escola britânica de artes criativas & tecnologia

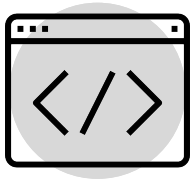
Profissão: Cientista de Dados

escola britânica de artes criativas & tecnologia

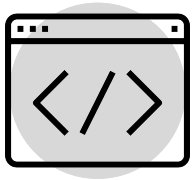
BOAS PRÁTICAS

Regressão I

- Interprete os parâmetros
- Utilize o statsmodels para Regressão
- Estime mínimos quadrados
- Avalie a qualidade do modelo
- Analise os resíduos



Regressão I



- Transforme as variáveis preditoras (x)
- Transforme a variável resposta (y)
- Multiplique matrizes
- Aplique álgebra matricial em modelos de regressão
- Conheça statsmodels e Patsy

Interprete os parâmetros

- Cada parâmetro em um modelo de regressão tem um significado específico. Por exemplo, o parâmetro alfa é o valor esperado de Y quando X é zero, enquanto beta é o aumento esperado para Y para cada unidade que X aumenta. Certifique-se de entender o que cada parâmetro representa.
- A modelagem é um processo iterativo. Esteja disposto a ajustar e melhorar o modelo com base nos resultados e nas limitações observadas.
- Lembre-se de que há sempre um erro aleatório ao prever o valor de Y com o modelo. Este erro pode ser para cima ou para baixo.
- Cada modelo tem suas limitações. Por exemplo, o modelo de regressão discutido presume que a variação é a mesma para todos os valores de X, o que nem sempre é verdade. No entanto, apesar de suas limitações, o modelo ainda pode ser útil.

Utilize statsmodels para Regressão

- Antes de rodar uma regressão, é útil fazer uma análise gráfica dos dados. Isso pode ajudar a identificar tendências, outliers ou outros aspectos interessantes dos dados.
- Lembre-se de que o statsmodels pode ser usado para fazer previsões para novas observações. Use o método 'predict' para isso.
- Não se esqueça de verificar os resíduos e o erro quadrado médio. Estes são indicadores importantes da qualidade do seu modelo.
- Depois de rodar a regressão, sempre verifique o resumo dos resultados. Preste atenção especial ao tamanho da amostra e aos parâmetros estimados para o intercepto e a inclinação.
- Você pode acessar os parâmetros do modelo diretamente do objeto de regressão. Use a tabulação para listar todos os atributos e métodos disponíveis.

Estime mínimos quadrados

- O objetivo do método de mínimos quadrados é encontrar os valores que minimizam a soma dos quadrados dos resíduos. Isso pode ser feito derivando a soma dos quadrados dos resíduos em relação aos parâmetros do modelo, igualando a zero e resolvendo para os parâmetros.
- A distribuição dos estimadores é conhecida, o que permite fazer inferências sobre eles. Isso é útil para estimar intervalos de confiança e realizar testes de hipóteses.
- As estimativas obtidas por meio do método de mínimos quadrados têm algumas propriedades importantes. Por exemplo, a soma dos resíduos é sempre zero, e os resíduos não têm correlação linear com as variáveis preditoras.

Avalie a qualidade do modelo

- Lembre-se de que a soma de quadrados total é igual à soma de quadrados do modelo mais a soma de quadrados do erro. Um modelo de alta qualidade terá uma soma de quadrados do modelo alta e uma soma de quadrados do resíduo baixa.
- Embora um R quadrado alto seja desejável, também é importante considerar possíveis influências e fragilidades no modelo. Não se baseie apenas em uma única métrica para avaliar a qualidade do seu modelo.
- Use o coeficiente de determinação, ou R quadrado, para avaliar a qualidade do seu modelo. O R quadrado é interpretado como a proporção da variância explicada pelo modelo. Quanto maior o R quadrado, melhor o modelo.
- Além do R quadrado, considere também o coeficiente de correlação, que é a raiz quadrada do R quadrado.

Analise os resíduos

- Em um modelo bem ajustado, espera-se que os resíduos não exibam um padrão uniforme nem apresentem qualquer relação evidente com a variável independente (X). A detecção de um padrão nos resíduos pode sugerir a necessidade de ajustes no modelo.
- Utilize programas de planilha eletrônica para ajudar a linearizar a relação entre as variáveis, usando uma função G de X. A ideia é fazer com que a variável dependente (Y) seja linear em relação ao G de X, melhorando assim o ajuste do modelo.
- Esteja atento aos padrões típicos de resíduos que indicam problemas no modelo. Por exemplo, se os resíduos aumentam à medida que a variável X aumenta, isso sugere que os dados têm um formato de cone. Se os resíduos são negativos em um trecho e positivos em outro, isso indica uma relação não linear que precisa ser linearizada.

