Data Visualization With R

Zack Treisman

Spring 2026

Philosophy

Exploration[1] ← **Visualizations** → Publication

Graphics systems in R

- Base R: intuitive, perhaps limited.
- `ggplot2`: robust and widely used.
- `lattice`: also nice, often older.

[1] Be wary of inference based on purely exploratory data analysis. If you look at your data until you find a pattern, and then test for that pattern, the significance levels of that test are inflated.

Loading data

Step 0 of visualizing your data with R is loading it.

- Clean your data spreadsheet:
  - Remove non-data (summaries, etc.)
  - Fix typos
  - Make good variable names
    - meaningful
    - not too long
    - no spaces - use under_score or camelCaps instead
    - don't start with a number
  - More good advice: Data Carpentry
- `csv` (comma separated variable)
- working directory
  - possibly *data* subdirectory.
- `read.csv` or `read_csv`.

Check the data loaded correctly

- `str()` Are variables coded correctly? (factors, dates)
- `head()` or `View()`

```
str(ReedfrogPred)
```

```
## 'data.frame':    48 obs. of  5 variables:
##  $ density : int  10 10 10 10 10 10 10 10 10 10 ...
##  $ pred    : Factor w/ 2 levels "no","pred": 1 1 1 1 1 1 1 1 2 2 ...
##  $ size    : Factor w/ 2 levels "small","big": 2 2 2 2 1 1 1 1 2 2 ...
##  $ surv    : num  9 10 7 10 9 9 10 9 4 9 ...
##  $ propsurv: num  0.9 1 0.7 1 0.9 0.9 1 0.9 0.4 0.9 ...
```

```
head(SeedPred)
```

```
##   station dist species       date seeds tcum tint taken available
## 1       1   10     psd 1999-03-25     5    0   NA    NA        NA
## 2       1   10     psd 1999-03-28     5    3    3     0         5
## 3       1   10     psd 1999-04-04     5   10    7     0         5
## 4       1   10     psd 1999-04-11     5   17    7     0         5
## 5       1   10     psd 1999-04-18     0   24    7     5         5
## 6       1   10     psd 1999-04-25     0   31    7     0         0
```

```
# data from R package emdbook
```

Exploration

Data are in R, now what?

- Check numerical summaries.

```r
summary(ReedfrogPred)
```

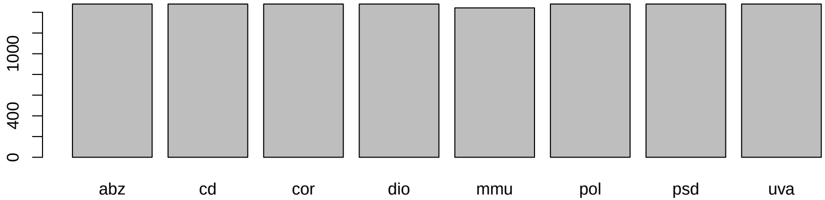
```
##     density         pred       size         surv          propsurv     
##  Min.   :10.00   no  :24   small:24   Min.   : 4.00   Min.   :0.1143  
##  1st Qu.:10.00   pred:24   big  :24   1st Qu.: 9.00   1st Qu.:0.4964  
##  Median :25.00                        Median :12.50   Median :0.8857  
##  Mean   :23.33                        Mean   :16.31   Mean   :0.7216  
##  3rd Qu.:35.00                        3rd Qu.:23.00   3rd Qu.:0.9200  
##  Max.   :35.00                        Max.   :35.00   Max.   :1.0000  
```

- Make some graphics!
  - patterns; expected/unexpected?
  - Data issues?

