

Show me your predictive model

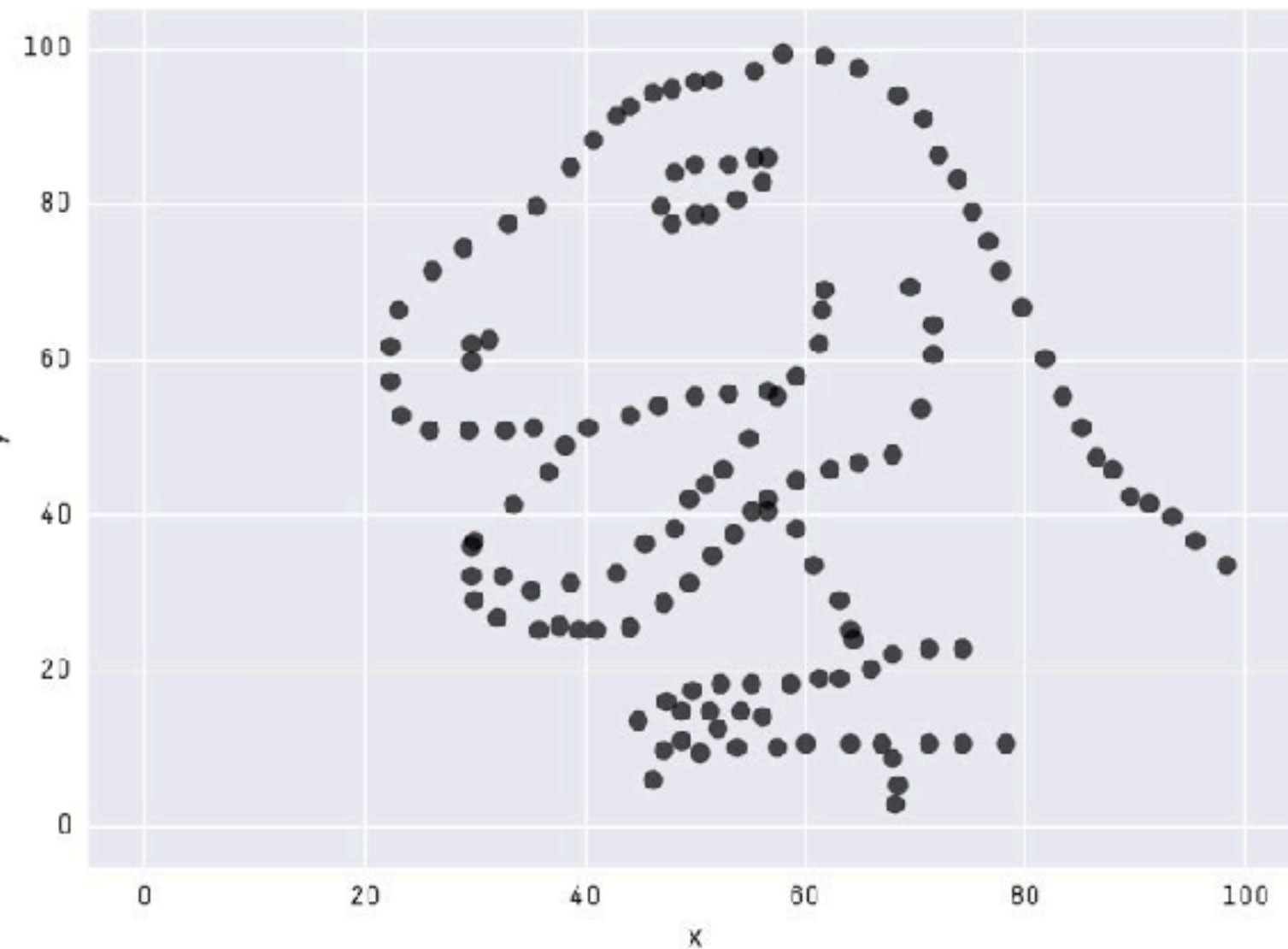
Why and how?

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?

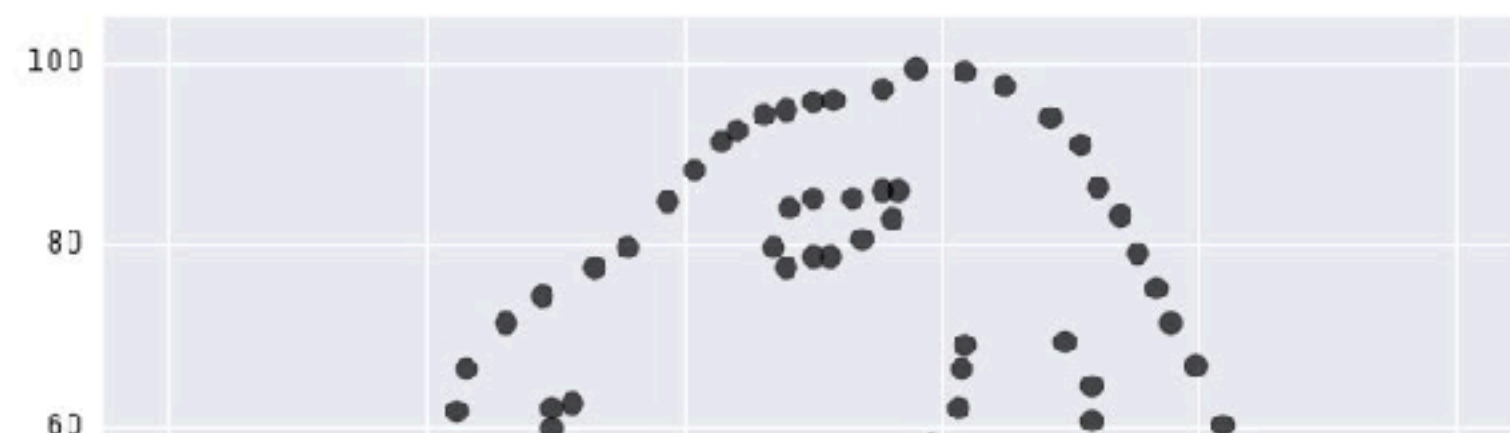


```
X Mean: 54.2632025
Y Mean: 47.8315781
X SD   : 16.7650109
Y SD   : 26.9353144
Corr.  : -0.0645195
```

Package datasauRus

<https://www.autodeskresearch.com/publications/samestats>

Do we really need plots to understand numbers?



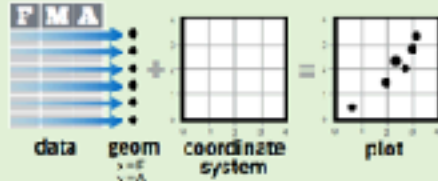
X Mean: 54.2632025
Y Mean: 47.8315781

Data Visualization with ggplot2 Cheat Sheet

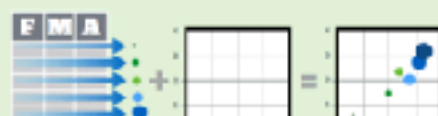


Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same few components: a **data** set, a set of **geoms**—visual marks that represent data points, and a **coordinate system**.



To display data values, map variables in the data set to aesthetic properties of the geom like **size**, **color**, and **x** and **y** locations.



Geoms - Use a geom to represent data points, use the geom's aesthetic properties to represent variables. Each function re

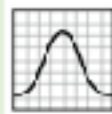
One Variable

Continuous

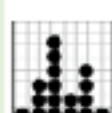
`a <- ggplot(mpg, aes(hwy))`



`a + geom_area(stat = "bin")`
x, y, alpha, color, fill, linetype, size
`b + geom_area(aes(y = ..density..), stat = "bin")`



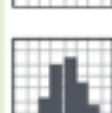
`a + geom_density(kernel = "gaussian")`
x, y, alpha, color, fill, linetype, size, weight
`b + geom_density(aes(y = ..county..))`



`a + geom_dotplot()`
x, y, alpha, color, fill



`a + geom_freqpoly()`
x, y, alpha, color, linetype, size
`b + geom_freqpoly(aes(y = ..density..))`



`a + geom_histogram(binwidth = 5)`
x, y, alpha, color, fill, linetype, size, weight
`b + geom_histogram(aes(y = ..density..))`

Discrete

`b <- ggplot(mpg, aes(fl))`

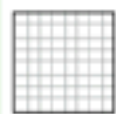


`b + geom_bar()`
x, alpha, color, fill, linetype, size, weight

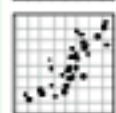
Two Variables

Continuous X, Continuous Y

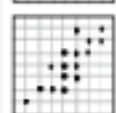
`f <- ggplot(mpg, aes(cty, hwy))`



`f + geom_blank()`



`f + geom_jitter()`
x, y, alpha, color, fill, shape, size



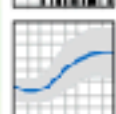
`f + geom_point()`
x, y, alpha, color, fill, shape, size



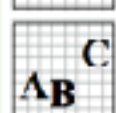
`f + geom_quantile()`
x, y, alpha, color, linetype, size, weight



`f + geom_rug(sides = "bl")`
alpha, color, linetype, size



`f + geom_smooth(model = lm)`
x, y, alpha, color, fill, linetype, size, weight



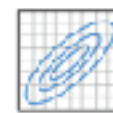
`f + geom_text(aes(label = cty))`
x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

Continuous Bivariate

`i <- ggplot(movies, aes(yr, rtg))`



`i + geom_bin2d(binwidth)`
xmax, xmin, ymax, ymin, linetype, size, weight



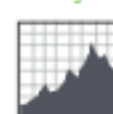
`i + geom_density2d()`
x, y, alpha, colour, linetype



`i + geom_hex()`
x, y, alpha, colour, fill size

Continuous Function

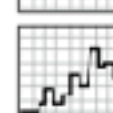
`j <- ggplot(economics, aes(dur, unemp))`



`j + geom_area()`
x, y, alpha, color, fill, line



`j + geom_line()`
x, y, alpha, color, linetype



`j + geom_step(direction)`
x, y, alpha, color, linetype

Visualizing error

`df <- data.frame(grp = c("A", "B"))`

Do we need plots to understand models?

Is it all about accuracy?

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

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

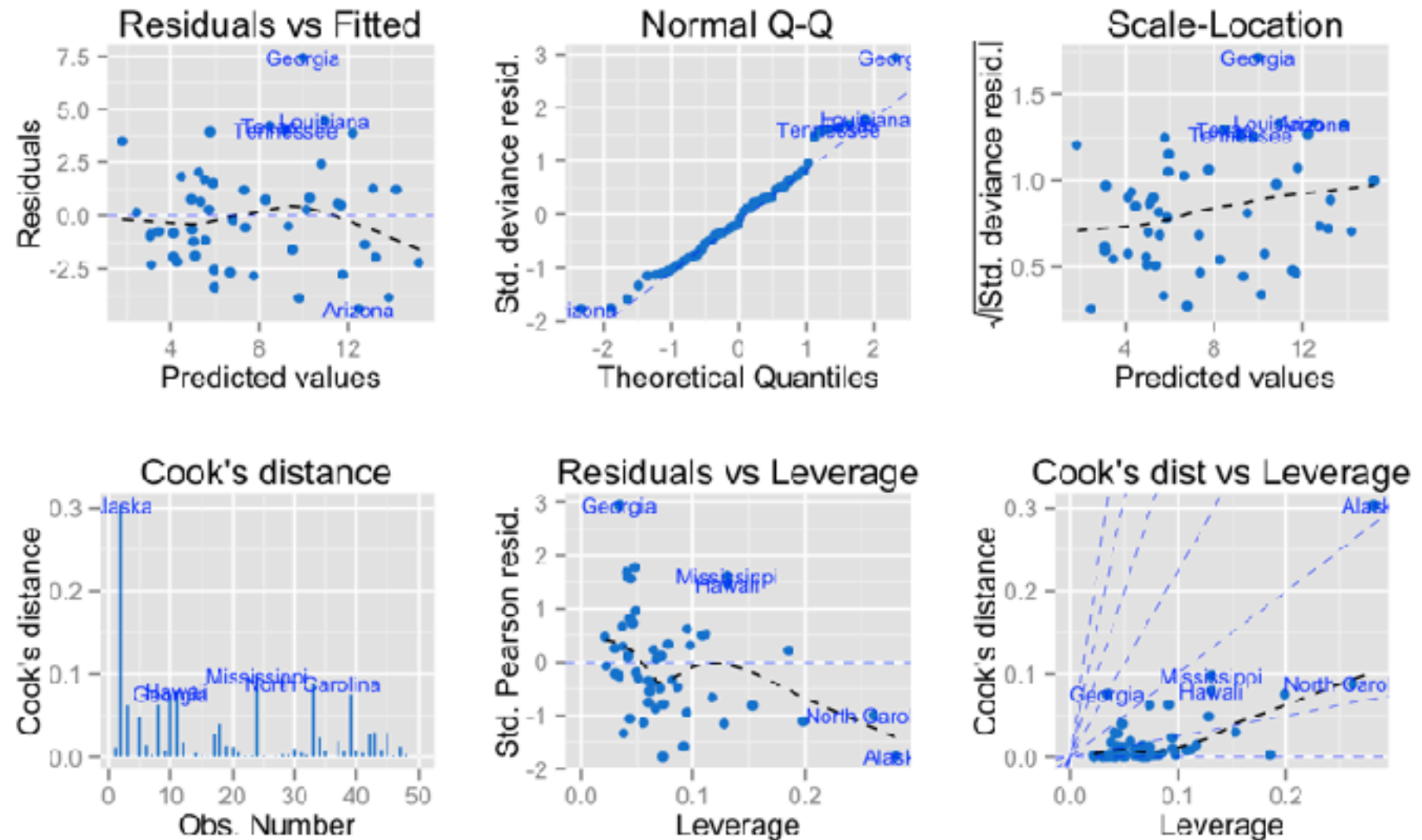
2) Understanding and trust - we need to **understand models that makes important decisions.**

