Supporting Student Learning with Predictive Tools: Using R in Regression Analysis

What is regression? Regression is a statistical technique that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).

The process of finding a mathematical model (an equation) that best fits the data (relationship) is part of a statistical technique known as regression analysis.

The two basic types of regression are

- linear regression - uses one independent variable to explain or predict the outcome of the dependent variable Y.
- multiple linear regression - uses two or more independent variables to predict the outcome.

The general form of each type of regression is:

- Linear regression: $Y = \beta_0 + \beta_1 X + \varepsilon$
- Multiple regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$

Where:

- Y = the variable that you are trying to predict (dependent variable).
- X = the variable that you are using to predict Y (independent variable).
- β_0 = the intercept.
- β_1 = the slope.
- ε = the regression residual.

Simple Linear Regression:

The straight line model for response y in terms of x:

$y = \beta_0 + \beta_1 x + \varepsilon$

The line of means: $E(y) = \beta_0 + \beta_1 x$

Fitted line we plan to find is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$\hat{y}$ is an estimator of the mean of y, E(y), and predictor of future value of y.

$\hat{\beta}_o$ and $\hat{\beta}_1$ are estimators of β_0 and β_1.

For a given point (x_i, y_i), the predicted value is obtained in $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

The deviation of the y_i and the predicted values ($\hat{y}_i$) is called the *i*th **residual**:

$$(y_i - \hat{y}_i) = [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]$$

Sum of the squares of residuals is

SSE $= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i-1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$

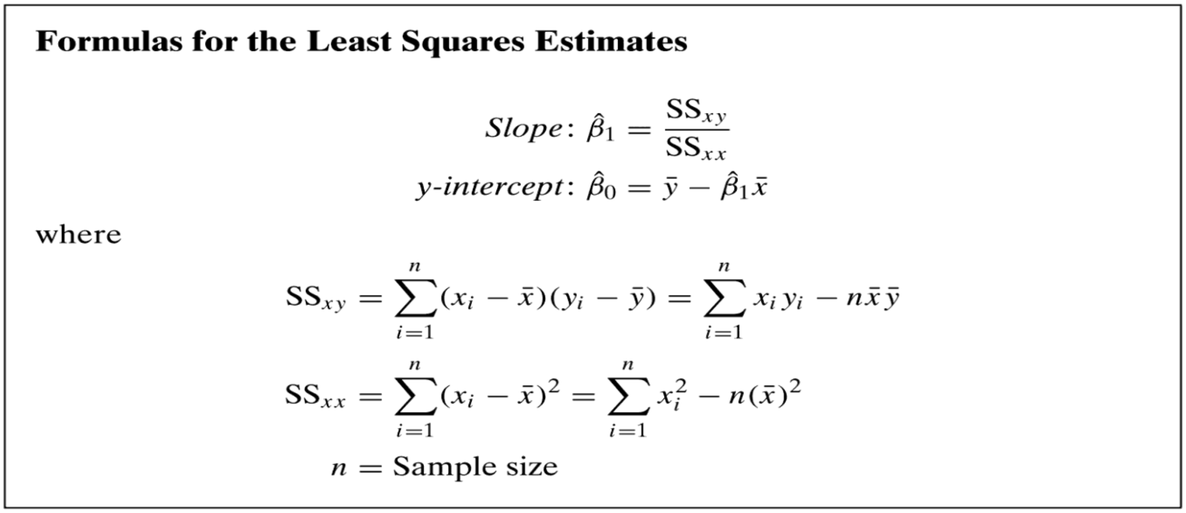
The quantities $\hat{\beta}_o$ and $\hat{\beta}_1$ that makes the SSE a minimum are called **least squares estimators** of β_0 and β_1.

The resulting perdition equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is called the **least squares regression line**.

The least squares line satisfies following:

1. SE $= \sum (y_i - \hat{y}_i) = 0$; sum of residuals is 0.
2. SSE = $\sum (y_i - \hat{y}_i)^2$ is smaller than for any other straight line model with SE = 0.

