Inspire…Educate…Transform.

Statistics and Probability in Decision Modeling

Linear Regression

Dr. Anand Narasimhamurthy

Outline

- General overview
- Basic understanding of Linear Regression with simple examples
- Brief review of basic concepts
    - Covariance, Correlation, p-value, hypothesis testing
- Hands-on exercise
- Detailed understanding of key ideas with a running example (Big Mac “Index”)
    - Testing how well the regression model fits the data, residual analysis
    - Major steps in building a Linear regression model
- Influential points : Leverage statistics, Cook’s distance
- Data transformations

Why linear regression?

- A regression based model can be used as a simple baseline model that can be built relatively easily.

- Interpretation of the model is often quite straightforward as compared to other more powerful models which tend to be black boxes.

- **A practical reason (often as good as any)**
  - Client's choice and mandate often dictates choice of model.
  - Interpretability is often very important even if it means having to trade-off some accuracy.

General overview

In many practical applications there is a need to predict quantities of interest with reasonable accuracy

Typically these quantities are either difficult to measure or are forecasts.

Forecasting quarterly sales of a product

Predicting length of patient stay in a hospital

Predicting whether a loan applicant is likely to default

Predicting stock prices

Common classes of practical learning problems

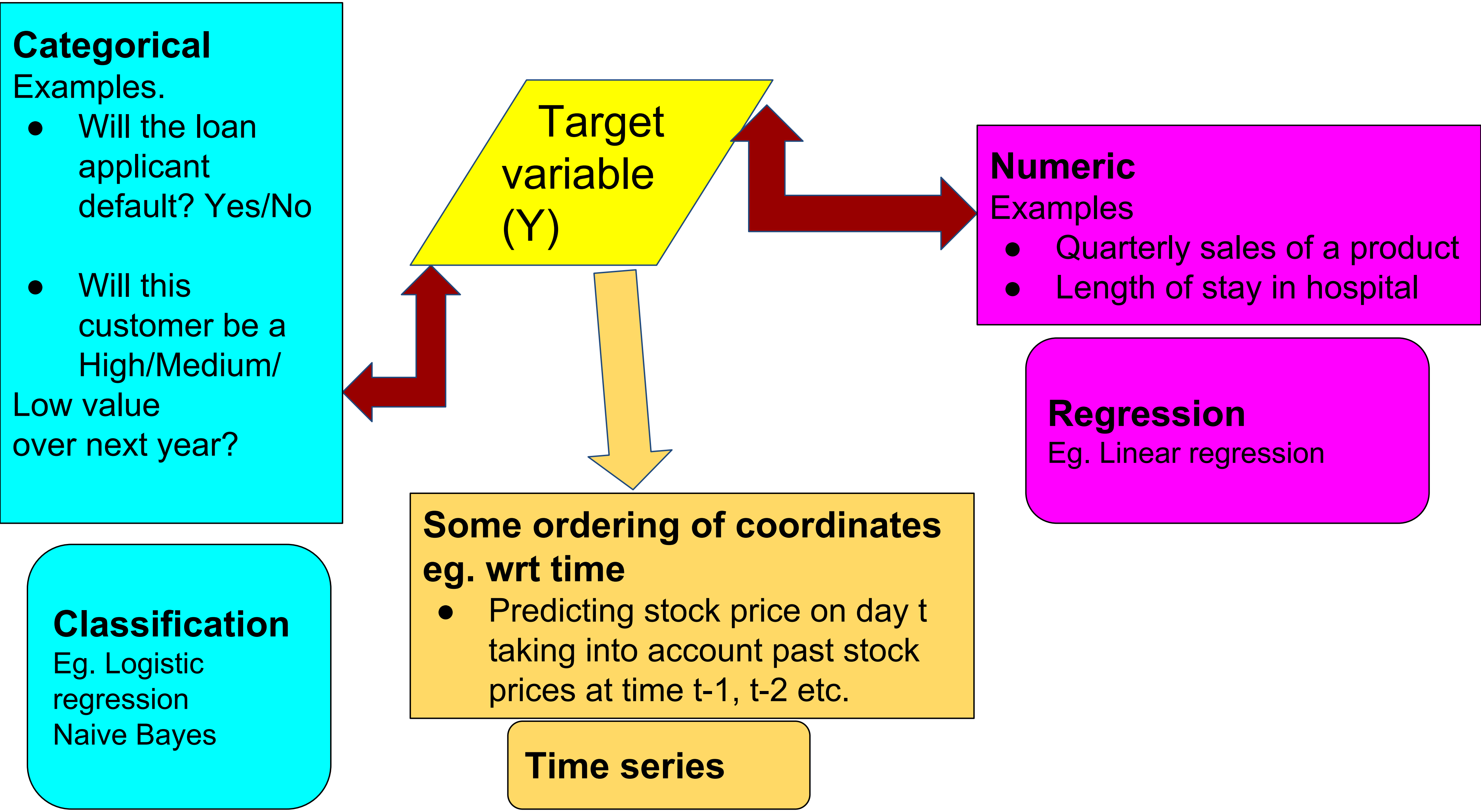
In many practical applications there are,

- some easy-to-measure quantities (generally called Xs)
  - Age; Gender; Income; Education level; etc.
- and a difficult-to-measure quantity (generally called the Y)
  - Amount of loan to give; Will she buy or not; How many days will he stay in the hospital; etc.

Common classes of practical learning problems

- Supervised learning is about computing the Y using the Xs, assuming availability of data samples with Xs and corresponding Ys (usually historical data)
- Unsupervised learning is about computing patterns within easy to measure attributes (the Xs)

Common supervised learning subtasks



Target variable (Y)

Categorical

Examples.

- Will the loan applicant default? Yes/No
- Will this customer be a High/Medium/

Low value over next year?

Classification

Eg. Logistic regression

Naive Bayes

Numeric

Examples

- Quarterly sales of a product
- Length of stay in hospital

Regression

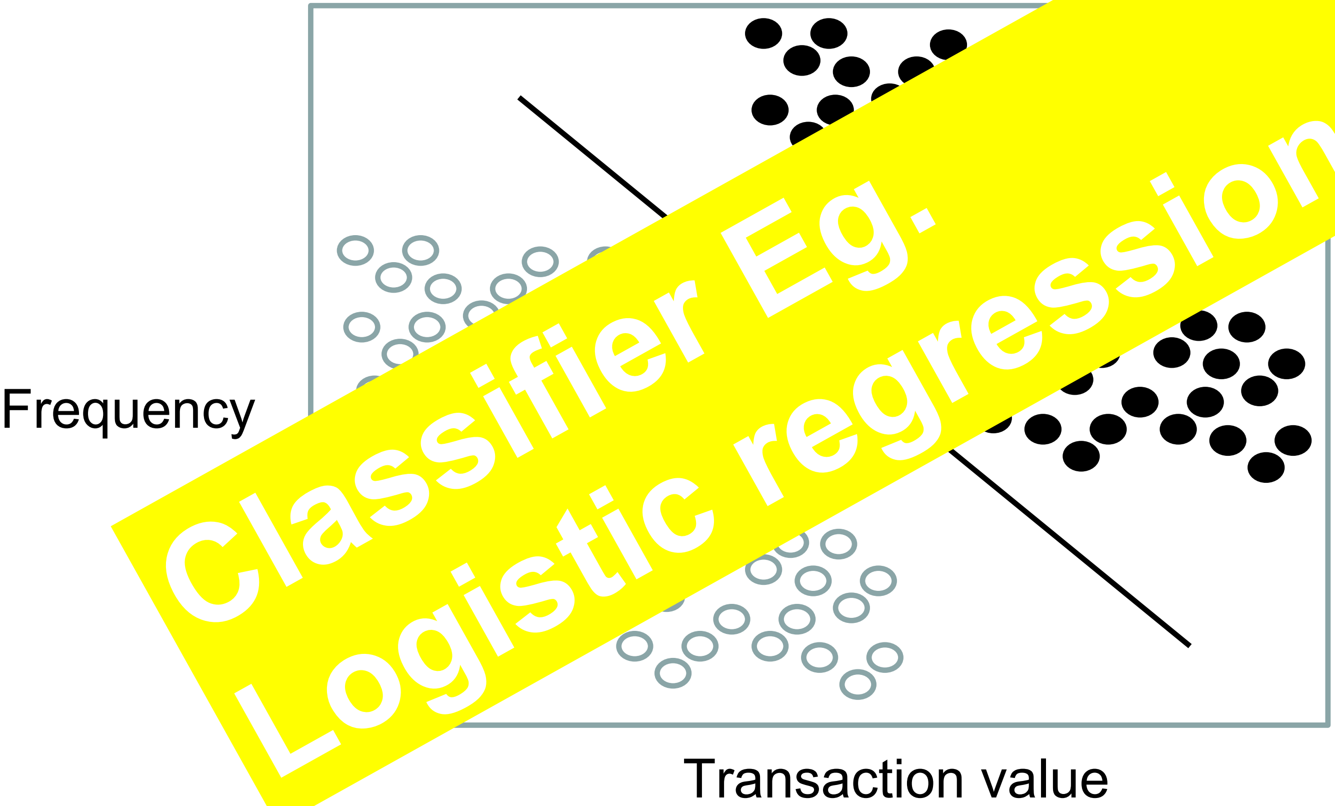
Eg. Linear regression

Some ordering of coordinates eg. wrt time

- Predicting stock price on day t taking into account past stock prices at time t-1, t-2 etc.

Time series

Learning Models: Classification



Classifier Eg.
Logistic regression

An illustration of classification with two attributes :

Xs : Transaction value, Frequency

Y (target variable) : High value/Low value customer

Learning Models: Regression

Generally,

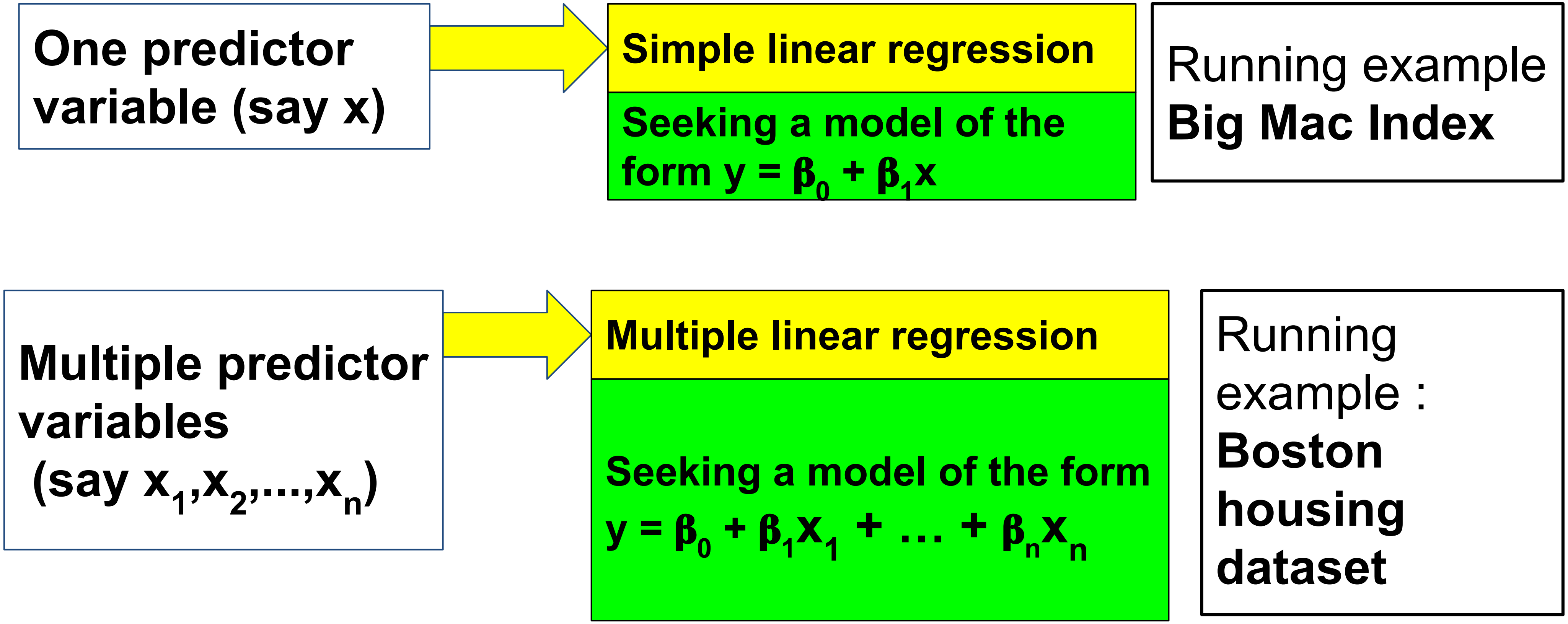
- **target variable** (also termed **response variable**) is a dependent variable representing something we are interested in predicting (and difficult to measure directly)

- **explanatory variables** (also termed **predictor variables**) are independent variables which are “easy” to measure

Suppose in the previous example,
Xs : Transaction value, Frequency
Y (the target variable) : Net worth of the individual

Learning Models: Linear Regression

Linear regression : one of the most commonly used method of regression.



One predictor variable (say x)

Simple linear regression

Seeking a model of the form $y = \beta_0 + \beta_1 x$

Running example **Big Mac Index**

Multiple predictor variables (say $x_1, x_2, \ldots, x_n$)

Multiple linear regression

Seeking a model of the form $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$

Running example : **Boston housing dataset**

Objectives : Linear Regression

To develop a good understanding of the following (with respect to Simple and Multiple Linear Regression) :

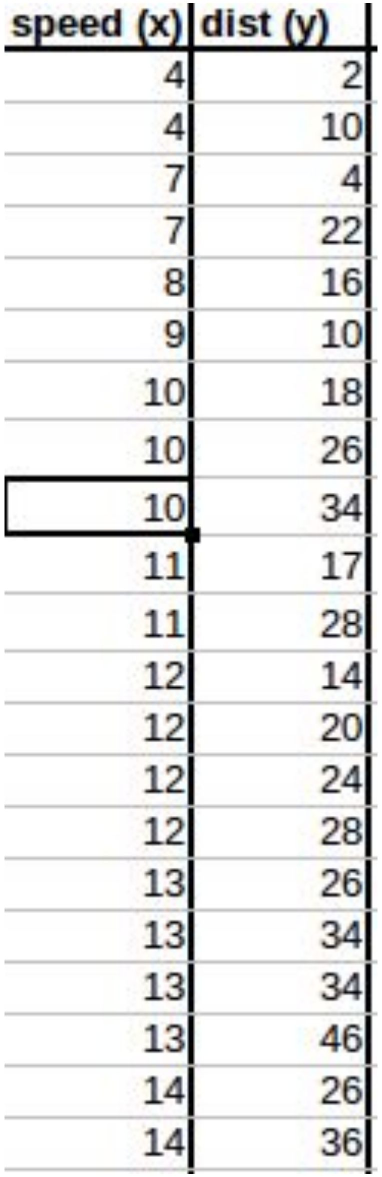
- Essential steps in **building and interpreting** a linear regression model
- **Diagnosing** and improving a model
- **Assumptions** made in linear regression and mechanisms to test whether these assumptions are violated in a given dataset
- **Awareness of common pitfalls**

Simple Linear Regression example

Problem : To predict stopping distance of a car given the speed.

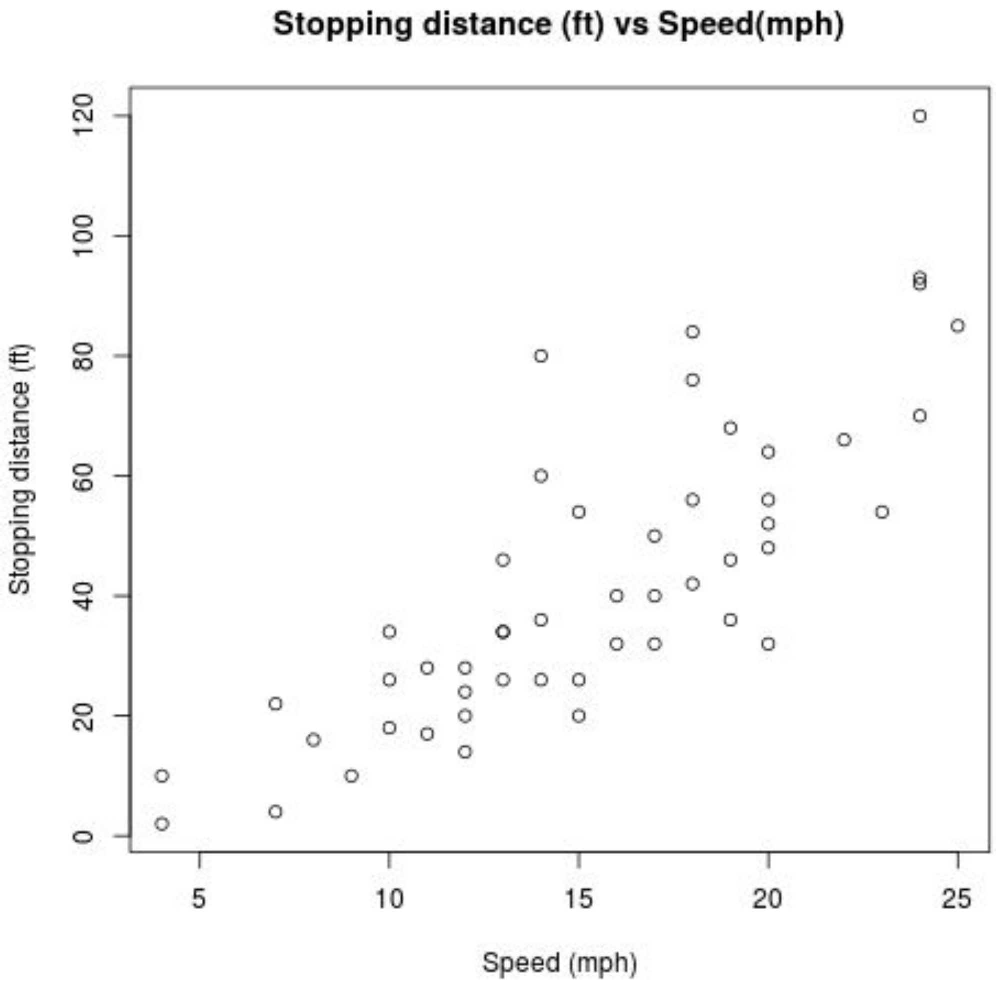
The “cars” dataset in R contains 50 pairs of datapoints for Speed(mph) vs stopping distance(ft), that were collected in 1920s.
See: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/cars.html

Snapshot of cars data

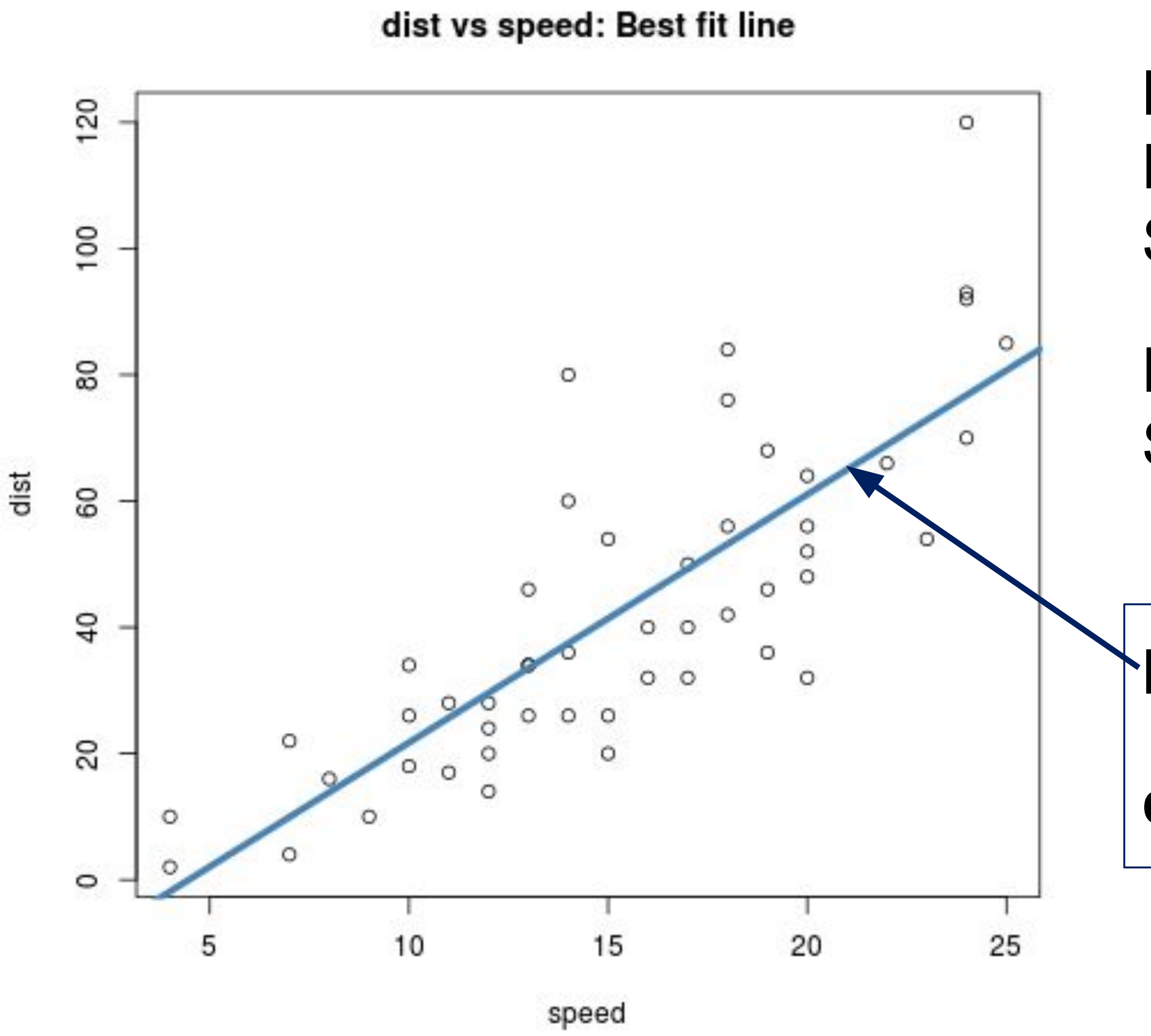


speed (x)	dist (y)
4	2
4	10
7	4
7	22
8	16
9	10
10	18
10	26
10	34
11	17
11	28
12	14
12	20
12	24
12	28
13	26
13	34
13	34
13	46
14	26
14	36

Plot of cars data



Simple Linear Regression example



Dataset : cars
Predictor variable (x-axis) :
Speed (mph)

Response variable (y-axis) :
Stopping distance (ft)

Line of best fit :

dist = 3.9324(speed) - 17.5791

Interpretation

For every 1 unit increase in speed (mph) , the stopping distance increases by approximately 4 units (ft).

