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
  - Major steps in building a Linear regression model

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

Forecasting quarterly sales of a product

Predicting length of patient stay in a hospital

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

Predicting whether a loan applicant is likely to default

Predicting stock prices

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

An example : Forecast Accuracy

In the real world, it is impossible to have 100% forecast accuracy.

- But is it possible to improve it by 5%?
- How much financial impact would that have?
- How do you measure it?
- How much effort would that take?

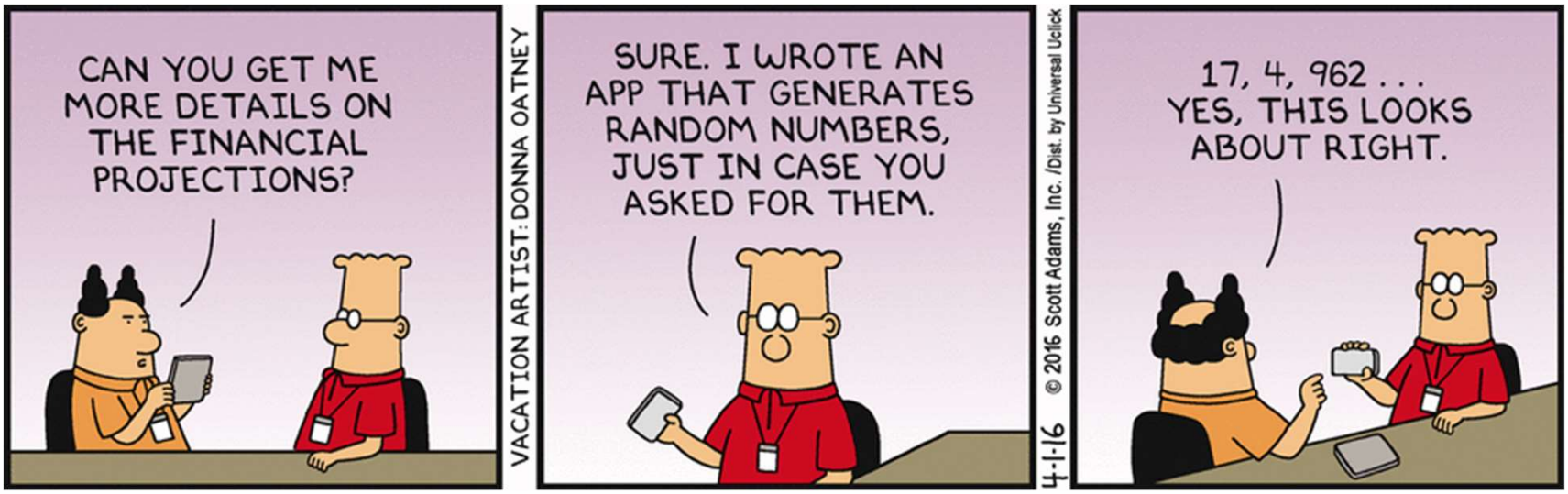
Forecasters should provide estimates of forecast errors along with their forecasts

Source: https://www.supplychain247.com/article/navigating_a_course_with_planning_forecasting/bristlecone

Importance of accurate Demand Forecasting

- A study of 67 companies conducted by the **Supply Chain Council** found a definite correlation between **higher forecast accuracy** and **higher fill rates**.

- When analysts studied companies that were best in class in demand forecasting, they found that every **3%** increase in forecast accuracy increased profit margin by **2%**.*



* *Source:* AMR Research *(http://www.thrivetech.com/forecast-accuracy-metrics/)*

Common classes of practical learning problems

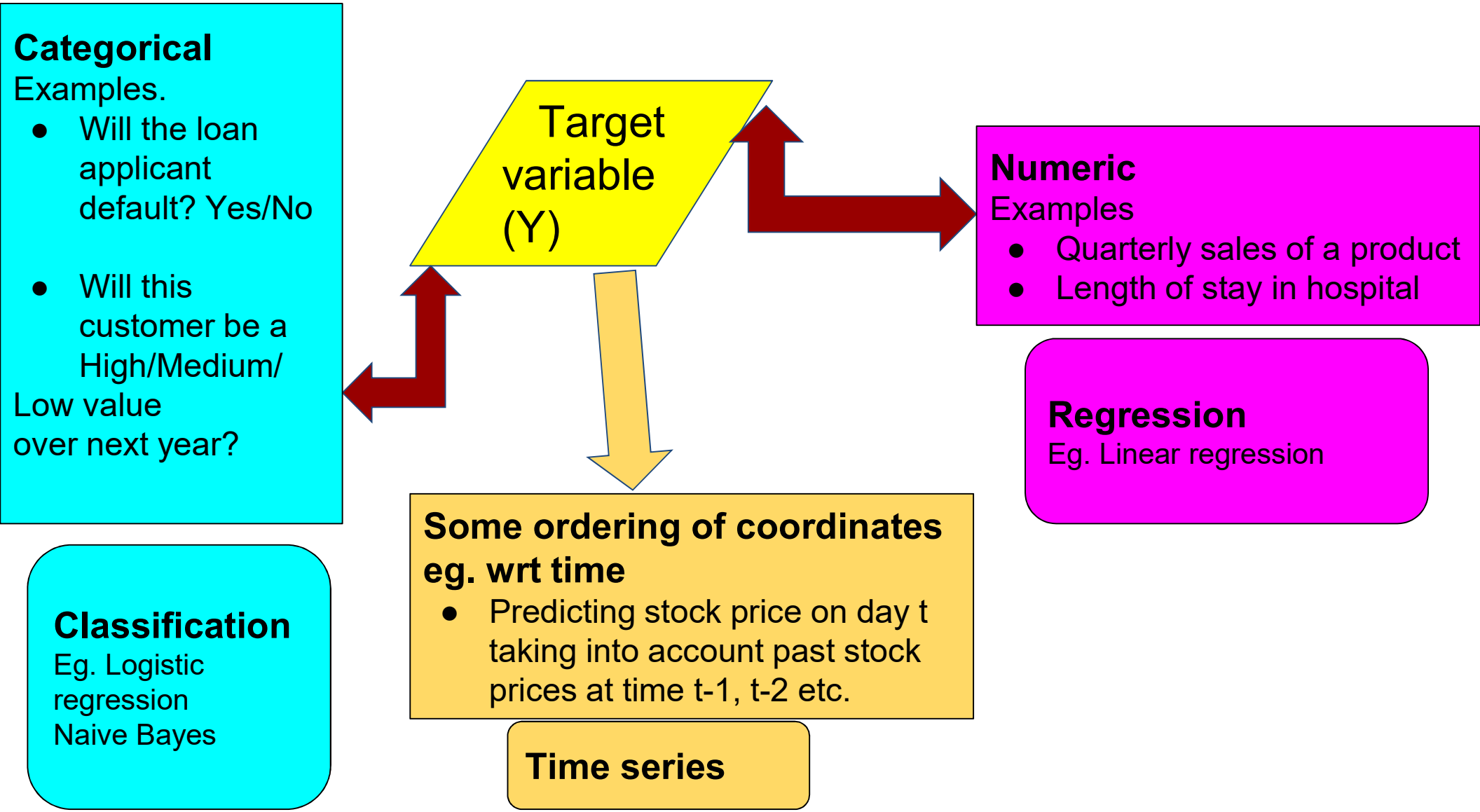
In many practical applications there are,

- 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.

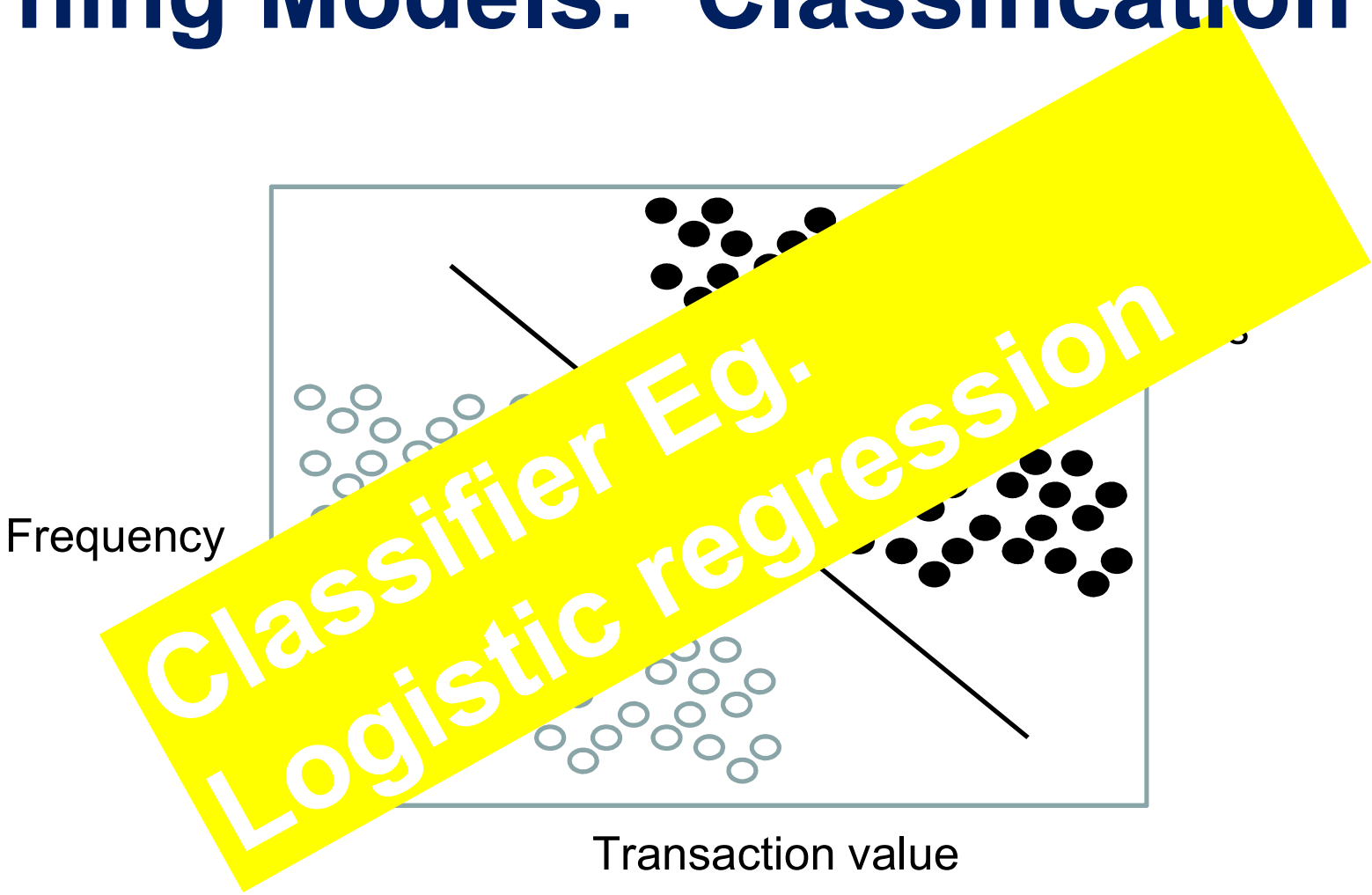
Common classes of practical learning problems

- 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



Learning Models: Classification



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

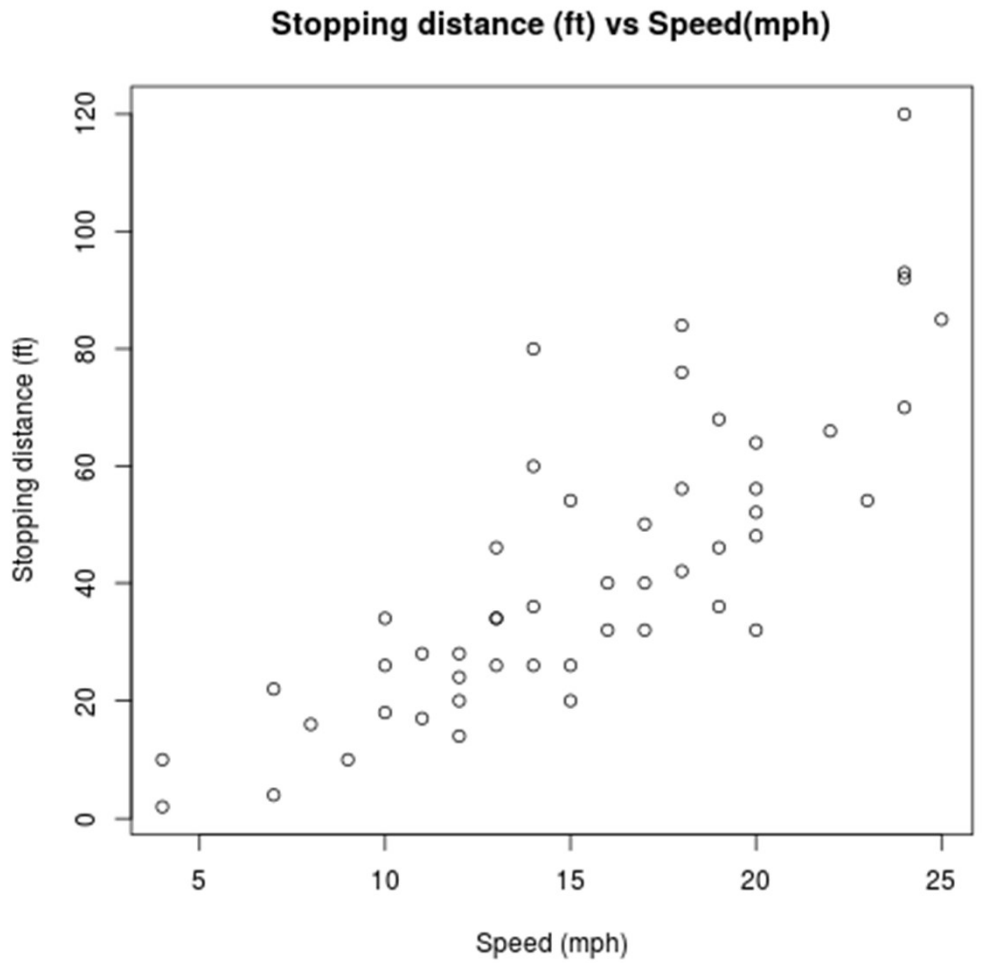
Problem : To predict stopping distance of a car given the speed.

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

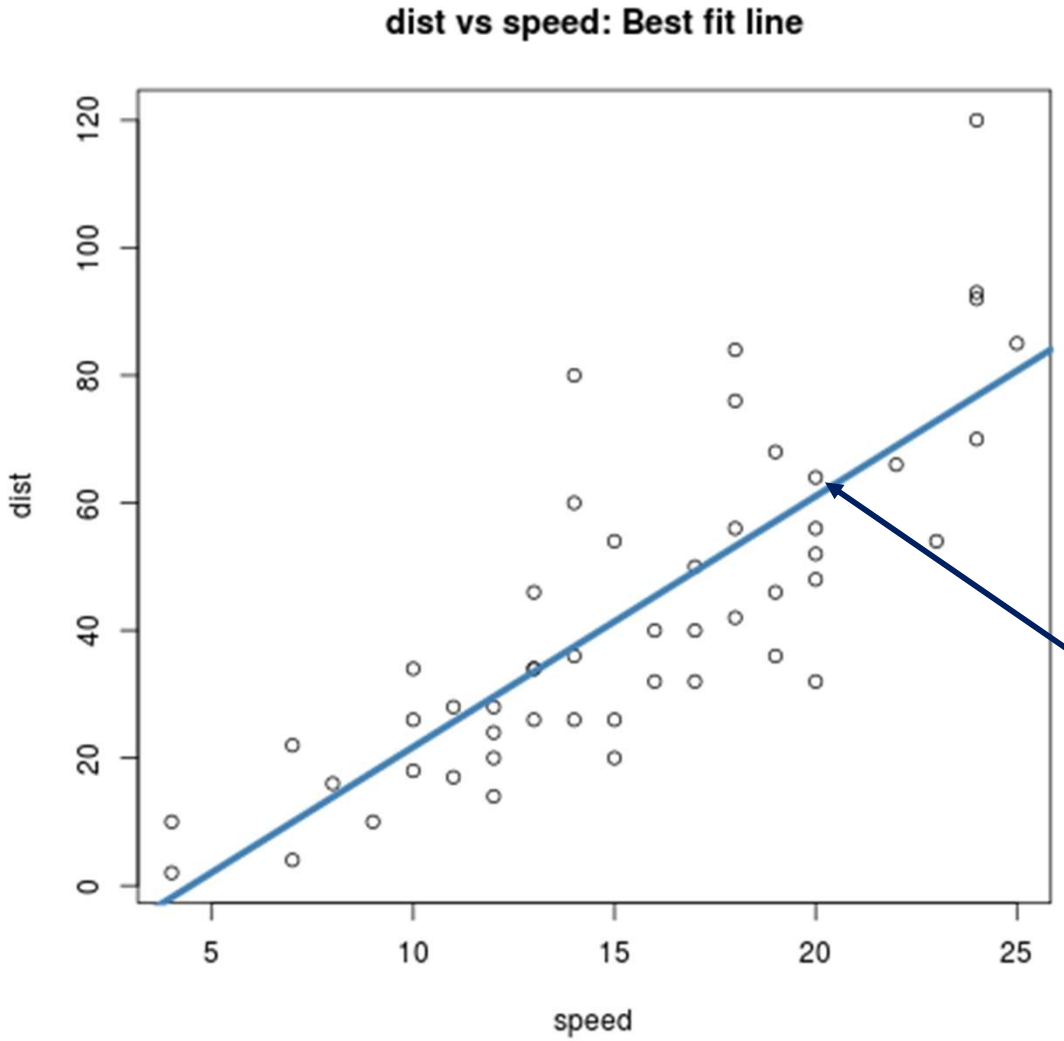
Snapshot of cars data

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

Linear regression model : Probabilistic relationship

In almost all practical examples, we :

- expect a **probabilistic relationship** rather than a **deterministic relationship**

- are using a sample to make inferences about the population.