The value of $\hat{\beta}_o$ and $\hat{\beta}_1$ that minimize the SSE are given by formulas below:



Formulas for the Least Squares Estimates

$$\textit{Slope}: \hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}}$$

$$\textit{y-intercept}: \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$\text{SS}_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$\text{SS}_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$$

n = Sample size

Steps in Linear Regression

- Check the utility of the hypothesized model, that is, whether x really contributes information for prediction of y using the model.
  - Hypotheses test for slope.
  - Confidence interval for slope.
  - Numerical descriptive measures of model adequacy: Coefficient of determination *r^2*, correlation coefficient *r*.
- If satisfied, use the model.
  - Predictions for given x.
  - Prediction intervals for y for a given x.
  - Confidence intervals for mean of y for a given x.

Example: Suppose a fire safety inspector wants to relate the amount of fire damage in major residential fires to distance between the residence and nearest fire station. The study is to be conducted in a large suburb of a major city. A sample of 15 recent fires in this suburb is selected and given in the data file **FIREDAM**.

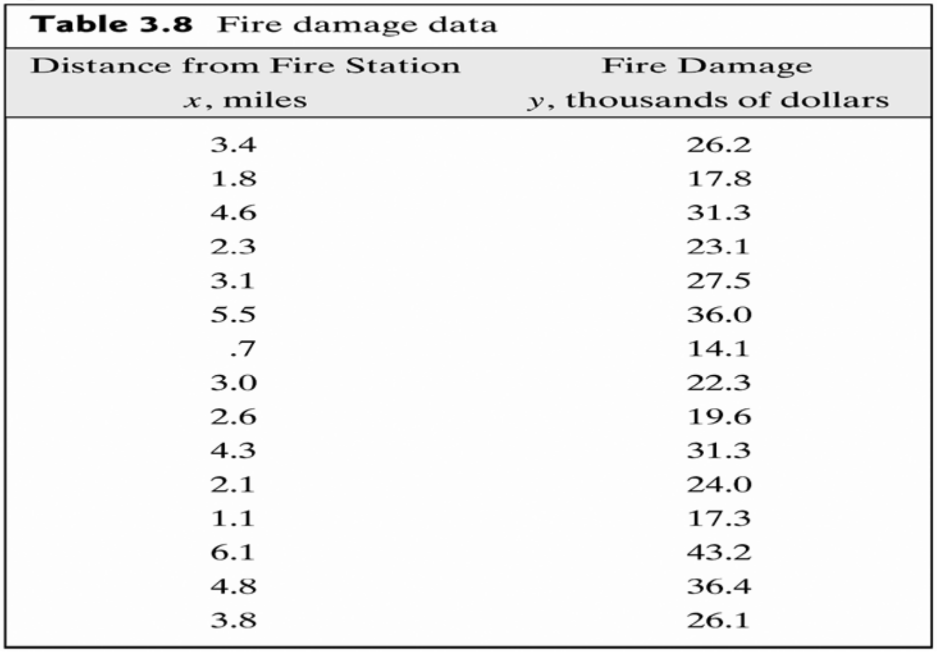


Table 3.8 Fire damage data

Distance from Fire Station x, miles	Fire Damage y, thousands of dollars
3.4	26.2
1.8	17.8
4.6	31.3
2.3	23.1
3.1	27.5
5.5	36.0
.7	14.1
3.0	22.3
2.6	19.6
4.3	31.3
2.1	24.0
1.1	17.3
6.1	43.2
4.8	36.4
3.8	26.1

Hypothesize the model.

x – distance from nearest fire station in miles.

y- fire damage in thousands of dollars.

Straight line probabilistic model: $y = \beta_0 + \beta_1 x + \varepsilon$

and the least squares regression model for estimating mean of y, E(y):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

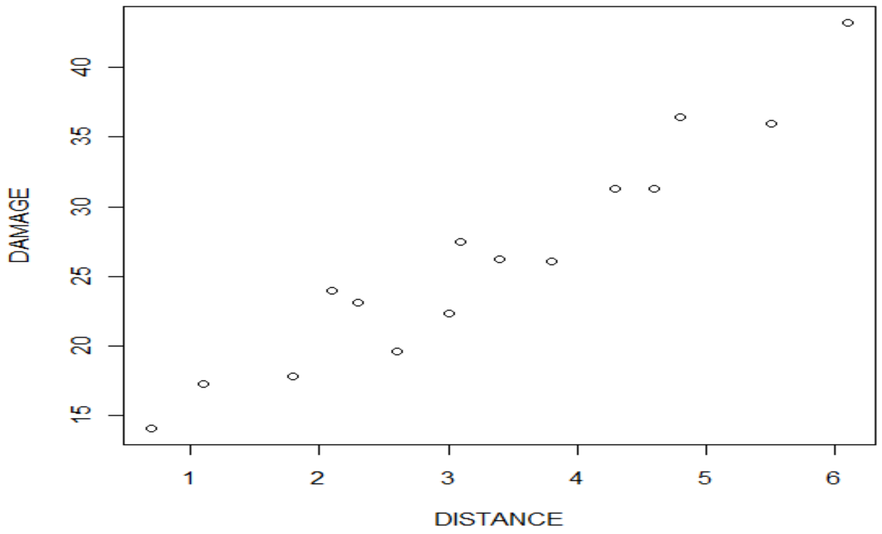
Fit the data to the least squares model using R: create scatterplot and regression outputs.

