CSCI E-106:Assignment 7

Due Date: November 9, 2020 at 7:20 pm EST

Instructions

Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document, word, or html generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

Solutions

Problem 1

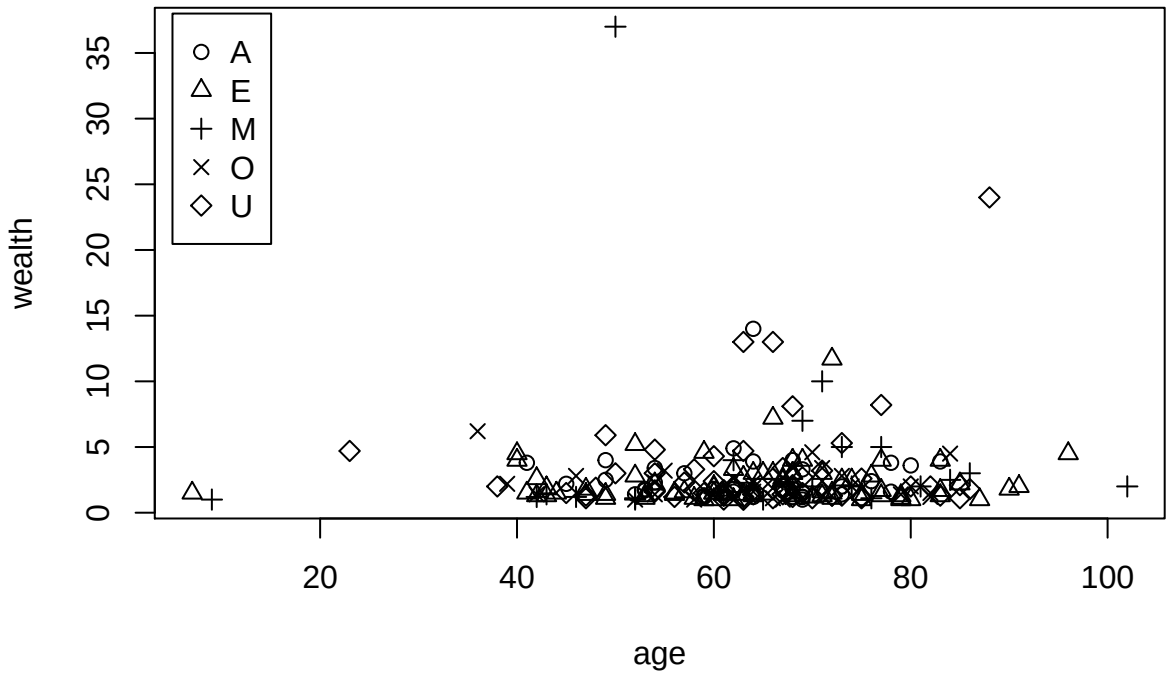
Use the fortune data under the faraway r library, data(fortune,package="faraway"). The wealth in billions of dollars for 232 billionaires is given in fortune. (50 points, 10 points each) (Hint: refer to the interaction.pdf and rmd files for details)

a-)Plot the wealth as a function of age using a different plotting symbol for the different regions of the world. All regions look similar and it is diffuclt to regions from each other.

```
data(fortune,package="faraway")
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
par(mfrow=c(1,1))
plot(wealth ~ age, fortune, pch=unclass(region))
legend(5,38,levels(fortune$region),pch=1:5)
```