Linear regression model : Probabilistic relationship

Probabilistic relationship formalized as :

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

Systematic component modelled as

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots$$

Random error component

Assumed to be normally distributed with 0 mean.

Linear regression model :

With only one predictor variable, the simple regression model is of the form

$$y = \beta_0 + \beta_1 x$$

or more precisely,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where

$$\epsilon_i \sim N(0, \sigma^2)$$

Linear regression model :

- The regression coefficients are estimated from the data
- Using the fitted model, we can estimate the value of the response variable for a given predictor as :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

where the residual error is

$$r_i = y_i - \hat{y}_i$$

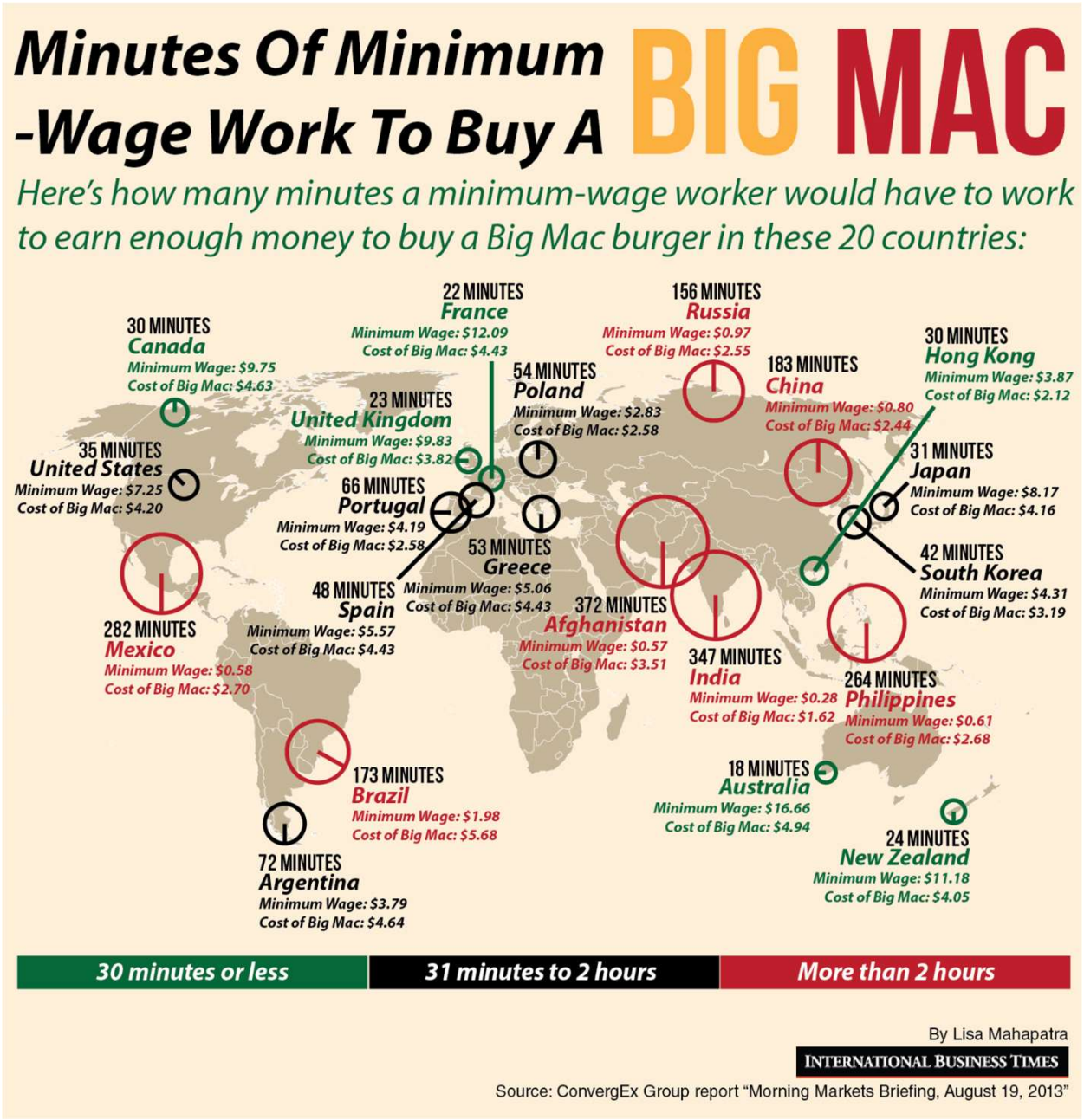
A running example : The Big Mac “Index”

THE BIG MAC INDEX

How many burgers you get for $50 USD?

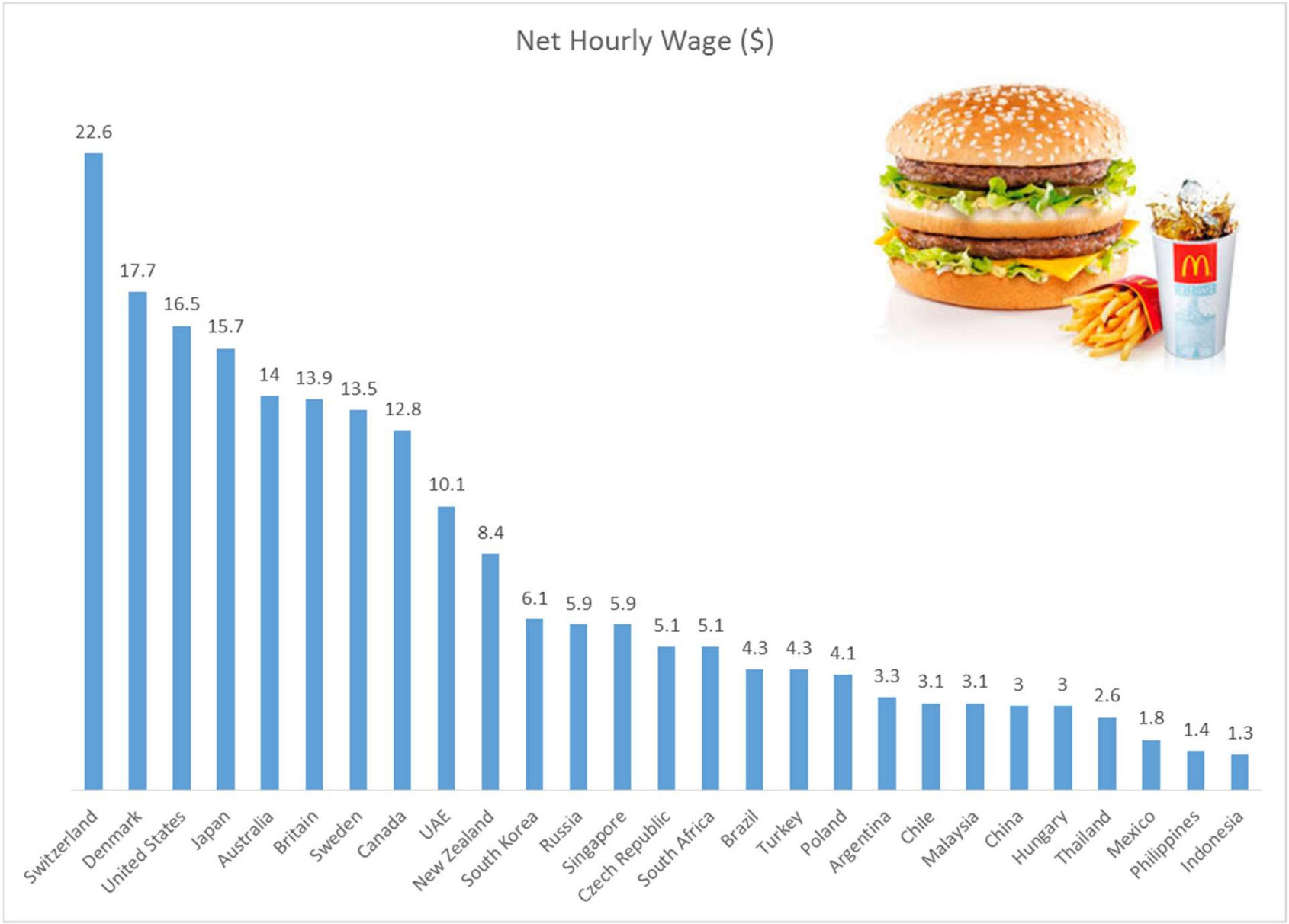
Source: The Economist (Jan 2012)

* Chicken burger



Burgernomics by UBS Wealth Management Research

Net hourly wage ($) in various countries



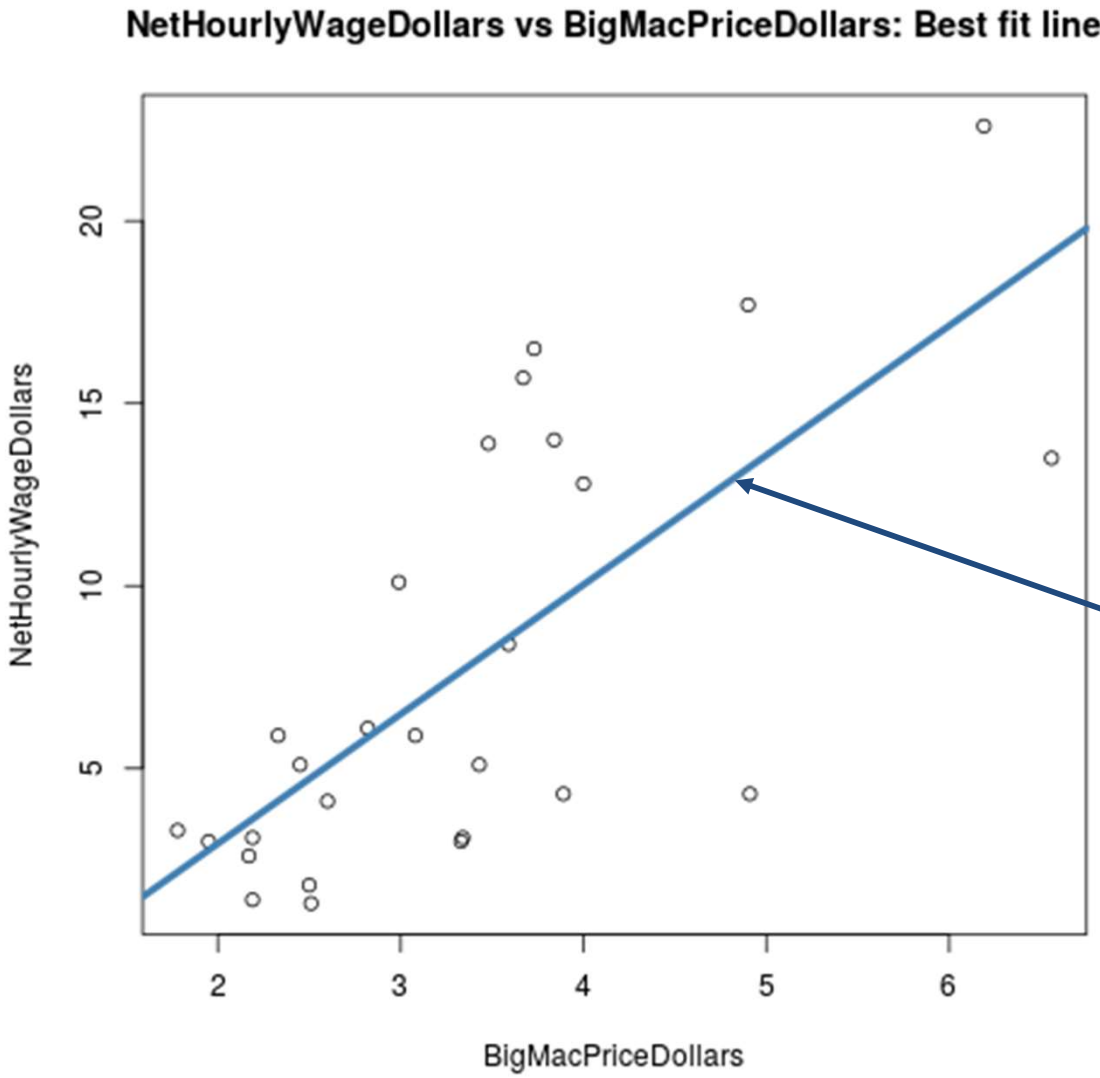
Problem :

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

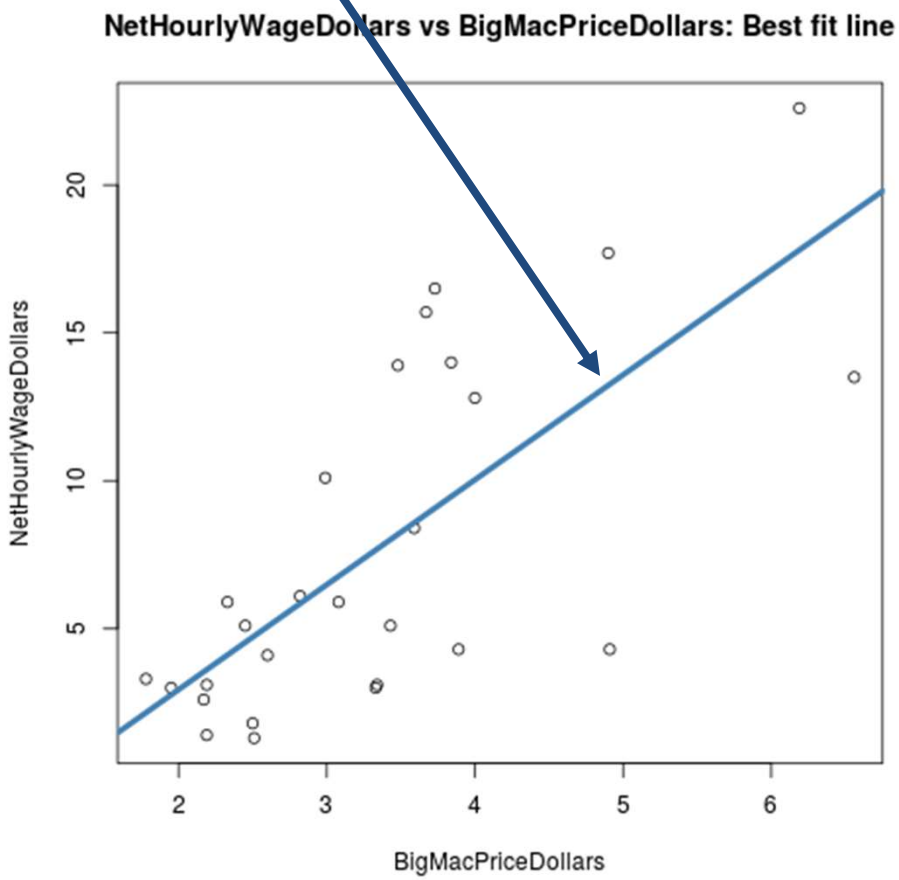


Equation of the best fit line

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, Standard error
- 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.**

Recall : Hypothesis test and p-value

Informally : p-value is the probability of obtaining a result equal to or more extreme than what was actually observed, assuming the **null hypothesis** is true.

A **small p-value** thus indicates **strong evidence against the null hypothesis**, i.e. sufficient evidence to reject the null hypothesis in favour of the alternate hypothesis.

A small, fixed but arbitrarily pre-defined threshold value α referred to as the level of significance (typically ≤ 0.05) is used to accept/reject the null hypothesis.

p-value and correlation

Null hypothesis : No correlation exists
i.e. correlation coefficient is 0.

Alternate hypothesis : Correlation coefficient is significantly different from 0 i.e. a significant correlation exists.

An α of **0.05** indicates that the risk of concluding that a correlation exists—when, actually, no correlation exists—is 5%.

