Final Exam

Nam Jun Lee

12/15/2021

The response variable (y) of interest here is wage. Answer the following questions towards the development of a predictive model for this dataset.

```r
# Wage dataset into wage variable
wage <- ISLR::Wage
# show wage data str
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~
```

a. Implement a multiple linear regression model with wage as the response (y) and the variables age, maritl, race, education, and jobclass as predictors/independent variables and print the summary table of you model.

```r
# fit multiple linear regression model
lm.fit <- lm(wage ~ age + maritl + race + education + jobclass, data = wage)
# summary model
summary(lm.fit)
```

```
##
## Call:
## lm(formula = wage ~ age + maritl + race + education + jobclass,
```

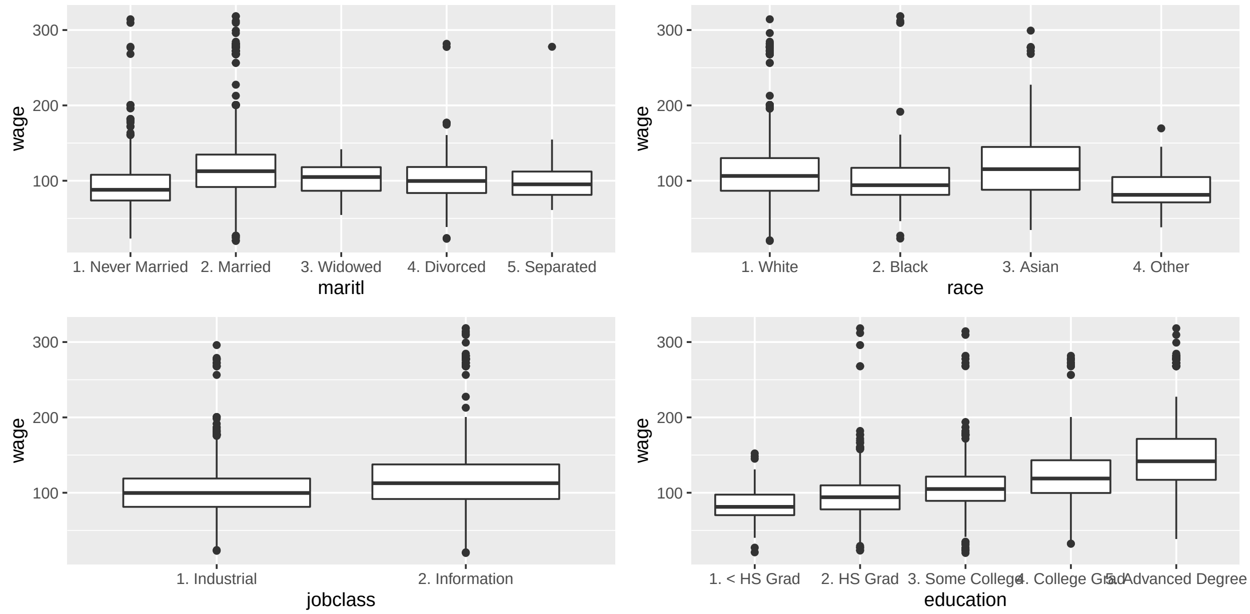
```
##     data = wage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -110.209  -19.671   -3.172   14.686  217.309
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  58.0305     3.2245  17.997  < 2e-16 ***
## age                           0.3119     0.0629   4.958 7.52e-07 ***
## maritl2. Married             18.1724     1.7696  10.269  < 2e-16 ***
## maritl3. Widowed              2.5091     8.2583   0.304 0.761278
## maritl4. Divorced             4.5890     2.9779   1.541 0.123422
## maritl5. Separated           12.2875     4.9997   2.458 0.014042 *
## race2. Black                 -5.4744     2.2141  -2.473 0.013471 *
## race3. Asian                 -4.0846     2.6836  -1.522 0.128101
## race4. Other                 -7.7962     5.8474  -1.333 0.182545
## education2. HS Grad          10.8470     2.4351   4.454 8.72e-06 ***
## education3. Some College     23.1818     2.5750   9.003  < 2e-16 ***
## education4. College Grad     37.3633     2.5882  14.436  < 2e-16 ***
## education5. Advanced Degree  61.0716     2.8494  21.433  < 2e-16 ***
## jobclass2. Information        5.1121     1.3610   3.756 0.000176 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.1 on 2986 degrees of freedom
## Multiple R-squared:  0.2955, Adjusted R-squared:  0.2924
## F-statistic: 96.32 on 13 and 2986 DF,  p-value: < 2.2e-16
```

b. Discuss your interpretation of various aspects the summary table that you obtain in part (a).

