Introduction to Statistical Models in Biomedical Research

ภาณุวัฒน์ เลิศสิทธิชัย

อบรมวิจัยสำหรับแพทย์ประจำบ้าน

ราชวิทยาลัยศัลยแพทย์แห่งประเทศไทย

๖ ก.ค. ๒๕๖๓

Corrected, 16 July 2020

Contents

- Biomedical research questions & relation to statistical models
- From the "Regression Model" point of view
- Total Variation = $\boxed{\text{Systematic}}$ + Random
- Lots of equations & numbers!
- *Research designs* will not be covered here, though just as important

Research Questions

Various biomedical research questions have corresponding appropriate statistical models, *given appropriate research designs*

- **Incidence**, **prevalence**, **average**/mean values
- **Treatment** (Impact of intervention)
- **Diagnosis** (Diagnostic accuracy)
- **Prevention** (Impact of intervention)
- **Risk**/etiologic/**prognostic**/predictive factors

Regression Point of View (Not New)

Many medical research questions can be rephrased in "Regression" terms, i.e., as *regression models* : in terms of *statistical relations* between Outcome/ Variate Y (*random*) and Predictors/Covariates X (*fixed*): *mean-value* relations

Questions of ***Incidence, Prevalence, Average*** (*no* X)

- $E(Y) = \alpha$

Questions of ***One Covariate or Risk Factor*** *or One Treatment Factor* (*one* X)

- $E(Y|X = x) = \alpha + \beta x$

Questions of ***Multiple Covariates*** (*many* $X_1, X_2, \ldots$)

- $E(Y|X_1 = x_1, X_2 \ldots) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots$

α, β's are *regression coefficients* or *regression parameters* (constants)*

*We will confine our discussions to the **Frequentist version/School** of Statistics

Example 1: Question of Average Value

Average (mean) value of BMI in a target population: Cohort Study

- $E(BMI) = \alpha$

From a random sample of 20 patients from the target population:

- $E(BMI) = \alpha \approx 30.5\ kg/m^2$; this is the *estimated average*
- If the sample $SD = 15.2\ kg/m^2$,

  the *standard error* of the average BMI is $SE \approx 15.2/\sqrt{20}\ kg/m^2$
- Thus, the 95% CI for BMI is $\left\{30.5 \pm \frac{1.96 \times 15.2}{\sqrt{20}} \frac{kg}{m^2}\right\} = \left\{30.5 \pm 6.6 \frac{kg}{m^2}\right\}$

Example 2: Question of Incidence

Incidence of postoperative infection in a population: Cohort Study

- Outcome $Y = 1 \; or \; 0$ (binary)*

Two common forms: firstly, the *linear* parametrization**,

- $E(Y) = \alpha \equiv \pi$

Secondly, by *reparametrizing*, using the *logistic transform*,

- $\log(\frac{\pi}{1-\pi}) = \theta$, i.e., the *log odds* parameter, then we have the *logistic* parametrization,

- $E(Y) = \pi = \frac{e^{\theta}}{1+e^{\theta}} = \frac{1}{e^{-\theta}+1}$

*Formally, we set an Indicator Function $I(Y)$, such that $I(Y) = 1$ for $Y = Infected$, etc.
**It is common to use π to denote the parameter of a Bernoulli/Binomial trial/process

Example 2: cont. 1

A sample of 10 patients with 2 infections*:

- $Y = \{1, 1, 0, 0, 0, 0, 0, 0, 0, 0\}$
- $E(Y) \approx \frac{1+1+0+0+0+0+0+0+0+0}{10} = 0.2$; this is the *estimated incidence***
- $\widehat{E(Y)} = \hat{\pi} = 0.2$; the standard error is $SE(\hat{\pi}) \approx \sqrt{\frac{0.2(1-0.2)}{10}} = 0.13$
- 95% CI for $\hat{\pi}$ is $\approx \{0.2 \pm 1.96 \times 0.13\} = \{0.2 \pm 0.25\} = \{-0.05, 0.45\}$
- Using *linear* parametrization, the incidence may have negative values!

* Again, formally, the set is defined for the indicator function $I(Y)$, not the random variable Y

**The "hat" in, e.g., $\hat{\pi}$ is commonly used to denote the "Maximum likelihood" estimator

Example 2: cont. 2

Using the *logistic* parametrization, on the other hand:

- $\theta = \log\left(\frac{\pi}{1-\pi}\right) \approx \hat{\theta} = \log\left(\frac{0.2}{1-0.2}\right) = \log(0.25) = -1.39$
- $SE(\hat{\theta}) \approx \sqrt{\frac{1}{2} + \frac{1}{(10-2)}} = 0.79$
- 95% CI for $\hat{\theta}$: $\{-1.39 \pm 1.96 \times 0.79\} = \{-2.94, 0.15\}$
- **Transforming back to** π, the 95% CI for incidence $\hat{\pi}$: $\{0.05, 0.54\}$
- No negative values!
- Also, the logistic scale is more convenient for multivariable regression models for Binary Outcomes (logistic regression models)

Maximum Likelihood Estimate (MLE)

