Regressão Linear Múltipla

Motivação: Elaborar modelo de regressão múltipla para prever as vendas de uma empresa de assentos infantis para carros.

**Professor:** BRUNO TEBALDI DE QUEIROZ BARBOSA

**Alunos:**

EDUARDO FLEISCHMANN

LUCAS LUCENA FALBO

MARIANE KIAM ITOCAZU

RAFAEL DA SILVA

ROBIN DANIEL HONSI

Base de dados: “Carseats.csv”

Campos disponíveis:

**• Sales:** Unit sales (in thousands) at each location

**• CompPrice:** Price charged by competitor at each location

**• Income:** Community income level (in thousands of dollars)

**• Advertising:** Local advertising budget for company at each location (in thousands

of dollars)

**• Population:** Population size in region (in thousands)

**• Price:** Price company charges for car seats at each site

**• ShelveLoc:** A factor with levels Bad, Good and Medium indicating the quality of the shelving location for the car seats at each site

**• Age:** Average age of the local population

**• Education:** Education level at each location

**• Urban:** A factor with levels No and Yes to indicate whether the store is in an urban or rural location

**• US:** A factor with levels No and Yes to indicate whether the store is in the US or not

**Questões:**

**1-) Fornecer um modelo de regressão múltipla para prever vendas usando as seguintes variáveis: Price, Urban e US.**

Antes de construirmos o modelo de regressão múltipla, é boa prática explorar os dados e entender seus comportamentos.

Para tal, iremos plotar alguns gráficos.

Histograma da variável “Sales”

Gráfico, Histograma

Descrição gerada automaticamente

Podemos notar que há um comportamento que indica uma distribuição bem próxima da normal para as observações das vendas.

Para entender se há uma correlação entre os dados, e entender qual a proporção dessa correlação, iremos plotar alguns gráficos de dispersão comparando algumas variáveis disponíveis.

Dispersão entre as variáveis Sales e Price:

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

O gráfico de dispersão acima indica que há uma forte correlação entre o preço e as vendas. Esta indicação faz sentido do ponto de vista econômico, ao aumentarmos o preço, as vendas tendem a diminuir.

Boxplot entre as variáveis Sales e Urban

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente

Podemos notar que não há uma diferença significativa entre as vendas realizadas por lojas urbanas ou não.

Boxplot entre Sales e US

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente

Podemos avaliar que há uma diferença clara entre as vendas de lojas dentro dos EUA e fora, porém, nenhuma conclusão deve ser tirada dessa relação, pois não temos ideia da quantidade de lojas nos EUA em comparação com as lojas de fora, se essa relação for desproporcional, indica que esse pode ser um fator relevante para a quantidade de vendas ser maior nas lojas dos EUA.

Com base nas análises anteriores, vamos construir o modelo de regressão múltipla considerando as variáveis indicadas pelo professor:

Texto

Descrição gerada automaticamente

Para construção deste modelo, nós julgamos mais adequado converter as variáveis US e Urban em variáveis dummy com os valores 0 e 1.

**2-) Fornecer uma interpretação de cada coeficiente no modelo;**

Analisando o modelo, podemos identificar os níveis de significância de cada coeficiente.

Para o coeficiente Price, o nível de relevância estatística é alto de acordo com o p valor calculado no modelo. Esse coeficiente possui um erro padrão baixo o que demostra que ele explica bem a variável alvo. Olhando o valor estimado, podemos concluir que a cada 5% de aumento no valor do preço as vendas caem na mesma proporção.

Para o coeficiente UrbanYes (dummy da variável Urban), o p valor indica que essa variável não é estatisticamente relevante para o modelo.

Para o coeficiente USYes (dummy da variável US), o p valor indica que é estatisticamente relevante para o modelo, porém, com um peso menor do que o coeficiente Price.

3-) Determinar se existem preditores significantes bem como explicar sobre a qualidade do modelo.

Para determinar se existem outros preditores com maior significância, nós realizamos uma análise gráfica comparando algumas outras variáveis da base de dados comparando com a variável “Sales”.

Dispersão entre Sales e Age:

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

A variável “Age” tem correlação negativa com as vendas, porém, numa proporção menor do que a que esperávamos. Nós esperávamos que as vendas diminuíssem mais de acordo com o aumento da idade, ou seja, que a correlação fosse “mais” negativa. Em função do exposto, não acreditamos que essa variável seja um preditor com alto grau de significância para o modelo.

Dispersão entre Sales e Advertising:

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Como podemos ver no gráfico de dispersão acima, existe uma forte correlação entre a variável “Advertising” e a variável “Sales” apesar de uma forte concentração de observações quando o valor da variável “Advertising” é 0. Em função deste comportamento, consideramos este preditor com alto grau de significância para o modelo.

Realizamos o plot de um gráfico Boxplot entre as variáveis “ShelveLoc” e “Sales” e obtivemos o seguinte resultado:

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente

Como podemos reparar, esta variável categórica influi no comportamento das vendas de maneira muito intensa. Portanto, consideramos essa variável como um preditor com alto grau de significância para o modelo.

A seguir, nós construímos um novo modelo considerando as novas variáveis levantadas nas análises anteriores, e obtivemos o seguinte resultado:

Tabela

Descrição gerada automaticamente

Uma das maneiras de medir a robustez de um modelo de regressão, é utilizando o coeficiente R2. Quanto mais próximo o R2 for do valor 1, mais robusto é o modelo.

O R2 do modelo anterior foi de 0.2393.

O novo modelo considerando as novas variáveis obteve um R2 de 0.6085, indicando que essas variáveis possuem uma maior significância para o modelo.

Gráfico, Gráfico de dispersão

Descrição gerada automaticamenteAbaixo plotamos dois gráficos de dispersão comparando os resultados de cada um dos modelos para visualizar suas diferenças:

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente