

An Update on Broom

Alex Hayes

2018-07-23

Outline

1. What is broom?
2. broom 0.5.0 release
3. Lessons learned
4. Moving forward

What is broom?

broom tidies model objects

Input: model object

Output: tidy tibble

- `tidy()` summarizes information about model components
- `glance()` reports information about the entire model
- `augment()` adds informations about observations to a dataset

Usage

```
fit <- lm(hp ~ ., mtcars)
```

```
tidy(fit)
```

```
## # A tibble: 11 x 5
```

```
##   term          estimate std.error statistic p.value
```

```
##   <chr>          <dbl>      <dbl>      <dbl>    <dbl>
```

```
## 1 (Intercept)    79.0        185.        0.428    0.673
```

```
## 2 mpg           -2.06         2.09       -0.987    0.335
```

```
## 3 cyl            8.20         10.1        0.813    0.425
```

```
## # ... with 8 more rows
```

```
glance(fit)
```

```
## # A tibble: 1 x 12
```

`broom 0.5.0`

broom 0.5.0: new features

- Tibble output
- New test suite
- New function documentation
- New vignette
- New tidiers
- Many bug fixes
- Many deprecations

broom 0.5.0: tibble output

```
tidy(prcomp(iris[, 1:4]), matrix = "d")
```

```
## # A tibble: 4 x 4
```

```
##       PC std.dev percent cumulative
```

```
##   <dbl>   <dbl>   <dbl>       <dbl>
```

```
## 1     1    2.06    0.925         0.925
```

```
## 2     2    0.493    0.0531        0.978
```

```
## 3     3    0.280    0.0171        0.995
```

```
## # ... with 1 more row
```


broom 0.5.0: tibble output

This was a breaking change. Common issues:

- subsetting with `[` and expecting a vector.
- setting rownames on a tibble.
- using `augment` on models making use of matrix covariates / outcomes.
 - i.e. `survival::Surv()`

Broom has 92 reverse dependencies. This (plus deprecations) broke 15 of them.

broom

CRAN 0.5.0 build passing  build passing coverage 80%

- Line coverage: 40% → 80%
- Higher in practice since we skip deprecated tests
- Revived the Travis CI build

broom 0.5.0: what gets tested

Test that

- `tidy()`, `glance()`, and `augment()` return tibbles.
- `glance()` returns a single row.
- Occasionally check dimensions of output

```
fit <- lm(hp ~ ., mtcars)
td <- tidy(fit)
check_tidy_output(td)
```

broom 0.5.0: new function documentation

- Gave each function it's own roxygen
 - Had started to document too much in one place
- Heavily cross-linked and aliased new docs
- Users requested more explicit doc for stuff like:

```
#' @rdname augment.lm
```

```
#' @export
```

```
augment.glm <- augment.lm
```

broom 0.5.0: dealing with repeated documentation

Many repeated arguments:

```
tidy.betareg <- function(x,  
  conf.int = FALSE,  
  conf.level = .95, ...)
```

```
tidy.ivreg <- function(x,  
  conf.int = FALSE,  
  conf.level = .95,  
  exponentiate = FALSE, ...)
```

Should share documentation for `conf.int`.

roxygen2 templates make this easy:

```
@template param_confint
```

Where `man-roxygen/param_confint.R` looks like:

```
#' @param conf.int Logical indicating  
#'   whether or not to include a  
#'   confidence interval in the tidied  
#'   output. Defaults to `FALSE`.  
#' @md
```


broom 0.5.0: what is templated

Templates currently used to generate:

- @title,
- @description,
- @params, and
- some @return

documentation sections.

broom 0.5.0: new README + pkgdown site

 **broom** part of the tidyverse

IntroReferenceArticles ▾News🔍

Overview

broom summarizes key information about models in tidy `tibble()` s. broom provides three verbs to make it convenient to interact with model objects:

- `tidy()` summarizes information about model components
- `glance()` reports information about the entire model
- `augment()` adds informations about observations to a dataset

For a detailed introduction, please see `vignette("broom")`.

broom tidies 100+ models from popular modelling packages and almost all of the model objects in the `stats` package that comes with base R. `vignette("available-methods")` lists method availability.

