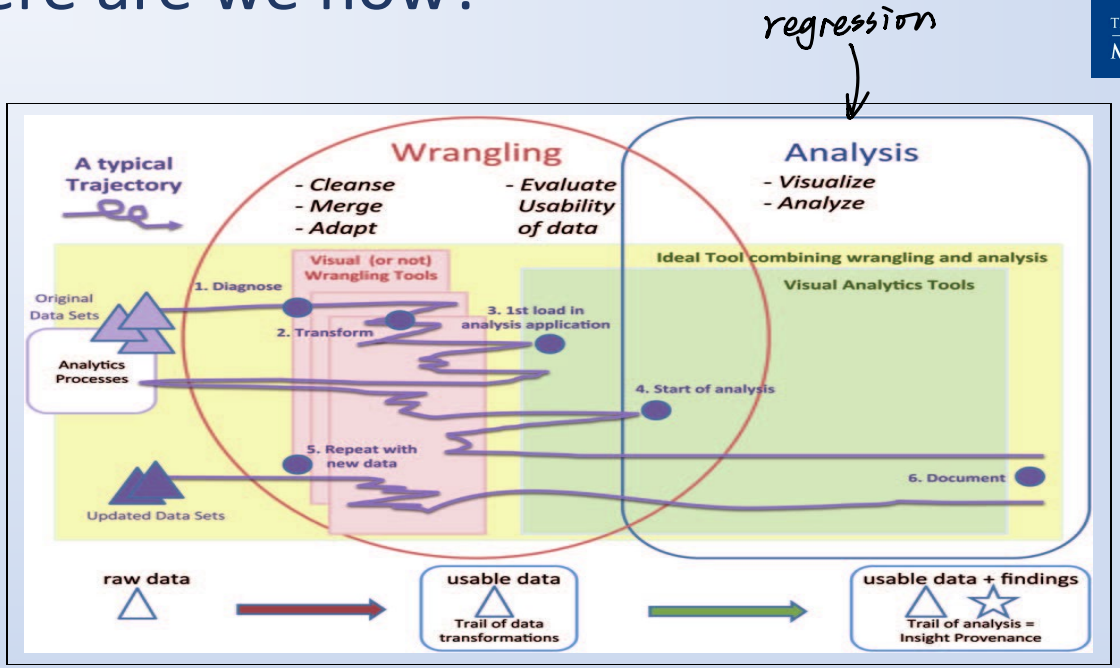
COMP20008 Elements of Data Processing

Regression

Semester 1, 2020

Contact: uwe.aickelin@unimelb.edu.au

Where are we now?



Learning Objectives

- How to use regression analysis to predict the value of a dependent variable based on independent variables
- Make inferences about the slope and correlation coefficient
- Evaluate the assumptions of regression analysis and know what to do if the assumptions are violated

Correlation vs. Regression

- A scatter diagram can be used to show the relationship between two variables
- Correlation analysis is used to measure strength / direction of the relationship between two variables
- No causal effect is implied with correlation

correlation { strength / direction } of the relationship
≠ the cause of relationship

Regression { dependent / independent

the dependent variable will change because of the independent variable
⇒ assume the cause of relationship

Introduction to Regression Analysis

Regression analysis is used to:

- Predict the value of a dependent variable based on the value of at least one independent variable
- Explain the impact of changes in an independent variable on the dependent variable
- Dependent variable: the variable we wish to predict or explain
- Independent variable: the variable used to explain the dependent variable

$y = f(x)$.
y depends on x
Y is dependent
X is independent

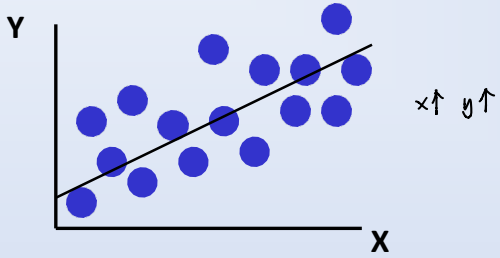
Linear Regression

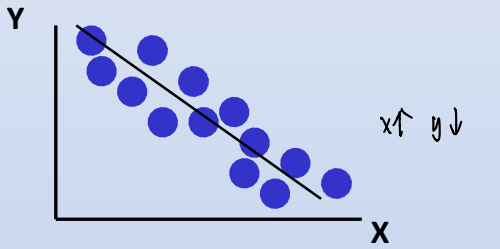
Simple Linear Regression Model

- Only one independent variable, X
- Relationship between X and Y is described by a linear function
- Changes in Y are assumed to be caused by changes in X

