PRO2514 - Pesquisa Quantitativa em Gestão de Operações

Regressão Linear

Prof. Dr. Renato de Oliveira Moraes

$\hat{Y}$ = Variável estatística de regressão

Variável dependente

$$y \cong \hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n$$

Variáveis independentes
Variáveis preditoras
Variáveis explicativas

Base de dados sobre uso de cartões de crédito

- Y: Número de cartões de crédito usados
- X1: Tamanho da família
- X2: Renda familiar ($ mil)
- X3: Número de automóveis

		Número de cartões de crédito usados (Y)	Tamanho da família (X1)	Renda familiar ($ mil) (X2)	Número de automóveis (X3)
Número de cartões de crédito usados (Y)	Pearson Correlation	1	,866	,829	,342
	Sig. (2-tailed)		,005	,011	,407
	N	8	8	8	8
Tamanho da família (X1)	Pearson Correlation	,866	1	,673	,192
	Sig. (2-tailed)	,005		,068	,649
	N	8	8	8	8
Renda familiar ($ mil) (X2)	Pearson Correlation	,829	,673	1	,301
	Sig. (2-tailed)	,011	,068		,469
	N	8	8	8	8
Número de automóveis (X3)	Pearson Correlation	,342	,192	,301	1
	Sig. (2-tailed)	,407	,649	,469	
	N	8	8	8	8

RStudio

1. Carregar o arquivo de dados (XLXS)
2. Fazer uma regressão com todas as variáveis

- lm (formula = Y ~ x1 +x2 + x3, data = Cartões_de_Credito)

OU, para ter mais informações (de forma semelhante ao que faz o MS Excel)

- Regr_Cartoes = lm (formula = Y ~ x1 +x2 + x3, data = Cartões_de_Credito)
- summary(Regr_Cartoes)

lm (formula = Y ~ x1 +x2 + x3, data = Cartões_de_Credito)

Call:

lm(formula = Y ~ x1 + x2 + x3, data = Cartões_de_Credito)

Coefficients:

(Intercept)	x1	x2	x3
0.2861	0.6346	0.1995	0.2716

summary(Regr_Cartoes)

Call:
lm(formula = Y ~ x1 + x2 + x3, data = Cartões_de_Credito)
Residuals:

1	2	3	4	5	6	7	8
-0.6202	0.7091	-0.1611	0.5120	0.1346	-1.1923	0.2428	0.3750

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2861	1.6059	0.178	0.8673
x1	0.6346	0.2710	2.341	0.0793 .
x2	0.1995	0.1194	1.671	0.1701
x3	0.2716	0.4702	0.578	0.5945

summary(Regr_Cartoes) - continuação

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8389 on 4 degrees of freedom

Multiple R-squared: 0.872, Adjusted R-squared: 0.7761

F-statistic: 9.087 on 3 and 4 DF, p-value: 0.02935

Colinearidade entre indicadores

- VIF – Fator de inflação da variação

  VIF > 5 ➔ problemas de colinearidade entre indicadores

- $x_1 \cong f(x_2; x_3\,; x_4 \cdots x_k\,)$ (regressão linear)
- R^2 (Coeficiente de determinação da regressão)
- $TOL = 1 - R^2$ (Tolerância)
- $VIF = \frac{1}{TOL}$

Multiconlinearidade no RStudio

car::vif(Regr_Cartoes)

x1	x2	x3
1.826923	1.934924	1.099760

Métodos de inclusão/remoção de variáveis dependentes no modelo de regressão linear

- Forward – começa com nenhuma variável preditora e vai acrescentando variáveis no modelo enquanto isso melhora a qualidade do modelo
- Backward – começa com todas as variáveis candidatas a preditora e vai retirando variáveis do modelo enquanto isso melhora a qualidade do modelo

- Stepwise – semelhante ao Forward, mas é possível que em algum passo, uma variável seja removida do modelo

Regressão com método Forward

- media = lm (Y ~ 1, data = Cartões_de_Credito)
- Regr_forward <- step(media, direction = "forward", scope = formula(Regr_Cartoes), trace = 0)
- summary(Regr_forward)

Para ver a evolução da construção dos modelos: **trace = 1**

- Regr_forward <- step(media, direction = "forward", scope = formula(Regr_Cartoes), **trace = 1**)

summary(Regr_forward)

Call:
lm(formula = Y ~ x1 + x2, data = Cartões_de_Credito)

