# Multiple Regression

Professor: Hammou El Barmi
Baruch College

- We consider the problem of regression when study variable depends on more than one explanatory or independent variables, called as multiple linear regression model.
- This model generalizes the simple linear regression in two ways.
- It allows the mean function of the response to depend on more than one explanatory variables and to have shapes other than straight lines, although it does not allow for arbitrary shapes

- Let y denotes the dependent (or study) variable whose mean is linearly related to k independent (or explanatory) variables $x_1, x_2, \ldots, x_k$, that

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

- $\beta_0$ is the mean of y when all the Xs are equal to zero
- $\beta_i$ is the change in the mean of y when we increase $x_i$ by one while holding all the other xs fixed

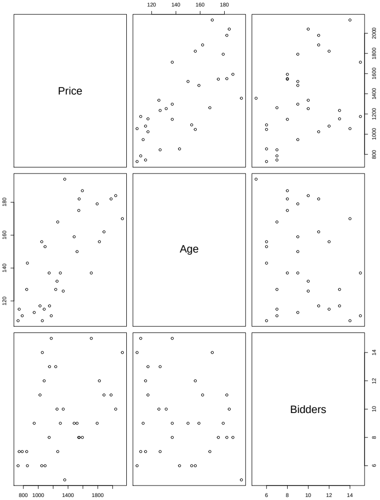
- The data give the selling price at auction of 32 antique grandfather clocks. Also recorded is the age of the clock and the number of people who made a bid.
- The variables are
  1. Age : Age of the clock (years)
  2. Bidders: Number of individuals participating in the bidding
  3. Price: Selling price (pounds sterling)

# Introduction

| | Age | Bidders | 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 |
| 7 | 156 | 12 | 1822 |
| 8 | 132 | 10 | 1253 |
| 9 | 137 | 9 | 1297 |
| 10 | 113 | 9 | 946 |
| 11 | 137 | 15 | 1713 |
| 12 | 117 | 11 | 1024 |
| 13 | 137 | 8 | 1147 |
| 14 | 153 | 6 | 1092 |
| 15 | 117 | 13 | 1152 |
| 16 | 126 | 10 | 1336 |
| 17 | 170 | 14 | 2131 |
| 18 | 182 | 8 | 1550 |
| 19 | 162 | 11 | 1884 |
| 20 | 184 | 10 | 2041 |
| 21 | 143 | 6 | 854 |
| 22 | 159 | 9 | 1483 |
| 23 | 108 | 14 | 1055 |
| 24 | 175 | 8 | 1545 |
| 25 | [illegible] | 6 | [illegible] |

First we plot the data

```
> pairs(~Price+Age+Bidders)
```



The estimates, $b_0, b_1, \ldots, b_k$. of $\beta_1, \beta_2, \ldots, \beta_k$ are the values of $\beta_1, \beta_2, \ldots, \beta_k$ that minimize

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_k x_{ki})^2$$

and the regression function is

$$\hat{y} = b_0 + b_1 x_1 + \ldots + b_p x_p$$

```
> attach(data)
> fit<-lm(Price~ Age+Bidders)
> fit

Call:
lm(formula = Price ~ Age + Bidders)

Coefficients:
(Intercept)          Age      Bidders
   -1336.72        12.74        85.82
```

The regression equation is

$$\widehat{Price} = -1336.72 + 12.74\text{Age} + 85.82\text{Bidders}$$

- A $100(1-\alpha)\%$ confidence interval for $\beta_i$ is

$$b_i \pm t_{[\alpha/2]}^{(n-p-1)} s_{b_i}$$

The interpretation of this confidence interval is: We are $100(1-\alpha)\%$ confident that when we increase $x_i$ by one unit while holding all the other xs fixed, on average, y changes by an amount in this interval.

# Example

```
> confint(fit)
                   2.5 %     97.5 %
(Intercept) -1691.27514 -982.16896
Age            10.89062   14.58177
Bidders        68.00986  103.62040
```

We are 95% that when we increase age by one year while holding the number of bidders fixed, on average the price goes by an amount between 10.89 and 12.58 pounds sterling.

- To test $H_0 : \beta_i = \beta_{i0}$ against $H_a : \beta_i \neq \beta_{i0}$, the test statistic is

$$t = \frac{b_i - \beta_{i0}}{s_{b_i}}$$

and we reject $H_0$ if

$$|t| > t_{[\alpha/2]}^{(n-p-1)} \quad \text{or if} \quad p - value < \alpha$$

- for the case where $\beta_{i0} = 0$, the p-values in printed in output

# Example

```
 > summary(fit)
lm(formula = Price ~ Age + Bidders)

Coefficients:
                Estimate         Std. Error    t value      Pr(>|t|)
(Intercept)    -1336.7221         173.3561     -7.711      1.67e-08 ***
Age               12.7362           0.9024     14.114      1.60e-14 ***
Bidders           85.8151           8.7058      9.857      9.14e-11 ***

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
```

p-values very small we reject $H_0 : \beta_i = 0$ against $H_a : \beta_i \neq 0$

# Analysis of Variance Table

- The ANOVA table is given by

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Model | k | SSR | MSR=SSR/p | MSR/MSE |
| Error | n-k-1 | SSE | MSE=SSE/(n-p-1) | |
| Total | n-1 | SST | | |

