Data Analytics and Mathematical Statistics
Day 4

Source: Towards Data Science, Spring Board, The Software Report

Agenda

- Investigating trends and relationships
  - Using models
  - Checking model assumptions
  - Hands-on model building

Investigating trends and relationships

Answering questions involving associations

Answering questions involving associations / trends / relationships

- E.g. does the size of my flat affect its return value?

Linear Regression

- Linear regression is a way for us to model the relationship between two variables of interest

- Say we are interested in the effect of studying time (# of waking hours) on marks obtained (% score)

- We can model this relationship using linear regression, which takes on the general form:

$$y = \beta_0 + \beta_1 x + \epsilon$$

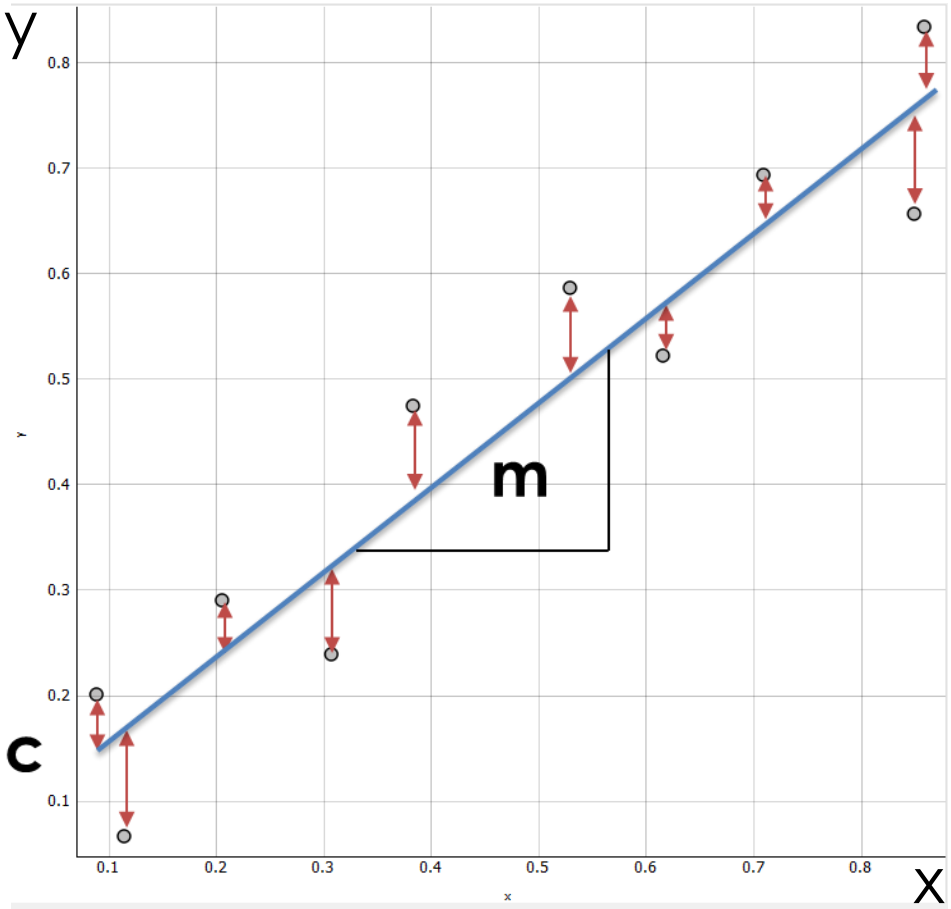
LR = Machine learning!

- The linear regression model that we're about to discuss further is an example of a machine learning model!

- "Branch of AI that provides systems with the ability to learn without being explicitly programmed…"

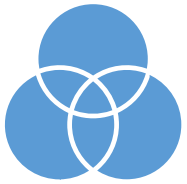
  - *Accenture*

LR = Machine learning!

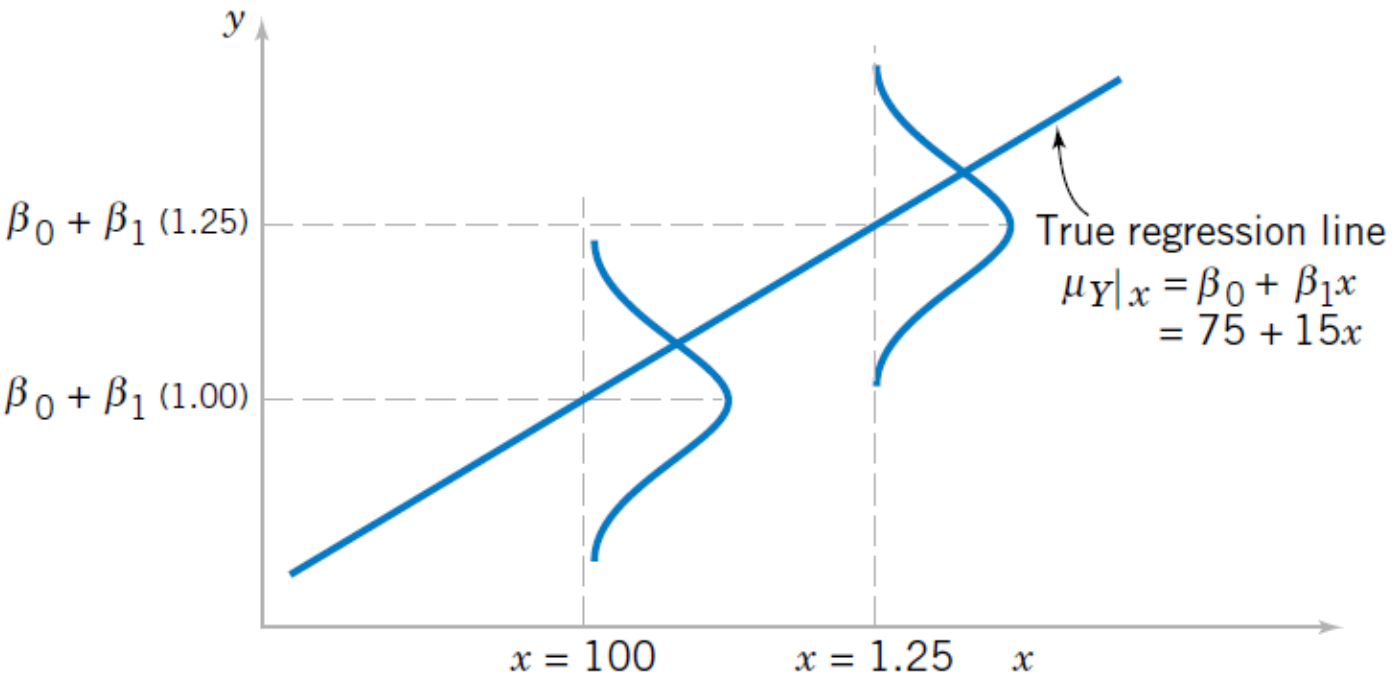


- Train the model using our sample data
  - Model "learns" the unknown parameters from our training data
- Assess model performance using test data
- So that ultimately, we can use the model, e.g. to make predictions, to draw conclusions

So why are linear models so popular?



Linear regression



Montgomery and Runger

Method of least squares

- To estimate the unknown coefficients β_1 and β_0, we find the "best-fit" line through the sample of points that we have collected

- This "best-fit" line is defined as the one that minimises the sum of squared errors, i.e.

$$\sum_{i=1}^{n} \varepsilon_i^2$$

Estimated line

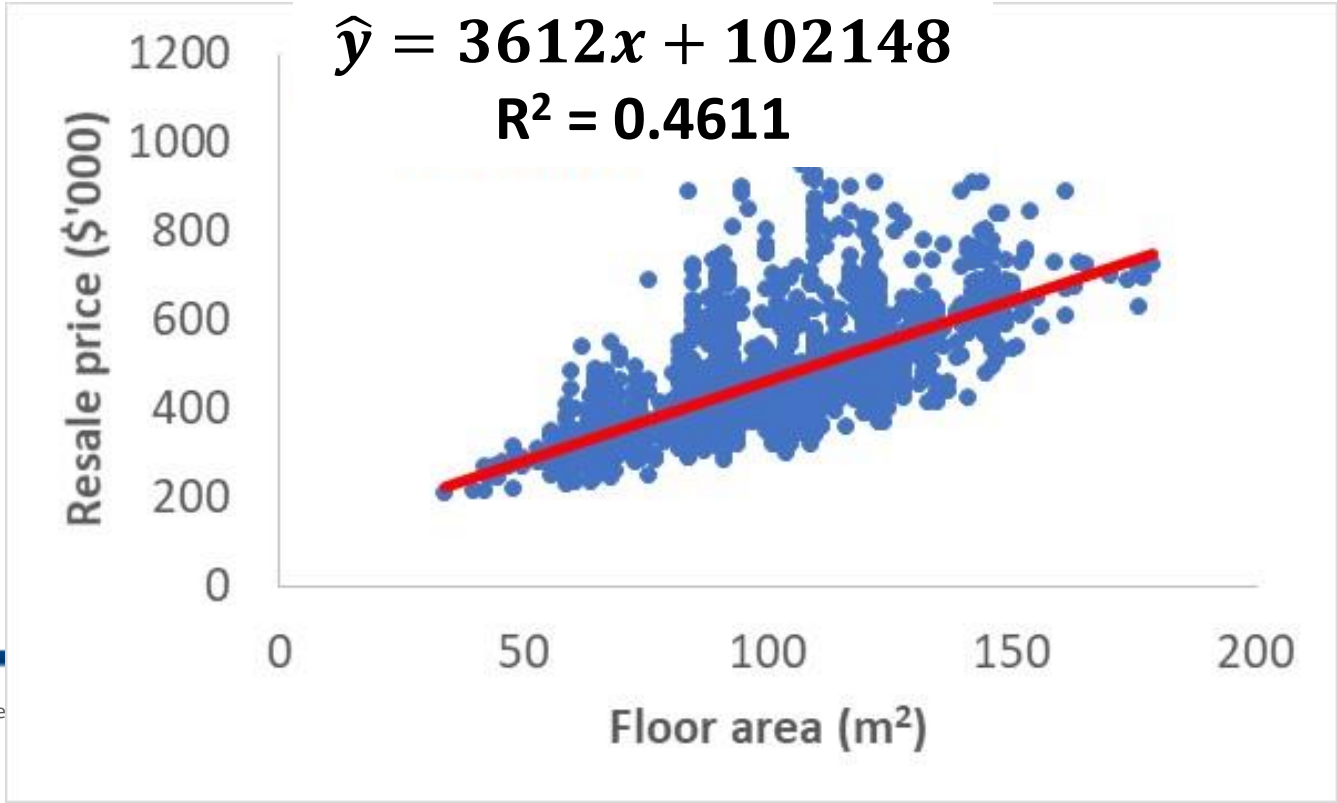
- The line that we have obtained is an estimated line, fitted to the one sample of data we have collected

