

**BUDT 758T**

**assignment #1**

The goal of this homework is to review the fundamental concepts of regression modeling learned in earlier classes. You will also review (or being working with) the statistical programming language R. So yes, you must complete this assignment in R and include all code used to solve the problems!

**The Data**

The data in the Airlines data file (“Airlines\_Data.csv” posted on Canvas) contains data from 3,999 airline customers enrolled in East-West Airlines’ customer rewards program. (Note that while East-West Airlines is clearly fictional, this is data from a real airlines reward program; names have been changed to protect the innocent and not-so-innocent alike.) East-West Airlines has two goals with this analysis: (1) identifying if a customer will claim a travel award using their rewards, and (2) identifying factors that lead to customers claiming a travel award. The data contains information about each customer’s history:

|  |  |
| --- | --- |
| **Field Name** | **Description** |
| *ID* | Unique ID |
| *Balance* | Number of miles eligible for award travel |
| *Qual\_miles* | Number of miles counted as qualifying for Topflight status |
| *cc1\_miles* | Number of miles earned with freq. flyer credit card in the past 12 months: |
| *cc2\_miles* | Number of miles earned with Rewards credit card in the past 12 months: |
| *cc3\_miles* | Number of miles earned with Small Business credit card in the past 12 months: |
| *note: miles are binned* | 1 = under 5,000 |
|  | 2 = 5,000 - 10,000 |
|  | 3 = 10,001 - 25,000 |
|  | 4 = 25,001 - 50,000 |
|  | 5 = over 50,000 |
| *Bonus\_miles* | Number of miles earned from non-flight bonus transactions in the past 12 months |
| *Bonus\_trans* | Number of non-flight bonus transactions in the past 12 months |
| *Flight\_miles\_12mo* | Number of flight miles in the past 12 months |
| *Flight\_trans\_12* | Number of flight transactions in the past 12 months |
| *Days\_since\_enroll* | Number of days since Enroll\_date |
| *Award* | Dummy variable for travel award claimed (1 = award claimed, 0 = not claimed) |

There are two goals for this assignment. The primary goal is to see if we can model the number of miles eligible for award travel (*Balance*) using linear regression. Secondly, we would like to consider modeling if someone will/will not claim a travel award (*Award*) using regression.

**Assignment**

Please answer all questions in the dedicated space and upload on Canvas. Please ensure that your numbering of questions matches those below. Include any R code you used to answer each question with your response. You are welcome to include any output you wish to provide either with the appropriate question or at the end of your assignment in an appendix.

Remember: you are allowed to consult with others in the class on this assignment, but all submitted work must be your own (and don’t forget to include the names of anyone you consulted in the last question!).

1. **30 points: Working with data and exploratory analysis**
   1. Load the Airlines data into R and attach it to your R session.

airlines = read.csv("Airlines\_Data.csv",header=T)

Alternatively, we could also do this by directly importing the data-set from the working directory.

* 1. Change the *cc3\_miles* variable to a factor variable using **as.factor()** in R. Use this new variable for the remainder of the assignment.

cc3\_miles\_factor = as.factor(airlines$cc3\_miles)

* + 1. Why is this necessary for the problem? (That is, in terms of the problem we are investigating, why do we need to change *cc3\_miles* to a factored variable?)

Because the values don’t specify the actual miles but the category in which they fall.

* + 1. Why is this necessary for R if we are going to run regressions with this data?

Changing it to a factor automically inputs the dummy variables in R when running the regressions and hence helps in looking at each category separately.

* 1. Create two new variables for the data set:
     1. *Years\_since\_enroll:* the number of years since the customer enrolled in the program (note this does not need to be a rounded number; you can do partial years, if you wish)

Years\_since\_enroll = (airlines$Days\_since\_enroll)/365

* + 1. *Loyal\_customer*: a variable that equals 1 if the customer enrolled more than five years ago and equals 0 if the customer did not

Loyal\_customer = ifelse(Years\_since\_enroll > 5,1,0)

* 1. Calculate the mean and median of the *Balance* variable. Does this concern you? Why or why not?

