Filippo Riva
MT4420 (Brandt)

Section Quiz 2

Please read these instructions carefully!

- This quiz is open note and open book. You may refer to any of the course materials on our Canvas course page, but you may not communicate with anyone else during this test. You may not use your phone or surf the internet during this test.

- Create a Word document and organize your work in report form as you do on the homework assignments. Please put your name and page numbers in the header of your Word document. Always include your code, any relevant output (plots, summaries, etc.), and your conclusions. Save your report as a pdf and submit it to Canvas.

- For coefficients and adjusted R2 values, round answers to 4 places after the decimal. Please do not state values in scientific notation. For the prediction and confidence interval in Question 1, round answers to the nearest dollar.

Here we will work with the Wage data set. Before you begin, review the book's discussion of the Wage data set on pages 1 and 2. Note that the wage values are given in thousands. Thus, for example, wage = 123.456 means $123,456. In this question, we will find and use some models for wage. From the plots given on page 2, we can see that there is significant spread to the data, so we should not expect our models to have R2 values very close to 1. Nevertheless, we can find significant variables and use them to model wage.

1. **(6 points) Modeling wage as a function of age and year.**

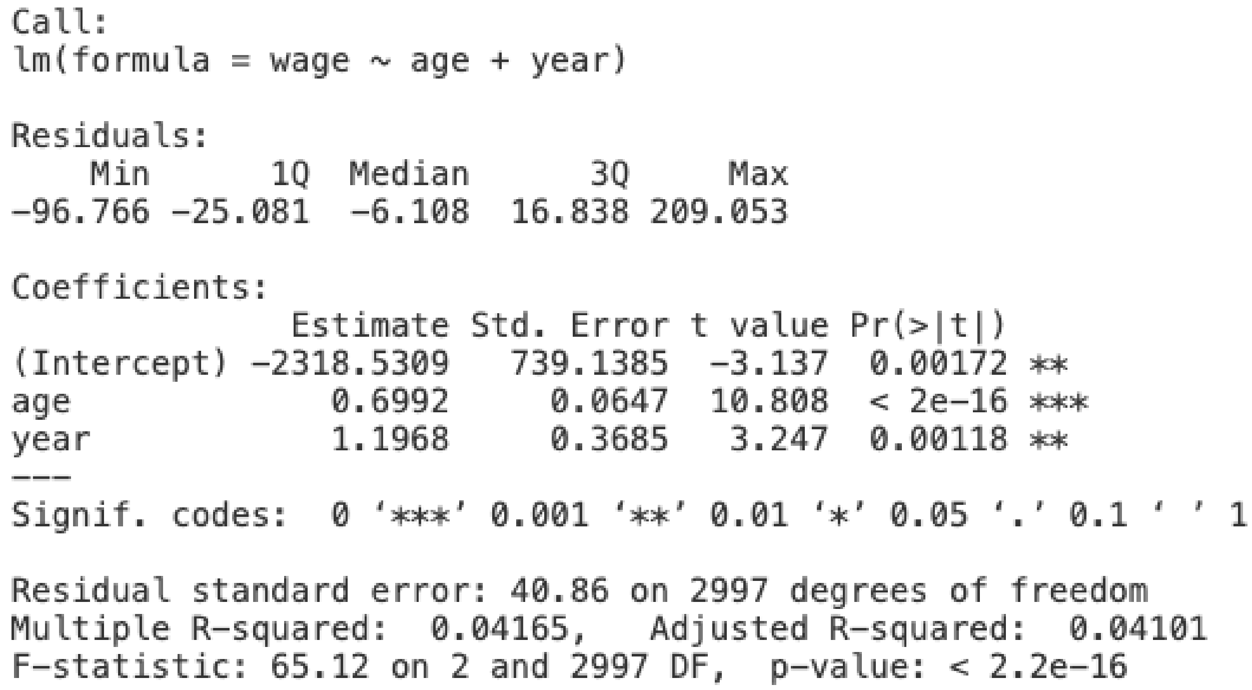
   a. **Find a linear model for wage as a function of age and year. Call this wageMod1. In your report, state your model in the form**

   Wage = β 0 + β 1 (age)+ β 2 (year) ,

   and include your commands and the summary command for your model.

