Regressão Linear Múltipla - Exemplo Wage

Julio Hsu, Guilherme Alberto Dutra Camelo, Fernando Souto Lima

2024-10-02

```
# Setup para o relatório Quarto

knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)
```

Introdução

O objetivo deste relatório é desenvolver um modelo de regressão linear múltipla para analisar a relação entre o salário e as características como idade, estado civil, raça, nível de educação, entre outras, de 3000 empregados masculinos na região do Atlântico.

Todas as análises são realizadas com base no conjunto de dados "Wage", editado manualmente por Steve Miller, da Inquidia Consulting (anteriormente Open BI), a partir do suplemento de março de 2011 da Pesquisa Atual de População.

Fonte: `https://www.re3data.org/repository/r3d100011860`

Os Dados

Com a inserção da base de dados mencionado acima, podemos observar que temos um relátorio de 3000 indíviduos representado por 'Rows' e suas respectivas 11 características representado por 'Columns', tal como:

```
library(ISLR)
library(dplyr)

glimpse(Wage)
```

```
Rows: 3,000
Columns: 11
$ year       <int> 2006, 2004, 2003, 2003, 2005, 2008, 2009, 2008, 2006, 2004,~
$ age        <int> 18, 24, 45, 43, 50, 54, 44, 30, 41, 52, 45, 34, 35, 39, 54,~
$ maritl     <fct> 1. Never Married, 1. Never Married, 2. Married, 2. Married,~
$ race       <fct> 1. White, 1. White, 1. White, 3. Asian, 1. White, 1. White,~
$ education  <fct> 1. < HS Grad, 4. College Grad, 3. Some College, 4. College ~
$ region     <fct> 2. Middle Atlantic, 2. Middle Atlantic, 2. Middle Atlantic,~
$ jobclass   <fct> 1. Industrial, 2. Information, 1. Industrial, 2. Informatio~
$ health     <fct> 1. <=Good, 2. >=Very Good, 1. <=Good, 2. >=Very Good, 1. <=~
$ health_ins <fct> 2. No, 2. No, 1. Yes, 1. Yes, 1. Yes, 1. Yes, 1. Yes, 1. Ye~
$ logwage    <dbl> 4.318063, 4.255273, 4.875061, 5.041393, 4.318063, 4.845098,~
$ wage       <dbl> 75.04315, 70.47602, 130.98218, 154.68529, 75.04315, 127.115~
```

Além disso, ao analisar as características ou variáveis correlacionadas à variável resposta "wage", temos os seguintes dados para cada indivíduo:

- `year`: ano em que os dados foram relatados (número inteiro);
- `age`: idade do empregado (número inteiro);
- `maritl`: estado civil (categoria): 1.Solteiro 2.Casado 3.Viúvo 4.Divorciado 5.Separado;
- `race`: raça do empregado (categoria): 1.Branco 2.Negro 3.Asiático 4.Outros;
- `education`: nível educacional (categoria): 1.Abaixo do ensino médio 2.Ensino médio completo 3.Ensino superior em andamento 4.Graduação/Bacharelado 5.Pós-graduação;
- `region`: região do país (apenas Meio-Atlântico);
- `jobclass`: tipo de emprego (categoria): 1.Industrial 2.Informação;
- `health`: nível de saúde do trabalhador (categoria): 1.Saúde intermediária ou inferior 2.Saúde superior ou excelente;
- `health_ins`: possui plano de saúde (categoria): 1.Sim 2.Não;
- `logwage`: logaritmo do salário do trabalhador (número ponto flutuante);
- `wage`: salário bruto do trabalhador (número ponto flutuante).

Análise Exploratória dos Dados

Em seguida, com as análises dos variáveis acima, podemos aprofundar mais para filtrar ou melhorar a base de dados fornecido, visando identicar possíveis ausências de dados, outliers, etc.

```
library(skimr)
skim(Wage)
```

Tabela 1: Data summary

Name	Wage
Number of rows	3000
Number of columns	11
_____	
Column type frequency:	
factor	7
numeric	4
_____	
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
maritl	0	1	FALSE	5	2. : 2074, 1. : 648, 4. : 204, 5. : 55
race	0	1	FALSE	4	1. : 2480, 2. : 293, 3. : 190, 4. : 37
education	0	1	FALSE	5	2. : 971, 4. : 685, 3. : 650, 5. : 426
region	0	1	FALSE	1	2. : 3000, 1. : 0, 3. : 0, 4. : 0
jobclass	0	1	FALSE	2	1. : 1544, 2. : 1456
health	0	1	FALSE	2	2. : 2142, 1. : 858
health_ins	0	1	FALSE	2	1. : 2083, 2. : 917

Variable type: numeric

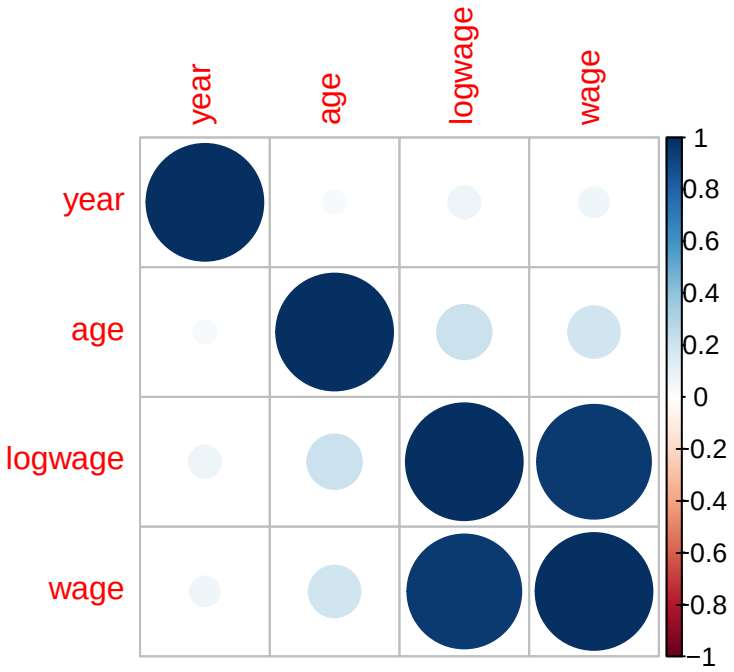
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
year	0	1	2005.79	2.03	2003.00	2004.00	2006.00	2008.00	2009.00	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	42.41	11.54	18.00	33.75	42.00	51.00	80.00	
logwage	0	1	4.65	0.35	3.00	4.45	4.65	4.86	5.76	
wage	0	1	111.70	41.73	20.09	85.38	104.92	128.68	318.34	

Analisando com o resumo de dados acima, podemos notar que a base de dados é divido em 2 dataframe: 1. dados categórico (7 variáveis) 2. dados numéricos (4 variáveis). Nenhum deles apresenta valores perdidos "n_missing". Logo, aproveitando esses variáveis podemos analisar suas respectivas correlações nesta conjuntura de dados...