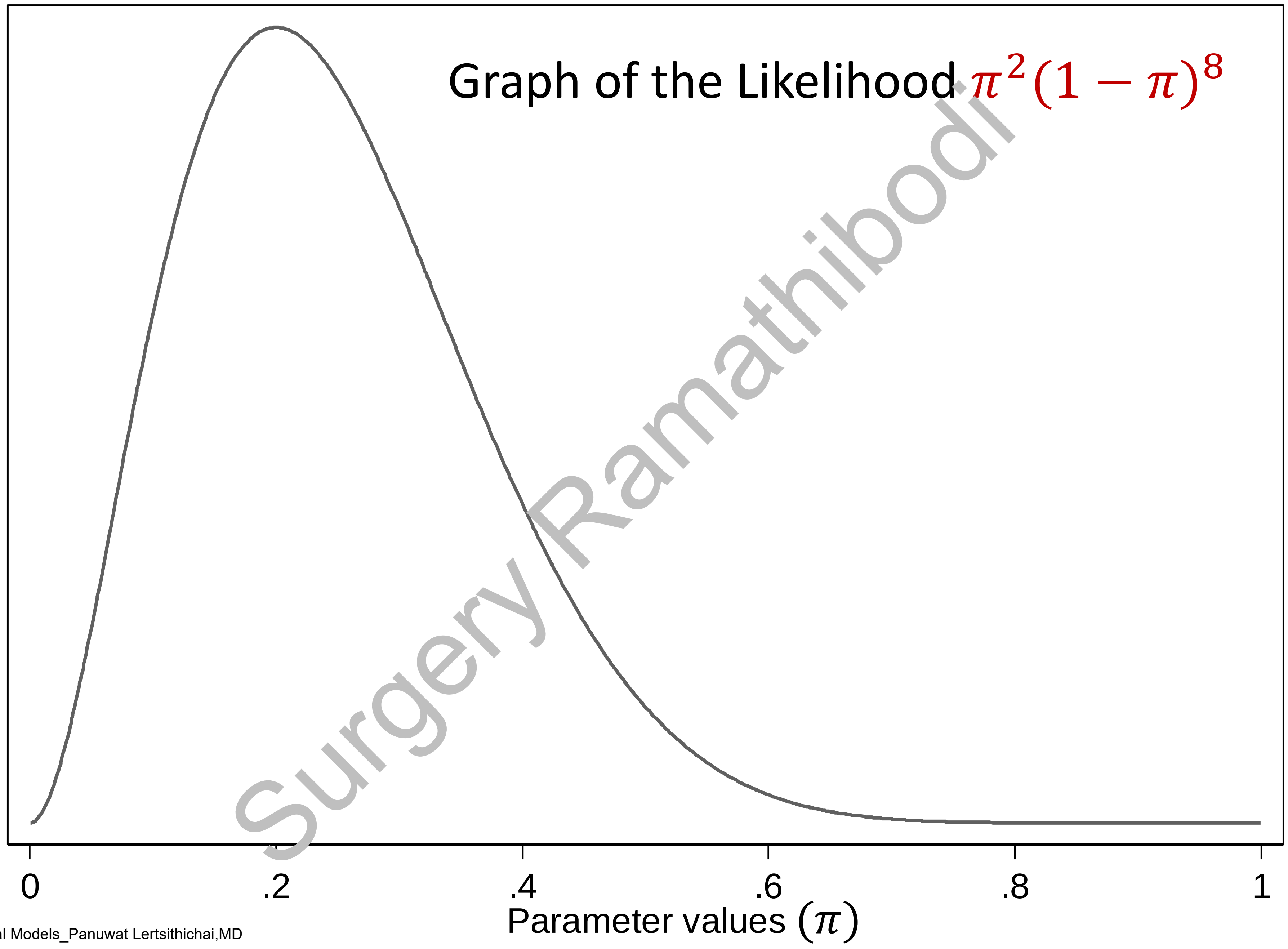
A sample of 10 patients with 2 infections:

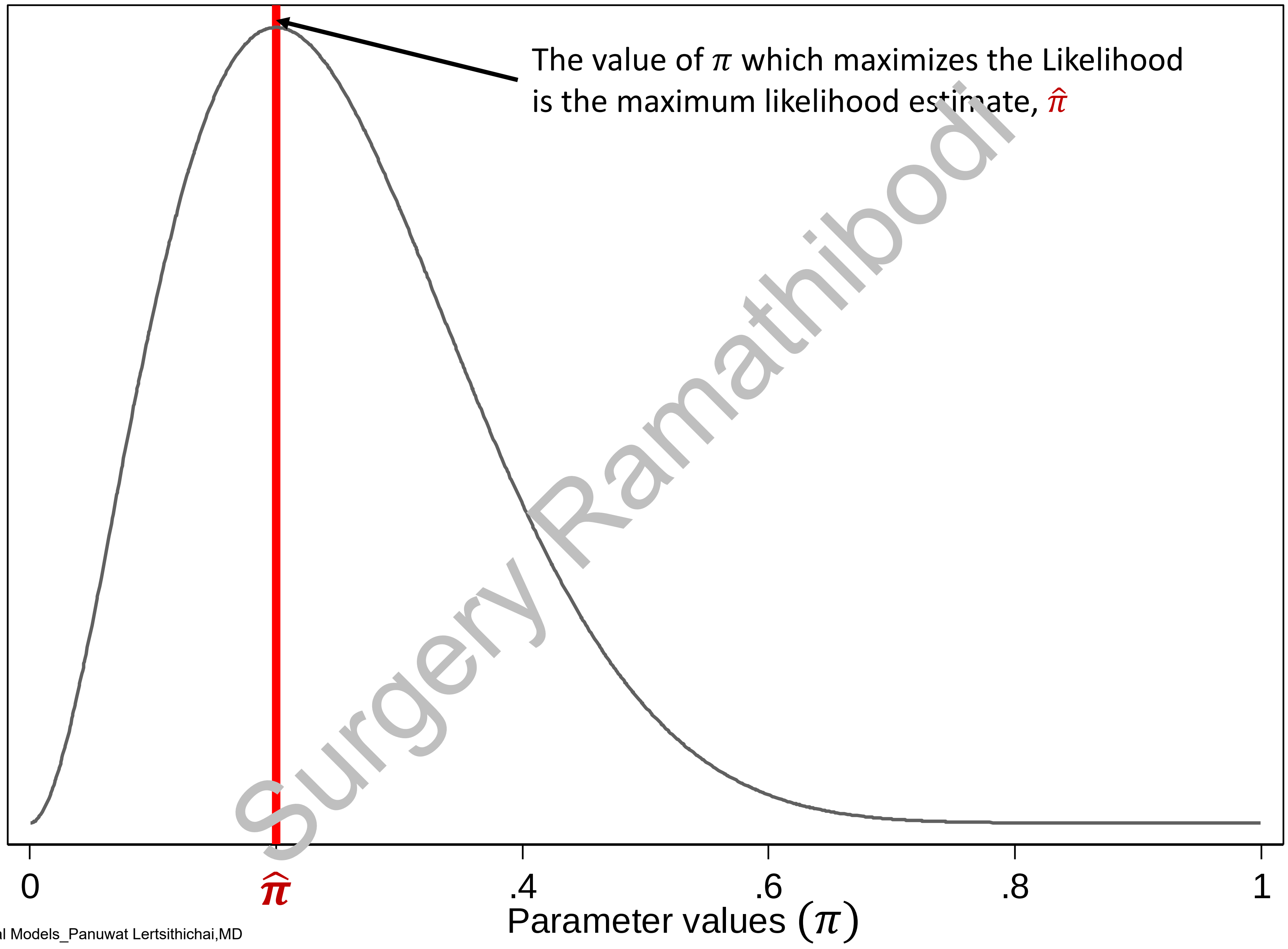
- $Y = \{1, 1, 0, 0, 0, 0, 0, 0, 0, 0\}$
- $E(Y) \approx \frac{1+1+0+0+0+0+0+0+0+0}{10} = 0.2 = \tilde{\pi}$
- This is the *estimated incidence* using the *Method of Moments**

