DEPARTMENT OF
ECONOMETRICS
AND BUSINESS
STATISTICS

CHUN MINSOO

MONASH University

MONASH BUSINESS SCHOOL

Multiple Linear Regression analysis

Part 2

ETW2001 Foundations of Data Analysis and Modeling

Outline

- ❑ **Regression – Part 2**
  - ❑ Residual analysis
  - ❑ Functional forms
  - ❑ Hypothesis testing
  - ❑ Prediction

MONASH University MALAYSIA

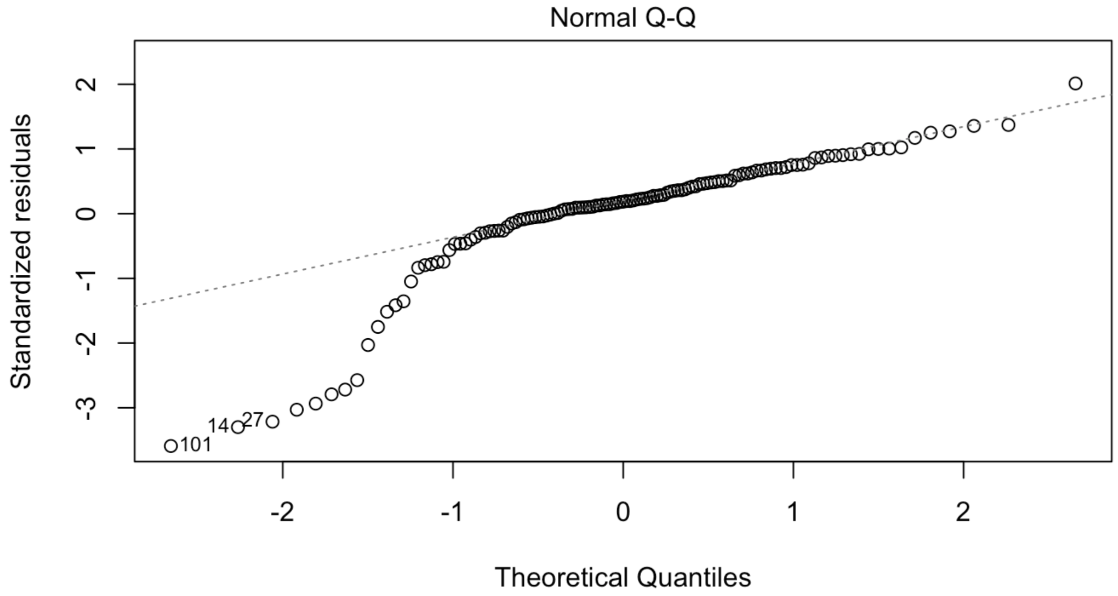
Residual diagnostics

- So far, we focused on interpretation and hypothesis testing of estimated coefficient.
- For rigorous modelling, you need to ensure the ‘robustness’ of the model.
- One way to improve the robustness is to check multicollinearity. (from previous week)
- Another diagnostic check is to examine patterns of residuals.
- Residual, $\hat{e}\ or\ \hat{u} =$ y (actual observation) - $\hat{y}$ (predicted value)
- There is a difference between error term and residual term.
    - Error term: all the excluded variables from your population model.
    - Residual term: discrepancy between actual and predicted value.
- You can visualize residuals to detect how it behaves
    1. Q-Q plot
    2. Residual vs fitted values plot
    3. Residual histogram