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…

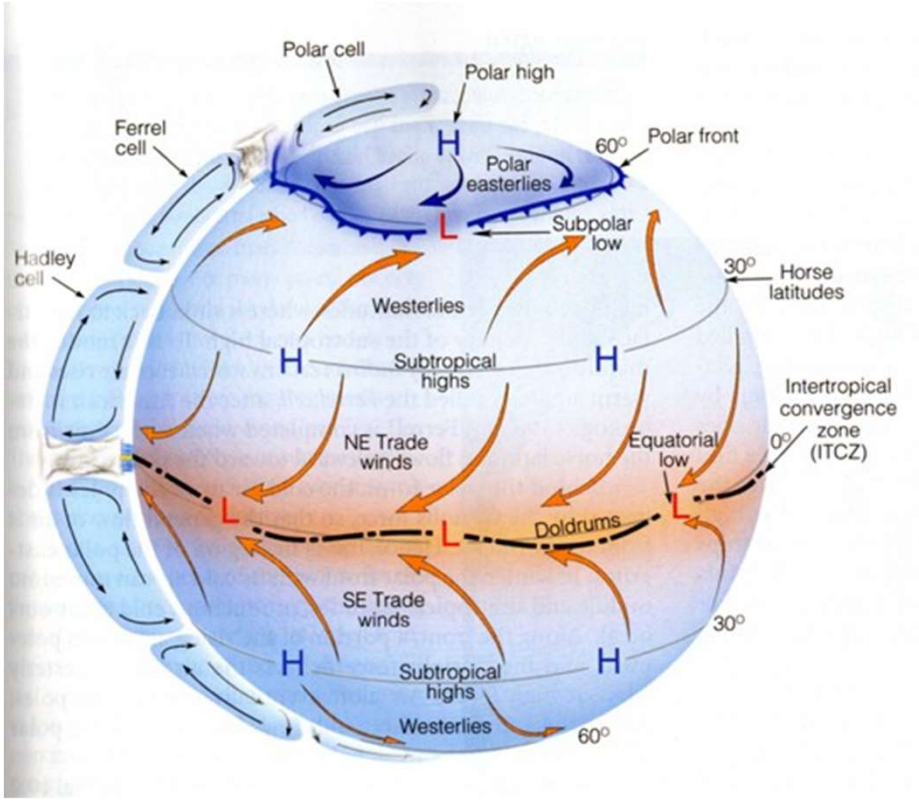


Image source :
http://slideplayer.com/slide/7972625/

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)

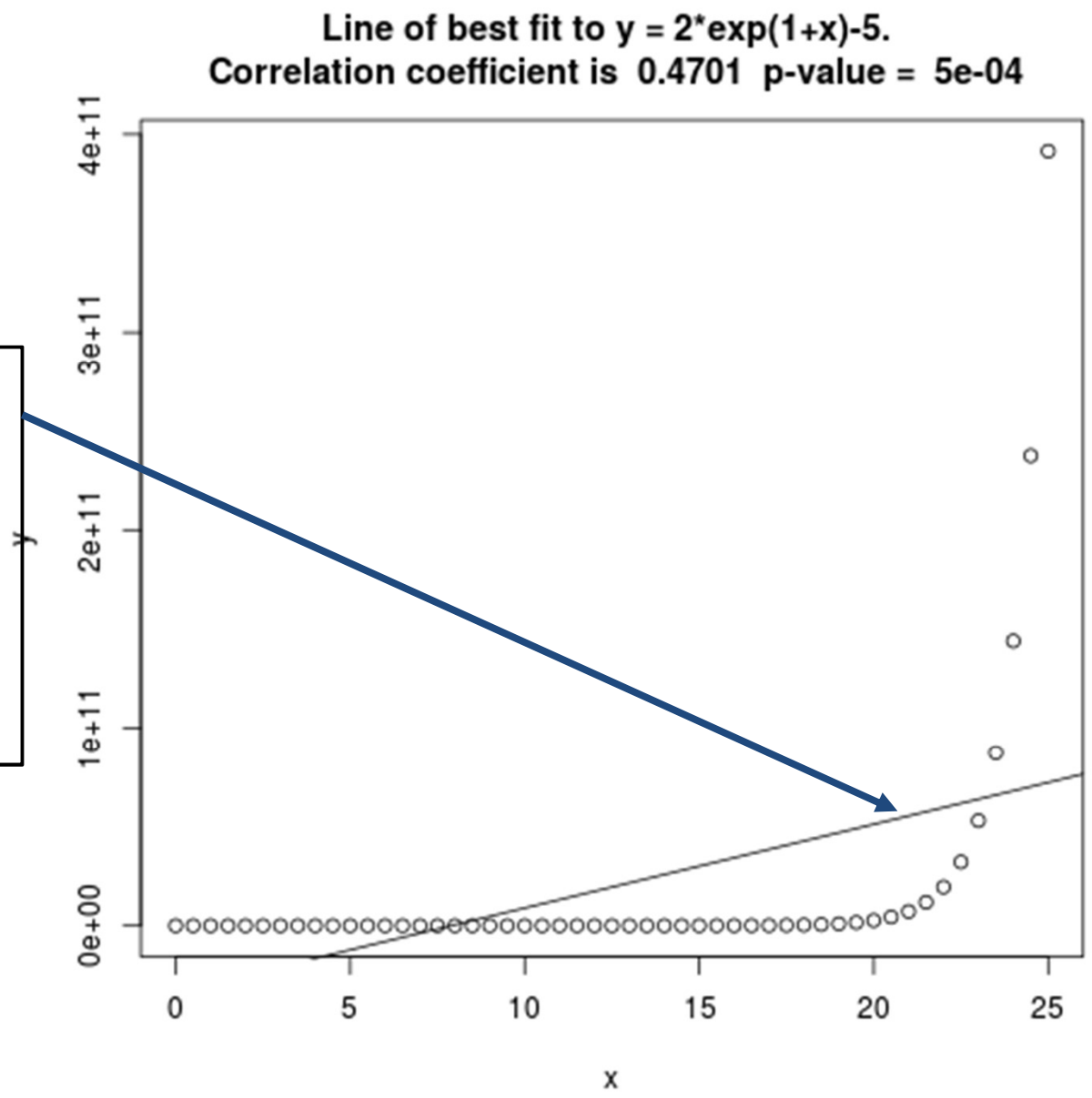
- Being able to fit a linear model does not necessarily mean model is good.

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

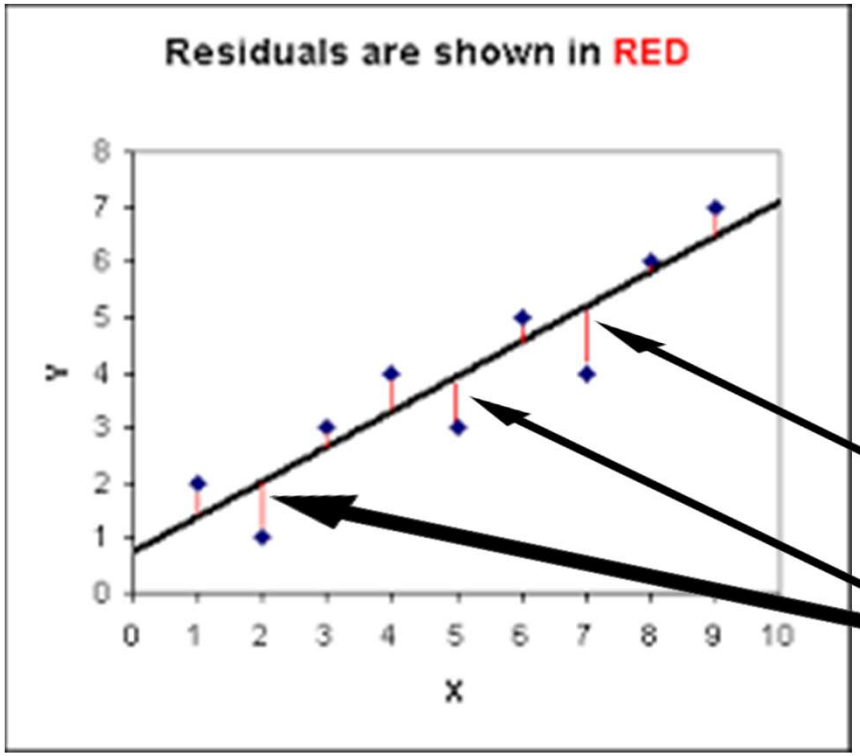
Line of best fit :
y = 4.224e+09 x - 3.329e+10

Correlation coefficient :
0.4701

p-value : 5e-04



A small detour : Computing the line of best fit



Differences b/w actual and predicted

Predictor variable (x)	Response variable (y)	Residual (r)
x_1	y_1	r_1
x_2	y_2	r_2
...	...	...
x_N	y_N	r_N

In **Ordinary Least Squares (OLS), line of best fit** is one that **minimizes the sum of squared differences between actual and predicted values**

Computing the line of best fit

The sum of squared differences between predicted and actual values (also referred to as **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.

Computing the line of best fit

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

and hence $SSE = \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2$

To minimize SSE take partial derivatives of SSE wrt to the unknown parameters and use first order minimization conditions to get :

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

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

Computing the line of best fit

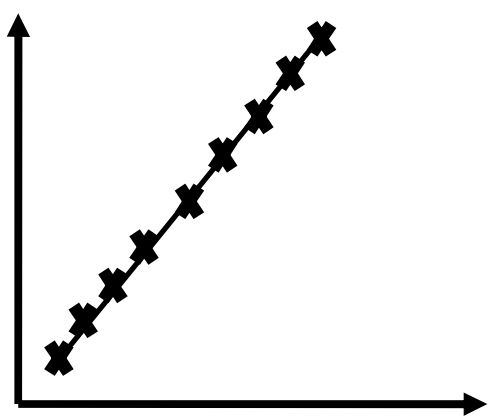
Solving the resulting system of equations and using some algebra (not shown) we get :

$$\beta_1 = \frac{\sum y_i x_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2} = \frac{cov(x,y)}{var(x)} \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

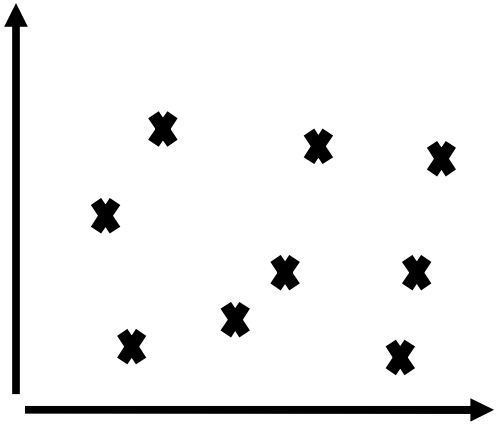
where,

- $\bar{x}$ is the sample mean of x.
- $\bar{y}$ is the sample mean of y.

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



Accurate Linear Correlation



No Linear Correlation

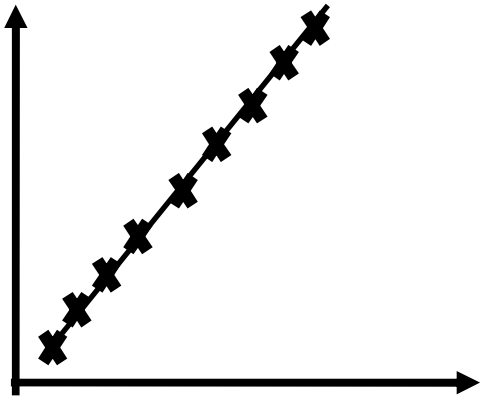
Basic measures of goodness of the fit :

- The **correlation coefficient**.
- C**oefficient 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.

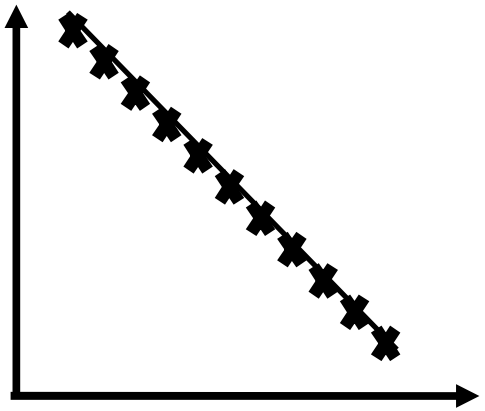
Correlation Coefficient and regression

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



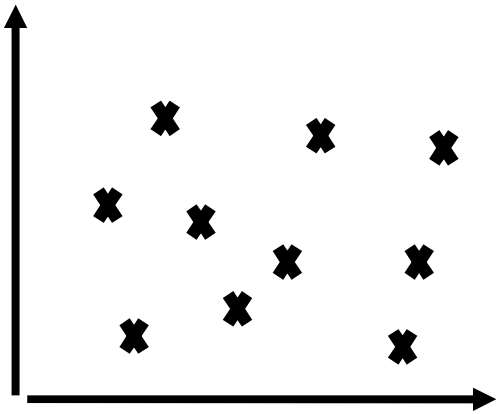
r = 1

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 :

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

Analysis and assessment of the model

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

- **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.

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

- **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**

Assessing 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. So F test for overall significance tests the same thing as t test.
    - (Null Hypothesis) $H_0 : \beta_1 = 0$
    - (Alternate hypothesis) $H_1 : \beta_1 \neq 0$

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

Assessing individual coefficients

- Inference about individual coefficients (including intercept) about the population slope can be made from the sample using either confidence intervals or hypothesis tests.
- The p-value is computed from the test statistic assuming a t-distribution.

Testing the slope

What is the Null Hypothesis?

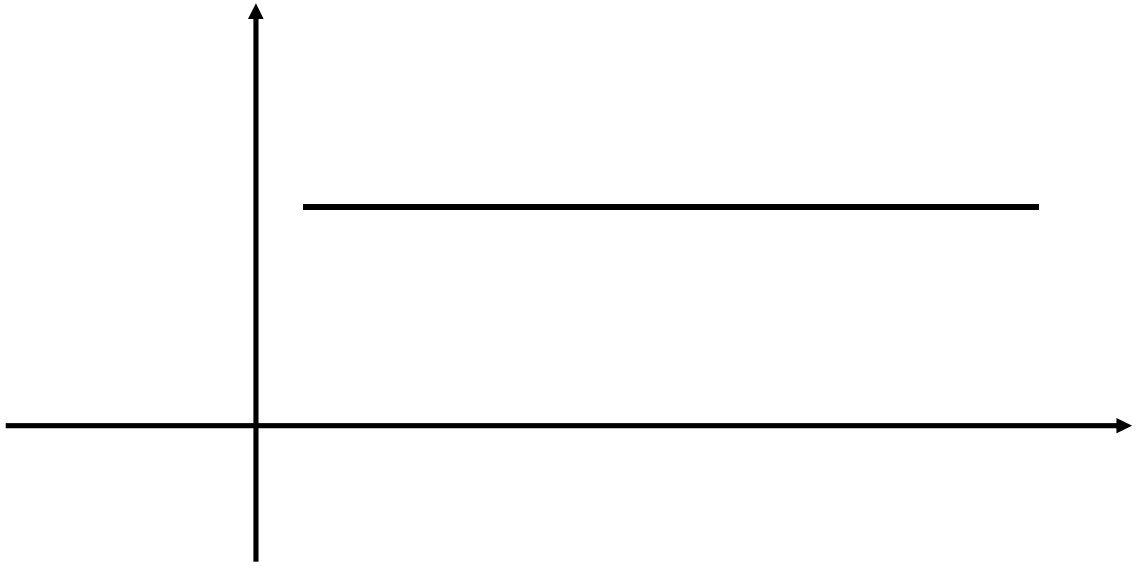
$$H_0: \beta_1 = 0$$

What is the Alternative Hypothesis?

$$H_1: \beta_1 \neq 0$$

Testing the Slope

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



Testing the Slope

Note : The regression coefficients are calculated from the given data points in the sample and hence may be different from the "true" values i.e. they may be considered as **point estimators of the “true coefficients”**

Since the population standard deviation is unknown, we instead use the corresponding MSE to calculate the respective standard errors, these are used as the test statistic in the respective t-tests.

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{s_{xx}}} \quad \text{and} \quad se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right]}$$

where,

- n = number of samples
- $\hat{\sigma}^2 = MSE = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}$
- $\bar{x} = \frac{\sum_i x_i}{n}$
- $s_{xx} = \sum_i (x_i - \bar{x})^2$

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

Hands-on exercise : Lab Part 1

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

Concepts emphasized

- 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

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

Regression diagnostics : Residual analysis

Residual plots

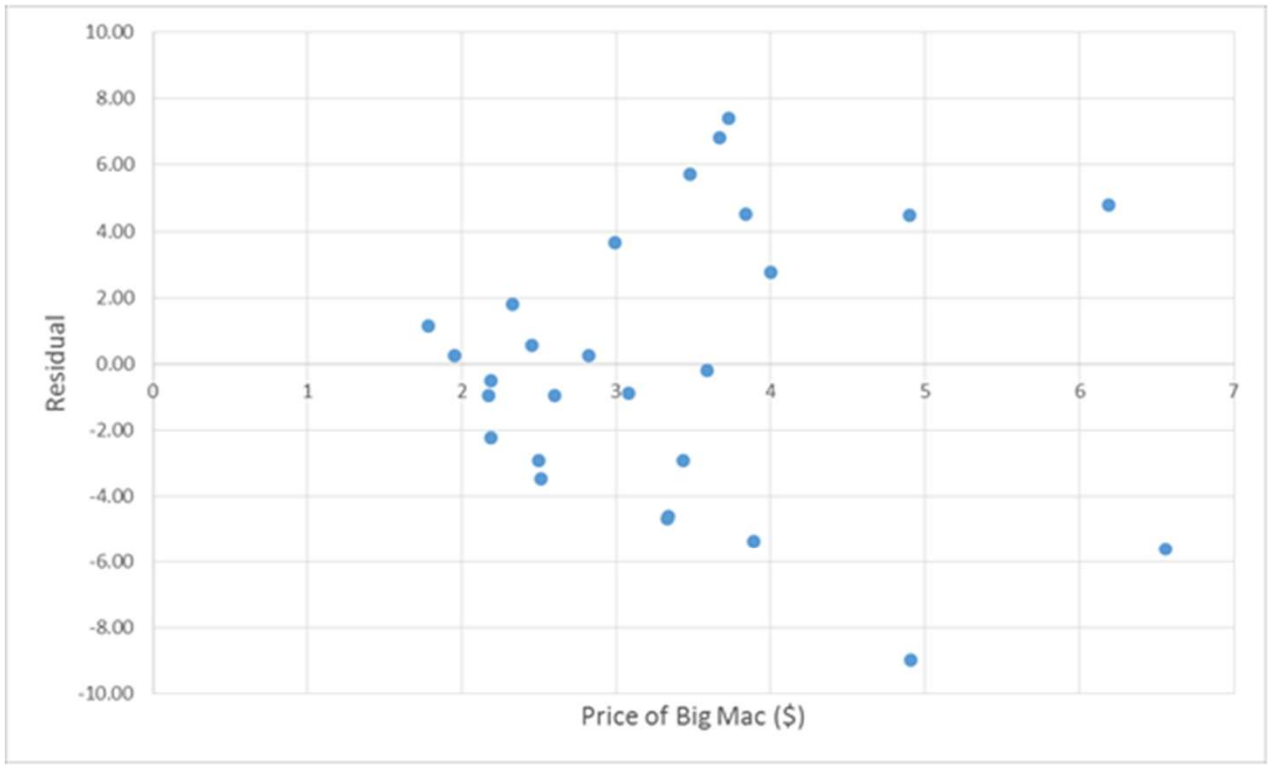
- A residual plot is a scatter plot of vertical residuals versus the independent variable.