Another method is that of **Maximum Likelihood**

- Consider the sequence of patients *as a sequence of Bernoulli Trials*
- With probability of infection (i.e. incidence) π
- The likelihood is defined as the probability of observed outcome sequence, **as a function of** π : $Likelihood(\pi) = \pi^2(1-\pi)^8$

*The “tilde” in $\tilde{\pi}$ is used to distinguish this from the Maximum likelihood estimator, $\hat{\pi}$


Graph of the Likelihood $\pi^2(1-\pi)^8$
0
.2
.4
.6
.8
1
Parameter values (π)


The value of π which maximizes the Likelihood
is the maximum likelihood estimate, $\hat{\pi}$
0
$\hat{\pi}$
.4
.6
.8
1
Parameter values (π)

Maximum Likelihood Estimate: cont.

Maximize the Likelihood $\pi^2(1-\pi)^8$ with respect to π

- Take the log first: $\log\{\pi^2(1-\pi)^8\} = 2 \times \log(\pi) + 8 \times \log(1-\pi)$
- Differentiate this with respect to π & set to 0: $\frac{2}{\hat{\pi}} - \frac{8}{(1-\hat{\pi})} = 0$
- Solve this equation and the "MLE" is: $\hat{\pi} = \frac{2}{10} = 0.2$
- The MLE is actually *equivalent* to the method of moments, in this case
- In general this may not be the case
- There are theoretical advantages of MLE, and also some disadvantages
- *We mention the MLE here because of its wide spread use*; see later*

*Used in likelihood-based models, especially useful in the case of Exponential Family of models

Example 3: Question of One Covariate

A RCT comparing operation A vs. B in terms of pain (VAS)

- One outcome: $Y = pain(VAS)$
- One binary Treatment factor: $X = 0$ for operation A; $X = 1$ for operation B

The regression equation is

- $E(Y|X = x) = \alpha + \beta x$

This can be expanded into 2 equations, one for each operation:

- $E(Y|X = 0) \equiv E(Y)_0 = \alpha$ for operation A with $X = 0$; and
- $E(Y|X = 1) \equiv E(Y)_1 = \alpha + \beta$ for operation B with $X = 1$

Example 3: cont. 1

- $E(Y)_0 = \alpha$ is the average pain for operation A
- $E(Y)_1 = \alpha + \beta$ is the average pain for operation B
- Therefore β is the difference in average pain between 2 operations
- The statistical *Null hypothesis* is: $H_0 : \beta = 0$
- Given that Y (pain VAS) has a Normal distribution,

  The test for $\boldsymbol{\beta = 0}$ is simply the *t*-test
- *Statistical tests can be interpreted as tests for regression parameters*
- Note*: $E(Y)_1 - E(Y)_0 \approx \bar{Y}_1 - \bar{Y}_0 = \hat{\beta}$

* Formally, $E(Y)$ is the **true** mean in a population, while $\bar{Y}$ is the **sample** mean, so $E(Y) \approx \bar{Y}$

Example 3: cont. 2

Result of a RCT with 20 patients per arm:

- Average pain (VAS) operation A: 5.6; SD: 2.3
- Average pain (VAS) operation B: 4.3; SD: 2.1
- $\hat{\beta} = 4.3 - 5.6 = -1.3$
- $SE(\hat{\beta}) \approx \sqrt{\frac{2.3^2}{20} + \frac{2.1^2}{20}} = 0.69$
- The t-test value = $\frac{\hat{\beta}}{SE(\hat{\beta})} = -\frac{1.3}{0.69} = -1.87;$ 2-sided p-value 0.070
- 95% CI for $\hat{\beta}$: $\{-1.3 \pm 2.02 \times 0.69\} = \{-2.69, 0.09\}$ ***covers 0***

Example 3: cont. 3 – Factors Related to Pain

- $E(Y)_0 = \alpha' \qquad + \beta_{age0}age_0 + \beta_{sex0}sex_0 + \cdots$ for operation A
- $E(Y)_1 = \alpha' + \beta + \beta_{age1}age_1 + \beta_{sex0}sex_1 + \cdots$ for operation B

- $E(Y)_1 - E(Y)_0 = \beta$ **is true** (on average) **for RCT's**

- *Non-randomized* (*Observational*) study designs cannot achieve this
- One way for observational studies to emulate RCT's is to do matching using some appropriate scoring system: known as *propensity scoring*

Example 4: Question of Many Covariates (a)

