INTRODUÇÃO À ECONOMETRIA

Análise de Regressão Múltipla

Aula 4

Escola Nacional de Administração Pública

Regressão Linear Múltipla

Análise de Regressão Múltipla

Modelo de Regressão Linear Múltipla

- O **modelo de regressão linear múltipla** é usado para estudar a relação entre uma variável dependente e uma ou mais variáveis independentes.

Análise de Regressão Múltipla
Modelo de Regressão Linear Múltipla

- O **modelo de regressão linear múltipla** com k variáveis toma a seguinte forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k + u \qquad (1)$$

- Onde Y é a variável dependente e X_1, X_2 , . . .,X_k são as variáveis independentes.

- O termo de erro ou pertubação (u) representa outros fatores, além do vetor X, que afetam Y. O termo u também captura erros de medida nas variáveis.

Análise de Regressão Múltipla
Modelo de Regressão Linear Múltipla

- Assume-se que cada observação da amostra (Y_i, X_{i1}, X_{i2} , . . .,X_{ik}), i = 1,…,n, é gerada por um processo descrito por

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_k X_{ik} + u_i \qquad (2)$$

- Assim, o valor observado de Y_i é soma duas partes, uma determinística e uma aleatória (u_i).

Modelo de Regressão Linear Múltipla
Baixando os dados – dados em corte

setwd("C:/diretorio1/diretorio2") – determinar o diretório: Não se esqueça, **inverter** as barras!!!

list.files() – quais arquivos que estão no seu diretório

X<-read.csv("paises.csv",sep=";", dec=".", head=TRUE) – baixando os dados.

Em Data, o X pode estar somente com uma coluna. Isso tem a ver com o decimal.

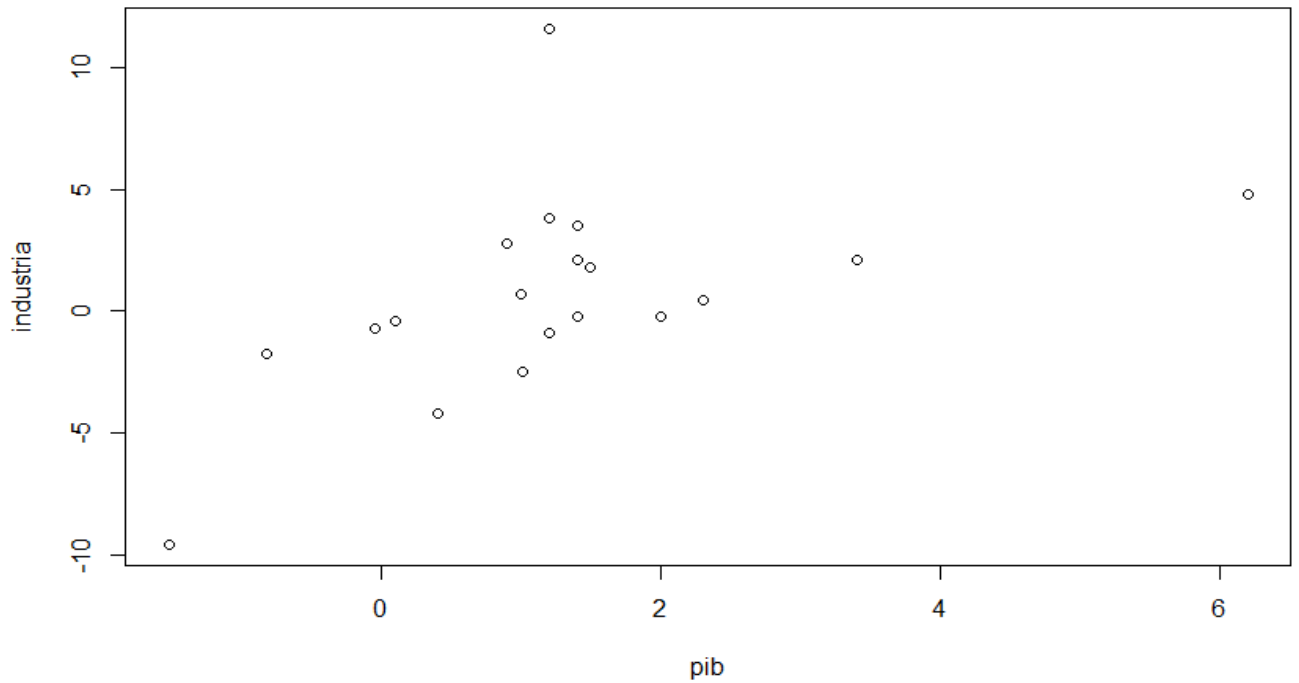
X <Enter> - mostra todos os dados do data.frame

X["PIB"] ou X$PIB

summary(X) ou summary(X$IPC)

