Data to be used can be found at the following link.

http://www.statsci.org/data/general/auction.html

1. Explain Data source and definitions.

The data are from Mendenhall and Sincich (1993, page 173). Professor WR Stephenson at Iowa State University observes that the currency should be pounds sterling and not dollars as stated in Mendenhall and Sincich (1993).

The variables used in this model are:

Age: age of the clock (Independent)
Bidders: Number of individuals participating in the bidding (Independent)
Price: selling price of the clock (Pounds sterling) (Dependent)

2. Main features of data set presented with appropriate graphics

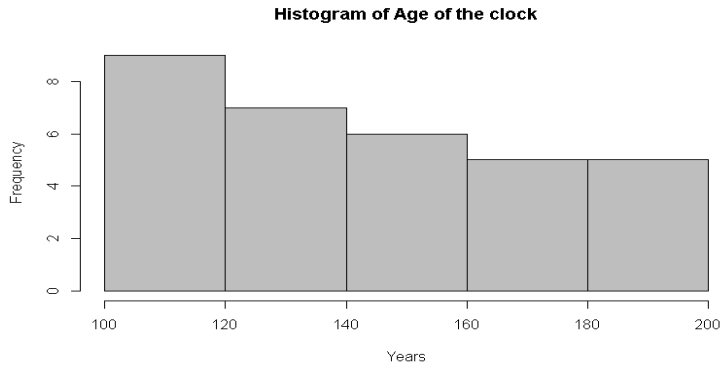


Figure 1. Histogram of Age of the clock

Figure 1 shows the histogram of age of the clock, it has a right skewed distribution, most of the data are in the range between 100 and 160 years old.

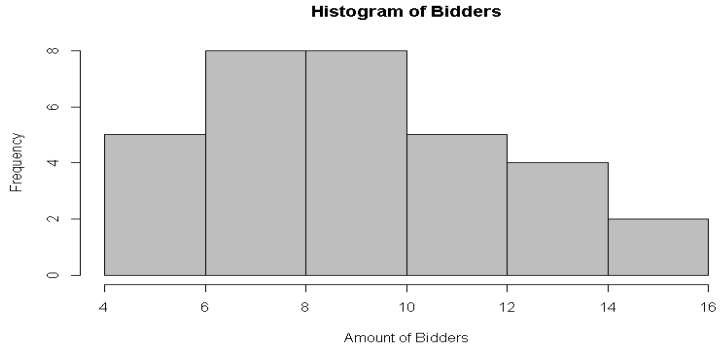


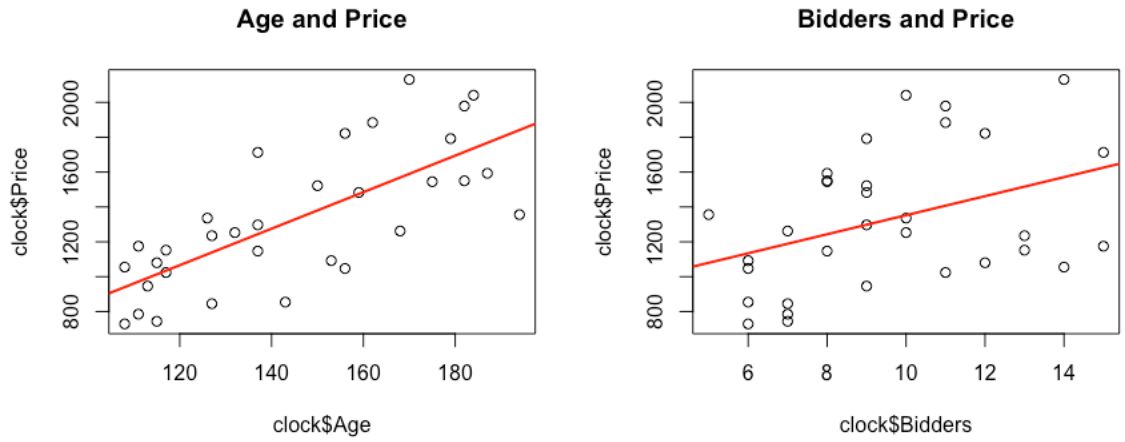
Figure 2. Histogram of Bidders

Figure 2 shows the histogram of numbers of bidders; this graph shows a close normal distribution with a slightly right skewed distribution.

