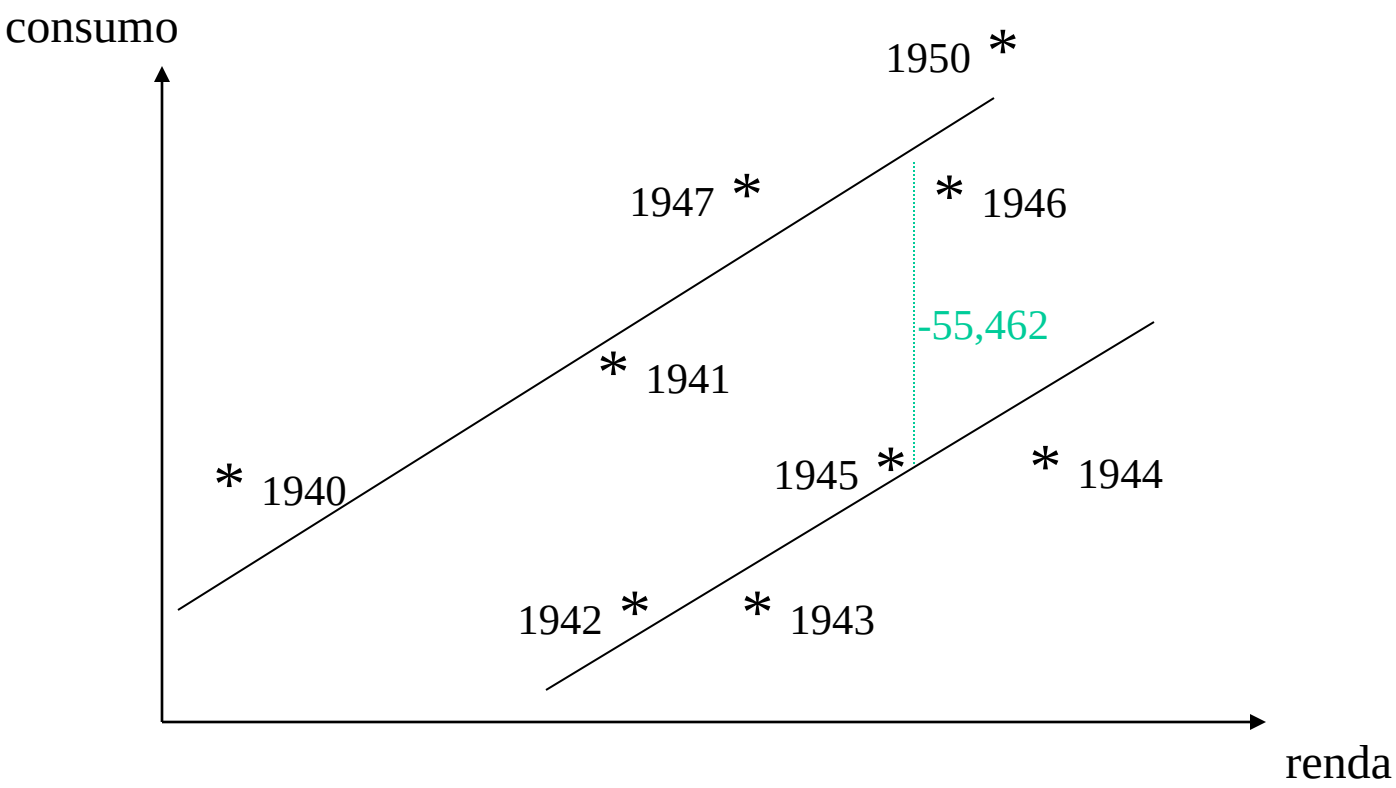
Variáveis Binárias na Regressão

Considere consumo pessoal e renda disponível de 1940 a 1950.

Observe que existem 4 pontos, 1942-1945, que estão abaixo da linha de regressão dos pontos remanescentes. A figura mostra como tratar deste problema.