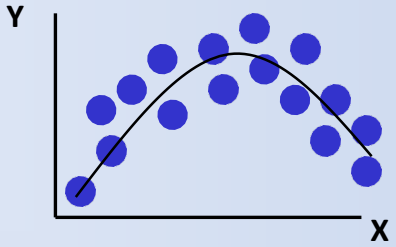
Types of Relationships

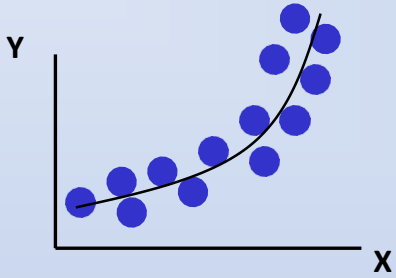
Linear relationships





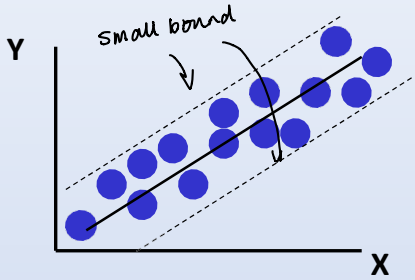
Non-linear relationships

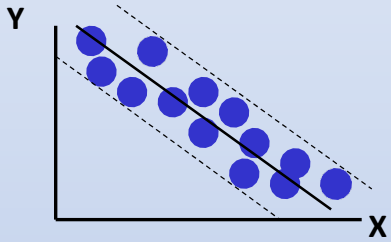




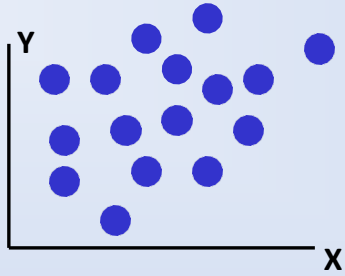
Types of Relationships (2)

