Os Fundamentos do Modelo de Regressão

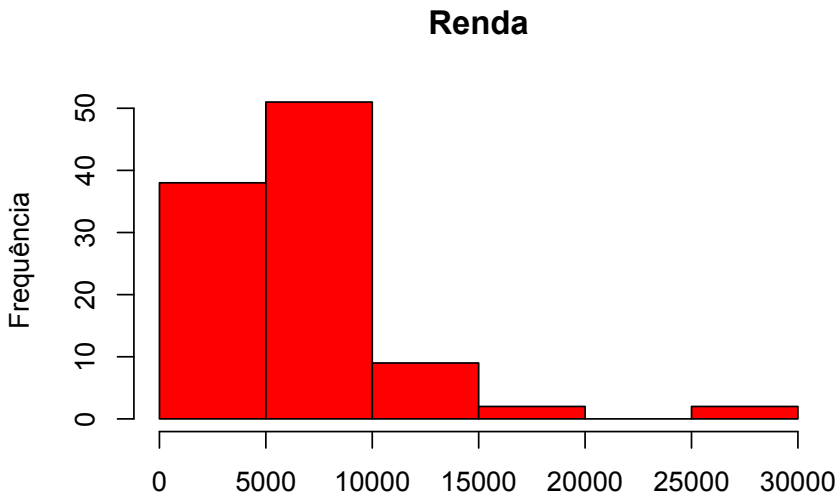
Ricardo Ceneviva
ceneviva@usp.br

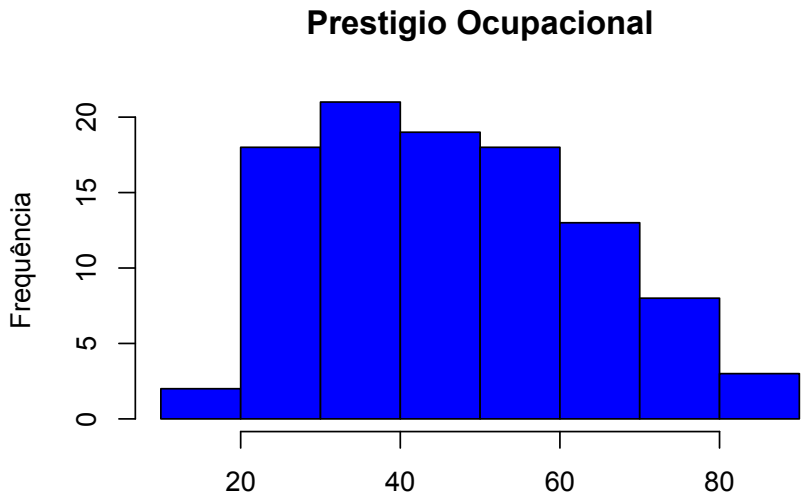
Universidade de São Paulo
Departamento de Ciência Política

12 de dezembro de 2012

Sobre o que vamos conversar hoje?

