Inspire...Educate...Transform.

Inspire…Educate…Transform.

Statistics and Probability in Decision Modeling

Linear Regression

Dr. Sridhar Pappu
Executive VP – Academics, INSOFE

December 31, 2017

CORRELATION, COVARIANCE AND REGRESSION

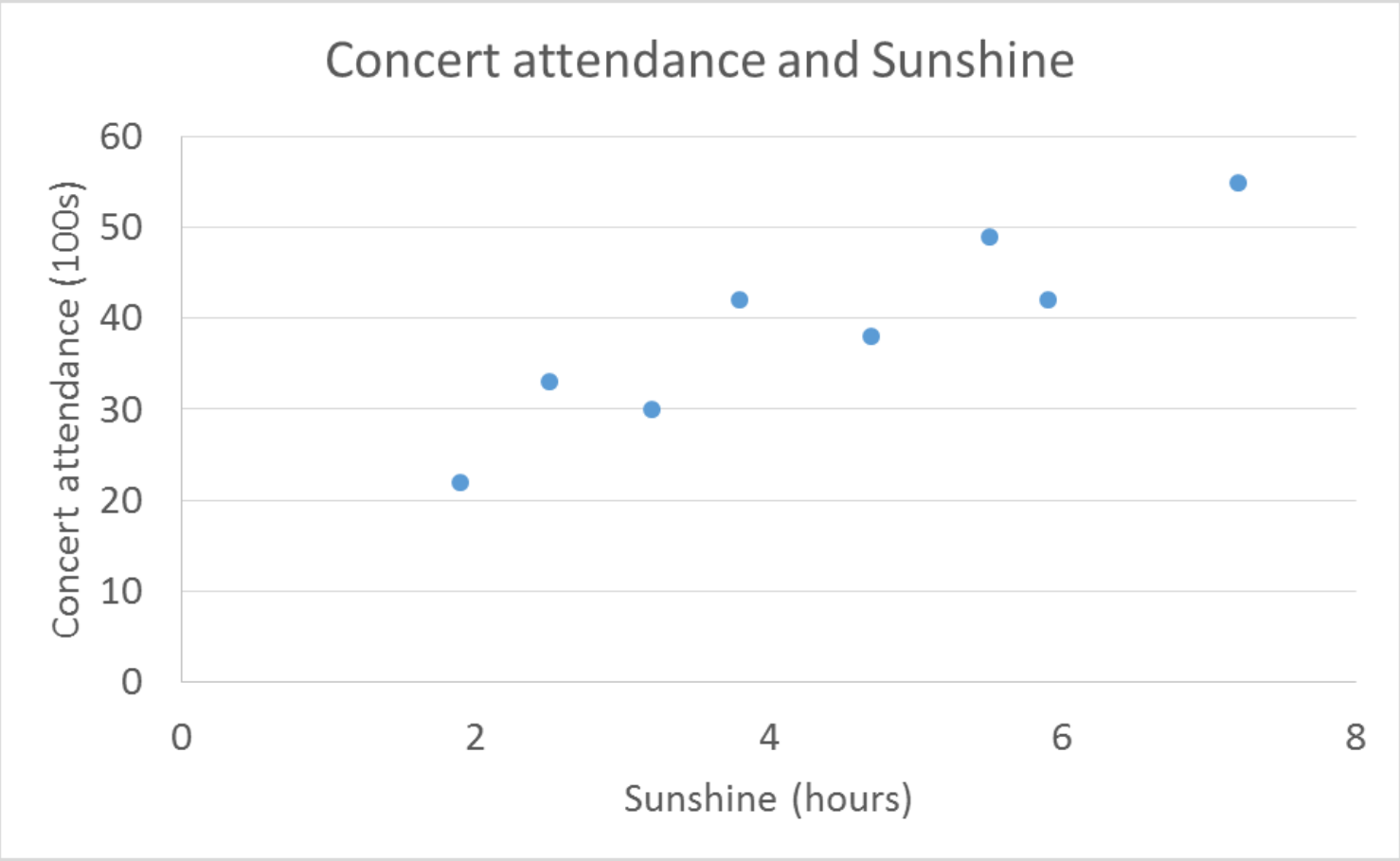
Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

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

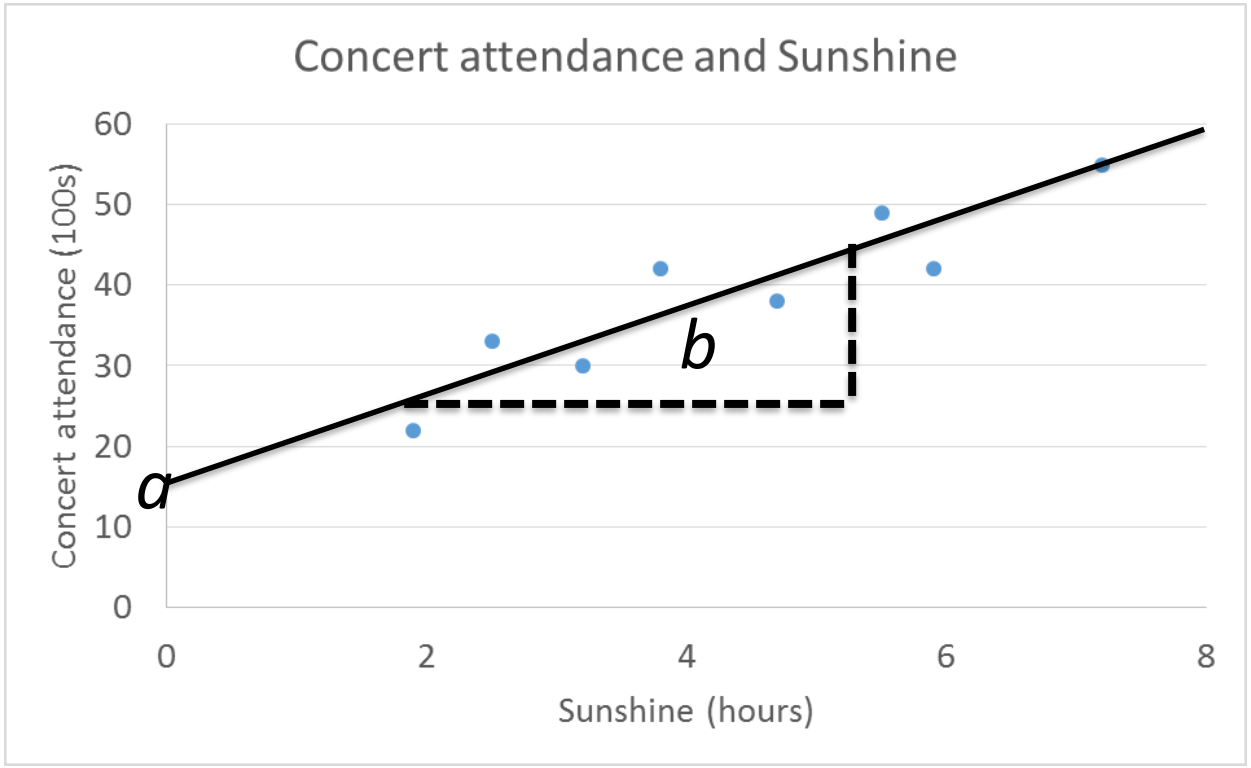
- The band makes a loss if less than 3500 people attend.
- Based on predicted hours of sunshine, can we predict ticket sales?
- Are sunshine and concert attendance correlated?

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.

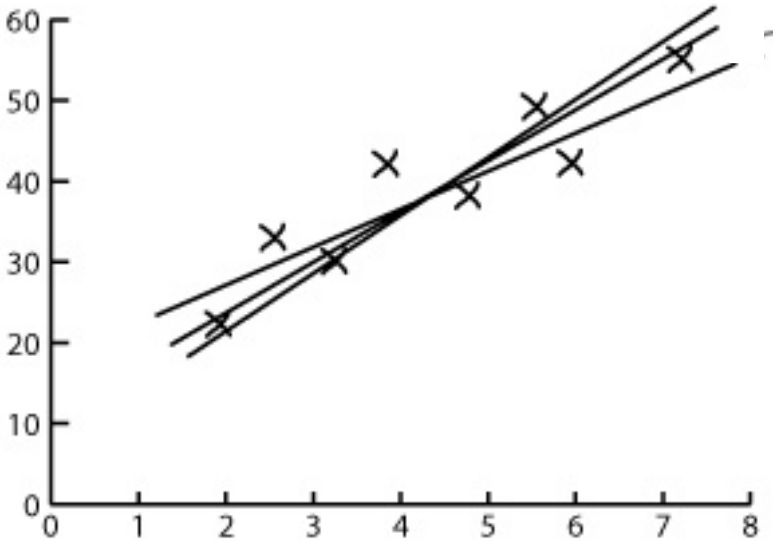


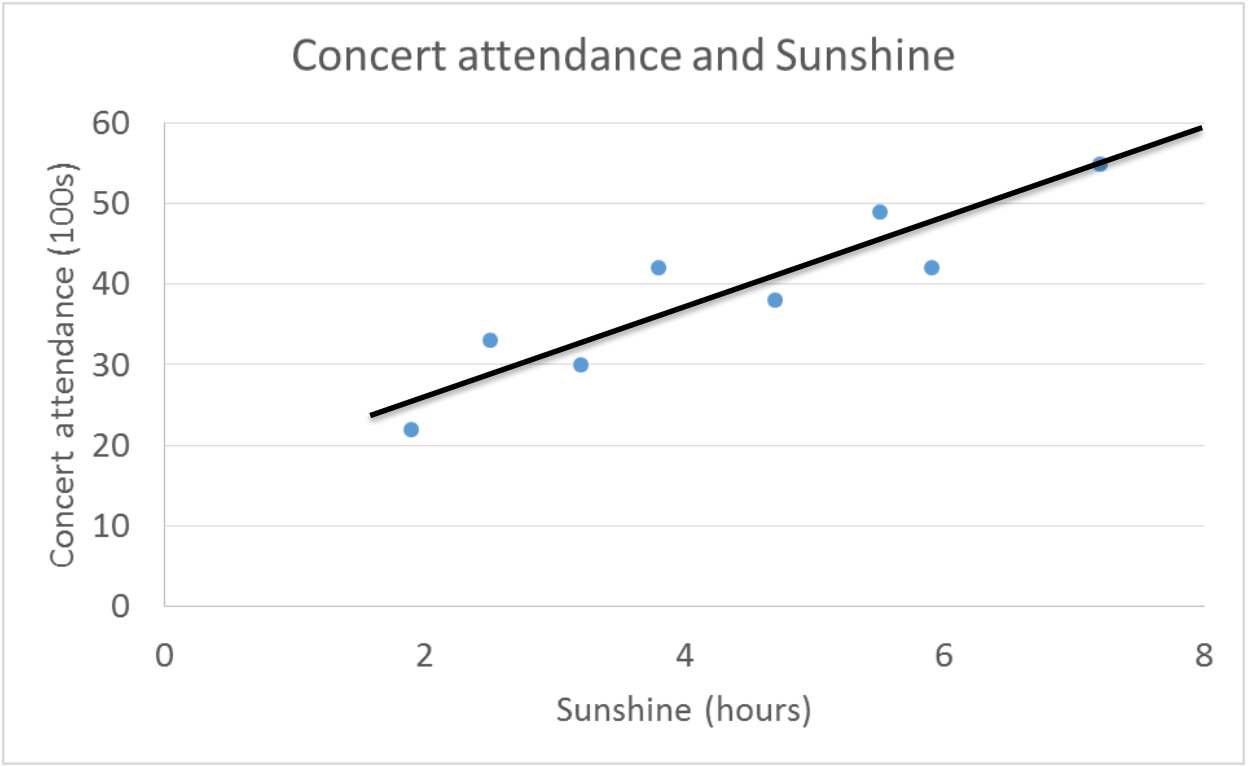
$y = a + bx$

CSE 7302c

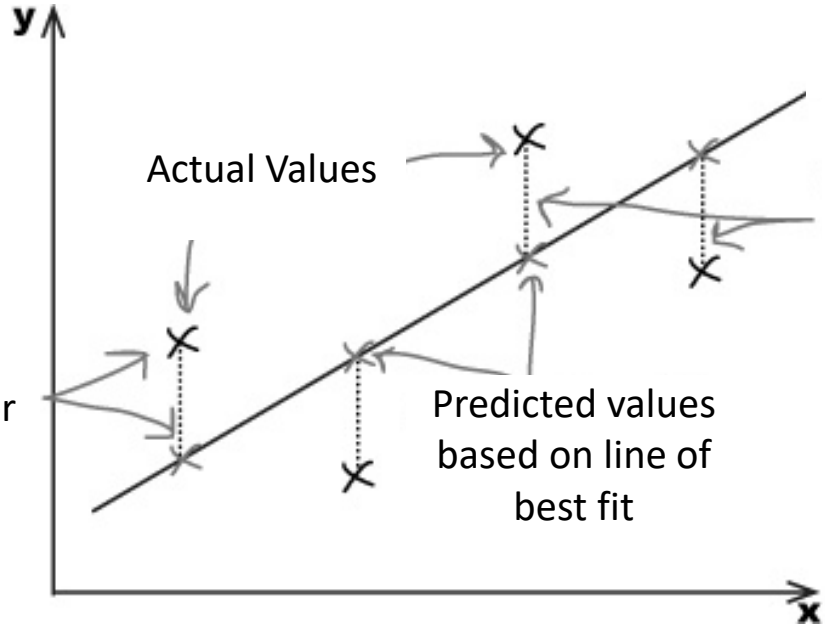
Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

- Line of best fit





We need to minimize errors.



Line of best fit is the line that minimizes all the distances (residuals) between the Actual and Estimated 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**.

CSE 7302c

We need to minimize errors.

Just as we did when finding variance, we find the **sum of squared errors** or SSE. *Note in variance calculations, we subtract mean, $\bar{y}$, not $\hat{y}_i$.*

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

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

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

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

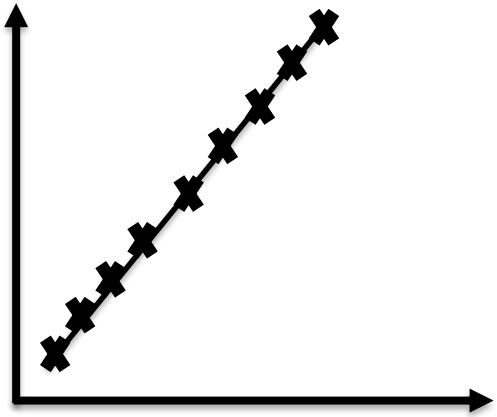
The value of *b*, the slope, that minimizes the SSE is given by $b = \frac{\sum((x-\bar{x})(y-\bar{y}))}{\sum(x-\bar{x})^2}$

How do you calculate *a* in y = *a* + *b*x?

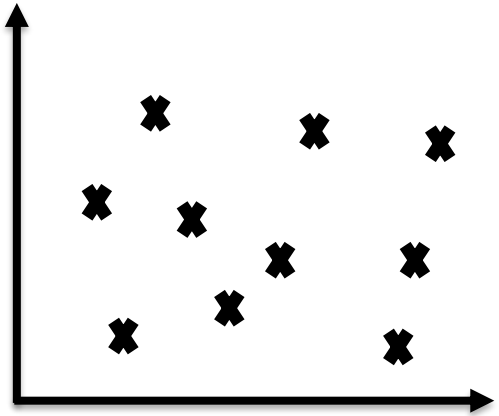
The line of best fit must pass through $(\bar{x}, \bar{y})$. Substituting in the equation $\bar{y} = a + b\bar{x}$, we can find *a*.

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**.

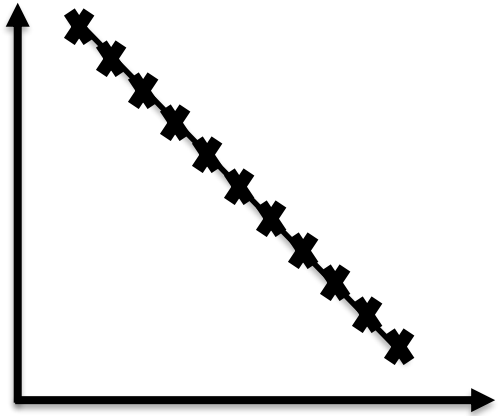
CSE 7302c

Correlation Coefficient

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

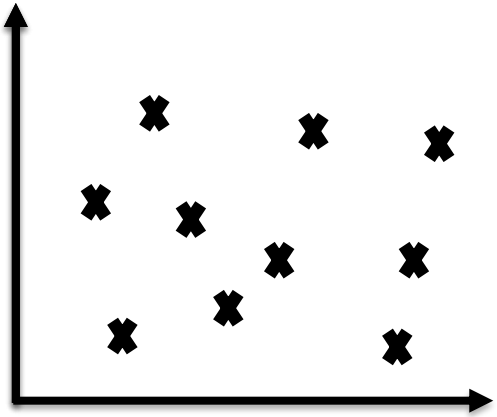
r = 1

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

CSE 7302c

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 covariance*.

* 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

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

- 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)

CSE 7302c

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**

CSE 7302c

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.

CSE 7302c

Coefficient of Determination

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.

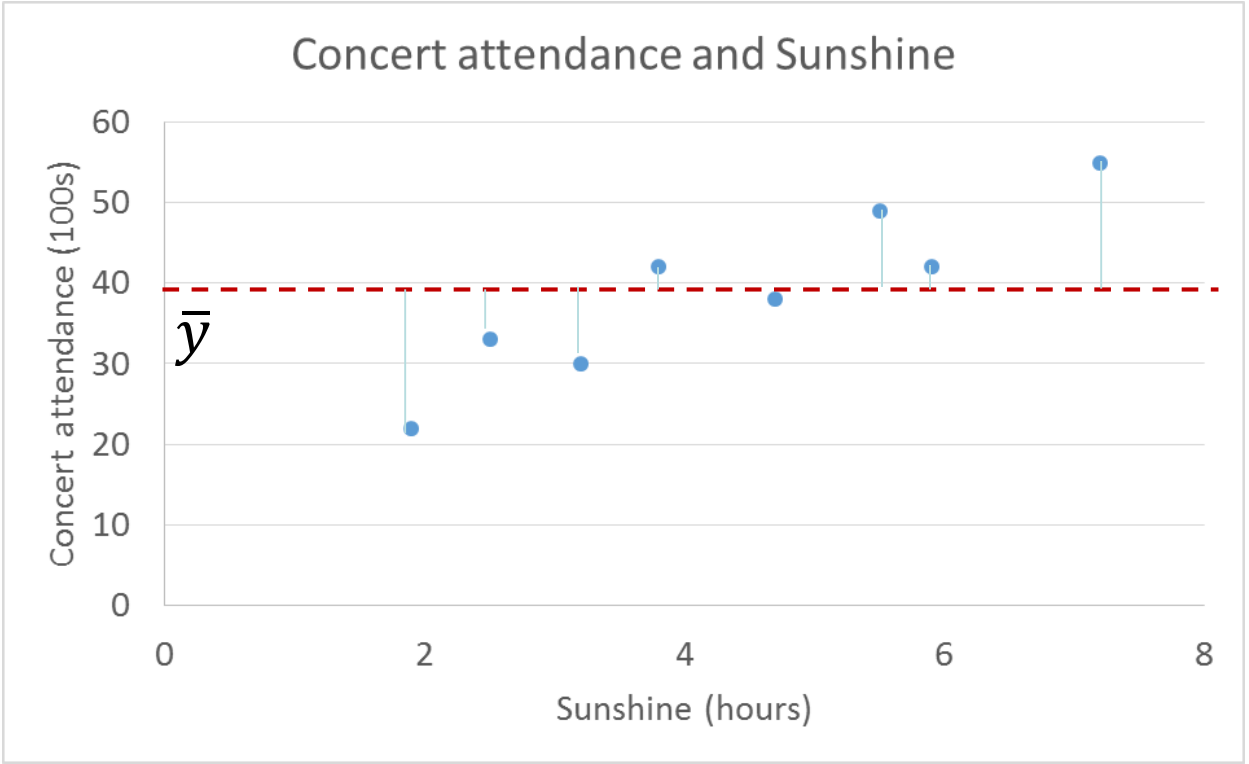
Usually, r^2 is between these two extremes.

CSE 7302c

Coefficient of Determination

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.

