**Assignment #1.1 Linear Regression**

|  |
| --- |
| **Submission Instructions**   * Submit the followingfile through Blackboard:  1. The completed **answer sheet** provided on the last page.  * If you do not follow the instructions, your assignment will be counted late. |

**Before you start**

For this assignment, you’ll be working with the **kc\_house\_data.csv** file and the **Lab1.py** script. This file reflects house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. Each row represents a selling record. The manager of a real estate company wants to better understand the determinants of housing price so that he can give more effect value evaluation in advance.

**Part 1. Simple linear regression**

In this part, you will build, and estimate with Python, a simple linear regression model to help the manager understand how the age of a house may influence its selling price.

* 1. Model Identification

1. What is the dependent variable and what is the explanatory variable in this business problem?

The dependent variable is price, and the explanatory variables are price, bedrooms, bathrooms, sqft\_lot, floors, waterfront, view, condition, sqft\_above, sqft\_basement, and AgeOfHouse.

1. What linear regression model can you use to answer this business problem?

We can use the y=b0+b1\*x regression model to answer this business problem where y is the dependent variable (house price), x is the independent variable (age of house), b0 is the intercept (the expected value of y when x is zero), and b1 is the regression coefficient (the change in y for a one-unit increase in x).

1. Estimate your above regression in Python and explain the results.

price bedrooms bathrooms sqft\_lot floors \

count 2.161300e+04 21613.000000 21613.000000 2.161300e+04 21613.000000

mean 5.401822e+05 3.370842 2.114757 1.510697e+04 1.494309

std 3.673622e+05 0.930062 0.770163 4.142051e+04 0.539989

min 7.500000e+04 0.000000 0.000000 5.200000e+02 1.000000

25% 3.219500e+05 3.000000 1.750000 5.040000e+03 1.000000

50% 4.500000e+05 3.000000 2.250000 7.618000e+03 1.500000

75% 6.450000e+05 4.000000 2.500000 1.068800e+04 2.000000

max 7.700000e+06 33.000000 8.000000 1.651359e+06 3.500000

waterfront view condition sqft\_above sqft\_basement \

count 21613.000000 21613.000000 21613.000000 21613.000000 21613.000000

mean 0.007542 0.234303 3.409430 1788.390691 291.509045

std 0.086517 0.766318 0.650743 828.090978 442.575043

min 0.000000 0.000000 1.000000 290.000000 0.000000

25% 0.000000 0.000000 3.000000 1190.000000 0.000000

50% 0.000000 0.000000 3.000000 1560.000000 0.000000

75% 0.000000 0.000000 4.000000 2210.000000 560.000000

max 1.000000 4.000000 5.000000 9410.000000 4820.000000

AgeOfHouse Unnamed: 11 Unnamed: 12 Unnamed: 13

count 21613.000000 0.0 0.0 1.0

mean 51.994864 NaN NaN 3025.0

std 29.373411 NaN NaN NaN

min 8.000000 NaN NaN 3025.0

25% 26.000000 NaN NaN 3025.0

50% 48.000000 NaN NaN 3025.0

75% 72.000000 NaN NaN 3025.0

max 123.000000 NaN NaN 3025.0

OLS Regression Results

==============================================================================

Dep. Variable: price R-squared: 0.003

Model: OLS Adj. R-squared: 0.003

Method: Least Squares F-statistic: 63.16

Date: Mon, 06 Mar 2023 Prob (F-statistic): 2.00e-15

Time: 19:03:29 Log-Likelihood: -3.0759e+05

No. Observations: 21613 AIC: 6.152e+05

Df Residuals: 21611 BIC: 6.152e+05

Df Model: 1

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

const 5.753e+05 5073.098 113.399 0.000 5.65e+05 5.85e+05

AgeOfHouse -675.1304 84.951 -7.947 0.000 -841.641 -508.620

==============================================================================

Omnibus: 19144.526 Durbin-Watson: 1.972

Prob(Omnibus): 0.000 Jarque-Bera (JB): 1142142.076

Skew: 4.030 Prob(JB): 0.00

Kurtosis: 37.689 Cond. No. 121.

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

OLS Regression Results

==============================================================================

Dep. Variable: price R-squared: 0.599

Model: OLS Adj. R-squared: 0.599

Method: Least Squares F-statistic: 3228.