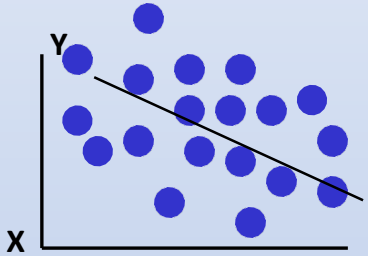
Strong relationships



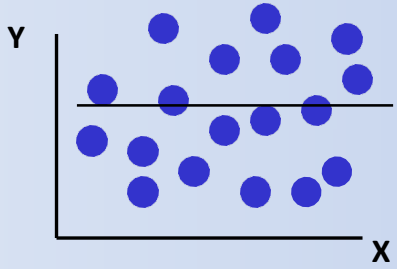


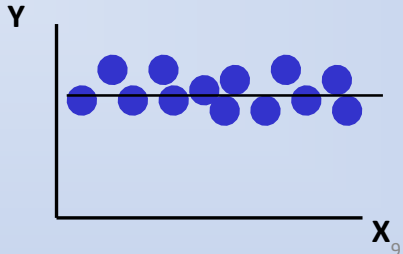
Weak relationships





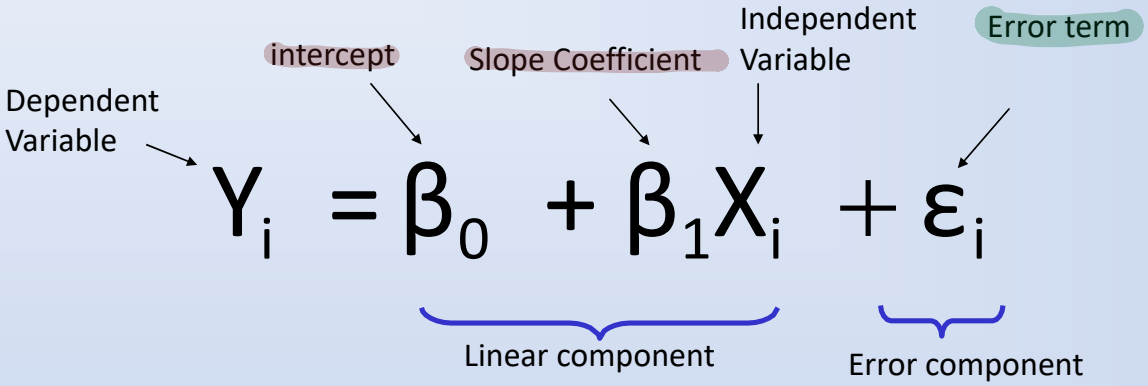
No relationship



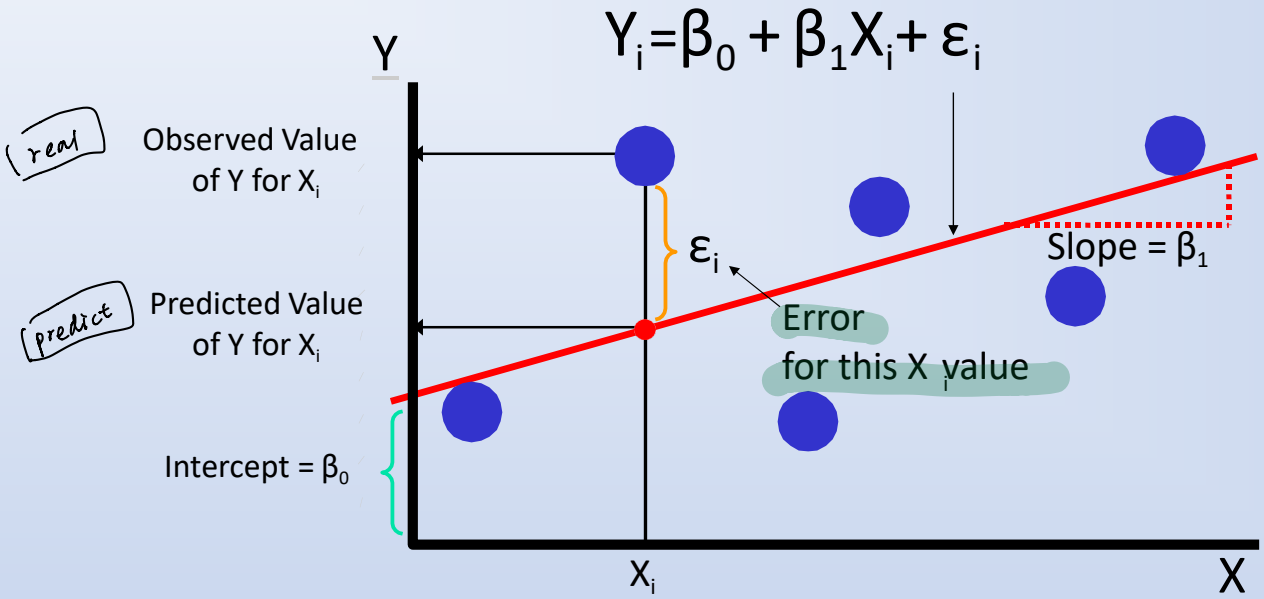


Simple Linear Regression Model

The simple linear regression equation provides an estimate of the population regression line



Simple Linear Regression Model (2)



Least Squares Method

b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared differences between Y and $\hat{Y}$

min (∑ (real y_i - predict y_i)^2).

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

Interpretation of Slope and Intercept

- b_0 is the estimated average value of Y when the value of X is zero (intercept)

- b_1 is the estimated change in the average value of Y as a result of a one-unit change in X (slope)

x change by 1 unit, what happens to y

the strength of the relationship → 1 to 1, 1 to 2, ... ?

Simple Linear Regression Example

A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

A random sample of 10 houses is selected

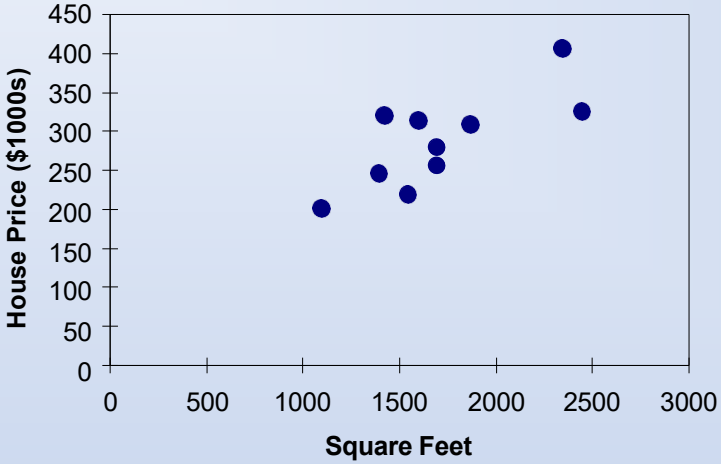
- Dependent variable (Y) = house price in $1000s
- Independent variable (X) = square feet

Sample Data for House Price Model

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Graphical Presentation

House price model: scatter plot



Calculation Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

"^" means predicted value

without "^" → actual value

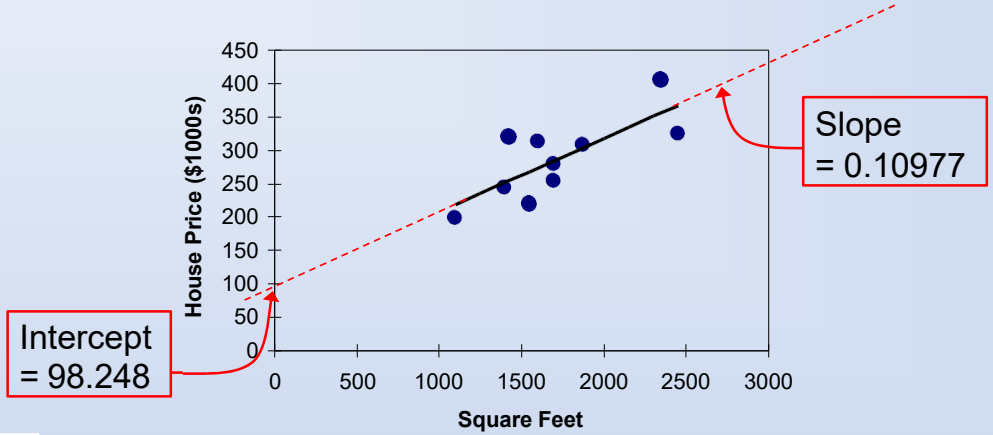
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Graphical Presentation