"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.**

Package ggfortify

Im



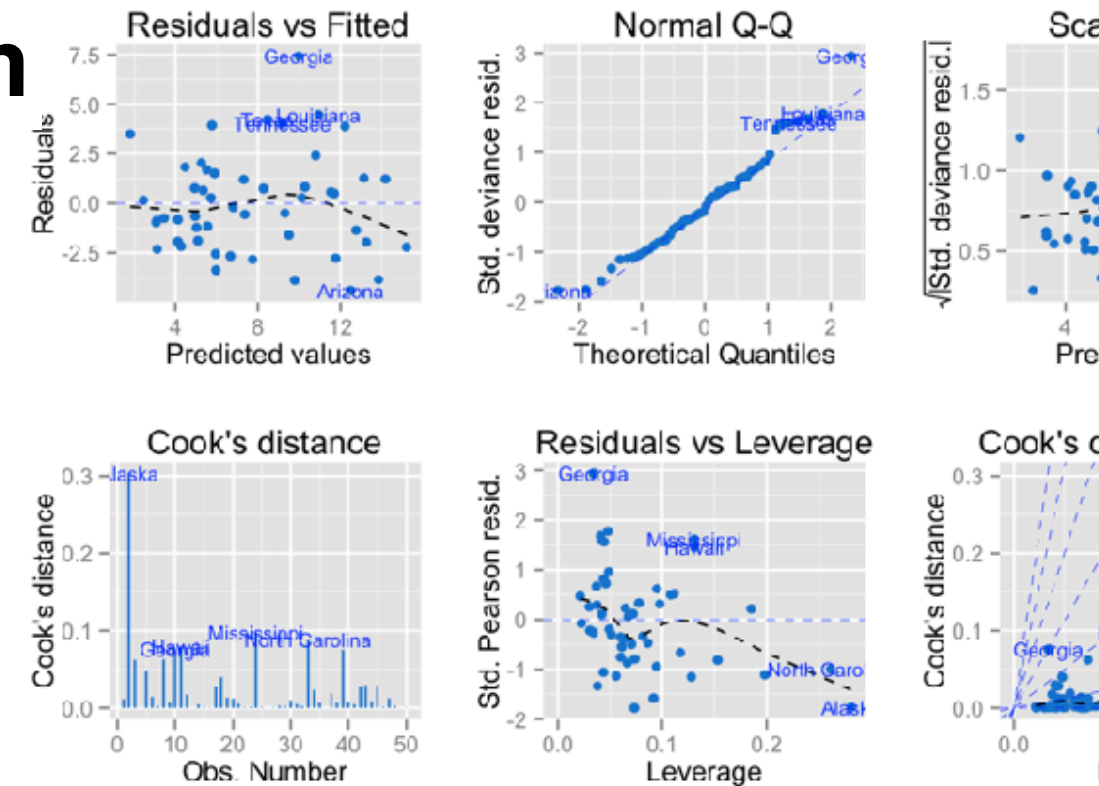
Yuan Tang, Masaaki Horikoshi, and Wenxuan Li.
"ggfortify: Unified Interface to Visualize Statistical
Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.

Package ggfortify

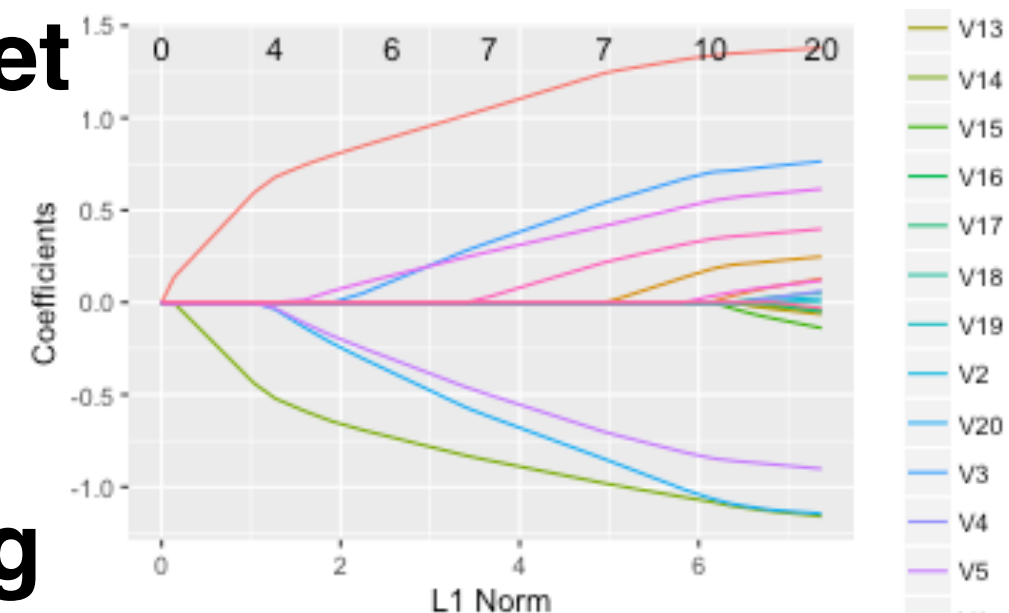
Table 1: Supported packages

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

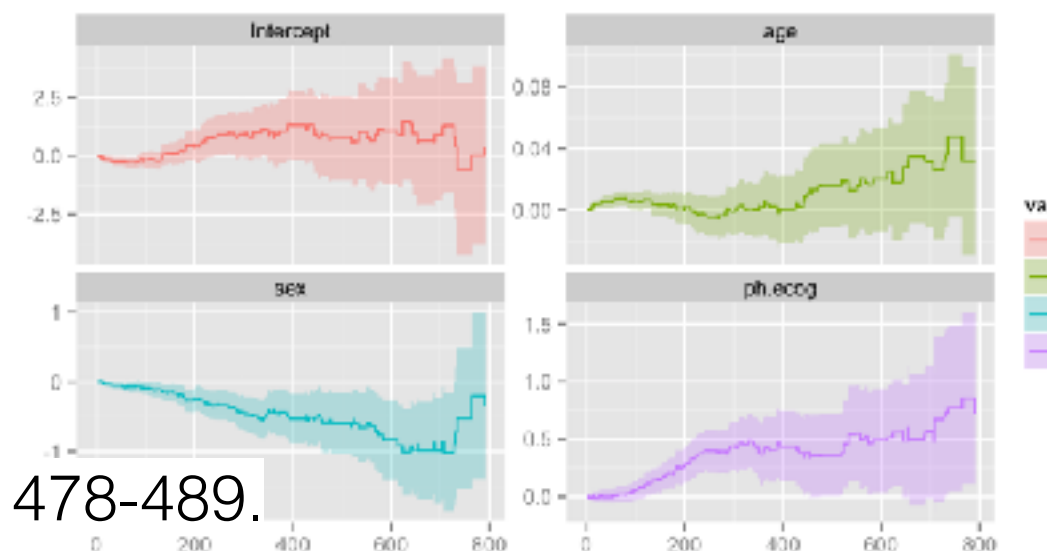
lm



glmnet

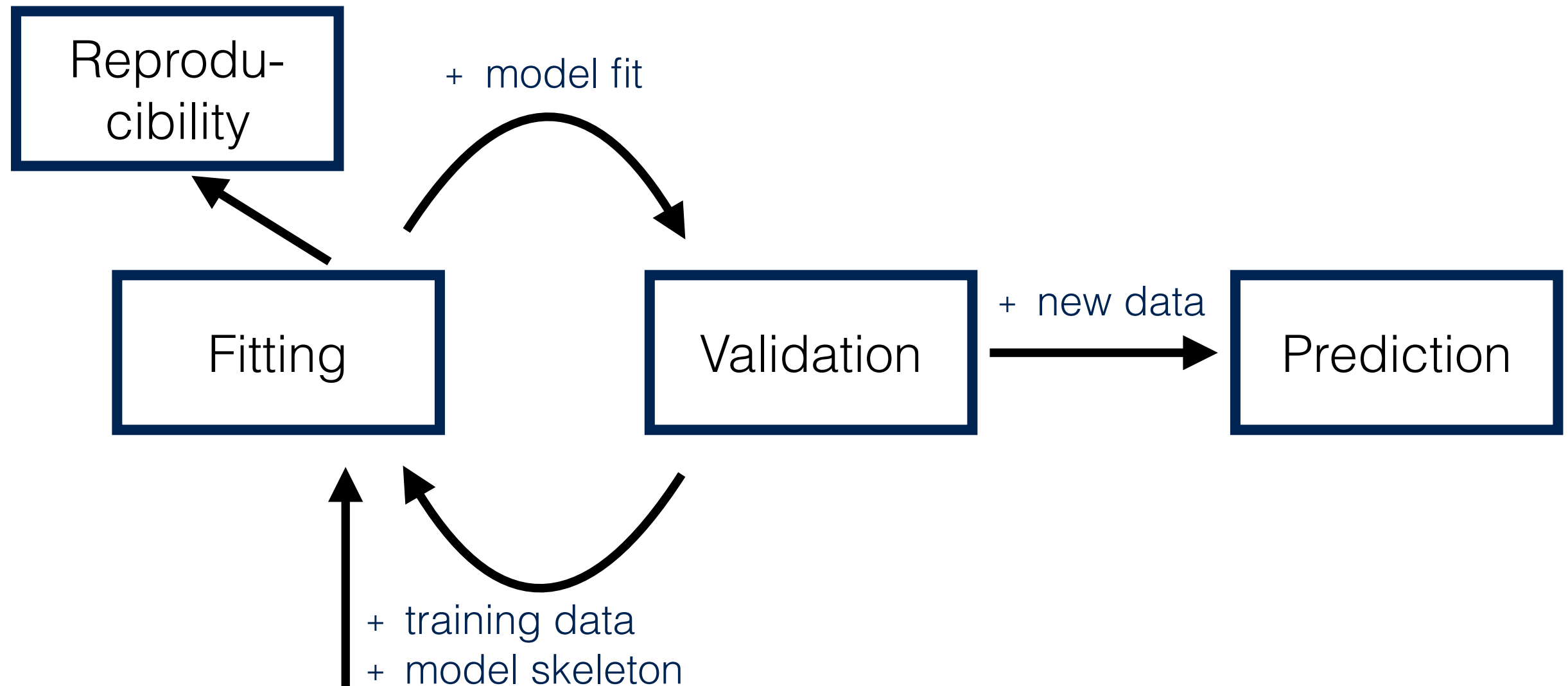


aareg

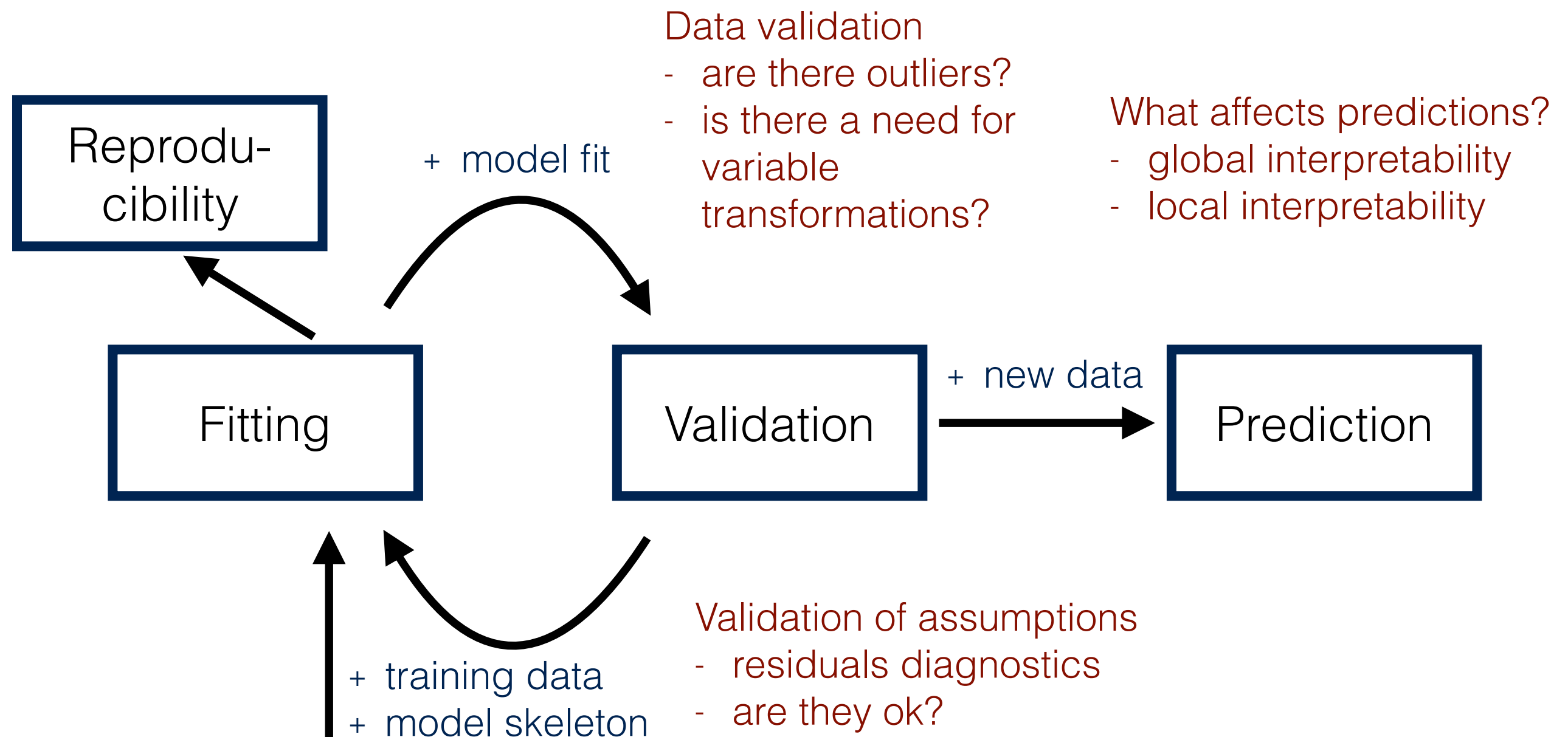


Yuan Tang, Masaaki Horikoshi, and Wenxuan Li.
 "ggfortify: Unified Interface to Visualize Statistical
 Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.