Ano	Renda	Consumo
1940	244.0	229.9
1941	277.9	243.6
1942	317.5	241.1
1943	332.1	248.2
1944	343.6	255.2
1945	338.1	270.9
1946	332.7	301.0
1947	318.8	305.8
1948	335.8	312.2
1949	336.8	319.3
1950	362.8	337.3

O modelo de regressão para os dados de 1940-1950 é:

$$C_t = \beta_1 + \beta_2 X_t + \beta_3 W_t + \varepsilon_t$$

onde X_t é renda, C_t é consumo e W_t é a variável binária.

W_t = 1 nos anos 1942-1945 $\longrightarrow$ $C_t = (\beta_1 + \beta_3) + \beta_2 X_t + \varepsilon_t$

W_t = 0 nos demais anos $\longrightarrow$ $C_t = \beta_1 + \beta_2 X_t + \varepsilon_t$

$$\hat{C}_t = -10{,}065 + 0{,}9595\, X_t - 55{,}4624 W_t$$

O resultado da regressão confirma que o consumo durante a 2ª. guerra mundial foi significativamente abaixo daquele que poderia ser esperado a uma dada renda. A produção é voltada para armamento e a demanda se restringe.

Outros exemplos que envolvem uma única variável binária:

Diferenças entre oferta de trabalho e estrutura de salários na indústria, para homens e mulheres.

Alteração nos modelos macro-econômicos antes e após um regime.

Nestes casos o modelo é:

$$y_i = \beta' X_i + \delta d_i + \varepsilon_i$$

Quando existem várias categorias, é necessário um conjunto de variáveis binárias. Corrigir para fatores sazonais em dados macroeconômicos é uma aplicação comum. A função consumo para dados trimestrais seria:

$$C_t = \beta_1 + \beta_2 X_t + \delta_2 D_{t2} + \delta_3 D_{t3} + \delta_4 D_{t4} + \varepsilon$$

$$\text{1o. trimestre} \rightarrow C_t = \beta_1 + \beta_2 X_t + \varepsilon$$
$$\text{2o. trimestre} \rightarrow C_t = \beta_1 + \delta_2 + \beta_2 X_t + \varepsilon$$
$$\text{3o. trimestre} \rightarrow C_t = \beta_1 + \delta_3 + \beta_2 X_t + \varepsilon$$
$$\text{4o. trimestre} \rightarrow C_t = \beta_1 + \delta_4 + \beta_2 X_t + \varepsilon$$

$$\mathbf{X} = \begin{array}{c} \begin{matrix} \text{cte} & D_2 & D_3 & D_4 & X \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & x_1 \\ 1 & 1 & 0 & 0 & x_2 \\ 1 & 0 & 1 & 0 & x_3 \\ 1 & 0 & 0 & 1 & x_4 \\ 1 & 0 & 0 & 0 & x_5 \\ 1 & 1 & 0 & 0 & x_6 \\ \vdots & & & & \\ 1 & 0 & 0 & 1 & x_T \end{bmatrix} \end{array}$$

Neste caso estamos omitindo a variável binária para o 1º. trimestre. Qualquer uma das 4 variáveis representando os trimestres poderia ser omitida, ou usada como período base.

Este procedimento é uma forma de desazonalisar os dados. Considere uma formulação alternativa,

$$C_t = \beta X_t + \delta_1 D_{1t} + \delta_2 D_{2t} + \delta_3 D_{3t} + \delta_4 D_{4t} + \varepsilon_t$$

No caso de dados mensais, teríamos:

$$C_t = \beta X_t + \sum_{i=1}^{12} \delta_i D_{it} + \varepsilon_t$$

D_i = 1 para o mês i e zero para os demais.

Dois ou mais conjuntos de variáveis binárias

Queremos incorporar duas variáveis qualitativas na equação de regressão, em que cada variável é representada por um conjunto de variáveis binárias.

Efeito dos níveis

A variável educação é tipicamente observada em níveis ao invés de anos de educação, por exemplo, suponha que estamos interessados na seguinte relação.

$$\text{Renda} = f(\text{idade, níveis de educação})$$

Suponha que nosso dado fornece o nível mais elevado de educação: ensino médio, faculdade, mestrado, doutorado.

