



# COMM215

First the Foundation, then Innovation

## **LESSON 12**

## **MULTIPLE LINEAR REGRESSION**

**SAMIE L.S. LY**

## 1. Multiple Regression Model

2. Least Squares Method
3. Multiple Coefficient of Determination
4. Model Assumptions
5. Testing for Significance
6. Using the Estimated Regression Equation for Estimation and Prediction

# MULTIPLE REGRESSION MODEL



## Multiple Regression Model

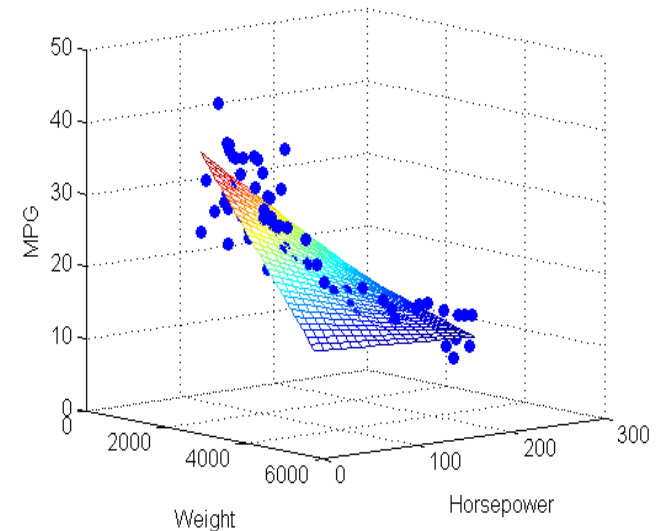
$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p + \varepsilon$$

## Multiple Regression Equation

$$E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p$$

## Estimated Multiple Regression Equation

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p$$



# INTERPRETATION OF COEFFICIENTS



$$\hat{y} = 100 + 20 x_1$$

In simple linear regression:  $b_1$  is an estimate of the change in  $y$  for one-unit change in the independent variable  $x_1$ .

$$\hat{y} = 100 + 20 x_1 + 3x_2 + 120 x_3$$

In multiple linear regression:  $b_i$  represents an estimate of the change in  $y$  corresponding to a one-unit change in  $x_i$   
**when all other independent variables are held constant.**

# ESTIMATION PROCESS

Multiple Regression Model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

Multiple Regression Equation

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Unknown parameters are

$$\beta_0, \beta_1, \beta_2, \dots, \beta_p$$

Sample Data:

$x_1$	$x_2$	$\dots$	$x_p$	$y$
$\cdot$	$\cdot$		$\cdot$	$\cdot$
$\cdot$	$\cdot$		$\cdot$	$\cdot$

Estimated Multiple  
Regression Equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$$

