INSOFE

Inspire...Educate...Transform.






Inspire…Educate…Transform.

Statistics and Probability in Decision Modeling

Linear Regression

Dr. Venkatesh Sunkad

Dec 30, 2018

MATERIAL CONTENT FROM Dr. SRIDHAR PAPPU

Analyzing relationships between attributes

CORRELATION, COVARIANCE AND R-SQUARED

CSE 7302c

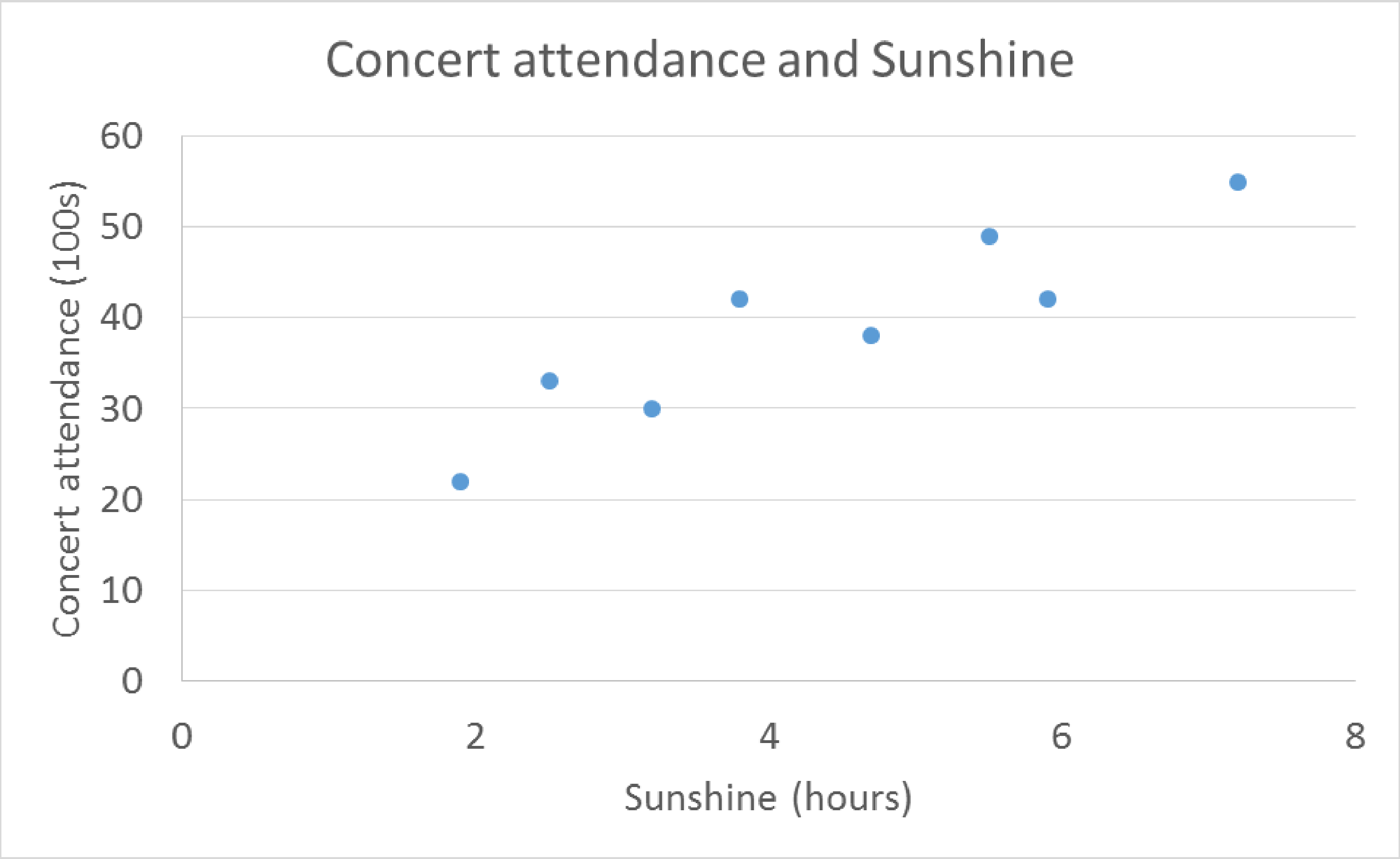
Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

- 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?

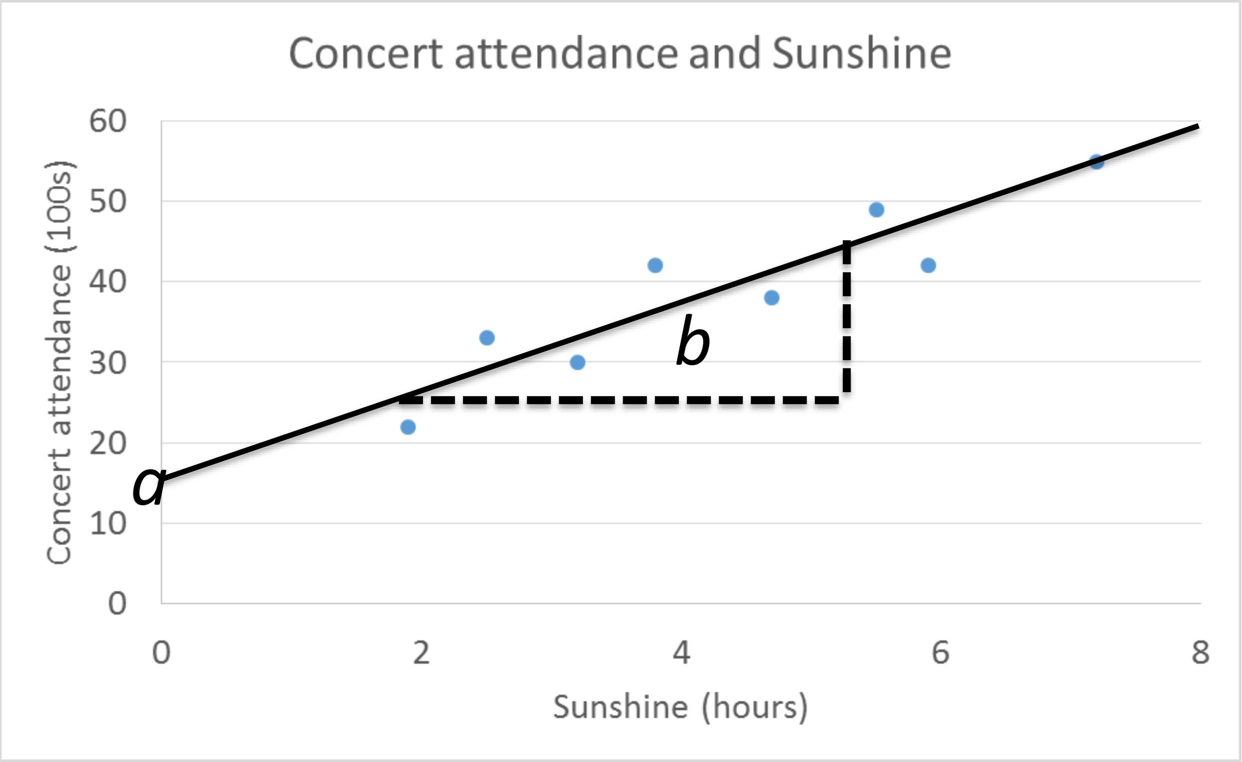
Image Source: http://blurtonline.com/wp-content/uploads/2013/06/Shaky-Knees-1514.jpeg;
Last accessed: May 1, 2014

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



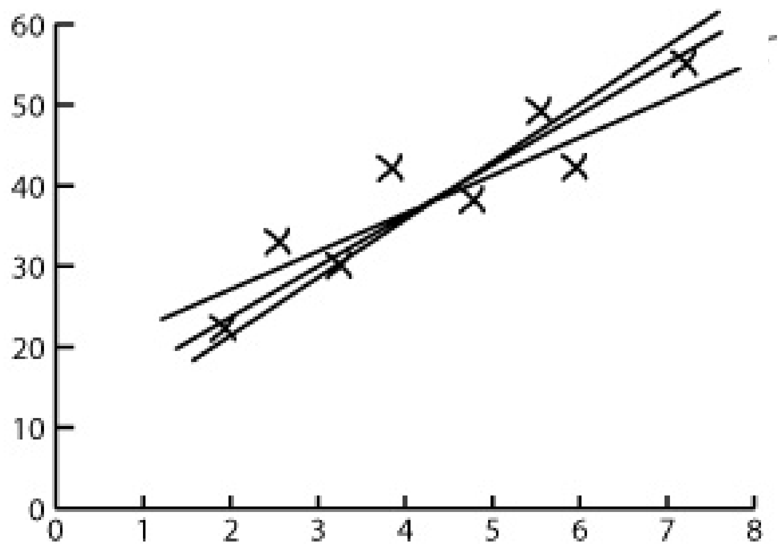
We need to find the equation of the line.

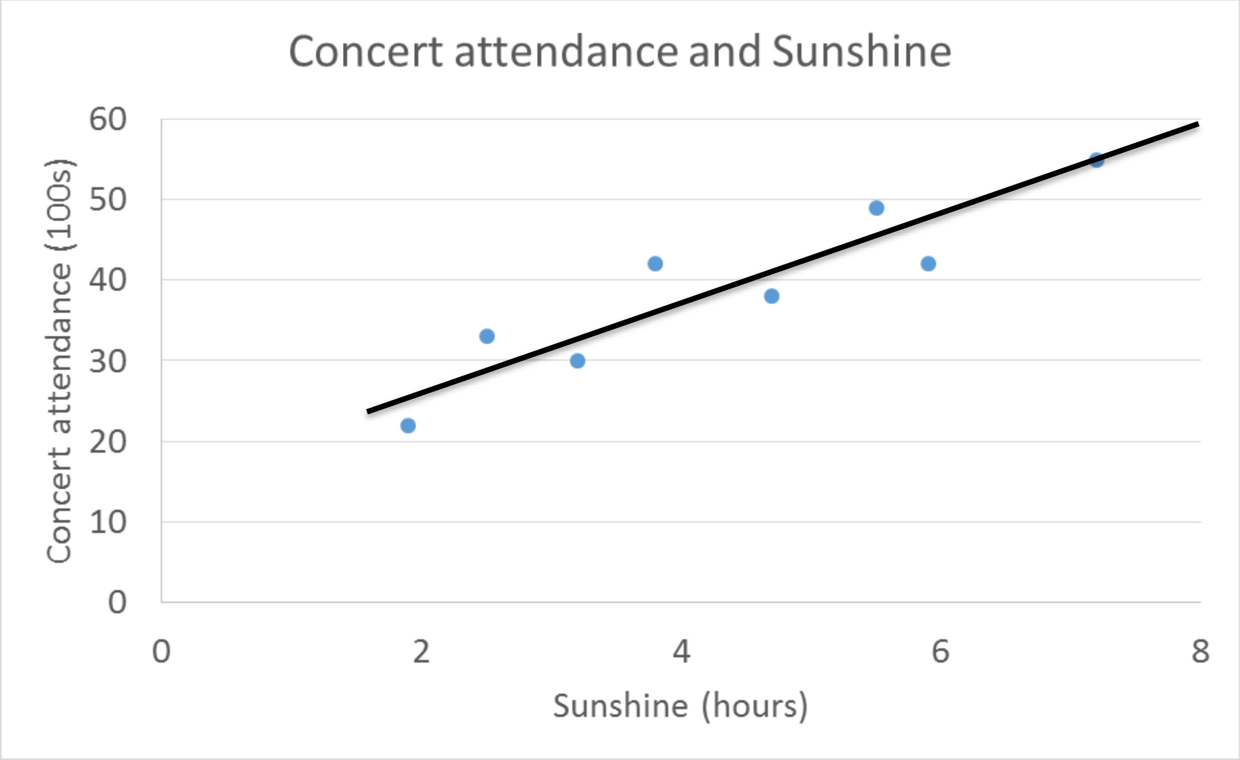


$y = a + bx$

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

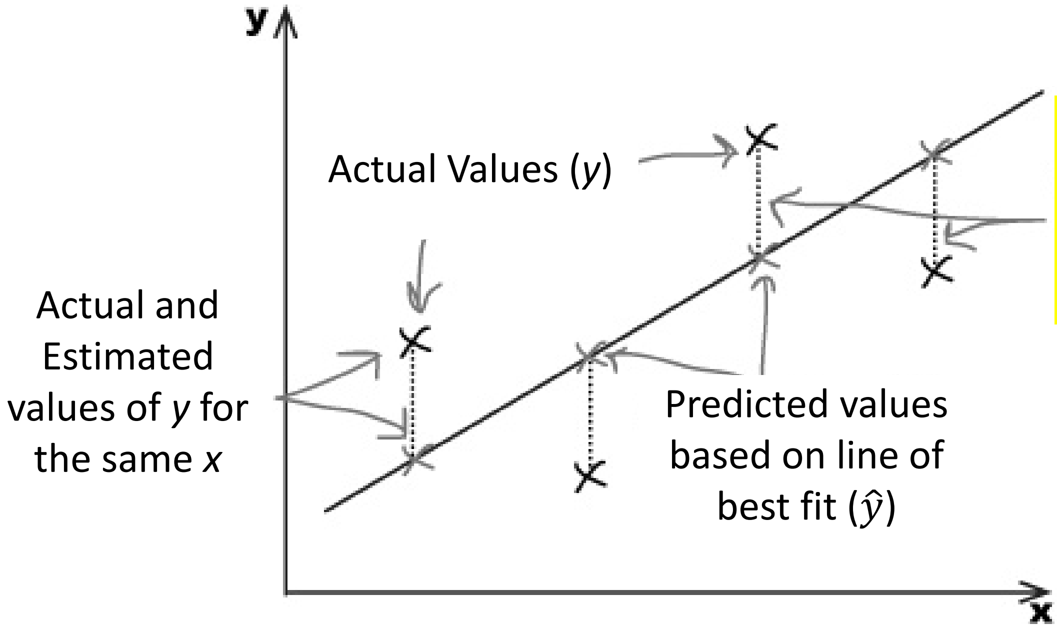
- Line of best fit





CSE 7302c

We need to minimize errors.



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

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

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

Where $\bar{x}$ and $\bar{y}$ are the means of x and y.

CSE 7302c

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

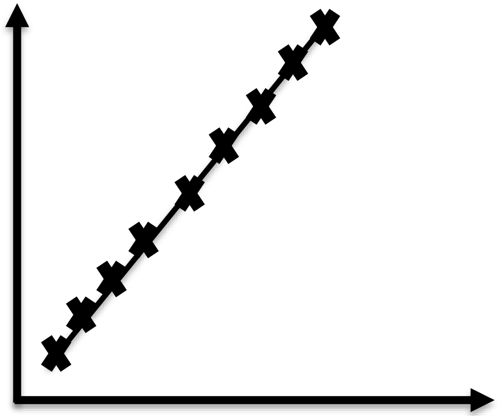
How do you calculate a in $\widehat{y}_i = a + bx$?

The line of best fit must pass through the average of the data.

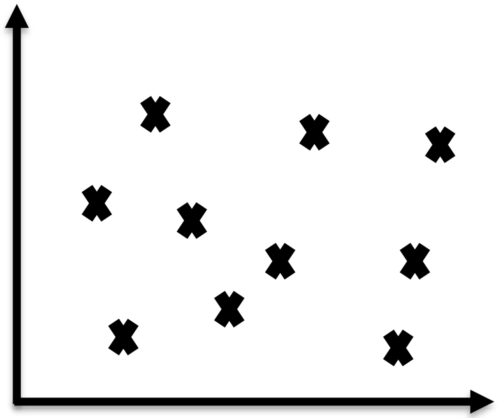
The line of best fit must pass through $(\bar{x}, \bar{y})$. Substituting in the equation $\bar{y} = a + b\bar{x}$, we can find a.

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

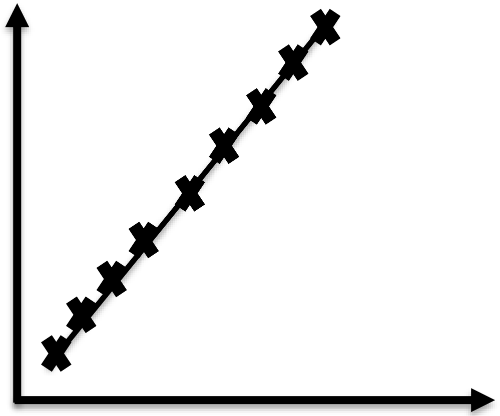


No Linear Correlation

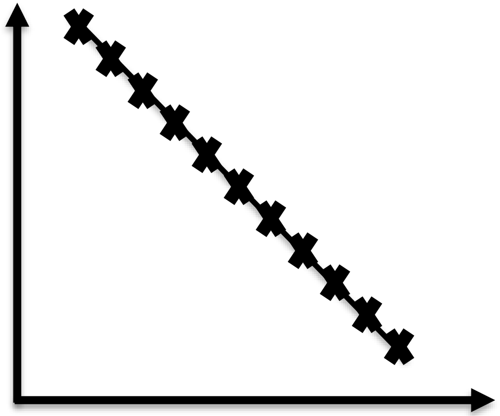
The fit of the line is given by **correlation coefficient**.

Correlation Coefficient

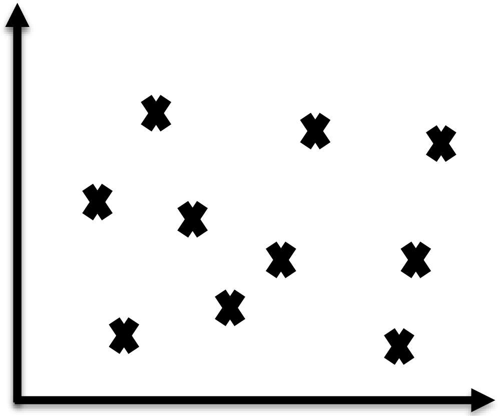
Correlation coefficient, *r*, is a number between -1 and 1 and tells us how well a regression line fits the data.



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

$$\textbf{Correlation Coefficient.}\ \boldsymbol{is\ represented\ as\ “r”\ \ and}$$

$$\boldsymbol{r = \frac{b s_x}{s_y}}$$

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.

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

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

Correlation Coefficient for our data is r = 0.916

Correlation Coefficient and Covariance – Excel*["Covariance Correlation" and Covariance Comparison"

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

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

$$b = \frac{(\sum((x-\bar{x})(y-\bar{y})))/(n-1)}{(\sum(x-\bar{x})^2)/(n-1)}$$

$$\boldsymbol{b = \frac{s_{xy}^2}{s_x{}^2}}$$

$$r = \frac{b s_x}{s_y}$$

$$r = \frac{s^2{}_{xy}}{s_x{}^2} \cdot \frac{s_x}{s_y}$$

$$\boldsymbol{r = \frac{s_{xy}^2}{s_x s_y}}$$

* 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

CSE 7302c

Correlation Coefficient and Covariance

$$\boldsymbol{b} = \frac{\boldsymbol{s}_{\boldsymbol{xy}}^{2}}{\boldsymbol{s}_{\boldsymbol{x}}{}^{2}} \textbf{ and so, } \boldsymbol{r} = \frac{\boldsymbol{s}_{\boldsymbol{xy}}^{2}}{\boldsymbol{s}_{\boldsymbol{x}}\boldsymbol{s}_{\boldsymbol{y}}}$$

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

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.