Residual diagnostics

Visualization of residuals

1. Q-Q plot

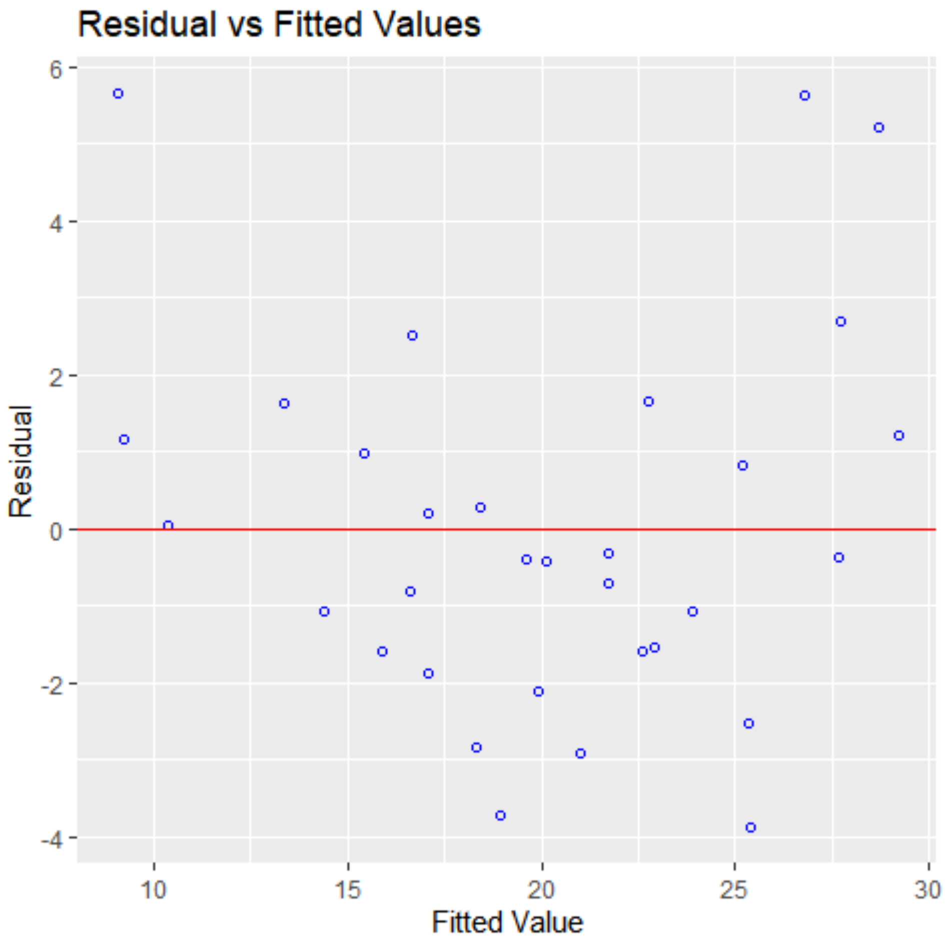


- Q-Q plot visualizes the normality of the variable.
- In this case, it examines the normality of the residuals.
- The residual follows a normality along the diagonal line.

MONASH BUSINESS

Residual diagnostics

Visualization of residuals
2. Residual vs fitted values plot



- x-axis: fitted value, y-axis: residuals.
- The residual 'behaves' well if the residuals are randomly scattered around the horizontal line, at 0.
- Any observed pattern (linear or non-linear) might indicate significant variable is omitted.
- The observed pattern in the residual plot influences the standard error of the estimated coefficients. Hence, it will affect the result of hypothesis testing.