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

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \end{aligned}$$

- Problems with linear regression are generally easier to identify by plotting the residuals rather than the original data.
  - If the data are heteroscedastic, nonlinearly associated, or have outliers, the regression line is not a good summary of the data, and it is not appropriate to use regression to summarize the data.

Residual Analysis



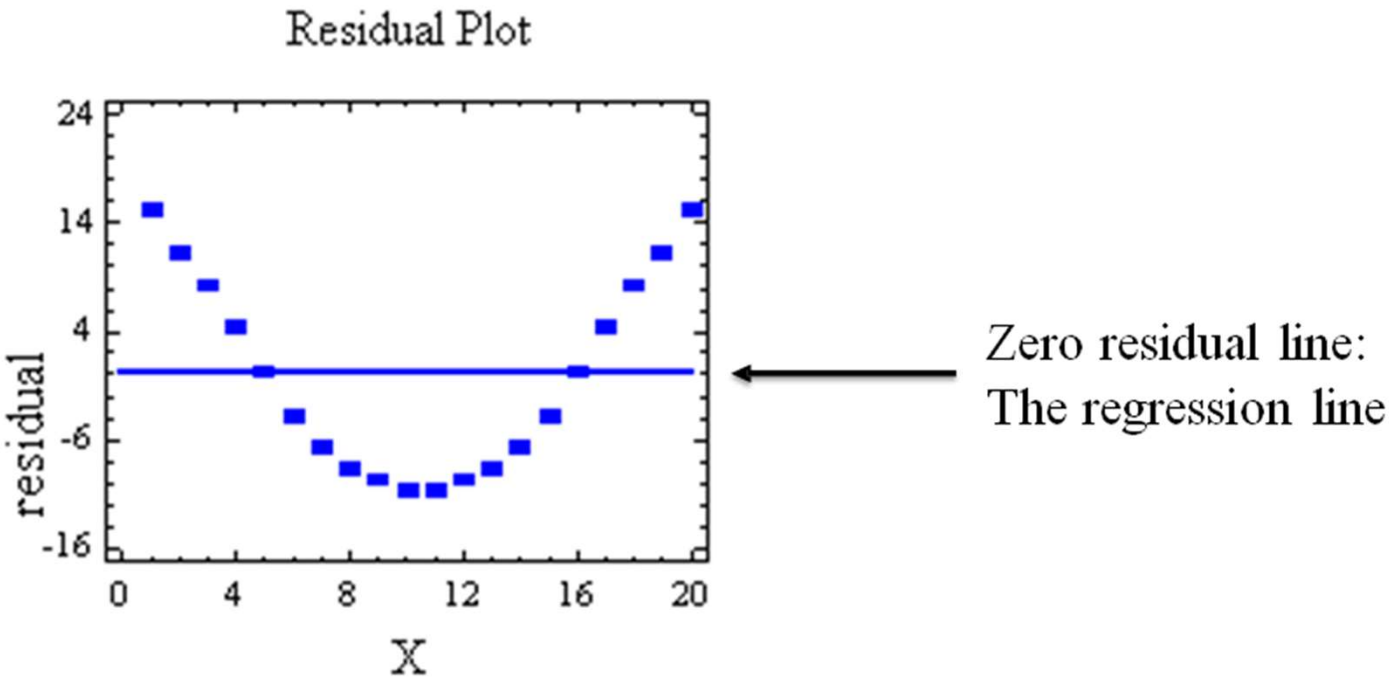
Can be used to locate outliers.

Verification of assumptions of linear regression

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

Assumptions of the Regression Model

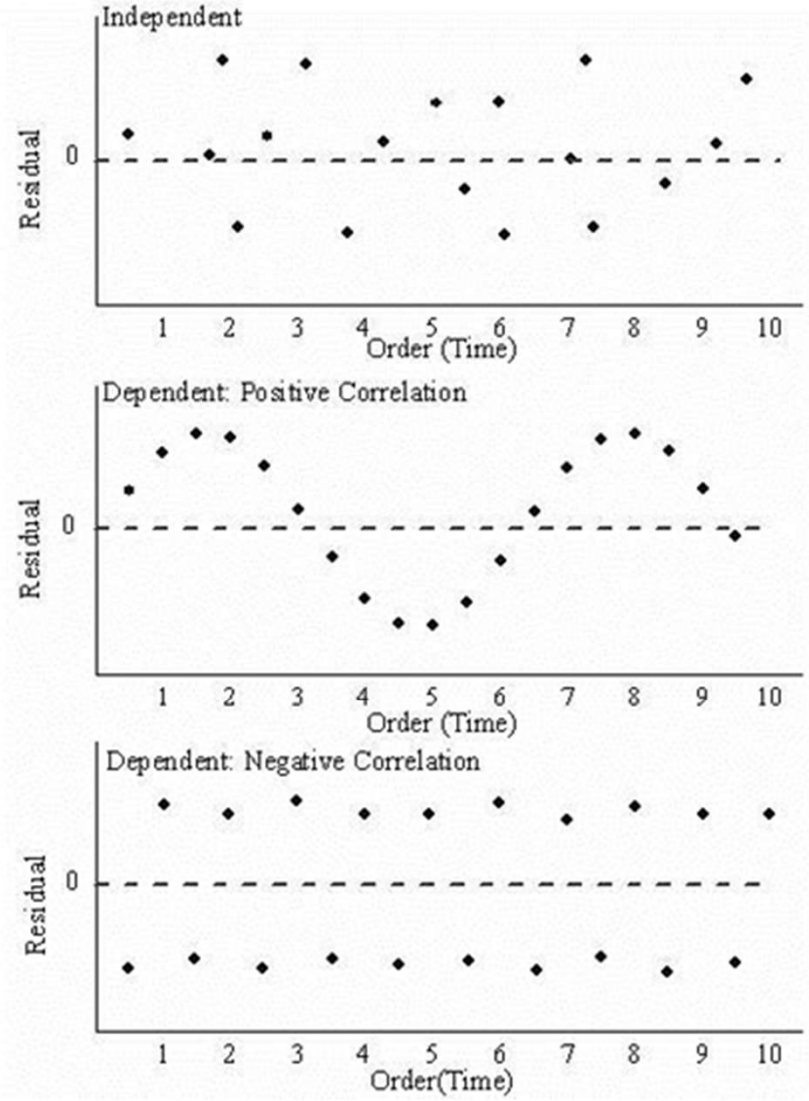
- The model is linear



In above example, a linear model does not seem to be a good fit.

Assumptions of the Regression Model

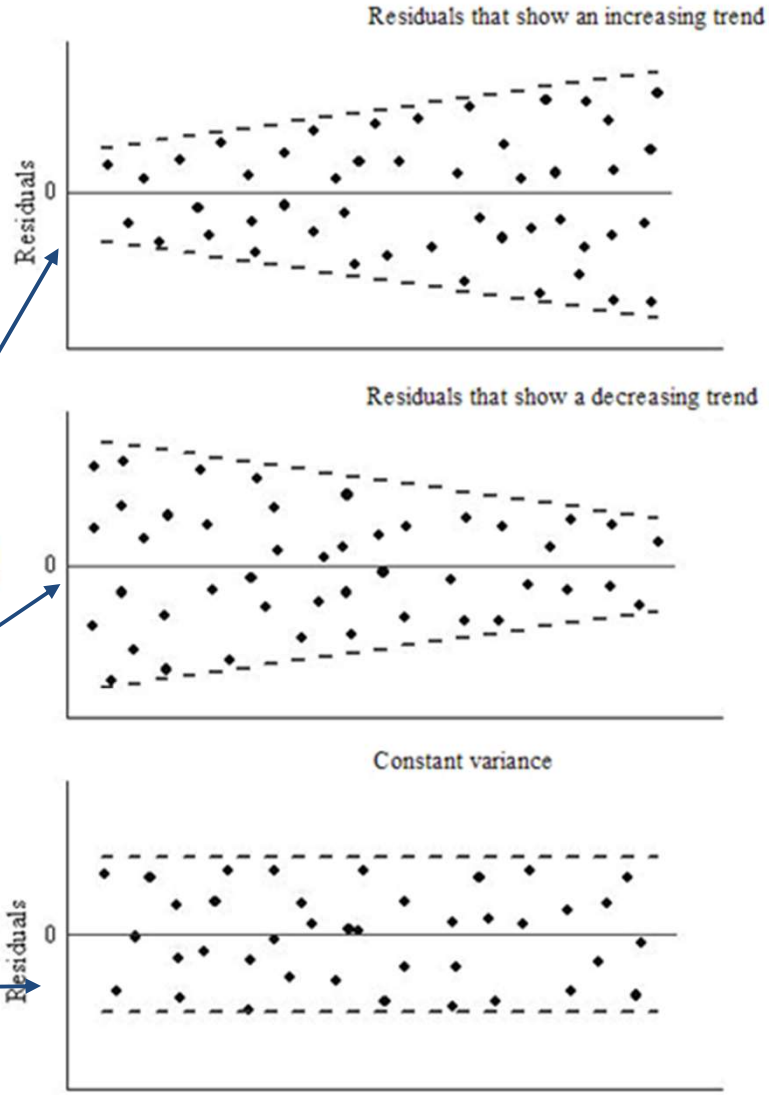
- The error terms are independent
  - Plot against any time (order of observation) or spatial variables preferably. Plots against independent variables may also detect independence.



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.

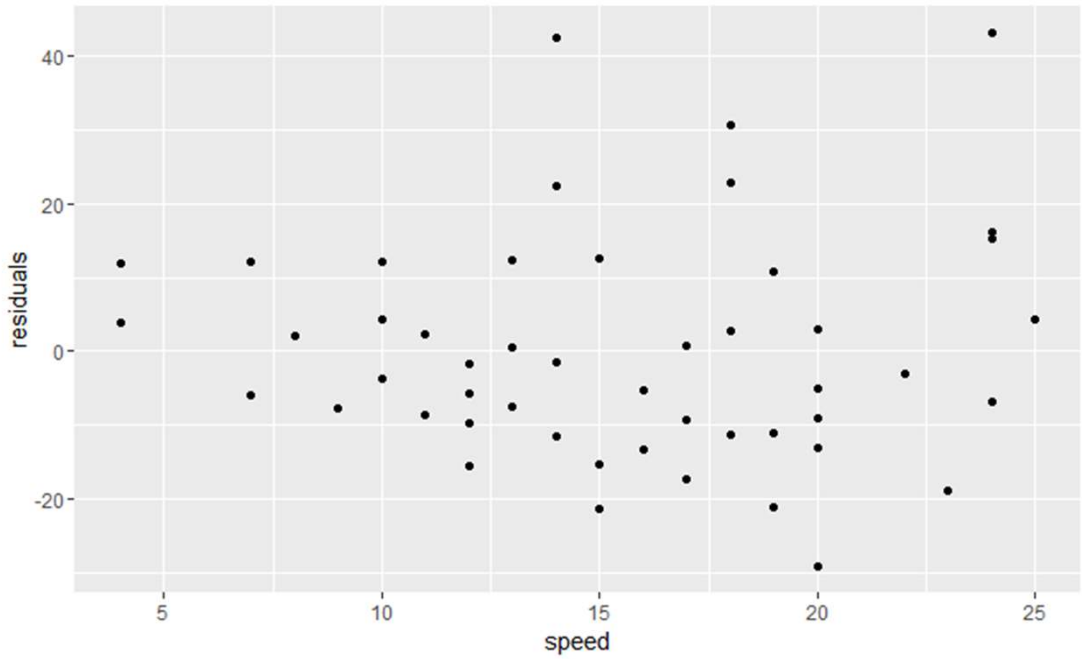


Heteroscedastic

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

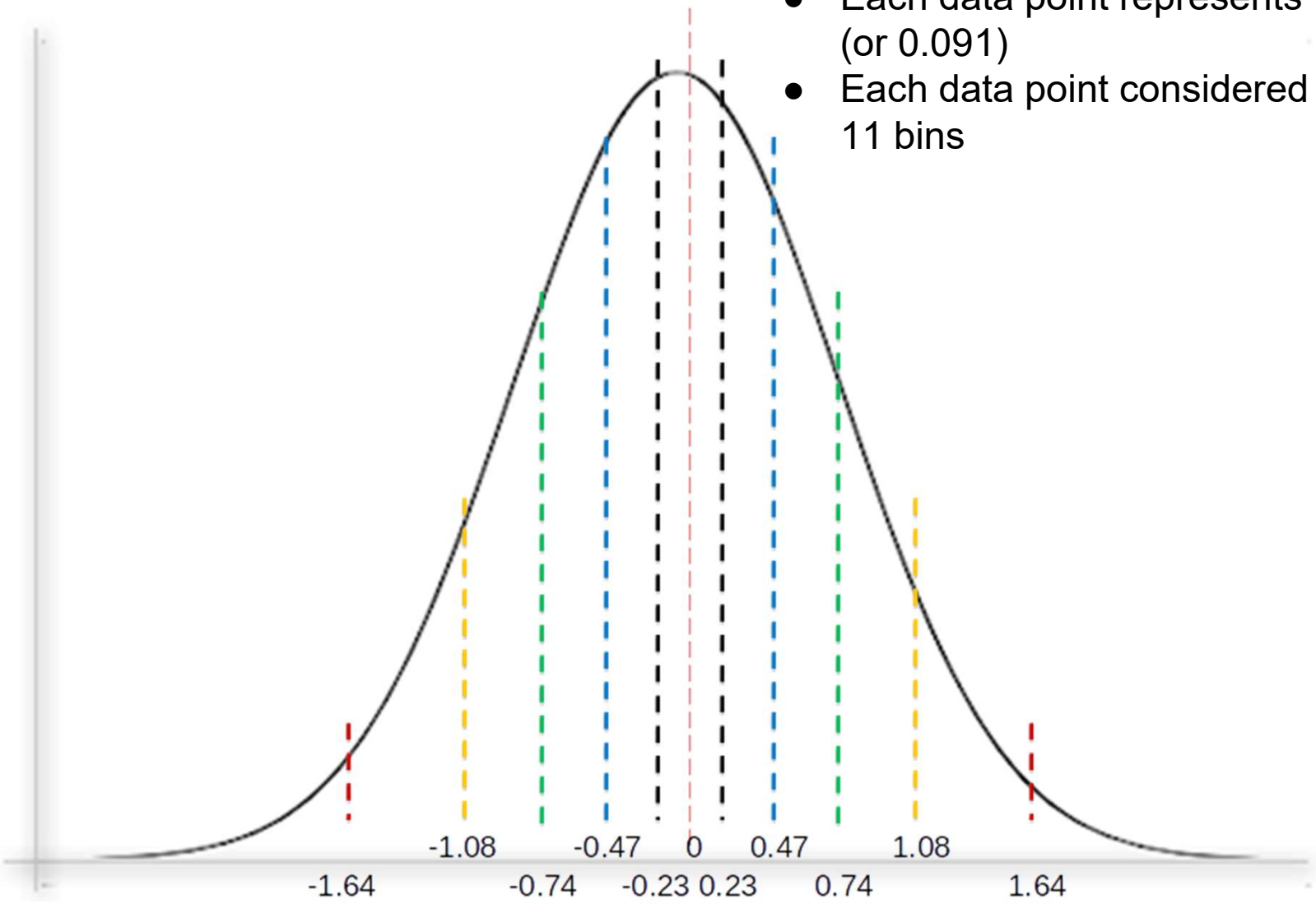
Mechanism to verify above : Q-Q plot

- 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.