Figure 3. Histogram of selling price

Figure 3 shows the histogram of selling price, which has a close to normal distribution.

We can see not all variables follow a normal distribution. However, being normally distributed is not a requirement for a multiple regression analysis, it actually won't affect the test. We only need to check if the residuals follow a normal distribution, which we will do later**.**



From the above two graph we can see that there is a linear relationship between price and age, as well as price and bidders. Therefore, it makes sense for us to fit multiple regression on our data.

3. Present what research questions can be asked for the given Data.
Are the ages of the clock and number of bidder predictive variables for selling price of the clock and If they do predict the selling price for the clock, how well do they predict it?
We are using regression analysis for multiple predictor variables to study the relationships between the variables.

4. Explaining how regression analysis can be used to address the research questions from question three.

A regression analysis is defined by Hair et al. (2009) as statistical technique than can be used to analyze the relationship between a single dependent variable and several independent variables. The objective of multiple regression analysis is to use the independent variables whose values are known to predict a single dependent value selected by the researcher.
In that sense, this technique is conducted in this analysis using variables age and bidders as independent variables and selling price of the clock as dependent variables with the objective to assess the prediction power of selling price of an antique grandfather clock using age of the clock and number of individuals in participating in the bidding as predictor variables.

5. Explain how the data given satisfies the requirement and assumptions for multiple linear regression.

Assumptions: 7 basic assumptions are needed to be met in order to conduct multiple regression tests

- **Assumption #1: dependent variable** should be measured on a continuous scale, in this analysis the dependent variable used is selling price of the clock, which is a continuous variable.
- **Assumption #2:** There are **two or more independent variables**, which are age of the clock and number of bidders for the clock.
- **Assumption #3: independence of observations**

```
Durbin-Watson test

data:  mod_1
DW = 1.8643, p-value = 0.3471
alternative hypothesis: true autocorrelation is greater than 0
```

This assumption was tested using Durbin-Watson test, which has a value of 1.864, the range to met this assumption is 1.5 to 2.5, moreover, the p value is greater than .05, meaning that we fail to reject the null hypothesis of no autocorrelation in the model. We can conclude that the residuals are not autocorrelated.

- **Assumption #4: The independent variables and residuals are uncorrelated**

```
        Pearson's product-moment correlation

data:  clock$Bidders and mod_1$residuals
t = -1.0457e-15, df = 30, p-value = 1
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.348694  0.348694
sample estimates:
          cor
-1.909095e-16


        Pearson's product-moment correlation

data:  clock$Age and mod_1$residuals
t = -1.1486e-15, df = 30, p-value = 1
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
 -0.348694  0.348694
sample estimates:
          cor
-2.097092e-16
```

We used the Pearson's product moment test for measuring the association between the independent variables. As we can see the p-value is high for both variables which leads to the decision that the null hypothesis that true correlation is 0 can't be rejected leading to the conclusion that assumption holds true for this model.

- **Assumption #5: No perfect multicolinearity**

```
> vif(mod_1)
Bidders     Age
1.06882 1.06882
```

We used the Variance inflation factors measure to measure the inflation in the variance of the parameter estimates due to collinearities that exits among the predictors. A VIF of 1 means that there is no correlation among the kth predictor and the remaining predictor variables, and hence the variance of βk is not inflated at all. The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction. This assumption is met since the values are lower than the conventional value of 4

- **Assumption #6:** Your data needs to show **homocedasticity**, which is where the variances along the line of best fit remain similar as you move along the line.
- **Assumption #7:** The residuals follow a **normal distribution**

