UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Estudos longitudinais para avaliação de custo na área da saúde: como tratar dados faltantes e censuras

Autor: Pricila Henkes Maciel
Orientador: Professora Dra. Patrícia Klarmann Ziegelmann

Porto Alegre, 21 de Dezembro de 2012.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Estudos longitudinais para avaliação de custo na área da saúde: como tratar dados faltantes e censuras

Autor: Pricila Henkes Maciel

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professora Dra. Patrícia Klarmann Ziegelmann (orientador)
Professora Dra. Luciana Neves Nunes

Porto Alegre, 21 de Dezembro de 2012.

"Enquanto suspiramos por uma vida sem dificuldades, devemos nos lembrar que o carvalho cresce forte através de ventos contrários e que os diamantes são formados sob pressão."

Peter Marshall

Agradecimentos

Agradeço aos meus pais, por terem me ensinado a sonhar e a lutar pelos meus sonhos. “A gente só não consegue aquilo que a gente não quer” foi uma frase que eu ouvi a minha vida inteira, e levo comigo eternamente. Por todo o amor, o incentivo, a paciência e também pela educação que me deram. Por trabalhar duro para que eu pudesse ter sempre o melhor. Palavras não descrevem todo o meu amor, orgulho e gratidão que eu sinto por ser filha de vocês.

Agradeço a vó Tereza, uma pessoa que me inspira muito. Um exemplo de determinação. Superou várias dificuldades sem nunca perder a esperança e o sorriso no rosto.

Agradeço a minha irmã Tássia e as amigas Natalia Giordani e Sabrina Grebin, por terem trilhado esse caminho ao meu lado em todos os momentos, me dando força quando estava difícil a batalha e vibrando junto comigo a cada vitória. Esse “Quarteto Fantástico” (nome dado pela professora Vanessa) continuará existindo muito além da faculdade. A amizade de vocês é uma das minhas maiores riquezas.

Agradeço ao meu melhor amigo e namorado Lucas, por nunca permitir que eu desanimasse, pelo amor, pela paciência, por ter estado do meu lado nos momentos em que eu mais precisei. Não tenho palavras para agradecer todo o apoio que sempre me deu. Sem tua ajuda teria sido ainda mais difícil superar todos os desafios que a vida me deu esse ano. Eu te amo, muito obrigada.

Agradeço aos demais colegas de curso e amigos da UFRGS, os formados e os que ainda estão na batalha, em especial aos amigos Letícia, Natália, Mariana, Mauro, Andrigo e Mateus Japa, pelos estudos em grupo, pela amizade e apoio de vocês, especialmente no início do curso. Aos professores da faculdade por todos os ensinamentos e conhecimento transmitido, em especial a professora Patrícia Ziegelmann, por ter me aceitado como tua orientanda.

A Ana Claudia, pela amizade verdadeira desde a infância, um carinho que a distancia nunca apagou e a saudade só faz crescer.

Aos demais familiares e amigos que me incentivaram a estudar na UFRGS, carinho de vocês e também pela compreensão todas as vezes que não pude estar tão presente quanto gostaria. Não posso cita todos, mas vocês sabem quem são e a importância que tem. A todos vocês, muito obrigada.

Resumo

A cada dia novas tecnologias são desenvolvidas na área da saúde para tratamento de doenças. Os recursos disponíveis para investir em tecnologias são limitados e precisam atender uma grande quantidade de pacientes, sendo necessário avaliar as tecnologias quanto aos custos envolvidos e a efetividade. Uma maneira de avaliar custos é através da realização de estudos longitudinais onde os custos da tecnologia avaliada são observados em várias consultas ao longo de um determinado tempo. Uma característica recorrente desse tipo de estudo e que prejudica a análise dos dados é a presença de observações faltantes e/ou casos de censura.

Esta monografia procura apresentar, de forma simples e didática, algumas técnicas para tratamento de dados faltantes oriundos de estudos longitudinais. São apresentadas a técnica da análise de casos completos, alguns métodos de imputação de dados e, também, técnicas para o tratamento de dados censurados tais como, o estimador KMSA, proposto por Lin *et al*. (1997), o estimador IPW, proposto por Bang e Tsiatis (2000) e o uso da ponderação IPW aliado a técnicas de regressão, como proposto por Lin (2000).

Para melhor ilustrar as técnicas apresentadas é utilizado um banco de dados adaptado de um estudo real que avalia os custos envolvidos em dois tratamentos para PSP (Progressive supranuclear palsy) e MSA (multiple system atrophy). A aplicação das técnicas é feita utilizando os programas Excel e Stata e o texto apresenta um detalhamento completo da análise. No exemplo apresentado não foi possível aplicar a técnica de ponderação proposta por Lin (2000) para a análise de regressão, pois foram violadas as suposições necessárias à aplicação deste modelo. Como alternativa a mesma técnica de ponderação proposta por Lin (2000) foi, então, aplicada utilizando Modelos Lineares Generalizados. Quanto à aplicação das técnicas de tratamento para dados censurados, observaram-se desempenhos semelhantes para estimar o custo total de três anos e meio de tratamento de PSP e MSA. Cabe ressaltar que o objetivo desta monografia não é comparar as técnicas entre si e sim destacar que escolha de qual técnica utilizar e o desempenho de cada uma dependem das características específicas de cada estudo e de cada de banco de dados.

Palavras-chave: Custo, estudos longitudinais, dados faltantes, técnicas de imputação, dados censurados, ponderação, *Kaplan-Meier Sample Average, Inverse Probability Weighting*.

Abstract

Every day new technologies are developed in healthcare to treat diseases. The resources available to invest in technologies are limited and required for a large number of patients, so it is necessary to evaluate the technologies for their effectiveness and the costs involved. One way to assess costs is through longitudinal studies which evaluate the costs of technology in multiple observations over a certain time. A recurring feature of this type of study that affects the data analysis is the presence of missing observations and / or cases of censorship.

This monograph attempts to present, in a simple and didactic way, some techniques for handling missing data from longitudinal studies. It presents complete cases analysis technique, some methods of data imputation, and also techniques to handle censored data, such as the KMSA estimator proposed by Lin *et al*. (1997), the IPW estimator proposed by Bang and Tsiatis (2000) and the use of IPW combined with regression techniques, as proposed by Lin (2000).

To better illustrate the techniques presented, we used a database adapted from a real study that evaluates the costs involved in two treatments for PSP (progressive supranuclear palsy) and MSA (multiple system atrophy). The application of the techniques is performed using softwares Excel and Stata, and the text presents full details on the analysis. In the example shown it was not possible to apply the weighting technique proposed by Lin (2000) for the regression analysis because the assumptions necessary to implement this model were violated. Alternatively, the same weighting technique proposed by Lin (2000) was then applied using Generalized Linear Models. As for the application of the techniques to handle censored data, we observed similar results to estimate the total mean cost of three and a half years of treatment for PSP and MSA. It should be noted that the purpose of this monograph is not to compare the techniques with each other and instead emphasize that the choice of which technique to use and the performance of each one of them depends on the characteristics that are specific to each study and each database.

Keywords: Cost, longitudinal studies, missing data, imputation techniques, censored data, weighting, Kaplan-Meier Sample Average, Inverse Probability Weighting.

Sumário

1. Introdução

A demanda de estudos de análise de custo na área da saúde é crescente conforme se desenvolvem novas tecnologias para o tratamento de enfermidades. Diferentes tecnologias que tratem da mesma doença precisam ser avaliadas e comparadas quanto aos custos envolvidos no tratamento e a efetividade (o quanto melhora a saúde e a qualidade de vida do paciente) visto que os recursos disponíveis são limitados e precisam atender uma grande quantidade de pacientes (Gray *et. al* (2011)). Para estimar o custo total por paciente em um pré-determinado período de tempo é comum a utilização de dados provenientes de estudos nos quais os pacientes são observados em várias consultas ao longo desse período, conhecidos como estudos longitudinais. Uma característica recorrente desse tipo de estudo e que prejudica a análise dos dados é a presença de dados faltantes e/ou casos de censura. A avaliação da efetividade das tecnologias é um capítulo a parte e não será discutido neste trabalho.

Quando não é possível avaliar o custo de um paciente em um ou mais dos tempos estudados ocorre um dado faltante. Vários métodos para trabalhar com esta limitação são encontrados na literatura. O método mais simples propõe que apenas os pacientes que foram observados de forma completa (apresentam observação do custo para todos os tempos planejados) sejam utilizados na estimação do custo total médio. A limitação deste método fica evidente pela diminuição no tamanho da amostra. Para contornar esta situação a literatura propõe métodos de imputação (utilizar um valor coerente para cada observação faltante). Este valor coerente pode ser uma média/mediana dos dados observados ou, ainda, o resultado de um modelo de regressão (utilizando os casos com dados completos) no qual variáveis relevantes tais como idade, gênero, intensidade da doença e etc. são consideradas, como descrito em Briggs *et al*.(2002). Para estudos longitudinais, Jean Mundahl Engels e Paula Diehr (2002) propõe algumas técnicas para imputar dados faltantes baseadas em informações de subgrupos (pacientes com perfil semelhante ao do paciente cujo custo está sendo imputado) tais como média/mediana condicional ou '*hot deck*', ou imputar os dados faltantes através de média/mediana dos valores de custo observados ao próprio paciente em outros períodos. Sabe-se que a escolha do método de imputação a ser utilizado deve ser realizada com critério visto que o método aplicado influencia diretamente no custo total estimado. Para tal certas características da ocorrência dos dados faltantes tais como, padrão e mecanismos de aparecimento, devem ser avaliadas visto que elas direcionam a escolha do método (Briggs *et al*. (2002)).

Um caso particular ocorre quando um paciente apresenta dados faltantes em todas as observações a partir de um determinado momento. Esta situação é conhecida na literatura por censura. Quando há casos de censura, a técnica mais simples de análise (que não consideram que os dados são censurados) subestima o custo total (Gray *et. al* (2011)). Técnicas como o estimador KMSA (*Kaplan-Meier Sample Average*), proposto por Lin *et al*. (1997), e o estimador IPW (*Inverse Probability Weighting*), proposto por Bang e Tsiatis (2000) visam ponderar os dados de

modo a corrigir esse erro de subestimação. Lin (2000) também propôs o uso da ponderação IPW aliado a técnicas de regressão, aplicando a ponderação na estimação dos coeficientes da regressão. A utilização de modelos de regressão com ponderação possibilita que os custos não observados sejam imputados de acordo com covariáveis relevantes e considerando a probabilidade de censura. Sabe-se que dados de custo comumente exibem distribuição assimétrica o que pode inviabilizar a aplicação de modelos de regressão. Como alternativa, este trabalho propõe a utilização da ponderação IPW aliado a modelos de regressão generalizados. Os modelos de regressão linear generalizados são mais flexíveis e, particularmente úteis nas situações onde os dados geram violações nas suposições dos modelos de regressão linear ordinal.

Apesar da crescente demanda de estudos de custo oriundos de dados longitudinais, ainda não existem pacotes estatísticos onde estes métodos estejam programados, e a divulgação destas técnicas é pequena. Motivado por atender essa demanda e contribuir para a escassa literatura acerca desse tema, especialmente em língua portuguesa, este trabalho tem como objetivo principal descrever de forma simples e didática algumas das técnicas aplicáveis para correção de bancos de dados com dados faltantes e censuras. Para tal, este trabalho apresenta e discute algumas técnicas de imputação, o estimador KMSA, proposto por Lin *et al*. (1997), o estimador IPW, proposto por Bang e Tsiatis (2000), o uso da ponderação IPW aliado a técnicas de regressão, como proposto por Lin (2000) e o uso da ponderação IPW aliado a modelos de regressão linear generalizados. As técnicas apresentadas são aplicadas a um banco de dados adaptado de um estudo real que avalia os custos envolvidos em dois tratamentos para PSP (Progressive supranuclear palsy) e MSA (multiple system atrophy). A aplicação das técnicas é feita utilizando os programas Excel e Stata e o texto apresenta um detalhamento completo da análise.

Este trabalho está estruturado da seguinte forma: a Seção 2 apresenta definições de estudos longitudinais e características de bancos de dados, a Seção 3 apresenta a definição de dados faltantes e algumas técnicas de análise e imputação de dados que podem ser utilizadas quando eles ocorrem. Na Seção 4 são definidas técnicas de estimação de custo na presença de dados censurados. A Seção 5 apresenta um exemplo prático para ilustrar e melhor explicar as técnicas propostas. Por fim, a Seção 6 apresenta as conclusões. Nos anexos são encontrados detalhamentos passo a passo de como proceder para aplicar as técnicas utilizando os pacotes estatísticos Excel 2010 e Stata 11.0.

2. Dados longitudinais de custo

Entende-se por estudos longitudinais estudos onde certa característica de interesse é observada ao longo de um período de tempo para diferentes indivíduos. Em estudos longitudinais envolvendo dados de custo, um número pré-determinado de tempos de observação é especificado e, em cada um destes tempos, é observado o valor gasto (custos diretos, tais como valor gasto em consultas médicas, medicação ou custo de internação hospitalar, e custos indiretos, tais como perda de produtividade) entre o último tempo observado e o tempo atual. Os dados de custo formam um

banco que deve ser em formato de matriz onde as linhas representam os pacientes e as colunas os tempos observados. Essas informações parciais devem ser utilizadas para estimar o custo total médio no período do estudo.

Suponha, por exemplo, um estudo cujo interesse seja avaliar o custo médio total em seis meses de tratamento para dez pacientes que tiveram custos observados mensalmente. A Tabela 1 a seguir apresenta como seria estruturado o banco de dados deste estudo ilustrando casos de dados faltantes (tais como o custo referente ao mês dois para os pacientes 7 e 10) e dados censurados (como observado nos pacientes 9 e 10).

Tabela 1: Exemplo fictício deum banco dedados de um estudo longitudinal

PACIENTE	MÊS 1	MÊS 2	MÊS 3	MÊS 4	MÊS 5	MÊS 6	TOTAL (6 meses)
1	15	30	27	24	24	19	139
2	20	33	26	--	30	29	138
3	22	25	27	25	25	27	151
4	27	29	--	27	28	25	136
5	30	24	26	--	29	23	132
6	10	20	17	16	16	19	98
7	17	--	23	22	22	30	114
8	19	29	27	25	24	28	152
9	14	31	19	--	--	--	64
10	18	--	26	23	--	--	67

Os métodos a serem descritos nas próximas seções são apropriados para análise dos dados e estimação de custo total médio em bancos de dados que apresentem observações faltantes e dados censurados tais como ilustrado na Tabela 1 acima.

3. Observações faltantes

As observações faltantes ocorrem quando não há dados de custo para algum paciente em um ou mais tempos observados, ou seja, alguns custos do paciente são desconhecidos, seja por recusa do paciente em informar ou por falha do pesquisador ao coletar a informação. Diversos métodos para contornar o problema causado pela presença de dados faltantes são encontrados na literatura, e cada um deles gera diferentes resultados para o custo total médio estimado. Portanto, a escolha de qual método utilizar é de suma importância, e é imprescindível que o pesquisador detalhe como foram tratados os dados faltantes ao reportar as suas conclusões e resultados. Certas características tais como o padrão e o mecanismo de aparecimento da observação faltante direcionam a escolha do método a ser utilizado para análise. O padrão é avaliado observando-se a forma como os dados faltantes estão distribuídos ao longo do banco de dados. Em alguns casos podem ocorrer dados faltantes para uma grande quantidade de pacientes em um ou dois tempos observados, enquanto os demais tempos apresentam observações completas. Em outros casos

alguns poucos pacientes apresentam um grande número de dados faltantes, em tempos diferentes. Uma terceira situação (em geral a mais frequente) é quando os dados faltantes estão espalhados, e não é possível identificar apenas um tempo observado que tenha apresentado problema ou apenas um pequeno grupo de pacientes que estejam faltando informações (Briggs *et al*. (2002)).

Considerando essa variedade de situações, Rubin (1976) definiu três mecanismos de dados faltantes: 1) faltantes completamente ao acaso (MCAR), quando é possível supor que o custo dos pacientes que não puderam ser avaliados não difere do custo de pacientes para os quais foi possível observar o valor exato, 2) faltantes ao acaso (MAR), quando é possível supor que o custo dos pacientes que não puderam ser observados pode estar correlacionado a outras variáveis que tenham sido observadas, como, por exemplo, sexo ou idade (nesse caso sugere-se o uso de técnicas que controlem estes fatores) e 3) faltantes não aleatoriamente (NMAR), quando o custo de pacientes não observados depende de outras variáveis também não observadas, ou seja, o custo para pacientes que contenham dados faltantes será diferente de forma imprevisível do custo de pacientes em que os dados foram coletados de forma completa.

Como não se sabe o custo verdadeiro das observações faltantes é impossível afirmar qual é o mecanismo que gerou os dados faltantes de um estudo. Porém, através de análise descritiva é possível supor qual mecanismo presente para, então, imputar os custos não observados aplicando uma técnica adequada. Se houver uma grande diferença no custo médio entre os homens e as mulheres que tiveram seus custos observados, por exemplo, há indícios de que os dados faltantes tenham mecanismo MAR em relação à variável sexo. Neste caso técnicas mais específicas que levem em consideração estes subgrupos talvez sejam mais adequadas. Por outro lado, se na análise descritiva, não há indícios de que possa haver diferença entre subgrupos, sejam eles de diferentes sexos ou passando por diferentes tratamentos, técnicas de imputação baseadas em todos os pacientes da amostra podem ser aplicadas, sem a necessidade de considerar subgrupos. O mecanismo NMAR, além de impossível de identificar, é o tipo de dado faltante mais grave, pois para qualquer que seja a técnica aplicada há um alto risco de obterem-se resultados viesados.

O método mais tradicional para o tratamento de dados faltantes e que é utilizado pela maior parte dos pacotes estatísticos é a Análise de Casos Completos. Este método consiste em omitir todos os pacientes que apresentem algum dado faltante. Embora simples de se aplicar esse método apresenta problemas especialmente quando uma grande proporção de pacientes é descartada. Além da perda de dados que podem ser informativos ao estudo, esse método pode acarretar em resultados viesados caso os dados faltantes sejam de mecanismo MAR.

Como método alternativo, especialmente útil no caso de dados faltantes tipo MAR, têm-se técnicas de imputação. A ideia básica é substituir (imputar) as observações faltantes de modo a se obter um banco de dados completo. Existem para isso diversas técnicas, algumas delas utilizando apenas informações referentes ao paciente que apresenta o dado faltante, como observações de tempos anteriores ou posteriores ao tempo não observado, outras utilizando informações dos demais pacientes que estejam no estudo.

