DSBD - Modelos Lineares

Slides: Cesar Taconeli, Apres.: José Padilha

18 de agosto, 2018

Aula 5 - Regressão linear com covariáveis categóricas

Introdução

- Neste módulo vamos tratar da inclusão de covariáveis categóricas (fatores) em modelos de regressão.

- Alguns exemplos de covariáveis categóricas:
  - Sexo (masculino ou feminino);
  - Estação do ano (primavera, verão, outono ou inverno);
  - Categoria de cliente de banco (platinum, gold, silver,...);
  - Escolaridade (sem escolaridade, ensino primário, ensino secundário,...).

- A forma usual de incorporar covariáveis categóricas a um modelo de regressão é através de variáveis indicadoras (variáveis *dummy*).

Regressão linear com covariáveis categóricas

- Considere uma covariável categórica (fator) com dois níveis, A e B;
- A inclusão dessa covariável ao modelo de regressão requer a incorporação de uma única variável indicadora:

$$x = \begin{cases} 0 & \text{se categoria A} \\ 1, & \text{se categoria B} \end{cases} \tag{1}$$

- Supondo que essa seja a unica covariável na análise, o modelo de regressão linear ficaria dado por:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2}$$

Regressão linear com covariáveis categóricas

- Sob as suposições de uma regressão linear temos que:

$$E(y|x) = \begin{cases} \beta_0 & \text{se categoria A} \\ \beta_0 + \beta_1 & \text{se categoria B} \end{cases} \tag{3}$$

- Desta forma, β_0 corresponde à média de y para indivíduos da categoria A;
- Ainda, β_1 corresponde à diferença na média de y dos indivíduos da categoria B para os indivíduos da categoria A.
- Se no modelo houver outras covariáveis (categóricas ou numéricas), então β_0 será a média de y quando todas as variáveis do modelo forem zero e a interpretação de β_1 se mantém, mas mantendo fixos os valores das demais covariáveis.

Regressão linear com covariáveis categóricas

- Por que não adicionar ao modelo uma variável indicadora para cada categoria da covariável?
- Seja x_1 a variável indicadora referente à categoria A e x_2 a variável indicadora referente a B;
- A matriz do modelo ficaria da seguinte forma:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \tag{4}$$

Regressão linear com covariáveis categóricas

- Ao inserir uma variável indicadora para cada categoria da covariável, a soma da segunda e da terceira coluna de $\boldsymbol{X}$ seria igual à primeira coluna (vetor de uns);

- Desta forma a matriz $\boldsymbol{X}$ não teria rank completo, de forma que haveria infinitas soluções para as equações de mínimos quadrados (os parâmetros do modelo não seriam identificáveis);

- Como a matriz do modelo não tem rank completo, a solução seria excluir uma de suas colunas (o intercepto ou uma das indicadoras das categorias de x).

Regressão linear com covariáveis categóricas

- Considere agora que a covariável categórica (fator) tenha k níveis $(A_1, A_2, ..., A_k)$.

- Se incluíssemos no modelo k variáveis indicadoras, então novamente a soma dessas variáveis resultaria num vetor de uns (dependência linear, matriz $\boldsymbol{X}$ de rank incompleto);

- Uma solução seria excluir da matriz do modelo uma das variáveis indicadoras (a categoria correspondente fica sendo a *categoria de referência*);
- Sejam:

$$x_1 = \begin{cases} 1 & \text{se categoria } A_2 \\ 0 & \text{se outra} \end{cases}, \; x_2 = \begin{cases} 1 & \text{se categoria } A_3 \\ 0 & \text{se outra} \end{cases}, \ldots, \; x_{k-1} = \begin{cases} 1 & \text{se categoria } A_k \\ 0 & \text{se outra} \end{cases}$$

