Valid Statistical Analysis for Logistic Regression with Multiple Sources

Rob Hall (Dept of Machine Learning, CMU)

Joint work with Yuval Nardi and Steve Fienberg

http://www.cs.cmu.edu/~rjhall rjhall+@cs.cmu.edu

Setting

Patient ID	Tobacco	Age	Weight	Heart Disease
0001	?	?	170	?
0002	?	?	150	N
0003	N	45	165	N

Patient ID	Tobacco	Age	Weight	Heart Disease
0001	Y	35	?	Y
0002	Y	40	?	?
0004	N	50	165	N

$$P(\text{heart disease} | \cdots ; \beta)$$

Logistic regression (or any glm)

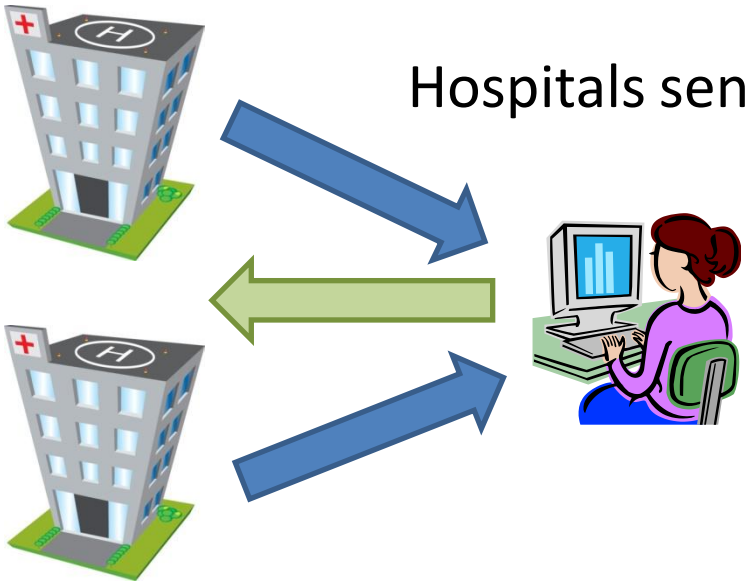
Alternatives

- Multiple organizations with databases want to do a statistical calculation (e.g., regression).

- Each would benefit by mining the pooled data.
- Not allowed/willing to share data (e.g., HIPAA).

- Share transformed data?
- Secure multiparty computation?

In an Ideal World



Hospitals send data to a “trusted party.”

“Trusted party” computes regression, sends same coefficients back to each hospital.

- This is an “ideal” scenario - trusted parties don’t exist.
- Using cryptography, we can do the computation as if they did.

Secure Multiparty Computation

- A protocol computes a "functionality:"

$$\{(X_1, y_y), (X_2, y_2) \cdots\} \rightarrow \{\hat{\beta}(1), \hat{\beta}(2), \cdots\}$$

Party 1's data Party 2's data Each party gets a copy of the output

- Messages are exchanged and coins are flipped, each party has a "view"
- It is secure whenever the messages can be simulated ("semi-honest" model):

$$S_i((X_i, y_i), \beta) \equiv_c \text{view}_i(\{(X_1, y_1)(X_2, y_2), \cdots\}, \beta)$$

Additive Random Shares

- Split a secret quantity so each party has a *share*:

$$\sum_{i=1}^{P} a_i \equiv a \mod n \qquad a_i \in \mathbb{Z}_n$$

- Marginally each share is uniformly distributed on $\mathbb{Z}_n$.
- Messages consisting of shares are easy to simulate.
- Finite precision reals only slightly trickier.

Multiplication

$$ab = (\sum_i a_i)(\sum_j b_j) = \underbrace{\sum_i a_i b_i}_{\text{Local product}} + \underbrace{\sum_i \sum_{j \neq i} a_i b_j}_{\text{Different parties}}$$

- Using homomorphic encryption:
  - i encrypts a_i
  - j computes: $E(s) = E(a_i b_j - r) = E(a_i)^{b_j} - E(r)$
  - i decrypts: $r + s \equiv a_i b_j \mod n$
- a_i is encrypted when sent, so message is easy to simulate.
- r, s are uniform in $\mathbb{Z}_n$.

