ACST 890 QUIZ 1

R program

```
#GitHub username:HualingzhuLong
#repositories name:ACST_s1_2019
#file name:ACST890 QUIZ1.R & ACST890 QUIZ1.PDF

#question 1
#c=coupon,n=number of coupon payment,t= time until coupon paid(half year as a period),F= face
value,P=price
#The yield of interest is effective semi-annually interest rate respect with tj
#assume the all the value of variables are known except p
p<-function(c,n,F,Yields){
  t=seq(1,n,by=1)
  y<-c(Yields)
  price<-0
  rec<- list(t,y)
  rec
  price=sum(c*exp(-y*t))+F*exp(-rec[[2]][n]*rec[[1]][n])
return(price)
}
#using assuming value of data
answer=p(100,10,1000,c(0.3,0.1,0.2,0.2,0.1,0.05,0.3,0.1,0.15,0.25))
answer

#question 3
#(a)
dataset<- read.csv(file.choose("singapore.economy.csv"), header=T)
dataset
#(b)
dataset1=na.omit(dataset)
dataset1
#(c)
attach(dataset1)
plot(time,gdp,xlab="Time",ylab="GDP(%)",main="Singapore GDP growth")
#(d)
GDP1=dataset1[period=="1","gdp"]
M1<- mean(GDP1)
M1
SD1<- sd(GDP1)
SD1
GDP2=dataset1[period=="2","gdp"]
M2<- mean(GDP2)
M2
SD2<- sd(GDP2)
SD2
GDP3=dataset1[period=="3","gdp"]
M3<-mean(GDP3)
M3
SD3<- sd(GDP3)
```

```
SD3
mean<- c(M1,M2,M3)
sd<- c(SD1,SD2,SD3)
period<- c(1,2,3)
stat.table<- data.frame(period,mean,sd)
stat.table
#(e)
pair<- pairs(dataset1[,3:10])
pair
#(f)
SLR.GDP=lm(gdp~exp)
SLR.GDP
summary(SLR.GDP)
#When exp=0, the gdp=1.19032%. An increase of 0.19076% of gdp is associated with the increase of 1 unit exp.
#As p value is small enough, we have strong evidence that we could reject null hypothesis which is beta1=0.
#(g)
MLR.GDP= lm(gdp~exp+epg+hpr+oil+gdpus+crd)
MLR.GDP
summary(MLR.GDP)
#The predictor of exp,epg abd hpr are in siginificant level 1%. The predictor oil,gdpus and crd shows insiginificant at level 1% as p-value is not samll enough.
#multiple R-squared provided that there are 0.372 proportion variability in gdp is explained by linear regression on prosictors.
#the F test show F>1, so that H1 is ture which is at least 1 betaj is non zero.
# the p-value shows that the linear regression function is siginificant as p is small enough.
#(h)
Quan<- quantile(gdp,0.05)
Quan
state<- ifelse(gdp<Quan,"crisis","normal")
state
econ.table<- data.frame(dataset1,state)
econ.table
trainingdata<- subset(econ.table, econ.table$period==1|econ.table$period==2)
trainingdata
testdata <- subset(econ.table, econ.table$period==3)
testdata
logisticstate<- glm(state~bci, data=trainingdata,family=binomial)
logisticstate
prediction<- predict(logisticstate,testdata,type="response")
glmpred <- rep("crisis", 38)
glmpred[prediction>0.5]="normal"
table(glmpred, testdata$state)
```

question 2

word count:420

Linear regression is an approach to predict response output by predictor variable input. According to the observed information, we have interest in relationship between variable input X and response output Y. Firstly, we assume there is a linear relationship between variable X and output Y.

