Linearidade

Alejandro Yopasá-Arenas, PhD

17/12/2020

## Resultados experimentais

Tabela de resultados

|  |  |  |
| --- | --- | --- |
| No. Observação | Concentração | Resposta |
| 1 | 1,998 | 91287 |
| 2 | 1,998 | 92635 |
| 3 | 1,998 | 87717 |
| 4 | 3,996 | 181620 |
| 5 | 3,996 | 183739 |
| 6 | 3,996 | 175633 |
| 7 | 5,994 | 288423 |
| 8 | 5,994 | 276837 |
| 9 | 5,994 | 271491 |
| 10 | 7,992 | 371431 |
| 11 | 7,992 | 378810 |
| 12 | 7,992 | 361988 |
| 13 | 8,991 | 445930 |
| 14 | 8,991 | 425366 |
| 15 | 8,991 | 440826 |
| 16 | 9,990 | 470969 |
| 17 | 9,990 | 453986 |
| 18 | 9,990 | 502596 |
| 19 | 10,989 | 543081 |
| 20 | 10,989 | 480102 |
| 21 | 10,989 | 529029 |
| 22 | 11,988 | 602909 |
| 23 | 11,988 | 523646 |
| 24 | 11,988 | 586989 |

## Ajuste do modelo linear

O nosso modelo é representado por:

Em que:

* representa o sinal analítico (área, absorbância. etc…);
* representa a concentração;
* representa o coeficiente linear ou intercepto;
* representa o coeficiente angular;
* representa o erro experimental;
* representa o número de réplicas do ponto de concentração;
* representa o número de pontos ou níveis.

O ajuste de modelos lineares é feito no R fácilmente com a função lm(). A função faz a regressão e a análise da variância e covariância. Obtemos o seguinte modelo

Tabela de coeficientes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimativa | Desvio Padrão | Valor t | Pr(>|t|) |
| Intercepto | -9443 | 10136 | -0,9316 | 0,3616 |
| Concentração | 48403 | 1206 | 40,1248 | 0,0000 |

Temos que os dados são ajustados por uma curva linear de intercepto= -9442,9682 e coeficiente angular= 48402,5767.

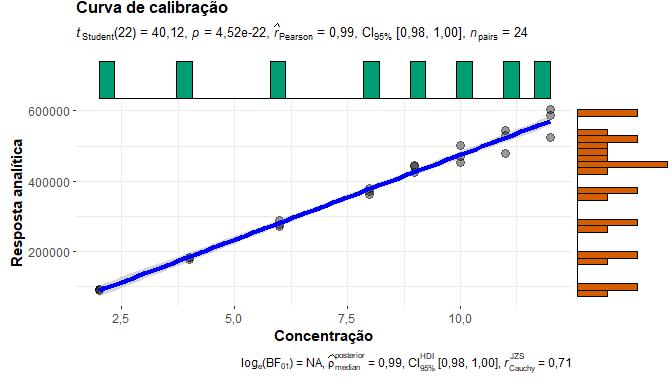


Tabela de resultados do ajuste

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. Observação | Concentração | Resposta | Resposta ajustada | Resíduo |
| 1 | 1,998 | 91287 | 87265 | 4021,9 |
| 2 | 1,998 | 92635 | 87265 | 5369,1 |
| 3 | 1,998 | 87717 | 87265 | 451,9 |
| 4 | 3,996 | 181620 | 183969 | -2348,8 |
| 5 | 3,996 | 183739 | 183969 | -229,7 |
| 6 | 3,996 | 175633 | 183969 | -8335,4 |
| 7 | 5,994 | 288423 | 280677 | 7745,4 |
| 8 | 5,994 | 276837 | 280677 | -3840,2 |
| 9 | 5,994 | 271491 | 280677 | -9185,8 |
| 10 | 7,992 | 371431 | 377381 | -5949,4 |
| 11 | 7,992 | 378810 | 377381 | 1429,5 |
| 12 | 7,992 | 361988 | 377381 | -15393,0 |
| 13 | 8,991 | 445930 | 425735 | 20195,4 |
| 14 | 8,991 | 425366 | 425735 | -368,6 |
| 15 | 8,991 | 440826 | 425735 | 15090,7 |
| 16 | 9,990 | 470969 | 474089 | -3119,8 |
| 17 | 9,990 | 453986 | 474089 | -20102,8 |
| 18 | 9,990 | 502596 | 474089 | 28507,0 |
| 19 | 10,989 | 543081 | 522438 | 20642,9 |
| 20 | 10,989 | 480102 | 522438 | -42336,7 |
| 21 | 10,989 | 529029 | 522438 | 6590,3 |
| 22 | 11,988 | 602909 | 570793 | 32116,8 |
| 23 | 11,988 | 523646 | 570793 | -47147,0 |
| 24 | 11,988 | 586989 | 570793 | 16196,1 |