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)

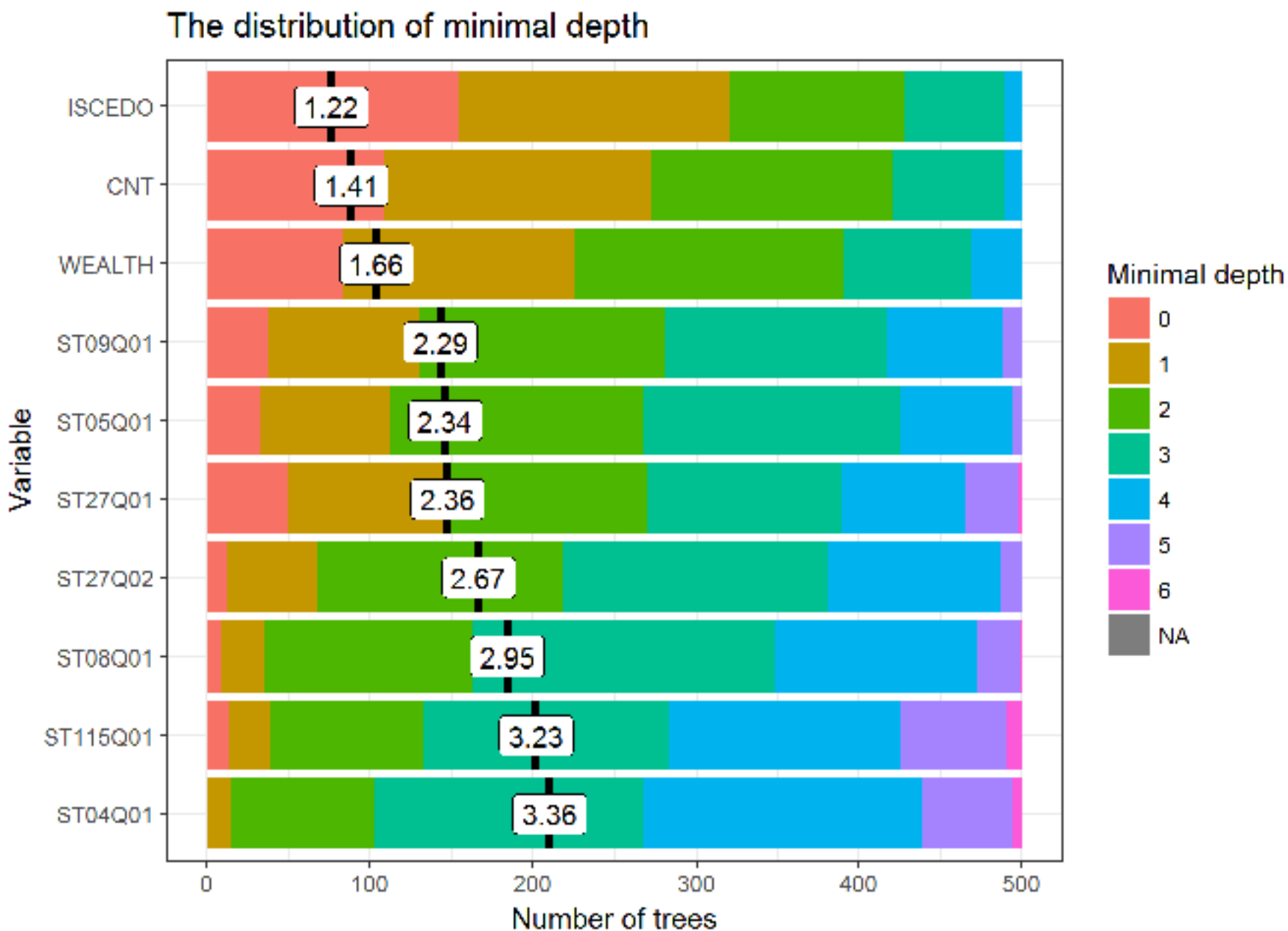
Prediction

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>

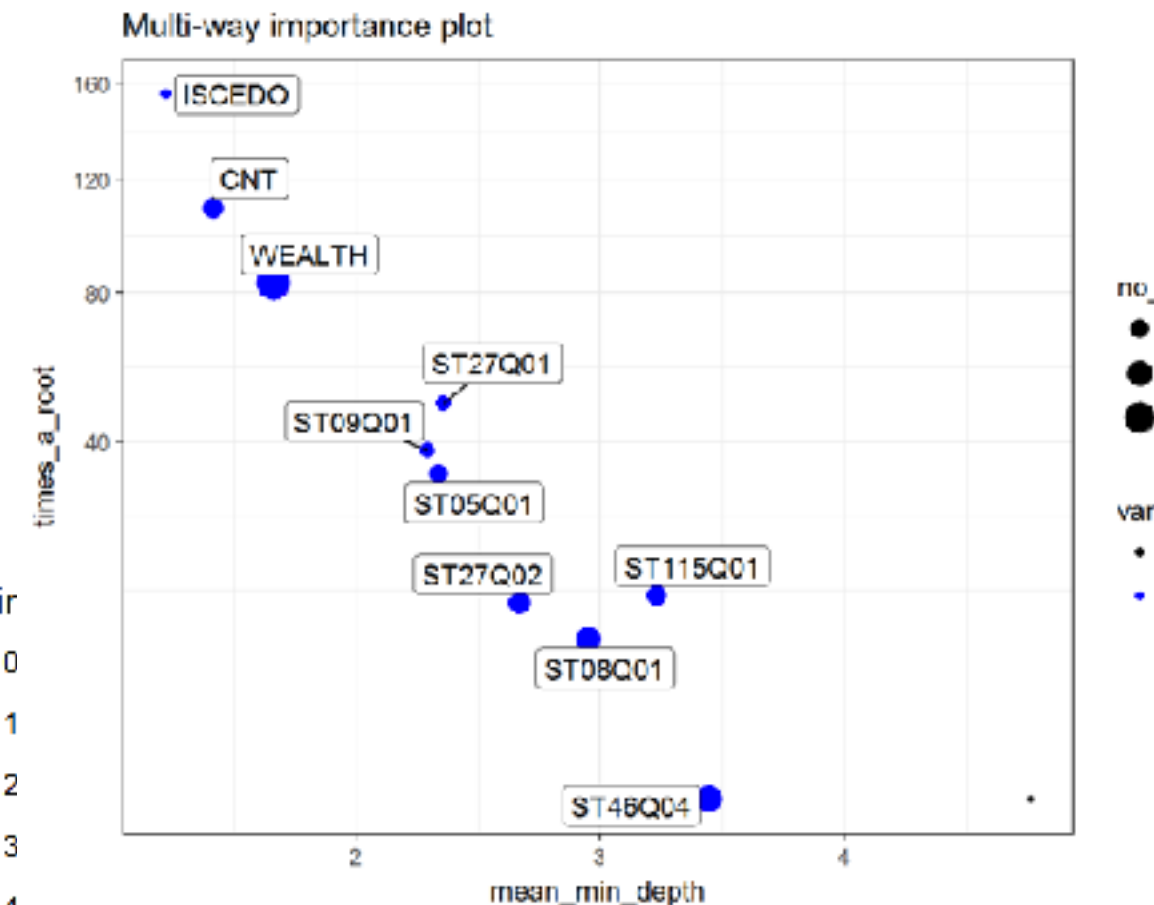
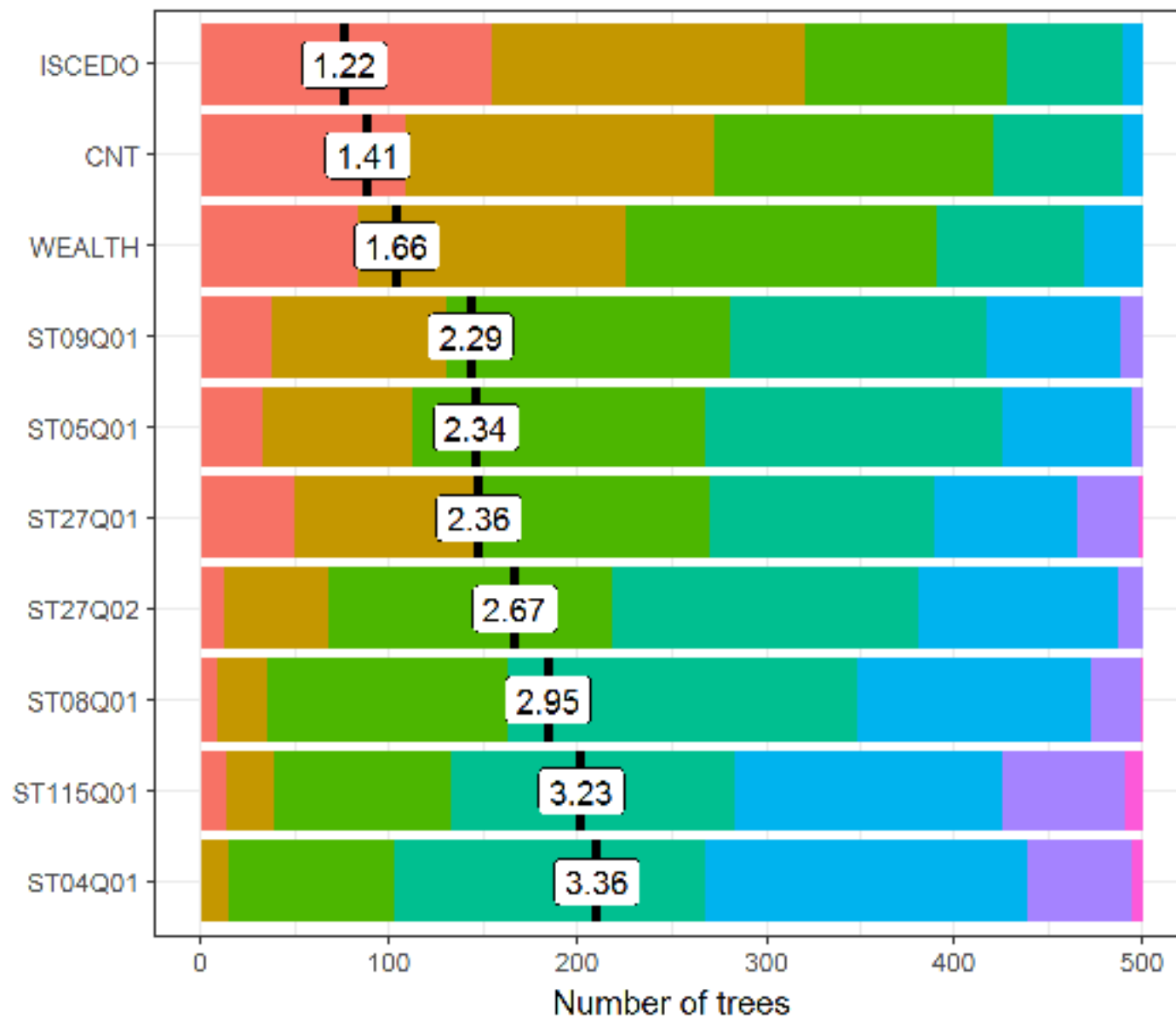
John Ehrlinger (2015)

ggRandomForests: Random Forests for Regression

Package randomForestExplainer

What is in a random forest?