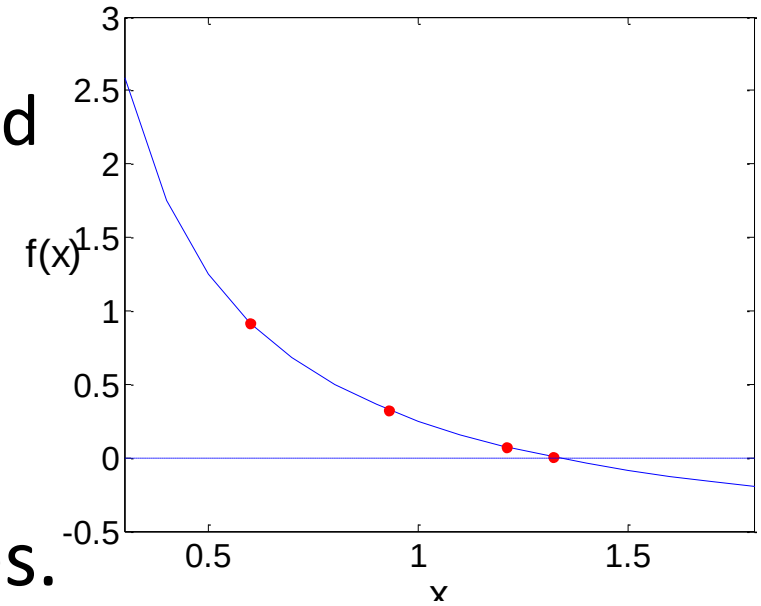
Linear Regression

- The MLE is: $\hat{\beta} = (X^T X)^{-1} X y$

1. Compute Shares of $X^T X, Xy$
2. Secure matrix inversion
   - Similar to Newton's method on the function:

$$f(x) = x^{-1} - a$$

3. Secure matrix multiply.
4. Modular addition of shares.



Logistic Regression (IRLS)

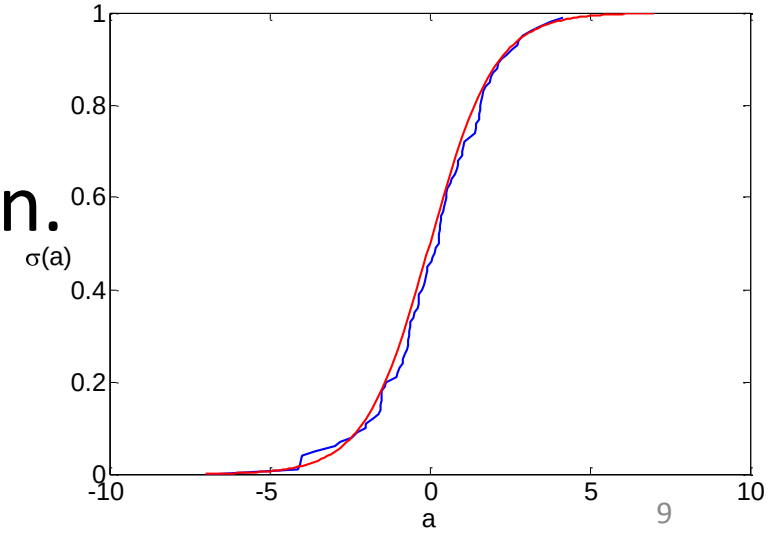
- Newton-Raphson iterates: $\beta' = \beta - \nabla^2(\beta)^{-1}\nabla(\beta)$

$$\nabla^2(\beta) = (X^T \Sigma X)^{-1} \quad \nabla(\beta) = X(y - \sigma)$$

- Approximate sigmoid by the empirical CDF:

$$\sigma(a) = (1 + \exp(-a))^{-1} \simeq L^{-1} \sum_{i=1}^{L} 1_{\{a > z_i\}}$$

- Secure computation of "greater than" is well known.
- Approximation error decreases with L.



CPS - Experimental Verification

	Non-Private	Private $L = 100$		Private $L = 500$	
		Mean	S.D.	Mean	S.D.
Intercept	-16.7401480	-16.5350768	1.4374467	-16.7433108	0.6123558
Alimony	0.0000314	0.0000304	0.0000031	0.0000314	0.0000011
Child Sup	0.0001695	0.0001646	0.0000168	0.0001704	0.0000058
Property Tax	0.0003021	0.0002937	0.0000306	0.0003043	0.0000105
Number in Household	0.9636042	0.9456159	0.0863145	0.9637204	0.0347013
Number of Children	-1.0408581	-1.0220314	0.0914965	-1.0407592	0.0372779
Number Married	0.0153436	0.0143678	0.0029414	0.0156506	0.0017629
Child Sup Individual	-0.0001062	-0.0001024	0.0000121	-0.0001071	0.0000038
Education	0.3199738	0.3142047	0.0290581	0.3195771	0.0116572
Social security Payments	-0.0000320	-0.0000311	0.0000032	-0.0000319	0.0000012
Age	1.4960356	1.5083476	0.1180312	1.4914760	0.0515902
	0.5940452	0.6294996	0.0791097	0.5884333	0.0258216
	1.1832524	1.2073569	0.0982567	1.1760422	0.0427014
	1.4186196	1.4273286	0.1133918	1.4162917	0.0479690
Marital Status	-0.0904335	-0.0813627	0.0229776	-0.0917471	0.0077264
	-0.2548489	-0.2451984	0.0283438	-0.2518805	0.0109333
	-0.4062043	-0.4066925	0.0300718	-0.3998577	0.0205781
	0.5283986	0.5232678	0.0455995	0.5190602	0.0255049
	0.5437736	0.5323785	0.0519936	0.5393087	0.0211449
	0.1140666	0.0806117	0.0644963	0.1195166	0.0131626
Race	0.1699793	0.1688992	0.0153372	0.1687002	0.0062255
	-0.0694353	-0.0682115	0.0090636	-0.0684738	0.0043338
	-0.3149339	-0.3058253	0.0334037	-0.3151108	0.0120143
Sex	-0.2873196	-0.2818015	0.0243565	-0.2872207	0.0100545