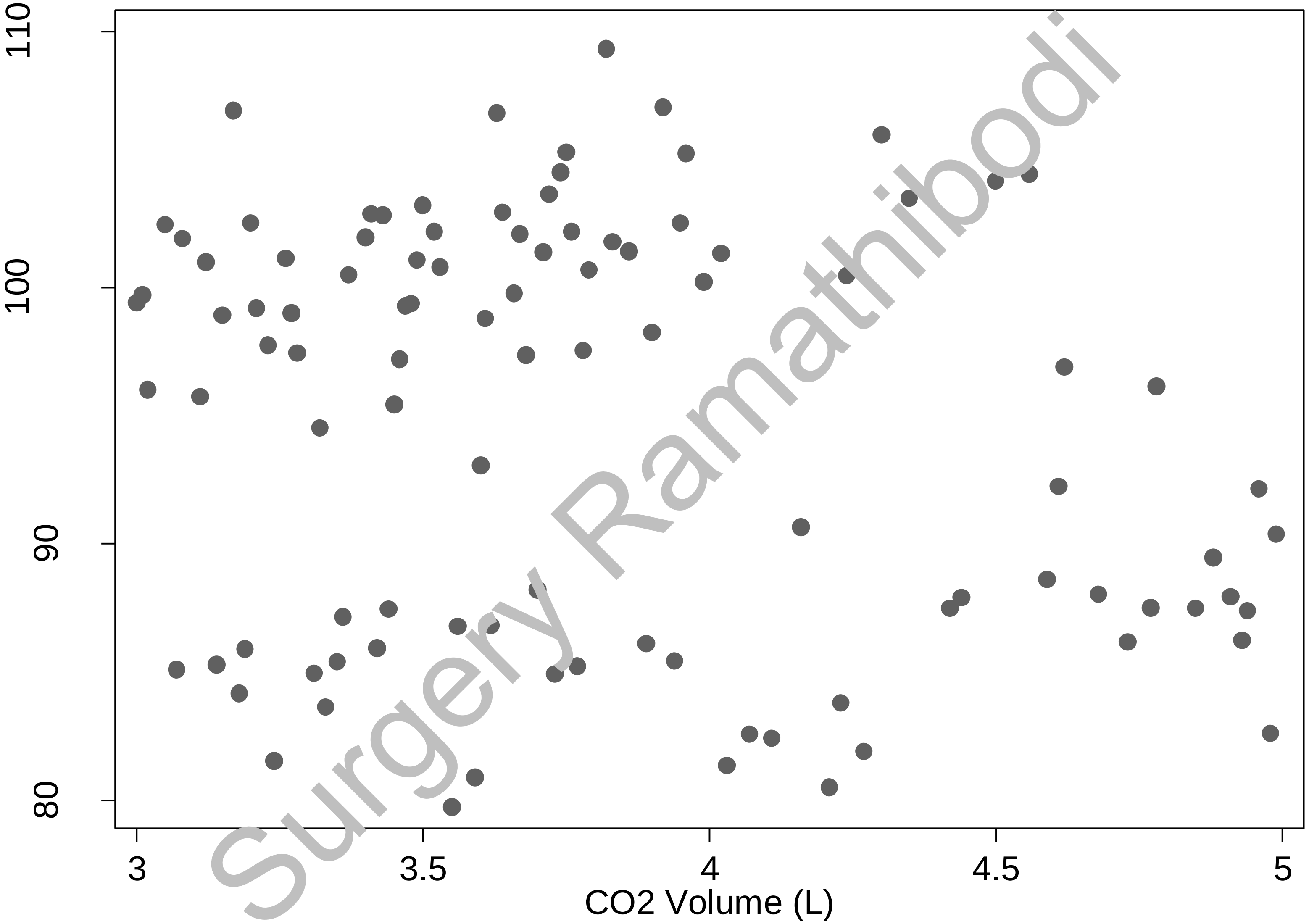
The Outcome Y may be affected by many factors $X_1, X_2, \dots$

Concentrating only on *one factor* X_1 without consideration of others may lead to erroneous conclusions in at least 2 ways:

- The prediction of Y **may not be accurate**
- The effect of X_1 on Y may depend on other $X's$ as well (***Confounding***); by ignoring this, the estimation of effect of of X_1 on Y **may be wrong**

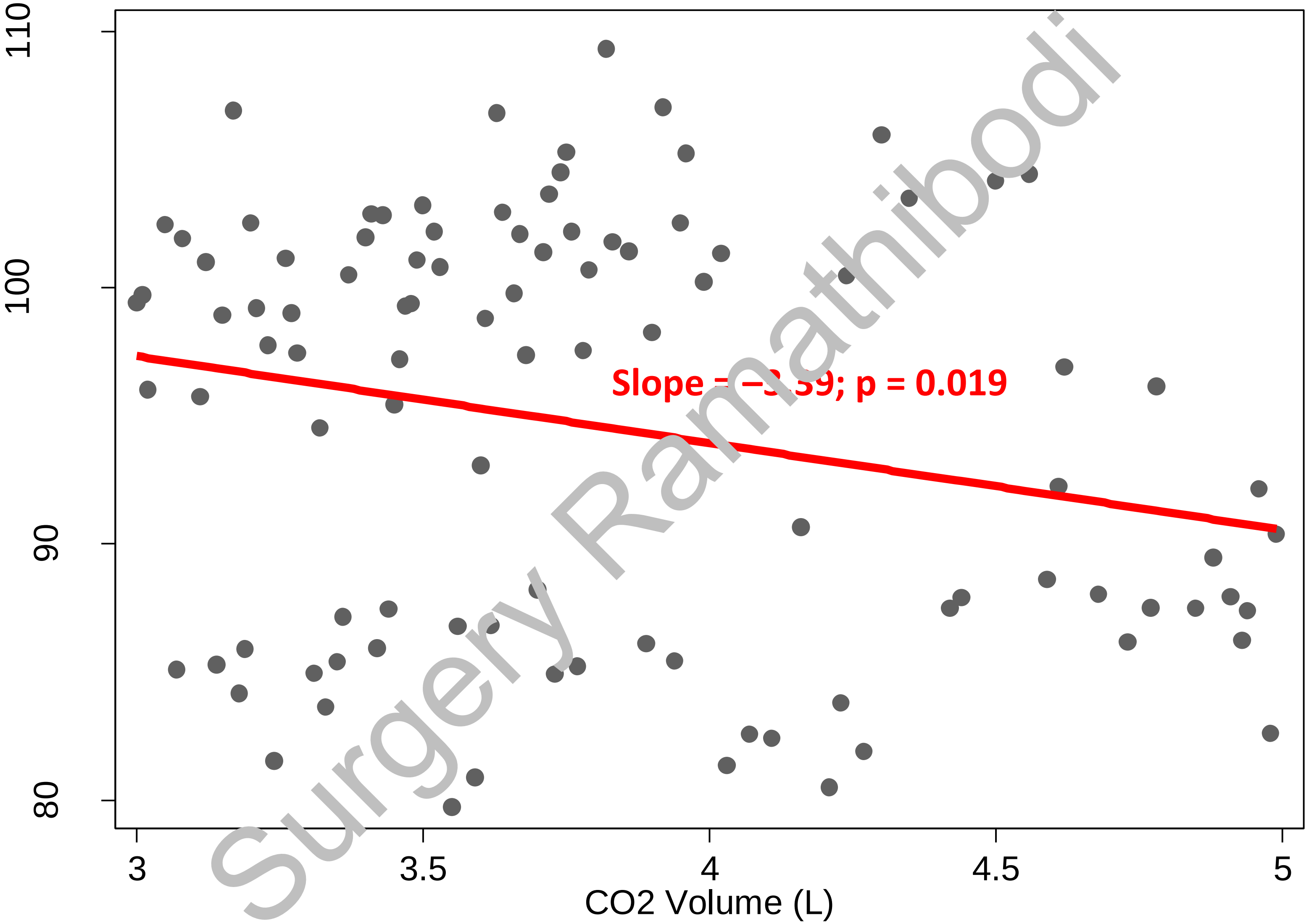
Study comparing MAP and CO_2 insufflation volume in LC

