```
## -- Attaching packages ---------------------------------------------------------------
------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------
- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Questão 4 - Análise de Dados

Suspeita-se que o número de divisões celulares possa se associar ao risco de desenvolvimento de câncer. Realize uma análise de dados, incluindo descritiva, culminando na proposição de um modelo de regressão que mostre a existência (ou não) de associação entre estas duas variáveis (risco de câncer deve ser a variável resposta). Apresente gráficos, intervalos de confiança, testes de hipótese e qualquer outro recurso estatístico para justificar suas decisões.

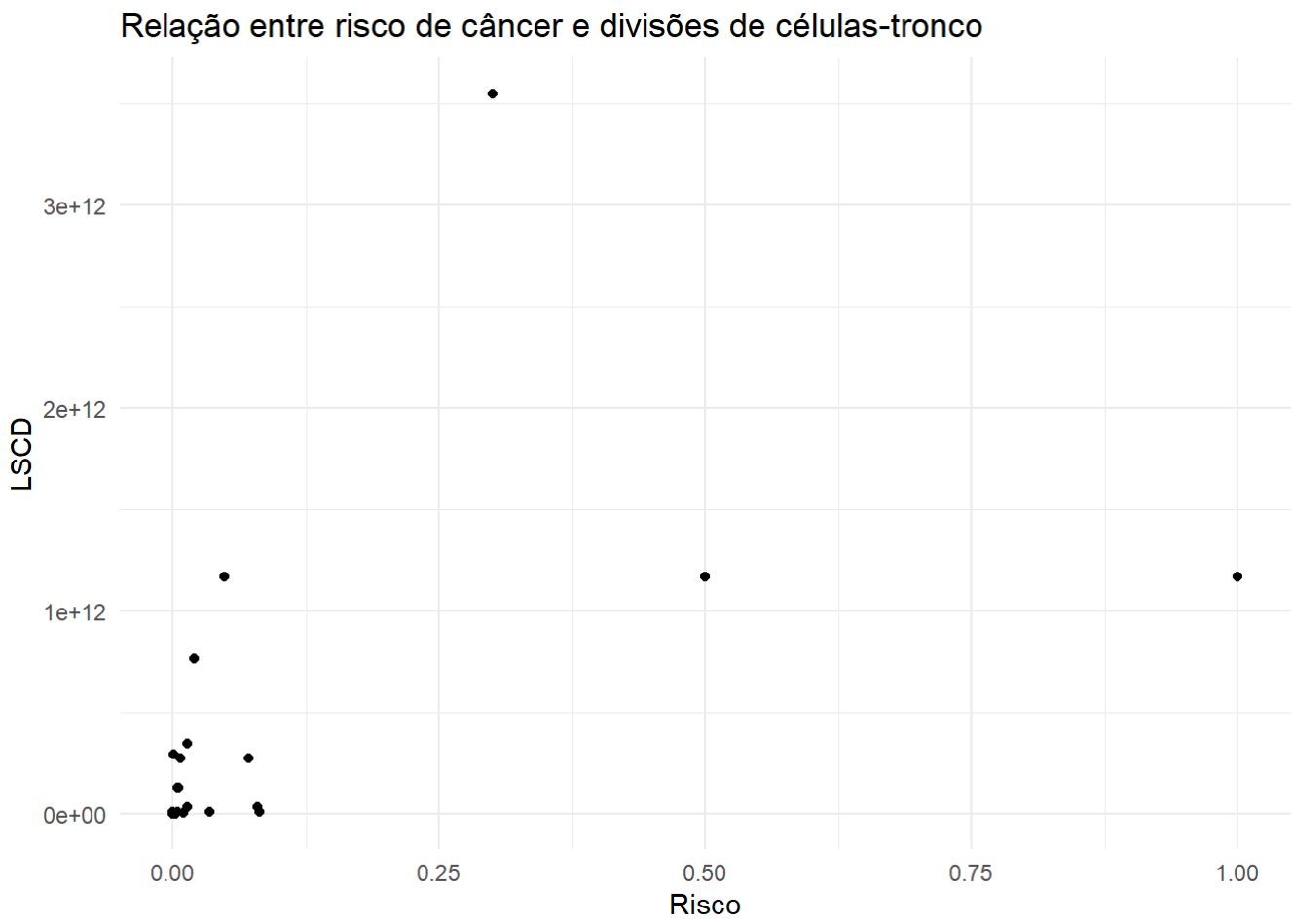
```
cancer <- read_excel("C:/Users/55199/Downloads/cancer.xlsx")
```

Estudando os dados preliminarmente

