fds <- read.csv(file.choose(), header=TRUE)

attach(fds)

a)

LE2013L= (1\*(LE2013 <=74))

LE2013H = (1\*(LE2013>74))

check = data.frame(LE2013,LE2013L,LE2013H)

fix(check)

model1= lm(HDI ~ LE2013L + LE2013H)

summary(model1)

Call:

lm(formula = HDI ~ LE2013L + LE2013H)

Residuals:

Min 1Q Median 3Q Max

-0.253047 -0.086047 0.008525 0.084239 0.243953

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.81348 0.01232 66.04 <2e-16 \*\*\*

LE2013L -0.22343 0.01628 -13.72 <2e-16 \*\*\*

LE2013H NA NA NA NA

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1102 on 185 degrees of freedom

Multiple R-squared: 0.5044, Adjusted R-squared: 0.5017

F-statistic: 188.3 on 1 and 185 DF, p-value: < 2.2e-16

> #We have HDI as the Y variable and LE2013 as the X variable.

>#we have model1 with the binary variable LE2013L (Low)

> #LE2013H (High)

> #They are all significant predictors since both p-value in intercept and

>#LE2013L are very low in P-value<.05(alpha)

b)

fds <- read.csv(file.choose(), header=TRUE)

attach(fds)

cut=cut(fds$MEANYRSCH,br=c(1,9,11,12,13))

table(cut)

cut

(1,9] (9,11] (11,12] (12,13]

109 46 18 14

> model3 = lm(HDI ~ cut)

> summary(model3)

Call:

lm(formula = HDI ~ cut)

Residuals:

Min 1Q Median 3Q Max

-0.25881 -0.08694 0.01094 0.08319 0.25619

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.59581 0.01071 55.632 < 2e-16 \*\*\*

cut(9,11] 0.18226 0.01966 9.271 < 2e-16 \*\*\*

cut(11,12] 0.24608 0.02845 8.650 2.58e-15 \*\*\*

cut(12,13] 0.28455 0.03174 8.964 3.62e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1118 on 183 degrees of freedom

Multiple R-squared: 0.495, Adjusted R-squared: 0.4867

F-statistic: 59.8 on 3 and 183 DF, p-value: < 2.2e-16

#Yes,they are all significant, since their p-value are all smaller than .05

#The coefficient of 1st term is the baseline value represents a nation

# which has been in cut(9,11] has 0.59581+ 0.18226in HDI.

c)

vars = data.frame(HDI, LE2013, MEANYRSCH, EYRSCH, GNI2013,HDI2012,CHINRANK,DL)

pairs(vars, upper.panel=NULL)



EY2 = EYRSCH\*EYRSCH

model4 = lm(HDI ~ LE2013 + MEANYRSCH + EYRSCH + EY2 + GNI2013 + HDI2012 +CHINRANK + DL)

summary(model4)

Call:

lm(formula = HDI ~ LE2013 + MEANYRSCH + EYRSCH + EY2 + GNI2013 +

HDI2012 + CHINRANK + DL)

Residuals:

Min 1Q Median 3Q Max

-0.0251828 -0.0007474 0.0001408 0.0009882 0.0039098

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -8.178e-03 4.531e-03 -1.805 0.07284 .

LE2013 4.503e-05 5.629e-05 0.800 0.42485

MEANYRSCH 2.960e-05 1.617e-04 0.183 0.85498

EYRSCH 1.019e-03 4.897e-04 2.081 0.03887 \*

EY2 -3.315e-05 1.859e-05 -1.783 0.07631 .

GNI2013 1.000e-08 1.896e-08 0.527 0.59856

HDI2012 9.980e-01 8.394e-03 118.888 < 2e-16 \*\*\*

CHINRANK 9.665e-04 1.509e-04 6.406 1.32e-09 \*\*\*

DLlow 2.108e-03 1.281e-03 1.646 0.10160

DLmedium 1.748e-03 6.638e-04 2.633 0.00922 \*\*

DLvery high -1.048e-03 7.774e-04 -1.348 0.17939

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002274 on 176 degrees of freedom

