Avaliação do desempenho de modelos preditivos no contexto de análise de sobrevivência

Tiago Mendonça dos Santos

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

Programa: Estatística

Orientadora: Prof[a]Dr[a]Gisela Tunes da Silva

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro do CNPq

São Paulo, abril de 2013

Avaliação do desempenho de modelos preditivos no contexto de análise de sobrevivência

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 17/05/2013. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Profª. Drª Gisela Tunes da Silva (orientadora) - IME-USP
- Profª. Drª. Viviana Giampaoli - IME-USP
- Prof. Dr. Enrico Antônio Colosimo - ICEx-UFMG

Agradecimentos

Agradeço primeiramente à minha orientadora, Profa Gisela Tunes da Silva, que tornou esse trabalho possível. Sua disponibilidade, dedicação, sugestões, confiança e paciência foram essenciais na execução desse trabalho. Agradeço, também, pelo ótimo convívio durante esse período.

Aos meus pais, Jurandir e Helyza, por todo o esforço que fizeram, desde a minha infância, para que eu tivesse condições de estudar. Por me ensinarem o valor da educação e por tudo aquilo que me proporcionaram. Ao meu irmão Jonatas pela amizade, por todas as conversas sobre os mais diversos assuntos e pela grande companhia em viagens.

À minha namorada, Luciana Morita, que me acompanha desde o início da graduação. Agradeço, também, pelo grande incentivo durante o mestrado, pelas sugestões no texto, pela compreensão e apoio durante esse período e por todos os bons momentos.

Um agradecimento especial deve ser feito ao fã número um de flamenco, Rafael Izbicki, pela contribuição inestimável a esse trabalho. Contribuição dada no texto, na parte estatística, computacional e ao me enviar alguns artigos. Além disso, pelas boas conversas quase diárias e por esses anos de amizade. Agradeço, também, ao grande amostrista, Raphael Nishimura, pelas conversas sobre estatística, corridas de rua e pela excelente amizade desde a Ipsos.

Aos meus grande companheiros durante esse período no IME: Enéas Nogueira e Leandro Nakao. Muito obrigado pela grande amizade e pela ótima companhia, seja ao comer yakissoba na Liberdade ou dentro de um ônibus sendo apedrejado.

Agradeço a todos os professores com os quais tive o privilégio de estudar durante a graduação e o mestrado no IME. Em especial ao professor Carlinhos por todo o apoio dado e por ajudar no meu desenvolvimento como estatístico.

Por fim, aos amigos que contribuíram diretamente com esse trabalho ou dando apoio em diversos momentos. Em especial ao Bruno Santos e Camila Moreira.

Resumo

Avaliação do desempenho de modelos preditivos no contexto de análise de sobrevivência.

Modelos estatísticos com objetivos preditivos são frequentemente aplicados como ferramentas no processo de tomadas de decisão em diversas áreas. Uma classe importante de modelos estatísticos é composta por modelos de análise de sobrevivência. Duas quantidades são de interesse nessa classe: o tempo até o instante do evento de interesse ou o status para um determinado instante de tempo fixado.

Aplicações importantes desses modelos incluem a identificação de novos marcadores para certas doenças e definição de qual terapia será mais adequada de acordo com o paciente. Os marcadores utilizados podem ser dados por biomarcadores, assim como por marcadores baseados em modelos de regressão. Um exemplo de marcador baseado em modelos de regressão é dado pelo preditor linear.

Ainda que a utilização de modelos de sobrevivência com objetivos preditivos seja de suma importância, a literatura nesse assunto é muito esparsa e não há consenso na forma de se avaliar o desempenho preditivo desses. Esse trabalho pretende reunir e comparar diferentes abordagens de se avaliar o desempenho preditivo de modelos de sobrevivência. Essa avaliação é feita principalmente utilizando-se funções de perda para o tempo de sobrevivência e quantidades associadas a diferentes definições de curva ROC para o status. Para a comparação dessas diferentes metodologias foi feito um estudo de simulação e no final aplicou-se essas técnicas em um conjunto de dados de um estudo do Instituto do Câncer de São Paulo.

Palavras-chave: análise de sobrevivência, predição, curva ROC e IPCW.

Abstract

Evaluation of predictive models in survival analysis.

In many fields, predictive models are often applied as a helpful tool in the decision making process. An important class of predictive models is composed by survival models. Two quantities of special interest in these class are: time until the occurrence of a specified event and survival status for a fixed moment of time.

Important applications of these models include new markers identification for certain diseases, as well as defining which therapy is the most appropriated for a patient. Markers can be given by biomarkers, but they can also be derived from regression models. An example of regression models based markers is the linear predictor.

Despite the importance of survival models applications with predictive goals, literature is this subject is very sparse and there is no agreement on the best methodology to evaluate predictive performance of these models. In this work we intend to assemble and to compare different methodologies for assessing the predictive performance of survival models. This assessment is made mainly with loss functions for the survival time and ROC curve associated quantities for status. An simulation study was done in order to compare these different methodologies, which were also applied to a study about survival of patients at ICU of ICESP (Instituto do Câncer de São Paulo).

Keywords: survival analysis, prediction, ROC curve and IPCW.

Sumário

Capítulo 1

Introdução

Modelos estatísticos de predição são aplicados frequentemente como ferramentas no processo de tomadas de decisão em diversas áreas. Esses modelos são utilizados em áreas como medicina, economia, marketing e controle de qualidade. Algumas aplicações de modelos estatísticos são dadas na identificação de quais tratamentos apresentam melhores resultados, para determinar a quantidade ótima de matéria-prima a ser comprada e para determinar a probabilidade de inadimplência dos clientes para tomar medidas preventivas e assim evitar maiores prejuízos.

A grande importância de fatores preditores, na medicina, pode ser verificada pelo resultado de mais de 160.000 artigos publicados com os termos "preditores"ou "fatores prognósticos"(PubMed, 2013). Para se utilizar um novo preditor para uma determinada doença, deve-se mostrar que este novo preditor apresenta resultados melhores do que os preditores previamente estabelecidos. Assim, é extremamente importante a avaliação preditiva desses fatores e a comparação entre eles.

Nessa dissertação, o foco será a avaliação de modelos preditivos no contexto de análise de sobrevivência. Em análise de sobrevivência, estuda-se dados de tempo ou instantes até a ocorrência de certos eventos (para mais detalhes ver Klein e Moeschberger (2003)). Algumas importantes aplicações de modelos preditivos no contexto de análise de sobrevivência podem ser na identificação de qual terapia apresentará um menor tempo de recuperação em determinada situação ou para determinar o tempo ótimo de manutenção de uma máquina. Além desses exemplos de aplicações, uma importante aplicação de modelo de predição nessa área está sendo desenvolvida num estudo do Instituto do Câncer de São Paulo (ICESP). Nesse estudo, observou-se o tempo de sobrevivência de pacientes com câncer a partir do instante de internação na UTI. Como os recursos para se manter um paciente internado na UTI são escassos, é de vital importância uma predição de sobrevivência acurada para otimizar a alocação desses recursos.

Ao utilizar os modelos de sobrevivência, pode-se estar interessado em duas formas de predição: na predição do tempo até o instante de evento ou no status predito para um determinado instante de tempo. Por exemplo, quando um paciente é diagnosticado com câncer, ele pode ser informado sobre sua probabilidade de estar vivo num período de dois anos. Portanto, é de grande importância avaliar a performance preditiva de modelos de sobrevivência nessas duas formas de predição.

Como citado anteriormente, a avaliação da performance preditiva dos modelos de sobrevivência pode ser feita utilizando-se como referências dois tipos de medida de natureza distintas. Uma é dada

pelo tempo de sobrevivência, que pode ser considerado como uma variável contínua, e outra por uma variável binária que é dada pelo status em um determinado instante de tempo. Para cada tipo de medida existem diversas abordagens propostas. Essas abordagens se dividem em dois principais grupos: variação explicada e curva ROC.

Medidas de variação explicada

As medidas de variação explicada são medidas semelhantes ao coeficiente de determinação R^2 utilizado em modelos de regressão linear (Montgomery *et al.*, 2001). No contexto de regressão linear considera-se

$$R^2 = 1 - \frac{\mathrm{E}[\mathrm{Var}(Y|\mathbf{Z})]}{\mathrm{Var}(Y)},$$

em que Y indica a variável resposta e $\mathbf{Z}$ um vetor de variáveis explicativas.

Essa medida pode ser interpretada como a proporção de variabilidade da variável resposta que é explicada ao utilizar-se as covariáveis em estudo. Baseadas em R^2, surgiram várias medidas no contexto de análise de sobrevivência.

O conceito de variação explicada (VE) recebe esse nome pois é uma classe de medidas que apresenta a redução proporcional, num erro de predição de sobrevivência, ao se utilizar um modelo preditor no lugar de um modelo *nulo*.

Uma abordagem geral que utiliza funções de perda foi proposta por Korn e Simon (1990). Nesse trabalho os autores apresentaram duas classes de medidas de variação explicada. A primeira classe requer a especificação de uma função de perda $L(y, p)$, em que y é o tempo de sobrevivência observado e p é o preditor desse tempo. A segunda classe é baseada na correlação de posto dos tempos de sobrevivência observados e preditos. Nas duas classes assume-se que o modelo de sobrevivência foi especificado corretamente.

Korn e Simon (1990) definiram o modelo *nulo* como $S_0(t) = \sum_i S(t_i|\mathbf{z_i})$, em que $S(\cdot|\mathbf{Z})$ é a função de sobrevivência associada à variável aleatória T, que representa o tempo de falha, condicional a um vetor de covariáveis preditoras $\mathbf{Z}$. Antes de definir a medida de VE é necessário definir a perda esperada, também chamada pelos autores de risco, da seguinte forma:

$$R_L[S(\cdot)] = \min_p \int L(t, p) d\overline{S}(t),$$

em que $\overline{S}(t)$ é dado por $1 - S(t)$. Com isso, os autores utilizam a seguinte definição de VE:

$$\mathrm{VE} = \frac{R_L[S_0(\cdot)] - \frac{1}{n}\sum_{i=1}^{n} R_L[S(\cdot|z_i)]}{R_L[S_0(\cdot)]}.$$

Novamente, essa classe de medidas recebe esse nome pois, assim como o coeficiente de determinação R^2, quantifica a variação reduzida ao se utilizar um modelo que considera conjunto de covariáveis $\mathbf{Z}$ no lugar do modelo *nulo*.

Note que a VE depende da função de perda utilizada. Nesse trabalho, Korn e Simon (1990) apresentaram diferentes funções de perdas, assim como o seus respectivos preditores ótimos. Utilizando uma específica função de perda obtém-se uma medida explorada por outros pesquisadores

conhecida como escore de Brier. Esse escore é obtido quando utiliza-se, para um instante de tempo fixado t_0, a seguinte função de perda:

$$L(y, p) = [\mathbb{I}(y > t_0) - S(t_0|\mathbf{z})]^2 .$$

Note, ainda, que a medida de VE proposta por Korn e Simon (1990) é totalmente baseada no modelo de sobrevivência especificado. Dessa forma, Gerds e Schumacher (2006) observaram que a especificação incorreta do modelo de sobrevivência leva a um viés e é particularmente problemática quando o objetivo está em comparar diferentes modelos de sobrevivência, pois o grau do viés pode depender do nível incorreto de especificação do modelo.

Para evitar a dependência da especificação correta do modelo de sobrevivência, Graf *et al.* (1999) propuseram um estimador de um caso específico da medida de variância explicada, dado pelo escore de Brier esperado, que utiliza o esquema de ponderação pelo inverso da probabilidade de censura (*inverse probability of censoring weighting* - IPCW) proposto por Robins e Rotnitzky (1992). Ainda utilizando estimadores IPCW, Uno *et al.* (2007) definiram um estimador para uma diferente função de perda considerando o status de sobrevivência e uma classe diferente de modelos de sobrevivência. No entanto, assim como no trabalho de Graf *et al.* (1999), considerou-se a censura aleatória completamente independente do tempo de sobrevivência e do conjunto de covariáveis.

Uma abordagem que considera a independência condicional da censura e do tempo de sobrevivência dadas as covariáveis foi proposta por Gerds e Schumacher (2006). Nesse trabalho os autores aprofundaram a análise, mostrando a consistência do estimador da VE proposto por Graf *et al.* (1999). Ainda, mostraram que uma versão modificada desse estimador é consistente mesmo quando a censura e o tempo de sobrevivência são, apenas, condicionalmente independentes dadas as covariáveis. Embora a modificação proposta por Gerds e Schumacher (2006) considere apenas funções de perda para o status de sobrevivência, Lawless e Yuan (2010) generalizaram esse estimador para qualquer função de perda arbitrária. Isso inclui o caso do tempo de sobrevivência. Nesse trabalho, além de generalizar as funções de perda e trabalhar com estimadores baseados no modelo de sobrevivência, Lawless e Yuan (2010) trabalharam com mais dois estimadores: perda aparente e validação cruzada.

Curva ROC

Nessa abordagem é necessário estender os conceitos de curva ROC, assim como as quantidades associadas a essa curva, ao contexto de análise de sobrevivência. Como o status de cada indivíduo depende do instante fixado, é necessário trabalhar com novas definições de sensibilidade e especificidade. Assim, diferentes autores trabalharam com três possíveis definições: *incidente/dinâmico*, *cumulativo/dinâmico* e *incidente/estático*.

Na abordagem *incidente/dinâmico* Heagerty e Zheng (2005) propuseram novas medidas de acurácia dependentes do tempo calculadas nos conjuntos de observações em risco. Os autores ainda propuseram um método de estimação para uma medida de concordância global, chamada de coeficiente de concordância, entre o marcador e o tempo de sobrevivência. Também, mostraram que quantidades utilizadas na estimação do modelo de regressão de Cox podem ser utilizadas para obter

as estimativas de sensibilidade e especificidade e, consequentemente, na curva ROC dependente do tempo. Ainda nessa abordagem, Cai *et al.* (2006) consideraram casos em que a acurácia preditiva pode depender da distância do tempo da medição do marcador e o evento. Com isso, os autores propuseram modelos semiparamétricos, que consideram marcadores contínuos e a presença de censura, para estimar a sensibilidade e especificidade.

Na abordagem *cumulativo/dinâmico*, Uno *et al.* (2007) propuseram estimadores IPCW da sensibilidade e especificidade utilizando como fator de ponderação a função de sobrevivência marginal. Nesse contexto, aplica-se a generalização proposta por Lawless e Yuan (2010) que ainda considera a distribuição do tempo de sobrevivência independente da censura condicional ao conjunto de covariáveis. Uma importante propriedade do estimador proposto por Lawless e Yuan (2010) é que, com a utilização desse estimador, garante-se a monotonicidade da curva ROC. Ainda nessa abordagem, mas com uma proposta diferente, Viallon e Latouche (2011) estabeleceram uma importante relação entre a curva ROC dependente do tempo e uma medida chamada curva preditiva. Essa relação mostra que medidas acuradas da distribuição acumulada do tempo de sobrevivência condicional ao marcador devem gerar estimativas acuradas da área sob a curva ROC dependente do tempo. A partir dessa relação os autores derivam diferentes estimadores da área sob a curva dependente do tempo.

1.1 Objetivos e organização do trabalho

Os trabalhos sobre métodos de avaliação do desempenho preditivo de modelos de sobrevivência ainda são muitos dispersos na literatura. Nessa dissertação, pretende-se avaliar os principais métodos de avaliação do desempenho preditivo no contexto de análise de sobrevivência considerando as abordagens com função de perda e curva ROC. No capítulo 2 será apresentada a abordagem com função de perda. Nesse capítulo, serão apresentados os estimadores propostos por Lawless e Yuan (2010) para a avaliação do desempenho preditivo para o tempo e status de sobrevivência. No capítulo 3 serão apresentadas as técnicas de avaliação do desempenho preditivo que utilizam a curva ROC ou medidas associadas a essa quantidade. É proposta, também, uma adaptação dos estimadores de Uno *et al.* (2007) da especificidade e sensibilidade que permite a utilização direta de um marcador biológico e também a utilização de um marcador obtido de um outro procedimento de modelagem estatística, assim como de suas respectivas variâncias. Note que, com essa adaptação, também é possível obter a curva ROC e, consequentemente, a área sob essa curva. No capítulo 4 serão apresentados estudos de simulação sob as abordagens função de perda e curva ROC. No capítulo 5 será feita uma aplicação de alguns dos métodos apresentados nos capítulos 2 e 3 num estudo do ICESP. Por fim, as considerações finais serão feitas no capítulo 6.

Capítulo 2

Função de Perda

Nesse capítulo será abordada a avaliação da performance preditiva de modelos utilizando funções de perda. Uma possível forma de se trabalhar com a função de perda nesse contexto é obter um conjunto de dados para gerar um modelo estatístico de predição e outro conjunto de dados independente para fazer a avaliação preditiva do modelo. No entanto, na maioria dos casos, não se dispõe desse segundo conjunto de dados. Para essas ocasiões em que não há um conjunto de dados de validação, serão propostas três abordagens: baseada em modelo, perda aparente e validação cruzada.

Inicialmente, essas abordagens serão apresentadas num contexto mais geral que inclui o caso da análise de sobrevivência sem censuras. Em seguida, será proposta uma abordagem específica para o contexto de análise de sobrevivência em que é possível se observar censuras. Essa abordagem específica utiliza um método de ponderação para compensar as censuras.

Considere, dessa forma, que o interesse está em avaliar a performance preditiva de um modelo estatístico em relação a uma variável resposta Y. Essa variável pode ser a concentração de determinada substância no sangue, um escore que indica a propensão ao cancelamento de um serviço ou o tempo até um novo episódio de depressão.

Para realizar o processo de avaliação da performance preditiva de um modelo, normalmente, selecionam-se n unidades amostrais, coletam-se informações referentes a variável de interesse Y e a um conjunto de covariáveis $\mathbf{Z}$, que acredita-se estar associado a essa variável resposta. O conjunto $D = \{(y_i, \mathbf{z}_i) : i = 1, ..., n\}$ dessas informações referente às n unidades amostrais será chamado de banco de dados de treinamento, pois será utilizado na construção de um modelo preditivo.

Após a coleta dos dados, aplica-se um procedimento de modelagem Q em D que pode incluir especificação do modelo, seleção de variáveis e estimação de parâmetros. Esse procedimento é utilizado para gerar preditores do valor futuro da variável de interesse Y dado o vetor de covariáveis $\mathbf{Z} = \mathbf{z}$. Esses preditores podem ser de duas naturezas: pontuais ou probabilísticos. O primeiro resulta em um preditor, que geralmente é dado pela média ou mediana, $\hat{Y}(\mathbf{Z}) = G(\mathbf{Z}; \hat{\theta})$ para Y, em que $\hat{\theta} = \hat{\theta}(D)$ é uma função do conjunto de dados observados. O segundo resulta em $\hat{F}(y|\mathbf{z})$, ou seja, na estimativa da função de probabilidade de Y dado o vetor de covariáveis $\mathbf{Z}$.

Ao utilizar um procedimento estatístico Q, é necessário avaliar a acurácia do preditor $\hat{Y}(\mathbf{Z})$ gerado por esse procedimento. Para isso, uma possível abordagem é trabalhar com funções de

perda $L(\cdot,\cdot)$. Geralmente as funções de perda são definidas de forma que atribuam perda 0 quando o valor predito é igual ao valor observado e quanto maior a diferença entre esses valores, maior será a perda. Alguns exemplos de funções de perdas para variáveis contínuas são as seguintes:

$$\text{Perda do erro quadrático: } L(Y,\hat{Y}) = (Y-\hat{Y})^2,$$
$$\text{Perda absoluta: } L(Y,\hat{Y}) = |Y-\hat{Y}|.$$

Logo, uma forma de avaliar a acurácia preditiva de um procedimento Q é trabalhar com a esperança da função de perda, ou seja,

$$\pi = E\left[L(Y, G(\mathbf{Z};\hat{\theta}(D)))\right], \tag{2.1}$$

em que a esperança é tomada em relação a D, Y e $\mathbf{Z}$. Essa esperança será chamada de erro de predição. Portanto, quanto menor essa esperança, ou seja, o erro de predição, espera-se que melhor será o desempenho do modelo preditivo. O preditor que minimiza o erro de predição será chamado de preditor ótimo.

É possível, utilizando a propriedade de distribuição condicional, escrever a verdadeira distribuição conjunta de (Y,$\mathbf{Z}$) como $F(y|\mathbf{z})H(\mathbf{z})$, em que $F(y|\mathbf{z})$ é a verdadeira distribuição condicional de Y dado $\mathbf{Z}$ e $H(\mathbf{z})$ é a distribuição marginal do vetor de covariáveis $\mathbf{Z}$. Então, o conjunto de dados de treinamento D = $\{(y_i, \mathbf{z}_i)$, i = 1,...,n$\}$ será considerado como uma amostra aleatória da distribuição conjunta (Y, $\mathbf{Z}$).

Em geral, a aplicação do procedimento Q especifica alguma família de modelos $F_\theta(y|\mathbf{z})$, indexado por um vetor de parâmetros θ, para a aproximação de $F(y|\mathbf{z})$. Portanto, sob $F_\theta(y|\mathbf{z})$, o preditor pontual ótimo será $G(\mathbf{Z};\theta)$ que minimiza o erro de predição $E_{F_\theta}\left[L\left(Y, G(\mathbf{Z})\right)\right]$ entre todas as funções $G(\mathbf{Z})$.

O procedimento Q gera $F_{\hat{\theta}}$ com base em D. Aqui, conforme já mencionado anteriormente, o preditor pontual será denotado por $\hat{Y}(\mathbf{Z}) = G(\mathbf{Z};\hat{\theta})$.

Para saber quão boa é a predição de $G(\mathbf{Z};\hat{\theta})$ em um conjunto de dados independentes, que será chamado de conjunto de dados de teste $D_{teste} = \{(y_j, \mathbf{z}_j), j = 1,...,m\}$, amostrado da mesma população (Y,$\mathbf{Z}$), pode-se utilizar a função de perda $L[y_i, G(\mathbf{z}_i, \hat{\theta}(D))]$ para cada realização de Y_j dado $\mathbf{z}_j$. Assim, uma medida intuitiva de acurácia é dada pela perda média:

$$\frac{1}{m}\sum_{j=1}^{m} L\left(y_j; G\left(\mathbf{z}_j;\hat{\theta}\right)\right), \tag{2.2}$$

em que y_j com $j = 1,...,m$ não pertence ao mesmo conjunto de dados utilizados para se obter $G(\cdot;\cdot)$, mas sim ao conjunto D_{teste}.

Como o interesse está em estimar o erro de predição (2.1), que é dado pela esperança tomada em relação às variáveis Y, $\mathbf{Z}$ e D, surgem, naturalmente, as seguintes quantidades:

$$\pi_1\left(Q;F,\mathbf{z},D\right) = E_Y\left[L\left(Y,\hat{Y}(\mathbf{Z})\right)|\mathbf{Z}=\mathbf{z},D\right] = \int L\left(y, G(\mathbf{z};\hat{\theta})\right) dF(y|\mathbf{z}), \tag{2.3}$$

$$\pi_2\left(Q;F,\mathbf{z}\right) = E_{Y,D}\left[L\left(Y,\hat{Y}(\mathbf{Z})\right)|\mathbf{Z}=\mathbf{z}\right] = E_D\left[\pi_1\left(Q;F,\mathbf{z}\right)\right] \text{ e} \tag{2.4}$$

$$\pi_3\left(Q; F, H_{\mathbf{Z}}\right) = E_{Y,D,\mathbf{Z}}\left[L\left(Y, \hat{Y}(\mathbf{Z})\right)\right] = E_{\mathbf{Z}}\left[\pi_2\left(Q; F, \mathbf{Z}\right)\right], \tag{2.5}$$

em que $Y \sim F(\cdot|\mathbf{z})$, e é independente do conjunto de dados de treinamento D e, portanto, também é independente de $\hat{Y}(\mathbf{Z})$. Fixados o conjunto de dados de treinamento D e o valor da covariável $\mathbf{Z}$, $\pi_1(Q; F, \mathbf{z}, D)$ mede a acurácia de predição do procedimento Q. Ao permitir a variação do conjunto de dados D, isto é, ao tomar a esperança em (2.3) em relação a D obtém-se a perda esperada $\pi_2(Q; F, \mathbf{z})$. Essa perda esperada mede a performance do procedimento Q para o valor fixado da covariável $\mathbf{Z} = \mathbf{z}$ sob a verdadeira distribuição F. Por fim, ao tomar a esperança de (2.4) em relação a variável $\mathbf{Z}$, isto é, considerando todas as variações de $\mathbf{Z}$ obtém-se (2.5). No entanto, a distribuição de probabilidade das covariáveis $H_{\mathbf{Z}}$ normalmente é desconhecida. Por isso, frequentemente estima-se o erro de predição utilizando-se a distribuição empírica de $\tilde{H}_{\mathbf{Z}}$ baseada em $(\mathbf{z}_1, ..., \mathbf{z}_n)$ de D, isto é,

$$\pi_3\left(Q; F, \tilde{H}_{\mathbf{Z}}\right) = \frac{1}{n}\sum_{i=1}^{n} \pi_2\left(Q; F, \mathbf{z}_i\right). \tag{2.6}$$

A partir de agora o foco será na estimação de $\pi_3(Q; F, H_{\mathbf{Z}})$, e muitas vezes será considerado (2.6), uma vez que $\pi_3(Q; F, H_{\mathbf{Z}})$ mede a performance média do procedimento de predição Q sob a verdadeira distribuição de F e $H_{\mathbf{Z}}$, mas (2.6) mede essa performance sob a distribuição empírica de $\mathbf{Z}$. Além disso, $\pi_3(Q, F, H_{\mathbf{z}})$ será denotado apenas por π ou $\pi(Q, F)$ para simplificação de notação.

Observe que a equação (2.2) poderia ser utilizada mesmo quando não se dispõe de um conjunto D_{teste} para estimar o erro de predição. Nesse caso, o conjunto de dados utilizado para construir a função $G(\cdot;\cdot)$ seria o mesmo utilizado para avaliar a acurácia dessa função. Dessa forma, espera-se que a perda média apresente um valor subestimado, o que não é desejável e, portanto, não recomendável.

Em situações que não se dispõe de um conjunto D_{teste} independente do conjunto de treinamento D é possível utilizar algumas abordagens que procuram minimizar esse viés. A seguir são descritas algumas dessas possíveis abordagens. Vale ressaltar que ainda será considerado o contexto sem censura.

2.1 Função de perda para variáveis contínuas

Nessa seção serão apresentados três estimadores para o valor esperado da função de perda quando se tem interesse em avaliar a performance preditiva em relação a uma variável contínua que pode ser dada pelo tempo até a ocorrência de algum evento, mas não se dispõe de um conjunto D_{teste} (Yuan, 2008), ainda considerando-se o contexto sem censuras.

2.1.1 Estimação baseada em modelo

Esse método é baseado na estimação de (2.4) e (2.5) ou (2.6) utilizando um estimador $\hat{F}$ de $F(Y|\mathbf{Z})$, que supõe-se que seja o mecanismo gerador de D e D_{teste}. O estimador baseado em modelo

é da seguinte forma:

$$\begin{aligned}\hat{\pi}^m =& \pi\left(Q; \hat{F}\right) \\ =& \frac{1}{n}\sum_{i=1}^{n} E_{\hat{F}}\left[\int_{-\infty}^{\infty} L\left(y, G(\mathbf{z}_i; \hat{\theta})\right) d\hat{F}(y|\mathbf{z}_i)\right],\end{aligned} \tag{2.7}$$

sendo que a esperança em relação a $\hat{\theta}$, que é uma função de D, é calculada utilizando-se $\hat{F}$ e não a verdadeira distribuição condicional $F(Y|\mathbf{Z})$.

Em algumas situações é conviente que $\hat{\pi}^m$ seja uma função de $\hat{F}_{Y|\mathbf{Z}}$. Por exemplo, quando a distribuição de $\mathbf{Z}$ da população para a qual se tem interesse em fazer previsões seja diferente da distribuição de $\mathbf{Z}$ no conjunto de dados de treinamento.

Para facilitar a compreensão de π_1, π_2, π_3 e $\hat{\pi}^m$, será utilizado um exemplo simples. Considere o contexto de um estudo sem censura em que se tem interesse em comparar o tempo de resposta a um determinado medicamento entre caso e controle. Para cada indivíduo será observado o tempo de resposta Y e a variável Z_i tal que $Z_i = 0$ se o indivíduo pertence ao grupo controle e $Z_i = 1$ caso contrário. Considere, também, que o tempo de resposta para cada grupo segue uma distribuição exponencial de parâmetro $\lambda(z)$. Além disso, considere que são observados n indivíduos no total, sendo que n_0 indivíduos são do grupo caso e n_1 são do grupo controle. Nesse exemplo, será utilizada a mediana de Y dado z_i como preditor do tempo de sobrevivência.

Inicialmente, serão calculados π_1, π_2 e π_3 segundo as fórmulas (2.3), (2.4) e (2.5). Em seguida, será calculado $\pi_2(Q, \hat{F}, z_i)$ para a obtenção de $\hat{\pi}^m$.

- **Cálculo de π_1, π_2 e π_3**

  Utilizando o verdadeiro valor de $\lambda(z)$ para calcular as quantidades (2.3), (2.4) e (2.5) obtém-se que:

$$\begin{aligned}\pi_1(Q; F; z; D) &= E_Y\left[L(Y; \hat{Y}|Z; D)\right] \\ &= \int_0^{\infty} L(y; G(z; \hat{\theta}))dF(y|z) \\ &= \int_0^{\infty} \left|y - \widehat{med}(y)\right| \lambda(z)\exp\left\{-\lambda(z)y\right\} dy \\ &= \int_0^{\infty} \left|y - \frac{\log 2}{\hat{\lambda}(z)}\right| \lambda(z)\exp\left\{-\lambda(z)y\right\} dy \\ &= 2\frac{\exp\left\{-\lambda(z)\frac{\log 2}{\hat{\lambda}(z)}\right\}}{\lambda(z)} - \frac{1}{\lambda(z)} + \frac{\log 2}{\hat{\lambda}(z)}.\end{aligned}$$

  Note que se $\hat{\lambda}(z)$ converge em probabilidade para $\lambda(z)$, então π_1 converge em probabilidade

para $\frac{\log 2}{\lambda(z)}$. Para obter π_2, calcula-se a esperança de π_1 em relação a D da seguinte forma:

$$\begin{aligned}
\pi_2(Q;F;z) &= E_D[\pi_1(Q;F;\mathbf{z};D)] \\
&= E_D\left[2\frac{\exp\left\{-\lambda(z)\frac{\log 2}{\hat{\lambda}(z)}\right\}}{\lambda(z)} - \frac{1}{\lambda(z)} + \frac{\log 2}{\hat{\lambda}(z)}\right] \\
&= E\left[2\frac{\exp\left\{-\lambda(z)\frac{\log 2}{\hat{\lambda}(z)}\right\}}{\lambda(z)}\right] - \frac{1}{\lambda(z)} + E\left[\frac{\log 2}{\hat{\lambda}(z)}\right] \\
&= \frac{2}{\lambda(z)}E\left[\exp\left\{-\lambda(z)\frac{\log 2}{\hat{\lambda}(z)}\right\}\right] - \frac{1}{\lambda(z)} + \log 2E\left[\frac{1}{\hat{\lambda}(z)}\right].
\end{aligned}$$

Lembre que $\hat{\lambda}(z) = \frac{n_z}{\sum_{i=1}^{n_z} Y_i}$. Então, chega-se a

$$\begin{aligned}
E\left[\frac{1}{\hat{\lambda}(z)}\right] =& E\left[\frac{\sum_{i=1}^{n_z} Y_i}{n_z}\right]] = \sum_{i=1}^{n_z}\frac{E\left[Y_i\right]}{n_z} \\
=& \frac{1}{n_z}\sum_{i=1}^{n_z} E\left[Y\right] = \frac{n_z}{n_z}E\left[Y\right] \\
=& \frac{1}{\lambda(z)}.
\end{aligned}$$

Lembre também que se $Y_i \sim \exp(\lambda(z))$, então $T = \sum_{i=1}^{n_z} Y_i \sim \Gamma(n_z, \lambda(z))$. Dessa forma,

$$\begin{aligned}
E\left[\exp\left\{-\lambda(z)\frac{\log 2}{\hat{\lambda}(z)}\right\}\right] =& E\left[\exp\left\{-\lambda(z)\frac{\log 2}{\frac{n_z}{\sum_{i=1}^{n_z} Y_i}}\right\}\right] \\
=& E\left[\exp\left\{-\lambda(z)\frac{\log 2}{n_z}T\right\}\right] \\
=& \int_0^\infty \frac{\exp\left\{-\lambda(z)\frac{\log 2}{n_z}t\right\}\lambda(z)\exp\left\{-\lambda(z)t\right\}(\lambda(z)t)^{n_z-1}}{\Gamma(n_z)}dt \\
=& \left(1+\frac{\log 2}{n_z}\right)^{-n_z}.
\end{aligned}$$

Assim, π_2 será dado por:

$$\begin{aligned}
\pi_2(Q;F;z) =& \frac{2}{\lambda(z)}\left(1+\frac{\log 2}{n_z}\right)^{-nz} - \frac{1}{\lambda(z)} + \log 2\frac{1}{\lambda(z)} \\
=& \frac{1}{\lambda(z)}\left[2\left(1+\frac{\log 2}{n_z}\right)^{-n_z} - 1 + \log 2\right].
\end{aligned}$$

Note que tomando o limite de π_2 para n_z tendendo ao infinito, obtém-se que:

$$\lim_{n_z\to\infty}\pi_2(M;F;z) = \frac{\log 2}{\lambda(z)},$$

que é a quantidade para a qual π_1 converge em probabilidade. Por fim, para obter π_3 calcula-se

a seguinte esperança:

$$\begin{aligned}\pi_3(Q;F;H_Z) &= E_Z[\pi_2(Q;F;Z)] \\ &= E_Z\left\{\frac{1}{\lambda(z)}\left[2\left(1+\frac{\log 2}{n_z}\right)^{-n_z} - 1 + \log 2\right]\right\} \\ &= \sum_{z=0}^{1}\frac{1}{\lambda(z)}\left[2\left(1+\frac{\log 2}{n_z}\right)^{-n_z} - 1 + \log 2\right]P(Z=z).\end{aligned}$$

Note que tomando o limite de π_3 para n_z tendendo ao infinito, obtém-se que:

$$\lim_{n_z\to\infty}\pi_3(Q;F;H_Z) = \sum_{z=0}^{1}\frac{\log 2}{\lambda(z)}P(Z=z).$$

- **Cálculo de $\pi_1(Q,\hat{F},z_i,D)$, $\pi_2(Q,\hat{F},z_i)$ e $\hat{\pi}^m$**

Primeiro será feito o cálculo de $\pi_1(Q,\hat{F},z_i,D)$ da seguinte forma:

$$\begin{aligned}\pi_1\left(Q;\hat{F},z,D\right) =& E_Y\left[L\left(Y,\hat{Y}(Z)\right)|Z=z,D\right] \\ =& \int L\left(y,G(z;\hat{\theta})\right)d\hat{F}(y|z) \\ =& \int\left|y-\widehat{med}(y)\right|d\hat{F}(y|z) \\ =& \int\left|y-\frac{\log 2}{\hat{\lambda}(z)}\right|\hat{\lambda}(z)\exp\left\{-\hat{\lambda}(z)y\right\}dy \\ =& \frac{\log 2}{\hat{\lambda}(z)}.\end{aligned} \tag{2.8}$$

Observe que $\pi_1(Q,\hat{F},Z,D)$ é diferente de $\pi_1(Q,F,Z,D)$. Aqui o estimador de $\lambda(z)$ é dado por $\left\{\sum_{i=1}^{n_z}\frac{y_i}{n_z}\right\}^{-1}$. Assim, utilizando-se esse estimador de $\lambda(z)$, π_2 será calculado pela esperaça de π_1 em relação a D da seguinte forma:

$$\begin{aligned}\pi_2(Q;\hat{F};z) &= E\left[\pi_1(Q;\hat{F};D)\right] = E\left[\frac{\log 2}{\hat{\lambda}(z)}\right] = \log 2E\left[\frac{1}{\hat{\lambda}(z)}\right] \\ &= \log 2E\left[\sum_{i=1}^{n_z}\frac{y_i}{n_z}\right] = \log 2\sum_{i=1}^{n_z}\frac{E\left[Y_i\right]}{n_z} = \frac{\log 2}{n_z}\sum_{i=1}^{n_z}E\left[Y_i\right] \\ &= \frac{\log 2}{n_z}n_zE\left[Y\right] = \log 2E\left[Y\right] = \frac{\log 2}{\lambda(z)}.\end{aligned} \tag{2.9}$$

Observe que, utilizando os resultados (2.8) e (2.9), o estimador baseado em modelo para essa situação será dado por:

$$\begin{aligned}\hat{\pi}^m =& \frac{1}{n}\sum_{i=1}^{n}E_{\hat{F}}\int_{-\infty}^{\infty}L\left(y,G(z_i;\hat{\theta})\right)d\hat{F}(y|z_i) \\ =& \frac{1}{n}\sum_{i=1}^{n}E_{\hat{F}}\left[\frac{\log 2}{\hat{\lambda}(z_i)}\right] \\ =& \frac{1}{n}\sum_{i=1}^{n}\frac{\log 2}{\hat{\lambda}(z_i)}.\end{aligned}$$

Note que, considerando a distribuição de Z, é possível obter o modelo preditivo em uma população com determinada distribuição de Z e aplicá-lo em uma população com uma diferente distribuição das covariáveis Z.

Por fim, tomando a esperança de π_2 em relação a distribuição das covariáveis Z obtém-se para π_3

$$\begin{aligned} E_Z\left[\pi_2\right] =& E\left[\frac{\log 2}{\lambda(z)}\right] = \log 2 E\left[\frac{1}{\lambda(z)}\right] \\ =& \log 2 \sum_{z=0}^{1} \frac{1}{\lambda(z)} P\left(Z=z\right). \end{aligned} \tag{2.10}$$

Embora tenha se calculado essas quantidades analiticamente para esse exemplo com a distribuição exponencial, em outros casos não é possível obter expressões analíticas. Para essas situações, pode-se utilizar métodos de reamostragem para calcular (2.7).

2.1.2 Perda Aparente

Essa técnica, como dito anteriormente, utiliza o mesmo conjunto de dados para gerar os preditores e para a avaliar a performance preditiva. O estimador perda aparente do erro de predição é definido por:

$$\hat{\pi}^{PA} = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, G(\mathbf{z}_i;\hat{\theta})\right) + \text{penalidade}.$$

Essa penalização é feita pois as observações y_i utilizadas na estimação do erro de predição são as mesmas que são utilizadas para gerar $G(\mathbf{z}_i;\hat{\theta})$. Dessa forma, o estimador perda aparente tende a subestimar o erro de predição porque Y_i e $G(\mathbf{z}_i;\hat{\theta})$ são positivamente correlacionados (Yuan, 2008).

A diferença entre o erro de predição e o valor esperado da perda aparente, ou seja, o viés, será denotado por Ω. Isso é,

$$\Omega = \pi - E_D(\hat{\pi}^{PA}).$$

Dessa forma, seria interessante obter uma estimativa de Ω. Um possível estimador é dado por:

$$\widehat{\Omega} = \frac{2}{n}\sum_{i=1}^{n} \text{cov}\left(y_i; f(\hat{y}_i)\right), \tag{2.11}$$

em que a forma funcional de $f(\cdot)$ depende da função de perda utilizada (Efron, 2004). Portanto, se for utilizado esse estimador tem-se que:

$$\hat{\pi} = \hat{\pi}^{PA} + \hat{\Omega} = \hat{\pi}^{PA} + \frac{2}{n}\sum_{i=1}^{n} \widehat{\text{cov}}\left(y_i; f(\hat{y}_i)\right).$$

Uma possível forma de se estimar a covariância entre y_i e $f(\hat{y}_i)$ utilizada por Efron (2004) é fazer uso de técnicas de bootstrap paramétrico. No entanto, para a função de perda absoluta não há nenhuma penalização proposta. Assim, será considerado o estimador perda aparente apenas como $\hat{\pi}^{PA} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y}_i)$, conforme Yuan (2008).

2.1.3 Validação Cruzada

Essa abordagem consiste na separação das n observações do banco de dados de treinamento D em V conjuntos de tamanho n_v $(v = 1, ..., V)$ e, para cada conjunto v, aplica-se o procedimento Q às $(n - n_v)$ observações restantes e calcula-se a perda observada para cada uma das unidades de v. Após a execução dessa rotina, calcula-se a média das funções de perda entre todas as observações. É importante ressaltar que não há nenhum critério definido para determinar o número de grupos V e, também, para selecionar quais n_v unidades serão incluídas no v-ésimo grupo.

Logo, o estimador validação cruzada é da forma

$$\hat{\pi}^{vc} = \frac{1}{n} \sum_{v=1}^{V} \sum_{i \in S_v} L\left(y_i, G\left(\mathbf{z}_i; \hat{\theta}_{(-v)}\right)\right), \tag{2.12}$$

em que S_v é o conjunto formado apenas pelas unidades amostrais do grupo v e $\hat{\theta}_{(-v)}$ é o valor estimado dos parâmetros do modelo sem as observações do grupo v.

Quando V é igual ao tamanho amostral, ou seja, quando todos os conjuntos são compostos por uma única unidade amostral, o estimador de validação cruzada é chamado de "leave-one-out". Dependendo do tamanho amostral e do número de conjuntos V utilizados, $\hat{\pi}^{vc}$ tende a superestimar o erro de predição, pois para cada um dos V conjuntos é gerado um modelo para o preditor $\hat{Y}(\mathbf{Z})$ com base em apenas $(n - n_v)$ unidades amostrais e não em n unidades como anteriormente.

2.2 Função de perda para variável binária

Nessa seção será abordada a situação em que o interesse está na predição de uma variável binária. Essa variável binária pode representar, por exemplo, uma situação em que se observa o tempo até a ocorrência de certo evento e há interesse em se predizer o status da observação num determinado instante de tempo t. Por exemplo, pode se ter interesse em determinar se após 4 meses em terapia um determinado sujeito não terá apresentado uma nova recaída. Nesse caso, o status é definido como uma variável binária denotada por W_t ou $W(t)$ que é dada pela função indicadora $\mathbb{I}(Y > t)$.

Utilizando a abordagem de função de perda para a variável binária W_t, pode-se definir as seguintes funções:

$$\text{Perda 0-1: } L(W(t), \hat{W}(t)) = I(W(t) \neq \hat{W}(t)) = |W(t) - \hat{W}(t)| \text{ e} \tag{2.13}$$

$$\text{Perda Quadrática: } L(W(t), \hat{W}(t)) = (W(t) - \hat{W}(t))^2. \tag{2.14}$$

O estimador $\hat{W}(t)$ pode ser dado utilizando-se uma estimativa da função de probabilidade, ou seja, pode ser de forma que $\hat{W}_t = \mathbb{I}(1 - F(t|\mathbf{z}) \geq 0,5)$. Esse é o estimador que minimiza a perda com a função (2.13) (Ver Apêndice B). Note que, na prática, a distribuição condicional do tempo de sobrevivência é desconhecida. Por isso, será utilizado um estimador de $F(\cdot|\mathbf{z})$ em $\hat{W}_t$.

