

Tempo, tempo, tempo: regressão de Cox na ciência política

Rodrigo Lins - Universidade Federal de Sergipe

Resumo

Este artigo tem como objetivo apresentar o modelo de riscos proporcional de Cox. Para tanto, será feito um debate teórico geral dos modelos de sobrevivência, assim como a apresentação de termos comuns para a técnica. Além disso, será feita uma apresentação passo a passo de como elaborar um modelo de Cox no software R. Espera-se que este trabalho ajude na difusão da técnica dentro das ciências sociais.

Palavras-chave: Análise de Sobrevivência; Modelo de Cox; R.

Abstract

This article aims to present Cox Proportional-Hazards Model. To do so, a general theoretical debate on survival models will be made, as well as the presentation of common terms for the technique. In addition, there will be a step-by-step presentation of how to develop a Cox model in the R software. It is hoped that this work will help to disseminate the technique within the social sciences.

Keywords: Survival Analysis; Cox Model; R.

1. Introdução¹

A análise de sobrevivência é uma técnica de vários nomes: análise de sobrevida; análise de evento histórico; modelo de duração; e tempo até evento (*time-to-event*) são apenas alguns dos nomes dados a essa ferramenta estatística. No entanto, apesar da variação de nome, todos eles apresentam uma característica chave: o tempo. Assim, não é surpresa que a preocupação dos modelos de sobrevivência não seja só o “se”, mas também “quando” um dado evento² acontece (GUO, 2010). Isto é: o interesse do pesquisador é a observação de uma mudança qualitativa da variável dependente em um ponto específico no tempo (ALLISON, 1984).

Dentro das possíveis abordagens de uma análise de sobrevivência, o modelo de Cox de riscos proporcionais tem lugar de destaque (BOX-STEFFENSMEIER & JONES, 2004). Conceituado como um modelo semiparamétrico³, ele é bastante maleável. Ele não assume uma distribuição estatística. Assume apenas que a ação das variáveis independentes sobre o risco é multiplicativa (CARVALHO ET AL., 2011), isto é, para cada incremento de unidade da variável explicativa, o risco é multiplicado pelo seu coeficiente exponenciado (ALLISON, 1984).

E como anda a utilização da análise de sobrevivência – e mais especificamente, do modelo de Cox – na ciência política brasileira? É possível notar que o uso ainda não decolou. Analisando os resumos de artigos disponíveis nos arquivos de cinco das principais revistas brasileiras

de ciência política⁴, só houve um retorno que indicasse o uso de uma análise de sobrevivência⁵. Assim, este trabalho pretende apresentar a técnica ao grande público das ciências sociais para que seja mais considerada.

O restante deste artigo está dividido da seguinte forma: primeiro focarei no modelo de Cox, com abordagem mais teórica; em seguida, apresentarei um breve glossário de termos comuns nas análises de sobrevivência, mas que não costumam ser tão utilizados em outras técnicas estatísticas. A seção seguinte apresentará as possíveis aplicações em diferentes áreas do conhecimento. Finalmente, na última etapa, apresentarei um passo a passo de como realizar uma análise de sobrevivência no software R e como interpretar os seus resultados.

2. O modelo de Cox: o que é, para que serve, quando usar?

O que é o modelo de riscos proporcionais de Cox? Antes, é preciso entender que as modelagens em análise de sobrevivência podem ser divididas em dois grupos: paramétricos e o semiparamétrico⁶. Os paramétricos são aqueles que pressupõem uma determinada distribuição, como o modelo de Weibull ou seu formato exponencial (BOX-STEFFENSMEIER & JONES, 2004). Enquanto a última espera que o risco seja constante no tempo, a de Weibull permite que o risco aumente ou diminua linearmente no tempo (ALLISON, 1984).

1 Agradeço aos pareceristas anônimos que me ajudaram a melhorar substancialmente este artigo. Erros remanescentes são de responsabilidade exclusiva do autor.

2 Na literatura de análise de sobrevivência, a o termo “evento” também pode ser chamado de “falha”. No entanto, como a mudança qualitativa nem sempre indica uma falha de fato, podendo ser apenas uma transição de governo, por exemplo, optei por sempre usar “evento”.

3 Modelos paramétricos e semiparamétricos são explicados na próxima seção.

4 A DADOS foi a única que apresentou um retorno, em artigo publicado em 2019. As outras quatro revistas incluídas na busca foram: Brazilian Political Science Review; Opinião Pública; Revista Brasileira de Ciências Sociais; e Revista de Sociologia e Política. Os termos de busca foram dois: Cox e sobrevivência. A pesquisa foi feita em 2 de junho de 2020.

5 Isso também pode indicar que os resumos dos artigos não apresentam claramente o desenho de pesquisa e, assim, não antecipam as ferramentas analíticas utilizadas.

6 Ainda é possível encontrar estimativas não-paramétricas. No entanto, não é possível estimar o efeito das variáveis explicativas nesses modelos. O estimador produto de Kaplan-Meier, por exemplo, compara a curva de sobrevivência de dois ou mais grupos, não considerando variáveis independentes. Ver Lee e Wang (2003), especialmente os capítulos 4 e 5.

Já o modelo semiparamétrico – o modelo Cox – não assume qualquer distribuição para a função de risco basal. Essa seria uma característica que poderia o colocar em um grupo não-paramétrico. No entanto, é esperado que suas variáveis ajam sempre com efeito multiplicativo sobre o risco. E esta é a característica que o credenciaria para o grupo paramétrico. É por isso, portanto, que o modelo de Cox é considerado um modelo semiparamétrico.

Respondendo, então, de forma mais direta o questionamento do início desta seção, podemos afirmar que o modelo de Cox: (1) assim como modelos paramétricos, permite a inclusão de variáveis explanatórias; (2) o efeito destas variáveis é multiplicativo; e (3) ele não carrega nenhum pressuposto no que diz respeito à distribuição estatística para a função de risco basal. O modelo de Cox é, então, uma técnica estatística para explorar a relação de um evento específico com uma série de variáveis exploratórias (WALTERS, 2009).

Ele foi desenvolvido por Cox (1972), de quem herdou o nome, e surgiu como forma de flexibilizar os pressupostos já mencionados nos modelos paramétricos. É chamado de modelo de riscos proporcionais porque para quaisquer duas observações em qualquer ponto no tempo, a razão de seus riscos deve ser constante. Em outras palavras, a razão dos riscos de uma observação A para uma observação B deve ser o mesmo tanto em t_0 quanto em t_{100} , por exemplo (ALLISON, 1984). Pode ser simbolizado pela equação a seguir:

$$\log h(t) = a(t) + b_1x_1 + b_nx_n \quad (1)$$

É possível notar que o lado esquerdo da equação é o logaritmo natural da função de risco. Do outro lado, a função de risco basal $a(t)$ pode ser qualquer função de tempo. Como já vimos, é justamente por não precisar ser especificado que o modelo de Cox é tido como parcialmente paramétrico ou semiparamétrico (ALLISON, 1984). Em b_1x_1 , b representa o coeficiente e x o

valor da variável independente 1. A fórmula (1) é aplicada ao tipo de dados que será visto na tabela 1, em que cada observação ocupa uma linha, pois as variáveis X_1 e X_n são constantes no tempo.

Modelo de Cox estendido⁷

É comum que quando se trabalhe com dados longitudinais, algumas das variáveis explicativas variem de acordo com o tempo. Digamos que o PIB per capita seja uma das variáveis independentes. É difícil imaginar que ela se mantenha constante com o passar dos anos. Dessa forma, como incluí-la em um modelo? É justamente nestes casos, quando existem variáveis dependentes do tempo (ou que variam no tempo), que o modelo de Cox estendido deve ser usado. Pode ser representado da seguinte forma:

$$\log h(t) = a(t) + b_1x_1 + b_nx_n(t) \quad (2)$$

Aqui, $a(t)$ segue podendo ser qualquer função de tempo e a variável X_1 segue constante no tempo. No entanto, a variável X_n varia, sim, em função do tempo. Para esse tipo de desenho, o pesquisador deve construir um banco de dados pelo processo de contagem, similar ao que será visto na tabela 2, onde uma nova linha é adicionada sempre que houver mudança nos valores de X_n .

Os resíduos de Schoenfeld

Existem diferentes formas de testar o ajuste do modelo de uma regressão de Cox. No entanto, considerando que a proporcionalidade dos riscos é uma das características definidoras do modelo, é importante avaliar se tal pressuposto está sendo cumprido. Os resíduos de Schoenfeld servem justamente para essa finalidade⁸.

⁷ Vale ressaltar que além de sua forma estendida, o modelo de Cox também se permite ser estratificado. Para uma visão mais completa do modelo de Cox estratificado, ver capítulo 6 de Carvalho et al. (2011).

⁸ Existem outras formas de analisar a proporcionalidade do risco. Um caminho comum é via resíduos martingale (ver Lin et al., 1993). Ainda existem as possibilidades de resíduos deviance e resíduos score (ver Carvalho et al., 2011).

Ele apresenta a proporcionalidade de cada covariável definido para cada indivíduo (CARVALHO ET AL., 2011). A equação que ilustra o cálculo dos resíduos é a seguinte:

$$r_{ik} = \delta_i(x_{ik} - a_{ik}) \quad (3)$$

onde o δ_i a ocorrência da falha para a observação i é simbolizada por δ_i . O valor da variável específica se dá por x_{ik} e, finalmente, a média ponderada das covariáveis em risco em um determinado tempo t é a_{ik} .

Embora existam diferentes formas de analisar os resíduos, a maneira mais intuitiva de compreender a análise, como argumentado por Schoenfeld (1982), é a de que os resíduos devem ser independentes do tempo. Assim, se uma correlação entre os resíduos e o tempo não for estatisticamente significativa, isso quer dizer que os coeficientes não são dependentes do tempo e, assim, o pressuposto foi observado. Caso seja estatisticamente significativa, é indicação de que a proporcionalidade dos riscos não foi atendida.

Como será visto na seção prática, a análise dos resíduos de Schoenfeld também pode ser vista graficamente, facilitando sua interpretação. Como é esperado que a correlação seja aleatória, a linha do gráfico de dispersão deve seguir uma reta horizontal. Caso exista alguma tendência, positiva ou negativa, a proporcionalidade dos riscos foi ferida.

Em caso de não observância da proporcionalidade de risco, o primeiro passo é verificar se isso acontece de forma natural ou artificial. Assim, é importante verificar a presença de *outliers*. Caso a quebra de proporcionalidade seja real, duas abordagens são possíveis: (1) buscar um novo modelo, como o modelo de Cox estratificado (ATA & SÖZER, 2007; BORUCKA, 2013); ou (2) adicionar um efeito de interação entre a

variável que fere o pressuposto de proporcionalidade e alguma função de tempo⁹ (BOX-STEFFENSMEIER & ZORN, 2001).

3. Pequeno glossário de sobrevivência

Antes de seguir adiante, é importante que seja feito um breve elucidário. Como os modelos de sobrevivência lidam com vocábulos que são pouco usuais a outras técnicas estatísticas, julgo que seja justificado apresentar ao menos os principais termos. Os conceitos apresentados nessa seção são válidos não só para os modelos de Cox, mas para análise de sobrevivência em geral. Esse pequeno glossário não tem intenção de ser exaustivo. Isso implica dizer que só serão apresentados aqui os termos mais comuns e relevantes para a compreensão dos modelos de sobrevivência¹⁰. Para começar, tais modelos podem lidar com dois tipos de dados incompletos: censurados e trucados.

Censurados/censored: Uma observação é considerada censurada quando seu início ou fim não são observados, fazendo com que seja difícil determinar um ponto específico no tempo. Existem três diferentes tipos de censura: (1) censura à direita; (2) censura à esquerda; e (3) censura intervalar. A primeira acontece quando a observação chega ao fim do estudo sem experimentar o evento ($T > t^-$)¹¹. Assim, o pesquisador sabe apenas que a observação sobreviveu até determinado momento. Mas não sabe quando,

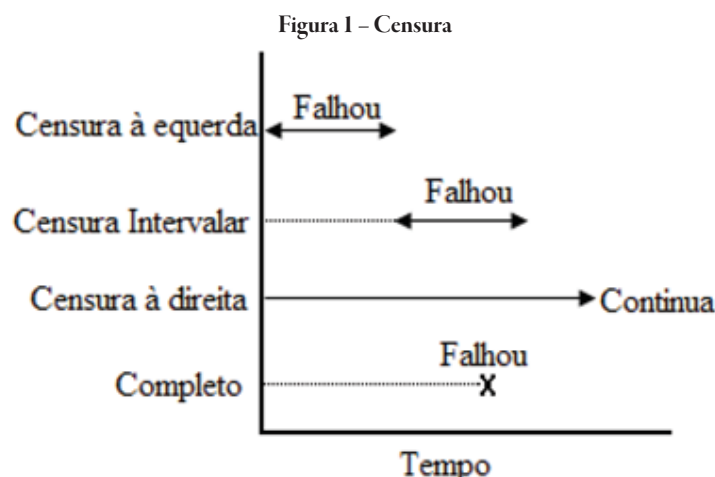
9 Em geral, é utilizado o logaritmo natural. Ver Box-Steffensmeier e Zorn (2001).

10 Os termos serão apresentados em português, com seu correspondente em inglês. O único objetivo é facilitar o estudo de quem quer trabalhar com análise de sobrevivência, evitando que o leitor se perca na tradução de termos técnicos.