Correlação Linear

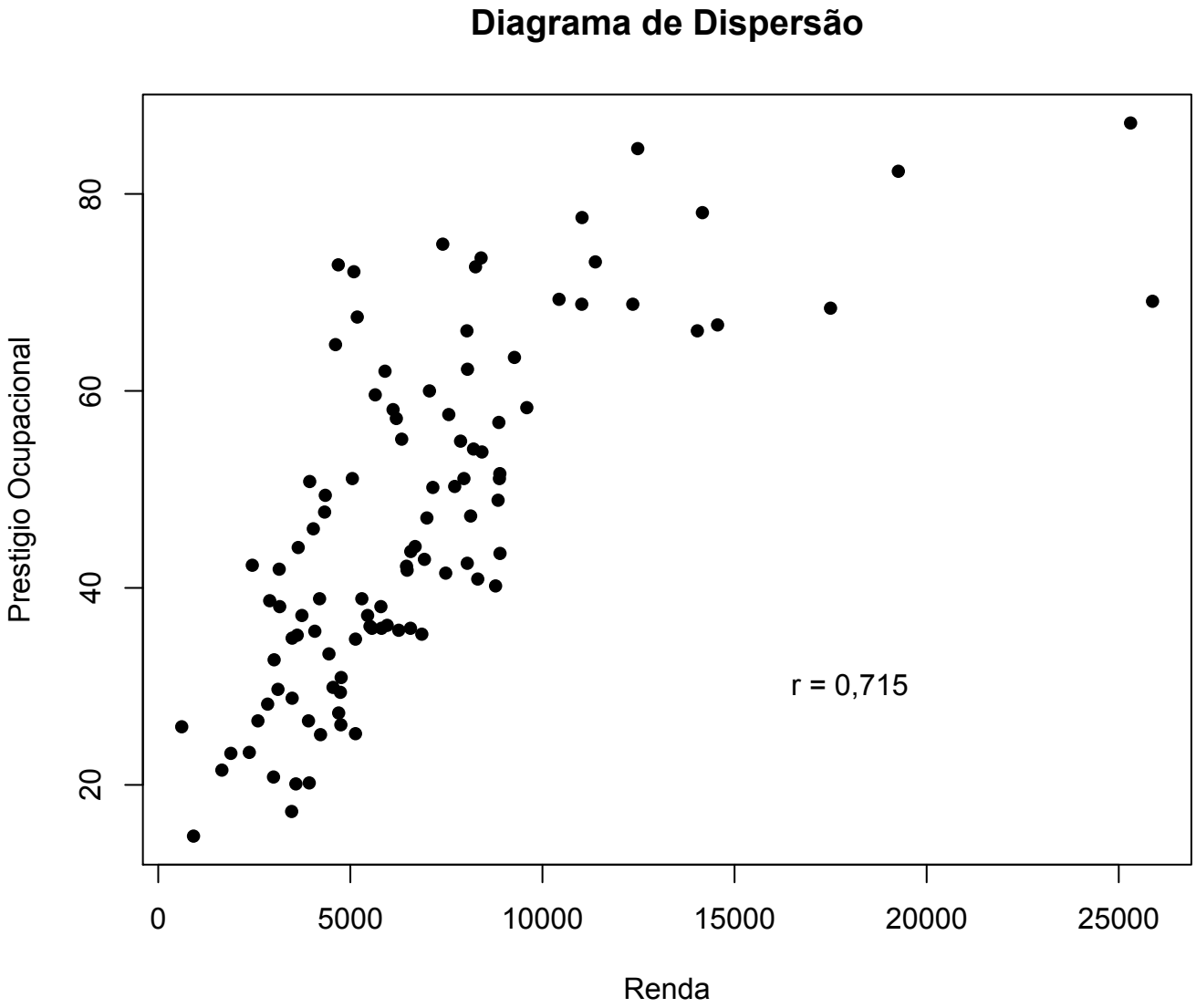




Correlação Linear

- Procedimento estatístico usado para medir e descrever a relação entre duas variáveis intervalares
- A correlação mede a força, ou grau, de relacionamento entre duas variáveis intervalares
- O coeficiente de correlação (r) é uma estatística amostral, nunca deve ser usada para se referir indivíduos (isolados) da amostra

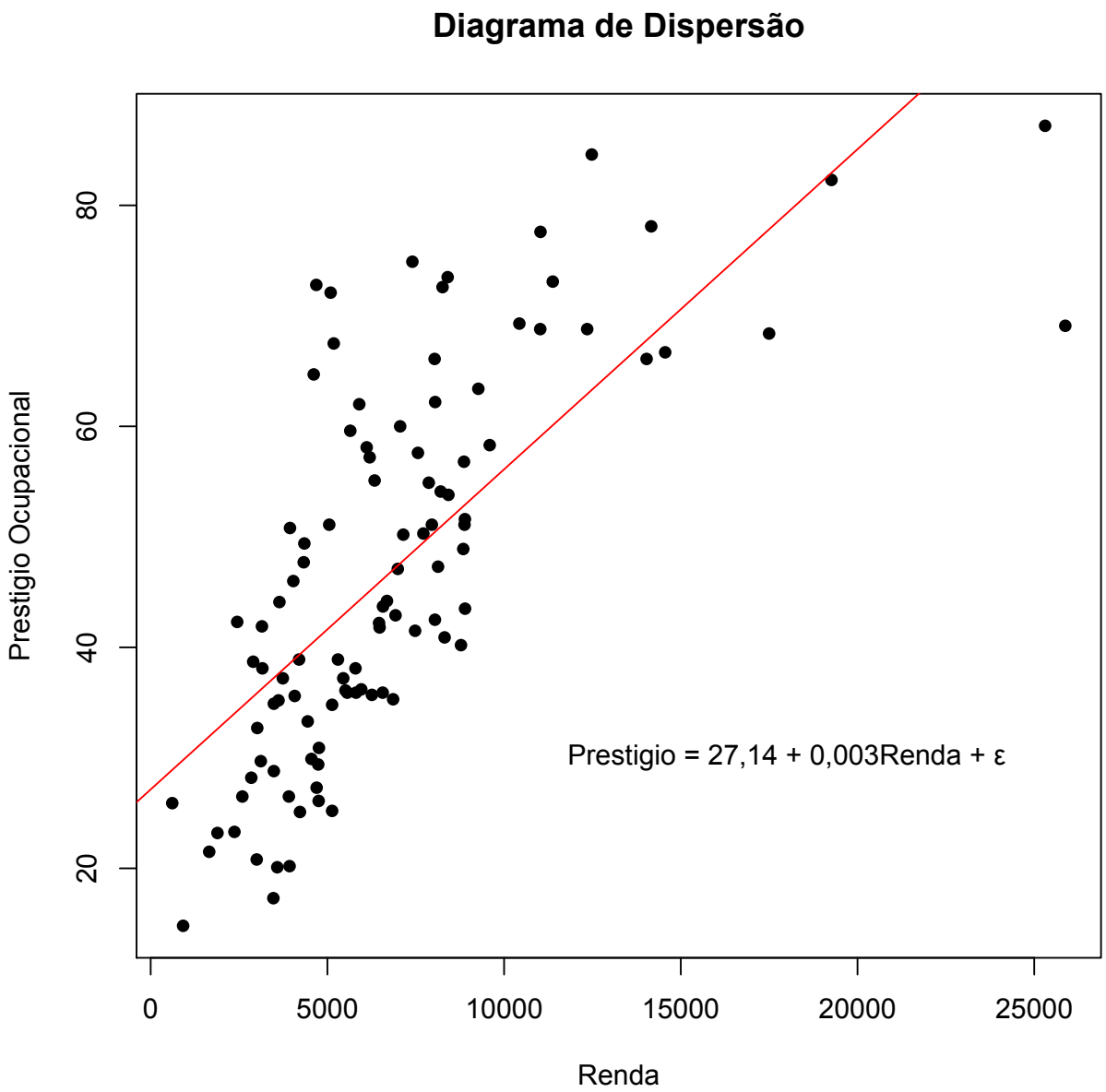
R scatterplot: plot(prestige, income)



Correlação Linear

- Precisão das predições dependerá da força (isto é, da magnitude) da correlação
- A magnitude da correlação é muito sensível a confiabilidade da mensuração de X e Y
- A validade das predições dependerá da validade de X e Y

Correlação Linear e Regressão Linear



Regressão Linear Simples

- A regressão linear simples constitui uma tentativa de estabelecer uma equação matemática linear (linha reta) que descreva o relacionamento entre duas variáveis.

