Estratégias para análise de dados de poluição do ar

William Nilson de Amorim

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Doutor em Ciências

Programa: Doutorado em Estatística

Orientador: Prof. Dr. Antonio Carlos Pedroso de Lima

Coorientador: Prof. Dr. Julio da Motta Singer

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES.

São Paulo, fevereiro de 2019

Estratégias para análise de dados de poluição do ar

Esta é a versão original da tese elaborada pelo candidato William Nilson de Amorim, tal como submetida à Comissão Julgadora.

Agradecimentos

Texto texto. Texto opcional.

Resumo

AMORIM, W. N. **Estratégias para análise de dados de poluição do ar**. 2019. ?? f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 201?.

Elemento obrigatório, constituído de uma sequência de frases concisas e objetivas, em forma de texto. Deve apresentar os objetivos, métodos empregados, resultados e conclusões. O resumo deve ser redigido em parágrafo único, conter no máximo 500 palavras e ser seguido dos termos representativos do conteúdo do trabalho (palavras-chave). Texto texto. Texto texto.

Palavras-chave:

Abstract

Amorim, W. N. **Survival analysis models with covariate subjected to non-random missingness**. 201?. ?? f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2010.

Elemento obrigatório, elaborado com as mesmas características do resumo em língua portuguesa. De acordo com o Regimento da Pós- Graduação da USP (Artigo 99), deve ser redigido em inglês para fins de divulgação. Text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text. Text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text.

Keywords:

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Na cidade de São Paulo, nas manhãs de dias secos, podemos observar uma grande faixa cinzenta entre o céu e a linha do horizonte. Saindo de casa, em vias movimentadas, vemos a harmonia dos escapamentos de motos, carros, ônibus e caminhões ao emitirem nuvens negras de gases sufocantes. Talvez, ao longo do trajeto, ainda passemos ao lado de alguma fábrica para testemunhar as incessantes colunas de fumaça tóxica que saem de suas chaminés.

Situações como essas fazem cada vez mais parte do nosso cotidiano. Sempre que nos deparamos com elas, pensamos inconscientemente em não respirar. Infelizmente, não é possível. Precisamos de ar. Em média, respiramos cerca de 30 mil vezes por dia — 50 mil se praticamos exercícios. Para piorar, os poluentes se dissipam rapidamente, tornando o problema invisível na maior parte do tempo. Vivemos sem saber exatamente o que estamos respirando e qual o mal que esse ar nos faz. Sofremos diariamente com os efeitos imediatos da poluição e nos acostumamos com eles. E costumamos nos preocupar com as consequências de longo prazo apenas quando já é tarde demais.

Se quisermos viver em um meio ambiente mais limpo, saudável e sustentável, sem abrir mão de diversos confortos, estudar esse tema é indispensável.

Talvez começar essa seção aqui

A poluição do ar, considerada pela Organização Mundial da Saúde (OMS) o maior risco ambiental à saúde humana, é responsável por aproximadamente 7 milhões de mortes por ano, um oitavo do total global (Jasarevic *et al.*, 2014). Poluentes como óxidos de carbono, nitrogênio e enxofre, ozônio e material particulado trazem diversos prejuízos a nossa qualidade de vida e ao equilíbrio do planeta. Eles são agentes sistemáticos no desenvolvimento de irritação dos olhos, obstrução nasal, tosse, asma e redução da função pulmonar. À exposição contínua estão associadas diversas doenças respiratórias e cardiovasculares, problemas digestivos e no sistema nervoso, câncer e o aumento da mortalidade infantil (European Commission, 1999). Além disso, vários ~~deles~~ poluentes estão diretamente ligados ao aquecimento global e ao efeito estufa.

As taxas elevadas de poluição do ar geralmente são ~~um~~ produto de políticas não-sustentáveis em setores como transporte, energia, saneamento e indústria. A escolha de estratégias favoráveis ao meio ambiente costuma esbarrar em fatores econômicos, mesmo quando a redução a longo prazo nos gastos com tratamentos de saúde poderia gerar números positivos para esse balanço. Nas últimas décadas, diversos estudos vêm sendo realizados para alertar sobre os riscos da poluição atmosférica. Seus principais objetivos compreendem a descrição dos níveis de poluição em uma determinada região, o acompanhamento das concentrações dos poluentes ao longo do tempo, a busca por associações entre a concentração de poluentes e mortalidade ou morbidade e o desenvolvimento de soluções mais limpas — ou menos poluentes — e ainda economicamente viáveis.

Carslaw *et al.* (2007), por exemplo, modelaram concentrações diárias de óxidos e dióxidos de nitrogênio, monóxido de carbono, benzeno e 1,3-butadieno para avaliar a tendência das concentrações desses poluentes durante o período de 1998 a 2005 no movimentado centro de Londres. Já Beer *et al.* (2011) analisaram dados de morbidade e mortalidade para estudar os impactos à saúde ao se utilizar etanol como aditivo na gasolina em regiões urbanas da Austrália, medindo níveis de ozônio, dióxido de nitrogênio e material particulado em câmaras de poluição. Kloog *et al.* (2012) utilizaram medidas de profundidade óptica de aerossóis feitas por satélites para prever concentrações diárias de material particulado na costa leste dos Estados Unidos. Belusic *et al.* (2015) estudaram a relação das concentrações horárias de monóxido de carbono, dióxido de enxofre, dióxido de nitrogênio e material particulado com as condições climáticas da cidade de Zagrebe, na Croácia.

Os estudos citados — e muitos outros, como Chang *et al.* (2017), Conceição *et al.* (2001a,b), Lin *et al.* (1999), Schwartz *et al.* (1996), Katsouyanni *et al.* (1996), Schwartz (1994, 1996), Saldiva *et al.* (1994, 1995), Schwartz e Dockery (1992), Schwartz e Marcus (1990) e Shumway e Stoffer (1982) — , embora abordem diferentes temas, concordam sobre a importância da diminuição da emissão de poluentes.

Devido à forte dependência de combustíveis fósseis, o setor de transportes é considerado pela União Europeia o mais resiliente aos esforços para a redução de emissões (European Commission, 2011). Como soluções que visam controlar o tamanho da frota de veículos ou restringir o seu uso são limitadas por fatores políticos e econômicos, os estudos nessa área buscam encontrar combustíveis menos poluentes, alternativas ao diesel e à gasolina.

Em um experimento controlado na cidade de Fairbanks, no Alasca, Mulawa *et al.* (1997) coletaram amostras de material particulado de carros a gasolina e as compararam com dados de emissões de carros abastecidos com E10 (gasolina com 10% de álcool). Os autores constataram que os carros com E10 emitiam menos material particulado e que os níveis desse poluente aumentavam conforme a temperatura dos dois combustíveis diminuía. Yoon *et al.* (2009) conduziram uma investigação similar e concluíram que a combustão de etanol e da mistura E85 (85% etanol e 15% gasolina) emitiam concentrações inferiores de hidrocarbonetos, monóxido de carbono e óxidos de nitrogênio quando comparados com a gasolina sem aditivos sob diversas condições experimentais. Já Pereira *et al.* (2004) expuseram ao sol câmaras contendo etanol puro e gasool (mistura de 22-24% etanol em gasolina) para estudar a formação do ozônio e concluíram que as concentrações máximas do poluente eram, em média, 28% maiores para o álcool do que para o gasool.

Associada à temperatura ambiental?