Research Design: prospective cross-sectional study



Scatter Plot
100 subjects

Study comparing MAP and CO_2 insufflation volume in LC



Appropriate Model?

$$Y = \underbrace{\alpha + \beta x}_{\text{Systematic}} + \underbrace{e}_{\text{Random}}$$

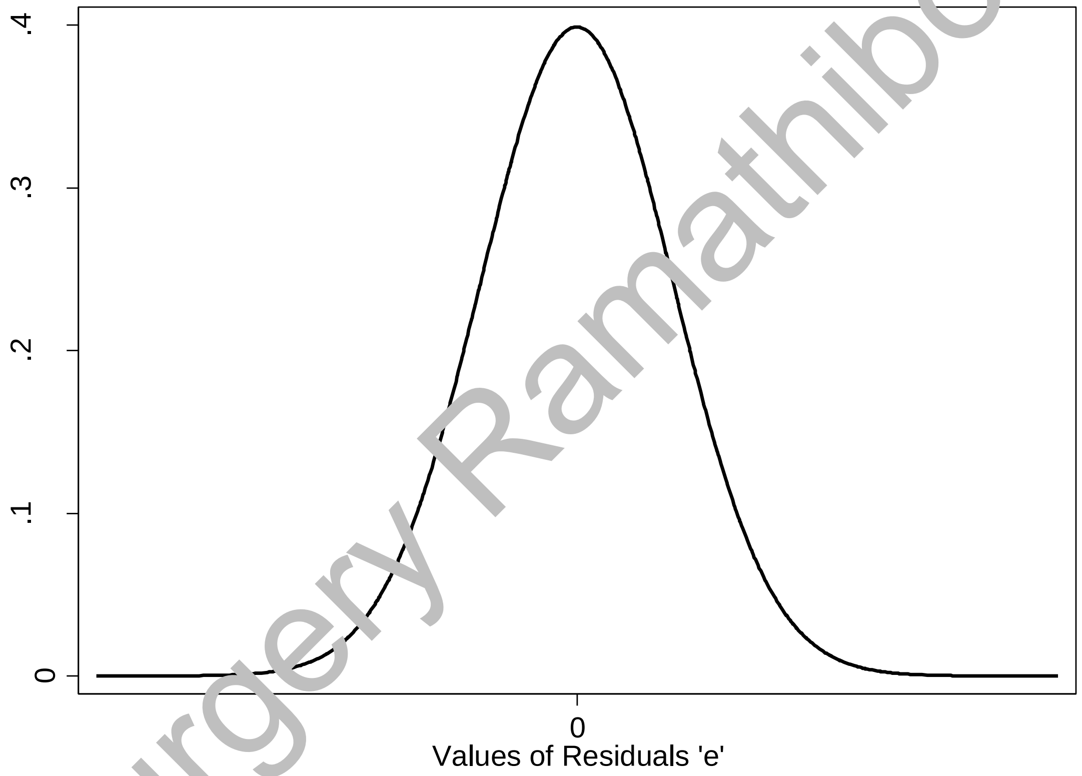
$$MAP|CO_2vol = \alpha + \beta_1 CO_2vol + e$$