Multiple R-squared: 0.9998, Adjusted R-squared: 0.9998

F-statistic: 8.762e+04 on 10 and 176 DF, p-value: < 2.2e-16

> #Actually I believe there is no variable in the model for

> #a squared term to better fit the model

> #and the best I could pick is EYRSCH

> #as we find EY2 has a even bigger p-value,so if we have alpha as .05

> #then it is insignificant for our model.

d)

LE = LE2013\*EYRSCH

model5 = lm(HDI ~ LE2013 + MEANYRSCH + EYRSCH + GNI2013 + HDI2012 +CHINRANK + DL+LE)

summary(model5)

Call:

lm(formula = HDI ~ LE2013 + MEANYRSCH + EYRSCH + GNI2013 + HDI2012 +

CHINRANK + DL + LE)

Residuals:

Min 1Q Median 3Q Max

-0.0239673 -0.0006992 0.0000954 0.0009770 0.0055156

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.555e-02 8.129e-03 -3.143 0.00196 \*\*

LE2013 3.658e-04 1.214e-04 3.012 0.00298 \*\*

MEANYRSCH -2.181e-05 1.591e-04 -0.137 0.89112

EYRSCH 2.211e-03 6.792e-04 3.255 0.00136 \*\*

GNI2013 9.911e-09 1.864e-08 0.532 0.59548

HDI2012 9.990e-01 8.228e-03 121.420 < 2e-16 \*\*\*

CHINRANK 9.530e-04 1.480e-04 6.438 1.11e-09 \*\*\*

DLlow 2.107e-03 1.254e-03 1.680 0.09475 .

DLmedium 1.506e-03 6.599e-04 2.283 0.02364 \*

DLvery high -3.764e-04 8.108e-04 -0.464 0.64304

LE -2.887e-05 9.463e-06 -3.051 0.00263 \*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002236 on 176 degrees of freedom

Multiple R-squared: 0.9998, Adjusted R-squared: 0.9998

F-statistic: 9.062e+04 on 10 and 176 DF, p-value: < 2.2e-16

qt(.95,176)

[1] 1.653557

#The interaction term LE is significant. as alpha = 0.05 > .00263 (our p-value)

H0: β10 = 0

H1: β10≠ 0

Since by absolute value comparison, our t score =3.051>1.653557 ,so we reject Ho and believe it is significant predictor of Y in our model.

The larger the t-value,whichever size it is,in absolute term,the higher the rejection power, if our t score is larger,p-value smaller,we reject Ho.

e)

model6 = lm(HDI ~ factor(CHINRANK))

anova(model6)

Analysis of Variance Table

Response: HDI

Df Sum Sq Mean Sq F value Pr(>F)

factor(CHINRANK) 9 0.1453 0.016147 0.6517 0.7515

Residuals 177 4.3854 0.024776

# Ho: Beta1 = 0;

#H1: Beta1 !=0;

#Since our p-value is significantly larger than critical p-value alpha =.05.

#so we can’t reject Ho.This model is insignificant

//The smaller the p,the more likely we reject Ho

> summary(model6)

Call:

lm(formula = HDI ~ factor(CHINRANK))

Residuals:

Min 1Q Median 3Q Max

-0.35438 -0.11204 0.02562 0.11812 0.24862

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.78400 0.15741 4.981 1.5e-06

factor(CHINRANK)-4 -0.09800 0.19278 -0.508 0.612

factor(CHINRANK)-3 -0.12980 0.17243 -0.753 0.453

factor(CHINRANK)-2 -0.07671 0.16827 -0.456 0.649

factor(CHINRANK)-1 -0.11330 0.16129 -0.702 0.483

factor(CHINRANK)0 -0.08862 0.15809 -0.561 0.576

factor(CHINRANK)1 -0.14046 0.16040 -0.876 0.382

factor(CHINRANK)2 -0.08122 0.16592 -0.490 0.625

factor(CHINRANK)3 0.01550 0.19278 0.080 0.936

factor(CHINRANK)4 -0.29200 0.22260 -1.312 0.191

(Intercept) \*\*\*

factor(CHINRANK)-4

factor(CHINRANK)-3

factor(CHINRANK)-2

factor(CHINRANK)-1

factor(CHINRANK)0

factor(CHINRANK)1

factor(CHINRANK)2

factor(CHINRANK)3

factor(CHINRANK)4

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1574 on 177 degrees of freedom

