TELKOM
DIGITAL TALENT
INCUBATOR 2020

Modul 4: Regression

Dr. Nurvita Trianasari, S.Si, M.Stat.

Organized by:
CAE

Module Overview

Topics

- Regression Analysis
- Type of Regression
- Simple Linear Regression
- Multiple Linear Regression

Activities

- Group Discussion
- Coding Practice

Module Objectives

- Understand what is Regression Analysis
- Create Regression Model using Python
- Use regression analysis to predict future values

What is Regression?

- Regression analysis is a tool for building statistical models that characterize relationships among a dependent variable and one or more independent variables.

Purpose of Regression

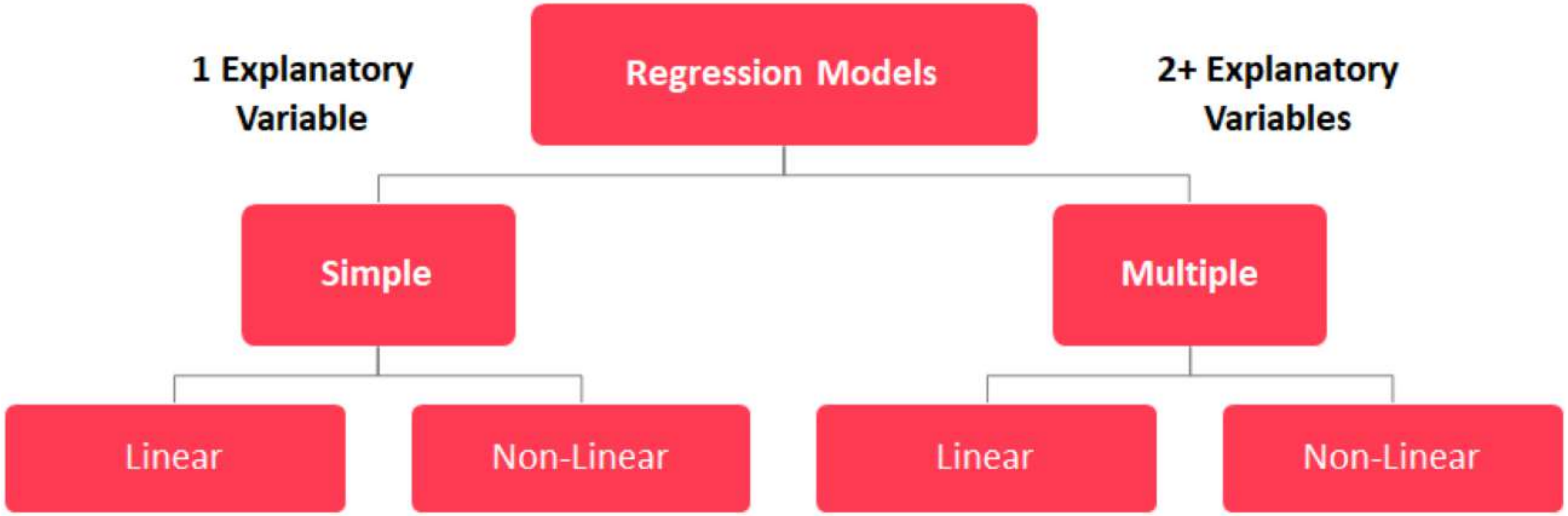
- The purpose of regression analysis is to analyze relationships among variables.
- The analysis is carried out through the estimation of a relationship and the results serve the following two purposes:
  - Answer the question of how much y changes with changes in each of the x's (x1, x2,...,xk), Y is the dependent variable
  - Forecast or predict the value of y based on the values of the X's. X is the independent variable

Step of Regression Analysis

A regression analysis can be broken down into 5 steps.

- Step 1 : state the hypothesis.
- Step 2 : test the hypothesis (estimate the relationship).
- Step 3 : interpret the test results. This step would enable us to answer the following questions,
- Step 4 : check for and correct common problems of regression analysis.
- Step 5 : evaluate the test results.

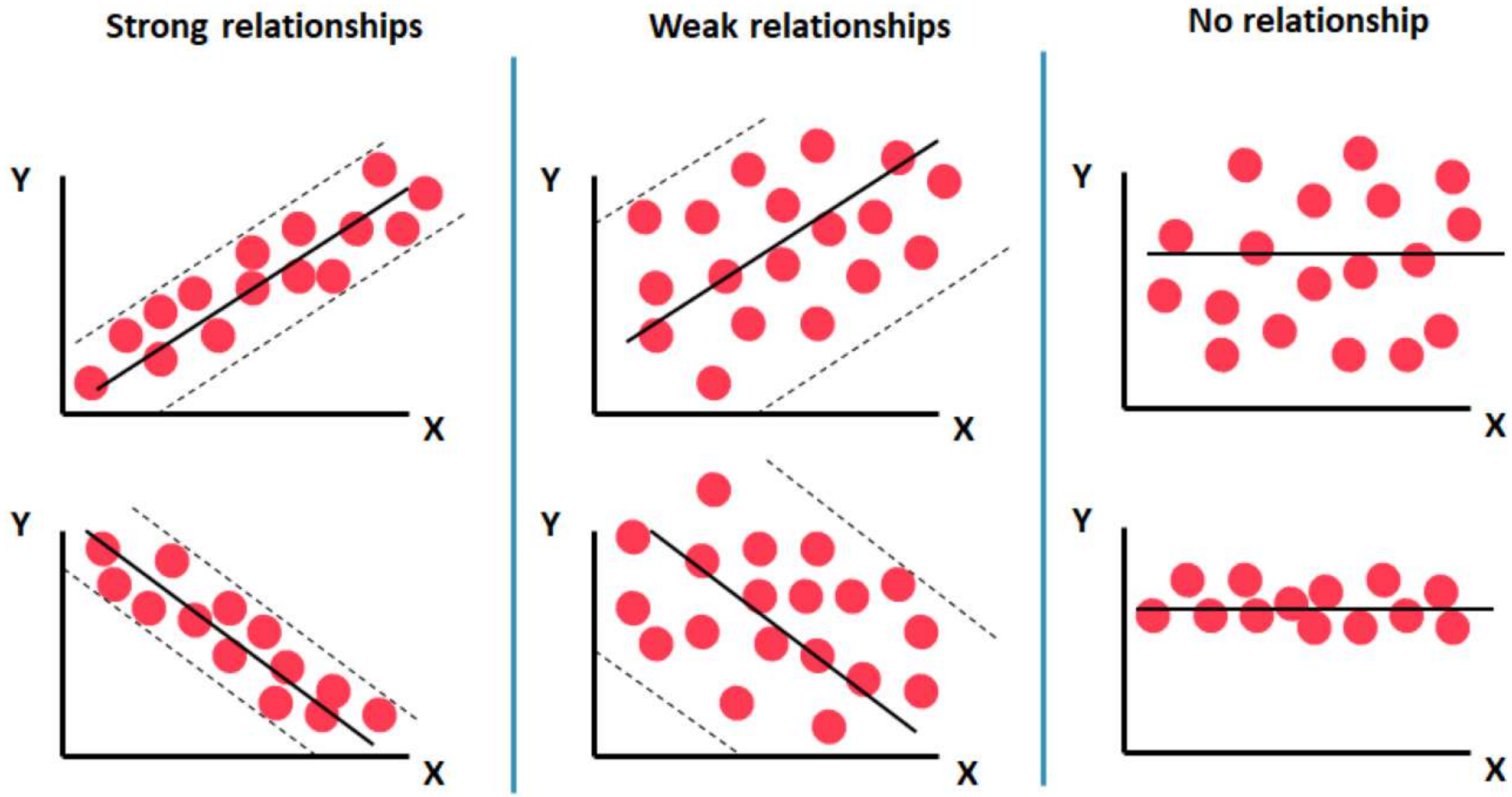
Type of Regression



Simple Linear Regression

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

Types of Relationship



Simple Linear Regression Model

Dependent Variable

Population Y intercept

Population Slope Coefficient

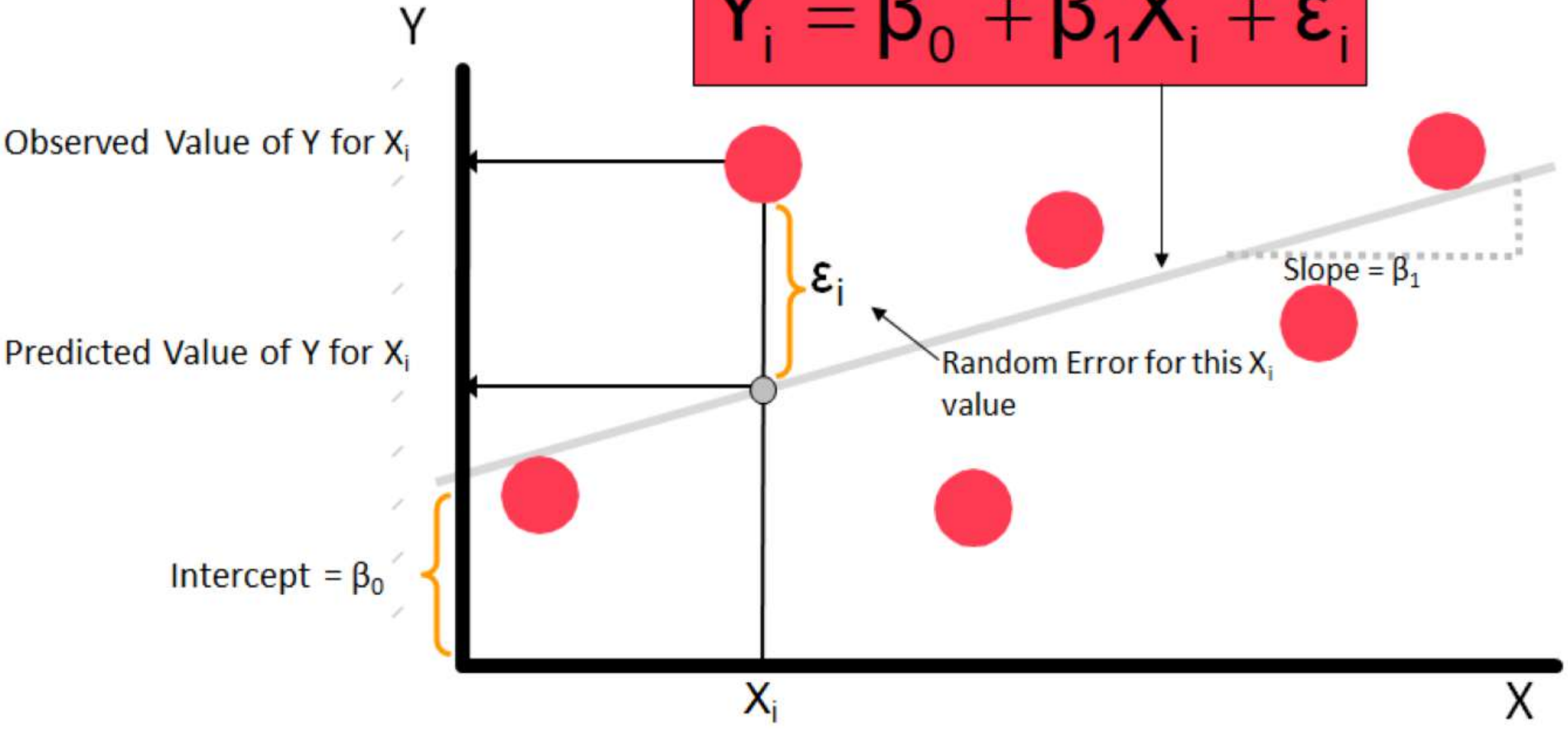
Independent Variable

Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



The Least Squares Method

- b0 and b1 are obtained by finding the values of that minimize the sum of the squared differences between Y and $\hat{Y}$:

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

Finding the Least Squares Equation

- The coefficients b0 and b1 , and other regression results in this chapter, will be found using python

Interpretation of Slope and Intercept

- b0 is the estimated average value of Y when the value of X is zero.
- b1 is the estimated change in the average value of Y as a result of a one-unit change in X.