11 T representa o tempo de sobrevivência ou tempo até o desfecho (CARVALHO et al., 2011). Ele é o intervalo entre a inserção da observação no estudo até o momento que ele experimenta a falha ou é censurado. t^- indica que o tempo de estudo é menor do que a sobrevivência da observação e $t+$ simboliza que o tempo de estudo supera o momento em que a observação foi censurada.

ou se chegou a experimentar o evento (GUJARATI, 2011). Dizemos que os dados têm censura a esquerda quando não sabemos quando o evento aconteceu, mas sabemos que ocorreu antes de um dado momento ($T < t_+$). Finalmente, a censura intervalar

($t^- < T < t_+$) indica que o evento aconteceu dentro de um período de tempo, mas o pesquisador não sabe exatamente quando (CARVALHO ET AL., 2011; KLEINBAUM & KLEIN, 2005). A figura 1 ilustra esses três tipos de censura. Teste



Fonte: Lins et al. (2017).

Existem ainda outros três tipos de censura: de tipo I, tipo II e aleatória, comuns em estudos clínicos (COLOSIMO & GIOLO, 2006). A censura tipo I, de maneira similar à censura à direita, acontece quando o estudo é encerrado após um tempo pré-estabelecido e, neste momento, algumas observações ainda não tinham sofrido o evento de interesse. A censura de tipo II ocorre quando o estudo é encerrado depois que um número pré-estabelecido de eventos acontece. Finalmente, a censura de tipo aleatório é aquela em que uma observação é retirada do estudo sem que a falha tenha acontecido (COLOSIMO & GIOLO, 2006).

Truncados/truncated: Os dados também podem ser truncados à direita e à esquerda. Truncamento à direita acontece quando são selecionados casos que sofreram o evento. Isso implica dizer que todas as observações sofreram o evento de interesse antes do limite final do estudo. Já o truncamento à esquerda conta com os casos que sofreram o evento após o limite inferior da janela de estudo (CARVALHO ET AL., 2011). Vale desta-

car também as observações com entradas atrasadas. São os casos que entraram no banco de dados uma vez que a pesquisa já havia iniciado. Esses casos costumam ser erroneamente tratados também como truncamento à esquerda (HARRELL, 2015).

Outros termos relevantes para a compreensão dos modelos de sobrevivência são:

Função de sobrevivência/survival function - $S(t)$: É a probabilidade de que uma observação sobreviva mais tempo do que um algum período específico t . Ela é representada pela seguinte fórmula:

$$S(t) = P(T > t) \quad (4)$$

Função de risco/hazard function - $h(t)$: A função de risco apresenta o potencial de falha em cada unidade de tempo, considerando que a observação sobreviveu até o tempo sendo observado t . Diferentemente da função de sobrevivência, ela não é uma probabilidade, e sim

uma taxa, podendo assumir qualquer valor maior do que zero. Pode ser representado como:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad (5)$$

Em termos visuais, a função de sobrevivência apresenta uma tendência de queda, indicando que a probabilidade de sobreviver cai à medida que o tempo passa (KLEINBAUM & KLEIN, 2005). Já a função de risco pode ser constante ou aumentar/diminuir com o tempo. A função de risco deve ter mais relevância porque os modelos de regressão em uma análise de sobrevivência analisam justamente o risco de um dado evento acontecer. Isso ficará claro mais à frente, quando trataremos especificamente dos modelos de Cox.

Verossimilhança parcial/partial likelihood. Os coeficientes dos modelos de Cox são estimados a partir de uma verossimilhança parcial¹². Diferentemente do formato tradicional¹³, a função do risco basal¹⁴ (*baseline hazard*) é retirada do cálculo para o formato parcial por não ser parametrizado (BOX-STEFFENSMEIER & JONES, 2004). São consideradas apenas as informações das observações que estão sob risco a cada tempo t .

$$PL = \prod_{j=1}^K L_j \quad (6)$$

onde K é o número total de eventos na amostra e L_j denota a verossimilhança de falhar no tempo j , dado que a observação sobreviveu até este ponto (KLEINBAUM & KLEIN, 2005).

Estrutura dos dados. Finalmente, também é importante falar sobre a estrutura dos dados de uma análise de sobrevivência. A construção correta do banco desempenha papel importante na hora da execução da análise. Quatro variáveis são obrigatórias ou esperadas¹⁵. A variável *início* e a variável *fim* indicam, respectivamente, o momento em que a observação entrou na pesquisa e o momento em que sofreu o evento ou foi censurado. Além disso, é recomendável a inserção da variável *tempo*, que indica justamente o período em que a observação esteve presente no banco. De forma simples, ela é a subtração da variável *fim* pela variável *início*. Finalmente, a variável binária *status*, que indica se a observação sofreu o evento ou não. Dentro do modelo tradicional, que é o utilizado em uma regressão de Cox, cada observação ocupará apenas uma linha dentro do banco. A tabela abaixo ilustra um banco tradicional. Ela indica um estudo hipotético que começa no tempo 0 e termina no tempo 30.

12 Uma apresentação básica de máxima verossimilhança é dada por Crawley (2015): “given the data, and given the model, what values of the parameters of that model make the observed data most likely?” (CRAWLEY, 2015, p. 6). No entanto, por ser considerado um método “intrincado” (GUJARATI & PORTER, 2011), não me aprofundarei aqui nos conceitos de verossimilhança.

13 Verossimilhança usual é utilizada no cálculo do coeficiente para os modelos paramétricos de sobrevivência. Ver apêndice A de Allison (1984) e capítulo 3 de Box-Steffensmeier e Jones (2004).

14 A função basal, ou função de base, é assim chamada pois nos casos em que as variáveis independentes têm valor 0, o valor da função de risco é igual ao valor da função basal (COLLETT, 1994).

15 O número exato das variáveis necessárias varia de acordo com o modelo. Para trabalhar com o modelo tradicional de Cox no software R, apenas a variável tempo e status são necessárias. Já para trabalhar com o modelo estendido, são necessárias início, fim e status. Por isso, o recomendável é que todas as quatro variáveis estejam presentes.

Tabela 1 – Banco de dados

ID	Início	Fim	Tempo	Status	X1	X2
A	0	30	30	0	1	1
B	0	27	27	1	1	0
C	1	15	14	1	0	0
D	2	30	28	0	1	1
E	3	15	12	1	0	1

Fonte: autor.

Pode-se perceber que temos cinco observações e uma linha para cada. A primeira observação faz parte do estudo desde o começo, pois seu início é 0. Chegou ao tempo 30 e não sofreu o evento, pois seu status também é 0. Já a terceira observação entrou no tempo 1 do estudo e saiu no tempo 15, tendo passado um total de 14 unidades de tempo na pesquisa. Sabemos que saída se deu por ter sofrido o evento, como indicado pela variável status de valor 1. As observações 1 e 4 servem de exemplo de censuras à direita, pois chegaram ao fim do estudo sem ter experimentado o

evento. As variáveis X1 e X2 da tabela 1 representam variáveis explicativas¹⁶ com valores dicotômicos e constantes no tempo.

Veremos mais adiante que o banco exemplificado na tabela 1 serve para as regressões tradicionais de Cox. No entanto, se entre as variáveis explicativas houver alguma que varie no tempo, a estrutura deve ser modificada. Neste caso, a observação ganhará uma nova linha sempre que o valor da variável independente mudar. A tabela 2 ilustra tal situação.

Tabela 2 – Banco de dados
(Variáveis dependente do tempo)

ID	Início	Fim	Status	X1	X2
A	0	11	0	1	100
A	11	22	1	1	120
B	0	15	0	0	40
B	12	23	0	0	42
B	23	30	0	0	47

Fonte: autor

Mais uma vez, a tabela indica um estudo hipotético com duração de 30 anos. Pela variável status, sabemos que a observação 1 sofreu o evento enquanto a observação 2 é censurado à direita. A variável independente X1 continua sendo dicotômica e constante no tempo. Mas agora a variável X2 apresenta mudanças. E a cada mudança, uma nova linha é inserida. Para observação 1, ela apresenta apenas uma alteração de valor durante a estada

da observação no banco. Por isso, só uma linha a mais é inserida. Já para a observação 2, existem duas mudanças no valor de X2. Dessa forma, duas novas linhas são inseridas. A quebra sempre acontece no tempo em que a alteração ocorre. Uma outra diferença importante é que a variável tempo não faz mais parte do banco. Ela

¹⁶ Variáveis explicativas são aquelas teorizadas como causa de variação na variável dependente (KELLSTEDT & WHITTEN, 2018).

contabilizaria cada quebra de linha como a inserção de uma nova observação, zerando o valor sempre que isso acontecesse. Dessa forma, ela é excluída do banco¹⁷. Esse banco é construído pelo chamado processo de contagem.

Existe ainda uma outra forma de estruturar os dados, caso o pesquisador esteja lidando com eventos recorrentes ou múltiplos. Neste caso, uma vez que uma determinada observação experimenta o evento de interesse, ele deve sair do banco de dados, voltando uma vez que retorna ao grupo de risco. Na ciência política, um exemplo claro – que apresentarei em maior detalhe

na seção prática – se dá nos estudos de transição de regime. Se o evento de interesse for a mudança de regime democrático para regime autocrático, o país sairá do banco assim que sofrer o revés institucional e permanecerá fora do grupo de risco enquanto for uma ditadura, pois não há como ele sofrer o evento novamente. No entanto, caso o país transite de volta para a democracia, ele é recolocado no grupo de risco e entra novamente no banco de dados. A tabela a seguir mostra como exemplo um excerto de um banco de dados com informações para a Argentina, construído a partir dos dados de Cheibub, Gandhi e Vreeland (2010).

Tabela 3 – Banco de dados
(Eventos recorrentes)

ID	Início	Fim	Status
ARG	1946	1955	1
ARG	1958	1962	1
ARG	1963	1966	1
ARG	1973	1976	1
ARG	1983	2008	0

Fonte: autor.

Percebe-se que entre 1966 e 1973 e entre 1976 e 1983, por exemplo, a Argentina esteve fora do banco de dados. Isso aconteceu porque durante esses períodos o país era uma ditadura. Dessa forma, não poderia estar em risco de se tornar um país autoritário. Ainda que tenha acontecido golpes dentro de golpes durante o período, todos os eventos ocorreram dentro de regimes autoritários. A tabela mostra que sempre que a observação sofre o evento, uma nova linha é criada. Mas ele só retorna ao banco se as condições iniciais forem restabelecidas. Também é importante ressaltar que para que os anos fossem apresentados da maneira que estão, o banco de dados deveria começar no ano 0.

¹⁷ A rigor, ela pode ser mantida. Mas o pesquisador precisa ficar atento ao fato de que ela contabilizaria o tempo de permanência da observação apenas para os intervalos em que não há alteração na variável que tem seu valor alterado. Para que isso não aconteça, o pesquisador precisa fazer as alterações “manualmente”. No entanto, como será visto na seção prática, ela pode ser descartada por completo por não ser necessária na construção de um modelo de Cox estendido no R.

Tendo compreendido o que são dados censurados e truncados, o que são as funções de sobrevivência e de risco, e visto como os dados devem ser estruturados, passo para a próxima seção: modelos de regressão de Cox.

4. Um modelo eclético: aplicações em diferentes áreas

Modelos de análise de sobrevivência são bastante ecléticos no que diz respeito às disciplinas em que podem ser aplicados. Apesar do conteúdo poder ser bastante diferente ao cruzar fronteiras de conhecimento, a técnica em si se adapta bem. Para ver como ela funciona em diferentes disciplinas, apresento trabalhos em três áreas diferentes: medicina, economia e ciência política.

Nesta última área, por ser o foco deste trabalho, apresentarei três exemplos. Procurei apresentar a hipótese e o modelo utilizado para testá-la. Por ser o foco deste trabalho, todos utilizaram o modelo de Cox. O quadro no fim da seção sumariza as variáveis de tempo e de status de cada estudo.

Medicina

A própria terminologia do nome da técnica já indica que ela surgiu na área de saúde. Especificamente, na biomedicina. O intuito inicial era estudar o tempo de sobrevivência de pacientes entre o momento do diagnóstico de uma dada doença até a morte do indivíduo (GUO, 2010). Assim, não é surpresa que o número de pesquisas que fazem uso da técnica na área de medicina seja elevado (COLOSIMO & GIOLO, 2006; MACHIN, CHEUNG & PARMAR, 2006).

Para exemplificar a utilização do modelo de Cox na área, vamos olhar o trabalho de Khambaty, Callahan e Stewart (2018). Os autores partem do pressuposto de que existe uma relação entre depressão e diabetes. Mais especificamente, afirmam que depressão poderia atuar como um fator de risco independente para incidência de diabetes. Dessa forma, eles levantam a seguinte pergunta: ao tratar pacientes para depressão também, os médicos também estariam reduzindo o excesso de risco de incidência de diabetes. Para tanto, eles testaram o impacto que um programa de cuidados colaborativos com duração de 12 meses teria sobre um grupo de idosos deprimidos e inicialmente livres de diabetes. O estudo fez *follow-up* por nove anos e contou com 119 pacientes.

É possível perceber, portanto, a variável que indica falha: desenvolvimento da doença. E ela é instrumentalizada com 0 se o paciente não desenvolver a doença e 1 caso desenvolva. A variável independente também é binária, indicando se o participante fez parte do

grupo de tratamento (1) ou do grupo de controle (0). Os autores também rodaram um segundo modelo, incluindo algumas variáveis pessoais como controles: idade, sexo, raça/etnia, histórico de hipertensão, de fumo, e o índice de massa corporal (IMC). Os resultados, no entanto, indicaram que apesar do grupo de tratamento indicar uma queda no nível de depressão, quando comparado com o controle, os grupos não divergiram em relação à incidência de diabetes. Assim, os autores argumentam que tratamento para depressão não é, por si só, suficiente para diminuir o risco dos pacientes desenvolverem a doença. Como uma das variáveis de controle tem seus valores alterados no tempo (IMC), os autores utilizaram um modelo de Cox estendido, como visto na fórmula (5).

