http://courses.had.co.nz

Modelling

Hadley Wickham

Assistant Professor / Dobelman Family Junior Chair

Department of Statistics / Rice University

July 2012

1. Introduction to modelling in R
2. Modelling algebra
3. Predictions, residuals and diagnostics
4. Machine learning

Introduction

Function	Model	Generalisation
`lm`	Linear model	
`glm`	Generalised linear models	Poisson, binomial, other exp. family
`gam`	Generalised additive models	Smooth functions
`rlm`	Robust linear models	Heavy-tailed errors
`lme`	Mixed effect models	Multiple sources of variation

Function	Package	Book
`lm`	`stats`	http://amzn.com/1584884258
`glm`	`stats`	http://amzn.com/158488424X
`gam`	`mgcv`	http://amzn.com/1584884746
`rlm`	`MASS`	http://amzn.com/0387954570
`lme`	`nlme`	http://amzn.com/1441903178
`lmer`	`lme4`	http://lme4.r-forge.r-project.org/book/

Getting help

```
?predict
?predict.lm
?predict.glm
?predict.nlme
```

Algebra

R's standard model notation

$$Y_{ij} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon$$

Y ~ A + B + A:B

$$Y_{ij} = \mu + a_i + (ab)_{ij} + \varepsilon$$

Y ~ A + A:B

Formula	Name	Alternative
`x + y`	addition	
`x:y`	interaction	
`x * y`	crossing	`x + y + x:y`
`x / y`	nesting	`x + x:y`
`(x * y) - y`	removal	`x + x:y`

Your turn

Have a go at simplifying the following model formulae:

```
x * y * z - x:y - y:z - z:x

(a + b + c)^2 + a:b:c

x + x^2
```

```
(x * y * z) - x:y - y:z - z:x
# =>
(x + y + z + x:y + x:z + y:z + x:y:z) -
  x:y - y:z - z:x
# =
x + y + z + x:y:z

(a + b + c)^2 + a:b:c
# =>
(a + b + c) * (a + b + c) + a:b:c
# =>
a + b + c + a + b + c + a:a + a:b + a:c + b:a + b:b + b:c +
c:a + c:b + c:c + a:b:c
# =
a + b + c + a:b + a:c + b:c + a:b:c
# =
a * b * c
```

```
# Watch out for this:
y ~ x + x^2
# equivalent to
y ~ x + (x + x + x:x)
# ie.
y ~ x

# If you want ^ interpreted as you expect:
y ~ x + I(x ^ 2)

# But it's better to do
y ~ poly(x, 2)

# or even better
library(splines)
y ~ ns(x, 2)
```

Formula	Meaning
`x:x`	x
`x + x`	x
`-1`	remove implicit intercept term
`(x + y + z)^2`	All second order (and lower) terms
`a %in% b`	`b:a`
`y ~ .`	y versus all other variables in data

Predictions, residuals, diagnostics

Fuel economy of 2011 mpg, as collected by the US Environmental protection agency.

What affects fuel consumption?

http://www.flickr.com/photos/27147/3411775886/

```
library(ggplot2)

# Convert to litres per 100 km
mpg$consump <- 235.214584 / mpg$comb
```

```
qplot(displ, consump, data = mpg)

mod1 <- lm(consump ~ displ, data = mpg)
mpg$pred <- predict(mod1)
mpg$resid <- resid(mod1)
```

```
qplot(displ, pred, data = mpg)
qplot(displ, resid, data = mpg) +
  geom_hline(yintercept = 0)
```

Your turn

Do you think the class variable should be included in the model? If so, then how?

Create some plots to illustrate the impact this has on the model

```
qplot(displ, consump, data = mpg) +
  facet_wrap(~ class)
qplot(displ, consump, data = mpg) +
  geom_point(aes(y = pred), colour = "red") +
  facet_wrap(~ class)

mod2 <- lm(consump ~ displ + class, data = mpg)
mpg$pred <- predict(mod2)
mpg$resid <- resid(mod2)
qplot(displ, consump, data = mpg) +
  geom_point(aes(y = pred), colour = "red") +
  facet_wrap(~ class)
```

```
# Let's look at the shape of the full model surface.
# expand.grid makes it easy to generate an evenly spaced
# grid of predictors over the entire design space.
grid <- expand.grid(
  class = unique(mpg$class),
  displ = seq(min(mpg$displ), max(mpg$displ),
    length = 20)
)
grid$consump <- predict(mod1, newdata = grid)

qplot(displ, consump, data = mpg) +
  geom_line(data = grid, colour = "grey50") +
  facet_wrap(~ class)

qplot(displ, resid, data = mpg) +  facet_wrap(~ class)
```

Function	Use
`predict(mod)`	Predicted values for original dataset
`predict(mod, newdata)`	Predicted values for new database
`resid(mod)`	Residuals
`lm.influence(mod)`	Influence measures

Books

Regression Modeling Strategies. *Frank Harrell*. http://amzn.com/0387952322

Mixed-Effects Models in S and S-PLUS. *Jose Pinheiro and Douglas Bates*. http://amzn.com/1441903178 (but also http://lme4.r-forge.r-project.org/book/)

Data Analysis Using Regression and Multilevel/Hierarchical Models. *Andrew Gelman and Jennifer Hill*. http://amzn.com/052168689X

Machine learning

	Packages	References
Support vector machines	`e1071, kernlab`	`http://bit.ly/LcseOe`
Shrinkage methods	`glmnet`	`http://bit.ly/LM3dzp`
Boosting	`mboost, gbm, GAMBoost`	
Random forests	`randomForest`	

http://www-stat.stanford.edu/~tibs/ElemStatLearn/

http://cran.r-project.org/web/views/MachineLearning.html