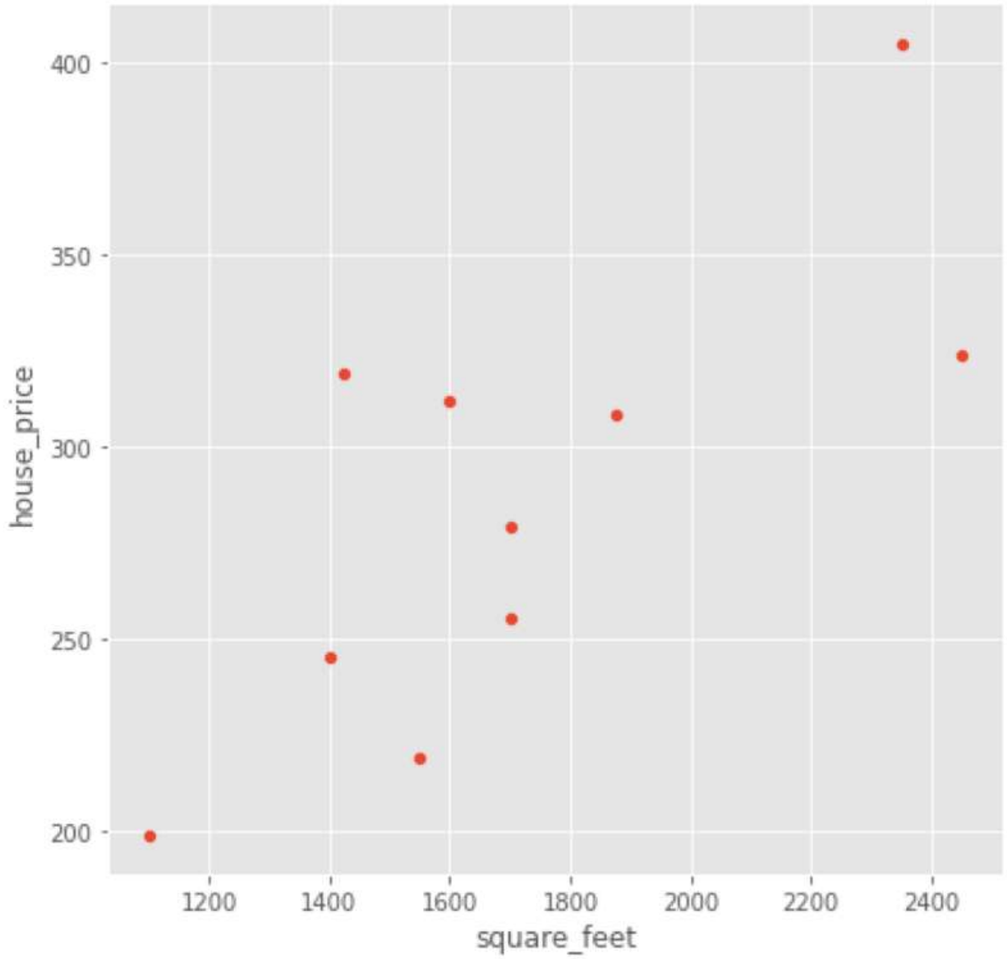
Example: Simple Linear Regression

- 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

Example: Data

	house_price	square_feet
0	245	1400
1	312	1600
2	279	1700
3	308	1875
4	199	1100
5	219	1550
6	405	2350
7	324	2450
8	319	1425
9	255	1700

Example: Scatter Plot

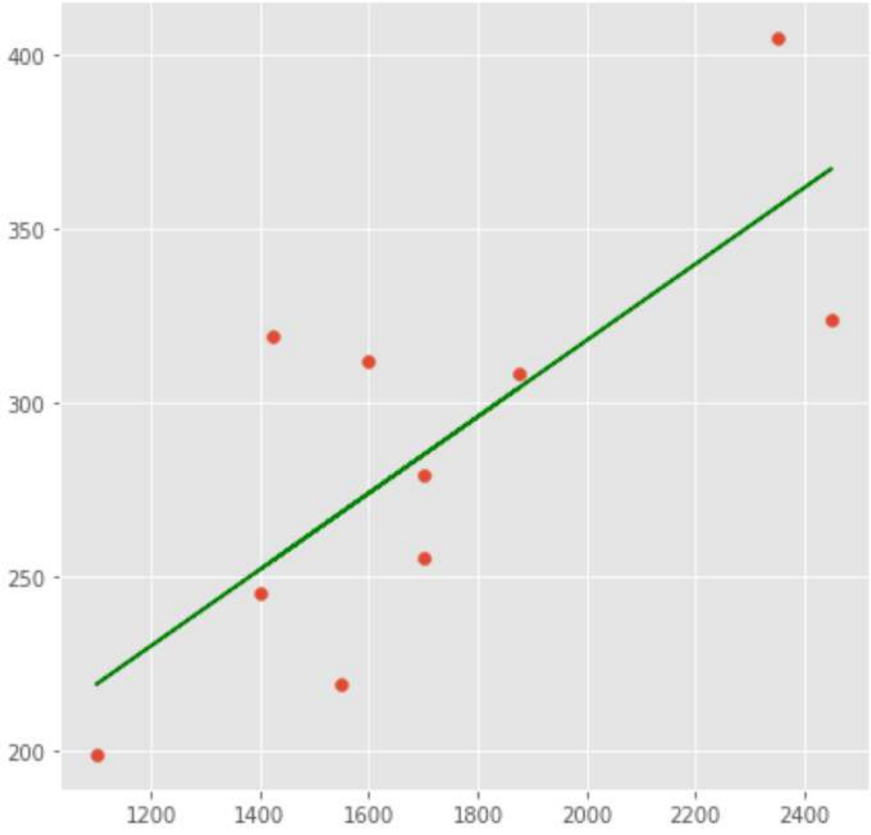


Example: Regression Equation

- Regression equation from Python
- **Slope** = 0.10976774
- **Intercept** = 98.24832962138078

```
print('Intercept: \n', regr.intercept_)
print('Coefficients: \n', regr.coef_)
```

```
Intercept:
 98.24832962138078
Coefficients:
 [0.10976774]
```



Example: Interpretation of b0

- b0 is the estimated average value of Y when the value of X is zero (if X = 0 is in the range of observed X values)
- Because a house cannot have a square footage of 0, b0 has no practical application

Example: Interpretation of b1

- b1 estimates the change in the average value of Y as a result of a one-unit increase in X
- Here, b1 = 0.10977 tells us that the mean value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Example: Making Predictions

- Predict the price for a house with 2000 square feet:
- House price = 98.25 + 0.1098(square feet)

  = 98.25 + 0.1098 (2000)

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

Measures of Variation

Total variation is made up of two parts:

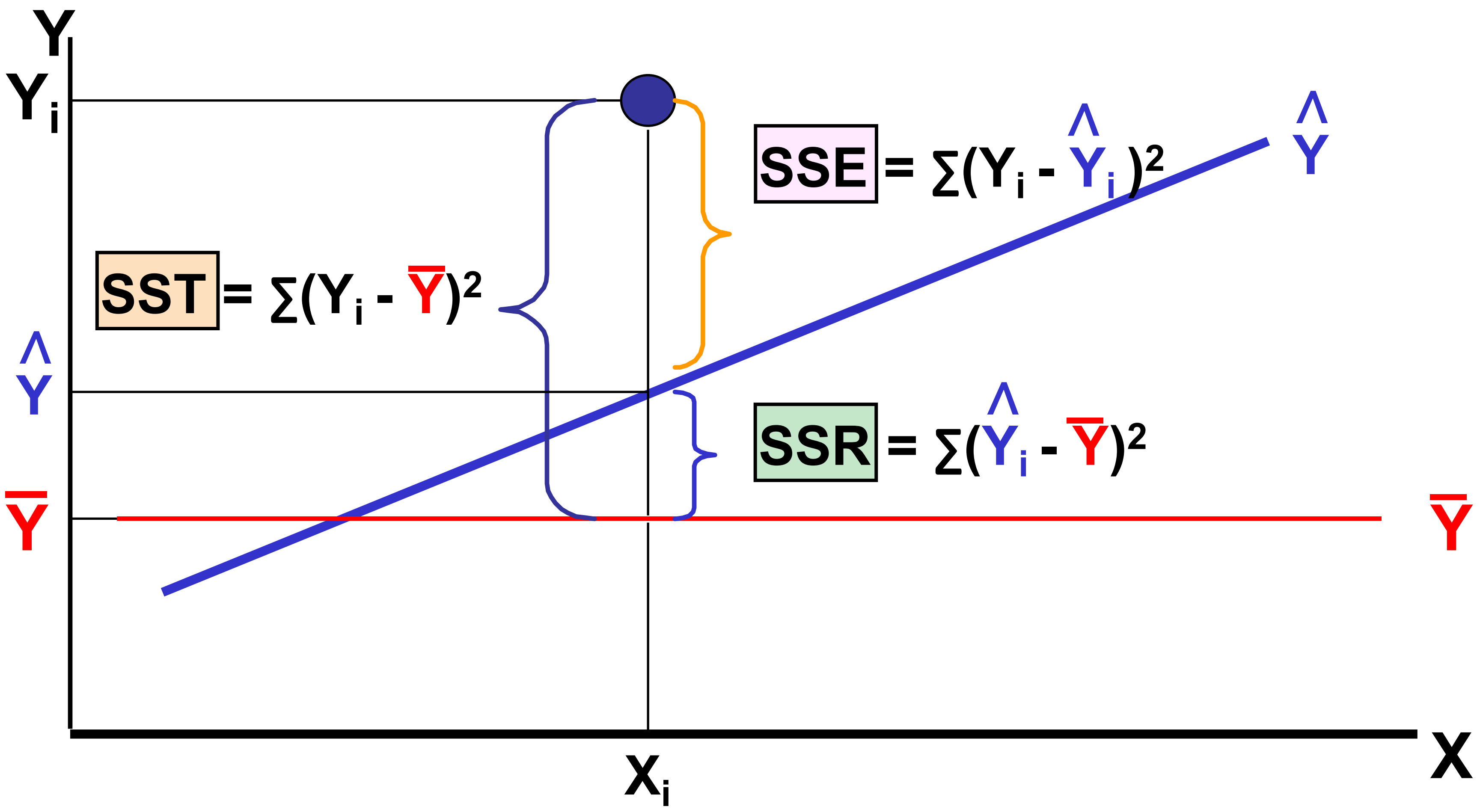
$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum (Y_i - \overline{Y})^2$	$SSR = \sum (\hat{Y}_i - \overline{Y})^2$	$SSE = \sum (Y_i - \hat{Y}_i)^2$

