Economic Card Dataset

Dylan Chou (dvchou)

11/23/2020

Reading in Card Data and GLM construction

```
library("boot")
```

```
set.seed(101)

card = read.csv("card.csv")
card = card[,c("X",
               "educ",
               "wage",
               "age",
               "black",
               "married",
               "region",
               "south",
               "kww",
               "iq",
               "exper")]
card$black = as.factor(card$black)
card$married = as.factor(card$married)
card$region = as.factor(card$region)
card$south = as.factor(card$south)
head(card)
```

```
##   X educ     wage age black married region south kww       iq exper
## 1 1    7 27.24656  29     1       1      1     0  15 102.4498    16
## 2 2   12 23.91532  27     0       1      1     0  35  93.0000     9
## 3 3   12 35.84813  34     0       1      1     0  42 103.0000    16
## 4 4   11 12.43000  27     0       1      2     0  25  88.0000    10
## 5 5   12 36.24588  34     0       1      2     0  34 108.0000    16
## 6 6   12 24.86000  26     0       1      2     0  38  85.0000     8
```

```
newLogWage = glm(log(wage)~educ+age+black+married+region+iq+kww+south,data=card)
cv.glm(card,newLogWage,K=10)$delta[1]
```

```
## [1] 0.139402
```

```r
mean(((log(card$wage)-fitted(newLogWage))/(1-hatvalues(newLogWage)))^2)
```

```
## [1] 0.1392593
```

```r
# LOOCV and CV errors similar for the model fit
```

Evaluating the new model

```r
set.seed(101)

interactionLogModel = glm(log(wage)~black+married+
                            region+iq+kww+south
                          +iq*educ
                          +poly(age,2),
                          data=card)
summary(interactionLogModel)
```

```
##
## Call:
## glm(formula = log(wage) ~ black + married + region + iq + kww +
##     south + iq * educ + poly(age, 2), data = card)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.61121  -0.23088   0.00944   0.24094   1.44834
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.9640213  0.3357764   5.849 5.47e-09 ***
## black1         -0.1111143  0.0195996  -5.669 1.57e-08 ***
## married1        0.1399185  0.0155635   8.990  < 2e-16 ***
## region2         0.1206504  0.0357929   3.371 0.000759 ***
## region3         0.1535727  0.0351086   4.374 1.26e-05 ***
## region4         0.0174129  0.0414217   0.420 0.674237
## region5         0.1327480  0.0416196   3.190 0.001440 **
## region6         0.1239633  0.0448165   2.766 0.005709 **
## region7         0.1429622  0.0446398   3.203 0.001376 **
## region8        -0.0639819  0.0513274  -1.247 0.212662
## region9         0.1606599  0.0388842   4.132 3.70e-05 ***
## iq              0.0047246  0.0032508   1.453 0.146229
## kww             0.0087879  0.0011002   7.988 1.94e-15 ***
## south1         -0.1710010  0.0257528  -6.640 3.71e-11 ***
## educ            0.0637878  0.0245623   2.597 0.009451 **
## poly(age, 2)1   4.6732540  0.4271917  10.939  < 2e-16 ***
## poly(age, 2)2 -0.6659543  0.3768321  -1.767 0.077289 .
## iq:educ        -0.0003294  0.0002330  -1.414 0.157482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1383322)
```

```
##
##     Null deviance: 592.64  on 3009  degrees of freedom
## Residual deviance: 413.89  on 2992  degrees of freedom
## AIC: 2607.9
##
## Number of Fisher Scoring iterations: 2
```

```
cv.glm(card,interactionLogModel,K=10)$delta[1]
```

```
## [1] 0.1394704
```

```
mean(((log(card$wage)-fitted(interactionLogModel))/(1-hatvalues(interactionLogModel)))^2)
```

```
## [1] 0.1392527
```