Barplots and Histograms - one variable

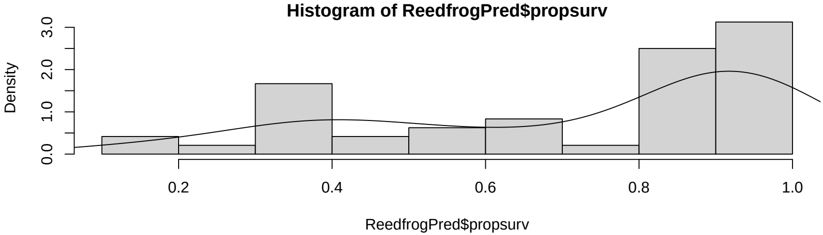
- Barplots ↔ categorical variables.

```
barplot(table(SeedPred$species))
```



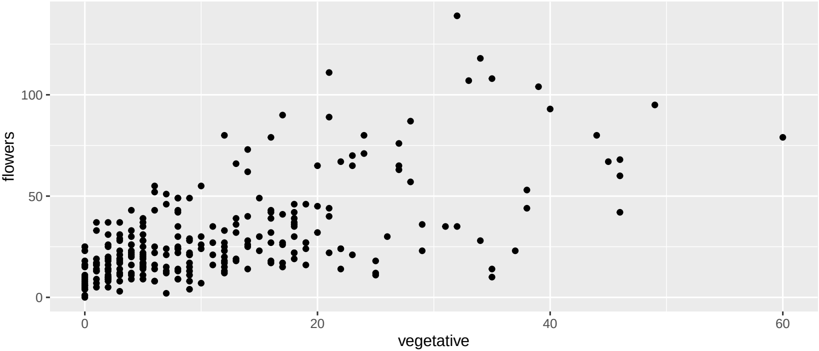
- Histograms, density estimates ↔ numeric variables.

```
hist(ReedfrogPred$propsurv,freq=F); lines(density(ReedfrogPred$propsurv))
```



Scatterplots - two numeric variables

```r
ggplot(Lily_sum, aes(vegetative, flowers)) + geom_point()
```

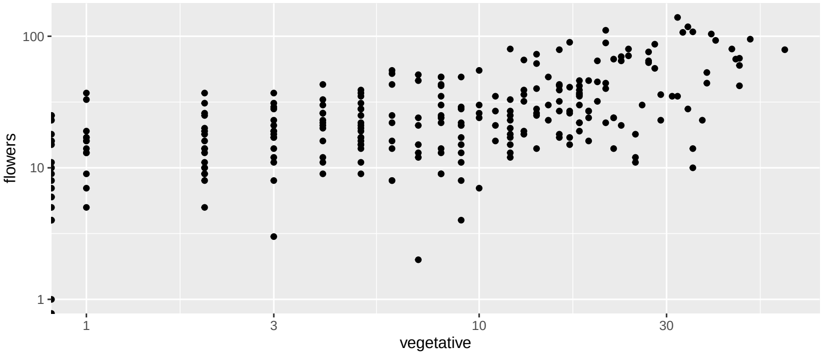


```r
# data in emdbook
```

Log scales

- Right skew
- Counts
- Dimensional data

```
ggplot(Lily_sum, aes(vegetative, flowers)) +
  geom_point() +
  scale_x_log10() +
  scale_y_log10()
```

