Pressupostos e análise de diagnósticos da Regressão Logística

PUC Minas Virtual

Pressupostos da Regressão Logística

- A variável reposta precisa ser qualitativa, dicotômica
- As preditoras podem ser quantitativas ou categóricas
- Relação linear entre o vetor das variáveis explicativas X e a variável independente Y;
- Ausência de correlação entre os resíduos
- Assume que as observações são independentes
- Ausência de multicolinearidade

Multicolinearidade

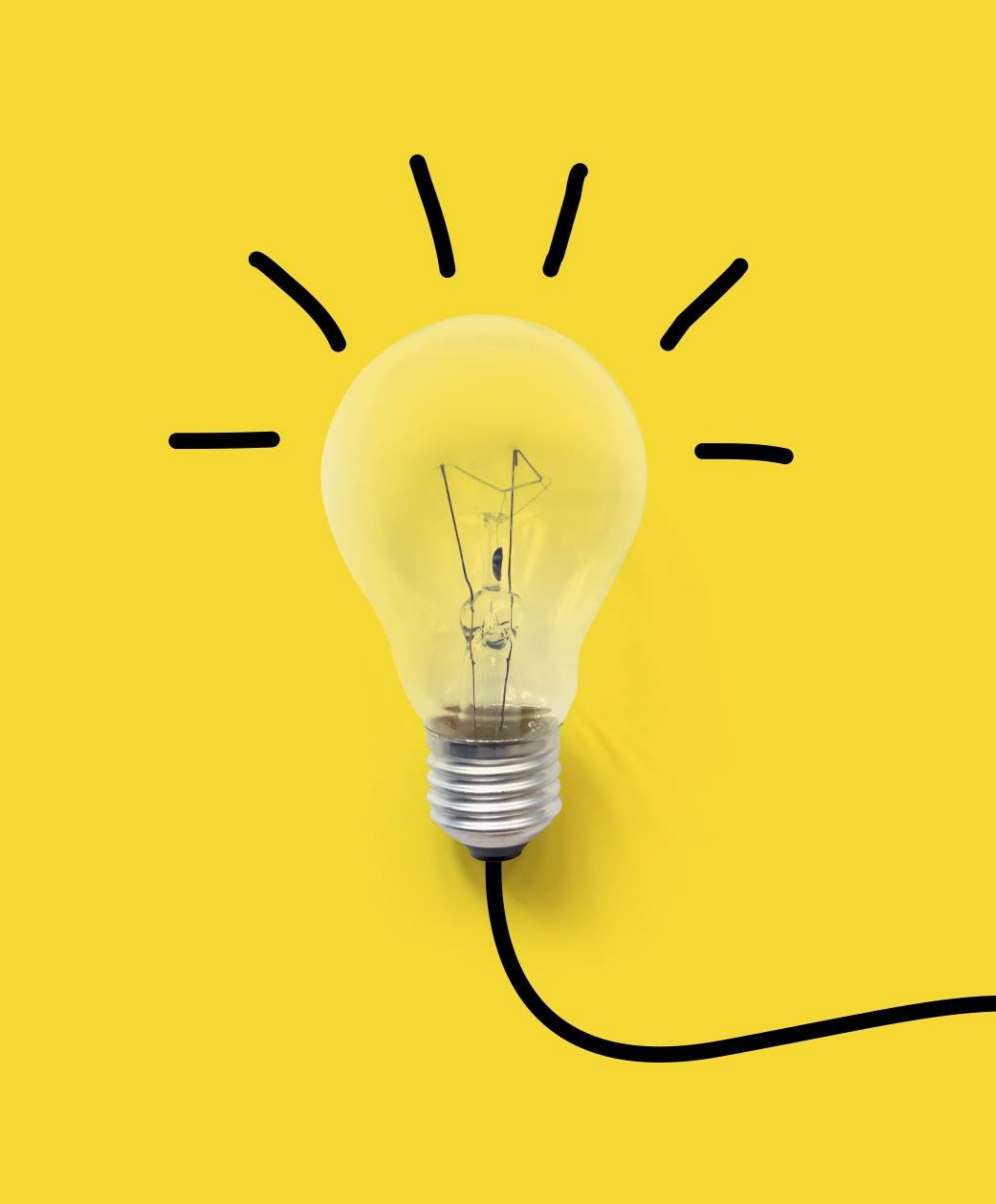
- **Preditoras correlacionadas com outras preditoras**, resulta quando você tem fatores que são, de certa forma, um pouco **redundantes**.
- Ou seja, quando **duas ou mais variáveis independentes em um modelo de regressão encontram-se altamente correlacionadas**
- Examinar a matriz de correlação das variáveis independentes.
  - ➢0,70 Altamente correlacionadas
- O valor do fator de inflação da variância (VIF), **que mede quanto a variância do coeficiente estimado para uma variável é inflada devido à multicolinearidade** com as outras variáveis independentes.
- VIFs maiores que 10 indicam alta multicolinearidade.

PARÂMETROS DOS MODELOS

Verificar a significância das variáveis do modelo

Teste de hipótese para determinar se a variável preditora do modelo é significativamente relacionada com variável resposta do modelo

- Teste de Wald
- Teste de Razão de verossimilhança



Teste de Razão de Verossimilhança (*Devaiance*)

Compara valores observados x preditos , com e sem determinadas variáveis.