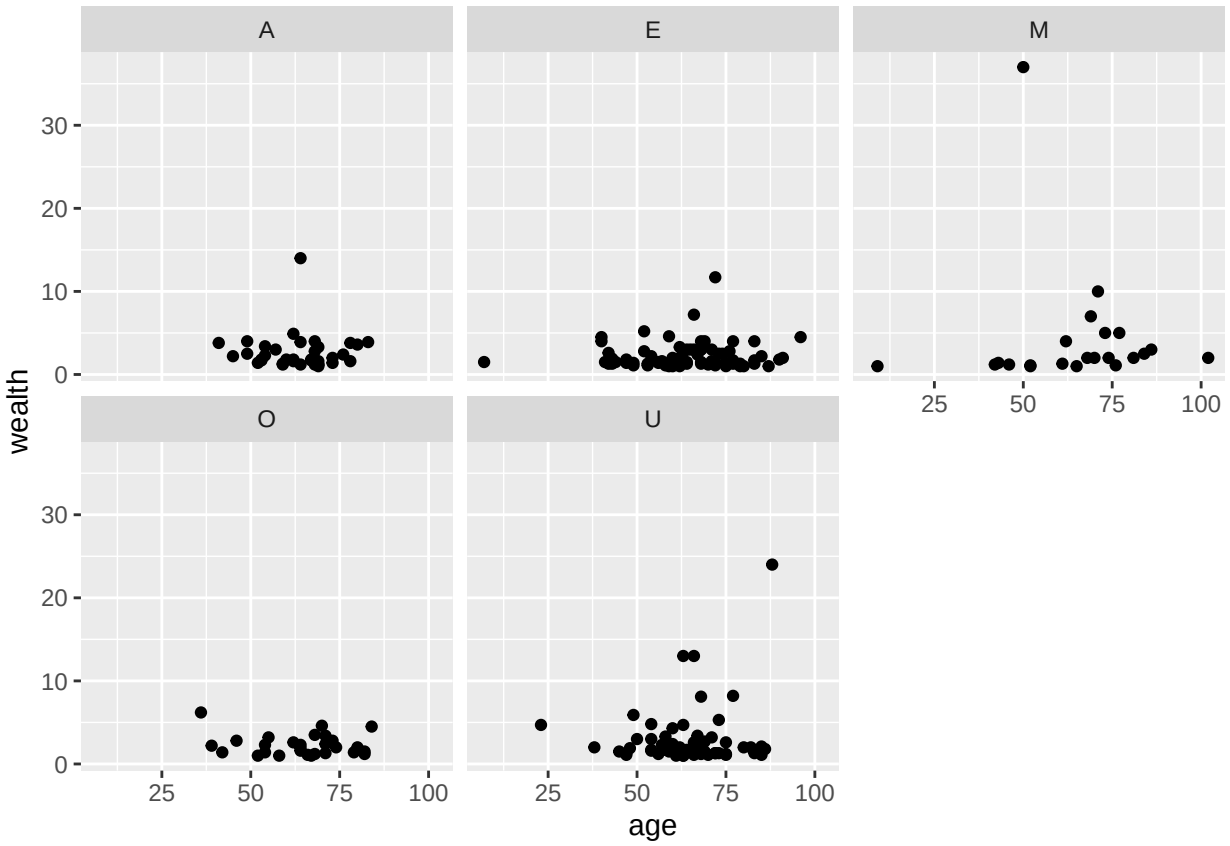


b-)Plot the wealth as a function of age with a separate panel for each region

Wealth and age distribution look very similar for each region. There are outliers in region M (with respect to age and wealth.)

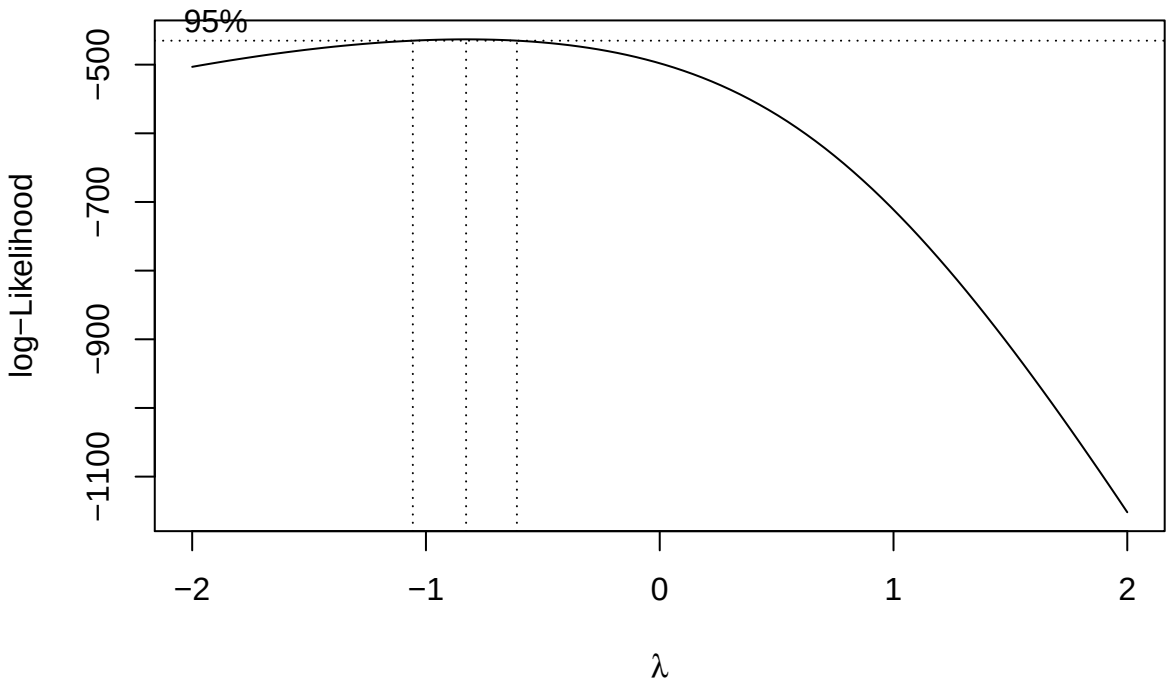
```
par(mfrow=c(1,1))
ggplot(aes(x=age,y=wealth),data=fortune) + geom_point() + facet_wrap( ~ region)
```

```
## Warning: Removed 7 rows containing missing values (geom_point).
```



