O problema dos erros correlacionados

Ian Kanda Bernucci - 12557421
Júlia de Lima Oliveira - 12567082
Matheus Laureano - 11965223

Falaremos sobre

- Remoção de Autocorrelação por Transformação
- Estimativa iterativa com erros autocorrelacionados
- Autocorrelação e variáveis ausentes
- Análise de Housing Starts
- Limitações da estatística Durbin-Watson
- Variáveis indicadoras para remover sazonalidade
- Regressão de duas séries temporais

Contexto

Para testar se os erros são correlacionados, podemos usar a estatística Durbin-Watson (d).

O teste é baseado na suposição de que erros sucessivos estão correlacionados, ou seja,

$$\varepsilon_t = \rho\varepsilon_{t-1} + \omega_t, \quad |\rho| < 1 \tag{1.1}$$

onde

- ρ é o coeficiente de correlação entre ε_t e ε_{t-1}
- $\omega_t \sim N(0, \sigma^2)$ independentes

Usamos d para testar H0: $\rho = 0$ x H1: $\rho > 0$, com $d = \dfrac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$

Se $\rho = 0$, os erros são não correlacionados.

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} \tag{1.2}$$

Estimativa para ρ

$$d \doteq 2(1 - \hat{\rho})$$

Relação aproximada entre d e estimativa de ρ

A presença de erros correlacionados pode distorcer:

- Estimativas de erros padrão
- Intervalos de confiança
- Testes estatísticos

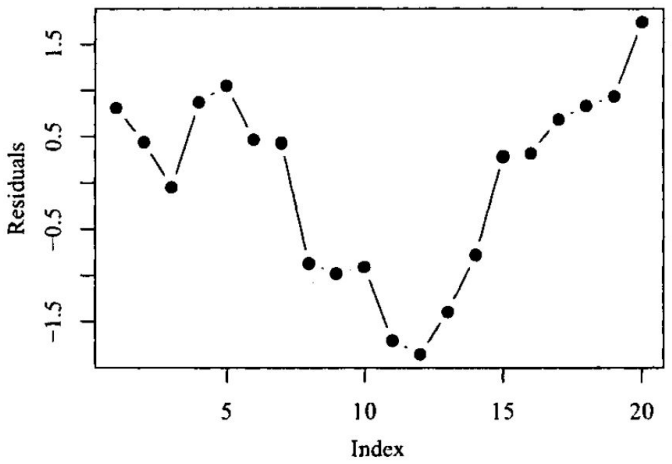

Figure 8.1 Index plot of the standardized residuals.

Quando os gráficos de resíduos e a estatística de Durbin-Watson indicarem a presença de erros correlacionados, a equação de regressão estimada deverá ser reajustada levando-se em conta a autocorrelação. Duas abordagens podem ser seguidas.

(1) trabalhar com variáveis transformadas
(2) introduzir variáveis adicionais que tenham efeitos ordenados no tempo.

Nos casos em que a estatística Durbin-Watson é inconclusiva, podemos estimar novamente a equação de regressão usando os métodos descritos a seguir para checar se ocorre alguma mudança.

Remoção de Autocorrelação por Transformação

Um método para ajustar o modelo é o uso de uma transformação que envolve o parâmetro de autocorrelação desconhecido ρ.

Vamos considerar o modelo $$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \tag{1.3}$$

Substituindo ε_t e ε_{t-1} em 1.1, obtemos $$\begin{array}{ccccccc} y_t - \rho y_{t-1} & = & \beta_0(1-\rho) & + & \beta_1(x_t - \rho x_{t-1}) & + & \omega_t \\ y_t^* & = & \beta_0^* & + & \beta_1^* \quad x_t^* & + & \omega_t \end{array} \tag{1.4}$$

que representa um modelo linear com erros não correlacionados.

Rodar regressão de mínimos quadrados ordinária usando y_t^* como variável de resposta e x_t^* como preditor.