Economia

A análise de sobrevivência em estudos econômicos pode ter várias aplicações. Desde estudos sobre falência de empresas (LEE, 2014) a pesquisas sobre estabilidade de bancos (PAPPAS ET AL., 2017), duração de desemprego (BIESZK-STOLORZ & MARKOWICZ, 2015) e processo de gentrificação (GRUBE-CAVERS & PATTERSON, 2015). No entanto, irei focar no trabalho de Gury (2011). Ele analisa os determinantes da taxa de abandono dos estudantes de ensino superior na França, i.e., o tempo de duração dos alunos no ensino superior. Assim, a variável status já pode ser desenhada: 0 se os estudantes não abandonam o ensino superior e 1 se abandonam. Para tanto, Gury (2011) utilizou dados do Ministério da Educação da França para acompanhar 5.383 estudantes que ingressaram no ensino médio em 1989 e, posteriormente, se matricularam em uma instituição de ensino superior.

Com o objetivo de explicar o comportamento dos estudantes, ele utiliza três grandes grupos de variáveis independentes: a de características pessoais (sexo e educação formal dos pais, por exemplo); as relacio-

nadas à educação secundária das observações (o tipo de diploma obtido¹⁸, por exemplo); e relacionadas ao ensino superior (se o estudante declarou ter algum projeto profissional).

Para testar as hipóteses, Gury (2011) desenvolve três modelos, indo do modelo mais simples (com apenas o primeiro grupo de variáveis independentes) ao mais completo (com todos os três grupos de variáveis explicativas). Como resultado, o autor encontra um padrão não-linear, indicando que existem dois grupos de estudantes. Os chamados *early leavers*, que abandonam o curso nos dois primeiros anos; e os *late leavers*, que abandonam a partir do terceiro ano. Entre os que sofrem o evento – abandono do ensino superior –, as mulheres tendem a sair já no primeiro ano enquanto os homens abandonam a universidade já no fim do curso. Outras características socioeconômicas, como o nível de educação formal dos pais, só têm efeito no início dos estudos. À medida que os anos passam, as diferenças tendem a desaparecer. Nesse estudo, as variáveis não mudam com o tempo, proporcionando ao autor a possibilidade de trabalhar com o modelo tradicional de Cox, como visto na fórmula (4).

Ciência Política

Para o terceiro exemplo da utilização de modelos de sobrevivência, apresento um estudo na ciência política. Vale destacar que a utilização de tais modelos na disciplina vem ocupando um espaço costumeiramente povoado por modelos logit e probit. Várias áreas dentro da disciplina servem como terreno fértil para os modelos de duração. Por exemplo, eles já foram usados em estudos de transição de regime político (SVO-LIK, 2008); para estudo de sobrevivência de governos (DIERMEIER & STEVENSON, 1999; SOMER-

-TOPCU & WILLIAMS, 2008); e a análise de difusão de políticas públicas (JONES & BRANTON, 2005; BATISTA, 2017).

O primeiro (1) exemplo que vamos olhar é o trabalho de Tansey, Koehler e Schmotz (2017), que estudam o efeito de ligações autocráticas e a sobrevivência de regimes ditatoriais. A hipótese é apresentada da seguinte maneira: quanto maior for o nível de ligações internacionais autocráticas, menor será o risco de quebra do regime autocrático. Tansey et al. (2017) instrumentalizam quatro dimensões da ligação autocrática. Elas são: comércio, migração, ligações diplomáticas e proximidade geográfica. Já a variável status é representada por (0) se um regime democrático sobreviveu ou (1) se quebrou e se tornou democrático. Entre os controles, estão PIB per capita, crescimento do PIB e uma variável *dummy* para indicar se era um período de Guerra Fria ou não. Os dados dos autores incluem 250 regimes autocráticos, entre 1949 e 2008. A variável tempo é ano, assim como nos exemplos anteriores.

A hipótese nula, de que as ligações autocráticas internacionais não tinham efeito na duração de regimes autocráticos, pôde ser rejeitada. Assim, o risco de sofrer uma transformação institucional é dirimido à medida que o grau de relações entre estados autocráticos se fortalecem. Levando em consideração que o trabalho utilizou o PIB per capita e o crescimento do PIB como variáveis de controle, os autores utilizaram o modelo de Cox estendido, como na fórmula (5).

O segundo (2) exemplo é o trabalho de Shipan e Shannon (2003). Os autores analisam o tempo que leva entre a indicação de um candidato à Suprema Corte americana e a confirmação do candidato pelo senado norte-americano. A variável que indica status, portanto, é: 0 quando a confirmação não ocorre e 1 quando ela acontece. O tempo, diferente dos trabalhos vistos anteriormente nessa seção, é mensurado em dias. Fazendo uso

18 De acordo com o autor, existem três tipos de diplomas possíveis no ensino secundário francês: o tecnológico, o profissional e o geral. O autor afirma que os detentores de diploma geral têm tendência maior a concluir o ensino superior, já que o diploma geral do ensino secundário teria baixo valor no mercado de trabalho.

da teoria espacial, a hipótese principal do trabalho é o de que quanto maior for a distância ideológica entre o presidente e o senado, maior também vai ser o tempo entre indicação e confirmação.

Apesar de a distância ideológica ser a principal hipótese que os autores pretendem testar, eles também consideram características do candidato, como: se serve ou já serviu no senado ou se tem experiência prévia com tribunais a nível estadual ou federal. Os resultados encontrados por Shipan e Shannon confirmam a hipótese levantada. Maior distância ideológica aumenta o tempo até a confirmação da indicação feita pelo presidente. Os autores também acharam resultados estatisticamente significantes para a experiência do candidato no senado. E como as variáveis explicativas não variam com o tempo, o modelo utilizado foi o Cox tradicional.

Finalmente, o terceiro (3) exemplo é o trabalho de Crespo-Tenorio, Jensen e Rosas (2014). O objetivo dos autores é compreender o impacto que crises bancárias têm sobre a chance de sobrevivência de incumbentes. A hipótese principal dos autores é a de que as crises bancárias terão impactos negativos sobre os incumbentes em regimes democráticos. O impacto tende a ser ainda maior em países que possuem poucos atores de veto. O evento de interesse, portanto, é a falha na manutenção de um mesmo partido no poder. De forma mais direta: quando eleições levam um outro partido ao poder. O tempo aqui volta a ser mensurado anualmente e a hipótese dos autores também se confirma.

Por fim, o quadro sumariza aspectos dos três estudos vistos aqui.

Quadro 1 – Resumo dos estudos

Disciplina	Variável status	Variável tempo	Variação no tempo	Modelo
Medicina	Desenvolver doença	Anos	Sim	Cox estendido
Economia	Abandonar faculdade	Anos	Não	Cox tradicional
Ciência Política (1)	Queda de regime	Anos	Sim	Cox estendido
Ciência Política (2)	Confirmar indicação	Dias	Não	Cox tradicional
Ciência Política (3)	Derrota do incumbente	Anos	Sim	Cox estendido

Fonte: autor.

Quatro dos cinco estudos analisados aqui utilizam o ano como mensuração do tempo. Três deles possuem variáveis independentes ou de controles cujo valores variam no tempo. Isso implica dizer que eles utilizaram o modelo de Cox estendido e, assim, construíram bancos com a estrutura de contagem, como visto no exemplo da tabela 2. Nesses três casos, a construção

estatística do modelo segue a fórmula (5). O estudo de economia e de ciência política sobre confirmação das indicações à Suprema Corte, que não possuíam variável tempo-dependente, utilizaram o modelo tradicional de Cox. Dessa forma, seus bancos devem possuir uma estrutura similar à exemplificada na tabela 1, sendo realizado via fórmula (4).

5. Como fazer? Um passo a passo computacional¹⁹

Após a abordagem teórica dos modelos de sobrevivência²⁰ e depois de apresentar três exemplos de utilização desses modelos, chegou a hora de um exemplo prático. Vamos seguir o passo a passo computacional com enfoque no *software* R²¹. O exemplo que utilizarei trata da sobrevivência de regimes democráticos e será elaborado a partir do banco de dados do projeto *Varieties of Democracy* (V-Dem)²². A análise focará na América Latina e terá um corte temporal de 1900 a 2016, e trabalhará com seguinte questão de pesquisa: qual o impacto da institucionalização de partidos políticos sobre a sobrevivência de regimes democráticos?

Duas variáveis são obrigatórias em qualquer regressão com modelos de sobrevivência: o tempo de sobrevivência e o evento (status). Esta última é, necessariamente, uma variável dicotômica, em que 1 indica a ocorrência do evento estudado e, por exclusão, 0 significa a manutenção do status. Já a variável tempo indica o tempo desde ou o início do estudo (ou inclusão da observação) até que uma das duas situações ocorra: (1) a observação experimentar o evento ou (2) o período do estudo

chegar ao fim. Qualquer mensuração de tempo (dias, meses ou anos, por exemplo) pode ser utilizada. Aqui, utilizaremos ano.

Já temos, portanto, clara a construção da variável dependente. Seguindo a explicação de Guo (2010), temos as duas informações necessárias: a de caráter dicotômico (a democracia sobreviveu (0) ou sofreu revés institucional (1)); e a variável contínua representando tempo. No entanto, o modelo possui variáveis tempo-dependentes. Assim, o banco é construído no formato de contagem, com uma variável indicando o início do período e outra indicando o fim, para a elaboração de uma regressão do modelo de Cox estendido.

Finalmente, a última característica do modelo é a de que os países podem entrar e sair do grupo de risco. Isso acontece porque uma democracia, ao se tornar uma ditadura, sai do grupo de risco (e, consequentemente, do banco de dados). No entanto, não existe impedimento para que esse mesmo país volte a ser uma democracia. Caso isso aconteça, ele volta ao grupo de risco formado por países democráticos que podem, eventualmente, se tornar ditaduras (voltando também ao banco de dados).

Dessa forma, o desenho final é:

Quadro 2 – Desenho do modelo

Modelo	Evento recorrente	Tempo	Evento	Unidade
Cox estendido	Sim	Ano	Quebra de regime	País-ano

Fonte: autor.

¹⁹ Com o objetivo de ser transparente e replicável, o script da análise está disponível no OSF (<https://osf.io/xp6mg/>) e no GitLab. Como o único banco utilizado foi o do V-Dem, recomendo que o download seja feito diretamente na página da iniciativa. Ver nota 21.

²⁰ Para uma visão teórica mais ampla de análise de sobrevivência, é possível aproveitar o material disponível on-line de uma disciplina pela professora Ronghui Xu em: <http://www.math.ucsd.edu/~rxu/math284/>.

²¹ O software R não é o único que realiza tal análise. Existe, por exemplo, vasto material sobre como realizar a análise de sobrevivência em STATA ou SPSS.

²² A versão do V-Dem utilizada neste trabalho é a 10, lançada em março de 2020. O download pode ser feito em: <https://www.v-dem.net/en/data/data/v-dem-dataset/>

Tendo compreendido a estrutura do modelo, apresento agora as variáveis incluídas. Elas são:

Quadro 3 – Variáveis

Variável	No banco
Regime político	democracy
Início	start
Fim	stop
Institucionalização partidária	v2xps_party
Presidencialismo	v2xnp_pres
PIB per capita (ln)	e_migdppln

Fonte: autor.

As três primeiras linhas são os componentes que formam a variável dependente. O regime político é a variável de status, indicando quando um país é democrático (0) ou quando sofre o evento (1). Ela é elaborada pelo pesquisador a partir de uma variável já existente no V-Dem (v2x_regime). A variável original é construída a partir da medida contínua de democracia do V-Dem. Lührmann, Lindberg e Tannenber (2017), antes de construir quatro categorias de regime, primeiro os divide em dois grandes grupos: democracias e não-democracias. Essa é a divisão utilizada aqui para a variável dicotômica de status.

Como estou lidando com variáveis que mudam no tempo, preciso construir também as variáveis início e fim. O início indica o começo do intervalo de tempo a partir do qual a variável assume um determinado valor; por sua vez, o fim indica justamente que o intervalo chegou ao fim, pois houve mais uma mudança de valor. Como a variável tempo-dependente deste trabalho muda anualmente, a estrutura fica parecida com a de um banco com dados em painel.

Antes de iniciar a seção prática do trabalho, uma última observação sobre o exemplo utilizado faz-se necessária. O número de eventos por variável (*events per variable*, EPV) é baixo. Entre 1900 e 2016, a região experimentou 14 eventos, tendo três variáveis independentes no modelo. Isso se explica pelo fato da região ter passado grande parcela desse período como regimes ditatoriais e, portanto, fora do grupo de risco²³. Como consequência, os resultados poderiam ser enviesados. No entanto, por se tratar de um exercício, não me preocuparei com isso aqui²⁴.

Aplicação no R

Finalmente com o R aberto e com o banco do V-Dem já no ambiente do software, começamos instalando os pacotes que precisaremos utilizar e, em seguida, fazendo o upload da biblioteca virtual. O *survival*²⁵ é o pacote mais completo para elaboração de modelos de sobrevivência no R. É com ele que faremos nossa regressão de Cox e também a análise dos resíduos de Schoenfeld.

