DEPARTMENT OF
ECONOMETRICS AND BUSINESS STATISTICS

CHUN MINSOO

MONASH University

MONASH BUSINESS SCHOOL

Multiple Linear Regression analysis Part 1

ETW2001 Foundations of Data Analysis and Modeling

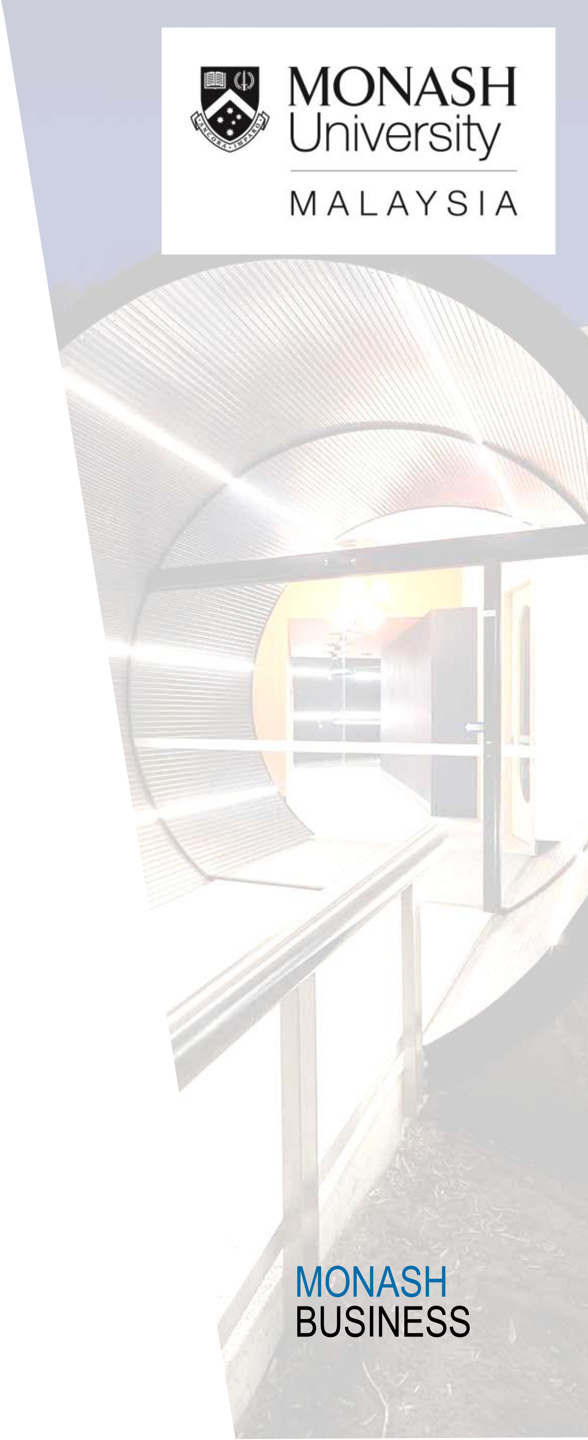
Outline

❑ Quick recap from the previous unit

- ❑ Simple linear regression
- ❑ Multiple linear regression
- ❑ Interpretation of coefficients
- ❑ Inferences

❑ Regression – Part 1

- ❑ Regression with constant
- ❑ Comparison between two groups
- ❑ Linear Probability Model
- ❑ Multicollinearity & Omitted variable bias



Recap: Simple linear regression

Basic format

❑ Population model:

$$y_i = \alpha + \beta_i x_i + u_i, i = 1,2,\ldots,n$$

Where:

y: *dependent varaible or explained variable*

a: *constant*

β_i: *coefficient of* x_i

x_i: *independent variable or explanatory variable*

u_i: *error term*

❑ Estimated model:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_i x_i, i = 1,2,\ldots,n$$

Where:

$\hat{y}$: ***estimated*** *dependent varaible or explained variable*

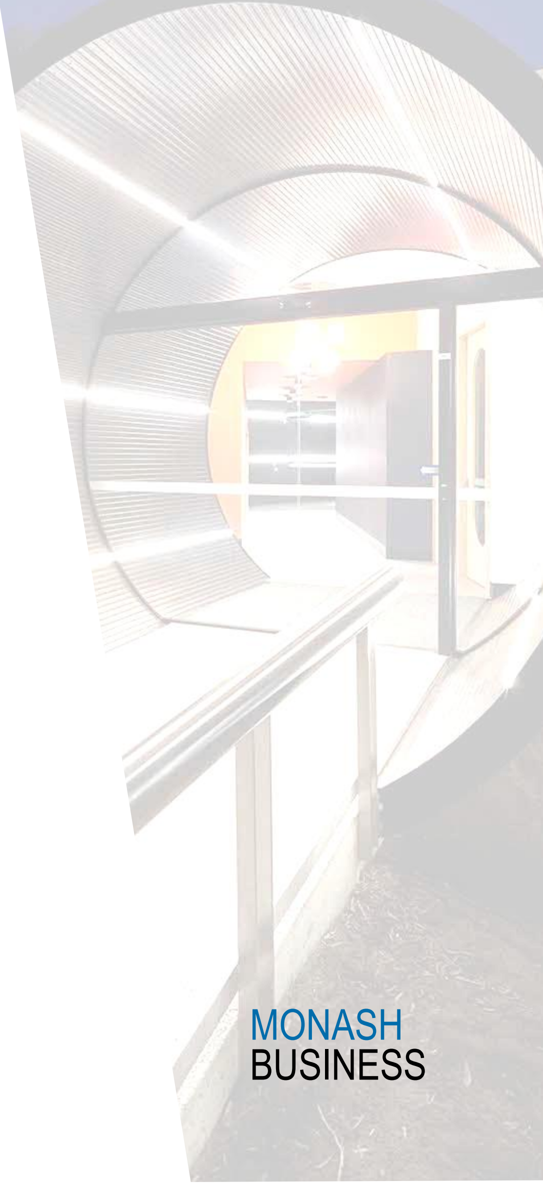
$\hat{a}$: ***estimated*** *constant*

$\hat{\beta}_i$: ***estimated*** *coefficient of* x_i

x_i: *independent variable or explanatory variable*

Note:

- Only 1 independent variable.
- i refers to No. of observations.
- There is an error term in the pop. Model, but not in the estimated model.
- Population model is a theoretical model.
- Estimated model is estimated by using sample.
- Notice that the hat indicates the estimated variable.