   CODE:

```
Wage = read.csv(file="Wage.csv",head=TRUE,sep=",",stringsAsFactors=FALSE,
na.strings="?")
Wage = na.omit(Wage)
wage=Wage$wage
age=Wage$age
year=Wage$year
wageMod1=lm(wage~age+year)
summary(wageMod1)
```



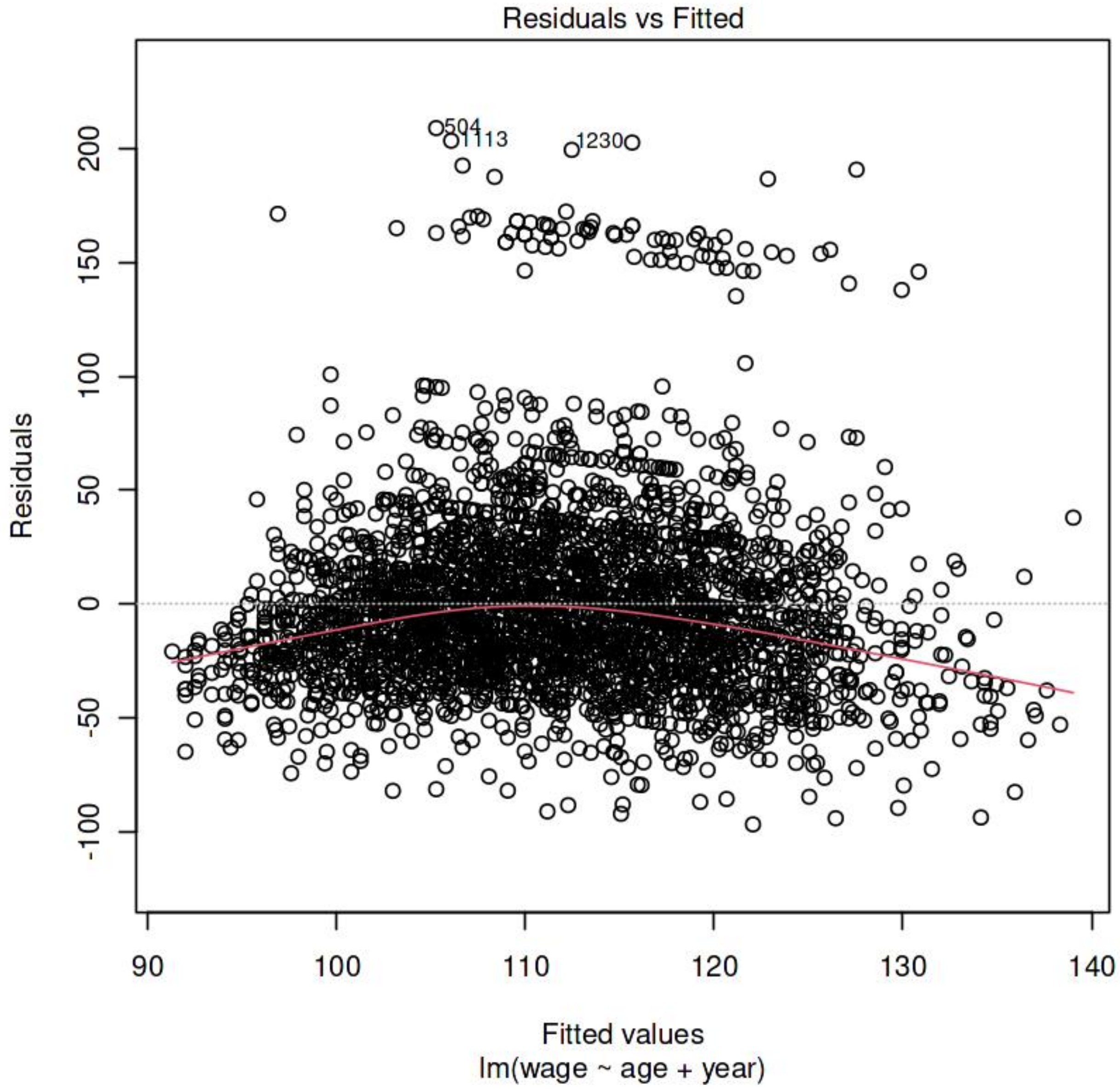
```
Call:
lm(formula = wage ~ age + year)

Residuals:
    Min      1Q  Median      3Q     Max 
-96.766 -25.081  -6.108  16.838 209.053 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2318.5309   739.1385  -3.137  0.00172 ** 
age             0.6992     0.0647  10.808  < 2e-16 ***
year            1.1968     0.3685   3.247  0.00118 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.86 on 2997 degrees of freedom
Multiple R-squared:  0.04165,	Adjusted R-squared:  0.04101 
F-statistic: 65.12 on 2 and 2997 DF,  p-value: < 2.2e-16
```