Embora a formação de poluentes envolva reações químicas complicadas, cuja análise demanda ambientes controlados, as grandes cidades funcionam como um laboratório natural para os estudos de poluição do ar. Com a disponibilidade de dados meteorológicos e de tráfego, é possível avaliar grande parte dos fatores que influenciam na formação dos poluentes. Muitos trabalhos vêm trocando os experimentos controlados pelos dados de poluição urbana. Jacobson (2007), por exemplo, estudou os efeitos da substituição da gasolina por etanol (E85) nas taxas de câncer, mortalidade e hospitalização em Los Angeles, em particular, e nos Estados Unidos como um todo. Os resultados do estudo mostraram que a utilização de E85, controlada pelas variáveis climáticas, aumentou as taxas de mortalidade, hospitalização e asma devido a maiores concentrações de ozônio. Salvo e Geiger (2014) utilizaram uma mudança real na preferência por gasolina ocasionada em flutuações de larga escala no preço do etanol para analisar a associação entre a proporções de carros a gasolina rodando

podem ser pensadas (?)

por(?)

na cidade de São Paulo[1] com os níveis de ozônio medidos no começo da tarde durante os anos de 2008 a 2011. Os autores concluíram que o uso do etanol em São Paulo está associado a maiores concentrações do poluente.

Uma das maiores dificuldades associadas ao estudo de dados de poluição atmosférica está no grande número de efeitos confundidores. Em áreas urbanas, pode existir uma complexa mistura de fatores que contribuem para a formação e dispersão dos poluentes. Diversas variáveis podem ser consideradas, como aquelas relacionadas ao clima, ao tráfego, à química atmosférica local, às mudanças climáticas sazonais, a eventos esporádicos na região (que podem alterar o fluxo do trânsito), ao tamanho e idade da frota de veículos, às emissões evaporativas, entre outras. Além disso, a relação entre essas variáveis pode não ser muito simples, o que exige o uso modelos menos restritivos, e não é rara a presença de dados omissos ou grande períodos sem observação, intrincando ainda mais a análise.

dificultando(?)

Embora diferentes técnicas estatísticas venham sendo empregadas na modelagem desses dados, nenhuma delas é robusta o suficiente para atacar sozinha todos esses problemas. Na maioria dos casos, a abordagem mais adequada seria a formulação de estratégias que envolvessem a combinação de duas ou mais técnicas, além de ferramentas estatísticas ainda não utilizadas neste contexto.

As principais metodologias para abordar dados de poluição e epidemiologia ambiental envolvem modelos lineares (Salvo e Geiger, 2014), modelos lineares generalizados (Conceição *et al.*, 2001b; Lin *et al.*, 1999; Saldiva *et al.*, 1994, 1995; Schwartz e Dockery, 1992) e modelos aditivos generalizados (Carslaw *et al.*, 2007; Conceição *et al.*, 2001a,b; Schwartz *et al.*, 1996; Schwartz, 1994, 1996) e séries temporais (Katsouyanni *et al.*, 1996; Schwartz e Marcus, 1990; Shumway e Stoffer, 1982). A escolha de uma estratégia de análise adequada é muito importante, pois dados de poluição do ar[2] usualmente violam as suposições associadas a esses modelos. Salvo e Geiger (2014), por exemplo, utilizam um modelo linear, que supõe independência entre as observações, para ajustar uma série horária de concentração de ozônio, possivelmente autocorrelacionada. No trabalho, a variável explicativa de maior interesse, a proporção de carros a gasolina na cidade de São Paulo, é estimada, isto é, existe um erro associado a essas observações, tornando um modelo com erros nas variáveis mais adequado para essa análise. Além disso, a utilização de variáveis indicadoras para a hora, o dia da semana e a semana do ano pode não ter sido adequada para controlar a sazonalidade da série. De uma forma geral, autocorrelação, heteroscedasticidade, superdispersão, tendência, sazonalidade, componentes espaciais, variáveis com erro de medida e grandes períodos sem observação, acarretando um grande número de observações omissas, são características comuns em dados de poluição do ar e precisam ser contempladas pelo modelo escolhido.

caberia uma referência aqui(?)

Modelos lineares são uma boa alternativa devido à facilidade de implementação e interpretação de seus coeficientes. Além disso, grandes intervalos sem observações não geram maiores problemas no processo de estimação, já que as observações não são interpretadas como uma série. Por outro lado, a concentração de poluentes é uma medida positiva e, em muitos casos, assimétrica e autocorrelacionada, tornando as suposições de normalidade e de independência muito restritivas. Os modelos lineares generalizados flexibilizam a suposição de normalidade e de homoscedasticidade, permitindo modelar também a dispersão dos dados, mas ainda estão restritos a observações independentes. Em geral, os efeitos temporais são representados por variáveis indicadoras para a hora

[1] Proporção de carros a gasolina entre os carros bicombustíveis.

[2] Consideraremos tanto séries de poluentes quanto dados epidemiológicos, como número de mortes ou casos de doenças associadas à poluição do ar.

do dia, dia da semana, semana do ano etc. Nessa abordagem, muitas vezes é difícil especificar quais termos devem ser incluídos no modelo e determinar se eles realmente controlam esses componentes.

Os modelos aditivos generalizados são uma boa alternativa nesse contexto. Eles permitem incluir termos não-paramétricos para ajustar uma curva suavizada da resposta em função do tempo, controlando os efeitos sazonais e de tendência, e um componente paramétrico para associar as variáveis explicativas à variável resposta. No entanto, a especificação do modelo pode ser complicada, pois a escolha dos parâmetros que controlam a suavização nem sempre é trivial. Grandes períodos sem observação também atrapalham o ajuste, pois a curva suavizada ao longo do tempo é considerada contínua durante todo o intervalo.

[REESCREVER PARTE DE SÉRIES TEMPORAIS]

[ESCREVER PARTE DE OUTROS TÓPICOS]

O principal objetivo desta tese é criar estratégias robustas para análise de dados de poluição do ar que considerem as seguintes situações:

1. presença de observações correlacionadas;
2. presença de dados faltantes; e
3. presença de variáveis com erro de medida.

Para isso, serão propostas generalizações dos modelos discutidos, como modelos com erro de medida e modelos espaço-temporais, ao lado de técnicas pouco ou nada? utilizadas neste contexto, como regularização, imputação e validação cruzada.

As estratégias serão avaliadas com a aplicação em conjuntos de dados reais e comparadas com as metodologias normalmente empregadas. Também deseja-se criar uma pacote no programa estatístico R para padronizar e facilitar a implementação dessas estratégias.

No Capítulo 2, serão discutidas com mais detalhes as metodologias que utilizam, respectivamente, modelos lineares normais, modelos lineares generalizados, modelos lineares aditivos e séries temporais. Os problemas da aplicação desses modelos a dados de poluição do ar serão apontados, assim como formas de contorná-los. Nos Capítulos 3, 4 e 5, serão apresentadas respectivamente as estratégias para dados correlacionados, dados faltantes e variáveis com erro de medida. No Capítulo 6, ilustraremos essas estratégias com aplicações a conjuntos de dados reais. No Capítulo 7, discutiremos os principais resultados encontrados neste trabalho.

Você acha realmente necessário esse parágrafo?

O texto a seguir busca um equilíbrio entre formalismo matemático, interpretação e aplicabilidade. O propósito dessa tentativa é produzir um trabalho acessível a pesquisadores de todas as áreas, tendo em vista os objetivos propostos. Apesar disso, um certo grau de conhecimento estatístico será exigido em muitos pontos. Hastie e Tibshirani (1990); James *et al.* (2013) são ótimas referências para consulta.

A parte computacional deste trabalho foi realizada integralmente no programa estatístico R (R Core Team, 2016).

Capítulo 2

Principais metodologias

2.1 Modelo linear

Dada uma variável de interesse Y e um vetor de variáveis explicativas $\mathbf{X}$, a modelagem estatística supervisionada (Hastie *et al.*, 2008) tem como objetivo encontrar funções f's tais que

$$Y \approx f(\mathbf{X}). \tag{2.1}$$

Essa relação implica que parte da variabilidade de Y pode ser explicada pelo componente sistemático $f(\mathbf{X})$. A partir desse modelo, é possível fazer predição — descobrir qual é o Y para um novo $\mathbf{X}$ — e inferência — investigar como $\mathbf{X}$ afeta Y.