Utilizando as definições (2.3) e (2.4) dos erros de predição para o caso de uma variável binária tem-se que:

$$\begin{aligned}\pi_{1t}(Q;F,\mathbf{z},D) &= E_{W_t}\left[I\left(W_t \neq \hat{W}_t\right)|\mathbf{Z}=\mathbf{z},D\right] \\ &= 0P\left(I\left(W_t \neq \hat{W}_t\right)=0|\mathbf{Z}=\mathbf{z},D\right)+1P\left(I\left(W_t \neq \hat{W}_t\right)=1|\mathbf{Z}=\mathbf{z},D\right) \\ &= P\left(W_t=1,\hat{W}_t=0|\mathbf{Z}=\mathbf{z},D\right)+P\left(W_t=0,\hat{W}_t=1|\mathbf{Z}=\mathbf{z},D\right) \\ &= P\left(W_t=1|\mathbf{Z}=\mathbf{z}\right)I\left(\hat{W}_t=0\right)+P\left(W_t=0|\mathbf{Z}=\mathbf{z}\right)I\left(\hat{W}_t=1\right).\end{aligned}$$

Tomando a esperança de π_{1t} em relação a D, chega-se a:

$$\begin{aligned}\pi_{2t}(Q;F,\mathbf{z}) &= E_D\left[\pi_{1t}(Q;F,\mathbf{z},D)\right] \\ &= E_D\left[P\left(W_t=1|\mathbf{Z}=\mathbf{z}\right)I\left(\hat{W}_t=0\right)+P\left(W_t=0|\mathbf{Z}=\mathbf{z}\right)I\left(\hat{W}_t=1\right)\right] \\ &= P\left(W_t=1|\mathbf{Z}=\mathbf{z}\right)E_D\left[I\left(\hat{W}_t=0\right)\right]+P\left(W_t=0|\mathbf{Z}=\mathbf{z}\right)E_D\left[I\left(\hat{W}_t=1\right)\right] \\ &= P\left(W_t=1|\mathbf{Z}=\mathbf{z}\right)P\left(\hat{W}_t=0|\mathbf{Z}=\mathbf{z}\right)+P\left(W_t=0|\mathbf{Z}=\mathbf{z}\right)P\left(\hat{W}_t=1|\mathbf{Z}=\mathbf{z}\right) \\ &= S(t|\mathbf{z})P\left(\hat{W}_t=0|\mathbf{Z}=\mathbf{z}\right)+\left(1-S(t|\mathbf{z})\right)P\left(\hat{W}_t=1|\mathbf{Z}=\mathbf{z}\right) \\ &= S(t|\mathbf{z})P\left(S_{\hat{\theta}}(t|\mathbf{z})\leq 0,5\right)+\left(1-S(t|\mathbf{z})\right)P\left(S_{\hat{\theta}}(t|\mathbf{z})>0,5\right).\end{aligned}$$

Além disso, o erro de predição definido em (2.6) será dado por

$$\begin{aligned}\pi_{3t}\left(Q;F,\widetilde{H}_Z\right) &= E_Z\left[\pi_{2t}(Q;F,\mathbf{z})\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[S(t|\mathbf{z}_i)P\left(S_{\hat{\theta}}(t|\mathbf{z}_i)\leq 0,5\right)+\left(1-S(t|\mathbf{z}_i)\right)P\left(S_{\hat{\theta}}(t|\mathbf{z}_i)>0,5\right)\right].\end{aligned}$$

Utilizando-se essas quantidades, é possível definir, como no caso para variáveis contínuas, os estimadores descritos a seguir.

2.2.1 Estimação baseada em modelo

O estimador do erro de predição para uma variável binária é obtido em duas etapas. Primeiro substitui-se $S(t|\mathbf{z})$ pelo estimador $S_{\hat{\theta}}(t|\mathbf{z})$ em π_{2t} da seguinte forma:

$$\hat{\pi}_{2t}^{m}=S_{\hat{\theta}}(t|\mathbf{z})\widehat{P}\left(S_{\hat{\theta}}(t|\mathbf{z})\leq 0,5\right)+\left(1-S_{\hat{\theta}}(t|\mathbf{z})\right)\widehat{P}\left(S_{\hat{\theta}}(t|\mathbf{z})>0,5\right). \tag{2.15}$$

Pode-se utilizar técnicas de bootstrap paramétricas e não-paramétricas para se obter as estimativas $\widehat{P}(S_{\hat{\theta}}(t|\mathbf{z})\leq 0,5)$. No procedimento paramétrico, gera-se K conjuntos $D_1^*,...D_K^*$ de dados de treinamento $\{(Y_i^*,\mathbf{z}_i),i=1,...,n\}$ utilizando a função estimada $F_{\hat{\theta}}(y|\mathbf{z})$ e aplica-se o procedimento de modelagem Q a D_k^* para produzir o preditor $\widehat{W}_t^{*k}=I\left[S_{\hat{\theta}_k^*}(t|\mathbf{z})>0,5\right]$. Então, a estimativa será dada por

$$\widehat{P}\left(S_{\hat{\theta}}(t|\mathbf{z})\leq 0,5\right)=\frac{1}{K}\sum_{k=1}^{K}I\left(S_{\hat{\theta}_k^*}(t|\mathbf{z})\leq 0,5\right), \tag{2.16}$$

para qualquer $\mathbf{z}$. Já o procedimento não-paramétrico consiste na retirada de K amostras com reposição do conjunto de dados de treinamento e na aplicação do procedimento de modelagem para cada amostra. Em seguida, calcula-se a estimativa de $\widehat{P}\left(S_{\hat{\theta}}(t|\mathbf{z}) \leq 0,5\right)$ como indicado em (2.16).

Na segunda etapa, utiliza-se o estimador de π_{2t} para se obter o estimador (2.6) do erro de predição baseado em modelo da seguinte forma:

$$\begin{aligned}\hat{\pi}_{3t}^{m} &= \frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_{2t}(Q;F;\mathbf{z}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} S_{\hat{\theta}}(t|\mathbf{z}_i)\widehat{P}\left(S_{\hat{\theta}}(t|\mathbf{z}_i) \leq 0,5\right) + (1 - S_{\hat{\theta}}(t|\mathbf{z}_i))\widehat{P}\left(S_{\hat{\theta}}(t|\mathbf{z}_i) > 0,5\right). \end{aligned} \tag{2.17}$$

2.2.2 Perda Aparente:

O estimador perda aparente do erro de predição é dado por:

$$\hat{\pi}_t^{PA} = \frac{1}{n}\sum_{i=1}^{n} I\left(w_{i,t} \neq \hat{w}_{i,t}\right) + \widehat{\Omega}_t, \tag{2.18}$$

em que $w_{i,t} = I(Y_i > t)$, $\hat{w}_{i,t} = I(S_{\hat{\theta}}(t|\mathbf{z}_i) > 0,5)$ e $\widehat{\Omega}_t = \frac{2}{n}\sum_{i=1}^{n}\text{cov}(w_{i,t}, f(\hat{w}_{i,t})$. Nesse caso, para a função de perda 0-1, $f(\hat{w}_{i,t})$ em (2.11) é dado por $\hat{w}_{i,t}$. Assim, para estimar a $\text{cov}(w_{i,t}, \hat{w}_{i,t})$, Yuan (2008) utilizou uma abordagem baseada no modelo, como sugerida por Efron (2004), utilizando simulações geradas a partir de $\hat{F}_T$.

2.2.3 Validação Cruzada

O estimador validação cruzada do erro de predição é dado por:

$$\hat{\pi}_t^{vc} = \frac{1}{n}\sum_{v=1}^{V}\sum_{i\in S_v} I\left(w_{i,t} \neq \hat{w}_{i(-v),t}\right), \tag{2.19}$$

em que $\hat{w}_{i(-v),t}$ é o status predito para i-ésima observação utilizando-se o modelo construído sem considerar as observações pertencentes ao v-ésimo grupo.

Nas seções anteriores foram apresentados estimadores do erro de predição para variáveis contínuas e binárias que podem representar, respectivamente, o tempo e o status de sobrevivência. No entanto, esses estimadores não consideram a possibilidade da observação de censuras. Na próxima seção serão apresentadas algumas alternativas para esses estimadores que consideram a presença de dados censurados.

2.3 Estimação do erro de predição na presença de censura

Na prática, quando se trabalha com dados de análise de sobrevivência, podem ser observadas censuras. Nos casos em que há apenas censura à direita podem ser observadas as seguintes quantidades:

$\mathbf{Z_i}$: vetor de covariáveis associadas ao evento de interesse referente à i-ésima unidade amostral,

X_i : mínimo entre o tempo de sobrevivência Y_i e o tempo de censura C_i associados a i-ésima unidade amostral e

δ_i: indicador de falha para a i-ésima unidade amostral, ou seja, $I(Y_i \leq C_i)$, $i = 1, ..., n$.

Note que, na presença de dados censurados, o estimador

$$\hat{\pi} = \frac{1}{m} \sum_{j=1}^{m} L(y_j; \hat{y}_j) \tag{2.20}$$

não poderá ser utilizado, pois não será possível calcular a função $L(\cdot, \cdot)$ para as observações censuradas. Nesses casos, em que não é possível calcular o valor da função de perda para algumas unidades, pode se utilizar algumas alternativas como ponderação e imputação.

2.3.1 Ponderação pelo inverso da probabilidade de censura (IPCW - *Inverse Probability of Censoring Weighting*)

Uma alternativa possível para se utilizar quando se observa censura é utilizar uma técnica de ponderação pelo inverso da probabilidade associada a censura. Essa técnica foi considerada inicialmente no estimador do total de Horvitz-Thompson (Horvitz e Thompson, 1952). Nesse estimador considera-se que cada uma unidade amostral será incluída na amostra com probabilidade η_i sendo que $i = 1, ..., N$. Considera-se, também, que o interesse está em estimar o total populacional $U = \sum_{i=1}^{N} u_i$.

O estimador não viesado do total populacional de Horvitz-Thompson é dado por:

$$\hat{U}_{HT} = \sum_{i=1}^{n} \frac{u_i}{\eta_i}.$$

Considere a_i uma variável binária tal que $a_i = 1$ se a i-ésima unidade pertence à amostra e $a_i = 0$ caso contrário. Dessa forma, a_i segue uma distribuição binomial com número de ensaios igual a 1 e probabilidade de sucesso igual a η_i, tal que $E[a_i] = \eta_i$ e $Var(a_i) = \eta_i(1 - \eta_i)$ (Casella e Berger, 2001). Portanto, considerando u_i fixo, tem-se que:

$$\begin{aligned} E[\hat{U}_{HT}] =& E\left[\sum_{i=1}^{n} \frac{u_i}{\eta_i}\right] = E\left[\sum_{i=1}^{N} \frac{a_i u_i}{\eta_i}\right] = \sum_{i=1}^{N} \frac{u_i}{\eta_i} E\left[a_i\right] \\ =& \sum_{i=1}^{N} \frac{u_i}{\eta_i} \eta_i = \sum_{i=1}^{N} u_i \\ =& U. \end{aligned}$$

A ideia do IPCW é utilizar uma versão do estimador de Horvitz-Thompson com o objetivo de estimar $L = \sum_{i=1}^{n} L(y_i; \hat{y}_i)$, pois para algumas unidades a função $L(\cdot, \cdot)$ não poderá ser calculada devido a censura. Dessa forma, agora considera-se a amostra selecionada como sendo a população para a qual se tem interesse em estimar o total L. Note que, nesse contexto, a probabilidade de inclusão na amostra será dada pela probabilidade da unidade não ser censurada, ou seja, será dada pela sobrevivência da censura, isto é $S^c(y) = P(C \geq y)$, no instante y_i.

Nesse método de ponderação, atribui-se peso α_j^{-1} para o j-ésimo indivíduo ao qual o tempo de sobrevivência y_j é observado, em que $\alpha_j = P(C \geq y_j) = S^c(y_j)$ é a probabilidade de que o tempo de censura da j-ésima unidade amostral seja maior do que o instante de falha observado y_j. Note que aqui é considerada a hipótese de que a censura é independente das covariáveis, por isso é utilizada a distribuição marginal do tempo de censura.

Os indivíduos que foram censurados recebem peso 0. Note que a contribuição dos indivíduos censurados se dá pelos pesos α_j's que são determinados com a utilização dos tempos de censura observados. Assim, o estimador ponderado será dado por:

$$\hat{\pi}^w = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\delta_j}{\alpha_j}L(y_j;\hat{y}_j)\right), \tag{2.21}$$

com $y_j \in D_{Teste}$ e j = 1,...,m.

Dado que C é completamente independente de Y e $\mathbf{Z}$, $E_C[\hat{\pi}^w]$ é igual a $\hat{\pi} = \frac{\sum_{j=1}^{m} L(y_j;\hat{y}_j)}{m}$ em (2.20), conforme pode ser verificado a seguir:

$$\begin{aligned}
E_C\left[\frac{1}{m}\sum_{i=1}^{m}\frac{\delta_i}{\alpha_i}L(y_i;\hat{y}_i)\right] =& \frac{1}{m}\sum_{i=1}^{m}E_C\left[\frac{\mathbb{I}(Y_i \leq C_i)}{P(C \geq y_i)}L(y_i;\hat{y}_i)\right] \\
=& \frac{1}{m}\sum_{i=1}^{m}\frac{L(y_i;\hat{y}_i)}{P(C \geq y_i)}E\left[\mathbb{I}(Y_i \leq C_i)\right] \\
=& \frac{1}{m}\sum_{i=1}^{m}\frac{L(y_i;\hat{y}_i)}{P(C \geq y_i)}P(y_i \leq C_i) \\
=& \frac{1}{m}\sum_{i=1}^{m}\frac{L(y_i;\hat{y}_i)}{P(C \geq y_i)}P(C_i \geq y_i) \\
=& \frac{1}{m}\sum_{i=1}^{m}L(y_i;\hat{y}_i).
\end{aligned}$$

Sob certas condições, Rosthoj e Keiding (2004) mostraram que $\hat{\pi}^w$ é um estimador consistente de π para (2.5) ou (2.6). Note que anteriormente foi feita a hipótese de que a censura era independente das covariáveis $\mathbf{Z}$. No entanto, essa hipótese é muito restritiva. Uma abordagem mais ampla e realista seria considerar que a distribuição do tempo de censura C condicional às covariaveis $\mathbf{Z}$, mas com o tempo de sobrevivência Y e o tempo de censura C condicionalmente independentes dadas as covariáveis $\mathbf{Z}$. Para esses casos em que se quer considerar a censura condicionalmente independente, foi proposta uma abordagem na qual se utiliza a função de perda $L(W_t.\hat{W}_t) = (W_t - \hat{W}_t)^2$ (Gerds e Schumacher, 2006). Nessa abordagem substitui-se α, que anteriormente foi considerado como a distribuição marginal da censura, por uma estimativa de $S^c(c|\mathbf{z}) = P(C > c|\mathbf{Z} = \mathbf{z})$ em (2.21), ou seja, agora se considera a função de sobrevivência condicional da censura, mas utilizando-se a função de perda quadrática para o status de sobrevivência. Dessa forma, obtém-se:

$$\hat{\pi}^w = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\delta_j}{\hat{S}^c(y_j|\mathbf{z}_j)}L(y_j;\hat{y}_j)\right). \tag{2.22}$$

A abordagem anterior foi proposta apenas para a função de perda quadrática para o status de sobrevivência (2.14). No entanto, é possível utilizar uma abordagem mais geral, na qual pode se

utilizar uma função de perda arbitrária (Yuan, 2008). Essa abordagem mais geral tem como caso específico a equação (2.22) e será apresentada a seguir.

Sejam

$$\Delta_j = I\left\{L(Y_j; G(\mathbf{Z}_j; \hat{\theta})) \text{ é conhecido}\right\} \text{ e } \alpha_j = P(\Delta_j = 1|Y_j, \mathbf{Z}_j, D).$$

Note que, dados $Y_j, \mathbf{Z}_j$ e D, a variável aleatória Δ_j depende apenas do tempo de censura C_j. Portanto, α_j é determinado pela distribuição condicional da censura $S^c(c|\mathbf{z})$.

Assim, o estimador IPCW proposto em Yuan (2008), para $\alpha_j > 0$ e função de perda arbitrária $L(Y, \hat{Y})$, é dado por:

$$\hat{\pi}^w = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\Delta_j}{\hat{\alpha}_j}L(y_j; \hat{y}_j)\right), \tag{2.23}$$

em que $\hat{\alpha}_j$ é uma estimativa de $P(\Delta_j = 1|Y_j, \mathbf{Z}_j, D)$ para o tempo apropriado a j-ésima unidade amostral. A motivação da equação (2.23) se dá em função da seguinte igualdade:

$$E_{Y,Z,C,D}\left[\frac{1}{m}\sum_{j=1}^{m}\frac{\Delta_j}{\alpha_j}L(Y_j, \hat{Y}_j)\right] = E_{Y,Z,D}\left[\frac{1}{m}\sum_{j=1}^{m}L(Y_j, \hat{Y}_j)\right].$$

Essa igualdade é válida pois

$$\begin{aligned}
E_{Y,Z,C,D}\left[\frac{1}{m}\sum_{j=1}^{m}\frac{\Delta_j}{\alpha_j}L(Y_j, \hat{Y}_j)\right] &= E_{Y,Z,D}\left\{E_{C|Y,Z,D}\left[\frac{1}{m}\sum_{j=1}^{m}\frac{\Delta_j}{\alpha_j}L(Y_j, \hat{Y}_j)\right]\right\} \\
&= E_{Y,Z,D}\left\{\frac{1}{m}\sum_{j=1}^{m}\frac{L(Y_j, \hat{Y}_j)}{\alpha_j}E_{C|Y,Z,D}\left[\Delta_j\right]\right\} \\
&= E_{Y,Z,D}\left\{\frac{1}{m}\sum_{j=1}^{m}\frac{L(Y_j, \hat{Y}_j)}{\alpha_j}P(\Delta_j = 1|Y, \mathbf{Z}, D)\right\}.
\end{aligned}$$

Mas, como $\alpha_j = P(\Delta_j = 1|Y, \mathbf{Z}, D)$, tem-se:

$$\begin{aligned}
E_{Y,Z,C,D}\left[\frac{1}{m}\sum_{j=1}^{m}\frac{\Delta_j}{\alpha_j}L(Y_j, \hat{Y}_j)\right] &= E_{Y,Z,D}\left\{\frac{1}{m}\sum_{j=1}^{m}\frac{L(Y_j, \hat{Y}_j)}{\alpha_j}\alpha_j\right\} \\
&= E_{Y,Z,D}\left\{\frac{1}{m}\sum_{j=1}^{m}L(Y_j, \hat{Y}_j)\right\}.
\end{aligned}$$

Na próxima seção serão apresentados os estimadores perda aparente e validação cruzada do erro de predição utilizando a técnica IPCW para o tempo e status de sobrevivência. Note que não há necessidade de ponderação para o estimador baseado em modelo pois nesse estimador não se utiliza os tempos de sobrevivência observados para se calcular diretamente $L(y_i, \hat{y}_i)$, mas sim para se determinar a função $F_{\hat{\theta}}$ que será utilizada na equação (2.7).

2.3.1.1 Aplicação a tempo de sobrevivência

Na seção 2.1 foram apresentados os estimadores baseado em modelo, perda aparente e validação cruzada fora do contexto de análise de sobrevivência. Inicialmente será apresentado o estimador baseado em modelo e posteriormente os estimadores perda aparente e validação cruzada utilizando a técnica IPCW (Yuan, 2008).

- **Baseado em Modelo**

  Para esse estimador não é necessário utilizar a técnica IPCW, pois a função de perda $L(\cdot,\cdot)$ não é calculada diretamente nos tempos de sobrevivência observados, mas nos tempos de sobrevivência gerados utilizando-se o modelo ajustado $F_{\hat{\theta}}(y|\mathbf{z})$. Esse modelo é obtido aplicando-se o procedimento de modelagem Q ao conjunto de dados D. Para obter o estimador baseado em modelo gera-se J conjuntos de dados de treinamento $(D_j : j = 1, ..., J)$ e de teste $(D_j^T : j = 1, ..., J)$ baseados em $F_{\hat{\theta}}(\cdot)$. Então, aplica-se o procedimento de modelagem Q a D_j para obter $F_{\hat{\theta}_j}(\cdot)$ que será utilizada como função preditora para cada elemento de D_j^T. Assim, obtém-se a perda relativa para cada valor de j

  $$PR_j = \frac{1}{n}\sum_{i=1}^{n} L(y_{ij}^*; \hat{y}_{ij}^*).$$

  Repete-se esse procedimento J vezes para se obter o estimador da perda baseado em modelo da seguinte forma:

  $$\hat{\pi}^m = \frac{1}{J}\sum_{j=1}^{J} PR_j = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{n}\sum_{i=1}^{n} L(y_{ij}^*; \hat{y}_{ij}^*). \tag{2.24}$$

  Esse procedimento é proposto porque, com exceção do caso exponencial apresentado na seção 2.1, não é possível obter uma forma analítica para a expressão dada em (2.7).

- **Perda Aparente**

  Utilizando (2.23) para o estimador perda aparente para o erro de predição obtém-se

  $$\hat{\pi}^{PA} = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{\alpha}_i} L(y_i, \hat{y}_i) + \hat{\Omega}.$$

  Note que, nesse contexto, a função de perda só será conhecida para os casos que não foram censurados. Por isso, $\hat{\alpha}_i = P(\Delta_i = \widehat{1|Y_j}, \mathbf{Z}_j, D)$ é dado por $\hat{S}^c(y_j|\mathbf{z}_j)$. Logo, o estimador de perda aparente é obtido da seguinte forma:

  $$\hat{\pi}^{PA} = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{S}^c(y_j|\mathbf{z}_j)} L(y_i, \hat{y}_i) + \hat{\Omega}. \tag{2.25}$$

  Como para o caso da função de perda absoluta não foi proposto nenhum termo de penalização por Efron (2004) e Yuan (2008), aqui será utilizado o estimador sem a penalização.

- **Validação Cruzada**

  O estimador de validação cruzada para o erro de predição utilizando (2.23) é dado por:

$$\hat{\pi}^{vc} = \frac{1}{n}\sum_{v=1}^{V}\sum_{i\in S_v}\frac{\Delta_i}{\hat{\alpha}_i}L(y_i, \hat{y}_{i(-v)}).$$

  Aqui, assim como no estimador perda aparente, $\hat{\alpha}_i$ é dado por $\hat{S}^c(y_j|\mathbf{z}_j)$. Logo, o estimador de validação cruzada é obtido da seguinte forma:

$$\hat{\pi}^{vc} = \frac{1}{n}\sum_{v=1}^{V}\sum_{i\in S_v}\frac{\Delta_i}{\hat{S}^c(y_j|\mathbf{z}_j)}L(y_i, \hat{y}_{i(-v)}). \tag{2.26}$$

2.3.1.2 Aplicação a status de sobrevivência

Na seção (2.2) foram apresentados os estimadores perda aparente e validação cruzada fora do contexto de análise de sobrevivência. Lembre que o status de sobrevivência da i-ésima unidade amostral para um instante de tempo fixado t_0 é dado pela variável $w_{i,t_0} = I(Y_i > t_0)$. A seguir serão apresentados os estimadores perda aparente e validação cruzada do erro de predição utilizando a técnica IPCW (Yuan, 2008).

- **Baseado em Modelo**

  O estimador baseado em modelo para o status de sobrevivência é calculado da mesma forma que foi apresentada na seção 2.2.1 utilizando-se as estimativas $S_{\hat{\theta}}(t|\mathbf{z})$ e $\widehat{P}\left(S_{\hat{\theta}}(t|\mathbf{z}) \leq 0,5\right)$ obtidas com a aplicação do procedimento de modelagem Q.

- **Perda Aparente**

  O estimador perda aparente para o status de sobrevivência é dado por:

$$\hat{\pi}^{PA}_{t_0} = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{\alpha}_i}I\left(w_{i,t_0} \neq \hat{w}_{i,t_0}\right),$$

  em que $\hat{w}_{i,t_0}$ é um estimador do status de sobrevivência no instante t_0 que pode ser dado por $I(S_{\hat{\theta}}(t_0|\mathbf{z}_i) > 0,5)$, Δ_i é a variável que indica se a função de perda é conhecida e α é a probabilidade da função de perda ser conhecida. Note que a função de perda será conhecida se a observação não for censurada ou se a censura for num tempo superior a t_0. Logo,

$$\alpha = P(\Delta = 1|Y, \mathbf{Z}, D) = P((C \geq Y) \cup (C < Y; C > t_0)) = P(C > \min(Y, t_0)).$$

  Assim, o estimador perda aparente para o status de sobrevivência é obtido da seguinte forma:

$$\hat{\pi}^{PA}_{t_0} = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{S}^c(\min(y_i, t_0)|\mathbf{z}_i)}I\left(w_{i,t_0} \neq \hat{w}_{i,t_0}\right), \tag{2.27}$$

  Conforme já discutido, é possível adicionar um termo de penalização (Efron, 2004). Utilizando esse termo o estimador ficaria da seguinte forma:

$$\begin{aligned}\hat{\pi}^{PA}_{t_0} =& \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{S}^c(\min(y_i, t_0)|z_i)}I\left(w_{i,t_0} \neq \hat{w}_{i,t_0}\right) + \hat{\Omega}_t \\ =& \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{S}^c(\min(y_i, t_0)|z_i)}I\left(w_{i,t_0} \neq \hat{w}_{i,t_0}\right) + \frac{2}{n}\sum_{i=1}^{n}\text{cov}(w_{i,t_0}; f(\hat{w}_{i,t_0})).\end{aligned}$$

No caso da função de perda 0-1, definida em (2.13), $f(\hat{w}_{t_0})$ é dado por $\hat{w}_{t_0}$. Uma possível forma de se estimar $\text{cov}(w_{t_0}, \hat{w}_{t_0})$ foi utilizada por Yuan (2008) com a abordagem baseada em modelo para a simulação de $\hat{F}_T$, como sugerido em Efron (2004).

- **Validação Cruzada**

  O estimador de validação cruzada do erro de predição é dado por:

$$\hat{\pi}_{t_0}^{vc} = \frac{1}{n}\sum_{v=1}^{V}\sum_{i\in S_v}\frac{\Delta_i}{\hat{\alpha}_i} I\left(w_{i,t_0} \neq \hat{w}_{i(-v),t_0}\right),$$

  em que $\hat{w}_{i(-v),t_0}$ é o status predito para i-ésima observação utilizando-se o modelo construído sem considerar as observações pertencentes ao v-ésimo grupo no instante t_0, Δ é uma função indicadora da perda ser conhecida e α é a probabilidade dessa função ser conhecida. Note que, assim como no estimador perda aparente, $\hat{\alpha}$ é dado por $P(C > \min(Y, t_0))$. Logo, o estimador validação cruzada é obtido da seguinte forma:

$$\hat{\pi}_{t_0}^{vc} = \frac{1}{n}\sum_{v=1}^{V}\sum_{i\in S_v}\frac{\Delta_i}{\hat{S}^c(\min(y_i, t_0)|z_i)} I\left(w_{i,t_0} \neq \hat{w}_{i(-v),t_0}\right). \tag{2.28}$$

No próximo capítulo serão apresentadas técnicas de avaliação do desempenho preditivo de modelos utilizando a curva ROC e medidas associadas à essa quantidade. Note que ao utilizar a curva ROC, avalia-se o status de cada unidade amostral num instante de tempo fixado. Nesse sentido, há uma intersecção das técnicas para o status de sobrevivência apresentadas no capítulo 2 com as técnicas apresentadas no capítulo 3. No entanto, para melhor organização do texto, optou-se por apresentá-las no capítulo seguinte.

Capítulo 3

Curva ROC

Inicialmente serão definidas algumas quantidades e funções utilizadas fora do contexto de análise de sobrevivência, ou seja, num contexto em que o status das unidades não muda de acordo com o tempo e também são não observadas censuras. Considere, novamente, que o objetivo é avaliar o grau de eficiência com que se consegue diferenciar dois grupos, caso e controle, baseado numa escala contínua. Por exemplo, pode se ter interesse em utilizar uma medida de concentração de determinada substância no sangue para se diferenciar sujeitos saudáveis de sujeitos doentes.

Posteriormente, essas quantidades serão estendidas para o contexto de análise de sobrevivência e, após essa extensão, serão apresentados alguns métodos de estimação para essas quantidades.

3.1 Curva ROC

Em muitas situações, há interesse em definir um critério de dicotomização para uma escala contínua. Por exemplo, dado um nível de colesterol, o paciente deve ser encaminhado ou não a um tratamento? Nessas situações, uma ferramenta útil para avaliar a performance dessa dicotomização é dada pela curva *receiver operating characteristc* (ROC).

Essa metodologia foi desenvolvida no contexto de análise de sinais de radares durante a segunda guerra mundial e na teoria de detecção de sinais na década de 1950 (Krzanowski e Hand, 2009). O termo *receiver operating characteristic* (ROC) apareceu primeiro em um relatório técnico em 1953 (Peterson *et al.*, 1953). Apesar do desenvolvimento da técnica ter sido na década de 1950, o seu potencial de utilização na área médica foi reconhecido apenas no início da década de 1960 (Lusted, 1960). Posteriormente, essa técnica começou a ser utilizada em diversas áreas como radiologia, epidemiologia, química e economia.

Nesse trabalho, a escala em que se tem interesse em definir um critério de dicotomização será dada por um marcador M. Esse marcador pode ser biológico ou dado por um escore obtido, por exemplo, por um modelo de regressão. Um exemplo de marcador biológico pode ser a concentração de determinada substância relacionada a uma doença no sangue. Outro tipo de marcador pode ser dado pelo preditor linear de um modelo de regressão em que se considera um determinado conjunto de covariáveis que espera-se que estejam associadas ao evento de interesse.

Para entender a utilização da curva ROC, considere o marcador contínuo M de forma que altos

valores estejam associados a um resultado positivo e utiliza-se a seguinte função para dicotomizar os resultados:

- Positivo se $M > c$;
- Negativo se $M \leq c$.

Ao utilizar essa regra de classificação, podem ocorrer quatro situações:

- **Verdadeiro Positivo**: o objeto é classificado como positivo quando, de fato, ele pertence ao grupo positivo;
- **Falso Positivo**: o objeto é classificado como positivo quando na verdade ele pertence ao grupo negativo;
- **Verdadeiro Negativo**: o objeto é classificado como negativo quando, de fato, ele pertence ao grupo negativo; e
- **Falso Negativo**: o objeto é classificado como negativo quando na verdade ele pertence ao grupo positivo.

Essas quatro situações podem ser resumidas na Tabela 3.1.

Tabela 3.1: Possíveis classificações

	Classificação	
Grupo	Positivo	Negativo
Positivo	Verdadeiro Positivo	Falso Negativo
Negativo	Falso Positivo	Verdadeiro Negativo

É necessário, antes de definir a curva ROC, determinar algumas quantidades como a taxa de falso positivo (TFP) e a taxa de verdadeiro positivo (TVP). Considere W uma variável tal que se W é igual a um, o objeto pertence ao grupo positivo, e se W é igual a zero o objeto pertence ao grupo negativo. A TFP é dada pela probabilidade de se classificar um objeto como positivo quando na realidade ele pertence ao grupo negativo, ou seja, $TFP(c) = P(M > c|W = 0)$. Já a TVP é dada pela probabilidade de se classificar um objeto como positivo quando na realidade ele de fato pertence ao grupo positivo, ou seja, $TVP(c) = P(M > c|W = 1)$.

No contexto de curva ROC, é muito comum se utilizar outras definições que são baseadas na TFP e na TVP. As principais são as definições de sensibilidade e especificidade. A sensibilidade é equivalente a taxa de verdadeiro positivo e a especificidade é a probabilidade de classificar um objeto como negativo quando ele, de fato, pertence ao grupo negativo (Vittinghof *et al.*, 2004). Utilizando a regra de classificação citada anteriormente, essas definições são dadas por:

$$\text{sensibilidade} = P(M > c|W = 1) = TVP(c) \text{ e} \tag{3.1}$$

$$\text{especificidade} = P(M \leq c|W = 0) = 1 - TFP(c). \tag{3.2}$$

Utilizando as definições anteriores, é possível agora definir mais precisamente a curva ROC. A curva ROC descreve o desempenho de classificação de acordo com vários valores de corte c. Essa curva é dada por um gráfico da TVP versus TFP para os diferentes valores de c. Portanto, uma forma de se definir a curva é dada por:

$$ROC(\cdot) = \{(TFP(c), TVP(c)), c \in (-\infty, \infty)\}.$$

Além disso, utilizando as definições de sensibilidade e especificidade, é possível definir a curva ROC da seguinte forma:

$$ROC(\cdot) = [\{(1 - \text{especificidade}(c)), \text{sensibilidade}(c)\}, c \in (-\infty, \infty)].$$

Uma outra forma de se definir a curva ROC utiliza a seguinte representação (Heagerty e Zheng, 2005):

$$ROC(a) = TVP\left\{TFP^{-1}(a)\right\}, a \in (0,1), \tag{3.3}$$

pois $TFP(c) = P(M > c|W = 0) = a$, então $TFP^{-1}(a) = c$. Assim, a taxa de verdadeiro positivo será $TVP\left\{TFP^{-1}(a)\right\} = TVP(c) = P(M > c|W = 1)$.

Propriedades da curva ROC

A curva ROC apresenta algumas propriedades descritas a seguir:

1. É uma função estritamente crescente, pois:

$$\begin{array}{lll} \lim\limits_{c\to-\infty} TVP(c) = 1 & \text{e} & \lim\limits_{c\to-\infty} TFP(c) = 1; \\ \lim\limits_{c\to\infty} TVP(c) = 0 & \text{e} & \lim\limits_{c\to\infty} TFP(c) = 0. \end{array}$$

2. É invariante a transformações estritamente crescentes em M (Pepe, 2003).

   Prova: Seja h uma transformação estritamente crescente, $J = h(M)$ e $s = h(c)$. Então, vale que

$$\begin{aligned} P(J > s|W = 0) =& P(h^{-1}(J) > h^{-1}(s)|W = 0) \\ =& P(h^{-1}(h(M)) > h^{-1}(h(c))|W = 0) \\ =& P(M > c|W = 0). \square \end{aligned}$$

   O mesmo vale quando $P(M > c|W = 1)$.

Exemplo de curva ROC

Para ilustrar uma aplicação da curva ROC, considere uma situação hipotética bem simples em que se quer diferenciar pessoas saudáveis de pessoas com determinada doença utilizando um determinado marcador M. Assim, foram geradas as informações presentes na Tabela 3.2 sobre os marcadores de 15 pessoas doentes e 15 pessoas saudáveis.

Tabela 3.2: Marcadores

Controle		Caso	
13,022	15,048	15,176	17,125
13,349	15,176	15,199	17,431
13,749	15,410	15,417	17,575
13,832	15,547	16,049	17,581
14,307	15,687	16,364	17,810
14,341	15,804	16,373	17,857
14,543	16,534	16,647	18,160
14,792		17,057	

Considere, também, a variável indicadora G que é igual a um se o indivíduo pertence ao grupo caso e zero, caso contrário. Uma forma de se obter a cuva ROC é utilizar os seguintes estimadores:

$$\hat{TVP}(c) = \frac{\sum_{i=1}^{30} \mathbb{I}(M_i > c; G_i = 1)}{\sum_{i=1}^{30} \mathbb{I}(G_i = 1)} \quad \text{e} \quad \hat{TFP}(c) = \frac{\sum_{i=1}^{30} \mathbb{I}(M_i > c; G_i = 0)}{\sum_{i=1}^{30} \mathbb{I}(G_i = 0)}.$$

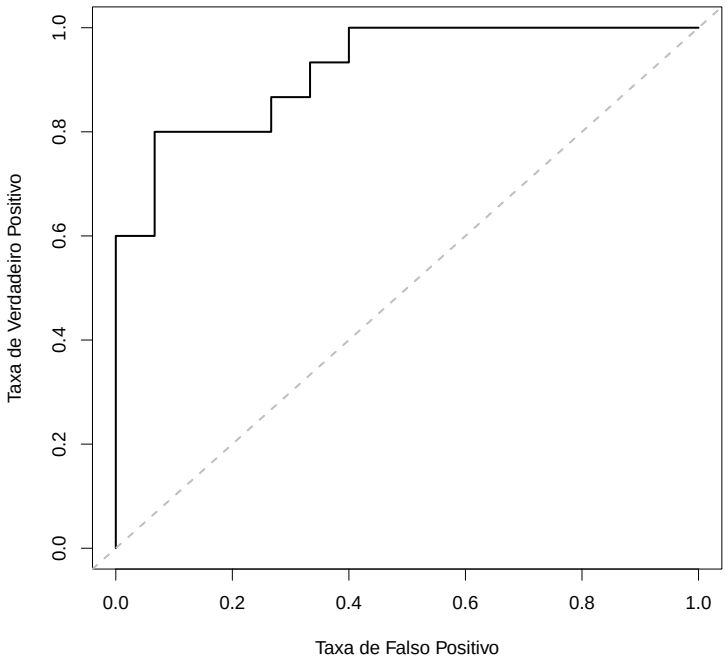
Utilizando esses estimadores para essa amostra obtém-se as estimativas apresentadas na Tabela 3.3:

Tabela 3.3: Estimadores da TFP e TVP para os respectivos valores de c

c	$\hat{TFP}$	$\hat{TVP}$		c	$\hat{TFP}$	$\hat{TVP}$
10,000	1,00	1,00		15,687	0,13	0,80
13,349	0,87	1,00		15,804	0,07	0,80
13,749	0,80	1,00		16,049	0,07	0,73
13,832	0,73	1,00		16,364	0,07	0,67
14,307	0,67	1,00		16,373	0,07	0,60
14,341	0,60	1,00		16,534	0,00	0,60
14,543	0,53	1,00		16,647	0,00	0,53
14,792	0,47	1,00		17,057	0,00	0,47
15,048	0,40	1,00		17,125	0,00	0,40
15,176	0,40	0,93		17,431	0,00	0,33
15,176	0,33	0,93		17,575	0,00	0,27
15,199	0,33	0,87		17,581	0,00	0,20
15,410	0,27	0,87		17,810	0,00	0,13
15,417	0,27	0,80		17,857	0,00	0,07
15,547	0,20	0,80		18,160	0,00	0,00

Assim, plotando e unindo os pontos dados pelas diferentes estimativas de $\hat{TFP}$ e $\hat{TVP}$ para os possíveis valores de c, obtém-se a curva ROC apresentada na Figura 3.1.

Figura 3.1: Curva ROC



Note que o ponto ideal é dado por (0,1). Isso porque, nesse caso, seria observada uma taxa de falso positivo igual a zero e uma taxa de verdadeiro positivo igual a um. Normalmente, o ponto de corte ótimo depende do problema em questão, pois deve se considerar a perda gerada por falsos positivos e falsos negativos. No entanto, o ideal é que a curva ROC esteja o mais próximo possível da ponto (0,1).

3.2 Área sob a curva (ASC) ROC

A área sob a curva ROC é uma medida resumo de performance do marcador dada por:

$$ASC = \int_0^1 ROC(a)da.$$

Essa medida pertence ao intervalo [0,1], pois é dada pela área sob o gráfico de TVP versus TFP. Como essas duas quantidades pertencem ao intevalo $[0, 1]$, então essa área também pertence ao intervalo $[0, 1]$ (Zhou *et al.*, 2002). Uma regra de decisão perfeita que utiliza o marcador M deve apresentar um valor de $ASC = 1$. Isso porque essa curva deve passar pelo ponto (0,1), ou seja, deve ter $TVP = 1$ enquanto a $TFP = 0$. Já um teste em que a TVP é igual a TFP para todos os valores possíveis do critério de corte c deve apresentar um valor de $ASC = 0{,}5$.

É possível mostrar que (Pepe, 2003):

$$\text{ASC} = \int_0^1 ROC(a)da = P(M_i > M_k | W_i = 1, W_k = 0). \tag{3.4}$$

Prova: Considere M_W o marcador dado que o indivíduo pertence ao grupo positivo e $M_{\overline{W}}$ o marcador dado que o indivíduo pertence ao grupo negativo. A área sob a curva ROC pode ser dada por:

$$ASC = \int_0^1 ROC(a)da = \int_0^1 TVP\left(TFP^{-1}(a)\right) da.$$

dos que não apresentaram o evento de interesse Utilizando a mudança de variável de a para $m = TFP^{-1}(a)$ na equação acima obtém-se que:

$$\begin{aligned} ASC &= \int_{-\infty}^{\infty} TVP\,(m)\, dTFP(m) \\ &= \int_{-\infty}^{\infty} P(M > m|W = 1) f(m|W = 0) dm \\ &= \int_{-\infty}^{\infty} P(M_W > m) f_{\overline{W}}(m) dm, \end{aligned}$$

em que $f(\cdot)$ é a função densidade do marcador. Portanto, utilizando a independência estatística entre M_W e $M_{\overline{W}}$, é possível escrever

$$\begin{aligned} ASC &= \int_{-\infty}^{\infty} P\left[M_W > m, M_{\overline{W}} = m\right] dm \\ &= P\left[M_W > M_{\overline{W}}\right] \\ &= P(M_i > M_k|W_i = 1, W_k = 0). \square \end{aligned}$$

Utilizando essa relação é possível interpretar a área sob a curva ROC como a probabilidade da ordenação correta dos marcadores dados os respectivos grupos. Por exemplo, considere um estudo com um grupo de pessoas saudáveis e outro grupo de pessoas doentes em que o marcador é dado por um escore de risco de forma que indivíduos doentes apresentem o escore maior do que indivíduos saudáveis. Assim, de acordo com (3.4), a ASC pode ser interpretada como a probabilidade de um indivíduo doente apresentar o escore maior do que um indivíduo do grupo saudável.

3.3 Curva Preditiva

A curva preditiva é uma ferramenta para avaliar a capacidade preditiva de um marcador (Huang *et al.*, 2007). Considere, novamente, as variáveis W e M definidas anteriormente. Considere também que $E(\cdot)$ é a função de distribuição acumulada do marcador e E^{-1} a sua função inversa. Dessa forma, ao se considerar um valor do marcador igual a m relacionado ao q-ésimo quantil (Casella e Berger, 2001), tem-se que $q = E(m)$. Portanto, $E^{-1}(q)$ será igual a m.

Antes de definir a curva preditiva, é necessário definir uma quantidade denominada risco. O risco associado ao q-ésimo quantil será dado por:

$$R^p(q) = P(W = 1|M = E^{-1}(q)). \tag{3.5}$$

Assim, valores de $R^p(q)$ próximos de um indicam que o q-ésimo quantil está associado ao grupo positivo denotado por $W = 1$ e valores de $R^p(q)$ próximos de zero indicam uma associação do q-ésimo quantil com o grupo negativo denotado por $W = 0$.

A curva preditiva será dada pelo gráfico $R^p(q)$ versus q e descreve a distribuição do risco $P(W = 1|M)$. Dessa forma, será possível visualizar como a probabilidade de ocorrência do evento varia de acordo com os quantis do marcador.

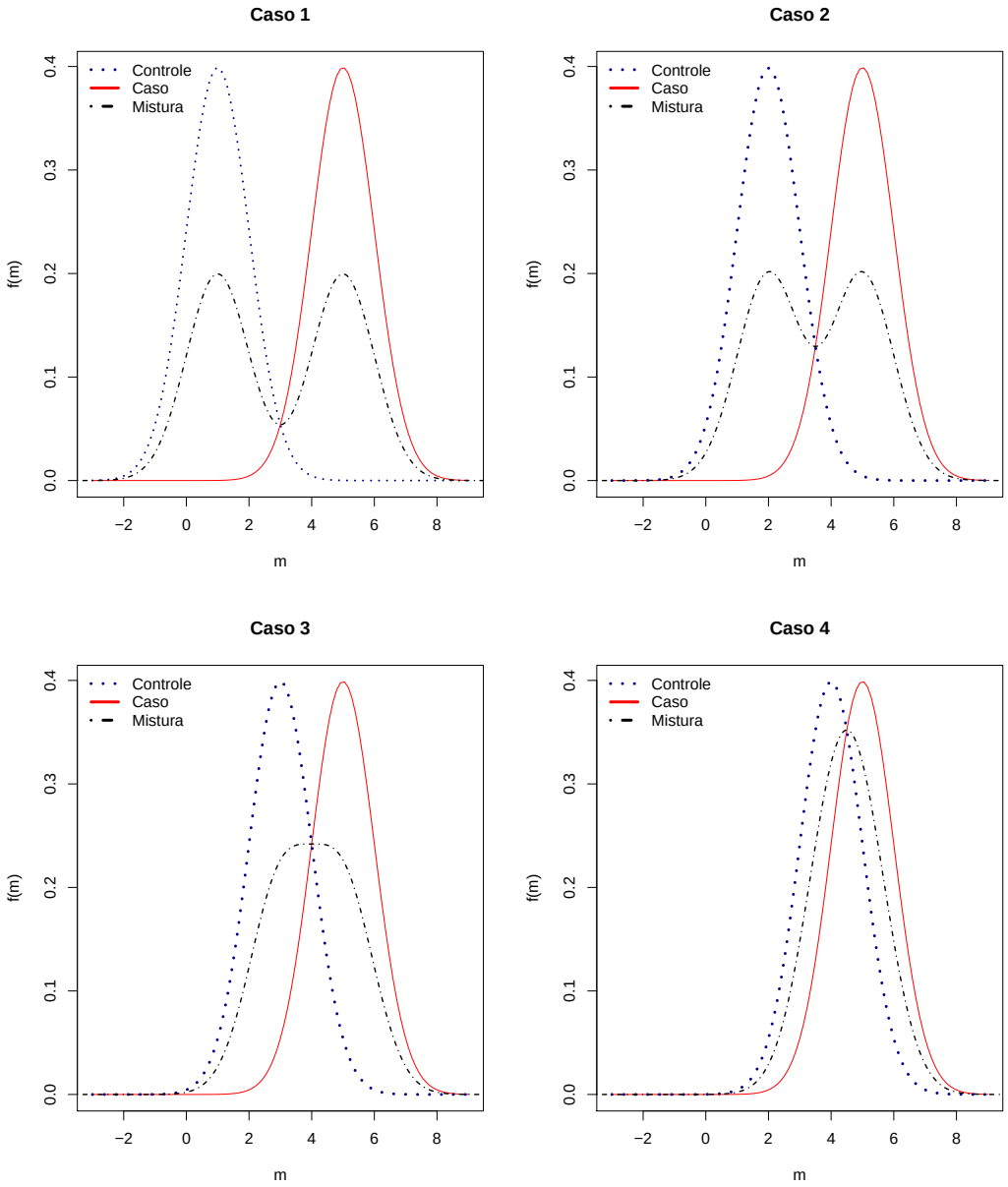
Uma boa característica da curva preditiva é que, ao utilizar a escala $q = E(m)$ no eixo das abscissas, os marcadores são transformados numa escala comum e isso facilita a comparação com outros marcadores.

