Universidade Estadual de Campinas - UNICAMP
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística

Validari - Linearidade

Ana Flávia Polisel 157672
Bruna Nascimento Marques 135143
Melissa Pelermo 140830
Welligton Takao Tanaka 093259

Setembro/2017

1 Introdução

A Validari é uma *start up* criada com o intuito de auxiliar os laboratórios farmacêuticos na validação de métodos analíticos. A proposta da empresa é criar um *software* online, que seja capaz de validar métodos analíticos de forma ágil e correta estatisticamente.

Resultados analíticos não confiáveis podem conduzir a decisões erradas e prejuízos financeiros irreparáveis para as indústrias, por isso a importância de validar de forma correta os métodos analíticos. Essa validação é dividida em etapas, existem vários critérios que precisam ser satisfeitos para que o método seja validado. Dentre esses critérios está a Linearidade, que corresponde à capacidade do método em fornecer resultados diretamente proporcionais à concentração do analito em amostras, dentro de uma determinada faixa de concentração. Todos os cálculos para a avaliação da Linearidade devem ser realizados a partir dos dados de concentrações reais e as respostas analíticas individuais (área).

A Agência Nacional de Vigilância Sanitária (Anvisa) possui normas com sugestões de análises para cada critério da validação de métodos analíticos. Porém, algumas sugestões são muito subjetivas, dificultando o trabalho de validação. Por isso, o objetivo do nosso projeto é avaliar e revisar a metodologia estatística usada para validação da Linearidade, explicando de forma clara o passo a passo da metodologia mais adequada. Todas as análises, gráficos e tabelas foram feitos no *software* R.

2 Material

Para exemplificar e facilitar o entendimento da análise descritiva e da metodologia estatística adequada, vamos utilizar um conjunto de dados simulados disponível em: http://www.portalaction.com.br/validacao-de-metodologia-analitica/112-linearidade (acesso em 24/10/2017). Temos duas variáveis: a concentração de um analito em uma amostra e a área (resultado liberado pelo equipamento). Na Tabela 1 a seguir, temos as 12 primeiras observações do banco de dados utilizado, de um total de 30:

Tabela 1: Tabela mostra os 12 primeiras obsevações do conjunto de dados utilizado. Temos duas variáveis: concentração e área, e para cada medida de concentração, há 6 medidas de área.

Concentração	Área
0,24	8597,85
0,24	8597,85
0,24	8596,78
0,24	8596,90
0,24	8597,30
0,24	8597,49
0,27	9607,39
0,27	9607,71
0,27	9607,44
0,27	9608,13
0,27	9607,17
0,27	9607,24

Temos 5 valores para a concentração, para cada um deles foi realizada a medição no equipamento 6

vezes, e o resultado dessa medição é a área apresentada. Observamos que para cada concentração temos valores muito próximos para a área, o que é o esperado, pois a variação o equipamento é muito pequena.

3 Análise Descritiva

Quando se tem um conjunto de dados, a análise descritiva é o primeiro passo que deve ser feito, pois fornece resumos simples sobre a amostra e sobre as variáveis observadas. Tal resumo pode ser quantitativo ou visual, com tabelas, gráficos e figuras.

Como no caso de validação de métodos analíticos as variáveis são concentração e área, sabemos que ambas são **positivas** e **não nulas**, ou seja, maiores que zero. Sugerimos então, fazer um gráfico de dispersão e calcular as medidas resumo para começar a entender melhor como os dados se comportam.

Pelo gráfico de dispersão (Figura 1), é possível observar como é a relação entre as duas variáveis, se há valores discrepantes ou se há alguma inconsistência nos dados (erro de digitação, por exemplo). Esse tipo de gráfico utiliza coordenadas cartesianas para ilustrar os valores do conjunto de dados. Estes são exibidos como uma coleção de pontos, cada um com os valores observados da concentração (que determina a posição no eixo horizontal) e o valor da área (que determina a posição no eixo vertical). Por exemplo, para uma concentração de 0,24 observamos 6 valores para a área (esses valores aparecem sobrepostos no gráfico devido a pequena variabilidade da variável área).