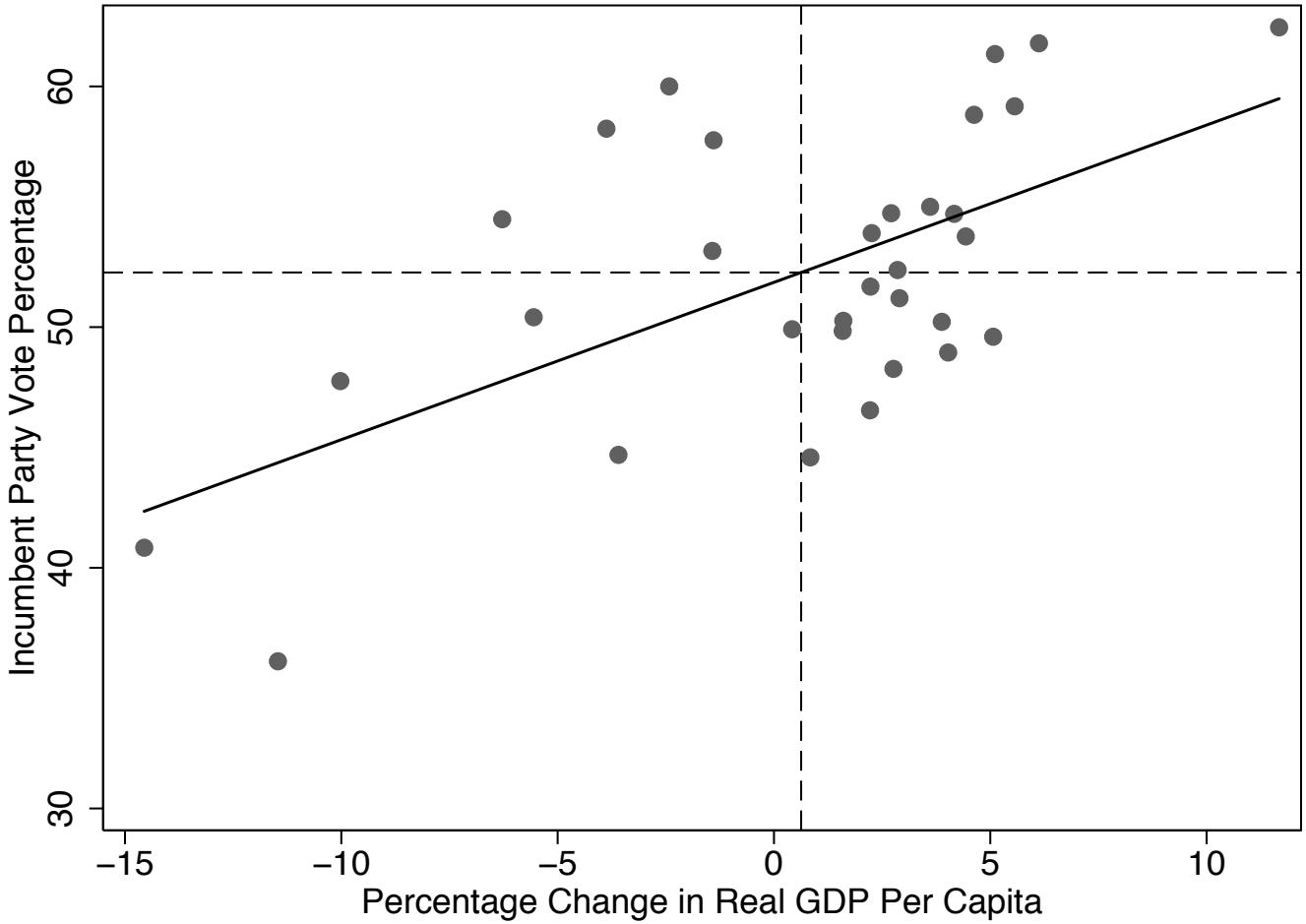
- Da mesma forma como usamos a média para resumir uma variável aleatória, a reta de regressão é usada para resumir a estimativa linear entre duas variáveis aleatórias (Freeman, 1997, p.344).

Regressão Linear Simples

- Modelo (linear): de explicação da relação (de causalidade) entre duas ou mais variáveis, Como se pode explicar o comportamento (i.e. a variação) de uma variável com base em valores conhecidos da outra: Modelo do Voto Econômico

- Predizer valores futuros de uma variável. Ex. aplicar testes para avaliar o sucesso de um ingressante na escola ou no emprego ou mesmo tentar prever as chances de sucesso eleitoral de um presidente dado o nível de crescimento do PIB naquele ano.

Regressão Linear: Modelo do Voto Econômico



A Equação Linear (a reta de regressão)

Principais características do modelo de regressão linear:

1. O coeficiente angular da reta é dado pela tangente da reta e se denomina ß (beta).
2. A cota da reta em determinado ponto é o coeficiente linear denominado (alfa), que é o valor de Y quando X=0.

