Notes on summary statistics and tables

ECON 6901

The following notes provide a discussion on creating tables and figures for your research papers. The sample data used for the discussion are 980 home sales from Santa Clara County in June 1987. The point is to illustrate some common mistakes when creating tables and figures, and also to explain how you can use them to help identify errors in your data.

Table 1 shows a typical "data dump table" from Stata.[1] I simply copied and pasted the information from Stata into Excel and then pasted a table into the MS Word document. While the table provides the required information, it does not display the data in any rational manner. The variable names are the same as they are in the Stata data file. While most of their meanings can be easily inferred from what is presented, I have no idea what "pprice" means.

Table 2 shows the same information, though the variables are labeled more clearly, the table is formatted a little better, and the caption provides some additional information. Table 2 is much better than Table 1, but can we improve Table 2?

First, note that I have removed some variables in Table 3. No one really cares too much about the summary statistics for the squares of living area, lot size, and age. They are included in the data set because they might be used to control for nonlinear relationships between these variables and price. No model is going to include the square of one of these variables without including the variable itself. Also, the means for these variables, particularly lot size and living area, are so much larger than the means for other variables that they overwhelm the table.

Second, I have rearranged the order of the variables in Table 3. You should not let the order of the variables in your data set dictate the order in which you present them in a table. You will note that I have separated the price variables, the continuous variables (age, living area, and lot size), the count variables, and the binary variables into different groups. I have done this separation to make it easier to compare and contrast different groups of variables. From Table 3 it is very clear to see that 84% of the houses had fireplaces, 14% had swimming pools, and 2% had spas. You should also be able to see that I have dropped the decimal places for all of the continuous variables as well as previous sales price and sales price. There are too many numbers otherwise.

Finally, I have added some spacing in Table 3 to separate the groups even further. This spacing can help draw attention to the distinct groups of variables.

Table 3 could likely be improved further. There are 980 observations for each variable, and I have that information in the caption. However, I included the number of observations for each variable to make sure that no records are missing any information.

[1] I know that I am violating our guidelines about one table or figure per page. These are notes, not a research paper, and I have been careful about how the pages appear.

Variable	Obs	Mean	Std. Dev.	Min	Max
pprice	980	71351.38	65705.37	0	624000
price	980	186365.6	83514.07	45000	970002
livingarea	980	1555.502	444.1384	544	4402
lotarea	980	7411.157	3947.853	1296	50529
stories	980	1.191837	0.399098	1	3
rooms	980	6.791837	1.52592	3	15
bedrooms	980	3.384694	1.11996	1	30
bathrooms	980	1.889796	0.493159	1	3
age	980	23.64694	12.45953	2	79
livingarea2	980	2616644	1598503	295936	1.94E+07
lotarea2	980	7.05E+07	1.62E+08	1679616	2.55E+09
age2	980	714.2592	843.2371	4	6241
fireplace	980	0.842857	0.364121	0	1
swimmingpool	980	0.137755	0.344819	0	1
spa	980	0.019388	0.137954	0	1
lnprice	980	12.06148	0.36751	10.71442	13.78505

Table 1 Summary statistics -- output dump from Stata

Variable	Obs	Mean	Std. Dev.	Min	Max
Previous price	980	71351.38	65705.37	0	624000
Sales price	980	186365.6	83514.07	45000	970002
Living area	980	1555.502	444.1384	544	4402
Lot size	980	7411.157	3947.853	1296	50529
# Stories	980	1.191837	0.399098	1	3
# Rooms	980	6.791837	1.52592	3	15
# Bedrooms	980	3.384694	1.11996	1	30
# Bathrooms	980	1.889796	0.493159	1	3
Age	980	23.64694	12.45953	2	79
$(\text{Living area})^2$	980	2616644	1598503	295936	19377604
$(\text{Lot size})^2$	980	7.05E+07	1.62E+08	1679616	2553179841
Age^2	980	714.2592	843.2371	4	6241
Fireplace – binary	980	0.842857	0.364121	0	1
Pool – binary	980	0.137755	0.344819	0	1
Spa – binary	980	0.019388	0.137954	0	1
ln(Sales price)	980	12.06148	0.36751	10.71	13.79

Table 2 Summary statistics for 980 single-family homes sold in Santa Clara County in June 1987

Variable	Obs	Mean	Std. Dev.	Min	Max
Previous price	980	71,351	65,705	0	624,000
Sales price	980	186,366	83,514	45,000	970,002
ln(Sales price)	980	12.06	0.37	10.71	13.79
Age	980	24	12	2	79
Living area	980	1,556	444	544	4,402
Lot size	980	7,411	3,948	1,296	50,529
# Stories	980	1.19	0.40	1	3
# Rooms	980	6.79	1.53	3	15
# Bedrooms	980	3.38	1.12	1	30
# Bathrooms	980	1.89	0.49	1	3
Fireplace	980	0.84	0.36	0	1
Swimming pool	980	0.14	0.34	0	1
Spa	980	0.02	0.14	0	1

Table 3 Summary statistics for 980 single-family homes sold in Santa Clara County in June 1987

When you look at Table 3, is there anything that stands out? To me, there are three things. One is that there are previous sales prices that are equal to 0. Those entries might be errors, or they might be entered correctly.