Pelas medidas resumo (Tabela 2), verificamos se existe pontos discrepantes, menores que zero, como é o desvio padrão das variáveis. Tais medidas consistem em valores máximo e mínimo, média e desvio padrão, para cada uma das variáveis. Máximo e Mínimo correspondem ao maior e ao menor valor observado, respectivamente, em relação à cada variável. Média é a média aritmética das observações de cada variável, e por fim o Desvio Padrão é uma medida de dispersão em torno da Média.

Se houver suspeita de que algum valor não esteja correto, deve-se realizar uma análise mais aprofundada para confirmar, lembrando que **não** devemos retirar um valor da amostra, por mais que ele seja muito diferente dos demais, exceto quando há **certeza** de que foi um erro humano ou falha do equipamento.

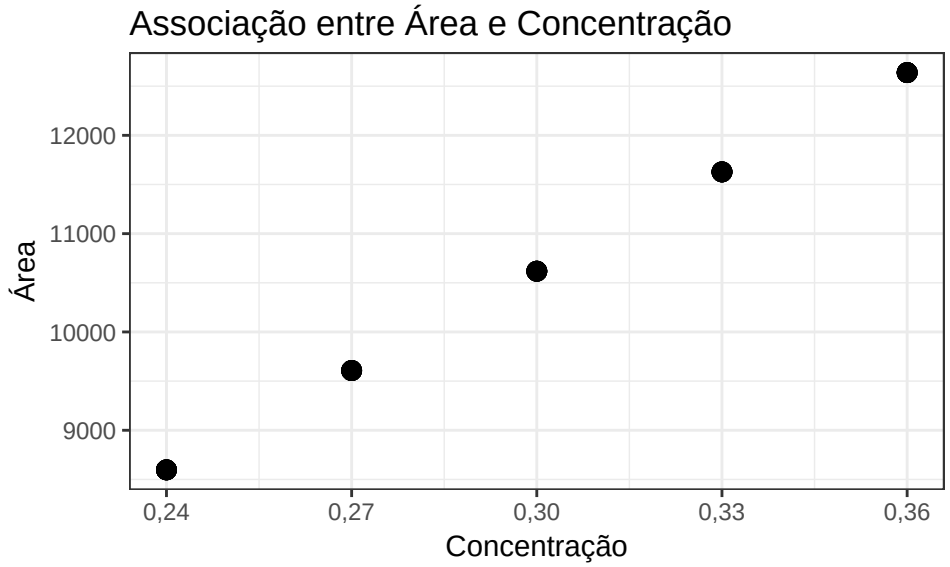

Figura 1: No gráfico de dispersão temos os valores da variável área no eixo *Y* e da variável concentração

no eixo X. Note que temos 6 pontos sobrepostos para cada concentração, pois a variabilidade da área é pequena, e que as variáveis tem uma relação linear, quanto maior a concentração, maior a área.

Tabela 2: Medidas resumo do conjunto de dados, que são: mínimo, máximo, média e desvio padrão. Não observamos *outliers*, todos os valores são maiores que zero.

	Concentração	Área
Mínimo	0,24	8596,78
Máximo	0,36	12638,63
Média	0,30	10617,82
Desvio Padrão	0,04	1453,16

4 Regressão Linear (Curva Analítica)

Estatisticamente, a melhor forma de avaliar a linearidade é através da regressão linear, pois é a metodologia que estuda a relação entre uma variável resposta (área) e outras variáveis (concentração). Com a regressão, vemos qual a melhor reta $(y = a + bx)$ a ser traçada que se ajusta melhor nesses dados, ou seja, qual reta faz com que a distância entre os pontos e a reta ajustada seja a menor possível, minimizando o erro e estimando com maior precisão, por isso a estimação é feita através do Método dos Mínimos Quadrados [1].

