Tristan Conde

ADTA 5160

Spring 2022

# Assignment 01

1. Open Stata. Open the Stata Data Editor and cut and paste your data from Excel into Stata.
   1. This was successfully completed by importing the data from Excel using the Copy and Paste Method.
2. Produce summary statistics for the following variables: seatqual, source, stdage, club and the year of sale indicator variables. Briefly describe findings that you think are relevant from this summary statistics table.
   1. Table

      Description automatically generated
   2. Findings:
      1. Seatquality is missing 3 observations
      2. 2008 has the highest mean observation.
      3. The average customer seat quality is much higher than expected, giving us almost a 4 if you round up in the average data.
      4. The stadium was on average around 10 years old during this dataset.
      5. Lower mean in club could infer that most dataset revolves around sales not always being made for a premium location
      6. Dataset provides that a lot of these sales are probably from 2007 and up, as 2005 and 2006 have significantly lower means than the other years.
3. What percentage of sales are for seat locations on the aisle? How do you know this?
   1. Table

      Description automatically generated
   2. 1.33% is in the aisle. The variable keys in the data specify “1 = seating location is on the aisle; 0 = otherwise”
4. How would you describe the distribution of stdage? Use a Stata command to produce the skewness value for stdage and explain what this value tells you.
   1. Table

      Description automatically generated
   2. Skewness value helps to show the measure of asymmetry of the distribution itself. The value we have is -1.63, which would indicate a skewed graph to the left.
5. Produce a frequency distribution for seatqual. What is the most common seat quality type in the data? Does this surprise you? Why or why not?
   1. Table

      Description automatically generated
   2. Most common seat quality type from our data would be quality 5.
   3. This is quite surprising, as it shows that we have at least 41% of people who really say the seats in the stadium they would rank as the highest quality available. We also see that over 50% at least think of the seats as a 4 or above, which could infer that most people really like the quality of the seats.
6. Produce a histogram for seatqual with a primary title of “Distribution of PSL Sales by Quality Level” and a subtitle of “2005-2009 Data”.
   1. Chart

      Description automatically generated
7. What is the mean for the variable source?  Is this value meaningful? What does it tell you? Should it be used when describing the distribution of this variable? What could be done to this variable to make it appropriate for use in a regression?
   1. Table

      Description automatically generated
   2. Is this value meaningful?
      1. I do find this value to be meaningful based on the numerical result given.
   3. What does it tell me?
      1. Since the mean is 1.93 and the two results of this variable can either be 1 or 2, having the number be incredibly close to 2 can infer that it is most likely that the option 2 is the more common answer for this variable.
   4. Should it be used when describing the distribution of the variable?
      1. Personally yes, since the mean, std. error and 95% confidence interval I do believe infers enough to say that it would be more likely for variable answer 2 to be the more common answer.
   5. What could be done to this variable to make it more appropriate for use?
      1. Instead of a 1 or 2, I would have done a 0 or 1 variable, since the source variable is just based on where the person bought their ticket, and a 0 or 1 could give us a more clear answer since it will instead base its mean on how many 1 answers there are.
8. Identify the mean and the standard deviation for ticketprice. Based on what you know about the mean and standard deviation of normal distributions, 95% of the face value ticket prices in this sample are theoretically supposed to fall in between what two values? Does this give you any indication of whether or not ticketprice is normally distributed? Please explain.
   1. Table

      Description automatically generated
   2. 95% of the face value ticket prices should fall between $113 and $122.
   3. While one could say yes since the mean of the ticketprice falls within our confidence interval, the data here alone I don’t believe is enough. A histogram or QQ plot should be made to help corroborate this analysis.
9. Produce a two-way frequency distribution for source and club. Describe what this table tells you.
   1. Table

      Description automatically generated
   2. Source option two (transaction occurred on seasonticketrights.com) is significantly more picked than source option 1, having 838 observations have this option as their source choice versus the 63 that have source option 1 (transaction occurred on Ebay.com)
   3. From source option 2, club option 0 (indicating sale was made not for a premium location) is the most common pick, having about 76% have this specific set of options selected versus the 24% who picked options 1 and 1 (implying bought on eBay and also for a premium location)
10. Estimate a linear regression model where the dependent variable is priceperseat. Include the following independent variables as predictors: trend, seatqual, stdage, stdcapacity, club, win3, localunemp
    1. Table