- If they are errors, can they be fixed?
- If they cannot be fixed, should you use that variable in analysis?
- How would you determine if they are errors and could be fixed?

When looking at the source data, it becomes apparent that there are many entries with previous sales price equal to 0. There are exactly 233, out of 980, observations with a previous sales price equal to 0. They are not all "young" homes – the first observation with a previous sales price of zero is a 41-year-old home; the second is a 16-year-old home; the third is a 30-year-old home. Because there is no external identification (address, parcel number, etc.) for the homes in this data set, it is impossible, or not very easy, to determine why the previous sales price is entered as a 0. Thus, I would not use previous sales price in my analysis – in fact, I probably would not include it in the table.

Now, suppose that you found that all of the homes that had a previous sales price of 0 also had an age of 2-3 years old. If that were true, you might infer that they had not been previously sold. In that case, you might want to structure your econometric model to incorporate homes that sold for the first time versus homes that had been previously sold.

One other entry stands out. How is it possible that the maximum number of bedrooms is greater than the maximum number of rooms? That cannot be true. It turns out the house is a 6 room house. It is likely that there are 3 bedrooms, and that entry was incorrectly entered. Once that observation is removed, the maximum number of bedrooms is 7, which is a more reasonable maximum number given the maximum number of rooms in a house.

The 15 room house has 6 bedrooms, a living area of 4402 sq. ft., and a lot size of 32,015 sq. ft. It is a two-story house. Everything in that record is large, so I would not doubt the accuracy of that record.

Finally, the maximum sales price is $970,002. Now, it is certainly possible that a house sold for $970,002 in California. However, when I look at the records for the three highest priced houses I see that they sold for $970,002, $920,006, and $750,002. They have living area sizes of 1040, 912, and 1040 square feet, respectively. They are the only three observations that did NOT have a zero as the last digit in the sales price of the house. The one that sold for $970,002 had a previous sales price of $13,000. It is very likely that these sales prices should be $97,000, $92,000, and $75,000, respectively.

While looking at minimums and maximums will not catch every error, you should catch the most egregious ones.

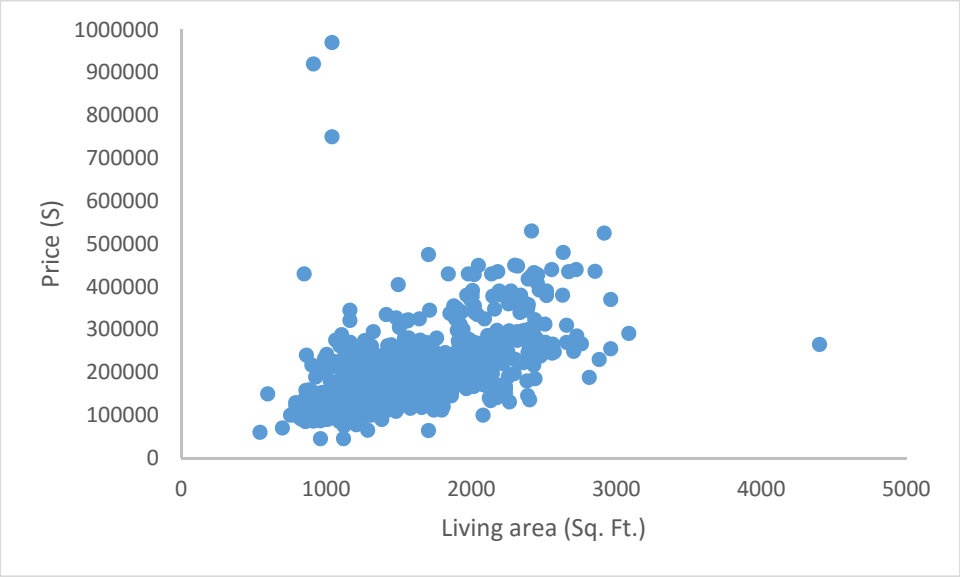

Figure 1 Scatterplot of living area vs. sales price for 980 homes sold in Santa Clara County in June 1987

Alternatively, we could examine plots of the data to determine unconditional relationships and to determine outliers. Our three tiny, but expensive, houses are the outliers in the upper left of Figure 1. There seems to be a positive relationship between price and the living area of the house, and the correlation coefficient is 0.52.[2]

Figure 2 shows what is essentially a density function of home sales. Figure 3 shows a histogram of home sales. Both figures show the same information but in a different format. Figure 3 is likely to be the more readily understood by most readers.

[2] I have left the likely errors in the data in the calculation of the correlation coefficient.