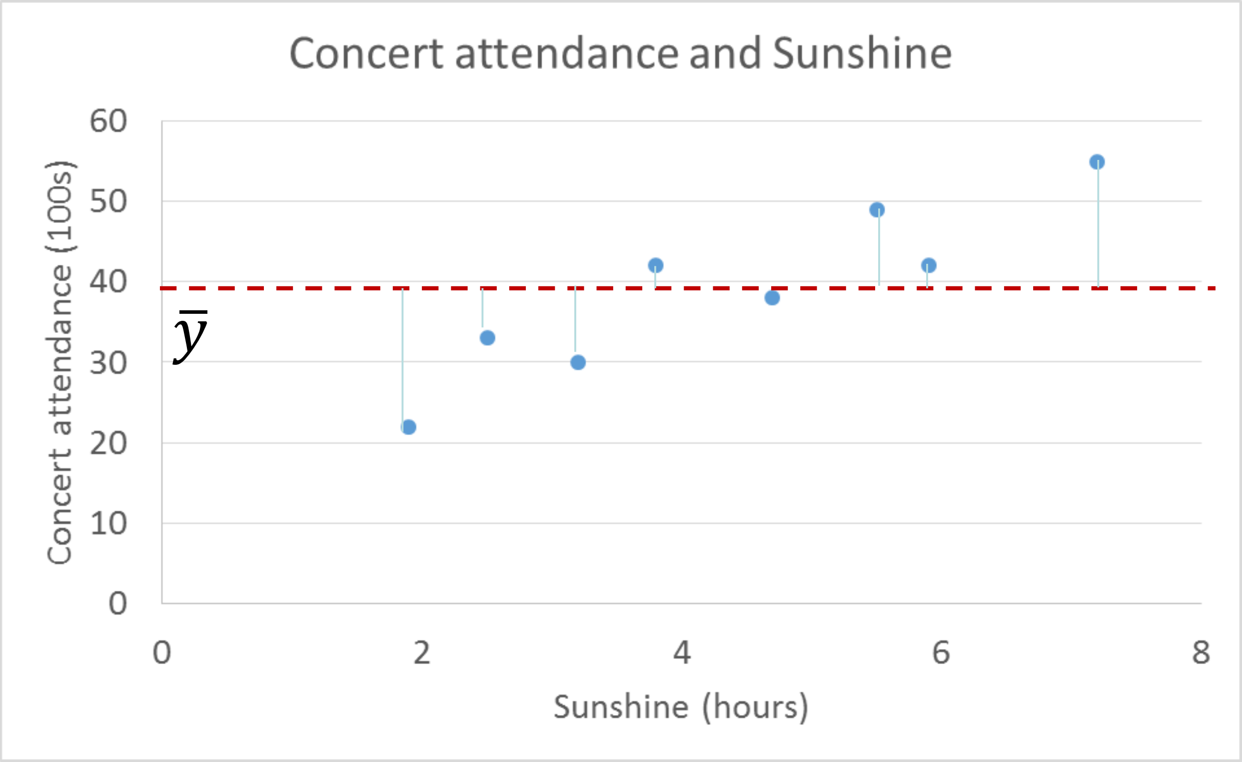
If $r^2 = 0$, it means you can't predict the *y* value from the *x* value.
If $r^2 = 1$, it means you can predict the *y* value from the *x* value without any errors.

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 <u>expected</u> concert attendance when sunshine is not considered. This "model" has **large** residuals.

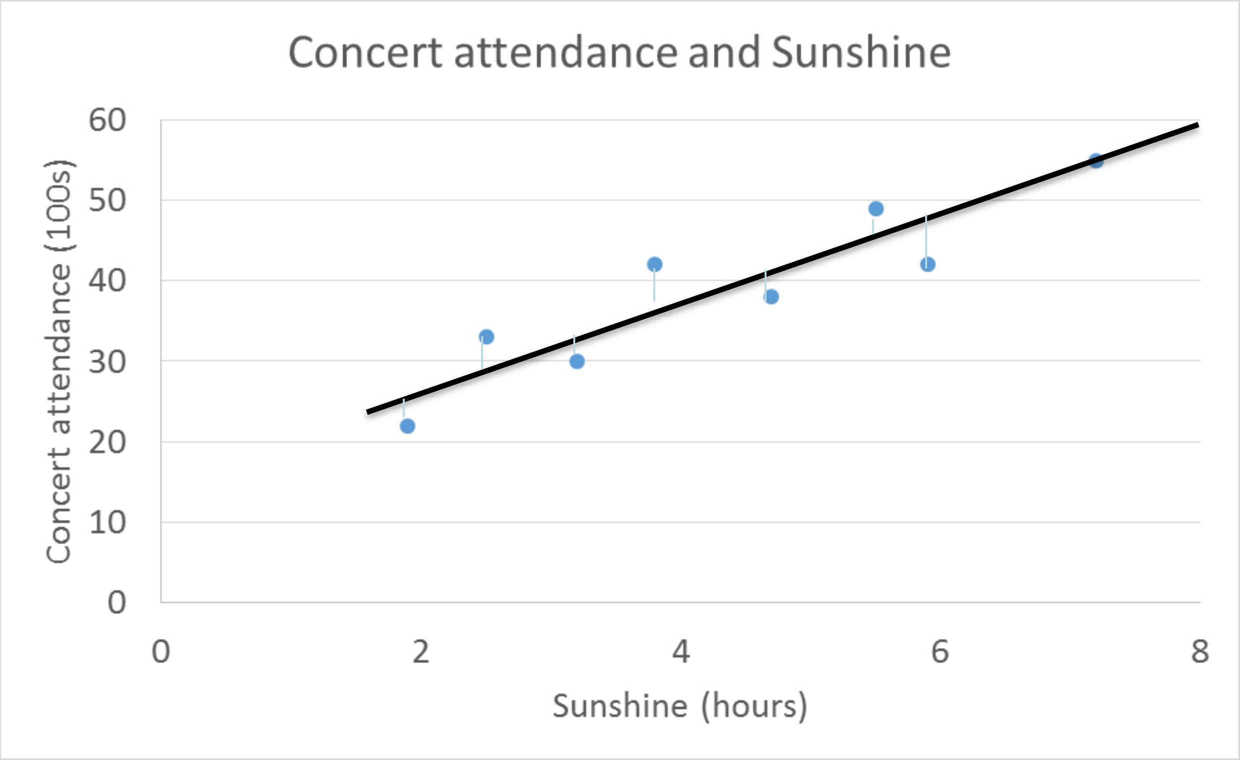
$$SST = \sum (y_i - \bar{y})^2$$



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.

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

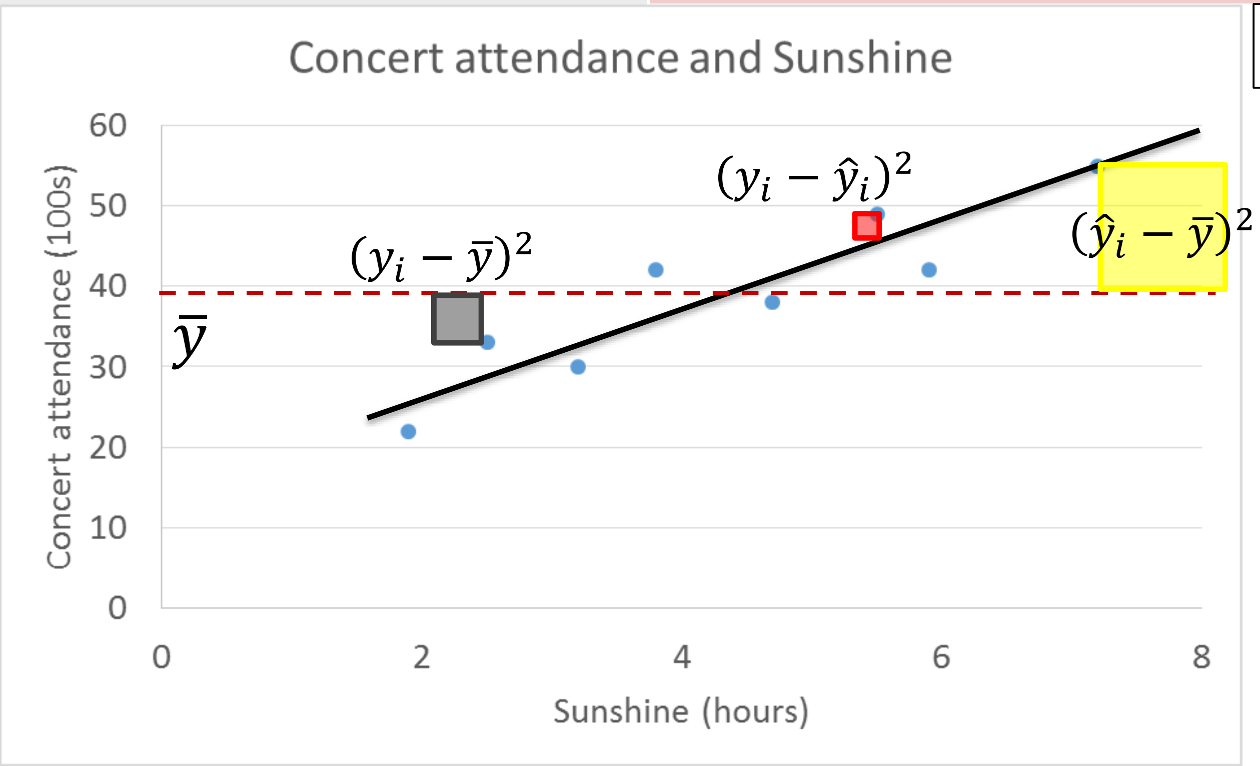


R-squared

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



$$SST = SSR + SSE$$

R-squared

$$SST = SSR + SSE$$

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

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

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

Dividing by SST we get

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

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

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

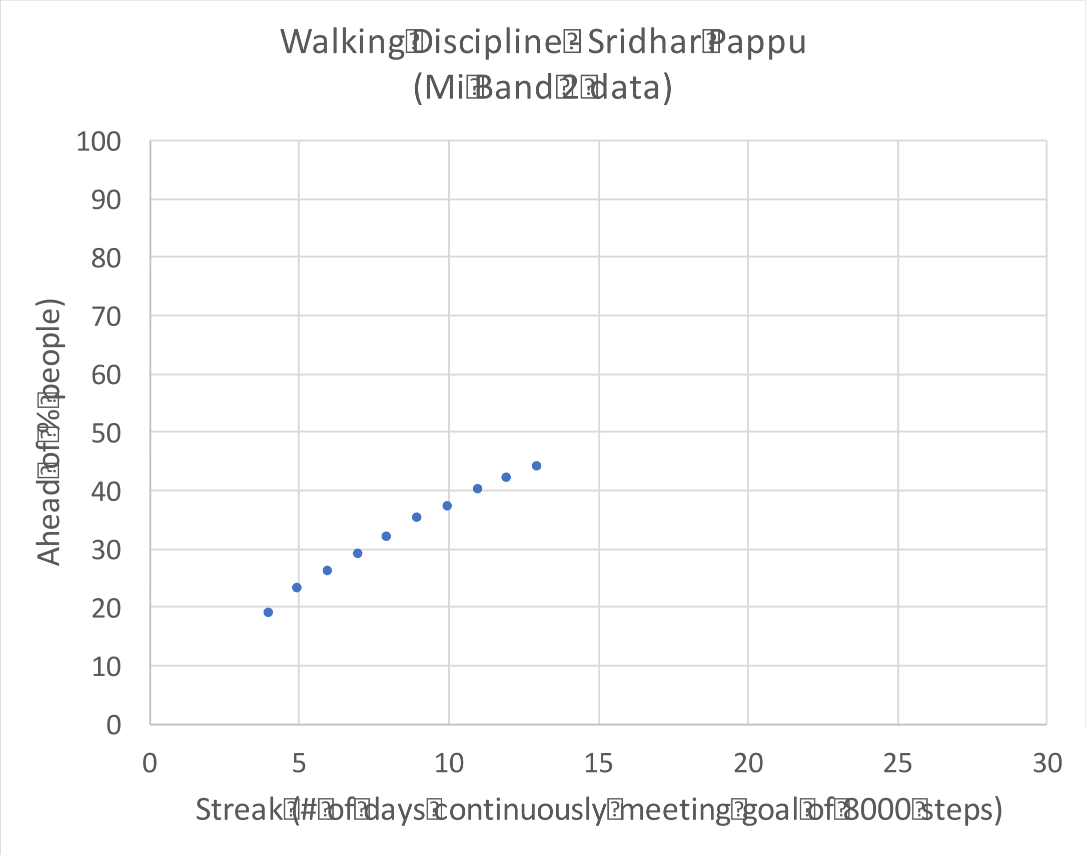
- Linear regression: A **regression** model where class/dependent/target variable is **numeric**

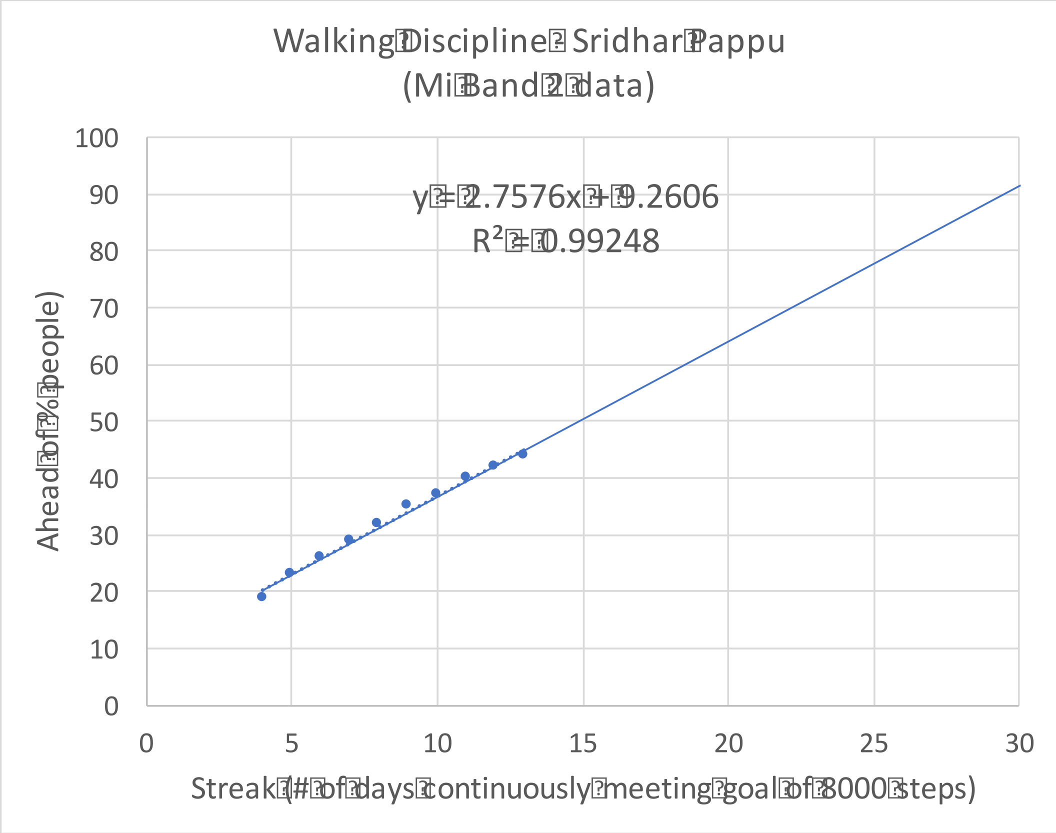
- Logistic regression: A **classification** model where class/dependent/target variable is **categorical**

CSE 7302c

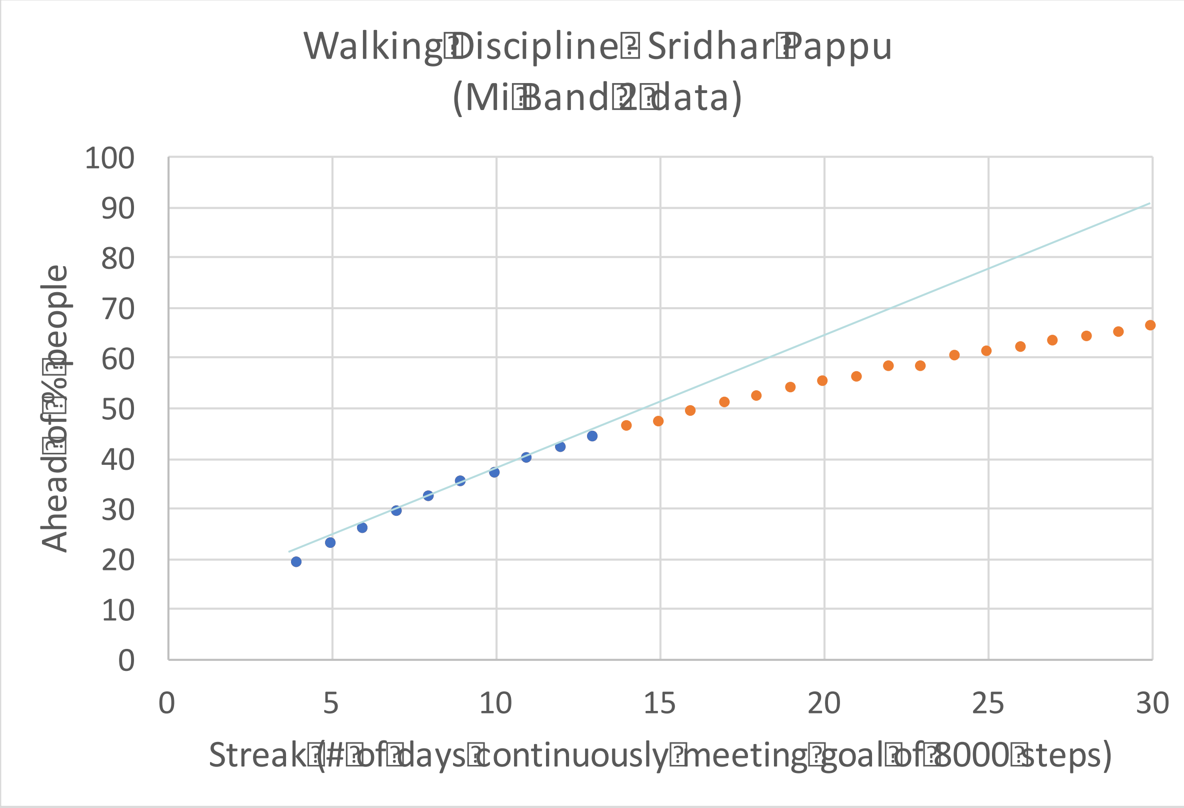
Linear Regression

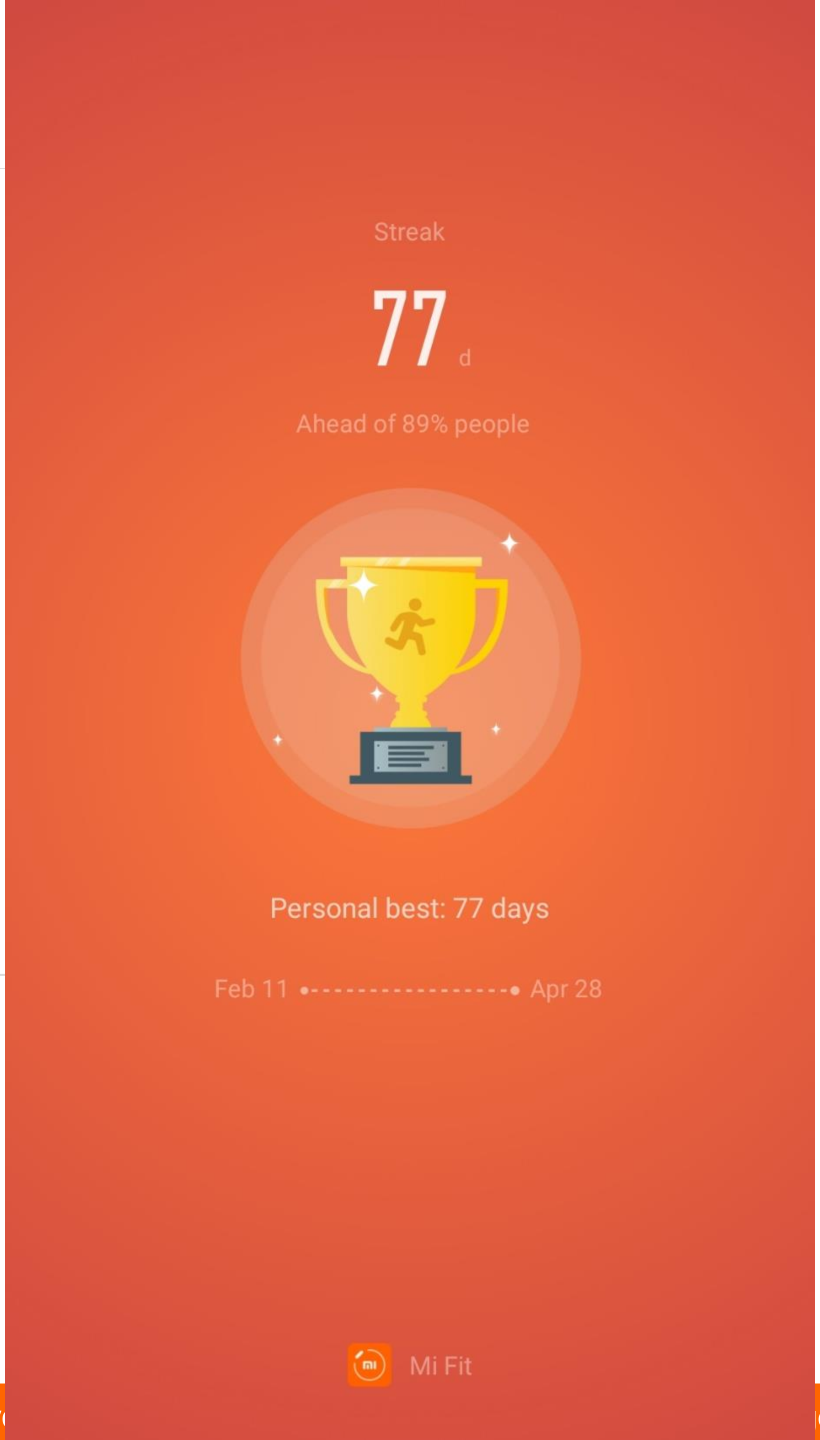
Linear Regression





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.