House price model: scatter plot and regression line



$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

Interpretation of the Intercept b_0

$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

- b_0 is the estimated average value of Y when the value of X is zero
- Here, no houses had 0 square feet, so b_0 = 98.24833 ($1000) just indicates that, for houses within the range of sizes observed, $98,248.33 is the portion of the house price not explained by square feet

Interpretation of the Slope Coefficient b_1

$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
- Here, b_1 = .10977 tells us that the average value of a house increases by .10977 ($1000) = $109.77, on average, for each additional one square foot of size

useful when extending the house
$ you want to pay for the extension of houses
by 1 square foot of size

Predictions using Regression

Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

$$= 98.25 + 0.1098\,(2000)$$

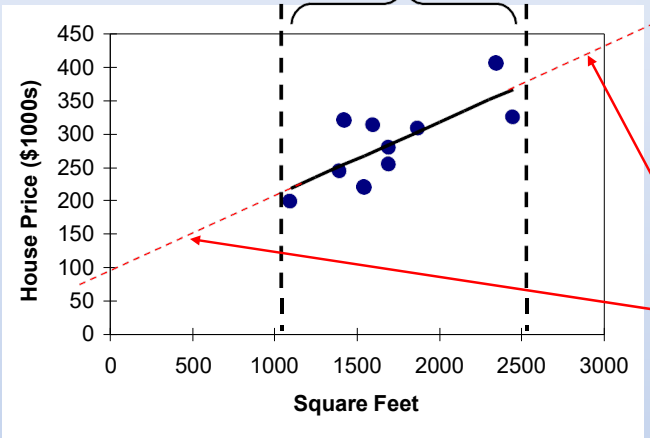
$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85 ($1,000s) = $317,850

Interpolation vs. Extrapolation 推断

When using a regression model for prediction, only predict within the relevant range of data

Relevant range for interpolation



we can't predict beyond the range

Do not try to extrapolate beyond the range of observed X's

Regression – Residual Analysis

Measures of Variation

Total variation is made up of two parts:

ERROR ↓ better regression

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares

$$SST = \sum (Y_i - \bar{Y})^2 \quad SSR = \sum (\hat{Y}_i - \bar{Y})^2 \quad SSE = \sum (Y_i - \hat{Y}_i)^2$$

where:

$\bar{Y}$ = Average value of the dependent variable

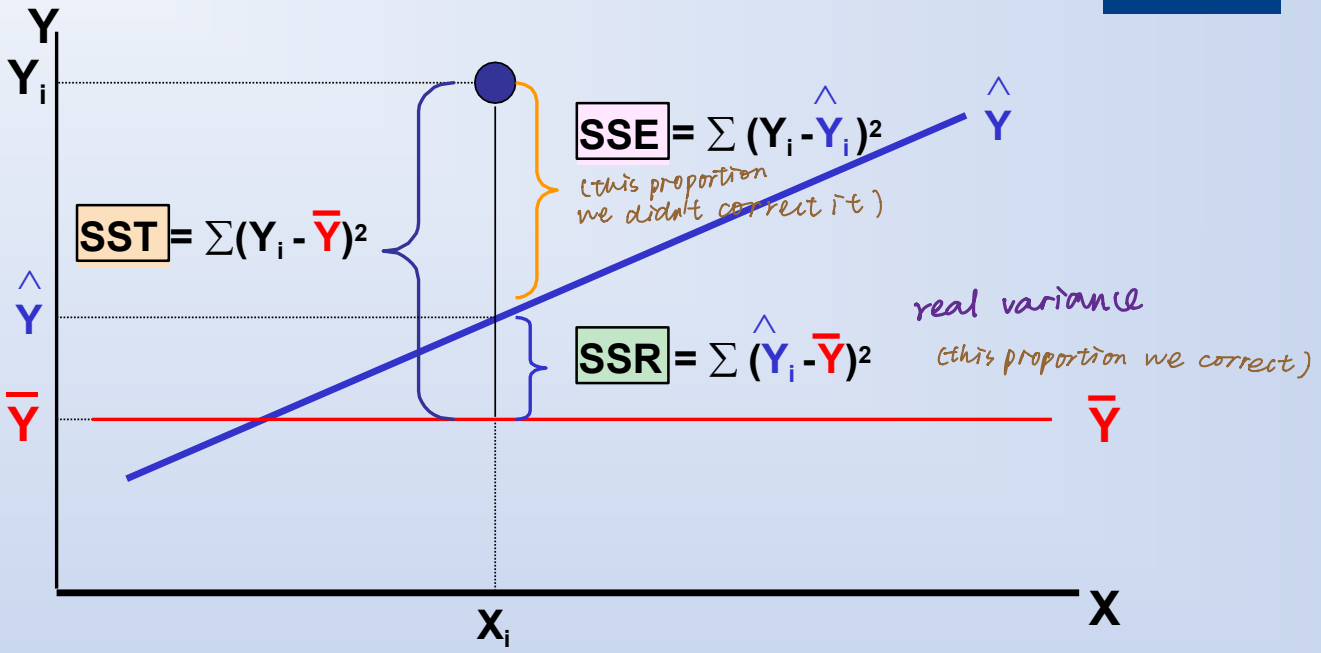
Y_i = Observed values of the dependent variable