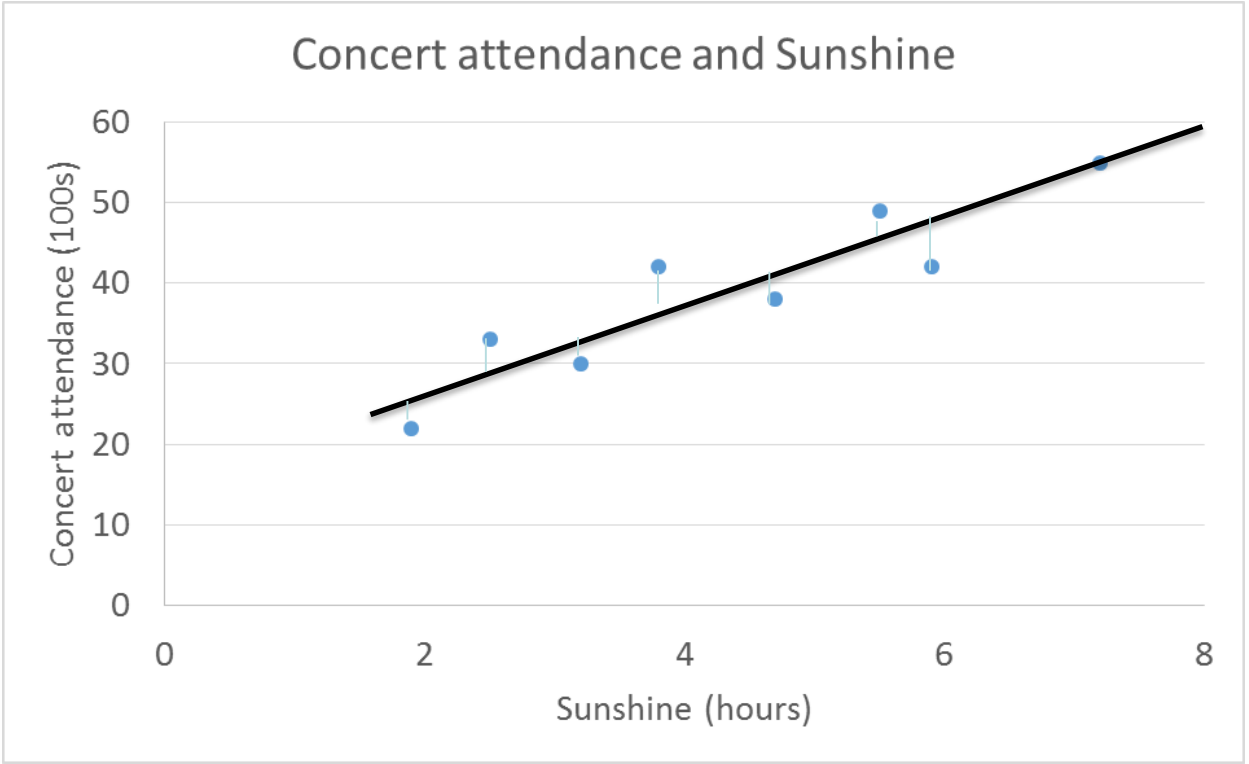
$$SST = \sum (y_i - \bar{y})^2$$



Coefficient of Determination

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.

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

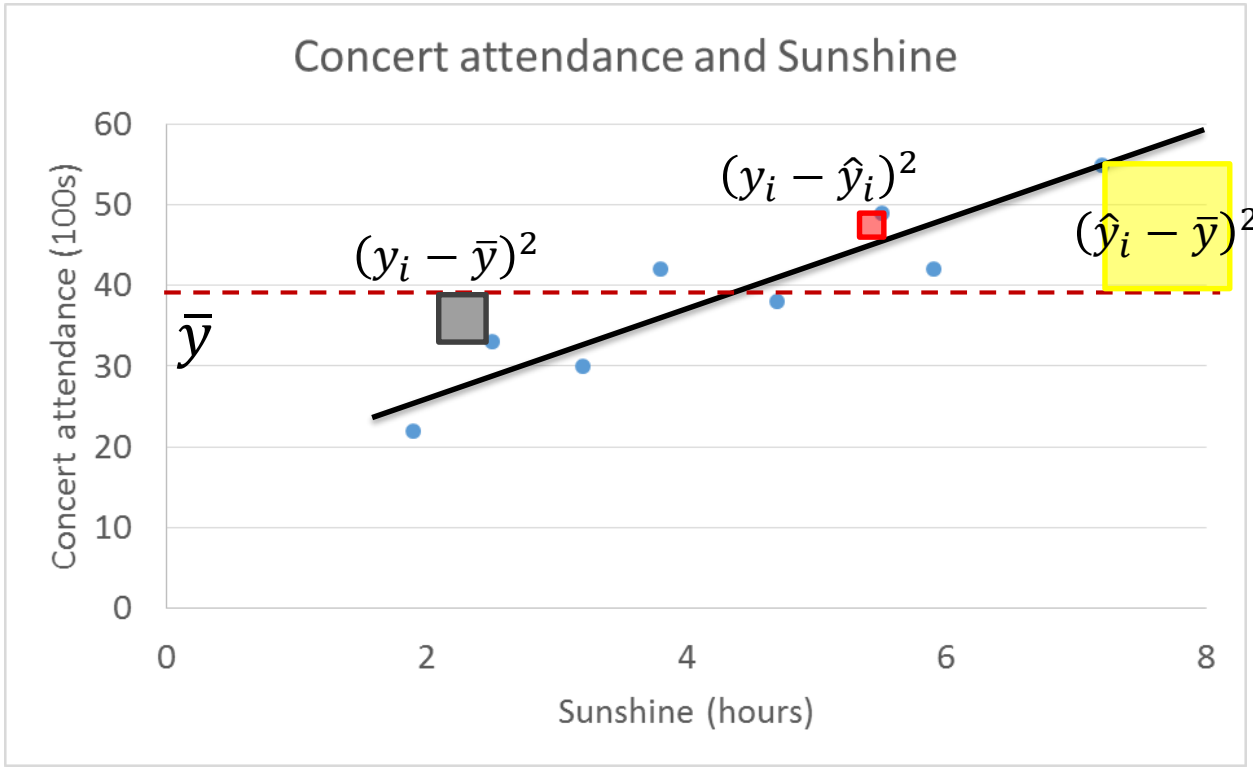


CSE 7302c

Coefficient of Determination

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

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



Considering sunshine

Not considering sunshine

CSE 7302c

Sums of Squares – Other Forms

$$SST = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$SSR = b(\sum xy - \frac{\sum x \sum y}{n})$$

$$SSE = SST - SSR$$

CSE 7302c

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			
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			
Mean	**7.74**	**227.08**			
StDev	**0.22**	**6.07**			

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

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

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

CSE 7302c

Welcome to the Learning Models

- Linear regression: A regression model (class variable is **numeric**)

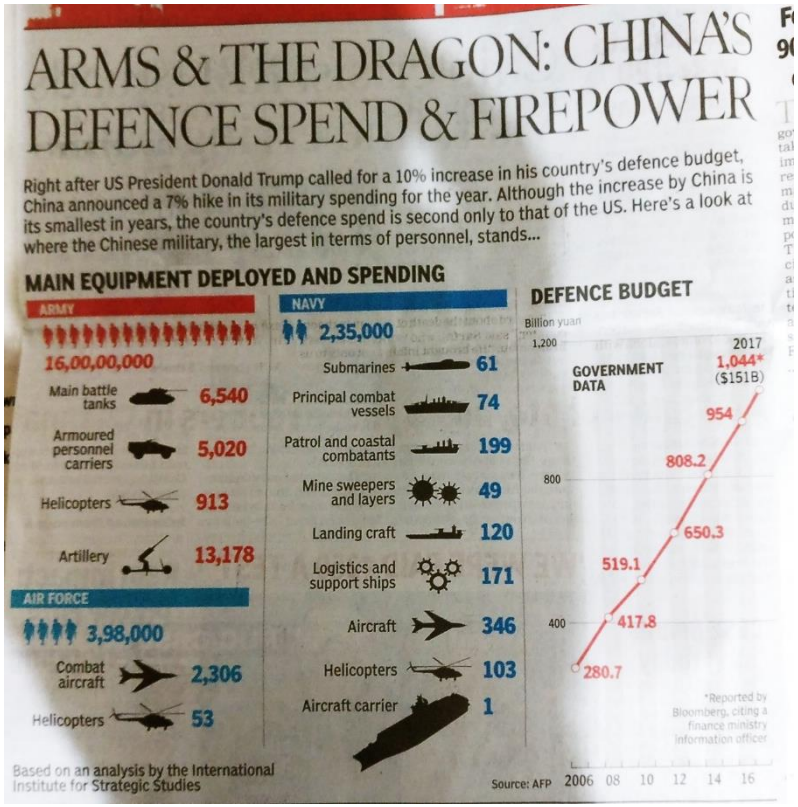
- Logistic regression: A classification model (class variable is **categorical**)

CSE 7302c

Linear Regression

CSE 7302c

Linear Regression



ARMS & THE DRAGON: CHINA'S DEFENCE SPEND & FIREPOWER

Right after US President Donald Trump called for a 10% increase in his country's defence budget, China announced a 7% hike in its military spending for the year. Although the increase by China is its smallest in years, the country's defence spend is second only to that of the US. Here's a look at where the Chinese military, the largest in terms of personnel, stands...

MAIN EQUIPMENT DEPLOYED AND SPENDING

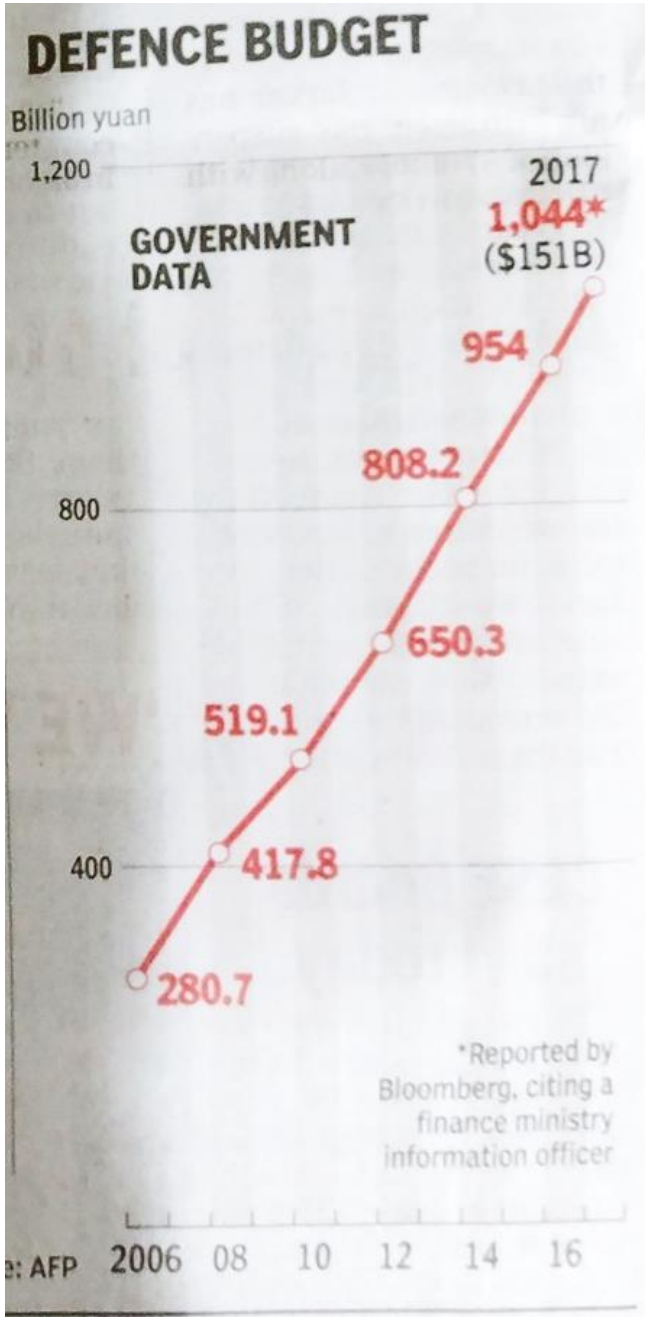
ARMY
16,00,00,000
Main battle tanks 6,540
Armoured personnel carriers 5,020
Helicopters 913
Artillery 13,178

AIR FORCE
3,98,000
Combat aircraft 2,306
Helicopters 53

NAVY
2,35,000
Submarines 61
Principal combat vessels 74
Patrol and coastal combatants 199
Mine sweepers and layers 49
Landing craft 120
Logistics and support ships 171
Aircraft 346
Helicopters 103
Aircraft carrier 1

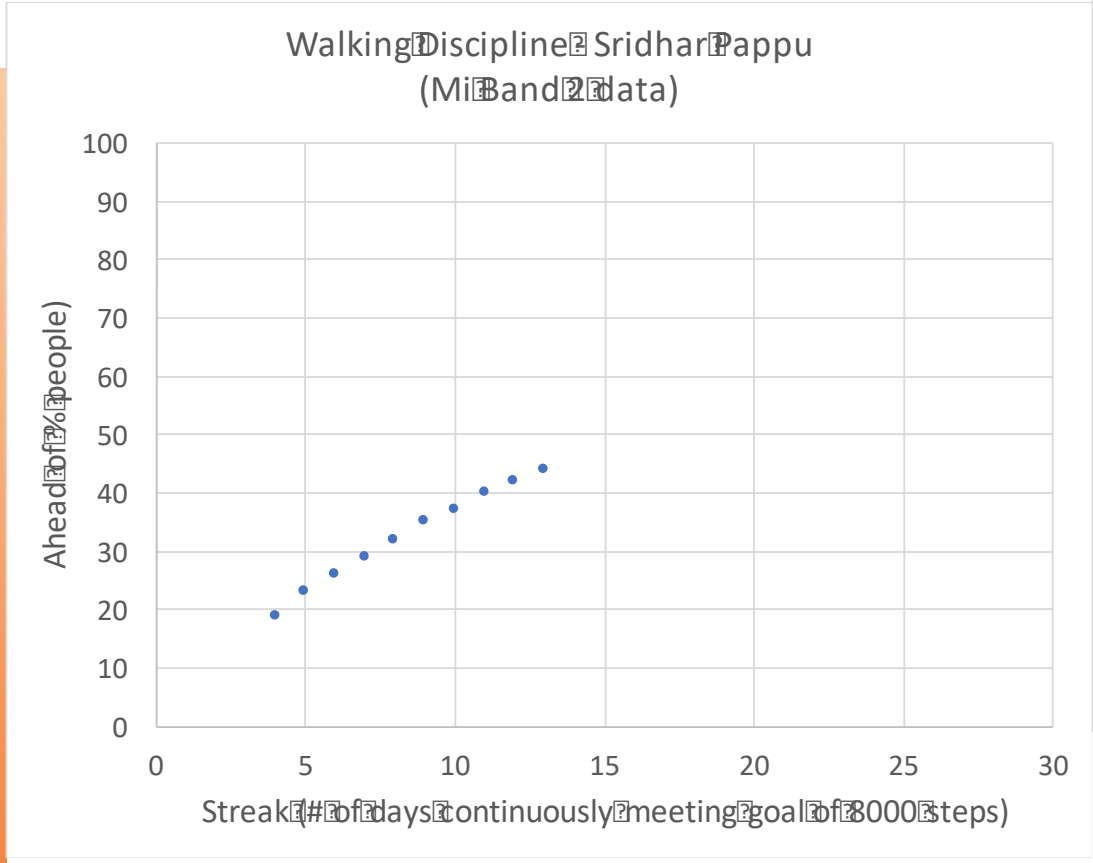
DEFENCE BUDGET
Billion yuan
1,200
800
400
GOVERNMENT DATA
2017 1,044* ($151B)
954
808.2
650.3
519.1
417.8
280.7
*Reported by Bloomberg, citing a finance ministry information officer
2006 08 10 12 14 16

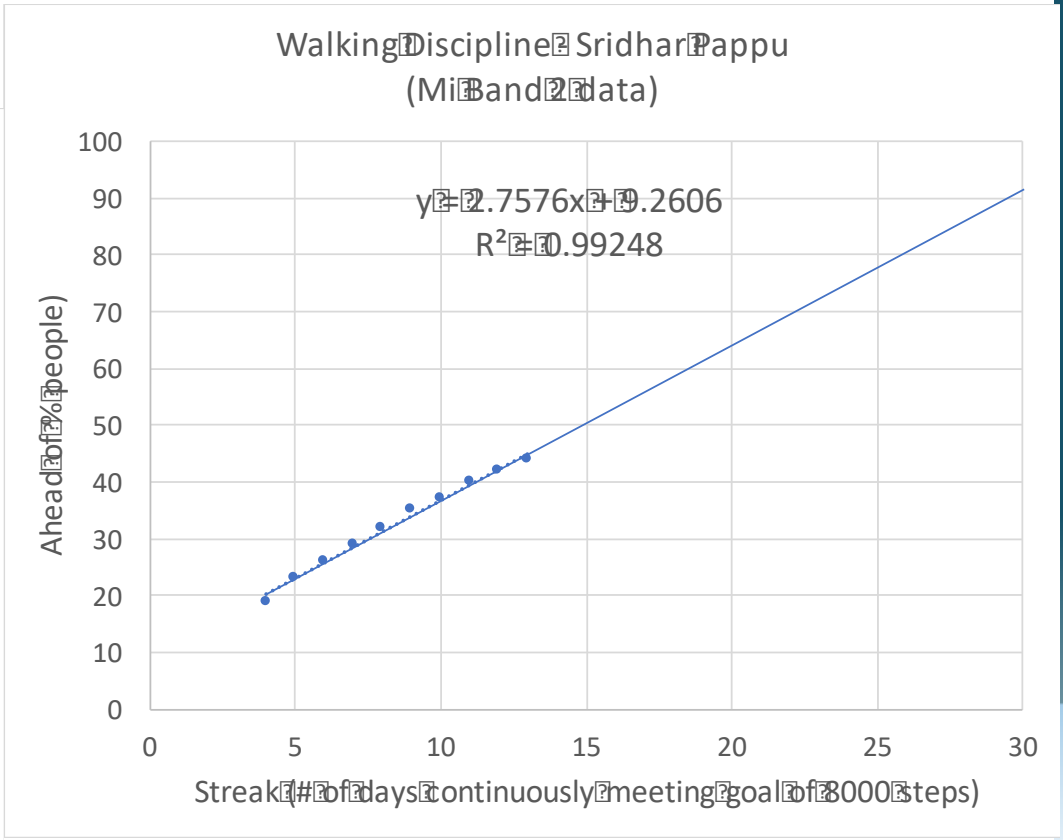
Based on an analysis by the International Institute for Strategic Studies
Source: AFP



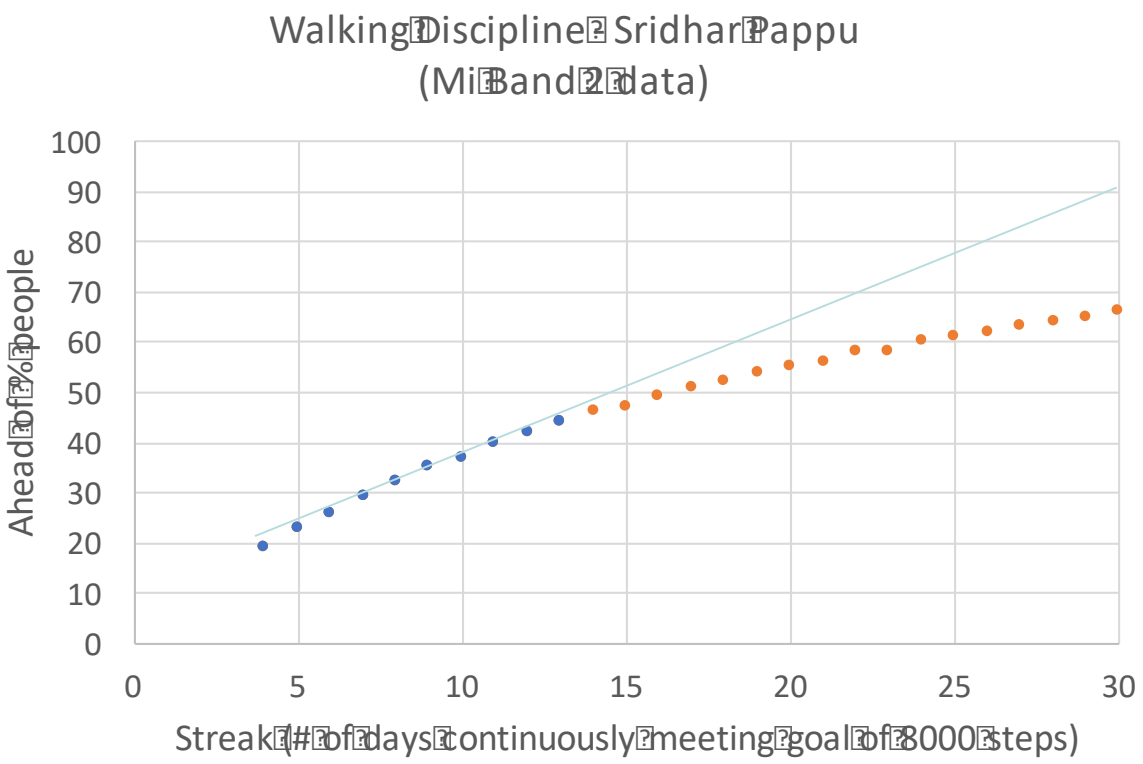
CSE 7302c

Linear Regression



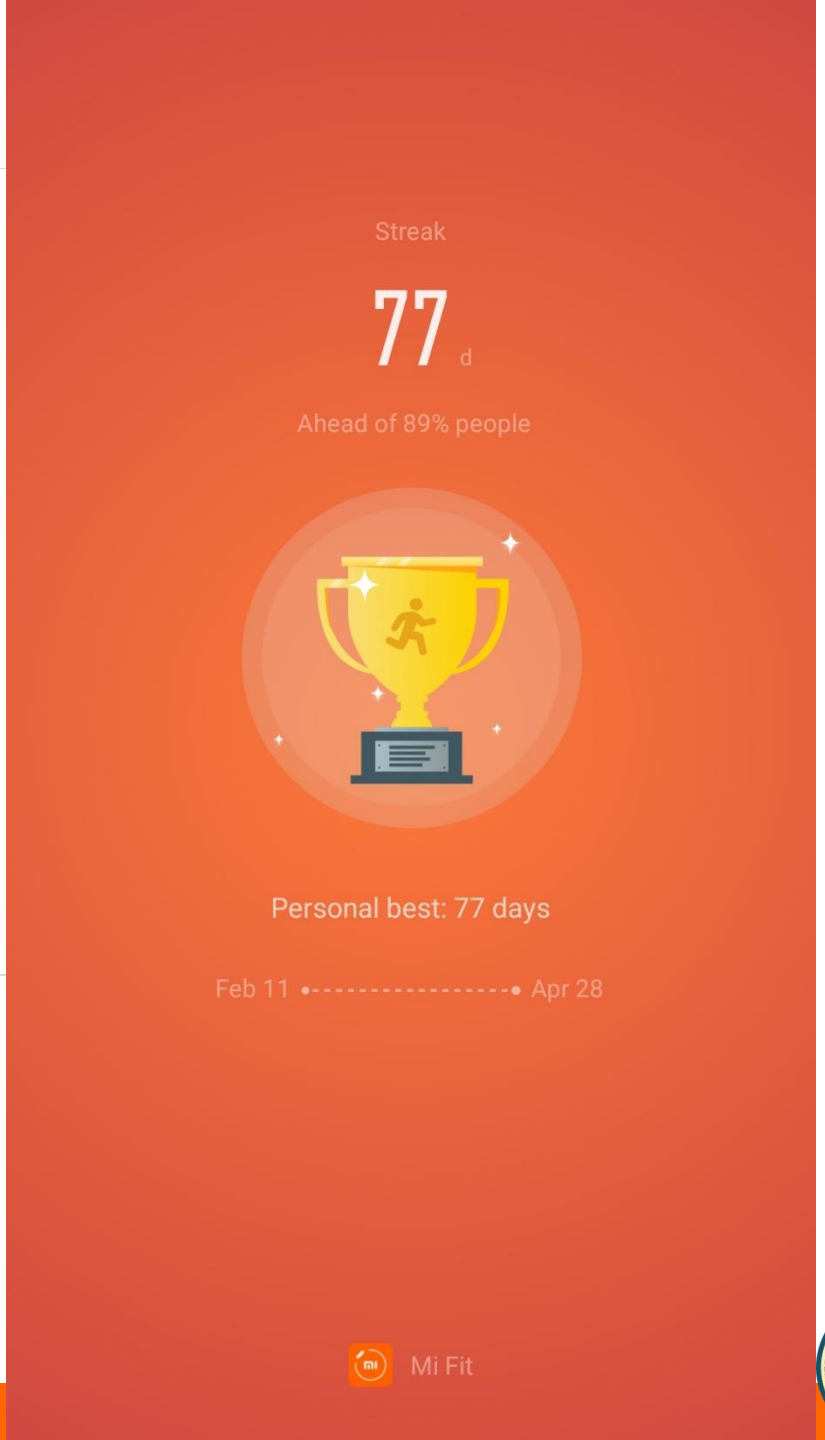


Linear Regression



Be careful when extrapolating.