Para ilustrar a aplicação da curva preditiva, considere a situação do exemplo anterior, em que o interesse é verificar a habilidade preditiva de um marcador contínuo em distinguir entre o grupo caso e o grupo controle. Considere ainda que o marcador segue uma distribuição normal, mas com média diferente entre os grupos e a proporção de unidades de cada grupo é a mesma. A seguir será considerado esse exemplo em quatro cenários diferentes:

- **Caso 1:** o marcador do grupo caso segue uma distribuição normal de média 5, o marcador do grupo controle segue uma distribuição normal de média 1 e a variância das duas distribuições é igual a um;
- **Caso 2:** o marcador do grupo caso segue uma distribuição normal de média 5, o marcador do grupo controle segue uma distribuição normal de média 2 e a variância das duas distribuições é igual a um;
- **Caso 3:** o marcador do grupo caso segue uma distribuição normal de média 5, o marcador do grupo controle segue uma distribuição normal de média 3 e a variância das duas distribuições é igual a um; e
- **Caso 4:** o marcador do grupo caso segue uma distribuição normal de média 5, o marcador do grupo controle segue uma distribuição normal de média 4 e a variância das duas distribuições é igual a um.

A Figura 3.2 apresenta a distribuição individual dos marcadores e as respectivas misturas para os quatro casos. Note que a melhor distinção deve ser dada pelo caso 1, pois é o caso em que as distribuições dos marcadores dos grupos caso e controle estão mais afastadas.

Figura 3.2: Distribuição dos marcadores dos grupos e respectivas misturas



Para construir o gráfico da curva preditiva para os quatro casos utiliza-se o teorema de Bayes na equação (3.5), obtendo-se:

$$R^p(q) = P(W = 1|M = E^{-1}(q)) = \frac{P(M = E^{-1}(q)|W = 1)P(W = 1)}{P(M = E^{-1}(q))}.$$

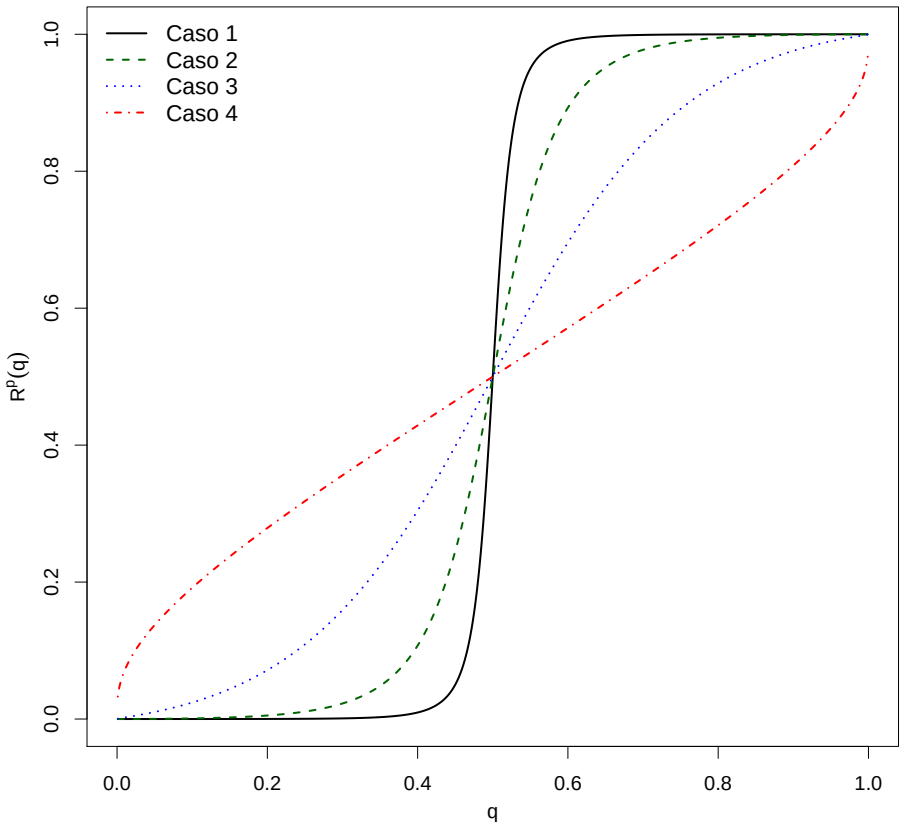
Considere nas quatro situações que a probabilidade de pertencer ao grupo controle é a mesma de se pertencer ao grupo caso, ou seja,

$$P(W = 1) = \tfrac{1}{2} = P(W = 0).$$

Logo, para calcular o risco associado ao q-ésimo quantil, o valor de $P(M = E^{-1}(q)|W = 1)$ será calculado utilizando os respectivos parâmetros da distribuição normal, $P(M = E^{-1}(q))$ será calculado utilizando a mistura de distribuições normais e $P(W = 1)$ será dado por $0,5$. Utilizando

esse procedimento para diferentes valores de q em cada uma das situações obtém-se as curvas preditivas na Figura 3.3.

Figura 3.3: Curva Preditiva



Note que, como era de se esperar, a melhor distinção é dada pelo caso 1, pois é a curva que mais se aproxima da função degrau $\mathbb{I}[(1-\rho) < q] = \mathbb{I}[P(W=0) < q]$, em que ρ é a proporção de casos (Ver Apêndice C). Já o pior marcador, ou seja, o que faz menor distinção entre os dois grupos, é dado pelo caso 4. Perceba que quanto mais próximo da função constante $R^p(q) = P(W = 1|M) = P(W = 1) = \rho$, menos informativo é o marcador. Isso porque, independente dos quantis do marcador, o risco preditivo será sempre o mesmo.

3.4 Extensão dos conceitos para o contexto de análise de sobrevivência

Note que as definições anteriores não foram feitas no contexto de sobrevivência, ou seja, foram feitas em situações nas quais o status denotado por W era fixo. Por isso, valores como sensibilidade, especificidade e área sob a curva também são fixos. No entanto, no contexto de análise de sobrevivência o status pode ser descrito como uma variável dependente do tempo. Por exemplo, a variável indicadora do status pode ser definida por $W(t) = \mathbb{I}(T \leq t)$. Nesse caso, para cada instante

de tempo fixado t, existiriam diferentes medidas de sensibilidade e especificidade, curva ROC e área sob a curva.

Além disso, no contexto de análise de sobrevivência, as observações podem ser censuradas. Aqui, novamente, será considerada censura aleatória à direita e o tempo de censura será denotado por C. Dessa forma, o tempo que se observa é o mínimo entre o tempo de censura e o tempo de falha. Esse tempo será denotado por $X = \min(T, C)$. Para diferenciar o tempo de falha do tempo de censura será utilizada a variável indicadora de falha, definida por $\delta = \mathbb{I}(T \leq C)$.

A seguir serão apresentadas diferentes definições de sensibilidade e especificidade, assim como um coeficiente de concordância apropriado para estudos envolvendo análise de sobrevivência.

3.4.1 Sensibilidade e Especificidade

Existem diferentes formas de se considerar o status dependente do tempo. Essas diferentes formas implicam em diferentes definições de sensibilidade e especificidade que, consequentemente, serão dependentes do tempo. Considerando o grupo que apresentou o evento de interesse, também chamado de grupo caso, utiliza-se a definição de *incidente* e *cumulativo*. Um objeto de estudo é chamado caso incidente se apresenta o evento no instante t e é chamado caso cumulativo se apresenta o evento num instante igual ou inferior a t. Já para o grupo controle são consideradas duas situações: *estático* e *dinâmico*. Os controles estáticos são os objetos que apresentam o evento de interesse num tempo superior a um instante pré fixado t^* e os controles dinâmicos sãos os objetos que apresentam o evento de interesse num instante superior ao instante de avaliação t. A seguir são formuladas algumas dessas definições, conforme discutido em Heagerty e Zheng (2005).

- Incidente/Dinâmico

  As definições de sensibilidade e especificidade, neste caso, são dadas por:

$$sensibilidade^{\mathbb{I}}(c,t) : P(M_i > c | T_i = t) \text{ e}$$
$$especificidade^{\mathbb{D}}(c,t) : P(M_i \leq c | T_i > t).$$

  Nessa abordagem, para cada instante t, divide-se as unidades em risco em dois grupos. O primeiro é composto pela unidades que apresentarem o evento de interesse naquele determinado instante de tempo t, e o segundo, composto pelas unidades que apresentarem um tempo de falha ou censura superior ao instante de avaliação t.

- Cumulativo/Dinâmico

  Nessa abordagem, as definições de sensibilidade e especificidade são dadas por:

$$sensibilidade^{\mathbb{C}}(c,t) : P(M_i > c | T_i \leq t) \text{ e}$$
$$especificidade^{\mathbb{D}}(c,t) : P(M_i \leq c | T_i > t).$$

  Nessa definição o objetivo é avaliar a performance do marcador em distinguir os indivíduos que já apresentaram o evento daqueles que não apresentaram. Note que nessa definição, independente do instante de tempo t, todos os indivíduos são classificados como caso ou controle.

O indivíduo é um caso cumulativo no instante de tempo t se já apresentou o evento de interesse ou um controle dinâmico se não apresentou o evento até esse instante de tempo (Saha, 2009).

- Incidente/Estático

  Utilizando essa abordagem, as definições de sensibilidade e especificidade são dadas por:

$$sensibilidade^{\mathbb{I}}(c,t) : P(M_i > c | T_i = t) \text{ e}$$
$$especificidade^{\overline{\mathbb{D}}}(c,t^*) : P(M_i \leq c | T_i > t^*).$$

  Nessa definição, considera-se como controle os indivíduos que apresentam um tempo de sobrevivência maior que um instante de pré fixado t^*. Assim, para todo instante de tempo t, o grupo controle permanece inalterado, ou seja, estático. Já o grupo caso é composto apenas pelos indivíduos que apresentam o evento de interesse naquele instante de tempo incidente t.

Note que a abordagem *incidente/estático* pode ser tratada como um caso específico da abordagem *incidente/dinâmico*. Apenas muda-se o instante de tempo utilizado para calcular o valor da especificidade$^{\mathbb{D}}$ e obtém-se a especificidade$^{\overline{\mathbb{D}}}$. Por isso, a abordagem *incidente/estático* não será estudada explicitamente nesse trabalho.

Uma possível forma de utilizar as definições de curva ROC e área sob a curva ROC no contexto de sobrevivência é fixar um instante de tempo t e utilizar uma das definições de sensibilidade e especificidade dependentes do tempo. Dessa forma, a curva ROC num determinado instante de tempo t será chamada de curva ROC dependente do tempo e denotada por ROC_t. Da mesma forma, a medida da área sob a curva será uma função do tempo denotada por $ASC(t)$.

A extensão do conceito de curva preditiva será feita da mesma forma, ou seja, fixando um instante de tempo t. A função da curva preditiva no instante de tempo t será denotada por R_t^p. Por exemplo, fixando um instante de tempo t_0, a curva preditiva dependente do tempo será dada por:

$$R_{t_0}^p(q) = P(W(t_0) = 1 | M = E^{-1}(q)).$$

3.4.2 Coeficiente de concordância C

Em muitas situações se tem interesse em numa medida resumo do desempenho discriminatório, independente do tempo, de um marcador. Por exemplo, ao trabalhar com a $ASC(t)$ verifica-se o desempenho apenas para o instante t fixado. Para se ter uma medida que não depende de t, se define o coeficiente de concordância. Dessa forma, é considerado como uma medida de resumo global, pois não está definido apenas para um instante de tempo t ou um determinado intervalo de tempo.

Inicialmente, o coeficiente de concordância C foi definido pela proporção de pares concordantes considerando o tempo de sobrevivência T e a probabilidade de sobrevivência (Harrel Jr *et al.*, 1996). No entanto, pode-se utilizar, no lugar da probabilidade de sobrevivência, um marcador M definido, por exemplo, como um preditor linear de um modelo de sobrevivência (Heagerty e Zheng, 2005). Normalmente, considera-se que altos valores dos marcadores estejam associados a longos períodos de sobrevivência. No entanto, aqui será considerado que altos valores de M estejam associados a menores períodos de sobrevivência para ficar de acordo com a associação definida na curva ROC.

Como motivação para definir o coeficiente de concordância, considere a situação, apresentada por Pencina e D'Agostino (2004), em que um conjunto de n observações do tempo de sobrevivência T e do marcador contínuo M serão observados até um determinado instante de tempo pré fixado e a única censura possível é a censura à direita devido ao término do estudo. Portanto, se duas observações chegarem ao final do estudo sem apresentar o evento de interesse, ambas apresentarão o mesmo valor para a variável T. Note que nessa situação $P(T_i = T_j) \neq 0$. Essas unidades que são censuradas devido ao término do estudo serão denominadas unidades inutilizáveis. Aqui serão considerados todos os diferentes pares (i, j), tal que $i < j$, i varia de 1 a $n-1$ e j varia de 2 a n.

Os pares concordantes serão pares tais que $m_i < m_j$ e $t_i > t_j$ ou, também, $m_i > m_j$ e $t_i < t_j$. Os casos em que $m_i < m_j$ e $t_i < t_j$ ou $m_i > m_j$ e $t_i > t_j$ serão chamados pares discordantes.

Para definir o coeficiente de concordância serão utilizadas as quantidades definidas por Pencina e D'Agostino (2004). No entanto, serão feitas as devidas adaptações para considerar a associação negativa entre o marcador e o tempo de sobrevivência. Assim, serão utilizadas as seguintes quantidades:

- π_c: é a probabilidade de pares concordantes. Essa probabilidade é dada por $P(M_i < M_j; T_i > T_j)$ ou $P(M_i > M_j; T_i < T_j)$. Note que, por simetria, essas quantidades são exatamente iguais. Portanto, desse ponto em diante será utilizada apenas a primeira quantidade.

- π_d: é a probabilidade de pares discordantes. Essa probabilidade é dada por $P(M_i > M_j; T_i > T_j) + P(M_i < M_j; T_i < T_j)$. Novamente, por simetria, essas quantidades são iguais. Portanto, por conveniência utilizaremos apenas a última quantidade.

- π_i: é a probabilidade de pares inutilizáveis, ou seja, pares que chegaram ao final do estudo sem apresentar o evento de interesse. Dessa forma, essa quantidade é definida como $P(T_i = T_j)$. Note que esse valor é dado por $1 - (\pi_c + \pi_d)$.

Logo, uma forma de definir o coeficiente C é dada por:

$$\begin{aligned} C &= P\left\{(T_i < T_j; M_i > M_j) | T_i \neq T_j\right\} \\ &= \frac{\pi_c}{1 - \pi_i} \\ &= \frac{\pi_c}{1 - \{1 - (\pi_c + \pi_d)\}} \\ &= \frac{\pi_c}{\pi_c + \pi_d} \\ &= \frac{P(M_i < M_j; T_i > T_j)}{P(M_i < M_j; T_i > T_j) + P(M_i > M_j; T_i > T_j)}. \end{aligned}$$

Supondo que o marcador M segue uma distribuição contínua, a última igualdade é equivalente a

$$C = \frac{P(M_i < M_j; T_i > T_j)}{P(T_i > T_j)} = P(M_i < M_j | T_i > T_j). \tag{3.6}$$

Note que esse coeficiente pertence ao intervalo [0,1] pois é dado por uma probabilidade. Se o coeficiente for maior que 0,5, então os maiores valores do marcador são preditores de tempos de

sobrevivência menores. Se for menor que 0,5, então os maiores valores do marcador são preditores de tempos maiores de sobrevivência. Se for igual a 0,5, então o marcador não é útil para diferenciar o tempo de sobrevivência.

A motivação apresentada é uma das possíveis motivações do coeficiente de concordância. Heagerty e Zheng (2005) mostraram outra motivação que é dada pela relação entre o coeficiente de concordância e a área sob a curva ROC ao longo do tempo sob a abordagem *incidente/dinâmico*. Mais detalhes e diferentes formas de se estimar esse coeficiente podem ser encontradas em Schmid e Potapov (2012).

A seguir serão apresentados alguns métodos de estimação de quantidades definidas anteriormente como curva ROC, área sob a curva ROC, coeficiente de concordância e curva preditiva, considerando diferentes definições de sensibilidade e especificidade.

3.5 Abordagem incidente/dinâmico

Nessa seção será utilizada a abordagem *incidente/dinâmico* para estimar a curva ROC, a área sob a curva e o coeficiente de concordância segundo a abordagem proposta por Heagerty e Zheng (2005). Utilizando essa abordagem, é possível definir a curva ROC dependente do tempo como a função $ROC_t^{\mathbb{I}/\mathbb{D}}(p)$, em que p é a taxa de falso positivo dinâmica ($TFP^{\mathbb{D}}$) relacionada a um valor de corte c^p, ou seja,

$$\begin{aligned} TFP^{\mathbb{D}}(c^p, t) &= P(M > c^p | T > t) \\ &= 1 - P(M \leq c^p | T > t) \\ &= 1 - especificidade^{\mathbb{D}}(c, t) \\ &= p. \end{aligned}$$

Além disso, $ROC_t^{\mathbb{I}/\mathbb{D}}(p)$, dada por (3.3), será a taxa de verdadeiro positivo incidente ($TVP^{\mathbb{I}}$). Note que $c^p = \left[TFP^{\mathbb{D}}\right]^{-1}(p, t)$. Para cada valor de c^p obtém-se a $TVP^{\mathbb{I}}(c^p, t)$. Assim, o valor de $ROC_t^{\mathbb{I}/\mathbb{D}}(p)$ será dado por:

$$\begin{aligned} ROC_t^{\mathbb{I}/\mathbb{D}}(p) &= TVP^{\mathbb{I}}\left\{c^p, t\right\} \\ &= TVP^{\mathbb{I}}\left\{\left[TFP^{\mathbb{D}}\right]^{-1}(p, t)\right\}, \end{aligned}$$

em que $\left[TFP^{\mathbb{D}}\right]^{-1}(p, t) = \inf_c \left\{c : TFP(c, t) \leq p\right\}$.

Nessa abordagem é possível demonstrar que o coeficiente de concordância está relacionado a ASC_t da seguinte forma:

$$C = P(M_j > M_k | T_j < T_k) = \int ASC(t)u(t)dt, \tag{3.7}$$

em que $u(t) = 2f(t)S(t)$, $f(t)$ é a função densidade do tempo de falha e $S(t)$ é a função de sobrevivência do tempo de falha (Ver Apêndice C). No entanto, na maioria dos casos, o interesse está num intervalo finito de tempo do tipo $(0, \tau)$. Nesse caso, é possível fazer uma ponderação em

u para que a soma desses pesos no intervalo $(0, \tau)$ seja igual a um. Essa versão modificada é dada por:

$$C^{\tau} = \int_0^{\tau} ASC(t)u^{\tau}(t)dt, \tag{3.8}$$

em que $u^{\tau}(t) = \frac{2f(t)S(t)}{U^{\tau}}$ e $U^{\tau} = \int_0^{\tau} 2f(t)S(t)dt = 1 - S^2(\tau)$ (Ver Apêndice C).

3.5.1 Estimação

Primeiro será apresentada a estratégia de estimação da $TVP^{\mathbb{I}}$. Note que essa quantidade pode ser dada por:

$$TVP^{\mathbb{I}}(c,t) = P(M > c|T = t) = \int_c^{\infty} f(m|t)dm. \tag{3.9}$$

Portanto, é de interesse se obter a distribuição do marcador M condicional ao instante de falha T. A seguir será apresentada a forma de estimar essa distribuição, sob o modelo de riscos proporcionais, proposta por Xu e O'Quigley (2000).

Utilizando algumas propriedades da probabilidade condicional é possível escrever:

$$\begin{aligned} f(m|T = t) &= \frac{f(m,t)}{f(t)} \\ &= \frac{f(t|m)f_M(m)}{\int f(t|m)f_M(m)dm}, \end{aligned}$$

em que $f_M(\cdot)$ é a função densidade do marcador.

Lembre que, sob o modelo de Cox,

$$\lambda(t|m) = \tfrac{f(t|m)}{S(t|m)} \Rightarrow f(t|m) = \lambda(t|m)S(t|m) = \lambda_0(t)\exp\{\gamma m\}S(t|m).$$

Então, pode-se escrever

$$\begin{aligned} f(m|T = t) &= \frac{f(t|m)f_M(m)}{\int f(t|m)f_M(m)dm} \\ &= \frac{\lambda_0(t)\exp\{\gamma m\}S(t|m)f_M(m)}{\int \lambda_0(t)\exp\{\gamma m\}S(t|m)f_M(m)dm} \\ &= \frac{\exp\{\gamma m\}P(M = m, T \geq t)}{\int \exp\{\gamma m\}P(M = m, T \geq t)dm} \\ &= \frac{\exp\{\gamma m\}P(M = m, T \geq t)}{\int \exp\{\gamma m\}P(M = m, T \geq t)dm}\frac{P(T \geq t)}{P(T \geq t)} \\ &= \frac{\exp\{\gamma m\}h(m|T \geq t)}{\int \exp\{\gamma m\}h(m|T \geq t)dm}, \end{aligned}$$

em que $h(m|T \geq t) = P(M = m|T \geq t)$.

Lembre que, conforme (3.9), o interesse está em estimar

$$P(M > c|T = t) = \int_c^{\infty} f(m|T = t)dm = \int_c^{\infty} \frac{\exp\{\gamma m\}h(m|T \geq t)}{\int_{-\infty}^{\infty}\exp\{\gamma m\}h(m|T \geq t)dm}dm. \tag{3.10}$$

Aqui, será utilizado o estimador empírico de $h(m|T \geq t)$ que é dado por:

$$\widehat{h(m|T \geq t)} = \widehat{P(M = m|T \geq t)} = \sum_{j=1}^{n}\mathbb{I}(M_j = m)\frac{R_j(t)}{\sum_{i=1}^{n}R_i(t)},$$

em que $R_i(t) = \mathbb{I}(T_i \geq t)$.

Da mesma forma será utilizada uma aproximação discreta para as integrais do numerador e denominador de (3.10). Assim, utilizando essa aproximação e $\widehat{h(m|T \geq t)}$ no numerador, obtém-se:

$$\int_c^{\infty} \widehat{\exp\{\gamma m\}\, h(m|T \geq t)}dm = \sum_{j=1}^{n} \exp\{\hat{\gamma} m_j\} \frac{\mathbb{I}(M_j \geq c) R_j(t)}{\sum_{i=1}^{n} R_i(t)}.$$

Analogamente, para o denominador, obtém-se o seguinte estimador:

$$\int_{-\infty}^{\infty} \widehat{\exp\{\gamma m\}\, h(m|T \geq t)}dm = \sum_{j=1}^{n} \exp\{\hat{\gamma} m_j\} \frac{R_j(t)}{\sum_{i=1}^{n} R_i(t)}.$$

Fazendo a razão dos estimadores do numerador e denominador de (3.10), obtém-se:

$$\begin{aligned}\frac{\int_c^{\infty} \widehat{\exp\{\gamma m\}\, h(m|T \geq t)}dm}{\int_{-\infty}^{\infty} \exp\{\gamma m\}\, h(m|T \geq t)dm} &= \frac{\sum_{j=1}^{n} \exp\{\hat{\gamma} m_j\} \frac{\mathbb{I}(M_j \geq c) R_j(t)}{\sum_{i=1}^{n} R_i(t)}}{\sum_{j=1}^{n} \exp\{\hat{\gamma} m_j\} \frac{R_j(t)}{\sum_{i=1}^{n} R_i(t)}} \\ &= \frac{\sum_{j=1}^{n} \mathbb{I}(M_j \geq c) R_j(t) \exp\{\hat{\gamma} m_j\}}{\sum_{j=1}^{n} R_j(t) \exp\{\hat{\gamma} m_j\}}.\end{aligned}$$

Logo, o estimador de $TVP^{\mathbb{I}}(c,t)$ é dado por:

$$\widehat{TVP}^{\mathbb{I}}(c,t) = P(\widehat{M > c|T = t}) = \sum_{j=1}^{n} \mathbb{I}(M_j > c) \frac{R_j(t) \exp\{\hat{\gamma} m_j\}}{\sum_{i=1}^{n} R_i(t) \exp\{\hat{\gamma} m_i\}} = \sum_{i=1}^{n} \mathbb{I}(M_j > c) \kappa_j(\hat{\gamma}, t). \tag{3.11}$$

Note que a quantidade $\kappa_j(\hat{\gamma}, t)$ é a mesma que aparece na função escore utilizada no modelo de Cox considerando o marcador M como covariável (Cox, 1972). Lembre que essa função é dada por:

$$U(\gamma) = \sum_i \delta_i \left[m_i - \sum_k m_k \left(\frac{R_k(t) \exp\{\hat{\gamma} m_k\}}{\sum_{j=1}^{n} R_j(t) \exp\{\hat{\gamma} m_j\}} \right) \right] = \sum_i \delta_i \left[m_i - \sum_k m_k \kappa_k(\boldsymbol{\gamma}, t) \right].$$

Para $TFP^{\mathbb{D}}(c,t)$, o seguinte estimador empírico pode ser utilizado:

$$\widehat{TFP}^{\mathbb{D}}(c,t) = P(\widehat{M > c|T > t}) = \sum_j \mathbb{I}(M_j > c) \frac{R_j(t^+)}{V^R(t^+)}, \tag{3.12}$$

em que $R_j(t^+) = \lim_{\delta \to 0} R_j(t + |\delta|)$ é a variável binária que indica se o indivíduo apresenta falha ou censura num instante posterior a t e $V^R(t^+) = \sum_j R_j(t^+)$ é a quantidade desses indivíduos. Logo, esse estimador é apenas a proporção de indivíduos em risco com o escore M_i maior que c no instante seguinte a t.

Dadas as estimativas $\widehat{TVP}^{\mathbb{I}}$ e $\widehat{TFP}^{\mathbb{D}}$, é possível estimar a curva ROC_t e consequentemente a área sob a curva (ASC), para cada instante de tempo. Um estimador da $ASC(t)$ é dado por $\widehat{ASC}(t) = \int \widehat{ROC}_t^{\mathbb{I}/\mathbb{D}}(p)dp$ que pode ser calculado utilizando-se um método de integração numérica.

Com ASC estimada, também é possível estimar o coeficiente de concordância C^τ. Utilizando a relação (3.8) é possível estimar o coeficiente de concordância da seguinte forma:

$$\hat{C}^\tau = \int^\tau \widehat{ASC}(t)\hat{u}^\tau(t)dt, \tag{3.13}$$

em que $\hat{u}^\tau(t) = 2\hat{f}(t)\hat{S}(t)/\left[1-\hat{S}^2(\tau)\right]$. Uma alternativa para a estimativa de $S(\cdot)$ é o estimador de Kaplan-Meier e para $f(\cdot)$ é possível utilizar uma aproximação discreta dada pelos incrementos do estimador de Kaplan-Meier. Assim, se o estimador de Kaplan-Meier for utilizado, então $\widehat{ASC}(t)$ só precisa ser avaliado nos tempos de falha observados para se calcular $\hat{C}^\tau$, pois nos outros instantes de tempo o valor de $\hat{u}(\cdot)$ será igual a zero.

3.6 Abordagem cumulativo/dinâmico

Nessa seção será utilizada a abordagem *cumulativo/dinâmico* com foco na área sob a curva ROC dependente do tempo proposta por Viallon e Latouche (2011). Será demonstrado que a área sob a curva ROC_t pode ser escrita em função da curva preditiva e o processo de estimação da área sob a curva ROC_t será feito utilizando-se essa relação.

Novamente, considere $E(\cdot)$ e $e(\cdot)$, respectivamente, como as funções de distribuição acumulada e densidade do marcador M. Lembre-se também que, nesse abordagem, a taxa de verdadeiro positivo ($TVP^{\mathbb{C}}$) e de falso positivo ($TFP^{\mathbb{D}}$) são dadas, respectivamente, por:

$$TVP^{\mathbb{C}}(c,t) = P(M > c|T \leq t) \text{ e}$$
$$TFP^{\mathbb{D}}(c,t) = P(M > c|T > t).$$

Então, uma forma de escrever a função da área sob a curva dependente do tempo é dada por:

$$ASC(t) = \int_0^1 TVP^{\mathbb{C}}\left\{\left\{TFP^{\mathbb{D}}\right\}^{-1}(q,t)\right\}dq. \tag{3.14}$$

Utilizando a mudança de variável de $\left\{TFP^{\mathbb{D}}\right\}^{-1}(q,t)$ para m em (3.14) é possível mostrar que (Ver Apêndice C):

$$ASC^{\mathbb{C}/\mathbb{D}}(t) = \int_{-\infty}^{\infty} P(M > m|T \leq t)P(M = m|T > t)dm.$$

Dessa forma, utilizando o teorema de Bayes com as equações acima de $TVP^{\mathbb{C}}$, $TFP^{\mathbb{D}}$ e $ASC^{\mathbb{C}/\mathbb{D}}$, obtém-se que (Ver ApêndiceC):

$$ASC^{\mathbb{C}/\mathbb{D}}(t) = \int_{-\infty}^{\infty}\int_c^{\infty} \frac{F(t|M=m)\left[1-F(t|M=c)\right]}{\left[1-F(t)\right]F(t)}e(m)e(c)dmdc. \tag{3.15}$$

Utilizando a equação (3.15), é possível obter a seguinte relação da ASC com a curva preditiva (Ver ApêndiceC):

$$ASC^{\mathbb{C}/\mathbb{D}}(t) = \frac{1}{F(t)\left[1-F(t)\right]}\left[\int_0^1 qR_t^p(q)dq - \frac{F^2(t)}{2}\right]. \tag{3.16}$$

3.6.1 Estimação

O processo de estimação, proposto por Viallon e Latouche (2011), será baseado na relação (3.16). Um dos termos dessa relação é dado por:

$$\int_0^1 qR_t^p(q)dq. \tag{3.17}$$

Para estimar esse termo, lembre que $R_t^p(q)$ é dado por $P(W(t)=1|M=E^{-1}(q))$, mas como $W(t)=\mathbb{I}(T\leq t)$, então $P(W(t)=1|M=E^{-1}(q))$ é igual a $P(T\leq t|M=E^{-1}(q))$ que será denotado por $F(t|M=E^{-1}(q))$. Assim, pode-se escrever

$$\int_0^1 qR_t^p(q)dq = \int_0^1 qP(W(t)=1|M=E^{-1}(q))dq = \int_0^1 qP(T\leq t|M=E^{-1}(q))dq.$$

Utilizando a mudança de variável $q=E(m)$, obtém-se:

$$\frac{dq}{dm} = \frac{dE(m)}{dm} = e(m).$$

Assim, chega-se a:

$$\begin{aligned}\int_0^1 qR_t^p(q)dq &= \int_0^1 qP(T\leq t|M=E^{-1}(q))dq \\ &= \int_{-\infty}^{\infty} E(m)P(T\leq t|M=E^{-1}\left[E(m)\right])e(m)dm \\ &= \int_{-\infty}^{\infty} E(m)P(T\leq t|M=m)e(m)dm \\ &= \int_{-\infty}^{\infty} E(m)F(t|m)e(m)dm. \end{aligned} \tag{3.18}$$

Assim, utilizando a relação (3.18), o estimador empírico de $\int_0^1 qR_t^p(q)dq$ é dado por:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{i}{n}\hat{F}_n(t|M_{(i)}),$$

em que $M_{(i)}$ denota a i-ésima estatística de ordem relacionada a amostra $\{M_1,...,M_n\}$ e $\hat{F}_n(t|M_{(i)})$ é um estimador da função de distribuição acumulada que pode ser dado, por exemplo, por um modelo de regressão. Para a distribuição marginal do tempo de sobrevivência em (3.16) serão consideradas duas opções. Na primeira, utiliza-se o estimador de Kaplan-Meier para obter $\hat{F}_1(t)$ (Kaplan e Meier, 1958). Na segunda, considera-se a relação:

$$F(t) = \int F(t;m)dm = \int F(t|m)e(m)dm.$$

Assim, uma alternativa é utilizar o seguinte estimador da distribuição marginal do tempo de sobrevivência:

$$\hat{F}_2(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{F}(t|M_i). \tag{3.19}$$

Portanto, é possível utilizar dois estimadores, com $k = 1$ ou $k = 2$, para $ASC_k^{\mathbb{C}/\mathbb{D}}(t)$ (Viallon e Latouche, 2011):

$$ASC_k^{\mathbb{C}/\mathbb{D}}(t) = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{i}{n}\hat{F}(t|M_{(i)}) - \hat{F}_k^2(t)/2}{\hat{F}_k(t)\left[1 - \hat{F}_k(t)\right]}. \tag{3.20}$$

3.7 Abordagem IPCW

Uma alternativa para estimar a sensibilidade e especificidade sob a abordagem cumulativo dinâmico foi proposta por Uno *et al.* (2007). Nesse abordagem utiliza-se, necessariamente, como marcador a função de distribuição acumulada dada por:

$$P(T \leq t_0|\mathbf{Z}) = g(\boldsymbol{\beta}\mathbf{Z}), \tag{3.21}$$

em que $\mathbf{Z}$ é um vetor de covariáveis com o primeiro elemento igual a um, $g(\cdot)$ é uma função conhecida, estritamente crescente, diferenciável e $\boldsymbol{\beta}$ é um vetor p-dimensional de parâmetros desconhecidos. O estimador de $\boldsymbol{\beta}$ é dado pela solução da seguinte função de estimação:

$$U(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\omega_i}{\hat{S}^c(\min(X_i, t_0))}\mathbf{Z}_i \left\{\mathbb{I}(X_i \leq t_0) - g(\boldsymbol{\beta}\mathbf{Z})\right\}, \tag{3.22}$$

em que $X_i = \min(T_i, C_i)$, $\omega_i = \mathbb{I}(\min(T_i, t_0) \leq C_i) = \mathbb{I}(X_i \leq t_0)\delta_i + \mathbb{I}(X_i > t_0)$, $\delta_i = \mathbb{I}(X_i = T_i)$ e $\hat{S}^c(\cdot)$ é o estimador de Kaplan-Meier para a censura. Após obter as estimativas do vetor de parâmetros $\boldsymbol{\beta}$, os autores propuseram os seguintes estimadores da sensibilidade e especificidade:

$$\widehat{SE}(c, t_0) : \frac{\sum_{i=1}^{n} \delta_i \mathbb{I}(g(\hat{\boldsymbol{\beta}}\mathbf{Z}_i) > c, X_i \geq t)/\hat{S}^c(X_i)}{\sum_{i=1}^{n} \delta_i \mathbb{I}(X_i \geq t)/\hat{S}^c(X_i)} \text{ e} \tag{3.23}$$

$$\widehat{ES}(c, t_0) : \frac{\sum_{i=1}^{n} \delta_i \mathbb{I}(g(\hat{\boldsymbol{\beta}}\mathbf{Z}_i) \leq c, X_i > t)}{\sum_{i=1}^{n} \delta_i \mathbb{I}(X_i > t)}. \tag{3.24}$$

Além dos estimadores, os autores obtiveram a distribuição assintótica e propuseram a utilização de um método de reamostragem com perturbação para a aproximação da distribuição. Eles mostraram que, para n grande, a distribuição de $\sqrt{n}\left\{\widehat{SE}(c) - SE(c)\right\}$ pode ser aproximada por $\sqrt{n}\left\{SE^*(c) - \widehat{SE}(c)\right\}$, sendo que SE^* é obtido substituindo-se $\boldsymbol{\beta}$ e $S^c(\cdot)$ em (3.23) por, respectivamente, $\boldsymbol{\beta}^*$ e $S^*(\cdot)$. Esses valores são obtidos da seguinte forma:

$$S^*(t) = \hat{S}^c(t) - \hat{S}^c(t)\sum_{i=1}^{n} V_i \int_0^t \left\{\sum_{j=1}^{n} \mathbb{I}(X_j > s)\right\}^{-1} d\hat{M}_i(s), \tag{3.25}$$

em que V_i é uma observação de uma variável aleatória positiva de distribuição conhecida com média e variância iguais a um, $\hat{M}_i(t) = \mathbb{I}(X_i \leq t, \delta_i = 0) - \int_0^t \mathbb{I}(X_i > s)d\hat{\Lambda}(s)$ e $\hat{\Lambda}(\cdot)$ é o estimador padrão

da função de risco acumulado de Nelson-Aalen para a variável de censura C. Ainda, $\boldsymbol{\beta}^*$ é obtido resolvendo-se a seguinte equação em relação a $\boldsymbol{\beta}$:

$$U^*(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\omega_i}{\hat{S}^*(\min(X_i, t_0))}\mathbf{Z}_i \left\{\mathbb{I}(X_i \leq t_0) - g(\boldsymbol{\beta}\mathbf{Z})\right\} V_i. \tag{3.26}$$

De forma análoga, se obtém a distribuição aproximada de $\sqrt{n}\left\{\widehat{ES}(c) - ES(c)\right\}$. No entanto, com essa abordagem não é possível utilizar diretamente um marcador biológico para o cálculo da sensibilidade e especificidade. Isso porque a equação de estimação dada em (3.22), que é a base dessa abordagem, impõe uma estrutura de modelagem dada pela função $g(\boldsymbol{\beta}\mathbf{Z})$. Por exemplo, se houvesse interesse em diferenciar pacientes diretamente de acordo com o nível de colesterol no sangue, não seria possível realizar essa análise sem incluir a estimação da estrutura dada por $g(\boldsymbol{\beta}\mathbf{Z})$. Dessa forma, será proposta aqui uma modificação ou adaptação do método de Uno *et al.* (2007), que consiste na utilização de equações de estimação generalizadas para se estimar a sensibilidade e especificidade numa abordagem em que, além de ser possível utilizar marcadores que sejam funções de um conjunto de covariáveis e outro modelo estatístico, também seja possível utilizar diretamente um marcador biológico.

Lembre que, sob a abordagem cumulativo/dinâmico, a sensibilidade e especificidade são dadas, respectivamente, por:

$$sens^{\mathbb{C}}(m, t_0): \quad P(M > m | T \leq t_0) = \frac{P(M > m; T \leq t_0)}{P(T \leq t_0)} \text{ e} \tag{3.27}$$

$$esp^{\mathbb{D}}(m, t_0): \quad P(M \leq m | T > t_0) = \frac{P(M \leq m; T > t_0)}{P(T > t_0)}. \tag{3.28}$$

Dessa forma, serão propostos estimadores de $P(M > m; T \leq t_0)$, $P(M \leq m; T > t_0)$ e $P(T > t_0)$ para estimar a sensibilidade e especificidade.

3.7.1 Estimação

Antes de apesentar os estimadores deve ser feita uma pequena distinção entre duas possíveis situações. A primeira é o caso em que o marcador é observado, ou seja, seu valor é conhecido para cada observação, como, por exemplo, a altura ou o peso de um indivíduo ou uma função fixada de determinadas quantidade e a segunda, em que o marcador é obtido, por exemplo, com a utilização de um modelo estatístico de forma que não se obtém verdadeiro marcador supostamente existente, mas um valor predito. Essa situação pode ser dada pelo caso em que se utiliza o resultado de um preditor linear de um modelo de regressão como marcador. Note que nesse caso não se conhece o marcador para cada unidade amostral, então trabalha-se com uma previsão dessa quantidade dada por $X\hat{\boldsymbol{\beta}}$. A seguir serão apresentados estimadores quando se considera o marcador fixado. Aqui será utilizada a mesma notação da seção 2.3.1. Lembre que a motivação do estimador IPCW proposto por Lawless e Yuan (2010) é dada pela seguinte igualdade:

$$E_{T,M,C}\left[\frac{\Delta}{\alpha} L(\cdot,\cdot)\right] = E_{T,M}\left[L(\cdot,\cdot)\right], \tag{3.29}$$

em que Δ é dado por $\mathbb{I}\{L(\cdot,\cdot) \text{ é conhecido}\}$ e α é dado por $P(\Delta = 1 | T, M, C)$. Considere agora as funções $L_1 = \mathbb{I}(T > t_0)$, $L_2 = \mathbb{I}(M > m; T \leq t_0)$ e $L_3 = \mathbb{I}(M \leq m; T > t_0)$ e as respectivas variáveis:

- Δ_1: $\mathbb{I}\{\mathbb{I}(T > t_0)$ é conhecido$\}$,
- Δ_2: $\mathbb{I}\{\mathbb{I}(M > m; T \leq t_0)$ é conhecido$\}$ e
- Δ_3: $\mathbb{I}\{\mathbb{I}(M \leq m; T > t_0)$ é conhecido$\}$.

Note que o conjunto em que essas funções indicadoras são conhecidas é o mesmo para as três variáveis e corresponde ao seguinte conjunto $\{(C \geq T) \cup (C < T; C > t_0)\}$.

Defina

$$\alpha = P\left[(C \geq t) \cup (C < t; C > t_0)\right] = \begin{cases} S^c(t) & \text{se} \quad t < t_0 \\ S^c(t_0) & \text{se} \quad t > t_0 \end{cases} = S^c(\min(t, t_0)).$$

Para L_2 é possível escrever

$$\begin{aligned} E_{T,M,C}\left[\frac{\Delta_2 L_2}{\alpha}\right] =& E_{T,M,C}\left[\frac{\Delta_2 \mathbb{I}(M > m; T \leq t_0)}{\alpha}\right] \\ =& E_{T,M}\left\{E_{C|T,M}\left[\frac{\mathbb{I}(M > m; T \leq t_0)\Delta_2}{\alpha}\right]\right\} \\ =& E_{T,M}\left\{\frac{\mathbb{I}(M > m; T \leq t_0)}{\alpha} E_{C|T,M}\left[\Delta_2\right]\right\}. \end{aligned}$$

Observe que, como Δ_2 é uma variável binária, $E_{C|T,M}\left[\Delta_2\right] = P(\Delta_2 = 1|T, M)$. Lembre que o conjunto em que Δ_2 é igual a um corresponde ao conjunto $\{(C \geq T) \cup (C < T; C > t_0)\}$. Assim,

$$E_{C|T,M}\left[\Delta_2\right] = P(\Delta_2 = 1|T, M) = P\left\{(C \geq t) \cup (C < t; C > t_0)\right\} = \alpha.$$

Portanto,

$$\begin{aligned} E_{T,M,C}\left[\frac{\Delta_2 \mathbb{I}(M > m; T \leq t_0)}{\alpha}\right] =& E_{T,M}\left\{\frac{\mathbb{I}(M > m; T \leq t_0)}{\alpha} E_{C|T,M}\left[\Delta_2\right]\right\} \\ =& E_{T,M}\left\{\frac{\mathbb{I}(M > m; T \leq t_0)}{\alpha}\alpha\right\} \\ =& E_{T,M}\left\{\mathbb{I}(M > m; T \leq t_0)\right\} \\ =& P(M > m; T \leq t_0). \end{aligned} \tag{3.30}$$

De forma análoga, é possível mostrar que

$$E_{T,M,C}\left[\frac{\Delta_1 L_1}{\alpha}\right] = E_{T,M,C}\left[\frac{\Delta_1 \mathbb{I}(T > t_0)}{\alpha}\right] = P(T > t_0) \text{ e} \tag{3.31}$$

$$E_{T,M,C}\left[\frac{\Delta_3 L_3}{\alpha}\right] = E_{T,M,C}\left[\frac{\Delta_3 \mathbb{I}(M \leq m; T > t_0)}{\alpha}\right] = P(M \leq m; T > t_0). \tag{3.32}$$

Note que, a partir das relações (3.30), (3.32) e (3.31) mostra-se que

$$E\left[\frac{\Delta_1 \mathbb{I}(M > m; T \leq t_0)}{\alpha} - P(M > m; T \leq t_0)\right] = 0,$$

$$E\left[\frac{\Delta_2 \mathbb{I}(M \leq m; T > t_0)}{\alpha} - P(M \leq m; T > t_0)\right] = 0 \text{ e}$$

$$E\left[\frac{\Delta_3 \mathbb{I}(T > t_0)}{\alpha} - P(T > t_0)\right] = 0.$$

Tendo como motivação essa três igualdades propõe-se a utilização da abordagem de equações de estimação generalizadas (EEG) proposta por Liang e Zeger (1986). Essa abordagem de EEG é uma generalização multivariada da abordagem de quase-verossimilhança proposta por Wedderburn (1974). A abordagem de quase-verossimilhança não requer a definição de uma distribuição para a variável resposta, mas apenas que a especificação do modelo para a média esteja correto para gerar estimadores consistentes e assintoticamente normais. Outra boa propriedade dos estimadores de quase-verossimilhança é que, mesmo quando a variância da variável resposta for especificada incorretamente, ainda assim, os estimadores são consistentes.