$\hat{Y}_i$ = Predicted value of Y for the given X_i value

Measures of Variation (2)

- SST = total sum of squares
  - Measures the variation of the Y_i values around their mean Y
- SSR = regression sum of squares
  - Explained variation attributable to the relationship between X and Y
- SSE = error sum of squares
  - Variation attributable to factors other than the relationship between X and Y

Measures of Variation (3)



Coefficient of Determination, r^2

- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called r-squared and is denoted as r^2

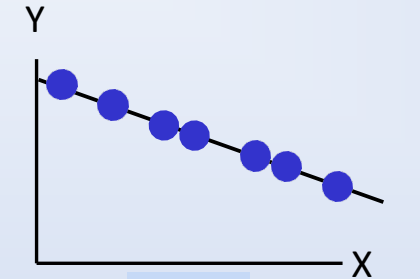
r^2 = SSR / SST

= regression sum of squares / total sum of squares

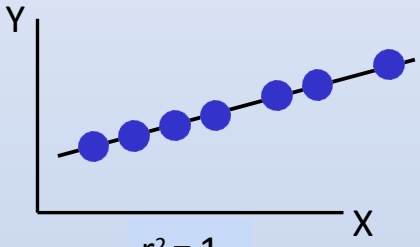
if SSR is large to SST → a good job!
↑ almost 1

note: $0 \leq r^2 \leq 1$

Examples of approximate r^2 values



$r^2 = 1$



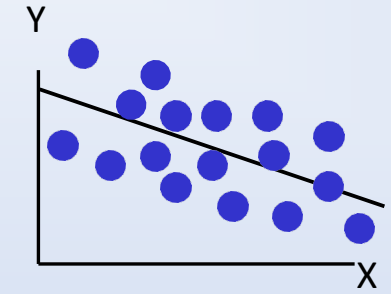
$r^2 = 1$

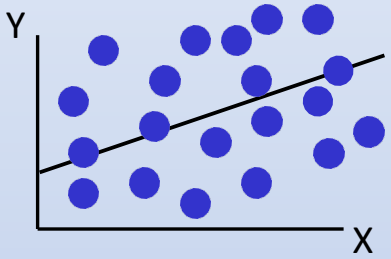
$r^2 = 1$

Perfect linear relationship between X and Y:

100% of the variation in Y is explained by variation in X

Examples of approximate r^2 values (2)



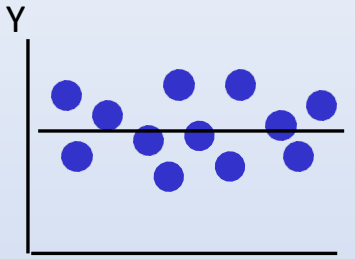


$0 < r^2 < 1$

Weaker linear relationships between X and Y:

Some but not all of the variation in Y is explained by variation in X

Examples of approximate r^2 values (3)



$r^2 = 0$

$r^2 = 0$

No linear relationship between X and Y:

The value of Y does not depend on X.

Calculation Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$r^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Assumptions of Regression

- Linearity: The underlying relationship between X and Y is linear
- Independence of residuals: Residual values (also known as fitting errors) are statistically sound and sum to zero

