Science For A Better Life

Logistic Regression Re-Modelled

Tina Müller and Hannes-Friedrich Ulbrich

Bayer HealthCare, Global Drug Discovery Statistics

NCSC 2012 Potsdam

Triggering Question

Ophtalmology research: Vascularization of the cornea

- Neovascularization = Development of new blood vessels
- Vascularization is experimentally induced in mice
- Research aim: Reducing neovascularization

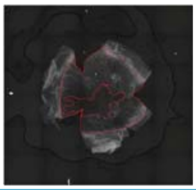
"Can we see differences in neovascularization of the cornea between treated and control animals?"

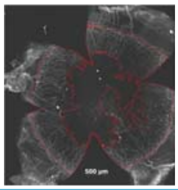
The Data

Data

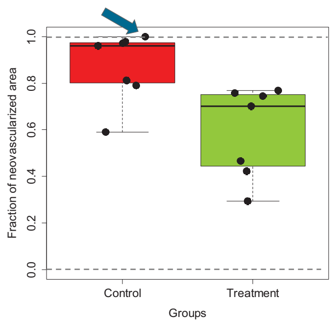
Proportion of neovascularized area compared to total cornea area

- Completely healthy: proportion = 0
- Completely diseased: proportion = 1





Modified Real-World Data Set



Data characteristics

- Two groups (control and treatment)
- Small sample size (7 per group)
- Bound between 0 and 1
- Boundary 1 is observed
- Aiming for reduction → 0 theoretically possible

Suitable Analysis Methods

Pre-requisite: Parametric method (Confidence interval of effect size desired, power important because of sample size)

First hits in literature search

1. Ignore boundaries: Linear model

2. Transform data: Linear model on transformed data
   - Popular: $y_{trans} = \log(\frac{y}{1-y})$

Similarity with Logistic Regression

Logistic regression

- Binary outcome y_b = 0 or 1, model $P(y_b = 1 \mid x) = \pi(x)$
- Probability ∈ [0, 1], use logit link function
- Conditional mean $y|x \sim Bin$

Proportions / Fractional regression

- Instead of binary outcome: „Probability" outcome $\pi(x)$ ∈ [0, 1]
- Conditional mean? Idea: $Bin \xrightarrow[n\to\infty]{} N$
- $y|x \sim N$
- Normal error distribution

Suitable Analysis Methods

Pre-requisite: Parametric method (Confidence interval of effect size desired, power important because of sample size)

First hits in literature search - Drawbacks

1. Ignore boundaries: Linear model
   - Predictions can lie outside possible value range
2. Transform data: Linear model on transformed data
   - Popular: $y_{trans} = \log(\frac{y}{1-y})$
   - How to transform value = 1 or value = 0?

Calculation in SAS® 9.2

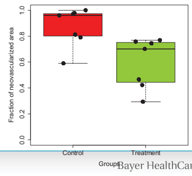
```
PROC GENMOD DATA = WORK.tmp;
   CLASS grp;
   MODEL neo = grp / LINK = logit DIST = normal;
   LSMEANS grp / ALPHA = .05 PDIFF=Control("Control");
   LSMESTIMATE grp "Treat - Control" -1 1 / CL EXP;
RUN;

WARNING: A link function appropriate for binomial data
         was selected but the binomial distribution was
         not used.
```

→ similar warnings for `probit` and `cloglog` link functions

Example Data Set

Method	Group	Mean estimate	Confidence interval	p-value
linear model (original)	Control	0.87	(0.73, 1.01)	0.012
	Treatment	0.59	(0.41, 0.77)	
linear model (logit)	Control			
	Treatment			
LogReg	Control	0.87	(0.70, 0.95)	0.010
	Treatment	0.59	(0.47, 0.71)	

- Confidence limit outside [0, 1]
- Transformed data not analyzed
- LogReg performs satisfactory

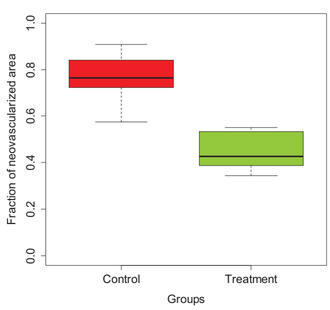


Example Data Set

Model estimates

Method	Group	Estimate	Standard Error	Mean	CI
linear model (logit)	Control	1.27	0.22	0.90	(0.80, 0.95)
	Treatment	-1.47	0.29	0.60	(0.41, 0.76)
LogReg	Control	1.16	0.23	0.87	(0.70, 0.95)
	Treatment	-1.35	0.28	0.59	(0.47, 0.70)

Are We Calculating What We Want?



Expanded Literature Research

Key words in literature inconsistent: percentage, fraction, proportion, …

Most literature found so far in econometrics, termed **Fractional Regression**

Expectation conditional mean $E(y|x) = G(x\theta)$

Choices of $G(\cdot)$

- monotonic, differentiable
- inverse of $G(\cdot) = h(\cdot)$ commonly known as link function

Model designation	Distribution function	$G(x\theta)$	$g(x\theta)$	$h(\mu)$
Cauchit	Cauchy	$\frac{1}{2}+\frac{1}{\pi}\arctan(x\theta)$	$\frac{1}{\pi}\frac{1}{(x\theta)^2+1}$	$\tan[\pi(\mu-0.5)]$
Logit	Logistic	$\frac{e^{x\theta}}{1+e^{x\theta}}$	$G(x\theta)[1-G(x\theta)]$	$\ln\frac{\mu}{1-\mu}$
Probit	Standard normal	$\Phi(x\theta)$	$\Phi(x\theta)$	$\Phi^{-1}(\mu)$
Loglog	Extreme maximum	$e^{-e^{-x\theta}}$	$e^{-x\theta}G(x\theta)$	$-\ln[-\ln(\mu)]$
Complementary loglog	Extreme minimum	$1-e^{-e^{x\theta}}$	$e^{x\theta}[1-G(x\theta)]$	$\ln[-\ln(1-\mu)]$

General quasi log likelihood $LL_i(\theta) = y_i \log[G(x_i\theta)] + (1-y_i)\log[1-G(x_i\theta)]$

Summary & Outlook

Our situation

- Until now: small number of experiments, each of limited sample size
- Logit-normal model:
  - seems to fit quite well
  - explainable to biologist
  - assumes a 'biological symmetry' between neovascularized and vessel-free area within total cornea area

Open questions

- Documentation of calculations in SAS® 9.2
- Other link functions more appropriate?
- Completely different approaches?

References

- L. Papke, J. Wooldridge (1996): Econometric methods for fractional response variables with an application to 401(K) plan participation rates. *Journal of Applied Econometrics*, 11, 619-632.
- E. Ramalho, J. Ramalho (2011): Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys*, 25, 1, 19-68.

Science For A Better Life

Thank you for your attention!