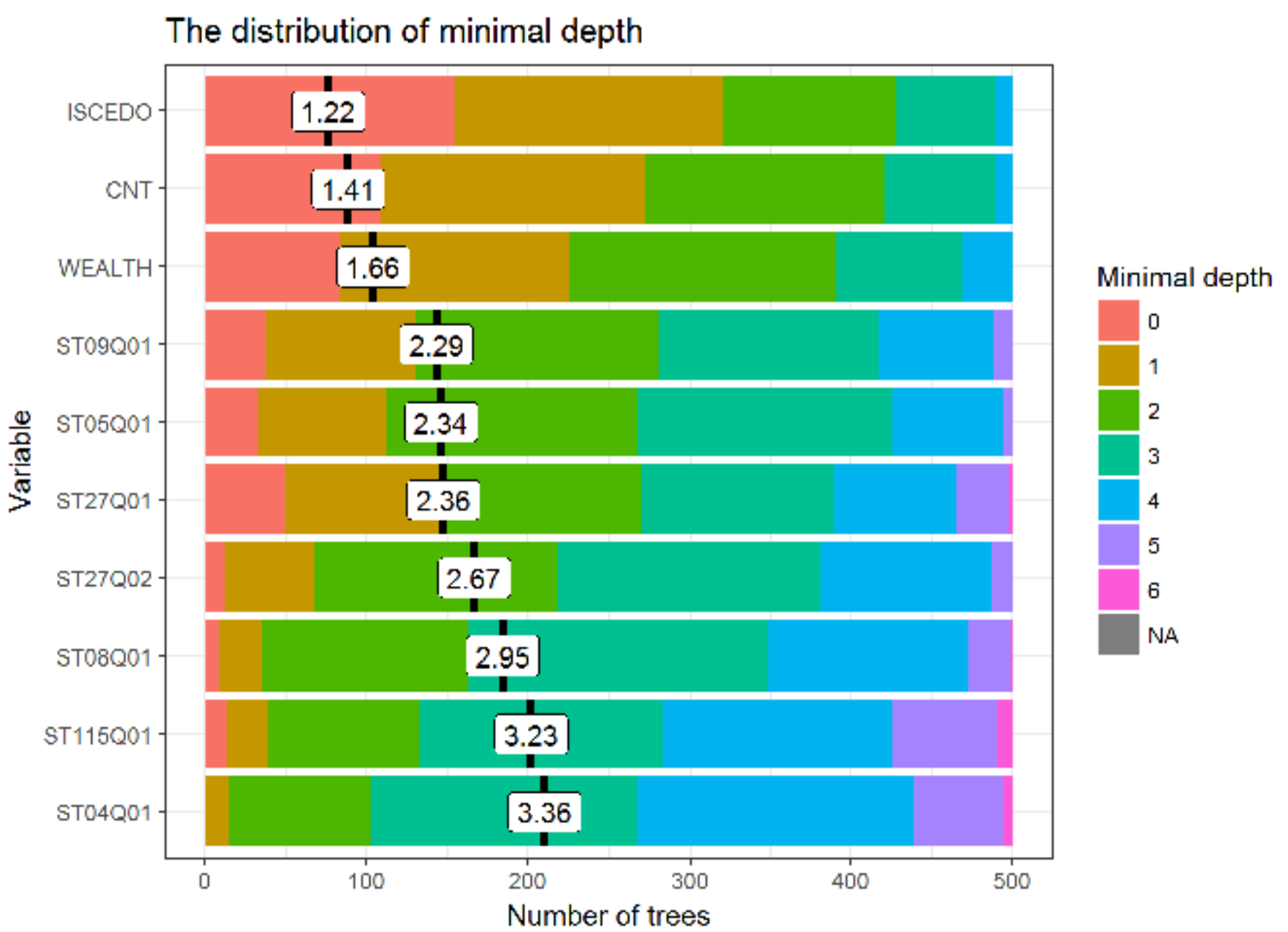
What affects the prediction for a single observation?
(local interpretability)

Prediction

Which elements of the model are the most important ones?
(global interpretability)

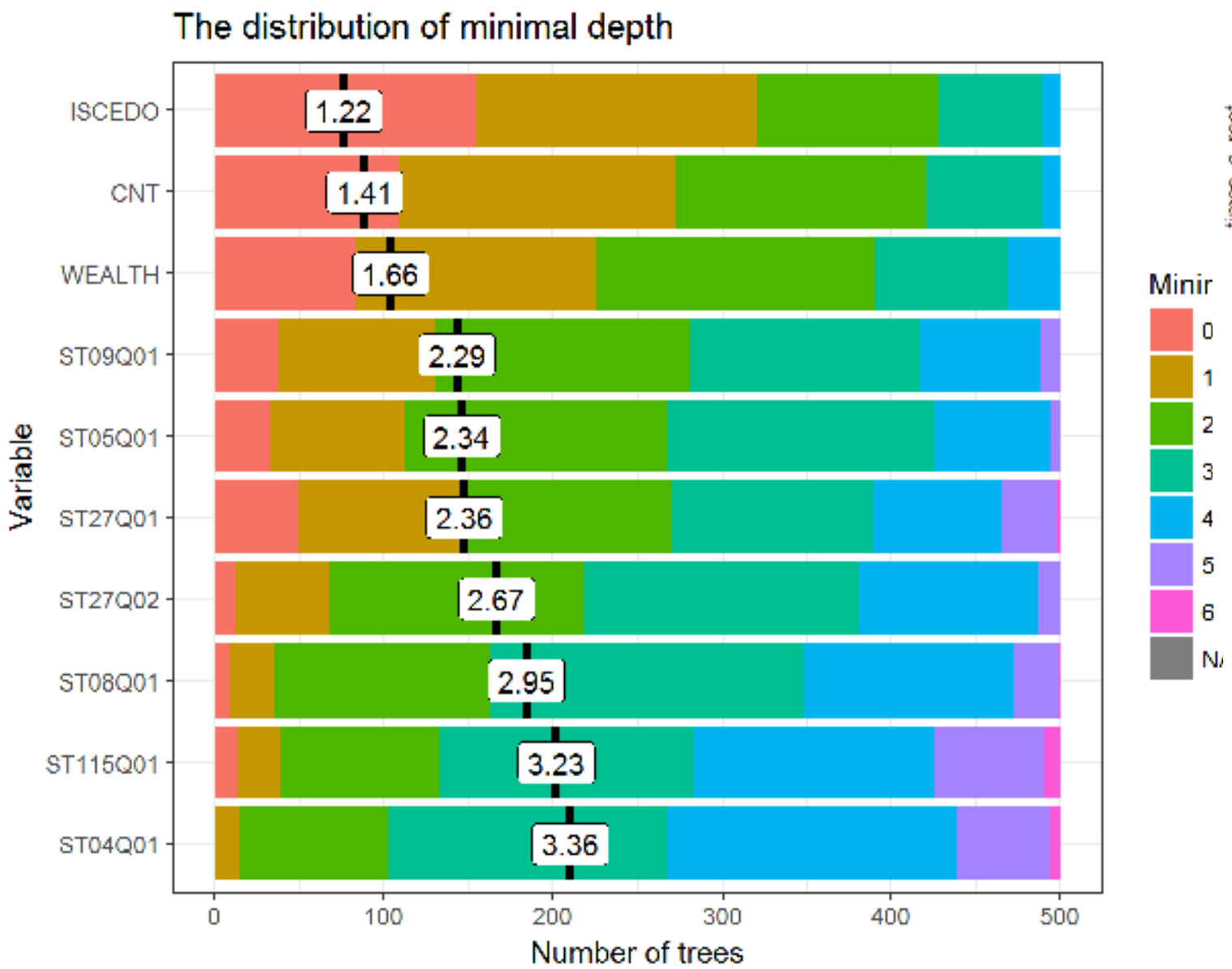
How a single variable affects the expected output?
(marginal interpretability)

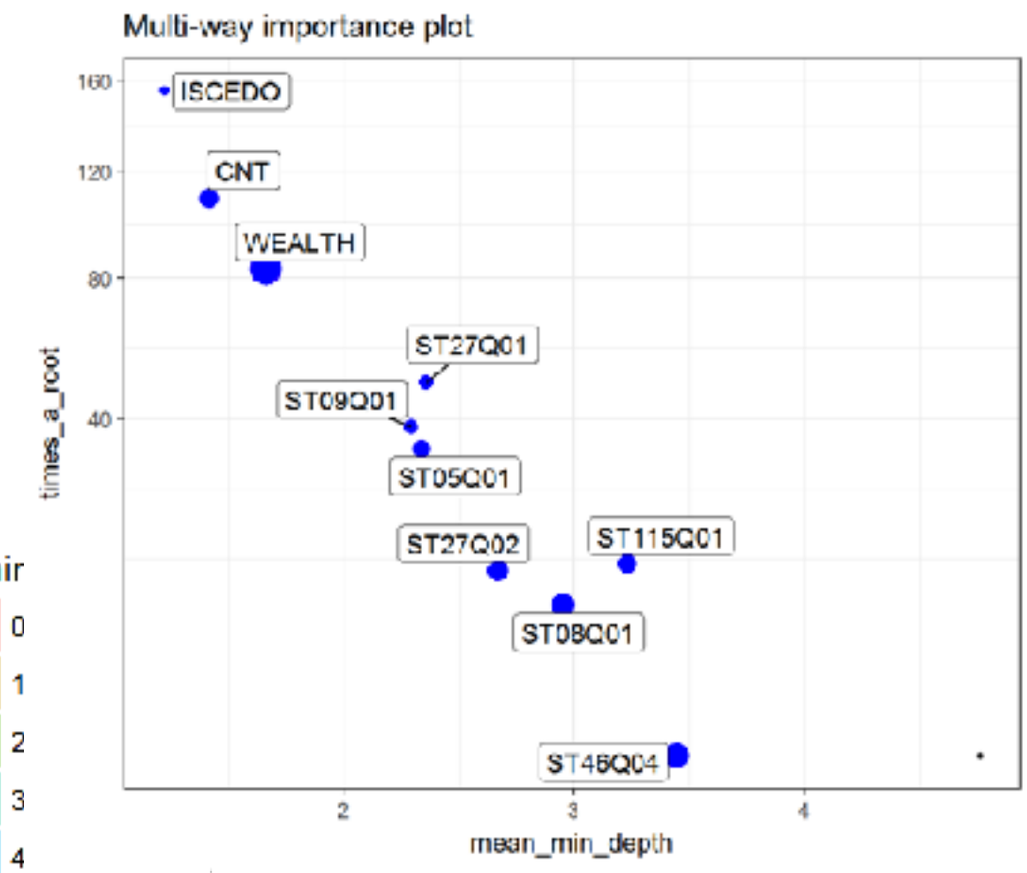
Package randomForestExplainer
What is in a random forest?

