Fábio Vieira

Análise de Sobrevivência - Trabalho Final
14 de outubro de 2017

O objetivo deste texto é realizar uma análise utilizando o banco de dados **colon** do pacote "survival". Esse conjunto de dados possui informações de morte e de recorrência de câncer de colon para 929 pessoas. Há duas linhas para cada pessoa, sendo que essas são caracterizadas pela variável "etype", onde "etype == 1" indica recorrência e "etype == 2" indica morte. Além disso, esse banco possuim outras 15 variáveis, que são:

id: que identifica o paciente;
study: vale 1 para todos os pacientes;
rx: tratament - Obs(ervation), Lev(amisole), Lev(amisole)+5-FU;
sex: 1 para masculino;
age: idade em anos;
obstruct: obstrução do colon pelo tumor;
perfor: perfuração do colon;
adhere: aderência do tumor à órgão adjacentes;
nodes: nº de nódulos linfáticos com câncer detectado;
time: tempo de sobrevivência;
status: status de censura;
differ: diferenciação do tumor (1 = boa, 2 = moderada, 3 = ruim);
extent:extenção da propagação local (1 = submucosa, 2 = músculo, 3 = serosa, 4 = contíguo);
sug: tempo da cirurgia até o registro de câncer;
node4: mais que 4 nódulos linfáticos positivos.

Vamos carregar o banco de dados:

```
names(colon)
```

```
##  [1] "id"       "study"    "rx"       "sex"      "age"      "obstruct"
##  [7] "perfor"   "adhere"   "nodes"    "status"   "differ"   "extent"
## [13] "surg"     "node4"    "time"     "etype"
```

Vamos então, primeiramente, realizar a análise apenas para o evento morte, isto é, "etype == 2".

Vamos olhar para as variáveis desses dados, que pela descrição aparentam, com exceção de **age** e **nodes**, ser todas categóricas. Com isso, vamos dicotomizar essas variáveis, separando **age** nos 65 anos, para comparar jovens e idosos e separando **nodes** em 4 nódulos com câncer detectado e vamos utilizar essa variável, aqui chamada de **nodesc** no lugar de **node4** uma vez que ambas querem dizer exatamente a mesma coisa (mais de quatro nódulos linfáticos com câncer).

```
## rx
##     Obs     Lev Lev+5FU
##     315     310     304
```

```
## sex
##   0   1
## 445 484
```

```
## obstruct
##   0   1
## 749 180
```

```
## perfor
##   0   1
## 902  27
```

```
## adhere
##   0   1
```

```
## 794 135
```

```
## differ
##   1   2   3
##  93 663 150
```

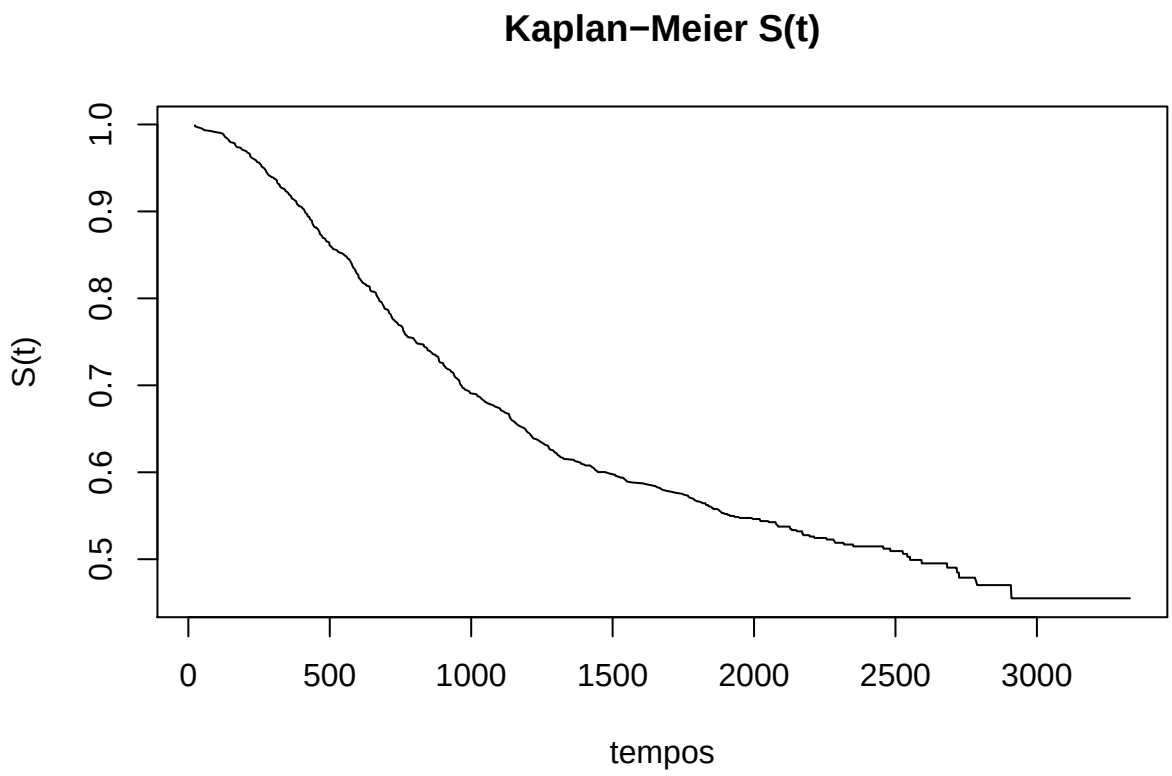
```
## extent
##   1   2   3   4
##  21 106 759  43
```

```
## surg
##   0   1
## 682 247
```

```
## agec
##   0   1
## 595 334
```