mean(airlines$Balance) 73601.33

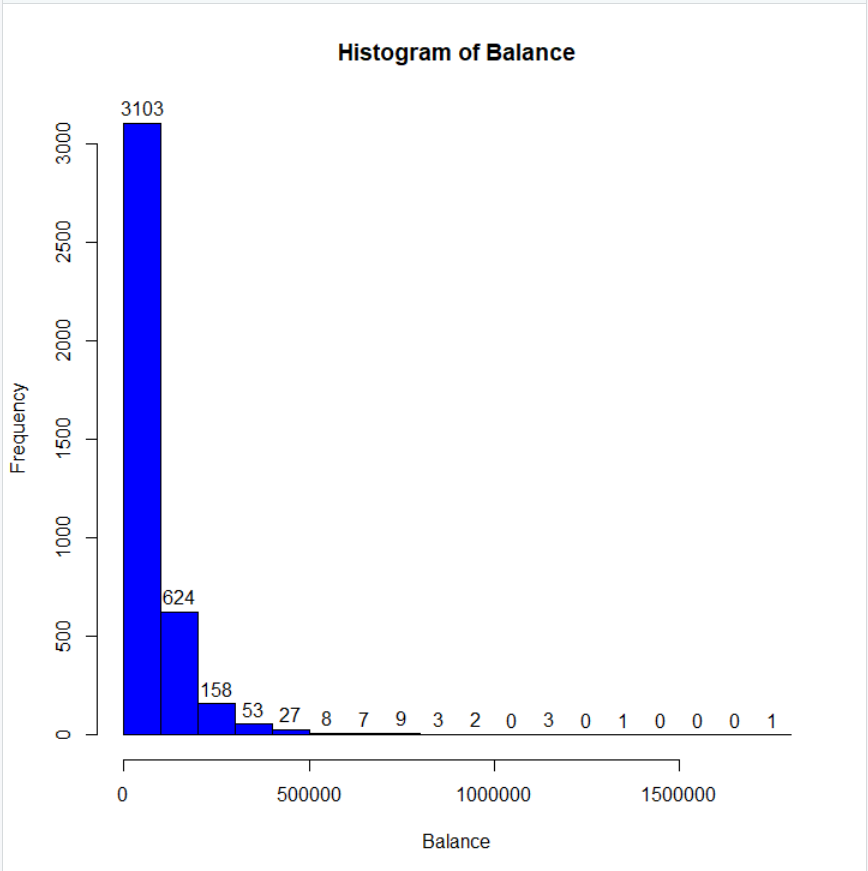
median(airlines$Balance) 43097

Yes, it does concern me because the mean seems to be very large compared to the median which gives us a hint that there could be some very large or wrong entries in the data-set.

* 1. Present a blue histogram of the *Balance* variable.

hist1 = hist(airlines$Balance, col='blue', breaks=15, xlab='Balance',

main='Histogram of Balance',labels = TRUE)



* + 1. Does this agree with what you found in part (d)? Why or why not?

Yes, because the Balance for the majority of the people is under 100000 miles and hence the median is small ; but the Balance of the remaining people is much higher, it has influenced and increased the mean.

* + 1. Is this likely to cause problems if we use *Balance* in a regression? Why or why not?

Yes, it may cause problems like predicting some very wrong Balances.

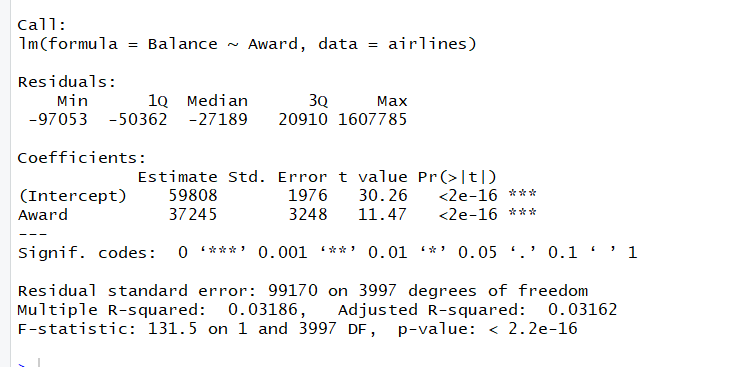
* 1. Calculate the mean of *Award*. What percentage of the customers in this data set did *not* claim a travel award?

