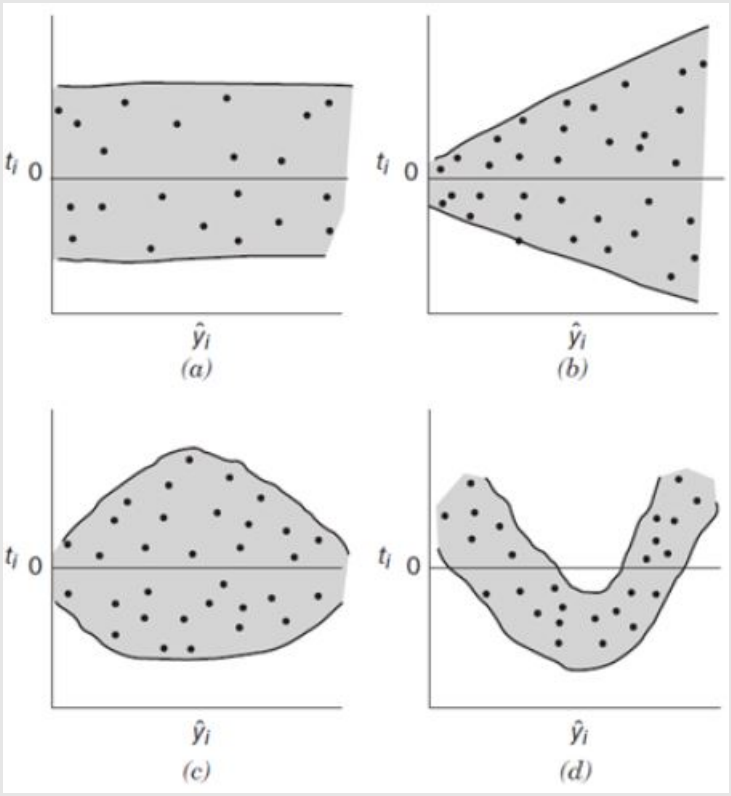
**Resolução de Atividades** **Avaliação 8**
Curso: Estatística para Ciência de Dados Perído: Segundo trimestre 2020
Tutores: Caio, Guilherme, Isis, Marina, Matheus e Tobias Criado em: 30 de Maio de 2020

Os 4 gráficos abaixo representam análises de resíduos em função dos valores preditos para verificação do ajuste de modelos de regressão linear.




Com base nessas ilustrações, responda as Questões de 1 a 5 abaixo.

**Questão 1**
O gráfico (a) exemplifica um padrão de comportamento desejado não revelando nenhum problema com as suposições do modelo.

Alternativas:

- Verdadeiro
- Falso

*Solução:* **Verdadeiro.**

Note que:

- o gráfico dos resíduos versus valores ajustados (valores preditos) é uma das técnicas para verificar as suposições feitas sobre os resíduos dos modelos;
- é possível detectar, por meio de tendências observadas nos pontos, indícios de heteroscedasticidade ou de que não existe uma relação linear entre as variáveis explicativas e a variável resposta;
- se os pontos estão aleatoriamente distribuídos em torno do zero, sem nenhum comportamento ou tendência, o comportamento é o esperado para distribuição dos erros e há indícios de que a variância dos resíduos é homoscedástica.

**Questão 2**
O gráfico (b) indica um problema de heterocedasticidade.

Alternativas:

- ○ Verdadeiro
- ○ Falso

*Solução:* **Verdadeiro.**

Note que:

- um gráfico em que os resíduos aumentam ou diminuem conforme os valores preditos indicam variância não constante dos erros (erros heterocedásticos; gráfico em forma de "funil");
  - um gráfico que apresenta uma tendência crescente sugere que a variação do erro aumenta com a variável independente;
  - um gráfico que revela uma tendência decrescente indica que a variação do erro diminui com a variável independente;
  - em nenhum dos casos as distribuições são padrões de variação constante dos resíduos;
- nessa situação, sugere-se transformar a variável resposta ou utilizar algum modelo linear generalizado.

**Questão 3**
O gráfico (c) representa um problema de associação não linear entre a resposta e os preditores.

Alternativas:

- ○ Verdadeiro
- ○ Falso

*Solução:* **Falso.**

Note que:

- a suposição de homocedasticidade é provavelmente violada se:
  - se os resíduos aumentam ou diminuem com os valores ajustados;
  - se os pontos formam uma curva ao redor de zero e não estão dispostos aleatoriamente.

