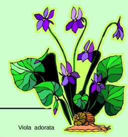
ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΓΕΩΠΟΝΙΚΗ ΣΧΟΛΗ
Α.Π.Θ.

Multiple Regression

Dr. George Menexes
Aristotle University of Thessaloniki
School of Agriculture, Lab of Agronomy

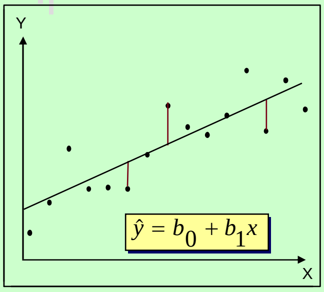


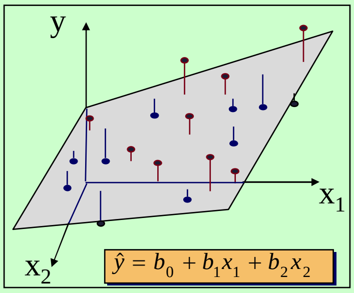
Learning Objectives

In this lecture, you learn:

- How to develop a multiple regression model
- How to interpret the regression coefficients
- How to determine which independent variables to include in the regression model
- How to determine which independent variables are most important in predicting a dependent variable
- How to use categorical variables in a regression model

Simple and Multiple Least-Squares Regression





In a **simple regression model**, the least-squares estimators minimize the sum of squared errors from the estimated regression **line**.

In a **multiple regression model**, the least-squares estimators minimize the sum of squared errors from the estimated regression **plane**.

The Multiple Regression Model

Idea: Examine the linear relationship between 1 dependent (Y) & 2 or more independent variables (X_i).

Multiple Regression Model with k Independent Variables:

$$Y_i = \underbrace{\beta_0}_{\text{Y-intercept}} + \underbrace{\beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}}_{\text{Population slopes}} + \underbrace{\varepsilon_i}_{\text{Random Error}}$$

Multiple Regression Equation

The coefficients of the multiple regression model are estimated using sample data

Multiple regression equation with k independent variables:

Estimated (or predicted) value of Y

Estimated intercept

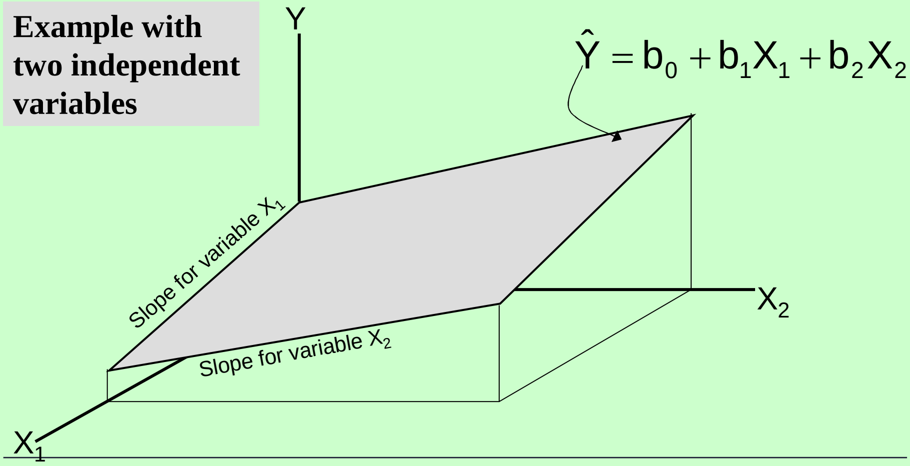
Estimated slope coefficients

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_k X_{ki}$$

In this lecture we will always use Excel to obtain the regression slope coefficients and other regression summary measures.

Multiple Regression Equation

Example with two independent variables



Multiple Regression Equation 2 Variable Example

- A distributor of frozen dessert pies wants to evaluate factors thought to influence demand
  - Dependent variable: Pie sales (units per week)
  - Independent variables: Price (in $)
    Advertising ($100's)
- Data are collected for 15 weeks

Multiple Regression Equation 2 Variable Example

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 equation:

- Sales = $b_0 + b_1$ (Price) + b_2 (Advertising)
- Sales = $b_0 + b_1X_1 + b_2X_2$

Where X_1 = Price

X_2 = Advertising

Multiple Regression Equation 2 Variable Example, Excel

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

ANOVA	df	SS	MS	F	Significance F
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

$$\text{Sales} = 306.526 - 24.975(X_1) + 74.131(X_2)$$

Multiple Regression Equation 2 Variable Example

$$\text{Sales} = 306.526 - 24.975(X_1) + 74.131(X_2)$$

where
Sales is in number of pies per week
Price is in $
Advertising is in $100's.