CSE 7302c

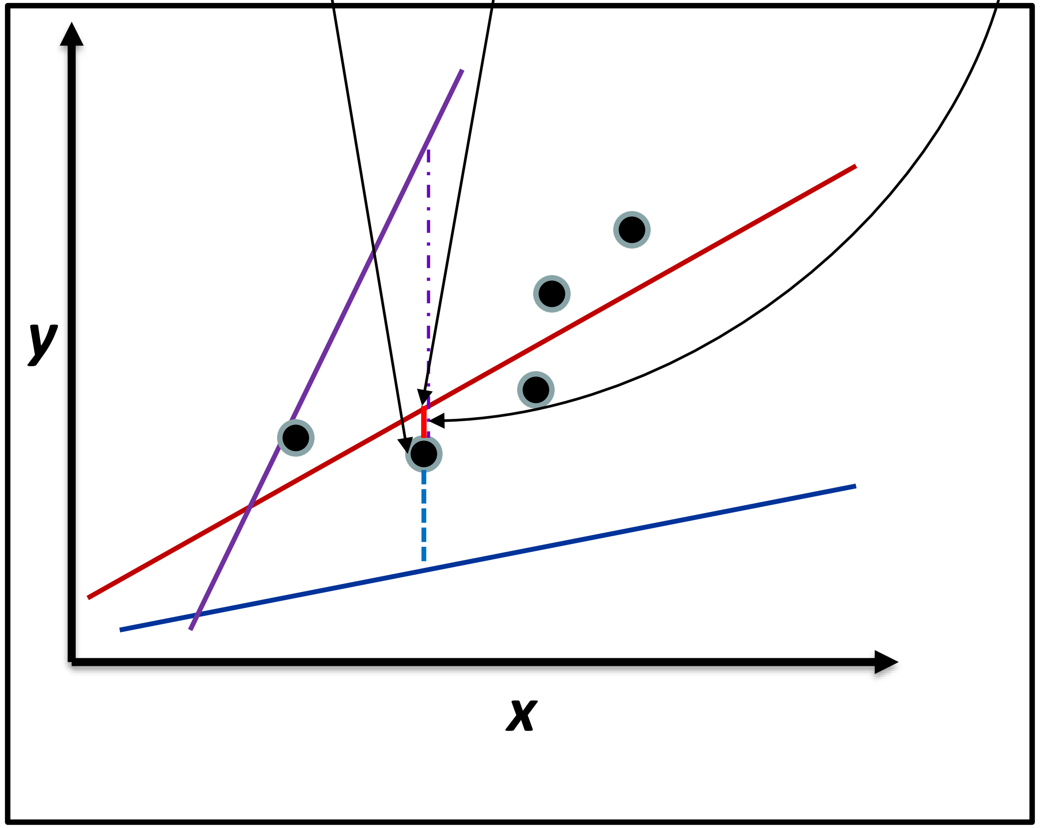
How to Pick the Best Model?

BREAK

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

30

India*	$1.62

23

Ukraine	$2.11
Hong Kong	$2.12

21

Malaysia	$2.34

20

China	$2.44
South Africa	$2.45
Indonesia	$2.46
Thailand	$2.46
Taiwan	$2.5

19

Russia	$2.55
Sri Lanka	$2.55
Egypt	$2.57
Poland	$2.58
Hungary	$2.63

18

Saudi Arabia	$2.67
Philippines	$2.68
Mexico	$2.7

17

Lithuania	$2.87
Pakistan	$2.89

16

Latvia	$3.0

15

South Korea	$3.19
UAE	$3.27

14

Czech Rep.	$3.45
Turkey	$3.54

13

Peru	$3.71
Singapore	$3.75
Britain	$3.82

12

Costa Rica	$4.02
Chile	$4.05
New Zealand	$4.05
Israel	$4.13
Japan	$4.16

11

USA	$4.2
Euro area	$4.43
Colombia	$4.54

10

Canada	$4.63
Uruguay	$4.63
Argentina	$4.64
Australia	$4.94

9

Denmark	$5.37

8

Brazil	$5.68
Sweden	$5.91

7

Norway	$6.79
Switzerland	$6.81

Source: The Economist (Jan 2012)

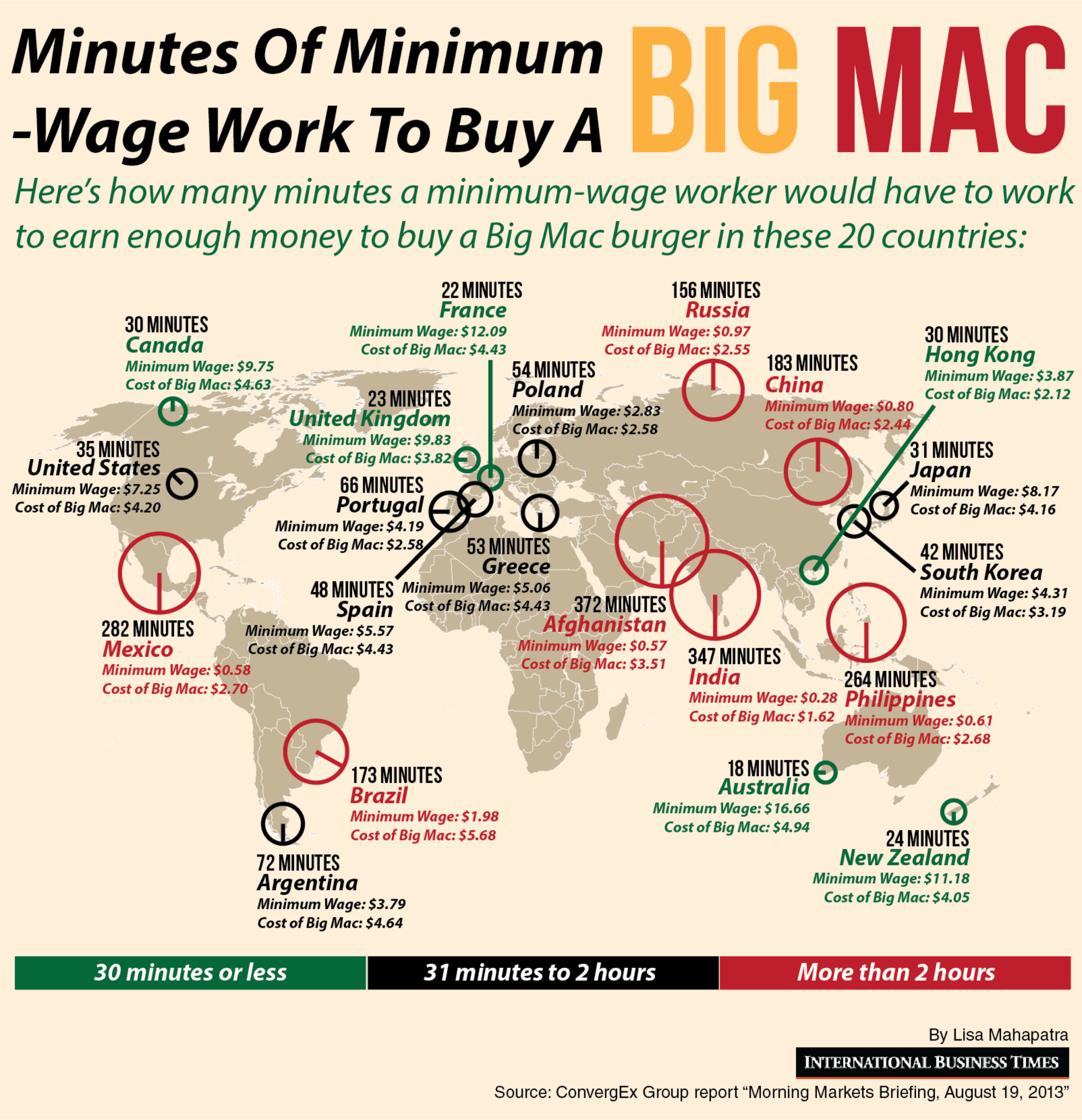
* Chicken burger

Burgernomics: Overvalued or Undervalued Currencies?

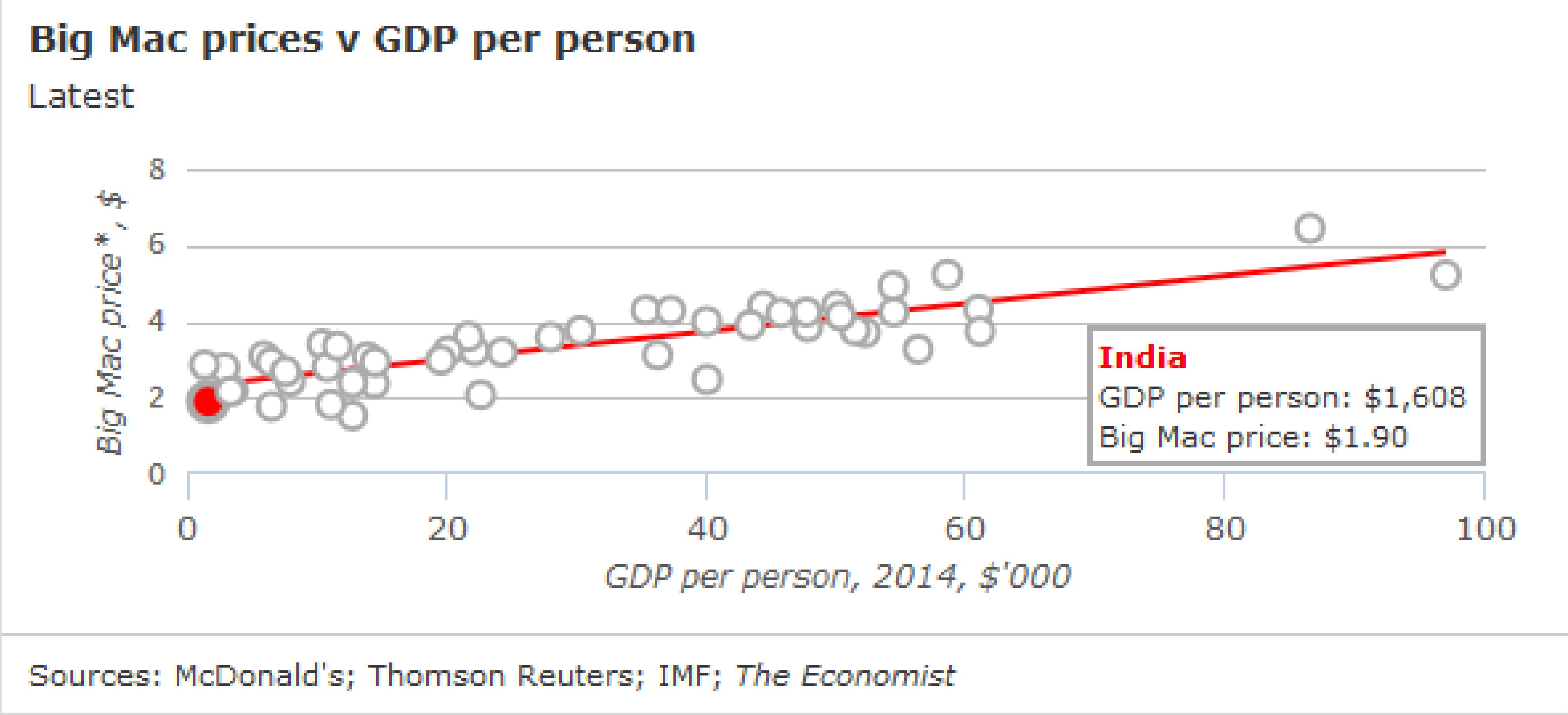
- Big Mac price in the US: $ 4.93
- Maharaja Mac price in India: Rs 155
- Implied PPP(Purchasing power Parity) is 155/4.93 = Rs 31.44/$
- Actual exchange rate = Rs 67.2959/$
- $\frac{31.44-67.2959}{67.2959} = -0.53$
- Rupee undervalued by 53% against the USD

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



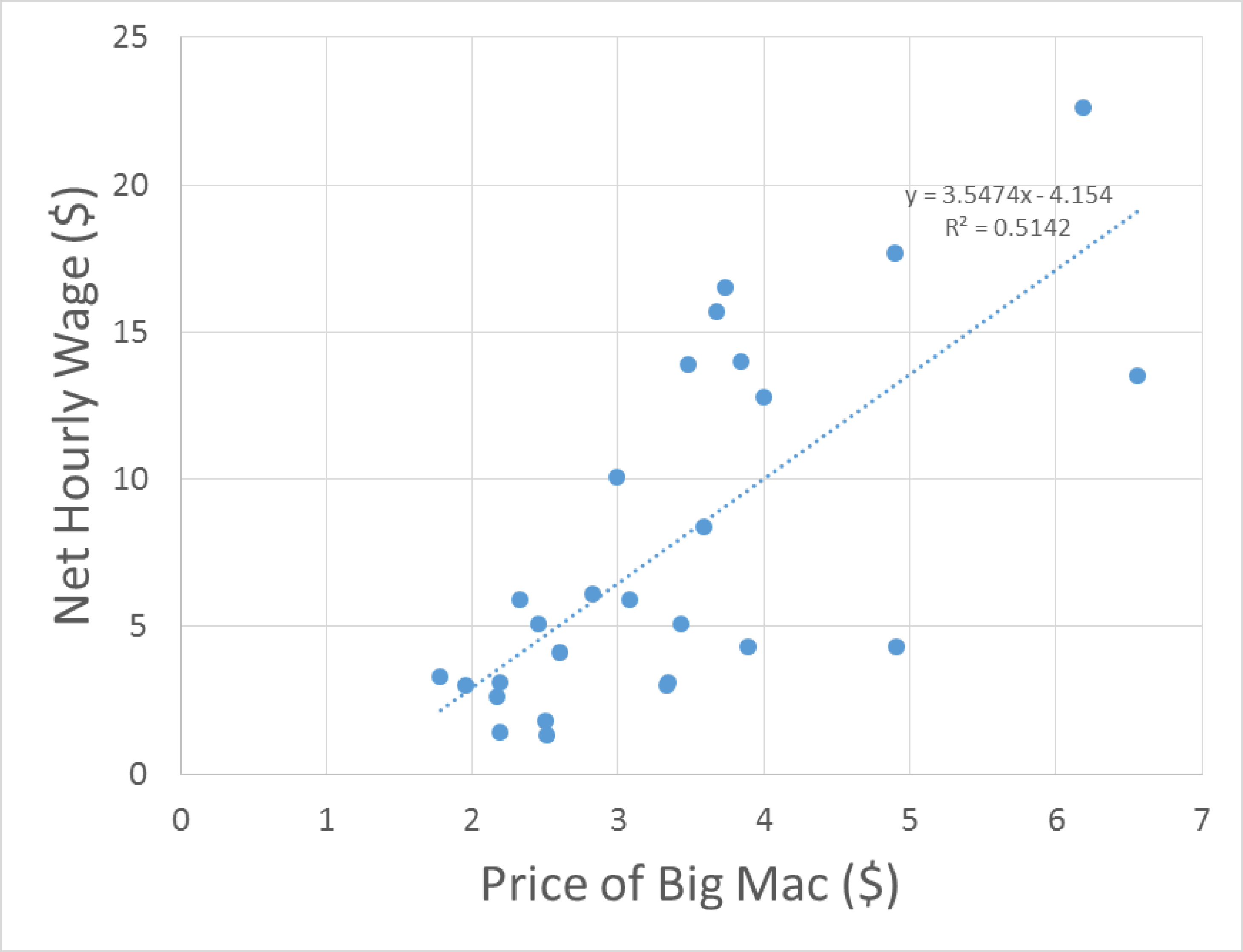
Burgernomics



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

Determining the Equation of the Regression Line – Excel [“Regression”]

Determining the Equation of the Regression Line - Excel



CSE 7302c

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			

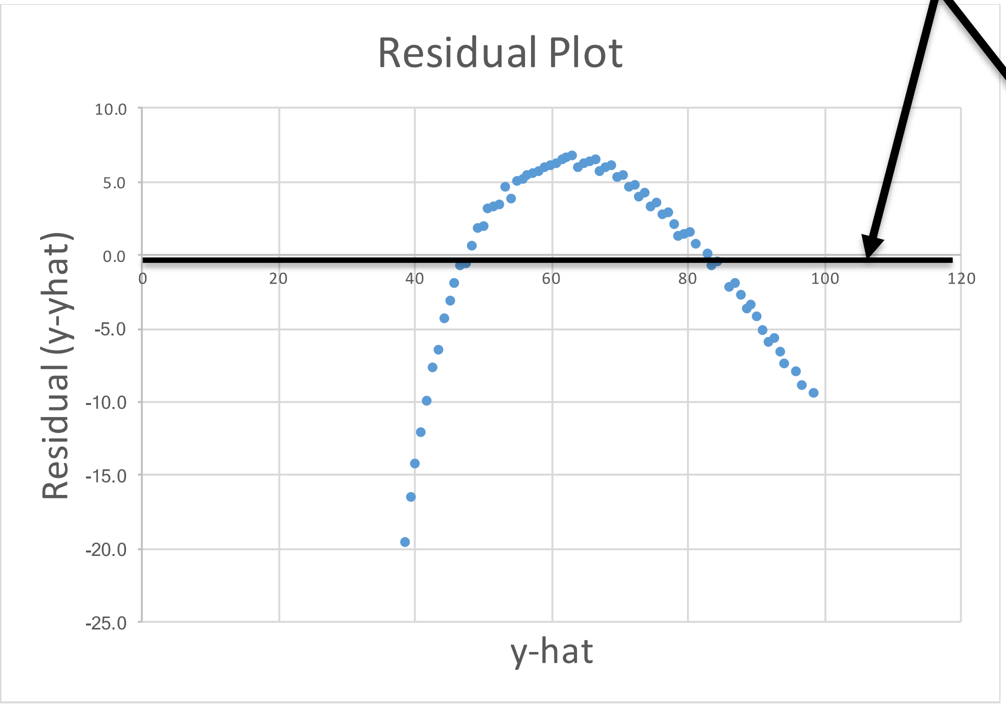
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0%	Upper 99.0%
Intercept	-4.154014573	2.447784673	-1.697050651	0.102104456	-9.195321476	0.88729233	-10.97705723	2.669028089
Big Mac Price ($)	3.547427488	0.689658599	5.143744297	2.57053E-05	2.127049014	4.967805962	1.625048409	5.469806567

