Análise de Regressão

Prof. Dr. Juliano Bortolini

Bacharelado em Estatística - UFMT

Período letivo: 2024/2

Avaliação 1 (arguição oral)

Bloco 1: Conceitos Fundamentais e Interpretações

1. O que é covariância e como interpretamos seu sinal?
A covariância mede a tendência de duas variáveis variarem juntas. Sinal positivo indica que ambas aumentam ou diminuem juntas; sinal negativo indica que uma aumenta enquanto a outra diminui.

2. Qual a principal limitação da covariância?
A covariância depende da escala das variáveis, dificultando a comparação entre diferentes conjuntos de dados.

3. Explique a diferença entre covariância e correlação.
A correlação é uma versão padronizada da covariância, variando entre -1 e 1, permitindo interpretação comparativa independente das unidades.

4. Como é definido o coeficiente de correlação de Pearson?
É a razão entre a covariância de duas variáveis e o produto dos seus desvios padrão:
$\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$.

5. O que é a correlação de Spearman e quando usá-la?
É a correlação de Pearson aplicada aos postos dos dados. É usada quando a relação entre as variáveis é monotônica mas não necessariamente linear, ou quando os dados não são normais.

6. Qual a interpretação geométrica do coeficiente de correlação?
É o cosseno do ângulo entre os vetores centrados das variáveis X e Y.

7. O que significa correlação perfeita positiva e perfeita negativa?
Correlação perfeita positiva ($\rho = 1$) indica que uma variável aumenta exatamente na mesma

proporção que a outra. Correlação perfeita negativa ($\rho = -1$) indica que uma variável aumenta enquanto a outra diminui exatamente na mesma proporção.

8. Qual a relação entre correlação e independência?
Independência implica correlação zero, mas correlação zero não implica independência.

9. O que é a função de distribuição acumulada empírica?
É uma função que, para cada valor x, retorna a proporção de observações na amostra menores ou iguais a x.

10. Enuncie o Teorema de Glivenko-Cantelli.
A função de distribuição empírica converge uniformemente, com probabilidade 1, para a função de distribuição verdadeira da população quando $n \to \infty$.

11. Quais são as propriedades desejáveis de um estimador?
Não-viesado, consistente, eficiente (menor variância) e suficiente.

12. Quais são os estimadores da covariância e correlação?
Covariância: $\widehat{\text{Cov}}(X, Y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$
Correlação: $\hat{\rho} = \frac{\widehat{\text{Cov}}(X,Y)}{s_X s_Y}$

Bloco 2: Modelagem, Estimação e Inferência

13. Como é definido o modelo de regressão linear simples?
$Y = \beta_0 + \beta_1 X + \varepsilon$, onde ε é o erro aleatório com média zero e variância constante.

14. Como é estimado o modelo por mínimos quadrados?
Minimizando a soma dos quadrados dos resíduos:
$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

15. Explique como calcular a Soma dos Quadrados dos Resíduos (SQRes) e sua importância.
$SQRes = \sum(y_i - \hat{y}_i)^2$. Mede a variabilidade não explicada pelo modelo. É usada na estimativa de σ^2.

16. Como é estimado σ^2 no modelo linear?
$\hat{\sigma}^2 = \frac{SQRes}{n-2}$

17. O que diz o Teorema de Gauss-Markov?
Sob os pressupostos do modelo linear, os estimadores de mínimos quadrados são os melhores lineares não-viesados (BLUE).

18. O que é homocedasticidade?
É a suposição de que a variância dos erros é constante para todos os valores de X.

19. O que representa o coeficiente de determinação R^2?
Representa a proporção da variância total de Y explicada linearmente pelo modelo.

20. Apresente a estatística t usada para testar $H_0 : \beta_1 = 0$.
$t = \frac{\hat{\beta}_1}{EP(\hat{\beta}_1)}$, onde EP é o erro padrão do estimador.

21. Qual é a implicação prática de rejeitar $H_0 : \beta_1 = 0$?
Há evidência de que X tem um efeito linear significativo sobre Y.

22. Qual a relação entre o teste t e o teste F na regressão simples?
Em regressão simples, $F = t^2$.

23. Quando usar o teste F na análise de regressão?
Para testar se o modelo explica significativamente a variabilidade de Y, comparando um modelo com e sem variáveis explicativas.