Open the file FIREDATA in R.

```
> firedata = read.table(file.choose( ), header = TRUE)

> attach(firedata)

> firedata

  DISTANCE DAMAGE
1      3.4   26.2
2      1.8   17.8
3      4.6   31.3
4      2.3   23.1
5      3.1   27.5
6      5.5   36.0
7      0.7   14.1
8      3.0   22.3
9      2.6   19.6
10     4.3   31.3
11     2.1   24.0
12     1.1   17.3
13     6.1   43.2
14     4.8   36.4
15     3.8   26.1
```

Scatterplot:

```
> plot(DISTANCE, DAMAGE)
```



Creating the model using R: lm(Y ~ X)

```
> model = lm(DAMAGE~DISTANCE)
> summary(model)
Call:
lm(formula = DAMAGE ~ DISTANCE)
Residuals:
   Min      1Q  Median      3Q     Max
-3.4682 -1.4705 -0.1311  1.7915  3.3915
Coefficients:
                Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)  10.2779     1.4203    7.237       6.59e-06 ***
DISTANCE      4.9193     0.3927   12.525      1.25e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.316 on 13 degrees of freedom
Multiple R-squared: 0.9235,     Adjusted R-squared: 0.9176
F-statistic: 156.9 on 1 and 13 DF,  p-value: 1.248e-08
```

Slope and intercept estimates are: $\hat{\beta}_0 = 10.2779, \quad \hat{\beta}_1 = 4.9193$

and the least squares line is: $\hat{y} = 10.2779 + 4.9193x$

Interpretation of regression coefficients:
The estimate of the slope coefficient, 4.92, implies that the estimated mean damage increase by $4,920 for each additional mile from the fire station. This interpretation is valid over the range of from 0.7 to 6.1 miles.
The estimated y intercept, 10.28, has the interpretation that a fire 0 miles from station has an estimated mean damage of $10,280.
Recall that y intercept is meaningfully interpreted only if x = 0 is in the range of in the sample values. Since x= 0 is outside the range of x (0.7, 6.1), this coefficient has no practical interpretation.

Checking the assumptions:
In this step, we specify the probability distribution of the random error term component ε.
Assumptions area:

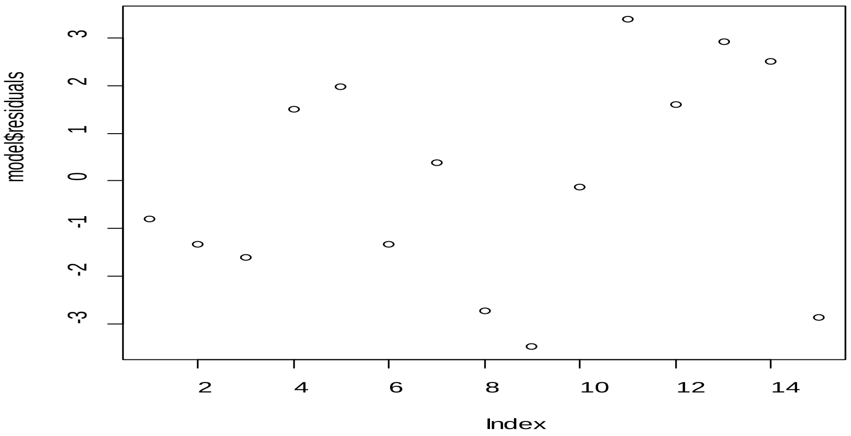
- E(ε) = 0.
- V(ε) = σ^2 for all values of x.
- The probability distribution of ε is normal.
- ε's are independent.

We can look at the residuals and get some idea about these assumptions.