       Description automatically generated
11. How many observations were included in the estimation of this regression based on the model statistics? Is this number different than the number of observations in the Excel file which includes the raw data? Explain why or why not.
    1. There is a mild difference between the number of observations. Our regression model has 898 observations versus the 901 total observations within the dataset.
    2. The one thing that would take into account is the degrees of freedom will have the number of observations minus 1, so we have to take that into account, and then the other option is that there are likely null options in our data that Stata is skipping over.
12. What does the model F statistic tell us? Is this value related to “Prob > F”? What does “Prob > F” tell us?
    1. The F stat tells us about the ratio between the Mean Square Model and Mean Square Residual. The numbers 7 and 890 represent the model and residual degrees of freedom.
    2. Yes, as the Prob > F value is dependent on how well the regression model can obtain the estimated F value or greater.
    3. Since the Prob > F value is equal to 0 (which Stata signifies as below 0.00005) which indicates that the independent variables can reliably predict the dependent variable and thus can reject the null hypothesis.
13. What is the R-squared value of the regression model? What does this value tell us?
    1. R-Squared value is 0.5077 and adjusted value of 0.5039
    2. This tells us that our independent variables can predict the output of the dependent variable (priceperseat) about 50.39% of the time (if you take into the adjusted r-squared value), otherwise it would be 50.77% of the time.
14. How many independent variables are included in the regression model? How many are significant at the 10% level and how many are not significant at the 10% level? Do you see any relationship between the statistical significance and the t-value for each independent variable?
    1. We had 7 independent variables in our regression model.
    2. We have 4 variables significant at the 10% level and 3 variables that are not significant at this level.
    3. Not entirely. Each variable seems to have a different t-value and doesn’t seem to be fully correlated or related. We have some variables that are statistically significant with incredibly low p-values but vastly different t-values. We can have one significant value with a t-value of 3.19 and then the next one at -18.
15. For each independent variable included in the regression, in two sentences or less, explain the result of each coefficient as if you were writing up the results in a report to a stakeholder. Your two sentences should tell the reader whether or not the independent variable is statistically significant and should explain the relationship between the independent variable and the dependent variable.
    1. Trend – For the trend variable, it’s p-value is extremely high at a 0.458 and also has a very low t-value of 0.74. With this, it is safe to say that trend is not statistically significant in relation to price per seat.
    2. Seatqual – For our seat quality variable, the p-value and t-value is the lowest of all our variables, with our p-value being below 0.005 and has a t-value of -18.08. This also has an extremely interesting co-efficient of -1201 and the t-value implying that the coefficient estimate is 18 times smaller than the standard error of 66.48. With this information alone, it is very safe to say that seat quality plays an extremely significant statistical role in relation to price per seat.
    3. Stdage – Our stdage variable has a relatively high p-value of 0.354, but inversely a t-value of -0.93. This stat also has a relatively high standard error around 343. That being said, due to its high p-value, it is safe to say it is not statistically significant to priceperseat.
    4. Stdcapacity – Stdcapacity follows pretty closely behind to seatqual in terms of p-value, also making a 0.000 (below 0.0005) value, as well as a negative t-value at -3.77. Stdcapacity also has the lowest standard error at a mere .181, which infers that this variable is not significantly different from zero in the slightest. Based on the p-value, we can conclude that it is statistically significant.
    5. Club – club also has a significantly low p-value, holding onto a 0.001 value, as well as a positive t-value at 3.19. Although it is not as low as stdcapacity or seatqual, the p-value of 0.001 still concludes that it is statistically significant.
    6. Win3 – Win3 has a very high p-value of 0.431 and also a t-value of 0.79. This variable also has an incredibly high standard error at around 2955, which is very high and shows it is significantly different from zero. That and the high p-value can conclude that this variable is not statistically significant.
    7. Localunemp – Localunemp has quite a low p-value at 0.003 and a t-value of -3.01. This variable also has a low standard error of 150, which is much lower than most of our other variables. While not the lowest p-value, it still passes at the 5% significant level, and therefore makes it statistically significant.

Grade: 55/60