

**University of Niagara Falls Canada**

**Master of Data Analytics**

[**Predictive Analytics (DAMO-510-1)**](https://learningspace.myunfc.com/d2l/home/9097)

**Winter 2025**

**Assignment #1**

**Submitted By**

**Aarmi Patel (NF1002881)**

**Professor: Zeeshan Ahmad**

**February 2025**

Table of Contents

[**1.** **PART 1** 3](#_Toc190117007)

[Using the “*statsmodels*” package with a focus on interpretation and explanation (PART I) 3](#_Toc190117008)

[1. Original Model: (30 points) 3](#_Toc190117009)

[2. Log-transformed Model: (30 points) 5](#_Toc190117010)

[**2.** **PART 2** 7](#_Toc190117011)

[Using the “*sklearn.model\_selection*” package with focus on prediction (PART II) 7](#_Toc190117012)

[**1**. **Compare the R² values between the original and transformed models. Why might the R² values not be directly comparable between these models? What additional metrics should we consider? (20 points)** 7](#_Toc190117013)

[**2.** **Using the Python code provided, add a new transformation to the model using the Box-Cox transformation for the price variable. Compare its performance with the log and square root transformations. When would you recommend using Box-Cox over a simple log transformation? (20 points)** 8](#_Toc190117014)

[**3.** **REFERENCES** 9](#_Toc190117015)

# **PART 1**

## Using the “*statsmodels*” package with a focus on interpretation and explanation (PART I)

### 1. Original Model: (30 points)

1. **Interpret the model (coefficients, p-values, R-sq and Fisher) (10 points)**

The original model was created using ordinary least squares (OLS) regression to predict home rates according to various factors such as square footage, age, and the number of bedrooms.

**Coefficients:**

* Square Footage: 95.3695 (positive, significant) - The coefficient of square footage is 95.36, which means that, for every further square foot, the home cost is raised by about $95.36, holding other factors constant.
* Age: −789.34 (negative, significant) - The coefficient for age is -789.34, which means that for all further years of age, the house price drops by about $789.34, considering all other factors stay the same.
* Bedrooms: The coefficient of the number of bedrooms is 20,010, which means that including one more bedroom will increase the price by about $20,010 when all other things are held constant.

**P**-**values**: Each predictor factors have p-values < 0.05, which assures that they are statistically significant in predicting house prices.

**R²**: The model shows 71.2% of the variance in home prices, according to the R² value of 0.712. This suggests a relatively good fit.

**F-statistic:** Generally, the model is statistically significant, according to the F-statistic of 819.1 (p-value = 2.63e-268).

1. **What problem(s) can you identify, and which transformation would be most appropriate to address these issue(s)? (10 points)**

There are several issues identified in the original model; they are given below:

1. Heteroscedasticity: The residual plot suggests that the variance of residuals rises with fitted values, which shows heteroscedasticity. heteroscedasticity is confirmed by the Breusch-Pagan test, Where a p-value < 0.05.
2. Non-normality of Residuals: The residuals have no normal distribution, based upon the results of the Jarque-Bera test where skewness = -0.581 and kurtosis = 8.174.
3. Right-skewed Distribution of Prices: The house prices are drastically skewed, which shows that a transformation might improve model performance.

Recommended Transformations:

* Applying a log transformation to price and square footage might help to normalize variation and make relationships more linear.
* Another approach, a Box-Cox transformation might be considered to figure out the ideal transformation parameter.

1. **Interpret the variance of the residuals. (10 points)**

The results of the Omnibus test and Jarque-Bera test show that the residuals differ significantly from normality. Residuals are the differences between what our model predicts and the actual prices. In our case, these differences are not uniform (they become larger for higher-priced houses). A log transformation of price can improve the model. After transforming, the errors become more consistent for all different price levels, helping the model meet the assumptions needed for a regression analysis. Based on the results of the Breusch-Pagan test, the model shows heteroscedasticity, which indicates that prediction errors are larger for higher-priced homes. The Durbin-Watson statistic of 1.976 shows that there is no significant autocorrelation in the residuals. The high condition number (8.18e+03) indicates the potential multicollinearity, which may lead to even higher variance. A non-normal residual distribution means that standard errors are biased leading to biased confidence intervals and biased p-values (Sladekova, Poupa, & Field, 2024).

### 2. Log-transformed Model: (30 points)

1. **When are log-transformed variables sometimes more suitable? (10 points)**

Log transformation is generally used when:

* The dependent variable is extremely skewed since it lessens right skewness and normalizes the distribution.
* There is heteroscedasticity in residuals because log transformation maintains variance.
* The relationship among predictors and the response variable is not linear, because it helps linearize exponential relationships.

The use of log-transformed variables is suitable when non-linearity, heteroscedasticity or skewed distributions are present in regression models. That helps to stabilize the variance by making residuals more homoscedastic and by making multiplicative effects more interpretable by converting them into additive effects. The log transformation of variables over multiple orders of magnitude reduces the influence of extreme values and makes them closer to normality (Sun & Xia,2024). It is especially useful for right-skewed data where some large are large.