**Questão 4**
O gráfico (d) indica desvios da normalidade da variável resposta.

Alternativas:

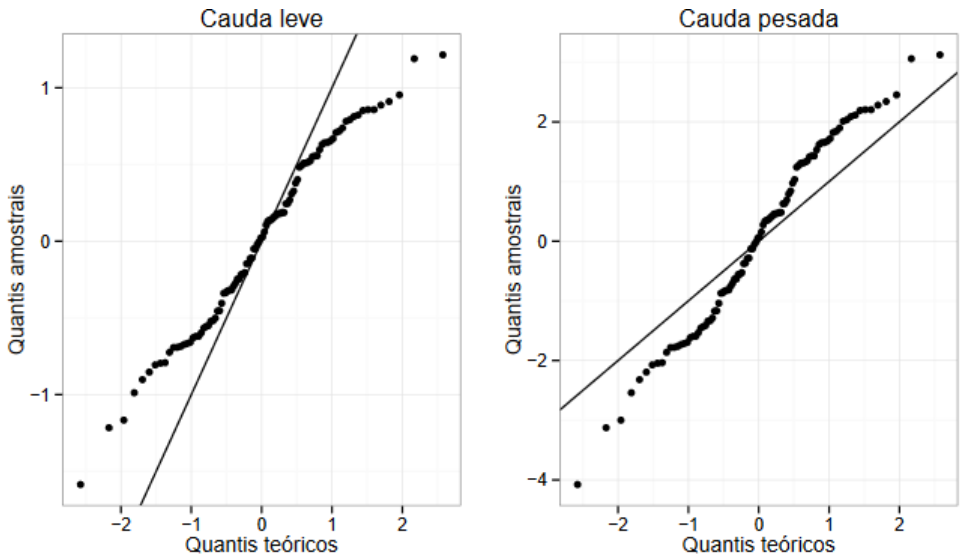
- ○ Verdadeiro
- ○ Falso

*Solução:* **Falso.**

Note que:

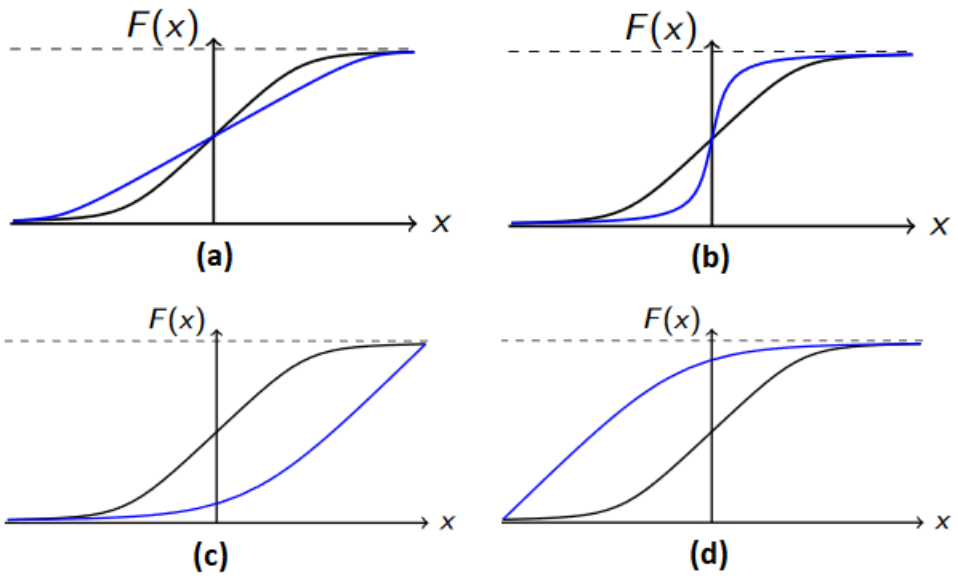
- o gráfico que permite detectar a (não) normalidade de resíduos é um gráfico de pontos de quantis amostrais dos resíduos versus quantis teóricos da distribuição normal-padrão;
- para um modelo bem ajustado, o gráfico de probabilidade normal dos resíduos deve seguir aproximadamente uma linha reta[1].

Figura 1: Exemplos de desvios de normalidade.




Fonte: https://www.ime.usp.br/~selian/mae328/AnaliseResiduos.pdf

Figura 2: Exemplos de desvios de normalidade. (a) Cauda mais pesada que a normal. (b) Cauda mais leve que a normal. (c) Assimetria positiva. (d) Assimetria negativa.




