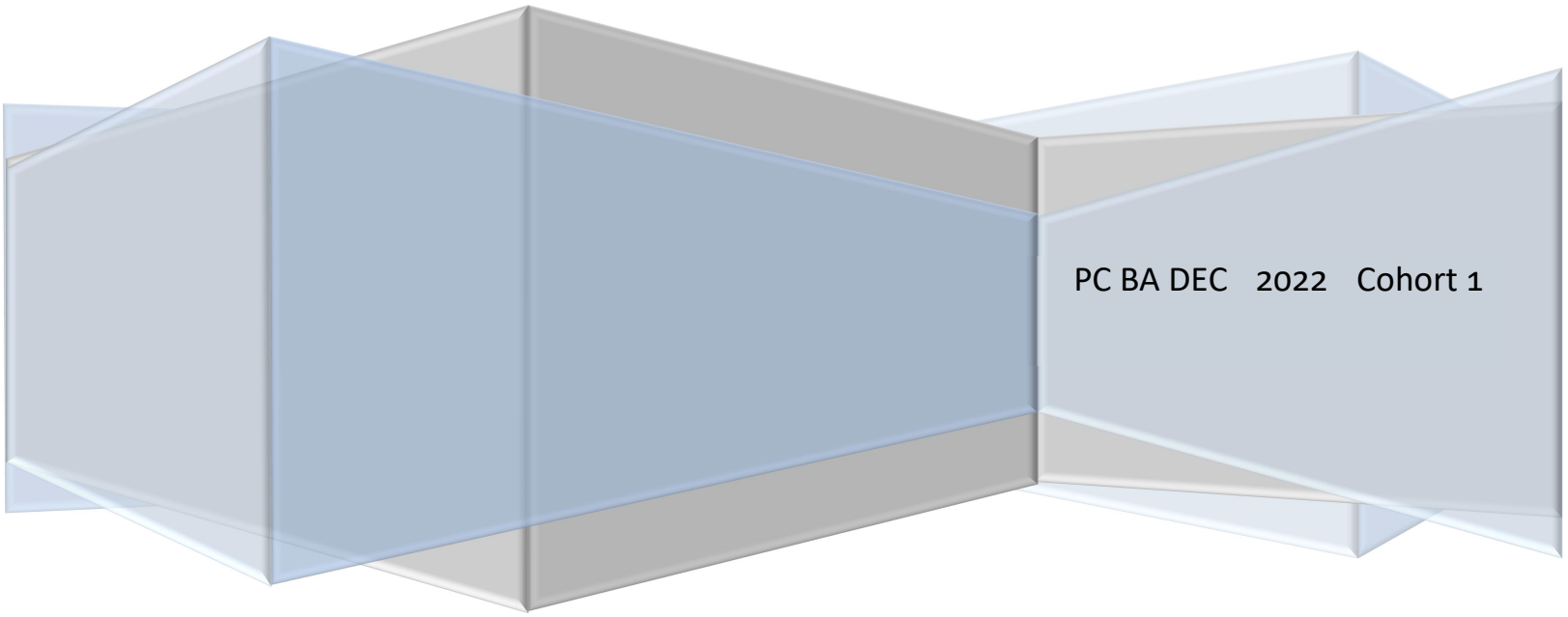
Simplilearn

Predicting Restaurant tips

Regression model using Excel

Analyst – Manohari Wijesooriya



Contents

1. Business case

A local restaurant gathered amount of tip their customer make on the bill. Additional customer data is gathered about customers dining in the restaurant. The management wants to know what variables motivate customers to give a tip.

Proposed solution: Build a multiple linear regression model to predict restaurant tip.

Following are the features in the dataset.

sex	Gender of the customer
smoker	Indicates if the customer is a smoker or not
day	Day of the restaurant visit
time	Indicates whether the tip was for lunch or dinner
size	Number of members dining
total bill	Bill amount in USD
tip	Tip amount in USD

First 6 records from the data file.

sex	smoker	day	time	size	total_bill	tip
Female	No	Sun	Dinner	2	16.99	1.01
Male	No	Sun	Dinner	3	10.34	1.66
Male	No	Sun	Dinner	3	21.01	3.5
Male	No	Sun	Dinner	2	23.68	3.31
Female	No	Sun	Dinner	4	24.59	3.61
Male	No	Sun	Dinner	4	25.29	4.71

2. Cleaning Data

Each column in Excel data file was examined to see if there are blank records. There were no blank records in the file.

