Modelos de Regressão com variáveis Dummy

Prof. Juvêncio Santos Nobre

Departamento de Estatística e Matemática Aplicada

Universidade Federal do Ceará-Brasil

http://www.dema.ufc.br/∼juvencio

DEMA-UFC

Capital do **Ceará**, novembro de 2022

Conteúdo

Níveis de medida

Hoffman (2016) e Gujarati e Porter (2011, Econometria básica, 5a edição) comentam sobre os diferentes tipos de níveis de medida (variáveis), em especial:

- Escala nominal: quando temos apenas uma classificação em categorias. Neste caso, se forem usados números para indicar as diferentes categorias, eles são apenas rótulos. Exs: sexo ou religião das pessoas.
- Escala ordinal: quando se tem uma variável qualitativa em que existe alguma ordem entre as categorias. Exs: Status social, nível de instrução.
- Escala intervalar: neste caso vale a ordem e também podemos comparar, numericamente, intervalos (diferenças) entre valores. Mas a razão entre valores não tem sentido porque a origem é arbitrária. Ex: temperatura medida em graus °C ou em °F e ano (data). Tem sentido dizer que o período 1982-1990 é três vezes mais longo do que o período 1979-1981, mas não tem sentido dizer que no ano de 2002 estávamos no dobro de 1001.
- Escala razão ou cardinal: que representam as variáveis quantitativas, ou seja, são válidas todas as operações álgebricas com os valores. Exs: peso, idade, valor monetário.

Níveis de medida

Hoffman (2016) e Gujarati e Porter (2011, Econometria básica, 5a edição) comentam sobre os diferentes tipos de níveis de medida (variáveis), em especial:

- Escala nominal: quando temos apenas uma classificação em categorias. Neste caso, se forem usados números para indicar as diferentes categorias, eles são apenas rótulos. Exs: sexo ou religião das pessoas.
- Escala ordinal: quando se tem uma variável qualitativa em que existe alguma ordem entre as categorias. Exs: Status social, nível de instrução.
- Escala intervalar: neste caso vale a ordem e também podemos comparar, numericamente, intervalos (diferenças) entre valores. Mas a razão entre valores não tem sentido porque a origem é arbitrária. Ex: temperatura medida em graus °C ou em °F e ano (data). Tem sentido dizer que o período 1982-1990 é três vezes mais longo do que o período 1979-1981, mas não tem sentido dizer que no ano de 2002 estávamos no dobro de 1001.
- Escala razão ou cardinal: que representam as variáveis quantitativas, ou seja, são válidas todas as operações álgebricas com os valores. Exs: peso, idade, valor monetário.

Níveis de medida

Hoffman (2016) e Gujarati e Porter (2011, Econometria básica, 5a edição) comentam sobre os diferentes tipos de níveis de medida (variáveis), em especial:

- Escala nominal: quando temos apenas uma classificação em categorias. Neste caso, se forem usados números para indicar as diferentes categorias, eles são apenas rótulos. Exs: sexo ou religião das pessoas.
- Escala ordinal: quando se tem uma variável qualitativa em que existe alguma ordem entre as categorias. Exs: Status social, nível de instrução.
- Escala intervalar: neste caso vale a ordem e também podemos comparar, numericamente, intervalos (diferenças) entre valores. Mas a razão entre valores não tem sentido porque a origem é arbitrária. Ex: temperatura medida em graus °C ou em °F e ano (data). Tem sentido dizer que o período 1982-1990 é três vezes mais longo do que o período 1979-1981, mas não tem sentido dizer que no ano de 2002 estávamos no dobro de 1001.
- Escala razão ou cardinal: que representam as variáveis quantitativas, ou seja, são válidas todas as operações álgebricas com os valores. Exs: peso, idade, valor monetário.

Níveis de medida

Hoffman (2016) e Gujarati e Porter (2011, Econometria básica, 5a edição) comentam sobre os diferentes tipos de níveis de medida (variáveis), em especial:

- Escala nominal: quando temos apenas uma classificação em categorias. Neste caso, se forem usados números para indicar as diferentes categorias, eles são apenas rótulos. Exs: sexo ou religião das pessoas.
- Escala ordinal: quando se tem uma variável qualitativa em que existe alguma ordem entre as categorias. Exs: Status social, nível de instrução.
- Escala intervalar: neste caso vale a ordem e também podemos comparar, numericamente, intervalos (diferenças) entre valores. Mas a razão entre valores não tem sentido porque a origem é arbitrária. Ex: temperatura medida em graus °C ou em °F e ano (data). Tem sentido dizer que o período 1982-1990 é três vezes mais longo do que o período 1979-1981, mas não tem sentido dizer que no ano de 2002 estávamos no dobro de 1001.
- Escala razão ou cardinal: que representam as variáveis quantitativas, ou seja, são válidas todas as operações álgebricas com os valores. Exs: peso, idade, valor monetário.

Tipos de regressão

- Os modelos de regressão usuais são utilizados em situações que as variáveis respostas são do tipo quantitativas contínuas com suporte em $\mathbb{R}$.
- Quando o suporte é $\mathbb{R}^+$ ou limitado do tipo $[0, 1]$ modelos com características especiais devem ser adotados, como por exemplo, os modelos de regressão gama, normal inversa, beta-prime, beta, simplex, beta-retangular, gama unitária, dentre várias outras alternativas.
- Em situações em que as variáveis respostas são do tipo **quantitativas discretas**, com suporte em $\mathbb{N}$ ou em um conjunto enumerável, pode-se trabalhar com modelos para dados de contagem ou modelos baseados em variáveis aleatórias discretas.
- Podemos ter modelos de regressão para variáveis resposta do tipo intervalar. ☺
- Para variáveis respostas do tipo qualitativas (nominais ou ordinais), pode-se considerar modelos do tipo Logístico, Probito, etc...

Tipos de regressão

- Os modelos de regressão usuais são utilizados em situações que as variáveis respostas são do tipo quantitativas contínuas com suporte em $\mathbb{R}$.
- Quando o suporte é $\mathbb{R}^+$ ou limitado do tipo $[0, 1]$ modelos com características especiais devem ser adotados, como por exemplo, os modelos de regressão gama, normal inversa, beta-prime, beta, simplex, beta-retangular, gama unitária, dentre várias outras alternativas.
- Em situações em que as variáveis respostas são do tipo **quantitativas discretas**, com suporte em $\mathbb{N}$ ou em um conjunto enumerável, pode-se trabalhar com modelos para dados de contagem ou modelos baseados em variáveis aleatórias discretas.
- Podemos ter modelos de regressão para variáveis resposta do tipo intervalar. ☺
- Para variáveis respostas do tipo qualitativas (nominais ou ordinais), pode-se considerar modelos do tipo Logístico, Probito, etc...

Tipos de regressão

- Os modelos de regressão usuais são utilizados em situações que as variáveis respostas são do tipo quantitativas contínuas com suporte em $\mathbb{R}$.
- Quando o suporte é $\mathbb{R}^+$ ou limitado do tipo $[0, 1]$ modelos com características especiais devem ser adotados, como por exemplo, os modelos de regressão gama, normal inversa, beta-prime, beta, simplex, beta-retangular, gama unitária, dentre várias outras alternativas.
- Em situações em que as variáveis respostas são do tipo **quantitativas discretas**, com suporte em $\mathbb{N}$ ou em um conjunto enumerável, pode-se trabalhar com modelos para dados de contagem ou modelos baseados em variáveis aleatórias discretas.
- Podemos ter modelos de regressão para variáveis resposta do tipo intervalar. ☺
- Para variáveis respostas do tipo qualitativas (nominais ou ordinais), pode-se considerar modelos do tipo Logístico, Probito, etc...