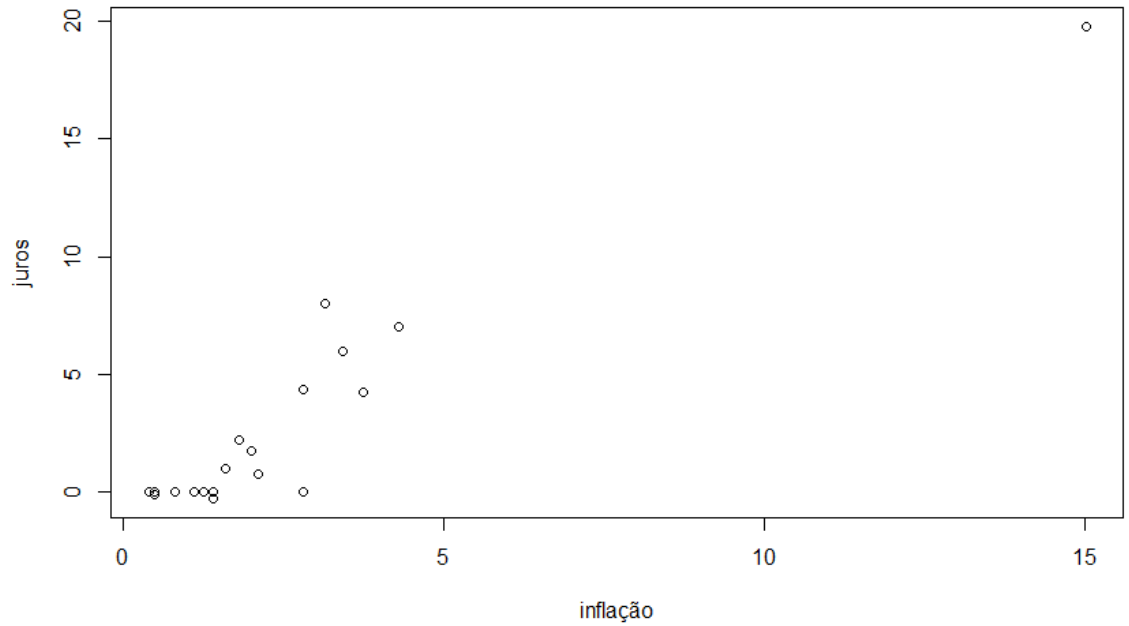
Modelo de Regressão Linear Múltipla
Visualização

```
plot(X$PIB,X$industria,type="p",xlab="pib",ylab = "industria")
```



Modelo de Regressão Linear Múltipla
Visualização

```
plot(X$IPC,X$juros,type="p",xlab="inflação",ylab = "juros")
```



Modelo de Regressão Linear Simples
Estimação

reg1_simples=lm(X$PIB~X$industria)
summary(reg1_simples)

```
Call:
lm(formula = X$PIB ~ X$industria)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2610 -0.7659 -0.1567  0.3413  4.1009

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.13486    0.32983   3.441  0.00312 **
X$industria  0.20088    0.07984   2.516  0.02221 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.417 on 17 degrees of freedom
Multiple R-squared:  0.2713,    Adjusted R-squared:  0.2285
F-statistic:  6.33 on 1 and 17 DF,  p-value: 0.02221
```

Significância!

Modelo de Regressão Linear Simples
Estimação

```
reg2_simples<-lm(X$juros~X$IPC)
summary(reg2_simples)
```

```
Call:
lm(formula = X$juros ~ X$IPC)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1154 -0.9097 -0.2310  0.4535  4.3671

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.9098     0.4885  -1.862   0.0799 .
X$IPC         1.4376     0.1195  12.030 9.69e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.627 on 17 degrees of freedom
Multiple R-squared:  0.8949,    Adjusted R-squared:  0.8887
F-statistic: 144.7 on 1 and 17 DF,  p-value: 9.691e-10
```

Modelo de Regressão Linear Simples
Resultado

Resultado da estimação – séries projetadas e o intervalo de confiança:

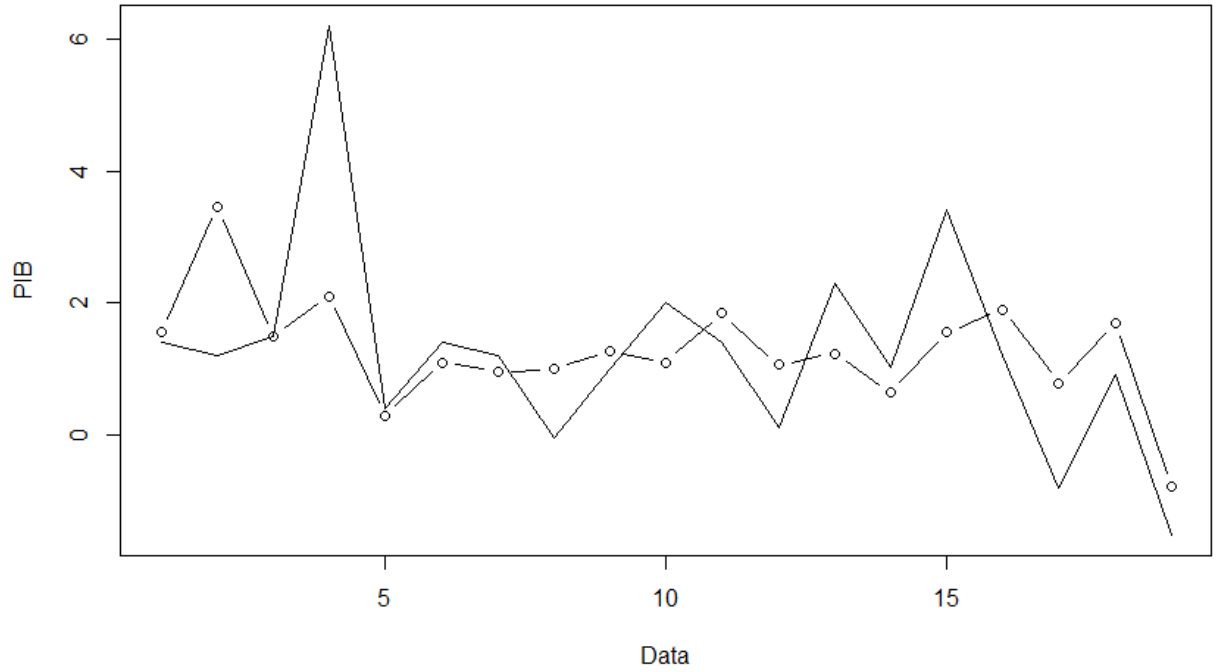
```
X1fit<-predict.lm(reg1_simples,interval="confidence") #resultado da estimação
```

Colocando os dados em um dataframe:

```
X1fit<-data.frame(X1fit)
```

Gráfico do valor real e estimado:

```
plot(X$PIB,type='l',xlab="Paises",ylab="PIB")
points(X1fit$fit,type='b')
```



Modelo de Regressão Linear Simples
Resultado

Resultado da estimação – séries projetadas e o intervalo de confiança:

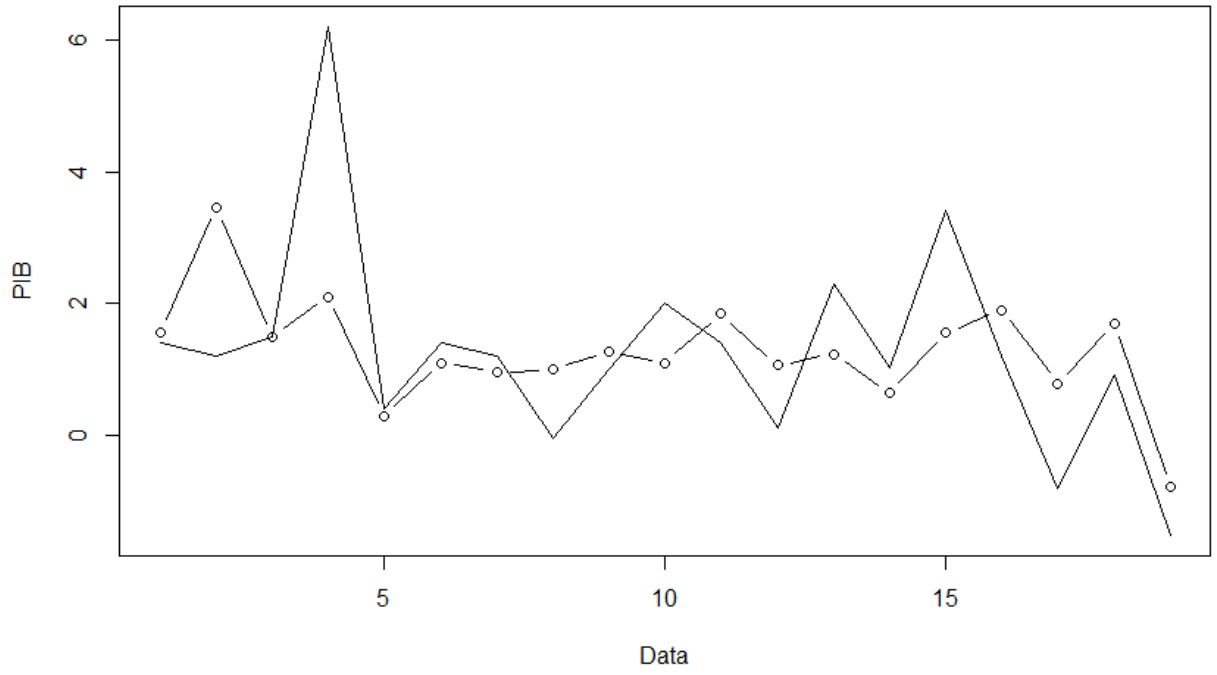
```
X2fit<-predict.lm(reg2_simples,interval="confidence") #resultado da estimação
```

Colocando os dados em um dataframe:

```
X2fit<-data.frame(X2fit)
```

Gráfico do valor real e estimado:

```
plot(X$juros,type='l',xlab="Paises",ylab="Juros")
points(X1fit$fit,type='b')
```



Modelo de Regressão Linear Múltipla *Estimação*

```
reg1_mult_a=lm(X$PIB~X$industria+X$varejo)
summary(reg1_mult_a)
```

```
reg1_mult_b=lm(X$PIB~X$industria+X$varejo+X$juros)
summary(reg1_mult_b)
```

O que aconteceu com os coeficientes? E a significância?

