

Jeffrey Leek, Assistant Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

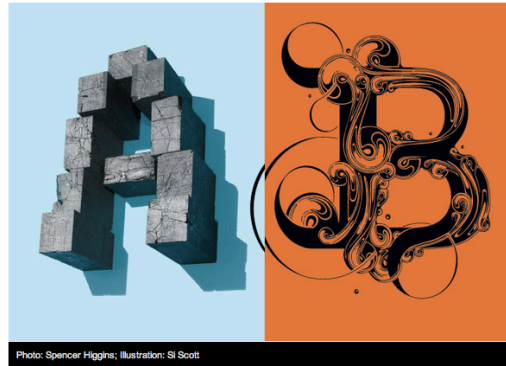
Key ideas

- Outcome is still quantitative
- You have multiple explanatory variables
- Goal is to identify contributions of different variables

A successful example

The A/B Test: Inside the Technology That's Changing the Rules of Business

BY BRIAN CHRISTIAN 04.25.12 8:47 PM



"For the button, an A/B test of three new word choices—"Learn More," "Join Us Now," and "Sign Up Now"—revealed that "Learn More" garnered 18.6 percent more signups per visitor than the default of "Sign Up." Similarly, a black-and-white photo of the Obama family outperformed the default turquoise image by 13.1 percent. Using both the family image and "Learn More," signups increased by a thundering 40 percent."

http://www.wired.com/business/2012/04/ff_abtesting/

Movie Data

```
download.file("http://www.rossmanchance.com/iscam2/data/movies03RT.txt",
              destfile="./data/movies.txt")
movies <- read.table("./data/movies.txt", sep="\t", header=T, quote="")
head(movies)
```

			X	score	rating	genre	box.office	running.time
1	2	Fast 2 Furious	48.9	PG-13	action/adventure	127.15	107	
2		28 Days Later	78.2	R	horror	45.06	113	
3		A Guy Thing	39.5	PG-13	rom comedy	15.54	101	
4		A Man Apart	42.9	R	action/adventure	26.25	110	
5		A Mighty Wind	79.9	PG-13	comedy	17.78	91	
6		Agent Cody Banks	57.9	PG	action/adventure	47.81	102	

<http://www.rossmanchance.com/>

Relating score to rating

$$S_i = b_0 + b_1 \mathbb{1}(Ra_i = "PG") + b_2 \mathbb{1}(Ra_i = "PG-13") + b_3 \mathbb{1}(Ra_i = "R") + e_i$$

The notation $\mathbb{1}(Ra_i = "PG")$ is a logical value that is one if the movie rating is "PG" and zero otherwise.

Average values

b_0 = average of the G movies

$b_0 + b_1$ = average of the PG movies

$b_0 + b_2$ = average of the PG-13 movies

$b_0 + b_3$ = average of the R movies

ANOVA in R

```
aovObject <- aov(movies$score ~ movies$rating)
aovObject
```

Call:

```
aov(formula = movies$score ~ movies$rating)
```

Terms:

	movies\$rating	Residuals
Sum of Squares	570	28149
Deg. of Freedom	3	136

Residual standard error: 14.39

Estimated effects may be unbalanced

ANOVA in R

```
aovObject$coeff
```

(Intercept)	movies\$ratingPG	movies\$ratingPG-13	movies\$ratingR
67.65	-12.59	-11.81	-12.02

Adding a second factor

$$S_i = b_0 + b_1 \mathbb{1}(Ra_i = "PG") + b_2 \mathbb{1}(Ra_i = "PG - 13") + b_3 \mathbb{1}(Ra_i = "R") \\ + \gamma_1 \mathbb{1}(G_i = "action") + \gamma_2 \mathbb{1}(G_i = "animated") + \dots + e_i$$

The notation $\mathbb{1}(Ra_i = "PG")$ is a logical value that is one if the movie rating is "PG" and zero otherwise.

Adding a second factor

$$S_i = b_0 + \underbrace{b_1 \mathbb{1}(Ra_i = " PG ") + b_2 \mathbb{1}(Ra_i = " PG - 13 ") + b_3 \mathbb{1}(Ra_i = " R ")}_{rating} + \underbrace{\gamma_1 \mathbb{1}(G_i = " action ") + \gamma_2 \mathbb{1}(G_i = " animated ") + \dots}_{genre} + e_i$$

There are only 2 variables in this model. They have multiple levels.

Second variable

```
aovObject2 <- aov(movies$score ~ movies$rating + movies$genre)
aovObject2
```

Call:

```
aov(formula = movies$score ~ movies$rating + movies$genre)
```

Terms:

	movies\$rating	movies\$genre	Residuals
Sum of Squares	570	3935	24214
Deg. of Freedom	3	12	124

Residual standard error: 13.97

Estimated effects may be unbalanced

ANOVA Summary

```
summary(aovObject2)
```

```

              Df Sum Sq Mean Sq F value Pr(>F)
movies$rating   3    570     190    0.97  0.408
movies$genre   12   3935     328    1.68  0.079 .
Residuals    124  24214     195
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Order matters

```
aovObject3 <- aov(movies$score ~ movies$genre + movies$rating)
summary(aovObject3)
```

```
              Df Sum Sq Mean Sq F value Pr(>F)
movies$genre   12   4222     352    1.80  0.055 .
movies$rating   3    284      95    0.48  0.694
Residuals    124  24214     195
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aovObject2)
```

```
              Df Sum Sq Mean Sq F value Pr(>F)
movies$rating   3    570     190    0.97  0.408
movies$genre   12   3935     328    1.68  0.079 .
Residuals    124  24214     195
---

```

12/15

Adding a quantitative variable

$$\begin{aligned}
 S_i = & b_0 + \underbrace{b_1 \mathbb{1}(Ra_i = " PG ") + b_2 \mathbb{1}(Ra_i = " PG - 13 ") + b_3 \mathbb{1}(Ra_i = " R ")}_{rating} \\
 & + \underbrace{\gamma_1 \mathbb{1}(G_i = " action ") + \gamma_2 \mathbb{1}(G_i = " animated ") + \dots}_{genre} + \eta_1 \underbrace{BO_i}_{box\ office} + e_i
 \end{aligned}$$

There are three variables in this model - box office is quantitative so only has one term.

ANOVA with quantitative variable in R

```
aovObject4 <- aov(movies$score ~ movies$genre + movies$rating + movies$box.office)
summary(aovObject4)
```

```

              Df Sum Sq Mean Sq F value    Pr(>F)
movies$genre    12   4222     352    2.19   0.016 *
movies$rating     3    284      95    0.59   0.624
movies$box.office  1   4421   4421   27.47 6.7e-07 ***
Residuals      123  19793     161
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Language and further resources

- Units - one observation
- Treatments - applied to units
- Factors - controlled by experimenters
- Replicates - multiple (independent) units with the same factors/treatments
- [Wikipedia on Experimental Design](#)
- [Wikipedia on ANOVA](#)
- [Wikipedia on A/B Testing](#)