Teaching Stats for Data Science

Danny Kaplan

Macalester College

A bridge is the wrong metaphor.

Bridges and roommates

Bridges

- A narrow path spanning a gulf between two disconnected places.
- Provides an opportunity to cross for those who are willing.
- But you don't have to!

Roommates

- Inhabit the same shared space.
- Generally requires compromise and mutual respect.
- Daily (and unavoidable) exchanges and common activities.

Should we be roomies?

- We’ll have to give up some space.
- We’ll need to adopt some good data habits.
- We’ll have to learn to talk with guests our roommate invites over:
    - large observational datasets
    - frequent interest in causation
    - guiding decision-making rather than rejecting hypothesis.
- Since our rooms are small, a lot of our stuff will be in the living room and kitchen for everyone to use.

Proposal for the move:

Pack up our stuff into ten boxes

Ten stat boxes

1. Data tables (K)
2. Data graphics (K)
3. Model functions (K)
4. Model training (K)
5. Effect size and covariates (LR)
6. Displays of distributions
7. Bootstrap replications
8. Prediction error (LR)
9. Comparing models (LR???)
10. Generalization and causality

K = for kitchen, LR = for living room

- Tidy data: every row is a unit of observation; every column is a variable.
- Meaningful *unit of observation*
- Data tables vs presentations

Not this…

1.1 Migraine and acupuncture. A migraine is a particularly painful type of headache, which patients sometimes wish to treat with acupuncture. To determine whether acupuncture relieves migraine pain, researchers conducted a randomized controlled study where 89 females diagnosed with migraine headaches were randomly assigned to one of two groups: treatment or control. 43 patients in the treatment group received acupuncture that is specifically designed to treat migraines. 46 patients in the control group received placebo acupuncture (needle insertion at nonacupoint locations). 24 hours after patients received acupuncture, they were asked if they were pain free. Results are summarized in the contingency table below.[47]

		Pain free		
		Yes	No	Total
Group	Treatment	10	33	43
	Control	2	44	46
	Total	12	77	89

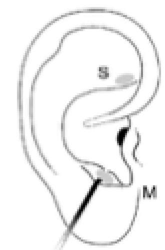


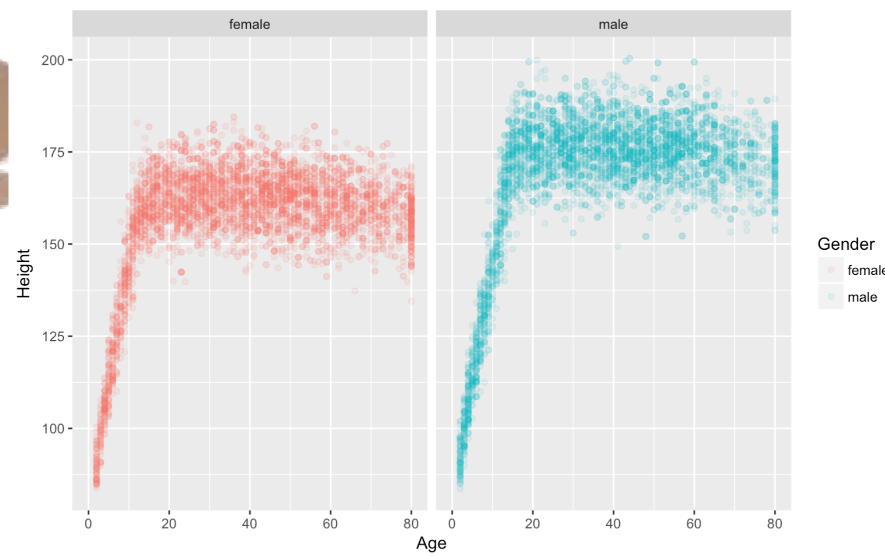
Figure from the original paper displaying the appropriate area (M) versus the inappropriate area (S) used in the treatment of migraine attacks.