23 Das 2.337 observações de país-ano, 1.713 foram sob regimes autoritários e 596 sob regimes democráticos.

24 De acordo com Peduzzi et al. (1995), um valor prudente de EPV é 10. Ver: Concato et al. (1995) e Peduzzi (1995).

25 Para uma visão mais panorâmica do pacote, ver o documento elaborado pelo seu idealizador, Terry Therneau, em: <https://cran.r-project.org/web/packages/survival/vignettes/survival.pdf>.

```
## Pacotes e opções ####
options(scipen = 999)
install.packages("survival")
library("survival")
```

A primeira linha indica a opção `scipen = 999`. Ela não é obrigatória, apenas evita que o software ofereça outputs em formato de notação científica. Isso é comum quando os valores são muito baixos, por exemplo. A segunda linha é o comando para instalar o pacote `survival`. Ele permite realizar praticamente todos os modelos dentro da família de análise de sobrevivência. A última linha faz o upload do pacote.

```
## Selecionando as variáveis que farão parte do banco ##
vdem <- `V-Dem-CY-Full+Others-v10`[, c("country_name", "country_text_id", "country_id",
                                     "year", "v2x_polyarchy", "v2x_regime", "v2xps_party",
                                     "v2xnp_pres", "e_migdppln", "e_regionpol")]

vdem <- vdem[ which(vdem$year > '1899'), ] ## Selecionando anos > 1899 ##
vdem <- vdem[ which(vdem$year < '2017'), ] ## Selecionando anos < 2017 ##
vdem <- vdem[ which(vdem$e_regionpol == '2'), ] ## Selecionando a América Latina ##
vdem$democracy <- ifelse(vdem$v2x_regime > 1, 0, 1) ## Criando variável 'democracy' ##
colnames(vdem)[1] <- "country" ## Renomeando a coluna "country_name" como "country" ##
```

Com isso feito, agora é a hora de estruturar o banco no formato de contagem para que ele esteja apto a ser utilizado em um modelo de Cox estendido. Nesse momento é preciso que o pesquisador esteja bastante atento aos dados. Como já dito na subseção anterior, esse trabalho lida com eventos recorrentes. Então o país precisará ser excluído do banco sempre que fizer a transição para um regime ditatorial, podendo retornar caso volte a ser uma democracia.

Vamos olhar um excerto da Argentina para facilitar a compreensão. O bloco mostra cada passo especifi-

catando os dados do V-Dem, podemos fazer um subgrupo para manter apenas aquelas variáveis que serão utilizadas. Também é possível fazer o recorte temporal, além de selecionar especificamente os países da América Latina. Neste exemplo, também é preciso criar a variável *dummy* que representa o evento. Tudo isso pode ser visto no bloco a seguir:

camente. De acordo com os dados do V-Dem, entre 1920 e 1929 a Argentina era uma democracia e, portanto, estava no banco. Sofreu o evento em 1930 e, por isso, foi excluído a partir de 1931. Em 1964 volta a ser uma democracia por um breve período e, dessa forma, volta aos grupos de países em risco. Mas já sofre uma nova falha em 1966 e é excluído da análise a partir de 1967²⁶. Em 1974 volta a ser democrático e a figurar no banco de dados.

Com isso feito, é hora de criar as variáveis de início e fim:

```
## Criando a variável 'start' ##
vdem$start <- as.numeric(as.factor(vdem$year))
## Criando a variável 'stop' ##
vdem$stop <- vdem$start + 1
```

26 O país sempre é excluído a partir do ano subsequente à falha, pois a observação do evento – por motivos óbvios – precisa estar no banco. Se o país deixar de ser uma democracia em um dado ano, mas já retornar ao estado inicial no ano seguinte, nenhuma alteração deve ser feita.

Depois dos ajustes, é possível usar o comando `head()` para olhar como ficaram as primeiras linhas do banco.

```
> head(vdem)
  country country_text_id country_id year democracy start stop v2x_polyarchy v2x_regime
207 Mexico             MEX          3 1995         0    73  74         0.514         2
208 Mexico             MEX          3 1996         0    74  75         0.528         2
209 Mexico             MEX          3 1997         0    75  76         0.546         2
210 Mexico             MEX          3 1998         0    76  77         0.596         2
211 Mexico             MEX          3 1999         0    77  78         0.606         2
212 Mexico             MEX          3 2000         0    78  79         0.619         2
  v2xps_party v2xnp_pres e_migdppcln e_regionpol
207      0.869      0.530      9.116         2
208      0.879      0.404      9.121         2
209      0.894      0.381      9.177         2
210      0.894      0.381      9.232         2
211      0.894      0.381      9.249         2
212      0.918      0.211      9.336         2
>
```

Finalmente, o comando para rodar a regressão de Cox:

```
## Regressao de Cox ###
modelo <- coxph(Surv(start, stop, democracy) ~ v2xps_party + v2xnp_pres + e_migdppcln,
               data=vdem)
modelo
```

A função `coxph` faz parte do pacote `survival` e indica justamente o modelo de riscos proporcionais de Cox. A partir de `Surv()`, é possível combinar os dois componentes necessários da variável dependente: tempo e evento. Como o modelo é estendido, ele segue a seguinte lógica: `Surv(início,fim,status)`. No entanto, caso não houvesse uma variável tempo-dependente e o banco fosse elaborado da maneira clássica, seria `Surv(tempo,status)`.

Resultados e interpretação

Depois de construir as variáveis necessárias para o nosso exemplo, ver uma amostra de como ficou o banco de dados e rodar o modelo de sobrevivência, agora é hora de olhar para os resultados encontrados e interpretá-los. O output gerado diretamente pelo comando do último bloco é:

```
Call:
coxph(formula = Surv(start, stop, democracy) ~ v2xps_party +
      v2xnp_pres + e_migdppcln, data = vdem)

              coef exp(coef)    se(coef)      z      p
v2xps_party   1.5569    4.7442    3.5509   0.438 0.66106
v2xnp_pres   10.1272  25014.0175    3.1730   3.192 0.00141
e_migdppcln  -0.6959     0.4986    1.0544  -0.660 0.50925

Likelihood ratio test=22.99 on 3 df, p=0.00004055
n= 609, number of events= 14
(1 observation deleted due to missingness)
```

O resultado vem com cinco informações: o coeficiente, o exponencial do coeficiente, erro padrão, a estatística *z* e o *p*-valor. Olhando para esta última coluna, vemos que apenas o índice de presidencialismo (`v2xnp_pres`) é estatisticamente significativo ao nível de 5% ($p < 0,05$).

Os coeficientes são estimados por meio da verossimilhança parcial e sua interpretação direta costuma ser

difícil. No entanto, existe uma forma bastante intuitiva de olhar para eles. Valores positivos indicam que as variáveis levam ao aumento no risco do evento; enquanto um valor negativo indica variável que ajuda na redução de risco. Se ignoramos o grau de significância e nos atermos apenas aos coeficientes, podemos afirmar que o PIB per capita atua como uma espécie de protetor de regimes democráticos. Em contrapartida, níveis altos de presidencialismo e de institucio-

nalização partidária podem prejudicar a estabilidade das democracias.

O output do R ainda apresenta o exponencial do coeficiente, que pode ser interpretado como a razão de risco. Valores entre 0 e 1 indicam proteção; acima de 1, maior risco. Ele ainda pode ser lido como percentual ao se aplicar a simples fórmula $(1 - \exp(-\text{coef})) \times 100$. Assim, voltando para o exemplo, cada unidade acrescida no valor do logaritmo natural do PIB per capita reduz a chance de quebra da democracia em 50,1%.

Resíduos de Schoenfeld

A proporcionalidade dos riscos é um pressuposto válido para o modelo de Cox tradicional, ou simples, e para o modelo estendido (BATISTA, 2017); e a interpretação dos resíduos de Schoenfeld ocorre da mesma forma tanto para a abordagem simples quanto para a que é formulado via processo de contagem (CARVALHO ET AL., 2011). A primeira etapa da análise é observar se há alguma correlação entre os coeficientes e o tempo. Para fazer isso no R, se utiliza a função `cox.zph`, também do pacote `survival`.

```
## Resíduos de Schoenfeld ###
res.modelo <- cox.zph(modelo, transform = "identity")
res.modelo
```

O output gerado pelo bloco anterior é:

```
> res.modelo
      chisq df    p
v2xps_party 0.0137 1 0.91
v2xnp_pres  0.9335 1 0.33
e_migdppln  0.2364 1 0.63
GLOBAL      2.0072 3 0.57
```

Olhando o bloco acima, é possível notar que nenhuma variável apresentou uma correlação significativa entre seus coeficientes e o tempo. Assim, não se exclui a hipótese nula que afirma não haver relação entre tempo e variáveis. No entanto, gráficos de dispersão permitem

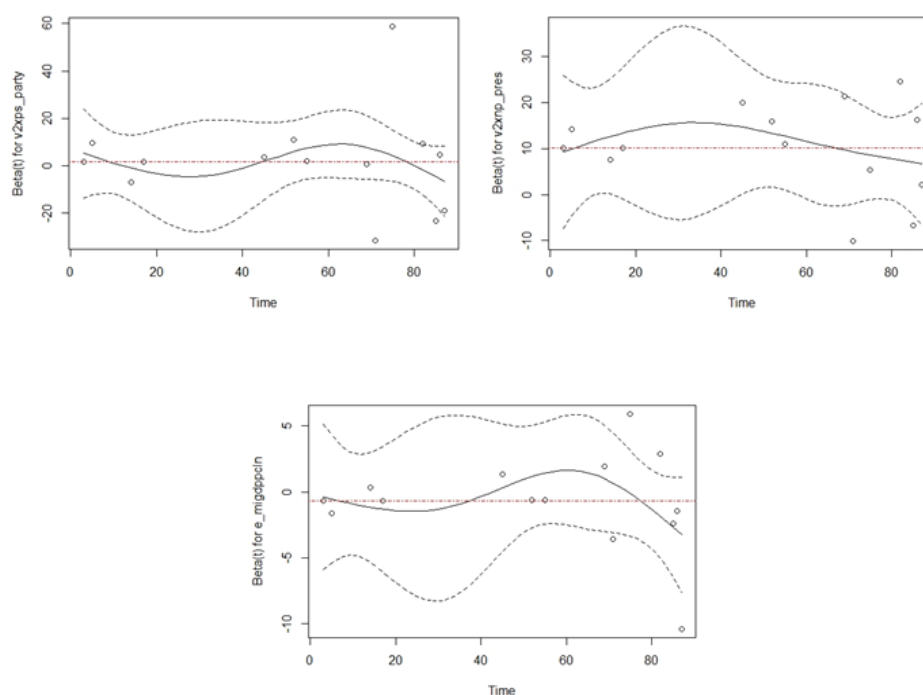
um formato visual para observar se o pressuposto dos riscos proporcionais se mantém. Será preciso criar um gráfico para quantas variáveis existirem no modelo. Neste caso, tenho três variáveis. O bloco a seguir mostra como gerar as figuras.

```
plot(res.modelo[1])
abline(h=modelo$coefficients[1], lty=4, col=2) ## Institucionalização (partidos) ##
plot(res.modelo[2])
abline(h=modelo$coefficients[2], lty=4, col=2) ## Presidencialismo ##
plot(res.modelo[3])
abline(h=modelo$coefficients[3], lty=4, col=2) ## Ln PIB per capita ##
```

A primeira linha de cada parâmetro constrói o gráfico e a segunda linha adiciona uma linha horizontal verme-

lha que representa os coeficientes de cada uma das variáveis independentes. Os três gráficos gerados são:

Figura 1 – Resíduos de Schoenfeld



Fonte: autor.

Se essa linha, assim como a linha suavizada, estiver dentro do intervalo de confiança, é possível atestar que o pressuposto foi observado. Nos três casos podemos afirmar que os parâmetros se encontram dentro dos intervalos de confiança. Além disso, não existe uma forte tendência positiva ou negativa.

6. Considerações finais

Apesar de ser uma técnica com um grande leque de possibilidades para problemas de pesquisa dentro das ciências sociais, a análise de sobrevivência ainda não decolou como uma ferramenta popular na realidade das sociais brasileira. Portanto, este trabalho tem três objetivos que se complementam: (1) difundir a técnica dentro da grande área das ciências sociais; (2) servir como material introdutório para aqueles que enxergaram valor na técnica para o desenvolvimento de suas pesquisas; e (3) apresentar um passo a passo de sua aplicação no R.

Para tanto, este trabalho apresentou um esforço para mostrar a parte teórica da análise de sobrevivência, junto com um pequeno dicionário de termos comuns para a técnica, para que seja possível compreender os diferentes tipos de desenhos de pesquisa que podem se valer da técnica. Em termos práticos, apresentei uma seção “mãos na massa”. Tomando o software R como base, mostrei o passo a passo de como elaborar um modelo de duração, desde a construção, passando pelo modelo propriamente dito, até a análise dos resíduos.

REFERÊNCIAS BIBLIOGRÁFICAS

- ALLISON, Paul. (1984), *Event History Analysis: regression for longitudinal event data*. Newbury Park, CA: SAGE Publications Inc.
- ATA, Nihal; SÖZER, M. Tekin. (2007), "Cox Regression Models with Nonparametrical Hazard Applied to Lung Cancer Survival Data". *Hacettepe Journal of Mathematics and Statistics*, vol. 3, nº 2: 157-167.
- BATISTA, Mariana. (2017), *A Difusão da Lei de Acesso à Informação nos Municípios Brasileiros: fatores internos e externos*. Brasília, DF: Enap.
- BOX-STEFFENSMEIER, Janet M.; JONES, Bradford S. (2004), *Event History Modeling: a guide for social scientists*. New York, NY: Cambridge University Press.
- BOX-STEFFENSMEIER, Janet M.; ZORN, Christopher J. W. (2001), "Duration Models and Proportional Hazards in Political Science". *American Journal of Political Science*, vol. 45, nº 4: 972-988.
- BIESZK-STOLORZ, Beata; MARKOWICZ, Iwona (2015), "Influence of Unemployment Benefit on Duration of Registered Unemployment Spells". *Equilibrium. Quarterly Journal of Economics and Economic Policy*, vol. 10, nº 3: 167-183.
- BORUCKA, Jadwiga. (2013), "Extensions of Cox Model for Non-Proportional Hazards Purpose". Disponível em: https://www.lexjansen.com/phuse/2013/sp/SP07_ppt.pdf. Acessado em: 10/06/2020.
- CARVALHO, Marília Sá; ANDREOZZI, Valeska Lima; CODEÇO, Claudia Torres; CAMPOS, Dayse Pereira; BARBOSA, Maria Tereza; SHIMAKURA, Silvia Emiko. (2011), *Análise de Sobrevida: Teoria e aplicações à saúde*. 2ª ed. Rio de Janeiro: Editora Fiocruz.
- CHEIBUB, Jose Antonio; GANDHI, Jennifer; VREELAND, James R. (2010), "Democracy and Dictatorship revisited". *Public Choice*, vol. 143, nº 2-1: 67-101.
- COLLETT, D. (1994), *Modelling Survival Data in Medical Research*. Springer-Science +Business Media.
- COLOSIMO, Enrico Antônio; GIOLO, Suely Ruiz. (2006), *Análise de Sobrevida Aplicada*. São Paulo, SP: Blucher.
- CONCATO, John; PEDUZZI, Peter; HOLFORD, Theodore R.; FEINSTEIN, Alvan R. (1995), "Importance of Event per Independent Variable in Proportional Hazard Analysis I. Background, Goals, and General Strategy". *Journal of Clinical Epidemiology*, vol. 48, nº 12: 1495-1501.
-

COX, David. (1972), "Regression Models and Life-Tables". *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, nº 2: 187-220.