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \tag{1}$$

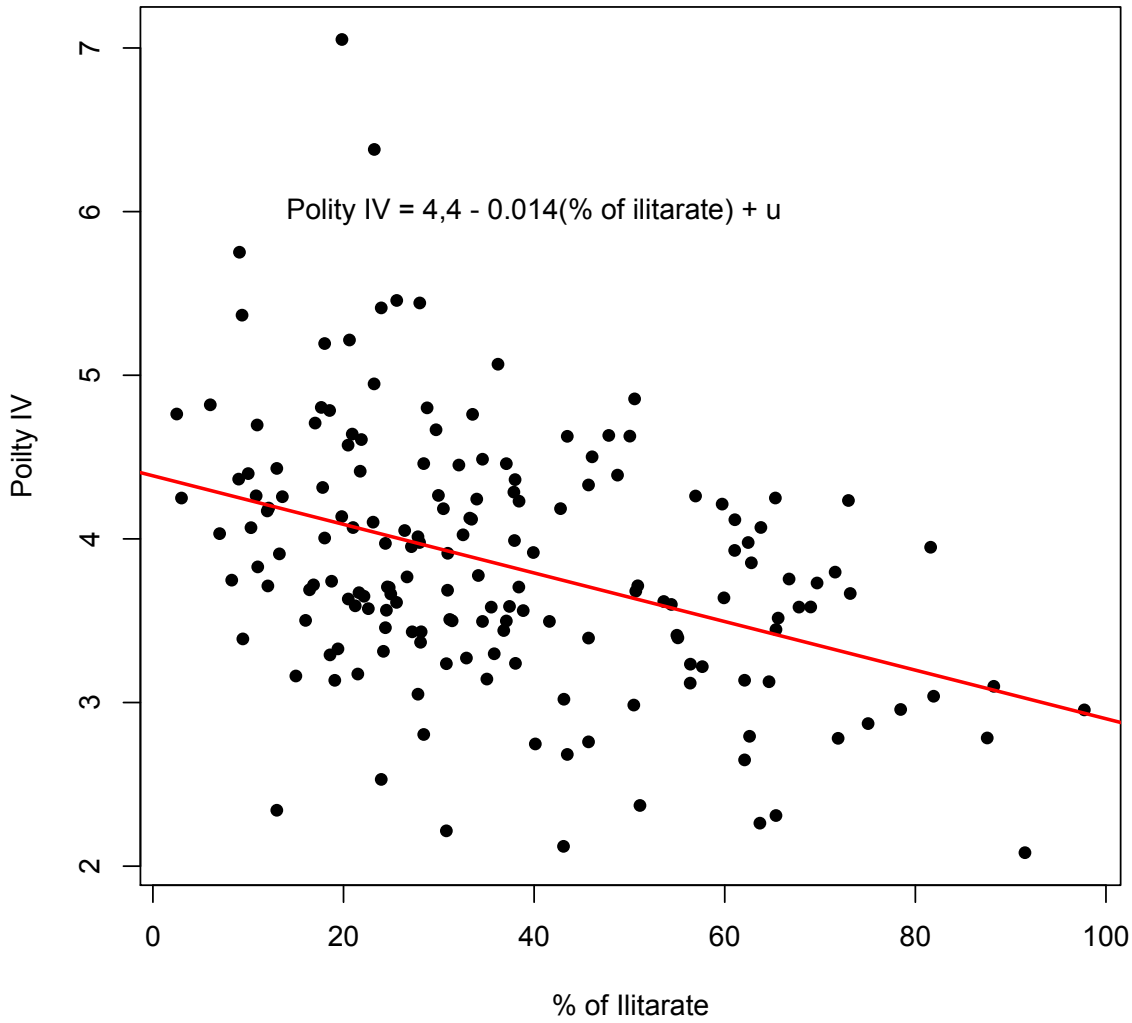
Regressão Linear

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \tag{2}$$

Nesse modelo se verifica que:

1. Para um valor Xi podem existir um ou mais valores de Yi amostrados.
2. Para esse mesmo valor Xi se terá apenas um valor projetado .
3. Para cada valor de Xi existirá um desvio di (ou erro ei) dos valores de , conforme indicado nas figuras da apresentação.
4. Sempre teremos observações que não são pontos da reta.

Regressão Linear

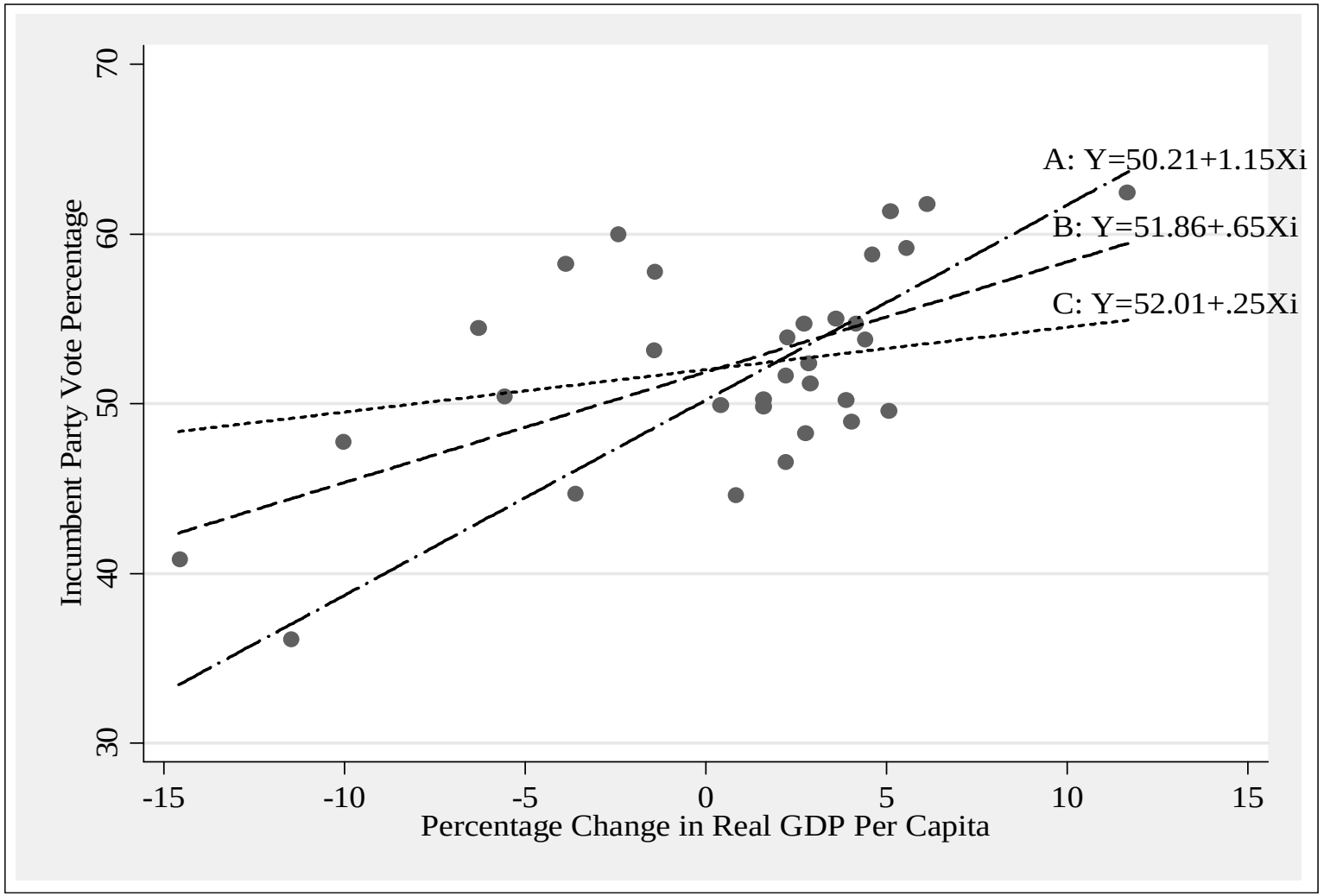


A Determinação da Equação Matemática

- Na regressão, os valores y são preditos com base em valores dados ou conhecidos de x. A variável y é chamada variável dependente, e a variável x, variável independente.