Tipos de regressão

- Os modelos de regressão usuais são utilizados em situações que as variáveis respostas são do tipo quantitativas contínuas com suporte em $\mathbb{R}$.
- Quando o suporte é $\mathbb{R}^+$ ou limitado do tipo $[0, 1]$ modelos com características especiais devem ser adotados, como por exemplo, os modelos de regressão gama, normal inversa, beta-prime, beta, simplex, beta-retangular, gama unitária, dentre várias outras alternativas.
- Em situações em que as variáveis respostas são do tipo **quantitativas discretas**, com suporte em $\mathbb{N}$ ou em um conjunto enumerável, pode-se trabalhar com modelos para dados de contagem ou modelos baseados em variáveis aleatórias discretas.
- Podemos ter modelos de regressão para variáveis resposta do tipo intervalar. 😉
- Para variáveis respostas do tipo qualitativas (nominais ou ordinais), pode-se considerar modelos do tipo Logístico, Probito, etc...

Tipos de regressão

- Os modelos de regressão usuais são utilizados em situações que as variáveis respostas são do tipo quantitativas contínuas com suporte em $\mathbb{R}$.
- Quando o suporte é $\mathbb{R}^+$ ou limitado do tipo $[0, 1]$ modelos com características especiais devem ser adotados, como por exemplo, os modelos de regressão gama, normal inversa, beta-prime, beta, simplex, beta-retangular, gama unitária, dentre várias outras alternativas.
- Em situações em que as variáveis respostas são do tipo **quantitativas discretas**, com suporte em $\mathbb{N}$ ou em um conjunto enumerável, pode-se trabalhar com modelos para dados de contagem ou modelos baseados em variáveis aleatórias discretas.
- Podemos ter modelos de regressão para variáveis resposta do tipo intervalar. 😉
- Para variáveis respostas do tipo qualitativas (nominais ou ordinais), pode-se considerar modelos do tipo Logístico, Probito, etc...

Tipos de regressão

- Vamos apresentar uma metodologia para ajustar modelos de regressão em que a variável resposta é do tipo quantitativa contínua com suporte em $\mathbb{R}$ no qual se tem também variáveis explicativas do tipo qualitativa.
- Modelos em que as variáveis explicativas são todas qualitativas são denominados modelos de ANOVA/planejamento.
- Infelizmente, dado o tempo e a ementa da disciplina, não teremos condição de discutir todas as extensões supracitadas. ☹

Tipos de regressão

- Vamos apresentar uma metodologia para ajustar modelos de regressão em que a variável resposta é do tipo quantitativa contínua com suporte em $\mathbb{R}$ no qual se tem também variáveis explicativas do tipo qualitativa.
- Modelos em que as variáveis explicativas são todas qualitativas são denominados modelos de ANOVA/planejamento.
- Infelizmente, dado o tempo e a ementa da disciplina, não teremos condição de discutir todas as extensões supracitadas. ☹

Tipos de regressão

- Vamos apresentar uma metodologia para ajustar modelos de regressão em que a variável resposta é do tipo quantitativa contínua com suporte em $\mathbb{R}$ no qual se tem também variáveis explicativas do tipo qualitativa.
- Modelos em que as variáveis explicativas são todas qualitativas são denominados modelos de ANOVA/planejamento.
- Infelizmente, dado o tempo e a ementa da disciplina, não teremos condição de discutir todas as extensões supracitadas. ☹

Exemplos

Em muitas aplicações de interesse, se faz necessário utilizar variáveis explicativas qualitativas, por exemplo:

- Rendimento do aluno vs. tipo de escola. Há diferença no rendimento dado o tipo de escola?
- Salário versus sexo. Há diferença de salário por gênero?
- Durabilidade bateria versus marca. A marca do tipo **A** realmente é mais duradoura?
- Redução de placa bacteriana versus tipo de escova. Será que a nova marca é tão duradoura e eficaz quanto a padrão?

Exemplos

Em muitas aplicações de interesse, se faz necessário utilizar variáveis explicativas qualitativas, por exemplo:

- Rendimento do aluno vs. tipo de escola. Há diferença no rendimento dado o tipo de escola?
- Salário versus sexo. Há diferença de salário por gênero?
- Durabilidade bateria versus marca. A marca do tipo **A** realmente é mais duradoura? ☺
- Redução de placa bacteriana versus tipo de escova. Será que a nova marca é tão duradoura e eficaz quanto a padrão?

Exemplos

Em muitas aplicações de interesse, se faz necessário utilizar variáveis explicativas qualitativas, por exemplo:

- Rendimento do aluno vs. tipo de escola. Há diferença no rendimento dado o tipo de escola?
- Salário versus sexo. Há diferença de salário por gênero?
- Durabilidade bateria versus marca. A marca do tipo **A** realmente é mais duradoura? ☺
- Redução de placa bacteriana versus tipo de escova. Será que a nova marca é tão duradoura e eficaz quanto a padrão?

Exemplos

Em muitas aplicações de interesse, se faz necessário utilizar variáveis explicativas qualitativas, por exemplo:

- Rendimento do aluno vs. tipo de escola. Há diferença no rendimento dado o tipo de escola?
- Salário versus sexo. Há diferença de salário por gênero?
- Durabilidade bateria versus marca. A marca do tipo **A** realmente é mais duradoura? 😉
- Redução de placa bacteriana versus tipo de escova. Será que a nova marca é tão duradoura e eficaz quanto a padrão?

Variáveis dummy

- Como incorporar este tipo de variável no modelo de regressão?
- Isso pode ser feito através da inclusão de variáveis indicadoras, também denominadas de variáveis *dummy*.
- Considere a situação hipotética em que o interesse é modelar o salário (y) em função da experiência (x_2 anos) no cargo. Inicialmente, pode-se considerar um MRLS do tipo

$$y_i = \beta_0 + \beta_2 x_{2i} + e_i, \ i = 1, \ldots, n.$$

- Posteriormente, pode-se ter interesse em avaliar se há diferença por gênero/sexo.
- Uma primeira alternativa, seria ajustar um MRLS para cada sexo.

Variáveis dummy

- Como incorporar este tipo de variável no modelo de regressão?
- Isso pode ser feito através da inclusão de variáveis indicadoras, também denominadas de variáveis *dummy*.
- Considere a situação hipotética em que o interesse é modelar o salário (y) em função da experiência (x_2 anos) no cargo. Inicialmente, pode-se considerar um MRLS do tipo

$$y_i = \beta_0 + \beta_2 x_{2i} + e_i,\ i = 1, \ldots, n.$$

- Posteriormente, pode-se ter interesse em avaliar se há diferença por gênero/sexo.
- Uma primeira alternativa, seria ajustar um MRLS para cada sexo.

Variáveis dummy

- Como incorporar este tipo de variável no modelo de regressão?
- Isso pode ser feito através da inclusão de variáveis indicadoras, também denominadas de variáveis *dummy*.
- Considere a situação hipotética em que o interesse é modelar o salário (y) em função da experiência (x_2 anos) no cargo. Inicialmente, pode-se considerar um MRLS do tipo

$$y_i = \beta_0 + \beta_2 x_{2i} + e_i, \;\; i = 1, \ldots, n.$$

- Posteriormente, pode-se ter interesse em avaliar se há diferença por gênero/sexo.
- Uma primeira alternativa, seria ajustar um MRLS para cada sexo.

Variáveis dummy

- Como incorporar este tipo de variável no modelo de regressão?
- Isso pode ser feito através da inclusão de variáveis indicadoras, também denominadas de variáveis *dummy*.
- Considere a situação hipotética em que o interesse é modelar o salário (y) em função da experiência (x_2 anos) no cargo. Inicialmente, pode-se considerar um MRLS do tipo

$$y_i = \beta_0 + \beta_2 x_{2i} + e_i, \; i = 1, \ldots, n.$$

- Posteriormente, pode-se ter interesse em avaliar se há diferença por gênero/sexo.
- Uma primeira alternativa, seria ajustar um MRLS para cada sexo.

