Multicolinariedade e Autocorrelação

Introdução

- Em regressão múltipla, se não existe relação linear entre as variáveis preditoras, as variáveis são ortogonais.
- Na maioria das aplicações os regressores não são ortogonais.
- Explicar preço de uma casa com regressão que tenha como variáveis explicativas a área da casa e o número de cômodos

Fontes de Muticolinariedade

- É conveniente que os preditores e variável resposta estejam centrados e padronizados com comprimento 1.
- $\mathbf{X}^T\mathbf{X}$ é uma matriz de correlação entre os regressores e $\mathbf{X}^t\mathbf{y}$ é o vetor de correlações entre os preditores e a variável resposta.
- Os regressores são linearmente dependentes se existe um conjunto de constantes $t_1,\ldots t_p$ não zero tal que

$$\sum_{j=1}^{p} t_j \boldsymbol{X}_j = \boldsymbol{0}$$

Fontes de Muticolinariedade

- Amostra de subespaço de uma região que existe dependência linear entre os regressores. Exemplo: estimando o tempo de *delivery* considerando número de casos e distância. Observações com um número pequeno de casos tem uma distância curta. Poderiam ser coletadas informações com número pequeno de casos e longa distância.
- Restrição na população. Exemplo, investigando o efeito de sálário e o tamanho da casa para predizer o consumo de ernergia. Famílias com salários altos geralmente têm grandes casas.

Fontes de Muticolinariedade

- Escolha de modelo. Adicionando termos polinomiais no modelo causa mal condicionamento na matriz $\mathbf{X}^T\mathbf{X}$. Se o range de x é pequeno, adicionando o termo x^2 pode resultar em multicolinariedade significativa. Retendo todos os regressores pode contribuir para multicolinariedade. Nesse caso, um subconjunto de regressores é preferível.
- Ter mais regressores que observações.

Usual abordagem para tratar multicolinariedade

- 1. Redefina o modelo em termos de um conjunto menor de regressores.
- 2. Realize estudo preliminar usando somente subconjuntos do conjunto de regressores originais.
- 3. Use regressão tipo componentes principais para decidir quais regressores devem ser removidos.

Efeitos de multicolinariedade

- 1. Variâncias e covariâncias grandes para os estimadores de mínimos quadrados da regressão. Isso implica que amostras diferentes para o mesmo range de x pode levar coeficientes muito diferentes.
- 2. Estimativas de mínimos quadrados grandes em valor absoluto. Enquanto o método de mínimos quadrados geralmente produz estimativas pobres quando existe multicolinariedade, isso não implica que o modelo ajustado é um preditor pobre. Isso ocorre quando as predições são feitas em regiões em que a multicolinariedade é moderada.

Se um modelo de regressão extrapola bem, boas estimativas dos coeficientes são requeridas.

Diagnóstico de multicolinariedade

- 1. Examinar a matriz de correlação é util em detectando dependência linear entre pares de regressores. Não é suficiente avaliar a correlação.
- 2. Fator de inflação de variância VIFs (Variance Inflation Factor) dos estimadores onde R_j^2 é o coeficiente de determinação obtido quando x_j ajustado considerando os p-1 regressores restantes. VIFs que excede o valor 5 ou 10, é uma indicação que coeficientes são pobremente estimados por causa de multicolinariedade.

$$VIF_j = (1 - R_j^{\,2})^{-1}$$

Diagnóstico de multicolinariedade: Análise de autovalores de $\mathbf{X}^T\mathbf{X}$

- Um ou mais autovalores (λ_j j=1,…,p) pequenos implica que existe dependências quase linear entre as colunas de $\mathbf{X}^T\mathbf{X}$. Nessa direção, é interessante observar o número condicional

$$\kappa = \lambda_{\max} / \lambda_{\min}$$

- Se $\kappa < 100$, existe fraca dependência linear, se $100 \leq \kappa \leq 1000$, existe moderada dependência e se $\kappa > 10000$, existe severa multicolinariedade.

Diagnóstico de multicolinariedade: Análise de autovalores de $\mathbf{X}^T\mathbf{X}$

- O índice condicional para a variável j (j=1,…,p) é

$$\kappa_j = \lambda_{\max} / \lambda_j$$

- Se um desses índices exceder 1000, existe pelo menos uma forte dependência linear.

Diagnóstico de multicolinariedade: Análise de autovalores de $\mathbf{X}^T\mathbf{X}$

- VIFs são afetados quando os regressores são centrados. Alguns autores recomendam não obter regressores centrados.
- Centrando os dados torna o regressor intercepto ser ortogonal aos outros regressores. Isso pode ser visto como uma operação removendo um mal condicionamento devido a uma constante.
- Se o intercepto no modelo tem uma interpretação física (o que não é o caso em muitas aplicações práticas), centrar os dados não é uma boa abordagem.

