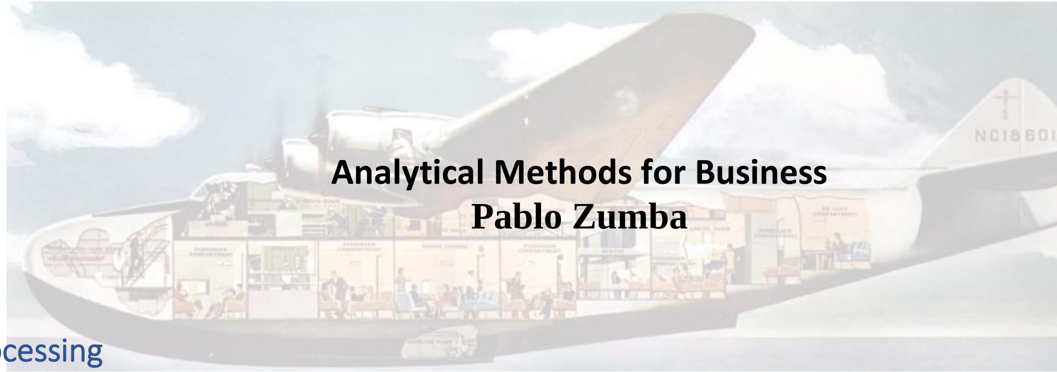


Analytical Methods for Business

Pablo Zumba

Preprocessing

1. Load the file “6304 Module 5 Assignment Data.xlsx” data set into R. This data shows airfares and passengers for certain U.S. Domestic Routes for the 4th quarter of 2002. This is not an exhaustive list of all flights.

```
#Pablo Zumba
#U54252888
#Processing 1:
rm(list=ls())
set.seed(54252888)
library(rio)
flights = import("6304 Module 5 Assignment Data.xlsx")
colnames(flights)=tolower(make.names(colnames(flights)))
attach(flights)
names(flights)
str(flights)
> names(flights)
[1] "origin"                  "destination"             "average.fare"
"distance"
[5] "avg.weekly.passengers"   "market.leading.airline" "route.market.share"
"low.price.airline"
[9] "price"
> str(flights)
'data.frame':    1000 obs. of  9 variables:
 $ origin                : chr  "CAK" "CAK" "ALB" "ALB" ...
 $ destination           : chr  "ATL" "MCO" "ATL" "BWI" ...
 $ average.fare          : num  114.5 122.5 214.4 69.4 158.1 ...
 $ distance              : num  528 860 852 288 723 ...
 $ avg.weekly.passengers : num  425 277 216 607 313 ...
 $ market.leading.airline: chr  "FL" "FL" "DL" "WN" ...
 $ route.market.share    : num  70.2 75.1 78.9 97 39.8 ...
 $ low.price.airline     : chr  "FL" "DL" "CO" "WN" ...
 $ price                 : num  111 118.9 167.1 68.9 145.4 ...
```

2. Create a random selection of flights of n=50. Be certain to include in your sample only the origin airports of LAS, LAX, BWI, LGA, MCI, MCO, ATL, and BNA. Make sure to convert any character (chr) variables to factor variables. This will be your primary data set for analysis.

```
subFlights =
subset(flights,origin=="LAS"|origin=="LAX"|origin=="BWI"|origin=="LGA"|origi
n=="MCI"|origin=="MCO"|origin=="ATL"|origin=="BNA")
subFlights = subFlights[sample(1:nrow(subFlights),50),]
subFlights$origin = as.factor(subFlights$origin)
subFlights$destination = as.factor(subFlights$destination)
subFlights$market.leading.airline =
as.factor(subFlights$market.leading.airline)
subFlights$low.price.airline = as.factor(subFlights$low.price.airline)
```