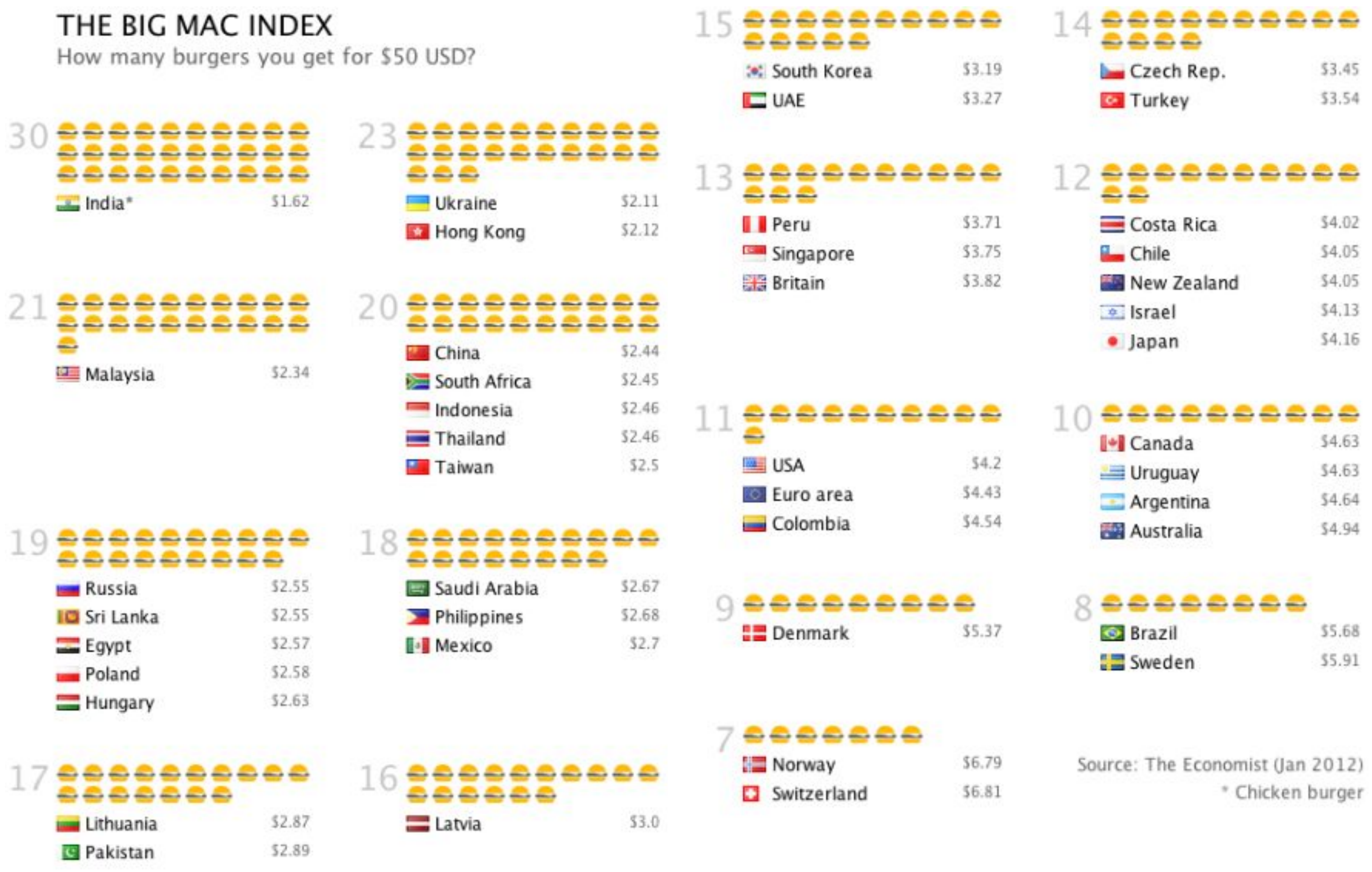
Sample output :R

```
Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791        6.7584  -2.601   0.0123 *
speed       3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Line of best fit :

dist = 3.9324(speed) - 17.5791

A running example : The Big Mac “Index”



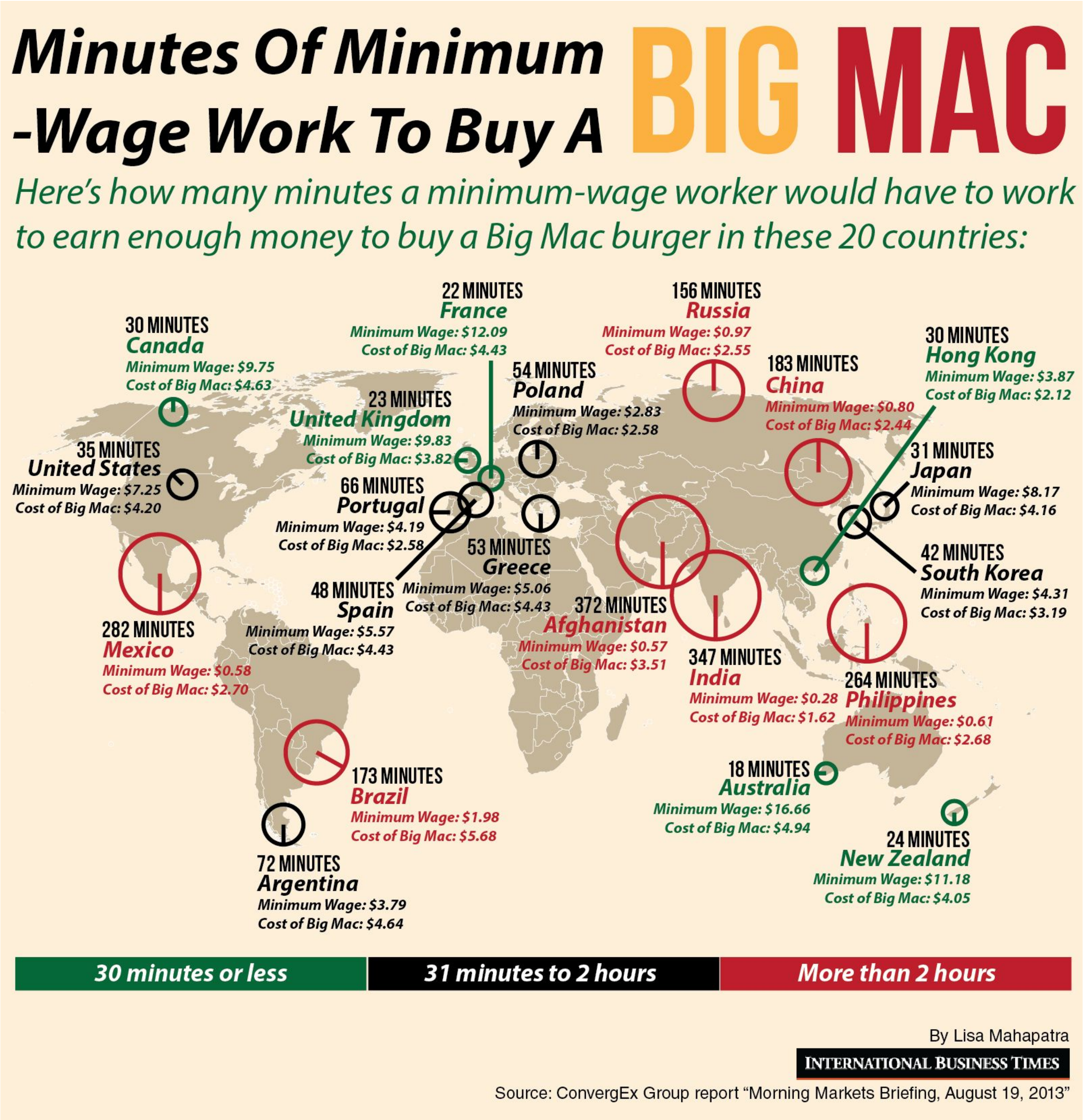
THE BIG MAC INDEX

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



Minutes Of Minimum -Wage Work To Buy A BIG MAC

Here's how many minutes a minimum-wage worker would have to work to earn enough money to buy a Big Mac burger in these 20 countries:

30 MINUTES **Canada**
Minimum Wage: $9.75
Cost of Big Mac: $4.63

22 MINUTES **France**
Minimum Wage: $12.09
Cost of Big Mac: $4.43

156 MINUTES **Russia**
Minimum Wage: $0.97
Cost of Big Mac: $2.55

183 MINUTES **China**
Minimum Wage: $0.80
Cost of Big Mac: $2.44

30 MINUTES **Hong Kong**
Minimum Wage: $3.87
Cost of Big Mac: $2.12

54 MINUTES **Poland**
Minimum Wage: $2.83
Cost of Big Mac: $2.58

23 MINUTES **United Kingdom**
Minimum Wage: $9.83
Cost of Big Mac: $3.82

35 MINUTES **United States**
Minimum Wage: $7.25
Cost of Big Mac: $4.20

66 MINUTES **Portugal**
Minimum Wage: $4.19
Cost of Big Mac: $2.58

31 MINUTES **Japan**
Minimum Wage: $8.17
Cost of Big Mac: $4.16

53 MINUTES **Greece**
Minimum Wage: $5.06
Cost of Big Mac: $4.43

42 MINUTES **South Korea**
Minimum Wage: $4.31
Cost of Big Mac: $3.19

48 MINUTES **Spain**
Minimum Wage: $5.57
Cost of Big Mac: $4.43

372 MINUTES **Afghanistan**
Minimum Wage: $0.57
Cost of Big Mac: $3.51

282 MINUTES **Mexico**
Minimum Wage: $0.58
Cost of Big Mac: $2.70

347 MINUTES **India**
Minimum Wage: $0.28
Cost of Big Mac: $1.62

264 MINUTES **Philippines**
Minimum Wage: $0.61
Cost of Big Mac: $2.68

173 MINUTES **Brazil**
Minimum Wage: $1.98
Cost of Big Mac: $5.68

18 MINUTES **Australia**
Minimum Wage: $16.66
Cost of Big Mac: $4.94

24 MINUTES **New Zealand**
Minimum Wage: $11.18
Cost of Big Mac: $4.05

72 MINUTES **Argentina**
Minimum Wage: $3.79
Cost of Big Mac: $4.64

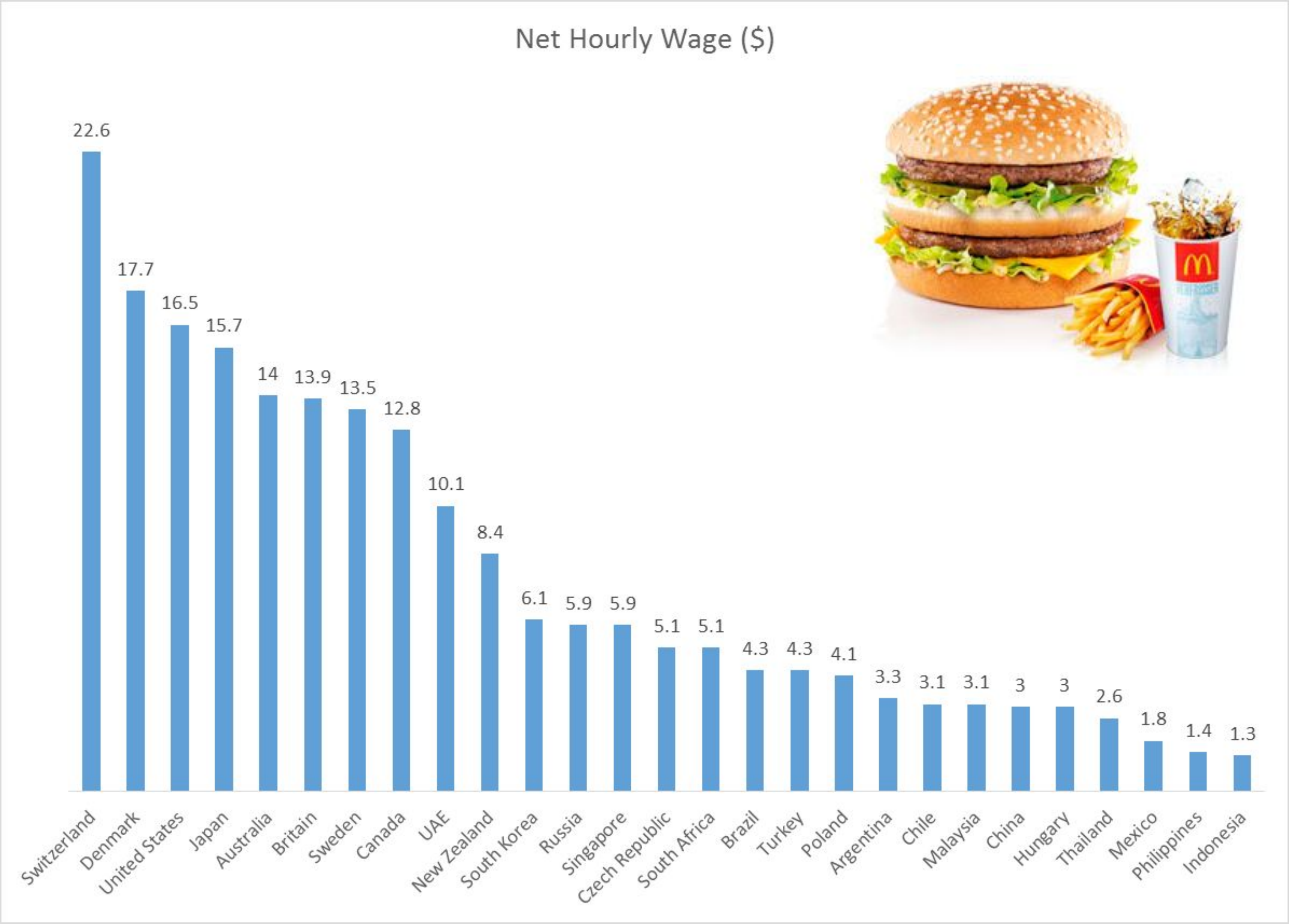
30 minutes or less	*31 minutes to 2 hours*	*More than 2 hours*

By Lisa Mahapatra
INTERNATIONAL BUSINESS TIMES

Source: ConvergEx Group report "Morning Markets Briefing, August 19, 2013"

Burgernomics by UBS Wealth Management Research

Net hourly wage ($) in various countries



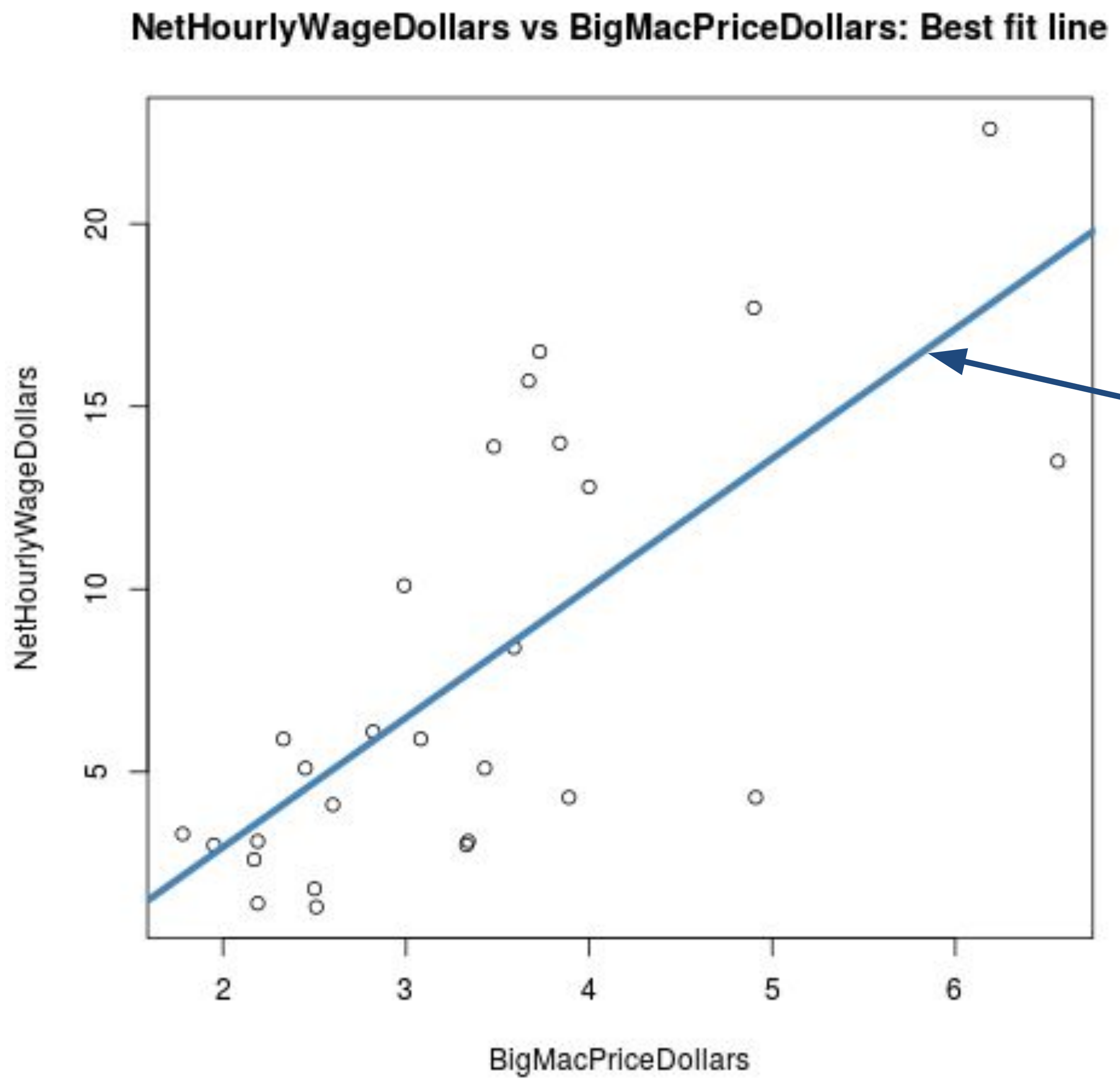
Problem :

How well can Net hourly wage be predicted from Big Mac prices?

Proposed technique :

Linear regression using Big Mac price ($) as a single predictor variable

Linear Regression : Predict net hourly wage

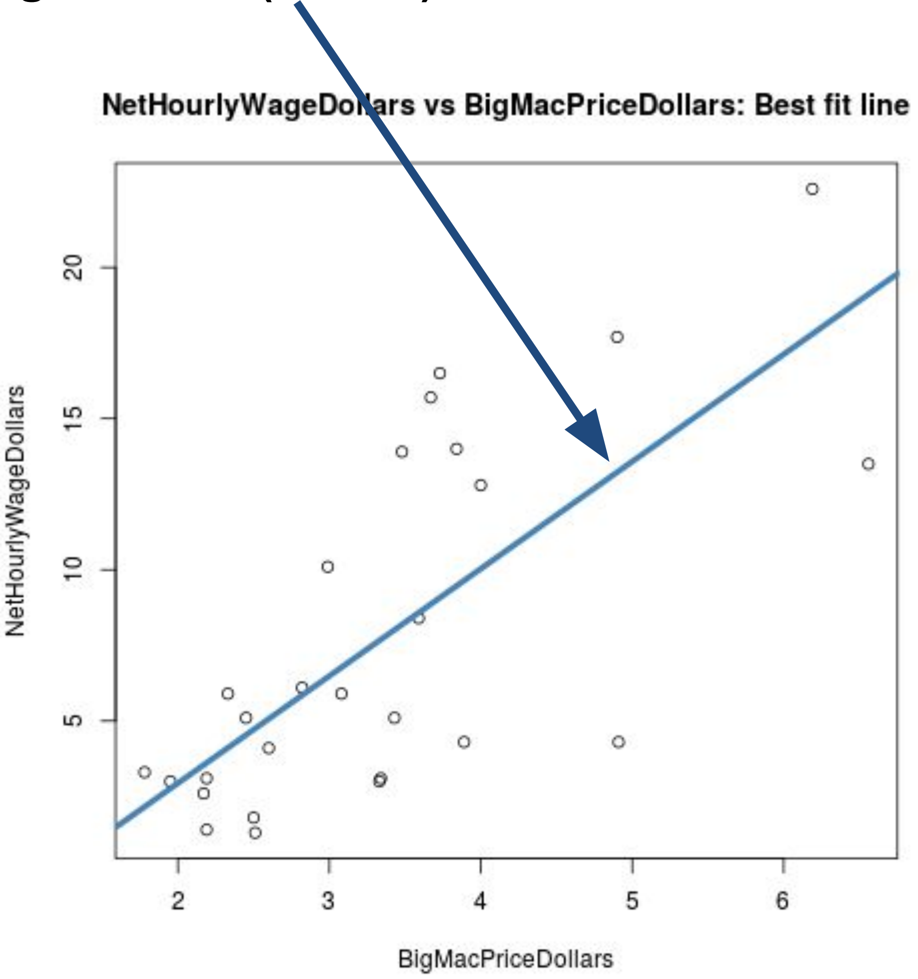


Equation of the best fit line

NetHourlyWage = BigMacPrice(3.5474)-4.1540

Pitfall : Extrapolating beyond scope of model

Line of best bit from regression :
NetHourlyWage = BigMacPrice(3.5474)-4.154091



Question : In the BigMac example, what is the predicted net hourly wage if Big Mac price is $1?

Substituting **BigMacPrice = 1** in regression equation yields
NetHourlyWage = 1(3.5474) - 4.154091
= **-0.6066 $**

Obtained answer is obviously incorrect.

Reason : Extrapolation done assuming the model holds even beyond the range of observed data

LINEAR REGRESSION : A few basic concepts

A few key terms

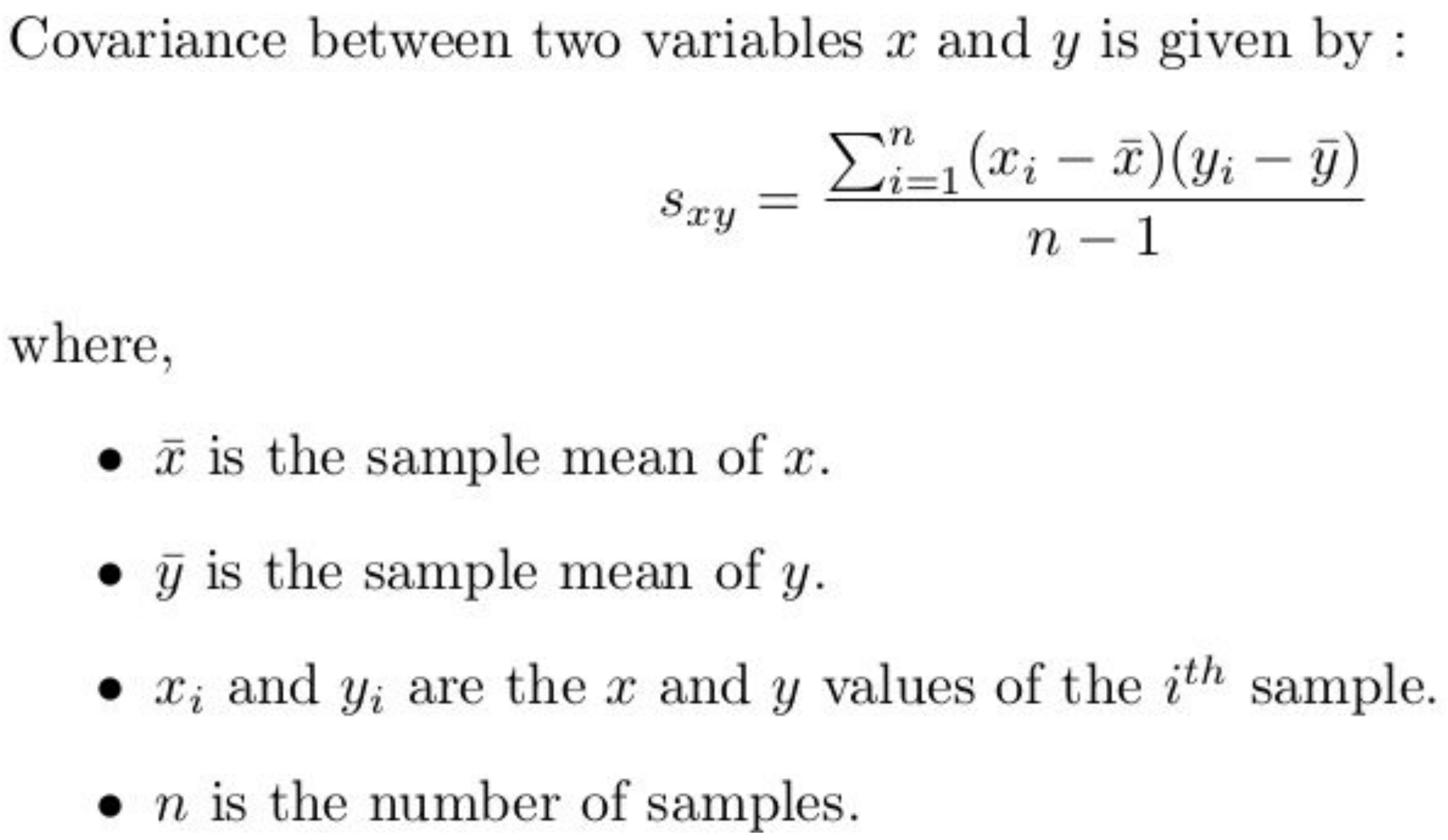
Background statistics terms :

- Sample, population
- Covariance and correlation
- Confidence intervals
- Hypothesis testing,p-value

Additional terms (today and next class)

- SSE,SST,SSR
- Coefficient of determination (Rsquared) and Adjusted R-Squared
- Residual errors
- Heteroscedasicity

Covariance



Covariance between two variables x and y is given by :

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

where,

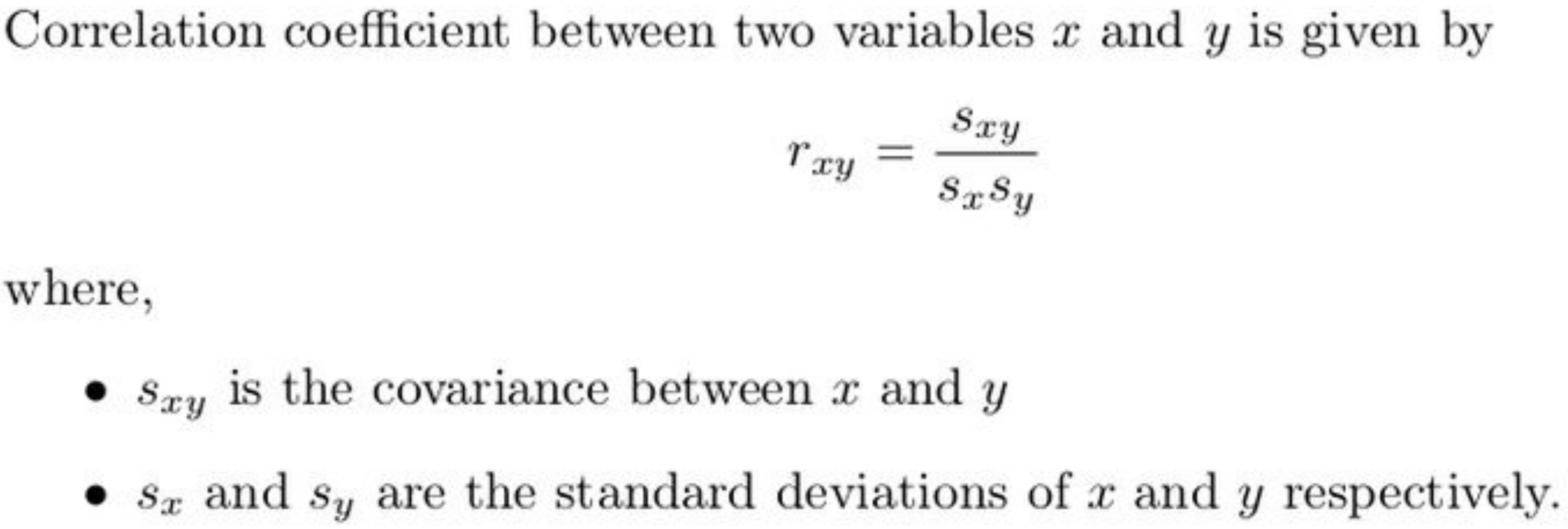
- $\bar{x}$ is the sample mean of x.
- $\bar{y}$ is the sample mean of y.
- x_i and y_i are the x and y values of the i^{th} sample.
- n is the number of samples.

Show and discuss Excel sheet computing covariance and correlation on cars data

Issues in interpreting covariance

- The value of the covariance only shows whether the variables vary in the same way (positive covariance) or in opposite directions (negative covariance).
- The value of the covariance depends heavily on the units used for measuring the variables and hence difficult to infer the strength of the relationship between the variables.
- Units are non-intuitive.

Correlation Coefficient



Correlation coefficient between two variables x and y is given by

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where,

- s_{xy} is the covariance between x and y
- s_x and s_y are the standard deviations of x and y respectively.

Correlation Coefficient

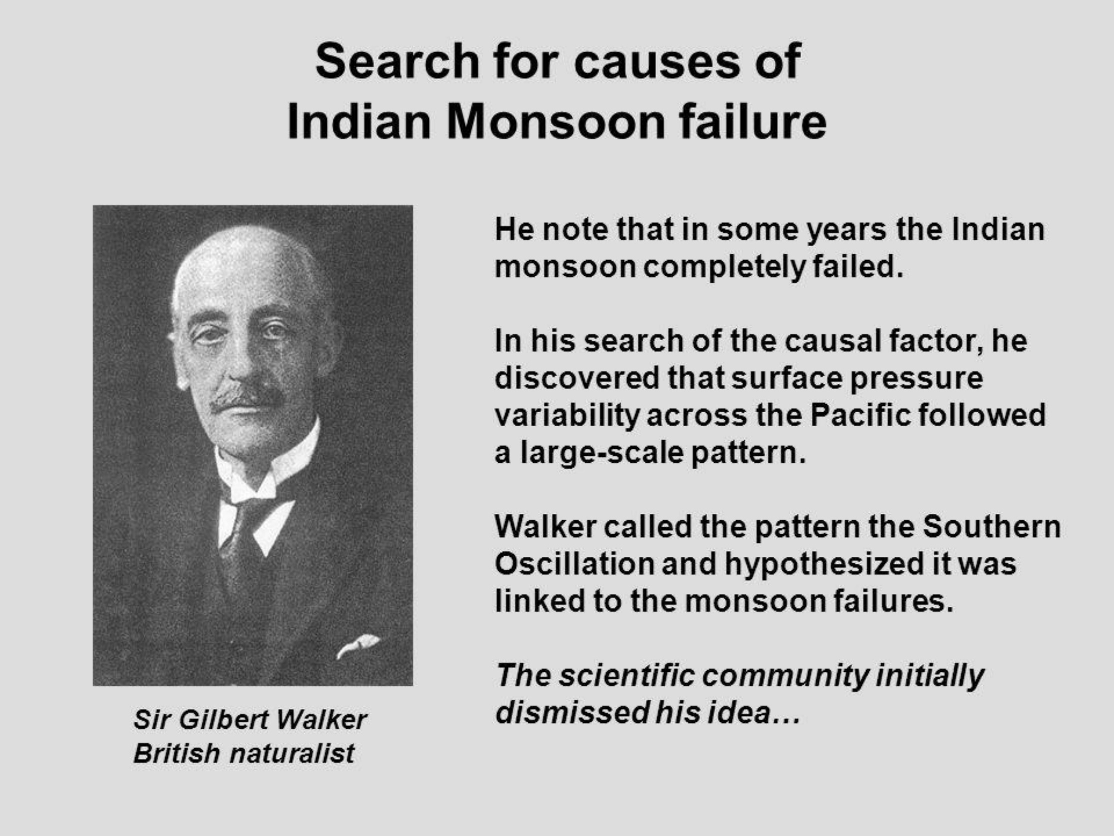
- Correlation coefficient r is a number between -1 and 1, whose magnitude indicates the strength of the relationship between the two variables.
    - Can be used to compare strength of relationship between different pairs of variables

- Correlation is dimensionless.
    - In fact covariance of standardized variables is the same as **correlation.**

Care to be exercised while applying correlation based analysis

Some correlations may indeed indicate a link between variables

- While establishing the correlation of the Indian rainfall with variables observed at various global locations, Walker (1923, 1924) discovered the Southern Oscillation, the North Atlantic Oscillation and the North Pacific Oscillation.



Search for causes of Indian Monsoon failure

Sir Gilbert Walker
British naturalist

He note that in some years the Indian monsoon completely failed.

In his search of the causal factor, he discovered that surface pressure variability across the Pacific followed a large-scale pattern.

Walker called the pattern the Southern Oscillation and hypothesized it was linked to the monsoon failures.

The scientific community initially dismissed his idea…

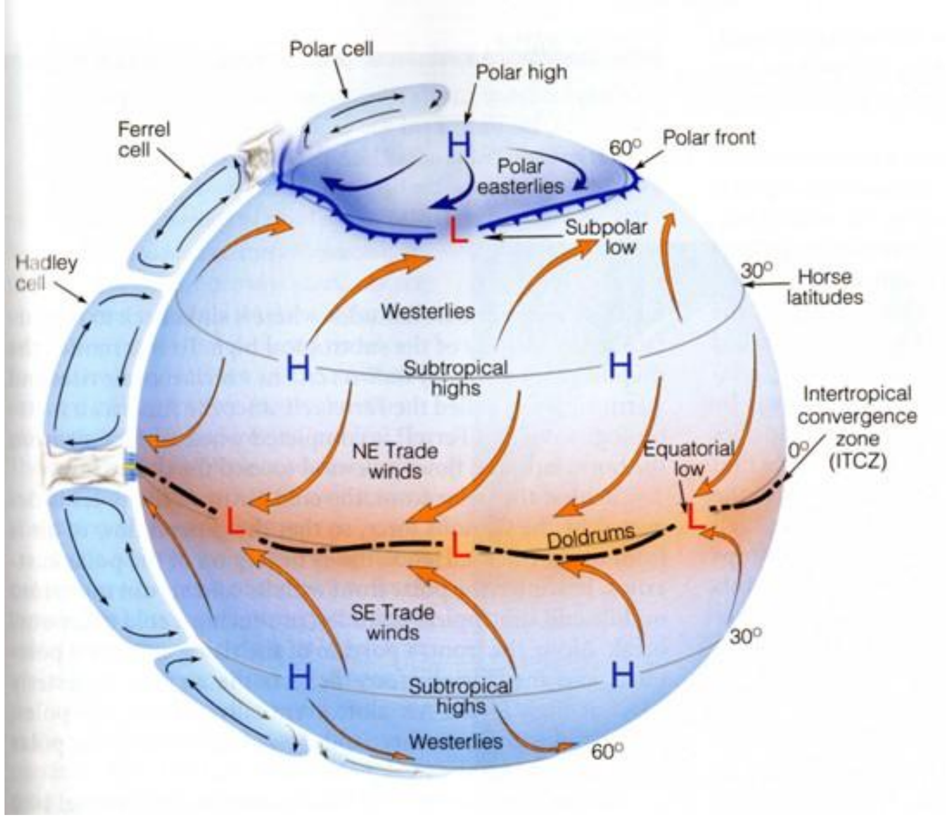


Image source :
http://slideplayer.com/slide/7972625/

Pitfall : Spurious correlations

- However it is possible to have spurious correlations as well.

Excerpt from an interview of Prof. Michael Jordan by Lee Gomes on behalf of IEEE Spectrum in October 2014

> **Michael Jordan:** I think data analysis can deliver inferences at certain levels of quality. But we have to be clear about *what* levels of quality. We have to have error bars around all our predictions. That is something that’s missing in much of the current machine learning literature.
>
> ***Spectrum***: What will happen if people working with data don’t heed your advice?
>
> **Michael Jordan:** I like to use the analogy of building bridges. If I have no principles, and I build thousands of bridges without any actual science, lots of them will fall down, and great disasters will occur. Similarly here, **if people use data and inferences they can make with the data without any concern about error bars, about heterogeneity, about noisy data, about the sampling pattern**, about all the kinds of things that you have to be serious about if you’re an engineer and a statistician—then you will make lots of predictions, and there’s a good chance that you will occasionally solve some real interesting problems. But you will occasionally have some disastrously bad decisions. And you won’t know the difference a priori. You will just produce these outputs and hope for the best.

Run Random correlations example code.

Note : Correlation only measures degree of linear dependence

- A low correlation or an inadequate fit of linear model **does not mean there is no functional relationship** between the variables.

(only means that the data is poorly explained by the linear model)

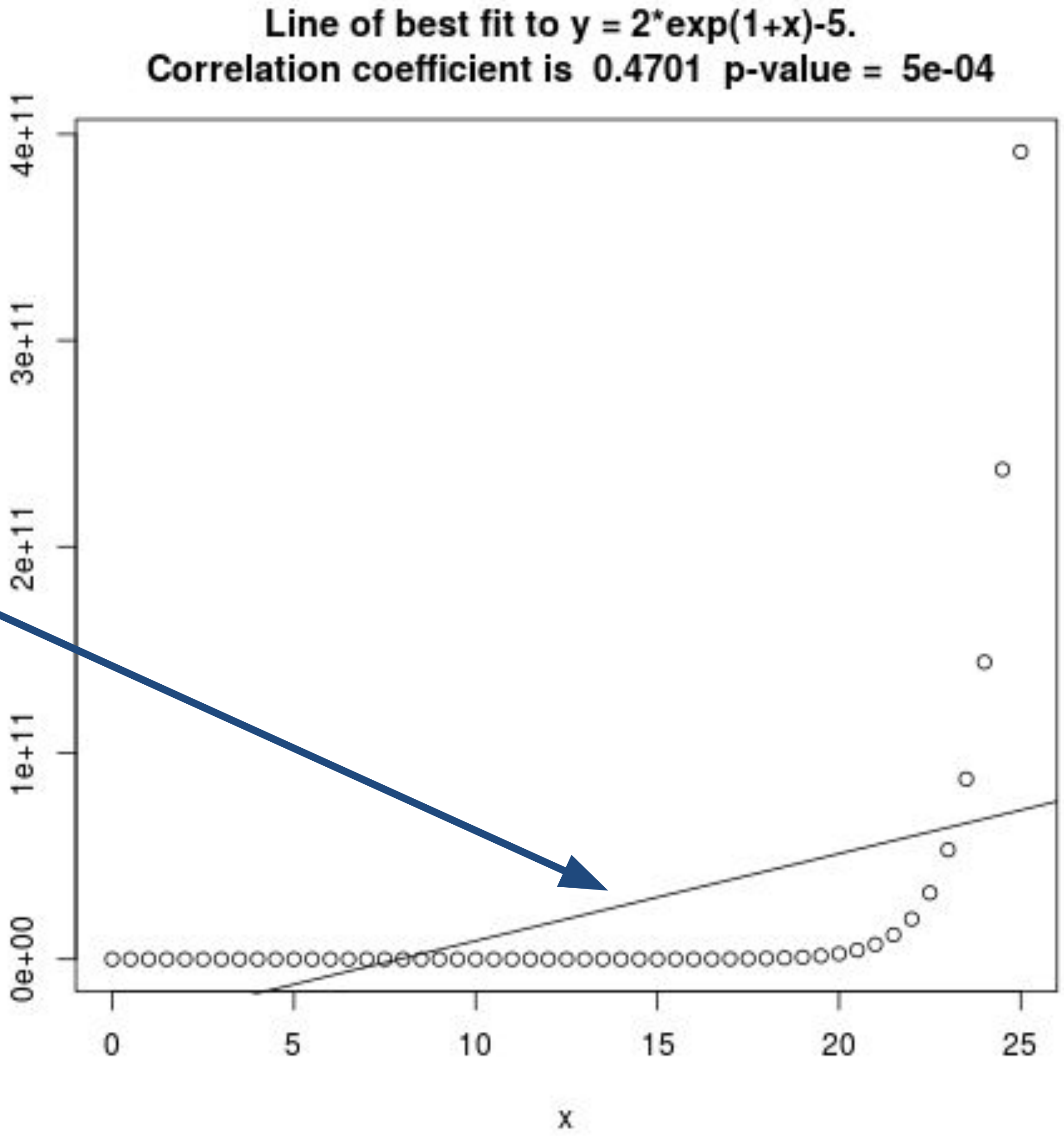
- Being able to fit a linear model does not necessarily mean model is good.

Example : Run code on fitting a line to y = e^(1+x)

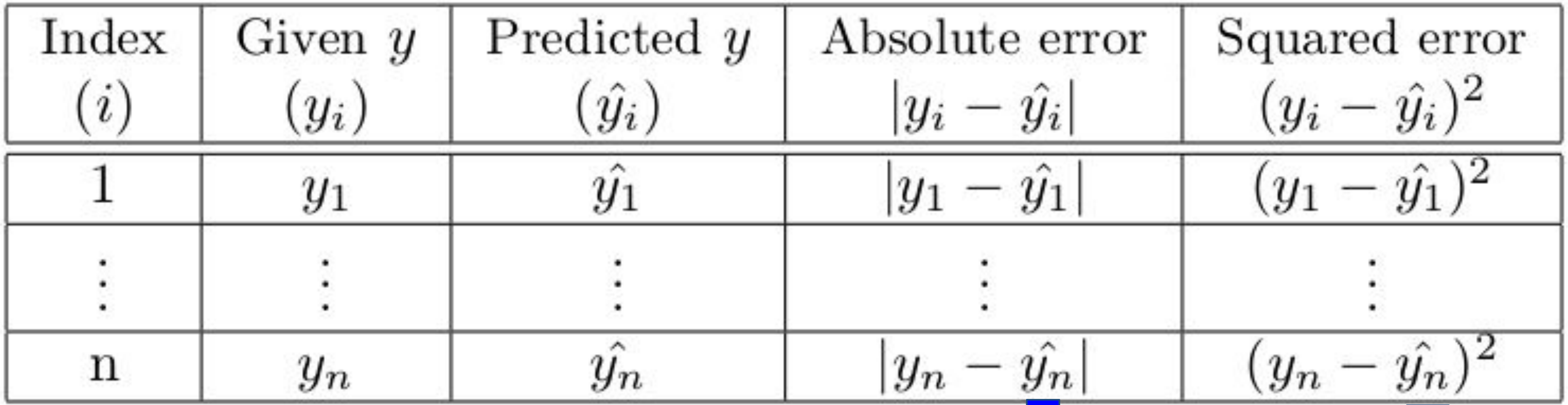
