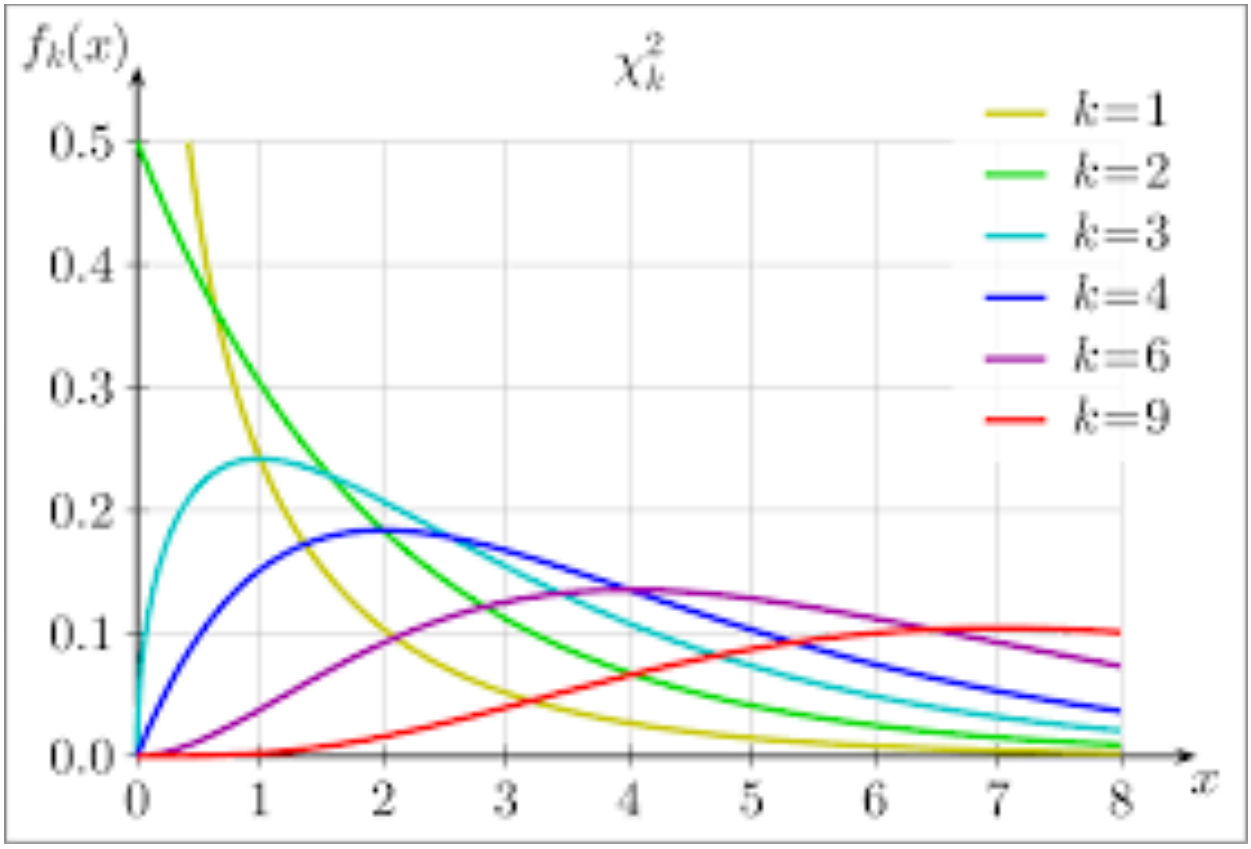
Project week 6

Regression & correlation analysis and Analysis of Variance (ANOVA)



Maryam Heidari

Winter 2019

Project week 6

Regression & correlation analysis and Analysis of Variance (ANOVA)

Introduction

In this project, we going to work with regression and chi-square distribution. we learn how to find a regression and correlation between data, how to measure how much our model fit the actual pattern and so on.