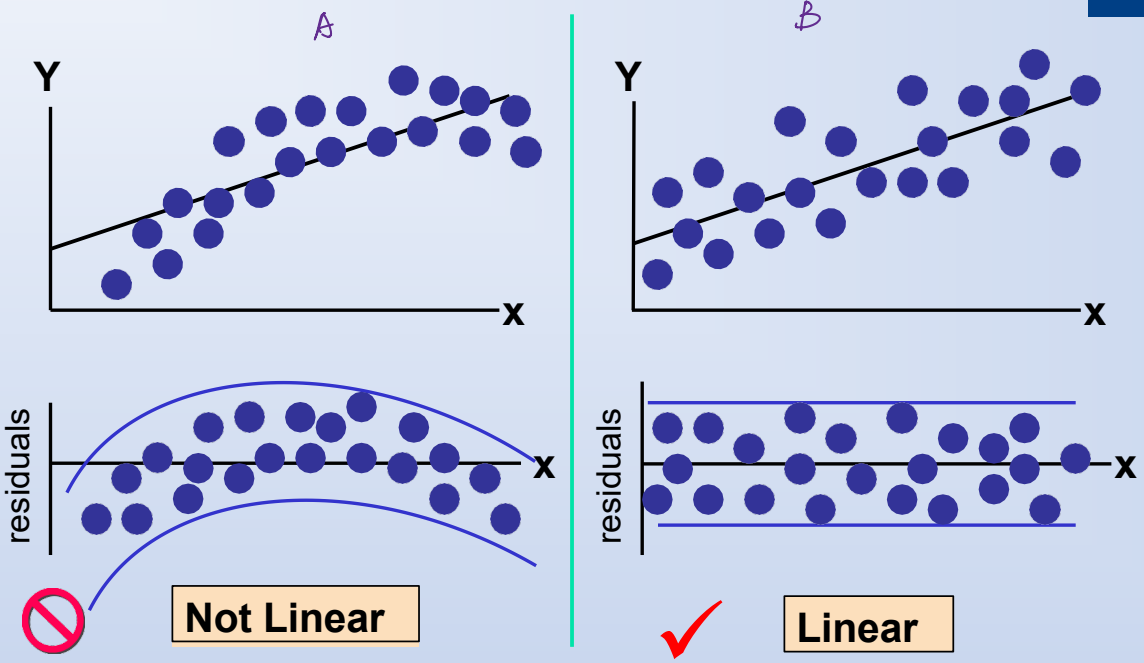
Residual Analysis

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

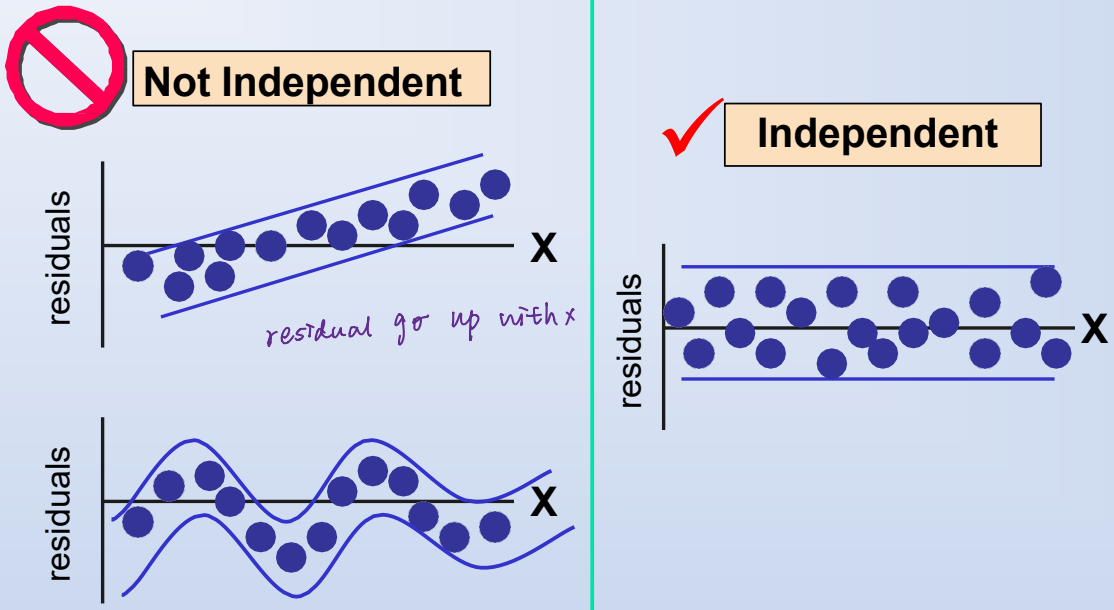
- The residual for observation i, e_i, is the difference between the observed and predicted value
- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Evaluate independence assumption — error should be independent
  - Examine for constant variance for all levels of X — error should not change with level of X
- Graphical Analysis of Residuals: plot residuals vs. X