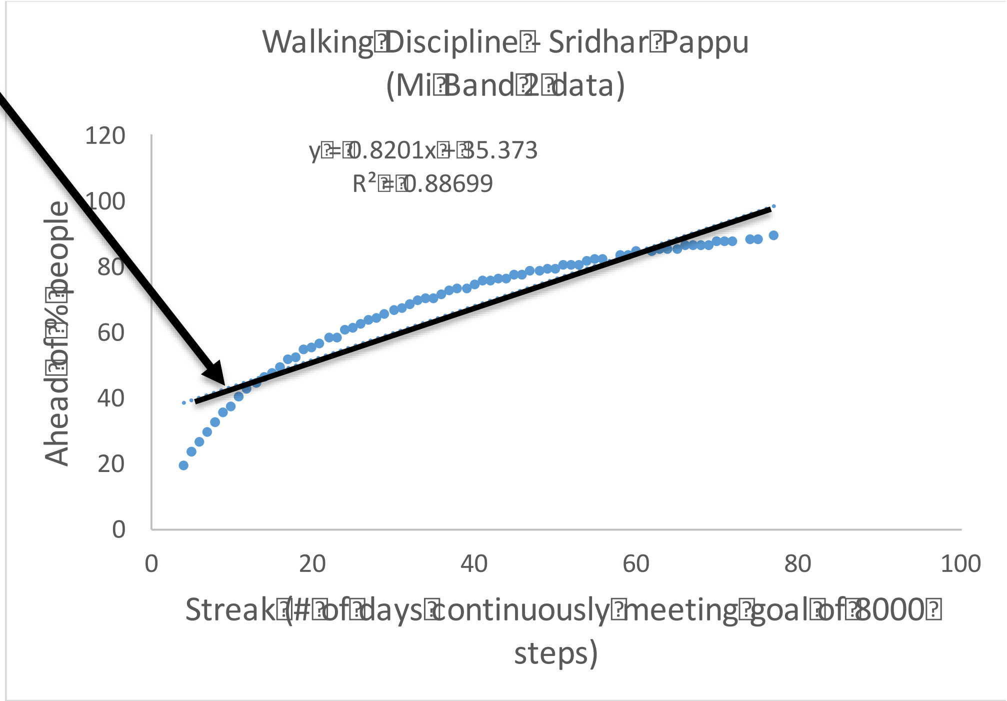
WAYS OF TESTING HOW WELL THE REGRESSION LINE FITS DATA

Assumptions of the Regression Model – Residuals Analysis

The model is linear

Zero residual line:
The regression line





Assumptions of the Regression Model – Excel ["Assumptions Error Dependence"]

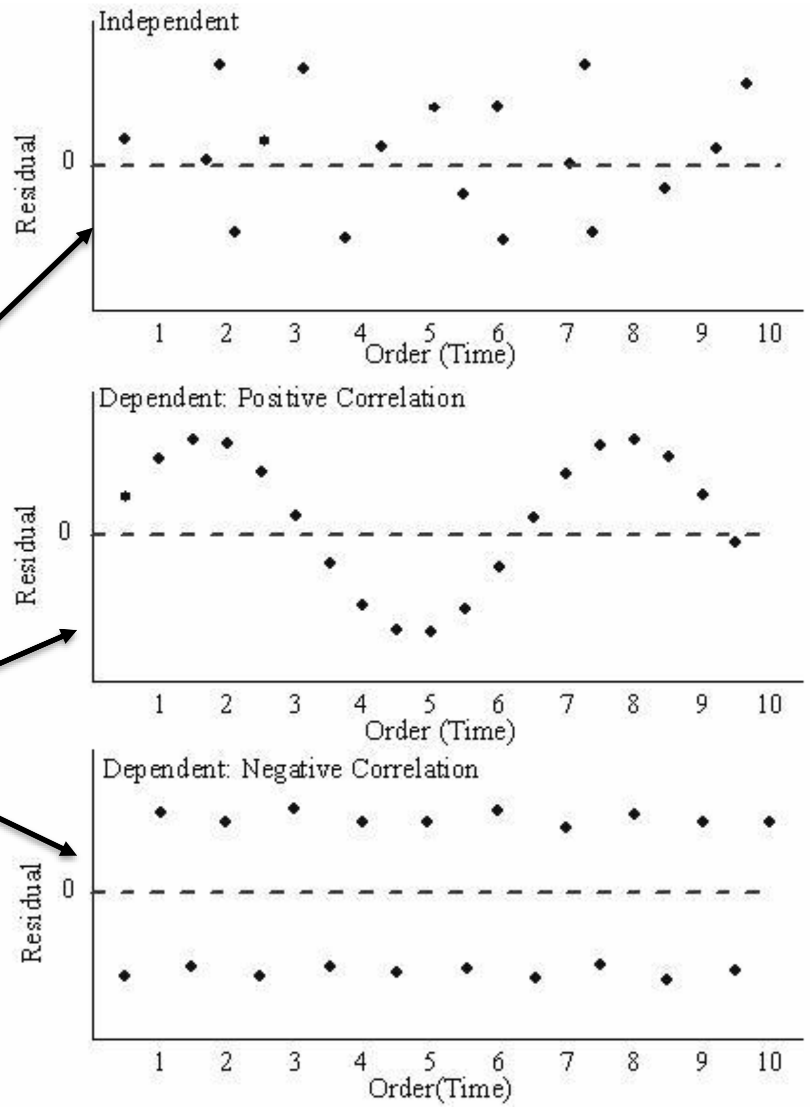
The error terms are independent

- 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

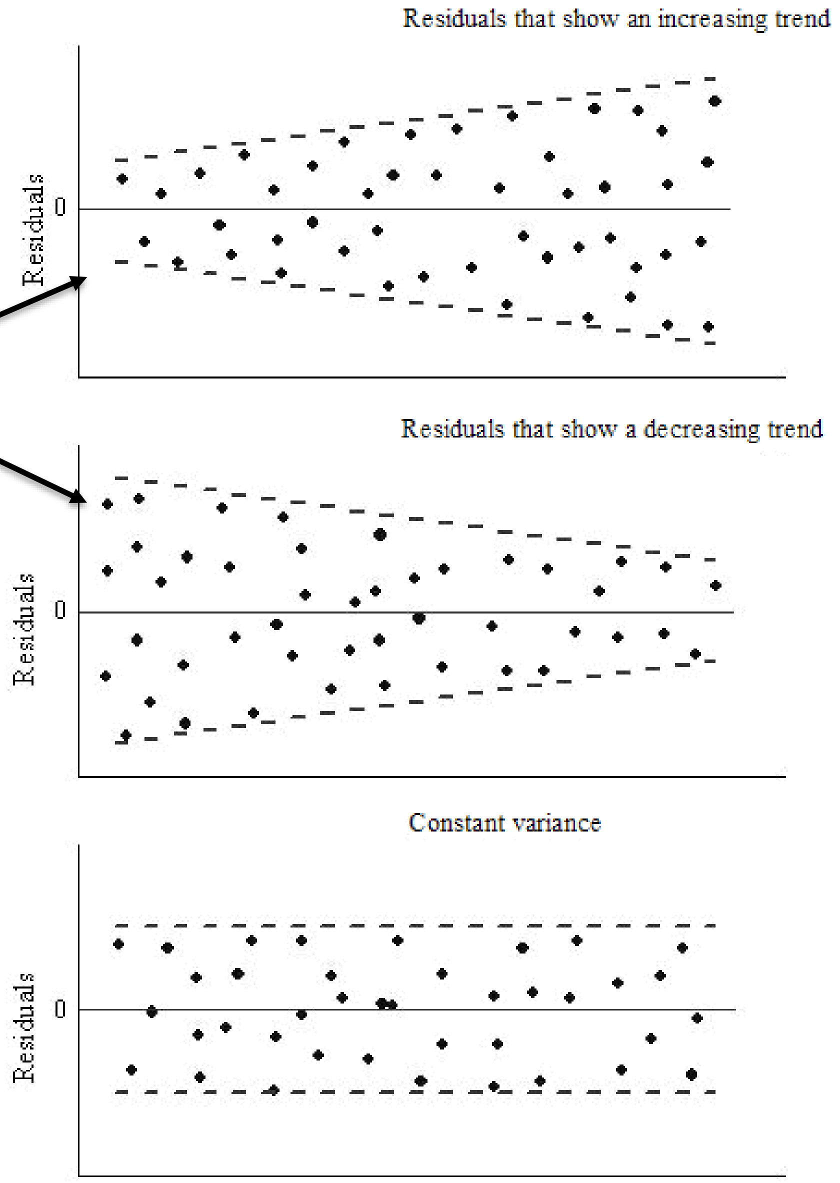


Assumptions of the Regression Model

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

Heteroscedastic

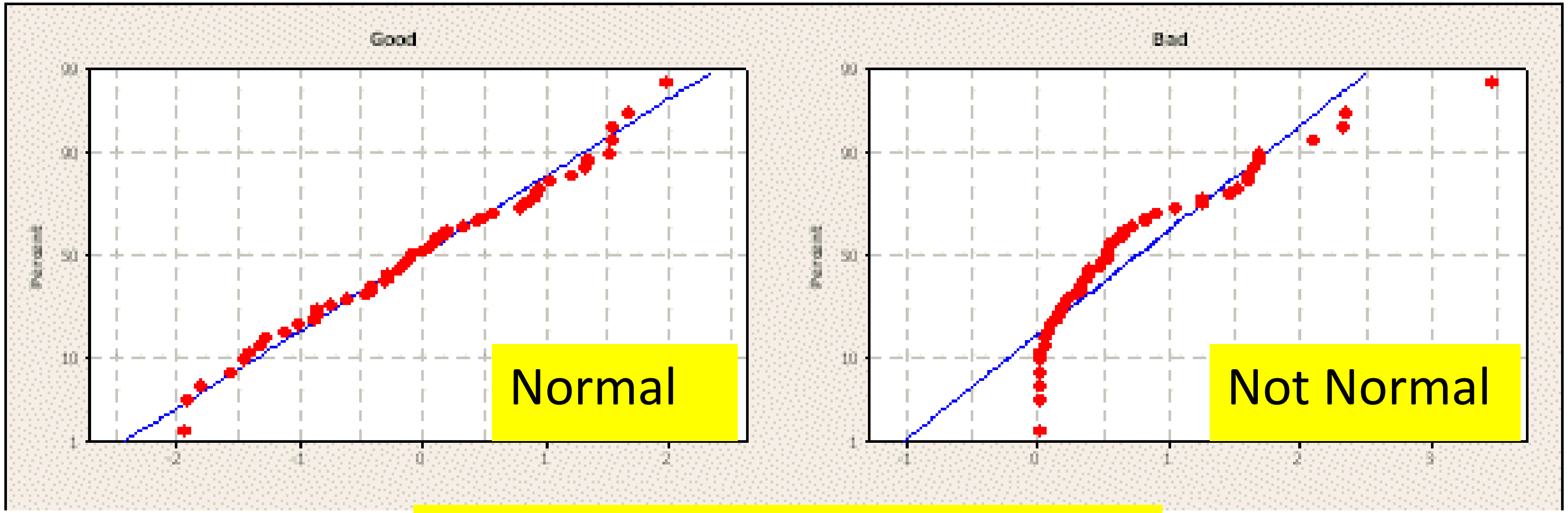
– RMSE (Root Mean Square Error) of Regression or Standard Error of the Estimate will be misleading as it will underestimate the spread for some x_i and overestimate for others.



Homoscedastic

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

CSE 7302c

Q-Q plot (Excel) [“Regression”]

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

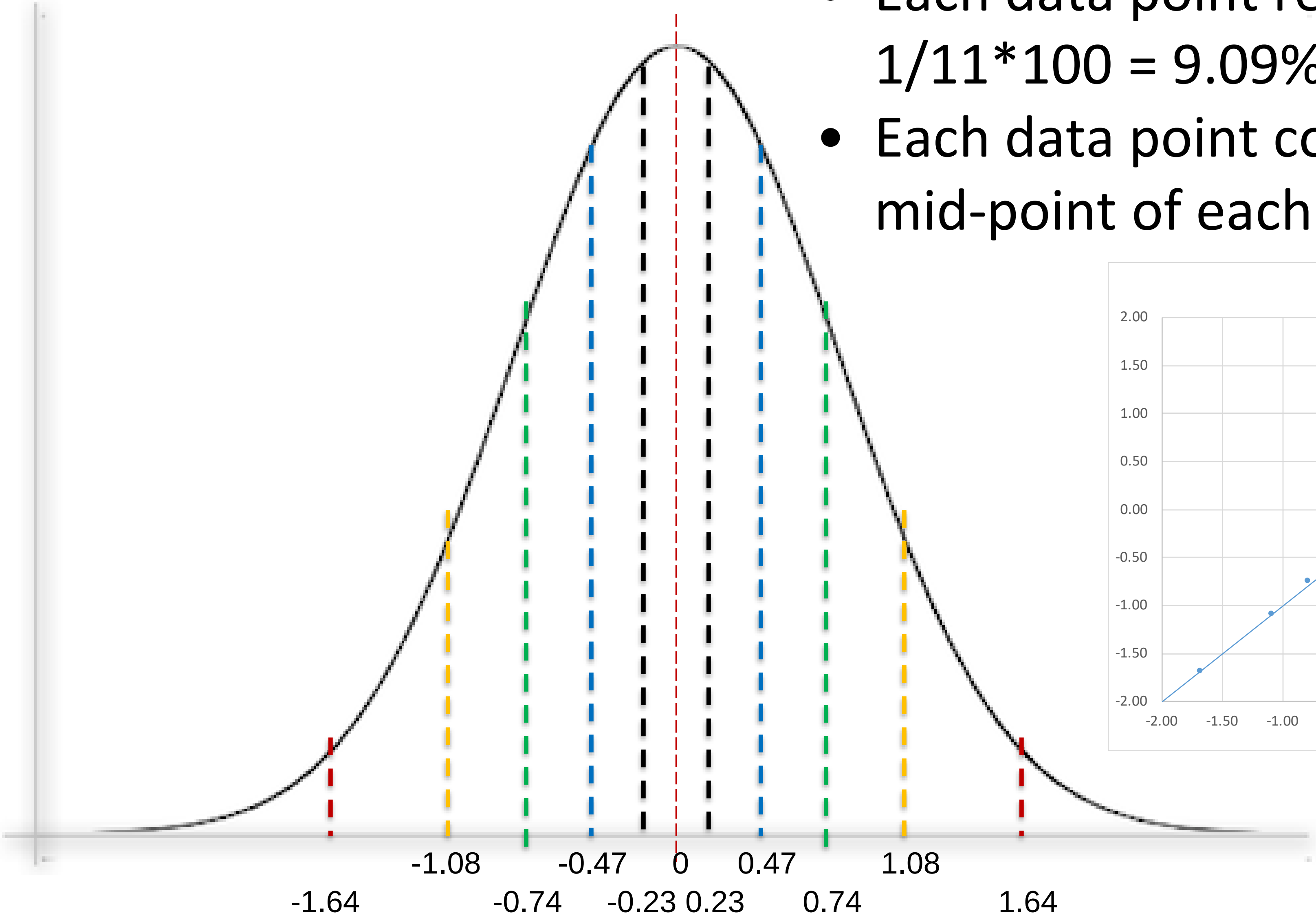
The quantile-quantile (q-q) plot is used to validate distributional assumptions of **a data set**.

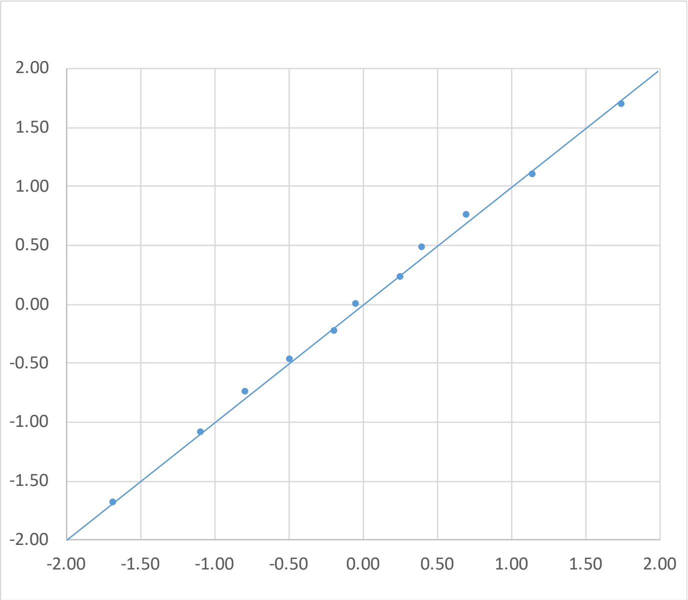
In linear regression, **this data set** is the residual errors.

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





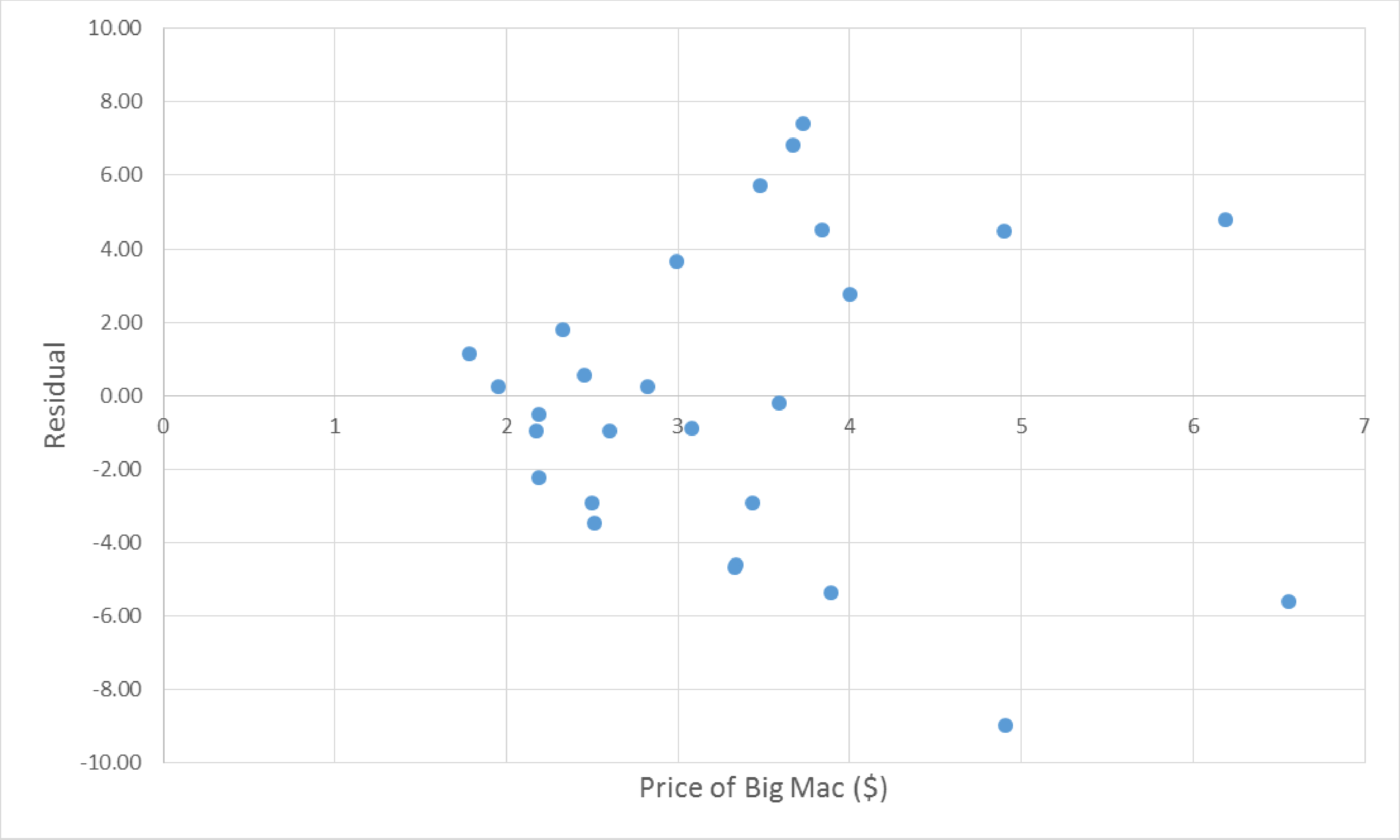
Interpreting Residuals

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

CSE 7302c

Residual Analysis – Big Mac

Which assumption is getting violated?

