Nome: Maria Clara Kaseker de Lara Pinheiro Goes, NºUSP: 10831396
EAD0830 - IA e ML Aplicados a Finanças
Atividade Computacional 1 - Regressão

i) Regressão:

· *Estatísticas das variáveis e suas correlações*: A amostra traz dados de idade, sexo, índice de massa corporal, número de filhos do segurado, gastos com seguro de vida e região de moradia de 1338 indivíduos. Destes, 662 são do sexo masculino e 676 do sexo feminino. A média de gastos com seguro para os indivíduos do sexo masculino é $12.569,58 e para indivíduos do grupo feminino é $13.956,75.

· *Estimando o modelo de regressão*: Para análise do problema proposto, foi utilizada a variável de gastos com seguro como variável dependente e as demais como variáveis explicativas. Desta forma o modelo que explica os gastos com seguro de vida por indivíduo pode ser calculado através da seguinte equação de regressão por OLS (*Ordinary Least Squares*):

$$\begin{aligned}\text{Expenses} = {} & -10.127{,}87 + 256{,}84\ \text{age} + 131{,}35\ \text{sex} + 339{,}29\ \text{BMI} + 23.847{,}48\ \text{smoker} + 475{,}69\ \text{children} \\ & -1.945{,}04\ \text{northeast} - 2.297{,}83\ \text{northwest} - 2.980{,}64\ \text{southeast} - 2.904{,}35\ \text{southwest}\end{aligned}$$

· *Análise de diagnóstico do modelo*: O coeficiente de determinação R^2 é 0,751. Isso significa que aproximadamente 75.1% da variação dos gastos com seguro de vida é explicada pelas variáveis independentes incluídas no modelo. O valor da estatística F é 500,9, com um p-valor de 0,00. Isso indica que o modelo global é estatisticamente significativo, ou seja, pelo menos uma das variáveis independentes tem um efeito significativo nos gastos com seguro. O coeficiente para a variável '*age*' é 256,84, o que significa que, em média, os gastos aumentam $256,84 para cada ano a mais de idade. Como o sexo foi determinado como 0 para feminino e 1 para masculino, o coeficiente 131,35 indica a diferença média nos gastos entre os sexos masculino e feminino. Apesar disso, o p-valor associado a essa variável é alto (0,693), indicando que não é estatisticamente significativa neste modelo. Para cada unidade de aumento no Índice de Massa Corporal (BMI), os gastos com seguro aumentam em média $339,29. Indivíduos fumantes gastam, em média, $23.850 a mais com seguro do que indivíduos não fumantes. Este coeficiente tem um p-valor muito baixo (0,000), indicando forte significância estatística. Já para cada filho adicional, os gastos com seguro aumentam em média $475,69. Os coeficientes das regiões *Northeast, Northwest, Southeast, Southwest* representam as diferenças médias nos gastos com seguro de vida em cada região. O coeficiente para *Northeast* é -1945,04, indicando que os gastos médios na região são $1945,04 menores. O coeficiente para *Northwest* é -2297,84. Já o coeficiente para *Southeast* é -2980,64, significando que os gastos médios na região são $2980,64 menores. Por fim, o coeficiente para *Southwest* é -2904,35, indicando que os gastos médios na região são $2904,35 menores. Todos esses coeficientes têm valores de significância estatística muito baixos (p-valor = 0.000), sugerindo que a região de residência dos segurados tem um impacto estatisticamente significativo nos gastos com seguro de vida, mesmo após controlar outras variáveis do modelo. É importante notar que os resíduos não seguem uma distribuição normal, como indicado pelo teste de normalidade Omnibus (Prob (Omnibus): 0,000) e o teste Jarque-Bera (Prob(JB): 6,14e-157), que sugerem que o modelo pode ser aprimorado considerando outras transformações ou inclusão de variáveis adicionais.