Discussion

Part 1:

We have data set about the location quotients of NY and LA. First and foremost, I collect random sample size 200 from both of these dates and then standardize them by using =STANDARDIZE(x, mean of the sample, SD of the sample). After that, I describe seven groups of range for them and count how many of them in each range by using =COUNTIF and =COUNTIFS. You can see the result in the below table:

Table A								
Observed	LOC QUOTIENTS							
	Z≤-0.5	-0.5<Z≤0	0<Z≤1	1<Z≤2	2<Z≤3	3<Z≤4	Z>4	TOTAL:
NY	65	52	55	18	7	2	1	200
LA	66	51	57	15	8	2	1	200
TOTAL:	131	103	112	33	15	4	2	400

Table 1

The next step is calculating the expected value, we can do that by this formula:
(sum of the row * sum of the column)/Total Sum
And the result is:

Table B								
Expected	LOC QUOTIENTS							
	Z≤-0.5	-0.5 <Z≤0	0 <Z≤1	1 <Z≤2	2 <Z≤3	3 <Z≤4	Z>4	TOTAL:
NY	65.5	51.5	56.0	16.5	7.5	2.0	1.0	200.0
LA	65.5	51.5	56.0	16.5	7.5	2.0	1.0	200.0
TOTAL:	131.0	103.0	112.0	33.0	15.0	4.0	2.0	400.0

Table 2

Now, it is time to perform a chi-squared test of independence in order to find if these two parameters dependent on each other or not. Our hypothesis is:

Ho: LOC QUOTIENTS and locations are independent factors.
Ha: LOC QUOTIENTS and locations are not independent factors.

picture 1

For this step, first of all, I calculated the chi-squared metric for each class by this formula:
χ^2 = (Observed f - Expected f)^2 / Expected f
The result is:

Table C								
χ^2								
	Z≤-0.5	-0.5 <Z≤0	0 <Z≤1	1 <Z≤2	2 <Z≤3	3 <Z≤4	Z>4	TOTAL:
NY	0.0038	0.0049	0.0179	0.1364	0.0333	0.0000	0.0000	0.1962
LA	0.0038	0.0049	0.0179	0.1364	0.0333	0.0000	0.0000	0.1962
TOTAL:	0.0076	0.0097	0.0357	0.2727	0.0667	0.0000	0.0000	0.3925

Table 3

By sum all of the values, we can find the Test statistic χ^2
and also, you can calculate the degree of the freedom by this formula
=(n1-1) *(n2-1)
Now, I can calculate the p-value by this formula = 1 - CHISQ.DIST (χ^2, DF, 1)

You can see the result of all of that in the table below:

Table D	
Test Statistic χ^2	0.4
Degrees of Freedom DF	39601
P-value	1.0000000000E+00
Conclusion:	
There is significant evidence indicating that the two factoes are dependent.	

Table 4

Part 2:

In this part, we have data consists of a sample of LOC QUOTIENTs for both NY and LA for 324 randomly selected.

First, I used =SLOP (range y, range x), =INTERCEPT(range y, range x) and =CORREL(range x, range y) in order to calculate the slope, intercept and a correlation between LA and NY. And for determination, I just power correlation to 2. You can see the result in the table below:

Q2 - (i): Table A	
Slope *m*	0.4806
Intercept *b*	0.8164
Correlation *R*	0.4265
Determination R^2	0.1819

Table 5

The determination R^2 is so small, so this model does not count as a good model.

For drawing the scatter plot, I select all the data of LA and select insert and chose to scatter plot. Then select data, then in x range I select all of the data of NY. Now we have our scatter plot, but we need the regression line and its equation too. For this part, I select one of the points, chose to add trend-line and check the last to box which are display the equation on the chart and display the R-squared value on the chart.

