****

**BUDT758T  
  
DATA MINING AND PREDICTIVE ANALYTICS**

**Homework 1**

**NAME (in capitals): ROHIT KOTHAVADE**

* Please see the instructions at <https://docs.google.com/document/d/1uwOFS-LVKDBAzjEonmfggJMAmNWhWxAJRbL71nsVg4A/edit?usp=sharing> and submit on Canvas.
* Your submission should consist of this document (with the answers filled in the appropriate places).
* Please ensure that answers are appropriately numbered and clearly legible.
* In the space below please enter the following text and initial below: “I pledge on my honor that I have not given or received unauthorized assistance on this assignment.”

|  |
| --- |
| HONOR PLEDGE: I pledge on my honor that I have not given or received unauthorized assistance on this assignment.    YOUR INITIALS: RVK |

This is an individual assignment. Your submission must represent your own work.

The goal of this homework is to review the fundamental concepts of regression modeling learned in earlier classes. You will also begin to develop facility with the statistical programming language *R*.

**The Assignment**

The data in the accompanying file “Airline data 2.csv” (posted on Canvas) was assembled by Professor Robert Windle of the Smith School with assistance from Oliver Yao. The file contains information on 627 air routes in the United States. A route refers to a pair of airports. Note that some cities are served by more than one airport. In such cases, the airports are distinguished by their 3-letter code. The data was collected for the third quarter of 1996 (3Q96). The variables in the data set are:

1. S\_CODE: starting airport’s code
2. S\_CITY: starting city
3. E\_CODE: ending airport’s code
4. E\_CITY: ending city
5. COUPON: average number of coupons (a one-coupon flight is a non-stop flight, a two-coupon flight is a one stop flight, etc.) for that route
6. NEW: number of new carriers entering that route between Q3-96 and Q2-97
7. VACATION: whether a vacation route (Yes) or not (No); Florida and Las Vegas routes are generally considered vacation routes
8. SW: whether Southwest Airlines serves that route (Yes) or not (No)
9. HI: Herfindahl Index – airlines use this as a measure of market concentration (a larger value indicates greater concentration)
10. S\_INCOME: starting city’s average personal income
11. E\_INCOME: ending city’s average personal income
12. S\_POP: starting city’s population
13. E\_POP: ending city’s population
14. SLOT: whether either endpoint airport is slot controlled or not; this is a measure of airport congestion
15. GATE: whether either endpoint airport has gate constraints or not; this is another measure of airport congestion
16. DISTANCE: distance between two endpoint airports in miles
17. PAX: number of passengers on that route during period of data collection
18. FARE: average fare on that route

There are two goals of the study. The first is to predict the FARE as a function of the other variables. The second is to determine how the presence of Southwest Airlines affects fares.

We will **not** use the first four attributes (S\_CODE, S\_CITY, E\_CODE, and E\_CITY) in our analysis.

**The Assignment**

Please answer all questions in the dedicated space and upload on Canvas. Please ensure that your numbering of questions matches those below.

1. **Working with data and regression in R**The VACATION, SW, GATE and SLOT variables will have type *Factor*. The dollar sign before FARE is likely to create a problem – you may choose to address this in *R*, or directly in *Excel* by changing the format from currency to number. Using the resulting new dataset run a multivariable regression for FARE, with all numerical variables (i.e. of type num or int) and the four factors above as independent variables.

setwd("C:/Users/Rohit/Documents/Spring Sem 2/Data mining/Assignment 1")  
dfKK <- read.csv("Individual Assignment 1 Airline Data v2.csv")  
dfKK$FARE<-gsub(pattern="\\$",replacement="",x=dfKK$FARE)  
dfKK$FARE <- as.integer(dfKK$FARE)  
dfKK$S\_INCOME<-gsub(pattern="\\$",replacement="",x=dfKK$S\_INCOME)  
dfKK$S\_INCOME<-gsub(pattern=",",replacement="",x=dfKK$S\_INCOME)  
dfKK$S\_INCOME <- as.integer(dfKK$S\_INCOME)  
dfKK$E\_INCOME<-gsub(pattern="\\$",replacement="",x=dfKK$E\_INCOME)  
dfKK$E\_INCOME<-gsub(pattern=",",replacement="",x=dfKK$E\_INCOME)  
dfKK$E\_INCOME <- as.integer(dfKK$E\_INCOME)  
  
