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

Yiming Cao, Sujan Kakumanu, Angel Sanaa Rutherford

2025-10-15

1 Introduction

2 Methods

2.1 Data Cleaning

2.2 Exploratory Data Analysis

2.3 Multiple Regression Analysis

Multiple regression models a dependent variable as a function of multiple predictors, rather than a single predictor such as in simple regression. These predictors each have a coefficient that represents their effect on a dependent variable, controlling for all other predictors. This approach improves model accuracy in situations where multiple variables better explain outcomes of a dependent variable.

This report regressed log-transformed median house value (LNMEDHVAL) on the proportion of housing units that are vacant (PCTVACANT), percent of housing units that are single family detached homes (PCTSINGLES), proportion of residents with at least a bachelor's degree (PCTBACHMOR), and log-transformed number of households with incomes below 100% poverty level (LNNBELPOV). This regression function can be expressed as follows:

$$\text{LNMEDHVAL} = \beta_0 + \beta_1 \text{PCTVACANT} + \beta_2 \text{PCTSINGLES} + \beta_3 \text{PCTBACHMOR} + \beta_4 \text{LNNBELPOV} + \varepsilon$$

Multiple regression relies on several key assumptions, most of which mirror the assumptions of simple regression. First, linear relationships should exist between the dependent variable and each predictor, assessed through scatterplots or residual plots and addressed via transformations if needed. Second, residuals should be approximately normally distributed, which can be assessed through a histogram. Third, residuals must be random — indicating that observations are not systematically related. Fourth, residuals must be homoscedastic, exhibiting constant variance across all values. Fifth, the dependent variable should be continuous.

A unique assumption for multiple regression is avoiding perfect multicollinearity: no predictor should be strongly correlated with others. Multicollinearity inflates standard errors and produces unstable coefficient estimates. This assumption can be checked by analyzing the correlation coefficients between all dependent variables, with anything greater than 0.9 generally being a cause for concern. Variance Inflation Factor (VIF) can be used to further inspect a suspicion of multicollinearity, with a VIF < 5 being generally acceptable and a VIF < 10 warranting more inspection. A VIF > 10 strongly indicates multicollineariy.

In the above multiple regression function, β_0 represents the depedent variable when all predictors are zero. The coefficients of the predictors $\beta_1, \beta_2, \beta_3, \beta_4$ each represent the change in the dependent variable with a one unit increase in the predictor, holding all other predictors constant.

These β coefficients in multiple regression are simultaneously estimated in order to minimize the Error Sum of Squares (SSE). The general formula and breakdown of what is to be minimized is provided below (with n being the number of observations, and k is the number of predictors):

$$SSE = \sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki} \right) \right]^2$$

This minimization works by finding the β coefficients that, when raw predictor (x_i) data is used, will minimize the residuals $(y_i - \hat{y}_i)$. SSE is also used to calculate Mean Squared Error (MSE), noted by the estimated parameter $\hat{\sigma}^2$. This is the estimate of the variance of the error term ϵ. The formula for MSE, in terms of SSE is noted below:

$$MSE = \frac{SSE}{n - (k+1)}$$

Another term in regression analysis is Total Sum of Squares (SST). It measures the total variation in the dependent variable around it's mean by using the following formula:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Using this formula for SST, and the previously stated formula for SSE, we can calculate R^2 — the coefficient of multiple determination. This is the proportion of variance in the model explained by all k predictors, and is the represented by the following:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

Multiple regression presents a unique dillema in comparison to simple regression, in that adding more predictors will generally increase R^2. Adjusting R^2, noted below, can account for additional predictors and determine whether or not they are improving the model.

$$R^2_{\text{adj}} = \frac{(n-1)R^2 - k}{n-(k+1)}$$

This report will conduct two tests to evaluate the model. First, there is the F-ratio — a model utility test. F-ratio tests the following null hypothesis H_0 and alternative hypothesis H_a:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_a : \text{At least one } \beta_i \neq 0$$