You can see the result of all of these steps below:

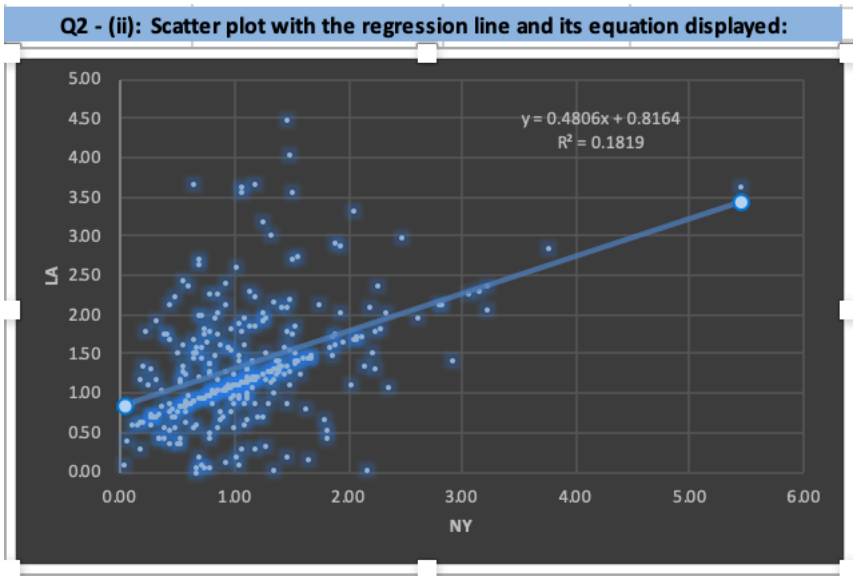


Figure 1

for the next part, I use the slop and intercept in table A to calculate the predicted value. Then, by using the formula of =predicted value- observed value, I calculated the residuals. And I completed the next table which is:

Q2 -(iv): Table B	
Residuals Mean	0.0000
Residuals SD	0.6587
Residuals Minimum	-1.8093
Residuals Maximum	2.9819
Residuals Count	324.0000

Table 6

In order to draw the normal probability plot for the residuals, we need to do some steps before that. I copied the residuals value in another column, so it does not change. For standardized these values, I used the formula
= (value- mean of residuals)/ residuals SD and gave each of them rank. Then I am calculated the cumulative on the left of each rank by using this formula =(i-0.5)/n. After that, I used =NORM.S.INV(x) in order to standardize these values. And Finally, I select this data and draw a scatter plot. In here after drawing the scatter plot for standard z-value, I select standard z-value column as y range and put the standardized residuals as x range, the result is the blue line. And I draw scatter plot one more time and put standard z-value in x range so we could have a red line in order to compare it to the blue line. You can see the result below:

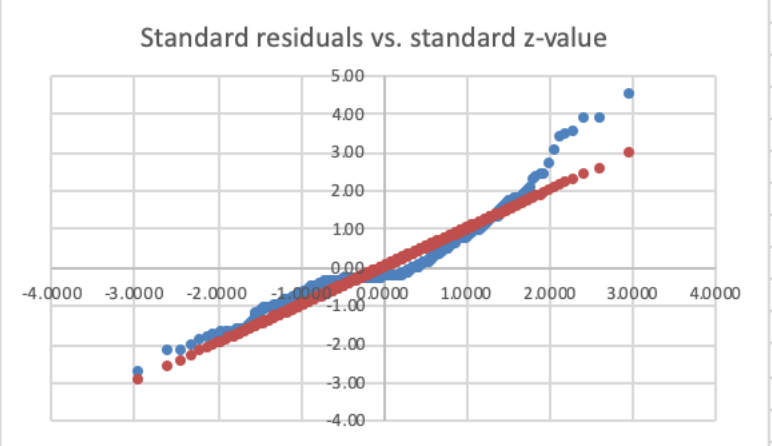


Figure 2

Residuals have to be four conditions:
1)Normally distributed
2)Have a zero mean
3)Have a constant variance (they are Homoscedastic)
4)Are independent
For Homoscedasticity, plot the residuals vs the predicted Y values and there should not exist and pattern in the plot. And For independence, plot the residuals vs the independent variable (or vs time if possible) and there should be any distinguishing pattern.

Because the values are standardized, they have a normal distribution and in table 6 we can see the mean is zero. So, we should just check the condition 3 and 4.
In the next step, I draw the independence and homoscedasticity for residuals.

