HOMEWORK 6 MATH 484-564, REGRESSION

DUE OCTOBER 11TH 2024, FRIDAY, 11:59PM. SEE THE SUBMISSION INSTRUCTIONS ON CANVAS.

(1) (6 points) This question should be answered using the Carseats data set as seen in Homework 4.

```r
library("ISLR2")

?Carseats
```

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

Please note that Urban and US are qualitative variables. When R encounters these types of variables, it recognizes them accordingly. However, for practice purposes, let's manually code these variables as shown below.

```r
library("ISLR2")

?Carseats

y<-Carseats$Sales
x1<-ifelse(Carseats$Urban=='Yes',1,0)
x2<-ifelse(Carseats$US=='Yes',1,0)
x3<-Carseats$Price

model1<-lm(y~x1+x2+x3)
summary(model1)
```

(b) From the summary, provide an interpretation of each coefficient in the model.

(c) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

(d) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

(e) How well do the model in (a) and (d) fit the data?

(2) (6 points) In this problem, we are going to explore the Swiss data set. See the information below:

```
> library(datasets); data(swiss)
> head(swiss)
             Fertility Agriculture Examination Education Catholic Infant.Mortality
Courtelary        80.2        17.0          15        12     9.96             22.2
Delemont          83.1        45.1           6         9    84.84             22.2
Franches-Mnt      92.5        39.7           5         5    93.40             20.2
Moutier           85.8        36.5          12         7    33.77             20.3
Neuveville        76.9        43.5          17        15     5.16             20.6
Porrentruy        76.1        35.3           9         7    90.57             26.6
```

The "Swiss" data has 47 observations on 6 variables:

(1) **Fertility**, I_g, using common standardized fertility measure

(2) **Agriculture**, % of males involved in agriculture as occupation

(3) **Examination**, % draftees receiving highest mark on army exam

(4) **Education**, % education beyond primary school for draftees.

(5) **Catholic**, % Catholic as opposed to Protestant

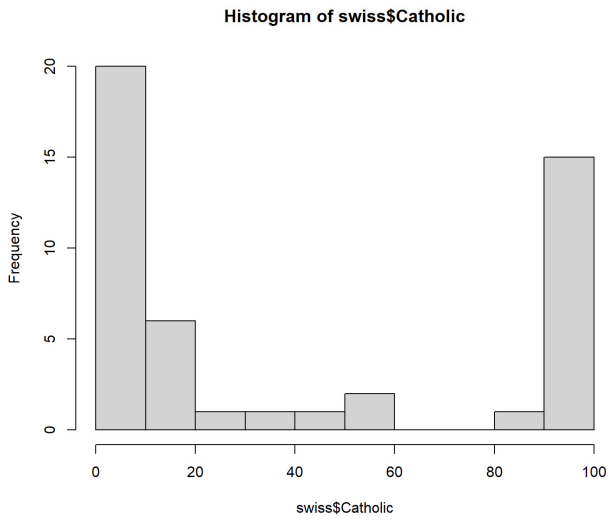
(6) **Infant Mortality**, % live births who live less than 1 year

The data collected are for 47 French-speaking provinces at about 1888.

I_g is equal to the total number of children born to married women divided by the maximum conceivable number of children, obtained from data on the Hutterites, an Anabaptist sect that does not practice any form of fertility limitations.

We are interested in "Agriculture", x_1, and "Catholic", x_2, as regressor variables and "Fertility" as an outcome variable, y.

Let us examine the data on "Catholic". Using hist(swiss$Catholic) we obtain



Because of the bimodel nature of "Catholic", we want to create an **indicator variable** x_2 as follows:

$$x_2 = \begin{cases} 1 & \text{if the province is over 50\% Catholic} \\ 0 & \text{otherwise} \end{cases}$$

We can approach this in a different way than what you've seen before. In this case, we will create a new column that assigns a value of 1 if the province has more than 50% Catholic population. We will do this in R using dplyr as shown below:

```
library(dplyr)
# Adding a column for 1 or 0 depending on Catholic %
swiss=mutate(swiss, CatholicBin=1*(Catholic>50))

head(swiss)
```

Notice the extra column known as CatholicBin.

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality	CatholicBin
Courtelary	80.2	17.0	15	12	9.96	22.2	0
Delemont	83.1	45.1	6	9	84.84	22.2	1
Franches-Mnt	92.5	39.7	5	5	93.40	20.2	1
Moutier	85.8	36.5	12	7	33.77	20.3	0
Neuveville	76.9	43.5	17	15	5.16	20.6	0
Porrentruy	76.1	35.3	9	7	90.57	26.6	1

Follow the steps outlined below to set up a regression model of y on x_1 and x_2. Note how x_2 is defined using the `factor` function in R. Additionally, you can limit the regression summary to only display the coefficients using the provided command.

```
x1<-swiss$Agriculture
y<-swiss$Fertility
plot(x1,y)


model1<-lm(y ~ x1 + factor(CatholicBin),data=swiss)
summary(model1)$coef
```

What can you conclude from this regression analysis?

