Question 8.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a linear regression model would be appropriate. List some (up to 5) predictors that you might use.

Answer:

Regression model can be used to describe a system or to predict values. One of the applications of regression can be predicting the success rate of a medical surgery, such as dental implant surgery. To be specific, we can use regression analysis to predict the success rate of a dental implant surgery.

As for predictors, we can think of the following five factors:

1) The age of the patients
2) How immediate the surgery was/ how long has the patient lived without a tooth
3) If the patient smokes or not
4) What kind of dental implant brand the patient is going to be implanted with
5) Which teeth is being implanted (maxillary or mandibular, incisor or molar)

Question 8.2

Using crime data from http://www.statsci.org/data/general/uscrime.txt (file `uscrime.txt`, description at http://www.statsci.org/data/general/uscrime.html), use regression (a useful R function is `lm` or `glm`) to predict the observed crime rate in a city with the following data:

M = 14.0, **So = 0,** Ed = 10.0, **Po1 = 12.0,** Po2 = 15.5, **LF = 0.640,** M.F = 94.0, **Pop = 150,** NW = 1.1, U1 = 0.120, U2 = 3.6, Wealth = 3200, Ineq = 20.1, Prob = 0.04, Time = 39.0

Show your model (factors used and their coefficients), the software output, and the quality of fit. **Note** that because there are only 47 data points and 15 predictors, you'll probably notice some overfitting. We'll see ways of dealing with this sort of problem later in the course.

Answer:

In this question, I tried to narrow down the relevant predictor variables, check the quality of the fit of the chosen variables and then used the chosen model to predict the observed crime rate using the given data.

First, I ran simple regressions 16 times with all and each of the 15 predictor and sorted out variables with p-value smaller than 0.05, which resulted in the following variables: Ed, Prob, Po1, Po2, Pop, and Wealth.

As for the second round, I ran a regression model with the all of the chosen variables: Ed, Prob, Po1, Po2, Pop, and Wealth, where most of the values show p-value higher than 0.05, as in the picture below. So, as for the next step, I tried getting rid of some of the variables and ran regression models with the new combinations of the predictors.

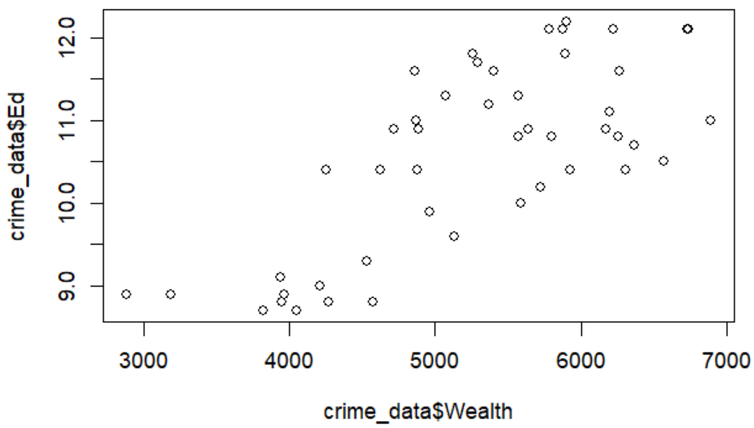
```
> com_fit_1 <- lm(Crime~ Ed+Prob+Po1+Po2+Pop+Wealth, data = crime_data)
> summary(com_fit_1)

Call:
lm(formula = Crime ~ Ed + Prob + Po1 + Po2 + Pop + Wealth, data = crime_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-597.05 -133.34   23.56  152.63  578.04 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)   532.1121   493.0890   1.079   0.2870  
Ed             64.4792    57.6568   1.118   0.2701  
Prob        -3996.2829  2180.1075  -1.833   0.0742 .
Po1           277.2720   121.6071   2.280   0.0280 *
Po2          -163.4147   129.6253  -1.261   0.2147  
Pop            -0.9084     1.3658  -0.665   0.5098  
Wealth         -0.2155     0.0932  -2.313   0.0260 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 271.8 on 40 degrees of freedom
Multiple R-squared:  0.5705,    Adjusted R-squared:  0.506 
F-statistic: 8.854 on 6 and 40 DF,  p-value: 3.692e-06
```

In eliminating the variables, I realized that some of the variables could have multicollinearity. Multicollinearity is when two or more independent variables are highly correlated to each other. We would not like this to happen, as we want the variables to be correlated to the response variable, not to each other. We can plot the variables or check VIF (Variance Inflation Factor), which measures the ratio of the whole model variance to the variance of the model that includes one single variable, to check if there is potential multicollinearity among the variables.

Here are the regression models with different combinations of the variables that I tried.

```
com_fit_1 <- lm(Crime~ Ed+Prob+Po1+Po2+Pop+Wealth, data = crime_data)
summary(com_fit_1) ## Ed and Wealth could have multicollinearity

com_fit_2 <- lm(Crime~ Ed+Prob+Po1+Po2+Pop, data = crime_data)
summary(com_fit_2) ## Dropped Wealth

com_fit_3 <- lm(Crime~ Prob+Po1+Po2+Pop+Wealth, data = crime_data)
summary(com_fit_3)    ### Dropeed Ed to check the alternative

com_fit_4 <- lm(Crime~ Prob+Po1+Pop+Wealth, data = crime_data)
summary(com_fit_4) ## Dropped Po2

com_fit_5 <- lm(Crime~ Prob+Po1+Wealth, data = crime_data)
summary(com_fit_5) ## Dropped Pop: Po1 and Pop could have multicollinearity
```

I suspected that wealth and education can have some correlation, as we could imagine that more economic power can lead to more education. The plotted result, as in the graph below, showed some convincing result.