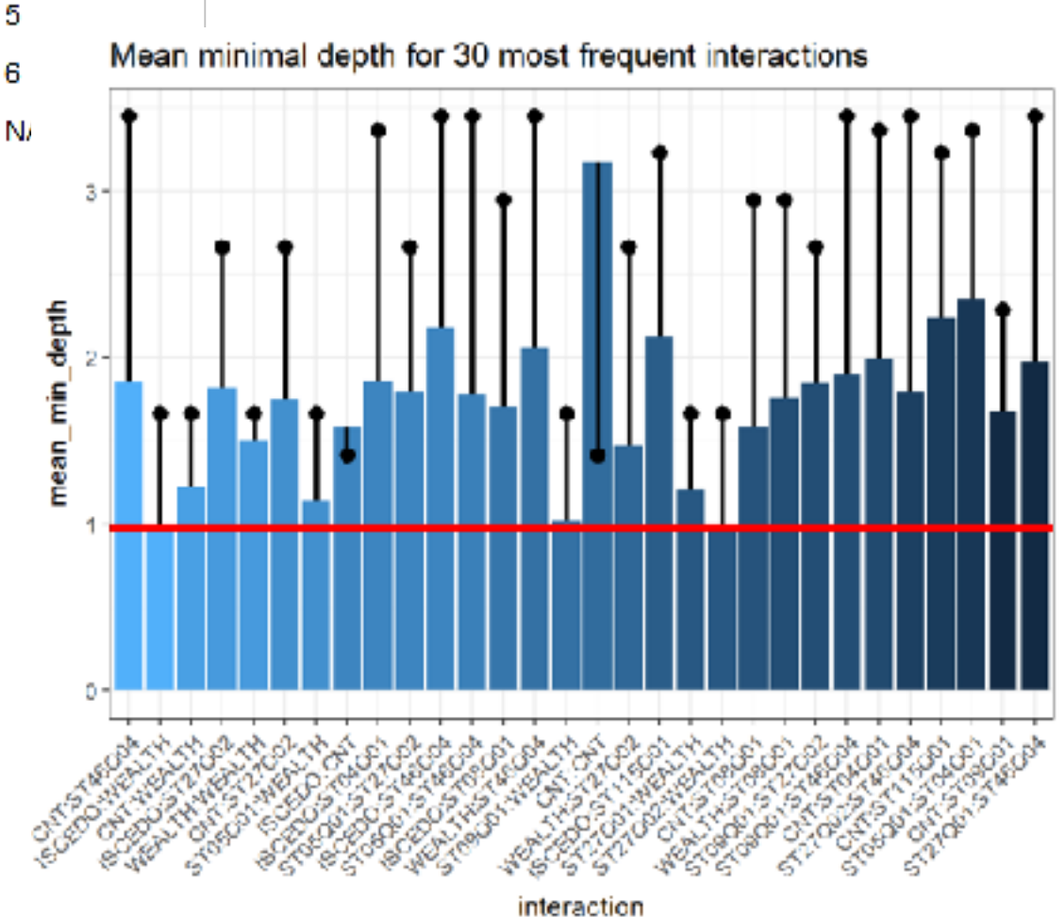


Aleksandra Paluszynska, Przemyslaw Biecek (2017)
https://github.com/geneticsMiNIng/BlackBoxOpener
John Ehrlinger (2015)
ggRandomForests: Random Forests for Regression

Package randomForestExplainer
What is in a random forest?

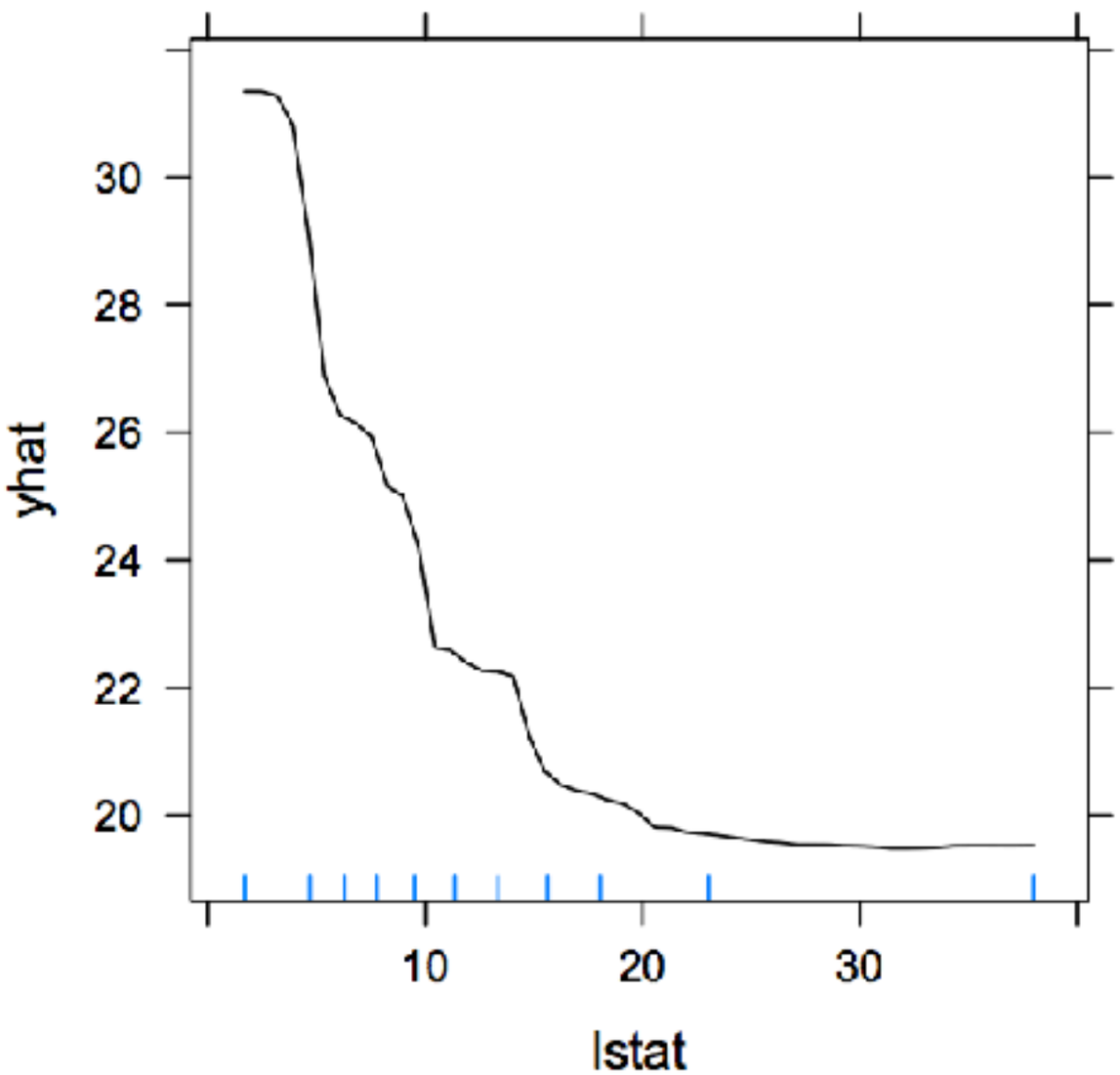


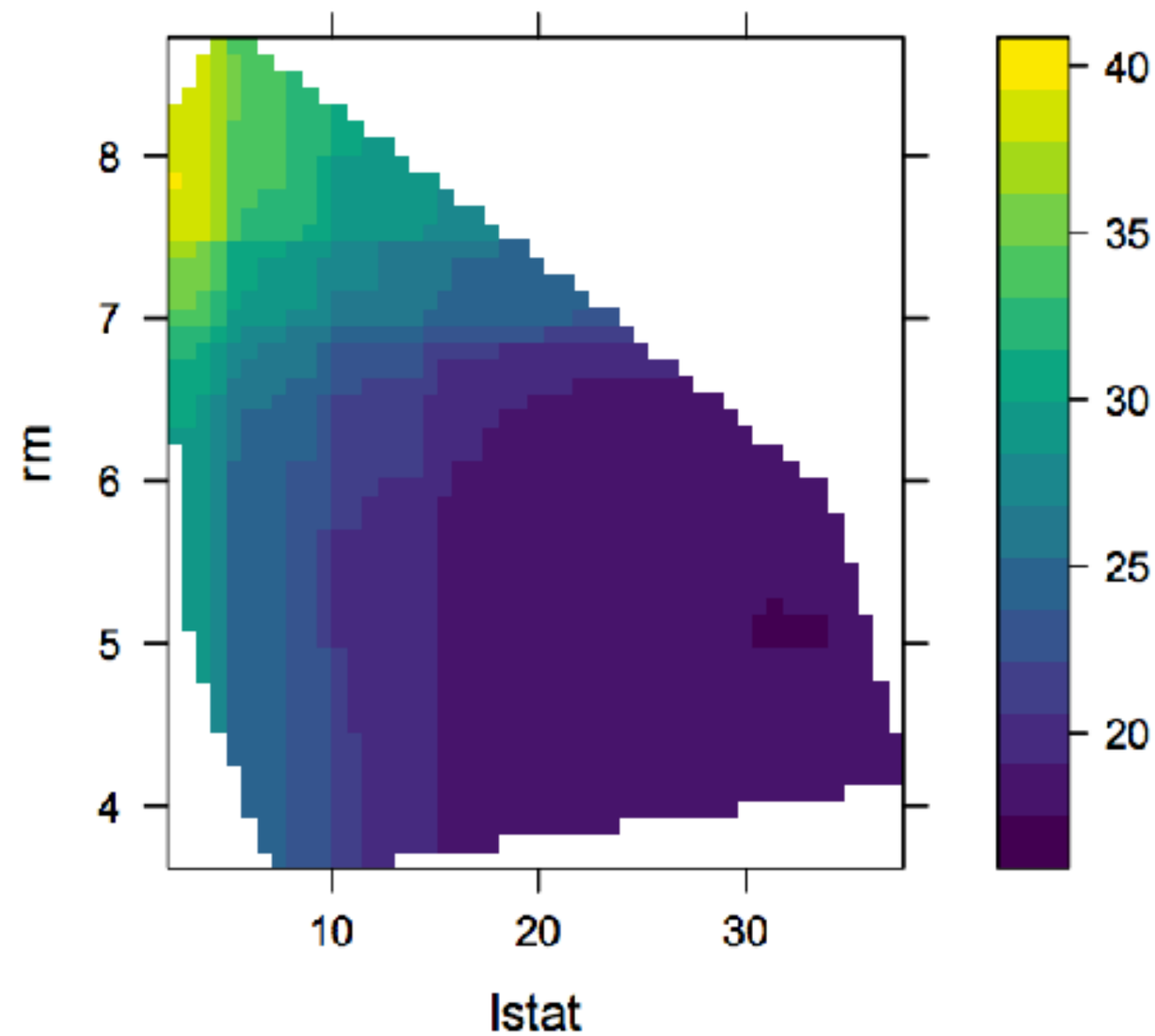




Aleksandra Paluszynska, Przemyslaw Biecek (2017)
https://github.com/geneticsMiNIng/BlackBoxOpener
John Ehrlinger (2015)
ggRandomForests: Random Forests for Regression

