Overview of “Wage” Dataset from “ISLR” Package

Julian Hatwell

February 29, 2016

This document provides a brief overview of the Wage dataset in the ISLR R package.

```
##       year           age               sex                    maritl
##  Min.   :2003   Min.   :18.00   1. Male  :3000   1. Never Married: 648
##  1st Qu.:2004   1st Qu.:33.75   2. Female:   0   2. Married      :2074
##  Median :2006   Median :42.00                    3. Widowed      :  19
##  Mean   :2006   Mean   :42.41                    4. Divorced     : 204
##  3rd Qu.:2008   3rd Qu.:51.00                    5. Separated    :  55
##  Max.   :2009   Max.   :80.00
##
##        race                   education                     region
##  1. White:2480   1. < HS Grad      :268   2. Middle Atlantic   :3000
##  2. Black: 293   2. HS Grad        :971   1. New England       :   0
##  3. Asian: 190   3. Some College   :650   3. East North Central:   0
##  4. Other:  37   4. College Grad   :685   4. West North Central:   0
##                  5. Advanced Degree:426   5. South Atlantic    :   0
##                                           6. East South Central:   0
##                                           (Other)              :   0
##           jobclass               health      health_ins      logwage
##  1. Industrial :1544   1. <=Good     : 858   1. Yes:2083   Min.   :3.000
##  2. Information:1456   2. >=Very Good:2142   2. No : 917   1st Qu.:4.447
##                                                            Median :4.653
##                                                            Mean   :4.654
##                                                            3rd Qu.:4.857
##                                                            Max.   :5.763
##
##       wage
##  Min.   : 20.09
##  1st Qu.: 85.38
##  Median :104.92
##  Mean   :111.70
##  3rd Qu.:128.68
##  Max.   :318.34
##
```

From the summary, and the associated help (not shown), the following observations can be made:

The dataframe contains 3000 rows and 12 columns.

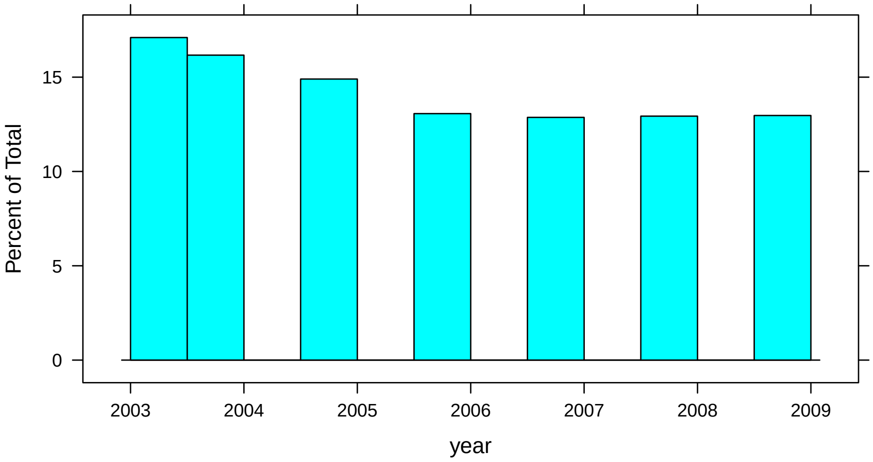


Figure 1: Histogram of the year variable

```
##
##  Welch Two Sample t-test
##
## data:  wage by jobclass
## t = -11.489, df = 2714.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -20.21940 -14.32378
## sample estimates:
##  mean in group 1. Industrial mean in group 2. Information
##                     103.3211                    120.5927
##
##
##  Welch Two Sample t-test
##
## data:  wage by health
## t = -9.2265, df = 1934.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -17.05452 -11.07524
## sample estimates:
##      mean in group 1. <=Good mean in group 2. >=Very Good
##                     101.6613                    115.7262
##
##
```

