Chapter 15: Multiple Regression and Model Building

STAT 2601 – Business Statistics

Esam Mahdi

School of Mathematics and Statistics
Carleton University

January 24, 2026

Multiple Linear Regression: Overview

Extension of Simple Linear Regression:

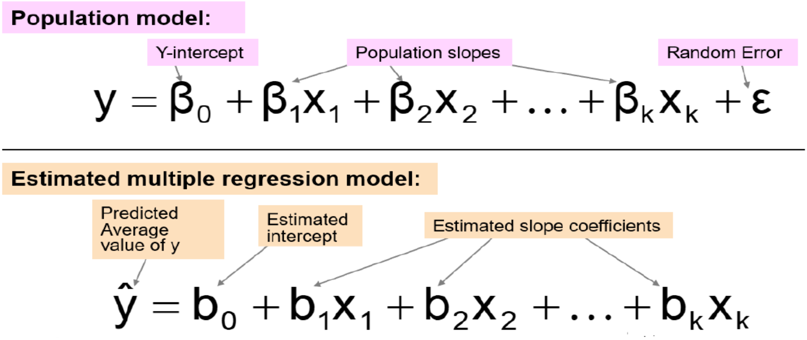
- One response variable y
- Multiple predictor variables $x_1, x_2, \ldots, x_k$
- Applications: Economics, biology, engineering, social sciences
- Goal: Model relationship and make predictions

Example: Modeling house prices based on:

- x_1: Square footage
- x_2: Number of bedrooms
- x_3: Age of house
- x_4: Location rating

Multiple Linear Regression Model

Idea: Examine the linear relationship between 1 dependent (y) and 2 or more independent variables (x_i).



Population model:

Y-intercept

Population slopes

Random Error

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

Estimated multiple regression model:

Predicted Average value of y

Estimated intercept

Estimated slope coefficients

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

Least Squares Estimates and Prediction

- $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$ is the point estimate of the mean value of the dependent variable when the values of the independent variables are $x_1, x_2, \ldots, x_k$.
- It is also the point prediction of an individual value of the dependent variable when the values of the independent variables are $x_1, x_2, \ldots, x_k$.
- $b_0, b_1, b_2, \ldots, b_k$ are the least squares point estimates of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$.
- $x_1, x_2, \ldots, x_k$ are specified values of the independent predictor variables.
- β_0 is the intercept (expected y when all $x_j = 0$ for $j = 1, 2, \cdots, k$).
- β_j are the partial regression coefficient (slope) for x_j where $j = 1, 2, \cdots, k$. item ε is the random error term.

Example 1: House Price (Simple and Multiple Models)

Example 1: House Price (Simple Linear Regression Model)

Variable	**Coeff**	**SE(Coeff)**	t**-ratio**	**P-value**
Intercept	14349.48	9297.69	1.5	0.1230
Bedrooms	48218.91	2843.88	16.96	$\leq$0.0001

$$\widehat{\text{Price}} = 14,349.48 + 48,218.91 \times \mathit{Bedrooms}$$

$$R^2 = 21.4\%, s_e = 68432.21 \text{ with } \mathit{df} = 1057 - 2 = 1055$$

The predictor *Bedrooms* can explain only 21.4% of the variation in *Price*.

Example 1: House Price (Multiple Linear Regression Model)

Variable	**Coeff**	**SE(Coeff)**	t**-ratio**	**P-value**
Intercept	20986.09	6816.3	3.08	0.0021
Bedrooms	-7483.10	2783.5	-2.69	0.0073
Living area	93.84	3.11	30.18	$\leq$0.0001

$$\widehat{\text{Price}} = 20,986.09 - 7,483.10 \times \mathit{Bedrooms} + 93.84 \times \mathit{Living\ Area}$$

$$R^2 = 57.8\%, s_e = 50142.4 \text{ with } \mathit{df} = 1057 - 3 = 1054$$

- Now the model (i.e., *Bedrooms* and *Living Area*) accounts for 57.8% of the variation in Price.
- Price drops with increasing bedrooms? Counter intuitive?

Example 1: House Price

Multiple regression coefficients must be interpreted in terms of the other predictors in the model.

- Simple Linear Regression Model (Predictor - Bedrooms only):

$$\widehat{\text{Price}} = 14,349.48 + 48,218.91 \times \textit{Bedrooms}$$

  On average, we'd expect the price to increase by \$48,218.91 for each additional bedroom in the house.