mean(airlines$Award) 0.3703426

(1 - sum(airlines$Award)/3999)\*100 62.97%

1. **15 points: Basic regression in R**
   1. Run a simple linear regression in R using *Award* to predict *Balance.* Report the summary of the regression. Do you consider this to be a useful model?

reg1 = lm(Balance ~ Award, data=airlines)



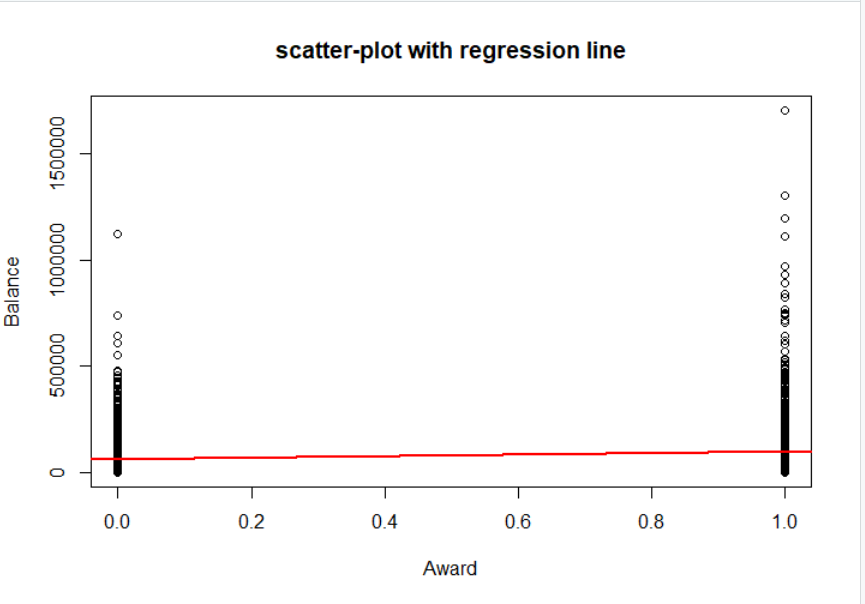
No, it’s not a useful model because it has very low adjusted R-squared value and may be we should consider other factors as well.

* 1. Create and report a scatterplot with the variable *Balance* on the y-axis and *Award* on the x-axis. Add the regression line from the model in part (a) to the plot in red with a line thickness of 2. Does your plot seem to agree with your conclusion about the usefulness of the model in part (a)?

plot(airlines$Award, airlines$Balance, xlab = 'Award', ylab = 'Balance',

main='scatter-plot with regression line')

abline(reg1, lwd=2, col='red')



Yes, the plot agrees with conclusion that the model is not that useful because the regression line is continuous but the actual Balance values are discrete taking just the values 0 and 1.

* 1. State precisely what effect the value of *Award* has on the predicted *Balance*.

On average and keeping everything else constant, the Balance increases by 37245 miles when the Award is claimed compared to the Balance when the Award is not claimed.

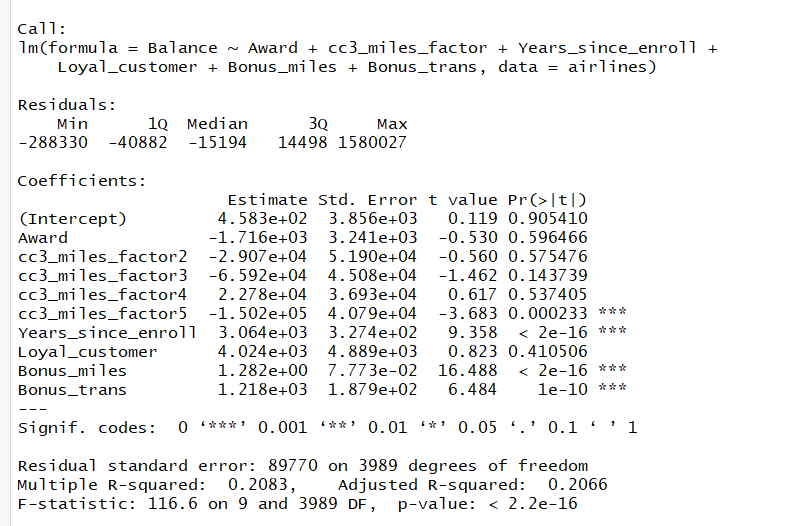
1. **40 points: Multivariate regression in R**
   1. Run a multiple linear regression in R using *Award, cc3\_miles, Years\_since\_enroll, Loyal\_customer, Bonus\_miles,* and *Bonus\_trans* to predict *Balance.* Report the summary of the regression.