Wage = -2318.5309 + 0.6992 age + 1.1968 year

b. **In your report, include the residual plot for your wageMod1. Does the plot suggest a non-linear trend in the data? Explain.**
The plot hints at a nonlinear trend, evident in the curvature of the red line. Initially, the line ascends until around 110, then reverses its trajectory, declining until approximately 140. Moreover, a substantial portion of the data falls outside the range of -50 to 50 along the y-axis, indicating a departure from linearity within the model. This deviation from linearity is emphasized by the significant concentration of data points beyond this range, underscoring the model's inability to accurately capture the underlying relationship between the variables.

CODE:

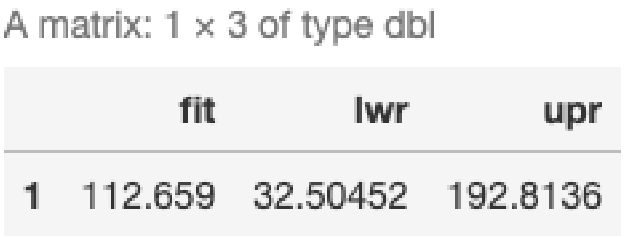
plot(wageMod1,1)



c. Make a prediction. In your report, include your commands and complete the following sentence: The predicted wage for a 40-year-old in 2008 is …

CODE:

```
Wage = read.csv(file="Wage.csv",head=TRUE,sep=",",stringsAsFactors=FALSE,
na.strings="?")
Wage = na.omit(Wage)
wage=Wage$wage
age=Wage$age
year=Wage$year
wageMod1=lm(wage~age+year)
predict(wageMod1,data.frame(age=40, year=2008),level=0.95, interval= "prediction")
```



A matrix: 1 × 3 of type dbl

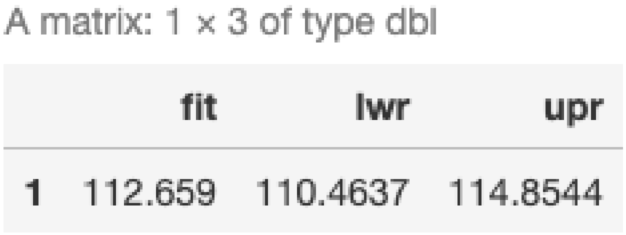
	fit	lwr	upr
1	112.659	32.50452	192.8136

The predicted wage for a 40-year-old in 2008 is 112.659$

d. Find a confidence interval (the default setting for the predict function gives a 95% confidence interval). In your report, include your commands and complete the following sentence: A 95% confidence interval for the predicted wage of a 40-year-old in 2008 is …

