Review on Linear Regression Analysis

Example: City X has recently conducted a travel survey of its 10 zones. The collected data are summarized in Table 1. Based on the given information,

- How to predict the total number of trips that will be produced by each zone in 10 years (assume zonal population in 10 years is known)?
- How to relate Y (total trips produced by a zone) to X_1 (zonal population)?

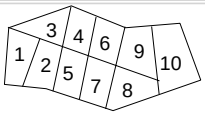


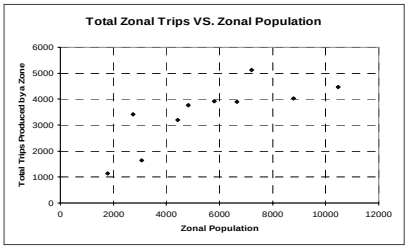
Table 1 Observed Trips and Zonal Activity

Zone #	Pop. X_1	Households X_2	Trips Y
1	7212	2488	5126
2	4818	2188	3764
3	8789	2423	4030
4	5805	2141	3921
5	3054	1241	1644
6	10463	3451	4467
7	2735	1857	3407
8	1784	905	1143
9	4418	1695	3202
10	6657	1960	3900

Our First Attempt (informal):

City X

- To answer the questions, we need to relate Y to the Xs!
- Plot the observed Y vs X_1 (scatter chart) and then Add a *Trend Line*



Outline:

- Simple Linear Regression Analysis
- Multiple Linear Regression
- Understanding the Regression Outputs
- Validating a Regression Model

Simple Linear Regression

- **Problem***: consider two variables, X and Y, and we have n paired observations on them (data): (x_1,y_1), (x_2,y_2), …, (x_n,y_n), what is the relationship between Y and X?*
- **Method***: Assumed Y and X are **linearly** related, that is*

$$Y = b_0 + b_1 X$$

where: ***X - independent*** variable (a factor that is considered to have impact on Y!)

Y - dependent variable

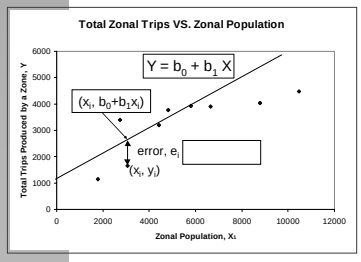
b_0, b_1 - ***regression coefficients***

- **Our Goals:**
  - Find the ***best*** value for b_0 and b_1 (but what do we mean by "best"?)
  - Find out whether or not the identified relationship is good, i.e., validate the model

Our Second Attempt (formal):

City X

- Plot Y vs. X_1 (scatter graph) and the straight line $Y = b_0 + b_1 X_1$ with different intercept (b_0) and slope (b_1)



We shall find b_0 and b_1 so as to *minimize* the total estimation error (E):

$E = \Sigma e_i^2$
$= \Sigma [y_i-(b_0+b_1x_i)]^2$

This is called the ***method of least squares***).

Results of Our Second Attempt:

City X

- b_0 and b_1 can be determined mathematically as follows:

Estimating the Regression Coefficients

- Data: $(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)$
- From data we can calculate the following statistics

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad , \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{YY} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$S_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad S_{Y|X} = \sqrt{\frac{S_{YY} - S_{XY}^2 / S_{XX}}{n-2}}$$

- b_0 and b_1 can then be determined by

$$b_0 = \bar{y} - b_1 \bar{x}, \qquad b_1 = \frac{S_{XY}}{S_{XX}}$$

Our Third Attempt (using Excel):

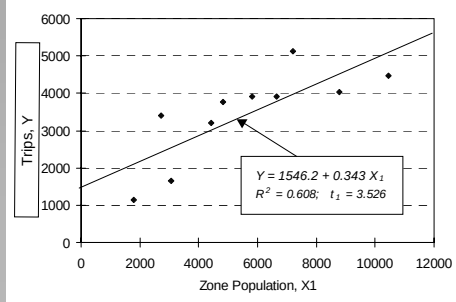
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.780
R^2	**0.608**
Adjusted R^2	0.560
Standard Error	808.122
Observations	10

	Coefficients	*Standard Error*	*t Stat*
Intercept	**1546.214**	600.034	2.577
X Variable 1	**0.343**	0.097	**3.526**

b_0 b_1

Scatter Plot and Regression Line



Multiple Regression

- Example 1 (Continued): it seems that the total number of trips produced by a zone should also be related the households of that zone (in addition to population), can we relate Y to both X_1(zonal population) and X_2(zonal households)?
- In general:
  - Data:$(y_1,x_{11},x_{12}, \ldots), (y_2,x_{21},x_{22},\ldots), \ldots, (y_n,x_{n1},x_{n2},\ldots)$
  - Assumed Relationship

    $$Y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_r x_r$$

  - Problem: how to find coefficients b_0, b_1, b_2 ,…
- The *principle* to solve the problem is the same as for the simple linear regression, that is, the ***method of least squares***.