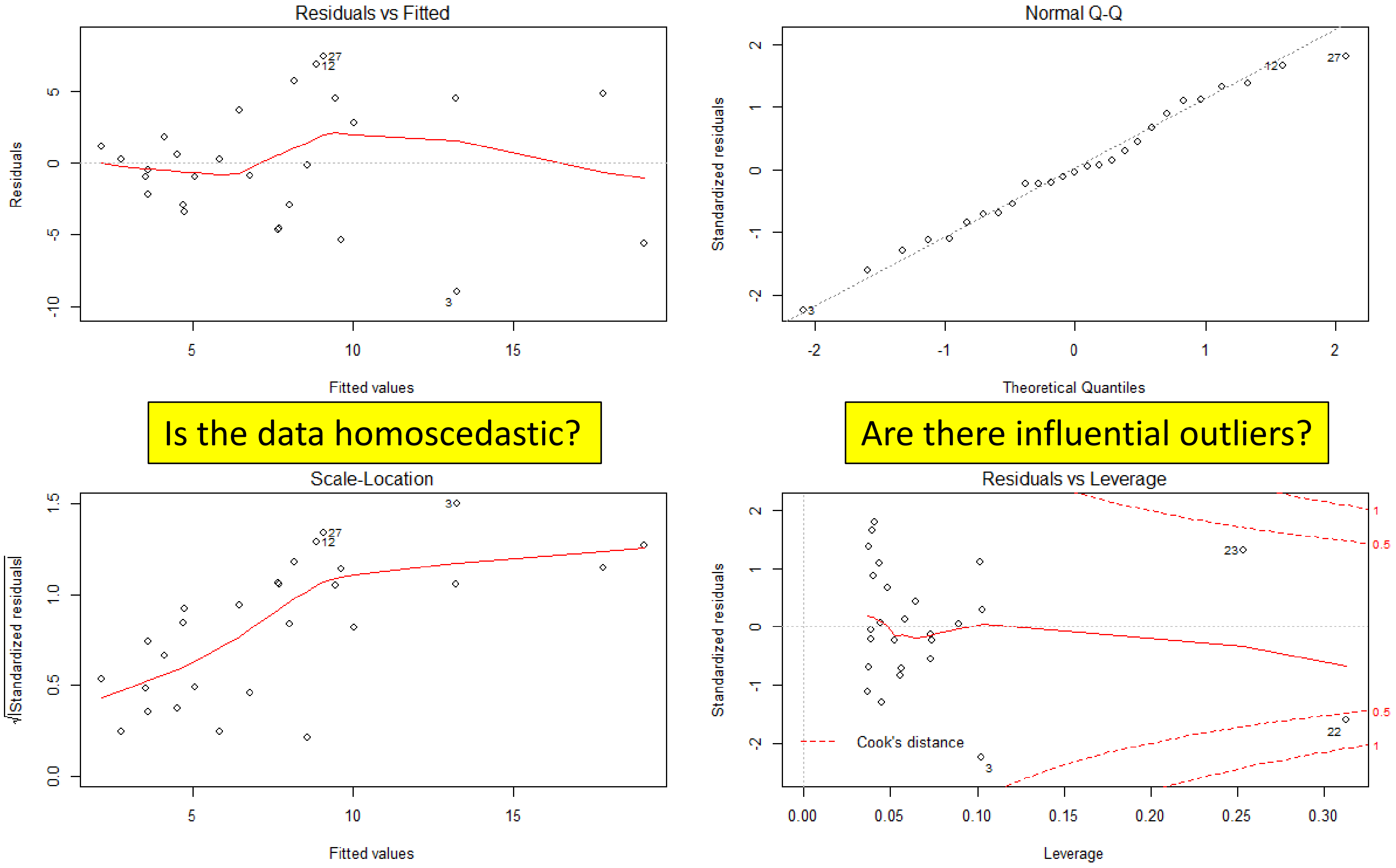


CSE 7302c

Residuals – Big Mac

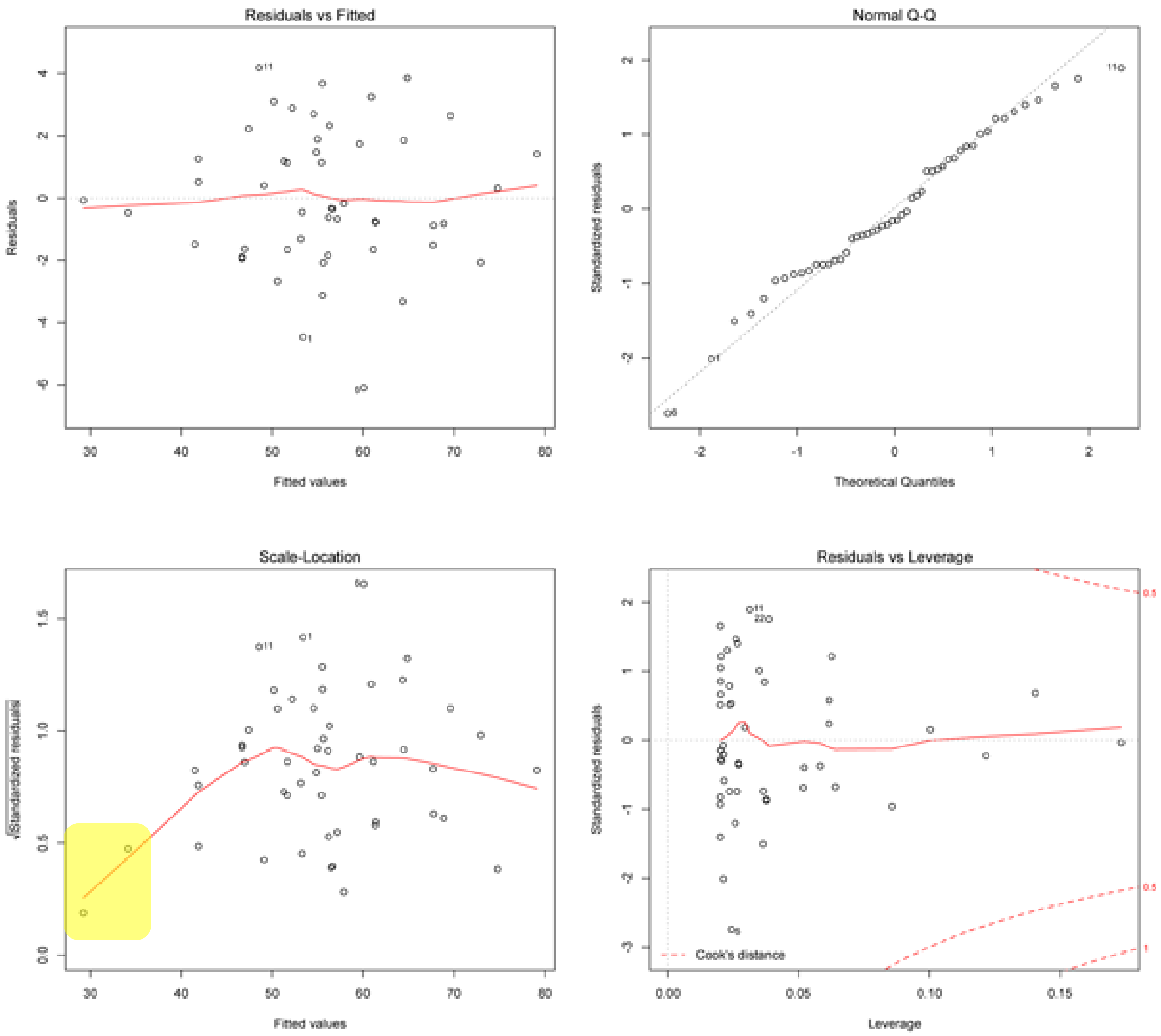
Is a wrong model fitted (linear or quadratic, etc.)? | Are the residuals normally distributed?

Is the data homoscedastic? | Are there influential outliers?



CSE 7302c

Caution – Is there heteroscedasticity here?



CSE 7302c

Fixing Non-normality and Heteroscedasticity

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

HYPOTHESIS TESTS

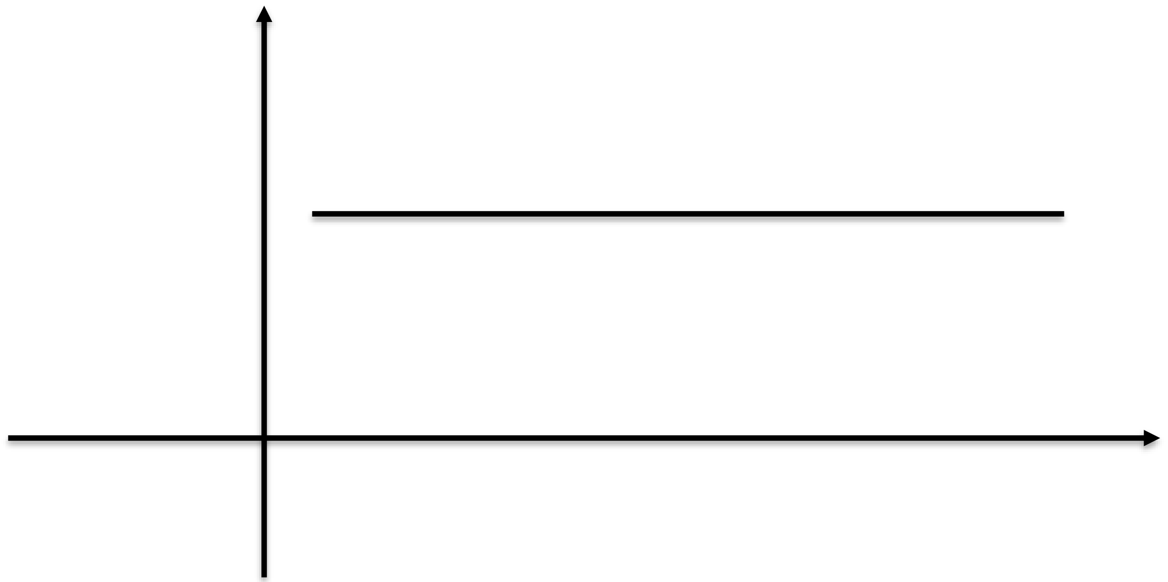
1) FOR THE SLOPE OF THE REGRESSION MODEL –T TEST

AND

2) TESTING THE OVERALL MODEL - ANOVA

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

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

We know that

SE = Standard Error

$$t = \frac{\bar{x} - \mu}{s/n}$$

Here we are dealing with slopes hence we use b_1 and β_1 Where

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

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

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

CSE 7302c

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 $\alpha/2$).

	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 Slope Output Calculations “t Test” – Big Mac – See Excel “Regression – Correlation and Regression tab”

$$t = \frac{b_1 - \beta_1}{s_b}. =. \frac{3.547427488}{0.6896586} = \mathbf{5.143744297} \qquad \text{Remember } \beta_1 = 0$$

$$b = \frac{\sum((x-\bar{x})(y-\bar{y}))}{\sum(x-\bar{x})^2} = 3.547427488$$

$$s_b = \frac{SE}{\sqrt{SS_{xx}}}. =. \frac{4.213191311}{\sqrt{37.32106667}} = 0.6896586$$

$$SE = \sqrt{\frac{\sum(e_i - \mu_e)^2}{df}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = 4.213191311$$

$$SS_{xx} = \sum(x - \bar{x})^2. = 37.32106667$$

CSE 7302c

Testing the Slope Output Calculations "t Test" – Big Mac See Excel "Regression – Correlation and Regression tab"

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

Since t is greater then $t_{\text{critical value}}$ we can reject the null hypothesis or t value is not significant which means alternate hypothesis of $\beta_1 \neq 0$ is correct.

Another way of doing this is to get the p value which is *R code: pt(5.1437,25,lower.tail = FALSE)* = 0.0000128. Since this is less than 0.025, we reject the null hypothesis.

However in R output of linear regression the value of p is doubled and we need to compare to 0.05 instead of 0.025. So the value of p that is output by the software is 0.0000128*2 = 2.57053E-05 which is significantly less than 0.05 and so we reject the null hypothesis

Business consequence: Big Mac price has an impact on "net hourly wage"

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

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

CSE 7302c

Testing the Overall Model – Big Mac - Excel

$F = 26.4581$

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

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 - Explanation

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

R=Correlation Coefficient

R^2 – How much of variance (Net Wages) can be explained by Big Mac Price

SE=Standard Error

n=Total Number of Observations

k=1 since only one β_1

Degrees of Freedom

n-k-1

Error

n-1

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

SSR = Sum of Squared Regression

SSE = Sum of Squared Errors

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

SST = Sum of Squared Total = SSR + SSE

MSR = Mean Squared Regression

MSR = $\frac{SSR}{df_{reg}}$

MSE = Mean Squared Error

MSE = $\frac{SSE}{df_{erro}}$

F = F-value = $\frac{MSR}{MSE}$

Significance F = p-value = *pf(26.4581,1,25,lower.tail = FALSE)*

CSE 7302c

BREAK

Sample Software Output (Equation = Net Wage = -4.154014573 + 3.547427488 BigMacPrice)

SUMMARY OUTPUT

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

R=Correlation Coefficient

R^2 – How of variance (Net Wages) can be explained by Big Mac Price

SE=Standard Error

n=Total Number of Observations

ANOVA

Degrees of Freedom

	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			

Error

	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

Y-Intercept

Slope = b

Confidence interval 95%

Confidence interval 99%

Standard Error of Slope

$$s_b = \frac{SE}{\sqrt{SS_{xx}}}$$

t test value for slope

$$t = \frac{b_1 - \beta_1}{s_b} \text{.remember } \beta_1 = 0$$

Significance t for slope = p-value = *pt(5.143744297,25,lower.tail = FALSE)*

CSE 7302c

Simple Linear Regression - Steps

Get familiar with data
- Plots
- Descriptive stats

Formulate a linear model and fit to data
- Do regression

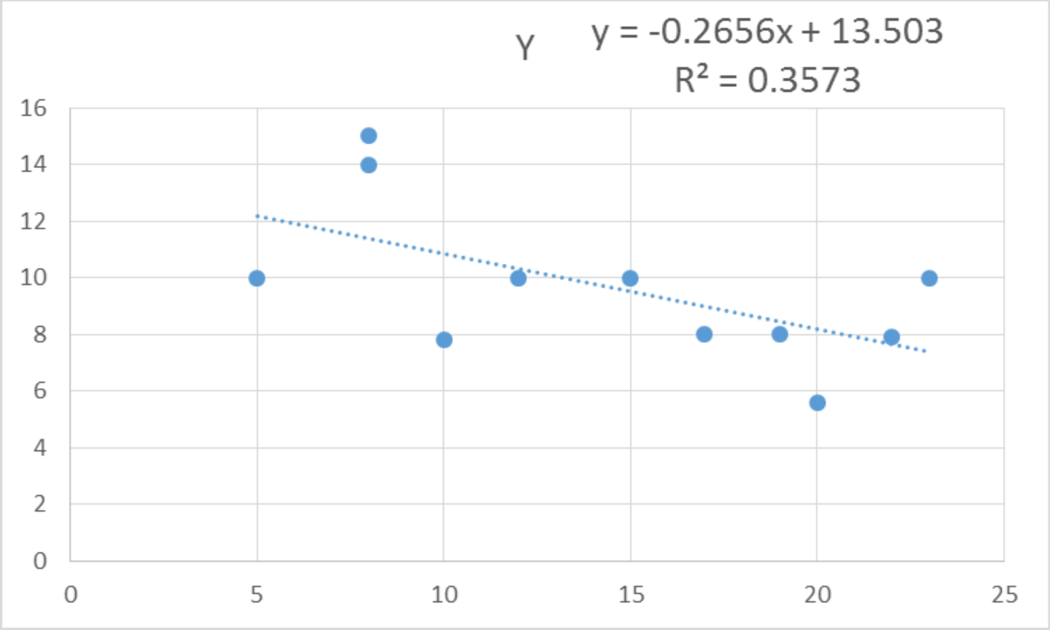
Inadequate fit

Check model and assumptions
- Look at residual plots
- Look at unusual observations
- Look at R-Squared
- Look at p-values

Good fit

Report results and equation
- Make predictions for values of interest

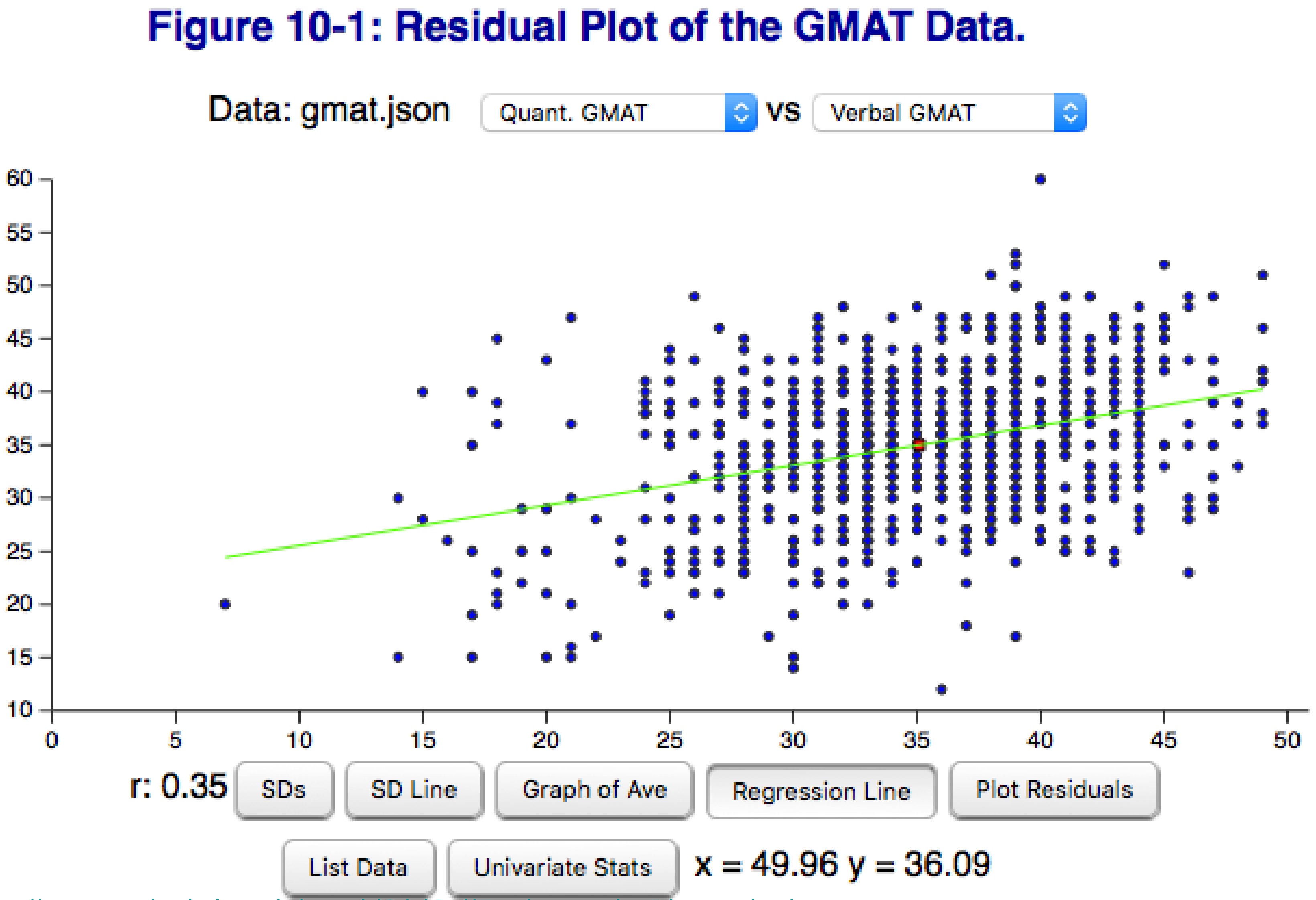
R-Squared and Significance - Caution



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

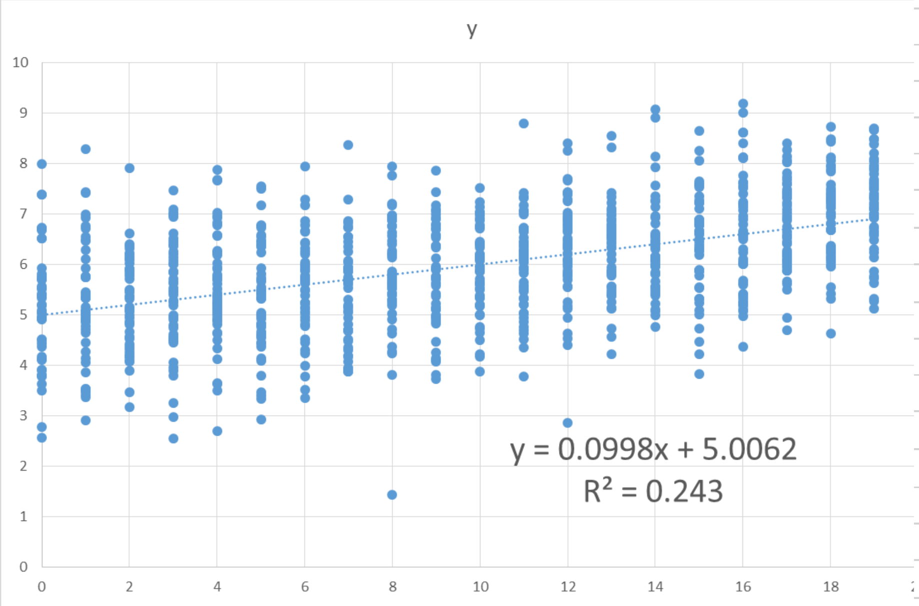
- R^2 suggests that 35% of variation in *y* can be explained by variation in *x*.
- *t* and *F* tests show that coefficient is <u>not significant</u> and null hypothesis cannot be rejected.
- The 95% confidence interval of the slope, $b_1 \pm t_{crit} * s_b$, is (-0.534,0.003).