Measures of Variation

- **SST = total sum of squares (Total Variation)**
  - Measures the variation of the Yi values around their mean $\bar{Y}$
- **SSR = regression sum of squares (Explained Variation)**
  - Variation attributable to the relationship between X and Y
- **SSE = error sum of squares (Unexplained Variation)**
  - Variation in Y attributable to factors other than X

Measures of Variation



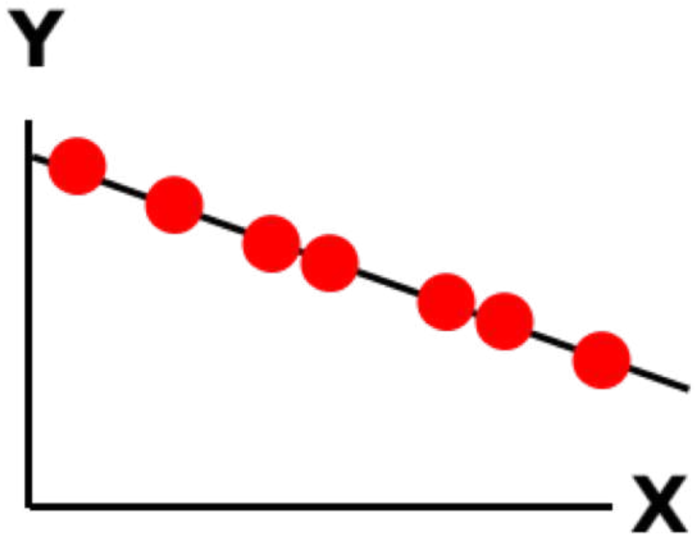
Coefficient of Determination, r2

- 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 r2
- Note $\boxed{0 \le r^2 \le 1}$

$$\boxed{r^2 = \frac{SSR}{SST} = \frac{\text{regression } \textit{sum} \text{ of squares}}{\textit{total} \text{ sum of squares}}}$$

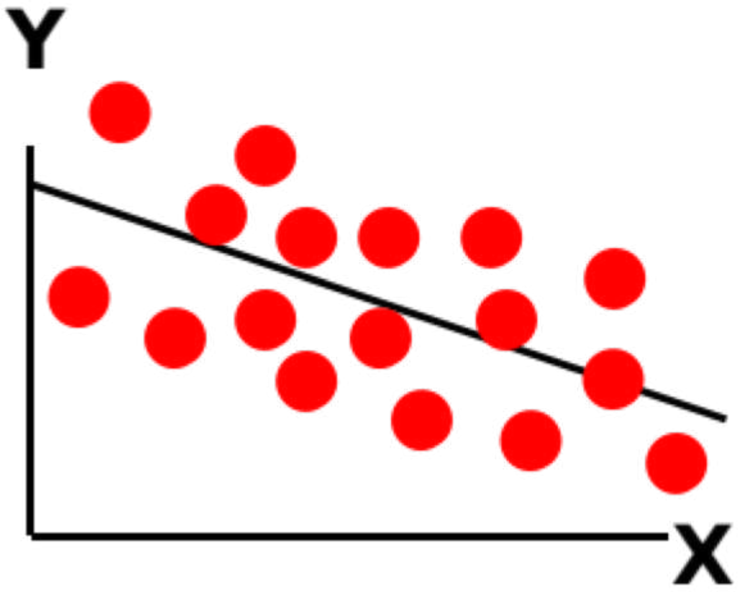
Examples of Approximate r2 Values

- **r2 = 1**
- Perfect linear relationship between X and Y:
- 100% of the variation in Y is explained by variation in X



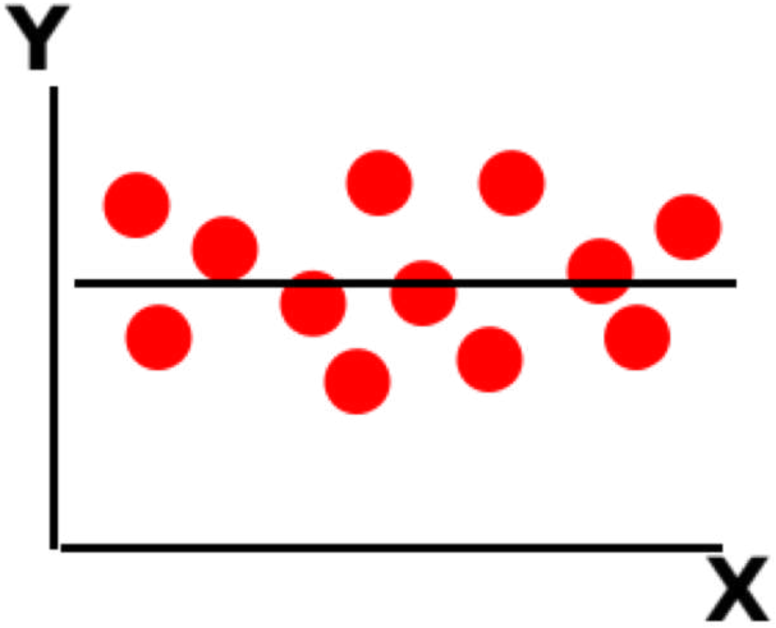
Examples of Approximate r2 Values

- **0 < r2 < 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 r2 Values

- **r2 = 0**
- No linear relationship between X and Y:
- The value of Y does not depend on X. (None of the variation in Y is explained by variation in X)



TELKOM DIGITAL TALENT INCUBATOR 2020

r2 in Python

```
model = sm.OLS(Y, X).fit()
predictions = model.predict(X)

print_model = model.summary()
print(print_model)
```