airlines = data.frame(airlines, Years\_since\_enroll, Loyal\_customer,

cc3\_miles\_factor)

reg2 = lm(Balance ~ Award + cc3\_miles\_factor + Years\_since\_enroll +

Loyal\_customer + Bonus\_miles + Bonus\_trans, data = airlines)



Is this a better or worse model than your model from Question 2? Justify your answer.

It is a better model than before because it has a better adjusted R-sqaure value

* + 1. Without any more calculations, do you have any potential concerns about multicollinearity with this model? Explain.

Yes, there could be potential concerns about multicollinearity with this model because as many as 5 out of 9 independent variables are statistically insignificant. Also because Loyal\_customer is derived from Years\_since\_enroll, so they tend to move together and collinearity could exist.

* 1. Consider *Years\_since\_enroll* in the regression summary.
     1. Is the number of years a customer has been enrolled significant in predicting their balance? Explain.

Yes, Years\_since\_enroll is significant in predicting the balance because it has very low p-value almost close to zero.

* + 1. State precisely the effect number of years since enrollment has on a customer’s predicted balance, according to this model.

On average and keeping everything else constant, the Balance increases by 3064 miles for every year that increases in the enrollment of a customer.

* 1. Consider *cc3\_miles* in the regression summary.
     1. Is the number of miles a customer earned with their Small Business credit card in the past 12 months significant in predicting their balance? Explain.

Yes, because cc3\_miles\_5 has a very small p-value and hence it is significant. (So, this means that when cc3\_miles becomes greater than 50000 miles, cc3\_miles becomes statistically significant)

* + 1. State precisely the effect of earning over 50,000 miles with a Small Business credit card has on predicted balance, according to this model.

On average and keeping everything else constant, the Balance decreases by 150200 miles for every one mile increased above 50,000 when earned with a Small Business credit card compared to increase of the Balance of a customer with less than 5,000 miles.

* 1. What is the predicted balance for a customer who has:

*Award = 0*

*cc3\_miles = 3*

*Years\_since\_enroll = 6*

*Loyal\_customer = 1*

*Bonus\_miles = 20000*

*Bonus\_trans = 20*

By plugging in the given values in the regression equation we get the required

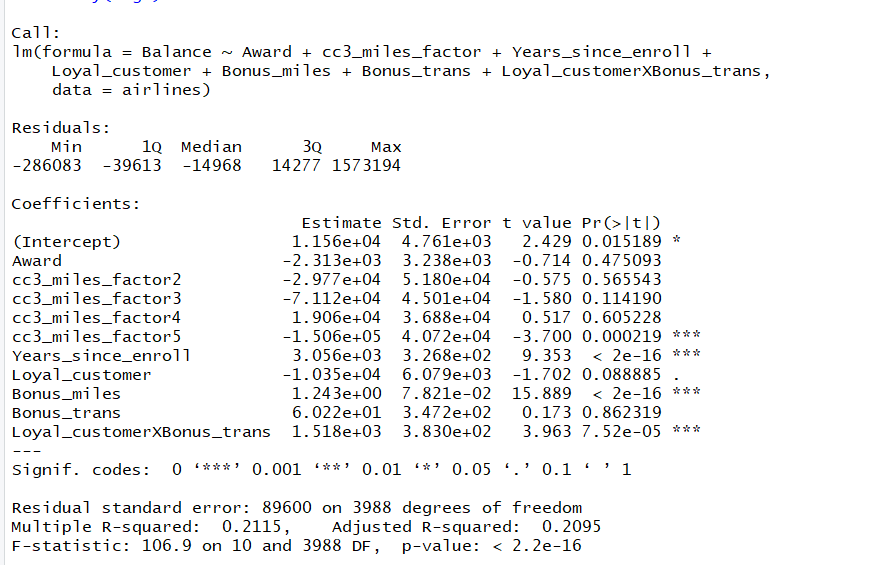
Predicted balance as 6946.3 miles.