3.1 Imputação utilizando informações de outros pacientes

A técnica mais simples de imputação de dados baseada em informações de outros pacientes consiste em utilizar medidas de tendência central (média ou mediana) de todos os pacientes que tenham o custo observado no período para imputar o custo dos pacientes com dado faltante no mesmo período. A escolha de qual medida de tendência central utilizar depende das características dos custos observadas em cada tempo onde a imputação é realizada. Na presença de valores atípicos (pacientes para os quais o custo foi muito superior ou muito inferior aos demais) ou no caso de distribuição assimétrica para custo, é mais recomendado o uso da mediana, uma vez que esta medida é mais robusta nestas situações. Nos demais casos é cabível utilizar a média aritmética. A imputação de valores utilizando medidas de tendência central, apesar de popular, apresenta limitações. Inserir a média ou mediana onde não há o custo exato faz com que o desvio padrão dos dados seja subestimado. Em um estudo no qual o interesse seja comparar os custos de dois ou mais tratamentos, por exemplo, essa técnica não estaria levando em consideração esta possível diferença entre os custos dos estudos. Assim, como não leva em consideração nenhuma covariável, essa técnica tem como suposição que os dados faltantes apresentem mecanismo MCAR.

No caso de dados faltantes com mecanismo MAR recomenda-se imputar a média ou mediana condicional de acordo com algumas covariáveis escolhidas. Por exemplo, suponha que há indícios de diferença do custo entre homens e mulheres. Nesta situação, o valor a ser imputado para dados faltantes de custo de homens seria a média/mediana dos valores observados entre os outros homens. E para as mulheres a média/mediana dos custos das mulheres. Através do uso de modelos de regressão linear é possível estimar a média de custo que considere diversas covariáveis simultaneamente, ajustando a cada dado faltante um valor obtido pela equação de regressão estimada baseada nas informações de custo que se tenha para aquele tempo. Mesmo considerando medidas de tendência central condicionada aos grupos, no entanto, o desvio padrão do custo continua sendo subestimado.

O método "*hot deck*" é uma alternativa à imputação de medida central, para não subestimar o desvio padrão. Consiste em identificar dentro do subgrupo um paciente com características de perfil semelhante as do paciente para o qual esta sendo feita a imputação e substituir o dado faltante pelo dado observado a este outro paciente no tempo sendo imputado.

Outra técnica utilizada para fazer imputação sem subestimar a variabilidade dos dados é a Imputação Múltipla, proposta por Rubin (1987). Essa técnica consiste em imputar cada dado faltante várias vezes, criando múltiplos bancos de dados. A imputação múltipla apresenta maior complexidade teórica e computacional e não será detalhada neste trabalho. Sugere-se como bibliografia acerca desta técnica o trabalho de NUNES (2007).

3.2 Imputação utilizando informações do próprio paciente:

Estudos longitudinais, por serem múltiplas observações espaçadas ao longo do tempo para cada paciente, apresentam também a possibilidade de fazer a imputação baseado apenas nas informações do próprio paciente, custos que tenham sido observados em tempos anteriores ou posteriores ao dado faltante. Algumas abordagens possíveis seriam aplicar a média ou mediana do custo de todos os tempos observados ao paciente, ou apenas média dos tempos imediatamente anterior e imediatamente posterior ao dado faltante. O método da última observação aplicada subsequentemente (LOCF) consiste repetir o custo do último tempo (anterior) observado para imputar custo do tempo que não teve observação, e NOCB (*Next Observation Carried Backward*) usa o próximo custo (seguinte) observado para substituir o dado faltante.

Os métodos de imputação mais aplicados são definidos e resumidos no Quadro 1 a seguir, adaptado de Jean Mundahl Engels e Paula Diehr (2002).

Quadro 1: Métodos de imputação mais comuns

Referência	Método	Valor imputado
Todos os demais pacientes	Média da coluna	Custo médio de todos os pacientes para o referido tempo.
	Mediana da coluna	Custo mediano de todos os pacientes para o referido tempo.
Pacientes de mesmo subgrupo	Média da classe	Custo médio de outros pacientes com características semelhantes para o referido tempo.
	Mediana da classe	Custo mediano de outros pacientes com características semelhantes para o referido tempo.
	Hot deck	Custo de um paciente com perfil semelhante que tenha sido observado para o referido tempo.
Próprio paciente	Média dos anteriores	Custo médio dos tempos anteriores observados do paciente
	Mediana dos anteriores	Custo mediano dos tempos anteriores observados do paciente
	LOCF	Custo do tempo imediatamente anterior que tenha sido observado
	NOCB	Custo do primeiro tempo seguinte que tenha sido observado
	Anterior e Posterior	Custo médio dos tempos imediatamente anterior e imediatamente posterior
	Média do paciente	Custo médio de todos os tempos observados do paciente
	Mediana do paciente	Custo mediano de todos os tempos observados do paciente

É importante ressaltar que a escolha do método está relacionada a características específicas do estudo e do banco de dados, como a possível relação dos custos observados com covariáveis que tenham sido avaliadas, os mecanismos envolvidos e com o padrão de distribuição dos dados faltantes pelo banco de dados. Dessa forma não é possível discriminar um único método como sendo o melhor método, o bom desempenho de cada técnica depende das circunstâncias do estudo. Por esse motivo indica-se fortemente a inclusão de análise de sensibilidade. A análise de sensibilidade tem como objetivo avaliar se os resultados finais do estudo são sensíveis à técnica de

imputação utilizada. Ou seja, diferentes técnicas são aplicadas na imputação dos dados faltantes para então avaliar o impacto de cada uma dessas técnicas nos resultados finais obtidos. Alguns dos principais métodos para análise de sensibilidade estão descritos em Briggs *et al*. (1994), Briggs (1999) e também em Briggs e Gray (1999 a,b).

4. Censura

Censura é o termo utilizado para descrever um particular tipo de dado faltante que ocorre quando os dados de custo não são observados a partir de um determinado tempo até o final do estudo. A censura pode ser originada de três diferentes mecanismos: censura do tipo I, quando o estudo longitudinal tem um tempo de duração previamente estabelecido e é necessário censurar os pacientes que chegaram até o fim do estudo sem apresentar o desfecho de interesse, censura do tipo II, quando os pacientes precisam ser censurados ao fim do estudo, que ocorreu após um determinado número de mortes (não havia data pré-determinada), e censura aleatória, quando um paciente sai no decorrer do estudo, porque desistiu de participar ou porque tenha falecido por outras causas que não a doença cujos custos estejam sendo avaliados (Bastos e Rocha (2006)). Além dos mecanismos as censuras também apresentam diferentes formas: censura à direita, censura à esquerda ou censura intervalar. Para este trabalho apenas serão estudadas técnicas adequadas para trabalhar com censura à direita, mais frequente em estudos de análise de custo. Censura à direita corre quando todos os pacientes foram acompanhados no começo do estudo, mas alguns precisaram ser censurados, pois deixaram o estudo sem que este tivesse chegado ao fim.

4.1 Análise não ajustada para censura

Uma maneira simplista de estimar o custo total de um período pré-determinado seria calcular para cada paciente a soma dos custos de todos os tempos observados e a partir deste total verificar o custo total médio a todos os pacientes. Ao proceder à soma dos custos de um paciente que tenha sido censurado antes do final do estudo, no entanto, há uma subestimação do custo total, uma vez que cada tempo que não tenha custo observado seria considerado na soma como custo zero. Embora esse valor seja verdadeiro para os casos dos pacientes que não foram observados devido a falecimento, quando se trata de um paciente censurado sabe-se que o paciente teve algum custo, embora não se saiba exatamente qual. Assim, quanto maior a proporção de pacientes censurados, maior seria o viés cometido por essa análise, subestimando o custo total.

Uma possível alternativa seria, então, desconsiderar os pacientes que tenham sido censurados e proceder à soma dos tempos observados apenas para os pacientes que não tenham sido censurados, ou seja, aqueles pacientes que tiveram todos os tempos observados ou que morreram no decorrer do estudo (pois estes sabe-se que tiveram de fato custo zero, após a morte). Essa medida, porém, além de diminuir o tamanho de amostra por excluir os pacientes censurados, provavelmente também resultaria em custos totais subestimados devido à maior proporção de

pacientes mortos (com custos zero) nessa nova amostra utilizada. Então os custos estariam viesados para pacientes com maior tempo de sobrevida. Por outro lado, o custo estimado considerando apenas os pacientes que tiveram todos os tempos observados (não foram censurados e nem faleceram no decorrer do estudo) não estaria levando em consideração a taxa de mortalidade da doença e a probabilidade do paciente morrer no decorrer do período do estudo, sendo superestimado e viesado para pacientes com tempo de sobrevida menor.

Algumas técnicas alternativas propõem ponderar os dados observados e ajustar a estimação do custo total médio em estudos nos quais haja censuras. Dentre as técnicas mais populares na literatura encontram-se o Estimador *Kaplan-Meier Sample Average* (KMSA), a ponderação dos custos pelo inverso da probabilidade (IPW) e também o uso da ponderação IPW aliado a técnicas de regressão, aplicando a ponderação na estimação dos coeficientes de regressão.

4.2 Ajuste para censura utilizando o estimador *Kaplan-Meier Sample Average*

O estimador de *Kaplan-Meier* (proposto em Kaplan & Meier, 1958), foi desenvolvido para incorporar os dados de pacientes censurados em estudos de análise de sobrevivência, de forma que fosse possível utilizar a informação parcial que se tem do paciente antes do momento de censura, mesmo sem saber o que aconteceu a partir deste momento. Utilizando este estimador é possível obter estimativas da probabilidade de sobrevivência (sobreviver à morte ou a qualquer outro desfecho de interesse) para cada fração de tempo avaliada, utilizando não só a informação oriunda de dados completos, mas também de dados censurados.

Devido à similaridade existente entre estudos longitudinais para avaliação de custos médicos e estudos longitudinais para a avaliação de tempo de sobrevivência, diversos pesquisadores tais como Quesenberry *et al*. (1989), Hiatt *et al* (1990) e Fenn *et al*. (1995) propuseram utilizar essa e outras técnicas de análise de sobrevivência em estudos de análise de custo. Essa estratégia, no entanto, apresenta limitações, como apontado por Lin *et al.* (1997). Ao aplicar técnicas padrões da análise de sobrevivência em estudos de avaliação de custos, supõe-se que os pacientes acumulem custo através de uma função comum a todos os pacientes, ou seja, que exista uma correspondência direta entre o tempo de sobrevivência e o custo total. No entanto, isso normalmente não acontece na prática. As funções de custo variam entre pacientes, e um paciente que acumula custos em taxas mais altas tende a gerar custos totais mais elevados, tanto no tempo de sobrevida quanto no tempo de censura, o que indica que o custo no tempo de sobrevivência tem correlação positiva com o custo no tempo de censura. Essa correlação implica que o custo total não deve ser analisado utilizando técnicas de análise de sobrevivência, pois todas tem como suposição a independência entre a variável de interesse e a variável de censura.

O método *Kaplan-Meier* foi, então, adaptado por Lin (1997) para estimar o custo médio total incorporando dados censurados, originando o Estimador KMSA (*Kaplan-Meier Sample Average Estimator*). Lin propõe que o período de interesse do estudo seja dividido em intervalos.

Em seguida é avaliado o custo médio dos pacientes não censurados até o fim de cada intervalo. Essas médias são ponderadas pelas estimativas de sobrevivência originando o estimador KMSA.

De maneira geral suponha, por exemplo, um estudo onde n pacientes tenham seus dados de custo coletados em K observações distintas distribuídas ao longo do tempo ($t_1 < t_2 < ... < t_K$). Considere CM_j o custo médio observado no tempo j ($1 \leq j \leq K$). Seja d_j o número de pacientes que morreram no tempo t_j, c_j o número de pacientes censurados no tempo t_j e n_j o número de pacientes que estavam em risco em t_j (não morreram e nem foram censurados até o início do tempo j). A estimativa da probabilidade de sobrevivência ($\widehat{R}(t_j)$) de cada período t_j é apresentada no Quadro 2 a seguir.

Quadro 2: Cálculo das estimativas utilizando o estimador KMSA

t_j	c_j	d_j	n_j (em risco)	$\widehat{R}(t_j)$	Custo médio	Custo médio KMSA
0	0	0	n_0	1	-	-
t_1	c_1	d_1	$n_0 - 0 - 0 = n_1$	$1*((n_0 - 0)/n_0) = r_1$	CM_1	CM_1*r_1
t_2	c_2	d_2	$n_1 - c_1 - d_1 = n_2$	$r_1((n_1-d_1)/n_1) = r_2$	CM_2	CM_2*r_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
t_k	c_k	d_k	$n_{(k-1)} - c_{(k-1)} - d_{(k-1)} = n_k$	$r_{(k-1)}((n_{k-1}-d_{k-1})/n_{k-1}) = r_k$	CM_k	CM_k*r_k

Ou, de forma generalizada:

$$\widehat{R}(t_j) = \left(\frac{n_1 - d_1}{n_1}\right)\left(\frac{n_2 - d_2}{n_2}\right)...\left(\frac{n_{j-1} - d_{j-1}}{n_{j-1}}\right)$$

Essa técnica pode ser facilmente aplicada utilizando diferentes programas estatísticos. Para este trabalho as análises foram feitas utilizando Excel 2010. Maior detalhamento sobre estes cálculos explicados com auxílio de exemplo podem ser encontrados no Anexo 2.

Note que, ao calcular o custo médio de cada período estamos considerando que todos os pacientes que morreram ou foram censurados neste período tiveram como custo o valor desta média. Desta maneira o custo médio está superestimando a estimativa do período. O cálculo da proporção de sobrevivência considera o número de mortes e o número de censuras no período e, portanto, é utilizada como fator de ponderação para corrigir a superestimação.

4.3 Ajuste para censura utilizando IPW

De modo semelhante ao KMSA, o método de ponderação pelo inverso da probabilidade (IPW) também utiliza o período estudado de forma particionada, porém ao invés da média este método utiliza para cada intervalo o custo total do tempo, soma de todos os pacientes observados no tempo. Note que ao somar os custos observados no tempo t_j, a todos os pacientes que não puderam ser observados naquele tempo está sendo atribuído o valor de custo 'zero', o que levaria a uma subestimação do custo médio. É necessário considerar que alguns dos pacientes que não

puderam ser observados faleceram, e nestes casos o custo zero é aplicado corretamente, porém outros dos pacientes não observados são pacientes que foram censurados e que tiveram custo, ainda que não se saiba exatamente qual. Assim o custo total de cada tempo é ponderado utilizando a probabilidade de não ser censurado até o início do referido período.

Considerando o exemplo descrito anteriormente e sendo CT_j o custo total observado ao tempo t_j, o IPW ($\hat{S}(t_j)$) de cada período t_j é apresentado no Quadro 3 a seguir.

Quadro 3: Cálculo das estimativas utilizando o estimador IPW

t_j	c_j	d_j	n_j (em risco)	$\hat{S}(t_j)$	Custo Total	Custo IPW	Custo por paciente
0	0	0	n_0	1	-	-	-
t_1	c_1	d_1	$n_0 - 0 - 0 = n_1$	$((n_0 - 0)/n_0) = s_1 = 1$	CT_1	CT_1/s_1	$\frac{CT_1/s_1}{n_0}$
t_2	c_2	d_2	$n_1 - c_1 - d_1 = n_2$	$s_1((n_1-c_1)/n_1) = s_2$	CT_2	CT_2/s_2	$\frac{CT_2/s_2}{n_0}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
t_k	c_k	d_k	$n_{(k-1)} - c_{(k-1)} - d_{(k-1)} = n_k$	$s_{(k-1)}((n_{(k-1)}-c_{(k-1)})/ n_{(k-1)}) = s_k$	CT_k	CT_k/s_k	$\frac{CT_k/s_k}{n_0}$

Ou, de forma generalizada, $\hat{S}(t_j) = \left(\frac{n_1 - c_1}{n_1}\right)\left(\frac{n_2 - c_2}{n_2}\right) \dots \left(\frac{n_{j-1} - c_{j-1}}{n_{j-1}}\right)$

Depois calculado o IPW de cada tempo t_j estudado, o custo total de cada tempo é dividido pelo IPW deste tempo, obtendo-se assim o custo estimado para todos os pacientes a cada período. É importante ressaltar que, diferentemente do estimador KMSA, esse método não resulta em estimações de custo por paciente, e sim um custo total para todos os pacientes observados na amostra. Para estimar o custo total médio por paciente seria necessário em seguida dividir o custo total encontrado pelo número de participantes do estudo.

4.4 Ajuste para censura utilizando covariáveis

Os métodos KMSA e IPW estimam custo total médio baseado em todos os pacientes que participaram do estudo, porém, não consideram covariáveis que possam ser relevantes na estimação do custo médio. Com o objetivo de estimar o custo total médio considerando diversas covariáveis simultaneamente, uma técnica alternativa foi proposta por Lin (2000), na qual as técnicas familiares de regressão linear e estimação de mínimos quadrados são modificadas através do uso do IPW na estimação de parâmetros dos modelos lineares.

A metodologia proposta pode ser aplicada tanto para estimar o custo total médio (para tanto utilizando apenas os pacientes que tenham dados completos observados), como também para a estimação de múltiplos intervalos de tempo, aplicando a regressão para cada intervalo separadamente e utilizando para cada um deles toda a informação disponível, mesmo de pacientes que tenham sido censurados em períodos subsequentes. A regressão quando aplicada a múltiplos

intervalos de tempo possibilita estimações mais eficientes por não haver perda de informação e pela possibilidade de comparar o efeito das covariáveis no custo total para cada tempo estudado.

Considere então que o período para o qual está sendo avaliado o custo total médio seja particionado em intervalos menores. Para cada período de tempo observado será possível, através de regressão linear ponderada, ajustar equações que permitam estimar o custo médio daquele período de tempo. O modelo é ajustado considerando os pacientes que tiveram custo observado naquele tempo, utilizando covariáveis e ponderando os coeficientes da regressão pela probabilidade de censura. Estimativas deste modelo são, então, utilizadas para imputar os custos de pacientes que foram censurados até este tempo. Por exemplo, suponha que para o tempo 2 (T2) tenha sido ajustado um modelo com as covariáveis sexo e idade. A regressão proporcionaria uma equação linear com um intercepto (um valor de custo comum a todos os pacientes) e coeficientes para as covariáveis sexo e idade, que indicariam as diferenças médias de custo entre homens e mulheres e a variação a cada ano de idade. Assim, todos os pacientes com custo observado no tempo T2 foram utilizados no modelo. E todos os pacientes que não tiveram seus custos observados no tempo T2 por terem sido censurados antes disto (neste caso T1), terão valores imputados utilizando a respectiva estimativa gerada pela equação de regressão. Assim, os valores imputados serão distintos conforme o sexo e a idade do paciente.

