Homework 3: Helena Mabey

Part 1

Question 1

There are four assumptions of linear regression. They are linearity, independence of errors, normality of errors, and an equal variance of errors (homoscedasticity). Each of these assumptions has unique methods of completing diagnostic analysis of the assumptions as well as how to overcome violations of assumptions once they are discovered.

The first assumption is linearity, is there a linear relationship between the dependent and independent variables. The independent variable should change constantly for every change in the dependent variable and this should show graphically as a straight-line relationship. To perform diagnostic analysis of this assumption, scatterplots and residual plots can be used. All of the data points in the scatterplot should follow a straight line. If the line is curved, this would show a violation of the assumption. The residual plot should have points randomly around the zero point. If this plot shows a curved pattern, it shows the assumption is violated. In order to overcome the violation, there are multiple possible solutions depending on the situation. A nonlinear transformation can be applied, such as using a log transformation when the data is positive (logs cannot be used on negative values). The Duke article states that natural log is often used since small changes in natural log are useful for percentage changes. Another suggested solution is adding another regressor that is a transformation of a variable value, such as squaring or cubing the dependent or independent variable. This assists when a value is negative and cannot be transformed using log. Another option is adding another variable that may further explain or clarify the initial relationship.

The next assumption is the independence of errors, meaning the errors are statistically independent and should not have any correlation. Performing diagnostic analysis can be completed on this assumption by reviewing residual time series plots or a table of residual correlation. A violation would occur if there were identifiable patterns in the data. Additionally, a Durbin-Watson test can be completed to determine residual autocorrelation. This test provides a result between zero and four. A result of two is the best outcome, showing no autocorrelation. If the value returned is less than two, it shows a positive autocorrelation, where residuals follow a trending pattern. If the value is greater than two, it shows a negative autocorrelation, where the residuals vary in a back-and-forth type of pattern. There is a package available in Python to run this test on residuals of a linear regression model. In order to overcome the autocorrelation in the model, different steps are taken depending on if the autocorrelation is positive or negative. Positive autocorrelation can be resolved by using a lagged variable for either the dependent variable (looking at past values to try and predict a future value) or the independent variable (looking at past independent variables to try and predict the dependent variable). This is often used in time-series data. There are also some different models that can be used, such as ARIMA, that may resolve the violation. If there is a negative autocorrelation, it can possibly be resolved by removing transformed data which has possibly been over correlated or reviewing for missing predictive data such as seasonality.

The next assumption is normality of errors, meaning the residuals are normally distributed. To perform diagnostic analysis for this assumption it is best to use a normal probability plot. This helps locate outliers to assure most of the values are near the average. If the line deviates from a straight diagonal line, this will show a violation of the assumption. There are different ways to resolve the violation of normality of errors. One would be to use a nonlinear transformation of the variables. This is useful when the variables are not normal, possibly heavily skewed. This can be solved similarly to a violation of linearity by using either a log transformation or a transformation of a variable by methods such as squaring. This should result in a more normal variable. Another option would be to review for removal of outliers. If a few extreme outliers are severely impacting the normality. These can be reviewed to see if they are potentially typographical errors or if they are legitimate errors. If they are legitimate, they can be analyzed to see how impactful they are to the overall model. If they are one-offs, they could potentially be removed which could resolve the violation.

The last assumption is the equal variance of the errors or homoscedasticity meaning that there is a constant variance throughout the probability distribution. Similar to other assumptions, a diagnosis analysis can be performed by using a scatterplot. If the errors are random near zero, this shows validation of the assumption. If the errors form a growing widespread pattern, this would be an indicator of a violation. There are also some tests that can be performed within the Python statsmodels package. The violation can be overcome using transformation methods. The log transformation can be used for this violation as seen in some prior assumption violations. Differently than some prior violations, a square root transformation is also suggested to correct the variance.

* AI Statement for question 1: I utilized the given PDF important self-reading PDFs as a starting point to research how to perform a diagnostic analysis and how to resolve the violation. This led to multiple ChatGPT searches for various testing methods, like graph types, and Python tests and packages used to determine violations. I needed to utilize ChatGPT to also interpret the articles presented as some of the provided information was beyond my general knowledge level from information provided in class. I needed to understand what they were conveying to determine which solutions were best for each assumption.