## Análise dos coeficientes do modelo linear:

### Intercepto

Com os resultados da tabela acima, além das estimativas dos parâmetros, podemos avaliar a significância dos parâmetros por meio do teste . O valor t é calculado dividindo o valor do coeficiente pelo erro padrão. Em relação ao parâmetro intercepto, temos que as hipóteses são dadas por:

: Intercepto é igual a zero ()

: Intercepto é diferente de zero ()

O *valor* para o intercepto é dado por:

no qual é o desvio padrão do intercepto dado nos resultados da tabela acima.

O valor crítico da distribuição ao nível de significância do 5% () e graus de liberdade é dado por . Como o valor associado a esse teste é maior que 0,05 aceitamos e concluimos que o intercepto é significativamente igual a zero ao nível de significância do 5%

### Coeficiente angular

Em relação ao coeficiente angular, temos que as hipóteses são:

: Coeficiente angular é igual a zero ()

: Coeficiente angular é diferente de zero ()

O estatístico T do teste é dado por:

O valor crítico da distribuição ao nível de significância do 5% () e graus de liberdade é dado por . Como o valor associado a esse teste () é menor que 0,05 rejeitamos e concluimos que o coeficiente angular é significativamente diferente de zero ao nível de significância do 5%

### Anova no método dos mínimos quadrados ordinários (MMQO)

Avaliamos também a significância do modelo por meio do teste F da ANOVA. Vale ressaltar que temos um modelo de regressão simples, desta forma o teste F da ANOVA é equivalente ao teste . A seguir, temos a Tabela da ANOVA.

Tabela de anova

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Graus de liberdade | Soma dos quadrados | Quadrado médio | valor-F | Pr(>F) |
| Concentração | 1 | 599694955424 | 599694955424 | 1610 | 0 |
| Resíduos | 22 | 8194586069 | 372481185 | NA | NA |

Para testarmos a significância do coeficiente angular do modelo com o teste F da ANOVA, apresentamos as seguintes hipóteses:

: Coeficiente angular é igual a zero.

: Coeficiente angular é diferente de zero.

A estatística de teste é dada pela divisão do quadrado médio da regressão (QMR) pelo quadrado médio dos resíduos ou erros (QME):

Onde é a soma dos quadrados da regressão, é a soma dos quadrados dos erros ou resíduos, é o número de parâmetros do modelo, é o número total de pontos, é a média aritmética dos valores de y no centroide, é um valor individual de y obtido experimentalmente em um determinado ponto (iésimo ponto) e é um valor individual de y calculado pelo modelo (equação) em um determinado ponto (i-ésimo ponto).

A região crítica para o teste F é dada por . Como a estatística observada é *maior* que o quantil da distribuição para a determinação da região crítica ( > ) e o valor associado a esse teste () é menor que 0,05 rejeitamos e concluimos que o coeficiente angular é significativamente diferente de zero ao nível de significância do 5%

## Análise de Resíduos

A tabela a seguir, apresenta a análise exploratória dos resíduos.