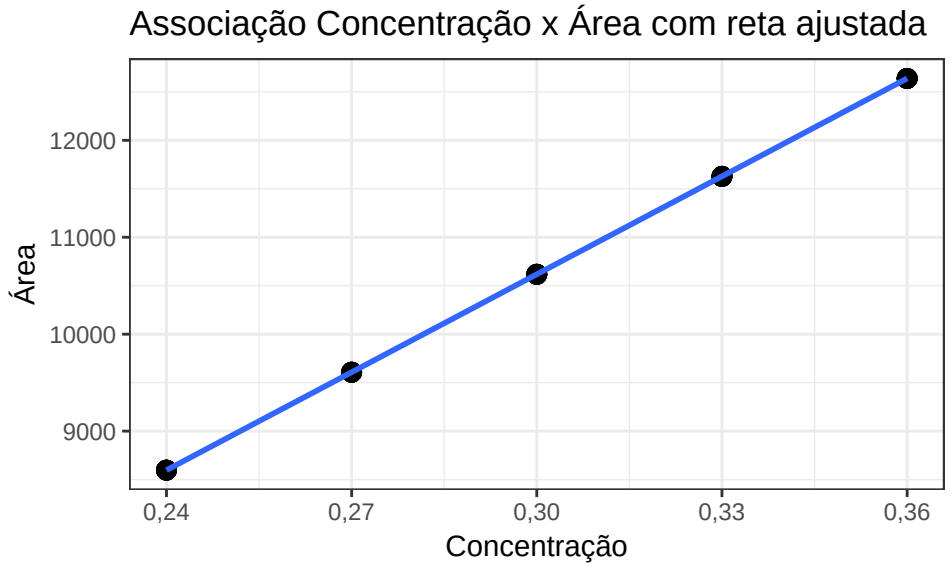

Figura 2: Gráfico de dispersão das variáveis área e concentração com a reta de regressão linear ajustada. Note que pela figura, a reta aparenta passar perfeitamente pelos pontos, mas existe uma distância muito pequena entre a reta e as observações sobrepostas.

Na Figura 2, temos a reta de regressão ajustada (curva analítica), podemos notar que os pontos se alinham quase perfeitamente a uma reta, porém existe uma distância muito pequena entre a reta e as observações sobrepostas. A equação da reta para o exemplo é dada por:

$$area = 515,12 + 33675,67 * Concentracao$$

ou seja, $a = 515,12$ e $b = 33675,67$. O coeficiente de correlação obtido é $R^2 = 0,999$, ou seja, $R^2 \approx 1$ lembrando que a Anvisa sugere que o R^2 esteja acima de $0,990$.

5 Significância dos Parâmetros

Após determinarmos a metodologia e ajustarmos o modelo, é necessário verificar a significância dos parâmetros, ou seja, se a e b da reta $y = a + bx$ são estatisticamente diferentes de zero. Para isso, foi realizado o teste *t-Student* [1]. Os resultados são as estimativas para os parâmetros a e b, onde o parâmetro a é chamado intercepto e representa o ponto em que a curva analítica corta o eixo dos y's, quando concentração = 0. E o parâmetro b representa a inclinação da curva analítica e é dito coeficiente angular. O p-valor dos parâmetros da reta ajustada é a probabilidade de se obter uma estatística de teste igual ou mais extrema que aquela observada em uma amostra, sob a hipótese nula. A hipótese nula do tste é de que os parâmetros são iguais a zero.

Considerando o nosso exemplo, os resultados são apresentados na Tabela 3:

Tabela 3: Resultados do teste *t-Student* - observamos que ambos os p-valores são aproximadamente zero, ou seja, há evidências de que a concentração tem uma associação linear com a área, em um modelo cujo coeficiente angular é diferente de zero.

Parâmetro	Estimativa	p-valor
a	515,12	0
b	33675,67	0

Para avaliar a significância dos parâmetros (a e b), ou seja, se eles são diferentes de zero, vamos observar o p-valor na Tabela 3. Dado que o nível de significância sugerido pela Anvisa é 5%, se o p-valor for menor que 0,05 há evidências de que a concentração tem uma associação linear com a área, em um modelo cujo coeficiente angular é diferente de zero (lembrando que dentre as sugestões da Anvisa, o coeficiente angular deve ser significativamente diferente de zero). Para o nosso exemplo, os dois parâmetros são significativos, uma vez que o p-valor de ambos é menor que 0,05.

6 Diagnóstico do Modelo - Análise de Resíduos

O ajuste de uma reta de regressão é feito baseado em suposições teóricas. Para verificar se a reta de regressão ajustada satisfaz as suposições teóricas da metodologia estatística, devemos avaliar três suposições relacionadas ao resíduo (diferença entre a variável resposta observada e a variável resposta estimada), são elas: Independência, Homocedasticidade e Normalidade. Podemos fazer análises gráficas e também testes estatísticos específicos para cada uma delas, para assim constatar se as três suposições são válidas.