Na generalização proposta por Liang e Zeger (1986), considera-se um vetor de parâmetros $\boldsymbol{\theta}$ de dimensão $(p \times 1)$, um vetor respostas $\mathbf{Y}_i$ e um vetor de médias $\boldsymbol{\mu}_i$, sendo que esses últimos são de dimensão $(n_i \times 1)$ e o vetor $\boldsymbol{\mu}$ é função de $\boldsymbol{\theta}$. Definidas essas quantidades, utiliza-se a seguinte equação de estimação:

$$\sum_{i=1}^{n} D_i^T K_i^{-1} \left(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})\right) = \mathbf{0}, \tag{3.33}$$

em que $D_i = \partial\boldsymbol{\mu}/\partial\boldsymbol{\theta}$ é uma matriz $(n_i \times p)$ e K_i é uma matriz de pesos de dimensão $(n_i \times n_i)$. Considere, ainda, que a matriz de pesos K_i pode ser decomposta da seguinte forma:

$$K_i = A_i^{\frac{1}{2}} R_i(\boldsymbol{\rho}) A_i^{\frac{1}{2}},$$

em que A_i é uma matriz diagonal de dimensão $(n_i \times n_i)$ com elementos na j-ésima diagonal dados por $\text{Var}(Y_{ij})$ e $R_i(\boldsymbol{\rho})$ é a matriz de correlações de $\mathbf{Y}_i$ que pode depender de um vetor adicional dado por $\boldsymbol{\rho}$. Note que, quando se considera os dados independentes, a matriz R_i é dada pela matriz identidade. Liang e Zeger (1986) chamaram $R_i(\boldsymbol{\rho})$ de matriz de correlação de trabalho, pois isso reflete o fato de que $R_i(\boldsymbol{\rho})$ não é necessariamente a verdadeira matriz de correlações de $\mathbf{Y}$.

Novamente, assumindo que o modelo de regressão das médias de $\mathbf{Y}$ foi especificado corretamente, $\hat{\boldsymbol{\theta}}$ é consistente para $\boldsymbol{\theta}$ e, também, $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ segue uma distribuição assintótica normal multivariada de média zero e matriz de covariâncias dada por C_θ, ou seja,

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N(\mathbf{0}, C_\theta), \tag{3.34}$$

em que

$$C_\theta = \lim_{N\to\infty} N \left(\sum_{i=1}^{N} D_i^T K_i^{-1} D_i\right)^{-1} \left\{\sum_{i=1}^{N} D_i^T K_i^{-1} \text{Cov}(\mathbf{Y}_i) K_i^{-1} D_i\right\} \left(\sum_{i=1}^{N} D_i^T K_i^{-1} D_i\right)^{-1}. \tag{3.35}$$

Note, que se a $\text{Cov}(\mathbf{Y}_i)$ for especificada corretamente, de forma que $K_i = \text{Cov}(\mathbf{Y_i})$, então (3.35) se reduz a:

$$\lim_{N\to\infty} N \left(\sum_{i=1}^{N} D_i^T K_i^{-1} D_i\right)^{-1}. \tag{3.36}$$

Por fim, C_θ em (3.35) pode ser consistentemente estimado substituindo-se $\boldsymbol{\theta}$ e $\boldsymbol{\rho}$ pelos seus respectivos estimadores e, também, $\text{Cov}(\mathbf{Y}_i)$ por $(\mathbf{Y}_i - \hat{\boldsymbol{\mu}})(\mathbf{Y}_i - \hat{\boldsymbol{\mu}})^T$ (Lipsitz e Fitzmaurice, 2008). Portanto, será proposta a utilização dos resultados de EEG para se obter estimativas da sensibilidade

e especificidade, assim como suas respectivas variâncias. Para facilitar o texto serão introduzidas as seguintes notações:

$$\begin{aligned} P(T > t_0) &= p_1, \\ P(M > m; T \leq t_0) &= p_2 \text{ e} \\ P(M \leq m; T > t_0) &= p_3. \end{aligned}$$

Note que as quantidades definidas acima, de acordo com as equações (3.23) e (3.24), são suficientes para estimar a sensibilidade e especificidade. Ainda, denote a sensibilidade e especificidade, respectivamente, por S e E. Dessa forma, mostra-se que:

$$S = P(M > m | T \leq t_0) = \frac{p_2}{1 - p_1} \Rightarrow p_2 = S(1 - p_1) \text{ e} \tag{3.37}$$

$$E = P(M \leq m | T > t_0) = \frac{p_3}{p_1} \Rightarrow p_3 = Ep_1. \tag{3.38}$$

Note que o interesse, para t_0 e m fixados, está em estimar S e E. Dessa forma, utilizando a notação introduzida no contexto de EEG, considere $\boldsymbol{\mu} = (p_1, S(1-p_1), Ep_1)$, $\boldsymbol{\theta} = (p_1, S, E)$ e $\mathbf{Y_i}$ que será o vetor de observações para cada unidade amostral de forma que

$$\mathbf{Y_i} = \left(\frac{\Delta_{1i}}{\alpha_i} \mathbb{I}(T_i > t_0), \frac{\Delta_{2i}}{\alpha_i} \mathbb{I}(M_i > m; T_i \leq t_0), \frac{\Delta_{3i}}{\alpha_i} \mathbb{I}(M_i \leq m; T_i > t_0) \right).$$

Portanto, as quantidade utilizadas na equação de estimação (3.33) serão dadas por:

$$D^T = \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} \right)^T = \begin{pmatrix} \frac{\partial \mu_1}{\partial \theta_1} & \frac{\partial \mu_2}{\partial \theta_1} & \frac{\partial \mu_3}{\partial \theta_1} \\ \frac{\partial \mu_1}{\partial \theta_2} & \frac{\partial \mu_2}{\partial \theta_2} & \frac{\partial \mu_3}{\partial \theta_2} \\ \frac{\partial \mu_1}{\partial \theta_3} & \frac{\partial \mu_2}{\partial \theta_3} & \frac{\partial \mu_3}{\partial \theta_3} \end{pmatrix} = \begin{pmatrix} \frac{\partial p_1}{\partial p_1} & \frac{\partial S(1-p_1)}{\partial p_1} & \frac{\partial E p_1}{\partial p_1} \\ \frac{\partial p_1}{\partial S} & \frac{\partial S(1-p_1)}{\partial S} & \frac{\partial E p_1}{\partial S} \\ \frac{\partial p_1}{\partial E} & \frac{\partial S(1-p_1)}{\partial E} & \frac{\partial E p_1}{\partial E} \end{pmatrix} = \begin{pmatrix} 1 & -S & E \\ 0 & 1-p_1 & 0 \\ 0 & 0 & p_1 \end{pmatrix},$$

$$(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\theta})) = \begin{pmatrix} \frac{\Delta_1}{\alpha} \mathbb{I}(T > t_0) - p_1 \\ \frac{\Delta_2}{\alpha} \mathbb{I}(M > m; T \leq t_0) - p_2 \\ \frac{\Delta_3}{\alpha} \mathbb{I}(M \leq m; T > t_0) - p_3 \end{pmatrix} = \begin{pmatrix} \frac{\Delta_1}{\alpha} \mathbb{I}(T > t_0) - p_1 \\ \frac{\Delta_2}{\alpha} \mathbb{I}(M > m; T \leq t_0) - S(1-p_1) \\ \frac{\Delta_3}{\alpha} \mathbb{I}(M \leq m; T > t_0) - Ep_1 \end{pmatrix} \text{ e}$$

$$K_i = \begin{pmatrix} \text{Var}(Y_{1i}) & \rho_1 & \rho_2 \\ \rho_1 & \text{Var}(Y_{2i}) & \rho_3 \\ \rho_2 & \rho_3 & \text{Var}(Y_{3i}) \end{pmatrix}.$$

Note que, no casos em questão, a matriz D_i será a mesma para todas as unidades amostrais. Além disso, aqui será considerado que a matriz K_i é conhecida e será dada pela matriz de covariâncias de $\mathbf{Y}$. Dessa forma, é possível substituir essas quantidades, respectivamente, por K e D e reescrever (3.33) como:

$$D^T K^{-1} \sum_{i=1}^{n} (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})) = \mathbf{0}. \tag{3.39}$$

Observe que, neste caso, $\left(D^T\right)^{-1}$ existe apenas se $det(D^T) = p_1(1-p_1)$ for diferente de zero, ou seja, se $p_1 \in (0,1)$. Se essa matriz existir, ela será dada por:

$$\left(D^T\right)^{-1} = \frac{1}{p_1(1-p_1)} \begin{pmatrix} p_1(1-p_1) & Sp_1 & -Ep_1 \\ 0 & p_1 & 0 \\ 0 & 0 & 1-p_1 \end{pmatrix}.$$

Assim, resolvendo (3.39) para $\boldsymbol{\theta}$, obtém-se:

$$\begin{aligned} \left(D^T\right)^{-1} D^T K^{-1} \sum_{i=1}^{n} (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})) &= \left(D^T\right)^{-1} \mathbf{0} \\ K^{-1} \sum_{i=1}^{n} (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})) &= \mathbf{0} \\ K K^{-1} \sum_{i=1}^{n} (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})) &= K\mathbf{0} \\ \sum_{i=1}^{n} (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})) &= \mathbf{0} \\ \boldsymbol{\mu}_i(\boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i \end{aligned}$$

Note que os estimadores não dependem da estrutura de correlação definida em K. Prosseguindo a resolução de (3.39), obtém-se os seguintes estimadores:

$$\hat{p_1} = \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_{1i}}{\alpha_i} \mathbb{I}(T_i > t_0), \tag{3.40}$$

$$\hat{S} = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_{2i}}{\alpha_i} \mathbb{I}(M_i > m; T_i \leq t_0)}{1-p_1} = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_{2i}}{\alpha_i} \mathbb{I}(M_i > m; T_i \leq t_0)}{1 - \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_{1i}}{\alpha_i} \mathbb{I}(T_i > t_0)} \text{ e} \tag{3.41}$$

$$\hat{E} = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_{3i}}{\alpha_i} \mathbb{I}(M_i \leq m; T_i > t_0)}{p_1} = \frac{\sum_{i=1}^{n} \frac{\Delta_{3i}}{\alpha_i} \mathbb{I}(M_i \leq m; T_i > t_0)}{\sum_{i=1}^{n} \frac{\Delta_{1i}}{\alpha_i} \mathbb{I}(T_i > t_0)}, \tag{3.42}$$

em que α é dado pela função de sobrevivência da censura no instante dado pelo mínimo entre o tempo de falha t e o instante fixado t_0, ou seja, $S^c(\min(t_i, t_0))$.

As estimativas das variâncias de $\hat{p}_1, \hat{S}$ e $\hat{E}$ serão dadas, respectivamente, pelos elementos da diagonal da matriz (3.35) substituindo-se as matrizes D, K e $\text{Cov}(\mathbf{Y}_i)$ pelas respectivas matrizes estimadas.

Para o caso em que M é estimado por meio de modelos de regressão, a proposta é utilizar os mesmos estimadores (3.40), (3.41) e (3.42). No entanto, as propriedades dos estimadores para essa situação não são bem conhecidas e, com intuito de avaliar empiricamente os estimadores propostos, foi feito um estudo de simulação apresentado no próximo capítulo.

Capítulo 4

Simulações

No intuito de conhecer melhor o comportamento das medidas de avaliação do desempenho preditivo do tempo e status de sobrevivência, utilizou-se dois grandes cenários de simulação. No primeiro, se estudou o comportamento dos três estimadores do erro de predição apresentados no segundo capítulo em diferentes situações. No segundo, considerou-se as adaptações dos estimadores de Uno *et al.* (2007) da sensibilidade e especificidade apresentados na seção 3.7 assim como os respectivos estimadores dos desvios padrão sob diferentes cenários. Considerou-se, também, situações em que o marcador foi estimado e em que foi fixado. Além disso, comparou-se o comportamento das estimativas da área sob a curva ROC considerando os estimadores adaptados da sensibilidade e especificidades, o estimador empírico na ausência de censura e o estimador empírico excluindo as observações censuradas.

4.1 Tempo de sobrevivência

Nesse contexto, considerou-se uma situação bem simples em que se tem interesse em verificar a performance preditiva de diferentes modelos para prever o tempo de sobrevivência representado pela variável T. Nessa simulação considerou-se como variável explicativa apenas uma variável indicadora. Essa situação pode representar, por exemplo, um estudo em que se observa o tempo até a presença de um novo episódio de depressão entre pacientes que participaram de uma terapia e pacientes que não participaram dessa terapia. A variável indicadora dos grupos será denotada por Z. Assim,quando Z for igual a zero o indivíduo não participa da terapia e quando Z for igual a 1 o indivíduo participa da terapia.

Considerando essa situação com uma variável indicadora, serão utilizados ainda dois cenários. No primeiro, será considerado que o tempo de recuperação segue uma distribuição exponencial que depende do valor de Z e no segundo, o tempo de recuperação segue uma distribuição lognormal que também depende do valor de Z. Para cada situação serão ajustados os seguintes modelos: exponencial, Weibull, lognormal e de riscos proporcionais de Cox.

Além disso, será utilizada a função de perda absoluta dada por $L(Y, \hat{Y}) = |Y - \hat{Y}(Z)|$, sendo que o preditor $\hat{Y}(Z)$ será dado pela mediana do tempo de sobrevivência condicional à covariável Z.

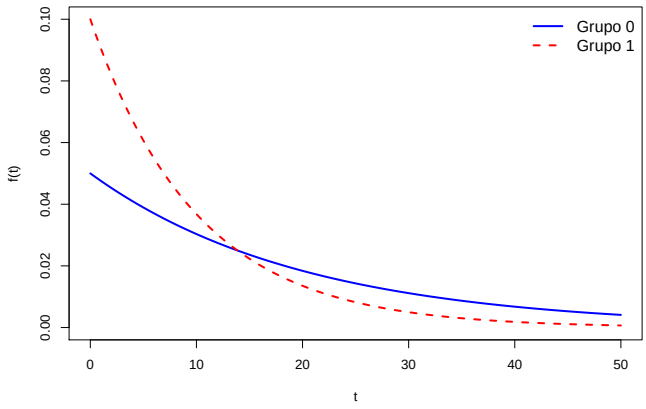
4.1.1 Exponencial

Nesse seção serão apresentados os resultados considerando que a distribuição do tempo de sobrevivência segue a distribuição exponencial com densidade dada por:

$$f(t) = \lambda_z \exp\{-\lambda_z t\}. \tag{4.1}$$

Conforme já discutido, a variável indicadora Z será utilizada para identificar os grupos. Para cada grupo considerou-se as densidades geradas com $\lambda_0 = 0,05$ e $\lambda_1 = 0,1$ em (4.1) (Figura 4.1).

Figura 4.1: Funções densidades das distribuições geradas na simulação



Novamente, no intuito de verificar o comportamento dos estimadores em diferentes cenários, utilizou-se diferentes proporções de censura e tamanhos amostrais. O tempo de censura C foi gerado utilizando-se a distribuição exponencial. Para conseguir fixar a proporção de censura esperada, suponha que $T \sim \exp(\lambda_z)$, $C \sim \exp(\alpha)$ e que esses tempos sejam independentes. Assim, utilizou-se a seguinte relação:

$$\begin{aligned} P(T > C|Z = z) &= \int_0^\infty \int_c^\infty \lambda_z \alpha \exp(-\lambda_z t) \exp(-\alpha c) dt dc \\ &= \frac{\alpha}{\alpha + \lambda_z} \int_0^\infty (\alpha + \lambda_z) \exp(-(\alpha + \lambda_z)c) dc \\ &= \frac{\alpha}{\alpha + \lambda_z}. \end{aligned}$$

No entanto, o interesse está em fixar p tal que $P(T > C)$ seja igual a p. Assim,

$$\begin{aligned} P(T > C) &= \sum_z P(T > C, Z = z) \\ &= \sum_z P(T > C|Z = z) P(Z = z) \\ &= \sum_z \frac{\alpha}{\alpha + \lambda_z} P(Z = z). \end{aligned}$$

Logo, fixando-se uma proporção p de censura, mostra-se que:

$$\begin{aligned} p =& P(T > C) \\ p =& \frac{\alpha}{\alpha + \lambda_0} P(Z = 0) + \frac{\alpha}{\alpha + \lambda_1} P(Z = 1) b \\ 0 =& \alpha^2 \{(P(Z = 0) + P(Z = 1)) - p\} \\ & + \alpha \{\lambda_1 P(Z = 0) + \lambda_0 P(Z = 1) - p(\lambda_0 + \lambda_1)\} - p\lambda_0\lambda_1. \end{aligned} \tag{4.2}$$

Portanto, fixando-se λ_0, λ_1, $P(Z = 0)$ e $P(Z = 1)$, é possível obter α tal que $p = P(T > C)$, resolvendo a equação (4.2). Utilizando-se essa relação, quando necessário, os dados foram gerados com 0%, 20% e 40% de censura. Além disso, foram gerados dois tamanhos de amostra: 100 e 250 observações por grupo.

Para estimar a perda esperada foram utilizados os estimadores baseado em modelo dado em (2.24), perda aparente dado em (2.25) e validação cruzada dado em (2.26). A distribuição do tempo de censura foi estimada utilizando-se o estimador de Kaplan-Meier. Ainda, os parâmetros necessários para o obtenção das estimativas foram dados por $K = 50$ para o estimador baseado em modelo e para o estimador validação cruzada os dados foram divididos em $V = 10$ conjuntos.

Nessa simulação foram geradas 1.000 amostras para cada situação. Para calcular a variância dos estimadores perda aparente e validação cruzada em cada uma dessas amostras foi utilizada uma técnica de reamostragem com reposição considerando $B = 150$. Para o estimador baseado em modelo foi utilizada uma técnica de reamostragem baseada no modelo ajustado considerando, também, $B = 150$.

Lembre que para a distribuição exponencial, de acordo com a equação (2.10), é possível obter o verdadeiro valor da perda analiticamente. No entanto, para as outras distribuições não é possível obter uma expressão para esse valor. Por isso, foram utilizados valores de referência para cada distribuição calculados da seguinte forma:

- Gerou-se um conjunto de dados de treinamento de 20.000 observações sendo metade de cada grupo. Note que os dados foram gerados sem censura e com os mesmos parâmetros da simulação.
- Aplicou-se o modelo especificado ao conjunto de dados de treinamento para se obter os preditores. Novamente, nessa simulação foram utilizadas as medianas dos grupos como preditores.
- Gerou-se um conjunto de dados de avaliação de 20.000 observações, independente dos dados de treinamento, com os mesmos parâmetros.
- Calculou-se o valor de referência que é dado pela média dos desvios absolutos de cada tempo dos dados de avaliação em relação aos valores preditos utilizando-se a mediana estimada pelo modelo ajustado com os dados de treinamento.

Após a conclusão da simulação obtiveram-se os resultados apresentados na Tabela 4.1. Nessa tabela $Var(\hat{\pi})$ é a variância das estimativas dos erros de predição da simulação e $\widehat{\text{Var}(\hat{\pi})}$ é a média

das variâncias dos erros de predição obtidos com a utilização de técnicas de amostragem. Observou-se nos resultados da simulação que o estimador perda aparente, independentemente da taxa de censura e modelo ajustado, subestimou o erro de predição (Tabela 4.1). Notou-se um viés com o seguinte padrão: quanto maior a proporção de censura, maior o viés no sentido de subestimar o erro de predição (Tabela 4.1, Figura A.1 e Figura A.7). Quando passou-se de 100 observações em cada grupo para 250 observações, houve uma tendência de diminuição do viés e, como esperado, a variância também diminuiu (Tabela 4.1, Figura A.2 e Figura A.8). Ainda, observou-se que há uma maior redução proporcional no viés para as maiores taxas de censura quando se passa de 100 para 250 observações por grupo (Tabela 4.1).

O estimador validação cruzada superestimou o erro de predição apenas para a situação sem censura e quando se utilizou o modelo exponencial (Tabela 4.1). Para as outras situações, o estimador validação cruzada subestimou o erro de predição, apresentando valores intermediários entre os erros de predição obtidos com o estimador perda aparente e os valores de referência (Tabela 4.1). Ainda, observou-se que há uma tendência do estimador validação cruzada semelhante à tendência observada com o estimador perda aparente, ou seja, quanto maior a proporção de censura observada, maior foi o viés no sentido de subestimar o erro de predição (Tabela 4.1, Figura A.3 e Figura A.9). No entanto, o viés observou-se uma diminuição no viés de acordo com o aumento do tamanho amostral. Aqui, também, como era de se esperar há uma redução na variância do estimador com o aumento da amostra (Tabela 4.1, Figura A.4 e Figura A.10).

Para o estimador baseado em modelo, quando utilizou-se os modelos exponencial e Weibull, observou-se valores bem próximos do verdadeiro erro de predição independentemente da proporção de censura (Tabela 4.1, Figura A.5 e Figura A.11). Além disso, observou-se que as variâncias são bem estáveis entre todas as proporções de censura, ou seja, não mudam tanto quanto os outros estimadores de acordo com a censura (Figura A.6). Considerando o modelo de Cox, o erro de predição estimado apresentou uma tendência crescente de subestimar o erro de predição conforme se aumentou a proporção de censura. Quando se utilizou um modelo incorreto, nesse caso o modelo lognormal, o erro de predição foi superestimado independente da proporção de censura (Tabela 4.1, Figura A.5 e Figura A.11). Ainda para esse modelo, observou-se um viés no sentido de que quanto maior a proporção de censura, maior o valor esperado do erro de predição. Além disso, considerando os quatro modelos ajustados, observou-se um aumento na variância de acordo com o aumento na proporção de censura (Tabela 4.1, Figura A.6 e Figura A.12). Ao passar de 100 observações por grupo para 250, observou-se uma pequena diminuição no viés do erro de predição, mas notou-se uma grande diminuição na variância do estimador (Tabela 4.1).

Baseado nos resultados da simulação, quando não houver graves desvios do modelo escolhido, indica-se a utilização do estimador baseado em modelo. Isso porque foram as estimativas que geralmente ficaram mais próximas do verdadeiro erro de predição, mesmo ao variar-se a proporção de censura. Além disso, geralmente, apresentaram a menor variância entre os estimadores propostos (Tabela 4.1).

Tabela 4.1: Resultados da simulação exponencial

			100 em cada grupo			250 em cada grupo		
Modelo	Censura	Estimador	$\hat{\pi}$	Var($\hat{\pi}$)	$\widehat{\text{Var}}(\hat{\pi})$	$\hat{\pi}$	Var($\hat{\pi}$)	$\widehat{\text{Var}}(\hat{\pi})$
Exponencial (10,3972)	0%	π^{PA}	10,3419	0,7927	0,8197	10,3526	0,3019	0,3269
		π^{VC}	10,4181	0,8075	0,8250	10,3786	0,3027	0,3226
		π^{M}	10,4247	0,6055	0,6253	10,3851	0,2518	0,2473
	20%	π^{PA}	9,9550	1,3566	1,3078	10,1513	0,5446	0,5532
		π^{VC}	10,0510	1,3880	1,3353	10,1911	0,5504	0,5595
		π^{M}	10,4552	0,8301	0,6374	10,4219	0,3181	0,2503
	40%	π^{PA}	8,8135	2,7053	2,0736	9,2920	1,5218	1,1846
		π^{VC}	8,9348	2,7926	2,1525	9,3402	1,5320	1,1960
		π^{M}	10,5449	1,2068	0,6449	10,4521	0,4756	0,2513
Weibull (10,4434)*	0%	π^{PA}	10,3343	0,7931	0,8124	10,3500	0,3019	0,3275
		π^{VC}	10,4247	0,8073	0,8327	10,3818	0,3035	0,3248
		π^{M}	10,3674	0,7663	0,7288	10,3662	0,2963	0,2957
	20%	π^{PA}	9,9504	1,3557	1,3111	10,1499	0,5442	0,5520
		π^{VC}	10,0558	1,3826	1,3369	10,1927	0,5466	0,5660
		π^{M}	10,3695	1,1254	0,7445	10,3912	0,4400	0,2940
	40%	π^{PA}	8,8123	2,7062	2,1058	9,2912	1,5210	1,1745
		π^{VC}	8,9370	2,7868	2,1248	9,3411	1,5379	1,1933
		π^{M}	10,4248	1,9908	0,7628	10,4567	0,7972	0,3041
Cox (10,5707)*	0%	π^{PA}	10,3192	0,7942	1,3118	10,3484	0,3015	0,6724
		π^{VC}	10,4449	0,8196	1,2132	10,3973	0,3063	0,6061
		π^{M}	10,2375	0,7776	-	10,3170	0,3056	-
	20%	π^{PA}	9,9328	1,3560	1,8607	10,1424	0,5441	0,9265
		π^{VC}	10,0762	1,3916	1,7710	10,1992	0,5498	0,8625
		π^{M}	10,0027	1,0585	-	10,2129	0,4285	-
	40%	π^{PA}	8,7844	2,7241	2,7759	9,2801	1,5215	1,6208
		π^{VC}	8,9581	2,8105	2,7321	9,3530	1,5292	1,5565
		π^{M}	9,3446	1,3724	-	9,7904	0,7753	-
Lognormal (10,6664)*	0%	π^{PA}	10,4610	0,7929	0,8194	10,4855	0,3050	0,3294
		π^{VC}	10,5484	0,8066	0,9362	10,5158	0,3067	0,3280
		π^{M}	15,2802	5,2390	5,5568	15,2745	1,7084	2,1095
	20%	π^{PA}	10,0282	1,3652	1,3312	10,2307	0,5475	0,5545
		π^{VC}	10,1283	1,3911	1,3433	10,2691	0,5501	0,5617
		π^{M}	18,3322	12,8853	9,5144	18,1365	5,3860	3,4653
	40%	π^{PA}	8,8268	2,7706	2,1377	9,3075	1,5522	1,2214
		π^{VC}	8,9455	2,8208	2,2085	9,3548	1,5618	1,2265
		π^{M}	23,3438	53,4946	21,720	23,1904	17,7044	6,9813

(*) Valores de referência.

Estudo Validação Cruzada

Para verificar a influência do número de grupos V no resultado final do estimador validação cruzada foi realizado um pequeno estudo em que se utilizou um cenário muito próximo ao utilizado na simulação descrita anteriormente. Nesse estudo trabalhou-se com dados gerados da mesma forma da simulação, no entanto foram analisadas somente as situações com 100 observações por grupo e as mesmas três proporções de censura. Nessa análise ajustou-se apenas o modelo exponencial e foram utilizados $V = 10$ e $V = 200$ grupos. Lembre que, quando o número de grupos V é igual ao número de observações, o estimador validação cruzada também é chamado de *leave-one-out* (LOO).

Observou-se que, para essa situação, o número de grupos parece não ter muita influência no valor estimado do erro de predição (Tabela 4.2). As estimativas do erro de predição, assim como as estimativas da variância do erro de predição, ficaram muito próximas.

Tabela 4.2: Resultados da simulação para o estimador validação cruzada

Número de grupos	Censura	$\hat{\pi}$	$\text{Var}(\hat{\pi})$	$\widehat{\text{Var}(\hat{\pi})}$
10 Grupos	0%	10,4037	0,7900	0,8250
	20%	10,0003	1,3902	1,3065
	40%	8,8689	2,6100	1,9729
LOO	0%	10,4009	0,7859	0,8239
	20%	9,9953	1,3933	1,3039
	40%	8,8618	2,5931	1,9688

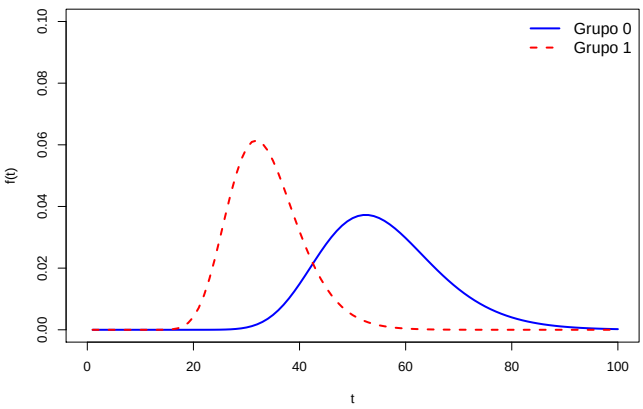
4.1.2 Lognormal

Nesse contexto gerou-se dados representando o tempo de sobrevivência utilizando-se a distribuição lognormal (Casella e Berger, 2001). Essa distribuição é dada pela seguinte função de densidade:

$$f(t) = \frac{1}{\sigma t\sqrt{2\pi}} \exp\left\{\frac{-(\log(t) - \mu_z)^2}{2\sigma^2}\right\}. \tag{4.3}$$

Prosseguindo na comparação de dois grupos, utilizou-se os seguintes parâmetros: $\mu_0 = 4$, $\mu_1 = 3,5$ e $\sigma = 0,2$ (Figura 4.2).

Figura 4.2: Funções densidades das distribuições geradas na simulação



Os tempos de censura foram gerados segundo uma distribuição uniforme de mínimo igual a 0 e máximo igual a 200. Para controlar a proporção de censura, utilizou-se o seguinte procedimento:

1. Gerava-se um tempo de censura c_i. Se não houvesse nenhum tempo de falha maior do que c_i, gerava-se outro instante de censura até se obter pelo menos um tempo de falha maior.

2. Entre os tempos de falha maiores do que c_i, sorteava-se aleatoriamente uma observação a qual seria atribuída esse instante de censura.

3. Repetia-se o procedimento até obter a porcentagem de censura desejada.

Aqui, serão considerados os mesmos cenários utilizados na simulação da distribuição exponencial, ou seja, serão consideradas as porcentagens de censuras de 0%, 20% e 40%, assim como tamanhos amostrais de 100 e 250 observações por grupo. Os valores de referência foram obtidos de forma análoga a da seção anterior. Note que, aqui, os modelos incorretos são os modelos: exponencial, Weibull e modelo de riscos proporcionais de Cox.

Após a conclusão da simulação obtiveram-se os resultados apresentados na Tabela 4.3. Observou-se nos resultados da simulação que, em todos os cenários, a menor estimativa do erro de predição foi dada quando utilizou-se o estimador perda aparente. Considerando esse estimador e a situação com 100 observações por grupo, com exceção da situação em que utilizou-se o modelo exponencial, esse estimador tende a aumentar o viés, no sentido de subestimar o erro de predição, a medida que se aumentou a proporção de censura (Tabela 4.3 e Figura A.13). Na presença de 20% de censura, considerando os dois números de observações por grupo, os modelos lognormal e de riscos proporcionais de Cox apresentaram uma distribuição do erro de predição muito semelhante (Figura A.13 e Figura A.19). Na presença de 40% de censura, as distribuições do erro de predição ao se utilizar esses dois modelos não apresentaram um comportamento tão semelhante como na situação com 20% de censura, sendo que a distribuição do erro de predição, ao se utilizar o modelo de riscos proporcionais de Cox, ficou mais deslocada para a esquerda no sentido de subestimar o erro de predição (Tabela 4.3, Figura A.13 e Figura A.19).

As variâncias do estimador perda aparente, considerando a situação sem censura, apresentaram valores muito semelhantes quando utilizou-se os modelos Weibull e lognormal (Tabela 4.3, Figura A.14 e Figura A.20). No cenário com 20% de censura, os modelos Weibull e lognormal, independentemente do número de observações por grupo, apresentaram um comportamento muito semelhante (Figura A.14 e Figura A.20). No entanto, quando considerou-se 40% de censura, observou-se que a distribuição da variância dos erros não foram tão semelhantes, sendo que a variância do erro de predição ao utilizar o modelo Weibull apresentou maiores valores. Já quando utilizou-se o modelo exponencial, a variância desse estimador na ausência de censura foi a menor entre todos os modelos, mas há um aumento considerável nessa variância conforme se aumenta a proporção de censura (Figura A.14 e Figura A.20).

O estimador validação cruzada, com exceção do modelo lognormal na ausência de censura, sempre apresentou valores intermediários entre os três estimadores, sendo que os valores menores eram obtidos com o estimador perda aparente e valores mais altos para o estimador baseado em modelo

(Tabela 4.3). As distribuições do erro de predição ao utilizar os modelos de riscos proporcionais de Cox e lognormal ficaram muito próximas, independentemente do número de observações por grupo e proporção de censura (Figura A.15 e Figura A.21). Embora essas distribuições ainda estejam próximas no cenário com 40% de censura, a distribuição do modelo de riscos proporcionais de Cox tende a se deslocar para esquerda no sentido de subestimar mais o erro de predição (Figura A.15 e Figura A.21). Considerando 20% e 40% de censura, observou-se que a distribuição do erro de predição ao utilizar o modelo Weibull ficou mais próxima do seu valor de referência do que as distribuições dos erros ao se utilizar o modelo de riscos proporcionais de Cox e lognormal (Figura A.15 e Figura A.21).

O comportamento da variância do estimador validação cruzada foi muito semelhante ao do estimador perda aparente. Na ausência de censura e com exceção do modelo de riscos proporcionais de Cox, todos os modelos apresentaram uma variância muito próxima sendo que a menor variância observada foi quando se utilizou o modelo exponencial (Tabela 4.3). Aqui, também, considerando 20% e 40% de censura, a variância do erro de predição ao considerar os modelos Weibull e lognormal apresentaram um comportamento muito semelhante independentemente do número de observações por grupo (Figura A.16 e Figura A.22). A maior discrepância entre a variância dos erros de predição, considerando esses dois modelos, ocorreu no cenário com 250 observações por grupo e 40% de censura. Nessa situação o erro de predição ao considerar o modelo lognormal apresentou uma variabilidade menor (Tabela 4.3 e Figura A.22)

O estimador baseado em modelo, na ausência de censura e considerando 100 observações por grupo, apresentou valores próximos dos valores de referência para todos os modelos com exceção do modelo exponencial (Tabela 4.3). O erro de predição mais próximo do valor de referência, como esperado, foi o erro de predição ao se utilizar o modelo lognormal, pois é o modelo especificado corretamente (Figura A.17). Ao se considerar os modelos de riscos proporcionais de Cox e Weibull, embora tenham ficado relativamente próximos dos valores de referência, observou-se estimativas superestimadas do erro de predição com o modelo Weibull apresentando os maiores valores (Tabela 4.3 e Figura A.17). Conforme aumentou-se a proporção de censura, ainda considerando 100 observações por grupo, observou-se que a distribuição do erro de predição ao utilizar o modelo lognormal tende a se deslocar no sentido de superestimar o erro de predição. Já ao se utilizar os modelos de riscos proporcionais de Cox e Weibull, o viés no sentido de superestimar o erro de predição tende a aumentar em menor proporção que ao se utilizar o modelo lognormal (Figura A.17). Note que, para todas as proporções de censura, o erro de predição ao se utilizar o modelo lognormal foi o que apresentou o menor viés entre todos os modelos utilizados (Tabela 4.3 e Figura A.17). Observou-se, também, que as distribuições das variâncias do estimador do erro de predição, ao considerar os modelos Weibull e lognormal, ficaram quase idênticas na ausência de censura (Figura A.18). Além disso, notou-se que a variância do estimador do erro de predição ao se utilizar o modelo lognormal aumenta mais rapidamente do que a variância do erro se utilizar o modelo Weibull conforme se aumenta da proporção de censura (Figura A.18).

Ao se considerar 250 observações por grupo, independentemente da proporção de censura e modelo ajustado, observou-se que o estimador baseado em modelo superestimou o erro de predição

(Tabela 4.3 e Figura A.23). Notou-se uma tendência, independentemente do modelo ajustado, de que quanto maior a proporção de censura maior o deslocamento na distribuição do erro de predição no sentido de superestimar o erro de predição (Figura A.23). Nas situações de ausência e 20% de censura, o modelo que obteve estimativas mais próximas do valor de referência foi o modelo lognormal seguido, respectivamente, pelos modelos de riscos proporcionais de Cox, Weibull e exponencial. Considerando 40% de censura, o erro de predição mais próximo do valor de referência se deu, novamente, ao utilizar o modelo lognormal (Figura A.23). Com essa proporção de censura, as distribuições dos erros de predição ao se utilizar os modelos de Weibull e de riscos proporcionais de Cox apresentaram um comportamento semelhante, mas com a distribuição do erro ao utilizar o modelo de riscos proporcionais de Cox um pouco mais deslocada a direita no sentido de superestimar o erro de predição (Figura A.23). De forma semelhante ao caso com 100 observações por grupo, a variância do erro de predição ao se utilizar os modelos Weibull e lognormal ficaram muito semelhantes para as situações de ausência e 20% de censura (Figura A.24). Com 40% de censura, a distribuição da variância do estimador do erro de predição ao se utilizar o modelo lognormal se desloca mais para a direita do que a distribuição da variância do erro de predição ao se utilizar o modelo Weibull, sendo que esse deslocamento é no sentido de aumentar a variância (Figura A.24).

Tabela 4.3: Resultados da simulação lognormal

			100 em cada grupo			250 em cada grupo		
Modelo	Censura	Estimador	$\hat{\pi}$	$\text{Var}(\hat{\pi})$	$\widehat{\text{Var}(\hat{\pi})}$	$\hat{\pi}$	$\text{Var}(\hat{\pi})$	$\widehat{\text{Var}(\hat{\pi})}$
Exponencial (13,8618)*	0%	π^{PA}	13,9231	0,0542	0,1098	16,3188	0,0291	0,0914
		π^{VC}	13,9328	0,0549	0,1114	16,3234	0,0292	0,0910
		π^{M}	31,1376	0,3484	5,2787	36,3755	0,1809	3,0333
	20%	π^{PA}	8,4343	0,4438	0,4696	8,2340	0,1447	0,1741
		π^{VC}	8,4673	0,4504	0,4610	8,2481	0,1436	0,1713
		π^{M}	37,6709	1,0134	8,1504	37,6819	0,4367	3,2628
	40%	π^{PA}	12,0299	3,3948	5,7170	12,3390	1,3043	2,3146
		π^{VC}	12,4186	3,6034	6,2565	12,4928	1,3474	2,3915
		π^{M}	51,2962	7,8053	16,5378	51,5849	3,1360	6,7056
Weibull (7,1985)*	0%	π^{PA}	7,1385	0,1942	0,2027	8,3827	0,1114	0,1219
		π^{VC}	7,2126	0,1978	0,2061	8,4179	0,1128	0,1232
		π^{M}	8,0891	0,3214	0,1463	9,5326	0,1702	0,0810
	20%	π^{PA}	6,4164	0,3083	0,1906	6,3249	0,0966	0,0648
		π^{VC}	6,4961	0,3174	0,1963	6,3559	0,0976	0,0656
		π^{M}	8,0891	0,3543	0,1479	8,2263	0,1383	0,0602
	40%	π^{PA}	6,1279	0,2193	0,2263	6,1716	0,0814	0,0964
		π^{VC}	6,2286	0,2229	0,2374	6,2205	0,0823	0,0983
		π^{M}	8,5850	0,4978	0,1654	8,7818	0,2044	0,0682
Cox (7,1435)*	0%	π^{PA}	7,0503	0,1807	21,6700	8,2935	0,1043	67,0180
		π^{VC}	7,1652	0,1907	17,6650	8,3466	0,1071	55,0869
		π^{M}	7,7357	0,2802	-	8,7918	0,1809	-
	20%	π^{PA}	6,2116	0,3430	18,7524	6,1198	0,1119	16,2546
		π^{VC}	6,3435	0,3653	15,3119	6,1719	0,1149	13,2879
		π^{M}	7,8560	0,3842	-	8,0835	0,1810	-
	40%	π^{PA}	5,3395	0,3335	12,4461	5,2735	0,0977	9,1841
		π^{VC}	5,4940	0,3401	10,1752	5,3359	0,0955	7,4691
		π^{M}	8,4757	0,6928	-	9,1029	0,4964	-
Lognormal (7,1339)*	0%	π^{PA}	7,0573	0,1797	0,1861	8,3010	0,1043	0,1143
		π^{VC}	7,1314	0,1843	0,1918	8,3361	0,1057	0,1156
		π^{M}	7,0903	0,1558	0,1468	8,3181	0,0849	0,0811
	20%	π^{PA}	6,2353	0,3285	0,2002	6,1297	0,1052	0,0701
		π^{VC}	6,3134	0,3385	0,2055	6,1604	0,1061	0,0708
		π^{M}	7,3882	0,2205	0,1606	7,4312	0,0761	0,0652
	40%	π^{PA}	5,4751	0,2011	0,1589	5,4285	0,0594	0,0529
		π^{VC}	5,5655	0,2109	0,1656	5,4659	0,0608	0,0539
		π^{M}	8,1203	0,3498	0,1988	8,2476	0,1295	0,0821

(*) Valores de referência.

4.2 Status de sobrevivência

Nessa seção serão apresentados os resultados da simulação considerando o status de sobrevivência ao utilizar a modificação dos estimadores de Uno *et al.* (2007) apresentada na seção 3.7 desse trabalho. O objetivo dessa simulação é estudar o comportamento dos estimadores da sensibilidade, especificidade e seus respectivos desvios padrão sob diferentes cenários. Esses cenários são compostos por combinações de diferentes tamanhos amostrais, taxas de censura e situações em que o marcador é fixado ou estimado.

Nessa simulação será considerada a situação em que se tem interesse em predizer o status de sobrevivência para um instante de tempo fixado t_0. Para isso, será considerada a situação em que se observa um vetor de covariáveis $\mathbf{Z} = (Z_1, Z_2)$, em que Z_1 é dado por uma variável indicadora e Z_2 por uma variável contínua. As medidas de sensibilidade e especificidade serão denotadas, respectivamente, por SE e ES.

Para a geração dos dados será utilizado um modelo de riscos proporcionais com função de risco dado por:

$$h(t|z) = h_0(t)\exp(\boldsymbol{\beta}\mathbf{z}), \tag{4.4}$$

em que $\boldsymbol{\beta}$ é um vetor de parâmetros, $\mathbf{z}$ é o vetor de covariáveis definido anteriormente e h_0 é a função de risco basal. Mais especificamente, para gerar o tempo de sobrevivência será utilizado o modelo Weibull com a seguinte função de densidade:

$$f(t) = \lambda\alpha t^{\alpha-1}\exp(-\lambda t^{\alpha}), \tag{4.5}$$

em que α é o parâmetro de forma e λ é o parâmetro de escala (Klein e Moeschberger, 2003). Dessa forma, a função de risco basal em (4.4) fica especificada por $\lambda\alpha t^{\alpha-1}$. Com isso, utilizando o método de geração sugerido por Bender *et al.* (2005), os dados representando o tempo de sobrevivência foram obtidos da seguinte forma:

$$T = \left(-\frac{\log(U)}{\lambda\exp(\boldsymbol{\beta}\mathbf{z})}\right)^{\frac{1}{\alpha}}, \tag{4.6}$$

em que U é uma observação de uma distribuição uniforme entre 0 e 1.