O modelo linear (ou regressão linear múltipla) corresponde à aproximação (2.1) mais simples e bem estabelecida. Ele vem sendo o pilar da Estatística nas últimas 4 décadas e ainda hoje é muito utilizado. Sua popularidade é justificada por se ajustar bem a diversos problemas reais, pela facilidade de interpretação dos resultados e por estar disponível em praticamente qualquer programa estatístico.

Em estudos de poluição do ar, modelos lineares são geralmente ajustados para investigar 1) a possível associação de variáveis sob estudo com a concentração de poluentes; e 2) a relação entre a concentração de poluentes com casos de certas doenças e mortalidade. Mesmo em cenários complexos, nos quais estes modelos tendem a não ser a técnica mais adequada, eles costumam ser utilizados como uma primeira estratégia de análise, mais simples, precedendo o ajuste de modelos mais sofisticados. Saldiva *et al.* (1995), por exemplo, utilizou modelos lineares para estudar o efeito de alguns poluentes nas taxas de mortalidade de idosos, controlando por condições climáticas e sazonais, e mostrou que os resultados encontrados não diferiam muito de modelos Poisson e de modelos aditivos generalizados.

Em uma formulação mais precisa, a...

A expressão (2.1) ~~também~~ pode ser escrita como

$$Y = f(\mathbf{X}) + \epsilon, \tag{2.2}$$

em que ϵ representa um erro aleatório, isto é, um conjunto de fatores desconhecidos ou não-observados a que será atribuída a variabilidade de Y não explicada por $\mathbf{X}$. Apesar de (2.1) ser mais intuitiva, (2.2) é mais conveniente para a formulação teórica dos modelos estatísticos e será utilizada a partir deste ponto.

Neste capítulo, discutiremos como adaptar o modelo linear para o ajuste de dados de poluição do ar e por que as suas suposições são tão restritivas em variadas situações práticas.

2.1.1 O modelo

Uma característica correntemente presente em estudos de poluição atmosférica é a existência de fatores confundidores, isto é, variáveis explicativas cuja relação com a variável resposta é conhecida, mas que precisam ser controladas pelo modelo por explicarem uma porção importante da sua variabilidade. Alguns fatores de confundimento muito comuns são os climáticos, os sazonais e os geográficos. Neste trabalho, como tentativa de manter o objetivo das análises sempre em foco, as variáveis explicativas sob investigação serão denotadas como $\mathbf{X}$ e as variáveis confundidoras como $\mathbf{Z}$, sendo que a união desses dois conjuntos de variáveis compõem o $\mathbf{X}$ em (2.1).

Ficou um pouco confuso

Mais formalmente, seja Y_t a variável sob estudo, $\mathbf{X_t} = (X_{1t}, ..., X_{pt})$ um vetor de variáveis explicativas cuja associação com Y_t estamos interessados em avaliar, $\mathbf{Z_t} = (Z_{1t}, ..., Z_{st})$ um vetor de variáveis de confundimento cuja associação com Y_t é, a priori, conhecida e t o instante no qual essas variáveis foram medidas, com $t = 1, ..., n$. Aqui não são feitas suposições sobre a natureza dos preditores $\mathbf{X_t}$ e $\mathbf{Z_t}$, isto é, essas variáveis podem ser fixas ou aleatórias, qualitativas ou quantitativas. Dado os vetores de parâmetros desconhecidos $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ e $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_s)$, o modelo linear pode ser definido por

$$Y_t = \alpha + \beta_1 X_{1t} + ... + \beta_p X_{pt} + \gamma_1 Z_{1t} + ... + \gamma_s Z_{st} + \epsilon_t, \tag{2.3}$$

com média zero

Em geral, supõe-se que os erros $(\epsilon_1, ..., \epsilon_n)$ sejam independentes, homoscedásticos (variância constante) e, em certos casos, ~~com~~ normalmente distribuídos. Além disso, a especificação (2.3) impõe que a relação entre a resposta Y_t e os preditores $\mathbf{X_t}$ e $\mathbf{Z_t}$ é linear e aditiva.

mantidas as outras fixas

A suposição de linearidade estabelece que a variação esperada em Y_t causada pelo acréscimo de uma unidade em X_{it} é constante e igual a β_i, independentemente do valor de X_{it}. O mesmo vale para as variáveis $\mathbf{Z_t}$ e os coeficientes $(\gamma_1, ..., \gamma_s)$, mas, em geral, não há interesse em interpretá-los. Neste trabalho, a interpretação dos coeficientes será abordada com mais detalhes nas aplicações do Capítulo **??**. Conceitos mais gerais podem ser encontrados em Hastie *et al.* (2008) e James *et al.* (2013).

A suposição de aditividade implica que a variação esperada na resposta Y_t causada por uma mudança no preditor X_{it} (ou Z_{jt}) independe do valor dos outros preditores. Essa suposição pode ser relaxada com a introdução de termos de interação (ver Seção 3.3 de James *et al.* (2013)), cuja interpretação será discutida no Capítulo **??**.

Daqui em diante, quando não houver possibilidade de confusão ou ambiguidade, omitiremos o índice t em Y_t, $\mathbf{X_t}$ e $\mathbf{Z_t}$.

Os coeficientes $(\beta_1, ..., \beta_p)$, $(\gamma_1, ..., \gamma_s)$ podem ser estimados pelo método dos mínimos quadrados (Hastie *et al.*, 2008) ou da máxima verossimilhança (Sprott, 2000). Nesse contexto, as duas abordagens são equivalentes. O método de mínimos quadrados não exige a suposição de normalidade no processo de estimação, mas pode ser necessária para se fazer inferência~~s~~ sobre os parâmetros (intervalos de confiança e testes de hipóteses) ~~em amostras pequenas~~.

A partir de (2.3), para $t = 1, ..., n$, podemos definir os resíduos como

$$r_t = Y_t - \hat{Y}_t, \tag{2.4}$$

em que $\hat{Y}_t$ representa o valor predito de Y_t com base nas estimativas dos coeficientes do modelo. Os resíduos medem o quanto os valores preditos se afastam dos valores observados, sendo muito úteis para medir a qualidade do ajuste e avaliar se as suposições foram satisfeitas.

Como o instante em que as observações omissas ocorreram não é considerado pelo processo de estimação, o modelo linear é uma alternativa para ajustar séries com "buracos"ou com grandes períodos sem informação. No entanto, a identificação da estrutura de tendência e sazonalidade pode ser mais difícil em séries com essas características. Na Seção 2.1.2, será abordado como incluir esses componentes no modelo.

Não é melhor especificar antes que Y_t é a variáver resposta?

Em alguns casos, as variáveis respostas podem ser medidas em diferentes locais de uma região. Salvo e Geiger (2014), por exemplo, mediu a concentração de ozônio (e outras medidas meteorológicas e de tráfego) em diversas estações ao longo da cidade de São Paulo. Como esse componente espacial pode ser um fator de confundimento, ele deve ser considerado pelo modelo. Na Seção 2.1.3, será abordada a incorporação de componentes espaciais no modelo linear para controlar a associação entre observações coletadas em localidades próximas.

resposta?

Em geral, os estudos de poluição do ar consideram variáveis positivas, assimétricas, possivelmente correlacionadas e heteroscedásticas. Essas características podem infringir as suposições do modelo linear, causando problemas no ajuste. Técnicas de diagnóstico — em especial, a análise dos resíduos — são uma boa opção para verificar se as suposições estabelecidas estão sendo violadas. Na Seção 2.1.4, será discutido como utilizar essas técnicas para avaliar se o modelo está bem ajustado e, em caso negativo, maneiras de contornar esse problema.