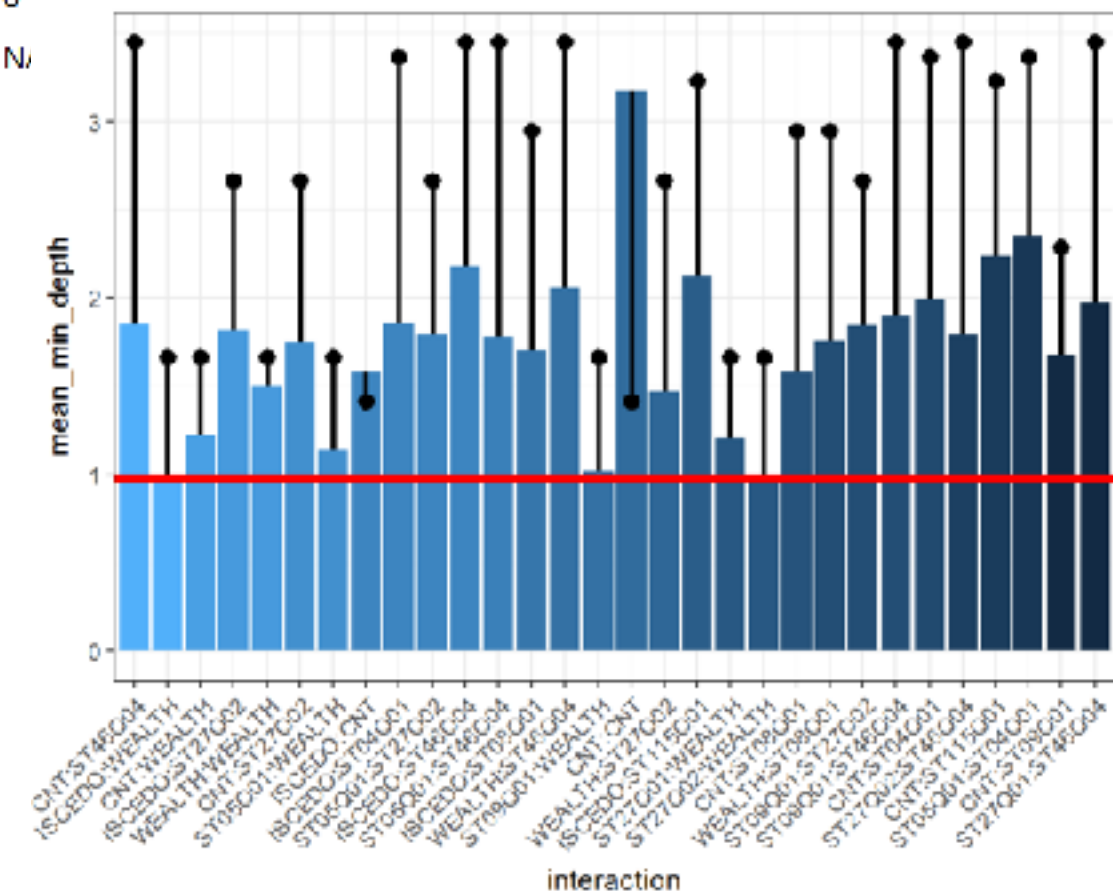
The distribution of minimal depth



Minir

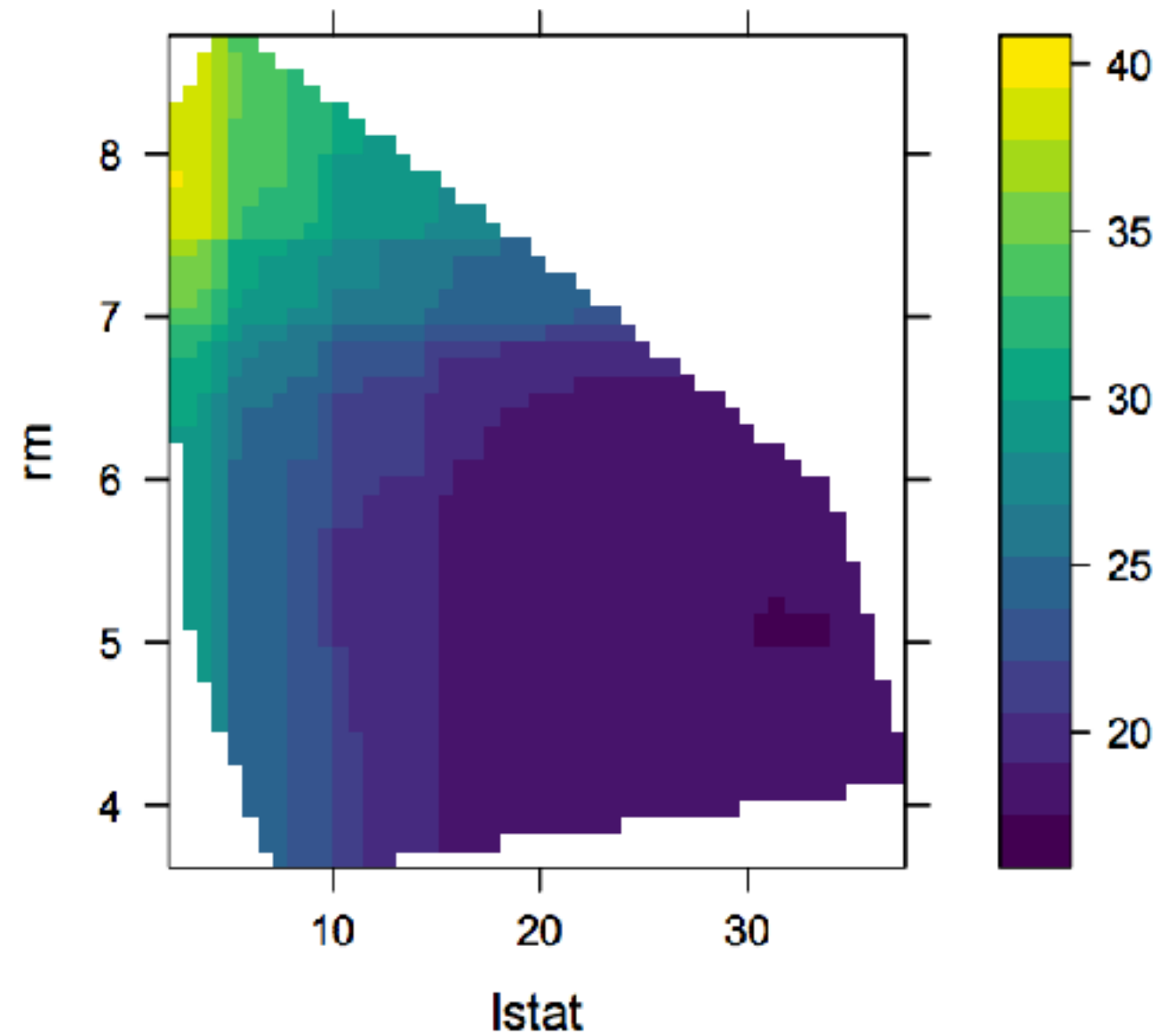
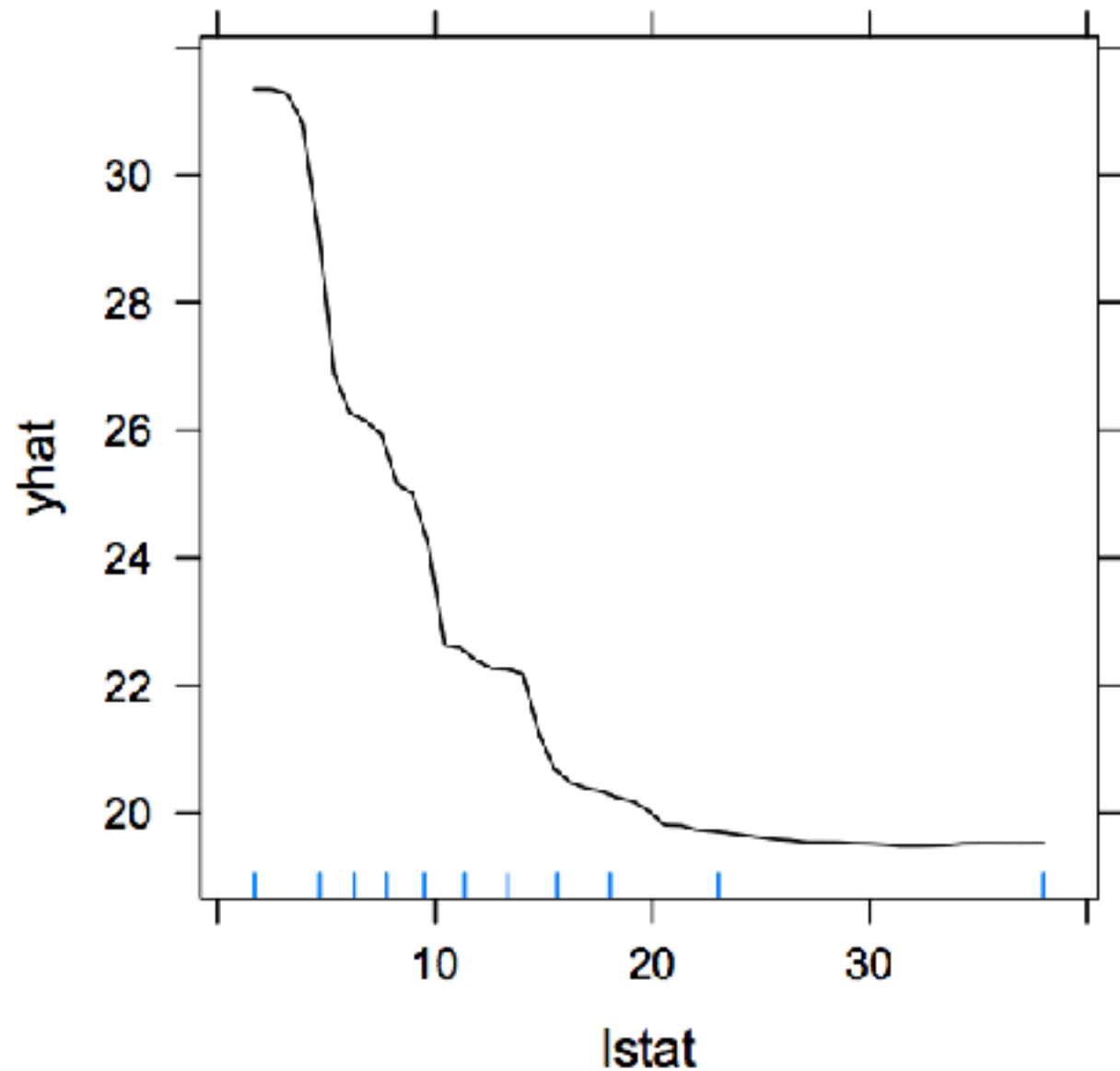


Mean minimal depth for 30 most frequent interactions



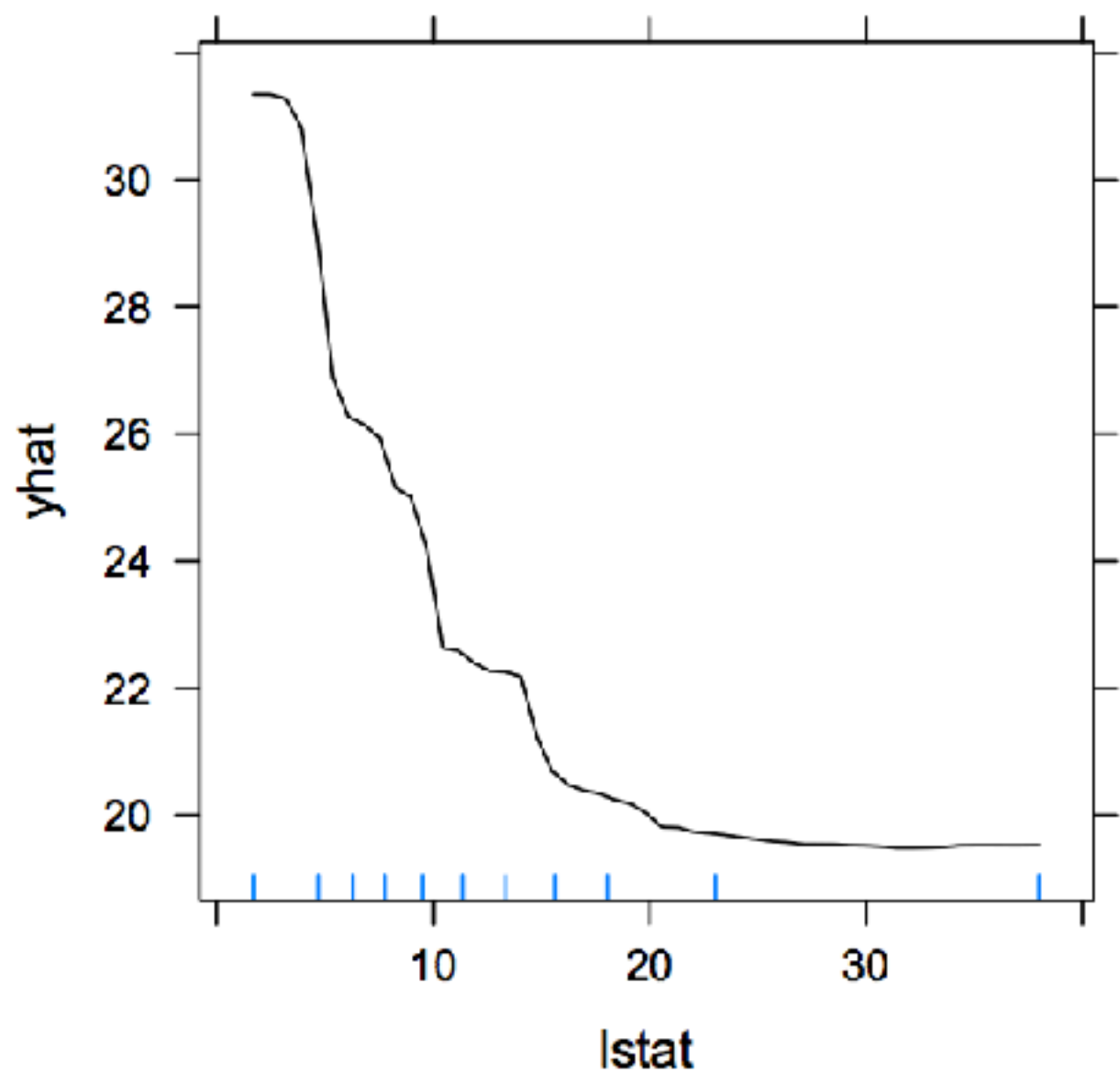
Aleksandra Paluszynska, Przemyslaw Biecek (2017)
<https://github.com/geneticsMiNIng/BlackBoxOpener>
John Ehrlinger (2015)
ggRandomForests: Random Forests for Regression

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

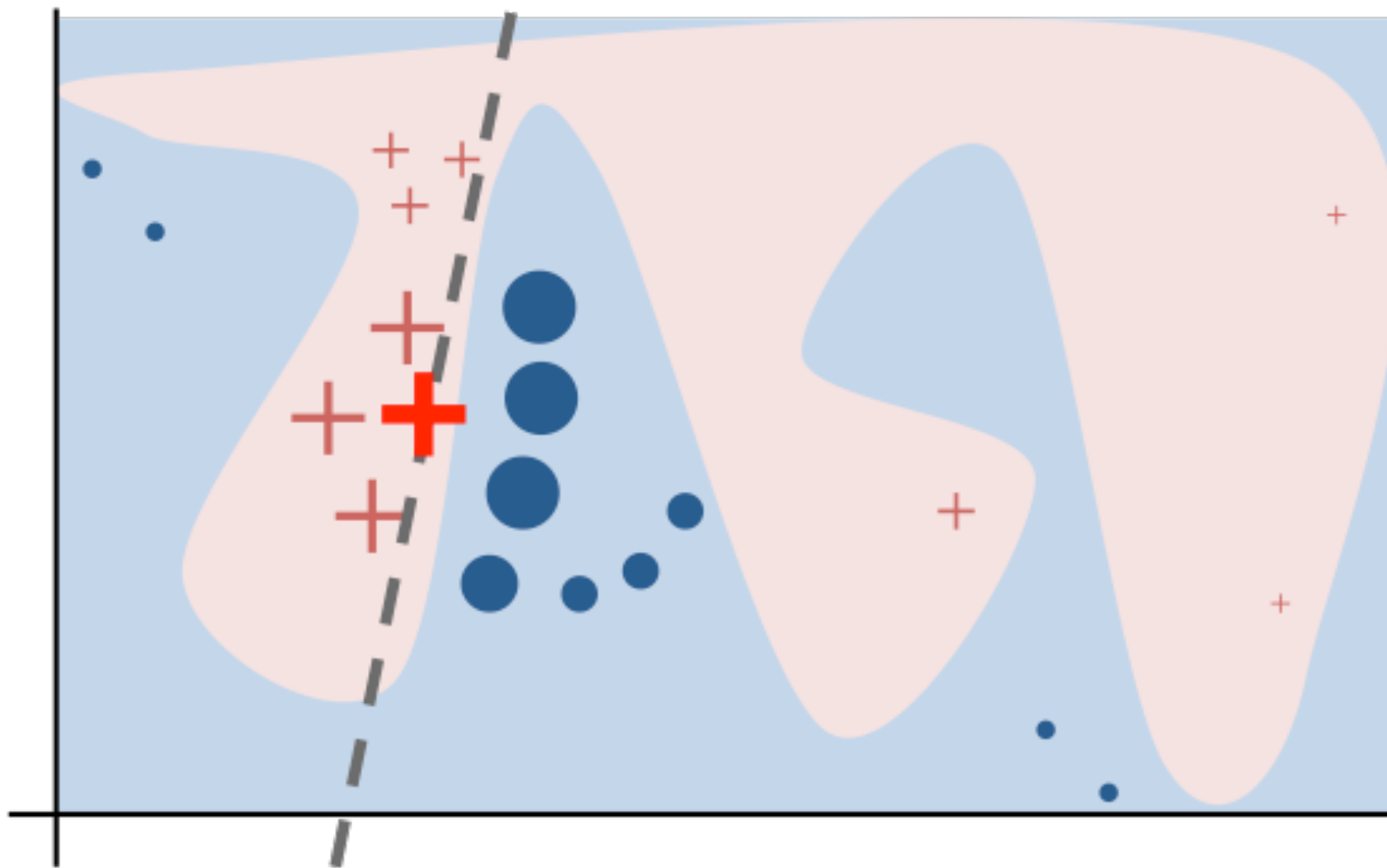


Type of model	R package	Object class
Decision tree	C50 (Kuhn et al., 2015) party partykit	"C5.0" "BinaryTree" "party"
Bagged decision trees	rpart (Therneau et al., 2015) adabag (Alfaro et al., 2013) ipred (Peters and Hothorn, 2015)	"rpart" "bagging" "classbagg" "regbagg"
Boosted decision trees	adabag (Alfaro et al., 2013) gbm xgboost	"boosting" "gbm" "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) mda (Leisch et al., 2016)	"earth" "mars"
Projection pursuit regression	stats	"ppr"
Random forest	randomForest party partykit ranger (Wright, 2016)	"randomFor" "RandomFor" "cforest" "ranger"
Support vector machine	e1071 (Meyer et al., 2015) kernlab (Karatzoglou et al., 2004)	"svm" "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`.