Variáveis dummy

- Como incorporar este tipo de variável no modelo de regressão?
- Isso pode ser feito através da inclusão de variáveis indicadoras, também denominadas de variáveis *dummy*.
- Considere a situação hipotética em que o interesse é modelar o salário (y) em função da experiência (x_2 anos) no cargo. Inicialmente, pode-se considerar um MRLS do tipo

$$y_i = \beta_0 + \beta_2 x_{2i} + e_i, \ i = 1, \ldots, n.$$

- Posteriormente, pode-se ter interesse em avaliar se há diferença por gênero/sexo.
- Uma primeira alternativa, seria ajustar um MRLS para cada sexo.

Variáveis dummy

- Admitindo que se existir diferença, seja somente no salário inicial, poderíamos considerar os dois modelos:

  Sexo masculino:

  $$y_i = \beta_0 + \beta_2 x_{2i} + e_i, i = 1, \ldots, n_1.$$

  Sexo feminino:

  $$y_i = \beta_0^* + \beta_2 x_{2i} + e_i, i = n_1 + 1, \ldots, n.$$

- Qual a desvantagem desta abordagem?
- Como podemos expressar a hipótese de que não existe diferença entre os gêneros?
- $\mathcal{H}_0 : \beta_0 = \beta_0^*$.

Variáveis dummy

- Admitindo que se existir diferença, seja somente no salário inicial, poderíamos considerar os dois modelos:

  Sexo masculino:

  $$y_i = \beta_0 + \beta_2 x_{2i} + e_i, i = 1, \ldots, n_1.$$

  Sexo feminino:

  $$y_i = \beta_0^* + \beta_2 x_{2i} + e_i, i = n_1 + 1, \ldots, n.$$

- Qual a desvantagem desta abordagem?
- Como podemos expressar a hipótese de que não existe diferença entre os gêneros?
- $\mathcal{H}_0 : \beta_0 = \beta_0^*.$

Variáveis dummy

- Admitindo que se existir diferença, seja somente no salário inicial, poderíamos considerar os dois modelos:

  Sexo masculino:

$$y_i = \beta_0 + \beta_2 x_{2i} + e_i, i = 1, \ldots, n_1.$$

  Sexo feminino:

$$y_i = \beta_0^* + \beta_2 x_{2i} + e_i, i = n_1 + 1, \ldots, n.$$

- Qual a desvantagem desta abordagem?
- Como podemos expressar a hipótese de que não existe diferença entre os gêneros?
- $\mathcal{H}_0 : \beta_0 = \beta_0^*$.

Variáveis dummy

- Admitindo que se existir diferença, seja somente no salário inicial, poderíamos considerar os dois modelos:

  Sexo masculino:

  $$y_i = \beta_0 + \beta_2 x_{2i} + e_i, i = 1, \ldots, n_1.$$

  Sexo feminino:

  $$y_i = \beta_0^* + \beta_2 x_{2i} + e_i, i = n_1 + 1, \ldots, n.$$

- Qual a desvantagem desta abordagem?
- Como podemos expressar a hipótese de que não existe diferença entre os gêneros?
- $\mathcal{H}_0 : \beta_0 = \beta_0^*$.

Variáveis dummy

- Pode-se definir a variável indicadora/dummy

$$x_{1i} := \begin{cases} 0 & \text{, se o } i\text{-ésimo indívíduo for do sexo masculino} \\ 1 & \text{, se o } i\text{-ésimo indívíduo for do sexo feminino,} \end{cases}$$

  e considerar o modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \ i = 1, \ldots, n. \tag{1}$$

- Desta forma, temos que

  Sexo masculino:

$$y_i = \beta_0 + \beta_2 x_{2i} + e_i, i = 1, \ldots, n_1.$$

  Sexo feminino:

$$y_i = (\beta_0 + \beta_1) + \beta_2 x_{2i} + e_i, i = n_1 + 1, \ldots, n.$$

- Quais suposições devemos fazer para conseguir ajustar e fazer inferências a respeito do modelo (1)?
- Qual a interpretação dos parâmetros do modelo (1)?

Variáveis dummy

- Pode-se definir a variável indicadora/dummy

$$x_{1i} := \begin{cases} 0 & \text{, se o } i\text{-ésimo indíviduo for do sexo masculino} \\ 1 & \text{, se o } i\text{-ésimo indíviduo for do sexo feminino,} \end{cases}$$

e considerar o modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \ i = 1, \ldots, n. \tag{1}$$

- Desta forma, temos que

Sexo masculino:

$$y_i = \beta_0 + \beta_2 x_{2i} + e_i, i = 1, \ldots, n_1.$$

Sexo feminino:

$$y_i = (\beta_0 + \beta_1) + \beta_2 x_{2i} + e_i, i = n_1 + 1, \ldots, n.$$

- Quais suposições devemos fazer para conseguir ajustar e fazer inferências a respeito do modelo (1)?
- Qual a interpretação dos parâmetros do modelo (1)?

Variáveis dummy

- Pode-se definir a variável indicadora/dummy

$$x_{1i} := \begin{cases} 0 & \text{, se o } i\text{-ésimo indívíduo for do sexo masculino} \\ 1 & \text{, se o } i\text{-ésimo indívíduo for do sexo feminino,} \end{cases}$$

  e considerar o modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i,\ i = 1, \ldots, n. \quad (1)$$

- Desta forma, temos que

  Sexo masculino:

$$y_i = \beta_0 + \beta_2 x_{2i} + e_i, i = 1, \ldots, n_1.$$

  Sexo feminino:

$$y_i = (\beta_0 + \beta_1) + \beta_2 x_{2i} + e_i, i = n_1 + 1, \ldots, n.$$

- Quais suposições devemos fazer para conseguir ajustar e fazer inferências a respeito do modelo (1)?

- Qual a interpretação dos parâmetros do modelo (1)?

Variáveis dummy

- Pode-se definir a variável indicadora/dummy

$$x_{1i} := \begin{cases} 0 & \text{, se o } i\text{-ésimo indíviduo for do sexo masculino} \\ 1 & \text{, se o } i\text{-ésimo indíviduo for do sexo feminino,} \end{cases}$$

e considerar o modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \; i = 1, \ldots, n. \tag{1}$$

- Desta forma, temos que

Sexo masculino:

$$y_i = \beta_0 + \beta_2 x_{2i} + e_i, i = 1, \ldots, n_1.$$

Sexo feminino:

$$y_i = (\beta_0 + \beta_1) + \beta_2 x_{2i} + e_i, i = n_1 + 1, \ldots, n.$$

- Quais suposições devemos fazer para conseguir ajustar e fazer inferências a respeito do modelo (1)?
- Qual a interpretação dos parâmetros do modelo (1)?

Variáveis dummy

- Como podemos expressar a hipótese de que não existe diferença entre os gêneros?
- $\mathcal{H}_0 : \beta_1 = 0.$
- Os estimadores dos parâmetros de localização (regressão) são **exatamente iguais** aqueles obtidos dos dois modelos marginais. Qual a vantagem desta abordagem?
- Ao se considerar a diferença somente no salário inicial, estamos ajustando duas retas paralelas.

Variáveis dummy

- Como podemos expressar a hipótese de que não existe diferença entre os gêneros?
- $\mathcal{H}_0 : \beta_1 = 0$.
- Os estimadores dos parâmetros de localização (regressão) são **exatamente iguais** aqueles obtidos dos dois modelos marginais. Qual a vantagem desta abordagem?
- Ao se considerar a diferença somente no salário inicial, estamos ajustando duas retas paralelas.

Variáveis dummy

- Como podemos expressar a hipótese de que não existe diferença entre os gêneros?
- $\mathcal{H}_0 : \beta_1 = 0$.
- Os estimadores dos parâmetros de localização (regressão) são **exatamente iguais** aqueles obtidos dos dois modelos marginais. Qual a vantagem desta abordagem?
- Ao se considerar a diferença somente no salário inicial, estamos ajustando duas retas paralelas.