Modelo de Regressão Linear Múltipla *Comparando modelos*

- Resultado da estimação – series projetadas e o intervalo de confiança:
  - X1fit_ma<-predict.lm(reg1_mult_a,interval="confidence")
  - X1fit_mb<-predict.lm(reg1_mult_b,interval="confidence")

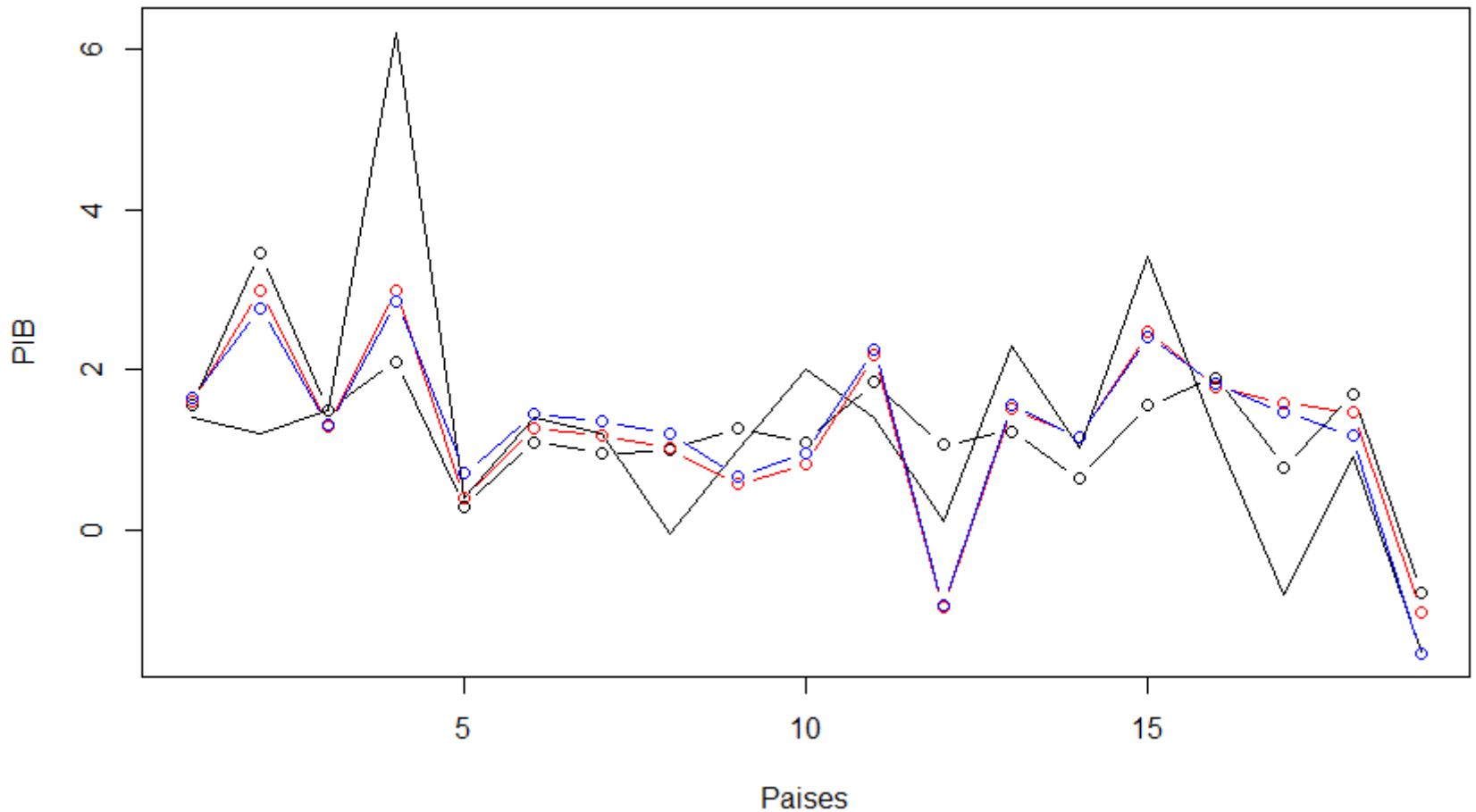
- Colocando os dados em um dataframe:
  - X1fit_ma <-data.frame(X1fit_ma)
  - X1fit_mb <-data.frame(X1fit_mb)

Modelo de Regressão Linear Múltipla *Comparando modelos*

Comparando os modelos estimados:

- plot(X$PIB,type='l',xlab="Paises",ylab="PIB")
- points(X1fit$fit,type='b')
- points(X1fit_ma$fit,type='b',col="red")
- points(X1fit_mb$fit,type='b',col="blue")

Modelo de Regressão Linear Múltipla
Comparando modelos



Modelo de Regressão Linear Múltipla Suposição Básica

- Todos os fatores considerados no termo de erro são não correlacionados com as variáveis independents.

$$E(u/X_1,X_2,....,X_k) = 0$$

- Isso implica que o modelo foi corretamente especificado.