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.



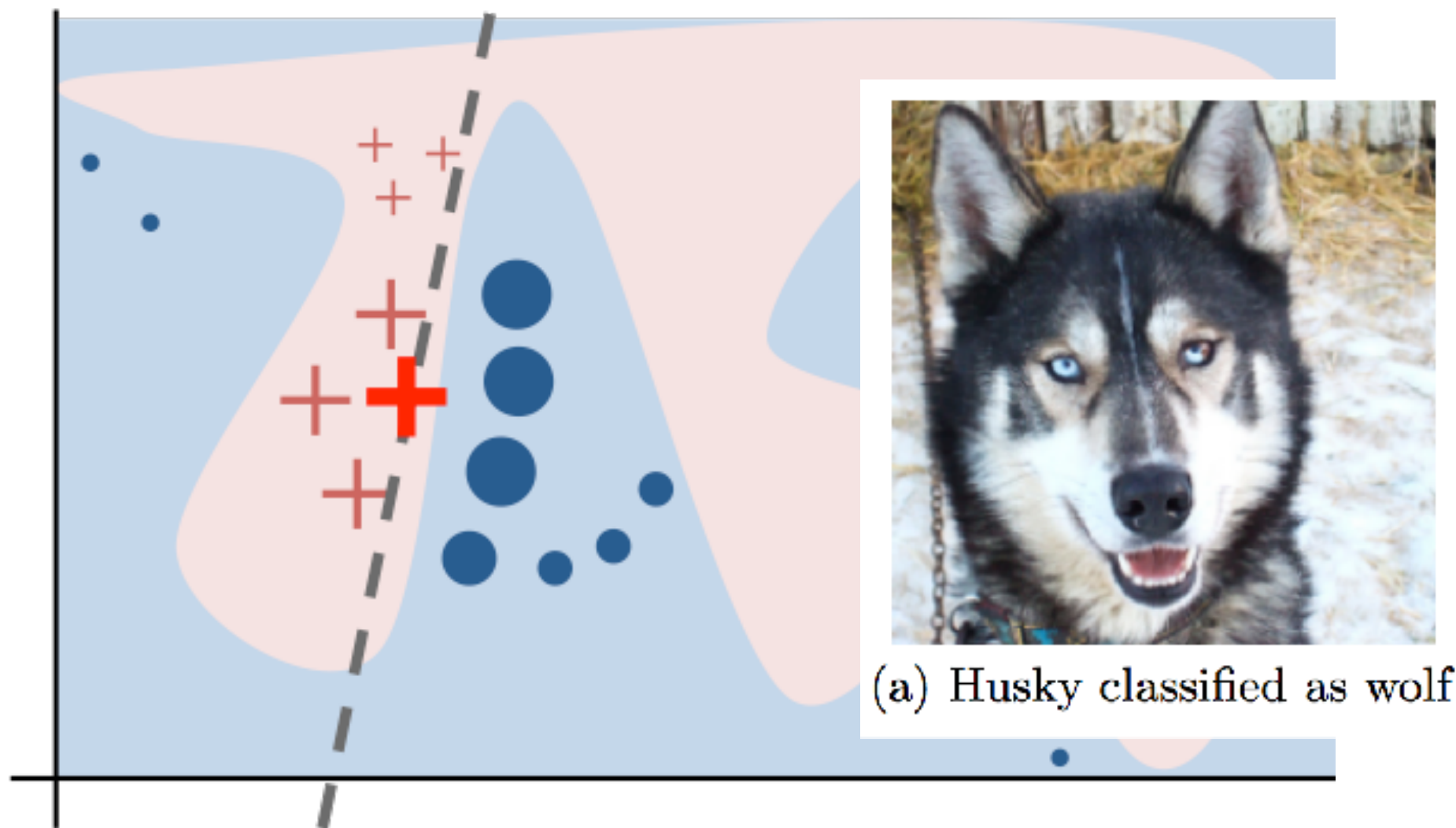
"Why Should I Trust You?" Explaining the Predictions of Any Classifier.

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.



"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>

What are estimates
of model parameters?









Are convergence
criteria satisfied?

Fitting

Are the assumptions satisfied?

Do we need
variable transformation?

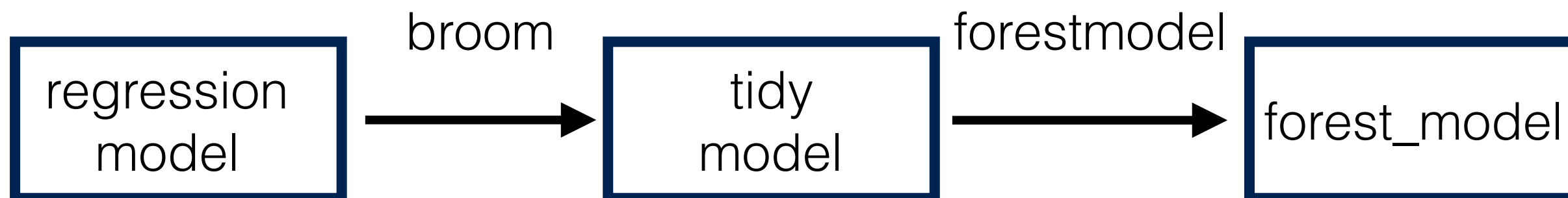
Package forestmodel

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, 45.70) 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

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, 45.70) 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

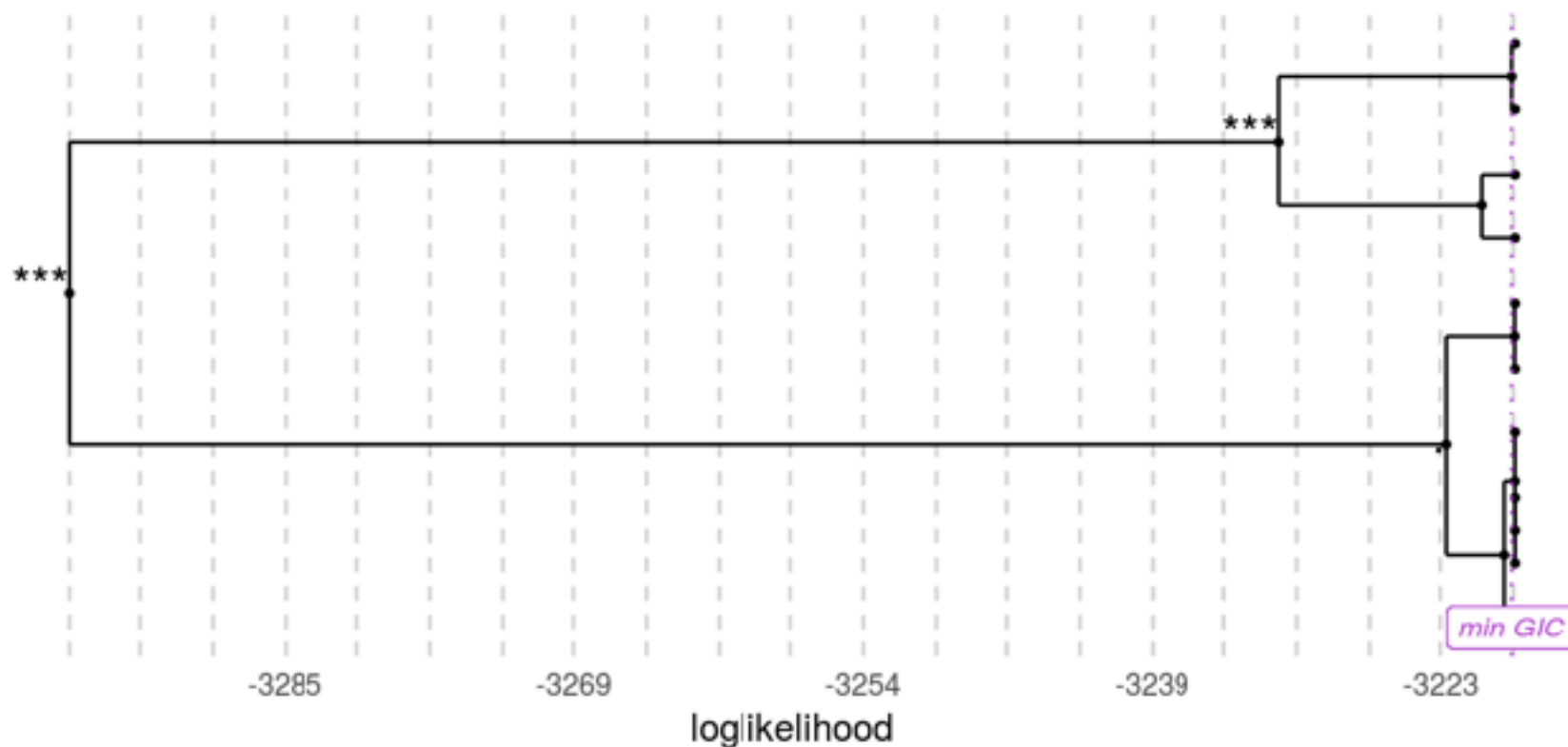
Nick Kennedy (2017) <https://github.com/NikNakk/forestmodel>

Package factorMerger

Visualisation for post-hoc testing

Merging path plot

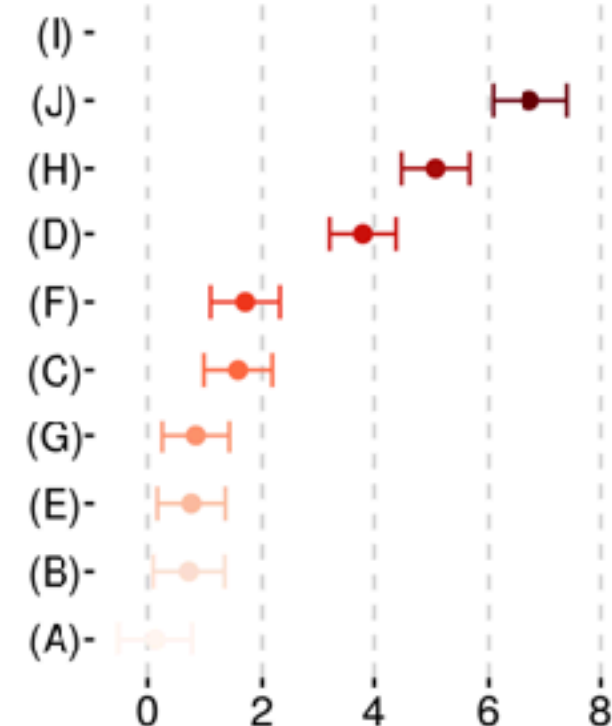
Optimal GIC partition: (A)(B)(E)(G):(C)(F):(D):(H):(J)(I)