Line of best fit :
y = 4.224e+09 x - 3.329e+10

Correlation coefficient :
0.4701

p-value : 5e-04



A small detour : Commonly used error metrics



Index (i)	Given y (y_i)	Predicted y $(\hat{y_i})$	Absolute error $\lvert y_i - \hat{y_i}\rvert$	Squared error $(y_i - \hat{y_i})^2$
1	y_1	$\hat{y_1}$	$\lvert y_1 - \hat{y_1}\rvert$	$(y_1 - \hat{y_1})^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	$\hat{y_n}$	$\lvert y_n - \hat{y_n}\rvert$	$(y_n - \hat{y_n})^2$

Table illustrating computation of error metrics

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |(y_i - \hat{y_i})|$$

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

Commonly used error metrics

- Mean Absolute Percent Error (MAPE)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(y_i - \hat{y}_i)}{y_i}\right|$$

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Root Mean Squared Error (RMSE)

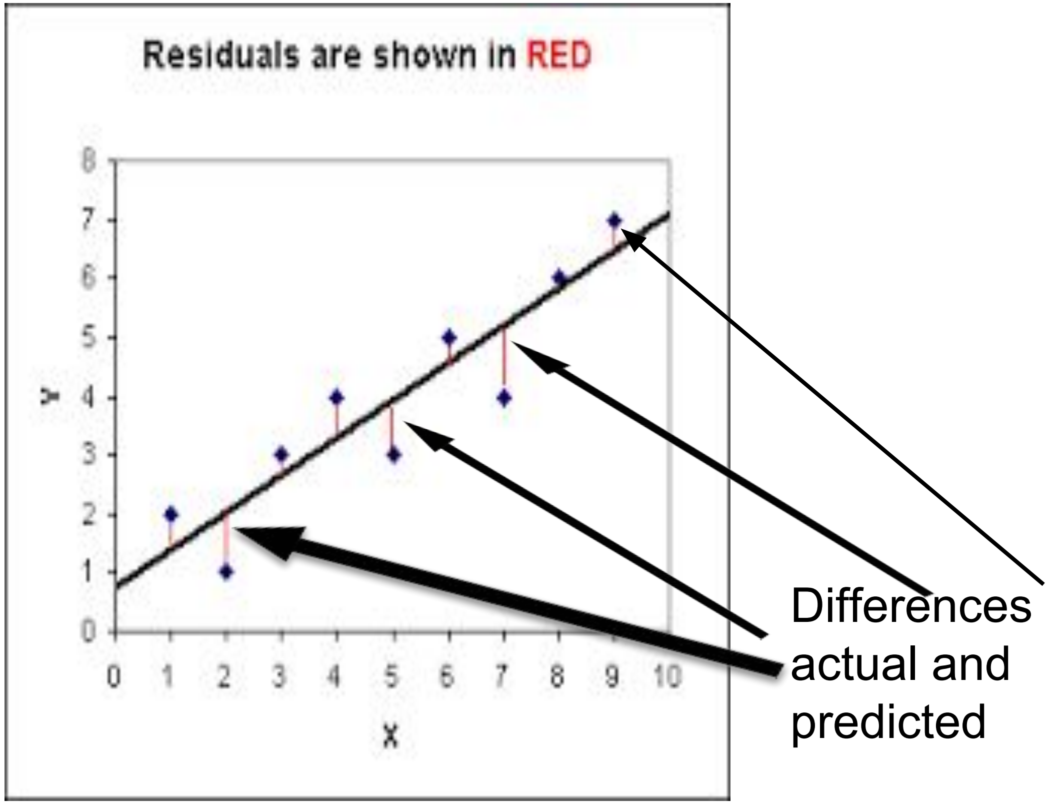
$$RMSE = \sqrt{MSE}$$

- Sum of Squared Errors (SSE)

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Caution : SSE is not a suitable error metric when comparing performance across data sets of different sizes

Computing the line of best fit



Differences b/w actual and predicted

Predictor variable (x)	Response variable (y)	Predicted value $\hat{y}$
x_1	y_1	$\beta_0+\beta_1 x_1$
x_2	y_2	$\beta_0+\beta_1 x_2$
…	…	…
x_N	y_N	$\beta_0+\beta_1 x_N$

In **Ordinary Least Squares (OLS), line of best fit** is one that **minimizes the sum of squared errors.**

Computing the line of best fit

The **Sum of Squared Errors** is given by

$$SSE = \sum_{i} (y_i - \hat{y_i})^2$$

where,

y_i is the actual y value of the i^{th} sample.

$\hat{y_i}$ is the predicted y value for the i^{th} sample.

When there is a single predictor variable, $\hat{y_i} = \beta_0 + \beta_1 x_i$

and hence

$$SSE = \sum_{i} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Computing the line of best fit

Taking partial derivatives of SSE wrt to the parameters and using first order minimization conditions we have :

$$\frac{\partial(SSE)}{\partial\beta_0} = 0$$

$$\frac{\partial(SSE)}{\partial\beta_1} = 0$$

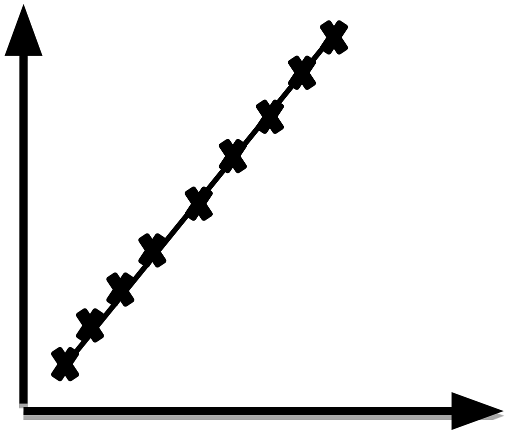
Solving the resulting system of equations and using some algebra (not shown) we get :

$$\beta_1 = \frac{\sum y_i x_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{cov(x,y)}{s_{xx}}$$

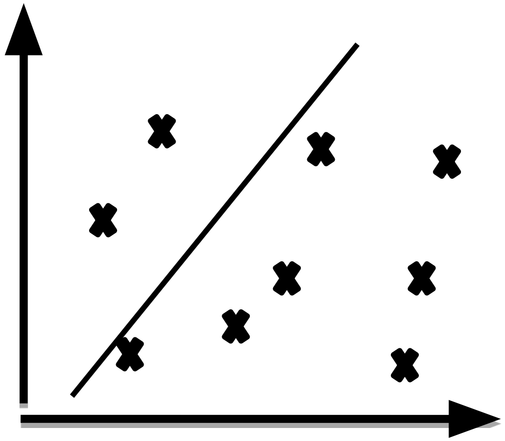
$$= r \times \frac{s_y}{s_x}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

But how do you know how good the best fit line is?



Accurate Linear Correlation



No Linear Correlation

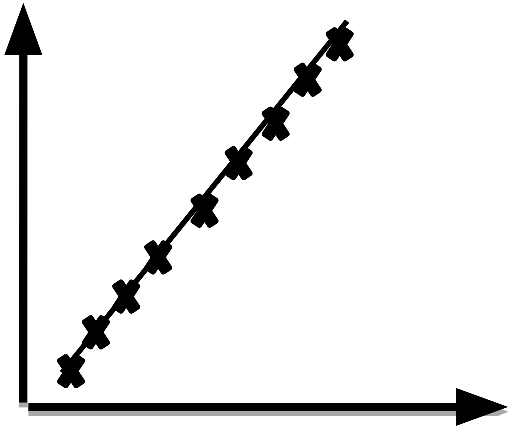
Basic measures of goodness of the fit :

- The **correlation coefficient (r)**
- **Coefficient of determination** (R^2)

Caveat : While above are indicative measures of goodness of fit, they are not sufficient for a systematic assessment of the model.

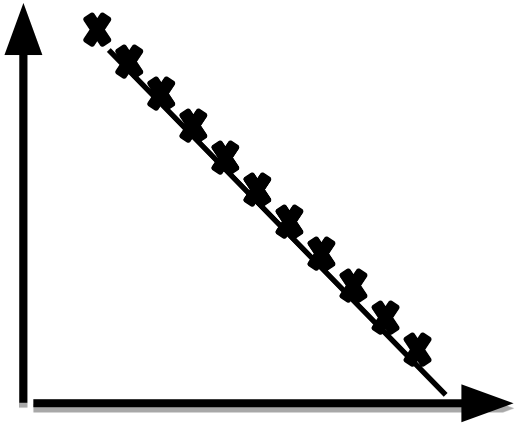
Correlation Coefficient and regression

- Correlation coefficient, *r*, is a number between -1 and 1.
- It gives the strength and direction of the relationship between two variables.



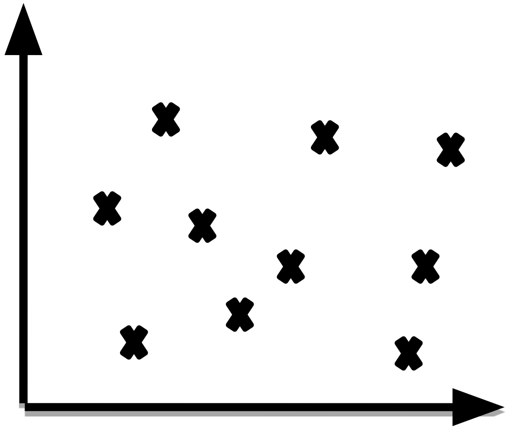
r = 1

Positive Linear Correlation



r = -1

Negative Linear Correlation



r =0

No Correlation

Coefficient of determination

- **Coefficient of determination** R^2 is the fraction (percentage) of variation in the response variable that is explainable by the predictor variable(s).

- R^2 ranges between 0 (no predictability) to 1 (or 100%) which indicates complete predictability

- A high R^2 indicates being able to predict response variable with less error.

Coefficient of determination

SST = Total variation in the data = Sum of Squares Total	SSR = Sum of Squares Regression = Variation explained by the model	SSE = Unexplained variation in the data = Sum of Squared Errors = Sum of Squares Within (from ANOVA)
$SST = \sum_i (y_i - \bar{y})^2$	$SSR = \sum_i (\hat{y}_i - \bar{y})^2$	$SSE = \sum_i (y_i - \hat{y}_i)^2$

where,

- y_i is the actual y value of the i^{th} sample.
- $\bar{y}$ is the sample mean of y.
- $\hat{y}_i$ is the predicted y value for the i^{th} sample.

Coefficient of determination

The coefficient of determination R^2 is given by : $$R^2 = \frac{SSR}{SST}$$

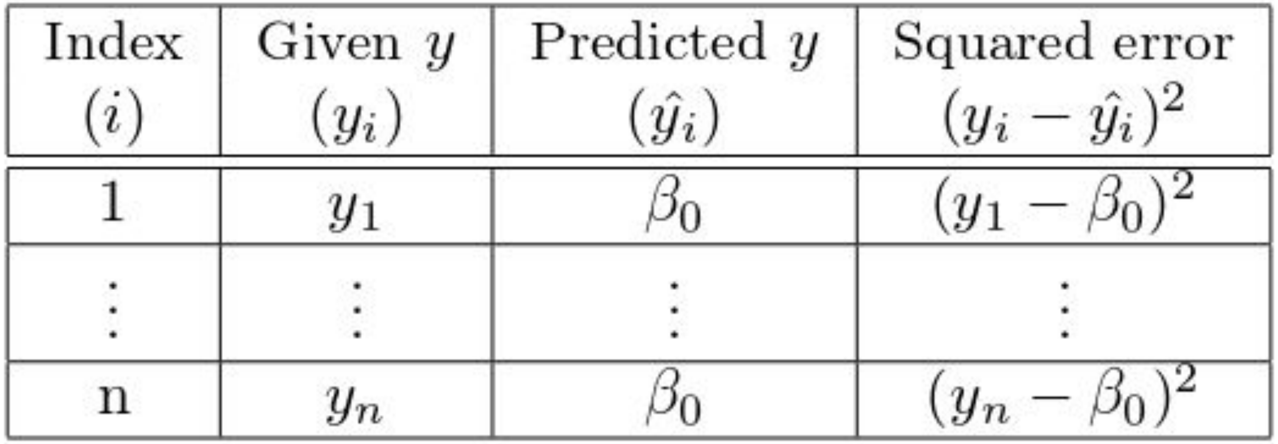
where SST,SSR and SSE are as specified previously.
Since SST = SSE + SSR (stated without proof), we have :

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

Show Excel file illustrating RSquared computation for cars data

Using a reduced model as a baseline

Suppose we seek to fit an **intercept only** model i.e.**a reduced model** of the form $y = \beta_0$



Index (i)	Given y (y_i)	Predicted y $(\hat{y}_i)$	Squared error $(y_i - \hat{y}_i)^2$
1	y_1	β_0	$(y_1 - \beta_0)^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	β_0	$(y_n - \beta_0)^2$

Sum of Squared Errors (SSE) is given by

$$SSE = \sum_{i=1}^{n} (y_i - \beta_0)^2$$

Differentiating wrt β_0 and equating to 0 it can be shown that the estimated value of β_0 that minimizes the Sum of Squared Errors above is given by :

$$\hat{\beta_0} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Analysis and assessment of the model

- **Question : How good is the model?**
- A basic assessment of the model can be obtained by reading off the R^2 and adjusted R^2 values.
- **Caution :** A good R^2 value alone can be misleading.

- **Question : Is the model significant?**
- **Important :** Both model significance as well as significance of the individual coefficients need to be considered.

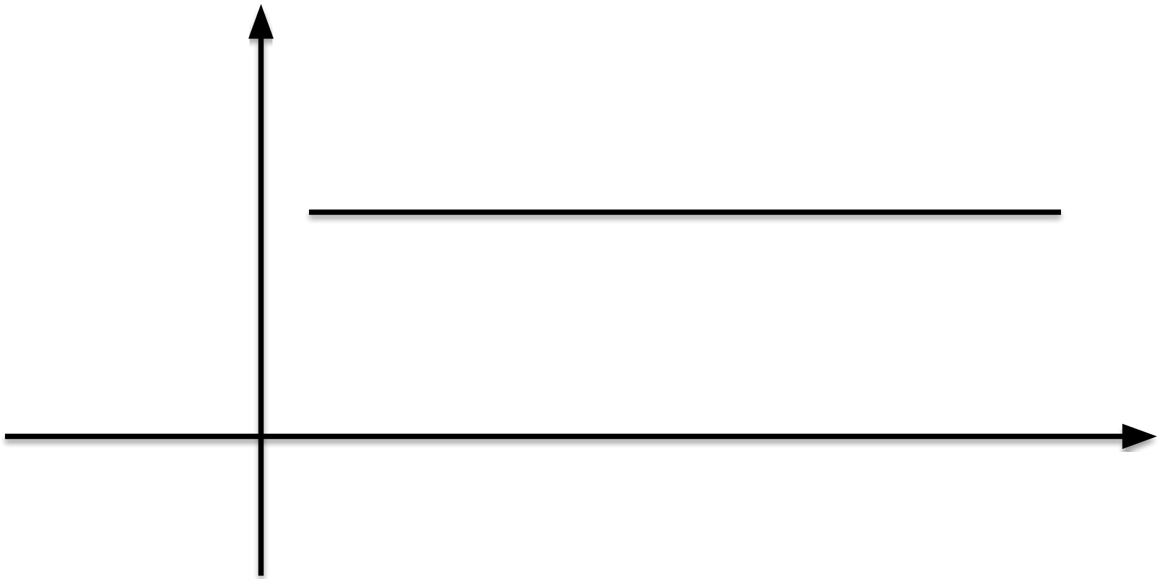
- **Important :** Verifying that the assumptions are not seriously violated is critical to interpreting the model.
- **Tool : Residuals.** Verify using the relevant residual plots.

- **Question : How to deal with outliers/influential points.**
- One possible approach : Build a model, identify outliers and rebuild
- **Mechanisms to identify outliers/influential points : Leverage statistics, Cook's distance**

Hypothesis tests for slope of 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.



t Test of the Slope

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

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

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

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

Standard Error of the Estimate

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

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

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

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

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

t Test of the Slope – Big Mac - Excel

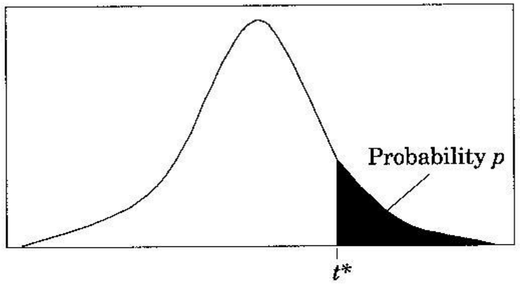
$t = 5.1437$

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^*.



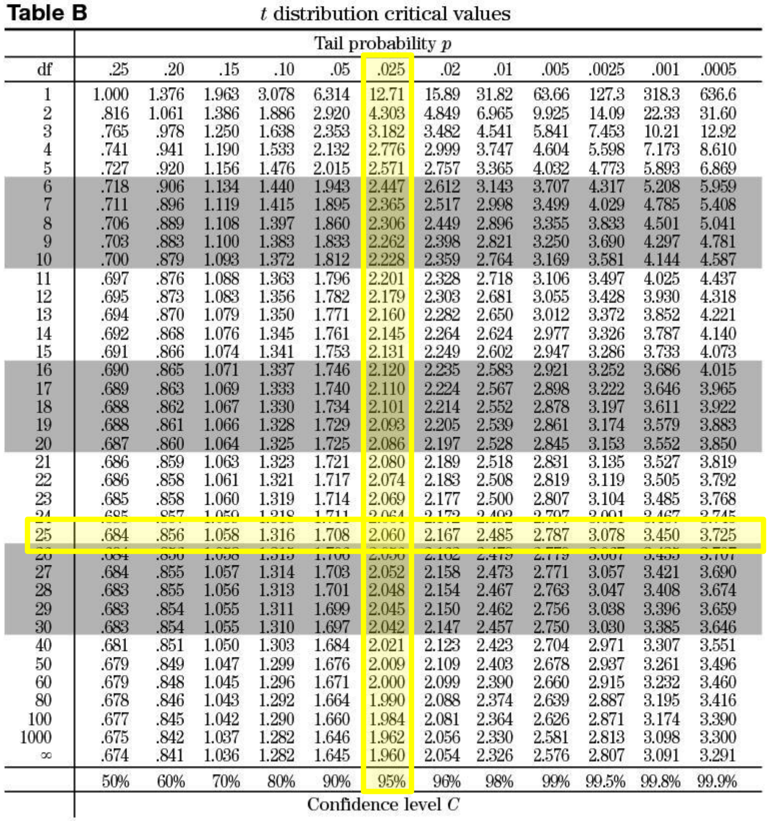


Table B *t* distribution critical values

	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											

Assessing the overall model

- **F test** and its associated **ANOVA table** is used to test the overall model.
  - In simple regression, we have only one coefficient. So F test for overall significance tests the same thing as t test.
    - (Null Hypothesis) $H_0 : \beta_1 = 0$
    - (Alternate hypothesis) $H_1 : \beta_1 \neq 0$
  - In multiple regression, it tests that at least one of the regression coefficients is different from 0.
    - (Null Hypothesis) $H_0 : \beta_1 = \beta_2 = \ldots \beta_k = 0$
    - (Alternate hypothesis) H_1 : At least one among $\beta_1, \beta_2 , \ldots\beta_k \neq 0$

Assessing the overall model

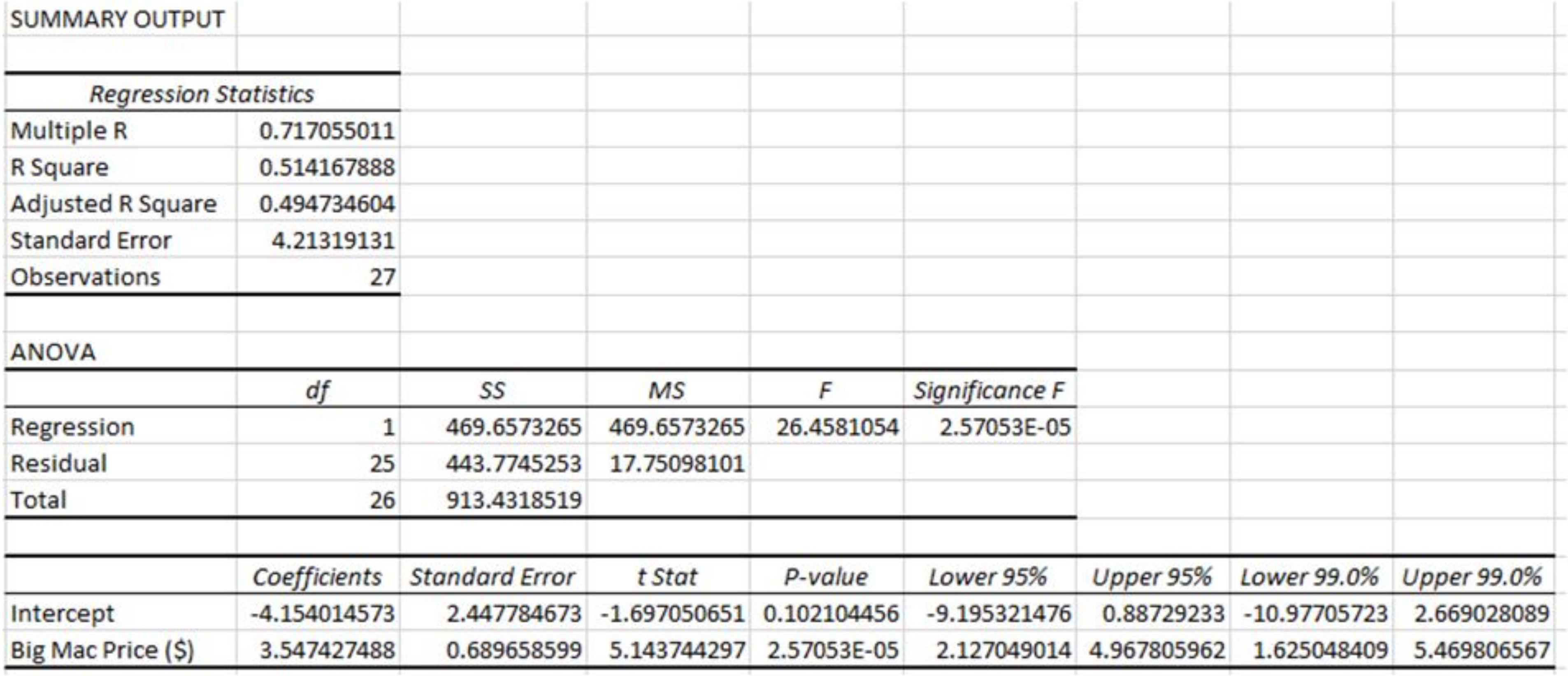
The F-statistic is given by

$$F = (SSR/df_{reg}) \;/\; (SSE/df_{err})$$

where

- k is the number of independent variables
- n is the number of samples
- $df_{reg} = k$
- $df_{err} = n-k-1$

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

R output : NetHourlyWage ($) vs BigMacPrice ($)

