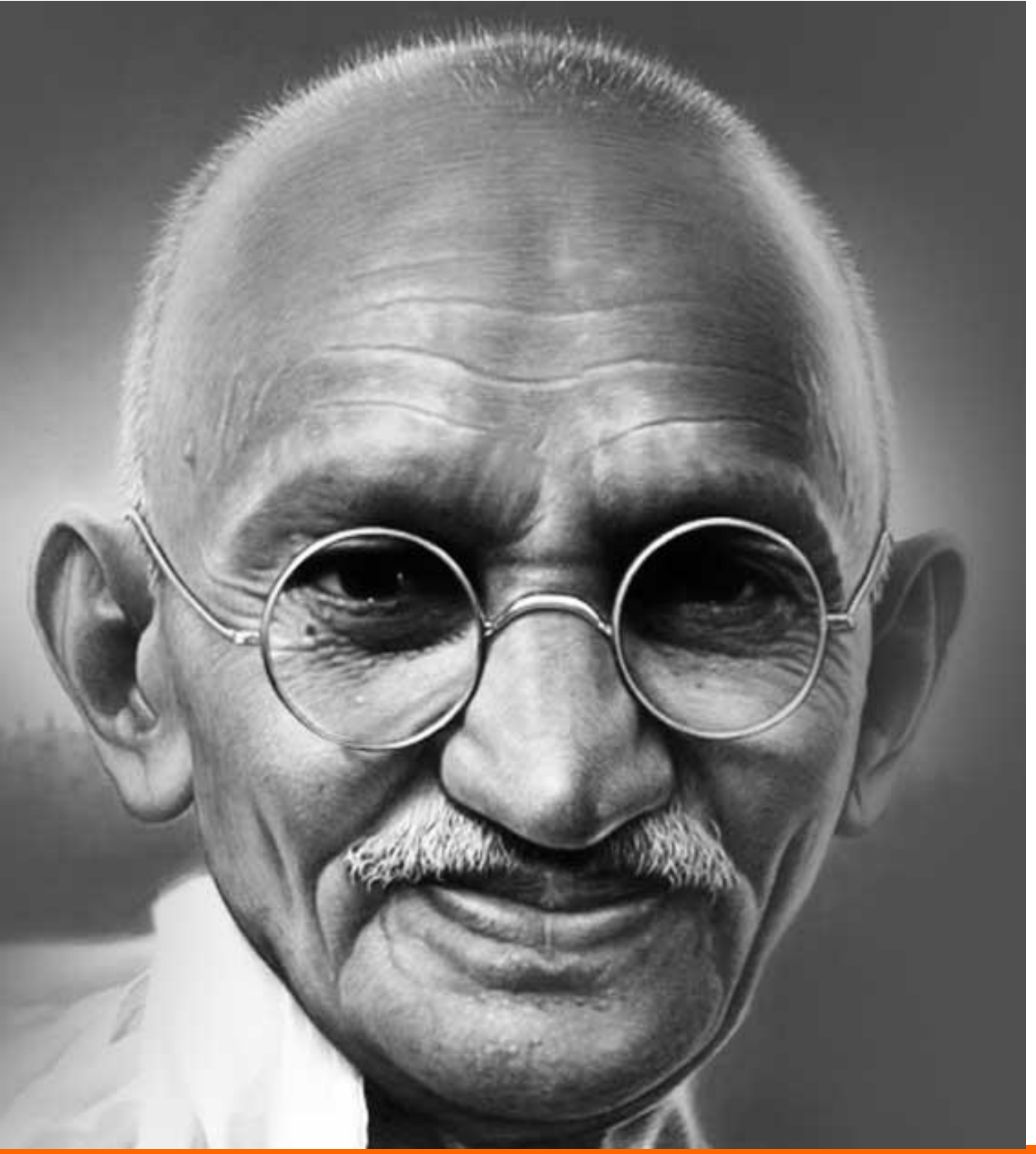
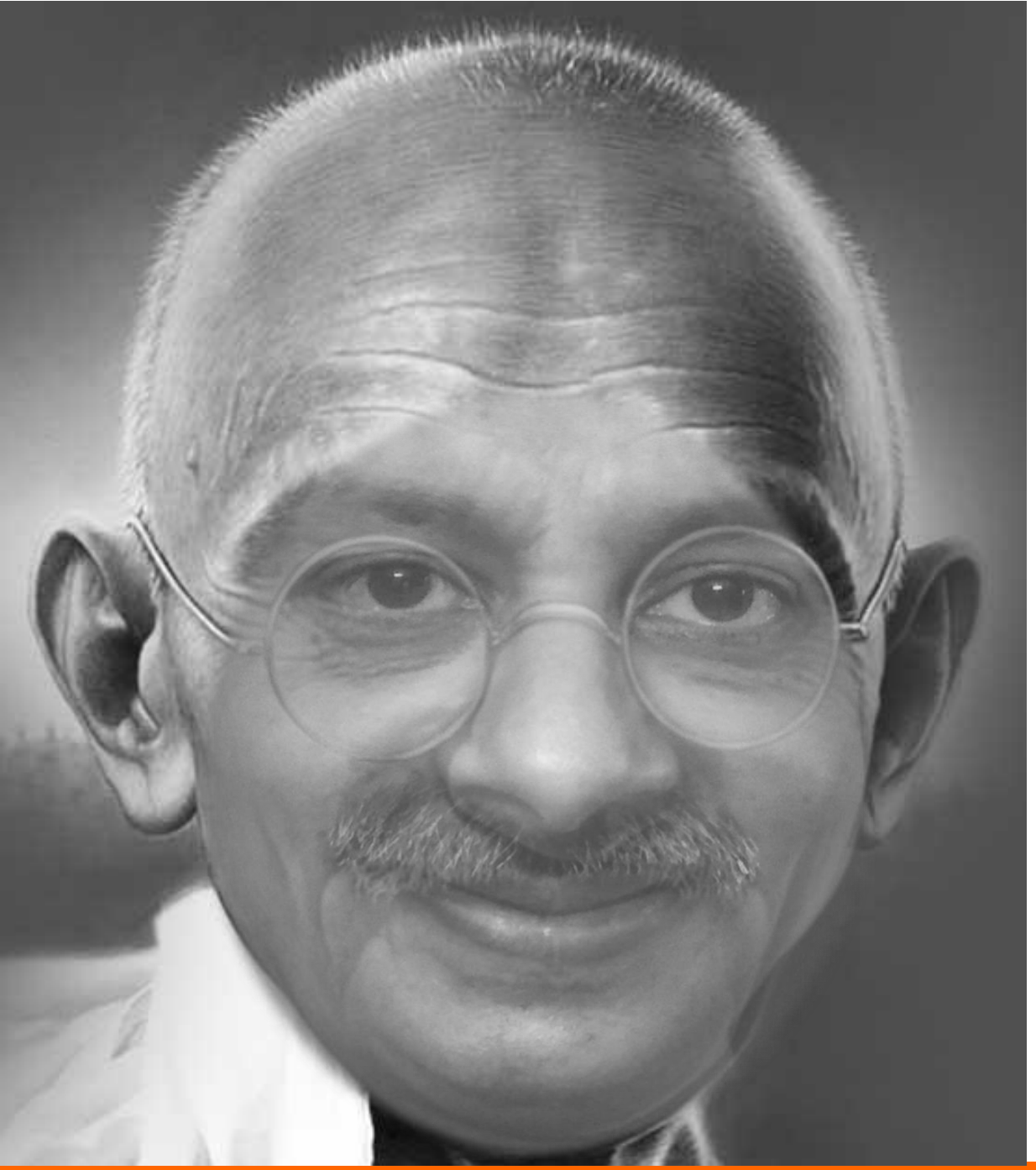


Live as if you were to die tomorrow.

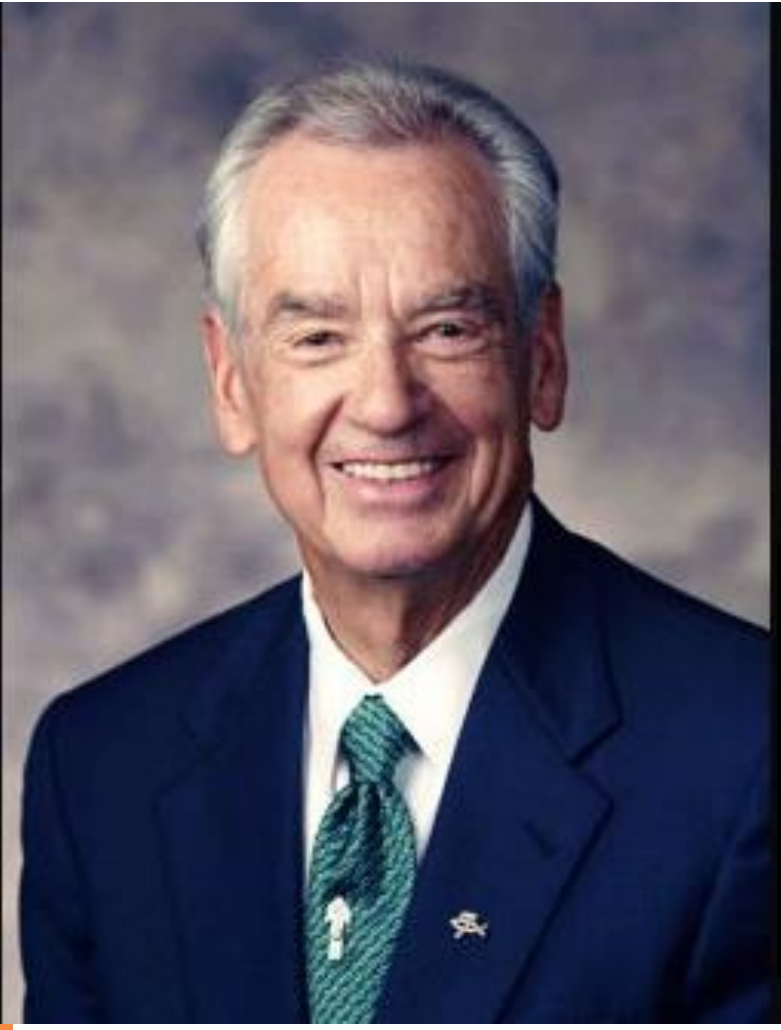
Learn as if you were to live forever.

CSE 7302c



There is nothing that wastes the body like worry, and one who has any faith in INSOFE should be ashamed to worry about anything whatsoever.

- Sridhar Pappu



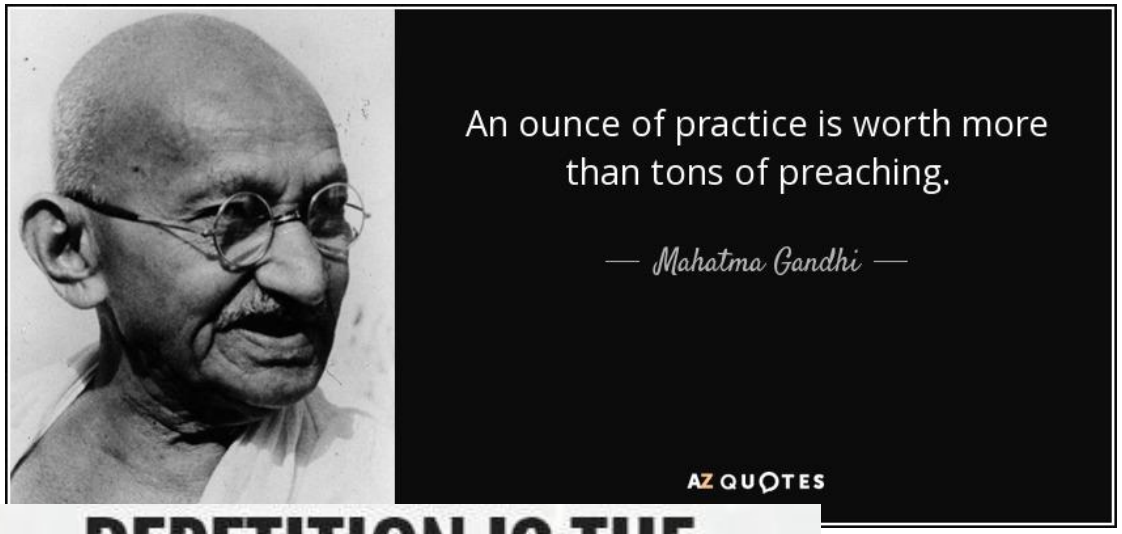
Positive thinking will let you do everything better than negative thinking will.

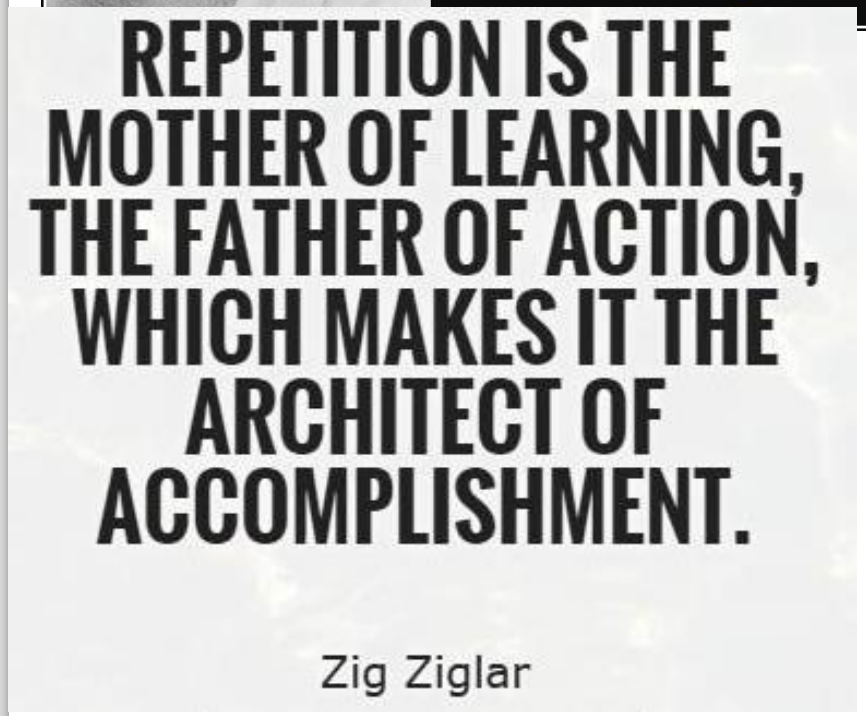
Your attitude, not your aptitude, will determine your altitude.

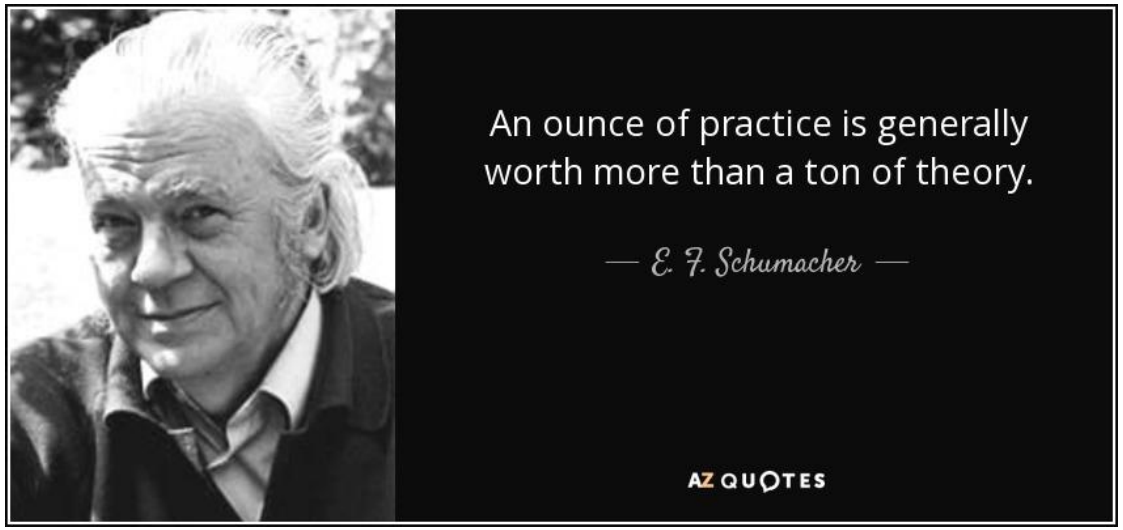
F-E-A-R has two meanings: “Forget Everything And Run” or “Face Everything And Rise”. The choice is yours.

- Zig Ziglar

E 73026


An ounce of practice is worth more than tons of preaching.
— Mahatma Gandhi —
AZ QUOTES


REPETITION IS THE MOTHER OF LEARNING, THE FATHER OF ACTION, WHICH MAKES IT THE ARCHITECT OF ACCOMPLISHMENT.
Zig Ziglar


An ounce of practice is generally worth more than a ton of theory.
— E. F. Schumacher —
AZ QUOTES

Regarding access to videos

Hi Sir,

I am not able to access the recorded lectures from more than a week. I am currently trying for data science opportunities and videos are very crucial for my preparation. Hence, I request you to extend my access till I get into data science role. Every time I see the videos, my understanding is becoming better and clear. Please do the needful.

Thanks in advance...

Inspire…Educate…Transform.

Statistics and Probability in Decision Modeling

Linear Regression

Dr. Sridhar Pappu
Executive VP – Academics, INSOFE

October 06, 2018

Analyzing relationships between attributes

CORRELATION, COVARIANCE AND R-SQUARED

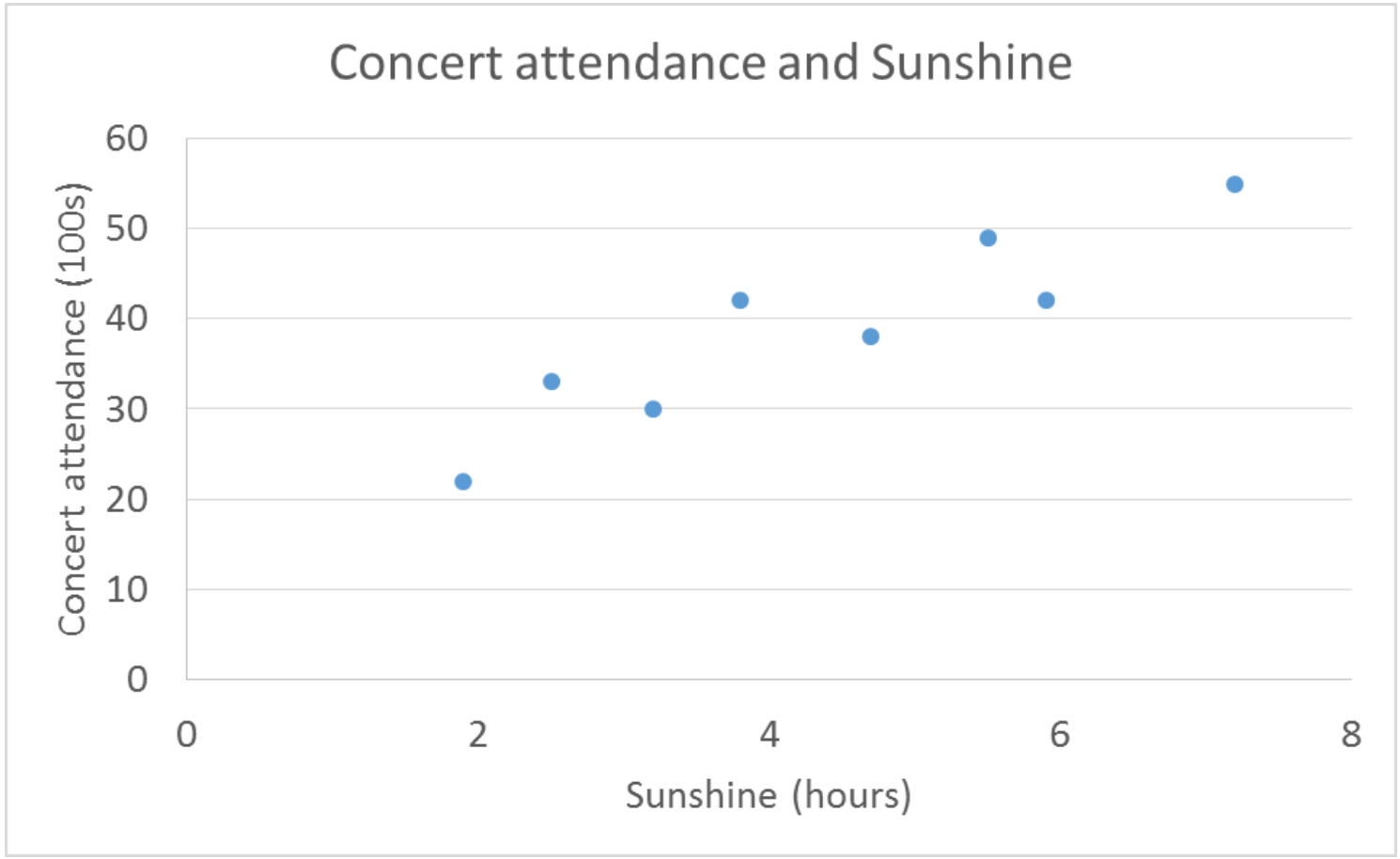
Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

- The band makes a loss if less than 3500 people attend.
- Based on predicted hours of sunshine, can we predict ticket sales?
- Are sunshine and concert attendance correlated?

