ISYE-6501
Fall 2024

Homework 4

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.

As a quality data analyst in a hospital, a linear regression model can be used to predict the length of stay (LOS) for patients with respiratory issues. This would help determine the correct staffing needs along with how many beds should be assigned to only patients with respiratory issues. Please find below 5 predictors that I might use to help determine this.

- Age (older patients tend to have more respiratory issues)
- Severity of illness (using a scoring method like apache 2 helps determine the mortality rate)
- Patient comorbidities (patients with pre-existing conditions will required longer care)
- Type of admission (patients who with respiratory issues that come through the emergency department tend to have longer LOS)
- Elective treatment protocol (a patient getting admitted as an ICU patient will tend to have a higher LOS)

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.

ANALYSIS:

When using all the predictors for the model from the data, I found that the adjusted R-squared value was about 0.71 which means that the model is pretty well fit for the data. I was able to predict using the test value predictors that the crime would be about 155.4. This value does not seem very accurate due to the fact that the lowest value of crime in all the data is 342. Which means that the model predicted a crime value less than half of the lowest observed crime rate. This could be a sign that the model was overfitting. I decided to make another model but only with the predictors that have a p-value of .05 or lower. Meaning only the predictors that are statistically significant. The 4 predictors with a p-value lower than .05 are M, Ed, Ineq, and Prob. The new model with only these 4 predictors showed a much lower adjusted R-squared value of about 0.19. However the model predicted a crime rate that seems much more plausible of 897.2.

This can help explain that R-squared should not be the only determinant to look at when seeing if a model fits best and if it will offer the best predictions. A lower R-squared could be a sign that there is just a lot of variability in the data, what is important to look at in my opinion is the statistical significance of each variable/predictor in the data and work a model with those statistically significant variables.

Below is the code/ R markdown, find the input in black, the comments in green, and the output in blue.

CODE:

```
> #load the data uscrime data into a table named data.
> data <- read.table("uscrime.txt", header = TRUE)
> #below I load the test/prediction parameters into a data frame
> crime_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 = 0.04, Time = 39.0)
> #below I use the lm function to load the data into a linear model
> model<-lm(Crime~.,data=data)
> #details and summary of the model
> summary(model)
Call:
lm(formula = Crime ~ ., data = data)

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
> #predict the crime rate using the model and the test/prediction parameters
> crime_prediction <- predict(model, crime_test)
> #details and summary of the prediction
> summary(crime_prediction)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  155.4   155.4   155.4   155.4   155.4   155.4 
> #checking min and max value of crime column
> min(data$Crime)
[1] 342
> max(data$Crime)
[1] 1993
> #using a new model with only predictors that have a p-value higher than .05, we can see from the model that the predictors with one or more "*" next to the p-value mean signify equal or less than 0.05.
> model2<-lm(Crime~M+Ed+Ineq+Prob,data=data)
> #details and summary of the new model
> summary(model2)
Call:
lm(formula = Crime ~ M + Ed + Ineq + Prob, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
```

```
-532.97 -254.03  -55.72  137.80  960.21

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1339.35    1247.01  -1.074  0.28893
M              35.97      53.39   0.674  0.50417
Ed            148.61      71.92   2.066  0.04499 *
Ineq           26.87      22.77   1.180  0.24458
Prob        -7331.92    2560.27  -2.864  0.00651 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 347.5 on 42 degrees of freedom
Multiple R-squared:  0.2629,	Adjusted R-squared:  0.1927
F-statistic: 3.745 on 4 and 42 DF,  p-value: 0.01077
> #new prediction using new model
> crime_prediction2 <-predict(model2, crime_test)
> #summary and details of new prediction
> summary(crime_prediction2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  897.2   897.2   897.2   897.2   897.2   897.2
```