Uma característica comum de dados de custo é o fato de sua distribuição ser assimétrica à direita, característica esta devido ao fato de não existirem custos negativos e também pela maior concentração de pacientes com valores baixos (ou até mesmo zero) e poucos pacientes com custos observados maiores. Essa característica prejudica a aplicação dos modelos de regressão linear, pois vai contra algumas suposições básicas do modelo, como a de que os resíduos sigam distribuição normal de média 0 e variâncias homogêneas.

Quando não é possível aplicar a regressão com mínimos quadrados ordinários devido às condições necessárias, uma alternativa cabível é aplicar a ponderação IPW proposta na estimação de coeficientes para modelos lineares generalizados (GLM). De modo bastante semelhante à análise de regressão, os modelos GLM também possibilitam ajustar equações para estimar os custos dos pacientes que foram censurados de acordo com o perfil desse paciente, porém, diferentemente da regressão com mínimos quadrados ordinários, os modelos GLM não tem as mesmas suposições de normalidade e homogeneidade de variâncias. Uma descrição mais detalhada de GLM foge do objetivo deste trabalho. Sugere-se como bibliografia nacional complementar deste assunto o trabalho de Gilberto A. Paula (2012) ou, de literatura internacional, J.K. Lindsey (2007).

5. Análise de dados

O banco de dados utilizado para ilustrar os métodos descritos é adaptado de um estudo real que avalia, através de 5 levantamentos espaçados ao longo de três anos e meio (42 meses), os custos envolvidos em dois tratamentos para PSP (Progressive supranuclear palsy) e MSA (multiple

system atrophy), doenças neurológicas progressivas com tempo de sobrevida curto. Os dados originais foram um pouco alterados de modo a preservar a originalidade de resultados a nível epidemiológico. Porém, tomou-se o cuidado de não alterar a estrutura dos dados de modo a trabalhar com um cenário realístico.

Para identificar preditores que pudessem estar relacionados com o custo direto (despesas envolvidas em medicamentos ou atenção de profissional da saúde) foram avaliadas variáveis sócio demográficas dos pacientes, tais como sexo e idade, e também características clínicas, como o tratamento sendo utilizado, tempo de doença (em dias), gravidade dos sintomas (variável quantitativa avaliada a cada tempo utilizando a escala Parkinson's Plus Symptom de índice global de severidade de doença) e cognição (medida pelo *Mini Mental State Examination* MMSE e separado em dois grupos, um com escores 0 a 27 e outro com escores de 28 a 30). A parte do banco utilizada é composta por 325 pacientes (41,5% do sexo masculino e 58,5% do sexo feminino), dos quais 160 (49,2%) morreram no decorrer do estudo e 53 foram censurados antes do fim do estudo. Apenas 112 pacientes (34,5%) tiveram o custo observado em todos os tempos avaliados (T1: 6 meses, T2: 12 meses, T3: 18 meses, T4: 30 meses e T5: 42 meses após o início do estudo). A Figura 1 a seguir apresenta a linha do tempo com o número de pacientes falecidos e censurados a cada etapa do estudo utilizado no exemplo.

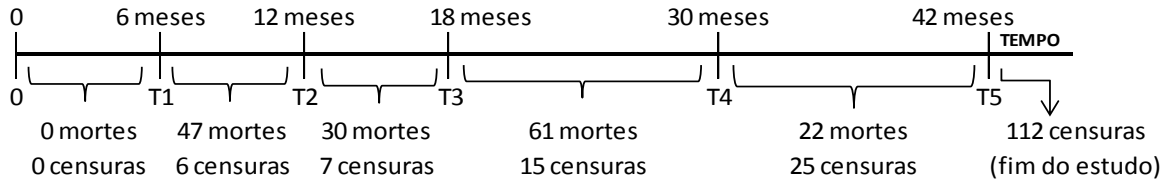

Figura 1: Linha do tempo do exemplo aplicado

À exceção das censuras, a parte do banco de dados utilizada no exemplo não apresentava dados faltantes. De modo a explorar as técnicas de imputação apresentadas neste trabalho optou-se por excluir alguns dados tratando-os como dados faltantes. Então foram selecionados de forma aleatória e proporcional ao número de observações em cada tempo 5% dos valores observados (um total de 54 observações) para exclusão. Estes valores foram imputados utilizando a média de todas as observações em cada tempo. Todas as análises feitas a seguir foram feitas no banco após a imputação. Ao final deste trabalho será feita uma comparação com as mesmas técnicas aplicadas ao banco de dados completo para verificar o impacto que teve a técnica de imputação utilizada sobre os resultados.

O custo total médio estimado para 3 anos e meio de tratamento considerando todos os 325 pacientes e sem ponderar os dados pela presença de censuras é de € 26099,92 (IC 95%: 23138,87;29060,97). O custo estimado utilizando apenas as informações dos não censurados, (casos completos ou pacientes que faleceram no decorrer do estudo) é de € 25584,91 (IC 95%: 22369,54;28800,28). Já o custo estimado considerando apenas os pacientes que tiveram todos os tempos avaliados (casos completos) é de € 33470,95 (IC 95%: 28098,93;38842,97).

A Tabela 2 apresenta o custo total médio estimado para cada período de tempo utilizando o estimador KMSA.

Tabela 2: Aplicação do estimador KMSA aos dados do exemplo

Tempo	Censuras	Mortes	Risco	R(tj)	Custo médio do período (€)	custo MÉDIO KMSA (€)
0	0	0	325	1,00		
T1	6	47	325	1,00	7122,24	7122,24
T2	7	30	272	0,86	6846,51	5856,40
T3	15	61	235	0,76	7293,23	5550,44
T4	25	22	159	0,56	9273,60	5225,62
T5	112	0	112	0,49	9973,96	4842,62

Com base nas estimativas apresentadas na Tabela 2, o custo total médio estimado para três anos e meio de tratamento médico pelo estimador KMSA é de € 28597,32. Note que quanto maior o tempo t observado, maior estará sendo o viés cometido se a média for utilizada, pois maior será o número de pacientes que já morreram em períodos anteriores para os quais o custo verdadeiro não é a média dos observados e sim o custo zero. O estimador KMSA reconhece que conforme o tempo avança menor é a probabilidade de sobrevivência do paciente, de forma que os dados estão ponderados corretamente (a diferença entre o total médio estimado e o total após a ponderação é maior para períodos de tempo mais avançados).

O custo total médio estimado para cada período utilizando o estimador IPW é apresentado na Tabela 3.

Tabela 3: Aplicação do estimador IPW aos dados do exemplo

Tempo	Censuras	Mortes	Risco	IPW	Custo Total (€)	Custo IPW (€)	Custo por paciente (€)
0	0	0	325	1,00			
T1	6	47	325	1,00	2314729	2314729	7122,24
T2	7	30	272	0,98	1862252	1897279	5837,78
T3	15	61	235	0,96	1713909	1792270	5514,68
T4	25	22	159	0,90	1474502	1647048	5067,84
T5	112	0	112	0,75	1117083	1480603	4555,70

No caso do IPW a ponderação é feita não no custo médio e sim no custo total de cada período (soma de todas as observações). Conforme esperado, a probabilidade de não ter sido censurado em tempos mais avançados é menor do que nos tempos iniciais. O custo por paciente estimado pelo IPW para o total de três anos e meio de tratamento é € 28098,24.

As estimações de custo feitas através da ponderação dos dados pelos coeficientes KMSA e IPW não consideram, no entanto, covariáveis de perfil clínico e demográfico do paciente. Para considerar essas covariáveis na estimação do custo é necessário aplicar a técnica de regressão linear ponderada pelo IPW conforme proposto por Lin (2000). Como é comum acontecer com dados de custo, não foi possível aplicar a técnica de regressão, pois as condições necessárias para a aplicação

da técnica não foram satisfeitas. Como alternativa foram aplicados modelos lineares generalizados (GLM). Tanto a técnica de regressão quanto os modelos lineares generalizados apresentam maior nível de dificuldade teórico e computacional quando comparadas com as técnicas dos estimadores KMSA e IPW apresentadas anteriormente. O programa estatístico Stata 11.0 tem os recursos necessários e com ele é possível proceder a essas análises. Mais detalhes a respeito de como proceder a análise computacionalmente utilizando o Stata 11.0 são apresentadas no Anexo 4 (análise de regressão linear com ponderação IPW) e no Anexo 5(Modelos Lineares Generalizados com ponderação IPW).

No exemplo aplicado o modelo melhor ajustado para estimação de custo para todos os tempos analisados foi o modelo GLM utilizando a distribuição Gamma e o link *identity* (identidade, os valores da variável não transformados). A Tabela 4 apresenta os coeficientes e nível de significância dos modelos feitos com cada variável individualmente para cada tempo observado. Esses modelos, porém, não serão utilizados para estimar o custo dos pacientes que foram censurados. O objetivo aqui é apenas observar de maneira geral como as covariáveis influenciam ou não nos custos observados ao longo do tempo.

Tabela 4: Modelos GLM com cada covariável individualmente

COVARIÁVEIS		T1	T2	T3	T4	T5
Tratamento	**Coef.**	246,20	1543,00	2480,83	3575,77	1999,53
	p-valor	0,821	0,156	0,050	0,151	0,330
Sexo	**Coef.**	-2062,91	-3691,46	-3618,15	-1243,67	-2625,18
	p-valor	0,069	0,001	0,008	0,610	0,241
Cognição	**Coef.**	-1964,86	-2430,93	-3372,82	1946,25	-3483,64
	p-valor	0,051	0,020	0,005	0,436	0,087
Idade	**Coef.**	168,35	216,65	287,36	124,52	192,29
	p-valor	0,001	0,000	0,000	0,330	0,265
Tempo doente	**Coef.**	294,67	202,11	77,91	-581,28	-73,05
	p-valor	0,359	0,559	0,838	0,333	0,909
Gravidade	**Coef.**	90,17	54,40	83,49	42,49	108,34
	p-valor	0,000	0,000	0,000	0,285	0,000

É possível observar, por exemplo, que para todos os tempos observados os pacientes de sexo feminino tiveram em média um custo observado inferior aos pacientes de sexo masculino, pois para todos os tempos o coeficiente estimado foi negativo, mas apenas para os custos observados em T2 e T3 essa diferença foi estatisticamente significativa a 5% de nível de significância, ou seja, apenas para esses dois tempos pode-se afirmar que o sexo do paciente influenciou no custo observado. A covariável gravidade apresentou coeficientes positivos em todos os tempos, e foi significativa para os tempos T1, T2, T3 e T5. Isso significa que a gravidade da doença influencia de forma significativa no custo observado, e quanto maior for a gravidade maior tende a ser o custo de tratamento do paciente.

A estratégia de modelagem adotada para obter as equações utilizadas na estimação de custo dos pacientes que tiveram seus dados censurados foi aplicar a técnica hierárquica com processo

retrospectivo. Isso significa que, para cada tempo, foi ajustado um modelo considerando apenas as variáveis que tenham sido estatisticamente significativas nos modelos individuais a 10% de nível de significância. As variáveis que deixaram de ser significativas foram retiradas do modelo de forma retrospectiva, ou seja, eliminando uma a uma do p-valor maior para o menor até que todas as variáveis remanescentes no modelo fossem significativas a 5% de nível de significância. A Tabela 5 apresenta os modelos resultantes com os coeficientes estimados e o seu respectivo nível de significância para cada tempo.

Tabela 5: Modelos GLM finais (banco com dados faltantes imputados)

COVARIÁVEIS		T1	T2	T3	T5
Tratamento	**Coef.**	-	-	1958,28	
	p-valor	-	-	0,008	
Sexo	**Coef.**	-	-2796,69	-2773,83	-
	p-valor	-	0,021	0,017	-
Idade	**Coef.**	-	119,90	95,66	-
	p-valor	-	0,003	0,000	-
Cognição	**Coef.**	-	-	-853,75	-4014,68
	p-valor	-	-	0,040	0,016
Gravidade	**Coef.**	90,16	37,77	42,34	88,67
	p-valor	0,000	0,000	0,000	0,000
Constante	**Coef.**	-1263,02	-3599,98	-3260,18	1093,31
	p-valor	0,116	0,249	0,108	0,597

Para obter a estimativa de custo a ser utilizada para imputar o custo dos pacientes que tiveram seus dados censurados é necessário aplicar estas equações aos dados de cada paciente. No tempo T1 todos os pacientes puderam ter seus custos observados, não havendo necessidade de utilizar a equação para imputar estes valores. Note apenas que o custo no T1 foi bastante influenciado pela variável gravidade, mas não teve influencia significativa de nenhuma das demais covariáveis. Para o tempo T4 não foi possível ajustar um modelo, nenhuma das covariáveis foi significativa. Isso significa que para o tempo T4 os custos observados não foram diferentes para pacientes de diferentes gêneros, ou sob diferentes tratamentos e nem tiveram relação com a idade, gravidade da doença ou quaisquer das demais variáveis. Então para esse tempo os pacientes que não tiveram seu custo observado por terem sido censurados em períodos anteriores tiveram seu custo estimado pela média dos pacientes que puderam ter o custo observado devidamente. Os demais pacientes censurados tiveram seu custo calculado a cada tempo de acordo com as equações:

$$custo\ T2 = -2796{,}69 * Sexo + 119{,}90 * Idade + 37{,}77 * Grav. - 3599{,}98$$

$$custo\ T3 = 1958{,}28 * trat - 2773{,}83 * sexo - 853.75 * cogn + 95{,}66 * idade + 42{,}35 * grav. - 3260{,}18$$

$$custo\ T5 = -4014{,}68 * cogn + 88{,}67 * grav. + 1093.31$$

A variável categórica (sexo) precisa estar codificada em 0 e 1 (neste exemplo, 1 corresponde a mulheres e 0 corresponde aos homens). Assim, por exemplo, o custo estimado para o tempo T2 de uma mulher com 70 anos e gravidade da doença avaliada em 150 seria € 7661,83 (custo T2 = -2796,69*1 + 119,90*70 + 37,77*150-3599,98).

Depois de calculado cada custo que não pode ser observado para todos os pacientes censurados o banco está completo (livre de dados faltantes e censuras) e uma análise descritiva direta com todos os pacientes resultará em uma estimação de custo total médio por paciente mais próximo da realidade. O custo estimado por paciente para 3 anos e meio de tratamento após essa análise foi de € 29099,19.

Os resultados de custo obtidos pelos métodos utilizados estão resumidos na Tabela 6.

Tabela 6: Custo total estimado por diferentes métodos (banco com dados faltantes imputados)

	MÉTODO	CUSTO TOTAL ESTIMADO (€)
ANÁLISE NÃO AJUSTADA DE DADOS	MÉDIA TODOS OS PACIENTES	26099,92 (23138,87 ; 29060,97)
	MÉDIA NÃO CENSURADOS	25584,91 (22369,54 ; 28800,28)
	MÉDIA CASOS COMPLETOS	33470,95 (28098,93 ; 38842,97)
ANÁLISE COM PONDERAÇÃO DE CENSURAS	ESTIMADOR KMSA	28597,32
	ESTIMADOR IPW	28098,24
	GLM	29099,19

Pela análise da tabela é possível perceber que os métodos com ponderação propostos tiveram desempenhos semelhantes e os custos calculados para três anos e meio de tratamento (total) foram bastante próximos. Conforme esperado os métodos com ponderação apresentaram custos de valor maior que os observados nas técnicas que consideram a média de todos os pacientes e a média dos pacientes não censurados, as quais se supõe estarem subestimando o custo total. Note também que os custos estimados com dados ponderados foram inferiores à média de casos completos, que se acredita superestimar o custo.

Para verificar o desempenho da técnica de imputação de dados faltantes aplicada (média dos custos de todos os pacientes observados para cada tempo), as mesmas técnicas de estimação de custos foram aplicadas no banco de dados completo. Os modelos GLM obtidos a partir da análise do banco sem dados faltantes estão apresentados na Tabela 7.

Tabela 7: Modelos GLM finais (banco sem dados faltantes)

COVARIÁVEIS		T1	T2	T3	T5
Tratamento	**Coef.**	-	-	2634,98	-
	p-valor	-	-	0,019	-
Sexo	**Coef.**	-	-2652,06	-3334,17	-
	p-valor	-	0,034*	0,023	-
Idade	**Coef.**	-	143,34	100,86	-
	p-valor	-	0,000*	0,012	-
Gravidade	**Coef.**	108,50	51,49	39,53	112,31
	p-valor	0,000*	0,000*	0,000	0,000*
Constante	**Coef.**	-2498,73	-6551,24	-3292,97	-3718,89
	p-valor	0,000*	0,017*	0,304	0,000*

A presença de dados imputados influenciou as variáveis que foram significativas em cada modelo. A covariável cognição foi estatisticamente significativa para os tempos T3 e T5 no banco de dados com observações faltantes imputadas, mas não foi significativa no banco de dados completo (original, com os valores reais). Também apareceram algumas pequenas diferenças nos coeficientes estimados para as variáveis que foram estatisticamente significativas em ambos os bancos de dados. A Tabela 8 a seguir apresenta um resumo dos resultados da aplicação das técnicas propostas no banco de dados completo, sem a presença de dados faltantes.

Tabela 8: Custo total estimado por diferentes métodos (banco sem dados faltantes)

	MÉTODO	CUSTO TOTAL ESTIMADO (€)
ANÁLISE NÃO AJUSTADA DE DADOS	MÉDIA TODOS OS PACIENTES	29996,44 (26732,00 ; 33260,88)
	MÉDIA NÃO CENSURADOS	26146,40 (22756,79 ; 29536,00)
	MÉDIA CASOS COMPLETOS	33961,26 (28187,07 ; 39735,47)
ANÁLISE COM PONDERAÇÃO DE CENSURAS	ESTIMADOR KMSA	29628,87
	ESTIMADOR IPW	29123,95
	GLM	30034,15

Os custos calculados pelo banco de dados com a presença de dados faltantes foram um pouco menores do que os custos calculados pelo banco de dados sem dados faltantes, mas os intervalos de confiança se sobrepõem, ou seja, essa diferença não foi significativa. Apesar das diferenças observadas nos coeficientes e nas variáveis estatisticamente significativas dos modelos de GLM, o custo total calculado pelo GLM foi semelhante em ambos os bancos, assim como os valores das demais técnicas. O método de imputação pela média teve desempenho satisfatório em supor os valores de custo que não puderam ser observados.