3. Variable features

To build the regression model, dependant and independent variables are identifies as follows.

Dependant variable: **tip**

Independent variables: **sex, smoker, day, time, size, total_bill**

4. Encoding data

Variables sex, smoker, day, time is categorical variables. We need to change those for numeric values to use these variables in regression analysis.

Formulas used for encoding:

sex_M =IF(C2="Male",1,0) , sex_F =IF(C2="Female",1,0)

smoker_Y =IF(F2="Yes",1,0) , smoker_N =IF(F2="No",1,0)

day_THU =IF(K2="Thur",1,0)

day_FRI =IF(K2="Fri",1,0)

day_SAT =IF(K2="Sat",1,0)

day_SUN =IF(K2="Sun",1,0)

time_D =IF(K2="Dinner",1,0)

time_L =IF(K2="Lunch",1,0)

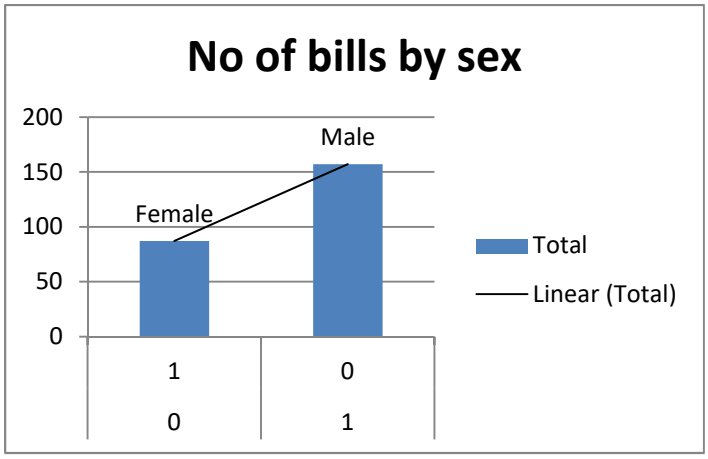
5. Exploratory Data Analysis

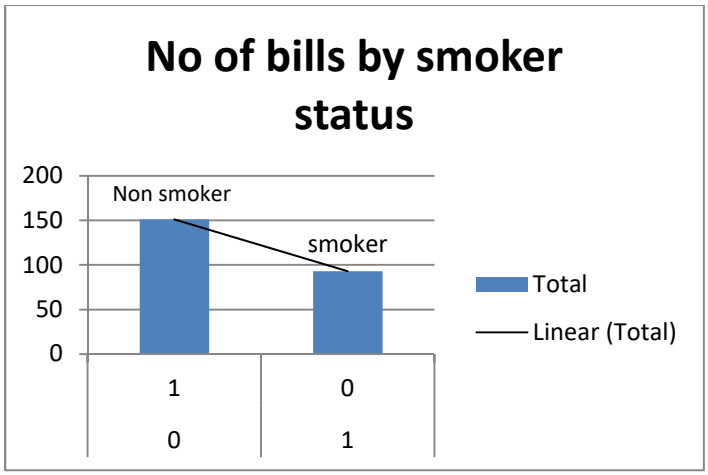
These charts are prepared to understand data and to identify patterns in data.