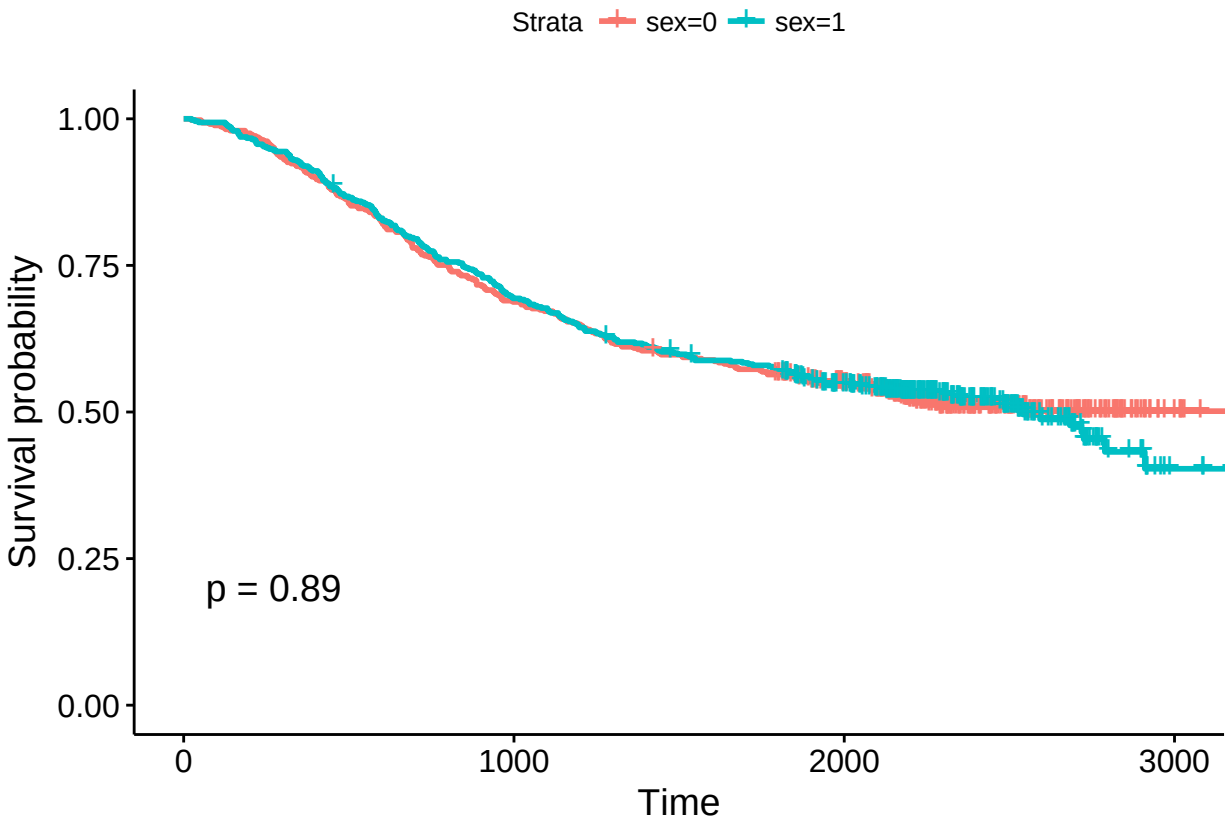
```
## nodesc
##   0   1
## 679 232
```

Com isso, a titulo de visualização vamos plotar a função de sobrevivência ajustada por Kaplan-Meier.



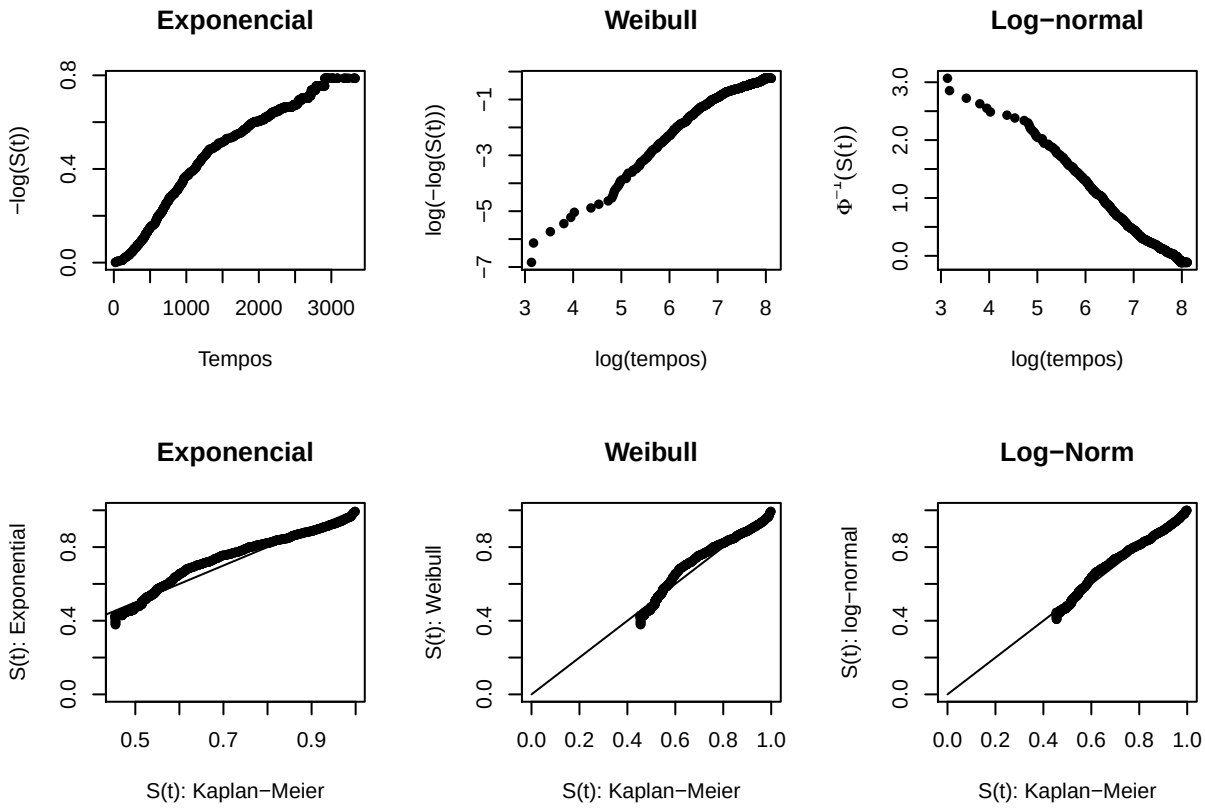
De imediato já é possível notar que, comparado com as situações vistas anteriormente no curso, agora temos uma quantidade de dados bastante grande, o que faz com a função fique mais suave e perca um pouco daquela forma de escada.

Novamente a titulo de visualização, vamos observar se há alguma diferença entre as sobrevivências de homens e mulheres para o evento morte por câncer de cólon.



O p-valor do teste logrank impresso no gráfico mostra que não há diferenças significativas entre as sobrevivências de homens e mulheres.

Vamos agora, verificar se é possível realizar o ajuste utilizando um modelo paramétrico. Para tanto, iremos plotar os gráficos de linearização e das funções de sobrevivência estimadas versus a função de sobrevivência de Kaplan-Meier para tentar decidir pelo modelo mais adequado.



Pelo gráficos aparentemente nenhum modelo é adequado para esse conjunto de dados. Assim, vamos realizar o teste da razão de verossimilhança com modelos encaixados, comparando esses três modelos com a Gama Generalizada para obter uma medida quantitativa da adequação desses modelos aos dados.

```
##                   Log-Verossimilhança     TRV p-valor
## Gama Generalizada               -4103
## Exponencial                     -4132 56.5179       0
## Weibull                         -4132 56.5175       0
## Log-Normal                      -4107  6.2149  0.0127
```

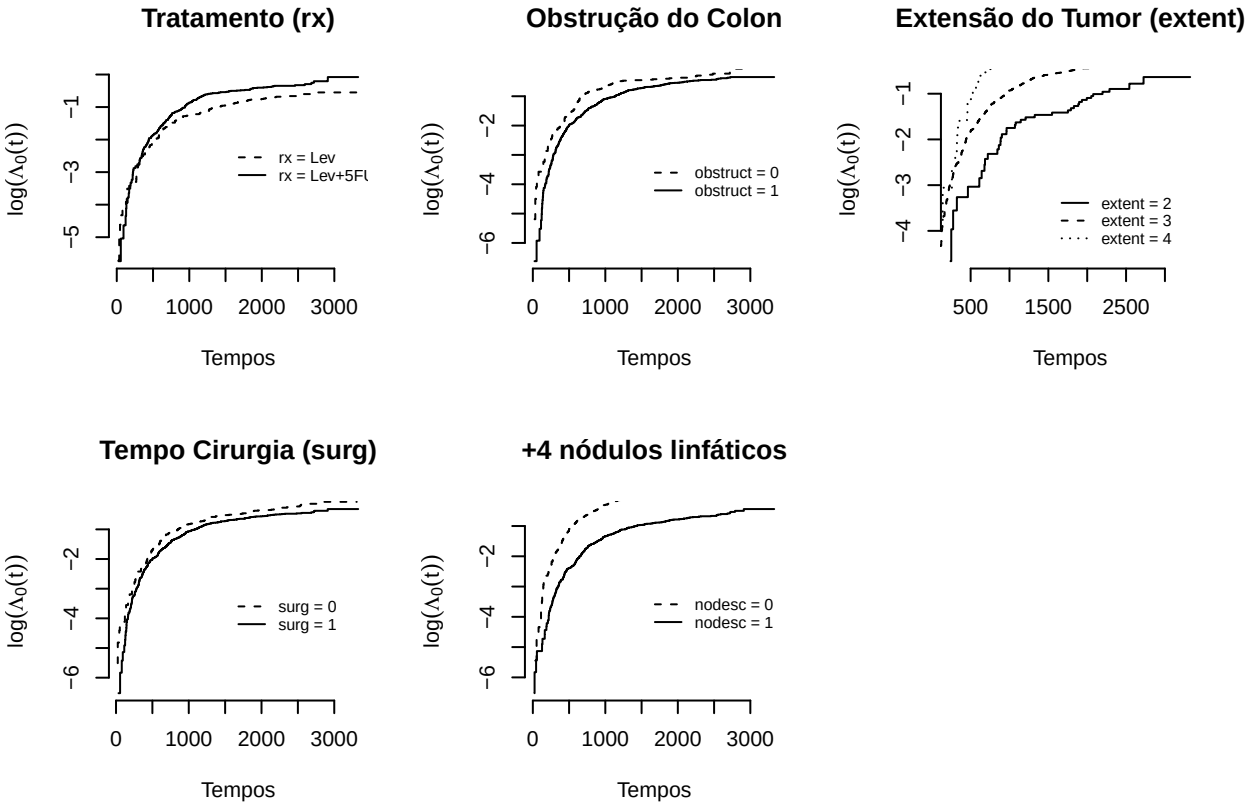
Pela tabela acima, olhando para a última coluna, vemos que todos os modelos foram rejeitados. Dessa forma, vamos prosseguir para o ajuste do modelo de Cox, já que a sua natureza semi paramétrica permite maior flexibilidade. Vamos então utilizar os passos para seleção de variáveis presentes no capítulo 4 do livro texto do curso. Primeiramente, vamos ajustar modelos com apenas uma variável.

```
##          log-veros  TRV  p-valor
## nulo         -2930
## rx           -2924 12.1  0.00231
## sex          -2930 0.02    0.887
## obstruct     -2928 5.05   0.0246
## perfor       -2930 0.33    0.565
## adhere       -2927 6.04    0.014
## nodesc       -2807 86.3        0
## differ       -2840 15.2 0.000487
## extent       -2916 29.2 2.03e-06
## surg         -2928 5.01   0.0253
```