Residual Analysis for Linearity

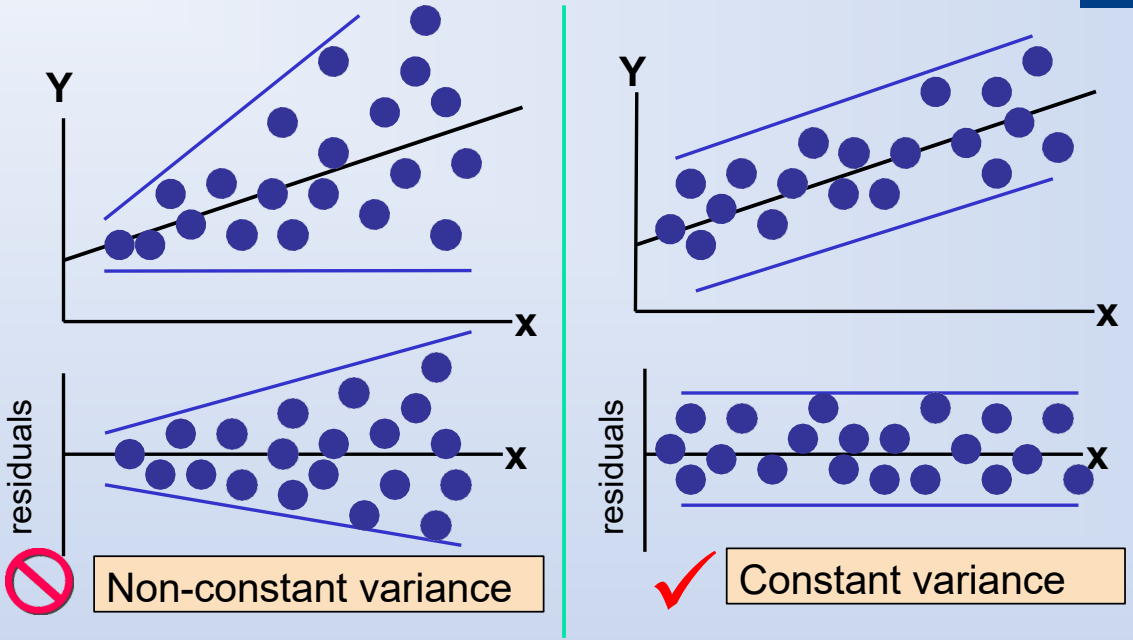
quality of regression of B is better than A



Residual Analysis for Independence

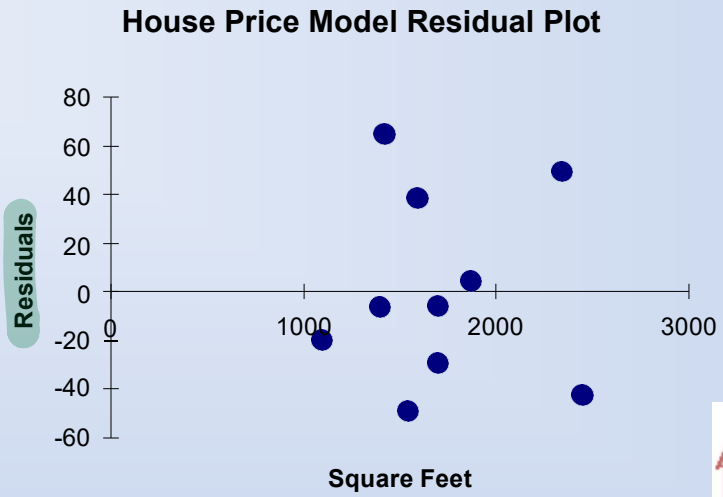


Residual Analysis for Equal Variance



House Price Residual Output

RESIDUAL OUTPUT		
	Predicted House Price	*Residuals*
1	251.92316	-6.923162
2	273.87671	38.12329
3	284.85348	-5.853484
4	304.06284	3.937162
5	218.99284	-19.99284
6	268.38832	-49.38832
7	356.20251	48.79749
8	367.17929	-43.17929
9	254.6674	64.33264
10	284.85348	-29.85348



Does not appear to violate any regression assumptions

Multiple Regression

Multiple Regression

- Multiple regression is an extension of simple linear regression
- It is used when we want to predict the value of a variable based on the value of two or more other variables
- The variable we want to predict is called the dependent variable
- The variables we are using to predict the value of the dependent variable are called the independent variables