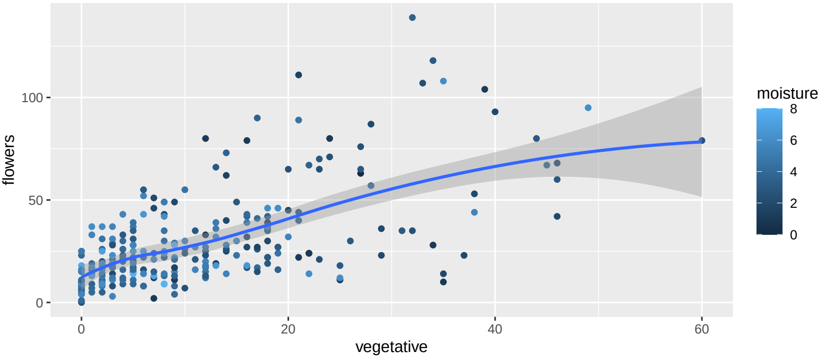


log(0) is undefined

Additional aesthetics

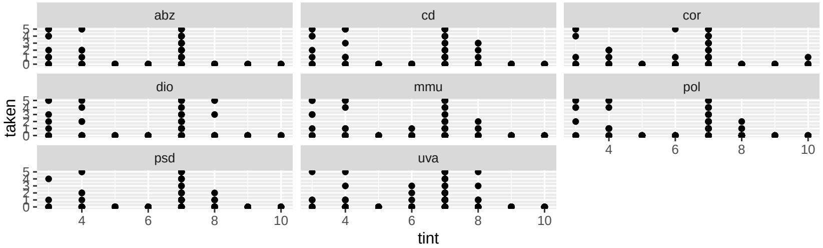
- color
- shape (categorical)
- size (numeric)
- trendlines or other model graphs.

```
ggplot(Lily_sum, aes(vegetative, flowers, color = moisture)) +
  geom_point() +
  geom_smooth()
```

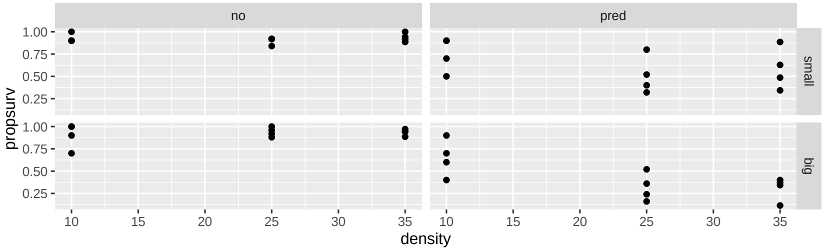


Categorical variables → facets

```
ggplot(SeedPred, aes(tint, taken)) + geom_point() + facet_wrap(~species)
```

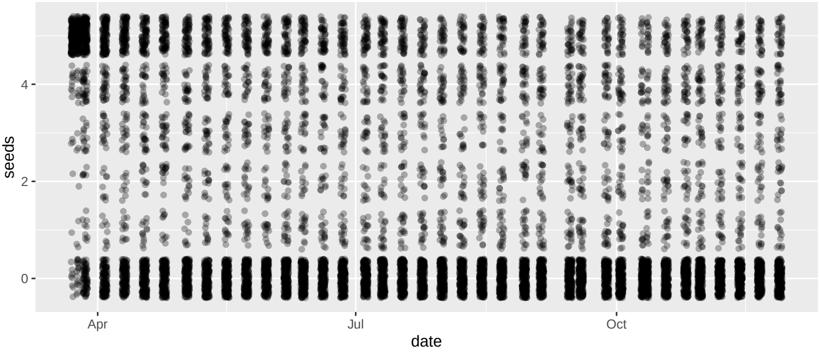


```
ggplot(ReedfrogPred, aes(density, propsurv)) + geom_point() +
  facet_grid(size~pred)
```



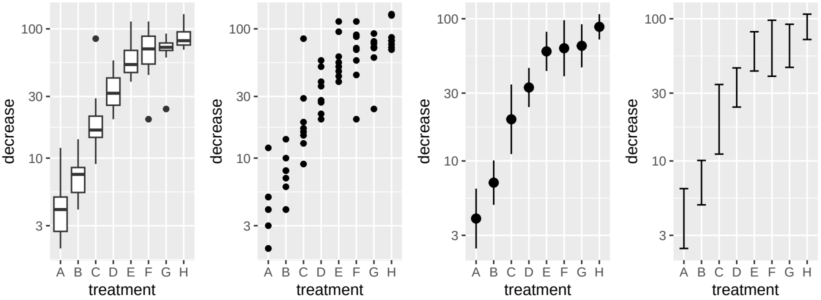
Jittering and transparency

```
ggplot(SeedPred, aes(date, seeds))+
  geom_jitter(alpha= 0.3)
```



