The Analytics Edge Summer 2022

Test your knowledge of Logistic Regression in R

Exercise: Week 3

1. In this question, we will use the data in `baseballlarge.csv` to investigate how well we can predict the World Series winner at the beginning of the playoffs. The dataset has the following fields:

   - `Team`: A code for the name of the team
   - `League`: The Major League Baseball league the team belongs to, either AL (American League) or NL (National League)
   - `Year`: The year of the corresponding record
   - `Games`: The number of games a team played in that year
   - `W`: The number of regular season wins by the team in that year
   - `RS`: The number of runs scored by the team in that year
   - `RA`: The number of runs allowed by the team in that year
   - `OBP`: The on-base percentage of the team in that year
   - `SLG`: The slugging percentage of the team in that year
   - `BA`: The batting average of the team in that year
   - `Playoffs`: Whether the team made the playoffs in that year (1 for yes, 0 for no)
   - `RankSeason`: Among the playoff teams in that year, the ranking of their regular season records (1 is best)
   - `RankPlayoffs`: Among the playoff teams in that year, how well they fared in the playoffs. The team winning the World Series gets a `RankPlayoffs` of 1.

   (a) Each row in the baseball dataset represents a team in a particular year. Read the data into a dataframe called `baseballlarge`.

      i. How many team/year pairs are there in the whole dataset?

      ii. Though the dataset contains data from 1962 until 2012, we removed several years with shorter-than-usual seasons. Using the `table()` function, identify the total number of years included in this dataset.

      iii. Since we are only analyzing teams that made the playoffs, use the `subset()` function to create a smaller data frame limited to teams that made the playoffs. Your subsetted data frame should still be called `baseballlarge`. How many team/year pairs are included in the new dataset?

iv. Through the years, different numbers of teams have been invited to the playoffs. Find the different number of teams making the playoffs across the seasons.

(b) It is much harder to win the World Series if there are 10 teams competing for the championship versus just two. Therefore, we will add the predictor variable `NumCompetitors` to the data frame. `NumCompetitors` will contain the number of total teams making the playoffs in the year of a particular team/year pair. For instance, `NumCompetitors` should be 2 for the 1962 New York Yankees, but it should be 8 for the 1998 Boston Red Sox. We want to look up the number of teams in the playoffs for each team/year pair in the dataset, and store it as a new variable named `NumCompetitors` in the data frame. Do this. How many playoff team/year pairs are there in the dataset from years where 8 teams were invited to the playoffs?

(c) In this problem, we seek to predict whether a team won the World Series; in our dataset this is denoted with a RankPlayoffs value of 1. Add a variable named `WorldSeries` to the data frame that takes value 1 if a team won the World Series in the indicated year and a 0 otherwise. How many observations do we have in our dataset where a team did NOT win the World Series?

(d) When we are not sure which of our variables are useful in predicting a particular outcome, it is often helpful to build simple models, which are models that predict the outcome using a single independent variable. Which of the variables is a significant predictor of the `WorldSeries` variable in a logistic regression model? To determine significance, remember to look at the stars in the summary output of the model. We'll define an independent variable as significant if there is at least one star at the end of the coefficients row for that variable (this is equivalent to the probability column having a value smaller than 0.05). Note that you have to build multiple models (`Year, RS, RA, W, OBP, SLG, BA, RankSeason, NumCompetitors, League`) to answer this question (you can code the `League` variable as a categorical variable). Use the dataframe `baseballlarge` to build the models.

(e) In this question, we will consider multivariate models that combine the variables we found to be significant in (d). Build a model using all of the variables that you found to be significant in (d). How many variables are significant in the combined model?

(f) Often, variables that were significant in single variable models are no longer significant in multivariate analysis due to correlation between the variables. Are there any such variables in this example? Which of the variable pairs have a high degree of correlation (a correlation greater than 0.8 or less than -0.8)?

(g) Build all of the two variable models from (f). Together with the models from (d), you should have different logistic regression models. Which model has the best AIC value (the minimum AIC value)?

(h) Comment on your results.

2. In this question, we will build on the `Parole.csv` dataset from the previous week's exercise to build and validate a model that predicts if an inmate will violate the terms of his or her parole. Such a model could be useful to a parole board when deciding to approve or deny an application for parole.

   (a) Load the dataset `Parole.csv` into a data frame called `Parole`. How many parolees are contained in the dataset?

