AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

(AIMS RWANDA, KIGALI)

Name: Stephen Kiilu
Course: Regression with R

Assignment Number: SRR2
Date: December 13, 2020

---

1 QUESTION 1

Descriptive analysis

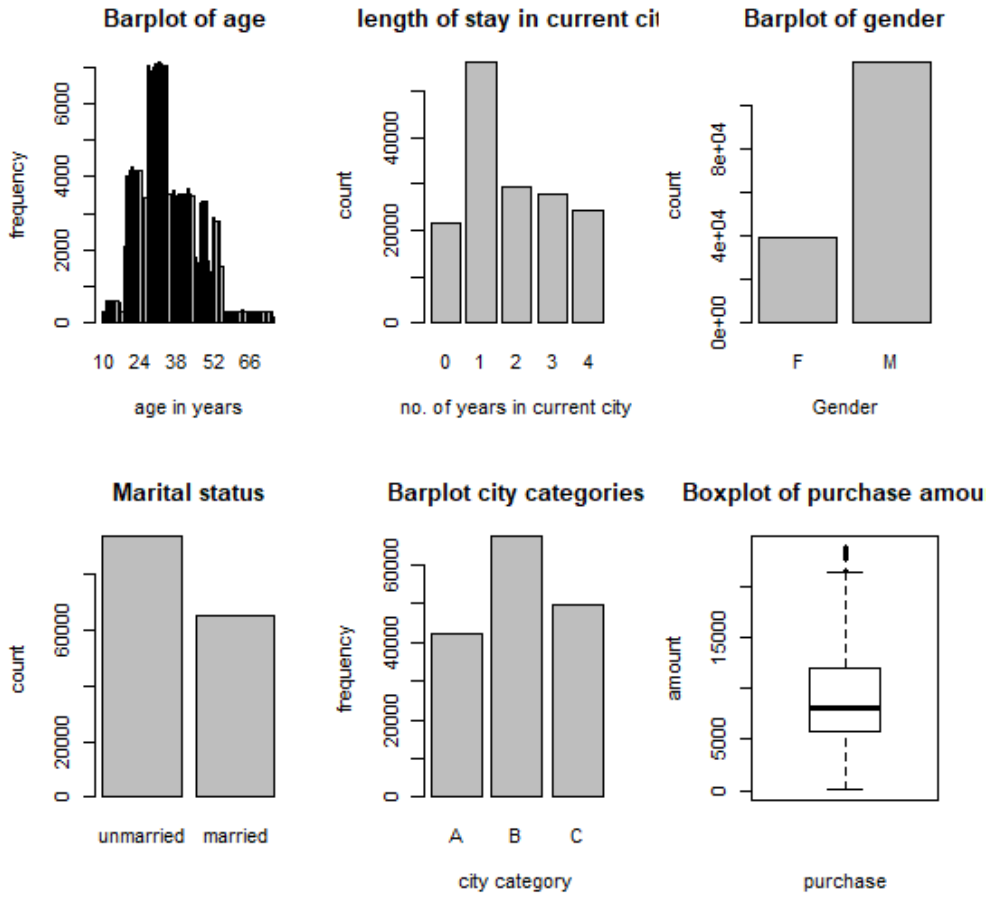

Figure 1: Descriptive analysis

Most of the customers are males, who are approximately three times the number of female customers. From the boxplot of purchase amount we see we have some extreme purchase, which

means some customers we spending abnormally more as compared to other customers. It also clear that most of customers are spending above middle purchase amount.
We can see from barplot of age that most of the customers are aged between 28 and 34 years. Most of the customers have stayed in their current city for one year, the city from which most customers are drawn is B. We can also state that most of the customers are unmarried.

2 QUESTION 2

We are going to built a regression model using backward elimination method. We first built a model with all variables and get this summary output.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8275.0003	51.3159	161.26	0.0000	***
cityB	232.8961	31.1291	7.48	0.0000	***
cityC	816.1364	33.3939	24.44	0.0000	***
GenderM	707.5733	29.0924	24.32	0.0000	***
Maritals	-44.7599	26.7599	-1.67	0.0944	
Age	3.8340	1.1270	3.40	0.0007	***
stay	-2.4060	9.7449	-0.25	0.8050	

Table 1: model 1

From model 1, all variables with exception of stay and marital status have a p-value <0.001. The variables with p-value >0.05, are insignficant to our model, they do not add any additional information to our model. Consequently we are going to eliminate marital status and stay in current city from our model and built new model.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8271.6583	48.2288	171.51	0.0000	***
cityB	232.3207	31.1195	7.47	0.0000	***
cityC	815.7573	33.3829	24.44	0.0000	***
GenderM	708.0186	29.0878	24.34	0.0000	***
Age	3.2739	1.0757	3.04	0.0023	**

Table 2: final model

We carefully examine the p-value against all our explanatory variables. We notice that all variables in our model have a p-value <0.05. This means all our explanatory variables are very significant to our model and conclude that we have our final model.

QUESTION 3

We have built a multiple regression model with age, city category and gender as our explanatory variables. Our regression model is

$$\hat{y} = 8271.65 + 232.32x_1 + 815.75x_2 + 708.01x_3 + 3.28x_4.$$

Where 8271.65 is the intercept of our model, $\hat{y}$ is the predicted purchase amount, x_1 is city category B, x_2 is city category C, x_3 Males and x_3 is age. The corresponding standard errors for city B, city C, Males and age are 31.12, 33.39, 29.09 and 1.08, which are relatively low as compared to the estimates.
When all other explanatory variables are held constant, a unit increase in age increases the expected purchase amount by 3.28 units, the purchase amount in male customers is 708.01 higher than in female customers,and customers from city category B and C have a higher purchase amount than customers from city category A by 232.32 and 815.75 units respectively.