Fonte: Adaptada de https://www.ime.usp.br/~selian/mae328/AnaliseResiduos.pdf

[1]Veja exemplos em: https://online.stat.psu.edu/stat501/lesson/4/4.6.

**Questão 5**
O gráfico (d) indica não linearidade, talvez devido à ausência de algum preditor importante no modelo.

Alternativas:

- Verdadeiro
- Falso

*Solução:* **Verdadeiro.**

Note que:

- se os pontos do gráfico dos resíduos versus valores ajustados (valores preditos) apresentam uma tendência não linear, pode ser necessário incorporar novas variáveis explicativas ao modelo, considerar alguma transformação nas variáveis explicativas e/ou preditoras, ou utilizar algum modelo de regressão não linear;
- no caso em que a distribuição dos resíduos apresenta uma forma de parábola, pode ser um indicativo que termos de segundo grau sejam necessários.

**Questão 6**
Com o conjunto de dados do estudo do Prestígio das ocupações profissionais usados na aula da semana 7 e ajuste os modelos de regressão linear múltipla com as seguintes variáveis explicativas (considerando Prestige como a variável resposta):

1. Education e Income
2. Education e Income e respectiva interação
3. Education e Income, ambas padronizadas
4. Education e Income, ambas padronizadas, e respectiva interação

Ajuste os modelos padronizados com e depois sem intercepto (basta incluir $-1$ no modelo, junto com os preditores, na função usada em aulas passadas). Calcule o VIF de cada preditor com o comando:

```
statsmodels.stats.outliers_influence.variance_inflation_factor
```

Assinale a alternativa INCORRETA.

Alternativas:

(a) Os coeficientes de determinação dos modelos 2 e 4 (com intercepto) são exatamente iguais.

(b) A padronização das variáveis resolveu o problema da multicolinearidade, quando este existia.

(c) O modelo sem interação com os preditores não padronizados não apresentou problema de multicolinearidade. Este problema apareceu quando o termo de interação foi inserido.

(d) O modelo de regressão padronizado deve ser ajustado sem intercepto quando as médias de todas as variáveis explicativas no modelo são iguais a 0, fazendo com que a função predita passe pela origem.

(e) No modelo de regressão sem interação, com as variáveis padronizadas e sem intercepto, pode-se dizer que a cada ano de estudo adicional, o prestígio da ocupação profissional aumenta em $0,69$ unidades, em média.

*Solução:* **Alternativa e.**

---

**Questão 7**
Use o código abaixo para simular os dados que usamos do trabalho do Sildenafil e encontrar um estimador bayesiano da proporção, na população, de homens com pelo menos 60% de tentativas bem sucedidas.

```
# Talvez tenha que instalar pelo Anaconda
import pymc3

# Simula o conjunto de dados do Sildenafil
data = np.concatenate((np.repeat(1, 183),np.repeat(0,379-183)),axis=0)

# Cria o modelo com priori Beta(alpha,beta) e verossimilhança Bernoulli(p)
def creat_model_pymc3(data):
  with pymc3.Model() as model:
    p = pymc3.Beta('theta', alpha=1, beta=1)
    bernoulli = pymc3.Bernoulli('bernoulli', p=p, observed=data)
  return model

model = create_model_pymc3(data)

# Estima a probabilidade de sucesso da Bernoulli por MAP
map_estimate = pymc3.find_MAP(model=model)
map_estimate
```

Assinale a alternativa CORRETA.

Alternativas:

(a) Se usarmos exatamente a mesma priori usada na vídeo aula (Beta(2,3), ao invés da Beta(1,1)), os estimadores MAP e EAP são iguais.

(b) A estimativa do MAP é mais parecida com a de MV do que com a da EAP.

(c) Os estimadores de MV e MAP maximizam exatamente a mesma função, independentemente dos parâmetros da priori.

(d) A estimativa do MAP é mais parecida com a da EAP do que com a de MV.

(e) A estimativa do MAP foi bem diferente das de MV e EAP.