We analyze the residuals to verify these assumptions. Residuals are accessed through *model$residuals* or *residuals(model)* in R.

```
> model$residuals
         1          2          3          4          5          6          7
-0.8036530 -1.3327239 -1.6068499  1.5076108  1.9721462 -1.3342475  0.3785399
         8          9         10         11         12         13         14
-2.7359207 -3.4681884 -0.1310507  3.3914769  1.6108077  2.9141540  2.5092840
        15
-2.8713853
```

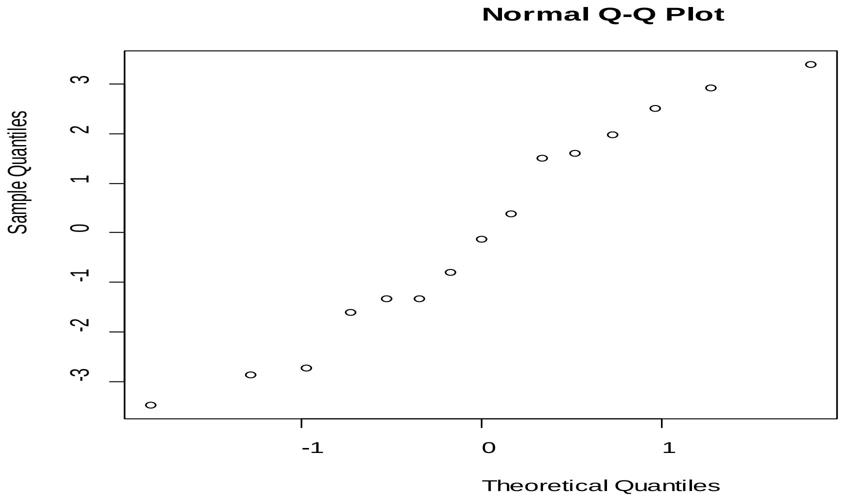
```
> plot(model$residuals)
```



```
> mean(model$residuals)
[1] 4.624484e-17
```

Mean of residuals is practically zero that implies that approximately E(ε) = 0.
To look at the normality of residuals we use qqnorm() command in R.

```
> qqnorm(model$residuals)
```



This plot will give some evidence that error terms are approximately normally distributed.

The estimate of σ^2, variance of the error term:

The estimated standard error of the regression model or the standard deviation (σ) of ε, denoted by **s is 2.316.** This is given in the model summary output under residual standard error.

Checking the utility of the model

Does x contribute information for the prediction of y using the straight line model?
Test of model utility: Test the hypothesis that the slope β_1 is 0. That is no linear association vs there is a positive linear association between distance (x) and damage (y). Use $\alpha = 0.05$.

H_0: $\beta_1 = 0$
H_a: $\beta_1 > 0$

From model summary output, note that *t* test statistic is 12.525 and two tailed test *p*-value is 1.25*e-08 = 0.
One tailed test *p*-value = two tailed *p*-value/2 = 1.25*e-08/2 = 0.
Since *p*-value < $\alpha = 0.05$, there is sufficient evidence to reject H_0 and conclude that distance between fire and station contributes information for the prediction of fire damage and mean fire damage increases as distance increases.

Confidence Interval for slope:

We gain additional information about the relationship by forming confidence intervals for β_1.
A 95% confident interval for β_1 can be found in following way:

```
> confint(model)
                2.5 %    97.5 %
(Intercept)  7.209605 13.346252
DISTANCE     4.070851  5.767811
```

We are 95% confident that the interval from $4,071 and $5,768 encloses the mean increase (β_1) in fire damage in per additional mile distance from station.

To obtain a confidence interval for a different confidence level (say 99%):

```
> confint(model, level = 0.99)
                 0.5 %    99.5 %
(Intercept) 5.999660 14.556197
DISTANCE     3.736266  6.102395
```

Numerical descriptive measures of model adequacy:
Coefficient of determination (r^2). From model summary output, we note that Multiple R-squared is 0.9235. This gives ***$r^2 = 0.9235$***.

This implies that 92% of the sample variation in fire damage (y) is explained by the distance x between fire and the station in the model.

The correlation coefficient, *r*:
The correlation coefficient measures the strength of the linear association between x and y.
$r = \text{sqrt}(r^2) = \text{sqrt}(0.9235) = 0.96$.

or use **cor(x,y)**.
This high correlation confirms our conclusion in hypothesis test of β_1.
It appears that fire damage and the distance from station are linearly correlated.
The results of the hypothesis test of β_1, the high value of r^2, and relatively small s value all point to a strong linear relationship between distance (x) and damage(y).