  
##Question 1a  
fit<-lm(FARE~COUPON+NEW+VACATION+SW+HI+S\_INCOME+E\_INCOME+S\_POP+E\_POP+SLOT+GATE+DISTANCE+PAX,data=dfKK)  
summary(fit)

##   
## Call:  
## lm(formula = FARE ~ COUPON + NEW + VACATION + SW + HI + S\_INCOME +   
## E\_INCOME + S\_POP + E\_POP + SLOT + GATE + DISTANCE + PAX,   
## data = dfKK)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -110.953 -23.033 -1.141 20.980 128.622   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.001e+01 2.861e+01 -0.350 0.726565   
## COUPON 6.678e+00 1.234e+01 0.541 0.588555   
## NEW -2.396e+00 1.898e+00 -1.263 0.207232   
## VACATIONYes -3.297e+01 3.779e+00 -8.727 < 2e-16 \*\*\*  
## SWYes -4.078e+01 3.766e+00 -10.829 < 2e-16 \*\*\*  
## HI 8.134e-03 1.005e-03 8.097 3.07e-15 \*\*\*  
## S\_INCOME 1.493e-03 5.264e-04 2.836 0.004719 \*\*   
## E\_INCOME 1.830e-03 4.132e-04 4.428 1.13e-05 \*\*\*  
## S\_POP 3.614e-06 6.589e-07 5.485 6.07e-08 \*\*\*  
## E\_POP 4.216e-06 7.621e-07 5.533 4.69e-08 \*\*\*  
## SLOTFree -1.497e+01 3.913e+00 -3.826 0.000144 \*\*\*  
## GATEFree -2.053e+01 4.048e+00 -5.073 5.22e-07 \*\*\*  
## DISTANCE 7.363e-02 3.631e-03 20.278 < 2e-16 \*\*\*  
## PAX -9.211e-04 1.475e-04 -6.247 7.87e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 35.46 on 609 degrees of freedom  
## (4 observations deleted due to missingness)  
## Multiple R-squared: 0.791, Adjusted R-squared: 0.7866   
## F-statistic: 177.3 on 13 and 609 DF, p-value: < 2.2e-16

* 1. What is the resulting R2?

Ans – 0.7866

* 1. State precisely what effect the value of SLOT has on the predicted FARE:

Ans – When all the other parameters are kept same, and SLOT = Free, the FARE reduces by $14.97

* 1. What is the predicted fare of a leg that has COUPON = 1, NEW = 3, VACATION = No, SW = No, HI =6000, S\_INCOME = $2000, E\_INCOME = $2000, S\_POP = 4,000,000, E\_POP=7,150,000, SLOT=Free and GATE = Constrained, DISTANCE = 1000, and PAX = 6000?

Ans – $142.66  
fit2<-lm(FARE~COUPON+NEW+VACATION+SW+HI+S\_INCOME+E\_INCOME+S\_POP+E\_POP+SLOT+GATE+DISTANCE+PAX,data=dfKK)  
newdata <- data.frame(COUPON = 1, NEW = 3, VACATION = "No" , SW ="No" , HI =6000, S\_INCOME = 2000, E\_INCOME = 2000, S\_POP = 4000000, E\_POP=7150000, SLOT="Free", GATE ="Constrained" , DISTANCE = 1000, PAX = 6000)  
pred<-predict(fit2,newdata)  
pred

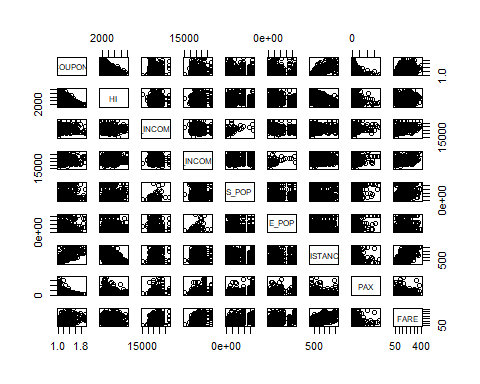
## 1   
## 142.6638

* 1. Do you have any reservations about your predicted fare? If so, explain why?