CODE:

```
Wage = read.csv(file="Wage.csv",head=TRUE,sep=",",stringsAsFactors=FALSE,
na.strings="?")
Wage = na.omit(Wage)
wage=Wage$wage
age=Wage$age
year=Wage$year
wageMod1=lm(wage~age+year)
predict(wageMod1,data.frame(age=40, year=2008),level=0.95, interval= "confidence")
```



A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	112.659	110.4637	114.8544

A 95 % confidence interval for the predicted wage of a 40-year-old in 2008 is (110.4637, 114.8544)

2. **4 points) Modeling wage as a function of age, age^2, and year. The top-left plot on page 2 suggests that there is a non-linear relationship between age and wage.**

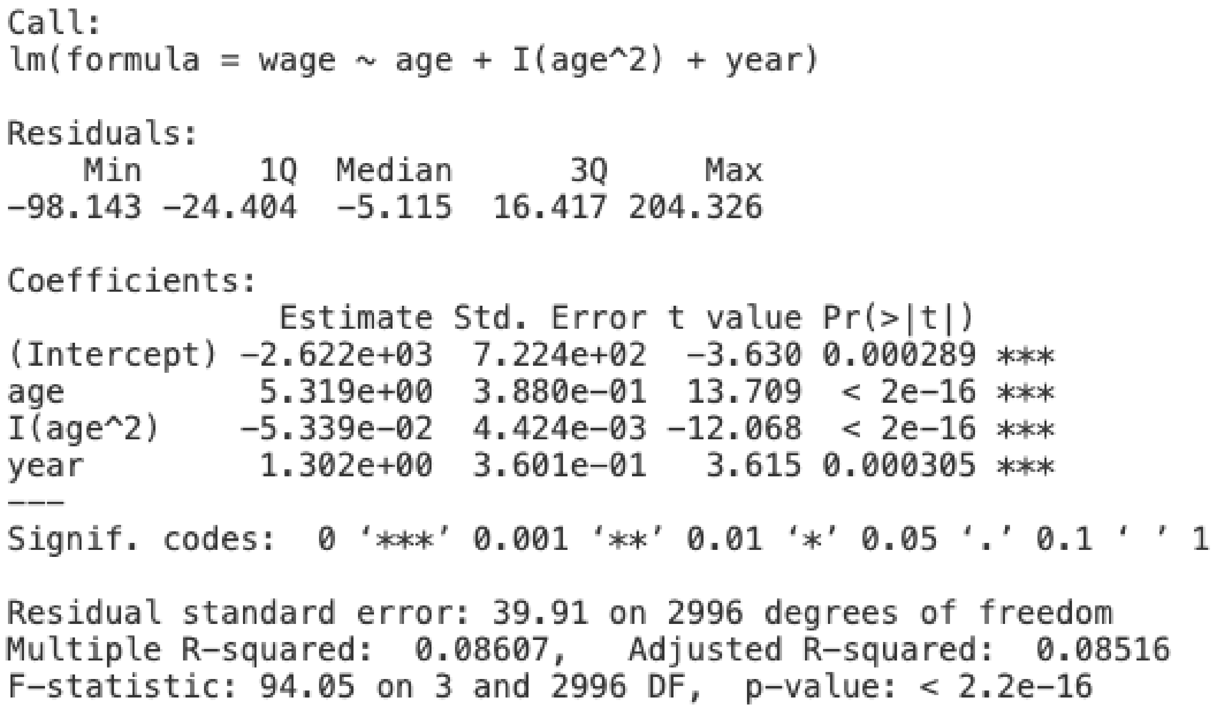
   a. **Find a model for wage as a function of age , age^2, and year. Call this wageMod2. In your report, state your model in the form.**

wage = β 0 + β 1 (age) + β 2 (age^2) + β 3 (year) ,

and include your commands and the summary command for your model. Is age^2 significant? Do the values for RSE and adjusted R2 suggest that wageMod2 is an improvement over wageMod1? Explain.