Sample statistics are

$$b_0, b_1, b_2, \dots, b_p$$

$b_0, b_1, b_2, \dots, b_p$   
provide estimates of

$$\beta_0, \beta_1, \beta_2, \dots, \beta_p$$

# LEAST SQUARES METHOD



## Least Squares Criterion

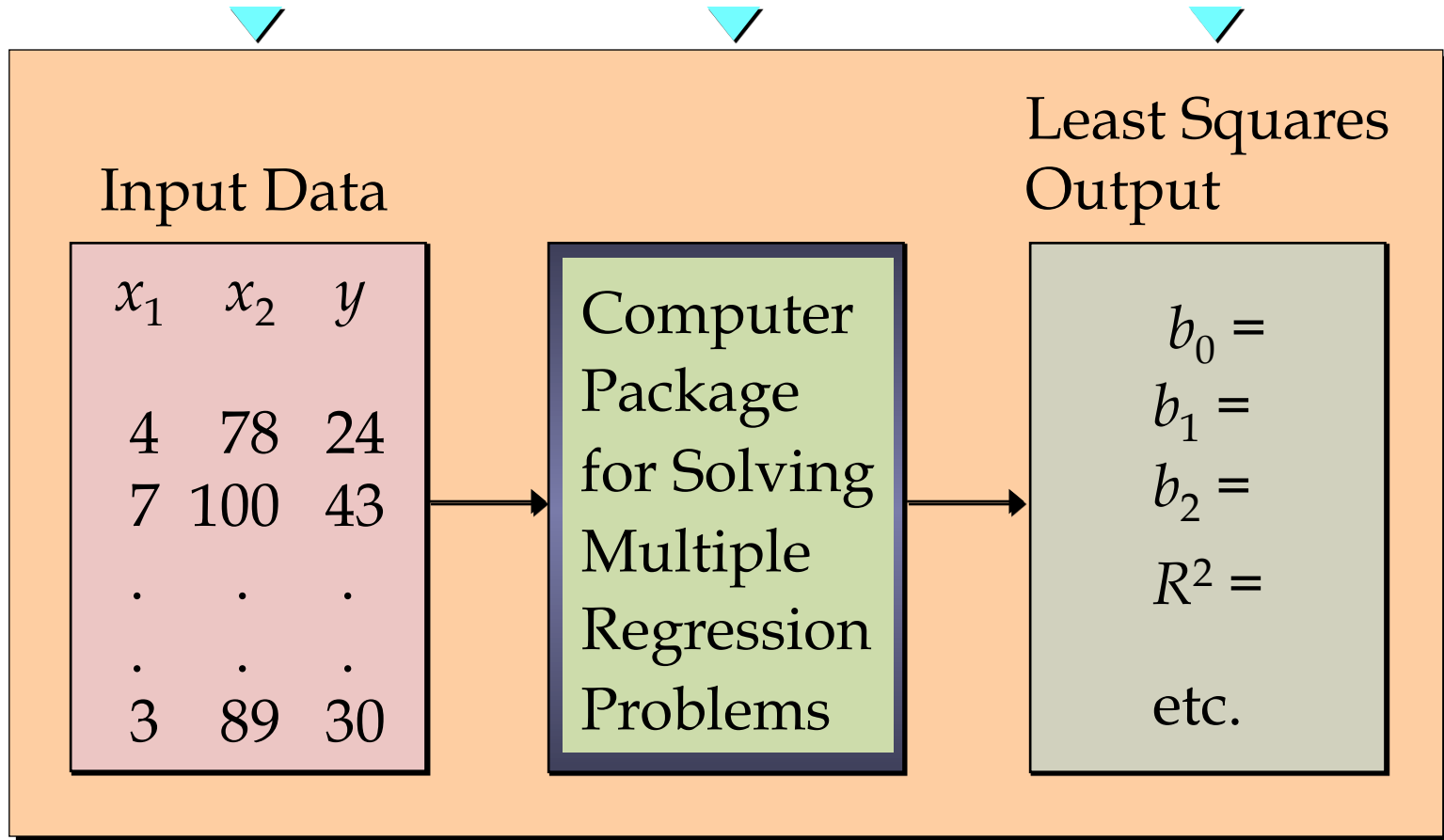


$$\min \sum (y_i - \hat{y}_i)^2$$

### ► ■ Computation of Coefficient Values

The formulas for the regression coefficients  $b_0, b_1, b_2, \dots, b_p$  involve the use of matrix algebra. We will rely on computer software packages to perform the calculations.

# Solving for the Estimates of $\beta_0, \beta_1, \beta_2$



# Solving for the Estimates of $\beta_0, \beta_1, \beta_2$

## ■ Excel's Regression Equation Output

	A	B	C	D	E
38					
39		<i>Coeffic.</i>	<i>Std. Err.</i>	<i>t Stat</i>	<i>P-value</i>
▶ 40	Intercept	3.17394	6.15607	0.5156	0.61279
41	Experience	1.4039	0.19857	7.0702	1.9E-06
42	Test Score	0.25089	0.07735	3.2433	0.00478
43					

Note: Columns F-I are not shown.






	A	B	C	D	E	F	G
14	SUMMARY OUTPUT			Visitors	Col-Inches	Discount	
15	<i>Regression Statistics</i>			23	4	100	
16	Multiple R	0.8465		30	7	20	
17	R Square	0.7165		20	3	40	
18	Adjusted R Square	0.6031		26	6	25	
19	Standard Error	3.3749		20	2	50	
20	Observations	8		18	5	30	
21				17	4	25	
22	ANOVA			31	8	80	
23		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
24	Regression	2	143.924	71.962	6.318	0.043	
25	Residual	5	56.951	11.390			
26	Total	7	200.875				
27							
28		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
29	Intercept	10.687	3.875	2.758	0.040	0.726	20.648
30	Col-Inches	2.157	0.628	3.434	0.019	0.542	3.771
31	Discount	0.042	0.044	0.949	0.386	-0.071	0.154

# MULTIPLE COEFFICIENT OF DETERMINATION

## ■ Relationship Among SST, SSR, SSE



$$SST = SSR + SSE$$


$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

where:

SST = total sum of squares

SSR = sum of squares due to regression

SSE = sum of squares due to error

# MULTIPLE COEFFICIENT OF DETERMINATION

## ■ Excel's ANOVA Output

	A	B	C	D	E	F
32						
33	ANOVA					
34		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
35	Regression	2	500.3285	250.1643	42.76013	2.32774E-07
36	Residual	17	99.45697	5.85041		
37	Total	19	599.7855			
38						

SST

SSR

# MULTIPLE COEFFICIENT OF DETERMINATION

$$R^2 = SSR/SST$$

$$\triangleright R^2 = 500.3285/599.7855 = .83418$$

# Assumptions About the Error Term $\varepsilon$

- ▶ The error  $\varepsilon$  is a random variable with mean of zero.
- ▶ The variance of  $\varepsilon$ , denoted by  $\sigma^2$ , is the same for all values of the independent variables.
- ▶ The values of  $\varepsilon$  are independent.
- ▶ The error  $\varepsilon$  is a normally distributed random variable reflecting the deviation between the  $y$  value and the expected value of  $y$  given by  $\beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p$ .

# TESTING FOR SIGNIFICANCE



In simple linear regression, the  $F$  and  $t$  tests provide the same conclusion.

In multiple regression, the  $F$  and  $t$  tests have different purposes.

# TESTING FOR SIGNIFICANCE: F-TEST



The  $F$  test is used to determine whether a significant relationship exists between the dependent variable and the set of all the independent variables.

The  $F$  test is referred to as the test for overall significance.

# TESTING FOR SIGNIFICANCE: *F* TEST



$$H_0: \beta_1 = \beta_2 = 0$$

1. Set up Hypotheses.  $H_a$ : One or both of the parameters is not equal to zero.
2. What is the appropriate test statistic to use?.
3. Calculate the test statistic value.  $F = \text{MSR}/\text{MSE}$
4. Find the critical value for the test statistic.  $\alpha = .05$
5. Define the decision rule
6. Make your decision
7. Interpret the conclusion in context



# F TEST FOR OVERALL SIGNIFICANCE



## ■ Excel's ANOVA Output

	A	B	C	D	E	F
32						
33	ANOVA					
34		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
35	Regression	2	500.3285	250.1643	42.76013	2.32774E-07
36	Residual	17	99.45697	5.85041		
37	Total	19	599.7855			
38						

*p*-value used to test for overall significance



### Step 3: Calculate $F_{\text{observed}}$

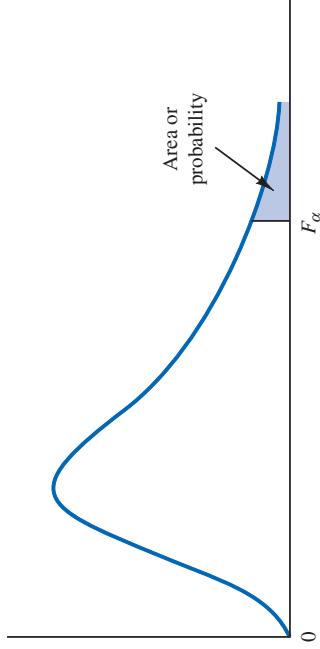
Source	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	SSR	p	$MSR = SSR/p$	$F = MSR/MSE$
Error	SSE	n-p-1	$MSE = SSE/n-p-1$	
Total	SST	n-1		



A	B	C	D	E	F
SUMMARY OUTPUT					
<i>Regression Statistics</i>					
Multiple R	0.8734				
R Square					
Adjusted R Square	0.7453				
Standard Error	3.75				
Observations	30				
ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance</i>
Regression	2	1223.2			0.0
Residual	27		1 611.59	43.43	
Total	29	160 380.2			
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	
Intercept	13.01	3.53		0.0010	
Assignment	0.194	0.200		0.3417	
Midterm	1.11	0.122		0.0000	



TION



$F_\alpha$  values, where  $\alpha$  is the area or probability in the upper tail of the  $F$  distribution. For example, with 4 numerator degrees of freedom, and a .05 area in the upper tail,  $F_{.05} = 3.84$ .