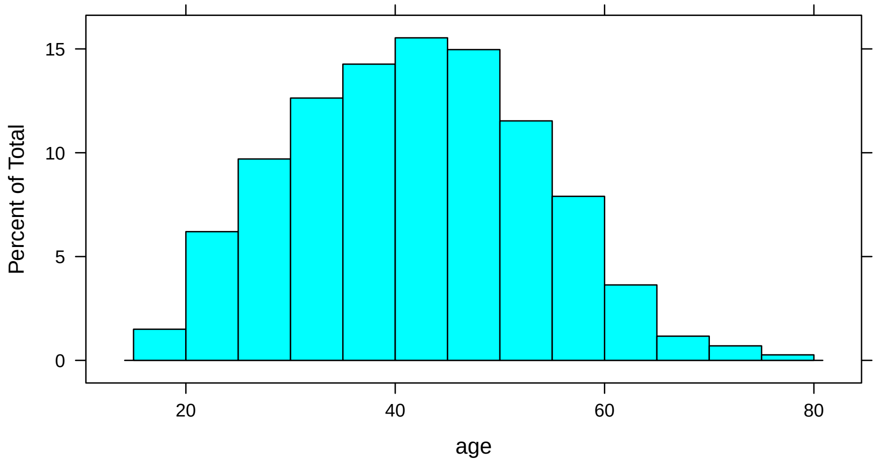


Figure 2: Histogram of the age variable

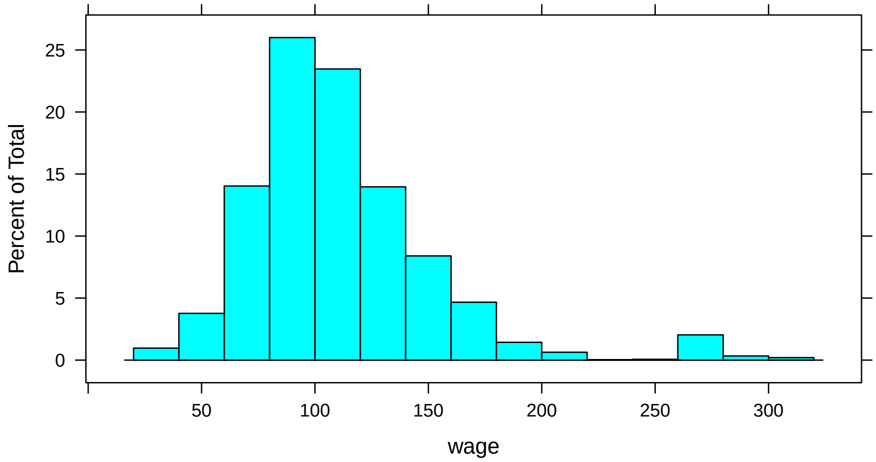


Figure 3: Histogram of the wage variable

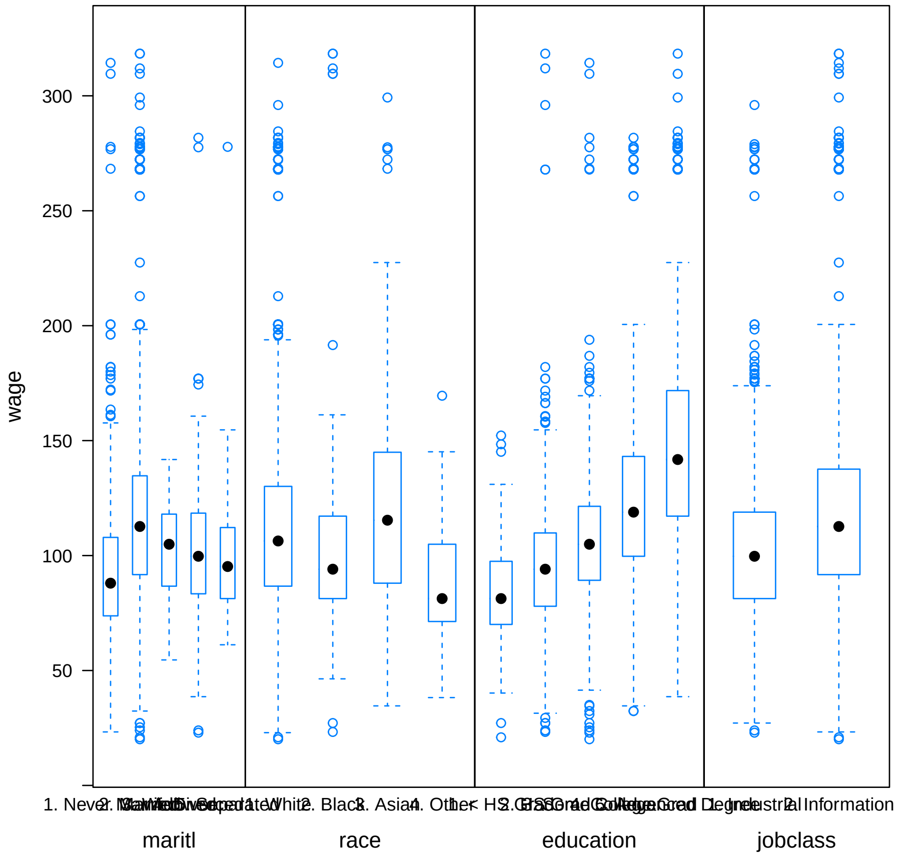


Figure 4: Boxplot of the dependent variable wage by each factor variable

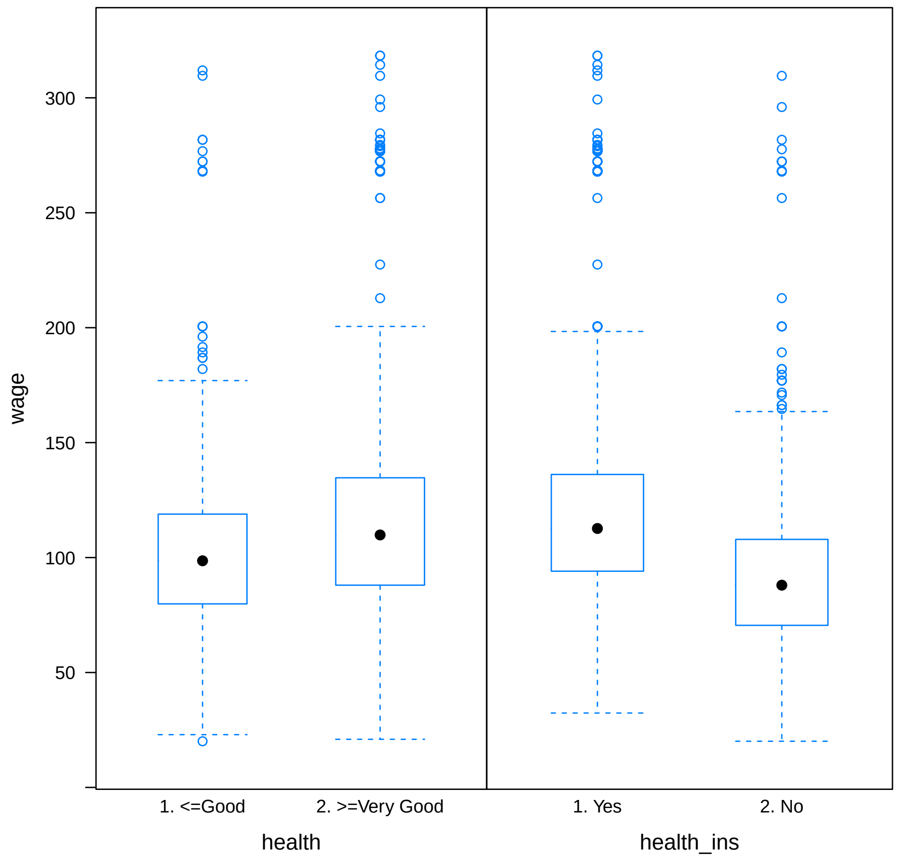


Figure 5: Boxplot of the dependent variable wage by each factor variable

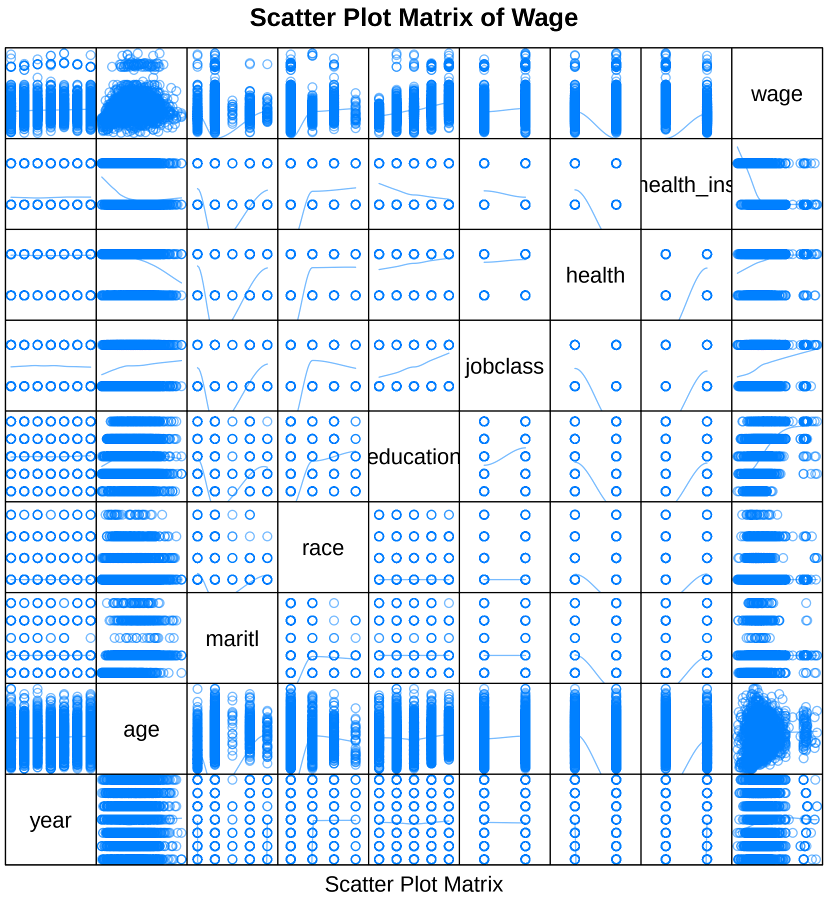


Figure 6: multi-variate comparisons

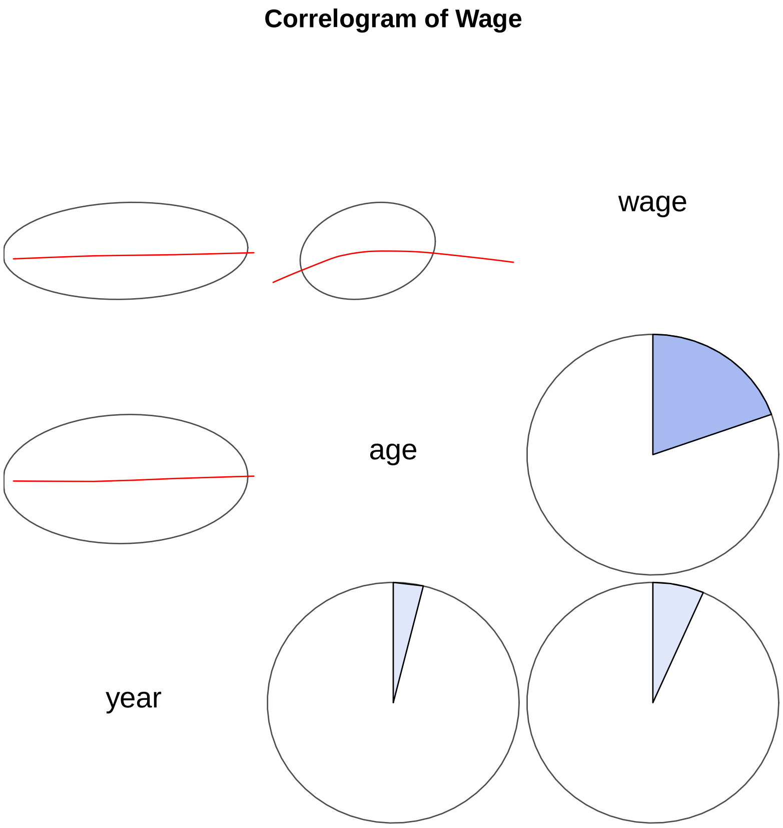


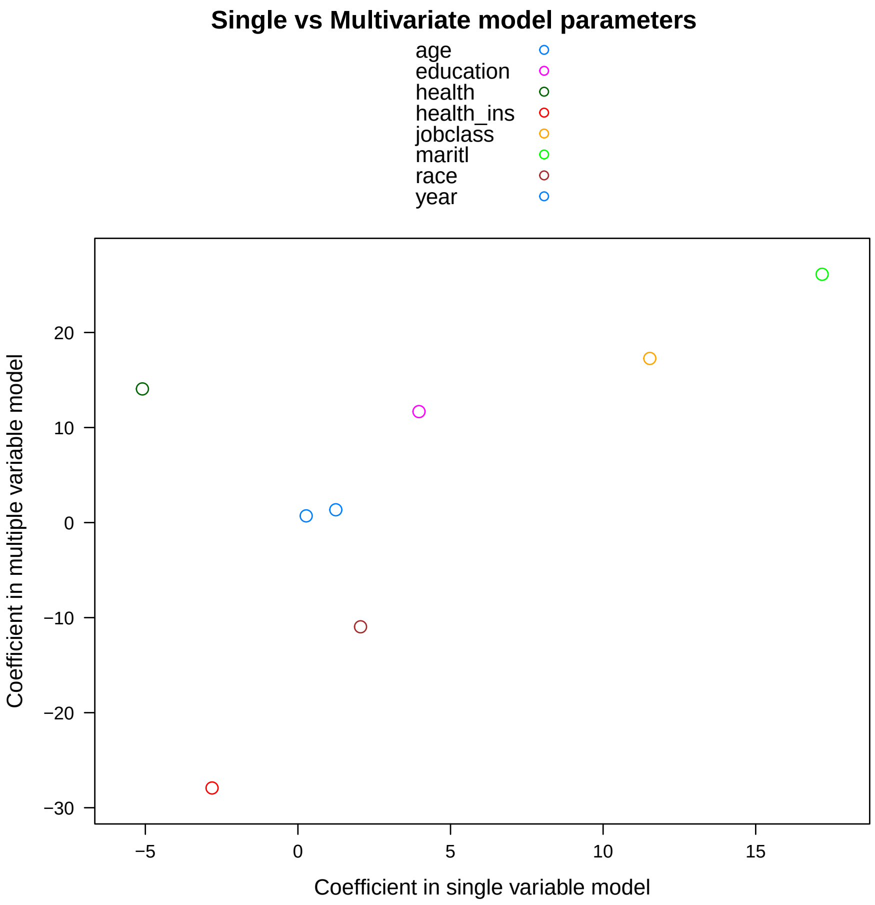
Figure 7: Correlogram

```
##  Welch Two Sample t-test
##
## data:  wage by health_ins
## t = 18.708, df = 1989.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  24.99464 30.84858
## sample estimates:
## mean in group 1. Yes  mean in group 2. No
##             120.2383             92.3167
```

```
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.550 -26.606  -6.415  17.830 206.393
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2595.8616   752.8243  -3.448 0.000572 ***