Variáveis dummy

- Como podemos expressar a hipótese de que não existe diferença entre os gêneros?
- $\mathcal{H}_0 : \beta_1 = 0$.
- Os estimadores dos parâmetros de localização (regressão) são **exatamente iguais** aqueles obtidos dos dois modelos marginais. Qual a vantagem desta abordagem?
- Ao se considerar a diferença somente no salário inicial, estamos ajustando duas retas paralelas.

Ilustração ajuste retas paralelas

Figura: Funções de regressão associadas ao modelo (1), considerando β_0, β_1 e β_2 positivos.

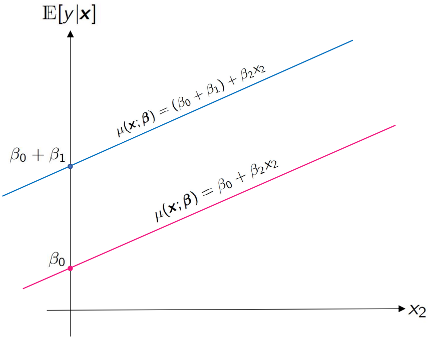


Ilustração ajuste retas paralelas

- Podemos fazer diferente, por exemplo definir duas variáveis dummy do tipo:

$$x_{1i} := \begin{cases} 0 & \text{, se o } i\text{-ésimo indíviduo for do sexo masculino} \\ 1 & \text{, se o } i\text{-ésimo indíviduo for do sexo feminino,} \end{cases}$$

e

$$x_{1i}^* := \begin{cases} 1 & \text{, se o } i\text{-ésimo indíviduo for do sexo masculino} \\ 0 & \text{, se o } i\text{-ésimo indíviduo for do sexo feminino,} \end{cases}$$

e considerar o modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_1^* x_{1i}^* + \beta_2 x_{2i} + e_i, \ i = 1, \ldots, n? \tag{2}$$

- A resposta é **não**, pois o modelo (2) é não identificável, dado que a matriz de especificação $\boldsymbol{X}$ correspondente não tem posto completo.
- Todavia se retirarmos o intercepto, podemos sim ajustar o modelo com as duas variáveis dummy

$$y_i = \beta_1 x_{1i} + \beta_1^* x_{1i}^* + \beta_2 x_{2i} + e_i, \ i = 1, \ldots, n. \tag{3}$$

- Neste caso, qual a interpretação dos parâmetros do modelo (3)?

Ilustração ajuste retas paralelas

- Podemos fazer diferente, por exemplo definir duas variáveis dummy do tipo:

$$x_{1i} := \begin{cases} 0 & \text{, se o } i\text{-ésimo indívíduo for do sexo masculino} \\ 1 & \text{, se o } i\text{-ésimo indívíduo for do sexo feminino,} \end{cases}$$

e

$$x_{1i}^{*} := \begin{cases} 1 & \text{, se o } i\text{-ésimo indívíduo for do sexo masculino} \\ 0 & \text{, se o } i\text{-ésimo indívíduo for do sexo feminino,} \end{cases}$$

e considerar o modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_1^{*} x_{1i}^{*} + \beta_2 x_{2i} + e_i,\ i = 1, \ldots, n? \tag{2}$$

- A resposta é **não**, pois o modelo (2) é não identificável, dado que a matriz de especificação $\boldsymbol{X}$ correspondente não tem posto completo.
- Todavia se retirarmos o intercepto, podemos sim ajustar o modelo com as duas variáveis dummy

$$y_i = \beta_1 x_{1i} + \beta_1^{*} x_{1i}^{*} + \beta_2 x_{2i} + e_i,\ i = 1, \ldots, n. \tag{3}$$

- Neste caso, qual a interpretação dos parâmetros do modelo (3)?

Ilustração ajuste retas paralelas

- Podemos fazer diferente, por exemplo definir duas variáveis dummy do tipo:

$$x_{1i} := \begin{cases} 0 & \text{, se o } i\text{-ésimo indíviduo for do sexo masculino} \\ 1 & \text{, se o } i\text{-ésimo indíviduo for do sexo feminino,} \end{cases}$$

e

$$x^*_{1i} := \begin{cases} 1 & \text{, se o } i\text{-ésimo indíviduo for do sexo masculino} \\ 0 & \text{, se o } i\text{-ésimo indíviduo for do sexo feminino,} \end{cases}$$

e considerar o modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta^*_1 x^*_{1i} + \beta_2 x_{2i} + e_i,\ i = 1, \ldots, n? \quad (2)$$

- A resposta é **não**, pois o modelo (2) é não identificável, dado que a matriz de especificação $\boldsymbol{X}$ correspondente não tem posto completo.
- Todavia se retirarmos o intercepto, podemos sim ajustar o modelo com as duas variáveis dummy

$$y_i = \beta_1 x_{1i} + \beta^*_1 x^*_{1i} + \beta_2 x_{2i} + e_i,\ i = 1, \ldots, n. \quad (3)$$

- Neste caso, qual a interpretação dos parâmetros do modelo (3)?

Ilustração ajuste retas paralelas

- Podemos fazer diferente, por exemplo definir duas variáveis dummy do tipo:

$$x_{1i} := \begin{cases} 0 & \text{, se o } i\text{-ésimo indíviduo for do sexo masculino} \\ 1 & \text{, se o } i\text{-ésimo indíviduo for do sexo feminino,} \end{cases}$$

e

$$x_{1i}^* := \begin{cases} 1 & \text{, se o } i\text{-ésimo indíviduo for do sexo masculino} \\ 0 & \text{, se o } i\text{-ésimo indíviduo for do sexo feminino,} \end{cases}$$

e considerar o modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_1^* x_{1i}^* + \beta_2 x_{2i} + e_i,\ i = 1, \ldots, n? \tag{2}$$

- A resposta é **não**, pois o modelo (2) é não identificável, dado que a matriz de especificação $\boldsymbol{X}$ correspondente não tem posto completo.
- Todavia se retirarmos o intercepto, podemos sim ajustar o modelo com as duas variáveis dummy

$$y_i = \beta_1 x_{1i} + \beta_1^* x_{1i}^* + \beta_2 x_{2i} + e_i,\ i = 1, \ldots, n. \tag{3}$$

- Neste caso, qual a interpretação dos parâmetros do modelo (3)?

Exercício (entregar próxima aula)

Exercício 1: Reescreva os modelos (1), (2) e (3) na forma matricial e interprete seus parâmetros.

Exercício 2: Considere o exemplo hipotético (y representando salário) e os seguintes modelos:

$$M_1: \; y_i = \beta_0 + \beta_1 x_{1i} + e_i \text{ e } M_2: \; y_i = \beta_1 x_{1i} + \beta_2 x_{1i}^* + e_i,$$

em que

$$x_{1i} := \begin{cases} 0 & \text{, se o } i\text{-ésimo indíviduo for do sexo masculino} \\ 1 & \text{, se o } i\text{-ésimo indíviduo for do sexo feminino,} \end{cases}$$

e

$$x_{1i}^* := \begin{cases} 1 & \text{, se o } i\text{-ésimo indíviduo for do sexo masculino} \\ 0 & \text{, se o } i\text{-ésimo indíviduo for do sexo feminino,} \end{cases}$$

i) Interprete os parâmetros dos modelos M_1 e M_2.

ii) Reescreva-os na forma matricial. Obtenha o EMQ dos parâmetros.

iii) Expresse a hipótese de que não existe diferença de salário por sexo.

iv) Através do teste t, obtenha as estatísticas de teste relacionadas ao item iii). Esta estatística você já conhecia? Comente.