Analysis Using Your Primary Data Set

1. Show the results of an str() command.

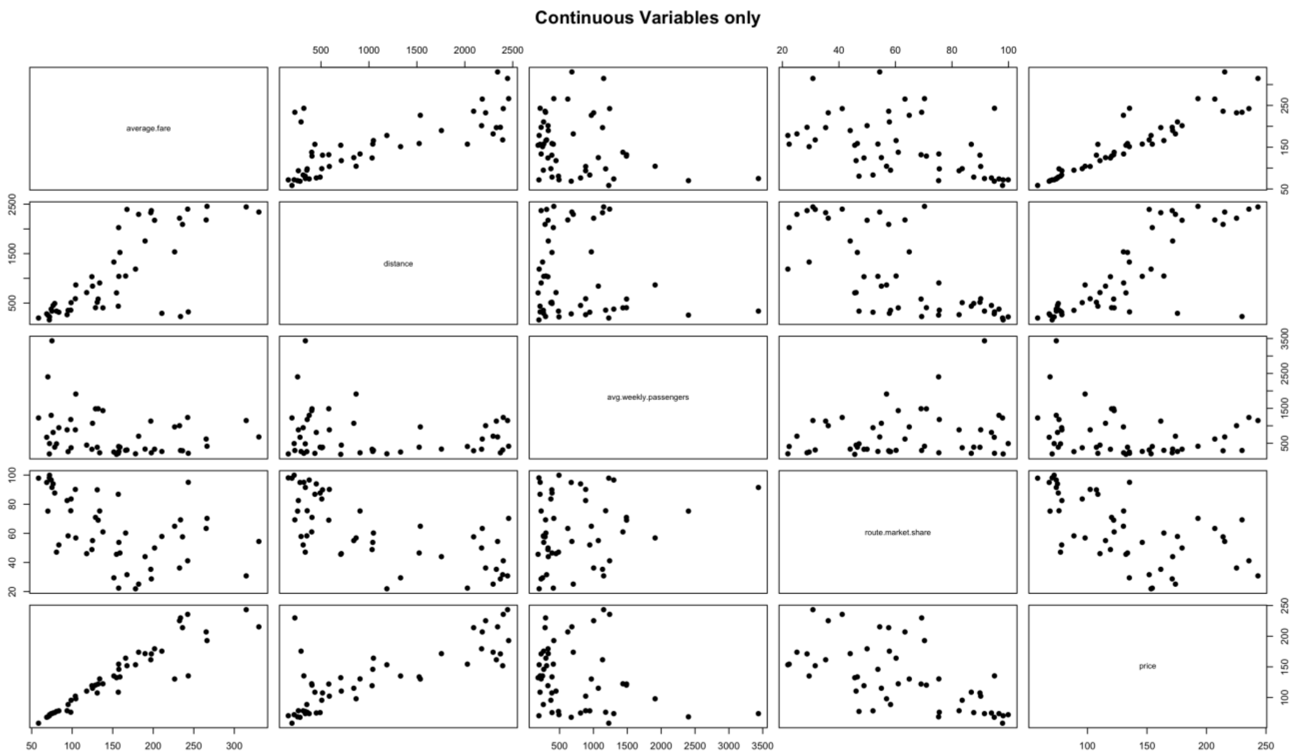
```
str(subFlights)
str(subFlights)
'data.frame':	50 obs. of  9 variables:
 $ origin                : Factor w/ 8 levels "ATL","BNA","BWI",..: 5 4 4 1
6 7 2 3 6 8 ...
 $ destination           : Factor w/ 35 levels "BNA","BUF","CLE",..: 17 30
24 24 33 13 23 5 27 28 ...
 $ average.fare          : num  151.4 104.4 157.3 98.2 165.9 ...
 $ distance              : num  1330 866 2027 356 1047 ...
 $ avg.weekly.passengers : num  248 1910 407 1180 297 ...
 $ market.leading.airline: Factor w/ 12 levels "AA","AS","B6",..: 9 2 5 5 4
12 11 7 4 11 ...
 $ route.market.share    : num  29.4 56.8 22.4 75.5 60.2 ...
 $ low.price.airline     : Factor w/ 12 levels "AA","B6","CO",..: 8 7 12 6 4
4 11 12 7 12 ...
 $ price                 : num  135.2 98 154.4 76.1 164.2 ...
```

2. Show the results of a table() command on the origin variable.

```
table(origin)
> table(origin)

ATL BNA BWI LAS LAX LGA MCI MCO
 12   1  12   6  10   4   3   2
```

