STAT 500

Multiple Linear Regression Models
Grandfather Clock Example

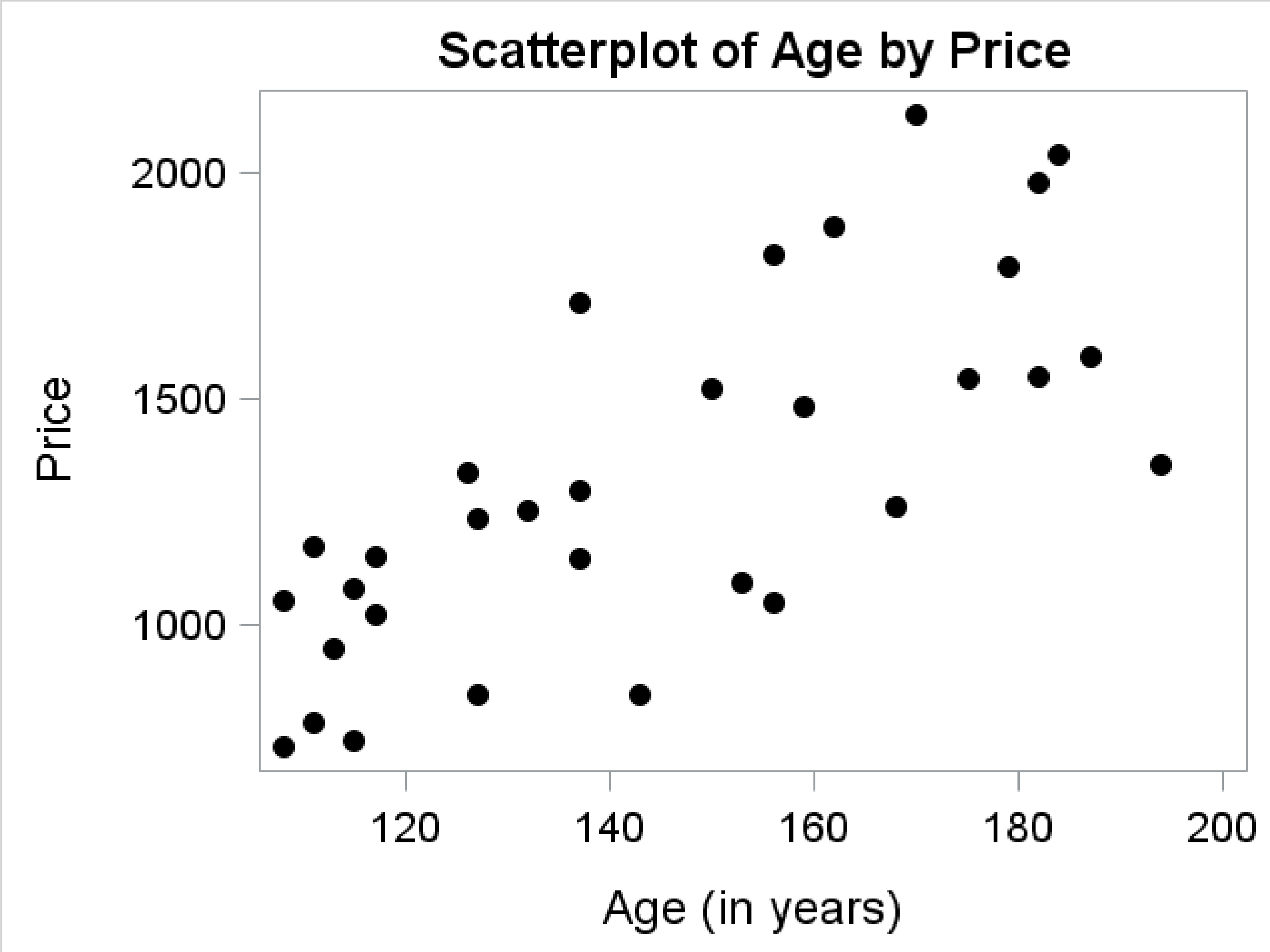
Grandfather Clock Example

- There were 32 antique (>100 years old) grandfather clocks sold at auction.
- Response variable: price at auction
- Two explanatory variables: age (in years) and the number of bidders

Grandfather Clock Example

Data:

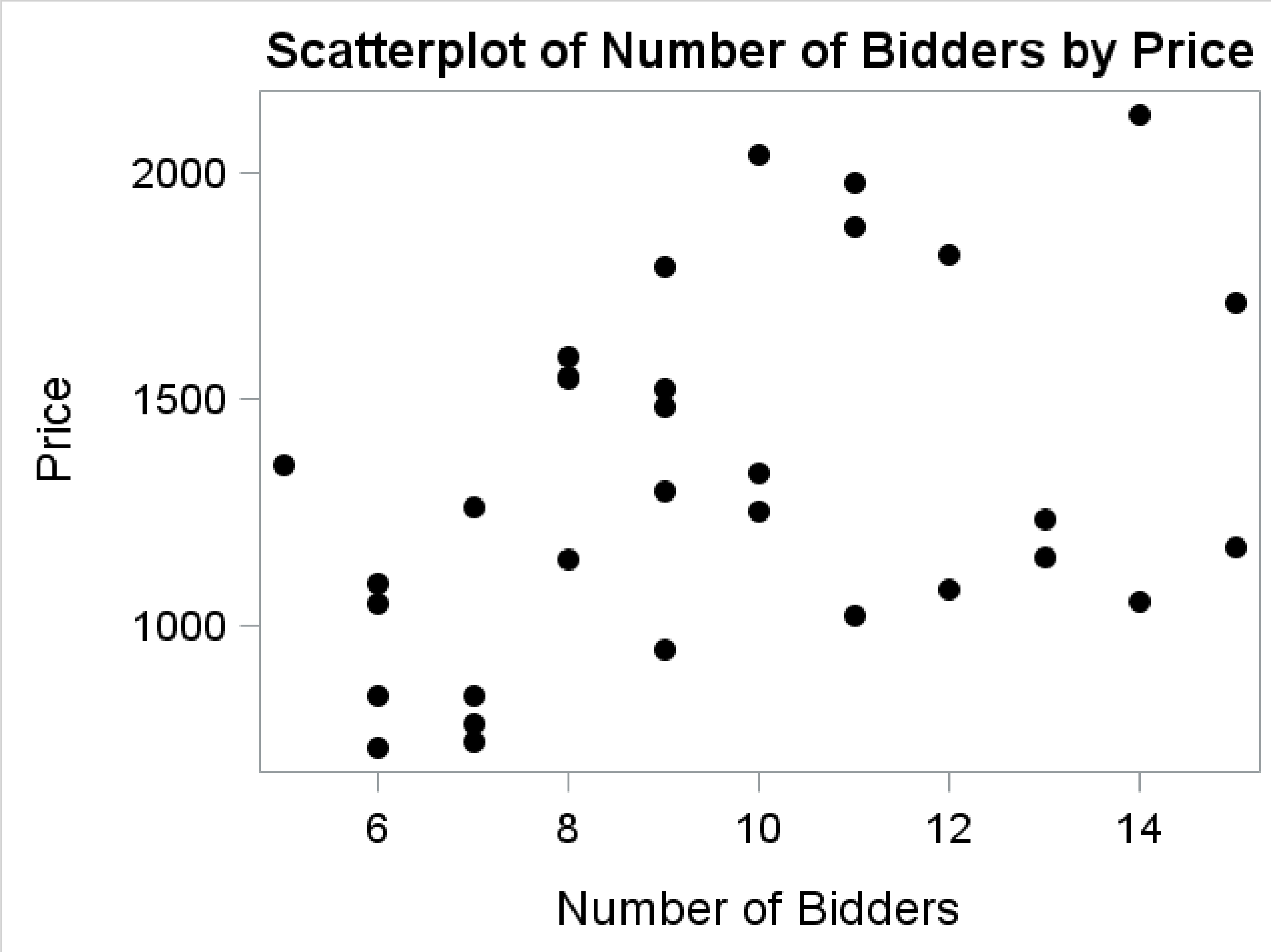
Price	Age (years)	NumBidder
Y	X_1	X_2
1235	127	13
1080	115	12
845	127	7
.	.	.
.	.	.
.	.	.
1262	168	7


Scatterplot of Age by Price
Price
2000
1500
1000
120
140
160
180
200
Age (in years)

SLR of Age on Price

$$\hat{Y}_i = -192.05 + 10.48 Age$$

- There is a significant linear relationship between age and price at auction (F-test p-value <0.0001).
- Each additional year of age is associated with a mean increase in price of 10.48 dollars.
- $R^2 = 53.24\%$ of the variation in price can be explained by the linear regression model with age.