$$E(MAP|CO_2vol) = \alpha + \beta_1 CO_2vol$$

$$e \sim N(0, \sigma^2)$$

σ^2 = Residual variance

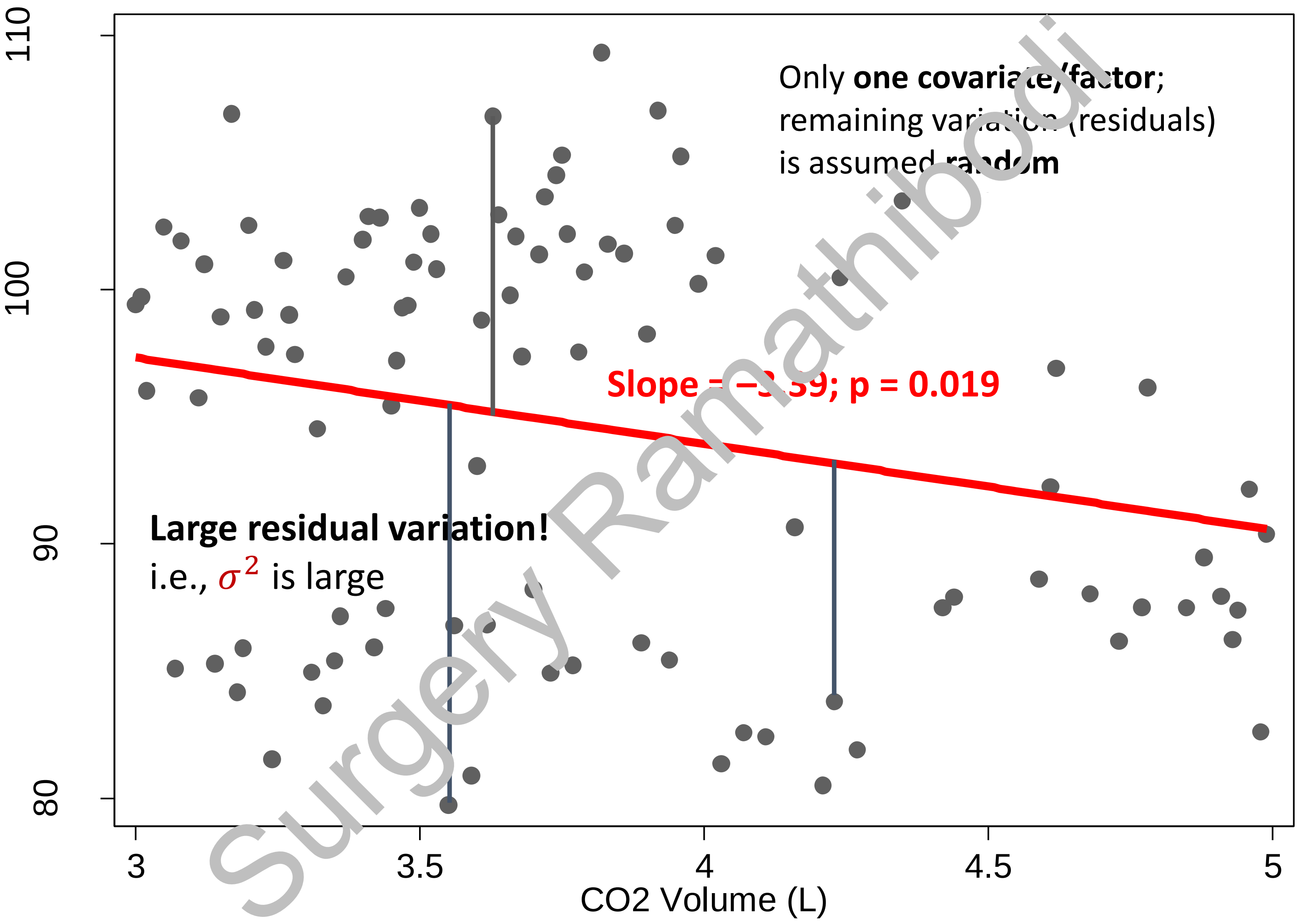
Appropriate Model?



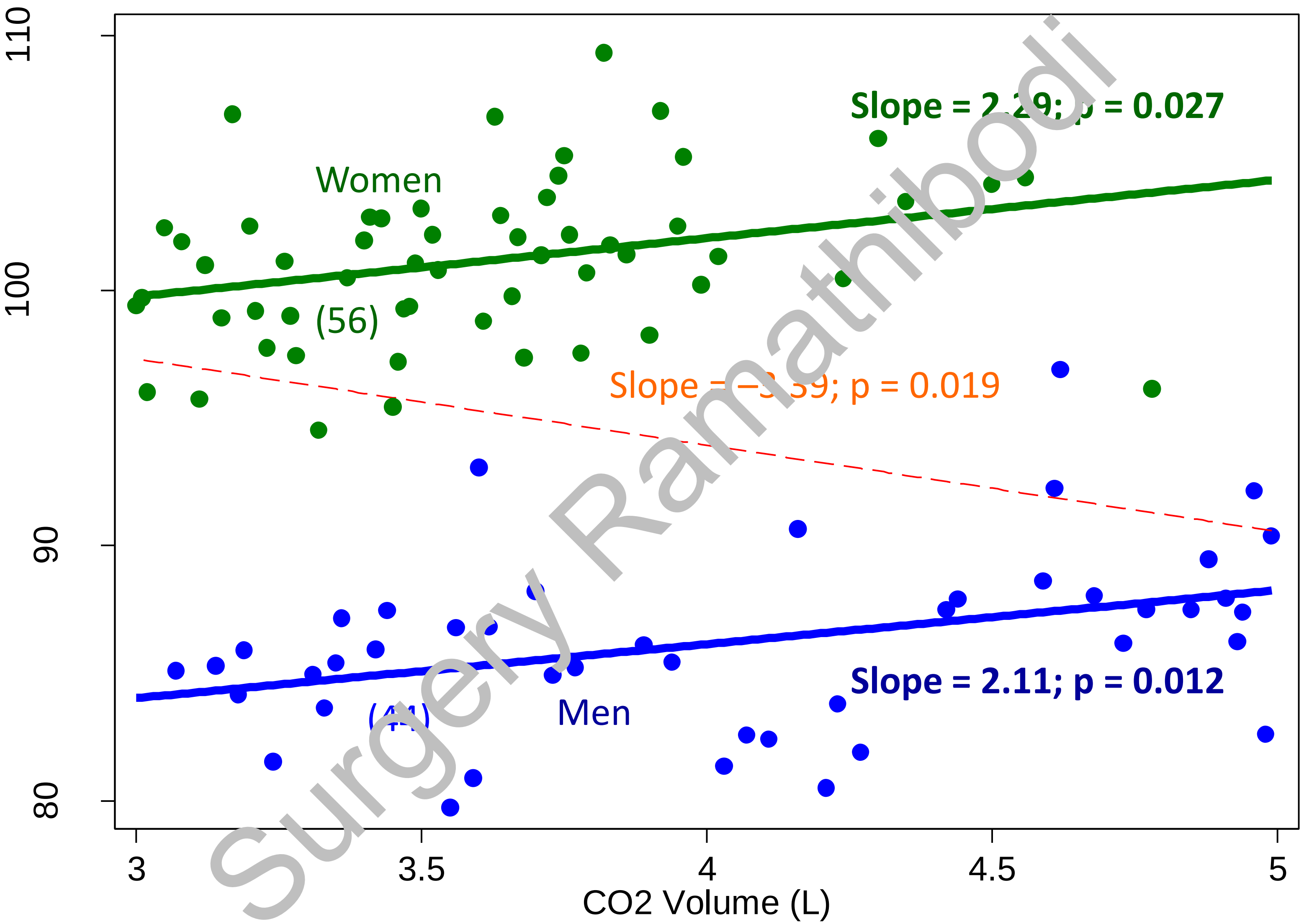
$$e \sim N(0, \sigma^2)$$

Although extremely important, we ignore modeling the random component of the statistical model for now, as the issue is more difficult to explain