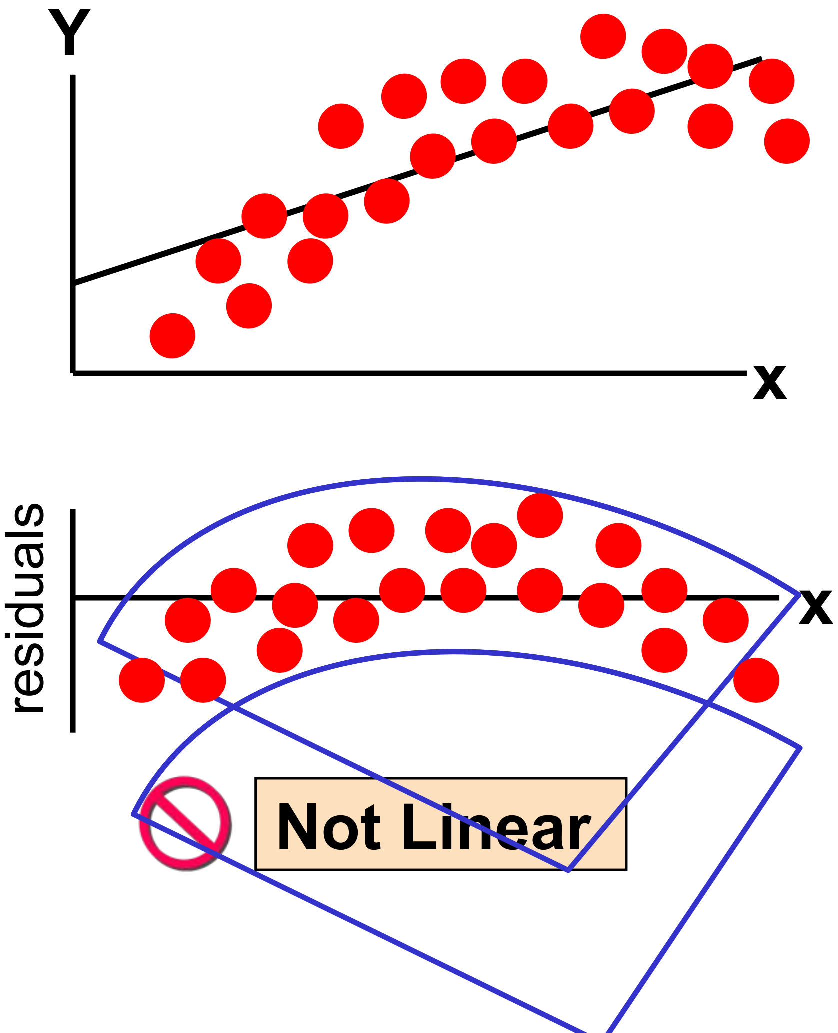
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            house_price   R-squared:                       0.581
Model:                            OLS   Adj. R-squared:                  0.528
Method:                 Least Squares   F-statistic:                     11.08
Date:                Wed, 14 Oct 2020   Prob (F-statistic):             0.0104
Time:                        16:28:58   Log-Likelihood:                -50.290
No. Observations:                  10   AIC:                             104.6
Df Residuals:                       8   BIC:                             105.2
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const          98.2483     58.033      1.693      0.129     -35.577     232.074
square_feet     0.1098      0.033      3.329      0.010       0.034       0.186
==============================================================================
Omnibus:                        1.066   Durbin-Watson:                   3.222
Prob(Omnibus):                  0.587   Jarque-Bera (JB):                0.779
Skew:                           0.399   Prob(JB):                        0.677
Kurtosis:                       1.890   Cond. No.                     7.82e+03
==============================================================================
```

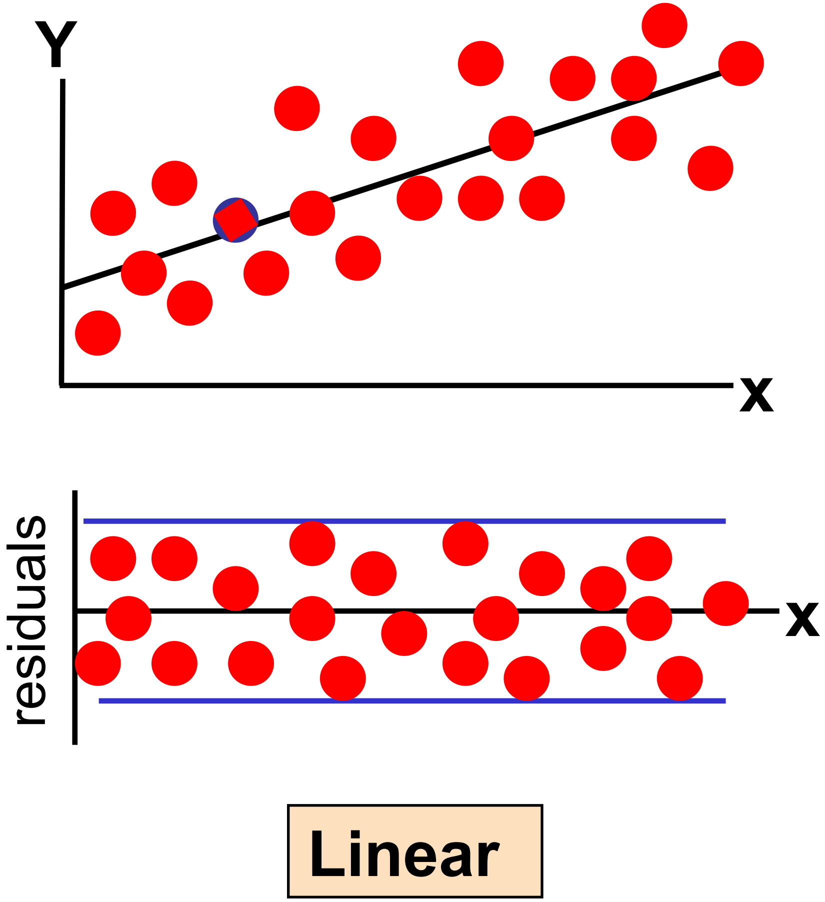
Organized by:
CAE

Residual Analysis

- The residual for observation i, ei, is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Evaluate independence assumption
  - Evaluate normal distribution assumption
  - Examine for constant variance for all levels of X (homoscedasticity)

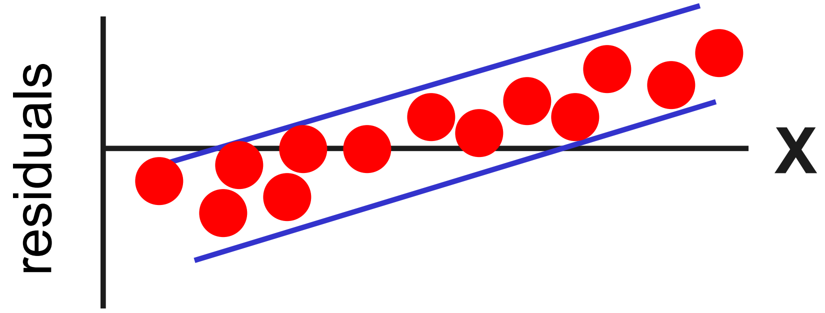
Residual Analysis for Linearity

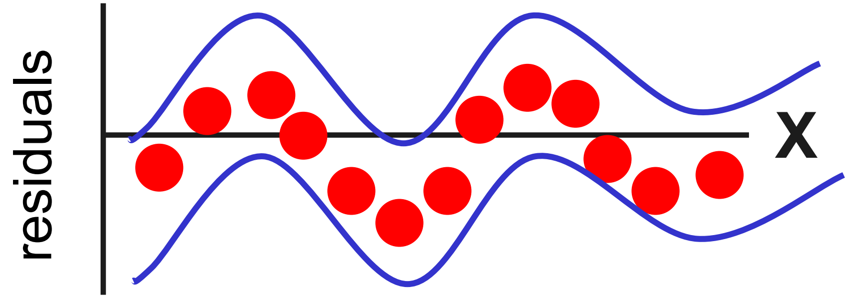




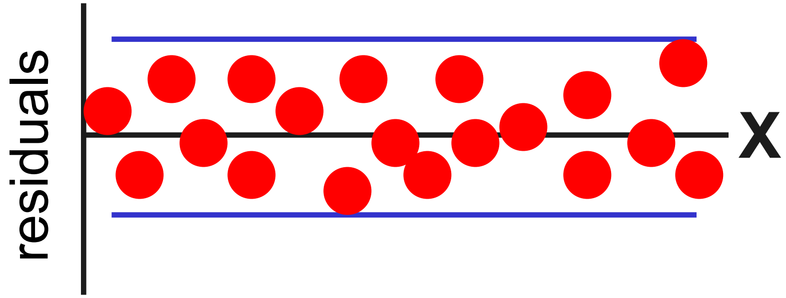
Residual Analysis for Independence

Not Independent





Independent

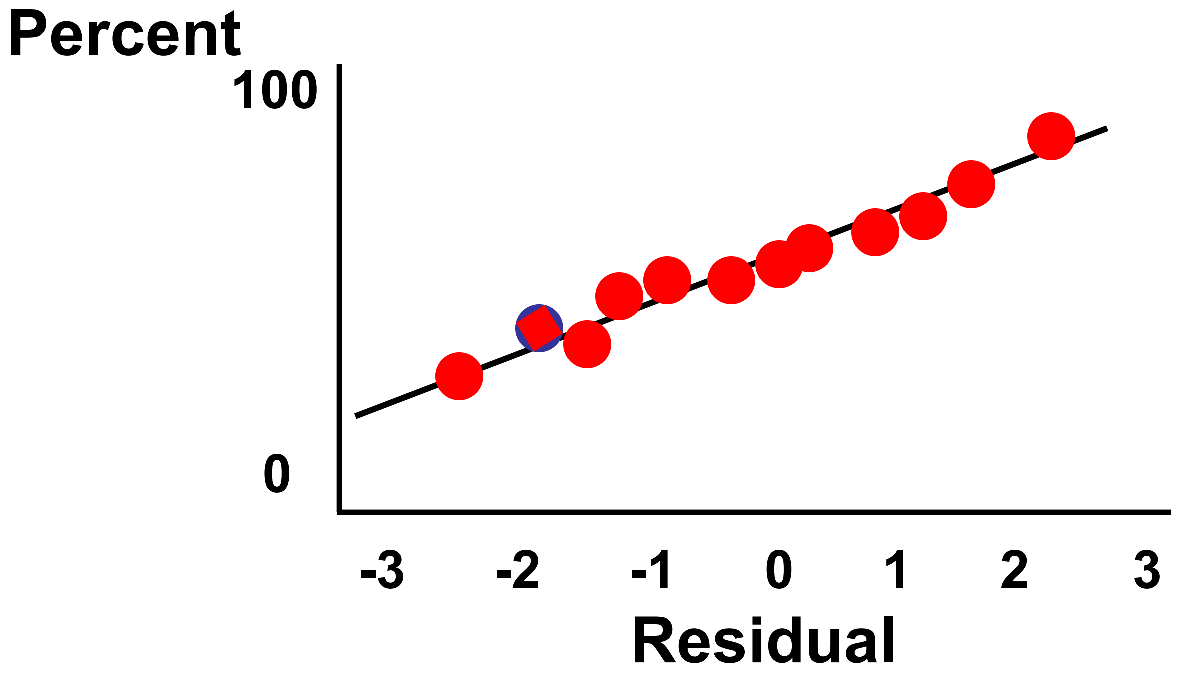


Checking for Normality

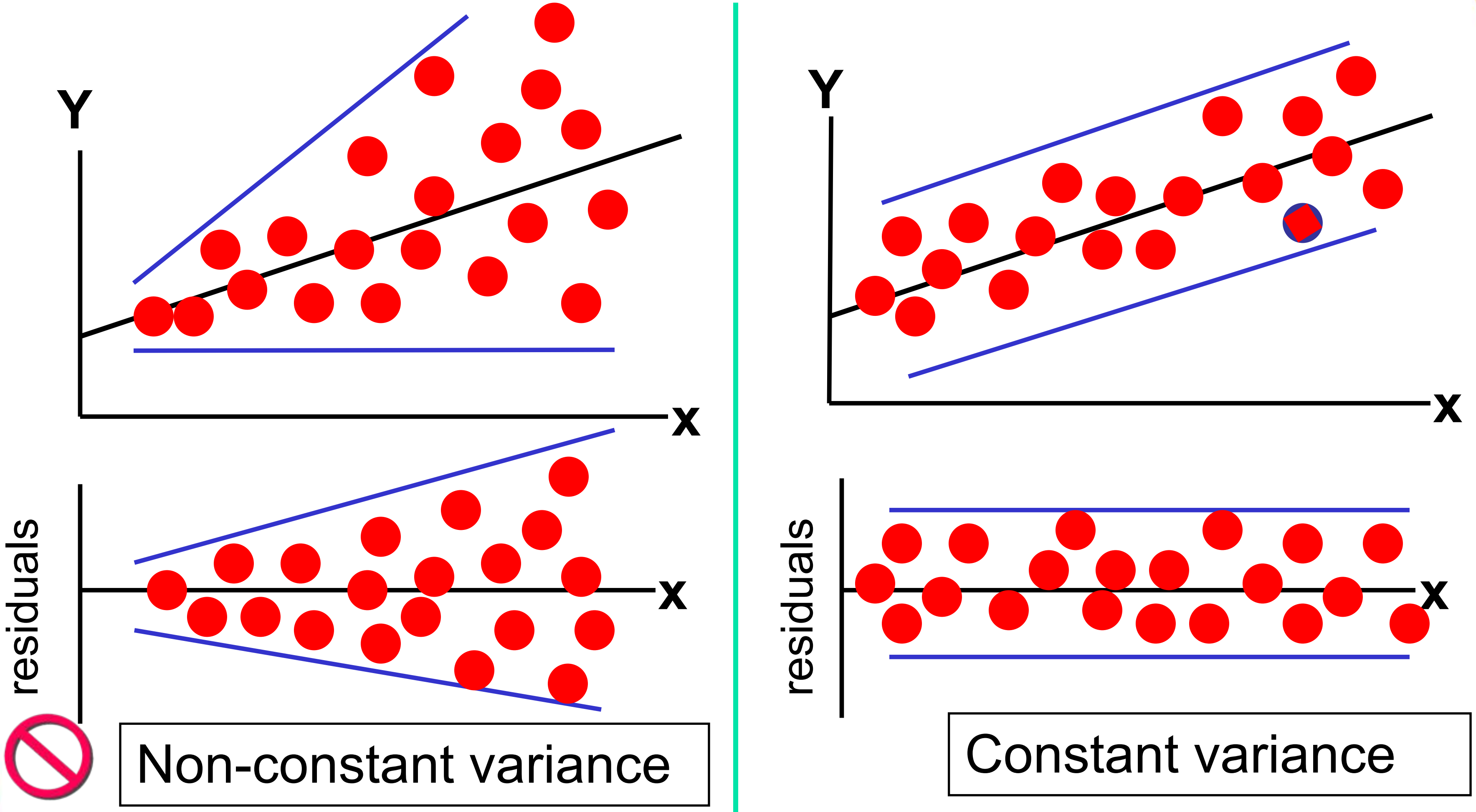
- Examine the Stem-and-Leaf Display of the Residuals
- Examine the Boxplot of the Residuals
- Examine the Histogram of the Residuals
- Construct a Normal Probability Plot of the Residuals

Residual Analysis for Normality

- When using a normal probability plot, normal errors will approximately display in a straight line.



Residual Analysis for Equal Variance

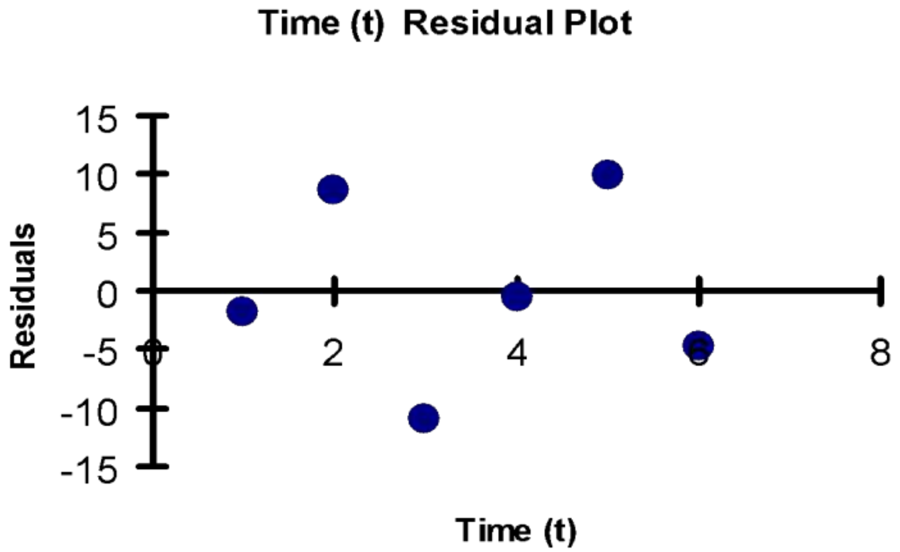


Measuring Autocorrelation

- Used when data are collected over time to detect if autocorrelation is present
- Autocorrelation exists if residuals in one time period are related to residuals in another period

Autocorrelation

- Autocorrelation is correlation of the errors (residuals) over time
- Here, residuals show a cyclic pattern, not random. Cyclical patterns are a sign of positive autocorrelation
- Violates the regression assumption that residuals are random and independent



The Durbin-Watson Statistic

- The Durbin-Watson statistic is used to test for autocorrelation
- H0: residuals are not correlated
- H1: positive autocorrelation is present

$$D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

- The possible range is $0 \le D \le 4$
- D should be close to 2 if H_0 is true
- D less than 2 may signal positive autocorrelation, D greater than 2 may signal negative autocorrelation

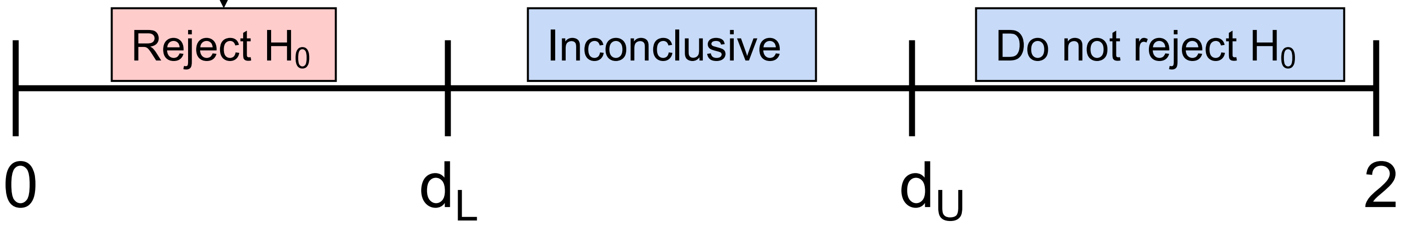
Testing for Positive Autocorrelation

H_0: positive autocorrelation does not exist

H_1: positive autocorrelation is present

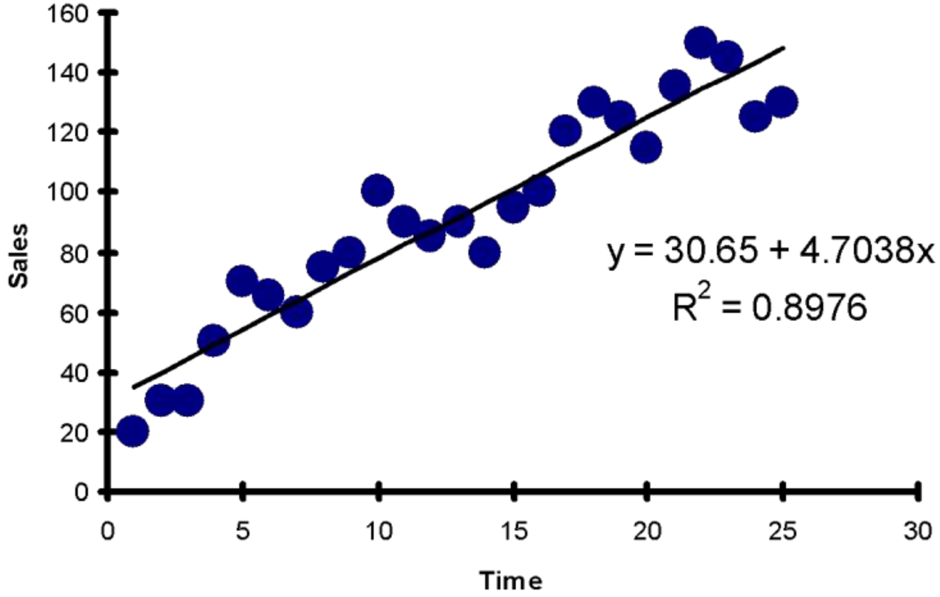
- Calculate the Durbin-Watson test statistic = D
  - (The Durbin-Watson Statistic can be found using Excel or Minitab)
- Find the values dL and dU from the Durbin-Watson table
  - (for sample size n and number of independent variables k)

Decision rule: reject H_0 if $D < d_L$



Testing for Positive Autocorrelation

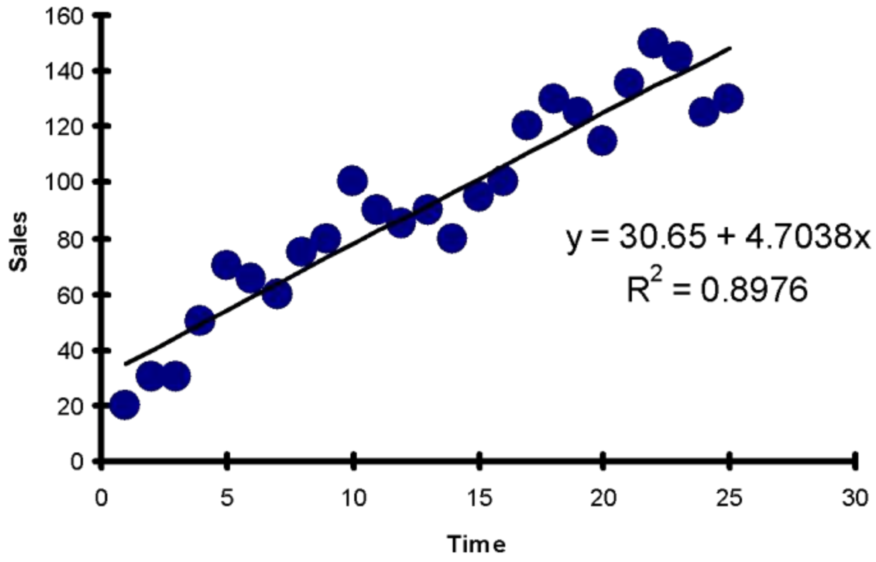
- Suppose we have the following time series data:
- Is there autocorrelation?



Testing for Positive Autocorrelation

- Example with n = 25:

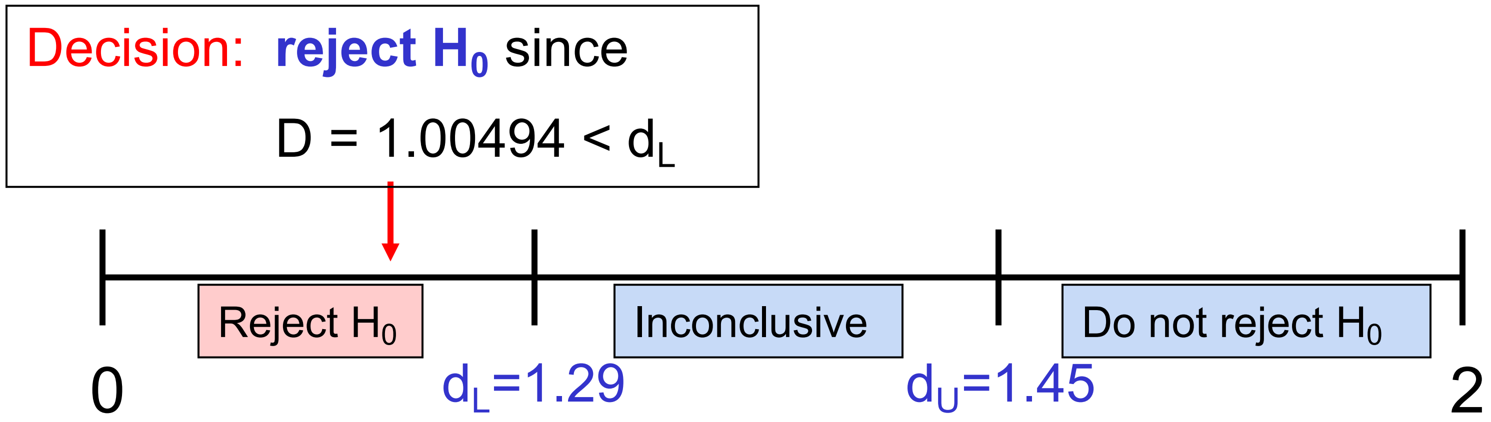
Durbin-Watson Calculations	
Sum of Squared Difference of Residuals	3296.18
Sum of Squared Residuals	3279.98
Durbin-Watson Statistic	**1.00494**



$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = \frac{3296.18}{3279.98} = 1.00494$$

Testing for Positive Autocorrelation

- Here, n = 25 and there is k = 1 one independent variable
- Using the Durbin-Watson table, dL = 1.29 and dU = 1.45
- D = 1.00494 < dL = 1.29, so reject H0 and conclude that significant positive autocorrelation exists



Inferences About the Slope

- The standard error of the regression slope coefficient (b1) is estimated by

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}} = \frac{S_{YX}}{\sqrt{\sum (X_i - \bar{X})^2}}$$

where:

S_{b_1} = Estimate of the standard error of the slope

$S_{YX} = \sqrt{\frac{SSE}{n-2}}$ = Standard error of the estimate

Inferences About the Slope: t Test

- t test for a population slope
  - Is there a linear relationship between X and Y?
- Null and alternative hypotheses
  - H0: β1 = 0 (no linear relationship)
  - H1: β1 ≠ 0 (linear relationship does exist)
- Test statistic

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}}$$

$$d.f. = n - 2$$

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

S_{b1} = standard error of the slope

TELKOM DIGITAL TALENT INCUBATOR 2020

Inferences About the Slope: t Test Example

	house_price	square_feet
0	245	1400
1	312	1600
2	279	1700
3	308	1875
4	199	1100
5	219	1550
6	405	2350
7	324	2450
8	319	1425
9	255	1700

Estimated Regression Equation:

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

The slope of this model is 0.1098

Is there a relationship between the square footage of the house and its sales price?

Organized by: CAE

t Test Example

- H0: β1 = 0
- H1: β1 ≠ 0

	coef	std err	t	P>|t|	[0.025	0.975]
const	98.2483	58.033	1.693	0.129	-35.577	232.074
square_feet	0.1098	0.033	3.329	0.010	0.034	0.186

b_1

S_{b_1}

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

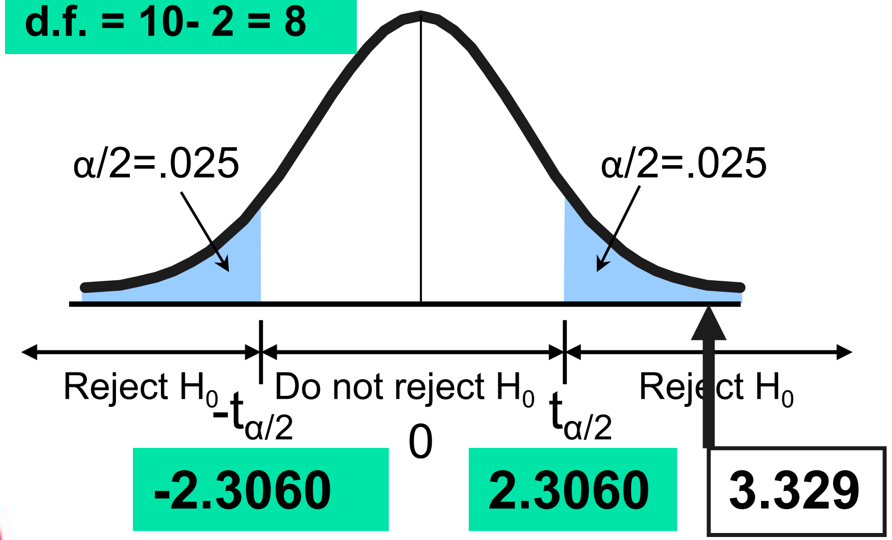
t Test Example

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

Test Statistic: $\mathbf{t_{STAT} = 3.329}$

d.f. = 10- 2 = 8



Decision: Reject H_0

There is sufficient evidence that square footage affects house price

Multiple Linear Regression

	coef	std err	t	P>|t|	[0.025	0.975]
const	98.2483	58.033	1.693	0.129	-35.577	232.074
square_feet	0.1098	0.033	3.329	0.010	0.034	0.186

p-value

Decision: Reject H_0, since p-value < α

There is sufficient evidence that square footage affects house price.

Multiple Linear Regression

- Idea: Examine the linear relationship between
- 1 dependent (Y) & 2 or more independent variables (Xi)

TELKOM DIGITAL TALENT INCUBATOR 2020

F Test for Significance

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            house_price   R-squared:                       0.581
Model:                            OLS   Adj. R-squared:                  0.528
Method:                 Least Squares   F-statistic:                     11.08
Date:                Wed, 14 Oct 2020   Prob (F-statistic):             0.0104
Time:                        16:28:58   Log-Likelihood:                -50.290
No. Observations:                  10   AIC:                             104.6
Df Residuals:                       8   BIC:                             105.2
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const         98.2483     58.033      1.693      0.129     -35.577     232.074
square_feet    0.1098      0.033      3.329      0.010       0.034       0.186
==============================================================================
Omnibus:                        1.066   Durbin-Watson:                   3.222
Prob(Omnibus):                  0.587   Jarque-Bera (JB):                0.779
Skew:                           0.399   Prob(JB):                        0.677
Kurtosis:                       1.890   Cond. No.                     7.82e+03
==============================================================================
```