Scatterplot of Number of Bidders by Price
Price
2000
1500
1000
6
8
10
12
14
Number of Bidders

SLR of Number of Bidders on Price

$$\hat{Y}_i = 804.91 + 54.76 NumBid$$

- There is a significant linear relationship between number of bidders and price at auction (F-test p-value=0.0252).
- Each additional additional bidder is associated with a mean increase in price of 54.76 dollars.
- $R^2 = 15.62\%$ of the variation in price can be explained by the linear regression model with number of bidders.

MLR: Grandfather Clock Example

With both explanatory variables in the MLR, the dimension of the design matrix $\boldsymbol{X}$ is $\mathbf{32} \times \mathbf{3}$.

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \mathbf{127} & \mathbf{13} \\ \mathbf{1} & \mathbf{115} & \mathbf{12} \\ \mathbf{1} & \mathbf{127} & \mathbf{7} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \mathbf{1} & \mathbf{168} & \mathbf{7} \end{bmatrix}$$

MLR: Grandfather Clock Example

With both explanatory variables in the MLR, the dimension of the design matrix $\boldsymbol{X}$ is $\mathbf{32} \times \mathbf{3}$.

$$\mathrm{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{Y} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \mathbf{-1338.95} \\ \mathbf{12.74} \\ \mathbf{85.95} \end{bmatrix}$$

Estimated Regression Model:
$\hat{Y}_i = \mathbf{-1338.95 + 12.74}\boldsymbol{Age} + \mathbf{85.95}\boldsymbol{NumBid}$

Grandfather Clock Example
Different Regression Analysis

$$\hat{Y}_i = -192.05 + 10.48 Age$$

$$\hat{Y}_i = 804.91 + 54.76 NumBid$$

$$\hat{Y}_i = -1338.95 + 12.74 Age + 85.95 NumBid$$

MLR: Grandfather Clock Example

$$MS_{error}(X^TX)^{-1} = 17818 \begin{bmatrix} 1.695 & -0.00773 & -0.057 \\ -0.00773 & 0.0000459 & 0.0001 \\ -0.057 & 0.0001 & 0.00428 \end{bmatrix}$$

$$= \begin{bmatrix} 30209 & -137.74 & -1016.58 \\ -137.74 & 0.8185 & 2.004 \\ -1016.58 & 2.004 & 76.186 \end{bmatrix}$$

Then

$$S_{b_0} = \sqrt{30209} = 173.81$$

$$S_{b_1} = \sqrt{0.8185} = 0.9047$$

$$S_{b_2} = \sqrt{76.186} = 8.7285$$

MLR: Grandfather Clock Example
Inference for β_1

β_1 represents the change in auction price when age is increased 1 year while the number of bidders is held constant.

A $(1-\alpha) \times 100\%$ confidence interval for β_1:

$$b_1 \pm t_{df_{error}, 1-\alpha/2}\ S_{b_1}$$

A 95% confidence interval is

$$12.74 \pm (2.045)(0.9047) \quad \Rightarrow \quad (10.89, 14.59)$$

MLR: Grandfather Clock Example
Inference for β_1

Hypothesis test:

$H_o : \beta_1 = 0$ or $E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_2 x_2$

versus

$H_a : \beta_1 \neq 0$ or $E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

$$t = \frac{b_1 - 0}{S_{b_1}} = \frac{12.74}{0.9047} = 14.08$$

on 29 df with p-value < 0.0001.

SAS code for MLR analysis of Grandfather Clock Example

```
/* compute the MLR with price and age and number of bidders */

proc reg data=set1;
  model price = age numbid;
  run;
```

MLR analysis for clock example

The REG Procedure
Model: MODEL1
Dependent Variable: price

Number of Observations Read	32
Number of Observations Used	32

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	4283063	2141531	120.19	<.0001
Error	29	516727	17818		
Corrected Total	31	4799790			

Root MSE	133.48467	**R-Square**	0.8923
Dependent Mean	1326.87500	**Adj R-Sq**	0.8849
Coeff Var	10.06008		

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	-1338.95134	173.80947	-7.70	<.0001	-1694.43162	-983.47106
age	1	12.74057	0.90474	14.08	<.0001	10.89017	14.59098
numbid	1	85.95298	8.72852	9.85	<.0001	68.10115	103.80482

MLR: Grandfather Clock Example

$$\hat{Y}_i = -1338.95 + 12.74Age + 85.95NumBid$$

- Model is statistically significant in explaining Price with $F = 120.9$ and p-value < 0.0001.

