The Model Representation Problem

And how broom makes life better

Alex Hayes

2018-09-12

Outline

- Who am I?
- What is the model representation problem?
- What is broom?
- What is tidy data?

About Me

- Summer intern at RStudio last summer
- Primary maintainer of `broom` package
- Just started a PhD in Statistics at UW-Madison

Active on #rstats Twitter and Github

The Model Representation Problem

What is a model?

Let x be values that live in some space $\mathcal{X}$, and let y be observations of interest that live in some space $\mathcal{Y}$. A *statistical model* is a set of probability distributions $\mathcal{P}(y|x)$ indexed by parameters $\theta \in \Theta$[1].

[1] In some cases we treat θ as itself random, which means that our model is a class of probability distributions $\mathcal{P}(y, \theta|x)$.

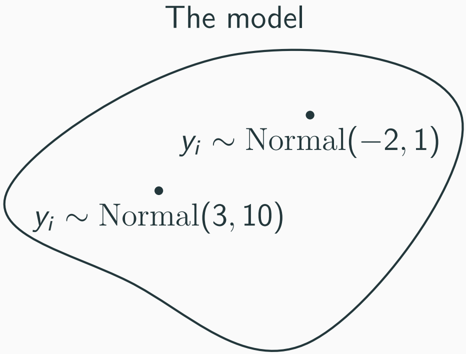
Example: the normal model

$$y_i \overset{\text{iid}}{\sim} \mathrm{Normal}(\mu, \sigma^2)$$

Here $\theta = (\mu, \sigma^2)$ and the parameter space is $\mathbb{R} \times \mathbb{R}^+$.

Visualizing the normal model

A model is a *set*.


The model
$y_i \sim \text{Normal}(-2, 1)$
$y_i \sim \text{Normal}(3, 10)$

We call a single element of a model a *fit*. The distribution with $\mu = -2, \sigma^2 = 1$ is a fit, for example.

Another parametric example: the linear model

Given response y and predictor variables x_1 and x_2, the linear model looks like:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \qquad \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{Normal}(0, \sigma^2)$$

This model says that y is i.i.d with a mean that depends on x and $\vec{\beta}$, and with fixed variance σ^2.

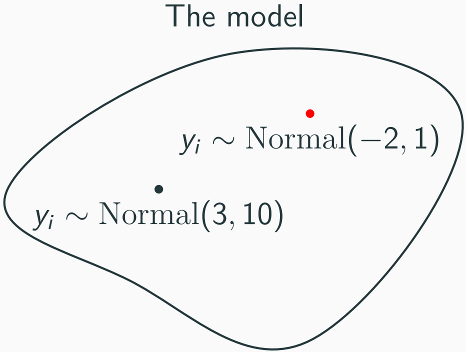
Model fitting

To learn a model, we have to find the best fit. This is equivalent to finding the best parameters in the parameter space.

For now, we won't worry about this. We will just assume we have a way to find the best fit.

Model fitting

Suppose our data comes from a Normal(-2, 1) distribution. We want:


The model
$y_i \sim \mathrm{Normal}(-2, 1)$
$y_i \sim \mathrm{Normal}(3, 10)$

The representation problem

Representing a fit in mathematical terms

In math terms, we can identify fits by their corresponding parameter vectors.

For example, for the normal model:

$\theta = (\mu, \sigma^2) = (-2, 1)$

This doesn’t work so well for code

- we want to be able to call methods a model

The representation problem

Nobody agreed on a standard way to represent fits as code objects!

What this means

The broom package

How should we represent models in code?

`broom` provides an *ad hoc* solution to the model representation problem

`broom` says that we should:

- use tidy data strucutres

and three generic functions that generate representations of fits that follow these principles

`broom` adopts the following philosophy:

Examples of model fitting in R

Example: fitting the normal model

```r
# simulate some normal data
# with mean of -2 and variance of 1
x <- rnorm(5000, -2, 1)

# fit a normal model to this data
normal_fit <- MASS::fitdistr(
  x,      # our data
  dnorm,  # use the normal model!
  start = list(mean = 0, sd = 1)
)

normal_fit$estimate
```

```
##       mean         sd
```

```r
# using the lm function tells R
# that we want the linear model
```

The `broom` package

broom: case studies

tidy use cases

Create a simple report of model

```
# knitr::kable(tidy(fit))
```

*pro-tip: use a fancy wrapper that does better stuff

visualize the sampling distribution of an estimate

sort terms by p-value

histogram of p-values

reread erle's blog post on batch adjustments and use similar exampels

volcano plot of terms in a omics regression

Mclust + PCA projection to lower dimension to visualize high dimensional clusters?

glance use cases

Selecting a model based on AIC

*please don't do this

Visualize the AIC of various models

-parallel coordinates plots

augment use case

Visualizing bootstrapped model fits

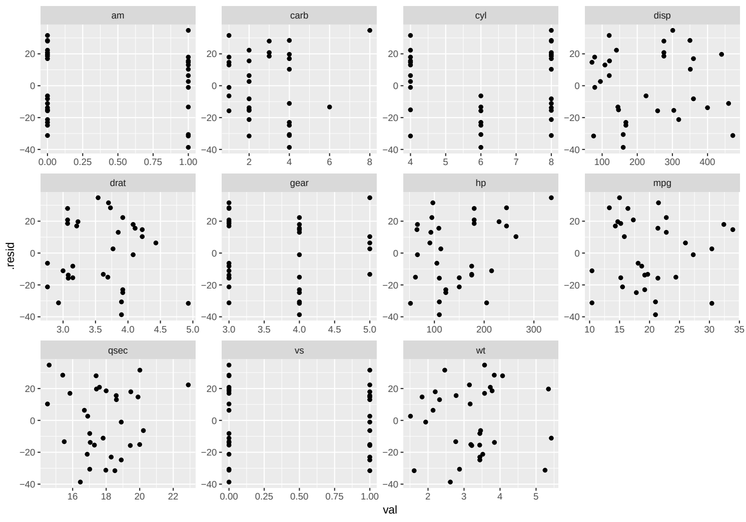
Inspecting residuals from multiple linear regression

```r
library(tidyverse)

fit <- lm(hp ~ ., mtcars)
au <- broom::augment(fit)

au %>%
  gather(x, val, -contains(".")) %>%
  ggplot(aes(val, .resid)) +
  geom_point() +
  facet_wrap(~x, scales = "free")
```

Inspecting residuals from multiple linear regression



Partial dependence plots

A grammar of modeling

The future of broom

tidymodels

Questions?

Read about the recent `broom` release on the tidyverse blog.

https://github.com/tidymodels/broom/

@alexpghayes alexpghayes@gmail.com