In the summary table, the null hypothesis is rejected because the p-value combined with **age**, **Marriage** in Maritl, **education**, and **jobclass** is 0.001. Since it is **Separated** in Maritl and the test p-value of the **Black** in race is 0.05, the null hypothesis can be rejected. In addition, the adjusted r-squared tells us the variance ratio explained by the independent variable. It can be seen that the model has **29.24%** explanatory power and the RSE value is **35.1**.

c. Note that the independent variables maritl, race, education, and jobclass are categorical variables. In view of this observation perform a hypothesis test to determine whether each of these variables are significantly associated with the response variable.

```r
# categorical vs response variabel plots
a <- ggplot(wage, aes(x = maritl, y = wage)) + geom_boxplot()
b <- ggplot(wage, aes(x = race, y = wage)) + geom_boxplot()
c <- ggplot(wage, aes(x = jobclass, y = wage)) + geom_boxplot()
d <- ggplot(wage, aes(x = education, y = wage)) + geom_boxplot()
grid.arrange(a, b, c, d, ncol = 2)
```




```r
# perform hypothesis test each categorical variables
summary(aov(wage ~ maritl, data = wage))
```

```
##               Df  Sum Sq Mean Sq F value Pr(>F)
## maritl         4  363144   90786   55.96 <2e-16 ***
## Residuals   2995 4858941    1622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov(wage ~ race, data = wage))
```

```
##               Df  Sum Sq Mean Sq F value   Pr(>F)
## race           3   63212   21071   12.24 5.89e-08 ***
## Residuals   2996 5158874    1722
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov(wage ~ jobclass, data = wage))
```

```
##               Df  Sum Sq Mean Sq F value Pr(>F)
## jobclass       1  223538  223538   134.1 <2e-16 ***
## Residuals   2998 4998547    1667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

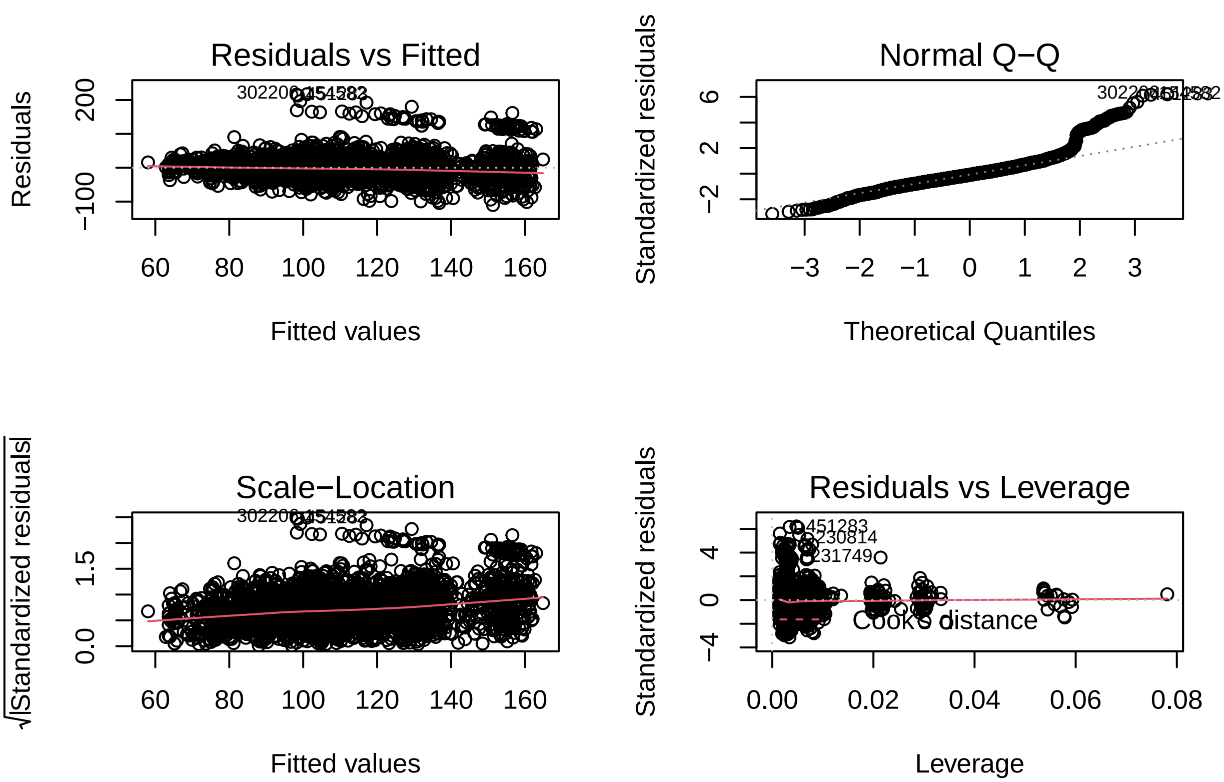
```r
summary(aov(wage ~ education, data = wage))
```

