BANA 7042 PROJECT

Name: UCID:

Background: Flight landing.

Motivation: To reduce the risk of landing overrun.

Goal: To study what factors and how they would impact the landing distance of a commercial flight.

Data: Landing data (landing distance and other parameters) from 950 commercial flights (not real data set but simulated from statistical models). See two Excel files 'FAA-1.xls' (800 flights) and 'FAA-2.xls' (150 flights).

Variable dictionary:

Aircraft: The make of an aircraft (Boeing or Airbus).

Duration (in minutes): Flight duration between taking off and landing. The duration of a normal flight should always be greater than 40min.

No_pasg: The number of passengers in a flight.

Speed_ground (in miles per hour): The ground speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.

Speed_air (in miles per hour): The air speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.

Height (in meters): The height of an aircraft when it is passing over the threshold of the runway. The landing aircraft is required to be at least 6 meters high at the threshold of the runway.

Pitch (in degrees): Pitch angle of an aircraft when it is passing over the threshold of the runway.

Distance (in feet): The landing distance of an aircraft. More specifically, it refers to the distance between the threshold of the runway and the point where the aircraft can be fully stopped. The length of the airport runway is typically less than 6000 feet.

Part 1. Practice of modeling the landing distance using linear regression.

Please write R programs to complete the following steps. In each step, provide

- The R code (how do you realize it?)
- The R output (Copy and paste only those relevant)
- Your observations (What do you observe from the output?)
- Your conclusion/decision

Initial exploration of the data

Step 1. Read the two files 'FAA-1.xls' (800 flights) and 'FAA-2.xls' into your R system. Please search "Read Excel files from R" in Google in case you do not know how to do that.

Step 2. Check the structure of each data set using the "str" function. For each data set, what is the sample size and how many variables? Is there any difference between the two data sets?

Step 3. Merge the two data sets. Are there any duplications? Search "check duplicates in r" if you do not know how to check duplications. If the answer is "Yes", what action you would take?

Step 4. Check the structure of the combined data set. What is the sample size and how many variables? Provide summary statistics for each variable.

Step 5. By now, if you are asked to prepare ONE presentation slide to summarize your findings, what observations will you bring to the attention of FAA agents?

Please list no more than five using “bullet statements”, from the most important to the least important.

Data Cleaning and further exploration

Step 6. Are there abnormal values in the data set? Please refer to the variable dictionary for criteria defining “normal/abnormal” values. Remove the rows that contain any “abnormal values” and report how many rows you have removed.

Step 7. Repeat Step 4.

Step 8. Since you have a small set of variables, you may want to show histograms for all of them.

Step 9. Prepare another presentation slide to summarize your findings drawn from the cleaned data set, using no more than five “bullet statements”.

Initial analysis for identifying important factors that impact the response variable “landing distance”

Step 10. Compute the pairwise correlation between the landing distance and each factor X. Provide a table that ranks the factors based on the size (absolute value) of the correlation. This table contains three columns: the names of variables, the size of the correlation, the direction of the correlation (positive or negative). We call it Table 1, which will be used for comparison with our analysis later.

Step 11. Show X-Y scatter plots. Do you think the correlation strength observed in these plots is consistent with the values computed in Step 10?

Step 12. Have you included the airplane make as a possible factor in Steps 10-11? You can code this character variable as 0/1.

Regression using a single factor each time

Step 13. Regress Y (landing distance) on each of the X variables. Provide a table that ranks the factors based on its significance. The smaller the p-value, the more significant the factor. This table contains three columns: the names of variables, the size of the p-value, the direction of the regression coefficient (positive or negative). We call it Table 2.

Step 14. Standardize each X variable. In other words, create a new variable

X'= {X-mean(X)}/sd(X).

The mean of X' is 0 and its standard deviation is 1.

Regress Y (landing distance) on each of the X' variables. Provide a table that ranks the factors based on the size of the regression coefficient. The larger the size, the more important the factor. This table contains three columns: the names of variables, the size of the regression coefficient, the direction of the regression coefficient (positive or negative). We call it Table 3.

Step 15. Compare Tables 1,2,3. Are the results consistent? At this point, you will meet with a FAA agent again. Please provide a single table than ranks all the factors based on their relative importance in determining the landing distance. We call it Table 0.

Check collinearity

Step 16. Compare the regression coefficients of the three models below:

Model 1: LD ~ Speed_ground

Model 2: LD ~ Speed_air

Model 3: LD ~ Speed_ground + Speed_air

Do you observe any significance change and sign change? Check the correlation between Speed_ground and Speed_air. You may want to keep one of them in the model selection. Which one would you pick? Why?

Variable selection based on our ranking in Table 0.

Step 17. Suppose in Table 0, the variable ranking is as follows: X1, X2, X3….. Please fit the following six models:

Model 1: LD ~ X1

Model 2: LD ~ X1 + X2

Model 3: LD ~ X1 + X2 + X3

…….

Calculate the R-squared for each model. Plot these R-squared values versus the number of variables p. What patterns do you observe?

Step 18. Repeat Step 17 but use adjusted R-squared values instead.

Step 19. Repeat Step 17 but use AIC values instead.

Step 20. Compare the results in Steps 18-19, what variables would you select to build a predictive model for LD?

Variable selection based on automate algorithm.

Step 21. Use the R function “StepAIC” to perform forward variable selection. Compare the result with that in Step 19.