Homework 1: Using OLS Regression to Predict Median House Values in Philadelphia

CPLN 671/MUSA 500

This assignment asks you to examine the relationship between median house values and several neighborhood characteristics, using Philadelphia data at the Census block group level. You will need R for this assignment. Remember that this report needs to be written with an introduction, methods/results, and discussion. Pointers for a successful report, as well as an example on how to interpret regression output, are included. Also included is an outline which you're asked to follow when writing your report. Your submission may be written in MS Word (submitted as a .docx or .pdf file) or in R Markdown format.

Data Description

The 2000 Philadelphia Census block group level dataset ***RegressionData.csv*** contains, among other variables, the variables below:

1) **POLY_ID:** Census Block Group ID
2) **MEDHVAL:** Median value of all owner occupied housing units
3) **PCBACHMORE:** Proportion of residents in Block Group with at least a bachelor's degree
4) **PCTVACANT:** Proportion of housing units that are vacant
5) **PCTSINGLES:** Percent of housing units that are detached single family houses
6) **NBELPOV100:** Number of households with incomes below 100% poverty level (i.e., number of households living in poverty)
7) **MEDHHINC:** Median household income

Note that the original Philadelphia block group dataset has 1816 observations. We clean the data by removing the following block groups:

1) Block groups where population < 40
2) Block groups where there are no housing units
3) Block groups where the median house value is lower than $10,000
4) One North Philadelphia block group which had a very high median house value (over $800,000) and a very low median household income (less than $8,000)

The final dataset which you are given contains 1720 block groups.

The shapefile ***Regression Data*** will also be included for your convenience.

INSTRUCTIONS

SUGGESTION: READ THE ENTIRE SET OF INSTRUCTIONS BEFORE STARTING TO WORK ON THE ASSIGNMENT

The first several steps will involve *exploratory data analysis*, which is needed to prepare the data for regression analysis.

1) Import the file ***RegressionData.csv*** into R using the read.csv command.

    a. Using the hist, mean, and sd commands in R, examine the distribution of the dependent variable, **MEDHVAL**, and predictors **PCBACHMORE**, **NBELPOV100**, **PCTVACANT**, and **PCTSINGLES**, and calculate the mean and standard deviation of each of these variables.

        i. Using the results you obtain, present the summary statistics (i.e., mean and standard deviation) of each of the variables in a table, such as the one below.

Variable	**Mean**	**SD**
Dependent Variable		
Median House Value		
Predictors		
# Households Living in Poverty		
% of Individuals with Bachelor's Degrees or Higher		
% of Vacant Houses		
% of Single House Units		

        ii. Also, observe from the histograms that none of the variables looks normal. This being the case, examine whether a logarithmic transformation of the variable helps achieve a normal distribution. In R, use the log command to create 5 new variables called **LNMEDHVAL, LNPCBACHMORE**, **LNNBELPOV100**, **LNPCTVACANT**, and **LNPCTSINGLES**, which are the natural logs of **MEDHVAL, PCBACHMORE**, **NBELPOV100**, **PCTVACANT**, and **PCTSINGLES**, respectively.

            a. Remember: If the variable has **any** zero values, use the **log(1+[VAR])** transformation instead of the **log([VAR])** transformation.

            b. In your report, you will be asked to present histograms of the original and transformed variables. If you're planning to submit a report in MS Word, you may simply print the screen (Ctrl + Prt Scn), and paste the screenshot into MS Paint. Then, you may cut

out the relevant part (i.e., the 5 histograms) and present them in your report.

c. Note that the dependent variable does look more or less normal after the transformation – hence, **LNMEDHVAL** will be used as the dependent variable in the regression analysis. You will also see that for the predictors, the logarithmic transformation only helps normalize the **NBELPOV100** variable (so we will use **LNNBELPOV100** in the subsequent analyses). The other variables have a large spike at zero (i.e., *zero-inflated distributions*) after the transformations, so we will use the original, untransformed **PCBACHMORE**, **PCTVACANT**, and **PCTSINGLES** variables in the regression,

b. Look at whether the relationship between the dependent variable and each of the predictors is linear. To do so, create four scatter plots – one for each predictor using the plot command (or any other command in R that yields a scatter plot).

   i. In your report, you will be expected to present all four scatter plots as a single figure.

c. Look at the Pearson correlations between all the predictors you will be including in your model, listed below. Use the cor command or another command in R that computes Pearson correlations.