- Similar to the estimate of $\bar{X}$ that we have obtained from our one sample earlier

Applications of regression

- Once we have obtained the coefficients of the estimated line, we can use it to predict future values of Y, based on specific values of X

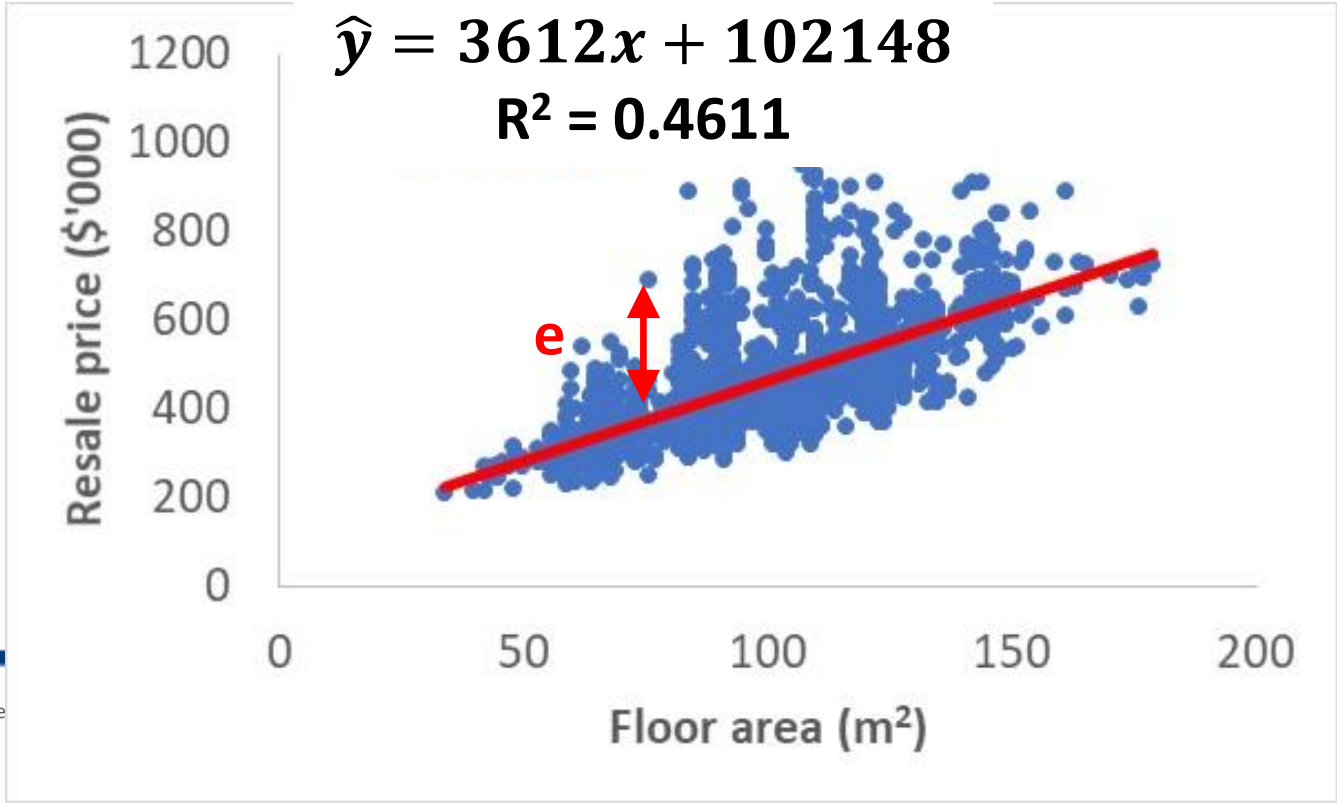
The fitted line $\hat{y}$ provides an estimate for what we predict Y to be, given a certain value of x.



Residuals

- The difference between each point and the estimated line is our residual, e

The residual can be thought of as an estimator for the error, ε



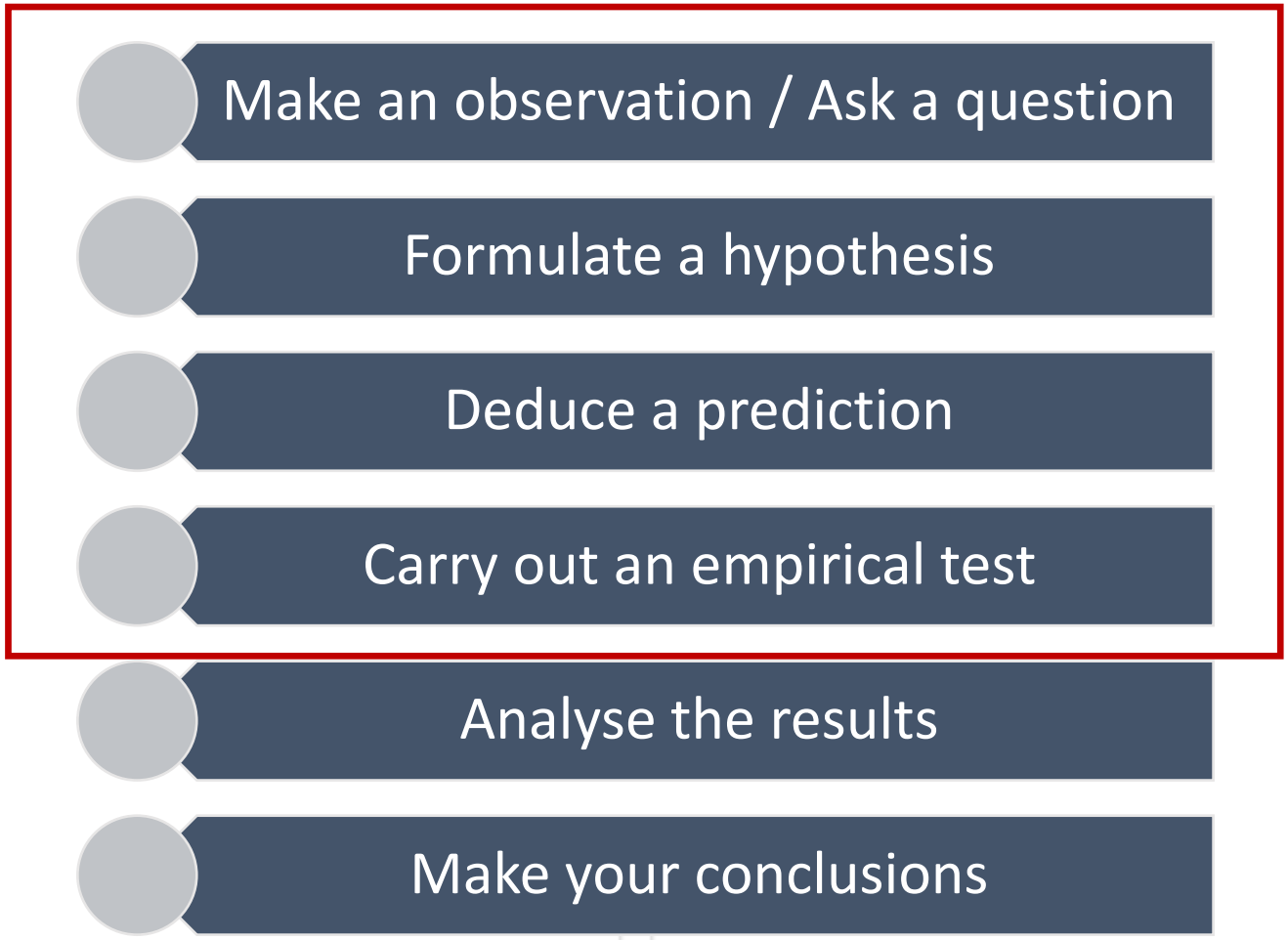
Reminder – Aug 2021

- Do get started on your group project
- Register for Qlik Sense Business account