CRAWLEY, Michael J. (2015) *Statistics: An introduction using R*. 2nd ed. West Sussex, UK: John Wiley & Sons.

CRESPO-TENORIO, Adriana; JENSEN, Nathan M.; ROSAS, Guillermo. "Political Liabilities: surviving banking crises". *Comparative Political Studies*, vol. 47, nº 7: 1047-1074.

DIERMEIER, Daniel; STEVENSON, Randy T. (1999), "Cabinet Survival and Competing Risks". *American Journal of Political Science*, vol. 43, nº 4: 1051-1068.

GRUBE-CAVERS, Annelise; PATTERSON, Zachary. (2015), "Urban rapid rail transit and gentrification in Canadian urban centres: a survival analysis approach". *Urban Studies*, vol. 52, nº 1: 178-194.

GUJARATI, Damodar. (2011), *Econometrics by example*. Palgrave Macmillan.

GUJARATI, Damodar N.; PORTER, Dawn C. (2011), *Econometria Básica* 5^a ed., Porto Alegre, RS: AMGH.

GURY, Nicolas. (2011), "Dropping Out of Higher Education in France: a micro-economic approach using survival analysis". *Education Economics*, vol. 19, nº 1: 51-64.

GUO, Shenyang. (2010), *Survival Analysis*. New York, NY: Oxford University Press.

HARRELL Jr.; Frank E. (2015), *Regression Modeling Strategies: with application to linear models, logistic and ordinal regression, and survival analysis*. Cham, Switzerland: Springer.

JONES, Bradford S.; BRANTON, Regina P. (2005), "Beyond Logit and Probit: Cox duration models of single, repeating, and competing events for state policy adoption". *State Politics and Policy Quarterly*, vol. 5, nº 4: 420-443.

KELLSTEDT, Paul M.; WHITTEN, Guy D. (2018), *The Fundamentals of Political Science Research*. 3rd ed. New York, NY: Cambridge University Press.

KHAMBATY, Tasneem; CALLAHAN, Christopher M.; STEWART, Jesse C. (2018), "Effect of collaborative depression treatment on risk of diabetes: a 9-year follow-up of the IMPACT randomized controlled trial". *PloS one*, vol. 13, nº 8.

KLEINBAUM, David G.; KLEIN, Mitchel. (2005), *Survival Analysis: a self-learning text*. 2^a ed. New York, NY: Springer.

LEE, Elisa T.; WANG, John Weny. (2003), *Statistical Methods for Survival Data Analysis*. 3rd ed. Hoboken, NJ: Wiley.

- LEE, Ming-Chang. (2014), "Business Bankruptcy Prediction Based on Survival Analysis Approach". *International Journal of Computer Science & Information Technology*, vol. 6, nº 2: 103-119.
- LIN, D. Y.; WEI, L. J.; YING, Z. (1993), "Checking the Cox model with cumulative sums of martingale-based residuals". *Biometrika*, vol. 80, nº 3: 557-572.
- LINS, Rodrigo; FIGUEIREDO FILHO, Dalson; ROCHA, Enivaldo. (2017), "Não temos tempo a perder: uma introdução à análise de sobrevivência". *Revista Política Hoje*, vol. 26, nº 1: 279-298.
- LÜHRMANN, Anna; LINDBERG, Staffan I.; TANNENBERG, Marcus. (2017), "Regimes in the World (RIW): a robust regime type measure based on V-Dem". *V-Dem Working Paper*, Series 2017:47.
- MACHIN, David; CHEUNG, Yin Bun; PARMAR, Mahesh K. B. (2006), *Survival Analysis: a practical approach*. 2nd ed. West Sussex, UK: John Wiley & Sons.
- PAPPAS, Vasileios; ONGENA, Steven; IZZELDIN, Marwan; FUERTES, Ana-Maria (2017), "A Survival Analysis of Islamic and Conventional Banks". *Journal of Financial Services Research*, vol. 51: 221-256.
- PEDUZZI, Peter; CONCATO, John; FEINSTEIN, Alvan R.; HOLFORD, Theodore R. (1995), "Importance of Event per Independent Variable in Proportional Hazard Analysis II. Accuracy and Precision of Regression Estimates". *Journal of Clinical Epidemiology*, vol. 48, nº 12: 1503-1510.
- SCHOENFELD, David. (1982), "Partial Residuals for the Proportional Hazard Regression Model". *Biometrika*, vol. 69, nº1: 239-241.
- SHIPAN, Charles R.; SHANNON, Megan L. (2003), "Delaying Justice(s): A duration analysis of Supreme Court Confirmations". *American Journal of Political Science*, vol. 47, nº 4: 654-668.
- SOMER-TOPCU, Zeynep; WILLIAMS, Laron K. (2008), "Survival of the Fittest? Cabinet duration in postcommunist Europe". *Comparative Politics*, vol. 40, nº 3: 313-329.
- SVOLIK, Milan. (2008), "Authoritarian Reversals and Democratic Consolidation". *American Political Science Review*, vol. 102, nº 2: 153-168.
- TANSEY, Oisín; KOEHLER, Kevin; SCHMOTZ, Alexander. (2017), "Ties to the Rest: autocratic linkage and regime survival". *Comparative Politics Studies*, vol. 50, nº 9: 1221-1254.
- WALTERS, Stephen J. (2009), *What is a Cox Model?* 2nd ed. Newmark, UK: Hayward Medical Communications.
-

All kinds of time: Cox regression in political science

Rodrigo Lins - Universidade Federal de Sergipe |

Resumo

Este artigo tem como objetivo apresentar o modelo de riscos proporcional de Cox. Para tanto, será feito um debate teórico geral dos modelos de sobrevivência, assim como a apresentação de termos comuns para a técnica. Além disso, será feita uma apresentação passo a passo de como elaborar um modelo de Cox no software R. Espera-se que este trabalho ajude na difusão da técnica dentro das ciências sociais.

Palavras-chave: Análise de Sobrevivência; Modelo de Cox; R.

Abstract

This article presents the Cox Proportional-Hazards Model. To do so, a general theoretical debate on survival models is made, as well as a presentation of common terms for the technique. In addition, there is a step-by-step presentation of how to develop a Cox model in R software. It is hoped that this work will help to disseminate the technique within the social sciences.

Keywords: Survival Analysis; Cox Model; R.

1 . Introduction¹

Survival analysis is a technique with several names: event history analysis, duration model, and time-to-event are just a few of the names given to this statistical tool. Despite the name variation, all have a key feature: time. Therefore, it is no surprise that survival models are concerned with not only “if” but also “when” a given event² occurs (GUO, 2010). That is, the researcher’s interest is the observation of a qualitative change in the dependent variable in a specific point in time (ALLISON, 1984).

Within the possible approaches for a survival analysis, the Cox proportional hazards model is highlighted (BOX-STEFFENSMEIER & JONES, 2004). Conceptualized as a semiparametric model, it is very malleable. It does not take on a statistical distribution. It only assumes that the action of the independent variables on the hazard is multiplicative (CARVALHO et al., 2011), that is, for each unit increase in the explanatory variable, the hazard is multiplied by its exponentiated coefficient (ALLISON, 1984).

And how goes the use of survival analysis – and more specifically, of the Cox model - in Brazilian political science? It is noticeable that its use has not taken off. While analyzing abstracts of available papers in the archives of the five main Brazilian political science journals³, there was only one result that indicated the use of survival analysis. Consequently, this work aims to demonstrate the technique to the larger audience of the social sciences so that it more widely considered.

¹ I would like to thank the anonymous peer reviewers who helped me to substantially improve this paper. All remaining errors are fully my responsibility.

² In survival analysis literature, the term “event” can also be named “failure”. However, since a qualitative change does not always mean an actual failure, and could be a government transition for example, I have opted to use event.

³ DADOS was the only journal with a result, a paper published in 2019. The other four journals included in the search were: the Brazilian Political Science Review; Opinião Pública; Revista Brasileira de Ciências Sociais; and Revista de Sociologia e Política. The search terms were Cox and survival. The search was conducted on 2 June 2020.

The remainder of this paper is divided thusly: first I focus on the Cox model, with a more theoretical approach; then, I present a brief glossary of common terms in survival analysis, but that are not so widely used in other statistical techniques. The following section shows possible applications in different areas of knowledge. Finally, in the last section, I demonstrate a step-by-step process of how to conduct a survival analysis in R and interpret its results.

2 . Cox model: what is it, what is it for, when to use it

What is the Cox proportional hazards model? Beforehand, it is important to understand that survival analysis modelling can be divided into two groups: parametric and semiparametric⁴. Parametric is the one that presuppose a given distribution, such as the Weibull model or its exponential format (BOX-STEFFENSMEIER & JONES, 2004). While the latter expects that the hazard is constant overtime, Weibull’s allows for the hazard to increase or lower linearly over time (ALLISON, 1984).

The semiparametric model – the Cox model – does not assume any distribution for the baseline hazard function. That would be a feature that could put it in a non-parametric group. Nevertheless, it is expected that its variables always act with a multiplicative fact on the hazard. That is the feature that would put it in the parametric group. That is why the Cox model is considered a semiparametric model.

Therefore, to answer more directly the question at the beginning of the section, we can affirm, regarding the Cox model: (1) like parametric models, it allows for

⁴ It is still possible to find non-parametric estimations. However, it is not possible to estimate the effects of explanatory variables in these models. The Kaplan-Meier estimate product, for example, compares the survival curve of two or more groups, not considering the independent variables. See Lee and Wang (2003), chapters four and five, in particular.

the inclusion of explanatory variables; (2) the effect of these variables is multiplicative; and (3) it does not hold assumptions when it comes to the statistical distribution of the baseline hazard function. The Cox model is, consequently, a statistical technique to explore the relationship of a specific event to a series of explanatory variables (WALTERS, 2009).

It was developed by Cox (1972), from whom it inherited the name, and arose as a way to soften the aforementioned assumptions of parametric models. It is called proportional hazards model because for any two observations at any point in time, the ratio between its hazards must be constant. In other words, the ratio of the hazards from observation A to observation B must be the same both on t_0 and on t_{100} , for example (ALLISON, 1984). It can be symbolised in the following equation:

$$\log h(t) = a(t) + b_1x_1 + b_nx_n \quad (1)$$

The left side of the equation is the natural logarithm of the hazard function. On the other side, the baseline hazard function $a(t)$ can be any function of time. As we have seen, it is specifically because it is unspecified that the Cox model is seen as partially parametric or semiparametric (ALLISON, 1984). In b_1x_1 , b represents the coefficient and x , the value of independent variable 1. Equation (1) is used on the type of data seen on Table 1, in which each observation occupies a line, because variables X_1 and X_n are constant in time.

The extended Cox model⁵

It is common, when working with longitudinal data, that some of the explanatory variables vary over time. Let's say the GDP per capita is one of the independent variables. It is hard to imagine that it stays constant over the years. Therefore, how to include it in a model? It is precisely for

these cases, when there are time-dependent variables (or variables that vary in time) that the extended Cox model must be used. It can be represented in the following way:

$$\log h(t) = a(t) + b_1x_1 + b_nx_n(t) \quad (2)$$

Here, $a(t)$ continues to be any function of time and variable X_1 continues to be constant over time. However, variable X_n does vary as a function of time. For this kind of design, the researcher must construct a database using a counting process, similar to what is seen on Table 2, where a new line is added whenever there are changes in the values of X_n .

The Schoenfeld residuals

There are different ways of testing the fit of a Cox regression model. However, considering that the proportionality of hazards is one of the defining features of the model, it is important to evaluate if this assumption is being met. That is the purpose of Schoenfeld residuals⁶. They give the proportionality of each covariate defined for each individual (CARVALHO et al., 2011). The equation that illustrates the calculation of residuals is:

$$r_{ik} = \delta_i(x_{ik} - a_{ik}) \quad (3)$$

where the occurrence of the event for observation i is symbolized by δ_i . The value of the specific variable is given by x_{ik} and, finally, the weighted average of the hazard covariates in a given time t is a_{ik} .

Although there are different ways of analyzing the residuals, the most intuitive way of understanding the analysis, as argued by Schoenfeld (1982), is that the residuals must be independent of time. Thus, if a correlation between the residuals and time is not statistically significant, that means that the coefficients are not

5 In addition to its extended form, the Cox model can also be stratified. For a more complete overview of the stratified Cox model, see chapter six in Carvalho et al. (2011).

6 There are other ways of analysing the proportionality of hazard. A common way is via martingale residuals (see Lin et al., 1993). There are still the possibilities of deviance residuals and score residuals (see Carvalho et al., 2011).

dependent on time and, consequently, the assumption is met. If it is statistically significant, it indicates that the proportional hazards assumption has not been met.