Nessa simulação foram considerados os seguintes valores: $\lambda = 0,005$, $\alpha = 4$, $\beta_1 = -2$ e $\beta_2 = 0,4$. Além disso, se atribuiu $Z_1 = 0$ para metade da amostra enquanto para outra metade $Z_1 = 1$ e Z_2 foi gerado segundo uma distribuição normal de média 7 e variância 1.

O tempo de censura foi gerado utilizando-se o mesmo procedimento empregado na seção 4.1.2, mas considerando uma distribuição uniforme com mínimo igual a 0 e máximo igual a 20. Ainda, considerou-se tamanhos de amostras de 200, 500, 1.000 e 5.000 observações, em que para cada tamanho amostral utilizou-se 20% e 40% de censura. Para cada cenário, assim como na simulação do tempo de sobrevivência, considerou-se 1.000 réplicas de cada situação na simulação.

Lembre-se que para a modificação do estimador proposto por Uno *et al.* (2007), apresentada nesse trabalho, é necessário a definição de um marcador. Assim, será utilizado o modelo de regressão de Weibull e de riscos proporcionais de Cox e o marcador será dado pelo preditor linear desses modelos. Dessa forma, fixado um conjunto de dados e os marcadores, obtiveram-se as estimativas

da SE, EP, assim como dos seus respectivos desvios padrão. As estimativas dos desvios padrão dessas quantidades foram obtidas com a utilização dos estimadores apresentados na seção 3.7 e serão denotadas, respectivamente, por $\widehat{DP}(SE)$ e $\widehat{DP}(ES)$. As estimativas dos desvios padrão obtidas com as réplicas da simulação serão denotadas por $DP(\widehat{SE})$ e $DP(\widehat{ES})$. Além disso, fixados um conjunto de dados e a regra de decisão criada, aplicou-se essa mesma regra num conjunto de dados independente gerado sem censura. Com isso, espera-se ter um valor de referência para comparar com o valor estimado da SE e ES. Esses valores de referência foram calculados da seguinte forma:

$$SE^v(c,t_0) = \frac{\sum_{i=1}^n \mathbb{I}(M > c; T \geq t_0)}{\sum_{i=1}^n \mathbb{I}(T \geq t_0)} \text{ e } ES^v(c,t_0) = \frac{\sum_{i=1}^n \mathbb{I}(M \leq c; T > t_0)}{\sum_{i=1}^n \mathbb{I}(T > t_0)}. \tag{4.7}$$

Os valores de cortes utilizados para se calcular a SE e ES foram determinados de forma a maximizar a estatística de Kolmogorov-Smirnov (da Silva, 2010). Essa estatística é dada por:

$$\text{KS} = \sup_m |F_E(m) - F_{\bar{E}}(m)|, \tag{4.8}$$

em que

$$F_E(m) = \frac{\text{nº de observações que apresentaram evento com marcador} > \text{m}}{\text{nº total de eventos}} \text{ e}$$

$$F_{\bar{E}}(m) = \frac{\text{nº de observações que não apresentaram evento com marcador} > \text{m}}{\text{nº total de observações sem eventos}}.$$

Note que essa duas quantidades podem ser consideradas como estimativas da sensibilidade e (1 - especificidade). Dessa forma, a estatística KS nessa situação seria dada por:

$$\text{KS} = \sup_m |SE(m) - (1 - ES(m))| = \sup_m |SE(m) + ES(m) - 1|. \tag{4.9}$$

Note que essas quantidades são definidas para um instante de tempo fixado. Com base em resultados de uma simulação preliminar para determinar o corte ótimo, optou-se, quando utilizou-se o modelo de regressão Weibull, por trabalhar com os cortes -0,85 e -0,9. Quando utilizou-se o marcador baseado no preditor linear do modelo de riscos proporcionais de Cox, optou-se por trabalhar com o valor de corte dado por 2,05.

Para as situações em que se utilizou o modelo Weibull com 20% de censura, observou-se que as estimativas de sensibilidade e especificidade ficaram muito próximas dos valores de validação, mesmo quando se considerou o menor tamanho amostral (Tabela A.1). Embora tenha sido observada uma pequena diferença, as sensibilidades estimadas ficaram mais próximas dos valores de validação do que as estimativas das especificidade. Ainda, notou-se que os desvios padrão estimados com a abordagem dada na seção 3.7 apresentaram valores maiores que os desvios padrão das simulações, sendo que essa diferença diminuiu com o aumento do tamanho da amostra (Tabela A.1) .

Nas situações com 40% de censura, observou-se o mesmo padrão de quando se considerou o caso com 20% de censura, ou seja, as estimativas da sensibilidade e especificidade ficaram muito próximas dos seus respectivos valores de validação, os desvios padrão obtidos com o estimador apresentado na seção 3.7 foram maiores que os desvios obtidos pela simulação, sendo que essa diferença tende a diminuir com o aumento do tamanho da amostra (Tabela A.2).

Fixando o tamanho amostral e variando a proporção de censura, observou-se, como esperado, que os desvios padrão dos estimadores da sensibilidade e especificidade, geralmente, são maiores quando se considera a maior proporção de censura (Tabela A.1 e Tabela A.2).

Para as situações em que se utilizou o modelo de riscos proporcionais de Cox, observou-se o mesmo padrão de quando foi considerado o modelo Weibull, com exceção do comportamento dos desvios (Tabela A.3). Notou-se que, para o modelo de riscos proporcionais de Cox, os desvios padrão estimados com a técnica de equação de estimação generalizada subestimaram os desvios padrão obtidos pela simulação (Tabela A.3). Assim, para as situações em que se utiliza o modelo de riscos proporcionais de Cox, propõe-se a utilização de técnicas de reamostragem para a estimação do desvio padrão. No entanto, quando se comparam os valores obtidos com o modelo Weibull com os valores obtidos com modelo de riscos proporcionais de Cox, observaram-se menores valores das sensibilidades e especificidades, independentemente do tamanho amostral e da censura, para o modelo de Cox. Ainda, observou-se que essa diferença diminui com o aumento do tamanho amostral. Além disso, notou-se que as medidas obtidas com o modelo de riscos proporcionais de Cox apresentaram desvios padrão maiores que as medidas obtidas com o modelo Weibull (Tabelas A.1, Tabela A.2 e Tabela A.3). Essa diferença pode ser devido a maior variabilidade dos parâmetros do modelo de riscos proporcionais de Cox que terá como consequência uma maior variabilidade no preditor linear, uma vez que o preditor linear é função dos parâmetros do modelo ajustado.

Para verificar a diferença do desempenho preditivo ao considerar o marcador fixado e estimado, utilizou-se outro estudo de simulação ainda com 1.000 réplicas de cada situação. Nesse estudo o marcador fixado foi dado pelo preditor linear obtido com os verdadeiros valores dos parâmetros utilizado na geração dos dados. Considerando o modelo Weibull com 20% de censura, observou-se que os valores da sensibilidade, especificidade e seus respectivos desvios padrão, quando se considerou os marcadores fixados, ficaram muito próximos dessas medidas ao se utilizar os marcadores estimados (Tabela A.4 e Tabela A.5). Além disso, tanto para o marcador fixado quanto para o marcador estimado, os valores da sensibilidade e especificidade ficaram próximos dos valores de validação. Já quando considerou-se o modelo de riscos proporcionais de Cox com 20% de censura, observou-se que as medidas de sensibilidade e especificidade com o marcador fixado também ficaram próximas das medidas obtidas com o marcador estimado (Tabela A.6 e Tabela A.7). No entanto, a distância entre essas medidas foi maior do que quando considerou-se o modelo Weibull. Além disso, observou-se que, tanto para sensibilidade quanto para a especificidade, o desvio padrão estimado com a simulação foi aproximadamente três vezes menor do que o desvio padrão da simulação quando considerou-se um tamanho amostral igual a 200. No entanto, essa diferença diminui com o aumento do tamanho amostral.

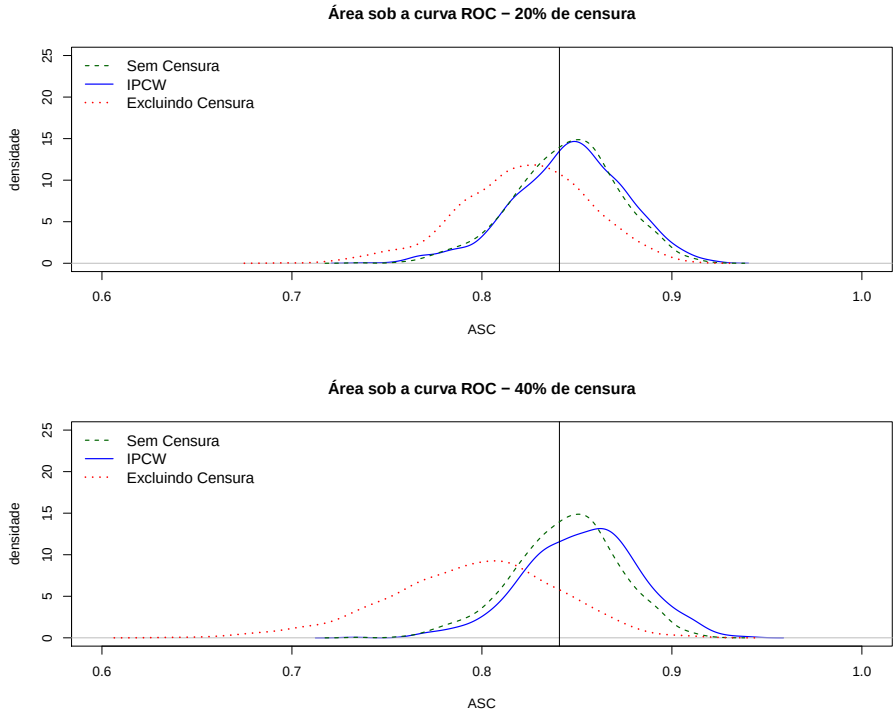
Estudo Área Sob a Curva ROC

Nesse estudo utilizou-se um esquema semelhante ao de seção anterior. Os dados foram gerados da mesma forma e utilizou-se o preditor linear como marcador. Note que aqui está sendo utilizado o marcador estimado. Assim, comparou-se as estimativas da área sob a curva ROC obtidas de três formas. O procedimento adotado foi: gerava-se um conjunto de dados sem censuras e calculava-se a

área sob a curva ROC utilizando os estimadores empírico da sensibilidade e especificidade dados em (3.27) e (3.28). Então, uma proporção fixada de dados era censurada, obtinham-se os estimadores adaptados da sensibilidade e especificidade com a utilização da técnica apresentada na seção 3.7 e calculava-se a área sob a curva. Por fim, excluíam-se as observações censuradas e calculava-se a área sob a curva com os estimadores empíricos citados anteriormente.

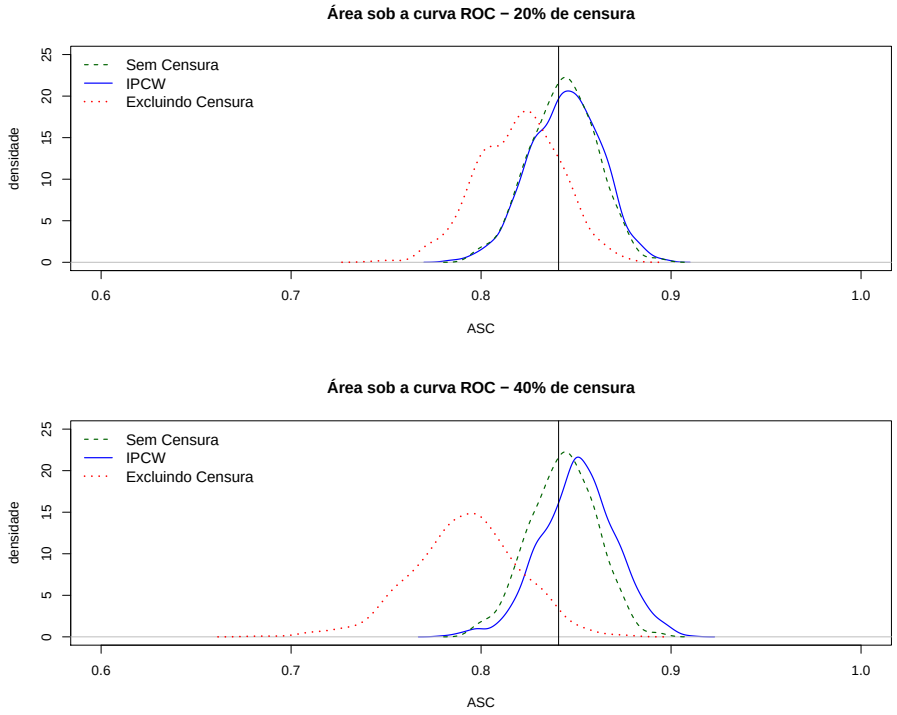
Observou-se, considerando um total de 200 observações, que o estimador apresentado nesse trabalho tende a compensar as obervações censuradas. Isso porque as estimativas da área sob a curva utilizando essa abordagem ficaram muito próximas das estimativas dos casos em que não havia censura (Figura 4.3). Já quando as obervações censuradas são excluídas, observou-se um viés no sentido de subestimar a área sob a curva ROC sendo que essa subestimação aumentou com o aumento da proporção de censura (Figura 4.3).

Figura 4.3: Funções densidades das distribuições da área sob a curva ROC considerando diferentes estimadores e um tamanho amostral de 200 observações



Considerando um total de 500 observações notou-se uma redução na variância das estimativas da área sob a curva para os três estimadores (Figura 4.4). No entanto, foi observado o mesmo padrão da situação em que se consideraram 200 observações, ou seja, as estimativas dadas pelo estimador apresentado nesse trabalho ficaram bem próximas das estimativas calculadas com o banco sem censura e quando as observações censuradas são excluídas observou-se um viés no sentido de subestimar a área sob a curva ROC (Figura 4.4). Além disso, novamente, quando as observações censuradas são excluídas o viés no sentido de subestimar a área sob a curva foi maior para a maior proporção de censura.

Figura 4.4: Funções densidades das distribuições da área sob a curva ROC considerando diferentes estimadores e um tamanho amostral de 500 observações



No próximo capítulo será feita uma aplicação das técnicas apresentadas no capítulo 3 em dados de um estudo do Instituto do Câncer de São Paulo.

Capítulo 5

Aplicação

A aplicação das técnicas apresentadas nesse trabalho será ilustrada com dados de um estudo que está sendo realizado no Instituto do Câncer do Estado de São Paulo (ICESP), que foi um dos problemas motivadores desse trabalho.

Esse estudo foi desenvolvido para a criação de um modelo de predição do tempo de sobrevivência ajustado pela qualidade de vida de pacientes severamente acometidos pelo câncer num estágio inicial de admissão na UTI e durante os 5 dias seguintes. Embora o objetivo do estudo tenha sido esse, o enfoque nesse trabalho será dado na predição do tempo de sobrevivência a partir da internação do paciente na UTI, independentemente da qualidade de vida. Para isso, serão considerados os dados de 721 pacientes com o seguinte conjunto de covariáveis:

- **idade**: medida em anos;
- **sexo**: masculino e feminino;
- **imc**: índice de massa corpórea dado pela divisão do peso em quilogramas pelo quadrado da altura em metros;
- **status do câncer**: ativo - diagnóstico recente, ativo - recaída/progressão e controlado - remissão;
- **extensão do câncer**: leucemias, limitado, totalmente avançado e metástase à distância;
- **cirurgia**: sim e não;
- **radioterapia**: sim e não;
- **quimioterapia**: sim e não;
- **ventilação**: espontânea, mecânica invasiva e não invasiva;
- **intervalo de dias entre a internação no hospital e a entrada na UTI**;
- **infecção respiratória na UTI**: sim e não;
- **sedação profunda**: sim e não;

- **alcoolismo**: sim e não;
- **tabagismo**: atual e nunca/prévio; e
- **local agrupado**: hematológicas e sólidas.

Conforme mencionado anteriormente, a amostra é composta por 721 indivíduos, sendo 418 (58%) homens e 303 (42%) mulheres (Tabela A.8). Para o status do câncer, observou-se 50,6% com diagnóstico recente, 43,6% com recaída ou progressão e 5,8% em remissão. Em relação à extensão do câncer, notou-se a proporção de 1,8%, 32,7%, 35,8% e 29,7% para as respectivas categorias, leucemia, limitado, localmente avançado e metástase à distância. Observou-se que 60,1% passaram por cirurgia, enquanto 39,9% não foram submetidos à cirurgia. Ainda, notou-se que 22,7% receberam radioterapia, enquanto 38,7% foram tratados com quimioterapia (Tabela A.8).

Em relação a ventilação, foi observada uma proporção de 75,9% de pacientes com ventilação espontânea, 20,2% com ventilação mecânica invasiva e apenas 3,9% com ventilação não invasiva. Do total de indivíduos observados, 86,5% tiveram infecção respiratória na UTI. Foi observada uma proporção de 14,8% de indivíduos que foram sedados profundamente. Já para alcoolismo, foram observados 11,5% dos pacientes com problemas de alcoolismo. Para tabagismo, foi observada uma proporção de 14,3% de tabagistas atuais e 85,7% que nunca foram tabagistas ou foram previamente. Ainda, para o local agrupado, foi observada uma proporção de 94,3% de sólidas e 5,7% de hematológicas (Tabela A.8).

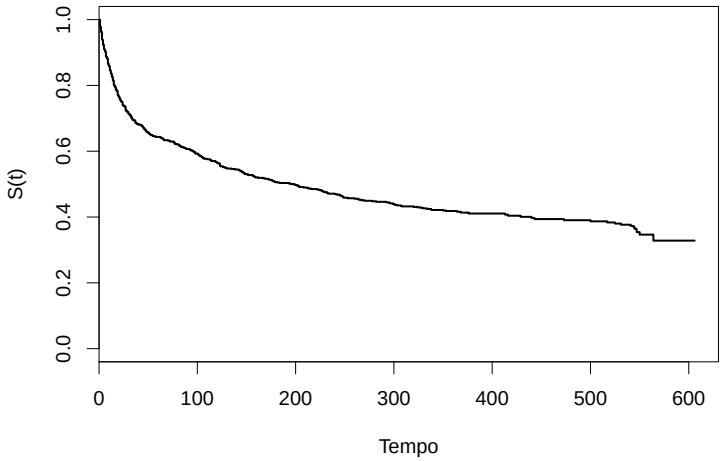
A idade mediana observada foi de 61,17 anos com intervalo interquartil de 17,64 anos. O IMC mediano foi de 23,49 kg/m^2 com intervalo interquartil de 6,65 kg/m^2. Já a mediana do número de dias entre a entrada no hospital até a entrada na UTI foi de 1 dia com intervalo interquartil de 3 dias (Tabela 5.1).

Tabela 5.1: Estatísticas Descritivas

	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo
Idade (anos)	16,86	53,43	61,17	71,07	91,21
IMC (kg/m^2)	11,94	20,55	23,49	27,20	49,31
Dias até entrada na UTI	0	0	1	3	366

O tempo mediano de sobrevivência foi de 198 dias e o maior tempo de acompanhamento foi de 564 dias (Figura 5.1). Dos 721 pacientes observados, 429 (59,5%) apresentaram óbito durante o estudo e 292 (40,5%) foram observados com o tempo de sobrevivência censurado. É importante ressaltar que esse estudo está em andamento, portanto os dados ainda estão sendo coletados. A análise apresentada nessa dissertação foi realizada com os dados mais recentes que são de junho de 2012.

Figura 5.1: Função de sobrevivência ICESP



Observou-se que os indivíduos que não foram submetidos à cirurgia apresentaram um menor tempo de sobrevivência em relação aos indivíduos que foram submetidos à cirurgia (Figura A.26). Para o status do câncer, o menor tempo de sobrevivência ocorreu para os indivíduos com recaída, já os indivíduos em remissão e com diagnóstico recente apresentaram um comportamento semelhante quanto ao tempo de sobrevivência, mas com a curva de sobrevivência dos pacientes do grupo em remissão um pouco abaixo da curva dos pacientes com diagnóstico recente (Figura A.27).

Os indivíduos submetidos à radioterapia apresentaram um tempo de sobrevivência menor em relação aos indivíduos que não receberam esse tratamento (Figura A.28). Observou-se esse mesmo padrão para os indivíduos submetidos à quimioterapia (Figura A.29). Em relação ao alcoolismo, até aproximadamente o centésimo dia, os indivíduos com problemas de alcoolismo apresentaram um comportamento na sobrevivência muito próximo aos indivíduos sem esse problema, mas, a partir do centésimo dia, os indivíduos com alcoolismo apresentaram um menor tempo de sobrevivência em relação aos indivíduos sem problemas de alcoolismo (Figura A.30). Para a variável tabagismo, a sobrevivência dos pacientes que são tabagistas atuais apresentou um comportamento muito semelhante a da curva de sobrevivência dos pacientes que foram tabagistas previamente ou nunca foram tabagistas (Figura A.31).

Observou-se que os indivíduos com ventilação não invasiva apresentaram o menor tempo de sobrevivência entre os três tipos de ventilação. A maior sobrevivência foi observada para os indivíduos com ventilação espontânea e a ventilação mecânica invasiva apresentou um comportamento intermediário entre a ventilação não invasiva e espontânea (Figura A.32).

Os indivíduos com infecção respiratória na UTI apresentaram um tempo de sobrevivência muito abaixo dos indivíduos sem infecção respiratória (Figura A.33). Em relação ao local agrupado, observou-se um comportamento muito semelhante entre os indivíduos com hematológicas e com sólidas. Esses dois grupos apresentaram um comportamento muito próximo até o dia 200 aproximadamente e, a partir desse tempo, os indivíduos com sólidas apresentaram uma sobrevivência um pouco maior (Figura A.34).

Inicialmente ajustou-se um modelo de riscos proporcionais de Cox, no entanto, para a covariável radioterapia, observou-se fortes indícios de que a hipótese de riscos proporcionais para essa variável não é válida. Por isso, optou-se por utilizar um modelo de riscos proporcionais de Cox estratificado pela covariável radioterapia. Assim, foram considerados dois modelos estratificados: um com todas as variáveis citadas anteriormente (Tabela A.9) e, o segundo, considerando um determinado método de seleção de variáveis (Tabela A.10). O método de seleção de variáveis utilizado num primeiro instante foi o AIC e a partir do modelo reduzido com esse método desconsiderou-se as variáveis com um valor-p menor que 0,1. Para verificar a possibilidade dos coeficientes do modelo não serem comuns para os estratos, ajustou-se modelos que consideram coeficientes distintos para cada estrato e utilizou-se o teste da razão de verossimilhanças para verificar essa hipótese (Colosimo e Giolo, 2006). Assim, não rejeitou-se a hipótese de que os coeficientes são comuns aos estratos (TRV = 20,19; valor-p = 0,32 e g.l. = 18). Além disso, o banco de dados foi separado em dois conjuntos: um de treinamento, com 80% das observações, e um de validação, com o restante dos casos.

No processo de redução de variáveis do modelo, algumas categorias das variáveis extensão do câncer e ventilação foram agrupadas. A variável extensão do câncer foi tratada com duas categorias: *limitado* e *outros*. Por fim, a variável ventilação foi agrupada em duas categorias: *não invasiva* e *outras*. No modelo reduzido, além dessas duas variáveis agrupadas, trabalhou-se com as variáveis IMC, cirurgia, infecção respiratória e sedação profunda (Tabela A.10).

Para verificar o desempenho preditivo do modelo reduzido em relação ao modelo completo foram aplicadas algumas técnicas com função de perda e, também, com curva ROC. Para a avaliação do desempenho preditivo em relação ao tempo de sobrevivência, utilizou-se a abordagem com função de perda considerando a mediana estimada como preditor e a função de perda absoluta. Ainda, utilizou-se para o estimador validação cruzada um número de grupos igual ao tamanho amostral e para o estimador baseado em modelo utilizou-se 200 reamostragens. O modelo completo, como esperado, apresentou a menor perda nos três estimadores (Tabela 5.2). Embora o modelo completo tenha apresentado a menor perda em todas as situações, o erro de predição foi bem próximo ao erro de predição com o modelo reduzido. Além disso, foram calculas as perdas esperadas, considerando os modelos ajustados com o banco de dados de treinamento, em relação aos indivíduos do conjunto de dados de validação que apresentaram óbito. Assim, obteve-se uma perda esperada de 203,94 com o modelo completo e uma perda de 200,22 com o modelo reduzido. Note que os valores das perdas esperadas considerando os modelos completo e reduzido ficaram muito próximos, tanto para os valores obtidos com o conjunto de dados de treinamento quanto para os valores obtidos com o conjunto de validação. Isso indica que, para fins preditivos, o modelo reduzido deve ser tão bom quanto o modelo completo.

Tabela 5.2: Estimadores do erro de predição

Modelo	Perda Aparente	Validação Cruzada	Baseado em Modelo
Completo	117,15	124,47	182,06
Reduzido	121,67	124,88	190,20

Para utilizar as técnicas apresentadas no Capítulo 3, foram fixados 9 instantes de tempo. Esses instantes foram definidos a partir dos quantis 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% e 90% da distribuição amostral do tempo de sobrevivência. Além disso, optou-se por trabalhar com os marcadores dados pelos valores da distribuição acumulada sob os determinados modelos ajustados nos instantes de tempo fixados.

Utilizando a abordagem *incidente/dinâmico* apresentada na seção 3.5, calculou-se a área sob a curva ROC, a estatística KS e o coeficiente de concordância para os instantes de tempos fixados. Sob essa abordagem, essas quantidades foram obtidas utilizando-se o pacote risksetROC do software R (Heagerty e Saha-Chaudhuri, 2012). Observou-se, considerando o conjunto de treinamento, que a área sob a curva ROC para os modelos completo e reduzido são muito próximas, sendo que essa área geralmente é maior para o modelo completo (Tabela A.11). Além disso, observou-se que as áreas sob a curva ROC calculadas com os banco de validação apresentaram valores próximos dos valores obtidos com o banco de dados de treinamento. Esse mesmo padrão foi observado quando considerou-se a estatística KS (Tabela A.11).

Para o coeficiente de concordância, observou-se que os coeficientes calculados com o modelo completo apresentaram valores muito próximos dos valores obtidos com o modelo reduzido (Tabela 5.3). Ainda, observou-se que os coeficientes calculados considerando os conjuntos de validação apresentaram valores muito próximos, embora um pouco menores, dos valores obtidos com os conjuntos de treinamento. Assim, considerando a abordagem *incidente/dinâmica*, observou-se que o modelo reduzido apresenta uma performance muito próxima do modelo completo.

Tabela 5.3: Coeficiente de concordância

	Modelo Completo		Modelo Reduzido	
	Treinamento	Validação	Treinamento	Validação
Coeficiente de Concordância	0,5346	0,5325	0,5321	0,5299

Para a abordagem *cumulativo/dinâmico* serão apresentadas estimativas da área sob a curva ROC utilizando o estimador proposto na seção 3.6, assim como estimativas da área sob a curva ROC e KS utilizando os estimadores apresentados na seção 3.7. Lembre que o estimador dado pela equação (3.20) da Seção 3.6 é dado por:

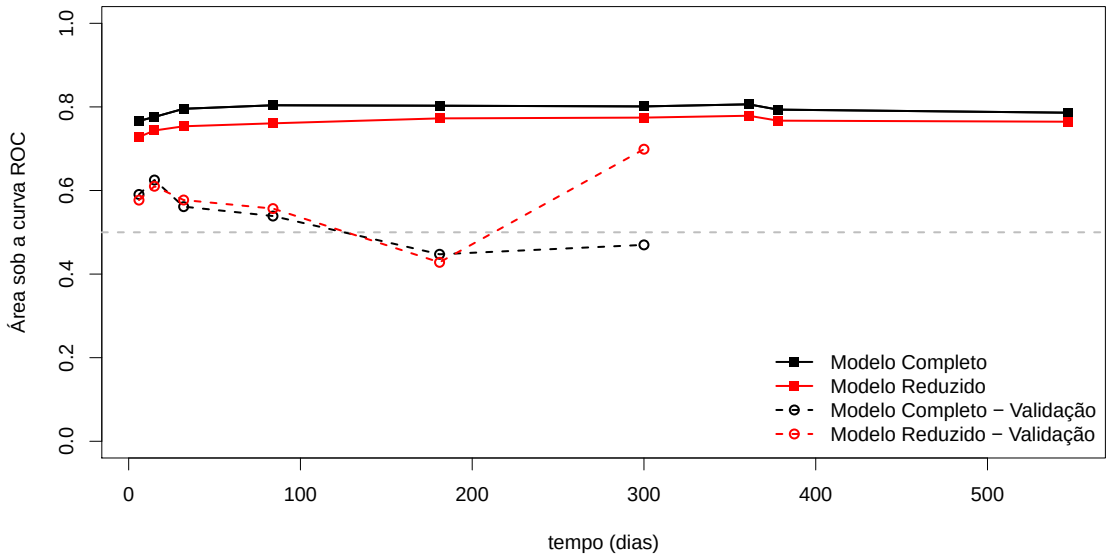
$$ASC_k^{\mathbb{C}/\mathbb{D}}(t) = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{i}{n}\hat{F}(t|M_{(i)}) - \hat{F}_2^2(t)/2}{\hat{F}_2(t)\left[1-\hat{F}_2(t)\right]}.$$

Assim, ajustou-se o modelo especificado para se obter o marcador que, como citado anteriormente, será dado pela função de distribuição acumulada do tempo de sobrevivência no instante fixado. Após obter esses marcadores, ajustou-se novamente um modelo de riscos proporcionais de Cox para se obter $\hat{F}(t|M_{(i)})$ e, consequentemente, $\hat{F}_2(t)$. É importante ressaltar que, neste caso, foi ajustado inicialmente um modelo de Cox estratificado e, dessa forma, os marcadores, dados pela função de distribuição acumulada, incorporam a diferença entre os estratos. As estimativas obtidas com o conjunto de validação foram obtidas utilizando-se a função de distribuição estimada com o conjunto de dados de treinamento para se obter o marcador e, posteriormente, utilizou-se esses marcadores

com a função $\hat{F}(t|M_{(i)})$ obtida com a utilização dos dados de treinamento. Assim, observou-se, para a situação com os dados de treinamento, que os valores da área sob a curva obtidos com o modelo reduzido ficaram muito próximos dos valores obtidos com o modelo completo, embora tenham sido sempre um pouco menores (Tabela A.12). Ainda, observou-se, tanto para o modelo completo quanto para o modelo reduzido, que a área sob a curva ROC calculada com o conjunto de validação apresentou valores menores do que os valores obtidos com o conjunto de treinamento. Além disso, notou-se um aumento da distância entre essas medidas considerando os valores calculados a partir de 300 dias após a internação.

Como citado anteriormente, utilizando a modificação do estimador proposto por Uno *et al.* (2007) na seção 3.7, construiu-se a curva ROC com as estimativas de sensibilidade e especificidade e calculou-se a área sob essa curva para os diferentes instantes de tempo fixados. Além disso, calculou-se a medida KS nesses instantes de tempo. Embora, em geral, não tenha sido uma grande diferença, notou-se que o modelo completo apresentou medidas melhores em relação ao modelo reduzido em todos os quantis (Figura 5.2 e Tabela A.13). Ainda, notou-se que as medidas são bem estáveis ao longo do tempo.

Figura 5.2: área sob a curva ROC ao longo do tempo obtida com o estimador apresentado na seção 3.7



Para os dados de validação, a área sob a curva ROC e a estatística KS foram calculadas utilizando-se as equações dadas em (4.7) na seção 4.2. Por isso, foram utilizados apenas os indivíduos que apresentaram óbito. Considerando o conjunto de validação, foi observado um total de 144 indivíduos, sendo que 84 óbitos apresentaram óbito. Além disso, o maior tempo observado de óbito foi de 328 dias após a internação na UTI, por isso calculou-se as medidas de desempenho até 300 dias após a internação. Assim, observou-se que a área sob a curva ROC ao longo do tempo para o modelo completo e reduzido ficaram muito próximas, com exceção do instante dado por 300 dias após a internação (Figura 5.2 e Tabela A.13). No entanto, existem apenas dois indivíduos com

tempo de óbito maior ou igual a 300 dias, por isso, para esse instante de tempo, essa medida deve ser analisada com cuidado. Além disso, as medidas obtidas com o conjunto de dados de validação apresentaram valores sensivelmente menores do que as medidas obtidas com o conjunto de dados de treinamento. A medidas KS apresentou o mesmo padrão obtido com a área sob a curva ROC com exceção das medidas obtidas 300 dias após a internação. Nesse instante, a medidas obtida com o conjunto de validação apresentou valores maiores do que a medida obtida com o conjunto de dados de treinamento (Tabela A.13).

Em todas as abordagens, notou-se que as medidas de desempenho dos modelos completo e reduzido, considerando o conjunto de treinamento, são muito próximas. Isso vale, também, para as medidas considerando o conjunto de validação. Entre esses modelos, para fins preditivos, não percebe-se um ganho significativo ao se utilizar o modelo completo. Por isso, entre esses dois modelos, indica-se a utilização do modelo reduzido. Além do motivo citado anteriormente, o modelo reduzido é um modelo mais simples e com isso pode se minimizar a chance de erros de digitação no banco de dados, de exclusão de pacientes com variáveis faltantes e o tempo de preenchimento dos dados. No entanto, considerando as medidas de desempenho, seria aconselhável trabalhar com outro conjunto de covariáveis com o qual se tenha um melhor desempenho preditivo. Embora alguns autores utilizem padrões de classificação para as medidas de área sob a curva ROC e KS, aqui não será utilizada essa classificação porque essas faixas devem depender exclusivamente do problema em questão, do custo de erros como falsos positivos ou acertos como, por exemplo, verdadeiros positivos. No entanto, valores de coeficiente de concordância próximos a 0,53, como foi observado sob a abordagem *incidente/dinâmico*, são valores baixos, ou seja, não há uma forte associação entre o marcador e o tempo de sobrevivência. Além disso, a grande maioria das medidas de área sob a curva ROC, sob todas as abordagens, ficaram abaixo de 0,6.

Capítulo 6

Considerações

Nesse trabalho foram apresentadas diferentes técnicas de métodos de avaliação do desempenho preditivo de modelos de análise de sobrevivência. Ainda que seja de grande interesse a utilização de modelos de análise de sobrevivência com objetivos de predição, a literatura nesse assunto é muito esparsa e não há consenso em quais técnicas são as mais adequadas. Por isso, um dos objetivos desse trabalho foi reunir as principais abordagens de avaliação desse desempenho e, quando possível, compará-las. Além disso, grande parte das técnicas apresentadas foram implementadas num software livre e serão disponibilizadas futuramente.

Nessa dissertação estudou-se duas principais abordagens: função de perda e curva ROC. Na abordagem com função de perda, foi feito um estudo de simulação em que observou-se que, na presença de censura, os estimadores perda aparente e validação cruzada tendem a subestimar o erro de predição. Observou-se uma tendência de aumentar o viés, no sentido de subestimar o erro de predição, com o aumento da proporção de censura. Já o estimador baseado em modelo tende a superestimar o erro de predição, mas a tendência do viés desse estimador é o contrário da tendência dos outros estimadores. Para o estimador baseado em modelo, com o aumento da proporção de censura, tende-se a superestimar o erro de predição. Por isso, para situações em que não há grandes desvios das suposições do modelo utilizado, aconselha-se a utilização do estimador baseado em modelo. Além disso, especificamente para o modelo de riscos proporcionais de Cox, mesmo quando a suposição de riscos proporcionais não seja válida, esse modelo apresentou valores do erro de predição próximos do valor de referência. Ainda nessa abordagem, considerou-se que a distribuição do tempo de censura era independente das covariáveis. Para trabalhos futuros pode-se estudar o comportamento desses estimadores para os casos em que a censura depende de alguma covariável utilizada no modelo de predição. Além disso, seria interessante estudar como definir um termo de penalização do erro de predição quando se considera o estimador perda aparente.

No contexto da curva ROC foram apresentadas diferentes técnicas de avaliação do desempenho preditivo de modelos de sobrevivência. No entanto, muitas dessas técnicas não são comparáveis, pois estimam quantidades diferentes ou são aplicadas em contextos distintos. Além disso, foi proposta uma adaptação dos estimadores, propostos por Uno *et al.* (2007), da sensibilidade e especificidade que permite se utilizar marcadores biológicos ou marcadores dados por um escore numérico e não apenas pela função de sobrevivência. É importante ressaltar que, com a adaptação apresentada

nesse trabalho, foram obtidos estimadores das variâncias da sensibilidade e da especificidade sob essa abordagem. Normalmente, para as técnicas apresentadas nesse contexto, não há estimadores das variâncias dessas quantidades. A avaliação da variância dessas técnicas são feitas utilizando-se métodos de reamostragem. Ainda sob a abordagem da curva ROC, foi feita uma simulação para verificar a diferença em se utilizar o marcador estimado e o marcador fixado. Com base nessa simulação, verificou-se uma grande proximidade dos valores da sensibilidade e especificidade e, também, que as respectivas variâncias, geralmente, são maiores quando utilizam-se os marcadores estimados. A ordem de grandeza dessa variância deve depender da variabilidade dos estimadores utilizado para gerar os marcadores. Por fim, comparou-se estimativas da área sob a curva ROC considerando estimativas obtidas a partir de estimadores empíricos de dados sem censura, com a aplicação da técnica proposta nesse trabalho censurando parte das observações e excluindo-se as observações censuradas. Notou-se que o estimador baseado na adaptação apresentada nesse trabalho gerou estimativas muito próximas das estimativas da área sob a curva ROC obtidas com dados sem censuras enquanto observou-se um grande viés no sentido de subestimar essa área ao se excluir as observações censuradas. Como foram obtidos resultados encorajadores com essa técnica, pretende-se estudar o comportamento desses estimadores sob diferentes cenários e as suas propriedades.

No capítulo de aplicações das técnicas apresentadas no conjunto de dados do ICESP, observou-se que um modelo com um número expressivamente menor de covariáveis foi capaz de predizer o tempo ou o status de sobrevivência tão bem quanto o modelo considerando todas as covariáveis. Além disso, observou-se que o desempenho preditivo do modelo de riscos proporcionais de Cox estratificado, para o modelo reduzido ou completo, não muda muito com o passar do tempo. Na prática, não só para esse exemplo, trabalhar com um modelo preditivo reduzido que tenha um desempenho tão bom quanto um modelo completo pode significar uma enorme economia de recursos.