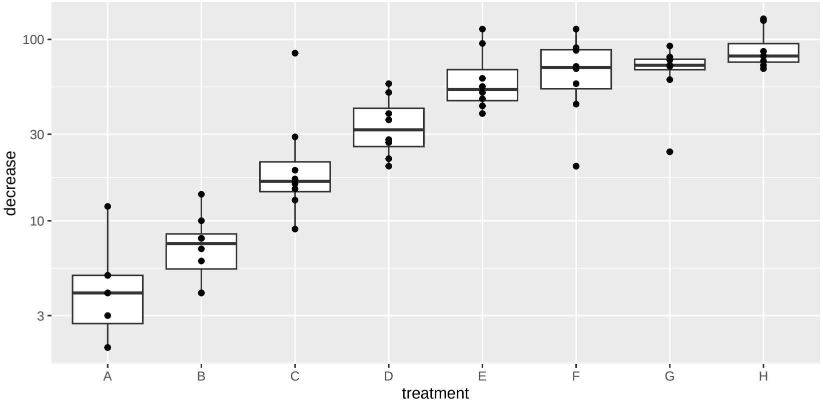
Boxplots - numeric response, categorical predictors

```r
g0 <- ggplot(OrchardSprays,aes(x=treatment,y=decrease)) + # data in MASS
  scale_y_log10()
g_boxplot <- g0 + geom_boxplot()
g_point <- g0 + geom_point()
g_ptrng <- g0 + stat_summary(fun.data=mean_cl_normal,geom="pointrange")
g_errbar <- g0 +
  stat_summary(fun.data=mean_cl_normal,geom="errorbar",width=0.5)
grid.arrange(g_boxplot,g_point,g_ptrng,g_errbar, nrow=1)
```



Combining layers

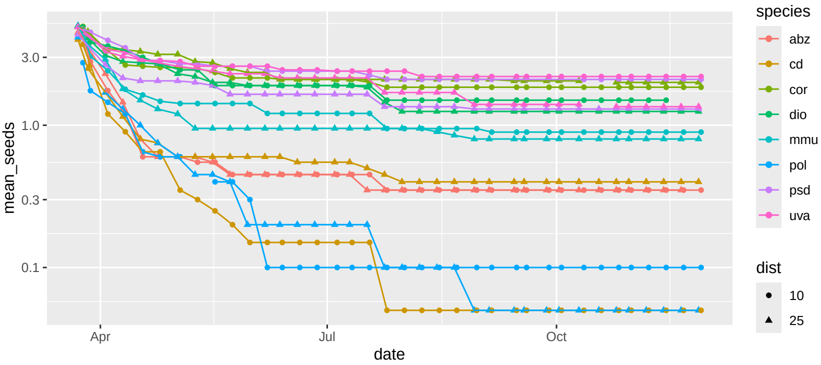
```r
g0 + geom_boxplot() + geom_point()
```



Reshape and Summarize

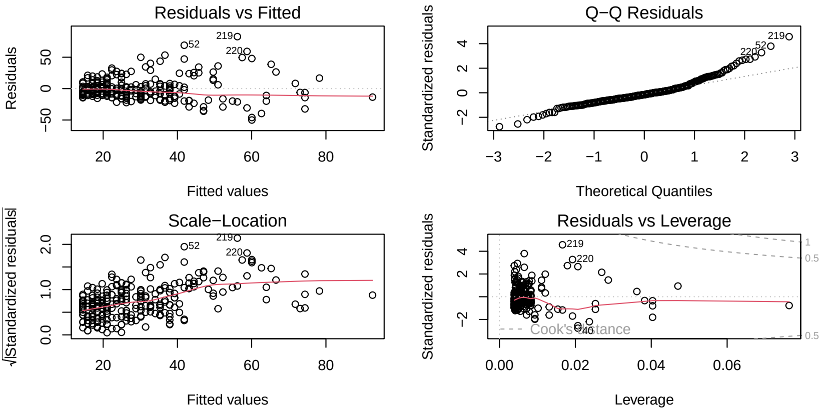
Figure 2.1 from Bolker (2008):

```
daily_avgs <- SeedPred %>%
  group_by(date, species, dist) %>%
  summarise(mean_seeds = mean(seeds))
ggplot(daily_avgs, aes(date, mean_seeds, color=species, shape=dist)) +
  geom_point() + geom_line() + scale_y_log10()
```



Diagnostics, assessment of model validity.