Analise os resíduos

- Em um modelo bem ajustado, espera-se que os resíduos não exibam um padrão uniforme nem apresentem qualquer relação evidente com a variável independente (X). A detecção de um padrão nos resíduos pode sugerir a necessidade de ajustes no modelo.
- Utilize programas de planilha eletrônica para ajudar a linearizar a relação entre as variáveis, usando uma função G de X. A ideia é fazer com que a variável dependente (Y) seja linear em relação ao G de X, melhorando assim o ajuste do modelo.
- Esteja atento aos padrões típicos de resíduos que indicam problemas no modelo. Por exemplo, se os resíduos aumentam à medida que a variável X aumenta, isso sugere que os dados têm um formato de cone. Se os resíduos são negativos em um trecho e positivos em outro, isso indica uma relação não linear que precisa ser linearizada.

Transforme as variáveis preditoras (x)

- Ao ajustar um modelo de regressão, não se baseie apenas no valor de R quadrado para avaliar o desempenho do modelo. Verifique também os resíduos para identificar qualquer padrão que possa indicar que o modelo pode ser melhorado.

- Experimente diferentes tipos de transformações nas variáveis preditoras para melhorar o ajuste do modelo. Isso pode incluir ajustar um polinômio de segundo grau, uma função exponencial ou uma função logarítmica.

- A transformação de variáveis é uma ferramenta poderosa, mas deve ser usada com cautela. Sempre verifique os resíduos após a transformação para garantir que a relação entre as variáveis preditoras e a variável resposta tenha sido adequadamente capturada.

Transforme a variável resposta (y)

- Se os dados apresentarem características multiplicativas, considere transformar a variável resposta em um logaritmo. Isso pode transformar uma característica multiplicativa em uma aditiva, tornando o modelo mais preciso.

- Diferentes pessoas podem resolver o mesmo problema de maneiras diferentes. Não há uma única "melhor" maneira de fazer as coisas, então seja criativo e pense fora da caixa.

Multiplique matrizes

- A multiplicação de matrizes segue uma regra específica - o número de colunas da primeira matriz deve ser igual ao número de linhas da segunda matriz. O resultado dessa multiplicação será uma matriz com o número de linhas da primeira matriz e o número de colunas da segunda matriz. Não seguir essa regra pode levar a erros.

- A multiplicação de matrizes pode parecer complexa no início, mas é uma sequência de passos que o computador pode resolver. O objetivo é entender o conceito, não necessariamente fazer os cálculos manualmente.

Aplique álgebra matricial em modelos de regressão

- A matriz de design é uma matriz de dados usada na regressão. A primeira coluna desta matriz é sempre composta por uns, pois está associada ao intercepto. As colunas subsequentes representam as variáveis do modelo.
- Cada linha da matriz representa uma observação e as colunas representam as variáveis. A multiplicação da matriz de design pelo vetor de parâmetros resulta nas previsões para cada observação.

Aplique álgebra matricial em modelos de regressão

- Quando há mais de uma variável explicativa, a matriz de design se expande para acomodar variáveis adicionais, e o vetor de parâmetros também cresce para incluir um coeficiente para cada variável.
- Variáveis qualitativas devem ser codificadas como variáveis dummy para que possam ser usadas em modelos de regressão. A multiplicação da matriz de design expandida pelo vetor de parâmetros expandido resulta nas previsões para uma regressão múltipla.

Conheça statsmodels e Patsy

- O Patsy também permite aplicar funções às variáveis, como padronização ou centralização. Isso pode ser útil para preparar seus dados para análise.
- Ao usar o statsmodels, você pode ajustar seu modelo e obter um resumo das estatísticas do modelo. Isso pode fornecer informações valiosas sobre o desempenho do seu modelo.
- O statsmodels oferece duas APIs: a API padrão e a API de fórmula. A API padrão requer a definição separada da matriz de design e do modelo, enquanto a API de fórmula permite definir tudo em uma única linha de comando. Escolha a que melhor se adapta às suas necessidades e preferências.

Conheça statsmodels e Patsy

- O Patsy também permite aplicar funções às variáveis, como padronização ou centralização. Isso pode ser útil para preparar seus dados para análise.
- Ao usar o statsmodels, você pode ajustar seu modelo e obter um resumo das estatísticas do modelo. Isso pode fornecer informações valiosas sobre o desempenho do seu modelo.
- O statsmodels oferece duas APIs: a API padrão e a API de fórmula. A API padrão requer a definição separada da matriz de design e do modelo, enquanto a API de fórmula permite definir tudo em uma única linha de comando. Escolha a que melhor se adapta às suas necessidades e preferências.

escola britânica de artes criativas & tecnologia

Bons estudos!