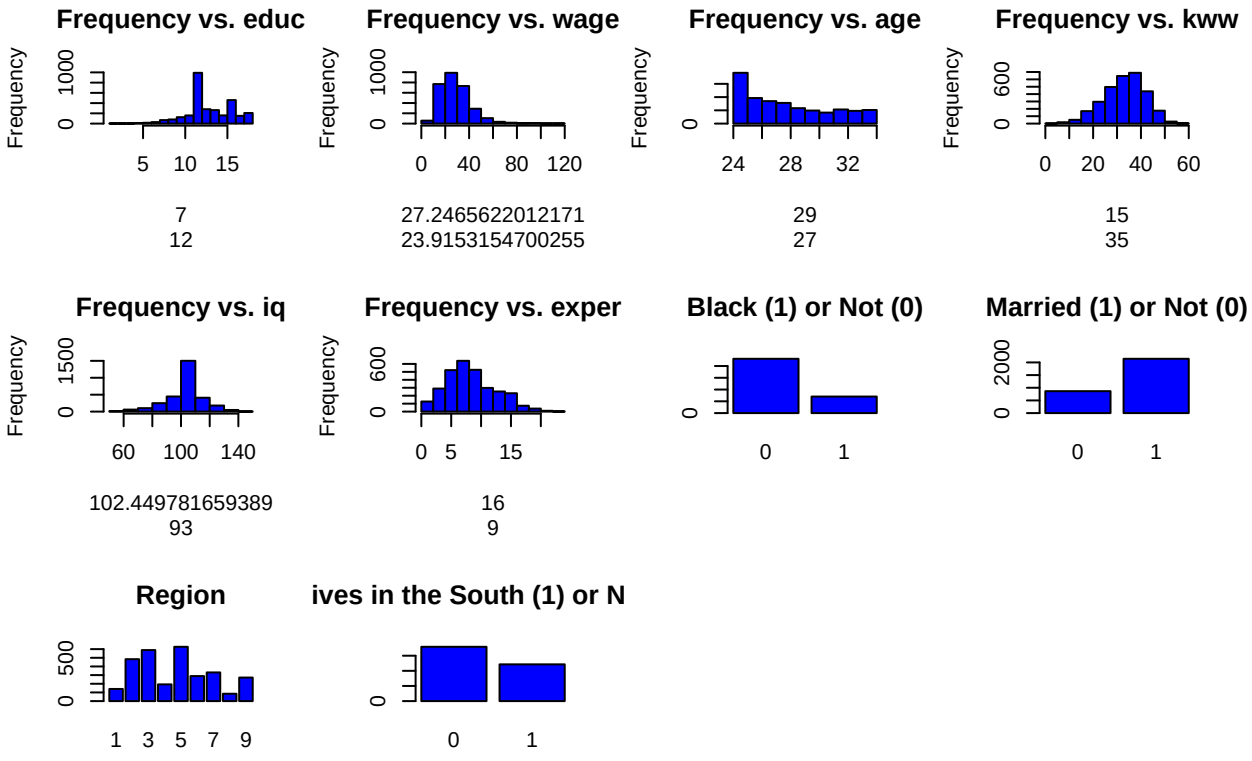
Further EDA on the Card Dataset

```
summary(card)
```

```
##        X                educ            wage                age         black
##  Min.   :   1.0   Min.   : 1.00   Min.   :  4.972   Min.   :24.00   0:2307
##  1st Qu.: 753.2   1st Qu.:12.00   1st Qu.: 19.602   1st Qu.:25.00   1: 703
##  Median :1505.5   Median :13.00   Median : 26.724   Median :28.00
##  Mean   :1505.5   Mean   :13.26   Mean   : 28.702   Mean   :28.12
##  3rd Qu.:2257.8   3rd Qu.:16.00   3rd Qu.: 35.239   3rd Qu.:31.00
##  Max.   :3010.0   Max.   :18.00   Max.   :119.527   Max.   :34.00
##
##  married      region    south         kww              iq
##  0: 866   5      :627   0:1795   Min.   : 4.00   Min.   : 50.0
##  1:2144   3      :589   1:1215   1st Qu.:28.00   1st Qu.: 98.0
##           2      :484            Median :34.00   Median :102.4
##           7      :331            Mean   :33.54   Mean   :102.4
##           6      :289            3rd Qu.:40.00   3rd Qu.:108.0
##           9      :272            Max.   :56.00   Max.   :149.0
##           (Other):418
##      exper
##  Min.   : 0.000
##  1st Qu.: 6.000
##  Median : 8.000
##  Mean   : 8.856
##  3rd Qu.:11.000
##  Max.   :23.000
##
```