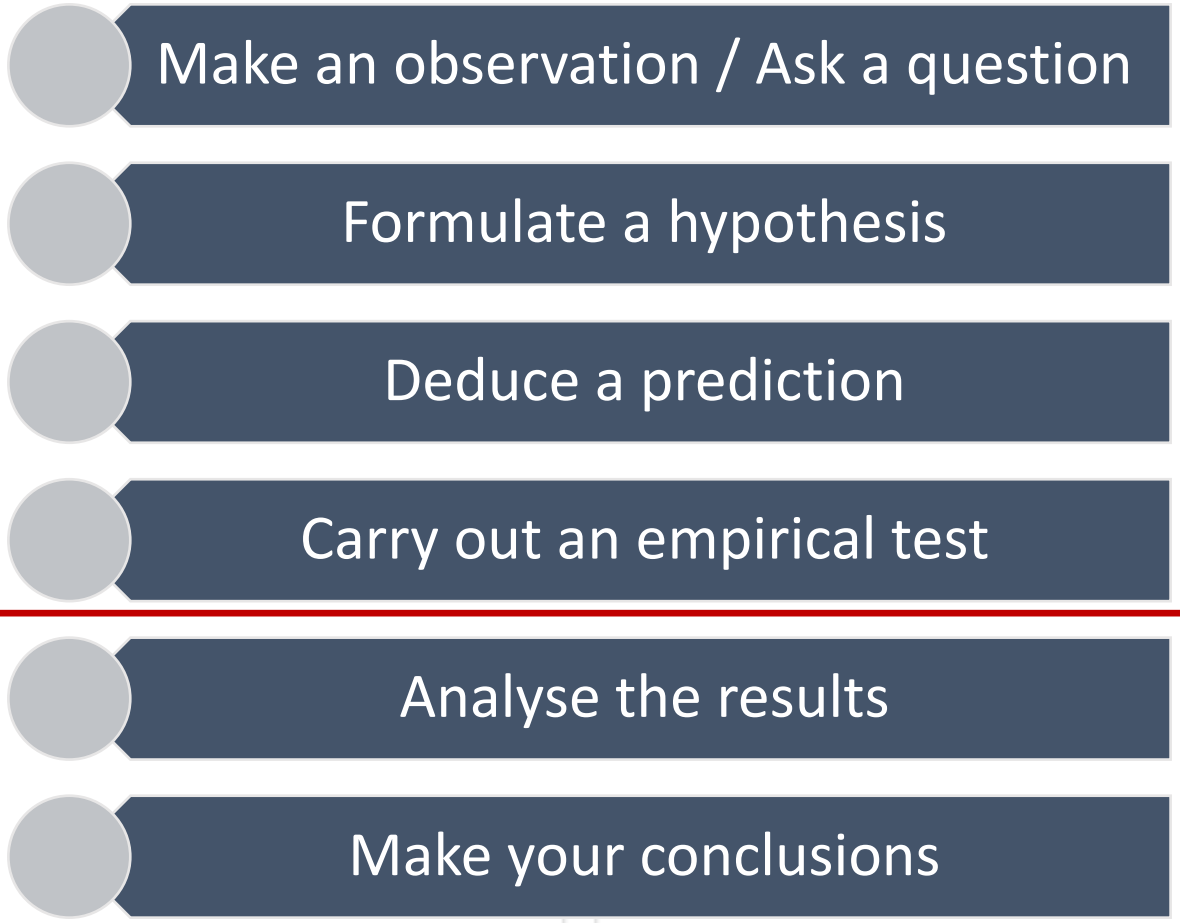
Tutorial 4

- Let's build our linear regression model from the "resale-sample.csv" file, predicting resale prices from floor area

Answering questions involving associations



Answering questions involving associations

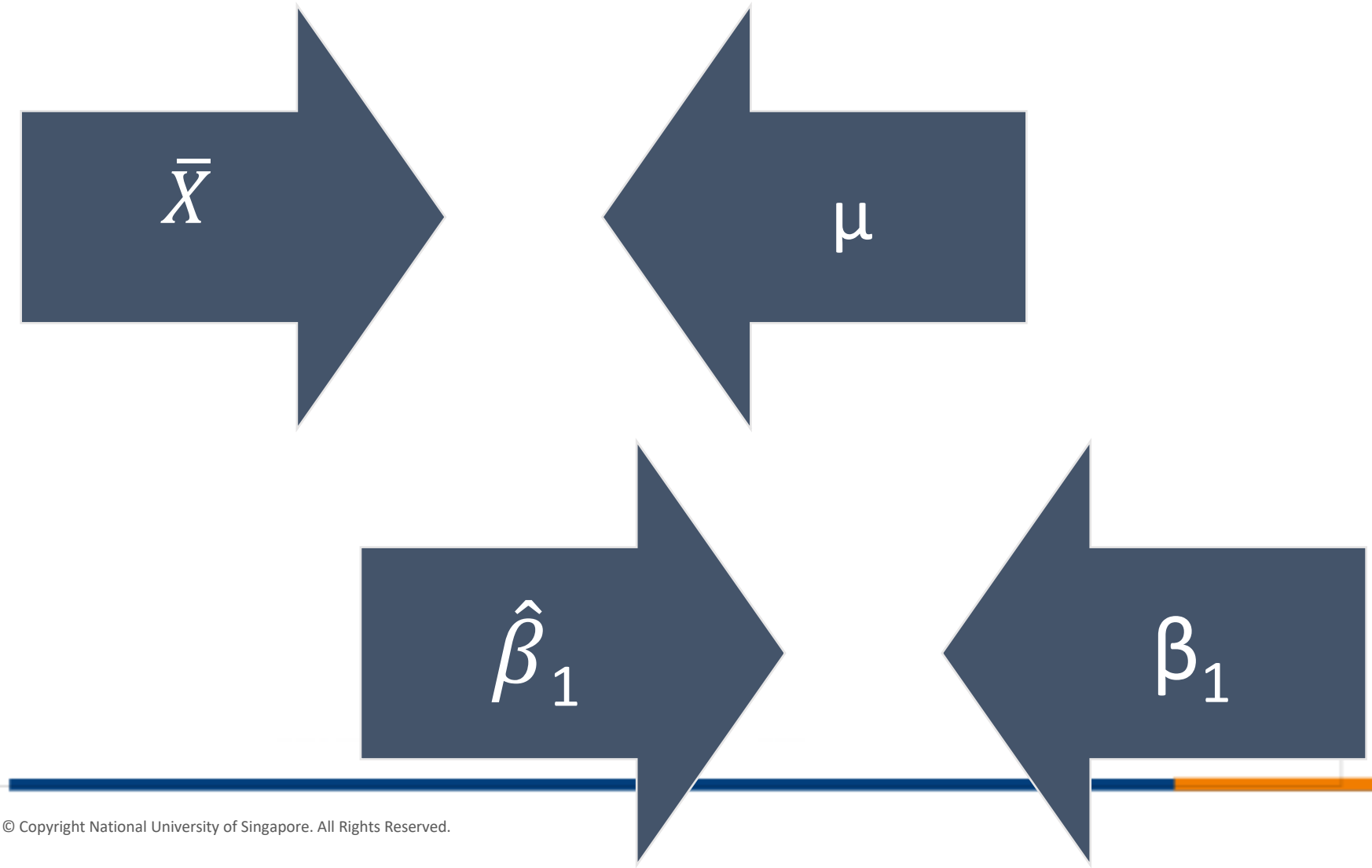


Significance of regression

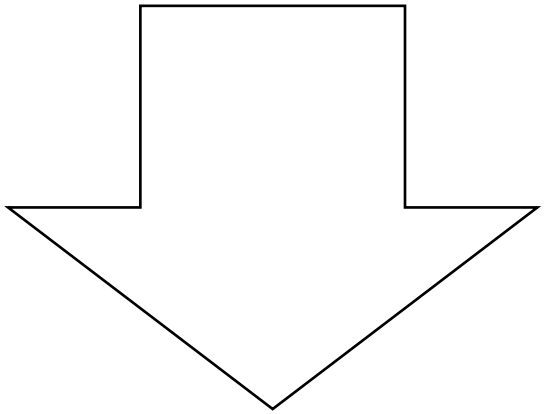
Is the sample telling us the actual mean is 60 kg?

Is the sample telling us that the slope of the actual line is 0?

Estimators

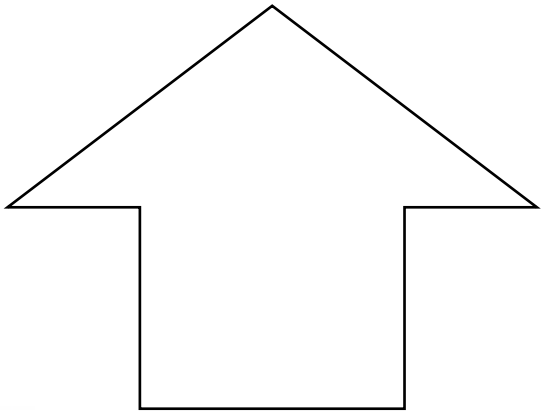


Significance of regression



If the slope of our fitted line is far from 0…

What does it say about our "population line"?