```
##               Df  Sum Sq Mean Sq F value Pr(>F)
## education      4 1226364  306591   229.8 <2e-16 ***
## Residuals   2995 3995721    1334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Independent variables maritl, race, education, and work classes are categorical variables, so first check the distribution through the plot. Married people have higher wages and Asian wages. It can be seen that the higher the educational background, the higher the wage. In addition, it can be seen that job class has higher wages when it is information than industrial. Also, it can be seen that there are many outliers at the top. Considering these observations, a one-way analysis of variance to see if **each categorical variable affects wage shows that the null hypothesis is rejected because the p-value of all categorical variables is much lower than 0.05**.

d. Analyze the residuals of the model that you implemented in part (a). Discuss your observations and propose suitable solutions to the problems that you observe. In particular, make sure to comment on your observations regarding the issues of Heteroskedasticity and Collinearity.

```
# show residuals plot
par(mfrow = c(2, 2))
plot(lm.fit)
```




```
# find collinearity
sqrt(vif(lm.fit)) > 2
```

```
##             GVIF    Df GVIF^(1/(2*Df))
## age        FALSE FALSE           FALSE
## maritl     FALSE FALSE           FALSE
```

```
## race      FALSE FALSE           FALSE
## education FALSE FALSE           FALSE
## jobclass  FALSE FALSE           FALSE
```

There are several outliers in the residual versus fit plot, and the variance seems to be constant, but it is slightly more distributed on the right. The normal Q-Q is a Q-Q diagram for determining whether the residual follows a normal distribution. This graph shows that the residual distribution is skewed to the right. The residual versus leverage plot affects the statistical model coefficient because a specific value is outside the chef distance.
In conclusion, it is heteroskedasticity and it can be seen from the vif function that there is no multicollinearity problem.

e. Now consider the variable logwage as the response (y). Comment on the distinctions/similarities that you observe with respect to the model in Part (a) and your observations of Part (c) and Part (d). Describe which of the two models Part (a) or Part (e) is better suited model and which of the two versions of the response variables wage or logwage would you utilize in practice.