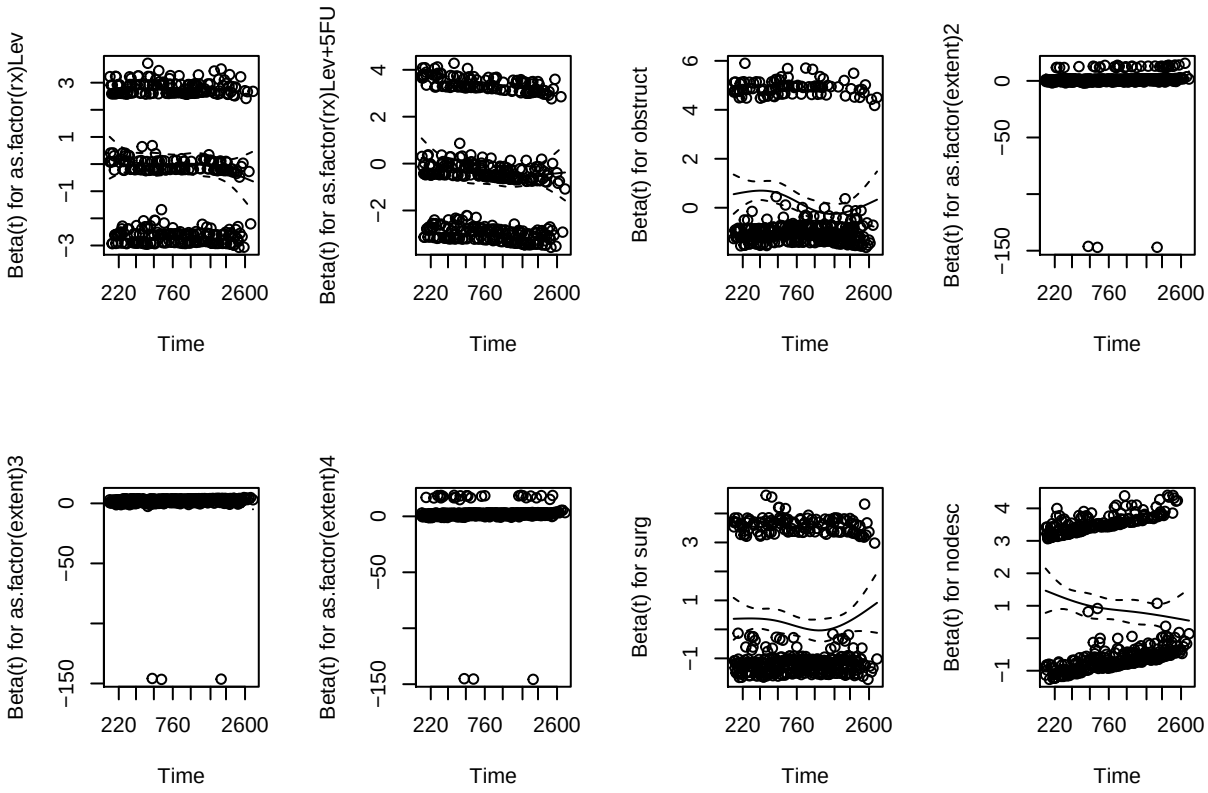
Pegando somente as variáveis que tiveram significância de pelo menos 10% e ajustando um modelo.

Então, após multiplos ajutes, excluindo e adicionando variáveis, para verificar a signifcância delas na presença e na ausência umas das outras, decidiu-se por retirar apenas as variáveis **adhere** e **differ**, pois eram as únicas que não se mostraram significativas quando se ajustou modelos com múltiplas covariáveis. Portanto o modelo final ficou com as variáveis: **rx**, **obstruct**, **extent**, **surg** e **nodesc**.

Vamos agora utilizar as técnicas gráficas e análise de resíduos para verificar a adequação do modelo de Cox.



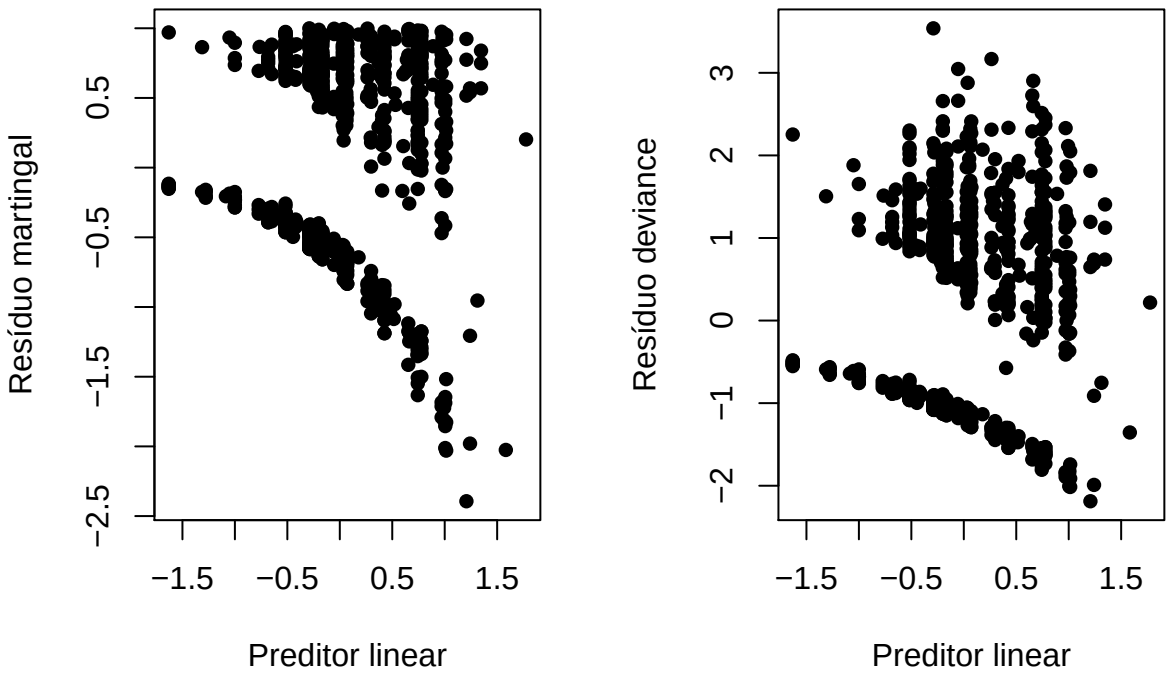
Por esses gráficos é possível notar que a variável **rx** viola a hipótese de taxas de falha proporcionais. Vamos então realizar a análise dos resíduos de Schoenfeld como mais uma forma de atestar a violação dessa hipótese para cada uma das variáveis.