Apêndice A

Gráficos e Tabelas

A.1 Capítulo 4 - Simulação do tempo de sobrevivência

Gráficos

Figura A.1: Resultados da simulação exponencial considerando 100 indivíduos em cada grupo para o estimador perda aparente

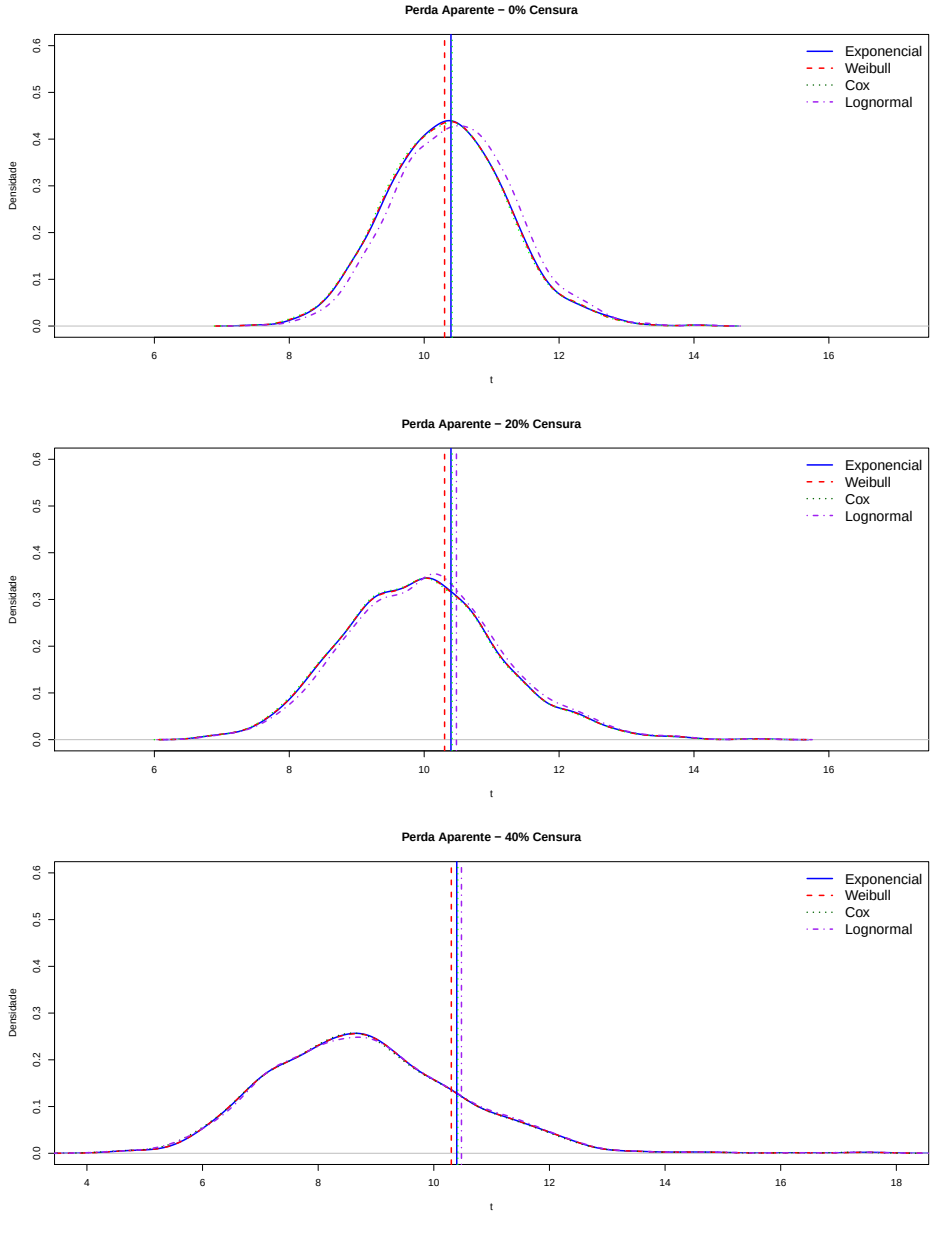


Figura A.2: Resultados da simulação exponencial considerando 100 indivíduos em cada grupo para a variância do estimador perda aparente

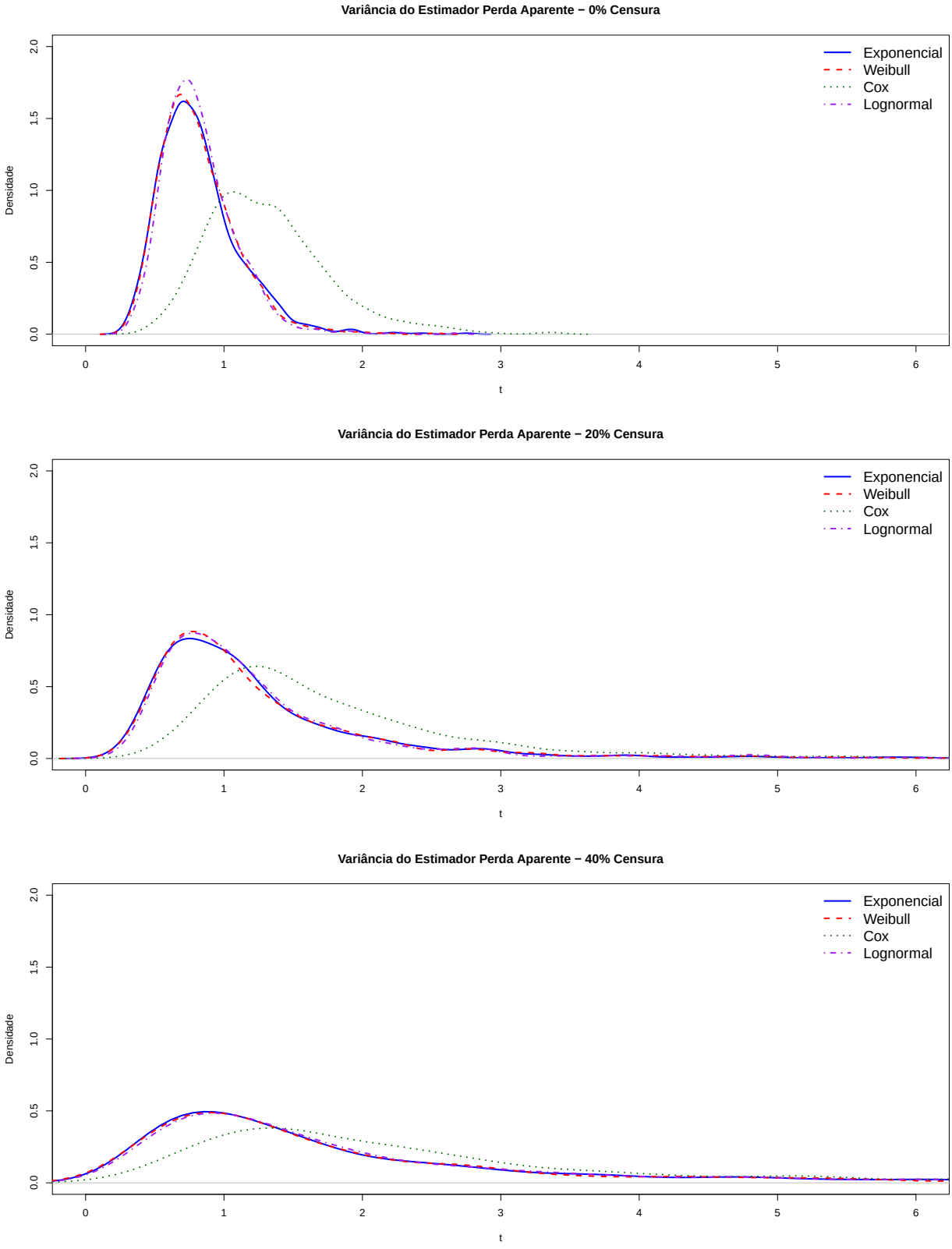


Figura A.3: Resultados da simulação exponencial considerando 100 indivíduos em cada grupo para o estimador validação cruzada

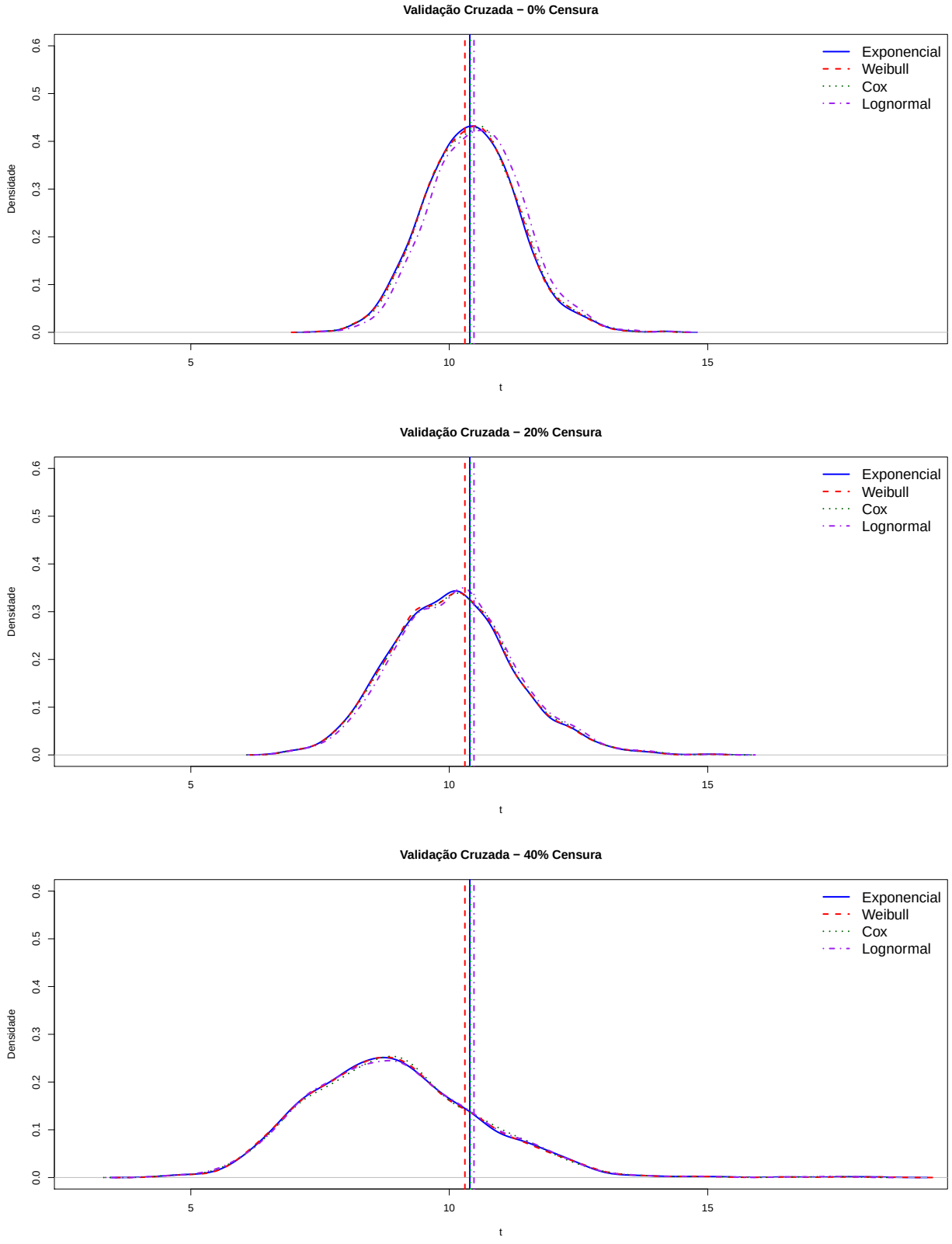


Figura A.4: Resultados da simulação exponencial considerando 100 indivíduos em cada grupo para a variância do estimador validação cruzada

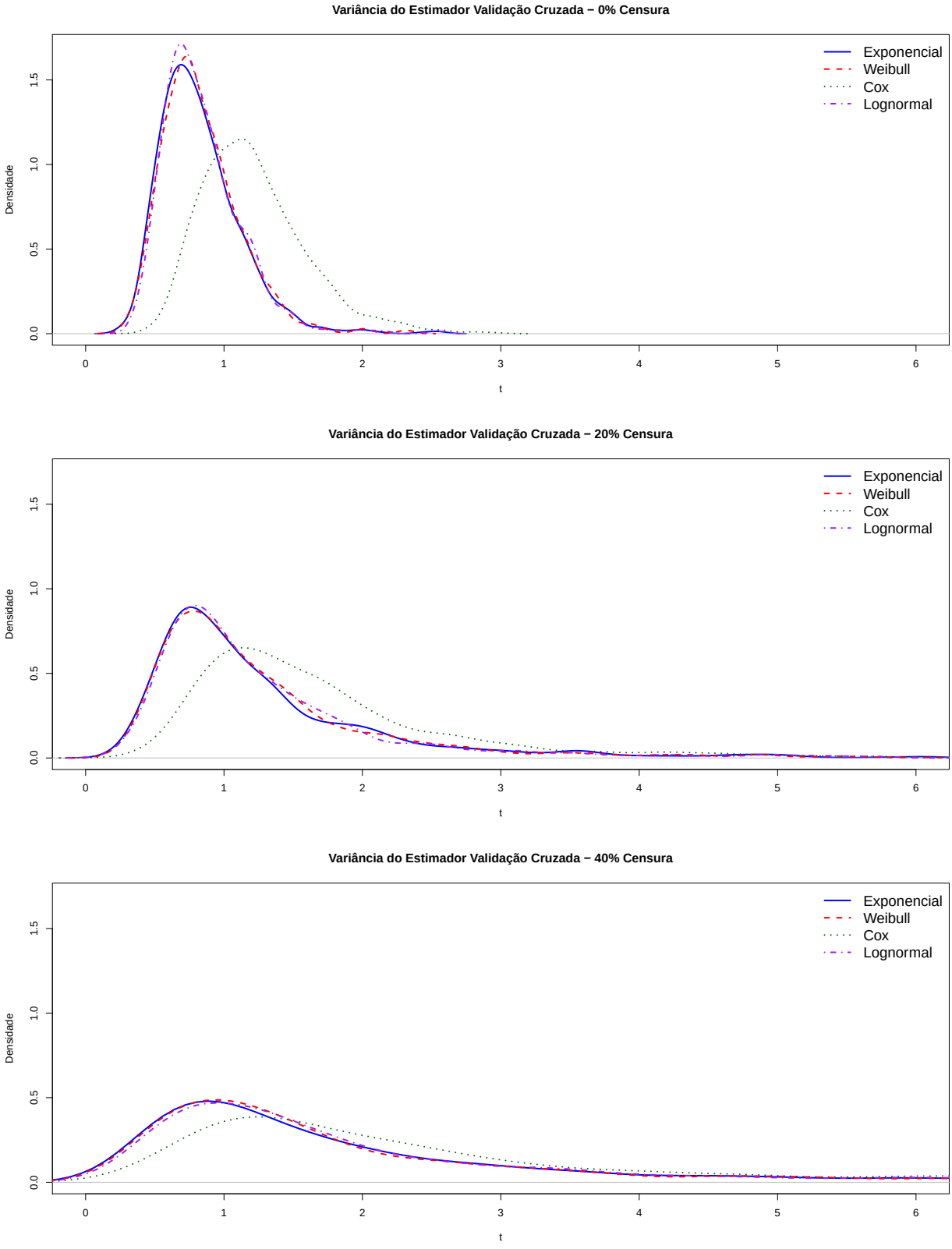


Figura A.5: Resultados da simulação exponencial considerando 100 indivíduos em cada grupo para o estimador baseado em modelo

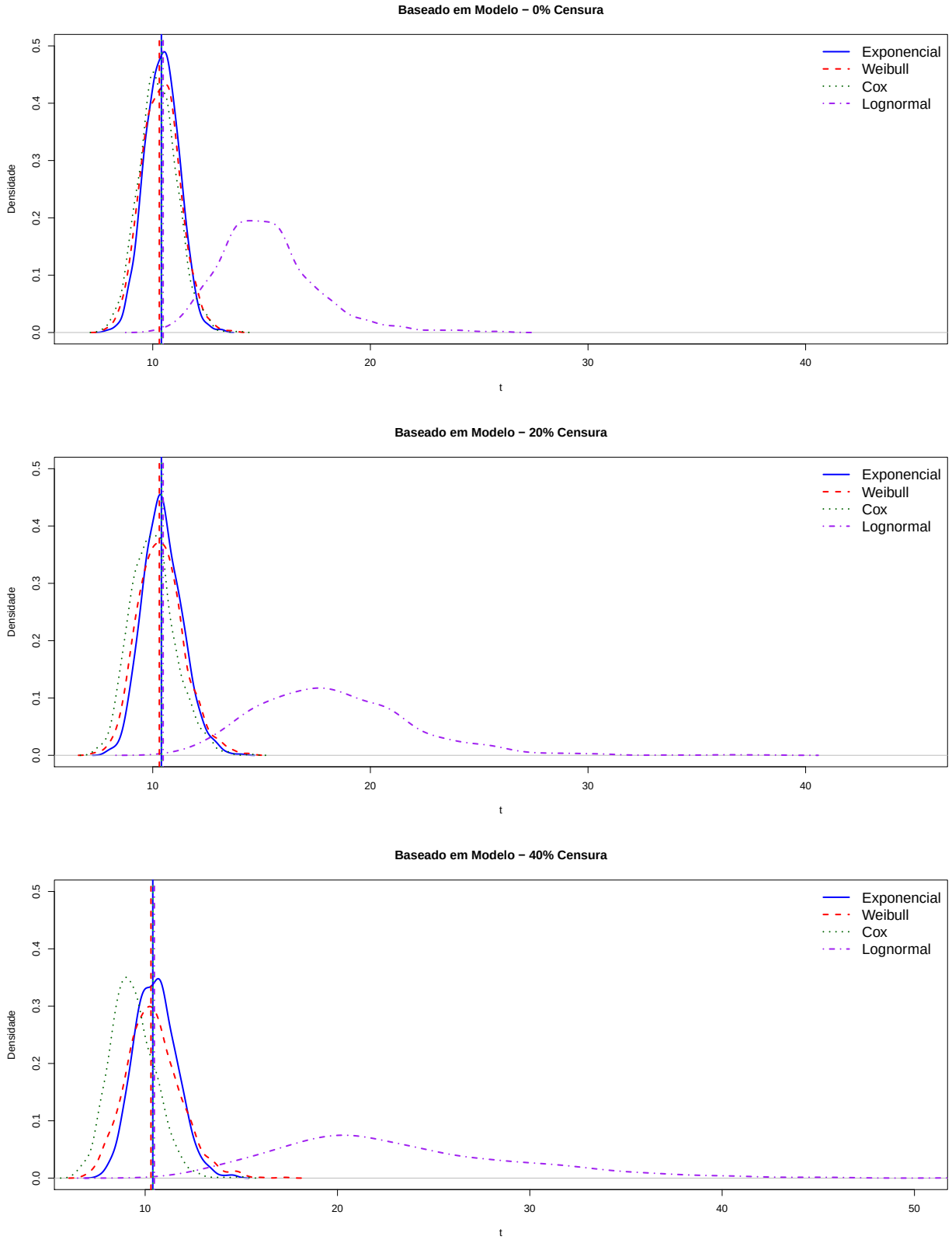


Figura A.6: Resultados da simulação exponencial considerando 100 indivíduos em cada grupo para a variância do estimador baseado em modelo

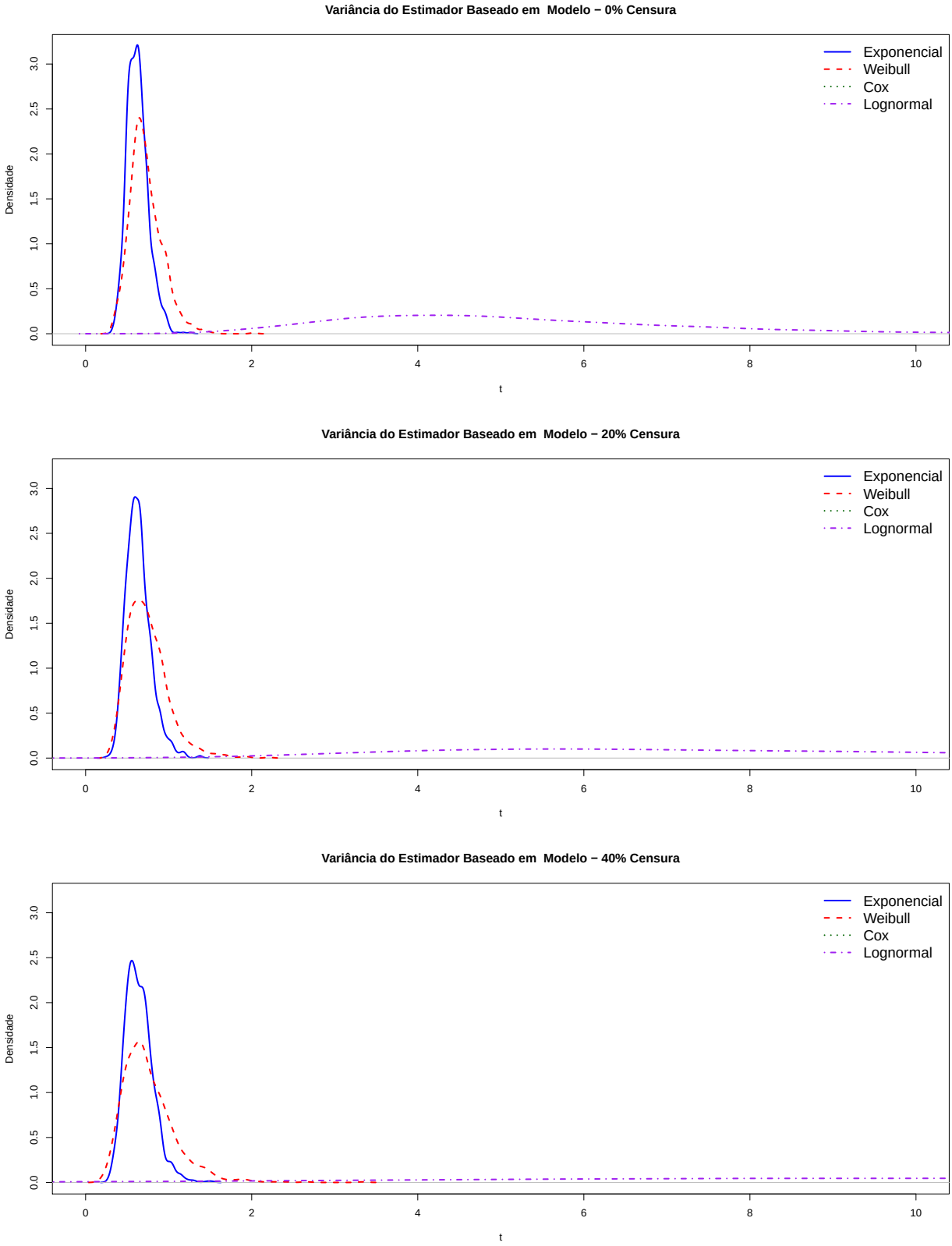


Figura A.7: Resultados da simulação exponencial considerando 250 indivíduos em cada grupo para o estimador perda aparente

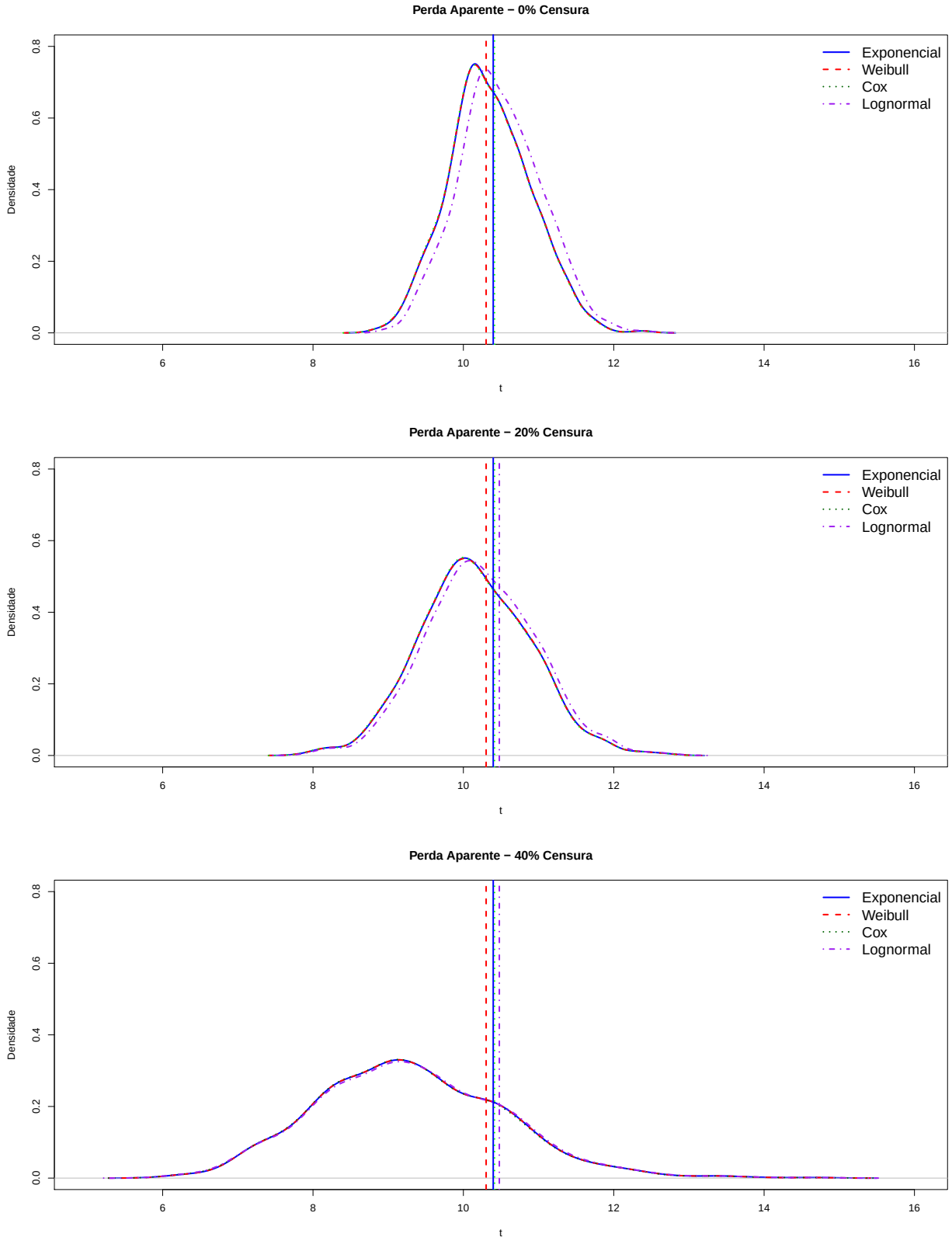


Figura A.8: Resultados da simulação exponencial considerando 250 indivíduos em cada grupo para a variância estimador perda aparente

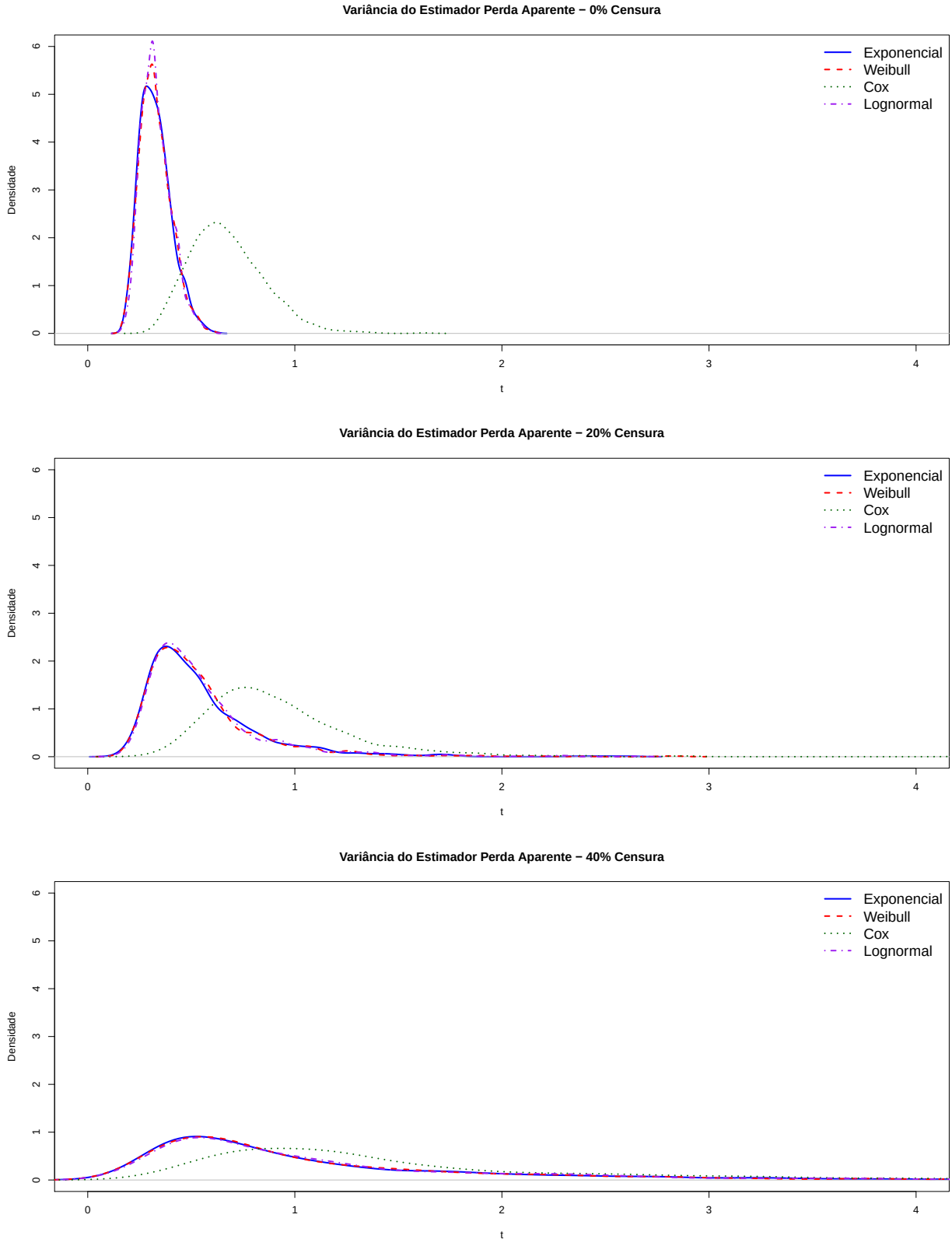


Figura A.9: Resultados da simulação exponencial considerando 250 indivíduos em cada grupo para o estimador validação cruzada

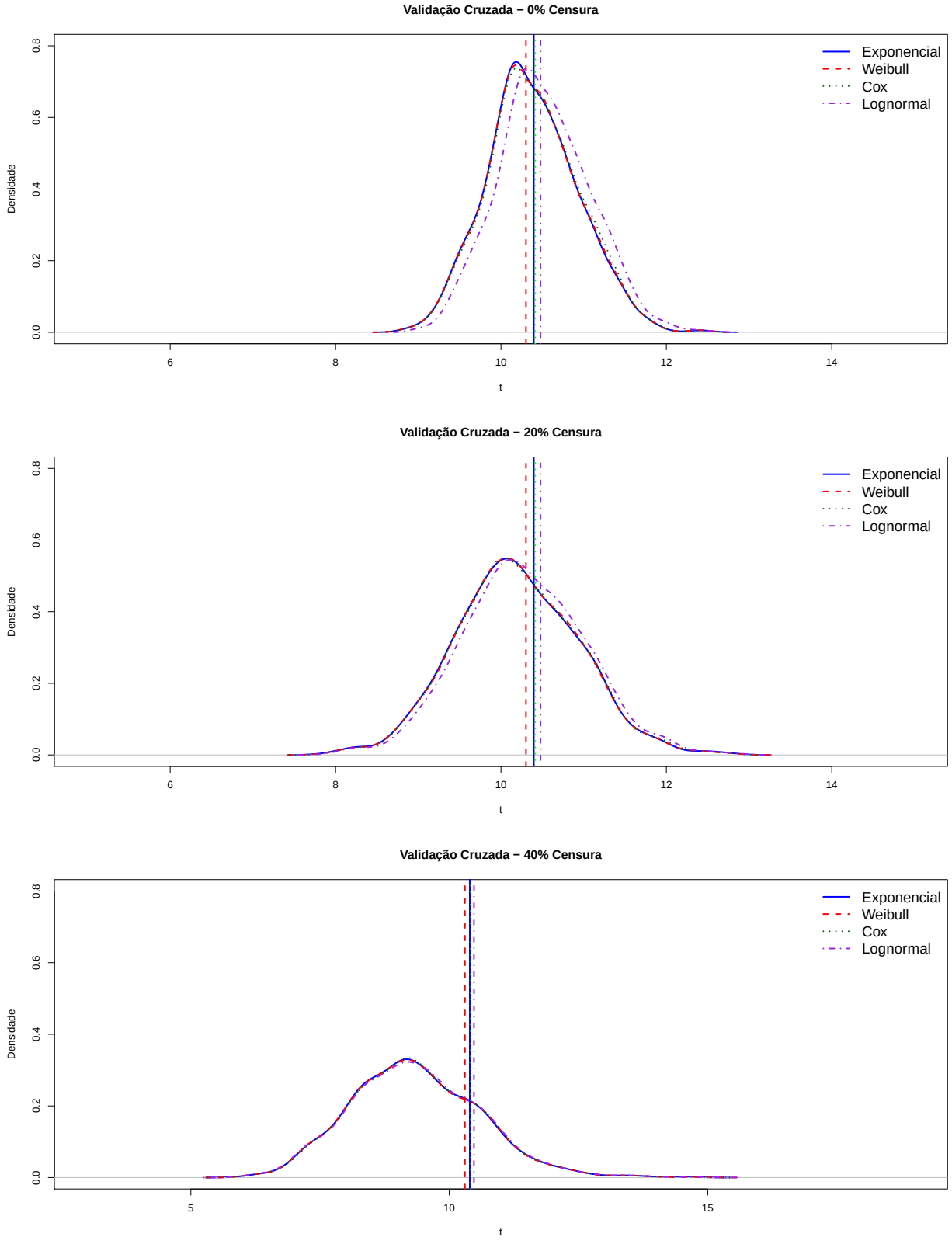


Figura A.10: Resultados da simulação exponencial considerando 250 indivíduos em cada grupo para a variância do estimador validação cruzada

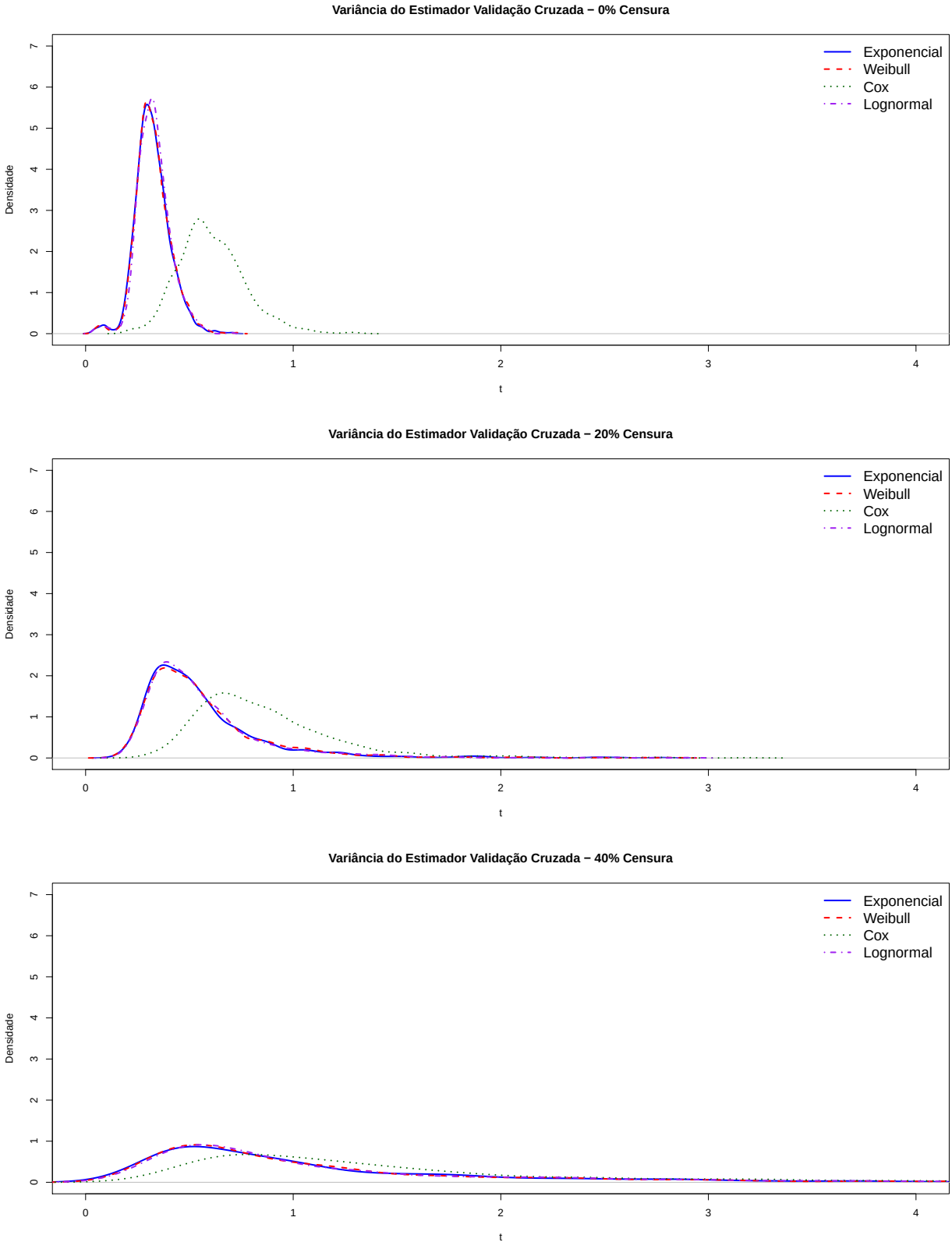


Figura A.11: Resultados da simulação exponencial considerando 250 indivíduos em cada grupo para o estimador baseado em modelo

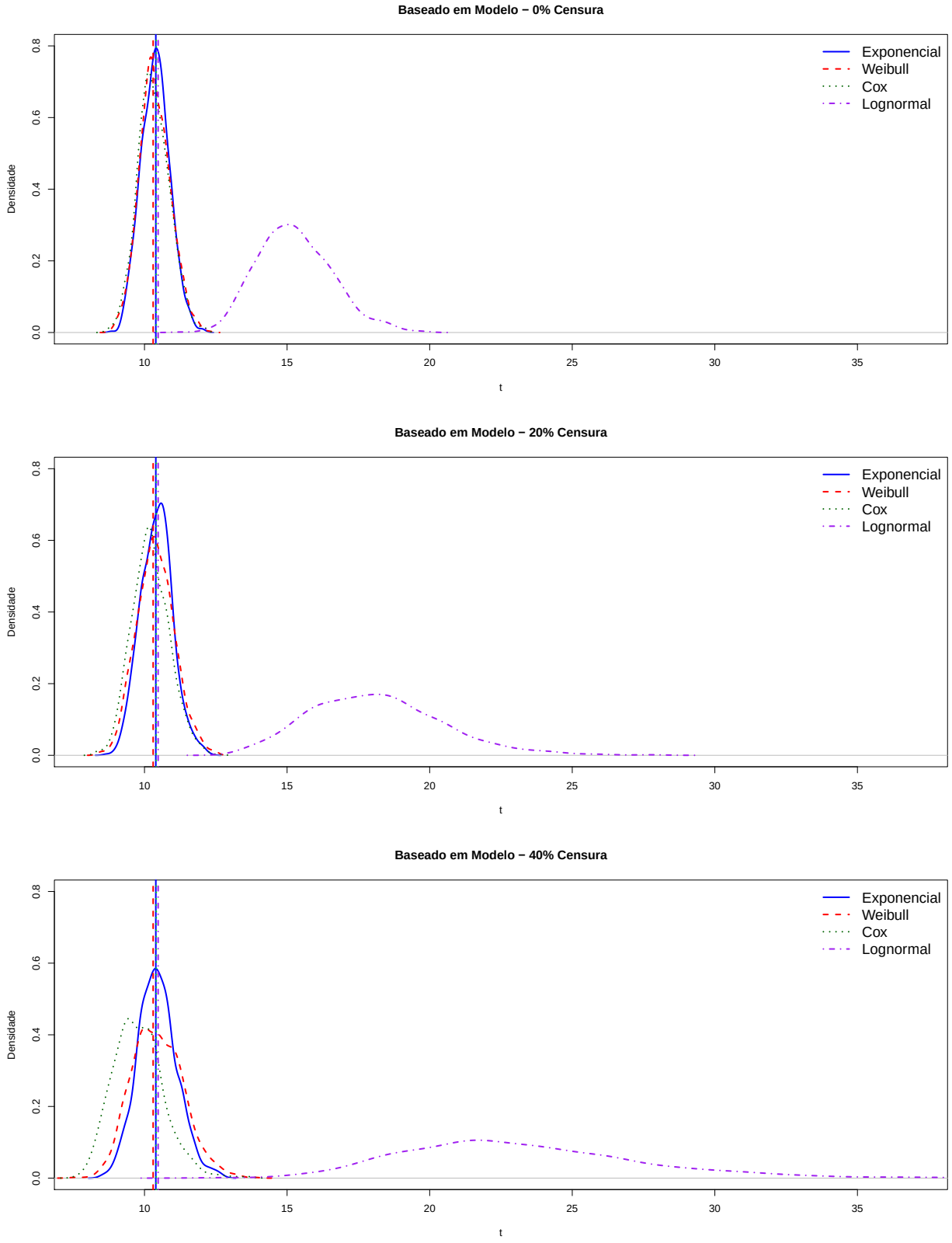


Figura A.12: Resultados da simulação exponencial considerando 250 indivíduos em cada grupo para a variância do estimador baseado em modelo

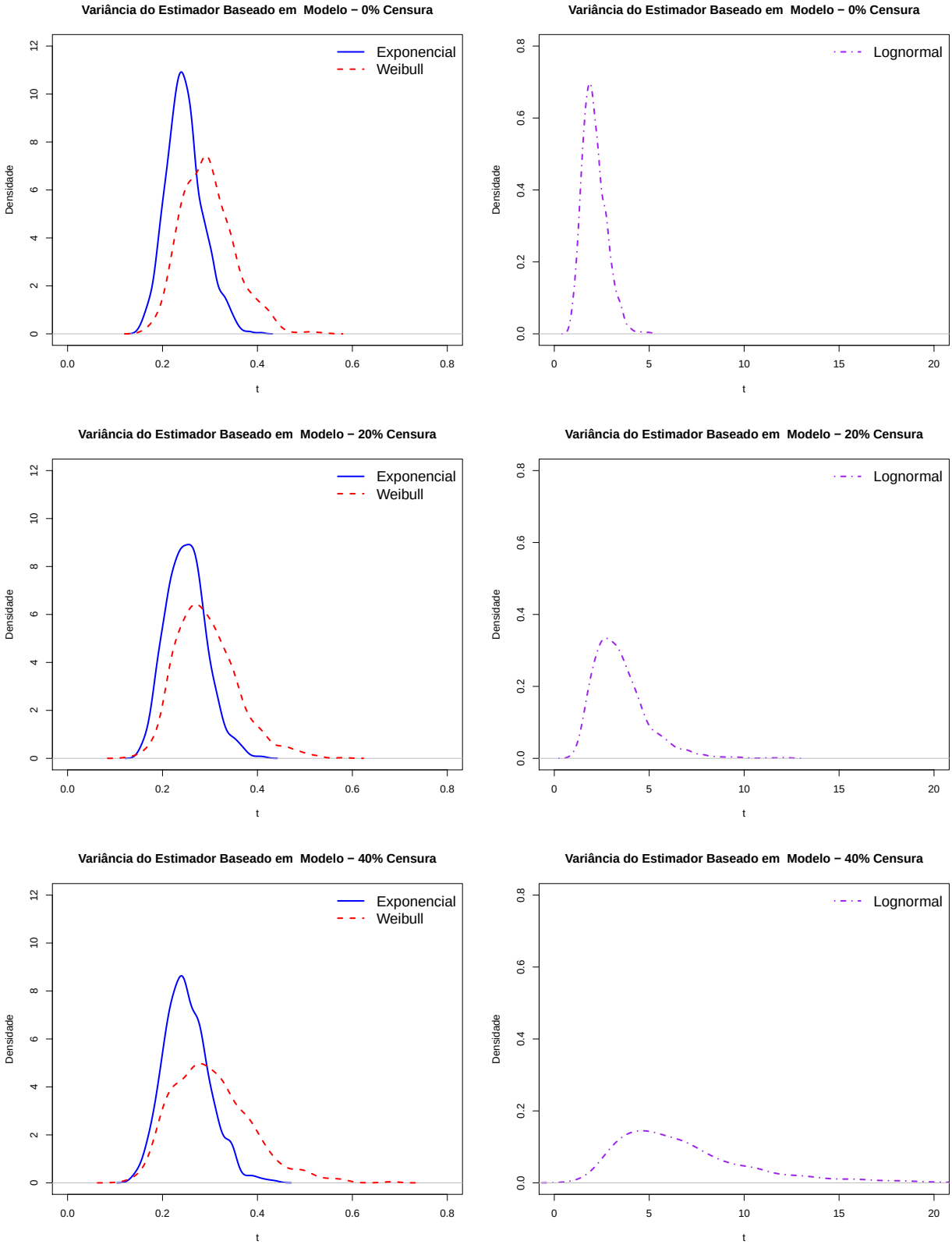


Figura A.13: Resultados da simulação lognormal considerando 100 indivíduos em cada grupo para o estimador perda aparente

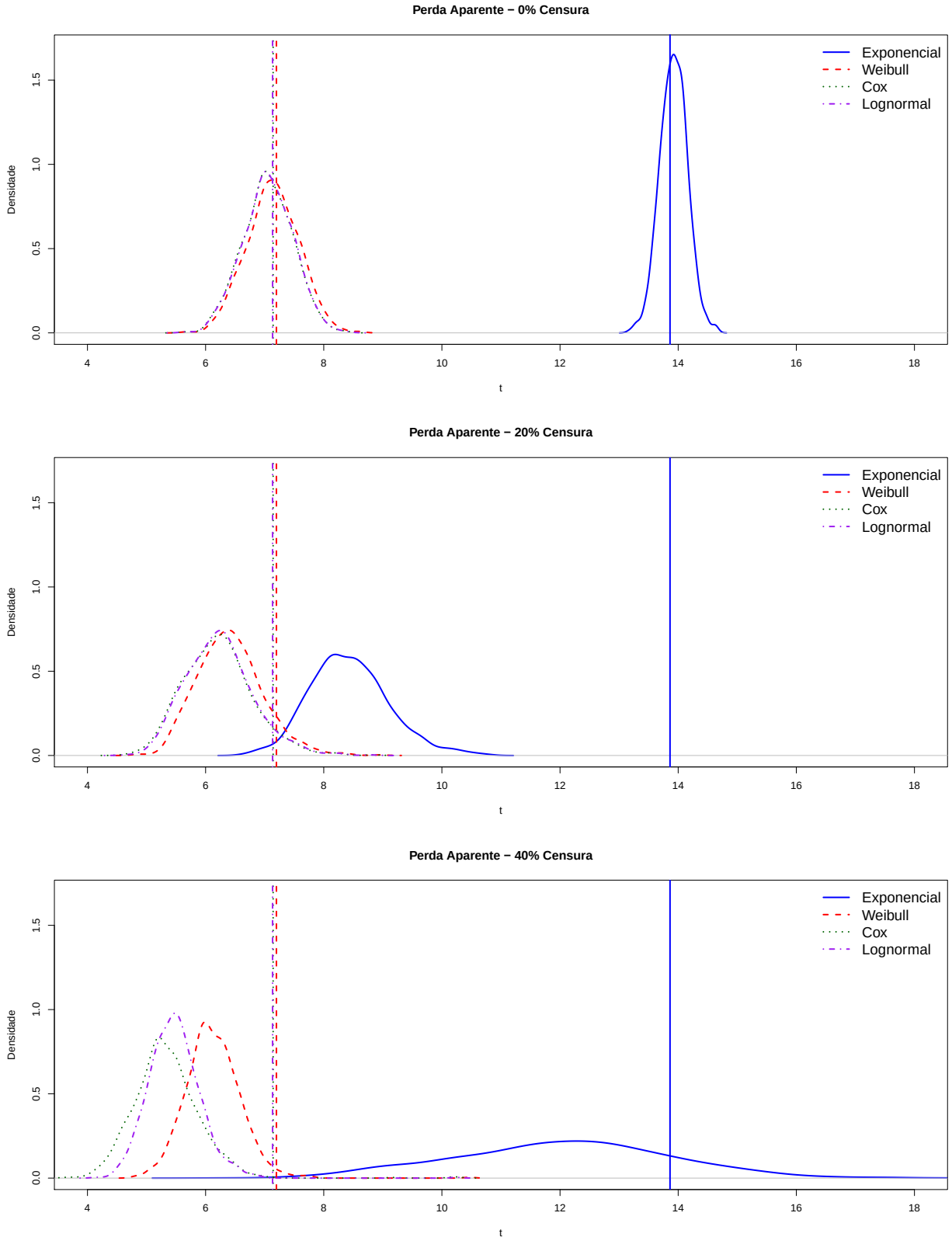


Figura A.14: Resultados da simulação lognormal considerando 100 indivíduos em cada grupo para a variância do estimador perda aparente