```
> summary(lineFit)
Call: lm(formula = NetHourlyWageDollars ~
BigMacPriceDollars, data = BigMac)

Residuals:
    Min      1Q  Median   3Q   Max
-8.9639 -2.9141 -0.1813  3.2058  7.4221

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -4.1540  2.4478  -1.697     0.102
BigMacPriceDollars   3.5474   0.6897   5.144 2.57e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.213 on 25 degrees of freedom

Multiple R-squared:  0.5142,    Adjusted R-squared:  0.4947

F-statistic: 26.46 on 1 and 25 DF,  p-value: 2.571e-05
```

Significance of individual coefficients

Goodness of fit : R-Squared, adjusted R-squared

Model significance

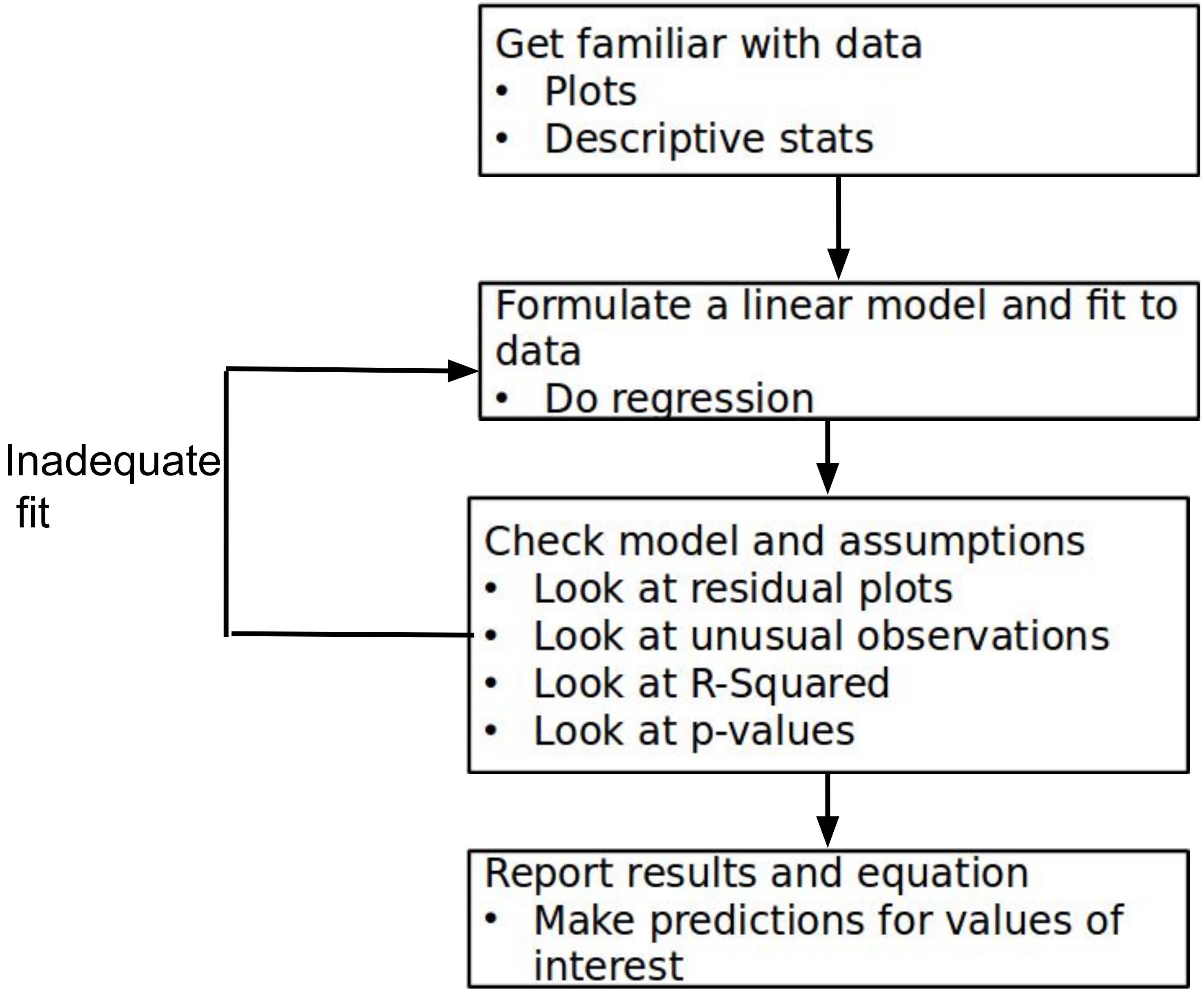
Lab Part 1

- Basic understanding of linear regression with examples (cars and Big Mac “Index)

Concepts emphasized

- Covariance and correlation with Excel and R.
- Assess goodness of fit and significance of the model
  - **Goodness of fit** : R squared, adjusted R squared
  - **Significance** of the model : ANOVA, p-value
  - **Significance** of each of the coefficients: Standard error, t statistic

Simple Linear regression : Typical flow



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

Report results and equation

- Make predictions for values of interest

Inadequate fit

Linear regression : Outline of steps

Step 1 : Building a linear regression model : Typically straightforward once data is available in the required format.

- R : lm
- python : Eg. LinearRegression from sklearn.linear_model

Step 2 : Testing of the model : Test whether a linear association exists between the predictor x and the response y in a simple linear regression model.

H0: $\beta1 = 0$ versus HA: $\beta1 \neq 0$.

Step 3 : Diagnose the model : A more detailed evaluation, this is generally the most time consuming portion of the overall analysis.

- R-Squared, adjusted R-squared values
- Examine residual plots, check whether the assumptions of linear regression are violated.

Caution : High R-squared alone is insufficient

- **Caveat :** Do not seek to improve R^2 alone in pursuit of a better model.
  - Perform systematic analyses using **residual plots**.
  - Adding more terms generally improves R-squared value, better to use adjusted R-squared value since it takes into account model complexity to some extent.

- A high R-squared value alone is insufficient to conclude that the model is good.
  - Also in some applications a low R-squared value is not necessarily bad.

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.

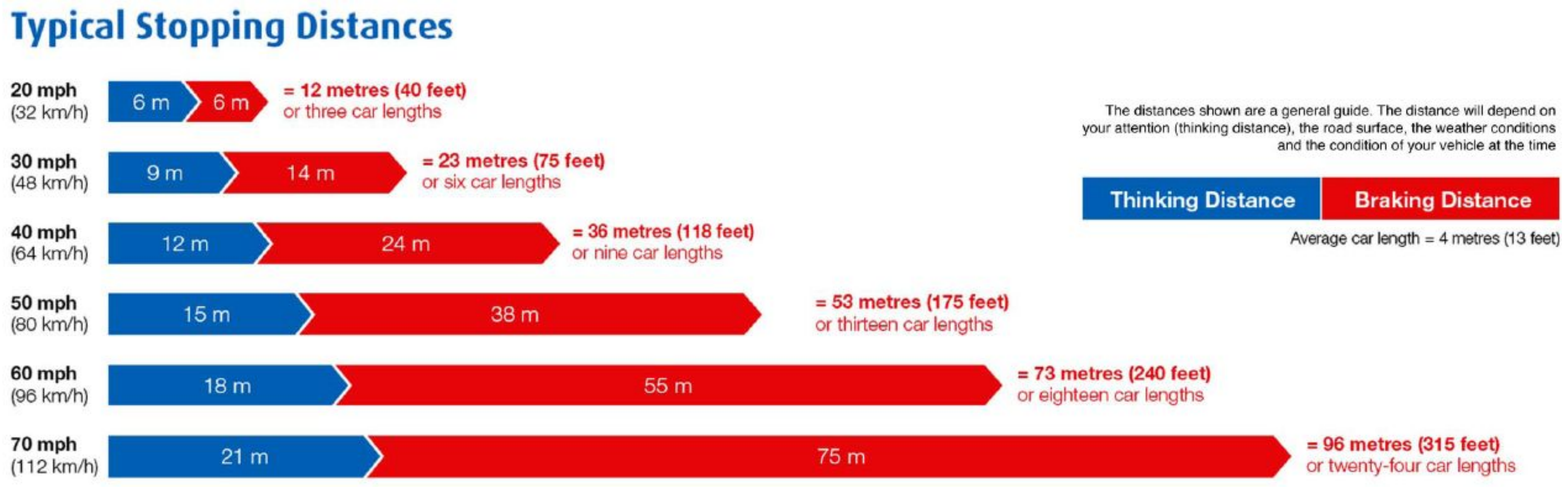
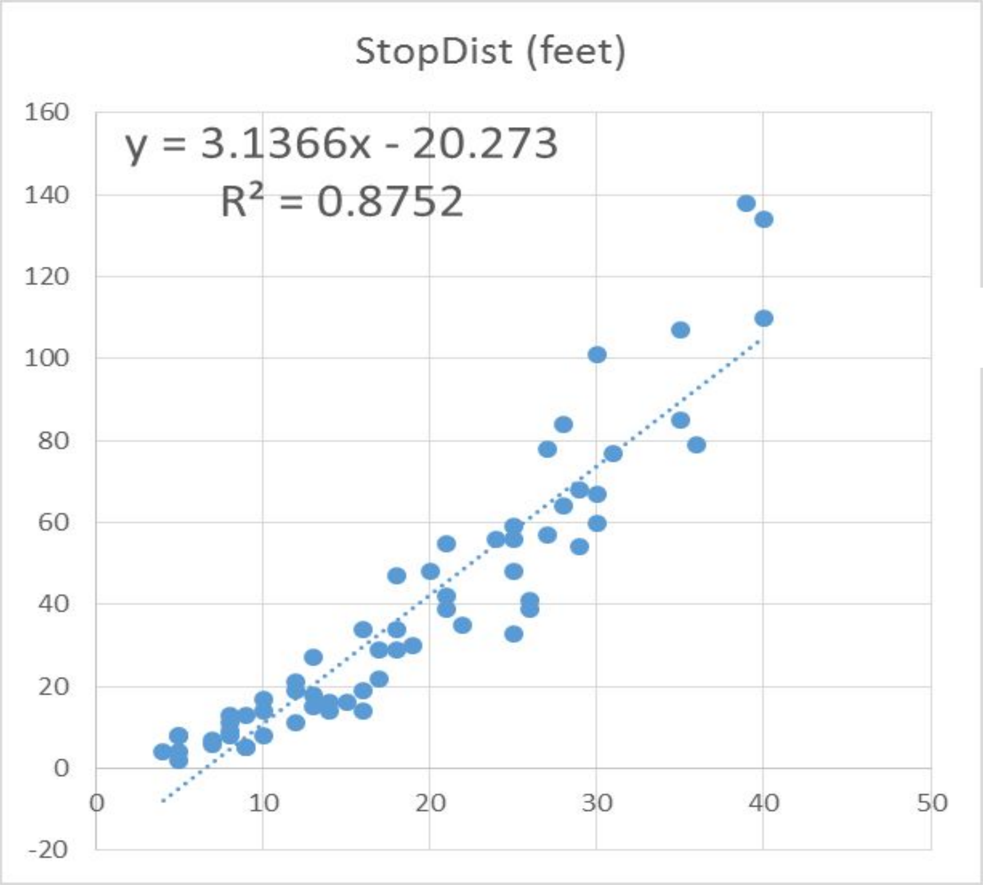


Image Source: http://streets.mn/2015/04/02/the-critical-ten/
Last accessed: November 20, 2015

R-Squared, Significance and Residuals - Caution

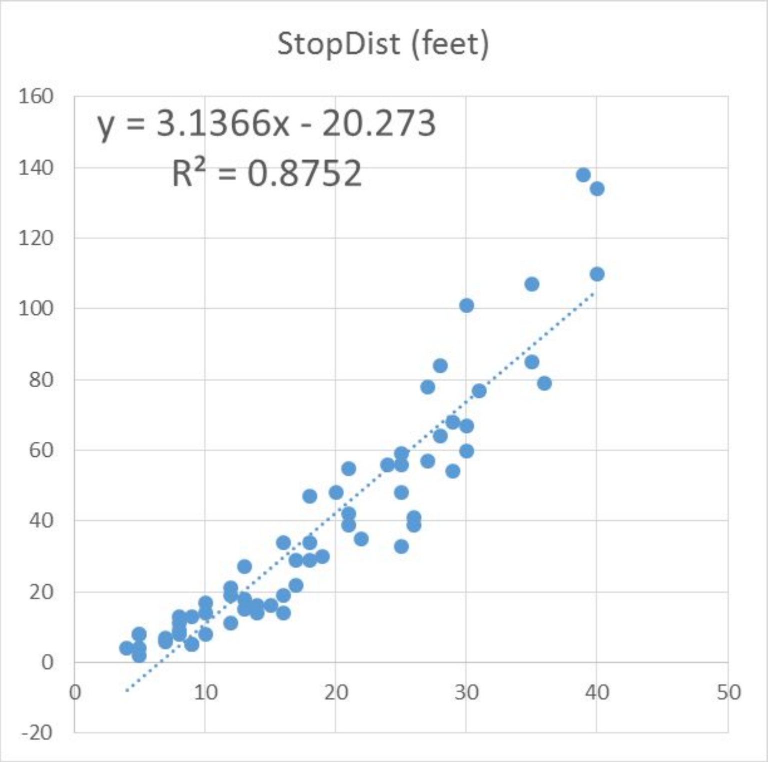
Does the estimated regression line fit the data well?

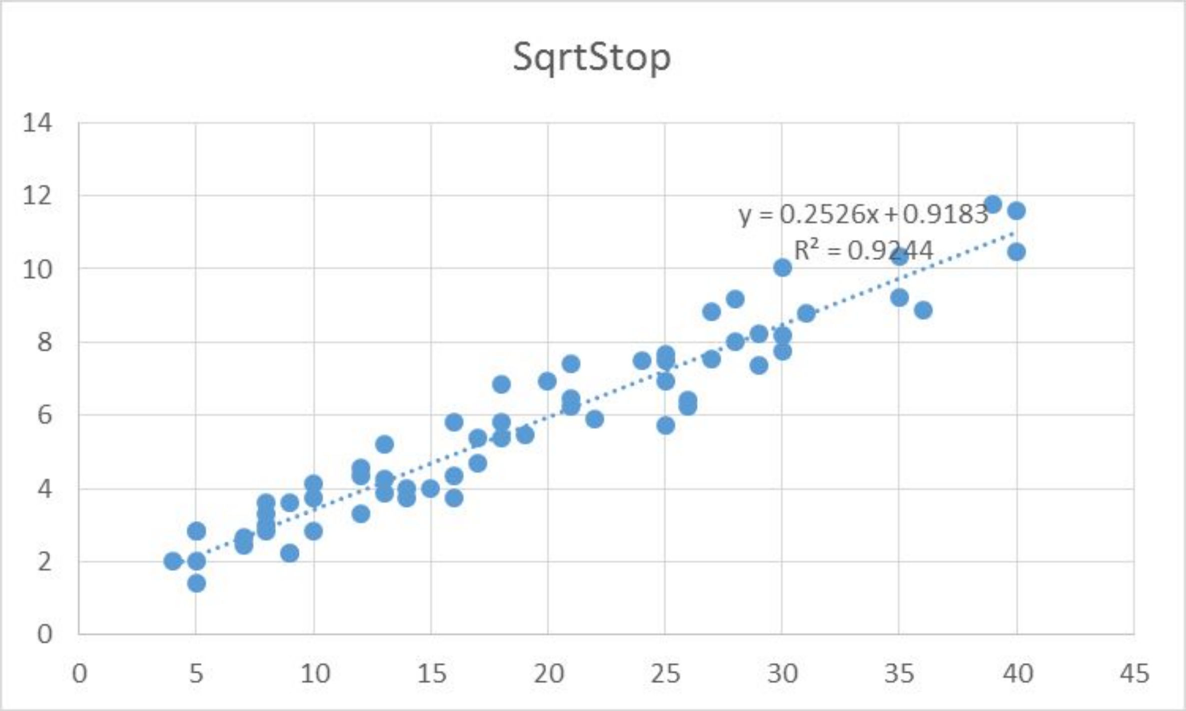


R-Squared, Significance and Residuals - Caution

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 well.



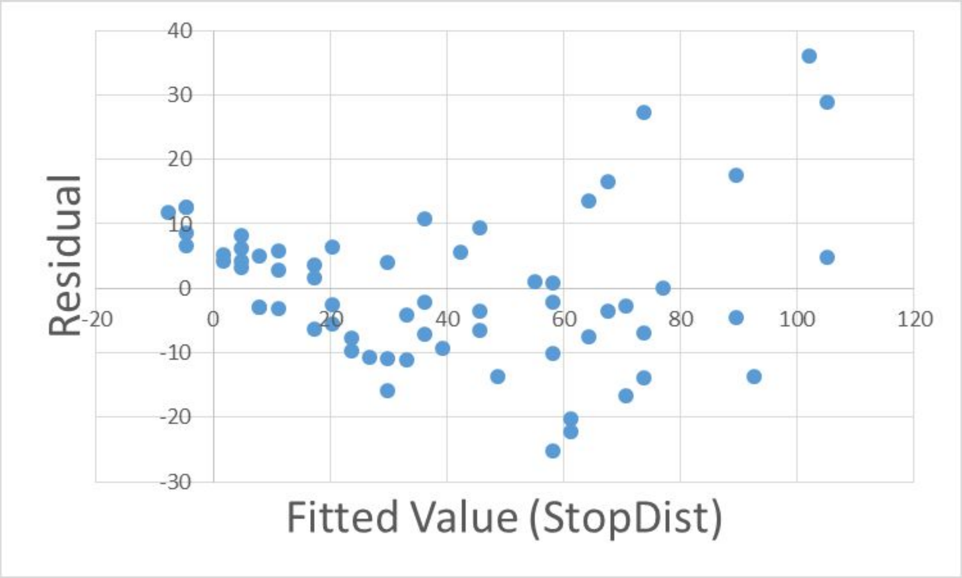


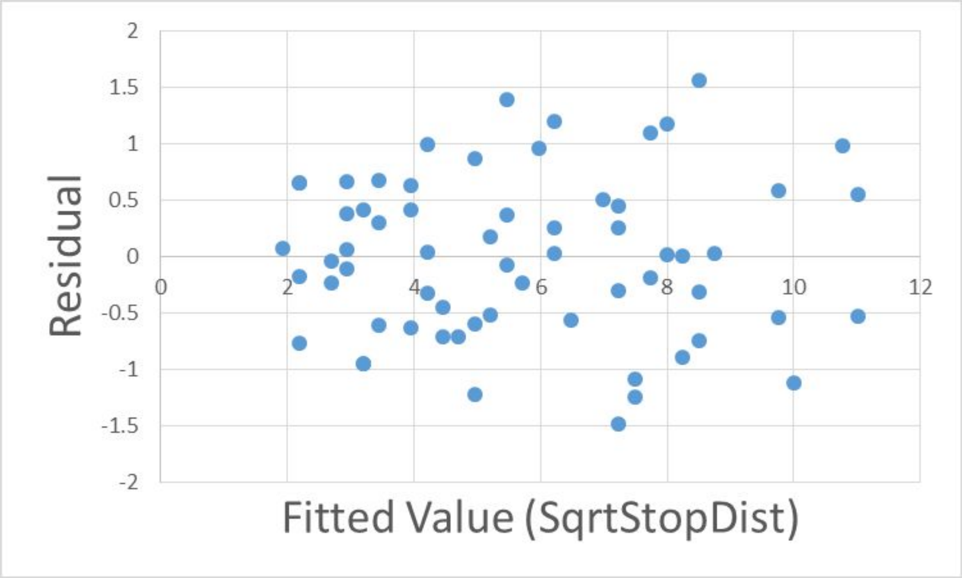
R-Squared, Significance and Residuals - Caution

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.





Assumptions in Linear regression

- Linearity

  The mean of the response , E(*Yi*), at each value of the predictor, *xi*, is a **Linear function** of the *xi*.

- Independence of errors :

  The errors, ε_i, are **Independent**

- Normality of errors :

  The errors, ε_i, at each value of the predictor, *xi*, are **Normally distributed**.