b_1 = -24.975: sales will decrease, on average, by 24.975 pies per week for each $1 increase in selling price, net of the effects of changes due to advertising

b_2 = 74.131: sales will increase, on average, by 74.131 pies per week for each $100 increase in advertising, net of the effects of changes due to price

Multiple Regression Equation 2 Variable Example

Predict sales for a week in which the selling price is $5.50 and advertising is $350:

$$\begin{aligned} \text{Sales} &= 306.526 - 24.975(X_1) + 74.131(X_2) \\ &= 306.526 - 24.975(5.50) + 74.131(3.5) \\ &= 428.62 \end{aligned}$$

Predicted sales is 428.62 pies

Note that Advertising is in $100's, so $350 means that $X_2 = 3.5$

Coefficient of Multiple Determination

- Reports the proportion of total variation in Y explained by all X variables taken together

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

Coefficient of Multiple Determination (Excel)

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$r^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$

52.1% of the variation in pie sales is explained by the variation in price and advertising

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Adjusted r^2

- r^2 never decreases when a new X variable is added to the model
  - This can be a disadvantage when comparing models
- What is the net effect of adding a new variable?
  - We lose a degree of freedom when a new X variable is added
  - Did the new X variable add enough independent power to offset the loss of one degree of freedom?

Adjusted r^2

- Shows the proportion of variation in Y explained by all X variables adjusted for the number of X variables used

$$r^2 = 1 - \left[(1 - r^2_{Y.12..k}) \left(\frac{n-1}{n-k-1} \right) \right]$$

(where n = sample size, k = number of independent variables)

- Penalizes excessive use of unimportant independent variables
- Smaller than r^2
- Useful in comparing models

Adjusted r^2

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$r^2_{adj} = .44172$

44.2% of the variation in pie sales is explained by the variation in price and advertising, taking into account the sample size and number of independent variables

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

F-Test for Overall Significance

- F-Test for Overall Significance of the Model
  - Shows if there is a linear relationship between all of the X variables considered together and Y
- Use F test statistic
  - Hypotheses:

    H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (no linear relationship)

    H_1: at least one $\beta_i \neq 0$ (at least one independent variable affects Y)

F-Test for Overall Significance

- Test statistic:

$$F = \frac{MSR}{MSE} = \frac{\dfrac{SSR}{k}}{\dfrac{SSE}{n-k-1}}$$

- where F has (numerator) = k and (denominator) = (n - k - 1) degrees of freedom

F-Test for Overall Significance

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$F = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

P-value for the F-Test

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

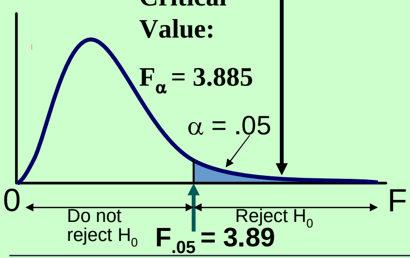
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

F-Test for Overall Significance

- H_0: $\beta_1 = \beta_2 = 0$
- H_1: β_1 and β_2 not both zero
- $\alpha = .05$
- $df_1 = 2$ $df_2 = 12$

Critical Value:

$F_\alpha = 3.885$



Test Statistic:

$$F = \frac{MSR}{MSE} = 6.5386$$

Decision:

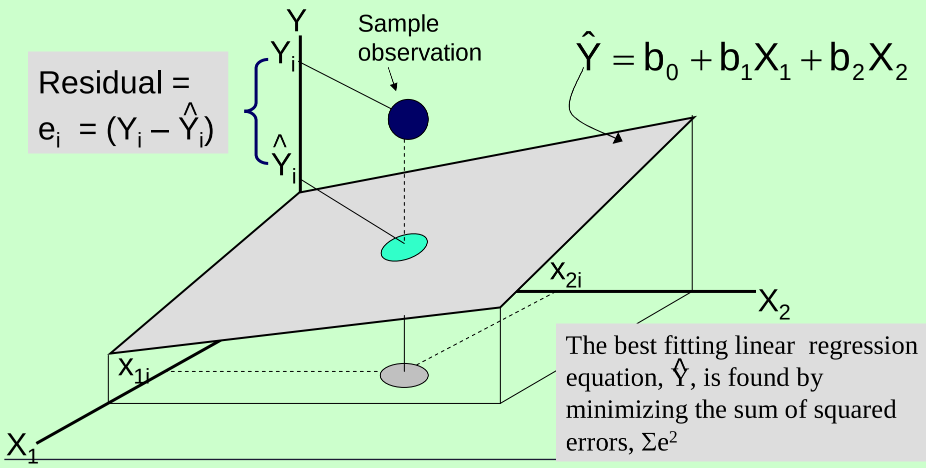
Since F test statistic is in the rejection region (p-value < .05), reject H_0