CPS - Experimental Verification

	Non-Private	Private $L = 100$		Private $L = 500$	
		Mean	S.D.	Mean	S.D.
Intercept	-16.7401480	-16.5350768	1.4374467	-16.7433108	0.6123558
Alimony	0.0000314	0.0000304	0.0000031	0.0000314	0.0000011
Child Sup	0.0001695	0.0001646	0.0000168	0.0001704	0.0000058
Property Tax	0.0003021	0.0002937	0.0000306	0.0003043	0.0000105
No. in Household	0.96	0.95	0.09	0.96	0.03
Number Married	0.0153436	0.0143678	0.0029414	0.0156506	0.0017629
Child Sup Individual	-0.0001062	-0.0001024	0.0000121	-0.0001071	0.0000038
Education	0.3199738	0.3142047	0.0290581	0.3195771	0.0116572
Social security Payments	-0.0000320	-0.0000311	0.0000032	-0.0000319	0.0000012
Age	1.4960356	1.5083476	0.1180312	1.4914760	0.0515902
	0.5940452	0.6294996	0.0791097	0.5884333	0.0258216
	1.1832524	1.2073569	0.0982567	1.1760422	0.0427014
	1.4186196	1.4273286	0.1133918	1.4162917	0.0479690
Marital Status	-0.0904335	-0.0813627	0.0229776	-0.0917471	0.0077264
	-0.2548489	-0.2451984	0.0283438	-0.2518805	0.0109333
	-0.4062043	-0.4066925	0.0300718	-0.3998577	0.0205781
	0.5283986	0.5232678	0.0455995	0.5190602	0.0255049
	0.5437736	0.5323785	0.0519936	0.5393087	0.0211449
	0.1140666	0.0806117	0.0644963	0.1195166	0.0131626
Race	0.1699793	0.1688992	0.0153372	0.1687002	0.0062255
	-0.0694353	-0.0682115	0.0090636	-0.0684738	0.0043338
	-0.3149339	-0.3058253	0.0334037	-0.3151108	0.0120143
Sex	-0.2873196	-0.2818015	0.0243565	-0.2872207	0.0100545

CPS - Experimental Verification

	Non-Private	Private $L = 100$		Private $L = 500$	
		Mean	S.D.	Mean	S.D.
Intercept	-16.7401480	-16.5350768	1.4374467	-16.7433108	0.6123558
Alimony	0.0000314	0.0000304	0.0000031	0.0000314	0.0000011
Child Sup	0.0001695	0.0001646	0.0000168	0.0001704	0.0000058
Property Tax	0.0003021	0.0002937	0.0000306	0.0003043	0.0000105
Number in Household	0.9636042	0.9456159	0.0863145	0.9637204	0.0347013
Number of Children	-1.0408581	-1.0220314	0.0914965	-1.0407592	0.0372779
Number Married	0.0153436	0.0143678	0.0029414	0.0156506	0.0017629
Child Sup Individual	-0.0001062	-0.0001024	0.0000121	-0.0001071	0.0000038
Education	0.3199738	0.3142047	0.0290581	0.3195771	0.0116572
Social security Payments	-0.0000320	-0.0000311	0.0000032	-0.0000319	0.0000012
	1.4960356	1.5083476	0.1180312	1.4914760	0.0515902
	0.5940452	0.6294996	0.0791097	0.5884333	0.0258216
Age(3)	**1.18**	**1.20**	**0.10**	**1.18**	**0.04**
Marital Status	-0.0904335	-0.0813627	0.0229776	-0.0917471	0.0077264
	-0.2548489	-0.2451984	0.0283438	-0.2518805	0.0109333
	-0.4062043	-0.4066925	0.0300718	-0.3998577	0.0205781
	0.5283986	0.5232678	0.0455995	0.5190602	0.0255049
	0.5437736	0.5323785	0.0519936	0.5393087	0.0211449
	0.1140666	0.0806117	0.0644963	0.1195166	0.0131626
Race	0.1699793	0.1688992	0.0153372	0.1687002	0.0062255
	-0.0694353	-0.0682115	0.0090636	-0.0684738	0.0043338
	-0.3149339	-0.3058253	0.0334037	-0.3151108	0.0120143
Sex	-0.2873196	-0.2818015	0.0243565	-0.2872207	0.0100545

Ongoing Work

- Faster approximations to logistic functions.
- Record linkage (assumed here).
- Imputation of missing data.
- Secure computation of goodness-of-fit statistics.
- Log-linear models.
- Other GLMs.

Questions

- For the technical details and a working implementation please see:

`http://www.cs.cmu.edu/~rjhall/slr`