Estimativas dos parâmetros nas equações originais: $$\hat{\beta}_0 = \frac{\hat{\beta}_0^*}{1-\hat{\rho}} \quad \text{and} \quad \hat{\beta}_1 = \hat{\beta}_1^* \tag{1.5}$$

Procedimento de Cochrane and Orcutt

O valor de ρ é desconhecido e deve ser estimado a partir dos dados. Procedimento:

1. Calcule as estimativas MQO de β_0 e β_1, ajustando o modelo 1.3 aos dados.
2. Calcule os resíduos e, a partir dos resíduos, estime ρ usando (1.2).
3. Ajuste a equação dada em (1.4) usando as variáveis $y_t - \hat{\rho} y_{t-1}$ e $x_t - \hat{\rho} x_{t-1}$ como variáveis de resposta e preditoras, respectivamente, e obter as estimativas de β_0 e β_1 usando (1.5).
4. Examine os resíduos da equação recém-ajustada. Se os novos resíduos continuarem a apresentar autocorrelação, repita todo o procedimento usando as estimativas de β_0 e β_1, em vez das estimativas originais de mínimos quadrados.
5. Por outro lado, se os novos resíduos não apresentarem autocorrelação, o procedimento acaba e a equação ajustada para os dados originais é $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$

Estimativa iterativa com erros autocorrelacionados

Abordagem mais direta: tentar estimar ρ, β_0 e β_1 simultaneamente.

- O modelo é formulado como antes, exigindo a construção de variáveis transformadas $y_t - \rho y_{t-1}$ $x_t - \rho x_{t-1}$
- Estimativas dos parâmetros obtidas minimizando a soma dos erros quadráticos

$$S(\beta_0, \beta_1, \rho) = \sum_{t=2}^{n} [y_t - \rho y_{t-1} - \beta_0(1-\rho) - \beta_1(x_t - \rho x_{t-1})]^2$$

Se ρ conhecido, β_0 e β_1 seriam facilmente obtidos fazendo a regressão de $y_t - \rho y_{t-1}$ em $x_t - \rho x_{t-1}$

- Estimativas finais $\tilde{\rho}, \tilde{\beta}_0$ e $\tilde{\beta}_1$ obtidas procurando vários valores de ρ até que seja encontrada uma combinação de ρ, β_0 e β_1 que minimize S(ρ, β_0 e β_1).

O erro padrão para $\tilde{\beta}_1$ é $\text{s.e.}(\tilde{\beta}_1) = \dfrac{\hat{\sigma}}{\sqrt{\sum [x_t - \tilde{\rho} x_{t-1} - \bar{x}(1-\tilde{\rho})]^2}}$, onde $\hat{\sigma} = S(\tilde{\rho}, \tilde{\beta}_0, \tilde{\beta}_1)/(n-2)$

Estimativa iterativa com erros autocorrelacionados

- Recomenda-se utilizar o método iterativo quando os recursos computacionais adequados estiverem disponíveis.

- Não se espera que as estimativas e os erros padrão para o método iterativo e o método Cochrane-Orcutt de dois estágios sejam significativamente diferentes.

Table 8.3 Comparison of Regression Estimates

Method	$\hat{\rho}$	$\hat{\beta}_0$	$\hat{\beta}_1$	s.e.$(\hat{\beta}_1)$
OLS	–	−154.700	2.300	0.115
Cochrane-Orcutt	0.874	−324.440	2.758	0.444
Iterative	0.824	−235.509	2.753	0.436

Autocorrelação e variáveis ausentes

Cuidado:
Características que sugerem autocorrelação dos resíduos também podem ser indicativos de outras falhas na especificação do modelo (como plot dos resíduos em clusters, coeficiente de autocorrelação estimado grande, e estatística de Durbin-Watson significativa)

- Gráfico de resíduos x potencial variável preditora, pode ser usado para explicar melhor a variação na variável de resposta
- Gráfico de resíduos x tempo mostra padrão do tipo descrito no exemplo anterior, pode ser devido à omissão de variáveis que mudam ao longo do tempo.

Sempre melhor explorar completamente a possibilidade de algumas variáveis preditoras adicionais.

As transformações que corrigem a autocorrelação pura ficam como último recurso.

Caso prático: Análise de Housing Starts

- Uma construtora quer compreender associação entre o novas construções e o crescimento populacional, para prever atividade de construção.