Ans – The R-squared value is pretty good. The p-value for NEW and COUPON is high which states that the possibility of their coefficients being not 0 is less. Also some t-values are negative as well.

1. **Exploratory Analysis** 
   1. Use the PAIRS function in *R* to depict pairwise scatterplots of all numerical variables in the data set.

Ans - new\_dfKK<-dfKK[c(5,9,10,11,12,13,16,17,18)]  
pairs(new\_dfKK)



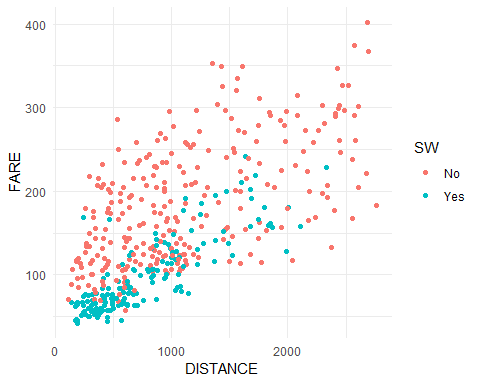
**Scatterplot Matrix**

* 1. Present a scatterplot of FARE (Y-axis) and DISTANCE (X-axis). Use different colors or symbols to distinguish routes where *Southwest* is present. Briefly describe what you observe.

Ans – Whenever Southwest is present on the routes, the FARE is less.

library(ggplot2)  
ggplot(data=dfKK)+aes(x=DISTANCE,y=FARE,color = SW)+geom\_point()+theme\_minimal()

## Warning: Removed 4 rows containing missing values (geom\_point).



**DISTANCE-FARE Scatterplot**

* 1. High correlations between independent variables can be problematic. Present a table of linear correlations and identify any values greater than 0.7.

Ans – There is one value greater than 0.7. Correlation between COUPON and DISTANCE is 0.7459  
library(corrplot)

## corrplot 0.84 loaded

corr<-cor(dfKK[,num],use="complete.obs")  
corr

## COUPON NEW HI S\_INCOME E\_INCOME  
## COUPON 1.00000000 0.01465463 -0.34735680 -0.08795449 0.04992242  
## NEW 0.01465463 1.00000000 0.05164550 0.03143026 0.12120600  
## HI -0.34735680 0.05164550 1.00000000 -0.01846331 0.09805658  
## S\_INCOME -0.08795449 0.03143026 -0.01846331 1.00000000 -0.17358284  
## E\_INCOME 0.04992242 0.12120600 0.09805658 -0.17358284 1.00000000  
## S\_POP -0.10655180 -0.01165029 -0.16278231 0.51300805 -0.16696171  
## E\_POP 0.09513191 0.05886380 -0.06547251 -0.26053371 0.51051764  
## DISTANCE 0.74599996 0.08123172 -0.30856540 0.02795266 0.18484601  
## PAX -0.33155163 0.02087204 -0.16452859 0.13480858 0.26974965  
## FARE 0.50132582 0.09305202 0.02419540 0.21886941 0.35665606  
## S\_POP E\_POP DISTANCE PAX FARE  
## COUPON -0.10655180 0.09513191 0.74599996 -0.33155163 0.50132582  
## NEW -0.01165029 0.05886380 0.08123172 0.02087204 0.09305202  
## HI -0.16278231 -0.06547251 -0.30856540 -0.16452859 0.02419540  
## S\_INCOME 0.51300805 -0.26053371 0.02795266 0.13480858 0.21886941  
## E\_INCOME -0.16696171 0.51051764 0.18484601 0.26974965 0.35665606  
## S\_POP 1.00000000 -0.27390954 0.01635429 0.28234359 0.15006073  
## E\_POP -0.27390954 1.00000000 0.11716227 0.31886253 0.28310058  
## DISTANCE 0.01635429 0.11716227 1.00000000 -0.09977940 0.67303486  
## PAX 0.28234359 0.31886253 -0.09977940 1.00000000 -0.08976502  
## FARE 0.15006073 0.28310058 0.67303486 -0.08976502 1.00000000

1. **More Regression Modeling**Run a simple regression to determine the effect of the presence of *Southwest* on FARE. Compare the coefficient with the corresponding value in (1) and explain discrepancies, if any.

