PRO3200 – EPUSP
Professora Celma de Oliveira Ribeiro
Assistente: Pedro Gerber Machado & Monitor: Vinícius Castanho

Estudo de Caso 9: Regressão Linear

Analisaremos a planilha "vendas.rds", extraída e editada do Kaggle, que contém informações sobre o investimento mensal, em milhares, de uma empresa em marketing para diversas plataformas e seu retorno em número de vendas. **Lembre-se de comentar e tirar conclusões dos exercícios realizados.**

1. Regressão Linear: Teoria

Podemos ajustar um modelo em que o número de vendas seja uma função linear do investimento em marketing pela TV. Nesse caso, podemos expressar o modelo da seguinte forma:

$$y = \beta_0 + \beta_1 x + e$$

Em que y é a variável dependente; β_0 é o coeficiente linear ou intercepto da reta de regressão; β_1 é o coeficiente angular ou inclinação (declive) da reta de regressão; x é a variável independente; e o erro aleatório referente a variabilidade em y que não pode ser explicada pela variável x.

Para estimação dos parâmetros β_0 e $\beta_1$são feitas as seguintes hipóteses: a primeira delas é que a variável X (investimento na TV) é controlada, ou seja, não está sujeita a variações aleatórias; a segunda hipótese é de que os erros se distribuem ao redor da média $\beta_0 + \beta_1 x_1$ com média zero, isto é,

$$E(e_i/x) = 0$$

Em terceiro lugar temos a suposição de que as variabilidades dos erros em torno dos níveis X sejam iguais, isto é,

$$Var(e_i/x) = \sigma_e^2$$

E em quarto lugar introduziremos a restrição de que os erros não sejam correlacionados. A partir das suposições feitas, devemos encontrar os estimadores $\widehat{\beta_0}$ e $\widehat{\beta_1}$, que estimam o coeficiente linear (intercepto) e coeficiente angular do modelo de regressão linear simples proposto. No presente estudo de caso daremos maior atenção à obtenção (via R) desses coeficientes bem como à utilização de inferência estatística para entender melhor nosso modelo de regressão.

a. O que é uma variável aleatória? O que são experimentos determinísticos? E probabilísticos?

b. O que são parâmetros, estimadores e estimativas?

c. Construa o gráfico de dispersão do investimento em TV em função do número de vendas. A partir do gráfico, há indicação de alguma relação entre o investimento na tv e o número de vendas?

2. Regressão Linear: Prática

Para checar nossa suspeita, podemos criar um modelo linear no R utilizando a função *lm( )* [1]com a seguinte notação lm(y ~ x).

a. Crie a variável modelo=lm(vendas ~tv, data=df). Quais foram os coeficientes obtidos?

b. Desenhe a reta obtida no diagrama de dispersão utilizando a função abline( ) ou geom_abline( ). A tendência da reta encontrada é a mesma da sugerida por você no item 1?

[1] O nome da função advém do termo em inglês linear model.

3. Regressão Linear: Teste de Hipóteses

Podemos testar os coeficientes encontrados a partir de testes de hipótese, uma vez que a nossa reta de regressão é uma variável aleatória. Sob a hipótese nula de que $\beta_0 = 0$ e $\beta_1 = 0$, podemos testar se o intercepto populacional é diferente de zero e também se o coeficiente angular é diferente de zero. Nesse último caso, ao rejeitarmos a hipótese nula podemos concluir que há uma relação (no caso, linear) entre as variáveis de estudo. Outro parâmetro importante a ser analisado é o coeficiente de determinação R^2, que representa a porcentagem da variabilidade total explicada pelo modelo.

Podemos também testar se o modelo de regressão é significativo para explicar a variável y a partir da variável x (de maneira linear). Como na análise de variância, podemos calcular somas de quadrados e desenvolver essas somas num teste com distribuição F, que compara as somas de quadrados relacionadas à Regressão e à parcela Residual (aleatoriedade não contemplada pelo modelo). No caso, o teste é definido de maneira que a hipótese nula afirme que "o modelo NÃO seja significativo para explicar a variável y" e a hipótese alternativa de que "o modelo SEJA significativo para explicar a variável y".