   (b) How many of the parolees in the dataset violated the terms of their parole?

   (c) Factor variables are variables that take on a discrete set of values and can be either unordered or ordered. Names of countries indexed by levels is an example of an unordered factor because there isn't any natural ordering between the levels. An ordered factor has a natural ordering between the levels (an example would be the classifications "large", "medium" and "small"). Which variables in this dataset are unordered factors with at least three levels? To deal with unordered factors in a regression model, the standard practice is to define one level as the "reference level" and create a binary variable for each of the remaining levels. In doing so, a factor with n levels is replaced by $n - 1$ binary variables. We will see this in question (e).

   (d) To ensure consistent training/testing set splits, run the following 5 lines of code: do not include the line numbers in the beginning)

   ```
   (1) > set.seed(144)
   (2) > library(caTools)
   (3) > split <- sample.split(Parole$Violator, SplitRatio = 0.7)
   (4) > train <- subset(Parole, split == TRUE)
   (5) > test <- subset(Parole, split == FALSE)
   ```

   Roughly what proportion of parolees have been allocated to the training and testing sets? Now, suppose you re-ran lines (1)-(5) again. What would you expect?

   - The exact same training/testing set split as the first execution of (1)-(5)
   - A different training/testing set split from the first execution of (1)-(5)

   If you instead ONLY re-ran lines (3)-(5), what would you expect?

   - The exact same training/testing set split as the first execution of (1)-(5)
   - A different training/testing set split from the first execution of (1)-(5)

   If you instead called 'set.seed()' with a different number and then re-ran lines (3)-(5), what would you expect?

   - The exact same training/testing set split as the first execution of (1)-(5)
   - A different training/testing set split from the first execution of (1)-(5)

   (e) If you tested other training/testing set splits in the previous section, please re-run the original 5 lines of code to obtain the original split. Using 'glm', train a logistic regression model on the training set. Your dependent variable is `Violator`, and you should use all the other variables as independent variables. What variables are significant in this model? Significant variables should have a least one star, or should have a p-value less than 0.05.

(f) What can we say based on the coefficient of the `MultipleOffenses` variable?

- Our model predicts that parolees who committed multiple offenses have 1.61 times higher odds of being a violator than the average parolee.
- Our model predicts that a parolee who committed multiple offenses has 1.61 times higher odds of being a violator than a parolee who did not commit multiple offenses but is otherwise identical.
- Our model predicts that parolees who committed multiple offenses have 5.01 times higher odds of being a violator than the average parolee.
- Our model predicts that a parolee who committed multiple offenses has 5.01 times higher odds of being a violator than a parolee who did not commit multiple offenses but is otherwise identical.

(g) Consider a parolee who is male, of white race, aged 50 years at prison release, from Kentucky, served 3 months, had a maximum sentence of 12 months, did not commit multiple offenses, and committed a larceny. According to the model, what are the odds this individual is a violator? According to the model, what is the probability this individual is a violator?

(h) Use the `predict()` function to obtain the model's predicted probabilities for parolees in the test set. What is the maximum predicted probability of a violation?

(i) In the following questions, evaluate the model's predictions on the test set using a threshold of 0.5. What is the model's sensitivity? What is the model's specificity? What is the model's accuracy?

(j) What is the accuracy of a simple model that predicts that every parolee is a non-violator?

(k) Consider a parole board using the model to predict whether parolees will be violators or not. The job of a parole board is to make sure that a prisoner is ready to be released into free society, and therefore parole boards tend to be particularily concerned with releasing prisoners who will violate their parole. Which of the following most likely describes their preferences and best course of action?

- The board assigns more cost to a false negative than a false positive, and should therefore use a logistic regression cutoff higher than 0.5.
- The board assigns more cost to a false negative than a false positive, and should therefore use a logistic regression cutoff less than 0.5.
- The board assigns equal cost to a false positive and a false negative, and should therefore use a logistic regression cutoff equal to 0.5.
- The board assigns more cost to a false positive than a false negative, and should therefore use a logistic regression cutoff higher than 0.5.
- The board assigns more cost to a false positive than a false negative, and should therefore use a logistic regression cutoff less than 0.5.

(l) Which of the following is the most accurate assessment of the value of the logistic regression model with a cutoff 0.5 to a parole board, based on the model's accuracy as compared to the simple baseline model?

