Relative Risk Regression

A Python Implementation

By David Marsden

Presentation Outline

- Statistical Advantages and Methods
- Python Implementation
- Illustration with External Dataset
- Monte Carlo Simulations of Relative Risk Regression

Relative Risk & Odds Ratio

- Both measure associations with binary outcomes
- Relative risk (RR) is preferred for many epidemiological analyses
- But logistic regression is commonly used
- Logistic regression estimates the odds ratio (OR)

Why is Relative Risk Better?

- Easy to understand
- Increase in probability of outcome vs. increase in odds
- ORs frequently mis-interpreted as RR (by journalists, the general public and sometimes academics who should know better!)
- ORs exaggerate the RR

Relative Risk Regression

- Two good statistical methods:
  - Log-binomial model
  - Poisson model with empirical variance (aka “robust” or “sandwich”)

- Potential problems with the log-binomial method:
  - May not converge properly
  - Confidence intervals may be too narrow

Poisson RR Model Details

- Generalized Linear Model (GLM)
  - Log link function
  - Poisson random component

- Empirical variance estimation
  - Can use software for generalized estimating equations (GEEs)

Poisson RR Model Diagnostics

- Can generate predicted probabilities greater than one
- Otherwise diagnostics are similar to other GLMs (e.g. outliers)

Software Considerations

- RR models can be fit with GEE software
- Many researchers are unfamiliar with GEEs and relative risk regression
- Convenient software encourages lesser-known statistical methods

Python Implementation

- Made a wrapper for the gee function from the statsmodel package
  - Designed to be easy to use
- Prints diagnostic info unique to the Poisson RR model
  - I.e. regarding observations with predicted probabilities greater than one
- Prints RRs and 95% confidence intervals for RR

File Edit Search Source Run Debug Consoles Projects Tools View Help

C:\Users\Work\Documents\UTSPH\Python

Editor - C:\Users\Work\Documents\UTSPH\Python\RelativeRiskExternalDataExample.py

RelativeRiskExternalDataExample.py | RelativeRisk.py | RelativeRiskSimulations.py | RelativeRiskCombinedScript.py

```
17 # Getting external data
18 carrot = pd.read_stata("https://stats.idre.ucla.edu/stat/stata/faq/eyestudy.dta")
19
20 # Checking external data
21 carrot.head(6)
22 carrot.mean(axis = 0) # It matches the SAS analysis
23
24 # Setting reference categories
25 # for replication to be same as other analysis
26 carrot["carrot0ref"] = 0
27 carrot.loc[carrot["carrot"]==0, ("carrot0ref")] = 1
28
29 carrot["gender2ref"] = 0
30 carrot.loc[carrot["gender"]==1, ("gender2ref")] = 1
31
32 RelativeRisk.RR("lenses ~ carrot0ref + gender2ref + latitude", "id", carrot)
33     # Replicates SAS analysis
34
35 # Running logit model for comparison
36 dv, ivs = patsy.dmatrices("lenses ~ carrot0ref + gender2ref + latitude",
37                           carrot, return_type='matrix')
38 logisticFit = sm.Logit(dv, ivs).fit(disp=0)
39 print("Carrot OR:", np.exp(logisticFit.params[1]))
40 print("Gender OR:", np.exp(logisticFit.params[2]))
41 print("Latitude OR:", np.exp(logisticFit.params[3]))
```

Help

Source Editor Object RelativeRisk.RR

RR

Definition : `RR(formula, idvar, df, printOutput=True)`

Type : Present in RelativeRisk module

Performs relative risk regression for dichotomous outcomes. Uses a working poisson model and an empirical ("robust") variance estimator.

Arguments:

1. formula - a formula expression for the model.
2. idvar - an identifier for each indepent observation of the data (typically a row).
3. df - the name of the pandas dataframe.
4. printOutput - a boolean argument for whether the function should print the output.

Example Code:

IPython console

Console 1/A

```
Relative Risk Regression
------------------------------------------------------------------------------
                           GEE Regression Results
==============================================================================
Dep. Variable:                    lenses   No. Observations:                  100
Model:                            GEE      No. clusters:                      100
Method:                   Generalized      Min. cluster size:                   1
                  Estimating Equations     Max. cluster size:                   1
Family:                        Poisson     Mean cluster size:                 1.0
Dependence structure:      Independence     Num. iterations:                     6
Date:                  Fri, 22 Jun 2018     Scale:                         1.000
Covariance type:                robust      Time:                        17:50:50
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.6521      0.490     -1.330      0.184      -1.613       0.309
carrot0ref     0.4832      0.195      2.473      0.013       0.100       0.866
gender2ref     0.2052      0.185      1.110      0.267      -0.157       0.567
latitude      -0.0100      0.013     -0.785      0.432      -0.035       0.015
==============================================================================
Skew:                         -0.1009   Kurtosis:                      -1.6893
Centered skew:                 0.0000   Centered kurtosis:             -3.0000
==============================================================================
Additional diagnostics:
0 observations have fitted probabilities greater than one
0.0 % of observations have fitted probabilities greater than one
==============================================================================
Relative Risk:
                RR
carrot0ref  1.621287
gender2ref  1.227772
latitude    0.990041
------------------------------------------------------------------------------
95% Confidence Intervals for Relative Risk:
                LCL       UCL
carrot0ref  1.105489  2.377746
gender2ref  0.854686  1.763715
latitude    0.965607  1.015093
==============================================================================
Carrot OR: 2.87974463773
Gender OR: 1.59558615687
Latitude OR: 0.977822965241

In [2]:
```

