Take Test: Fall 2023 Midterm Exam

Test Information

Question Completion Status:

There are 506 observations and 25 variables in the correctHousing dataset.

2 points Saved

QUESTION 2

Choose between and enter the variable types; discrete, continuous, nominal, or ordinal for the following variables in the correctHousing dataset. Be sure to enter only one of these types for each variable. What types of variables are; "CRIM," continuous "CHAS," discrete "RM," continuous "AGE," continuous and "CMEDV" continuous?

5 points Saved

QUESTION 3

Calculate descriptive statistics for the variables in the housing dataset. What is the mean of the "CMEDV" variable?

26.27

1 points Saved

QUESTION 4

Calculate descriptive statistics for the variables in the housing dataset. What is the value of the CMEDV variable in dollars and cents (USD)?

26270

2 points Saved

QUESTION 5

Generate a frequency plot (histogram) for the variable “CMEDV,” i.e. the median value of owner-occupied homes, and answer the following questions.

The distribution of (corrected) home values appears to be normal.

- ○ True
- ◉ False

3 points Saved

QUESTION 6

Generate a frequency plot (histogram) for the variable “CMEDV,” i.e. the median value of owner-occupied homes, and answer the following questions.

The distribution representing CMEDV is obviously skewed.

- ◉ True
- ○ False

3 points Saved

QUESTION 7

Generate a frequency plot (histogram) for the variable “CMEDV,” i.e. the median value of owner-occupied homes, and answer the following questions.

There appear to be "outliers" in the (corrected) median home values data.

- ○ True
- ◉ False

4 points Saved

QUESTION 8

The variable INDUS in the housing dataset appear to be [A] skewed? Enter either right or left.

right

5 points Saved

QUESTION 9

What is the value of the Spearman correlation coefficient between CRIM and LSTAT?

0.70

3 points Saved

QUESTION 10

The relationship between the variables crim and lstat is [A]. Enter positive or negative.

positive

2 points Saved

QUESTION 11

Apologies this is a long question.

One of the problems you will encounter many times over the course of your careers is the issue of not having enough information about a dataset to make sense of it. This is actually one of those cases. I don't think that many people have paid too much attention to all the variables in the dataset or else we would have many more papers making corrections such as the one I have provided for you by Gilley and Kelley Pace (Gilley & Kelley Pace, 1996). I have also provided the original paper written on this dataset for you, "Hedonic Housing Prices and the Demand for Clean Air" (Harrison, Jr. & Rubinfeld, 1978). To get a sense of what I'm talking about I've substituted the variables "crim," "black," and "lstat," from the original data set for the variables "CRIM," "B," and "LSTAT," in the dataset with corrections by Gilley and Kelley-Pace. Using this, create two scatter plots of crime as a function of the percentage of lower status population. You can just copy/paste the commands in gretl for this which are:

gnuplot CRIM LSTAT --output=display

gnuplot crim lstat --output=display

Considering these plots, you can see that the data from the original dataset with lower case variable names makes some sense, i.e. a growing crime rate with increasing percentage of lower socio-economic status citizens living in the area. The plot of presumably the same variables including corrected data and now with uppercase variable names, CRIM and LSTAT, doesn't really make sense.

The values have obviously changed but Gilley and Kelley-Pace do not mention making any such change for these variables in their journal article. So… we will use the values for these variables from the original dataset with the caveat that something still seems wrong. That is, think about what a proportion is. Think about what the range of a proportion is.

My remarks are wrong. These values are truly proportions.

- ○ True
- ○ False

5 points Save Answer

QUESTION 12

Apologies this is a long question. I've repeated the text here so you don't have to go back and forth between parts of questions for details.

One of the problems you will encounter many times over the course of your careers is the issue of not having enough information about a dataset to make sense of it. This is actually one of those cases. I don't think that many people have paid too much attention to all the variables in the dataset or else we would have many more papers making corrections such as the one I have provided for you by Gilley and Kelley Pace (Gilley & Kelley Pace, 1996). I have also provided the original paper written on this dataset for you, "Hedonic Housing Prices and the Demand for Clean Air" (Harrison, Jr. & Rubinfeld, 1978). To get a sense of what I'm talking about I've substituted the variables "crim," "black," and "lstat," from the original data set for the variables "CRIM," "B," and "LSTAT," in the dataset with corrections by Gilley and Kelley-Pace. Using this, create two scatter plots of crime as a function of the percentage of lower status population. You can just copy/paste the commands in gretl for this which are:

```
gnuplot  CRIM LSTAT --output=display

gnuplot  crim lstat --output=display
```

Considering these plots, you can see that the data from the original dataset with lower case variable names makes some sense, i.e. a growing crime rate with increasing percentage of lower socio-economic status citizens living in the area. The plot of presumably the same variables including corrected data and now with uppercase variable names, CRIM and LSTAT, doesn't really make sense.

The values have obviously changed but Gilley and Kelley-Pace do not mention making any such change for these variables in their journal article. So… we will use the values for these variables from the original dataset with the caveat that something still seems wrong. That is, think about what a proportion is. Think about what the range of a proportion is.