c-)Determine a transformation on the response to facilitate linear modeling. it is difficult to see the transformation visually. I used boxcox which is suggesting 1/Wealth, however log(wealth) and $\sqrt{wealth}$ are common choices with this type of data sets.

```
f<-lm(wealth ~ age*region,data=fortune)
library(MASS)
boxcox(f,lambda=seq(-2,2,0.1))
```



d-)What is the relationship of age and region to wealth?

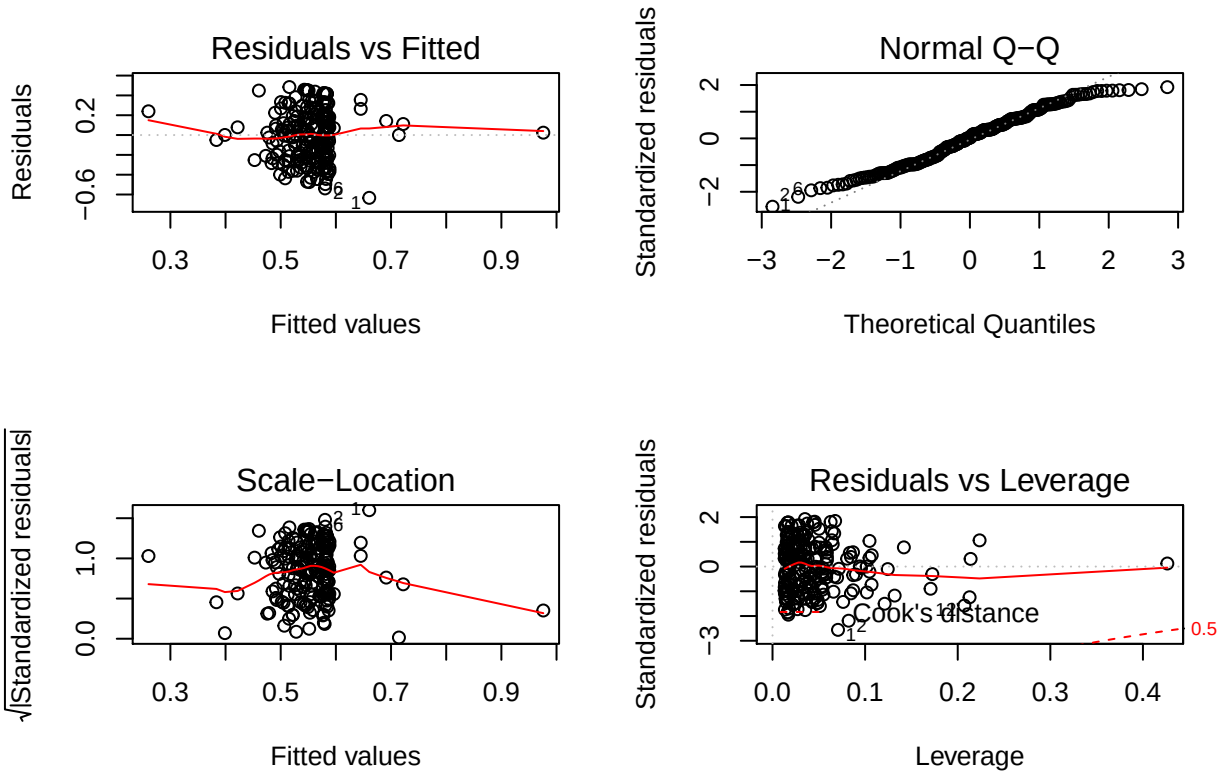
There is no relationship

```
summary(f)
```

```
##
## Call:
## lm(formula = wealth ~ age * region, data = fortune)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.737 -1.243 -0.736  0.499 32.357
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.721453   3.626188   0.750    0.454
## age         -0.001101   0.056305  -0.020    0.984
## regionE     -0.920828   4.020305  -0.229    0.819
## regionM      3.255513   4.415284   0.737    0.462
## regionO      0.526521   4.861590   0.108    0.914
## regionU     -2.894546   4.329271  -0.669    0.504
## age:regionE  0.008243   0.062202   0.133    0.895
## age:regionM -0.025575   0.067704  -0.378    0.706
## age:regionO -0.014012   0.074971  -0.187    0.852
## age:regionU  0.050569   0.066972   0.755    0.451
##
## Residual standard error: 3.364 on 215 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.04241,    Adjusted R-squared:  0.002325
## F-statistic: 1.058 on 9 and 215 DF,  p-value: 0.3953
```