Regressão linear com covariáveis categóricas

- Considerando que não haja outras covariáveis no modelo:

$$E(y|x) = \begin{cases} \beta_0 & \text{se categoria } A_1 \\ \beta_0 + \beta_1 & \text{se categoria } A_2 \\ \beta_0 + \beta_2 & \text{se categoria } A_3 \\ \vdots & \\ \beta_0 + \beta_{k-1} & \text{se categoria } A_k \end{cases} \tag{5}$$

- Neste caso β_0 é a média de y para a categoria de referência (A_1);
- β_j é a diferença na média de y da categoria A_{j+1} para a categoria A_1, para $j = 1, 2, ..., k-1$;
- $\beta_j - \beta_l$ é a diferença na média de y da categoria A_j para a categoria A_l ($j, l \in \{1, 2, ..., k-1\}$; $j \neq l$).

Regressão linear com covariáveis categóricas

- Se o modelo tiver ainda outras covariáveis, então:
  - β_0 é a média de y quando todas as variáveis do modelo forem zero;
  - Os demais $\beta's$ associados a essa covariável correspondem à diferença na média de y da categoria A_j para a categoria A_1, para $j \neq 1$, considerando fixos os valores das demais variáveis;
  - $\beta_j - \beta_l$ é a diferença na média de y em duas categorias da covariável (digamos A_j e A_l $(j, l \in \{1, 2, ..., k-1\}; j \neq l)$ considerando fixos os valores das demais variáveis.

Regressão linear com covariáveis categóricas

- A categoria A_1 a ser designada como referência fica a critério do analista;
- Independente da escolha, o modelo ajustado é o mesmo, apenas as interpretações dos parâmetros mudam;
- No R, se você não especificar a categoria de referência, por default ele utiliza o primeiro nível do fator;
- Há outras formas de incorporar variáveis categóricas a modelos de regressão. Cuidado ao usar outros softwares, consulte sempre a documentação!

Regressão linear com covariáveis categóricas e quantitativas

- Vamos considerar um modelo de regressão com uma variável quantitativa (x) e uma variável categórica com k níveis.

- Para simplificar a notação, vamos considerar D_2, D_3, ..., D_k as variáveis indicadoras para os níveis 2, 3,..., k da variável categórica.

- Um primeiro modelo a ser considerado é o de efeitos aditivos (sem interação), em que:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 d_{i2} + \beta_3 d_{i3} + ... + \beta_k d_{ik} + \epsilon_i, \quad (6)$$

em que $d_{ij} = 1$, se o indivíduo pertence à categoria j, e $d_{ij} = 0$, caso contrário.

Regressão linear com covariáveis categóricas e quantitativas

- Para modelos com variáveis categóricas e quantitativas é sempre útil escrever a equação do modelo para cada nível da variável categórica.

- O modelo de regressão com efeitos aditivos pode ser expresso, para cada nível da covariável categórica, por:

$$E(y|\boldsymbol{x}) = \begin{cases} \beta_0 + \beta_1 x_1 & \text{se categoria } A_1 \\ (\beta_0 + \beta_2) + \beta_1 x_1 & \text{se categoria } A_2 \\ (\beta_0 + \beta_3) + \beta_1 x_1 & \text{se categoria } A_3 \\ \vdots & \\ (\beta_0 + \beta_k) + \beta_1 x_1 & \text{se categoria } A_k \end{cases} \tag{7}$$

Regressão linear com covariáveis categóricas e quantitativas

- Observe que para o modelo aditivo, as retas de regressão para os k níveis da variável categórica apresentam somente interceptos diferentes. As inclinações são as mesmas.

- Para o modelo com efeitos aditivos, β_j corresponde à diferença na média de y da categoria j para a categoria 1 (referência) da covariável categórica, fixado o valor de x_1.