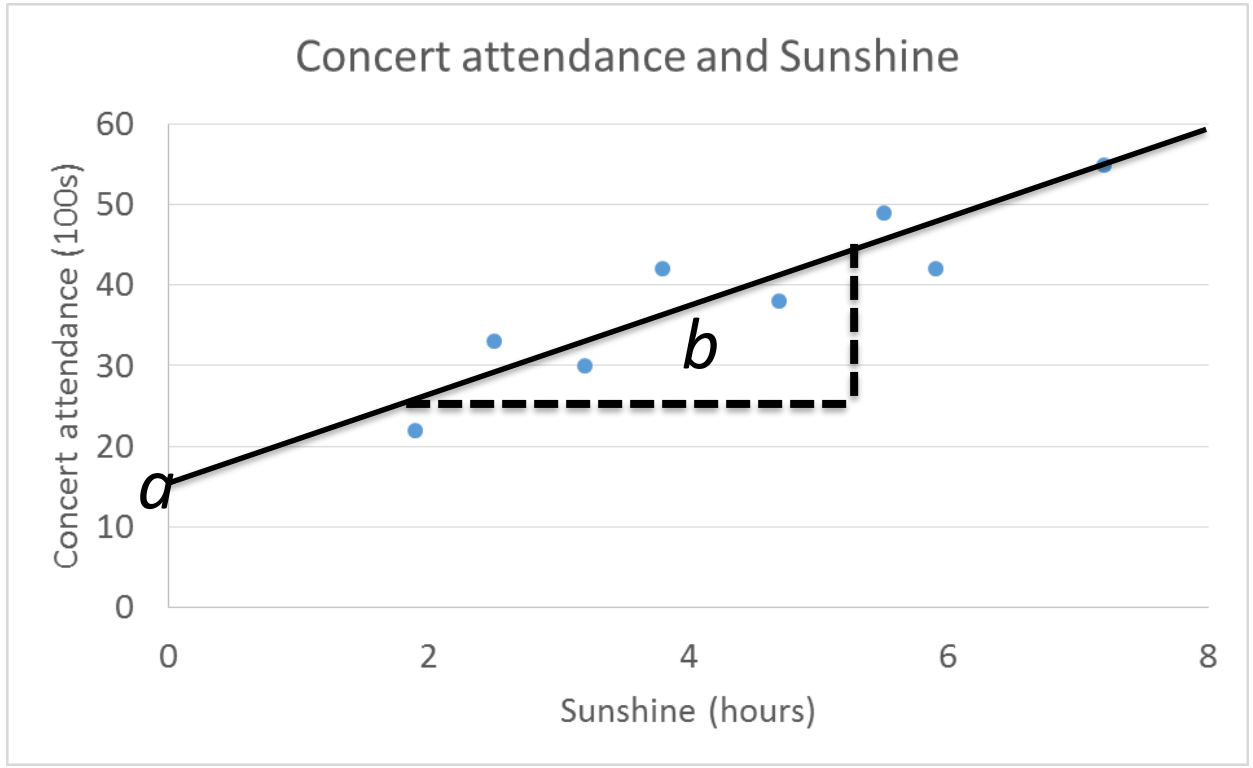
Image Source: http://blurtonline.com/wp-content/uploads/2013/06/Shaky-Knees-1514.jpeg;
Last accessed: May 1, 2014

Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

- Independent variable (explanatory) – Sunshine – Plotted on X-axis
- Dependent variable (response) – Concert attendance – Plotted on Y-axis



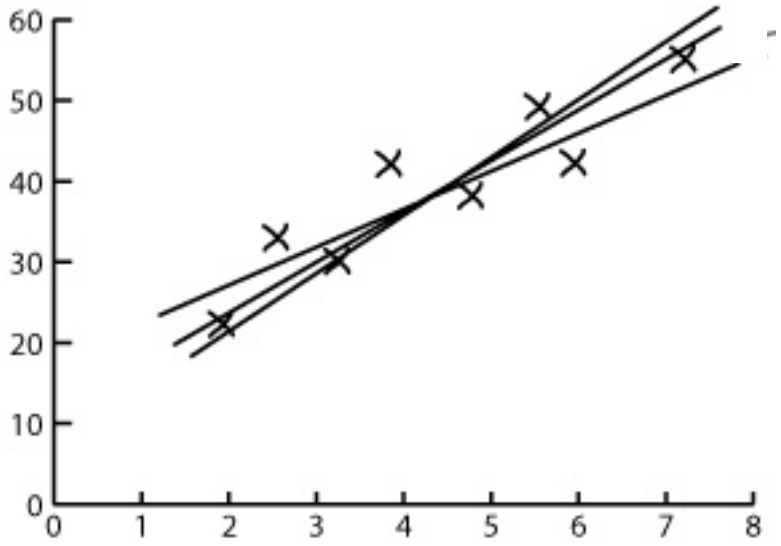
We need to find the equation of the line.

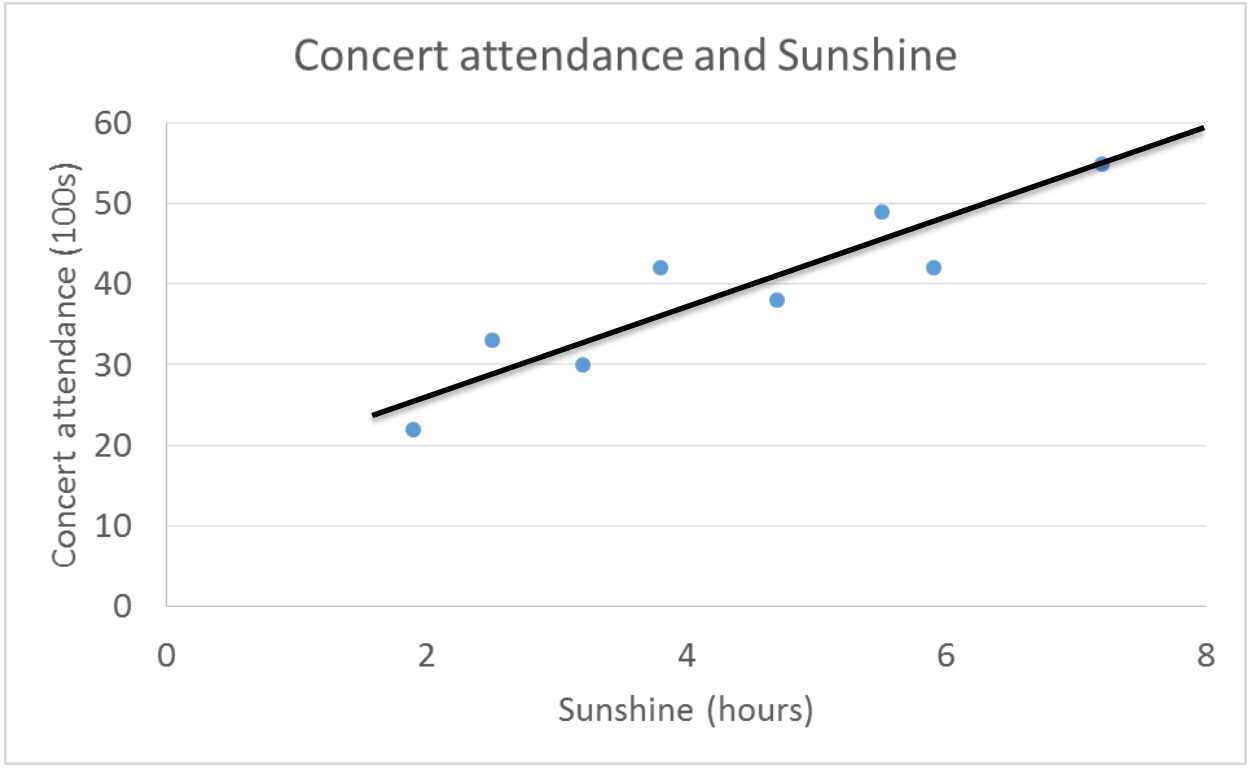


$y = a + bx$

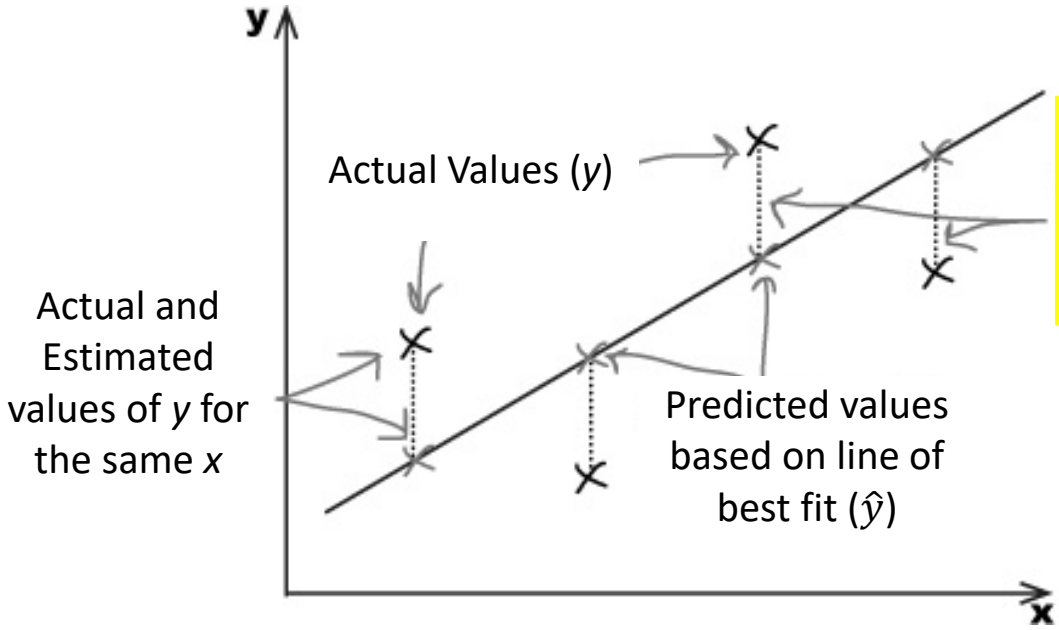
Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

- Line of best fit





We need to minimize errors.



$$\hat{y} = a + bx$$

Line of best fit is the line that minimizes all the distances (residuals) between the Actual (y) and Estimated ($\hat{y}$) values

We could do that by minimizing $\sum(y_i - \hat{y}_i)$, where $y_i$is the actual value and $\hat{y}_i$ its estimate. $(y_i - \hat{y}_i)$ is also known as the **residual**.

But $\sum(y_i - \hat{y}_i) = 0$.

Just as we did when finding variance, we find the **sum of squared errors** or SSE.

$$SSE = \sum (y_i - \hat{y}_i)^2$$

The value of *b*, the slope, that minimizes the SSE is given by

$$b = \frac{\sum((x - \bar{x})(y - \bar{y}))}{\sum(x - \bar{x})^2}$$

Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

The value of b, the slope, that minimizes the SSE is given by $b = \frac{\sum((x-\bar{x})(y-\bar{y}))}{\sum(x-\bar{x})^2}$

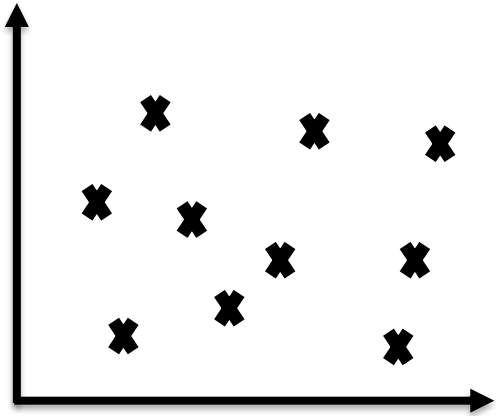
How do you calculate a in $\hat{y}_i = a + bx$?

The line of best fit must pass through $(\bar{x}, \bar{y})$. Substituting in the equation $\bar{y} = a + b\bar{x}$, we can find a.

This method of fitting the line of best fit is called **Least Squares Regression** or **Ordinary Least Squares Regression** or **OLS Regression**.

But how do you know how accurate this line is?

Accurate Linear Correlation

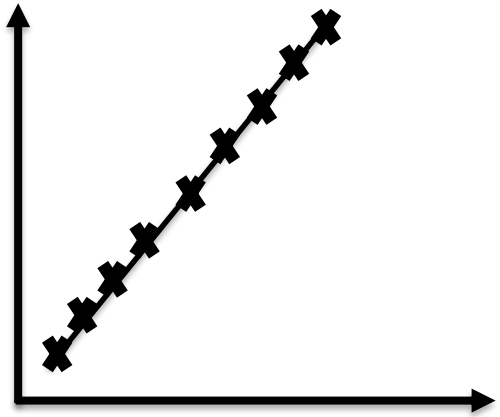


No Linear Correlation

The fit of the line is given by **correlation coefficient**.

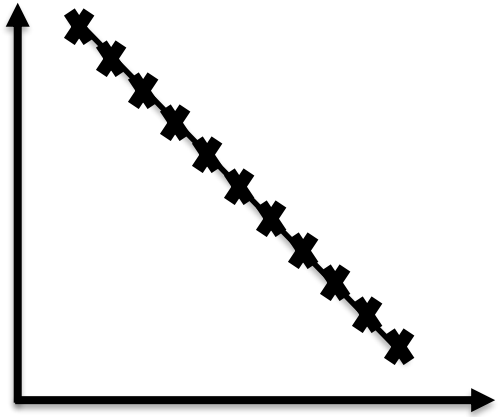
Correlation Coefficient

Correlation coefficient, *r*, is a number between -1 and 1 and tells us how well a regression line fits the data.



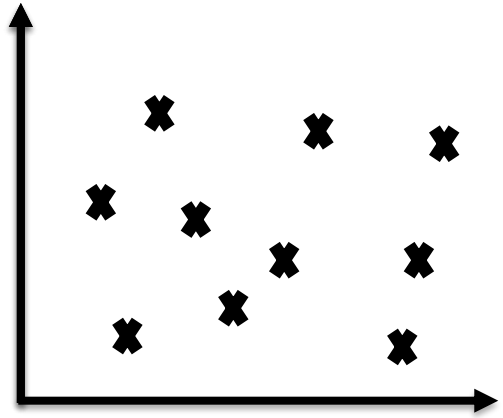
r = 1

Positive Linear Correlation



r = -1

Negative Linear Correlation



r =0

No Correlation

It gives the **strength** and **direction** of the relationship between two variables.

Correlation Coefficient

$r = \frac{bs_x}{s_y}$ where *b* is the slope of the line of best fit, s_x is the standard deviation of the *x* values in the sample, and s_y is the standard deviation of the *y* values in the sample.

$s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ and $s_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n-1}}$.

Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

Find *r* for this data. r = 0.916

Correlation Coefficient and Covariance – Excel*

$s_x{}^2 = \frac{\sum(x-\bar{x})^2}{n-1}$, $s_y{}^2 = \frac{\sum(y-\bar{y})^2}{n-1}$, $s_{xy}^2 = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$, where $s_x{}^2$ is the sample variance of the *x* values, $s_y{}^2$ is the sample variance of the *y* values and s_{xy}^2 is the covariance.

$b = \frac{s_{xy}^2}{s_x{}^2}$ and so, $r = \frac{s_{xy}^2}{s_x s_y}$ *(Recall* $b = \frac{\sum((x-\bar{x})(y-\bar{y}))}{\sum(x-\bar{x})^2}$ *and* $r = \frac{bs_x}{s_y}$*)*.

So, correlation coefficient is simply *standardized (or scaled) covariance*. And covariance of *standardized variables (z-scores)* is the same as their correlation coefficient.

* Height and weight data generated randomly using Excel.
Oil prices from http://www.macrotrends.net/1369/crude-oil-price-history-chart
Potato prices from https://data.gov.in/catalog/dailyweekly-retail-prices-potato
Last accessed: October 28, 2017

Covariance and Correlation

$$s_{xy}^2 = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}, r = \frac{s_{xy}^2}{s_x s_y}$$

- If both x and y are large distance away from their respective means, the resulting covariance will be even larger.
  - The value will be positive if both are below the mean or both are above.
  - If one is above and the other below, the covariance will be negative.
- If even one of them is very close to the mean, the covariance will be small.
- Cov(x,x)=Var(x)

Covariance and Correlation

$$s_{xy}^2 = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}, r = \frac{s_{xy}^2}{s_x s_y}$$

- The value of covariance itself doesn't say much. It only shows whether the variables are moving together (positive value) or opposite to each other (negative value).
  - **Affected by scale (measuring height in ft vs mm)**
  - **Not intuitive comparing covariance values between 2 sets of variables (how does height-weight covariance compare with oil price($)-potato price (Rupee) covariance)**
  - **Unintuitive units**

Covariance and Correlation

$$s_{xy}^2 = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}, r = \frac{s_{xy}^2}{s_x s_y}$$

- To know the strength of how the variables move together, covariance is standardized to the dimensionless quantity, correlation.

Coefficient of Determination – R^2

The coefficient of determination is given by r^2 or R^2. **It is the percentage of variation in the *y* variable that is explainable by the *x* variable.**

For example, what percentage of the variation in open-air concert attendance is explainable by the number of hours of predicted sunshine.

If $r^2 = 0$, it means you can't predict the *y* value from the *x* value.

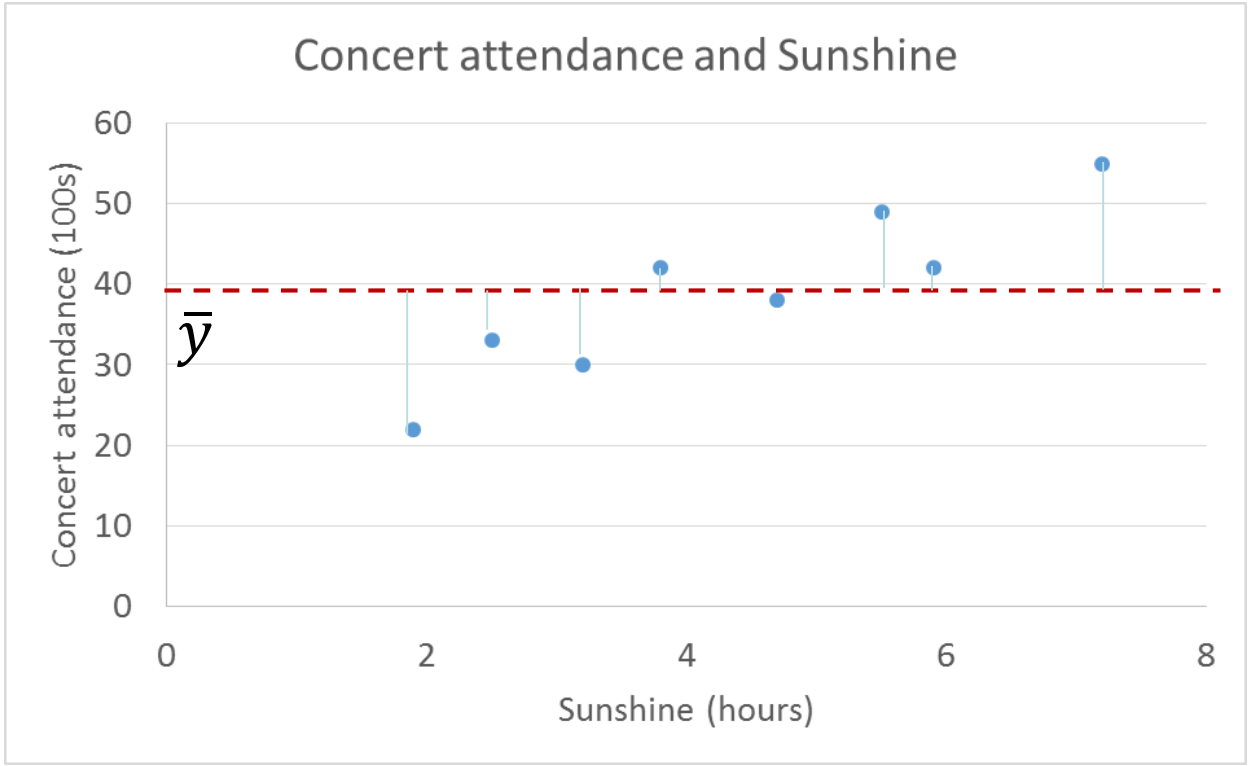
If $r^2 = 1$, it means you can predict the *y* value from the *x* value without any errors.

Usually, r^2 is between these two extremes.

R-squared

SST (Recall Sum of Squares Total from ANOVA) – This is the total variation in data. The horizontal line at $\bar{y}$ indicates expected concert attendance when sunshine is not considered. This "model" has **large** residuals.

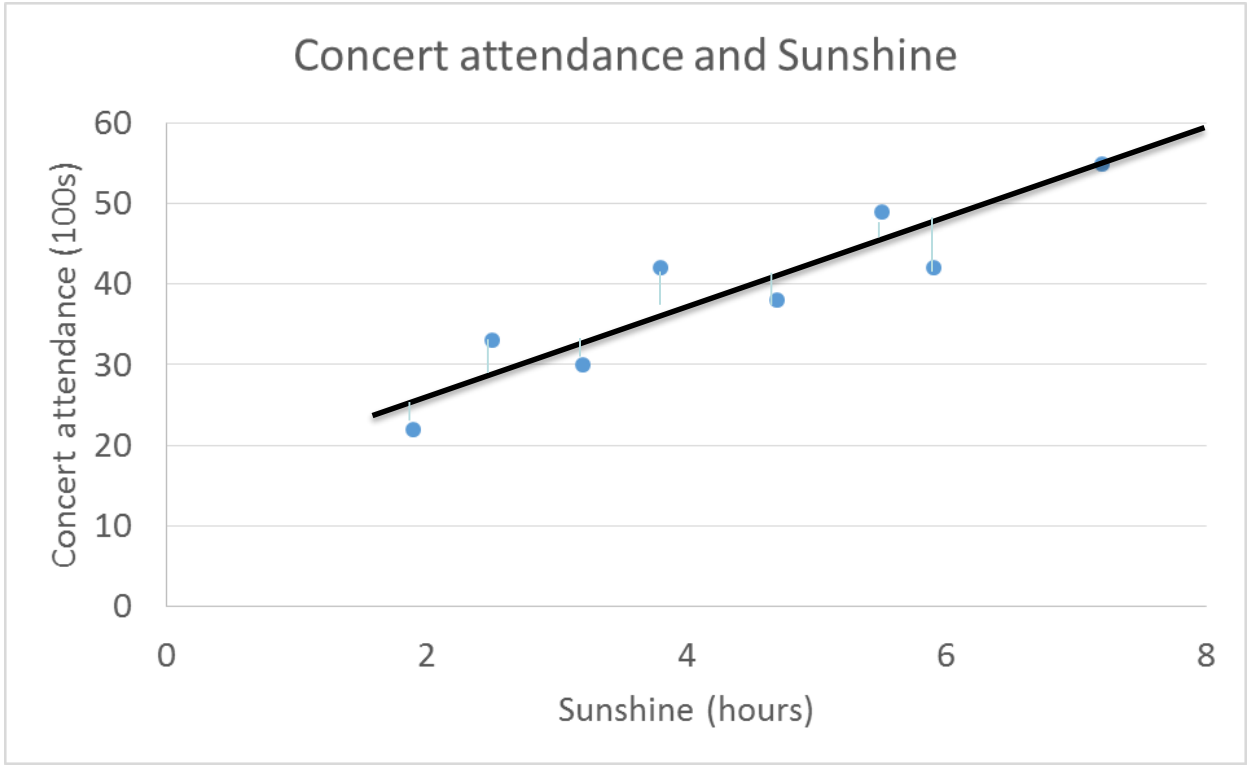
$$SST = \sum (y_i - \bar{y})^2$$



R-squared

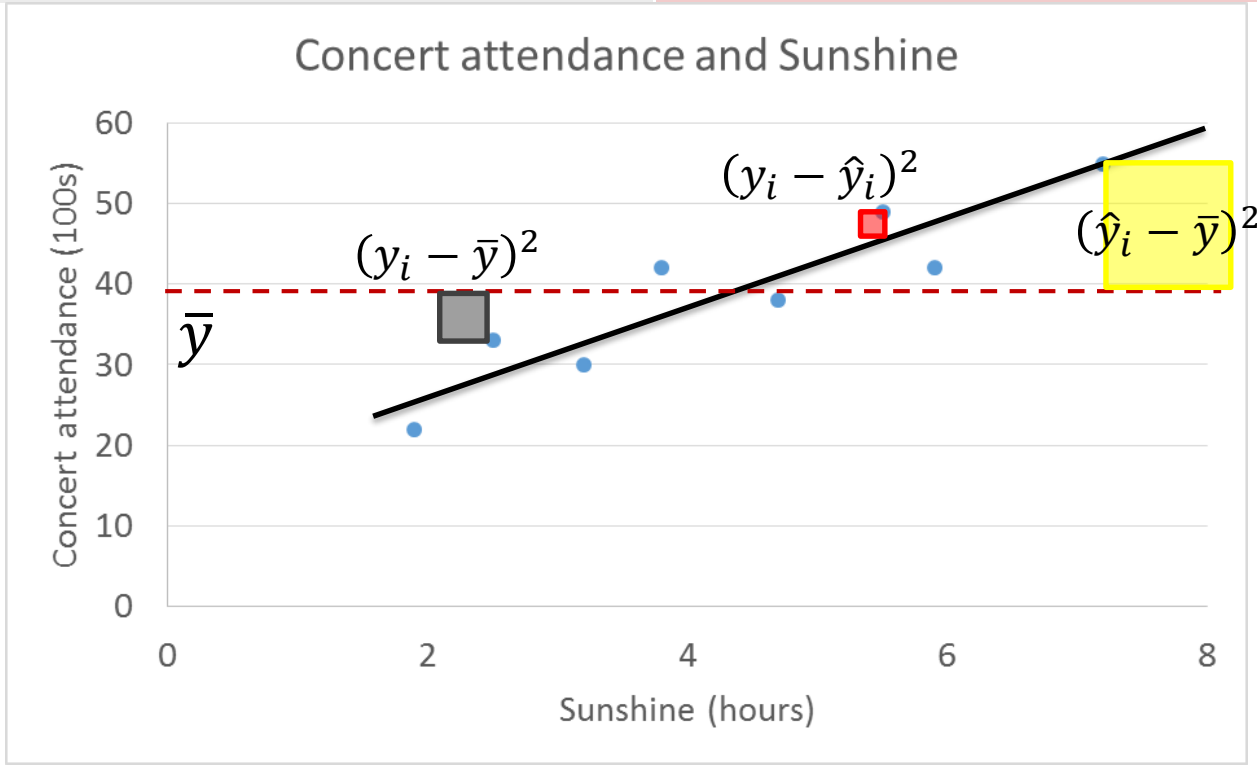
SSE (Recall Sum of Squares Within from ANOVA – the inherent noise) – This is the unexplained variation in data. The line indicates expected concert attendance when sunshine is ~~not~~ considered. This "model" has **small** residuals.

$$SSE = \sum (y_i - \hat{y}_i)^2$$



R-squared

Total Variation	Unexplained Variation	Explained Variation
$SST = \sum (y_i - \bar{y})^2$	$SSE = \sum (y_i - \hat{y}_i)^2$	$SSR = \sum (\hat{y}_i - \bar{y})^2$



Considering sunshine

Not considering sunshine

$$\boldsymbol{SST = SSR + SSE}$$

R-squared

$$SST = SSR + SSE$$

$$\Rightarrow \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = R^2$$

$$SST = \sum (y_i - \bar{y})^2$$

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

Covariance, Correlation and R^2

How do the interest rates of federal funds and the commodities futures index co-vary and correlate?