R-Squared and Significance - Caution



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 <u>significant</u> 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*.

CSE 7302c

R-Squared, Significance and Residuals – Caution Excel Activities [”Rsquared Distance” Tab-Car Stopping

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

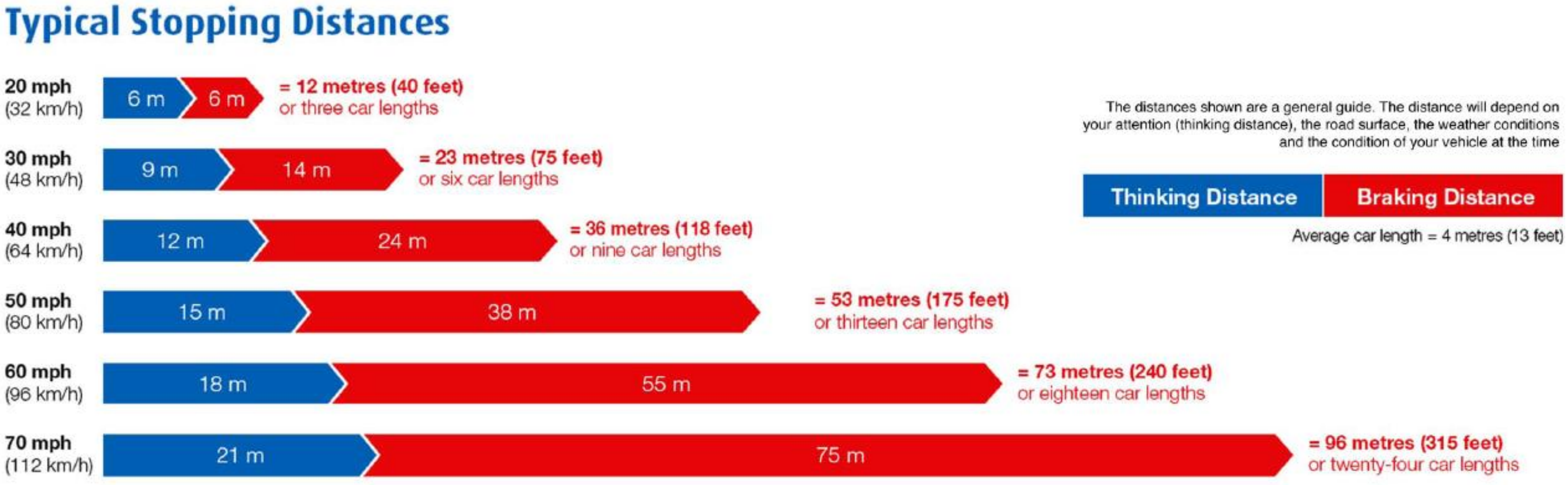
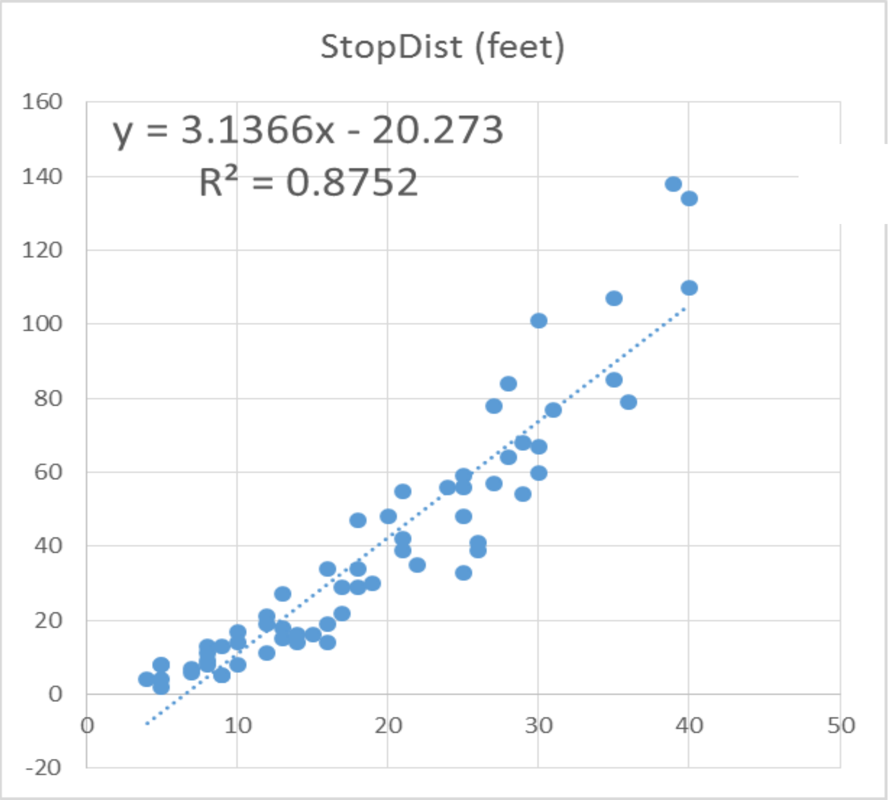


Image Source: http://streets.mn/2015/04/02/the-critical-ten/
Last accessed: November 20, 2015

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.

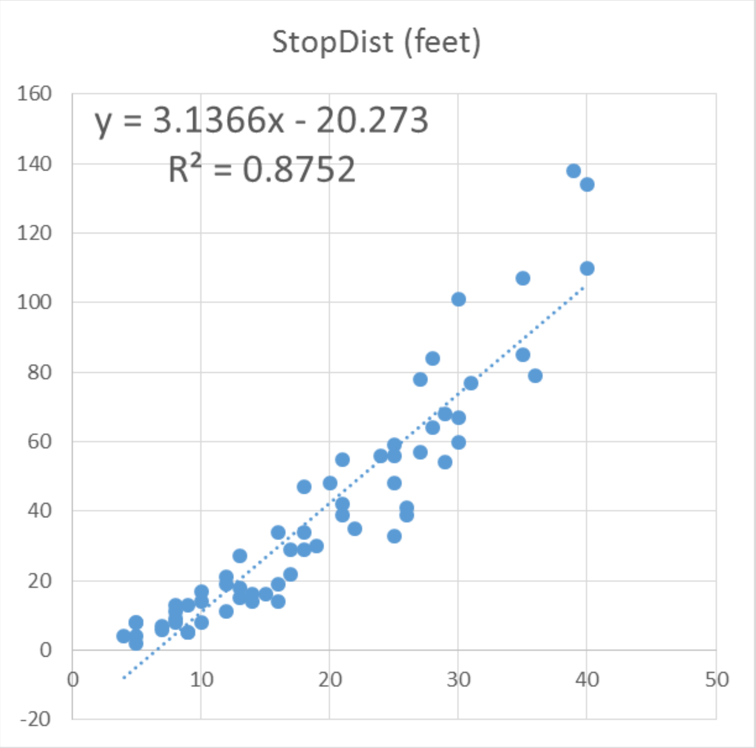
Does the estimated regression line fit the data well?

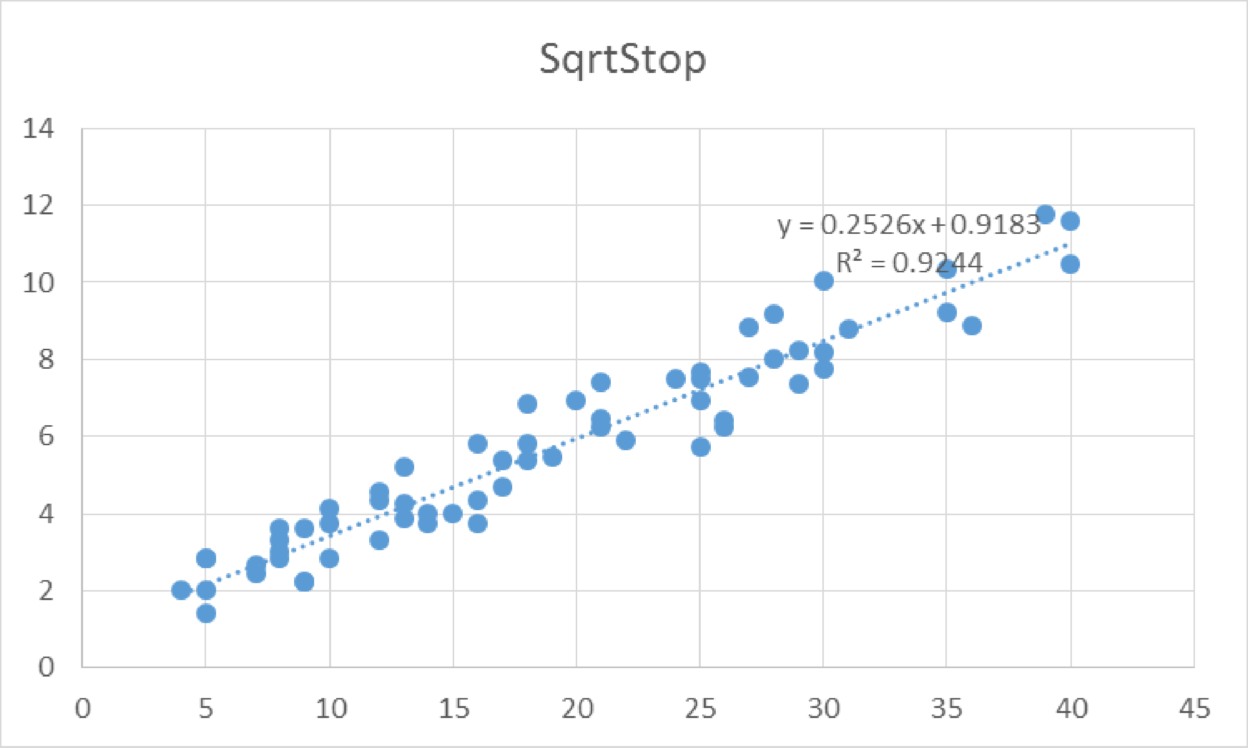


R-Squared, Significance and Residuals - Caution

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.





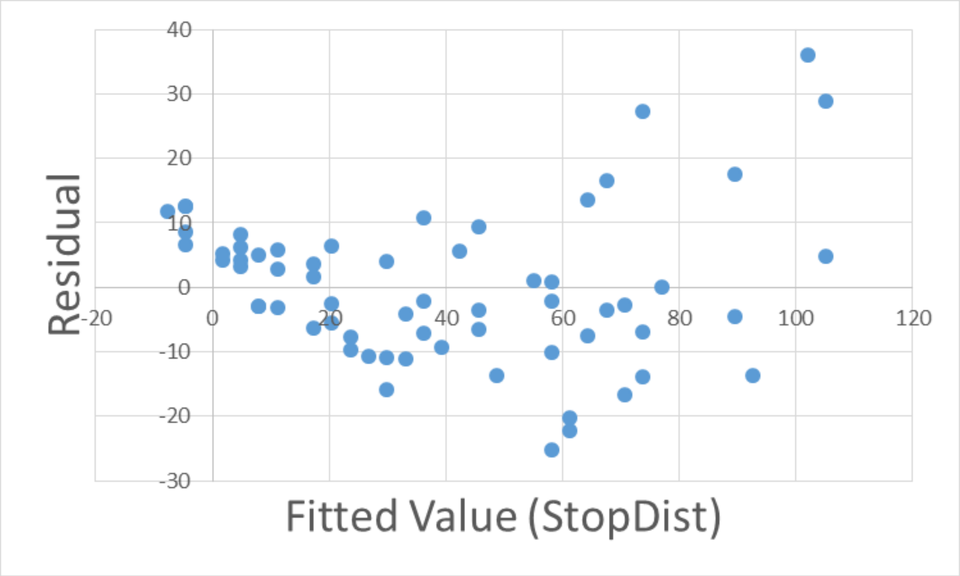
CSE 7302c

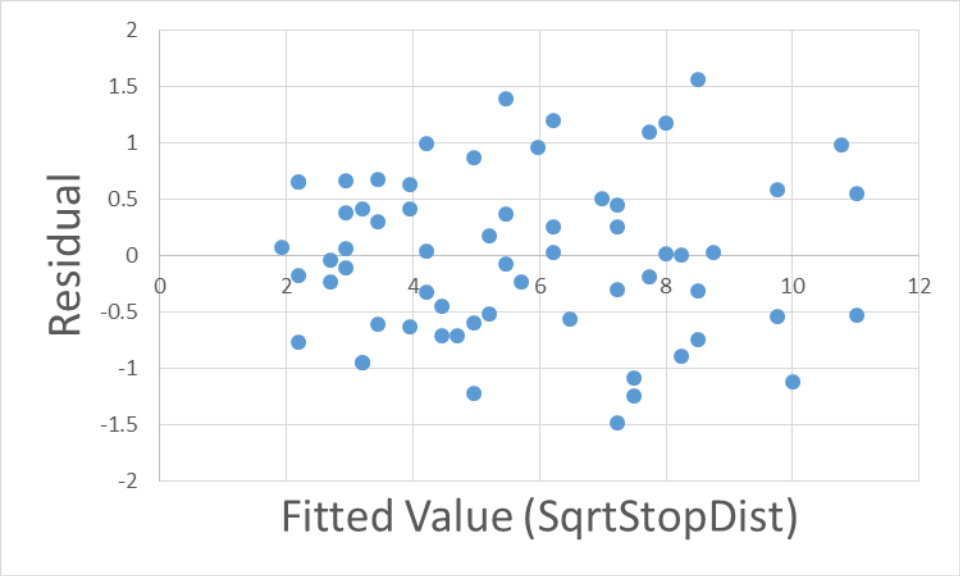
R-Squared, Significance and Residuals - Caution

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.



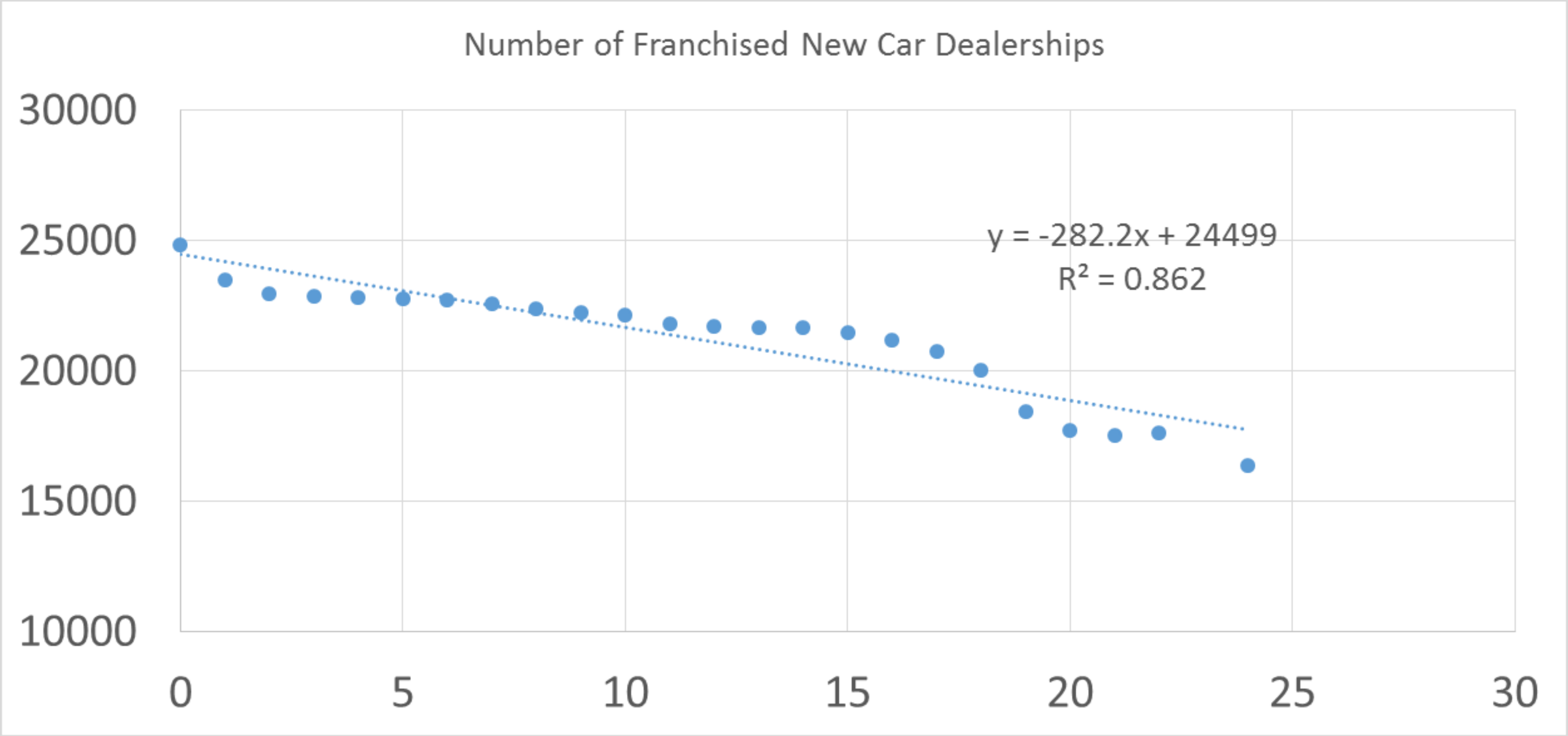


R-Squared, Significance and Residuals - Caution

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 [“Rsquared Distance” NADA Dealership Tab]

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?