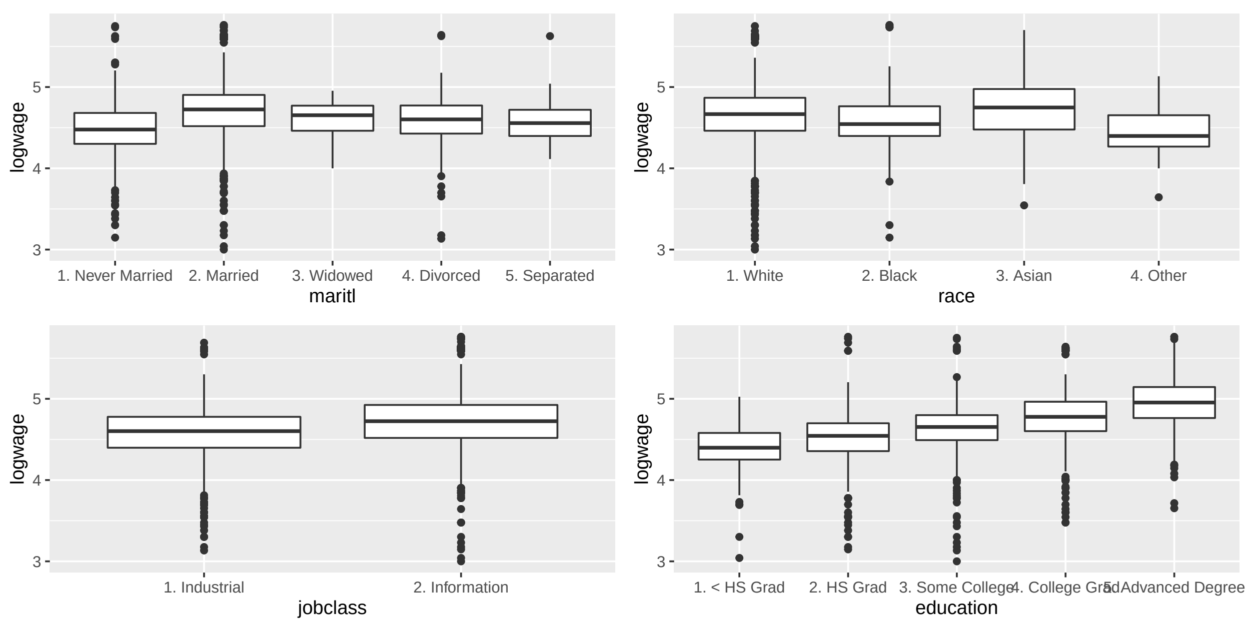
```r
# fit multiple linear regression model (response = logwage)
lm.fit.log <- lm(logwage ~ age + maritl + race + education + jobclass, data = wage)
# summary model
summary(lm.fit.log)
```

```
##
## Call:
## lm(formula = logwage ~ age + maritl + race + education + jobclass,
##     data = wage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72781 -0.15535  0.00945  0.16857  1.21271
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.1429484  0.0270302 153.271  < 2e-16 ***
## age                         0.0030781  0.0005273   5.837 5.87e-09 ***
## maritl2. Married            0.1746901  0.0148341  11.776  < 2e-16 ***
## maritl3. Widowed            0.0543165  0.0692274   0.785 0.432745
## maritl4. Divorced           0.0534642  0.0249634   2.142 0.032298 *
## maritl5. Separated          0.1313648  0.0419118   3.134 0.001739 **
## race2. Black               -0.0444655  0.0185602  -2.396 0.016648 *
## race3. Asian               -0.0349513  0.0224960  -1.554 0.120370
## race4. Other               -0.0775632  0.0490179  -1.582 0.113677
## education2. HS Grad         0.1156124  0.0204130   5.664 1.62e-08 ***
## education3. Some College    0.2352059  0.0215854  10.897  < 2e-16 ***
## education4. College Grad    0.3479349  0.0216966  16.036  < 2e-16 ***
## education5. Advanced Degree 0.5088003  0.0238863  21.301  < 2e-16 ***
## jobclass2. Information      0.0426104  0.0114091   3.735 0.000191 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.2943 on 2986 degrees of freedom
## Multiple R-squared:  0.3032, Adjusted R-squared:  0.3002
## F-statistic: 99.97 on 13 and 2986 DF,  p-value: < 2.2e-16
```

In the summary table, the null hypothesis is rejected because the p-value combined with **age**, **Marriage** in Maritl, **education**, and **jobclass** is 0.001. **Separated** in Maritl p-value is 0.01, so it can be rejected the null hypothesis. Since it is **Divorced** in Maritl and the test p-value of the **Black** in race is 0.05, the null hypothesis can be rejected. In addition, the adjusted r-squared tells us the variance ratio explained by the independent variable. It can be seen that the model has **30.02%** explanatory power and the RSE value is **0.2943**.

```r
# categorical vs response variabel plots
e <- ggplot(wage, aes(x = maritl, y = logwage)) + geom_boxplot()
f <- ggplot(wage, aes(x = race, y = logwage)) + geom_boxplot()
g <- ggplot(wage, aes(x = jobclass, y = logwage)) + geom_boxplot()
h <- ggplot(wage, aes(x = education, y = logwage)) + geom_boxplot()
grid.arrange(e, f, g, h, ncol = 2)
```




```r
# perform hypothesis test each categorical variables
summary(aov(logwage ~ maritl, data = wage))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## maritl         4   31.2   7.789   68.63 <2e-16 ***
## Residuals   2995  339.9   0.113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov(logwage ~ race, data = wage))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## race           3    4.5  1.5129   12.37 4.88e-08 ***
```

```
## Residuals   2996  366.5  0.1223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov(logwage ~ jobclass, data = wage))
```