6. Considerações Finais

Este trabalho ilustrou através do uso de teoria e com auxílio de um exemplo aplicado algumas das técnicas mais utilizadas para a correção de dados faltantes e dados censurados em estudos longitudinais para estimação de custos na área da saúde. Foram apresentados alguns métodos simples de imputação de dados faltantes baseados em dados de outros pacientes da amostra e também técnicas alternativas utilizando apenas informações de custo que tenham sido observadas ao próprio paciente para o qual está sendo feita a imputação, em períodos anteriores e/ou posteriores ao tempo para o qual o custo está sendo imputado.

Para tratamento de dados censurados foram apresentadas técnicas tais como o estimador KMSA e o estimador IPW, ambas de aplicação simples, porém incapazes de considerar covariáveis de perfil do paciente que possam ser relevantes na estimação de custos. Este trabalho também apresentou a técnica de regressão linear ponderada pelo IPW proposta por Lin (2000), uma técnica mais complexa, mas capaz de considerar diversas covariáveis simultaneamente. Devido à dificuldade de se aplicar regressão linear em dados de custo (com frequência dados de custo

apresentam assimetria e violam as suposições necessárias à aplicação do modelo de regressão), este trabalho propôs a utilização de Modelos Lineares Generalizados ao invés de modelos lineares ordinais.

Um exemplo aplicado foi apresentado utilizando dados adaptados de um estudo longitudinal real que teve por objetivo estimar o custo total médio de três anos e meio de tratamento de PSP (*Progressive supranuclear palsy*) e MSA (*multiple system atrophy*). Vale ressaltar que os dados utilizados neste trabalho foram alterados em relação ao estudo origina e que a utilização deles neste trabalho foi realizada apenas para fins ilustrativos. Com o auxílio do exemplo foi possível evidenciar a importância de se escolher cuidadosamente a técnica de imputação de dados faltantes e de se ponderar os dados censurados para evitar subestimação ou superestimação dos custos. Este exemplo foi utilizado para detalhar as técnicas apresentadas e mostrar como elas podem ser replicadas computacionalmente com os pacotes estatísticos Excel 2010 e Stata versão 11.0

Neste exemplo as técnicas com ponderação apresentaram desempenhos semelhantes ao estimar o custo total médio de três anos e meio de tratamento. Todas as três técnicas, no entanto, tem como limitação o fato de fazerem estimativas pontuais, não há medida de incerteza, variabilidade ou intervalos de confiança descritos na literatura existente acerca destas técnicas atualmente. Importante ressaltar que, apesar de observada a proximidade dos custos estimados por diferentes técnicas, não é intenção deste trabalho comparar o desemprenho das técnicas entre si. Essa postura foi adotada devido ao fato de não haver técnicas melhores ou piores, o desempenho de cada uma depende de características que são específicas de cada estudo e cada de banco de dados.

Replicando as análises efetuadas em dois bancos de dados, um com dados completos (à exceção das censuras) e outro com presença de dados faltantes, foi possível verificar que os resultados obtidos foram próximos e a técnica de imputação de dados pela média de todos os pacientes observados teve desempenho satisfatório. Mas, como já mencionado, neste trabalho o exemplo foi aplicado para fins ilustrativos. Para um estudo mais completo cujo interesse fosse estudar de forma detalhada o custo envolvido no tratamento das doenças PSP e MSA e os efeitos de covariáveis, seria necessário efetuar uma análise de sensibilidade, ou seja, replicar as técnicas de ponderação de dados censurados em diversos bancos, cada um tendo utilizado uma técnica diferente de imputação dos dados faltantes, a fim de verificar o impacto da técnica de imputação utilizada sobre os resultados obtidos.

7. Referências Bibliográficas

BANG H.; TSIATIS A.A. Estimating medical costs with censored data. *Biometrika*, 2000, 87,2,pp.329-343

BASTOS, J.; ROCHA, C.; Análise de sobrevivência Conceitos Básicos. 2006. Arquivos de Medicina, Vol. 20, Nº 5/6.

BRIGGS, A.; Economics notes: handling uncertainty in economic evaluation. *British Medical Journal*, 1999. 319, 120.

BRIGGS, A.; CLARK, T.; WOLSTENHOLME, J.; CLARKE, P. Missing.... presumed at random: cost-analysis of incomplete data. *Health Economics*, 2002. 12(5):pp. 377-392.

BRIGGS, A.H.; GRAY, A.M.; Handling uncertainty in economic evaluations of healthcare interventions. *British Medical Journal*, 1999a. 319, 635-8.

BRIGGS, A.H.; GRAY, A.M.; Handling uncertainty when performing economic evaluation of healthcare interventions. *Health Technology Assessment*, 1999b. 3, 1-134.

BRIGGS, A.; SCUMPHER,M.; BUXTON,M. ; Uncertainty in the economic evaluation of health care technologies: the role of sensitivity analysis. *Health Economics*, 1994. 3, 95-104.

ENGELS JM, DIEHR P. Imputation of missing longitudinal data: a comparison of methods, *Journal of Clinical Epidemiology*, 2003; 56(10):968-76.

FENN, P.; MCGUIRE, A.; PHILLIPS, V.; BACKHOUSE,M.; JONES, D.; The analysis of censored treatment cost data in economic evaluation. *Medical Care*, 1995. 33, 851-863.

GRAY,A.M.; CLARKE, P.M.; WOLSTENHOLME, J.L.; WORDSSWORTH,S. Applied Methods of Cost-effectiveness Analysis in Health Care. Editora Oxford, 2011.

HIATT, R.A.; QUESENBERRY, C.P.; SELBY, J.V.; FIREMAN, B.H.; KNIGHT, A.; The cost of acquired immunodeficiency syndrome in Northern California: The experience of a large prepaid health plan. *Archives of International Medicine*, 1990. 150, 833-838.

KAPLAN, E. L.; MEIER, P.; Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 1958. 53, 457-48 1.

LIN, D.Y. Linear regression analysis of censored medical costs. *Biostatistics*, 2000;1,1,pp.35-47

LIN, D.Y.; FEUER,E.J.; ETZIONI,R; WAX,Y. Estimating medical costs from incomplete follow-up data. *Biometrics*,1997; 53:419-434.

LINDSEY, J.K. Applying generalized linear models. Limburgs Universitair Centrum, Diepenbeek, 2007.

NUNES, L. N.; Métodos de imputação de dados aplicados na área da saúde. Tese de doutorado em medicina: Epidemiologia. Faculdade de medicina. Universidade federal do Rio Grande do Sul, Porto Alegre. 2007.

PAULA, G.A., Modelos de regressão com apoio computacional. São Paulo: IME/USP, 2012.

QUESENBERRY, C.P.; FIREMAN, B.; HIATT, R.A.; SELBY, J.V.; A survival analysis of hospitalization among patients with acquired immunodeficiency syndrome. American Journal of Public Health, 1989. 79, 1643-1647.

RUBIN,D.B. Inference and missing data. *Biometrika*, 1976. 63,3,*pp*.581-592.

RUBIN, D.B.; Multiple Imputation for Nonresponde in Surrveys. New York,: Wiley (1987)

Anexos

1. Banco de Dados

Os dados de custo observados formam um banco de dados que deve ser em formato de matriz onde as linhas representam os pacientes e as colunas os tempos observados. Para o exemplo o banco de dados terá 325 linhas e 5 colunas (além das colunas de covariáveis). A coluna 1 representa os dados de custo observados a cada paciente nos primeiros 6 meses do estudo (T1). A coluna 2 representa os custos observados no período entre 6 e 12 meses de estudo (T2). A coluna 3 representa os custos observados no período entre 12 e 18 meses de estudo (T3), a coluna 4 os custos observados no período entre 18 e 30 meses de estudo (T4) e a coluna 5 os custos observados entre 30 e 42 meses após o início do estudo (T5). A Tabela 9 apresenta as primeiras 10 linhas do banco de dados. Para a montagem do banco de dados foi utilizado o Excel versão 2010.

Tabela 9: Exemplo banco de dados

PACIENTE	T1	T2	T3	T4	T5	TOTAL
1	1478,51	5148,18				**6626,69**
2	4412,55	1782,07	13857,15			**20051,77**
3	-	23932,84	13912,31	14312,30	19072,04	**71229,49**
4	2066,88	4718,00	5431,13	-		**12216,01**
5	1564,23	3364,20	2626,30	-		**7554,73**
6	18608,80	32303,62	3187,91	17390,75	12344,69	**83835,77**
7	5292,67	-	17263,55			**22556,22**
8	4596,78	263,58	241,40			**5101,76**
9	1073,02	1306,92	2270,26	4255,64	1737,06	**10642,90**
10	7041,11	12459,67	24216,51			**43717,29**
⋮	⋮	⋮	⋮	⋮	⋮	⋮
TOTAL (soma)	**2314728,93**	**1862251,94**	**1713908,99**	**1474501,75**	**1117083,16**	**8482474,77**
MÉDIA	**7122,24**	**6846,51**	**7293,23**	**9273,60**	**9973,96**	**26099,92**

2. Cálculo do Estimador Kaplan-Meier Sample Average

Para calcular o estimador *Kaplan Meier Sample Average* é necessário primeiramente identificar no banco de dados para cada tempo t_j observado qual o número de pacientes tiveram seu custo observado no início do tempo (em risco), quantos desses morreram (d_j) e quantos foram censurados antes do tempo seguinte. Um esquema destas informações é apresentado na Figura 2.

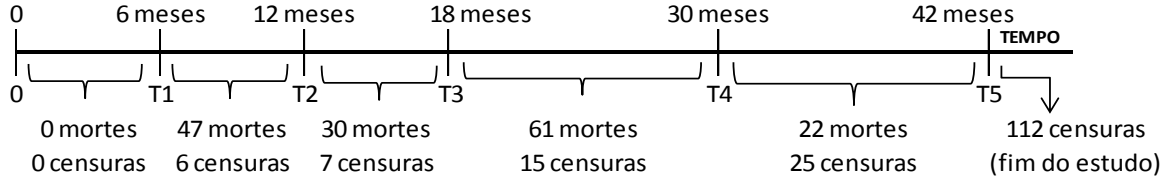

Figura 2: Linha do tempo do exemplo aplicado

Estas informações devem ser digitadas em uma tabela conforme a Tabela 10:

Tabela 10: Exemplo generalizado

$\mathbf{t_j}$	$\mathbf{c_j}$ (número de censuras)	$\mathbf{d_j}$ (número de mortes)	$\mathbf{n_j}$ (número em risco)
0	0	0	n_0
t_1	c_1	d_1	$n_0 - 0 - 0 = n_1$
t_2	c_2	d_2	$n_1 - c_1 - d_1 = n_2$
⋮	⋮	⋮	⋮
t_k	c_k	d_k	$n_{(k-1)} - c_{(k-1)} - d_{(k-1)} = n_k$

A Tabela 11 apresenta as informações observadas no exemplo aplicado

Tabela 11: Informações do exemplo aplicado

tj	cj(censuras)	dj(falhas ou mortes)	nj (risco)
0	0	0	325
T1	6	47	325
T2	7	30	272
T3	15	61	235
T4	25	22	159
T5	112	0	112

Avaliando os dados da Tabela 11 é possível observar que o estudo iniciou acompanhando um número total de 325 pacientes, ou seja, no tempo T1 foram observados 325 valores de custo. No decorrer dos próximos 6 meses (intervalo de tempo entre T1 e T2) observa-se que 47 dos pacientes morreram em função da doença e 6 deles saíram do estudo (deixaram de comparecer as consultas ou morreram por causas externas a doença estudada). Deste modo, no tempo 2 (T2) o número de pacientes observados passou a ser 272 (325 – 6 – 47 = 272). Nos próximos 6 meses (de 1 ano a 1 ano e meio a partir do início do estudo) mais 30 pacientes morreram e outros 7 foram censurados, de modo que para o T3 o número de pacientes observados passou a ser 235 (272 – 30 – 7 = 235). Entre os tempos T3 e T4 foram observados 61 falecimentos e 15 casos de censura, totalizando ao tempo T4 159 observações de custo. Entre o tempo T4 e a observação final (T5) 25 pacientes foram censurados e outros 22 faleceram, se modo que ao término do estudo (na última avaliação, obtida aos 42 meses) apenas o custo de 112 pacientes pode ser observado, dos demais pacientes 53 foram censurados e 160 morreram antes de chegar aos 42 meses. No final do estudo todos os pacientes que ainda não morreram ou foram censurados em períodos anteriores são censurados visto que não serão mais observados os custos.

A partir dessas informações calcula-se a proporção de sobrevivência ($\hat{R}(tj)$) e a estimativa KMSA para o custo médio para cada tempo avaliado no exemplo (Tabela 12).

Tabela 12: KMSA aplicado ao exemplo

Tempo	Censuras	Mortes	Risco	$\hat{R}(tj)$	Custo médio do período (€)	custo MÉDIO KMSA (€)
0	0	0	325	1,00		
T1	6	47	325	1,00	7122,24	7122,24
T2	7	30	272	0,86	6846,51	5856,40
T3	15	61	235	0,76	7293,23	5550,44
T4	25	22	159	0,56	9273,60	5225,62
T5	112	0	112	0,49	9973,96	4842,62

Note na Tabela 12 que para o tempo 0 (que antecede o início do estudo) a proporção de sobrevivência é igual a 1 (isso acontecerá em todos os estudos). Como no tempo T1 o custo de todos os pacientes foi observado, o custo médio deste tempo não precisa de ponderação, e a probabilidade de sobrevivência continua sendo 1. As fórmulas a seguir apresentam um detalhamento de como foram obtidas as estimativas KMSA para cada tempo.

$$\hat{R}(t1) = \hat{R}(t0)\left[\frac{(n_0 - d_0)}{n_0}\right] = 1{,}00\left[\frac{(325 - 0)}{325}\right] = 1{,}00$$

$$\hat{R}(t2) = \hat{R}(t1)\left[\frac{(n_1 - d_1)}{n_1}\right] = 1{,}00\left[\frac{(325 - 47)}{325}\right] = 0{,}86$$

$$\hat{R}(t3) = \hat{R}(t2)\left[\frac{(n_2 - d_2)}{n_2}\right] = 0{,}86\left[\frac{(272 - 30)}{272}\right] = 0{,}76$$

$$\hat{R}(t4) = \hat{R}(t3)\left[\frac{(n_3 - d_3)}{n_3}\right] = 0{,}76\left[\frac{(235 - 61)}{235}\right] = 0{,}56$$

$$\hat{R}(t5) = \hat{R}(t4)\left[\frac{(n_4 - d_4)}{n_4}\right] = 0{,}56\left[\frac{(159 - 22)}{159}\right] = 0{,}49$$

$$Custo\ KMSA\ (t1) = custo\ médio\ (t1) * \hat{R}(t1) = 7122{,}24 * 1{,}00 = 7122{,}24$$

$$Custo\ KMSA\ (t2) = custo\ médio\ (t2) * \hat{R}(t2) = 6846{,}51 * 0{,}86 = 5856{,}40$$

$$Custo\ KMSA\ (t3) = custo\ médio\ (t3) * \hat{R}(t3) = 7293{,}23 * 0{,}76 = 5550{,}44$$

$$Custo\ KMSA\ (t4) = custo\ médio\ (t4) * \hat{R}(t4) = 9273{,}60 * 0{,}56 = 5225{,}62$$

$$Custo\ KMSA\ (t5) = custo\ médio\ (t5) * \hat{R}(t5) = 9973{,}96 * 0{,}49 = 4842{,}62$$

$$Custo\ médio\ por\ paciente = 7122{,}24 + 5856{,}40 + 5550{,}44 + 5225{,}62 + 4842{,}62$$

$$Custo\ médio\ por\ paciente = 28597{,}32$$

Note que quanto maior o tempo t observado, maior seria o erro de superestimação, pois maior seria o número de pacientes que já morreram em períodos anteriores para os quais o custo verdadeiro não é a média dos observados e sim o custo zero. Como se pode perceber o estimador KMSA

reconhece que conforme o tempo avança menor é a probabilidade de sobrevivência do paciente, de forma que os dados estão ponderados corretamente (a diferença entre o total médio estimado e o total após a ponderação é maior para períodos de tempo mais avançados).

3. Cálculo do estimador IPW (Inverse Probability Weighting)

O início do cálculo do IPW consiste em identificar no banco de dados para cada tempo t_j qual o número de pacientes tiveram seu custo observado no início do tempo (em risco), quantos desses morreram d_j e quantos foram censurados antes do tempo seguinte, igual feito para o cálculo do estimador KMSA (na Tabela 10).

A partir dessas informações foram calculadas as estimativas de IPW e de custo total e ponderado para cada tempo avaliado no exemplo (Tabela 13).

Tabela 13: IPW aplicado ao exemplo

Tempo	Censuras	Mortes	Risco	IPW ($\hat{S}(tj)$)	Custo Total (€)	Custo IPW (€)	Custo por paciente (€)
0	0	0	325	1,00			
T1	6	47	325	1,00	2314729	2314729	7122,24
T2	7	30	272	0,98	1862252	1897279	5837,78
T3	15	61	235	0,96	1713909	1792270	5514,68
T4	25	22	159	0,90	1474502	1647048	5067,84
T5	112	0	112	0,75	1117083	1480603	4555,70

Note na Tabela 13 que para o tempo 0 (que antecede o início do estudo) o valor do IPW é igual a 1 (isso acontecerá em todos os estudos). Como no tempo T1 o custo de todos os pacientes foi observado, o custo médio deste tempo não precisa de ponderação, e o valor do IPW continua sendo 1. As fórmulas a seguir apresentam um detalhamento de como foram obtidas as estimativas do custo IPW para cada tempo.

$$\hat{S}(t1) = \hat{S}(t0)\left[\frac{(n_0 - c_0)}{n_0}\right] = 1{,}00\left[\frac{(325 - 0)}{325}\right] = 1{,}00$$

$$\hat{S}(t2) = \hat{S}(t1)\left[\frac{(n_1 - c_1)}{n_1}\right] = 1{,}00\left[\frac{(325 - 6)}{325}\right] = 0{,}98$$

$$\hat{S}(t3) = \hat{S}(t2)\left[\frac{(n_2 - c_2)}{n_2}\right] = 0{,}98\left[\frac{(272 - 7)}{272}\right] = 0{,}96$$

$$\hat{S}(t4) = \hat{S}(t3)\left[\frac{(n_3 - c_3)}{n_3}\right] = 0{,}96\left[\frac{(235 - 15)}{235}\right] = 0{,}90$$

$$\hat{S}(t5) = \hat{S}(t4)\left[\frac{(n_4 - c_4)}{n_4}\right] = 0{,}90\left[\frac{(159 - 25)}{159}\right] = 0{,}75$$

$$Custo\ IPW\ (t1) = \frac{custo\ total(t1)}{\hat{S}(t1)} = \frac{2314729}{1{,}00} = 2314729$$

$$Custo\ IPW\ (t2) = \frac{custo\ total(t2)}{\hat{S}(t2)} = \frac{1862252}{0{,}98} = 1897279$$

$$Custo\ IPW\ (t3) = \frac{custo\ total(t3)}{\hat{S}(t3)} = \frac{1713909}{0{,}96} = 1792270$$

$$Custo\ IPW\ (t4) = \frac{custo\ total(t4)}{\hat{S}(t4)} = \frac{1474502}{0{,}90} = 1647048$$

$$Custo\ IPW\ (t5) = \frac{custo\ total(t5)}{\hat{S}(t5)} = \frac{1117083}{0{,}75} = 1480603$$

$$Custo\ médio\ por\ paciente = \frac{soma\ dos\ custos\ IPW}{número\ de\ pacientes\ no\ estudo}$$

$$Custo\ médio\ por\ paciente = \frac{2314729 + 1897279 + \ 1792270 + 1647048 + \ 1480603}{325}$$

$$Custo\ médio\ por\ paciente = \ 28098{,}24$$

Note que apenas o custo total (soma) dos pacientes que foram observados estaria subestimando o custo total, pois implicaria que os pacientes que não tiveram custo observado apresentaram o valor zero, o que está errado dado que alguns pacientes foram censurados e, portanto, continuam tendo custo diferente de zero, ainda que não se saiba exatamente qual. Assim o valor ponderado deve ser maior do que o observado.