Todos os gráficos nesse caso, com exceção de nodesc, não apresentam nenhum tendência forte, crescente ou decrescente, que seria evidência contra a suposição de taxas de falha proporcionais. Outra medida que pode auxiliar nessa decisão é o coeficiente de correlação de Pearson entre os resíduos de Schoenfeld (ρ) padronizador e o tempo (uma função do tempo $g(t)$ que no caso é a identidade). Valores de ρ próximos de zero indicam que não há evidência para a rejeição da suposição.

```
##                          rho    chisq      p
## as.factor(rx)Lev     -0.05945  1.57438 0.2096
## as.factor(rx)Lev+5FU -0.06135  1.67639 0.1954
## obstruct             -0.10374  4.80314 0.0284
## as.factor(extent)2    0.06067  1.62744 0.2021
## as.factor(extent)3    0.02584  0.29598 0.5864
## as.factor(extent)4    0.02043  0.18476 0.6673
## surg                  0.00242  0.00259 0.9594
## nodesc               -0.09732  4.09235 0.0431
## GLOBAL                     NA 18.91386 0.0153
```

Veja que a variável **rx** que apresentou violação no gráfico descrivito, não teria a suposição rejeitada no teste com os resíduos. O mesmo já não se pode dizer para as variáveis **obstruct** e **nodesc**. Porém, como fica evidente nos gráficos descritivos que os riscos de ambas aparentam ser proporcionais, vamos manter essas variáveis no modelo. Por fim, vamos realizar o gráfico com os resíduos martingal e deviance a fim de detectar a existência de pontos influentes.



Veja que embora esses gráficos não apresentem nenhum ponto muito discrepante que pudesse ser considerado ponto influente, pode-se que dizer que eles apresentam um padrão no mínimo estranho. Pelo menos, não visto durante as aulas ou nos exemplos do livro. Dessa forma, vamos então prosseguir com a análise dessa vez filtrando os dados para o evento recorrência, isto é "etype == 1".

Novamente, dicotomizando as variáveis **age** nos 65 anos e **nodes** em 4 nódulos com câncer detectado. Temos:

```
## rx
##     Obs     Lev Lev+5FU
##     315     310     304

## sex
##   0   1
## 445 484

## obstruct
##   0   1
## 749 180

## perfor
##   0   1
## 902  27

## adhere
##   0   1
## 794 135

## differ
##   1   2   3
##  93 663 150
```

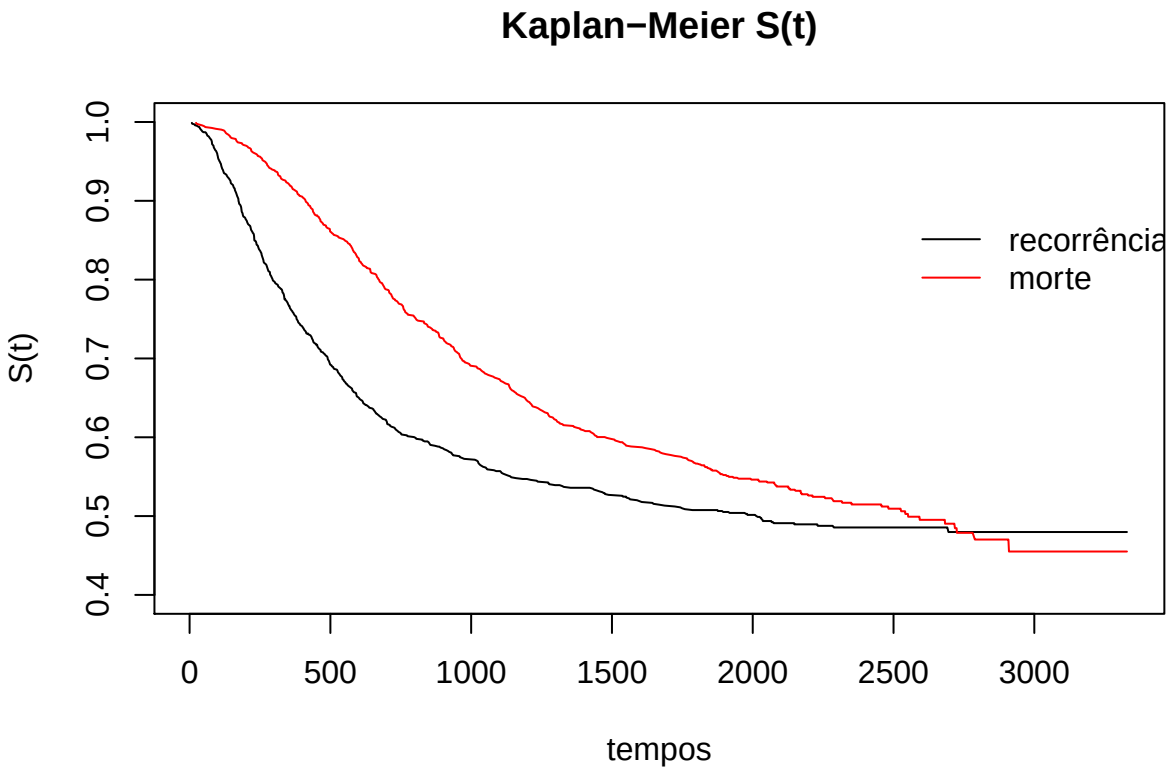
```
## extent
##   1   2   3   4
##  21 106 759  43
```

```
## surg
##   0   1
## 682 247
```

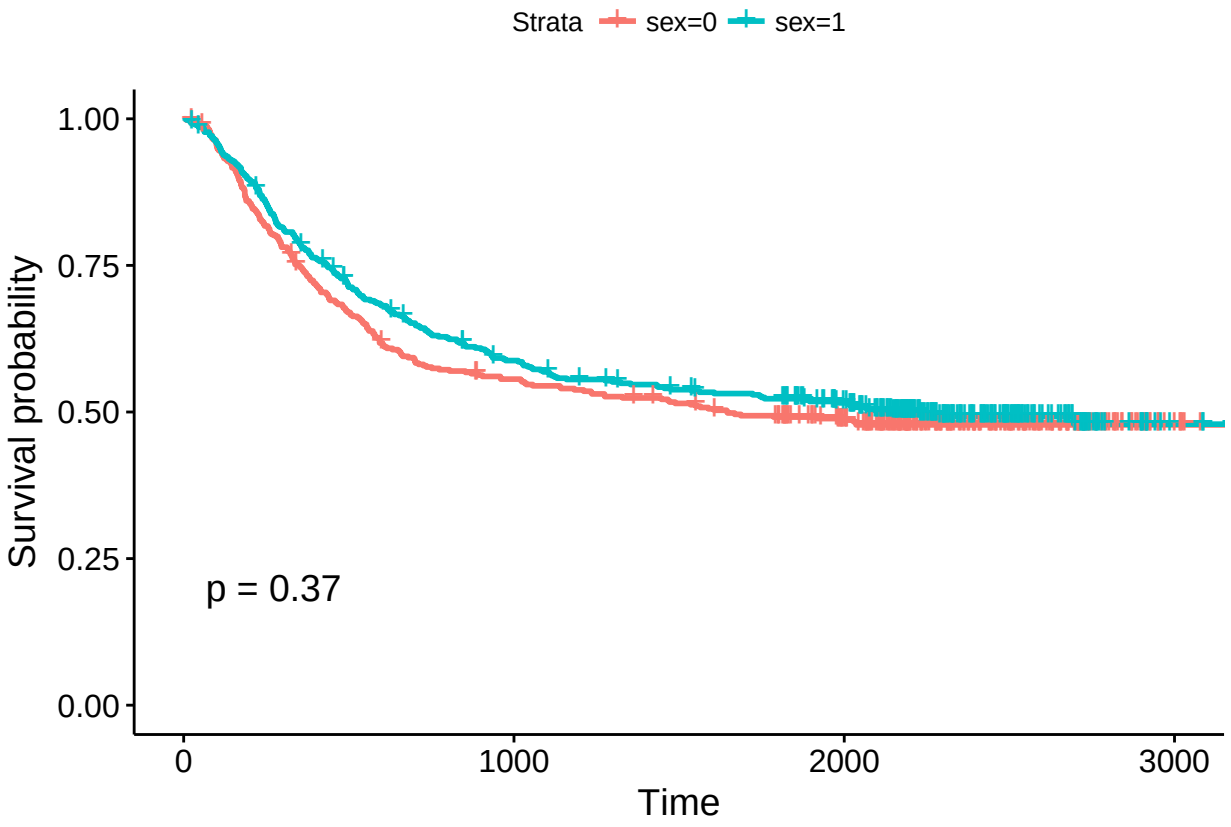
```
## agec
##   0   1
## 595 334
```

```
## nodesc
##   0   1
## 679 232
```