Variáveis dummy

- Se tivermos mais do que duas categorias? Como proceder?
- Por exemplo, imagine que desejamos incluir a variável explicativa que representa o grau de instrução, no qual seus níveis são: até segundo grau, superior completo e pós-graduado, ou seja, com 3 níveis.
- Neste caso, precisamos de duas variáveis dummy, como no esboço abaixo:

Nível	x_1	x_2
até segundo grau	0	0
superior completo	1	0
pós-graduado	0	1

- Considerando y como sendo o salário, poderíamos adotar o seguinte MRLM:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, i = 1, \ldots, n. \tag{4}$$

Variáveis dummy

- Se tivermos mais do que duas categorias? Como proceder?
- Por exemplo, imagine que desejamos incluir a variável explicativa que representa o grau de instrução, no qual seus níveis são: até segundo grau, superior completo e pós-graduado, ou seja, com 3 níveis.
- Neste caso, precisamos de duas variáveis dummy, como no esboço abaixo:

Nível	x_1	x_2
até segundo grau	0	0
superior completo	1	0
pós-graduado	0	1

- Considerando y como sendo o salário, poderíamos adotar o seguinte MRLM:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, i = 1, \ldots, n. \tag{4}$$

Variáveis dummy

- Se tivermos mais do que duas categorias? Como proceder?
- Por exemplo, imagine que desejamos incluir a variável explicativa que representa o grau de instrução, no qual seus níveis são: até segundo grau, superior completo e pós-graduado, ou seja, com 3 níveis.
- Neste caso, precisamos de duas variáveis dummy, como no esboço abaixo:

Nível	x_1	x_2
até segundo grau	0	0
superior completo	1	0
pós-graduado	0	1

- Considerando y como sendo o salário, poderíamos adotar o seguinte MRLM:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, i = 1, \ldots, n. \tag{4}$$

Variáveis dummy

- Se tivermos mais do que duas categorias? Como proceder?
- Por exemplo, imagine que desejamos incluir a variável explicativa que representa o grau de instrução, no qual seus níveis são: até segundo grau, superior completo e pós-graduado, ou seja, com 3 níveis.
- Neste caso, precisamos de duas variáveis dummy, como no esboço abaixo:

Nível	x_1	x_2
até segundo grau	0	0
superior completo	1	0
pós-graduado	0	1

- Considerando y como sendo o salário, poderíamos adotar o seguinte MRLM:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, i = 1, \ldots, n. \tag{4}$$

Variáveis dummy

- Note que neste caso a casela de referência, associada ao intercepto β_0, é justamente o primeiro nível.
- Qual a interpretação dos parâmetros do modelo (4)? Como podemos expressar a hipótese de que o salário médio não difere por nível de instrução? Como podemos expressar a hipótese de que o salário médio cresce com o nível de instrução?
- Podemos usar outras parametrizações, como por exemplo, desvios com restrição, em que os parâmetros representam o desvio com relação a média geral marginal.

Variáveis dummy

- Note que neste caso a casela de referência, associada ao intercepto β_0, é justamente o primeiro nível.
- Qual a interpretação dos parâmetros do modelo (4)? Como podemos expressar a hipótese de que o salário médio não difere por nível de instrução? Como podemos expressar a hipótese de que o salário médio cresce com o nível de instrução?
- Podemos usar outras parametrizações, como por exemplo, desvios com restrição, em que os parâmetros representam o desvio com relação a média geral marginal.

Variáveis dummy

- Note que neste caso a casela de referência, associada ao intercepto β_0, é justamente o primeiro nível.
- Qual a interpretação dos parâmetros do modelo (4)? Como podemos expressar a hipótese de que o salário médio não difere por nível de instrução? Como podemos expressar a hipótese de que o salário médio cresce com o nível de instrução?
- Podemos usar outras parametrizações, como por exemplo, desvios com restrição, em que os parâmetros representam o desvio com relação a média geral marginal.

Variáveis dummy

- No exemplo em questão, continuaríamos com duas variáveis dummy e para garantir a identificabilidade do modelo, impomos a restrição

$$\beta_1 + \beta_2 + \beta_3 = 0,$$

de forma que

Nível	x_1	x_2
até segundo grau	1	0
superior completo	0	1
pós-graduado	-1	-1

- Neste caso, a média de salário para quem possui até segundo grau é dada por $\beta_0 + \beta_1$, para quem possui nível superior completo é $\beta_0 + \beta_2$ e para quem é pós-graduado é de $\beta_0 + \beta_3 = \beta_0 - (\beta_1 + \beta_2)$.

- O parâmetro β_0 representa a média salarial geral, desconsiderando o grau de instrução.

Variáveis dummy

- No exemplo em questão, continuaríamos com duas variáveis dummy e para garantir a identificabilidade do modelo, impomos a restrição

$$\beta_1 + \beta_2 + \beta_3 = 0,$$

de forma que

Nível	x_1	x_2
até segundo grau	1	0
superior completo	0	1
pós-graduado	-1	-1

- Neste caso, a média de salário para quem possui até segundo grau é dada por $\beta_0 + \beta_1$, para quem possui nível superior completo é $\beta_0 + \beta_2$ e para quem é pós-graduado é de $\beta_0 + \beta_3 = \beta_0 - (\beta_1 + \beta_2)$.

- O parâmetro β_0 representa a média salarial geral, desconsiderando o grau de instrução.

Variáveis dummy

- No exemplo em questão, continuaríamos com duas variáveis dummy e para garantir a identificabilidade do modelo, impomos a restrição

$$\beta_1 + \beta_2 + \beta_3 = 0,$$

de forma que

Nível	x_1	x_2
até segundo grau	1	0
superior completo	0	1
pós-graduado	-1	-1

- Neste caso, a média de salário para quem possui até segundo grau é dada por $\beta_0 + \beta_1$, para quem possui nível superior completo é $\beta_0 + \beta_2$ e para quem é pós-graduado é de $\beta_0 + \beta_3 = \beta_0 - (\beta_1 + \beta_2)$.
- O parâmetro β_0 representa a média salarial geral, desconsiderando o grau de instrução.

Variáveis dummy

- O parâmetro β_1 representa o desvio da média salarial de quem possui até segundo grau, comparativamente a população geral. Se $\beta_1 < 0$, isso implica que em média as pessoas com no máximo segundo grau ganham menos que a média da população geral.
- Como podemos expressar a hipótese de que o salário médio não difere por nível de instrução? Como podemos expressar a hipótese de que o salário médio cresce com o nível de instrução?
- Qual a interpretação de β_0 nas duas parametrizações?
- O que significa testar $\mathcal{H}_0 : \beta_1 = 0$ nas duas parametrizações?
- Ainda tem uma terceira parametrização, que a parametrização de médias, na qual não se considera o intercepto no modelo e utilizamos três variáveis dummy (uma indicadora para cada nível).
- No software `R` a parametrização casela de referência é a padrão. Para usar desvios com restrição, deve-se usar o comando `'contr.sum'`. No `SAS` tem-se uma flexibilidade maior para considerar a parametrização mais conveniente.

Variáveis dummy

- O parâmetro β_1 representa o desvio da média salarial de quem possui até segundo grau, comparativamente a população geral. Se $\beta_1 < 0$, isso implica que em média as pessoas com no máximo segundo grau ganham menos que a média da população geral.
- Como podemos expressar a hipótese de que o salário médio não difere por nível de instrução? Como podemos expressar a hipótese de que o salário médio cresce com o nível de instrução?
- Qual a interpretação de β_0 nas duas parametrizações?
- O que significa testar $\mathcal{H}_0 : \beta_1 = 0$ nas duas parametrizações?
- Ainda tem uma terceira parametrização, que a parametrização de médias, na qual não se considera o intercepto no modelo e utilizamos três variáveis dummy (uma indicadora para cada nível).
- No software `R` a parametrização casela de referência é a padrão. Para usar desvios com restrição, deve-se usar o comando `'contr.sum'`. No `SAS` tem-se uma flexibilidade maior para considerar a parametrização mais conveniente.

Variáveis dummy

- O parâmetro β_1 representa o desvio da média salarial de quem possui até segundo grau, comparativamente a população geral. Se $\beta_1 < 0$, isso implica que em média as pessoas com no máximo segundo grau ganham menos que a média da população geral.
- Como podemos expressar a hipótese de que o salário médio não difere por nível de instrução? Como podemos expressar a hipótese de que o salário médio cresce com o nível de instrução? 😇
- Qual a interpretação de β_0 nas duas parametrizações?
- O que significa testar $\mathcal{H}_0 : \beta_1 = 0$ nas duas parametrizações?
- Ainda tem uma terceira parametrização, que a parametrização de médias, na qual não se considera o intercepto no modelo e utilizamos três variáveis dummy (uma indicadora para cada nível).
- No software `R` a parametrização casela de referência é a padrão. Para usar desvios com restrição, deve-se usar o comando `'contr.sum'`. No `SAS` tem-se uma flexibilidade maior para considerar a parametrização mais conveniente.