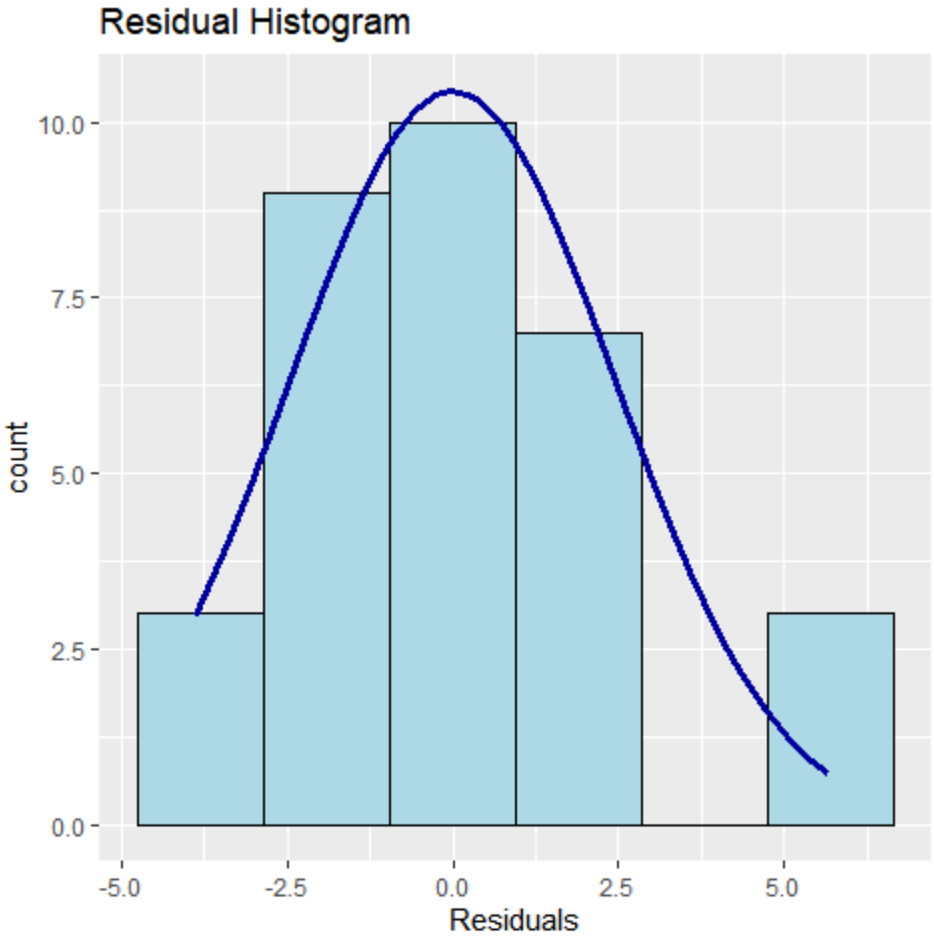
MONASH University MALAYSIA

MONASH BUSINESS

Residual diagnostics

Visualization of residuals

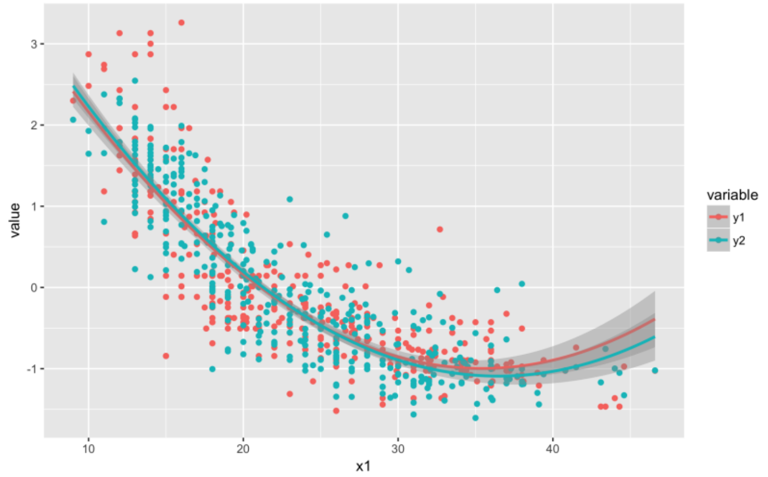
3. Residual histogram



- This is another visualization to observe normality of residuals.
- If the residuals are not distributed around 0, it also indicates omitted variable bias.
- We will not cover formal statistical test for the diagnostic analysis in this unit. ETW2510 will teach you further techniques 😀 (it is called heteroskedasticity).

Functional forms – quadratic term

- So far, we include variables as it is from the dataset.
- You can modify variables mathematically to improve the predictability.
- For instance:



- The scatterplot indicates quadratic relationship between independent and dependent variables.
- A typical regression model may not provide accurate prediction.

- It affects both estimation and interpretation.
- Note that our usual interpretation takes increases in 1 unit of x, y is estimated to change by its corresponding $\hat{\beta}$.
- However, the interpretation is no longer the same as the variable x does not take a linear form.