Tratamento para multicolinariedade: Adicionar dados

- Não é sempre possível pois pode existir restrição econômica.
- Não é viável quando a multicolinariedade é devido a restrição no modelo ou na população. Exemplo: considere os regressores salário e tamanho da casa. Adicionar dados é de pouco valor por pois a relação entre salário e tamanho da casa é uma característica estrutural da população. Todos os dados na população exibirão esse comportamento.

Tratamento para multicolinariedade: Reespecificação do modelo

- Redefinir variáveis. Por exemplo, se x_1, x_2 e x_3 são linearmente dependente, pode existir alguma função tal que $x=(x_1+x_2)/2$ ou $x=(x_1 \times x_2 \times x_3)$ que reduza o mal condicionamento .
- Eliminar variáveis do modelo. Pode não ser muito eficaz se os regressores eliminados tem um alto poder de influência na variabilidade da variável resposta. Não existe garantia que o modelo final exibirá um grau de multicolinariedade menor que o modelo original. A análise é feita usando métodos stepwise e usando as estatísticas R^2 e VIFs.
- É importante investigar o poder desempenho dos modelos com dados não usados no modelo (taxa de erro de predição).

Tratamento para multicolinariedade: Regressão *Ridge*

- Quando o método de mínimos quadrados é aplicado para dados não ortogonais, implica que os valores absolutos das estimativas são grandes e esses são instáveis (variâncias inflacionadas).
- Teorema de Gauss-Markov afirma que estimadores de mínimos quadrados tem variância mínima na classe de estimadores não enviesados mas não garante que essa variância não será pequena.
- Variância pequena para um estimador viesado implica que esse é mais estável que o estimador não enviesado.

Tratamento para multicolinariedade: Regressão *Ridge*

- O método regressão *ridge* obtém estimadores viesados

$$\hat{\beta} = (X^T X + kI)^{-1} X^T y$$

- k é uma constante selecionada pelo analista. Quando k=0, o estimador *ridge* é o estimador de mínimos quadrados.
- Esse estimador é uma transformação linear do estimador de mínimos quadrados.
- O viês cresce quando k cresce contudo a variância descresce.

Tratamento para multicolinariedade: Regressão *Ridge*

- Erro quadrático médio

$$MSE = Var(\hat{\beta}) + vies(\hat{\beta})$$
$$= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + k^2 \beta^T \left((X^T X + kI)^{-1} \right)^2 \beta$$

- A regressão *ridge* pode apresentar melhores resultados de predição com observações futuras que a regressão mínimos quadrados.
- K pode ser determinado por um gráfico dos elementos de β estimado e os valores de k.
- Alguns autores sugerem que sejam considerados 25 valores de espaçados logaritmamente sobre o intervalo [0,1]. K é escolhido de tal forma que os coeficientes não mudam drasticamente e não produza viês grande.

Tratamento para multicolinariedade: Regressão *Ridge*

- Regressão ridge em geral não requer uma nova teoria de distribuição. É assumido distribuição normal para valores pequenos de k.
- Uma escolha analítica para k é: $k = \frac{p\hat{\sigma}^2}{\hat{\beta}^T \hat{\beta}}$

Estimadores de mínimos quadrados

Tratamento para multicolinariedade: Regressão componentes principais

- Forma do modelo

$$y = Z\alpha + \varepsilon \qquad Z = XT \qquad \alpha = T^T \beta$$

- λ=diag(λ_1,…, λ_p) é uma matriz diagonal p$\times$ p dos autovalores de X^TX e T é uma matriz p$\times$ p cujas colunas são autovetores associados com os autovalores λ_1,…, λ_p.
- Essa abordagem combate a multicolinariedade usando menos que um conjunto completo de componentes principais no modelo.
- Asssuma que os regressores são ordenados em ordem decrescentes de autovalores $\lambda_1 > \lambda_2 > \ldots > \lambda_p$. Suponha que os últimos s desses autovalores são aproximadamente zero. As componentes principais correspodentes a esses autovalores são removidas e a regressão é aplicada considerando os p-s regressores componentes principais restantes.
- ***Z*** =***XT*** transforma os regressore originais padronizados em regressores ortogonais.

Tratamento para multicolinariedade: Regressão componentes principais

- A regressão componentes principais reduz o efeito de multicolinariedade usando um subconjunto de componentes principais.
- Diferentes números de componentes principais no modelo produzem diferentes estimativas dos coeficientes.
- Um subconjunto de componentes podem produzir estimativas não muito diferentes das estimativas viesadas (regressão *ridge*)

Tratamento para multicolinariedade

- Estimadores viesados são úteis em tratando situações de muticolinariedade.
- Estimadores viesados é comparavelmente melhor do que o método de eliminação de variável.
- É melhor usar alguma da informação com todos os regressores (regressão *ridge*) do que usar toda a informação em alguns regressores e nenhuma informação em outros regressores.
- Montgomery recomenda usar dados centrados e padronizados.