If you aren't familiar with tidy data structures and want to know how they can make your life easier, we highly recommend reading Hadley Wickham's [Tidy Data](#).

Installation

```
# we recommend installing the entire tidyverse, which includes broom:
install.packages("tidyverse")
```

Links

Download from CRAN at <https://cloud.r-project.org/package=broom>

Browse source code at <http://github.com/tidyverse/broom>

Report a bug at <http://github.com/tidyverse/broom/issues>

License

MIT + file [LICENSE](#)

Developers

David Robinson
Author, maintainer

Alex Hayes
Author

[All authors...](#)

Dev status

broom 0.5.0: new vignette

Adding new tidiers to broom

Thank you for your interest in contributing to broom! This document is a **work in progress** describing the conventions that you should follow when adding tidiers to broom.

General guidelines:

- Try to reach a minimum 90% test coverage for new tidiers. To check your test coverage we recommend using `covr::report()`.
- `tidy`, `glance` and `augment` methods **must** return tibbles.
- Update `NEWS.md` to reflect the changes you've made
- Follow the [tidyverse style conventions](#). You can use the `styler` package to reformat your code according to these conventions, and the `lintr` package to check that your code meets the conventions.
- Use new tidyverse packages such as `dplyr` and `tidyr` over older packages such as `plyr` and `reshape2`.
- It's better to have a predictable number of columns and unknown number rows than an unknown number of columns and a predictable number of rows.
- It's better for users to need to `tidyr::spread` than `tidyr::gather` data after it's been tidied.
- Add yourself as a contributor to `DESCRIPTION`.
- Pull requests must pass the AppVeyor and Travis CI builds to be merged.

broom 0.5.0: new tidiers

- `lavaan` objects from the `lavaan` package
- `ivreg` objects from the `AER` package
- `Kendall` objects from the `Kendall` package
- `garch` objects from the `tseries` package
- `irlba` lists from the `irlba` package
- `durbinWatsonTest` objects from the `car` package
- `confusionMatrix` objects from the `caret` package
- `glmnet` and `cv.glmnet` objects from the `glmnetUtils` package
- `clm` and `clmm` objects from the `ordinal` package
- `svyolr` objects from the `survey` package, and
- `polr` objects from the `MASS` package.

broom 0.5.0: Bug fixes and pull requests

Start of internship: 134 issues, 34 pull requests

- Triaged two year backlog of issues
- Closed ~80 issues
- Merged 40 pull requests

Current status:

The screenshot shows the GitHub repository page for `tidymodels / broom`. At the top, there are buttons for 'Unwatch' (57), 'Star' (730), 'Fork' (186), and 'Edit'. Below this is a navigation bar with links for 'Code', 'Issues' (78), 'Pull requests' (13), 'Projects' (0), 'Wiki', 'Insights', and 'Settings'. The main content area features the repository description: 'Convert statistical analysis objects from R into tidy format' with a link to <https://broom.tidyverse.org>. Below the description are tags for 'r', 'tidy-data', and 'modeling', along with a 'Manage topics' link. At the bottom, a summary bar displays statistics: '783 commits', '2 branches', '10 releases', and '63 contributors'.

tidymodels / broom

Unwatch 57 Star 730 Fork 186 Edit

Code Issues 78 Pull requests 13 Projects 0 Wiki Insights Settings

Convert statistical analysis objects from R into tidy format <https://broom.tidyverse.org>

r tidy-data modeling Manage topics

783 commits 2 branches 10 releases 63 contributors

broom 0.5.0: Bug fixes and pull requests

- Contributors are enthusiastic and fun to work with
- Lots of issues, but generally easy to fix
- *Beginner Friendly* tag has been immensely popular:

<input type="checkbox"/>		Check if tidy.density works for multivariate densities	beginner-friendly	documentation	2	
#438 opened 2 days ago by alexpghayes						
<input type="checkbox"/>		Update tidy.anova to play well with lme4 anova results	beginner-friendly	feature-request	1	
#434 by alexpghayes was closed 2 days ago 0.7.0						
<input type="checkbox"/>		order of model terms changes in tidy dfs for `clm` and `clmm` when `conf.int = TRUE`	beginner-friendly	bug	4	
#420 opened 16 days ago by IndrajeetPatil 0.7.0						
<input type="checkbox"/>		Adding statistic and p-values to `rq` class objects	beginner-friendly	consistency/specification	feature-request	2
#404 opened 24 days ago by IndrajeetPatil 0.7.0						
<input type="checkbox"/>		`broom::augment` doesn't work if weights are provided for glm	beginner-friendly	bug	1	
#381 opened on Jun 17 by IndrajeetPatil						
<input type="checkbox"/>		`MASS::rlm` tidy output doesn't contain confidence intervals	beginner-friendly	bug	4	
#380 opened on Jun 16 by IndrajeetPatil						
<input type="checkbox"/>		broom::tidy throws error on quantreg::rq model when run on a constant	beginner-friendly	bug	2	
#354 by lboller was closed on Jun 18						
<input type="checkbox"/>		Throw an informative error for `glance.rq` with multiple tau	beginner-friendly	bug		
#330 by alexpghayes was closed on Jun 9						
<input type="checkbox"/>		augment.smooth.spline doesn't accept new data in the data argument	beginner-friendly	consistency/specification	feature-request	2
#318 opened on Jun 5 by alexpghayes						

broom 0.5.0: deprecations: tidy statistical objects only

- Some people were using broom like `as_tibble()`
- Deprecations to prevent this:
 - `tidy.data.frame()`
 - `tidy.matrix()`
 - `tidy.numeric()`
 - `tidy.character()`
 - `tidy.logical()`

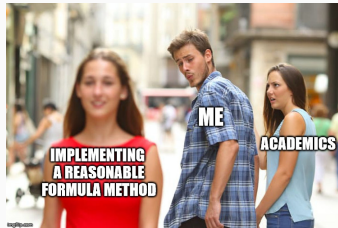
Due to high maintenance burden, moving tidiers for

- lme, lme4 and nmle models,
- brms models,
- rstanarm models, and
- mcmc objects

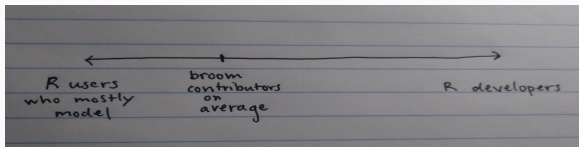
to Ben Bolker's `broom.mixed` (hopefully on CRAN soon).

Lessons learned

Lesson: writing and sanity checking tidiers is hard



Lesson: broom depends on high quality PRs



Key: empower contributors to make high quality PRs

- Document behavioral expectations for tidying methods
- Provide consistent documentation
- Automate as much as possible in tests

Merge now, fix later: community involvement far more important than perfect code

Lesson: `augment()` is hard

Original thought: `tidy()` is most ambiguous method, will be hardest to work with

Incorrect: `augment()` is hard

- Need different behavior for `data` and `newdata` args
- People often don't implement it
- Have to deal with both model input and output

Lesson: there are many ways to represent a model

Representations of a fit model:

- Mathematical: $y \sim \mathcal{N}(X\hat{\beta}, \sigma^2)$
- Code object: `fit <- lm(hp ~ . , mtcars); fit`
- Relational: `tidy(fit)`, `glance(fit)`, `augment(fit)`
- ???