Regressão linear com covariáveis categóricas e quantitativas

- O modelo com efeitos multiplicativos (com interação), por sua vez, é definido por:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 d_{i2} + ... + \beta_k d_{ik} + \beta_{k+1} d_{i2} x_{i1} + ... + \beta_p d_{ik} x_{i1} + \epsilon_i,$$

podendo se expresso, para cada nível da covariável categórica, por:

$$E(y|\boldsymbol{x}) = \begin{cases} \beta_0 + \beta_1 x_1 & \text{se categoria } A_1 \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_{k+1}) x_1 & \text{se categoria } A_2 \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_{k+2}) x_1 & \text{se categoria } A_3 \\ \vdots & \\ (\beta_0 + \beta_k) + (\beta_1 + \beta_p) x_1 & \text{se categoria } A_k \end{cases} \quad (8)$$

Regressão linear com covariáveis categóricas e quantitativas

- Observe que para o modelo com efeitos multiplicativos as retas de regressão têm interceptos e inclinações diferentes para cada nível da variável categórica.

- Para o modelo com efeitos multiplicativos, β_2, β_3, ..., β_k expressam as diferenças de intercepto das retas de regressão para os níveis 2, 3,.., k da variável categórica em relação ao intercepto para o nível de referência;

- Já β_{k+1}, β_{k+2}, ..., β_p expressam as diferenças das inclinações das retas de regressão para os níveis 2, 3,.., k da variável categórica em relação à inclinação para o nível de referência.

Regressão linear com covariáveis categóricas e quantitativas

- Outra forma de explorar os efeitos das covariáveis em modelos contendo covariáveis categóricas e quantitativas é através de gráficos de efeitos;

- Para avaliar o efeito de uma covariável quantitativa x_1, plota-se o gráfico da estimativa de $E(y|\boldsymbol{x})$ ao longo de x_1, separado para cada nível da covariável categórica;

- Se houver mais covariáveis no modelo, elas podem ser fixadas em valores típicos (por exemplo em suas médias).

Regressão linear com covariáveis categóricas e quantitativas

- A seleção de modelos quando se tem efeito de interação deve obedecer ao princípio hierárquico;
- Basicamente, para um modelo ajustado com preditor `x1 + x2 + x1:x2`, em que `x1:x2` representa o termo de interação, não se deve remover os termos de menor ordem (`x1` e `x2`) na presença do termo de maior ordem `x1:x2`;
- Vamos considerar novamente que x_1 é uma covariável quantitativa e x_2 uma covariável categórica com k níveis;
- Pode-se proceder a seleção de modelos com base na seguinte sequência de modelos encaixados:

Regressão linear com covariáveis categóricas e quantitativas

- **Modelo nulo:** sem efeito de covariáveis (`y ~ 1`);
- **Modelo de retas coincidentes:** sem efeito da covariável categórica (`y ~ x1`);
- **Modelo de retas paralelas:** modelo de efeitos aditivos, sem interação (`y ~ x1 + x2`);
- **Modelo de retas concorrentes:** modelo de efeitos multiplicativos, com interação (`y ~ x1 * x2`).