Conclusion:

There is evidence that at least one independent variable affects Y

Residuals in Multiple Regression

Two variable model



Multiple Regression Assumptions

Errors (residuals) from the regression model:

$$e_i = (Y_i - \hat{Y}_i)$$

Assumptions:

- The errors are independent
- The errors are normally distributed
- Errors have an equal variance

Multiple Regression Assumptions

- These residual plots are used in multiple regression:
  - Residuals vs. $\hat{Y}_i$
  - Residuals vs. X_{1i}
  - Residuals vs. X_{2i}
  - Residuals vs. time (if time series data)

Use the residual plots to check for violations of regression assumptions

Individual Variables Tests of Hypothesis

- Use t-tests of individual variable slopes
- Shows if there is a linear relationship between the variable X_i and Y
- Hypotheses:

  H_0: $\beta_i = 0$ (no linear relationship)

  H_1: $\beta_i \neq 0$ (linear relationship does exist between X_i and Y)

Individual Variables Tests of Hypothesis

H_0: β_j = 0 (no linear relationship)
H_1: $\beta_j \neq 0$ (linear relationship does exist between X_i and Y)

- Test Statistic:

$$t = \frac{b_j - 0}{S_{b_j}}$$

(df = n – k – 1)

Individual Variables Tests of Hypothesis

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

t-value for Price is t = -2.306, with p-value .0398

t-value for Advertising is t = 2.855, with p-value .0145

ANOVA	df	SS	MS	F	Significance F
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Individual Variables Tests of Hypothesis

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$

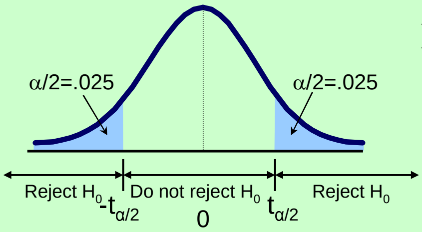
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Price	-24.97509	10.83213	-2.30565	0.03979
Advertising	74.13096	25.96732	2.85478	0.01449

d.f. = 15 - 2 - 1 = 12

α = .05

$t_{\alpha/2}$ = 2.1788

The test statistic for each variable falls in the rejection region (p-values < .05)



Decision:

Reject H_0 for each variable

Conclusion:

There is evidence that both Price and Advertising affect pie sales at α = .05

Confidence Interval Estimate for the Slope

Confidence interval for the population slope β_i

$$b_i \pm t_{n-k-1} S_{b_i}$$

where t has (n – k – 1) d.f.

	Coefficients	*Standard Error*
Intercept	306.52619	114.25389
Price	-24.97509	10.83213
Advertising	74.13096	25.96732

Here, t has (15 – 2 – 1) = 12 d.f.

Example: Form a 95% confidence interval for the effect of changes in price (X_1) on pie sales, holding constant the effects of advertising:

-24.975 ± (2.1788)(10.832): So the interval is (-48.576, -1.374)

Confidence Interval Estimate for the Slope

Confidence interval for the population slope β_i

	Coefficients	*Standard Error*	…	*Lower 95%*	*Upper 95%*
Intercept	**306.52619**	**114.25389**	**…**	**57.58835**	**555.46404**
Price	**-24.97509**	**10.83213**	**…**	**-48.57626**	**-1.37392**
Advertising	**74.13096**	**25.96732**	**…**	**17.55303**	**130.70888**

Example: Excel output also reports these interval endpoints:

Weekly sales are estimated to be reduced by between 1.37 to 48.58 pies for each increase of $1 in the selling price, holding constant the effects of advertising.

Testing Portions of the Multiple Regression Model

- Contribution of a Single Independent Variable X_j

SSR(X_j | all variables except X_j)
= SSR (all variables) – SSR(all variables except X_j)

- Measures the contribution of X_j in explaining the total variation in Y (SST)

Testing Portions of the Multiple Regression Model

Contribution of a Single Independent Variable X_j, assuming all other variables are already included (consider here a 3-variable model):

SSR(X_1 | X_2 and X_3)
= **SSR (all variables) – SSR(X_2 and X_3)**

From ANOVA section of regression for

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

From ANOVA section of regression for

$$\hat{Y} = b_0 + b_2X_2 + b_3X_3$$

Measures the contribution of X_1 in explaining SST

The Partial F-Test Statistic

Consider the hypothesis test:

H_0: variable X_j does not significantly improve the model after all other variables are included

