Exploring trends

Hadley Wickham

Assistant Professor / Dobelman Family Junior Chair
Department of Statistics / Rice University

July 2010

1. Line plots
2. Intro to modelling
3. Many small models
4. One big model

Getting started

```
library(ggplot2)
tb <- read.csv(”tb.csv”)

info <- read.csv(”world-info.csv”)
info$income <- factor(info$income,
  c(””, ”Low”, ”Lo-mid”, ”Up-mid”, ”Hi”))

tb <- merge(tb, info, by = ”iso2”)
tb$country <- NULL
```

Trends over time

We are also interested in how things are changing over time. Typically, changes over time are best display with a line plot (`geom = ”line”`).

Must remember to set the **group** aesthetic, to get the correct number of lines.

Your turn

Use facetting and aesthetics to explore the relationship between region, income and tb trends.

What problems do you encounter?

Problems

Rates are very noisy, so it’s hard to see any global trends.

Instead, can fit models and look the coefficients. (I can’t find a particularly compelling story with this data, but it is useful technique in general)

We’ll do this first graphically and then more formally

```
qplot(year, rate, data = tb, geom = "line", group = iso2) +
  geom_smooth()
qplot(year, rate, data = tb, geom = "line", group = iso2) +
  geom_smooth(se = F)
qplot(year, rate, data = tb, geom = "line", group = iso2) +
  geom_smooth(method = lm, se = F)

qplot(year, rate, data = tb, geom = "line", group = iso2) +
  facet_wrap(~ income) +
  geom_smooth(se = F)

qplot(year, rate, data = tb, geom = "line", group = iso2) +
  facet_wrap(~ income) +
  geom_smooth(aes(group = 1), se = F, size = 2)
```

Your turn

Using what you know about grouping, create a plot that shows smoothed overall trends by region and income, with one variable displayed with facetting and the other with aesthetics.

```
ggplot(tb, aes(year, rate)) +
  geom_smooth(aes(colour = income), se = F, size = 2) +
  facet_wrap(~ region) +
  scale_colour_brewer(pal = "YlOrRd")

ggplot(tb, aes(year, rate)) +
  geom_smooth(aes(colour = region), se = F, size = 2) +
  facet_wrap(~ income) +
  scale_colour_brewer(pal = "YlOrRd")

ggplot(tb, aes(year, rate)) +
  geom_smooth(aes(colour = income), method = lm,
    size = 2) +
  facet_wrap(~ region) +
  scale_colour_brewer(pal = "YlOrRd")
```

Modelling

```
za <- subset(tb, iso2 == "ZA")
qplot(year, rate, data = za, geom = "line")

# Explore model for additive change
model <- lm(rate ~ year, data = za)
model
summary(model)
coef(model)
coef(summary(model))
model <- lm(rate ~ I(year - 1999), data = za)

# See predictions
za$pred <- predict(model)
qplot(year, rate, data = za, geom = "line") +
  geom_line(aes(y = pred), colour = "red")

# What does this model tell us about TB in Zaire?
```

Your turn

Fit a similar model to the US. What does the model tell you about TB in the US? Is it a good summary?

For all countries?

Need to repeat this process for all countries.

Three options: split + for loop, split + lapply, dlply

Important skills to gain in the long-term, but usually mystifying the first time you see them.

A fundamental programming virtue is **laziness**: you want to do as little work as possible, and have the computer do all the heavy lifting

```
library(plyr)

models <- dlply(tb, ”iso2”, function(df) {
  lm(rate ~ I(year - 1999), data = df)
})

length(models)
models[[1]]
```

```
coefs <- ldply(models, coef)
names(coefs)[2:3] <- c(”intercept”, ”slope”)
# add in country info
coefs <- merge(coefs, info, by = ”iso2”)
```

Your turn

Is there any relationship between slope and intercept and income and region?

Use your visualisation skills to explore.

A better model?

```
tb$ystart <- tb$year - 1999
tb$healthy <- tb$cases
tb$sick <- tb$pop - tb$healthy

model <- glm(cbind(healthy, sick) ~
ystart * iso2, data = tb, family =
”binomial”)
```

Other models

These models are just two ends of a continuum—completely separate and completely pooled—and there are many models in between. However, describing and fitting these is much more complicated, so it’s a topic for another time. See Andrew Gelman's “Data analysis using regression and multilevel/hieararchical models”.

More about plyr

http://had.co.nz/plyr
and tomorrow