Significance of regression

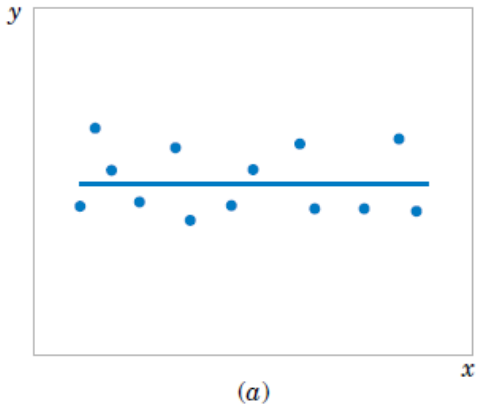
$$H_0: \beta_1 = 0$$

$$H_0: \beta_1 \neq 0, > 0, < 0$$

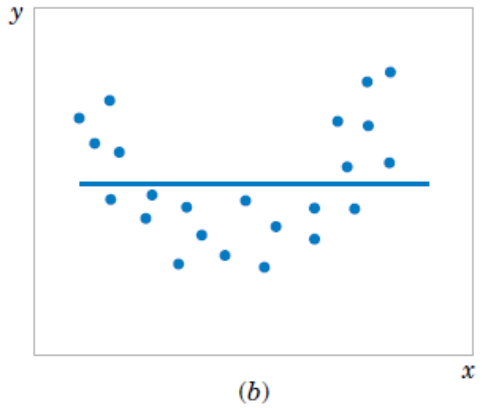
Why 0 specifically?

- If H_0: $\beta_1 = 0$, what does it mean?

- If we do not reject H_0, then it implies either
  - X is of little value in explaining the variation in Y and the best estimator for Y is $\bar{Y}$
  - The true relationship between X and Y is not linear



(a)



(b)

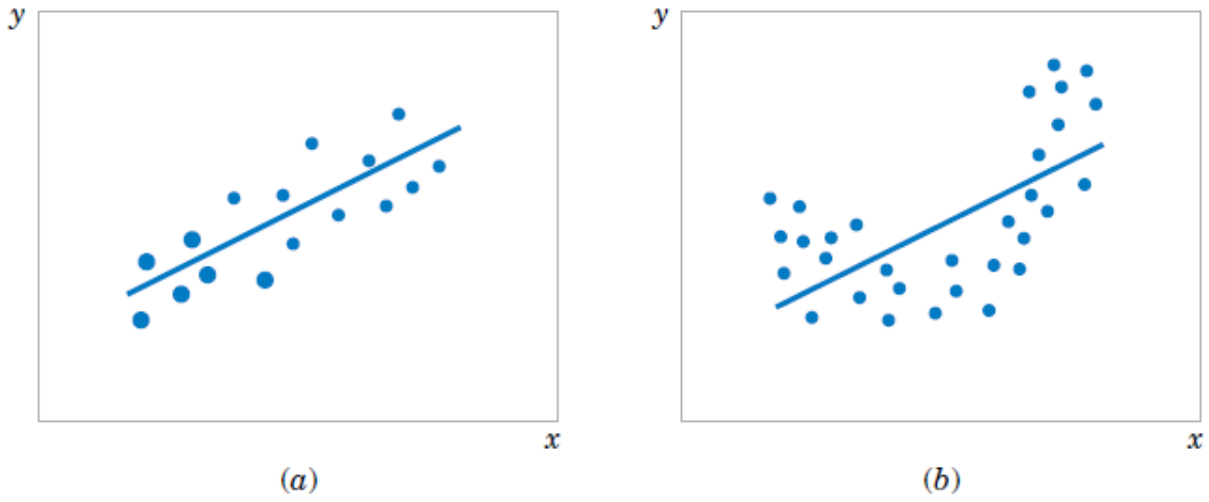
H_0 not rejected

Montgomery and Runger

Why 0 specifically?

- If we reject H_0, then it implies that X is of value in explaining the variability in Y, and means either
  - The linear model is adequate
  - Although there is a linear effect of X, better results could be obtained with the addition of higher polynomial terms in x

H_0 rejected



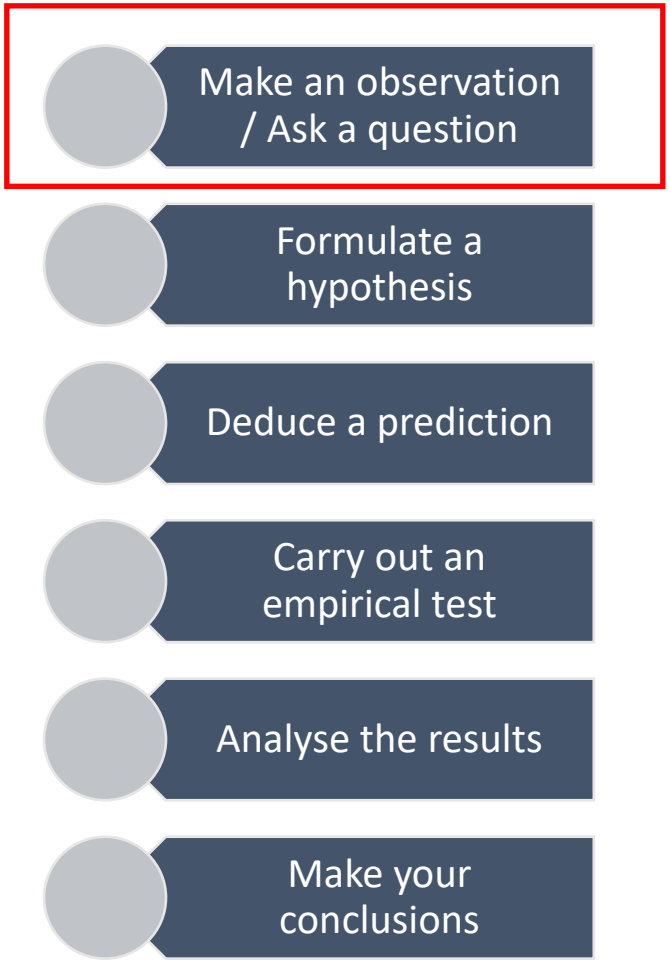
Tutorial 4

- Let’s take a look now at whether our sample line is actually saying if there is a relationship between resale price and floor area

Recap

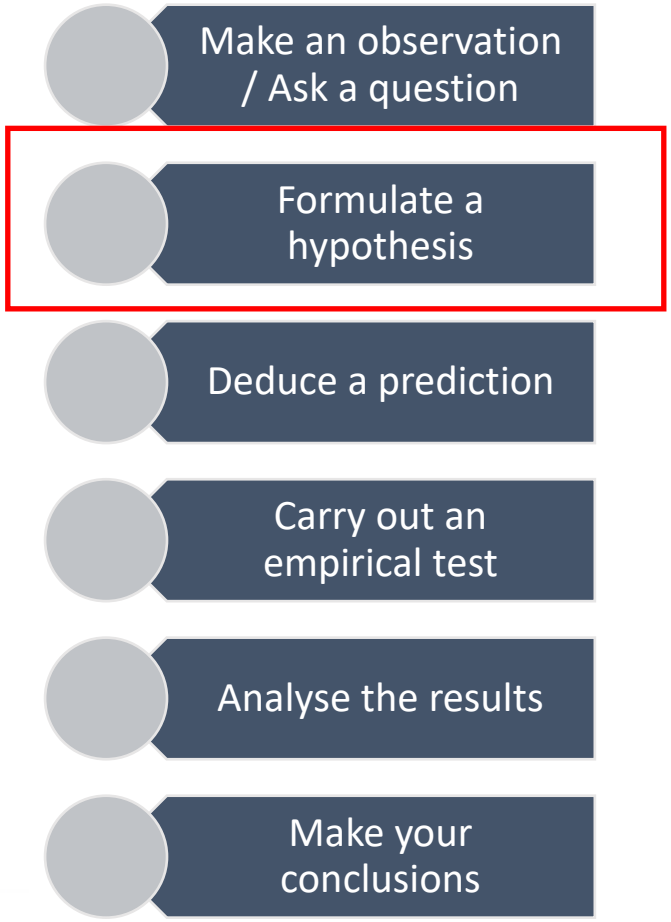
- Used a linear regression model to answer the following question:

- How much can I sell my flat?



Recap

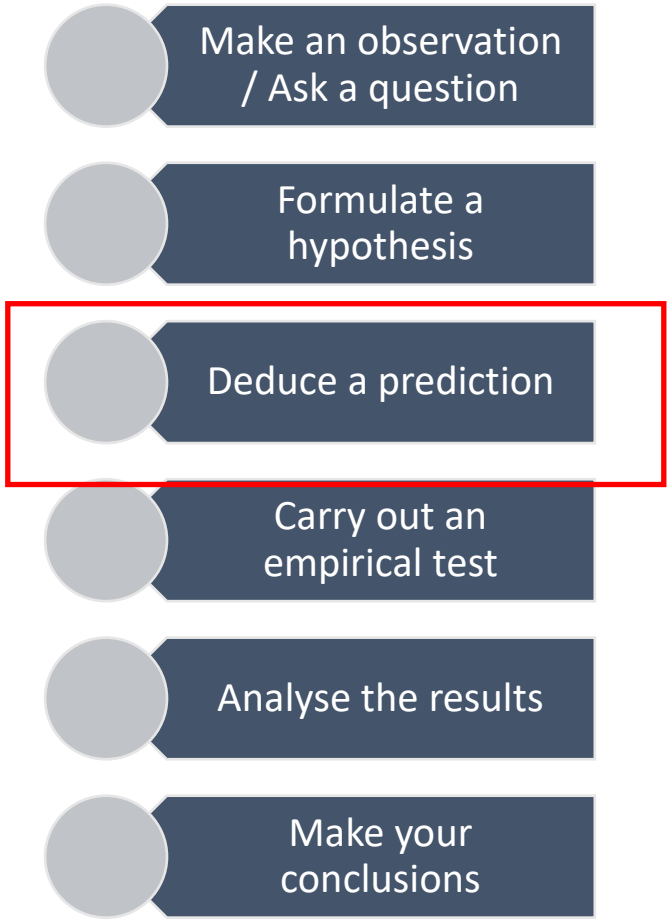
- (based on my observations) The bigger my flat, the more valuable it will be
- Note that my hypothesis now involves a relationship