Opinion: need a *tidy modelling* paper to clarify the key objects in play like *tidy data* did

Moving forward

What's happening next

- More tools for contributors
- More deprecations
- Integrating `broom` and friends into `tidymodels`
- I will likely take over as `broom` maintainer

Tools for contributors: tests for argument names

```
check_arguments(tidy.lm)
```

- Checks arguments against master list
- Checks default arguments
 - Shouldn't be missing
 - `conf.int = FALSE`
 - `conf.level = 0.95`
 - `conf.int` and `conf.level` always come as a pair
- Tests written, but not yet passing

Goal: enforce consistency, especially in new PRs

Tools for contributors: tests for column names

```
library(lavaan)

cfa.fit <- cfa(
  "F =~ x1 + x2 + x3 + x4 + x5",
  data = HolzingerSwineford1939, group = "school"
)

select(glance(cfa.fit), 1:5)

## # A tibble: 1 x 5
##   agfi    AIC    BIC    cfi chisq
##   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 0.971 4473. 4584. 0.766 99.3
```

Tools for contributors: column_glossary approach

- Describe acceptable column names in tidy.yaml:
 - column: AIC
 - description: Akaike's Criterion.
 - used_by:
 - ivreg
- Compile tidy.yaml into a column_glossary tibble
- Export column_glossary (downstream package maintainers have asked for this)
- Test output column names against column_glossary
- Populate documentation from column_glossary

Tools for contributors: more vignettes

- Second draft of *Adding new tidiers* vignette
 - Detailed and explicit behavioral specification
 - Will write tests for as many of these as possible
- New vignette on adding tidiers to packages other than broom
 - Based on reexport generics from `modelgenerics`

More deprecations

- Hard deprecate mixed model tidiers in favor of `broom.mixed`
- Soft deprecate time series tidiers in favor of `sweep`
- `tidy.table()`
- `tidy.ftable()`
- etc

tidymodels integration: vision

- Finish any missing aspects of the tidier behavior specification
- Document this clearly
- Develop tests for as much of the spec as possible
- Reach out to package maintainers
- Invite them to join tidymodels once they meet the spec??
- Some system for keeping track of where tidiers live
- Potentially break boom into smaller pieces

```
library(tidymodels)  # load everything
```

tidymodels integration: possible collaborations

- `sweep`: time series
- `tidytext`: natural language processing
- `broomstick`: trees
- `broom.mixed`: mixed models, bayesian models
- `biobroom`: bioconductor objects
- `schoenberg`: gams
- `tidybayes`: bayesian models
- `broom.base`: broom infrastructure
- `mlbroom`: doesn't exist yet but demand seems high

Priorities for the next 3 weeks

1. Vignette on addings tidiers to packages other than broom
2. Infrastructure for auto-building `@return` documentation
3. Write `tidy.yaml`, `glance.yaml` and `augment.yaml`
4. Move time series stuff into sweep
5. Reach out to potential collaborators
6. Make sure existing tidiers meet behavioral specification

Questions?

Read more about [broom 0.5.0 release](#) on the tidyverse blog.

You can follow broom development on our [Github page](#).

@alexpghayes on Twitter

alexpghayes@gmail.com

Appendix

broom 0.5.0: matrix column and augment example

```
y <- rnorm(5)
x <- matrix(rnorm(10), nrow = 5)

df <- data.frame(x, y)    # ok
tibble::tibble(x, y)      # errors

fit <- lm(y ~ x, df)      # problem: this works
augment(fit)              # this goes kaboom
```

Passing data argument can help:

```
augment(fit, data = df)  # happy again
```


Aside: model coverage

```
# glance.arima coverage was 100 percent.
```

```
# tested output of:
```

```
glance(arima(lh, order = 1:3))
```

```
# but this was broken until recently:
```

```
glance(arima(lh, order = 1:3, method = "CSS"))
```

- Same class can correspond to many varied model objects
- Hard to write varied tests for unfamiliar model objects

Aside: arguments disappearing into ...

```
fit <- lm(hp ~ ., mtcars)

# misspelled argument
td <- tidy(fit, conf.int = TRUE, comf.level = 0.9)

# no error, output looks exactly like
# you might expect
```