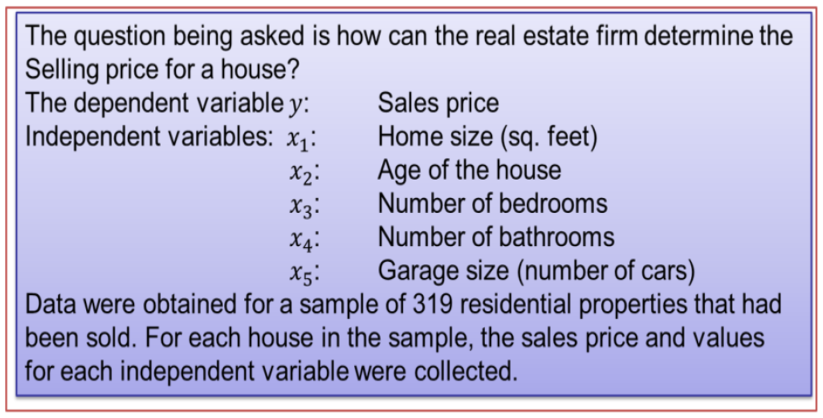
- Multiple Linear Regression Model (Predictors - Bedrooms and Living Area):

$$\widehat{\text{Price}} = 20,986.09 - 7,483.10 \times \textit{Bedrooms} + 93.84 \times \textit{Living Area}$$

  On average, we'd expect the price to decrease by \$7,483.10 for each additional bedroom in the house holding constant the effect of the variable living area.

Example 2: House Price (Multiple Regression Model)

Example 2: House Price



The question being asked is how can the real estate firm determine the Selling price for a house?

The dependent variable y: Sales price

Independent variables: x_1: Home size (sq. feet)

x_2: Age of the house

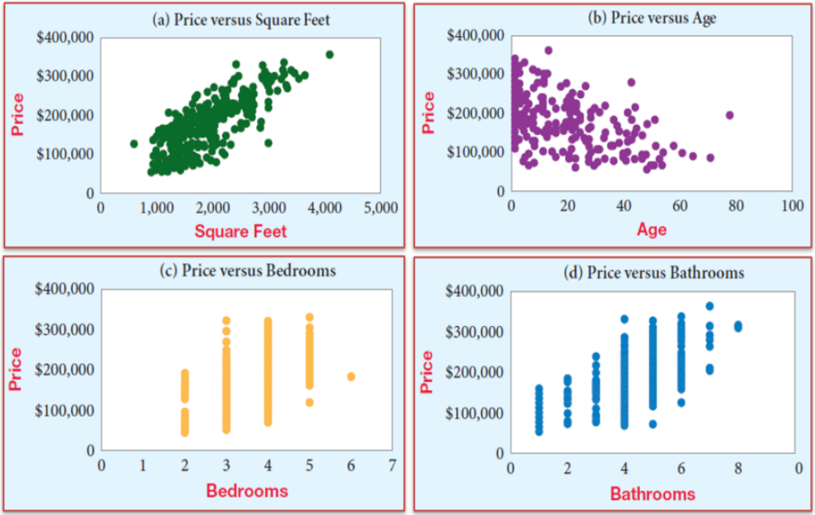
x_3: Number of bedrooms

x_4: Number of bathrooms

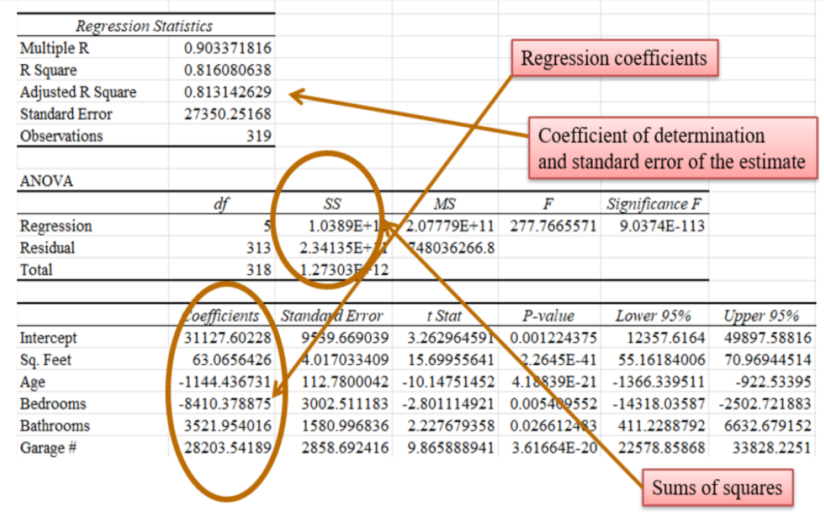
x_5: Garage size (number of cars)

Data were obtained for a sample of 319 residential properties that had been sold. For each house in the sample, the sales price and values for each independent variable were collected.