Predictions and prediction intervals.
Now we use the least squares model to predict the fire damage if a fire were to occur 3.5 miles from the nearest fire station (x = 3.5).
Use R to compute the predicted value and 95% prediction interval:

```
> predict(model, newdata = data.frame(DISTANCE = 3.5))
    1
27.49559
```

This means predicted fire damage, = $27,496.

The corresponding 95% prediction interval is:

```
> predict(model, newdata = data.frame(DISTANCE = 3.5), interval = "prediction" )
     fit        lwr       upr
1  27.49559  22.32394   32.66723
```

The prediction interval is (22.3239, 32.6673).
We predict (with 95% confidence) that the fire damage for a fire 3.5 miles from the nearest station will fall between $22,324 and $32,667.

Multiple Regression

In many situations, the dependent variable (y) could be related to **several** independent variables (x's). The dependent variable y is now written as a function of *k* independent variables, x_1, x_2, x_3,….x_k.

The multiple regression model is of the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \text{………} + \beta_k x_k + \varepsilon$$

where y is the response (dependent) variable that we want to predict. x_1, x_2, ….,x_k are independent information contributing variables, β_0, β_1, ……, β_k are unknown parameters, and ε is a random error component.

The estimated model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots\ldots. + \hat{\beta}_k x_k$

Example: A collector of antique grandfather clocks sold at auction believes that the price received for the clocks depend on both the age of the clock and the number of bidders at the auction. The hypothesized model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where y = Auction price (dollars)

x_1 = Age of the clock (years)

x_2 = Number of bidders.

A sample of 32 auction prices of grandfather clocks, alone with their age and the number of bidders are given in the data file **GFCLOCKSDATA.**