As the practical section shows, the analysis of Schoenfeld residuals can also be done graphically, making its interpretation easier. As it is expected that the correlation is random, the line in the dispersion graph should be horizontal. If there are any trends, positive or negative, the proportionality of hazards has not been met.

In case of not meeting the proportionality of hazards, the first step is to verify if this happens naturally or artificially. Thus, it is important to check for outliers. If the break in the assumption is natural, two approaches are possible: (1) to try a new model, such as the stratified Cox model (ATA & SÖZER, 2007; BORUCKA, 2013); or (2) to add an interaction effect between the variable that does not satisfy the assumption of proportionality and some function of time⁷ (BOX-STEFFENSMEIER & ZORN, 2001).

3. Small survival glossary

Before moving on, it is important to write out a brief glossary. Given that survival models deal with vocabulary that are not common to other statistical

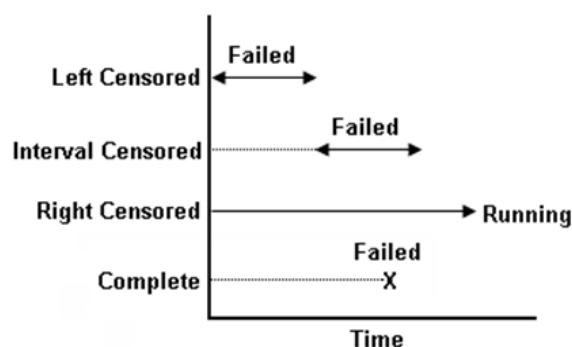
techniques, I believe it is justified to show at least the main terms. The concepts presented in this section are valid not only for Cox models, but for survival analysis in general. The small glossary does not intend to be exhaustive. This means that only the most common and relevant terms are indicated here, to enable the understanding of survival models. First off, these models can deal with two types of incomplete data, censored and truncated.

Censored: An observation is considered censored when its beginning or end are not observed, making it difficult to determine a specific point in time. There are three different types of censoring: (1) right-censored; (2) left-censored; and (3) interval-censored. The first occurs when the observation arrives at the end of the study without experiencing the event ($T > \tau^-$)⁸. Thus, the researcher only knows if this observation survived up to a given point. But they do not know when or if the event was ever experienced (GUJARATI, 2011). We say that data is censored to the left when we do not know when the event happened, but we do know that it happened before a given moment ($T < \tau^+$). Finally, interval-censored ($\tau^- < T < \tau^+$) indicates that the event happened within a period of time, but the researcher does not know exactly when (CARVALHO et al., 2011; KLEINBAUM & KLEIN, 2005). Figure 1 illustrates these three types of censoring.

7 Generally, the natural logarithm is used. See Box-Steffensmeier and Zorn (2001).

8 T represents the survival time or time to completion (CARVALHO et al., 2011). It is the interval between the insertion of the observation in the study until the moment it experiences the failure or is censored. τ^- indicates that the time of the study is shorter than the survival of the observation and τ^+ symbolizes that the time of the study surpasses the moment in which the observation was censored.

Figure 1 – Censoring



Source: Lins et al. (2017).

There are still three other types of censoring: type I, type II, and random, common in clinical studies (COLOSIMO & GIOLO, 2006). Type I censoring, similarly to right-censoring, occurs when the study is ended after a pre-established timeframe and, at that point, some observations had not yet gone through the event of interest. Type II censoring occurs when the study is ended after a pre-established number of events takes place. Finally, random censoring is the one where an observation is removed from the study without the event having happened (*idem*).

Truncated: The data can also be truncated to the right and to the left. Truncation to the right occurs when cases that have gone through the event are selected. This means all observations that went through the event of interest before the final limit of the study. Truncation to the left includes cases that experienced the event after the inferior limit of the study's timeframe (CARVALHO et al., 2011). It is also worth highlighting observations with delayed entries. These are the cases that were included in a database once the research had already started. These cases are usually mistakenly treated as truncated to the left (HARRELL, 2015).

Other terms relevant for comprehending survival models are:

Survival function - $S(t)$: It is the probability that an observation survives longer than a specific time t . It is represented by the following equation:

$$S(t) = P(T > t) \quad (4)$$

Hazard function - $h(t)$: The hazard function represents the potential of failure for each unit of time, considering that the observation survived until the time of observation t . Unlike the survival function, it is not a probability but a rate, taking any value higher than zero. It can be represented as:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad (5)$$

Visually, the survival function has declining trend, indicating that the probability of survival decreases as time passes (KLEINBAUM & KLEIN, 2005). The hazard function can be constant or increase/decrease over time. The hazard function should take precedence because regression models in a survival analysis analyze the risk of a given event occurring. This will become clearer as we deal specifically with Cox models.

Partial likelihood: The coefficients of Cox models are estimated based on a partial likelihood⁹. Unlike the

⁹ A basic presentation of maximum likelihood is given by Crawley (2015): "given the data, and given the model, what values of the parameters of that model make the observed data most likely?" (CRAWLEY, 2015, p. 6). However, because it is considered an "intricate" method (GUJARATI & PORTER, 2011), I will not explore the concepts of likelihood here.

traditional form¹⁰, the baseline hazard function¹¹ is removed from the calculation for the partial form because it is not parameterized (BOX-STEFFENSMEIER & JONES, 2004). Only the information from the observations that are under risk at each time t are considered.

$$PL = \prod_{j=1}^K L_j \quad (6)$$

where K is the total number of events in the sample and L_j is the likelihood of failure at time j , given that the observation survived until this point (KLEINBAUM & KLEIN, 2005).

Data structure: Finally, it is also important to talk about the data structure in a survival analysis. The proper

construction of the database has an important role when executing the analysis. Four variables are mandatory or expected. The *start* variable and the *end* variable respectively indicate the moment in which the observation entered the research and the moment at which it experienced the event or was censored. In addition, it is advisable to insert the *time* variable, which indicates the period in which the variable was present in the dataset. Simply put, it is the subtraction of the *end* variable by the *start* variable. Lastly, the binary variable *status*, which indicates if the observation experienced the event or not. Within the traditional model, which is the one used in a Cox regression, each observation will occupy a single line within the dataset. The table below illustrates a traditional dataset. It demonstrates a hypothetical study that begins in time 0 and ends in time 30.

Table 1 – Database

ID	Start	End	Time	Status	X1	X2
A	0	30	30	0	1	1
B	0	27	27	1	1	0
C	1	15	14	1	0	0
D	2	30	28	0	1	1
E	3	15	12	1	0	1

Source: the author.

We have five observations and one line for each. The first observation is part of the study since the beginning, given that its start is at 0. It reached time 30 without experiencing the event, since its status is also 0. The third observation started at time 1 of the study, and left at time 15, having experienced a total of 14 units of time in the research. We know that its exit happened because it experienced the event, as indicated by the status variable having the value 1. Observations 1 and 4 are right-censored examples, because they reached the

end of the study without having experienced the event. Variables X1 and X2 of Table 1 represent explanatory variables¹² with dichotomous values and that are constant in time.

Further we will see that the dataset represented in Table 1 can be used in traditional Cox regressions. However, if among the explanatory variables there is one that varies in time, the structure must be modified. In that case, the observation will gain a new line every time the value of the independent variable changes. Table 2 illustrates this situation.

¹² Explanatory variables are the ones theorized as the cause of the variation in the dependent variable (KELLSTEDT & WHITTEN, 2018).

¹⁰ The general likelihood is used in the calculation of the coefficient for the parametric survival models. See Appendix A in Allison (1984) and Chapter 3 of Box-Steffensmeier and Jones (2004).

¹¹ The baseline function is thus named because in the cases where the independent variables take on the value of 0 the value of the hazard function equals the value of the baseline function (COLLETT, 1994).

Table 2 – Database
(Time-dependent variables)

ID	Start	End	Status	X1	X2
A	0	11	0	1	100
A	11	22	1	1	120
B	0	15	0	0	40
B	12	23	0	0	42
B	23	30	0	0	47

Source: the author.

Again, the table indicates a hypothetical study lasting 30 years. Given the status variable, we know that observation 1 experienced the event, while observation 2 is right-censored. Independent variable X1 remains dichotomous and constant in time. But now variable X2 shows changes. And for every change, a new line is inserted. For observation 1, it shows only one value alteration during the whole time it is present in the dataset. Therefore, only one additional line is inserted. For observation 2, there are two changes in the value of X2. Consequently, two new lines are inserted. The break always occurs at the time the alteration happens. Another important difference is that the time variable is no longer a part of the database. It would count each line break as the input of a new observation, nullifying the value every time it happened. Thus, it is excluded from the dataset¹³. This dataset is built through the so-called counting process.

There is still another way of structuring the data, for instance if the researcher is dealing with recurring or multiple events. In this case, once a given observation experiences the event of interest, it must leave the dataset, returning once it goes back to the risk set. In political science, a clear example – which I will present with greater detail in the how-to section – is the study of regime transition. If the event of interest is the change from democratic regime to autocratic, the country will leave the dataset as soon as it experiences the institutional setback and it will remain out of the risk set as long as it is a dictatorship, because it cannot experience the event again. However, should it transition back into democracy, it is placed back in the risk set and re-enters the dataset. The following table is an excerpt of a dataset with information regarding Argentina, built from the data by Cheibub, Gandhi, and Vreeland (2010).

Table 3 – Database
(Recurring events)

ID	Start	End	Status
ARG	1946	1955	1
ARG	1958	1962	1
ARG	1963	1966	1
ARG	1973	1976	1
ARG	1983	2008	0

Source: the author.

¹³ As a rule, it can be kept. But the researcher needs to be cognizant of the fact that it would count the time of the variable's permanence only for the intervals for which there are no alterations in the variable that has its value altered. For that not to happen, the research must make the alterations "manually". However, as the example section shows, it can be completely excluded because it is not necessary to construct an extended Cox model in R.

Between 1966 and 1973 and again between 1976 and 1983, for example, we can see that Argentina was excluded from the dataset. This was done because during those years, the country was a dictatorship. Thus, it was not at risk of becoming an authoritarian country. Although coups-within-coups happened throughout that period, all events occurred during authoritarian regimes. The table shows that every time an observation experiences the event, a new line is created. But it only returns to the dataset if the initial conditions are re-established. It is also important to highlight that for the years to be shown as they are, the database should also begin at year 0.

Having understood what are censored and truncated date, what are survival and hazard functions, and seen how data must be structured, I move on to the next section, showing Cox regression models.

4. An eclectic model: applications in different fields

Survival analysis models are very eclectic when it comes to the disciplines to which they can be applied. Despite the content being very different when crossing knowledge borders, the technique itself adapts well. To see how it works across different disciplines, I present works from three areas: medicine, economics, and political science. For the latter, since it is the focus of this paper, I will show three examples. I have sought to lay out the hypothesis and the model used to test it. Being the focus of this paper, all used the Cox model. The box at the end of the section summarizes the time and status variables of each study.

Medicine

The terminology of the technique's name itself indicates that it arose in the healthcare field. Specifically,

in biomedicine. The original goal was to study patients' survival time between the diagnosis of a given illness until death (GUO, 2010). Consequently, it is unsurprising the elevated number of studies in medicine that make use of the technique (COLOSIMO & GIOLO, 2006; MACHIN, CHEUNG & PARMAR, 2006).

To demonstrate the use of the Cox model in the field, we will use the work of Khambaty, Callahan, and Stewart (2018). The authors start from the assumption that there is a relationship between depression and diabetes. More specifically, they argue that depression could be an independent risk factor for the incidence of diabetes. Thus, they wonder if by treating patients for depression, doctors would be reducing the added risk for diabetes. They tested the impact a collaborative treatment program lasting 12 months had on a group of elderly individuals with depression and initially free of diabetes. The study followed-up for nine years and had 119 patients.

It is possible to see, therefore, the variable that indicates failure: developing the illness. It is instrumentalized as 0 if the patient does not develop the illness and 1 if they do. The independent variable is also binary, indicating if the participant was part of the treatment group (1) or the control group (0). The authors also ran a second model, including some personal variables as controls: age, gender, race/ethnicity, history of hypertension, of smoking, and body mass index (BMI). The results, however, indicate that despite the treatment group indicating a drop in depression levels, when compared to the control, the groups did not diverge regarding the incidence of diabetes. Consequently, the authors argue that treatment for depression is not, on its own, sufficient to lower the risk of patients developing the illness. Since one of the control variables has its values altered over time (BMI), the authors used an extended Cox model, as seen in equation (5).

Economics

Survival analysis in economics may have several applications. From studies on the failure of businesses (LEE, 2014), to studies on the stability of banks (PAPPAS et al., 2017), the duration of unemployment (BIESZK-STOLORZ & MARKOWICZ, 2015), and gentrification processes (GRUBE-CAVERS & PATTERSON, 2015). However, I will focus on the work of Gury (2011). He analyzes the determining factors in the dropout rate of higher education students in France, i.e., the duration time of students in higher education. Thus, the status variable can already be designed: 0 if the students do not abandon higher education and 1 if they do. To conduct the study, Gury (2011) use data from France's Ministry of National Education, to follow the 5383 students that started secondary school in 1989 and later registered in a higher education institution.

With the goal of explaining the students' behaviour, he uses three large groups of independent variables: one related to personal characteristics (gender and formal education of parents, for instance); one related to the observations' secondary education (such as the type of degree obtained¹⁴); and one related to higher education (if the student declared having some professional project).

To test the hypotheses, Gury (2011) developed three models, going from the simplest (with only the first group of independent variables) to the most complete (with all three groups of explanatory variables). As a result, the author finds a non-linear pattern, indicating that there are two groups of students. The *early leavers*, who abandoned the course in the first two years; and the *late leavers*, who abandoned it from the third year on. Among the ones who experience the event – abandon higher education –, women tend

to leave in the first year, while men abandoned university towards the end of the course. Other social economic characteristics, such as the parents' level of formal education, only show effect in the first years of schooling. As years go by, differences tend to disappear. In this study, the variables do not change with time, allowing the author to work with the traditional Cox model, as seen in equation (4).