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

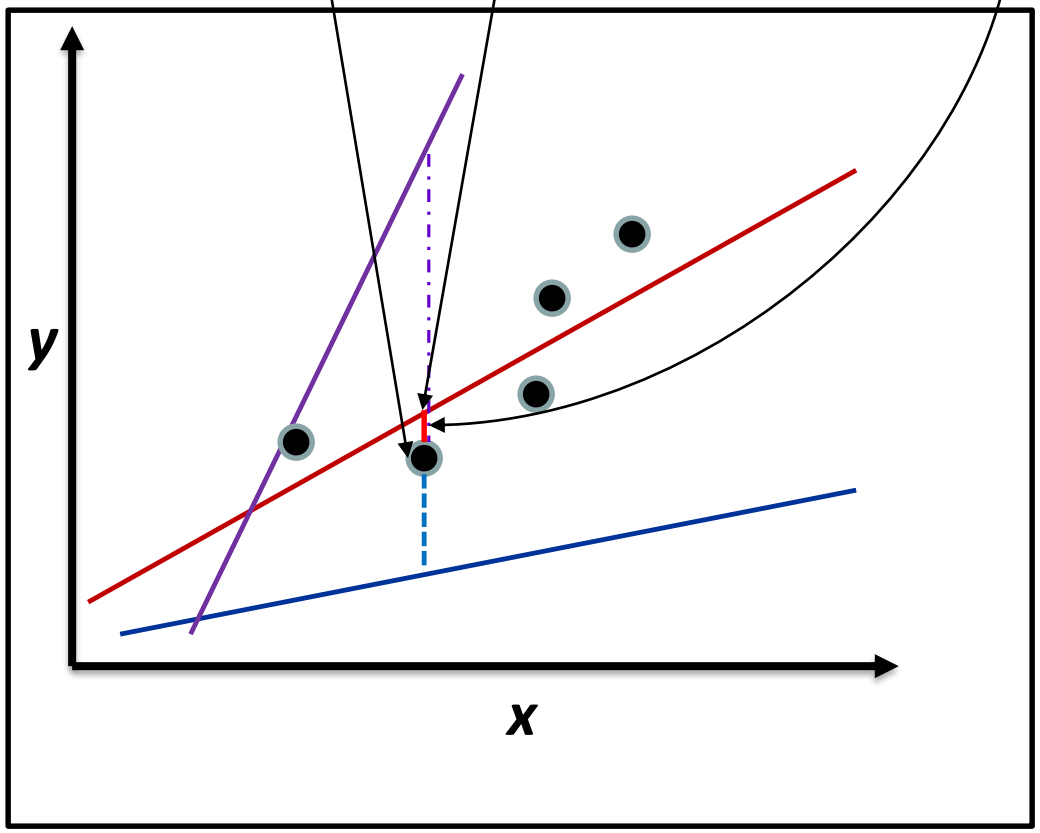


How to Pick the Best Model?

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

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

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



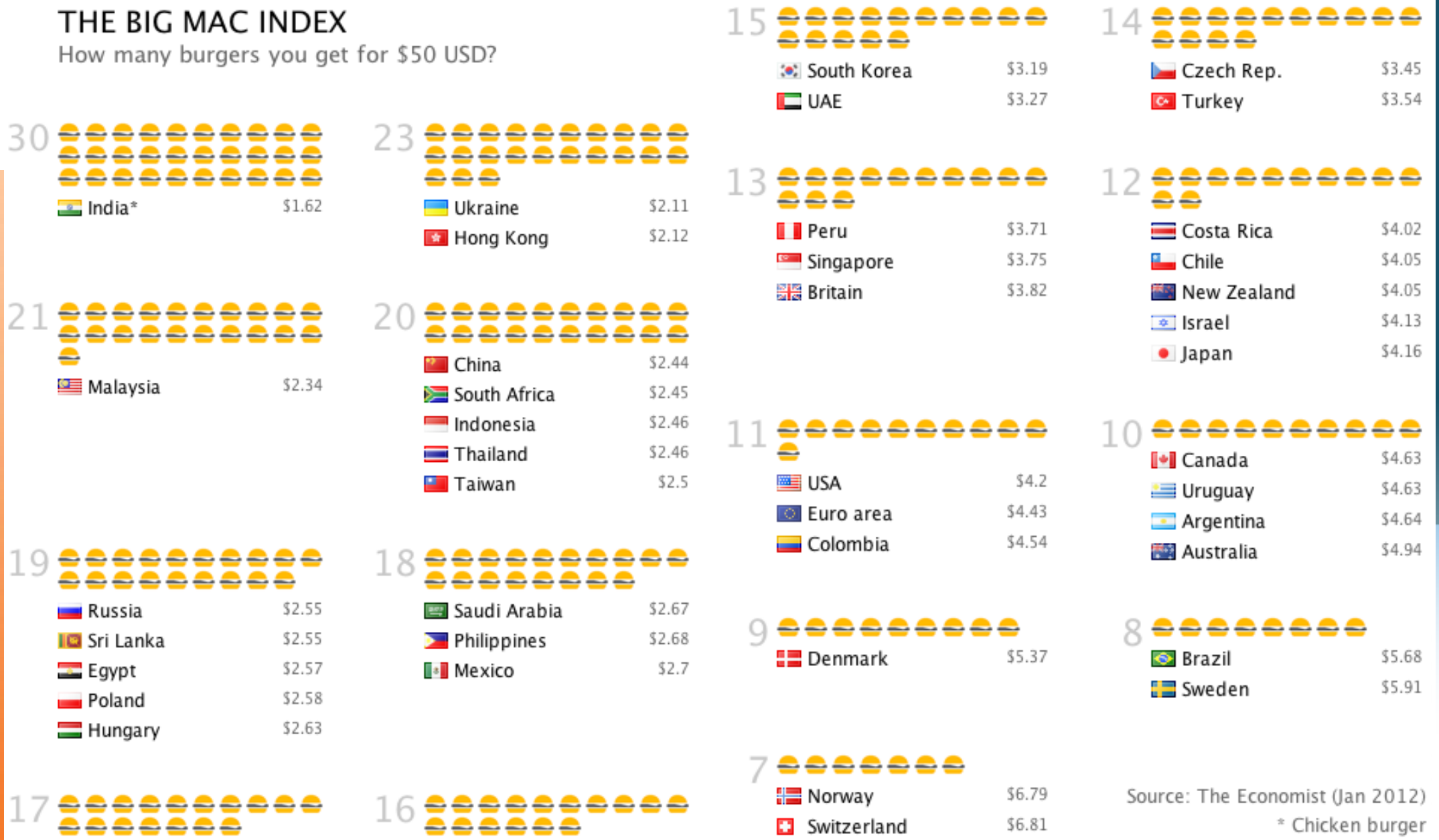
The lines whose residual error on all points is the least is the best line.

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

CSE 7302c

THE BIG MAC INDEX

How many burgers you get for $50 USD?

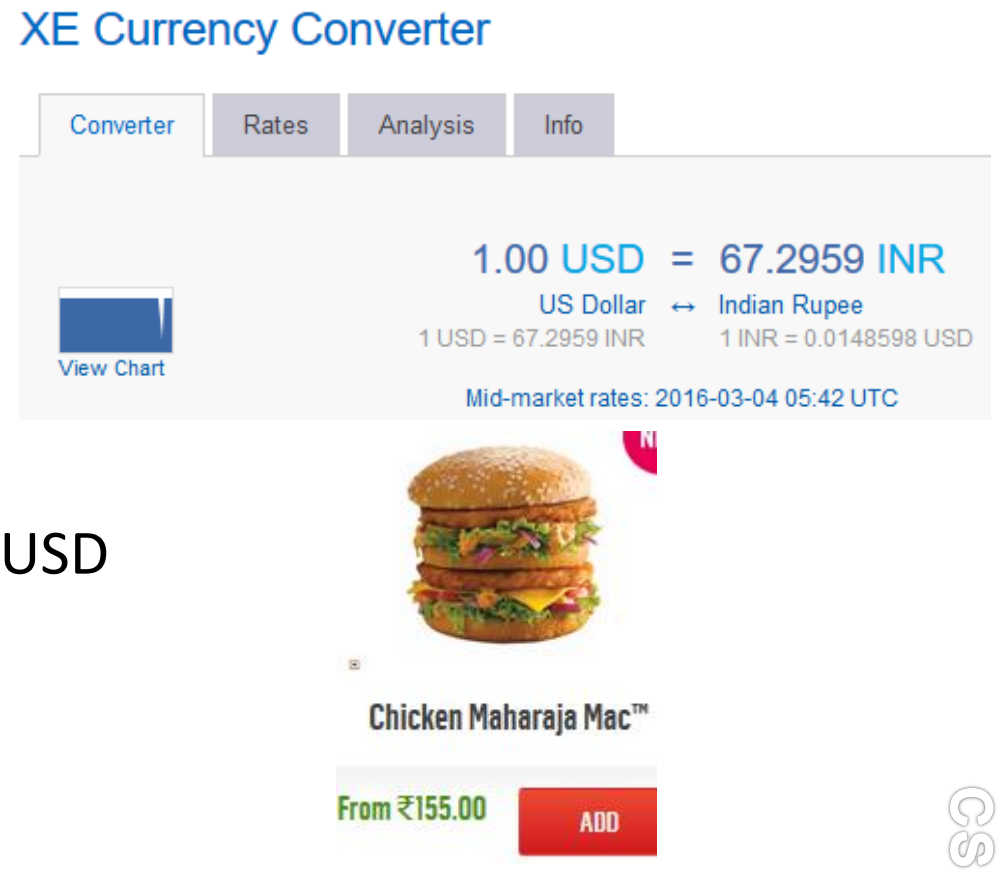


Source: The Economist (Jan 2012)

* Chicken burger

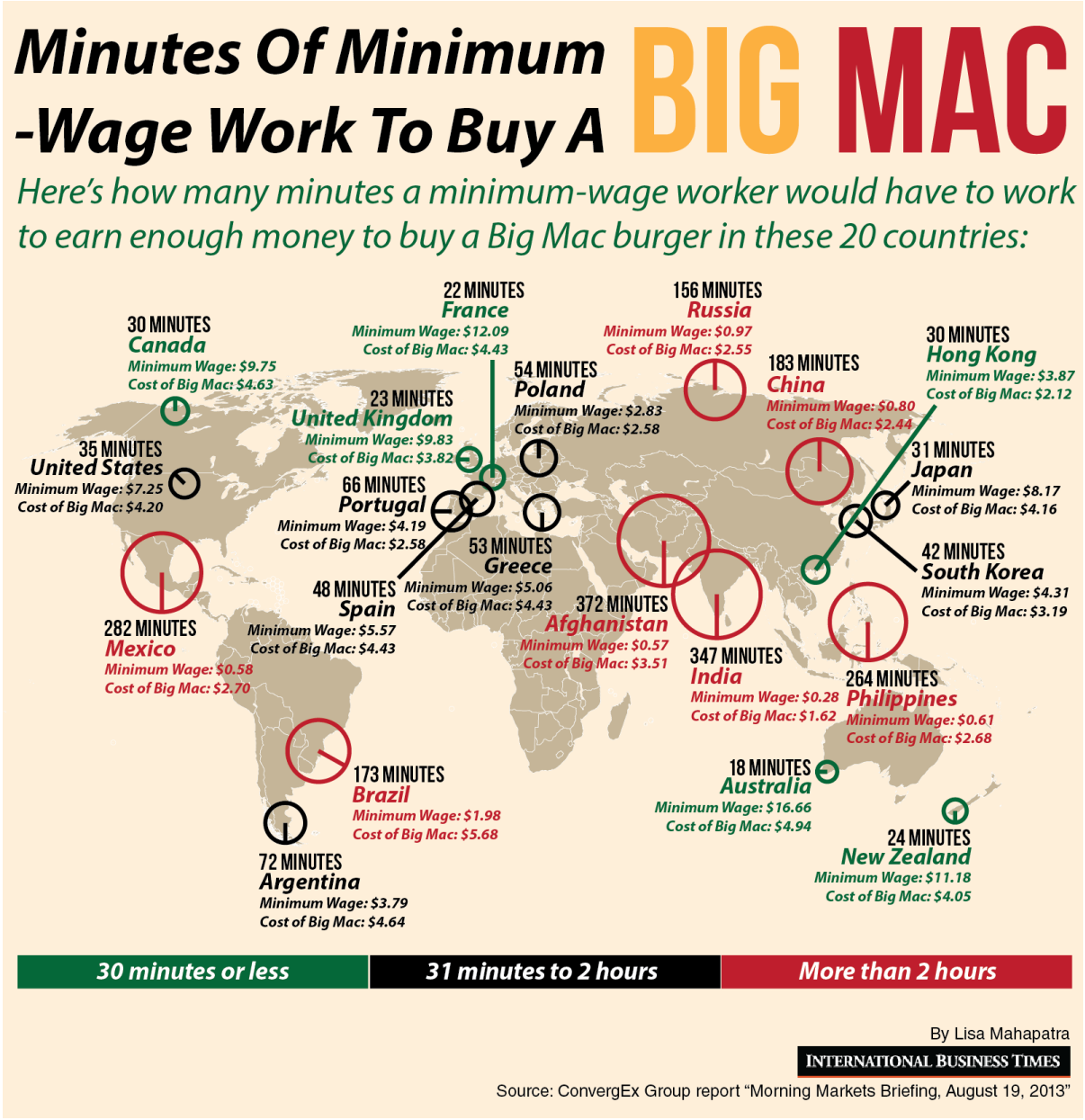
Burgernomics: Overvalued or Undervalued Currencies?

- Big Mac price in the US: $ 4.93
- Maharaja Mac price in India: Rs 155
- Implied PPP 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



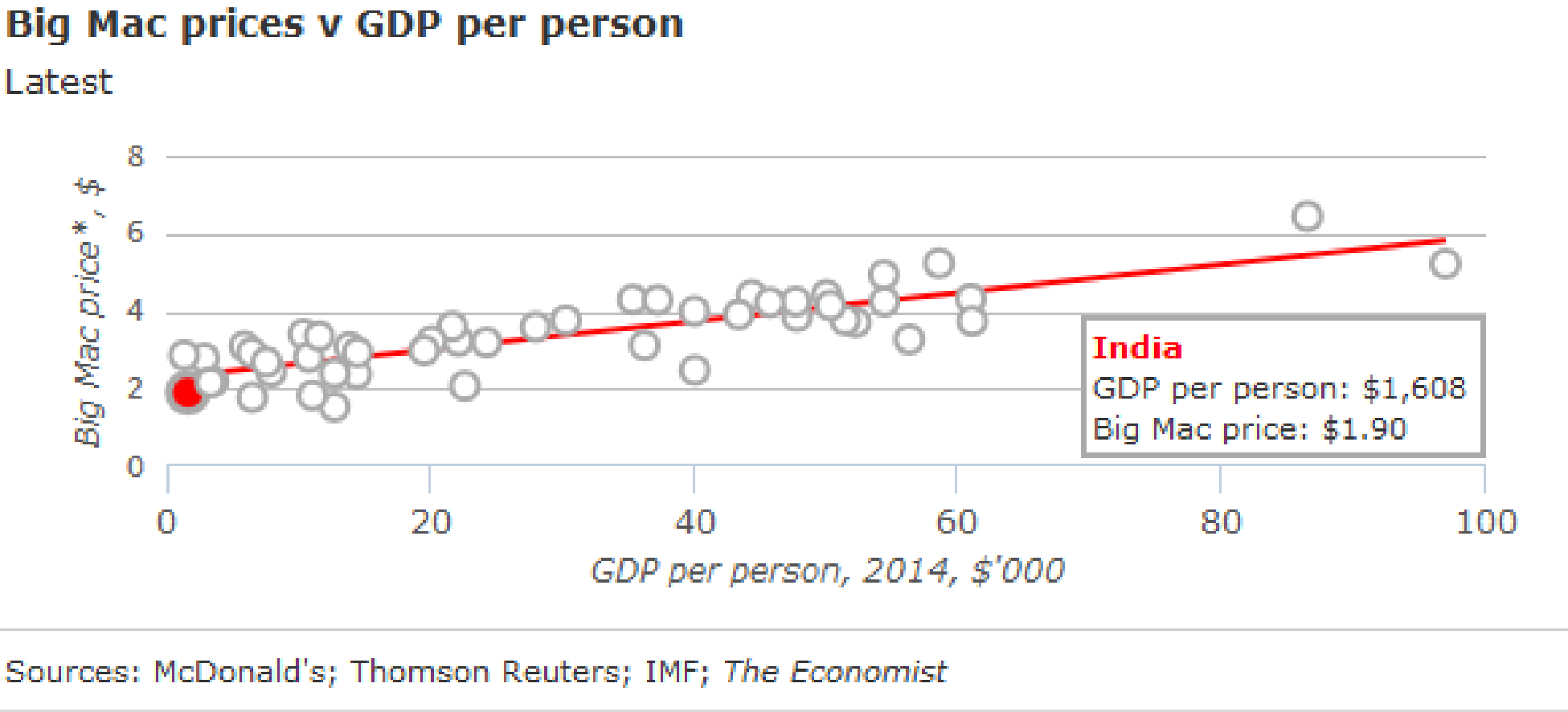
Global prices for a Big Mac in July 2016 based on a survey conducted in January 2016 by IMF, McDonald's, Thomson Reuters and The Economist