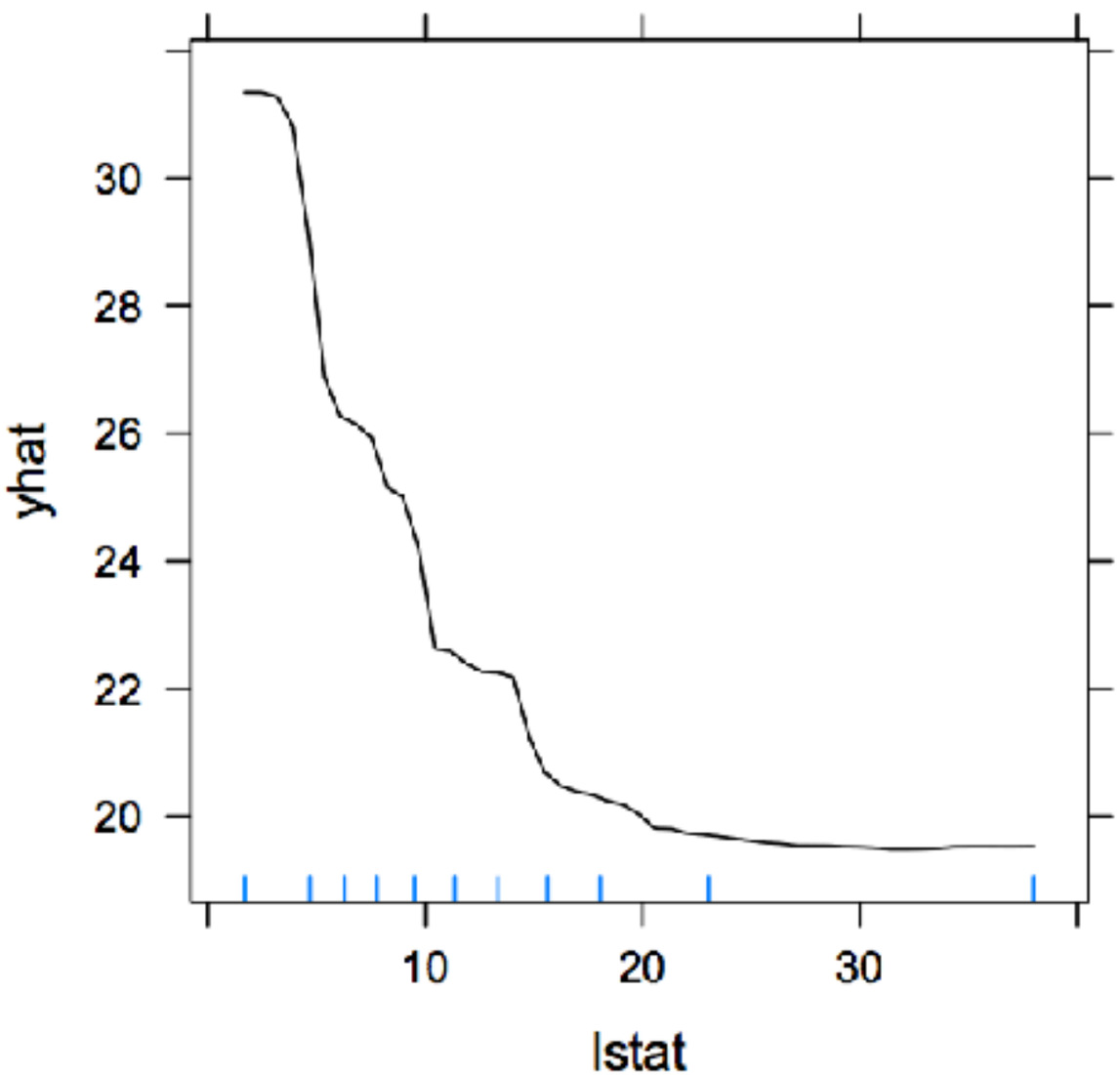
Package pdp - Partial Dependence Plots





pdp: An R Package for Constructing Partial Dependence Plots. Brandon M. Greenwell (2017)
https://journal.r-project.org/archive/2017/RJ-2017-016/index.html

Package pdp - Partial Dependence Plots



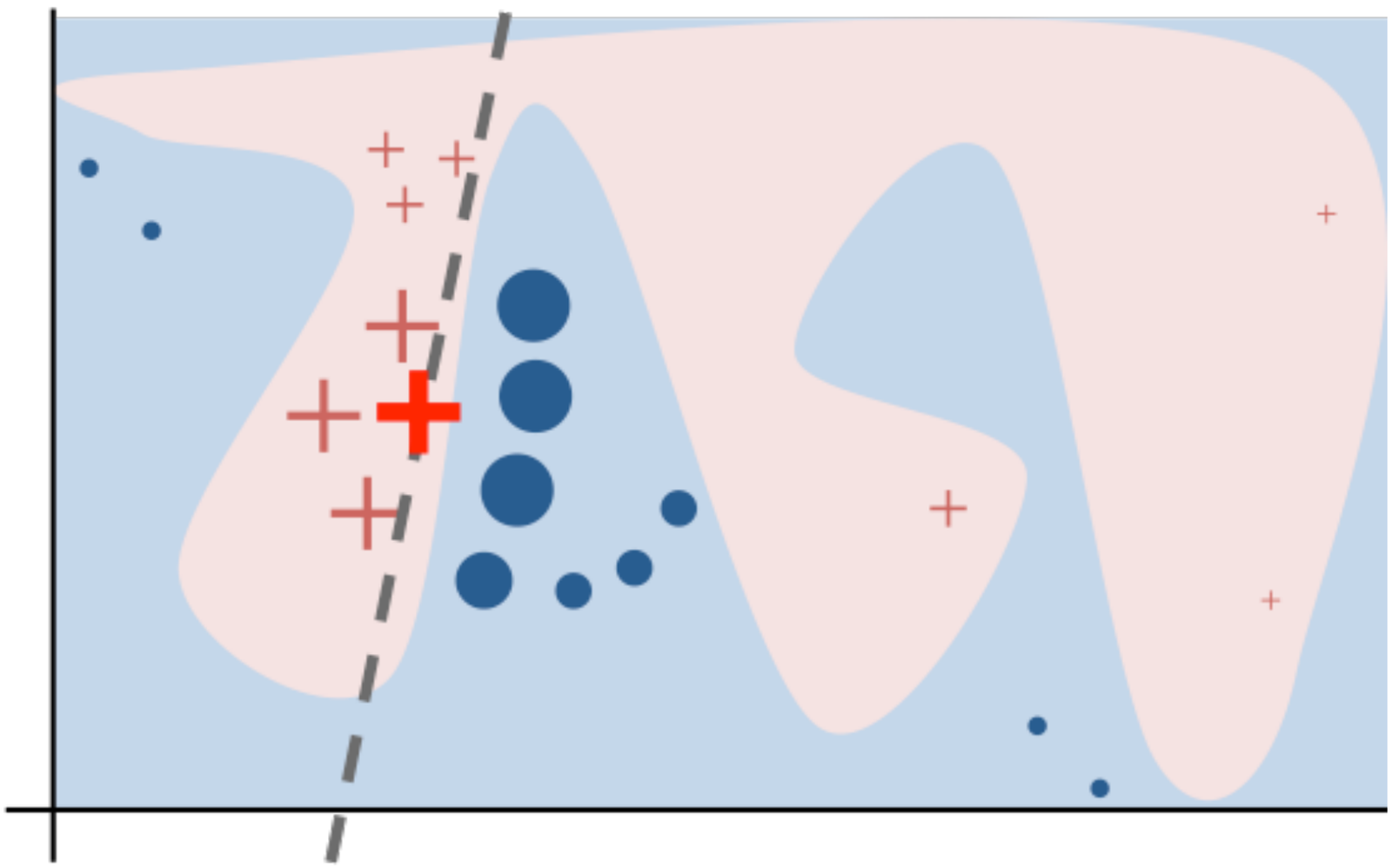
Type of model	R package	Object class
Decision tree	C50 (Kuhn et al., 2015)	"C5.0"
	party	"BinaryTre
	partykit	"party"
	rpart (Therneau et al., 2015)	"rpart"
Bagged decision trees	adabag (Alfaro et al., 2013)	"bagging"
	ipred (Peters and Hothorn, 2015)	"classbagg
		"regbagg"
Boosted decision trees	adabag (Alfaro et al., 2013)	"boosting"
	gbm	"gbm"
	xgboost	"xgb.Boost
Cubist	Cubist (Kuhn et al., 2014)	"cubist"
Discriminant analysis	MASS (Venables and Ripley, 2002)	"lda", "qda
Generalized linear model	**stats**	"glm", "lm"
Linear model	**stats**	"lm"
Nonlinear least squares	**stats**	"nls"
Multivariate adaptive regression splines (MARS)	earth (Milborrow, 2016)	"earth"
	mda (Leisch et al., 2016)	"mars"
Projection pursuit regression	**stats**	"ppr"
Random forest	**randomForest**	"randomFor
	party	"RandomFor
	partykit	"cforest"
	ranger (Wright, 2016)	"ranger"
Support vector machine	e1071 (Meyer et al., 2015)	"svm"
	kernlab (Karatzoglou et al., 2004)	"ksvm"

Table 1: Models specifically supported by the **pdp** package. **Note:** for some of these ca
may still need to supply additional arguments in the call to `partial`.

pdp: An R Package for Constructing Partial Dependence Plots. Brandon M. Greenwell (2017)
https://journal.r-project.org/archive/2017/RJ-2017-016/index.html

LIME: Local Interpretable Model-agnostic Explanations

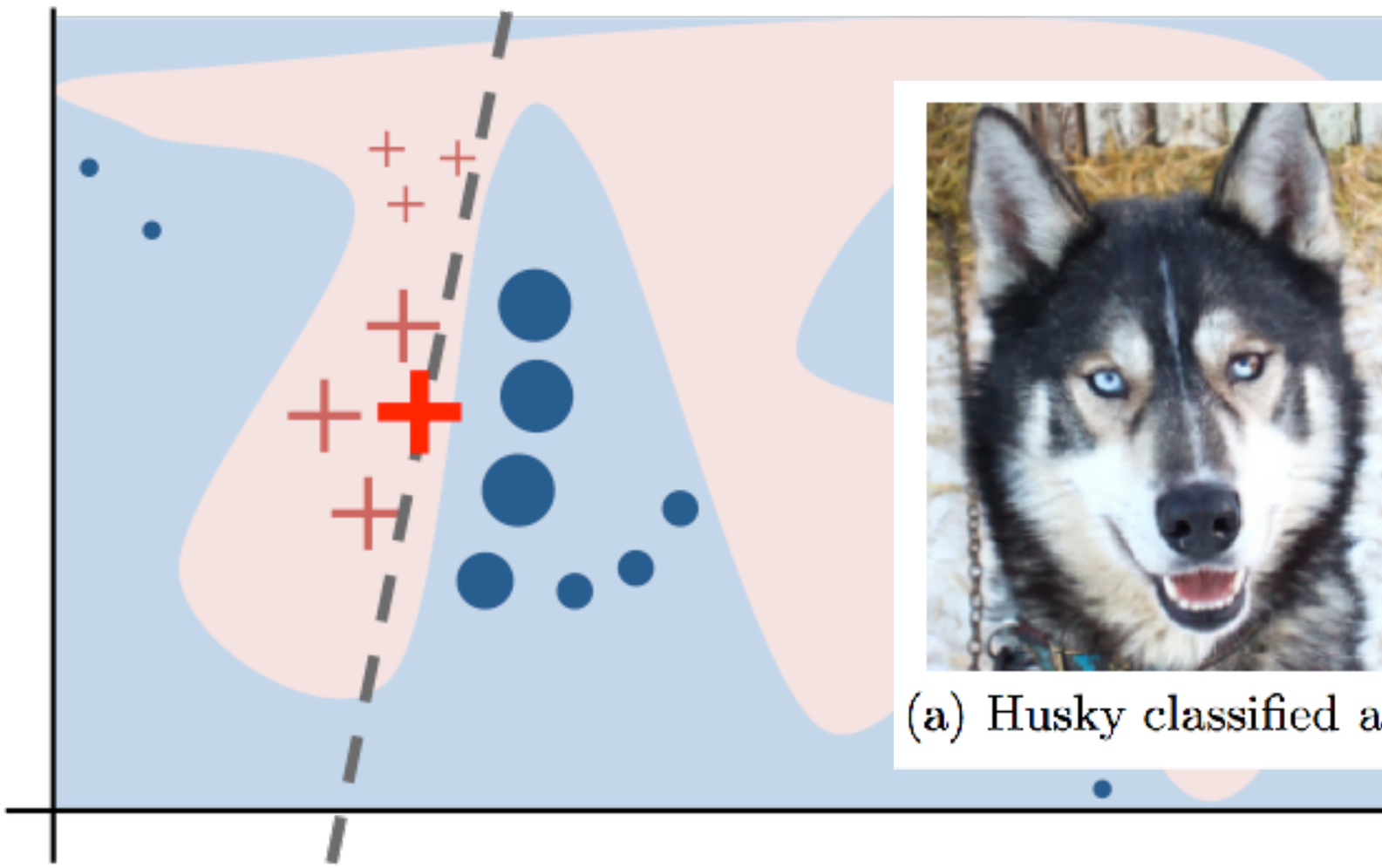
1. Generate a fake dataset around x.
2. Use black-box estimator to get target values y.
3. Train a new white-box estimator for (y,x).
4. Check prediction quality of a white-box classifier.
5. Use white-box estimator as an explanation of black-box model.