Enter **yes** if you have read and understand the content of this question. Enter **no** if you have read but do not understand what is going on. Right now, either answer is correct. [yes; no], I understand what is going on!

yes

5 points ✔ Saved

QUESTION 13

Apologies, this is a long question.

The variable TOWN includes the names of the towns in the data. There may be more than one census tract per town. The variable TOWN_aaa includes coding for each of the towns in the data so that, in general, all census tracks in a single town have a single code.

However, some of this coding also seems to be incorrect. If you display the data for the variables TOWN and TOWN_aaa you will see that rather than an integer the code for "North" is "Reading," another character string. Beyond census tract #356 there is a lot more of this type "miscoding".

The variable crim (note the lowercase letters) seems to represent the crime rate by census tract. Perhaps it is the per capita crime rate. Note that Harrison and Rubinfeld just say it is the crime rate by town. The data says otherwise. The variable lstat (again note the lowercase letters) represents the proportion of the population that is lower socio-economic status. You should take time to look at the data to make sure you understand it. The following questions help you consider how the different variables change by town and demographic.

- This bullet point is just something for you to think about when you setup your analysis. It might be easier to answer the following questions for you to rename the observations that were miscoded, i.e. with character strings rather than integer coding. For example, you would need to rename "Reading" to be "15". Strictly speaking, this is not necessarily because all of the TOWN_aaa codes are unique. However, it might be easier for all of them to be integer codes. Unfortunately, the numbering doesn't work for Boston so if you do this you might just want to use a dummy integer such as 99. Even this is not truly necessary if you setup your analysis using samples from the census track numbers, i.e. 1 through and including 506.
- For this part of the question consider crime as a function of the number of people in lower socio-economic status. For your final answer, you want to determine if there are statistically significant differences in crime between the towns in the dataset depending on the number of citizens in lower socio-economic status. Let's consider the simplest cases first, i.e. is there a significant difference in crime across the towns separate from the existence of lower socio-economic citizens? And second, is there a significant difference in the number of citizens in the lower socio-economic status across towns? To do this we'll consider just the towns Newton encoded as TOWN_aaa = 40 or census tracks 221 through and including 238, and Boston encoded as TOWN_aaa =
    - Allston-Brighton
    - Back
    - Beacon
    - North
    - Charlestown
    - East
    - South
    - Downtown
    - Roxbury
    - Savin
    - Dorchester
    - Mattapan
    - Forest
    - West
    - Hyde

  and census tracks 347 through and including 488.

Again, since there are more census tracks here than could be condensed to one integer that would fit in the TOWN_aaa coding list, you can setup a dummy integer for Boston you can remember, e.g. 999. Or, you can do the analysis using a range of census tracts.

The reason you can condense your analysis across all towns is because Newton is considered a very affluent community whereas big areas of Boston such as Allston and Roxbury are not affluent at all. In fact, nearly 22% or almost 1 out of 4 people in Boston live in poverty (Boston Redevelopment Authority, 2014). That is, you can decide if there is a difference across the towns in the dataset by looking at the two extremes represented by the data.

For the crim variable, determine if there is a statistically significant variation between Newton and Boston using a 0.05 significance level.

Such a statistically significant variation exists and we must reject the null hypothesis.

- (●) True
- ( ) False

10 points ✓ Saved

QUESTION 14

Apologies, this is the second part of a long question. I repeated the entire background so you would not have to go back and forth between questions for details.

The variable TOWN includes the names of the towns in the data. There may be more than one census tract per town. The variable TOWN_aaa includes coding for each of the towns in the data so that, in general, all census tracks in a single town have a single code. However, some of this coding also seems to be incorrect. If you display the data for the variables TOWN and TOWN_aaa you will see that

rather than an integer the code for "North" is "Reading," another character string. Beyond census tract #356 there is a lot more of this type "miscoding".

The variable crim (note the lowercase letters) seems to represent the crime rate by census tract. Perhaps it is the per capita crime rate. Note that Harrison and Rubinfeld just say it is the crime rate by town. The data says otherwise. The variable lstat (again note the lowercase letters) represents the proportion of the population that is lower socio-economic status. You should take time to look at the data to make sure you understand it. The following questions help you consider how the different variables change by town and demographic.

- This bullet point is just something for you to think about when you setup your analysis. It might be easier to answer the following questions for you to rename the observations that were miscoded, i.e. with character strings rather than integer coding. For example, you would need to rename "Reading" to be "15". Strictly speaking, this is not necessarily because all of the TOWN_aaa codes are unique. However, it might be easier for all of them to be integer codes. Unfortunately, the numbering doesn't work for Boston so if you do this you might just want to use a dummy integer such as 99. Even this is not truly necessary if you setup your analysis using samples from the census track numbers, i.e. 1 through and including 506.
- For this part of the question consider crime as a function of the number of people in lower socio-economic status. For your final answer, you want to determine if there are statistically significant differences in crime between the towns in the dataset depending on the number of citizens in lower socio-economic status. Let's consider the simplest cases first, i.e. is there a significant difference in crime across the towns separate from the existence of lower socio-economic citizens? And second, is there a significant difference in the number of citizens in the lower socio-economic status across towns? To do this we'll consider just the towns Newton encoded as TOWN_aaa = 40 or census tracks 221 through and including 238, and Boston encoded as TOWN_aaa =
    - Allston-Brighton
    - Back
    - Beacon
    - North
    - Charlestown
    - East
    - South
    - Downtown
    - Roxbury
    - Savin
    - Dorchester
    - Mattapan
    - Forest
    - West
    - Hyde

  and census tracks 347 through and including 488.