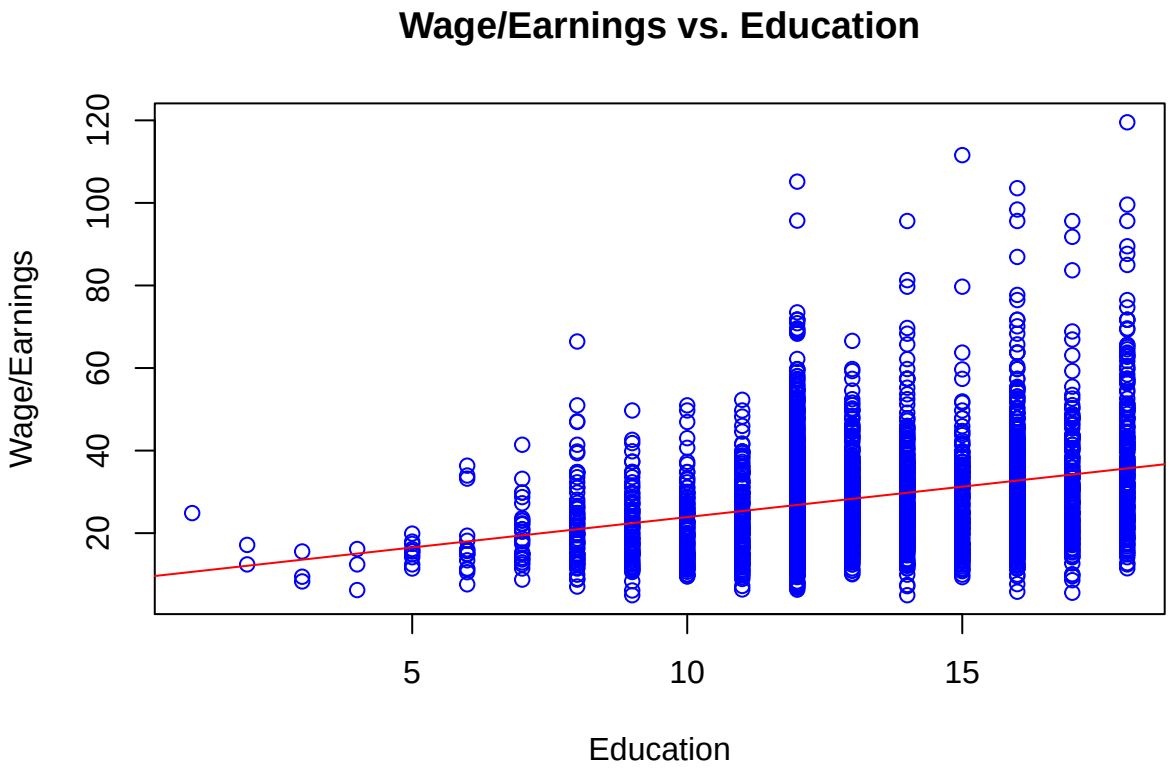
```
par(mfrow=c(3,4))
hist(card[["educ"]],
     xlab=card$educ,
     main=sprintf("Frequency vs. %s", "educ"),
     col="blue")
```

```r
hist(card[["wage"]],
     xlab=card$wage,
     main=sprintf("Frequency vs. %s", "wage"),
     col="blue")
hist(card[["age"]],
     xlab=card$age,
     main=sprintf("Frequency vs. %s", "age"),
     col="blue")
hist(card[["kww"]],
     xlab=card$kww,
     main=sprintf("Frequency vs. %s", "kww"),
     col="blue")
hist(card[["iq"]],
     xlab=card$iq,
     main=sprintf("Frequency vs. %s", "iq"),
     col="blue")
hist(card[["exper"]],
     xlab=card$exper,
     main=sprintf("Frequency vs. %s", "exper"),
     col="blue")
barplot(table(card$black),col="blue",main="Black (1) or Not (0)")
barplot(table(card$married),col="blue",main="Married (1) or Not (0)")
barplot(table(card$region),col="blue",main="Region")
barplot(table(card$south),col="blue",main="Lives in the South (1) or Not (0)")
```



Earnings vs. Education Regression and Plot

```
plot(x=card$educ,
     y=card$wage,
     xlab="Education",
     ylab="Wage/Earnings",
     main="Wage/Earnings vs. Education",
     col="blue")
educEarnModel = lm(wage~educ,data=card)
abline(educEarnModel,col="red")
```




```
summary(educEarnModel)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = card)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.643  -8.619  -1.697   6.355  83.841
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.1459     1.1487   7.962 2.38e-15 ***
## educ          1.4745     0.0849  17.368  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

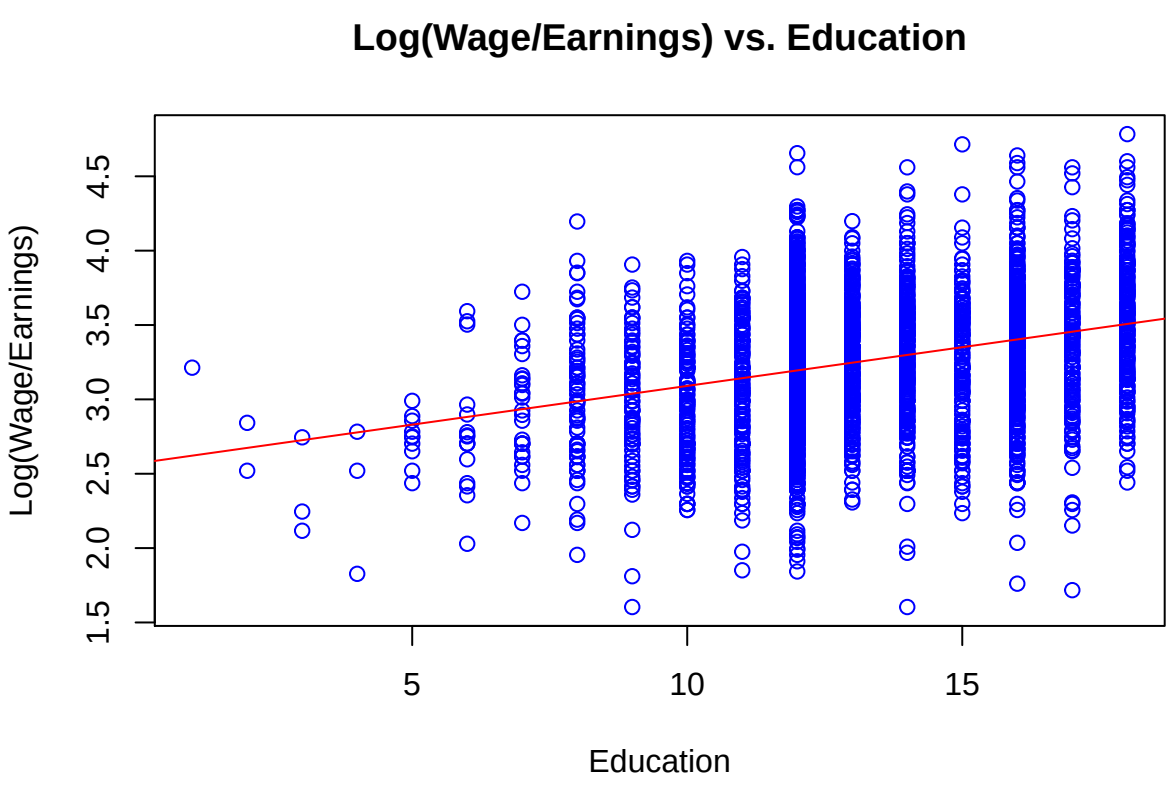
```
##
## Residual standard error: 12.47 on 3008 degrees of freedom
## Multiple R-squared:  0.09114,    Adjusted R-squared:  0.09084
## F-statistic: 301.6 on 1 and 3008 DF,  p-value: < 2.2e-16
```

