Homework 5: ISYE 6501 - Introduction to Analytics Modeling

Question 8.1

Prompt

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.

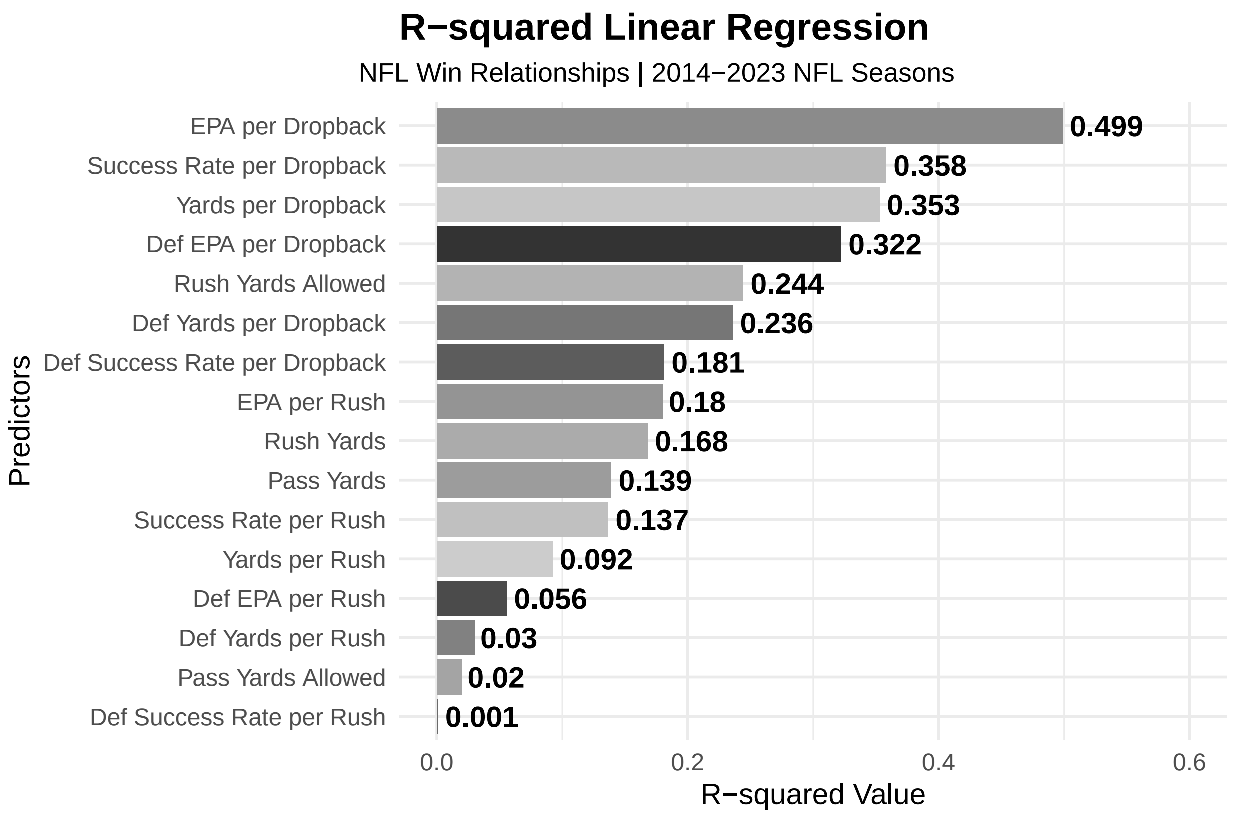
Solution

Considering american football in the National Football League (NFL), the most important outcome is winning. NFL teams or analysts may consider what are the greatest factors to win games. Linear regression would be appropriate to provide insights on the predictors that have the greatest impact on win or loss outcomes. This can be particularly useful when considering the importance of players to retain on the roster, where salary cap budgets should be distributed, or how to construct a team roster with drafting players.

Modeling with linear regression to determine metrics that have the greatest impact on win or loss outcomes should consider predictors from both offense and defense. Using play-by-play data from `nflfastR` [1], several predictors can be obtained dating back to the 1999 season that will assist in using a linear regression model. Some predictors that should be considered include:

- **EPA per Dropback:** Expected Points Added (EPA) per dropback is an advanced statistic that measures the efficiency of a quarterback per-dropback. This allows for providing insight into the effectiveness of a QB's contribution towards the team's scoring potential.
- **Pass Yards:** Evaluates the amount of yards thrown through the air and can provide insight on the success of both the QB and well as the receiver unit on offense.
- **Rush Yards:** Evaluates the amount yards run on the ground and can evaluate the performance of the running back as well as offensive line play.
- **Pass Yards Allowed:** Amount of yards allowed through the air, which can provide insight on the performance of the secondary on defense to prevent scoring.
- **Rush Yards Allowed:** Amount of yards allowed on the ground, which can assist in determining the performance of the defensive line preventing scoring.