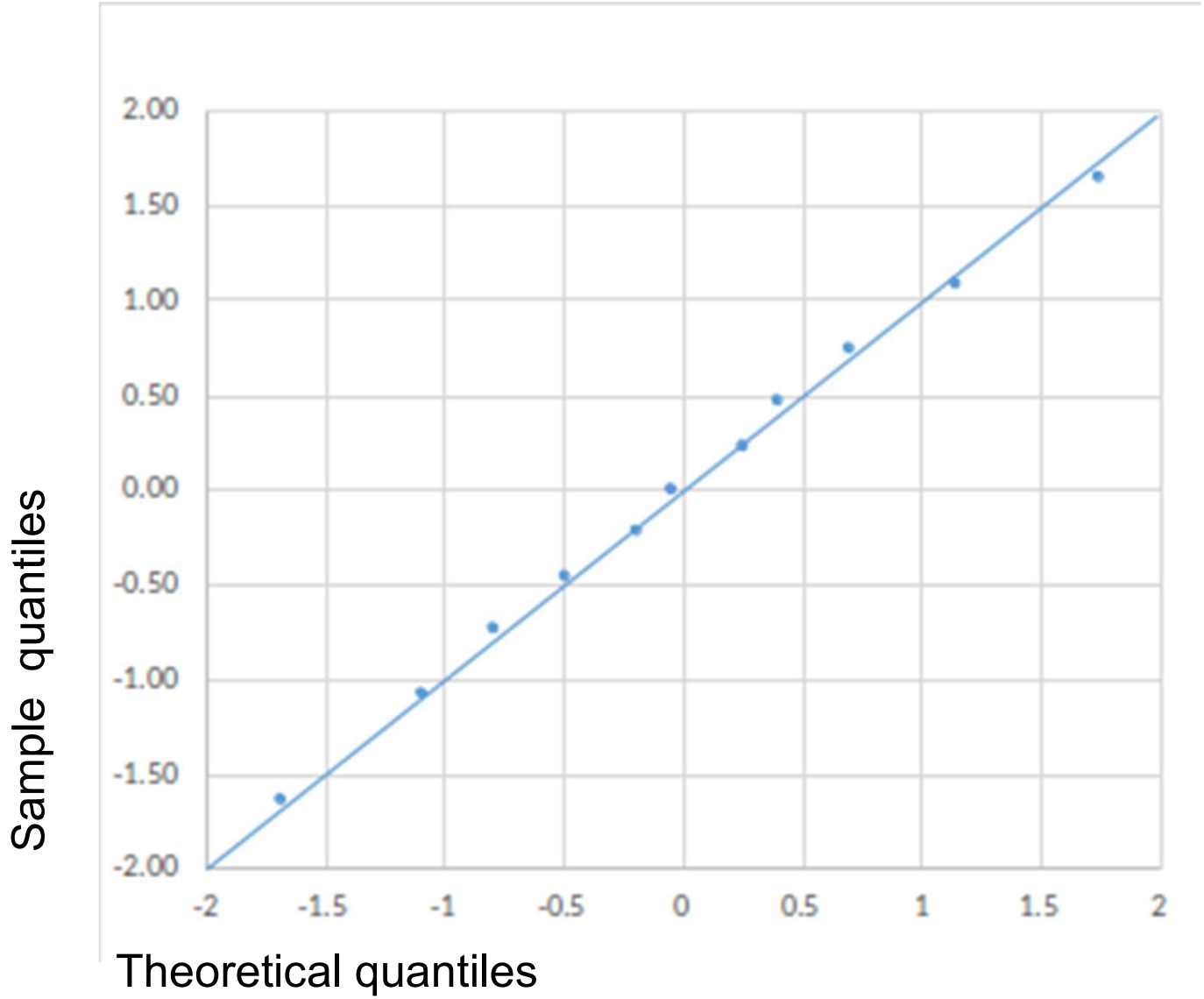
- In linear regression, this data set is the residual errors.
- If the normality assumption holds true, then the z-scores of the residuals should be equal to the expected z-scores at corresponding quantiles.

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



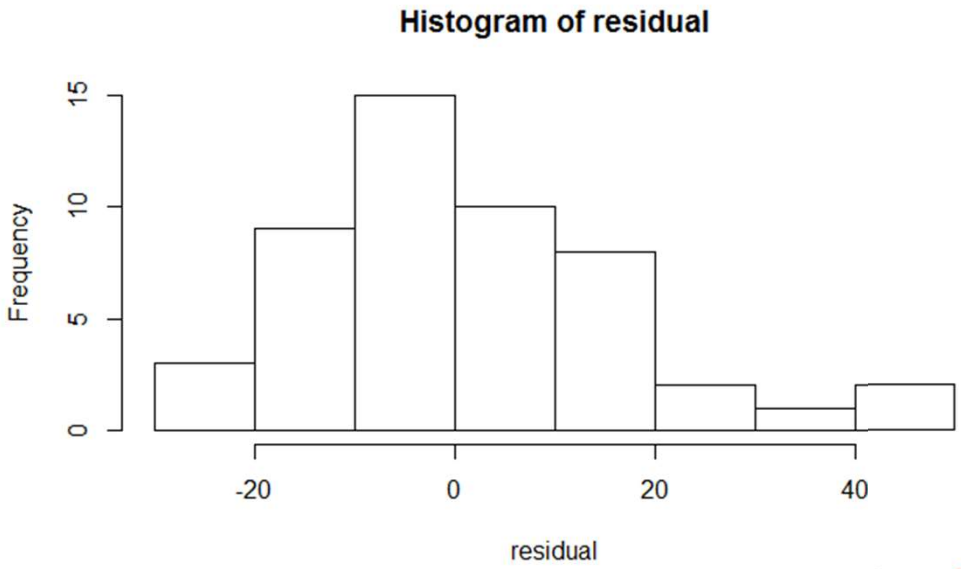
An example of a Q-Q plot

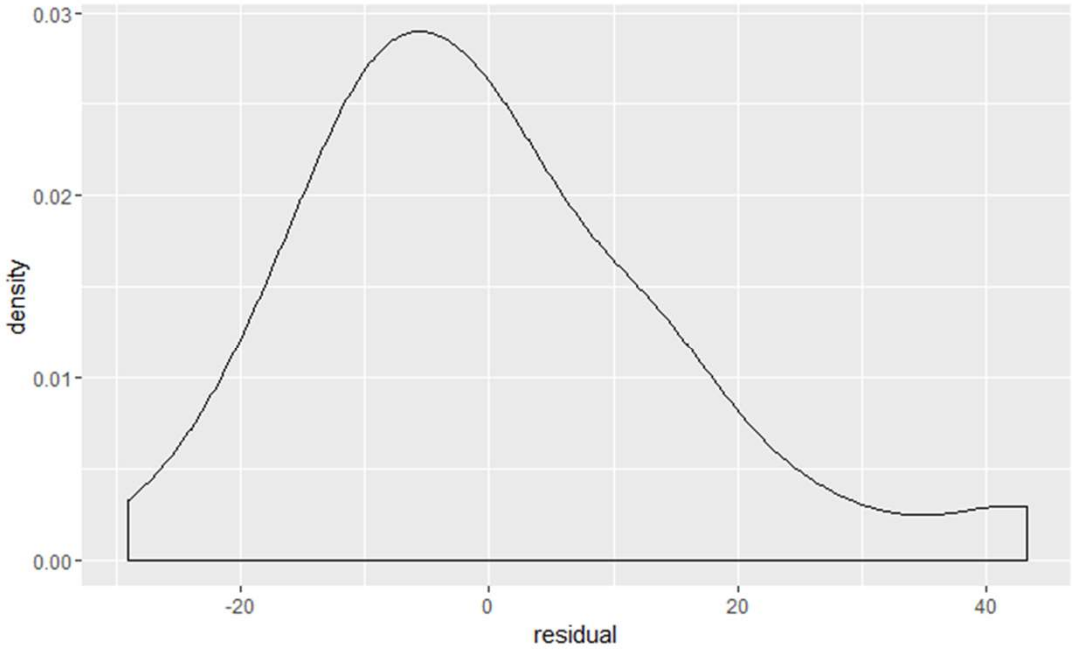


Checking for Normality

- Start by plotting the data

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





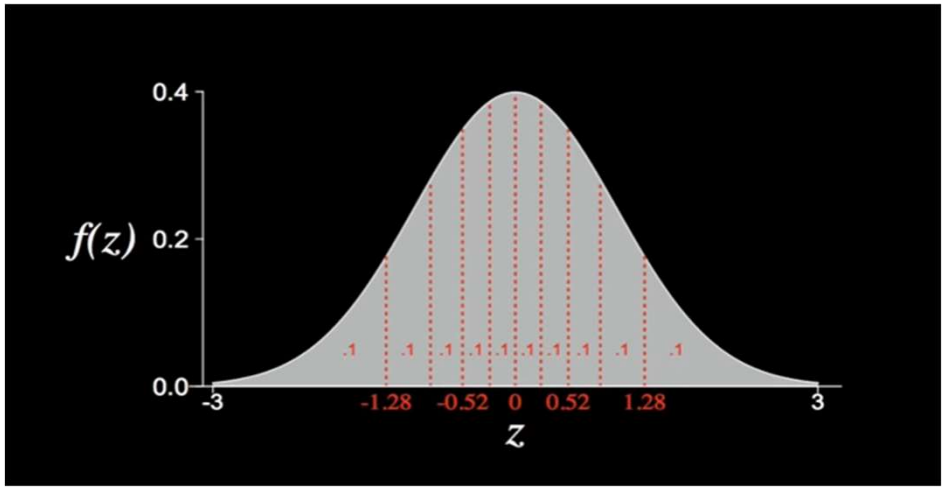

```
> ggplot() + geom_density(aes(residual))  # density plot. Requires ggplot2
```

Is there a better

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



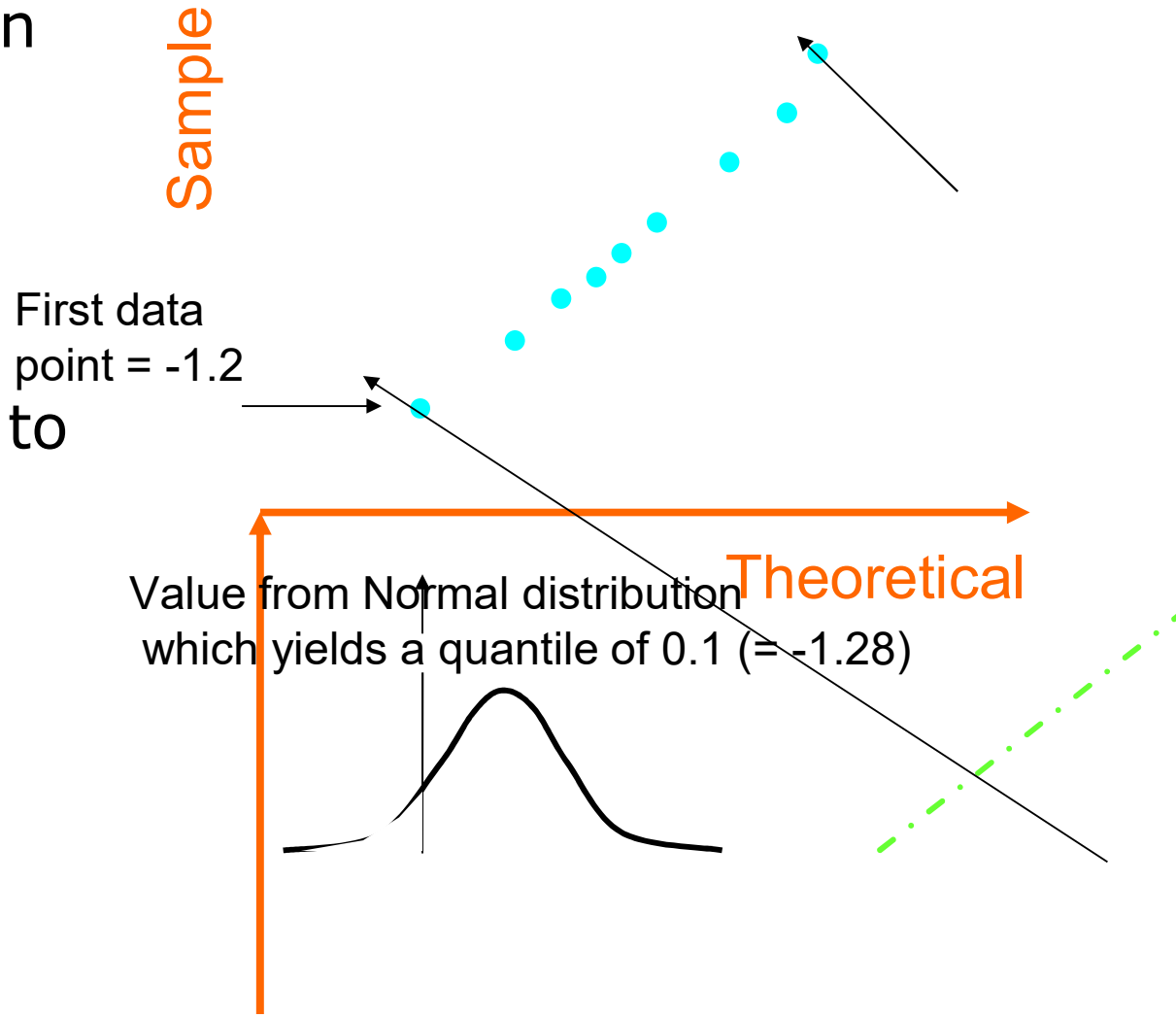
- Then one might expect the smallest of the 9 data points to be from the lowest quantile (0.1)
- Similarly, the largest value would be from the largest quantile (0.9) of the normal distribution

Quantile Quantile Plot

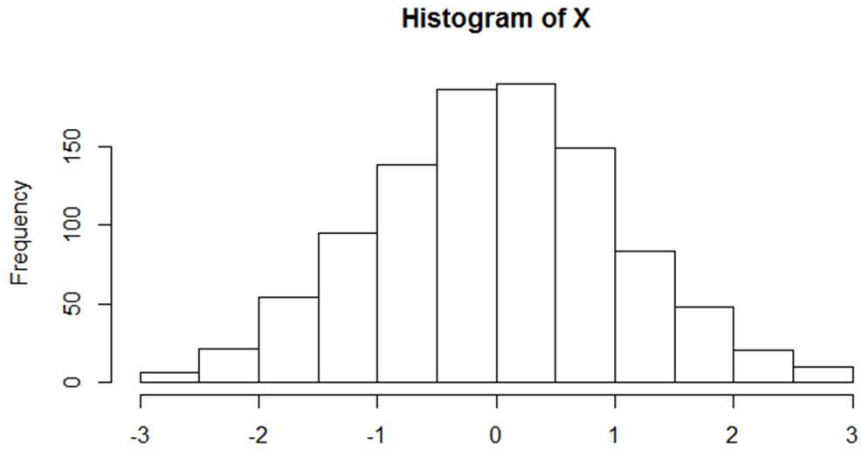
-1.2, -1.11, -1.08, -0.28, -0.25, 0.33, 0.41, 1.37, 1.41

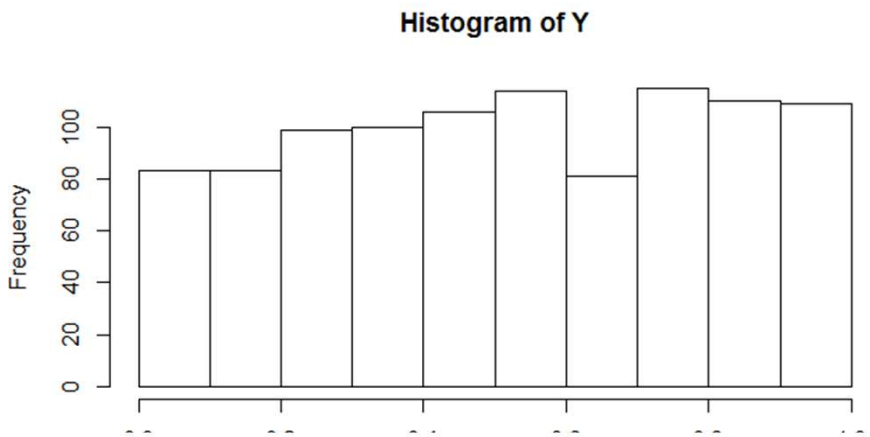
We plot the quantile values for the distribution on the x-axis and the values of the sample on the y-axis

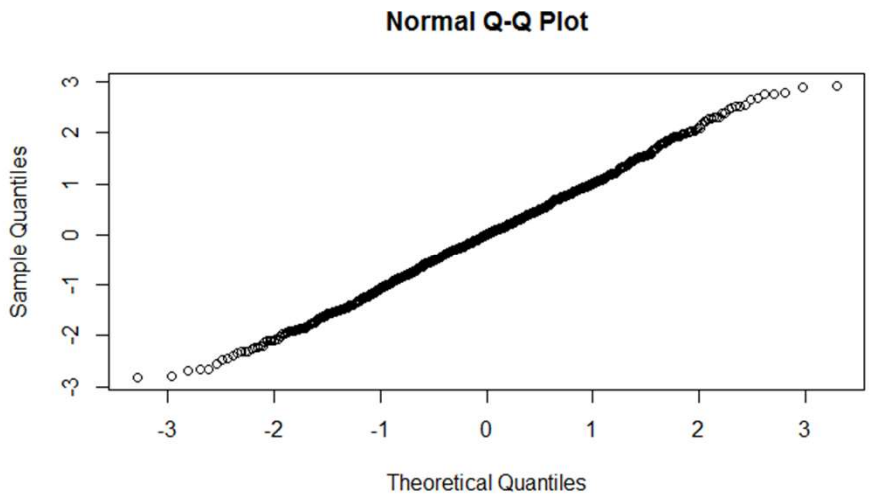
If the points lie on close to a straight line, then the sample is normal