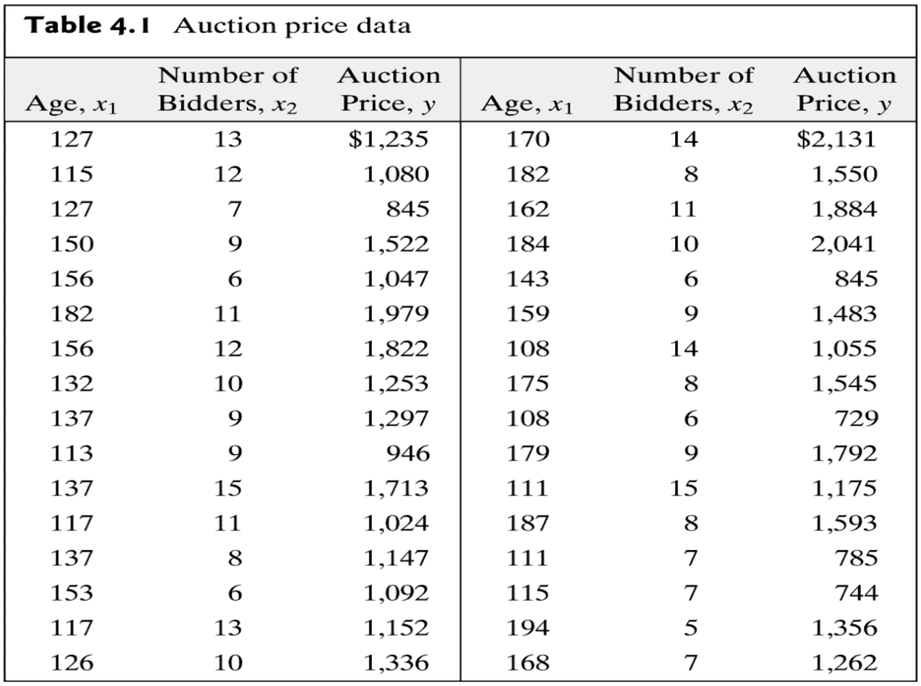


Table 4.1 Auction price data

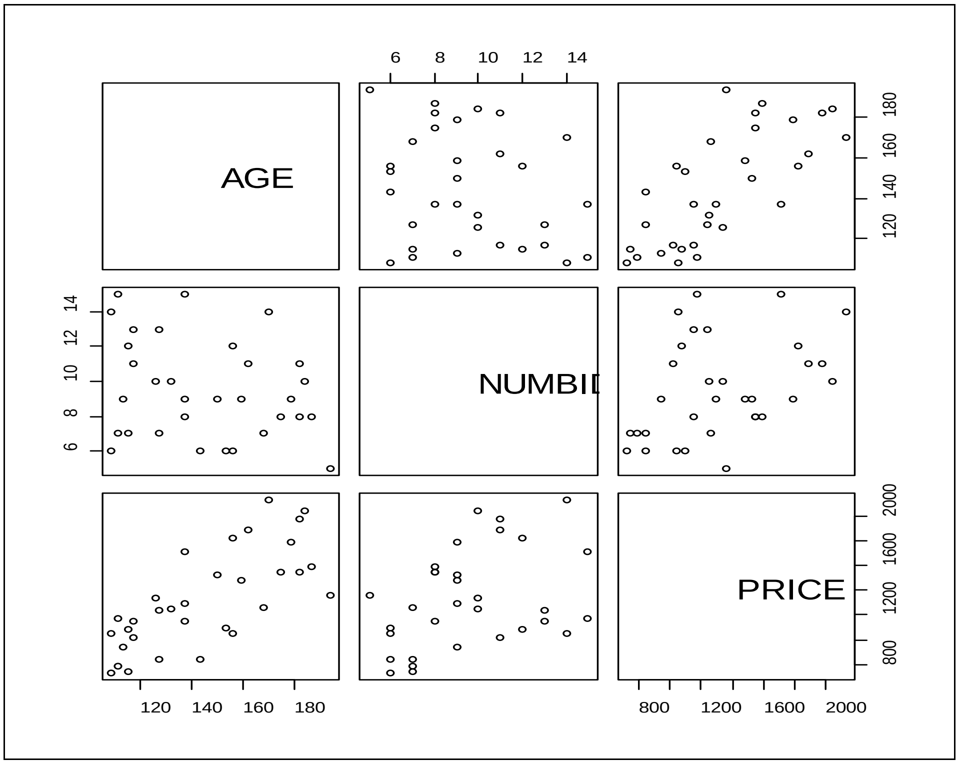
Age, x_1	Number of Bidders, x_2	Auction Price, y	Age, x_1	Number of Bidders, x_2	Auction Price, y
127	13	$1,235	170	14	$2,131
115	12	1,080	182	8	1,550
127	7	845	162	11	1,884
150	9	1,522	184	10	2,041
156	6	1,047	143	6	845
182	11	1,979	159	9	1,483
156	12	1,822	108	14	1,055
132	10	1,253	175	8	1,545
137	9	1,297	108	6	729
113	9	946	179	9	1,792
137	15	1,713	111	15	1,175
117	11	1,024	187	8	1,593
137	8	1,147	111	7	785
153	6	1,092	115	7	744
117	13	1,152	194	5	1,356
126	10	1,336	168	7	1,262

Open the GFCLOCKSDATA data file (text file) using R:

```
> clockdata = read.table(file.choose(),header = TRUE)
> attach(clockdata)
> clockdata
   AGE NUMBIDS PRICE
1  127      13  1235
2  115      12  1080
3  127       7   845
4  150       9  1522
5  156       6  1047
6  182      11  1979
```

Scatterplots:

```
> plot(clockdata)
```



To find the least squares estimates for multiple regression:

```
> model = lm(PRICE~AGE+NUMBIDS)
> summary(model)
Call:
```

```
lm(formula = PRICE ~ AGE + NUMBIDS)
Residuals:
   Min      1Q  Median     3Q    Max
-206.49 -117.34   16.66  102.55  213.50
Coefficients:
                  Estimate      Std. Error    t value      Pr(>|t|)
(Intercept)      -1338.9513   173.8095    -7.704      1.71e-08 ***
AGE               12.7406       0.9047      14.082     1.69e-14 ***
NUMBIDS           85.9530       8.7285        9.847      9.34e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 133.5 on 29 degrees of freedom
Multiple R-squared: 0.8923,     Adjusted R-squared: 0.8849
F-statistic: 120.2 on 2 and 29 DF,  p-value: 9.216e-15
```

From the regression output we can read the following least squares estimates:
$\hat{\beta}_0$ = -1,338.95, $\hat{\beta}_1$ = 12.74, and $\hat{\beta}_2$ = 85.95

and the least squares prediction equation: $\hat{y} = -1{,}338.95 + 12.74x_1 + 85.95x_2$

Interpretation of regression coefficients:
$\hat{\beta}_1$ = 12.74: We estimate the mean auction price E(y) of an antique clock to increase $12.74 for every 1-year increase in age (x_1) when the number of bidders (x_2) is held constant.