- The model is of limited value to the board because it cannot outperform a simple baseline, and using a different logistic regression cutoff is unlikely to improve the model's value.
- The model is of limited value to the board because it cannot outperform a simple baseline, and using a different logistic regression cutoff is likely to improve the model's value.
- The model is likely of value to the board, and using a different logistic regression cutoff is unlikely to improve the model's value.
- The model is likely of value to the board, and using a different logistic regression cutoff is likely to improve the model's value.

(m) Using the `ROCR` package, what is the AUC value for the model?

(n) Describe the meaning of AUC in this context.

- The probability the model can correctly differentiate between a randomly selected parole violator and a randomly selected parole non-violator.
- The model's accuracy at logistic regression cutoff of 0.5.
- The model's accuracy at the logistic regression cutoff at which it is most accurate.

(o) Our goal has been to predict the outcome of a parole decision, and we used a publicly available dataset of parole releases for predictions. In this final problem, we will evaluate a potential source of bias associated with our analysis. It is always important to evaluate a dataset for possible sources of bias. The dataset contains all individuals released from parole in 2004, either due to completing their parole term or violating the terms of their parole. However, it does not contain parolees who neither violated their parole nor completed their term in 2004, causing non-violators to be underrepresented. This is called "selection bias" or "selecting on the dependent variable," because only a subset of all relevant parolees were included in our analysis, based on our dependent variable in this analysis (parole violation). How could we improve our dataset to best address selection bias?

- There is no way to address this form of biasing.
- We should use the current dataset, expanded to include the missing parolees. Each added parolee should be labeled with Violator=0, because they have not yet had a violation.
- We should use the current dataset, expanded to include the missing parolees. Each added parolee should be labeled with Violator=NA, because the true outcome has not been observed for these individuals.
- We should use a dataset tracking a group of parolees from the start of their parole until either they violated parole or they completed their term.

3. Credit scoring rules are used to determine if a new applicant should be classified as a good credit risk or a bad credit risk, based on values for one or more of the predictor variables. Lenders such as banks and credit card companies use credit scores to determine if money should be lent to consumers. The file `germandcredit.csv` contains data on 1000 past credit applicants from Germany. This data is obtained from the UCI Machine Learning Repository. In this question, we want to develop a model that may be used to determine if new applicants present a good credit risk or a bad credit risk. The dataset contains the following variables:

    - `chkacct`: Status of existing checking account. Variable is coded as:
      0: < 0 DM (Deutsche Mark)
      1: $0 \leq \ldots < 200$ DM
      2: ≥ 200 DM/salary assignments for at least 1 year
      3: no checking account
    - `dur`: Duration of credit in months
    - `hist`: Credit history. Variable is coded as:
      0: no credits taken
      1: all credits at this bank paid back duly
      2: existing credits paid back duly till now
      3: delay in paying off in the past
      4: critical account
    - `newcar`: Purpose of credit (new car). Variable is coded as: 0: No, 1: Yes
    - `usedcar`: Purpose of credit (used car). Variable is coded as: 0: No, 1: Yes
    - `furn`: Purpose of credit (used furniture/equipment). Variable is coded as: 0: No, 1: Yes
    - `radiotv`: Purpose of credit (radio/television). Variable is coded as: 0: No, 1: Yes
    - `educ`: Purpose of credit (education). Variable is coded as: 0: No, 1: Yes
    - `retrain`: Purpose of credit (retraining). Variable is coded as: 0: No, 1: Yes
    - `amt`: Credit amount
    - `sav`: Average balance in savings account. Variable is coded as:
      0: < 100 DM
      1: $100 \leq \ldots < 500$ DM
      2: $500 \leq \ldots < 1000$ DM
      3: ≥ 1000 DM
      4: unknown/ no savings account
    - `emp`: Present employment since. Variable is coded as:
      0: unemployed
      1: < 1 year
      2: $1 \leq \ldots < 4$ years
      3: $4 \leq \ldots < 7$ years
      4: ≥ 7 years

