Universidade Federal do Paraná – Departamento de Estatística
Disciplina CE071 – Análise de Regressão Linear
Prof. Cesar Augusto Taconeli
Lista de exercícios

Exercício 1 - Sobre a análise de regressão linear múltipla:

a) O que a diferencia em relação à regressão linear simples?
b) Enuncie o modelo de regressão linear múltipla, identificando cada um de seus componentes;
c) Quais pressuposições são feitas com respeito à distribuição dos erros? Qual a implicação dessas pressuposições em relação à distribuição da variável resposta condicional aos valores das covariáveis?
d) Qual a interpretação dos parâmetros do modelo de regressão linear múltipla?
e) Qual o método usado na estimação dos parâmetros? Qual o princípio desse método?
f) Quais as principais propriedades de $\hat{\boldsymbol{\beta}}$, o vetor de estimadores de mínimos quadrados para os parâmetros do modelo?

Exercício 2 - Considere os seguintes dados:

$$\mathbf{Y} = \begin{bmatrix} 3 \\ 1 \\ 8 \\ 3 \\ 5 \end{bmatrix} \quad ; \quad \mathbf{X'Y} = \begin{bmatrix} 20 \\ 76 \\ 109 \end{bmatrix} \quad ; \quad (\mathbf{X'X})^{-1} = \begin{bmatrix} 26,7 & 4,5 & -8,0 \\ 4,5 & 1,0 & -1,5 \\ -8,0 & -1,5 & 2,5 \end{bmatrix}.$$

a) Apresente o modelo de regressão linear múltipla ajustado;
b) Interprete as estimativas obtidas para os parâmetros do modelo;
c) Calcule uma estimativa para a variância do erro aleatório;
d) Calcule o erro padrão para cada coeficiente do modelo;
e) Sabendo que a soma dos desvios quadráticos em torno da média é igual a 26.5, calcule e interprete o valor do coeficiente de determinação;
f) Teste a significância do modelo de regressão ao nível de 1% de significância.

Exercício 3 - Pretende-se ajustar um modelo de regressão linear múltipla para explicar a variação da viscosidade de um polímero (y) em função da temperatura de reação, x_1, e da taxa de alimentação do catalisador, x_2. Um experimento realizado com esse objetivo produziu os seguintes resultados:

Polímero	y	x_1	x_2	Polímero	y	x_1	x_2
1	2256	80	8	9	2364	94	12
2	2340	93	9	10	2379	93	11
3	2426	100	10	11	2440	97	13
4	2293	82	12	12	2364	95	11
5	2330	90	11	13	2404	100	8
6	2368	99	8	14	2317	85	12
7	2250	81	8	15	2309	86	9
8	2409	96	10	16	2328	87	12

a) Ajuste o modelo de regressão linear múltipla $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. Interprete a estimativa de cada parâmetro;
b) Represente graficamente o modelo ajustado;
c) Apresente uma estimativa para σ^2;
d) Apresente a matriz de covariâncias estimada de $\hat{\boldsymbol{\beta}}$;
e) Estime a viscosidade do polímero para uma temperatura de reação igual a 90 e taxa de alimentação do catalisador igual a 10;
f) Calcule o valor ajustado e o resíduo para o polímero de número 5;
g) Calcule o valor do coeficiente de determinação do ajuste e comente o resultado obtido.
h) Teste a significância do modelo ajustado ao nível de 5%;
i) Teste a significância de cada parâmetro do modelo;
j) Estime os parâmetros do modelo de regressão por meio de intervalos com 95% de confiança;
k) Estime a viscosidade média de polímeros sob temperatura de reação igual a 90 e taxa de alimentação do catalisador igual a 10 por meio de um intervalo com 95% de confiança;
l) Preveja a viscosidade de um polímero sob temperatura de reação igual a 90 e taxa de alimentação do catalisador igual a 10 por meio de um intervalo com 95% de confiança;
m) Apresente intervalos de confiança simultâneos (95%) para β_1 e β_2.

Nota – Recomendo que o exercício 3 seja resolvido, num primeiro momento, sem usar a função `lm` e seus recursos. Depois, você poderá usá-la para conferir os resultados.