*Solução:* **Alternativa b.**

Note que:

- EMV:
  - o EMV é o valor que maximiza a verossimilhança;
  - a função de verossimilhança da Bernoulli é dada por:

$$L(y|\theta) = \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{N-\sum_{i=1}^{n} y_i};$$

  - o EMV para $\theta$, denotado por $\hat{\theta}_{MV}$, é o valor que maximiza a (log-)verossimilhança. Assim, a

função de log-verossimilhança é dada por:

$$l(y|\theta) = N - \sum_{i=1}^{n} y_i \log\theta + (N - \sum_{i=1}^{n} y_i)\log(1-\theta);$$

- diferenciando-se a expressão em relação ao parâmetro e igualando-se a zero, tem-se:

$$\frac{\partial l}{\partial \theta} = \frac{\sum_{i=1}^{n} y_i}{\theta} - \frac{N - \sum_{i=1}^{n} y_i}{1-\theta} = 0;$$

- o EMV é, portanto:

$$\hat{\theta}_{MV} = \frac{\sum_{i=1}^{n} y_i}{N}.$$

- MAP:
  - o máximo a posteriori é o valor que maximiza a posteriori;
  - quando a priori é uniforme, o MAP é igual ao EMV;
  - usando uma verossimilhança Bernoulli e uma priori Beta, o MAP para $\theta$, denotado por $\hat{\theta}_{MAP}$, é o valor que maximiza a posteriori. Assim, a log-posteriori é dada por:

$$l(p(\theta|y)) \propto \left(\sum_{i=1}^{n} y_i + \alpha - 1\right)\log\theta + \left(N - \sum_{i=1}^{n} y_i + \beta - 1\right)\log(1-\theta);$$

  - diferenciando-se a expressão em relação ao parâmetro $\theta$ e igualando-se a zero, tem-se:

$$\frac{\partial l}{\partial \theta} = \frac{\sum_{i=1}^{n} y_i}{\theta} - \frac{N - \sum_{i=1}^{n} y_i}{1-\theta} + \frac{\alpha - 1}{\theta} - \frac{\beta - 1}{1-\theta} = 0;$$

  - o MAP é, portanto:

$$\hat{\theta}_{MAP} = \frac{\sum_{i=1}^{n} y_i + \alpha - 1}{N + \beta - 1 + \alpha - 1}.$$

- EAP:
  - a esperança da posteriori $Beta(\alpha + \sum_{i=1}^{n} y_i,\ \beta + N - \sum_{i=1}^{n} y_i)$ é dada por:

$$\hat{\theta}_{EAP} = \frac{\alpha + \sum_{i=1}^{n} y_i}{\alpha + \sum_{i=1}^{n} y_i + \beta + N - \sum_{i=1}^{n} y_i} = \frac{\alpha + \sum_{i=1}^{n} y_i}{\alpha + \beta + N}.$$

Assim, ao usar $Beta(1,1)$ como priori, com $N = 379$ e $\sum_{i=1}^{n} y_i = 183$, obtém-se:

$$\hat{\theta}_{MV} = \frac{183}{379} \approx 0,4828;$$

$$\hat{\theta}_{MAP} = \frac{183 + 1 - 1}{379 + 1 - 1 + 1 - 1} \approx 0,4828;$$

$$\hat{\theta}_{EAP} = \frac{1 + 183}{1 + 1 + 379} \approx 0,4829.$$

Ao usar $Beta(2,3)$ como priori, obtém-se:

$$\hat{\theta}_{MAP} = \frac{183 + 2 - 1}{379 + 3 - 1 + 2 - 1} \approx 0,4817;$$

$$\hat{\theta}_{EAP} = \frac{2 + 183}{2 + 3 + 379} \approx 0,4818.$$

---

**Questão 8**
Complemente o código do exercício anterior com:

```
# Usa MCMC para gerar observações da posteriori (método de cálculo da posteriori por
    simulação)
with model:
  trace = pymc3.sample(1000, tune=1000, cores=1)
# Faz o gráfico da posteriori, calcula sua média e o intervalo de credibilidade
pymc3.plot_posterior(trace);
```

Qual a afirmativa correta?

Alternativas:

(a) Os resultados pela simulação de MCMC (com 1000 observações geradas, como o código do enunciado, pelo menos) são similares aos obtidos teoricamente (apresentados em aula).

(b) O intervalo de credibilidade calculado tem 99% de probabilidade de conter o verdadeiro valor da probabilidade de sucesso, $p$.

(c) Se gerarmos poucas observações da posteriori simuladas por MCMC (10, por exemplo, ao invés de 1000), os resultados ficam bem piores. Faça isso com a seguinte modificação:

```
trace = pymc3.sample(10, tune=1000, cores=1)
```

(d) A média da posteriori calculada no MCMC com o código acima é o estimador de MV de $p$.

(e) Na inferência frequentista, o parâmetro $p$ da Bernoulli segue distribuição $Beta(1, 1)$, ou seja, Uniforme.

*Solução:* **Alternativa a.**

**Alternativa c.**

---