Econ4274\_problem\_set\_2

Man Adrian Ho Hin

2023-03-14

rm(list=ls())  
setwd("/Users/adrian/Desktop/econ4274\_problem\_set\_2")

# Question 1a

library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

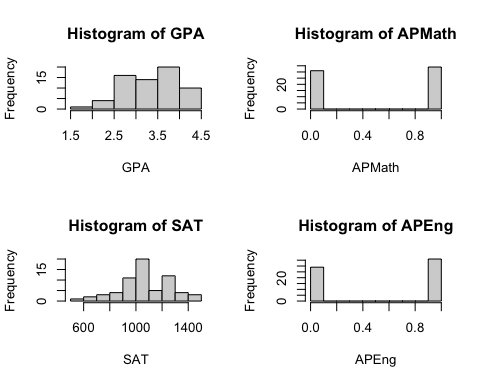
sat = read.csv("SAT.csv")  
stargazer(sat, type="html",out="sat.html")

Summary Statistic for SAT

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Obs | Mean | St. Dev. | Min | Max |
| AP | 65 | 0.677 | 0.471 | 0 | 1 |
| APEng | 65 | 0.554 | 0.501 | 0 | 1 |
| APMath | 65 | 0.523 | 0.503 | 0 | 1 |
| ESL | 65 | 0.4 | 0.494 | 0 | 1 |
| gender | 65 | 0.492 | 0.504 | 0 | 1 |
| GPA | 65 | 3.362 | 0.613 | 1.64 | 4.38 |
| prep | 65 | 0.738 | 0.443 | 0 | 1 |
| race | 65 | 0.323 | 0.471 | 0 | 1 |
| SAT | 65 | 1,075.54 | 191.361 | 590 | 1,430 |

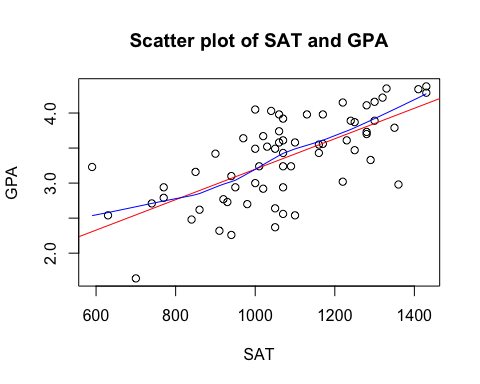
# Question 1b

par(mfcol = c(2, 2))  
with(sat, hist(GPA))  
with(sat, hist(SAT))  
with(sat, hist(APMath))  
with(sat, hist(APEng))



# Question 1c

with(sat, plot(SAT, GPA, main = "Scatter plot of SAT and GPA"))  
abline(lm(GPA~SAT, data = sat), col = "red")  
lines(lowess(sat$SAT, sat$GPA), col="blue")



# Question 1d

# i  
x\_1 = rep(1,length(sat$SAT))  
  
# ii   
X = with(sat, matrix(c(x\_1,SAT,APMath,APEng,ESL,gender,race),nrow=length(SAT)))  
  
# iii  
y = sat$GPA  
  
# iv  
beta\_hat = solve(t(X) %\*% X) %\*% t(X) %\*% y  
beta\_hat = t(beta\_hat)  
colnames(beta\_hat)=c("Intercept","SAT","APMath","APEng","ESL","gender","race")  
  
# v   
lm1 = with(sat, lm(GPA~SAT+APMath+APEng+ESL+gender+race))  
cat("By matrix algebra: \n")

## By matrix algebra:

print(beta\_hat)

## Intercept SAT APMath APEng ESL gender  
## [1,] 1.66719 0.001319036 0.1408304 0.5032422 -0.07609953 -0.1477924  
## race  
## [1,] 0.08437138

cat("\n")

cat("By lm() function: \n")

## By lm() function:

print(lm1$coefficients)