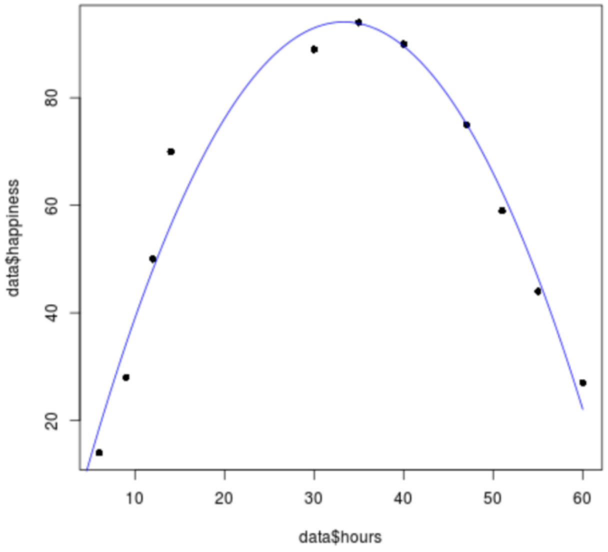
MONASH University MALAYSIA

MONASH BUSINESS

Functional forms – quadratic term

- Quadratic term: applying x^2 as independent variable.
- Here, the level of satisfaction increases until around 30 hours, reaching the maximum.
- If we use No. of hours to predict the happiness using linear function, the effect of time on happiness may not be significant – the best fit would be horizontally constant.
- Hence, the appropriate regression model would be:

$$happiness_{\mathrm{i}} = \beta_0 + \beta_1 hours_i + \beta_2 hours_i^2 + u_i$$



- Here, we can do a bit of maths to compute the turning point:

$$\frac{\delta happiness}{\delta hours} = \hat{\beta}_1 + 2 \cdot \hat{\beta}_2 \cdot hours_i = 0$$

- The pattern increases, then decreases. Here, we can estimate the direction of the estimated coefficient.

$$\hat{\beta}_1 = +ve, and\ \hat{\beta}_2 = -ve$$

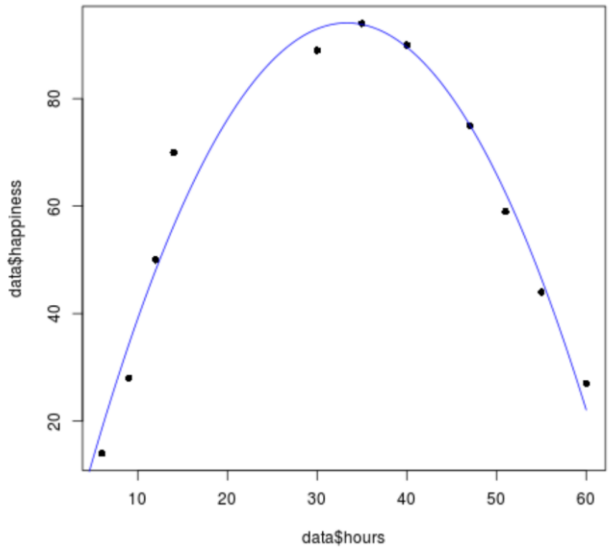
MONASH University MALAYSIA

MONASH BUSINESS

Functional forms – quadratic term

- Continuing from the previous example, the usual interpretation of constant increase does not work.

$$happiness_{\mathrm{i}} = \beta_0 + \beta_1 hours_i + \beta_2 hours_i^2 + u_i$$



- The estimated coefficient $\hat{\beta}_1$ would be relatively larger than the estimated coefficient $\hat{\beta}_2$.
- As the value of x (hours) increases in the beginning, the effect of estimated coefficient $\hat{\beta}_1$ dominates as the effect of $\hat{\beta}_2$ is relatively small ($\hat{\beta}_1 hours_i > \hat{\beta}_2 hours_i^2$).
- However, as hours increases further, the effect of $\hat{\beta}_2$ increases at a faster rate due to the quadratic term, diminishing the effect of hours on happiness. ($\hat{\beta}_1 hours_i < \hat{\beta}_2 hours_i^2$)
- The incremental effect of each hour on happiness is not constant, thus, the usual interpretation does not apply.
- Usually, we focus on the pattern, and when the effect reach out either maximum or minimum.

Functional forms – quadratic term

- Let's look at one example,

$$\underset{}{\widehat{wage}_{\mathrm{i}}} = \underset{(0.19)}{3.73} + \underset{(0.066)}{0.298} exper_i - \underset{(0.019)}{0.0061} exper_i^2$$

- The aggregate effect of experience cannot be interpreted directly.

The turning point:

$$0.298 - 2 * 0.0061 * exper = 0$$

$$exper = \frac{-0.298}{2 * -0.0061}$$

$$= 24.426$$

This indicates that the effect of experience decreases average income after 24 years of experience.