24. O que indica um intervalo de confiança para β_1 que contém o valor zero?
Que, ao nível de confiança adotado, não há evidência suficiente para afirmar que $\beta_1 \neq 0$.

25. Como interpretar um intervalo de confiança para β_1?
Intervalo que, com um nível de confiança (ex: 95%), contém o valor verdadeiro de β_1.

26. Qual a diferença entre predição da média e predição de nova observação?
A predição da média estima o valor médio de Y para um dado X; a predição de nova observação inclui a variabilidade do erro aleatório.

27. O que representa o intervalo de predição para nova observação?
Uma faixa onde se espera que uma nova observação de Y, para dado X, caia com certa confiança (ex: 95%).

28. Como interpretar um valor-p alto no teste de correlação?
Indica que não há evidência suficiente para rejeitar a hipótese de correlação nula entre as variáveis.

Bloco 3: Análise Crítica e Discussão de Modelos

29. Quais são os pressupostos do modelo de regressão linear múltipla?
Linearidade, independência dos erros, homocedasticidade, normalidade dos erros e ausência de multicolinearidade entre as variáveis explicativas.

30. Como verificar a adequação do modelo ajustado?
Por meio de análise gráfica dos resíduos, testes de normalidade, testes de heterocedasticidade e avaliação de medidas como R^2 e R^2 ajustado.

31. O que é multicolinearidade e como detectá-la?
É a presença de correlação alta entre variáveis explicativas. Pode ser detectada por meio de matriz de correlação, FIV (fator de inflação da variância) ou análise de autovalores da matriz $X'X$.

32. O que é o R^2 ajustado?
É uma medida que ajusta o R^2 levando em conta o número de variáveis explicativas e o tamanho da amostra. Penaliza o acréscimo de variáveis que não melhoram o modelo.

33. O que são efeitos de interação?
São efeitos em que a influência de uma variável explicativa sobre a resposta depende do nível de outra variável explicativa.

34. Como comparar modelos completo e reduzido usando o teste F?
Usa-se a fórmula:

$$F = \frac{(SQ_{res\ reduzido} - SQ_{res\ completo})/(df_{reduzido} - df_{completo})}{SQ_{res\ completo}/df_{completo}}$$

para verificar se o modelo completo explica significativamente mais que o modelo reduzido.

35. O que é heterocedasticidade e como ela afeta os testes?
É a não-constância da variância dos erros. Pode tornar inválidos os testes estatísticos e os intervalos de confiança, pois os erros padrão estarão incorretos.

36. Como detectar pontos influentes em um modelo?
Usando métricas como distância de Cook, alavancagem (leverage), resíduos padronizados e studentizados.

37. Qual a importância da análise dos resíduos?
Permite verificar os pressupostos do modelo, detectar outliers, influências e possíveis problemas de especificação do modelo.

38. O que são coeficientes padronizados e para que servem?
São estimativas obtidas após padronizar as variáveis. Permitem comparar a importância relativa de cada variável explicativa no modelo.

39. Como interpretar um intervalo de confiança que contém o valor zero para β_j?
Indica que, ao nível de confiança adotado, não há evidência suficiente para afirmar que o coeficiente β_j é diferente de zero.

40. Em que situações o modelo linear não deve ser utilizado?
Quando os pressupostos não são atendidos (ex: não-linearidade, heterocedasticidade, outliers fortes, dependência serial), ou quando a relação entre as variáveis não é bem representada por uma reta.

41. Quando é justificável usar regressão pela origem?
Quando há justificativa teórica de que a variável resposta deve ser zero quando a variável explicativa é zero, e essa hipótese pode ser sustentada com base no contexto do estudo.

42. Qual é a interpretação de um coeficiente angular $\hat{\beta}_1$ negativo?
Indica que, à medida que X aumenta, a variável Y tende a diminuir, ou seja, existe uma relação linear decrescente entre as variáveis.

43. Em que situações o modelo de regressão linear simples pode ser inadequado, mesmo com R^2 alto?
Quando há violação dos pressupostos, presença de variáveis omitidas relevantes, ou quando a relação é espúria. Um alto R^2 não garante validade inferencial.