A independência pode ser verificada, se no gráfico de resíduos versus ordem da coleta, os pontos estiverem distribuídos de forma aleatória em torno de zero e dentro dos limites, assim como na Figura 3. Podemos fazer também o teste *Durbin-Watson* [2], que é utilizado para detectar a presença de autocorrelação (dependência), representada por ρ . Temos que os resíduos são independentes se, dada as hipóteses:

$$H_0 : \rho = 0 \text{ versus } H_1 : \rho \neq 0$$

não rejeitamos H_0 se o p-valor obtido for maior que o nível de significância, isso significa que não

rejeitamos a hipótese de que a correlação é zero, ou seja, há evidências de independência. O p-valor encontrado para nosso exemplo foi de 0,3554, ou seja, temos evidências de que os resíduos são independentes, com nível de significância de 5%. Na Figura 3, utilizamos resíduos padronizados (subtraímos a média e dividimos pelo desvio padrão) para que todos os resíduos possam ser comparados dentro de um mesmo limite de confiança, que é [-1,96;1,96] ou aproximadamente [-2;2], assim espera-se que 95% dos resíduos estejam dentro desses limites [3]. Podemos observar que os pontos estão dispersos aleatoriamente, o que é mais uma evidência de que os resíduos são independentes.

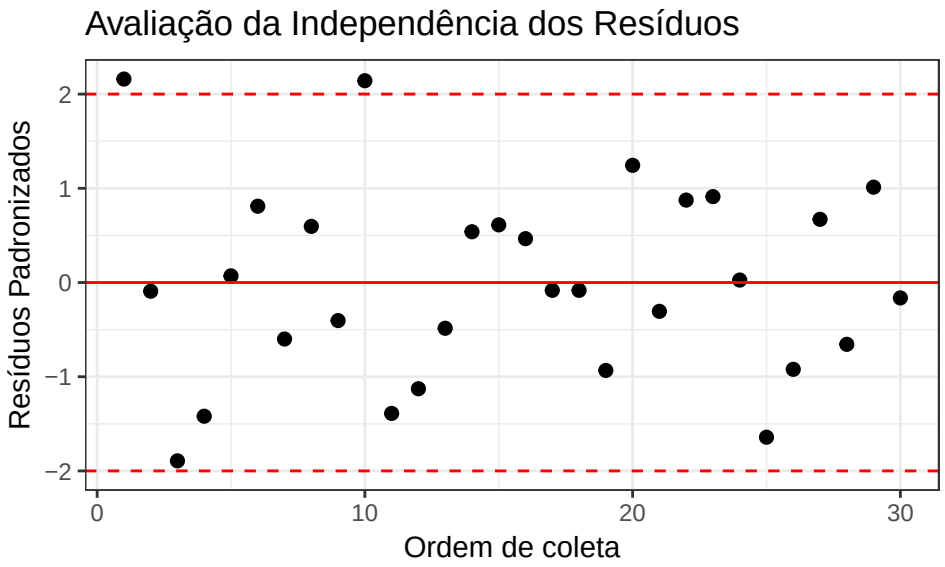

Figura 3: Gráfico de resíduos padronizados versus ordem de coleta. Os valores dos resíduos são apresentados seguindo a ordem da coleta dos dados, então se os pontos estiverem distribuídos de forma aleatória em torno de zero e dentro dos limites, é uma evidência de independência dos resíduos, como foi observado para esse gráfico.

A segunda suposição a ser verificada é a homocedasticidade, isto é, se a variância dos resíduos é constante. Uma forma de fazer isso é gerando o gráfico de resíduos versus valores ajustados. Se os pontos estiverem distribuídos de forma aleatória, sem uma tendência de aumento ou diminuição de variabilidade, há indícios de que a homocedasticidade é satisfeita (Figura 4). Para complementar a análise, um teste estatístico que pode ser feito é o Teste de *Cochran* [3], que compara a maior variância com as demais. As hipóteses são dadas por:

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 \text{ versus } H_1 : \text{pelo menos um } \sigma_i^2 \text{ é diferente, } i = 1, ...k, \text{ onde } k \text{ é o número de observações}$$

Do mesmo modo como teste anterior, os resíduos são homocedásticos se p-valor for maior que o nível de significância. No nosso exemplo $k = 30$, e o p-valor encontrado foi de 0,409, então há evidência estatística de que os resíduos do modelo são homocedásticos, a nível de 5% de significância. Ou seja, não detectamos diferença significativa entre as variâncias dos resíduos do modelo.