```
head(cancer)
```

```
## # A tibble: 6 x 3
##   TYPE                                              LSCD   RISK
##   <chr>                                            <dbl>  <dbl>
## 1 Acute myeloid leukemia                        129900000000 0.0041
## 2 Basal cell carcinoma                         3550000000000 0.3
## 3 Chronic lymphocytic leukemia                  129900000000 0.0052
## 4 Colorectal adenocarcinoma                    1168000000000 0.048
## 5 Colorectal adenocarcinoma with FAP           1168000000000 1
## 6 Colorectal adenocarcinoma with Lynch syndrome 1168000000000 0.5
```

```
cancer %>%
  ggplot(aes(RISK, LSCD)) +
  geom_point() +
  labs(title = "Relação entre risco de câncer e divisões de células-tronco") +
  xlab("Risco") + ylab("LSCD") +
  theme_minimal()
```



Risco de câncer por divisão celular

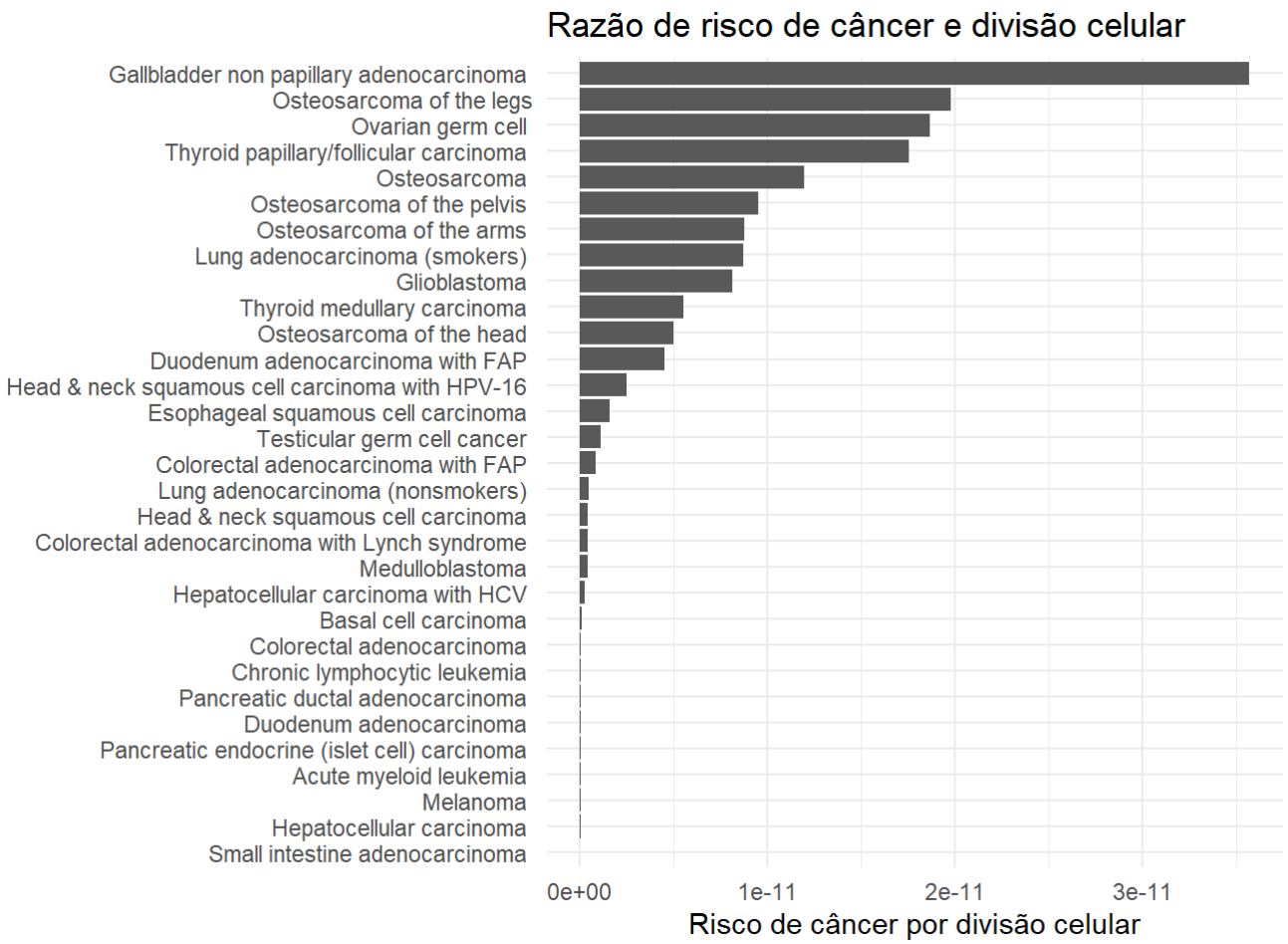
Agora irei estudar a razão entre os riscos de desenvolvimento para cada câncer e o número total de divisões de células-tronco ao longo da vida. Para tanto, criarei uma nova coluna na tabela com essa razão calculada.