H_1: variable X_j significantly improves the model after all other variables are included

Test using the F-test statistic:
(with 1 and n-k-1 d.f.)

$$F = \frac{\text{SSR}\,(X_j \mid \text{all variables except j})}{\text{MSE}}$$

Testing Portions of Model: Example

Example: Frozen dessert pies

Test at the α = .05 level to determine whether the price variable significantly improves the model given that advertising is included

Testing Portions of Model: Example

H_0: X_1 (price) does not improve the model with X_2 (advertising) included

H_1: X_1 does improve model

α = .05, df = 1 and 12

F critical Value = 4.75

(For X_1 and X_2)

ANOVA			
	df	*SS*	*MS*
Regression	2	**29460.02687**	14730.01343
Residual	12	27033.30647	**2252.775539**
Total	14	56493.33333	

(For X_2 only)

ANOVA		
	df	*SS*
Regression	1	**17484.22249**
Residual	13	39009.11085
Total	14	56493.33333

Testing Portions of Model: Example

(For X_1 and X_2)

ANOVA			
	df	*SS*	*MS*
Regression	2	**29460.02687**	14730.01343
Residual	12	27033.30647	**2252.775539**
Total	14	56493.33333	

(For X_2 only)

ANOVA		
	df	*SS*
Regression	1	**17484.22249**
Residual	13	39009.11085
Total	14	56493.33333

$$F = \frac{SSR(X_1 \mid X_2)}{MSE(all)} = \frac{29{,}460.03 - 17{,}484.22}{2252.78} = 5.316$$

Conclusion: Reject H_0; adding X_1 does improve model

Relationship Between Test Statistics

- The partial *F* test statistic developed in this section and the *t* test statistic are both used to determine the contribution of an independent variable to a multiple regression model.
- The hypothesis tests associated with these two statistics always result in the same decision (that is, the *p*-values are identical).

$$t_a^2 = F_{1,a}$$

Where a = degrees of freedom

Coefficient of Partial Determination for k Variable Model

$$r^2_{Yj.(\text{all variables except } j)} = \frac{SSR(X_j \mid \text{all variables except } j)}{SST - SSR(\text{all variables}) + SSR(X_j \mid \text{all variables except } j)}$$

Measures the proportion of variation in the dependent variable that is explained by X_j while controlling for (holding constant) the other independent variables

Using Dummy Variables

- A dummy variable is a categorical independent variable with two levels:
  - yes or no, on or off, male or female
  - coded as 0 or 1
- Assumes equal slopes for other variables
- If more than two levels, the number of dummy variables needed is (number of levels - 1)

Dummy Variable Example

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2$$

Let:

Y = pie sales

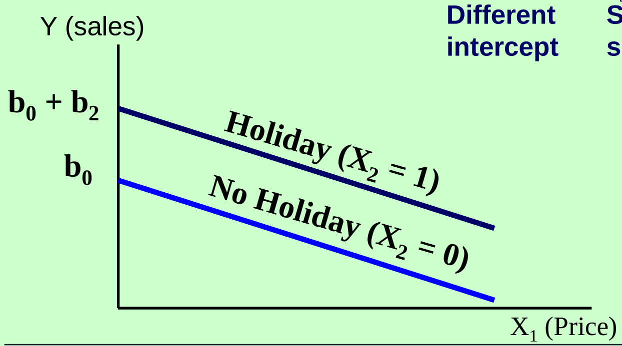
X_1 = price

X_2 = holiday (X_2 = 1 if a holiday occurred during the week)
(X_2 = 0 if there was no holiday that week)

Dummy Variable Example

$\hat{Y} = b_0 + b_1X_1 + b_2(1) = (b_0 + b_2) + b_1X_1$	**Holiday**
$\hat{Y} = b_0 + b_1X_1 + b_2(0) = \qquad b_0 \qquad + b_1X_1$	**No Holiday**

Different intercept **Same slope**



If H_0: $\beta_2 = 0$ is rejected, then "Holiday" has a significant effect on pie sales

Dummy Variable Example

$$\text{Sales} = 300 - 30(\text{Price}) + 15(\text{Holiday})$$

Sales: number of pies sold per week
Price: pie price in $

$$\text{Holiday: } \begin{cases} 1 & \text{If a holiday occurred during the week} \\ 0 & \text{If no holiday occurred} \end{cases}$$

b_2 = 15: on average, sales were 15 pies greater in weeks with a holiday than in weeks without a holiday, given the same price

Interaction Between Independent Variables

- Hypothesizes interaction between pairs of X variables
  - Response to one X variable may vary at different levels of another X variable
- Contains a two-way cross product term
  - $$\begin{aligned} \hat{Y} &= b_0 + b_1X_1 + b_2X_2 + b_3X_3 \\ &= b_0 + b_1X_1 + b_2X_2 + b_3(X_1X_2) \end{aligned}$$