Example 2: House Price (Scatter Plots)



Example 2: House Price (EXCEL Output)



Regression Statistics	
Multiple R	0.903371816
R Square	0.816080638
Adjusted R Square	0.813142629
Standard Error	27350.25168
Observations	319

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	1.0389E+1	2.07779E+11	277.7665571	9.0374E-113
Residual	313	2.34135E+11	748036266.8		
Total	318	1.27303E+12			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	31127.60228	9539.669039	3.262964591	0.001224375	12357.6164	49897.58816
Sq. Feet	63.0656426	4.017033409	15.69955641	2.2645E-41	55.16184006	70.96944514
Age	-1144.436731	112.7800042	-10.14751452	4.18839E-21	-1366.339511	-922.53395
Bedrooms	-8410.378875	3002.511183	-2.801114921	0.005409552	-14318.03587	-2502.721883
Bathrooms	3521.954016	1580.996836	2.227679358	0.026612483	411.2288792	6632.679152
Garage #	28203.54189	2858.692416	9.865888941	3.61664E-20	22578.85868	33828.2251

Regression coefficients

Coefficient of determination and standard error of the estimate

Sums of squares

Example 2: House Price (EXCEL Output)

- The estimate of the multiple regression model:

$$\hat{y} = 31,128 + 63.07x_1 - 1,144x_2 - 8,410x_3 + 3,522x_4 + 28,204x_5$$

- Point estimate (price prediction)

x_1 - Square feet = 2,100
x_2 - Age = 15
x_3 - Number of bedrooms = 4
x_4 - Number of bathrooms = 3
x_5 - Size of garage = 2

$$\hat{y} = 31,128 + (63.07 \times 2,100) - (1,144 \times 15) - (8,410 \times 4) + (3,522 \times 3) \\ + (28,204 \times 2) = \$179,749$$

Quality of Fit

- Coefficient of Determination: R^2

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

- Interpretation: Fraction of the total variation in y accounted for by the model (all the predictor variables included). It measures how much variability in y can be explained by the multiple regression model.
- Valid Range: $0 \leq R^2 \leq 1$.
- **Problem with R^2:** Adding new predictor variables (even if they are completely irrelevant to the dependent variable) to a model never decreases R^2 and may increase it. For example, modeling weight and height

$$\widehat{Weight} = 103.40 + 6.38\, Height(\text{over 5 ft}), \quad R^2 = 0.74.$$

Adding a new variable (completely irrelevant, say campus post office box number, Box#) might give

$$\widehat{Weight} = 102.35 + 6.36\, Height(\text{over 5 ft}) + 0.02\, Box\#, \quad R^2 = 0.75.$$

Adjusted Coefficient of Determination: Adjusted R^2

Adjusted Coefficient of Determination: Adjusted R^2

$$\bar{R}^2 = (R^2 - \frac{k}{n-1})(\frac{n-1}{n-k-1}) = 1 - \frac{\frac{SSE}{n-k-1}}{\frac{SSTO}{n-1}}$$

- $\bar{R}^2$ permits a more equitable comparison between models of different sizes.
- Each additional variable results in the loss of one degree of freedom $(n - k - 1)$. The lower the degrees of freedom, the less reliable the estimates are likely to be.
- Thus, the increase in the quality of fit needs to be compared to the decrease in the degrees of freedom. The $\bar{R}^2$ takes into account this cost and adjusts the R^2 value accordingly.
- When comparing models, an increase in $\bar{R}^2$ indicates that the marginal benefit of adding a variable exceeds the cost, while a decrease in $\bar{R}^2$ indicates that the marginal cost exceeds the benefit.
- $\bar{R}^2$ will always be less than R^2.

The Overall F Test

Test 1: Is the *Overall* Model Significant?