Multiple Regression Example

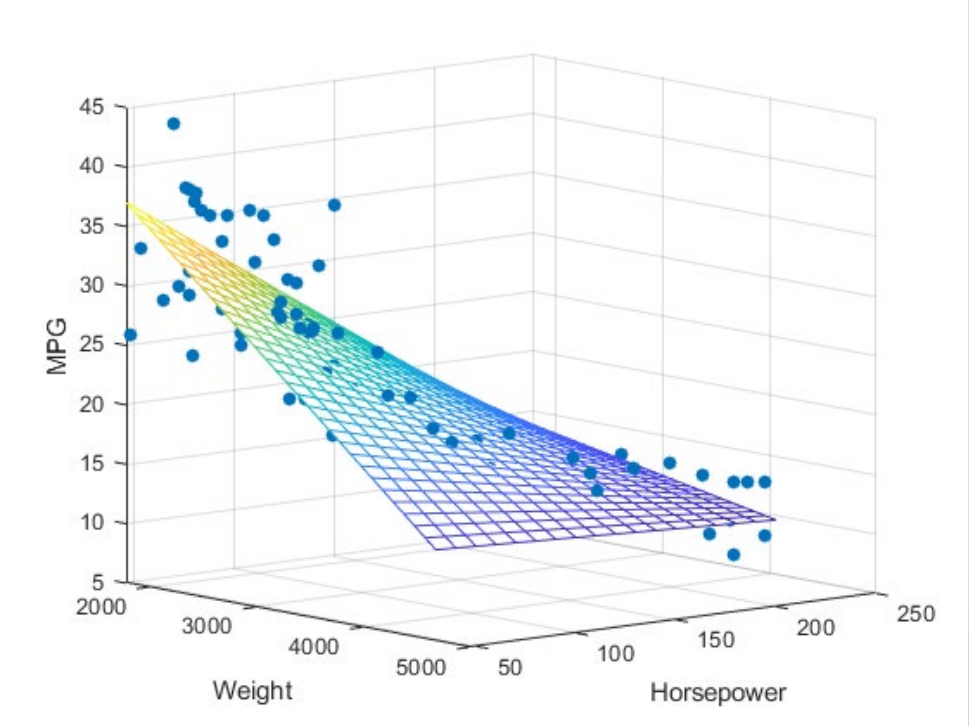
A researcher may be interested in the relationship between the weight of a car, engine size of a car and petrol consumption.

- Independent Variable 1: weight
- Independent Variable 2: horsepower
- Dependent Variable: miles per gallon

Multiple Regression Fitting

- Linear regression is based on fitting a line as close as possible to the plotted coordinates of the data on a two-dimensional graph
- Multiple regression with two independent variables is based on fitting a plane as close as possible to the plotted coordinates of your data on a three-dimensional graph: $Y = a + b_1X_1 + b_2X_2$
- More independent variables extend this into higher dimensions
- The plane (or higher dimensional shape) will be placed so that it minimises the distance (sum of squared errors) to every data point

Multiple Regression Graphic



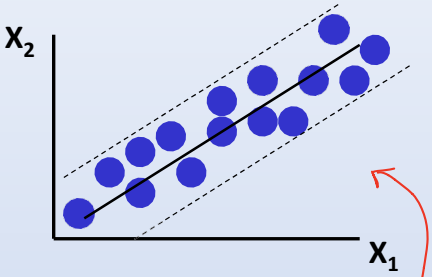
Multiple Regression Assumptions

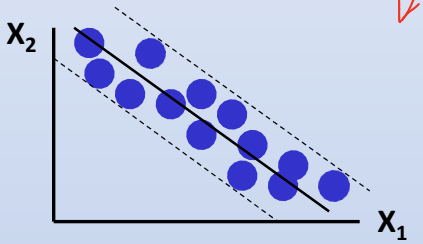
- Multiple regression assumes that the independent variables are not highly correlated with each other
- Use scatter plots to check

otherwise
you will increase the error
put too much weight on the same factor

Multiple Regression Assumption

BAD

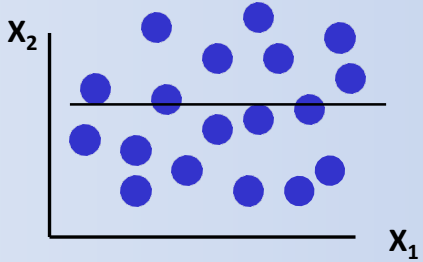


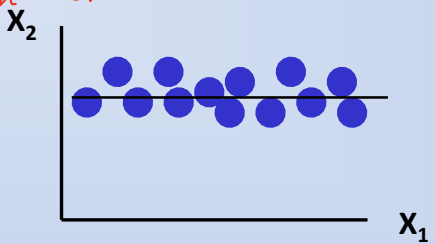


- Multiple regression assumes that the independent variables are not highly correlated with each other
- Use scatter plots to check

we could use a single factor to represent x1 & x2 if they are highly correlated with each other

GOOD





Pitfalls of Regression Analysis

- Lacking an awareness of the assumptions underlying least-squares regression
- Using a regression model without knowledge of the subject matter
- Extrapolating outside the relevant range
- For multiple regression remember the importance of independence assumptions

Avoiding the Pitfalls of Regression

- Start with a scatter plot of X vs. Y to observe possible relationships
- Perform residual analysis to check the assumptions: Plot residuals vs. X to check for violations
- If there is no evidence of assumption violation, test for significance of the regression coefficients
- Avoid making predictions or forecasts outside the relevant range

Acknowledgements

- Materials are partially adopted from …
  - Previous COMP2008 slides including material produced by James Bailey, Pauline Lin, Chris Ewin, Uwe Aickelin and others
  - Sutikno, Department of Statistics, Faculty of Mathematics and Natural Sciences Sepuluh Nopember Institute of Technology (ITS), Surabaya