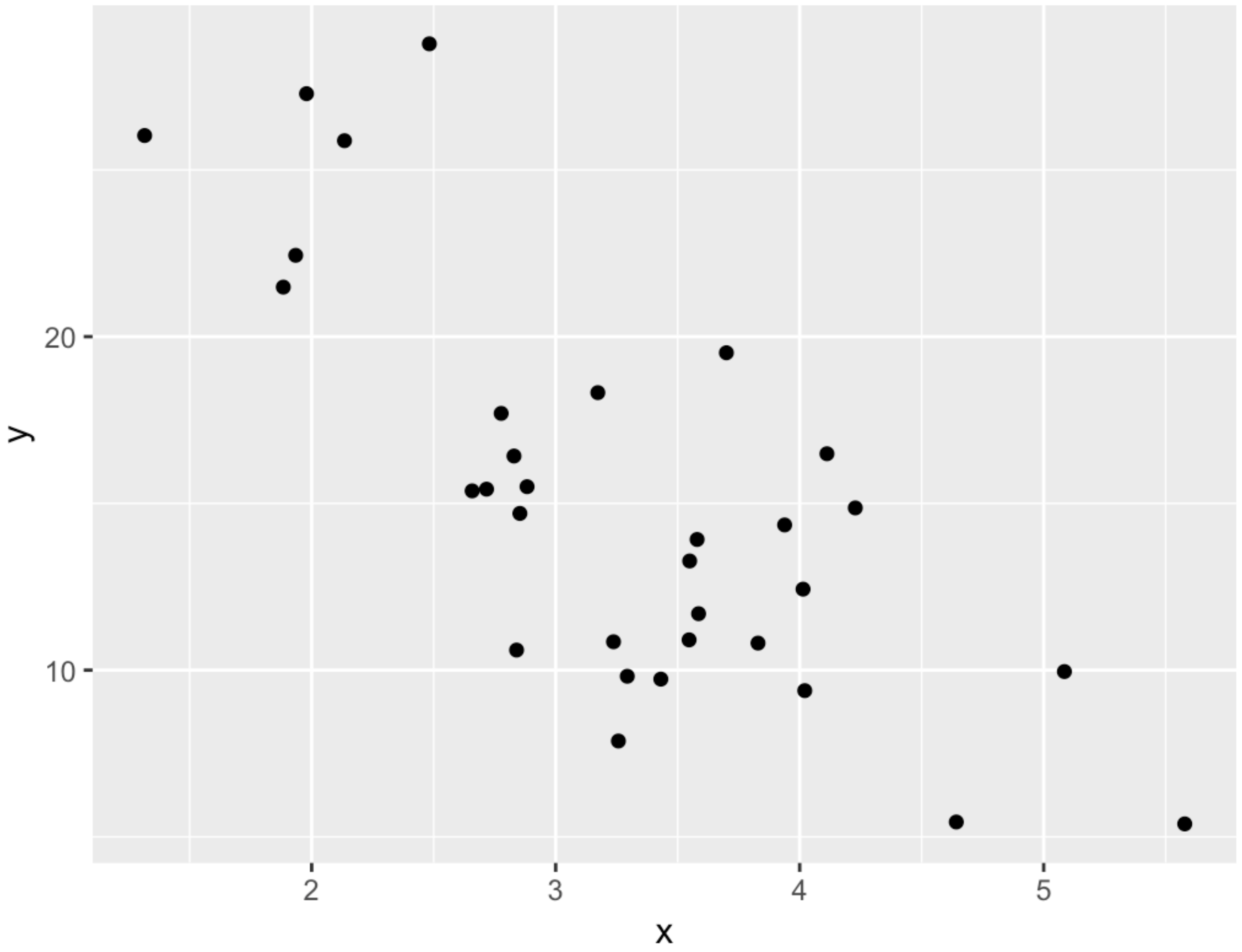


MONASH University MALAYSIA

MONASH BUSINESS

Recap: Simple linear regression

Visualization of simple linear regression



MONASH
BUSINESS

Recap: Multiple linear regression

Basic format

- Population model:

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_{ki} + u_i, i = 1,2, \ldots, n$$

Where:

$$y: dependent\ varaible\ or\ explained\ variable$$
$$a: constant$$
$$\beta_{ki}: coefficient\ of\ x_i$$
$$x_{ki}: independent\ variable\ or\ explanatory\ variable$$
$$u_i: error\ term$$

- Estimated model:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_{ki}, i = 1,2, \ldots, n$$

Where:

$$\hat{y}: \textbf{estimated}\ dependent\ varaible\ or\ explained\ variable$$
$$\hat{a}: \textbf{estimated}\ constant$$
$$\hat{\beta}_{ki}: \textbf{estimated}\ coefficient\ of\ x_i$$
$$x_{ki}: independent\ variable\ or\ explanatory\ variable$$

Note:

- Multiple independent variables.
- k refers to No. of coefficients and variables.

MONASH University MALAYSIA

MONASH BUSINESS

Recap: Interpretation

❑ Simple linear Regression

Example 1

$$\widehat{Salary} = 2382.32 + 28.92WAM_i, i = 1,2, \dots, n$$

$$R^2 = 0.108, n = 1{,}000$$

Example 2

$$\widehat{WAM} = 30.82 + 0.047min\_video_i, i = 1,2, \dots, n$$

$$R^2 = 0.287, n = 384$$

Note:

Salary: Graduates' salary in RM
WAM: Final WAM upon graduation

WAM: WAM of the semester
Min_video: Minutes to watch lecture video

❑ Interpretation

Example 1

- The model predicts that increase in 1 point of WAM would increase salary, by RM 28.92, on average.

Example 2

- The model predicts that students watching 1 extra minute of lecture video would increase WAM by 0.047, on average.