* 1. A senior consultant has indicated that a customer’s balance is more affected by being a loyal customer when that customer also has a large number of bonus non-flight transactions in the last 12 months. Add this domain knowledge to your regression model from (a) and run a new multivariate linear regression.

Loyal\_customerXBonus\_trans = airlines$Loyal\_customer\*airlines$Bonus\_trans

airlines = data.frame(airlines, Loyal\_customerXBonus\_trans)

reg3 = lm(Balance ~ Award + cc3\_miles\_factor + Years\_since\_enroll + Loyal\_customer + Bonus\_miles + Bonus\_trans + Loyal\_customerXBonus\_trans, data = airlines)



* + 1. Does the consultant appear to be right? Justify your answer.

Yes, the consultant appears to be right because the variable Loyal\_customerXBonus\_trans has a very low p-value and hence is statistically significant and also the new model has a better adjusted R-square than before.

* + 1. State precisely what effects the number of bonus non-flight transactions now has on *Balance*.

Keeping everything else constant and on average, for a loyal customer Balance increases by 1578.22 miles and for a non loyal customer, Balance increases by 60.22 miles for every unit increase in bonus non-flight transcations.

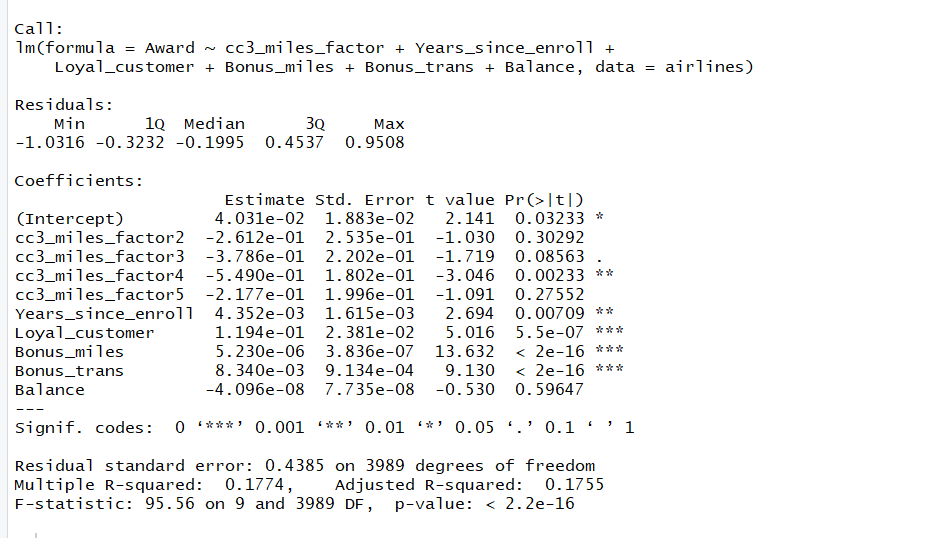
1. **15 points: Categorical Prediction in R**
   1. The airline CEO is now wondering if you can predict *Award* using the other variables in the data set. Consider the scatterplot you created in Question 2, part (b). Does it seem appropriate to use linear regression to model this problem? Why or why not?

No, it doesn’t make sense because we are predicting a categorical variable that can take only values 1 and 0 and hence logistic regression would be more suitable.

* 1. Run a multiple linear regression to predict *Award* using the variables from Question 3: *cc3\_miles, Years\_since\_enroll, Loyal\_customer, Bonus\_miles, Bonus\_trans,* and *Balance* and report the summary.

reg4 = lm(Award ~ cc3\_miles\_factor + Years\_since\_enroll + Loyal\_customer +

Bonus\_miles + Bonus\_trans + Balance, data=airlines)



c. What would your model predict for an entirely new customer (0 values for every variable) claiming a travel award? Does this make sense? Explain your answer.

Predicted value is 0.04. No it doesn’t make sense because Award can only take value of 0 or 1 and can’t take the value of 0.4