## year            1.3499     0.3753   3.597 0.000328 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.65 on 2998 degrees of freedom
## Multiple R-squared:  0.004296,Adjusted R-squared:  0.003964
## F-statistic: 12.94 on 1 and 2998 DF,  p-value: 0.0003277
##
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -100.265  -25.115   -6.063   16.601  205.748
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 81.70474    2.84624   28.71   <2e-16 ***
## age          0.70728    0.06475   10.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 40.93 on 2998 degrees of freedom
## Multiple R-squared:  0.03827,Adjusted R-squared:  0.03795
## F-statistic: 119.3 on 1 and 2998 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.775 -24.788  -4.754  15.845 221.595
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          92.735      1.582  58.608  < 2e-16 ***
## maritl2. Married     26.126      1.813  14.413  < 2e-16 ***
## maritl3. Widowed      6.804      9.375   0.726  0.46804
## maritl4. Divorced    10.425      3.234   3.224  0.00128 **
## maritl5. Separated    8.481      5.657   1.499  0.13392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.28 on 2995 degrees of freedom
## Multiple R-squared:  0.06954,Adjusted R-squared:  0.0683
## F-statistic: 55.96 on 4 and 2995 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -92.478 -24.708  -6.251  17.283 216.741
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  112.5637     0.8333 135.088  < 2e-16 ***
## race2. Black -10.9625     2.5634  -4.276 1.96e-05 ***
## race3. Asian   7.7246     3.1236   2.473  0.01345 *
## race4. Other -22.5903     6.8726  -3.287  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.5 on 2996 degrees of freedom
## Multiple R-squared:  0.0121,Adjusted R-squared:  0.01112
## F-statistic: 12.24 on 3 and 2996 DF,  p-value: 5.89e-08
##
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -112.31  -19.94   -3.09   15.33  222.56
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   84.104      2.231  37.695  < 2e-16 ***