- Homoscedasticity (constant variance)

  The errors, ε_i, at each value of the predictor, x_i, have **Equal variances** (denoted σ^2) i.e. the variance of the error term is constant for all values of x and does not depend on x_i.

An alternative way to describe all four above assumptions :
The errors, ϵ_i, are independent normal random variables with mean zero and constant variance, σ^2

Linear regression model : Probabilistic relationship

In almost all practical examples, we :

- are using a sample to make inferences about the population.
- expect a **probabilistic relationship** rather than a **deterministic relationship**

Probabilistic relationship formalized as :

$$y = \underbrace{E(Y/X=x)}_{} + \underbrace{\varepsilon}_{}$$

Systematic component modelled as

$$\boldsymbol{\beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots}$$

Random error component

Assumed to be normally distributed with 0 mean.

Regression diagnostics : Residual analysis

Residual plots

- For the i^{th} sample the vertical residual is given by :

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - (\beta_0 + \beta_1 x_i) \end{aligned}$$

- Residual plots typically plot residuals or standardized residuals along the y axis.

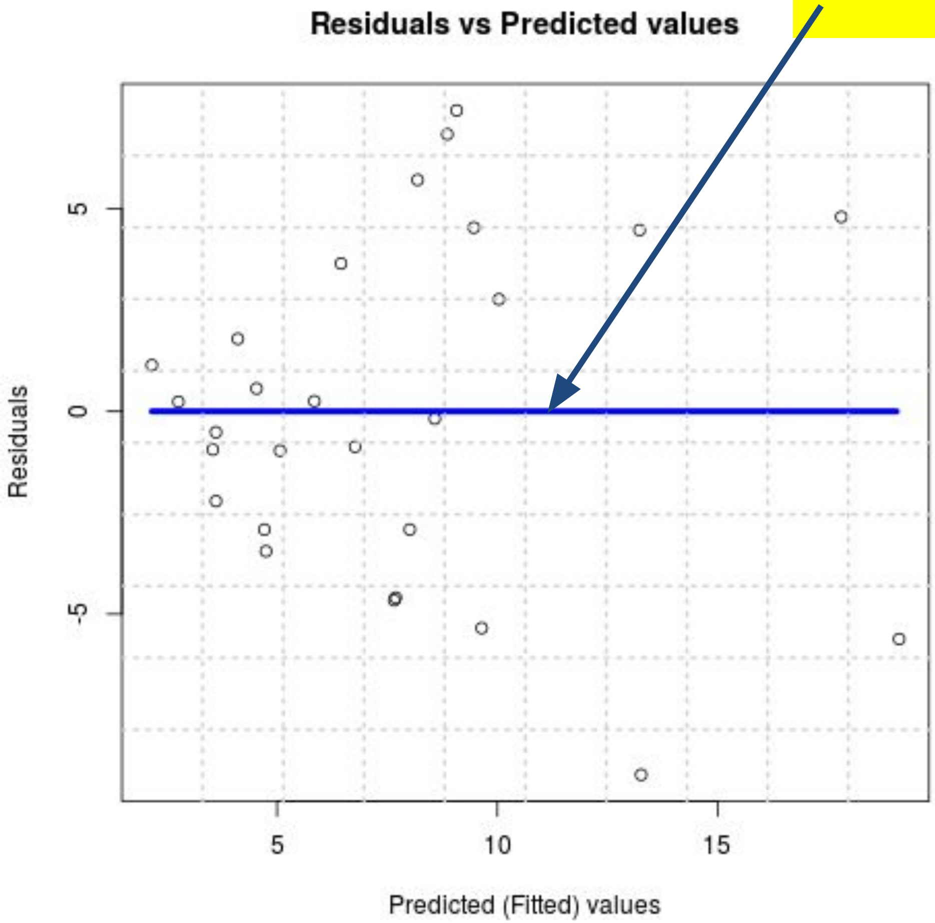
- Problems with linear regression are generally easier to identify via the residual plots rather than the scatter plots of the original data.

A general rule of thumb : Patterns in the residual plots usually indicate that some useful information is not captured by the model.

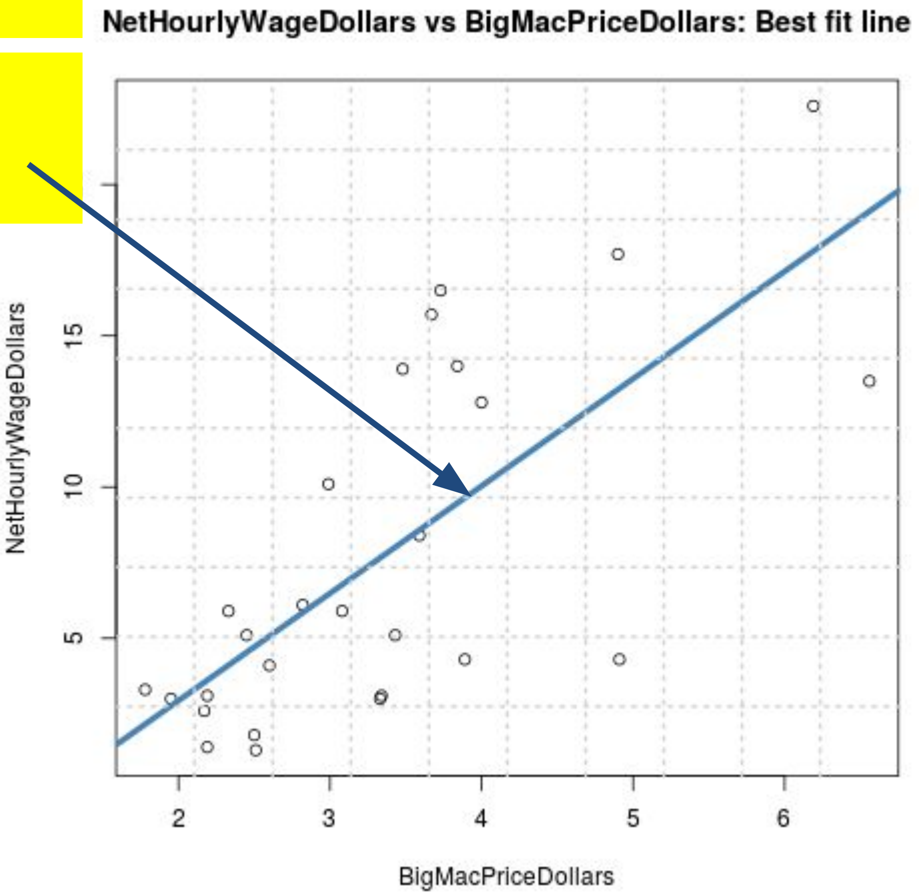
Analysis of residuals : Big Mac example

Zero residual line

Line of best fit



Big Mac example :
Plot of residuals vs fitted values



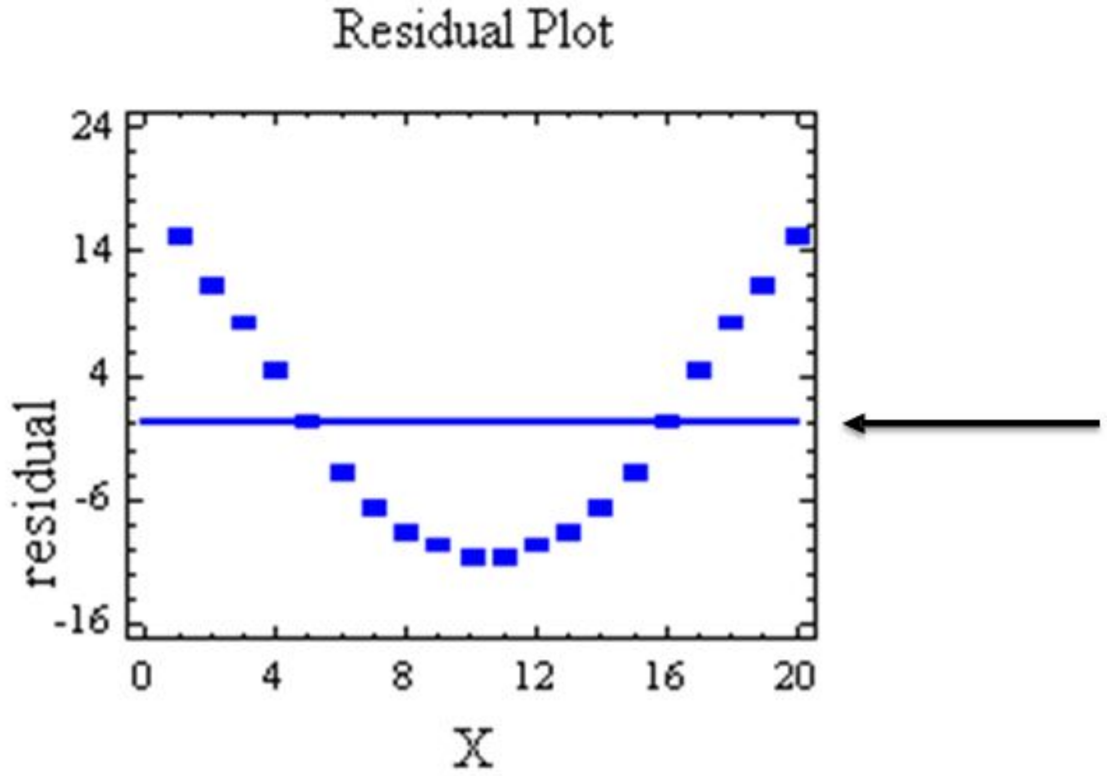
Big Mac example :
Data samples and regression line

Verification of assumptions of linear regression

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

Assumptions of the Regression Model

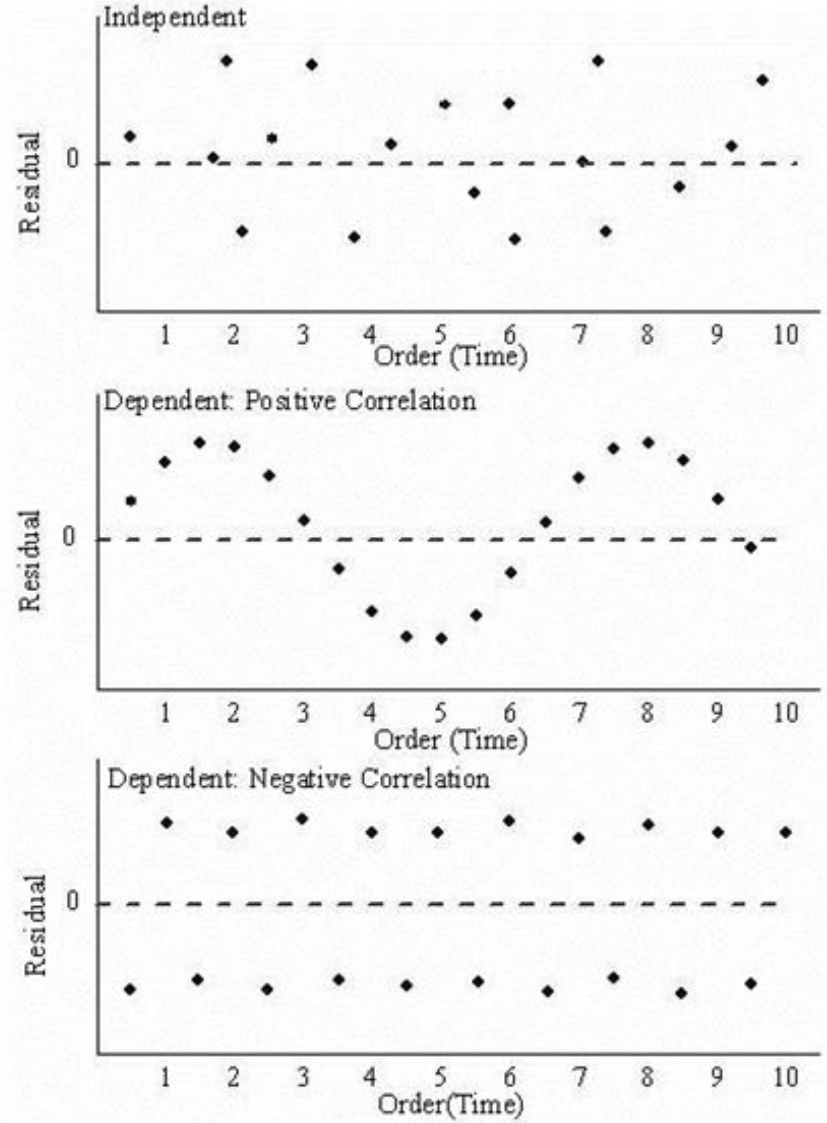
- The model is linear



Zero residual line:
The regression line

Assumptions of the Regression Model

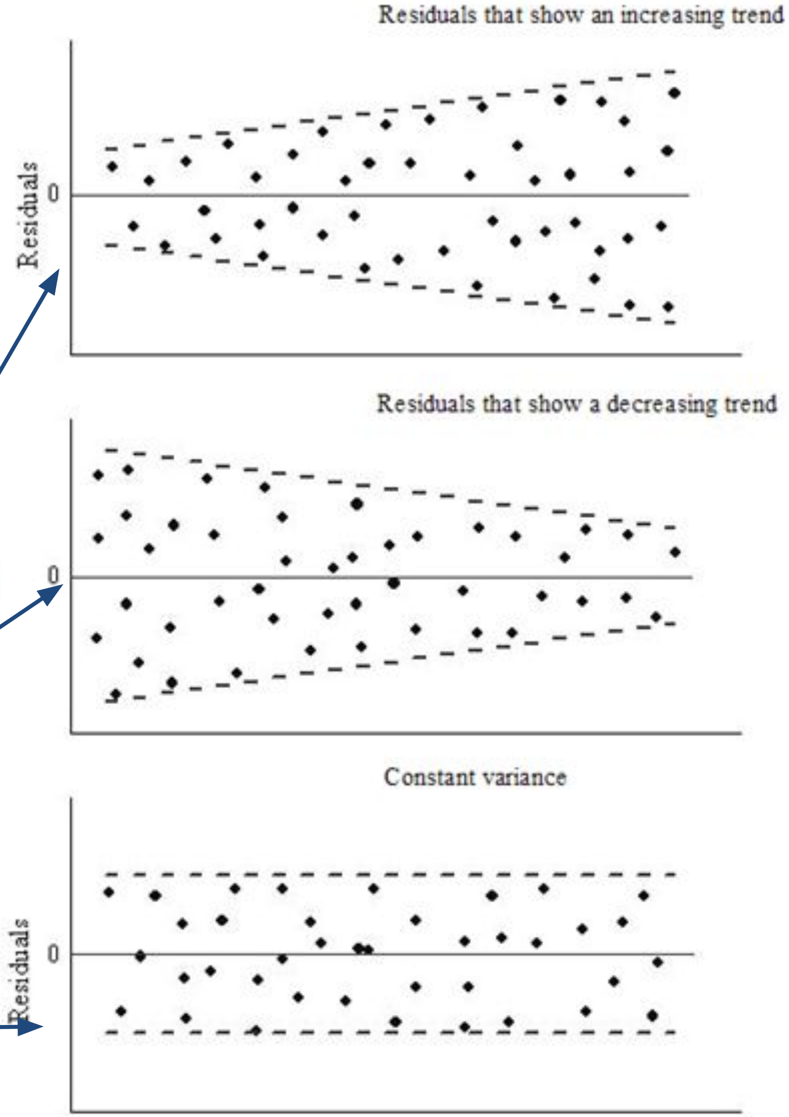
- The error terms are independent
  - Plot against any time (order of observation) or spatial variables preferably. Plots against independent variables may also detect independence.



Assumptions of the Regression Model

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

  - 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.

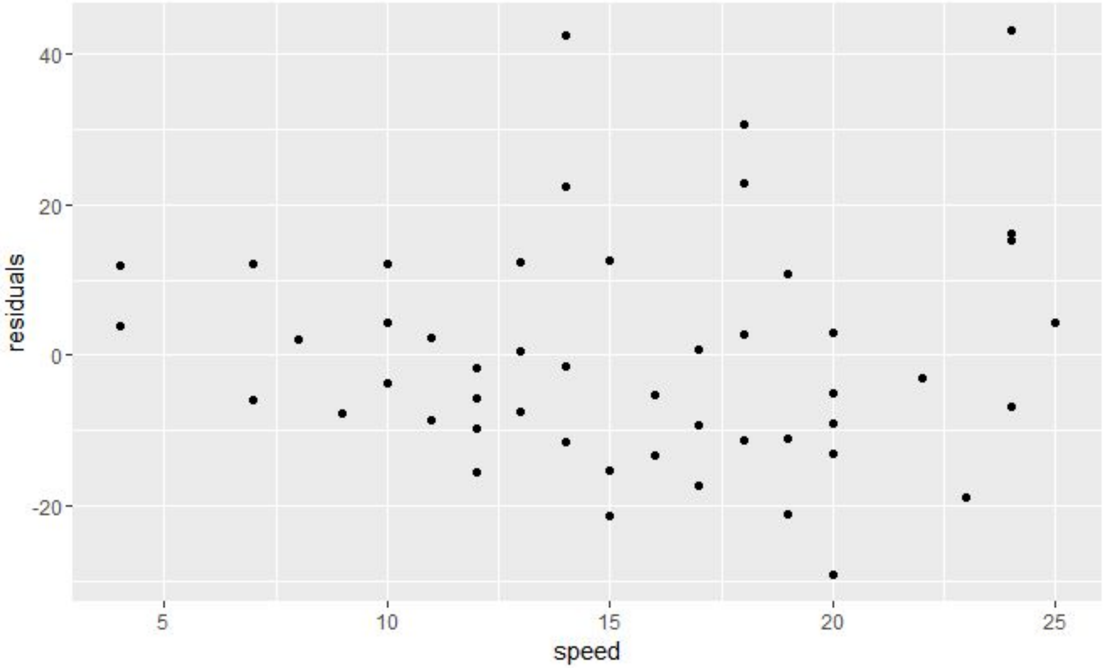


Heteroscedastic

Homoscedastic

Assumptions of the Regression Model

- The residual errors are normally distributed



But, how do we know if something is normally distributed?

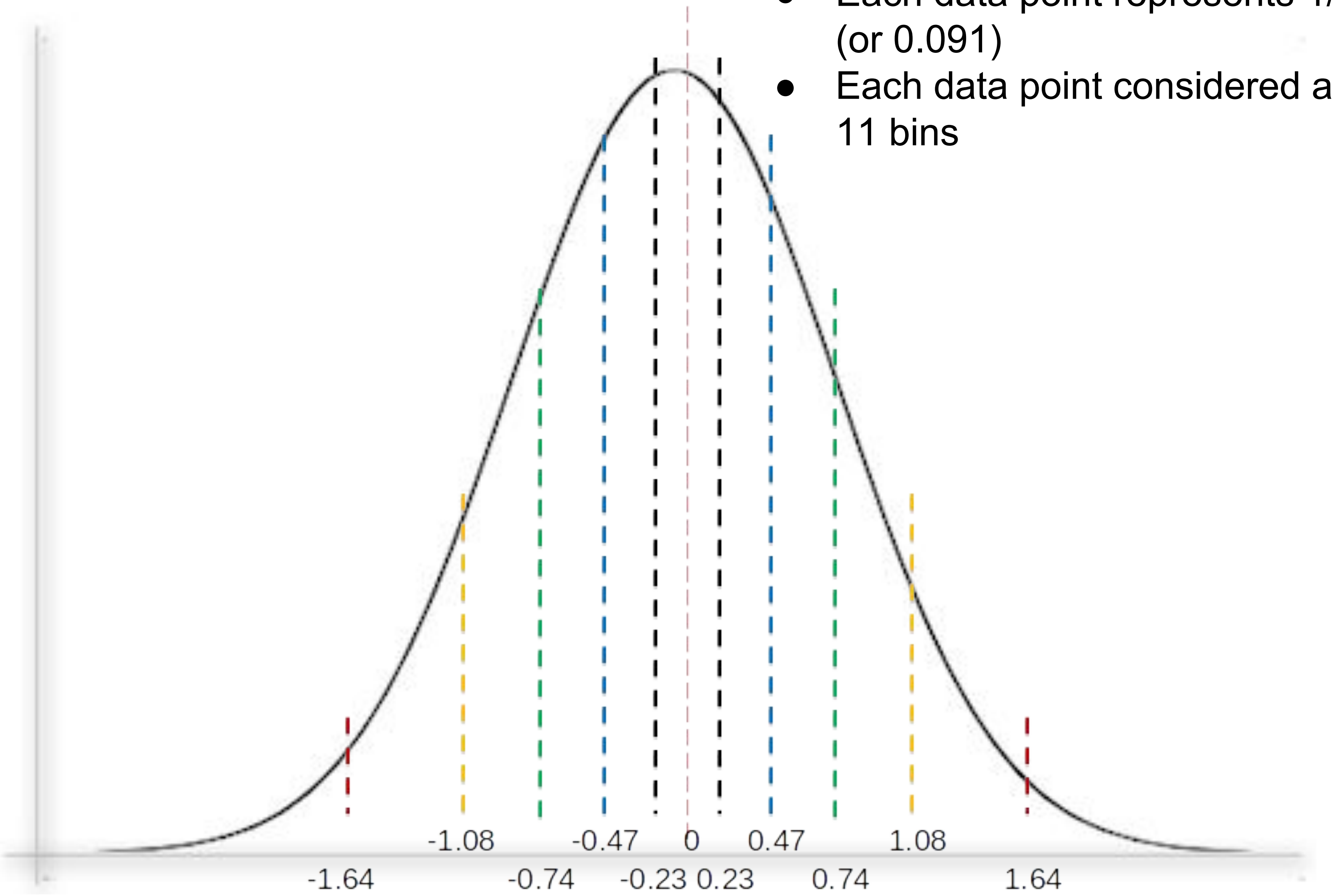
Assumption : Errors are normally distributed

Mechanism to verify above : Q-Q plot

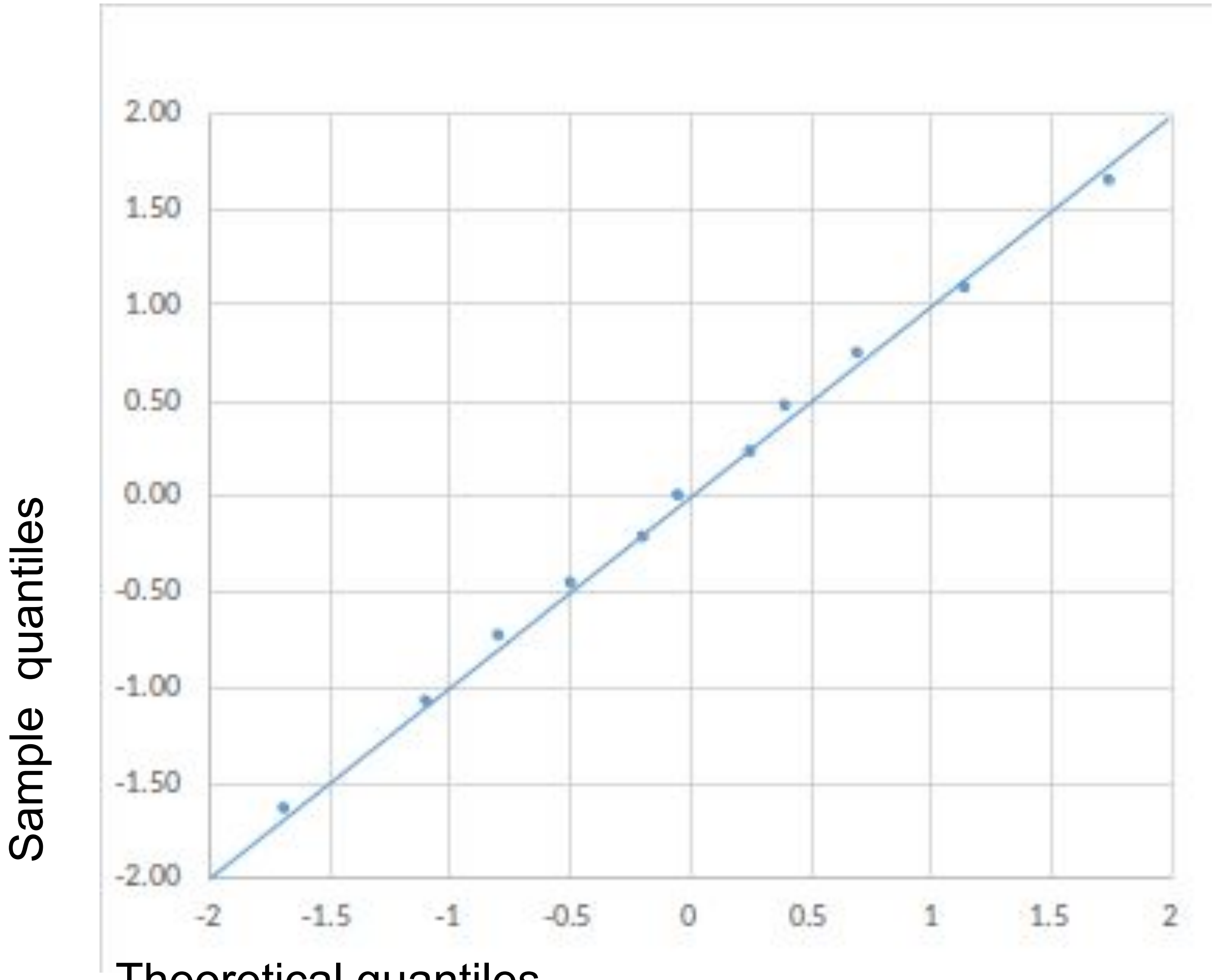
- Quantiles are cutpoints dividing the range of a probability distribution into contiguous intervals with equal probabilities, or dividing the observations in a sample in the same way. https://en.wikipedia.org/wiki/Quantile
- The **quantile-quantile (q-q)** plot is used to validate distributional assumptions 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

- 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



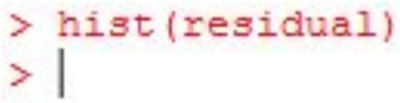
An example of a Q-Q plot



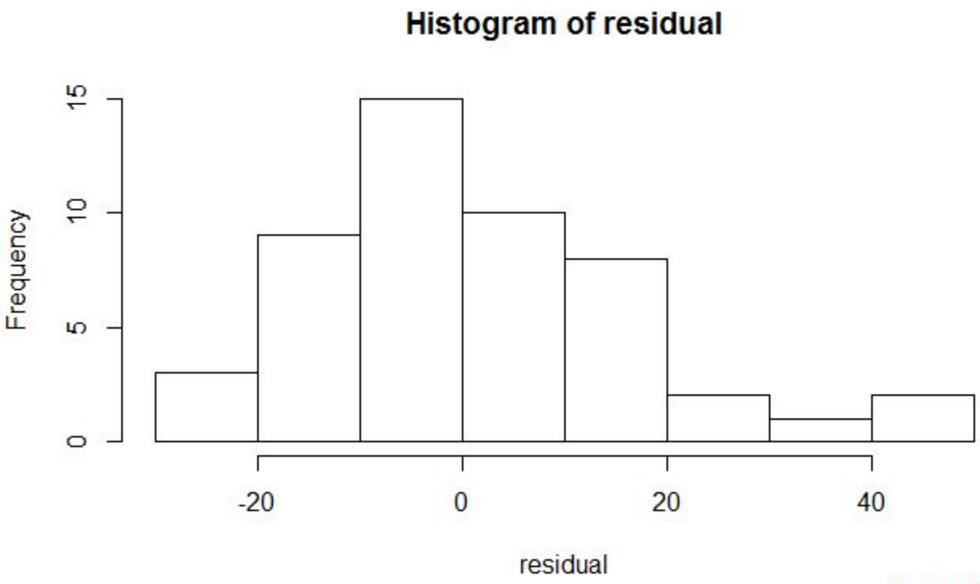
Theoretical quantiles

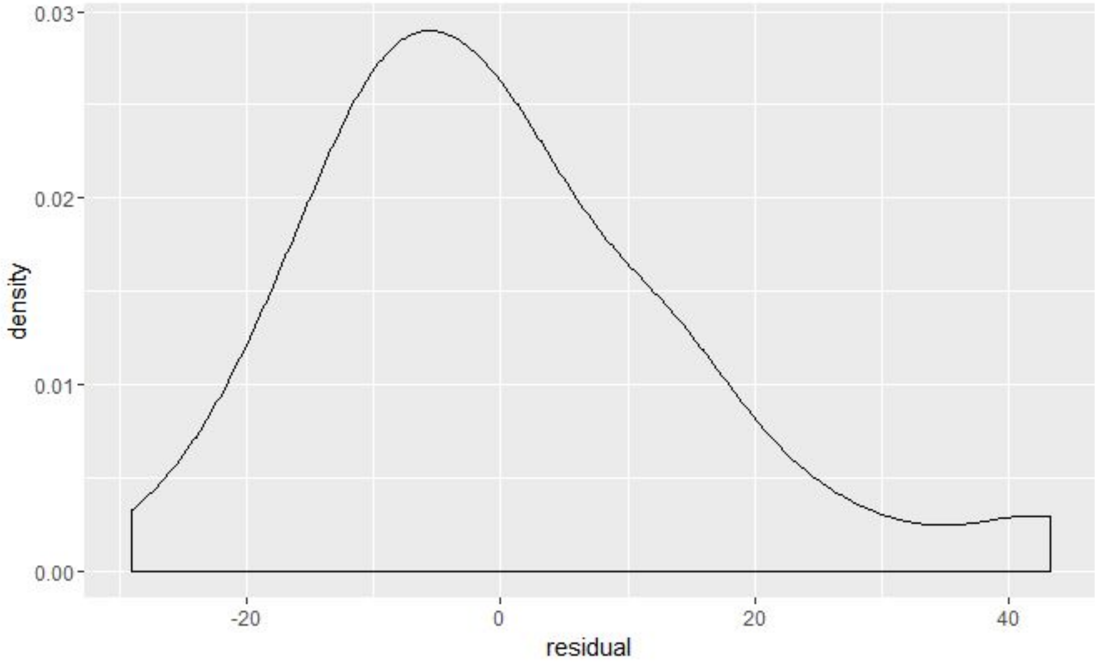
Checking for Normality

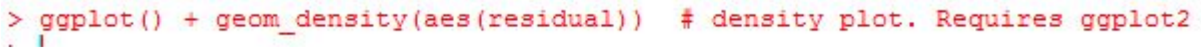
- Start by plotting the data