```
confint(educEarnModel, 'educ', level=0.95)
```

```
##         2.5 %   97.5 %
## educ 1.308006 1.640931
```

Log Earnings vs. Education Regression and Plot

```
plot(x=card$educ,y=log(card$wage),
     xlab="Education",
     ylab="Log(Wage/Earnings)",
     main="Log(Wage/Earnings) vs. Education",
     col="blue")
logeducEarnModel = lm(log(wage)~educ,data=card)
abline(logeducEarnModel,col="red")
```




```
summary(logeducEarnModel)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ, data = card)
##
```

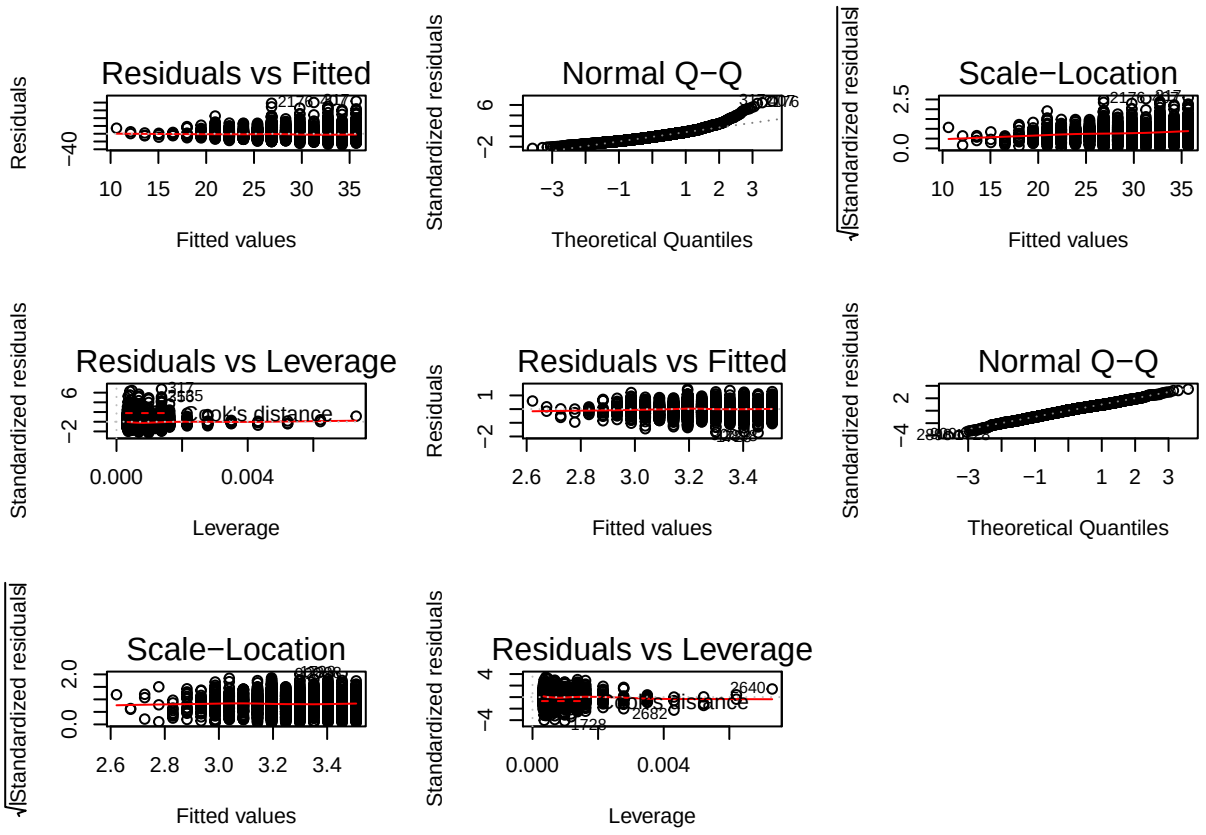
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73799 -0.27764  0.02373  0.28839  1.46080
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.56953    0.03883   66.17   <2e-16 ***
## educ         0.05209    0.00287   18.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4214 on 3008 degrees of freedom
## Multiple R-squared:  0.09874,    Adjusted R-squared:  0.09844
## F-statistic: 329.5 on 1 and 3008 DF,  p-value: < 2.2e-16
```

```
confint(logeducEarnModel, 'educ', level=0.95)
```

```
##           2.5 %     97.5 %
## educ 0.04646744 0.05772102
```