Recap

- Operationalise
  - Bigger → flat area
  - Valuable → resale price

- Prediction: we will observe a higher resale price with higher flat area

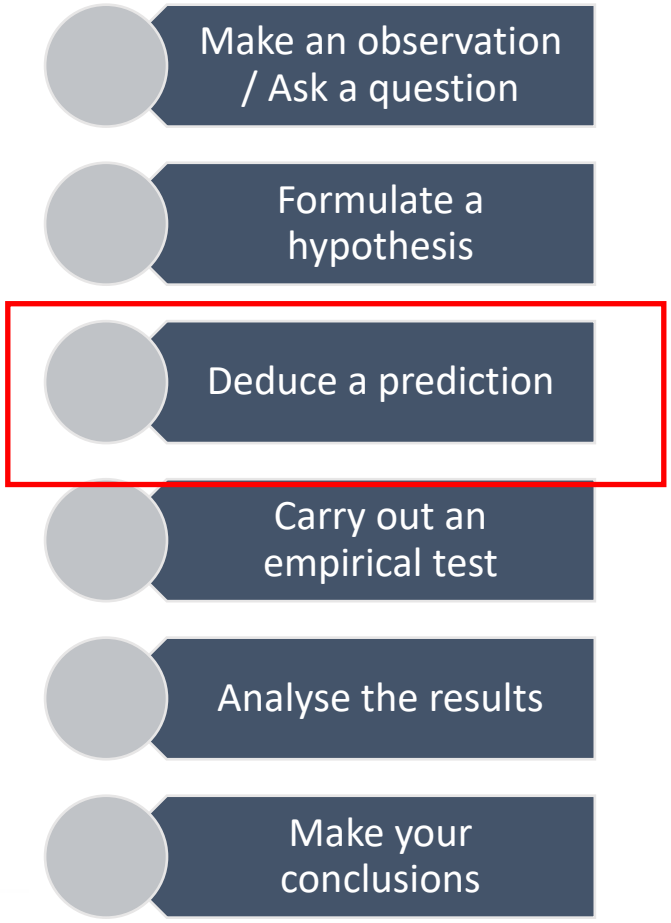


Recap

- We can use a linear regression model to mathematically represent this relationship:

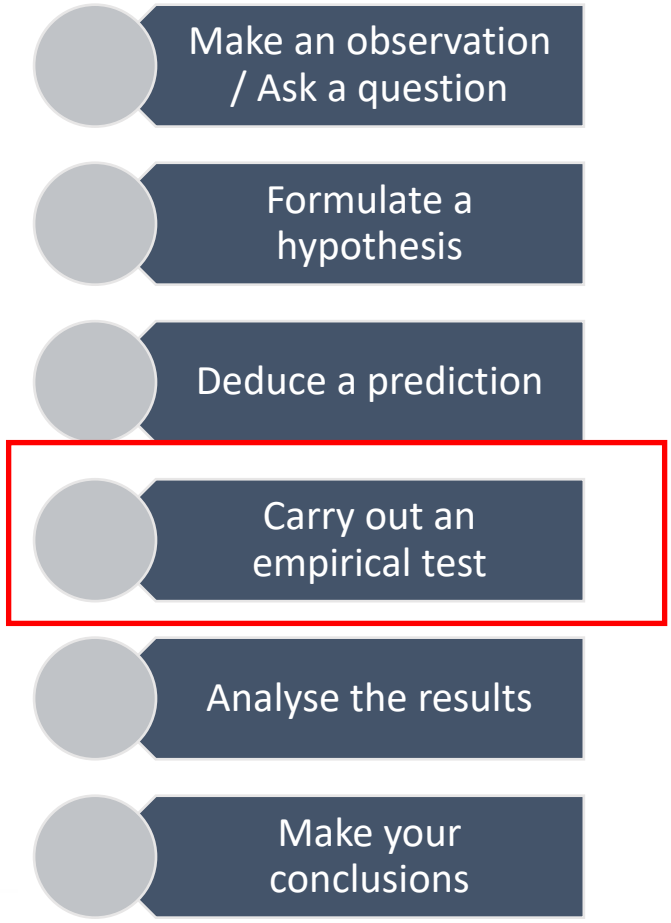
$$y = \beta_0 + \beta_1 x + \epsilon$$

- So that we can use the data we collect to assess if the prediction is accurate



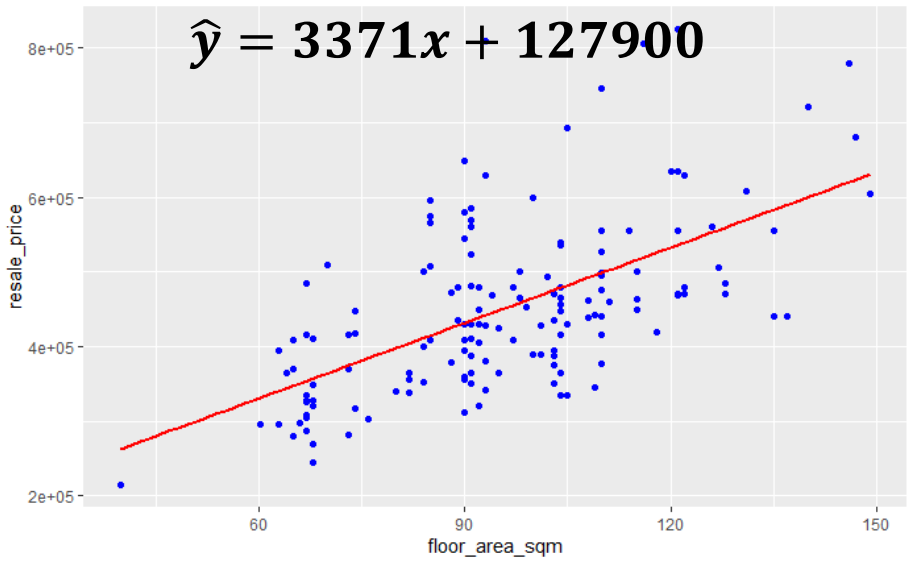
Recap

- Collecting data from a sample of 150 resale transactions from 2012 – 2017, as described in the “resale-sample-small.csv” dataset
- This is a sample from our population of interest
- What type of empirical test is this?



Recap

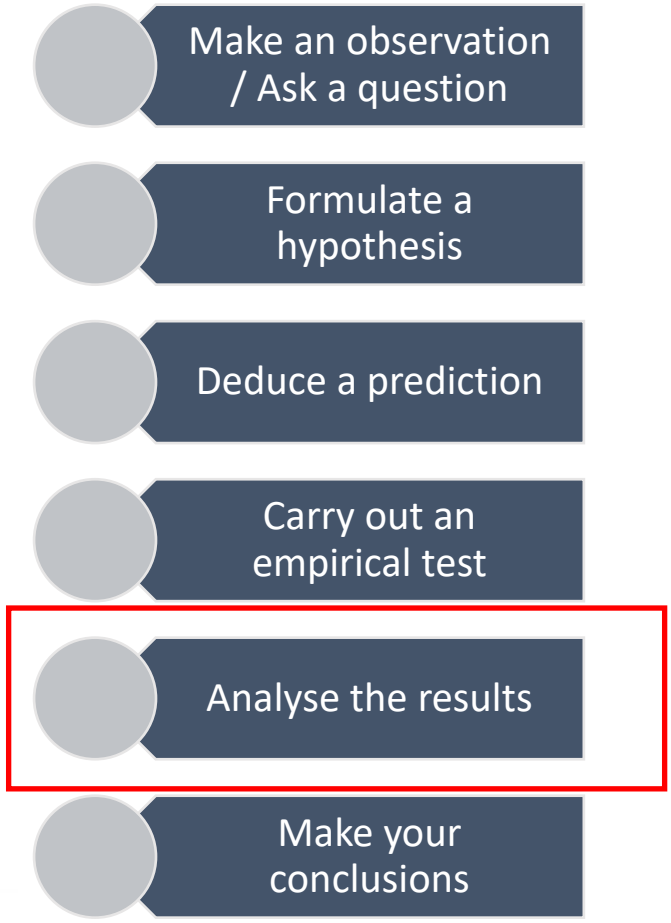
- Use the method of least squares to construct the best-fit sample line



- Make an observation / Ask a question
- Formulate a hypothesis
- Deduce a prediction
- Carry out an empirical test
- Analyse the results
- Make your conclusions

Recap

- Carried out a hypothesis test to assess if the value of $\hat{\beta}_1 = 3371$ is telling us which of the following statements is true:
  - $H_0: \beta_1 = 0$
  - $H_1: \beta_1 > 0$