Quanto maior o tempo t observado, maior seria o viés subestimando o custo, pois maior o número de pacientes censurados em períodos anteriores para os quais o custo verdadeiro não é zero. Como pode-se perceber o estimador IPW reconhece que conforme o tempo avança maior é a probabilidade de censura do paciente, de forma que os dados estão ponderados corretamente (a diferença entre o total e o total após a ponderação é maior para períodos de tempo mais avançados).

4. Análise de Regressão

As análises de regressão foram realizadas utilizando o pacote estatístico Stata 11.0.

Passo 1: O primeiro passo para aplicar os modelos regressão é a criação das variáveis ponderadoras chamadas aqui de peso. Para cada tempo observado é criada uma variável peso. O valor dessas variáveis será zero, caso o paciente não tenha custo observado no tempo referido, ou $\frac{1}{IPW(T^*)}$, ou seja, 1 dividido pelo IPW do tempo em que o paciente saiu do estudo, seja por censura ou por morte. Assim cada paciente vai apresentar valores de 'peso' diferentes de acordo com o tempo em que foi censurado ou faleceu.

Passo 2: A forma mais simples de importar o banco de dados para o Stata a partir da planilha Excel é utilizando o editor. No Stata, menu *Data* > *Data Editor* > *Data Editor (Edit)*. Nessa interface é possível digitar diretamente os dados ou então colar os dados copiados do programa Excel (Ctrl C + Ctrl V). O programa abrirá então uma janela na qual se deve informar se o banco de dados colado possui ou não o nome das variáveis listados na primeira linha.

Os dados de custo formam um banco de dados que deve ser em formato de matriz onde as linhas representam os pacientes e as colunas os tempos observados, os pesos calculados, e as covariáveis observadas. Para facilitar o cálculo dos pesos é importante acrescentar também uma coluna com o tempo de morte ou censura do paciente.

Tabela 14: Exemplo banco de dados

PACIENTE	T1	T2	...	EVENTO	Peso T1	Peso T2	...	SEXO	IDADE
1	1478,51	5148,18	...	morte T2	1,33	1,33	...	0	70,8
2	4412,55	1782,07	...	morte T3	1,33	1,33	...	1	78,1
3	7122,24	23932,84	...	censura T5	1,33	1,33	...	0	54,6
4	2066,88	4718,00	...	morte T4	1,05	1,05	...	0	67,3
5	1564,23	3364,20	...	censura T4	1,33	1,33	...	1	77,0
6	18608,80	32303,62	...	censura T5	1,02	1,02	...	0	76,3
7	5292,67	6846,51	...	morte T3	1,05	1,05	...	0	74,8
8	4596,78	263,58	...	morte T3	1,33	1,33	...	0	73,7
9	1073,02	1306,92	...	censura T5	1,33	1,33	...	1	70,8
10	7041,11	12459,67	...	morte T2	1,05	1,05	...	0	60,1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Passo 3: Depois que o banco de dados estiver armazenado no programa Stata, antes de começar a modelar é interessante fazer um histograma da variável custo que pretendemos analisar, para verificar a distribuição dos dados. No menu *Graphics* > *Histogram* e na janela que abrir selecionar a variável (no caso c_formal0), com mostrado na Figura 3. A Figura 4 apresenta o histograma resultante dessa operação.

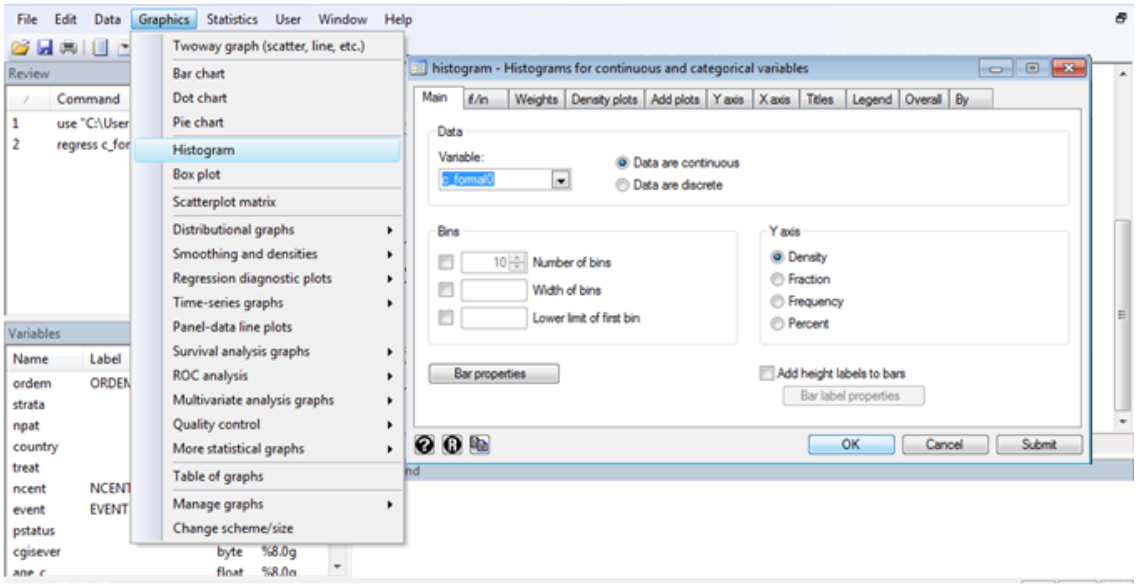

Figura 3: Como fazer um histograma

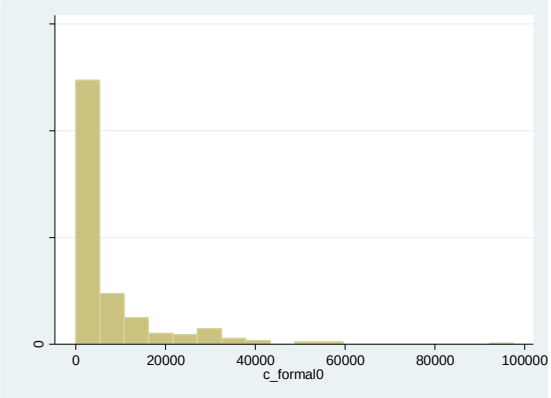

Figura 4: Histograma dados de custo T1

O comando *summarize c_formal0, detail* resume as estatísticas descritivas da variável como mostrado na Figura 5.

```
. summarize c_formal0, detail
```

c_formal0

Percentiles		Smallest		
1%	153.72	40.51		
5%	307.92	106.78		
10%	765.59	146.52	Obs	325
25%	1495.46	153.72	Sum of Wgt.	325
50%	3296.3		Mean	7122.243
		Largest	Std. Dev.	10462.29
75%	7255.39	52109.58		
90%	18182.89	53015.94	Variance	1.09e+08
95%	29658.42	59324.18	Skewness	3.713251
99%	52109.58	97612.29	Kurtosis	23.63536

Figura 5: Medidas descritivas de custo T1

Os dados apresentam uma distribuição bastante inclinada para a direita, com muita concentração de valores menores e alguns poucos valores extremos superiores. É provável que a análise de regressão não seja bem ajustada devido às suposições de normalidade e homogeneidade de variâncias, mas vamos verificar.

A análise de regressão para o tempo inicial do estudo, utilizando como explicativas as variáveis sexo (sex) e gravidade da doença no tempo observado (incl_pps), podem ser feitas utilizando diretamente o comando

regress c_formal0 sex incl_pps [pweight = pesot1] ou então clicando conforme representado na Figura 6, Figura 7 e Figura 8 a seguir:

Menu *Statistics > Linear models and related > Linear Regression*

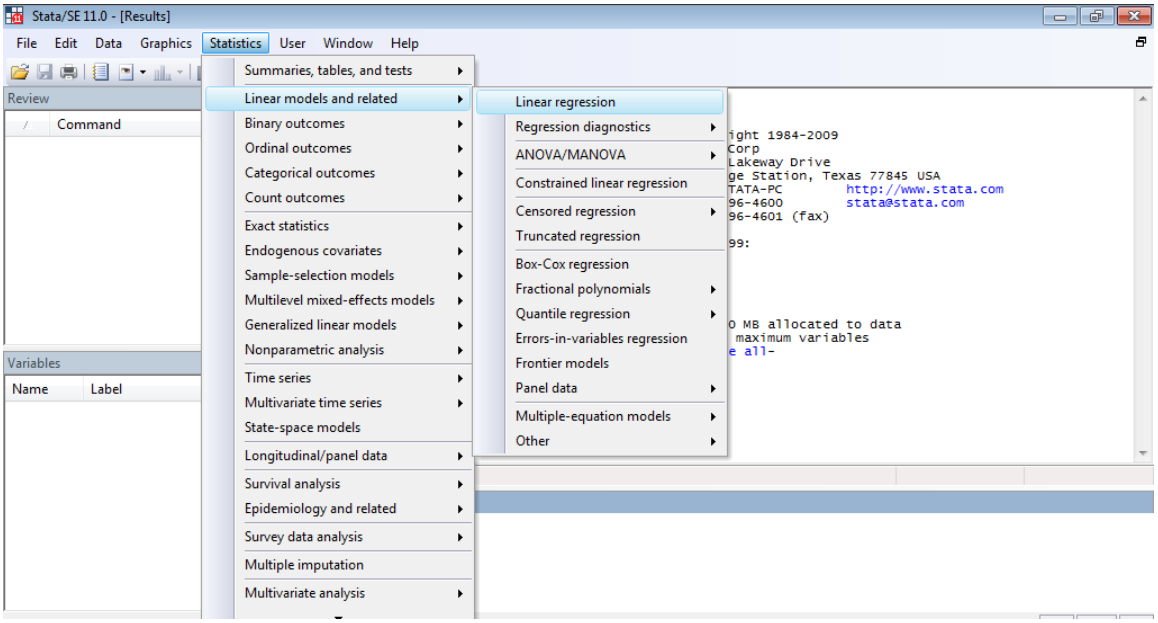

Figura 6: Como fazer regressão linear

Na aba *Model* informar a variável dependente (no caso c_formal0) e também as variáveis explicativas a serem incluídas no modelo:

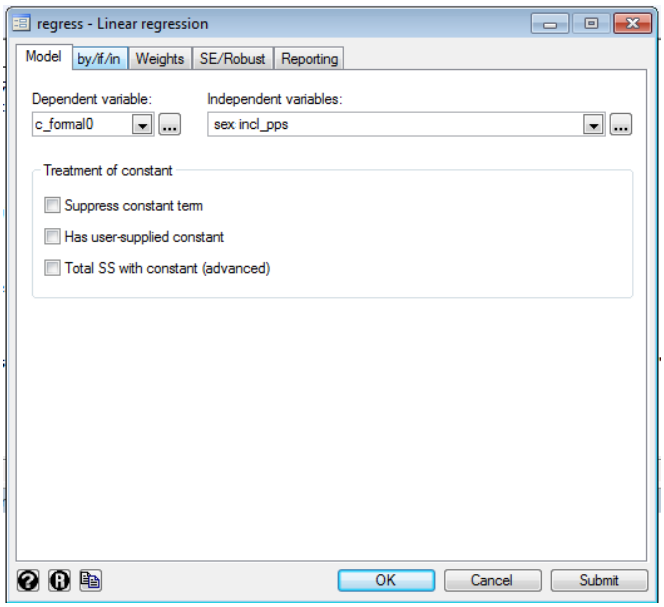

Figura 7: Inserindo as variáveis

Na aba *weights* selecionar a opção 'samplig weights' e em seguida a variável de ponderação referente ao tempo analisado (no caso, pesot1) e em seguida 'OK'.

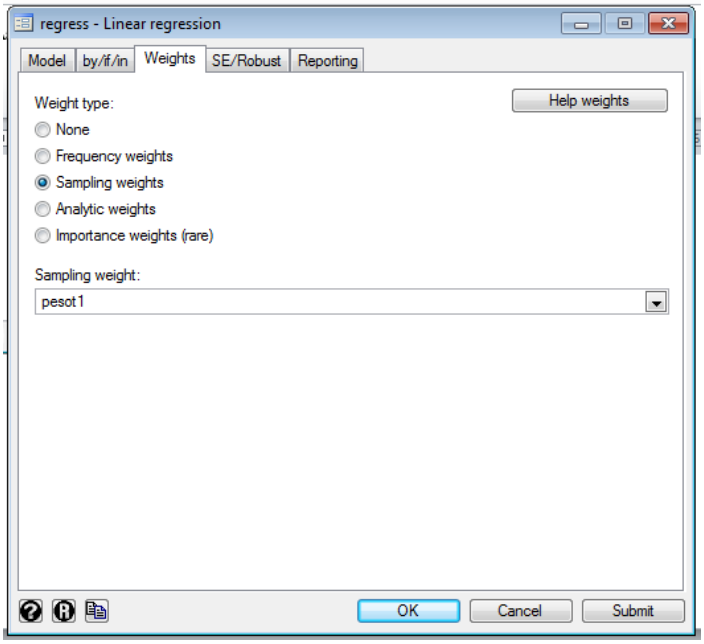

Figura 8: Acrescentando a variável 'peso' ao modelo de regressão

O modelo ajustado é:

```
. regress c_formal0 sex incl_pps [pweight = pesot1]
(sum of wgt is   3.2500e+02)

Linear regression                               Number of obs =     325
                                                F(  2,   322) =   10.75
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1274
                                                Root MSE      =   10600

------------------------------------------------------------------------------
             |               Robust
   c_formal0 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -2681.418   1220.033    -2.20   0.029    -5081.661   -281.1742
     incl_pps |   126.1444   27.57308     4.57   0.000     71.89831    180.3906
       _cons |  -2234.883   2022.317    -1.11   0.270    -6213.506    1743.74
------------------------------------------------------------------------------
```

Figura 9: Modelo de regressão para custo T1 com covariáveis sexo e gravidade

Ou seja, de acordo com o modelo de regressão da Figura 9 o custo formal no tempo inicial dos pacientes poderia ser estimado pela equação

$$custoformal0 = -2681{,}418 * Sex + 126{,}1444 * incl\_pps - 2234{,}883$$

Mas será que esse modelo está bem ajustado? Para verificar se estão satisfeitas as suposições da regressão linear é necessário salvar os resíduos da regressão e testar a normalidade dos mesmos. O gráfico de dispersão dos resíduos *versus* valores preditos também dá uma noção do ajuste do modelo. Os comandos para essas verificações são:

predict yestimado, xb

predict residuos, resid

rvfplot, yline(0)

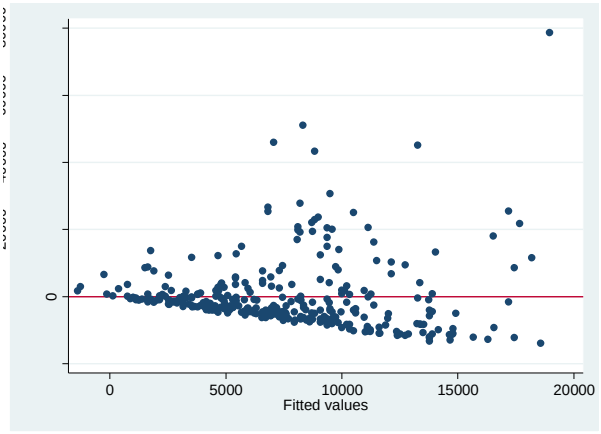

Figura 10: Gráfico de resíduos e preditos, testando suposições da regressão linear

O gráfico da Figura 10 apresenta padrão funil, indicando que o modelo não está bom. Os resíduos não apresentam variância homogênea, para valores preditos mais altos a variabilidade dos resíduos também é maior.

swilk residuos

```
. swilk residuos
```

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
residuos	325	0.74950	57.299	9.539	0.00000

Figura 11: Testando a normalidade dos resíduos

O teste de Shapiro-wilk (Figura 11) rejeitou a hipótese nula de normalidade dos resíduos, ou seja, conforme previsto o modelo de regressão não é bom para estimação dos custos.

Da mesma forma como para o tempo inicial T1, para os demais tempos observados também não foi possível aplicar análise de regressão. Optou-se, então, por utilizar a ponderação IPW proposta em modelos lineares generalizados com a distribuição gamma e o link identity para estimar as equações de previsão para os dados censurados.

5. Modelos Lineares Generalizados

Para o tempo T1 serão apresentadas descrições mais detalhadas a respeito dos procedimentos até o modelo final, para os outros tempos a metodologia é análoga e portanto serão apresentados apenas comandos e resultados.