Análise de Correlação (Gráfico & Tabela)

```
library(corrplot)

num_col <- Wage[sapply(Wage, is.numeric)]

corr <- cor(num_col, use = 'pairwise.complete.obs')

corrplot(corr, method = 'circle')
```



Com o gráfico da correlação dos variáveis numéricos, podemos notar em que existe muita pouca correlação entre as variáveis indenpendentes, porém, especialmente a variável "logwage" podemos notar uma forte correlação com a variável "wage", ou seja, a variável resposta dos nossos dados.

```
library(vcd)

categorical_columns <- Wage[sapply(Wage, is.factor)]

association_results <- data.frame(
  Var1 = character(),
  Var2 = character(),
  CramerV = numeric(),
  stringsAsFactors = FALSE
)

for (i in 1:(ncol(categorical_columns) - 1)) {
  for (j in (i + 1):ncol(categorical_columns)) {
    contingency_table <- table(categorical_columns[[i]], categorical_columns[[j]])  # Correçã
    cramer_v <- assocstats(contingency_table)$cramer
    association_results <- rbind(
      association_results,
      data.frame(
        Var1 = colnames(categorical_columns)[i],
        Var2 = colnames(categorical_columns)[j],
        CramerV = cramer_v
      )
    )
  }
}

association_results
```

```
        Var1       Var2    CramerV
1     maritl       race 0.08275370
2     maritl  education 0.06927421
3     maritl     region        NaN
4     maritl   jobclass 0.04397398
5     maritl     health 0.06540070
6     maritl health_ins 0.10035994
7       race  education 0.12384347
8       race     region        NaN
```

```
9       race   jobclass 0.09887955
10      race     health 0.04361799
11      race health_ins 0.04575151
12 education     region        NaN
13 education   jobclass 0.30694372
14 education     health 0.17541130
15 education health_ins 0.21728212
16    region   jobclass        NaN
17    region     health        NaN
18    region health_ins        NaN
19  jobclass     health 0.06703049
20  jobclass health_ins 0.14918956
21    health health_ins 0.07643685
```

Em seguida, nesta tabela de correlação entre as variáveis categóricos independentes, podemos visualizar também a fraca correlação dos variáveis por meio dos valores de correlação calculado.

Por final, com base do análise do gráfico (variáveis numéricos) e da tabela (variáveis categóricos), podemos concluir que a correlação existente entre as variáveis é mínima. Extraindo sinais sobre as variáveis tal como...

1. A variável dependente é "wage".
2. Não apresenta multicolinearidade para variável "year".
3. Não apresenta multincolinearidade para variável "age".

Além disso, nas correlações entre as variáveis categóricos independente é mínima, logo, podemos inferir uma baixa de existência da multicolinearidade através do Fator de Inflação da Variância (VIF) abaixo.

Análise da Multicolinearidade (VIF)

```
sapply(Wage[, sapply(Wage, is.factor)], levels)
```

```
$maritl
[1] "1. Never Married" "2. Married"       "3. Widowed"       "4. Divorced"
[5] "5. Separated"

$race
[1] "1. White" "2. Black" "3. Asian" "4. Other"
```

```
$education
[1] "1. < HS Grad"       "2. HS Grad"         "3. Some College"
[4] "4. College Grad"    "5. Advanced Degree"

$region
[1] "1. New England"        "2. Middle Atlantic"    "3. East North Central"
[4] "4. West North Central" "5. South Atlantic"     "6. East South Central"
[7] "7. West South Central" "8. Mountain"           "9. Pacific"

$jobclass
[1] "1. Industrial"  "2. Information"

$health
[1] "1. <=Good"      "2. >=Very Good"

$health_ins
[1] "1. Yes" "2. No"
```

```
table(Wage$year)
```

```
2003 2004 2005 2006 2007 2008 2009
 513  485  447  392  386  388  389
```

```
table(Wage$age)
```

```
 18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37
 11  14  20  15  38  45  32  56  47  53  59  58  74  63  78  87  76  75  66  77
 38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57
 83  89 113  92  88  98  93  95  80  98  93  83  95  82  69  62  68  65  62  42
 58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77
 57  39  37  33  30  27  11   8  13   7   4   5   6   8   3   5   3   2   3   1
 80
  4
```

```
table(Wage$maritl)
```

```
1. Never Married       2. Married       3. Widowed      4. Divorced
             648             2074               19              204
    5. Separated
              55
```

```
table(Wage$race)
```

```
1. White 2. Black 3. Asian 4. Other
    2480      293      190       37
```

```
table(Wage$education)
```

```
      1. < HS Grad         2. HS Grad    3. Some College    4. College Grad
               268                971                650                685
5. Advanced Degree
               426
```

```
table(Wage$region)
```

```
       1. New England    2. Middle Atlantic 3. East North Central
                    0                  3000                     0
4. West North Central     5. South Atlantic 6. East South Central
                    0                     0                     0
7. West South Central           8. Mountain            9. Pacific
                    0                     0                     0
```

```
table(Wage$jobclass)
```

```
 1. Industrial 2. Information
          1544           1456
```

```
table(Wage$health)
```

```
     1. <=Good 2. >=Very Good
           858           2142
```

```r
table(Wage$health_ins)
```

```
1. Yes  2. No
  2083    917
```

Modelo

```r
library(car)
dados_filtrados <- Wage %>% select(-c(region, logwage))
modelo <- lm(wage ~ ., data = dados_filtrados)
vif(modelo)
```

```
               GVIF Df GVIF^(1/(2*Df))
year       1.006588  1        1.003289
age        1.338316  1        1.156856
maritl     1.320888  4        1.035400
race       1.095774  3        1.015360
education  1.259657  4        1.029275
jobclass   1.136374  1        1.066008
health     1.069775  1        1.034299
health_ins 1.084694  1        1.041486
```

Logo, podemos concluir que todas as variáveis realmente como sinalizados anteriormente não existe uma correlação forte, em que seus respectivos valores de VIF apresentaram abaixo de 10. Portanto, fica evidente que as variáveis independente explicam separadamente a variável resposta/dependente "wage" sem interferência dos outros.

```r
summary(modelo)
```

```
Call:
lm(formula = wage ~ ., data = dados_filtrados)

Residuals:
    Min      1Q  Median      3Q     Max
-100.33  -18.70   -3.26   13.29  212.79
```

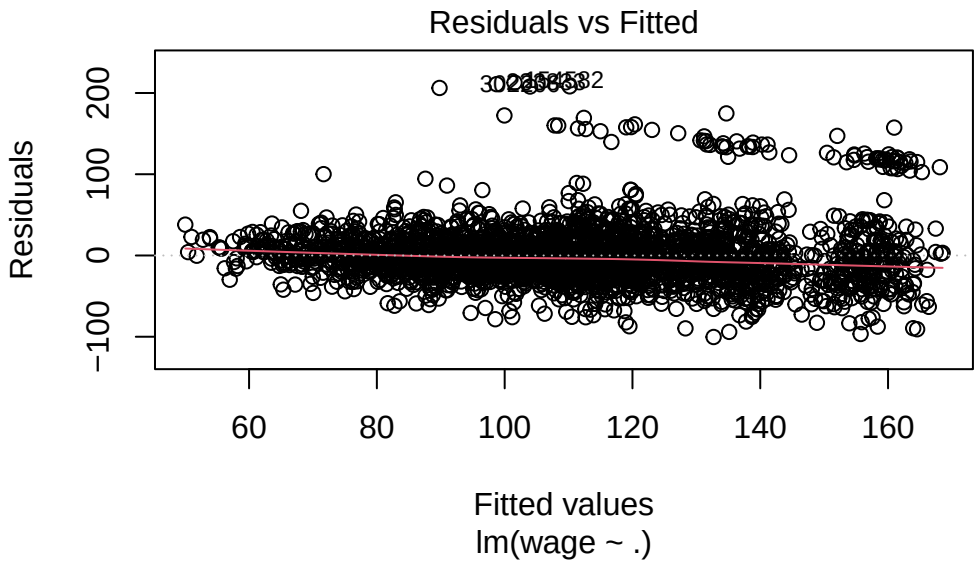
```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -2.423e+03  6.165e+02  -3.931 8.67e-05 ***
year                         1.241e+00  3.074e-01   4.037 5.54e-05 ***
age                          2.707e-01  6.223e-02   4.350 1.41e-05 ***
maritl2. Married             1.718e+01  1.720e+00   9.985  < 2e-16 ***
maritl3. Widowed             2.052e+00  8.005e+00   0.256  0.79774
maritl4. Divorced            3.967e+00  2.887e+00   1.374  0.16951
maritl5. Separated           1.153e+01  4.844e+00   2.380  0.01736 *
race2. Black                -5.096e+00  2.146e+00  -2.375  0.01760 *
race3. Asian                -2.814e+00  2.603e+00  -1.081  0.27978
race4. Other                -6.059e+00  5.666e+00  -1.069  0.28505
education2. HS Grad          7.759e+00  2.369e+00   3.275  0.00107 **
education3. Some College     1.834e+01  2.520e+00   7.278 4.32e-13 ***
education4. College Grad     3.124e+01  2.548e+00  12.259  < 2e-16 ***
education5. Advanced Degree  5.395e+01  2.811e+00  19.190  < 2e-16 ***
jobclass2. Information       3.571e+00  1.324e+00   2.697  0.00704 **
health2. >=Very Good         6.515e+00  1.421e+00   4.585 4.72e-06 ***
health_ins2. No             -1.751e+01  1.403e+00 -12.479  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34 on 2983 degrees of freedom
Multiple R-squared:  0.3396,    Adjusted R-squared:  0.3361
F-statistic: 95.89 on 16 and 2983 DF,  p-value: < 2.2e-16
```