Modelo de Regressão Linear Múltipla *Correlação entre os resíduos e o X's*

Salvando os resíduos:

- X$erro <- residuals(reg1_mult_b)
- Plot(X$erro)

Estimando a correlação:

- cor(X$erro,X$varejo)
- cor(X$erro,X$industria)
- cor(X$erro,X$PIB)

Modelo de Regressão Linear Múltipla Mínimos Quadrados Ordinários

- Como no caso da regressão simples, a função de regressão da população não é diretamente observável.
- Nós a estimamos a partir da **função de regressão amostral**:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \ldots. + \hat{\beta}_k X_{ik} \quad (6)$$

- Como obtemos os estimadores $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots., \hat{\beta}_k$?

Modelo de Regressão Linear Múltipla
Mínimos Quadrados Ordinários

- O método de **mínimos quadrados ordinários** escolhe as estimativas que minimizam a soma dos resíduos ao quadrado.
- Considerando um modelo com duas variávies independentes, o objetivo do MQO é **miniminar** a expressão abaixo:

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2})^2$$

Análise de Regressão Múltipla

Mínimos Quadrados Ordinários

- Diferenciando-se $\sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2})^2$ com relação à $\hat{\beta}_0, \hat{\beta}_1 \, e \, \hat{\beta}_2$ e igualando a zero, temos:

$$\frac{\partial\left(\sum \hat{u}_i^2\right)}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2}\right) = -2\sum \hat{u}_i = 0$$

$$\frac{\partial\left(\sum \hat{u}_i^2\right)}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2}\right)X_{i1} = -2\sum_{i=1}^{n}\hat{u}_i X_{i1} = 0$$

$$\frac{\partial\left(\sum \hat{u}_i^2\right)}{\partial \hat{\beta}_2} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2}\right)X_{i2} = -2\sum_{i=1}^{n}\hat{u}_i X_{i2} = 0$$

Modelo de Regressão Linear Simples
Relembrando

O modelo de regressão linear simples é

$$y = \alpha + \beta_1 x + \varepsilon$$

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum(x_i - \bar{x})^2}$$

$$= \frac{cov(x, y)}{var(x)}$$

Análise de Regressão Múltipla

Mínimos Quadrados Ordinários

- Resolvendo as equações normais, obtemos:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_{i1} \sum_{i=1}^{n} x_{i2}^2 - \sum_{i=1}^{n} y_i x_{i2} \sum_{i=1}^{n} x_{i1} x_{i2}}{\sum_{i=1}^{n} x_{i1}^2 \sum_{i=1}^{n} x_{i2}^2 - (\sum_{i=1}^{n} x_{i1} x_{i2})^2}$$

Covariância

Variância

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} y_i x_{i2} \sum_{i=1}^{n} x_{i1}^2 - \sum_{i=1}^{n} y_i x_{i1} \sum_{i=1}^{n} x_{i2} x_{i1}}{\sum_{i=1}^{n} x_{i1}^2 \sum_{i=1}^{n} x_{i2}^2 - (\sum_{i=1}^{n} x_{i1} x_{i2})^2}$$

Modelo de Regressão Linear Múltipla *Coeficiente*

Relembrando as equações

- summary(reg1_simples)
- summary(reg1_mult_a)
- *Compare os betas...*

Estimando o beta 1 da equação simples

- cov(X$industria,X$PIB)/var(X$industria)
- *Os betas são iguais da regressão linear simples!*

Mínimos Quadrados Ordinários Propriedades Numéricas

- A soma e, consequentemente, o valor médio dos resíduos é zero

$$\sum \hat{u}_i = 0$$

- Os resíduos não têm correlação com X_i

$$\sum_{i=1}^{n} \hat{u}_i X_{i1} = 0, \quad \sum_{i=1}^{n} \hat{u}_i X_{i2} = 0$$

- A reta de regressão passa pela média de Y e pela média de X

$$Y = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}_1 + \hat{\beta}_2 \overline{X}_2$$

Modelo de Regressão Linear Múltipla *Propriedades numéricas*

Média e soma dos resíduos igual a zero

- mean(X$erro)
- sum(X$erro)

Covariância entre os X's e o resíduo igual a zero

- cov(X$industria,X$erro)
- mean(X$industria*X$erro)
- cov(X$juros,X$erro)