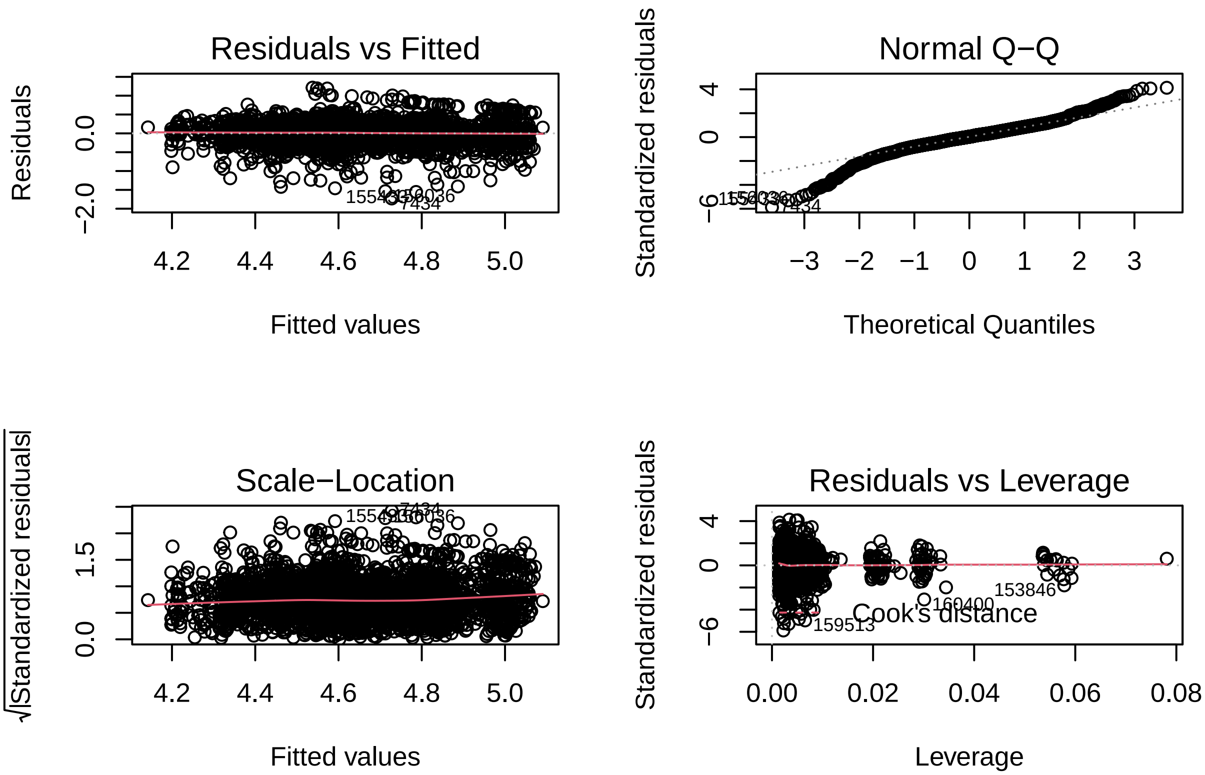
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## jobclass       1   15.7  15.656   132.1 <2e-16 ***
## Residuals   2998  355.4   0.119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov(logwage ~ education, data = wage))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## education      4  83.92  20.979   218.8 <2e-16 ***
## Residuals   2995 287.15   0.096
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As a result of visualizing the categorical dependent variable for logwage, it shows similar results to wage vs categorical variables, but it can be seen that there are many outliers at the bottom.
Considering these observations, a one-way analysis of variance to see if **each categorical variable affects logwage shows that the null hypothesis is rejected because the p-value of all categorical variables is much lower than 0.05**.

```r
# show residuals plot
par(mfrow = c(2, 2))
plot(lm.fit.log)
```




```r
# find collinearity
sqrt(vif(lm.fit.log)) > 2
```

```
##              GVIF    Df GVIF^(1/(2*Df))
## age         FALSE FALSE           FALSE
## maritl      FALSE FALSE           FALSE
## race        FALSE FALSE           FALSE
## education FALSE FALSE           FALSE
## jobclass    FALSE FALSE           FALSE
```