$$F_{STAT} = \frac{MSR}{MSE} = \frac{18934.9348}{1708.1957} = 11.0848$$

With 1 and 8 degrees of freedom

p-value for the F-Test

F Test for Significance

$H_0: \beta_1 = 0$

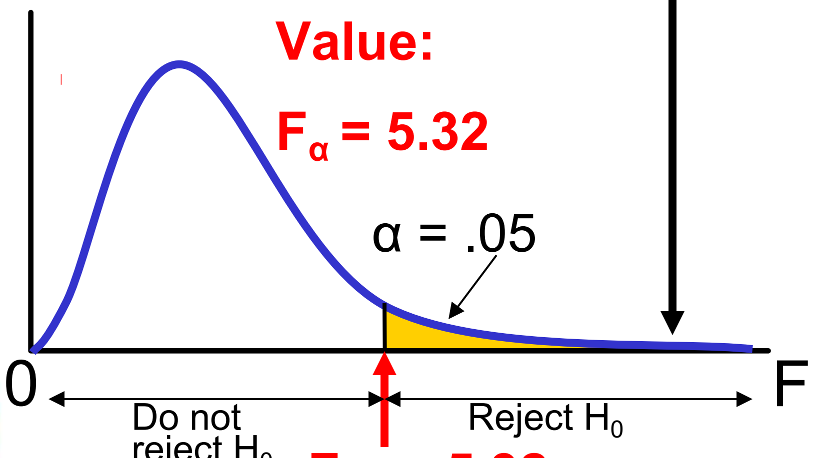
$H_1: \beta_1 \neq 0$

$\alpha = .05$

$\underbrace{df_1 = 1 \qquad df_2 = 8}$

Critical Value:

$F_\alpha = 5.32$



Test Statistic:

$$F_{STAT} = \frac{MSR}{MSE} = 11.08$$

Decision:

Reject H_0 at $\boldsymbol{\alpha} = 0.05$

Conclusion:

There is sufficient evidence that house size affects selling price

Organized by: CAE

Confidence Interval Estimate for the Slope

- Confidence Interval Estimate of the Slope:

$$b_1 \pm t_{\alpha/2} S_{b_1}$$

d.f. = n - 2

	coef	std err	t	P>|t|	[0.025	0.975]
const	98.2483	58.033	1.693	0.129	-35.577	232.074
square_feet	0.1098	0.033	3.329	0.010	0.034	0.186

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)

Confidence Interval Estimate for the Slope

	coef	std err	t	P>|t|	[0.025	0.975]
const	98.2483	58.033	1.693	0.129	-35.577	232.074
square_feet	0.1098	0.033	3.329	0.010	0.034	0.186

Since the units of the house price variable is \$1000s, we are 95% confident that the average impact on sales price is between \$33.74 and \$185.80 per square foot of house size

This 95% confidence interval does not include 0.

Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance

Organized by: CAE

t Test for a Correlation Coefficient

- Hypotheses
  - H0: ρ = 0 (no correlation between X and Y)
  - H1: ρ ≠ 0 (correlation exists)
- Test statistic
  - (with n – 2 degrees of freedom)

$$t_{STAT} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

where

$$r = +\sqrt{r^2} \quad \text{if } b_1 > 0$$

$$r = -\sqrt{r^2} \quad \text{if } b_1 < 0$$

t Test for a Correlation Coefficient

Is there evidence of a linear relationship between square feet and house price at the .05 level of significance?

H_0: ρ = 0 (No correlation)

H_1: ρ ≠ 0 (correlation exists)

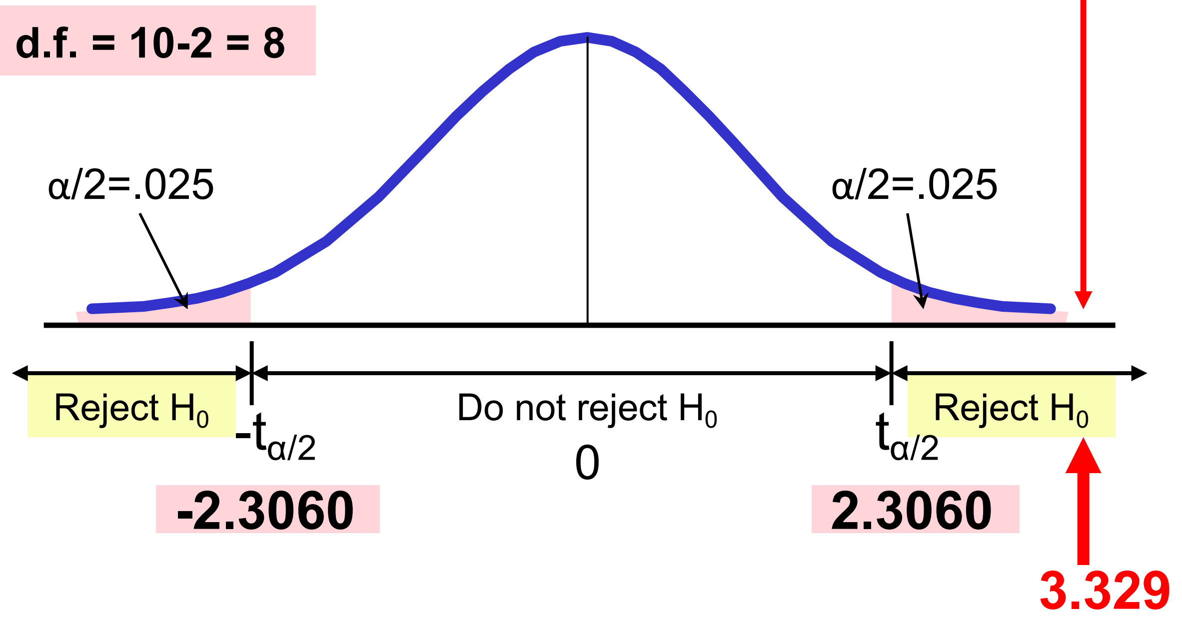
α =.05 , df = 10 - 2 = 8

$$t_{STAT} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{.762 - 0}{\sqrt{\frac{1 - .762^2}{10 - 2}}} = 3.329$$

