

# Inference for the Regression Coefficient in a Bivariate Normal Model



## when Data have been Subjected to Censoring



Jennifer Boyko, Haverford College, Summer REU 2007 – Ohio Wesleyan University

Faculty Mentor: Dr. Scott Linder, Ohio Wesleyan University, Dept. of Mathematics and Computer Science

### Censoring of Data

Start with  $n$  subjects, but record measurements for only  $p$

#### Type I

- Stop experiment at specified time (number of observations is random)

#### Type II

- Stop experiment after specified number of observations (duration of experiment is random)

### Example (Type II Censoring)

- Begin with  $n$  monkeys
- Feed each a high-fat diet
- Record  $X$  = time of death  
 $Y$  = amount of plaque in aorta
- Stop experiment after  $p$  monkeys die
- $n-p$  observations have been "censored" (but we know something about them)

### Linear regression in bivariate normal model

Given  $X = x$ ,

$$Y = \alpha + \beta x + \varepsilon$$

where  $\varepsilon \sim N(0, \sigma^2)$ ,

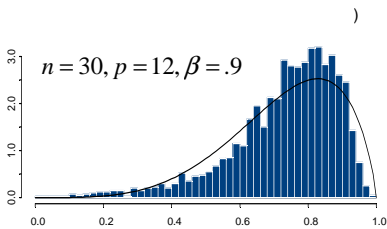
The regression coefficient ( $\beta$ ) is dependent on the correlation coefficient between  $X$  and  $Y$

Independence of  $X$  and  $Y \Leftrightarrow \beta = 0$

### The Maximum Likelihood Estimator

$$\hat{\beta} = \frac{\sum_{i=r}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=r}^n (x_i - \bar{x})^2}$$

This estimator's sampling distribution is mathematically intractable and must be approximated.



### Acknowledgements:

This work was sponsored by grants from the National Science Foundation (0648751) and from the Ohio Wesleyan University Summer Science Research Program.

### Objectives

- I. Develop a method for approximating the power of the test for independence  $H_0 : \beta = 0$ . based on  $\hat{\beta}$ .
- II. Examine properties of a "quick estimator" of  $\beta$

#### I. Power

- The power of a test is the probability of rejecting a false hypothesis. If  $\beta = \beta_1$ ,

$$P\left(\frac{|\hat{\beta}|\sqrt{(p-2)SS_{XX}}}{\sqrt{SSE}} > t_{p-2}(\alpha/2) | \beta = \beta_1\right)$$

where  $SS_{XX} = \sum (x_i - \bar{x})^2$

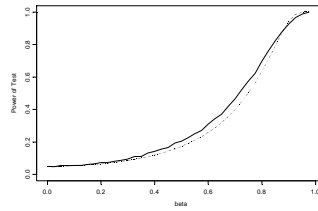
- Needs to be approximated via simulation.

#### Alternative approximation

- We replace the random  $x$  with its average value (tabulated), replacing  $SS_{XX}$  with  $SS_{uu}$
- By doing this, we eliminate a source of variability and we can approximate power without simulation

$$\frac{|\hat{\beta}|\sqrt{(p-2)SS_{uu}}}{\sqrt{SSE}} \sim t_{p-2}(\delta)$$

$$\delta = \frac{\beta\sqrt{SS_{uu}}}{\sqrt{1-\beta^2}}$$



Exact power: \_\_\_\_\_  
Approximate power: -----

	$\beta$	Simulated Power	Approximated Power	Relative Error
$n = 20, p = 8$	.1	.047	.052	.093
	.2	.061	.057	-.075
	.3	.070	.066	-.060
	.4	.083	.081	-.031
	.5	.106	.104	-.015
	.6	.176	.142	-.192
	.7	.223	.209	-.064
	.8	.387	.342	-.116
	.9	.654	.664	.015
$n = 20, p = 12$	.1	.056	.054	-.050
	.2	.074	.065	-.118
	.3	.095	.086	-.094
	.4	.142	.120	-.160
	.5	.205	.173	-.154
	.6	.310	.259	-.166
	.7	.466	.400	-.142
	.8	.698	.636	-.090
	.9	.928	.942	.015