MONASH University MALAYSIA

MONASH BUSINESS

Recap: Interpretation

MONASH University MALAYSIA

- Multiple linear Regression

$$\widehat{WAM} = 42.74 + 0.057min\_video_i - 1.37hrs\_insta_i - 0.163classsize_i,$$
$$i = 1,2,\ldots,n$$
$$R^2 = 0.375, n = 384$$

- Interpretation

<u>Min_video</u>

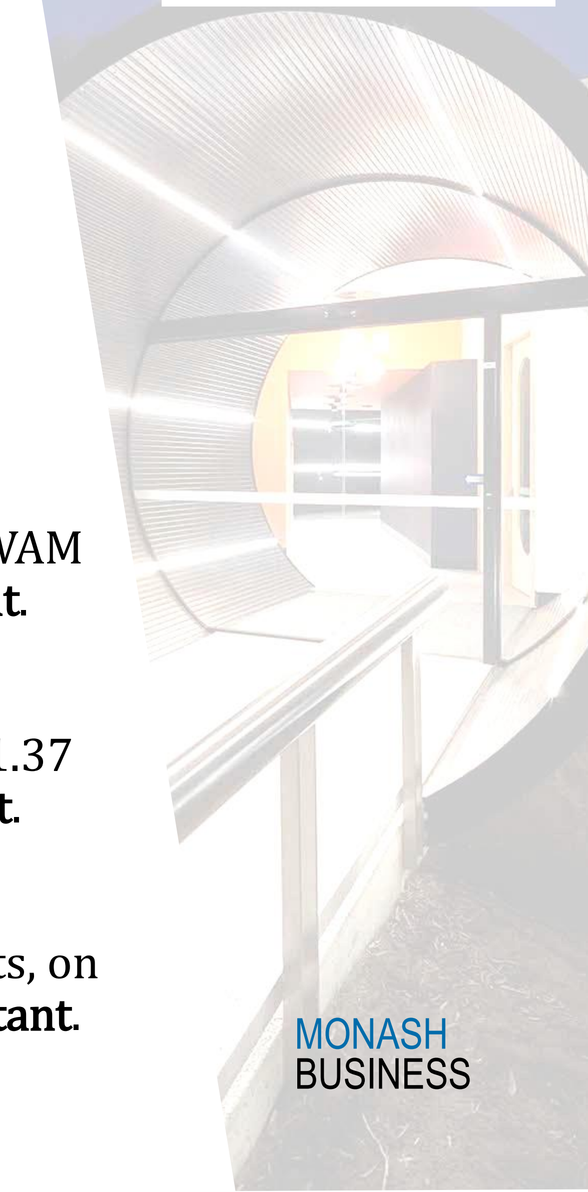
- The model predicts that students watching extra minute of lecture video would increase WAM by 0.057 points, on average, while **holding** hours of using Instagram and class size **constant**.

<u>hrs_insta</u>

- The model predicts that students using extra hour of Instagram would decrease WAM by 1.37 points, on average, while **holding** minutes of watching lecture video and class size **constant**.

<u>class_size</u>

- The model predicts that having extra student in class would decrease WAM by 0.163 points, on average, while **holding** minutes of watching lecture video and hours using Instagram **constant**.



MONASH BUSINESS

Recap: Inference

- Simple linear Regression

$$\widehat{Salary} = 2382.32 + \underset{(15.42)}{28.92}WAM_i, i = 1,2, \dots, n$$
$$R^2 = 0.108, n = 1{,}000$$

Note:

Salary: Graduates' salary in RM
WAM: Final WAM upon graduation

- **Inference**

Step 1: Set hypothesis

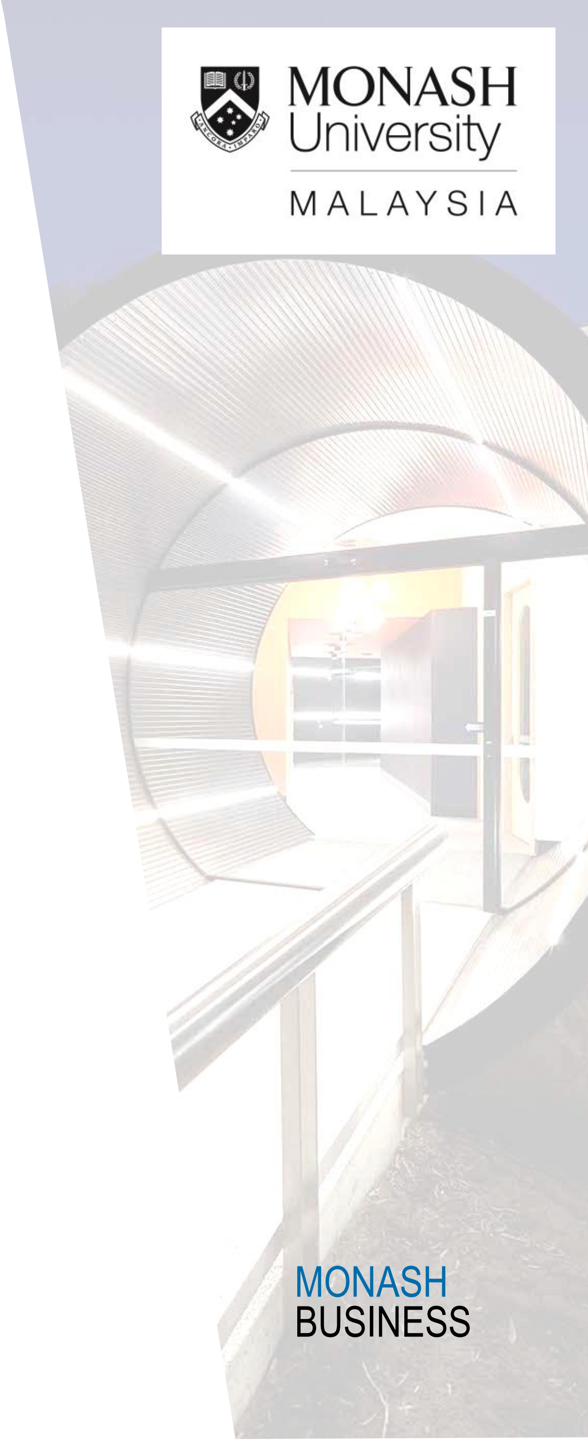
- $H_0: \beta_{WAM} = 0$
- $H_1: \beta_{WAM} \neq 0$

What we want to find out:

Is WAM upon graduation important to determine the salary?