Hypotheses:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_a : \text{At least one } \beta_j \neq 0, \text{ for some } j = 1, \cdots, k$$

Test Statistic:

$$F_0 = \frac{\frac{\text{Explained Variation}}{k}}{\frac{\text{Unexplained Variation}}{n-k-1}} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{MSR}{MSE}$$

where

$SSR =$ Sum of Squares Regression

$SSE =$ Sum of Squares Error

$n =$ Sample Size

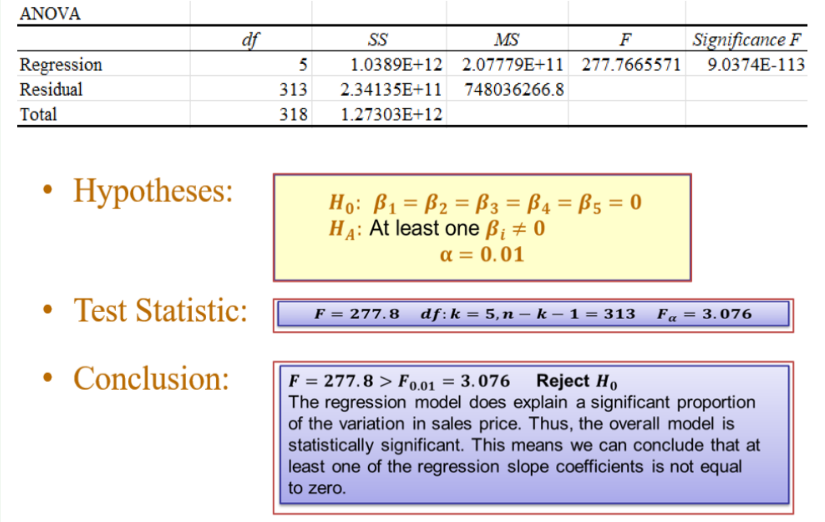
$k =$ Number of Independent Variables

ANOVA Table:

Source	**SS**	**df**	**MS**	**F-ratio**
Regression	SSR	k	MSR	MSR/MSE
Residual	SSE	$n-k-1$	MSE	
Total	$SSTO$	$n-1$		

Test 1: Is the Overall Model Significant? (House Price Model)

Example 2: House Price (EXCEL Output)



ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	1.0389E+12	2.07779E+11	277.7665571	9.0374E-113
Residual	313	2.34135E+11	748036266.8		
Total	318	1.27303E+12			

- Hypotheses:

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
H_A: At least one $\beta_i \neq 0$
$\alpha = 0.01$

- Test Statistic:

$F = 277.8 \quad df: k = 5, n - k - 1 = 313 \quad F_\alpha = 3.076$

- Conclusion:

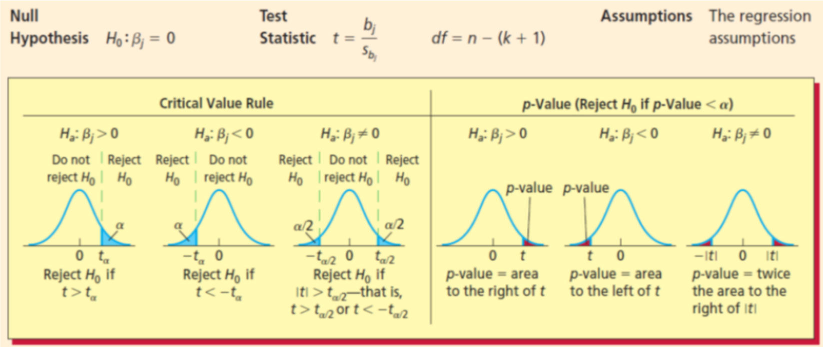
$F = 277.8 > F_{0.01} = 3.076$ **Reject H_0**
The regression model does explain a significant proportion of the variation in sales price. Thus, the overall model is statistically significant. This means we can conclude that at least one of the regression slope coefficients is not equal to zero.

Test 2: Testing the Significance of The individual Independent Variables

Test 2: Are the individual Variables Significant?