Question 2

1. The first set of features reviewed was house age and house price of unit area (price). The first step of the comparison was to review the non-graphical statistics. The house age data contains 414 records with a minimum year of 0 and a maximum value of 43.8. Based on the median and mean, it appears that the data may be relatively normal distributed but further investigation is required. A similar review of the price shows a similar spread between the values with the median and mean very similar. In this case, the minimum and maximum values have a large variance. This could imply that there are outliers at the high end of the values. Next, we reviewed the skewness and kurtosis values of both features. The house age skewness value of 0.383 is slightly right-skewed which could indicated a relatively normal distribution. The kurtosis value of -0.877 below 3 indicates a distribution with a flatter peak than a normal distribution and possibly fewer outliers. The price skewness of 0.5888 appears to be more moderately right-skewed than the house age, indicated there may be some higher outliers causing the skew in distribution. The price kurtosis value of 2.179 above 3 shows that the distribution is more sharply peaked than a normal distribution with quite a few outliers. A correlation value of -0.211 shows a negative correlation between the two features but it appears to be very small. House prices do fall as the age increases but there may be other features with more impact. This theory is also shown with the P-Value of 0.00 which is less than 0.05 meaning that there is a correlation but it is very slight.  
   A graphical EDA was performed to validate the results found above. A boxplot was created for each of the features. The house age boxplot confirmed a slight right-skewed distribution where the median value is slightly left of the average value and the tail is longer on the right side. The price boxplot also shows a right-skew but it is considerably more defined with multiple large outliers on the far right. Next, a scatterplot was created to compare the relationship of the features. This visual also showed that while there may be a relationship between house age and price, it is not very strong. There are a few outliers but, generally, the price is similar regardless of house age. Another suggested visualization was a multiple box plot with grouping. This also confirmed that a few high outliers may be impacting the distribution of the data. It does show an interesting pattern when a house reaching the maximum age reviewed, the prices increase higher than at any other age. Lastly, a regression plot was created to show the relationship with a regression line and confidence interval (CI). This again shows the slight negative relationship between the features. This confirms the correlation results previously reviewed indicating that age has a minimal impact on price. The CI stays relatively thin along the regression line meaning there is relatively strong confidence in the trend with the confidence decreasing slightly at the low and high ends of the age range. Overall, the results show that the house age has a minimal impact on the price of the house.
2. The next features reviewed were the distance to the nearest MRT station (MRT) and the house price of unit area (price). The first step was to review the non-graphical statistics. The MRT data have a minimum value of 23.38 and a maximum value of 6488.02. This shows a large distribution of values. This is confirmed a median value of 492.23 and a mean value of 1083.89 showing that the majority of the values are on the lower end of the range. As stated above, the price median and mean values are similar but there is also a large range between the minimum and maximum values, implying the possibility of high outliers. A review of the skewness value of 1.89 for the MRT distribution shows a pretty defined right-skew with many outliers on the right side. The kurtosis value of 3.21 above 3 is very high indicating the distribution has a high peak with many outliers and heavy tails. As stated above, the price skewness of 0.5888 appears to be more moderately right-skewed than the house age, indicated there may be some higher outliers causing the skew in distribution. The price kurtosis value of 2.179 above 3 shows that the distribution is more sharply peaked than a normal distribution with quite a few outliers. The correlation value of these features is -0.674. This shows a relatively moderate negative correlation between these features. As the house value increases the MRT value decreases. The P-Value of 0.00 for this set of features is again under 0.05 meaning these is a relationship but it may be very slight.   
   A graphical EDA was performed to validate the results found above. A boxplot was created for each of the features. The boxplot for MRT clearly shows the defined right-skew of the distribution and a large number of outliers on the right side of the graph. The price boxplot also shows a right-skew but it has considerable outliers on the far right. Next, a scatterplot was created to compare the relationship of the features. This shows a relatively strong negative relationship between MRT and price. Where the MRT value is less, the price of the homes appears to increase. A multiple boxplot was created for these features as well. The first iteration of the boxplot was quite compressed due to the number of outliers in the shorter distance range. A log transformation was suggested to try and decompress the results for improved readability and clarity. This graph shows a slightly negative relationship as the distance increases. Lastly, a regression plot was created to show the relationship with a regression line and confidence interval (CI). The negative relationship is clearly defined with agrees with the previous analysis. Also, the CI is very strong where the majority of the results are seen (short MRT distance) but becomes less defined as the distance increases. The results show that homes closer to a MRT station generally are higher priced than those farther away from a station.
3. The last features reviewed were the number of convenience stores (stores) and the house price of unit area (price). The first step was to review the non-graphical statistics. The stores data is very categorical with distinct values between 0 and 10. The mean and median are nearly identical indicated the data may be normally distributed. Again, we review the price data. The price median and mean values are similar but there is also a large range between the minimum and maximum values, implying the possibility of high outliers. The stores skewness value of 0.155 shows a very slight right-skew but it is nearly normal. There may be a few outliers on the right. The kurtosis value of -1.066 shows that the distribution has flatter peaks and fewer extreme values. We again compare the price values as well. The price skewness of 0.5888 appears to be more moderately right-skewed than the house age, indicated there may be some higher outliers causing the skew in distribution. The price kurtosis value of 2.179 above 3 shows that the distribution is more sharply peaked than a normal distribution with quite a few outliers. The correlation value of 0.571 shows a moderate positive correlation between these two features. The P-Value of 0.00 for this set of features is again under 0.05 meaning these is a relationship but it may be very slight.  
   A graphical EDA was performed to validate the results found above. A boxplot was created for each of the features. A boxplot was created for the stores data which confirmed the very slight right-skew of the data. The price boxplot also shows a much more defined right-skew with considerable outliers on the far right. A scatterplot was created for these features as well. There is a positive relationship between these features. It shows that as the number of stores increases, the price of the homes increases as well. There are a few outliers but they don’t appear to have a strong impact on the relationship. The multiple boxplot graph also shows the positive relationship between the features. The outliers are more visible within this graph. Lastly, a regression plot was created for these features as well. The positive relationship is clearly defined with agrees with our previous analysis. The confidence interval is very strong within the median number of stores where many of the points are located. These results show that homes near more convenience stores are generally priced higher than those with fewer convenience stores.
4. The simple linear regression model was completed and annotated within the provided notebook.
5. The regression function for this data is house value = 42.58 – 0.27\*X. The intercept is the base value at zero years and it is reduced by -0.27 for every year greater. This confirms that even though house age does impact and decrease the house price, it is by a very minimal amount. Other factors may have more impact than just the house age in the price.
6. 1. State Hypothesis: The null hypothesis states that the house age has no significant impact on the house price. The alternative hypothesis is that the house age has a significant impact on the house price.
   2. The required significance level is 0.05.
   3. A screenshot of a computer