Step 2: Level of significance

- 5% Level of Significance, $\alpha = 0.05$



Recap: Inference

MONASH University MALAYSIA

- Simple linear Regression

$$\widehat{Salary} = \underset{(15.42)}{2382.32 + 28.92WAM_i}, i = 1,2,\dots,n$$

$$R^2 = 0.108, n = 1{,}000$$

Note:

Salary: Graduates' salary in RM
WAM: Final WAM upon graduation

- **Inference**

Obtained from H_0

Step 3: t-stat approach

$$t_{stat}: \frac{\hat{\beta}_{WAM} - \beta_{WAM}}{se(\hat{\beta}_{WAM})}$$

$$= t_{stat}: \frac{28.92 - 0}{15.42}$$

$$\approx 1.875$$

See what happens if lower.tail=TRUE

Obtain t-crit value

Using $qt(\alpha/2, df, lower.tail = FALSE)$
$= 1.962341$

Step 4: Conclusion

- Since the t-stat < t-crit, we **DO NOT** reject the null hypothesis.
- There is an **insufficient** evidence to conclude that WAM has a significant effect on graduate's salary.

MONASH BUSINESS

Recap: Inference

- Simple linear Regression

$$\widehat{Salary} = \underset{(15.42)}{2382.32 + 28.92} WAM_i, i = 1,2,\dots,n$$
$$R^2 = 0.108, n = 1,000$$

Note:

Salary: Graduates' salary in RM
WAM: Final WAM upon graduation

- **Inference**

Step 1: Set hypothesis

- $H_0: \beta_{WAM} = 0$
- $H_1: \beta_{WAM} > 0$

What if we want to find whether WAM has a **positive** impact on salary?

Step 2: decision rule

- 5% Level of Significance, $\alpha = 0.05$

Step 3
The t-stat value is still the same
Using $qt(\alpha, df, lower.tail = FALSE)$
$= 1.646382$

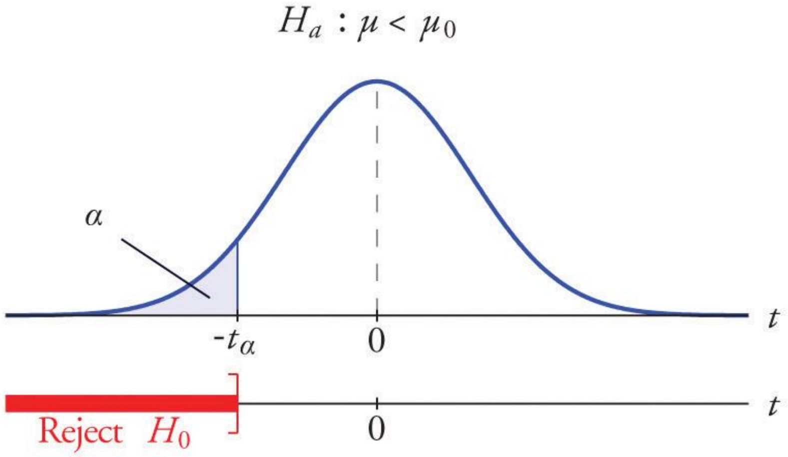
Step 4: Conclusion

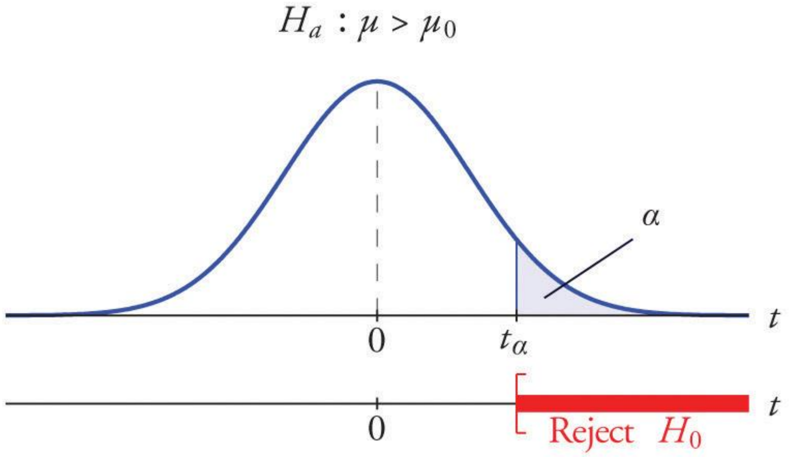
- Since the t-stat < t-crit, we reject the null hypothesis.
- There is a sufficient evidence to conclude that WAM has a positive effect on graduate's salary.

MONASH BUSINESS

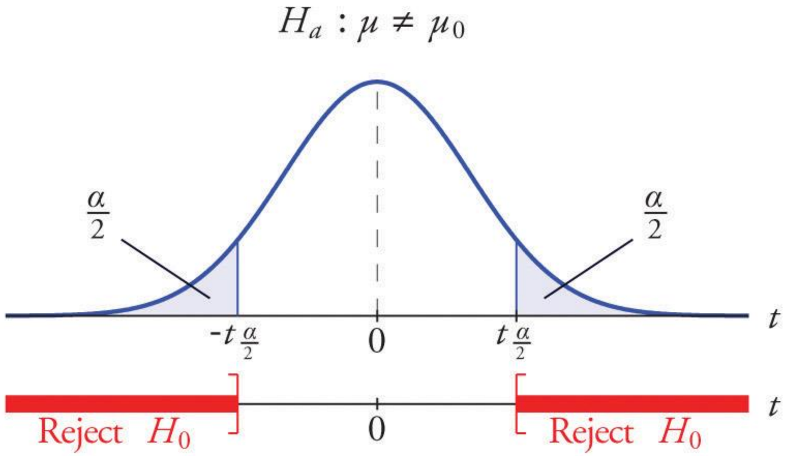
Recap: Inference

One-sided test





Two-sided test



Recap: Model evaluation

- R^2: % variation of dependent variable explained by % variation of independent variables.
- It is called R-squared or Coefficient of Determination.
- $0 < R^2 < 1$
- Adding an extra independent variable only increases R^2
- If the added independent variable is statistically significant (an important variable), then the R^2 would increase much.
- $R^2 = 1 - \frac{SSR}{SST} \; or \frac{SSE}{SST}$
- $SSR = Sum\ of\ Squared\ Residual, SST = Sum\ of\ squared\ total$
- $SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$
- $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$

Recap: Model evaluation

- Since the R^2 value increases as we include more variables, it appears that adding more variables seems good.