- `instrate`: Installment rate in percentage of disposable income
- `malediv`: Applicant is male and divorced. Variable is coded as: 0: No, 1: Yes
- `malesingle`: Applicant is male and single. Variable is coded as: 0: No, 1: Yes
- `malemarwid`: Applicant is male and married or a widower. Variable is coded as: 0: No, 1: Yes
- `coapp`: Applicant has a co-applicant. Variable is coded as: 0: No, 1: Yes
- `guar`: Applicant has a guarantor. Variable is coded as: 0: No, 1: Yes
- `presres`: Present resident since in years. Variable is coded as:
  0: ≤ 1 year
  1: $1 < \ldots \leq 2$ years
  2: $2 < \ldots \leq 3$ years
  3: > 4 years
- `realest`: Applicant owns real estate. Variable is coded as: 0: No, 1: Yes
- `propnone`: Applicant owns no property (or unknown). Variable is coded as: 0: No, 1: Yes
- `age`: Age in years
- `other`: Applicant has other installment plan credit. Variable is coded as: 0: No, 1: Yes
- `rent`: Applicant rents. Variable is coded as: 0: No, 1: Yes
- `ownres`: Applicant owns residence. Variable is coded as: 0: No, 1: Yes
- `numcred`: Number of existing credits at this bank
- `job`: Nature of job. Variable is coded as:
  0: unemployed/unskilled - non-resident
  1: unskilled - resident
  2: skilled employee/official
  3: management/self-employed/highly qualified employee/officer
- `numdep`: Number of people for whom liable to provide maintenance
- `tel`: Applicant has phone in his or her name. Variable is coded as: 0: No, 1: Yes
- `foreign`: Foreign worker. Variable is coded as: 0: No, 1: Yes
- `resp`: Credit rating is good. Variable is coded as: 0: No, 1: Yes

(a) Read the data into the dataframe `germancredit`. We are interested in predicting the `resp` variable. Obtain a random training/test set split with:

```
> set.seed(2019)
> library(caTools)
> spl <- sample.split(germancredit$resp, 0.75)
```

Split the data frame into a training data frame called "training" using the observations for which `spl` is TRUE and a test data frame called "test" using the observations for which spl is FALSE. Why do we use the sample.split() function to split into a training and testing set?

- It is the only method in R to randomly split the data.
- It balances the independent variables between the training and testing sets.
- It balances the dependent variable between the training and testing sets.

Select the best option.

(b) We start with the simplest logistic regression model to predict credit risk in the training set using no predictor variables except the constant (intercept). Write down the fitted model.

(c) Provide a precise mathematical relationship between the estimated coefficient and the fraction of respondents with a good credit rating in the training set.

(d) We now develop a logistic regression model to predict credit card default using all the possible predictor variables. Identify all variable that are significant at the 10% level.

(e) What is the log likelihood value for this model?

(f) Compute the confusion matrix on the test set. For the logistic regression model use a threshold of 0.5.

(g) What is the accuracy of the model?

(h) Redo the logistic regression model to predict credit risk using only the predictor variables that were significant at the 10% level in (d). What is the AIC value for this model?

(i) Based on the AIC, which model is preferable?

(j) Compute the confusion matrix on the test set for the model in (h). For the logistic regression model use a threshold of 0.5.

(k) Based on the fraction of people who are predicted as good credit risk but are actually bad credit risk in the test set, which model is preferable?

(l) Based on the fraction of people who are predicted as bad credit risk but are actually good credit risk in the test set, which model is preferable?

(m) Based on the area under the curve in the test set, which model is preferable?

(n) From this point onwards, we use the model with all the predictor variables included. We now consider a more sophisticated way to evaluate the consequence of misclassification. The consequences of misclassification by the credit company is assessed as follows: the costs of incorrectly saying an applicant is a good credit risk is 300 DM while the profit of correctly saying an applicant is a good credit risk is 100 DM. In terms of profit this can be considered in terms of a table as follows:

	Actual Bad	Actual Good
Predicted Bad	0	0
Predicted Good	-300 DM	100 DM

What is the total profit incurred by the credit company on the test set?

(o) To see if we can improve the performance by changing the threshold, we will use the predicted probability of credit risk from the logistic regression as a basis by selecting the good credit risks first, followed by poorer risk applicants. Sort the test set on the predicted probability of good credit risk from high to low (Hint: You can use the `sort()` command). What is the duration of credit in months for the individual with the lowest predicted probability of good credit risk?

(p) For each observation in the sorted test set, calculate the actual profit of providing credit (use the table in (n)). Compute the net profit by adding a new variable that captures the cumulative profit. How many far down the test set do you need to go in order to get the maximum profit? (Hint. You can use the index from the `index.return` argument in the `sort` function and use the `cumsum` function)

(q) If the logistic regression model from (p) is scored to future applicants, what "probability of good credit risk" cutoff should be used in extending credit?