Date: Mon, 06 Mar 2023 Prob (F-statistic): 0.00

Time: 19:04:07 Log-Likelihood: -2.9774e+05

No. Observations: 21613 AIC: 5.955e+05

Df Residuals: 21602 BIC: 5.956e+05

Df Model: 10

Covariance Type: nonrobust

=================================================================================

coef std err t P>|t| [0.025 0.975]

---------------------------------------------------------------------------------

const -2.812e+05 1.25e+04 -22.539 0.000 -3.06e+05 -2.57e+05

bedrooms -5.623e+04 2154.605 -26.096 0.000 -6.04e+04 -5.2e+04

bathrooms 6.372e+04 3709.414 17.177 0.000 5.64e+04 7.1e+04

sqft\_lot -0.3501 0.039 -8.903 0.000 -0.427 -0.273

floors 4.529e+04 3995.040 11.335 0.000 3.75e+04 5.31e+04

waterfront 5.401e+05 2e+04 26.994 0.000 5.01e+05 5.79e+05

view 6.212e+04 2404.542 25.836 0.000 5.74e+04 6.68e+04

condition 1.806e+04 2650.482 6.814 0.000 1.29e+04 2.33e+04

sqft\_above 285.1920 3.079 92.617 0.000 279.156 291.228

sqft\_basement 246.2398 4.746 51.888 0.000 236.938 255.542

AgeOfHouse 2918.3514 70.996 41.106 0.000 2779.194 3057.509

==============================================================================

Omnibus: 13032.917 Durbin-Watson: 1.981

Prob(Omnibus): 0.000 Jarque-Bera (JB): 467399.280

Skew: 2.329 Prob(JB): 0.00

Kurtosis: 25.301 Cond. No. 5.58e+05

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 5.58e+05. This might indicate that there are

strong multicollinearity or other numerical problems.

The results show that there is a here is a negative correlation between the age of the house and its price, with a coefficient of -675.1304 and a p-value of 0.000. The intercept is 575,300, which means that the predicted price of a new house is 575,300. The results also show that age is not a strong predictor of price as the R-squared value is very low.

**Part 2. Multiple Linear Regression**

1. Look at the data file. What other factors in the table you also want to include in the model to avoid omitted variable bias? Why?

Factors that can also influence the price of a house are:

* Cost of materials used: Higher cost of materials used during construction means a higher price for the house.
* Location of the house: If the property is in a posh location, then the prices are bound to go higher and if the location is in an area that is underdeveloped it can go down as well.
* Type of house: Prices may vary if the house is a loft, a bungalow, or a part of an apartment complex.

1. Estimate your new model in Python. Which variables are statistically significant and interpret the coefficients of these variable.

Bedrooms, bathrooms, sqft\_lot, floors, waterfront, view, condition, sqft\_above, sqft\_basement, and AgeOfHouse are all statistically significant for the model and influence the price of a house since the p-value for all the variables is 0.000 suggesting that the null hypothesis is true. The coefficients of these variables are:

* bedrooms -5.623e+04
* bathrooms 6.372e+04
* sqft\_lot -0.3501
* floors 4.529e+04
* waterfront 5.401e+05
* view 6.212e+04
* condition 1.806e+04
* sqft\_above 285.1920
* sqft\_basement 246.2398
* AgeOfHouse 2918.3514

1. Compare the adjusted R2 of the multiple regression model and the simple linear model. What can we know from these two R square values?

The R-squared value for the simple linear regression model is very low (0.003), indicating that age is not a strong predictor of price, whereas the R-squared value of the multiple linear regression model is 0.599. This suggests that the additional variables in the multiple regression model contribute towards the explanation of variance in the dependent variable.

1. What are the four critical diagnostics for linear regression we talked about in the class? Is the estimates reliable based on the results of these critical diagnostics?

* **Regression diagnostic I:** multicollinearity: The classical linear regression model (CLRM) assumes that there is no perfect linear correlation between the independent variables. The given model can be multicollinear since there is a statistically significant negative relationship between the age of the house and its price.
* **Regression diagnostic 2:** heteroscedasticity: heteroscedasticity occurs when the variance of the residuals is not the same across the range of predicted values. Based on the results of the white test, this model does not have a Heteroscedasticity issue, as the p-value of the test is less than 0.00.
* **Regression diagnostic 3:** autocorrelation: occurs when the residuals are correlated with one another, indicating that there is some pattern or structure in the errors that the model has failed to capture. Since it is not time series data, it is not possible.
* **Regression diagnostic 4:** model specification errors: one assumption of the classical linear regression model (CLRM) is that the model used in the analysis is “correctly specified”. By correct specification we mean one or more of the following:

1. The model does not exclude any “core” variables. Not many relevant factors are missing from the model including the cost of materials used, the Location of the house, and the Type of house
2. The model does not include superfluous variables. This is true.
3. The functional form of the model is suitably chosen. Yes, as the dependent variable is a continuous variable, linear regression is suitable.
4. There are no errors of measurement in the regress and regressors. Not sure how variables are measured.
5. Outliers in the data, if any, are considered. According to the summary statistics, the concern about outliers is minimal in this model.
6. The probability distribution of the error term is well specified. We check whether the residuals are normally distributed. As the p-value is 0.00 evidence shows that the residuals are not normally distributed.
7. Can you claim a causal effect of age of a house on its selling price based on your above results? Why

No, we cannot claim a causal effect of the age of a house on its selling price, because linear regression can help establish a statistically significant relationship between these two variables. To establish a causal relationship, we need a clear cause and effect relationship.