Modelo de Regressão Linear Múltipla *Propriedades numéricas*

Equação que passa pela média

- reg1_mult_a
- summary(X)
- Estime o seguinte resultado:
  - 0.8083219+0.1698498*0.6963+0.1804849*1.929 = **Média do PIB**

Modelo de Regressão Linear Múltipla
Interpretação da Equação

- Da função de regressão amostral, obtemos

$$\Delta\hat{Y} = \hat{\beta}_1\Delta X_1 + \hat{\beta}_2\Delta X_2$$

- Se X_1 é mantido fixo, ΔX_1=0, então

$$\Delta\hat{Y} = \hat{\beta}_2\Delta X_2$$

Mínimos Quadrados Ordinários
Interpretação da Equação

- Mantendo as outras variáveis fixas, se X_1 aumentar em uma unidade, Y aumentará em $\hat{\beta}_1$ unidades. A interpretação é idêntica para $\hat{\beta}_2$.

$$\Delta \hat{Y} = \hat{\beta}_1 \Delta X_1$$

Análise de Regressão

Regressão Simples vs. Regressão Múltipla

- Há duas situações específicas nas quais as estimativas do modelo de regressão simples são iguais as estimativas do modelo de regressão múltipla.

$$\tilde{Y} = \tilde{\beta}_0 + \tilde{\beta}_1 X_1 \qquad \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_2 + \hat{\beta}_2 X_2$$

- Considerando um modelo com duas variáveis, temos:
  - Quando o efeito parcial de X_2 sobre Y é igual a zero, $\hat{\beta}_2 = 0$;
  - Quando X_1 e X_2 são não correlacionados.

$$\hat{\beta}_1 = \tilde{\beta}_1$$

Modelo de Regressão Linear Múltipla
Baixando os dados

list.files() – quais arquivos que estão no seu diretório

teorico<-read.csv("teorico.csv",sep=";", dec=".", head=TRUE) – baixando os dados.

teorico <Enter>

summary(teorico)

Modelo de Regressão Linear Múltipla *Aplicando os conceitos*

Estimando a matriz de correlação

- round(cor(teorico),2)

Calculando a equação simples

- tr_simples=lm(teorico$x~teorico$y)
- summary(tr_simples)

Calculando a equação múltipla (Y e Z)

- tr_mult_a=lm(teorico$x~teorico$y+teorico$z)
- summary(tr_mult_a)

Calculando a equação múltipla (Y e W)

- tr_mult_b=lm(teorico$x~teorico$y+teorico$w)
- summary(tr_mult_b)

O que aconteceu com os coeficientes? Está relacionado com o viés e multicolinearidade

Coeficiente de Determinação - R^2: Uma medida do grau de ajuste

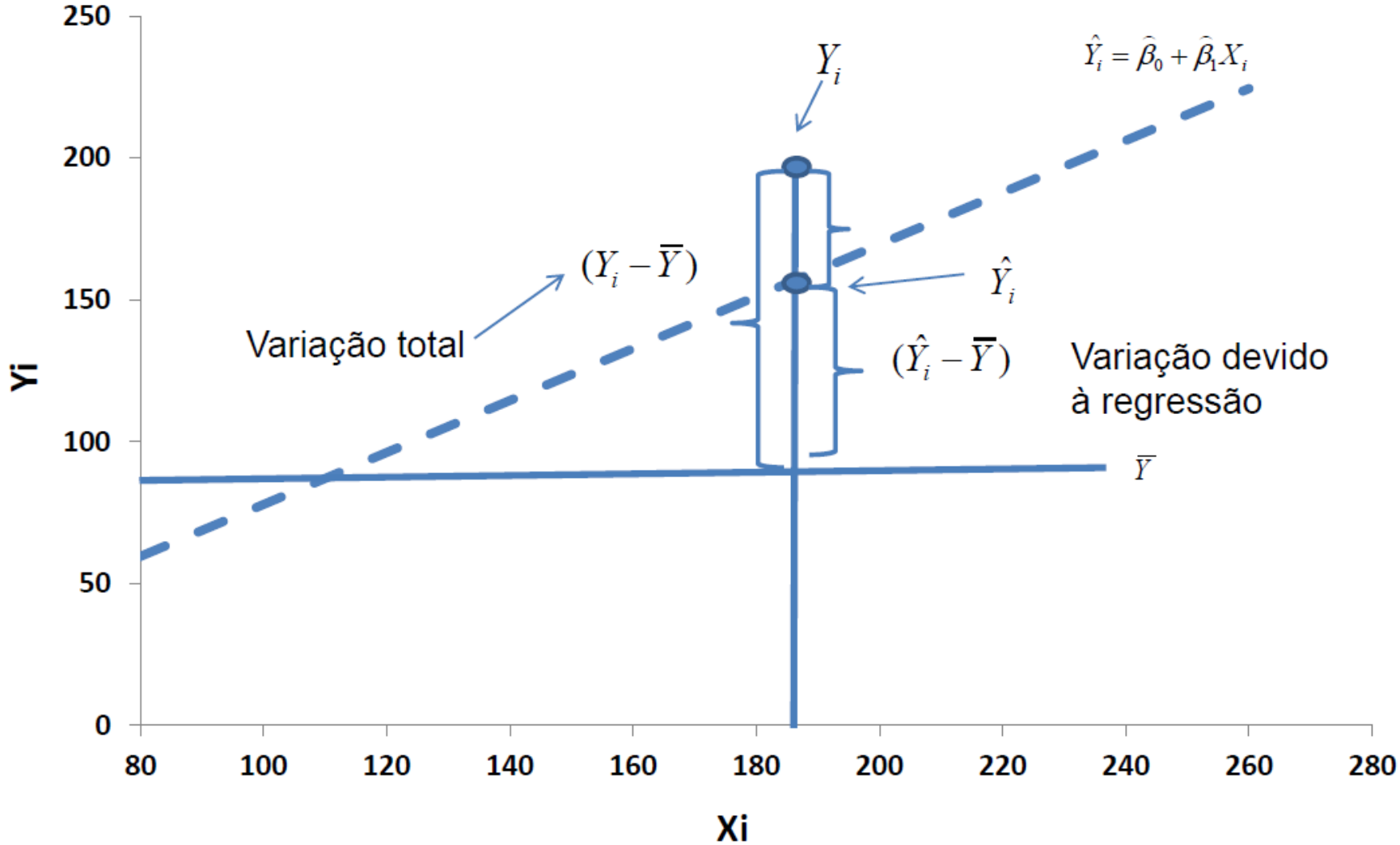
- O **coeficiente de determinação** R^2 é uma medida que diz quão bem a reta de regressão da amostra se ajusta ao dados.