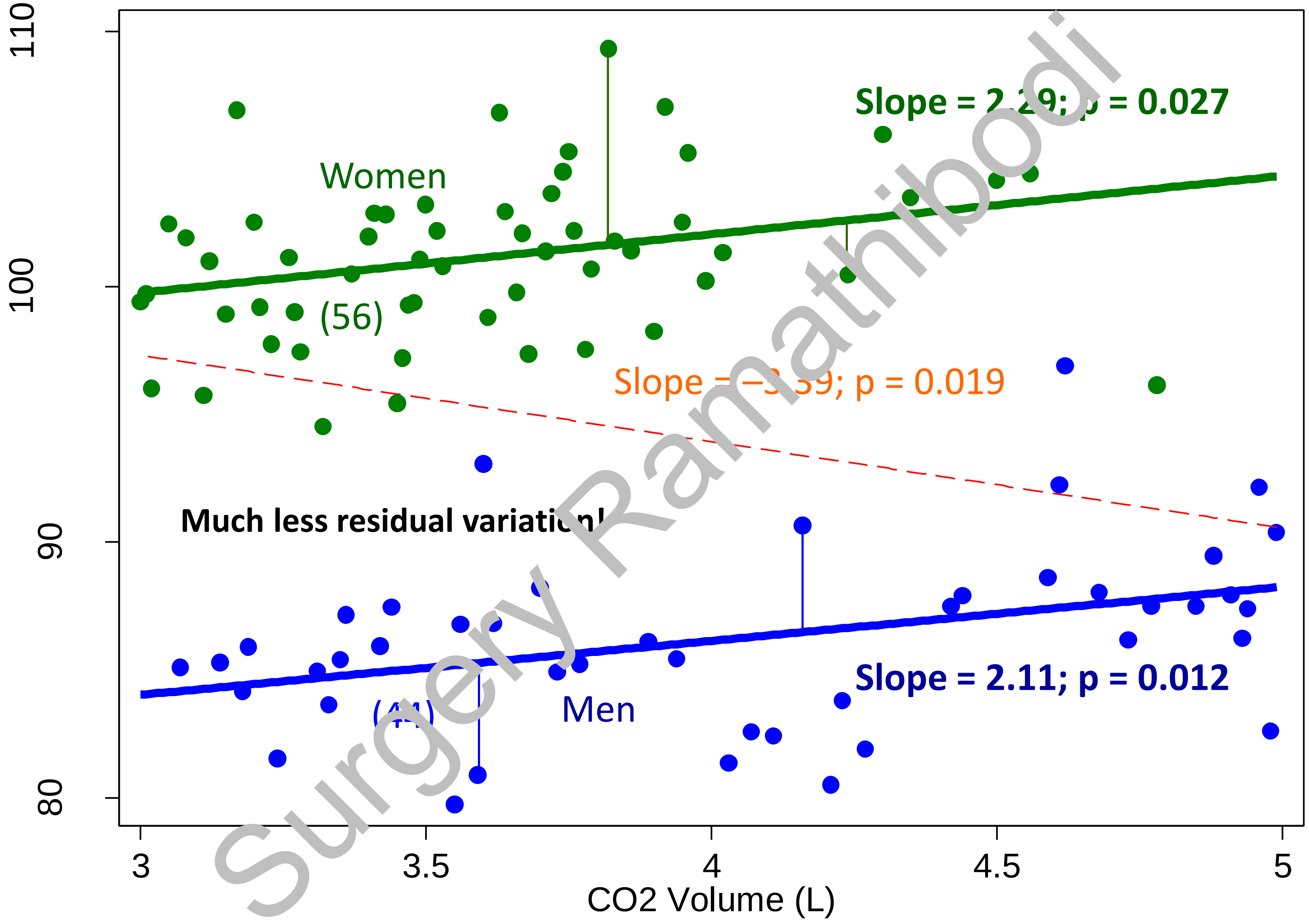
Study comparing MAP and CO_2 insufflation volume in LC



Study comparing MAP and CO_2 insufflation volume in LC



Study comparing MAP and CO_2 insufflation volume in LC



Linear Regression Analysis

Covariates	Univariable Analysis		Multivariable Analysis	
	Mean Effect (95% CI)	**P-value**	**Mean Effect (95% CI)**	**P-value**
CO_2 volume	**–3.39** (–6.22 to –0.56)	0.019	**2.18** (0.95 to 3.40)	0.001
Gender (men)	–14.9 (–16.3 to –13.6)	< 0.001	–15.9 (–17.3 to –14.5)	< 0.001

For every 1 L of CO2 volume increase, a 3.39 mmHg **reduction** in MAP is expected **if everything else is unknown**;

For every 1 L of CO2 volume increase, a 2.18 mmHg **increase** in MAP is expected **after adjustment for effect of gender**

“Simpson’s Paradox”

Uni- vs. Multi-variable Analysis*

Univariable Analysis

107.5 **-3.39**

$$E(MAP) = \alpha + \beta_1 CO_2vol$$

$$E(MAP) = \alpha + \beta_2 Gender$$

Men = 1; Women = 0

101.2 -14.9

Multivariable Analysis

$$E(MAP) = \alpha + \beta_1 CO_2vol + \beta_2 Gender$$

93.3 **2.18** -15.9

*We drop the *conditional notation* for convenience: thus, $E(MAP)\ is\ E(MAP|CO_2vol)$, etc.

