Prior to completing this discussion, read the [Four Assumptions of Multiple Regression That Researchers Should Always Test](https://sandiego.instructure.com/courses/4795/files/651117?wrap=1)

[Actions](https://sandiego.instructure.com/courses/4795/discussion_topics/57511?module_item_id=278149)

 article. Respond to the following prompt by Day 4 of the learning week. Cite resources and references appropriately in APA format.

* Choose a data-related problem or application in your professional or academic career, social life, or news where a regression model would be useful.
* Describe what assumption(s) from the article you studied would be important to verify for the problem you chose. For instance, would a linear assumption between the dependent and independent variables be valid? Justify your answer.

One data-related problem where a regression model would be useful is in predicting housing prices in a real estate market.

In real estate, buyers, sellers, and agents are often interested in predicting the selling price of a house based on various features such as its size, location, number of bedrooms, number of bathrooms, and other amenities. A regression model can be trained using historical data on sold properties, where the selling prices and corresponding features are known. Once trained, the model can then be used to predict the selling price of new or unsold properties based on their features.

This regression model can help sellers determine a suitable listing price for their property, assist buyers in making informed decisions about the value of a property they are interested in, and guide real estate agents in pricing strategies and negotiations. Additionally, it can provide insights into which features have the most significant impact on housing prices in a particular market, helping stakeholders understand market dynamics and trends.

Overall, a regression model for predicting housing prices can be a valuable tool in the real estate industry, facilitating more accurate pricing decisions and improving the efficiency of property transactions.

One example for a regression model would be predicting the selling price data of a house in the real-estate market.

There are various factors which can affect the selling price of a house – for example, size, location, new or pre-owned house, number of bedrooms, number of bathroom, and school district. For regression model, the selling price would be the dependent variable, while the factors affecting the price (size, location, new or pre-owned house, number of bedrooms, number of bathroom, and school district) would be the independent variables.

Once the regression model is trained, it can be tested on a section of data which is not used as the training data. Then, if the accuracy score is relatively good, one can use the regression model to predict the price of unsold properties. Thus, it can assist buyers in making informed decision about the value of a property and its expected price. It can also tell the real-estate agents about pricing strategies and negotiating points.

Here is a list of assumptions for regression model:

1. **Variables are normally distributed**: Regression models assume a normal distribution of the variables. If the variables are not normally distributed, it can distort the relationship between the independent and dependent variables. Thus, it would be important to verify the normality assumption by plotting bar charts and histograms, and performing tests like Shapiro-Wilk test on the data. It is also important to remove the outliers from the data by visual inspection of the data or using statistical software. For the example chosen here, the size of the house, number of bedrooms and bathrooms will probably be normally distributed. However, there may be some outliers for the size of the house data for really high-priced houses.
2. **Linear relationship between dependent and independent variables**: Regression models also assumes that there is a linear relationship between the dependent and independent variables. There are a few methods available to detect a nonlinear relationship between the dependent and independent variables. One of the methods is visual inspection of the residual plot of the data Another approach will be performing regression analysis using quadratic and cubic terms or using nonlinear regression options in statistical software packages. For the example chosen here, there is possible linear relationship between the size of the house, number of bedrooms and bathrooms and the selling price.
3. **Variables are measured without errors**: Making accurate measurements is an important consideration in regressions. If the measurements are not accurate, it can lead to under-estimation or over-estimation. Statisticians need to correct for low-reliability data which can be done using statistical software. For our example, if the size of the houses is not accurately measured, it can lead to misinterpretation of the regression model.
4. **Reliability and Multiple Regression**: The reliability assumption states that in the regression model, the independent variables are measured with good reliability and accuracy. If the reliability assumption is not met, it can lead to under-estimation or over-estimation. The model is these cases will not lead to accurate predictions. Statisticians should perform careful assessment the reliability of the measurements for multiple regression. Reliability tests should be used when the data is susceptible to human errors and the number of data is small. Researchers can use various types of tests for reliability tests for multiple regression. These can be *remeasurements* within raters, across raters, split halves; *absolute difference tests* like mean difference, range of differences; *Pearson correlation test and Cronbach’s alpha* estimating average inter-item reliability. For our example, if the size of the types of location is small and is prone to human errors, then it can lead to less reliability of the data and multiple regression can give incorrect results.
5. **Homoscedasticity**: Homoscedasticity means that the variance of errors is the same across all levels of the independent variables – otherwise it is called heteroscedasticity. If the heteroscedasticity is large, it can lead to severe distortions of the results for multiple regression. Homoscedasticity can be checked by visual inspection of the standardized residual plots. If the residual data is randomly scattered around zero, it leads to the conclusion of homoscedasticity. Statistical tests like Goldfeld-Quandt test and Glejser test can be used to determine heteroscedasticity which can be done using standard software. For our example of predicting selling-price of a house, this would mean that the variability in prediction errors is consistent for all levels of house features. If the variance of errors changes has different levels for the independent variables it could indicate that the model is not appropriately captured using the multiple regression model.