- $R^2 = 89.23\%$ of the variation in price can be explained by the multiple linear regression model with both age and number of bidders.

- Given number of bidders in the model, age is statistically significant with $t = 14.08$ and p-value < 0.0001.

- Given age in the model, number of bidders is statistically significant with $t = 9.85$ and p-value < 0.0001.

MLR: Grandfather Clock Example

$\hat{Y}_i = -1338.95 + 12.74 Age + 85.95 NumBid$

- This analysis indicates that changes in either Age (X_1) or Number of Bidders (X_2) affect the auction price.
  - Holding the number of bidders constant, a 1 year increase in age increases price by 12.74 dollars.
  - Holding age constant, a 1 additional bidder increase increases auction price by 85.95 dollars.
  - What if you change both age and number of bidders?
  - How should the intercept be interpreted?
- The significance of each coefficient does not necessarily imply that the model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$ is correct

MLR: Grandfather Clock Example

Estimate the mean price of a clock when

$\boldsymbol{X_1}$ = Age = 150 years
$\boldsymbol{X_2}$ = NumBid = 10

In this case

$$\boldsymbol{x^T} = (\mathbf{1} \quad \mathbf{150} \quad \mathbf{10})$$

The least squares estimate of the mean yield under these conditions is

$$\hat{\boldsymbol{Y}} = \boldsymbol{x^T b} = (\mathbf{1} \quad \mathbf{150} \quad \mathbf{10}) \begin{bmatrix} \mathbf{-1338.95} \\ \mathbf{12.74} \\ \mathbf{85.95} \end{bmatrix} = \mathbf{1431.55}$$

MLR: Grandfather Clock Example

Compute the standard error of the estimated mean

$$
\begin{aligned}
\boldsymbol{S^2_{\hat{Y}}} &= \boldsymbol{MS_{error} x^T (X^T X)^{-1} x} \\
&= \boldsymbol{x^T \left[MS_{error} (X^T X)^{-1} \right] x} \\
&= \mathbf{(1\ 150\ 10)} \begin{bmatrix} \mathbf{30209} & \mathbf{-137.74} & \mathbf{-1016.58} \\ \mathbf{-137.74} & \mathbf{0.8185} & \mathbf{2.004} \\ \mathbf{-1016.58} & \mathbf{2.004} & \mathbf{76.186} \end{bmatrix} \begin{bmatrix} \mathbf{1} \\ \mathbf{150} \\ \mathbf{10} \end{bmatrix} \\
&= \mathbf{604.04}
\end{aligned}
$$

The standard error is $\boldsymbol{S_{\hat{Y}}} = \sqrt{\mathbf{604.04}} = \mathbf{24.58}$

MLR: Grandfather Clock Example

A $(1-\alpha) \times 100\%$ confidence interval for the mean price under the conditions specified by $x = (1\ 150\ 10)$ is

$$\hat{Y} \pm t_{df_{error},1-\alpha/2}\ S_{\hat{Y}}$$

A 95% confidence interval is

$$1431.55 \pm (2.045)(24.58) \quad \Rightarrow \quad (1381.28, 1481.82)$$

MLR: Grandfather Clock Example

Predict price of a clock to be sold at a future auction when

$\boldsymbol{X}_1$ = Age = 150 years
$\boldsymbol{X}_2$ = NumBid = 10

In this case

$$\boldsymbol{x}^T = (\mathbf{1} \quad \mathbf{150} \quad \mathbf{10})$$

The predicted value of the random error is zero and the predicted price under the conditions specified by $\boldsymbol{x}$ is

$$\hat{\boldsymbol{Y}} = \boldsymbol{x}^T\boldsymbol{b} + \mathbf{0} = (\mathbf{1} \quad \mathbf{150} \quad \mathbf{10})\begin{bmatrix} \mathbf{-1338.95} \\ \mathbf{12.74} \\ \mathbf{85.95} \end{bmatrix} = \mathbf{1431.55}$$