Ans – The coefficient of SW in (1) is -40.78 and in the current model is -91.94. This means that in (1), the FARE is not reducing significantly when SW=Yes and other variables remain constant, whereas in current model, the FARE reduces significantly.

fit3<- lm(FARE~SW,data=dfKK)  
summary(fit3)

##   
## Call:  
## lm(formula = FARE ~ SW, data = dfKK)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -132.32 -46.32 -14.32 41.68 212.68   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 189.324 3.071 61.64 <2e-16 \*\*\*  
## SWYes -91.946 5.527 -16.64 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 63.84 on 623 degrees of freedom  
## (2 observations deleted due to missingness)  
## Multiple R-squared: 0.3076, Adjusted R-squared: 0.3065   
## F-statistic: 276.8 on 1 and 623 DF, p-value: < 2.2e-16

1. **Further Analysis**
   1. A senior consultant in the airline industry has indicated that the presence of *Southwest* on Vacation routes has been driving prices down on these legs. Add this domain knowledge to your regression model and run a new multivariable linear regression. Describe how you added this feature to the model.

Ans – I’ve added this feature using interaction between Southwest and Vacation.

fit5<- lm(FARE~COUPON+NEW+VACATION+SW+(SW\*VACATION)+HI+S\_POP+E\_POP+SLOT+GATE+DISTANCE+PAX,data=dfKK)  
summary(fit5)

##   
## Call:  
## lm(formula = FARE ~ COUPON + NEW + VACATION + SW + (SW \* VACATION) +   
## HI + S\_POP + E\_POP + SLOT + GATE + DISTANCE + PAX, data = dfKK)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -106.459 -20.211 0.497 21.343 122.281   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.018e+02 1.805e+01 5.641 2.60e-08 \*\*\*  
## COUPON -1.381e+01 1.190e+01 -1.161 0.246   
## NEW -6.004e-01 1.832e+00 -0.328 0.743   
## VACATIONYes -5.751e+01 4.196e+00 -13.704 < 2e-16 \*\*\*  
## SWYes -6.200e+01 3.947e+00 -15.706 < 2e-16 \*\*\*  
## HI 8.202e-03 9.558e-04 8.581 < 2e-16 \*\*\*  
## S\_POP 3.494e-06 6.108e-07 5.719 1.68e-08 \*\*\*  
## E\_POP 4.367e-06 6.800e-07 6.422 2.71e-10 \*\*\*  
## SLOTFree -1.702e+01 3.690e+00 -4.612 4.86e-06 \*\*\*  
## GATEFree -2.060e+01 3.888e+00 -5.299 1.63e-07 \*\*\*  
## DISTANCE 8.093e-02 3.400e-03 23.806 < 2e-16 \*\*\*  
## PAX -8.089e-04 1.380e-04 -5.860 7.57e-09 \*\*\*  
## VACATIONYes:SWYes 5.722e+01 6.793e+00 8.423 2.62e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 34.18 on 610 degrees of freedom  
## (4 observations deleted due to missingness)  
## Multiple R-squared: 0.8056, Adjusted R-squared: 0.8017   
## F-statistic: 210.6 on 12 and 610 DF, p-value: < 2.2e-16

* 1. What is the resulting R2?

Ans – 0.8017

* 1. Now how would you quantify the effect of SW on the fare?

Ans – 1. When there is interaction between SW and VACATION, the FARE increases by $57.22.

2. Considering the variable SW separately, it reduces the FARE by $62

1. **Comparing two different models**

We will now build a model for FARES using the following explanatory variables: Factor variables for VACATION and SW, HI, S\_INCOME, E\_INCOME, S\_POP, E\_POP, DISTANCE, PAX. Then we will compare this with the model *without* the two INCOME variables.

* 1. Run the regression with and without the INCOME variables.