Day	Interest Rate	Futures Index
1	7.43	221
2	7.48	222
3	8.00	226
4	7.75	225
5	7.60	224
6	7.63	223
7	7.68	223
8	7.67	226
9	7.59	226
10	8.07	235
11	8.03	233
12	8.00	241

Covariance, Correlation and R^2

Day	Interest Rate	Futures Index	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x}) * (y - \bar{y})$
1	7.43	221	-0.314	-6.083	1.911
2	7.48	222	-0.264	-5.083	1.343
3	8.00	226	0.256	-1.083	-0.277
4	7.75	225	0.006	-2.083	-0.012
5	7.60	224	-0.144	-3.083	0.445
6	7.63	223	-0.114	-4.083	0.466
7	7.68	223	-0.064	-4.083	0.262
8	7.67	226	-0.074	-1.083	0.080
9	7.59	226	-0.154	-1.083	0.167
10	8.07	235	0.326	7.917	2.580
11	8.03	233	0.286	5.917	1.691
12	8.00	241	0.256	13.917	3.560
Mean	**7.74**	**227.08**		**Sum**	**12.216**
StDev	**0.22**	**6.07**			

$$Cov = \frac{12.216}{11} = 1.111$$

$$r = \frac{1.111}{0.22 * 6.07} = 0.815$$

$$R^2 = 0.815^2 = 0.665$$

Covariance, Correlation and R^2 - SUMMARY

- **Covariance**

  Tells you the direction of relationship between 2 variables

- **Correlation Coefficient**

  Tells you the direction AND strength of linear relationship between 2 variables

- **R^2**

  Tells you what percentage of the variation in y can be explained by the model (or equivalently, by the independent variable(s)).

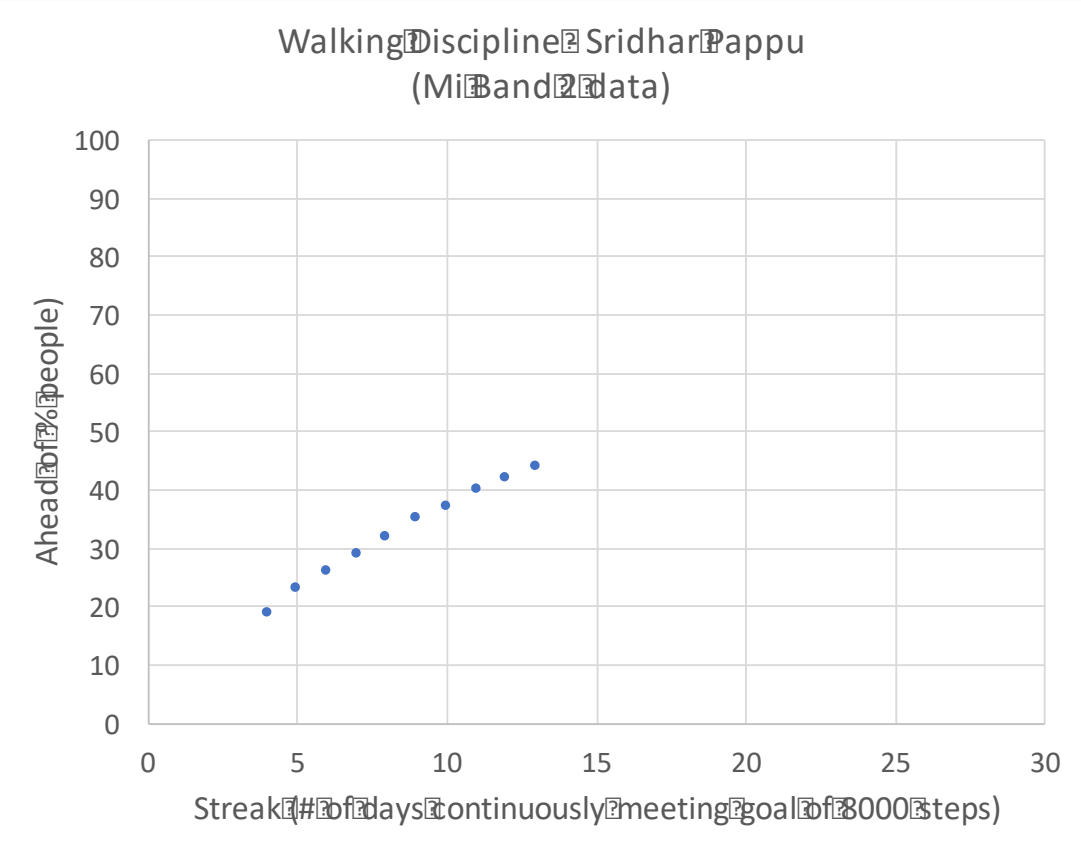
Welcome to the Learning Models

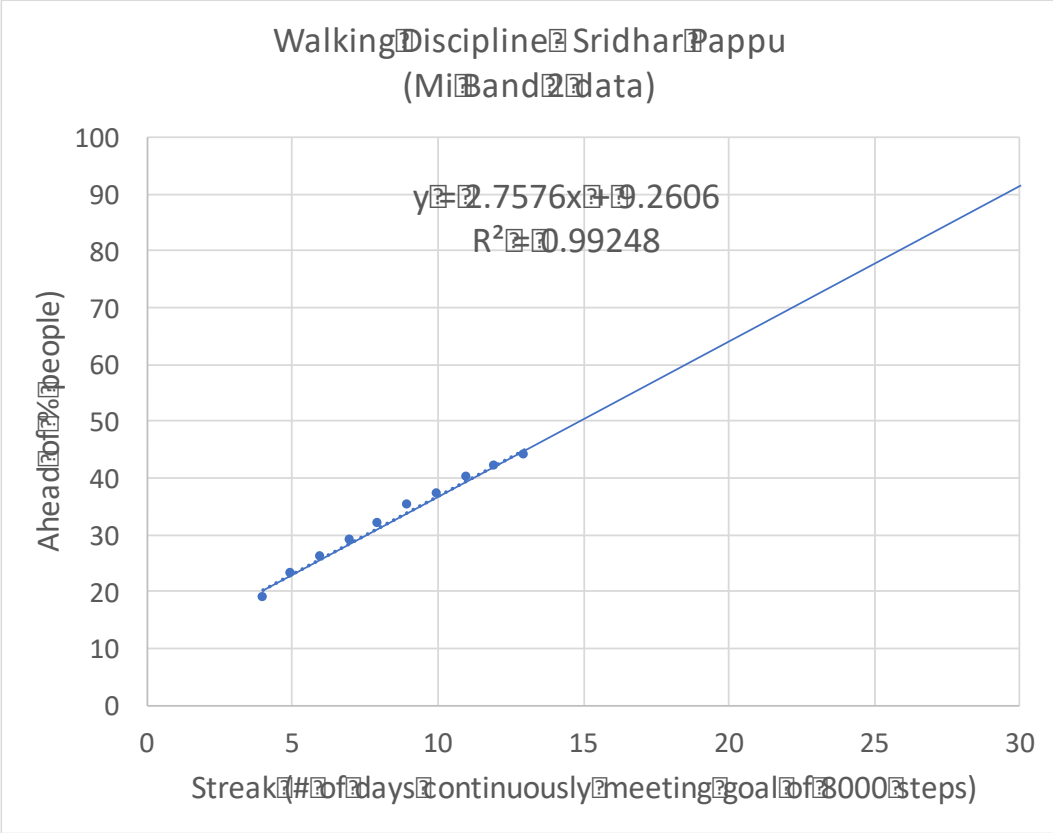
- Linear regression: A **regression** model where class/dependent/target variable is **numeric**

- Logistic regression: A **classification** model where class/dependent/target variable is **categorical**

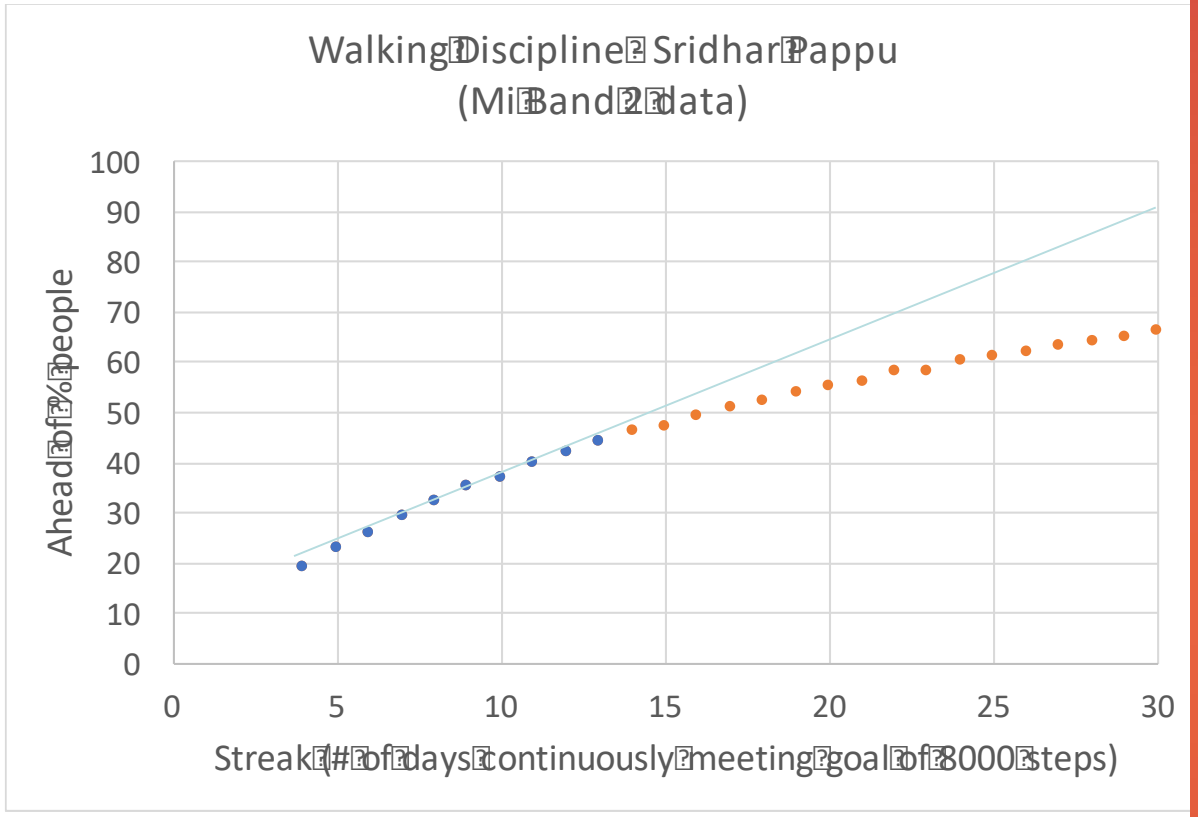
Linear Regression

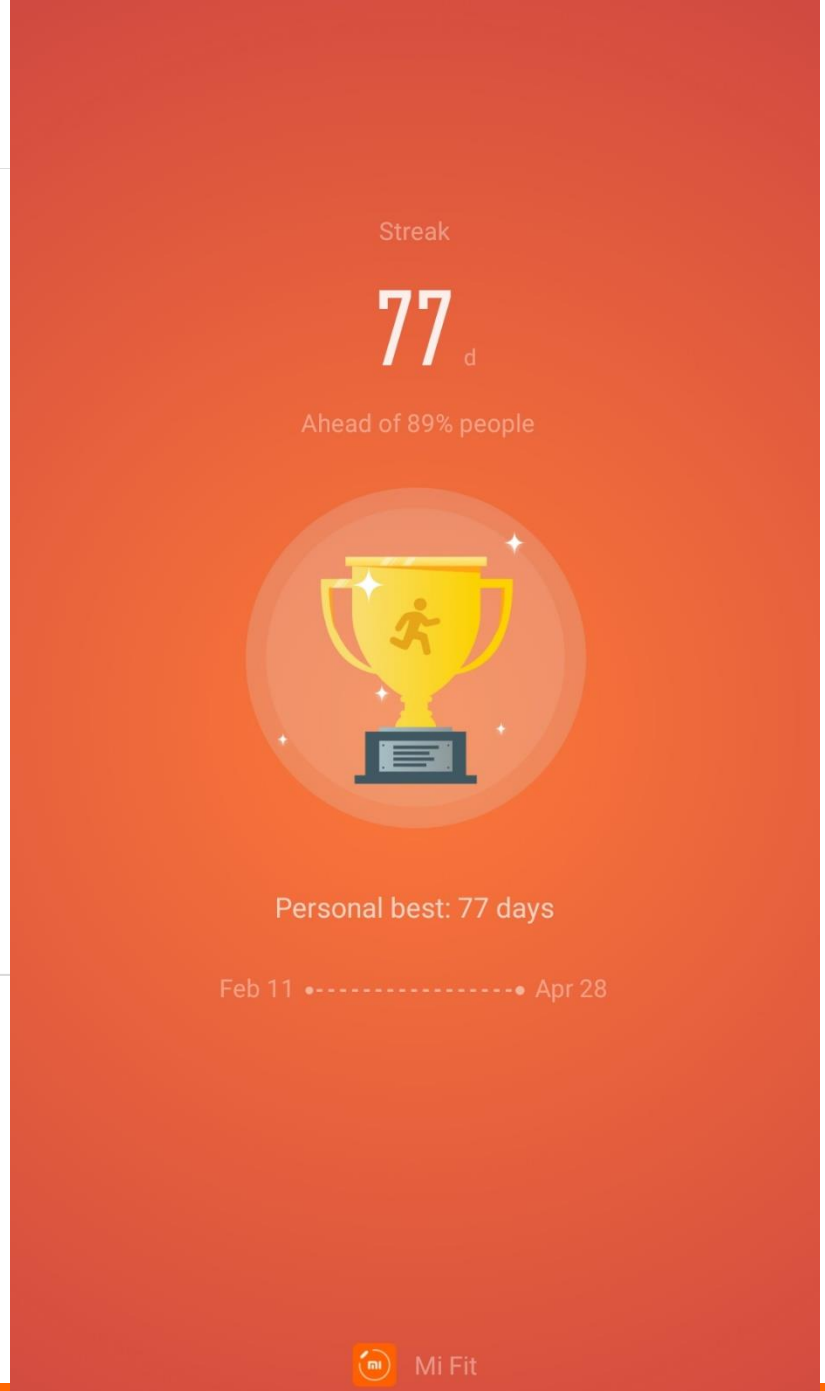
Linear Regression





Linear Regression





Be careful when extrapolating.

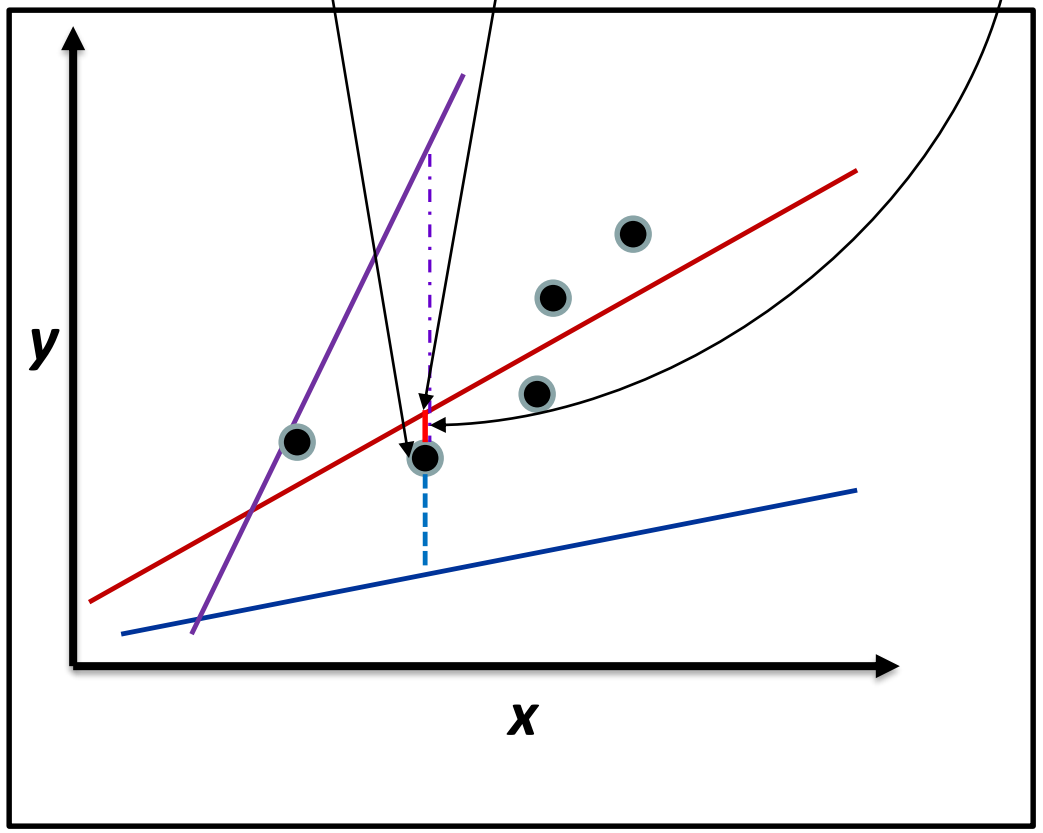
Extrapolation is done assuming that the same process that generated observed data is continuing in the unseen region as well.

How to Pick the Best Model?

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ (Probabilistic model)}$$

$$y = E(Y|X = x) + \varepsilon$$

Recall: Conditional Expected Value…Conditional Expectation of a Random Variable…Conditional Mean of a Random Variable

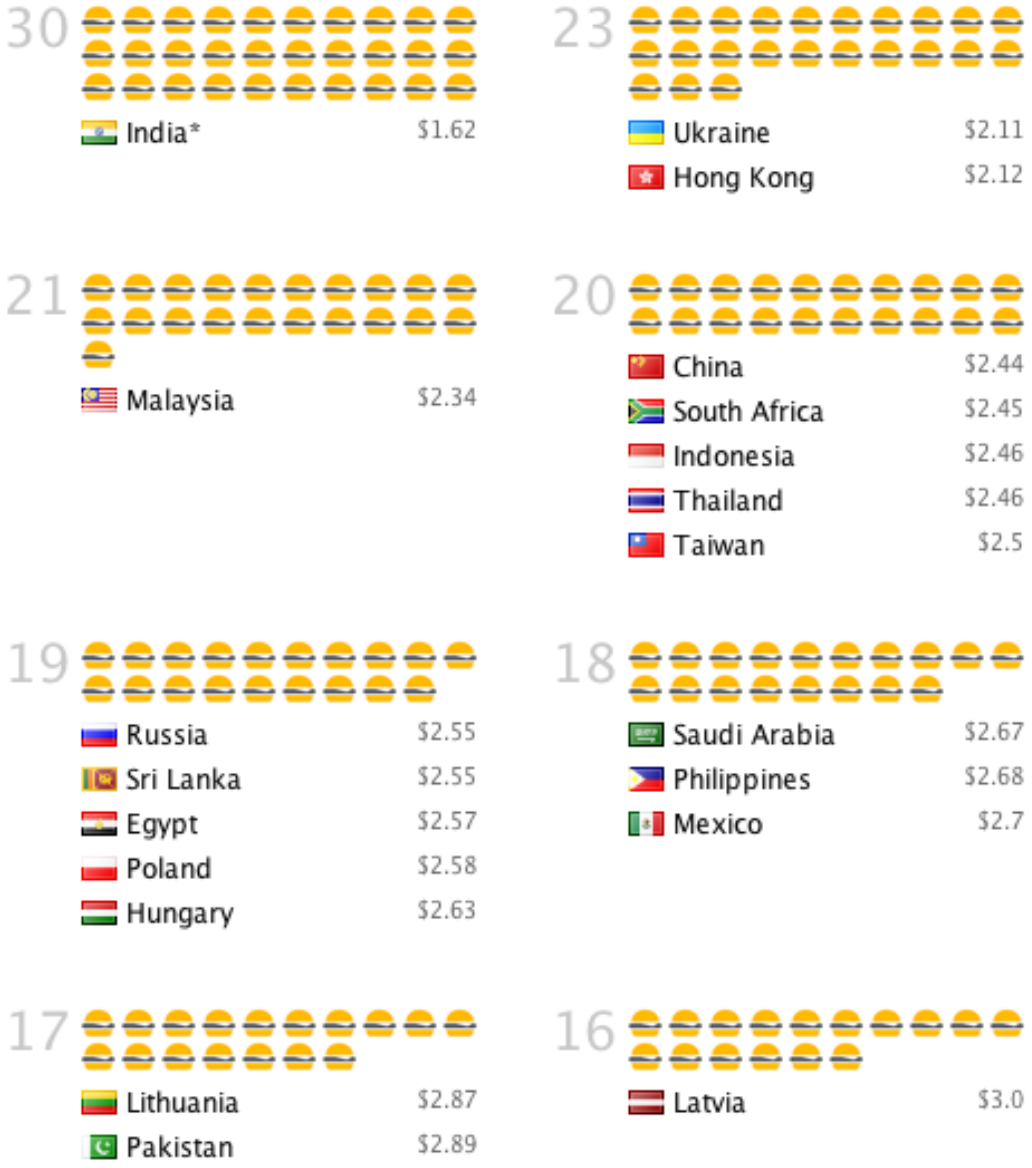


The lines whose residual error on all points is the least is the best line.