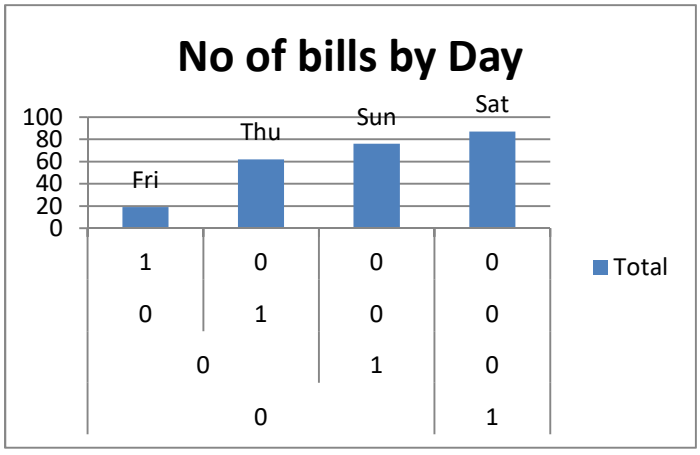
Correlation													
	sex_M	*sex_F*	*smoker_Y*	*smoker_N*	*day_THU*	*day_FRI*	*day_SAT*	*day_SUN*	*time_D*	*time_L*	*size*	*total_bill*	*tip*
sex_M	1												
sex_F	-1	1											
smoker_Y	0.002816	-0.00282	1										
smoker_N	-0.00282	0.002816	-1	1									
day_THU	-0.19444	0.194445	-0.12853	0.128534	1								
day_FRI	-0.07106	0.07106	0.244316	-0.24432	-0.16961	1							
day_SAT	0.053957	-0.05396	0.155744	-0.15574	-0.43448	-0.21632	1						
day_SUN	0.168106	-0.16811	-0.18162	0.181624	-0.39257	-0.19545	-0.50068	1					
time_D	0.205231	-0.20523	0.054921	-0.05492	-0.918	-0.05816	0.462709	0.418071	1				
time_L	-0.20523	0.205231	-0.05492	0.054921	0.917996	0.058159	-0.46271	-0.41807	-1	1			
size	0.086195	-0.08619	-0.13318	0.133178	-0.0726	-0.14218	-0.04112	0.193054	0.103411	-0.10341	1		
total_bill	0.144877	-0.14488	0.085721	-0.08572	-0.13817	-0.08617	0.054919	0.122953	0.183118	-0.18312	0.598315	1	
tip	0.088862	-0.08886	0.005929	-0.00593	-0.09588	-0.05546	-0.00279	0.125114	0.121629	-0.12163	0.489299	0.675734	1

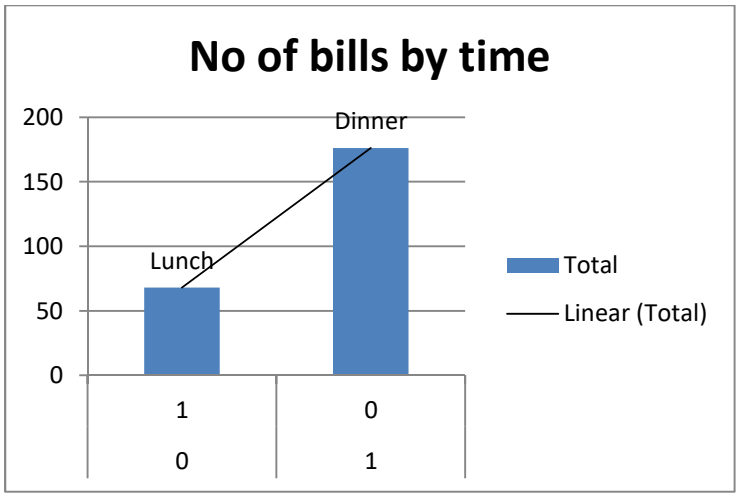
- Tip has high correlation with size and total bill.
- Restaurant has more clients on Thursday for lunch. Also more clients on Saturday and Sunday for dinner.
- High correlation in total bill (as expected)
- no direct correlation with sex and smoker variables

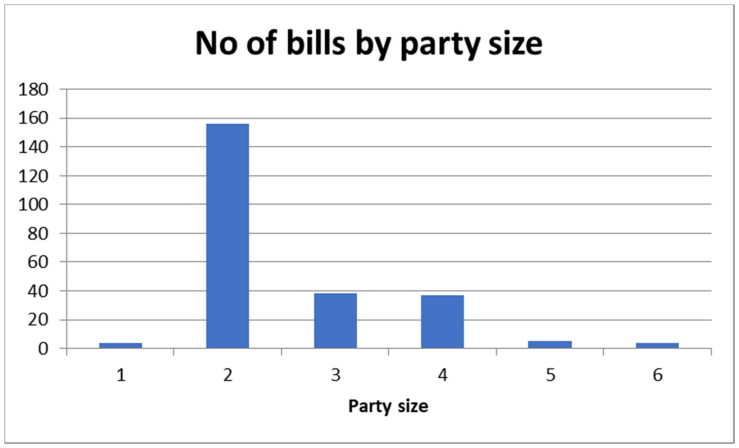
Visualisations to understand data patterns –