Coeficiente de Determinação - R^2: Uma medida do grau de ajuste

- Nós podemos dividir a variação em Y em dois componentes, uma parte explicada pelo modelo de regressão e uma parte não explicada.

- O **coeficiente de determinação** R^2 mede a proporção ou a porcentagem da variação total em Y explicada pelo modelo de regressão.

Ilustração
Regressão Linear Simples



Coeficiente de Determinação - R^2: Como calcular?

Definindo :

$\sum \left(Y_i - \bar{Y}\right)^2$ como Soma dos Quadrados Total (SQT)

$\sum \left(\hat{Y}_i - \bar{Y}\right)^2$ como Soma dos Quadrados Explicada (SQE)

$\sum \hat{u}_i^2$ como Soma dos Quadrados dos Resíduos (SQR)

É possível mostrar que

$$SQT = SQE + SQR$$

Coeficiente de Determinação - R^2: Propriedades

- É uma quantidade não negativa
- Seus limites são $0 \leq R^2 \leq 1$.
  - Um $R^2 = 1$ significa um perfeito ajuste, isto é, $\hat{Y}_i = Y_i$ para todo i.
  - Um $R^2 = 0$ significa que não há relação entre o regredido e o regressor. Nesse caso, $\hat{Y}_i = \hat{\beta}_1 = \bar{Y}$, ou seja, a melhor estimativa para Y é sua média.

Coeficiente de Determinação Ajustado

R^2 Ajustado

- Um fato importante sobre o R^2 é que ele nunca diminui e, geralmente, aumenta quando outra variável independente é adicionada à regressão.

- Esse fato algébrico ocorre por definição, pois a soma dos resíduos quadrados nunca aumenta quando regressores adicionais são acrescentados ao modelo.

Coeficiente de Determinação Ajustado

R^2 Ajustado

- Para solucionar esse aspecto, considera-se o **coeficiente de determinação ajustado**, que leva em conta o número de variáveis independentes presentes no modelo.

$$\overline{R}^2 \equiv 1 - \frac{SQR/(n-k)}{SQT/(n-1)}$$

- Onde k = número de parâmetros do modelo (*incluindo o termo intercepto*)

Modelo de Regressão Linear Múltipla
Coeficiente de determinação

- summary(tr_simples)
- summary(reg1_mult_a)
- summary(reg1_mult_b)