To ensure residual errors don't cancel, we take squares of residual errors.

THE BIG MAC INDEX

How many burgers you get for $50 USD?



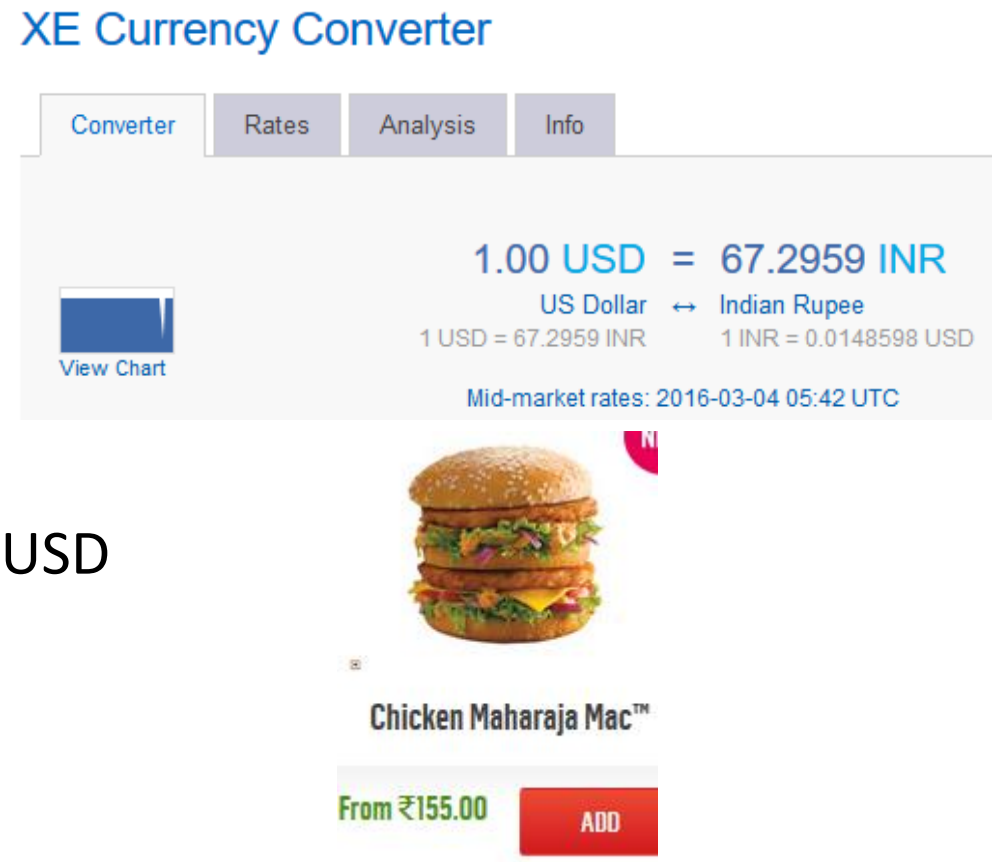
Source: The Economist (Jan 2012)

* Chicken burger

CSE 73026

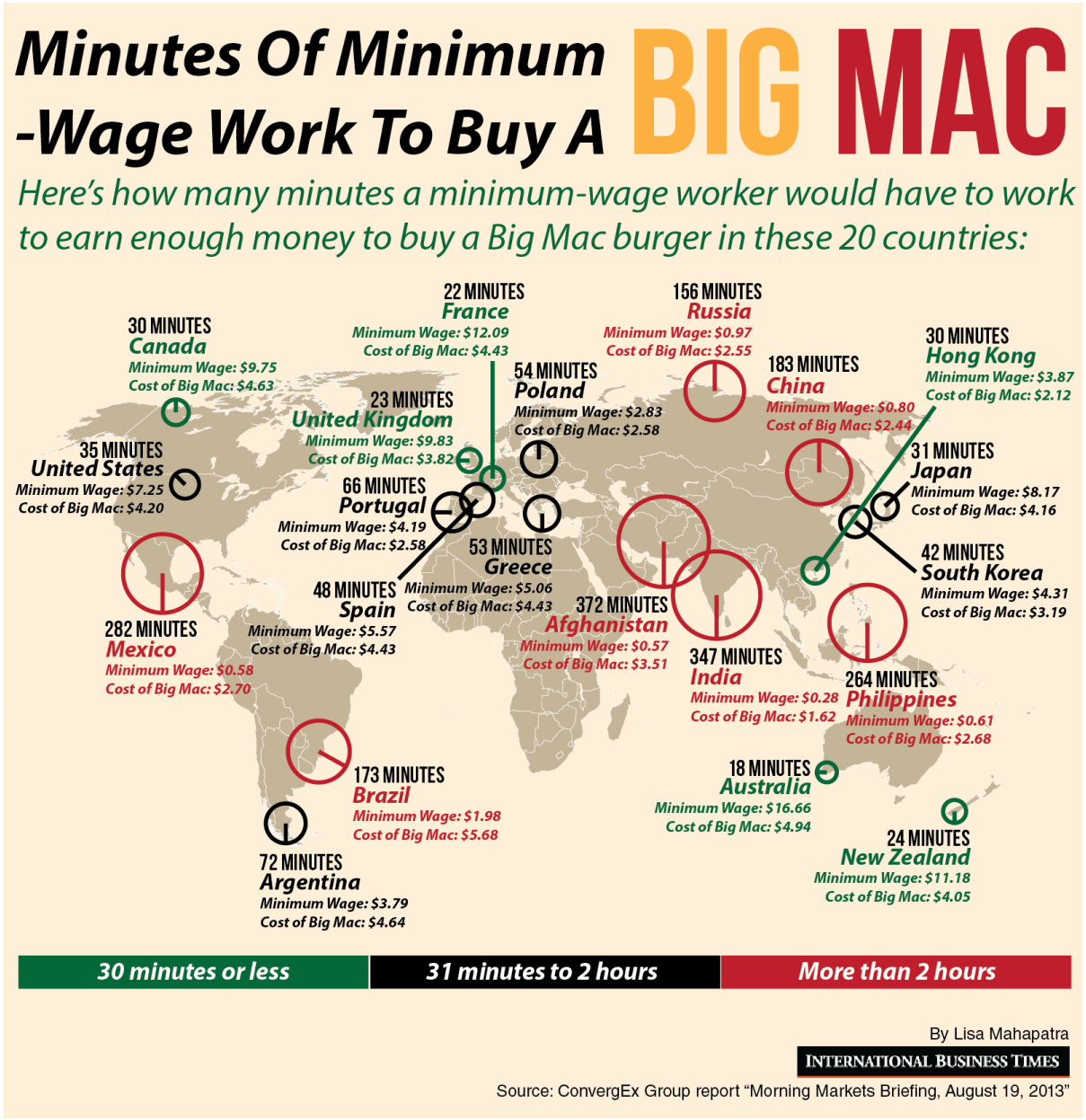
Burgernomics: Overvalued or Undervalued Currencies?

- Big Mac price in the US: $ 4.93
- Maharaja Mac price in India: Rs 155
- Implied PPP is 155/4.93 = Rs 31.44/$
- Actual exchange rate = Rs 67.2959/$
- $\frac{31.44-67.2959}{67.2959} = -0.53$
- Rupee undervalued by 53% against the USD

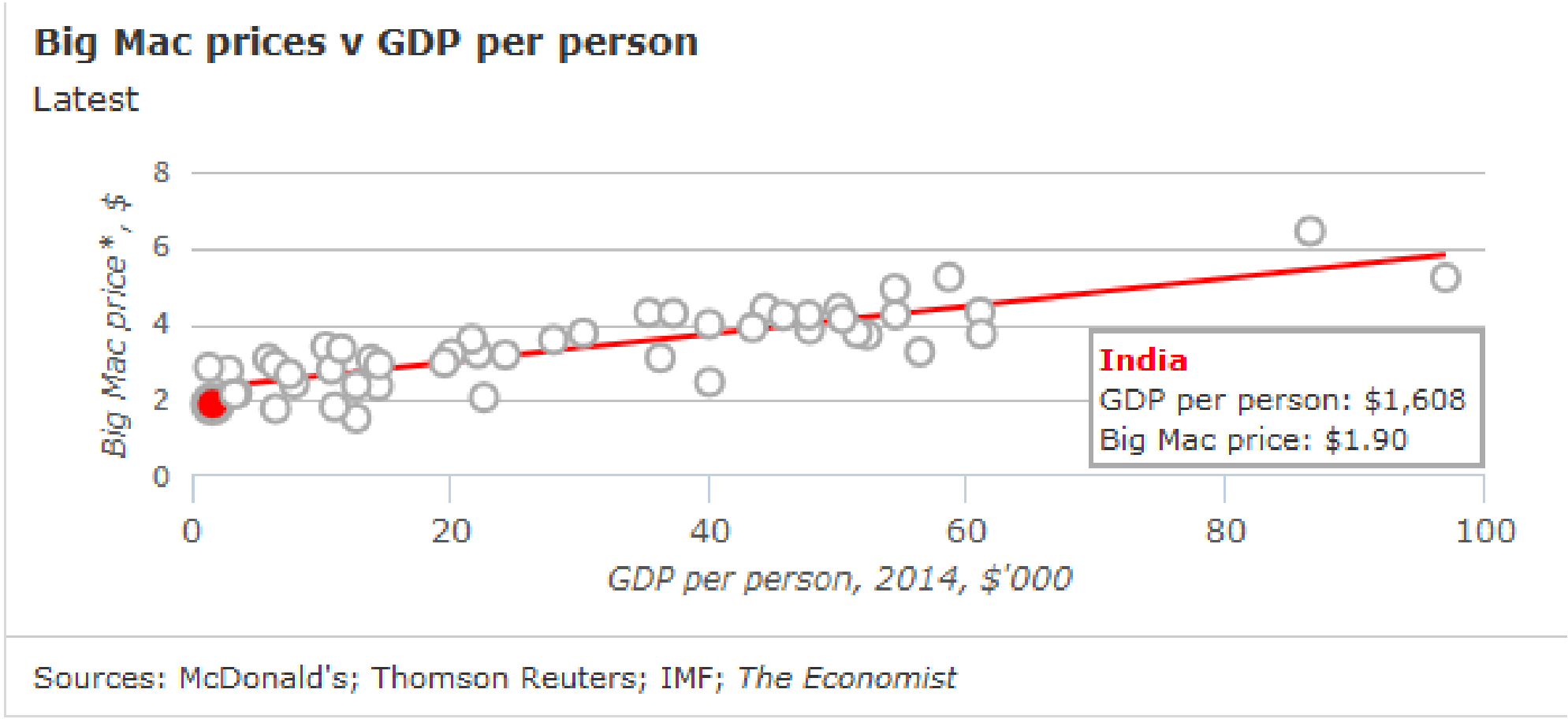


Global prices for a Big Mac in July 2016 based on a survey conducted in January 2016 by IMF, McDonald's, Thomson Reuters and The Economist

Burgernomics by UBS Wealth Management Research

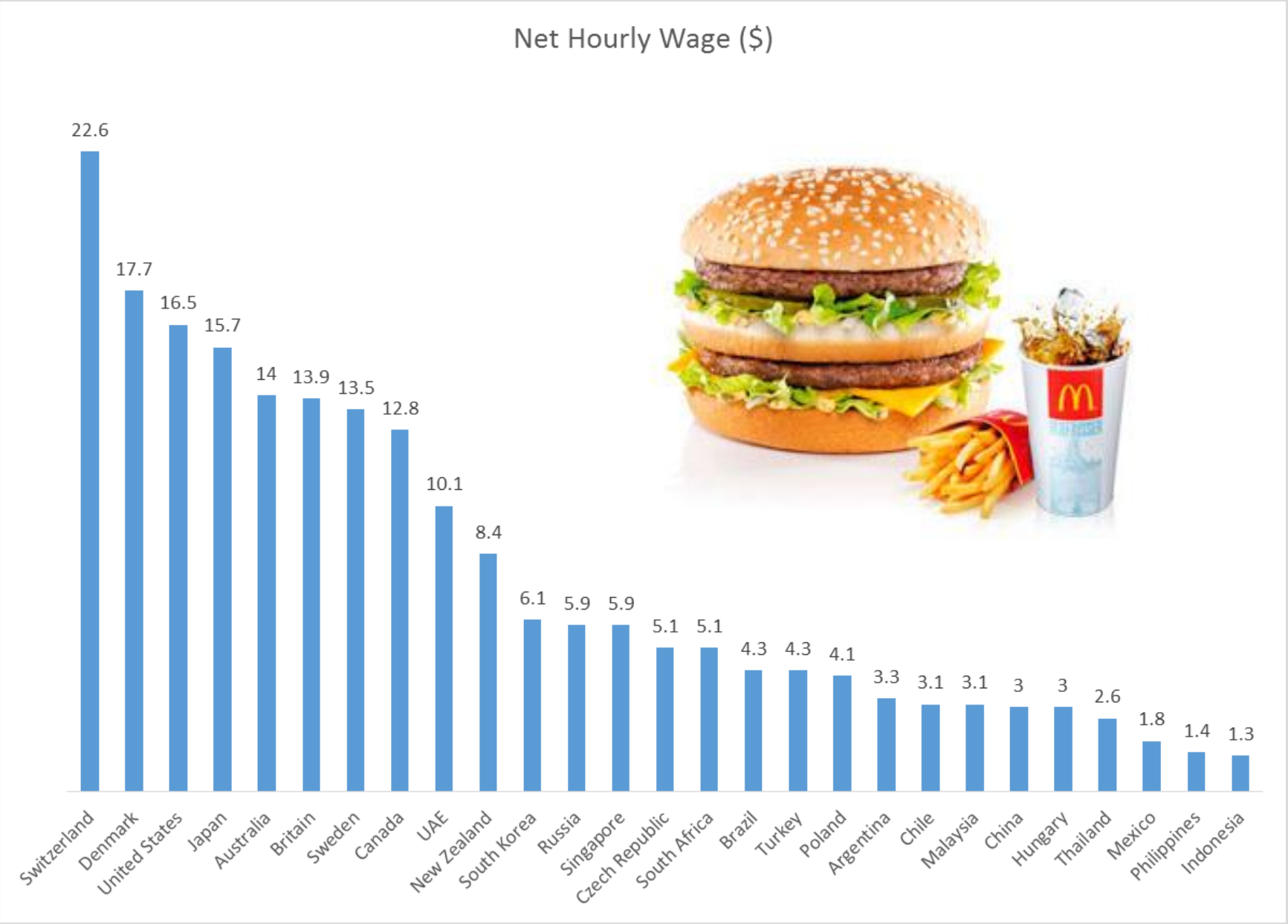


Burgernomics

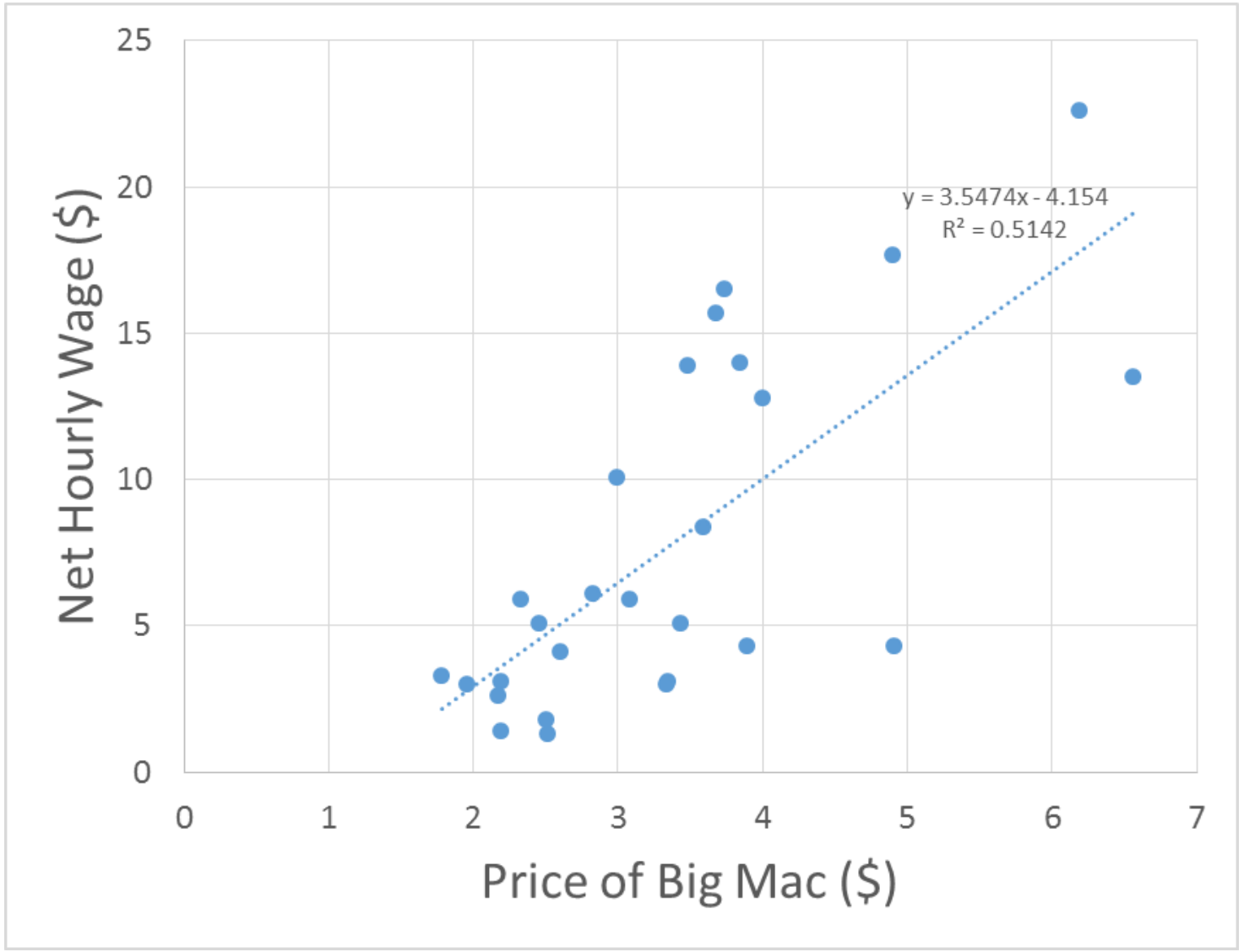


Source: http://www.economist.com/content/big-mac-index
Last accessed: March 04, 2016

Determining the Equation of the Regression Line - Excel



Determining the Equation of the Regression Line - Excel



Sample Software Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.717055011
R Square	0.514167888
Adjusted R Square	0.494734604
Standard Error	4.21319131
Observations	27

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	469.6573265	469.6573265	26.4581054	2.57053E-05
Residual	25	443.7745253	17.75098101		
Total	26	913.4318519			

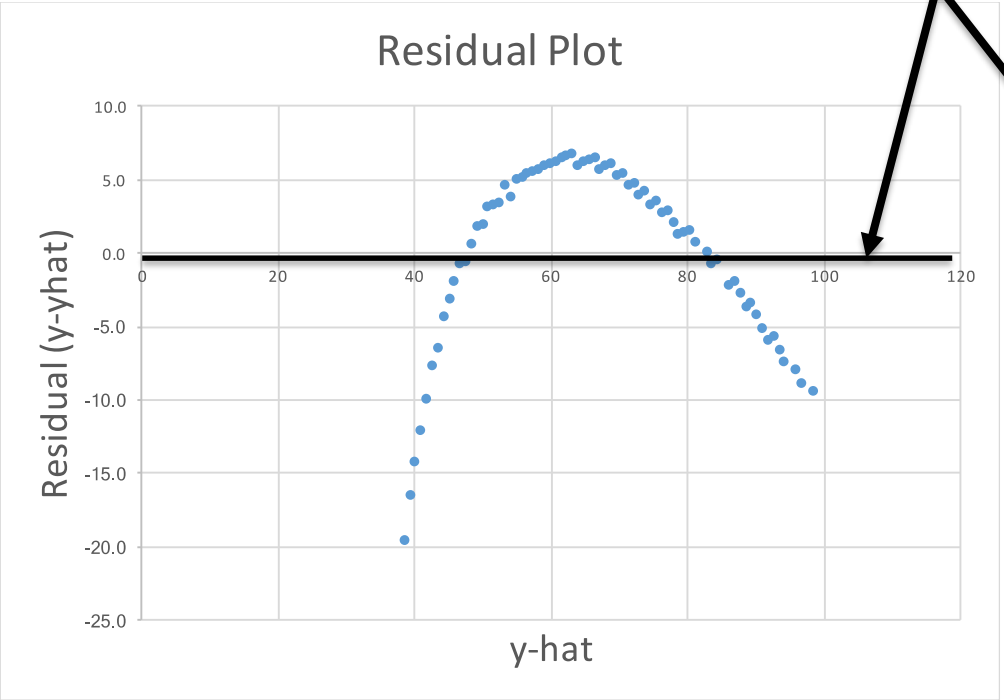
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	-4.154014573	2.447784673	-1.697050651	0.102104456	-9.195321476	0.88729233	-10.97705723	2.669028089
Big Mac Price ($)	3.547427488	0.689658599	5.143744297	2.57053E-05	2.127049014	4.967805962	1.625048409	5.469806567

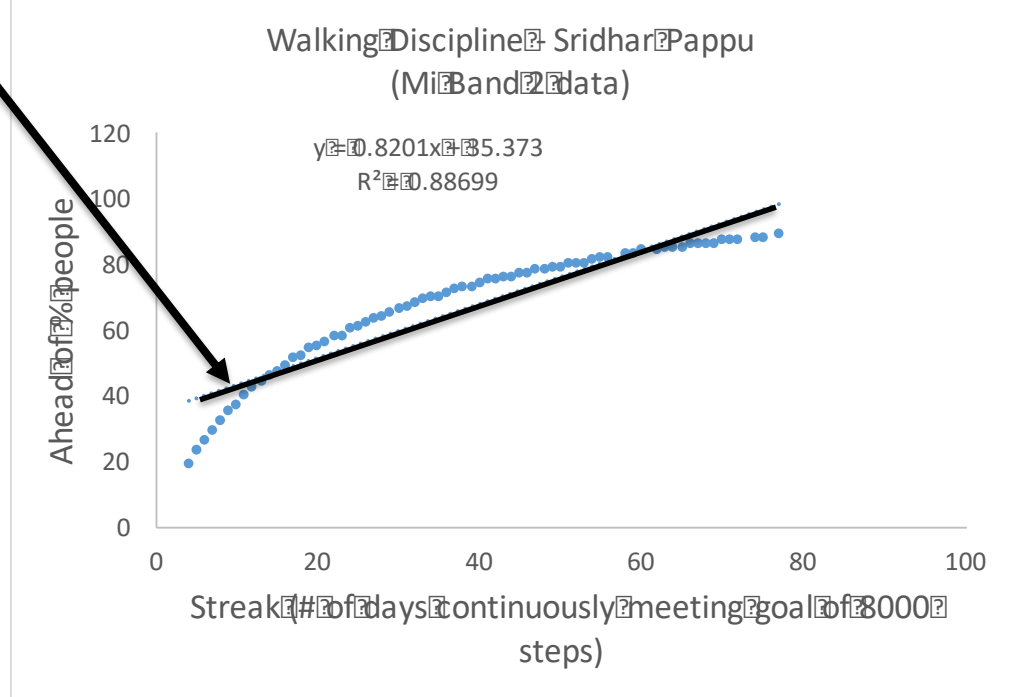
WAYS OF TESTING HOW WELL THE REGRESSION LINE FITS DATA