Variáveis dummy

- O parâmetro β_1 representa o desvio da média salarial de quem possui até segundo grau, comparativamente a população geral. Se $\beta_1 < 0$, isso implica que em média as pessoas com no máximo segundo grau ganham menos que a média da população geral.
- Como podemos expressar a hipótese de que o salário médio não difere por nível de instrução? Como podemos expressar a hipótese de que o salário médio cresce com o nível de instrução? 😇
- Qual a interpretação de β_0 nas duas parametrizações?
- O que significa testar $\mathcal{H}_0 : \beta_1 = 0$ nas duas parametrizações?
- Ainda tem uma terceira parametrização, que a parametrização de médias, na qual não se considera o intercepto no modelo e utilizamos três variáveis dummy (uma indicadora para cada nível).
- No software `R` a parametrização casela de referência é a padrão. Para usar desvios com restrição, deve-se usar o comando `'contr.sum'`. No `SAS` tem-se uma flexibilidade maior para considerar a parametrização mais conveniente.

Variáveis dummy

- O parâmetro β_1 representa o desvio da média salarial de quem possui até segundo grau, comparativamente a população geral. Se $\beta_1 < 0$, isso implica que em média as pessoas com no máximo segundo grau ganham menos que a média da população geral.
- Como podemos expressar a hipótese de que o salário médio não difere por nível de instrução? Como podemos expressar a hipótese de que o salário médio cresce com o nível de instrução?
- Qual a interpretação de β_0 nas duas parametrizações?
- O que significa testar $\mathcal{H}_0 : \beta_1 = 0$ nas duas parametrizações?
- Ainda tem uma terceira parametrização, que a parametrização de médias, na qual não se considera o intercepto no modelo e utilizamos três variáveis dummy (uma indicadora para cada nível).
- No software R a parametrização casela de referência é a padrão. Para usar desvios com restrição, deve-se usar o comando 'contr.sum'. No SAS tem-se uma flexibilidade maior para considerar a parametrização mais conveniente.

Variáveis dummy

- O parâmetro β_1 representa o desvio da média salarial de quem possui até segundo grau, comparativamente a população geral. Se $\beta_1 < 0$, isso implica que em média as pessoas com no máximo segundo grau ganham menos que a média da população geral.
- Como podemos expressar a hipótese de que o salário médio não difere por nível de instrução? Como podemos expressar a hipótese de que o salário médio cresce com o nível de instrução?
- Qual a interpretação de β_0 nas duas parametrizações?
- O que significa testar $\mathcal{H}_0 : \beta_1 = 0$ nas duas parametrizações?
- Ainda tem uma terceira parametrização, que a parametrização de médias, na qual não se considera o intercepto no modelo e utilizamos três variáveis dummy (uma indicadora para cada nível).
- No software `R` a parametrização casela de referência é a padrão. Para usar desvios com restrição, deve-se usar o comando `'contr.sum'`. No `SAS` tem-se uma flexibilidade maior para considerar a parametrização mais conveniente.

Ajuste de retas com intercepto e inclinação diferentes

- Retornemos ao nosso exemplo hipotético em que o interesse consiste em modelar o salário (y) em função da experiência (x_2 em anos) no cargo e do gênero/sexo. Agora, vamos permitir que exista efeito do gênero tanto no salário inicial como na taxa de variação do salário médio ao longo dos anos. Novamente, considerando a variável dummy

$$x_{1i} := \begin{cases} 0 & \text{, se o } i\text{-ésimo indíviduo for do sexo masculino} \\ 1 & \text{, se o } i\text{-ésimo indíviduo for do sexo feminino,} \end{cases}$$

pode-se considerar um MRLS do tipo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + e_i, \; i = 1, \ldots, n. \tag{5}$$

- Desta forma, temos que

<u>Sexo masculino</u>:

$$y_i = \beta_0 + \beta_2 x_{2i} + e_i, i = 1, \ldots, n_1.$$

<u>Sexo feminino</u>:

$$y_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) x_{2i} + e_i, i = n_1 + 1, \ldots, n.$$

Ajuste de retas com intercepto e inclinação diferentes

- Retornemos ao nosso exemplo hipotético em que o interesse consiste em modelar o salário (y) em função da experiência (x_2 em anos) no cargo e do gênero/sexo. Agora, vamos permitir que exista efeito do gênero tanto no salário inicial como na taxa de variação do salário médio ao longo dos anos. Novamente, considerando a variável dummy

$$x_{1i} := \begin{cases} 0 & \text{, se o } i\text{-ésimo indíviduo for do sexo masculino} \\ 1 & \text{, se o } i\text{-ésimo indíviduo for do sexo feminino,} \end{cases}$$

pode-se considerar um MRLS do tipo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + e_i, \; i = 1, \ldots, n. \tag{5}$$

- Desta forma, temos que

Sexo masculino:

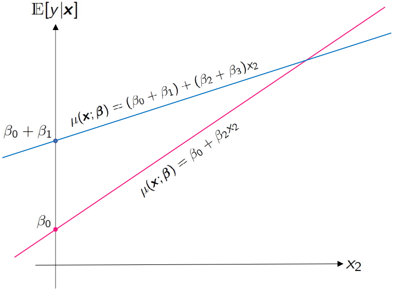
$$y_i = \beta_0 + \beta_2 x_{2i} + e_i, i = 1, \ldots, n_1.$$

Sexo feminino:

$$y_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) x_{2i} + e_i, i = n_1 + 1, \ldots, n.$$

Ilustração ajuste retas com intercepto e inclinação diferentes

Figura: Funções de regressão associadas ao modelo (5), considerando $\beta_0, \beta_1, \beta_2 > 0$ e $\beta_3 < 0$.



Variáveis dummy

- Com base no modelo (5), como podemos expressar as hipóteses de que:
- Não há efeito de sexo?
- $\mathcal{H}_0 : \beta_1 = \beta_3 = 0$
- Testar que não existe interação, i.e., a relação entre o salário médio e os anos de experiência é exatamente o mesmo para ambos os gêneros, implicando que as retas são paralelas.
- $\mathcal{H}_0 : \beta_3 = 0$
- Testar se ao longo do tempo, o aumento salarial é maior para os homens.
- $\mathcal{H}_0 : \beta_3 < 0$.
- Para ajustar este modelo no software `R` basta fazer

  `lm(y~sexo+x2+x2*sexo)`.
- Cabe salientar que podemos ter mais que uma variável dummy no modelo, especialmente, considerando efeito de interação entre elas.

Variáveis dummy

- Com base no modelo (5), como podemos expressar as hipóteses de que:
- Não há efeito de sexo?
- $\mathcal{H}_0 : \beta_1 = \beta_3 = 0$
- Testar que não existe interação, i.e., a relação entre o salário médio e os anos de experiência é exatamente o mesmo para ambos os gêneros, implicando que as retas são paralelas.
- $\mathcal{H}_0 : \beta_3 = 0$
- Testar se ao longo do tempo, o aumento salarial é maior para os homens.
- $\mathcal{H}_0 : \beta_3 < 0$.
- Para ajustar este modelo no software R basta fazer

  `lm(y~sexo+x2+x2*sexo).`
- Cabe salientar que podemos ter mais que uma variável dummy no modelo, especialmente, considerando efeito de interação entre elas.