## (Intercept) SAT APMath APEng ESL gender   
## 1.667190138 0.001319036 0.140830421 0.503242246 -0.076099529 -0.147792393   
## race   
## 0.084371380

# Question 2

rm(list=ls())  
drug\_price = read.csv("drug\_price.csv")  
  
# Base Model  
lm1 = with(drug\_price, lm(p.r~GDP.r))  
# Model with consumption related factors  
lm2 = with(drug\_price, lm(p.r~GDP.r+cv+cv.r))  
# Model with pricing related factors  
lm3 = with(drug\_price, lm(p.r~GDP.r+p.control+p.comp))  
# Model with legal restriction factors  
lm4 = with(drug\_price, lm(p.r~GDP.r+patent))  
# Full Model  
lm5 = with(drug\_price, lm(p.r~GDP.r+cv+cv.r+p.control+p.comp+patent))  
  
library(stargazer)  
stargazer(lm1,lm2,lm3,lm4,lm5,type="html",out = "models1.html")

Regression Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Dependent variable: p.r | | | | |
|  | (1) | (2) | (3) | (4) | (5) |
| GDP.r | 0.923\*\*\* | 1.650\*\*\* | 0.974\*\*\* | 0.871\*\*\* | 1.407\*\*\* |
|  | (0.14) | (0.25) | (0.12) | (0.14) | (0.22) |
| cv |  | (0.11) |  |  | (0.20) |
|  |  | (0.25) |  |  | (0.23) |
| cv.r |  | -0.822\*\*\* |  |  | -0.483\* |
|  |  | (0.28) |  |  | (0.26) |
| p.control |  |  | -21.518\*\*\* |  | -17.832\*\* |
|  |  |  | (7.40) |  | (7.43) |
| p.comp |  |  | -14.386\* |  | (9.09) |
|  |  |  | (7.87) |  | (7.68) |
| patent |  |  |  | 13.944\* | 7.66 |
|  |  |  |  | (8.04) | (6.17) |
| Observations | 32.00 | 32.00 | 32.00 | 32.00 | 32.00 |
| R2 | 0.58 | 0.71 | 0.75 | 0.62 | 0.82 |
| Adjusted R2 | 0.56 | 0.68 | 0.72 | 0.59 | 0.77 |
| Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | | | | |

# The GDP.r is highly significant among all the models and its coefficient values having some level of variations, therefore we may compare the adjusted r^2 for the model improvement. The full model has the highest r^2 and the model with pricing factor has the 2nd highest r^2

# Model with pricing related factors  
lm5 = with(drug\_price, lm(p.r~GDP.r+p.control+p.comp))  
# Model with pricing related factors & consumption related factors  
lm6 = with(drug\_price, lm(p.r~GDP.r+p.control+p.comp+cv+cv.r))  
# Model with pricing related factors & legal restriction factors  
lm7 = with(drug\_price, lm(p.r~GDP.r+p.control+p.comp+patent))  
# Full Model  
lm8 = with(drug\_price, lm(p.r~GDP.r+p.control+p.comp+cv+cv.r+patent))  
  
stargazer(lm5,lm6,lm7,lm8,type="html",out = "models2.html")

Regression Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Dependent variable: p.r | | | |
|  | (1) | (2) | (3) | (4) |
| GDP.r | 0.974\*\*\* | 1.451\*\*\* | 0.943\*\*\* | 1.407\*\*\* |
|  | (0.12) | (0.22) | (0.12) | (0.22) |
| p.control | -21.518\*\*\* | -18.343\*\* | -20.626\*\*\* | -17.832\*\* |
|  | (7.40) | (7.49) | (7.38) | (7.43) |
| p.comp | -14.386\* | (10.61) | (12.97) | (9.09) |
|  | (7.87) | (7.66) | (7.89) | (7.68) |
| cv |  | (0.18) |  | (0.20) |
|  |  | (0.23) |  | (0.23) |
| cv.r |  | -0.508\* |  | -0.483\* |
|  |  | (0.26) |  | (0.26) |
| patent |  |  | 8.11 | 7.66 |
|  |  |  | (6.77) | (6.17) |
| Observations | 32.00 | 32.00 | 32.00 | 32.00 |
| R2 | 0.75 | 0.81 | 0.76 | 0.82 |
| Adjusted R2 | 0.72 | 0.77 | 0.72 | 0.77 |
| Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | | | |

# Again, the GDP.r is highly significant among all the models. The best model seems to be either lm6 or the full model, although they has higher standard error in GDP.r coefficient, they has the highest 2 r^2 value.  
  
# Use anova to verify the result  
anova(lm6,lm8)

## Analysis of Variance Table  
##   
## Model 1: p.r ~ GDP.r + p.control + p.comp + cv + cv.r  
## Model 2: p.r ~ GDP.r + p.control + p.comp + cv + cv.r + patent  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 26 7288.6   
## 2 25 6864.8 1 423.8 1.5434 0.2256

# So anova suggest that lm6 is the best model as not rejecting the F-test, which implies that including the patent variable cannot create significant difference.  
  
# Compare to the remaining model  
modelname = c("lm1","lm2","lm3","lm4","lm7")  
  
for(i in modelname){  
 cat("\n", i, "vs lm6:\n")  
 print(anova(get(i), lm6))  
}

##   
## lm1 vs lm6:  
## Analysis of Variance Table  
##   
## Model 1: p.r ~ GDP.r  
## Model 2: p.r ~ GDP.r + p.control + p.comp + cv + cv.r  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 30 15831.9   
## 2 26 7288.6 4 8543.3 7.619 0.0003345 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## lm2 vs lm6:  
## Analysis of Variance Table  
##   
## Model 1: p.r ~ GDP.r + cv + cv.r  
## Model 2: p.r ~ GDP.r + p.control + p.comp + cv + cv.r  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 28 10998.1   
## 2 26 7288.6 2 3709.5 6.6163 0.004756 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## lm3 vs lm6:  
## Analysis of Variance Table  
##   
## Model 1: p.r ~ GDP.r + p.control + p.comp  
## Model 2: p.r ~ GDP.r + p.control + p.comp + cv + cv.r  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 28 9461.2   
## 2 26 7288.6 2 2172.6 3.875 0.03366 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## lm4 vs lm6:  
## Analysis of Variance Table  
##   
## Model 1: p.r ~ GDP.r + patent  
## Model 2: p.r ~ GDP.r + p.control + p.comp + cv + cv.r  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 29 14343.3   
## 2 26 7288.6 3 7054.7 8.3885 0.0004579 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## lm7 vs lm6:  
## Analysis of Variance Table  
##   
## Model 1: p.r ~ GDP.r + p.control + p.comp + patent  
## Model 2: p.r ~ GDP.r + p.control + p.comp + cv + cv.r  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 27 8983.8   
## 2 26 7288.6 1 1695.2 6.0473 0.0209 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Which support our claim which lm6 is the best model as rejecting all the F-test above.  
  
par(mfcol = c(2, 2))  
plot(p.r~GDP.r,data=drug\_price,pch=20,main="Base Model")  
points(drug\_price$GDP.r,lm1$fitted.values,pch=17,col=2)  
  
plot(p.r~GDP.r,data=drug\_price,pch=20,main="Full Model")  
points(drug\_price$GDP.r,lm8$fitted.values,pch=17,col=2)  
  
plot(p.r~GDP.r,data=drug\_price,pch=20,main="lm2")  
points(drug\_price$GDP.r,lm2$fitted.values,pch=17,col=2)  
  
plot(p.r~GDP.r,data=drug\_price,pch=20,main="lm6 (Best Model)")  
points(drug\_price$GDP.r,lm6$fitted.values,pch=17,col=2)