Tempo 1 (T1)

Inicialmente serão descritos para estimação de custo no tempo T1 modelos com cada covariável individualmente. O primeiro modelo para o tempo T1 (c_formal0) foi ajustado utilizando a variável explicativa gravidade da doença (incl_pps). A análise pode ser feita através do comando

glm c_formal0 incl_pps [pweight = pesot1], family(gamma) link(identity)

ou então clicando conforme apresentado na Figura 12, Figura 13 e Figura 14:

Statistics > Generalized linear models > Generalized linear models (GLM)

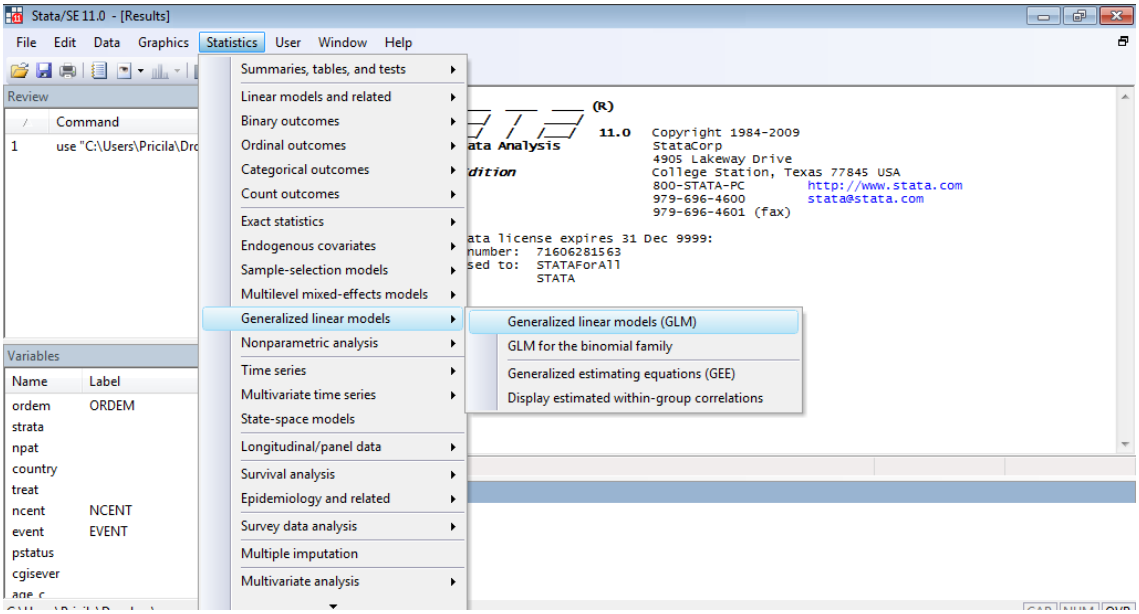

Figura 12: Como fazer GLM

Na janela que abrir, na aba *Model* identificar o modelo gamma com link identidade, e nos locais indicados o custo sendo estimado e a variável explicativa do modelo.

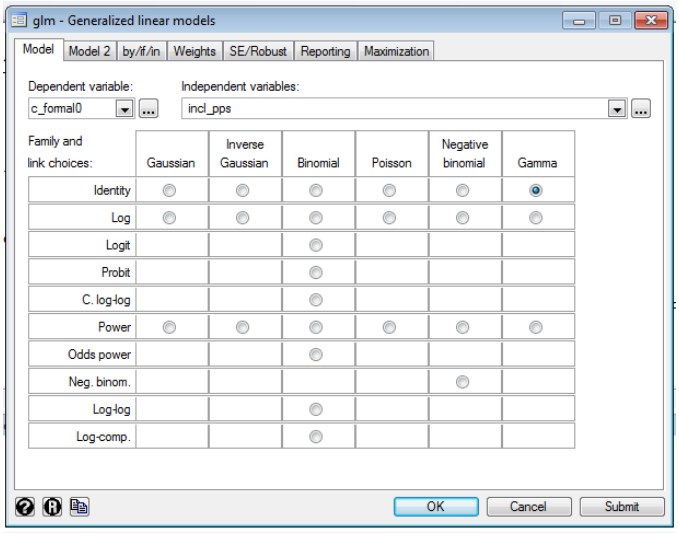

Figura 13: Selecionando modelo e variáveis do GLM

Na aba *weights* selecionar a opção ‘samplig weights’ e a variável de ponderação referente ao tempo analisado (no caso, pesot1). Em seguida, ‘OK’.

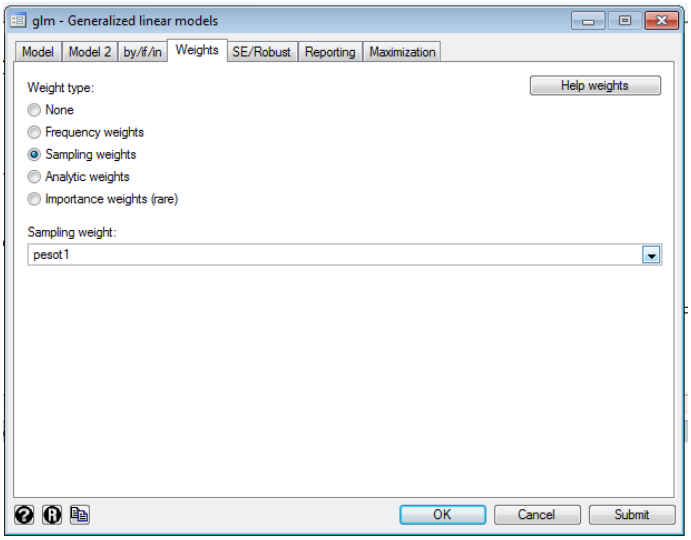

Figura 24: Acrescentando a variável 'peso' ao modelo GLM

O modelo resultante foi:

```
. glm c_formal0 incl_pps [pweight = pesot1], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood =  -3919.323
Iteration 1:   log pseudolikelihood = -3740.1432
Iteration 2:   log pseudolikelihood = -3734.7366
Iteration 3:   log pseudolikelihood = -3734.2993
Iteration 4:   log pseudolikelihood = -3734.2989
Iteration 5:   log pseudolikelihood = -3734.2989

Generalized linear models                         No. of obs      =       325
Optimization     : ML                             Residual df     =       323
                                                  Scale parameter =  1.785011
Deviance         =  485.2403458                   (1/df) Deviance =  1.502292
Pearson          =  576.5586356                   (1/df) Pearson  =  1.785011

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  22.99261
Log pseudolikelihood = -3734.298889               BIC             = -1382.935

------------------------------------------------------------------------------
             |               Robust
  c_formal0  |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    incl_pps |     90.5188   12.22084     7.41   0.000     66.5664    114.4712
       _cons |   -1241.789   822.2324    -1.51   0.131    -2853.335    369.7567
------------------------------------------------------------------------------

.
```

Figura 15: Modelo GLM para custo T1 com covariável gravidade

Na parte em destaque da Figura 15 podemos perceber o coeficiente da reta estimado pelo modelo (coluna Coef.) e o nível de significância associado à variável (coluna P>|z|).

Os modelos para a estimação do custo no tempo T1 com as demais covariáveis estão apresentadas a seguir.

Tratamento (covariável *treat*):

glm c_formal0 treat [pweight = pesot1], family(gamma) link(identity)

```
. glm c_formal0 treat [pweight = pesot1], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -3888.4848
Iteration 1:   log pseudolikelihood = -3661.9485
Iteration 2:   log pseudolikelihood = -3660.0051
Iteration 3:   log pseudolikelihood = -3659.8351
Iteration 4:   log pseudolikelihood = -3659.8345
Iteration 5:   log pseudolikelihood = -3659.8345

Generalized linear models                         No. of obs      =       325
Optimization     : ML                             Residual df     =       323
                                                  Scale parameter =  2.458436
Deviance         =  547.1468624                   (1/df) Deviance =  1.693953
Pearson          =  794.0747877                   (1/df) Pearson  =  2.458436

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  22.53437
Log pseudolikelihood = -3659.834507               BIC             = -1321.029

------------------------------------------------------------------------------
             |               Robust
  c_formal0  |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   246.2036   1087.538     0.23   0.821    -1885.331    2377.738
       _cons |   6748.091   690.2043     9.78   0.000     5395.315    8100.866
------------------------------------------------------------------------------
```

Figura 16: Modelo GLM para custo T1 com covariável tratamento

Idade (covariável *age_c*):

glm c_formal0 age_c [pweight = pesot1], family(gamma) link(identity)

```
. glm c_formal0 age_c [pweight = pesot1], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -3873.2604
Iteration 1:   log pseudolikelihood = -3656.5726
Iteration 2:   log pseudolikelihood = -3654.9268
Iteration 3:   log pseudolikelihood = -3653.4736
Iteration 4:   log pseudolikelihood = -3653.4441
Iteration 5:   log pseudolikelihood = -3653.4441

Generalized linear models                         No. of obs      =       325
Optimization     : ML                             Residual df     =       323
                                                  Scale parameter =  2.483291
Deviance         =  534.3660327                   (1/df) Deviance =  1.654384
Pearson          =  802.1028989                   (1/df) Pearson  =  2.483291

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  22.49504
Log pseudolikelihood = -3653.444092               BIC             =  -1333.81

------------------------------------------------------------------------------
             |               Robust
  c_formal0  |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age_c |   168.3496   53.00076     3.18   0.001     64.47005    272.2292
       _cons |  -4571.923   3465.292    -1.32   0.187    -11363.77    2219.925
------------------------------------------------------------------------------
```

Figura 37: Modelo GLM para custo T1 com covariável idade

Sexo (covariável *sex*):

glm c_formal0 sex [pweight = pesot1], family(gamma) link(identity)

```
. glm c_formal0 sex [pweight = pesot1], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -3876.3667
Iteration 1:   log pseudolikelihood = -3658.4651
Iteration 2:   log pseudolikelihood = -3656.1891
Iteration 3:   log pseudolikelihood =  -3655.927
Iteration 4:   log pseudolikelihood = -3655.9265
Iteration 5:   log pseudolikelihood = -3655.9265

Generalized linear models                         No. of obs      =       325
Optimization     : ML                             Residual df     =       323
                                                  Scale parameter =  2.464935
Deviance         =  539.3308691                   (1/df) Deviance =  1.669755
Pearson          =  796.1740282                   (1/df) Pearson  =  2.464935

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  22.51032
Log pseudolikelihood =  -3655.92651               BIC             = -1328.845

------------------------------------------------------------------------------
             |               Robust
  c_formal0  |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -2062.913   1132.719    -1.82   0.069    -4283.001    157.1743
       _cons |   8082.548   932.9878     8.66   0.000     6253.925     9911.17
------------------------------------------------------------------------------
```

Figura 48: Modelo GLM para custo T1 com covariável sexo

Tempo doente (covariável *disdur_c*):

glm c_formal0 disdur_c [pweight = pesot1], family(gamma) link(identity)

```
. glm c_formal0 disdur_c [pweight = pesot1], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -3887.3305
Iteration 1:   log pseudolikelihood = -3661.3928
Iteration 2:   log pseudolikelihood = -3659.3926
Iteration 3:   log pseudolikelihood = -3658.9585
Iteration 4:   log pseudolikelihood = -3658.9582

Generalized linear models                         No. of obs      =       325
Optimization     : ML                             Residual df     =       323
                                                  Scale parameter =  2.493672
Deviance         =  545.3942063                   (1/df) Deviance =  1.688527
Pearson          =  805.4561802                   (1/df) Pearson  =  2.493672

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  22.52897
Log pseudolikelihood = -3658.958179               BIC             = -1322.781

------------------------------------------------------------------------------
             |               Robust
  c_formal0  |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     disdur_c |   294.675   321.3706     0.92   0.359    -335.1997    924.5498
       _cons |   5687.287   1363.537     4.17   0.000     3014.804    8359.769
------------------------------------------------------------------------------
```

Figura 19: Modelo GLM para custo T1 com covariável tempo doente

Cognição (covariável *mmse*)

glm c_formal0 mmse [pweight = pesot1], family(gamma) link(identity)

```
. glm c_formal0 mmse [pweight = pesot1], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -3759.5831
Iteration 1:   log pseudolikelihood = -3558.1105
Iteration 2:   log pseudolikelihood = -3555.9431
Iteration 3:   log pseudolikelihood = -3554.5036
Iteration 4:   log pseudolikelihood = -3554.4943
Iteration 5:   log pseudolikelihood = -3554.4943

Generalized linear models                         No. of obs      =       317
Optimization     : ML                             Residual df     =       315
                                                  Scale parameter =   2.45759
Deviance         =  510.9360419                   (1/df) Deviance =  1.622019
Pearson          =  774.1408241                   (1/df) Pearson  =   2.45759

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  22.43845
Log pseudolikelihood =  -3554.49426               BIC             = -1303.118

------------------------------------------------------------------------------
             |               Robust
  c_formal0  |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mmse |  -1964.863   1008.665    -1.95   0.051    -3941.811     12.08455
       _cons |   7289.393   744.7266     9.79   0.000     5829.756      8749.03
------------------------------------------------------------------------------
```

Figura 20: Modelo GLM para custo T1 com covariável cognição

Para o modelo a ser utilizado na estimação dos custos dos pacientes censurados serão utilizadas apenas as covariáveis que nos modelos individuais tenham apresentado nível de significância menor que 10%. Dessa forma para o tempo T1 as covariáveis a serem incluídas no modelo são sexo (sex), cognição (mmse) idade (age_c) e gravidade da doença (incl_pps):

glm c_formal0 sex mmse age_c incl_pps [pweight = pesot1], family(gamma) link(identity)

```
. glm c_formal0  sex  mmse age_c incl_pps [pweight = pesot1], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -3683.1689
Iteration 1:   log pseudolikelihood = -3526.1637
Iteration 2:   log pseudolikelihood = -3522.1134
Iteration 3:   log pseudolikelihood = -3521.9464
Iteration 4:   log pseudolikelihood = -3521.9457
Iteration 5:   log pseudolikelihood = -3521.9457

Generalized linear models                         No. of obs      =       317
Optimization     : ML                             Residual df     =       312
                                                  Scale parameter =  1.838904
Deviance         =  445.8389954                   (1/df) Deviance =  1.428971
Pearson          =  573.7381357                   (1/df) Pearson  =  1.838904

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  22.25202
Log pseudolikelihood = -3521.945736               BIC             = -1350.938

------------------------------------------------------------------------------
             |               Robust
  c_formal0  |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -775.1375    874.754    -0.89   0.376    -2489.624     939.3488
        mmse |   677.9576   808.0589     0.84   0.401    -905.8087     2261.724
       age_c |    96.7397   59.92799     1.61   0.106    -20.71699     214.1964
    incl_pps |   79.63061   16.20498     4.91   0.000     47.86943     111.3918
       _cons |  -6958.229   4558.954    -1.53   0.127    -15893.61     1977.156
------------------------------------------------------------------------------
```

Figura 21: Modelo GLM para custo T1 com covariáveis sexo, cognição, idade e gravidade da doença (incl_pps)

Quando incluídas simultaneamente no modelo as variáveis idade, sexo e gravidade deixaram de ser significante ($p > 0,05$), então serão ordenadamente retiradas do modelo até que apenas variáveis significantes permaneçam. Assim, retirando a variável de maior p-valor (mmse) o modelo fica:

glm c_formal0 sex age_c incl_pps [pweight = pesot1], family(gamma) link(identity)

```
. glm c_formal0  sex  age_c incl_pps [pweight = pesot1], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -3790.4402
Iteration 1:   log pseudolikelihood =   -3623.25
Iteration 2:   log pseudolikelihood = -3618.6039
Iteration 3:   log pseudolikelihood = -3618.3808
Iteration 4:   log pseudolikelihood = -3618.3789
Iteration 5:   log pseudolikelihood = -3618.3789

Generalized linear models                         No. of obs      =       325
Optimization     : ML                             Residual df     =       321
                                                  Scale parameter =  1.761136
Deviance         =  464.2356988                   (1/df) Deviance =  1.446217
Pearson          =  565.3245719                   (1/df) Pearson  =  1.761136

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  22.29156
Log pseudolikelihood = -3618.378925               BIC             = -1392.372

------------------------------------------------------------------------------
             |               Robust
   c_formal0 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -539.5662   830.2466    -0.65   0.516    -2166.82    1087.687
       age_c |   79.82332   62.85945     1.27   0.204    -43.37894   203.0256
    incl_pps |   80.27409   13.34259     6.02   0.000     54.1231    106.4251
       _cons |  -5555.344   4297.689    -1.29   0.196    -13978.66   2867.971
------------------------------------------------------------------------------
```

Figura 22: Modelo GLM para custo T1 com covariáveis sexo, idade e gravidade da doença (incl_pps)

Retirando a próxima variável de maior p-valor (sexo), o modelo fica

glm c_formal0 age_c incl_pps [pweight = pesot1], family(gamma) link(identity)

```
. glm c_formal0  age_c incl_pps [pweight = pesot1], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -3795.7073
Iteration 1:   log pseudolikelihood = -3624.1958
Iteration 2:   log pseudolikelihood = -3619.1797
Iteration 3:   log pseudolikelihood = -3618.7353
Iteration 4:   log pseudolikelihood = -3618.7345
Iteration 5:   log pseudolikelihood = -3618.7345

Generalized linear models                         No. of obs      =       325
Optimization     : ML                             Residual df     =       322
                                                  Scale parameter =  1.725091
Deviance         =  464.9467926                   (1/df) Deviance =  1.443934
Pearson          =  555.4793321                   (1/df) Pearson  =  1.725091

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   22.2876
Log pseudolikelihood = -3618.734472               BIC             = -1397.445

------------------------------------------------------------------------------
             |               Robust
   c_formal0 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age_c |   86.81657    57.2741     1.52   0.130    -25.4386    199.0717
    incl_pps |   83.54005   12.43677     6.72   0.000     59.16442   107.9157
       _cons |  -6596.987   3545.141    -1.86   0.063    -13545.33   351.3607
------------------------------------------------------------------------------
```

Figura 23: Modelo GLM para custo T1 com covariáveis idade e gravidade da doença (incl_pps)

Retirando a última covariável de p-valor não significante temos que o modelo final para estimação do custo no tempo T1 é:

glm c_formal0 incl_pps [pweight = pesot1], family(gamma) link(identity)

```
. glm c_formal0 incl_pps [pweight = pesot1], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -3800.8429
Iteration 1:   log pseudolikelihood = -3626.4706
Iteration 2:   log pseudolikelihood = -3621.0144
Iteration 3:   log pseudolikelihood = -3620.5294
Iteration 4:   log pseudolikelihood = -3620.5288
Iteration 5:   log pseudolikelihood = -3620.5288

Generalized linear models                         No. of obs      =        325
Optimization     : ML                             Residual df     =        323
                                                  Scale parameter =   1.725673
Deviance         =  468.5354402                   (1/df) Deviance =   1.450574
Pearson          =   557.392461                   (1/df) Pearson  =   1.725673

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   22.29248
Log pseudolikelihood = -3620.528796               BIC             =  -1399.64

------------------------------------------------------------------------------
             |               Robust
   c_formal0 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    incl_pps |    90.1682   11.96529     7.54   0.000     66.71665    113.6197
       _cons |  -1263.022   803.4652    -1.57   0.116    -2837.785    311.7405
------------------------------------------------------------------------------
```

Figura 24: Modelo GLM para custo T1 com covariável gravidade da doença (incl_pps)

$$custo\ T1 = 90{,}16 * gravidade - 1263{,}02$$

Tempo 2 (T2)

Inicialmente, modelos com as covariáveis individualmente.