Com isso, plotando a função de sobrevivência estimada por Kaplan-Meier:



Olhando para o gráfico fica evidente a diferença entre as sobrevivências dos dois tipos de evento. Como esperado a recorrência acontece mais rápido do que a morte. Vamos então, mais uma vez, verificar se exite alguma diferença entre as sobrevivências de homens e mulheres.

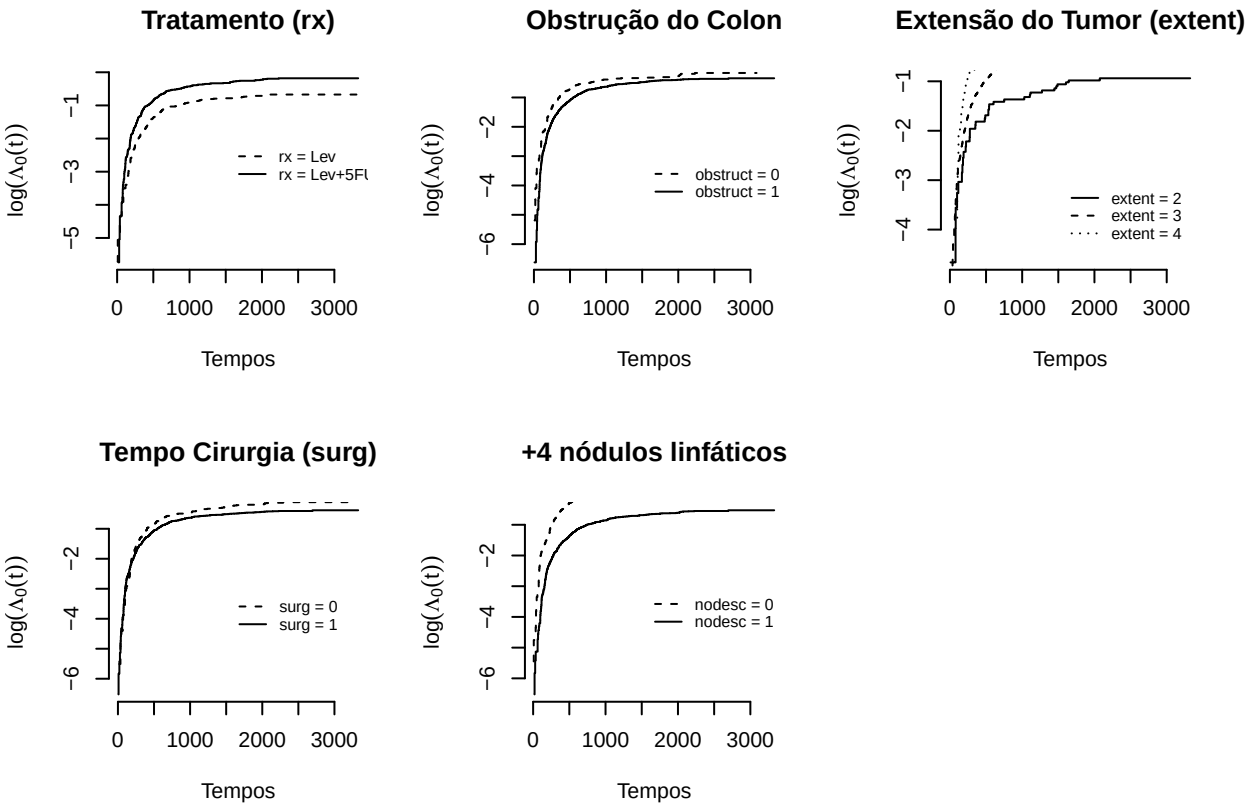


Outra vez, pelo p-valor impresso no gráfico, mostra-se que não há diferenças significativas entre homens e mulheres. Como já na primeira análise todos os modelos paramétricos que foram testados, foram rejeitados. E como o objetivo agora é verificar se há diferenças entre os modelos de Cox para o evento **morte** e para o evento **recorrência**. Vamos partir diretamente para a seleção de variáveis e ajuste do modelo de Cox para esse conjunto de dados.

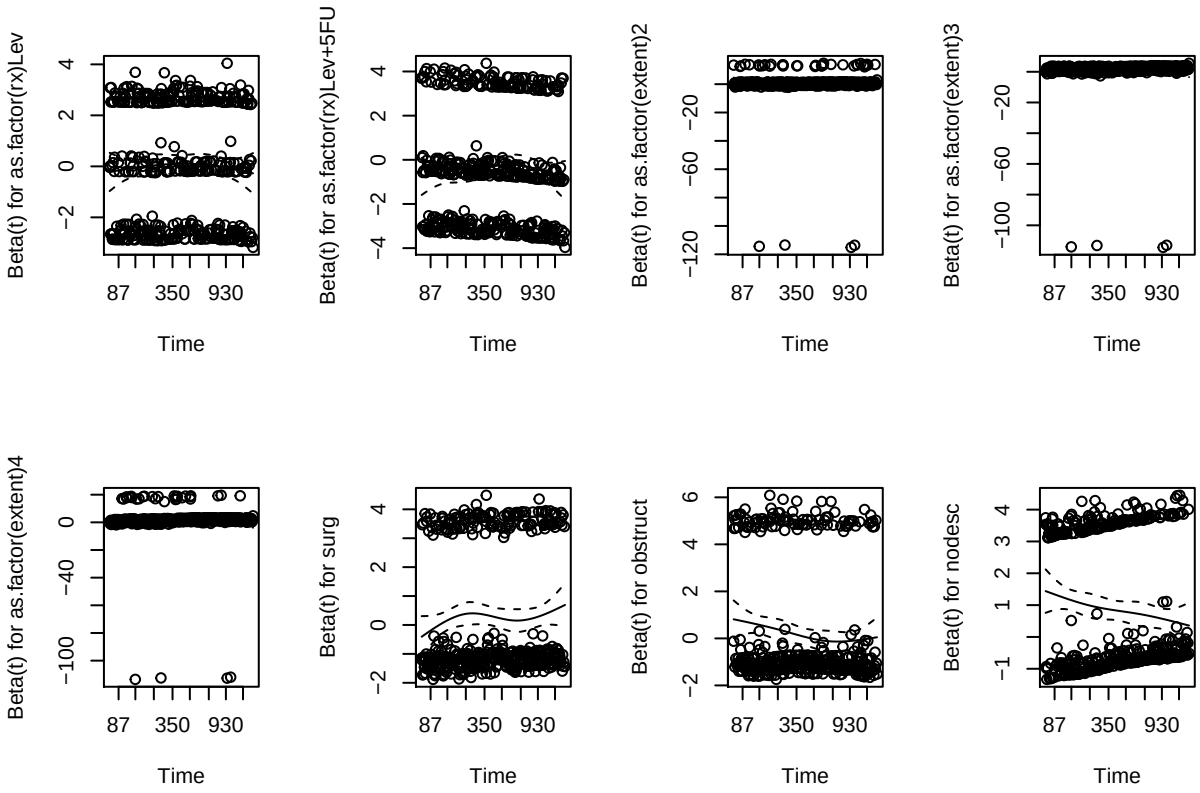
```
##          log-veros  TRV  p-valor
## nulo         -3040
## rx           -3028 24.3 5.23e-06
## sex          -3040 0.81    0.367
## obstruct     -3038 4.18    0.041
## perfor       -3039 1.89    0.169
## adhere       -3037  6.2   0.0127
## nodesc       -2917 73.9        0
## differ       -2957 13.1  0.00147
## extent       -3024 32.6 3.85e-07
## surg         -3037 6.17    0.013
```

