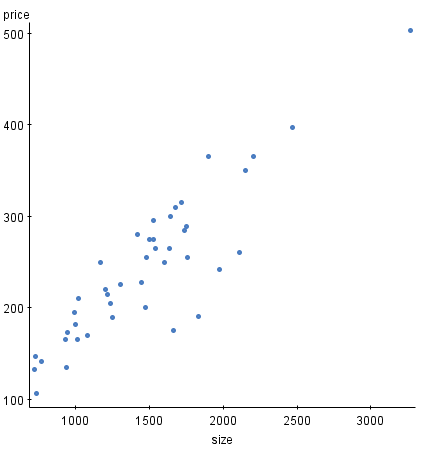
1. Explanatory = size

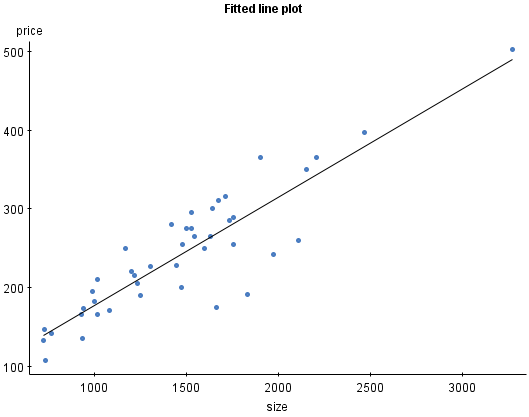
Response = price

1. As the size of the houses go up, so do the prices. There seems to be a linear correlation between the size and price. Though the relationship isn’t perfectly linear, the price of the house tends to increase as the size of the house increases.



3) The R (correlation coefficient) = 0.8911. This means that there is a strong correlation between the size of the house and the price of the house, since r is close to 1. Thus, there is a strong relationship between the two variables. The size of the house strongly affects how much the cost of the house. Not only is there a functional relationship between the two variables, but there is also a causal relationship. The size of the house would naturally result in the price increasing, since more materials would be needed for the house to be built, more time and resources would be needed, etc.

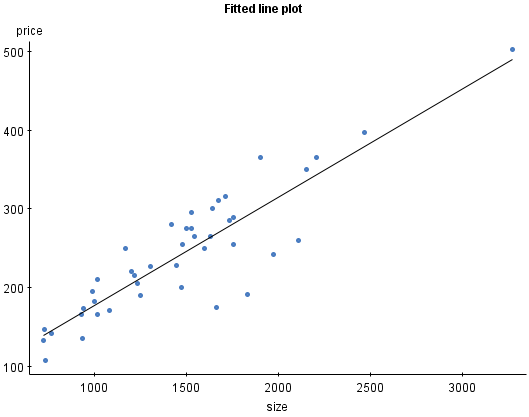
4)



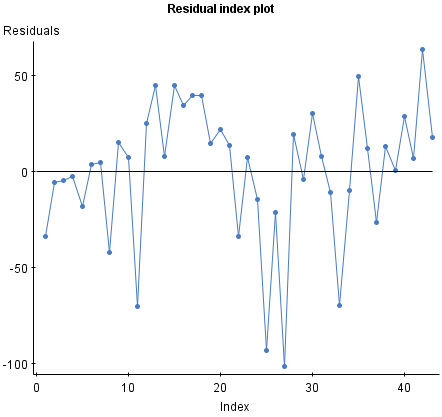
5) b1 = 0.13762467 size is the average rate of change. For ever increase in the size by 1, we would expect the price to be 0.13762467 greater. b0 = 39.96205, which illustrates the average size value of y when x = 0. . When the size of the house is 0, we would expect the price to be approximately $39,000, though realistically this isn’t the case, and this does not make a lot of sense. Thus, it is important for the size of the house to be greater than 1 for the equation to make sense.

6) 0.13762467 (2200) + 39.96205 = 342.736324. Approximately $342,736.32!

7) There are a few irregularities on the fitted plot line. Between a size of approximately 1500 and 2250, there are four data points that fall significantly below the fitted line plot. Those points fall below the price of some of the houses that are smaller, which are more expensive. This raises some questions. Are there other variables that affect the price of the house, besides the house? Is the house in good condition? Were the houses on sale? These are thing to consider in order to determine whether those specific points fit into this data set, or if they belong somewhere else.



The residual plot is more or less random, which gives the line a good fit. The data points are all fitted around the horizontal access, and there are no discernable patterns. Thus, the data is a good fit. There are two points that fall far below the horizontal axis (the two points between the index of 20 and 30). As with the fitted line plot outliers, these data points need to be analyzed to see if they fit with this data set. Otherwise, the data is close to the horizontal axis, and the pattern is more or less random.

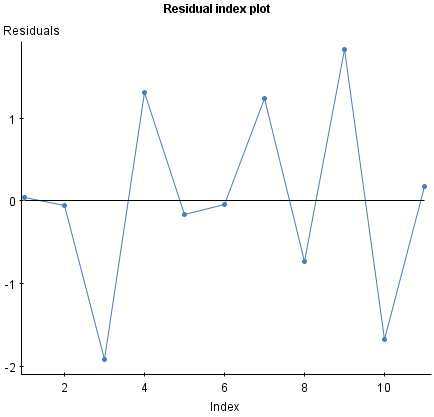
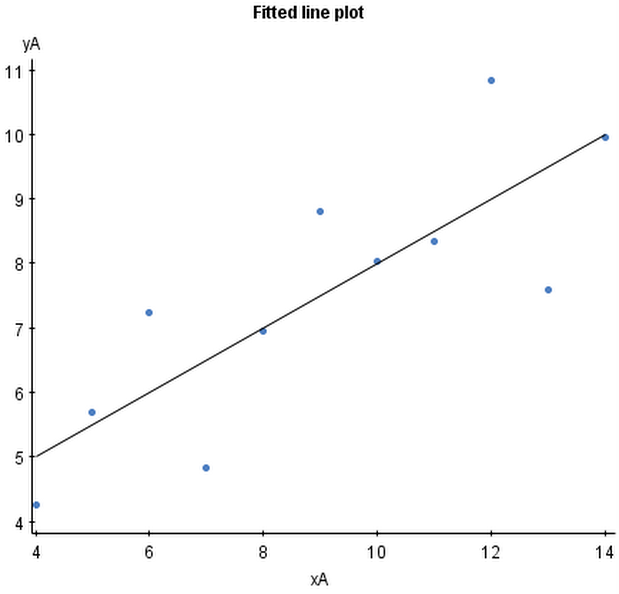


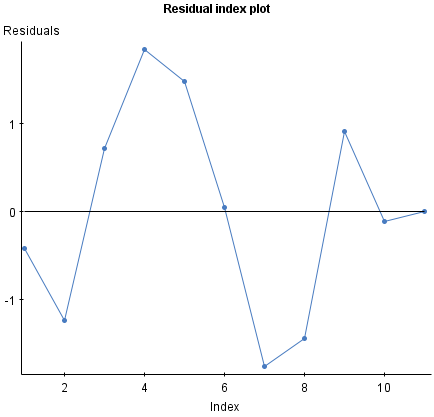
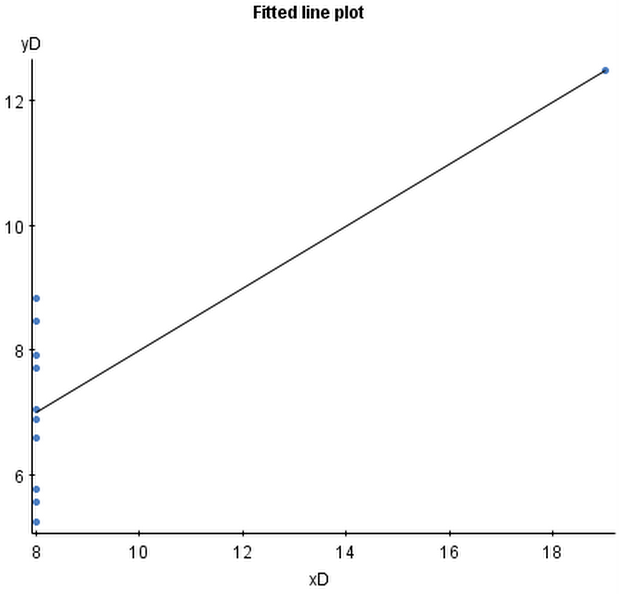
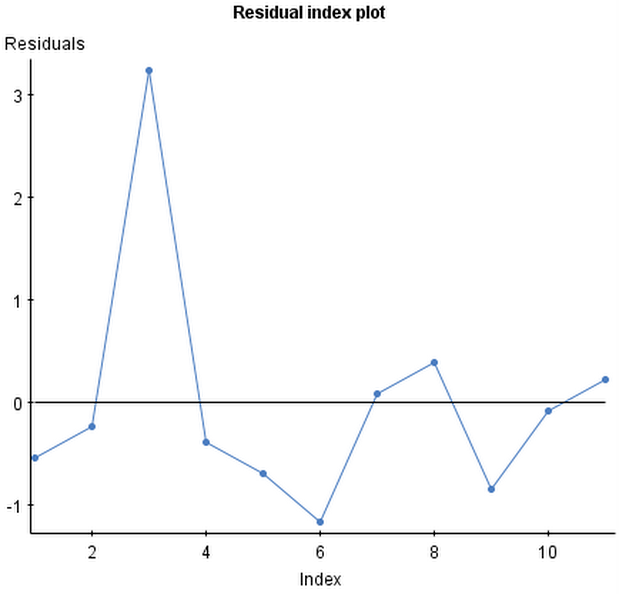
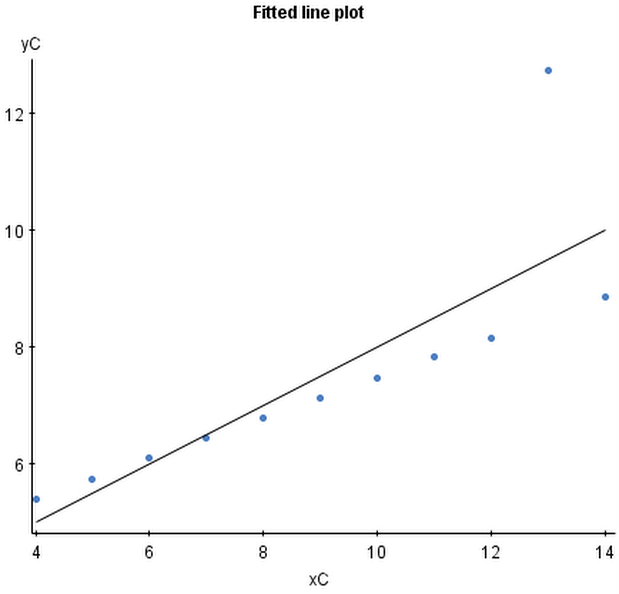
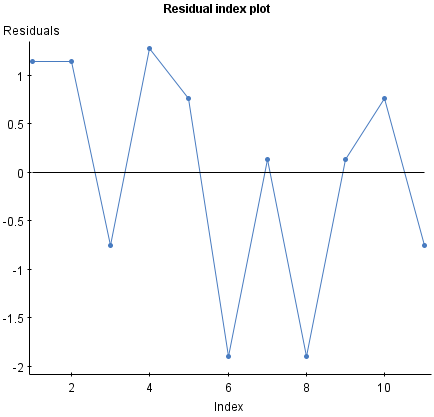
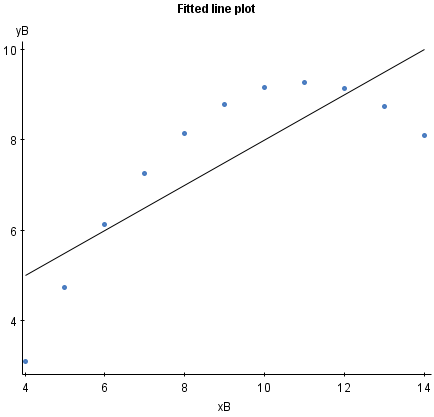


|  |  |  |  |
| --- | --- | --- | --- |
| Regress | Least squares regression line | r | R-squared |
| yA on xA | 3.0000908 + 0.5000909 xA | 0.8164 | 0.6665425 |
| yB on xB | 3.000909 + 0.5 xB | 0.8162 | 0.66624206 |
| yC on xC | 3.0024545 + 0.49972728 xC | 0.8163 | 0.666324 |
| yD on xD | 3.0017273 + 0.4999091 xD | 0.8165 | 0.6667073 |

The results are all nearly identical, and the values of r and r-squared are all very similar across all of the values. The equations of the least squares regression line are also similar across the board.

1. If the points on a residual plot are randomly dispersed around the horizontal access, then a linear model is fitting for the data. If the residual plot follows a set pattern, then the data would be better fitted for a non-linear model. The data sets would probably be better fitted for a non-linear model. That residual plots have patterns, and some of them (such as the first and fourth graphs) are relatively symmetric, making them not as suitable for linear models. So even though the variables may have a high r value and a relatively high r2 value, a linear model may not necessarily be appropriate for each model. Only the third graph (yC on xC) seems to fit well with the linear model, as the residual graph has no dicernable pattern.





1. Blindly running regressions on any set of data can have very varying results. The table above shows that the regression line has a moderately good fit with the data. However, the residual data can illustrate that we have not correctly specified the relationship between the explanatory and response variable. Thus, it is important that we make sure that there is a causal relationship between the variables. Just because a functional relationship exists, it does not mean that the two variables can causally be linked, making it important to make sure that there is a causal relationship between the explanatory and response variables in order to make sure that analyzing the data makes sense. Just because there is a strong r value and the linear regression line is strongly fitted, it does not mean that regression is appropriate for that set of data.