- Que critério devemos aplicar para obter os valores dos coeficientes Alfa e Beta?

Regressão Linear: Modelo do Voto Econômico



A Determinação da Equação Matemática

Existem dois critérios:

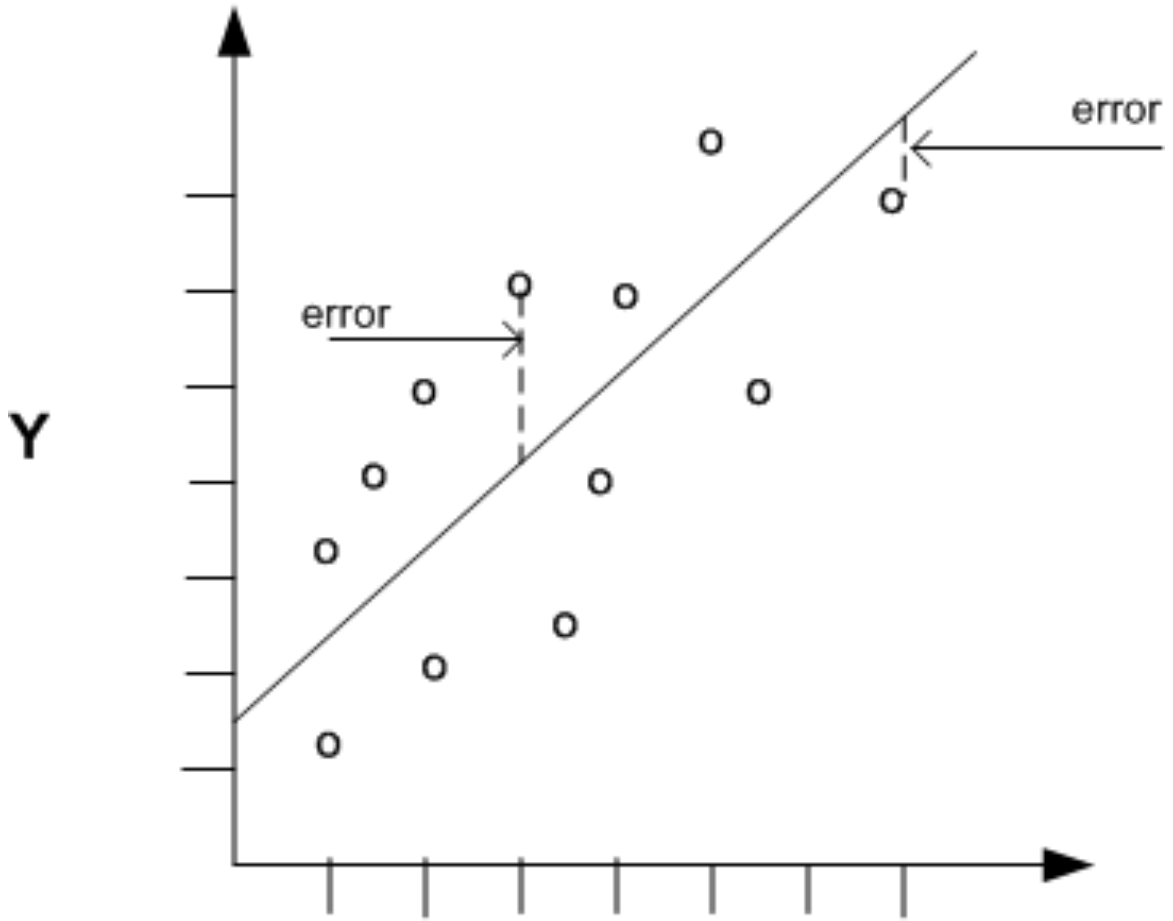
- Ajustar um reta horizontal de valor igual à média dos valores de y, pois a média é uma reta de regressão com Beta = 0.

- Ajustar um reta que divida os pontos observados de forma que a soma dos desvios seja nula. No entanto, a simples soma dos desvios leva à compensação dos desvios positivos e negativos, como já se viu no cálculo da variância.

O Método dos Mínimos Quadrados Ordinários (OLS)

- O critério é encontrar os coeficientes Alfa e Beta da reta de regressão que minimizem a soma dos quadrados dos desvios.
- A soma dos desvios verticais dos pontos em relação à reta é zero
- A soma dos quadrados desses desvios é mínima (isto é, nenhuma outra reta daria menor soma de quadrados de tais desvios).

O Método dos Mínimos Quadrados Ordinários (OLS)



O Método dos Mínimos Quadrados Ordinários (OLS)

- Simbolicamente, o valor que é minimizado é:

$$\sum d_i^2 = \sum \left(y_i - y_c \right)^2$$

- Onde:
- yi = valor observado de y
- yc = o valor calculado de y utilizando-se a equação de mínimos quadrados com o valor de x correspondente a yi.

O Método dos Mínimos Quadrados Ordinários (OLS)

- Os coeficientes são calculados pelas fórmulas abaixo.