Numerator Degrees of Freedom

NUMERATOR

1	2	3	4	5	6	8	9	10	15	20	25	30
39.86	49.50	53.59	55.83	57.24	58.20	59.44	59.86	60.19	61.22	61.74	62.05	62.26
161.45	199.50	215.71	224.58	230.16	233.99	238.88	240.54	241.88	245.95	248.02	249.26	250.10
647.79	799.48	864.15	899.60	921.83	937.11	956.64	963.28	968.63	984.87	993.08	998.09	1001.40
052.18	4999.34	5403.53	5624.26	5763.96	5858.95	5900.95	6022.40	6055.93	6156.97	6208.66	6239.86	6260.35
8.53	9.00	9.16	9.24	9.29	9.33	9.37	9.38	9.39	9.42	9.44	9.45	9.46
18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.38	19.40	19.43	19.45	19.46	19.46
38.51	39.00	39.17	39.25	39.30	39.33	39.37	39.39	39.40	39.43	39.45	39.46	39.46
98.50	99.00	99.16	99.25	99.30	99.33	99.38	99.39	99.40	99.43	99.45	99.46	99.47
5.54	5.46	5.39	5.34	5.31	5.28	5.25	5.24	5.23	5.20	5.18	5.17	5.17
10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.81	8.79	8.70	8.66	8.63	8.62
17.44	16.04	15.44	15.10	14.88	14.73	14.54	14.47	14.42	14.25	14.17	14.12	14.08
34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.34	27.23	26.87	26.69	26.58	26.50
4.54	4.32	4.19	4.11	4.05	4.01	3.95	3.94	3.92	3.87	3.84	3.83	3.82
7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.80	5.77	5.75
12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.56	8.50	8.46
21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.02	13.91	13.84
4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.324	3.21	3.17
6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.50
10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43	6.27	6.23
16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.72	9.55	9.38

# F TABLE : 3 PIECES OF INFORMATION



**F**  $\alpha, k, n-k-1$

- Alpha  $\alpha$
- Numerator: df of SSR
- Denominator: df of SSE

	Denominator Degrees of Freedom	Area in Upper Tail	NUMERATOR							
			1	2	3	4	5	6	7	8
DENOMINATOR	25	.10	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93
		.05	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34
		.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75
		.01	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32
	26	.10	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92
		.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32
		.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73
		.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29
	27	.10	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91
		.05	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31
		.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71
		.01	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26
	28	.10	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90
		.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29
		.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69
		.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23
	29	.10	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89
		.05	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28
		.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67
		.01	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20

# TESTING FOR SIGNIFICANCE : T TEST



- ▶ If the  $F$  test shows an overall significance, the  $t$  test is used to determine whether each of the individual independent variables is significant.
- ▶ A separate  $t$  test is conducted for each of the independent variables in the model.
- ▶ We refer to each of these  $t$  tests as a test for individual significance.

## $t$ Test for Significance of Individual Parameters

### ■ Excel's Regression Equation Output

	A	B	C	D	E
38					
39		<i>Coeffic.</i>	<i>Std. Err.</i>	<i>t Stat</i>	<i>P-value</i>
40	Intercept	3.17394	6.15607	0.5156	0.61279
41	Experience	1.4039	0.19857	7.0702	1.9E-06
42	Test Score	0.25089	0.07735	3.2433	0.00478
43					

Note: Columns F-I are not shown.

$t$  statistic and  $p$ -value used to test for the individual significance of "Experience"

# TESTING FOR SIGNIFICANCE:



▶ The term multicollinearity refers to the correlation among the independent variables.

▶ When the independent variables are highly correlated (say,  $|r| > .7$ ), it is not possible to determine the separate effect of any particular independent variable on the dependent variable.

$$E(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$





# TESTING FOR SIGNIFICANCE




- If the estimated regression equation is to be used only for predictive purposes, multicollinearity is usually not a serious problem.
- Every attempt should be made to avoid including independent variables that are highly correlated.



The **standard errors** for the partial regression coefficients ( $s_{\beta_1}, s_{\beta_2} \dots$ ) become very large

and the coefficients are **statistically unreliable** and **difficult to interpret**.

Multicollinearity is a problem when we are trying to interpret the partial regression coefficients.  $\beta_1, \beta_2$  etc..

$$E(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$


# USING THE ESTIMATED REG EQ



- ▶ The procedures for estimating the mean value of  $y$  and predicting an individual value of  $y$  in multiple regression are similar to those in simple regression.
- ▶ We substitute the given values of  $x_1, x_2, \dots, x_p$  into the estimated regression equation and use the corresponding value of  $y$  as the point estimate.

# USING THE ESTIMATED REG EQ



- ▶ The formulas required to develop interval estimates for the mean value of  $\hat{y}$  and for an individual value of  $y$  are beyond the scope of the textbook.
- ▶ Software packages for multiple regression will often provide these interval estimates.