(c)

```
par(mfrow=c(3,3))
plot(educEarnModel)

plot(logeducEarnModel)
```



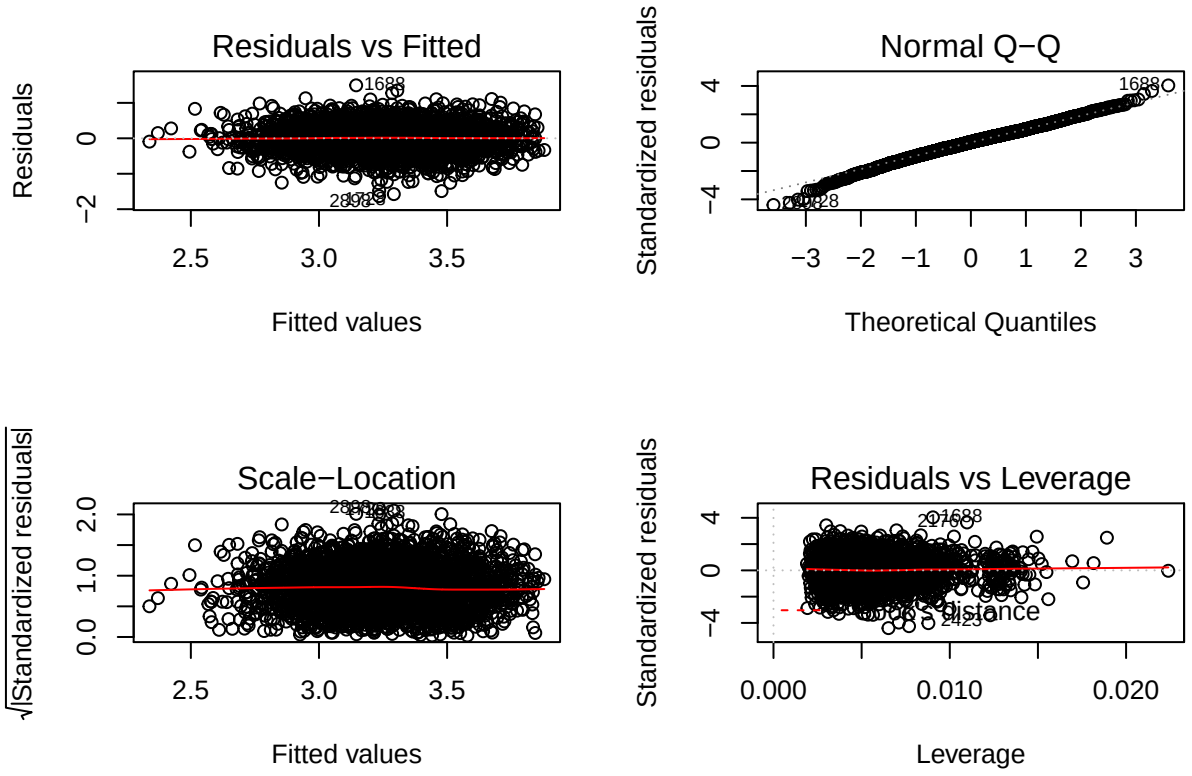
Log Earnings Regression vs Education, Age, Race, Marriage Status, Region, IQ, KWW, South

```
par(mfrow=c(2,2))
newLogWage = lm(log(wage)~educ+age+black+married+region+iq+kww+south,data=card)
summary(newLogWage)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + age + black + married + region +
##     iq + kww + south, data = card)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63072 -0.22937  0.01542  0.24118  1.49509
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.6763037  0.0972067  17.245  < 2e-16 ***
## educ         0.0295873  0.0030650   9.653  < 2e-16 ***
## age          0.0266427  0.0024679  10.796  < 2e-16 ***
## black1      -0.1102251  0.0196038  -5.623 2.05e-08 ***
## married1     0.1426869  0.0155156   9.196  < 2e-16 ***
## region2      0.1200313  0.0358065   3.352 0.000812 ***
## region3      0.1547695  0.0351023   4.409 1.07e-05 ***
## region4      0.0196517  0.0414124   0.475 0.635152
```