If a multiple regression F-test leads to a rejection of the null hypothesis, then perform t-test for each regression coefficient.

$$H_0 : \beta_j = 0, \text{ given all other variables are in the model}$$
$$H_a : \beta_j \neq 0, \text{ given all other variables are in the model}$$



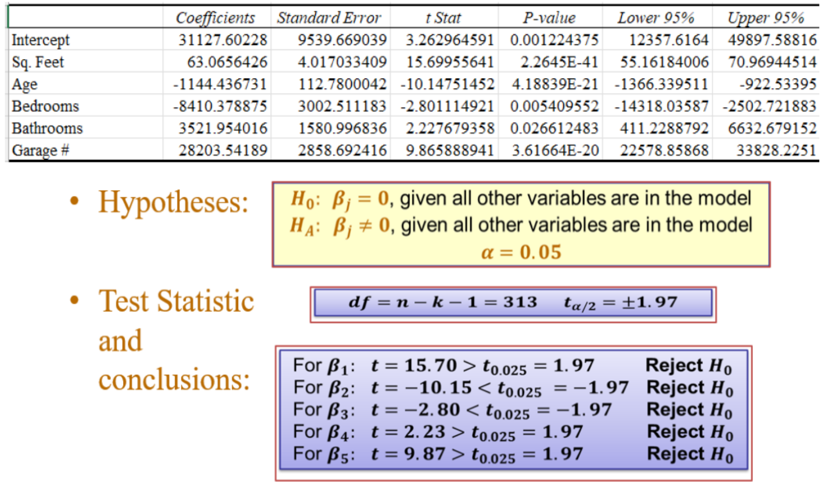
Null Hypothesis $H_0: \beta_j = 0$ **Test Statistic** $t = \frac{b_j}{s_{b_j}}$ $df = n - (k+1)$ **Assumptions** The regression assumptions

Critical Value Rule			p-Value (Reject H_0 if p-Value $< \alpha$)		
$H_a: \beta_j > 0$	$H_a: \beta_j < 0$	$H_a: \beta_j \neq 0$	$H_a: \beta_j > 0$	$H_a: \beta_j < 0$	$H_a: \beta_j \neq 0$
Reject H_0 if $t > t_\alpha$	Reject H_0 if $t < -t_\alpha$	Reject H_0 if $\lvert t\rvert > t_{\alpha/2}$—that is, $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$	p-value = area to the right of t	p-value = area to the left of t	p-value = twice the area to the right of $\lvert t\rvert$

Testing the Significance of an Independent Variable

Example 2: House Price

Are the individual Variables Significant?



	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	31127.60228	9539.669039	3.262964591	0.001224375	12357.6164	49897.58816
Sq. Feet	63.0656426	4.017033409	15.69955641	2.2645E-41	55.16184006	70.96944514
Age	-1144.436731	112.7800042	-10.14751452	4.18839E-21	-1366.339511	-922.53395
Bedrooms	-8410.378875	3002.511183	-2.801114921	0.005409552	-14318.03587	-2502.721883
Bathrooms	3521.954016	1580.996836	2.227679358	0.026612483	411.2288792	6632.679152
Garage #	28203.54189	2858.692416	9.865888941	3.61664E-20	22578.85868	33828.2251

- Hypotheses:

H_0: $\beta_j = 0$, given all other variables are in the model

H_A: $\beta_j \neq 0$, given all other variables are in the model

$\alpha = 0.05$

- Test Statistic and conclusions:

$df = n - k - 1 = 313 \qquad t_{\alpha/2} = \pm 1.97$

For β_1: $t = 15.70 > t_{0.025} = 1.97$ Reject H_0

For β_2: $t = -10.15 < t_{0.025} = -1.97$ Reject H_0

For β_3: $t = -2.80 < t_{0.025} = -1.97$ Reject H_0

For β_4: $t = 2.23 > t_{0.025} = 1.97$ Reject H_0

For β_5: $t = 9.87 > t_{0.025} = 1.97$ Reject H_0

Testing the Significance of an Independent Variable

Tricky Parts of the t test:

- SE's are harder to compute (let technology do it!).
- The meaning of a coefficient depends on the other predictors in the model.
  - If we fail to reject $H_0 : \beta_j = 0$ based on it's t-test, it does not mean that x_j has no linear relationship to y.
  - Rather, it means that x_j contributes nothing to modeling y after allowing for the other predictors.

Example 3: Pie Sale

Example 3: Pie Sale

A distributor of frozen desert pies wants to evaluate factors thought to influence demand. Data are collected for 15 weeks.