"Why Should I Trust You?" Explaining the Predictions of Any Classifier.
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin (2016). https://arxiv.org/pdf/1602.04938.pdf
Port to R: Thomas Lin Pedersen (2017) https://github.com/thomasp85/lime

LIME: Local Interpretable Model-agnostic Explanations

1. Generate a fake dataset around x.
2. Use black-box estimator to get target values y.
3. Train a new white-box estimator for (y,x).
4. Check prediction quality of a white-box classifier.
5. Use white-box estimator as an explanation of black-box model.



(a) Husky classified as wolf

(b) Explanation

"Why Should I Trust You?" Explaining the Predictions of Any Classifier.
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin (2016). https://arxiv.org/pdf/1602.04938.pdf
Port to R: Thomas Lin Pedersen (2017) https://github.com/thomasp85/lime

Fitting

What are estimates
of model parameters?

Are convergence
criteria satisfied?

Are the assumptions satisfied?

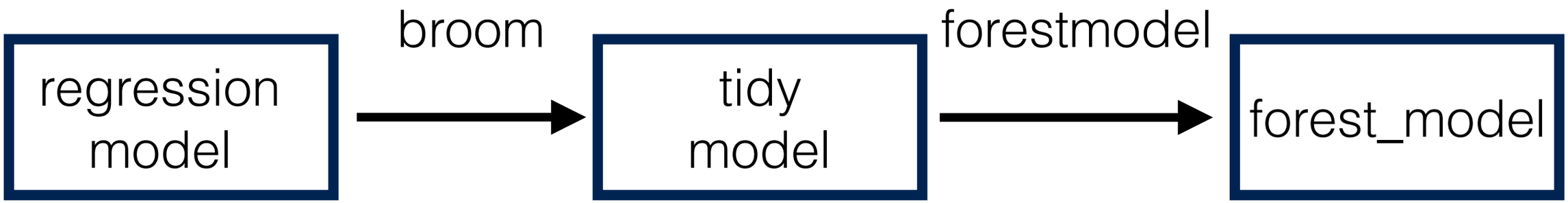
Do we need
variable transformation?

Package forestmodel

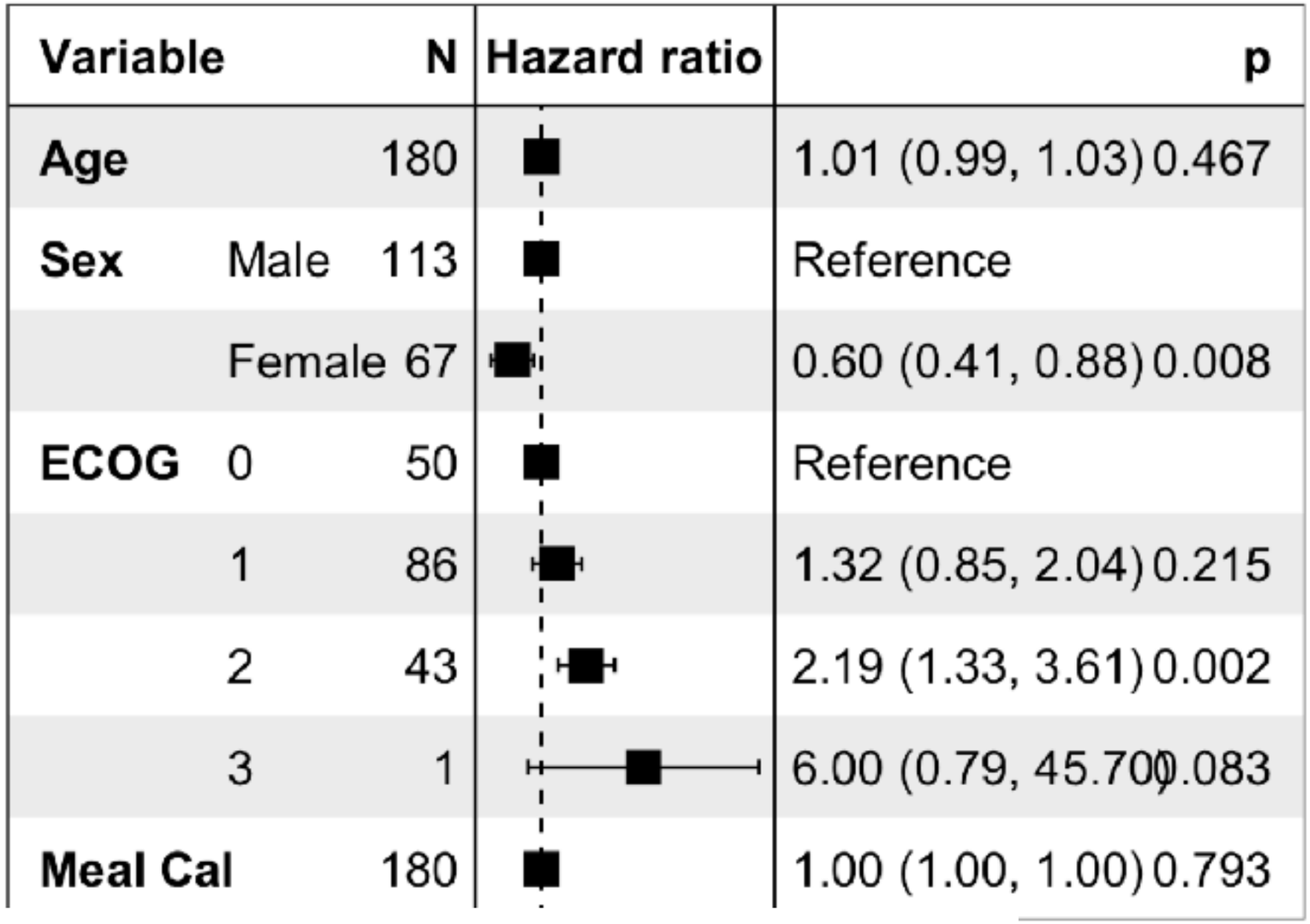
Variable		N	Hazard ratio	p
Age		180		1.01 (0.99, 1.03) 0.467
Sex	Male	113		Reference
	Female	67		0.60 (0.41, 0.88) 0.008
ECOG	0	50		Reference
	1	86		1.32 (0.85, 2.04) 0.215
	2	43		2.19 (1.33, 3.61) 0.002
	3	1		6.00 (0.79, 45.700) 0.083
Meal Cal		180		1.00 (1.00, 1.00) 0.793

broom: An R Package for Converting Statistical Analysis Objects Into Tidy Data Frames,
David Robinson (2014) arXiv:1412.3565v2
Nick Kennedy (2017) https://github.com/NikNakk/forestmodel

Package forestmodel



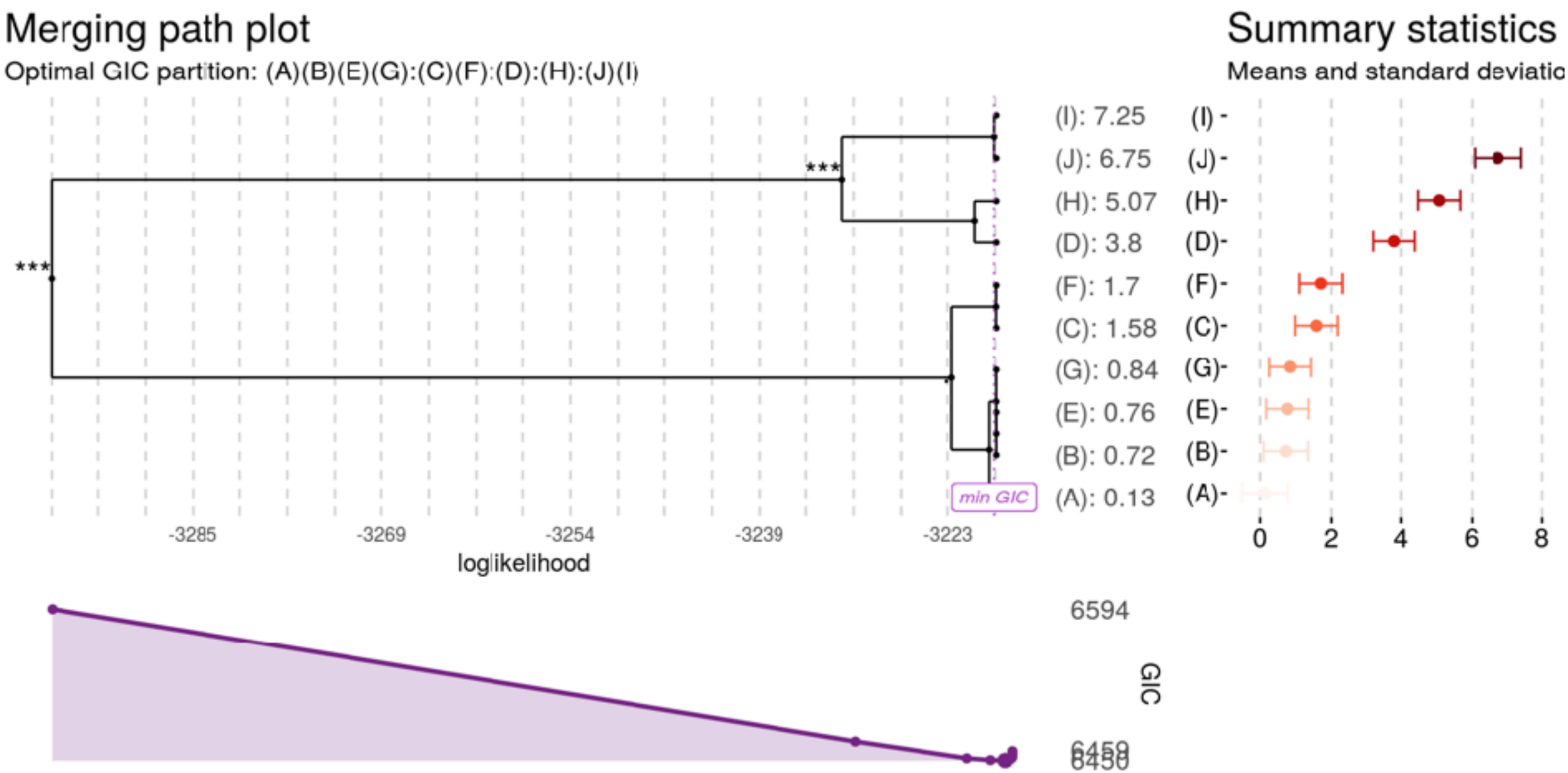
Class	tidy
aareg	x
acf	x
anova	x
aov	x
aovlist	x
Arima	x
betareg	x
biglm	x
binDesign	x
binWidth	x
boot	x
brmsfit	x



broom: An R Package for Converting Statistical Analysis Objects Into Tidy Data Frames, David Robinson (2014) arXiv:1412.3565v2