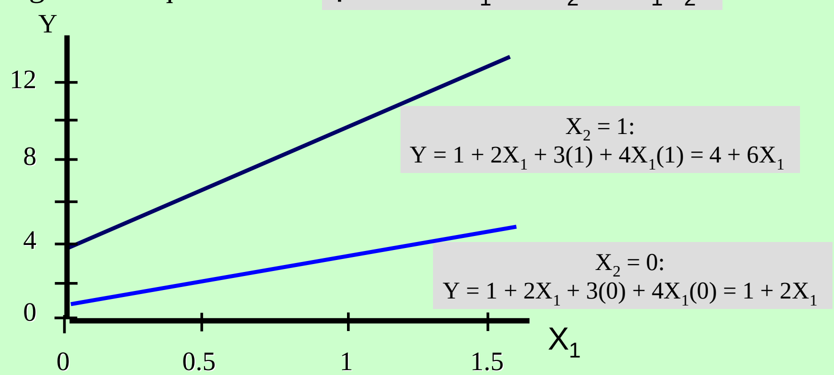
Effect of Interaction

- Given: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$

- Without interaction term, effect of X_1 on Y is measured by β_1
- With interaction term, effect of X_1 on Y is measured by $\beta_1 + \beta_3 X_2$
- Effect changes as X_2 changes

Interaction Example

Suppose X_2 is a dummy variable and the estimated regression equation is $\hat{Y} = 1 + 2X_1 + 3X_2 + 4X_1X_2$



Slopes are different if the effect of X_1 on Y depends on X_2 value

Significance of Interaction Term

- Can perform a partial F-test for the contribution of a variable to see if the addition of an interaction term improves the model

- Multiple interaction terms can be included
  - Use a partial F-test for the simultaneous contribution of multiple variables to the model

Simultaneous Contribution of Independent Variables

- Use partial F-test for the simultaneous contribution of multiple variables to the model
  - Let m variables be an additional set of variables added simultaneously
  - To test the hypothesis that the set of m variables improves the model:

$$F = \frac{[\text{SSR(all)} - \text{SSR (all except new set of m variables)}] / \text{m}}{\text{MSE(all)}}$$

(where F has m and n-k-1 d.f.)

Lecture Summary

In this lecture, we have

- Developed the multiple regression model
- Tested the significance of the multiple regression model
- Discussed adjusted r^2
- Discussed using residual plots to check model assumptions

Lecture Summary

In this lecture, we have

- Tested individual regression coefficients
- Tested portions of the regression model
- Used dummy variables
- Evaluated interaction effects

Some Special Topics

The F Test of a Multiple Regression Model

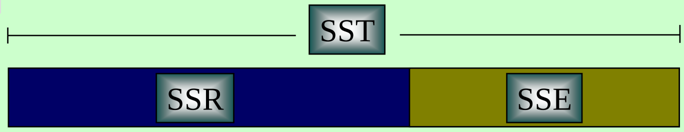
A statistical test for the existence of a linear relationship between Y and any or all of the independent variables X_1, x_2, ..., X_k:

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

H_1: Not all the β_i (i=1,2,...,k) are equal to 0

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Regression	SSR	k	$MSR = \frac{SSR}{k}$	
Error	SSE	n - (k+1)	$MSE = \frac{SSE}{(n-(k+1))}$	
Total	SST	n-1	$MST = \frac{SST}{(n-1)}$	

Decomposition of the Sum of Squares and the Adjusted Coefficient of Determination



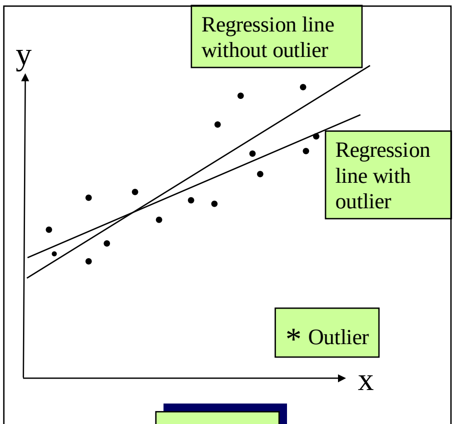
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

The **adjusted multiple coefficient of determination**, $\overline{R}^2$, is the coefficient of determination with the SSE and SST divided by their respective degrees of freedom:

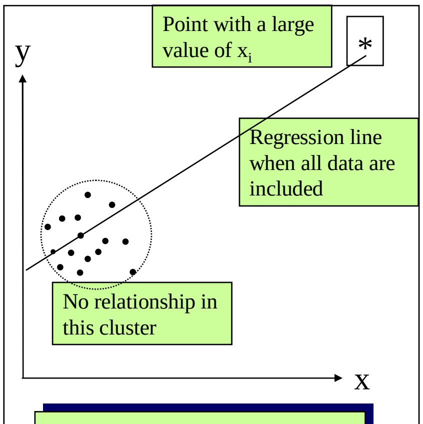
$$\overline{R}^2 = 1 - \frac{\frac{SSE}{(n-(k+1))}}{\frac{SST}{(n-1)}}$$

Example : s = 1.911 R-sq = 96.1% R-sq(adj) = 95.0%

Investigating the Validity of the Regression: Outliers and Influential Observations

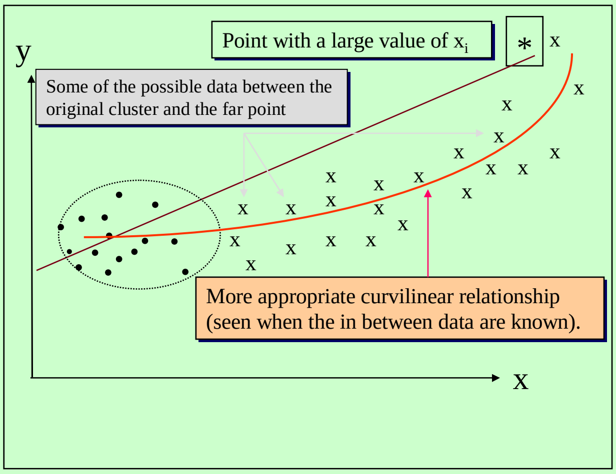


Outliers



Influential Observations

Possible Relation in the Region between the Available Cluster of Data and the Far Point



Prediction in Multiple Regression

A (1-α) 100% prediction interval for a value of Y given values of X_i:

$$\hat{y} \pm t_{(\frac{\alpha}{2},(n-(k+1)))}\sqrt{s^2(\hat{y})+MSE}$$

A (1-α) 100% prediction interval for the conditional mean of Y given values of X_i:

$$\hat{y} \pm t_{(\frac{\alpha}{2},(n-(k+1)))} s[\hat{E}(Y)]$$

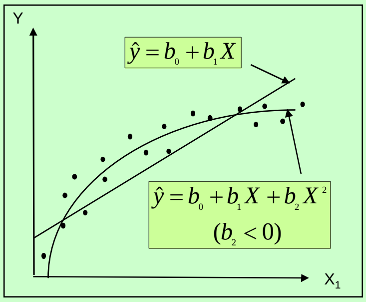
19	Prediction Interval							
20								
21	Given X		M1	Lend	Price	Exch.		
22			5	10	150	2		
23								
24		1-α	(1-α) P.I. for Y for given X			1-α	(1-α) P.I. for E[Y | X]	
25		95%	3.88621	+ or -	0.8509	95%	3.8862 + or -	0.523
26								

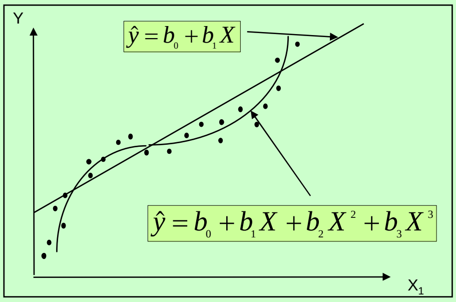
Polynomial Regression

One-variable polynomial regression model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots + \beta_m X^m + \varepsilon$$

where *m* is the *degree* of the polynomial - the highest power of X appearing in the equation. The degree of the polynomial is the ***order*** of the model.





Nonlinear Models and Transformations

The **multiplicative model**:

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} X_3^{\beta_3} \varepsilon$$

The **logarithmic transformation**:

$$\log Y = \log \beta_0 + \beta_1 \log X_1 + \beta_2 \log X_2 + \beta_3 \log X_3 + \log \varepsilon$$

	A	B	C	D	E	F	G	H	I	J	K	L
1	Multiple Regression Results					LogSales						
2												
3		0	1	2	3	4	5	6	7	8	9	10
4		Intercept	ogAdver									
5	*b*	1.70082	0.55314									
6	*s(b)*	0.05123	0.03011									
7	*t*	33.2006	18.3727									
8	*p-value*	0.0000	0.0000									
9												
10		VIF	#REF!									
11												
12	ANOVA Table											
13		Source	SS	df	MS	*F*	$F_{Critical}$	*p*-value				
14		Regn.	4.27217	1	4.2722	337.56	4.3808	0.0000	*s*	0.1125		
15		Error	0.24047	19	0.0127							
16		Total	4.51263	20		R^2	0.9467		Adjusted R^2	0.9439		
17												
18												