4. There was significant interest in the results of the U.S. elections in November 2020. In this question you will build a model that would have helped predict the winner of the U.S. presidential elections using data which was available before the elections. The model will use past election outcomes and the state of the economy to predict how people might vote. The data is provided in the file `presidential.csv` and contains the following variables:

   - `YEAR`: Year of the U.S. presidential election
   - `DEM`: Name of Democratic nominee
   - `REP`: Name of Republican nominee
   - `INC`: Incumbent (party in power) leading up to that election (1 = Democratic, -1 = Republican)
   - `RUN`: Variable to indicate if the incumbent president is running for the presidential election again (1 = Democratic incumbent president is running, -1 = Republican incumbent president is running, 0 = otherwise. For example after becoming president in 2008, Obama ran for presidential elections again as a Democrat in 2012 implying that the corresponding entry is 1.)
   - `DUR`: Duration of the current party in power in the White House (0 = Incumbent has been in power only for one term before the election, 1 (-1) if the Democratic (Republican) party has been in the White House for two consecutive terms, 1.25 (-1.25) if the Democratic (Republican) party has been in the White House for three consecutive terms, 1.50 (-1.50) if the Democratic (Republican) party has been in the White House for four consecutive terms, and so on)
   - `GROWTH`: Growth rate of the real per capita GDP in the year of the election (%)
   - `GOOD`: Number of good quarters (in terms of performance in the growth rate of the real per capita GDP) in the first fifteen quarters of the current administration
   - `WIN`: Winner of the presidential election (1 = Democratic, -1 = Republican)

   (a) Read the dataset into the dataframe `pres`. In the elections starting from 1916 up to and including 2016, which party has won more presidential elections? How many elections has that party won?

   (b) Who among the nominees have represented the Democrats and the Republicans in the most number of presidential elections? How many times have they respectively done so?

   (c) Use a two-sided t-test to verify if there is evidence to show that the number of good quarters when the president is Republican is different from the number of good quarters when the president is Democratic. What is the p-value of the test and your conclusion?

   (d) Define a new variable `WININC` that takes a value of 1 if the presidential nominee of the incumbent party wins and 0 otherwise.

   (e) How many times did the presidential nominee from the incumbent party win and how many times did the presidential nominee from the incumbent party lose?

(f) Perform a simple logistic regression to predict the WININC variable using the GROWTH variable and the constant. What is the log-likelihood value for the model?

(g) The GROWTH variable is:

- Significant at the 0.001 level
- Significant at the 0.01 level
- Significant at the 0.05 level
- Significant at the 0.1 level
- Insignificant

(h) Unlike questions (d) to (g) which looked at the incumbent party's winning chances, from this point onwards, we are going to predict the chances of the Democratic party nominee winning in the presidential election. To do this, transform the variables as follows:

i. Transform the WIN variable to be 1 when the presidential winner is a Democrat and 0 when the winner is a Republican.

ii. Transform the GROWTH variable as follows: When the growth rate is positive (say 4.623) and the Republican party is incumbent, we should transform it to a negative value -4.623 since this should have a negative effect on the Democratic nominee's chances of winning while if the growth rate is negative (say -4.623) and the Republican party is incumbent, we should transform it to positive 4.623 since this should have a positive effect on the Democratic nominee's chances of winning.

(i) Repeat step ii in question (h) for the GOOD variable. You are now ready to develop a logistic regression model for the WIN variable using the predictor variables INC, RUN, DUR, GROWTH, GOOD and the constant (intercept). Use all the observations to build your model. What is the AIC of the model?

(j) Among the predictor variables INC, RUN, DUR, GROWTH, GOOD and the constant (intercept), identify the three least significant variables?

(k) Drop the three variables identified in question (j) and rebuild you logistic regression model. What is the AIC of the new model?

(l) In this new model, what is the smallest significance level at which you would reject the null hypothesis that the coefficient for the variable DUR is zero? Suppose, we now decide to use a level of 0.10, what would your suggestion be?

(m) Which among the two models that you have developed in questions (i) and (k) do you prefer? Explain your reasons briefly.

(n) We will now evaluate the probability of Biden winning the 2020 election with this model where Biden is the Democratic nominee and Trump is the Republican nominee. What should be the corresponding INC, RUN and DUR variables?

(o) The projected growth rate for the US economy for 2020 is -5% (possibly worse). Based on this, what is the probability of Joe Biden winning in the 2020 election based on the model you developed in question (k)?