Again, since there are more census tracks here than could be condensed to one integer that would fit in the TOWN_aaa coding list, you can setup a dummy integer for Boston you can remember, e.g. 999. Or, you can do the analysis using a range of census tracts.

The reason you can condense your analysis across all towns is because Newton is considered a very affluent community whereas big areas of Boston such as Allston and Roxbury are not affluent at all. In fact, nearly 22% or almost 1 out of 4 people in Boston live in poverty (Boston Redevelopment Authority, 2014). That is, you can decide if there is a difference across the towns in the dataset by looking at the two extremes represented by the data.

For the lstat variable, determine if there is a statistically significant variation between Newton and Boston using a 0.05 significance level.

Such a statistically significant variation exists and we must reject the null hypothesis.

- (●) True
- ( ) False

10 points ✓ Saved

QUESTION 15

Next to last, consider how the variables, crim and lstat, are related. What type of analysis have we conducted that can be used to evaluate the relationship between two variables?

- ( ) Test of Independence

- ○ ANOVA
- ◉ Correlation
- ○ Goodness of Fit

5 points ✓ Saved

QUESTION 16

What is the value of the correlation coefficient for crim and lstat?

0.46

5 points ✓ Saved

QUESTION 17

There is a [A] relationship between crim and lstat. Enter positive or negative.

positive

5 points ✓ Saved

QUESTION 18

There is a (very) strong relationship between crim and lstat.

- ○ True
- ◉ False

5 points ✓ Saved

QUESTION 19

Last, consider how different variables effect the median value of homes. Harrison and Rubinfeld's original purpose in collecting these data were to show that people were moving to the Boston area and making home-buying decisions based on seeking cleaner air. What they found

was that crime and other factors equally or more than the desire for cleaner air affected home-buyers' decisions about which homes to actually purchase.

So first, consider the median value of homes as a function of NOX (nitric oxides). Remember that the median value of homes is one of the variables that Gilley and Kelley-Pace corrected so use the CMEDV variable in the dataset for these values.

Compute the correlation coefficient and, for your own enlightenment, generate a plot of CMEDV as a function of air pollution in the form of NOX. Enter the value of the correlation coefficient.

0.03

5 points Saved

QUESTION 20

Last, consider how different variables effect the median value of homes. Harrison and Rubinfeld's original purpose in collecting these data were to show that people were moving to the Boston area and making home-buying decisions based on seeking cleaner air. What they found was that crime and other factors equally or more than the desire for cleaner air affected home-buyers' decisions about which homes to actually purchase.

Next, consider the median value of homes as a function of crim (crime). Remember that the median value of homes is one of the variables that Gilley and Kelley-Pace corrected so use the CMEDV variable in the dataset for these values.

How does the impact of crime affect home-buyers' decision? Compute the correlation coefficient between CMEDV and crim and enter that value.

0.06

5 points Saved

QUESTION 21

Last, consider how different variables effect the median value of homes. Harrison and Rubinfeld's original purpose in collecting these data were to show that people were moving to the Boston area and making home-buying decisions based on seeking cleaner air. What they found was that crime and other factors equally or more than the desire for cleaner air affected home-buyers' decisions about which homes to actually purchase.

Now, consider the median value of homes as a function of lstat (the presence of lower socio-economic status citizens/neighbors). Remember that the median value of homes is one of the variables that Gilley and Kelley-Pace corrected so use the CMEDV variable in the dataset for these values.

How does the presence of lower socio-economic status citizens affect home-buyers' decisions? Compute the correlation coefficient between CMEDV and lstat and enter that value.

-0.22

5 points Saved

QUESTION 22

Last, consider how different variables effect the median value of homes and hence (presumably) home-buyers' decisions.. Harrison and Rubinfeld's original purpose in collecting these data were to show that people were moving to the Boston area and making home-buying decisions based on seeking cleaner air. What they found was that crime and other factors equally or more than the desire for cleaner air affected home-buyers' decisions about which homes to actually purchase.

Remember that the median value of homes is one of the variables that Gilley and Kelley-Pace corrected so use the CMEDV variable in the dataset for these values.

Which of the following factors seems to most affect home values and therefore (presumably) home-buyers' decision to purchase a home?

- ○ Crime
- ◉ The presence of lower socio-economic status citizens/neighbors
- ○ Air pollution (NOX)
- ○ None of these has any effect

5 points ✓ Saved

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Save All Answers | Save and Submit