e-)Check the assumptions of your model using appropriate diagnostics. Heavy tails and unequal variances

```r
f1<-lm(1/wealth ~ age*region,data=fortune)
par(mfrow=c(2,2))
plot(f1)
```



Problem 2

Refer to the CDI data set. A regression model relating serious crime rate (Y, total serious crimes divided by total population) to population density (X_1, total population divided by land area) and unemployment rate (X_3) is to be constructed. (30 points, 10 points each)

a-) Fit second-order regression model. Plot the residuals against the fitted values. How well does the second-order model appear to fit the data? What is R^2?

The R^2 is 24%. The only X1 and X1*X3 (interaction term) are significant.

```r
CDI<- read.csv("/cloud/project/Fall 2020/CDI Data.csv")
CDI[,18]=CDI$Total.serious.crimes/CDI$Total.population
dimnames(CDI)[[2]][18]<-"Serious.crime.rate"

CDI[,19]=CDI$Total.population/CDI$Land.area
dimnames(CDI)[[2]][19]<-"Population.density"

Y=CDI$Serious.crime.rate
X1=CDI$Population.density
X3=CDI$Percent.unemployment

#centering the variables, subtracting the mean
x1<-scale(X1, scale = FALSE)
```

```r
x3<-scale(X3, scale = FALSE)

x11=x1^2
x33=x3^2
x13=x1*x3


f<-lm(Y~x1+x3+x11+x33+x13)
summary(f)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x3 + x11 + x33 + x13)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.055642 -0.016851 -0.002889  0.014810  0.085485
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.629e-02  1.260e-03  44.662  < 2e-16 ***
## x1           4.585e-06  9.841e-07   4.659 4.23e-06 ***
## x3          -8.800e-05  6.276e-04  -0.140   0.8886
## x11          2.698e-12  5.932e-11   0.045   0.9637
## x33          1.629e-04  9.541e-05   1.708   0.0884 .
## x13          8.334e-07  4.091e-07   2.037   0.0423 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02383 on 434 degrees of freedom
## Multiple R-squared:  0.2485, Adjusted R-squared:  0.2398
## F-statistic:  28.7 on 5 and 434 DF,  p-value: < 2.2e-16
```

b-) Test whether or not all quadratic and interaction terms can be dropped from the regression model; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

H_0: The variables can be dropped H_a: The variables can NOT be dropped

P value is $0.02 > 0.01$. Accept,H_0. The variables can be dropped from the model.

```r
fr<-lm(Y~x1+x3)
summary(fr)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x3)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.053806 -0.016940 -0.003898  0.014680  0.084508
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 5.729e-02  1.144e-03  50.054   <2e-16 ***
## x1          5.973e-06  5.222e-07  11.439   <2e-16 ***
## x3          3.618e-04  4.902e-04   0.738    0.461
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02401 on 437 degrees of freedom
## Multiple R-squared:  0.2318, Adjusted R-squared:  0.2283
## F-statistic: 65.92 on 2 and 437 DF,  p-value: < 2.2e-16
```

```
anova(fr,f,test="F")
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ x1 + x3
## Model 2: Y ~ x1 + x3 + x11 + x33 + x13
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1    437 0.25186
## 2    434 0.24638  3  0.005477 3.2159 0.02278 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c-) Instead of the predictor variable population density, total population (X_1 and land area (X_2) are to be employed as separate predictor variables, in addition to unemployment rate (X_3). The regression model should contain linear and quadratic terms for total population, and linear terms only for land area and unemployment rate. (No interaction terms are to be included in this model.) Fit this regression model and obtain R^2. Is this coefficient of multiple determination substantially different from the one for the regression model in part (a)?

The R^2 is 15%. The only X1 and $X1^2$ are significant.The model in part a has a higher R^2.