- 25 anos de dados do índice de housing starts regional e população de potenciais compradores

- Objetivo: obter uma relação de regressão simples entre o início da construção de moradias e a população

$$H_t = \beta_0 + \beta_1 P_t + \varepsilon_t$$

Table 8.4 Data for Housing Starts (H), Population Size (P) in millions, and Availability for Mortgage Money Index (D)

Row	H	P	D
1	0.09090	2.200	0.03635
2	0.08942	2.222	0.03345
3	0.09755	2.244	0.03870
4	0.09550	2.267	0.03745
5	0.09678	2.280	0.04063
6	0.10327	2.289	0.04237
7	0.10513	2.289	0.04715
8	0.10840	2.290	0.04883
9	0.10822	2.299	0.04836
10	0.10741	2.300	0.05160
11	0.10751	2.300	0.04879
12	0.11429	2.340	0.05523
13	0.11048	2.386	0.04770
14	0.11604	2.433	0.05282
15	0.11688	2.482	0.05473
16	0.12044	2.532	0.05531
17	0.12125	2.580	0.05898
18	0.12080	2.605	0.06267
19	0.12368	2.631	0.05462
20	0.12679	2.658	0.05672
21	0.12996	2.684	0.06674
22	0.13445	2.711	0.06451
23	0.13325	2.738	0.06313
24	0.13863	2.766	0.06573
25	0.13964	2.793	0.07229

Análise de Housing Starts

- Modelo ajustado:

Table 8.5 Regression on Housing Starts (H) Versus Population (P)

Variable	Coefficient	s.e.	t-test	p-value
Constant	-0.0609	0.0104	-5.85	< 0.0001
P	0.0714	0.0042	16.90	< 0.0001
$n = 25$	$R^2 = 0.925$	$d = 0.621$	$\hat{\sigma} = 0.0041$	$d.f. = 23$

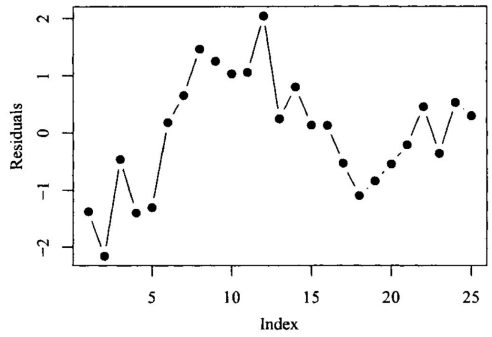

Figure 8.3 Index plot of standardized residuals from the regression of H_t on P_t for the Housing Starts data.

Porém, outras variáveis podem explicar melhor o número de novas construções, e podem ser responsáveis pelo aparecimento de autocorrelação, como:

- Taxa de desemprego
- Tendências sociais no casamento e na formação da família
- Programas governamentais de habitação
- Disponibilidade de fundos para construção e hipotecas

Análise de Housing Starts

- Novo modelo com variável de disponibilidade empréstimo hipotecário para a região:

$$H_t = \beta_0 + \beta_1 P_t + \beta_2 D_t + \varepsilon_t$$

Table 8.6 Results of the Regression of Housing Starts (H) on Population (P) and Index (D)

Variable	Coefficient	s.e.	t-test	p-value
Constant	−0.0104	0.0103	−1.01	0.3220
P	0.0347	0.0064	5.39	< 0.0001
D	0.7605	0.1216	6.25	< 0.0001
$n = 25$	$R^2 = 0.973$	$d = 1.85$	$\hat{\sigma} = 0.0025$	$d.f. = 22$

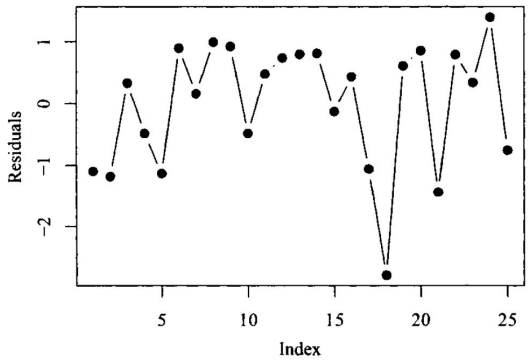

Figure 8.4 Index plot of the standardized residuals from the regression of H_t on P_t and D_t for the Housing Starts data.