Burgernomics by UBS Wealth Management Research



CSE 7302c

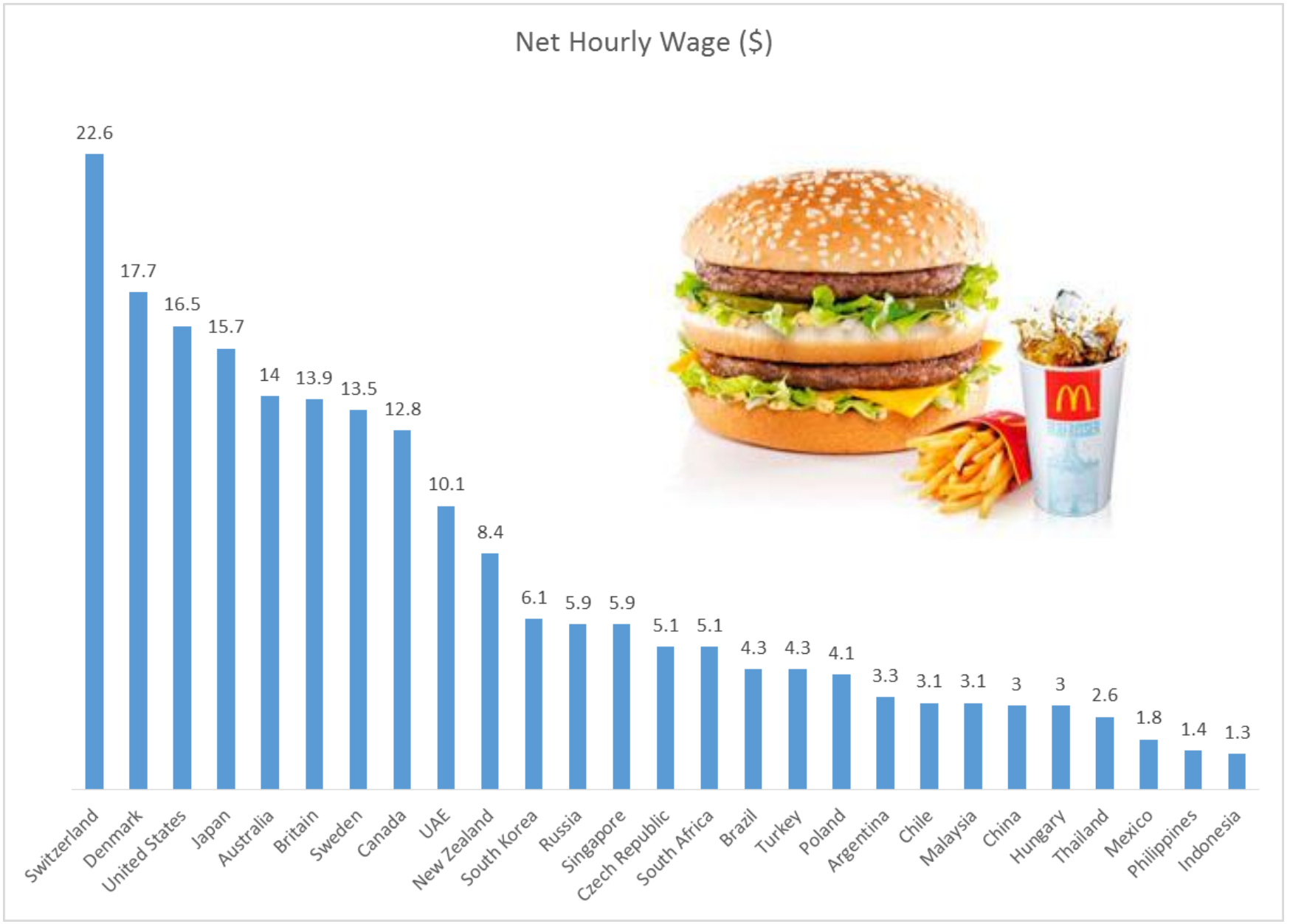
Burgernomics



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

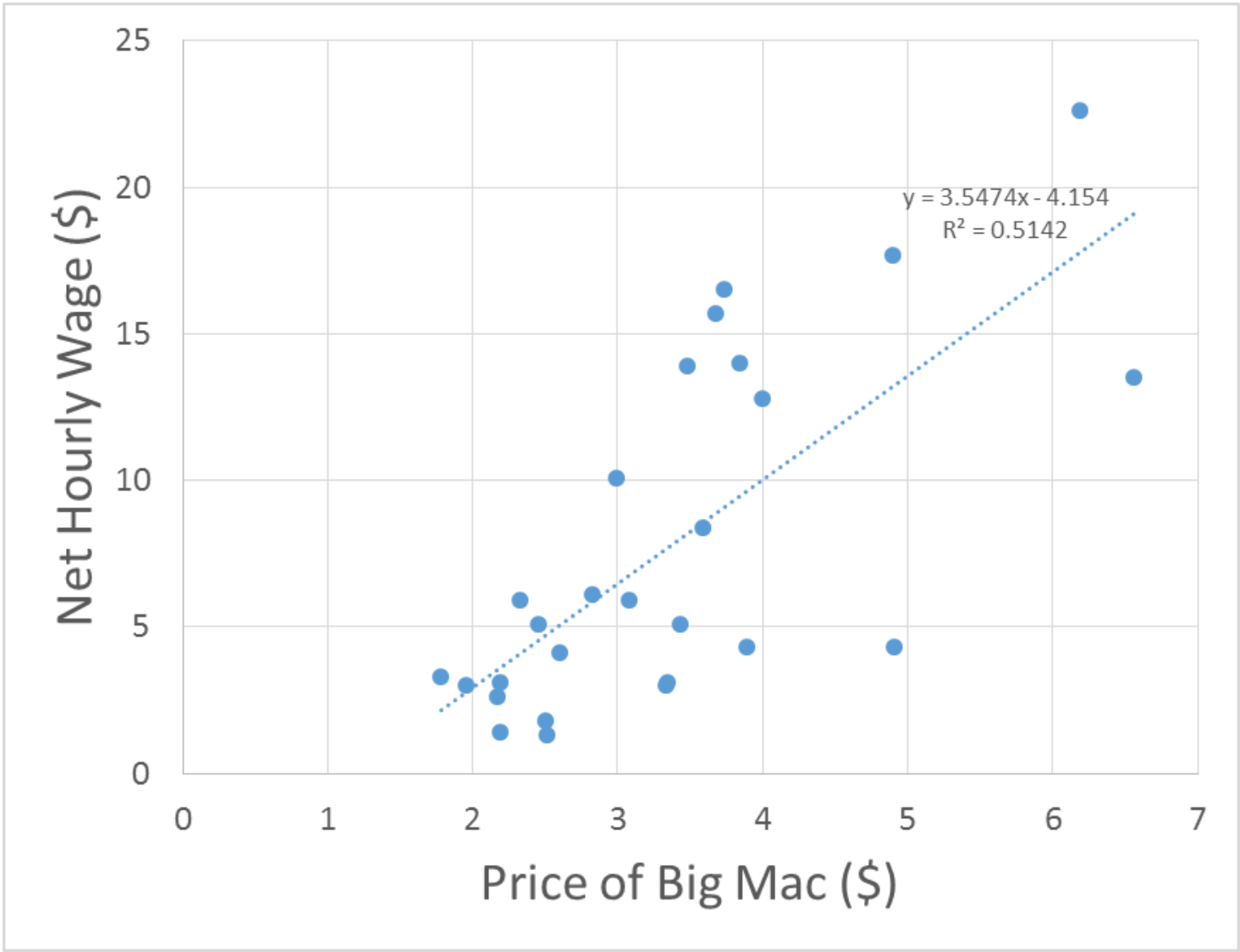
CSE 7302c

Determining the Equation of the Regression Line - Excel



CSE 7302c

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

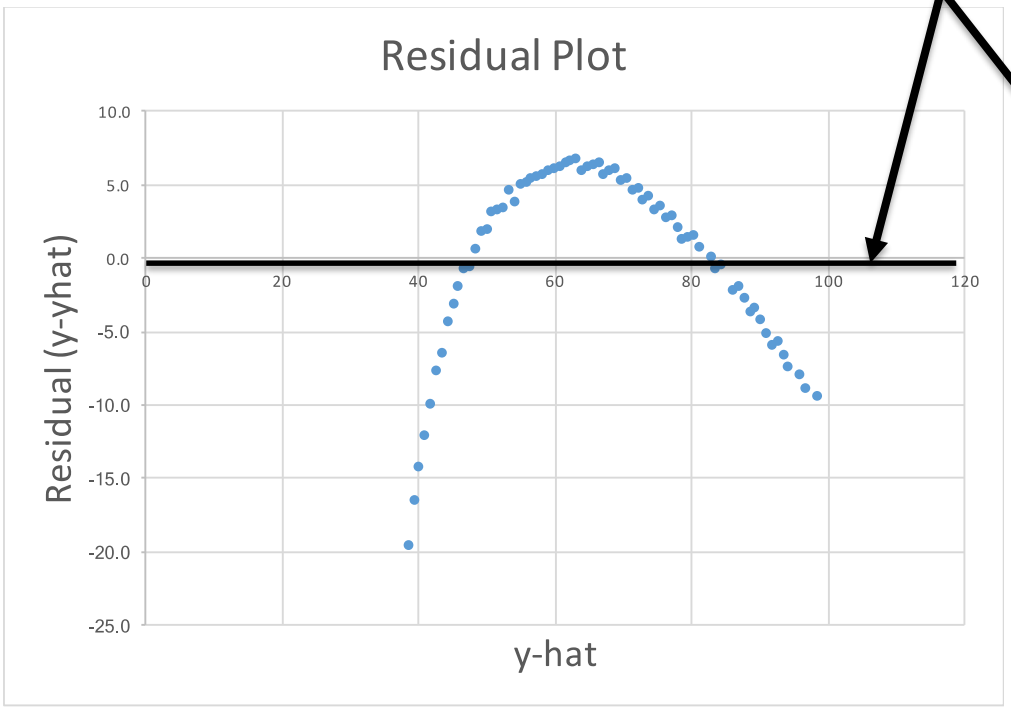
WAYS OF TESTING HOW WELL THE REGRESSION LINE FITS DATA

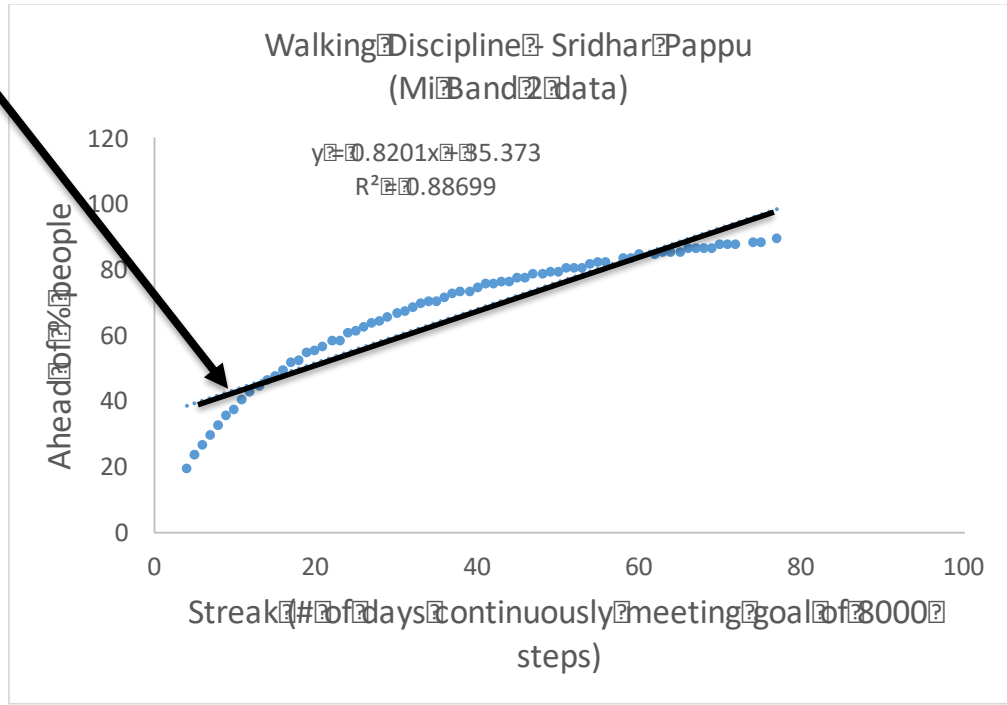
CSE 7302c

Assumptions of the Regression Model – Residuals Analysis

The model is linear

Zero residual line:
The regression line

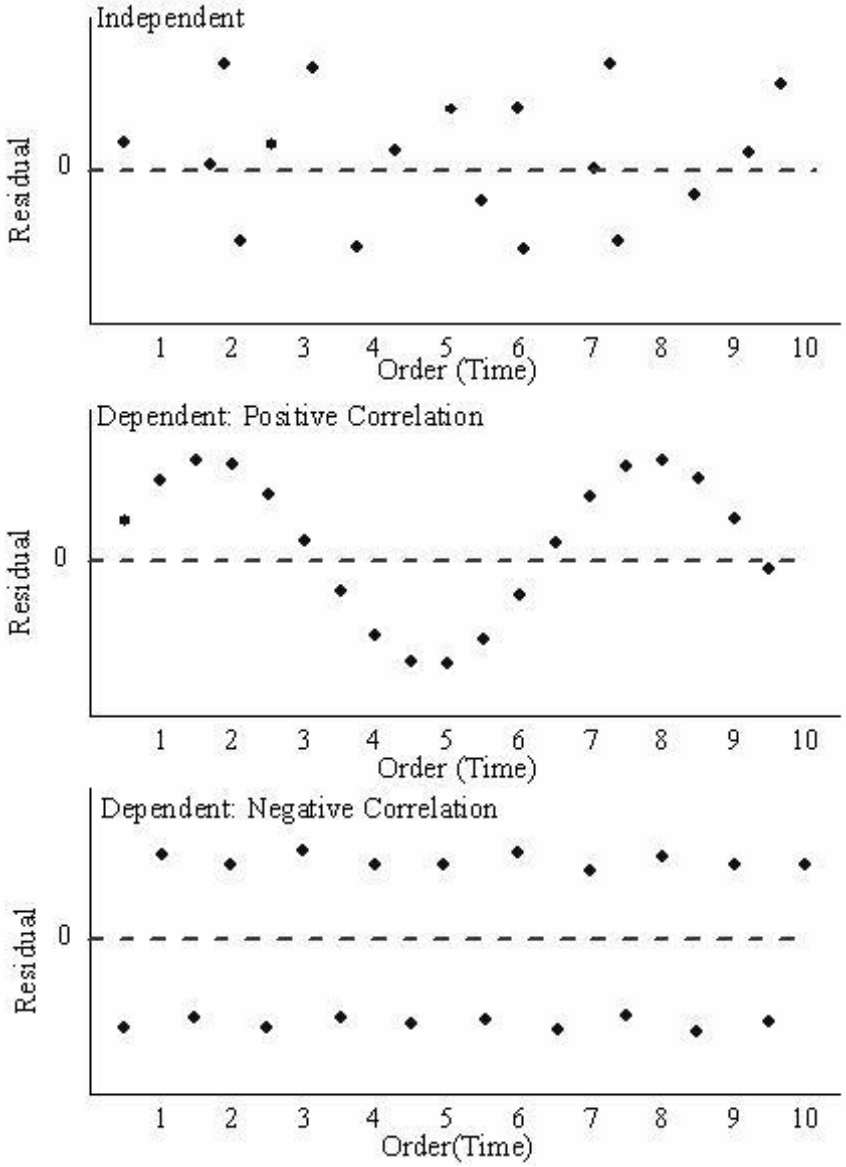




Assumptions of the Regression Model

The error terms are independent

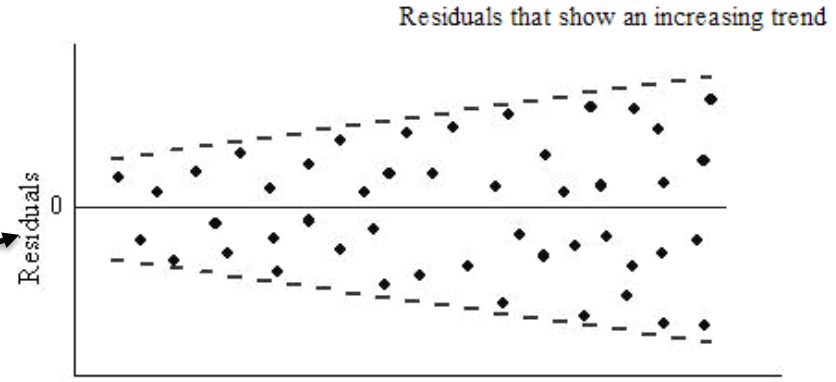
- Plot against any time (order of observation) or spatial variables preferably. Plots against independent variables may also detect independence.
- Time series methods are more appropriate in such situations than regular regression.

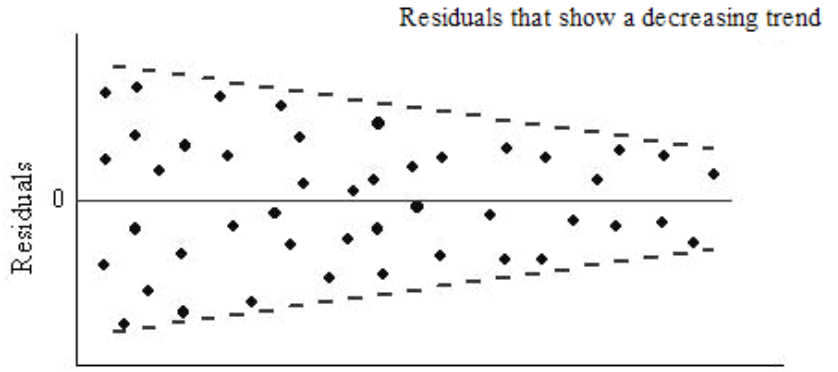


Assumptions of the Regression Model

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

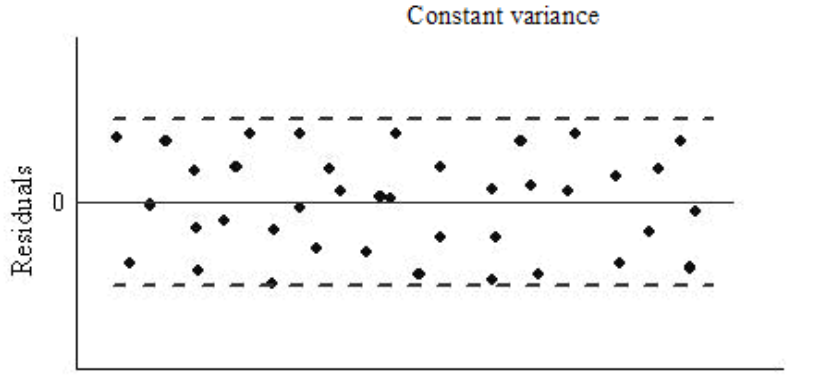
Heteroscedastic





- 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.

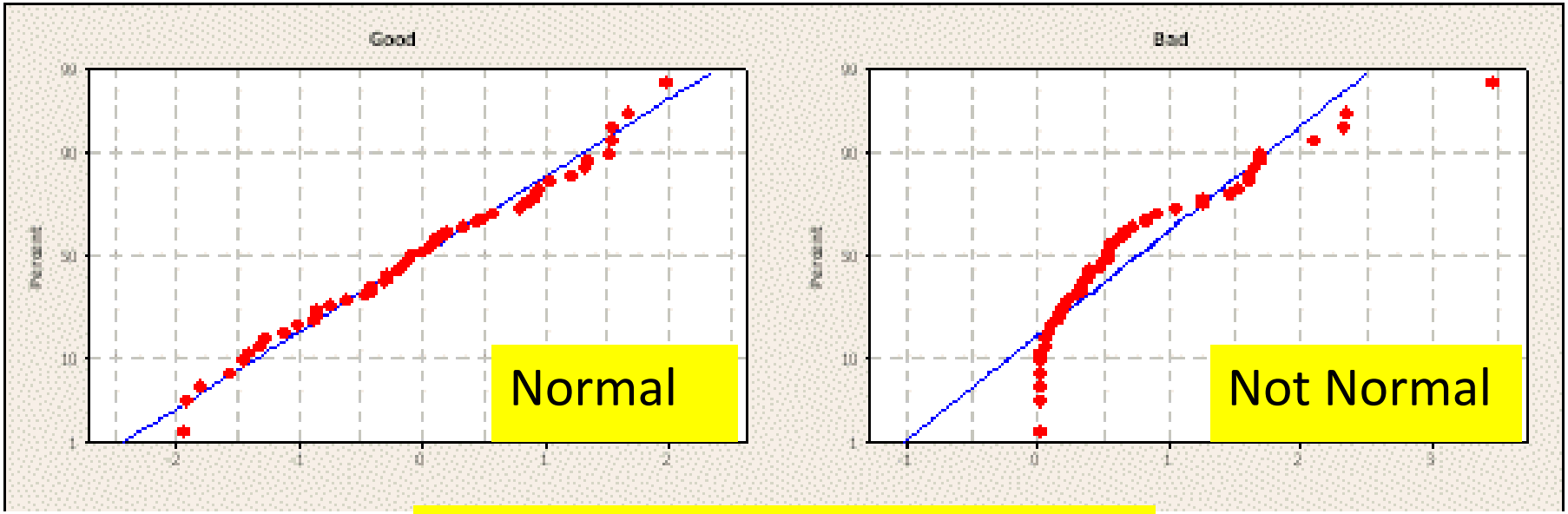
Homoscedastic →



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

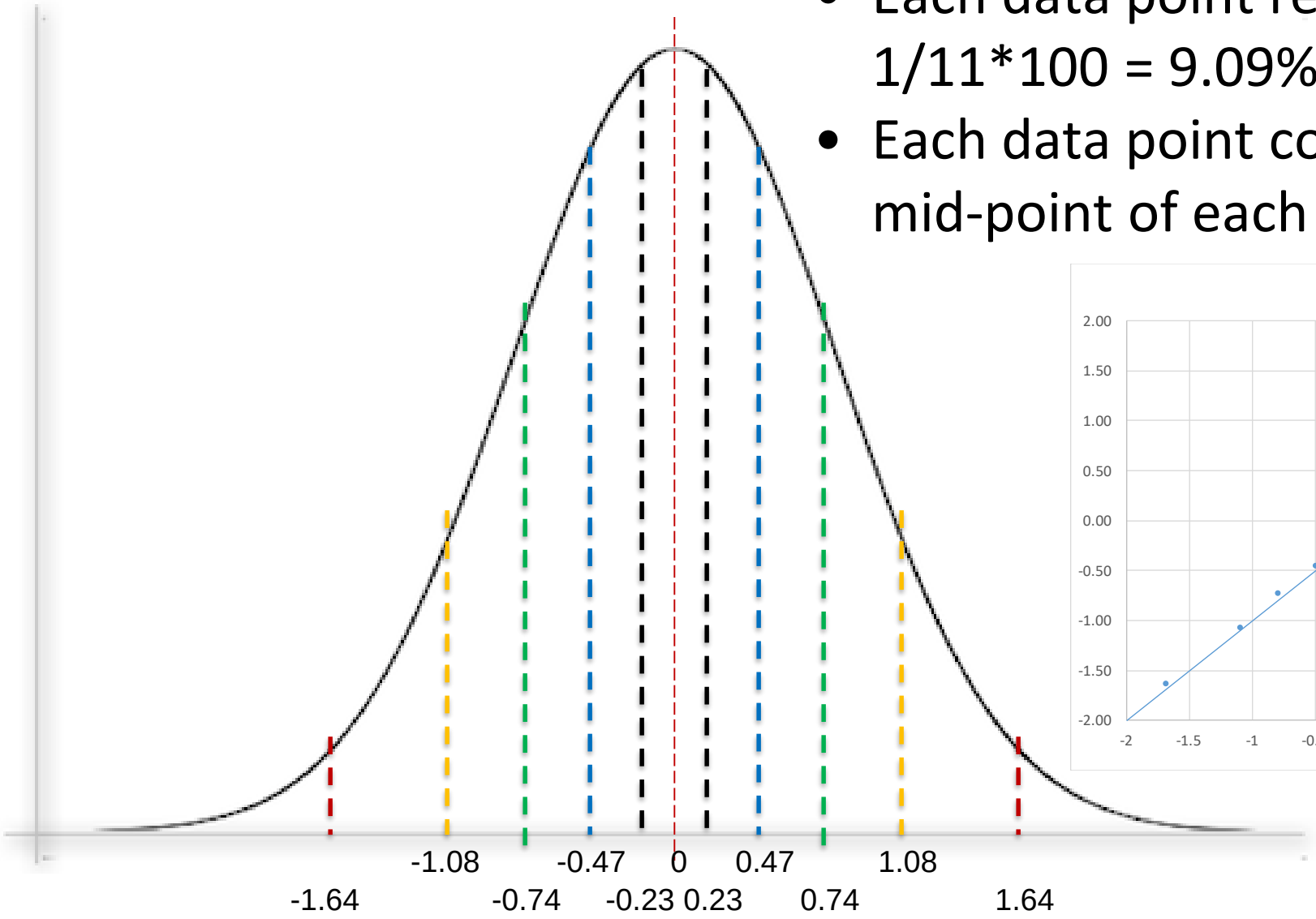
The quantile-quantile (q-q) plot is used to validate distributional assumptions of **a data set**.