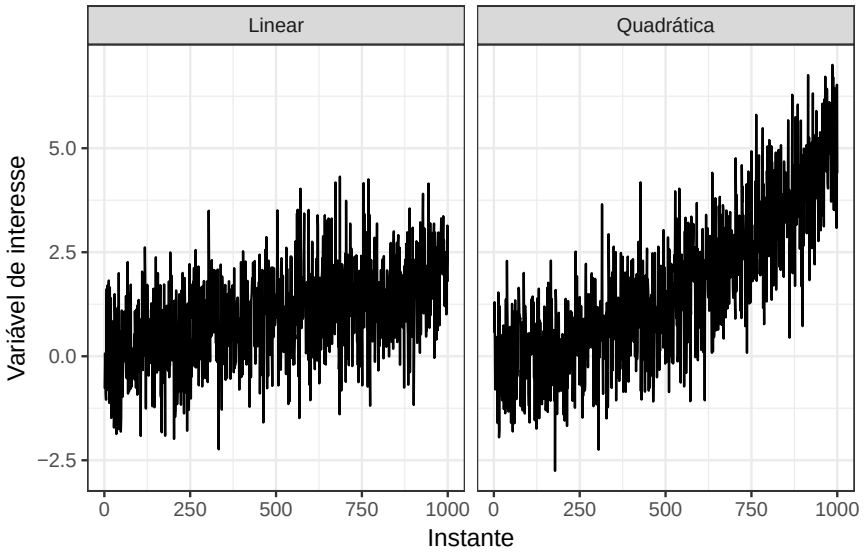
2.1.2 Incorporando tendência e sazonalidade

É natural pressupor que séries de variáveis ambientais ou epidemiológicas não sejam estacionárias, isto é, que as suas médias não são constantes ao longo do tempo. (EXEMPLO). Muitas vezes, essa tendência é causada por fatores que, por questões práticas, não podem ser controlados, como, por exemplo, crescimento populacional ou industrial, mudanças climáticas, aumento da frota de automóveis, novas leis de regulamentação de combustíveis, desenvolvimento de novos medicamentos, etc. Essa “tendência natural” precisa ser controlada pelo modelo, pois, caso contrário, o confundimento com outras variáveis pode gerar viés na estimação dos parâmetros.

O gráfico da variável resposta contra o tempo pode ser utilizado para investigar a estacionariedade de uma série. No entanto, a partir dele, não é possível avaliar se uma eventual tendência já está sendo controlada por uma das variáveis explicativas. Além disso, o efeito conjunto de duas ou mais variáveis pode mascarar a tendência da série, não sendo possível detectá-la pelo gráfico. Uma alternativa é construir o gráfico dos resíduos do modelo (2.3) contra o tempo. Como a variação dos resíduos representa toda a variação de Y não explicada pelas variáveis $\mathbf{X}$ e $\mathbf{Z}$, qualquer tendência dos resíduos ao longo do tempo indicaria uma tendência causada por fenômenos não observados. Por esse gráfico, também é possível julgar qual a melhor forma para o termo de tendência, isto é, linear, quadrática, logarítmica etc.

Para acrescentar um termo de tendência linear ao modelo (2.3), podemos especificar $Z_{1t} = t$, $t = 1, ..., n$. Assim, um coeficiente γ_1 positivo indica que a resposta cresce linearmente com o tempo,

enquanto um coeficiente negativo indica que a resposta decresce linearmente com o tempo. Podemos definir outras formas para a tendência, como quadrática, $Z_{1t} = t^2$, ou logarítmica, $Z_{1t} = log(t)$. A Figura 2.1 mostra um exemplo de séries com tendências linear e quadrática.



O que você acha de incluir uma linha de tendência em cada um desses gráficos para reforçar a tendência linear/ quadrática?

Figura 2.1: *Exemplos de séries com tendência linear e quadrática, ambas positivas.*

Uma restrição nessa abordagem é a homogeneidade da tendência ao longo do período observado. Em alguns casos, a tendência pode mudar em diferentes intervalos de tempo (Figura 2.2). Uma alternativa seria definir um termo de tendência para cada intervalo, por exemplo:

$$Z_{1t} = \begin{cases} t, & \text{se } t \text{ pertence ao conjunto } \{1, 2, ..., j\}; \text{ e} \\ 0, & \text{caso contrário.} \end{cases}$$

e

$$Z_{2t} = \begin{cases} t - j, & \text{se } t \text{ pertence ao conjunto } \{j+1, j+2, ..., n\}; \text{ e} \\ 0, & \text{caso contrário.} \end{cases}$$

Uma outra maneira de se controlar a tendência de uma série é transformar os dados originais para gerar uma série estacionária. A transformação mais comum é dada por

$$\Delta Y_t = Y_t - Y_{t-1}. \tag{2.5}$$

Caso o cálculo dessa primeira diferença não tenha sido o suficiente para alcançar a estacionariedade, diferenças de maior grau costumam ser tomadas, isto é,

$$\Delta^2 Y_t = \Delta[\Delta Y_t] = \Delta[Y_t - Y_{t-1}]$$

e, no caso geral, para $j < n$,

$$\Delta^j Y_t = \Delta[\Delta^{j-1} Y_t].$$

Usualmente, uma ou duas diferenças tornam a série estacionárias (Morettin e Toloi, 2004). Voltaremos a tratar de estacionariedade no Capítulo **??**.

Da mesma forma que a tendência, a média da variável de interesse pode variar segundo efeitos

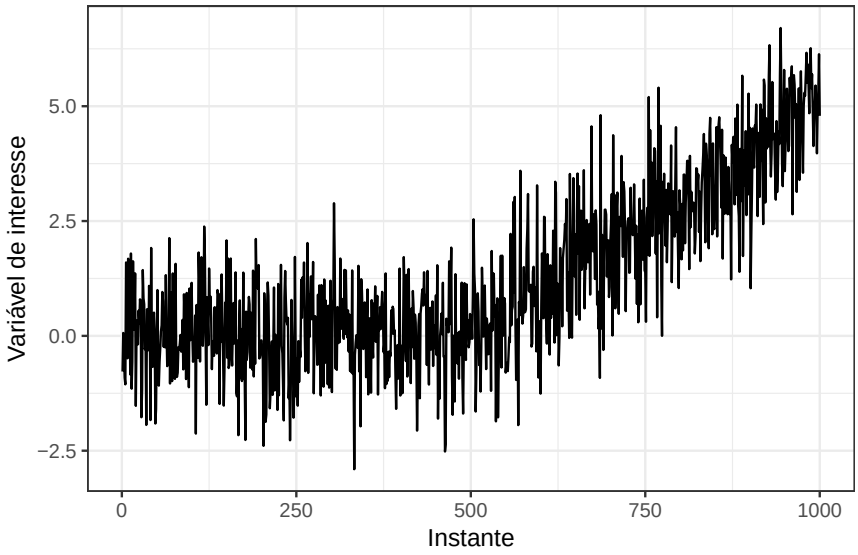


Figura 2.2: *Exemplos de uma série com tendência não-constante.*

que variam periodicamente dentro de um intervalo. Os níveis de ozônio, por exemplo, cresce no verão e diminui no inverno; o número de casos de problemas respiratórios tende a aumentar nos meses mais seco; a concentração de diversos poluentes são menores durante os fins de semana, devido ao menor tráfego. (EXEMPLO).

Esses padrões, conhecidos como sazonalidade, se repetem ao longo das semanas, meses e anos, e também precisam ser controlados pelo modelo. De uma maneira geral, podemos classificar a sazonalidade como *determinística* — o padrão é constante ao longo do tempo — ou *estocástica* — o padrão muda ao longo do tempo.