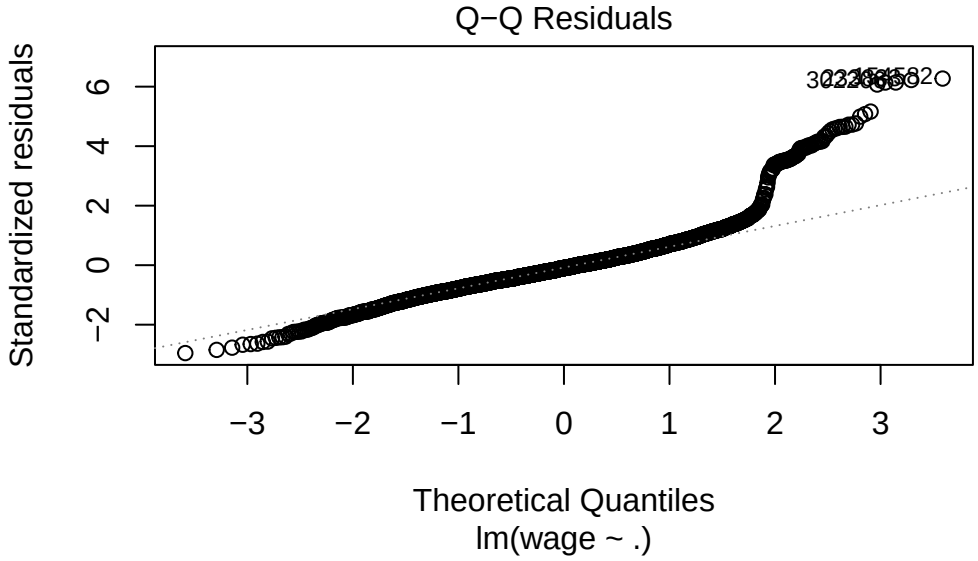
```
step(modelo)
```

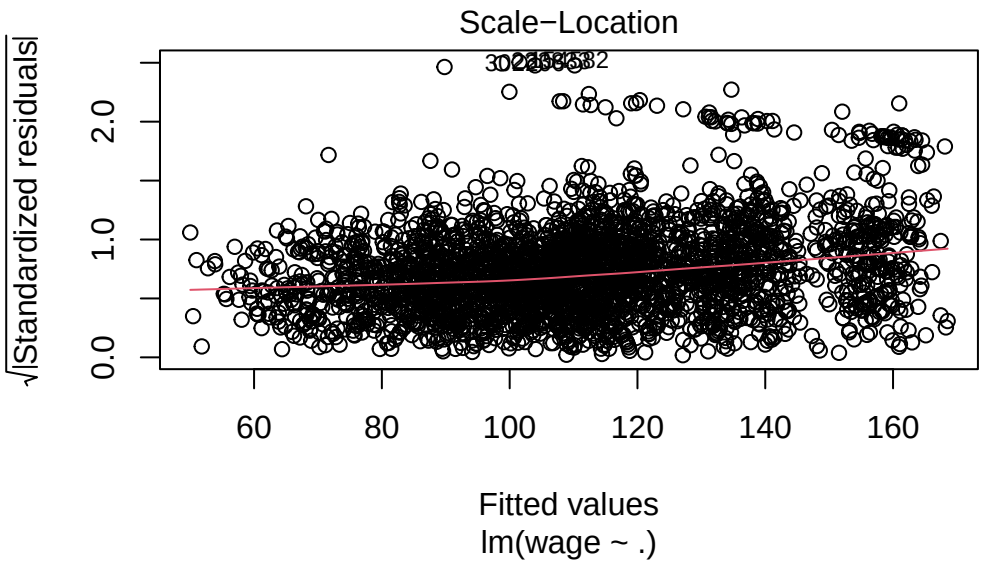
```
Start:  AIC=21175.25
wage ~ year + age + maritl + race + education + jobclass + health +
    health_ins

             Df Sum of Sq     RSS   AIC
<none>                    3448498 21175
- race        3      8520 3457018 21177
- jobclass    1      8407 3456906 21181
- year        1     18844 3467342 21190
- age         1     21875 3470373 21192
- health      1     24307 3472805 21194
- maritl      4    138458 3586956 21285
- health_ins  1    180023 3628521 21326
- education   4    683716 4132214 21710
```

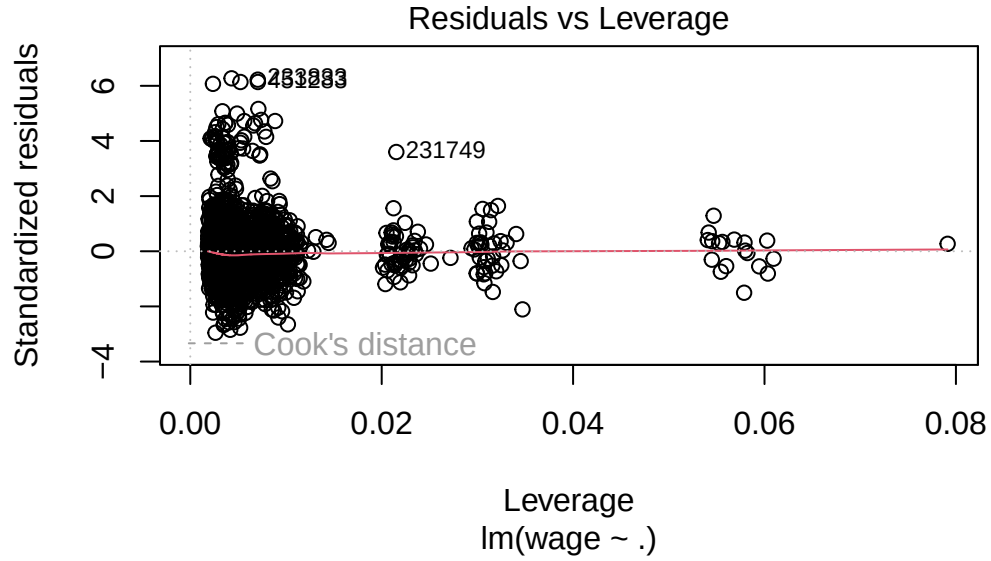
```
Call:
lm(formula = wage ~ year + age + maritl + race + education +
    jobclass + health + health_ins, data = dados_filtrados)

Coefficients:
                (Intercept)                         year
                  -2423.3291                       1.2412
                        age              maritl2. Married
                     0.2707                      17.1767
           maritl3. Widowed             maritl4. Divorced
                     2.0517                       3.9666
         maritl5. Separated                 race2. Black
                    11.5301                      -5.0963
               race3. Asian                 race4. Other
                    -2.8141                      -6.0588
        education2. HS Grad     education3. Some College
                     7.7592                      18.3405
   education4. College Grad  education5. Advanced Degree
                    31.2398                      53.9485
     jobclass2. Information         health2. >=Very Good
                     3.5707                       6.5151
            health_ins2. No
                   -17.5125
```

```
plot(modelo)
```


Residuals vs Fitted
Residuals
−100
0
100
200
60
80
100
120
140
160
Fitted values
lm(wage ~ .)