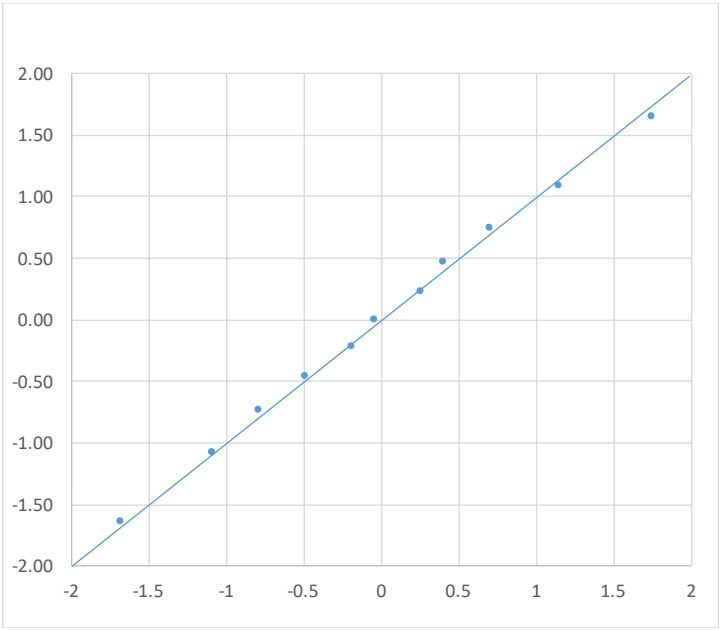
In linear regression, **this data set** is the residual errors.

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)

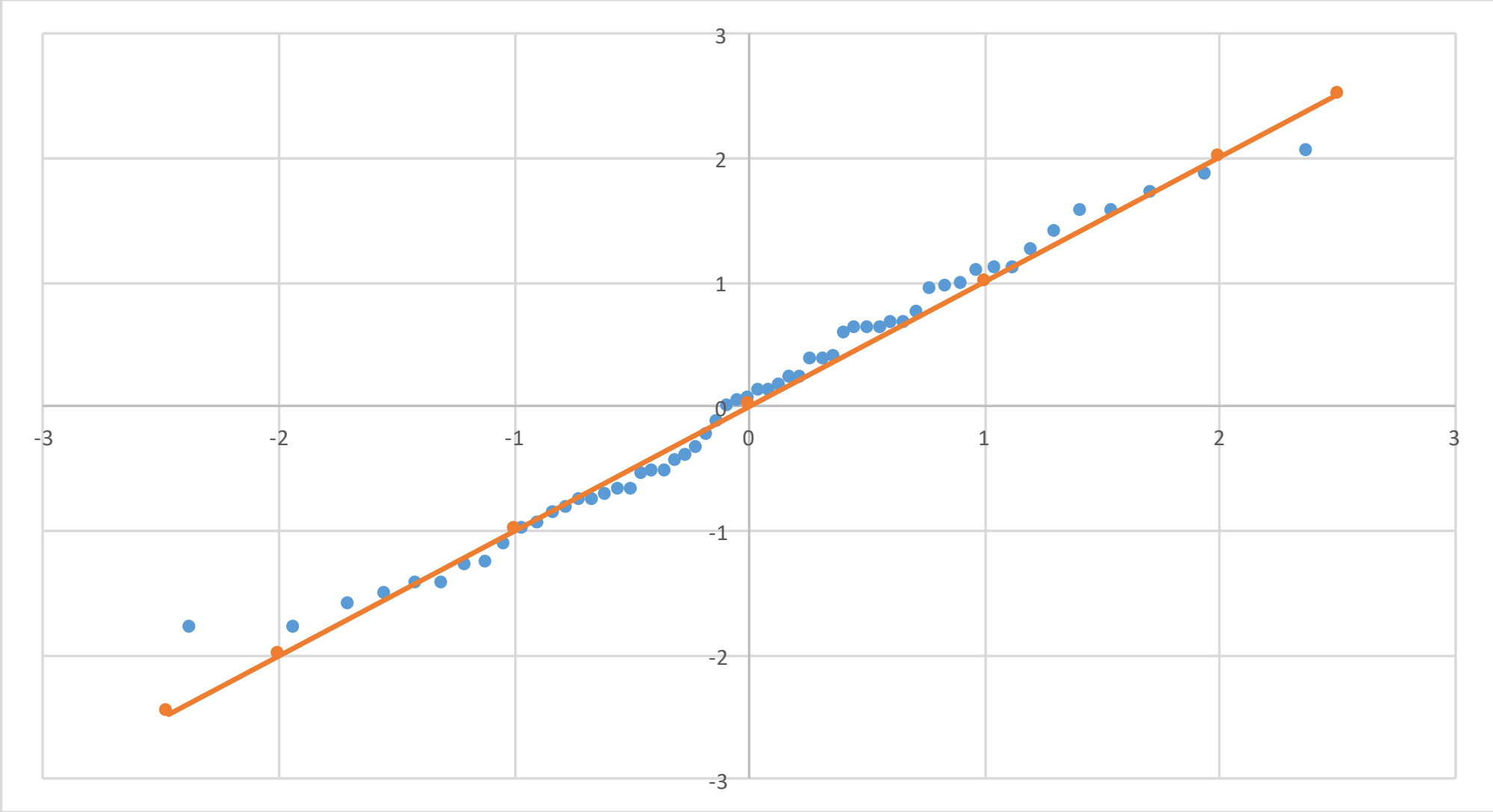
- 11 data points cover 100% area
- Each data point represents 1/11*100 = 9.09% area (or 0.091)
- Each data point considered as mid-point of each of 11 bins





CSE 7302c

Q-Q plot for FEV1 Example (Excel)



CSE 7302c

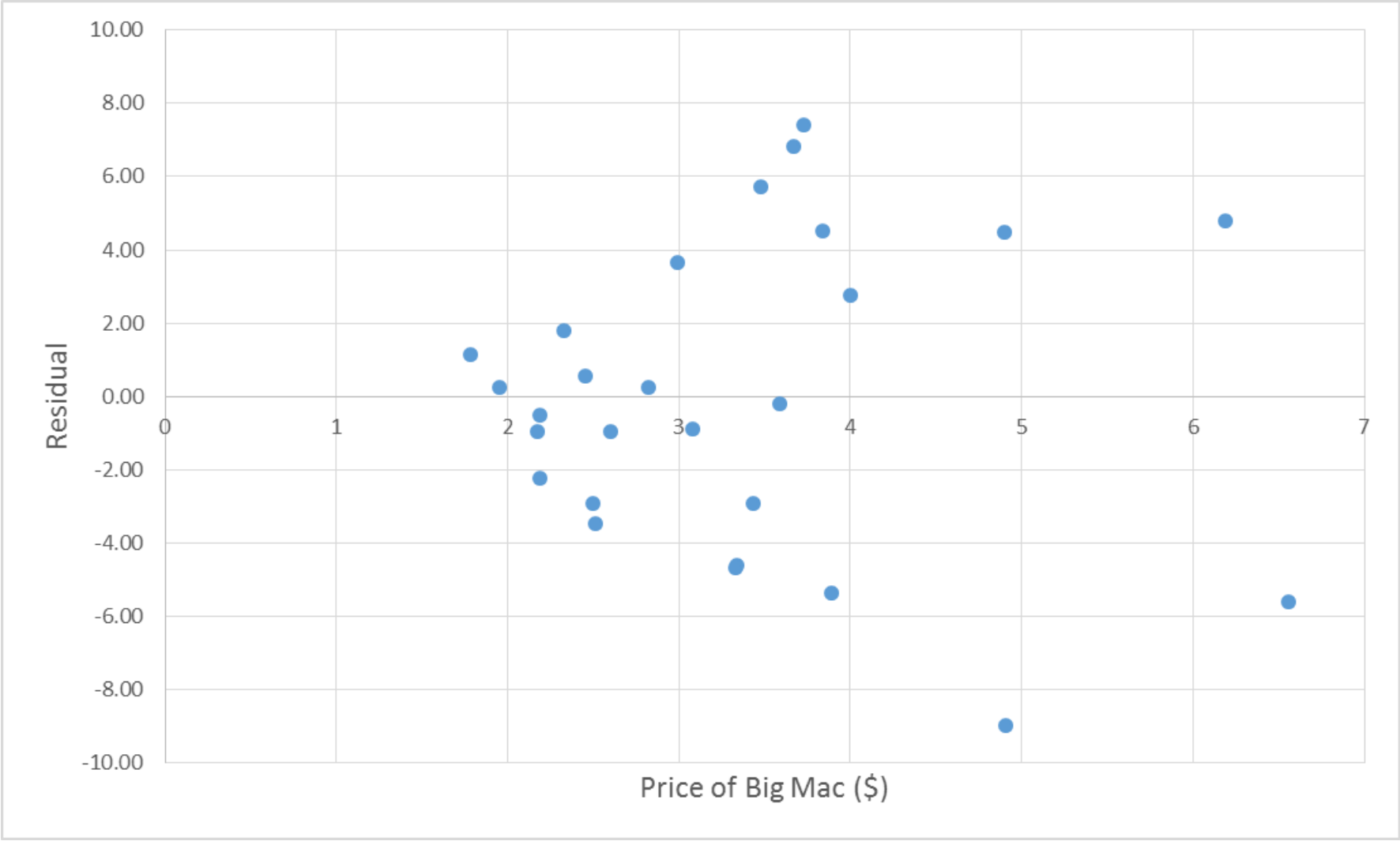
Interpreting Residuals

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

CSE 7302c

Residual Analysis – Big Mac

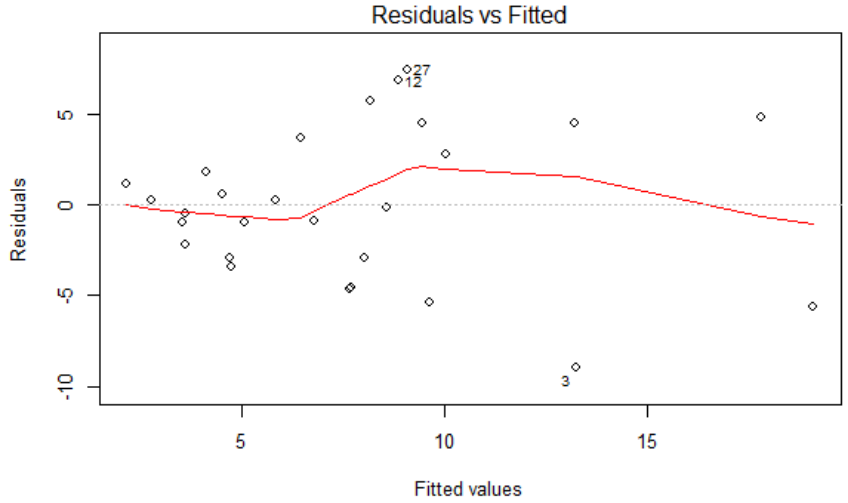
Which assumption is getting violated?



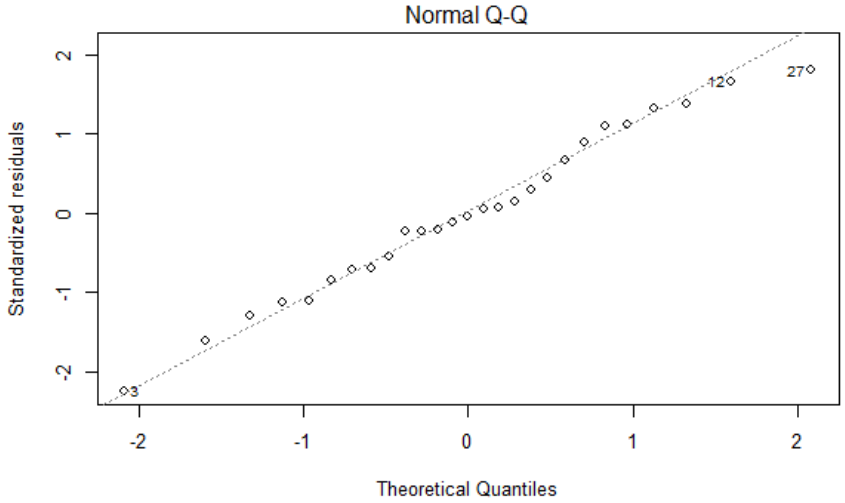
CSE 7302c

Residuals – Big Mac

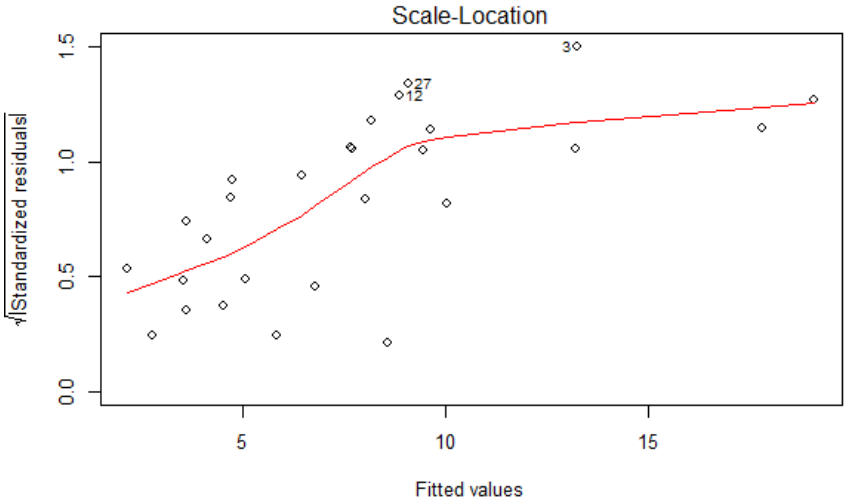
Is a wrong model fitted (linear or quadratic, etc.)?



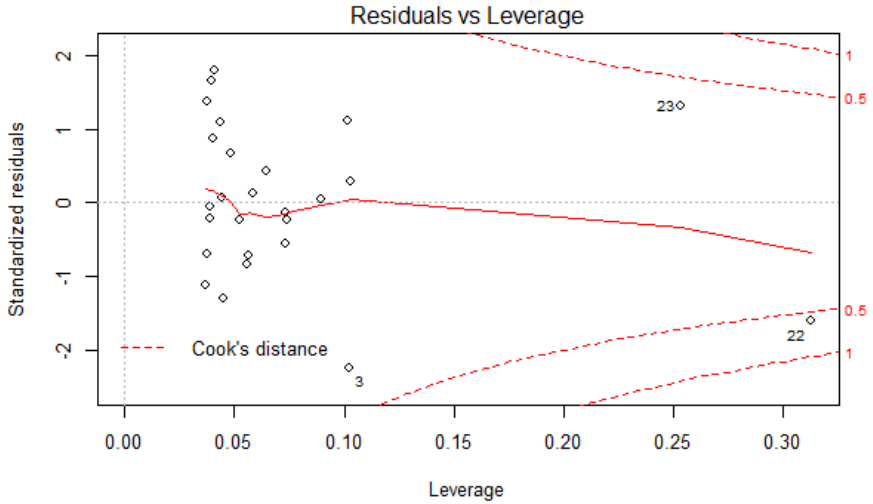
Are the residuals normally distributed?



Is the data homoscedastic?

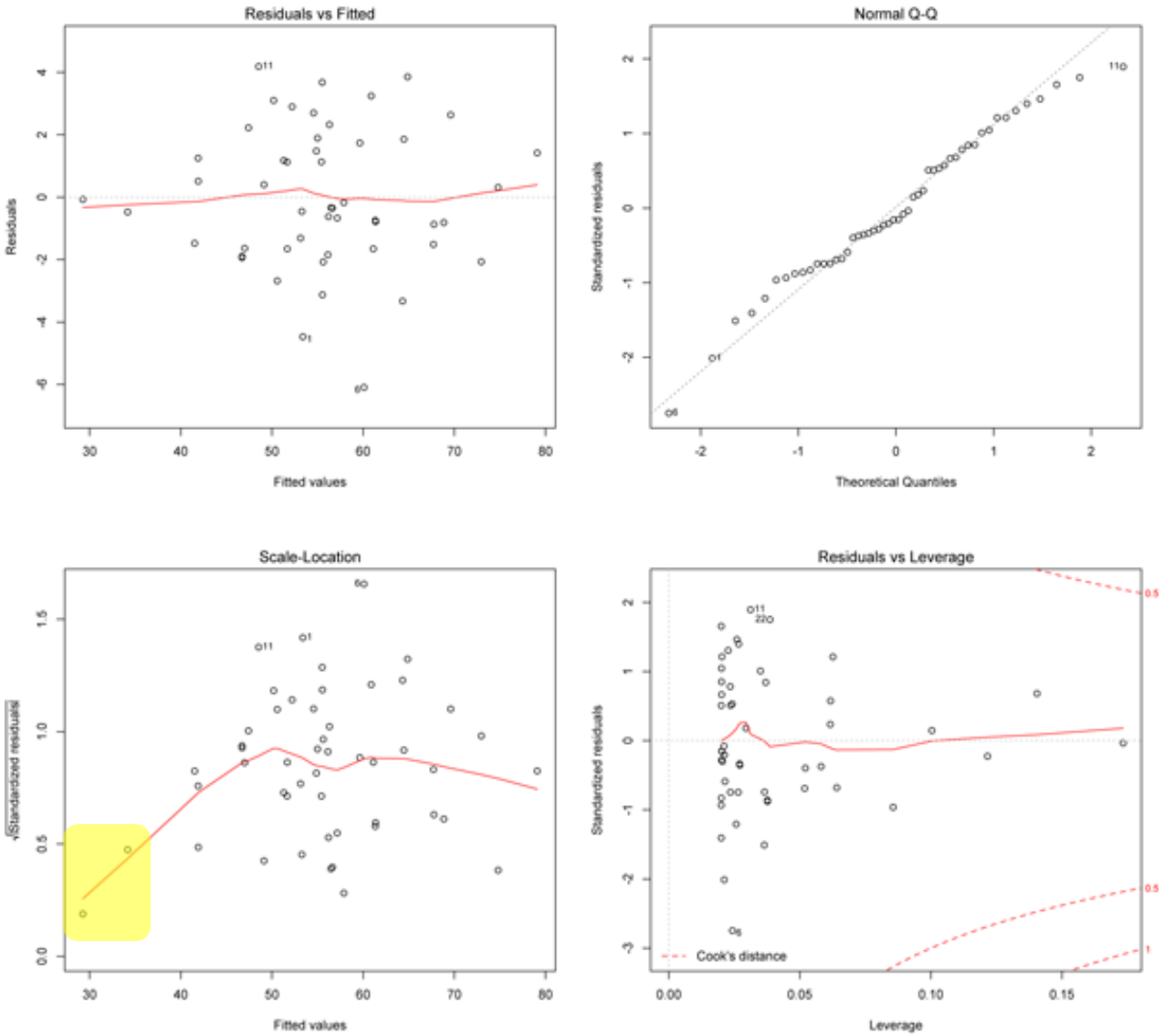


Are there influential outliers?



CSE 7302c

Caution – Is there heteroscedasticity here?



CSE 7302c

Fixing Non-normality and Heteroscedasticity

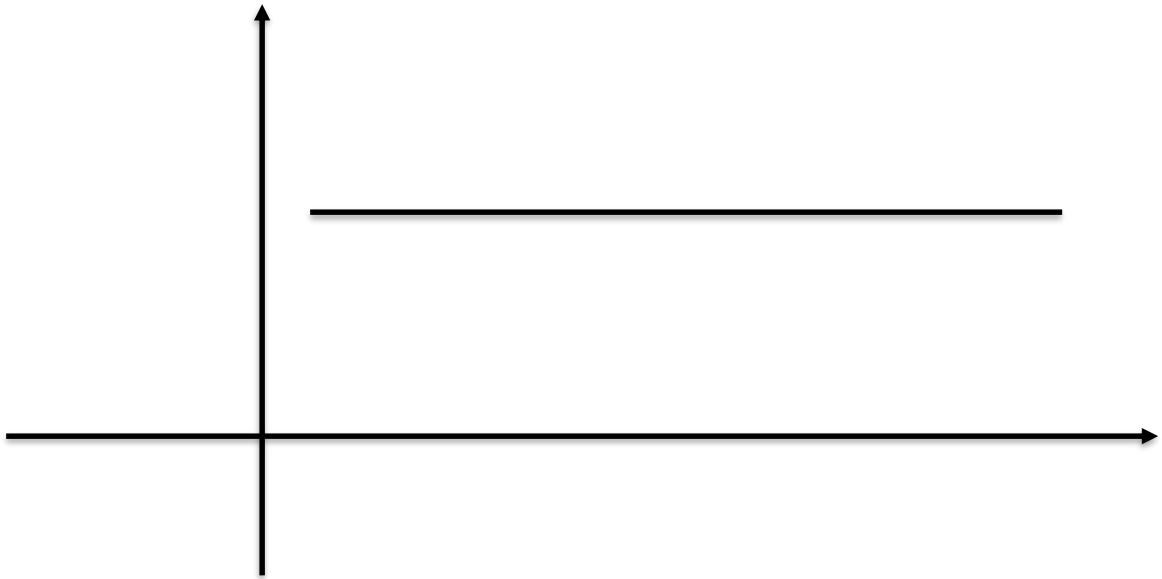
Transformation of data (square root, logarithm, etc.) can help correct normality and unequal variances problems.