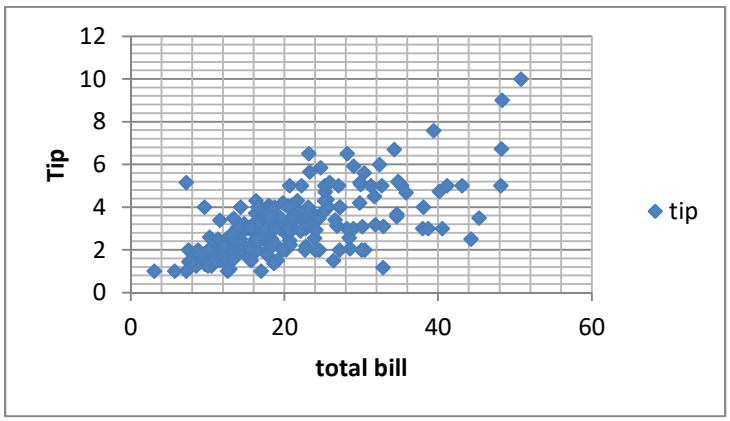




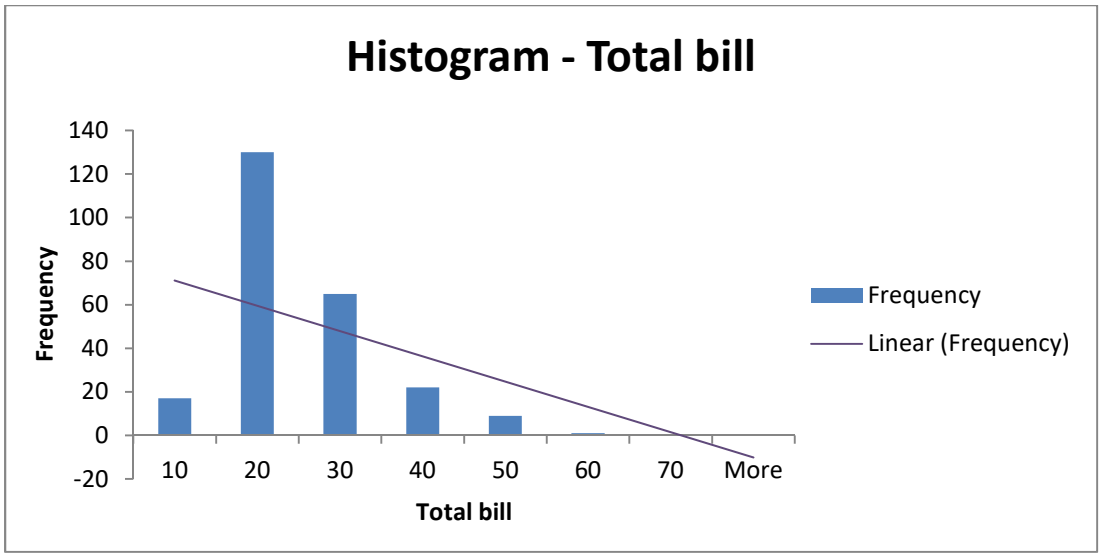








Tip amount increases as the total bill increase



In most of the times, the bill amount lies in the range $11 - $20

6. Linear Regression Model

A linear regression model is built to predict restaurant tip.

Dependant variable: **tip**

Independent variables: **sex, smoker, day, time, size, total_bill**

Regression Statistics	
Multiple R	0.685622435
R Square	0.470078123
Adjusted R Square	0.435016953
Standard Error	1.02423042
Observations	244

R squared is Coefficient of determination. In this model 47% of values fit the model.

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	12	218.6862081	18.22385068	26.05769921	1.46506E-36
Residual	235	246.5262689	1.049047953		
Total	247	465.212477			

Significance F value is the significance associated P-value. Lower the p-value the better model.