- Alterações na disponibilidade de dinheiro hipotecário para um nível fixo de população tem efeito maior em housing starts que uma alteração semelhante no tamanho da população.

Análise de Housing Starts

- Se cada variável na equação de regressão for substituída pela versão padronizada da variável (as variáveis transformadas de modo a ter média 0 e variância 1), a equação de regressão resultante é

$$\tilde{H}_t = 0.4668\tilde{P}_t + 0.5413\tilde{D}_t$$

onde $\tilde{H}$ denota o valor padronizado de H, $\tilde{H} = (H - \bar{H})/s_H$

Aprendizados:

- Um valor grande de R2 não implica que os dados tenham sido bem ajustados e explicados
- A estatística de Durbin-Watson e os gráficos residuais podem indicar a presença de autocorrelação entre os erros quando, na verdade, os erros são independentes, mas a omissão de variáveis deu origem à situação observada.

Limitações da estatística Durbin-Watson

- Se a dependência temporal nos resíduos for diferente de primeira ordem, o gráfico dos resíduos ainda será informativo. Porém, a estatística Durbin-Watson não foi concebida para medir a dependência temporal de ordem superior e pode não produzir muita informação valiosa.

- Exemplo: relação entre vendas trimestrais de ski nos EUA e indicador econômico PDI (personal disposable income, em bilhões de dólares).

Modelo inicial: $S_t = \beta_0 + \beta_1 \text{PDI}_t + \varepsilon_t$

S_t = vendas de esqui no período t em milhões de dólares
PDI_t = renda pessoal disponível para o mesmo período

Table 8.7 Ski Sales Versus PDI

Variable	Coefficient	s.e.	t-test	p-value
Constant	12.3921	2.539	4.88	< 0.0001
PDI	0.1979	0.016	12.40	< 0.0001
$n = 40$	$R^2 = 0.801$	$d = 1.968$	$\hat{\sigma} = 3.019$	$d.f. = 38$

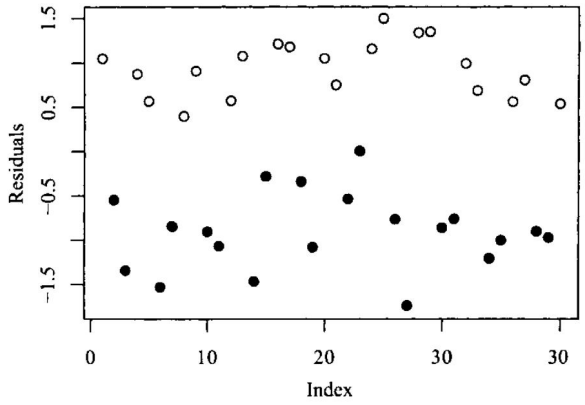

Figure 8.5 Index plot of the standardized residuals. (Quarters 1 and 4 are indicated by an open circle and Quarters 2 and 3 are indicated by a solid circle.)

Limitações da estatística Durbin-Watson

Table 8.7 Ski Sales Versus PDI

Variable	Coefficient	s.e.	t-test	p-value
Constant	12.3921	2.539	4.88	< 0.0001
PDI	0.1979	0.016	12.40	< 0.0001
$n = 40$	$R^2 = 0.801$	$d = 1.968$	$\hat{\sigma} = 3.019$	$d.f. = 38$

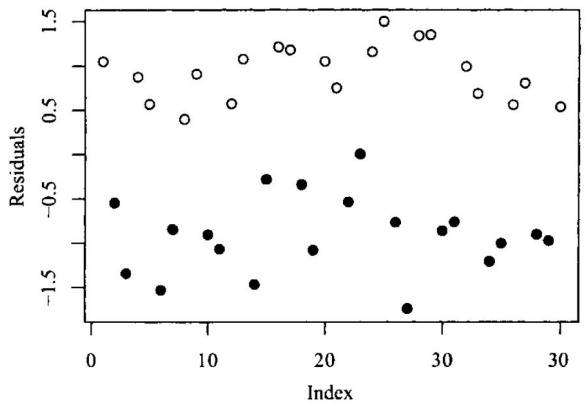

Figure 8.5 Index plot of the standardized residuals. (Quarters 1 and 4 are indicated by an open circle and Quarters 2 and 3 are indicated by a solid circle.)