QUESTION 4

We are going to assume that, the independent variables are non-stochastic (fixed predictors), the regression parameters are constant and carry out residual analysis to check for assumptions of;linear association, homoskedasticity, normality and outliers.

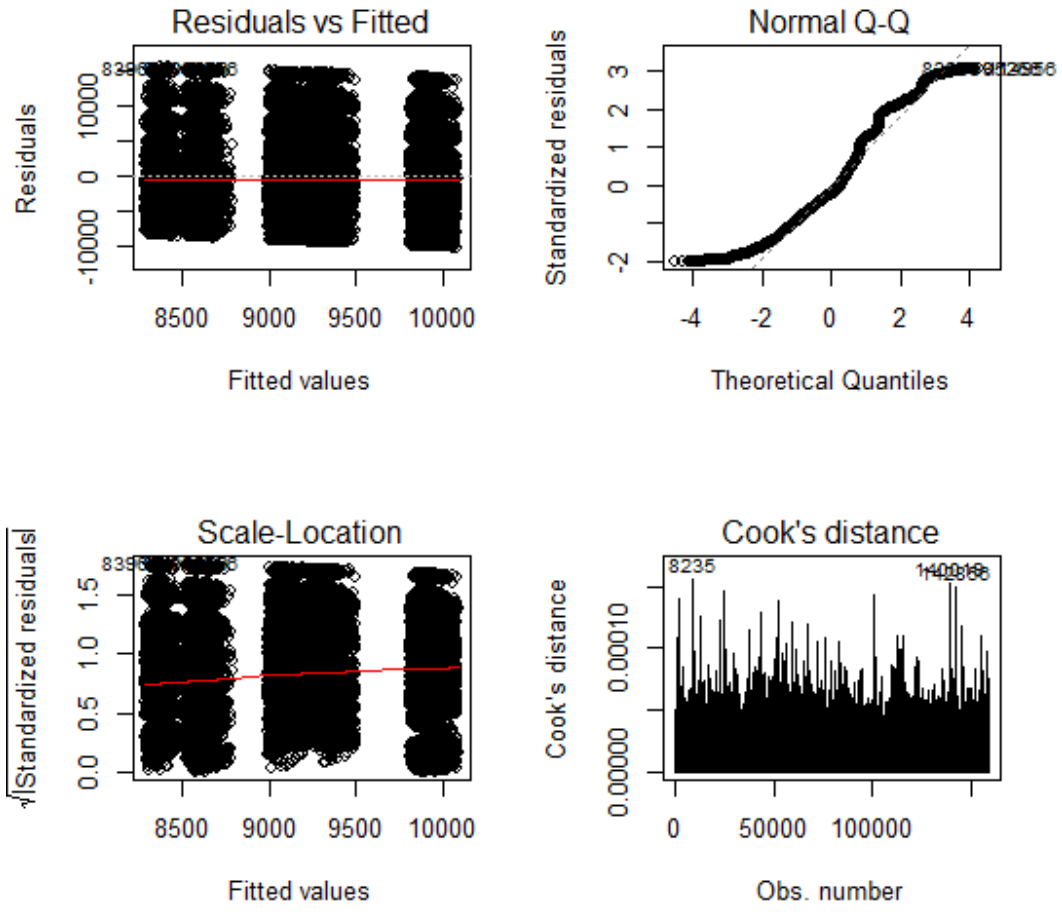

Figure 2: Residue analysis

Linear association - The first plot from left is a plot of residuals against fitted values, it shows some gaps in the fitted values which indicates that the residuals are not random. The assumption of linear association does not hold.

Normality- for the assumption of normality to be true, the residuals should follow a straight line, from our second plot many data points deviate away from a straight line which show that the residues are not normally distributed, consequently the assumption of normality fails.

Homoskedasticity - for this assumption, we expect that the residuals to have a constant variance, but according to our third plot, there is increasing spread of the residuals from left to right with some gaps, and as a result the assumption of homoskedasticity is violated.

Check for outliers - we observe from our fourth plot that are significant outliers in our data. The outliers need to be investigated and a decision reached whether to remove or retail them in our data.

QUESTION 5

From this model, city category C is a very important driver towards maximization of the company's sales. The company should consider employing more strategies so that it continues enjoying more purchase from customers in this city category. The company is doing poorly in terms of purchase from female customers, there is a need to come up with products which attract more female customers for the company to realize increase in its sales.

QUESTION 6

Some possible ways of improving our model include;

- Adding appropriate variables
- Checking for outliers
- Data transformation e.g by taking natural log of the purchase amount.

QUESTION 7

We are going to use ANOVA to compare the two models, model 1 and model 2.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	
Model 1	158997	4001520005780.23					
Model 2	158995	3984905477192.59	2	16614528587.64	331.45	0.0000	***

The model built on gender and age is model 1 and my final model is model 2. The p-value of model 2 is very significant, which means that my final model is better. The difference between the two models is inclusion of city category variable in my model, this shows that the city category variable adds additional information to the model. This means that the amount of total variation explain by my final model is higher compared to amount variation explain by model built on age and gender only.