Poderíamos usar:

$$\text{Renda} = \beta_1 + \beta_2 \,\text{idade} + \beta_3 E + \varepsilon$$

onde E = 0 para o grupo 1, 1 para o grupo 2, 2 para o grupo 3 e 3 para o grupo 4.

Esta não é uma forma satisfatória, pois assume que o incremento na renda para cada nível de educação é o mesmo.

β_3 é a diferença entre renda com doutorado e renda com mestrado, entre renda com mestrado e com faculdade, e entre renda com faculdade e colegial. Isto não ocorre na prática.

Um modelo mais flexível usaria 3 ou 4 binárias, uma para cada nível de educação.

$$\text{Renda} = \beta_1 + \beta_2 \,\text{idade} + \delta_1 W_1 + \delta_2 W_2 + \delta_3 W_3 + \varepsilon$$

Colegial: $\text{Renda} = \beta_1 + \beta_2 \,\text{idade}$

Faculdade: $\text{Renda} = \beta_1 + \beta_2 \,\text{idade} + \delta_1$

Mestrado: $\text{Renda} = \beta_1 + \beta_2 \,\text{idade} + \delta_2$

Doutorado: $\text{Renda} = \beta_1 + \beta_2 \,\text{idade} + \delta_3$

$$
\begin{array}{c}
\\
\textit{colegial} \\
\textit{faculdade} \\
\textit{mestrado} \\
\textit{doutorado}
\end{array}
\begin{array}{c}
\begin{array}{ccccc} \text{cte.} & W_1 & W_2 & W_3 & \end{array} \\
\begin{bmatrix}
i_C & 0 & 0 & 0 & idade_C \\
i_F & i_F & 0 & 0 & idade_F \\
i_M & 0 & i_M & 0 & idade_M \\
i_D & 0 & 0 & i_D & idade_D
\end{bmatrix}
\end{array}
$$

Nosso interesse está na diferença entre δ_3 e δ_2 e entre δ_2 e δ_1, que é simples de computar.

Uma forma alternativa de formular esta equação que revela estas diferenças diretamente é feita através da redefinição das variáveis binárias.

$$\text{Renda} = \beta_1 + \beta_2 \,\text{idade} + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3 + \varepsilon$$

$$
\begin{array}{c}
\\
\textit{colegial} \\
\textit{faculdade} \\
\textit{mestrado} \\
\textit{doutorado}
\end{array}
\begin{array}{c}
\begin{array}{ccccc} \text{cte.} & Z_1 & Z_2 & Z_3 & \end{array} \\
\begin{bmatrix}
i_C & 0 & 0 & 0 & idade_C \\
i_F & i_F & 0 & 0 & idade_F \\
i_M & i_M & i_M & 0 & idade_M \\
i_D & i_D & i_D & i_D & idade_D
\end{bmatrix}
\end{array}
$$

Colegial: Renda = $\beta_1 + \beta_2$ idade

Faculdade: Renda = $\beta_1 + \beta_2$ idade + γ_1

Mestrado: Renda = $\beta_1 + \beta_2$ idade + $\gamma_1 + \gamma_2$

Doutorado: Renda = $\beta_1 + \beta_2$ idade + $\gamma_1 + \gamma_2 + \gamma_3$

No 1º. modelo, δ_3 era a diferença entre o doutorado e o caso básico, neste modelo γ_3 é o valor marginal do doutorado.

Observe que:

$\gamma_1 = \delta_1$

$\gamma_1 + \gamma_2 = \delta_2$ ou $\gamma_2 = \delta_2 - \delta_1$

$\gamma_1 + \gamma_2 + \gamma_3 = \delta_3$ ou $\gamma_3 = \delta_3 - (\delta_2 - \delta_1) - \delta_1 = \delta_3 - \delta_2$

Outro exemplo:

Demanda de sorvete em função das estações do ano.