Exercício 4 - O arquivo `atletas.csv`, disponível na página da disciplina, contém dados de 200 atletas de determinada modalidade esportiva. As variáveis que compõem a base são as seguintes:

`idade:` Idade da atleta (em anos);
`imc:` Índice de massa corporal (kg/cm^2);
`horas:` Horas de treino semanal;
`massa:` Percentual de massa gorda;
`gasto:` Gasto semanal com suplementos alimentares (reais);
`estudo:` Anos de estudo;
`escore:` Escore de desempenho nas últimas provas (numa escala de 0 a 100).

a) Faça uma análise exploratória dos dados. Construa gráficos uni e bivariados, calcule medidas de locação e dispersão para cada variável, obtenha as correlações;
b) Ajuste um modelo de regressão linear múltipla tomando o escore de desempenho como a variável resposta. Escreva a expressão do modelo ajustado;
c) Qual o valor ajustado pelo modelo para a primeira atleta da base? Qual o resíduo correspondente?

```
   escore idade  imc   horas  massa  gasto  estudo
1   79.6    30   20.5    19   21.66    72     15
```

d) Qual a probabilidade estimada de um atleta com características semelhantes à 1ª atleta da base ter escore de desempenho superior a 80? E entre 65 e 75?
e) Quais variáveis são estatisticamente significativas ao nível de 5% de significância?
f) Qual a variação esperada no escore de desempenho para um aumento de 1% de massa gorda, mantendo fixas as demais variáveis?
g) Qual a variação esperada no escore de desempenho para 1 ano a mais de idade, mantendo fixas as demais variáveis?
h) Qual a variação esperada no escore de desempenho para 20 reais semanais a mais gastos com suplementos alimentares?
i) Obtenha intervalos de confiança (95%) para os parâmetros do modelo;
j) Com base nos intervalos de confiança obtidos, você rejeitaria, ao nível de significância de 5%, a hipótese de não efeito da massa muscular no desempenho dos atletas? Justifique sua resposta;
k) Ainda com base nos intervalos de confiança, você rejeitaria a hipótese de que o desempenho caia, em média, dois pontos para 1% a mais de massa gorda? Justifique.
l) Obtenha a elipse de 95% de confiança para os parâmetros referentes aos efeitos de massa gorda e idade. Adicionalmente, obtenha intervalos de confiança simultâneos, com nível de confiança conjunto igual a 95%. Represente num único gráfico a região de confiança e os intervalos de confiança 95% simultâneos e individuais;
m) Obtenha o quadro da análise de variância e conduza o teste da hipótese nula de não significância do modelo de regressão;
n) Qual a estimativa para a variância dos erros aleatórios?
o) Obtenha o coeficiente de determinação e o coeficiente de determinação ajustado. Interprete-os;
p) Ajuste um novo modelo de regressão, desta vez desconsiderando a variável massa gorda. Compare os resultados dos dois ajustes. O que houve com o efeito da variável `imc`? Justifique.
q) Voltando ao modelo original, teste a significância conjunta das variáveis `imc` e `estudo`. De acordo com o resultado do teste, poderíamos excluir ambas as variáveis do modelo ao nível de 5% de significância?
r) Considere os seguintes perfis de atletas:

Perfil	Idade	IMC	Horas	Massa	Gasto
1	20	20	20	17	100
2	28	23	13	22	60
3	25	21	15	20	85

Para cada um dos perfis apresente um intervalo de confiança (95%) para o escore médio de desempenho e um intervalo de predição (95%) para o escore de um novo atleta com as características apresentadas;

s) Faça um gráfico de escores observados versus escores ajustados pelo modelo. Acrescente ao gráfico a reta identidade;
t) Faça um gráfico de resíduos versus escores ajustados. Como os resíduos estão distribuídos?
u) Padronize cada uma das variáveis explicativas e ajuste um novo modelo de regressão, substituindo as variáveis originais pelas padronizadas. Qual das variáveis apresenta maior efeito no escore de desempenho? Qual a interpretação das estimativas obtidas?