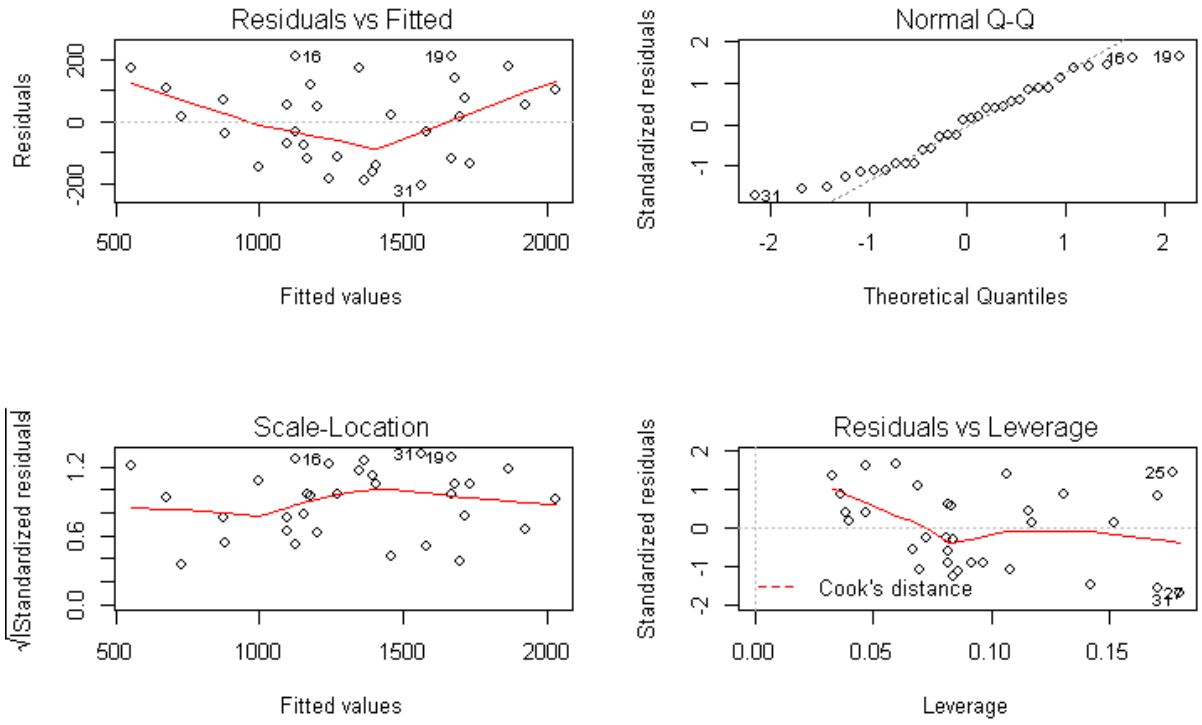


Figure 7. Assumption graphs

Figure 7 shows the scatter plot using Regression standardized residual vs. regression standardized predicted value as a way to assess the homocedasticity, in the top right, as the scatter plot follow a pattern the assumption is not met.

The normality of the residuals are evaluated through the Normal QQ plot (top right), this follow a fair normal distribution with some deviation. On the bottom right, we can see the Residual vs Leverage plot which uses the Cook's distance to see if there are any influential cases that if we take it out it might affect the regression line. As we can see from the plot, we can barely see the Cook's distance lines because all cases are well inside of the Cook's distance line.

Summarizing all the assumption are met except for homocedasticity

6. Apply the methods to the data and interpret It correctly.

```
Call:
lm(formula = Price ~ Bidders + Age, data = clock)

Residuals:
   Min     1Q Median     3Q    Max
-207.2 -117.8   16.5  102.7  213.5
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1336.7221   173.3561  -7.711 1.67e-08 ***
Bidders        85.8151     8.7058   9.857 9.14e-11 ***
Age            12.7362     0.9024  14.114 1.60e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 133.1 on 29 degrees of freedom
Multiple R-squared:  0.8927,   Adjusted R-squared:  0.8853
F-statistic: 120.7 on 2 and 29 DF,  p-value: 8.769e-15
```

A multiple regression was run to predict selling price of clock from number of bidders for the clock and age of the clock. These variables statistically significantly predicted the selling price, $F(29) = 120.7$, $p < .0005$, $R^2 = .893$. The two variables added statistically significantly to the prediction, $p < .05$.

The regression equation is:

$$Y = -1336.7221 + 85.8151(Bidders) + 12.7362(Age)$$

This means that for every bidder the predicted price will increase by 85.82 pounds sterling with a constant age, on the other hand, for every year add to the clock it will be an increase of 12.74 pounds sterling with a constant bidder. We can also notice that both the Multiple R-squared and the Adjusted R-squared are approximately 89% which shows how well the regression model explains the data.

References:

Mendenhall, W, and Sincich, TL (1993). *A Second Course in Statistics: Regression Analysis, 6th Edition*, Prentice-Hall.

International Journal of Pediatrics Volume 2009, Article ID 952042.

Hair, J., Black, W., Babin, B., & Anderson, R. (2009). Multivariate Data Analysis. 7th edition. Prentice Hall.