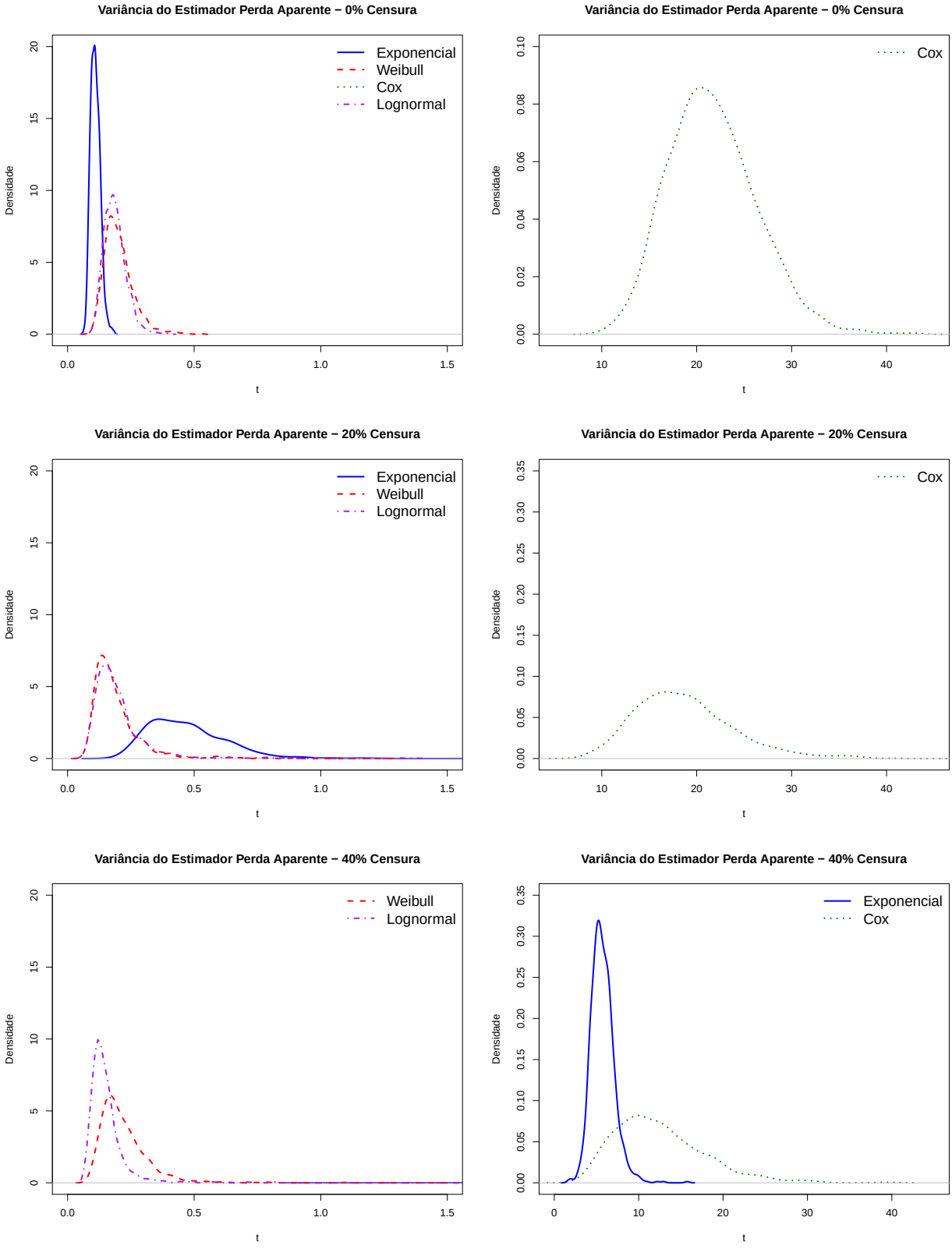


Figura A.15: Resultados da simulação lognormal considerando 100 indivíduos em cada grupo para o estimador validação cruzada

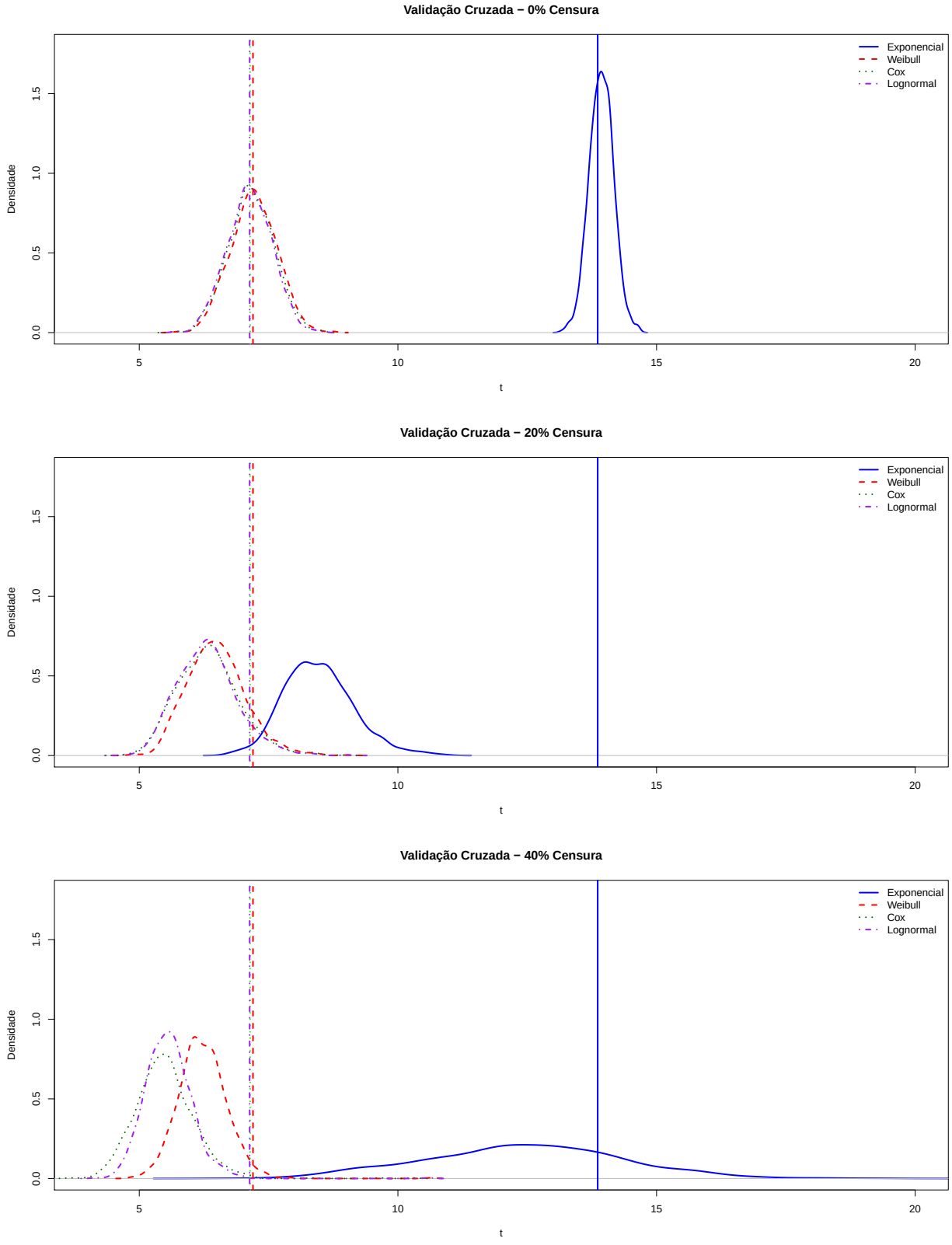


Figura A.16: Resultados da simulação lognormal considerando 100 indivíduos em cada grupo para a variância do estimador validação cruzada

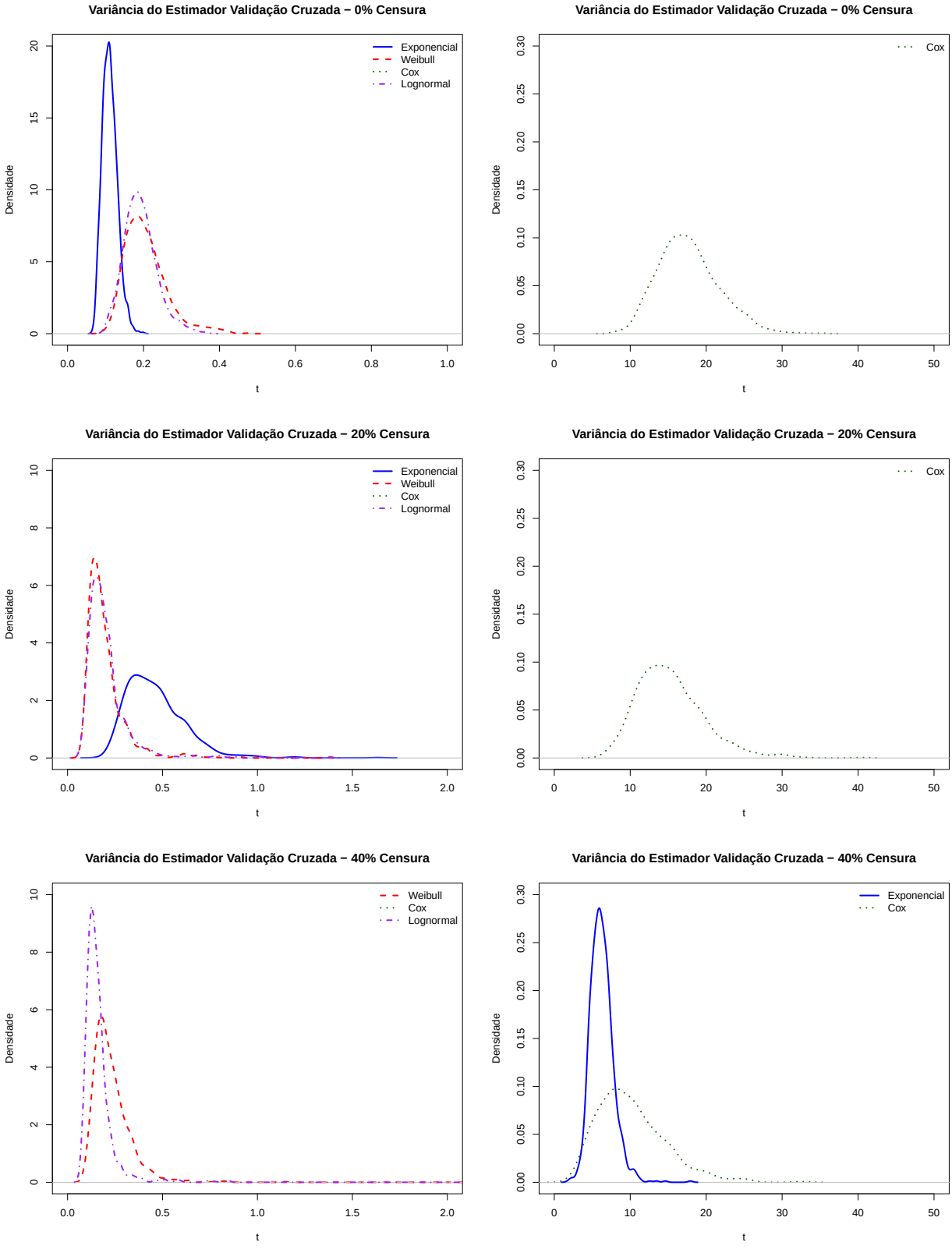


Figura A.17: Resultados da simulação lognormal considerando 100 indivíduos em cada grupo para o estimador baseado em modelo

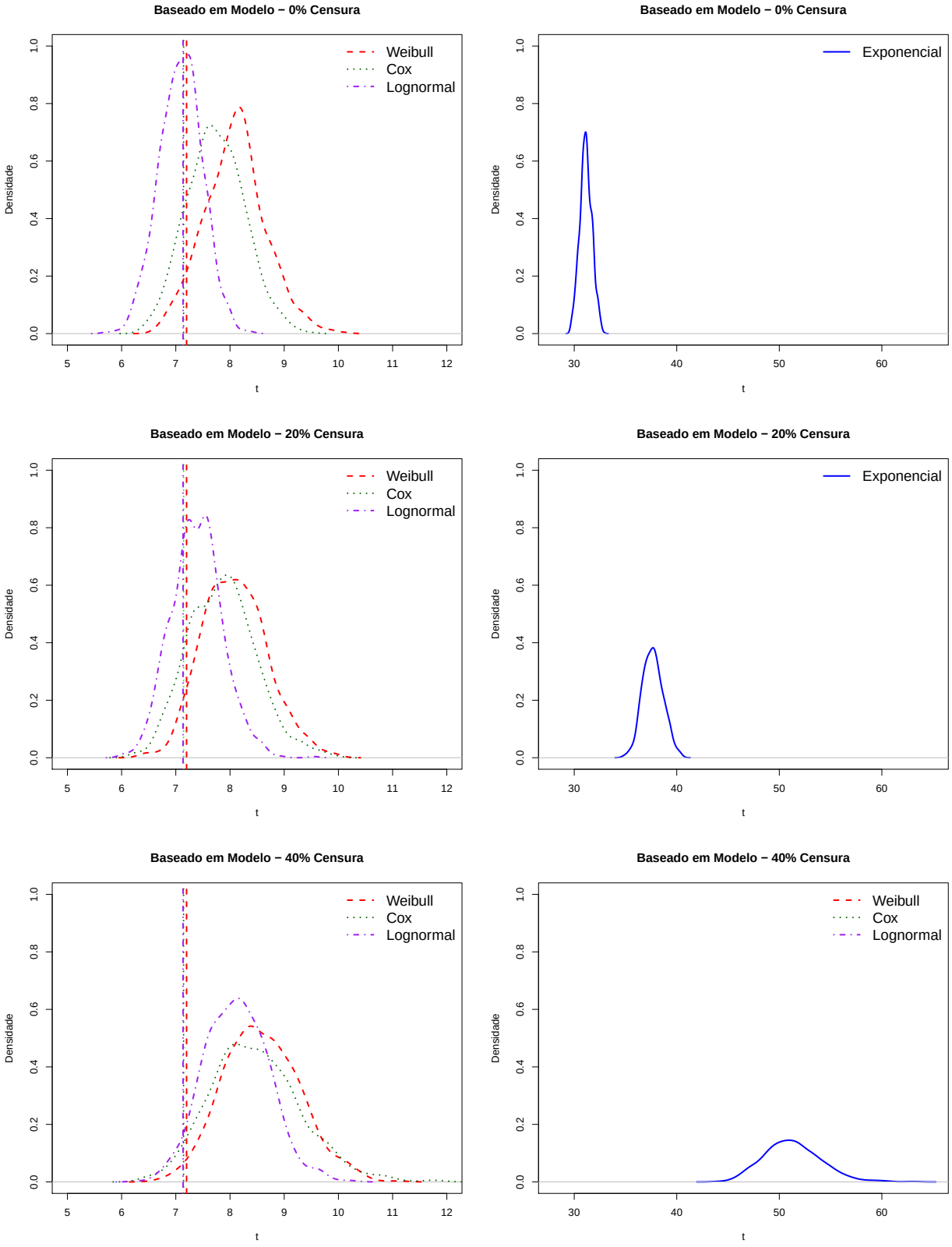


Figura A.18: Resultados da simulação lognormal considerando 100 indivíduos em cada grupo para a variância do estimador baseado em modelo

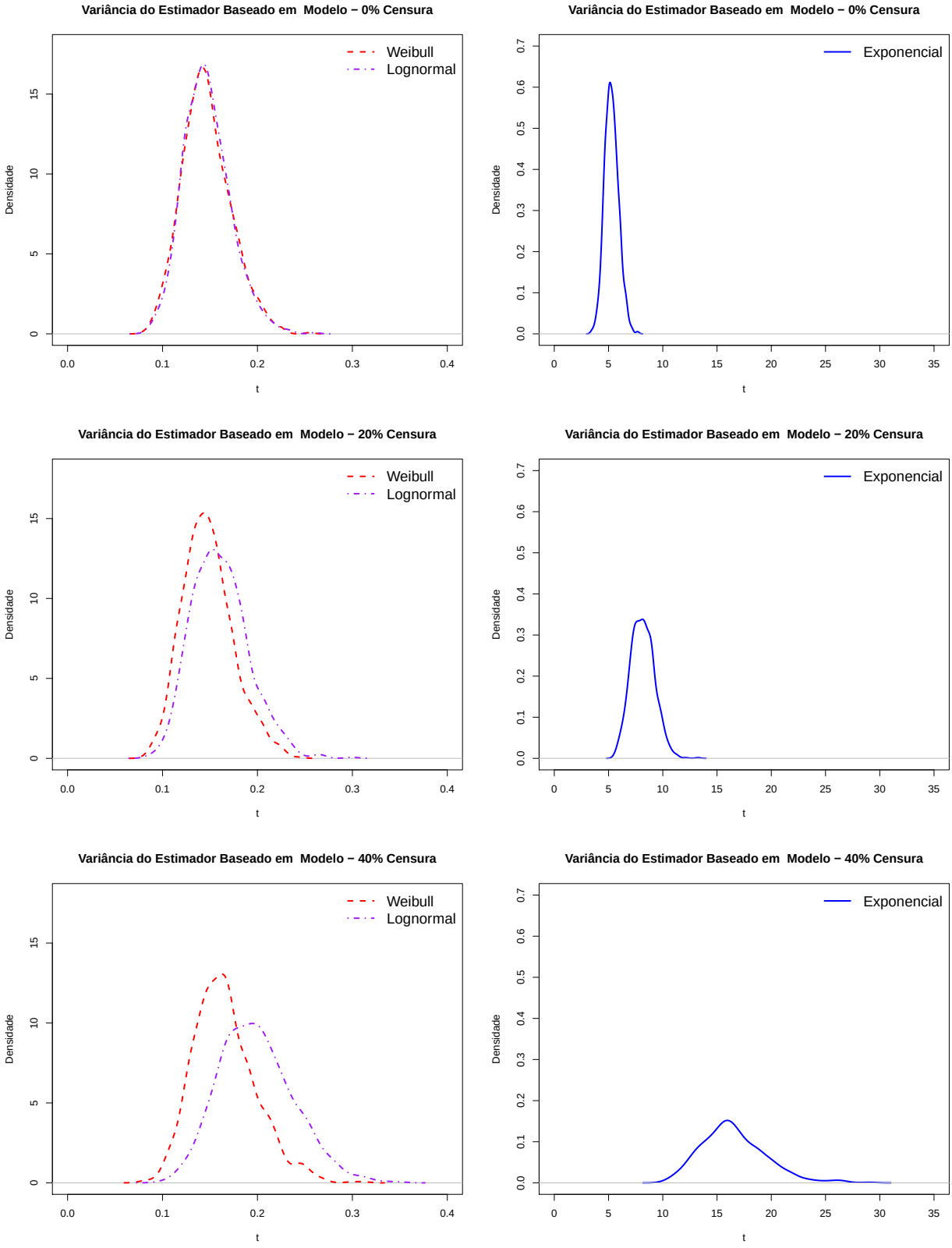


Figura A.19: Resultados da simulação lognormal considerando 250 indivíduos em cada grupo para o estimador perda aparente

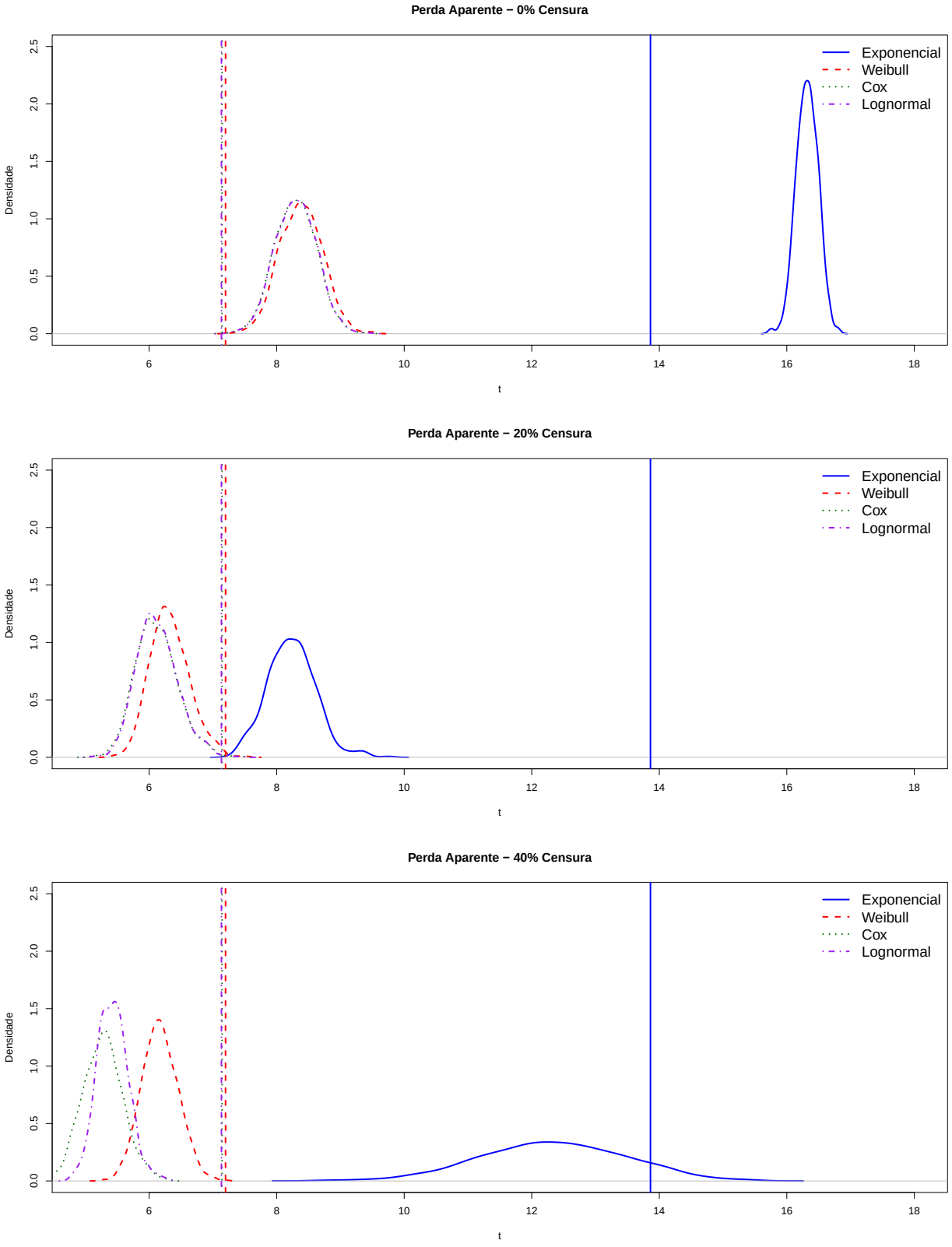


Figura A.20: Resultados da simulação lognormal considerando 250 indivíduos em cada grupo para a variância do estimador perda aparente

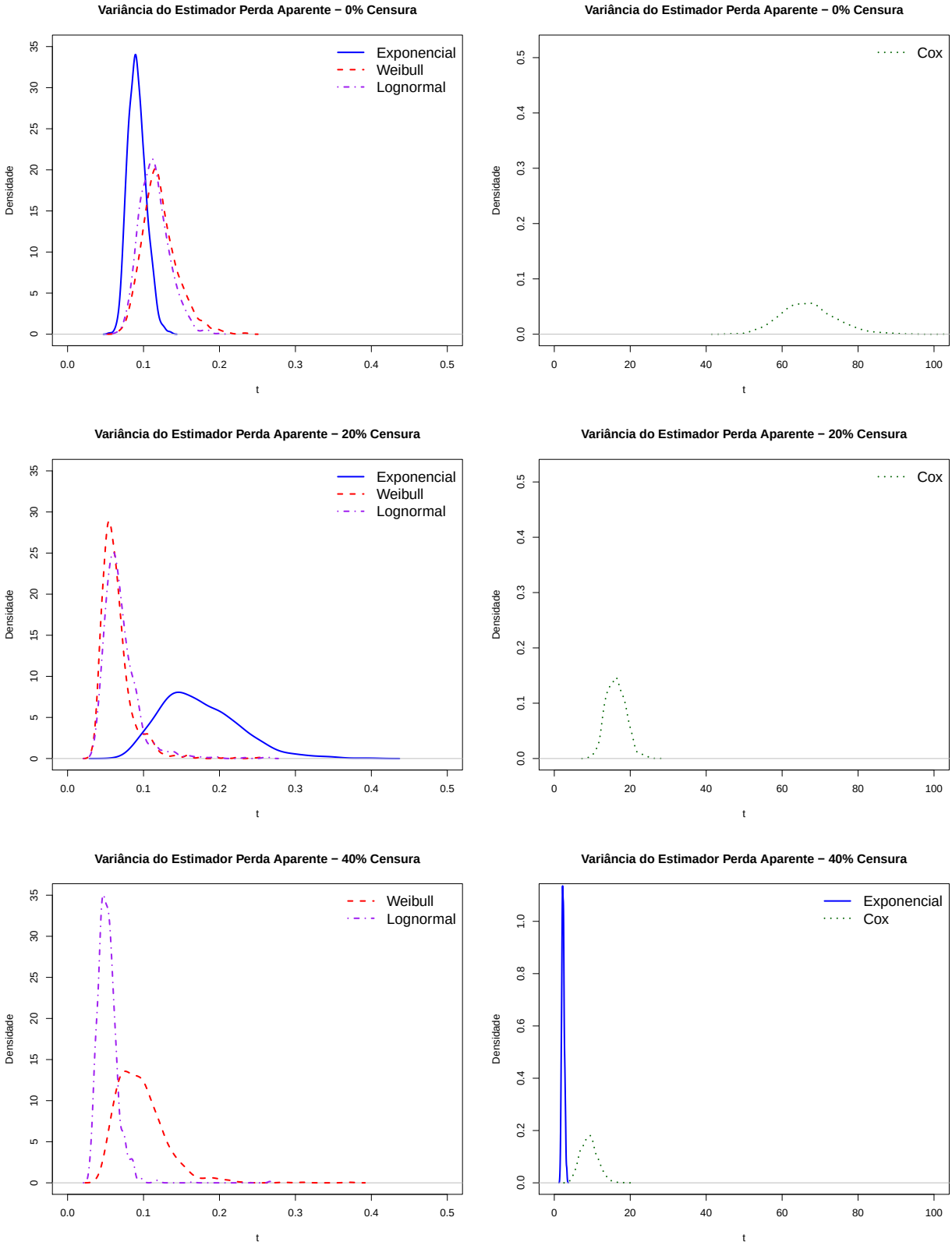


Figura A.21: Resultados da simulação lognormal considerando 250 indivíduos em cada grupo para o estimador validação cruzada

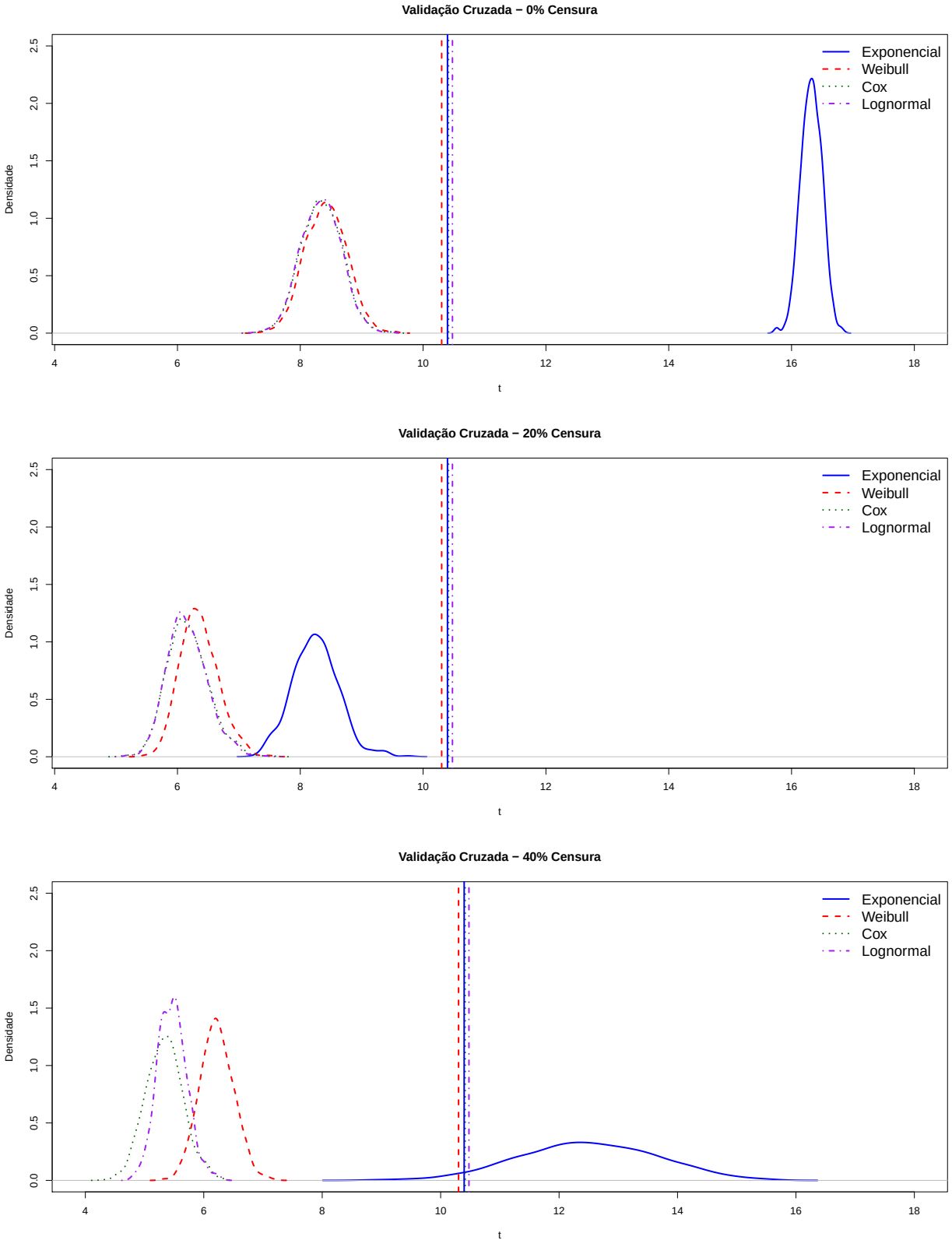


Figura A.22: Resultados da simulação lognormal considerando 250 indivíduos em cada grupo para a variância do estimador validação cruzada

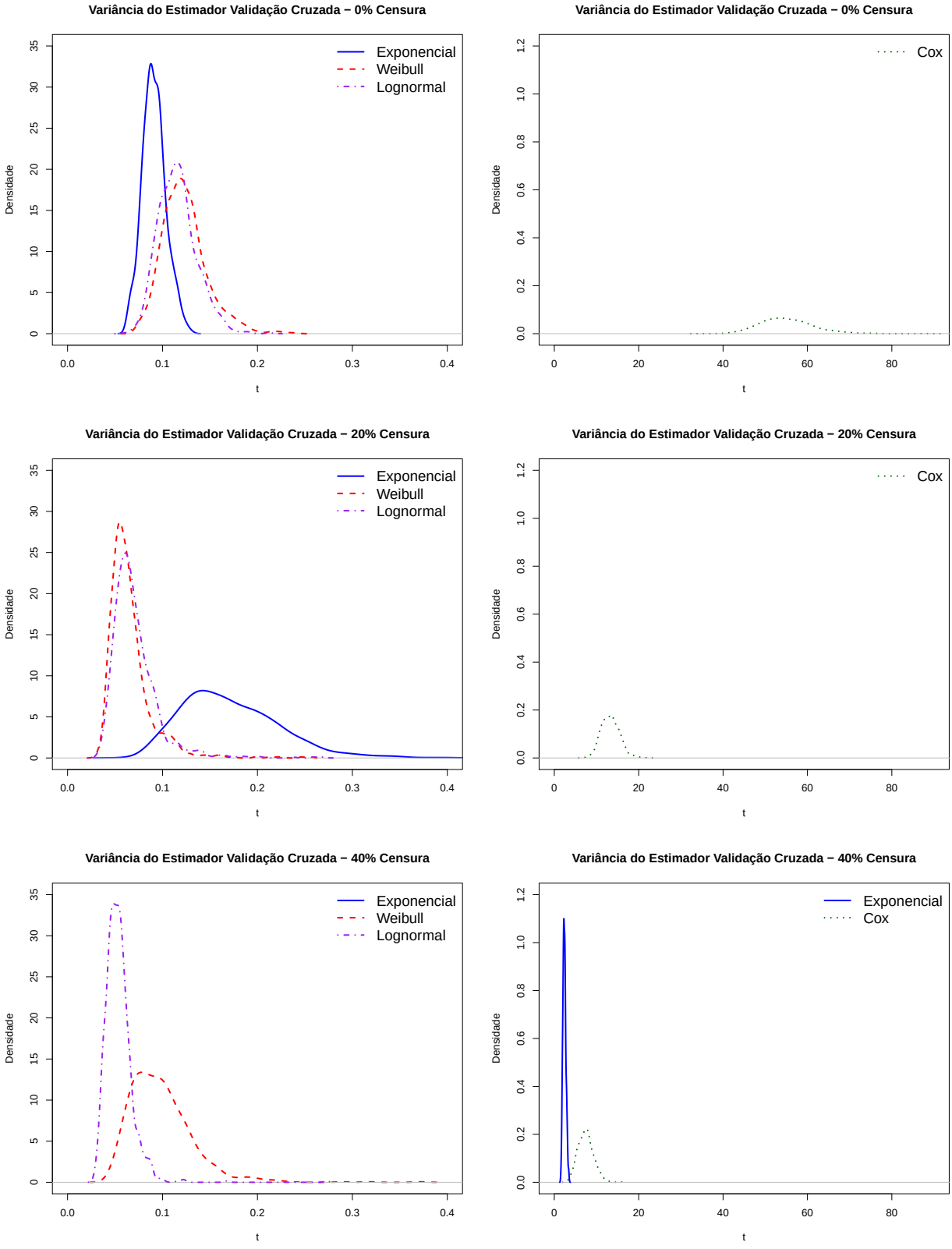


Figura A.23: Resultados da simulação lognormal considerando 250 indivíduos em cada grupo para o estimador baseado em modelo

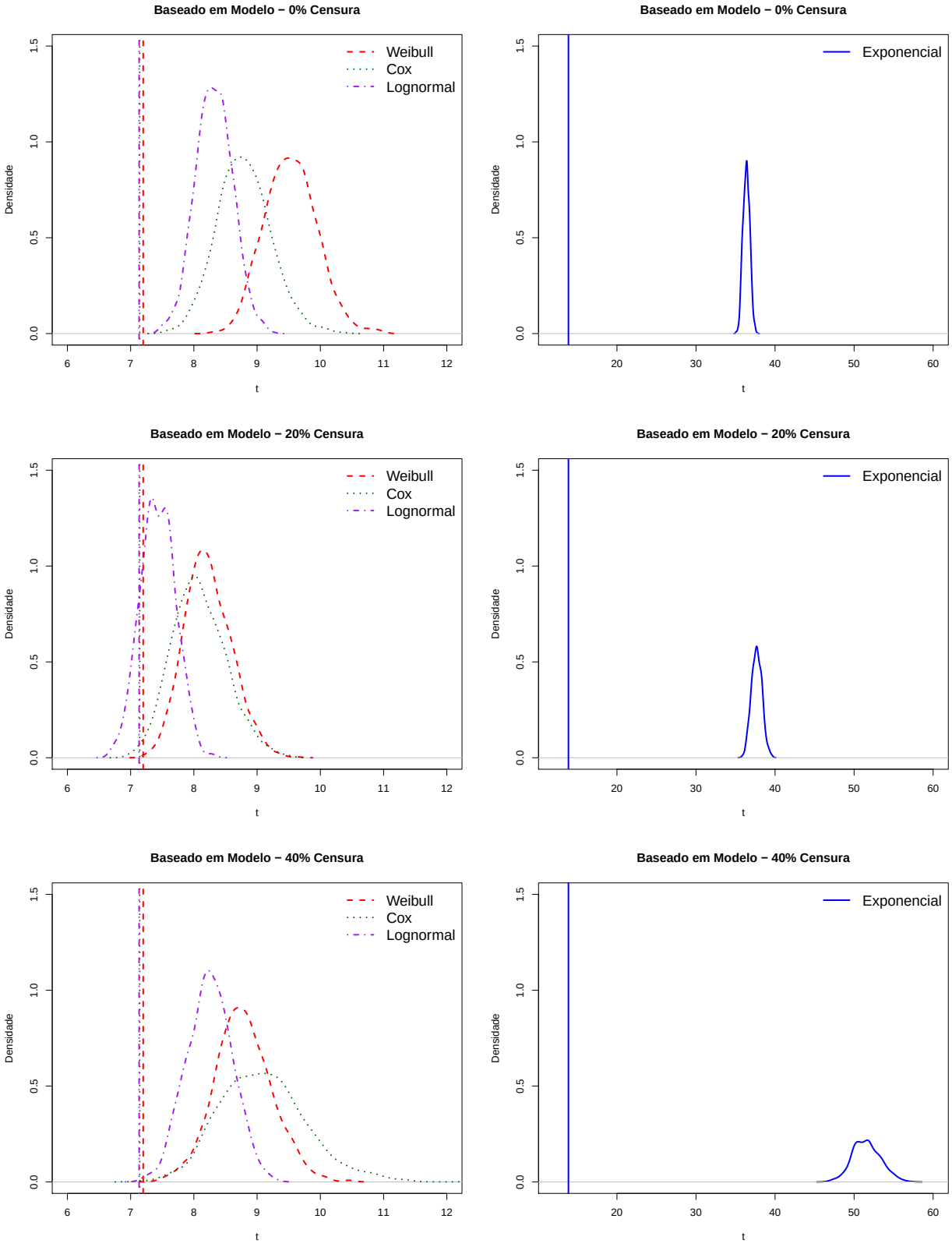
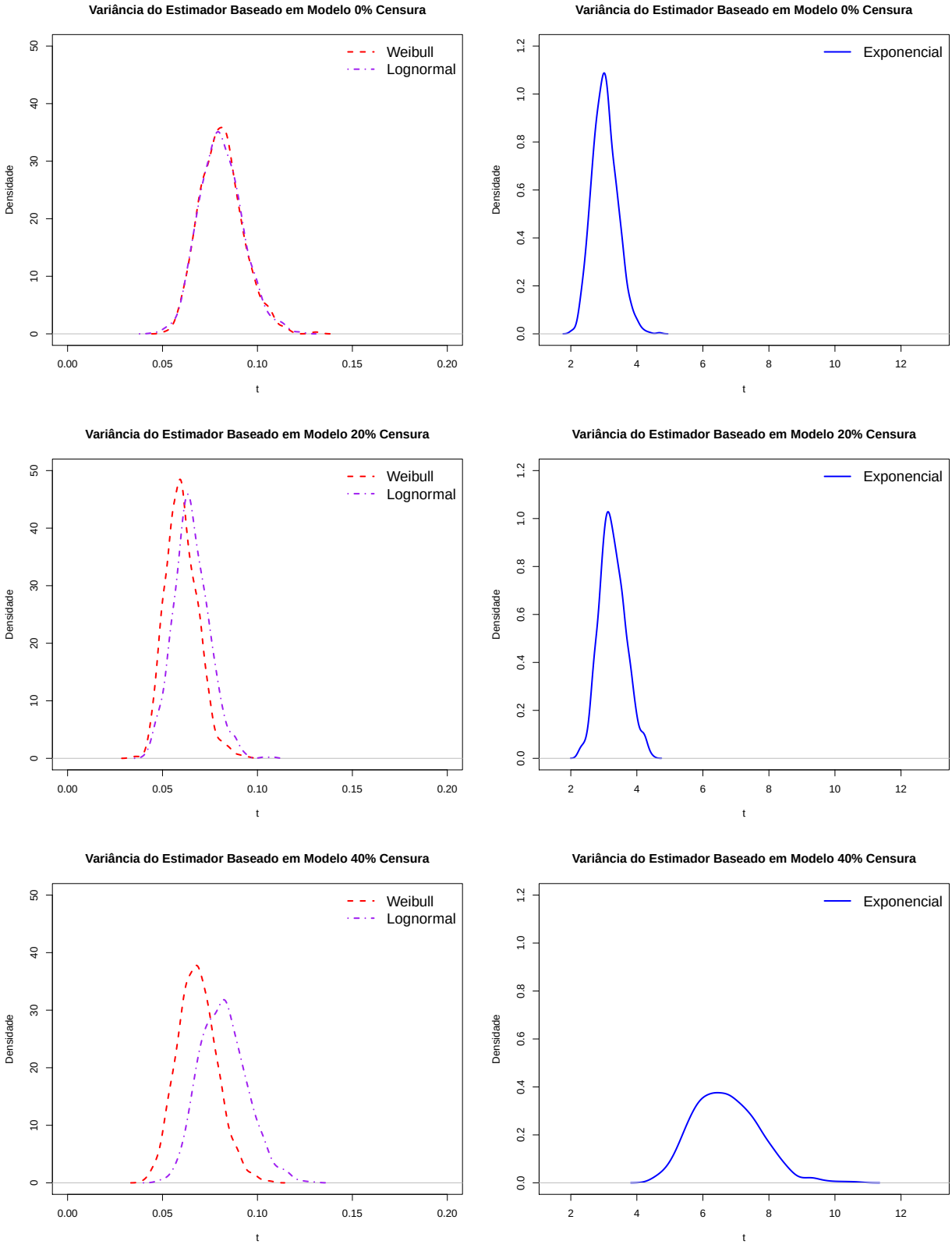


Figura A.24: Resultados da simulação lognormal considerando 250 indivíduos em cada grupo para a variância do estimador baseado em modelo



A.2 Capítulo 4 - Simulação do status de sobrevivência

Tabelas

Tabela A.1: Resultados da simulação utilizando a abordagem apresentada na seção 3.7 considerando o preditor linear do modelo Weibull com 20% de censura

Corte	N	$\widehat{SE}$	SE^v	$\widehat{DP}(SE)$	$DP(\widehat{SE})$	$\widehat{ES}$	ES^v	$\widehat{DP}(ES)$	$DP(\widehat{ES})$
-0,85	200	0,7856	0,7839	0,05675	0,03866	0,7998	0,7957	0,04197	0,03422
	500	0,7849	0,7852	0,03547	0,02445	0,7983	0,7950	0,02666	0,02114
	1000	0,7839	0,7843	0,02523	0,01647	0,7976	0,7955	0,01890	0,01485
	5000	0,7846	0,7845	0,01125	0,00799	0.7983	0,7953	0,00844	0,00655
-0,90	200	0,7972	0,7958	0,05467	0,03507	0,7862	0,7819	0,04296	0,03862
	500	0,7961	0,7960	0,03557	0,02200	0,7863	0,7831	0,02723	0,02333
	1000	0,7949	0,7947	0,02510	0,01519	0,7861	0,7833	0,01929	0,01625
	5000	0,7948	0,7948	0,01120	0,00731	0,7867	0,7838	0,00861	0,00732

Tabela A.2: Resultados da simulação utilizando a abordagem apresentada na seção 3.7 considerando o preditor linear do modelo Weibull com 40% de censura

Corte	N	$\widehat{SE}$	SE^v	$\widehat{DP}(SE)$	$DP(\widehat{SE})$	$\widehat{ES}$	ES^v	$\widehat{DP}(ES)$	$DP(\widehat{ES})$
-0,85	200	0,7790	0,7762	0,07622	0,04727	0,8194	0,8019	0,04322	0,03482
	500	0,7770	0,7747	0,04786	0,03890	0,8170	0,8050	0,02743	0,02296
	1000	0,7766	0,7755	0,03371	0,02072	0,8179	0,8031	0,01941	0,01518
	5000	0,7767	0,7757	0,01502	0,00942	0,8181	0,8035	0,00866	0,00687
-0,90	200	0,7915	0,7889	0,07645	0,04234	0,8073	0,7899	0,04431	0,03648
	500	0,7895	0,7876	0,04799	0,02690	0,8058	0,7934	0,02807	0,02425
	1000	0,7886	0,7877	0,03379	0,01818	0,8069	0,7917	0,01985	0,01585
	5000	0,7888	0,7879	0,01505	0,00828	0,8073	0,7922	0,00886	0,00712