- Explanatory power vs parsimony

- We need to consider important **explanatory** variables (x variables) to explain the variation of **explained** variable (y variable).

- Selection of variables: “What do you intend to find out?”
    - Using common sense (when you have obvious variables)
    - Educational guess (what you ‘think’ it is important’)
    - Theory driven ideas (explore important variables)

- Between several models having **the same** dependent variable but with different combination of independent variables, we use $Adjusted\ R^2$ to compare.

- $Adjusted\ R^2 = 1 - (1 - R^2)\ * \left(\frac{n-1}{n-k-1}\right)$

- Rather than using R^2 value directly, we compare using $adjusted\ R^2$

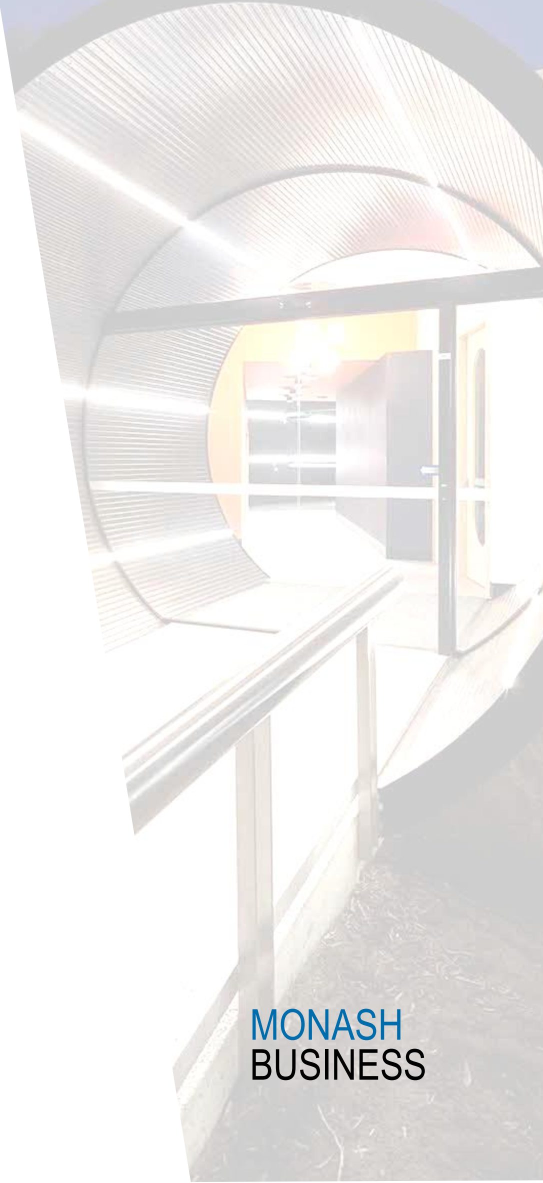
MONASH BUSINESS

Regression with Constant

MONASH University MALAYSIA

- What happens if we regress dependent variable with constant?
- $y_i = \alpha + u_i, i = 1,2, \dots , n$
- The dataframe looks like:
- $y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_i \end{bmatrix} \; and \; a = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \dots \\ 1 \end{bmatrix},$
- INSERT regress y on constant lm(y~1), see what happens
- It returns a mean of the dependent variable!
- Think back how we usually interpret the coefficients – ‘average’ effect of x on y.
- Regression is a **linear** function that minimizes ‘residual’ between y and $\hat{y}$.

Linear in parameters



MONASH BUSINESS

Comparison between two groups

- We can assign numerical value to the categorical variables.

- It is useful when the observations are categorized into distinctive groups:
  - Gender: Male or Female
  - Race: Malay, Chinese, or Indian
  - Quarters: Q1, Q2, Q3 or Q4
  - Industry sectors: Primary, Secondary or Tertiary

- If you have conducted any survey asking demographic characteristics, those data are most likely used to separate observations by groups.

- Finding differences between groups impose interesting stories:
  - Average sleeping hour between male and female.
  - University enrolment percentage between races.
  - Quarterly change in unemployment rate.
  - Average wage between workers in three sectors.

- A simple way to incorporate the difference between group is including binary (dummy) variable into the regression model.

Comparison between two groups

- For any category, you can create a variable and assign value of 1.
- For instance:

Gender

Male

=1 if the obsn is male, =0 otherwise

Female

=1 if the obsn is female, =0 otherwise

- You only need to create 1, since one gender variable can categorize all observations.
- If the variable includes 3 categories, then you need to create 2 dummy variables.

Comparison between two groups

- Hence, you need to omit at least one variable as a base dummy variable

- The estimated coefficients of dummy variables will be compared with base condition.

- For example:
  - Dependent variable = Income in RM
  - Independent variables = Primary, Secondary and Tertiary sector