Multiple Regression Using Excel

City X

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.891
R^2	**0.794**
Adjusted R^2	0.735
Standard Error	627.121
Observations	10

	Coefficients	*Standard Error*	*t Stat*
Intercept	**491.361**	637.458	0.771
X Variable 1	**0.804**	0.575	**1.398**
X Variable 2	**0.552**	0.398	**1.385**

Resulting Equation: $Y = 491.361 + 0.804\ X_1 + 0.552\ X_2$

Understanding Regression Output

- ***We obtained $Y = b_0 + b_1 X$, but***
  - how good is the final regression equation?
  - How good are the estimates b_0 and b_1?
  - ...
- *Regression coefficients* are estimates of the "true" population coefficients. If the true relationship is

  $$Y = \beta_0 + \beta_1 x$$

  then b_0 and b_1 are estimates of β_0 and β_1 and a different sample would result in different b_0 and b_1 !!!

Regression Output from Excel

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.7801					
R Square	0.6085					
Adjusted R Square	0.5595					
Standard Error	808.12					
Observations	10					
ANOVA						
	df	*SS*	*MS*	*F*	*gnificance F*	
Regression	1	8119632.892	8119633	12.43318	0.007777	
Residual	8	5224491.608	653061.5			
Total	9	13344124.5				
	Coefficien	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1546.2	600.0341482	2.576876	0.032775	162.5315	2929.896
X Variable 1	0.3435	0.097406453	3.526072	0.007777	0.118842	0.568082

Understand Regression Output (1):

Regression Coefficients, b_0, b_1, ...

Example 1: $Y = 1546.2 + 0.343\ X_1$

$b_0 = 1546.2$ interpreted as: amount of trips would be produced if there is no population in a zone (X_1=0)?!)

$b_1 = 0.343$ interpreted as: one additional person in a zone would produce additional 0.343 trips!)

14

Regression Output from Excel

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.7801					
R Square	0.6085					
Adjusted R Square	0.5595					
Standard Error	808.12					
Observations	10					
ANOVA						
	df				*gnificance F*	
Regression	1	81		18	0.007777	
Residual	8	5224491.608	653061.5			
Total	9	13344124.5				
	Coefficien	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1546.2	600.0341482	2.576876	0.032775	162.5315	2929.896
X Variable 1	0.3435	0.097406453	3.526072	0.007777	0.118842	0.568082

This is the S_{yix} value !

***This is the "Standard error (deviation) of regression coefficient* of b_1" value !**

Understand Regression Output:

Confidence Interval and Hypothesis Testing for b_i

- A (1- α)100% confidence interval for the parameter β_1 in the true regression line
- Test the hypothesis H_0: $\beta_1 = 0$. If this hypothesis is *not* rejected, then X_1 is *not* a significant factor influencing Y and should not be included in the function!

If $|t| < t_{\alpha/2,\, n-k-1}$, reject H_0 at (1- α)100% confidence level

17

Regression Output from Excel

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.7801					
R Square	0.6085					
Adjusted R Square	0.5595					
Standard Error	808.12					
Observations	10					
ANOVA						
			MS	*F*	*gnificance F*	
Regression			8119633	12.43318	0.007777	
Residual	8	5224491.608	653061.5			
Total	9	13344124.5				
	Coefficien	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1546.2	600.0341482	2.576876	0.032775	162.5315	2929.896
X Variable 1	0.3435	0.097406453	3.526072	0.007777	0.118842	0.568082

This is the 95% confidence interval of b_1

***This is the t* value !**

Understand Regression Output (3):

Coefficient of Determination, R^2

- R^2 is a measure of overall quality of the regression, specifically, it is the *percentage* of total variation in Y that is accounted by the sample regression line.
- A high value of R^2 means that most of the variation we observed in Y (y_i) can be *attributed* to their corresponding X values
- There is however NO standard on how high a R^2 value is "good" enough - it depends on the application! In transportation planning, a R^2 score over 0.3 is often acceptable!