- Proporção da variação das vendas contabilizada pelo PDI é de 0,80.
- A contribuição marginal de uma unidade adicional em dólares de PDI para as vendas está entre US$ 165.420 e US$ 230.380 ($\hat{\beta}_1$ = 0,1979) com um coeficiente de confiança de 95%.
- Estatística Durbin-Watson é 1.968, indicando ausência de autocorrelação de primeira ordem.

Limitações da estatística Durbin-Watson

Table 8.7 Ski Sales Versus PDI

Variable	Coefficient	s.e.	t-test	p-value
Constant	12.3921	2.539	4.88	< 0.0001
PDI	0.1979	0.016	12.40	< 0.0001
$n = 40$	$R^2 = 0.801$	$d = 1.968$	$\hat{\sigma} = 3.019$	$d.f. = 38$

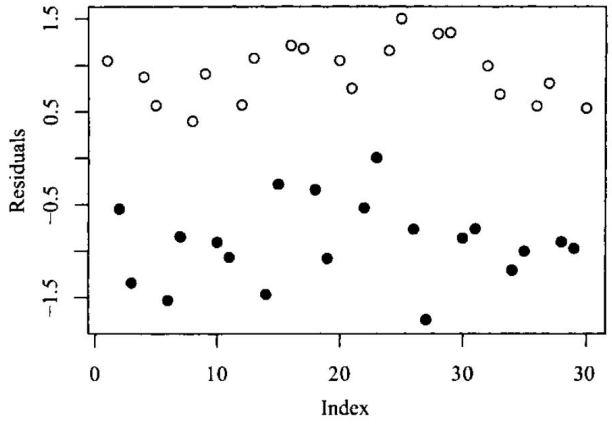

Figure 8.5 Index plot of the standardized residuals. (Quarters 1 and 4 are indicated by an open circle and Quarters 2 and 3 are indicated by a solid circle.)

- Embora o valor de R2 de 0,80 seja bom, não deve ser tomado como avaliação final do modelo.
- O valor de Durbin-Watson está na faixa aceitável, mas fica claro na Figura 8.5 que existe algum tipo de dependência temporal dos resíduos.
  → 1º e 4º trimestres positivos
  → 2º e 3º trimestres negativos
- Este efeito sazonal pode ser caracterizado pela definição de uma variável indicadora (dummy) que assume o valor 1 para cada trimestre de inverno e 0 para cada trimestre de verão.

Variáveis indicadoras para remover sazonalidade

Usando a variável sazonal adicional, o modelo é expandido para

$$S_t = \beta_0 + \beta_1 \text{PDI}_t + \beta_2 Z_t + \varepsilon_t$$

β_2 → parâmetro que mede o efeito sazonal

Z_t → variável 0/1

Pode também ser representado por 2 modelos

$$\begin{aligned} \text{Winter season}: \quad S_t &= (\beta_0 + \beta_2) + \beta_1 \text{PDI}_t + \varepsilon_t \\ \text{Summer season}: \quad S_t &= \beta_0 + \beta_1 \text{PDI}_t + \varepsilon_t \end{aligned}$$

Variáveis indicadoras para remover sazonalidade

Representa a suposição de que as vendas podem ser aproximadas por uma função linear do PDI, em uma linha para o inverno e outra para o verão, paralelas.

Table 8.9 Ski Sales Versus PDI and Seasonal Variables

Variable	Coefficient	s.e.	t-test	p-value
Constant	9.5402	0.9748	9.79	0.3220
PDI	0.1987	0.0060	32.90	< 0.0001
Z	5.4643	0.3597	15.20	< 0.0001
$n = 40$	$R^2 = 0.972$	$d = 1.772$	$\hat{\sigma} = 1.137$	$d.f. = 37$

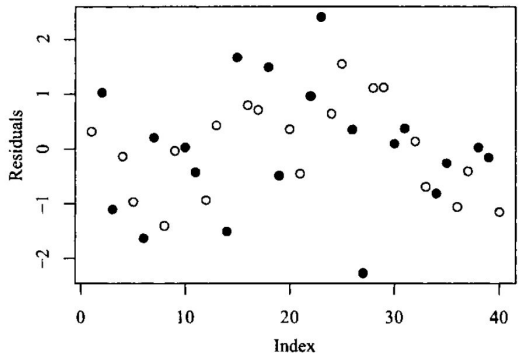

Figure 8.7 Index plot of the standardized residuals with seasonal variables (quarters indicated). (Quarters 1 and 4 are indicated by an open circle and Quarters 2 and 3 are indicated by a solid circle.)