```
## Criando coluna da razão entre risco e divisão celular
## RISK/LSCD
cancer = cancer %>%
  mutate(RISK_LSCD = RISK/LSCD) %>%
  arrange(RISK_LSCD)
head(cancer)
```

```
## # A tibble: 6 x 4
##   TYPE                                          LSCD    RISK RISK_LSCD
##   <chr>                                        <dbl>   <dbl>     <dbl>
## 1 Small intestine adenocarcinoma           292200000000 7.00e-4  2.40e-15
## 2 Hepatocellular carcinoma                 270900000000 7.10e-3  2.62e-14
## 3 Melanoma                                 763800000000 2.03e-2  2.66e-14
## 4 Acute myeloid leukemia                   129900000000 4.10e-3  3.16e-14
## 5 Pancreatic endocrine (islet cell) carcino~   6068000000 1.94e-4  3.20e-14
## 6 Duodenum adenocarcinoma                      7796000000 3.00e-4  3.85e-14
```

```
## Criando gráfico de barras para identificar as razões e os tipos de câncer

ggplot(cancer, aes(x = reorder(TYPE, RISK_LSCD), y = RISK_LSCD)) +
  geom_bar(stat = "identity") +
  labs(title = "Razão de risco de câncer e divisão celular") +
  xlab("") + ylab("Risco de câncer por divisão celular") +
  coord_flip() +
  scale_x_discrete(position = "bottom") +
  theme_minimal()
```



Ajustando o modelo

Ao estudarmos preliminarmente a distribuição da relação entre risco de câncer e número total de divisões de células-tronco percebemos que os pontos não estão distribuídos linearmente. Dessa forma, precisaremos fazer uma transformação nos dados. Aqui aplicarei a transformação log nas variáveis preditora e resposta.