Chart, scatter chart

Description automatically generated

# Question 3a

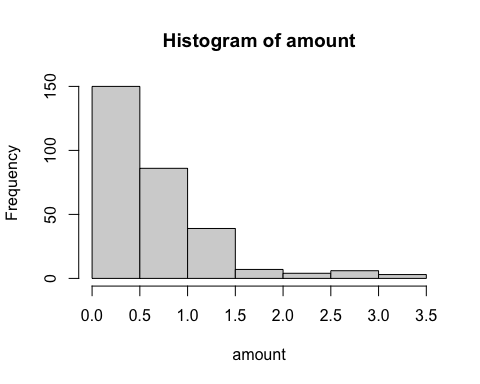
rm(list=ls())  
library(stargazer)  
tsssu = read.csv("TSSSU.csv")  
  
stargazer(tsssu,type="html",out="tsssu.html")

Summary Statistic for TSSSU

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | N | Mean | St. Dev. | Min | Max |
| amount | 295 | 0.708 | 0.599 | 0.1 | 3.28 |
| survive | 295 | 0.844 | 0.363 | 0 | 1 |
| social\_media | 294 | 0.33 | 0.471 | 0 | 1 |
| phone\_call | 294 | 0.156 | 0.364 | 0 | 1 |
| No\_members | 295 | 3.763 | 1.991 | 0 | 15 |
| No\_alumni | 295 | 1.4 | 1.244 | 0 | 8 |
| No\_professor | 295 | 0.708 | 0.776 | 0 | 5 |
| No\_undergrad | 295 | 0.217 | 0.66 | 0 | 4 |
| No\_postgrad | 295 | 0.519 | 1.026 | 0 | 10 |
| No\_other | 295 | 0.919 | 1.424 | 0 | 11 |
| Employee | 88 | 6.034 | 6.934 | 1 | 51 |
| private\_fund | 295 | 0.125 | 0.332 | 0 | 1 |
| Sciencepark | 295 | 0.363 | 0.482 | 0 | 1 |

# Question 3b

with(tsssu, hist(amount))



with(tsssu, plot(Employee, amount))

# Chart, scatter chart Description automatically generated

# Question 3c

library(stargazer)  
  
tsssu$f.university=as.factor(tsssu$university)  
tsssu$f.year=as.factor(tsssu$year)  
tsssu$f.area=as.factor(tsssu$area)  
  
m1 = lm(survive~amount+No\_undergrad+No\_postgrad+No\_professor+f.university+f.year+f.area,data = tsssu)  
m2 = lm(Employee~amount+No\_undergrad+No\_postgrad+No\_professor+f.university+f.year+f.area,data = tsssu)  
m3 = lm(social\_media~amount+No\_undergrad+No\_postgrad+No\_professor+f.university+f.year+f.area,data = tsssu)  
m4 = lm(phone\_call~amount+No\_undergrad+No\_postgrad+No\_professor+f.university+f.year+f.area,data = tsssu)  
m5 = lm(private\_fund~amount+No\_undergrad+No\_postgrad+No\_professor+f.university+f.year+f.area,data = tsssu)  
m6 = lm(Sciencepark~amount+No\_undergrad+No\_postgrad+No\_professor+f.university+f.year+f.area,data = tsssu)  
  