Assumptions of the Regression Model – Residuals Analysis

The model is linear

Zero residual line:
The regression line





CSE 73026

Assumptions of the Regression Model

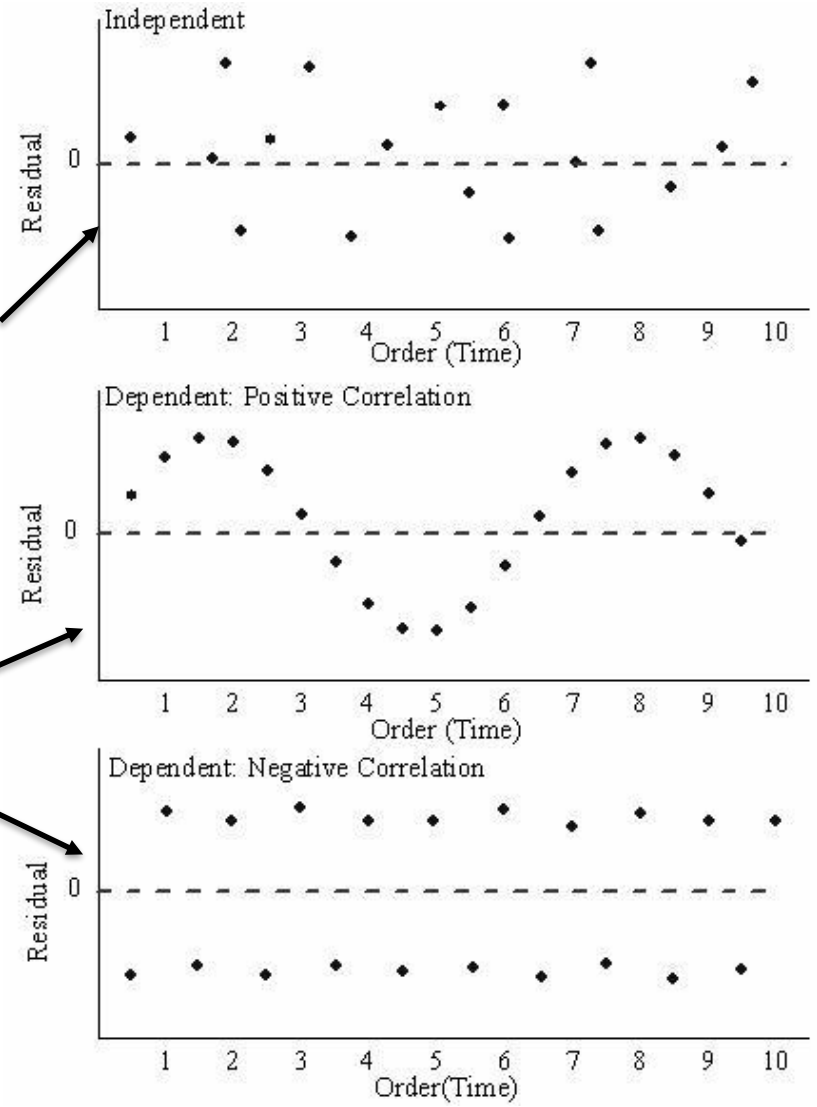
The error terms are independent

- Plot against any time or spatial variables where **order of observation is important**.

Independent

- Time series methods are more appropriate in such situations than regular regression.

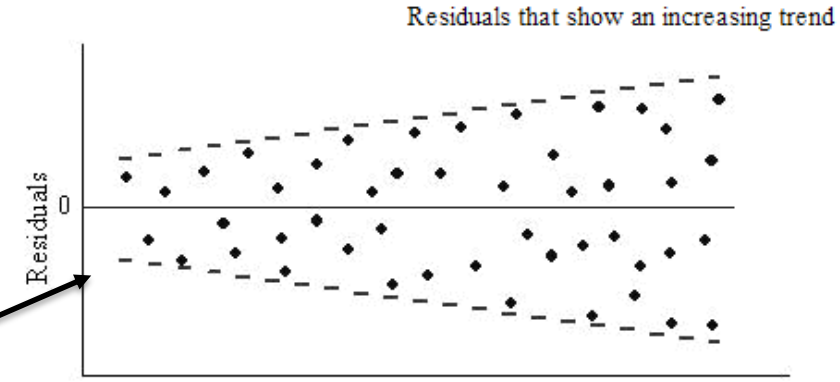
Dependent

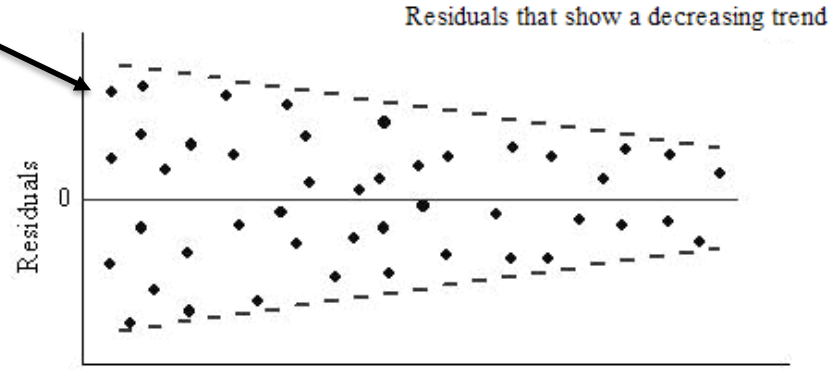


Assumptions of the Regression Model

The error terms have constant variances (homoscedasticity as opposed to heteroscedasticity)

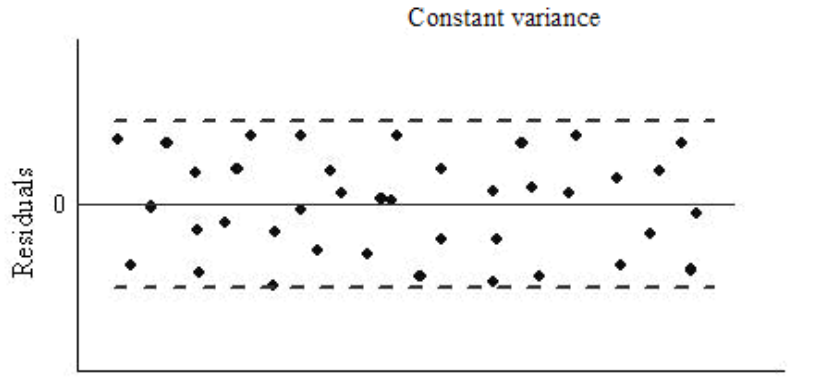
Heteroscedastic





- RMSE (Root Mean Square Error) of Regression or Standard Error of the Estimate will be misleading as it will underestimate the spread for some x_i and overestimate for others.

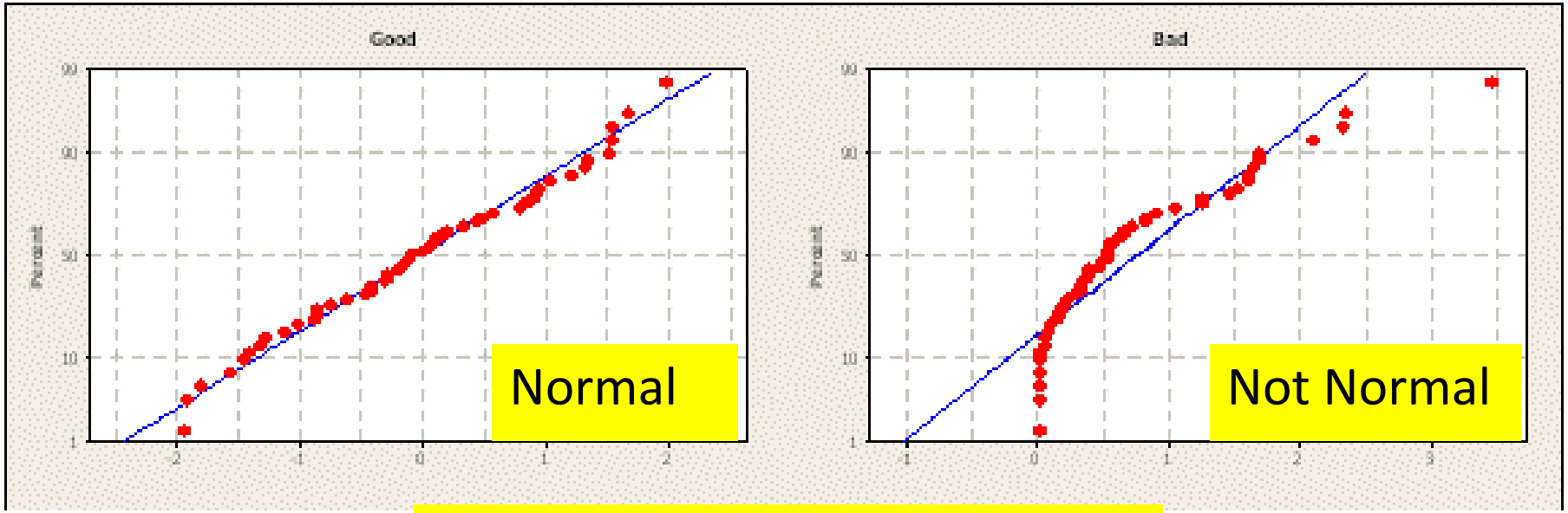
Homoscedastic



CSE 7302c

Assumptions of the Regression Model

The error terms are normally distributed



The quantile-quantile (q-q) plot

x-axis: Theoretical quantiles in a standard normal distribution

y-axis: Observed quantiles in the sample

Q-Q plot (Excel)

Quantiles are cutpoints dividing the range of a probability distribution into contiguous intervals with equal probabilities, or dividing the observations in a sample in the same way.
https://en.wikipedia.org/wiki/Quantile

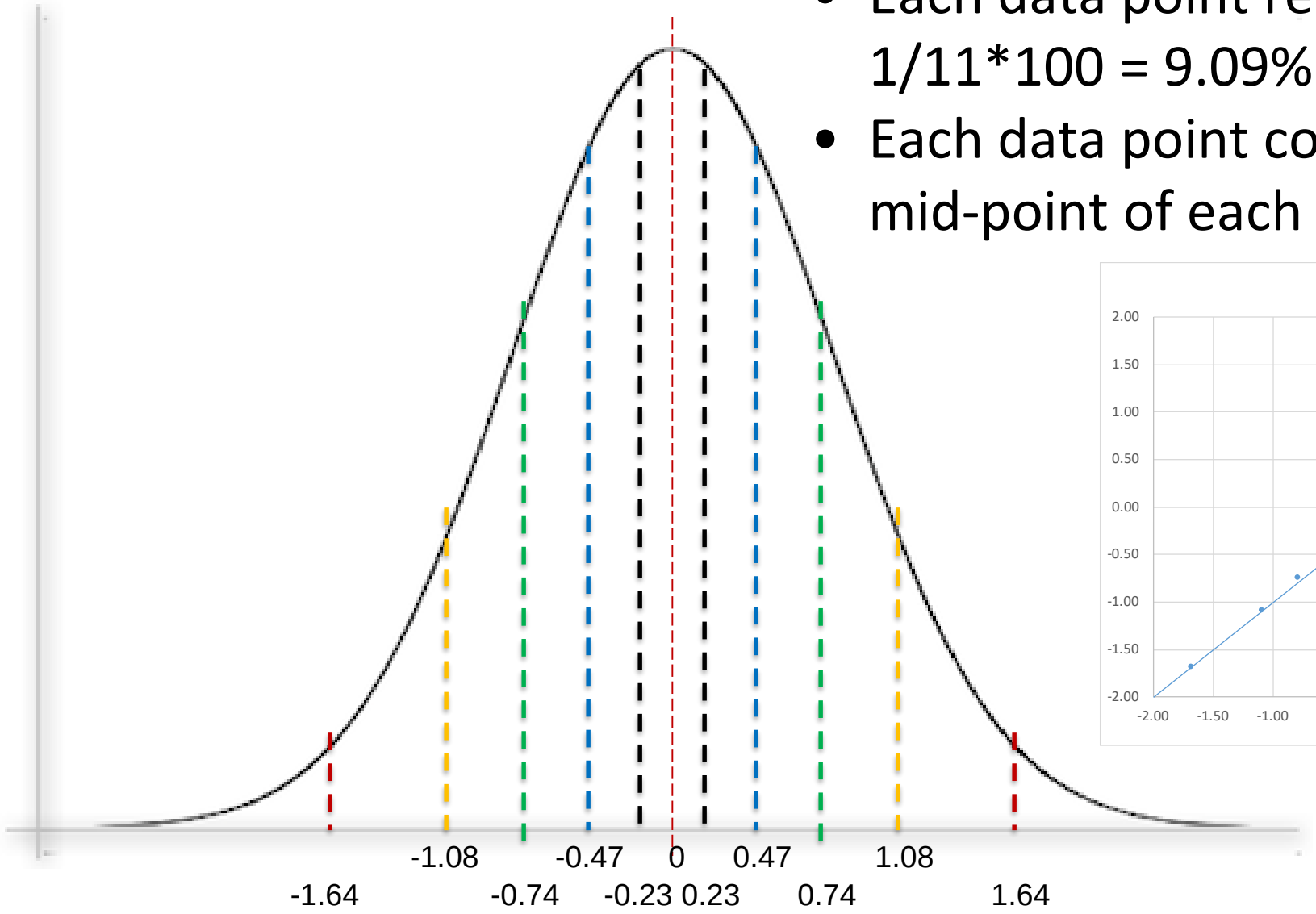
The quantile-quantile (q-q) plot is used to validate distributional assumptions of **a data set**.

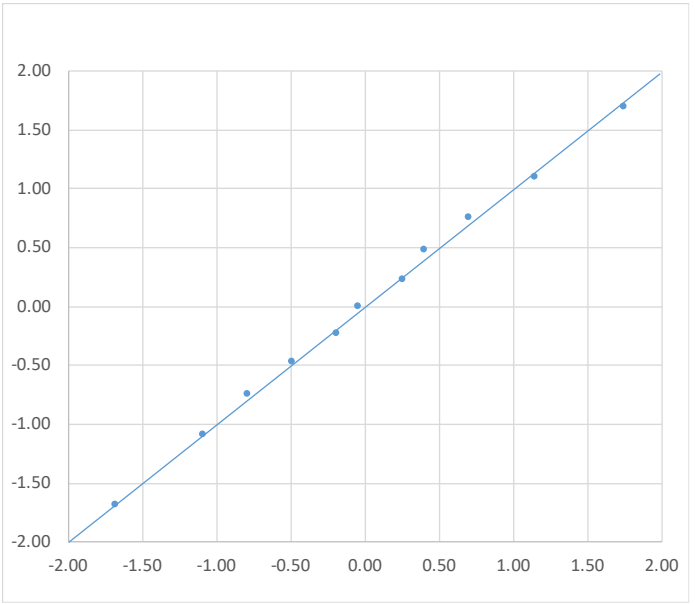
In linear regression, **this data set** is the residual errors.

If the normality assumption holds true, then the z-scores of the residuals should be equal to the expected z-scores at corresponding quantiles.

Q-Q plot (Excel)

- 11 data points cover 100% area
- Each data point represents 1/11*100 = 9.09% area (or 0.091)
- Each data point considered as mid-point of each of 11 bins



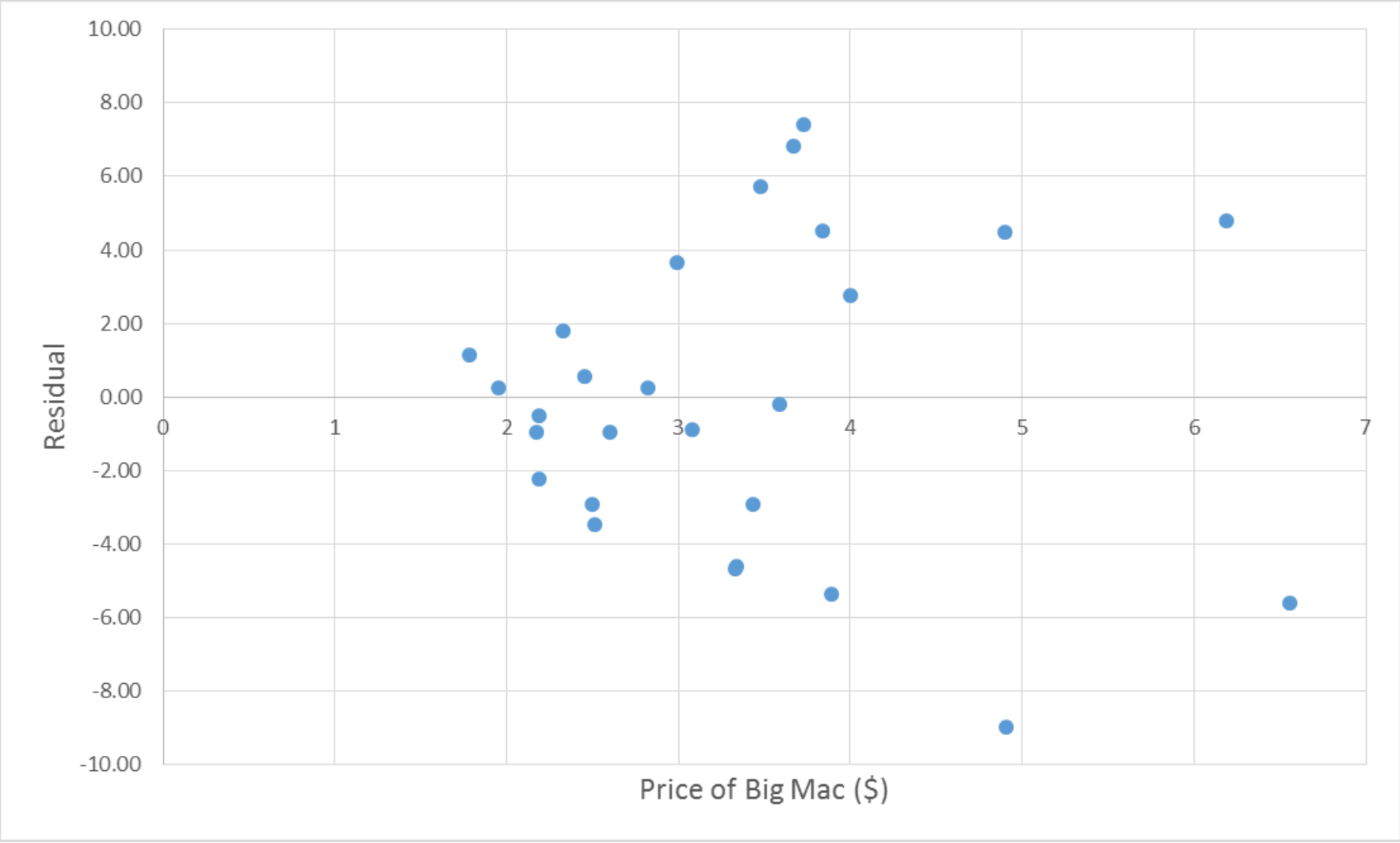


Interpreting Residuals

http://www.stat.berkeley.edu/~stark/SticiGui/Text/regressionDiagnostics.htm

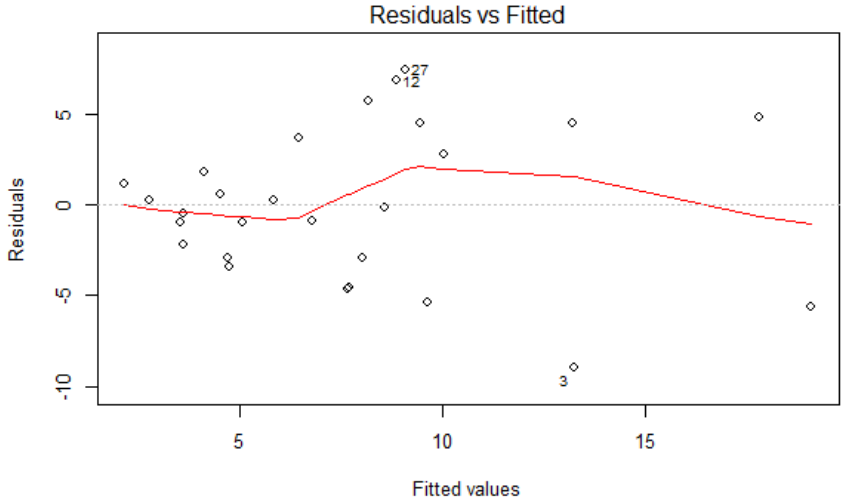
Residual Analysis – Big Mac

Which assumption is getting violated?

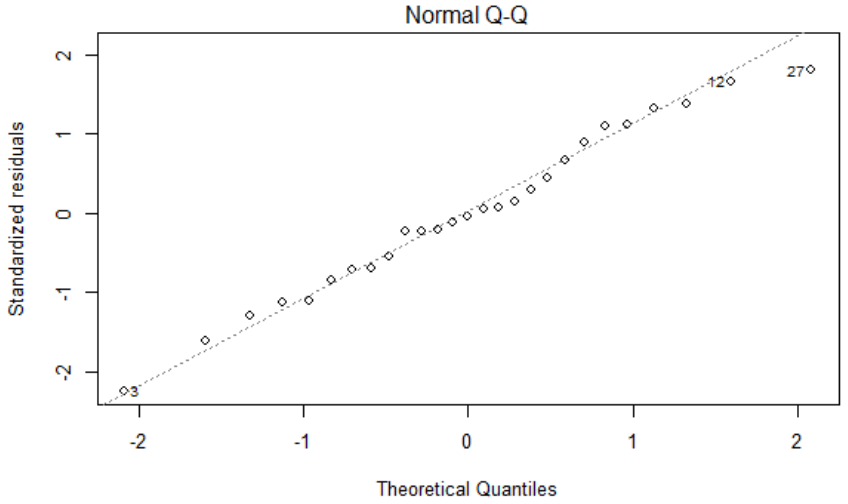


Residuals – Big Mac

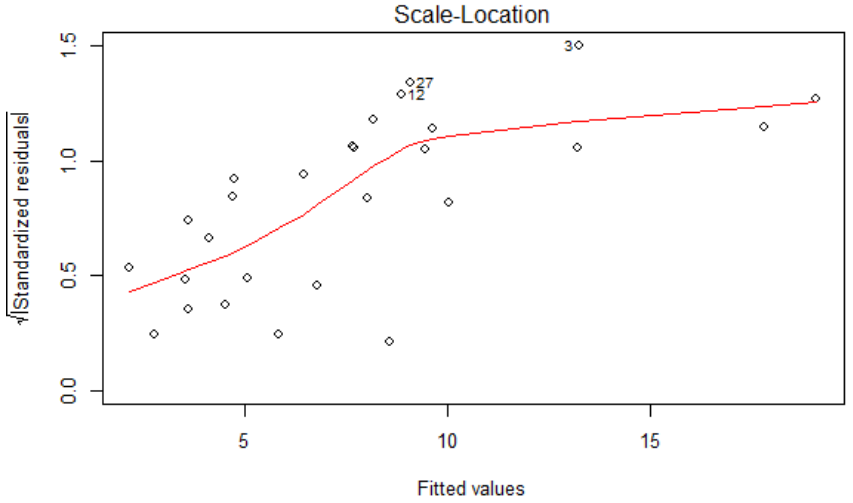
Is a wrong model fitted (linear or quadratic, etc.)?