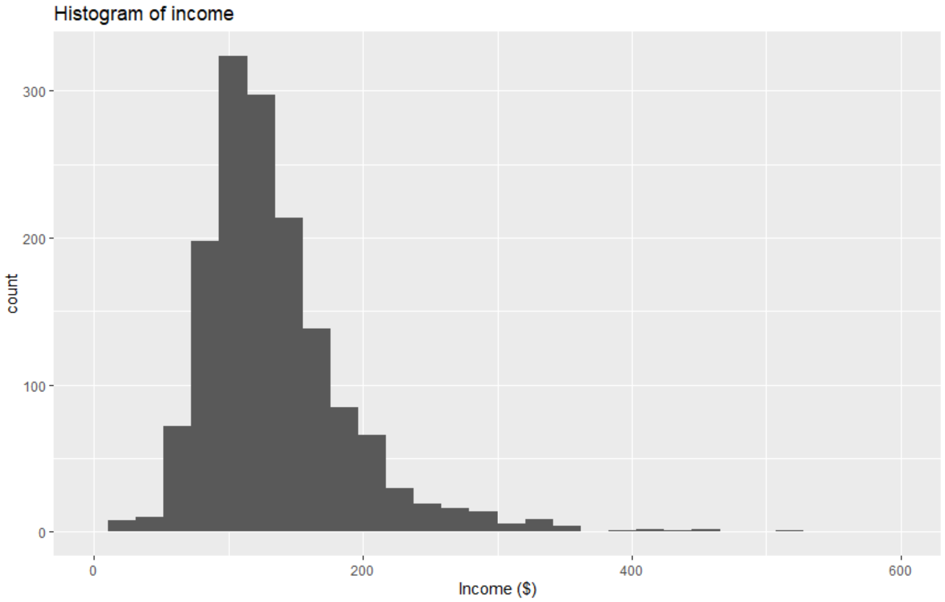
MONASH BUSINESS

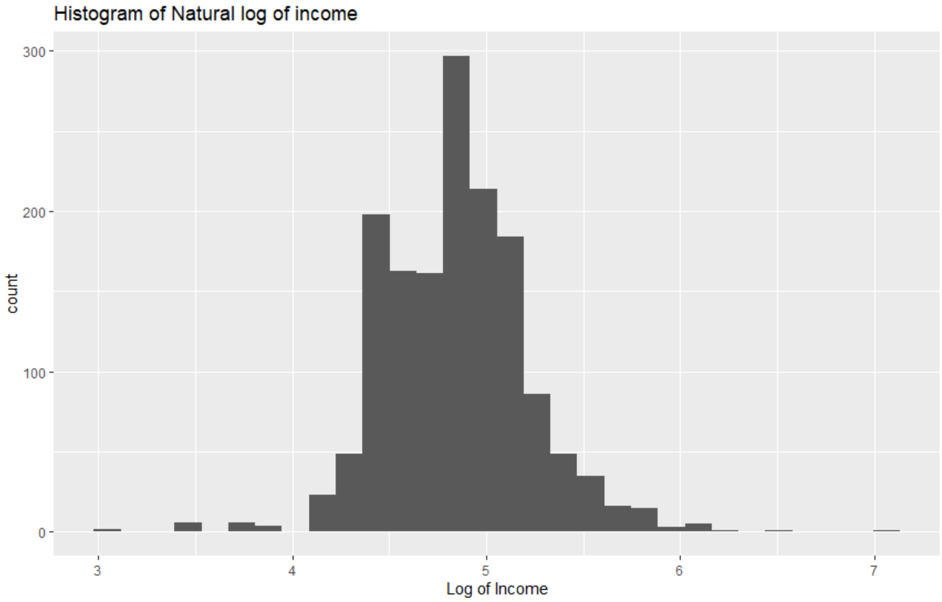
Functional forms – natural log

- Natural logarithm: applying natural $\log_e x$, $(\ln x)$.

- In statistics/econometrics, log(x) usually refers to natural logarithm.

- It is quite common to apply log function to the variables for few reasons:
    1. When the scale of variable is too large.
    2. When the distribution of the variable is not normal.
    3. When the relationship appears to be non-linear.
    4. If you wish to interpret the variables in terms of %.