```
> hist(residual)
> |
```







```
> ggplot() + geom_density(aes(residual))  # density plot. Requires ggplot2
```

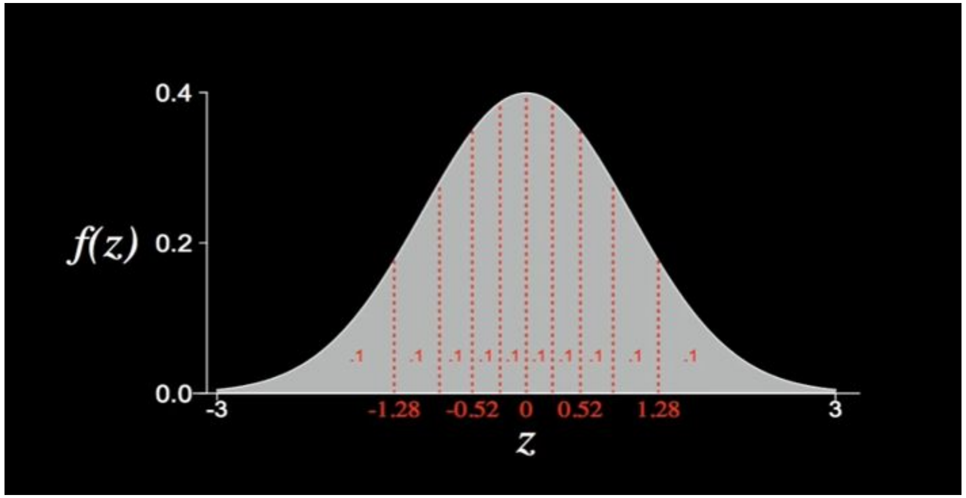
Is there a better way?

Quantile Quantile-Plot

- Its used to assess if the given data-set follows a particular distribution
- For example is the 9-point (sorted) data-set below normal?
  -1.2,-1.11, -1.08, -0.28, -0.25, 0.33, 0.41, 1.37, 1.41

- Lets start with assumption that the data is from normal distribution.
- Lets divide the normal distribution into 9+1 equal areas.
- The boundary point would represent a 0.1 quantile

Quantile-Quantile Plot

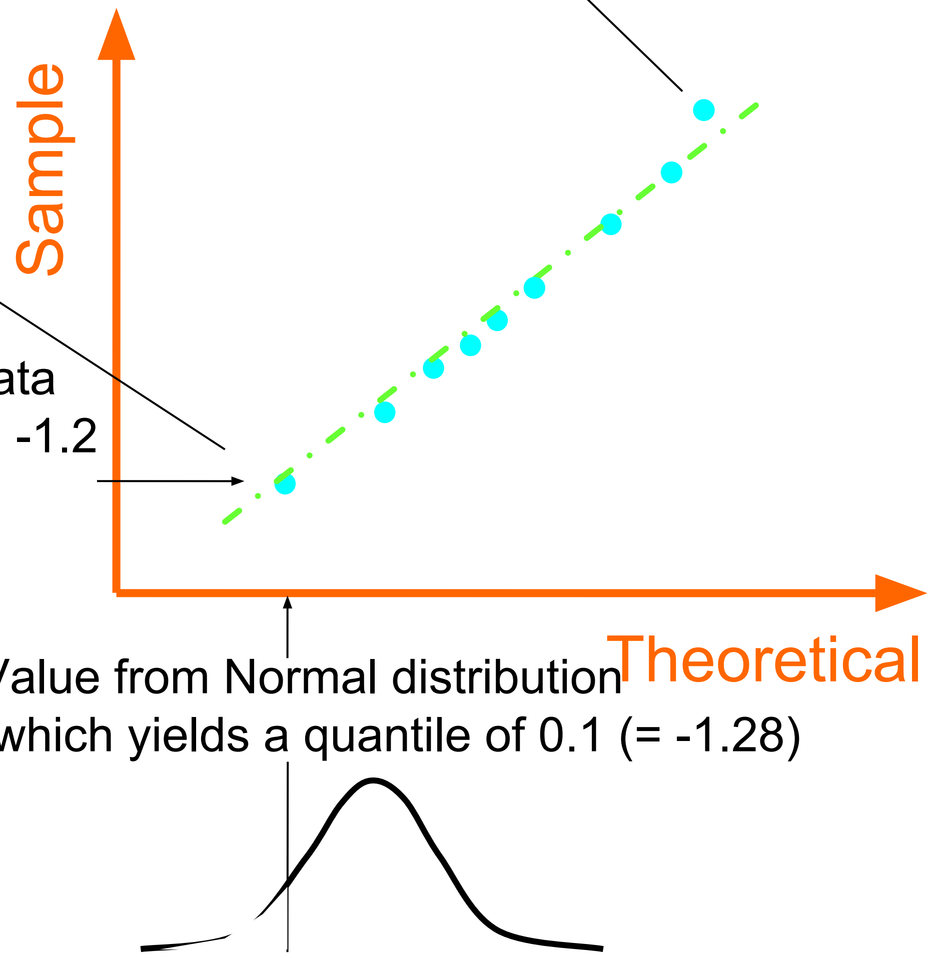


- Then one might expect the smallest of the 9 data points to be from the lowest quantile (0.1)
- Similarly, the largest value would be from the largest quantile (0.9) of the normal distribution

Quantile Quantile Plot

-1.2, -1.11, -1.08, -0.28, -0.25, 0.33, 0.41, 1.37, 1.41

We plot the quantile values for the distribution on the x-axis and the values of the sample on the y-axis

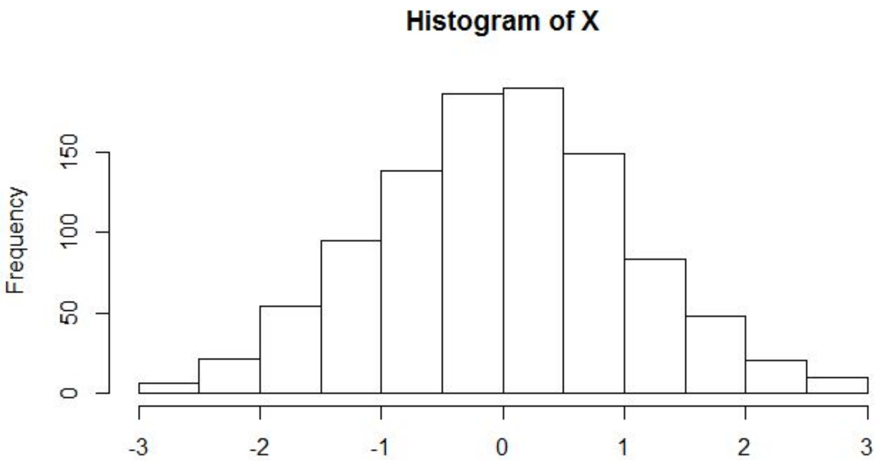


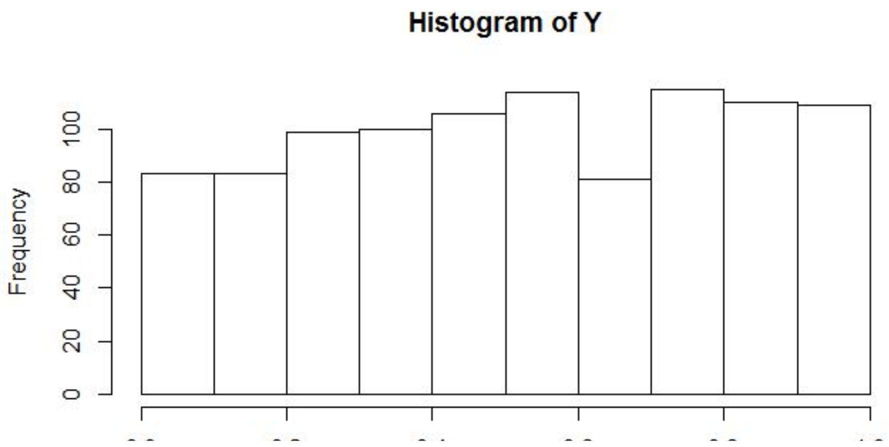
First data point = -1.2

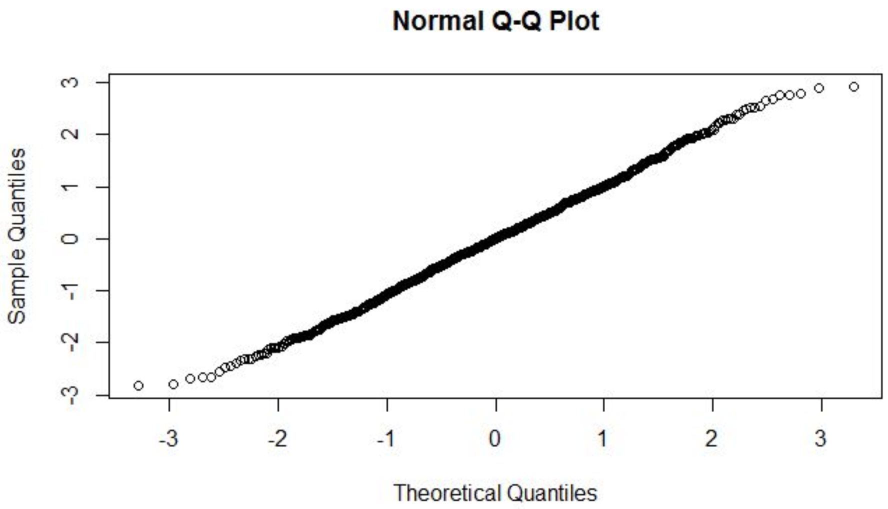
Value from Normal distribution which yields a quantile of 0.1 (= -1.28)

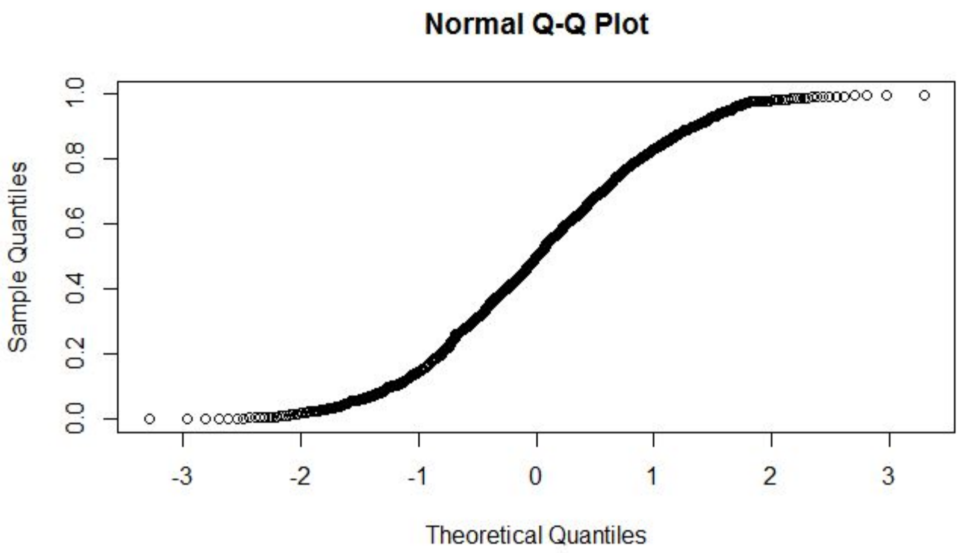
If the points lie on close to a straight line, then the sample is normal

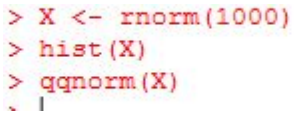
QQ Plot for Normal vs Uniform Distribution



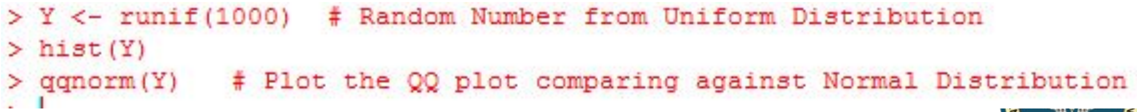








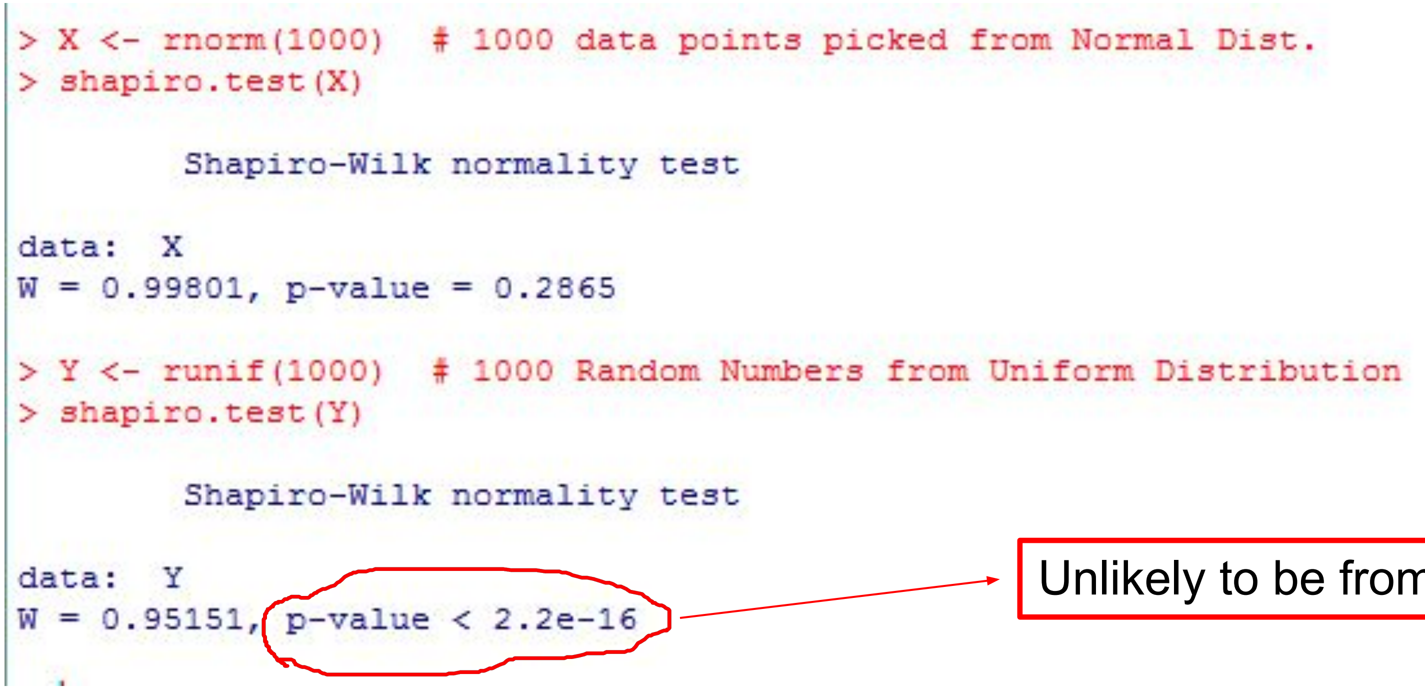
```
> X <- rnorm(1000)
> hist(X)
> qqnorm(X)
```