Q−Q Residuals
Standardized residuals
−2
0
2
4
6
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(wage ~ .)





Seguindo as observações dos gráficos da análise das relações entre variáveis e seus respectivos dispersão e padronização dos resíduos, podemos concluir que nosso modelo de regressão linear precisa de ajuste ainda, devido a falta da uniformidade/linearidade da distribuição do nosso resíduos.

Primeiramente, deveríamos testar cada variável do nosso modelo para inferir seu respectivo influência no modelo.

```r
modelo1 <- update(modelo, ~. -year)

summary(modelo1)
```

```
Call:
lm(formula = wage ~ age + maritl + race + education + jobclass +
    health + health_ins, data = dados_filtrados)

Residuals:
     Min       1Q   Median       3Q      Max
-103.663  -18.706   -3.473   13.853  211.966

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  65.8356     3.5407  18.594  < 2e-16 ***
age                           0.2837     0.0623   4.554 5.47e-06 ***
maritl2. Married             16.9253     1.7234   9.821  < 2e-16 ***
maritl3. Widowed              0.9009     8.0206   0.112 0.910578
maritl4. Divorced             3.6329     2.8929   1.256 0.209287
maritl5. Separated           11.5439     4.8563   2.377 0.017512 *
race2. Black                 -4.8977     2.1505  -2.277 0.022830 *
race3. Asian                 -2.5041     2.6087  -0.960 0.337193
race4. Other                 -5.9525     5.6809  -1.048 0.294809
education2. HS Grad           7.8432     2.3754   3.302 0.000972 ***
education3. Some College     18.3040     2.5265   7.245 5.49e-13 ***
education4. College Grad     31.3257     2.5547  12.262  < 2e-16 ***
education5. Advanced Degree  54.1677     2.8180  19.222  < 2e-16 ***
jobclass2. Information        3.4806     1.3273   2.622 0.008775 **
health2. >=Very Good          6.5454     1.4244   4.595 4.51e-06 ***
health_ins2. No             -17.4482     1.4069 -12.402  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.09 on 2984 degrees of freedom
Multiple R-squared:  0.336, Adjusted R-squared:  0.3327
F-statistic: 100.7 on 15 and 2984 DF,  p-value: < 2.2e-16
```

```r
modelo1 <- update(modelo, ~. -age)

summary(modelo1)
```

```
Call:
lm(formula = wage ~ year + maritl + race + education + jobclass +
    health + health_ins, data = dados_filtrados)

Residuals:
     Min       1Q   Median       3Q      Max
-101.028  -18.900   -3.358   13.585  214.700

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -2552.9992   617.6690  -4.133 3.67e-05 ***
year                           1.3107     0.3079   4.256 2.14e-05 ***
maritl2. Married              20.4251     1.5543  13.141  < 2e-16 ***
maritl3. Widowed               6.8259     7.9534   0.858 0.390829
maritl4. Divorced              7.8410     2.7541   2.847 0.004443 **
maritl5. Separated            14.4241     4.8124   2.997 0.002746 **
race2. Black                  -4.4165     2.1463  -2.058 0.039706 *
race3. Asian                  -3.0956     2.6102  -1.186 0.235733
race4. Other                  -6.9354     5.6799  -1.221 0.222161
education2. HS Grad            7.7235     2.3765   3.250 0.001167 **
education3. Some College      17.9939     2.5264   7.122 1.32e-12 ***
education4. College Grad      31.3337     2.5558  12.260  < 2e-16 ***
education5. Advanced Degree   54.4011     2.8178  19.306  < 2e-16 ***
jobclass2. Information         3.8891     1.3260   2.933 0.003383 **
health2. >=Very Good           5.4586     1.4041   3.888 0.000103 ***
health_ins2. No              -18.1708     1.3994 -12.985  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.1 on 2984 degrees of freedom
Multiple R-squared:  0.3354,    Adjusted R-squared:  0.3321
F-statistic: 100.4 on 15 and 2984 DF,  p-value: < 2.2e-16
```

```r
modelo1 <- update(modelo, ~. -maritl)

summary(modelo1)
```

```
Call:
lm(formula = wage ~ year + age + race + education + jobclass +
    health + health_ins, data = dados_filtrados)

Residuals:
    Min      1Q  Median      3Q     Max
-98.554 -19.127  -3.888  14.112 217.335

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -2.256e+03  6.276e+02  -3.595 0.000329 ***
year                        1.159e+00  3.130e-01   3.703 0.000217 ***
age                         5.078e-01  5.664e-02   8.965  < 2e-16 ***
race2. Black               -7.628e+00  2.169e+00  -3.516 0.000445 ***
race3. Asian               -1.902e+00  2.648e+00  -0.718 0.472633
race4. Other               -7.166e+00  5.770e+00  -1.242 0.214348
education2. HS Grad         7.598e+00  2.410e+00   3.153 0.001632 **
education3. Some College    1.805e+01  2.562e+00   7.045 2.30e-12 ***
education4. College Grad    3.110e+01  2.590e+00  12.008  < 2e-16 ***
education5. Advanced Degree 5.456e+01  2.854e+00  19.120  < 2e-16 ***
jobclass2. Information      3.491e+00  1.348e+00   2.589 0.009671 **
health2. >=Very Good        7.740e+00  1.443e+00   5.362 8.85e-08 ***
health_ins2. No            -1.780e+01  1.430e+00 -12.454  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.65 on 2987 degrees of freedom
Multiple R-squared:  0.3131,    Adjusted R-squared:  0.3104
F-statistic: 113.5 on 12 and 2987 DF,  p-value: < 2.2e-16
```

```r
modelo1 <- update(modelo, ~. -race)

summary(modelo1)
```

```
Call:
lm(formula = wage ~ year + age + maritl + education + jobclass +
    health + health_ins, data = dados_filtrados)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-99.816 -18.755  -3.178  13.436 214.025

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -2.373e+03  6.166e+02  -3.849 0.000121 ***
year                         1.216e+00  3.075e-01   3.954 7.86e-05 ***
age                          2.639e-01  6.203e-02   4.254 2.17e-05 ***
maritl2. Married             1.761e+01  1.710e+00  10.299  < 2e-16 ***
maritl3. Widowed             1.430e+00  8.005e+00   0.179 0.858202
maritl4. Divorced            4.444e+00  2.882e+00   1.542 0.123144
maritl5. Separated           1.155e+01  4.845e+00   2.384 0.017199 *
education2. HS Grad          8.016e+00  2.365e+00   3.389 0.000710 ***
education3. Some College     1.847e+01  2.516e+00   7.341 2.72e-13 ***
education4. College Grad     3.168e+01  2.535e+00  12.498  < 2e-16 ***
education5. Advanced Degree  5.429e+01  2.792e+00  19.446  < 2e-16 ***
jobclass2. Information       3.204e+00  1.316e+00   2.435 0.014970 *
health2. >=Very Good         6.599e+00  1.421e+00   4.643 3.57e-06 ***
health_ins2. No             -1.764e+01  1.403e+00 -12.570  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.03 on 2986 degrees of freedom
Multiple R-squared:  0.338, Adjusted R-squared:  0.3351
F-statistic: 117.3 on 13 and 2986 DF,  p-value: < 2.2e-16
```

```r
modelo1 <- update(modelo, ~. -education)

summary(modelo1)
```

```
Call:
lm(formula = wage ~ year + age + maritl + race + jobclass + health +
    health_ins, data = dados_filtrados)

Residuals:
    Min      1Q  Median      3Q     Max
-97.946 -22.190  -5.033  14.650 213.099

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -2.698e+03  6.741e+02  -4.002 6.44e-05 ***
year                    1.383e+00  3.362e-01   4.115 3.97e-05 ***
```

```
age                     3.535e-01  6.783e-02   5.212 2.00e-07 ***
maritl2. Married        1.789e+01  1.881e+00   9.512  < 2e-16 ***
maritl3. Widowed        1.241e+00  8.755e+00   0.142   0.8873
maritl4. Divorced       2.900e+00  3.157e+00   0.919   0.3584
maritl5. Separated      3.953e+00  5.279e+00   0.749   0.4541
race2. Black           -9.666e+00  2.331e+00  -4.146 3.47e-05 ***
race3. Asian            6.121e+00  2.812e+00   2.177   0.0296 *
race4. Other           -1.271e+01  6.178e+00  -2.058   0.0397 *
jobclass2. Information  1.263e+01  1.389e+00   9.097  < 2e-16 ***
health2. >=Very Good    1.172e+01  1.536e+00   7.634 3.03e-14 ***
health_ins2. No        -2.226e+01  1.512e+00 -14.717  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.19 on 2987 degrees of freedom
Multiple R-squared:  0.2087,    Adjusted R-squared:  0.2055
F-statistic: 65.65 on 12 and 2987 DF,  p-value: < 2.2e-16
```

```r
modelo1 <- update(modelo, ~. -jobclass)

summary(modelo1)
```

```
Call:
lm(formula = wage ~ year + age + maritl + race + education +
    health + health_ins, data = dados_filtrados)

Residuals:
    Min      1Q  Median      3Q     Max
-98.979 -18.817  -3.335  13.331 214.619

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                -2394.6168   617.0991  -3.880 0.000107 ***
year                           1.2272     0.3077   3.988 6.81e-05 ***
age                            0.2800     0.0622   4.501 7.01e-06 ***
maritl2. Married              17.1438     1.7219   9.956  < 2e-16 ***
maritl3. Widowed               1.5070     8.0111   0.188 0.850800
maritl4. Divorced              3.9752     2.8897   1.376 0.169031
maritl5. Separated            11.8594     4.8475   2.447 0.014481 *
race2. Black                  -4.4246     2.1333  -2.074 0.038164 *
race3. Asian                  -2.7997     2.6059  -1.074 0.282749
```

```
race4. Other                    -5.8708     5.6720  -1.035 0.300729
education2. HS Grad              7.9116     2.3712   3.336 0.000859 ***
education3. Some College        18.9039     2.5140   7.519 7.24e-14 ***
education4. College Grad        32.2754     2.5218  12.799  < 2e-16 ***
education5. Advanced Degree     55.5215     2.7530  20.167  < 2e-16 ***
health2. >=Very Good             6.6025     1.4220   4.643 3.58e-06 ***
health_ins2. No                -17.8413     1.3995 -12.748  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.04 on 2984 degrees of freedom
Multiple R-squared:  0.338, Adjusted R-squared:  0.3347
F-statistic: 101.6 on 15 and 2984 DF,  p-value: < 2.2e-16
```

```r
modelo1 <- update(modelo, ~. -health)

summary(modelo1)
```

```
Call:
lm(formula = wage ~ year + age + maritl + race + education +
    jobclass + health_ins, data = dados_filtrados)

Residuals:
     Min       1Q   Median       3Q      Max
-101.665  -19.054   -3.385   13.475  214.698

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -2.433e+03  6.186e+02  -3.932 8.61e-05 ***
year                        1.249e+00  3.085e-01   4.048 5.30e-05 ***
age                         2.219e-01  6.152e-02   3.607 0.000314 ***
maritl2. Married            1.772e+01  1.722e+00  10.293  < 2e-16 ***
maritl3. Widowed            2.418e+00  8.032e+00   0.301 0.763409
maritl4. Divorced           3.966e+00  2.896e+00   1.369 0.170986
maritl5. Separated          1.197e+01  4.859e+00   2.464 0.013799 *
race2. Black               -5.169e+00  2.153e+00  -2.401 0.016408 *
race3. Asian               -3.078e+00  2.611e+00  -1.179 0.238541
race4. Other               -6.677e+00  5.684e+00  -1.175 0.240206
education2. HS Grad         7.977e+00  2.377e+00   3.356 0.000801 ***
education3. Some College    1.892e+01  2.525e+00   7.492 8.89e-14 ***
education4. College Grad    3.234e+01  2.545e+00  12.705  < 2e-16 ***
```

```
education5. Advanced Degree  5.539e+01  2.803e+00  19.762  < 2e-16 ***
jobclass2. Information       3.709e+00  1.328e+00   2.793 0.005261 **
health_ins2. No             -1.789e+01  1.406e+00 -12.731  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.11 on 2984 degrees of freedom
Multiple R-squared:  0.335, Adjusted R-squared:  0.3316
F-statistic: 100.2 on 15 and 2984 DF,  p-value: < 2.2e-16
```