MLR: Grandfather Clock Example

Compute the standard error of the predicted price

$$\begin{aligned} S^2_{pred} &= MS_{error} + MS_{error} x^T (X^T X)^{-1} x \\ &= MS_{error} + S^2_{\hat{Y}} \\ &= 17818 + 604.04 \\ &= 18422.04 \end{aligned}$$

The standard error is

$$S_{pred} = \sqrt{18422.04} = 135.73$$

MLR: Grandfather Clock Example

$(1-\alpha) \times 100\%$ prediction interval for the price under the conditions specified by $x = (1\ 150\ 10)$ is

$$\hat{Y} \pm t_{df_{error}, 1-\alpha/2}\ S_{pred}$$

A 95% prediction interval is

$$1431.55 \pm (2.045)(135.73) \quad \Rightarrow \quad (1153.98, 1709.12)$$

MLR: Grandfather Clock Example

$(1-\alpha)\times 100\%$ simultaneous prediction region for the auction price

$$\hat{Y} \pm \sqrt{(k+1)F_{(k+1,df_{error}),1-\alpha}}\; S_{pred}$$

Simultaneous 95% prediction intervals are

$$\hat{Y} \pm \sqrt{3F_{(3,29),0.95}} S_{pred}$$

$\Rightarrow$

$$\hat{Y} \pm \sqrt{(3)(2.934)} S_{pred}$$

$\Rightarrow$

$$\hat{Y} \pm (2.9668) S_{pred}$$

MLR: Grandfather Clock Example
Effect Test for β_2 (NumBid)

Source	d.f.	SS	MS	F	p-val
Model with Age	$\mathbf{1}$	$\mathbf{2555224}$	$\mathbf{2555224}$	34.15	$< \mathbf{0.0001}$
Error	$\mathbf{30}$	$\mathbf{2244565}$	$\mathbf{74819}$		
corrected total	$\mathbf{31}$	$\mathbf{4799790}$			

Source	d.f.	SS	MS	F	p-val
Model with Age and NumbBid	$\mathbf{2}$	$\mathbf{4283063}$	$\mathbf{2141531}$	120.19	$< \mathbf{0.0001}$
Error	$\mathbf{29}$	$\mathbf{516727}$	$\mathbf{17818}$		
corrected total	$\mathbf{31}$	$\mathbf{4799790}$			

MLR: Grandfather Clock Example Effect Test for β_2 (NumBid)

- Adding Number of Bidders to the SLR model with Age reduces the SS_{Error} for the model.
- For SLR with Age, $SS_{Error} = 2244565$.
- For MLR with Age and NumBid, $SS_{Error} = 516727$.
- Difference = 2244565 - 516727 = 1727838.

MLR: Grandfather Clock Example
Effect Test for β_2 (NumBid)

$$
\begin{aligned}
\boldsymbol{F} &= \frac{(\boldsymbol{SSE}_{\text{r.model}} - \boldsymbol{SSE}_{\text{f.model}})/\boldsymbol{m}}{\boldsymbol{MSE}_{\text{f.model}}} \\
&= \frac{(\boldsymbol{SSE}_{\text{SLRage}} - \boldsymbol{SSE}_{\text{MLR}})/\boldsymbol{m}}{\boldsymbol{MSE}_{\text{MLR}}} \\
&= \frac{\mathbf{1727838/1}}{\mathbf{17818}} \\
&= \mathbf{96.97} \\
&= \mathbf{9.85^2}
\end{aligned}
$$

MLR: Grandfather Clock Example

$$\hat{Y}_i = -1338.95 + 12.74 Age + 85.95 NumBid$$

- This model is additive.
  - The effect of age on the price of a clock is the same for each number of bidders.
  - The effect of number of bidders on the price of a clock is the same for every value of age.

MLR with Interaction: Grandfather Clock Example

- Allows for the effect of one explanatory variable on the response variable to be different depending on the value of another explanatory variable.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$$

- Effect on Response Variable
  - The effect of increasing x_{i1} by 1 is $\beta_1 + \beta_3 x_{i2}$.
  - The effect of increasing x_{i2} by 1 is $\beta_2 + \beta_3 x_{i1}$.