Political Science

For the third example on using survival models, I demonstrate a study in political science. It is important to highlight that these models have been used in the discipline where logit and probit models would usually be found. Several areas in the discipline are ideal for duration models. For example, they have been used in political regime transition studies (SVOLIK, 2008); in the study of government survival (DIERMEIER & STEVENSON, 1999; SOMER-TOPCU & WILLIAMS, 2008); and in the analysis of public policy diffusion (JONES & BRANTON, 2005; BATISTA, 2017).

The first (1) example we will be looking at is the work by Tansey, Koehler, and Schmotz (2017), who studied the effect of autocratic links and the survival of dictatorships. The hypothesis is presented thusly: the higher the level of international autocratic links, the lower the risk of autocratic regime breakdown. Tansey et al. (2017, p. 2) instrumentalize four dimensions of autocratic linkage. They are: trade, migration, diplomatic ties, and geographic proximity. The status variable is represented by (0) if a democratic regime survived or (1) if it broke down and became democratic. Among the controls are the GPD per capita, GDP growth, and a dummy to indicate whether it was during the Cold War. The authors' data include 250 autocratic regimes between 1949 and 2008. The time variable is year, like in the previous examples.

¹⁴ According to the author, there are three types of possible degrees in French secondary education: technical, professional, and general. The author argues that holders of the general degree are more likely to finish higher education since the secondary education general degree is of low value in the labour market.

The null hypothesis, that international autocratic linkages did not have an effect on the duration about autocratic regimes, was rejected. Thus, the risk of suffering an institutional transformation is lessened as the degree of relationships among autocratic states is strengthened. Considering that the work used GDP per capita and GDP growth as control variables, the authors used the extended Cox model as in equation (5).

The second (2) example is the work of Shipan and Shannon (2003). The authors analyzed the time it takes between nominating a candidate to the United States Supreme Court and the confirmation of that candidate by the Senate. The status variable, therefore, is: 0 when the confirmation does not occur and 1 when it does. Time, unlike the studies seen previously in this section, is measured in days. Using spatial theory, the main hypothesis in the study is that the larger the ideological distance between the president and the Senate, the longer will be the time between nomination and confirmation.

Despite ideological distance being the main hypothesis the authors intended to test, they also consider candidate characteristics, such as: if they serve or have served

in the Senate or if they had previous experience with state or federal courts. The results found by Shipan and Shannon confirmed the hypothesis. A larger ideological distance increases the time of confirmation of the president's nomination. The authors also find statistically significant results for the candidates experience in the Senate. And since the explanatory variables do not vary with time, the traditional Cox model was used.

Finally, the third (3) example is the work by Crespo-Tenorio, Jensen, and Rosas (2014). The authors' goal is to understand the impact banking crises have on the chance of incumbent survival. The authors' main hypothesis is that banking crises will have a negative impact on incumbents in democratic regimes. That impact tends to be even stronger in countries where there are few veto players. The event of interest, therefore, is failure in maintaining the same party in power. To put it more directly: when elections lead another party to power. Time is again measured annually, and the authors' hypothesis is also confirmed.

Lastly, the box below summarizes the features of the three studies seen here.

Box 1 – Summary of studies

Disciplina	Variável status	Variável tempo	Variação no tempo	Modelo
Medicine	Develop the illness	Years	Yes	Extended Cox
Economics	Abandon university	Years	No	Traditional Cox
Political Science (1)	Regime breakdown	Years	Yes	Extended Cox
Political Science (2)	Nominee confirmed	Days	No	Traditional Cox
Political Science (3)	Incumbent defeat	Years	Yes	Extended Cox

Source: the author.

Four out of the five studies analyzed here used year as a measure of time. Three of them have independent or control variables with values varying over time. This means that they used the extended Cox model and thus constructed datasets with the counting structure,

as seen in Table 2. In these three cases, the statistical construction of the model follows equation (5). The economics study and the political science study on the confirmation of nominees to the Supreme Court, which did not have a time-dependent variable, used the

traditional Cox model. Therefore, their datasets should possess a similar structure to the one shown in Table 1 and would have used equation (4).

5. How to – a computational step-by-step guide¹⁵

After the theoretical approach on survival models¹⁶ and presenting three examples on using those models, it is time for a practical example. We will follow a computational step-by-step using R¹⁷. The example is about the survival of democratic regimes, elaborated with a database from the *Varieties of Democracy* (V-Dem) project¹⁸. The analysis will focus on Latin America and with a timeframe between 1900 and 2016, working with the following research question: what is the impact of the institutionalization of political parties on the survival of democratic regimes?

Two variables are mandatory in any regression with survival models: the survival time and the event (status). The latter is, necessarily, a dichotomous variable, in which 1 indicates the occurrence of the event and, by exclusion, 0 means status quo. The time variable indicates the time since the beginning of the study

(or inclusion of the observation) until one of two situations occurs: (1) the observation experiences the event or (2) the period of the study reaches an end. Any measure of time (days, months, or years, for example) can be used. Here, we will use year.

We have, therefore, a clear construction of the dependent variable. Following the explanation of Guo (2010), we have the two pieces of information needed: one dichotomous (democracy survived (0) or suffered an institutional setback (1)); and one continuous variable representing time. However, the model has time-dependent variables. Thus, the dataset is constructed in the counting format, with a variable indicating the start of the period and another indicating the end, for the elaboration of an extended Cox regression model.

Finally, the last feature of the model is that countries can enter and exit the risk set. This happens because a democracy, by becoming a dictatorship, leaves the risk set (and, consequently, the database). However, there is no impediment for that same country to become a democracy again. Should this happen, it returns to the risk set comprised of democratic countries that can, eventually, become dictatorships (returning again to the database).

Thus, the final design is:

Box 2 – Design of the model

Model	Recurring event	Time	Event	Unit
Extended Cox	Yes	Year	Regime breakdown	Country-year

Source: the author.

¹⁵ Aiming to be transparent and replicable, the script for the analysis is available at OSF (<https://osf.io/xp6mg/>) and GitLab. Given that the only data set used was V-Dem, I recommend that the download be done directly from their webpage. See footnote 21.

¹⁶ For a broader theoretical perspective on survival analysis, one can use the material available online from professor Ronghui Xu's class: <http://www.math.ucsd.edu/~rxu/math284/>.

¹⁷ R is not the only software that can conduct this analysis. There is, for instance, ample material on how to run survival analyses on STATA or SPSS.

¹⁸ The V-Dem dataset used here is the version 10, released in March 2020. The download can be made here: <https://www.v-dem.net/en/data/data-v-dem-dataset/>.

Having understood the structure of the model, I now present the variables included. They are:

Box 3 – Variables

Variable	In the dataset
Political regime	democracy
Start	start
End	stop
Party institutionalisation	v2xps_party
Presidentialism	v2xnp_pres
GDP per capita (ln)	e_migdppln

Source: the author.

The first three lines are the components that form the dependent variable. The political regime is the status variable, indicating when a country is democratic (0) or when it experiences the event (1). It is elaborated by the researcher from a variable already present in V-Dem (v2x_regime). The original variable is constructed from the continuous measure of democracy from V-Dem. Lührmann, Lindberg, and Tannenberg (2017), before constructing the four regime categories, first divide them into two large groups: democratic and non-democratic. That is the division used here for the dichotomous status variable.

As I am dealing with variables that change over time, I also need to construct the variables start and end. The start indicates the beginning of the time interval from which the variable takes on a determined value; the end indicates precisely that the interval has reached an end, because there was another change in value. Since the time-dependent variable in this work changes annually, the structure looks like that of a panel dataset.

Before beginning the practical section of the work, one final observation on the example is needed. The number of events per variable (EPV) is low. Between 1900 and 2016, the region experienced 14 events, with three independent variables in the model. This is due to the fact that the region spent a large part of that period as dictatorships and, therefore, outside of the risk set¹⁹. Consequently, the results could be biased. Nevertheless, because this is an exercise, it will not be a concern²⁰.

Application in R

At last, with R open and the V-Dem database already loaded in the environment, we begin by installing the packages we will need and, next, uploading the virtual library. The survival²¹ package is the most complete for elaborating survival models in R. That is the one we will use to run our Cox regression and our Schoenfeld residuals analysis.

19 From the 2337 country-year observations, 1713 were under authoritarian regimes and 596 under democratic regimes.

20 According to Peduzzi et al. (1995), a prudent EPV value is 10. See Concato et al. (1995) and Peduzzi (1995).

21 For a broader view of the package, see the document elaborated by its creator, Terry Therneau, at: <https://cran.r-project.org/web/packages/survival/vignettes/survival.pdf>.

```
## Pacotes e opções ####
options(scipen = 999)
install.packages("survival")
library("survival")
```

The first line indicates the option `scipen = 999`. It is not mandatory; it simply avoids the software from offering outputs in the format of scientific notation. It is common when values are too low, for example. The second line is the command to install the `survival` package. It makes possible to run practically all the models within the family of survival analysis. The last line uploads the package.

While treating the V-Dem data, we can make a subgroup to keep only those variables that will be used. It is also possible to choose a timeframe, as well as specifically select the Latin American countries. In this example, it is also necessary to create the dummy variable that represents the event. All of this can be seen below:

```
## Selecionando as variáveis que farão parte do banco ##
vdem <- `V-Dem-CY-Full+Others-v10`[, c("country_name", "country_text_id", "country_id",
                                     "year", "v2x_polyarchy", "v2x_regime", "v2xps_party",
                                     "v2xnp_pres", "e_migdppln", "e_regionpol")]

vdem <- vdem[ which(vdem$year > '1899'), ] ## Selecionando anos > 1899 ##
vdem <- vdem[ which(vdem$year < '2017'), ] ## Selecionando anos < 2017 ##
vdem <- vdem[ which(vdem$e_regionpol == '2'), ] ## Selecionando a América Latina ##
vdem$democracy <- ifelse(vdem$v2x_regime > 1, 0, 1) ## Criando variável 'democracy' ##
colnames(vdem)[1] <- "country" ## Renomeando a coluna "country_name" como "country" ##
```

This being done, it is now time to structure the database in the counting format, so it is possible to use it in an extended Cox model. At this point it is necessary that the researcher pays close attention to the data. As mentioned in the previous subsection, this work deals with recurring events. Therefore, the country will need to be excluded from the database every time it transitions into a dictatorship, with the permission to return in case it goes back to being a democracy.

Let's look at an excerpt of Argentina to make understanding this easier. The code shows each step specifi-

cally. According to the V-Dem data, between 1920 and 1929, Argentina was a democracy and, therefore, was in the database. It experienced an event in 1930 and, consequently, was excluded from 1931. In 1964, it goes back to being a democracy for a brief period and thus, returns to the group of countries at risk. But it experiences a new breakdown in 1966 and it is excluded from the analysis from 1967²². In 1974, it returns to being a democracy and to the database.

With this done, it is time to create the start and end variables:

```
## Criando a variável 'start' ##
vdem$start <- as.numeric(as.factor(vdem$year))
## Criando a variável 'stop' ##
vdem$stop <- vdem$start + 1
```

22 The country is always excluded from the subsequent year to the failure since the observation of an event – for obvious reasons – needs to be in the database. If the country stops being a democracy in a given year, but already returns to its initial status in the following year, no alteration must be made.

After adjustments, it is possible to use the `head()` command to see how the first lines in the dataset look.

```
> head(vdem)
  country country_text_id country_id year democracy start stop v2x_polyarchy v2x_regime
207 Mexico             MEX          3 1995         0    73   74         0.514         2
208 Mexico             MEX          3 1996         0    74   75         0.528         2
209 Mexico             MEX          3 1997         0    75   76         0.546         2
210 Mexico             MEX          3 1998         0    76   77         0.596         2
211 Mexico             MEX          3 1999         0    77   78         0.606         2
212 Mexico             MEX          3 2000         0    78   79         0.619         2
  v2xps_party v2xnp_pres e_migdppln e_regionpol
207      0.869      0.530      9.116         2
208      0.879      0.404      9.121         2
209      0.894      0.381      9.177         2
210      0.894      0.381      9.232         2
211      0.894      0.381      9.249         2
212      0.918      0.211      9.336         2
> |
```

Finally, the command to run the Cox regression:

```
## Regressao de Cox ###
modelo <- coxph(Surv(start, stop, democracy) ~ v2xps_party + v2xnp_pres + e_migdppln,
               data=vdem)
modelo
```

The `coxph` function is part of the `survival` package and it precisely results in the Cox proportional hazards model. Using `Surv()`, it is possible to combine the two components needed in the dependent variable: time and event. As the model is extended, it uses the following logic: `Surv(start,end,status)`. However, if there were no time-dependent variable and the database were elaborated in the traditional way, it would be `Surv(time,status)`.

Results and interpretation

After construction all the necessary variables for our example, seeing a sample of how the dataset looks, and running the survival model, it is now time to look at the results found and interpret them. The output directly generated by the command in the last code section is:

```
Call:
coxph(formula = Surv(start, stop, democracy) ~ v2xps_party +
      v2xnp_pres + e_migdppln, data = vdem)

              coef exp(coef)    se(coef)      z      p
v2xps_party  1.5569    4.7442    3.5509  0.438 0.66106
v2xnp_pres   10.1272  25014.0175    3.1730  3.192 0.00141
e_migdppln   -0.6959     0.4986    1.0544 -0.660 0.50925

Likelihood ratio test=22.99 on 3 df, p=0.00004055
n= 609, number of events= 14
(1 observation deleted due to missingness)
```

The output comes with five pieces of information: the coefficient, exponentiated coefficient, the standard error, the z-score, and the p-value. Looking at the latter, we see that the presidentialism index (`v2xnp_pres`) is statistically significant at 5% ($p < 0.05$).