Depois de ter analisado a influência de cada um dos variáveis do nosso modelo, podemos concluir que alguns deles tem pouca influência sobre o modelo, ou melhor uma influência negativa diminuindo o "R-squared".

Por conseguinte, deveríamos olhar e redefinir para o nosso caso base, aonde definimos o parâmetro do nosso modelo inicialmente, excluindo alguns variáveis que não explicam profundamente e de forma uníssona sobre a variável resposta tal como race, jobclass, etc.

Além disso, podemos mudar o olhar da nossa variável de resposta "wage" para o "logwage", desde que percebemos uma não-linearidade dos pontos de dados residuais que provavelmente pode ser causado pela dispersão do intervalo da variável de resposta.

Modelo 2

```r
dados_filtrados <- Wage %>% select(-c(region, jobclass, race, health_ins, wage))
modelo2 <- lm(logwage ~ ., data = dados_filtrados)
summary(modelo2)
```

```
Call:
lm(formula = logwage ~ ., data = dados_filtrados)

Residuals:
     Min       1Q   Median       3Q      Max
-1.63723 -0.14962  0.00968  0.16721  1.25542

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -1.940e+01  5.300e+00  -3.661 0.000256 ***
year                        1.171e-02  2.643e-03   4.430 9.74e-06 ***
```

```
age                          3.561e-03  5.287e-04   6.736 1.94e-11 ***
maritl2. Married             1.740e-01  1.469e-02  11.841  < 2e-16 ***
maritl3. Widowed             4.920e-02  6.880e-02   0.715 0.474601
maritl4. Divorced            6.101e-02  2.477e-02   2.463 0.013821 *
maritl5. Separated           1.310e-01  4.164e-02   3.146 0.001674 **
education2. HS Grad          1.172e-01  2.024e-02   5.791 7.74e-09 ***
education3. Some College     2.377e-01  2.134e-02  11.141  < 2e-16 ***
education4. College Grad     3.503e-01  2.128e-02  16.465  < 2e-16 ***
education5. Advanced Degree  5.106e-01  2.317e-02  22.036  < 2e-16 ***
health2. >=Very Good         7.299e-02  1.219e-02   5.988 2.38e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2925 on 2988 degrees of freedom
Multiple R-squared:  0.311, Adjusted R-squared:  0.3084
F-statistic: 122.6 on 11 and 2988 DF,  p-value: < 2.2e-16
```