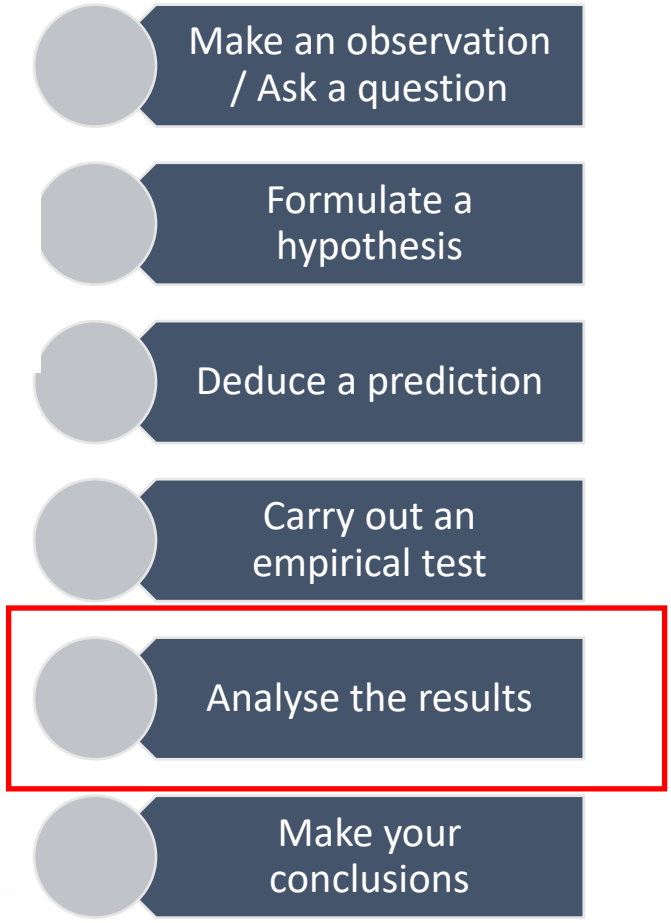


Recap

- Since p-value is small

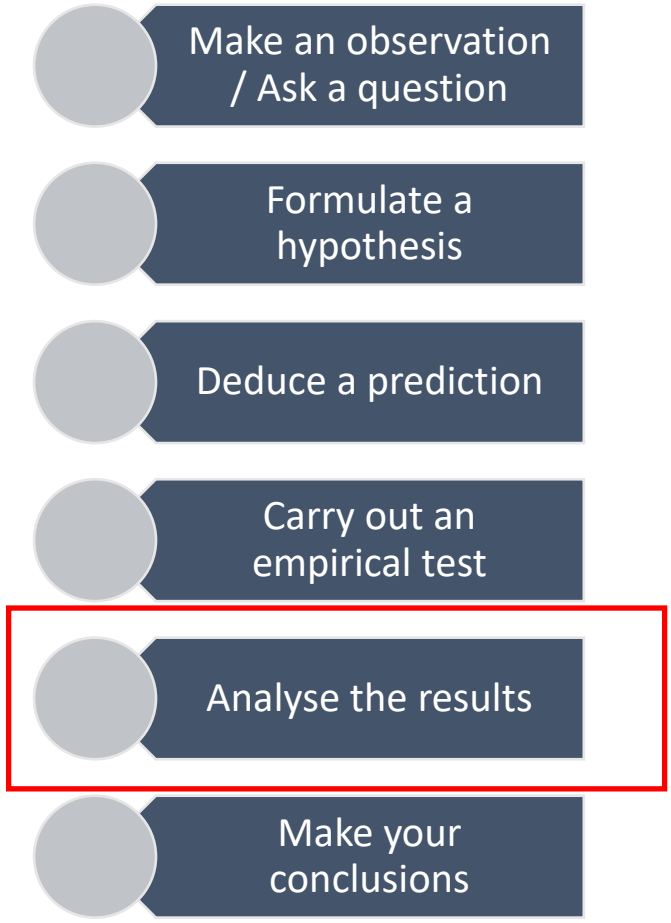
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    127900.9    37040.4   3.453 0.000723 ***
floor_area_sqm   3371.3      378.2   8.914 1.69e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- We reject H_0 → our prediction is accurate → there is support for our hypothesis



Recap

- However, as we have made use of the linear regression model in our analysis
- Before we can conclude, we need to check the model's assumptions
- To ensure our analysis is valid



Assumptions of regression

- Functional form:
  - The model is linear in the coefficients

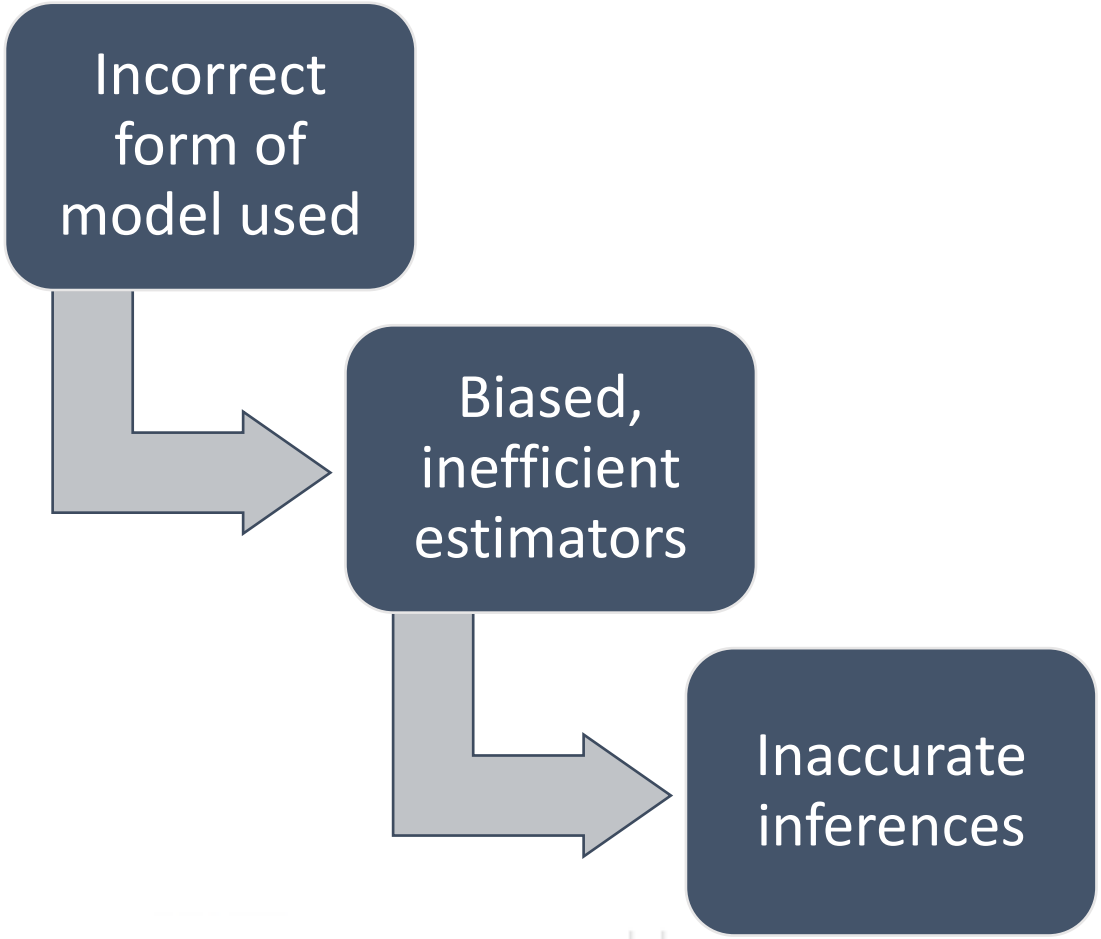
$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Exogeneity:
  - The independent variable, X, is uncorrelated with the error term, ε

Assumptions of regression

- The error term is normally distributed
- Observations of the error term are uncorrelated with each other (no autocorrelation)
- The error term has a population mean of 0, and has a constant variance (homoscedasticity)

Implications of assumption violations



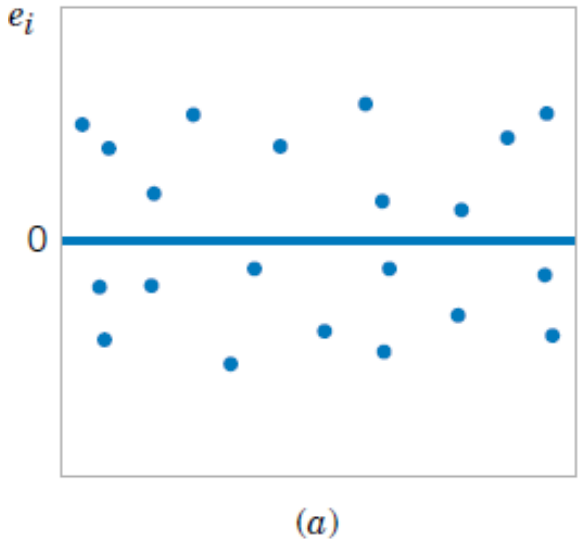
How to check these assumptions?

