Análise de Regressão Linear

Bruno Gondim Toledo | 150167636

17 de julho de 2023

Atividade 2

1 Questão 1

As saídas do programa vêm em anexo ao documento

2 Questão 2

Analisando as saídas do programa, podemos primeiramente nos debruçar sobre uma análise dos coeficientes R-quadrado (*R-squared*) e R-quadrado ajustado (*Adjusted R-squared*). Esta escala varia de 0 a 1, sendo 1 o melhor modelo possível (explica o máximo da variabilidade dos dados) e 0 o pior modelo possível. Levando ainda em conta os princípios de parcimônia, tentaremos encontrar as melhores opções de modelo com o menor número de variáveis possíveis. Notamos que o modelo contendo as variáveis X_1, X_2, X_3, X_5, X_6 *e* x_7a apresenta o maior coeficiente R-quadrado ajustado, seguido pelos modelos contendo as variáveis X_1, X_2, X_3, X_6 *e* x_7a, e $X_1X_2X_3X_4X_5X_6x_7a$ respectivamente. A variável x_7a, no caso, é uma transformação 'dummy' da variável x7, que continha 3 fatores. Portanto, esta foi transformada em duas variáveis dummy: x_7a e x_7b

Analisando estes três modelos, notamos que é comum em todos as variáveisX_1, X_2, X_3, X_6 *e* x_7a, com as demais suprimidas em pelo menos um dos modelos. Entretanto, se compararmos o R-quadrado ajustado dos modelos de três variáveis, com o modelo de cinco variáveis, notamos que o ganho no coeficiente foi desprezível. Portanto, este modelo aparenta já ser suficientemente robusto. Pelo princípio da navalha de Occam, podemos inferir que este modelo aparenta ser suficiente neste caso.

Os métodos de validação automática, entretanto, chegaram a resultados diferentes, no caso, o modelo contendo X_1, X_2, x_3, x_7a para o método *Backward*; o modelo contendo X_1, X_2, X_3, X_5, x_7a *e* x_7b pelo método *Forward*. Finalmente, o método mais acurado *Stepwise* retornou o modelo contendo x_1, x_2, x_3, x_6 *e* x_7a. Analisando este último modelo, notamos que é o mesmo visto anteriormente. Quando avaliamos o p-valor do teste F de cada coeficiente deste modelo, notamos que todos são convincentemente pertencentes ao modelo, à excessão da variável x_6, que entra no modelo somente sob um nível de significância $\alpha = .15$, que costuma estar fora do que "gostamos"de aceitar. Entretando, analisando exploratoriamente, notamos que esta variável é o gênero do candidato. Ou seja, é provavelmente a variável mais simples e barata de se coletar. Portanto, se ela consegue trazer algum ganho pro modelo, sem trazer confusão e sendo extremamente simples e barata de coletar, acredito que seja melhor mantê-la no modelo do que remove-la.

3 Questão 3

Nesta etapa, iremos selecionar o modelo 163 (o que contém as variáveis X_1, X_2, X_3, X_6 *e* x_7a) para realizar os procedimentos de validação. Analisando as saídas do programa para o banco de dados de validação, podemos observar medidas que nos ajudam a tomar essas decisões.

4 Questão 4

4.1 a)

Verificar saídas do programa em anexo

4.2 b)

Analisando as saídas do programa, nada salta aos olhos indicando grandes falhas no modelo. Portanto, aparenta ser válido o modelo.

Verificando $E(ln(y))$; comparando com o $ln(y)$ real, calculando as diferenças do esperado para o observado e verificando o ganho com a variável x_6, confirmamos todas as suspeitas e hipóteses levantadas anteriormente. O modelo é bom e explica bem a variável; de fato a variável x_6 é pouco influente mas será mantida pelos motivos já explicados. Importante destacar que logo ao início, a variável resposta fora transformada em logaritmo. Portanto, a resposta final "comercial"deve ser aplicada a exponencial, para apresentar os resultados ao "chefe"!

5 Questão 5

Com o modelo validado, produzindo predições próximas tanto para o modelo de treino quanto o de teste, com um MSPR aceitável, temos fortes indícios para acreditar que este é um bom modelo.

6 Saídas do programa