Regressão linear com covariáveis categóricas e quantitativas

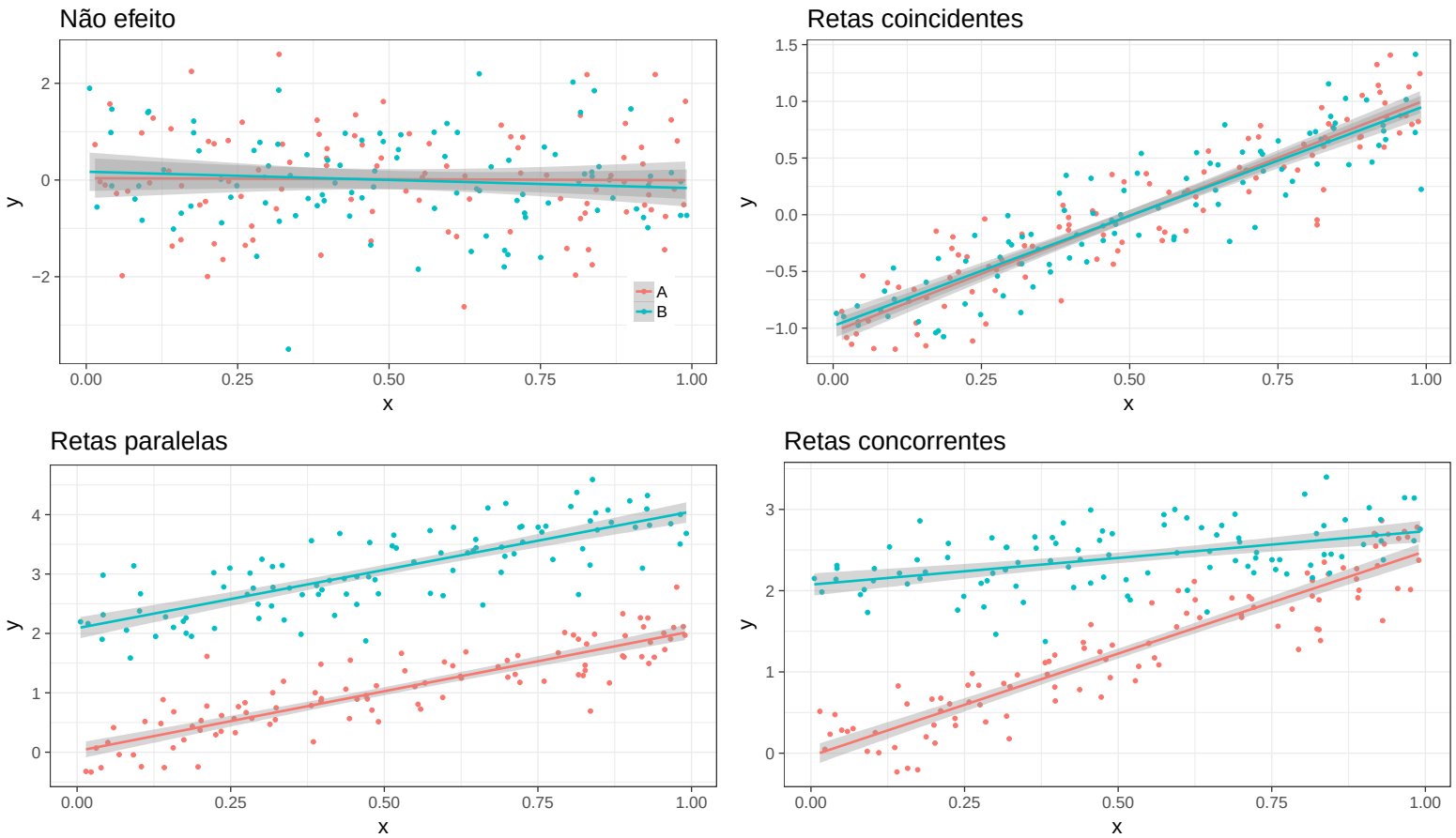

Figura 1: Ilustração para os cenários correspondentes aos quatro modelos incluindo uma covariável quantitativa e outra categórica

Regressão linear com covariáveis categóricas e quantitativas

- Como os quatro modelos sugeridos são encaixados, podemos compará-los, sequencialmente, usando o teste F baseado na variação da soma de quadrados de resíduos;
- Caso se disponha de mais covariáveis para análise, pode-se aplicar algum algoritmo de seleção de covariáveis. Além disso, interações entre outras covariáveis podem ser consideradas;
- Outra possibilidade é avaliar modelos em que as retas de regressão são paralelas ou coincidentes apenas para algum subconjunto dos níveis da variável categórica;
- Como exemplo poderíamos ter retas de mesma inclinação descrevendo a relação entre índice de massa corporal e consumo calórico para adultos e idosos, mas de menor inclinação para a população de crianças.