Homework-5

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 the Human Resource (HR) domain, the measurement and balance of staff happiness is an important indicator for the company's healthy development. Low happiness level could cause long term issues such as high turnover rates, lack of working morale, and even suppressed creativity in working domains. Using a simple linear regression model for this instance, we could try to model staffs' indicated happiness level working in the company. Some of the predictors that could be used on top of that are:

1. Productivity level (working hours)
2. Lunch hours (coupled with working hours for work-life balance indicators)
3. Remuneration level
4. Opinions towards upper management
5. Employee welfare

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.

```
library(GGally)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(boot)

set.seed(42069)

# load data
data = read.table("C:/Users/Admin/Desktop/MM/Homework 5/uscrime.txt",
                  stringsAsFactors = FALSE,
                  header = TRUE)
head(data)
```

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

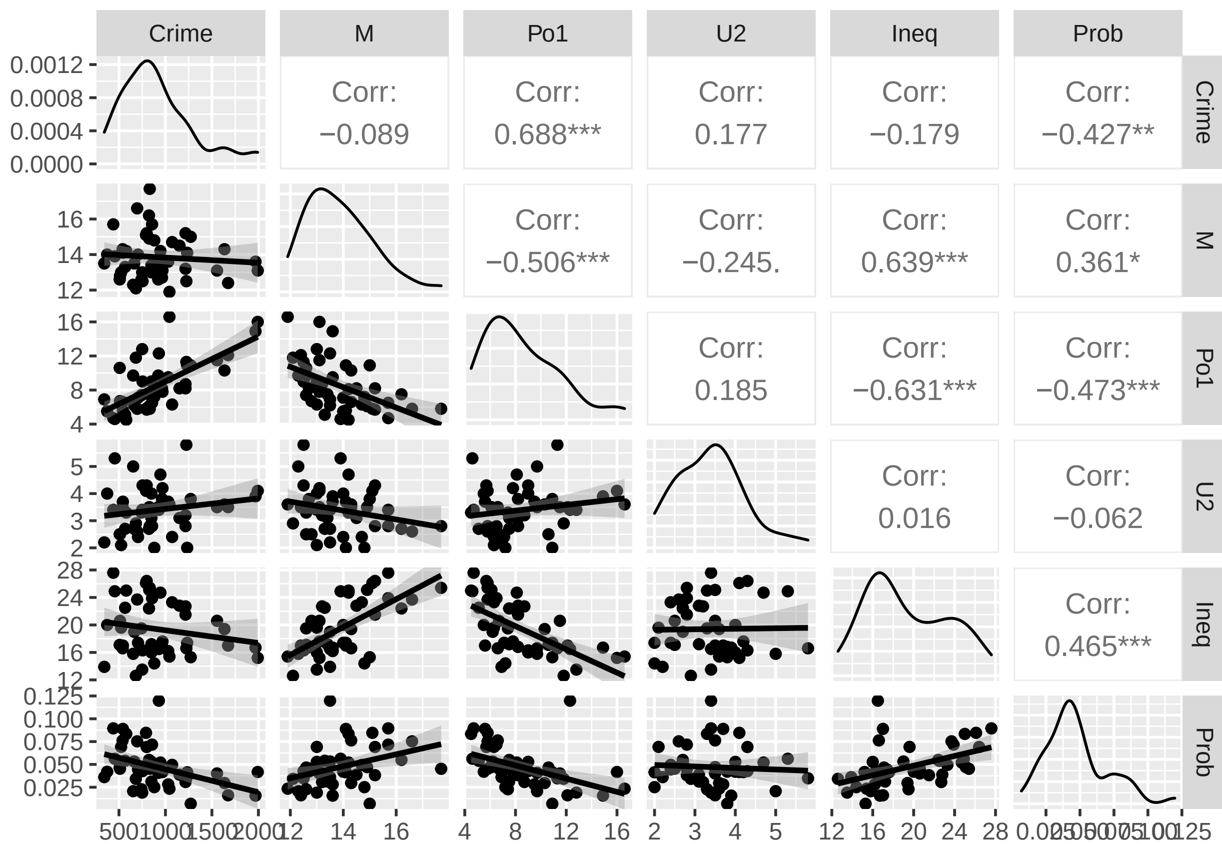
```
# test data
test = data.frame(M=14.0, So=0, Ed=10.0, Po1=12.0, Po2=15.5, LF=0.640,
                  M.F=94.0, Pop=150, NW=1.1, U1=0.120, U2=3.6,
                  Wealth=3200, Ineq=20.1, Prob=0.040, Time = 39.0)
```

```
# build model
model = glm(Crime~., data=data)
summary(model)
```

```
##
## Call:
## glm(formula = Crime ~ ., data = data)
##
## Deviance 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
##
```

```
## (Dispersion parameter for gaussian family taken to be 43707.93)
##
##     Null deviance: 6880928  on 46  degrees of freedom
## Residual deviance: 1354946  on 31  degrees of freedom
## AIC: 650.03
##
## Number of Fisher Scoring iterations: 2
```

```
data %>% ggpairs(columns = c('Crime', 'M', 'Po1', 'U2', 'Ineq', 'Prob'),
                 lower=list(continuous='smooth'))
```




```
# model fit
tss = sum((data$Crime - mean(data$Crime))^2) # total sum of squared
rss = sum((model$residuals)^2) # residual sum of squared
rsq = 1 - rss/tss
```

Model with all predictors yield $R^2 = 0.8030868$, explaining 80.3086758 % of the data's variability. Using insignificant variables might overfit the data.

```
# building model with significant variables only
cmodel = glm(Crime~M+Ed+Po1+U2+Ineq+Prob, data=data)
summary(cmodel)
```

```
##
```

```
## Call:
## glm(formula = Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -470.68   -78.41   -19.68   133.12   556.23
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5040.50     899.84  -5.602 1.72e-06 ***
## M             105.02      33.30   3.154  0.00305 **
## Ed            196.47      44.75   4.390 8.07e-05 ***
## Po1           115.02      13.75   8.363 2.56e-10 ***
## U2             89.37      40.91   2.185  0.03483 *
## Ineq           67.65      13.94   4.855 1.88e-05 ***
## Prob        -3801.84    1528.10  -2.488  0.01711 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 40276.42)
##
##     Null deviance: 6880928  on 46  degrees of freedom
## Residual deviance: 1611057  on 40  degrees of freedom
## AIC: 640.17
##
## Number of Fisher Scoring iterations: 2
```

```r
crss = sum((cmodel$residuals)^2) # residual sum of squared
crsq = 1 - crss/tss
```

Model with significant predictors yield R^2=0.7658663 , explaining 76.5866329 % of the data's variability. AIC is lower for this model as compared the previous.

```r
# prediction
predict(cmodel, test)
```

```
##        1
## 1304.245
```