O próximo passo mais uma vez seria olhar para a tabela acima e ajustar um modelo com todas as variáveis com pelo menos 10% de significância. Novamente, após múltiplos ajustes decidiu-se por manter as variáveis **rx**, **extent**, **obstruct**, **surg** e **nodesc**.

Dessa forma, vamos então novamente verificar o adequação do modelo de Cox verificando a suposição de taxas de falha proporcionais.



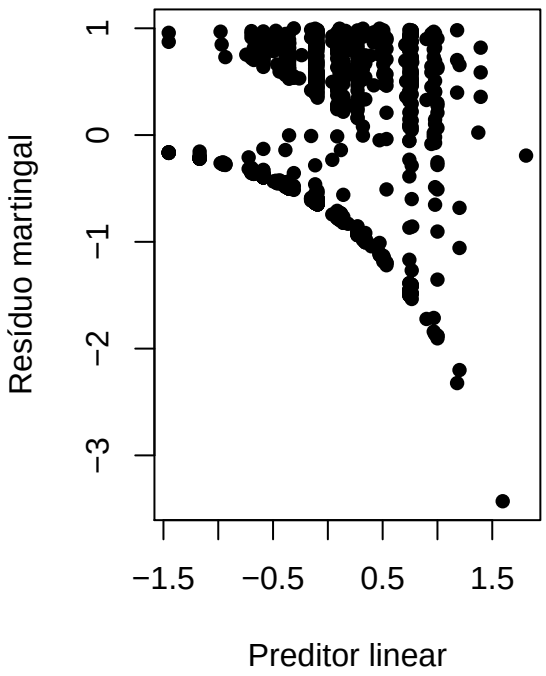
Dessa vez, tanto a variável **surg** quanto a variável **extent** parecem violar a hipótese de taxas de falha proporcionais. Vamos realizar a análise de resíduos para verificar se esse realmente é o caso.

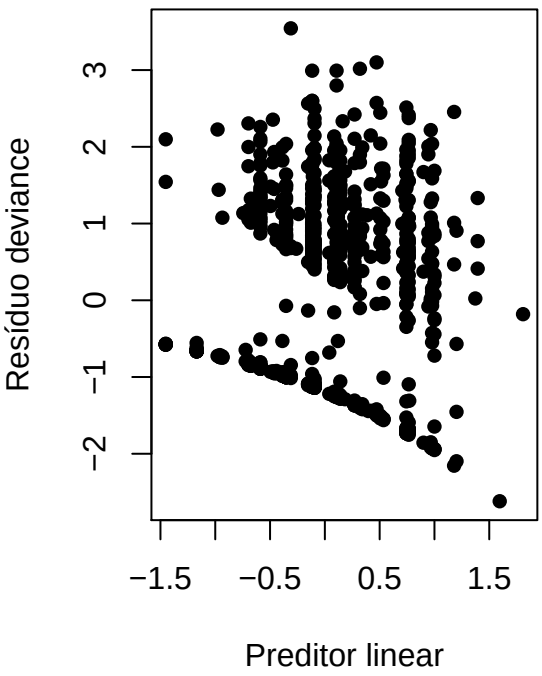


Pelos gráficos dos resíduos de Schoenfeld, apenas a variável **nodesc** apresenta um padrão decrescente nítido que poderia estar causando a violação da hipótese. vamos então realizar o teste da correlação de Pearson entre os resíduos padronizados de Schoenfeld e o tempo.

```
##                           rho   chisq      p
## as.factor(rx)Lev     -0.01561  0.1131 0.7366
## as.factor(rx)Lev+5FU -0.01005  0.0466 0.8290
## as.factor(extent)2    0.01571  0.1128 0.7370
## as.factor(extent)3    0.01762  0.1430 0.7053
## as.factor(extent)4   -0.00697  0.0223 0.8813
## surg                  0.05897  1.6066 0.2050
## obstruct             -0.06759  2.1389 0.1436
## nodesc               -0.10962  5.3556 0.0207
## GLOBAL                     NA 10.8586 0.2098
```

De fato, esse resultado corrobora o que foi apresentado no gráfico dos resíduos, isto é, apenas a variável **nodesc** apresentou violação da hipótese. Mesmo assim, vamos plotar os resíduos martigal e deviance para verificar a existência de pontos influentes e, então, vamos estratificar o banco de dados por sexo, fazendo a análise para os homens, para verificar se a hipótese de taxas de falha proporcionais se mantem para os dados estratificados.





Mais uma vez os resíduos apresentaram aquele padrão estranho. Dessa vez o gráfico dos resíduos martingal apresentou um ponto próximo de -3.5 que parece ser um ponto influente, mas vemos que o correspondente desse ponto no gráfico para os resíduos deviance mostra que esse ponto se encontra compreendido em uma região aceitável dentro da variação observada para esses resíduos.

Estratificando os dados por sexo, no caso vamos utilizar a estratificação para "sex == 1", para o evento recorrência, isto é "etype == 1".

Mais uma vez vamos separa as variáveis **age** e **nodes** em dos grupos, sendo **age** nos 65 anos e **nodes** em 4 nódulos com câncer detectado.

```
## rx
##     Obs     Lev Lev+5FU
##     166     177     141
```

```
## sex
##   1
## 484
```

```
## obstruct
##   0   1
## 396  88
```

```
## perfor
##   0   1
## 470  14
```

```
## adhere
##   0   1
## 417  67
```

```
## differ
##   1   2   3
##  42 349  77
```

```
## extent
##   1   2   3   4
##   8  59 393  24
```

```
## surg
##   0   1
## 354 130
```

```
## agec
##   0   1
## 311 173
```