Variáveis dummy

- Com base no modelo (5), como podemos expressar as hipóteses de que:
- Não há efeito de sexo?
- $\mathcal{H}_0 : \beta_1 = \beta_3 = 0$
- Testar que não existe interação, i.e., a relação entre o salário médio e os anos de experiência é exatamente o mesmo para ambos os gêneros, implicando que as retas são paralelas.
- $\mathcal{H}_0 : \beta_3 = 0$
- Testar se ao longo do tempo, o aumento salarial é maior para os homens.
- $\mathcal{H}_0 : \beta_3 < 0$.
- Para ajustar este modelo no software `R` basta fazer

  `lm(y~sexo+x2+x2*sexo).`
- Cabe salientar que podemos ter mais que uma variável dummy no modelo, especialmente, considerando efeito de interação entre elas.

Variáveis dummy

- Com base no modelo (5), como podemos expressar as hipóteses de que:
- Não há efeito de sexo?
- $\mathcal{H}_0 : \beta_1 = \beta_3 = 0$
- Testar que não existe interação, i.e., a relação entre o salário médio e os anos de experiência é exatamente o mesmo para ambos os gêneros, implicando que as retas são paralelas.
- $\mathcal{H}_0 : \beta_3 = 0$
- Testar se ao longo do tempo, o aumento salarial é maior para os homens.
- $\mathcal{H}_0 : \beta_3 < 0$.
- Para ajustar este modelo no software `R` basta fazer

  `lm(y~sexo+x2+x2*sexo)`.
- Cabe salientar que podemos ter mais que uma variável dummy no modelo, especialmente, considerando efeito de interação entre elas.

Variáveis dummy

- Com base no modelo (5), como podemos expressar as hipóteses de que:
- Não há efeito de sexo?
- $\mathcal{H}_0 : \beta_1 = \beta_3 = 0$
- Testar que não existe interação, i.e., a relação entre o salário médio e os anos de experiência é exatamente o mesmo para ambos os gêneros, implicando que as retas são paralelas.
- $\mathcal{H}_0 : \beta_3 = 0$
- Testar se ao longo do tempo, o aumento salarial é maior para os homens.
- $\mathcal{H}_0 : \beta_3 < 0$.
- Para ajustar este modelo no software R basta fazer

  `lm(y~sexo+x2+x2*sexo)`.
- Cabe salientar que podemos ter mais que uma variável dummy no modelo, especialmente, considerando efeito de interação entre elas.

Variáveis dummy

- Com base no modelo (5), como podemos expressar as hipóteses de que:
- Não há efeito de sexo?
- $\mathcal{H}_0 : \beta_1 = \beta_3 = 0$
- Testar que não existe interação, i.e., a relação entre o salário médio e os anos de experiência é exatamente o mesmo para ambos os gêneros, implicando que as retas são paralelas.
- $\mathcal{H}_0 : \beta_3 = 0$
- Testar se ao longo do tempo, o aumento salarial é maior para os homens.
- $\mathcal{H}_0 : \beta_3 < 0.$
- Para ajustar este modelo no software `R` basta fazer

  `lm(y~sexo+x2+x2*sexo).`
- Cabe salientar que podemos ter mais que uma variável dummy no modelo, especialmente, considerando efeito de interação entre elas.

Variáveis dummy

- Com base no modelo (5), como podemos expressar as hipóteses de que:
- Não há efeito de sexo?
- $\mathcal{H}_0 : \beta_1 = \beta_3 = 0$
- Testar que não existe interação, i.e., a relação entre o salário médio e os anos de experiência é exatamente o mesmo para ambos os gêneros, implicando que as retas são paralelas.
- $\mathcal{H}_0 : \beta_3 = 0$
- Testar se ao longo do tempo, o aumento salarial é maior para os homens.
- $\mathcal{H}_0 : \beta_3 < 0.$
- Para ajustar este modelo no software R basta fazer

  `lm(y~sexo+x2+x2*sexo).`
- Cabe salientar que podemos ter mais que uma variável dummy no modelo, especialmente, considerando efeito de interação entre elas.

Variáveis dummy

- Com base no modelo (5), como podemos expressar as hipóteses de que:
- Não há efeito de sexo?
- $\mathcal{H}_0 : \beta_1 = \beta_3 = 0$
- Testar que não existe interação, i.e., a relação entre o salário médio e os anos de experiência é exatamente o mesmo para ambos os gêneros, implicando que as retas são paralelas.
- $\mathcal{H}_0 : \beta_3 = 0$
- Testar se ao longo do tempo, o aumento salarial é maior para os homens.
- $\mathcal{H}_0 : \beta_3 < 0$.
- Para ajustar este modelo no software `R` basta fazer

  `lm(y~sexo+x2+x2*sexo).`
- Cabe salientar que podemos ter mais que uma variável dummy no modelo, especialmente, considerando efeito de interação entre elas.

Variáveis dummy

- Com base no modelo (5), como podemos expressar as hipóteses de que:
- Não há efeito de sexo?
- $\mathcal{H}_0 : \beta_1 = \beta_3 = 0$
- Testar que não existe interação, i.e., a relação entre o salário médio e os anos de experiência é exatamente o mesmo para ambos os gêneros, implicando que as retas são paralelas.
- $\mathcal{H}_0 : \beta_3 = 0$
- Testar se ao longo do tempo, o aumento salarial é maior para os homens.
- $\mathcal{H}_0 : \beta_3 < 0$.
- Para ajustar este modelo no software `R` basta fazer

  `lm(y~sexo+x2+x2*sexo).`
- Cabe salientar que podemos ter mais que uma variável dummy no modelo, especialmente, considerando efeito de interação entre elas.

Ajuste de Poligonais

- Em muitas situações práticas, podemos ter interesse em ajustar modelos de regressão segmentadas, i.e., modelos de regressão em que a forma funcional se modifica. Em Hoffman (2016) ele usa a denominação poligonal.
- Em livros de Economia, é comum usar o termo mudança estrutural. Tanto modelos de regressão segmentadas como modelos de mudança estrutural são bem mais gerais.
- Aqui, usaremos as variáveis dummy para captar a mudança na inclinação entre segmentos consecutivos da poligonal.
- O modelo geral para uma poligonal com k vértices ($k+1$ segmentos) é

$$y_i = \beta_0 + \beta_1 x_i + \sum_{h=1}^{k} \gamma_h z_{hi}(x_i - \theta_h) + e_i, \; i = 1, \ldots, n, \tag{6}$$

em que θ_h é a abcissa do h-ésimo vértice (que pressupomos conhecida) e z_{hi} é uma variável indicadora definida por

$$z_{hi} = \mathbb{1}(x_i > \theta_h). \tag{7}$$

Ajuste de Poligonais

- Em muitas situações práticas, podemos ter interesse em ajustar modelos de regressão segmentadas, i.e., modelos de regressão em que a forma funcional se modifica. Em Hoffman (2016) ele usa a denominação poligonal.
- Em livros de Economia, é comum usar o termo mudança estrutural. Tanto modelos de regressão segmentadas como modelos de mudança estrutural são bem mais gerais.
- Aqui, usaremos as variáveis dummy para captar a mudança na inclinação entre segmentos consecutivos da poligonal.
- O modelo geral para uma poligonal com k vértices ($k+1$ segmentos) é

$$y_i = \beta_0 + \beta_1 x_i + \sum_{h=1}^{k} \gamma_h z_{hi}(x_i - \theta_h) + e_i, \ \ i = 1, \ldots, n, \tag{6}$$

em que θ_h é a abcissa do h-ésimo vértice (que pressupomos conhecida) e z_{hi} é uma variável indicadora definida por

$$z_{hi} = \mathbb{1}(x_i > \theta_h). \tag{7}$$

Ajuste de Poligonais

- Em muitas situações práticas, podemos ter interesse em ajustar modelos de regressão segmentadas, i.e., modelos de regressão em que a forma funcional se modifica. Em Hoffman (2016) ele usa a denominação poligonal.
- Em livros de Economia, é comum usar o termo mudança estrutural. Tanto modelos de regressão segmentadas como modelos de mudança estrutural são bem mais gerais.
- Aqui, usaremos as variáveis dummy para captar a mudança na inclinação entre segmentos consecutivos da poligonal.
- O modelo geral para uma poligonal com k vértices ($k+1$ segmentos) é

