



UNIVERSIDADE
DE ÉVORA

Resposta ao Quarto Trabalho de Casa de Estatística Aplicada

Trabalho realizado por

Jorge Couto
Miguel Grilo

58656
58387

Colégio Luís António Verney

Como trabalho de casa a ser feito antes do miniteste que será realizado na aula do dia 11 de maio, foi-nos pedido para realizar um estudo quanto à digitalização das empresas através da adoção de IA, com base no nível de educação dos trabalhadores. Os dados que nos foram oferecidos indicam a quantidade de empresas com diferentes porcentagens de funcionários que têm ensino superior e, para as empresas de cada nível, quantas delas começaram a adotar medidas de IA.

Pelos dados oferecidos, percebemos que estamos diante de um problema de Modelo Resposta. Portanto, para começar a estudar os dados, criamos antes de tudo um *data frame* com as três variáveis mencionadas pelo professor (*perc_edu*, *n_empresas* e *n_adotaram*). Todavia, não sendo suficiente, criamos também a variável *n_nao_adotaram*, uma variável que representa o ‘evento negativo’, ou seja, o número de empresas que não adotaram IA. Criamos, também, a variável *prop*, que representa a porcentagem de empresas que adotaram IA face o número de empresas para cada nível de *perc_edu*.

De seguida, criamos os 3 modelos *logit*, *probit* e *cloglog*. Para avaliar os modelos e descobrir qual seria o mais adequado, utilizamos três critérios: O *AIC*, o *BIC* e o R quadrado, sendo um melhor modelo dado por um menor valor nos dois primeiros, e por um valor maior no terceiro.

Tanto no *AIC* quanto no *BIC*, o modelo *probit* foi aquele com melhores resultados, tendo o menor valor dos três nos dois casos, de aproximadamente 43.35 no *AIC* e 45.96 no *BIC*. No R quadrado, também, demonstrou o melhor valor, dando uma porcentagem de explicabilidade de 97%. Portanto, não resta dúvidas que o modelo que melhor explica a variabilidade dos resultados é o modelo *probit*.

Depois do ajuste e avaliação dos modelos, foi-nos pedido pelo professor para determinar a porcentagem mínima de trabalhadores com ensino superior que garantisse uma probabilidade de adoção de IA superior ou igual a 90%. Para isso, criamos as variáveis *ld_logit*, *ld_probit* e *ld_cloglog*, que explicam, usando o comando *dose.p*, o valor em porcentagem de trabalhadores com ensino superior necessário para que a probabilidade de adoção de IA seja superior ou igual a 90%, como pedido.

Naturalmente, o valor varia de modelo a modelo. Seguindo a análise pelo modelo *probit*, podemos concluir que pelo menos 72.4% dos trabalhadores deveriam ter ensino superior para que a taxa de adoção de IA fosse pelo menos 90%. Todavia, pensando nisto de modo mais versátil, e não apenas como um estatístico, poderíamos concluir que seriam precisos entre 70% a 75% dos trabalhadores com ensino superior para que a taxa de adoção de IA fosse, pelo menos, 90%.

Finalmente, a última questão colocada pede-nos para obter a estimação pontual e intervalar da probabilidade de uma empresa adotar IA quando 50% dos trabalhadores têm ensino superior. Podemos obter essa estimação através da prefição, criando uma variável *grid* que, para melhor facilidade na interpretação, criamos como um pequeno intervalo para a variável do *data frame* *perc_edu* de 49% a 51%. Usando a função *get_preds*, uma função oferecida em aula pelo professor que permite calcular a estimação pontual e intervalar para cada nível de *perc_edu* presente na variável *grid*, vemos:

- Para o modelo *logit*, a probabilidade de uma empresa adotar IA quando 50% dos trabalhadores têm ensino superior é de 51.7%, sendo o intervalo da probabilidade dado por [45.3%, 57.9%];
- Para o modelo *probit*, a probabilidade da empresa adotar IA quando 50% dos trabalhadores têm ensino superior é de 51%, estando a verdadeira probabilidade entre 45.4% e 56.7%;
- Para o modelo *cloglog*, existe uma probabilidade de 44.3% de uma empresa adotar IA quando 50% dos trabalhadores têm ensino superior, e o intervalo é dado por [39%, 50%].

Contudo, não podemos dizer que acabamos por aqui. Devemos confirmar se o modelo a se usar, tanto na alínea b quanto na alínea c, de facto é o modelo *probit*. Apesar do modelo *probit* ser, em termos globais, o mais adequado, pode ser que em intervalos específicos um dos outros três modelos seja mais eficaz. Podemos verificar isso recriando a variável *grid* para que, agora, assuma valores cobrindo todo o domínio da variável

perc_edu. Com a nova variável grid, recriamos as predições, mas agora ao invés de estudarmos as predições criamos um gráfico com as *bandas de confiança*. Podemos, pelo gráfico, verificar qual modelo é o mais adequado em certos momentos com base na aproximação do modelo às amostras de facto mostradas.

Começando pela alínea b, devemos verificar se, para uma taxa de adoção de IA de 90% (eixo das ordenadas), o modelo *probit* realmente é o mais adequado. A nosso ver, para uma taxa de adoção de IA de 90% não parece haver preferência entre qualquer um dos três modelos (apesar de, logo antes, haver preferência pelo modelo *cloglog*). Portanto, podemos manter a análise que tivemos quanto a essa questão.

Para a alínea c, deveríamos investigar o eixo oposto, verificando se, para uma porcentagem de 50% de trabalhadores com ensino superior (eixo das abcissas), o modelo *probit* também é o mais adequado. Quanto a este, não nos parece haver dúvidas que os modelos *probit* e *logit* são os mais adequados, não havendo preferência em nenhum dos dois, visto que seguem mais fielmente as amostras reais observadas mais próximas. Portanto, também podemos considerar essa análise como correta.

Com base nos resultados obtidos, e pela análise gráfica, podemos ver que empresas com mais empregados com o ensino superior parecem vir a ter maiores chances de adotar IA. Caso uma empresa queira vir a adotar IA para melhor ajudar no seu desempenho lucrativo e competitividade de mercado, poderia dar incentivos aos seus trabalhadores para que sigam o ensino superior, seja por incentivo monetário ou, para estudantes do ensino superior que se possam vir a interessar na empresa, por estágios.