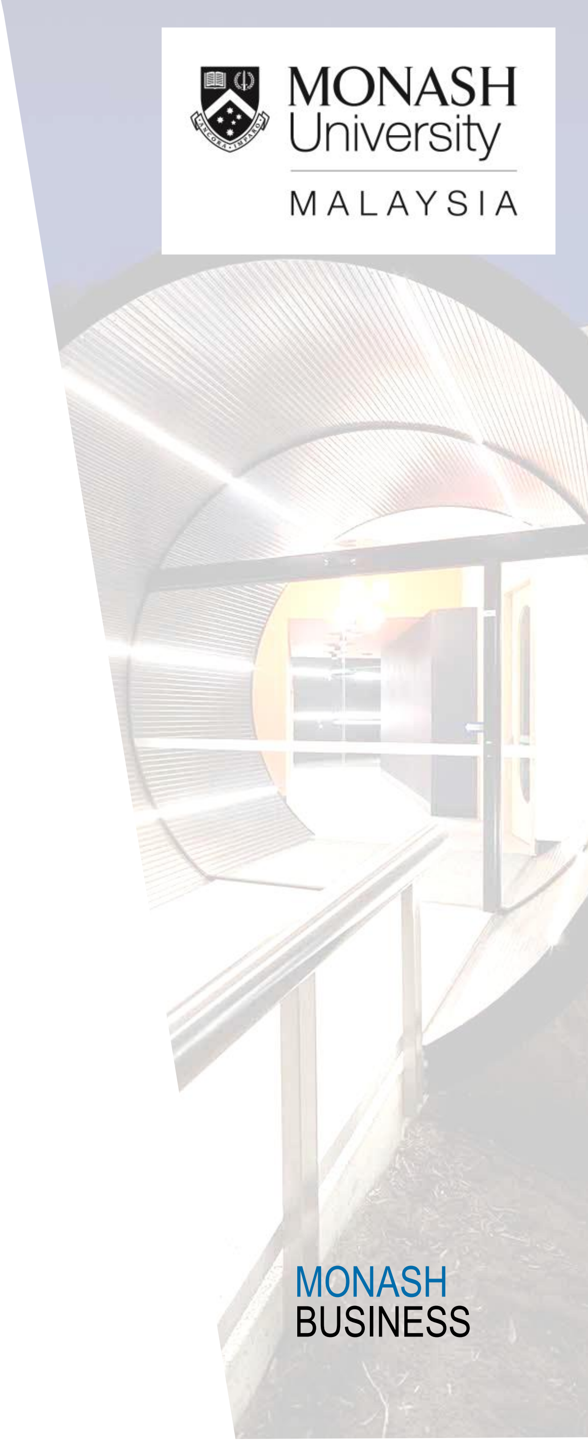
$$\widehat{income} = 4246.37 + 385.21Secondary_i + 769.38Tertiary_i$$

  - Average income for primary sector is RM 4246.37.
  - Average income for secondary sector earns RM 385.21 more than **primary sector**.
  - Average income for tertiary sector earns RM 769.38 more than **primary sector.**

- Note that the comparison is made between estimated coeff and the base dummy.

$$\widehat{income} = 4631.58 - 385.21Primary_i + 384.17Tertiary_i$$

  - Two changes: Base dummy changed, and the estimated coefficients changed.



Comparison between two groups

MONASH University MALAYSIA

- It is possible to incorporate multiple categorical variables.

- Previous example only includes one category: economic sector.

- You can also take gender into account across the sector:
  - Dependent variable = Income in RM
  - Independent variables = Primary, Secondary, Tertiary sector, Male and Female

$$\widehat{income} = 3974.53 + 147.83Secondary_i + 353.48Tertiary_i + 328.42Male_i$$

  - Average income for female who works in primary sector is RM 3974.53.
  - Average income for female who works in secondary sector earns RM 147.83 more than the female who works in primary sector.
  - Average income for female who works in tertiary sector earns RM 353.48 more than the female who works in primary sector.
  - Across all economic sectors, male workers earn 328.42RM more than female workers on average.
  - Identification of the characteristics of the base dummy variable is important.

MONASH BUSINESS

Comparison between two groups

$$\widehat{income} = 3974.53 + 147.83Secondary_i + 353.48Tertiary_i + 328.42Male_i$$

- Here, the base dummy is a female who works in the primary sector.

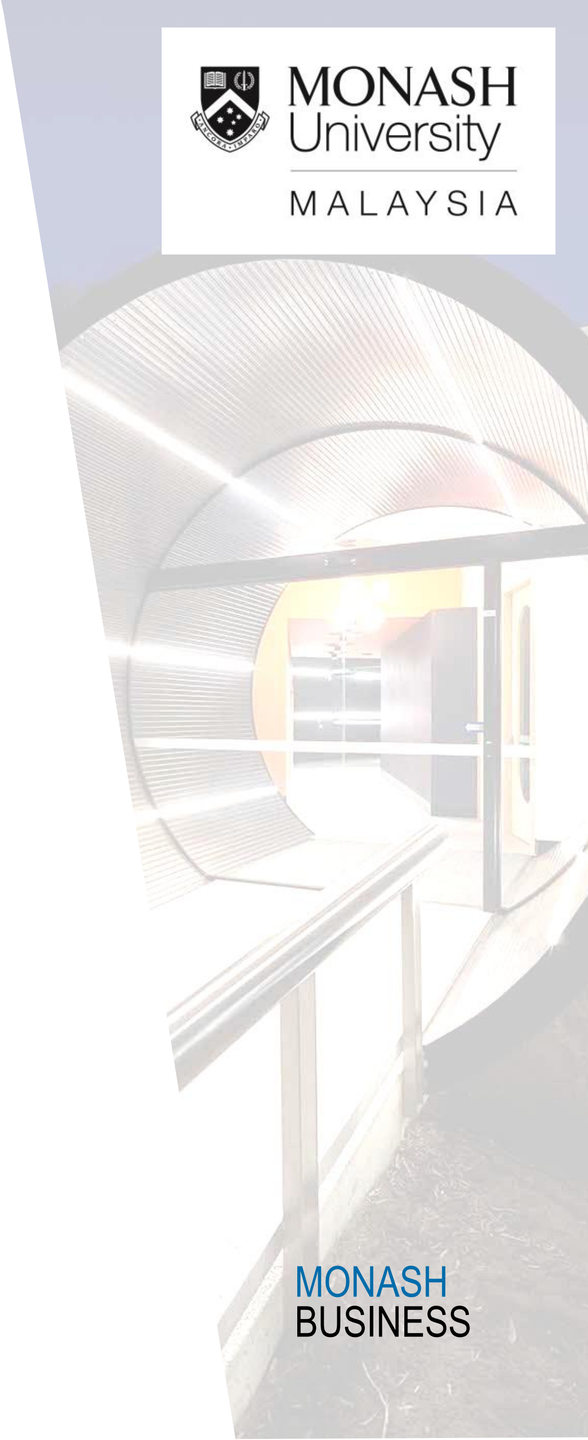
- Just like how we identified average income across economic sector, we can also identify average income across economic sector **between gender.**

- The estimated coefficients of 'secondary' and 'tertiary' indicate the difference in average income between female workers in 'primary' sector.