O gráfico da série é uma boa forma de se detectar a sazonalidade. Com uma análise descritiva adequada, é quase sempre possível identificar quando este componente é gerado por uma das variáveis explicativas. O gráfico dos resíduos contra o tempo do modelo (2.3) também pode ser usado para identificar a sazonalidade da série.

É possível controlar a sazonalidade determinística no modelo (2.3) a partir de variáveis indicadoras. Se, por exemplo, acredita-se haver um efeito sazonal de mês, pode-se então adicionar ao modelo 11 variáveis indicadoras Z_{it}, $i = 1, ..., 11$ tais que

$$Z_{it} = \begin{cases} 1, & \text{se a observação } t \text{ pertence ao } i\text{-ésimo mês do ano; e} \\ 0, & \text{caso contrário.} \end{cases} \tag{2.6}$$

Com essa formulação, o mês de dezembro será tomado como referência. Para mais informações sobre a utilização de variáveis indicadoras, consultar a Seção 3.3.1 de James *et al.* (2013).

Se a sazonalidade for estocástica, procedimentos um pouco mais sofisticados serão necessários para controlá-la. Discutiremos esses métodos no capítulo **??**.

2.1.3 Incorporando um componente espacial

2.1.4 Problemas com as suposições do modelo

A seguir, será discutido por que as suposições de linearidade, normalidade, independência e homoscedasticidade podem ser um problema para a utilização do modelo linear normal na análise

de dados de poluição e epidemiologia ambiental.

2.1.5 Linearidade

Não seria importante comentar que a linearidade diz respeito aos parâmetros?

O modelo linear normal assume que a relação entre a resposta e os preditores é uma reta. Se a verdadeira relação tiver outra forma (polinomial ou exponencial, por exemplo), as conclusões feitas a partir do modelo podem ser inválidas. Gráficos são uma boa maneira de identificar relações não-lineares entre as variáveis.

Gráficos de dispersão dos preditores contra a resposta (quando ambos são quantitativos) podem ser utilizados como primeiro recurso. Se existir associação entre as variáveis, espera-se que a nuvem de pontos siga uma reta.

Os resíduos definidos pela expressão (2.4) também podem ser utilizados para essa tarefa. A ideia é, após o ajuste do modelo, construir o gráfico dos resíduos contra os valores preditos ($\hat{Y}$) e avaliar se a nuvem de pontos apresenta algum padrão. Nuvens em forma de "U", por exemplo, são fortes indicativos de não-linearidade (veja Figura 2.3).

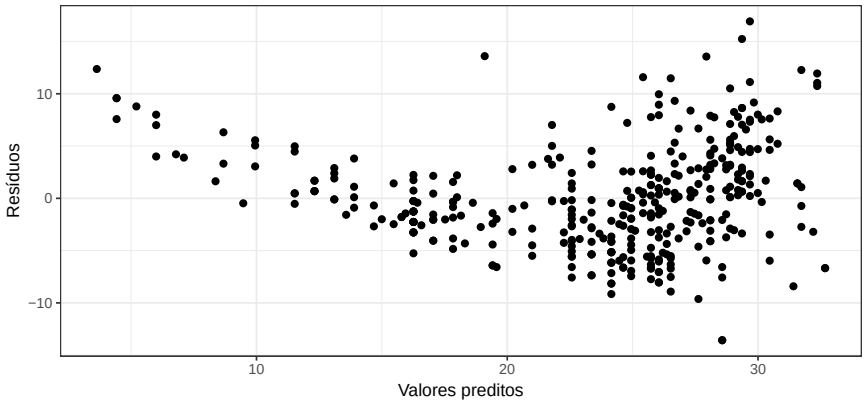


Figura 2.3: *Gráfico dos resíduos contra os valores preditos. Exemplo de nuvem de pontos com forma "U" indicando não-linearidade.*

Existem algumas maneiras de contornar problemas com a suposição de linearidade do modelo (2.3). A mais simples consiste em aplicar transformações nos preditores, tais como $\log X$, $sqrt(X)$ e X^2. Quando há interesse inferencial, é preciso ter cuidado com a nova interpretação dos parâmetros após aplicar essas transformações. Mais informações podem ser encontradas em (REF) .

Outra forma de lidar com a não-linearidade é considerar modelos polinomiais. Esses modelos estendem o modelo linear utilizando variáveis do tipo X, X^2, X^3 como preditores.

Uma terceira maneira utiliza funções escadas para representar o preditor cuja associação com a resposta é supostamente não-linear. Essa variável é dividida em M categorias, transformando-se em uma variável qualitativa. Isso é equivalente a ajustar uma função constante em cada subintervalo.

2.1.6 Normalidade

A distribuição Normal assume que a variável aleatória está definida na reta real, isto é, pode assumir valores positivos e negativos, e é simétrica em relação a sua média. Intuitivamente, não há motivos para acreditar que essa distribuição se ajustaria bem a dados naturalmente positivos e assimétricos, como a concentração de poluentes e número de casos de uma doença ou mortalidade. Mesmo assim, pela facilidade de implementação e interpretação, o modelo normal é muitas vezes

utilizado como uma primeira tentativa na estratégia de análise, de tal forma que, se os dados não se afastarem muito desta distribuição, pequenos vieses podem ser negligenciado em favor de um modelo mais simples.

Como mencionado anteriormente, o método de mínimos quadrados não supõe normalidade no processo de estimação. Apesar de essa suposição ser feita na construção de intervalos de confiança e testes de hipóteses para os parâmetros, para amostras grandes[1], a teoria assintótica garante estimadores com distribuição aproximadamente normal (REF) . No entanto, como a velocidade dessa convergência depende da natureza (desconhecida) da variável sob estudo, não temos como quantificar o que é uma "amostra grande", o que justifica a importância de se avaliar a qualidade do ajuste em relação à suposição de normalidade.

Uma primeira ideia para checar essa suposição seria construir o histograma da variável resposta e avaliar se ele se aproxima de uma distribuição normal. Contudo, a suposição de normalidade se refere à variável $Y|\mathbf{X}, \mathbf{Z}$ e não a Y simplesmente. Seria necessário construir histogramas para cada combinação de valores de $\mathbf{X}$ e $\mathbf{Z}$, o que é quase sempre impossível.

Gráficos Q-Q (quantil-quantil) e envelopes (REF) são técnicas muito úteis para investigar possíveis desvios da suposição de normalidade. Eles se baseiam nos resíduos studentizados (REF) e avaliam o quanto a distribuição empírica desses resíduos se afasta da distribuição teórica. (EXEMPLO)

A transformação da variável Y pode ser uma alternativa para casos em que a suposição de normalidade não é válida. As transformações $\log Y$ e $\sqrt{Y}$ são as mais utilizadas. Na Seção 2.2, veremos como generalizar o modelo linear para outras distribuições.

2.1.7 Independência

O modelo linear normal assume que as variáveis $Y_1, ..., Y_n$ são independentes, isto é, saber que Y_i assume um valor alto não traz informação sobre o valor de Y_{i+1}. Em estudos de poluição do ar, essa suposição é, na maioria dos casos, inadequada. Se a concentração de poluente está alta em um certo dia, esperamos encontrá-la alta também no dia seguinte. A mesma interpretação pode ser feita para o número de casos de uma determinada doença.

Como as estimativas do modelo são calculadas com base na suposição de independência, variáveis $Y_1, ..., Y_n$ muito correlacionadas podem levar a estimativas enviesadas e conclusões equivocadas. Mais especificamente, se a amostra apresentar correlação, os erros-padrão estimados tenderão a subestimar os verdadeiros erros, o que comprometeria toda a inferência feita a partir deste modelo, já que os valores p associados seriam menores do que deveriam ser.