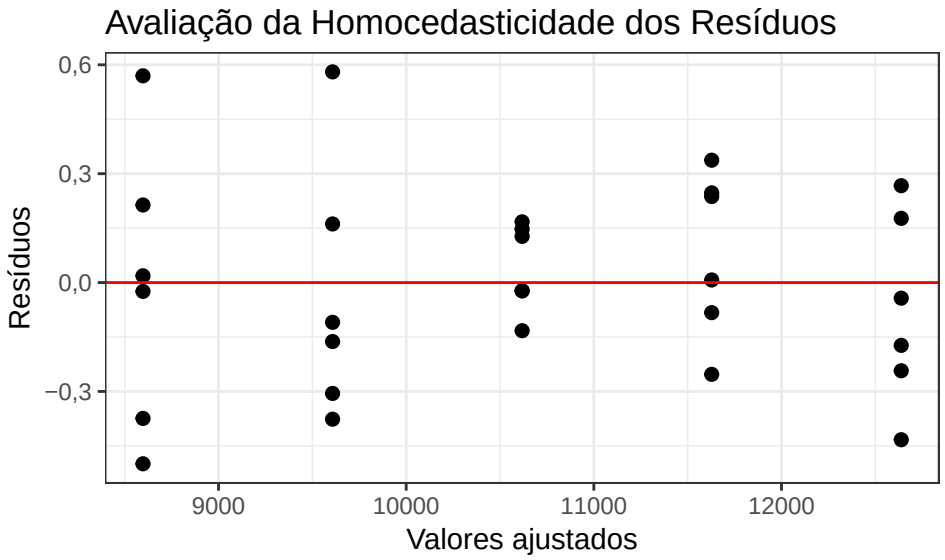

Figura 4: Gráfico de resíduos versus valores ajustados. Uma vez que os pontos estão distribuídos de forma aleatória, sem uma tendência de aumento ou diminuição de variabilidade dos resíduos quando comparamos os valores ajustados, há evidência de que a variância desses resíduos seja constante.

Por fim, vamos verificar terceira suposição: a normalidade dos resíduos por meio do gráfico quantil-quantil (q-q) que é um método gráfico para comparar duas distribuições de probabilidade, traçando seus quantis uns contra os outros. Primeiro, o conjunto de intervalos para os quantis é escolhido, neste caso os quantis de uma distrbuição Normal. Um ponto (x, y) no gráfico corresponde a um dos quantis da amostra (coordenada y) plotadas contra o mesmo quantil da distribuição Normal (coordenada x).

Se os resíduos são normais, os pontos do gráfico devem estar próximos e em torno da reta, de forma aleatória, sem seguir nenhum padrão, como é mostrado na Figura 5. Conjuntamente com a análise gráfica, vamos utilizar um dos testes estatísticos mais conhecidos para verificar a normalidade, o teste de *Shapiro-Wilk*, em que as hipóteses são:

H_0: A amostra vem de uma distribuição Normal versus H_1: A amostra não vem de uma distribuição Normal

E assim, para o nosso exemplo, chegamos em conclusões similares aos outros testes: como p-valor encontrado é 0,7476 e é maior que o nível de significância, temos evidência estatística de normalidade dos resíduos. Além disso, na Figura 5 os pontos estão bem dispostos aleatoriamente em torno da reta, ou seja, há indícios de que os resíduos seguem uma distribuição normal, e assim a última suposição é satisfeita para o exemplo.