Overall, the goal is to win, and the outcome is determined by scoring. Dependent upon the performance of the offense, points can be scored to win games. Dependent upon the defense, scoring can be prevented to win games. To elaborate on this, using NFL play-by-play data from the last decade (2014-2023), as well literature on exploring wins with `nflfastR` [2], a linear regression model is formed using `lm()` from the stats package to identify the most important predictors for NFL wins. The code is developed with R, but is redacted due to it not being required and to keep the document relatively short.



Data: @nflfastR | Linear Regression Model: lm {stats}

Pass efficiency predictors are shown in the figure to have the strongest relationships to winning. This makes sense as typically the leader of the offense, the QB, must have above-average efficiency to win a game. From the R-squared values in the figure above, we can determine that the win and loss outcomes are most impacted by the performance of the quarterback (QB) and passing efficiency predictors.

Question 8.2

Prompt

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.

Solution

First before utilizing the data set, the description should be recognized to identify exactly what is being worked with in the problem. The data set for `uscrime.txt` uses aggregate data on 47 states of the USA for 1960. The data set provided by [3] contains the following columns:

- `M`: Percentage of males aged 14–24 in total state population
- `So`: Indicator variable for a southern state
- `Ed`: Mean years of schooling of the population aged 25 years or over
- `Po1`: Per capita expenditure on police protection in 1960
- `Po2`: Per capita expenditure on police protection in 1959
- `LF`: Labour force participation rate of civilian urban males in the age-group 14-24
- `M.F`: Number of males per 100 females
- `Pop`: State population in 1960 in hundred thousands
- `NW`: Percentage of nonwhites in the population
- `U1`: Unemployment rate of urban males 14–24
- `U2`: Unemployment rate of urban males 35–39
- `Wealth`: Wealth: median value of transferable assets or family income
- `Ineq`: Income inequality: percentage of families earning below half the median income
- `Prob`: Probability of imprisonment: ratio of number of commitments to number of offenses
- `Time`: Average time in months served by offenders in state prisons before their first release
- `Crime`: Crime rate: number of offenses per 100,000 population in 1960

As described in the prompt, useful R functions include `lm` and `glm`. These regression packages are contained in the `stats` package, which is part of base R and are loaded by default.

Initial Approach

Initially to approach the problem, I set my current working directory to `HW5` and load the `uscrime.txt` data set into a table. Using `head()` I display the data to determine if the import was successful and observe the data set. Observing the data, there are 16 columns (variables) that contain 47 rows of data (for all 47 states in 1960) that are described in the list of variables stated above.

```r
# Import libraries
# lm and glm are in the stats package loaded by default
library(ggplot2) # Plot functions

# Set the working directory
setwd("~/projects/ISYE6501/HW5")

# Load the crime data into a table
data <- read.table("data/uscrime.txt", stringsAsFactors = FALSE, header = TRUE)

# View the imported data to check import but only first few rows
head(data, 5)
```

```
##      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
##      Time Crime
## 1 26.2011   791
## 2 25.2999  1635
## 3 24.3006   578
## 4 29.9012  1969
## 5 21.2998  1234
```

Both packages, `lm` and `glm` are functions for linear regression. The purpose of `lm` is to fit linear models, used to carry out regression, single stratum analysis of variance, and analysis of covariance [4]. The `glm` function is used to fit generalized linear models, specified by giving a symbolic description of the linear predictor and a description of the error distribution [5].

Using the Crime data, it should be noted the response variable is `Crime` and the remainding 15 variables are predictors. The first approach is implementation of the `lm` model from the R stats package. The `lm` model is developed by considering the US crime data and several arguments that are available. By referencing the `lm {stats}` documentation, a formula and data are required arguments (`lm(formula, data)`).

Forming the Linear Regression Model

The model formula defines the **model factors** for prediction. Essentially, this indicates multiple linear regression, where the relationship of crime rate or `Crime` is determined with the 15 variables or predictors. The model factors used to predict crime are defined as:

```
Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 + Wealth + Ineq + Prob + Time
```

In the code below, the model factors or `formula` arguments is defined with a shorthand expression that is equivalent to the model factors. The formula is defined by the expression `Crime~.`, which refers to `Crime` as the response variable to predict, or the crime rate. The ~ indicates separating `Crime` from all other variables in the dataframe, indicated by `.`.