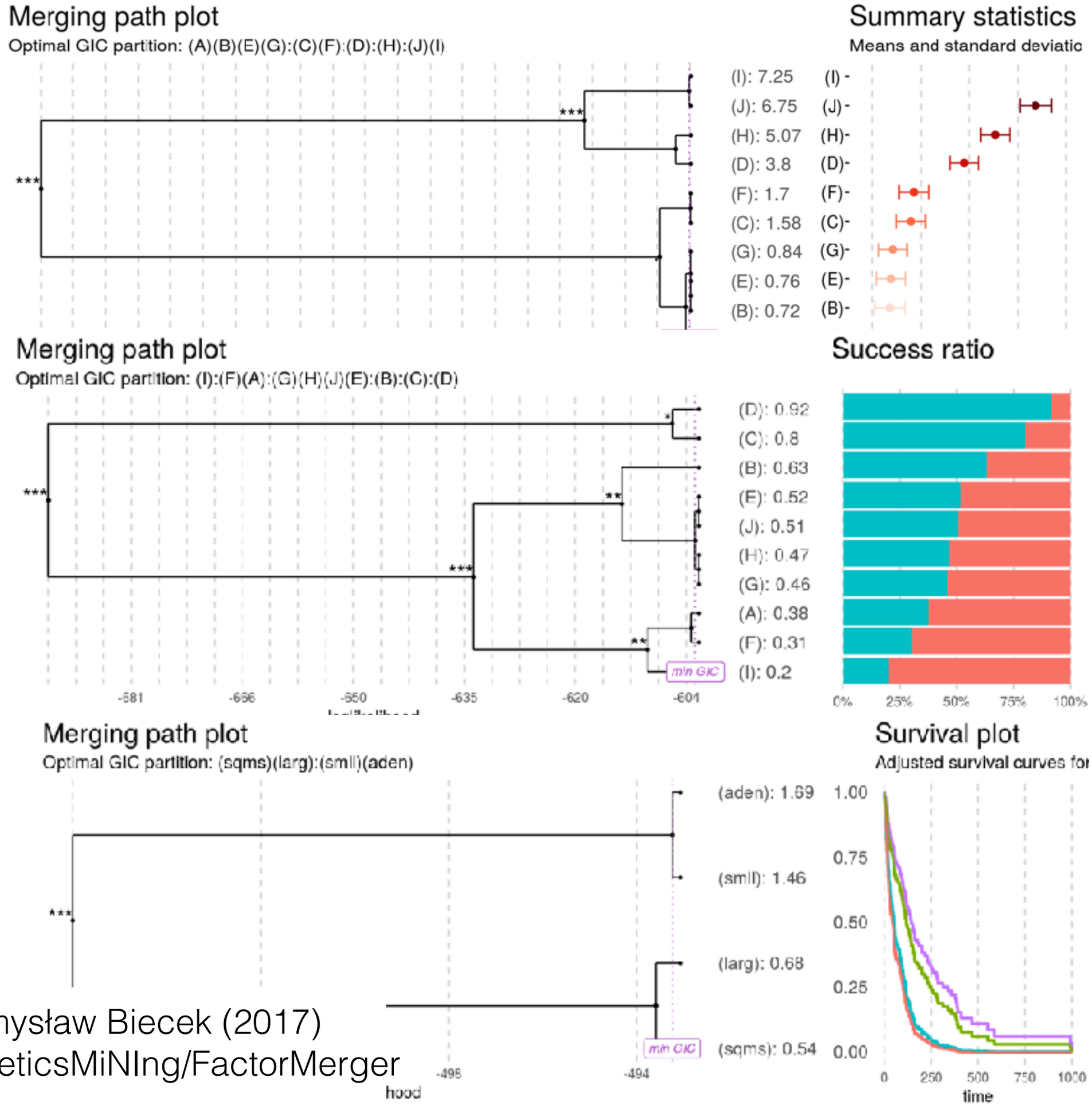
Nick Kennedy (2017) https://github.com/NikNakk/forestmodel

Package factorMerger
Visualisation for post-hoc testing



Agnieszka Sitko, Przemysław Biecek (2017)
https://github.com/geneticsMiNIng/FactorMerger

Package factorMerger for post-hoc testing



Agnieszka Sitko, Przemysław Biecek (2017)
https://github.com/geneticsMiNIng/FactorMerger

Validation

Performance charts,
is it a good model?

- model comparisons
- what is the performance?

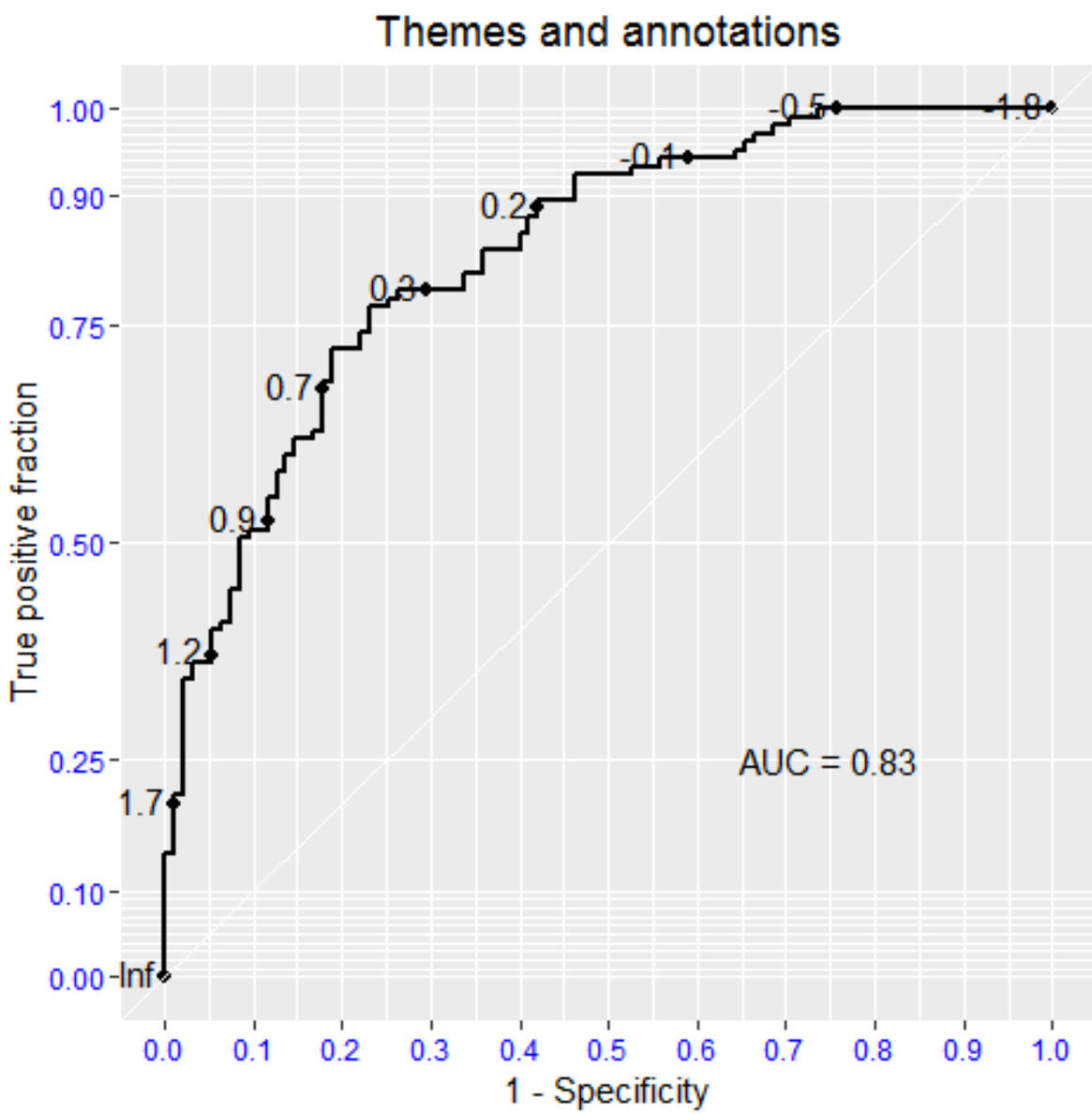
Validation of assumptions

- residuals diagnostics
- are they ok?

Data validation

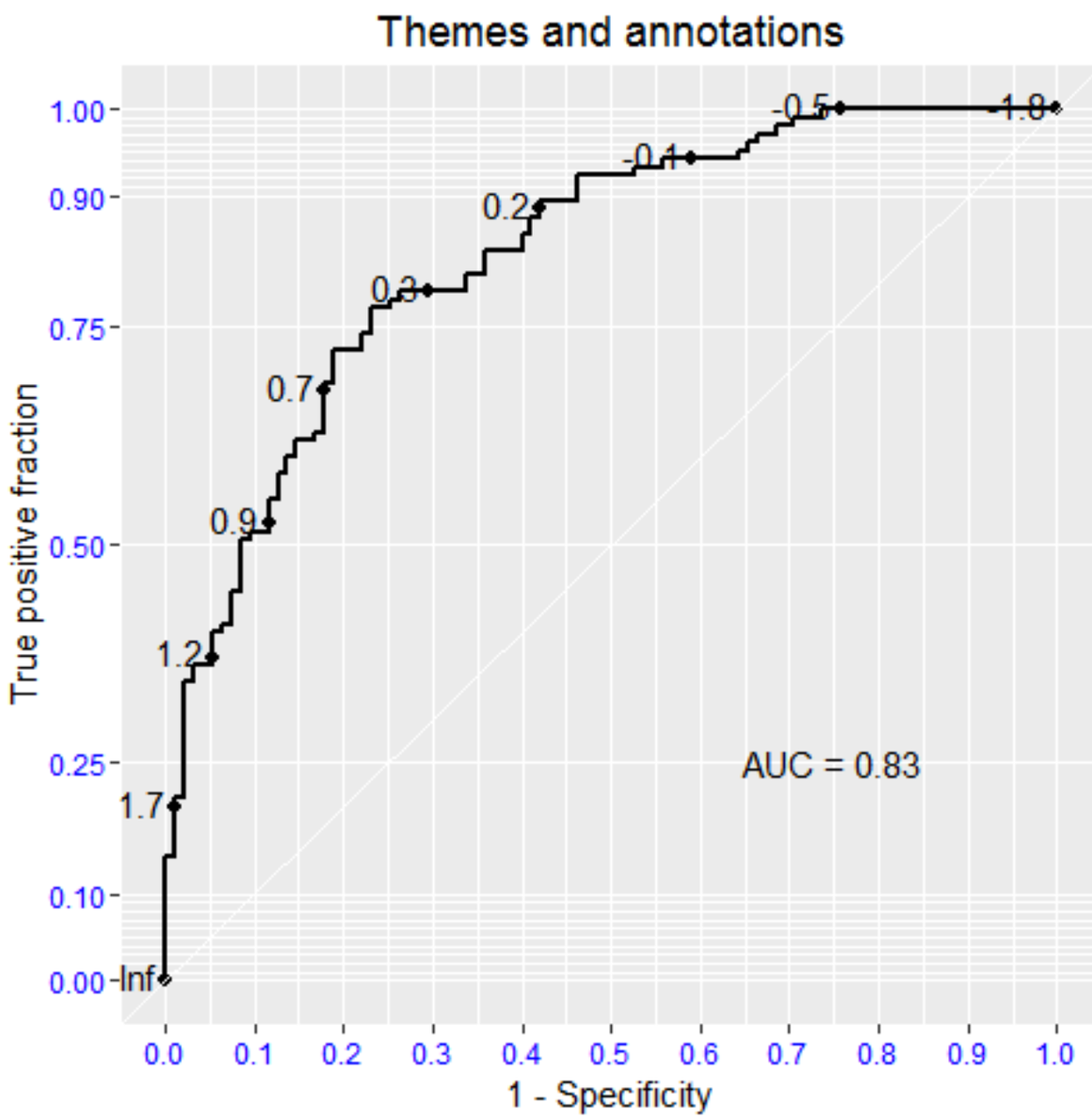
- are there outliers?
- is there a need for variable transformations?