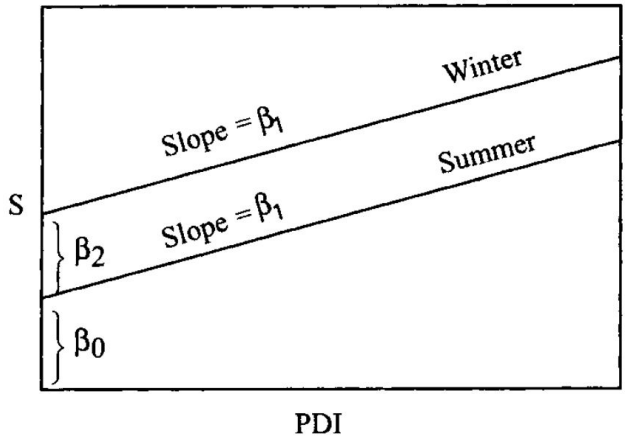

Figure 8.6 Model for Ski Sales and PDI adjusted for season.

Variáveis indicadoras para remover sazonalidade

- Todas as indicações do padrão sazonal foram removidas
- Precisão do efeito marginal estimado do PDI aumentou
- O intervalo de confiança agora é de US$ 186.520 a US$ 210.880
- O efeito sazonal foi quantificado e para um nível fixo de PDI a temporada de inverno traz entre US$ 4.734.109 e US$ 6.194.491 em relação à temporada de verão (com 95% de confiança)

Aprendizados:

- Estatística Durbin-Watson só é sensível a erros correlacionados de primeira ordem. Nos dados de ski:
  Correlação de 1ª ordem: -0,001
  Correlações de 2ª, 4ª, 6ª e 8ª ordem: -0,81, 0,76, -0,71 e 0,73
- O modelo deve ser reajustado quando é indicada autocorrelação
- Se as observações não forem ordenadas a tempo, a estatística Durbin-Watson não é estritamente relevante.
  Porém, por exemplo, uma lista de cidades pode ser ordenada por tamanho.
  Um valor baixo da estatística Durbin-Watson indicaria a presença de um efeito de tamanho significativo.

Regressão de duas séries temporais

- O conceito de autocorrelação não é relevante em dados transversais, já que a ordem das observações muitas vezes é arbitrária.
  Já para dados de séries temporais, a autocorrelação geralmente é um fator significativo, indicando estruturas ocultas nos dados que não foram detectadas (muitas vezes relacionados com o tempo)

- A maioria dos dados de séries temporais apresentam sazonalidade, que deve ser investigada.
  Para dados trimestrais ou mensais, a introdução de variáveis indicadoras é um solução satisfatória.

- Numa tentativa de encontrar uma relação entre y_t e $x_{1t}, x_{2t}, \ldots, x_{pt}$
  podemos tentar expandir as variáveis preditoras, incluindo seus valores defasados, como em:

  $$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{1t-1} + \beta_3 x_{2t} + \varepsilon_t$$

  O que é útil para dados de séries temporais, mas não dados transversais.

Regressão de duas séries temporais

- Os dados de séries temporais provavelmente contêm tendências.
  Podemos incluir variáveis que são funções diretas do tempo (t, t^2, etc)
  Ou usar a primeira diferença simples $(y_t - y_{t-1})$ou lag mais complexo do tipo $(y_t - ay_{t-1})$ como no procedimento de Cochrane-Orcutt

Resumindo:
Ao realizar análise de regressão com dados de séries temporais, se atentar a autocorrelação e aos efeitos sazonais. Explorar a utilização de variáveis preditoras defasadas.

Conclusão e Referências

CHATTERJEE, S.; HADI, A. S. **Regression analysis by example**. Fourth edition; John Wiley & Sons, 2015.