HW3-proposed solutions

Problem 1

Refer to the CDI data set. Using R2 as the criterion, which predictor variable accounts for the largest reduction in the variability in the number of active physicians?

Solution:

```
library(knitr)
CDI <- read.csv("CDI.csv")
Temp<-CDI[,4:17]
X<- Temp[,-c(5)]
Y<- Temp[,5]

prg1 <-function(X,Y){
out<-matrix(0,nrow=13,ncol=1)
for (i in 1:13){
f<-lm(Y~X[,i])
f1<-summary(f)
out[i,1]<- f1$r.squared
}
data.frame(names(X) ,out)}

kable(prg1(X,Y))
```

names.X.	out
Land.area	0.0060957
Total.population	0.8840674
Percent.of.population.aged.18.34	0.0143279
Percent.of.population.65.or.older	0.0000098
Number.of.hospital.beds	0.9033826
Total.serious.crimes	0.6731538
Percent.high.school.graduates	0.0000180
Percent.bachelor.s.degrees	0.0560579
Percent.below.poverty.level	0.0041135
Percent.unemployment	0.0025519
Per.capita.income	0.0999411
Total.personal.income	0.8989137
Geographic.region	0.0006074

```
names(X[which.max(prg1(X,Y)[,2])])
```

```
## [1] "Number.of.hospital.beds"
```

Number of Hospital beds has the highest R Square.

Problem 2

Refer to the CDI data set in Appendix C.2 and Project 1.44. Obtain a separate interval estimate of β_1, for each region. Use a 90 percent confidence coefficient in each case. Do the regression lines for the different regions appear to have similar slopes?

Solution:

They re different, the confidence intervals do not overlap.

```r
f1<-lm(Per.capita.income~Percent.bachelor.s.degrees,data= CDI[CDI[,17]==1,])
f2<-lm(Per.capita.income~Percent.bachelor.s.degrees,data= CDI[CDI[,17]==2,])
f3<-lm(Per.capita.income~Percent.bachelor.s.degrees,data= CDI[CDI[,17]==3,])
f4<-lm(Per.capita.income~Percent.bachelor.s.degrees,data= CDI[CDI[,17]==4,])

confint(f1,level=0.9)
```

```
##                                  5 %     95 %
## (Intercept)                7809.8077 10637.82
## Percent.bachelor.s.degrees  460.5177   583.80
```

```r
confint(f2,level=0.9)
```

```
##                                   5 %      95 %
## (Intercept)                12627.0363 14535.774
## Percent.bachelor.s.degrees   193.4858   283.853
```

```r
confint(f3,level=0.9)
```

```
##                                  5 %       95 %
## (Intercept)                9516.0773 11543.4929
## Percent.bachelor.s.degrees  285.7076   375.5158
```

```r
confint(f4,level=0.9)
```

```
##                                  5 %       95 %
## (Intercept)                6862.6967 10367.4086
## Percent.bachelor.s.degrees  364.7585   515.8729
```

Problem 3

Refer to GPA data:

a) Set up the ANOVA table.

b) b) What is estimated by MSR in your ANOVA table? by MSE? Under what condition do MSR and MSE estimate the same quantity?

c) Conduct an F test of whether or not $\beta_1 = 0$. Control the α risk at .01. State the alternatives, decision rule, and conclusion.

d) What is the absolute magnitude of the reduction in the variation of Y when X is introduced into the regression model? What is the relative reduction? What is the name of the latter measure?

e) Obtain r and attach the appropriate sign.

f) Which measure, R^2 or r, has the more clear-cut operational interpretation? Explain.

Solution:

a)

```
GPA <- read.csv("GPA.csv")
f<-lm(GPA~ACT,data=GPA)
anova(f,test="Chi")
```

```
## Analysis of Variance Table
##
## Response: GPA
##            Df Sum Sq Mean Sq F value   Pr(>F)
## ACT         1  3.588  3.5878  9.2402 0.002917 **
## Residuals 118 45.818  0.3883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b)

$$\sigma^2 + \beta_1^2 \sum \left(X_i - \hat{X} \right), \sigma^2, \text{ when } \beta_1 = 0$$

c)

From the ANOVA table above, Fstat=9.2402 and P Value= 0.002917. Reject Null and accept Ha. β_1 is significant.

d)

SSR = 3.588, 3.588/(3.588+45.818) = 7.26% or 0.0726, coefficient of determination

e)

Sign is positive since the slope is positive.

```
sqrt(0.0726)
```

```
## [1] 0.2694439
```

f)

R^2

Problem 4

Refer to Crime rate data.

a) Compute the Pearson product-moment correlation coefficient r_{12}.

b) Test whether crime rate and percentage of high school graduates are statistically independent in the population; use a $\alpha = .01$. State the alternatives, decision rule, and conclusion.

c) Compute the Spearman rank correlation coefficient rs.

d) Test by means of the Spearman rank correlation coefficient whether an association exists between crime rate and percentage of high school graduates. State the alternatives, decision rule, and conclusion.

Solution:

a)

```
Crime.Rate <- read.csv("Crime Rate.csv")

cor.test(Crime.Rate$X,Crime.Rate$Y, method ="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  Crime.Rate$X and Crime.Rate$Y
## t = -4.1029, df = 82, p-value = 9.571e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5761223 -0.2175580
## sample estimates:
##        cor
## -0.4127033
```

r12=-0.4127033

b)

$$H_0 : \rho = 0$$

$$H_A : \rho \neq 0$$

From above, the p-value = 9.571e-05, reject null, ρ is significant

c)

```r
cor.test(Crime.Rate$X,Crime.Rate$Y, method ="spearman")
```

```
## Warning in cor.test.default(Crime.Rate$X, Crime.Rate$Y, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  Crime.Rate$X and Crime.Rate$Y
## S = 140839, p-value = 5.359e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.4259324
```

rs=-0.4259324

d)

$$H_0 : \rho = 0$$

$$H_A : \rho \neq 0$$

From above, the p-value = 5.359e-05, reject null, ρ is significant