Instead, this

patient	accupuncture	pain	date	technician
A2322	control	yes	2014-03-15	Audrey
A2397	treatment	yes	2014-03-17	Audrey
A3213	treatment	no	2014-03-17	Bill
B8732	treatment	no	2014-03-18	Audrey
C6920	control	yes	2014-03-18	Bill
⋮	⋮	⋮	⋮	⋮

- Rich graphics, incl. color, tranparency, faceting, …
- Relationships among multiple variables

```
NHANES %>%
  gf_point(Height ~ Age | Gender, color = ~ Gender, alpha = 0.1)
```

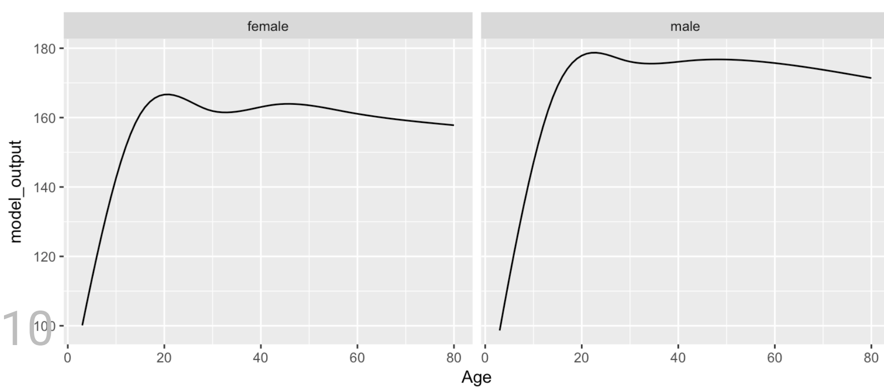


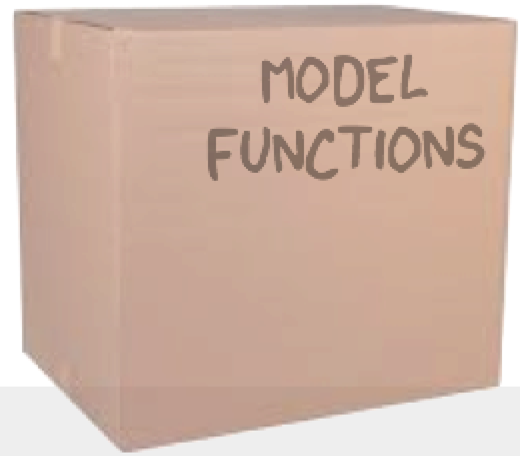
- Inputs and output, explanatory and response variables

```
height(Age = 25, Gender = "female")
```

```
##   Gender Age model_output
## 1 female  25     164.8456
```

```
height(Age = 3:80, Gender = c("female", "male")) %>%
  gf_line(model_output ~ Age | Gender)
```





Model Training: tools for building functions that look like your data

```
hmod1 <- lm(Height ~ Gender * ns(Age, 5), data = NHANES)
```

- Make it about different architectures, e.g. CART, Random Forest, Logistic regression, …
- Both regression models and classifiers
- Encourage nonlinearity (`ns()` == “not straight”?)

```r
wmod1 <- glm(outcome == "Dead" ~ smoker, data = Whickham,
mod_effect(wmod1, ~ smoker, age = c(40, 50, 60))
```

```
##        change smoker to:smoker
## 1 -0.07537604     No       Yes
```

