term <- read.csv(choose.files(), header=TRUE)

#Data file is "termlife"

term1 = subset(term,subset=FACE>0)

attach(term1)

hist(FACE)

logFACE=log(FACE)

hist(INCOME)

logINCOME=log(INCOME)

dim(term1)

[1] 275 18

hist(EDUCATION)

hist(NUMHH)

Vars = data.frame(FACE,INCOME,EDUCATION,NUMHH,logFACE,logINCOME)

# There are were 18 variables in the original dataset,this will

#take a subset of only a few of them

summary(Vars)

FACE INCOME EDUCATION NUMHH

Min. : 800 Min. : 260 Min. : 2.00 Min. :1.00

1st Qu.: 50000 1st Qu.: 38000 1st Qu.:13.00 1st Qu.:2.00

Median : 150000 Median : 65000 Median :16.00 Median :3.00

Mean : 747581 Mean : 208975 Mean :14.52 Mean :2.96

3rd Qu.: 590000 3rd Qu.: 120000 3rd Qu.:16.50 3rd Qu.:4.00

Max. :14000000 Max. :10000000 Max. :17.00 Max. :9.00

logFACE logINCOME

Min. : 6.685 Min. : 5.561

1st Qu.:10.820 1st Qu.:10.545

Median :11.918 Median :11.082

Mean :11.990 Mean :11.149

3rd Qu.:13.288 3rd Qu.:11.695

Max. :16.455 Max. :16.118

par(mfrow=c(1,2))

plot(INCOME,FACE)

plot(log(INCOME),logFACE)

Varb = data.frame(logINCOME,NUMHH,EDUCATION,logFACE)

pairs(Varb, upper.panel=NULL)

#List the Y variable last in the data.frame command to get it in the bottom

#row as the y axis to every other variable's x axis

#helps us see unusual points, nonlinear patterns,etc.quickly

cor(cbind(logFACE,logINCOME,NUMHH,EDUCATION))

logFACE logINCOME NUMHH EDUCATION

logFACE 1.0000000 0.4818427 0.2876115 0.3828489

logINCOME 0.4818427 1.0000000 0.1793354 0.3427036

NUMHH 0.2876115 0.1793354 1.0000000 -0.0635292

EDUCATION 0.3828489 0.3427036 -0.0635292 1.0000000

model1 = lm(logFACE ~ logINCOME + EDUCATION + NUMHH)

summary(model1)

Call:

lm(formula = logFACE ~ logINCOME + EDUCATION + NUMHH)

Residuals:

Min 1Q Median 3Q Max

-5.7420 -0.8681 0.0549 0.9093 4.7187

Coefficients:

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

(Intercept) 2.58408 0.84643 3.053 0.00249 \*\*

logINCOME 0.49353 0.07754 6.365 8.32e-10 \*\*\*

EDUCATION 0.20641 0.03883 5.316 2.22e-07 \*\*\*

NUMHH 0.30605 0.06333 4.833 2.26e-06 \*\*\*

---

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

Residual standard error: 1.525 on 271 degrees of freedom

Multiple R-squared: 0.3425, Adjusted R-squared: 0.3353

F-statistic: 47.07 on 3 and 271 DF, p-value: < 2.2e-16

**# ALL individual X p-value are small.Since, for example,NUMHH has a**

**# p-value = 2.26e-06 < alpha = .05,so NUMHH is a significant predictor of**

**# logFACE**

**#R-SQR are R- sQR adjusted are close,which is good ,but low which is bad**

**# Then being close means that the X's we have included are providing us**

**# with enough benefit to outweigh the cost of including them in the model**

#**then being low means that there is a lot of variation in Y that we are not explaining**

#so **there are other X variables that we should also be including** that we are not

#Interpret R2a: we explaining 33.53%of the variation in logFACE with the 3X’s we

#are using once we have taken the penalty for the # of Xs in the model

**#Interpret the slope for NUMHH: for every 1 person added to a household,**

**# they will purchase an additional .306 of log dollors to the FACE value of**

**# their policy,holding the other X variable constant.**

refrig <- read.csv(choose.files(), header=TRUE)