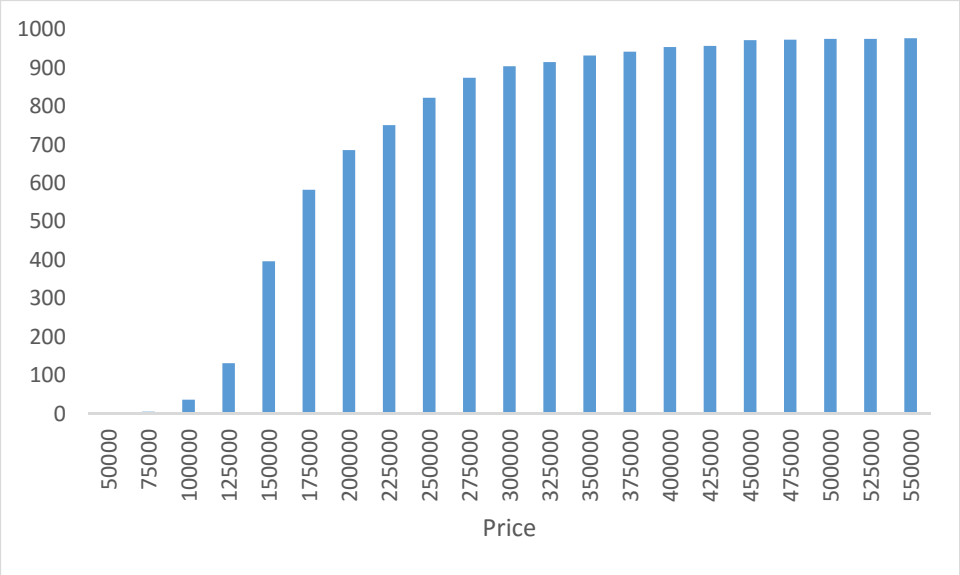

Figure 2 Histogram of sales prices for the 980 homes sold in Santa Clara County in June 1987. The figure shows the number of homes that sold for a price less than the stated price on the horizontal axis.

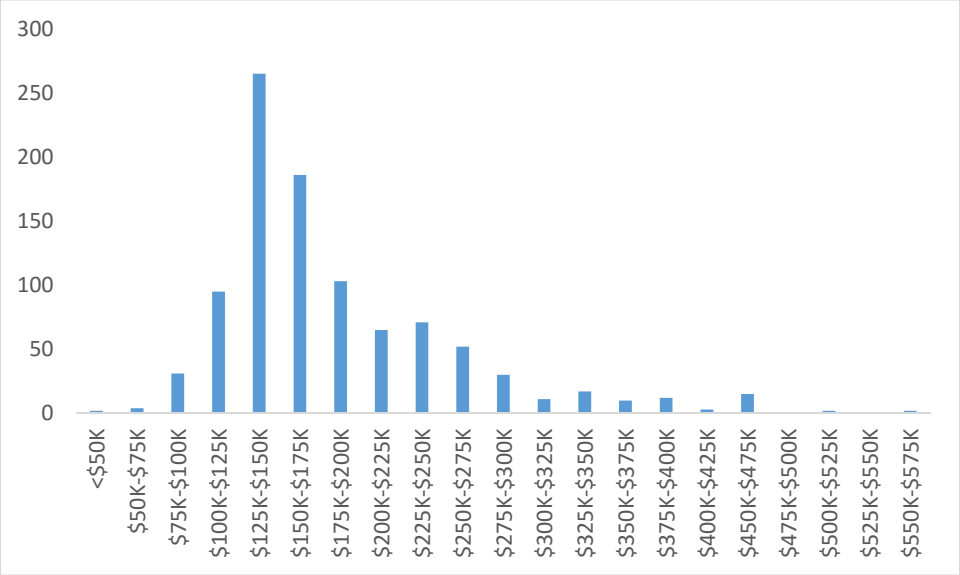

Figure 3 Histogram of sales prices for the 980 homes sold in Santa Clara County in June 1987. The figure shows the number of homes that sold for a price in the stated price range on the horizontal axis.

Some figures and tables you may create just for your own use – they may help you identify relationships. The ones that you put in your paper should be related to your research question and help to illustrate a point you would like to make. If you are attempting to build a hedonic model of home prices, the variables that you would like to include in your model should determine what tables and graphs you use. You may want to know whether there is an unconditional difference in sales price for each of the binary variables (fireplace, swimming pool, and spa).

	Fireplace		Swimming pool		Spa	
	with	w/o	with	w/o	with	w/o
Mean Sales Price	188,648	174,122	242,175	177,449	239,079	185,323
Std. Dev.	78,432	106,202	99,094	77,158	75,633	83,362
N	826	154	135	845	19	961
p-value	0.0475		<0.0001		0.0054	

Table 4 Mean unconditional sales price for homes sold with and without each of the specified variables (fireplace, swimming pool, and spa) in Santa Clara County in June 1987. Standard deviation and number of observations also included. The p-value is for a t-test of the equality of means of sales price for each variable (i.e. comparing the mean sales prices of homes with and without a fireplace).

Table 4 presents basic t-tests for the equality of mean sales price by each binary variable. The caption provides all of the information needed to understand the table. The unconditional mean sales price differs for each binary variable, which suggests that they should be included in a hedonic model of home sales prices.