

Data Exploration and Analysis: Three Algorithms

Marco Morales
mam2519@columbia.edu

GR5069
Topics in Applied Data Science
for Social Scientists
Spring 2017
Columbia University

Housekeeping

- ▶ Today:
 - ▶ three algorithms
 - ▶ Weekly progress report
 - ▶ Second data challenge
- ▶ next week:
 - ▶ your third progress report
 - ▶ Data challenge due at 6PM

Three Algorithms

Algorithm 1: OLS

- ▶ what is a regression?

$$E[y|x] = f(x)$$

- ▶ where $f(x)$ is a conditional mean function, such that

$$y = E[y|x] + \epsilon$$

- ▶ empirically: what do we get from a regression?

Three Algorithms

Algorithm 1: OLS

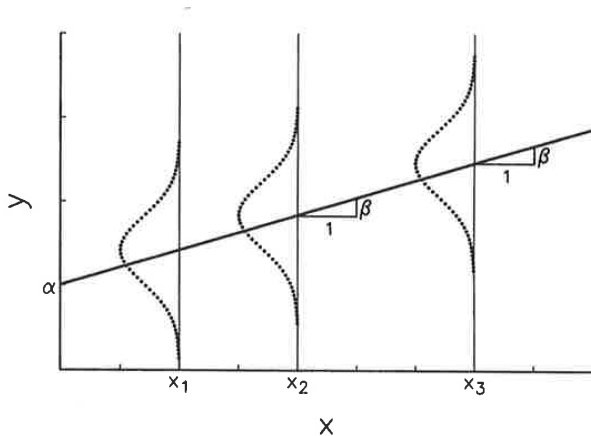


Figure 2.1. Simple Linear Regression Model With the Distribution of y Given x

Figure: Long (1997)

Three Algorithms

Algorithm 1: OLS

Call:

```
lm(formula = organized.crime.dead ~ organized.crime.wounded +  
    afi + army + navy + federal.police + long.guns.seized + small.arms.seized +  
    clips.seized + cartridge.seized, data = AllData)
```

Residuals:

Min	1Q	Median	3Q	Max
-11.6058	-0.7274	-0.4506	0.2192	27.3262

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.4505553	0.0332307	13.558	< 2e-16 ***
organized.crime.wounded	0.3736900	0.0239171	15.624	< 2e-16 ***
afi	-0.2261752	0.4210396	-0.537	0.5912
army	0.3066898	0.0532594	5.758	8.96e-09 ***
navy	0.7150402	0.1389449	5.146	2.75e-07 ***
federal.police	-0.1271515	0.0773309	-1.644	0.1002
long.guns.seized	0.1478424	0.0085972	17.197	< 2e-16 ***
small.arms.seized	-0.0437447	0.0184592	-2.370	0.0178 *
clips.seized	0.0004374	0.0003152	1.388	0.1653
cartridge.seized	-0.0001690	0.0000193	-8.760	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.731 on 5386 degrees of freedom

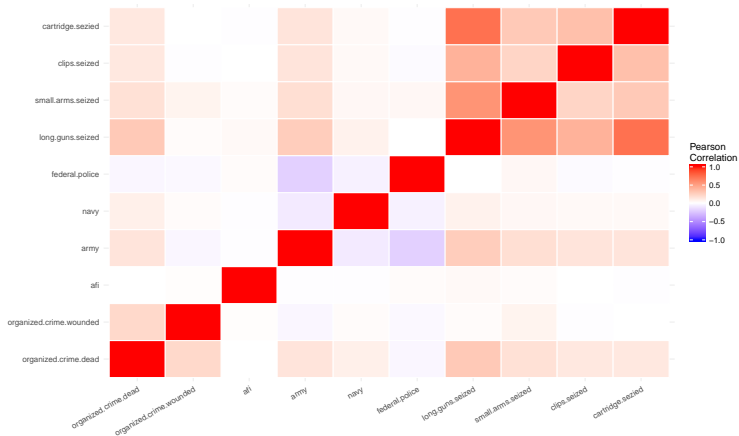
Multiple R-squared: 0.1413, Adjusted R-squared: 0.1398