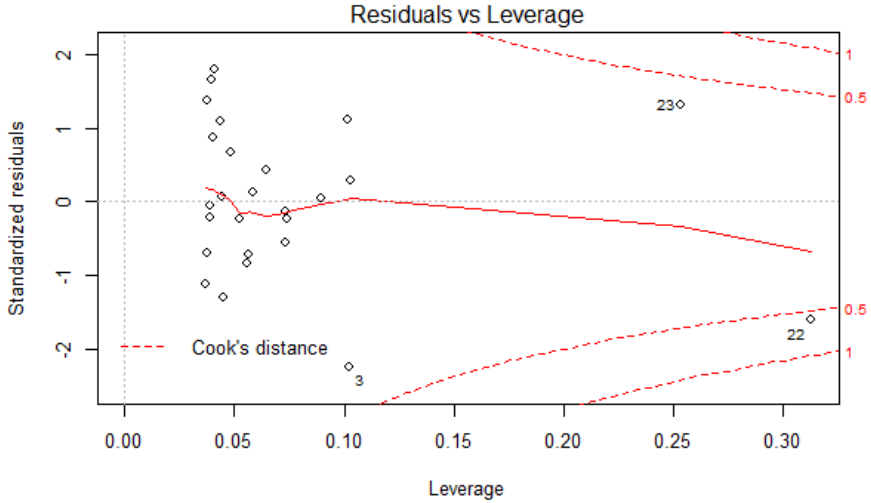
Are the residuals normally distributed?



Is the data homoscedastic?

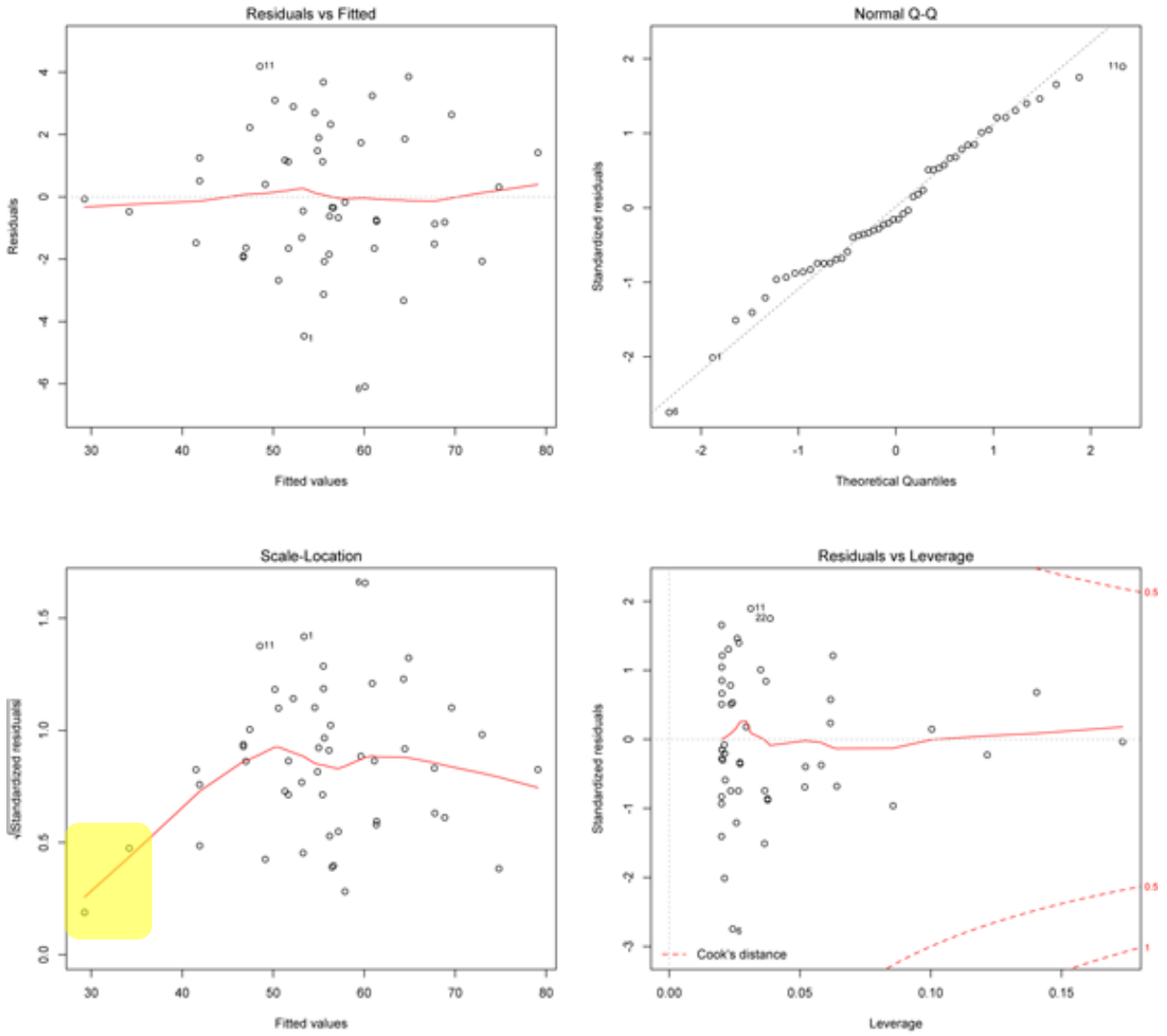


Are there influential outliers?



CSE 73026

Caution – Is there heteroscedasticity here?



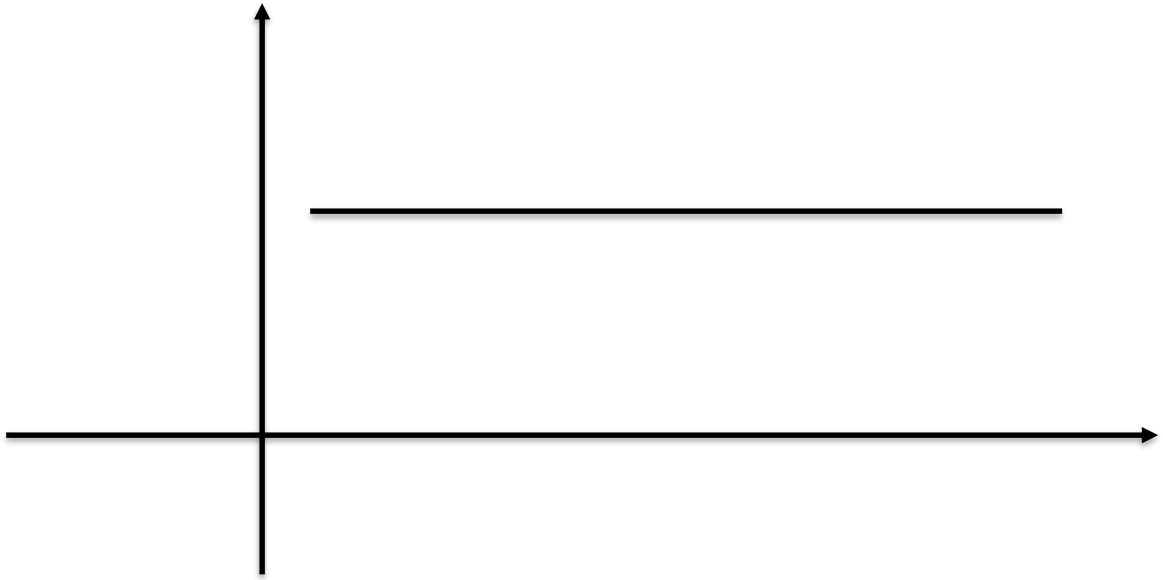
Fixing Non-normality and Heteroscedasticity

Transformation of data (square root, logarithm, etc.) can help correct normality and unequal variances problems.

HYPOTHESIS TESTS FOR THE SLOPE OF THE REGRESSION MODEL AND TESTING THE OVERALL MODEL

Testing the Slope

If the Net Hourly Wage is NOT dependent on the Big Mac price, we could use its mean value as predictor of the *y* for all values of *x*, i.e., slope is 0. As slope deviates from 0, the model adds more predictability.



Testing the Slope

What is the Null Hypothesis?

$H_0: \beta_1 = 0$

What is the Alternative Hypothesis?

$H_1: \beta_1 \neq 0$

t Test of the Slope

$$t = \frac{b_1 - \beta_1}{s_b}$$

Where $s_b, the\ standard\ error\ of\ the\ slope = \frac{SE}{\sqrt{SS_{xx}}}$

$$SS_{xx} = \sum (x - \bar{x})^2$$

$$\beta_1 = the\ hypothesized\ slope$$

Standard Error of the Estimate

Standard error of the estimate, SE, is the standard deviation of the errors of the regression model.

$$SE = \sqrt{\frac{\sum(e_i - \mu_e)^2}{df}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}},$$

$$where\ e_i = (y_i - \hat{y}_i)\ and\ \ \mu_e = 0.$$

$$SE = \sqrt{MSE}, where\ MSE = \frac{SSE}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$$

Degrees of freedom, *df* = *n-k-1* where *k* is the number of regressors or independent variables

t Test of the Slope – Big Mac - Excel

t = 5.1437 from $t = \frac{b_1 - \beta_1}{s_b}$ where $s_b = \frac{SE}{\sqrt{SS_{xx}}}$

At $\alpha = 0.05$, the critical region for a 2-tailed test is

$t_{0.025,25} = \pm 2.060$ *R code: qt(0.025,25)*

Since *t* value calculated from the sample slope is in the rejection region, we reject the null hypothesis.

The *p*-value corresponding to the t-statistic for this sample is 0.0000128 *R code: pt(5.1437,25,lower.tail = FALSE)*. Since this is less than 0.025, we reject the null hypothesis.

(Note: All software output double this value for 2-tailed tests to allow easier comparison with α instead of with α/2).

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-4.154014573	2.447784673	-1.697050651	0.102104456
Big Mac Price ($)	3.547427488	0.689658599	5.143744297	2.57053E-05

Testing the Overall Model

F test and its associated ANOVA table is used to test the overall model. In multiple regression, it tests that at least one of the regression coefficients is different from 0. In simple regression, we have only one coefficient, β_1. So *F* test for overall significance tests the same thing as *t* test.

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

Testing the Overall Model

$$F = \frac{\frac{SSR}{df_{reg}}}{\frac{SSE}{df_{err}}} = \frac{MSR}{MSE}$$

where $df_{reg} = k, df_{err} = n - k - 1$

and $k = the\ number\ of\ independent\ variables$

Testing the Overall Model – Big Mac - Excel

$F = 26.4581$

Critical Fvalue, $F_{.05,1,25} = 4.2417$

R code: qf(0.05,1,25,lower.tail = FALSE)

Reject the null hypothesis for overall significance.

The *p*-value corresponding to the *F* statistic of this sample is 0.0000257

R code: pf(26.4581,1,25,lower.tail = FALSE)

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	469.6573265	469.6573265	26.4581054	2.57053E-05
Residual	25	443.7745253	17.75098101		
Total	26	913.4318519			

Sample Software Output

SUMMARY OUTPUT

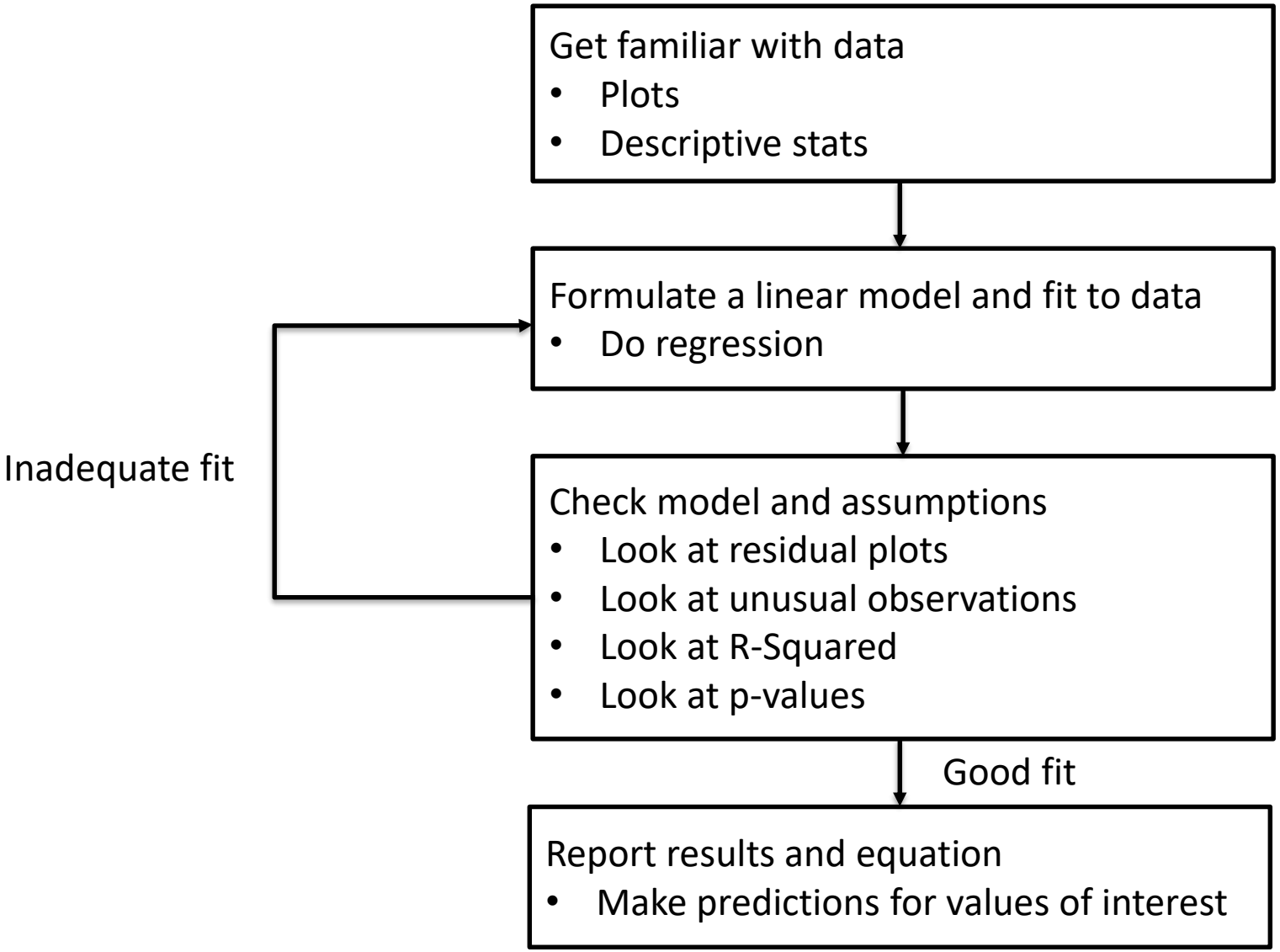
Regression Statistics	
Multiple R	0.717055011
R Square	0.514167888
Adjusted R Square	0.494734604
Standard Error	4.21319131
Observations	27

ANOVA

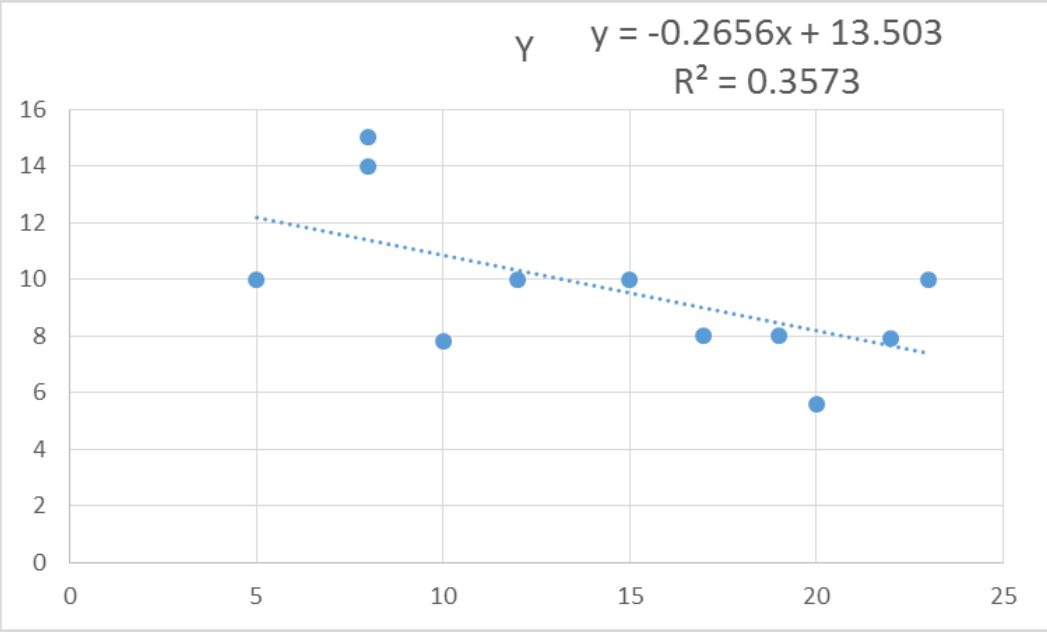
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	469.6573265	469.6573265	26.4581054	2.57053E-05
Residual	25	443.7745253	17.75098101		
Total	26	913.4318519			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	-4.154014573	2.447784673	-1.697050651	0.102104456	-9.195321476	0.88729233	-10.97705723	2.669028089
Big Mac Price ($)	3.547427488	0.689658599	5.143744297	2.57053E-05	2.127049014	4.967805962	1.625048409	5.469806567

Simple Linear Regression - Steps



R-Squared and Significance - Caution



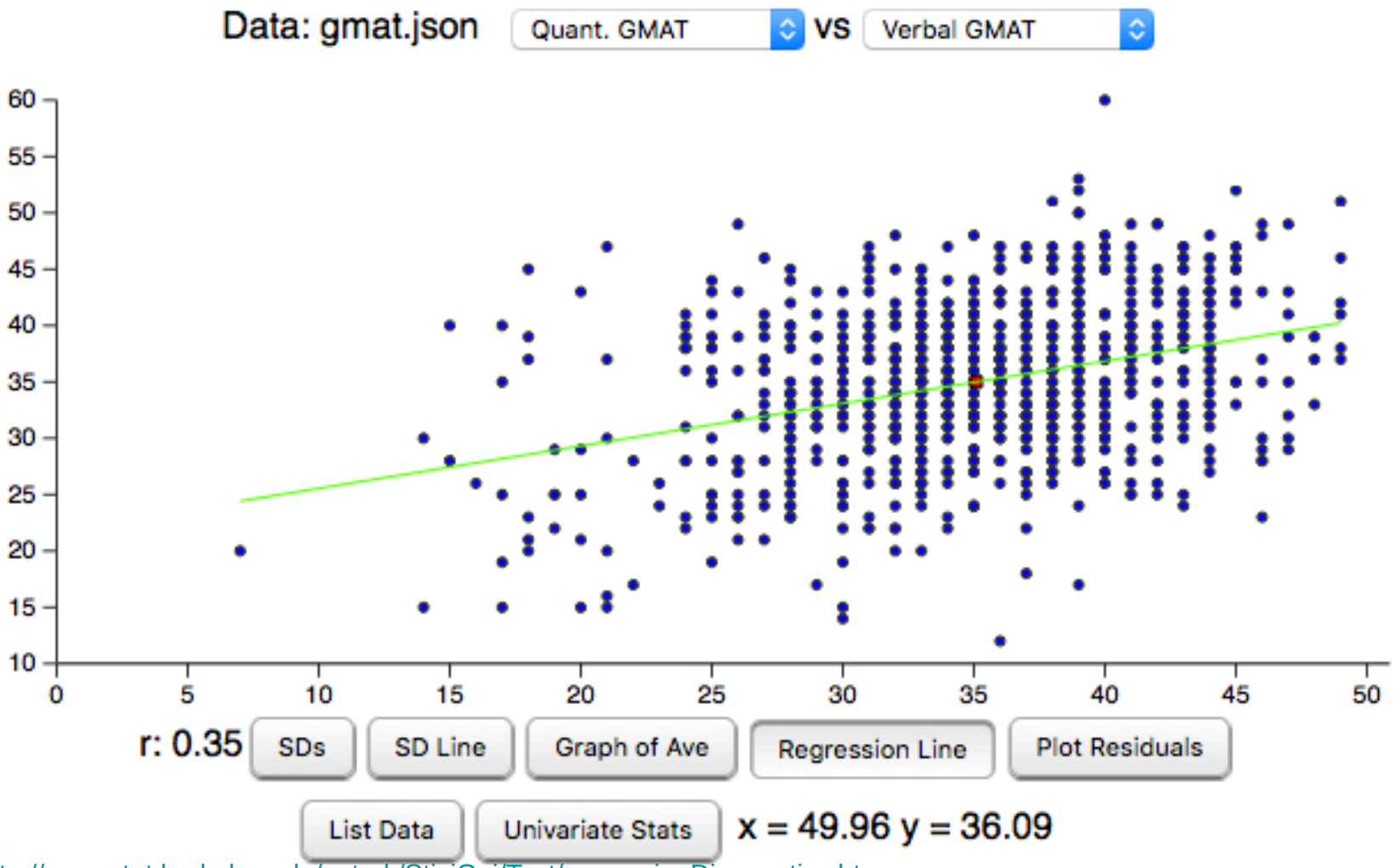
23	Regression Analysis						
24							
25	OVERALL FIT						
26	Multiple R	0.597718					
27	R Square	0.357267					
28	Adjusted R Square	0.285852					
29	Standard Error	2.335299					
30	Observations	11					
31							
32	ANOVA				Alpha	0.05	
33		*df*	*SS*	*MS*	*F*	*p-value*	*sig*
34	Regression	1	27.282857	27.28285699	5.002704	0.052125754	no
35	Residual	9	49.082598	5.453621951			
36	Total	10	76.365455				
37							
38		*coeff*	*std err*	*t stat*	*p-value*	*lower*	*upper*
39	Intercept	13.50289	1.8553076	7.277980002	4.67E-05	9.305894086	17.69988871
40	X	-0.26561	0.1187518	-2.23667255	0.052126	-0.53424402	0.003026342

- R^2 suggests that 35% of variation in *y* can be explained by variation in *x*.
- *t* and *F* tests show that coefficient is not significant and null hypothesis cannot be rejected.
- The 95% confidence interval of the slope, $b_1 \pm t_{crit} * s_b$, is (-0.534,0.003).