A summary of the `lm` model is provided by using `summary(lm_crime)`.

```
# Predict observed crime rate
# Crime~. separates Crime (response variable) from predictors
model <- lm(Crime~., data=data)
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
```

Evaluating the Linear Regression Model

First, the prompt discusses to show the model coefficients. The summary of the model, `summary(model)` provides the **model coefficients** or estimated effects of each predictor on crime rate in the `Estimate` column.

There is additional important information in the summary, or **software output** that is noteworthy. Standard error of each estimate is tabulated in `Std. Error`. The `t value` is the estimate divided by the standard error, or the t-statistic. `Pr(>|t|)` is the p-value for each coefficient. It is also important to note that values with asterisks (`*`) indicate higher significance with more asterisks, such as `***`.

The `Multiple R-Squared` value of 0.8031 is displaying a 80.31% variance in the crime rate. R-Squared values range from 0 to 1, with 1 indicating the fit is perfect. `Multiple R-Squared` has a value of 0.8031 which suggests a good fit, while `Adjusted R-squared` has a lower value of 0.7078, which is not as good, but can still be a reasonable fit. In this case of predicting the crime rate, `Adjusted R-squared` is important. `Adjusted R-squared` normalizes `Multiple R-Squared` taking into account the number of data points and predictors that were utilized in the model.

For the linear regression model, the hypothesis for testing whether each variable is related to the crime rate is if $p \leq 0.05$, where linearity is assumed [6]. Observing the data, predictors `So, LF, M.F, Pop, NW, Wealth, Time` do not appear to have a significant impact on the crime rate. These predictors are determined to have insignificance due to the high p-value, indicated in the summary as `Pr(>|t|)`. The p-value can be interpreted as a probability where the smaller values indicate it is unlikely to observe a significant association between the predictor and response due to chance [7]. This also indicates that the very low p-value (3.539e-07) for `F-statistic: 8.429 on 15 and 31 DF` indicates the model is statistically significant. In summary, observing smaller p-values infers there is an association between the predictor and the response.

Observing the Relationship between Crime and Predictors

To observe the relationship between `Crime` and each of the 15 predictors, a matrix scatter plots is developed with `ggplot` and `gridExtra`. Remember, `ggplot` was imported earlier, so `gridExtra` is the only library that will need to be loaded. First, a list of the predictor variables is defined as `predictors`. The list of predictors is then utilized to generate scatter plots. A function is created to create a single scatter plot so that mapping can be utilized to create all 15 plots at once and be placed in a matrix for a single image.

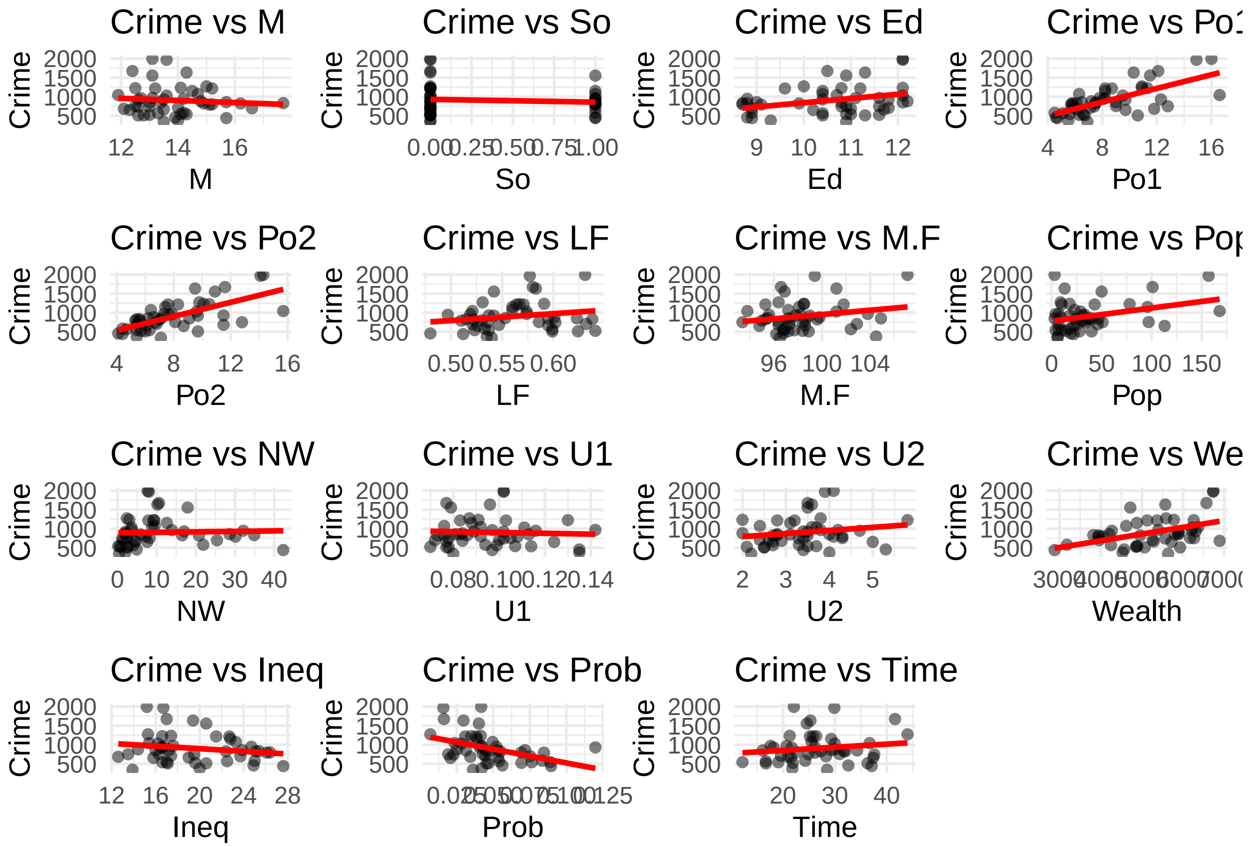
```r
library(gridExtra)

# Create a list of predictor variables
predictors <- c("M", "So", "Ed", "Po1", "Po2", "LF", "M.F", "Pop", "NW", "U1",
                "U2", "Wealth", "Ineq", "Prob", "Time")

# Function to create a scatter plot with regression line for each predictor
plot_predictor <- function(predictor) {
  ggplot(data, aes_string(x = predictor, y = "Crime")) +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "lm", se = FALSE, color = "red") +
    theme_minimal() +
    labs(title = paste("Crime vs", predictor))
}

# Create plots for all predictors
plots <- map(predictors, plot_predictor)

# Arrange plots in a grid
scatter <- grid.arrange(grobs = plots, ncol = 4)
scatter
```