Baseada na log verossimilhança

$$\mathbf{D = -\,2\,ln}\left(\frac{\boldsymbol{Verossimilhança\ modelo\ ajustado}}{\boldsymbol{Verossimilhança\ modelo\ saturado}}\right)$$

Modelo Saturado -> Modelo que se ajusta perfeitamente os dados

Teste de Razão de Verossimilhança

G = D(modelo sem a variável) – D(modelo com a variável)

Modelo saturado é o mesmo para os 2.

$$\mathbf{G = -2\ ln} \left(\frac{Verossimilhança\ sem\ a\ variavel}{Verossimilhança\ com\ a\ variavel} \right)$$

Utilizando a *deviance* para comparação de modelos que não sejam saturados.

Teste de Razão de Verossimilhança

H_0 : A hipótese nula afirma que o modelo nulo (mais simples) é verdadeiro, ou seja, a inclusão das variáveis adicionais no modelo completo não melhora significativamente o ajuste do modelo.

H_1 : A hipótese alternativa rejeita a hipótese nula e afirma que o modelo completo é significativamente melhor do que o modelo nulo.

H_0 : Verossimilhança do modelo Nulo = Verossimilhança do Modelo Completo

H_1 : Verossimilhança do modelo Completo > Verossimilhança do Modelo Nulo

Teste Wald

Obtido por comparação entre a estimativa de máxima verossimilhança do parâmetro $\widehat{\beta}_j$ e a estimativa de seu erro padrão.

$$H_0 : \widehat{\beta}_1 = 0$$
$$H_1 : \widehat{\beta}_1 \neq 0$$

$$\hat{Y}_i = \widehat{\beta}_o + \widehat{\beta}_1 X_i$$

A estatística do Teste Wald para a regressão logística é dada por:

$$W_j = \frac{\widehat{\beta}_1}{SE\left(\widehat{\beta}_1\right)}$$

Se não rejeitarmos H_0 , temos que a variável X não explica a variável resposta.

Medidas de qualidade do ajuste do modelo

Para analisar o desempenho geral do modelo ajustado podemos utilizar vários tipos de Testes de qualidade de ajuste.

Testes que necessitam dados replicados (múltiplas observações com os mesmos valores para todos os preditores):

- χ^2 de Pearson
- Deviance

No aprendizado de máquina é raro utilizar essas técnicas pois geralmente se avalia o desempenho do modelo pelo conjunto de teste

Deviance

Pequenos valores de *Deviance* (ou elevado valor p) implicam que o modelo fornece um ajuste satisfatório aos dados , enquanto grandes valores de *deviance* implicam que o modelo atual não é adequado.

- Podemos dividir o *deviance* pelo graus de liberdade.

  - Se $\frac{D}{n-k} \gg 1 \;\rightarrow O\ Modelo\ não\ é\ adequado\ aos\ dados$
  - Onde $n\ -k$ é o gruas de liberdade. k é o número de parâmetro do modelo
  - D é dado por:

$$D = 2\ \sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{n_i\hat{p}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i-y_i}{n_i(1-\hat{p}_i)}\right)\right]$$

χ^2 de Pearson

Compara as probabilidade de sucesso e fracasso observadas e esperadas em cada grupo de observações

- Nº esperado de sucesso : $n_i \hat{p}_i$
- Nº esperado de fracasso: $n_i(1 - \hat{p}_i)$

- A estatística de de Pearson é dada por:

$$\chi^2_{n-k} = \sum_{i=1}^{n} \frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i (1 - \hat{p}_i)}$$

Valores pequenos da estatística de teste ou um grande valor de p , implica que o modelo fornece um ajuste satisfatório aos dados

Análise dos resíduos

- Resíduos de Pearson
- Resíduos de Deviance
- Resíduos de Pseudo –Valor
- Gráfico de Resíduos

Resíduos de Pearson

1.Os resíduos de Pearson (r_i) são calculados para cada observação (*i*) da seguinte maneira:

$$r_i = \frac{y_i - \widehat{p_i}}{\sqrt{\widehat{p_i}(1 - \widehat{p_i})}}$$

Onde:

•*yi* é o valor observado da variável dependente (0 ou 1).

•$\widehat{p_i}$ é a probabilidade prevista da variável dependente ser igual a 1(sucesso), estimada pelo modelo.

•1- $\widehat{p_i}$: é a probabilidade prevista da variável dependente ser igual a 0.

Resíduos de Deviance

Os resíduos de Deviance (D_i) são calculados para cada observação (*i*) da seguinte maneira:

$$D_i = sign\,(y_i - \widehat{p_i}) \cdot \sqrt{2\,(y_i \log\left(\frac{y_i}{\widehat{p_i}}\right) + (1 - y_i)\log\left(\frac{1 - y_i}{1 - \widehat{p_i}}\right)}$$

Os resíduos de deviance representam a diferença entre a log-verossimilhança dos modelos completo e nulo.

Um valor absoluto grande indica uma má adequação do modelo aos dados.

Resíduos de Pseudo-Valor

Os resíduos de pseudo-valor (v_i) são calculados para cada observação (*i*) da seguinte maneira:

$$v_i = y_i - \widehat{p_i}$$

Os resíduos de pseudo-valor são úteis para identificar pontos de influência nos dados.

PUC Minas
Virtual