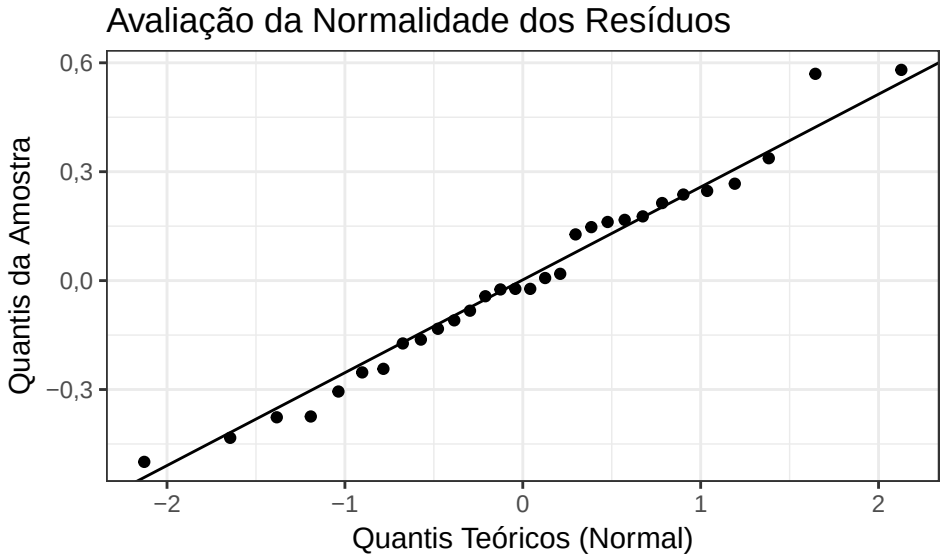

Figura 5: Gráfico Quantil-Quantil, observamos que os pontos estão bem dispostos aleatoriamente em torno da reta, ou seja, há indícios de que os resíduos seguem uma distribuição normal.

Se alguma das suposições acima não ocorrer, a linearidade não é satisfeita, uma vez que as suposições teóricas para ajuste da reta não são válidas. Para o nosso exemplo, a linearidade é satisfeita, pois a reta está bem ajustada e os resíduos satisfazem as suposições do modelo de regressão, e também para todos os testes o p-valor foi maior que o nível de significância estabelecido. Um outro teste que complementa a avaliação do modelo de regressão é o Teste da Falta de Ajuste [4], que verifica se o modelo linear é adequado para o conjunto de dados ou não, testando as seguintes hipóteses:

$$H_0 : \text{Modelo linear é adequado versus } H_1 : \text{Modelo linear não é adequado}$$

Este teste assume que a normalidade, independência e homocedasticidade dos resíduos sejam satisfeitas (o que é válido no nosso exemplo). Assim, após o ajuste e análise de resíduos, é importante verificar se o modelo linear é adequado. Como podemos ver na Tabela 4, o p-valor encontrado para a falta de ajuste foi 0,7676, sendo é maior que o nível de significância 0,05, ou seja, como não rejeita-se H_0, há evidências de que o modelo linear é adequado para esse conjunto de dados.

Tabela 4: Resultados do Teste de Falta de Ajuste. Como temos um p-valor muito grande para a falta de ajuste (0,7676), não rejeita-se a hipótese de que o modelo linear é adequado, isto é, há evidência de que o modelo linear ajustado.

	Estatística	p-valor
Concentração	739650308	0
Falta de ajuste	0,38	0,7676

7 Conclusão

Através de métodos estatísticos temos condição de validar a Linearidade. Partindo inicialmente com uma análise descritiva, podemos ter uma visão geral dos dados e verificar se há possíveis valores muito discrepantes (*outliers*), erro da amostra, erro de digitação ou alguma inconsistência. Seguindo com uma

análise de regressão linear, com o objetivo de traçar uma reta que matematicamente interpreta os pontos amostrados, essa reta denomidada de curva analítca possui um coeficiente linear que também foi calculado conforme é pedido pelas normas de validação da Linearidade pelo nosso país.

Por fim, é feita uma análise para verificar a qualidade desta curva e se ela de fato pode ser utilizada para representar a amostra, o que inclui os testes para verificação de independência, normalidade e homocedasticidade dos resíduos do modelo. A análise dos gráficos é de certa forma subjetiva, por isso cada conjunto de dados deverá ser analisado por um avaliador experiente e ele quem deverá fazer as devidas conclusões. Já para as análises que possuem p-valor, após um nível de significância pré definido, o resultado será obtido diretamente.

Os códigos computacionais no *software* R podem ser acessados no *link*: https://github.com/brumarques/ME810 - arquivo "codigo.Rmd"

8 Referências Bibliográficas

[1] Michael H Kutner, Christopher J. Nachtsheim, John Neter, William Li. Applied Linear Statistical Models.

[2] J. Durbin G. S. Watson. Testing for serial correlation in least squares regression.III

[3] SHESKIN, David J. Handbook of Parametric and Nonparametric Statistical Procedures. 4th ed.

[4] Norman R. Draper, Harry Smith. Applied Regression Analysis.