Transformations: Exponential Model

The **exponential model:**

$$Y = \beta_0 e^{\beta_1 X} \varepsilon$$

The **logarithmic transformation:**

$$\log Y = \log \beta_0 + \beta_1 X_1 + \log \varepsilon$$

Multiple Regression Results

	0	1	2	3	4	5	6	7	8	9	10
	Intercept	LogAdver									
b	3.66825	6.784									
s(b)	0.40159	0.23601									
t	9.13423	28.7443									
p-value	0.0000	0.0000									
VIF		#REF!									

ANOVA Table

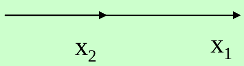
Source	SS	df	MS	F	$F_{Critical}$	p-value
Regn.	642.622	1	642.62	826.24	4.3808	0.0000
Error	14.7777	19	0.7778			
Total	657.4	20				

s 0.8819

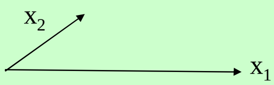
R^2 0.9775 Adjusted R^2 0.9763

Multicollinearity

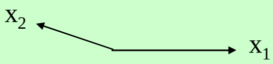
Orthogonal X variables provide information from independent sources. **No multicollinearity.**



Perfectly collinear X variables provide identical information content. **No regression.**



Some degree of collinearity. Problems with regression depend on the degree of collinearity.



A high degree of **negative collinearity** also causes problems with regression.

Effects of Multicollinearity

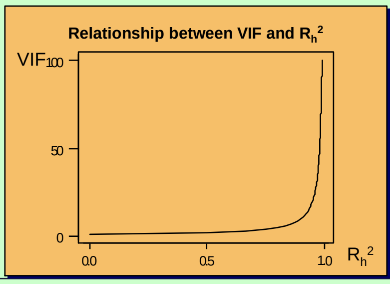
- Variances of regression coefficients are inflated.
- Magnitudes of regression coefficients may be different from what are expected.
- Signs of regression coefficients may not be as expected.
- Adding or removing variables produces large changes in coefficients.
- Removing a data point may cause large changes in coefficient estimates or signs.
- In some cases, the *F* ratio may be significant while the *t* ratios are not.

Variance Inflation Factor

The **variance inflation factor** associated with X_h:

$$VIF(X_h) = \frac{1}{1 - R_h^2}$$

where R_h^2 is the R^2 value obtained for the regression of X on the other independent variables.



Variance Inflation Factor (VIF)

Multiple Regression Results — Exports

	0	1	2	3	4	5	6	7	8	9	10
	Intercept	M1	Lend	Price	Exch.						
b	-4.0155	0.36846	0.0047	0.0365	0.2679						
s(b)	2.7664	0.06385	0.0492	0.0093	1.1754						
t	-1.4515	5.7708	0.0955	3.9149	0.2279						
p-value	0.1517	0.0000	0.9242	0.0002	0.8205						
VIF		3.2072	5.3539	6.2887	1.3857						

Observation: The VIF (Variance Inflation Factor) values for both variables Lend and Price are both greater than 5. This would indicate that some degree of multicollinearity exists with respect to these two variables.

Partial *F* Tests and Variable Selection Methods

Full model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Reduced model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Partial *F* test:

H_0: $\beta_3 = \beta_4 = 0$

H_1: β_3 and β_4 not both 0

Partial F statistic:

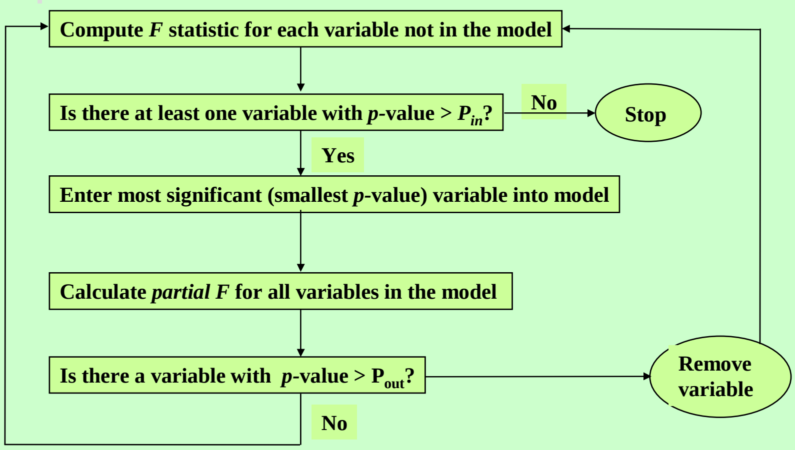
$$F_{(r,(n-(k+1))} = \frac{(SSE_R - SSE_F)/r}{MSE_F}$$

where SSE_R is the sum of squared errors of the reduced model, SSE_F is the sum of squared errors of the full model; MSE_F is the mean square error of the full model [$MSE_F = SSE_F/(n-(k+1))$]; *r* is the number of variables dropped from the full model.