(I): 7.25
(J): 6.75
(H): 5.07
(D): 3.8
(F): 1.7
(C): 1.58
(G): 0.84
(E): 0.76
(B): 0.72
(A): 0.13

Summary statistics

Means and standard deviation



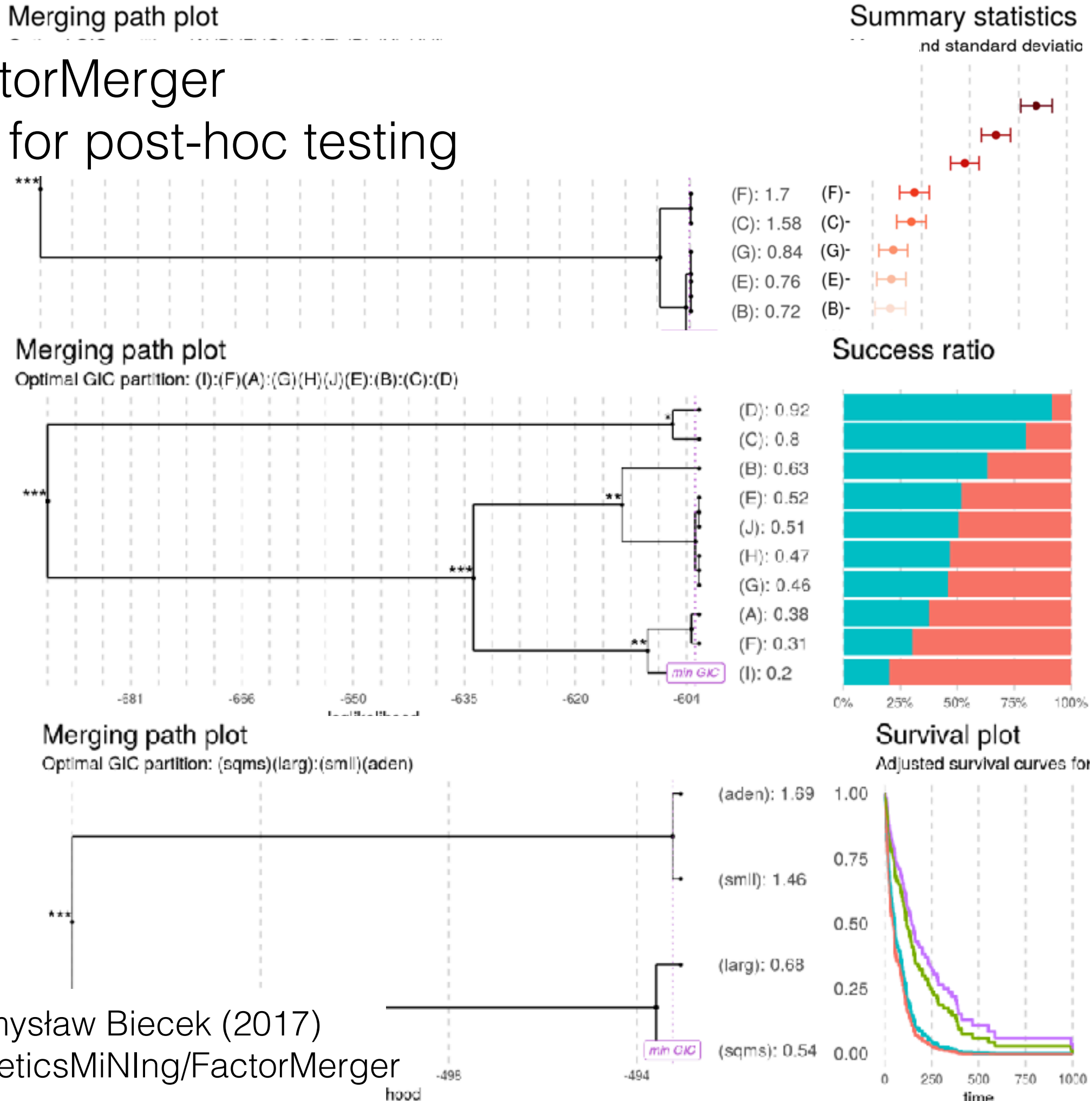
6594

GIC

6458

Package factorMerger

Visualisation for post-hoc testing



Performance charts,
is it a good model?

- model comparisons
- what is the performance?

Validation of assumptions

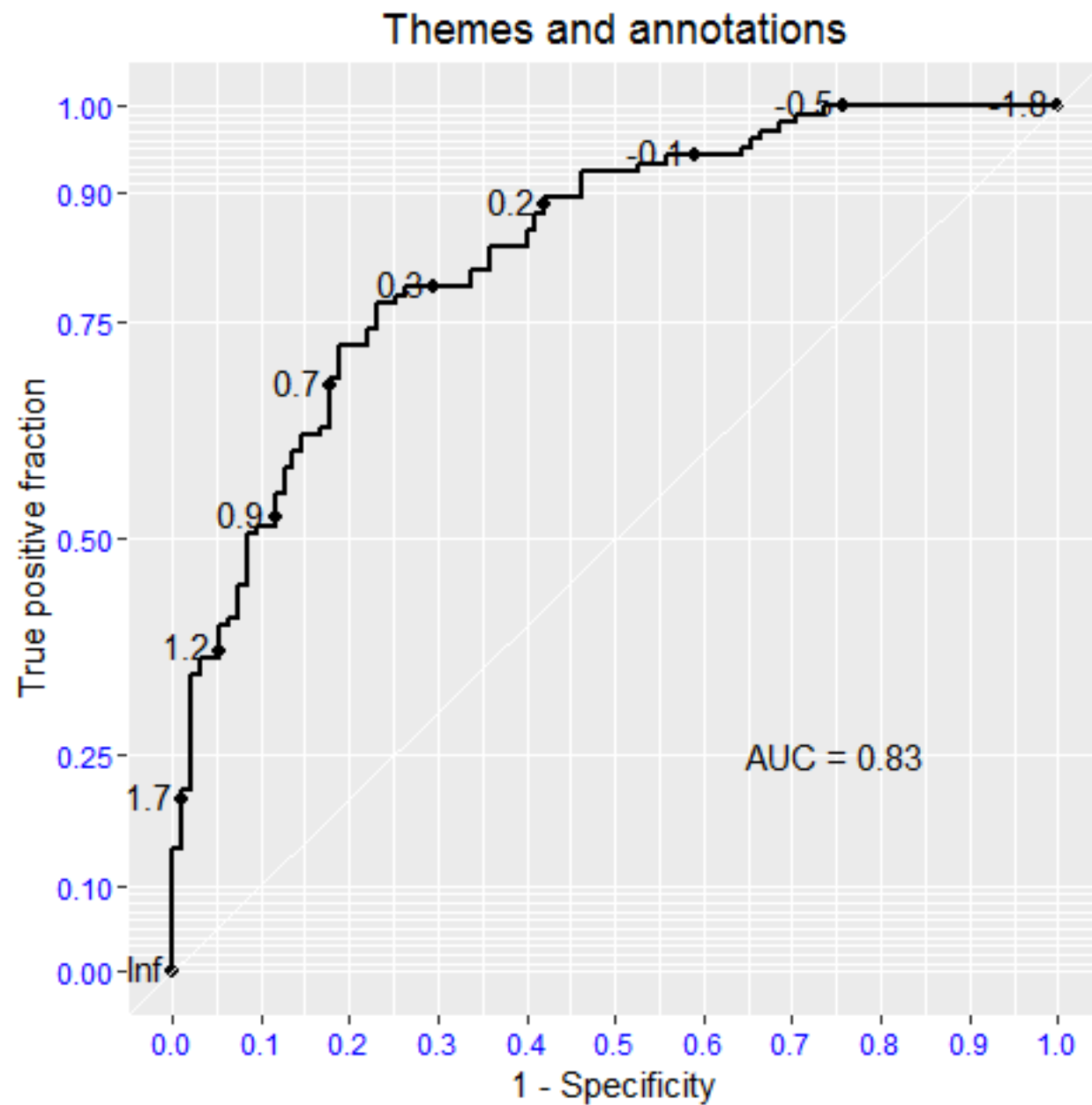
- residuals diagnostics
- are they ok?

Validation

Data validation

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

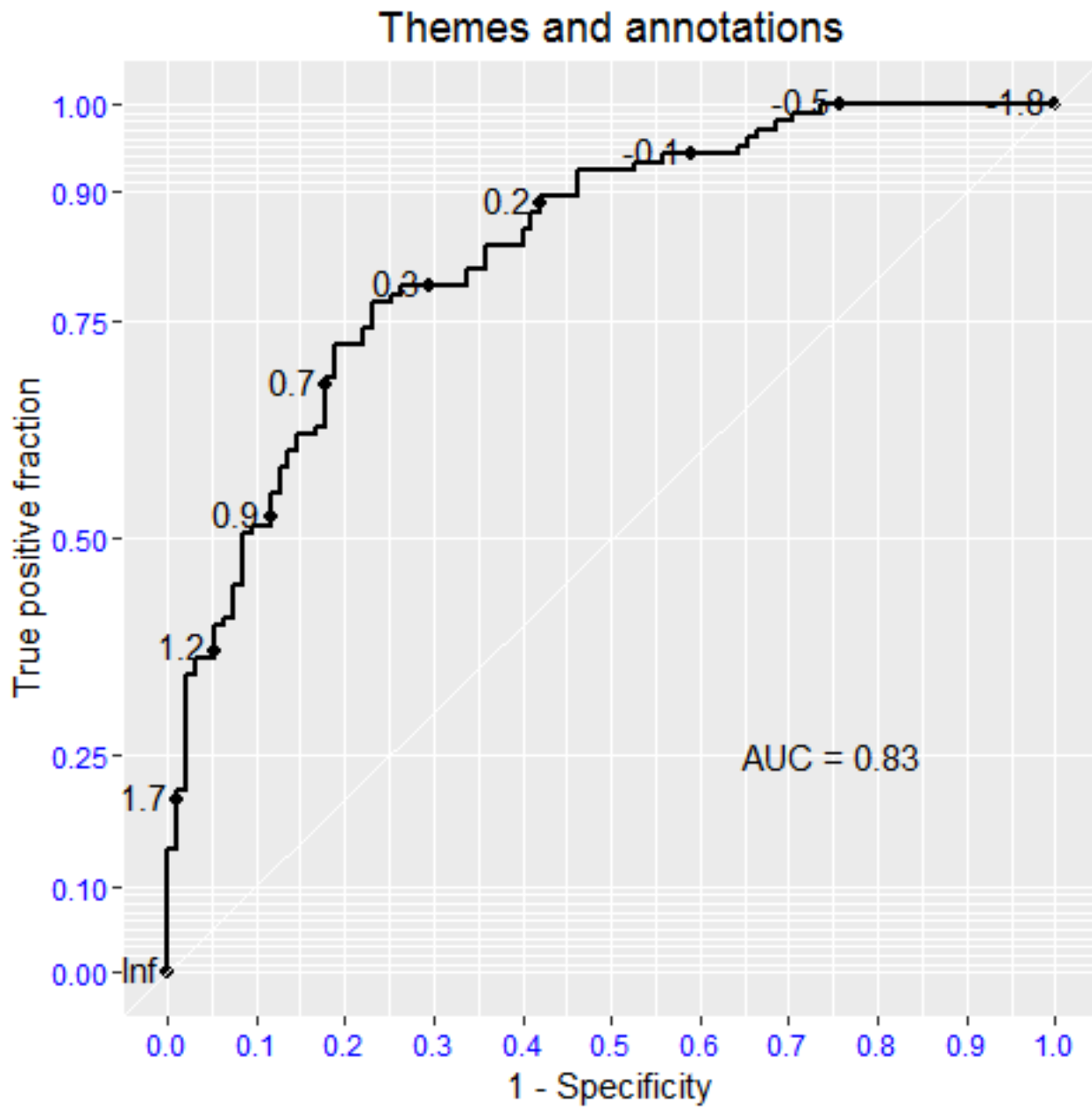
Package plotROC



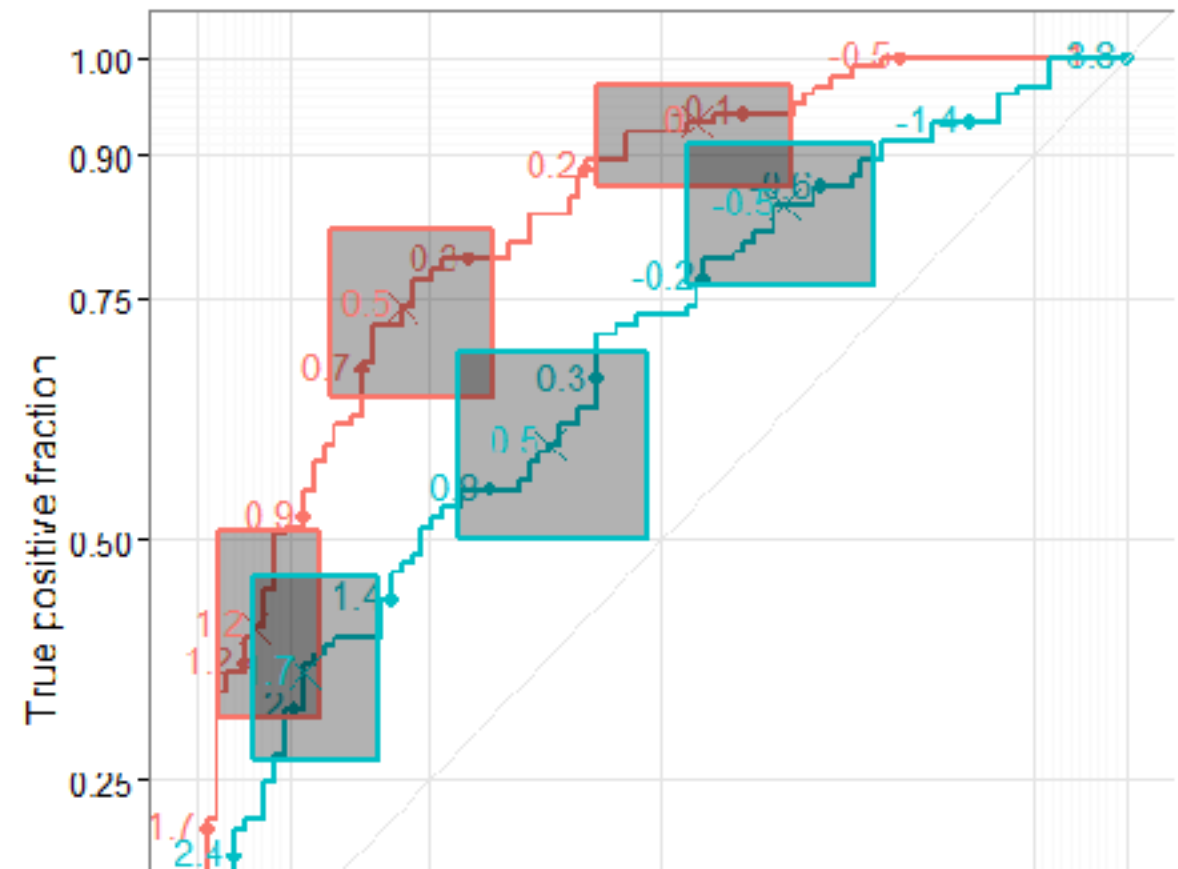
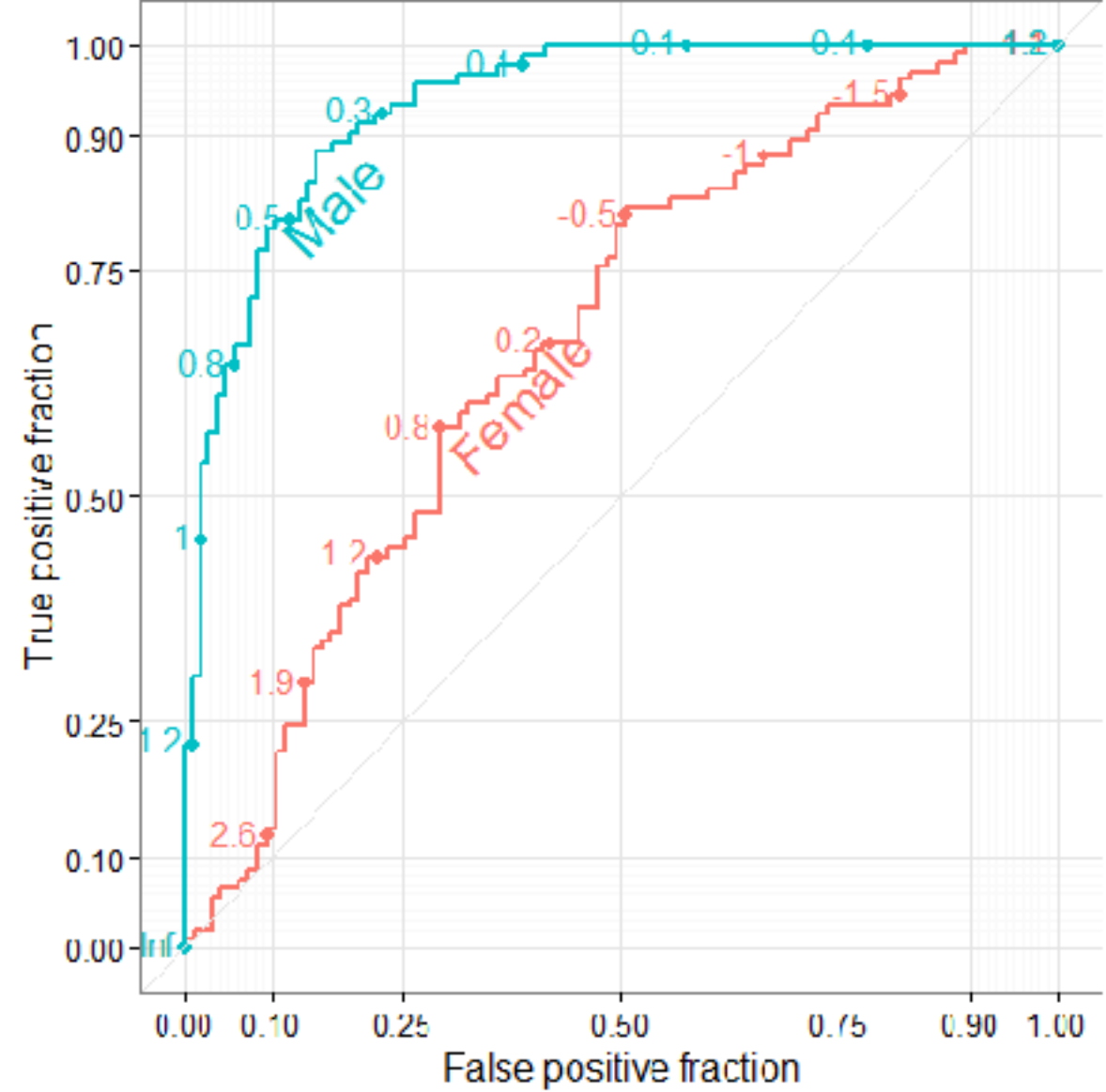
Michael Sachs (2016)

