

Stat 442 - Oct. 23. 16

Review - use of constructed variables

- interaction type terms

→ upto 1985 - $\begin{cases} 0 & \text{year} > 1985 \\ \text{year} & \text{year} \leq 1985 \end{cases}$

↳ (year ≤ 1985) × year

dummy

→ ~~year~~ after 1985 - $\begin{cases} 0 & \text{year} \leq 1985 \\ \text{year} & \text{year} > 1985 \end{cases}$

= (year > 1985) × year

= year - upto 1985

Two points → interpretation of interaction terms depends on all terms in model

with both upto 1985 & after 1985

β_0 → slope upto 1985 slope after 1985

2nd → best way to interpret is to look at predictions

- effects library in R.

↳ complicated
model

Applications of Model

- Tests of hypotheses
in terms of β 's

- Problems with collinearity

- collinear occurs
when two X 's are highly
correlated

⊙

- when any of X 's can be
predicted ~~to~~ reasonably well
with linear combination of
other X 's

- i.e. if regress X_j on
remainder
- get high R^2

Consequences

- $\hat{\beta}$'s will tend to be
"unreliable" - large standard
errors.

Artificial collinearity example

shark data
fuzzy Year
= year + "noise"
↓
N(0,σ²)
Cor > .8

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-75.7705	8.6650	-8.74	<2e-16 ***
year	0.0206	0.1636	0.13	0.90
fuzzyYear	0.0106	0.1633	0.06	0.95
Null deviance:	176.93	on 53	degrees of freedom	
Residual deviance:	119.11	on 51	degrees of freedom	

Model output for the segmented Model

```
glm(formula = attacks ~ offset(log(population)) + (year > 1985) + upTo1985 + after1985, family = poisson)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-5.102e+01	1.590e+01	-3.208	0.00134 **
year > 1985TRUE	-1.232e+02	3.838e+01	-3.210	0.00133 **
upTo1985	1.861e-02	8.063e-03	2.309	0.02096 *
after1985	8.054e-02	1.752e-02	4.598	4.27e-06 ***

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

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 176.93 on 53 degrees of freedom
Residual deviance: 108.59 on 50 degrees of freedom
AIC: 282.24

→ n - # param estimate
4 param n=54

Model output for the Combined Model

```
glm(formula = attacks ~ offset(log(population)) + (year > 1985) + upTo1985 + after1985 + poly(year, 3), family = poisson)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.317	-1.083	-0.457	0.650	4.018

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-39.6273	34.4355	-1.15	0.25
year > 1985TRUE	-209.2984	187.8747	-1.11	0.27
upTo1985	0.0128	0.0175	0.73	0.47
after1985	0.1181	0.0794	1.49	0.14
poly(year, 3)1	NA	NA	NA	NA
poly(year, 3)2	-1.8564	1.6333	-1.14	0.26
poly(year, 3)3	0.2618	1.1200	0.23	0.82

→ = year

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 176.93 on 53 degrees of freedom
Residual deviance: 105.91 on 48 degrees of freedom
AIC: 283.6

↳ df: 48 → 6 param in model

eg. fuzzy year
→ neither are significant
→ both are significant without
other in model

While can't identify which
variable is significant, "after
controlling for the other"

can test to see if any of
two is significant by testing

$H: \text{all } \beta\text{'s} = 0$ - null model
↳ composite hypothesis
↳ global hypothesis

linear case - overall F-test for
model

- generalized linear models

- can test using
likelihood ratio test

$$LRT = -2 \left[\log \left(\frac{\text{lik}(\text{Null model})}{\text{lik}(\text{Full model})} \right) \right]$$

$\log(\mu) = \eta = \beta$

- note that in R. glmer → reports
Deviance

Linear Regression

test H_0 : all $\beta_j = 0$

$$F = \frac{(\text{Total SS}_y - \text{Resid SS}_y / 9)}{(\text{Resid SS}_y / \text{df. error})}$$

Generalized Linear model

LRT = difference in deviances

↳ No error term in denominator
- don't divide by d.f.

$$\text{LRT} = 126.9 - 119.1 = 57.8$$

refer to χ^2

- $p < .0001$

→ at least one of ~~β_j~~
 X 's is predictive

since in separate regressions
year & fuzzy year are
signif → having both is
redundant

- recall deviance = $-2 \left[\log(\text{Lik}(M)) - \log(\text{Lik}(\text{saturated model})) \right]$

M = any model of interest

- $\log(\text{Lik}(\text{saturated model}))$

← saturated model - parameter for each case
 $\rightarrow \hat{y} = y \cdot 0$ - perfect, non-informative

LRT = Deviance (Null Model) ✓

↓
 H_0 : all β 's = 0 - Deviance (Full Model)

under H_0 • LRT $\sim \chi^2(q)$

↓
 approximately
 q = # β 's setting to 0

→ LRT : → deviances are generalization of notion of residual SS_q

Null deviance → Total SS_q

Residual deviance → Resid. SS_q

- no evidence that having both improve prediction relative to just using one

- to choose which one,

use AIC \rightarrow bigger for fuzzy-year model.

In general LRT can be used to compare a subset model to bigger model

$M_r \subset M_F \rightarrow$
 \downarrow reduced \leftarrow Full

$M = \{X_i\}$ in models

More precisely if any X in M_r can be produced as linear combination of X_i in M_F

M_r is a subset model of M_F

so shark $M_r = \mathcal{N} = \beta_0 + \beta_1 \text{year}$

$M_F = \mathcal{N} = \beta_0' + \beta_1' (\text{Year} < 1995) + \beta_2' \text{upto } 1985 +$

+ β 's after 1985

$M_r \leq M_F$ since year
= upto 1985
+ after 1985

→ Test: H_0 : M_r is sufficient
in relation to
 M_F

in case when X 's in M_r
are actual subset
of X 's in M_F

ie: M_r is result of dropping
 X 's from M_F

LRT: $Dev(M_r) - Dev(M_F)$

follow (approximate)

$\chi^2_{(q)}$

$q = \#$ variables
dropped

= β 's set to 0 in
 M_F

H_0 : β 's = 0 (corresp. to
 X 's not in M_r)

easy example - artificial

shark example

- two models - segmented
- Polynomial fit

how about combining terms

in a "super" model - M_F

- is there evidence M_F has additional predictive power,

→ do test of signif

two choices of M_r → segm.
→ poly.

Dev(M_r) = ~~175.7~~ 108.86

Dev(M_F) = 105.94

LRT = 3.3 2.7

want $p = \text{Prob}(\chi_{(df)}^2) > 3.3$

df = 3 ?? df = 2 just

$p \hat{=} .05$??

$p > .1$

subtract separate d.f. for resid. dev.

So indication via AIC:

- bigger model is worse in context of given sample size
- might be better with suff. data.