Multiple R-squared: 0.03207, Adjusted R-squared: -0.01714

F-statistic: 0.6517 on 9 and 177 DF, p-value: 0.7515

> TukeyHSD(aov(HDI ~ factor(CHINRANK)))

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = HDI ~ factor(CHINRANK))

$`factor(CHINRANK)`

diff lwr upr p adj

-4--5 -0.098000000 -0.71577325 0.51977325 0.9999644

-3--5 -0.129800000 -0.68235319 0.42275319 0.9990812

-2--5 -0.076714286 -0.61595099 0.46252242 0.9999860

-1--5 -0.113300000 -0.63016618 0.40356618 0.9994732

0--5 -0.088622807 -0.59524005 0.41799443 0.9999185

1--5 -0.140461538 -0.65447995 0.37355688 0.9970054

2--5 -0.081222222 -0.61291678 0.45047234 0.9999742

3--5 0.015500000 -0.60227325 0.63327325 1.0000000

4--5 -0.292000000 -1.00534311 0.42134311 0.9498141

-3--4 -0.031800000 -0.45381947 0.39021947 0.9999999

-2--4 0.021285714 -0.38314181 0.42571324 1.0000000

-1--4 -0.015300000 -0.38938028 0.35878028 1.0000000

0--4 0.009377193 -0.35040945 0.36916384 1.0000000

1--4 -0.042461538 -0.41259710 0.32767402 0.9999978

2--4 0.016777778 -0.37753746 0.41109302 1.0000000

3--4 0.113500000 -0.39090975 0.61790975 0.9993492

4--4 -0.194000000 -0.81177325 0.42377325 0.9915721

-2--3 0.053085714 -0.24226639 0.34843782 0.9998977

-1--3 0.016500000 -0.23570487 0.26870487 1.0000000

0--3 0.041177193 -0.18929552 0.27164991 0.9999027

1--3 -0.010661538 -0.25697755 0.23565447 1.0000000

2--3 0.048577778 -0.23276854 0.32992409 0.9999270

3--3 0.145300000 -0.27671947 0.56731947 0.9838421

4--3 -0.162200000 -0.71475319 0.39035319 0.9948622

-1--2 -0.036585714 -0.25809979 0.18492836 0.9999499

0--2 -0.011908521 -0.20832355 0.18450651 1.0000000

1--2 -0.063747253 -0.27853265 0.15103814 0.9944249

2--2 -0.004507937 -0.25870656 0.24969068 1.0000000

3--2 0.092214286 -0.31221324 0.49664181 0.9992762

4--2 -0.215285714 -0.75452242 0.32395099 0.9570350

0--1 0.024677193 -0.09760647 0.14696086 0.9997320

1--1 -0.027161538 -0.17718553 0.12286245 0.9998912

2--1 0.032077778 -0.17038548 0.23454103 0.9999648

3--1 0.128800000 -0.24528028 0.50288028 0.9838375

4--1 -0.178700000 -0.69556618 0.33816618 0.9833698

1-0 -0.051838731 -0.16146342 0.05778596 0.8844512

2-0 0.007400585 -0.16724690 0.18204807 1.0000000

3-0 0.104122807 -0.25566384 0.46390945 0.9953793

4-0 -0.203377193 -0.70999443 0.30324005 0.9555299

2-1 0.059239316 -0.13583925 0.25431789 0.9933997

3-1 0.155961538 -0.21417402 0.52609710 0.9401450

4-1 -0.151538462 -0.66555688 0.36247995 0.9946954

3-2 0.096722222 -0.29759302 0.49103746 0.9987052

4-2 -0.210777778 -0.74247234 0.32091678 0.9589070

4-3 -0.307500000 -0.92527325 0.31027325 0.8489676

#when the p-value > alpha (0.05) or when the CI does contain 0, we believe that there is not a significant difference between these two groups.

f)

> anova(model1,model3,model4,model5,model6)

Analysis of Variance Table

Model 1: HDI ~ LE2013L + LE2013H

Model 2: HDI ~ cut

Model 3: HDI ~ LE2013 + MEANYRSCH + EYRSCH + EY2 + GNI2013 + HDI2012 +

CHINRANK + DL

Model 4: HDI ~ LE2013 + MEANYRSCH + EYRSCH + GNI2013 + HDI2012 + CHINRANK +

DL + LE

Model 5: HDI ~ factor(CHINRANK)

Res.Df RSS Df Sum of Sq F Pr(>F)

1 185 2.2456

2 183 2.2879 2 -0.0423

3 176 0.0009 7 2.2870 63197 < 2.2e-16 \*\*\*

4 176 0.0009 0 0.0000

5 177 4.3854 -1 -4.3846 848102 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1For model1 = lm(HDI~ LE2013L + LE2013H),the s=.1102, is really small in

> 185 df,the R2 = .5044 is small and R2 adj = .5017 changed very little

> #and p-value is almost 0.