Uma forma de avaliar a quebra dessa suposição é construir um gráfico dos resíduos do modelo pelo tempo. A presença de padrões na sequência de pontos, isto é, resíduos adjacentes tendem a apresentar valores próximos, é um indício de correlação.

? Variáveis indicadoras a unidade de tempo em que as medidas foram realizadas (hora, dia da semana, semana do mês etc) podem ajudar a reduzir o efeito da correlação.

[1] Comum em estudos de poluição do ar.

Anteriormente você usou erros-padrão

2.1.8 Homoscedasticidade

A partir da primeira expressão em (??), verificamos que a mesma variância σ^2 é atribuída a todas as variáveis $Y_t|\mathbf{X_t}, \mathbf{Z_t}$, $t = 1, ..., n$. Essa é a suposição de homoscedasticidade, e a sua violação também pode afetar o cálculo dos erros-padrões, resultando em inferências não-confiáveis. ??

Para séries de tempo, não é raro encontrarmos variâncias não-constantes ao longo de grandes períodos. Da mesma forma que algum fator altera a média de uma variável, ele pode também aumentar ou diminuir a sua variância. Uma nova regulamentação, por exemplo, pode limitar a quantidade de um determinado componente nos combustíveis, diminuindo a variação das concentrações de um determinado poluente gerado pela interação deste componente com outros compostos.

em função dos

Os gráficos dos resíduos pelos valores preditos são uma boa ferramenta para identificar a quebra dessa suposição. Nuvens de pontos em forma de funil são indícios de heteroscedasticidade.

Uma maneira de controlar a variância de observações heteroscedásticas é transformar a variável Y. As transformações $\log Y$ e $\sqrt{Y}$ são, em geral, boas tentativas. Uma outra alternativa consiste em ponderar as observações com pesos proporcionais ao inverso da sua variância, mas se limita aos casos em que essa medida pode ser estimada com precisão.

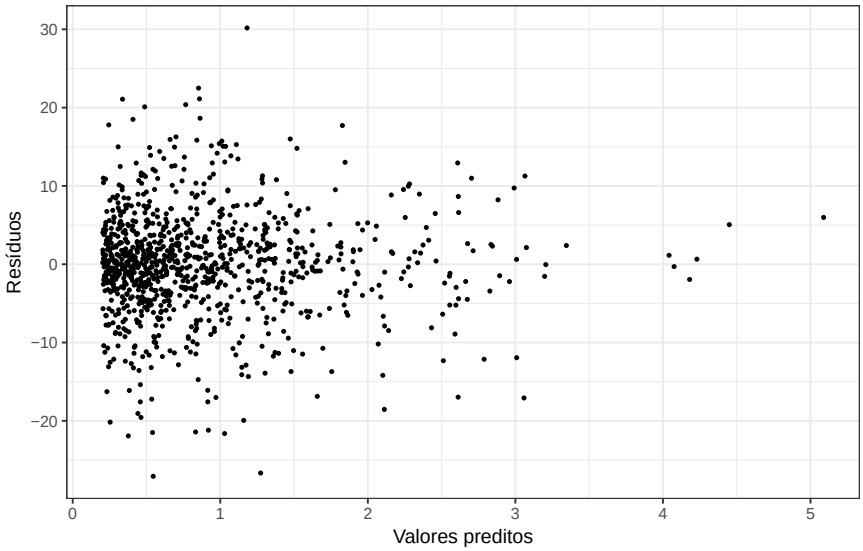


Figura 2.4: *Gráfico dos resíduos contra os valores preditos. Exemplo de nuvem de pontos em forma de funil, indicando heteroscedasticidade.*

2.2 Modelos lineares generalizados

O modelo linear definido em (2.3) tem duas importantes restrições:

repetitivo

i. para se fazer inferência, principalmente em amostras pequenas, a variável Y precisa ter distribuição Normal; e

ii. a variância de Y é considerada constante durante todo o período de observação.

Formalmente, o modelo admite que $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$, $t = 1, ..., n$, isto é, os erros aleatórios são normalmente distribuídos e homoscedásticos.

Como discutido anteriormente, essas suposições podem não ser compatíveis com dados de poluição atmosférica. A concentração de um poluente é uma medida positiva, muitas vezes assimetricamente distribuída e a sua variabilidade por não ser constante ao longo do tempo. Já o número de casos de uma doença ou a mortalidade são variáveis de contagem, isto é, assumem valores positivos discretos. A distribuição Normal pode não se ajustar bem a essas variáveis, principalmente quando o tamanho amostral é pequeno. Sendo assim, é importante flexibilizar o modelo para contemplar distribuições mais adequadas a dados dessa natureza, sem perder a sua interpretabilidade. Além disso, há casos em que a variância da variável resposta está associada a uma ou mais variáveis explicativas, exigindo um modelo que permita ajustar também a dispersão da variável.

Os modelos lineares generalizados foram introduzidos por (Nelder e Wedderburn, 1972) e englobam uma série de modelos de regressão. Eles são uma das principais alternativas para o ajuste de dados não-normais, permitindo a utilização de distribuições como a Gama e a Log-normal para dados assimétricos, a Poisson e a Binomial negativa para dados de contagem e Binomial para dados categorizados. Neste capítulo, será discutido como utilizar esta classe de modelos para o ajuste de dados de poluição do ar.

2.2.1 O modelo

Sejam Y_t, X_t e Z_t, $t = 1, ..., n$, definidos assim como na Seção 2.1.1. O modelo linear generalizado pode ser definido como

$$Y_t|\mathbf{X_t}, \mathbf{Z_t} \overset{ind}{\sim} \mathcal{D}(\mu_t, \phi)$$

$$g(\mu_t) = \alpha + \beta_1 X_{1t} + ... + \beta_p X_{pt} + \gamma_1 Z_{1t} + ... + \gamma_s Z_{st}, \tag{2.7}$$

em que $\mathcal{D}$ é uma distribuição pertencente à família exponencial[2], μ_t é um parâmetro de posição e ϕ é um parâmetro de precisão. Se $\mathbf{D}$ é a distribuição Normal, (2.7) se resume ao modelo linear (2.3).

Os parâmetros deste modelo podem ser estimados por máxima verossimilhança, e os cálculos envolvem o uso de procedimentos iterativos, como o Newton-Raphson e escore de Fisher (Dobson, 1990).

A especificação dos termos de tendência e sazonalidade podem ser feita da mesma forma que no modelo linear, assim como a inclusão de componentes espaciais.

A utilização dos resíduos em técnicas que avaliam a qualidade do ajuste pode ser conduzida de forma análoga à apresentada no Capítulo 2.1. Todavia, sem a suposição de normalidade, as propriedades do resíduo padronizado já não são mais verificadas. O resíduo mais utilizado em modelos lineares generalizados são definidos a partir da função desvio. Para mais informações, consultar (REF) . Outras técnicas de diagnóstico, como a detecção de pontos de alavanca e influência local, também foram generalizadas para esta classe de modelos. Para mais detalhes, veja (REF) .

Os modelos Gama e Log-normal são usualmente utilizados para ajustar dados positivos assimétricos, sendo, em geral, mais adequados para concentrações de poluentes do que a distribuição Normal. Esse tópico será discutido na Seção 2.2.2.

Dados de contagem, como o número de casos de uma doença ou mortalidade, são usualmente ajustados pelo modelo Poisson. Conceição *et al.* (2001b), por exemplo, utilizaram esse modelo para

[2] A família exponencial corresponde a uma classe de distribuições de probabilidade que, sob certas condições de regularidade, apresentam algumas características em comum. Para mais informações, consulte (REF) .

avaliar a associação entre poluição atmosférica e marcadores de mortalidade em idosos na cidade de São Paulo. No entanto, a distribuição Poisson impõe que a média e a variância das observações são iguais e pode não se ajustar bem quando os dados apresentam sobredispersão (variância maior que a média). O modelo com resposta binomial negativa é uma alternativa nesses casos, já que permite a modelagem conjunta dos parâmetros de posição e dispersão. Esses pontos serão melhor discutidos na Seção 2.2.5.