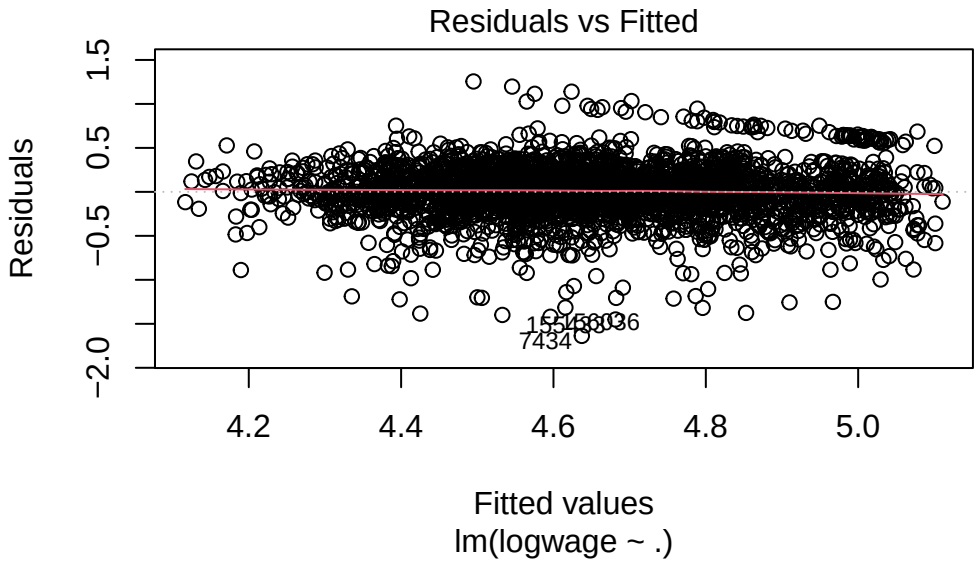
```
step(modelo2, direction='backward')
```

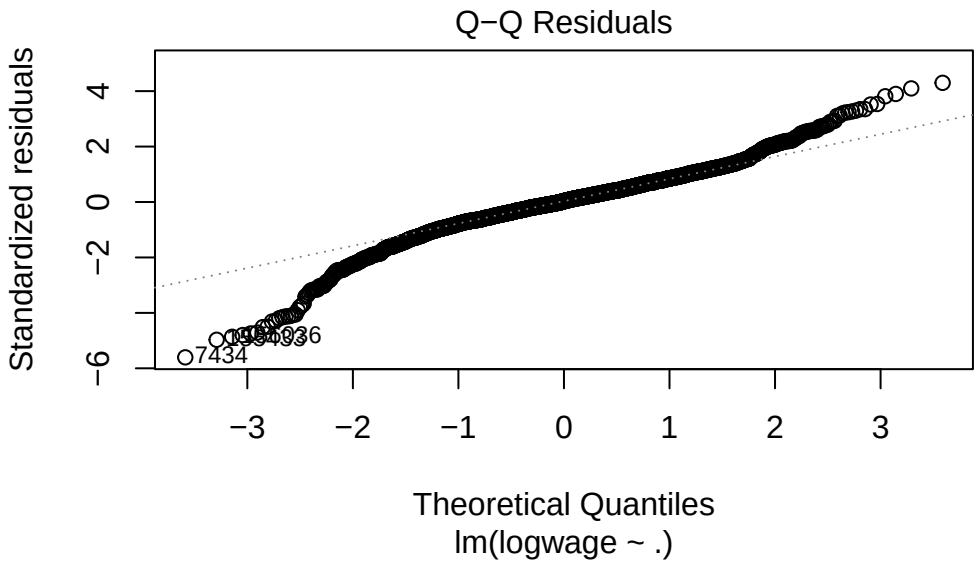
```
Start:  AIC=-7363.31
logwage ~ year + age + maritl + education + health

            Df Sum of Sq    RSS     AIC
<none>                  255.68 -7363.3
- year       1     1.680 257.36 -7345.7
- health     1     3.068 258.75 -7329.5
- age        1     3.883 259.56 -7320.1
- maritl     4    13.546 269.23 -7216.4
- education  4    66.301 321.98 -6679.6


Call:
lm(formula = logwage ~ year + age + maritl + education + health,
    data = dados_filtrados)

Coefficients:
                (Intercept)                         year
                 -19.402749                     0.011708
                        age              maritl2. Married
                   0.003561                     0.173978
           maritl3. Widowed             maritl4. Divorced
                   0.049200                     0.061009
```

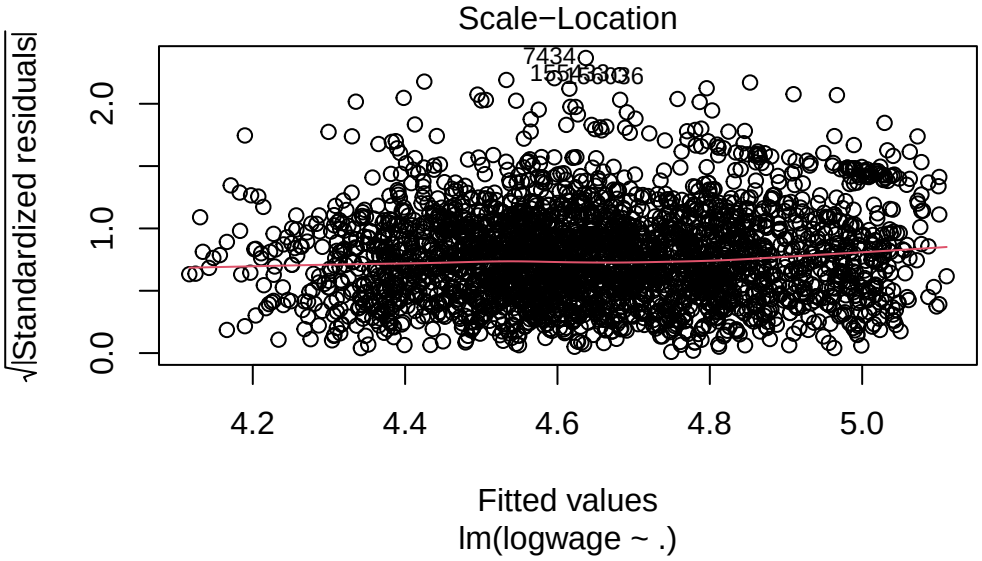
```
         maritl5. Separated          education2. HS Grad
                   0.130984                     0.117178
   education3. Some College     education4. College Grad
                   0.237709                     0.350309
education5. Advanced Degree         health2. >=Very Good
                   0.510647                     0.072995
```

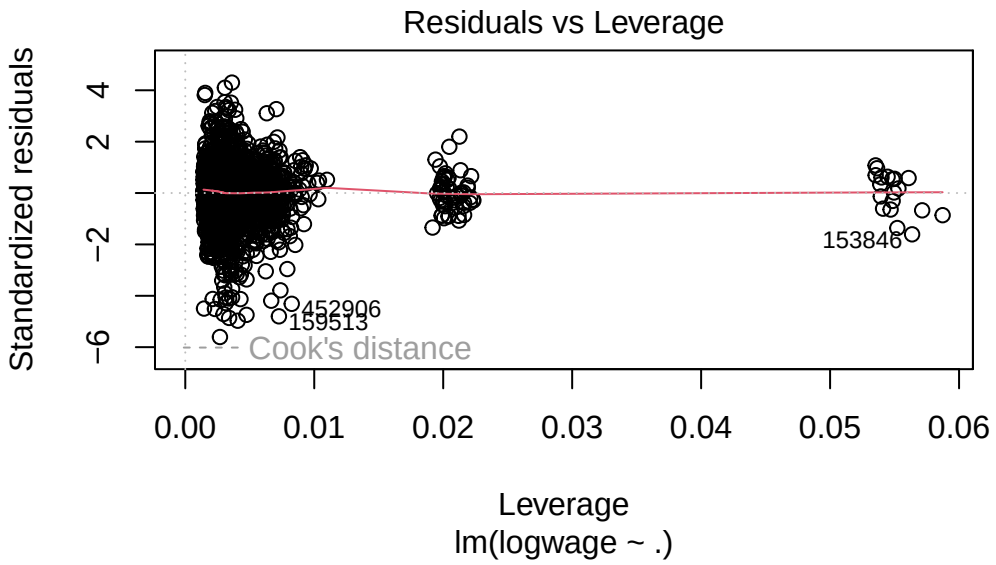
Pressupostos do MRLM

```
plot(modelo2)
```




Q−Q Residuals
Standardized residuals
−6
−2
0
2
4
7434
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(logwage ~ .)