CODE:

```
Wage = read.csv(file="Wage.csv",head=TRUE,sep=",",stringsAsFactors=FALSE,
na.strings="?")
Wage = na.omit(Wage)
wage=Wage$wage
age=Wage$age
year=Wage$year
wageMod2=lm(wage~age+I(age^2)+year)
summary(wageMod2)
```



```
Call:
lm(formula = wage ~ age + I(age^2) + year)

Residuals:
    Min      1Q  Median      3Q     Max 
-98.143 -24.404  -5.115  16.417 204.326 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.622e+03  7.224e+02  -3.630 0.000289 ***
age          5.319e+00  3.880e-01  13.709  < 2e-16 ***
I(age^2)    -5.339e-02  4.424e-03 -12.068  < 2e-16 ***
year         1.302e+00  3.601e-01   3.615 0.000305 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.91 on 2996 degrees of freedom
Multiple R-squared:  0.08607,	Adjusted R-squared:  0.08516 
F-statistic: 94.05 on 3 and 2996 DF,  p-value: < 2.2e-16
```

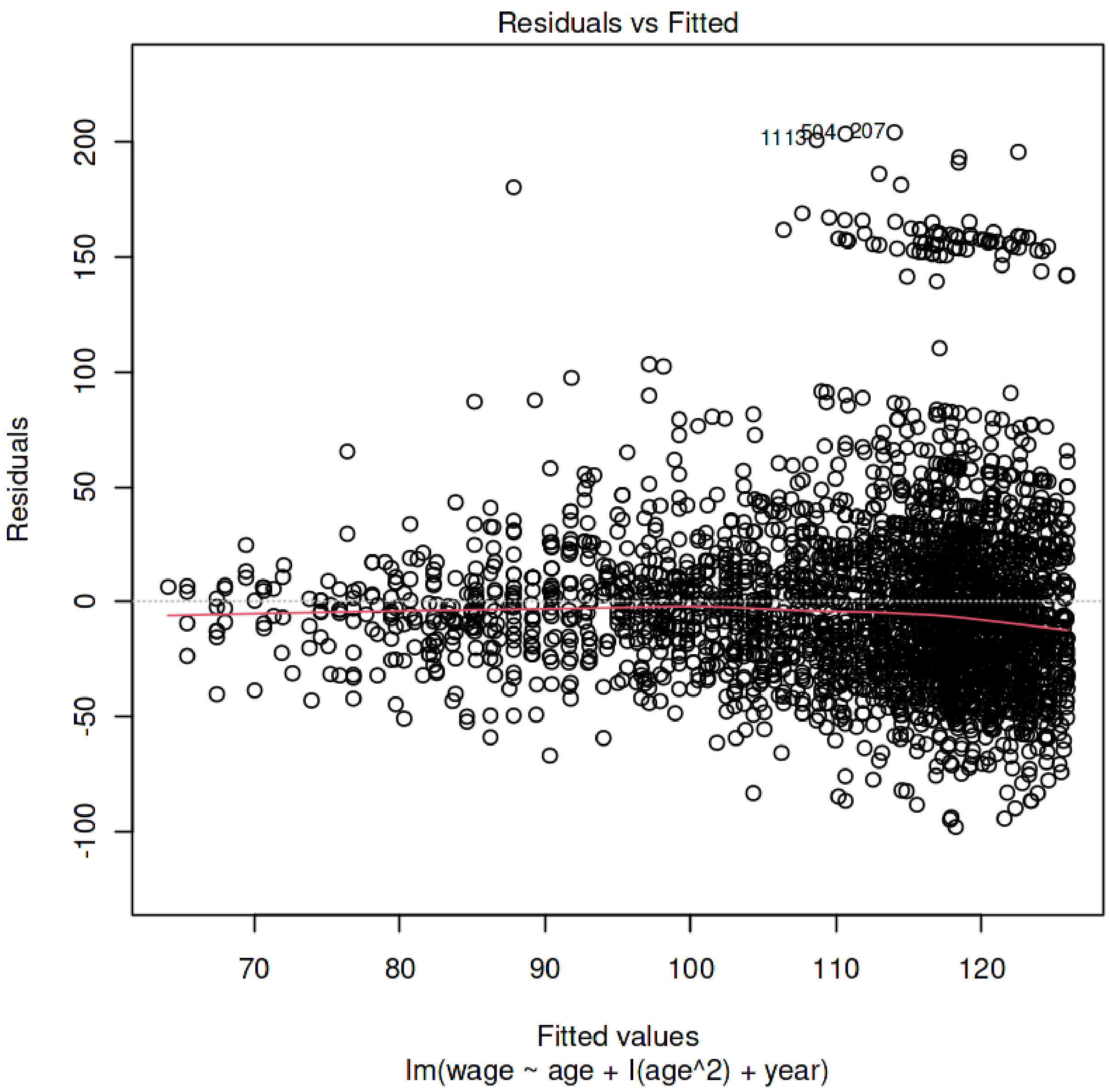
Wage = -2622 + 5.319 age – 533.9 age^2 + 1.302 year