<http://sachsmc.github.io/plotROC>

Package plotROC



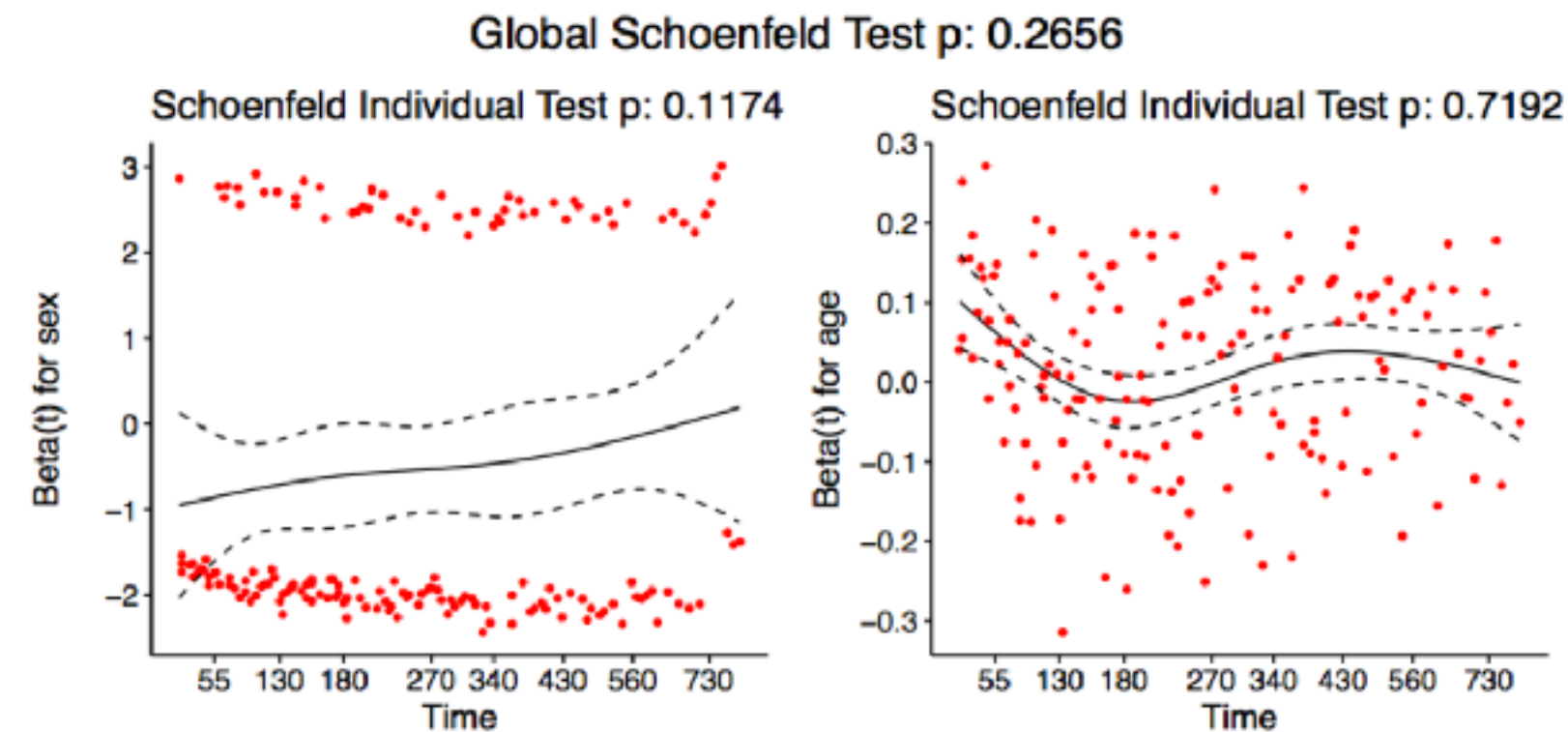
Michael Sachs (2016)
<http://sachsmc.github.io/plotROC>



Package survminer

Diagnostic plots for various residuals:

"martingale", "deviance", "score", "schoenfeld",
"dfbeta", "dfbetas" and "scaledsch"

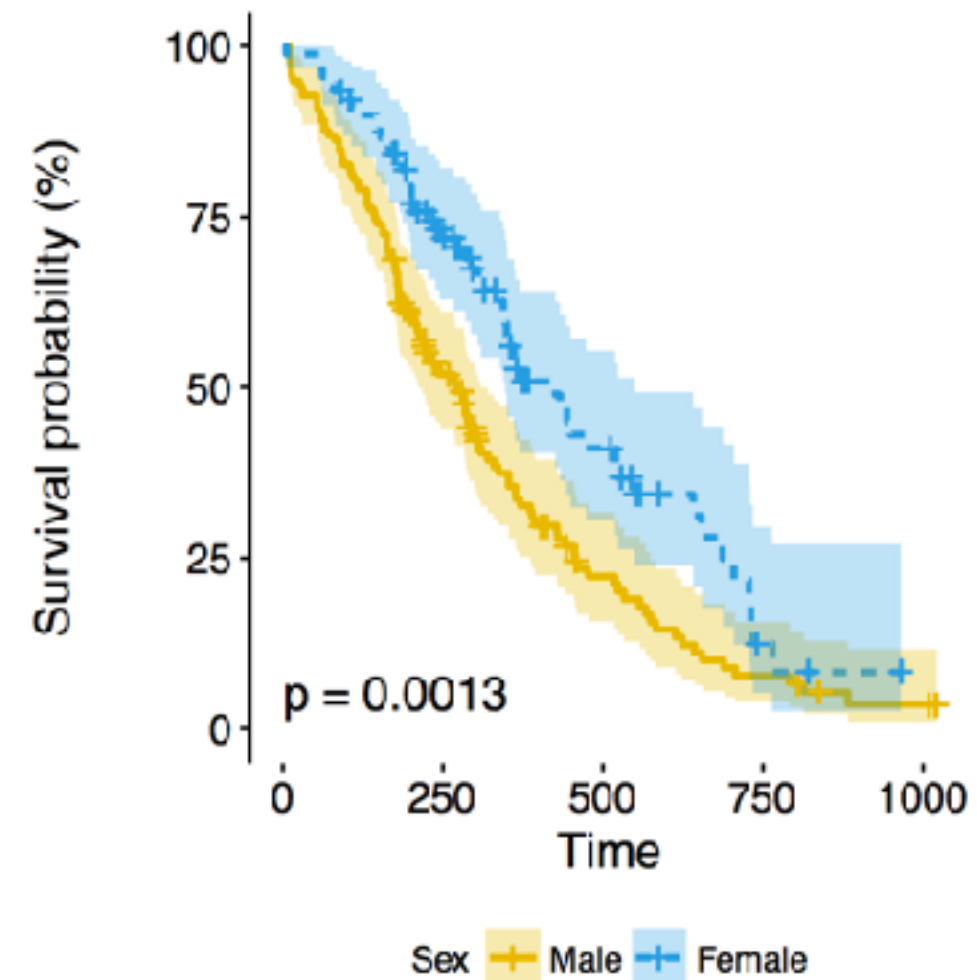
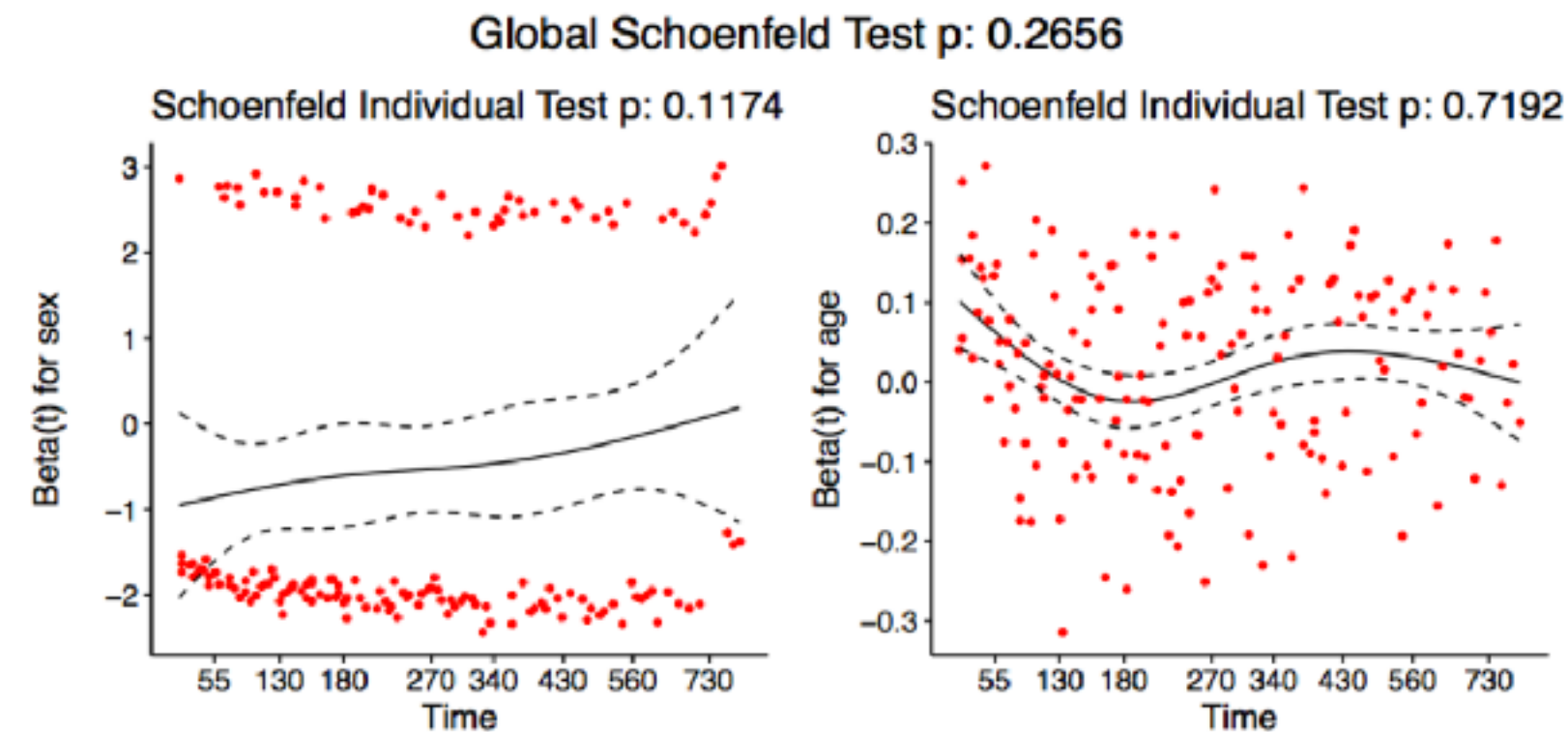
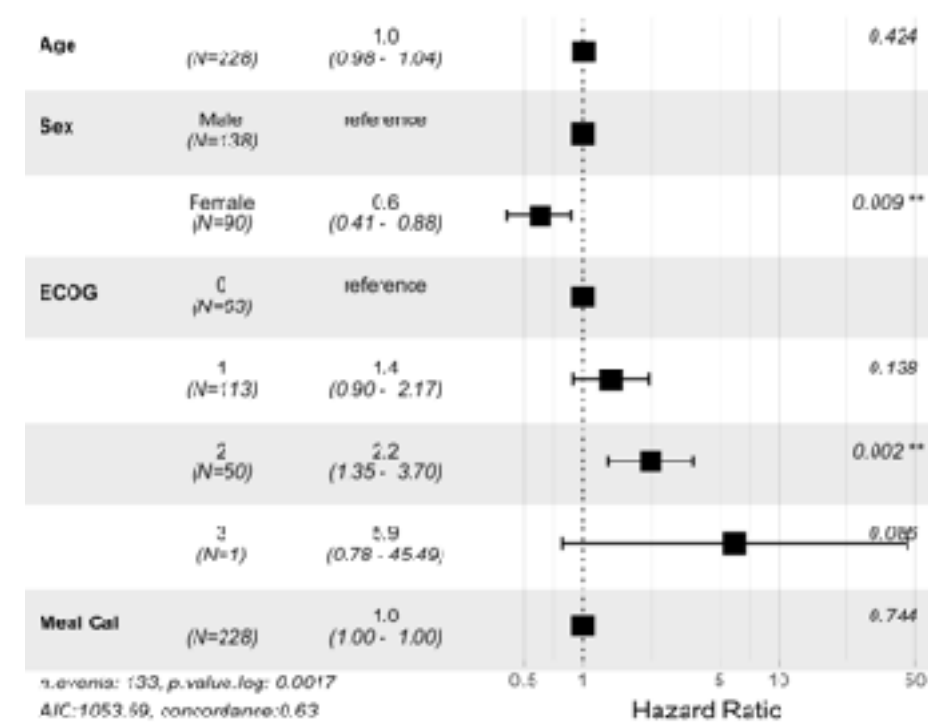