Week	Pie Sales	Price ($)	Advertising ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

Multiple regression model:

$$\widehat{\text{Sales}} = b_0 + b_1\,(\text{Price}) + b_2\,(\text{Advertising})$$

Correlation matrix:

	Pie Sales	Price	Advertising
Pie Sales	1		
Price	-0.44327	1	
Advertising	0.55632	0.03044	1

Example 3: Pie Sales (Excel Multiple Regression Output)

Regression Statistics

Regression Statistics	
Multiple R	0.722134292
R Square	0.521477936
Adjusted R Square	0.441724259
Standard Error	47.46341263
Observations	15

ANOVA

	df	SS	MS	F	Significance F
Regression	2	29460.02687	14730.01343	6.538606789	0.012006372
Residual	12	27033.30647	2252.775539		
Total	14	56493.33333			

Model Coefficients

	Coefficients	Standard Error	t Stat	P-value
Intercept	306.5261933	114.2538935	2.682851182	0.019931591
Price	-24.97508952	10.83212512	-2.305650022	0.039788461
Advertising	74.13095749	25.96731792	2.854779139	0.014493627

Regression Equation:

$\widehat{\text{Sales}} = 306.53 - 24.98(\text{Price}) + 74.13(\text{Advertising})$

Example 3: Pie Sale (Interpreting the Multiple Regression Results)

- **Interpreting the Slope (b_1):**

  For the **Price** variable, the coefficient of approximately -25 indicates that the average value of sales (y) is expected to **decrease by 25 pies** for each \$1 increase in price, provided advertising is held constant.

- **Interpreting the Slope (b_2):**

  For the **advertising** variable, the coefficient of approximately 74 indicates that the average value of sales (y) is expected to **increase by 74 pies** for every additional \$100 spent on advertising, assuming the price remains constant.

- **Interpreting the y-Intercept (b_0):**

  The intercept of 306.53 represents the estimated average number of pies sold per week if both the price were set to zero and no money were spent on advertising, provided these values are within the range of the observed data.

- **Model Quality (R^2):**

  The R^2 value of **0.5215** means that **52.15%** of the total variation in pie sales is accounted for by the combined effects of Price and Advertising.

- **Adjusted Fit ($\bar{R}^2$):**

  The Adjusted R^2 of **0.4417** provides a more equitable comparison for model quality by accounting for the loss of degrees of freedom when adding predictors.

- **Overall Significance:**

  The "Significance F" value of **0.0120** is less than **0.05**, so we reject the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$. This means the overall model is **statistically significant**.

Summary of Key Formulas

1. Multiple Regression Model

$$\boxed{\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k}$$

- b_0: Estimated intercept (expected y when all $x_j = 0$)
- b_j: Estimated slope for x_j ($j = 1, 2, \ldots, k$)

2. Model Quality Measures

- **Coefficient of Determination:**

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

- **Adjusted R^2 (accounts for model size):**

$$\bar{R}^2 = 1 - \frac{\frac{SSE}{n-k-1}}{\frac{SSTO}{n-1}} = (R^2 - \frac{k}{n-1})\left(\frac{n-1}{n-k-1}\right)$$

  - n: Sample size, k: Number of predictors
  - $\bar{R}^2 \leq R^2$ always

Summary of Key Formulas

3. ANOVA Table for Regression

Source	**Sum of Squares (SS)**	**df**	**Mean Square (MS)**	**F**
Regression	$SSR = \sum(\hat{y}_i - \bar{y})^2$	k	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Residual	$SSE = \sum(y_i - \hat{y}_i)^2$	$n-k-1$	$MSE = \frac{SSE}{n-k-1}$	
Total	$SSTO = \sum(y_i - \bar{y})^2$	$n-1$		

$$\boxed{SSTO = SSR + SSE}$$

4. Hypothesis Testing

Overall Model (F-test):

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$
- H_a: At least one $\beta_j \neq 0$

$$F_0 = \frac{MSR}{MSE} \sim F_{k,n-k-1}$$

Reject H_0 if $F_0 > F_{\alpha,k,n-k-1}$ or p-value $< \alpha$

Individual Predictor (t-test):

- $H_0 : \beta_j = 0$ (given other predictors)
- $H_a : \beta_j \neq 0$

$$t_0 = \frac{b_j}{SE(b_j)} \sim t_{n-k-1}$$

Reject H_0 if $|t| > t_{\alpha/2,n-k-1}$ or p-value $< \alpha$