Descriptive analysis of statistical data

Business

Final Project: Statistics II Descriptive analysis of statistical data INTRODUCTION There have always been crimes, from a treachery to an assassination. Happens in every country you can think of, and every government has to deal with it. It is really stressful to try to understand the nature of the crimes: why are they done and where could they happen next. Out of this preoccupation is that we found studies gathering data from communities; we focused on one specific crime: murders. In several communities, it is thought that the murder rate is somehow related to several factors.

For instance, it is common to hear that the murder rate depends on poverty and unemployment.

Starting from this hypothesis, the database found to make this analysis relates the number of murders per year per 1, 000, 000 inhabitants with the number of inhabitants, the percentage of families' incomes below $5000, and the percentage unemployed. OBJECTIVE OF THE STUDY Trying to estimate how many murders will happen in a year in a specific place is difficult, but not impossible. This is why we are using the dataset found with the variables mentioned above, with which we'll be able to find a formula.

So, after this project, if we want to know how many murders will be on a city, for example Monterrey, we'd just plug in the data from that city (the inhabitants, the percentage of families income below $5000, and the percentage unemployed) and we'll get a number, which would be the predicted number of murders in that specific year. DESCRIPTION OF THE PROCEDURES USED For this study, we used a computer package named

Minitab. Minitab is a PC program designed to launch statistical functions, both basic and advanced.

Also, we used Microsoft's Excel. First of all, using Excel we did a scatterplot (the graph that connects each X and Y) of each variable paired with murders, and we added a trend line on each graph to see if there was a linear trend in them that would mean a linear relation. In other words, to see if when the percentage of any of the independent variables rose, the murder rate rose as well. In each of the three graphs of each variable with a line marking the trend the data follows: Scatterplot relating Inhabitants (Y) with murders (X). We can observe that there is not a linear trend which most probably means that the variable Inhabitants is not linearly related with the number of murders.) Scatterplot relating % of people with income below 5000 (Y) and murders (X).

Scatterplot relating the % of people unemployed (Y) with murders (X) Using Minitab, we wanted to test if the factors (% of unemployment, % of families with income below 5000, and the inhabitants), in general, were really influent on the murder rates. This means that we had to see if the three variables helped us to estimate the number of murders that happened in a year.

So, the first thing we did was to use a procedure called regression, which finds the best linear equation between the variables observed. This equation is the one that can be used to do later predictions about the murders with a specific given percentage of unemployment, number of inhabitants and percentage of families with income below $5000. Minitab gave us the equation, but also many other information.

We saw some data that gave us the idea that maybe some of the factors (variables) used were not being helpful.

So, because our measure of how much evidence we have against the hypothesis " it is not useful" (called P-value) was big, we decided to test if the factor " inhabitants" was being useful; this means we wanted to see if it is a significant variable in the role of estimating the murder rate. After running another test not including inhabitants (called partial F test), we compared both tests (the one with inhabitants and this new one) and we found out that the number of inhabitants was not an important determinant of the murder rates; in other words, the murder rate were not related to the number of people living in one specific area.

Finally, we decided to drop this variable, so we could perform an even better predictive equation. Now we are using the equation given at the partial F test. It is common that some times one of the independent variables (in this case, percentage of families with income below $5000 and percentage people unemployed) may be linearly related with another independent variable.

In other words, there may be a pair of explanatory (independent) variables that measure the same thing or that they depend on each other, so it is recommended to drop one of them because having both causes instability in the model.

This means that the equation wouldn't be as precise as it should. This issue is called multicollinearity. To prove if this is the case in our study, we calculated some numbers called correlation coefficients and variance

inflation factors. Basically, the rule is that if one of the correlation coefficient between a pair of explanatory variables is higher than 0.

9 there is evidence of multicollinearity. Another criteria is that if the variance inflation factor of each explanatory variable is higher than 10, there is a case of multicollinearity, as well.

In our study, multicollinearity was not an issue. Another issue when having more than 1 independent variable as predictors is that the presence of one of them may affect the prediction capacity of the other. For being sure that this wasn’t happening, we proceeded to perform a stepwise procedure. In this procedure one variable is entered to the equation at a time, and simultaneously, the prediction capacity of the variable is being tested, this helps to see which variables should be included and should not be included in the final equation.

Given that in this case there were only 2 explanatory variables, and after doing this procedure, we decided to keep both variables, because mainly, they didn’t affect their prediction capacity. This stepwise procedure helped us conclude that the equation was better when both variables were included. This means that the equation was going to be a better murder predictor when having both variables in it. Additionally, when doing this type of equations, there are several assumptions that should be satisfied in order for the equation to be valid.

These assumptions are: (Before explaining these assumptions, it is necessary to explain the meaning of statistical error.