Permissions: RW | End-of-lines: CRLF | Encoding: ASCII | Line: 32 | Column: 14 | Memory: 28 %

Illustration with Carrot Data

- Fictional data set with risk factors for needing glasses
- Used to illustrate relative risk regression in SAS and Stata
- Successfully replicated this analysis with my python implementation
- Example function call:

```
RelativeRisk.RR("lenses ~ carrot0ref + gender2ref + latitude", "id", carrot)
```

Sample Output from Carrot Analysis

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.6521	0.490	-1.330	0.184	-1.613	0.309
carrot0ref	0.4832	0.195	2.473	0.013	0.100	0.866
gender2ref	0.2052	0.185	1.110	0.267	-0.157	0.567
latitude	-0.0100	0.013	-0.785	0.432	-0.035	0.015

Sample Output (cont.)

Additional diagnostics:

0 observations have fitted probabilities greater than one

0.0 % of observations have fitted probabilities greater than one

Sample Output (cont.)

Relative Risk:

	RR
carrot0ref	1.621287
gender2ref	1.227772
latitude	0.990041

95% Confidence Intervals for Relative Risk:

	LCL	UCL
carrot0ref	1.105489	2.377746
gender2ref	0.854686	1.763715
latitude	0.965607	1.015093

Carrot Analysis Results

	RR	95% CI	p-value	OR
Carrot	1.62	1.11 to 2.38	.01	2.88
Gender	1.23	0.85 to 1.76	.27	1.60
Latitude	0.99	0.97 to 1.02	.43	0.98

Simulation Methods

- 2 independent variables, 1 error term and 1 binary dependent variable
  - 1 dichotomous IV: X_1 ~ Bernoulli (0.25). True RR = 1.75
  - 1 continuous IV: X_2 ~ Normal (0, 1). True RR = 1.25
  - Error term: ε ~ Normal (0,1). True RR = 1.75
- Set maximum risk per individual to be 1
- Checked for quasi-complete separation
- Different unexposed risks (i.e. risk when IVs and $\varepsilon = 0$) and sample sizes
- Generated 10,000 data sets for each scenario

