HW5

Alan Lo, Arjun Goyal, James Trawick, Richard Hardis

9/20/2019

# ISYE 6501 HW5 Response

## 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.

In daily life, one common activity for many people is working out. A person who is trying to lose weight or track general health metrics may be interested in knowing how many calories he/she burns in a given period of time working out. There are many factors that can be used to predict the total calorie expenditure of a workout for a given length. Some example factors include resting metabolic rate, average heart rate over the workout period, a categorical predictor describing the type of exercises performed (bike vs swim vs bicep curls), temperature of the surrounding workout environment, bodyfat content, daily diet, and .

A linear regression model incorporating a combination of these factors and others could yield a useful prediction of caloric expenditure for a given workout.

## 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.

crime\_df = read.table("uscrime.txt", header = TRUE)  
model = lm(Crime ~ ., crime\_df)  
summary(model)

##   
## Call:  
## lm(formula = Crime ~ ., data = crime\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -395.74 -98.09 -6.69 112.99 512.67   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.984e+03 1.628e+03 -3.675 0.000893 \*\*\*  
## M 8.783e+01 4.171e+01 2.106 0.043443 \*   
## So -3.803e+00 1.488e+02 -0.026 0.979765   
## Ed 1.883e+02 6.209e+01 3.033 0.004861 \*\*   
## Po1 1.928e+02 1.061e+02 1.817 0.078892 .   
## Po2 -1.094e+02 1.175e+02 -0.931 0.358830   
## LF -6.638e+02 1.470e+03 -0.452 0.654654   
## M.F 1.741e+01 2.035e+01 0.855 0.398995   
## Pop -7.330e-01 1.290e+00 -0.568 0.573845   
## NW 4.204e+00 6.481e+00 0.649 0.521279   
## U1 -5.827e+03 4.210e+03 -1.384 0.176238   
## U2 1.678e+02 8.234e+01 2.038 0.050161 .   
## Wealth 9.617e-02 1.037e-01 0.928 0.360754   
## Ineq 7.067e+01 2.272e+01 3.111 0.003983 \*\*   
## Prob -4.855e+03 2.272e+03 -2.137 0.040627 \*   
## Time -3.479e+00 7.165e+00 -0.486 0.630708   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 209.1 on 31 degrees of freedom  
## Multiple R-squared: 0.8031, Adjusted R-squared: 0.7078   
## F-statistic: 8.429 on 15 and 31 DF, p-value: 3.539e-07

From the above R output from the lm linear modeling function, we can see all of the factors and their coefficients. The adjusted R-squared value is 0.7078, which is a high value for human-related modeling scenarios. Additionally, based on the data summary seen below, we can see that all of the values of the input data lie within the ranges of each factor.

Although the dataset is small, potentially leading to overfitting, the high value along with the all of the input data not requiring extrapolation gives us confidence in the predictive capability of the model.

summary(crime\_df)

## M So Ed Po1   
## Min. :11.90 Min. :0.0000 Min. : 8.70 Min. : 4.50   
## 1st Qu.:13.00 1st Qu.:0.0000 1st Qu.: 9.75 1st Qu.: 6.25   
## Median :13.60 Median :0.0000 Median :10.80 Median : 7.80   
## Mean :13.86 Mean :0.3404 Mean :10.56 Mean : 8.50   
## 3rd Qu.:14.60 3rd Qu.:1.0000 3rd Qu.:11.45 3rd Qu.:10.45   
## Max. :17.70 Max. :1.0000 Max. :12.20 Max. :16.60   
## Po2 LF M.F Pop   
## Min. : 4.100 Min. :0.4800 Min. : 93.40 Min. : 3.00   
## 1st Qu.: 5.850 1st Qu.:0.5305 1st Qu.: 96.45 1st Qu.: 10.00   
## Median : 7.300 Median :0.5600 Median : 97.70 Median : 25.00   
## Mean : 8.023 Mean :0.5612 Mean : 98.30 Mean : 36.62   
## 3rd Qu.: 9.700 3rd Qu.:0.5930 3rd Qu.: 99.20 3rd Qu.: 41.50   
## Max. :15.700 Max. :0.6410 Max. :107.10 Max. :168.00   
## NW U1 U2 Wealth   
## Min. : 0.20 Min. :0.07000 Min. :2.000 Min. :2880   
## 1st Qu.: 2.40 1st Qu.:0.08050 1st Qu.:2.750 1st Qu.:4595   
## Median : 7.60 Median :0.09200 Median :3.400 Median :5370   
## Mean :10.11 Mean :0.09547 Mean :3.398 Mean :5254   
## 3rd Qu.:13.25 3rd Qu.:0.10400 3rd Qu.:3.850 3rd Qu.:5915   
## Max. :42.30 Max. :0.14200 Max. :5.800 Max. :6890   
## Ineq Prob Time Crime   
## Min. :12.60 Min. :0.00690 Min. :12.20 Min. : 342.0   
## 1st Qu.:16.55 1st Qu.:0.03270 1st Qu.:21.60 1st Qu.: 658.5   
## Median :17.60 Median :0.04210 Median :25.80 Median : 831.0   
## Mean :19.40 Mean :0.04709 Mean :26.60 Mean : 905.1   
## 3rd Qu.:22.75 3rd Qu.:0.05445 3rd Qu.:30.45 3rd Qu.:1057.5   
## Max. :27.60 Max. :0.11980 Max. :44.00 Max. :1993.0

The predicted crime rate for the given factor values can be found as follows:

new\_data = data.frame(  
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  
)  
  
predict(model, new\_data)

## 1   
## 155.4349

The crime rate with the given input values is 155 offenses per 100,000 population in 1960.