Modelo 1 $$Y = \alpha + \beta X + \delta_1 Z_1 + \delta_2 Z_2 + \delta_3 Z_3 + \varepsilon$$

	Z_1	Z_2	Z_3	
Inverno	0	0	0	$Y = \alpha + \beta X$
Outono	1	0	0	$Y = \alpha + \beta X + \delta_1$
Primavera	0	1	0	$Y = \alpha + \beta X + \delta_2$
Verão	0	0	1	$Y = \alpha + \beta X + \delta_3$

Modelo 2 $$Y = \alpha + \beta X + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3 + \varepsilon$$

	Z_1	Z_2	Z_3	
Inverno	0	0	0	$Y = \alpha + \beta X$
Outono	1	0	0	$Y = \alpha + \beta X + \delta_1$
Primavera	1	1	0	$Y = \alpha + \beta X + \gamma_1 + \gamma_2$
Verão	1	1	1	$Y = \alpha + \beta X + \gamma_1 + \gamma_2 + \gamma_3$

Para testar a hipótese de que a demanda por sorvete não varia na primavera e verão.

Modelo 1 $H_0: \delta_2 = \delta_3$

Modelo 2 $H_0: \gamma_3 = 0$

Interações:

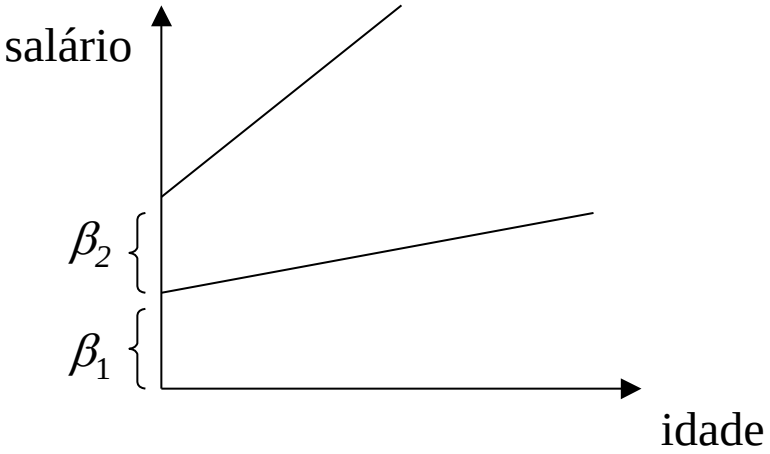
Considere, por exemplo, um modelo de regressão que mostra que não são só os salários dos estudantes com faculdade que são maiores dos que os salários dos estudantes sem faculdade, a uma dada idade, mas também que os salários dos estudantes formados aumentam mais rapidamente quando os mesmos adquirem mais idade.

$$S = \beta_1 + \beta_2 idade + \beta_3 d + \beta_4 d.idade + \varepsilon$$

onde *S* é o salário e *d* = 1 se o estudante terminou a faculdade e 0 caso contrário

Estudante com faculdade $\longrightarrow$ $S = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) idade$

Estudante sem faculdade $\longrightarrow$ $S = \beta_1 + \beta_2 idade$



Neste caso a matriz de regressores é:

$$X = \begin{bmatrix} i_1 & 0 & X_1 & 0 \\ i_2 & i_2 & X_2 & X_2 \end{bmatrix}$$

ou

$$Z = \begin{bmatrix} i & d & X & dX \end{bmatrix}$$

No caso de duas variáveis (Z é a binária):

$$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma Z + \delta_1 Z X_1 + \delta_2 Z X_2$$

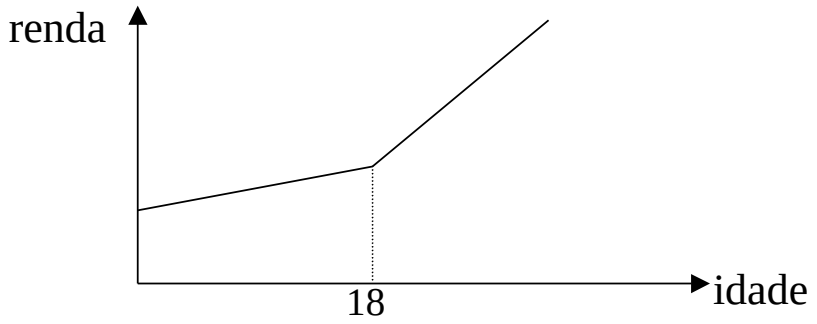
Pode ocorrer de haver diferenças em diferentes níveis de idade. Por exemplo, ao completar 18 anos, idade típica para o término do colegial, e 22 anos para o término da faculdade, a inclinação a partir destas idades deve se alterar, apesar de a renda estar sempre aumentando com a idade.

