The Analytics Edge 23 October, 2019

Midterm Examination

NAME: _____

STUDENT ID: _____

HONOR CODE: As a member of the SUTD community, I pledge to always uphold honorable conduct. I will be accountable for my words and actions, and be respectful to those around me. All work turned in for this test is solely my own and I take pride in this.

SIGNATURE: _____

- The exam is 1 hour and 45 minutes in duration.
- There are a total of 30 questions, each carrying one point.
- Answer all the questions in the boxes provided. You do NOT need to write your R commands, or any justification in the answer sheet unless they are specifically asked for.
- The data sets are available in the e-dimension course content folder at the link `Mid-term`.
- For your work, you may use an R-script, R-markdown or just an R session. Immediately at the end of the exam, save your R-file.
  - To do this go to `File` > `Save as` > `"Yourname_IDnumber.R"`, or, `"Yourname_IDnumber.txt"`, or, `"Yourname_IDnumber.Rmd"`.
  - Upload the files via the `Mid-term` link on e-dimension.
  - If the link does not work send the files to `bikram@sutd.edu.sg`.
  - This will be used only to validate that your work is original.
- You are allowed to use only the notes from the lectures (associated paper, book, slides) and R for this exam.
- The use of the Internet and mobilephones are NOT permitted.
- Good luck!

1. (16 points) Criminologists are interested in the effect of punishment regimes on crime rates. This has been studied using aggregate data on 47 states of the USA for the year 1960. The data set `USCrime.csv` contains the following columns:

Variable	Description
`M`	percentage of males aged 14–24 in total state population
`So`	indicator variable for a southern state (binary variable)
`Ed`	mean years of schooling of the population aged 25 years or over
`Po1`	per capita expenditure on police protection in 1960 (in US $)
`Po2`	per capita expenditure on police protection in 1959 (in US $)
`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 aged 14–24
`U2`	unemployment rate of urban males aged 35–39
`Wealth`	wealth: median value of transferable assets or family income (in US $)
`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
`States`	two letter code for the state name

Our goal in this study is to find which factors affect crime rate using this dataset.

(a) Read the data into the dataframe `UScrime`. Which state has the lowest and which state has the highest crime rates respectively?

(b) Find the average crime rate (`Crime`) in states where per capita expenditure on police protection in 1960 was above US$ 8. Also find the average crime rate for states where per capita expenditure on police protection in 1960 was less than or equal to US$ 8.

(c) Run a two-sample t-test to verify if there is any difference in crime rates for states spending more than US $ 8 per month and states spending less than or equal to US $ 8 per month on police protection (both for the year 1960). Write down the null hypothesis for your test.

(d) For the test conducted in part (c) write down the p-value and your conclusion.

(e) In this question, you will develop a simple linear regression model. Create a training set with the first 42 states (first 42 rows) and remove the `States` variable as follows:
`UStrain = subset(UScrime[1:42,], select=-States)`

Create a simple linear regression model using `UStrain` to fit the Crime rates of a state against their per capita expenditure on police protection in 1960.

What is the R-squared for your model?

(f) Create a test dataset by choosing the 43rd to 47th states (`VA, WA, WV, WI, WY`). Call this dataset `UStest` (remove the `States` variable as before). What is the 99% confidence interval for the predicted crime rate for Washington (`WA`) using the model from part (e)? Is the actual crime rate value within this interval?

(g) What is the 99% confidence interval for the predicted crime rate for Wisconsin (`WI`) using the same model? Is the actual crime rate value within this interval?

(h) Now develop a multiple linear regression model with the dataset `UStrain` to predict Crime rates using all available predictor variables (and intercept). Which variables are significant at 0.05 level and what is the R-squared for this model?

(i) Using only the variables that you found significant at 0.05 level (p-value less than 0.05) in the previous model in part (h), create a new linear regression model on the `UStrain` data set. What is the R-squared for the new model?

(j) Among the models you created in parts (e), (h) and (i) which one would you prefer? What comparison criteria did you use to come up with your conclusion?

(k) We now use the `regsubsets()` function in the `leaps` package for a best subset selection. How many variables are included in the model if you use adjusted R-squared to pick your model (use the `UStrain` data set)?

(l) Another criterion for selecting models that is often used is Bayesian information criterion (BIC). For a model with n observations and k parameters where $\hat{\boldsymbol{\beta}}$ is the least-squares estimate of the model parameter $\boldsymbol{\beta}$, it is defined as

$$\text{BIC} = k \log(n) - 2 \log(\text{likelihood}(\hat{\boldsymbol{\beta}})).$$

Lower BIC values are preferred to higher BIC values. BIC is provided as an output of `regsubsets()` called `bic`. How many variables are included in the model if you use BIC to pick your model with best subset selection (use the `UStrain` data set)?

(m) How many variables are included in the model if you use BIC to pick your model with forward stepwise selection (use the `UStrain` data set)?

(n) Now using `UStest` set for your out-of-sample validation, you are asked to choose one of the models from parts (k), (l), and (m). Create separate multiple linear regression models using the predictors found in parts (k), (l) and (m). Use these models to find the sum-of-squared errors (SSE) for the `UStest` data set.
What is the test set SSE for the model in part (k)?

(o) What are the test set SSEs for the models in parts (l) and (m)?

(p) Which model do you prefer among the ones created in parts (k), (l) and (m) based on your findings above?