3. Show a scatterplot matrix of the continuous variables only. From this matrix which pair of variables do you believe would have the strongest linear relationship? How did you arrive at this conclusion?

```
plot(subFlights[,c(3,4,5,7,9)],pch=19,main="Continuous Variables only")
judge_cor = round(cor(subFlights[,c(3,4,5,7,9)]),3)
library(corrplot)
corrplot(judge_cor,method="number")
```



Interpretation: Based on the scatterplot matrix, the variables "price" and "average.fare" seems to be the most positively correlated due to the pattern tends to follow a straight line. However, to make sure they are correlated I used the corrplot() function. On the corrplot we can see that the most relevant correlations are as follows:

1. "price vs average.fare" 0.93
2. "price vs distance" 0.78
3. "average.fare vs distance" 0.75

4. Parameterize a full regression model with y=price. Include all other continuous variables as well as the origin variable. Show the R summary of this model.

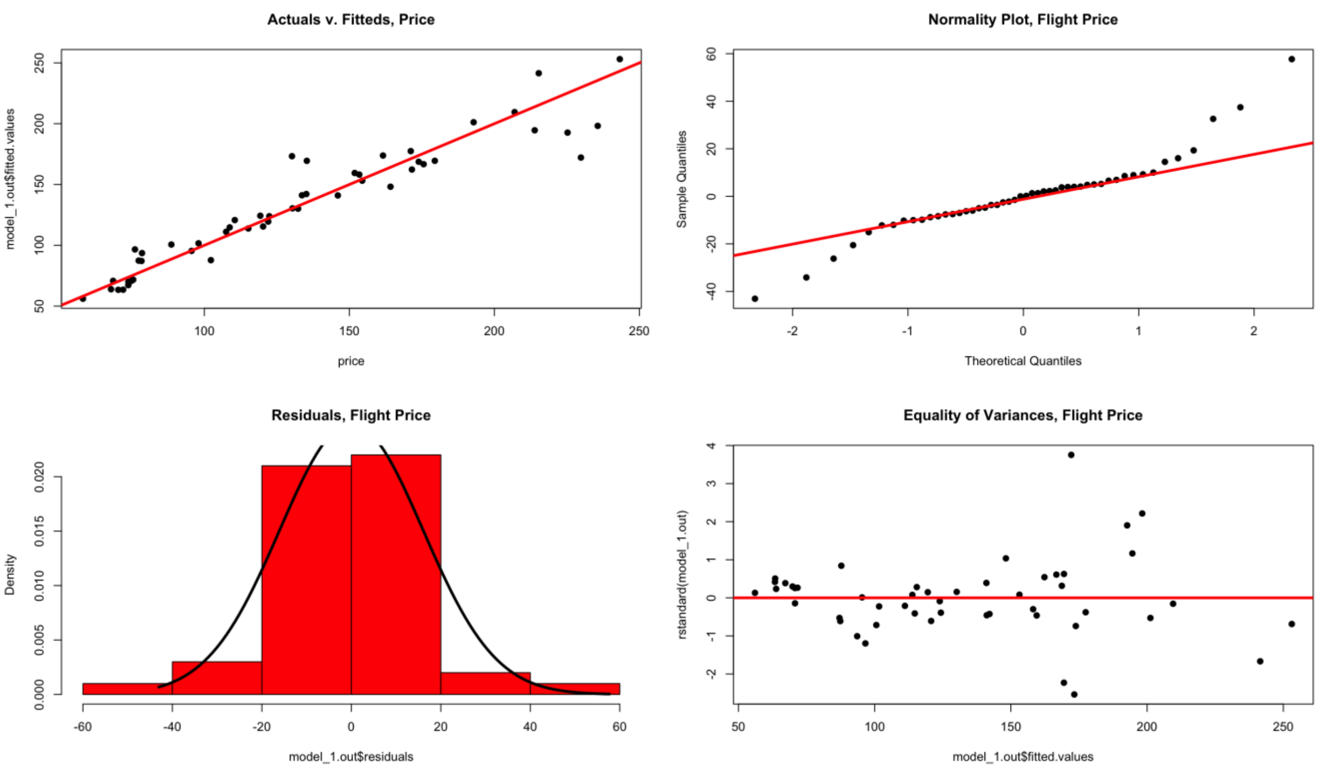
```
model_1.out =
lm(price~origin+average.fare+distance+avg.weekly.passengers+route.market.sha
re,data = subFlights)
summary(model_1.out)
Call:
lm(formula = price ~ origin + average.fare + distance +
avg.weekly.passengers + route.market.share, data = subFlights)

Residuals:
    Min      1Q  Median      3Q     Max
-43.064  -7.581   0.098   5.162  57.701

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            6.501e+01  1.620e+01   4.012 0.000272 ***
originBNA              9.086e+00  1.947e+01   0.467 0.643428
originBWI             -9.037e+00  8.630e+00  -1.047 0.301646
originLAS             -9.480e+00  1.049e+01  -0.904 0.371604
originLAX             -8.396e+00  8.766e+00  -0.958 0.344202
originLGA              2.733e+00  1.073e+01   0.255 0.800291
originMCI             -1.100e+01  1.212e+01  -0.908 0.369779
originMCO             -1.952e+00  1.446e+01  -0.135 0.893339
average.fare           5.513e-01  7.296e-02   7.555 4.42e-09 ***
distance               9.156e-03  6.935e-03   1.320 0.194685
avg.weekly.passengers  4.974e-05  4.717e-03   0.011 0.991642
route.market.share    -3.428e-01  1.640e-01  -2.090 0.043340 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.24 on 38 degrees of freedom
Multiple R-squared:  0.9032,     Adjusted R-squared:  0.8752
F-statistic: 32.24 on 11 and 38 DF,  p-value: 6.888e-16
```