```
Y=CDI$Serious.crime.rate
X1=CDI$Total.population
X2=CDI$Land.area
X3=CDI$Percent.unemployment

#centering the variables, subtracting the mean
x1<-scale(X1, scale = FALSE)
x2<-scale(X2, scale = FALSE)
x3<-scale(X3, scale = FALSE)

x11=x1^2
x22=x2^2
x33=x3^2

g<-lm(Y~x1+x2+x3+x11+x22+x33)
summary(g)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + x3 + x11 + x22 + x33)
##
## Residuals:
```

```
##       Min        1Q    Median        3Q       Max
## -0.059310 -0.016657 -0.003452  0.013547  0.191792
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.815e-02  1.397e-03  41.618  < 2e-16 ***
## x1           2.969e-08  3.573e-09   8.309 1.25e-15 ***
## x2          -1.901e-07  1.412e-06  -0.135    0.893
## x3           3.005e-04  6.674e-04   0.450    0.653
## x11         -3.388e-15  5.928e-16  -5.714 2.05e-08 ***
## x22         -4.592e-11  1.151e-10  -0.399    0.690
## x33          8.585e-05  1.007e-04   0.852    0.395
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02542 on 433 degrees of freedom
## Multiple R-squared:  0.1464, Adjusted R-squared:  0.1345
## F-statistic: 12.37 on 6 and 433 DF,  p-value: 7.075e-13
```

Problem 3

Refer to the CDI data set. The number of active physicians (Y) is to be regressed against total population (X_1, total personal income (X_2), and geographic region (X_3, X_4 , X_5). (20 points, 10 points each)

("Geographic region classification is that used by the U.S. Bureau of the Census, where: 1 = NE, 2 = NC, S = 3, W=4")

a-) Fit a first-order regression model. Let $X_3 = 1$ if NE and 0 otherwise, $X_4 = 1$ if NC and 0 otherwise, and $X_5 = 1$ if S and 0 otherwise. Examine whether the effect for the northeastern region on number of active physicians differs from the effect for the north central region by constructing an appropriate 90 percent confidence interval. Interpret your interval estimate.

R^2 is 90%. X1,X2, and X5 are significant variables. When we look at the simultaneous confidence interval for β_3 and β_4, they are very close to each other and the coefficients are also very close to each other (149 vs 145). Their impacts are similar.

```
Y=CDI$Number.of.active.physicians
X1=CDI$Total.population
X2=CDI$Total.personal.income

X3=as.numeric(CDI$Geographic.region==1)
X4=as.numeric(CDI$Geographic.region==2)
X5=as.numeric(CDI$Geographic.region==3)

g<-lm(Y~X1+X2+X3+X4+X5)
summary(g)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1866.8  -207.7   -81.5    72.4  3721.7
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.075e+02  7.028e+01  -2.952  0.00332 **
## X1           5.515e-04  2.835e-04   1.945  0.05243 .
## X2           1.070e-01  1.325e-02   8.073  6.8e-15 ***
## X3           1.490e+02  8.683e+01   1.716  0.08685 .
## X4           1.455e+02  8.515e+01   1.709  0.08817 .
## X5           1.912e+02  8.003e+01   2.389  0.01731 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 566.1 on 434 degrees of freedom
## Multiple R-squared:  0.9011, Adjusted R-squared:  0.8999
## F-statistic: 790.7 on 5 and 434 DF,  p-value: < 2.2e-16
```

```
confint(g,level=1-0.10/2)
```

```
##                     2.5 %        97.5 %
## (Intercept) -3.456304e+02 -69.361106971
## X1          -5.828310e-06   0.001108748
## X2           8.095958e-02   0.133063482
## X3          -2.164623e+01 319.685375829
## X4          -2.183695e+01 312.889843723
## X5           3.391581e+01 348.516796471
```

b-) Test whether any geographic effects are present; use $\alpha = .10$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

H_0: Geographic variables can be dropped H_a: Geographic variables can NOT be dropped

P value os 0.12. Accept H_0. Geographic variables can be dropped

```
g1<-lm(Y~X1+X2)
anova(g1,g)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2
## Model 2: Y ~ X1 + X2 + X3 + X4 + X5
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1    437 140967081
## 2    434 139093455  3   1873626 1.9487  0.121
```