Tabela dos resíduos

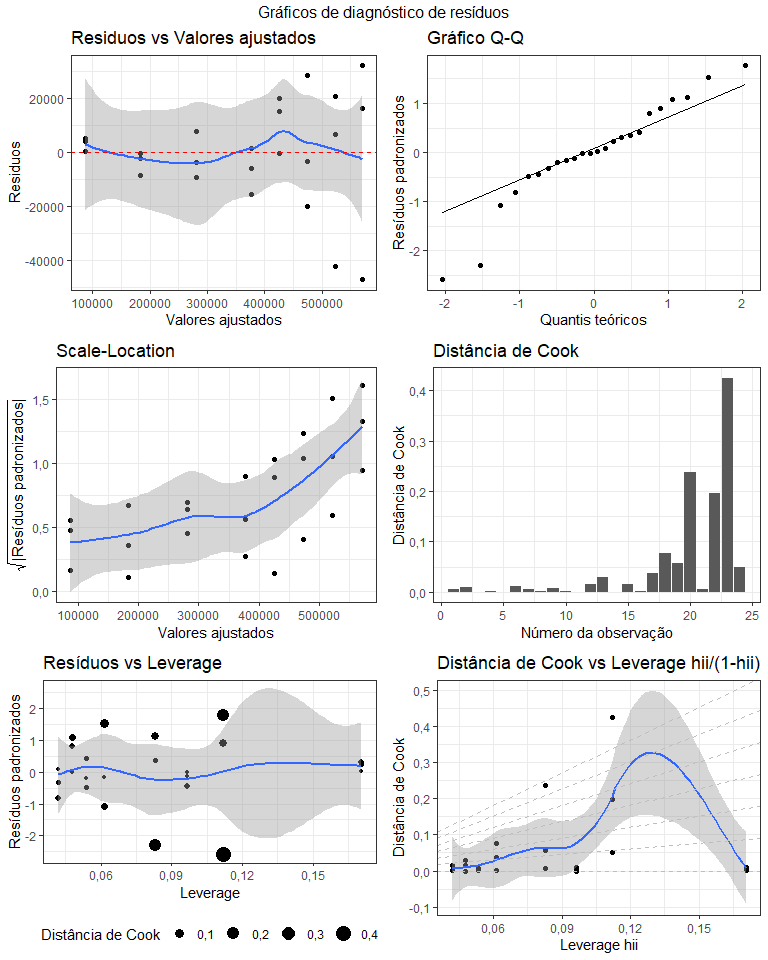
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mínimo | 1Q | Mediana | Média | 3Q | Máximo |
| -47147 | -6546 | 111,1 | 0 | 9582 | 32117 |

Observando a tabela acima, podemos estudar se os valores de mínimo e máximo, em módulo apresentam ou não uma diferença notável, assim como a mediana e a média, o que nos dá indícios de que a distribuição dos resíduos é simétrica.

O coeficiente de correlação de Pearson mede o grau de proporcionalidade entre a variável explicativa (concentração) e a varíavel resposta (área). Temos que o coeficiente de determinação R2 é dado pela divisão da soma de quadrados da regressão (SQR) pela soma de quadrados total (SQT=SQR+SQE):

Logo o critério da RDC em relação ao coeficiente é satisfeto, visto que 0,9932 está acima do valor especificado pela agência reguladora (0,990). Note que o coeficiente de determinação representa a relação sinal/ruído, em que SQR está relacionada ao sinal analítico e o ruído está relacionada ao SQT.

### Gráficos de diagnóstico



* No gráfico de residuos vs valores ajustados observamos se existe comportamento não linear, dados dispersos homogenamente no redor de uma linha horizontal sem padrão aparente é um indicador de relação linear.
* O gráfico Q-Q indica se os resíduos estão distribuidos normalmente. Se os resíduos seguem a linha reta é um indício de que a suposição de normalidade para os erros experimentais é satisfeita.
* O gráfico de *scale-location* indica se os resíduos estão distribuídos igualmente ao longo dos intervalos dos preditores, uma linea horizontal com pontos dispersos aleatóriamente na volta é um indicativo de que a suposição de homocedasticidade é satisfeita.
* O gráfico de distância de Cook indica se existe algum punto influente, alguns indicam uma distância maior do que 1 para considerar um punto influente. [Outras referências](https://stats.stackexchange.com/questions/22161/how-to-read-cooks-distance-plots#:~:text=Cook%27s%20distance%20can%20be%20contrasted,dropped%20from%20the%20data%20set.) indicam como limite um valor de 4/n ou 4/(n-k-1), onde n=número de observações e k é o número de variáveis explicativas.
* No gráfico de resíduos vs *leverage* observamos a dispersão dos resíduos em função do *leverage*. O *leverage* é uma medida de quão distantes os valores das variáveis independentes de uma observação estão daqueles das outras observações. O gráfico é utilizado principalmente para detectar heteroscedasticidade e não linearidade. A propagação de resíduos padronizados não deve mudar em função do leverage numa situação de homocedasticidade (variância constante).
* O gráfico de distância de cook vs leverage indica se os pontos com alto *leverage* podem ter influência: ou seja, se excluí-los mudaria muito o modelo. Para isso, podemos observar a distância de Cook, que mede o efeito da exclusão de um ponto no vetor de parâmetro combinado. Uma diretriz aproximada, para tamanhos de amostra grandes, é considerar os valores de distância de Cook acima de 1 para indicar pontos altamente influentes e valores de *leverage* maiores que 2 vezes o número de preditores dividido pelo tamanho da amostra para indicar observações de alto *leverage* (no nosso caso esté limite sería ). Observações de alto *leverage* são aquelas que possuem valores preditores muito distantes de suas médias, o que pode influenciar muito o modelo ajustado.