In essence, the null hypothesis states that all of the model β parameters are zero, and the alternative states that at least one of those parameters is not zero. Failure to reject the null hypothesis suggests that the model is incredibly weak, and should be reevaluated. If the null hypothesis is rejected, the second test can be conducted with the following hypotheses.

$$H_0 : \beta_i = 0$$

$$H_a : \beta_i \neq 0$$

In this test, we evaluate the performance of each predictor i (in the case of this report, the 4 predictors stated earlier). A t-test can be conducted, where the t-statistic for each predictor is calculated as the estimated coefficient divided by its standard error:

$$t_i = \frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}}$$

Each predictor has its own p-value calculated using the above t-statistic. If the p-value is less than 0.05, we reject the null hypothesis for that predictor and conclude that it is a statistically significant predictor of the dependent variable. If the p-value is greater than or equal to 0.05, we fail to reject the null hypothesis and conclude that the predictor is not statistically significant.

2.4 Additional Analysis

2.5 Software Used

3 Results

3.1 Exploratory Results

3.2 Regression Results

The output of the regression model (LNMEDHVAL $=$ $\beta_0 + \beta_1$PCTVACANT $+$ β_2PCTSINGLES $+ \beta_3$PCTBACHMOR $+ \beta_4$LNNBELPOV $+ \varepsilon$) in R is as follows.

```
Call:
lm(formula = LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR +
    LNNBELPOV, data = Regression_shpData)

Residuals:
     Min       1Q   Median       3Q      Max
-2.25817 -0.20391  0.03822  0.21743  2.24345

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept) 11.1137781  0.0465318 238.843 < 0.0000000000000002 ***
PCTVACANT   -0.0191563  0.0009779 -19.590 < 0.0000000000000002 ***
PCTSINGLES   0.0029770  0.0007032   4.234            0.0000242 ***
PCTBACHMOR   0.0209095  0.0005432  38.494 < 0.0000000000000002 ***
LNNBELPOV   -0.0789035  0.0084567  -9.330 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3665 on 1715 degrees of freedom
Multiple R-squared:  0.6623,    Adjusted R-squared:  0.6615
F-statistic: 840.9 on 4 and 1715 DF,  p-value: < 0.00000000000000022
```

We regressed the natural log of median house value (LNMEDHVAL) on the percentage of vacant houses (PCTVACANT), percentage of single-family houses (PCTSINGLES), percentage of residents with a bachelor's degree or higher (PCTBACHMOR), and the natural log of the neighborhood poverty rate (LNNBELPOV). All four predictors are statistically significant with p-values far below a threshold of $p < 0.05$.

The log-transformation of median house value (LNMEDHVAL, the dependent variable) means that we can interpret the coefficients as percent changes in median home value for a one unit change in the predictor. A one percentage point increase in vacant houses (PCTVACANT) is associated with an approximate 1.92% decrease in median home value. A one percentage point increase in single-family houses (PCTSINGLES) is associated with an approximate 0.30% increase in median home value. A one percentage point increase in residents with a bachelor's degree or higher (PCTBACHMOR) is associated with a roughly 2.09% increase in median home value. For LNNBELPOV — a log-transformed predictor — a 1% increase in the poverty rate corresponds to an approximate 7.9% decrease in median home value.

The very low p-values indicate that if there were actually no relationship between each predictor and median home value (i.e., $H_0 : \beta_i = 0$), the probability of observing the estimated coefficients we see would be very close to zero. Therefore, we can reject the null hypotheses for all predictors $H_0 : \beta_i = 0$.

The model explains a substantial portion of the variance in median home values, with $R^2 = 0.6623$ and $R^2_{adj} = 0.6615$. The F-statistic is highly significant, with $F = 840.9$ and a p-value of $p < 0.00000000000000022$, allowing us to reject the H_0 that all coefficients in the model are 0.

3.3 Regression Assumption Checks

3.4 Additional Models

4 Discussion & Limitations