      AI-generated content may be incorrect.  
      The test statistic is determined by taking the coefficient value of -0.270 – 0 and dividing it by the standard error of the slope which is 0.071. The result is -3.806. The P-Value is 0.00.
   4. The critical value for this is +/- 1.968 as outlined in the recorded class session on hypothesis testing and can also be found using the query listed in the notebook (provided by the class and ChatGPT).
   5. The absolute value of the test statistic, 3.806, is greater than the critical value of 1.968 so therefore we reject the null hypothesis. Additionally, the P-Value of 0.00 is less than the alpha value of 0.05 which also leads to rejecting the null hypothesis.
   6. Based on our analysis, we can conclude that house age does have an impact on house prices. As homes age (the house age increases), the house value decreases.
7. The R^2 value of the training set is slightly higher than the value for the test set but both values are very low. This is not a characteristic of overfitting. Based on what we’ve seen about the relationship between house age and house price, this possibly confirms that additional features should be reviewed to determine what impacts house prices.
8. Using the house age as the only predictor in house prices is very limiting. We have concluded that house age has a very minimal impact on house prices. It would be beneficial to add in other features, such as distance to MRT station or number of convenience stores, to supplement this analysis for a fully rounded picture of impacts on house prices. Limiting an analysis to on feature when others are available is not an efficient use of the data provided.
9. I utilized ChatGPT to assist in creating a multiple linear regression model by gathering the full sample code required. The code is annotated in the attached notebook.
10. The multiple regression model has significantly better performance than the simple regression model. In a review of the R^2 values, this is very clear. Using house age only accounts for just 2.44% of the impact in the test data while using both the house age and distance to the MRT station that percentage increases to 54.01%. A search online then using ChatGPT finds that a higher adjusted R^2 value adds to the predictive value of a model and the multiple regression model’s value is considerably higher than the simple model. Additional research showed that lower Root mean squared error and mean squared error values show that a model have better predications with less chance of error. The multiple regression model was lower than the simple model for both of those values. This proves that adding an additional value to the model significantly increasing the performance of the model.

Question 3

* 1. The 95% confidence interval is (6.315, 7.872).
  2. We can say with 95% confidence that the mean amount spent for lunch will fall within the confidence interval range of 6.315 and 7.872 ($).
  3. The population distribution should be normal.
  4. The distribution is not fully normal as there is a very slight right-skew. Because the data is nearly normal, it should be acceptable to use in determining the confidence interval.

Question 4

1. No, there is not enough evidence that the mean amount spent for lunch is different than $6.50.
   1. State Hypothesis: The null hypothesis states that the mean amount spent for lunch is $6.50 The alternative hypothesis is that the mean amount spend for lunch is not $6.50.
   2. The required significance level is 0.05. The sample size is 15.
   3. The test statistics for this population a t distribution with 14 degrees of freedom (1 less than sample size).
   4. The test statistic and p-value calculations are found in the notebook. The test statistic is 1.634 and the p-value is 0.1245.
   5. The p-value is greater than 0.05 significance so we cannot reject the null hypothesis.
2. The population distribution should be normal.
3. The population size of 15 is small so the slight right-skew of the data is not reason for concern. If the population size was greater or the shape of the distribution was more skewed, that could prove reason for concern about the population.

AI Statement for part 2

* I utilized ChatGPT and multiple web searches for the code used for these hypothesis testing scenarios. Because I was not familiar with the python code required to calculate the test statistics as well as how to complete the evaluations of the hypothesis tests, I needed to use the codes provided by ChatGPT to have success in this assignment. I also used ChatGPT to assist with the histogram needed to test the shape of the population distribution. I used it for test code for the majority of the queries for question 2 in addition to the sample notebooks provided by the professor for simple regression modeling.