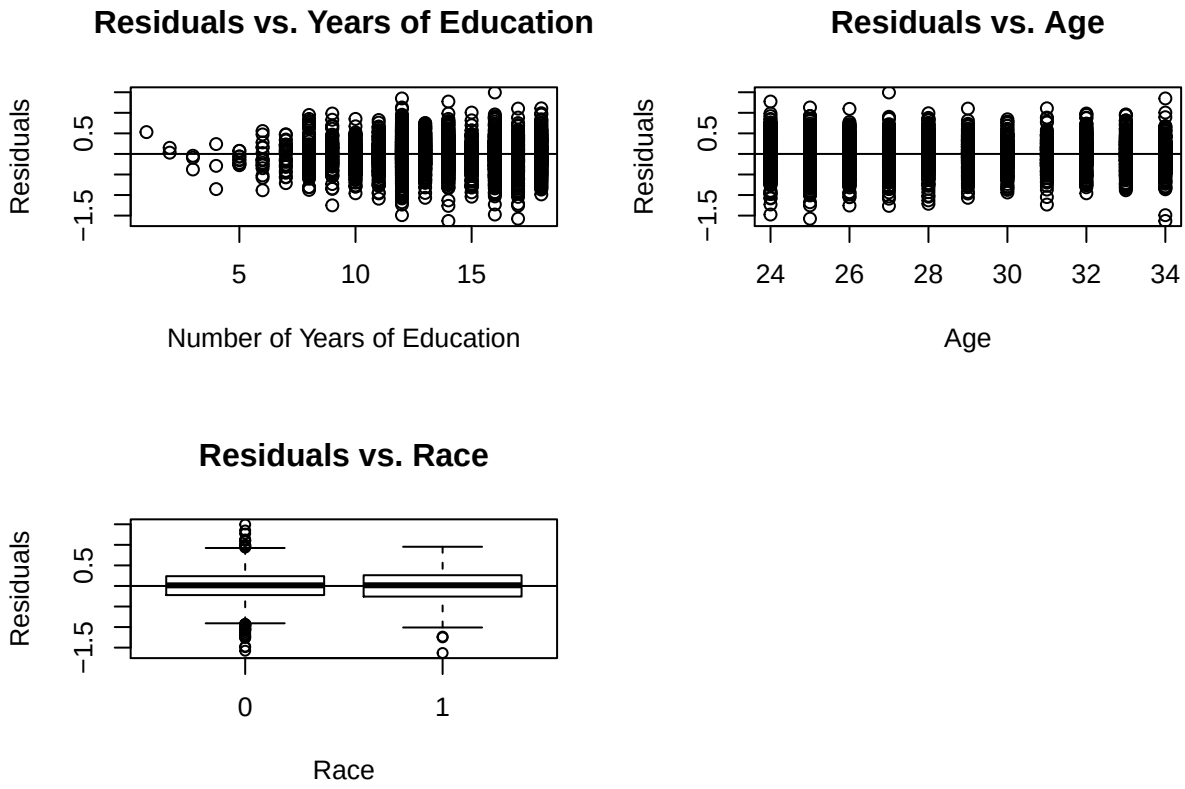
```
## region5      0.1338119  0.0416350   3.214 0.001323 **
## region6      0.1289298  0.0447631   2.880 0.004002 **
## region7      0.1458298  0.0446395   3.267 0.001100 **
## region8     -0.0601312  0.0513092  -1.172 0.241315
## region9      0.1617159  0.0388762   4.160 3.28e-05 ***
## iq           0.0001470  0.0005928   0.248 0.804118
## kww          0.0089672  0.0010976   8.170 4.49e-16 ***
## south1      -0.1714418  0.0257626  -6.655 3.36e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3721 on 2994 degrees of freedom
## Multiple R-squared:  0.3005, Adjusted R-squared:  0.297
## F-statistic: 85.76 on 15 and 2994 DF,  p-value: < 2.2e-16
```

```
plot(newLogWage)
```

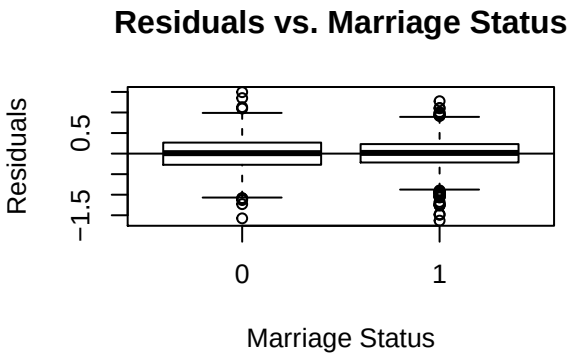


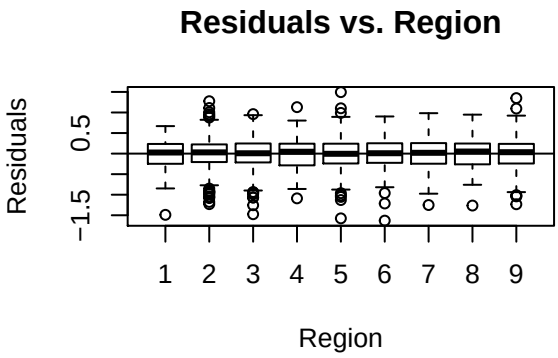

```
# aside from plotting residuals vs fitted we look at residual vs. covariate plots

plot(card$educ,resid(newLogWage),
     xlab="Number of Years of Education",
     ylab="Residuals",
     main="Residuals vs. Years of Education")
abline(a=0,b=0)
plot(card$age,resid(newLogWage),
     xlab="Age",
     ylab="Residuals",
     main="Residuals vs. Age")
```

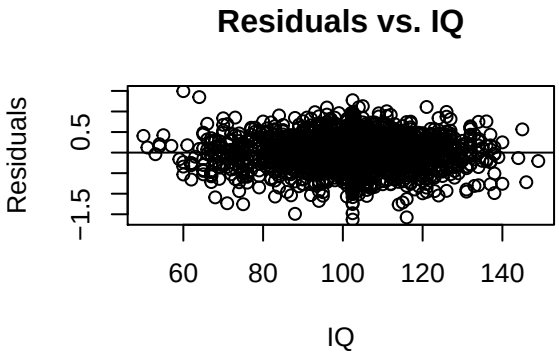
```
abline(a=0,b=0)
plot(card$black,resid(newLogWage),
     xlab="Race",
     ylab="Residuals",
     main="Residuals vs. Race")
abline(a=0,b=0)
```