There are several outliers in the residual versus fit plot, and the variance seems to be constant. Normal Q-Q graph shows that the residual distribution is skewed to the left. The residual versus leverage plot affects the statistical model coefficient because a some specific value is outside the chef distance.
In conclusion, it is homogeneity and it can be seen from the vif function that there is no multicollinearity problem.

Compare two models:
Exploratory power (higher is better model)

$$wage(0.2924) < logwage(0.3002)$$

MSE (lower is better model)

$$wage(35.1) < logwage(0.2943)$$

In addition, when comparing the residual graphs of the two models, the residual of the model with logwage as the response variable follows the normal distribution better.
Therefore, **it can be seen that the model in which the response variable is set to logwage is a more suitable model**.

f. The models considered so far have been linear regression models. Use the poly() function to fit polynomial regression (of degree 3) to the above data (recall that age is the only continuous predictor variable), use the response variable that you recommended in Part e.

It is judged that it would be more appropriate to use the logwage response variable, so logwage is set as the response variable.

```r
# fit polynomial regression model (degree 3)
fit.poly = lm(logwage ~ poly(age, 3) + maritl + race + education + jobclass, data = wage)
# summary model
summary(fit.poly)
```

```
##
## Call:
## lm(formula = logwage ~ poly(age, 3) + maritl + race + education +
##     jobclass, data = wage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76595 -0.15696  0.01227  0.16878  1.15431
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.3251235  0.0221223 195.510  < 2e-16 ***
## poly(age, 3)1               2.5094958  0.3319391   7.560 5.33e-14 ***
## poly(age, 3)2              -3.1228780  0.3064853 -10.189  < 2e-16 ***
## poly(age, 3)3               0.8995463  0.2932474   3.068 0.002178 **
## maritl2. Married            0.1208827  0.0154716   7.813 7.66e-15 ***
## maritl3. Widowed            0.0261443  0.0681167   0.384 0.701141
## maritl4. Divorced          -0.0007254  0.0250684  -0.029 0.976916
## maritl5. Separated          0.0731504  0.0415423   1.761 0.078364 .
## race2. Black               -0.0401691  0.0182478  -2.201 0.027791 *
## race3. Asian               -0.0351357  0.0221108  -1.589 0.112150
## race4. Other               -0.0722229  0.0481546  -1.500 0.133769
## education2. HS Grad         0.1112183  0.0200600   5.544 3.21e-08 ***
## education3. Some College    0.2275371  0.0212174  10.724  < 2e-16 ***
## education4. College Grad    0.3340959  0.0213545  15.645  < 2e-16 ***
## education5. Advanced Degree 0.4925230  0.0235155  20.945  < 2e-16 ***
## jobclass2. Information      0.0397095  0.0112134   3.541 0.000404 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2891 on 2984 degrees of freedom
## Multiple R-squared:  0.3281, Adjusted R-squared:  0.3247
## F-statistic: 97.14 on 15 and 2984 DF,  p-value: < 2.2e-16
```

In the summary table, the null hypothesis is rejected because the p-value combined with **age**, **age^2**, **Marriage** in Maritl, **education**, and **jobclass** is 0.001. **age^3** p-value is 0.01, so it can be rejected the null hypothesis. Since p-value of the **Black** in race is 0.05, the null hypothesis can be rejected. In addition, the adjusted r-squared tells us the variance ratio explained by the independent variable. It can be seen that the model has **32.47%** explanatory power and the RSE value is **0.2943**.