```
> Y <- runif(1000)   # Random Number from Uniform Distribution
> hist(Y)
> qqnorm(Y)    # Plot the QQ plot comparing against Normal Distribution
```

Checking for Normal Distribution

- Other objective methods of checking for normality also exist
- Shapiro-Wilk Test gives a probability value (p-value) that the given data sample is actually from a Normal distribution
- If p-value is less than 0.05, then its unlikely to be from Normal distribution



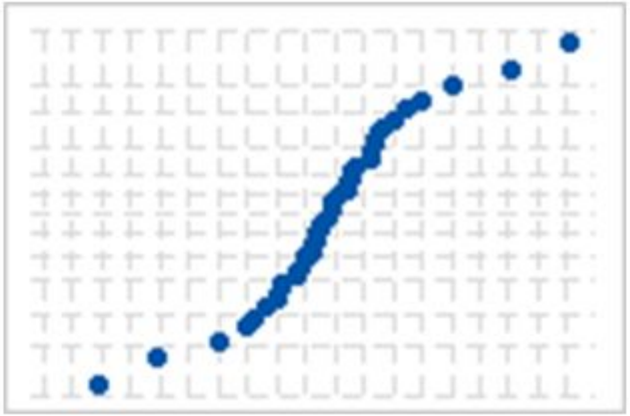
```
> X <- rnorm(1000)  # 1000 data points picked from Normal Dist.
> shapiro.test(X)

        Shapiro-Wilk normality test

data:  X
W = 0.99801, p-value = 0.2865

> Y <- runif(1000)  # 1000 Random Numbers from Uniform Distribution
> shapiro.test(Y)