According to the summary command age^2 is significant since it has a really small R value. The values of RSE and R^2 in the new model are showing an improvement from wageMod1. The value of RSE decreased from 40.82 to 39.91 showing a smaller average error and the value of R^2 increased from 0.04101 to 0.08516 showing a more linear model, conformed to the actual model.

b. In your report, include the residual plot for wageMod2. Does the plot suggest an improvement from wageMod1? Explain.

CODE:

```
plot(wageMod2,1)
```

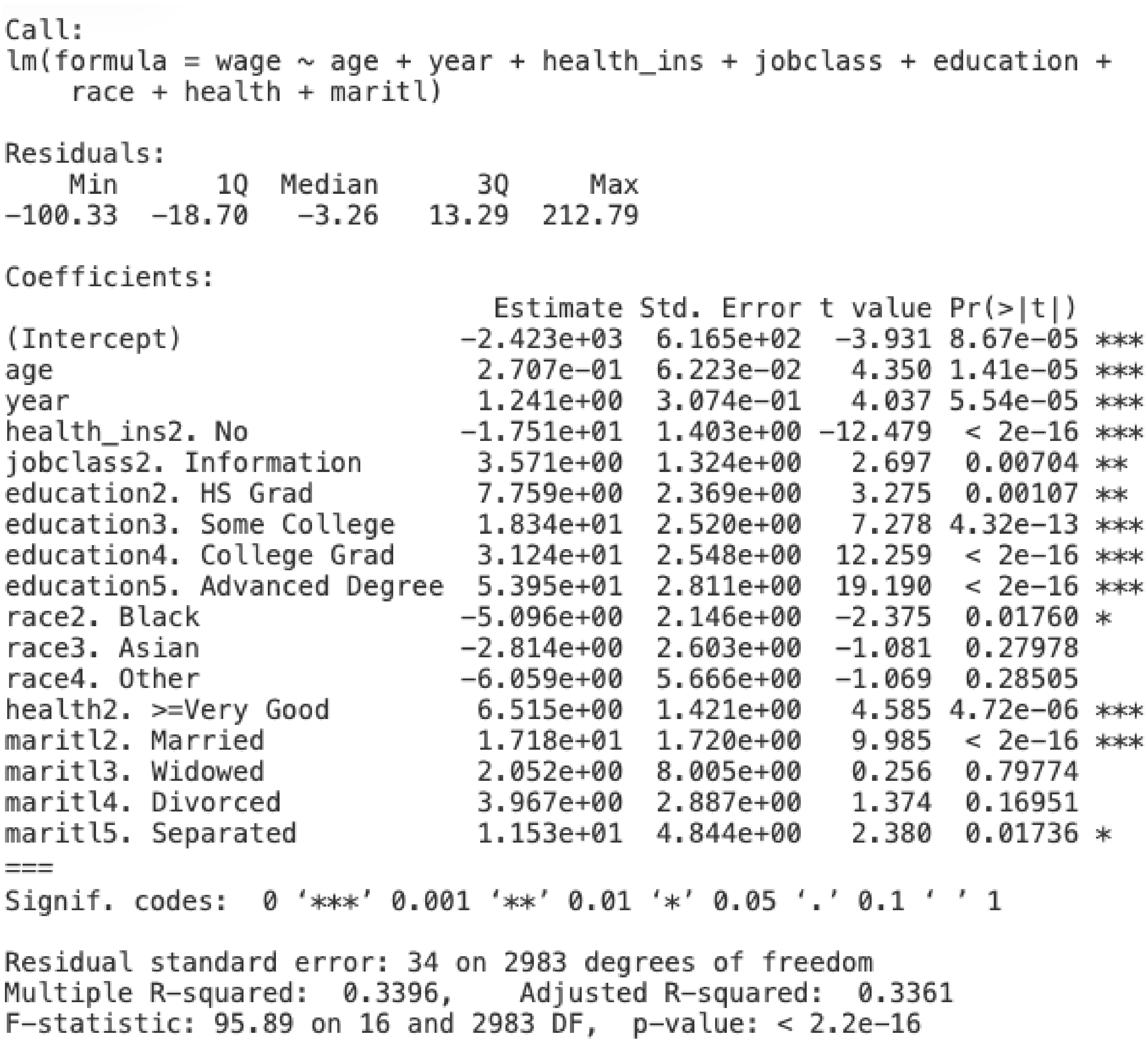


The plot indicates an improvement as the red line displays a more linear trend compared to the red line of the previous model, wageMod1. Additionally, the majority of values are clustered within the range of -50 to 50 on the y-axis. Notably, there is a reduction in the number of data points lying outside this range in comparison to the previous model, suggesting a better alignment with linearity.

3. (3 points) Find a better model. There are several other variables in the Wage data set. Consider the other variables, interaction terms, and transformations of the predictors to find your best model for wage as a function of additional variables. Call this model wageMod3. In your report, explain which variables you used and include your commands, the summary command for wageMod3, and the residual plot for wageMod3. You should be able to improve on the RSE and adjusted R2 from your first 2 models. Include the following table in your report:

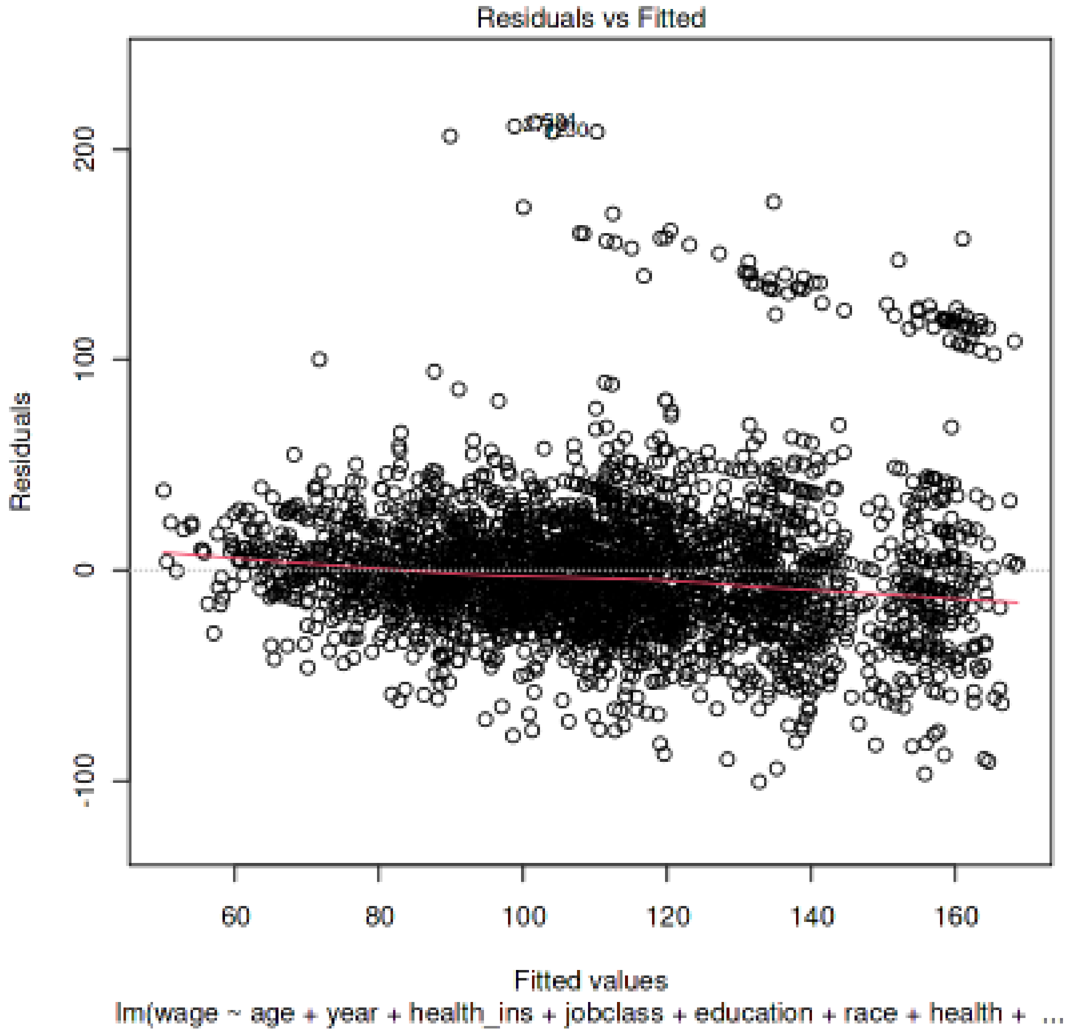
CODE:

```
Wage = read.csv(file="Wage.csv",head=TRUE,sep=",",stringsAsFactors=FALSE,
na.strings="?")   Wage = na.omit(Wage)
wage=Wage$wage
age=Wage$age
year=Wage$year
maritl=Wage$maritl
education=Wage$education
race=Wage$race
jobclass=Wage$jobclass
health=Wage$health
health_ins=Wage$health_ins
wageMod3=lm(wage~age+year+health_ins+jobclass+education+race+health+maritl)
summary(wageMod3)
```



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

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 ***
age                          2.707e-01  6.223e-02   4.350 1.41e-05 ***
year                         1.241e+00  3.074e-01   4.037 5.54e-05 ***
health_ins2. No             -1.751e+01  1.403e+00 -12.479  < 2e-16 ***
jobclass2. Information       3.571e+00  1.324e+00   2.697  0.00704 **
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 ***
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
health2. >=Very Good         6.515e+00  1.421e+00   4.585 4.72e-06 ***
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 *
===
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
```

RESIDUAL PLOT:



During my analysis, I carefully selected several variables for inclusion in my best model: Age, Health, Health_insurance status (Health_ins), Marital status (Maritl), Education level (education), Year, Race, and Job classification (jobclass).Upon testing various iterations of the model, it became evident that whenever any of these variables were excluded, the Residual Standard Error (RSE) increased, while the Adjusted R-squared decreased. This observation indicates that each of these variables contributes significantly to explaining the variation in the response variable.Furthermore, I explored potential interactions between these variables but found none that improved the model's performance.

	RSE	Adjusted R squared
wageMod3	34	0.3361

Notes for Question 3:

- Do not include region and logwage as predictors.

- R will not let you perform transformations on the qualitative predictors. For example, you cannot square education.

- Remember the basic adage: If two models have very close RSE and R2 , choose the simpler of the two.