# Further modify in the excel  
stargazer(m1,m2,m3,m4,m5,m6,type="html",out="regtab.html")

Regression Table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Dependent variable: | | | | | |
|  | survive | Employee | social\_media | phone\_call | private\_fund | Sciencepark |
|  | (1) | (2) | (3) | (4) | (5) | (6) |
| amount | 0.081\* | 2.74 | 0.06 | 0.07 | 0.06 | 0.163\*\*\* |
|  | (0.05) | (1.89) | (0.06) | (0.05) | (0.05) | (0.06) |
| No\_undergrad | 0.01 | (0.80) | (0.01) | 0.03 | (0.02) | 0.05 |
|  | (0.03) | (1.42) | (0.05) | (0.04) | (0.03) | (0.04) |
| No\_postgrad | (0.01) | (0.87) | (0.02) | 0.00 | 0.01 | 0.04 |
|  | (0.02) | (1.12) | (0.03) | (0.02) | (0.02) | (0.03) |
| No\_professor | (0.02) | 0.47 | (0.01) | (0.04) | -0.088\*\*\* | 0.01 |
|  | (0.03) | (1.29) | (0.04) | (0.03) | (0.03) | (0.04) |
| f.universityCUHK | 0.185\*\* | 0.21 | 0.213\*\* | 0.07 | 0.04 | 0.164\* |
|  | (0.07) | (3.60) | (0.10) | (0.08) | (0.07) | (0.09) |
| f.universityHKBU | 0.02 | (2.12) | 0.19 | 0.03 | 0.05 | (0.21) |
|  | (0.12) | (4.50) | (0.16) | (0.12) | (0.11) | (0.14) |
| f.universityHKU | 0.293\*\*\* | 3.01 | 0.03 | 0.06 | 0.09 | 0.145\* |
|  | (0.07) | (3.19) | (0.10) | (0.08) | (0.07) | (0.09) |
| f.universityPoly | 0.178\*\* | 0.71 | 0.04 | 0.06 | 0.02 | (0.07) |
|  | (0.08) | (3.34) | (0.10) | (0.08) | (0.07) | (0.09) |
| f.universityUST | 0.151\*\* | 1.37 | (0.04) | (0.06) | (0.02) | 0.08 |
|  | (0.07) | (3.15) | (0.10) | (0.08) | (0.07) | (0.09) |
| Observations | 295.00 | 88.00 | 294.00 | 294.00 | 295.00 | 295.00 |
| R2 | 0.14 | 0.25 | 0.12 | 0.07 | 0.10 | 0.28 |
| Adjusted R2 | 0.05 | (0.02) | 0.03 | (0.02) | 0.01 | 0.21 |
| Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | | | | | |

# Question 3d

area = aggregate(amount~area, data=tsssu, sum)  
area = area[order(area$amount, decreasing = FALSE),]  
  
png("barchart.png",width=900,height=500)  
par(mar=c(3,20,1,1))   
barplot(area$amount,names.arg = area$area,   
 las=1,  
 horiz=TRUE)  
dev.off()

## quartz\_off\_screen   
## 2

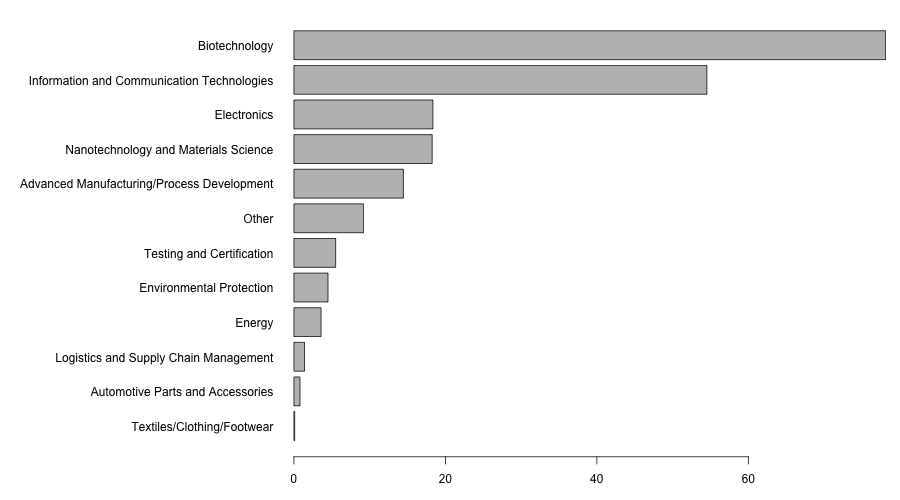


Figure 1 TSSSU Funding Distribution Acorss Areas