TELKOM DIGITAL TALENT INCUBATOR 2020

t Test for a Correlation Coefficient

$$t_{STAT} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{.762 - 0}{\sqrt{\frac{1 - .762^2}{10 - 2}}} = 3.329$$



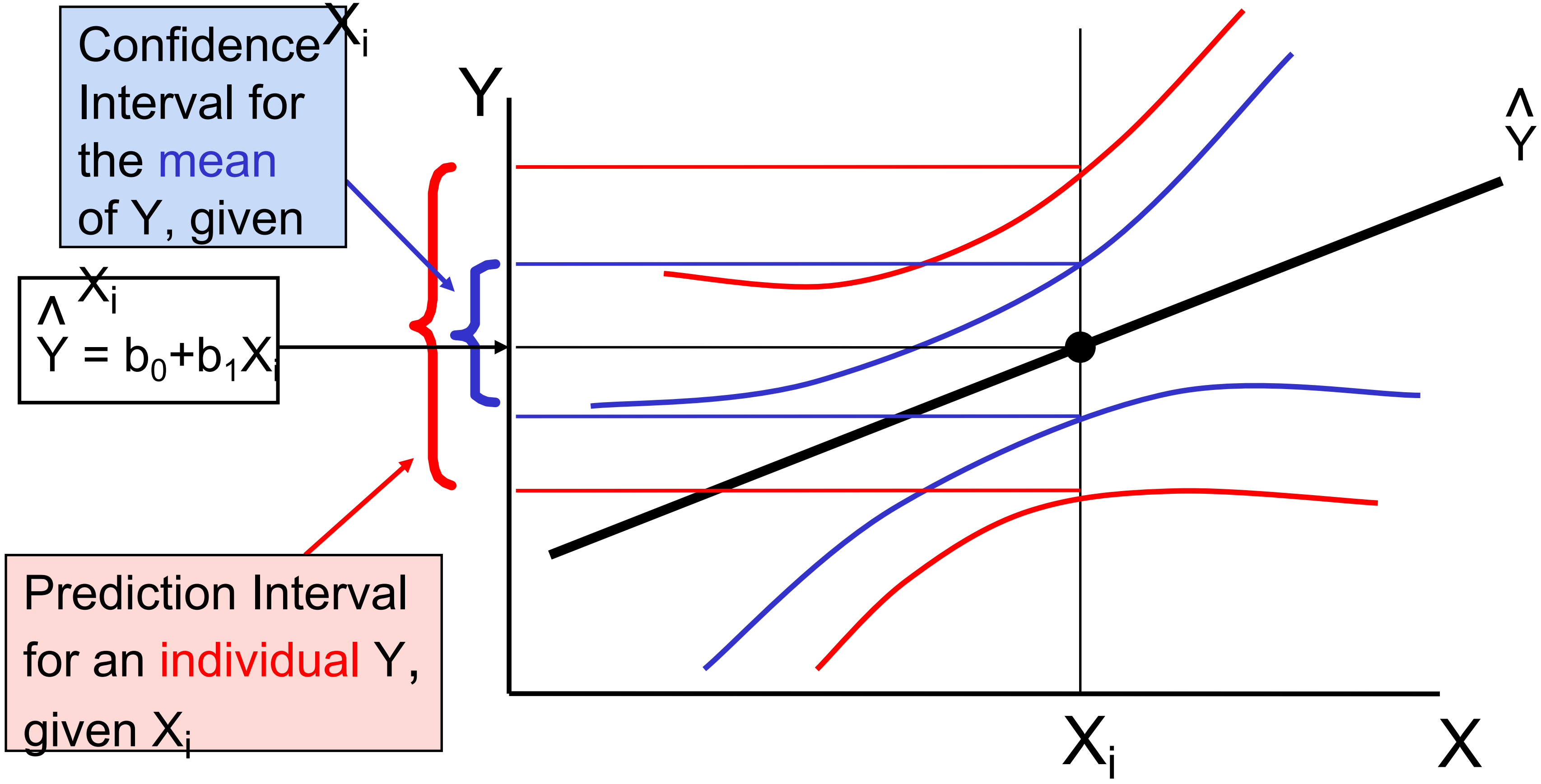
Decision:
Reject H_0

Conclusion:
There **is evidence** of a linear association at the 5% level of significance

Organized by: CAE

Estimating Mean Values and Predicting Individual Values

Goal: Form intervals around Y to express uncertainty about the value of Y for a given X_i



Confidence Interval for the Average Y, Given X

Confidence interval estimate for the **mean value of Y** given a particular X_i

Confidence interval for $\mu_{Y|X=X_i}$:

$$\hat{Y} \pm t_{\alpha/2} S_{YX} \sqrt{h_i}$$

Size of interval varies according to distance away from mean, $\bar{X}$

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}$$

Organized by: CAE

Prediction Interval for an Individual Y, Given X

Confidence interval estimate for an **Individual value of Y** given a particular X_i

Confidence interval for $Y_{X=X_i}$:

$$\hat{Y} \pm t_{\alpha/2} S_{YX} \sqrt{1 + h_i}$$

This extra term adds to the interval width to reflect the added uncertainty for an individual case

Estimation of Mean Values: Example

Confidence Interval Estimate for $\mu_{Y|X=X}$

Find the 95% confidence interval for the mean price of 2,000 square-foot houses

Predicted Price Y_i = 317.85 ($1,000s)

$$\hat{Y} \pm t_{0.025} S_{YX} \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}} = 317.85 \pm 37.12$$

The confidence interval endpoints are 280.66 and 354.90, or from $280,660 to $354,900

Estimation of Individual Values: Example

Prediction Interval Estimate for $Y_{X=X}$

Find the 95% prediction interval for an individual house with 2,000 square feet

Predicted Price $\hat{Y}_i$ = 317.85 ($1,000s)

$$\hat{Y} \pm t_{0.025} S_{YX} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}} = 317.85 \pm 102.28$$

The prediction interval endpoints are 215.50 and 420.07, or from $215,500 to $420,070

Multiple Linear Regression

- Idea: Examine the linear relationship between
- 1 dependent (Y) & 2 or more independent variables (Xi)

Multiple Linear Regression

Idea: Examine the linear relationship between
1 dependent (y) & 2 or more independent variables (x_i)

Population model:

Y-intercept
Population slopes
Random Error

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

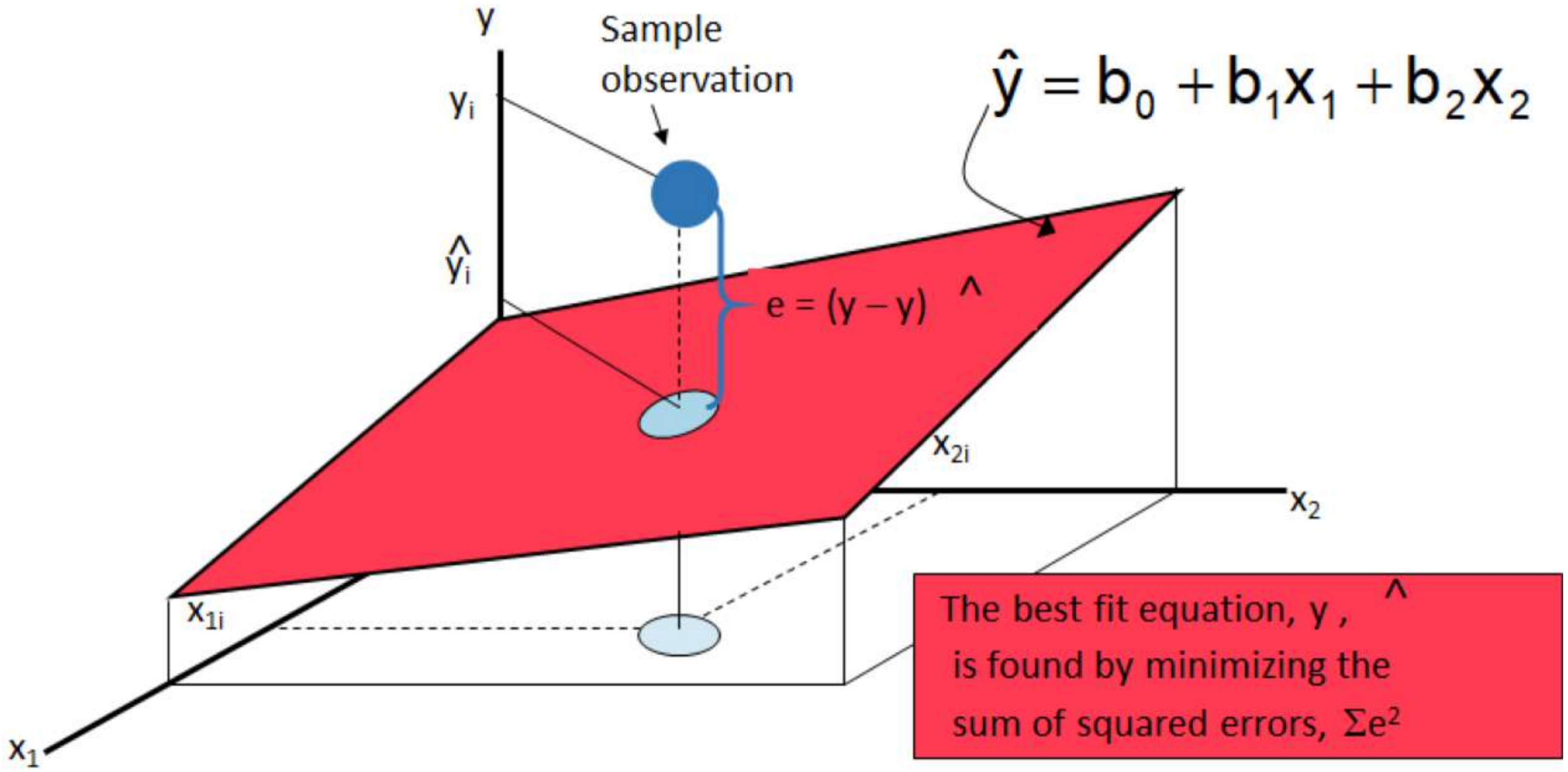
Estimated multiple regression model:

Estimated (or predicted) value of y
Estimated intercept
Estimated slope coefficients