CSE 7302c

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.



CSE 7302c

Testing the Slope

What is the Null Hypothesis?

$H_0: \beta_1 = 0$

What is the Alternative Hypothesis?

$H_1: \beta_1 \neq 0$

CSE 7302c

t Test of the Slope

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

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

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

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

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

$t = 5.1437$

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

$t_{25,0.025} = \pm 2.060$

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

Table entry for p and C is the point t^* with probability p lying above it and probability C lying between $-t^*$ and t^*.

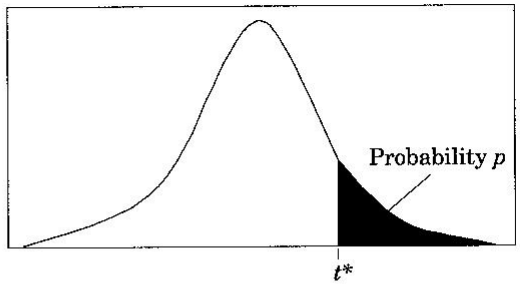


Table B t distribution critical values

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

CSE 7302c

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.

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

CSE 7302c

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

CSE 7302c

Testing the Overall Model – Big Mac - Excel

$F = 26.4581$

Critical Fvalue,

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

Reject the null hypothesis for overall significance

F - Distribution ($\alpha = 0.05$ in the Right Tail)

df_2 \ df_1 (Denominator Degrees of Freedom)	Numerator Degrees of Freedom 1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123
4	7.7086	9.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	6.9988
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799

Sample Software Output

SUMMARY OUTPUT

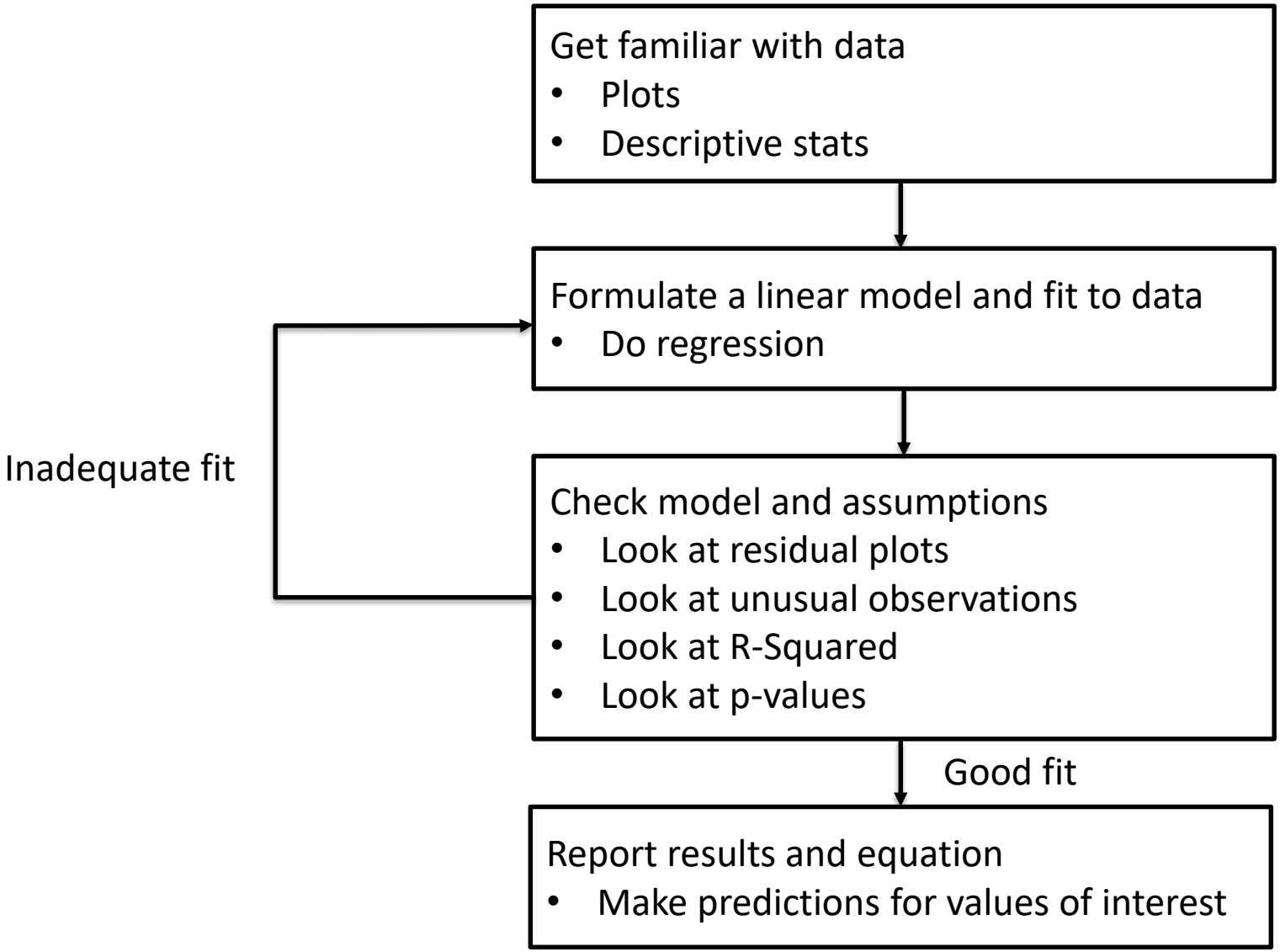
Regression Statistics	
Multiple R	0.717055011
R Square	0.514167888
Adjusted R Square	0.494734604
Standard Error	4.21319131
Observations	27

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



CSE 7302c

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.

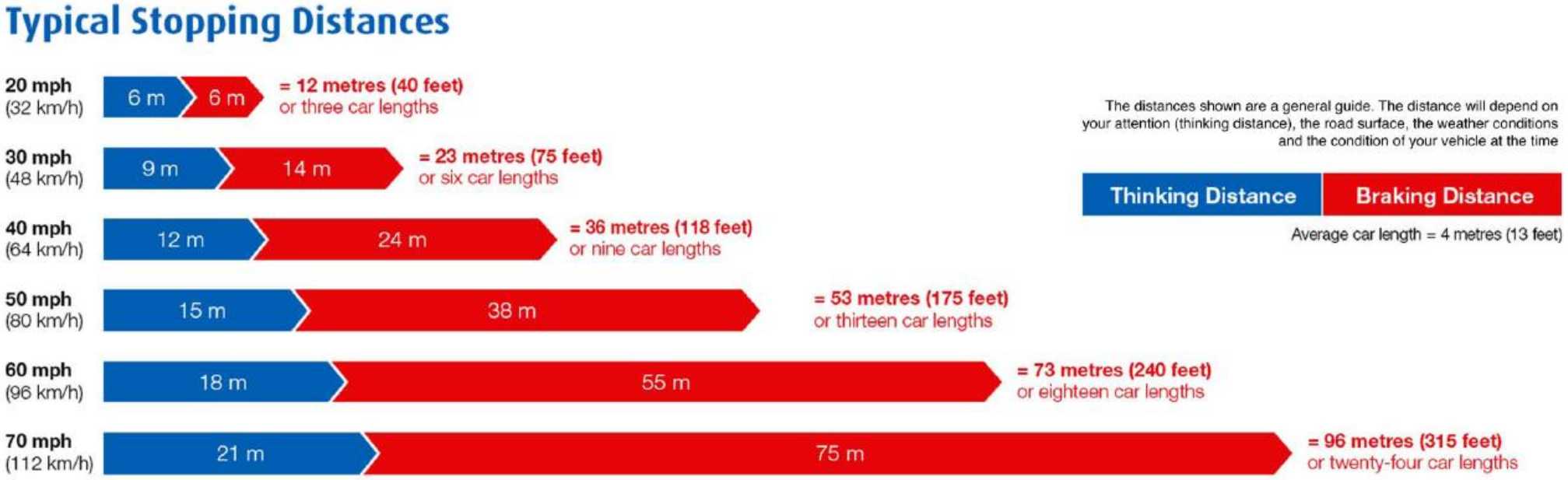


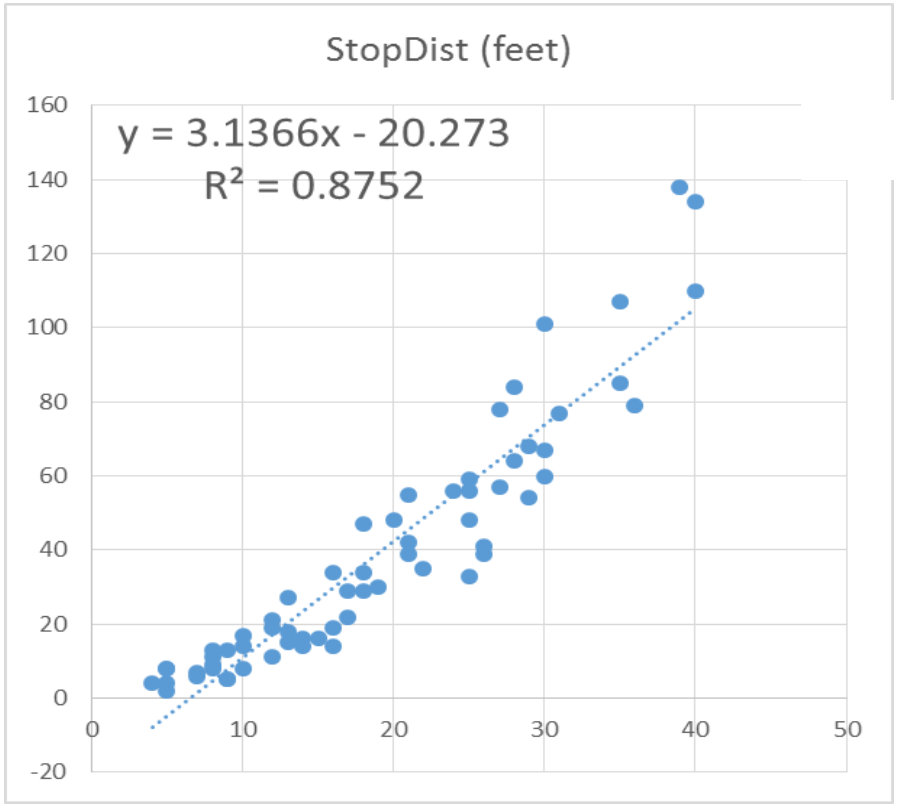
Image Source: http://streets.mn/2015/04/02/the-critical-ten/
Last accessed: November 20, 2015

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.

Does the estimated regression line fit the data well?

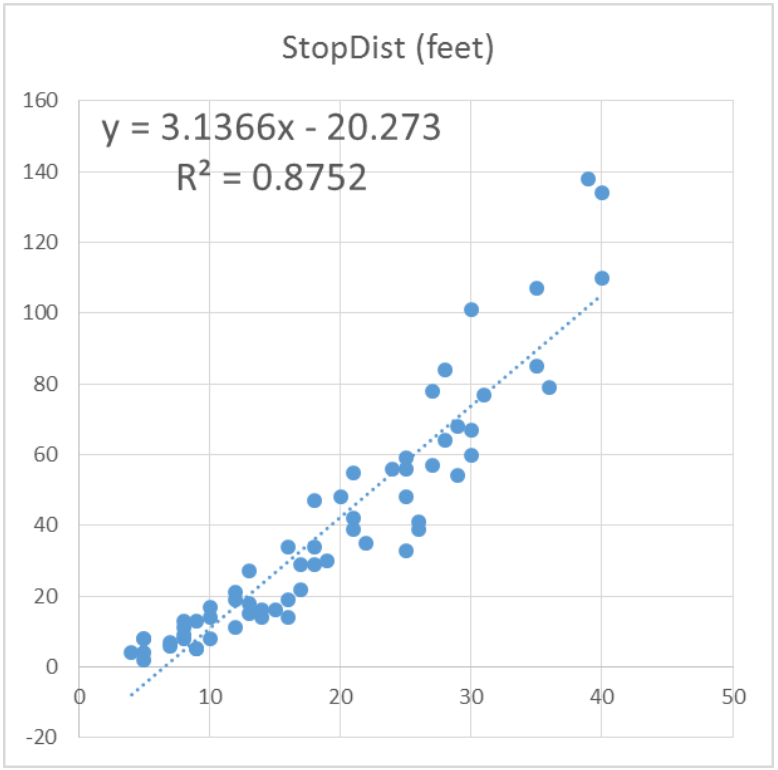


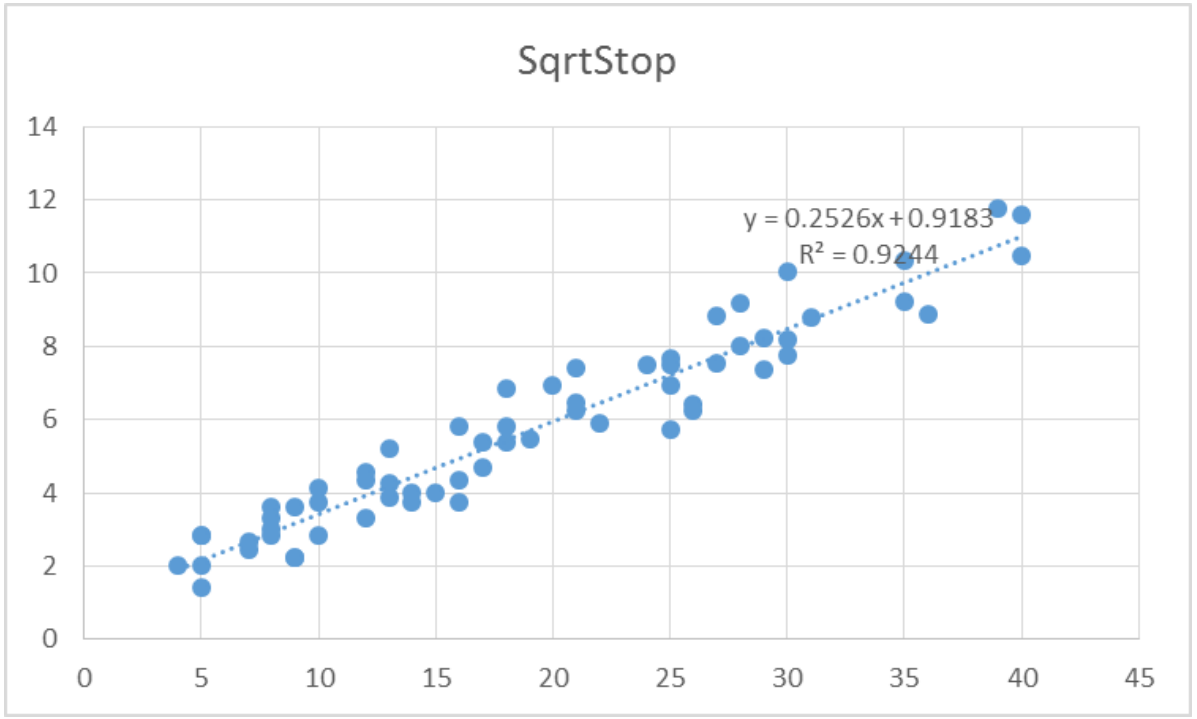
CSE 7302c

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 well.



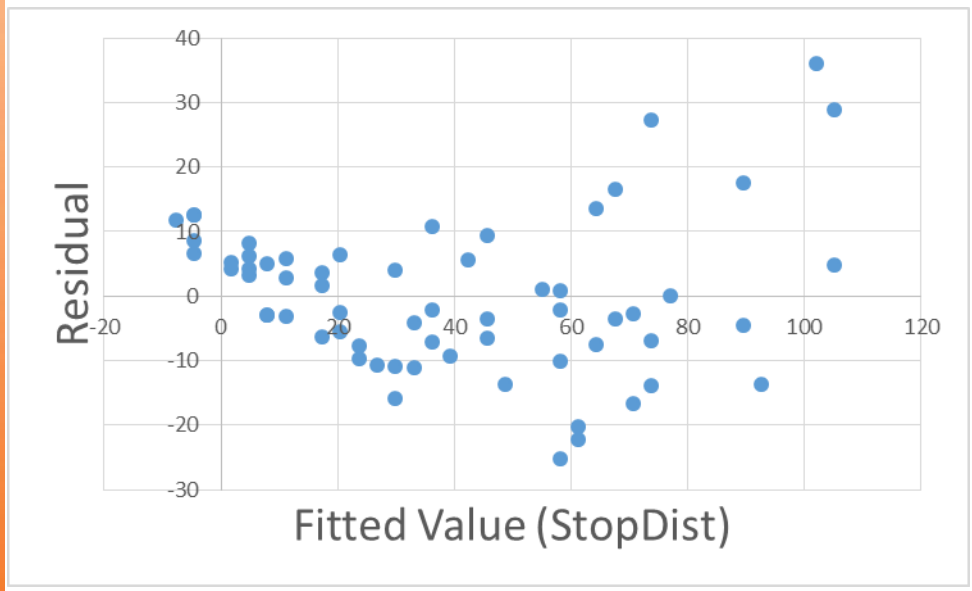


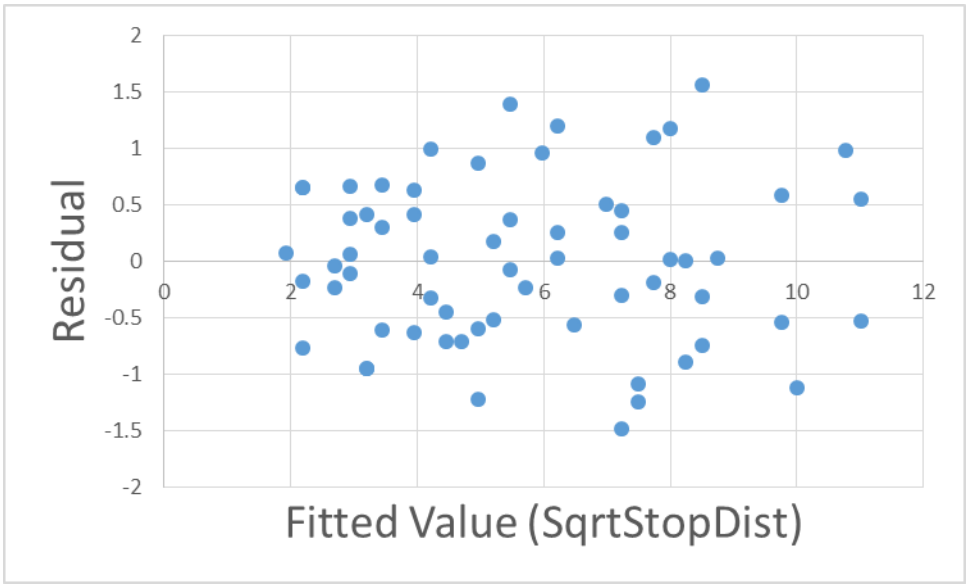
R-Squared, Significance and Residuals - Caution

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 well.

Check the residuals.