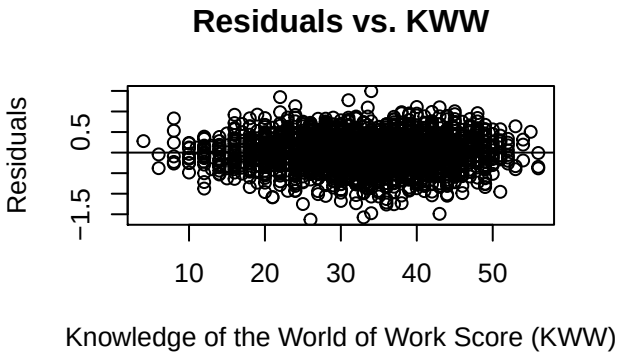



```
par(mfrow=c(2,2))
plot(card$married,resid(newLogWage),
     xlab="Marriage Status",
     ylab="Residuals",
     main="Residuals vs. Marriage Status")
abline(a=0,b=0)
plot(card$region,resid(newLogWage),
     xlab="Region",
     ylab="Residuals",
     main="Residuals vs. Region")
abline(a=0,b=0)
plot(card$iq,resid(newLogWage),
     xlab="IQ",
     ylab="Residuals",
     main="Residuals vs. IQ")
abline(a=0,b=0)
plot(card$kww,resid(newLogWage),
     xlab="Knowledge of the World of Work Score (KWW)",
     ylab="Residuals",
     main="Residuals vs. KWW")
abline(a=0,b=0)
```

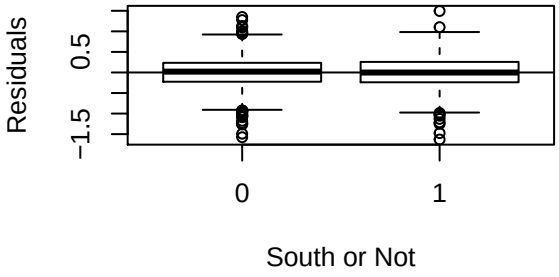









```
plot(card$south,
     resid(newLogWage),
     xlab="South or Not",
     ylab="Residuals",
     main="Residuals vs. South")
abline(a=0,b=0)
```



Confidence Interval of Race Coefficient

```
summary(newLogWage)$coefficients[4,] # race row
```

```
##      Estimate    Std. Error       t value      Pr(>|t|)
## -1.102251e-01  1.960375e-02 -5.622654e+00  2.053554e-08
```

```r
confint(newLogWage, "black1",level=0.95)
```

```
##             2.5 %      97.5 %
## black1 -0.1486633 -0.07178693
```

Confidence Interval of Education Coefficient

```r
summary(newLogWage)$coefficients[2,] # race row
```

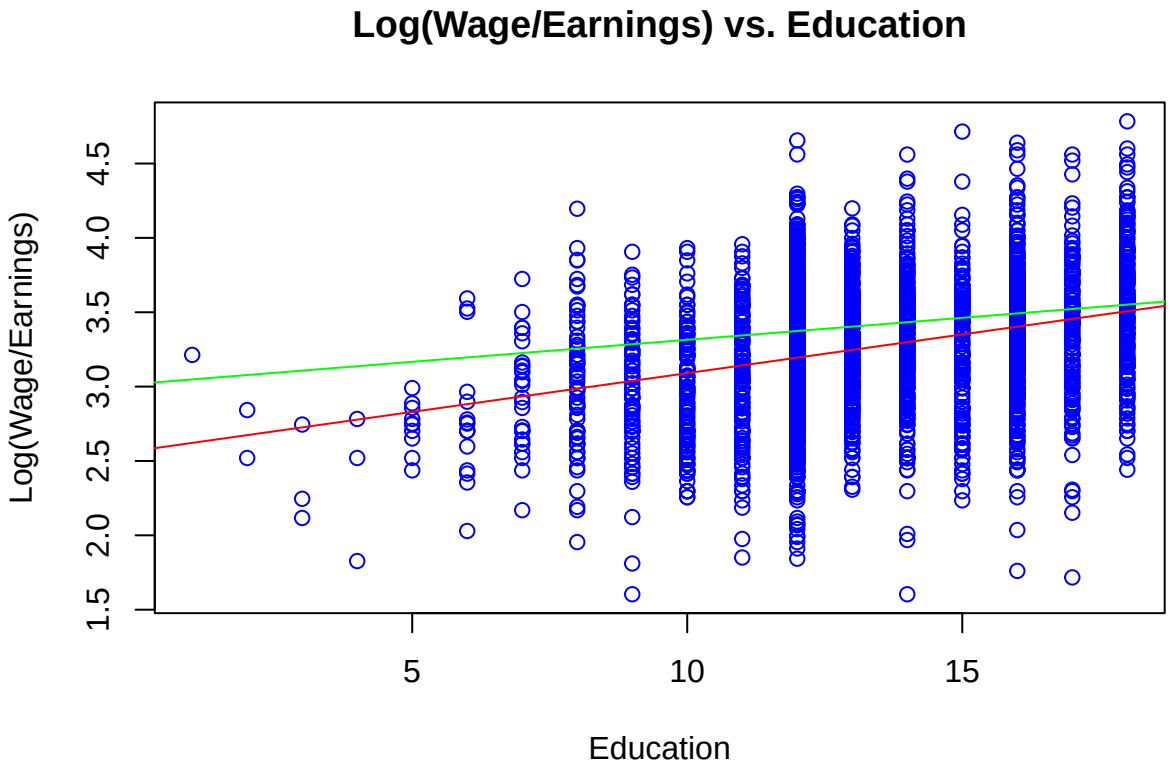
```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 2.958727e-02 3.065042e-03 9.653134e+00 9.849542e-22
```

```r
confint(newLogWage, "educ",level=0.95)
```

```
##           2.5 %     97.5 %
## educ 0.02357746 0.03559707
```