5. Drawing on Step 4, give a verbal interpretation of the impact of the levels of the “origin” variable on price in your model.

Interpretation: Assuming that the “price” variable is in USD dollars, we can interpret the Beta values as follows:

- For every additional flight with origin in BNA, we expect the price variable to *increase* by 9.08 USD
- For every additional flight with origin in BWI, we expect the price variable to *decrease* by -9.04 USD
- For every additional flight with origin in LAS, we expect the price variable to *decrease* by -9.48 USD
- For every additional flight with origin in LAX, we expect the price variable to *decrease* by -8.40 USD
- For every additional flight with origin in LGA, we expect the price variable to *increase* by 2.73 USD
- For every additional flight with origin in MCI, we expect the price variable to *decrease* by -11 USD
- For every additional flight with origin in MCO, we expect the price variable to *decrease* by -1.95 USD

6. Drawing on Step 4, determine whether your model meets the LINE assumptions of regression.



Interpretation: As a result of applying the square root to the Multiple R-squared value in step 4, we obtain an “r” value of 0.95 (very close to 1), indicating a highly positive linear relationship. Look at the "Actual v. Fitted" plot to confirm this. The Normality plot shows that the quantiles and theoretical quantiles follow the QQLine reasonably well. Additionally, the residuals seem to be normally distributed so we can confirm Normality. Finally, there are no obvious patterns in the Equality of variances plot, however, due to the nature of our business problem, it is likely to see patterns caused by seasonal price variations e.g., during holidays like Christmas or Spring break. Concluding, the model meets Linearity, Normality, and Equality of variance. Note: further analysis needs to be performed to assess Independence and autocorrelation.

7. Drawing on Step 4, report in a single vector the origin and destination airports and the original price of the flight for which the actual price deviates *most* from your model's regression line.

```
subFlights[which.max(abs(model_1.out$residuals)),c(1,2,9)]
> #Analysis 7
> subFlights[which.max(abs(model_1.out$residuals)),c(1,2,9)]
   origin destination  price
35    ATL         CLT 229.85
```

8. Drawing on Steps 4 and 7, report in a single vector the origin and destination airports and the original price of the flight for which the actual price deviates *least* from your model's regression line.

```
subFlights[which.min(abs(model_1.out$residuals)),c(1,2,9)]
> #Analysis 8
> subFlights[which.min(abs(model_1.out$residuals)),c(1,2,9)]
    origin destination  price
799    BNA         PVD 130.38
```