### Minimum Sample Size for Specified Power

Power	$\beta$								
	.2	.3	.4	.5	.6	.7	.8	.9	
.1	(-,24)	(-,14,18)	(8,11,12)	(7,8,10)	(6,7,7)	(5,7,7)	(5,7,7)	(5,7,7)	
.2	(-,)	(-,27)	(-,16,22)	(-,14,16)	(8,11,12)	(7,8,10)	(6,7,7)	(5,7,7)	
.3	(-,)	(-,29)	(-,24)	(-,16,21)	(-,14,16)	(8,11,12)	(7,8,9)	(5,7,7)	
.4	(-,)	(-,)	(-,27)	(-,18,24)	(-,16,20)	(-,12,15)	(8,10,11)	(6,7,8)	
.5	(-,)	(-,)	(-,29)	(-,24)	(-,16,23)	(-,14,18)	(8,11,12)	(6,7,8)	
.6	(-,)	(-,)	(-,27)	(-,18,24)	(-,16,20)	(-,12,15)	(7,8,9)		
.7	(-,)	(-,)	(-,)	(-,28)	(-,24)	(-,16,22)	(-,14,16)	(8,9,11)	
.8	(-,)	(-,)	(-,)	(-,)	(-,27)	(-,18,24)	(-,15,19)	(8,10,12)	
.9	(-,)	(-,)	(-,)	(-,)	(-,29)	(-,24)	(-,16,21)	(-,12,14)	

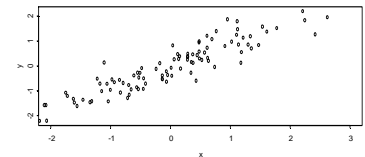
The ordered triples represent the minimum value of  $p$  required to achieve desired power when  $n=10, 20$ , and  $30$ , respectively. This table provides researchers with guidance for choosing their sample size in an experiment.

#### II. Barton and Casley Estimator

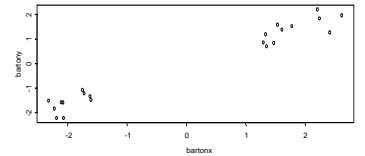
- This estimator uses the average of the first  $k$  values of the ordered pairs and the last  $k$  values to calculate

$$\beta^* = \frac{(\bar{y}_k^1 - \bar{y}_k)}{(\bar{x}_k^1 - \bar{x}_k)}$$

- What is the optimal value of  $k$ ?



Full sample when  $n=100$



Barton and Casley sample when  $n=100$  and  $k=10$

$n = 20$  and  $p = 12$

$k$	Single Censoring		Symmetric Censoring	
	$\text{Var}(\beta^*)$	$\varepsilon(k)$	$\text{Var}(\beta^*)$	$\varepsilon(k)$
1	.612	.437	1.091	.425
2	.394	.680	.684	.678
3	.344	<b>.778</b>	.569	.815
4	.355	.754	.549	<b>.844</b>
5	.368	.727	.591	.785
6	.420	.637	.694	.668
$\text{Var}(\hat{\beta}) = .268$		$\text{Var}(\hat{\beta}) = .464$		

### Optimal value of $k$

	$p$	Single Censoring			Symmetric Censoring					
		$k$	$\text{Var}(\beta^*)$	$\varepsilon(k)$	$k$	$\text{Var}(\beta^*)$	$\varepsilon(k)$			
$n = 10$	$p = 6$	2	1.108	0.819	0.740	2	1.594	1.315	0.825	
		8	0.422	0.326	0.772	2	0.494	0.393	0.794	
	$n = 20$	$p = 6$	10	3	1.994	1.611	0.808	2	7.403	6.047
8			2	0.963	0.686	0.712	3	2.529	2.066	0.817
10		3	0.530	0.419	0.789	3	1.095	0.906	0.827	
		12	3	0.344	0.268	0.778	4	0.581	0.481	0.827
14		4	0.235	0.189	0.807	5	0.321	0.269	0.838	
		16	4	0.166	0.132	0.793	4	0.197	0.162	0.821
18	5	0.112	0.091	0.805	5	0.121	0.100	0.827		
	$n = 30$	$p = 6$	2	2.527	2.048	0.811	2	17.744	14.155	0.798
10			3	0.762	0.569	0.746	3	2.614	2.169	0.830
14		4	0.364	0.296	0.812	5	0.832	0.693	0.833	
		18	5	0.210	0.158	0.754	5	0.344	0.297	0.863
22		8	0.129	0.103	0.797	8	0.170	0.142	0.839	
		28	7	0.062	0.051	0.810	7	0.065	0.054	0.823