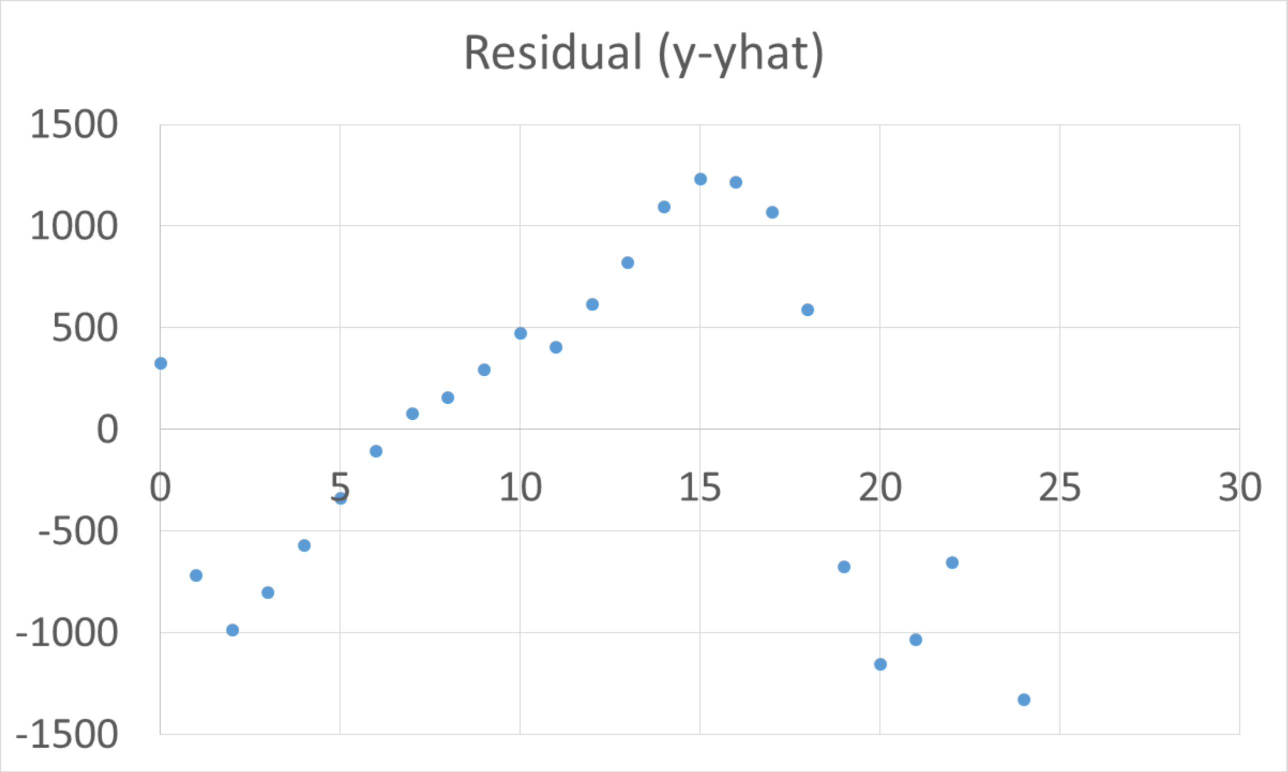
CSE 7302c

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	

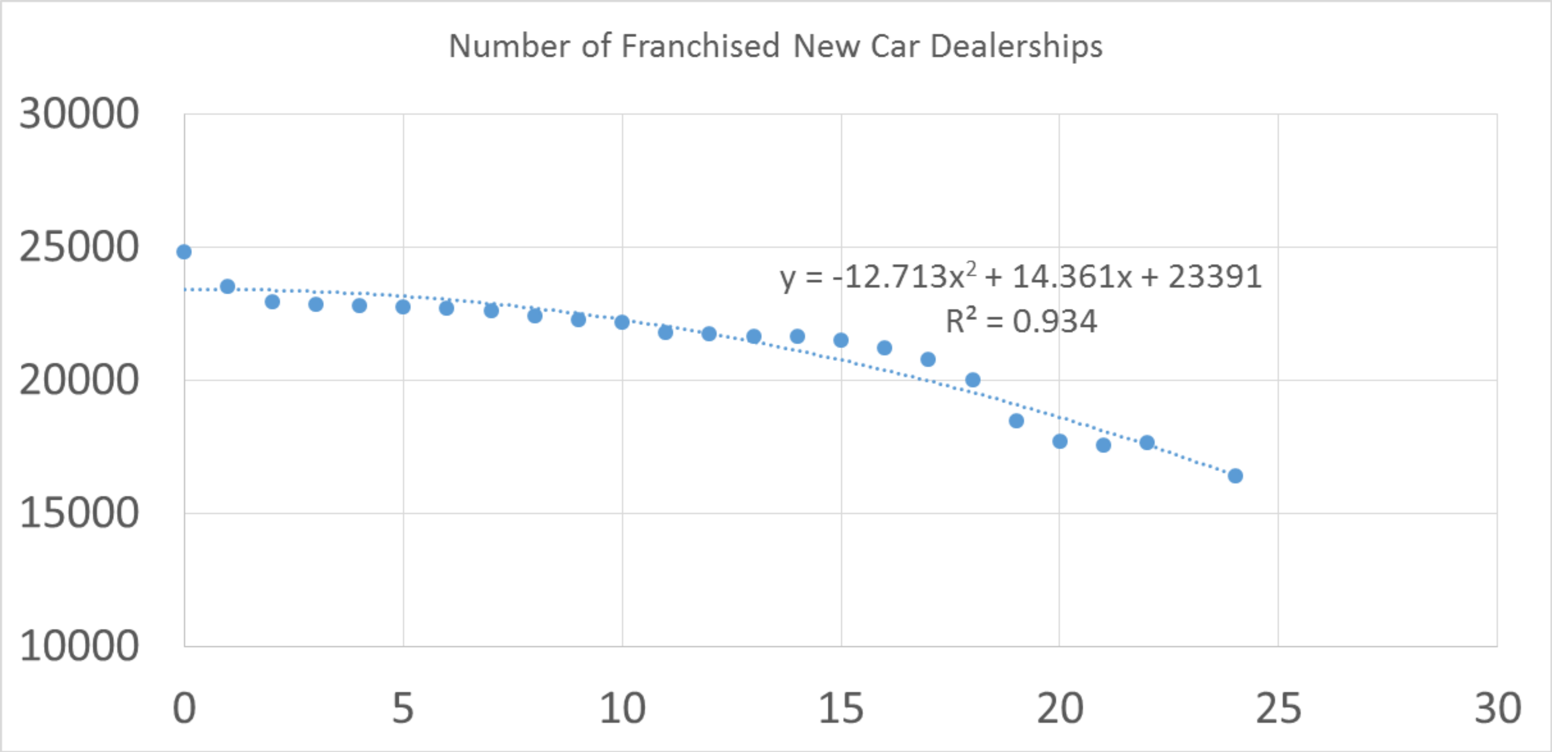
- Is the slope significant?
- Is the model significant?

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



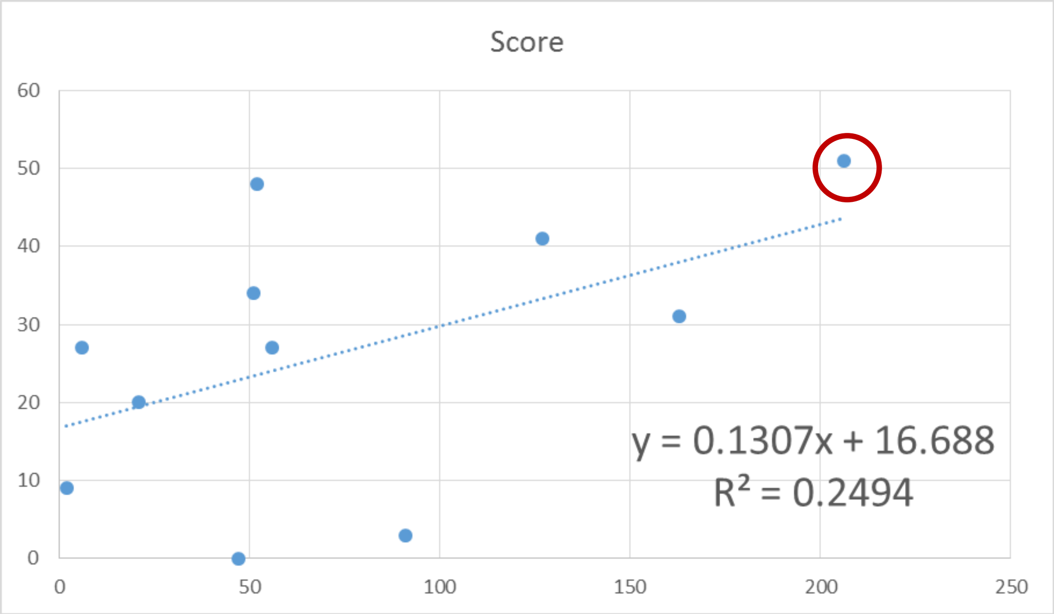
NOTE

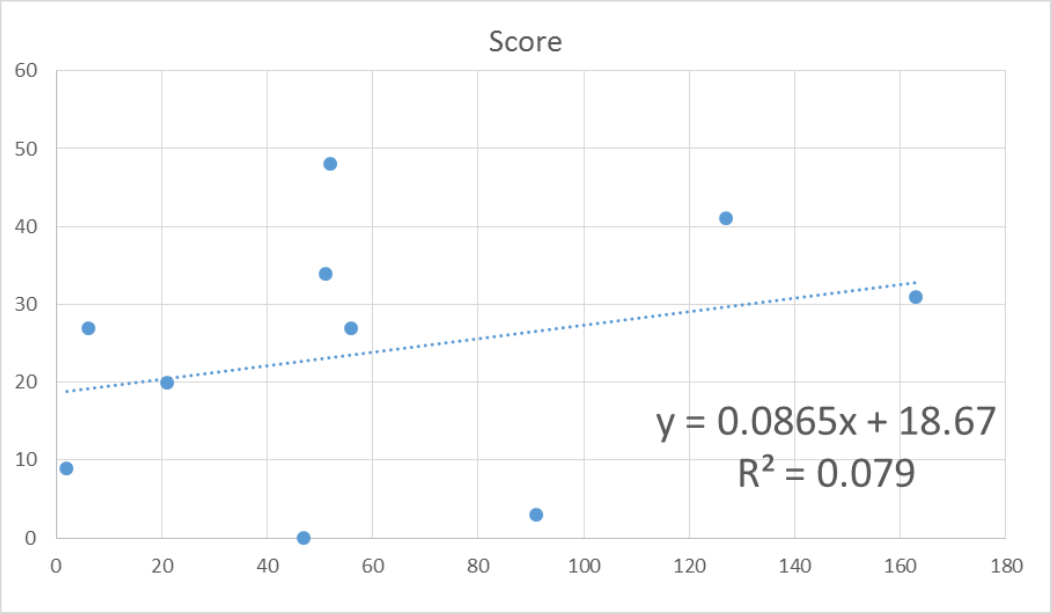
Since the data here are ordered (time is a factor), the residual plot also indicates violation of the *independence* assumption. Time series analysis becomes the right approach for this dataset instead of OLS Regression.

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.

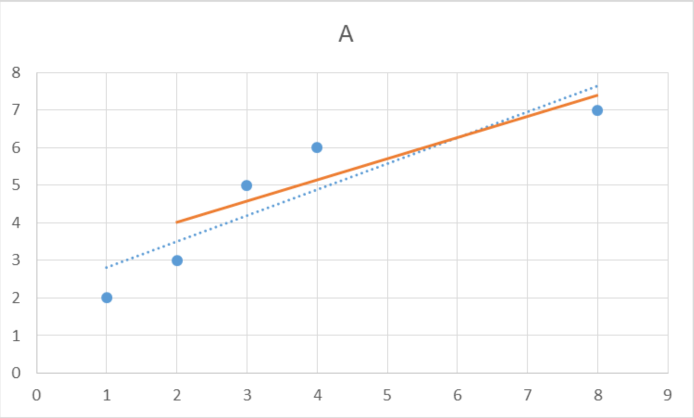
Influence is a function of **leverage** and **distance (or 'residuality' or 'outlierness')**.

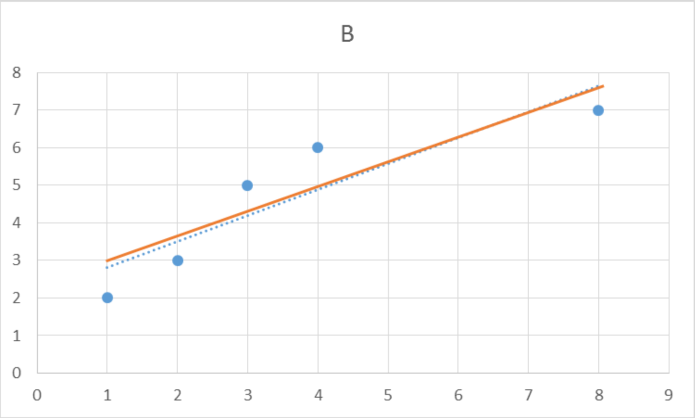
Influential Observations

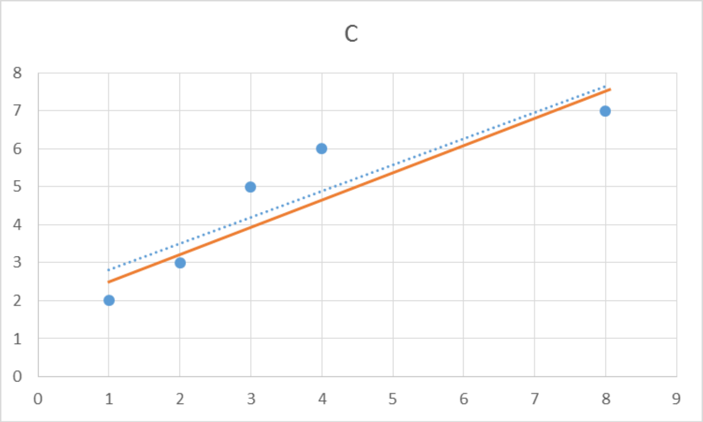
ID	X	Y	h	R	D
A	1	2	0.39	-1.02	0.4
B	2	3	0.27	-0.56	0.06
C	3	5	0.21	0.89	0.11
D	4	6	0.2	1.22	0.19
E	8	7	0.73	-1.68	8.86
h is the leverage, R is the studentized residual, and D is Cook's measure of influence.					

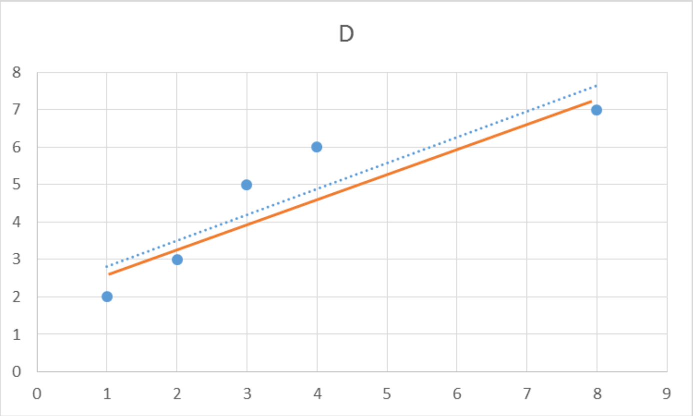
Source: http://onlinestatbook.com/2/regression/influential.html
Last accessed: June 30, 2017

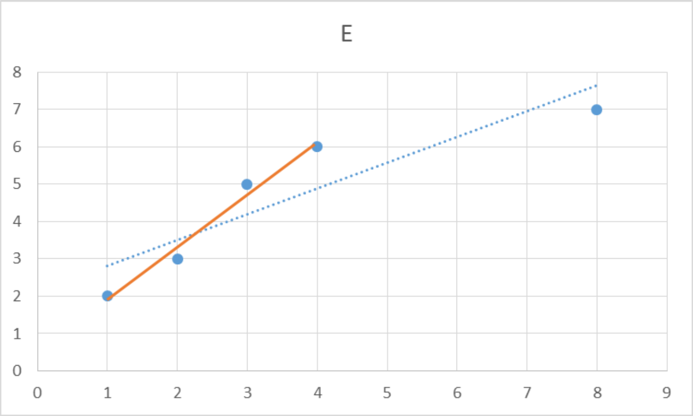
Influential Observations





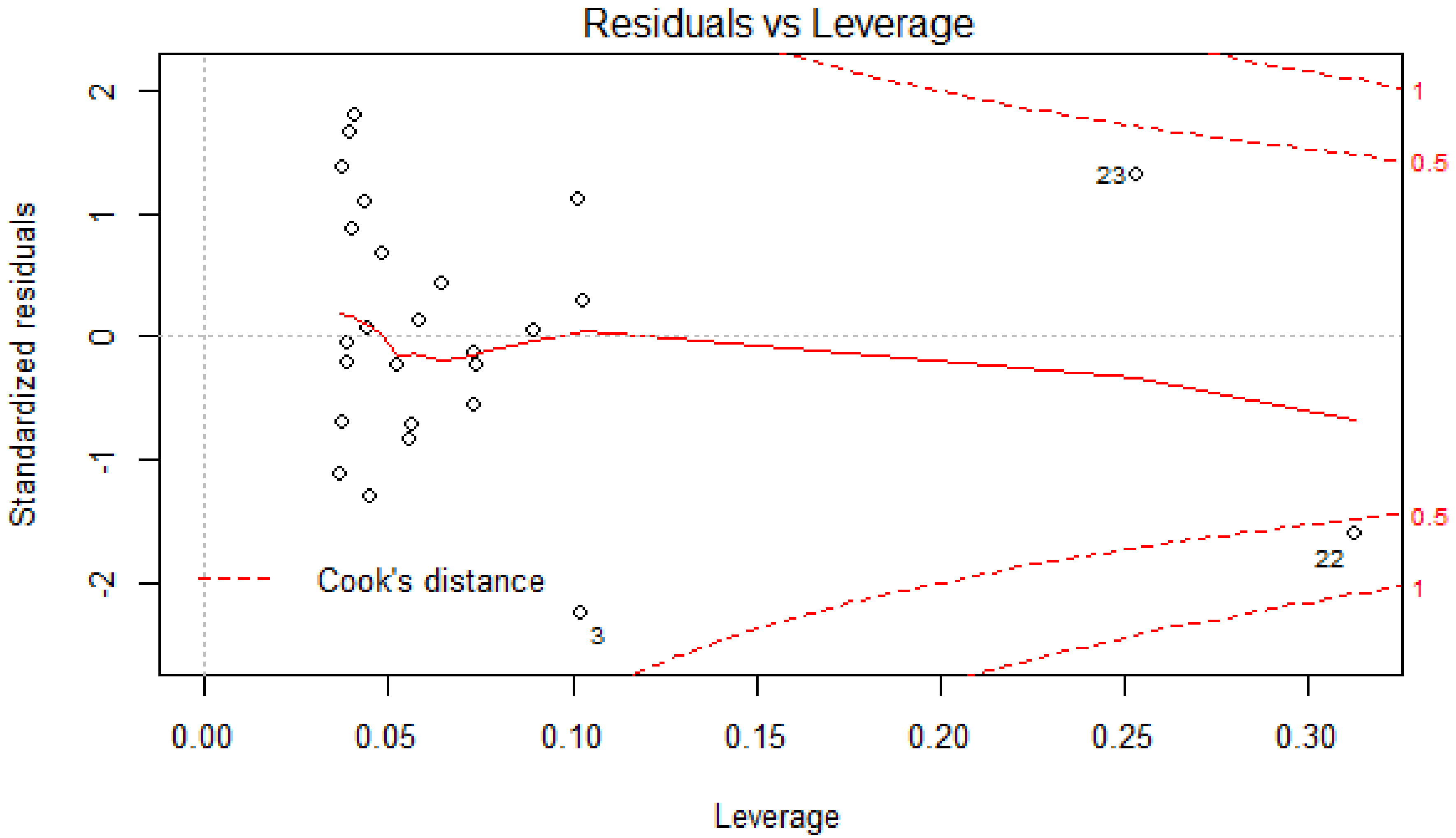






Influential Observations

Are there influential outliers?



Influential Observations – Rules of Thumb

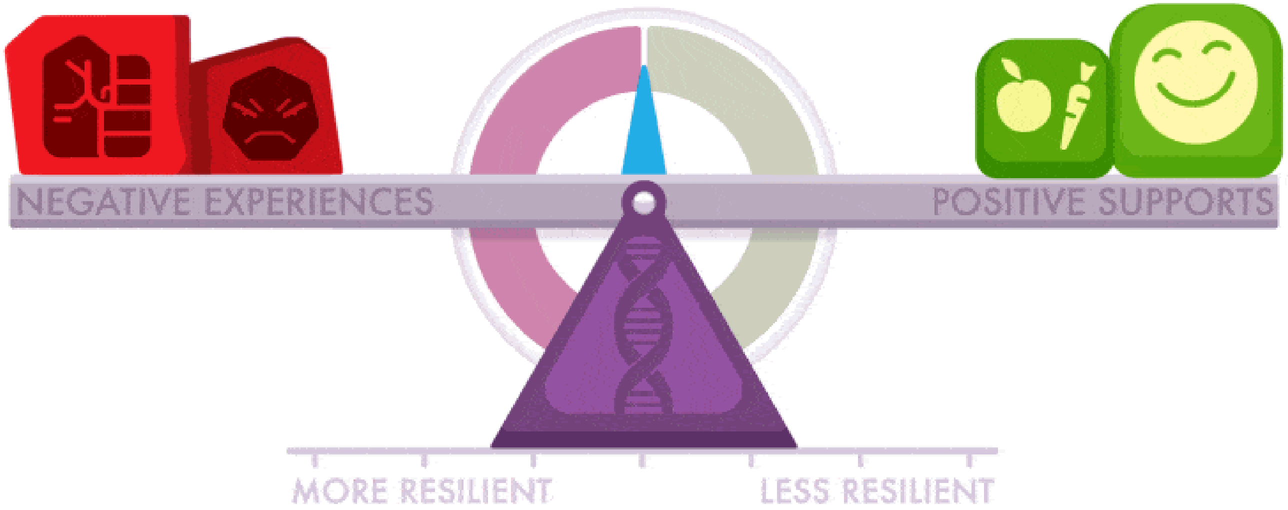
- If Cook's D of any observation (D_i) > 1, that observation can be considered as having too much influence, but investigate values greater than 0.5 also.

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

Influential Observations - Leverage

How much the observation's value on the **predictor variable** differs from the mean of the **predictor variable**.