> #For model3 =lm(formula = HDI ~ cut) the s is .1118, really small in

> #185 df,the R2 = .495 is small and R2 adj = .4867 changed very little

> #and p-value is almost 0.

#For model4 = lm(formula = HDI ~ LE2013 + MEANYRSCH + EYRSCH + EY2 + GNI2013 +

HDI2012 + CHINRANK + DL),the s = 0.002274,is really small in

#176 df,the R2 = .9998, is big and R2 adj remains the same

#and p-value is almost 0.

#For model5 = lm(formula = HDI ~ LE2013 + MEANYRSCH + EYRSCH + GNI2013 + HDI2012 +

CHINRANK + DL + LE),the s= 0.002236,is really small in

#176 df,the R2 = .9998 is big and R2 adj remains the same

#and p-value is almost 0.

#For model6 = lm(formula = HDI ~ factor(CHINRANK)) the s=.1576 ,is really small in

#177 df,the R2 = .03207 is small and R2 adj is -0.01714

#and p-value is .7515.

#I prefer model5 = lm(formula = HDI ~ LE2013 + MEANYRSCH + EYRSCH + GNI2013 + HDI2012 +

CHINRANK + DL + LE),because it has the smallest residual std error, and R2 is .9998 and is no different from R2 adj and p-value is 0 in 176 df.

g)  
model9 = lm(HDI ~ LE2013 + MEANYRSCH + GNI2013 + HDI2012 +CHINRANK + DL)

> anova(model9,model4)

#anova(reduced,full)

Analysis of Variance Table

Model 1: HDI ~ LE2013 + MEANYRSCH + GNI2013 + HDI2012 + CHINRANK + DL

Model 2: HDI ~ LE2013 + MEANYRSCH + EYRSCH + EY2 + GNI2013 + HDI2012 + CHINRANK + DL

Res.Df RSS Df Sum of Sq F Pr(>F)

1 178 0.00093430

2 176 0.00090989 2 2.4407e-05 2.3605 0.09735 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> dim(fds)

[1] 187 9

Our model has 8 variables

Df2=187 –(8+1)=176

> Ho = 0;

>H1!=0;

#since we have p-value =0.09735 > alpha =.05

#So we can’t reject the Ho and believe EY2 and EYRSCH are jointly insignificant predictors.

#By hand.

#**I have got Diff(Risiduals-reduced, Risduals-full) equal to** 2.4407e-05, thus in

**#F=(S(reduced)-S(full))/p/(S(full))^2, F=0,**

> dim(fds)

[1] 187 9

N=187,K=8 in our case

> 2.4407e-05/(2\* 0.002274\* 0.002274)

[1] 2.359954

> qf(0.95,2,178)

3.046721

#since our F-ratio = 2.359954< 3.04672

Thus we can’t reject the Ho

#Thus, we have p-value =0.09735 > alpha =.05 we can’t reject Ho

#So we have the same result

#and believe EY2 and EYRSCH are not jointly significant

h)

Originally,

LE = LE2013\*EYRSCH

model5 = lm(HDI ~ LE2013 + MEANYRSCH + EYRSCH + GNI2013 + HDI2012 +CHINRANK + DL+ LE)

summary(model5)

Call:

lm(formula = HDI ~ LE2013 + MEANYRSCH + EYRSCH + GNI2013 + HDI2012 +

CHINRANK + DL + LE)

Residuals:

Min 1Q Median 3Q Max

-0.0239673 -0.0006992 0.0000954 0.0009770 0.0055156

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.555e-02 8.129e-03 -3.143 0.00196 \*\*

LE2013 3.658e-04 1.214e-04 3.012 0.00298 \*\*

MEANYRSCH -2.181e-05 1.591e-04 -0.137 0.89112

EYRSCH 2.211e-03 6.792e-04 3.255 0.00136 \*\*

GNI2013 9.911e-09 1.864e-08 0.532 0.59548

HDI2012 9.990e-01 8.228e-03 121.420 < 2e-16 \*\*\*

CHINRANK 9.530e-04 1.480e-04 6.438 1.11e-09 \*\*\*

DLlow 2.107e-03 1.254e-03 1.680 0.09475 .