$$y_i = \beta_0 + \beta_1 x_i + \sum_{h=1}^{k} \gamma_h z_{hi}(x_i - \theta_h) + e_i, \; i = 1, \ldots, n, \tag{6}$$

em que θ_h é a abcissa do h-ésimo vértice (que pressupomos conhecida) e z_{hi} é uma variável indicadora definida por

$$z_{hi} = \mathbb{1}(x_i > \theta_h). \tag{7}$$

Ajuste de Poligonais

- Em muitas situações práticas, podemos ter interesse em ajustar modelos de regressão segmentadas, i.e., modelos de regressão em que a forma funcional se modifica. Em Hoffman (2016) ele usa a denominação poligonal.
- Em livros de Economia, é comum usar o termo mudança estrutural. Tanto modelos de regressão segmentadas como modelos de mudança estrutural são bem mais gerais.
- Aqui, usaremos as variáveis dummy para captar a mudança na inclinação entre segmentos consecutivos da poligonal.
- O modelo geral para uma poligonal com k vértices ($k+1$ segmentos) é

$$y_i = \beta_0 + \beta_1 x_i + \sum_{h=1}^{k} \gamma_h z_{hi}(x_i - \theta_h) + e_i,\ \ i = 1, \ldots, n, \tag{6}$$

em que θ_h é a abcissa do h-ésimo vértice (que pressupomos conhecida) e z_{hi} é uma variável indicadora definida por

$$z_{hi} = \mathbb{1}(x_i > \theta_h). \tag{7}$$

Ajuste de Poligonais

- Os parâmetros γ_h representam a mudança na inclinação do h-ésimo segmento da poligonal em relação à inclinação do segmento anterior.
- Para uma poligonal com 3 segmentos, o modelo (6) fica igual a

$$y_i = \beta_0 + \beta_1 x_i + \gamma_1 z_{1i}(x_i - \theta_1) + \gamma_2 z_{2i}(x_i - \theta_2) + e_i, \; i = 1, \ldots, n, \quad (8)$$

- Na próxima figura mostramos uma ilustração hipotética da função de regressão associada ao modelo acima.

Ajuste de Poligonais

- Os parâmetros γ_h representam a mudança na inclinação do h-ésimo segmento da poligonal em relação à inclinação do segmento anterior.

- Para uma poligonal com 3 segmentos, o modelo (6) fica igual a

$$y_i = \beta_0 + \beta_1 x_i + \gamma_1 z_{1i}(x_i - \theta_1) + \gamma_2 z_{2i}(x_i - \theta_2) + e_i,\ i = 1, \ldots, n, \tag{8}$$

- Na próxima figura mostramos uma ilustração hipotética da função de regressão associada ao modelo acima.

Ajuste de Poligonais

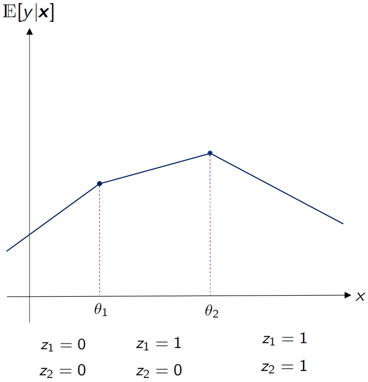
- Os parâmetros γ_h representam a mudança na inclinação do h-ésimo segmento da poligonal em relação à inclinação do segmento anterior.
- Para uma poligonal com 3 segmentos, o modelo (6) fica igual a

$$y_i = \beta_0 + \beta_1 x_i + \gamma_1 z_{1i}(x_i - \theta_1) + \gamma_2 z_{2i}(x_i - \theta_2) + e_i, \ i = 1, \ldots, n, \tag{8}$$

- Na próxima figura mostramos uma ilustração hipotética da função de regressão associada ao modelo acima.

Ilustração ajuste de poligonais

Figura: Função de regressão associada a uma poligonal com 3 segmentos (8) em que $\beta_0 > 0, \beta_1 > 0, \gamma_1 < 0, \gamma_2 < 0, \beta_1 + \gamma_1 > 0$ e $\beta_1 + \gamma_1 + \gamma_2 < 0$.



Ajuste de Poligonais

- Pelo modelo (8), tem-se que:
- Se $x_i \leq \theta_1$, então

$$y_i = \beta_0 + \beta_1 x_i + e_i.$$

- Se $\theta_1 < x_i \leq \theta_2$, então

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \gamma_1(x_i - \theta_1) + e_i \\ &= (\beta_0 - \gamma_1\theta_1) + (\beta_1 + \gamma_1)x_i + e_i. \end{aligned}$$

- Por fim, Se $x_i > \theta_2$, então

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \gamma_1(x_i - \theta_1) + \gamma_2(x_i - \theta_2) + e_i \\ &= (\beta_0 - \gamma_1\theta_1 - \gamma_2\theta_2) + (\beta_1 + \gamma_1 + \gamma_2)x_i + e_i. \end{aligned}$$

Ajuste de Poligonais

- Pelo modelo (8), tem-se que:
- Se $x_i \leq \theta_1$, então

$$y_i = \beta_0 + \beta_1 x_i + e_i.$$

- Se $\theta_1 < x_i \leq \theta_2$, então

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \gamma_1(x_i - \theta_1) + e_i \\ &= (\beta_0 - \gamma_1\theta_1) + (\beta_1 + \gamma_1)x_i + e_i. \end{aligned}$$

- Por fim, Se $x_i > \theta_2$, então

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \gamma_1(x_i - \theta_1) + \gamma_2(x_i - \theta_2) + e_i \\ &= (\beta_0 - \gamma_1\theta_1 - \gamma_2\theta_2) + (\beta_1 + \gamma_1 + \gamma_2)x_i + e_i. \end{aligned}$$

Ajuste de Poligonais

- Pelo modelo (8), tem-se que:
- Se $x_i \leq \theta_1$, então

$$y_i = \beta_0 + \beta_1 x_i + e_i.$$

- Se $\theta_1 < x_i \leq \theta_2$, então

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \gamma_1(x_i - \theta_1) + e_i \\ &= (\beta_0 - \gamma_1\theta_1) + (\beta_1 + \gamma_1)x_i + e_i. \end{aligned}$$

- Por fim, Se $x_i > \theta_2$, então

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \gamma_1(x_i - \theta_1) + \gamma_2(x_i - \theta_2) + e_i \\ &= (\beta_0 - \gamma_1\theta_1 - \gamma_2\theta_2) + (\beta_1 + \gamma_1 + \gamma_2)x_i + e_i. \end{aligned}$$

Ajuste de Poligonais

- Pelo modelo (8), tem-se que:
- Se $x_i \leq \theta_1$, então

$$y_i = \beta_0 + \beta_1 x_i + e_i.$$

- Se $\theta_1 < x_i \leq \theta_2$, então

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \gamma_1 (x_i - \theta_1) + e_i \\ &= (\beta_0 - \gamma_1 \theta_1) + (\beta_1 + \gamma_1) x_i + e_i. \end{aligned}$$

- Por fim, Se $x_i > \theta_2$, então

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \gamma_1 (x_i - \theta_1) + \gamma_2 (x_i - \theta_2) + e_i \\ &= (\beta_0 - \gamma_1 \theta_1 - \gamma_2 \theta_2) + (\beta_1 + \gamma_1 + \gamma_2) x_i + e_i. \end{aligned}$$

Exercício (entregar próxima aula)

Exercício 3: Reproduzir o exemplo 8.1 (The tool life data) de Montgomery et al. (2012) adicionando a parte de diagnóstico.

Exercício 4: Fazer o exercício 8.13 de Montgomery et al. (2012).

Lista III

Lista III: Fazer todos os exercícios do Cap 5 de Hoffman (2016) e exercícios 8.1-8.11 de Montgomery et al (2012).