Simulation Results – Categorical Coefficient

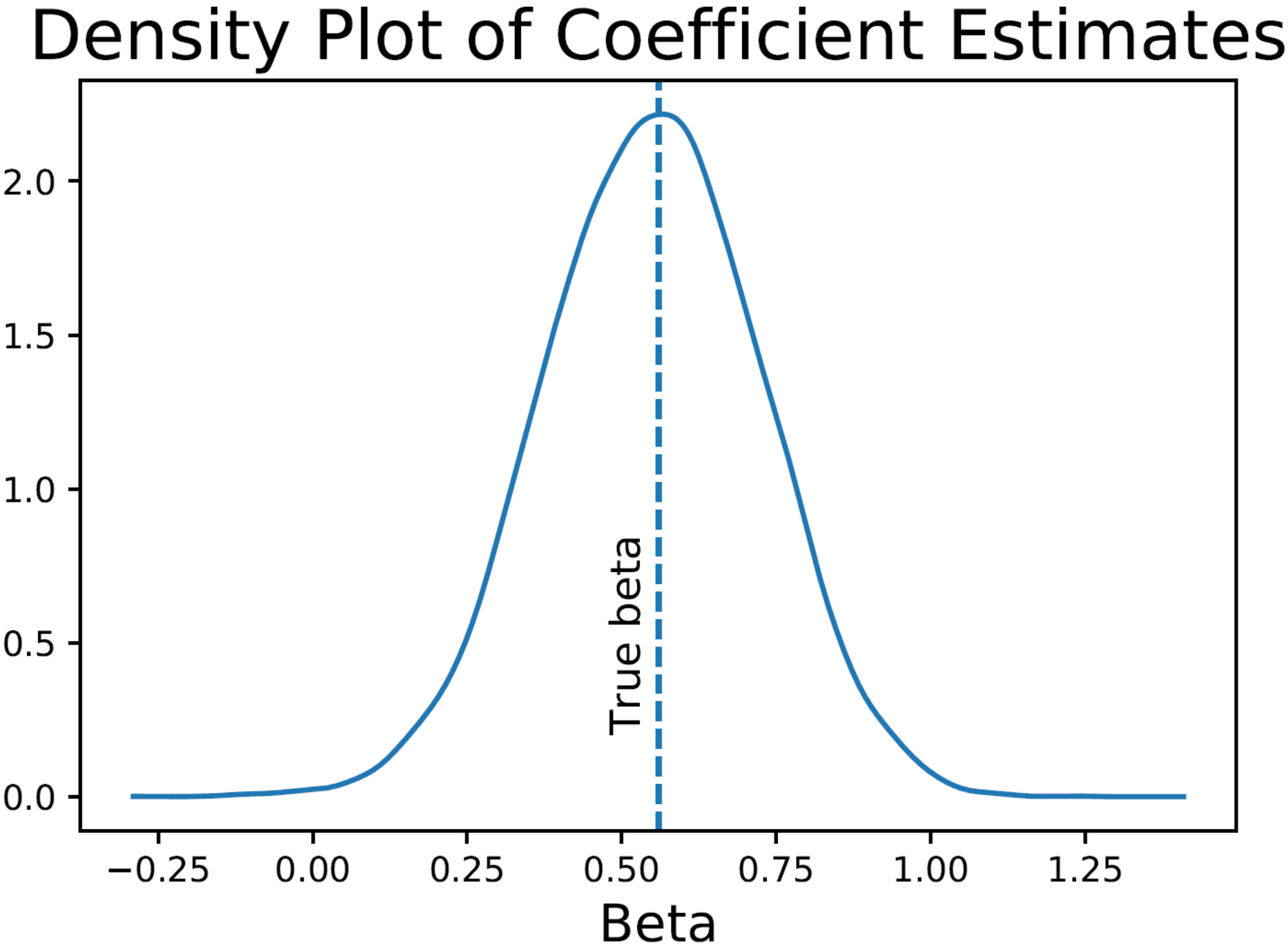
Sample Size (n)	Unexposed Risk	Mean Exposed Cases	RMSE	95% CI Coverage	Bias	OR Bias[a]
500	25%	55	0.13	93.7%	-0.051	0.30
500	15%[b]	33	0.18	95.3%	-0.011	0.18
500	10%	22	0.23	95.0%	-0.007	0.11
500	5%	11	0.35	95.6%	-0.010	0.04
1,000	15%	66	0.12	95.2%	-0.012	0.18
250	15%	16	0.25	95.4%	-0.018	0.17

a. The bias from using the OR as an estimate of the RR, or the mean difference between log OR and log RR
b. Estimates depicted in the following figure

Simulation Results – Continuous Coefficient

Sample Size (n)	Unexposed Risk	Mean Exposed Cases	RMSE	95% CI Coverage	Bias	OR Bias[a]
500	25%	55	0.06	93.7%	-0.019	0.11
500	15%	33	0.09	94.0%	-0.003	0.07
500	10%	22	0.11	94.5%	-0.002	0.04
500	5%	11	0.17	93.5%	-0.004	0.02
1,000	15%	66	0.06	94.8%	-0.003	0.07
250	15%	16	0.12	93.8%	-0.003	0.07

a. The bias from using the OR as an estimate of the RR, or the mean difference between log OR and log RR


Density Plot of Coefficient Estimates
2.0
1.5
1.0
0.5
0.0
True beta
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Beta

Simulation Discussion

- 95% CI coverage probabilities are very close to 95% (range: 93.5% to 95.6%)
- Relative risk regression is approximately unbiased in almost all cases
- Relative risk regression bias is always smaller than “OR bias”
  - I.e. unsurprisingly, the OR is not an unbiased estimator of the RR
- Estimates for the continuous IV are better than for the categorical IV
  - Smaller bias and RMSE

Concluding Remarks

- Relative risk regression is a capable and feasible statistical method
- Relative risk regression is already possible in standard statistical packages
- Made even more accessible with convenient software implementations

Recommended Resources

Stata tutorial: https://stats.idre.ucla.edu/stata/faq/how-can-i-estimate-relative-risk-using-glm-for-common-outcomes-in-cohort-studies/

SAS tutorial: https://stats.idre.ucla.edu/sas/faq/how-can-i-estimate-relative-risk-in-sas-using-proc-genmod-for-common-outcomes-in-cohort-studies/

Detailed statistical discussion: Lumley, T., Kronmal, R., & Ma, S. (2006). Relative risk regression in medical research: models, contrasts, estimators, and algorithms.

Intro and bibliography: https://www.mailman.columbia.edu/research/population-health-methods/relative-risk-regression

Source code

Available online:

https://github.com/MarsdenDavidG/RelativeRiskRegression

Any questions?