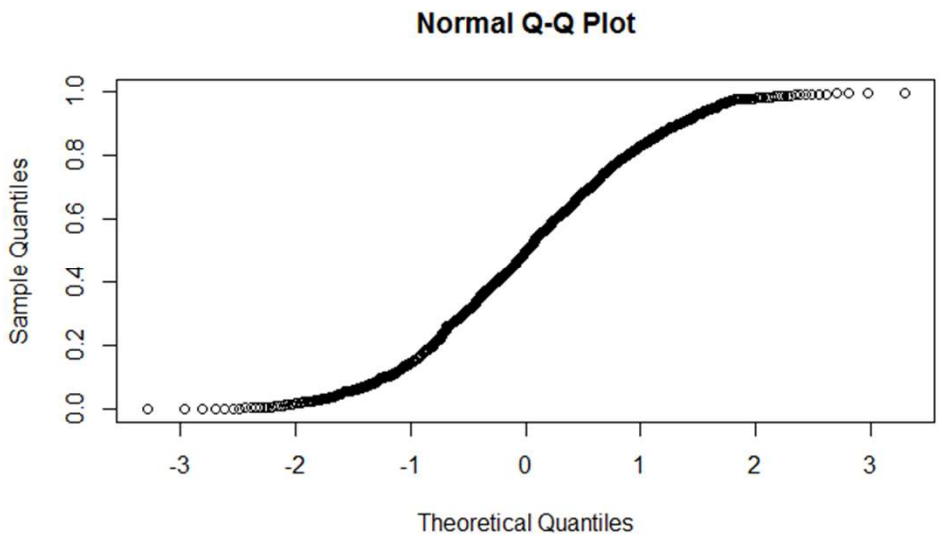


QQ Plot for Normal vs Uniform Distribution










```
> 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

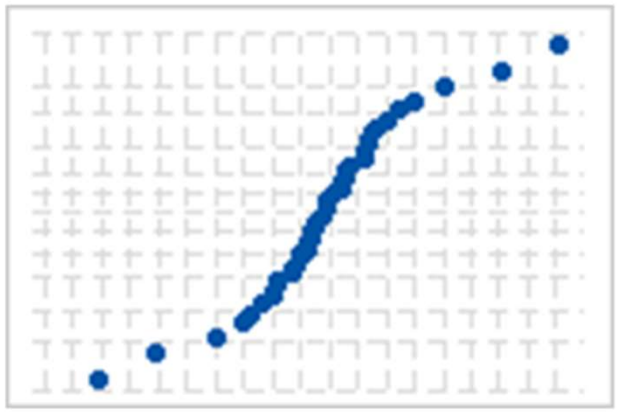
```
> X <- rnorm(1000)  # 1000 data points picked from Normal Dist.
> shapiro.test(X)

        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

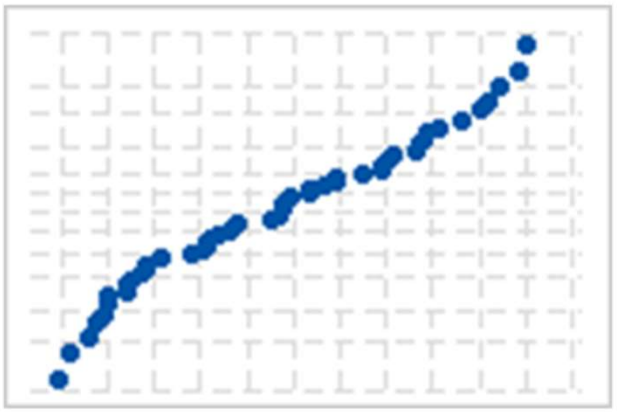
Interpreting Residuals

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

Interpreting Residuals – Non-normality

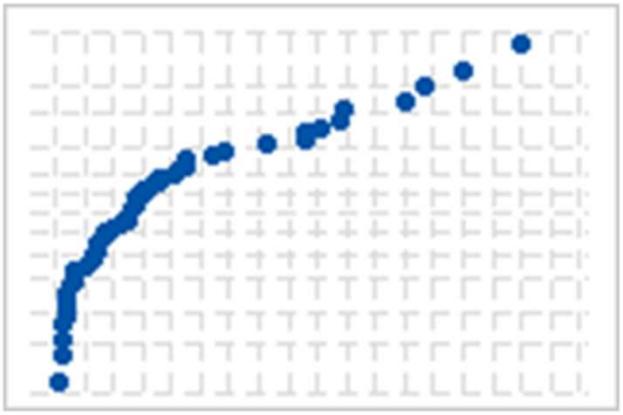


S-curve implies a distribution with long tails

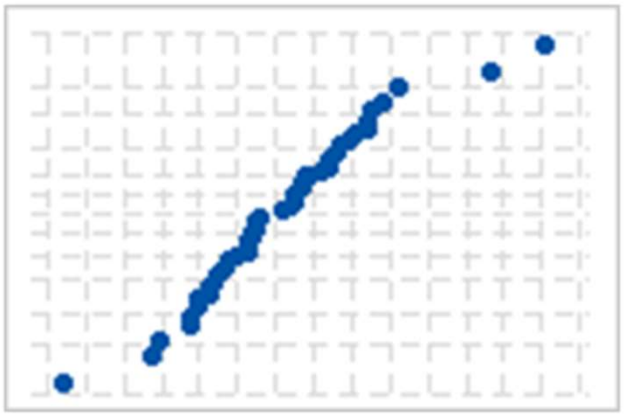


Inverted S-curve implies a distribution with short tails

Interpreting Residuals – Non-normality



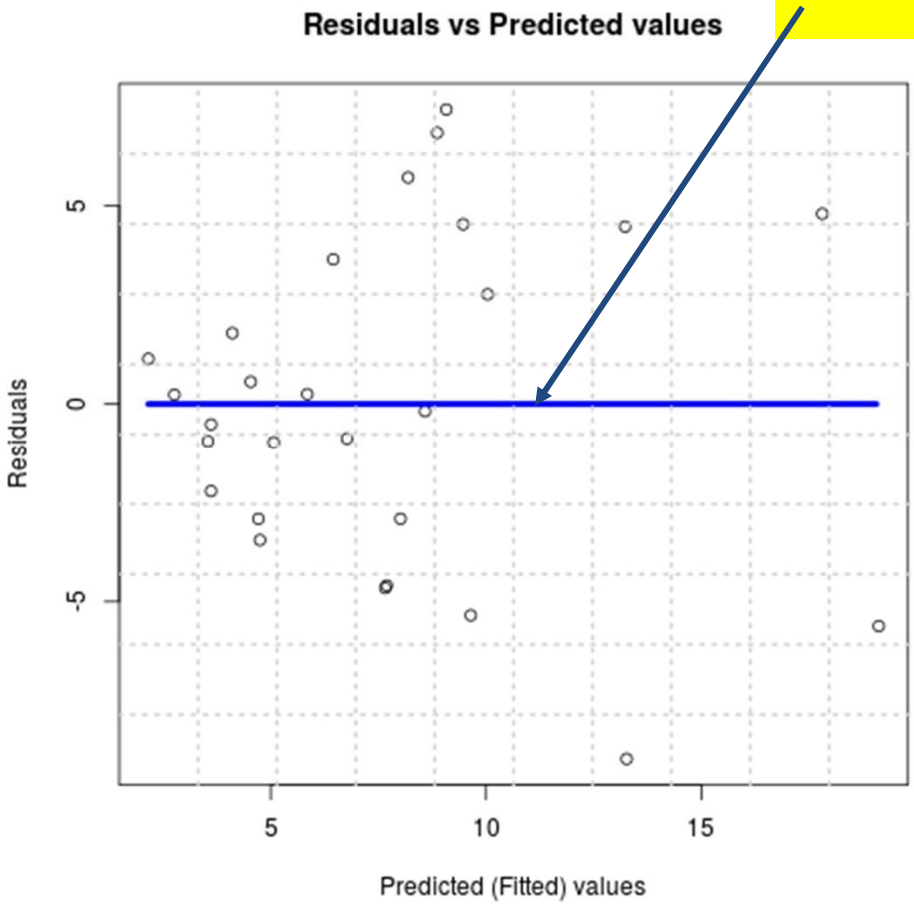
Downward curve implies an asymmetric distribution



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

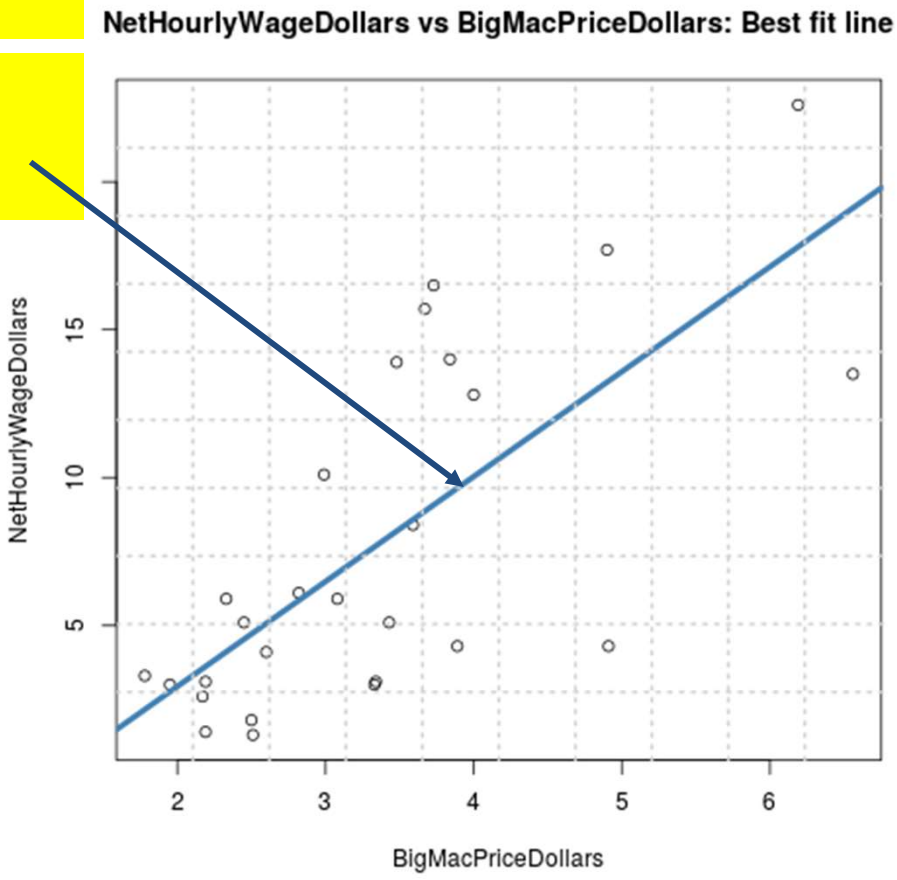
Analysis of residuals : Big Mac example

Zero residual line



Big Mac example :
Plot of residuals vs fitted values

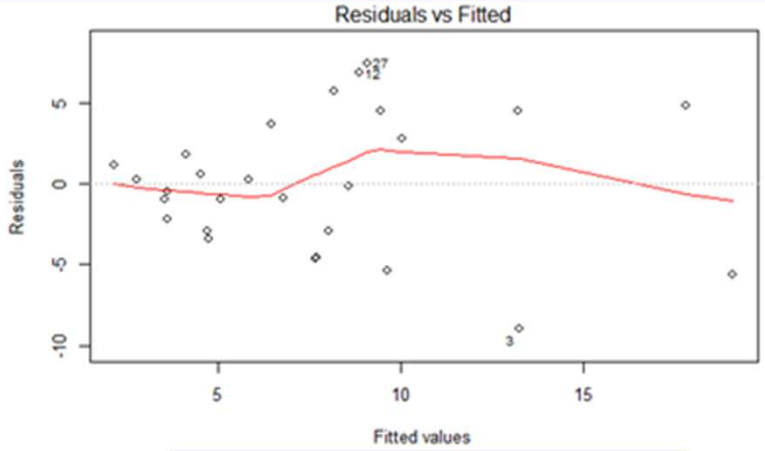
Line of best fit



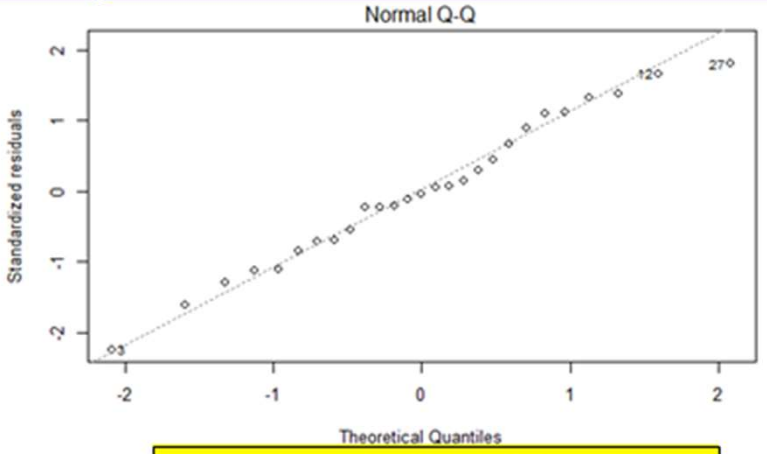
Big Mac example :
Data samples and regression line

Residuals – Big Mac

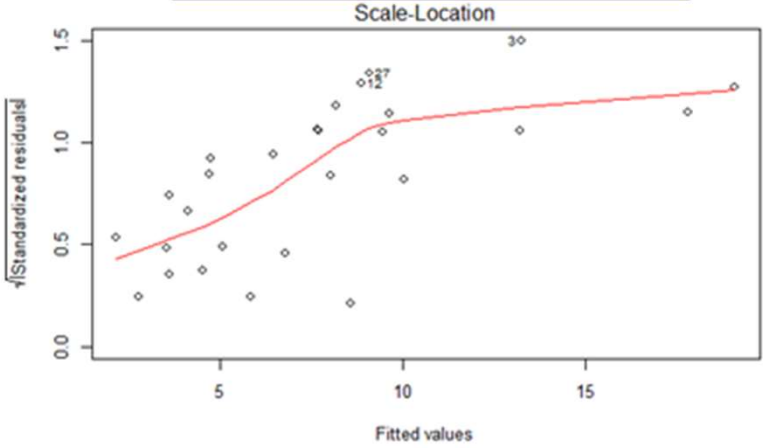
Is a wrong model fitted (linear or quadratic, etc.)?



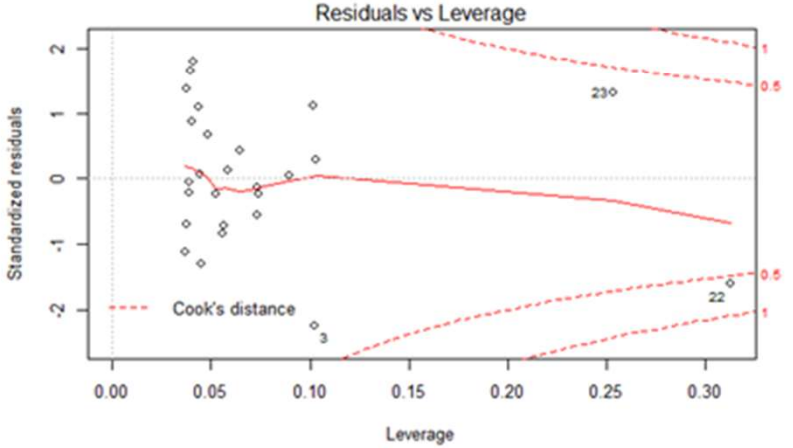
Are the residuals normally distributed?



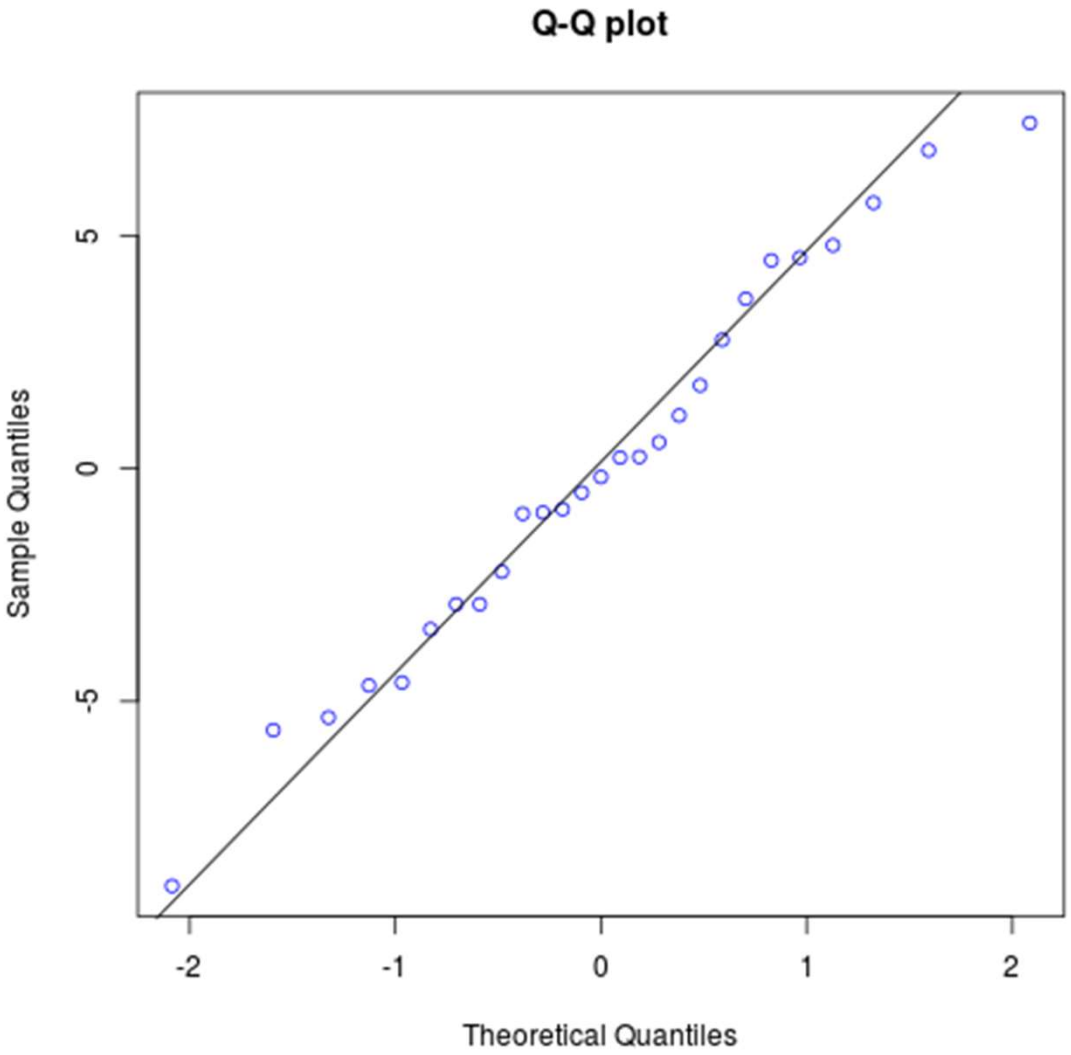
Is the data homoscedastic?