Example 5: Question of Many Covariates (b)

Important risk factors for post-operative infection; a Cohort Study

- Outcome is occurrence of SSI: Y (yes =1, No = 0)*
- Multiple risk factors: $X_1, X_2, \ldots$

The regression equation is

- $E(Y|X_1 = x_1, X_2 \ldots) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots = \pi$
- *Since for binary data π cannot be negative, we use the logistic scale***
- $\log\left(\frac{\pi}{1-\pi}\right) = \theta = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots$
- *This is the multiple logistic regression model*

* Again, formally, the binary values are for the indicator function $I(SSI)$

** There are theoretical reasons for using this parametrization, e.g. the log-odds is the **canonical** or **natural parameter** for Generalized Linear Model (GLM) with the Bernoulli/ Binomial Probability Distribution

Table 1. Characteristics of patients and operations, with and without SSI

Characteristics[a]	Total (n = 458 operations)[b]	Without SSI (n = 423 operations)	With SSI (n = 35 operations)
Age (years): mean (SD)		59.6 (14.4)	57.5 (17.2)
Gender (males)		205 (49%)	18 (51%)
Preoperative stay (days):			
Median (range)		2 (1 to [illegible])	2 (1 to 68)
ASA class			
I		17 (4%)	0
II		182 (43%)	8 (23%)
III		174 (41)	21 (60%)
IV		[illegible] (12%)	6 (17)
V		1 (0)	0
Wound classification			
Clean[d]		37 (9%)	4 (11%)
Clean-contaminated		348 (82%)	23 (66%)
Contaminated		23 (5%)	3 (9%)
Dirty		15 (4%)	5 (14%)
Duration of surgery (minutes):			
Median (range)		120 (30 to 550)	150 (60 to 615)
NNIS index			
0		139 (33%)	1 (3%)
1		203 (48%)	17 (49%)
2		74 (17%)	14 (40%)
3		7 (2%)	3 (9%)
Cancer (yes)		235 (56%)	18 (51%)
Operations on Organs			
Gall bladder		121 (31%)	8 (23%)
Biliary tract		38 (9%)	10 (29%)
Colon		205 (49%)	15 (43%)
Liver		35 (8%)	2 (6%)
Pancreas		16 (4%)	0
Older operating theaters (> 12 yrs)		282 (67%)	25 (71%)

SSI rate:
35/458=0.076

J Med Assoc Thai 2007; 90 (7): 1356-62

Testing for Significant Difference

Finding important factors related to SSI can be done by looking for significant differences between patients *with* SSI and those *without* SSI *in terms of these factors*

- We can do this one by one – "univariable" analysis
- In terms of regression models, for factor X_1:
- $\log\left(\frac{\pi}{1-\pi}\right) = \theta = \alpha + \beta_1 x_1$
- By testing for $\beta_1 = 0$ in a univariable logistic regression model
- If we *decide* $\beta_1 \neq 0$, then factor X_1 is significantly related to SSI

Interpretation of β in Logistic Regression

Consider a binary factor X (1 or 0)

- $\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$ can be written as 2 equations:
- $\log\left(\frac{\pi_1}{1-\pi_1}\right) = \alpha + \beta$ if $X = 1$
- $\log\left(\frac{\pi_0}{1-\pi_0}\right) = \alpha$ if $X = 0$; thus
- $\log\left(\frac{\pi_1}{1-\pi_1}\right) - \log\left(\frac{\pi_0}{1-\pi_0}\right) = \beta$; or $\log\left(\frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}\right) = \beta \equiv \log(OR)$; or
- $OR = e^{\beta}$: the **Odds Ratio**

Example 5: cont.

- Factors identified to be important on **univariable analysis** are inserted into a **multivariable analysis** to obtain a final multivariable model with a set of *independent* risk factors
- This is a commonly used strategy & reasonable in many situations
- Estimates of $\beta' s$ can be obtained by *maximizing the Likelihood function* and using asymptotic properties (point & interval estimates)*:

$$\max_{\beta_1=\widehat{\beta}_1,\beta_2=\widehat{\beta}_2,\ldots} Likelihood(\beta_1,\beta_2,\ldots)$$

*Lachin JM. Biostatistical methods: the assessment of relative risk. 2nd ed. Wiley; 2011

Table: Logistic Regression Models for SSI

Factor	Odds Ratio (95% CI) Univariable	p-value	Odds Ratio (95% CI) Multivariable	P-value
Age (years)	0.99 (0.97, 1.01)	0.416	-	
Gender (m=1; f=0)	1.23 (0.56, 2.24)	0.736	-	
Preop stay (d)	1.03 (0.99, 1.06)	0.162	-	
ASA Class	**1.78 (1.14, 2.84)**	**0.012**	**1.88 (1.15, 3.09)**	**0.013**
Wound Class				
1	**1**	**0.030**	**1**	**0.018**
2	**0.61 (0.20, 1.86)**		**0.94 (0.28, 3.14)**	
3	**1.21 (0.25, 5.89)**		**1.65 (0.31, 8.90)**	
4	**3.03 (0.73, 13.1)**		**5.92 (1.21, 28.9)**	
Duration of Surg (hr)	**1.42 (1.20, 1.67)**	**<0.001**	**1.47 (1.23, 1.76)**	**<0.001**
Cancer (yes)	0.85 (0.42, 1.69)	0.637	-	

N = 458

A "Predictive/Prognostic Score"

- From $\log\left(\frac{\pi}{1-\pi}\right) = \theta = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots$
- Construct a "score" for SSI = $\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots$ "Linear predictor"
- In this case*: $S = -6.88 + 0.622\mathrm{ASA} + 0.698\mathrm{Wound} + 0.394\mathrm{Optime}$
- The risk of SSI is calculated as

$$Risk = \frac{1}{1 + e^{-S}}$$

- *Example*: a patient with ASA = 2; Wound class = 2; op time = 1.5 hr
- $S = -3.651$ $\quad Risk = \frac{1}{1+e^{3.651}} = 0.025$

*For this scoring system, Wound class was entered as a quantitative variable

Extensions of Statistical Models

The idea of regression modeling extends to more complicated data structures or generating mechanisms (and alternatives):

- Other categorical outcomes: > 2 categories; ordinal outcomes
- Correlated and longitudinal data: multiple measurements on 1 person
- Robust regression
- Non-parametric regression
- Bayesian version of regression models
- etc.