1. Somente uma mudança na inclinação, para indivíduos maiores ou menore de 18 anos.

$$y = \alpha + \beta X + \gamma d(X - 18) + \varepsilon$$

$$d = 0 \text{ para } X < 18 \Rightarrow y = \alpha + \beta X + \varepsilon$$

$$d = 1 \text{ para } X \geq 18 \Rightarrow y = (\alpha - 18\gamma) + (\beta + \gamma) X + \varepsilon$$



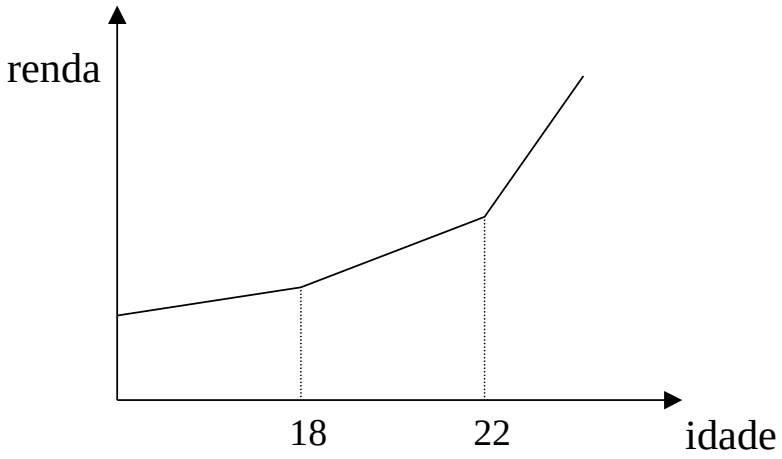
Mudança em duas inclinações.

$$y = \alpha + \beta X + \gamma_1 d_1 (X - 18) + \gamma_2 d_2 (X - 22) + \varepsilon$$

se idade $< 18, d_1 = d_2 = 0 \Rightarrow y = \alpha + \beta X + \varepsilon$

se $18 \leq$ idade $< 22, d_1 = 1$ e $d_2 = 0 \Rightarrow y = (\alpha - \gamma_1 18) + (\beta + \gamma_1) X + \varepsilon$

se idade $\geq 22, d_1 = d_2 = 1 \Rightarrow y = (\alpha - \gamma_1 18 - \gamma_2 22) + (\beta + \gamma_1 + \gamma_2) X + \varepsilon$



Exemplos:

$$y = \alpha + \beta x + \gamma Z + \varepsilon$$

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \\ 1 & 4 & 0 \end{bmatrix} \qquad y = \begin{bmatrix} 8 \\ 7 \\ 7 \\ 6 \\ 6 \\ 5 \\ 3 \\ 2 \end{bmatrix}$$

$$(X'X) = \begin{bmatrix} 8 & 20 & 4 \\ 20 & 60 & 10 \\ 4 & 10 & 4 \end{bmatrix} \qquad (X'X)^{-1} = \begin{bmatrix} 0{,}875 & -0{,}25 & -0{,}25 \\ -0{,}25 & 0{,}1 & 0 \\ -0{,}25 & 0 & 0{,}5 \end{bmatrix}$$

$$X'y = \begin{bmatrix} 44 \\ 100 \\ 28 \end{bmatrix} \qquad b = \begin{bmatrix} 6{,}5 \\ -1 \\ 3 \end{bmatrix}$$

$Z = 1 \qquad y = 9{,}5 - x$

$Z = 0 \qquad y = 6{,}5 - x$

Se $Z' = \begin{bmatrix} -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \end{bmatrix}$

$$(X'X)^{-1} = \begin{bmatrix} 0{,}75 & -0{,}25 & 0 \\ -0{,}25 & 0{,}1 & 0 \\ 0 & 0 & 0{,}125 \end{bmatrix} \qquad X'y = \begin{bmatrix} 44 \\ 100 \\ -12 \end{bmatrix} \qquad b = \begin{bmatrix} 8 \\ -1 \\ -1{,}5 \end{bmatrix}$$