Are there influential outliers?



Q-Q plot for the Big Mac example dataset



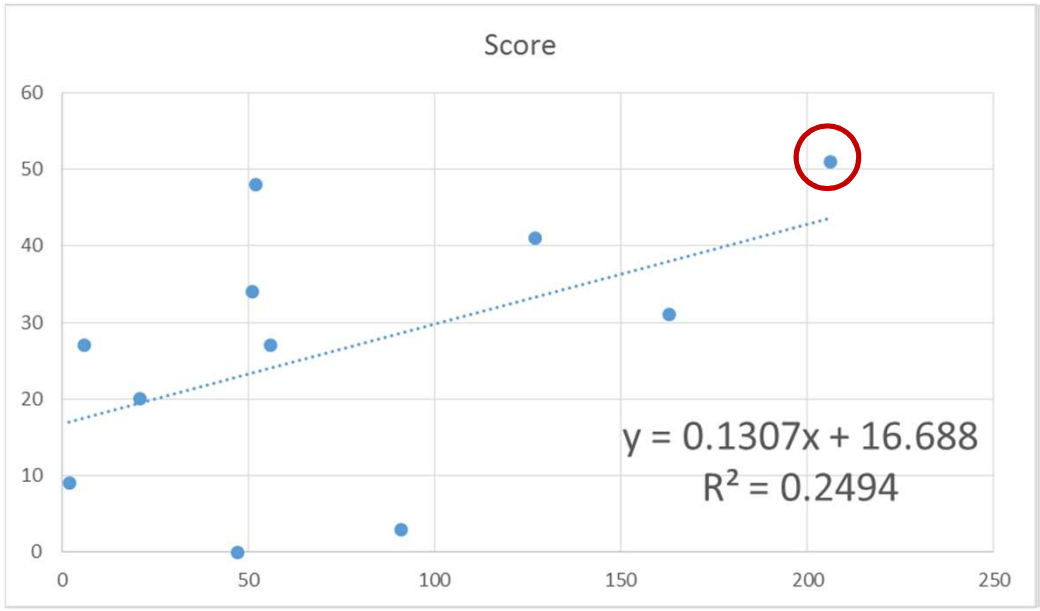
Does the Q-Q plot show the residuals to be approximately normal?

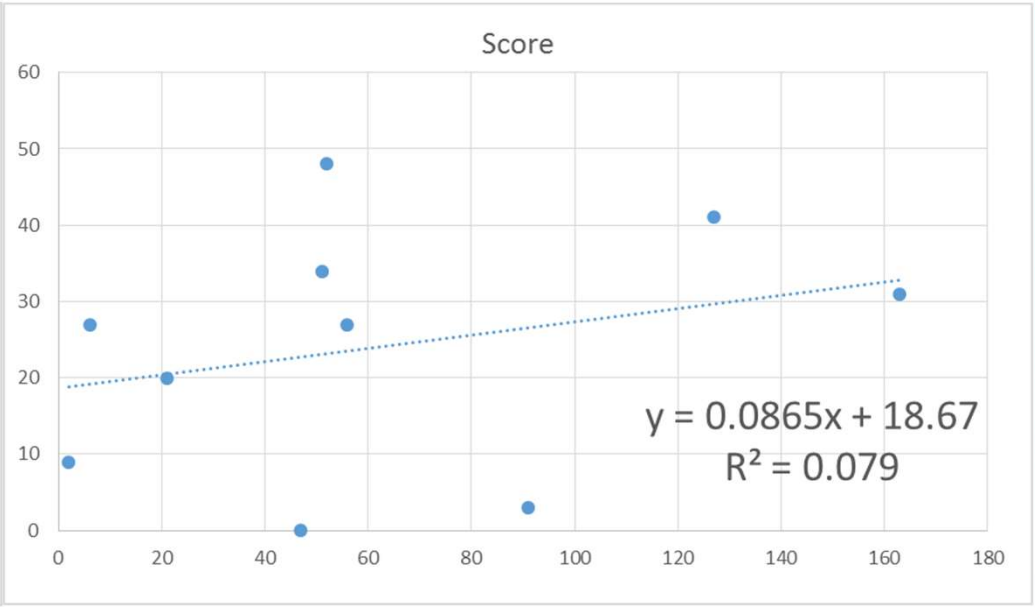
Influential observations

R-Squared, Significance and Residuals - Caution

Why it is important to plot.

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





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

Influential observations

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

Influence generally measured by

- **Leverage :**
  - Calculated only from the independent variables.
  - A data point has **high leverage** if it has "extreme" predictor x values.

- **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.
- That is, 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.**

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

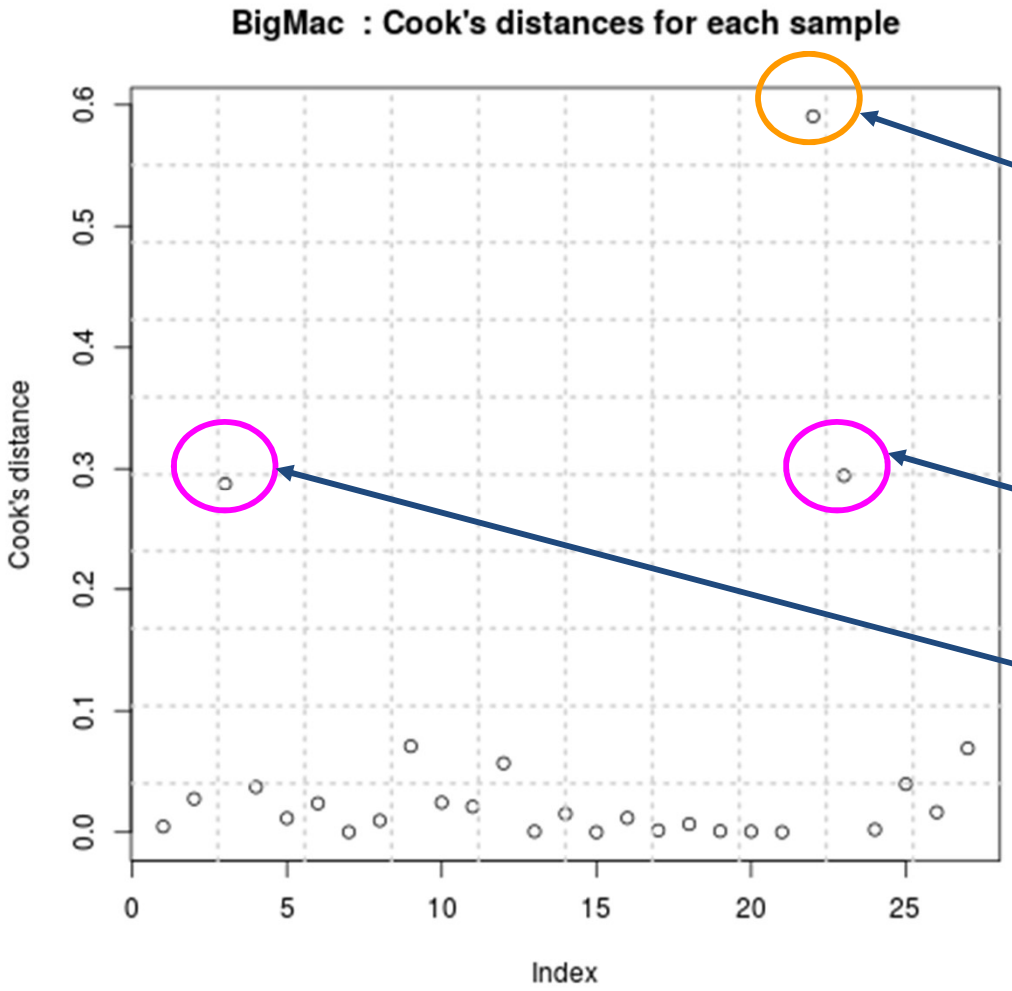
- An observation i can be considered as having too much influence if its Cooks distance (D_i) > 1.
  - Investigate observations with Cooks distances > 0.5 also.

- Relative size interpretation :

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

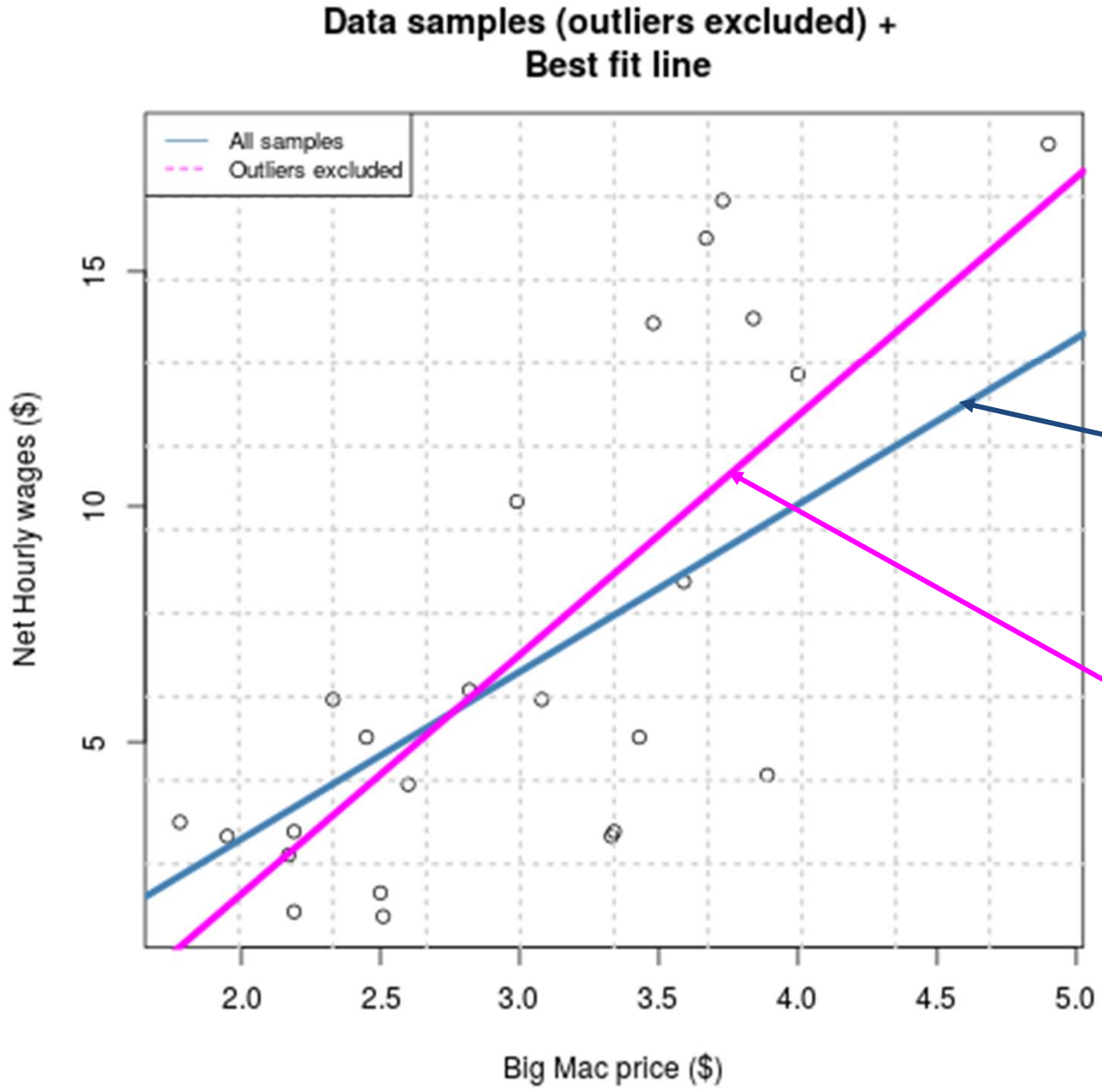
Identifying influential points using Cook's distances