Scale−Location
√|Standardized residuals|
0.0
1.0
2.0
7434
4.2
4.4
4.6
4.8
5.0
Fitted values
lm(logwage ~ .)



Diante do que foi ajustado com os variáveis, descartando variáveis que impacta negativamente o modelo, podemos observar que foi obtido uma uniformidade dos nossos resíduos que anteriormente estava formando uma parábola.

Além disso, é notório que existe alguns outliers no nosso base de dados, logo o sugerido para aprimorar o modelo seria a remoção dos outliers conforme mostrada nos passos abaixo.

```
outliers <- outlierTest(modelo2)

outliers
```

```
        rstudent unadjusted p-value Bonferroni p
7434   -5.633223         1.9332e-08   5.7995e-05
155433 -4.990365         6.3718e-07   1.9115e-03
156036 -4.880845         1.1118e-06   3.3353e-03
159513 -4.817937         1.5229e-06   4.5686e-03
86679  -4.758999         2.0380e-06   6.1140e-03
160130 -4.726695         2.3876e-06   7.1629e-03
160269 -4.528016         6.1870e-06   1.8561e-02
228764 -4.511889         6.6732e-06   2.0020e-02
452906 -4.326002         1.5682e-05   4.7045e-02
```

A partir dos dados acima podemos notar alguns outliers, com o valores identificados na tabela

são aqueles com valores de resíduos padronizados (rstudent) extremos e p-valores ajustados por Bonferroni menores que 0.05. Então no próximo passo é remover eles dos nossos dados.

```
outliers_indices <- c(7434, 155433, 156036, 159513, 86679, 160130, 160269, 228764, 452906, 2

wage_sem_outliers <- Wage %>% slice(-outliers_indices)

glimpse(wage_sem_outliers)
```

```
Rows: 2,996
Columns: 11
$ year       <int> 2006, 2004, 2003, 2003, 2005, 2008, 2009, 2008, 2006, 2004,~
$ age        <int> 18, 24, 45, 43, 50, 54, 44, 30, 41, 52, 45, 34, 35, 39, 54,~
$ maritl     <fct> 1. Never Married, 1. Never Married, 2. Married, 2. Married,~
$ race       <fct> 1. White, 1. White, 1. White, 3. Asian, 1. White, 1. White,~
$ education  <fct> 1. < HS Grad, 4. College Grad, 3. Some College, 4. College ~
$ region     <fct> 2. Middle Atlantic, 2. Middle Atlantic, 2. Middle Atlantic,~
$ jobclass   <fct> 1. Industrial, 2. Information, 1. Industrial, 2. Informatio~
$ health     <fct> 1. <=Good, 2. >=Very Good, 1. <=Good, 2. >=Very Good, 1. <=~
$ health_ins <fct> 2. No, 2. No, 1. Yes, 1. Yes, 1. Yes, 1. Yes, 1. Yes, 1. Ye~
$ logwage    <dbl> 4.318063, 4.255273, 4.875061, 5.041393, 4.318063, 4.845098,~
$ wage       <dbl> 75.04315, 70.47602, 130.98218, 154.68529, 75.04315, 127.115~
```

```
# checkar se ainda existe outliers ou não
any(outliers_indices %in% rownames(wage_sem_outliers))
```

```
[1] TRUE
```

Modelo 3

```
dados_filtrados <- wage_sem_outliers %>% select(-c(region, jobclass, race, health_ins, wage)
modelo3 <- lm(logwage ~ ., data = dados_filtrados)
summary(modelo3)
```

```
Call:
lm(formula = logwage ~ ., data = dados_filtrados)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.45752 -0.15173  0.00797  0.16506  1.25347

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -1.911e+01  5.241e+00  -3.646 0.000271 ***
year                         1.156e-02  2.613e-03   4.424 1.00e-05 ***
age                          3.540e-03  5.224e-04   6.777 1.48e-11 ***
maritl2. Married             1.757e-01  1.452e-02  12.103  < 2e-16 ***
maritl3. Widowed             5.062e-02  6.797e-02   0.745 0.456455
maritl4. Divorced            7.483e-02  2.455e-02   3.048 0.002327 **
maritl5. Separated           1.324e-01  4.114e-02   3.219 0.001301 **
education2. HS Grad          1.166e-01  1.999e-02   5.835 5.97e-09 ***
education3. Some College     2.425e-01  2.109e-02  11.502  < 2e-16 ***
education4. College Grad     3.509e-01  2.102e-02  16.695  < 2e-16 ***
education5. Advanced Degree  5.144e-01  2.290e-02  22.461  < 2e-16 ***
health2. >=Very Good         6.971e-02  1.205e-02   5.783 8.09e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.289 on 2984 degrees of freedom
Multiple R-squared:  0.3172,    Adjusted R-squared:  0.3147
F-statistic:   126 on 11 and 2984 DF,  p-value: < 2.2e-16
```

Com os pequenos ajustes acima percemos que o 'R-squared' foi aprimorado, então o sugerido seria continuar com a eliminação dos variáveis que comprometem com as seguintes características: Linearidade, Independência dos Erros, Homoscedasticidade, Normalidade dos Erros, Ausência de Multicolinearidade, Independência das Observações;

As observações no conjunto de dados devem ser independentes umas das outras. Isso é especialmente importante em dados de séries temporais ou dados agrupados.