That is, it tells us about extreme x values, which have the potential to highly influence the regression in certain conditions. *Remember Eric McCoo.*



CSE 7302c

Influential Observations – Leverage [Excel "Rsquared-Significance" Tab Leverage]

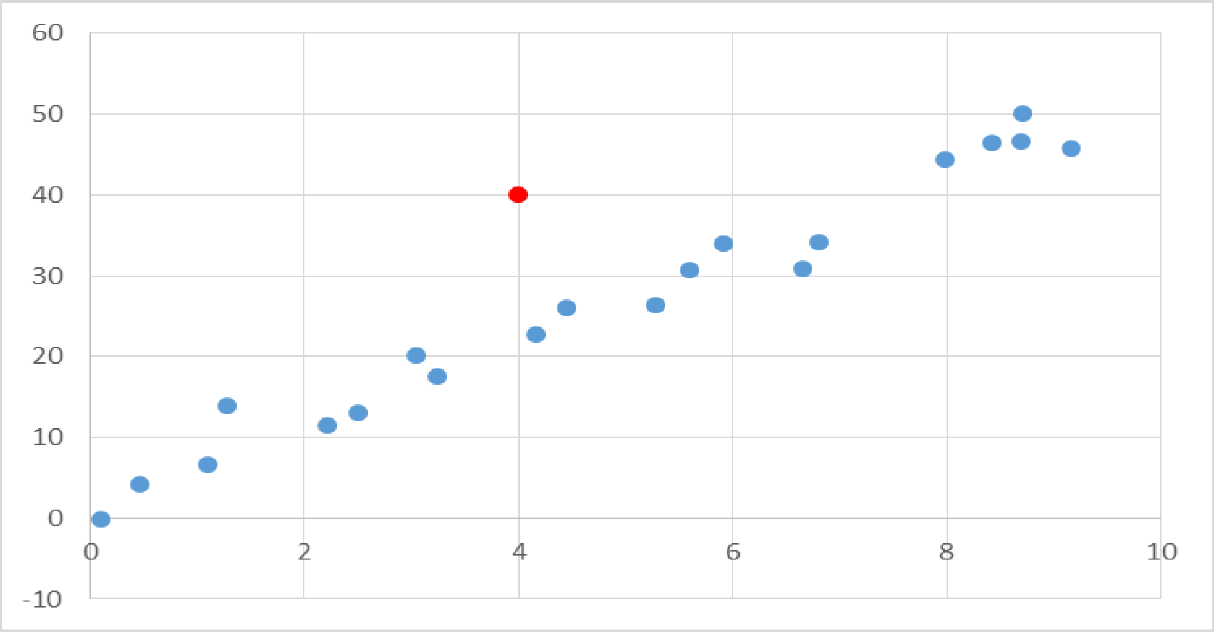
Leverage of the i^{th} data point is given by:

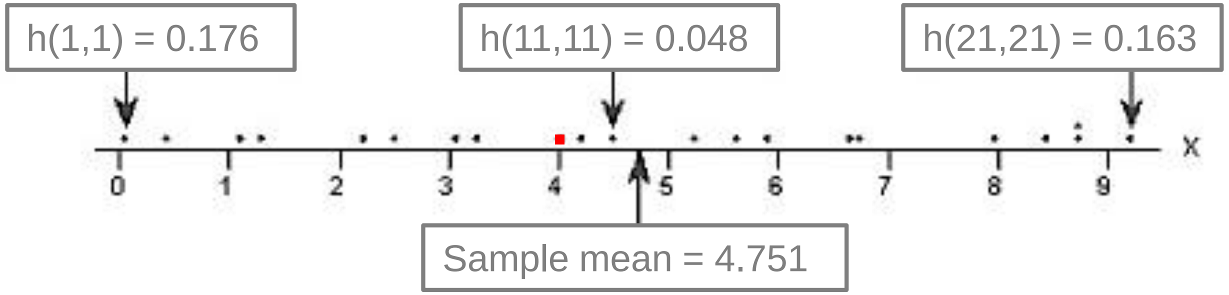
$$h_i = \frac{1+z^2}{n}$$

The sum of leverages = # of parameters, p (regression coefficients **including intercept**).

EXCEL ACTIVITY

Influential Observations - Leverage





Flag observations whose h > 3* avg(h) or h > 2* avg(h)

$$Avg(h) = \frac{sum(h)}{n} = \frac{p}{n}$$

CSE 7302c

Influential Observations - Distance

Based on error of prediction and is measured by Studentized Residual. This is calculated on the **dependent** variable and is a measure of ‘outlierness’.

Recall Student’s t-test. So, Studentizing is related to calculating the t-statistic of the metric in question, i.e., it is related to error of prediction of that observation divided by the standard deviation of the errors of prediction.

Influential Observations - Distance [Excel “RSquared-Significance” Tab Influence]

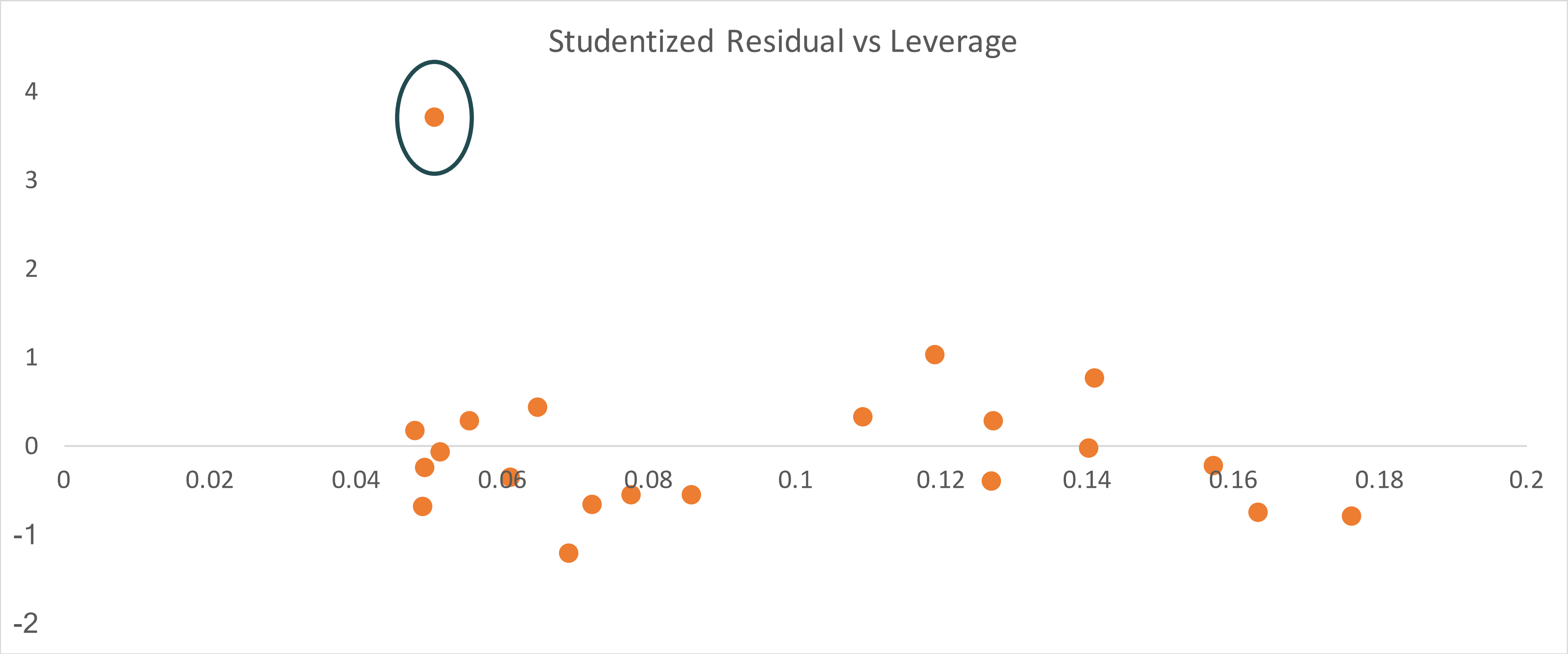
$$stdres_i = \frac{e_i}{\sqrt{MSE(1 - h_i)}}$$

Investigate observations with internally studentized residuals smaller than -2 or larger than 2.

Recall the empirical rule for normal distribution and the assumption that residuals follow normal distribution.

EXCEL ACTIVITY

Influential Observations - Distance

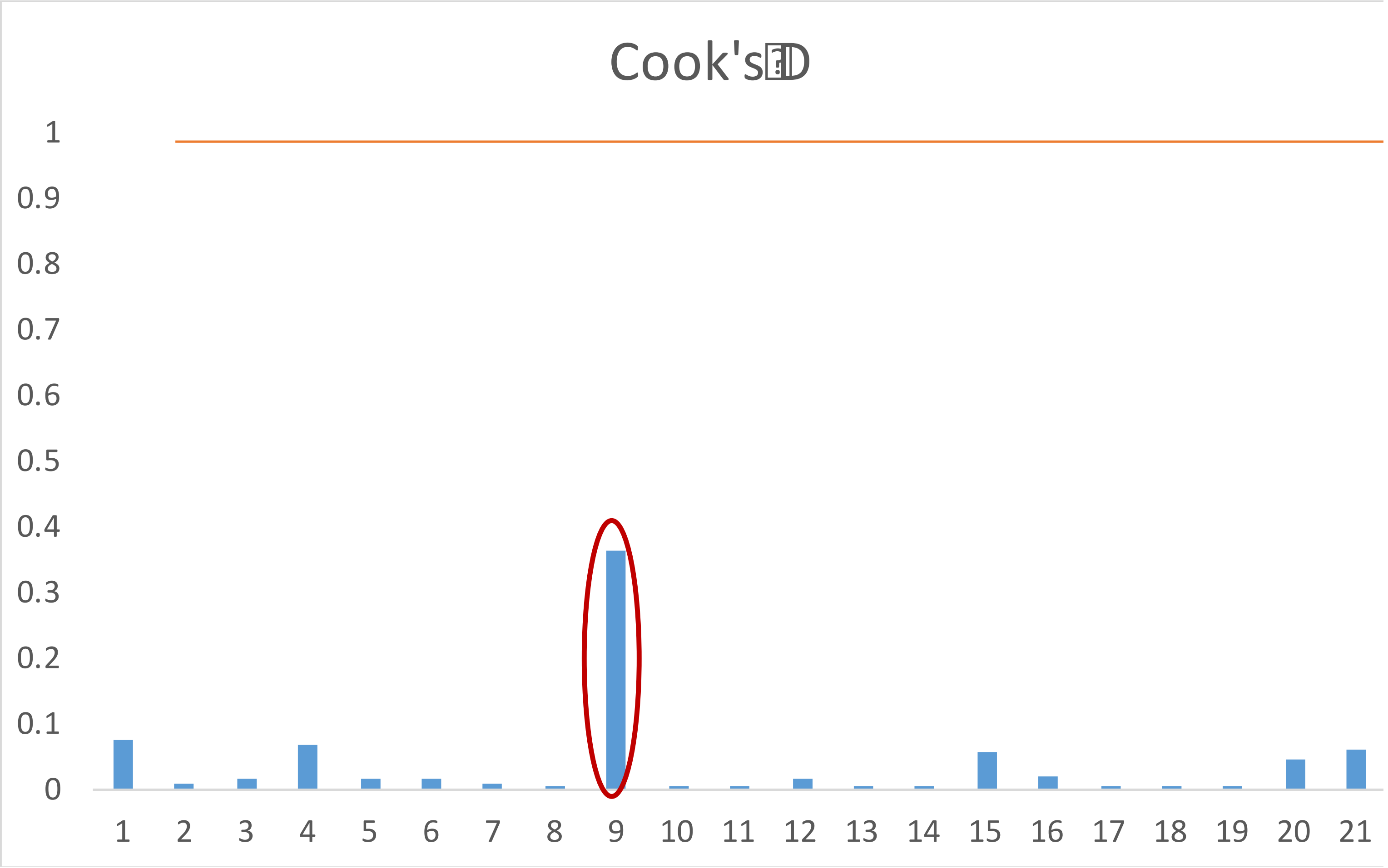


Influential Observations – Cook's D

Measures overall influence of an observation by seeing the impact on the regression coefficients when this observation is omitted. It accounts both for **leverage** and **residual**.

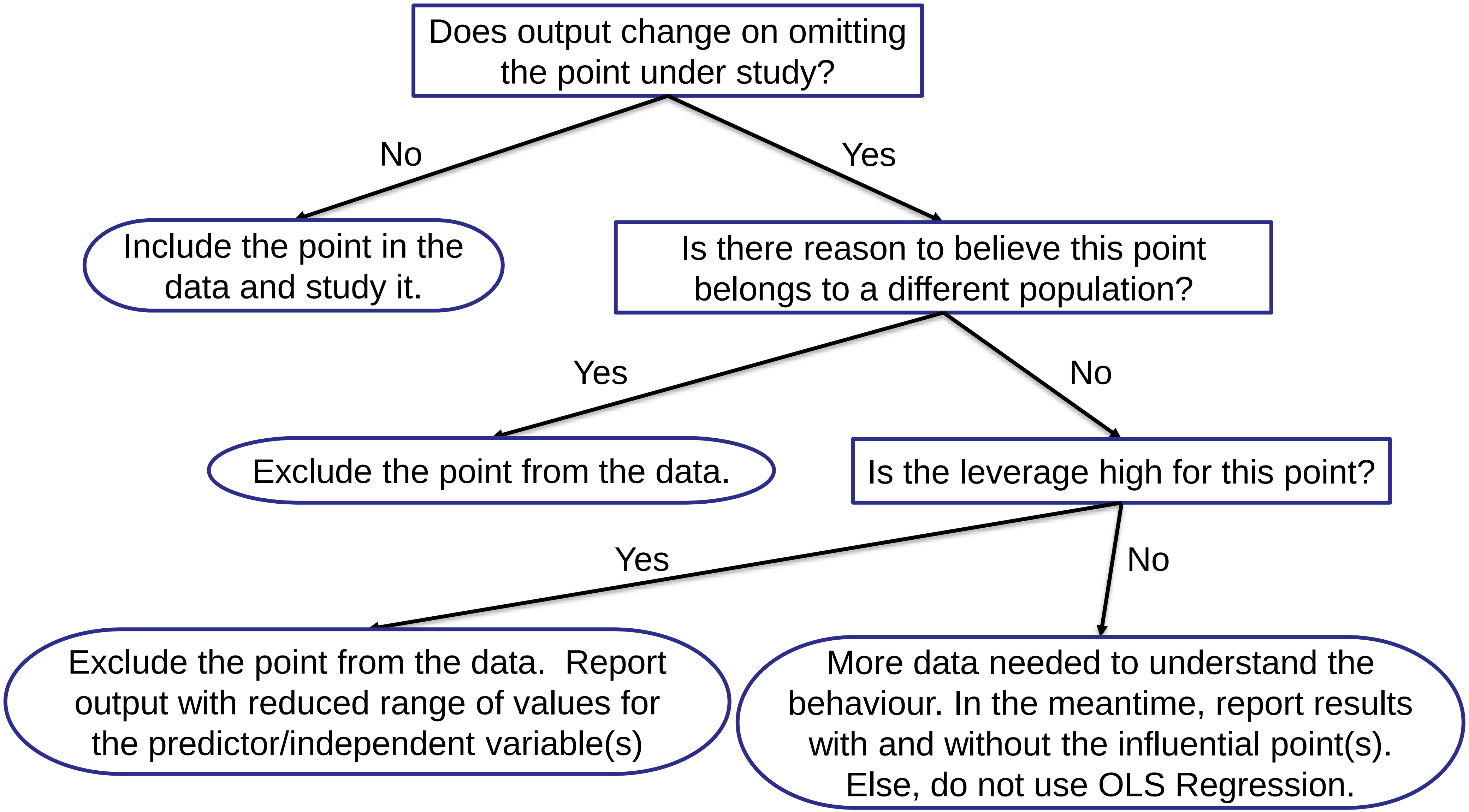
$$D_i = \frac{1}{p}(stdres_i)^2\left(\frac{h_i}{1-h_i}\right)$$

Influential Observations – Cook's D



CSE 7302c

Handling Influential Observations



Recall Day 1

End of Day 1

- Definition of Linear Equation – Y=a+bx
  - Independent Variable – x, Dependent Variable – y, Slope is b, y-intercept is a
- Line of Best Fit - **Least Squares Regression** or **Ordinary Least Squares Regression** or **OLS Regression**
  - $b = \frac{\sum((x-\bar{x})(y-\bar{y}))}{\sum(x-\bar{x})^2}$
  - The line of best fit <u>must</u> pass through $(\bar{x}, \bar{y})$. Substituting in the equation $\bar{y} = a + b\bar{x}$, we can find a.
- **Covariance .** $s_{xy}^2 = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$

  Tells you the direction of relationship between 2 variables
- **Correlation Coefficient-** $r = \frac{s_{xy}^2}{s_x s_y}$

  Tells you the direction AND strength of linear relationship between 2 variables
- **R^2.** $SST = SSR + SSE$

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

CSE 7302c

Recall Day 1

End of Day 1

- Assumptions of the Regression Model
  - R^2
  - **Residuals Analysis**
    - Is the Model Linear
    - The error terms are independent – More for time series
    - The error terms have constant variances (homoscedasticity as opposed to heteroscedasticity)
    - The error terms are normally distributed – Q-Q Plot
  - Hypothesis Tests
    - t-test for slope
    - Anova for Entire Model
- R-Squared and Significance - Caution

Simple Linear Regression - Steps

End of Day 1

Get familiar with data
- Plots
- Descriptive stats

Formulate a linear model and fit to data
- Do regression

Check model and assumptions
- Look at residual plots
- Look at unusual observations
- Look at R-Squared
- Look at p-values

Inadequate fit

Good fit

Report results and equation
- Make predictions for values of interest

CSE 7302c

Recall for Day 1

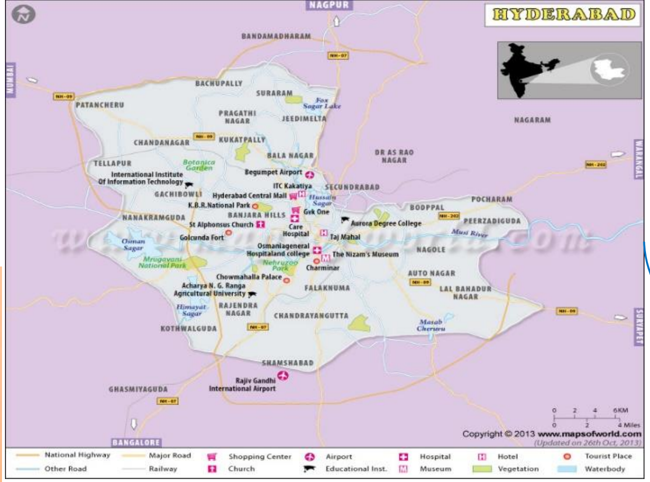
End of Day 1

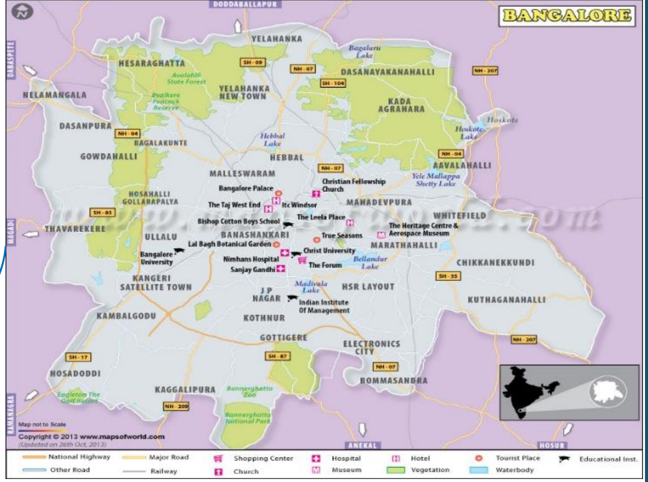
- Influential Data Points
  - **Influence** is a function of **leverage** and **distance (or 'residuality' or 'outlierness')**.
  - Cook's Distance (Cook D) Cook's D is a measure of the influence
  - If Cook's D of any observation (D_i) > 1, that observation can be considered as having too much influence, but investigate values greater than 0.5 also

CSE 7302c

Inspire...Educate...Transform.

WISHES YOU





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