Tratamento (covariável *treat*):

glm c_formal6 treat [pweight = pesot2], family(gamma) link(identity)

```
. glm c_formal6 treat [pweight = pesot2], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -3366.3595
Iteration 1:   log pseudolikelihood = -3129.3483
Iteration 2:   log pseudolikelihood = -3128.5338
Iteration 3:   log pseudolikelihood = -3128.4714
Iteration 4:   log pseudolikelihood = -3128.4712

Generalized linear models                         No. of obs      =        272
Optimization     : ML                             Residual df     =        270
                                                  Scale parameter =   2.052035
Deviance         =  518.2602427                   (1/df) Deviance =   1.919482
Pearson          =  554.0493602                   (1/df) Pearson  =   2.052035

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   23.01817
Log pseudolikelihood =  -3128.47123               BIC             =  -995.3063

------------------------------------------------------------------------------
             |               Robust
   c_formal6 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   1543.009   1088.536     1.42   0.156    -590.4817    3676.499
       _cons |    5965.83   649.6291     9.18   0.000     4692.58     7239.079
------------------------------------------------------------------------------
```

Figura 25: Modelo GLM para custo T2 com covariável tratamento

Idade (covariável *age_c*):

glm c_formal6 age_c [pweight = pesot2], family(gamma) link(identity)

```
. glm c_formal6 age_c [pweight = pesot2], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood =  -3339.856
Iteration 1:   log pseudolikelihood = -3120.6878
Iteration 2:   log pseudolikelihood = -3120.1717
Iteration 3:   log pseudolikelihood = -3119.2053
Iteration 4:   log pseudolikelihood =  -3119.176
Iteration 5:   log pseudolikelihood = -3119.1759

Generalized linear models                         No. of obs      =       272
Optimization     : ML                             Residual df     =       270
                                                  Scale parameter =  1.970021
Deviance         =  499.8407461                   (1/df) Deviance =  1.851262
Pearson          =  531.9057219                   (1/df) Pearson  =  1.970021

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  22.94982
Log pseudolikelihood = -3119.175902               BIC             = -1013.726

------------------------------------------------------------------------------
             |               Robust
   c_formal6 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age_c |   216.6478   48.55978     4.46   0.000     121.4724    311.8233
       _cons |  -8054.767   3107.887    -2.59   0.010    -14146.11   -1963.421
------------------------------------------------------------------------------
```

Figura 26: Modelo GLM para custo T2 com covariável idade

Sexo (covariável *sex*):

glm c_formal6 sex [pweight = pesot2], family(gamma) link(identity)

```
. glm c_formal6 sex [pweight = pesot2], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -3340.7254
Iteration 1:   log pseudolikelihood = -3121.2302
Iteration 2:   log pseudolikelihood = -3119.6539
Iteration 3:   log pseudolikelihood = -3119.3318
Iteration 4:   log pseudolikelihood = -3119.3318

Generalized linear models                         No. of obs      =       272
Optimization     : ML                             Residual df     =       270
                                                  Scale parameter =  2.062415
Deviance         =  500.3393813                   (1/df) Deviance =  1.853109
Pearson          =  556.8521781                   (1/df) Pearson  =  2.062415

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  22.95097
Log pseudolikelihood = -3119.331814               BIC             = -1013.227

------------------------------------------------------------------------------
             |               Robust
   c_formal6 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -3691.465    1148.69    -3.21   0.001    -5942.856   -1440.075
       _cons |    8906.12   994.0693     8.96   0.000     6957.78    10854.46
------------------------------------------------------------------------------
```

Figura 27: Modelo GLM para custo T2 com covariável sexo

Tempo doente (covariável *disdur_c*):

glm c_formal6 disdur_c [pweight = pesot2], family(gamma) link(identity)

```
. glm c_formal6 disdur_c [pweight = pesot2], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -3373.8805
Iteration 1:   log pseudolikelihood =  -3131.789
Iteration 2:   log pseudolikelihood = -3130.6045
Iteration 3:   log pseudolikelihood = -3130.2238
Iteration 4:   log pseudolikelihood = -3130.2228
Iteration 5:   log pseudolikelihood = -3130.2228

Generalized linear models                         No. of obs      =       272
Optimization     : ML                             Residual df     =       270
                                                  Scale parameter =  2.127527
Deviance         =  521.5530225                   (1/df) Deviance =  1.931678
Pearson          =   574.432369                   (1/df) Pearson  =  2.127527

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  23.03105
Log pseudolikelihood = -3130.222811               BIC             = -992.0135

------------------------------------------------------------------------------
             |               Robust
   c_formal6 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     disdur_c |   202.1144   346.0346     0.58   0.559    -476.101    880.3298
       _cons |   5890.918   1520.613     3.87   0.000     2910.57    8871.265
------------------------------------------------------------------------------
```

Figura 28: Modelo GLM para custo T2 com covariável tempo doente

Cognição (covariável *mmse*):

glm c_formal6 mmse [pweight = pesot2], family(gamma) link(identity)

```
. glm c_formal6 mmse [pweight = pesot2], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -3282.8205
Iteration 1:   log pseudolikelihood = -3056.5377
Iteration 2:   log pseudolikelihood = -3055.9235
Iteration 3:   log pseudolikelihood = -3055.8932
Iteration 4:   log pseudolikelihood = -3055.8931

Generalized linear models                         No. of obs      =        266
Optimization     : ML                             Residual df     =        264
                                                  Scale parameter =   2.133779
Deviance         =  504.5884338                   (1/df) Deviance =    1.91132
Pearson          =  563.3177612                   (1/df) Pearson  =   2.133779

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   22.99168
Log pseudolikelihood = -3055.893137               BIC             =  -969.4546

------------------------------------------------------------------------------
             |               Robust
   c_formal6 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mmse |  -2430.931   1043.8     -2.33   0.020    -4476.741   -385.1209
       _cons |    7560.95   763.8927    9.90   0.000     6063.748    9058.153
------------------------------------------------------------------------------
```

Figura 29: Modelo GLM para custo T2 com covariável cognição

Gravidade (covariável *m6_pps*):

glm c_formal6 m6_pps [pweight = pesot2], family(gamma) link(identity)

```
. glm c_formal6  m6_pps [pweight = pesot2], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -3313.5647
Iteration 1:   log pseudolikelihood = -3098.4463
Iteration 2:   log pseudolikelihood = -3097.2172
Iteration 3:   log pseudolikelihood = -3097.1442
Iteration 4:   log pseudolikelihood = -3097.1441

Generalized linear models                         No. of obs      =        270
Optimization     : ML                             Residual df     =        268
                                                  Scale parameter =   2.129945
Deviance         =  497.1836567                   (1/df) Deviance =   1.855163
Pearson          =  570.8252978                   (1/df) Pearson  =   2.129945

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   22.95662
Log pseudolikelihood = -3097.144069               BIC             =  -1003.193

------------------------------------------------------------------------------
             |               Robust
   c_formal6 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      m6_pps |   54.40584   12.52124     4.35   0.000     29.86466    78.94701
       _cons |   1413.302    1135.61     1.24   0.213    -812.4533    3639.057
------------------------------------------------------------------------------
```

Figura 30: Modelo GLM para custo T2 com covariável gravidade

Para o modelo a ser utilizado na estimação dos custos dos pacientes censurados serão utilizadas apenas as covariáveis que nos modelos individuais tenham apresentado nível de significância menor que 10%. Dessa forma para o tempo T2 as covariáveis a serem incluídas no modelo são idade (age_c), sexo (sex), cognição (mmse) e gravidade da doença (m6_pps).

glm c_formal6 sex age_c mmse m6_pps [pweight = pesot2], family(gamma) link(identity)

```
. glm c_formal6  sex age_c mmse m6_pps [pweight = pesot2], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -3207.0564
Iteration 1:   log pseudolikelihood = -3015.1131
Iteration 2:   log pseudolikelihood =  -3013.117
Iteration 3:   log pseudolikelihood = -3010.7575
Iteration 4:   log pseudolikelihood =  -3010.706
Iteration 5:   log pseudolikelihood = -3010.7058

Generalized linear models                         No. of obs      =        264
Optimization     : ML                             Residual df     =        259
                                                  Scale parameter =   1.945103
Deviance         =  456.6201411                   (1/df) Deviance =   1.763012
Pearson          =  503.7816135                   (1/df) Pearson  =   1.945103

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   22.84626
Log pseudolikelihood =  -3010.70576               BIC             =  -987.5507

------------------------------------------------------------------------------
             |               Robust
   c_formal6 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -3053.592   1229.816    -2.48   0.013    -5463.986   -643.1969
       age_c |   104.1215   45.36479     2.30   0.022     15.20816    193.0349
        mmse |  -881.7136   1471.543    -0.60   0.549    -3765.885    2002.458
      m6_pps |   27.59674   16.01381     1.72   0.085    -3.789761    58.98323
       _cons |  -1174.652   4627.261    -0.25   0.800    -10243.92    7894.613
------------------------------------------------------------------------------
```

Figura 31: Modelo GLM para custo T2 com covariáveis sexo, idade, cognição e gravidade

Quando incluídas as variáveis simultaneamente no modelo as variáveis gravidade e cognição deixaram de ser significante (p > 0,05), então serão ordenadamente retiradas do modelo até que apenas variáveis significantes permaneçam. Assim, retirando a variável de maior p-valor (mmse) o modelo fica:

```
. glm c_formal6  sex age_c m6_pps [pweight = pesot2], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -3277.4687
Iteration 1:   log pseudolikelihood =  -3084.639
Iteration 2:   log pseudolikelihood = -3082.8569
Iteration 3:   log pseudolikelihood = -3080.8357
Iteration 4:   log pseudolikelihood = -3080.7861
Iteration 5:   log pseudolikelihood = -3080.7859

Generalized linear models                         No. of obs      =       270
Optimization     : ML                             Residual df     =       266
                                                  Scale parameter =  1.920032
Deviance         =  465.3275138                   (1/df) Deviance =  1.749352
Pearson          =  510.7284353                   (1/df) Pearson  =  1.920032

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  22.85027
Log pseudolikelihood = -3080.785875               BIC             = -1023.853

------------------------------------------------------------------------------
             |               Robust
    c_formal6 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -2796.689   1208.953    -2.31   0.021    -5166.193   -427.1857
       age_c |   119.9017   40.64314     2.95   0.003     40.24264    199.5608
      m6_pps |   37.77323   10.47063     3.61   0.000     17.25118    58.29528
       _cons |  -3599.985   3123.366    -1.15   0.249    -9721.669      2521.7
------------------------------------------------------------------------------
```

Figura 32: Modelo GLM para custo T2 com covariáveis sexo, idade e gravidade

Assim, o modelo final para estimação do custo no tempo T2 é:

$$custo\ T2 = 119{,}90 * idade - 2796{,}68 * sexo + 37{,}77 * grav. - 3599{,}98$$

Tempo 3 (T3)

Inicialmente, modelos com as covariáveis individualmente.

Tratamento (covariável *treat*):

glm c_formal12 treat [pweight = pesot3], family(gamma) link(identity)

```
. glm c_formal12 treat [pweight = pesot3], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -2954.5061
Iteration 1:   log pseudolikelihood = -2772.8941
Iteration 2:   log pseudolikelihood = -2771.5297
Iteration 3:   log pseudolikelihood = -2771.1346
Iteration 4:   log pseudolikelihood = -2771.1344

Generalized linear models                         No. of obs      =       235
Optimization     : ML                             Residual df     =       233
                                                  Scale parameter =  2.247836
Deviance         =  426.7564549                   (1/df) Deviance =  1.831573
Pearson          =  523.7457687                   (1/df) Pearson  =  2.247836

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  23.60114
Log pseudolikelihood = -2771.134439               BIC             =  -845.327

------------------------------------------------------------------------------
             |               Robust
   c_formal12 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        treat |   2480.835   1265.668     1.96   0.050      .170368    4961.499
        _cons |   5868.404   717.1538     8.18   0.000     4462.809      7274
------------------------------------------------------------------------------
```

Figura 33: Modelo GLM para custo T3 com covariável tratamento

Idade (covariável *age_c*):

glm c_formal12 age_c [pweight = pesot3], family(gamma) link(identity)

```
. glm c_formal12  age_c [pweight = pesot3], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -2923.6765
Iteration 1:   log pseudolikelihood = -2762.2832
Iteration 2:   log pseudolikelihood = -2760.0815
Iteration 3:   log pseudolikelihood = -2759.4632
Iteration 4:   log pseudolikelihood = -2759.3647
Iteration 5:   log pseudolikelihood = -2759.3628
Iteration 6:   log pseudolikelihood = -2759.3628

Generalized linear models                         No. of obs      =        235
Optimization     : ML                             Residual df     =        233
                                                  Scale parameter =   2.014356
Deviance         =  403.2132347                   (1/df) Deviance =   1.730529
Pearson          =  469.3449026                   (1/df) Pearson  =   2.014356

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  23.50096
Log pseudolikelihood = -2759.362829               BIC             = -868.8702

------------------------------------------------------------------------------
             |               Robust
  c_formal12 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age_c |    287.3664   39.19191     7.33   0.000     210.5516    364.1811
       _cons |   -12428.76   2234.488    -5.56   0.000    -16808.27   -8049.239
------------------------------------------------------------------------------
```

Figura 34: Modelo GLM para custo T3 com covariável idade

Sexo (covariável *sex*):

glm c_formal12 sex [pweight = pesot3], family(gamma) link(identity)

```
. glm c_formal12   sex [pweight = pesot3], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -2944.4599
Iteration 1:   log pseudolikelihood = -2769.3356
Iteration 2:   log pseudolikelihood = -2767.6981
Iteration 3:   log pseudolikelihood =  -2766.813
Iteration 4:   log pseudolikelihood = -2766.8088
Iteration 5:   log pseudolikelihood = -2766.8088

Generalized linear models                         No. of obs      =        235
Optimization     : ML                             Residual df     =        233
                                                  Scale parameter =    2.18829
Deviance         =    418.10508                   (1/df) Deviance =   1.794442
Pearson          =   509.871675                   (1/df) Pearson  =    2.18829

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  23.56433
Log pseudolikelihood = -2766.808752               BIC             = -853.9783

------------------------------------------------------------------------------
             |               Robust
  c_formal12 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   -3618.15   1355.973    -2.67   0.008    -6275.808   -960.4921
       _cons |   9194.561   1193.063     7.71   0.000       6856.2    11532.92
------------------------------------------------------------------------------
```

Figura 35: Modelo GLM para custo T3 com covariável sexo

Tempo doente (covariável *disdur_c*):

glm c_formal12 disdur_c [pweight = pesot3], family(gamma) link(identity)

```
. glm c_formal12   disdur_c [pweight = pesot3], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -2964.5611
Iteration 1:   log pseudolikelihood = -2776.5502
Iteration 2:   log pseudolikelihood = -2775.5775
Iteration 3:   log pseudolikelihood = -2775.4468
Iteration 4:   log pseudolikelihood = -2775.4467

Generalized linear models                         No. of obs      =        235
Optimization     : ML                             Residual df     =        233
                                                  Scale parameter =   2.348869
Deviance         =  435.3809081                   (1/df) Deviance =   1.868588
Pearson          =  547.2865449                   (1/df) Pearson  =   2.348869

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  23.63784
Log pseudolikelihood = -2775.446666               BIC             = -836.7025

------------------------------------------------------------------------------
             |               Robust
  c_formal12 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     disdur_c |   77.91489   381.2881     0.20   0.838    -669.3961    825.2259
       _cons |   6758.249   1726.154     3.92   0.000     3375.049    10141.45
------------------------------------------------------------------------------
```

Figura 36: Modelo GLM para custo T3 com covariável tempo doente

Cognição (covariável *mmse*):

glm c_formal12 mmse [pweight = pesot3], family(gamma) link(identity)

```
. glm c_formal12  mmse [pweight = pesot3], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -2895.8368
Iteration 1:   log pseudolikelihood = -2723.9852
Iteration 2:   log pseudolikelihood = -2722.6134
Iteration 3:   log pseudolikelihood = -2721.9009
Iteration 4:   log pseudolikelihood =  -2721.897
Iteration 5:   log pseudolikelihood =  -2721.897

Generalized linear models                         No. of obs      =        231
Optimization     : ML                             Residual df     =        229
                                                  Scale parameter =   2.297048
Deviance         =  413.6496881                   (1/df) Deviance =   1.806331
Pearson          =  526.0239497                   (1/df) Pearson  =   2.297048

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   23.58352
Log pseudolikelihood = -2721.896989               BIC             =   -832.664

------------------------------------------------------------------------------
             |               Robust
  c_formal12 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mmse |  -3372.827   1214.815    -2.78   0.005    -5753.821    -991.833
       _cons |   8528.725   963.3465     8.85   0.000       6640.6    10416.85
------------------------------------------------------------------------------
```

Figura 37: Modelo GLM para custo T3 com covariável cognição

Gravidade (covariável *m12_pps*):

glm c_formal12 m12_pps [pweight = pesot3], family(gamma) link(identity)

```
. glm c_formal12   m12_pps [pweight = pesot3], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -2850.5096
Iteration 1:   log pseudolikelihood = -2708.1115
Iteration 2:   log pseudolikelihood = -2705.0503
Iteration 3:   log pseudolikelihood = -2704.3398
Iteration 4:   log pseudolikelihood = -2704.2699
Iteration 5:   log pseudolikelihood = -2704.2696
Iteration 6:   log pseudolikelihood = -2704.2696

Generalized linear models                         No. of obs      =        231
Optimization     : ML                             Residual df     =        229
                                                  Scale parameter =    2.18374
Deviance         =  379.9694934                   (1/df) Deviance =   1.659255
Pearson          =  500.0765451                   (1/df) Pearson  =    2.18374

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   23.43091
Log pseudolikelihood = -2704.269621               BIC             =  -866.3442

------------------------------------------------------------------------------
             |               Robust
  c_formal12 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      m12_pps |   83.48771   10.32041     8.09   0.000     63.26008    103.7153
       _cons |  -1796.836   681.7161    -2.64   0.008    -3132.975   -460.6973
------------------------------------------------------------------------------
```

Figura 38: Modelo GLM para custo T3 com covariável gravidade

Para o modelo a ser utilizado na estimação dos custos dos pacientes censurados serão utilizadas apenas as covariáveis que nos modelos individuais tenham apresentado nível de significância menor que 10%. Dessa forma para o tempo T3 as covariáveis a serem incluídas no modelo são tratamento (treat), idade (age_c), sexo (sex), cognição (mmse) e gravidade da doença (m12_pps).

glm c_formal12 treat sex mmse age_c m12_pps [pweight = pesot3], family(gamma) link(identity)

```
. glm c_formal12  treat sex mmse age_c m12_pps [pweight = pesot3], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -2767.7552
Iteration 1:   log pseudolikelihood = -2644.8012
Iteration 2:   log pseudolikelihood = -2640.9462
Iteration 3:   log pseudolikelihood = -2640.1979
Iteration 4:   log pseudolikelihood = -2640.1874
Iteration 5:   log pseudolikelihood = -2640.1873

Generalized linear models                         No. of obs      =       227
Optimization     : ML                             Residual df     =       221
                                                  Scale parameter =  1.669375
Deviance         =  337.1730597                   (1/df) Deviance =   1.52567
Pearson          =  368.9319485                   (1/df) Pearson  =  1.669375

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  23.31443
Log pseudolikelihood = -2640.187337               BIC             = -861.7409

------------------------------------------------------------------------------
             |               Robust
   c_formal12 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
        treat |   1958.278   742.8186     2.64   0.008      502.38    3414.175
          sex |  -2773.834   1166.307    -2.38   0.017    -5059.753   -487.9144
         mmse |   -853.753   415.5546    -2.05   0.040    -1668.225   -39.28098
        age_c |   95.66062   18.52436     5.16   0.000     59.35353    131.9677
      m12_pps |   42.3494    6.706179     6.31   0.000     29.20553    55.49327
        _cons |  -3260.183   2029.109    -1.61   0.108    -7237.163    716.7973
------------------------------------------------------------------------------
```

Figura 39: Modelo GLM para custo T3 com covariáveis tratamento, sexo, cognição, idade e gravidade

Assim, o modelo final para estimação do custo no tempo T3 é:

$custo\ T3 = 1958{,}28 * trat - 2773{,}83 * sexo - 853.75 * cogn + 95{,}66 * idade$

$+42{,}35 * grav. - 3260{,}18$

Tempo 4 (T4)

Inicialmente, modelos com as covariáveis individualmente.