Interpretações do modelo selecionado

```
library(report)
report(modelo2)
```

```
We fitted a linear model (estimated using OLS) to predict logwage with year,
age, maritl, education and health (formula: logwage ~ year + age + maritl +
```

```
education + health). The model explains a statistically significant and
substantial proportion of variance (R2 = 0.31, F(11, 2988) = 122.59, p < .001,
adj. R2 = 0.31). The model's intercept, corresponding to year = 0, age = 0,
maritl = 1. Never Married, education = 1. < HS Grad and health = 1. <=Good, is
at -19.40 (95% CI [-29.79, -9.01], t(2988) = -3.66, p < .001). Within this
model:

  - The effect of year is statistically significant and positive (beta = 0.01,
95% CI [6.53e-03, 0.02], t(2988) = 4.43, p < .001; Std. beta = 0.07, 95% CI
[0.04, 0.10])
  - The effect of age is statistically significant and positive (beta = 3.56e-03,
95% CI [2.52e-03, 4.60e-03], t(2988) = 6.74, p < .001; Std. beta = 0.12, 95% CI
[0.08, 0.15])
  - The effect of maritl [2. Married] is statistically significant and positive
(beta = 0.17, 95% CI [0.15, 0.20], t(2988) = 11.84, p < .001; Std. beta = 0.49,
95% CI [0.41, 0.58])
  - The effect of maritl [3. Widowed] is statistically non-significant and
positive (beta = 0.05, 95% CI [-0.09, 0.18], t(2988) = 0.72, p = 0.475; Std.
beta = 0.14, 95% CI [-0.24, 0.52])
  - The effect of maritl [4. Divorced] is statistically significant and positive
(beta = 0.06, 95% CI [0.01, 0.11], t(2988) = 2.46, p = 0.014; Std. beta = 0.17,
95% CI [0.04, 0.31])
  - The effect of maritl [5. Separated] is statistically significant and positive
(beta = 0.13, 95% CI [0.05, 0.21], t(2988) = 3.15, p = 0.002; Std. beta = 0.37,
95% CI [0.14, 0.60])
  - The effect of education [2. HS Grad] is statistically significant and
positive (beta = 0.12, 95% CI [0.08, 0.16], t(2988) = 5.79, p < .001; Std. beta
= 0.33, 95% CI [0.22, 0.45])
  - The effect of education [3. Some College] is statistically significant and
positive (beta = 0.24, 95% CI [0.20, 0.28], t(2988) = 11.14, p < .001; Std.
beta = 0.68, 95% CI [0.56, 0.79])
  - The effect of education [4. College Grad] is statistically significant and
positive (beta = 0.35, 95% CI [0.31, 0.39], t(2988) = 16.46, p < .001; Std.
beta = 1.00, 95% CI [0.88, 1.11])
  - The effect of education [5. Advanced Degree] is statistically significant and
positive (beta = 0.51, 95% CI [0.47, 0.56], t(2988) = 22.04, p < .001; Std.
beta = 1.45, 95% CI [1.32, 1.58])
  - The effect of health [2. >=Very Good] is statistically significant and
positive (beta = 0.07, 95% CI [0.05, 0.10], t(2988) = 5.99, p < .001; Std. beta
= 0.21, 95% CI [0.14, 0.28])

Standardized parameters were obtained by fitting the model on a standardized
version of the dataset. 95% Confidence Intervals (CIs) and p-values were
```

```
computed using a Wald t-distribution approximation.
```

Previsões

Um breve resumo sobre os dados observados:

```
summary(Wage)
```

```
      year           age                     maritl           race
 Min.   :2003   Min.   :18.00   1. Never Married: 648   1. White:2480
 1st Qu.:2004   1st Qu.:33.75   2. Married      :2074   2. Black: 293
 Median :2006   Median :42.00   3. Widowed      :  19   3. Asian: 190
 Mean   :2006   Mean   :42.41   4. Divorced     : 204   4. Other:  37
 3rd Qu.:2008   3rd Qu.:51.00   5. Separated    :  55
 Max.   :2009   Max.   :80.00

              education                     region               jobclass
 1. < HS Grad      :268   2. Middle Atlantic   :3000   1. Industrial :1544
 2. HS Grad        :971   1. New England       :   0   2. Information:1456
 3. Some College   :650   3. East North Central:   0
 4. College Grad   :685   4. West North Central:   0
 5. Advanced Degree:426   5. South Atlantic    :   0
                          6. East South Central:   0
                          (Other)              :   0
           health      health_ins      logwage           wage
 1. <=Good     : 858   1. Yes:2083   Min.   :3.000   Min.   : 20.09
 2. >=Very Good:2142   2. No : 917   1st Qu.:4.447   1st Qu.: 85.38
                                     Median :4.653   Median :104.92
                                     Mean   :4.654   Mean   :111.70
                                     3rd Qu.:4.857   3rd Qu.:128.68
                                     Max.   :5.763   Max.   :318.34
```

E agora vamos criar um dataframe para calcular a estimação pontual e intervalar para dois valores médios da variável resposta a explicativa do nosso modelo...

```
novos_dados_media <- data.frame(
  year = c(2007, 2008),
  age = c(30, 45),
  maritl = factor(c("1. Never Married", "2. Married"),
```

```
                    levels = c("1. Never Married", "2. Married", "3. Widowed", "4. Divorced",
  education = factor(c("2. HS Grad", "3. Some College"),
                       levels = c("1. < HS Grad", "2. HS Grad", "3. Some College", "4. College
  health = factor(c("1. <=Good", "2. >=Very Good"),
                    levels = c("1. <=Good", "2. >=Very Good"))
)

estimativas <- predict(modelo3, newdata = novos_dados_media, interval = "confidence")

estimativas
```

```
       fit      lwr      upr
1 4.319290 4.287592 4.350988
2 4.755296 4.727991 4.782600
```

Agora, faremos previsões pontuais e intervalares para duas observações específicas...

```
novos_dados_previsao <- data.frame(
  year = c(2009, 2008),
  age = c(50, 35),
  maritl = factor(c("4. Divorced", "1. Never Married"),
                  levels = c("1. Never Married", "2. Married", "3. Widowed", "4. Divorced",
  education = factor(c("4. College Grad", "2. HS Grad"),
                       levels = c("1. < HS Grad", "2. HS Grad", "3. Some College", "4. College
  health = factor(c("2. >=Very Good", "1. <=Good"),
                  levels = c("1. <=Good", "2. >=Very Good"))
)

previsoes <- predict(modelo3, newdata = novos_dados_previsao, interval = "prediction")

previsoes
```

```
       fit      lwr      upr
1 4.792018 4.223358 5.360677
2 4.348553 3.781000 4.916107
```

Conclusão

Conclusão de Estimação

Primeira Estimação: Estimação Pontual (fit): 4.319290 Intervalo de Confiança (lwr, upr): [4.287592, 4.350988] Interpretação: Para uma combinação específica de características (ano = 2007, idade = 30, estado civil = nunca casado, educação = ensino médio completo, saúde = <=Good), o valor médio esperado de logwage é 4.319290. Estamos 95% confiantes de que o valor médio verdadeiro de logwage para essa combinação de características está entre 4.287592 e 4.350988.

Segunda Estimação: Estimação Pontual (fit): 4.755296 Intervalo de Confiança (lwr, upr): [4.727991, 4.782600] Interpretação: Para outra combinação de características (ano = 2008, idade = 45, estado civil = casado, educação = algum curso superior, saúde = >=Very Good), o valor médio esperado de logwage é 4.755296. Estamos 95% confiantes de que o valor médio verdadeiro de logwage para essa combinação de características está entre 4.727991 e 4.782600.

Conclusão de Previsão

Primeira Previsão: Previsão Pontual (fit): 4.792018 Intervalo de Previsão (lwr, upr): [4.223358, 5.360677] Interpretação: Para uma observação específica (ano = 2009, idade = 50, estado civil = divorciado, educação = graduação completa, saúde = >=Very Good), o valor esperado de logwage é 4.792018. Estamos 95% confiantes de que o valor verdadeiro de logwage para essa observação estará entre 4.223358 e 5.360677. O intervalo de previsão é mais amplo do que o intervalo de confiança, refletindo a maior incerteza associada a prever um valor individual em vez de uma média.

Segunda Previsão: Previsão Pontual (fit): 4.348553 Intervalo de Previsão (lwr, upr): [3.781000, 4.916107] Interpretação: Para outra observação específica (ano = 2008, idade = 35, estado civil = nunca casado, educação = ensino médio completo, saúde = <=Good), o valor esperado de logwage é 4.348553. Estamos 95% confiantes de que o valor verdadeiro de logwage para essa observação estará entre 3.781000 e 4.916107. Novamente, o intervalo de previsão é mais amplo, refletindo a incerteza na previsão de um valor individual.

Por conseguiente…

As sstimações Pontuais e Intervalares: Fornecem uma faixa de valores esperados para a média da população com certas características, com um nível de confiança de 95%.

E as previsões Pontuais e Intervalares: Fornecem uma faixa de valores esperados para observações individuais, com um nível de confiança de 95%, mas com maior incerteza devido à variabilidade individual.