The results are:

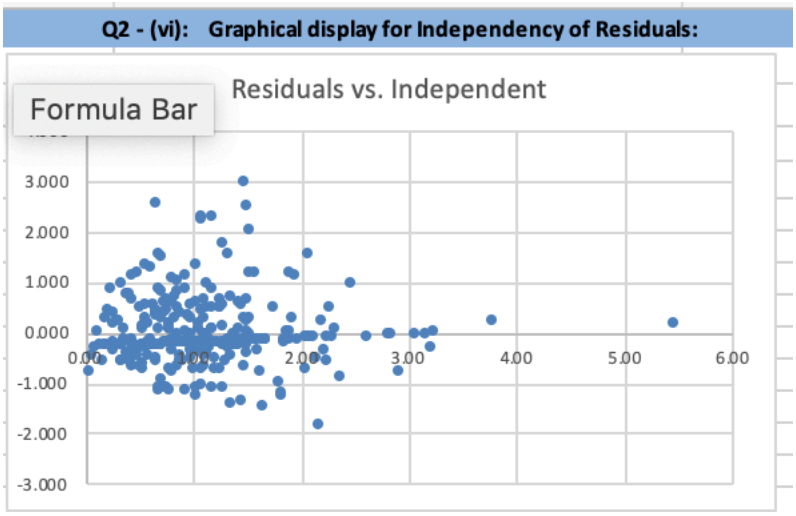


Figure 3

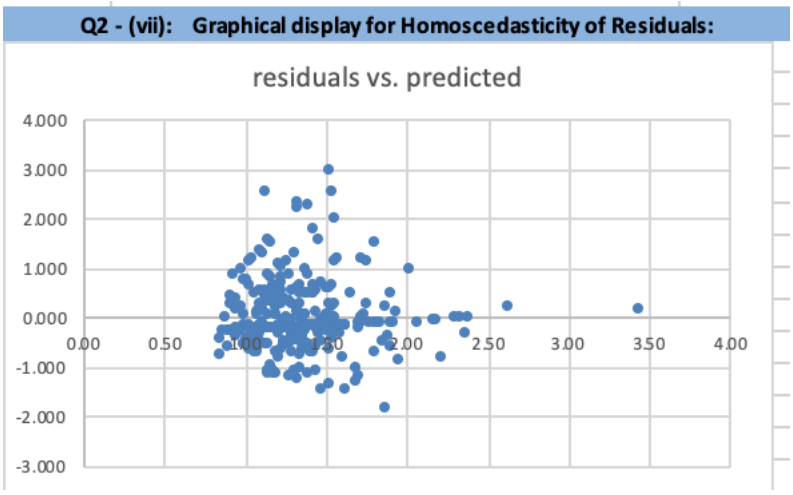


Figure 4

Based on both of these plots shape, we can say our data are satisfied with the homoscedasticity condition, but not the independence condition.

Now it is time to calculate the frequency distribution and chi-squared goodness of fit test for normality of residuals. First and foremost, I used the histogram in data analytics package in order to calculate the frequency distribution and bins. The result is:

	Q2 - (viii): Frequency distribution & Chi-squared Goodness of Fit test for normality of residuals:
Bin	*Frequency*
-1.81	1
-1.54	0
-1.28	3
-1.01	13
-0.74	6
-0.48	27
-0.21	55
0.05	115
0.32	32
0.59	25
0.85	16
1.12	9
1.38	10
1.65	4
1.92	1
2.18	1
2.45	3
2.72	2
More	1

Table 7

And then I draw the histogram graph for it:

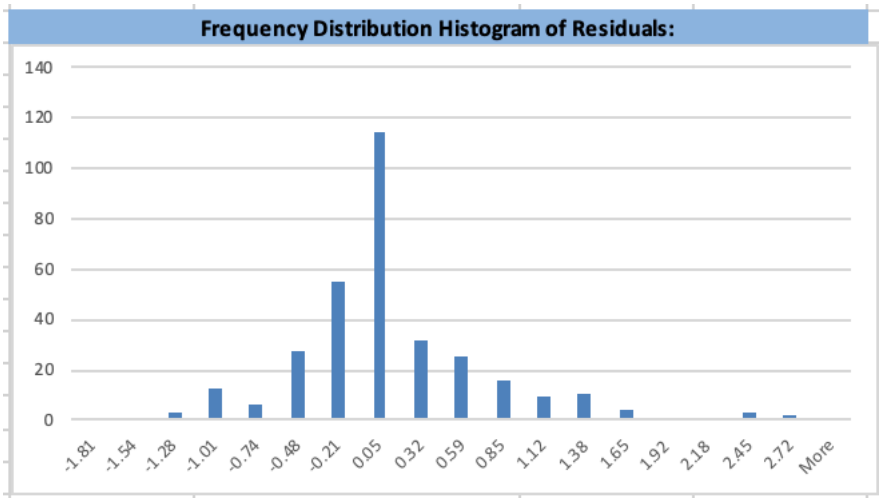


Figure 5

In order to calculate the left and right class of end, I calculated half of the difference between each bin. The lift class end is equal to bin mines this value and the right class end is equal to bin plus this value. In here we have 2 exceptions which are left and right class end of the last bin. The left class of last bin equals to the right class of the bin before that and the right class can calculate by adding the same distance to the left class. The next step is calculating the normal probability. For the First one =NORM.DIST(right end,0, residual SD,1), for others we used the =NORM.DIST(right end,0,SD,1)- NORM.DIST(left end,0,SD,1), and for the last one =NORM.DIST(left end,0,SD,1). For checking, if you sum all of them the result should be one. After that, we need to calculate the Expected Frequency which is total observe frequency multiple by normal probability. And finally fo

calculates the χ2, we should use this formula (observed frequency- expected frequency) ^2/expected frequency. The results of all of these steps are:

	AF	AG	AH	AI	AJ	AK	AL
	Class Left End	Class Right End	Class Midpoint	Observed Frequency	Normal Probabilities	Expected Frequency	(Observed - Expected)^2 / Expected
1	-1.9424	-1.6762	-1.8093	1	0.0054692	1.766558	0.33
2	-1.6762	-1.4100	-1.5431	0	0.0106854	3.451373	3.45
3	-1.4100	-1.1438	-1.2769	3	0.0250858	8.102719	3.21
4	-1.1438	-0.8776	-1.0107	13	0.0501281	16.191376	0.63
5	-0.8776	-0.6115	-0.7446	6	0.0852626	27.539829	16.85
6	-0.6115	-0.3453	-0.4784	27	0.1234437	39.872309	4.16
7	-0.3453	-0.0791	-0.2122	55	0.1521309	49.138280	0.70
8	-0.0791	0.1871	0.0540	115	0.1595907	51.547807	78.11
9	0.1871	0.4533	0.3202	32	0.1425084	46.030200	4.28
10	0.4533	0.7194	0.5863	25	0.1083213	34.987765	2.85
11	0.7194	0.9856	0.8525	16	0.0700847	22.637366	1.95
12	0.9856	1.2518	1.1187	9	0.0385978	12.467094	0.96
13	1.2518	1.5180	1.3849	10	0.0180935	5.844201	2.96
14	1.5180	1.7841	1.6511	4	0.0072192	2.331816	1.19
15	1.7841	2.0503	1.9172	1	0.0024516	0.791879	0.05
16	2.0503	2.3165	2.1834	1	0.0007086	0.228878	2.60
17	2.3165	2.5827	2.4496	3	0.0001743	0.056300	153.91
18	2.5827	2.8489	2.7158	2	0.0000441	0.014250	276.72
				323	1.0000	323.00000	554.91

Table 8

Now we have χ2 which is 554.91. The degree of freedom is also DF = n - K - 1 where n is the number of classes in the distribution (18) and the number of parameter estimates used to calculate the normal probabilities (1), so the DF is 16. Now, we can calculate the p-value by using = 1 - CHISQ.DIST (χ^2, DF, 1).
The result is:

Table D	
χ^2	554.91
DF	16
P-value	0.0000000
Decision:	Null hypothesis is rejected

Table 9

The null hypothesis is rejected, so there is sufficient evidence to reject the fact that the residuals belong to the Normal distribution.

Conclusion

In this assignment, we learn about regression and how to calculate that. Moreover, we learn about chi-square and the goodness of the fit of our model.