19

Regression Output from Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.7801
R Square	0.6085
Adjusted R Square	0.5595
Standard Error	808.12
Observations	10

This is the R^2 value !

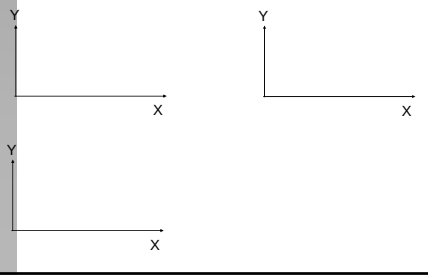
ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	1	8119632.892	8119633	12.43318	0.007777
Residual	8	5224491.608	653061.5		
Total	9	13344124.5			

	Coefficien	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1546.2	600.0341482	2.576876	0.032775	162.5315	2929.896
X Variable 1	0.3435	0.097406453	3.526072	0.007777	0.118842	0.568082

Understand Regression Output (3):
Coefficient of Determination, R^2

- R^2 is a measure of overall quality of the regression: Examples



21

Example 1 (continued):

- Regression Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.780
R^2	**0.608**
Adjusted R^2	0.560
Standard Error	808.122
Observations	10

	Coefficients	*Standard Error*	*t Stat*
Intercept	**1546.214**	600.034	2.577
X Variable 1	**0.343**	0.097	**3.526**

- Questions:
  - R^2 ?
  - Is variable X_1(population) significant at 95% confidence level?
  - What is 95% confidence interval of the slope (β_1)?

22

Validate Regression Model

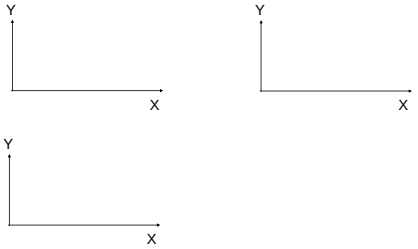
(pitfalls in regression analysis!)

- Important Assumptions Used in Linear Regression Analysis
  - Y is *linearly* related to x
  - For any given x, Y is *Normally* distributed
  - Residual Error has *constant* standard deviation (for all x)
  - Independent variables are NOT *linearly* related (multiple regression only!)
- How to find out whether or not these conditions are met?

23

Check for Linearity

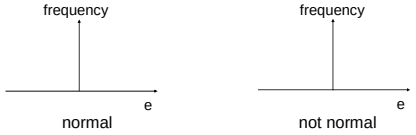
- Objective: is Y really linearly related to x ?
- Method: check visually from scatter plot



24

Check for Normality

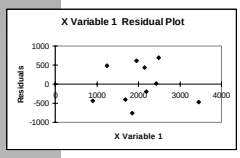
- Objective: is the residual normally distributed (e_i) ?
- Method: plot a histogram of the residuals $[e_i = y_i - (b_0 + b_1 x_i)]$. (note that a more formal procedure is available, but will not be covered in this course!)



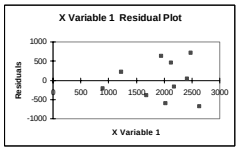
25

Check for Homogeneous Variance

- Objective: is the standard deviation of the residual (e) constant across all X values?
- Method: check scatter plot of the residuals VS. each X variable



homogeneous variance

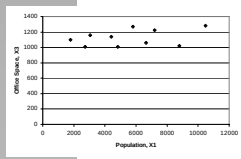


non-homogeneous variance

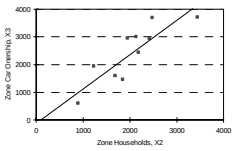
26

Check for Multicolinearity

- Objective: is independent variable X_i linearly correlated to independent variable X_j?
- Method: check scatter plot of X_i vs. X_j



No evidence of multicolinearity



Evidence of multicolinearity

27

Check for Multicolinearity - cont.

- Sample Correlation Coefficient (r): is a measure of the degree of linear-relationship between two variables.
- $-1 \le r \le 1$:
  - if r = +1, X_1 and X_2 have a perfect *positive* correlation
  - if r = -1, X_1 and X_2 have a perfect *negative* correlation
  - if r = 0, X_1 and X_2 have *no* linear relationship
  - if $|r| > 0.4$ for two independent variables, it may run into the multicolinearity problem if both variables are included in a regression equation.
- Correlation Coefficient can also be directly obtained using Excel function

28