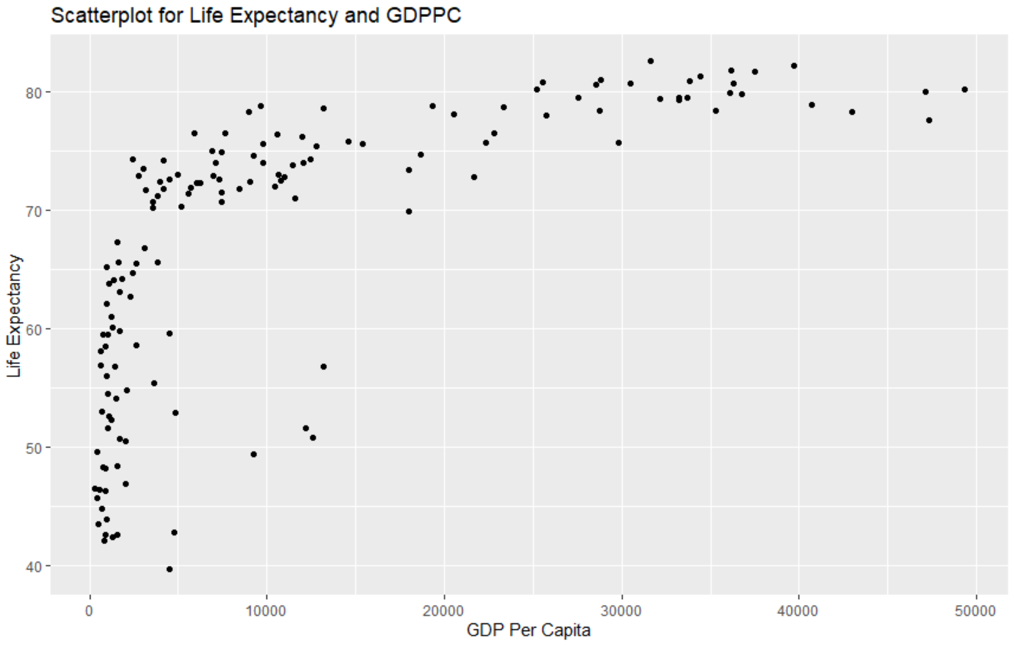


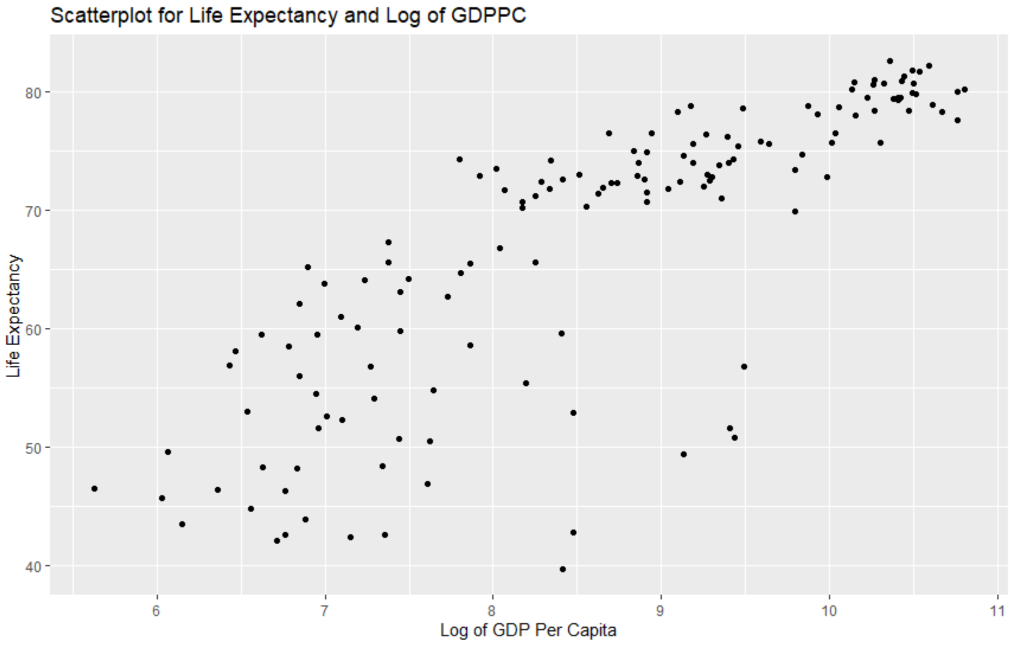


MONASH BUSINESS

Functional forms – natural log

- Diagrams below show the relationship between GDP per capita and life expectancy.
- One used the data as it is (left) and the other applied natural log on GDP per capita (right).
- For the goodness fit of the model, which one would be better?