A primeira expressão em (2.7) impõe que as variáveis sejam homoscedásticos e independentes. A suposição de homoscedasticidade pode ser relaxada modelando-se a dispersão dos dados, como será visto na Seção 2.2.8. Uma extensão para dados correlacionados é dada pelas equações de estimação generalizadas, que serão abordadas na Seção 2.2.9.

2.2.2 Modelos para dados assimétricos

2.2.3 Modelo Gama

2.2.4 Modelo Log-normal

2.2.5 Modelos para dados de contagem

2.2.6 Modelo Poisson

2.2.7 Modelo Binomial Negativa

2.2.8 Modelo linear generalizado duplo

2.2.9 Equações de estimação generalizadas

2.3 Modelos aditivos generalizados

repetitivo

Modelos lineares têm um papel muito importante na análise de dados, provendo técnicas de predição e inferência computacionalmente simples e fáceis de serem interpretadas. Contudo, a suposição de linearidade é quase sempre uma aproximação, o que se torna um grande problema se levarmos em conta que, na vida real, o efeito dos preditores na variável resposta frequentemente não é linear. Em estudos de poluição do ar especificamente, a interferência do tempo nos dados gera efeitos sazonais, cuja relação com a variável reposta é, em geral, muito melhor representada por curvas senoidais do que por retas.

No Capítulo 2.1, foram introduzidas algumas técnicas que estendem o modelo linear na tentativa de modelar relações não-lineares, como a transformação de variáveis, a regressão polinomial e as funções escada. Outras, como regressão por splines, splines de alisamento e regressão local, também podem ser empregadas neste contexto. No entanto, os modelos aditivos generalizados estendem esses métodos, permitindo o ajuste de múltiplos preditores. (REF) (EXEMPLO)

Neste Capítulo, os modelos aditivos generalizados serão introduzidos como um método integrado, automático e flexível para identificar e caracterizar a não-linearidade de preditores em estudos de poluição atmosférica.

2.4 O modelo

Modelos aditivos generalizados são uma extensão do modelo linear generalizado que permite associar cada um dos proditores à variável resposta a partir de funções não-lineares, mantendo a suposição de aditividade. Formalmente, como nos capítulos anteriores, sejam Y_t, $\mathbf{X}_t$ e $\mathbf{Z}_t$, respectivamente, a variável resposta, as variáveis explicativas e as variáveis confundidoras medidas nos instantes $t = 1, ..., n$. O modelo aditivo generalizado pode ser escrito como

$$Y_t|\mathbf{X_t}, \mathbf{Z_t} \overset{ind}{\sim} \mathcal{D}(\mu_t, \phi)$$

$$g(\mu_t) = \beta_0 + f_1(X_{1t}) + ... + f_p(X_{pt}) + f_{p+1}(Z_{1t}) + ... + f_{p+s}(Z_{st}), \tag{2.8}$$

em que $\mathcal{D}$ é uma distribuição pertencente à família exponencial. No caso mais simples, assim como nos modelos lineares generalizados, supõe-se que as variáveis Y_t são homoscedásticas, independentes e normalmente distribuídas.

O ajuste dos modelos aditivos generalizados consiste em incorporar um método de alisamento ao procedimento de estimação de um modelo linear generalizado. Para mais informações, consulte Hastie e Tibshirani (1990) ou Hastie *et al.* (2008).

A expressão (2.8) considera que a associação com todos os preditores será estimada a partir de uma função alisada. No entanto, em muitos casos pode ser interessante considerar o alisamento de apenas algumas variáveis explicativas, enquanto outra são tratadas com a estrutura paramétrica. Em estudos de poluição, costuma-se alisar o tempo e, em alguns casos, as variáveis climáticas. (EXEMPLO). Esta abordagem semiparamétrica será formulada na Seção 2.5.

Existem diversas propostas a respeito de como f deve ser representada, incluindo o uso de splines naturais, splines penalizados, splines suavizados e loess (Hastie e Tibshirani, 1990). Outra questão importante é sobre o quão suave f deve ser. Vários métodos para determinar a suavidade ótima de f já foram desenvolvidos (Hastie *et al.*, 2008). Na Seção 2.6, serão abordadas as principais técnicas.

repetitivo

Um ponto importante no ajuste de modelos lineares generalizados é a escolha de quais variáveis serão alisadas. Em estudos de poluição atmosférica, o tempo se torna uma opção óbvia para o alisamento, mas nem sempre fica claro se as variáveis confundidoras, como temperatura, umidade, congestionamento etc, devem ou não ser suavizadas. (EXEMPLO). Esse tema será discutido na Seção 2.7.

Nesta abordagem, a escolha dos parâmetros de suavização é crucial. Em geral, os algorítimos de estimação permitem que os próprios dados escolham os parâmetros que melhor se ajustem aos dados. No entanto, há casos em que a estrutura dos dados (sazonal por exemplo) pede que esses valores sejam fixados. Uma visão geral deste tópico será apresentada na Seção 2.8.

Na Seção 2.9, serão discutidos os problemas gerados pelos dados omissos no processo de estimação, assim como formas de contornar esse problema.

2.5 Modelo semiparamétrico

2.6 Métodos de alisamento

2.7 Quais variáveis devem ser suavizadas?

2.8 Os parâmetros de suavização

2.9 Dados omissos

2.10 Séries temporais

Referências Bibliográficas

Beer *et al.*(**2011**) Tom Beer, John Carras, David Worth, Nick Coplin, Peter K. Campbell, Bin Jalaludin, Dennys Angove, Merched Azzi, Steve Brown, Ian Campbell, Martin Cope, Owen Farrell, Ian Galbally, Stephen Haiser, Brendan Halliburton, Robert Hynes, David Jacyna, Melita Keywood, Steven Lavrencic, Sarah Lawson, Sunhee Lee, Imants Liepa, James McGregor, Peter Nancarrow, Michael Patterson, Jennifer Powell, Anne Tibbett, Jason Ward, Stephen White, David Williams e Rosemary Wood. The health impacts of ethanol blend petrol. *Energies*, (4): 352–367. Citado na pág. 2

Belusic *et al.*(**2015**) Andreina Belusic, Ivana Herceg-Bulic e Zvjezdana Bencetic Klaic. Using a generalized additive model to quantify the influence of local meteorology on air quality in zagreb. *Geofizika*, 32: 48–78. Citado na pág. 2

Carslaw *et al.*(**2007**) David C. Carslaw, Sean D. Beevers e James E. Tate. Modelling and assessing trends in traffic-related emissions using a generalised additive modelling approach. *Atmospheric Environment*, 41: 5289–5299. Citado na pág. 1, 3

Chang *et al.*(**2017**) Shih Ying Chang, William Vizuete, Marc Serre, Lakshmi Pradeepa Vennam, Mohammad Omary, Vlad Isakov, Michael Breen e Saravanan Arunachalam. Finely resolved on-road PM2.5 and estimated premature mortality in central north carolina. *Risk Analysis*. doi: 10.1111/risa.12775. URL http://dx.doi.org/10.1111/risa.12775. Citado na pág. 2

Conceição *et al.*(**2001a**) Gleice M.S. Conceição, Simone G.E.K. Miraglia, Humberto S. Kishi, Paulo Hilário Nascimento Saldiva e Julio da Motta Singer. Air pollution and child mortality: a time-series study in São Paulo, Brazil. *Environmental Health Perspectives*, 109(3): 347–350. Citado na pág. 2, 3

