Week 5

ANOVA with multiple variables / factors

outcome is still quantative

you have multiple explanatory variables

goal:
to identify contributions of different variables

A/B test - Obama Campaing

Relating score to rating (sinemas)

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

$\mathbb{1}(Ra_i = \text{"PG"})$ - logical value ($Ra_i == \text{"PG"}$)
„indicator factors"

average values:

b_0 = avg for G

$b_0 + b_1$ - avg for PG

$b_0 + b_2$ - avg for PG-13

$b_0 + b_3$ - avg for R

ANOVA in R

aov
instead of lm

aovObj = aov(movies$score ~ movies$rating)

aovObj$coeff

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

avg for G

Adding a second factor

$$S_i = \underbrace{b_0 + b_1 \mathbb{I}(Ra_i = \text{"PG"}) + b_2 \mathbb{I}(Ra_i = \text{"PG-13"}) + \dots}_{\text{rating}} +$$

$$+ y_1 \mathbb{I}(G_i = \text{"action"}) + \underbrace{y_2 \mathbb{I}(G_i = \text{"animated"})}_{\text{lots of genres}} + \dots + e$$

genre

only 2 variables!

aov2 = aov(score ~ rating + genre)

summary(aov2)

F value

↓

variation explained
by factor

for 1st (rating) –
only by itself

for 2nd (genre) –
only what wasn't
explained by rating
(after taking into account
rating)

order matters!

⇓

aov2 = aov(score ~ genre + rating) ← better

- you can add quantitative variables

→ so we may add box office

aov3 = aov(score ~ genre + rating + box office)

↑

only 1 degree of freedom

Language:

unit - one observation

treatment - applied to units

factors - controlled by experementers

replicates - multiple (indipendent) units with the same factors / treatments

Useful resourses

(wiki pages: Experimental Design
ANOVA,
A/B Testing)

Binary outcomes

like:

- Dead / Alive
- Win / Loss
- Success / Failure
- etc

binary outcomes
or
0/1 outcomes

Linear regression - may be not the best

Ravens data

win num   win (W/L)   score   opponent. score

Linear model:

$$RW_i = b_0 + b_1 \cdot S_i + e_i$$

RW_i 1 if W, 0 if not

S_i - number of points they scored

b_0 - probability of a Ravens win with (0 score)

b_1 - increase in probability by each 1 point

e_i - error

lm(win num ~ raven. score)

not good

(sometimes $Pr > 1$ which is impossible)

etc

Binary Outcomes: RW_i
0/1

Probability (0,1)

$$Pr(RW_i \mid RS_i, b_0, b_1)$$

Odds $(0, \infty)$

$$\frac{Pr(RW_i \mid RS_i, b_0, b_1)}{1 - Pr(RW_i \mid RS_i, b_0, b_1)}$$

Log odds - $(-\infty, \infty)$ log (odd)

Linear vs Logistic Regression

Linear

Logistic

instead of modelling RW_i

we model the probability based on odds

$$\Pr(RW_i \mid RS_i, b_0, b_1) = \frac{e^{b_0 + b_1 RS_i}}{1 + e^{b_0 + b_1 RS_i}}$$

or

$$\underbrace{\log\left(\frac{\Pr(RW_i \mid RS_i, b_0, b_1)}{1 - \Pr(RW_i \mid RS_i, b_0, b_1)}\right)}_{\text{log odds}} = \underbrace{b_0 + b_1 RS_i}_{\text{Ravens score}}$$

any number

$(-\infty; +\infty)$

not $[0, 1]$

interpretation

b_0 - log odds of win if they score 0 points

b_1 - log odds ratio of win probability for each point scored (compared to 0 points)

$\exp(b_1)$ - odds ratio of win probability for each point scored (compared to 0 points) in R:

glm command

$$\text{logRegRavens} = \text{glm}(\underbrace{\text{winNum}}_{\text{outcome}} \sim \underbrace{\text{score}}_{\text{covariants}}, \text{family} = \text{"binomial"})$$

logistic regression

coefficients are interpreted differently

it gets log!

exp(logRegRavens$coeff)