## education2. HS Grad           11.679      2.520   4.634 3.74e-06 ***
## education3. Some College      23.651      2.652   8.920  < 2e-16 ***
## education4. College Grad      40.323      2.632  15.322  < 2e-16 ***
## education5. Advanced Degree   66.813      2.848  23.462  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.53 on 2995 degrees of freedom
## Multiple R-squared:  0.2348,Adjusted R-squared:  0.2338
## F-statistic: 229.8 on 4 and 2995 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -100.507  -25.362   -6.117   15.697  197.750
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             103.321      1.039   99.43   <2e-16 ***
## jobclass2. Information   17.272      1.492   11.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.83 on 2998 degrees of freedom
## Multiple R-squared:  0.04281,Adjusted R-squared:  0.04249
## F-statistic: 134.1 on 1 and 2998 DF,  p-value: < 2.2e-16
##
##
## Call:
```

```
## lm(formula = fmla1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.792 -26.618  -5.892  17.223 210.273
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           101.661      1.408   72.19   <2e-16 ***
## health2. >=Very Good   14.065      1.667    8.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.25 on 2998 degrees of freedom
## Multiple R-squared:  0.02321,Adjusted R-squared:  0.02288
## F-statistic: 71.23 on 1 and 2998 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -87.872 -25.355  -5.763  15.919 217.255
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     120.2383     0.8699  138.22   <2e-16 ***
## health_ins2. No -27.9216     1.5734  -17.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.7 on 2998 degrees of freedom
## Multiple R-squared:  0.09505,Adjusted R-squared:  0.09475
## F-statistic: 314.9 on 1 and 2998 DF,  p-value: < 2.2e-16
##
## Call:
## lm(formula = fmla, data = df)
##
## 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
```



```
df <- Wage %>% select(-sex, -region, -logwage)
```