*1.**Simple linear regression*** (quantitative response Y based on single predictor variable X)

$Y \approx \beta_0 + \beta_1 X + \varepsilon$ where β_0 is intercept and β_1 is slope; ε is N~ $(0,\sigma^2)$ random error term

- Estimate the coefficients (by minimizing the least squares $e_i = y_i - \hat{y}_i$)

  Based on $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$, we can obtain residual sum of squares (RSS)

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2 = \left(y_1 - \widehat{\beta_0} - \widehat{\beta_1} x_1\right)^2 + \left(y_2 - \widehat{\beta_0} - \widehat{\beta_1} x_2\right)^2 + \cdots (y_n - \widehat{\beta_0} - \widehat{\beta_1} x_n)^2$$

  Using least squares approach to minimize the RSS, we have $\widehat{\beta_0}$ and $\widehat{\beta_1}$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\ ,\ \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} \quad \text{where } \bar{y} = \frac{\sum_{i=1}^n y_i}{n}\ ,\ \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

- Accuracy of coefficient estimates

  More sample data sets we observed, $n \to \infty$, more accurate estimates we obtained. Calculating standard error of $\widehat{\beta_0}$ and $\widehat{\beta_1}$ to see how close it is from true value:

  $\mathrm{SE}\left(\widehat{\beta_0}^2\right) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]$, $\mathrm{SE}\left(\widehat{\beta_1}^2\right) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$ where $\sigma^2 = VAR(\epsilon)$

  Standard error is used to find confidence interval $\widehat{\beta_1} \pm 2 \times \mathrm{SE}(\widehat{\beta_1})$ and $\widehat{\beta_0} \pm 2 \times \mathrm{SE}(\widehat{\beta_0})$

  It is also used in hypothesis test (X&Y have relationship or not): $H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$. By evaluating t test $\mathrm{t} = \frac{\widehat{\beta_1} - 0}{\mathrm{SE}(\widehat{\beta_1})}$, when probability of observing number $\geq |t|$, $\beta_1 = 0$.

  This probability is called p-value: small p-value, reject H_0.

- Accuracy of model (how model fits the data) by RSE & R^2 statistic

  By evaluating residual standard error (RSE) (**measure of lack of fit**), we obtained

  $\mathrm{RSE} = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \widehat{y_i})^2}{n-2}}$ if $y_i \approx \widehat{y_i}$ for i=1, ..., n, RSE is small, the data fit model well.

  R^2 statistic is explained proportion of variance. $\mathrm{R}^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$ will lie in (0,1),

  where TSS is the total sum of squares $\sum_{i=1}^n (y_i - \overline{y_i})^2$. If it is close to 0, it means regression did not explain the variability of response. Hard to determine the good value.

  Can use r=Cor (X, Y) instead of R^2. $\mathrm{R}^2 = \mathrm{r}^2$ $\mathrm{Cor}(\mathrm{X},\mathrm{Y}) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$

*2.**Multiple linear regression*** (more than 1 predictors response to output) **interaction**

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$ with p distinct predictors and $\beta_1, \ldots, \beta_p$ associate X_j& Y

- Estimate the regression coefficient (minimize the sum of squared residuals)

  Use $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \cdots + \widehat{\beta_p} x_p$, $\mathrm{RSS} = \sum_{i=1}^n (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_{i1} - \widehat{\beta_2} x_{i2} - \cdots - \widehat{\beta_p} x_{ip})^2$

- Accuracy of coefficient estimates (using hypothesis)

  $H_0: \beta_1 = \beta_2 = \cdots = \beta_P = 0,\ H_1: at\ least\ one\ \beta_j\ is\ non\ zero$ Test by computing F-statistic $\mathrm{F} = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$. If H_0 is true, $\mathrm{E}\{RSS/(n-p-1)\} = \sigma^2\ and\ \mathrm{E}\{(TSS - RSS)/p\} = \sigma^2$. If H_1 is true, $\mathrm{E}\{(TSS - RSS)/p\} > \sigma^2$, so that F>1.

- Accuracy of model (3 classical approaches) 1. Forward selection (begin with H_0, find small p) 2. Backward selection (remove large p) 3. Mixed selection

- Model fit (Using RSE & R^2) $\mathrm{RSE} = \sqrt{\frac{RSS}{n-p-1}}$

- Prediction (3 uncertainty associated)

  1. $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \cdots + \widehat{\beta_p} x_p$ only estimate for true population regression $f(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$. 2. Model bias. 3. We cannot predict perfectly as irreducible error.