↳ should see if
score is bigger than 1
(more chances to win)

exp(confint(logRegRavens))

for confidence intervals

Anova for logistic regression

anova(logRegRavens, test = "Chisq")

analysis of Deviance Table

Simpson's Paradox - take a look at wiki

Interpreting Odds Ratios

- not probabilities
- odds ratio of 1 = no difference in odds
- log odds ratio of 0 = no difference in odds
  - 0.5 < odds ratio < 2 - „moderate effect"
- Relative Risk $\dfrac{Pr(RW_i \mid RS_i = 10)}{Pr(RW_i \mid RS_i = 0)}$

  often easier to interpret

wiki

$\boxed{\text{Odds Ratio}}$

- for small probabilities $RR \approx OR$, but they are not the same

Count Outcomes

- many data take form of counts
    - calls to a call center
    - number of flu cases in area
    - numbers of cars that cross a bridge

- data may also be in the form of rates
    - percent of students passing ~~the~~ a test
    - percent of hits to a website from a country

Linear regression is an option

Poisson distribution can be used to model this data

```
set.seed(3433)
~~pois~~
rpois(100, lambda = 100)
```

↑ rate

(an avg number of calls coming to a call center)

Spread is bigger for bigger lambdas

and it controls both mean and var

mean(poisd) ≈ var(poisd)

very close

Web site traffic:

possible to fit regression here

$$NH_i = b_0 + b_1 JD_i + e_i$$

NH_i - number of hits

JD_i - day of the year (Julian day (from 01.01.70))

b_0 - number of hits on Julian day 0

b_1 - increase in number of hits per unit day

e_i - error

lm(visits ~ julian)

Linear vs Poisson Regression (log-linear)

$$\log(E(NH_i \mid JD_i, b_0, b_1)) = b_0 + b_1 JD_i$$

or

$$E[NH_i \mid JD_i, b_0, b_1] = e^{b_0 + b_1 JD_i} = e^{b_0} \cdot e^{b_1 JD_i}$$

If JD_i is increased by 1, $E[NH_i \mid JD_i, b_0, b_1]$ is multiplied by $\exp(b_1)$

glm(visits ~ julian, family = "poisson")

poisson regression

To model rates

$$\log(E[NHSS \mid JD_i, b_0, b_1) = \log(WH_0) + b_0 + b_1 JD_i$$

(number of hits from a specific website)

more information

Wiki on Poisson Regression

Model Checking and Model Selection

Basic assumptions for linear regressions

- Variance is constant
- trend is linear
- no big outliers
- no biases

- what to do if variance grows?
  - see if other variable explains the growth
  - sandwich library

    library(sandwich)

    lm1 <- ~~fda~~ lm(data1 ~ data2)

    vcovHC(lm1)

- the trend is not linear.
  - use Poisson regression
  - use data transformation (log, etc)
  - use linear regression

    + vcovHC

- missing covariate
  - check covariates carefully - use exploratory analysis
  - report unexplained patter

- outliers

how much influence do they have?

- fit regression with the outliers and then without

- if the two slopes are very different, then the outliers have a major effect

↓

more caution needed

So if you know for sure they are mistakes - remove and document them

If they are real - consider reporting how sensitive your estimate is to the outlier

Consider using a robust linear model fit

like rlm {MASS}

(down weights outliers)

Model Checking

- Default Plots

after fitting lm, try to see it's plot

~~dd~~ plot(lm1)

↑

Residuals vs Fittes

- Deviance

Commonly reported measure

might tell you that the model is wrong

- R^2

may be a bad summary

Model Selection

Usually you have a lot of variables -
you have to do some sort of filtering

How to choose correct variables?

- have domain-specific knowledge
  (prev. experience)

- exploratory analysis
  (make plots, make plots of residuals, coloring them by different variables,)

- Statistical selection

  • Step-wise (add/remove one var at a time)

  • AIC / BIC

  • etc

may bias your inference, so don't overdo statistical selection

Error measures

- R^2 (not always good enough)
- Adjusted R^2 - takes into account the number of estimated parameters
- AIC   Information criteria
- BIC

Model Selection - Step

all the terms

lm 1 = lm(score ~ . , data = movies)

aic Formula = step(lm1)

recomputes, finds better orders, adds, deletes, etc.

$\Downarrow$

score ~ box.office + running.time

Regsubsets

library(leaps)

regSub = regsubsets(score ~ . , data = movies)

↑

calculates BIC score
for all possible subsets

plot(regSub) ↓

Goal: minimize BIC

Notes And Resources

- exploratory/visual analysis – is a key
- automatic selection produces an answer,
  but it may bias inference (outfit – fits yours,
  but won't fit other sample)
- you may think of separating the sample
  into 2 groups

  use 1st to estimate
  2nd to do the inference

- goal – not to get „causal" model.

LARS package

Elements of /Statistical Machine Learning –
(book from the TO-READ list)

model selection
choosing variables