   PCTVACANT PCTSINGLES PCTBACHMORE LNNBELPOV100

   Again, you may simply print the screen (Ctrl + Prt Scrn), and paste it into MS Paint. Then, you may cut out the relevant part (i.e., the correlation matrix) and paste it into your report.

   Note whether you observe severe multicollinearity, and whether it's appropriate to include all 4 variables as predictors.

   Keep in mind that when you look at multicollinearity, you shouldn't be including the dependent variable in the correlation matrix – that is, in a good predictive model, you want the correlation between each predictor and the dependent variable to be strong, and that's not an issue.

d. Now use the readOGR command in the rgdal library in R (or another command in another R library of your choice) to import the shapefile and create choropleth maps of the following variables:

LNMEDHVAL PCTVACANT PCTSINGLES PCTBACHMOR LNNBELPOV100

In your report, present the map of **LNMEDHVAL** as a single figure, and then combine the maps of the four predictors into a single figure (i.e., all 4 maps should be smaller and fit on 1 page as a single figure). Please use any color/classification scheme you see fit, as long as it is consistent for these five maps, and the remaining maps in the report.

2) You are now done with the *exploratory analysis* of the data for this assignment. Note that here, you received data that have been cleaned. Typically, when you work with data, you need to make sure that you examine the variables for outliers and incorrectly coded/entered values. But now, you're ready for regression analysis.

    a. Assuming there's no severe multicollinearity, use the lm command to run the regression where **LNMEDHVAL** is the dependent variable and **PCTVACANT, PCTSINGLES, PCTBACHMOR**, and **LNNBELPOV100** are predictors.

    b. In your report, be sure to present the summary of the fit as well as the ANOVA table containing the regression and error sum of squares (use the summary and anova commands). The only thing you should be looking at in the output from the anova command is the error sum of squares, and not any of the p-values.

    c. Use the fitted, residuals and rstandard commands to save the predicted values, residuals and standardized residuals, respectively.

    d. Create a scatter plot with *Standardized Residuals* on the y-axis and *Predicted Values* on the x-axis. You will be asked to present this scatter plot in your report, so take a screenshot of it if you plan to use MS Word.

3) Use the step and step$anova commands in the MASS library to run stepwise regression and determine the best model based on the Akaike Information Criterion. Take a screenshot of the step$anova output if you plan to use MS Word.

4) Perform k-fold cross-validation (in which k = 5) using the CVlm command in the DAAG library and calculate the root mean square error (RMSE). Then re-run the regression model only using **PCTVACANT** and **MEDHHINC** as predictors, and again perform k-fold cross-validation in which k = 5. You will be asked to present the RMSE of both this model and the original model in your report.

5) Finally, create a histogram and a choropleth map of standardized regression residuals that you saved using the rstandard command earlier. Use the same classification/color scheme as in your earlier maps.

REGRESSION EXAMPLE

As an example, I am providing what I am expecting in your write-up. Here, I'm describing the following regression model: **MEDHHINC** regressed on **PCTKITCHEN** and **PCTSINGLES.** Our regression equation is:

$$MEDHHINC = \beta_0 + \beta_1 PCTSINGLES + \beta_2 PCTKITCHENS + \varepsilon$$

Results (this is output from GeoDa, and uses a different dataset, but the idea is the same)

SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION

Dependent Variable	**MEDHHINC**	Number of Observations	1720
Mean dependent var	31541.8	Number of Variables	3
S.D. dependent var	16293.7	Degrees of Freedom	1717
R-squared	0.252898	F-statistic	145.134
Adjusted R-squared	0.251155	Prob(F-statistic)	0
Sum squared residual	3.41152e+011	Log likelihood	-18871.3
Sigma-square	1.98922e+008	Akaike info criterion	37752.6
S.E. of regression	14104	Schwarz criterion	37779.9
Sigma-square ML	1.98344e+008		
S.E of regression ML	14083.5		

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	-46520.32	5441.647	-8.54894	0.0000000
PCTSINGLES	48544.42	2578.256	18.8284	0.0000000
PCTKITCHEN	76876.91	5614.012	13.69375	0.0000000

Write-up (***ALSO LOOK AT THE 'REGRESSION COEFFICIENT INTERPRETATION' FILE)***:

We regressed median household income (MEDHHINC) on the % of households with kitchen facilities (PCTKITCHEN), and % of single family houses (PCTSINGLES). The regression output tells us that the percentage of houses that are single and percentage of houses with kitchen facilities are highly significant and are positively associated with median household income (p<0.0001 for both variables). A one unit (i.e., percentage point) increase in the % of single houses in the block group is associated with a β_1 *= $48,544.42 increase in median household income. Similarly, a one unit (i.e., 1%) increase in the % of houses with kitchen facilities is associated with a* β_2 *= $76,876.91 increase in median household income. The p-value of less than 0.0001 for PCTSINGLES tells us that* ***if*** *there is actually no relationship between PCTSINGLES and the dependent variable MEDHHINC (i.e., if the null hypothesis that* β_1*=0 is actually true), then the probability of getting a* β_1 *coefficient estimate of 48,544.42 is less than 0.0001. Similarly, the p-value of less than 0.0001 for PCTKITCHEN tells us that* ***if*** *there is actually no relationship between PCTKITCHEN and the dependent variable MEDHHINC (i.e., if the null hypothesis that* β_2*=0 is actually true), then the probability of getting a* β_2 *coefficient estimate of 76,876.91 is less than 0.0001. These low probabilities indicate that we can safely reject* H_0*:* β_1 *= 0 for* H_a*:* $\beta_1 \neq 0$ *and* H_0*:* β_2 *= 0 for* H_a*:* $\beta_2 \neq 0$ *(at most reasonable levels of* α *= P(Type I error)).*

A *little over a quarter of the variance in the dependent variable is explained by the model (*R^2 *and Adjusted* R^2 *are 0.253 and 0.251, respectively*). *The low p-value associated with the F-ratio shows that we can reject the null hypothesis that all coefficients in the model are 0.*

REPORT OUTLINE

A successful report will address the points presented in this outline. You are strongly encouraged to use the outline as a backbone for your report.

Note that this outline separates the Methods and Results sections. If you prefer to combine them, that is certainly fine as well – however, be sure to address every point in this outline.

The outline here is structured as an outline for a journal article. That is, in the Methods section, only talk about the techniques that you use, present the formulas, etc. Do not present any results in the methods section. In the Results section, actually present the output from R, any figures/screenshots, etc, and describe your output.

1) **Introduction (~2 paragraphs)** *Section Title*
   - **a)** State the problem and the setting of the analysis (i.e., Philadelphia).
   - **b)** Present either a brief review of the literature (use Google Scholar) or simply speculate as to why the predictors we're using might be related with the response variable.

2) **Methods (~2-5 pages)** *Section Title*
   - **a) Data Cleaning** *Subsection Title*
     i. Simply state that the original dataset had 1816 observations (i.e., block groups) and was cleaned in order to achieve a dataset with 1720 observations (that is, basically include the information about data cleaning that is on the very the first page of this assignment).
   - **b) Exploratory Data Analysis** *Subsection Title*
     i. State that you will examine the summary statistics and distributions of variables.
     ii. Also state that as part of your *exploratory data analysis*, you will examine the correlations between the predictors.
       1. Explain what a correlation is, and provide the formula for the sample correlation coefficient *r*. Also mention the possible range of *r* values, and what correlation of 0 means.
   - **c) Multiple Regression Analysis** *Subsection Title*
     i. Describe the method of regression in several sentences. I.e., what is it used for, what does it do?
     ii. State the equation for y for this problem. The equation should be in the form:
     $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon.$
     However, in your report, instead of y and $x_1 \ldots x_k$, fill in the actual variable names (as in the regression example given above). Be sure to mention what βi's and ε are as well. If the variables are log transformed, be sure to indicate that in the formulas.

iii. State and explain regression assumptions (e.g., linearity; independence of observations; normality of residuals; homoscedasticity; no multicollinearity).
iv. Mention the parameters that need to be estimated in multiple regression (σ^2, β_0 ,…, β_k). State what σ^2 is (you should have already talked about β_i in (ii) above).
v. Talk about the way of estimating the parameters. (Hint: present the equation on the slide 'β Coefficient Estimation – Least Squares' for multiple regression and briefly discuss what the equation does).
vi. Talk about the coefficient of multiple determination R^2, and the adjusted R^2. Present *and* explain the relevant formulas and all the terms that are used in the formulas.
vii. State the hypotheses you test. Specifically, talk about the F-ratio and the H_0 and H_a associated with it, as well as the hypotheses you test about each of the individual βi's (again, state H_0 and H_a).

d) Additional Analyses *Subsection Title*

i. Talk about stepwise regression – discuss what it does and its limitations
ii. Talk about k-fold cross-validation (mentioning that k = 5) – discuss what it is used for, describe how it is operationalized and mention that the RMSE is used to compare models (explain what the RMSE is and how it is calculated, presenting and describing any relevant formulas).

e) Software

i. State that you're using R for your data analysis.

3) <u>Results (~2-3 pages, excluding maps)</u> *<u>Section Title</u>*

a) Exploratory Results *Subsection Title*

i. Present and briefly talk about the table with *summary statistics* which includes the dependent variable and the predictors (i.e., mean, standard deviation).
ii. Also state whether the variables are normal before and after the logarithmic transformation
    1. Present the *histograms* of the original variables alongside the histograms of the log-transformed variables, and clearly state whether you're using the log-transformed or original variable in your regression.
    2. State that the other regression assumptions will be examined in a separate section below (Regression Assumption Checks).
iii. Present the *choropleth maps* of the dependent variable and the predictors.
    1. Refer to the maps in the text, and talk about the following:

a. Which maps look similar? Which maps look different? That is, which predictors do you expect to be strongly associated with the dependent variable based on the visualization? Also, given your examination of the maps, are there any predictors that you think will be strongly inter-correlated? That is, do you expect severe multicollinearity to be an issue here? Discuss this in a paragraph.

iv. Present the *correlation matrix* of the predictors which you obtained from R.
    1. Talk about whether the correlation matrix shows that there is severe multicollinearity.
    2. Does the correlation matrix support your conclusions based on your visual comparison of predictor maps?

b) Regression Results *Subsection Title*

i. Present the *regression output* from R. Be sure that your output presents the parameter estimates (and associated standard errors, t-statistics and p-values), as well as the R^2, the adjusted R^2, and the relevant F-ratio and associated p-value.

ii. Referencing the regression output in (i) above, interpret the results as in the example included above this report outline.

NOTE: YOUR DEPENDENT VARIABLE (AND SOME PREDICTORS) WOULD BE LOG-TRANSFORMED, UNLIKE IN THE EXAMPLE HERE. LOOK AT THE SLIDES FOR EXAMPLES OF INTERPRETING REGRESSION OUTPUT WITH LOG-TRANSFORMED VARIABLES.

c) Regression Assumption Checks *Subsection Title*

i. First state that in this section, you will be talking about testing model assumptions. State that you have already looked at the variable distributions earlier.

ii. Present *scatter plots of the dependent variable and each of the predictors*. State whether each of the relationships seems to be linear, as assumed by the regression model. *[Hint: they will not look linear.]*

iii. Present the histogram of the standardized residuals. State whether the residuals look normal.

iv. Present the *'Standardized Residual by Predicted Value' scatter plot*. What conclusions can you draw from that? Does there seem to be heteroscedasticity? Do there seem to be outliers? Anything else? Discuss.
    1. Mention what standardized residuals are.

v. Referencing the maps of the dependent variable and the predictors that you presented earlier, state whether there seems to be *spatial autocorrelation* in your variables. That is, does it seem that the

observations (i.e., block groups) are independent of each other? Briefly discuss.

vi. Now, present the *choropleth map of the standardized regression residuals*. Do there seem to be any noticeable spatial patterns in them? That is, do they seem to be spatially autocorrelated?
    1. You will examine the spatial autocorrelation of the variables and residuals and run spatial regressions in the next assignment.

d) Additional Models *Subsection Title*

i. Present the results of the stepwise regression and state whether all 4 predictors in the original model are kept in the final model.

ii. Present the cross-validation results – that is, compare the RMSE of the original model that includes all 4 predictors with the RMSE of the model that only includes **PCTVACANT** and **MEDHHINC** as predictors.

4) <u>Discussion and Limitations (~1 page)</u> *Section Title*

a) Recap what you did in the paper and your findings. Discuss what conclusions you can draw, which variables were significant and whether that was surprising or not.

b) Talk about the quality of the model – that is, state if this is a good model overall (e.g., R^2, F-ratio test), and what other predictors that we didn't include in our model might be associated with our dependent variable.

i. Looking at the stepwise regression results, did the final model include all 4 predictors or were some dropped? What does that tell you about the quality of the model?

ii. Looking at the cross-validation results, was the RMSE better for the 4 predictor model or the 2 predictor model?

c) If you haven't done that in the Results section, talk *explicitly* about the limitations of the model – that is, mention which assumptions were violated, and if applicable, how that may affect the model/parameter estimation/estimated significance.

i. In addition, talk about the limitations of using the **NBELPOV100** variable as a predictor – that is, what are some limitations of using the raw number of households living in poverty rather than a percentage?

d) Would it make sense to run Ridge or LASSO regression here? Explain briefly (~4-5 sentences) what these methods are, when they're used, and why they would or would not be appropriate here.