CSE 73026

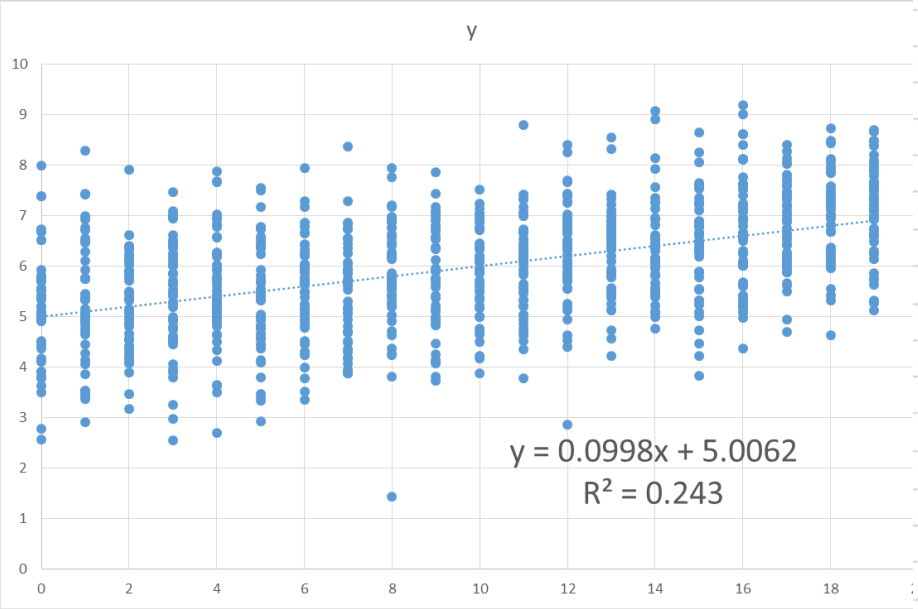
R-Squared and Significance - Caution

Figure 10-1: Residual Plot of the GMAT Data.



Source: http://www.stat.berkeley.edu/~stark/SticiGui/Text/regressionDiagnostics.htm
Last accessed: May 31, 2016

R-Squared and Significance - Caution



Regression Statistics	
Multiple R	0.492914799
R Square	0.242964999
Adjusted R Square	0.242206447
Standard Error	1.016805138
Observations	1000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	331.1568641	331.1568641	320.3010019	2.43789E-62
Residual	998	1031.824904	1.033892689		
Total	999	1362.981768			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.006235417	0.061969128	80.78595852	0	4.88463068	5.127840154
x	0.09979782	0.005576246	17.8969551	2.43789E-62	0.088855309	0.110740332

- R-Sq suggests that 24% of variation in *y* can be explained by variation in *x*.
- *t* and *F* tests show that coefficient is significant and null hypothesis should be rejected.
- The 95% confidence interval of the slope, $b_1 \pm t_{crit} * s_b$, is (0.089,0.111).
- *Statistical significance* doesn't necessarily mean *practical significance*.

R-Squared, Significance and Residuals - Caution

Excel Activities

American Automobile Association (AAA) publishes data that looks at the relationship between average stopping distance and the speed of car.

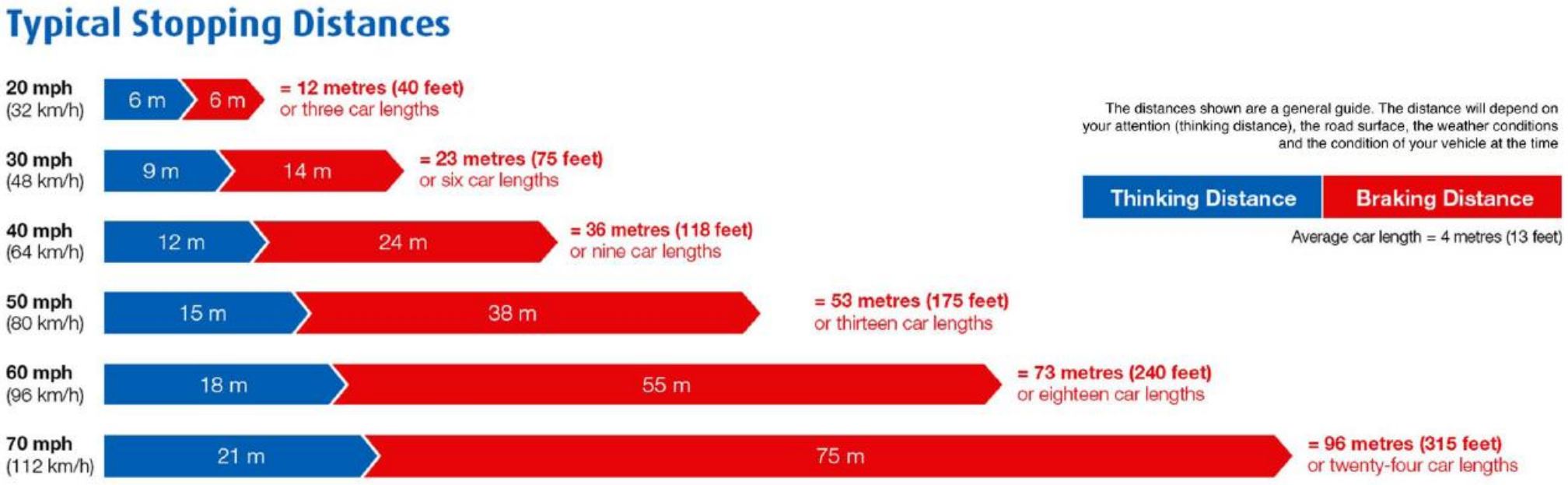
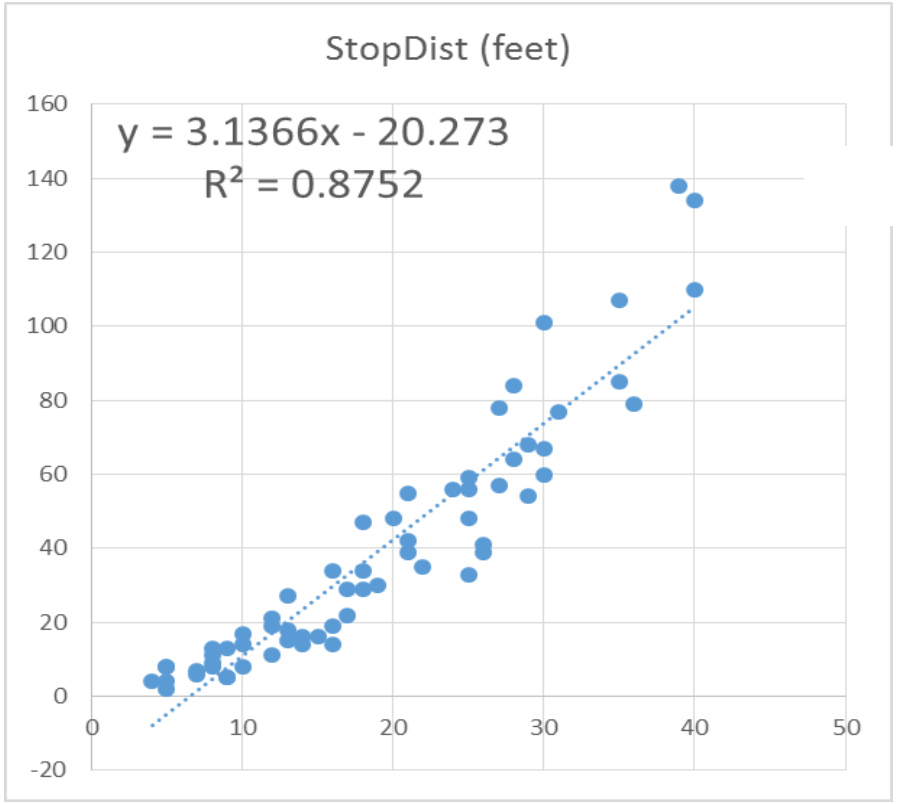


Image Source: http://streets.mn/2015/04/02/the-critical-ten/
Last accessed: November 20, 2015

R-Squared, Significance and Residuals - Caution

American Automobile Association (AAA) publishes data that looks at the relationship between average stopping distance and the speed of car.

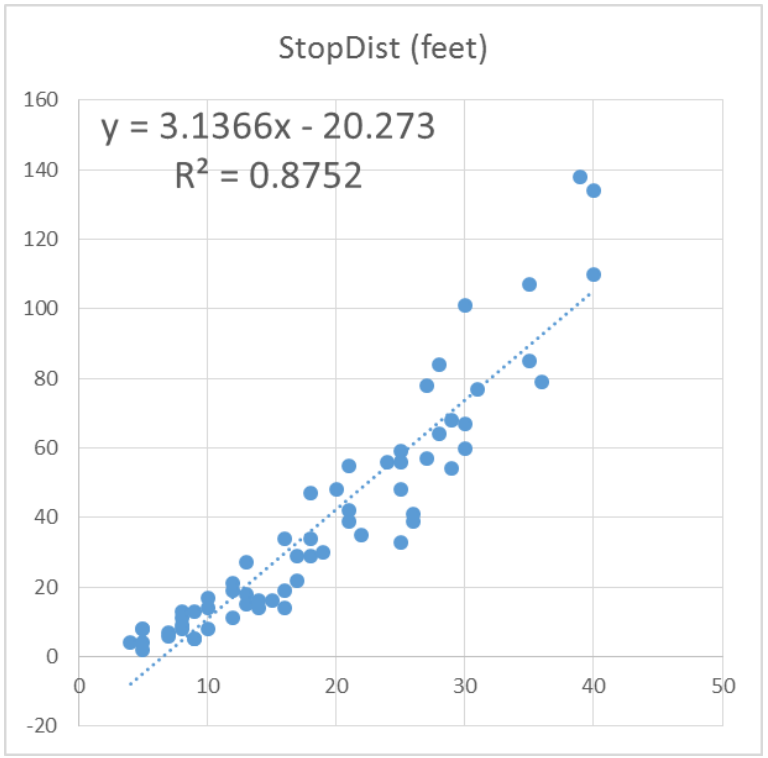
Does the estimated regression line fit the data well?

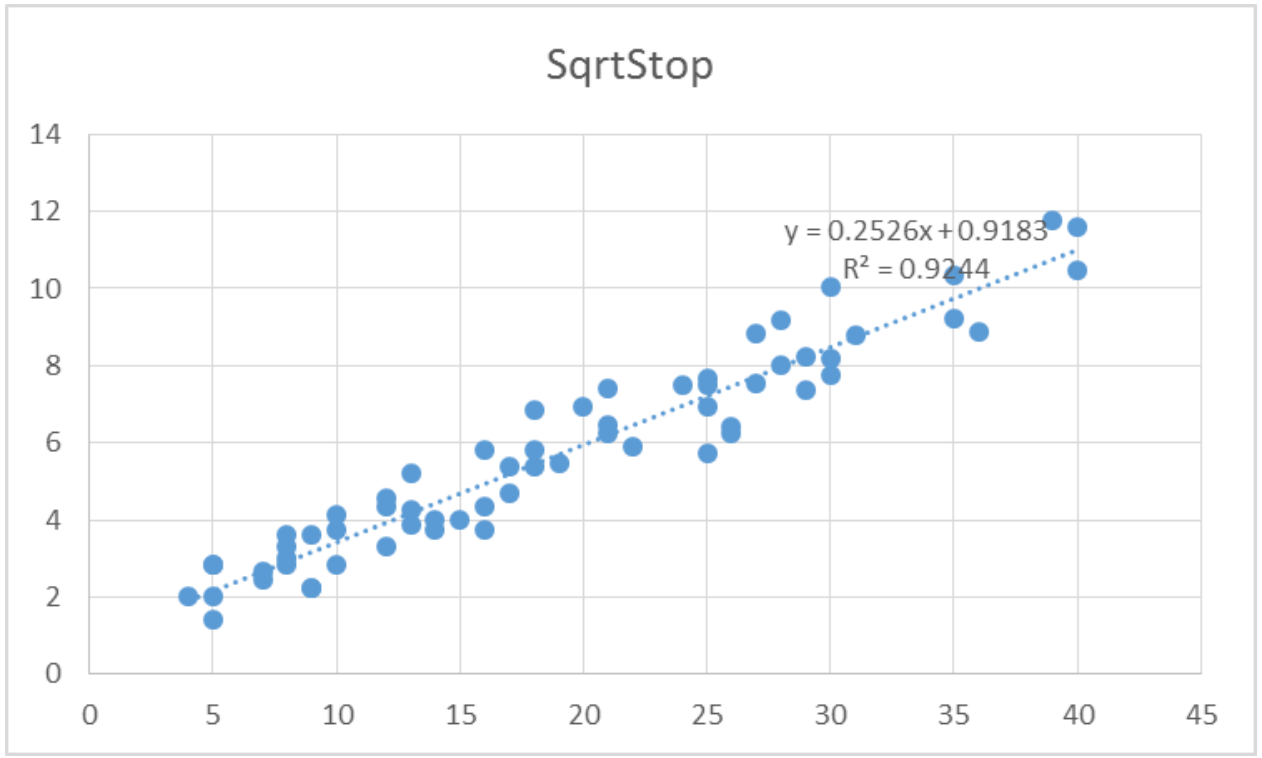


R-Squared, Significance and Residuals - Caution

American Automobile Association (AAA) publishes data that looks at the relationship between average stopping distance and the speed of car.

A large R-Sq does not imply that the estimated regression line fits the data best.



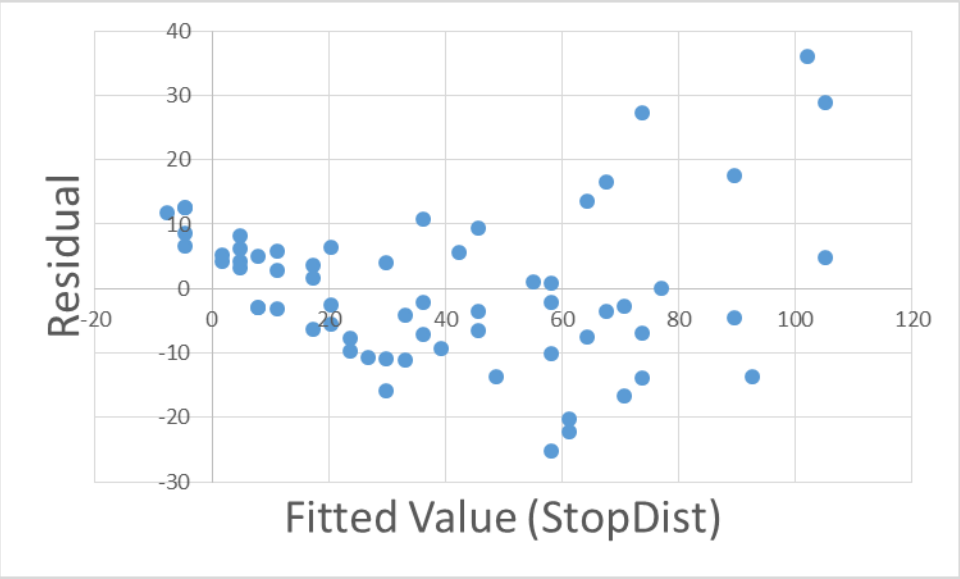


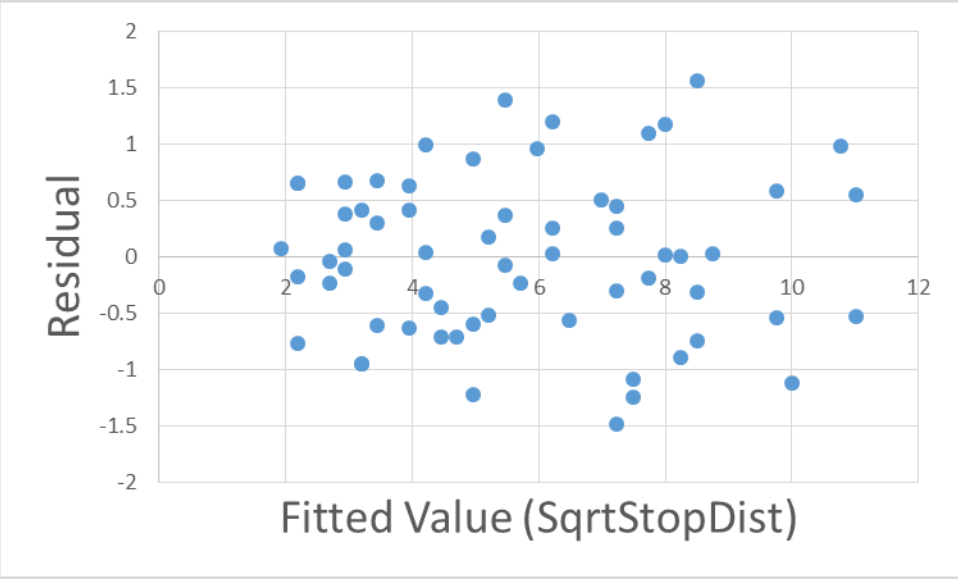
R-Squared, Significance and Residuals - Caution

American Automobile Association (AAA) publishes data that looks at the relationship between average stopping distance and the speed of car.

A large R-Sq does not imply that the estimated regression line fits the data best.

Check the residuals.



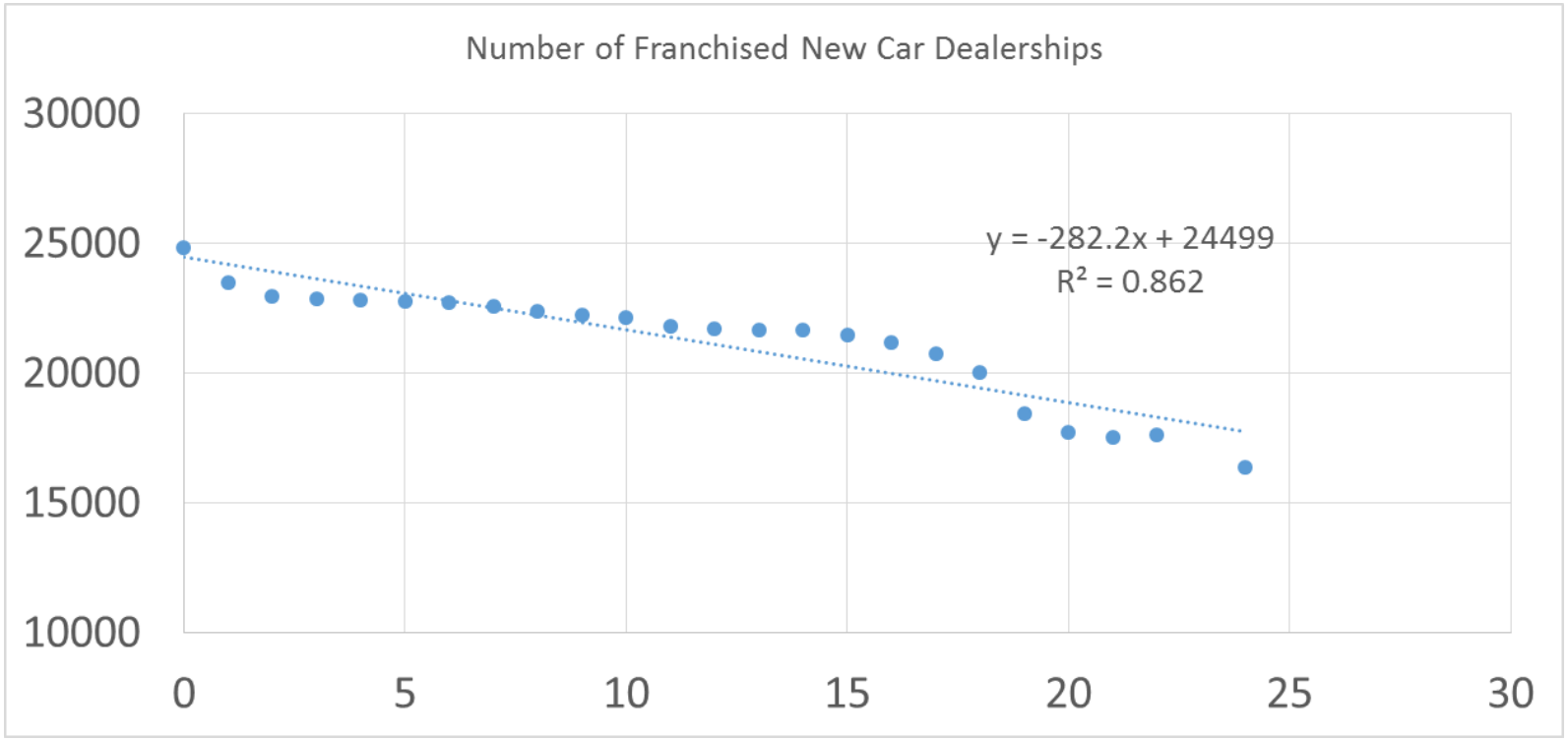


R-Squared, Significance and Residuals - Caution

National Automotive Dealers Association (NADA) of US publishes state-of-the-industry report each year. You want to know if there is any linear relationship between the time since 1990 and the number of franchised new car dealerships.

EXCEL ACTIVITY

R-Squared, Significance and Residuals - Caution



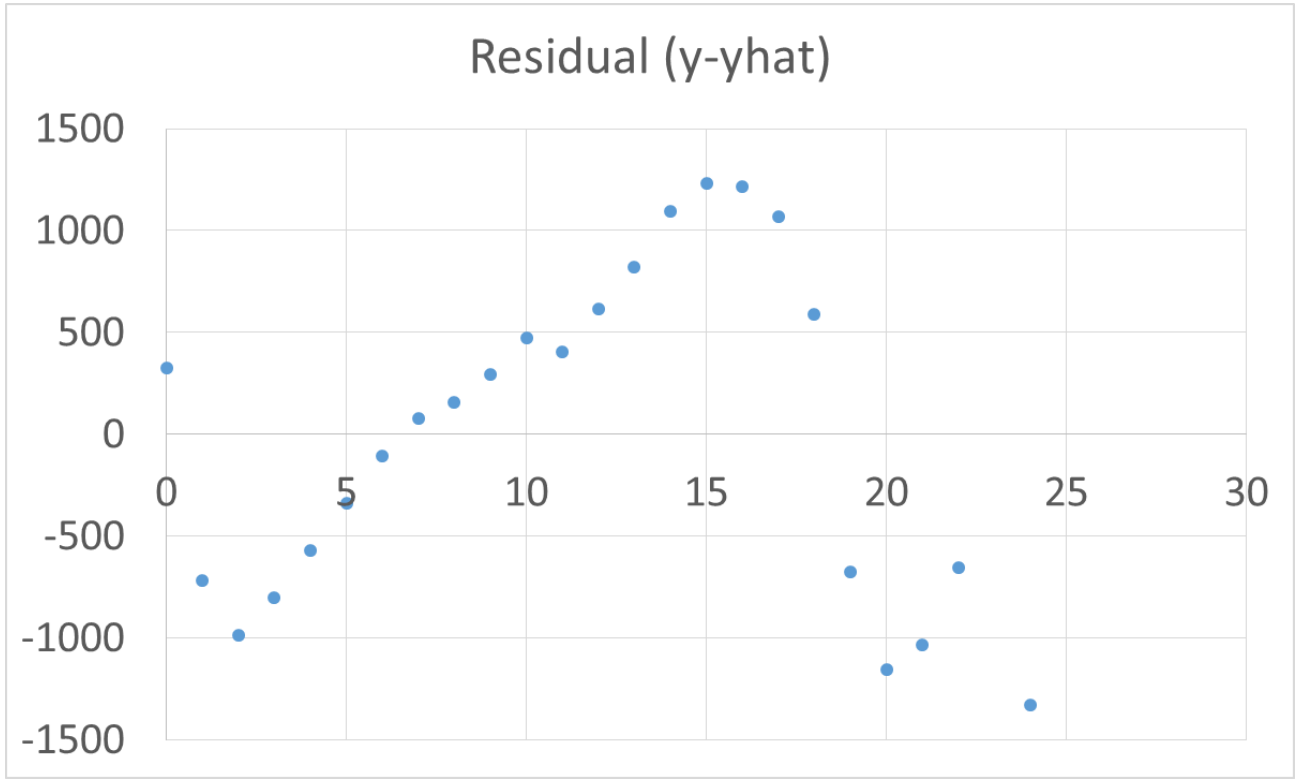
- Based on the shape of the scatter plot, do you think a linear fit looks good?
- Does R^2 imply a good fit?
- What can you infer from the intercept and the slope?