DLmedium 1.506e-03 6.599e-04 2.283 0.02364 \*

DLvery high -3.764e-04 8.108e-04 -0.464 0.64304

LE -2.887e-05 9.463e-06 -3.051 0.00263 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002236 on 176 degrees of freedom

Multiple R-squared: 0.9998, Adjusted R-squared: 0.9998

F-statistic: 9.062e+04 on 10 and 176 DF, p-value: < 2.2e-16

After the log:

lnHDI= log(HDI)

model8 = lm(lnHDI ~ LE2013 + MEANYRSCH + EYRSCH + GNI2013 + HDI2012 +CHINRANK + DL+LE)

summary(model8)

Call:

lm(formula = lnHDI ~ LE2013 + MEANYRSCH + EYRSCH + GNI2013 +

HDI2012 + CHINRANK + DL + LE)

Residuals:

Min 1Q Median 3Q Max

-0.067947 -0.007862 0.000455 0.008586 0.048747

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.331e+00 5.205e-02 -44.787 < 2e-16 \*\*\*

LE2013 9.431e-03 7.776e-04 12.129 < 2e-16 \*\*\*

MEANYRSCH -5.219e-03 1.019e-03 -5.124 7.82e-07 \*\*\*

EYRSCH 5.927e-02 4.349e-03 13.629 < 2e-16 \*\*\*

GNI2013 -7.319e-07 1.193e-07 -6.133 5.52e-09 \*\*\*

HDI2012 1.998e+00 5.268e-02 37.924 < 2e-16 \*\*\*

CHINRANK 1.171e-03 9.478e-04 1.235 0.218

DLlow 3.447e-02 8.033e-03 4.291 2.93e-05 \*\*\*

DLmedium 2.980e-02 4.225e-03 7.052 3.88e-11 \*\*\*

DLvery high -2.118e-02 5.192e-03 -4.080 6.83e-05 \*\*\*

LE -8.907e-04 6.059e-05 -14.700 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01432 on 176 degrees of freedom

Multiple R-squared: 0.9969, Adjusted R-squared: 0.9968

F-statistic: 5744 on 10 and 176 DF, p-value: < 2.2e-16

#p-value < 2e-16 for EY2 is smaller than alpha = .05,other variables are getting better

# For the HDI model: Multiple R-squared: 0.9998, Adjusted R-squared: 0.9998

# For the logHDI model: Multiple R-squared: 0.9969, Adjusted R-squared: 0.9968

# Via observation, we see HDI is originally very high in R2 and changes little in adjusted R2

#Also, logHDI model decreases a bit from the original model, yet it deceases very insignificantly by

#.0001. Thus,overall, logY has improved the result

> log(.92)

[1] -0.08338161

newdata = data.frame(lnHDI = -0.08338161, LE2013 = 80, MEANYRSCH = 12 , EYRSCH = 17,GNI2013 = 52000, HDI2012=-.08255, CHINRANK =-2, DL= “very high” , LE= 1280)

newdata = data.frame(LE2013 = 80, MEANYRSCH = 12 , EYRSCH = 17,GNI2013 = 52000, HDI2012=-.08255, CHINRANK =-2, DL= "high",LE= 1280)

predict(model8, newdata, interval="confidence", level=.95)

fit lwr upr

1 -1.977114 -2.080657 -1.87357

#the predicted y (under the log) is -1.977114

3)

x=-1.977114

y = exp(x)

y

[1] 0.1384683

> x = -2.080657

> y = exp(x)

> y

[1] 0.1248482

> x = -1.87357

> y = exp(x)

> y

[1] 0.1535744

If a nation has all of the qualities listed above, we are 95% confident that they have an average predicted interval between .124848 and .1535744