$\hat{\beta}_2$ = 85.95: We estimate the mean auction price E(y) of an antique clock to increase $85.95 for every one bidder increase in the number of bidders(x_2) when the age (x_1) is held constant.

$\hat{\beta}_0$= - 1,338.95 does not have a meaningful interpretation in this example. It says when $x_1 = x_2 = 0$, the estimated average auction price is $-1,339.95. It is not practical in this example.

Testing the Utility of a Model: The Analysis of Variance of F-Test

This overall (global) test of significance about all β parameters at the same time is called *F* test.
H_0: $\beta_1 = \beta_2 = 0$
H_a: At least one of the two coefficients is nonzero

From R output:
Test Statistic: *F* = 120.2, *p*-value = 0.000

Conclusion: **p-value < α = 0.05**, we reject the null hypothesis. **Data provides strong evidence that at least one of the model coefficients is nonzero. The overall model appears to be statistically useful for predicting auction prices.**

Inferences about individual β parameters.

The inferences about individual β parameters in a model are obtained using either a confidence interval or a hypothesis test.

- Test the hypothesis that mean auction price of a clock increases as the number of bidders increases when age is held constant, that is $\beta_2 > 0$. Use $\alpha = 0.05$.

- Form a 95% confidence interval for β_1 and interpret the results.

To test the hypothesis that the mean auction price of a clock increases as the number of bidders increases when age is held constant:

H_0: $\beta_2 = 0$

H_a: $\beta_2 > 0$

From regression output, we can directly read the test statistic value and the *p*-value for β_2.

t- test statistic = 9.847 = 85.9350/8.7285.

Two sided *p*-value = 9.34*e-11.

p-value of our test ($\beta_2 > 0$) = 9.34*e-11/2 = 0.00.

Since *p*-value = 0 < α = 0.05, we reject H_0.

Based on data, at 0.05 significance level, we have sufficient evidence to conclude that mean auction price of a clock increases as the number of bidders increases, when age is held constant

From R output directly we read the confidence interval as:

```
> confint(model)
                  2.5 %          97.5 %
(Intercept)   -1694.43162   -983.47106
AGE            10.89017      14.59098
NUMBIDS       68.10115       103.80482
```

The confidence interval for β_1 is (10.89, 14.59).

Interpretation of the 95% confidence interval:

We are 95% confident that β_1 falls between 10.89 and 14.59. We conclude that price increases between \$10.89 and \$14.59 for every 1-year increase in age, holding number of bidders (x_2) constant.

To compute confidence interval for a specific variable (say NUMBIDS) use

```
> confint(model, parm = 'NUMBIDS')
```

Prediction

Estimate the average auction price for all 150 year old clocks sold at an auction with 10 bidders using 95% confidence interval. Interpret the result.

Here key words *average* and *for all* imply we want to estimate the mean y, E(y). We need 95% confidence Interval for E(y) when x_1 = 150 years and x_2 = 10 bidders.

```
> predict(model, data.frame(AGE = 150, NUMBIDS = 10), interval = "confidence")
     fit        lwr        upr
1  1431.665  1381.398   1481.931
```

From R output, we can read the estimated average price is **\$1431.67** and the confidence interval is **(1381.398,   1481.931)**.
We are 95% confident that the mean auction price for all 150 year old clocks sold at an auction with 10 bidders lies between \$1,381.40 and \$1,481.90.

Predict the auction price for a single 150 year old clock at an auction with 10 bidders using a 95% prediction interval. Interpret the result.

The key words *predict* and *for a single* imply that we want 95% prediction interval for y when x_1 = 150 years and x_2 = 10 bidders.

```
> predict(model, data.frame(AGE = 150, NUMBIDS = 10), interval = "prediction")
     fit          lwr         upr
1  1431.665    1154.069    1709.26
```

From R output, we can read the predicted auction price is **\$1431.67** and the prediction interval is **(1154.069, 1709.26)**.

We are 95% confident that the auction price for a single 150 year old clocks sold at an auction with 10 bidders falls between \$1,154.10 and \$1,709.30.