A statistical error is the difference between the observation from a sample (in this case the number of murders that really happened) and the predicted observation (gotten from the equation, in this case the value of murders that would result from the equation, given specific values for % of unemployment and % of families with income below $5000).) 1.

At any given value of independent variables, the average of the error values should be 0. 2. Normality assumption. At any given combination of independent variables, all the error term values have a normal distribution.

In other words, if plotted into a histogram it should have the form of a bell curve. This is the form of a normal distribution. 3. Constant variance assumption. At any combination of independent variables, all the error term values have a variance that does not depend on the combination of the values.

The variance is a measure of how far apart the set of numbers are from each other.

4. Independence Assumption: At any given combination of the independent variables, the error term is independent from any other error term. In the case of this specific study, it was necessary to confirm if the normality and constant variance assumptions were being satisfied. To prove the normality assumption, we performed a histogram of the error terms (residuals), which has to have a bell curve shape in order for this assumption to be satisfied.

To be more objective, we performed a probability plot, which related the percentage of the residuals observed with the theoretical percentages.

For the assumption to be satisfied this plot has to have a linear form. To prove the constant variance assumption, we performed a scatter plot relating each error versus each predicted value of murders given specific percentages of unemployment and of families with an income below $5000. If this scattered plot looks like a horizontal band of equal width then the constant variance assumption would be satisfied.

According to the results obtained (shown in the Results section), we concluded that the normality assumption was satisfied, which means that all the observations of the sample are distributed in a way that the middle observations (if arranged in a descending or ascending order) are the most frequent and the highest and lowest observations are the least frequent. On the other hand, there were doubts about constant variance assumption, so to be safe, a new equation was calculated with an adjustment made to the values of murders observed from the sample to be able to get constant variances.

This means that we were not sure if all of the results are equally apart from each other in distance, and this was our goal. Before being able to use the equation as a confident predictor for murders, two final procedures should be performed. It is necessary to make sure that there are no observations of murders, % of families with income below $5000, or % of unemployment from the sample that fall out of the general linear pattern. To detect any case of this issue we had Minitab calculate a table with several values for each observation (leverage and studentized residuals).

Basically, the rule to detect outliers for any value of observations of the independent variables is that if a leverage of an observation is greater than

twice the average of all leverages, then there is evidence for that specific observation falling out of pattern. The rule to detect outliers for any value of the observations of the number of murders is that if any of the studentized residuals is greater than 2, then there is evidence for that specific value falling out of pattern as well.

In this case, any leverage higher than 0. 3 would be a sign of an influential value in the independent variables.

These types of sample observations are undesired for calculating our equation because they are not part of the " usual". In other words, it may have been that one of the observations recorded was an exception or an unusual case. Finally after all of this long procedure, the equation is ready to be used confidently. It is important to note that the values obtained in a prediction are not necessarily the exact values; it is just an approximate prediction.

In the next section, the results obtained from all the procedures performed are shown to demonstrate and validate how we got to the best final model.

RESULTS First of all, when doing the regression with all three factors, Minitab gave us this output: Regression Analysis: Murders versus Inhabitants, %below5000, %unemployed The regression equation is Murders = – 36. 8 + 0. 000001 Inhabitants + 1. 19 %below5000 + 4.

72 %unemployed Predictor Coef SE Coef T P Constant -36. 765 7. 011 -5. 24 0. 000 Inhabitants 0.

00000076 0. 00000064 1. 20 0. 248 %below5000 1. 1922 0.

5617 2. 12 0. 050 %unemployed 4. 720 1. 530 3.

08 0. 007

In this table is where we saw that the P-value from Inhabitants was larger than the others, giving us a reason to test its efficiency on the equation. When doing this test, the partial F test, we had to compare the F that resulted in the test (which in this case was 1. 4373) and the F from the statistical tables, which was 4. 49. Having these results means that the factor inhabitants, was indeed, not relevant for making a precise test; this is the reason we dropped the factor, because of this information.

So now, we are using the equation without the factor inhabitants: The regression equation is

Murders = – 34. 1 + 1. 22 %below5000 + 4. 40 %unemployed Predictor Coef SE Coef T P Constant -34. 073 6.

727 -5. 07 0. 000 %below5000 1. 2239 0. 5682 2.

15 0. 046 %unemployed 4. 399 1. 526 2. 88 0. 010 What we did next was to calculate the correlation coefficient.

This is the result: Pearson correlation of %below5000 and %unemployed = 0. 815 We only have one correlation coefficient because these numbers are calculated with each pair of variables, and here we only have one. Because it is smaller than 0. we conclude that both variables (%unemployed and % of families with income below 5000) are important to the test and that this is not a case of multicollinearity. In the stepwise procedure, we had this output:

Stepwise Regression: Murders versus %below5000, %unemployed Alpha-to-Enter: 0.

05 Alpha-to-Remove: 0. 05 Response is Murders on 2 predictors, with N = 20 Step 1 2 Constant -28. 53 -34. 07 %unemployed 7. 08 4.