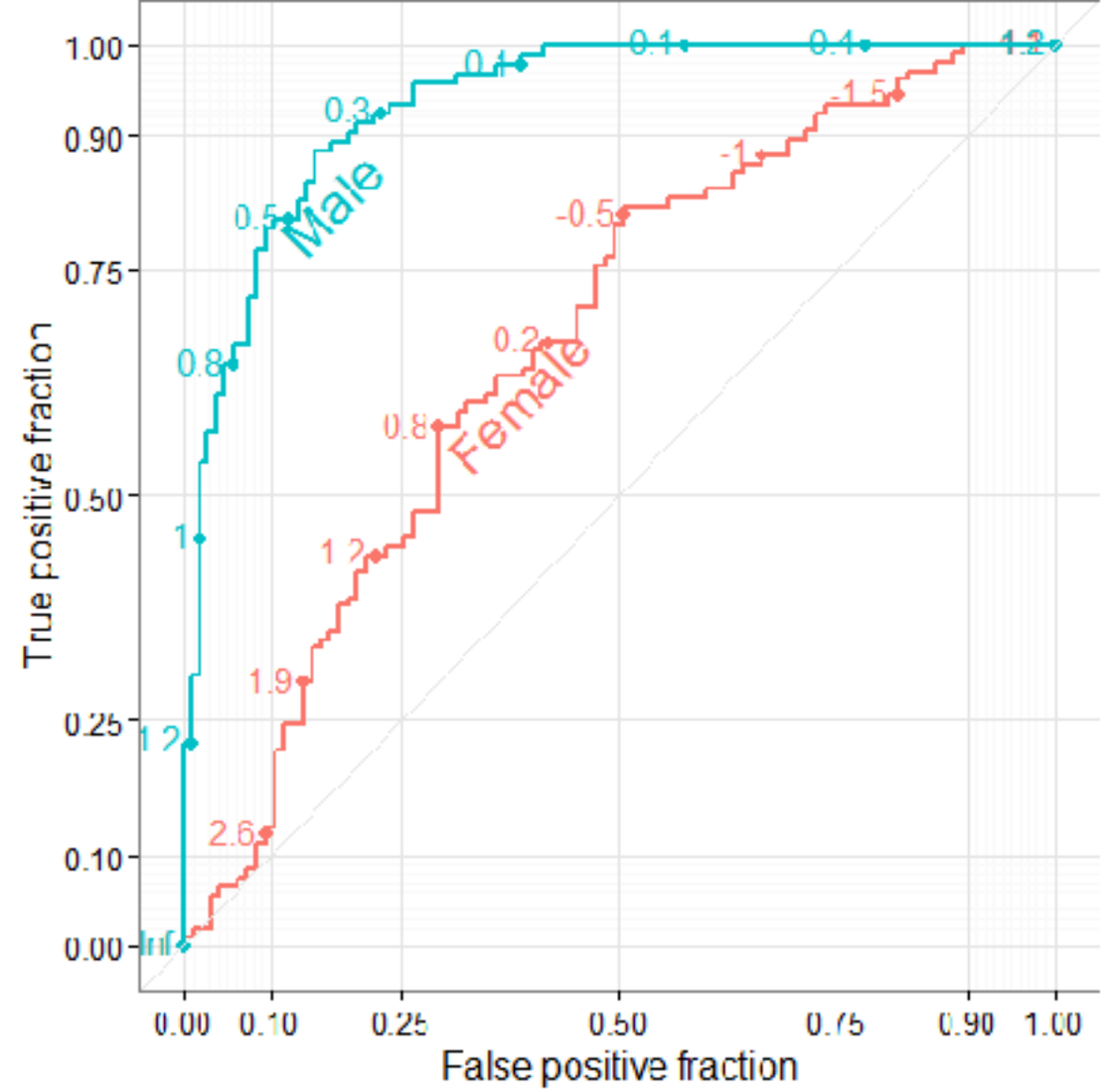
Package plotROC

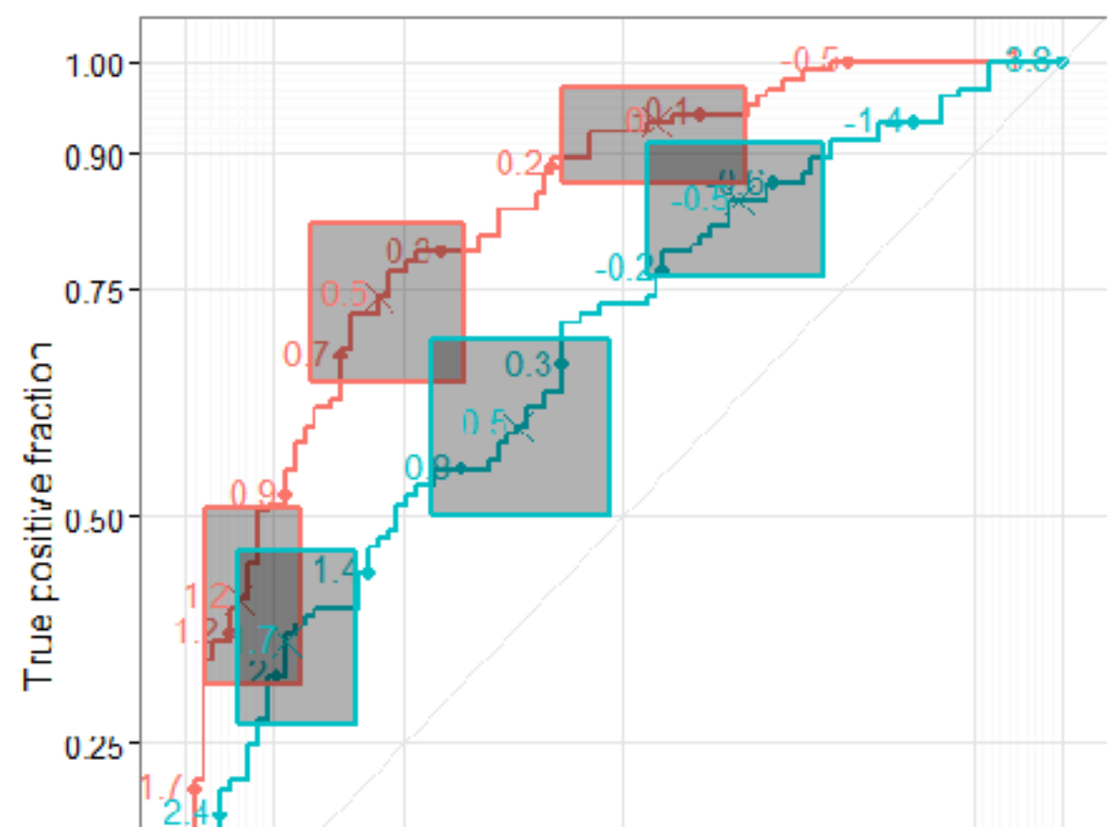


Michael Sachs (2016)
http://sachsmc.github.io/plotROC

Package plotROC



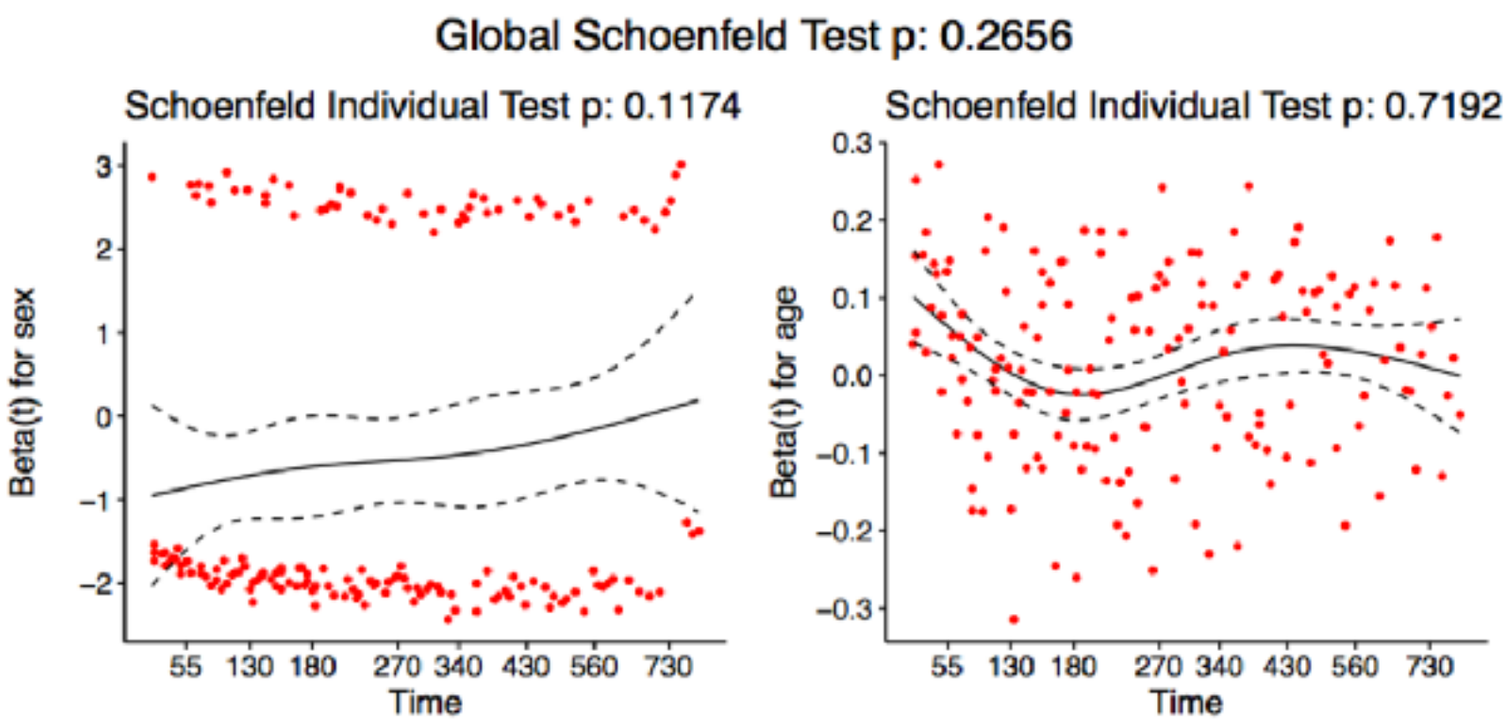




Michael Sachs (2016)
http://sachsmc.github.io/plotROC

Package survminer

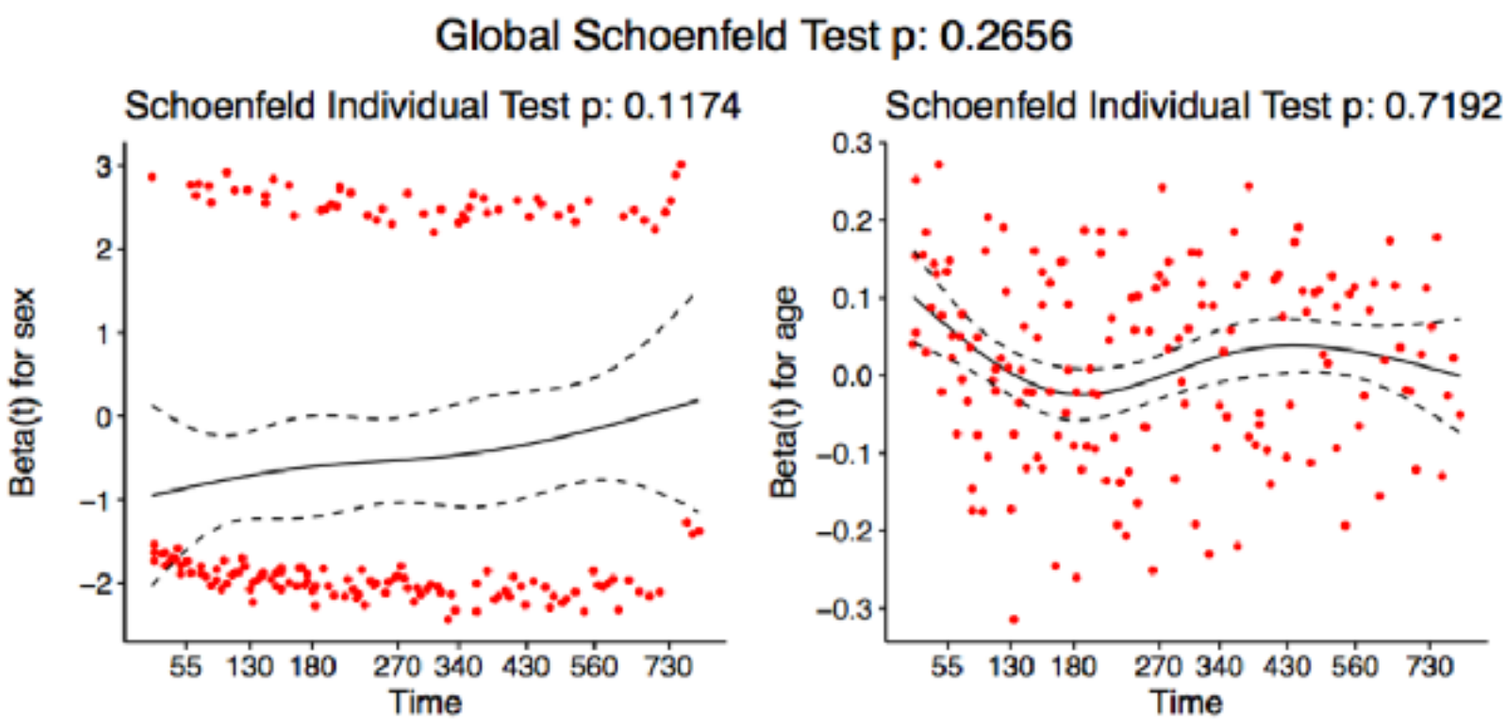
Diagnostic plots for various residuals:
"martingale", "deviance", "score", "schoenfeld",
"dfbeta", "dfbetas" and "scaledsch"

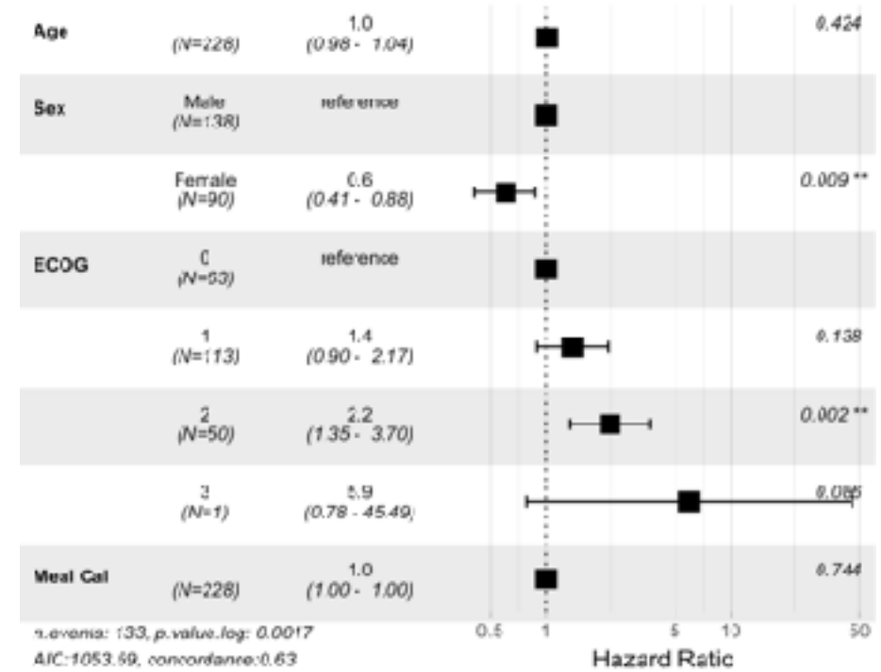


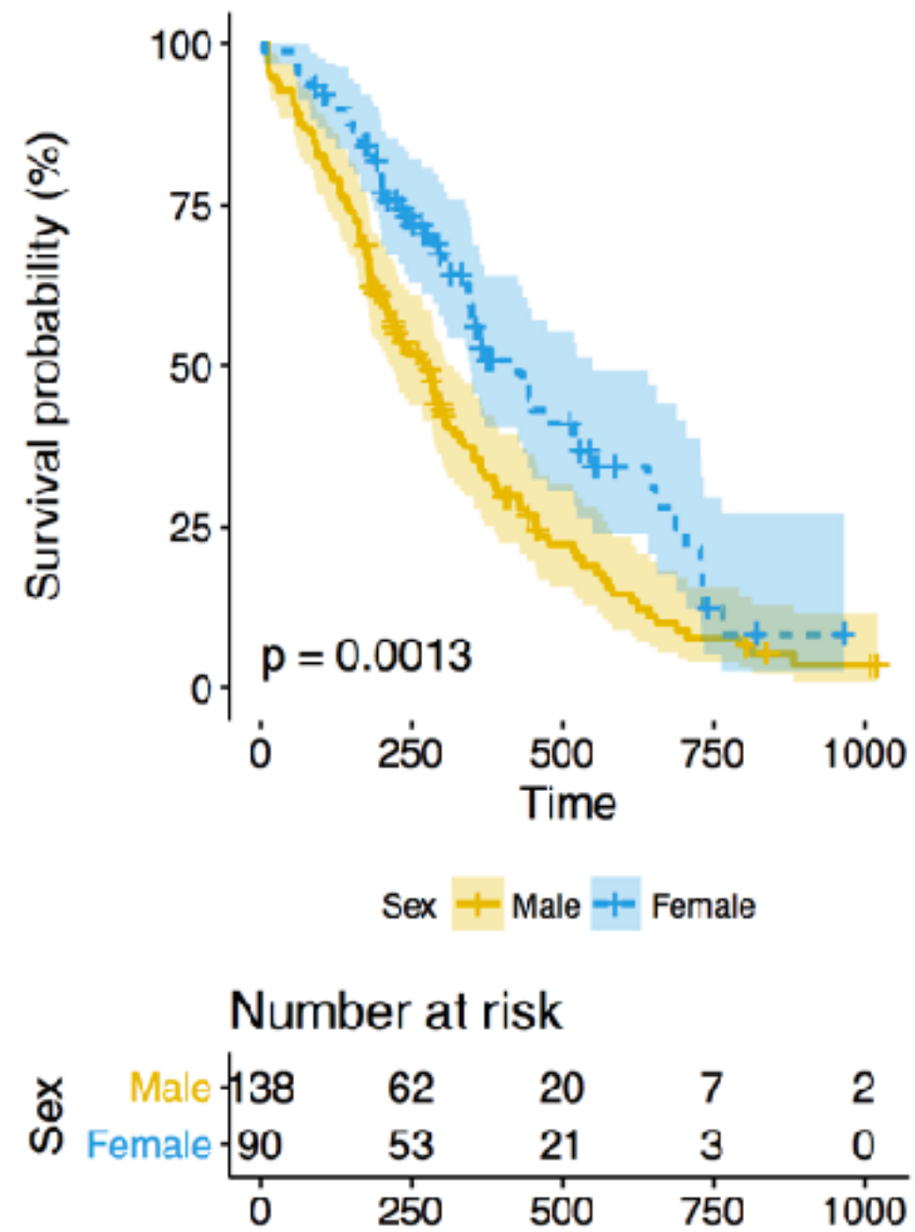
Alboukadel Kassambara, Marcin Kosiński (2016)
https://github.com/kassambara/survminer

Package survminer

Diagnostic plots for various residuals: "martingale", "deviance", "score", "schoenfeld", "dfbeta", "dfbetas" and "scaledsch"







Alboukadel Kassambara, Marcin Kosiński (2016)
https://github.com/kassambara/survminer

?? packrat

?? knitr

Reproducibility

?? docker

archivist - reproducible and recordable research

```
library("archivist")
model <- lm(Sepal.Length ~ Sepal.Width, data=iris)
saveToLocalRepo(model)

models <- asearch("pbiecek/graphGallery", patterns = "class:lm")
modelsBIC <- sapply(models, BIC)
sort(modelsBIC)

## 990861c7c27812ee959f10e5f76fe2c3 2a6e492cb6982f230e48cf46023e2e4f
##                         39.05577                         67.52735
## 0a82efeb8250a47718cea9d7f64e5ae7 378237103bb60c58600fe69bed6c7f11
##                        189.73593                        189.73593
## 7f11e03539d48d35f7e7fe7780527ba7 c1b1ef7bcddefb181f79176015bc3931
##                        189.73593                        189.73593
```