        Shapiro-Wilk normality test

data:  Y
W = 0.95151, p-value < 2.2e-16
```

Unlikely to be from Normal Dist

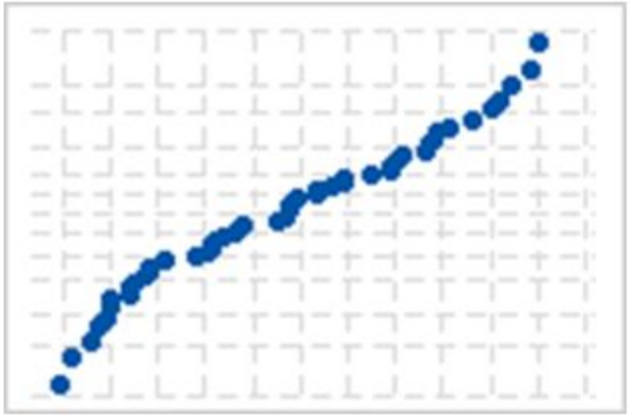
Interpreting Residuals

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

Interpreting Residuals – Non-normality

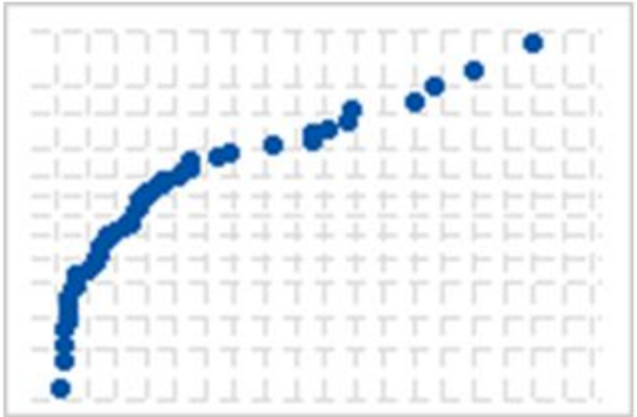


S-curve implies a distribution with long tails

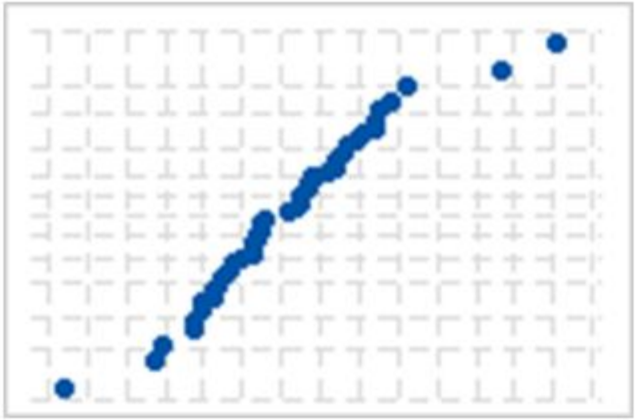


Inverted S-curve implies a distribution with short tails

Interpreting Residuals – Non-normality



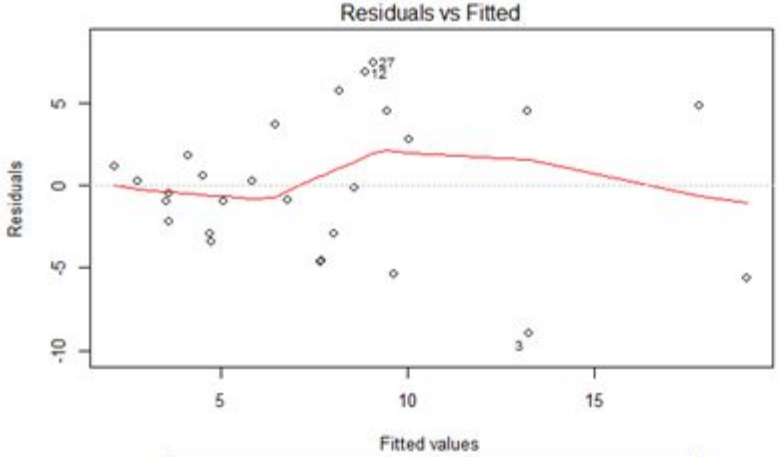
Downward curve implies an asymmetric distribution



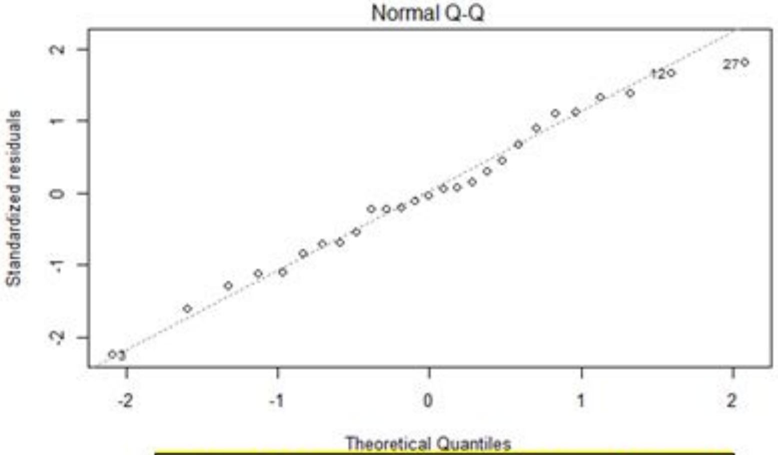
A few points lying away from the line implies a distribution with outliers

Residuals – Big Mac

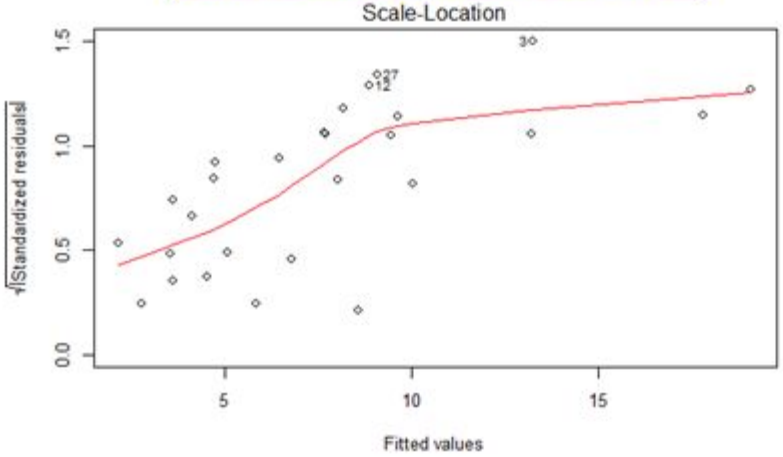
Is a wrong model fitted (linear or quadratic, etc.)?


Residuals vs Fitted

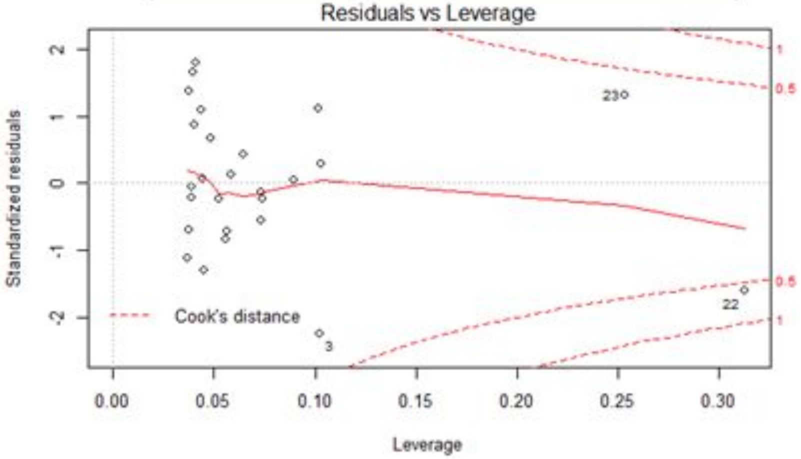
Are the residuals normally distributed?


Normal Q-Q

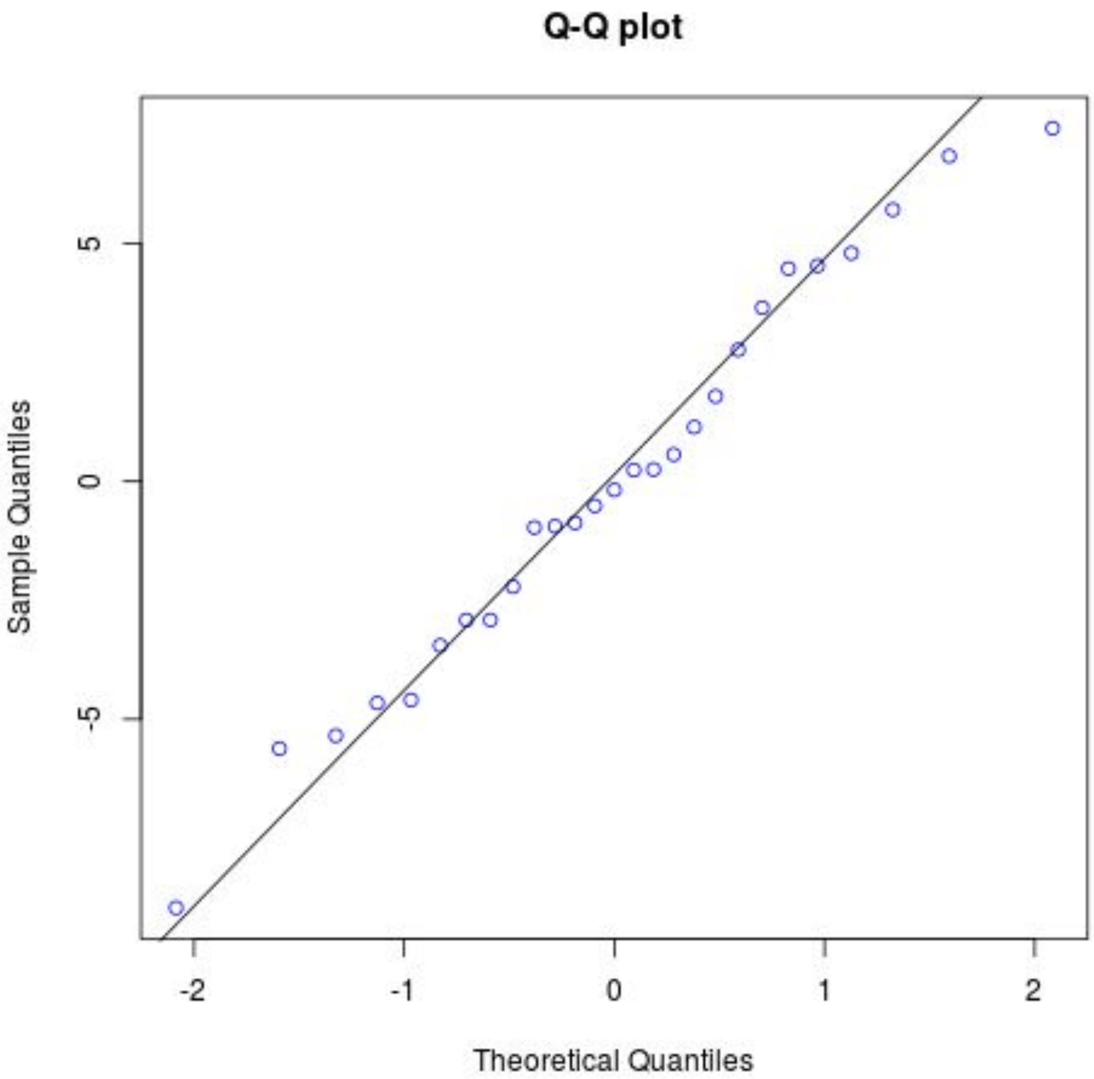
Is the data homoscedastic?


Scale-Location

Are there influential outliers?


Residuals vs Leverage

Q-Q plot for the Big Mac example dataset



Does the Q-Q plot show the residuals to be approximately normal?

Influential observations

Influential observations

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

Influence generally measured by

- **Leverage :**
  - Calculated only from the independent variables.

- **Distance** (or ‘residuality’ or ‘outlierness’)
  - Calculated from the y values (through residuals).

Influence is a function of leverage and distance (“**residuality or outlierness**”)

Influential observations : Leverage

- How much the observation’s value on the predictor variable differs from the mean of the predictor variable.

- It tells us about extreme x values, which have the potential to highly influence the regression in certain conditions.

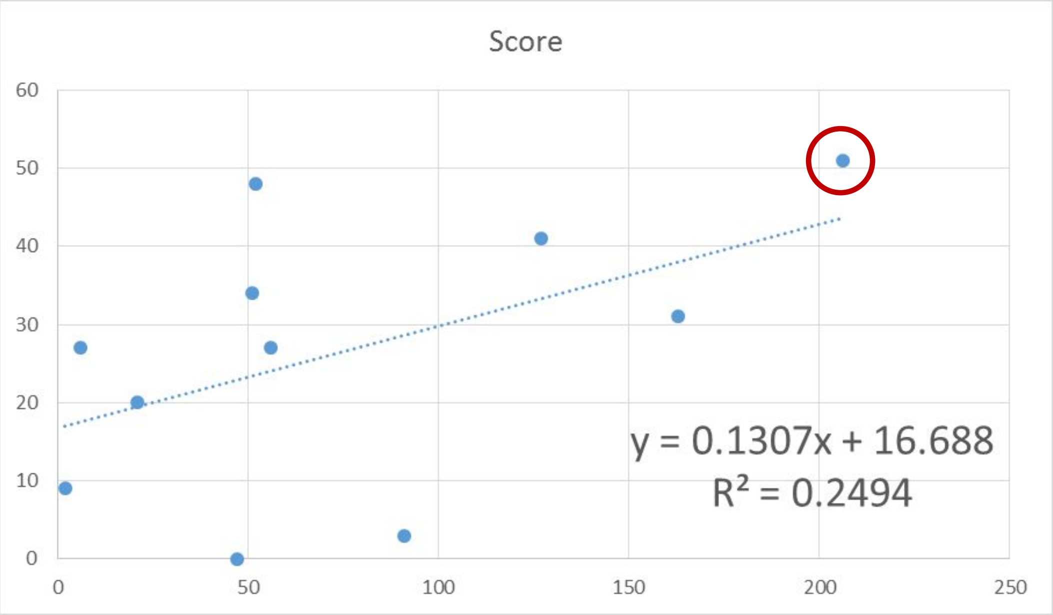
- **A distinction is often made between outliers and influential data points.**

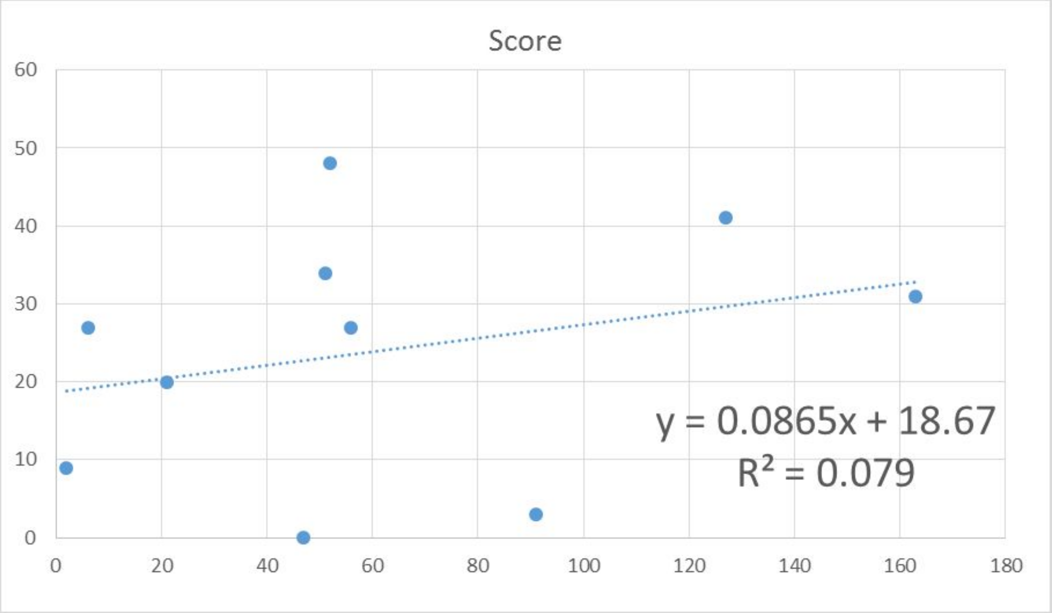
- **High leverage points may or may not be outliers.**

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.

Slide credit : Dr. Sridhar Pappu, Data available at https://onlinecourses.science.psu.edu/stat501/node/258/

Influential observations : Leverage

The leverage for point i in the data sample is given by :

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

where the standardized residual z_i is given by

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

where

- x_i = x value corresponding to i^{th} observation.
- $\bar{x}$ = Mean of the x-values
- σ = standard deviation of the x values

Cook’s distance

- Cook’s Distance measures overall influence of an observation by seeing the impact on the regression coefficients when this observation is omitted.

- It is a measure of the influence of a data point that **accounts both for leverage and residual**.

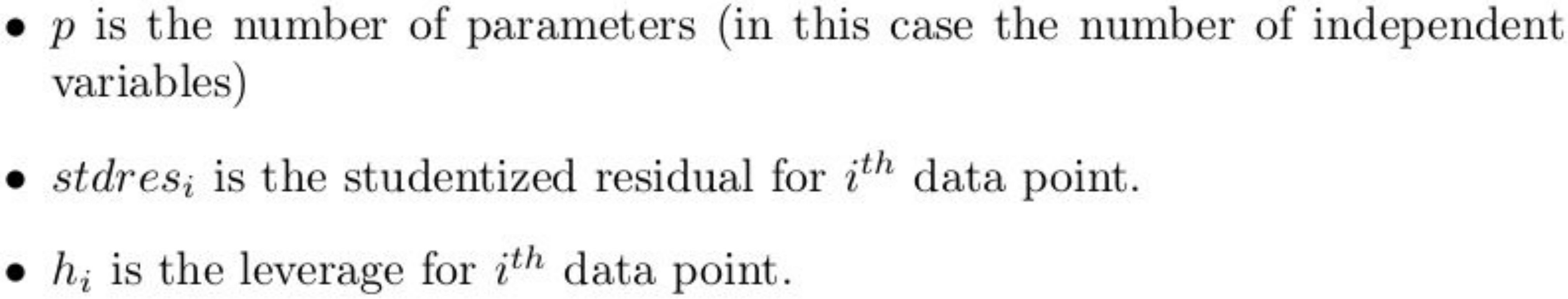
- It is proportional to the sum of the squared differences between predictions made with all observations in the analysis and predictions made leaving out the observation in question.

Dealing with influential observations using Cooks distance

The **Cooks distance** for point i in the data sample is given by :

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

where



- p is the number of parameters (in this case the number of independent variables)
- $stdres_i$ is the studentized residual for i^{th} data point.
- h_i is the leverage for i^{th} data point.

Dealing with influential observations using Cooks distance

Rules of thumb

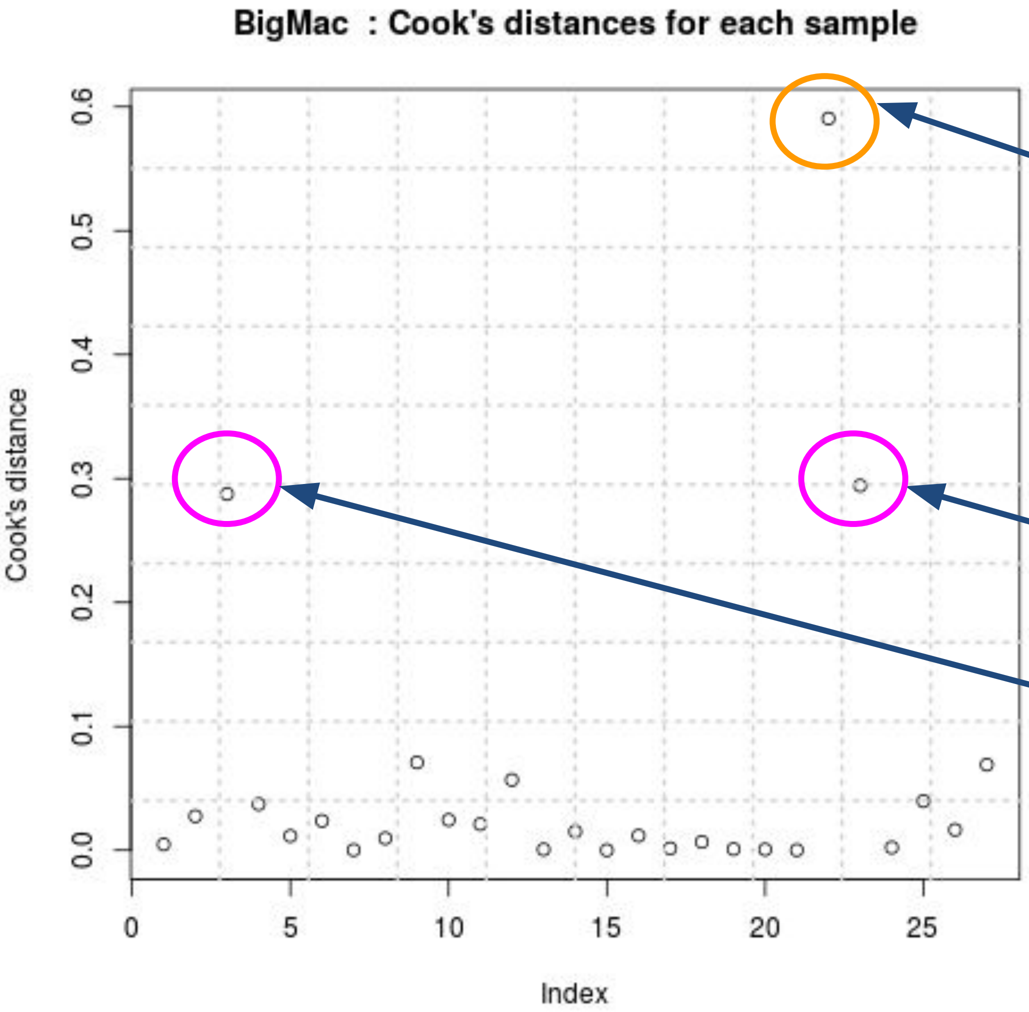
- An observation i can be considered as having too much influence if its Cooks distance (D_i) > 1.
  - Investigate observations with Cooks distances > 0.5 also.

- Relative size interpretation :

In general, investigate any observation whose Cooks distance is significantly different from the rest.

Identifying influential points using Cook's distances

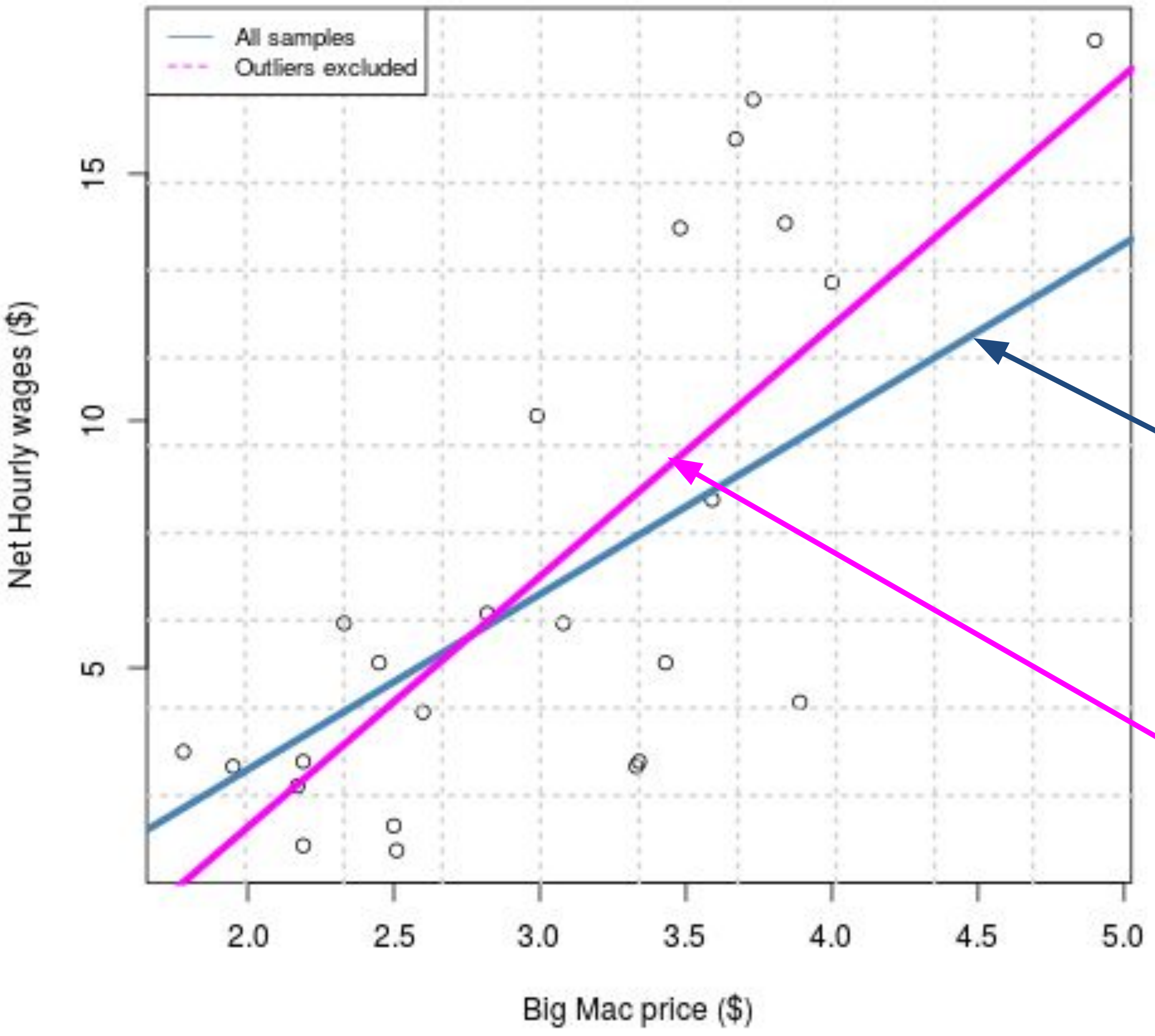
Big Mac example : Cook's distances for each sample



Cook's distance > 0.5

Cook's distances "significantly different" from those of other samples

Data samples (outliers excluded) + Best fit line



All data samples included: Equation of the best fit line

NetHourlyWage = BigMacPrice(3.5474)-4.1540

R^2=0.5142, adjusted R^2 = 0.4947

Outliers excluded: Best fit line

NetHourlyWage = BigMacPrice(5.0745)-8.3760

R^2=0.5714, adjusted R^2=0.552

Fixing Non-normality and Heteroscedasticity

Transformation of data can help correct normality and unequal variances problems

Data Transformations

- **Main aim of applying transformations in linear regression** To ensure that after transformation, the assumptions of linear regression are violated to a much lesser extent.

- Commonly used transformations :

log, power, square root etc.

- Transformations may be applied to the predictor variables (Xs) or to response variable (y) or both.

Data Transformations commonly used : The log transformation

Problem diagnosed	Recommended transform
Non-linearity is the only problem — the independence, normality and equal variance conditions are met.	Log transform the x (predictor)
Non-normality and/or unequal variances	Log transform the x (predictor)
When the regression function is **not linear** and the error terms are **not normal** and have **unequal variances**.	Log transform both x and y

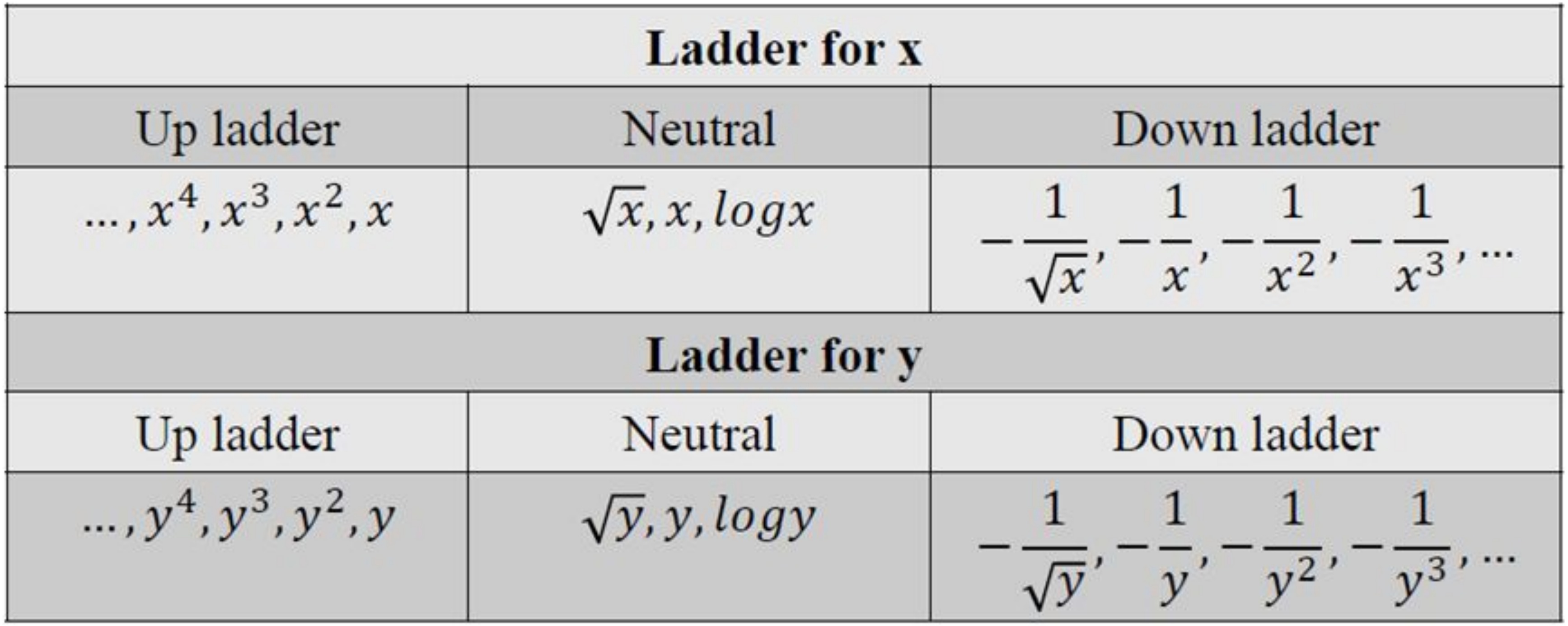
Source : https://onlinecourses.science.psu.edu/stat501/

Other suggested data transformations

Problem diagnosed	**Recommended transform**
Primary problem with the model is **non-linearity.**	**Single predictor :** Look at a scatter plot of the data **Multiple predictors :** Look at residual plots to suggest transformations that might help.
If the variances are unequal and/or error terms are not normal,	**Power transformation** on *y*. i.e. $y^{*}=y^{\lambda}$
If the response *y* is a **Poisson count**	**Square root** transformation on y

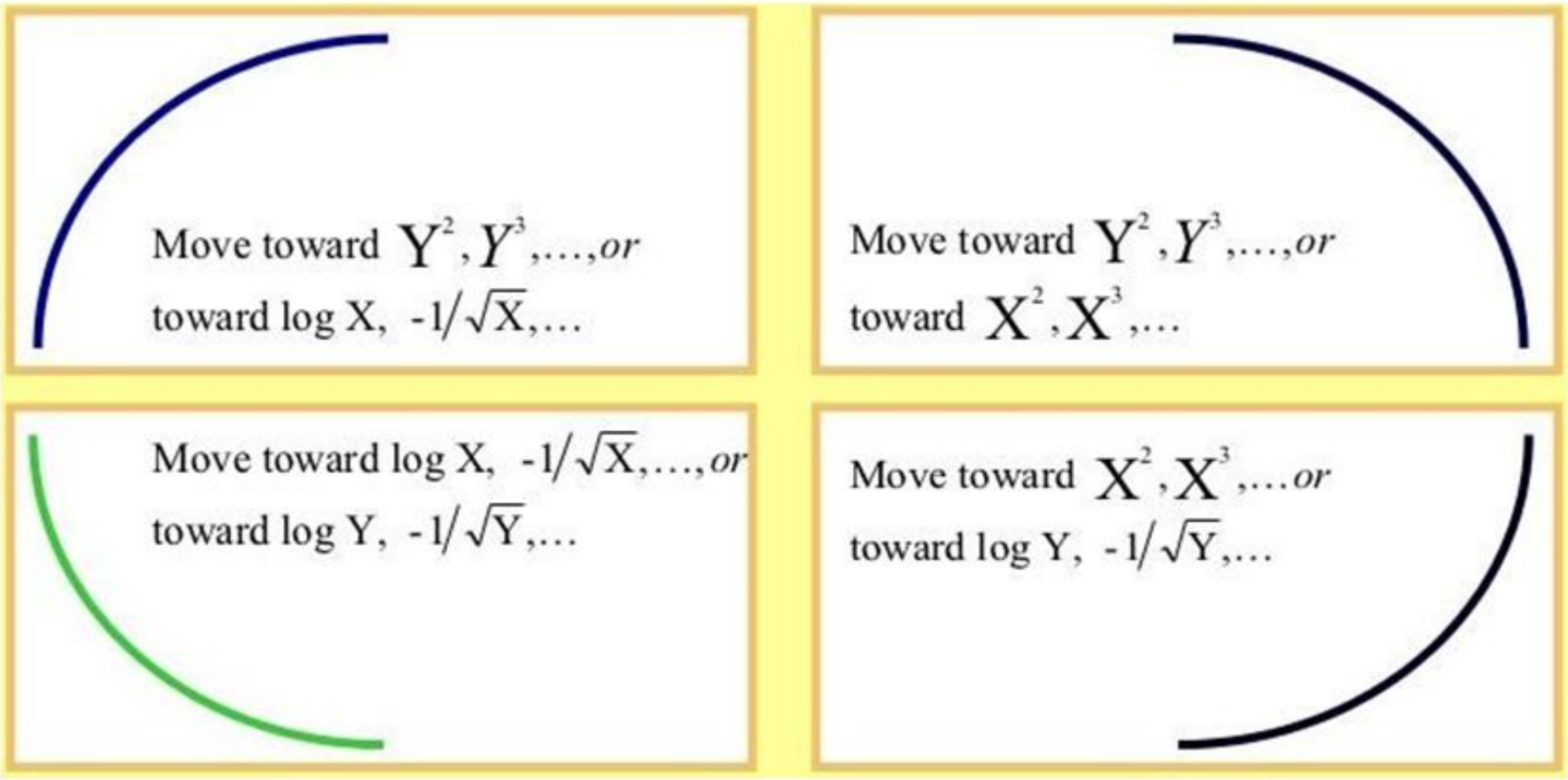
Source : https://onlinecourses.science.psu.edu/stat501/

Tukey's Ladder of Transformations



Ladder for x		
Up ladder	Neutral	Down ladder
$\ldots, x^4, x^3, x^2, x$	$\sqrt{x}, x, logx$	$-\frac{1}{\sqrt{x}}, -\frac{1}{x}, -\frac{1}{x^2}, -\frac{1}{x^3}, \ldots$
Ladder for y		
Up ladder	Neutral	Down ladder
$\ldots, y^4, y^3, y^2, y$	$\sqrt{y}, y, logy$	$-\frac{1}{\sqrt{y}}, -\frac{1}{y}, -\frac{1}{y^2}, -\frac{1}{y^3}, \ldots$

Tukey's Four-Quadrant Approach



Move toward $Y^2, Y^3, \ldots, or$
toward log X, $-1/\sqrt{X}, \ldots$

Move toward $Y^2, Y^3, \ldots, or$
toward $X^2, X^3, \ldots$

Move toward log X, $-1/\sqrt{X}, \ldots, or$
toward log Y, $-1/\sqrt{Y}, \ldots$

Move toward $X^2, X^3, \ldots or$
toward log Y, $-1/\sqrt{Y}, \ldots$

Hands on exercise

- Verify whether linear regression assumptions are satisfied
  - Analysis of residuals.
  - For detailed reference on interpreting residuals refer http://www.stat.berkeley.edu/~stark/SticiGui/Text/regressionDiagnostics.htm

- Identify influential points. Remove influential points and rebuild the model if necessary.
  - Use Cook’s distance to identify influential points

- Split data into train,validation and test buckets.
  - Report final performance metrics on test set only.

Outline of major steps in building a linear regression model

Commonly employed preprocessing steps in linear regression model building

Data exploration and understanding

- Scatter plots and other visualizations
- Descriptive statistics, correlations etc.
- Missing value imputation

Outlier identification and removal

Possible approaches

- Identify and reject outliers upfront

  OR
- Build a model, identify and discard influential points, rebuild model

Transformations

- **Aim in linear regression**

To ensure that assumptions are violated to a significantly lesser extent after applying transformations.

- Common types of transformations include

square root, log etc.

Outlier identification

Dealing with Outliers

- **Outlier :** An anomalous data sample (possible measure : data point that is distant from/dissimilar to other similar points)
- An outlier could be due to :
  - An anomalous instance
  - Variability in the measurement
  - Experimental errors.
- Rules of thumb available, but what data sample is an outlier is best left to the judgement of the one investigating the data.

Outlier detection approaches

Approach 1 :
Identify and reject outliers upfront, build model without these outliers.

- **Univariate** case : **boxplot**
  Commonly used criterion : a data sample with value outside of 1.5 times inter-quartile range is considered an outlier.

- **Bivariate** case :
  - ○ Boxplots of combinations of variables (useful especially for a numerical and categorical variable combination)
  - ○ **Scatterplot** with confidence ellipse.
    Criterion : outside of a (say, 95%) confidence ellipse is considered an outlier.

Possible problem with above strategies:
The features based on which the outlier data samples were identified, themselves may not be significant (i.e. excluding these attributes yields a better model)

A partial remedy : A multivariate strategy i.e. use all predictors to build an initial model and then reject one or more predictors.

Outlier detection approaches

Approach 2 : Build a model using all data, identify outliers using the model and a statistical criterion, rebuild model without outliers

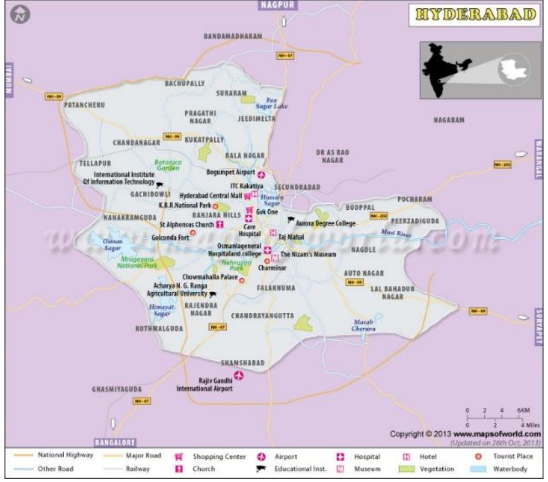
Statistical tests that can be used as a basis for exclusion

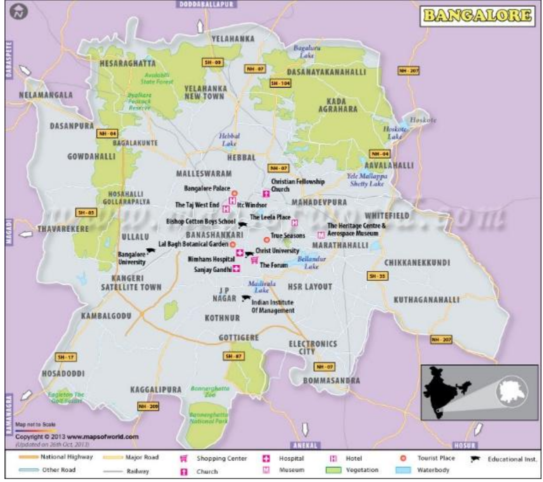
- Standardised residuals
- **Leverage statistics**
- **Cook's distance**, which can be viewed as a combination of the two above.

Approach 3 : Use a robust estimation procedure that is less influenced by outliers

A few examples of robust estimation techniques:

- Weighted least squares
- RANSAC
- Regularization based regression methods. Eg. LASSO, ridge regression





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:	
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	
SlideShare:	
LinkedIn:	

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.

Additional links and resources

- Home page of Regression methods online lecture notes from the course offered by the Eberly College of Science, Penn State
  https://onlinecourses.science.psu.edu/stat501/

For detailed description on specific topics you may refer the following specific pages :

- Hypothesis test related to slopes, in the context of Linear Regression (Lesson 2: SLR Model Evaluation)
  https://onlinecourses.science.psu.edu/stat501/node/260/

- Influential points : Leverage and influence, Cook's distance (Lesson 11 : Influential points)
  https://onlinecourses.science.psu.edu/stat501/node/336/

- Data transformations(Lesson 9 : Data transformations)
  https://onlinecourses.science.psu.edu/stat501/node/318/