Alboukadel Kassambara, Marcin Kosiński (2016)

<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>

		Number at risk				
Sex	Male	138	62	20	7	2
	Female	90	53	21	3	0
		0	250	500	750	1000

?? 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
```

archivist - reproducible and recordable research

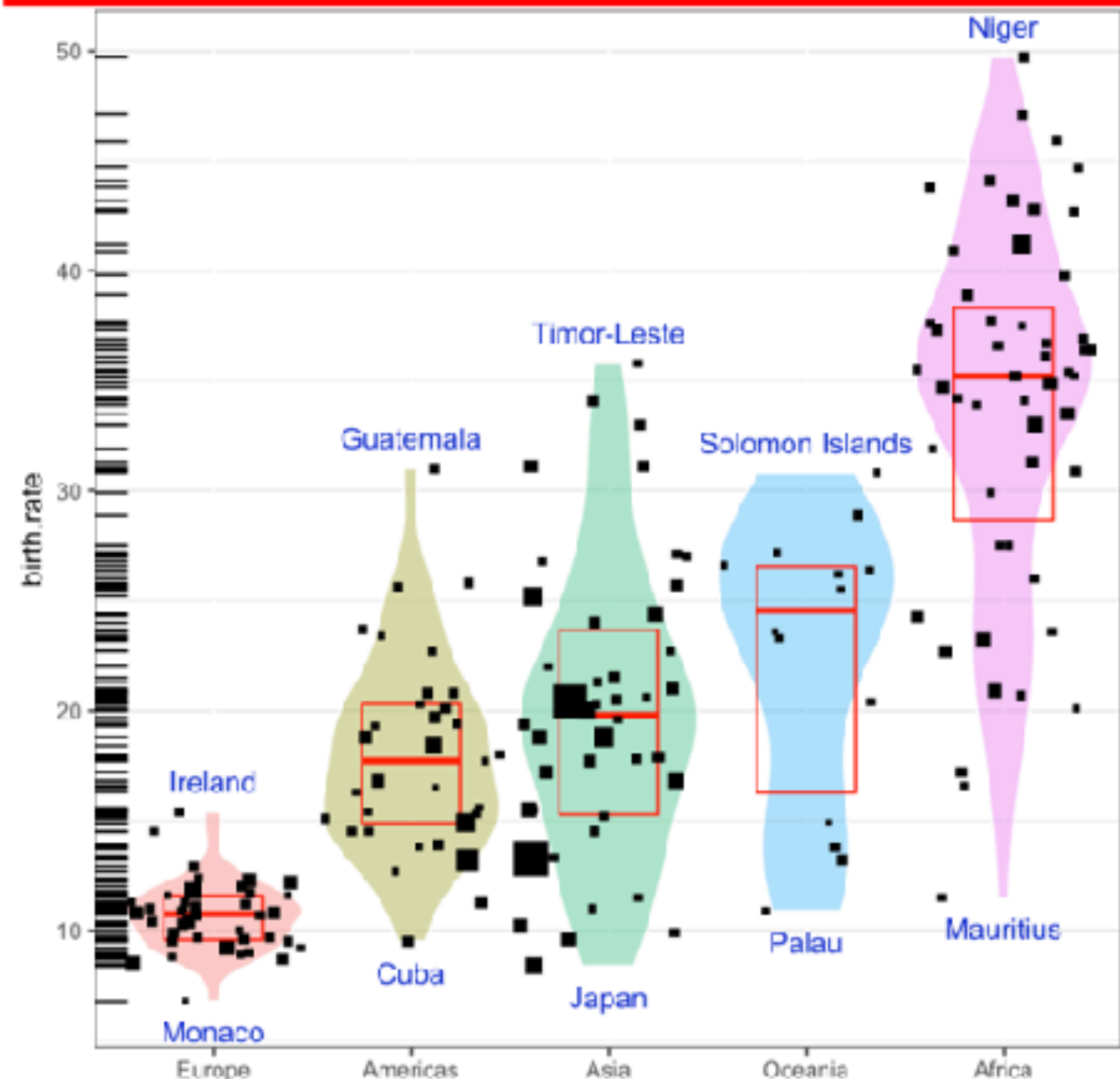
```
library("archivist")  
model <- lm(Sepal.Length ~ Sep  
saveToLocalRepo(model)
```

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

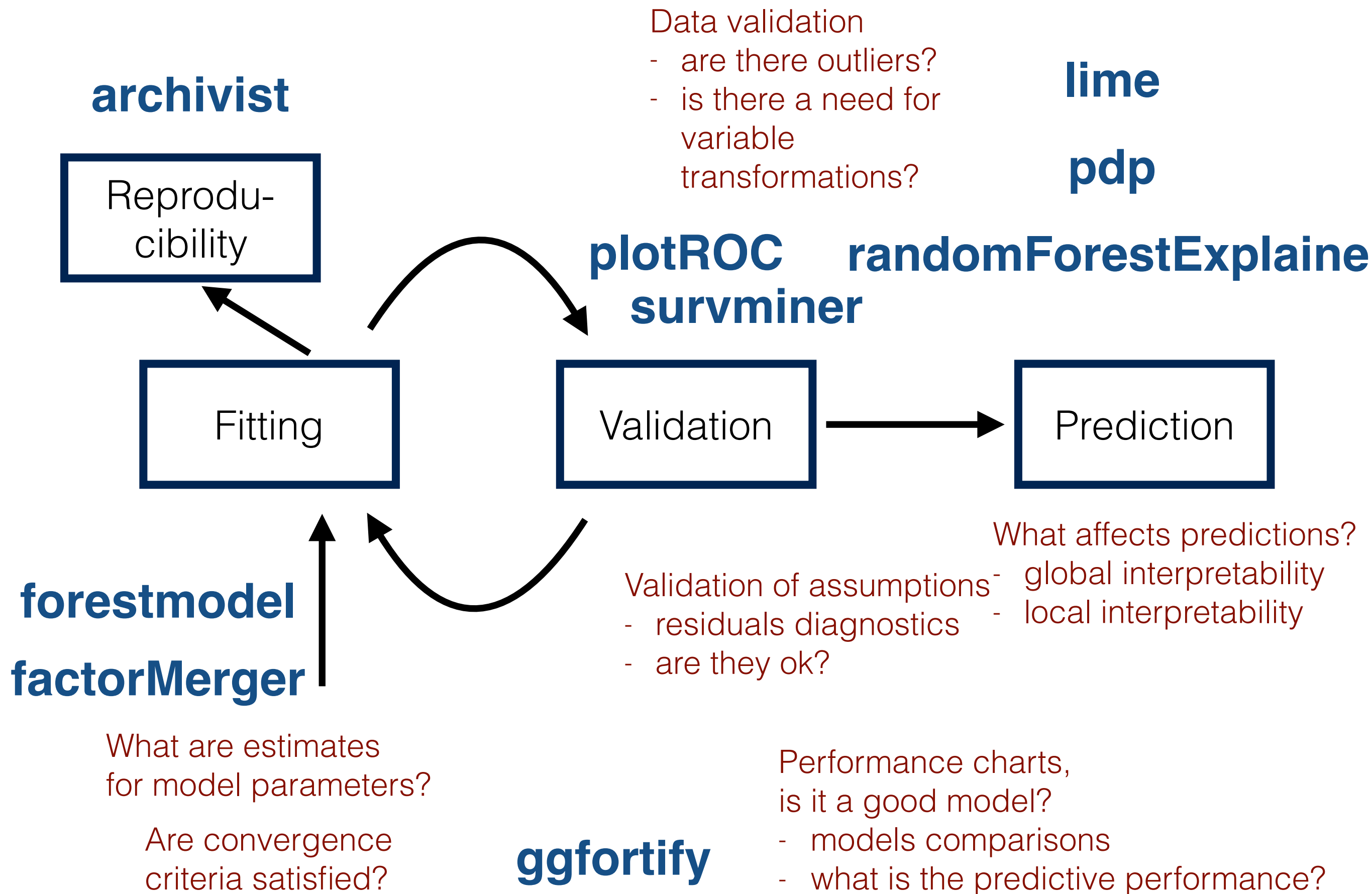
```
## 990861c7c27812ee959f10e5f70  
## 39  
## 0a82efeb8250a47718cea9d7f6  
## 189  
## 7f11e03539d48d35f7e7fe7780  
## 189
```

```
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"))
```

```
load(archivist::aread('pbiecek/Rseje/arepo/ba7f58fafa7373420e3ddce039558140'))
```



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.

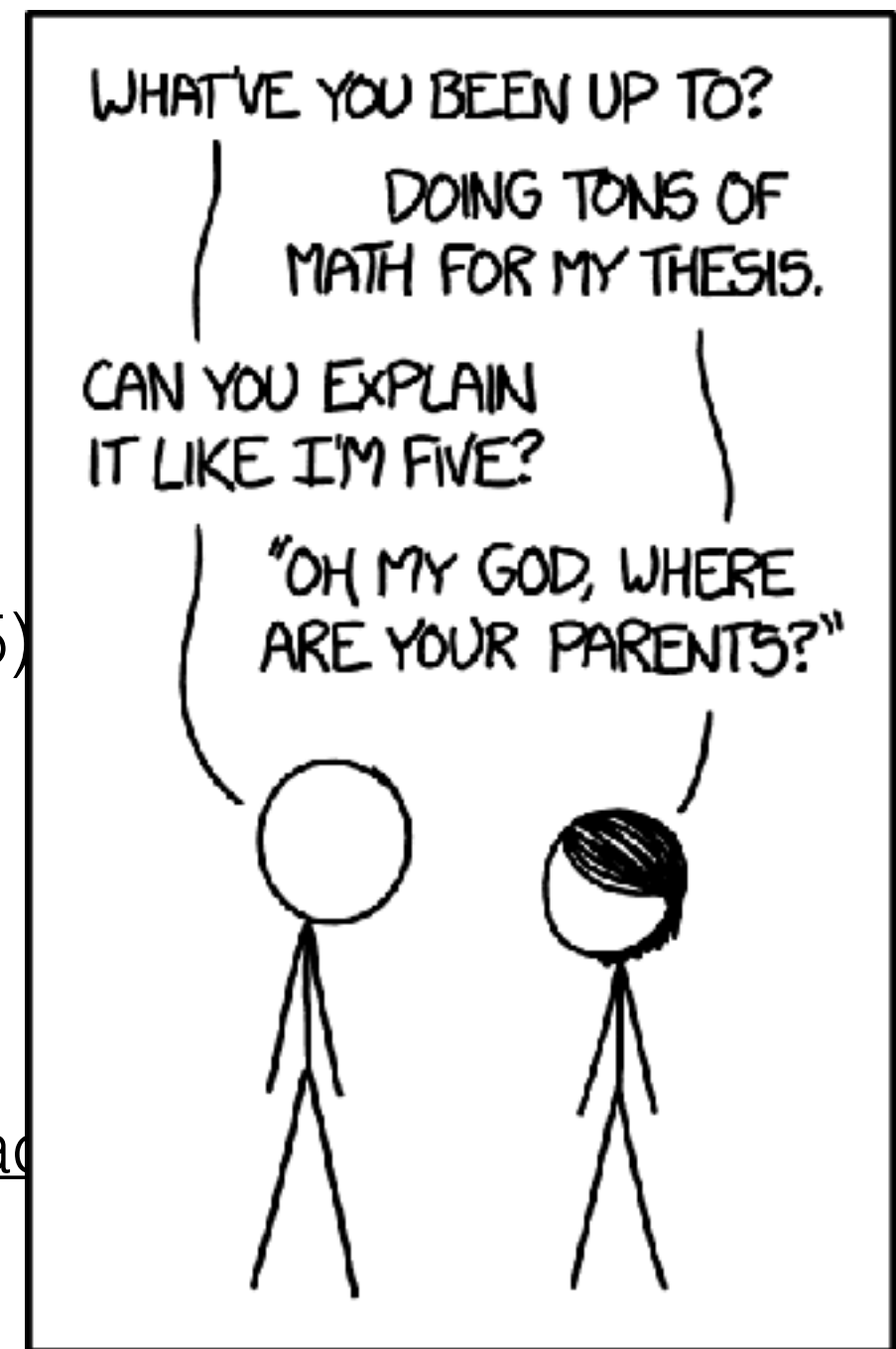
Patrick Hall, Wen Phan, SriSatish Ambati (2017)

<https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning>

rms: Regression Modeling Strategies.

Frank E Harrell (2009-2017) CRAN

<https://cran.r-project.org/web/packages/rms/index.html>



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

Thank you for your attention!