```
fit_cancer = lm(log(RISK) ~ log(LSCD), data=cancer)
summary(fit_cancer)
```

```
##
## Call:
## lm(formula = log(RISK) ~ log(LSCD), data = cancer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8019 -1.0722  0.1420  0.9942  2.7870
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.52454    1.66458 -10.528 2.03e-11 ***
## log(LSCD)     0.53264    0.07317   7.279 5.12e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.725 on 29 degrees of freedom
## Multiple R-squared:  0.6463, Adjusted R-squared:  0.6341
## F-statistic: 52.99 on 1 and 29 DF,  p-value: 5.124e-08
```

O modelo linear ajustado é $Y = \beta_0 + \beta_1 * LSCD$, sendo LSCD o número total de divisões de células-tronco ao longo da vida. Ele testa a relação entre o risco de desenvolvimento de câncer e a variável LSCD.

A regressão nos dá que o valor de β_0 é -17.52454, e β_1 é 0.53264.

Podemos interpretar esses parâmetros da seguinte forma:

$\beta_1 = 0.53264$:

O β_1 é o incremento médio esperado na resposta, sendo a resposta o risco, em toda a vida, de desenvolver algum câncer, quando comparamos dois indivíduos cuja diferença é apenas uma unidade na variável preditora (número total de divisões de células-tronco ao longo da vida).

O parâmetro β_1 representa a mudança esperada em y quando x aumenta em uma unidade. Em outras palavras, se pegarmos dois indivíduos e um deles tiver uma unidade a mais de LSCD que o outro, ele terá 0.53264 risco de desenvolver algum tipo de câncer em toda sua vida.

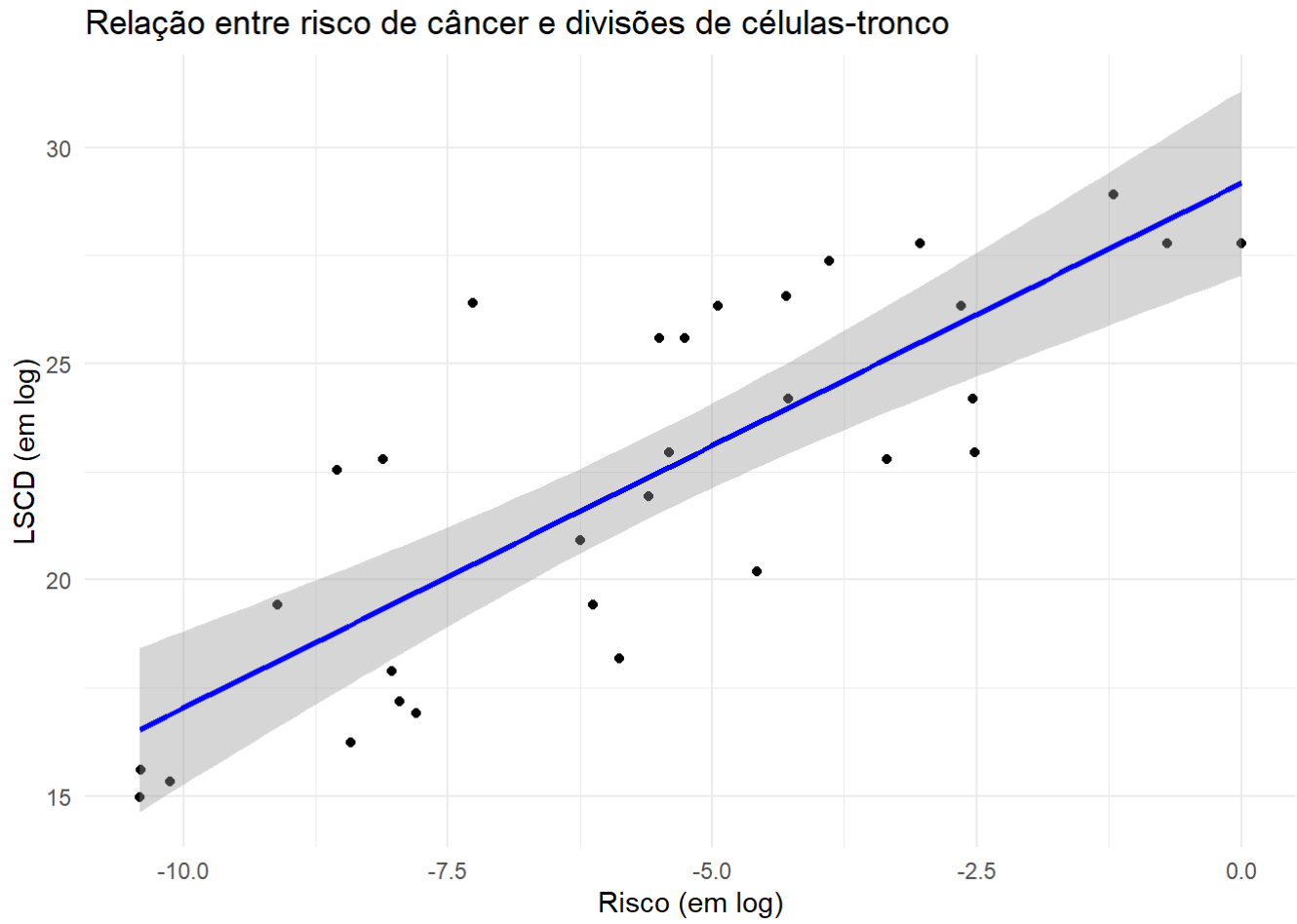
$\beta_0 = -17.52454$:

O parâmetro β_0 é a resposta média para quando x é zero. Em outras palavras, o modelo diz que o risco de desenvolvimento de câncer, em toda a vida, de um indivíduo com 0 LSCD é -17.52454.

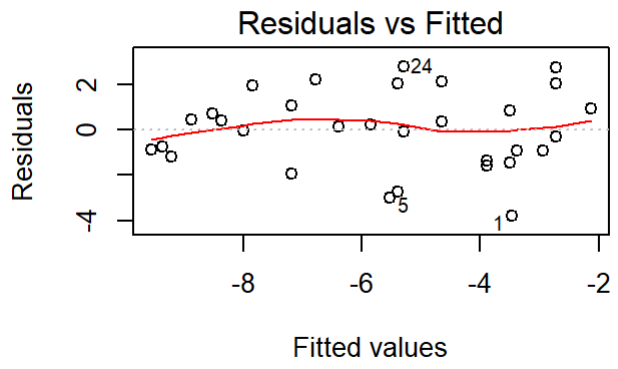
Como não é possível um indivíduo ter um número total de divisões de células-tronco ao longo da vida negativo, podemos assumir que o modelo está matematicamente correto mas que não é coerente com a realidade do fenômeno nessa condição destacada.

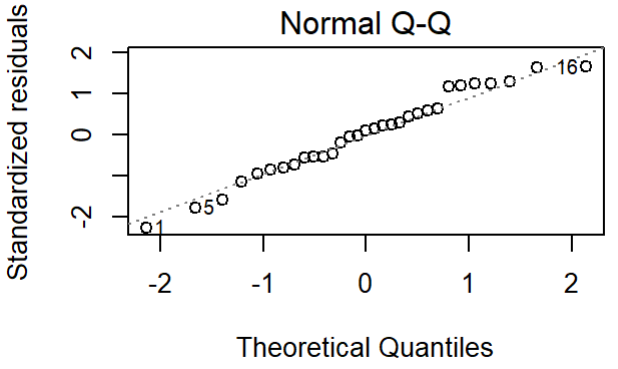
Gráficos da regressão e dos resíduos

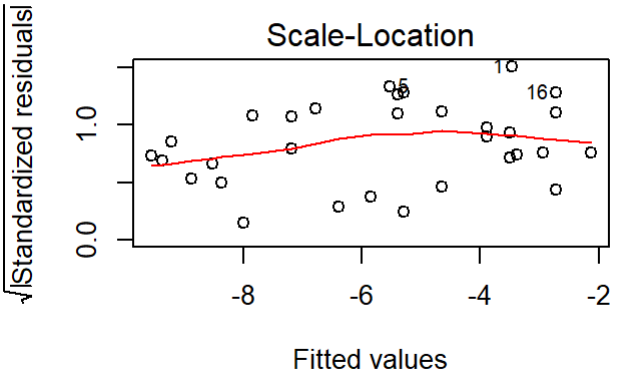
```
cancer %>%
  ggplot(aes(log(RISK), log(LSCD))) +
  geom_point() +
  labs(title = "Relação entre risco de câncer e divisões de células-tronco") +
  xlab("Risco (em log)") +  ylab("LSCD (em log)") +
  geom_smooth(method = "lm", se = TRUE, formula = y ~ poly(x, 1, raw = TRUE), colour = "blue"
) +
  theme_minimal()
```

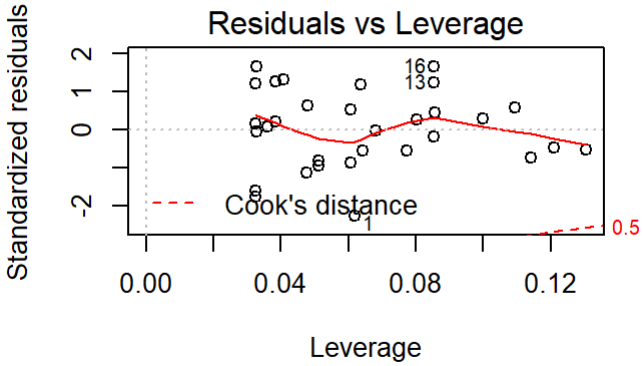