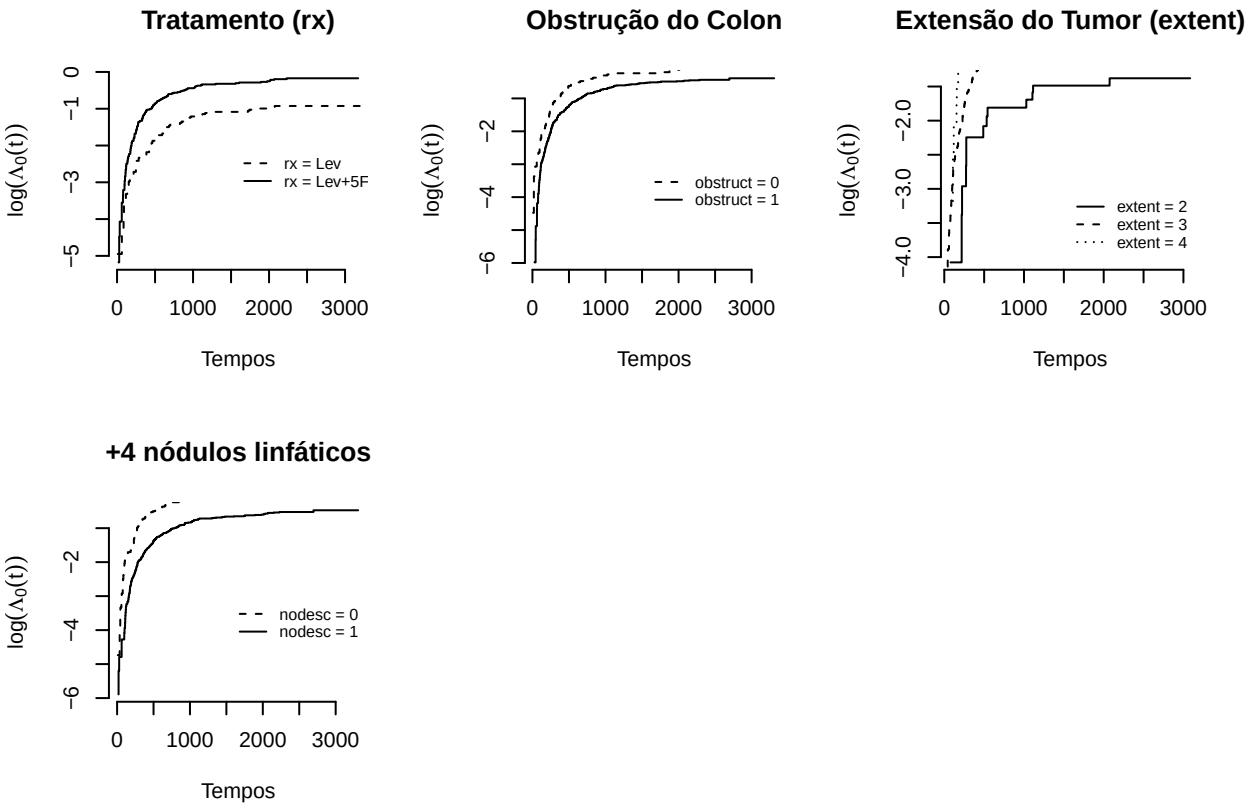
```
## nodesc
##   0   1
## 362 114
```

Não vamos perder tempo dessa vez olhando para as funções de sobrevivência estimadas por Kaplan-Meier, pois as mesmas já foram apresentadas nas análises anteriores.

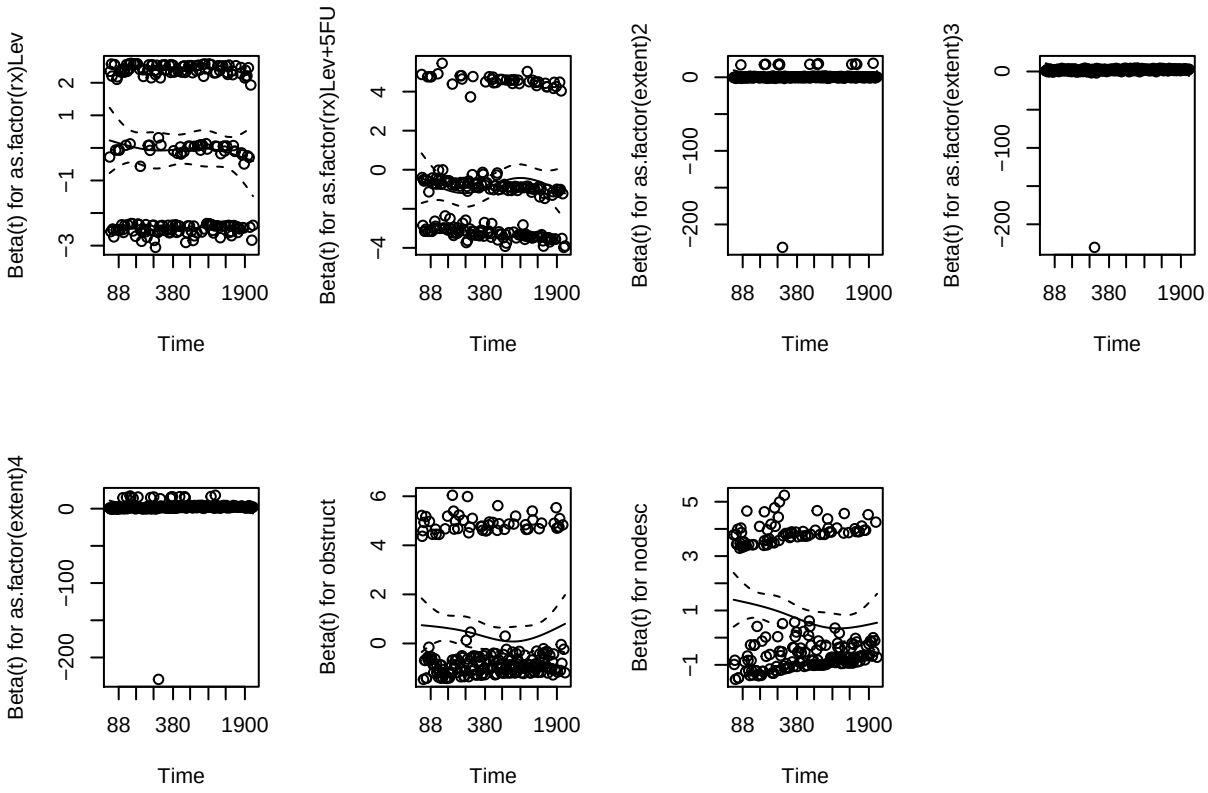
```
##          log-veros  TRV  p-valor
## nulo         -1396
## rx           -1382 27.2 1.26e-06
## obstruct     -1392 8.18  0.00425
## perfor       -1396    0    0.978
## adhere       -1395 2.59    0.108
## nodesc       -1341 22.8 1.83e-06
## differ       -1336 10.3  0.00578
## extent       -1379 32.7 3.74e-07
## surg         -1395 1.41    0.235
```

Olhando para a tabela acima é interessante notar que variáveis que antes eram significativas, agora não são mais. Como é o caso de **adhere** e **surg**. Dessa maneira, como anteriormente, após múltiplos ajustes decidiu-se por manter as variáveis **rx**, **obstruct**, **nodesc** e **extent**.

Dessa forma, vamos então novamente verificar o adequação do modelo de Cox verificando a suposição de taxas de falha proporcionais.