Tabela A.3: Resultados da simulação utilizando a abordagem apresentada na seção 3.7 considerando o preditor linear do modelo de riscos proporcionais de Cox

Censura	N	$\widehat{SE}$	SE^v	$\widehat{DP}(SE)$	$DP(\widehat{SE})$	$\widehat{ES}$	ES^v	$\widehat{DP}(ES)$	$DP(\widehat{ES})$
20%	200	0,7423	0,7429	0,05585	0,15448	0,7839	0,7826	0,04051	0,12589
	500	0,7620	0,7622	0,03570	0,08890	0,8001	0,7989	0,02612	0,06075
	1000	0,7747	0,7757	0,02528	0,04007	0,8006	0,7990	0,01872	0,02990
	5000	0,7811	0,7811	0,01127	0,01175	0,8009	0,7979	0,00839	0,01080
40%	200	0,7605	0,7580	0,07501	0,14471	0,7782	0,7642	0,04329	0,14312
	500	0,7714	0,7707	0,04764	0,08845	0,8005	0,7900	0,02782	0,06612
	1000	0,7791	0,7788	0,03370	0,04367	0,8082	0,7933	0,01969	0,03548
	5000	0,7853	0,7845	0,01504	0,01112	0,8099	0,7952	0,00881	0,01091

Tabela A.4: Resultados da simulação utilizando a abordagem apresentada na seção 3.7 considerando o preditor linear estimado e o verdadeiro valor do preditor linear sob o modelo Weibull com 20% de censura

	Marcador Estimado				Marcador Fixado			
N	$\widehat{SE}$	SE^v	$\widehat{DP}(SE)$	$DP(\widehat{SE})$	$\widehat{SE}$	SE^v	$\widehat{DP}(SE)$	$DP(\widehat{SE})$
200	0,7862	0,7827	0,05668	0,03993	0,7900	0,7891	0,05661	0,03763
500	0,7857	0,7838	0,03570	0,02463	0,7893	0,7883	0,03564	0,02338
1000	0,7849	0,7850	0,02521	0,01667	0,7888	0,7889	0,02517	0,01553
5000	0,7878	0,7846	0,01124	0,00737	0,7890	0,7889	0,01122	0,00698

Tabela A.5: Resultados da simulação utilizando a abordagem apresentada na seção 3.7 considerando o preditor linear estimado e o verdadeiro valor do preditor linear sob o modelo Weibull com 20% de censura

	Marcador Estimado				Marcador Fixado			
N	$\widehat{ES}$	ES^v	$\widehat{DP}(ES)$	$DP(\widehat{ES})$	$\widehat{ES}$	ES^v	$\widehat{DP}(ES)$	$DP(\widehat{ES})$
200	0,7995	0,7961	0,04197	0,03408	0,7942	0,7925	0,04235	0,03584
500	0,7990	0,7953	0,02665	0,02060	0,7948	0,7917	0,02684	0,02277
1000	0,7983	0,7953	0,01887	0,01530	0,7944	0,7915	0,01900	0,01672
5000	0,7986	0,7958	0,00843	0,00647	0,7943	0,7916	0,00850	0,00718

Tabela A.6: Resultados da simulação utilizando a abordagem apresentada na seção 3.7 considerando o preditor linear estimado e o verdadeiro valor do preditor linear sob o modelo de riscos proporcionais de Cox com 20% de censura

	Marcador Estimado				Marcador Fixado			
N	$\widehat{SE}$	SE^v	$\widehat{DP}(SE)$	$DP(\widehat{SE})$	$\widehat{SE}$	SE^v	$\widehat{DP}(SE)$	$DP(\widehat{SE})$
200	0,7515	0,7517	0,05600	0,14034	0,7791	0,7780	0,05688	0,03897
500	0,7663	0,7659	0,03568	0,08219	0,7787	0,7776	0,03581	0,02377
1000	0,7744	0,7747	0,02527	0,04138	0,7783	0,7786	0,02528	0,01604
5000	0,7816	0,7814	0,01126	0,01108	0,7786	0,7784	0,01127	0,00712

Tabela A.7: Resultados da simulação utilizando a abordagem apresentada na seção 3.7 considerando o preditor linear estimado e o verdadeiro valor do preditor linear sob o modelo de riscos proporcionais de Cox com 20% de censura

	Marcador Estimado				Marcador Fixado			
N	$\widehat{ES}$	ES^v	$\widehat{DP}(ES)$	$DP(\widehat{ES})$	$\widehat{ES}$	ES^v	$\widehat{DP}(ES)$	$DP(\widehat{ES})$
200	0,7797	0,7794	0,04106	0,12241	0,8040	0,8023	0,04156	0,03576
500	0,7995	0,7971	0,02624	0,05626	0,8048	0,8015	0,02635	0,02237
1000	0,8014	0,7991	0,01868	0,03273	0,8043	0,8015	0,01865	0,01654
5000	0,8009	0,7984	0,00839	0,01046	0,8042	0,8016	0,00834	0,00701

A.3 Capítulo 5 - Aplicação

Gráficos

Figura A.25: Curva de sobrevivência por sexo

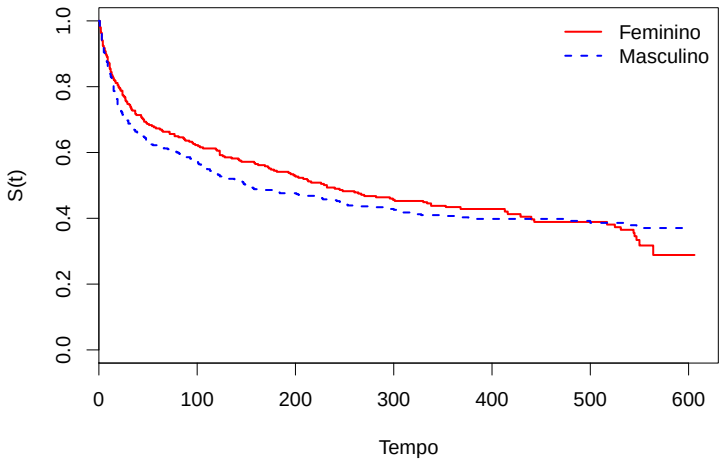


Figura A.26: Curva de sobrevivência por cirurgia

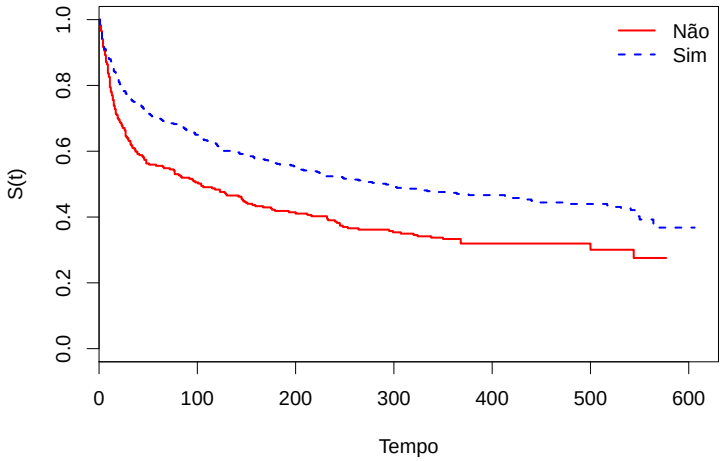


Figura A.27: Curva de sobrevivência por status do câncer

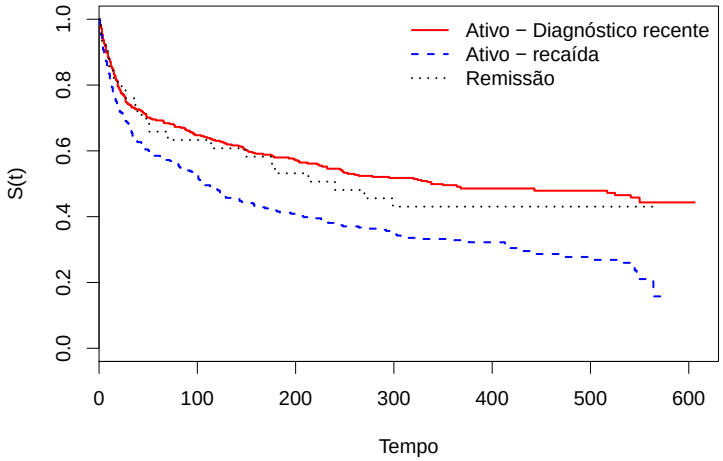


Figura A.28: Curva de sobrevivência por radioterapia

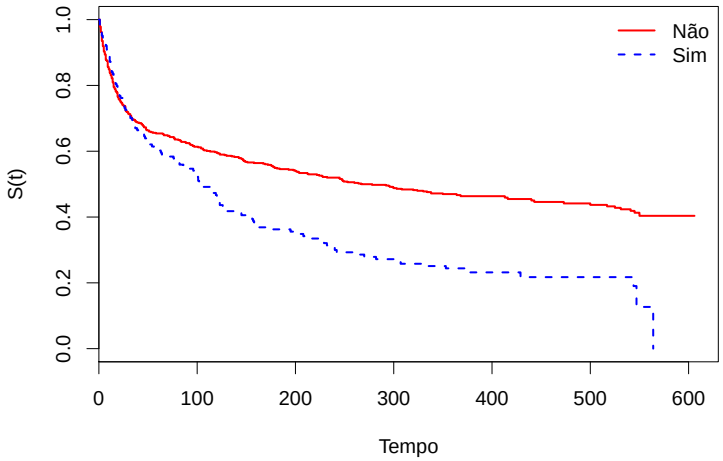


Figura A.29: Curva de sobrevivência por quimioterapia

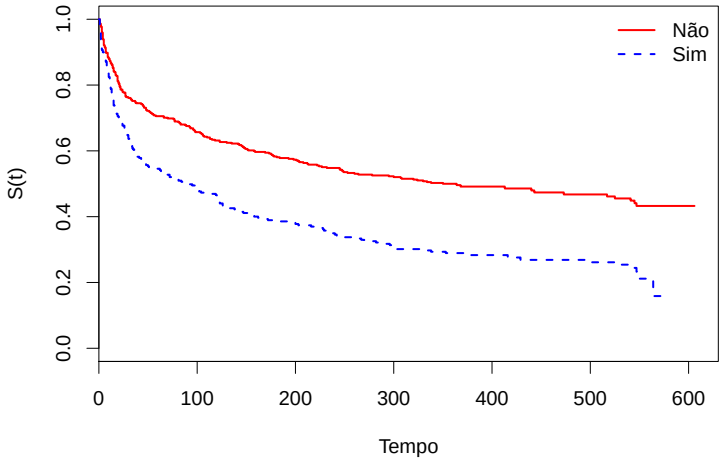


Figura A.30: Curva de sobrevivência por alcoolismo

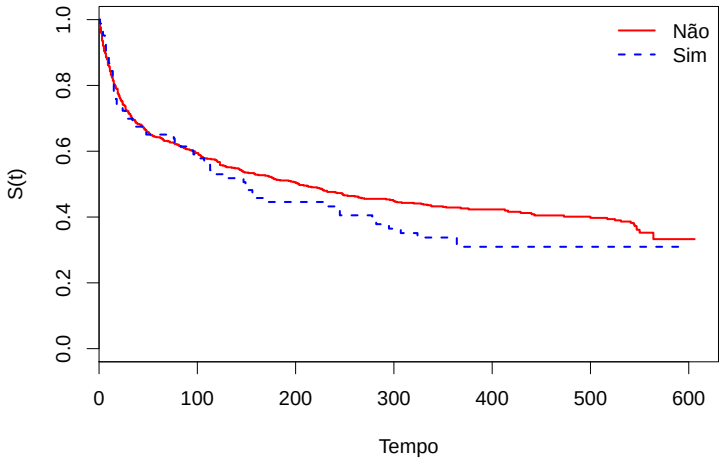


Figura A.31: Curva de sobrevivência por tabagismo

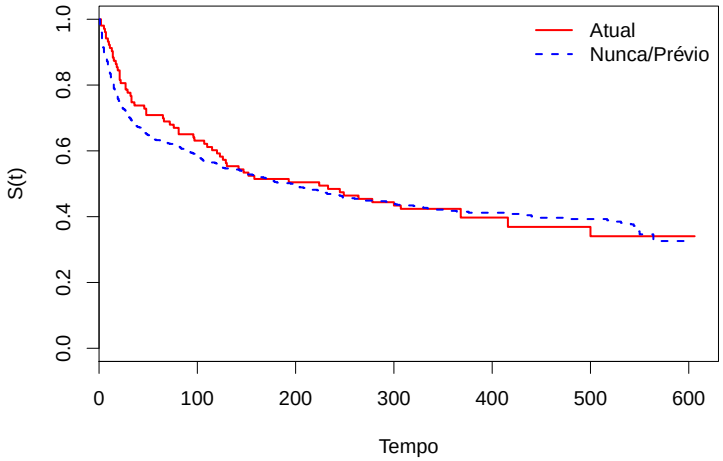


Figura A.32: Curva de sobrevivência por ventilação

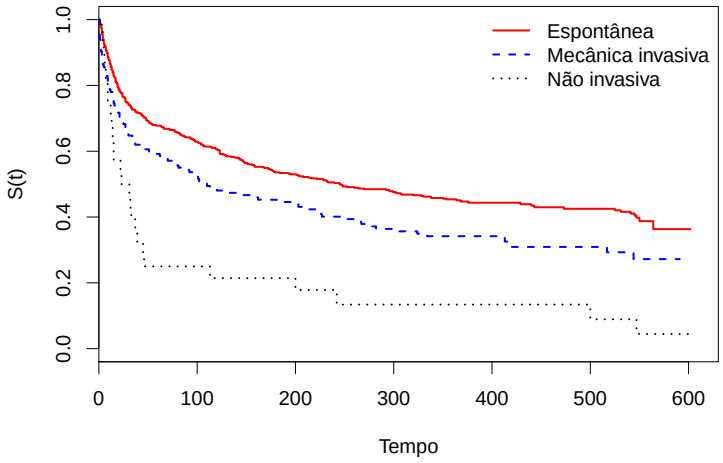


Figura A.33: Curva de sobrevivência por infecção respiratória na UTI

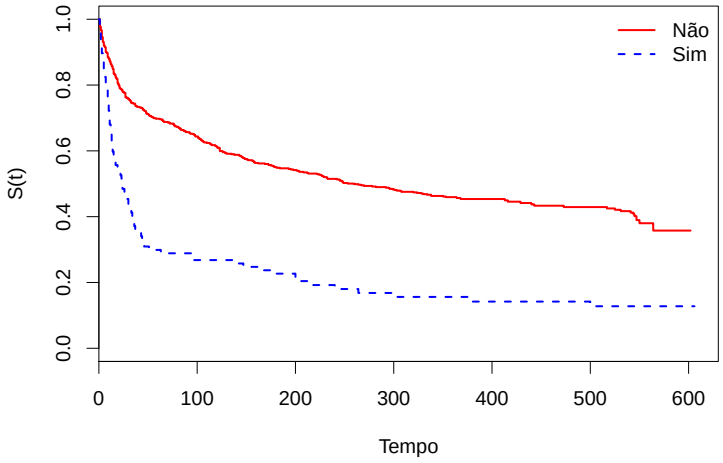
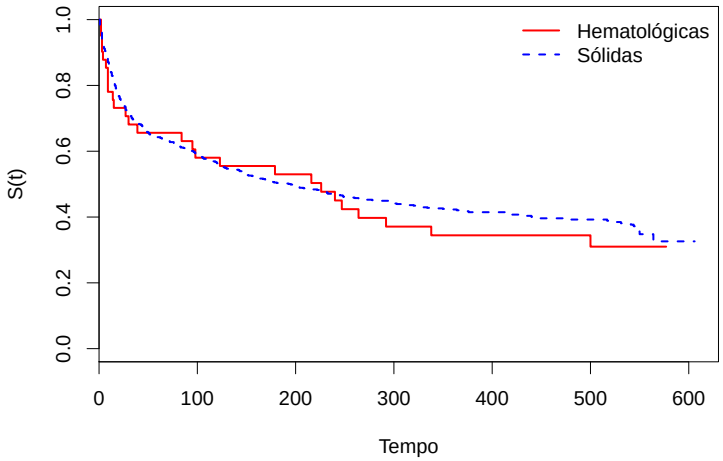


Figura A.34: Curva de sobrevivência por local agrupado



Tabelas

Tabela A.8: tabela de frequências

Variável	Categorias	Frequência	Porcentagem
Sexo	Masculino	418	58,0%
	Feminino	303	42,0%
Status câncer	Ativo - diagnostico recente	365	50,6%
	Ativo - recaída/progressão	314	43,6%
	Controlado - remissão	42	5,8%
Extensão do câncer	Leucemias	13	1,8%
	Limitado	236	32,7%
	Localmente avançado	258	35,8%
	Metástase a distância	214	29,7%
Cirurgia	Não	288	39,9%
	Sim	433	60,1%
Radioterapia	Não	557	77,3%
	Sim	164	22,7%
Quimioterapia	Não	442	61,3%
	Sim	279	38,7%
Ventilação	Espontânea	547	75,9%
	Mecânica invasiva	146	20,2%
	Não invasiva	28	3,9%
Infecção respiratória na UTI	Não	624	86,5%
	Sim	97	13,5%
Sedação profunda	Não	614	85,2%
	Sim	107	14,8%
Alcoolismo	Não	638	88,5%
	Sim	83	11,5%
Tabagismo	Atual	103	14,3%
	Nunca/Prévio	618	85,7%
Local agrupado	hematológicas	41	5,7%
	sólidas	680	94,3%

Tabela A.9: Modelo de Cox estratificado completo

Variável	β	exp(β)	SE(β)	z	Valor-P
Idade	0,0008	1,001	0,004	0,204	0,8385
Sexo - Masculino	0,0437	1,045	0,117	0,373	0,7091
IMC	-0,0536	0,948	0,012	-4,536	$<$ 0,0001
Status câncer - Ativo - recaída/progressão	0,0161	1,016	0,145	0,110	0,9120
Status câncer - Controlado remissão	0,0258	1,026	0,246	0,105	0,9166
Extensão câncer - Limitado	-0,7926	0,453	0,399	-1,984	0,0472
Extensão câncer - Localmente avançado	-0,6101	0,543	0,389	-1,567	0,1172
Extensão câncer - Metástase a distancia	-0,0402	0,961	0,386	-0,104	0,9171
Cirurgia - Sim	-0,3597	0,698	0,121	-2,968	0,0030
Quimioterapia - Sim	0,1696	1,185	0,138	1,230	0,2187
Ventilação - Mecânica invasiva	0,0661	1,068	0,183	0,362	0,7174
Ventilação - Não invasiva	0,5726	1,773	0,242	2,363	0,0181
Dias até entrada na UTI	0,0028	1,003	0,002	1,175	0,2400
Infecção respiratória UTI - Sim	0,8141	2,257	0,156	5,221	$<$ 0,0001
Sedação profunda - Sim	0,7375	2,091	0,187	3,949	$<$ 0,0001
Alcoolismo - Sim	0,2223	1,249	0,181	1,230	0,2186
Tabagismo - Atual	-0,2572	0,773	0,167	-1,536	0,1246
Local agrupado - sólidas	0,4633	1,589	0,270	1,713	0,0867

Tabela A.10: Modelo de Cox estratificado reduzido

Variável	β	exp(β)	SE(β)	z	Valor-P
IMC	-0,0512	0,950	0,012	-4,398	$<$ 0,0001
Extensão câncer - Outros	0,4545	1,575	0,131	3,481	0,0005
Cirurgia - Sim	-0,3247	0,723	0,112	-2,888	0,0039
Ventilação - Outras	-0,4160	0,660	0,241	-1,729	0,0838
Infecção respiratória UTI - Sim	0,7861	2,195	0,149	5,274	$<$ 0,0001
Sedação profunda - Sim	0,7907	2,205	0,136	5,835	$<$ 0,0001

Tabela A.11: Medidas da área sob a curva ROC e KS sob a abordagem *incidente/dinâmico*

Tempo	Modelo Completo				Modelo Reduzido			
	Treinamento		Validação		Treinamento		Validação	
	ASC	KS	ASC	KS	ASC	KS	ASC	KS
6	0,5217	0,0324	0,5280	0,0418	0,5195	0,0288	0,5251	0,0370
15	0,5318	0,0458	0,5487	0,0739	0,5294	0,0415	0,5438	0,0681
32	0,5388	0,0570	0,5460	0,0676	0,5365	0,0527	0,5429	0,0632
84	0,5403	0,0586	0,5463	0,0652	0,5374	0,0527	0,5470	0,0688
181	0,5452	0,0643	0,5472	0,0654	0,5437	0,0628	0,5494	0,0713
300	0,5462	0,0669	0,5554	0,0749	0,5433	0,0605	0,5577	0,0786
361	0,5477	0,0694	-	-	0,5450	0,0634	-	-
378	0,5517	0,0770	-	-	0,5505	0,0726	-	-
547	0,5641	0,1062	-	-	0,5645	0,1021	-	-

Tabela A.12: Medidas da área sob a curva ROC sob a abordagem *cumulativo/dinâmico* apresentada na seção 3.6

Tempo	Modelo Completo		Modelo Reduzido	
	Treinamento	Validação	Treinamento	Validação
6	0,6727	0,5286	0,6572	0,5255
15	0,6830	0,5471	0,6679	0,5423
32	0,6988	0,5592	0,6826	0,5537
84	0,7271	0,5682	0,7063	0,5625
181	0,7591	0,5736	0,7308	0,5694
300	0,7813	0,5736	0,7493	0,5710
361	0,7879	0,5728	0,7550	0,5706
378	0,7909	0,5726	0,7578	0,5706
547	0,8134	0,5695	0,7783	0,5698

Tabela A.13: Medidas da área sob a curva ROC e KS sob a abordagem *cumulativo/dinâmico* apresentada na seção 3.7

Tempo	Modelo Completo				Modelo Reduzido			
	Treinamento		Validação		Treinamento		Validação	
	ASC	KS	ASC	KS	ASC	KS	ASC	KS
6	0,7664	0,4358	0,5905	0,3784	0,7280	0,3953	0,5770	0,2865
15	0,7760	0,4404	0,6251	0,3315	0,7435	0,4041	0,6102	0,2502
32	0,7955	0,4888	0,5614	0,2045	0,7536	0,4234	0,5773	0,1932
84	0,8039	0,4880	0,5389	0,1607	0,7607	0,4134	0,5568	0,1607
181	0,8005	0,4650	0,4474	0,1111	0,7725	0,4364	0,4281	0,1111
300	0,8011	0,4647	0,4699	0,4699	0,7744	0,4655	0,6988	0,6988
361	0,8062	0,4880	-	-	0,7788	0,4694	-	-
378	0,7935	0,4432	-	-	0,7671	0,4491	-	-
547	0,7861	0,4786	-	-	0,7646	0,5124	-	-

Apêndice B

Demonstrações de resultados do capítulo Função de Perda

Nesse apêndice serão demonstrados alguns resultados utilizados no capítulo Função de Perda.

(**Página** 12) Mostrar que $\hat{W}(t) = \mathbb{I}(1 - F(y|\mathbf{z}) \geq 0,5)$ é o preditor que minimiza a perda absoluta:

Lembre que a variável que define o status é tal que $W(t) = \mathbb{I}(Y > t)$. Agora suponha que $P(W(t) = 1) = P(\mathbb{I}(Y > t) = 1) = P(Y > t) > 0,5$. Assim, se $\hat{W}(t) = 0$, a perda é dada por:

$$E(\mathbb{I}(W(t) \neq 0)) = P(W(t) = 1) > 0,5.$$

Alternativamente, se $\hat{W}(t) = 1$, a perda é dada por:

$$E(\mathbb{I}(W(t) \neq 1)) = P(W(t) = 0) < 0,5.$$

Portanto, nesse caso, deve se decidir por 1. Analogamente, se $P(W(t) = 1) < 0,5$, deve se decidir por 0. Logo, a decisão ótima é dada por $\mathbb{I}(P(W(t) = 1) > 0,5)$.

Apêndice C

Demonstrações de resultados do capítulo Curva ROC

Nesse apêndice serão demonstrados alguns resultados utilizados no capítulo Curva ROC.

(**Página** 33) Mostrar, no contexto *incidente/dinâmico*, que $ASC(t) = P(M_j > M_k|(T_j = t, T_k > t))$.

Considere as observações i e k independentes. Assim, pode-se escrever

$$\begin{aligned}
P(M_j > M_k|T_j = t, T_k > t) &= \int_{-\infty}^{\infty} P(M_j > c, M_k = c|T_j = t, T_k > t)dc \\
&= \int_{-\infty}^{\infty} \frac{P(M_j > c, M_k = c|T_j = t, T_k > t)P(T_j = t)P(T_k > t)}{P(T_j = t)P(T_k > t)}dc \\
&= \int_{-\infty}^{\infty} \frac{P(M_j > c, M_k = c|T_j = t, T_k > t)P(T_j = t, T_k > t)}{P(T_j = t)P(T_k > t)}dc \\
&= \int_{-\infty}^{\infty} \frac{P(M_j > c, T_j = t, M_k = c, T_k > t)}{P(T_j = t)P(T_k > t)}dc \\
&= \int_{-\infty}^{\infty} \frac{P(M_j > c, T_j = t)}{P(T_j = t)} \frac{P(M_k = c, T_k > t)}{P(T_k > t)}dc \\
&= \int_{-\infty}^{\infty} P(M_j > c|T_j = t)P(M_k = c|T_k > t)dc. \qquad \text{(C.1)}
\end{aligned}$$

Lembre que a área sob a curva ROC pode ser dada pela seguinte equação:

$$ASC(t) = \int_0^1 TVP_t \left\{ [TFP_t]^{-1}(p) \right\} dp,$$

em que p é igual a $P(M > c^p|T > t)$. Assim, pode-se escrever

$$\frac{dp}{dc} = \frac{dP(M > c^p|T > t)}{dc} = P(M = c|T > t).$$

Logo, fazendo a mudança de variável de p para c, é possível escrever

$$ASC(t) = \int_{-\infty}^{\infty} P(M > c|T = t)P(M = c|T > t)dc. \qquad \text{(C.2)}$$

Logo, utilizando (C.2) em (C.1), mostra-se que

$P(M_j > M_k|T_j = t, T_k > t) = \int_{-\infty}^{\infty} P(M_j > c|T_j = t)P(M_k = c|T_k > t)dc = ASC(t)$ □.

(**Página** 33) Mostrar que $C = \int ASC(t)u(t)dt$.

Por definição, $C = P(M_i > M_j|T_i < T_j)$. Assim, pode se escrever

$$\begin{aligned} C = P(M_i > M_j|T_i < T_j) =& \frac{P(M_i > M_j, T_i < T_j)}{P(T_i < T_j)} \\ =& 2P(M_i > M_j, T_i < T_j) \\ =& 2P(M_i > M_j|T_i < T_j)P(T_i < T_j) \\ =& 2\int P(M_i > M_j|(T_i = t, t < T_j))P(T_i = t, t < T_j)dt. \end{aligned}$$

Utilizando a relação obtida em (C.1), pode-se escrever

$$\begin{aligned} C = P(M_i > M_j|T_i < T_j) =& 2\int ASC(t)P(T_i = t, t < T_j)dt \\ =& 2\int ASC(t)P(T_i = t)P(T_j > t)dt \\ =& \int ASC(t)2P(T_i = t)P(T_j > t)dt \\ =& \int ASC(t)2f(t)S(t)dt \\ =& \int ASC(t)u(t)dt. \ \square \end{aligned}$$

(**Página** 34) Mostrar que $U^\tau = 1 - S^2(\tau)$.

Lembre que U^τ é dado por:

$$\begin{aligned} U^\tau =& \int_0^\tau 2f(t)S(t)dt \\ =& -\int_0^\tau -2f(t)S(t)dt \\ =& -\left[S^2(t)\Big|_0^\tau\right]. \end{aligned}$$

Nesse trabalho, considera-se, por definição, que a função de sobrevivência no instante zero é igual a um. Portanto, pode-se escrever que

$$\begin{aligned} U^\tau =& -\left[S^2(\tau) - 1\right] \\ =& 1 - S^2(\tau). \ \square \end{aligned}$$

(**Pagina** 36) Mostrar que $ASC^{\mathbb{C}/\mathbb{D}}(t) = \int_{-\infty}^{\infty} P(M > m|T \leq t)P(M = m|T > t)dm$.

A área sob a curva pode ser dada pela seguinte equação:

$$ASC^{\mathbb{C}/\mathbb{D}}(t) = \int_0^1 TVP\left\{TFP^{-1}(q,t);t\right\} dq. \tag{C.3}$$

Para mostrar o resultado, será feita a mudança de variável de q para M. Assim, considere $\{TFP\}^{-1}(q,t) = m_q$, em que m_q é o valor do marcador M associado ao q-ésimo quantil. Então, pode-se escrever

$$TFP\left[\{TFP\}^{-1}(q,t)\right] = TFP\left[m_q,t\right] = q.$$

Assim, tomando a derivada da função acima em relação a q, obtém-se:

$$\frac{dTFP\left[\{TFP\}^{-1}(q,t)\right]}{dq} = \frac{dTFP\left[m_q,t\right]}{dq} = \frac{dq}{dq} = 1.$$

Utilizando a regra da cadeia no primeiro termo da igualdade acima, obtém-se:

$$\frac{dTFP\left[\{TFP\}^{-1}(q,t)\right]}{dm}\frac{d\{TFP\}^{-1}(q,t)}{dq} = 1.$$

Daí, vem que

$$\begin{aligned}
\frac{d\{TFP\}^{-1}(q,t)}{dq} &= \left\{\frac{dTFP\left[\{TFP\}^{-1}(q,t)\right]}{dm}\right\}^{-1} \\
&= \left\{\frac{dTFP\left[m,t\right]}{dm}\right\}^{-1} \\
&= \left\{\frac{dP(M>m|T>t)]}{dm}\right\}^{-1} \\
&= \{P(M=m|T>t)\}^{-1} = \frac{1}{P(M=m|T>t)}.
\end{aligned}$$

Logo, pode-se concluir que

$$\frac{dm}{dq} = \frac{d\{TFP\}^{-1}(q,t)}{dq} = \frac{1}{P(M=m|T>t)}.$$

Portanto, utilizando a mudança de variável de q para M em (C.3), obtém-se que:

$$\begin{aligned}
ASC^{\mathbb{C}/\mathbb{D}}(t) &= \int_0^1 TVP\left\{TFP^{-1}(q,t);t\right\} dq \\
&= \int_{-\infty}^{\infty} TVP(m,t)P(M=m|T>t)dm \\
&= \int_{-\infty}^{\infty} P(M>m|T\leq t)P(M=m|T>t)dm. \ \square
\end{aligned}$$

(**Página** 36) Combinando $TVP^{\mathbb{C}}$, $TFP^{\mathbb{D}}$ e $ASC^{\mathbb{C}/\mathbb{D}}(t)$ mostra-se que:

$$\begin{aligned} ASC^{\mathbb{C}/\mathbb{D}}(t) &= \int_{-\infty}^{\infty} P(M > c|T \leq t) d\left[P(M > c|T > t)\right] \\ &= \int_{-\infty}^{\infty}\int_{c}^{\infty} \frac{P(T \leq t|M = m)P(M = m)}{P(T \leq t)} \frac{P(T > t|M = c)P(M = c)}{P(T > t)} dmdc \\ &= \int_{-\infty}^{\infty}\int_{c}^{\infty} \frac{F(t|M = m)e(m)}{F(t)} \frac{\left[1 - F(t|M = c)\right]e(c)}{1 - F(t)} dmdc \\ &= \int_{-\infty}^{\infty}\int_{c}^{\infty} \frac{F(t|M = m)\left[1 - F(t|M = c)\right]}{F(t)\left[1 - F(t)\right]} e(m)e(c)dmdc. \end{aligned} \tag{C.4}$$

(**Página** 37) Mostrar que $ASC^{\mathbb{C}/\mathbb{D}}(t) = \frac{\int_0^1 vF(t|M=E^{-1}(v))dv - \frac{F^2(t)}{2}}{F(t)[1-F(t)]}$.

Esse demonstração se encontra no artigo de Viallon e Latouche (2011). Utilizando a mudança de variável de m para $E^{-1}(u)$ e c para $E^{-1}(v)$ no numerador de (C.4), pode-se escrever:

$$\begin{aligned} &\int_{-\infty}^{\infty}\int_{c}^{\infty} F(t|M = m)\left[1 - F(t|M = c)\right]e(m)e(c)dmdc \\ &= \int_{0}^{1}\int_{v}^{1} F(t|M = E^{-1}(u)\left[1 - F(t|M = E^{-1}(v)\right]dudv \\ &= \int_{0}^{1}\int_{v}^{1} \left[1 - S(t|M = E^{-1}(u))\right]S(t|M = E^{-1}(v))dudv \\ &= \int_{0}^{1}\int_{v}^{1} \left[S(t|M = E^{-1}(v)) - S(t|M = E^{-1}(u))S(t|M = E^{-1}(v))\right]dudv \\ &= \int_{0}^{1}(1 - v)S(t|M = E^{-1}(v))dv - \int_{0}^{1}\int_{v}^{1} S(t|M = E^{-1}(u))S(t|M = E^{-1}(v))dudv \\ &= \int_{0}^{1}(1 - v)S(t|M = E^{-1}(v))dv - \int_{0}^{1}\int_{0}^{1} \mathbb{I}(u \geq v)S(t|M = E^{-1}(u))S(t|M = E^{-1}(v))dudv. \end{aligned}$$

Seja $L(u, v) = S(t|M = E^{-1}(u))S(t|M = E^{-1}(v))$, então $L(u, v) = L(v, u)$. Assim,

$$\int_{0}^{1}\int_{0}^{1} \mathbb{I}(u \geq v)S(t|M = E^{-1}(u))S(t|M = E^{-1}(v))dudv = \frac{1}{2}\int_{0}^{1}\int_{0}^{1} S(t|M = E^{-1}(u))S(t|M = E^{-1}(v))dudv.$$

Dando continuidade à equação anterior, utilizando essa última igualdade, obtém-se que

$$\int_0^1 (1-v)S(t|M=E^{-1}(v))dv - \frac{1}{2}\int_0^1\int_0^1 S(t|M=E^{-1}(u))S(t|M=E^{-1}(v))dudv$$
$$= \int_0^1 (1-v)S(t|M=E^{-1}(v))dv - \frac{1}{2}\int_0^1 S(t|M=E^{-1}(u))du \int_0^1 S(t|M=E^{-1}(v))dv$$
$$= \int_0^1 (1-v)S(t|M=E^{-1}(v))dv - \frac{1}{2}\left[\int_0^1 S(t|M=E^{-1}(v))dv\right]^2$$
$$= \int_0^1 (1-v)S(t|M=E^{-1}(v))dv - \frac{1}{2}\left[S(t)\right]^2.$$

Logo, substituindo essa igualdade no numerador de (C.4), mostra-se que

$$\begin{aligned}
ASC^{\mathbb{C}/\mathbb{D}}(t) &= \frac{\int_0^1 (1-v)S(t|M=E^{-1}(v))dv - \frac{1}{2}\left[S(t)\right]^2}{F(t)\left[1-F(t)\right]} \\
&= \frac{\int_0^1 (S(t|M=E^{-1}(v))dv - \int_0^1 vS(t|M=E^{-1}(v))dv - \frac{1}{2}\left[S(t)\right]^2}{F(t)\left[1-F(t)\right]} \\
&= \frac{S(t) - \int_0^1 vS(t|M=E^{-1}(v))dv - \frac{1}{2}\left[S(t)\right]^2}{F(t)\left[1-F(t)\right]} \\
&= \frac{S(t) - \int_0^1 v\left[1-F(t|M=E^{-1}(v))\right]dv - \frac{1}{2}\left[S(t)\right]^2}{F(t)\left[1-F(t)\right]} \\
&= \frac{S(t) - \int_0^1 vdv + \int_0^1 vF(t|M=E^{-1}(v))dv - \frac{1}{2}\left[S(t)\right]^2}{F(t)\left[1-F(t)\right]} \\
&= \frac{S(t) - \frac{v^2}{2} + \int_0^1 vF(t|M=E^{-1}(v))dv - \frac{1}{2}\left[S(t)\right]^2}{F(t)\left[1-F(t)\right]} \\
&= \frac{S(t) - \frac{1}{2} + \int_0^1 vF(t|M=E^{-1}(v))dv - \frac{1}{2}\left[S(t)\right]^2}{F(t)\left[1-F(t)\right]} \\
&= \frac{1-F(t) - \frac{1}{2} + \int_0^1 vF(t|M=E^{-1}(v))dv - \frac{1}{2}\left[1-F(t)\right]^2}{F(t)\left[1-F(t)\right]} \\
&= \frac{1-F(t) - \frac{1}{2} + \int_0^1 vF(t|M=E^{-1}(v))dv - \frac{1}{2}\left[1-2F(t)+F^2(t)\right]}{F(t)\left[1-F(t)\right]} \\
&= \frac{1-F(t) - \frac{1}{2} + \int_0^1 vF(t|M=E^{-1}(v))dv - \frac{1}{2} + F(t) - \frac{F^2(t)}{2}}{F(t)\left[1-F(t)\right]} \\
&= \frac{\int_0^1 vF(t|M=E^{-1}(v))dv - \frac{F^2(t)}{2}}{F(t)\left[1-F(t)\right]}.
\end{aligned}$$

Note que, no contexto cumulativo/dinâmico, $F(t|M=E^{-1}(v))$ é igual a $R_t^p(v)$. Assim, mostra-se que

$$ASC^{\mathbb{C}/\mathbb{D}}(t) = \frac{\int_0^1 vR_t^p(v)dv - \frac{F^2(t)}{2}}{F(t)\left[1-F(t)\right]}. \square$$

(**Página** 29) Mostrar que se $R(q) = \mathbb{I}(q \geq 1-p)$, então $ASC = 1$.

$$
\begin{aligned}
ASC^{\mathbb{I}/\mathbb{D}}(t) &= \frac{1}{p(1-p)}\left[\int_0^1 q\mathbb{I}(q \geq 1-p)dq - \frac{p^2}{2}\right] \\
&= \frac{1}{p(1-p)}\left[\int_{1-p}^1 q\mathbb{I}(q \geq 1-p)dq - \frac{p^2}{2}\right] \\
&= \frac{1}{p(1-p)}\left[\frac{q^2}{2}\Big|_{1-p}^1 - \frac{p^2}{2}\right] \\
&= \frac{1}{p(1-p)}\left[\frac{1}{2} - \frac{(1-p)^2}{2} - \frac{p^2}{2}\right] \\
&= \frac{1}{p(1-p)}\left[\frac{1}{2} - \frac{1}{2} + p - \frac{p^2}{2} - \frac{p^2}{2}\right] \\
&= \frac{1}{p(1-p)}\left[p - p^2\right] \\
&= \frac{1}{p(1-p)}\left[p(1-p)\right] \\
&= 1 \ \square.
\end{aligned}
$$

Referências Bibliográficas

Bender, R., Augustin, T. e Blettner, M. Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24:1713–1723, 2005. 55

Cai, T., Pepe, M., Zheng, Z., Lumley, T. e Jenny, N. The sensitivity and specificity of markers for event times. *Biostatistics*, 7(2):182–197, 2006. 4

Casella, G. e Berger, R. *Statistical Inference*. Duxbury, 2nd edition, 2001. 15, 26, 50

Colosimo, E. A. e Giolo, S. R. *Análise de sobrevivência aplicada*. Edgard Blücher LTDA, 2006. 64

Cox, D. Regression models and life-tables. *Journal of the Royal Statistical Society*, 34:187–220, 1972. 35

da Silva, L. T. *Modelos baseados em pseudo-valores e sua aplicabilidade em credit scoring*. 2010. Dissertação, Instituto de Matemática e Estatística da Universidade de São Paulo, 2010. 56

Efron, B. The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99:619–632, 2004. 11, 14, 18, 19, 20

Gerds, T. e Schumacher, M. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48:1029–1040, 2006. 3, 16

Graf, E., Schmoor, C., Sauerbrei, W. e Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18:2529–2545, 1999. 3

Harrel Jr, F. E., Lee, K. L. e Mark, D. B. Tutorial in biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing error. *Statistics in Medicine*, 15:361–387, 1996. 31

Heagerty, P. e Saha-Chaudhuri, P. *Riskset ROC curve estimation from censored survival data*, 2012. URL `http://cran.r-project.org/web/packages/risksetROC`. 65

Heagerty, P. e Zheng, Y. Survival model predictive accuracy and roc curves. *Biometrics*, 61:92–105, 2005. 3, 23, 30, 31, 33

Horvitz, D. e Thompson, D. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. 15

Huang, Y., Pepe, M. e Feng, Z. Evaluating the predictiveness of a continuous marker. *Biometrics*, 63:1181–1188, 2007. 26

Kaplan, E. e Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958. 37

Klein, J. e Moeschberger, M. *Survival Analysis Techniques for Censored and Truncated Data*. Springer, 2nd edition, 2003. 1, 55

Korn, E. e Simon, R. Measures of explained variation for survival data. *Statistics in Medicine*, 9: 487–503, 1990. 2, 3

Krzanowski, W. e Hand, D. *ROC curves for continuous data*. Chapman & Hall/CRC, 2nd edition, 2009. 21

Lawless, J. F. e Yuan, Y. Estimation of prediction error for survival models. *Statistics in Medicine*, 29:262–274, 2010. 3, 4, 39

Liang, K. Y. e Zeger, S. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986. 41

Lipsitz, S. e Fitzmaurice, G. Generalized estimating equations for longitudinal data analysis. In Fizmaurice, G., Davidian, M., Verbeke, G. e Molenberghs, G., editors, *Longitudinal data analysis*. Chapman & Hall/CRC, 2008. 41

Lusted, L. Logical analysis in roentgen diagnosis. *Radiology*, 74:178–193, 1960. 21

Montgomery, D., Peck, E. e Vining, G. *Introduction to linear regression analysis*. Wiley, 3nd edition, 2001. 2

Pencina, M. e D'Agostino, R. Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 23:2109–2123, 2004. 32

Pepe, M. S. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2003. 23, 25

Peterson, W., Birdsall, T. e Fox, W. The theory of signal detectability. Technical report, University of Michigan, 1953. 21

Robins, J. e Rotnitzky, A. Confirmar recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS Epidemiology - Methodological Issues*, pages 297–331, 1992. 3

Rosthoj, S. e Keiding, N. Explained variation and predictive accuracy in general parametric statistical models: the role of model misspecification. *Lifetime Data Analysis*, 10(4):461–472, 2004. 16

Saha, P. *Time-dependent Predictive Accuracy: Extending Binary Classification Accuracy Methods for Censored Survival Data*. Doctoral thesis, University of Washington, 2009. 31

Schmid, M. e Potapov, S. A comparison of estimators to evaluate the discriminatory power of time-to-event models. *Statistics in Medicine*, 31:2588–2609, 2012. 33

Uno, H., Cai, T., Tian, L. e Wei, L. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537, 2007. 3, 4, 38, 39, 45, 55, 66, 69

Viallon, V. e Latouche, A. Discrimination measures for survival outcomes: Connection between the auc and the predictiveness curve. *Biometrical Journal*, 53:217–236, 2011. 4, 36, 37, 38, 114

Vittinghof, E., Shiboski, S., Glidden, D. e McCulloch, E. *Regression Methods in Biostatistics: Linear, Logistic, Survival and Repeated Measures Models*. Springer, 2004. 22

Wedderburn, R. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61:439–447, 1974. 41

Xu, R. e O'Quigley, J. Proportional hazards estimate of the conditional survival function. *Journal of the Royal Statistical Society: Series B*, 62:667–680, 2000. 34

Yuan, Y. *Prediction Performance of Survival Models*. Doctoral thesis, University of Waterloo, Ontario, Canada, 2008. DisponÃvel em http://uwspace.uwaterloo.ca/bitstream/10012/3974/1/thesis.pdf. 7, 11, 14, 17, 18, 19, 20

Zhou, X., Obuchowski, N. e McClish, D. *Statistical Methods in Diagnostic Medicine*. Wiley, 2002. 25