Tratamento (covariável *treat*):

glm c_formal24 treat [pweight = pesot4], family(gamma) link(identity)

```
. glm c_formal24 treat [pweight = pesot4], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -2187.9214
Iteration 1:   log pseudolikelihood =  -2037.459
Iteration 2:   log pseudolikelihood = -2037.1087
Iteration 3:   log pseudolikelihood = -2037.0928
Iteration 4:   log pseudolikelihood = -2037.0928

Generalized linear models                         No. of obs      =       159
Optimization     : ML                             Residual df     =       157
                                                  Scale parameter =  3.361551
Deviance         =  329.7721802                   (1/df) Deviance =   2.10046
Pearson          =  527.7634805                   (1/df) Pearson  =  3.361551

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  25.64897
Log pseudolikelihood = -2037.092825               BIC             = -466.0458

------------------------------------------------------------------------------
             |               Robust
   c_formal24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
        treat |    3575.77   2488.768     1.44   0.151    -1302.125    8453.665
        _cons |   7394.537   1222.234     6.05   0.000     4999.003    9790.071
------------------------------------------------------------------------------
```

Figura 40: Modelo GLM para custo T4 com covariável tratamento

Idade (covariável *age_c*):

glm c_formal24 age_c [pweight = pesot4], family(gamma) link(identity)

```
. glm c_formal24  age_c [pweight = pesot4], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood =  -2199.097
Iteration 1:   log pseudolikelihood = -2040.6535
Iteration 2:   log pseudolikelihood = -2039.6585
Iteration 3:   log pseudolikelihood = -2039.5999
Iteration 4:   log pseudolikelihood = -2039.5998

Generalized linear models                         No. of obs      =       159
Optimization     : ML                             Residual df     =       157
                                                  Scale parameter =   3.96247
Deviance         =  334.7861227                   (1/df) Deviance =  2.132396
Pearson          =  622.1078048                   (1/df) Pearson  =   3.96247

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   25.6805
Log pseudolikelihood = -2039.599796               BIC             = -461.0318

------------------------------------------------------------------------------
             |               Robust
  c_formal24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age_c |   124.5211   127.7947     0.97   0.330    -125.9519     374.994
       _cons |    793.148     8884.5     0.09   0.929    -16620.15    18206.45
------------------------------------------------------------------------------
```

Figura 41: Modelo GLM para custo T4 com covariável idade

Sexo (covariável *sex*):

glm c_formal24 sex [pweight = pesot4], family(gamma) link(identity)

```
. glm c_formal24   sex [pweight = pesot4], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -2198.5311
Iteration 1:   log pseudolikelihood = -2040.9556
Iteration 2:   log pseudolikelihood = -2040.5833
Iteration 3:   log pseudolikelihood = -2040.5568
Iteration 4:   log pseudolikelihood = -2040.5568

Generalized linear models                         No. of obs      =       159
Optimization     : ML                             Residual df     =       157
                                                  Scale parameter =  3.876481
Deviance         =  336.7000436                   (1/df) Deviance =  2.144586
Pearson          =  608.6075896                   (1/df) Pearson  =  3.876481

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  25.69254
Log pseudolikelihood = -2040.556757               BIC             = -459.1179

------------------------------------------------------------------------------
             |               Robust
  c_formal24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   -1243.67   2437.694    -0.51   0.610    -6021.462     3534.122
       _cons |    9927.75   1719.182     5.77   0.000     6558.215     13297.29
------------------------------------------------------------------------------
```

Figura 42: Modelo GLM para custo T4 com covariável sexo

Tempo doente (covariável *disdur_c*):

glm c_formal24 disdur_c [pweight = pesot4], family(gamma) link(identity)

```
. glm c_formal24   disdur_c [pweight = pesot4], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -2198.0183
Iteration 1:   log pseudolikelihood = -2040.1689
Iteration 2:   log pseudolikelihood = -2039.4349
Iteration 3:   log pseudolikelihood = -2039.3039
Iteration 4:   log pseudolikelihood = -2039.3038

Generalized linear models                         No. of obs      =       159
Optimization     : ML                             Residual df     =       157
                                                  Scale parameter =  3.273394
Deviance         =  334.1942079                   (1/df) Deviance =  2.128626
Pearson          =  513.9227807                   (1/df) Pearson  =  3.273394

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  25.67678
Log pseudolikelihood = -2039.303839               BIC             = -461.6238

------------------------------------------------------------------------------
             |               Robust
  c_formal24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    disdur_c |   -581.2782   600.7845    -0.97   0.333    -1758.794     596.2377
       _cons |    11482.02   3302.616     3.48   0.001     5009.015     17955.03
------------------------------------------------------------------------------
```

Figura 43: Modelo GLM para custo T4 com covariável tempo doente

Cognição (covariável *mmse*):

glm c_formal24 mmse [pweight = pesot4], family(gamma) link(identity)

```
. glm c_formal24   mmse [pweight = pesot4], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -2161.1995
Iteration 1:   log pseudolikelihood = -2012.3806
Iteration 2:   log pseudolikelihood = -2012.1598
Iteration 3:   log pseudolikelihood = -2012.1572
Iteration 4:   log pseudolikelihood = -2012.1572

Generalized linear models                         No. of obs      =       157
Optimization     : ML                             Residual df     =       155
                                                  Scale parameter =  4.123233
Deviance         =  326.9952866                   (1/df) Deviance =  2.109647
Pearson          =   639.101048                   (1/df) Pearson  =  4.123233

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  25.65805
Log pseudolikelihood = -2012.157212               BIC             = -456.7228

------------------------------------------------------------------------------
             |               Robust
  c_formal24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mmse |  -1946.248   2496.797    -0.78   0.436    -6839.881    2947.385
       _cons |   9879.684   1466.127     6.74   0.000     7006.127    12753.24
------------------------------------------------------------------------------
```

Figura 44: Modelo GLM para custo T4 com covariável cognição

Gravidade (covariável *m12_pps*):

glm c_formal24 m24_pps [pweight = pesot4], family(gamma) link(identity)

```
. glm c_formal24  m24_pps [pweight = pesot4], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -2160.6072
Iteration 1:   log pseudolikelihood = -2018.3287
Iteration 2:   log pseudolikelihood = -2017.8011
Iteration 3:   log pseudolikelihood =   -2017.78
Iteration 4:   log pseudolikelihood = -2017.7799

Generalized linear models                         No. of obs      =       158
Optimization     : ML                             Residual df     =       156
                                                  Scale parameter =    3.6358
Deviance         =  318.9218942                   (1/df) Deviance =  2.044371
Pearson          =  567.1848554                   (1/df) Pearson  =    3.6358

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  25.56683
Log pseudolikelihood = -2017.779939               BIC             = -470.8429

------------------------------------------------------------------------------
             |               Robust
  c_formal24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      m24_pps |   42.49186    39.7168     1.07   0.285    -35.35163    120.3354
       _cons |   3853.454   4474.099     0.86   0.389     -4915.62    12622.53
------------------------------------------------------------------------------
```

Figura 45: Modelo GLM para custo T4 com covariável gravidade

Para o tempo T4 não foi possível ajustar um modelo pois nenhuma covariável foi significante na estimação do custo. Os pacientes censurados tiveram seus custos estimados para T4 através da média dos pacientes observados neste tempo.

Tempo 5 (T5)

Inicialmente, modelos com as covariáveis individualmente.

Tratamento (covariável *treat*):

glm c_formal36 treat [pweight = pesot5], family(gamma) link(identity)

```
. glm c_formal36 treat [pweight = pesot5], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -1597.8233
Iteration 1:   log pseudolikelihood =  -1521.522
Iteration 2:   log pseudolikelihood = -1520.3002
Iteration 3:   log pseudolikelihood = -1519.7946
Iteration 4:   log pseudolikelihood = -1519.7918
Iteration 5:   log pseudolikelihood = -1519.7918

Generalized linear models                         No. of obs      =       112
Optimization     : ML                             Residual df     =       110
                                                  Scale parameter =  1.605109
Deviance         =  199.8151419                   (1/df) Deviance =  1.816501
Pearson          =  176.5620198                   (1/df) Pearson  =  1.605109

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  27.17485
Log pseudolikelihood = -1519.791776               BIC             = -319.2197

------------------------------------------------------------------------------
             |               Robust
  c_formal36 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   1999.532   2051.824     0.97   0.330    -2021.969    6021.032
       _cons |   8974.191   1373.036     6.54   0.000     6283.089    11665.29
------------------------------------------------------------------------------
```

Figura 46: Modelo GLM para custo T5 com covariável tratamento

Idade (covariável *age_c*):

glm c_formal36 age_c [pweight = pesot5], family(gamma) link(identity)

```
. glm c_formal36  age_c [pweight = pesot5], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -1596.3383
Iteration 1:   log pseudolikelihood = -1521.1404
Iteration 2:   log pseudolikelihood = -1519.4105
Iteration 3:   log pseudolikelihood =  -1519.408
Iteration 4:   log pseudolikelihood =  -1519.408

Generalized linear models                         No. of obs      =       112
Optimization     : ML                             Residual df     =       110
                                                  Scale parameter =   1.65187
Deviance         =  199.0476602                   (1/df) Deviance =  1.809524
Pearson          =  181.7057197                   (1/df) Pearson  =   1.65187

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =    27.168
Log pseudolikelihood = -1519.408035               BIC             = -319.9872

------------------------------------------------------------------------------
             |               Robust
  c_formal36 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age_c |   192.2943   172.5116     1.11   0.265    -145.8222    530.4108
       _cons |  -2867.741   11482.45    -0.25   0.803    -25372.92    19637.44
------------------------------------------------------------------------------
```

Figura 47: Modelo GLM para custo T5 com covariável idade

Sexo (covariável *sex*):

glm c_formal36 sex [pweight = pesot5], family(gamma) link(identity)

```
. glm c_formal36  sex [pweight = pesot5], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -1597.4332
Iteration 1:   log pseudolikelihood = -1521.1982
Iteration 2:   log pseudolikelihood = -1520.0317
Iteration 3:   log pseudolikelihood =  -1519.396
Iteration 4:   log pseudolikelihood = -1519.3913
Iteration 5:   log pseudolikelihood = -1519.3913

Generalized linear models                         No. of obs      =       112
Optimization     : ML                             Residual df     =       110
                                                  Scale parameter =  1.577741
Deviance         =   199.014148                   (1/df) Deviance =   1.80922
Pearson          =  173.5515327                   (1/df) Pearson  =  1.577741

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   27.1677
Log pseudolikelihood = -1519.391279               BIC             = -320.0207

------------------------------------------------------------------------------
             |               Robust
  c_formal36 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -2625.18   2240.953    -1.17   0.241    -7017.367    1767.008
       _cons |   11638.13   1903.985     6.11   0.000     7906.391    15369.87
------------------------------------------------------------------------------
```

Figura 48: Modelo GLM para custo T5 com covariável sexo

Tempo doente (covariável *disdur_c*):

glm c_formal36 disdur_c [pweight = pesot5], family(gamma) link(identity)

```
. glm c_formal36   disdur_c [pweight = pesot5], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -1599.9468
Iteration 1:   log pseudolikelihood = -1522.1215
Iteration 2:   log pseudolikelihood = -1520.9186
Iteration 3:   log pseudolikelihood = -1520.5355
Iteration 4:   log pseudolikelihood = -1520.5335
Iteration 5:   log pseudolikelihood = -1520.5335

Generalized linear models                         No. of obs      =        112
Optimization     : ML                             Residual df     =        110
                                                  Scale parameter =    1.60245
Deviance         =  201.2985566                   (1/df) Deviance =   1.829987
Pearson          =  176.2694513                   (1/df) Pearson  =    1.60245

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =    27.1881
Log pseudolikelihood = -1520.533483               BIC             = -317.7363

------------------------------------------------------------------------------
             |               Robust
  c_formal36 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    disdur_c |   -73.04707   638.8101    -0.11   0.909    -1325.092    1178.998
       _cons |    10269.94   2843.776     3.61   0.000     4696.245    15843.64
------------------------------------------------------------------------------
```

Figura 49: Modelo GLM para custo T5 com covariável tempo doente

Cognição (covariável *mmse*):

glm c_formal36 mmse [pweight = pesot5], family(gamma) link(identity)

```
. glm c_formal36  mmse [pweight = pesot5], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -1593.9136
Iteration 1:   log pseudolikelihood = -1520.2821
Iteration 2:   log pseudolikelihood = -1518.2434
Iteration 3:   log pseudolikelihood =  -1518.228
Iteration 4:   log pseudolikelihood =  -1518.228

Generalized linear models                         No. of obs      =        112
Optimization     : ML                             Residual df     =        110
                                                  Scale parameter =   1.761992
Deviance         =   196.687504                   (1/df) Deviance =   1.788068
Pearson          =  193.8191532                   (1/df) Pearson  =   1.761992

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   27.14693
Log pseudolikelihood = -1518.227957               BIC             = -322.3474

------------------------------------------------------------------------------
             |               Robust
  c_formal36 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mmse |   -3483.639   2037.155    -1.71   0.087    -7476.389    509.1103
       _cons |    11684.67   1367.576     8.54   0.000      9004.27    14365.07
------------------------------------------------------------------------------
```

Figura 50: Modelo GLM para custo T5 com covariável cognição

Gravidade (covariável *m12_pps*):

glm c_formal36 m36_pps [pweight = pesot5], family(gamma) link(identity)

```
. glm c_formal36   m36_pps [pweight = pesot5], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -1573.2626
Iteration 1:   log pseudolikelihood = -1510.9983
Iteration 2:   log pseudolikelihood = -1509.8299
Iteration 3:   log pseudolikelihood = -1509.7241
Iteration 4:   log pseudolikelihood = -1509.7232
Iteration 5:   log pseudolikelihood = -1509.7232

Generalized linear models                         No. of obs      =        112
Optimization     : ML                             Residual df     =        110
                                                  Scale parameter =   1.693333
Deviance         =  179.6779676                   (1/df) Deviance =   1.633436
Pearson          =  186.2666674                   (1/df) Pearson  =   1.693333

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   26.99506
Log pseudolikelihood = -1509.723188               BIC             = -339.3569

------------------------------------------------------------------------------
             |               Robust
  c_formal36 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      m36_pps |   108.3374   16.44164     6.59   0.000     76.11233    140.5624
       _cons |  -3107.158   1430.119    -2.17   0.030     -5910.14   -304.1765
------------------------------------------------------------------------------
```

Figura 51: Modelo GLM para custo T5 com covariável gravidade

Para o modelo a ser utilizado na estimação dos custos dos pacientes censurados serão utilizadas apenas as covariáveis que nos modelos individuais tenham apresentado nível de significância menor que 10%. Dessa forma para o tempo T5 as covariáveis a serem incluídas no modelo são gravidade (m36_pps) e cognição (mmse).

```
. glm c_formal36 mmse  m36_pps [pweight = pesot5], family(gamma) link(identity)

Iteration 0:   log pseudolikelihood = -1563.1378
Iteration 1:   log pseudolikelihood =   -1504.81
Iteration 2:   log pseudolikelihood =    -1504.1
Iteration 3:   log pseudolikelihood = -1504.0947
Iteration 4:   log pseudolikelihood = -1504.0947

Generalized linear models                         No. of obs      =       112
Optimization     : ML                             Residual df     =       109
                                                  Scale parameter =  1.602038
Deviance         =  168.4209667                   (1/df) Deviance =  1.545146
Pearson          =  174.6221422                   (1/df) Pearson  =  1.602038

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  26.91241
Log pseudolikelihood = -1504.094688               BIC             = -345.8954
```

c_formal36	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
mmse	-4014.679	1666.822	-2.41	0.016	-7281.589	-747.7684
m36_pps	88.66879	14.93942	5.94	0.000	59.38807	117.9495
_cons	1093.313	2069.897	0.53	0.597	-2963.609	5150.236

Figura 52: Modelo GLM para custo T5 com covariáveis gravidade e cognição

Assim, o modelo final para estimação do custo no tempo T5 é:

$$custo\ T5 = -4014{,}68 * cogn + 88{,}67 * grav. + 1093.31$$