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

O Método dos Mínimos Quadrados Ordinários (OLS)

- Tendo presente que

$$Cov(x, y) = rxy * \sigma x * \sigma y$$

, o coeficiente b será igual a estas quatro fórmulas possíveis:

$$b = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{Cov(x, y)}{Var(x)} = \frac{r_{xy}\sigma_X\sigma_Y}{\sigma_X^2} = r_{xy}\frac{\sigma_Y}{\sigma_X}$$

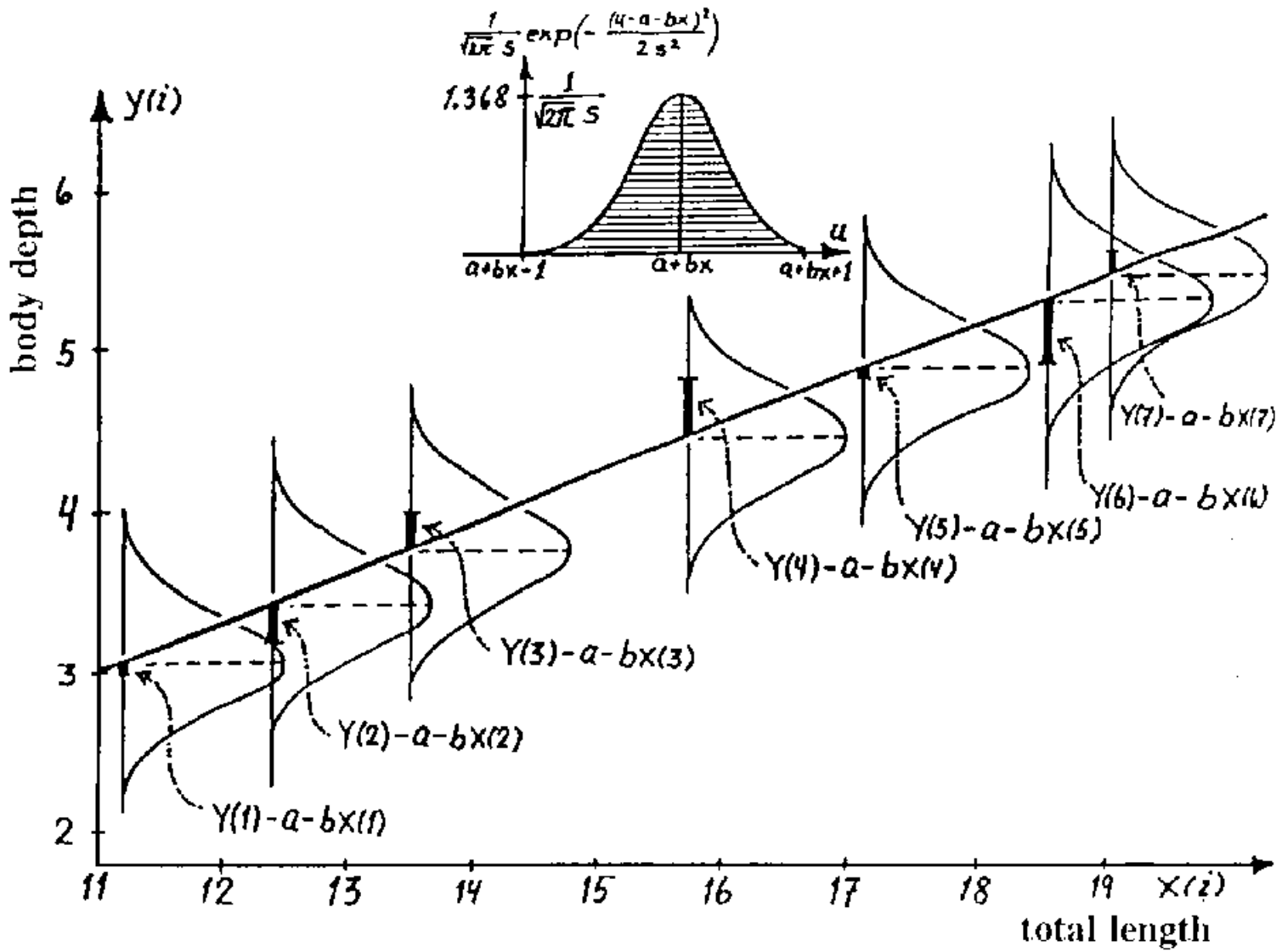
$$a = \frac{\sum y - b\sum x}{n} = \bar{Y} - b\bar{X}$$

O Método dos Mínimos Quadrados Ordinários (OLS)

Fatos importantes da equação de regressão:

- Trata-se de uma média
- Seria arriscado extrapolar essa equação para fora do âmbito dos dados
- A reta de regressão tem a interessante propriedade de passar sempre pelo ponto (x, y).

O Método dos Mínimos Quadrados Ordinários (OLS)



O Método dos Mínimos Quadrados Ordinários (OLS)

$$r^2 = \frac{SS_{(regression)}}{SS_{(total)}}$$

J. WOOLDRIDGE - CAPÍTULO 3: ANÁLISE DE REGRESSÃO MÚLTIPLA: ESTIMAÇÃO

O Modelo de Regressão Linear Múltipla

Perguntas Frequentes:

- Como nunca há uma relação exata entre duas variáveis, como consideramos outros fatores que afetam y?
- Qual é a relação funcional entre y e x?
- Como podemos estar certos de que estamos capturando uma relação ceteris paribus (outros fatores constantes) entre y e x?

Os Pressupostos do Método de MQO

1. A relação entre a variável dependente e as variáveis independentes deve ser linear;
2. as variáveis foram medidas adequadamente, ou seja, assume-se que não há erro sistemático de mensuração;
3. a expectativa da média do termo de erro é igual a zero;
4. homocedasticidade, ou seja, a variância do termo de erro é constante para os diferentes valores da variável independente;
5. ausência de autocorrelação, ou seja, os termos de erros são independentes entre si;

Os Pressupostos do Método de MQO

1. a variável independente não deve ser correlacionada com o termo de erro;
2. nenhuma variável teoricamente relevante para explicar Y foi deixada de fora do modelo e nenhuma variável irrelevante para explicar Y foi incluída no modelo;
3. as variáveis independentes não apresentam alta correlação, o chamado pressuposto da não multicolinearidade;
4. assume-se que o termo de erro tem uma distribuição normal; e
5. há uma adequada proporção entre o número de casos e o número de parâmetrosestimados.