2. (14 points) Record labels often face a decision problem of which musical releases to support to maximize their financial success. In this question our goal is to predict whether a song will reach a spot in the **Top 10** of the Billboard Hot 100 Chart. Taking an analytics approach, we aim to use information about a song's properties to predict its popularity. The dataset `songs.csv` consists of all songs which made it to the **Top 10** of the Billboard Hot 100 Chart from 1990–2010 plus a sample of additional songs that didn't make the **Top 10**.

   The variables included in the dataset either describe the artist or the song, or they are associated with the following song attributes: time signature, loudness, key, pitch, tempo, and timbre. Here's a detailed description of the variables:

   - `year` = the year the song was released;
   - `songtitle` = the title of the song;
   - `artistname` = the name of the artist of the song;
   - `songID` and `artistID` = identifying variables for the song and artist;
   - `timesignature` and `timesignature_confidence` = a variable estimating the time signature of the song, and the confidence in the estimate;
   - `loudness` = a continuous variable indicating the average amplitude of the audio in decibels;
   - `tempo` and `tempo_confidence` = a variable indicating the estimated beats per minute of the song, and the confidence in the estimate;
   - `key` and `key_confidence` = a variable with twelve levels indicating the estimated key of the song (C, C#, . . ., B), and the confidence in the estimate;
   - `energy` = a variable that represents the overall acoustic energy of the song, using a mix of features such as loudness;
   - `pitch` = a continuous variable that indicates the pitch of the song;
   - `timbre_0_min, timbre_0_max, timbre_1_min, timbre_1_max, ..., timbre_11_min`, and `timbre_11_max` = variables that indicate the minimum/maximum values over all segments for each of the twelve values in the timbre vector (resulting in 24 continuous variables);
   - `Top10` = a binary variable indicating whether or not the song made it to the Top 10 of the Billboard Hot 100 Chart (1 if it was in the top 10, and 0 if it was not).

(a) Read the data into dataframe `songs`. How many songs appear in the chart by Michael Jackson, and how many of them reached Top 10?

(b) Note that `time_signature` is a discrete variable. What are the values taken by `time_signature` and what are the frequency of occurrences of each value in the dataset?

(c) We would like to predict if a song makes it to the Top 10 chart. The outcome is listed in the variable `Top10`. Create a training set `SongsTrain` with observations up to and including 2008 releases and a testing set `SongsTest` consisting of 2009 and 2010 song releases.

How many observations (songs) are in the test set?

(d) Now we remove the variables not used for prediction: "year", "songtitle", "artistname", "songID" or "artistID". Define

```
nonvars = c("year", "songtitle", "artistname", "songID", "artistID").
```

Create training and test sets as follows:

```
SongsTrain = SongsTrain[ ,!(names(SongsTrain) %in% nonvars)]
SongsTest = SongsTest[ ,!(names(SongsTest) %in% nonvars)]
```

Now create a logistic regression model to predict `Top10` using all other available variables (using `SongsTrain`). We will call this **Model 1**. What is the value of Akaike's Information Criterion (AIC)?

(e) Songs with heavy instrumentation are supposed to be louder (higher "loudness" value). Sometimes they also tend to be more energetic (high "energy" value). From the coefficients obtained in **Model 1**, which one of the following do you think is more likely:

(i) Mainstream listeners prefer songs with heavy instrumentation and high energy.

(ii) Mainstream listeners prefer songs with light instrumentation and high energy.

(iii) Mainstream listeners prefer songs with heavy instrumentation and low energy.

(iv) Mainstream listeners prefer songs with light instrumentation and low energy.

(f) Find the correlation between loudness and energy (using the data set `SongsTrain`).

(g) Create a new logistic regression model called **Model 2** with the `loudness` component removed (using the data set `SongsTrain`). What does your **Model 2** suggest? Choose one of the following answers and give a short justification for your choice.

(i) Mainstream listeners prefer songs with high energy, contradicting Model 1.

(ii) Mainstream listeners prefer songs with low energy, as we saw in Model 1.

(h) Create **Model 3** like **Model 1** but just removing `energy` now (instead of `loudness` which was removed in **Model 2**). Still use the data set `SongsTrain`. What is the AIC of this model?

(i) Now make predictions on your test set `SongsTest` (use **Model 3**). Compute the accuracy of the model using a threshold value of 0.40. Also find the baseline accuracy (this is the accuracy if we pick the most frequent outcome as our prediction).

(j) What are the sensitivity and specificity values you obtain using **Model 3** on the `SongsTest` data with threshold 0.40?

(k) Use the command `performance` in the `ROCR` package to find the AUC value for **Model 3** on the test data.

(l) Let us check the robustness of your findings by trying an alternative to logistic regression, namely, probit regression where the equation to be estimated is given

$$\Pr(Y = 1) = \Phi(\beta_0 + \beta_1 x_1 + \ldots + \beta_p)$$

where Φ is the standard normal cumulative distribution function. Build this model using the data set `SongsTrain`, which we call **Model 4** with only the statistically significant predictors (at level 0.05) that you identified while building **Model 3** on `SongsTrain`.
Hint: You can fit the model by using `glm` by modifying the family argument with `family = binomial(link="probit")`. Write down the accuracy for the probit fit with a threshold of 0.40 on the test data set `SongsTest`.

(m) What are the sensitivity and specificity values you obtain using **Model 4** on the `SongsTest` data with threshold 0.40?

(n) From your findings in parts (h)-(l), would you prefer **Model 3** or **Model 4** as your choice, or, are they indistinguishable? Give a short justification.

END OF EXAM