#data file "Refrigerator"

attach(refrig)

dim(refrig)

[1] 37 8

Vars =data.frame(PRICE,ECOST,RSIZE,FSIZE,SHELVES,FEATURES)

summary(Vars)

PRICE ECOST RSIZE FSIZE

Min. : 460.0 Min. :60.00 Min. :12.6 Min. :4.100

1st Qu.: 545.0 1st Qu.:66.00 1st Qu.:12.9 1st Qu.:4.400

Median : 590.0 Median :68.00 Median :13.2 Median :5.100

Mean : 626.4 Mean :70.51 Mean :13.4 Mean :5.184

3rd Qu.: 685.0 3rd Qu.:75.00 3rd Qu.:13.9 3rd Qu.:5.700

Max. :1200.0 Max. :94.00 Max. :14.7 Max. :7.400

SHELVES FEATURES

Min. :1.000 Min. : 1.000

1st Qu.:2.000 1st Qu.: 2.000

Median :2.000 Median : 3.000

Mean :2.514 Mean : 3.459

3rd Qu.:3.000 3rd Qu.: 5.000

Max. :5.000 Max. :12.000

> cor(cbind(PRICE,ECOST,RSIZE,FSIZE,SHELVES,FEATURES))

PRICE ECOST RSIZE FSIZE SHELVES FEATURES

PRICE 1.0000000 0.52221565 -0.02399240 0.7200542 0.3995884 0.69743175

ECOST 0.5222156 1.00000000 -0.03340599 0.8548806 0.1877225 0.33423817

RSIZE -0.0239924 -0.03340599 1.00000000 -0.2346965 -0.3631354 -0.09575734

FSIZE 0.7200542 0.85488064 -0.23469650 1.0000000 0.2511044 0.43930765

SHELVES 0.3995884 0.18772247 -0.36313539 0.2511044 1.0000000 0.16048117

FEATURES 0.6974317 0.33423817 -0.09575734 0.4393077 0.1604812 1.00000000

> cor(Vars)

PRICE ECOST RSIZE FSIZE SHELVES FEATURES

PRICE 1.0000000 0.52221565 -0.02399240 0.7200542 0.3995884 0.69743175

ECOST 0.5222156 1.00000000 -0.03340599 0.8548806 0.1877225 0.33423817

RSIZE -0.0239924 -0.03340599 1.00000000 -0.2346965 -0.3631354 -0.09575734

FSIZE 0.7200542 0.85488064 -0.23469650 1.0000000 0.2511044 0.43930765

SHELVES 0.3995884 0.18772247 -0.36313539 0.2511044 1.0000000 0.16048117

FEATURES 0.6974317 0.33423817 -0.09575734 0.4393077 0.1604812 1.00000000

> model =lm(PRICE ~ECOST + RSIZE + FSIZE + SHELVES + FEATURES)

> summary(model1)

Call:

lm(formula = logFACE ~ logINCOME + EDUCATION + NUMHH)

Residuals:

Min 1Q Median 3Q Max

-5.7420 -0.8681 0.0549 0.9093 4.7187

Coefficients:

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

(Intercept) 2.58408 0.84643 3.053 0.00249 \*\*

logINCOME 0.49353 0.07754 6.365 8.32e-10 \*\*\*

EDUCATION 0.20641 0.03883 5.316 2.22e-07 \*\*\*

NUMHH 0.30605 0.06333 4.833 2.26e-06 \*\*\*

---

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

Residual standard error: 1.525 on 271 degrees of freedom

Multiple R-squared: 0.3425, Adjusted R-squared: 0.3353

F-statistic: 47.07 on 3 and 271 DF, p-value: < 2.2e-16

# Notice that the ECOST slope is neg and significant but its conrrelation

# is a decent medium strength positive relationship

# The diff is because correlation pretends tha these are the only 2 #

#VARIABLES that exist.

#Regression takes the effect of the other X's into account