Normal probability plot, histogram

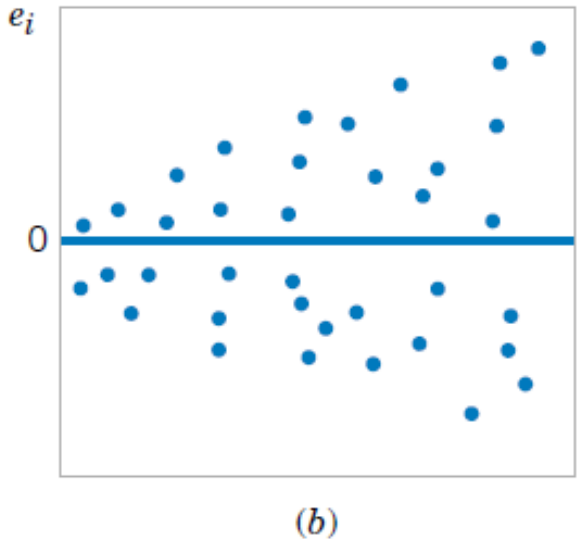
- Normality assumption

Residual plots, R^2, r

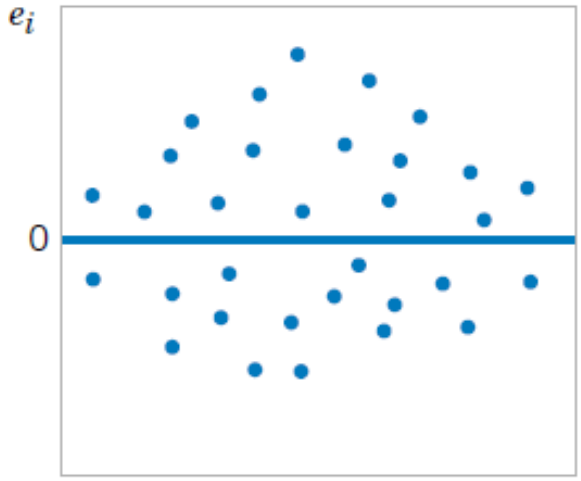
- Functional form, exogeneity, independence, constant variance



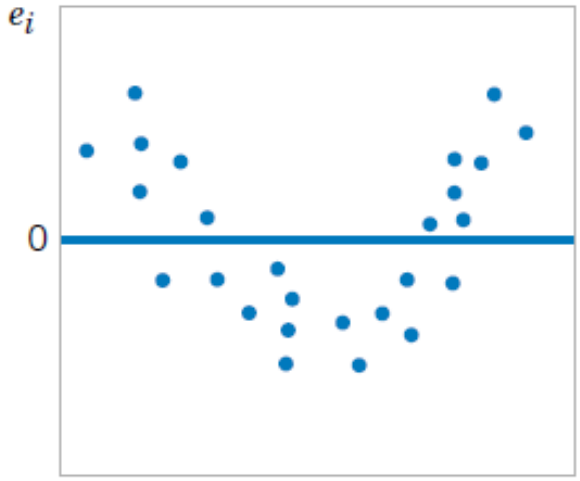
(a)



(b)



(c)



(d)

Adapted from Montgomery and Peck (1992)

Model adequacy checks

- The metric R^2 that we saw on the previous slide, is the coefficient of determination
- It varies from 0 to 1, and describes how much of the variability in Y (the dependent variable) is accounted for by the regression model
- An R^2 value closer to 1 generally means a better model

Coefficient of determination

- Always possible to make R^2 equal 1 by adding more terms to the model
  - E.g. a perfect fit can be obtained for n data pts with a polynomial of degree (n-1)

- R^2 can be large, even if x and y are related in a non-linear fashion

- Though R^2 may be large, it does not necessarily mean that model will provide accurate predictions of future observations

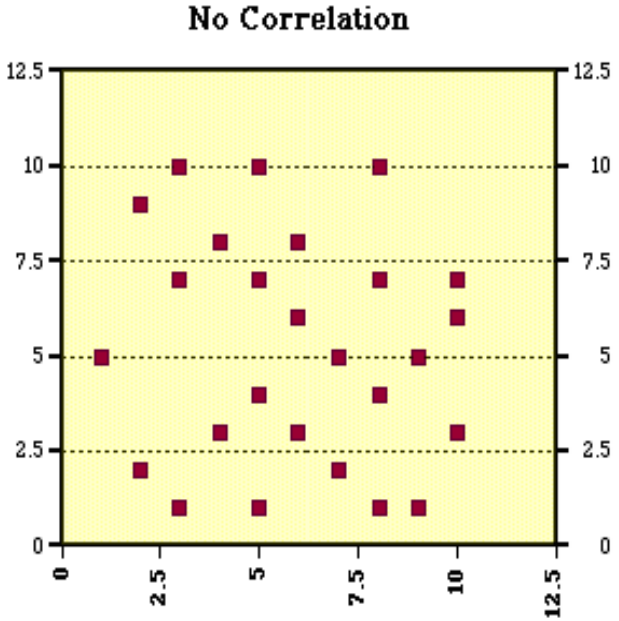
Coefficient of determination

- What if R^2 is low? Is it bad?
  - Depending on the specific case, it may be entirely expected that R^2 values will be low(<50%), e.g. human behavior studies
  - Even if we have low R^2 values, if we have statistically significant coefficients, we can still draw important conclusions about relationship between IVs and DV

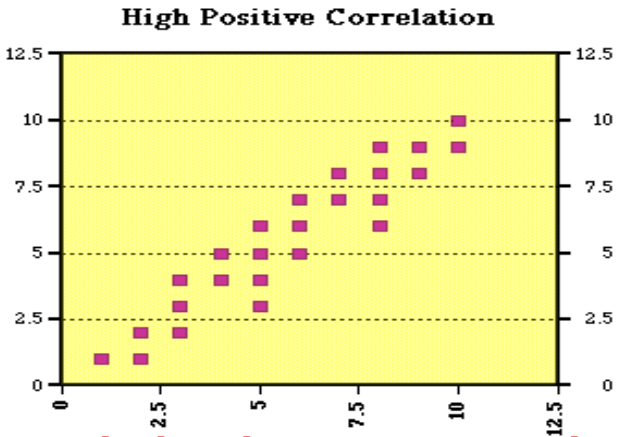
Model adequacy checks

- Another metric to assess model adequacy is the correlation, r
- The correlation describes the strength of the linear association between two variables
- It varies from -1 to 1. The closer the correlation is to -1 or 1, the stronger the negative and positive linear association

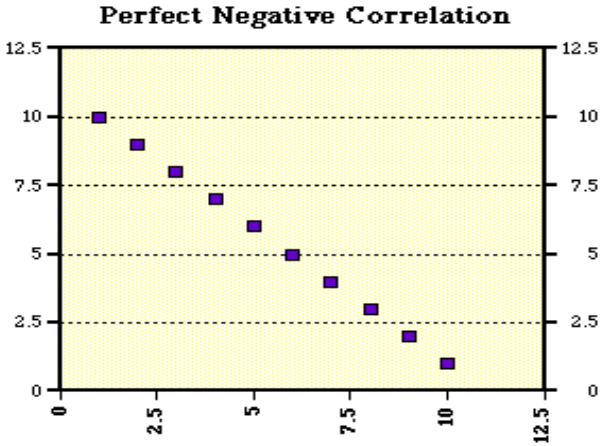
Correlation



Correlation Value is close to Zero, -0.3 < r <0.3



Correlation is NOT Causation!



Correlation and regression

Correlation

- Tells us how much one variable tends to change, when the other changes

Linear regression

- Find best fit line for predicting Y from X

Tutorial 4

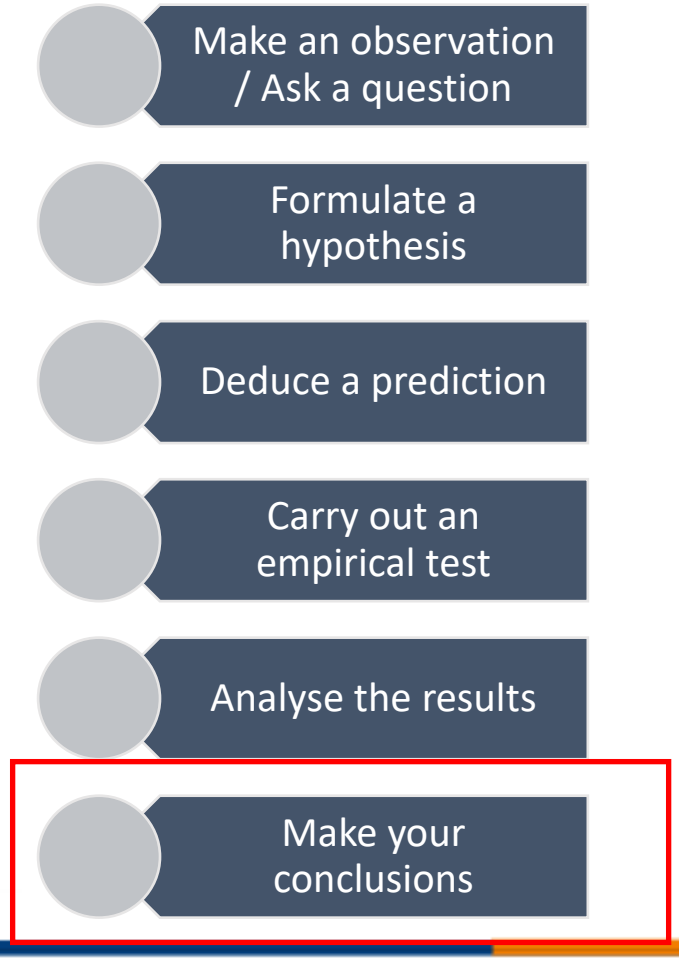
- Before we conclude, we need to check our assumption to see if the linear model is adequate

- What do the results say?