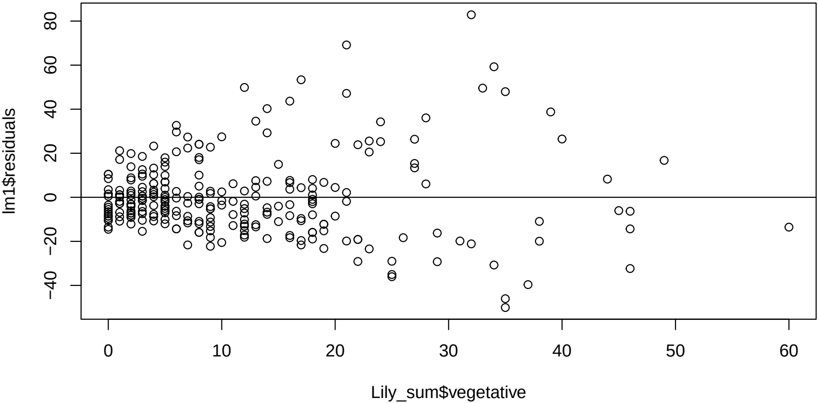
```
lm1 <- lm(flowers~vegetative, data = Lily_sum)
par(mfrow=c(2, 2), mar = c(4, 4, 2, 2)) # see all 4 plots at once
plot(lm1)
```



```
par(mfrow=c(1, 1), mar = c(4, 4, 0.75, 0.5)) # restore graphics parameters
```

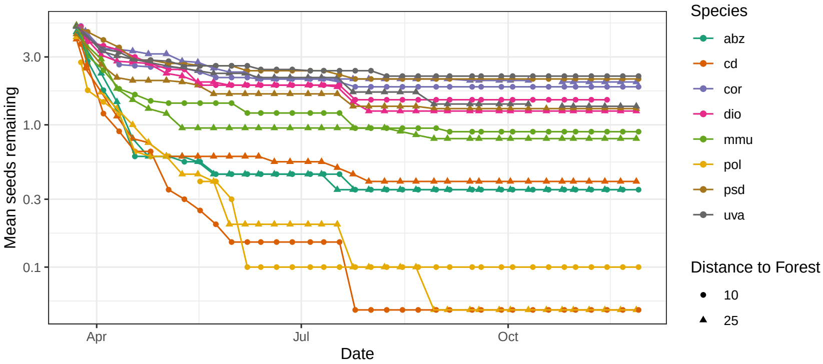
Residuals v. predictors

```
plot(lm1$residuals~Lily_sum$vegetative)
abline(h=0)
```



Fine tune and save graphics for presentation

```
emd2.1<-ggplot(daily_avgs,aes(date,mean_seeds,color=species,shape=dist))+
  geom_point() + geom_line() + scale_y_log10() +
  labs(y="Mean seeds remaining", x = "Date",
       color = "Species", shape = "Distance to Forest") +
  scale_color_brewer(palette = "Dark2") +
  theme_bw()
emd2.1
```



```
ggsave("figures/BolkerFig2.1.tiff", plot=emd2.1,
       width = 10, height = 4, units = "cm", dpi = 800)
```

Opinions on graphical style

Plenty of people with good ideas about style.

- Leland Wilkinson
- Edward Tufte
- William Cleaveland
- Andrew Gelman

Some graph types are controversial. That doesn't mean never use them, but if you do, be aware of the criticisms.

- Pie charts, dynamite plots, dual-axes plots

References

Bolker, Benjamin M. 2008. *Ecological Models and Data in R*. Princeton University Press.