The coefficients are estimated by partial likelihood and their direct interpretation can be difficult. Nev-

ertheless, there is a very intuitive way of looking at them. Positive values indicate that the variables lead to an increase in risk of the event, while a negative value indicates that the variable aids in reducing risk. If we ignore the level of significance and stick only with the coefficients, we can state that the GDP per capita acts as a form of shield in democratic regimes. On the other hand, high levels of presidentialism

and party institutionalization can harm the stability of democracies.

R's output also shows the exponentiated coefficient, which can be interpreted as the risk ratio. Values between 0 and 1 indicate protection; above 1, greater risk. It can also be read as a percentage by applying the simple equation $(1 - \exp(\text{coef})) \times 100$. Therefore, returning to the example, each unit added in the value of the natural logarithm of the GDP per capita reduces the chance of a breakdown in democracy in 50.1%.

Schoenfeld residuals

The proportional hazards assumption is valid for the traditional or simple Cox model, and for the extended model (BATISTA, 2017); and the interpretation of the Schoenfeld residuals occurs in the same way for both the simple approach and for when it is formulated via the counting process (CARVALHO et al., 2011). the first step in the analysis is to observe if there is any correlation between the coefficients and time. To do that in R, the `cox.zph` function is used, also in the survival package.

```
## Resíduos de Schoenfeld ###
res.modelo <- cox.zph(modelo, transform = "identity")
res.modelo
```

The output generated by the previous code is:

```
> res.modelo
              chisq df    p
v2xps_party 0.0137  1 0.91
v2xnp_pres  0.9335  1 0.33
e_migdppln  0.2364  1 0.63
GLOBAL      2.0072  3 0.57
```

Looking at the code above, it is possible to see that no variable presented a significant correlation between their coefficients and time. Thus, the null hypothesis that states that there is no relationship between time and the variables is not excluded. However, dispersion

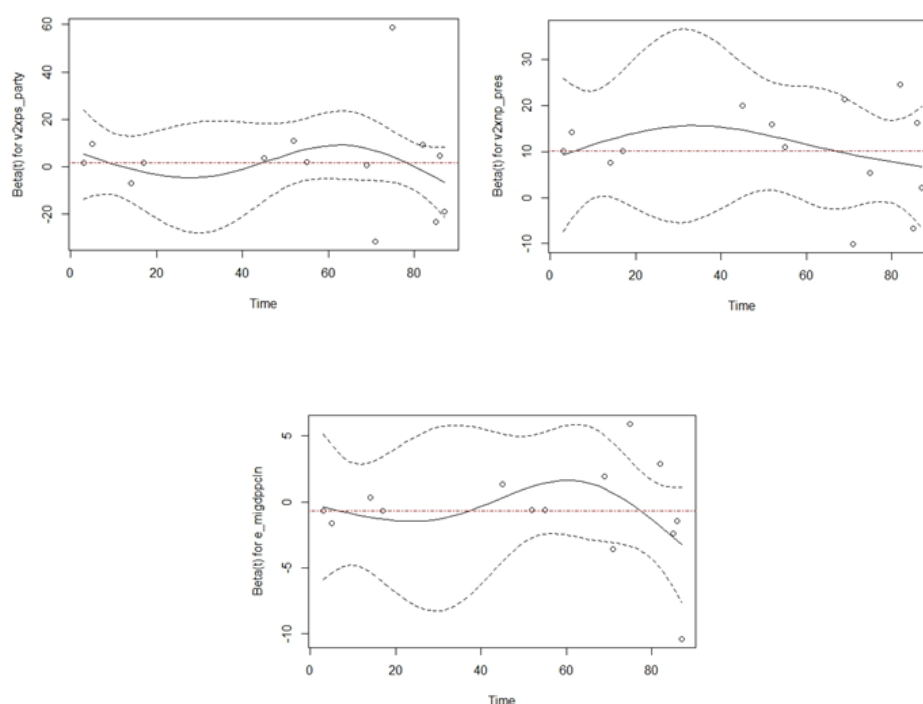
graphs allow a visual way to observe if the proportional hazards assumption is upheld. It is necessary to create a graph for however many variables exist in the model. In this case, I have three variables. The following code shows how to generate the figures.

```
plot(res.modelo[1])
abline(h=modelo$coefficients[1], lty=4, col=2) ## Institucionalização (partidos) ##
plot(res.modelo[2])
abline(h=modelo$coefficients[2], lty=4, col=2) ## Presidencialismo ##
plot(res.modelo[3])
abline(h=modelo$coefficients[3], lty=4, col=2) ## Ln PIB per capita ##
```

The first line of each parameter constructs the graph, and the second line adds a horizontal, red line that rep-

resents the coefficients of each of the independent variables. The three graphs generated are:

Figure 1 – Schoenfeld Residuals



Source: the author.

If that line, similarly to the softened line, is within the confidence interval, it is possible to affirm that the assumption is upheld. In the three cases we can state that the parameters can be found within the confidence intervals. In addition, there are no strong positive or negative trends.

6. Discussion

Despite being a technique with a wide range of possibilities for research problems within the social sciences, survival analysis has not yet taken off as a popular tool in Brazilian social sciences. Therefore, this paper has three complementary goals: (1) to share the technique within the large field of the social sciences; (2) to serve as introductory material for those who see value in the technique in the development of their research; and (3) to present a step-by-step of its application in R.

To do so, this paper demonstrated the theoretical background of survival analysis, along with a small glossary of common terms for the technique, so that it is possible to understand the different types of research design that might use it. In a practical perspective, I presented a “hands on” section. Using R, I demonstrated a step-by-step on how to elaborate a duration model, from construction, through the model, until the residuals analysis.

BIBLIOGRAPHICAL REFERENCES

- ALLISON, Paul. (1984), *Event History Analysis: regression for longitudinal event data*. Newbury Park, CA: SAGE Publications Inc.
- ATA, Nihal; SÖZER, M. Tekin. (2007), "Cox Regression Models with Nonparametrical Hazard Applied to Lung Cancer Survival Data". *Hacettepe Journal of Mathematics and Statistics*, vol. 3, nº 2: 157-167.
- BATISTA, Mariana. (2017), *A Difusão da Lei de Acesso à Informação nos Municípios Brasileiros: fatores internos e externos*. Brasília, DF: Enap.
- BOX-STEFFENSMEIER, Janet M.; JONES, Bradford S. (2004), *Event History Modeling: a guide for social scientists*. New York, NY: Cambridge University Press.
- BOX-STEFFENSMEIER, Janet M.; ZORN, Christopher J. W. (2001), "Duration Models and Proportional Hazards in Political Science". *American Journal of Political Science*, vol. 45, nº 4: 972-988.
- BIESZK-STOLORZ, Beata; MARKOWICZ, Iwona (2015), "Influence of Unemployment Benefit on Duration of Registered Unemployment Spells". *Equilibrium. Quarterly Journal of Economics and Economic Policy*, vol. 10, nº 3: 167-183.
- BORUCKA, Jadwiga. (2013), "Extensions of Cox Model for Non-Proportional Hazards Purpose". Disponível em: https://www.lexjansen.com/phuse/2013/sp/SP07_ppt.pdf. Acessado em: 10/06/2020.
- CARVALHO, Marília Sá; ANDREOZZI, Valeska Lima; CODEÇO, Claudia Torres; CAMPOS, Dayse Pereira; BARBOSA, Maria Tereza; SHIMAKURA, Silvia Emiko. (2011), *Análise de Sobrevida: Teoria e aplicações à saúde*. 2ª ed. Rio de Janeiro: Editora Fiocruz.
- CHEIBUB, Jose Antonio; GANDHI, Jennifer; VREELAND, James R. (2010), "Democracy and Dictatorship revisited". *Public Choice*, vol. 143, nº 2-1: 67-101.
- COLLETT, D. (1994), *Modelling Survival Data in Medical Research*. Springer-Science + Business Media.
- COLOSIMO, Enrico Antônio; GIOLO, Suely Ruiz. (2006), *Análise de Sobrevida Aplicada*. São Paulo, SP: Blucher.
- CONCATO, John; PEDUZZI, Peter; HOLFORD, Theodore R.; FEINSTEIN, Alvan R. (1995), "Importance of Event per Independent Variable in Proportional Hazard Analysis I. Background, Goals, and General Strategy". *Journal of Clinical Epidemiology*, vol. 48, nº 12: 1495-1501.
- COX, David. (1972), "Regression Models and Life-Tables". *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, nº 2: 187-220.
- CRAWLEY, Michael J. (2015) *Statistics: An introduction using R*. 2nd ed. West Sussex, UK: John Wiley & Sons.

- CRESPO-TENORIO, Adriana; JENSEN, Nathan M.; ROSAS, Guillermo. "Political Liabilities: surviving banking crises". *Comparative Political Studies*, vol. 47, nº 7: 1047-1074.
- DIERMEIER, Daniel; STEVENSON, Randy T. (1999), "Cabinet Survival and Competing Risks". *American Journal of Political Science*, vol. 43, nº 4: 1051-1068.
- GRUBE-CAVERS, Annelise; PATTERSON, Zachary. (2015), "Urban rapid rail transit and gentrification in Canadian urban centres: a survival analysis approach". *Urban Studies*, vol. 52, nº 1: 178-194.
- GUJARATI, Damodar. (2011), *Econometrics by example*. Palgrave Macmillan.
- GUJARATI, Damodar N.; PORTER, Dawn C. (2011), *Econometria Básica* 5ª ed., Porto Alegre, RS: AMGH.
- GURY, Nicolas. (2011), "Dropping Out of Higher Education in France: a micro-economic approach using survival analysis". *Education Economics*, vol. 19, nº 1: 51-64.
- GUO, Shenyang. (2010), *Survival Analysis*. New York, NY: Oxford University Press.
- HARRELL Jr.; Frank E. (2015), *Regression Modeling Strategies: with application to linear models, logistic and ordinal regression, and survival analysis*. Cham, Switzerland: Springer.
- JONES, Bradford S.; BRANTON, Regina P. (2005), "Beyond Logit and Probit: Cox duration models of single, repeating, and competing events for state policy adoption". *State Politics and Policy Quarterly*, vol. 5, nº 4: 420-443.
- KELLSTEDT, Paul M.; WHITTEN, Guy D. (2018), *The Fundamentals of Political Science Research*. 3rd ed. New York, NY: Cambridge University Press.
- KHAMBATY, Tasneem; CALLAHAN, Christopher M.; STEWART, Jesse C. (2018), "Effect of collaborative depression treatment on risk of diabetes: a 9-year follow-up of the IMPACT randomized controlled trial". *PloS one*, vol. 13, nº 8.
- KLEINBAUM, David G.; KLEIN, Mitchel. (2005), *Survival Analysis: a self-learning text*. 2ª ed. New York, NY: Springer.
- LEE, Elisa T.; WANG, John Wen. (2003), *Statistical Methods for Survival Data Analysis*. 3rd ed. Hoboken, NJ: Wiley.
- LEE, Ming-Chang. (2014), "Business Bankruptcy Prediction Based on Survival Analysis Approach". *International Journal of Computer Science & Information Technology*, vol. 6, nº 2: 103-119.
- LIN, D. Y.; WEI, L. J.; YING, Z. (1993), "Checking the Cox model with cumulative sums of martingale-based residuals". *Biometrika*, vol. 80, nº 3: 557-572.
- LINS, Rodrigo; FIGUEIREDO FILHO, Dalson; ROCHA, Enivaldo. (2017), "Não temos tempo a perder: uma introdução à análise de sobrevivência". *Revista Política Hoje*, vol. 26, nº 1: 279-298.
-

- LÜHRMANN, Anna; LINDBERG, Staffan I.; TANNENBERG, Marcus. (2017), "Regimes in the World (RIW): a robust regime type measure based on V-Dem". *V-Dem Working Paper*, Series 2017:47.
- MACHIN, David; CHEUNG, Yin Bun; PARMAR, Mahesh K. B. (2006), *Survival Analysis: a practical approach*. 2nd ed. West Sussex, UK: John Wiley & Sons.
- PAPPAS, Vasileios; ONGENA, Steven; IZZELDIN, Marwan; FUERTES, Ana-Maria (2017), "A Survival Analysis of Islamic and Conventional Banks". *Journal of Financial Services Research*, vol. 51: 221-256.
- PEDUZZI, Peter; CONCATO, John; FEINSTEIN, Alvan R.; HOLFORD, Theodore R. (1995), "Importance of Event per Independent Variable in Proportional Hazard Analysis II. Accuracy and Precision of Regression Estimates". *Journal of Clinical Epidemiology*, vol. 48, nº 12: 1503-1510.
- SCHOENFELD, David. (1982), "Partial Residuals for the Proportional Hazard Regression Model". *Biometrika*, vol. 69, nº1: 239-241.
- SHIPAN, Charles R.; SHANNON, Megan L. (2003), "Delaying Justice(s): A duration analysis of Supreme Court Confirmations". *American Journal of Political Science*, vol. 47, nº 4: 654-668.
- SOMER-TOPCU, Zeynep; WILLIAMS, Laron K. (2008), "Survival of the Fittest? Cabinet duration in postcommunist Europe". *Comparative Politics*, vol. 40, nº 3: 313-329.
- SVOLIK, Milan. (2008), "Authoritarian Reversals and Democratic Consolidation". *American Political Science Review*, vol. 102, nº 2: 153-168.
- TANSEY, Oisín; KOEHLER, Kevin; SCHMOTZ, Alexander. (2017), "Ties to the Rest: autocratic linkage and regime survival". *Comparative Politics Studies*, vol. 50, nº 9: 1221-1254.
- WALTERS, Stephen J. (2009), *What is a Cox Model?* 2nd ed. Newmark, UK: Hayward Medical Communications.
-