```
par(mfrow=c(2,2))
plot(fit_cancer)
```









O gráfico de Residuals vs Fitted não está centrado em média 0. O ideal é que os pontos formem uma nuvem aleatória (sejam distribuídos de forma aleatória em torno do eixo horizontal); entretanto, a linha vermelha que ultrapassa o gráfico horizontalmente indica a existência da presença de mais pontos positivos entre os valores -5 a -8. Dessa forma, não temos indícios de que a variância dos resíduos seja homoscedástica.

O gráfico Normal Q-Q exibe a relação entre os resíduos normalizados e os quantis teóricos. Se os dados seguissem uma distribuição normal eles iriam seguir a linha traçada diagonalmente, no entanto tem alguns fora dessa reta. Os dados são razoavelmente normais, apesar de notar alguns outliers na cauda superior (cauda superior pesada).

No gráfico de Scale-Location, a linha horizontal tem inclinação entre os valores iniciais do intervalo de preditores (-9 a -6, aproximadamente), e suaviza após isso. A linha indica crescimento inicialmente porque os resíduos desses valores estão mais dispersos. Isso indica que os dados não apresentam variação uniforme nas extremidades do intervalo dos preditores (suspeita de heterocedasticidade).

Testando a normalidade dos resíduos

Como a interpretação dos gráficos levantam a suspeita de heterocedasticidade dos resíduos, irei testar a normalidade dos resíduos com o teste de Shapiro-Wilk. As hipóteses a serem testadas são:

H_0: Resíduos seguem distribuição normal H_1: Resíduos de distribuição não normal