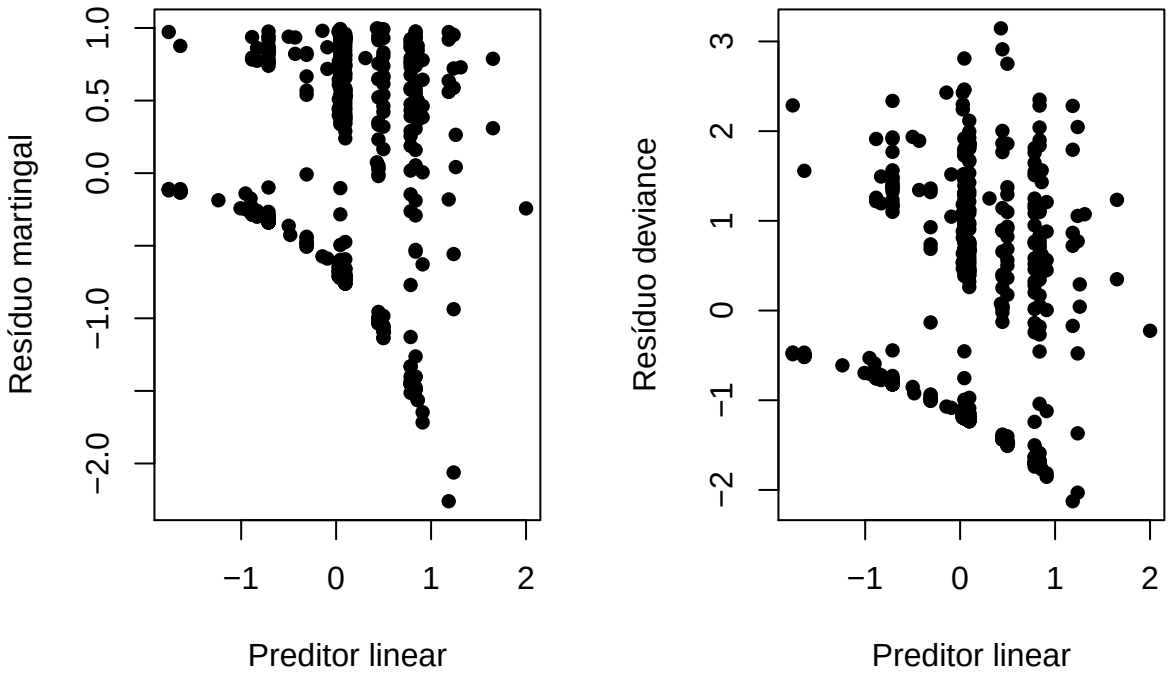
Dessa vez, parece que nenhuma variável está violando a hipótese de taxa de falha prorporcional, no entanto dessa vez temos muito menos observações. Vamos prosseguir para a análise dos resíduos.



Mais uma vez apenas a variável **nodesc** apresenta um padrão decrescente nítido, embora essa angulação esteja bem menor do que antes. Vamos então realizar o teste com os coeficientes de correlação.

```
##                          rho  chisq      p
## as.factor(rx)Lev     -0.0409 0.3898 0.5324
## as.factor(rx)Lev+5FU  0.0132 0.0404 0.8406
## as.factor(extent)2    0.0463 0.4934 0.4824
## as.factor(extent)3    0.0576 0.7649 0.3818
## as.factor(extent)4    0.0347 0.2775 0.5984
## obstruct              0.0140 0.0461 0.8299
## nodesc               -0.1097 2.7817 0.0953
## GLOBAL                   NA 5.5241 0.5963
```

Outra vez se vê um problema com a variável **nodesc**. O mais interessante de tudo é que essa variável em momentou algum apresentou o cruzamento das funções de taxa de falha acumulada para os dois níveis dessa variável, enquanto que outras variáveis, como o caso do **rx** na primeira análise, apresentaram esse cruzamento, mas o teste com o coeficiente de correlação não apontou evidências para a rejeição da suposição de taxas de falha proporcionais.



Novamente, os resíduos martingal e deviance apresentam aquele padrão perturbador. No entando, no livro texto não explicita se esses resíduos precisam apresentar um padrão aleatório. O que chama a atenção é que os autores comentam sobre "a forma funcional das variáveis" e nesse caso aparentemente esse pode ser o problema. Talvez uma transformação das variáveis resolvesse esse problema dos resíduos. No entanto, dada a extensão da análise e o grande número de variáveis, teríamos dificuldade em saber qual delas está causando esse problema. Portanto, aqui não iremos perseguir esse caminho para tentar resolver essa questão.

Vamos então realizar a interpretação dos coeficientes para esse último modelo, no caso estaremos falando da recorrência do câncer de cólon em homens.

```
## Call:
## coxph(formula = Surv(time, status) ~ as.factor(rx) + as.factor(extent) +
##     obstruct + nodesc, data = menstrat, x = T, method = "breslow")
##
##   n= 476, number of events= 232
##    (8 observations deleted due to missingness)
##
##                          coef exp(coef) se(coef)     z Pr(>|z|)
## as.factor(rx)Lev     -0.0521    0.9492   0.1458 -0.36    0.721
## as.factor(rx)Lev+5FU -0.8085    0.4455   0.1848 -4.37  1.2e-05 ***
## as.factor(extent)2    0.1219    1.1296   1.0388  0.12    0.907
## as.factor(extent)3    1.0508    2.8600   1.0050  1.05    0.296
## as.factor(extent)4    1.8655    6.4589   1.0314  1.81    0.071 .
## obstruct              0.4012    1.4936   0.1590  2.52    0.012 *
## nodesc                0.7404    2.0967   0.1438  5.15  2.6e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##                       exp(coef) exp(-coef) lower .95 upper .95
## as.factor(rx)Lev          0.949      1.054     0.713      1.26
## as.factor(rx)Lev+5FU      0.446      2.245     0.310      0.64
## as.factor(extent)2        1.130      0.885     0.147      8.65
## as.factor(extent)3        2.860      0.350     0.399     20.50
## as.factor(extent)4        6.459      0.155     0.855     48.77
## obstruct                  1.494      0.670     1.094      2.04
## nodesc                    2.097      0.477     1.582      2.78
##
## Concordance= 0.667  (se = 0.019 )
## Rsquare= 0.165   (max possible= 0.997 )
## Likelihood ratio test= 85.6  on 7 df,   p=9.99e-16
## Wald test            = 78.3  on 7 df,   p=3e-14
## Score (logrank) test = 84.8  on 7 df,   p=1.44e-15
```

Com isso, olhando para a exponencial dos coeficientes vemos que a recorrência em indivíduos que receberam o Lev(amisole)+5-FU é reduzida em aproximadamente 56% comparada com a recorrência do grupo de observação. Enquanto que os pacientes que receberam apenas o tratamento Lev(amisole) a recorrência caiu apenas 5% em relação ao grupo de observação.

Já em relação à extensão da propagação do câncer é possível notar que a recorrência aumenta muito para conforme a extensão do câncer vai se agravando. Sendo crítico para àqueles em que o câncer atingiu as estruturas contíguas ("extent == 4"), esses paciente tem um risco de apromadamente 6.5 vezes maior do que os pacientes onde a extensão atingiu apenas a submucosa ("extent == 1").

Para a obstrução do cólon pelo tumor, a recorrência nos indivíduos que apresentam a obstrução é por volta de 50% maior do que a recorrência nos pacientes que não apresentam a obstrução.

Por fim, para os pacientes que possuem mais de 4 nódulos linfáticos com câncer detectável a recorrência é aproximadamente 2.1 vezes maior do que a dos indivíduos com 3 nódulos ou menos.

Conclusão

Nessa análise deu para sentir realmente os problemas que seriam enfrentados em uma análise clínica real. O processo de seleção de variáveis se torna bastante enfadonho a medida que o número de variáveis cresce.

Além disso, foi interessante notar que mesmo tendo uma quantidade enorme dados, o que intuitivamente faz pensar que um modelo paramétrico poderia ser bem ajustado, se a estrutura do modelo paramétrico não for compátivel com os dados de sobrevivência, não importa o quanto se tente ajustar, esse modelo jamais ficará bom. Nesse caso é mais indicado utilizar o modelo de Cox, pois sua natureza semi paramétrica permite maior flexibilidade. Isso talvez explique a popularidade desse modelo em análise de sobrevivência.

Outro ponto interessante que poderia ser explorado é a questão da forma funcional das covariáveis. O livro em si não traz nenhum exemplo de como iriamos ajustar essa forma funcional quando temos apenas variáveis categóricas.

Por fim, em relação aos dados em si, vários aspectos surpreenderam, por exemplo a variável **adhere** não ser significativa é algo extremamente curioso, pois intuitivamente a aderência do câncer à órgãos adjacentes deveria impactar tanto na recorrência como na morte do indivíduo, no caso do câncer se espalhar pelo organismo a ponto de se tornar incurável.

Portanto, fazer esse trabalho possibilitou não somente analisar esses aspectos dos dados em si e das estruturas do modelo. Mas também possibilitou adquirir uma melhor compreensão do poder da modelagem de sobrevivência em estudos clínicos.