Para validar as indicações sugeridas a partir da análise gráfica, vamos verificar as hipótese levantadas por meio dos testes estatísticos:

### Teste de normalidade dos resíduos

A seguir, avaliamos a normalidade dos erros experimentais por meio do teste de Shapiro-Wilk, no qual as hipóteses são:

: A distribuição dos erros experimentais é normal

: A distribuição dos erros experimentais não é normal

Teste de normalidade de Shapiro-Wilk

|  |  |
| --- | --- |
| Estatística w | p-valor |
| 0,9363 | 0,1346 |

Como resultado temos um valor estatístico w=0,9363 e que o valor associado a esse teste (0,1346) é maior que 0,05, sendo assim, aceitamos a hipótese nula de normalidade dos erros experimentais ao nível de significância de 5%. Note que o resultado do teste de Shapiro-Wilk está em conformidade com a análise gráfica do gráfico Q-Q.

### Teste de homocedasticidade

A seguir, analisamos a homoscedasticidade por meio do teste de Breusch-Pagan, no qual as hipóteses são:

: As variâncias são iguais.

: Pelo menos uma variância difere.

Teste de homoscedasticidade de Breusch-Pagan

|  |  |
| --- | --- |
| Estatística BP | p-valor |
| 7,569 | 0,0059 |

Como resultado temos um valor estatístico BP=7,5689 e que o valor associado a esse teste (0,0059) é menor que 0,05, sendo assim, rejeitamos a hipótese nula de igualdade das variâncias ao nível de significância de 5%. Note que o resultado do teste está em conformidade com a análise gráfica dos resíduos X valores ajustados. Logo, temos um modelo heterocedástico.

Conforme o [portal action](http://www.portalaction.com.br/validacao-de-metodologia-analitica/exemplo-linearidade-hplc) O teste de Breusch-Pagan é o que melhor se adequa neste caso, visto que assumimos a suposição de normalidade para os erros experimentais. Os teste de Cochran e de Brown-Forsythe não se adequam ao nosso objetivo pois necessitam de grupos e, como os dados do exemplo foram coletados de forma independente, os testes em questão não poderiam ser realizados. Já o teste de Goldfeld-Quandt tem como limitação a exigência de amostras relativamente grandes.

### Teste de autocorrelação

Para confirmar que não há dependência das observações, isto é, que não temos sequências de pontos decrescentes ou crescentes vamos aplicar o teste de Durbin-Watson. As hipóteses do teste são:

: As observações são independentes.

: As observações não são independentes.

Teste de autocorrelação de Durbin-Watson

|  |  |
| --- | --- |
| Estatística DW | p-valor |
| 2,825 | 0,9731 |

Como resultado temos um valor estatístico DW=2,8255 e que o valor associado a esse teste (0,9731) é maior que 0,05, sendo assim, aceitamos a hipótese nula de independência das observações ao nível de significância de 5%.

## Conclusões

Critérios da RDC 166:

* Coeficiente linear não significativo ao nível de significância de 5%;
* Coeficiente angular significativo ao nível de significância de 5%;
* Normalidade dos erros experimentais;
* Diferências nas variâncias ao nível de significância de 5% (modelo heterocedástico);
* As observações são independentes.