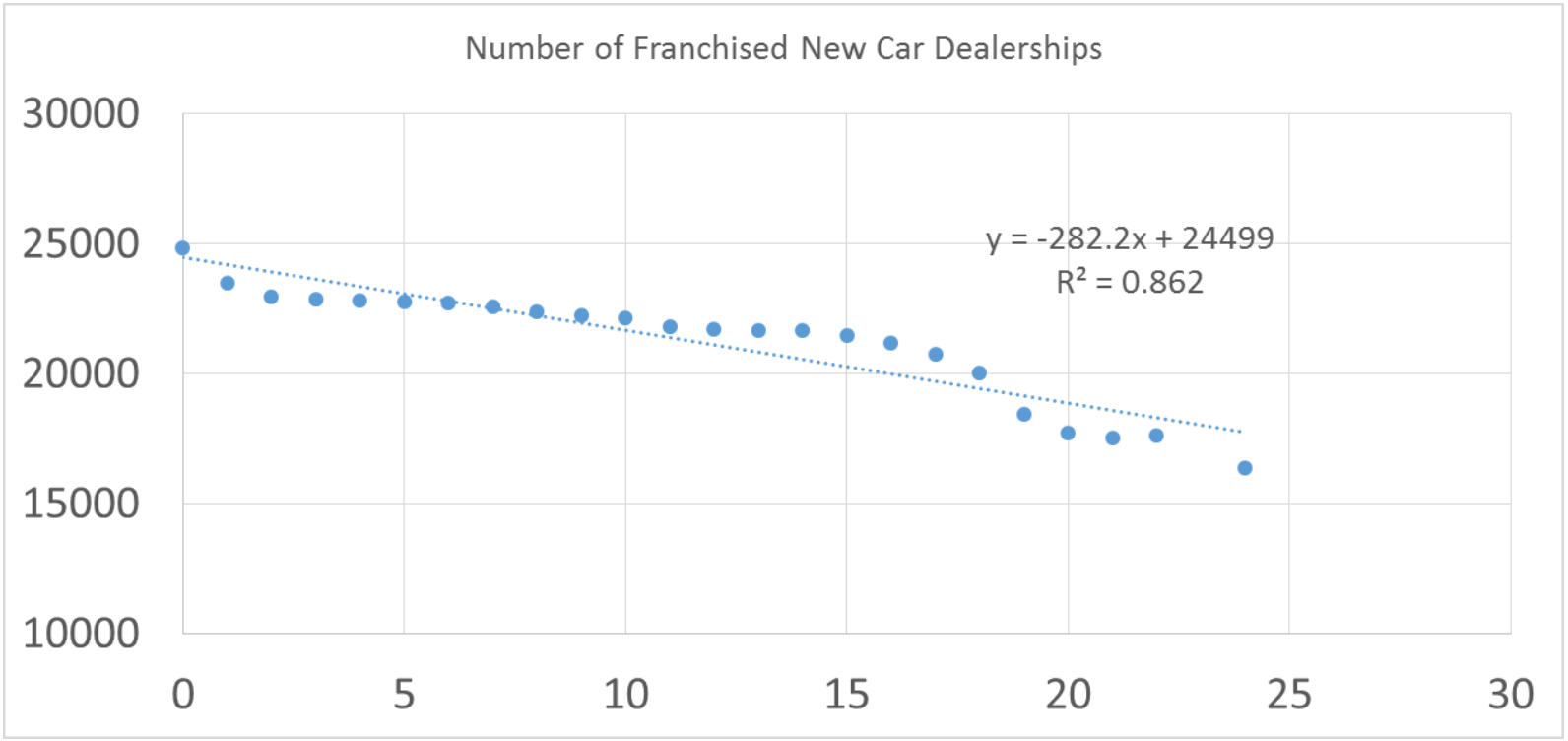
CSE 7302c

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



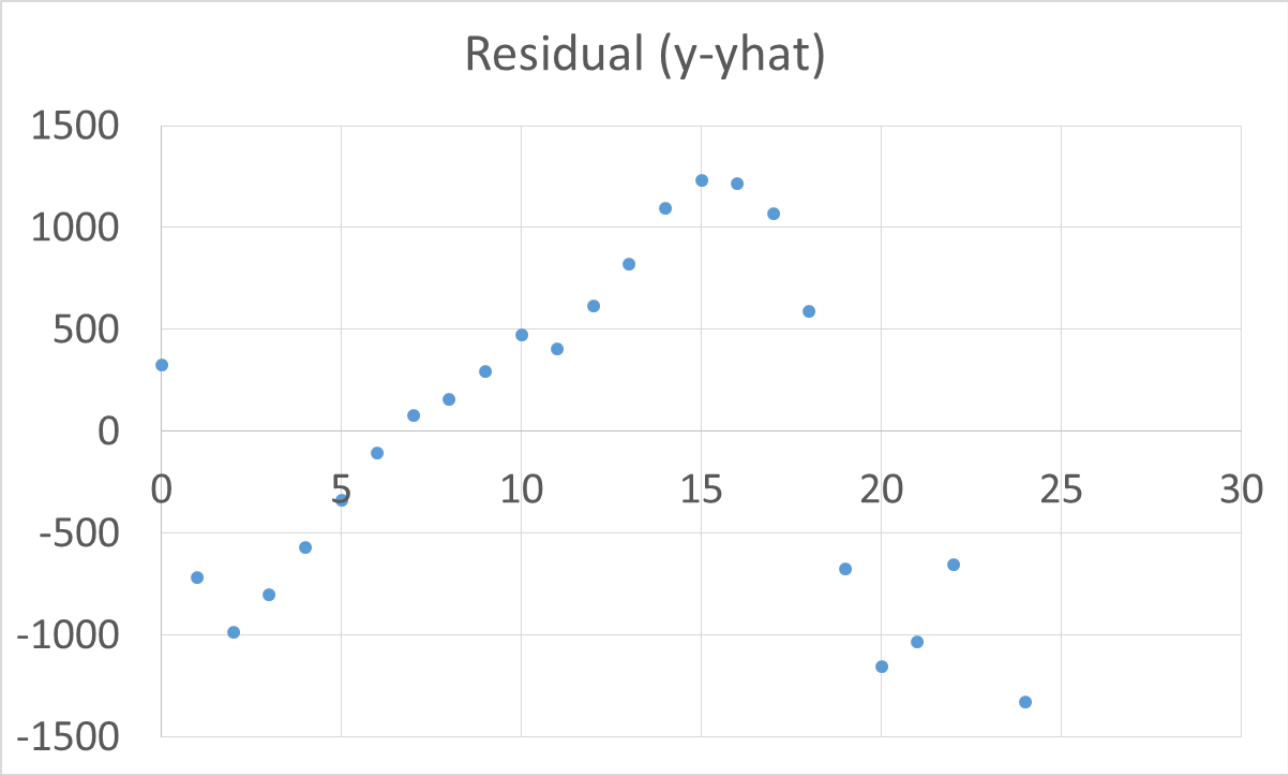
- Based on the shape of the scatter plot, do you think a linear fit looks good?
- Does R^2 imply a good fit?
- What can you infer from the intercept and the slope?

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

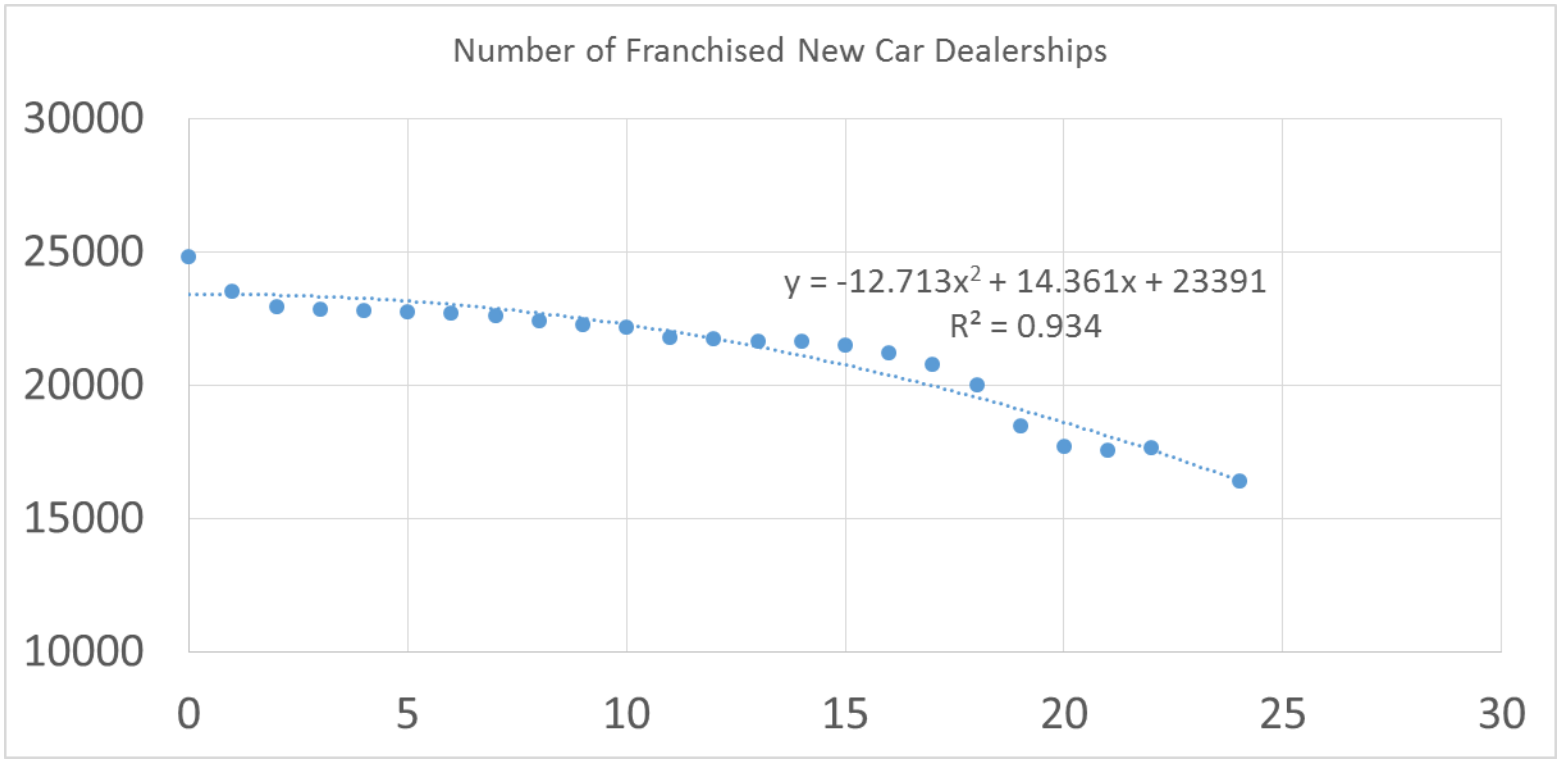
- Is the slope significant?
- Is the model significant?

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

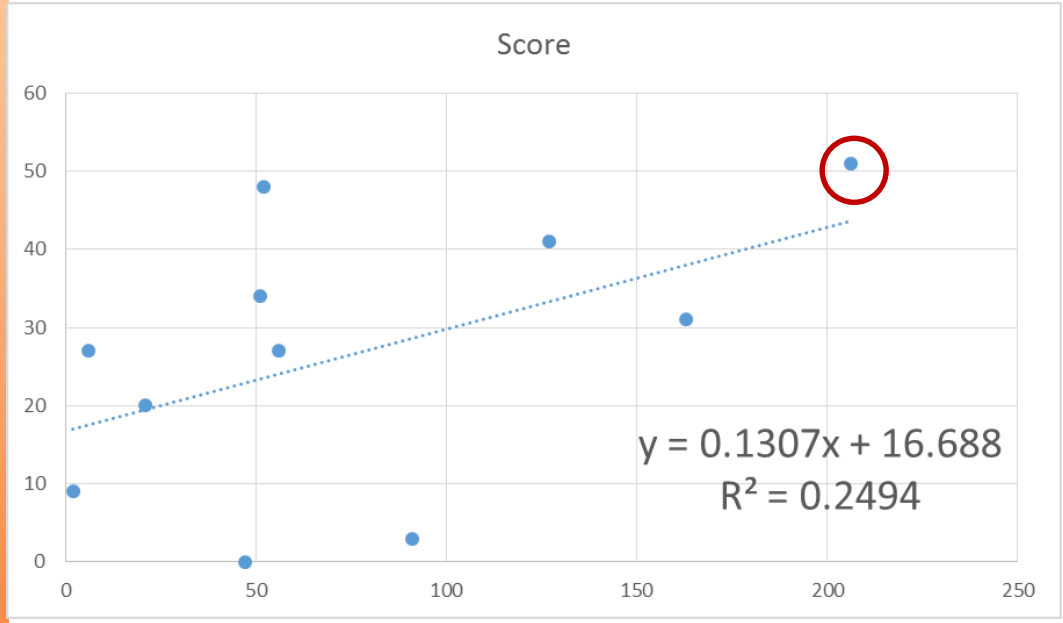


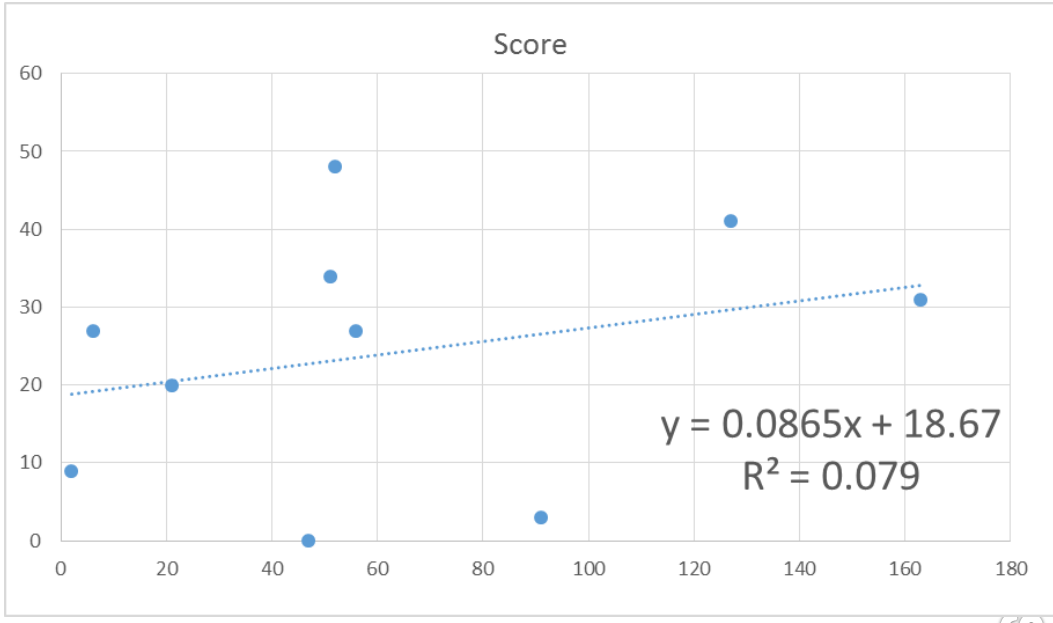
CSE 7302c

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.

SE 73026

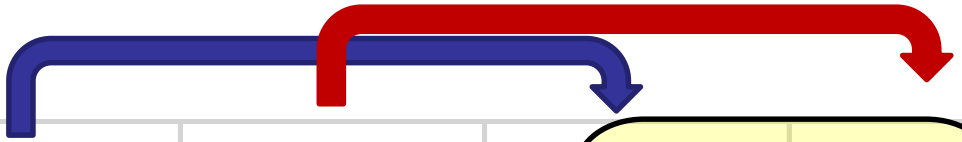
Influential Observations

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



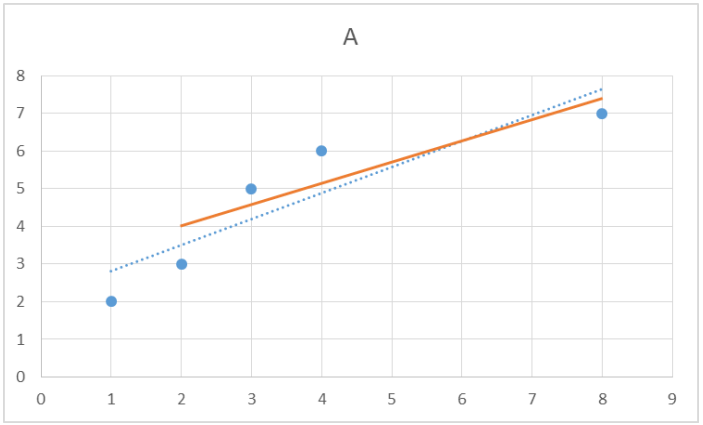
ID	X	Y	h	R	D
A	1	2	0.39	-1.02	0.4
B	2	3	0.27	-0.56	0.06
C	3	5	0.21	0.89	0.11
D	4	6	0.2	1.22	0.19
E	8	7	0.73	-1.68	8.86

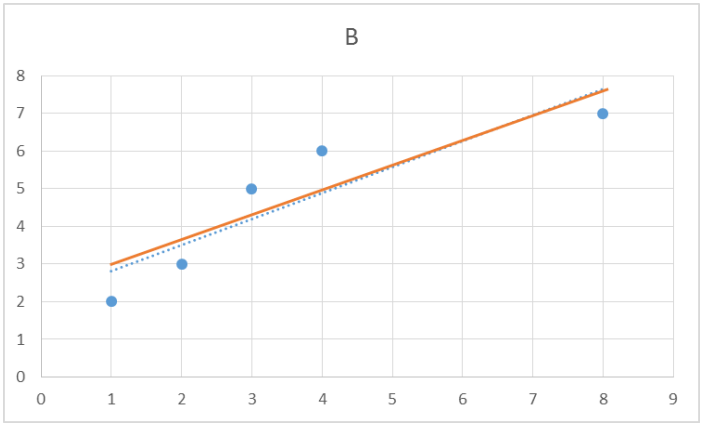
h is the leverage, R is the studentized residual, and D is Cook's measure of influence.

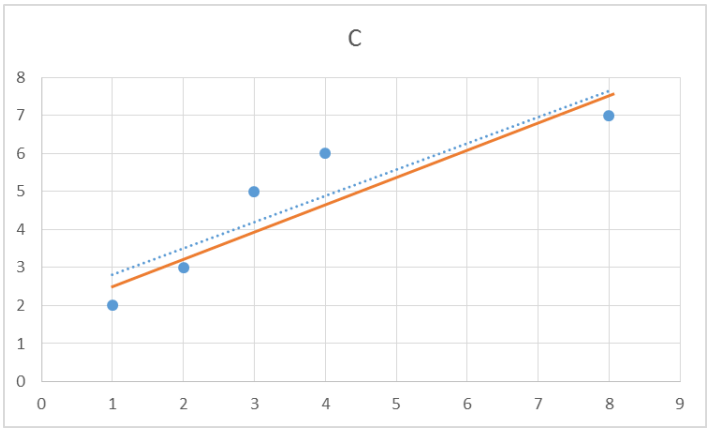
Source: http://onlinestatbook.com/2/regression/influential.html
Last accessed: June 30, 2017

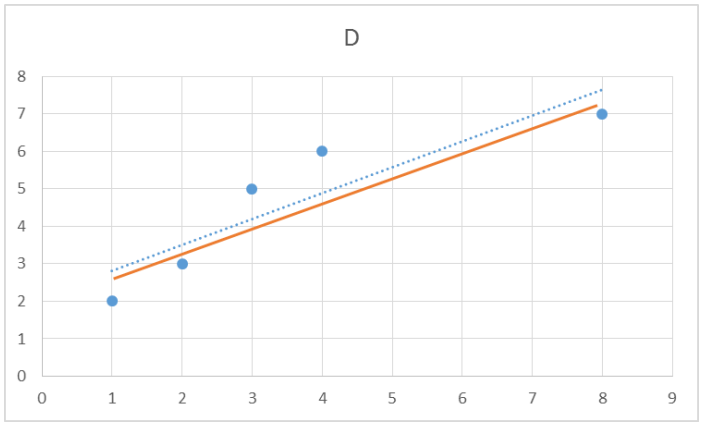
CSE 7302c

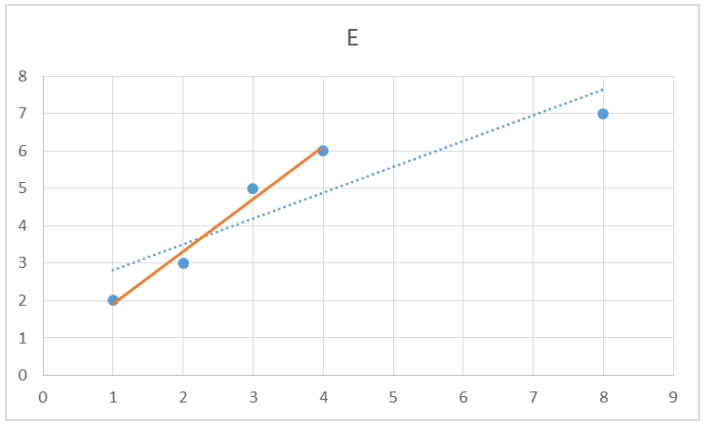
Influential Observations







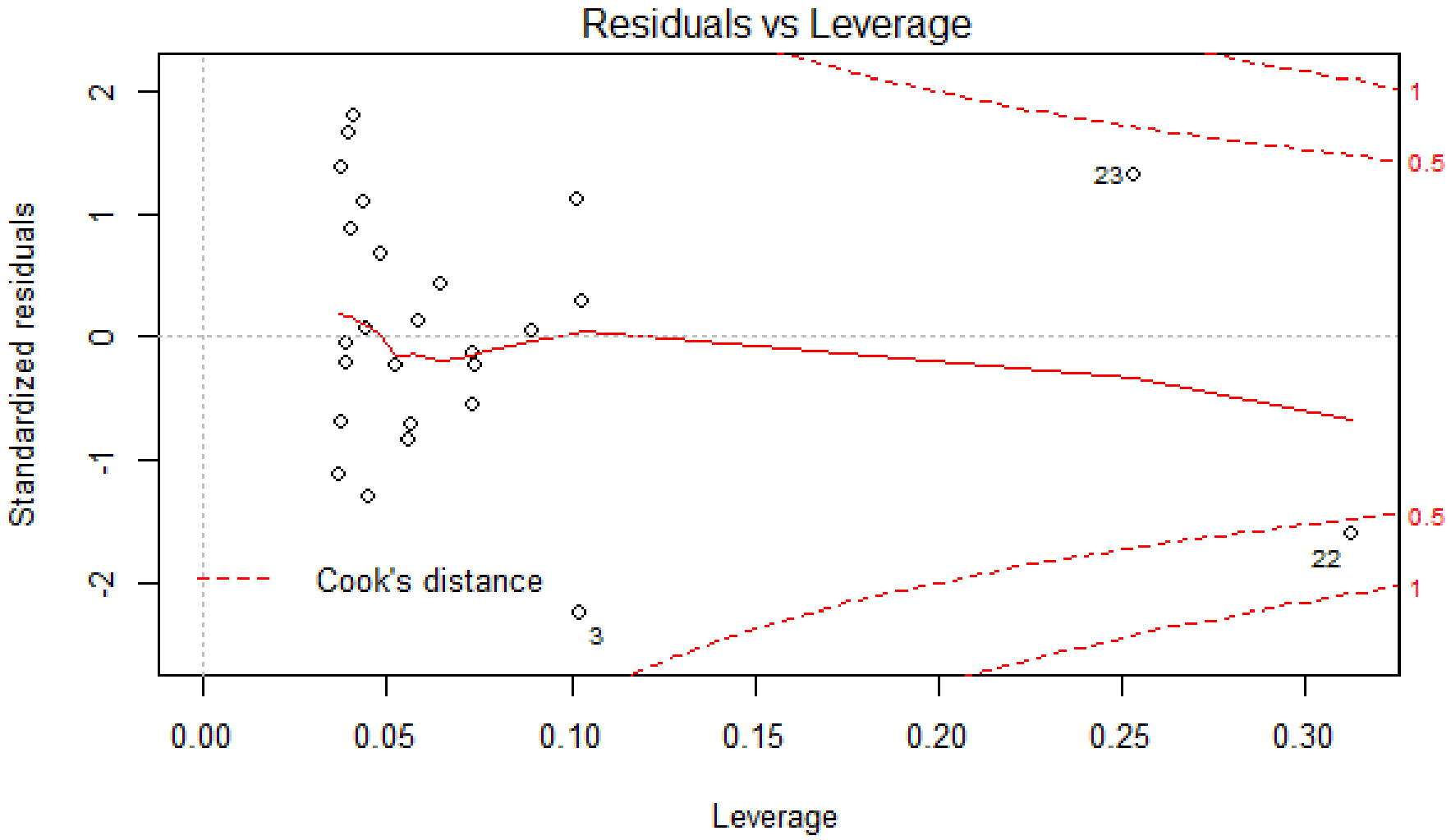




CSE 7302c

Influential Observations

Are there influential outliers?



