ISYE 6501 Week 5 Homework

2023-09-20

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.

Linear regression can be used to estimate the relationship between two or more variables. One application of linear regression could be to estimate the sale price of a home. Relevant predictors may include the square footage, number of bedrooms and bathrooms, location, lot size, and age of property.

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.

After performing exploratory data analysis, I fitted a linear regression model using all of the columns in the crime_data data frame.

```
#load the data
crime_data <- read.table("http://www.statsci.org/data/general/uscrime.txt", header=TRUE)

#fit linear regression model
lm_model <- lm(Crime~., data=crime_data)
```

After fitting the model, I used the summary() function to observe the r-squared value, adjusted r-squared value, p-values, and t-stats. I also viewed the model's coefficients.

```
#view summary
summary(lm_model)
```

```
##
## Call:
## lm(formula = Crime ~ ., data = crime_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
```

```
#view coefficients
lm_model$coefficients
```

```
##   (Intercept)             M            So            Ed           Po1
## -5.984288e+03  8.783017e+01 -3.803450e+00  1.883243e+02  1.928043e+02
##           Po2            LF           M.F           Pop            NW
## -1.094219e+02 -6.638261e+02  1.740686e+01 -7.330081e-01  4.204461e+00
##            U1            U2        Wealth          Ineq          Prob
## -5.827103e+03  1.677997e+02  9.616624e-02  7.067210e+01 -4.855266e+03
##          Time
## -3.479018e+00
```

Next, I calculated the Sum of Square Errors (SSE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) to get a baseline evaluation of model output. I will compare these values to future models to better understand model performance.

```r
#evaluate output of new model(SSE, AIC, BIC)
SSE <- sum(lm_model$residuals ^ 2)
SSE
```

```
## [1] 1354946
```

```r
AIC <- AIC(lm_model)
AIC
```

```
## [1] 650.0291
```

```r
BIC <- BIC(lm_model)
BIC
```

```
## [1] 681.4816
```

Then, I used this model to make a prediction on new data.

```r
#create prediction data frame
new_data <- data.frame(M=14, So=0, Ed=10, Po1=12, Po2=15.5, LF=0.640, M.F=94, Pop=150,
                       NW=1.1, U1=0.12, U2=3.6, Wealth=3200, Ineq=20.1, Prob=0.04, Time=39, Crime=NA)

#use lm_model to predict crime rate on new data
prediction <- predict(lm_model, newdata=new_data)
prediction
```

```
##        1
## 155.4349
```

This prediction is not very probable since it does not fall within the range of existing crime rate observations (see range below). Thus, it indicates that the linear regression model is likely overfitting the data.

```r
#calculate range
crime_range <- range(crime_data$Crime)
crime_range
```

```
## [1]  342 1993
```

To create a better model, I performed feature selection using a $p < 0.05$ significance threshold. Using this approach, I refit the linear regression model using only significant attributes (M, Ed, Ineq, and Prob).

```r
#refit linear regression model
lm_model2 <- lm(Crime~M+Ed+Ineq+Prob, data=crime_data)
```

Like before, I observed the r-squared value, adjusted r-squared value, p-values, t-stats, and coefficients of my new model. The first model has a higher r-squared value, but this doesn't necessarily mean it is a better model. Including irrelevant variables in a model can artificially inflate r-squared without contributing to a better understanding of the dependent variable.

However, my chosen p-values are no longer beneath the $p < 0.05$ significance threshold, which signals that they are not providing enough information to reject the null hypothesis.

```r
#view model summary
summary(lm_model2)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Ineq + Prob, data = crime_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
```

```r
#view coefficients
lm_model2$coefficients
```

```
## (Intercept)           M          Ed        Ineq        Prob
## -1339.34621    35.97296   148.60531    26.87457 -7331.91531
```

Next, I calculated the Sum of Square Errors (SSE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) to compare with model 1. It does not surprise me that these values are larger compared to model 1 since model 1 was overfitting the data. However, the magnitude of the difference raised alarm and signaled that I may want to try a different model.

```r
#evaluate output of new model(SSE, AIC, BIC)
SSE2 <- sum(lm_model2$residuals ^ 2)
SSE2
```

```
## [1] 5071868
```

```r
AIC2 <- AIC(lm_model2)
AIC2
```

```
## [1] 690.0666
```

```r
BIC2 <- BIC(lm_model2)
BIC2
```

```
## [1] 701.1675
```

I used model 2 to make a prediction on new data. Compared to prediction 1, prediction 2 seems like a much more realistic estimate of crime rate.

```
#use lm_model to predict crime rate on new data
prediction2 <- predict(lm_model2, newdata=new_data)
prediction2
```

```
##        1
## 897.2307
```

While model 2 is likely a better estimate than model 1, it still may not be the best model. It is possible we left out important features that are actually meaningful to the model. For this reason, I built a third model but increased my significance threshold for feature selection to $p < 0.1$. Using this approach, I refit the linear regression model using only significant attributes (M, Ed, Po1, U2, Ineq, and Prob).

```
#refit linear regression model
lm_model3 <- lm(Crime~M+Ed+Po1+U2+Ineq+Prob, data=crime_data)
```

Once again, I observed the r-squared value, adjusted r-squared value, p-values, t-stats, and coefficients of my new model. Adding the Po1 and U2 predictors significantly increased the r-squared value. Additionally, model 3's p-values are still beneath my pre-determined significance threshold.

```
#view model summary
summary(lm_model3)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data = crime_data)
##
## 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
##
## Residual standard error: 200.7 on 40 degrees of freedom
## Multiple R-squared:  0.7659, Adjusted R-squared:  0.7307
## F-statistic: 21.81 on 6 and 40 DF,  p-value: 3.418e-11
```

```
#view coefficients
lm_model3$coefficients
```

```
## (Intercept)           M          Ed         Po1          U2        Ineq
## -5040.50498   105.01957   196.47120   115.02419    89.36604    67.65322
##        Prob
## -3801.83628
```

Next, I calculated the Sum of Square Errors (SSE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) to compare with model 2. These values are much lower in model 3 compared to model 2, indicating a better fit to the data.

```
#evaluate output of new model(SSE, AIC, BIC)
SSE3 <- sum(lm_model3$residuals ^ 2)
SSE3
```

```
## [1] 1611057
```

```
AIC3 <- AIC(lm_model3)
AIC3
```

```
## [1] 640.1661
```

```
BIC3 <- BIC(lm_model3)
BIC3
```

```
## [1] 654.9673
```

Finally, I used model 3 to make a prediction on new data. While this is on the higher end of predictions, it still falls within a realistic range.

```
#use lm_model to predict crime rate on new data
prediction3 <- predict(lm_model3, newdata=new_data)
prediction3
```

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

All things considered, I think model 3 is the best of my 3 models. It has a high R-squared value, significant p-values, and SSE/AIC/BIC similar to the baseline model.