$Z = -1 \qquad y = 9{,}5 - x$

$Z = 1 \qquad y = 6{,}5 - x$

Se $Z' = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$

$$(X'X)^{-1} = \begin{bmatrix} 0{,}875 & -0{,}25 & -0{,}25 \\ -0{,}25 & 0{,}1 & 0 \\ -0{,}25 & 0 & 0{,}5 \end{bmatrix} \qquad x'y = \begin{bmatrix} 44 \\ 100 \\ 16 \end{bmatrix} \qquad b = \begin{bmatrix} 9{,}5 \\ -1 \\ -3 \end{bmatrix}$$

$Z = 0 \qquad y = 9{,}5 - x$

$Z = 1 \qquad y = 6{,}5 - x$

Modelo 1–Regressores comuns para ambos os períodos (intercepto+inclinações)

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} i_1 & X_1 \\ i_2 & X_2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + u \qquad SQres \text{ I}$$

Modelo II – Interceptos diferentes e inclinações iguais

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} i_1 & 0 & X_1 \\ 0 & i_2 & X_2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix} + u \qquad SQres \text{ II}$$

Modelo III – Interceptos diferentes e inclinações diferentes

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} i_1 & X_1 & 0 & 0 \\ 0 & 0 & i_2 & X_2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{bmatrix} + u \qquad SQres \text{ III}$$

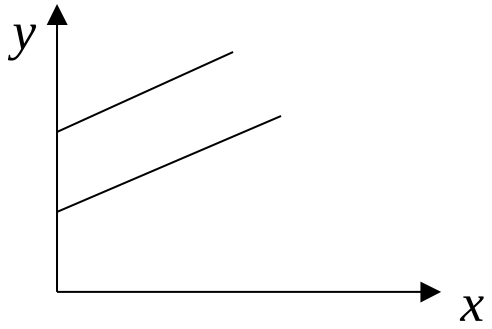
Modelo IV – Interceptos iguais e inclinações diferentes

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} i_1 & X_1 & 0 \\ i_2 & 0 & X_2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix} + u$$

SQres IV

1) Testar se os interceptos são diferentes

$$F = \frac{SQresI - SQresII/1}{SQresII/n-3}$$



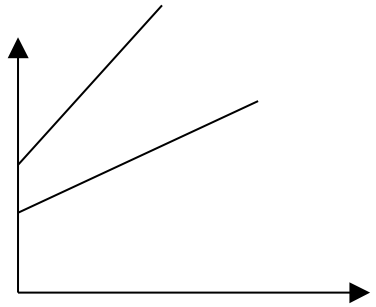
ou

$$y = \alpha + \beta X + \delta D + \varepsilon$$

$$H_0 : \delta = 0$$

2) Testar se todos os coeficientes são diferentes

$$F = \frac{SQresI - SQresIII/2}{SQresIII/n-4}$$

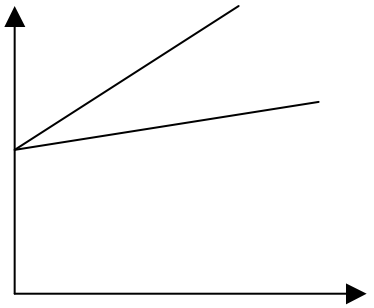


ou

$$y = \alpha + \beta X + \delta D + \gamma DX + \varepsilon$$

$$H_0 : \delta = 0 \text{ e } \gamma = 0$$

3) Testar se as inclinações são diferentes



$$F = \frac{SQresI - SQresIV/1}{SQresIV/n-3}$$

ou

$$y = \alpha + \beta X + \gamma DX + \varepsilon \qquad H_0 : \gamma = 0$$

4) Testar que as inclinações são diferentes dado que os interceptos são diferentes

$$F = \frac{SQresII - SQresIII/1}{SQresIII/n-4}$$

ou

$$y = \alpha + \beta X + \delta D + \gamma DX + \varepsilon \qquad H_0 : \gamma = 0$$