g. Construct a Generalized additive model with all predictor variables while utilizing a natural cubic spline to the above data using the ns() function wherever appropriate. You can utilize 3 knots.

```
# Sets the value of an attribute on the specified element knots
attr(ns(wage$age, 4), "knots")
```

```
##   25%   50%   75%
## 33.75 42.00 51.00
```

```
# fit generalized addtive model with all predictor variables
fit.gam1 = gam(logwage ~ ns(year, 4) + ns(age, 4) + maritl + race + education + jobclass +
    health + health_ins, data = wage)
# fit generalized addtive model with set predictor variables
# (age,maritl,race,education,jobclass)
fit.gam2 = gam(logwage ~ ns(age, 4) + maritl + race + education + jobclass, data = wage)
# compare best fit gam model
anova(fit.gam1, fit.gam2)
```

```
## Analysis of Deviance Table
##
## Model 1: logwage ~ ns(year, 4) + ns(age, 4) + maritl + race + education +
##     jobclass + health + health_ins
## Model 2: logwage ~ ns(age, 4) + maritl + race + education + jobclass
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      2977     225.38
## 2      2983     249.31 -6  -23.936 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The knots were divided into 25%, 50%, and 75%, and model1 using all explanatory variables and model2 using age, maritl, race, education, and jobclass were compared.
Through the anova test, it can be seen that Model 2 is a more suitable model.

```
# summary gam model
summary(fit.gam2)
```

```
##
## Call: gam(formula = logwage ~ ns(age, 4) + maritl + race + education +
##     jobclass, data = wage)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.75989 -0.15774  0.01124  0.17052  1.15738
##
## (Dispersion Parameter for gaussian family taken to be 0.0836)
##
##     Null Deviance: 371.0659 on 2999 degrees of freedom
## Residual Deviance: 249.3108 on 2983 degrees of freedom
## AIC: 1086.629
##
## Number of Local Scoring Iterations: 2
```

```
##
## Anova for Parametric Effects
##               Df  Sum Sq Mean Sq  F value    Pr(>F)
## ns(age, 4)     4  43.279 10.8197 129.4580 < 2.2e-16 ***
## maritl         4   9.065  2.2662  27.1146 < 2.2e-16 ***
## race           3   2.238  0.7461   8.9272 6.915e-06 ***
## education      4  66.140 16.5350 197.8416 < 2.2e-16 ***
## jobclass       1   1.033  1.0330  12.3598 0.0004452 ***
## Residuals   2983 249.311  0.0836
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis is rejected because the p-value of all variables is much lower than 0.05 in generalized additive model.

h. Comment on which one of the several models that you constructed in the above exercises is the best suited for the data under consideration and support your conclusions with numerical evidence.

```r
# find AIC
AIC(lm.fit.log, fit.poly, fit.gam2)
```

```
##            df      AIC
## lm.fit.log 15 1189.713
## fit.poly   17 1084.727
## fit.gam2   18 1086.629
```

```r
# find BIC
BIC(lm.fit.log, fit.poly, fit.gam2)
```

```
##            df      BIC
## lm.fit.log 15 1279.808
## fit.poly   17 1186.835
## fit.gam2   18 1194.744
```

```r
# compare three models
anova(lm.fit.log, fit.poly, fit.gam2)
```

```
## Analysis of Variance Table
##
## Model 1: logwage ~ age + maritl + race + education + jobclass
## Model 2: logwage ~ poly(age, 3) + maritl + race + education + jobclass
## Model 3: logwage ~ ns(age, 4) + maritl + race + education + jobclass
##   Res.Df    RSS Df Sum of Sq       F Pr(>F)
## 1   2986 258.54
## 2   2984 249.32  2    9.2240 55.1824 <2e-16 ***
## 3   2983 249.31  1    0.0081  0.0968 0.7558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obtained AIC and BIC of several models constructed in the above exercise (the smaller the AIC and BIC, the better the model).
AIC:
Linear regression: **1189.713**
Polynomial regression: **1084.727**
Generalized additive model: **1086.629**

$$Linear < GAM < Polynomial$$

BIC:
Linear regression: **1279.808**
Polynomial regression: **1186.835**
Generalized additive model: **1194.744**

$$Linear < GAM < Polynomial$$

As such, the values of AIC and BIC of the polynomial regression model are the smallest, and additional three models are compared through anova test, indicating that the **polynomial regression model is the most suitable model**.