- The coefficient of determination is

$$r^2 = \frac{SSR}{SST}$$

- Adjusted

$$r^2_{adj} = 1 - \frac{n-1}{n-p-1}\frac{SSE}{SST}$$

  can be used for model selection
- MSE is an estimate of $\sigma^2$

- $y =$ volume of sales in July of some electronic store (in thousands of dollars)
- $x =$ number of households in the location
- Location of the store $= \begin{cases} \text{Mall} \\ \text{Downtown} \\ \text{Street} \end{cases}$

| number of household | location | sales |
|---|---|---|
| 161 | street | 157.27 |
| 99 | street | 93.28 |
| 135 | street | 136.81 |
| 120 | street | 123.79 |
| 164 | street | 153.51 |
| 221 | mall | 241.74 |
| 179 | mall | 201.54 |
| 204 | mall | 206.71 |
| 214 | mall | 229.78 |
| 101 | mall | 135.22 |
| 231 | downtown | 224.71 |
| 206 | downtown | 195.29 |
| 248 | downtown | 242.16 |
| 107 | downtown | 115.21 |
| 205 | downtown | 197.82 |

```
>fit<-lm(sales~nhousehold+factor(location))
> fit
Call:
lm(formula = sales ~ nhousehold + factor(location))

Coefficients:
           (Intercept)              nhousehold    factor(location)mall
               21.8415                  0.8686                 21.5100
factor(location)street
               -6.8638
```

```
> summary(fit)

Call:
lm(formula = sales ~ nhousehold + factor(location))

Residuals:
    Min      1Q  Median      3Q     Max
-13.834  -2.999   2.225   4.357   6.431

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            21.84147    8.55848   2.552 0.026898 *
nhousehold              0.86859    0.04049  21.452 2.52e-10 ***
factor(location)mall   21.50998    4.06509   5.291 0.000256 ***
factor(location)street -6.86378    4.77048  -1.439 0.178047

Residual standard error: 6.349 on 11 degrees of freedom
Multiple R-squared:  0.9868,Adjusted R-squared:  0.9833
F-statistic: 275.1 on 3 and 11 DF,  p-value: 1.268e-10
```

# Example

```
> confint(fit)
                              2.5 %     97.5 %
(Intercept)               3.0043933 40.6785468
nhousehold                0.7794707  0.9577061
factor(location)mall     12.5627722 30.4571864
factor(location)street  -17.3635248  3.6359712
```

# Partial F test

- Suppose we want to test $H_0 : \beta_{g+1} = \beta_{g+2} = \ldots = \beta_k = 0, g < k$ against $H_a$ : at least one of of $\beta_{g+1}, \beta_{g+2}, \ldots, \beta_k$ is not equal to zero.
- In this case we have two models:
  - a reduced model(the model in which $\beta_{g+1} = \beta_{g+2} = \ldots = \beta_k = 0$) and
  - a full model in which we have all the $\beta$s
- The test statistic is given by

$$F = \frac{(SSE_R - SSE_C)/(df_R - df_C)}{SSE_C/df_C)} = \frac{(SSE_R - SSE_R)/(k-g)}{SSE_C/(n-k-1)}$$

and we reject $H_0$ if

$$F > F(\alpha, k-g, n-k-1)$$

or if $p - value < \alpha$.

Full Model

```
> fitC<-lm(sales~nhousehold+factor(location))
> anova(fitC)
Analysis of Variance Table

Response: sales
                 Df  Sum Sq Mean Sq F value    Pr(>F)
nhousehold        1 31244.4 31244.4 775.006 1.502e-11 ***
factor(location)  2  2024.3  1012.2  25.107 7.944e-05 ***
Residuals        11   443.5    40.3
```

# Example

Reduced Model

```
> fitR<-lm(sales~nhousehold)
> anova(fitR)
Analysis of Variance Table

Response: sales
           Df  Sum Sq Mean Sq F value    Pr(>F)
nhousehold  1 31244.4 31244.4  164.59 9.339e-09 ***
Residuals  13  2467.8   189.8
```

# Example

Partial F test

```
> anova(fitR,fitC)
Analysis of Variance Table

Model 1: sales ~ nhousehold
Model 2: sales ~ nhousehold + factor(location)
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     13 2467.81
2     11  443.47  2    2024.3 25.107 7.944e-05 ***
```