P2 de Análise Exploratória

Como ocorreu na P1, aqui há questões práticas a serem resolvidas com o uso do R:

- Ora com base em dados reais;
- Ora gerando os dados por simulação.

Espera-se que os códigos utilizados sejam anexados ao seu trabalho, para maior clareza e também para facilitar a correção.

Mas, desta vez, há também uma questão teórica.

A sua prova deverá ser entregue pelo Google Classroom até as 12:00 da 3ª feira 19/07/2022, em formato pdf.

Um aspecto que será bastante valorizado na correção é a clareza com a qual vocês explicarão o que foi feito.

IMPORTANTE: Esta prova é para ser resolvida em grupo. Então espera-se que cada aluno do grupo participe de todas as etapas do trabalho. Ou seja, a ideia não é incumbir cada membro do grupo de resolver somente uma parte da tarefa total. A riqueza do aprendizado consiste em a equipe como um todo discutir a solução a ser apresentada.

1. Quando se deseja aplicar a metodologia da Tabela de duas entradas (TWT) a um conjunto de dados, pode ser importante transformar previamente a variável resposta. A regra apresentada em Tukey-EDA, que determina a escolha dessa transformação (também está no Classroom em "Transformação da variável em tabela de duas entradas") é a seguinte:
   (a) Inicialmente, aplique o algoritmo TWT (Ver "Explicando Two-way table: Tabela de duas entradas" no Classroom) para estimar o valor central μ , os efeitos de linha α_i, os efeitos de coluna β_j e os resíduos e_{ij}, como se esse modelo se ajustasse sem necessidade de transformar a variável y.
   (b) Em seguida, calcule um novo termo chamado valor de comparação $c_{ij} = \frac{\alpha_i \beta_j}{\mu}$, para cada par (i,j).
   (c) Plote os m x n pares (c_{ij}, e_{ij}) em um gráfico bidimensional, onde e_{ij} é o erro do modelo original (sem transformação), ou seja, $e_{ij} = y_{ij} - (\mu + \alpha_i + \beta_j)$.
   (d) Suponha que, para alguma constante k, aos pares (c_{ij}, e_{ij}) desse plot possa ser ajustada uma reta passando pela origem, do tipo: $e_{ij} = k.\, c_{ij} = k\frac{\alpha_i \beta_j}{\mu}$.
   (e) Nessas condições a melhor forma de estimar o coeficiente angular k dessa reta é calcular a mediana do quociente $\frac{\text{resíduo}}{\text{valor de comparação}}$ para todos os pares (i,j) em que ele está definido.
   (f) Segundo Tukey-EDA, isso permite estabelecer uma regra para se escolher a melhor transformação a ser aplicada à variável resposta y, em função de k. Para que realmente se obtenha um modelo aditivo de tabela de duas entradas, aplique previamente à variável resposta y a transformação:

y^2,	se k = - 1;
$y^{3/2}$,	se k = - 0.5;
y (sem transformação),	se k = 0;
$\sqrt{y}$,	se k = 0.5;
log(y),	se k = 1;
$\frac{1}{\sqrt{y}}$,	se k = 1.5;
$\frac{1}{y}$,	se k = 2.

Mostre que, se existem constantes $\mu, \alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_n$ tais que

$$\log(y_{ij}) = \mu + \alpha_i + \beta_j + e_{ij}, \quad i = 1,2, \ldots, m, \quad j = 1,2,\ldots,n,$$

onde e_{ij} se comporta como um ruido branco (ou seja, uma variável aleatória centrada em zero), então teremos um coeficiente angular k = 1, quando no passo (d) acima for ajustada a reta, passando pela origem, aos m x n pontos de coordenadas (c_{ij}, e_{ij}).

2. Com base no dataset Wage do ISLR foi montada a tabela de contingência a seguir, que contem o número de trabalhadores em cada um dos 5 x 3 = 15 cruzamentos de nível educacional (de 1 a 5) com faixa etária em anos (até 34, de 35 a 49, 50 ou +):

Faixa etária ↓\ nível educ →	1	2	3	4	5
Até 34	79	276	215	176	80
35 a 49	115	417	275	314	202
50 ou +	74	278	160	195	144

A tabela a seguir contem a mediana dos salários dos trabalhadores em cada um desses 15 cruzamentos:

Faixa etária ↓\ nível educ →	1	2	3	4	5
Até 34	73.77574	81.28325	94.07271	104.92151	118.88436
35 a 49	86.69515	95.23071	109.83399	127.11574	148.41316
50 ou +	86.68249	98.59934	112.64397	123.08970	136.94859

Tomando esta última matriz como ponto de partida, use a metodologia proposta em Tukey-EDA para:

- Deduzir que transformação deve ser aplicada previamente à variável resposta wage (ou seja, salário), para que tenhamos um modelo de TWT puramente aditivo, capaz de explicar o salário como função de nível educacional e faixa etária simultaneamente.
- Com a variável resposta devidamente transformada, ajustar o modelo aos dados, calculando o valor central, os efeitos de linha e de coluna e os resíduos.
- Obter um gráfico com os efeitos-linha e os efeitos-coluna representados por retas a 45º (como vimos na teoria), que permita visualizar os resultados finais dessa análise, e incluindo eventualmente uma representação dos maiores resíduos em valor absoluto.
- Interpretar os resultados obtidos, extraindo as conclusões cabíveis.

3. Neste exercício, o objetivo é:
   - gerar datasets, por simulação, a partir de funções cuja expressão matemática é conhecida e;
   - através da função loess para a suavização de curvas, tentar obter uma boa aproximação para essa relação de dependência.

   a) Polinômio do 3º grau
      - Considere um polinômio do terceiro grau definido por

$$\text{estr}(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3,$$

$$\text{onde } a_0 = 500,\ a_1 = -30,\quad a_2 = 2,\ \ a_3 = -0.02$$

      - Gere por simulação n = 1000 observações e_i, i = 1,2,...,n, a partir da distribuição Normal com média 0 e desvio padrão 300. (Para isso, use a função rnorm do R)
      - Obtenha n pares (x_i,y_i), onde:

$$x_i = i/10, \qquad y_i = \text{estr}(x_i) + e_i, \qquad \text{para cada } i = 1,2,...,n$$

        Ou seja, $\text{estr}(x_i)$ é a parte estrutural da relação de dependência e e_i é um ruido perturbador.

      - Usando a função loess do R, e uma particular combinação dos seus parâmetros family (symmetric ou gaussian), degree (1 ou 2) e span (0.1, 0.3 ou 0.5), ajuste uma curva suave a esses n pares (x_i,y_i). Como são ao todo 12 = 2 x 2 x 3 combinações possíveis desses 3 parâmetros, você obterá ao todo 12 diagramas de dispersão, e em cada um deles será ajustada uma curva diferente.
      - Para cada uma dessas 12 opções, calcule $\sum_{i=1}^{100}[\text{fit}(i) - \text{estr}(i)]^2$: uma distância quadrática entre a curva ajustada via loess e a função estr(.). Qual das 12 combinações dos 3 parâmetros minimiza essa distância? Na sua opinião, por que essa particular combinação foi a vencedora? Era ou não previsível esse resultado? Por que?

   b) Exponencial negativa vezes Senóide
      - Considere uma dependência estrutural definida por

$$\text{estr}(x) = e^{-0.2x}\,\text{sen}(x),$$

      - Gere por simulação n = 1400 observações e_i, i = 1,2,...,n, a partir da distribuição Normal com média 0 e desvio padrão 0.1, que correspondem ao ruido. (Para isso, use a função rnorm do R)
      - Obtenha n pares (x_i,y_i), onde:

$$x_i = i/100, \qquad y_i = \text{estr}(x_i) + e_i, \qquad \text{para cada } i = 1,2,...,n$$

      - Usando a função loess do R, e uma particular combinação dos seus parâmetros family (symmetric ou gaussian), degree (1 ou 2) e span (0.1, 0.3 ou 0.5), ajuste uma curva a esses n pares (x_i,y_i). Como são ao todo 12 = 2 x 2 x 3 combinações possíveis desses 3 parâmetros, você obterá ao todo 12 diagramas de dispersão, e em cada um deles será ajustada uma curva diferente.

- Para cada uma dessas 12 opções, calcule $\sum_{j=1}^{140}[\text{fit}(j/10) - \text{estr}(j/10)]^2$: uma distância quadrática entre a curva ajustada via loess e a função estr(.). Qual das 12 combinações dos 3 parâmetros (family, degree, span) minimiza essa distância? Na sua opinião, por que essa particular combinação foi a vencedora? Era ou não previsível esse resultado? Por que?

4. Neste exercício será usado o dataset Auto do ISLR e o objetivo é obter o gráfico de uma função suave que descreva a forma como o "Tempo necessário para acelerar de 0 até 60 mph" (variável acceleration) depende da "potência do veículo" (variável horsepower). Teste todas as 12 possíveis combinações dos parâmetros family (symmetric ou gaussian), degree (1 ou 2) e span (0.1, 0.3 ou 0.5) na função loess e selecione entre elas aquela que lhe parece ser a mais adequada no caso. Justifique a sua escolha.