F-statistic: 98.44 on 9 and 5386 DF, p-value: < 2.2e-16

Three Algorithms

Algorithm 1: OLS

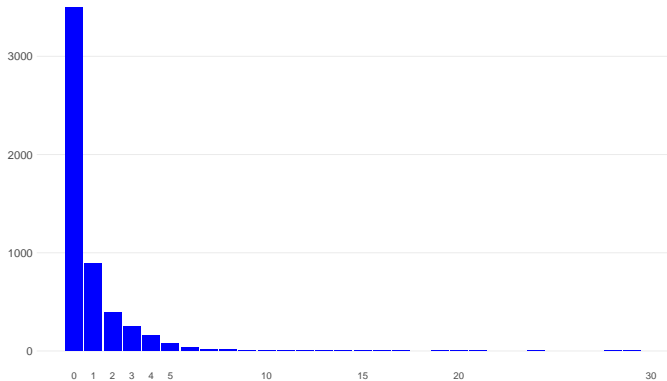
- ▶ are these "real" results, or just a mirage from reiterated information in our variables?



Three Algorithms

Algorithm 1: OLS

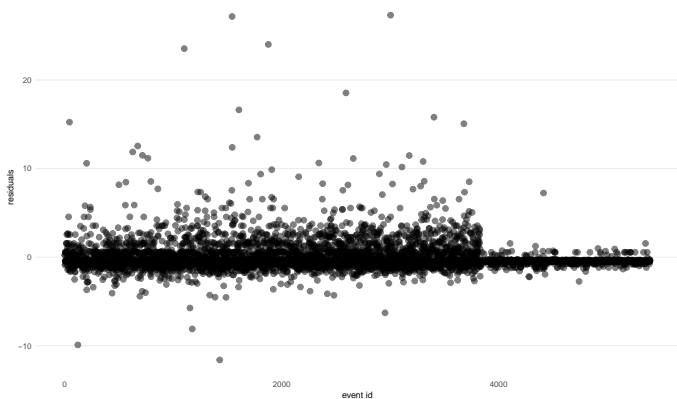
- ▶ but wait... what does my DV look like?



Three Algorithms

Algorithm 1: OLS

- ▶ what's the problem with this?



Algorithm 1: OLS

Number of non-zero elements in x	Number of vectors
0	~3300
1	~900
2	~400
3	~250
4	~150
5	~100
6	~50
7	~30
8	~20
9	~10
10	~5

Three Algorithms

a quick detour: conditional effects

- ▶ assume for a moment that this is not problematic
 - ▶ (it is! but assume...)
- ▶ when analyzing people and behaviors, we're not only concerned about levels
 - ▶ we typically care about behaviors *conditional* on something else happening
 - ▶ note that this is different from "holding the rest constant"
 - ▶ these can be easily computed through **multiplicative interactions**

Three Algorithms

a quick detour: conditional effects

- ▶ from the model (with multiplicative interactions)

$$Y = \beta_0 + \beta_X \mathbf{X} + \beta_Z \mathbf{Z} + \beta_{XZ} \mathbf{XZ} + \epsilon$$

- ▶ we'd be interested in the marginal effect of Z given X on Y

$$\frac{\partial E[Y|X, Z]}{\partial \mathbf{X}} = \beta_X + \beta_{XZ} \mathbf{Z}$$