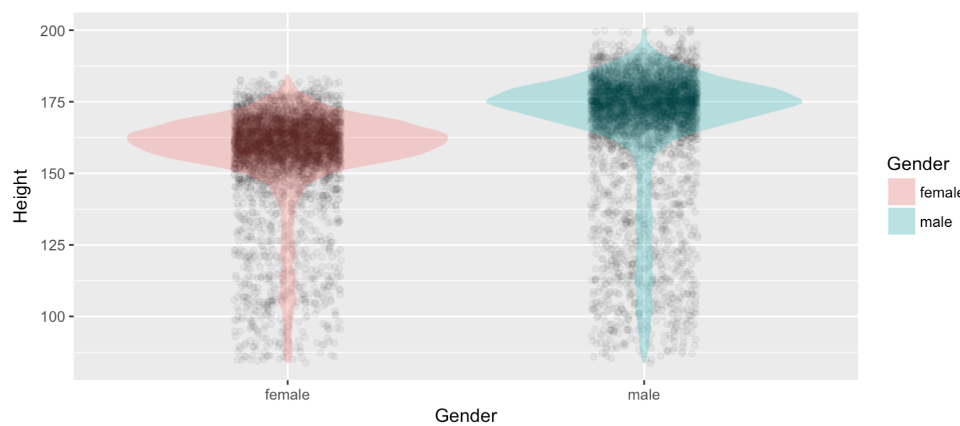
```r
wmod2 <- glm(outcome == "Dead" ~ smoker + age, data = Whickham,
mod_effect(wmod2, ~ smoker, age = c(40, 50, 60))
```

```
##       change smoker to:smoker age
## 1 0.01377155     No       Yes  40
## 2 0.03419996     No       Yes  50
## 3 0.05105680     No       Yes  60
```

```r
NHANES %>% df_stats(Height ~ Gender, coverage(0.95))
```

```
##   Gender  lower upper
## 1 female 102.43 176.1
## 2   male 100.90 190.3
```

```r
NHANES %>%
  gf_jitter(Height ~ Gender, alpha = 0.05, width = 0.15) %>%
  gf_violin(alpha = 0.3, fill = ~ Gender, color = NA)
```



```r
hmod_ensemble <- mod_ensemble(hmod1, nreps = 4)
mod_effect(hmod_ensemble, ~ Age, Age = 5, step = 1,
           Gender = c("male", "female")) %>%
  arrange(Gender)
```

```
##      slope Age to:Age Gender bootstrap_rep
## 1 6.442775   5      6 female             1
## 2 6.394813   5      6 female             2
## 3 6.415638   5      6 female             3
## 4 6.396559   5      6 female             4
## 5 7.254420   5      6   male             1
## 6 7.258021   5      6   male             2
## 7 7.248615   5      6   male             3
## 8 7.233795   5      6   male             4
```

```
mod_eval(hmod1, data = NHANES) %>%
  df_stats( ~ I((model_output - Height)^2), mean)
```

```
## [1] 52.14731
```

Or make it a fundamental operation.

```
mod_error(hmod1, testdata = NHANES)
```

```
## [1] 52.14731
```

Let's try another model that's more flexile

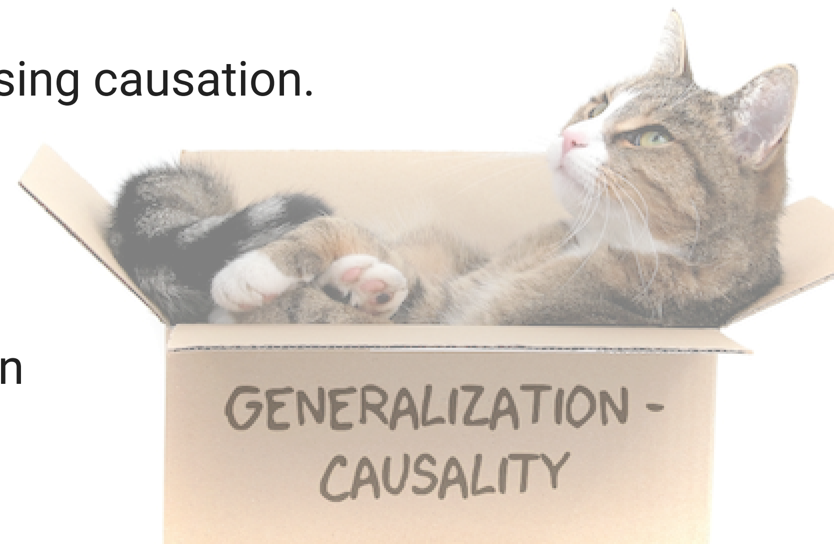
```
hmod2 <- lm(Height ~ Gender * ns(Age, 25), data = NHANES)
```

How does it compare to the original?

```
mod_cv(hmod1, hmod2, ntrials = 50) %>%
  df_stats(mse ~ model, coverage(0.95))
```

```
##   model    lower    upper
## 1 hmod1 52.21514 52.30473
## 2 hmod2 49.72174 49.86336
```

- It’s still important to talk about how to collect meaningful data to form conclusions that generalize outside the sample at hand.
- Many data-science applications involve reasoning about causal influences.
  - We need to come down from the mathematical high horse of “no causation without experimentation.”
  - Recognize responsible methods for addressing causation.



e.g. the Judea Pearl award in causality education

… and of course