1. fit4<- lm(FARE~COUPON+NEW+VACATION+SW+HI+S\_POP+E\_POP+S\_INCOME+E\_INCOME+DISTANCE+PAX,data=dfKK)  
   summary(fit4)
2. ##   
   ## Call:  
   ## lm(formula = FARE ~ COUPON + NEW + VACATION + SW + HI + S\_POP +   
   ## E\_POP + S\_INCOME + E\_INCOME + DISTANCE + PAX, data = dfKK)  
   ##   
   ## Residuals:  
   ## Min 1Q Median 3Q Max   
   ## -126.093 -24.048 -0.172 22.584 117.531   
   ##   
   ## Coefficients:  
   ## Estimate Std. Error t value Pr(>|t|)   
   ## (Intercept) -6.454e+01 2.715e+01 -2.377 0.017749 \*   
   ## COUPON 1.182e+01 1.256e+01 0.941 0.347097   
   ## NEW -2.032e+00 1.941e+00 -1.047 0.295393   
   ## VACATIONYes -3.251e+01 3.866e+00 -8.410 2.89e-16 \*\*\*  
   ## SWYes -4.631e+01 3.721e+00 -12.446 < 2e-16 \*\*\*  
   ## HI 8.201e-03 1.011e-03 8.115 2.69e-15 \*\*\*  
   ## S\_POP 4.623e-06 6.420e-07 7.202 1.75e-12 \*\*\*  
   ## E\_POP 5.525e-06 7.377e-07 7.489 2.43e-13 \*\*\*  
   ## S\_INCOME 1.924e-03 5.297e-04 3.632 0.000304 \*\*\*  
   ## E\_INCOME 2.015e-03 4.181e-04 4.820 1.81e-06 \*\*\*  
   ## DISTANCE 6.962e-02 3.621e-03 19.224 < 2e-16 \*\*\*  
   ## PAX -9.608e-04 1.507e-04 -6.373 3.64e-10 \*\*\*  
   ## ---  
   ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
   ##   
   ## Residual standard error: 36.3 on 611 degrees of freedom  
   ## (4 observations deleted due to missingness)  
   ## Multiple R-squared: 0.7803, Adjusted R-squared: 0.7764   
   ## F-statistic: 197.3 on 11 and 611 DF, p-value: < 2.2e-16
3. fit6<-lm(FARE~COUPON+NEW+VACATION+SW+HI+S\_POP+E\_POP+DISTANCE+PAX,data=dfKK)  
   summary(fit6)
4. ##   
   ## Call:  
   ## lm(formula = FARE ~ COUPON + NEW + VACATION + SW + HI + S\_POP +   
   ## E\_POP + DISTANCE + PAX, data = dfKK)  
   ##   
   ## Residuals:  
   ## Min 1Q Median 3Q Max   
   ## -116.116 -23.917 -1.567 23.321 113.832   
   ##   
   ## Coefficients:  
   ## Estimate Std. Error t value Pr(>|t|)   
   ## (Intercept) 4.278e+01 1.821e+01 2.349 0.0191 \*   
   ## COUPON 4.460e+00 1.278e+01 0.349 0.7272   
   ## NEW -1.317e+00 1.981e+00 -0.665 0.5066   
   ## VACATIONYes -3.838e+01 3.780e+00 -10.152 < 2e-16 \*\*\*  
   ## SWYes -5.332e+01 3.519e+00 -15.153 < 2e-16 \*\*\*  
   ## HI 8.887e-03 1.020e-03 8.709 < 2e-16 \*\*\*  
   ## S\_POP 4.838e-06 6.229e-07 7.767 3.40e-14 \*\*\*  
   ## E\_POP 5.827e-06 6.980e-07 8.347 4.64e-16 \*\*\*  
   ## DISTANCE 7.408e-02 3.611e-03 20.512 < 2e-16 \*\*\*  
   ## PAX -7.579e-04 1.496e-04 -5.067 5.37e-07 \*\*\*  
   ## ---  
   ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
   ##   
   ## Residual standard error: 37.15 on 613 degrees of freedom  
   ## (4 observations deleted due to missingness)  
   ## Multiple R-squared: 0.7691, Adjusted R-squared: 0.7657   
   ## F-statistic: 226.9 on 9 and 613 DF, p-value: < 2.2e-16­­­­  
   1. On the basis of the output evaluate which model is better for predicting FARES. *Explain carefully how you made this determination*.

Ans – On the basis of the output, the R-squared value for initial model including the income variables is 0.7764

and for the second model without the income variables is 0.7657.

As the R-squared value for the first model is higher it is better for predicting FARES.