Does neighborhood with crime affect housing prices?

1. **Introduction**

Housing demand is an essential and there are many people who do not yet owned a home because the housing prices exceed their financial capacity. What are the factors that driving the house prices so high, this model is to analyze which factor affects the house prices directly the from a random selected state in America. Although the model may not be perfect, but it contributes to explaining some of the fluctuation in house prices. This model will help the real estate investors or consumers to predict the prices of a house in certain neighborhood. In this specific model, I want to evaluate how the crime rates has any effect to house prices.

My hypothesis is that if the poverty rate is high in a specific neighborhood, house prices tend to fall, and a neighborhood with low crime rates, house prices tend to be higher.

1. **Theory and Hypotheses**

The principle variables used and expected relationships in the model are:

Price: Median value of owner-occupied homes in the census tract

Crime: Per capita crime rate in the community

Nox: Average concentration (parts per 100 million) of nitrogen oxides in the air (a measure of air pollution)

Rooms: Average number of rooms per owner-occupied home

Dist. : the weighted distances to five major employments centers in the metropolitan area. Larger values indicate the tract is farther from the major employment centers

Radial:an index of accessibility to radial highways. Calculated on a community basis. Larger values represent better highway access.

Prop tax:Property tax rate in dollars per $1,000of property value. This measures costs paid by homeowners to maintain schools and public services in each community

Stratio: Pupil-teacher ratio in each school district. Lower values may indicate higher quality public schools.

Lowstat : percent of people  “lower  status”

Before performing regression analyses for the aforementioned variables, the Linear Model (LM) assumptions were checked for bias, variance, and a normal distribution. These assumptions are as follows:

1. Assumption 1: The model must be linear in parameters  
   The model in this study passes the assumption linear in parameter on the equation:

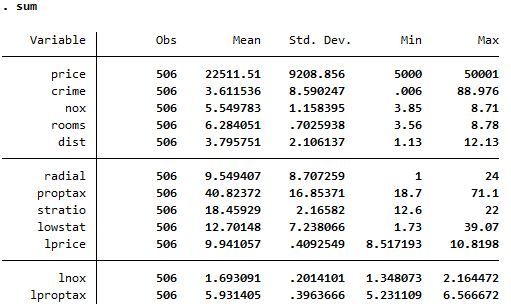
y = B0 + B1X1+ B2X2 + ... + BkXk + u

1. Assumption 2: Random Sampling  
   All data is drawn from random populations and samples around the globe and therefore the model satisfies this assumption
2. Assumption 3: No perfect collinearity  
   To test for perfect collinearity the model was placed in STATA software for the data collected. Correlation statistics performed in the STATA
3. Assumption 4: Expected value of error term, u, is zero.
4. Assumption 5: Homoscedasticity  
   This assumption requires the expected variance of the error term, u, to be constant given any dependent variable. Given that there are variables within the error term not included in this analysis, this assumption also cannot be met with certainty. Due to these uncertainties, the model below will be interpreted accordingly and with caution
5. Assumption 6: Standard Normal Distribution  
   The standard normal distribution is also assumed for this model in order to compute simple and multiple regression analysis.

The significance levels I will use in my tests: alpha = 0.05

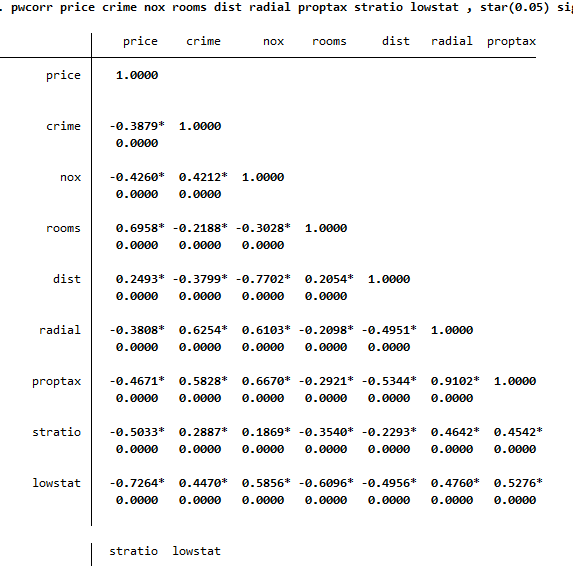
1. **Empirical Model and Data**

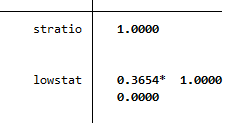
Exploring Data:



Since variable price, nox, proptax has value min – max large, so we will log transform this variable.

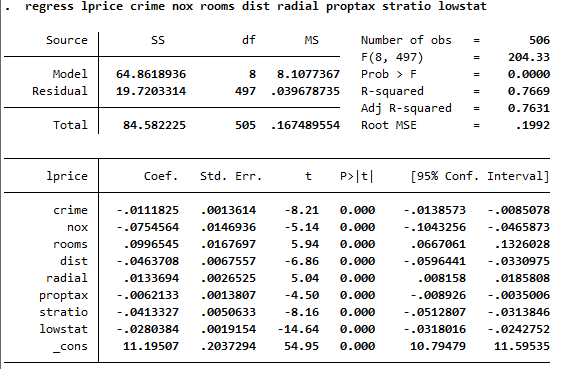
We create a correlation matrix for all variables in the model. Numbers are Pearson correlation coefficients, go from -1 to 1. Closer to 1 means strong correlation. A negative value indicates an inverse relationship (roughly when one goes up the other goes down).



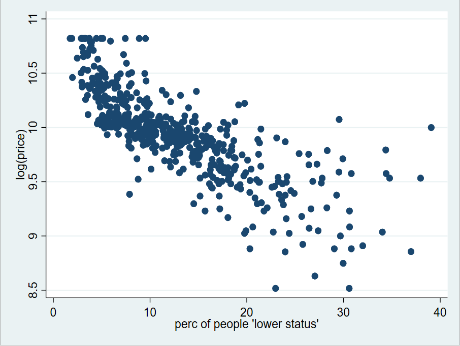
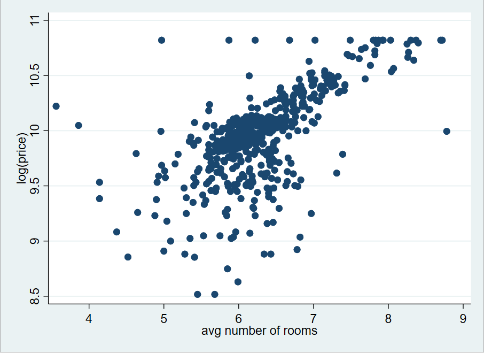
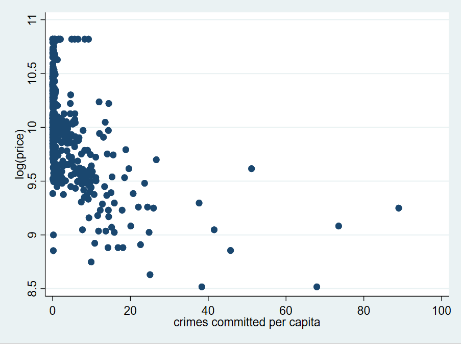


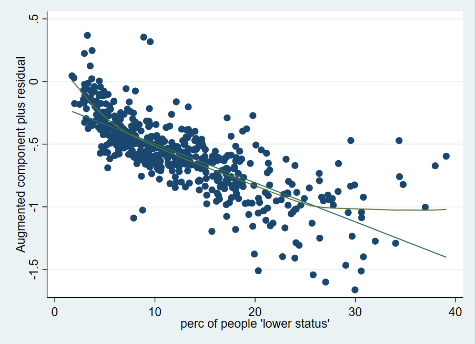
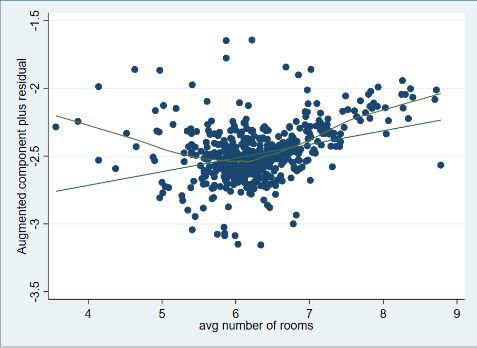
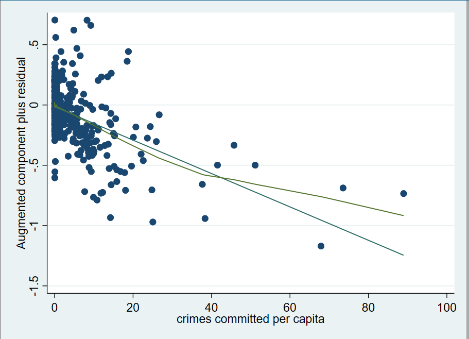
**MODEL 1: Full model (with no squared terms or interaction)**

**lprice= 11.195– 0.111(crime) – 0.075(nox) – 0.099(room) -0.046(dist)+0.13(radial)-0.006(proptax) -0.041(stratio)-0.028 (lowstat)**



Exploring the relationship graphically:

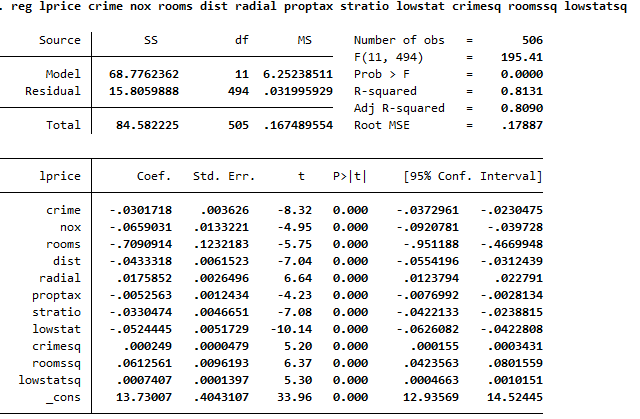




There is seemed to be a curvilinear relationship between lprice and crime, lprice and rooms, and lprice and lowstat. To deal with U-shaped curves we need to add a square version of the variable.

We use another graphical way to examine the relationship between variables, it does provide a good testing for linearity.

Create linear regression model has squared terms (crimesq, roomssq, lowstatsq)



We remove variables that are not important with the hypothesis. Finally, we obtain final model

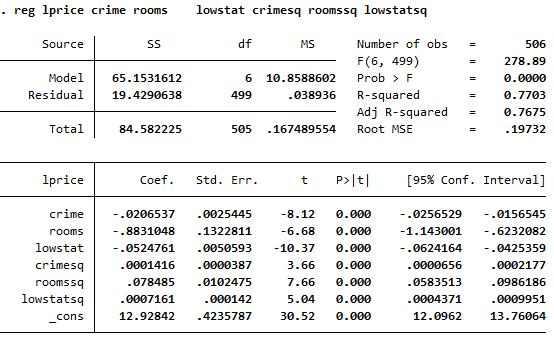
Model only has linear and squared terms:

**Model 2: no interaction term included**

**lprice= 12.928-0.02 (crime)-0.883(rooms)+ 0.052(lowstat)+0.001(crimesq)**

**+ 0.078(roomssq) +0.0007(lowstatsq)**

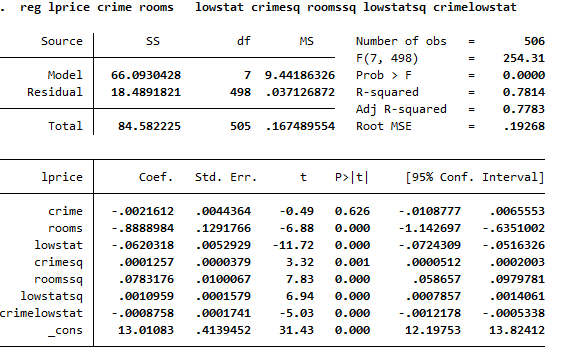
**(n= 506, Adj R squared = 76.8%)**



**Model 3 : interaction term included (crimelowstat)**

**lprice= 13.010 - 0.002(crime) – 0.888(rooms) – 0.062(lowstat) +0.0001(crimesq)+0.078(roomssq) +0.001(lowstatsq) – 0.0008(crimelowstat)**

**(n= 506, Adj R square = 77.8%)**



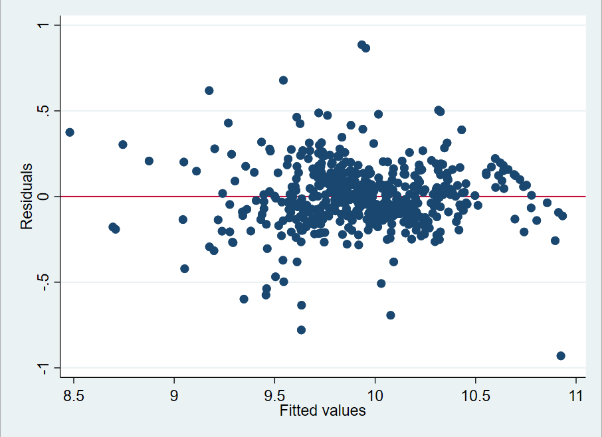
In model 1: All variables have statistical significance (Use t-test, p- value < alpha=0.05)

In model 2:  All variables have statistical significance (Use t-test, p- value < alpha=0.05), but “crime” variable has p- value = 0.626> alpha, so variable crime has no statistical significance.

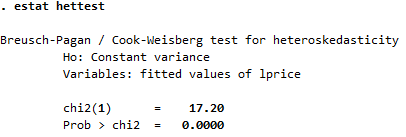
1. **Violation and Assumptions**

An important assumption is that the variance in the residuals has to be homoscedastic or constant.

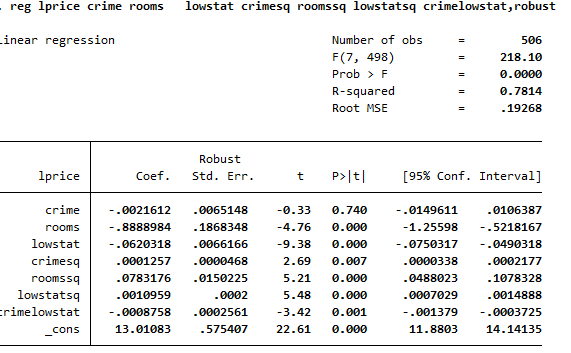
There are two ways to check for heteroscedastic, graphically and Breush-Pagan.

Graphically ****

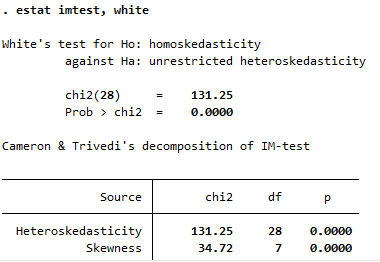
A non-graphical way to detect heteroskedasticiy is the Breusch-Pagan test. The null hypothesis is that residuals are homoscedastic:

****

The graphical and the Breush-Pagan test suggest the possible presence of heteroskedasticity in our model. The problem with this is that we may have the wrong estimates of the standard errors for the coefficients and therefore their t-values. There are two ways to deal with this problem, one is using heteroskedasticity-robust standard errors

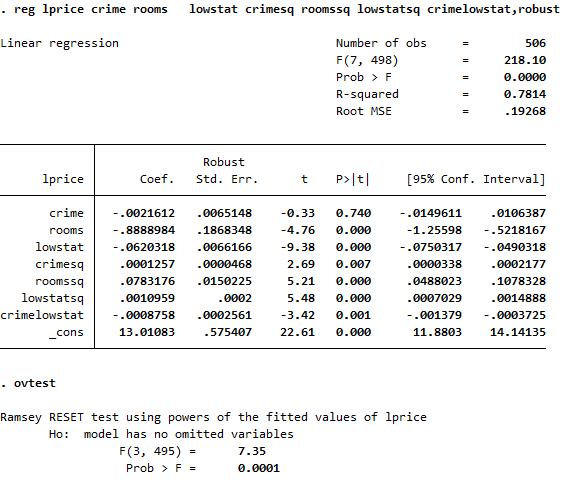
****

After trying to fix the heteroscedasticity errors of the model with robus option, the results of with test still indicates that this model still has heteroscedasticity.

**** Because p-value=0.0000<0.05, reject Ho, this model is heteroskedasticity.

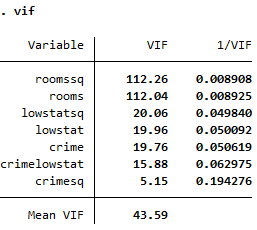
**How do we know we have included all variables we need to explain Y?**

Testing for omitted variable bias is important for our model since it is related to the assumption that the error term and the independent variables in the model are not correlated (E(e|X) = 0)

****

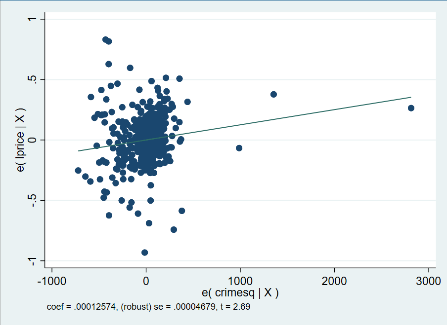
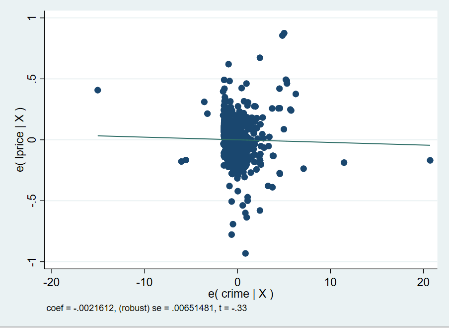
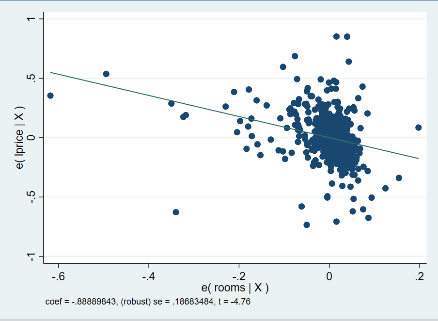
The null hypothesis is that the model does not have omitted-variables bias, the p-value is less than the usual threshold of 0.05 (95% significance), so we reject the null and conclude that we do need more variables.

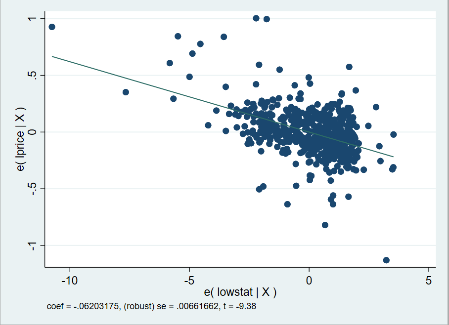
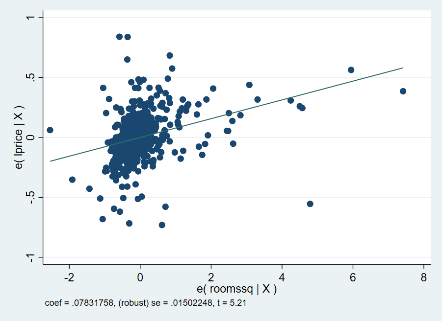
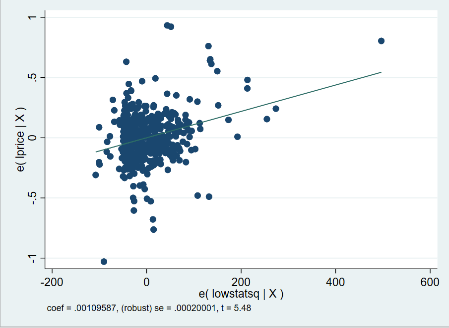
An important assumption for the multiple regression model is that independent variables are *not perfectly multicollinear*. One regressor should not be a linear function of another.

****

A vif > 10 or a 1/vif < 0.10 indicates trouble. We know that crime-crimesq, rooms-roomssq, and lowstat-lowstatsq are related since one is the square of the other. They are ok since crime, rooms, and lowstat has a quadratic relationship with Y, but this would be an example of multicolinearity.

Check outlier (Outliers are data points with extreme values that could have a negative effect on our estimators)

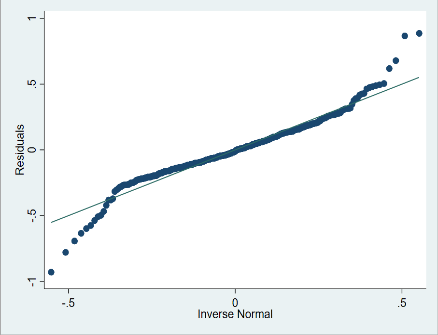
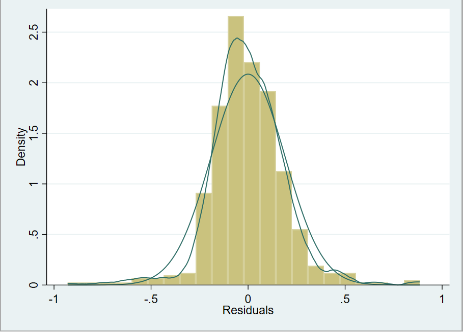
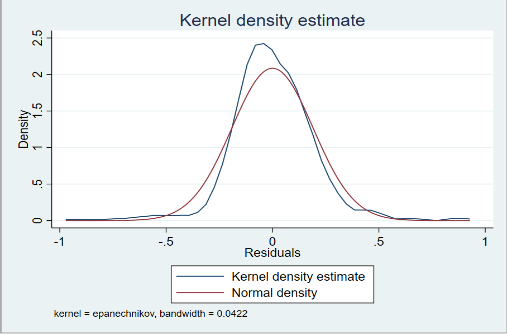
****

****

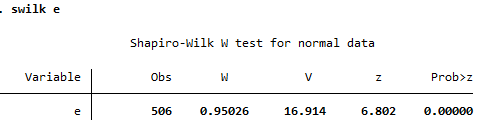
These plots regress each variable against all others, notice the coefficients on each. All data points seem to be in range, no outliers observed.

Another assumption of the regression model (OLS) that impact the validity of all tests (p, t and F) is that residuals behave ‘normal’. Residuals (here indicated by the letter “e”) are the difference between the observed values (Y) and the predicted values (Yhat): e = Y – Yhat.

Three graphs will help us check for normality in the residuals: kdensity, pnorm and qnorm.



A non-graphical test is the Shapiro-Wilk test for normality. It tests the hypothesis that the distribution is normal, in this case the null hypothesis is that the distribution of the residuals is normal.



From result of Shapiro Wilk W test, we say that residuals is not normal distribution.

1. **Results**

**MODEL 1:**

**lprice= 12.928 - 0.020(crime) – 0.883(rooms) – 0.524(lowstat) + 0.0001(crimesq) +0.078(roomssq) + 0.0007 (lowstatsq)**

**(n= 506, Adj R squared = 76.8%)**

In model 1, there is not yet an interaction between “crime” and “lowstat”. In model 1, the coefficients of the variable’s crime, rooms, lowstat has negative relationship for the dependent var “lprice”, and variables crime sq, room ssq, lowstatsq has a positive relationship with “lprice”.

In model 2, we are interested in the interaction between “crime” and “lowstat”, and now we analyze the meaning of the coefficients in the model 2.

**MODEL 2:**

**lprice= 13.010 - 0.002(crime) – 0.888(rooms) – 0.062(lowstat) + 0.0001(crimesq)**

**+ 0.078(roomssq) + 0.001(lowstatsq) – 0.0008(crimelowstat)**

**(n= 506, Adj R square = 77.8%)**

The variable crime, rooms, lowstat, “crime\*lowstat” indicates a negative relationship to the “lprice”, whereas variable crimesq, roomssq, lowstatsq has a positive relationship with “lprice”.

However, the final forcasting model violates a few assumptions of OLS (for example: the residuals is not normal which causes heteroscedasticity to occurs, and the model is lack variables). Therefore, this model needed to be adjusted for a better outcome, otherwise, it is not a good model to predict house prices.

1. **Conclusion**

This model is built based on OLS assumptions. It violates too many OLS assumptions, therefore, the data does not help to solve the original question. If there are additional time given, I will choose a better forcasting model to provide a better forecast for house prices than this model.