Predicting Crime Rate

To predict the crime rate, the linear regression model coefficients need to be extracted. The coefficients observed in the summary's `Estimate` column are extracted by using `coef(model)`. The coefficients represent the change in predicted crime rate. Using the extracted coefficients with the provided city data, the predicted crime rate can be quantified. The new city data provided in the prompt is defined in a dataframe to be utilized with the model predictions. Model predictions are performed with `predict(object, ...)` from the R stats package. To use `predict`, the `object` argument is for the `lm` model to utilize for prediction. In this case, `object` is the `lm` model defined as `model`. An additional argument, `newdata`, is defined to specify the first place to look for explanatory variables to be used for prediction [8].

```r
# Extract the coeffieciens
coefficients <- coef(model)

# Create a data frame with the given city data from the prompt
new_city <- 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 the crime rate with predict {stats}
predicted_crime_rate <- predict(model, newdata = new_city)
paste0("Predicted Crime Rate per 100,000 Population:", predicted_crime_rate)
```

```
## [1] "Predicted Crime Rate per 100,000 Population:155.434896887446"
```

Utilizing the model coefficients, a prediction of the crime rate for a new city is determined. The predicted crime rate result is 155 offenses per 100,000 population in 1960 considering the provided city data.

Conclusion

In conclusion, a linear regression model is developed to evaluate crime rate. The model suggested a reasonable fit after evaluating the Multiple R-Squared value and the Adjusted R-Squared value. Additional evaluation of resulting p-values of the linear regression model suggested which variables are correlated to crime rate. The high p-values for predictors `So, LF, M.F, Pop, NW, Wealth, Time` provided insight that they do not appear to have a significant impact on the crime rate.

References

[1] S. Carl and B. Baldwin, *nflfastR: Functions to efficiently access NFL play by play data*. 2024.

[2] A. Ryan, "Open source football: Exploring wins with nflfastR." 2020, [Online]. Available: https://www.opensourcefootball.com/posts/2020-08-23-exploring-wins-with-nflfastr/.

[3] O. -Australasian Data and S. Library, "Effect of punishment regimes on crime rates." http://www.statsci.org/data/general/uscrime.html, 2024.

[4] "Lm: Fitting linear models - r documentation — rdocumentation.org." https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm.

[5] "Glm: Fitting generalized linear models - r documentation — rdocumentation.org." https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm.

[6] K. Okoye and S. Hosseini, *R programming: Statistical data analysis in research*. Springer Nature, 2024.

[7] J. Gareth, W. Daniela, H. Trevor, and T. Robert, *An introduction to statistical learning: With applications in r*. Spinger, 2013.

[8] "Predict: Model predictions - r documentation — rdocumentation.org." https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/predict.