Residuals:

1	2	3	4	5	6	7	8
-0.76803	0.80027	-0.03251	0.31995	0.47186	-1.17568	0.05546	0.32869

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4817	1.4614	0.330	0.7551
x1	0.6322	0.2523	2.506	0.0541 .
x2	0.2158	0.1080	1.998	0.1021

Mudança de método de inclusão/remoção de variáveis no modelo

Alterar o valor do parâmetro direction. As opções possíveis são:

- "forward"
- “backward“
- “both“ (método *stepwise*)

Maior controle sobre os critérios de inclusão e remoção de variáveis

- Usar a biblioteca **olsrr**. Esta biblioteca oferece suporte a modelos de regressão baseados em mínimos quadrados ordinários (ordinary least squares – OLS)
  library(olsrr)
- Para construir um modelo de regressão pelo método forward é preciso definir a probabilidade para entrada de um anova variável no modelo (parâmetro penter)
  ols_step_forward_p(Regr_Cartoes, penter = 0.05, details = TRUE)
- Observe, contudo, que com o valor usual de probabilidade (5%) o modelo resultante tem apenas uma variável preditora.

Diferentes métodos e inclusão/remoção de variáveis no modelo

- library(olsrr)

Método Forward

- m = ols_step_forward_p(Regr_Cartoes, penter = 0.05, details = TRUE)

Método Backward

- ols_step_backward_p(Regr_Cartoes, prem = 0.10, details = TRUE)

Método Stepwise

- ols_step_both_p(Regr_Cartoes, penter = 0.05, prem = 0.10, details = TRUE)

Dados originais				Média		F1		F2	
Y	x1	x2	x3	Y médio	Erro^2	2,871+0,971*X1	Erro^2	0,482+0,632*X1+0,216*X2	Erro^2
4	2	14	1	7	9	4,813	0,66	4,77	0,59
6	2	16	2	7	1	4,813	1,41	5,202	0,64
6	4	14	2	7	1	6,755	0,57	6,034	0,00
7	4	17	1	7	0	6,755	0,06	6,682	0,10
8	5	18	3	7	1	7,726	0,08	7,53	0,22
7	5	21	2	7	0	7,726	0,53	8,178	1,39
8	6	17	1	7	1	8,697	0,49	7,946	0,00
10	6	25	2	7	9	8,697	1,70	9,674	0,11
					22		**5,49**		**3,05**

Tipos de problemas de pesquisa

- Previsão
- Explicação

- Tamanho da amostra
  - Mínimo 5 x número de variáveis
  - Ideal: entre 15 e 20 x número e variáveis

Seleção das variáveis

- Erro de medida: grau em uma variável é uma medida precisa e consistente do conceito em estudo
- Erro de especificação:
  - Inclusão de variáveis irelevantes
  - Omissão de de variáveis relevantes

Suposições do modelo de regressão linear

- Linearidade do fenômeno medido
- Variância constante dos termos de erro
- Independência dos termos de erro
- Normalidade da distribuição dos termos de erro

Análise da significância do modelo

- Significância geral – coeficiente de determinação
  - R, R^2 e R^2 ajustado

    H_0: $R^2 = 0$

    H_1: $R^2 \neq 0$
- Significância dos coeficientes de regressão

    H_0: $b_i = 0$

    H_1: $b_i \neq 0$

Multicolinearidade

- Interpretação do papel das variáveis no modelo de explicação
- Análise
  - Matriz de correlação
  - Tolerância: o quanto de uma variável independente não é explicada pelas demais variáveis independentes

    Fator de inflação da variância (VIF): inverso da tolerância – 1/Tolerância

Correção da multicolinearidade

- Omitir uma ou mais variáveis independetes (aumenta o risco de erro de especificação)
- Usar o modelo apenas para previsão, e não para explicação
- Usar a matriz de correlações para explicar o comportamento da variável dependente
- Usar métodos mais sofisticados de análise, como regressão Bayesiana (além do nosso escopo)

Validação dos resultados

- Amostras adicionais
  - Avaliar o poder preditivo do modelo gerado
  - Comparação com novo modelo
- Amostras parcionadas
- Comparação de modelos de regressão – uso de R^2 e do R^2 ajustado