1. **Interpret the model (coefficients, p-values, R-sq and Fisher). (10 points)**

The model is interpreted by assessing the coefficients, p-values, R squared and F statistic. Predictors having a strong influence on the outcome will yield significant p-values (<0.05). The higher the value of r-squared, the greater the model explained the variance. However, adjusted R squared is a better measure if we have multiple predictors because that accounts for model complexity. The model significance is tested by the F statistic, and a large value and small p-value imply that at least one predictor is important. After applying a log transformation to both price and square footage, the results of the model change as follows:

**Coefficients**:

* Square Footage: The coefficient is about 0.4422, which means that a 1% rise in square footage results in a 0.442% growth in cost.
* Age: The coefficient is almost -0.0021, which shows that for every extra year of age, the log-transformed price falls by 0.21%.
* Bedrooms: The coefficient is 0.0504, which means that adding every additional bedroom will increase the log price by 5.04%. P-values: All variables remain significant.

**Statistical Significance:** Every predictor continues to be statistically significant., with a p-value less than 0.05.

**R²:** The value of R² is 0.696, which is just less than the original model’s 0.712, but still, it is strong.

**F-statistic:** The model's F-statistic is 760.1, which shows that it is still highly significant.

1. **Interpret the variance of the residuals? (10 points)**

It helps to interpret the variance of the residuals monotonically and assess model fit and assumptions. High variance represents poor model fit while low variance shows better predictive accuracy. Variance patterns are diagnosed using graphs such as residual plots and statistical tests like Breusch-Pagan. However, if heteroscedasticity is present, transformations or robust standard errors can make the model more reliable. We have an improved homoscedasticity compared to the original model which means that it shows a more consistent variance after log-transformation.

1. Lower Heteroscedasticity: The residuals display a more consistent variance after log transformation. Heteroscedasticity is still recognized by the Breusch-Pagan test(p < 0.05), but the magnitude is significantly less.
2. Improved Normality: As compared to the original model, the Jarque-Bera test results show less skewness and kurtosis.
3. Better Model Fit for High-Value Houses: The log transformation helps by reducing large errors for costly homes, enhancing prediction accuracy.

# **PART 2**

## Using the “*sklearn.model\_selection*” package with focus on prediction (PART II)

### **1**. **Compare the R² values between the original and transformed models. Why might the R² values not be directly comparable between these models? What additional metrics should we consider? (20 points)**

* **The Original Model R² value is**0.6296, and the **Log-Transformed Model R² value is**0.6894. Comparing **the R2** values between the original and the transformed models shows that the log-transformed model explains a higher percentage of the variation in the data when the price is transformed, which suggests it fits the data better on that scale.
* The R² values may not be directly comparable because the log-transformed model predicts log(price), rather than actual cost, which Direct comparison is inaccurate or misleading. A change in the log scale represents a percentage change, not an absolute change in dollars. This means you can’t directly compare the numbers because they are measured on different scales.

Other Metrics to Consider:

* The Mean Absolute Error (MAE) measures the average absolute difference between the predicted and actual values, giving another way to assess prediction accuracy in practical terms.
* Root Mean Squared Error (RMSE) should be used as an additional metric, because it, shows prediction accuracy in original price units. For the original model, RMSE is 60,598, meaning the predictions are off by about $60,598 on average. For the log-transformed model, RMSE is 0.128 (in log units), which isn’t directly comparable to the RMSE of the original model. The Residual Plots show whether the variance is stabilized following transformation. The Out-of-sample Predictions compare models using cross-validation.

### **Using the Python code provided, add a new transformation to the model using the Box-Cox transformation for the price variable. Compare its performance with the log and square root transformations. When would you recommend using Box-Cox over a simple log transformation? (20 points)**

1. **Box-Cox performed worse than log-transformation (R² = 0.67).**
2. **When to use Box-Cox?**

The original Model R² value is 0.712, the Log-transformed Model R² value is 0.696 and the Box-Cox Model R² value is 0.670. The Box-Cox transformation evaluates an optimal lambda parameter, which occasionally produces a transformation that is not very consistent with the original data. Here, in this present scenario, the log transformation was already almost ideal, so Box-Cox did not add much improvement. Different transformation powers could be required if the data has varied levels of skewness. If log transformation does not completely normalize the distribution, Bo-Cox may alter the transformation accordingly. The Box-Cox transformation is a more flexible power transformation than approximating skewed data to normality, making it useful where simple log or square root transformations do not sufficiently stabilize the variance. The price variable is next transformed into its Box-Cox form to compare performance with the log and the square root transformation. Log transformations are useful when data has the right skewness and no zero or negative values as the large values are compressed and the small values are expanded. Log transformations are more aggressive in reducing skewness, but square root transformations also help reduce skewness. While these transformations are often applied when the data is small, Box-Cox generalizes these transformations allowing an estimation of an optimal lambda parameter, which is then used to determine the optimal power transformation to normalize the data. R², RMSE and residual diagnostics are used to compare model performance. The Box-Cox transformation is preferred if it has higher R², lower RMSE, and homoscedastic residuals than log or square root transformation. A box-cox transformation is recommended to a simple log transformation if the optimal lambda is not 0, log transformation, for improved normality, and variance stabilization. In particular, it can be helpful when standard log transformation does not sufficiently alleviate the heteroscedasticity of the dependent variable.

# **REFERENCES**

Sladekova, M., Poupa, V. L., & Field, A. P. (2024, June). *Sources of bias in general linear models: Evaluating the analytic practice in psychological research*. <https://files.de-1.osf.io/v1/resources/wc42b/providers/osfstorage/667e8cbe1c69c0038627487b?format=pdf&action=download&direct&version=2>

Sun, J., & Xia, Y. (2024). Pretreating and normalizing metabolomics data for statistical analysis. *Genes & Diseases*, *11*(3), 100979. <https://www.sciencedirect.com/science/article/am/pii/S2352304223002246>