Interpret Regression co-efficient

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.727615964	0.504516987	1.442203101	0.150576486	-0.26633803	1.721569957	-0.26633803	1.721569957
sex_M	0	0	65535	#NUM!	0	0	0	0
sex_F	0.03244094	0.141612158	0.229083017	#NUM!	-0.246550597	0.311432478	-0.246550597	0.311432478
smoker_Y	0	0	65535	#NUM!	0	0	0	0
smoker_N	0.08640832	0.146587073	0.589467534	#NUM!	-0.202384346	0.375200987	-0.202384346	0.375200987
day_THU	-0.136778538	0.471696302	-0.289971613	0.772093651	-1.066072169	0.792515094	-1.066072169	0.792515094
day_FRI	0.02548066	0.321297781	0.079305433	0.936857175	-0.607511331	0.658472651	-0.607511331	0.658472651
day_SAT	-0.095977717	0.165881584	-0.578591753	0.563418769	-0.422782695	0.230827262	-0.422782695	0.230827262
day_SUN	0	0	65535	#NUM!	0	0	0	0
time_D	-0.068128601	0.444616858	-0.15322991	#NUM!	-0.944072746	0.807815544	-0.944072746	0.807815544
time_L	0	0	65535	#NUM!	0	0	0	0
size	0.175992003	0.089527743	1.965781748	#NUM!	-0.000387504	0.35237151	-0.000387504	0.35237151
total_bill	0.094487006	0.009601399	9.84096224	2.34253E-19	0.075571193	0.113402819	0.075571193	0.113402819

We see lower p-values for intercept and total bill, which is indication of good fit. T stat 65535 and #NUM for p-value indicates Excel couldn't out put the extremely smaller p-value and it is showing as error.

RESIDUAL OUTPUT

Observation	*Predicted tip*	*Residuals*
1	2.73565486	-1.72565486
2	2.250867333	-0.590867333
3	3.259043687	0.240956313
4	3.33533199	-0.02533199
5	3.805740111	-0.195740111
6	3.839440075	0.870559925
7	1.926530731	0.073469269

In the regression co-efficient chart, Days THU,FRI and SAT shows higher p-value.

We can drop Day field from the independent variable list, run the regression model and check regression output.

Model2 –

Dependant variable: **tip**

Independent variables: **sex, smoker, time, size, total_bill**

Regression Statistics								
Multiple R	0.6846667							
R Square	0.4687685							
Adjusted R Square	0.4450031							
Standard Error	1.0190116							
Observations	244							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	8	218.07694	27.259618	42.003115	2.913E-41			
Residual	238	247.13554	1.0383846					
Total	246	465.21248						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.6132397	0.2255569	2.7187804	0.0070348	0.1688968	1.0575827	0.1688968	1.0575827
sex_M	0	0	65535	#NUM!	0	0	0	0
sex_F	0.0281106	0.1400537	0.2007134	#NUM!	-0.2477925	0.3040137	-0.2477925	0.3040137
smoker_Y	0	0	65535	#NUM!	0	0	0	0
smoker_N	0.0839021	0.1386962	0.6049341	#NUM!	-0.1893269	0.3571311	-0.1893269	0.3571311
time_D	0	0	65535	#NUM!	0	0	0	0
time_L	-0.0049474	0.1507022	-0.0328291	#NUM!	-0.301828	0.2919331	-0.301828	0.2919331
size	0.1802742	0.0881646	2.0447461	0.0419802	0.0065916	0.3539567	0.0065916	0.3539567
total_bill	0.0940681	0.009504	9.8977781	1.446E-19	0.0753455	0.1127907	0.0753455	0.1127907

Compare 2 models

Model	**R Square**	**Root mean square error**
Model1 - considering all independent variables	0.470078123	1.00516345
Model2 - considering variables sex_M, sex_F, smoker_Y, smoker_N, time_D, time_L, size, total_bill	0.46876847	1.006404768

We select model 1 as the best model considering higher R squared and lower root mean square error.

7. Conclusion

Independent variables sex, smoker, day, time, size and total_bill can be used to predict tip with 47% accuracy using linear regression model.

Model	R Square	Root mean square error
Model1 - considering all independent variables	0.470078123	1.00516345

	Coefficients
Intercept	0.727615964
sex_M	0
sex_F	0.03244094
smoker_Y	0
smoker_N	0.08640832
day_THU	-0.136778538
day_FRI	0.02548066
day_SAT	-0.095977717
day_SUN	0
time_D	-0.068128601
time_L	0
size	0.175992003
total_bill	0.094487006