SAS code for MLR analysis of Grandfather Clock Example

```
/* compute the MLR with price and age and number of bidders */
data set1;
  infile 'clocks.csv' dlm = ',' firstobs = 2;
  input age numbid price;
  agexnumbid=age*numbid;
  cagexcnumbid=(age-144.9375)*(numbid-9.53125);
run;

proc reg data=set1 ;
  model price = age numbid agexnumbid ;
  run;
```

The REG Procedure

Dependent Variable: price

Number of Observations Read	32
Number of Observations Used	32

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	4578427	1526142	193.04	<.0001
Error	28	221362	7905.79047		
Corrected Total	31	4799790			

Root MSE	88.91451	**R-Square**	0.9539
Dependent Mean	1326.87500	**Adj R-Sq**	0.9489
Coeff Var	6.70105		

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	320.45799	295.14128	1.09	0.2868	-284.11152	925.02751
age	1	0.87814	2.03216	0.43	0.6690	-3.28454	5.04083
numbid	1	-93.26482	29.89162	-3.12	0.0042	-154.49502	-32.03462
agexnumbid	1	1.29785	0.21233	6.11	<.0001	0.86290	1.73279

Test for Significance of Interaction Term

- $\boldsymbol{T}$-test: $\boldsymbol{t = 6.11}$, p-value $\boldsymbol{< 0.0001}$.

- Effect test:
  - For MLR with Age and NumBid, $\boldsymbol{SS_{Error}}$ = 516727.
  - For MLR with Age and NumBid and interaction, $\boldsymbol{SS_{Error}}$ = 221362.

Test for Significance of Interaction Term

- The partial $\boldsymbol{F}$-test:

$$
\begin{aligned}
\boldsymbol{F} &= \frac{(\boldsymbol{SSE}_{\text{r.model}} - \boldsymbol{SSE}_{\text{f.model}})/\boldsymbol{m}}{\boldsymbol{MSE}_{\text{f.model}}} \\
&= \frac{(\mathbf{516727} - \mathbf{221362})/\mathbf{1}}{\mathbf{7905.79}} \\
&= \mathbf{37.36} \\
&= \mathbf{6.11}^{\mathbf{2}}
\end{aligned}
$$

- Interaction Term is statistically significant in model.

Tests for Component Explanatory Variables

- Do not perform significance tests for component explanatory variables when corresponding interaction terms exist in the model.

- These tests no longer have any meaning.
  - Test for significance of variable given the other variables in the model.
  - The component variable is already in the model through its presence in the interaction term.
  - Cannot separate significance of component variable from its interaction term.

Alternative Parameterization of Interaction Term

$$\begin{aligned}
\boldsymbol{Y_i} &= \boldsymbol{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) + \epsilon_i} \\
&= \boldsymbol{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} - \beta_3 x_{i1} \bar{x}_2 - \beta_3 x_{i2} \bar{x}_1 + \beta_3 \bar{x}_1 \bar{x}_2 + \epsilon_i} \\
&= \boldsymbol{\beta_0 + \beta_3 \bar{x}_1 \bar{x}_2 + (\beta_1 - \beta_3 \bar{x}_2) x_{i1} + (\beta_2 - \beta_3 \bar{x}_1) x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i}
\end{aligned}$$

SAS code for MLR analysis of Grandfather Clock Example

```
/* compute the MLR with price and age and number of bidders */

proc reg data=set1;
  model price = age numbid cagexcnumbid /  p clm cli clb;
  run;
```

The REG Procedure

Dependent Variable: price

Number of Observations Read	32
Number of Observations Used	32

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	4578427	1526142	193.04	<.0001
Error	28	221362	7905.79047		
Corrected Total	31	4799790			

Root MSE	88.91451	**R-Square**	0.9539
Dependent Mean	1326.87500	**Adj R-Sq**	0.9489
Coeff Var	6.70105		

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	-1472.43236	117.81657	-12.50	<.0001	-1713.76866	-1231.09606
age	1	13.24824	0.60835	21.78	<.0001	12.00209	14.49438
numbid	1	94.84170	5.99320	15.82	<.0001	82.56519	107.11822
cagexcnumbid	1	1.29785	0.21233	6.11	<.0001	0.86290	1.73279

Alternative Parameterization of Interaction Term

- Estimated coefficient for interaction term does not change.
- Estimated coefficients for intercept and component explanatory variables change.
  - Different std. errors, t-test statistics and p-values.
- Correlation between component explanatory variables and interaction term is reduced.