MONASH BUSINESS

Functional forms – natural log

- Applying natural log alters how we interpret the estimated coefficient.
- The following table shows how we interpret.

Model	Estimated Equation	Interpretation
Level – Level	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$	One unit change in X leads to $\hat{\beta}$ unit change in Y
Log – Level	$Log(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 X$	One unit change in X leads to $100 * \hat{\beta}$ percent change in Y
Level – Log	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 Log(X)$	One percent change in X leads to $\hat{\beta}/100$ unit change in Y
Log – Log	$Log(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 Log(X)$	One percent change in X leads to $\hat{\beta}$ percent change in Y

Functional forms – natural log

- Let’s look at one example.

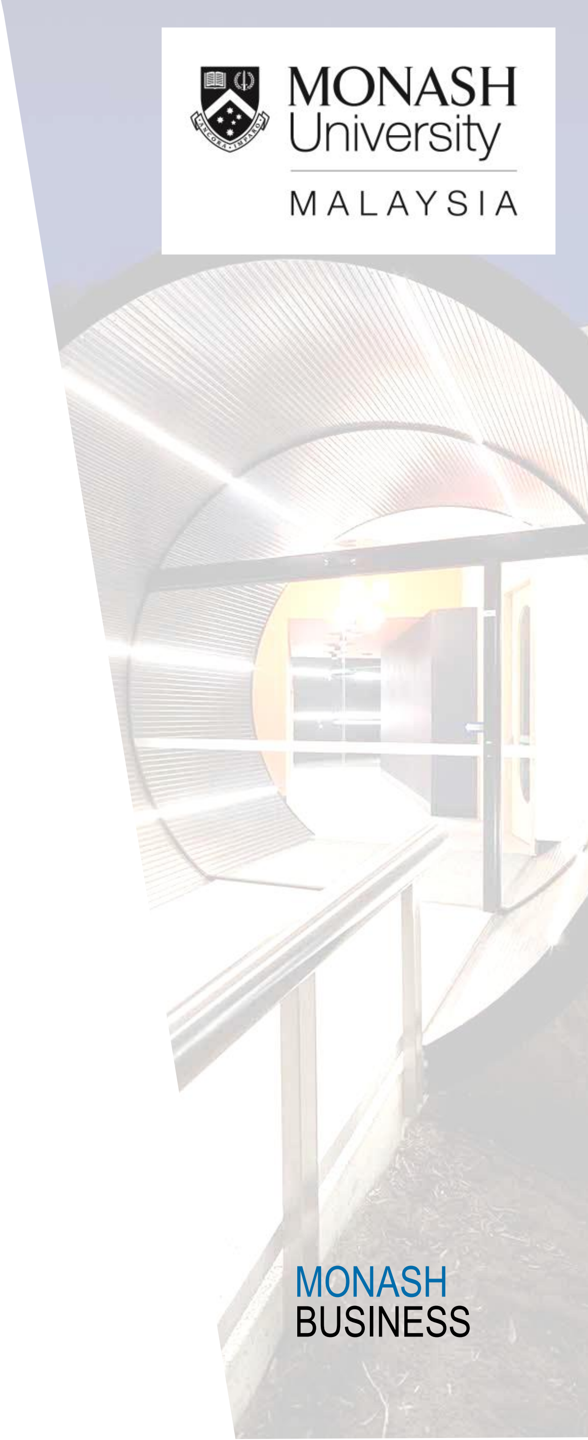
- $\widehat{Log(price)} = 9.23 - 0.718\log(nox) + 0.306rooms$

  $\qquad\qquad(0.19)\quad(0.066)\qquad\qquad(0.019)$

  - price = housing price
  - nox = amount of nitrogen oxide (pollution)
  - Rooms: number of rooms

- When the nitrogen oxide level increases by 1%, the housing price is expected to decrease by 0.72%, on average while holding the number of rooms constant.