R-Squared, Significance and Residuals - Caution

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.928448566							
R Square	0.862016739							
Adjusted R Square	0.855744773							
Standard Error	824.748263							
Observations	24							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	93487768.66	93487768.66	137.4396293	6.21261E-11			
Residual	22	14964613.34	680209.6973					
Total	23	108452382						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	24498.51368	324.8477406	75.41537349	4.68438E-28	23824.8207	25172.20666	23582.84714	25414.18022
Time Since 1990 (in years)	-282.1961313	24.07105183	-11.7234649	6.21261E-11	-332.1164374	-232.2758252	-350.0465546	-214.3457081

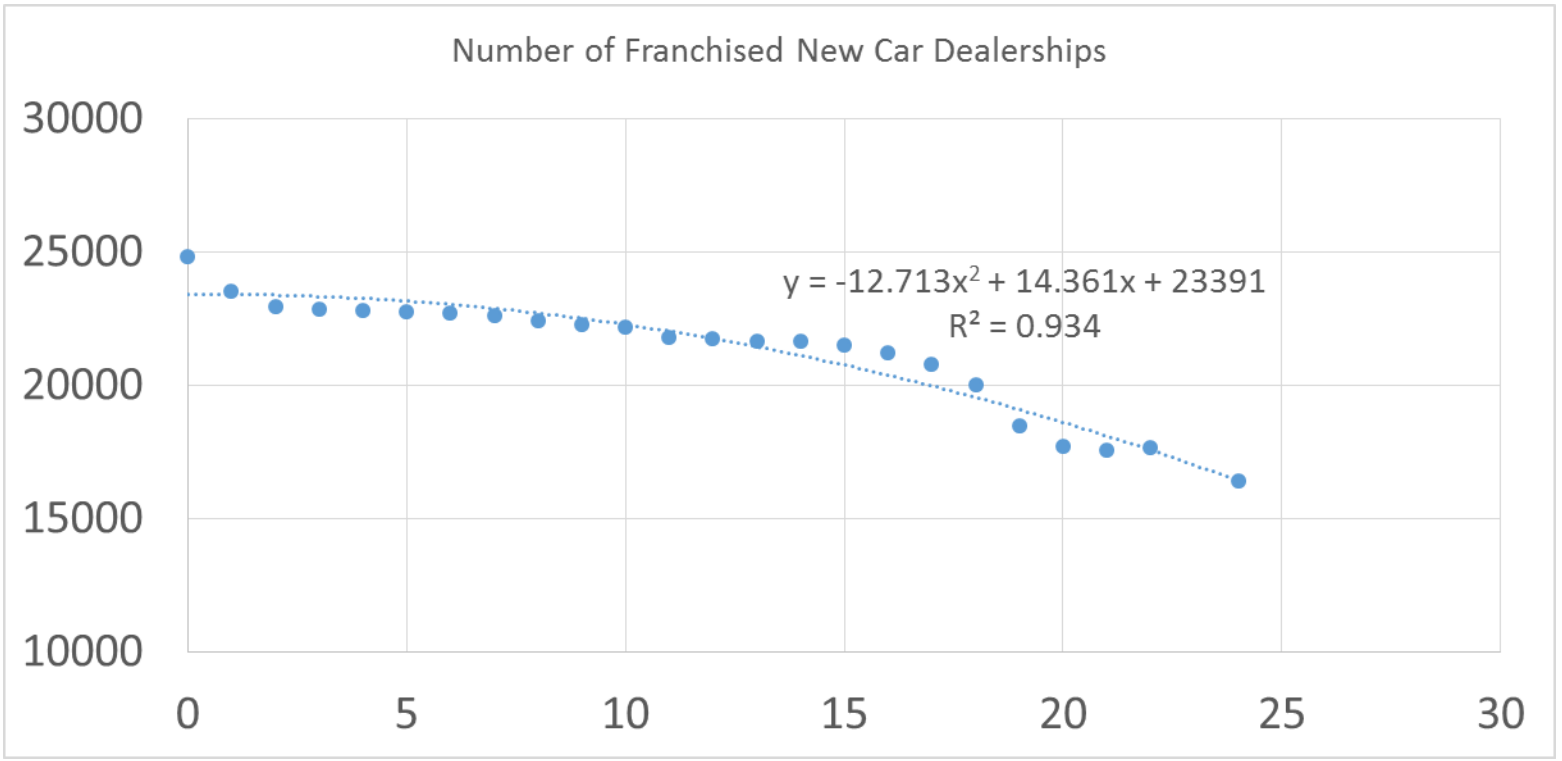
- Is the slope significant?
- Is the model significant?

R-Squared, Significance and Residuals - Caution



- Based on the residual plot, do you think a linear model is a good fit?

R-Squared, Significance and Residuals - Caution



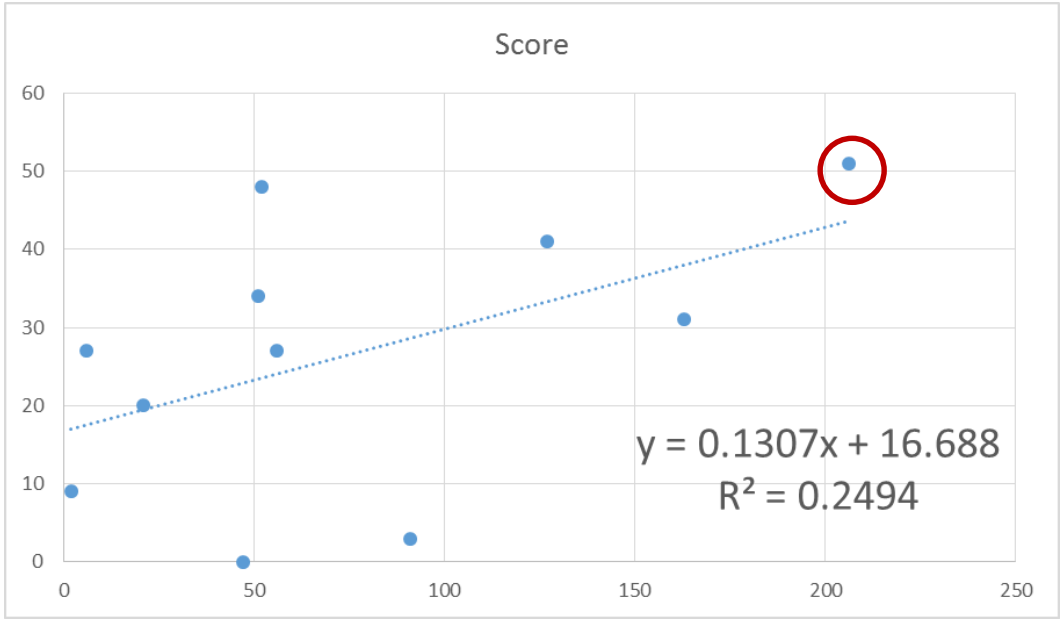
NOTE

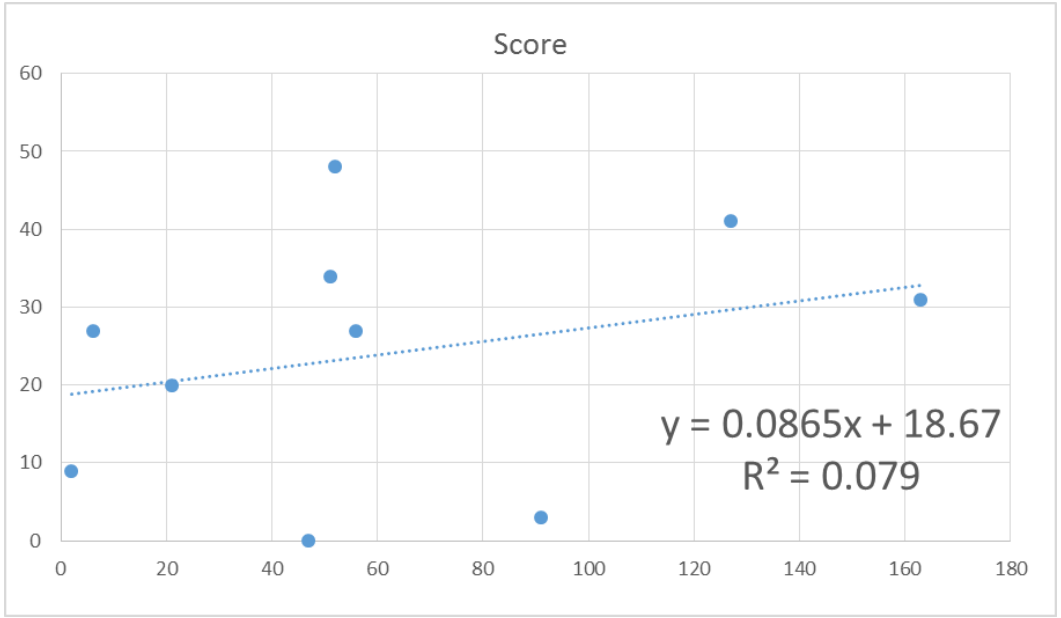
Since the data here are ordered (time is a factor), the residual plot also indicates violation of the *independence* assumption. Time series analysis becomes the right approach for this dataset instead of OLS Regression.

R-Squared, Significance and Residuals - Caution

Why it is important to plot.

1998 Penn State Football season – Eric McCoo's rushing yards vs the final score.





The last data point is *influencing* the regression line significantly.

Influential Observations

An observation which, when not included, greatly alters the predicted scores of other observations.

Cook's D is a measure of the influence and is proportional to the sum of the squared differences between predictions made with all observations in the analysis and predictions made leaving out the observation in question.

Influence is a function of **leverage** and **distance (or 'residuality' or 'outlierness')**.

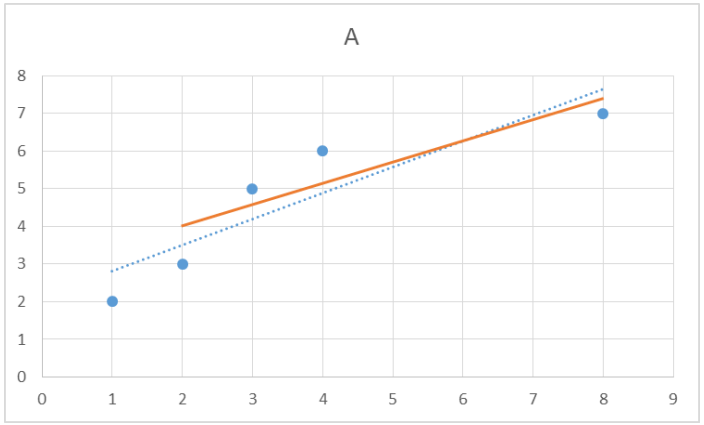
Influential Observations

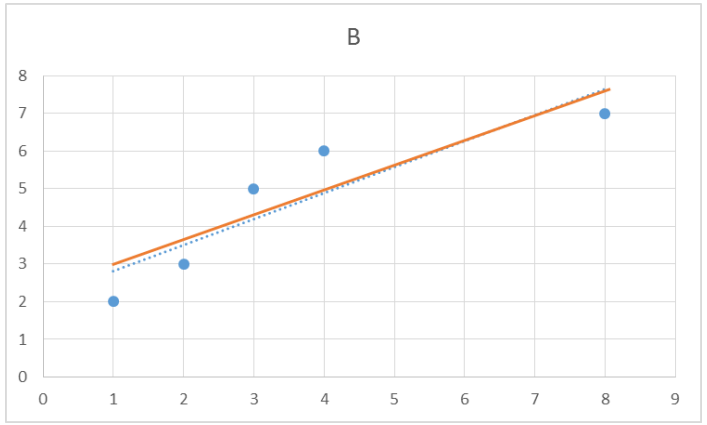
ID	X	Y	h	R	D
A	1	2	0.39	-1.02	0.4
B	2	3	0.27	-0.56	0.06
C	3	5	0.21	0.89	0.11
D	4	6	0.2	1.22	0.19
E	8	7	0.73	-1.68	8.86

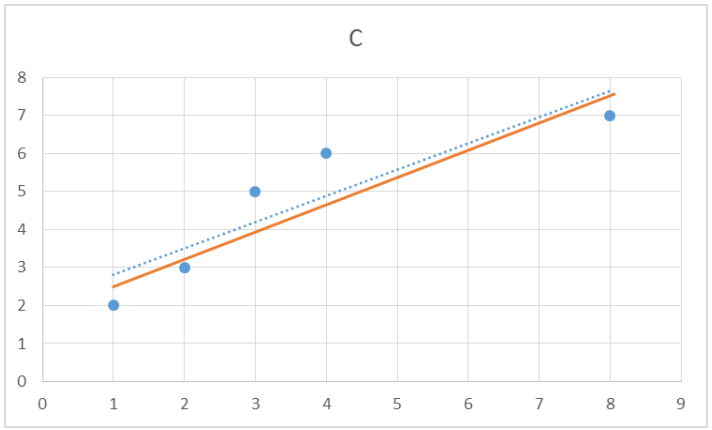
h is the leverage, R is the studentized residual, and D is Cook's measure of influence.

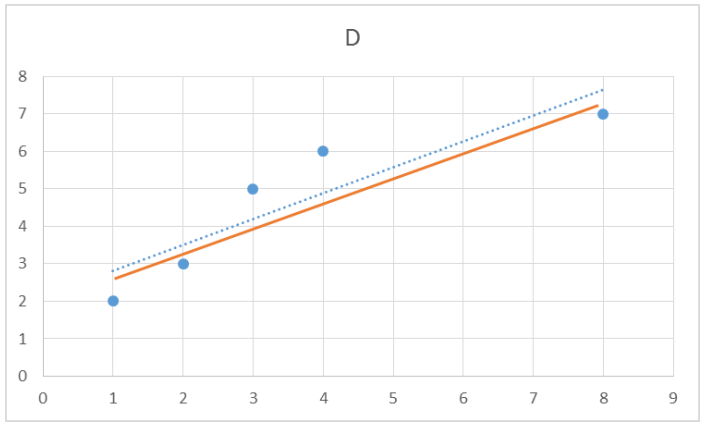
Source: http://onlinestatbook.com/2/regression/influential.html
Last accessed: June 30, 2017

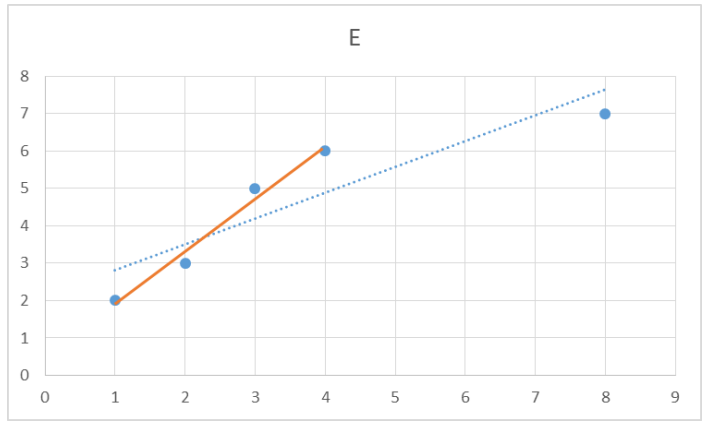
Influential Observations







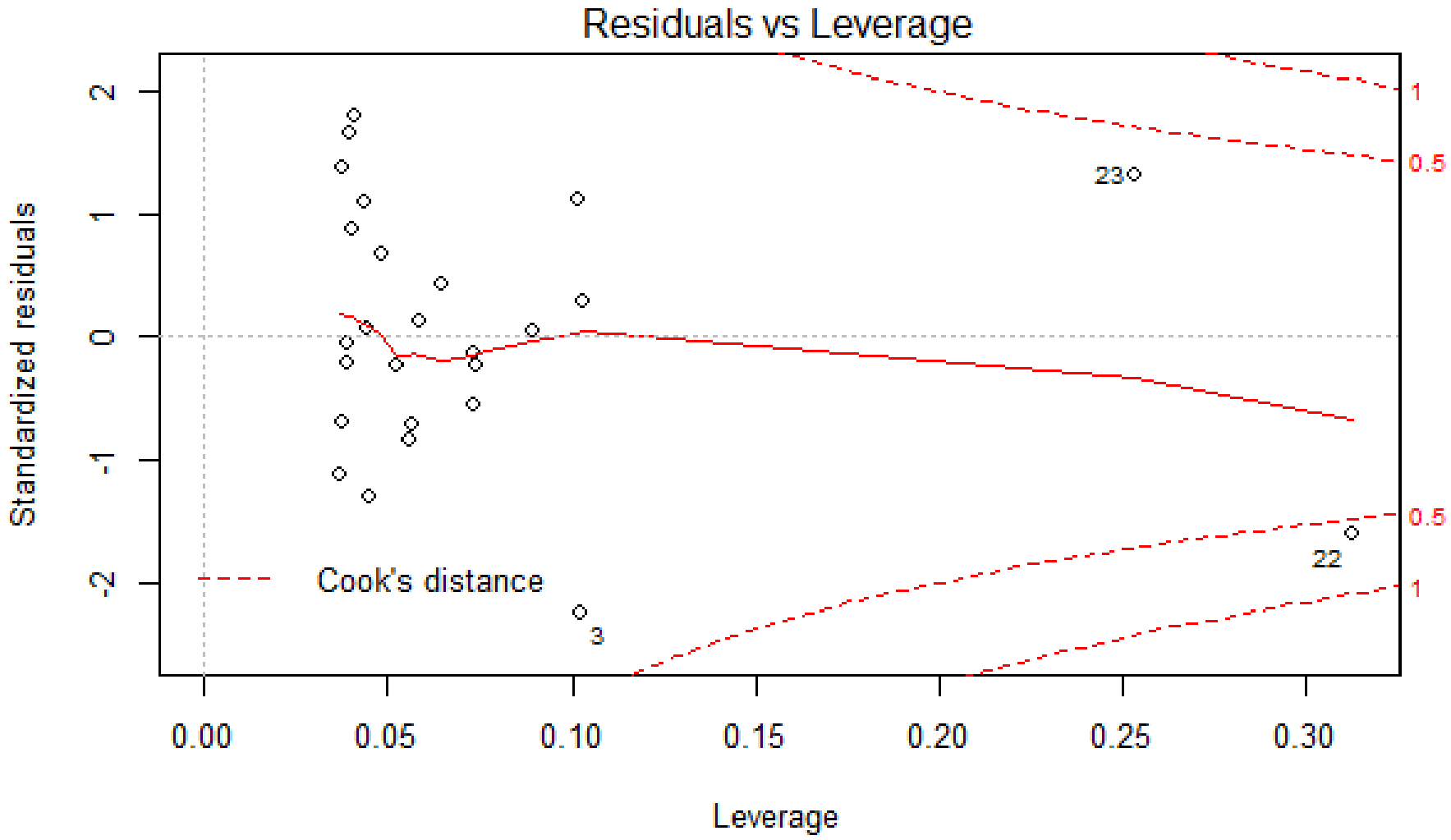




CSE 73026

Influential Observations

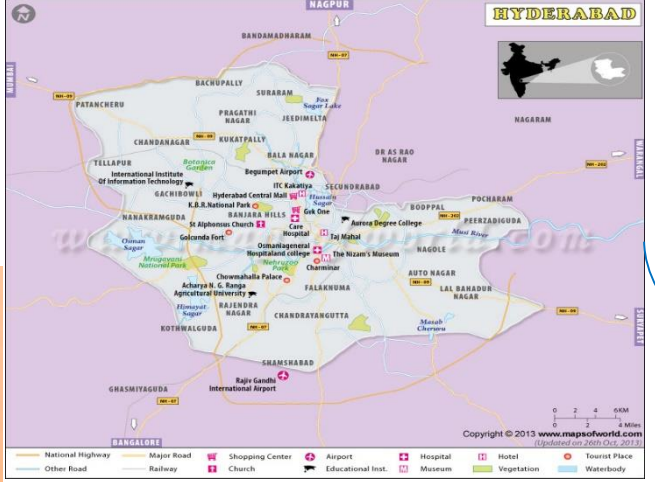
Are there influential outliers?



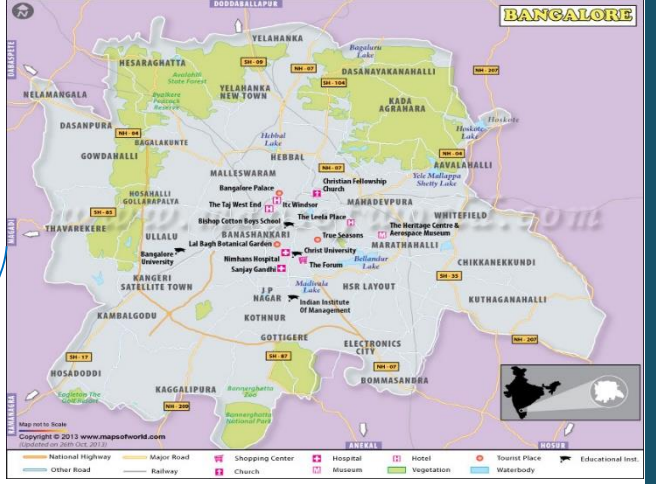
Influential Observations – Rules of Thumb

- If Cook's D of any observation (D_i) > 1, that observation can be considered as having too much influence, but investigate values greater than 0.5 also.

- ***Relative size interpretation*: In general, investigate any value that is very different from the rest.**



INSOFE
Inspire...Educate...Transform.



HYDERABAD

2nd Floor, Jyothi Imperial, Vamsiram Builders, Old Mumbai Highway, Gachibowli, Hyderabad - 500 032
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

BENGALURU

Floors 1-3, L77, 15th Cross Road, 3A Main Road, Sector 6, HSR Layout, Bengaluru – 560 102
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Social Media

Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.