A linearidade dos parâmetros

- Esperara-se que a relação entre as variáveis independentes e a variável dependente possa ser representada por uma função linear

- Na estimação do modelo, a linearidade implica que o aumento de uma unidade em X1 gera o mesmo efeito sobre Y, independente do valor inicial de X1 (Wooldridge, 2009)

A Mensuração das Variáveis

- Para Tabachnick e Fidell (2007), "a análise de regressão assume que as variáveis são medidas sem erro, uma clara impossibilidade em muitas pesquisas nas ciências sociais e comportamentais" (Tabachnick e Fidell, 2007: 122)

- De acordo com Lewis-Beck (1980), a importância de incluir variáveis bem medidas no modelo é evidente: variáveis mal medidas produzirão estimativas inconsistentes. Em particular, se as variáveis independentes são medidas com erro, as estimativas (intercepto e coeficiente de regressão) serão viesadas.

O Termo Aleatório de Erro (u)

- A violação desse pressuposto compromete a consistência da estimativa do intercepto.
- Dessa forma, enquanto o coeficiente de regressão (slope) não é afetado, o pesquisador deve ter cuidado com a interpretação substantiva da constante.
- Para Kennedy (2009), "o erro pode ter uma média diferente de zero devido a presença de erros de mensuração sistematicamente positivos ou negativos no cálculo da variável dependente" (Kennedy, 2009: 109).

Homocedasticidade

- A homogeneidade da variância é um pressuposto central do modelo de regressão de mínimos quadrados ordinários. Os resíduos, ou seja, a diferença entre os resultados observados e os resultados preditos pelo modelo devem variar uniformemente.

- Hair et al (2009) afirmam que "homocedasticidade refere-se ao pressuposto de que a variável dependente exibe níveis iguais de variância em toda a gama de variável preditora. Homocedasticidade é desejável porque a variância da variável dependente a ser explicada na relação de dependência não deve ser concentrada em apenas uma gama limitada de valores independentes"(Hair et al, 2009: 83).

- Para Lewis-Beck (1980), "violar a suposição da homocedasticidade é mais grave. Isso porque mesmo que as estimativas dos mínimos quadrados continuem a ser não-viesados, os testes de significância e intervalos de confiança estariam errados" (Lewis-Beck, 1980: 28).

A Ausência de Autocorrelação entre os Casos

- O valor de uma observação medida em determinado período (t1) não influencia o valor de uma observação medida em um momento posterior (t2).
- Significa dizer que as observações são independentes, ou seja, que não existe correlação entre os termos de erro.
- Enquanto os valores dos coeficientes permanecem não-viesados, tem-se problemas na confiabilidade dos testes de significância e intervalos de confiança.

A Ausência de Correlação as Variáveis Independentes e o Termo de Erro

- Para Lewis-Beck (1980) é difícil satisfazer esse pressuposto em desenhos de pesquisa não experimentais.
- Se, por exemplo, uma variável X1 está correlacionada com outra variável explicativa X2, mas o pesquisador não incluir está última em seu modelo, as estimativas serão viesadas.
- Como o pesquisador não pode manipular o valor da variável independente, é importante que todas as variáveis teoricamente importantes sejam incorporadas ao modelo explicativo.

A Especificação Adequada do Modelo

- Primeiro, todas as variáveis independentes teoricamente relevantes devem ser incluídas na equação de regressão.

- Segundo, nenhuma variável teoricamente irrelevante deve ser incluída no modelo já que isso produz ineficiência nos estimadores, aumentando o erro padrão da estimativa.

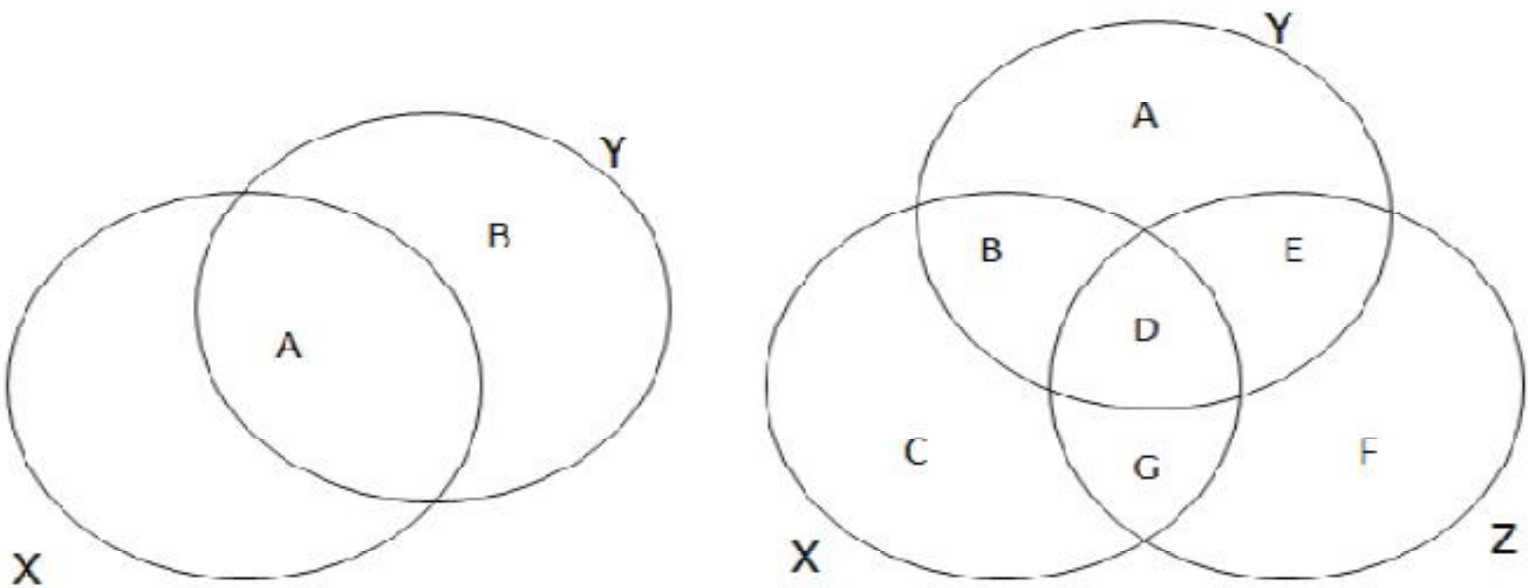
Ausência de Multicolinearidade

- Kennedy (2009) argumenta que "o estimador OLS na presença de multicolinearidade permanece não viesado e, de fato, ainda é o melhor estimador linear não viesado (BLUE)