40 T-Value 7. 31 2. 88 P-Value 0. 000 0. 010 %below5000 1.

22 T-Value 2. 15

P-Value 0. 046 S 5. 10 4. 65 R-Sq 74. 80 80.

20 R-Sq(adj) 73. 39 77. 87 Mallows Cp 5. 6 3. 0 The important thing here is to see that at the last step (in red) both variables are included, meaning that the test and the equation is better with both variables, and that they don’t interfere with the other’s prediction capacity.

After this, we were testing for the assumptions. The first thing we did was the histogram: But because we are not very sure if has, or not, the bell curve shape, we did what is called the probability plot:

And here, it is clearly seen that it has a linear shape (marked with the line between the dots), so we can conclude that the normality assumption was being satisfied. When trying to see if the constant variance assumption was being satisfied, we did the scatter plot (previously explained), but as seen, we are again, not sure of the shape it has. Scatter plot before adjustmentScatter plot after adjustment After doing the adjustment to the observation “ murders”, we got the new scatter plot which has a same width size, and this output from Minitab: The regression equation is

Y* = 0. 085 + 0.

0527 %below5000 + 0. 256 %unemployed Predictor Coef SE Coef T P VIF Constant 0. 0847 0. 4101 0. 21 0. 839 %below5000 0.

05273 0. 03464 1. 52 0. 146 2. 984 %unemployed 0. 25619 0.

09305 2. 75 0. 014 2. 984 After this, our final step was to make sure there weren’t any values falling out of the linear pattern, so we made Minitab calculate their studentized values and leverages: RESI1 SRES1TRES1 HI1COOK1 -0. 127102-0.

47488 -0. 463790 . 1082530. 09125 -0. 209901-0. 78823 -0.

779060. 1172540. 027509 -0. 147683-0. 61464 -0.

603020. 2813220. 049293 -0. 644740-2. 46517 -2.

983590. 1485030. 353287 0. 2436610. 89669 0.

891250. 0808230. 023566 0. 0754420. 28093 0.

273180. 1022670. 002997 0. 2504110. 93186 0.

928050. 1011020. 032556 0. 4203881. 53489 1.

604380. 0662010. 055673 -0. 083556-0. 34170 -0.

332640. 2556400. 013366 -0. 216495-0. 93317 -0.

929430. 3299940. 142965 0. 0980090. 36137 0. 35193 0.

0843120. 004008 0. 0936950. 35671 0. 34736 0. 1411380.

006970 0. 1760330. 7308 0. 661870. 1485470.

026346 -0. 001159-0. 00451 -0. 004370. 1766960. 000001 0.

1653150. 60072 0. 58907 0. 0572530. 007305 -0.

160065-0. 62710 -0. 61554 0. 1889870. 030546 -0.

043565-0. 16003 -0. 15537 0. 0774520. 000717 -0. 218575-0.

84348 -0. 835980. 1640940. 046555 -0. 224403-0. 90745 -0.

902490. 2387630. 086094 0. 5542902. 09837 2.

36489 0. 1313990. 222032 In this procedure we could find out that there was 1 influential observation in the independent variables (in purple), and 2 in the murders observations (marked in red).

So we took them out of the test to get a more satisfying equation. So, the final equation (without those unusual observations) is Y* = 0. 369 + 0.

0611 %below5000 + 0. 194 %unemployed NOTE: This equation had to be adjusted, so when using the equation, we have to elevate E to the answer. CONCLUSIONS It is worth mentioning that making a census is quite a difficult, costly and a time consuming task. Therefore, we have statistical procedures to be able to generalize conclusions based on a mere sample.

This is why to reach the final equation we had to go through a lot of steps, so the sample could be generalized, although we have to keep on mind that the number that may result from the equation is an estimation, so the real number, in this case the murder rate that actually happened, may be around

our estimation, not necessarily the exact same number. These statistical procedures are very useful, not only to know how many murders will there be in a certain place, but also to do any other type of investigations.

To conclude, we'd like to use our equation to estimate the murder rate in Monterrey, Mexico this year (2012) based on the 2011, 5. 7 % unemployment and % of families with income below 5000. Y* = 0. 369 + 0. 0611 (51. 93 %below5000) + 0.

194 (5. 57 %unemployed) Y*= 0. 41153503 Y* adjusted = 1. 5091326 REFERENCES INEGI. (2012, Enero 31).

Instituto Nacional de Estadistica y Geografia. Retrieved from Poblacion, Hogares y Vivienda: http://www. inegi. org. mx/Sistemas/temasV2/Default. aspx? s= est&c= 17484 Kleinbaum, D.

G. (1978). Applied Regression Analysis and Other Multivariable Methods. Duxbury Press. Spaeth, H. (1991).

Mathematical Algorithms for Linear Regression. Academic Press.