Three Algorithms

a quick detour: conditional effects

- ▶ going back to our example:
 - ▶ **are there more expected deaths when combat is heavier?**
 - ▶ let's look at the case of events where the Navy is involved
 - ▶ we'd need to assume that more seized heavy weapons indicate heavier combat and compute

$$\beta_{navy} + \beta_{navy, long.guns.seized} * long.guns.seized$$

- ▶ **are there less expected number of deaths when no weapons are seized?**
 - ▶ let's look at the case of the Army
 - ▶ we maintain the same assumption and compute

$$\beta_{army}$$

Three Algorithms

a quick detour: conditional effects

- ▶ but in addition to the marginal effect

$$\frac{\partial E[Y|X, Z]}{\partial \mathbf{X}} = \beta_X + \beta_{XZ}\mathbf{Z}$$

- ▶ we also need to compute appropriate standard errors

$$\text{Var}\left(\frac{\partial \hat{E}[Y|X, Z]}{\partial \mathbf{X}}\right) = \text{Var}[\hat{\beta}_X] + \mathbf{Z}^2 \text{Var}[\hat{\beta}_{XZ}] + 2\mathbf{Z} \text{Cov}[\hat{\beta}_X, \hat{\beta}_{XZ}]$$

Three Algorithms

a quick detour: conditional effects

Call:

```
lm(formula = organized.crime.dead ~ organized.crime.wounded +  
  afi * long.guns.seized + army * long.guns.seized + navy *  
  long.guns.seized + federal.police * long.guns.seized + afi *  
  cartridge.seized + army * cartridge.seized + navy * cartridge.seized +  
  federal.police * cartridge.seized + small.arms.seized + clips.seized,  
  data = AllData)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.6509	-0.7385	-0.4189	0.1933	27.2187

Residual standard error: 1.714 on 5378 degrees of freedom

Multiple R-squared: 0.1587, Adjusted R-squared: 0.156

F-statistic: 59.67 on 17 and 5378 DF, p-value: < 2.2e-16

Three Algorithms

a quick detour: conditional effects

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.4188645	0.0336777	12.437	< 2e-16	***
organized.crime.wounded	0.3624050	0.0237796	15.240	< 2e-16	***
afi	-0.0419271	0.5040535	-0.083	0.9337	
long.guns.seized	0.1713811	0.0172327	9.945	< 2e-16	***
army	0.4244453	0.0556353	7.629	2.78e-14	***
navy	0.2772627	0.1567621	1.769	0.0770	.
federal.police	-0.1113463	0.0801781	-1.389	0.1650	
cartridge.seized	0.0002292	0.0000968	2.368	0.0179	*
small.arms.seized	-0.0452969	0.0186014	-2.435	0.0149	*
clips.seized	0.0003127	0.0003146	0.994	0.3202	
afi:long.guns.seized	0.0229013	0.0784035	0.292	0.7702	
long.guns.seized:army	-0.0459567	0.0181403	-2.533	0.0113	*
long.guns.seized:navy	0.1761160	0.0421782	4.176	3.02e-05	***
long.guns.seized:federal.police	-0.0253811	0.0190541	-1.332	0.1829	
afi:cartridge.seized	-0.0050516	0.0031231	-1.617	0.1058	
army:cartridge.seized	-0.0003911	0.0000981	-3.987	6.78e-05	***
navy:cartridge.seized	-0.0006909	0.0001728	-3.998	6.47e-05	***
federal.police:cartridge.seized	-0.0001518	0.0001102	-1.377	0.1685	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Three Algorithms

a quick detour: conditional effects

- ▶ marginal effect of 5 seized long guns on the expected number of dead on events that involve the Navy

$$1.16$$
$$[-1.17, 3.48]$$

- ▶ marginal effect on the expected number of dead of events that involve the Army when no long guns (zero) are seized

$$0.42$$
$$[0.32, 0.53]$$

Three Algorithms

a quick detour: conditional effects

- ▶ Always, always, always remember (Brambor et al. 2006):
 1. Use multiplicative interaction models **whenever one's hypothesis is conditional** in nature.
 2. Include **all constitutive terms** in the model specification.
 3. **Do not interpret the coefficients on constitutive terms as if they are unconditional marginal effects.**
 4. Do not forget to **calculate substantively meaningful marginal effects and standard errors.**
- ▶ ... or face the wrath of econometricians

Three Algorithms

Algorithm 2: logistic regression

- ▶ different question: **did something happen or not?**
 - ▶ essentially, binary outcome classification
 - ▶ why not just use OLS?
- ▶ one way to think about this: let y^* be a continuous (latent) variable

$$y^* = x\beta + \epsilon$$

- ▶ for which we only observe two outcomes

$$y_i = \begin{cases} 1 & \text{if } y_i^* > \tau \\ 0 & \text{if } y_i^* \leq \tau \end{cases}$$

Three Algorithms

Algorithm 2: logistic regression

- ▶ we're interested in the probability that $y = 1$

$$\pi_i = \text{Pr}(y = 1) = F(\beta x)$$

- ▶ in the case of a logit, we estimate

$$\pi_i = \Lambda(\beta x) = \frac{e^{\beta x}}{1 + e^{\beta x}}$$

- ▶ but there's also additional "flavors" (i.e. probit)

Three Algorithms

Algorithm 2: logistic regression

- ▶ going back to our example:
 - ▶ we have a natural dual category: **events with deaths / no deaths**
 - ▶ **could we learn something about correlates to events with organized crime deaths?**
 - ▶ we have information on federal forces involved
 - ▶ also on materiel seizures
 - ▶ **can this relationship ever be causal?**

Three Algorithms

Algorithm 2: logistic regression

Call:

```
glm(formula = organized.crime.death ~ organized.crime.wounded +  
    afi + army + navy + federal.police + long.guns.seized + small.arms.seized +  
    clips.seized + cartridge.seized, family = binomial(link = "logit"),  
    data = AllData)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-4.5396	-0.6657	-0.4731	-0.4592	2.7612

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.1337831	0.0599578	-35.588	< 2e-16 ***
organized.crime.wounded	0.2839835	0.0376519	7.542	4.62e-14 ***
afi	-0.6960636	0.7234004	-0.962	0.336
army	0.7395036	0.0812191	9.105	< 2e-16 ***
navy	0.9292565	0.1827726	5.084	3.69e-07 ***
federal.police	-0.0628413	0.1331772	-0.472	0.637
long.guns.seized	0.1544432	0.0141145	10.942	< 2e-16 ***
small.arms.seized	-0.0137429	0.0271923	-0.505	0.613
clips.seized	-0.0004430	0.0004284	-1.034	0.301
cartridge.seized	-0.0002413	0.0000510	-4.730	2.25e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

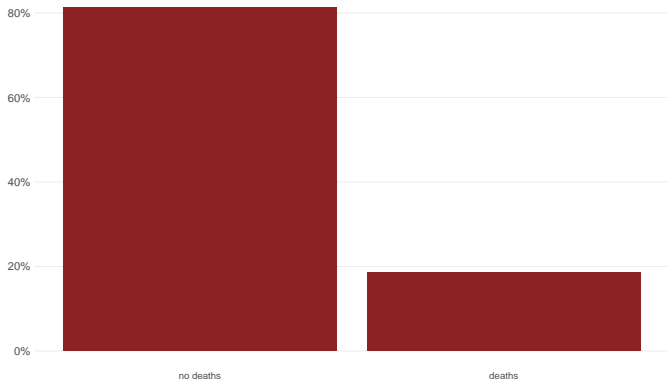
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5185.2 on 5395 degrees of freedom
Residual deviance: 4721.3 on 5386 degrees of freedom
AIC: 4741.3

Three Algorithms

Algorithm 2: logistic regression

- ▶ but wait again, what does my DV look like?



- ▶ what does your "plain vanilla" logistic regression assume?

Weekly Progress Review

Data Exploration and Analysis: Three Algorithms

Marco Morales
mam2519@columbia.edu

GR5069
Topics in Applied Data Science
for Social Scientists
Spring 2017
Columbia University