- When the number of rooms increases by 1 unit, the housing price is expected to increase by 30.6% (100*0.306), on average while holding the nitrogen oxide level constant.

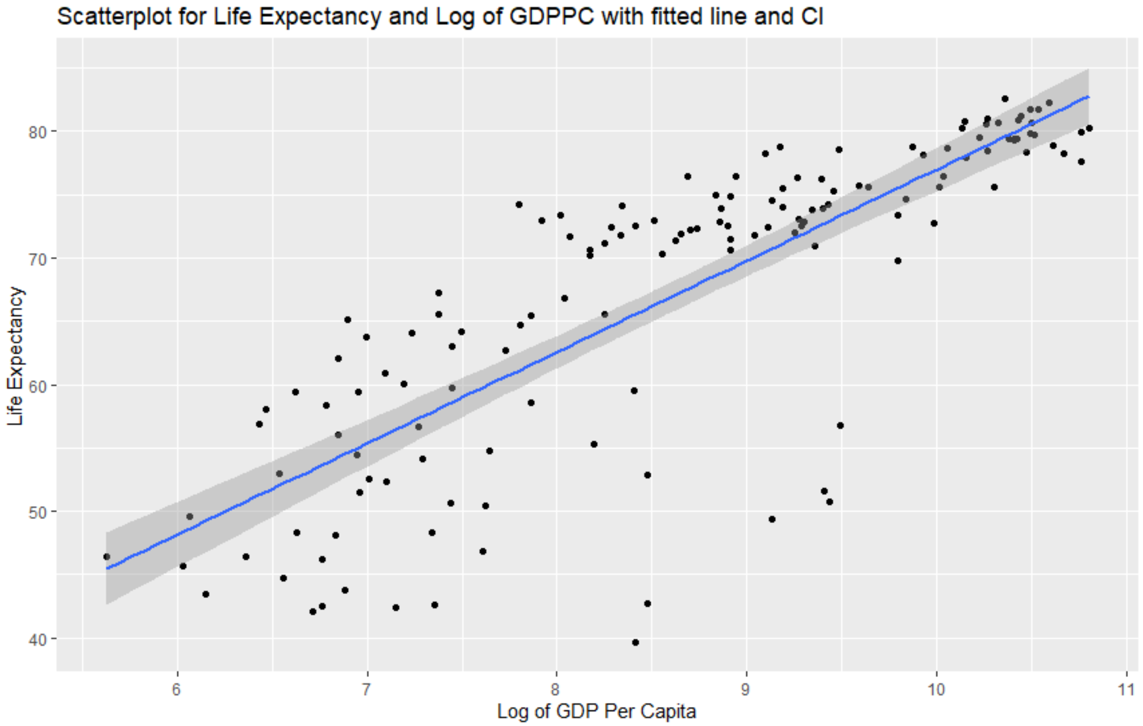


Prediction

- Once your model seems robust, making predictions is fairly simple.

- Using your estimated model, just substitute the desired values into independent variables, and the $\hat{y}$ is your predicted value for dependent variable.

- In a simple linear regression, you can visualize the prediction line with its confidence interval.



- From the scatterplot in slide 12, the fitted line is added showing the changes in life expectancy for changes in log of GDP per capita.

- The grey area shows the 95% confidence interval of the prediction.

MONASH BUSINESS

Prediction

- It would be difficult to visualize predicted value for multiple linear regression.

- There are two ways to predict for the estimated equation below:

$$\widehat{LifeExp}_i = 4.589 + 7.217\log(gdpPercap_i) + 0.000000005449pop_i$$

- Let's predict for a country that GDP per capita is \$21,583 and population is 13 mil.

1. You can substitute the values into the estimated model manually.

$$4.589 + 7.217\log(21583) + 0.000000005449 * 13000000$$
$$= 76.68$$

2. Alternatively, you can create a data frame of the values using tibble()

```
predict_data <- tibble(gdpPercap=21583, pop=13000000)
predict(lm2, predict_data)
```

Using predict(regression model, values for indep. Variable) to automatically substitute the values for you.

MONASH BUSINESS