Para o estudo de caso atual utilizaremos a função summary( ) aplicada à variável Modelo1 para obter as informações sobre os testes de hipótese e sobre o coeficiente de determinação

a. Utilize summary(Modelo1) e encontre os resultados para os testes de β_0 e β_1 , do teste sobre a regressão, bem como o valor de R^2. Explique o resultado de cada um deles.

b. Faça o histograma dos resíduos encontrados. Discorra sobre a distribuição dos resíduos[2].

c. Calcule a média dos resíduos. Ela está de acordo com a suposição feita na definição do modelo de regressão linear simples?

[2] Perceba que a variável Modelo1 possui o atributo *residuals.*

Podemos também fazer um teste de hipótese sobre o parâmetro coeficiente de correlação. Esse parâmetro é representado pela letra grega ρ e seu estimador é denotado por r. Podemos utilizar a função cor( ) para obtenção do coeficiente de correlação r a partir dos dados amostrais obtidos.

d. Utilize a função cor( ) e determine o coeficiente de correlação linear. O que ele significa?

Podemos testar o coeficiente de correlação, a partir de um teste em que a hipótese nula é definida como $\rho = 0$ e a hipótese alternativa pode conter $\rho \neq 0, \rho > 0 \; ou \; \rho < 0$. No ambiente R, a função cor.test( ) retorna o teste de hipótese para o coeficiente de correlação.

e. Faça um teste de hipótese sobre o coeficiente de correlação, considerando a hipótese alternativa $H_1: \rho \neq 0$. Comente o resultado do teste.

4. Regressão Linear: Intervalo de Confiança e ANOVA

Agora que você já testou os estimadores da reta de regressão encontrada, analisou os resíduos da regressão, testou o coeficiente de correlação linear de Pearson, vamos aprender sobre intervalos de confiança para média de y e intervalos de predição.

Podemos estar interessados, para um determinado x, em calcular um intervalo de confiança para o *valor médio* de y. Também há a possibilidade de nos interessarmos em estimar, para um determinado x, o intervalo de predição de y. Nesse caso, o intervalo que obteremos conterá os possíveis valores que Y pode assumir (e não a média de Y).

É importante não confundir os dois conceitos. No aspecto mais prático podemos observar que o intervalo de predição costuma ser mais espaçado (com maior amplitude) que o intervalo de confiança para média de y.

Podemos obter o resultado de cada intervalo com a função *predict( )* no R. Para um investimento em tv podemos proceder da seguinte maneira[3]:

```
67  tv=data.frame(tv=sort(df$tv))
```

Para o intervalo de confiança para a média do valor de y utilizamos o seguinte comando:

```
69  estimados_IC=predict(modelo,tv,interval = "confidence")
```

Para o intervalo de predição de y utilizamos o seguinte comando:

```
70  previstos_IP=predict(modelo,tv,interval = "prediction")
```

Os resultados retornados pela função *predict( )* são: valor de y na reta (fit), limite inferior (lwr) e superior (upr) do intervalo analisado.

a. Desenvolva o gráfico dos intervalos de predição e confiança (para o investimento em tv).

b. Faça uma Análise de Variâncias (ANOVA) para seu modelo de regressão linear e comente seus resultados.

[3] Certifique-se de criar a variável do modelo linear dessa maneira, não fazer assim pode atrapalhar o uso da função predict().

3. Entrega no Moodle.

Os cases devem ser enviados no e-disciplinas em um arquivo .pdf com o script do R anexo ao final do próprio PDF, de forma a possibilitar o Ctrl c, Ctrl v do mesmo para efeitos de correção.

Lembre-se de que dissertações e conclusões acerca dos resultados são mais importantes que a própria construção do código em R. Indique todos os resultados da maneira mais expositiva possível.

O prazo de entrega é domingo, 25/06, às 23h59.