CSE 7302c

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.**

CSE 7302c

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.

CSE 7302c

Influential Observations - Leverage

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

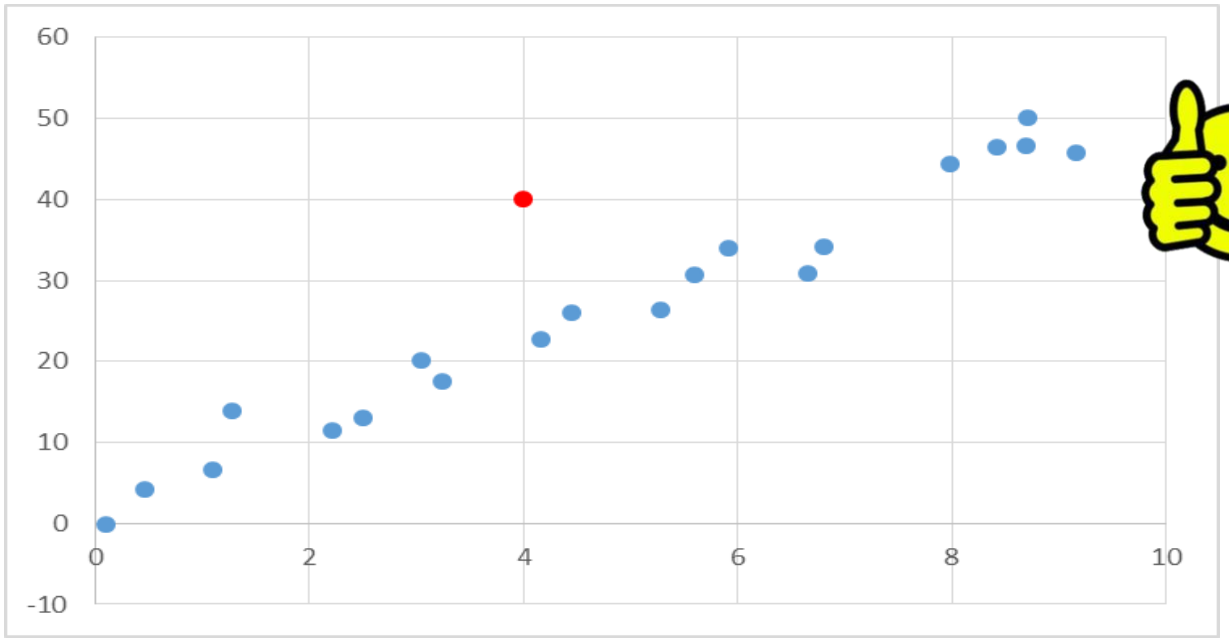
$$h_i = \frac{1+z^2}{n}$$

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

EXCEL ACTIVITY

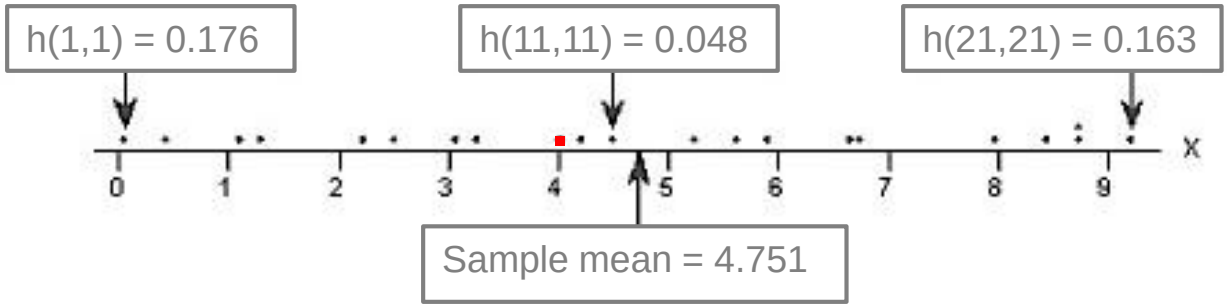
CSE 7302c

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

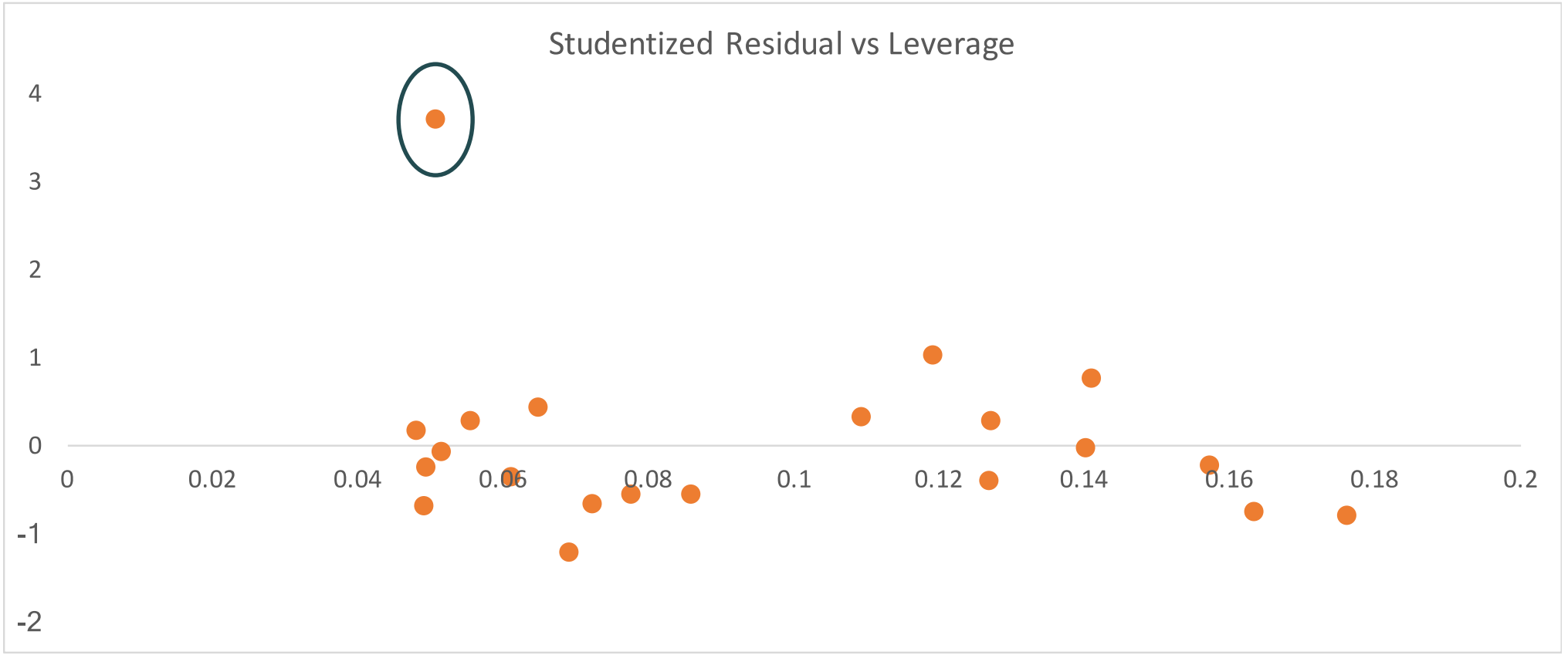
$$stdres_i = \frac{e_i}{\sqrt{MSE(1 - h_i)}}$$

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



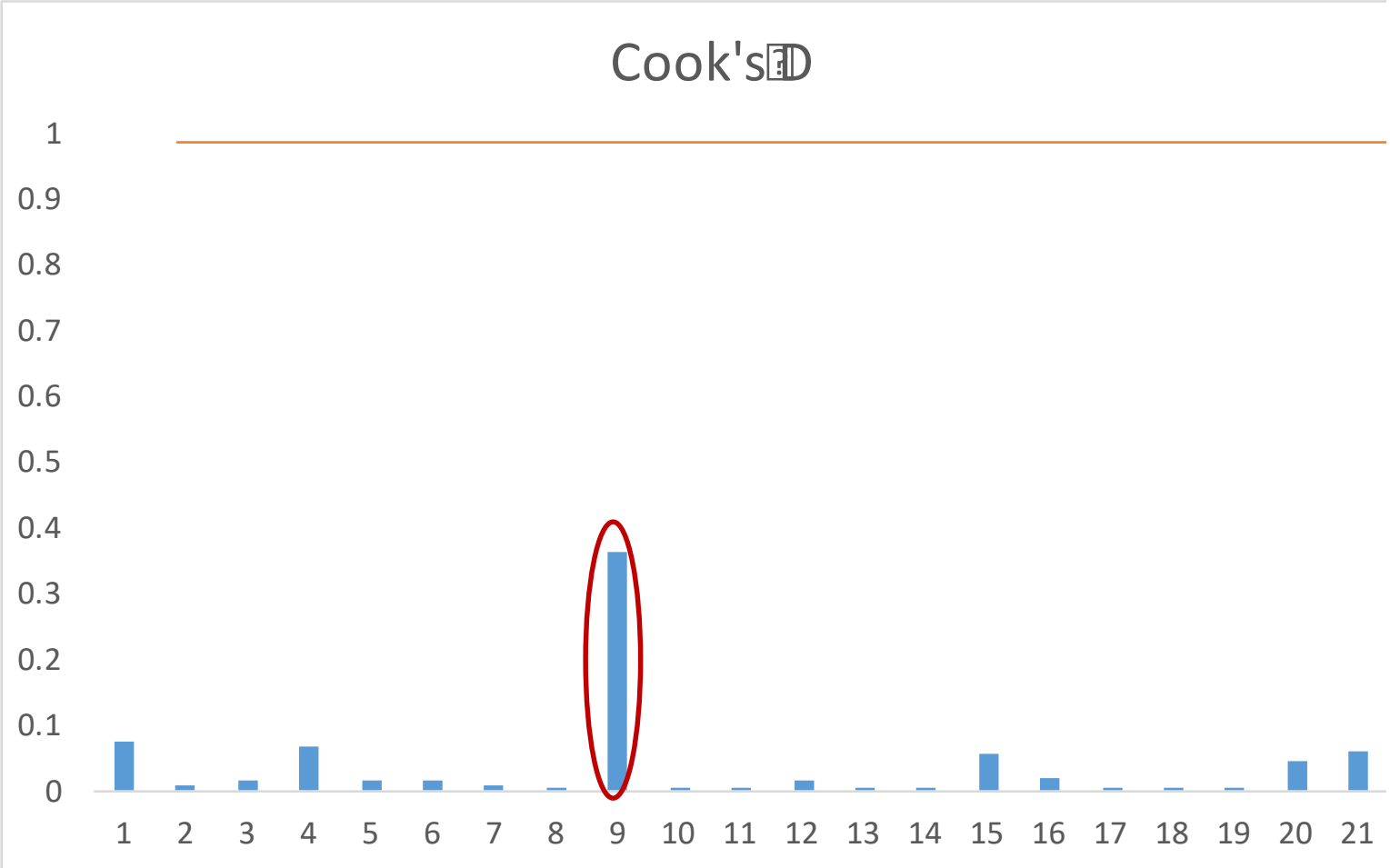
Influential Observations – Cook's D

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**.

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

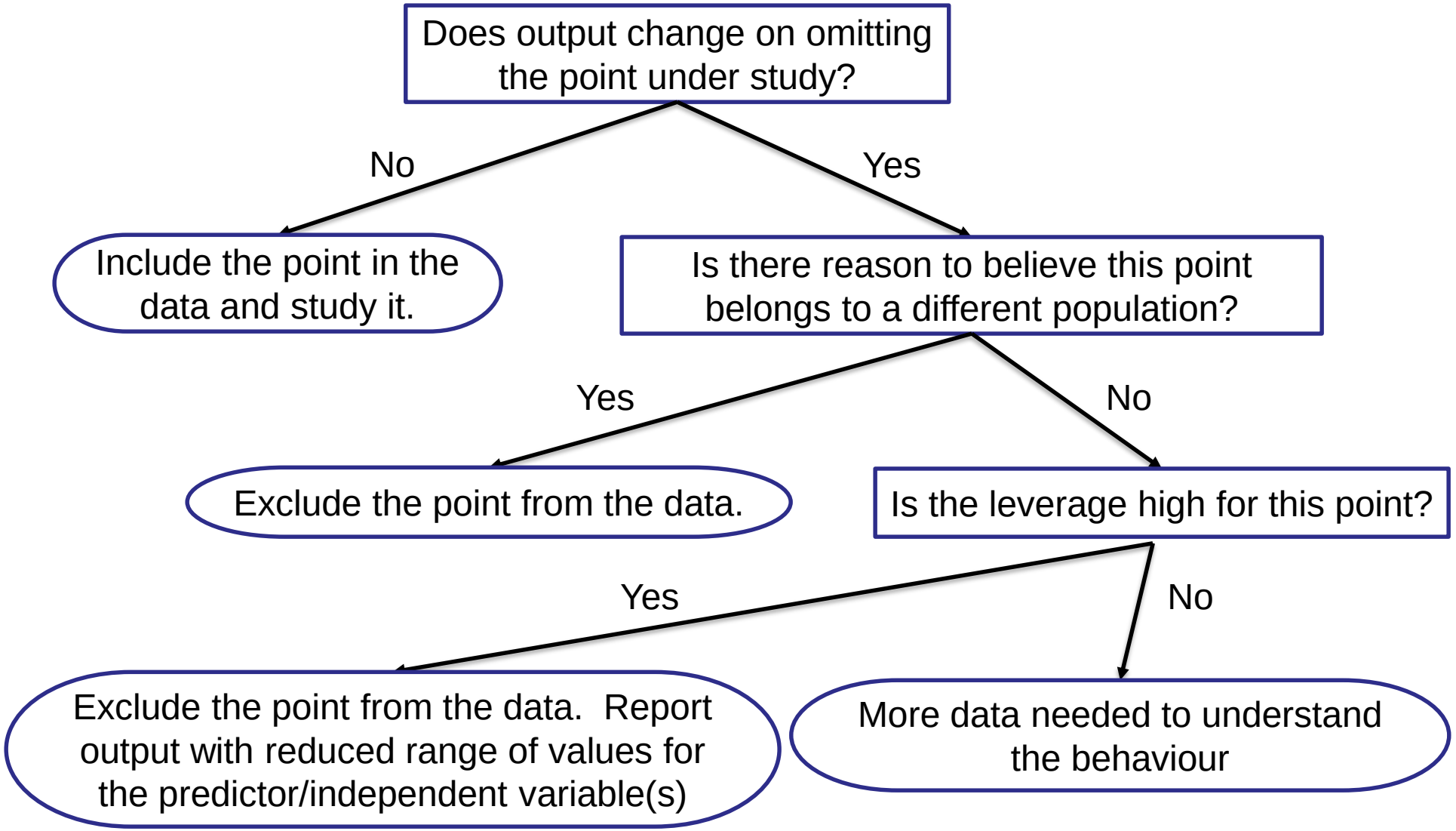
CSE 7302c

Influential Observations – Cook's D

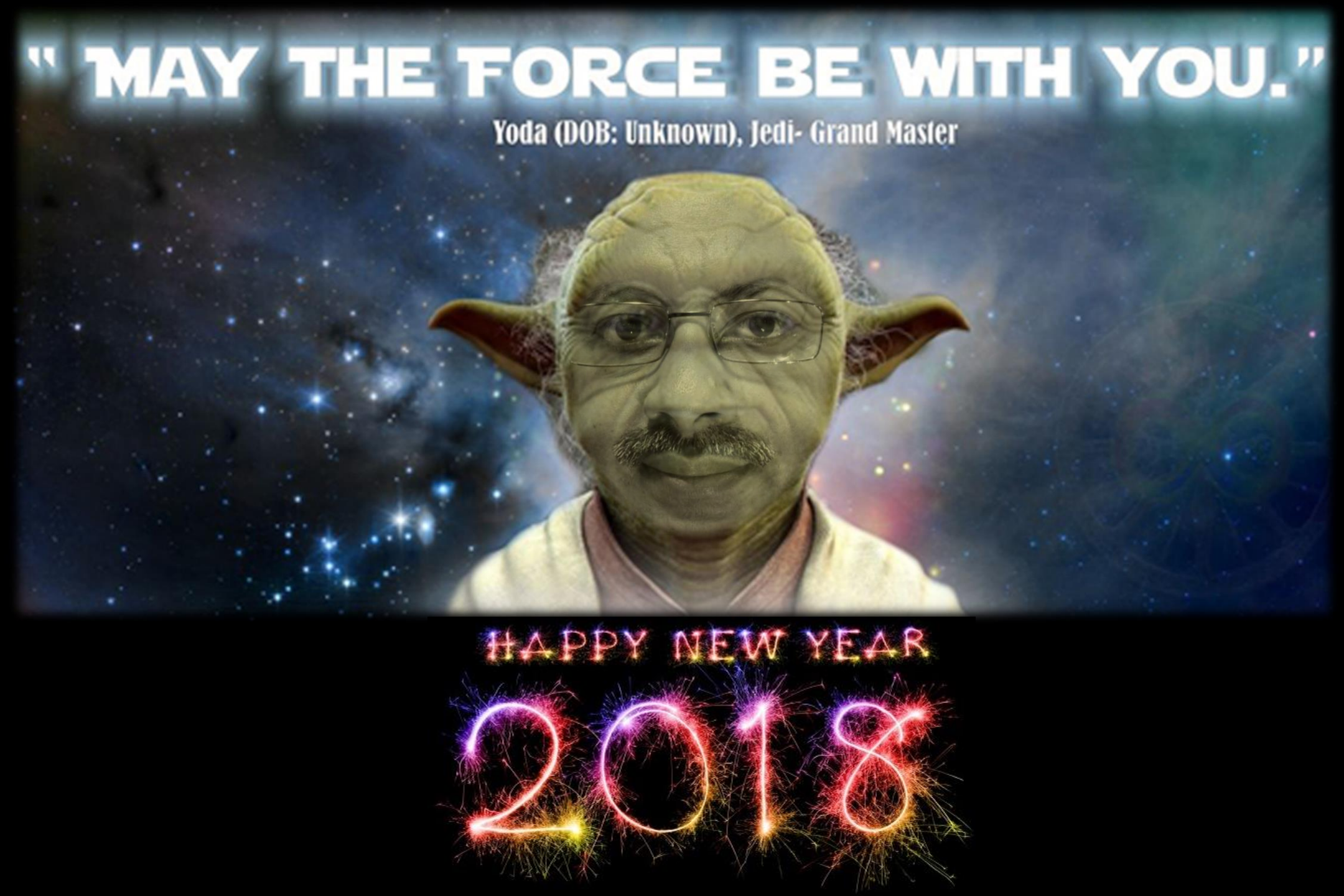


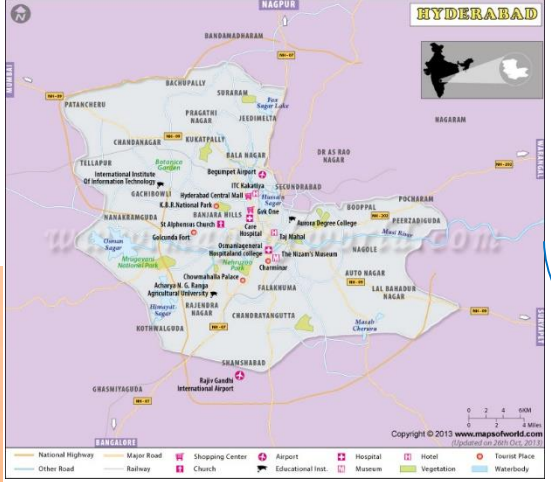
CSE 7302c

Handling Influential Observations

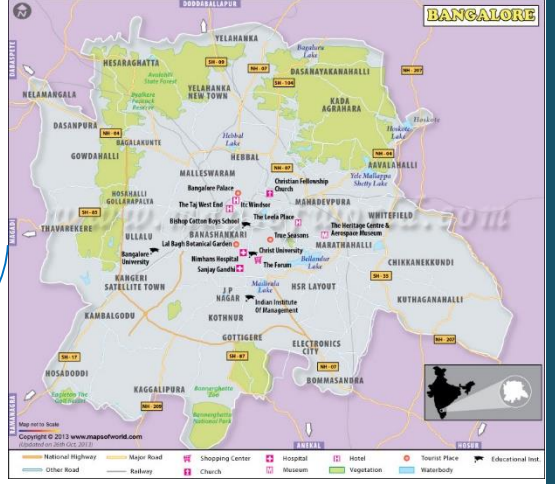


CSE 7302c


" MAY THE FORCE BE WITH YOU."
Yoda (DOB: Unknown), Jedi- Grand Master
HAPPY NEW YEAR
2018



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.