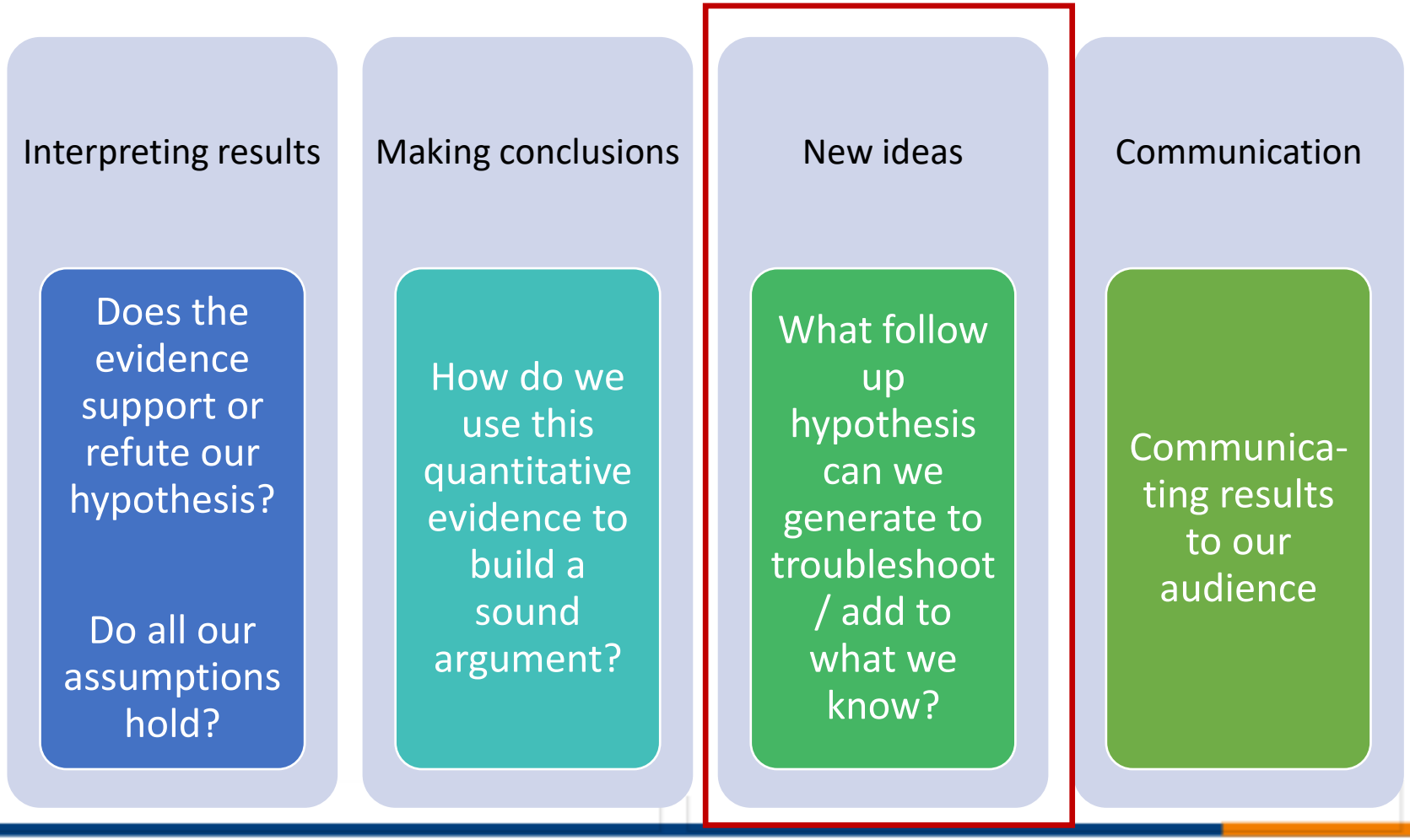
Making our conclusions

What if assumptions are not met?

- Then the use of the model won't be valid
- Analysis won't yield accurate results
- We won't be able to say if the prediction is true or not, and thus if the hypothesis has quantitative support
- Then we might need to come up with new ideas (see next slide)



Make our conclusions



Making our conclusions

- Let's investigate the lease commence date as another possible predictor
- Earlier date → lower resale price and vice versa
- More than 1 predictor → MLR

Tutorial 4

- What are the coefficients of the model?
- What are the results of the hypothesis testing on each coefficient of the model?
- What do we notice about the R^2?

Adjusted R^2

$$R^2{}_{adj} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

- Adjusted R^2 is "adjusted" based on the number of predictors in our model, p
  - N is the sample size
- It doesn't increase from the scenario when we have less predictors, just because we have more predictors
- But only if the additional predictor improves the model more than what we would expect by chance

Multicollinearity

- All the assumptions that come with regression using one predictor also holds for multiple predictors

- In addition, for MLR, we also need to check multicollinearity

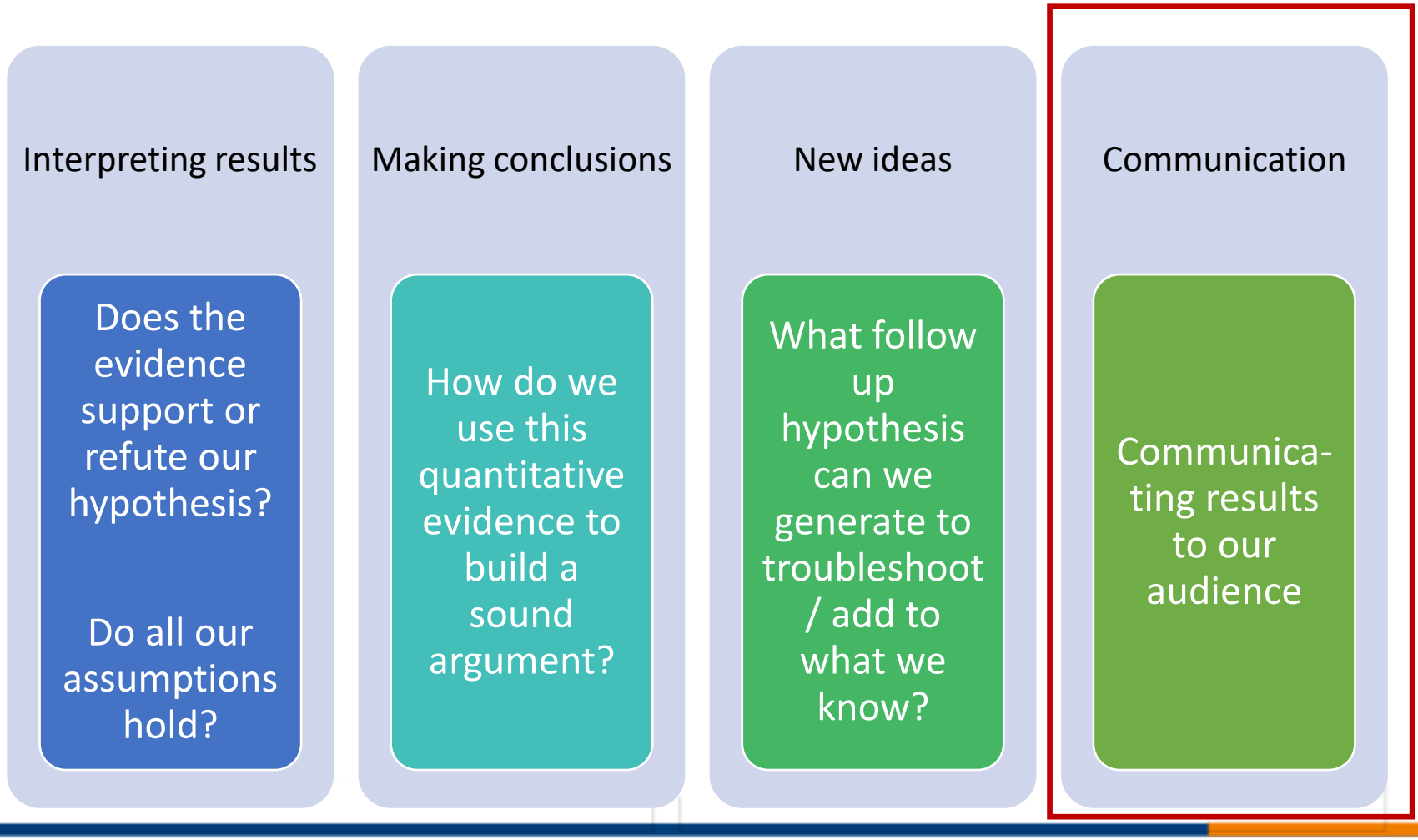
Multicollinearity

- "multicollinearity" refers to predictors that are correlated with other predictors.
- Variance inflation factor (VIF) assesses how much the variance of an estimated regression coefficient increases if your predictors are correlated. If no factors are correlated, the VIFs will all be 1.
- 5< VIF <10 indicates high correlation that may be problematic; if the VIF > 10, you can assume that the regression coefficients are poorly estimated due to multicollinearity.

Tutorial 4

- Let’s compute the variance inflation factor for our MLR model, and assess if multicollinearity is present

Make our conclusions



Epilogue

Strengths	Concerns
• More precise in our definition of the problem, and in the conclusions we can draw • Reliable basis upon which to decide, convince, and predict* • Allows us to track performance across place and time	• Can distort the analysis and conclusions if an unrepresentative / convenient sample is taken • May derive overly simplistic conclusions, ignoring richer details

Epilogue

- With the concepts and skillsets that we have seen…

Active, effective "engagers" of data

Informed consumers of data that we encounter

End of Day 4