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

Multiple Linear Regression

Two variable model



Example: Multiple Linear Regression

- A distributor of frozen dessert pies wants to evaluate factors thought to influence demand
- Dependent variable: Pie sales (units per week)
- Independent variables: Price (in $), Advertising ($100's)
- Data are collected for 15 weeks

Example: Data

	week	pie_sales	price	advertising
0	1	350	5.5	3.3
1	2	460	7.5	3.3
2	3	350	8.0	3.0
3	4	430	8.0	4.5
4	5	350	6.8	3.0
5	6	380	7.5	4.0
6	7	430	4.5	3.0
7	8	470	6.4	3.7
8	9	450	7.0	3.5
9	10	490	5.0	4.0
10	11	340	7.2	3.5
11	12	300	7.9	3.2
12	13	440	5.9	4.0
13	14	450	5.0	3.5
14	15	300	7.0	2.7

$$\widehat{\text{Sales}} = b_0 + b_1\,(\text{Price}) + b_2\,(\text{Advertising})$$

Example: Regression Equation

$$\text{Sales} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

where
Sales is in number of pies per week
Price is in $
Advertising is in $100’s.

b_1 = -24.975: sales will decrease, on average, by 24.975 pies per week for each $1 increase in selling price, net of the effects of changes due to advertising

b_2 = 74.131: sales will increase, on average, by 74.131 pies per week for each $100 increase in advertising, net of the effects of changes due to price

Example: Making Predictions

Predict sales for a week in which the selling price is \$5.50 and advertising is \$350:

$$\text{Sales} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

$$= 306.526 - 24.975\,(5.50) + 74.131\,(3.5)$$

$$= 428.62$$

Predicted sales is 428.62 pies

Note that Advertising is in \$100's, so \$350 means that X_2 = 3.5

Adjusted r2

- r2 never decreases when a new X variable is added to the model
  - This can be a disadvantage when comparing models
- What is the net effect of adding a new variable?
  - We lose a degree of freedom when a new X variable is added
  - Did the new X variable add enough explanatory power to offset the loss of one degree of freedom?

Adjusted r2

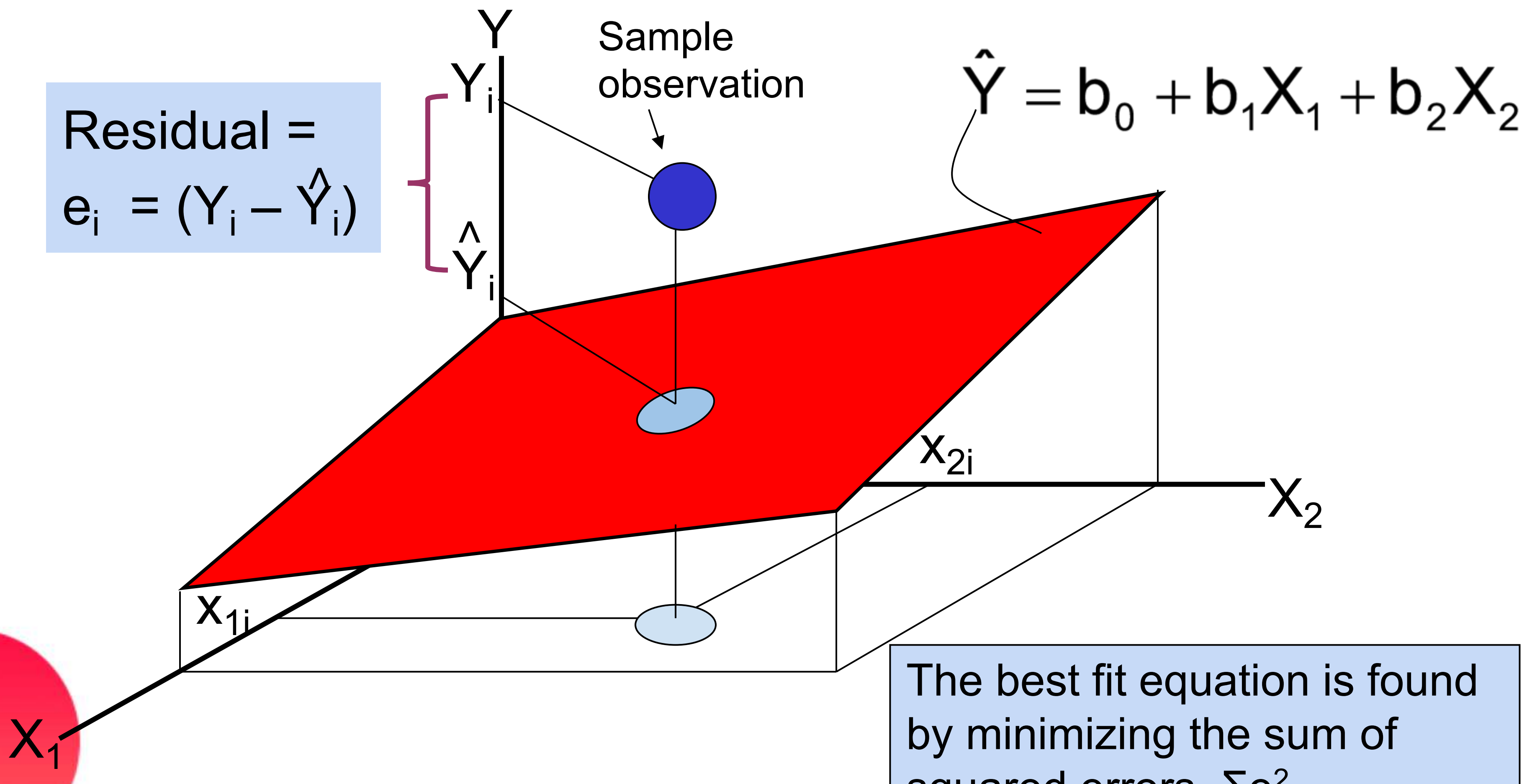
- Shows the proportion of variation in Y explained by all X variables adjusted for the number of X variables used

$$r_{adj}^2 = 1 - \left[(1 - r^2) \left(\frac{n-1}{n-k-1} \right) \right]$$

(where n = sample size, k = number of independent variables)

  - Penalize excessive use of unimportant independent variables
  - Smaller than r2
  - Useful in comparing among models

Residuals in Multiple Regression



The best fit equation is found by minimizing the sum of squared errors, Σe^2

Organized by: CAE

Pitfalls of Regression Analysis

- Lacking an awareness of the assumptions underlying least-squares regression
- Not knowing how to evaluate the assumptions
- Not knowing the alternatives to least-squares regression if a particular assumption is violated
- Using a regression model without knowledge of the subject matter
- Extrapolating outside the relevant range

Strategies for Avoiding the Pitfalls of Regression

- Start with a scatter plot of X vs. Y to observe possible relationship
- Perform residual analysis to check the assumptions
- Plot the residuals vs. X to check for violations of assumptions such as homoscedasticity
- Use a histogram, stem-and-leaf display, boxplot, or normal probability plot of the residuals to uncover possible non-normality

Strategies for Avoiding the Pitfalls of Regression

- If there is violation of any assumption, use alternative methods or models
- If there is no evidence of assumption violation, then test for the significance of the regression coefficients and construct confidence intervals and prediction intervals
- Avoid making predictions or forecasts outside the relevant range

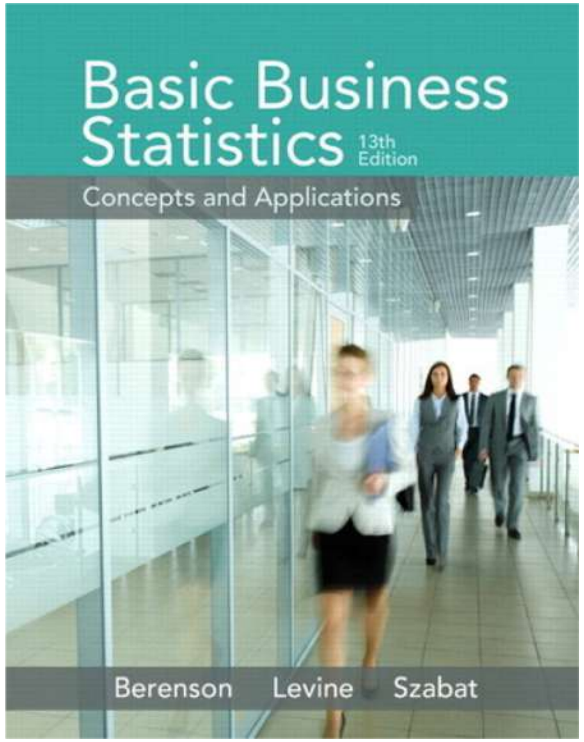
Practice with Python

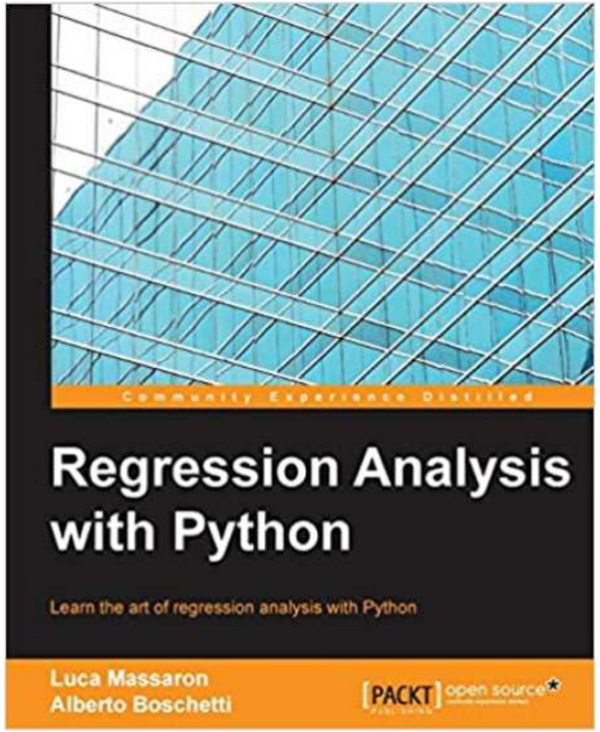
- Practice Link: https://github.com/rc-dbe/dti

Organized by:

References/Additional Resources

- Basic Business Statistics 13th Edition by Mark Berenson
- Regression Analysis with Python by Luca Massaron





TELKOM DIGITAL TALENT INCUBATOR 2020

Assignment Week 4

- Create a multiple linear regression model using pie sales data in https://github.com/rc-dbe/dti
- Use Google Collab (or Jupyter Notebook if you want)
- Put the code in your github
- Make it informative as possible

Organized by: CAE

TELKOM
DIGITAL TALENT
INCUBATOR 2020

Assignment Week 4

Organized by:
CAE