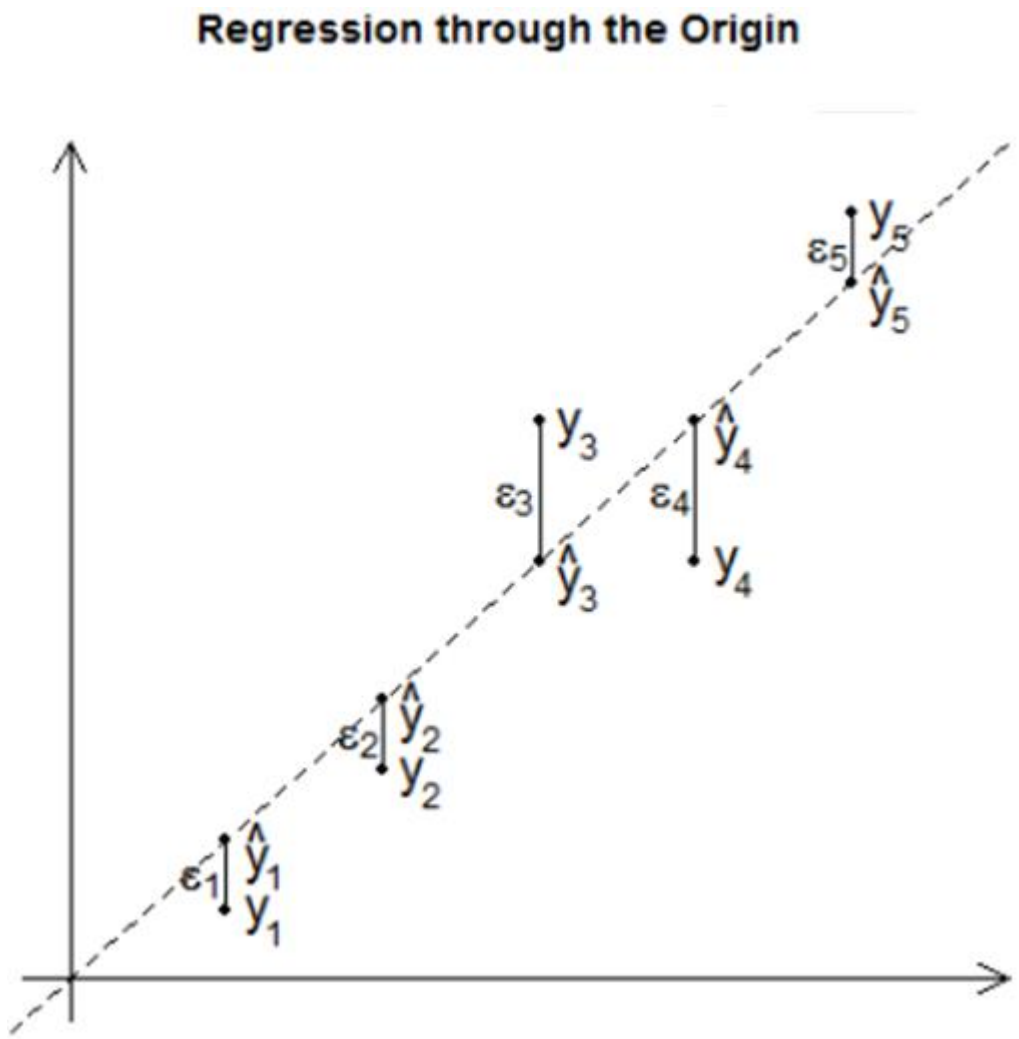
- *Compare os coeficientes de determinação e os ajustados.*

Modelo de Regressão Linear Múltipla
Coeficiente de determinação

- summary(reg1_simples)
- summary(tr_mult_a)
- summary(tr_mult_b)

- *O que aconteceu com os coeficientes de correlação quando o indicador tinha correlação próxima a zero?*

Modelo de Regressão Linear Múltipla
Regressão pela Origem



Modelo de Regressão Linear Múltipla Regressão pela Origem

- O termo intercepto β_0 é igual a zero.

$$\tilde{Y} = \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_2 + \ldots + \tilde{\beta}_k X_k \qquad (7)$$

- A soma dos resíduos não necessariamente é igual a zero
- O R^2 pode ser negativo
- Se o intercepto no modelo populacional é diferente de zero, as estimativas de (6) serão viesadas.
- O custo de estimar β_0 quando ele é igual a zero é que a variância dos estimadores será maior.

Modelo de Regressão Linear Múltipla
Regressão pela Origem

Extraindo as médias

- mean(X$pibm)
- X$pibm=X$PIB-1.274737
- mean(X$varejo)
- X$varejom=X$varejo-1.928947
- mean(X$industria)
- X$industriam=X$industria-0.6963158

Modelo de Regressão Linear Múltipla
Regressão pela Origem

Estimando as equações

- reg_orig=lm(X$pibm~X$varejom+X$industriam)
- summary(reg_orig)

- reg_s_interc=lm(X$pibm~X$varejom+X$industriam-1)
- summary(reg_s_interc)

Modelo de Regressão Linear Múltipla
O que acontece se retirar o intercepto?

Estimando as equações

- summary(reg1_mult_a)

- reg1_mult_sa=lm(X$pib~X$varejo+X$industria-1)
- summary(reg1_mult_sa)

Exercício

Modelo de Regressão Linear Múltipla *Baixando os dados – série de tempo*

list.files() – quais arquivos que estão no seu diretório

```
Z<-read.csv("serie_temporal.csv",sep=";", dec=".", head=TRUE)
head(Z)
```

Modelo de Regressão Linear Múltipla
Exercício

1. Estime uma equação simples do PIB brasileiro (Y) e alguma variável explicativa da base;
2. Faça uma regressão com duas variáveis explicativas;
3. Faça uma regressão com três variáveis explicativas;
4. O que aconteceu com os coeficientes?