ii) *Regressão com regularização*: Fazendo a regressão do modelo usando *OLS* no Python, duas notas merecem atenção. São elas:
[1] *Standard Errors assume that the covariance matrix of the errors is correctly specified*.
[2] *The smallest eigenvalue is 6.05e-29. This might indicate that there are strong multicollinearity problems or that the design matrix is singular*.
Sobre a primeira, podemos dizer que os erros padrão assumem que a matriz de covariância dos erros foi especificada corretamente, seguindo uma distribuição normal com covariância constante. Já que essa suposição foi atendida, os erros padrão fornecem uma medida confiável da precisão dos coeficientes estimados. Sobre a segunda, o menor autovalor da matriz do modelo é extremamente próximo de zero, podendo indicar problemas de multicolinearidade forte ou que a *design matrix* é singular. Multicolinearidade nos diz que duas ou mais variáveis independentes no modelo estão altamente correlacionadas entre si. Isso pode dificultar a interpretação dos coeficientes individuais e aumentar a instabilidade das estimativas. Já uma matriz singular significa que uma ou mais variáveis independentes podem ser combinações lineares das outras variáveis independentes, levando a problemas na inversão da matriz e afetar a estabilidade e a confiabilidade das estimativas dos coeficientes. Nesse caso, a alternativa que a atividade propõe é o uso de regularização, uma abordagem diferente de aumentar o tamanho da amostra ou remover variáveis redundantes do modelo.

Regularização é um método que introduz penalidades nos coeficientes do modelo, podendo reduzir a sensibilidade a pequenas mudanças nos dados e ajudar a estabilizar as estimativas dos coeficientes. Decidiu-se pelo uso da *Ridge Regression*, que adiciona uma penalidade à função de perda *(loss function*) durante a estimativa dos coeficientes, onde essa penalidade é proporcional à soma dos quadrados dos coeficientes, multiplicada por um hiper parâmetro chamado de "parâmetro de regularização" (alpha). Ao suavizar os coeficientes, a regularização por Ridge ajuda a reduzir a variância do modelo, o que pode melhorar sua capacidade de generalização para novos dados. Embora a *Ridge* seja eficaz na redução da multicolinearidade e na estabilização dos coeficientes, ela não realiza seleção de variáveis. Ou seja, todas as variáveis explicativas permaneceram no modelo, mesmo que algumas delas tenham pouco impacto nos resultados, como sexo. Em casos em que a seleção de variáveis é importante, outras técnicas como a regularização *Lasso* (L1) ou *Elastic Net* podem ser mais adequadas. Um valor muito baixo de alpha pode não ter efeito significativo na redução da variância, enquanto um valor muito alto pode levar à subestimação dos coeficientes. O próprio código faz uma busca através de *cross-validation* para encontrar o valor ótimo de alpha que melhor equilibra o viés e a variância do modelo, que para este caso é 1,0. A equação de regressão que melhor representa o gasto com seguros passa a ser:

$$\text{Expenses} = 256{,}76\,\text{age} + 10{,}64\,\text{sex} + 337{,}13\,\text{BMI} + 23.514{,}21\,\text{smoker} + 426{,}17\,\text{children} + 458{,}70\,\text{northeast} + 85{,}79\,\text{northwest} - 193{,}27\,\text{southeast} - 351{,}27\,\text{southwest}$$

Para esse modelo, o R^2 é de aproximadamente 0,783, o que significa que o modelo explica cerca de 78.3%, 3,2% mais preciso que o modelo sem regularização. Tendo o *F-Test* resultado em 3340,33, com o p-valor associado de aproximadamente 0,0 o modelo com regularização (*Ridge Regression*) se mostra estatisticamente significativo, pois o valor de F é bastante alto e o valor-p é muito baixo. Isso indica que a hipótese nula de que todos os coeficientes das variáveis explicativas são iguais a zero é rejeitada, e pode-se concluir que o modelo como um todo é útil para explicar os gastos com seguros com base nas variáveis incluídas.