Przemysław Biecek, Marcin Kosiński (2015) https://github.com/pbiecek/archivist

archivist - reproducible and recordable research

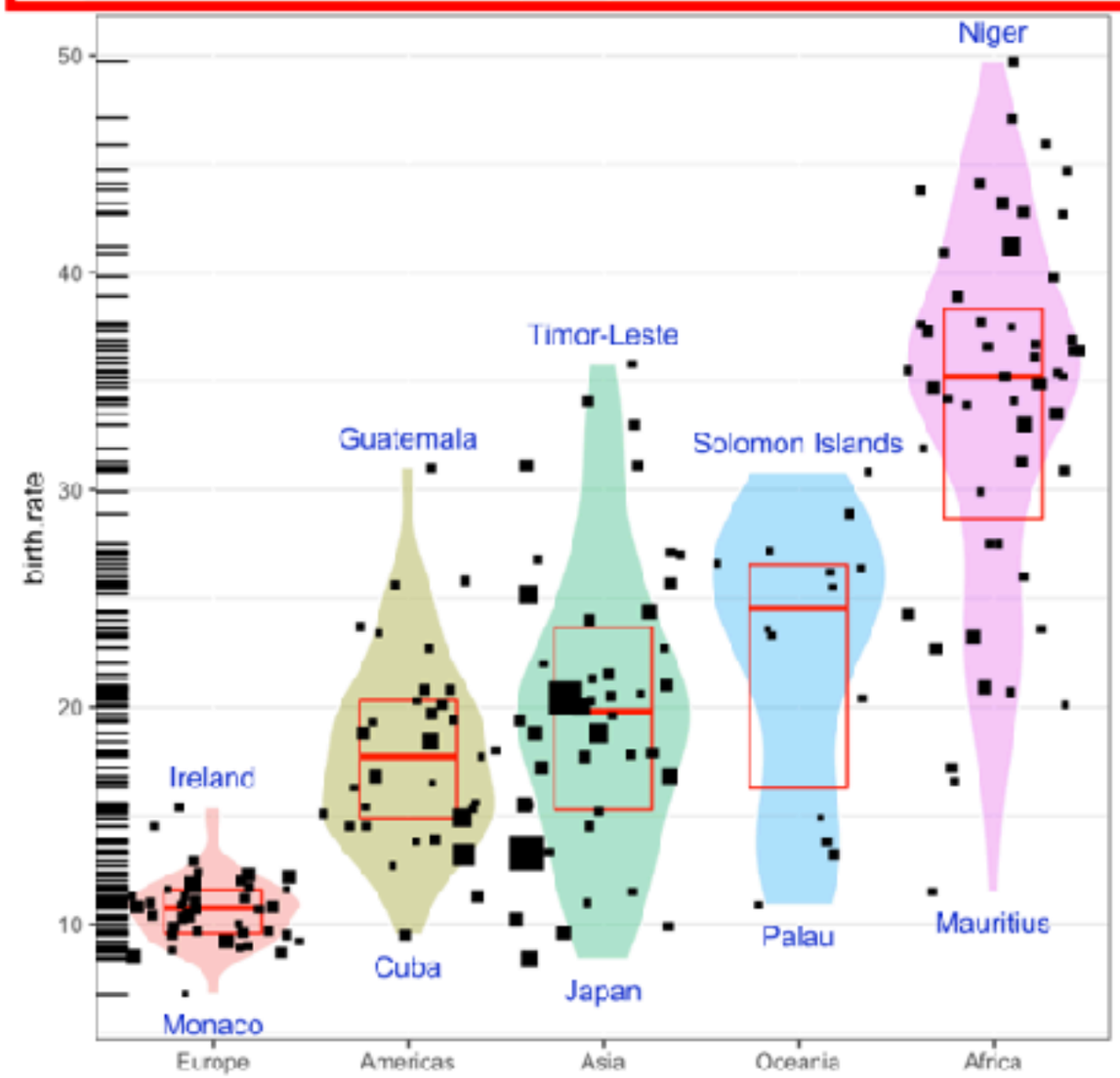
```
library("archivist")
model <- lm(Sepal.Length ~ Se
saveToLocalRepo(model)

models <- asearch("pbiecek/gr
modelsBIC <- sapply(models, B
sort(modelsBIC)

## 990861c7c27812ee959f10e5f7
##                          39
## 0a82efeb8250a47718cea9d7f6
##                         189
## 7f11e03539d48d35f7e7fe7780
##                         189
```

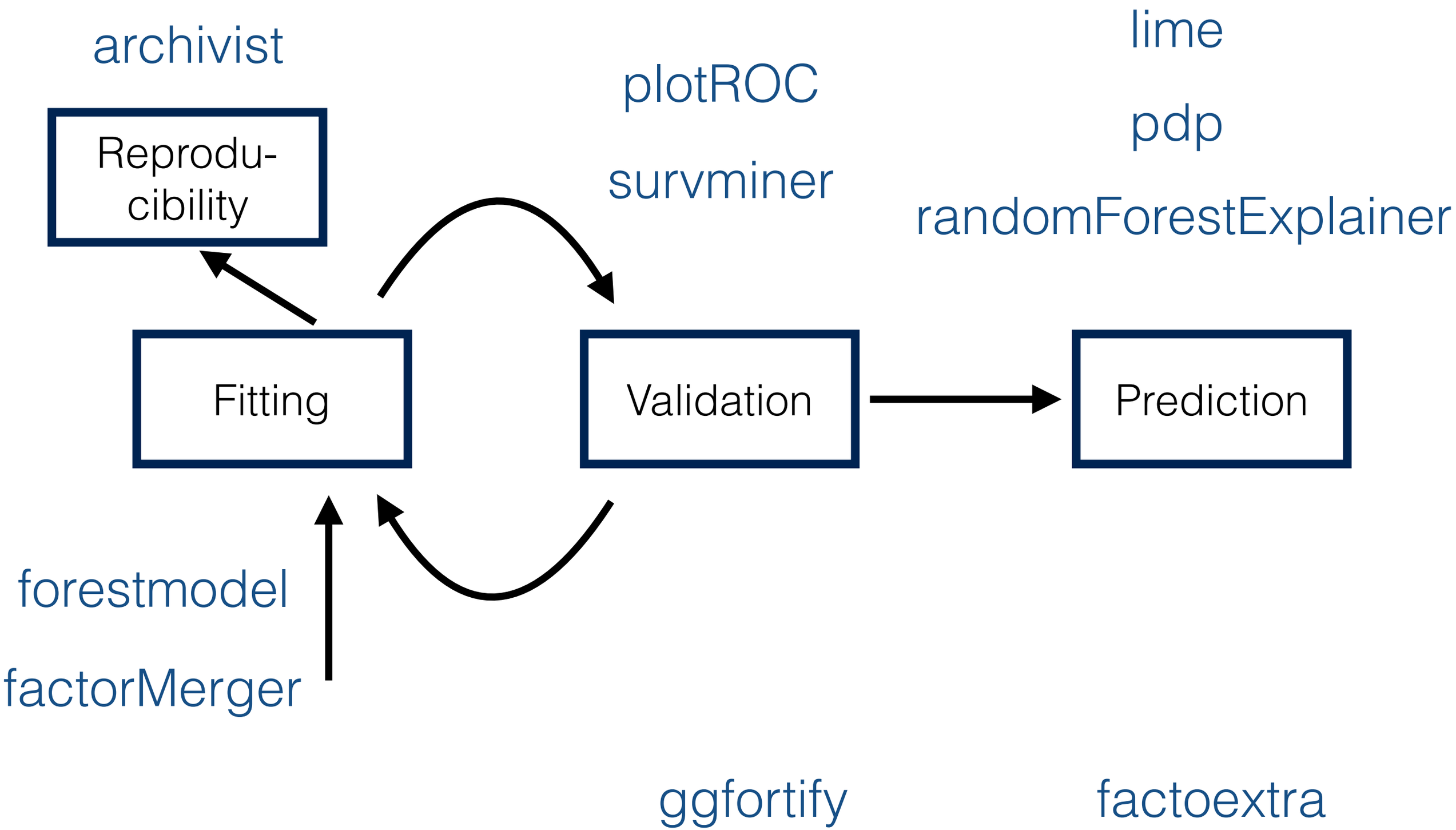
```
ggplot(countries, aes(x=continent, y=birth.rate, label=country)) +
  geom_violin(scale='width', aes(fill=continent), color="white", alpha=0.4) +
  stat_summary(fun.data = 'q3', geom = 'crossbar',
               colour = "red", width = 0.5) +
  geom_jitter(aes(size=(population)^0.9),position=position_jitter(width = .45, height =
              shape=15) +
  geom_rug(sides = "l") +
  geom_text(data=countriesMin, vjust=2, color='blue3') +
  geom_text(data=countriesMax, vjust=-1, color="blue3") +
  theme_bw() + xlab("") + theme(legend.position="none", panel.grid.major.x = element_li
te"))
```

```
pad: archivist::aread('pbiecek/Rseje/arepo/ba7f58fafe7373420e3ddce039558140')
```



Przemysław Biecek, Marcin Kosiński (2015) https://github.com/pbiecek/archivist

Life-cycle of a typical prognostic model



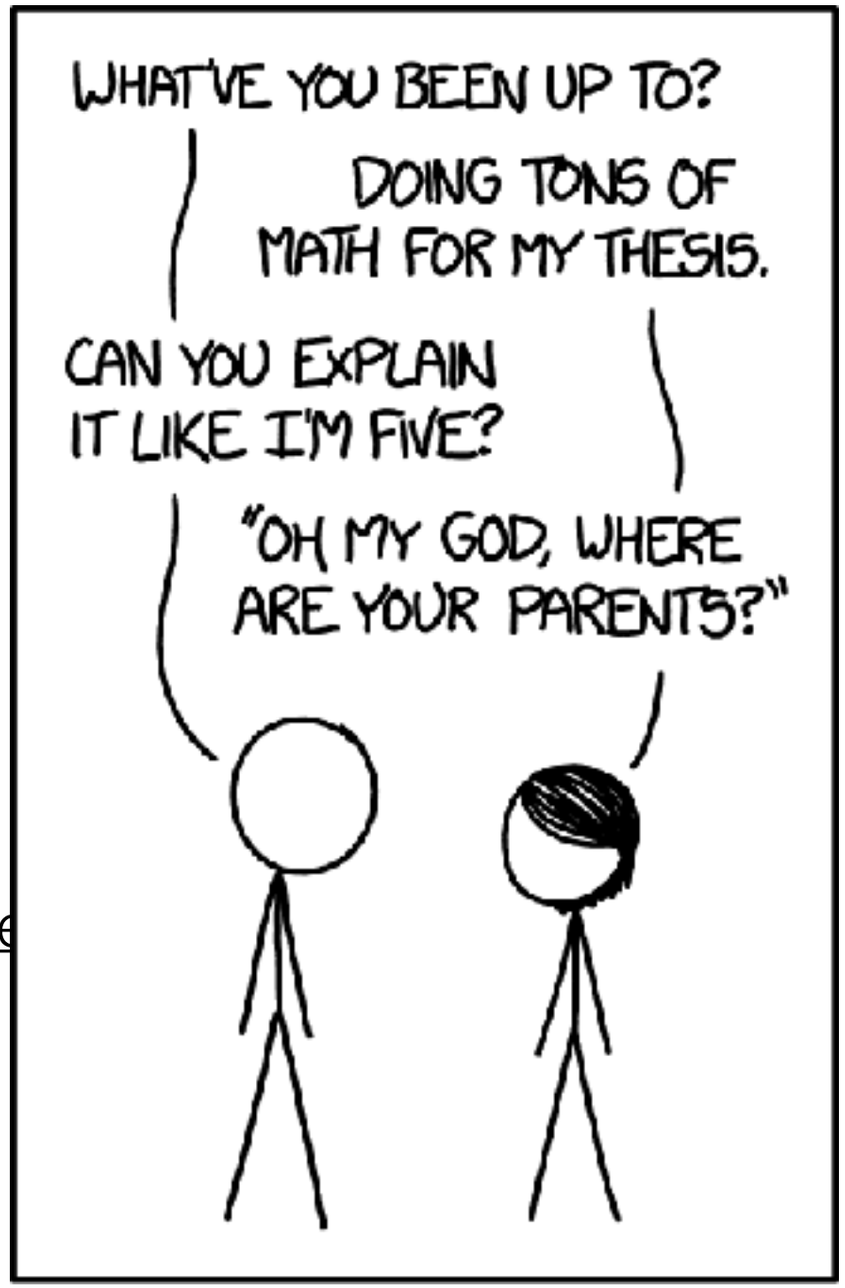
ELI5 is a Python library which allows to visualize and debug various Machine Learning models
http://eli5.readthedocs.io/en/latest/index.html

Visualizing statistical models: Removing the blindfold.
Hadley Wickham, Dianne Cook, Heike Hofmann (2015)
Statistical Analysis and Data Mining
http://had.co.nz/stat645/model-vis.pdf

Ideas on interpreting machine learning.
Patrick Hall, Wen Phan, SriSatish Ambati (2017)
https://www.oreilly.com/ideas/ideas-on-interpreting-machine-le

rms: Regression Modeling Strategies.
Frank E Harrell (2009-2017) CRAN
https://cran.r-project.org/web/packages/rms/index.html

ggRandomForests: Random Forests for Regression
John Ehrlinger (2015)
https://arxiv.org/pdf/1501.07196.pdf



https://imgs.xkcd.com/comics/like_im_five.png

Thank you for your attention!