- To compute the average income for male, we can substitute 1 into respective variable.
- For example, average income for male who works in secondary sector;

$$\widehat{income} = 3974.53 + 147.83(1) + 353.48(0) + 328.42(1)$$

Average income across sector and gender		**Gender**	
		Male	**Female**
Economic Sector	**Primary**	4302.95	3974.53
	Secondary	4450.78	4122.36
	Tertiary	4656.43	4328.01



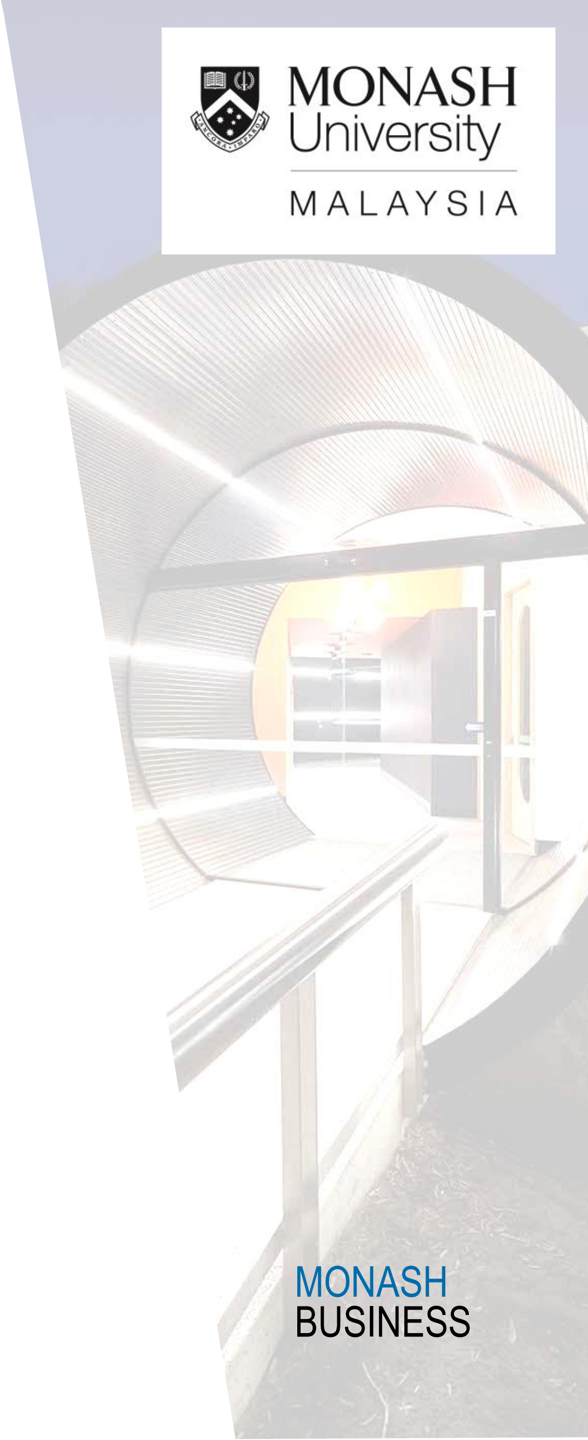
Linear Probability Model

- We can also use dummy variable for the dependent variable.
- In this case, the estimated coefficients refer to the probability of the dependent variable being =1.
- For example:

$$\widehat{EMP}_i = 0.072 + 0.0051WAM_i + 0.047Stat_i + 0.001Male_i$$

$$EMP_i: 1 = employed, 0 = otherwise$$
$$WAM_i: Weighted\ Average\ Mean$$
$$Stat_i: 1 = Stat\ Major, 0 = Other\ Majors$$
$$Male_i: 1 = Male, 0 = Otherwise$$

- The estimated intercept represents that there is an average of 7.2% chance of getting employed when the person is female, non-stat major, and WAM is zero.
- In this case, the model makes sense as WAM is less than 0.01.
- If the estimated coefficient for WAM >0.01, then the predicted value for the probability of getting employed might exceed 1 mathematically.



