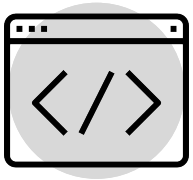
escola britânica de artes criativas & tecnologia

Profissão Cientista de Dados

escola britânica de artes criativas & tecnologia

BOAS PRÁTICAS

Regressão IV



- Apresente a Introdução
- Avalie as Suposições
- Identifique Outliers
- Aborde Casos de Correções
- Analise a Correlação nas Variáveis Explicativas - Multicolinearidade
- Explore Variáveis Ortogonais

Apresente a Introdução

- Treine o modelo com um objetivo específico em mente: Assim como um atleta treina para uma competição específica, um modelo de regressão ou qualquer modelo preditivo ou estatístico deve ser treinado com um objetivo específico em mente. Isso ajudará a garantir que o modelo seja eficaz e preciso.
- Esteja ciente do risco de modelo: O risco de modelo pode ocorrer quando se faz uma aplicação cruzada, ou seja, quando se usa um modelo para um propósito para o qual ele não foi originalmente projetado. Esteja ciente desse risco e tome as precauções necessárias.

Apresente a Introdução

- Monitore constantemente o modelo: As populações podem mudar com o tempo, e isso pode afetar a eficácia do modelo. É importante monitorar constantemente o modelo para detectar quando ele começa a deteriorar e fazer os ajustes necessários.
- Considere todos os componentes do modelo: Um modelo tem três componentes - entrada de informações, processamento e report. Certifique-se de considerar todos esses componentes ao construir e usar um modelo.

Avalie as Suposições

- Esteja ciente de que a variância pode aumentar com o valor da variável dependente, indicando heterocedasticidade. Isso pode ser um problema e deve ser corrigido.
- A distribuição normal dos resíduos é importante. Se a distribuição for muito assimétrica ou não se parecer com uma curva de sino, pode ser necessário usar uma técnica diferente que não dependa dessa suposição.

Avalie as Suposições

- As suposições a serem verificadas incluem a independência dos resíduos, a normalidade da distribuição dos resíduos com média zero e variância constante, e a homocedasticidade.
- Use exemplos práticos para verificar se as suposições são atendidas. Por exemplo, crie um modelo de regressão simples e analise os resíduos.
- A suposição de homocedasticidade é particularmente importante. Se ela falhar, pode comprometer as estimativas do modelo. Portanto, sempre verifique a homocedasticidade ao criar um modelo de regressão.

Identifique Outliers

- Sempre verifique a presença de outliers em seus dados: Outliers podem distorcer significativamente a análise de regressão e enviesar todo o modelo. Portanto, é crucial detectar e lidar com outliers antes de prosseguir com a análise.

- Preste atenção aos resíduos: O valor do resíduo é um indicador importante para determinar se um ponto é um outlier. Valores discrepantes podem indicar a presença de um outlier.

Identifique Outliers

- Esteja ciente dos pontos de alavanca: Um ponto de alavanca é um ponto que tem uma grande influência na regressão e pode alterar significativamente a reta de regressão. A presença de um ponto de alavanca com um resíduo alto é um sinal de um ponto influente que pode ser perigoso para a análise.

- Não se esqueça de tratar os outliers: A correção básica é eliminar o outlier, mas outras correções podem incluir trucagem ou transformação.

- Esteja preparado para lidar com a heteroscedasticidade no modelo: A próxima aula abordará este tópico, mas é importante estar ciente de que a heteroscedasticidade pode afetar a análise de regressão.

Identifique Outliers

- Esteja ciente dos pontos de alavanca: Um ponto de alavanca é um ponto que tem uma grande influência na regressão e pode alterar significativamente a reta de regressão. A presença de um ponto de alavanca com um resíduo alto é um sinal de um ponto influente que pode ser perigoso para a análise.

- Não se esqueça de tratar os outliers: A correção básica é eliminar o outlier, mas outras correções podem incluir trucagem ou transformação.

- Esteja preparado para lidar com a heteroscedasticidade no modelo: A próxima aula abordará este tópico, mas é importante estar ciente de que a heteroscedasticidade pode afetar a análise de regressão.

Aborde Casos de Correções

- Avalie a distribuição normal dos resíduos. Embora possa haver alguns outliers, a distribuição deve ter uma aparência razoavelmente normal para um bom modelo de regressão.

- Não hesite em remover outliers e reavaliar o modelo. No entanto, é importante entender o motivo desses outliers e considerar se eles são representativos de uma tendência real nos dados.

Aborde Casos de Correções

- Sempre verifique a independência dos dados ao avaliar um modelo de regressão. Isso pode ser feito usando o resíduo estudante.

- Lembre-se de que a adequação do modelo é mais importante do que a variância. Embora a variância possa aumentar ligeiramente ao remover outliers, a adequação geral do modelo pode melhorar.

- Ao identificar pontos de dados problemáticos ou outliers, considere a possibilidade de transformar os dados para melhorar a adequação do modelo. No exemplo da aula, o professor sugeriu a modelagem no log da gorjeta e no log do total da conta.

Analise a Correlação nas Variáveis Explicativas - Multicolinearidade

- Esteja ciente da multicolinearidade: Como cientista de dados, é importante estar ciente da multicolinearidade ao construir modelos de regressão. A multicolinearidade pode levar a resultados instáveis e deve ser tratada adequadamente.
- Use o Variance Inflation Factor (VIF): O VIF é uma medida útil que indica o quanto a variância dos coeficientes de regressão estimados é aumentada devido à multicolinearidade. Calcule o VIF para cada variável explicativa no seu modelo de regressão.
- Trate a multicolinearidade adequadamente: Se você identificar um problema de multicolinearidade, considere usar técnicas como a análise de componentes principais (PCA) ou a remoção do efeito de uma variável em outras variáveis para tratar a multicolinearidade.

Explore Variáveis Ortogonais

- Resolução de multicolinearidade: Ao lidar com dados multicolineares, considere a construção de variáveis ortogonais. Isso pode ajudar a obter informações de maneira mais eficaz e evitar problemas de multicolinearidade.

- Centralização de variáveis: Ao analisar dados de diferentes grupos, considere subtrair a média de cada grupo de cada variável. Isso pode ajudar a remover a informação linear da variável de grupo e alterar a interpretação da variável para ser quanto um indivíduo é maior ou menor que a média de seu grupo.

Explore Variáveis Ortogonais

- Correlação entre variáveis: Ao construir modelos, é importante calcular e entender a correlação entre as variáveis. Isso pode ajudar a entender como as variáveis estão relacionadas e como elas podem afetar a variável de resposta.
- Ajuste de modelos: Os modelos devem ser ajustados conforme necessário para garantir que eles sejam consistentes e interpretáveis. Isso pode envolver a adição ou remoção de variáveis, a alteração da forma do modelo ou a alteração dos parâmetros do modelo.

escola britânica de artes criativas & tecnologia

Bons estudos!