Plotting Model Log Earnings vs. Education Holding Other Covariates Constant

```r
newEduModeld = lm(log(wage)~educ+region+age+iq+kww+married+black+south,data=card)
interceptval = newEduModeld$coefficients["(Intercept)"]+
  newEduModeld$coefficients["age"]*median(card$age)+
      newEduModeld$coefficients["married1"]+
        newEduModeld$coefficients["iq"]*median(card$iq)+
        newEduModeld$coefficients["kww"]*median(card$kww)+
        newEduModeld$coefficients["region5"]
educ = newEduModeld$coefficients["educ"]

plot(x=card$educ,y=log(card$wage),
     xlab="Education",
     ylab="Log(Wage/Earnings)",
     main="Log(Wage/Earnings) vs. Education",
     col="blue")
abline(a=interceptval,b=educ,col="green")
abline(logeducEarnModel,col="red")
```



Log Earnings vs Education, Region, Age, IQ, KWW, Marriage Status, Experience, Living in the South and Race

```
newMedModele = lm(log(wage)~educ+region+age+iq+kww+married+black+south+exper,data=card)
summary(newMedModele)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + region + age + iq + kww + married +
##     black + south + exper, data = card)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63072 -0.22937  0.01542  0.24118  1.49509
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.6763037  0.0972067  17.245  < 2e-16 ***
## educ         0.0295873  0.0030650   9.653  < 2e-16 ***
## region2      0.1200313  0.0358065   3.352 0.000812 ***
## region3      0.1547695  0.0351023   4.409 1.07e-05 ***
## region4      0.0196517  0.0414124   0.475 0.635152
## region5      0.1338119  0.0416350   3.214 0.001323 **
## region6      0.1289298  0.0447631   2.880 0.004002 **
## region7      0.1458298  0.0446395   3.267 0.001100 **
## region8     -0.0601312  0.0513092  -1.172 0.241315
## region9      0.1617159  0.0388762   4.160 3.28e-05 ***
```

```
## age          0.0266427  0.0024679  10.796  < 2e-16 ***
## iq           0.0001470  0.0005928   0.248 0.804118
## kww          0.0089672  0.0010976   8.170 4.49e-16 ***
## married1     0.1426869  0.0155156   9.196  < 2e-16 ***
## black1      -0.1102251  0.0196038  -5.623 2.05e-08 ***
## south1      -0.1714418  0.0257626  -6.655 3.36e-11 ***
## exper               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3721 on 2994 degrees of freedom
## Multiple R-squared:  0.3005, Adjusted R-squared:  0.297
## F-statistic: 85.76 on 15 and 2994 DF,  p-value: < 2.2e-16
```

```
summary(lm(age~exper,data=card))
```

```
##
## Call:
## lm(formula = age ~ exper, data = card)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.827 -1.625 -0.157  1.375  5.219
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.001009   0.087262  263.59   <2e-16 ***
## exper        0.577971   0.008926   64.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.028 on 3008 degrees of freedom
## Multiple R-squared:  0.5823, Adjusted R-squared:  0.5821
## F-statistic:  4193 on 1 and 3008 DF,  p-value: < 2.2e-16
```

```
summary(lm(educ~exper,data=card))
```

```
##
## Call:
## lm(formula = educ ~ exper, data = card)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.827 -1.625 -0.157  1.375  5.219
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.001009   0.087262  194.83   <2e-16 ***
## exper       -0.422029   0.008926  -47.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.028 on 3008 degrees of freedom
```

```
## Multiple R-squared:  0.4264, Adjusted R-squared:  0.4262
## F-statistic:  2236 on 1 and 3008 DF,  p-value: < 2.2e-16
```

```
summary(lm(iq~exper,data=card))
```

```
##
## Call:
## lm(formula = iq ~ exper, data = card)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.512  -5.054   0.118   6.762  45.126
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 109.70661    0.52948  207.19   <2e-16 ***
## exper        -0.81941    0.05416  -15.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.3 on 3008 degrees of freedom
## Multiple R-squared:  0.07072,    Adjusted R-squared:  0.07041
## F-statistic: 228.9 on 1 and 3008 DF,  p-value: < 2.2e-16
```