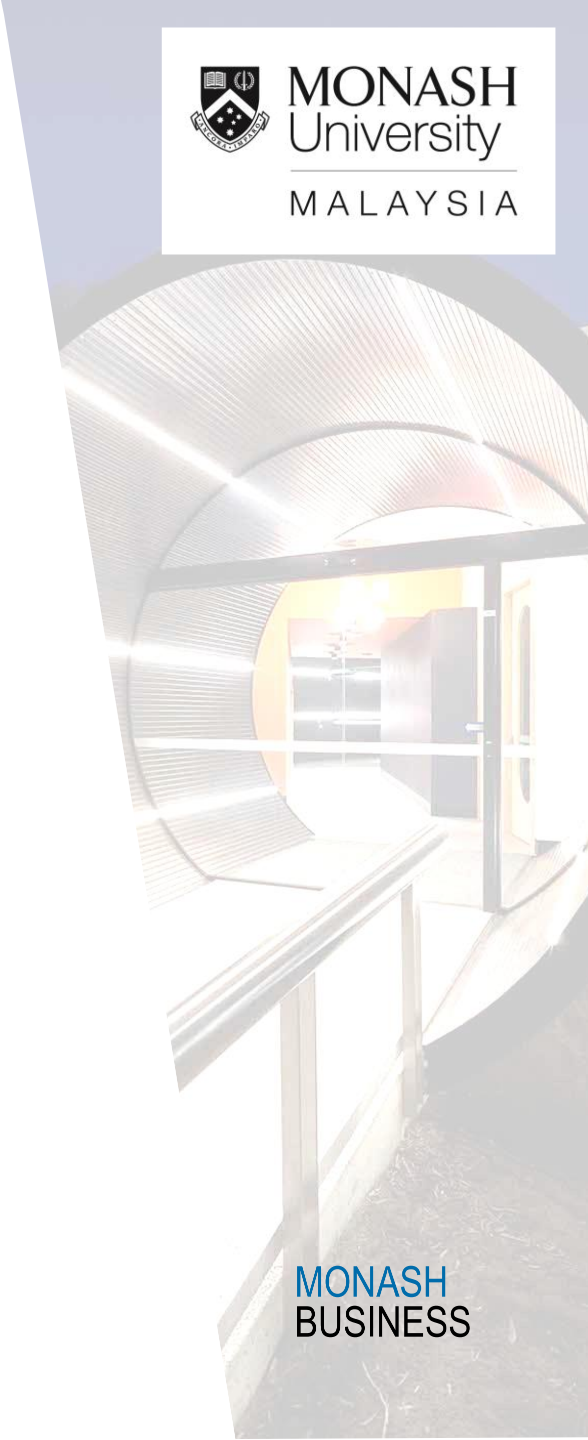
Linear Probability Model

$$\widehat{EMP}_i = 0.072 + 0.0051WAM_i + 0.047Stat_i + 0.001Male_i$$

- Interpretation
  - $\hat{\beta}_{WAM} = 0.0051$: The model estimates that a person whose WAM is 1 point higher have 0.0051 chance more to be employed on average, while holding the major and gender constant
  - Alternatively, you can also interpret as: between two people whose major and gender are the same, a person with 1 point higher WAM has 0.0051 higher chance to be employed.
  - $\hat{\beta}_{Stat} = 0.047$: The model estimates that the proportion of employed people who are stat major is 0.047 higher than non-stat major while their WAM and gender are the same.
- Note that the interpretation slightly changed from probability to the proportion as we are interpreting dummy variable.
- You will practice hypothesis testing during tutorial.

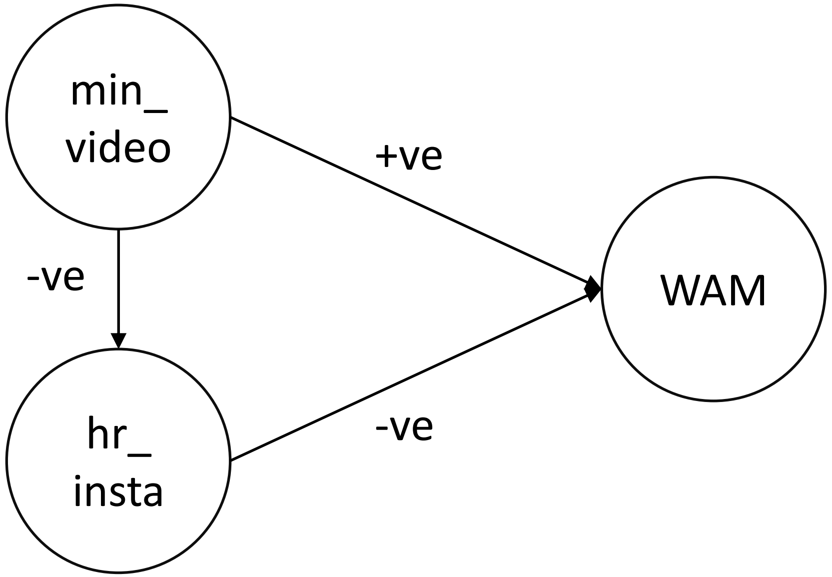
Multicollinearity

- Estimated coefficients are calculated based on how independent and dependent variables are correlated.
- We ‘hold’ other variables constant when we interpret: isolating the effect of x on y.
- The above statement assumes that other independent variables do not interfere.
- Try to look at the example below:
- $\widehat{WAM} = 42.74 + 0.057min\_video_i - 1.37hrs\_insta_i - 0.163classsize_i$
- By now, you would know how to interpret each of estimated coefficient.
- We assume that the minutes spend on lecture video positively affect on WAM while time (hours) spend on Instagram and class-size are held constant.
- Are time spent on both lecture and Instagram totally irrelevant?
- If they are relevant, the estimated coefficient are affected hence its significance.



Multicollinearity

- Let's compare two models below:
- $\widehat{WAM} = 30.82 + 0.047min\_video_i, i = 1,2, \ldots, n$
- $\widehat{WAM} = 42.74 + 0.057min\_video_i - 1.37hrs\_insta_i - 0.163classsize_i$
- If the true effect of time spend on watching lecture video is 0.047, inclusion of other variables would not affect the coefficient.



- Both variables are statistically significant on WAM.
- 0.047 is underestimated as the effect of time spent on Instagram on WAM is partially included in min_video in simple linear regression.
- The degree of interference is influenced by the degree of correlation between x1 and x2.
- If x1 and x2 are perfectly correlated, coefficients cannot be calculated mathematically.

MONASH
BUSINESS

Multicollinearity

- ❑ There is a simple way to check whether model includes severe multicollinearity
- ❑ Variance Inflation Factor (VIF) detects severity of multicollinearity which estimates how much the variance of the regression coefficient is inflated due to the multicollinearity.
- ❑ $VIF = \frac{1}{1-R^2}$
- ❑ We all know that R^2 represents the coefficient of determination
  - % of variation of dependent variable is explained by the variation of independent variables.
- ❑ This is why high R^2 value is not necessarily good.
- ❑ Again, the context of the model is important.
- ❑ Conservative views <3, lower the better.
- ❑ Time-series data vs cross-sectional data

R^2 and VIF

R^2	VIF
0.5	0.5
0.67	0.67
0.86	0.86
0.9	10

❑ Thompson, C. G., Kim, R. S., Aloe, A. M., & Becker, B. J. (2017). Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. *Basic and Applied Social Psychology*, *39*(2), 81-90. (read pg. 81 – 84 for your own reference)

Omitted variable bias

MONASH University MALAYSIA

- Then, should we include more variables? Or the least variables?
- This is where 'modelling' technique comes.
- The ideal regression model would be 'Parsimonious' and 'Robust'.
- Inclusion of too many variables may cause problem of multicollinearity
- Inclusion of too little variables may cause omitted variable bias.
- Omitted variable bias: Biasedness caused from an exclusion of important variables that is correlated with independent variables.
- What can we do?

    1. Think about the nature of dependent variables.
    2. What would be the important variables to explain / cause the dependent variable? (see slide 13)
    3. Check whether similar variables are included.
    4. Check the significance of coefficients to keep / withdraw – review residuals of the model.
    5. See whether the model makes sense in general. Can you make a meaningful interpretation and implication out of it?