- A maior dificuldade de modelos com problemas de multicolinearidade é o aumento da magnitude da variância dos parâmetros estimados. Isso porque a presença de altos níveis de correlação entre as variáveis independentes impossibilita estimar, com precisão, o efeito de cada variável sobre a variável dependente.

Ausência de Multicolinearidade

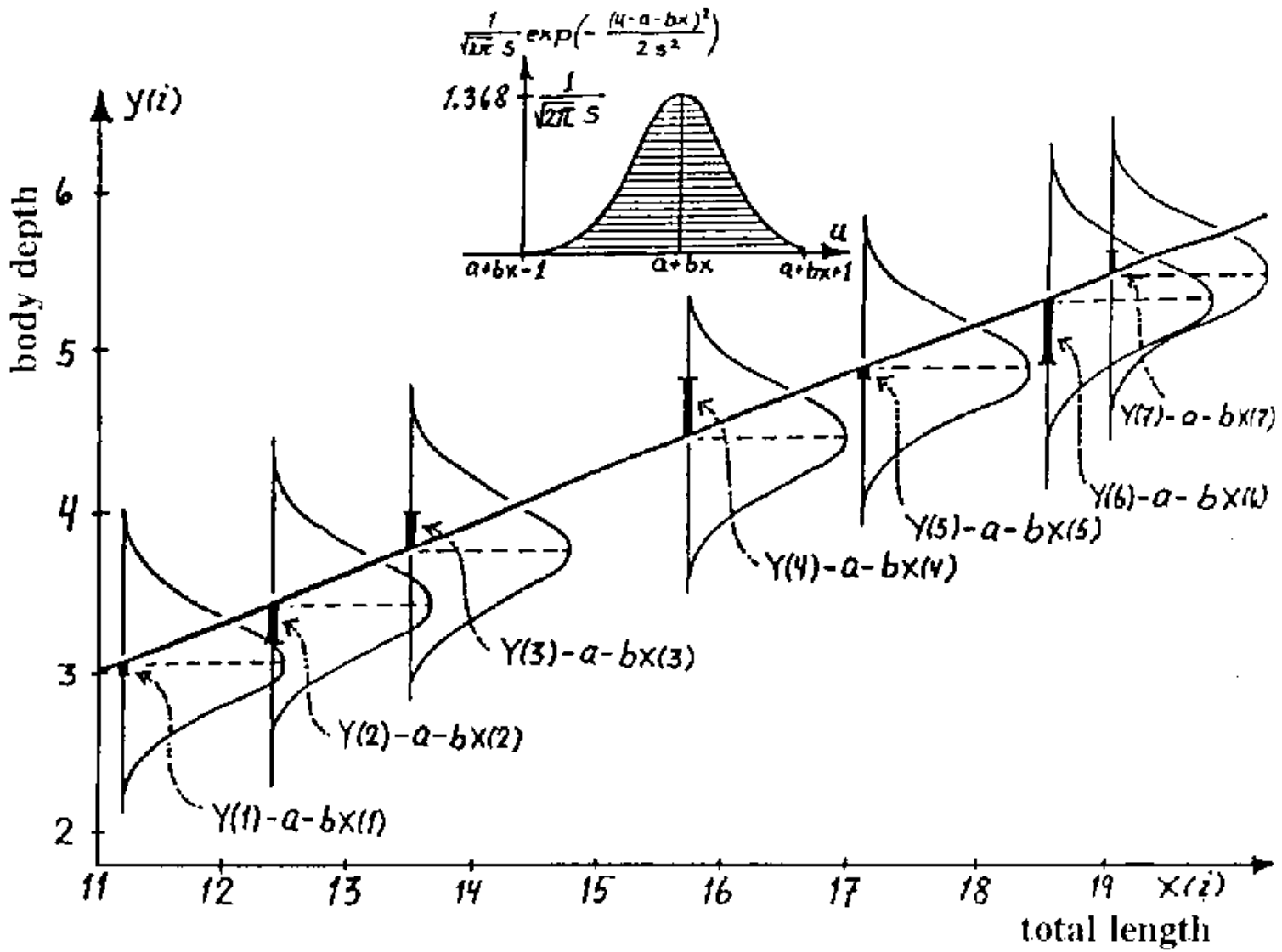
FIGURA 2
Multicolinearidade utilizando diagrama de Vein



A Distribuição do Termo de Erro

- De acordo com as premissas do teorema de Gauss-Markov, o erro amostral deve seguir uma distribuição aproximadamente normal para que os estimadores de Beta1, Beta2 e Sigma encontrados a partir do método de mínimos quadrados ordinários sejam não-viesados e eficientes.

A Distribuição do Termo de Erro



O Número de Casos e de Parâmetros

- Essa é uma condição matemática básica. Como o algoritmo computacional inverte a matriz para realizar os cálculos, caso o número de parâmetros a serem estimados supere a quantidade de observações, a estimação torna-se matematicamente impossível.

- O pesquisador deve maximizar o número de observações em sua análise dada as propriedades desejáveis de amostras grandes. Isso porque a partir do Teorema Central do Limite sabe-se que a distribuição amostral de variáveis aleatórias converge para a distribuição normal quando o tamanho da amostra aumenta.