3. (6 points) Let's revisit the Swiss data set from the previous problem. This time, we will define two indicator variables, x_2 and x_3, as follows:

$$x_2 = \begin{cases} 1 & \text{if the province is over 50\% Catholic} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if the province is over 50\% Protestant} \\ 0 & \text{otherwise} \end{cases}$$

Using these indicators, we can express our regression model as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

However, this straightforward approach of creating an indicator variable for each category of the qualitative predictor can lead to computational issues.

To illustrate, suppose we look at 4 observations from our data set, in which the first two were majority Catholic ($x_2 = 1, x_3 = 0$) and the second two being majority non-Catholic ($x_2 = 0, x_3 = 1$). Hence the matrix $\mathbf{X}$ would be

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & 1 & 0 \\ 1 & x_{21} & 1 & 0 \\ 1 & x_{31} & 0 & 1 \\ 1 & x_{41} & 0 & 1 \end{bmatrix}$$

Observe that the first column of $\mathbf{X}^T\mathbf{X}$ is the sum of the last two columns. Hence the columns of $\mathbf{X}^T\mathbf{X}$ are linearly dependent.

$$\begin{aligned} \mathbf{X}^T\mathbf{X} &= \begin{bmatrix} 1 & 1 & 1 & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & x_{11} & 1 & 0 \\ 1 & x_{21} & 1 & 0 \\ 1 & x_{31} & 0 & 1 \\ 1 & x_{41} & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 4 & \sum_{i=1}^4 x_{i1} & 2 & 2 \\ \sum_{i=1}^4 x_{i1} & \sum_{i=1}^4 x_{i1}^2 & \sum_{i=1}^2 x_{i1} & \sum_{i=3}^4 x_{i1} \\ 2 & \sum_{i=1}^2 x_{i1} & 2 & 0 \\ 2 & \sum_{i=3}^4 x_{i1} & 0 & 2 \end{bmatrix} \end{aligned}$$

What is the implication of this linear dependence for determining the coefficients β? (See Lecture Notes (Sept Version) Page 60).

Note: In general, we shall follow the principle:

A qualitative variable with c classes will be represented by $c - 1$ indicator variables, each taking the values 0 and 1.

4. (6 points) Tool wear is the gradual failure of cutting tools due to regular operation. Let us consider a regression model of tool wear y, on tool speed, x_1, and tool models A, B, C and D.

   Since we have a qualitative variable with four classes, using the principle mentioned at the end of Problem 3, we therefore require three indicator variables as follow:

$$x_2 = \begin{cases} 1 & \text{for tool model A} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{for tool model B} \\ 0 & \text{otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{for tool model C} \\ 0 & \text{otherwise} \end{cases}$$

   The regression model will be of the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$$

   and the data input for x variables are given as follow:

Model	x_1	x_2	x_3	x_4
A	x_{i1}	1	0	0
B	x_{i1}	0	1	0
C	x_{i1}	0	0	1
D	x_{i1}	0	0	0

   Read the CSV data file named Cat1, run the following code and interpret the result.

```r
Catdata<-read.csv("Cat1.CSV", header=TRUE, sep=",")
plot(Catdata$x1,Catdata$y)

#create indicator variable
x2<-ifelse(Catdata$group=='A',1,0)
x3<-ifelse(Catdata$group=='B',1,0)
x4<-ifelse(Catdata$group=='C',1,0)
df_new<-data.frame(toolwear=Catdata$y, speed=Catdata$x1, x2, x3, x4)
df_new
model1<-lm(toolwear~speed+factor(x2)+factor(x3)+factor(x4),data=df_new)
summary(model1)
summary(model1)$coef
```

5. (6 points) You are studying the relationship between two continuous variables, temperature (x_1) and humidity (x_2), and their effect on a plant's growth rate (y). The interaction between temperature and humidity is hypothesized to affect plant growth in a significant way. The model is specified as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \times x_2) + \varepsilon$$

Where:

- y: Plant growth rate (in cm/day)
- x_1: Temperature (in degrees Celsius)
- x_2: Humidity (in percentage)
- $x_1 \times x_2$: Interaction term between temperature and humidity
- ε: Error term

You estimate the following regression model from the collected data:

$$\hat{y} = 2 + 0.8x_1 + 0.5x_2 - 0.01(x_1 \times x_2)$$

(a) Without the interaction term, what is the average change in plant growth when the temperature increases by 1 degree Celsius while holding the humidity at 60%?

(b) How will your answer in (a) change with the inclusion of the interaction term?