Also, I did VIF calculation and it showed high numbers at Po1 and Po2, over the value of 5, which we could suspect correlations. In order to solve multicollinearity, I made a new regression by dropping Ed, Wealth and Po2 from the predictors, as in com_fit_2 and com_fit_3 and com_fit_5.

```
> vif(com_fit_1)
       Ed      Prob       Po1       Po2       Pop    Wealth 
 2.589956  1.529624 81.310980 81.782189  1.683166  5.034290 
```

After checking the result of com_fit_3 as in the picture below, I dropped the variable Pop, as it had p-value greater than 0.05.

```
> com_fit_3 <- lm(Crime~ Prob+Po1+Po2+Pop+Wealth, data = crime_data)
> summary(com_fit_3)

Call:
lm(formula = Crime ~ Prob + Po1 + Po2 + Pop + Wealth, data = crime_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-591.83 -117.90   27.69  152.76  554.26 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)  9.229e+02  3.490e+02   2.645   0.0115 *
Prob        -4.133e+03  2.183e+03  -1.893   0.0654 .
Po1          2.625e+02  1.213e+02   2.165   0.0363 *
Po2         -1.490e+02  1.294e+02  -1.152   0.2561  
Pop         -1.388e+00  1.301e+00  -1.067   0.2923  
Wealth      -1.537e-01  7.528e-02  -2.042   0.0476 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 272.7 on 41 degrees of freedom
Multiple R-squared:  0.557,	Adjusted R-squared:  0.503 
F-statistic: 10.31 on 5 and 41 DF,  p-value: 1.888e-06
```

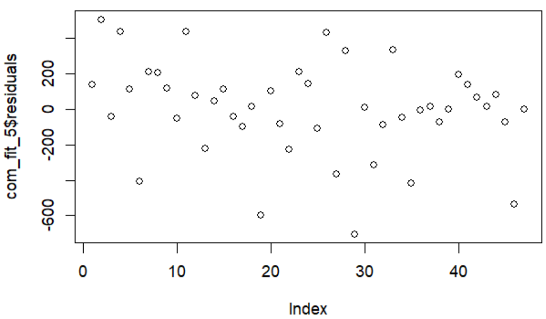
Among the combinations, com_fit_5 was selected as the best, with all of the variables showing p-value small enough. Adjusted R-square shows that the model can explain almost 50% of what is happening to the crime rate with the chosen predictor variables.

```
> com_fit_5 <- lm(Crime~ Prob+Po1+Wealth, data = crime_data)
> summary(com_fit_5)

Call:
lm(formula = Crime ~ Prob + Po1 + Wealth, data = crime_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-706.64  -85.35   14.62  137.45  506.92 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.875e+02  3.418e+02   2.597   0.0128 *  
Prob        -3.709e+03  2.140e+03  -1.734   0.0901 .  
Po1          1.137e+02  2.208e+01   5.152 6.15e-06 ***
Wealth      -1.474e-01  7.202e-02  -2.047   0.0469 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 273.7 on 43 degrees of freedom
Multiple R-squared:  0.5318,	Adjusted R-squared:  0.4991 
F-statistic: 16.28 on 3 and 43 DF,  p-value: 3.246e-07
```

In order to check the quality of the fit, I used AIC to compare the result of the five combinations. AIC test selects the model that can explain most with fewer independent variable as the best-fit model. K means the number of predictors (plus 2) in the model. AICc is the score of the model, where the smaller number signifies the better fit. AICcWt is the predictive power of the model, where the best model of the combinations could explain 54% of what is happening between the predictors and the response variable. As in the picture below, AIC result selected com_fit_5 as the best model among the 5.

```
> models <- list(com_fit_1, com_fit_2, com_fit_3, com_fit_4, com_fit_5)
> model_names <- c('com_fit_1', 'com_fit_2', 'com_fit_3', 'com_fit_4', 'com_fit_5')
> aictab(cand.set = models, modnames = model_names)

Model selection based on AICc:

          K   AICc Delta_AICc AICcWt Cum.Wt      LL
com_fit_5 5 668.20       0.00   0.54   0.54 -328.37
com_fit_4 6 669.73       1.53   0.25   0.79 -327.82
com_fit_3 7 671.01       2.81   0.13   0.92 -327.07
com_fit_1 8 672.48       4.28   0.06   0.99 -326.34
com_fit_2 7 675.46       7.26   0.01   1.00 -329.29
```

I also checked the residuals of the chosen regression model, which seems random.



And with the chosen regression model, we could predict that the crime rate, the number of offenses per 100,000 population in 1960 to be about 1632.17.

```
> new_data <- list(Prob = 0.04, Po1 = 12.0, Wealth = 3200)
> prediction <- predict(com_fit_5, new_data )
> prediction
       1
1632.165
> head(crime_data)
     M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq
1 15.1  1  9.1  5.8  5.6 0.510  95.0  33 30.1 0.108 4.1   3940 26.1
2 14.3  0 11.3 10.3  9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4
3 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180 25.0
4 13.6  0 12.1 14.9 14.1 0.577  99.4 157  8.0 0.102 3.9   6730 16.7
5 14.1  0 12.1 10.9 10.1 0.591  98.5  18  3.0 0.091 2.0   5780 17.4
6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6
      Prob    Time Crime
1 0.084602 26.2011   791
2 0.029599 25.2999  1635
3 0.083401 24.3006   578
4 0.015801 29.9012  1969
5 0.041399 21.2998  1234
6 0.034201 20.9995   682
> prediction
       1
1632.165
```

Reference

https://www.scribbr.com/statistics/akaike-information-criterion/