Exemplo

Modelo de previsão

- Variável dependente: x9: Nível de uso - quanto do produto total da empresa é comprado da HATCO, medido numa escala de 100 pontos percentuais
- Variáveis independentes

  x1: Velocidade de entrega

  x2: Nível de preço

  x3: Flexibilidade de preço

  x4: Imagem do fabricante

  x5: Serviço geral

  x6: Imagem da força de vedas

  x7:Qualidade do produto

Coeficiente de Correlação de Pearson

	x9 - Nível de uso Velocidade	x1 - Velocidade de entrega	x2 - Nível de preço	x3 - Flexibilidade de preço	x4 - Imagem do fabricante	x5 - Serviço geral	x6 - Imagem da força de vendas
x1 - Velocidade de entrega	0,676 (0,000)						
x2 - Nível de preço	0,082 (0,418)	-0,349 (0,000)					
x3 - Flexibilidade de preço	0,559 (0,000)	0,509 (0,000)	-0,487 (0,000)				
x4 - Imagem do fabricante	0,224 (0,025)	0,05 (0,618)	0,272 (0,006)	-0,116 (0,250)			
x5 - Serviço geral	0,701 (0,000)	0,612 (0,000)	0,513 (0,000)	0,067 (0,510)	0,299 (0,003)		
x6 - Imagem da força de vendas	0,256 (0,001)	0,077 (0,446)	0,186 (0,064)	-0,034 (0,735)	0,788 (0,000)	0,241 (0,016)	
x7 - Qualidade do produto	-0,192 (0,055)	-0,483 (0,000)	0,470 (0,000)	-0,448 (0.000)	0,200 (0,046)	-0,055 (0,586)	0,177 (0,078)

Construção do modelo com todas os possíveis preditores

```
Regr_todos = lm(formula = x9 ~ x1+x2+x3+x4+x5+x6+x7, data = Hatco)

summary(Regr_todos)
```

summary(Regr_todos)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-10.18687	4.97678	-2.047	0.0435 *
x1	-0.05758	2.01266	-0.029	0.9772
x2	-0.69691	2.09017	-0.333	0.7396
x3	3.36822	0.41123	8.191	1.44e-12 ***
x4	-0.04220	0.66681	-0.063	0.9497
x5	8.36914	3.91815	2.136	0.0353 *
x6	1.28067	0.94717	1.352	0.1797
x7	0.56693	0.35543	1.595	0.1141

Verificando problema de multicolinearidade

- car::vif(Regr_todos)

x1	x2	x3	x4	x5
35.746811	31.597377	1.644718	2.879474	43.834228

x6	x7
2.696942	1.606105

Construção pelo método forward

- modelo_1 = ols_step_forward_p(Regr_todos, penter = 0.05, details = TRUE)

- summary (modelo_1$model)

Parameter Estimates

---

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-6.520	3.247		-2.008	0.047	-12.965	-0.075
x5	7.621	0.607	0.637	12.547	0.000	6.416	8.827
x3	3.376	0.320	0.521	10.562	0.000	2.742	4.010
x6	1.406	0.591	0.121	2.378	0.019	0.232	2.579

summary (modelo_1$model)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.5201	3.2467	-2.008	0.0474 *
x5	7.6214	0.6074	12.547	<2e-16 ***
x3	3.3760	0.3196	10.562	<2e-16 ***
x6	1.4056	0.5910	2.378	0.0194 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

summary (modelo_1$model) - continuação

Residual standard error: 4.394 on 96 degrees of freedom

Multiple R-squared: 0.7683, Adjusted R-squared: 0.7611

F-statistic: 106.1 on 3 and 96 DF, p-value: < 2.2e-16

Multicolinearidade

car::vif(modelo_1$model)

x5	x3	x6
1.067922	1.007181	1.064437

Suposições do modelo de regressão linear

- Linearidade do fenômeno medido
- Variância constante dos termos de erro
- Independência dos termos de erro
- Normalidade da distribuição dos termos de erro

Tarefa

O que explica a satisfação com o restaurante?

Variáveis independentes:

X1 -- Excellent Food Quality
X2 -- Attractive Interior
X3 -- Generous Portions
X4 -- Excellent Food Taste
X5 -- Good Value for Money
X6 -- Friendly Employees
X7 -- Appears Clean and Neat
X8 -- Fun Place to Go
X9 -- Wide Variety of Menu Items
X10 -- Reasonable Prices
X11 -- Courteous Employees
X12 -- Competent Employees

Variável dependente:

X17 – Satisfaction