```
shapiro.test(fit_cancer$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit_cancer$residuals
## W = 0.97375, p-value = 0.6271
```

No teste de Shapiro-Wilk obtemos W = 0.97375 e p-value = 0.6271. Dessa forma, temos evidências suficientes para não rejeitar H_0 de que os resíduos estão normalmente distribuídos.

b. Leia a reportagem da BBC e escreva um parecer técnico a respeito da reportagem.

O estudo da reportagem foi conduzido por pesquisadores da Universidade Johns Hopkins e da Escola de Saúde Pública Bloomberg, no qual eles afirmaram existir aleatoriedade no desenvolvimento da maioria dos tipos de câncer devido a maneira de como os tecidos do corpo se regeneram: células danificadas são constantemente substituídas por células-tronco, que se dividem para formar novas células. Entretanto, a cada divisão realizada há o risco de que ocorra mutação, o que aumenta a chance do desenvolvimento de algum tipo de câncer naquele tecido.

O modelo aqui ajustado para o banco de dados fornecido é uma regressão linear simples da forma $Y = \beta_0 + \beta_1 * LSCD$, no qual LSCD (variável preditiva) representa o número total de divisões de células-tronco ao longo da vida e Y (variável resposta) é o risco de desenvolvimento de algum tipo de câncer.

Ao calcular a regressão, descobrui-se que o modelo ajustado é $Y = -17.52454 + 0.53264 * LSCD$. Ao analisar os gráficos de diagnósticos apresentados anterioremente é possível notar uma forte associação positiva entre a incidência de câncer e o número de divisões de células-tronco.

É possível identificar que os resíduos não estão centrados em média igual a 0, portanto não temos indícios de que a variância dos resíduos seja homoscedástica. O gráfico Normal Q-Q dos resíduos do modelo indica que eles são normais, apesar de apresentar uma cauda superior pesada. A normalidade dos resíduos veio a ser comprovada através do teste de Shapiro-Wilk.

A análise do modelo pode indicar que mutações aleatórias devido ao processo biológico de divisão celular, esta inerente a natureza do homem, podem explicar a variação no risco entre os tipos de câncer. Em outras palavras, o modelo ajustado associa que quanto mais divisões de células tronco na vida um indivíduo apresenta, maior é a sua chance de desenvolver algum tipo de câncer.

É válido ressaltar, no entanto, que o estudo não considera agravantes para o desenvolvimento do câncer que já são de conhecimento coletivo, tais como estilo de vida ou fatores genéticos hereditários.

Portanto, a variação na incidência de câncer no conjunto de dados pode ser explicada pelo número total de divisões de células-tronco. Dito isto, considerando apenas os cânceres não relacionados à hereditariedade, doenças ou outros fatores, a quantidade LSCD explica a variação residual. Então, sim, de certa forma, o desenvolvimento de alguns tipos de câncer está associada a “má sorte”.