Variable Selection Methods

- Stepwise procedures
  - ✓ Forward selection
    - Add one variable at a time to the model, on the basis of its *F* statistic
  - ✓ Backward elimination
    - Remove one variable at a time, on the basis of its *F* statistic
  - ✓ Stepwise regression
    - Adds variables to the model and subtracts variables from the model, on the basis of the *F* statistic

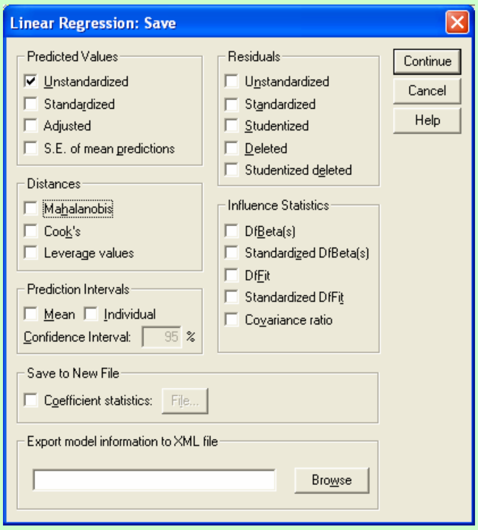
Stepwise Regression



Influential Points

- Outliers (univariate, multivariate)
- Leverage Points (Distances)
- Influence Statistics

Influential Points continued…



Distances

- **Mahalanobis**: A measure of how much a case's values on the independent variables differ from the average of all cases. A large Mahalanobis distance identifies a case as having extreme values on one or more of the independent variables.
- **Cook’s**: A measure of how much the residuals of all cases would change if a particular case were excluded from the calculation of the regression coefficients. A large Cook's D indicates that excluding a case from computation of the regression statistics, changes the coefficients substantially.
- **Leverage values**: Measures the influence of a point on the fit of the regression. The centered leverage ranges from 0 (no influence on the fit) to (N-1)/N.

Influence Statistics (1)

- **DfBeta(s):** The difference in beta value is the change in the regression coefficient that results from the exclusion of a particular case. A value is computed for each term in the model, including the constant.
- **Std. DfBeta(s):** Standardized difference in beta value. The change in the regression coefficient that results from the exclusion of a particular case. You may want to examine cases with absolute values greater than 2 divided by the square root of N, where N is the number of cases. A value is computed for each term in the model, including the constant.
- **DfFit:** The difference in fit value is the change in the predicted value that results from the exclusion of a particular case.

Influence Statistics (2)

- **Std. DfFit:** Standardized difference in fit value. The change in the predicted value that results from the exclusion of a particular case. You may want to examine standardized values which in absolute value exceed 2 divided by the square root of p/N, where p is the number of independent variables in the equation and N is the number of cases.
- **Covariance Ratio:** The ratio of the determinant of the covariance matrix with a particular case excluded from the calculation of the regression coefficients to the determinant of the covariance matrix with all cases included. If the ratio is close to 1, the case does not significantly alter the covariance matrix.

Bibliography

- **Steel, R. & Torrie, J. (1986). *Principles and Procedures of Statistics: A Biometrical Approach*. Singapore: McGraw-Hill Book Company.**
- **Gomez, K. & Gomez, A. (1984). *Statistical Procedures for Agricultural Research*. Singapore: John Willey & Sons, Inc.**
- **Kuehl, R. (2000). *Designs of Experiments: Statistical Principles of Research Design and Analysis*. Pacific Grove: Duxbury Thomson Learning.**
- **Jacoby, W. (2000). Loess: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies*, 19, 577-613.**
- **Zar, J. (1996). *Biostatistical Analysis*. New Jersey: Prentice-Hall International, Inc.**
- **Kirk, R. (1995). *Experimental Design: Procedures for the Behavioral Sciences*. Pacific Grove: Brooks/Cole Publishing Company.**
- **Kleinbaum, D., Kupper, L., Muller, K. & Nizam, A. (1998). *Applied Regression Analysis and Other Multivariable Methods*. Pacific Grove: Duxbury Press.**


Thank you very much for your attention