Conceição *et al.*(**2001b**) Gleice M.S. Conceição, Paulo Hilário Nascimento Saldiva e Julio da Motta Singer. Modelos MLG e MAG para análise da associação entre poluição atmosférica e marcadores de morbi-mortalidade: uma introdução baseada em dados da cidade de São Paulo. *Revista Brasileira de Epidemiologia*, 4(3): 206–219. Citado na pág. 2, 3, 13

Dobson(1990) A. J. Dobson. *An introduction to generalized linear models*. Chapman and Hall, New York. Citado na pág. 13

European Commission(1999) European Commission. EU focus on clean air. *Office for Official Publications of the European Communities*. Citado na pág. 1

European Commission(2011) European Commission. Climate action. https://ec.europa.eu/clima/policies/strategies/2050_en, 2011. [Online; acessado 15-03-2017]. Citado na pág. 2

Hastie e Tibshirani(1990) Trevor Hastie e Robert Tibshirani. *Generalized additive models*. London:Chapman & Hall. Citado na pág. 4, 15

Hastie *et al.*(**2008**) Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer. Citado na pág. 5, 6, 15

Jacobson(2007) Mark Z. Jacobson. Effects of ethanol (E85) versus gasoline vehicles on cancer and mortality in the united states. *Environmental Science & Technology*, 41(11): 4150–4157. Citado na pág. 2

James *et al.***(2013)** Gareth James, Daniela Witten, Trevor Hastie e Robert Tibshirani. *An Introduction to Statistical Learning*. Springer Series in Statistics. Springer, New York. Citado na pág. 4, 6, 9

Jasarevic *et al.***(2014)** Tarik Jasarevic, Glenn Thomas e Nada Osseiran. 7 million premature deaths annually linked to air pollution. http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/, 2014. [Online; acessado 13-03-2017]. Citado na pág. 1

Katsouyanni *et al.***(1996)** K. Katsouyanni, J. Schwartz, C. Spix, G. Touloumi, D. Zmirou, A. Zanobetti, B. Wojtyniak, J. M. Vonk, A. Tobias, A. Pönkä, S. Medina, L. Bachárová e H. R. Anderson. Short term effects of air pollution on health: a european approach using epidemiologic time series data: the aphea protocol. *Journal of Epidemiology & Community Health*, 50(Suppl 1): S12–S18. ISSN 0143-005X. doi: 10.1136/jech.50.Suppl_1.S12. URL http://jech.bmj.com/content/50/Suppl_1/S12. Citado na pág. 2, 3

Kloog *et al.***(2012)** Itai Kloog, Francesco Nordio, Brent A. Coull e Joel Schwartz. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal PM2.5 exposures in the mid-atlantic states. *American Chemical Society*, 46: 11913–11921. Citado na pág. 2

Lin *et al.***(1999)** C. A. Lin, M. A. Martins, S. C. Farhat, C. A. Pope, G. M. Conceição, V. M. Anastácio, M. Hatanaka, W. C. Andrade, W. R. Hamaue, G. M. Bohm e P. H. Saldiva. Air pollution and respiratory illness of children in São Paulo, Brazil. *Paediatric and Perinatal Epidemiology*, 13(4): 475–488. ISSN 1365-3016. doi: 10.1046/j.1365-3016.1999.00210.x. URL http://dx.doi.org/10.1046/j.1365-3016.1999.00210.x. Citado na pág. 2, 3

Morettin e Toloi(2004) Pedro A. Morettin e Clelia M.C. Toloi. *Análise de Series Temporais*. ABE - Projeto Fisher e Editora Edgard Blucher, São Paulo. Citado na pág. 8

Mulawa *et al.***(1997)** Patricia A. Mulawa, Steven H. Cadle, Kenneth Knapp, Roy Zweidinger, Richard Snow, Randy Lucas e Joseph Goldbach. Effect of ambient temperature and E10 fuel on primary exhaust particulate matter emissions from light-duty vehicles. *American Chemical Society: Environ. Sci. Technol.*, 31 (5): 1302–1307. Citado na pág. 2

Nelder e Wedderburn(1972) J. A. Nelder e R. W. M. Wedderburn. Generalized linear models. *Stat Soc A*, 135: 370–384. Citado na pág. 13

Pereira *et al.***(2004)** Pedro Afonso Pereira, Leilane Maria B. Santos, Eliane Teixeira Sousa e Jailson B. de Andrade. Alcohol- and gasohol-fuels: a comparative chamber study of photochemical ozone formation. *Journal of the Brazilian Chemical Society*, 15(5): 646–651. Citado na pág. 2

R Core Team(2016) R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/. Citado na pág. 4

Saldiva *et al.***(1994)** P. H. N. Saldiva, A. J. F. C. Lichtenfels, P. S. O. Paiva, I. A. Barone, M. A. Martins, E. Massad, J. C. R. Pereira, V. P. Xavier, J. M. Singer e G. M. Bohm. Association between air pollution and mortality due to respiratory diseases in children in São Paulo, Brazil: a preliminary report. *Environmental Research*, 65(2): 218 – 225. ISSN 0013-9351. doi: http://dx.doi.org/10.1006/enrs.1994.1033. URL http://www.sciencedirect.com/science/article/pii/S0013935184710334. Citado na pág. 2, 3

Saldiva *et al.***(1995)** Paulo H. N. Saldiva, C. Arden Pope, Joel Schwartz, Douglas W. Dockery, Ana Julia Lichtenfels, Joao Marcos Salge, Ivana Barone e Gyorgy Miklos Bohm. Air pollution and mortality in elderly people: a time-series study in São Paulo, Brazil. *Archives of Environmental Health: An International Journal*, 50: 159–163. Citado na pág. 2, 3, 5

Salvo e Geiger(2014) Alberto Salvo e Franz M. Geiger. Reduction in local ozone levels in urban São Paulo due to a shift from ethanol to gasoline use. *Nature Geoscience*, 7: 450–458. Citado na pág. 2, 3, 7

Schwartz e Dockery(1992) J. Schwartz e D. W. Dockery. Particulate air pollution and daily mortality in Steubenville, Ohio. *Am J Epidemiol.*, 1(135): 12–19. Citado na pág. 2, 3

Schwartz *et al.***(1996)** J. Schwartz, D. W. Dockery e L. M. Neas. Is daily mortality associated specifically with fine particles? *J Air Waste Manag Assoc*, 10(46): 927–939. Citado na pág. 2, 3

Schwartz(1994) Joel Schwartz. Nonparametric smoothing in the analysis of air pollution and respiratory illness. *Statistical Society of Canada*, 22(4): 471–487. Citado na pág. 2, 3

Schwartz(1996) Joel Schwartz. Air pollution and hospital admissions for respiratory disease. *Epidemiology*, 1(7): 20–28. Citado na pág. 2, 3

Schwartz e Marcus(1990) Joel Schwartz e Allan Marcus. Mortality and air pollution in London: a time series analysis. *American Journal of Epidemiology*, 131(1): 185. doi: 10.1093/oxfordjournals.aje.a115473. URL +http://dx.doi.org/10.1093/oxfordjournals.aje.a115473. Citado na pág. 2, 3

Shumway e Stoffer(1982) R. H. Shumway e D. S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(4): 253–264. ISSN 1467-9892. doi: 10.1111/j.1467-9892.1982.tb00349.x. URL http://dx.doi.org/10.1111/j.1467-9892.1982.tb00349.x. Citado na pág. 2, 3

Sprott(2000) D. A. Sprott. *Statistical Inference in Science*. Springer Series in Statistics. Springer. Citado na pág. 6

Yoon *et al.***(2009)** S. H. Yoon, S. Y. Ha, H. G. Roh e C. S. Lee. Effect of bioethanol as an alternative fuel on the emissions reduction characteristics and combustion stability in a spark ignition engine. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 223: 941–951. Citado na pág. 2