Big Mac example : Cook's distances for each sample



Cook's distance > 0.5

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



All data samples included:
Equation of the best fit line

NetHourlyWage = BigMacPrice(3.5474)-4.1540

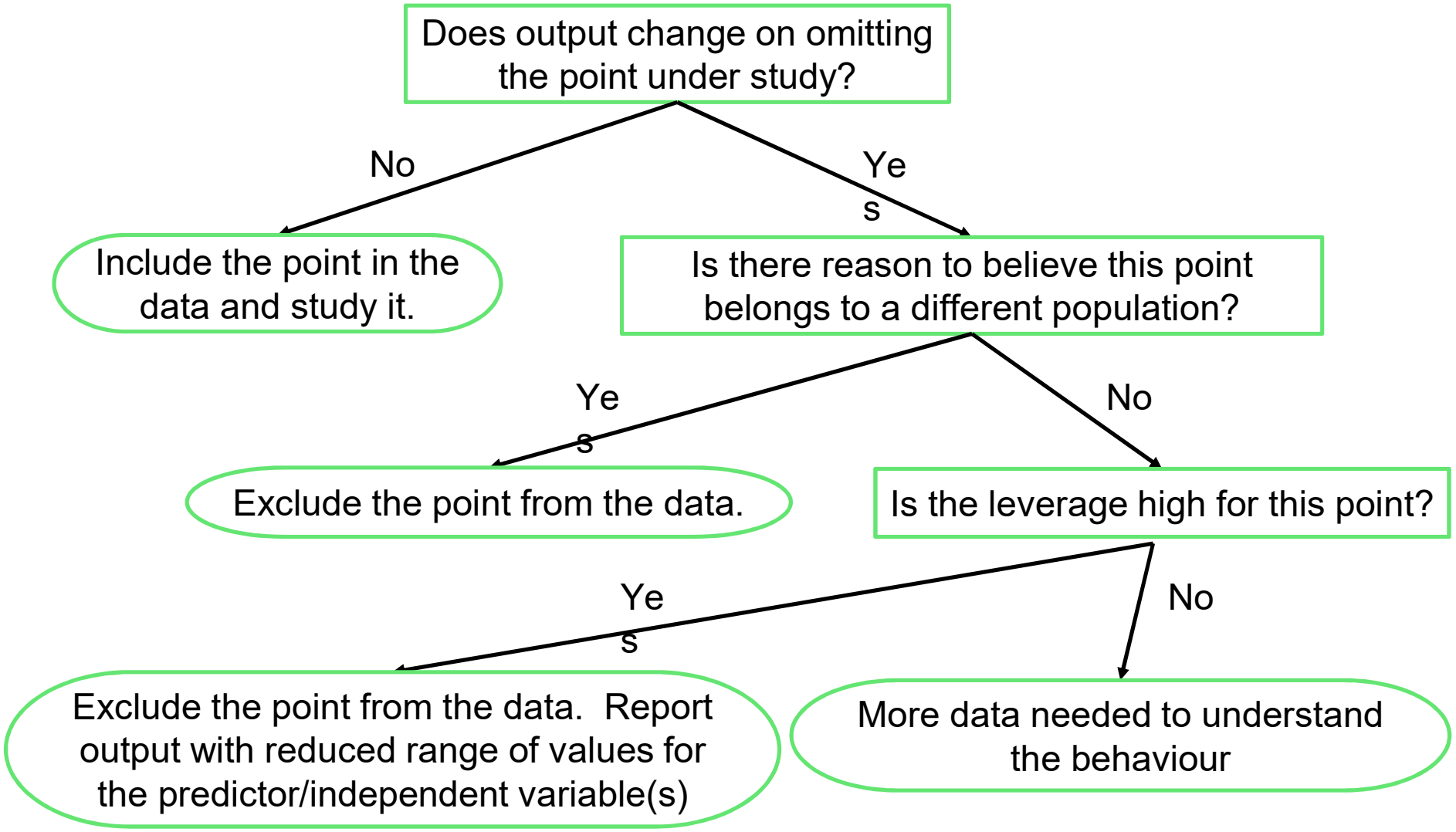
R^2=0.5142, adjusted R^2 = 0.4947

Influential points excluded:
Best fit line

NetHourlyWage = BigMacPrice(5.0745)-8.3760

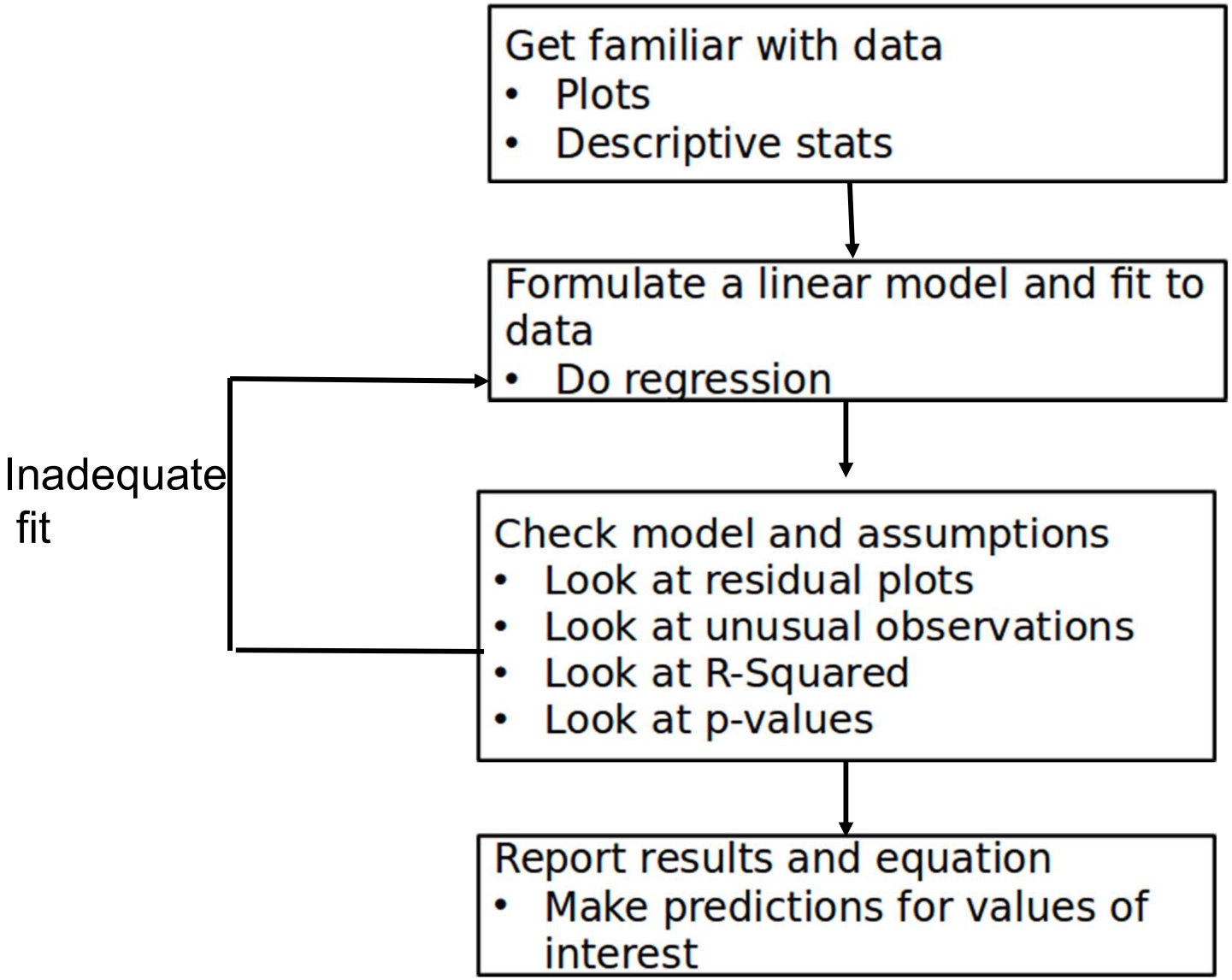
R^2=0.5714, adjusted R^2=0.552

Handling Influential Observations



Outline of major steps in building a linear regression model

Simple Linear regression : Typical flow



Linear regression : Outline of steps

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

- R : Eg. 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: $\beta 1 = 0$ versus HA: $\beta 1 \neq 0$.

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.

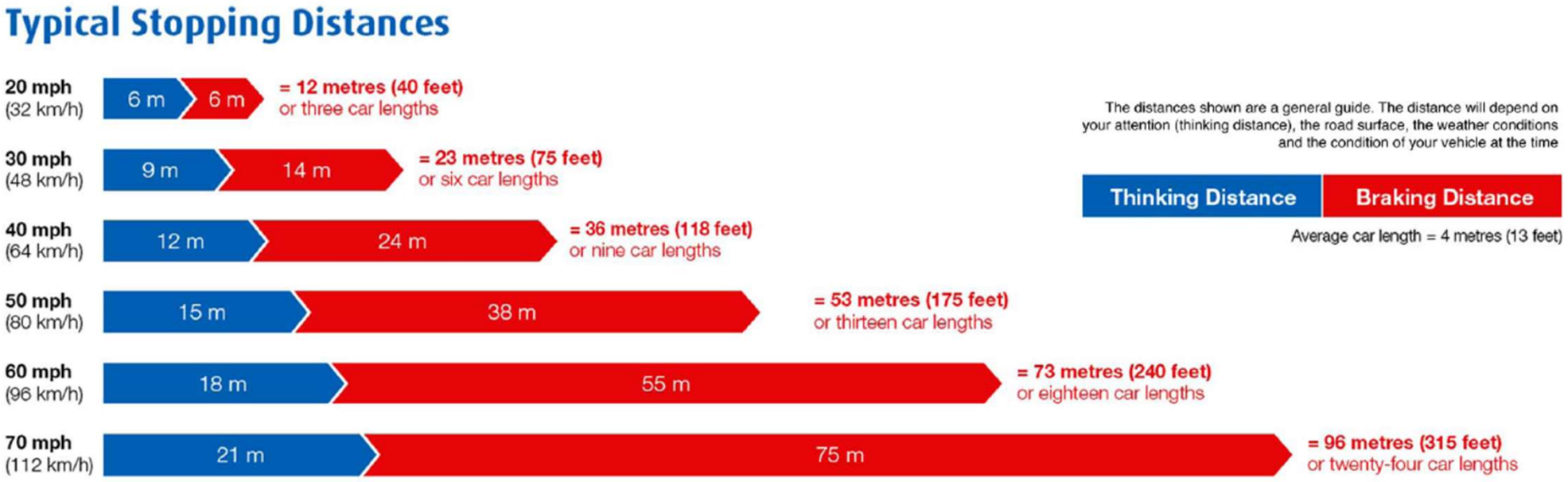
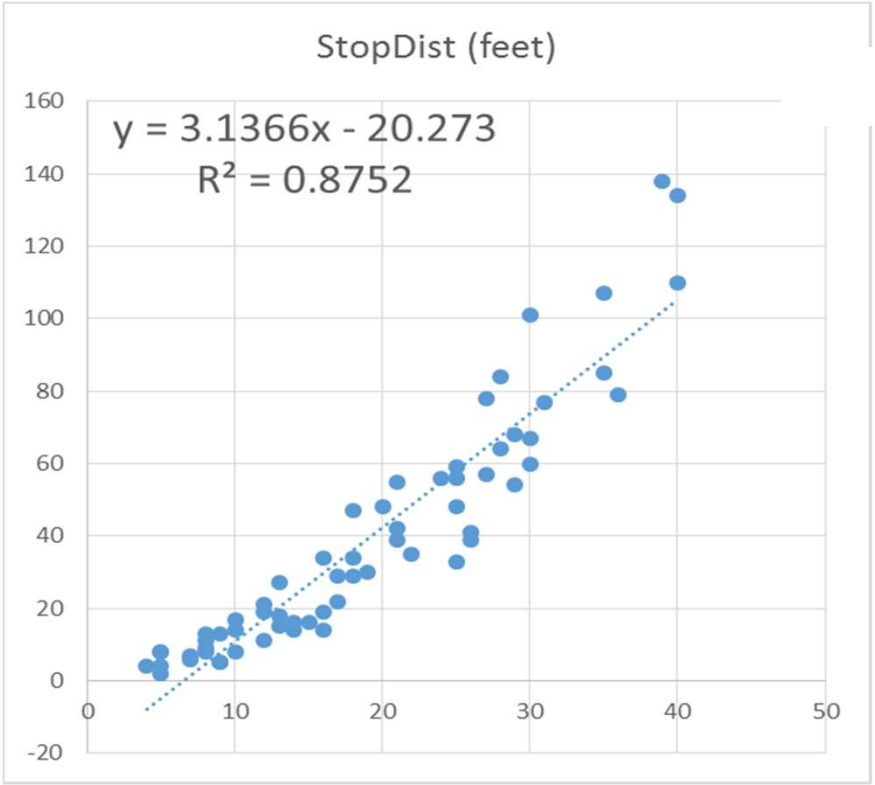


Image Source: http://streets.mn/2015/04/02/the-critical-ten/
Last accessed: November 20, 2015

R-Squared, Significance and Residuals - Caution

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

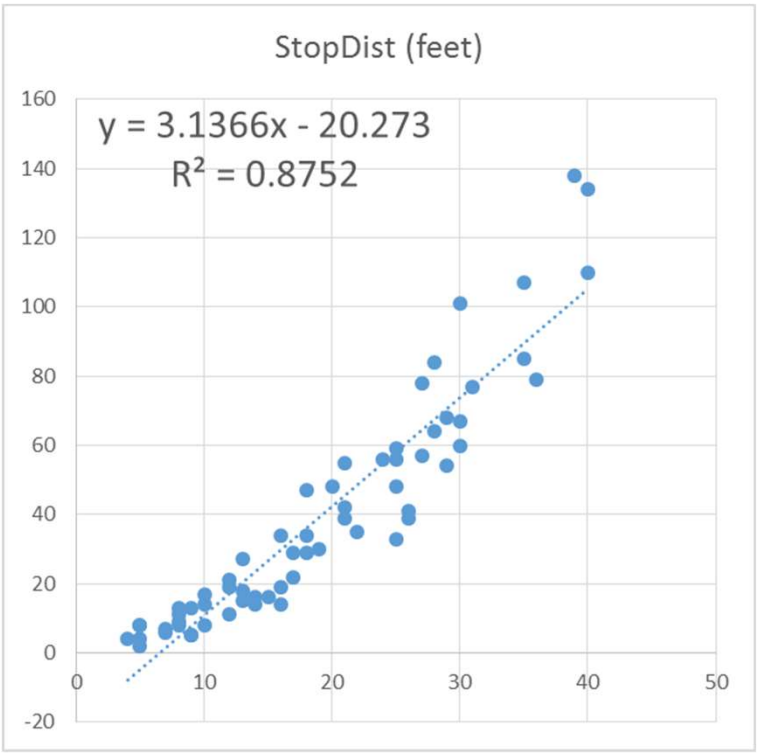
Does the estimated regression line fit the data well?

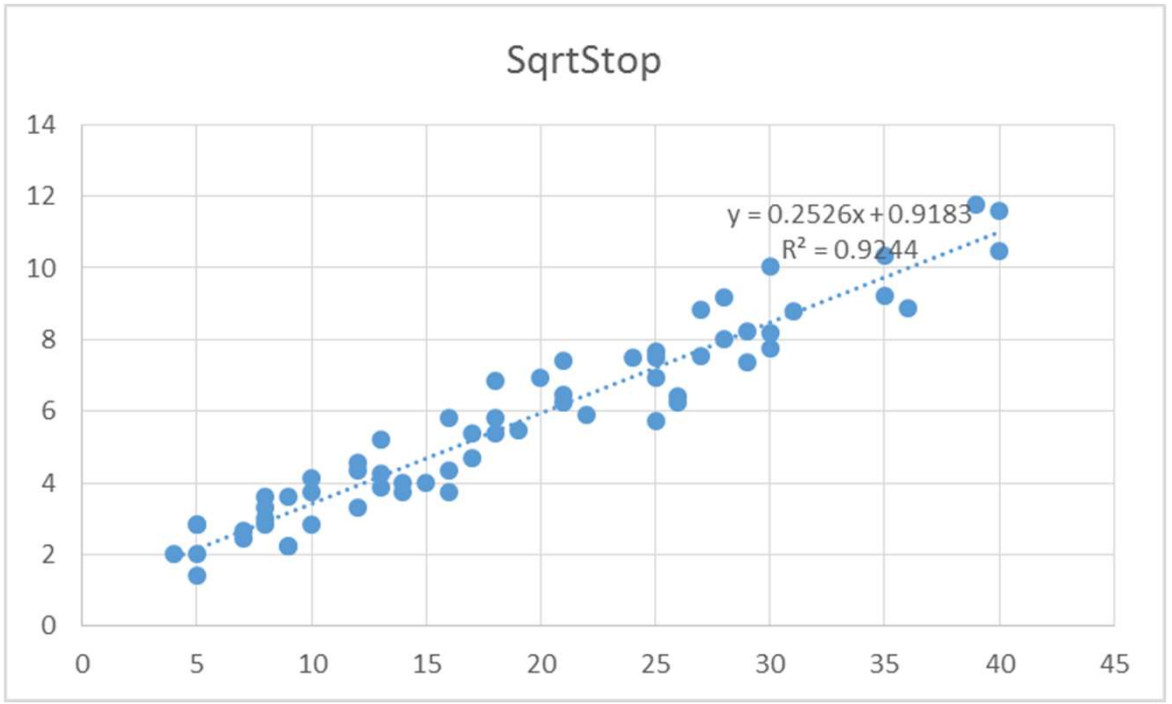


R-Squared, Significance and Residuals - Caution

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

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



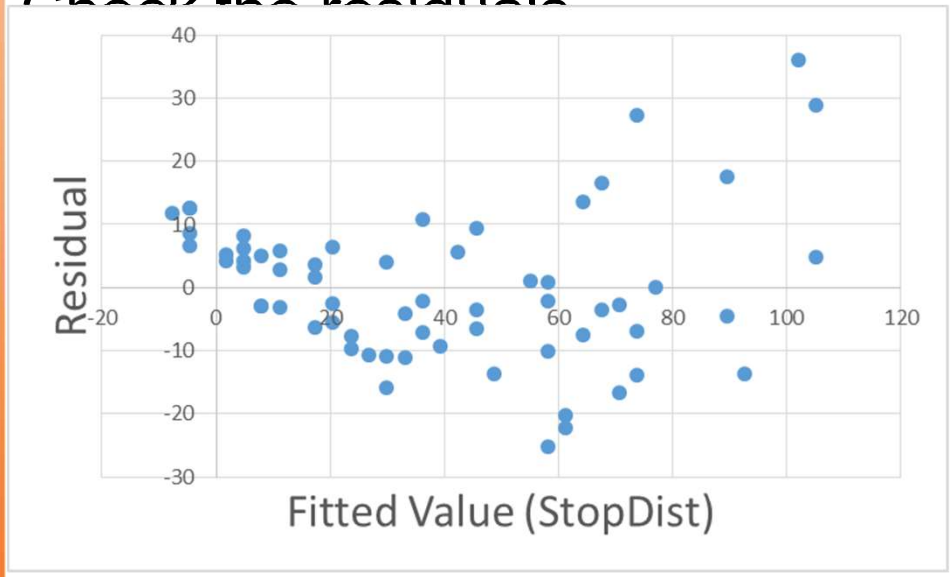


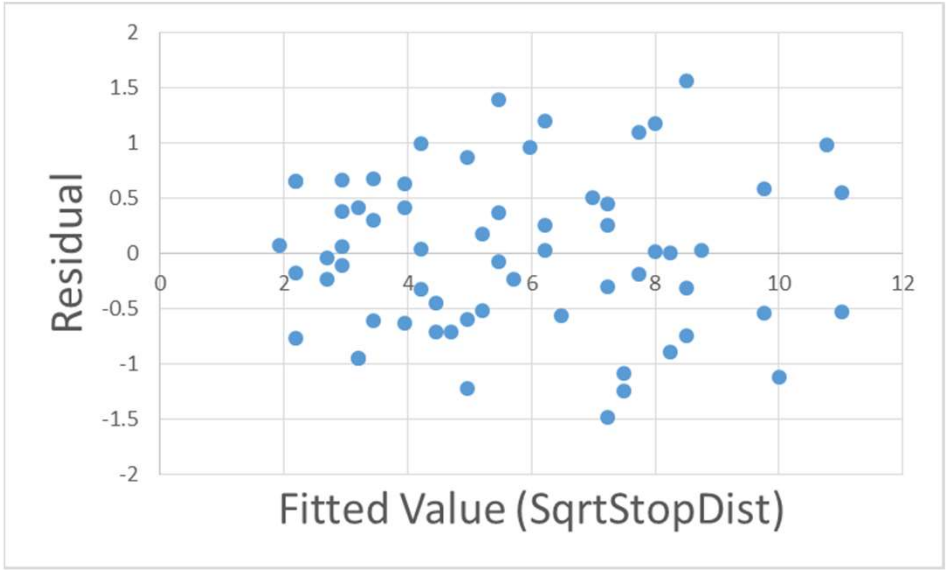
R-Squared, Significance and Residuals - Caution

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

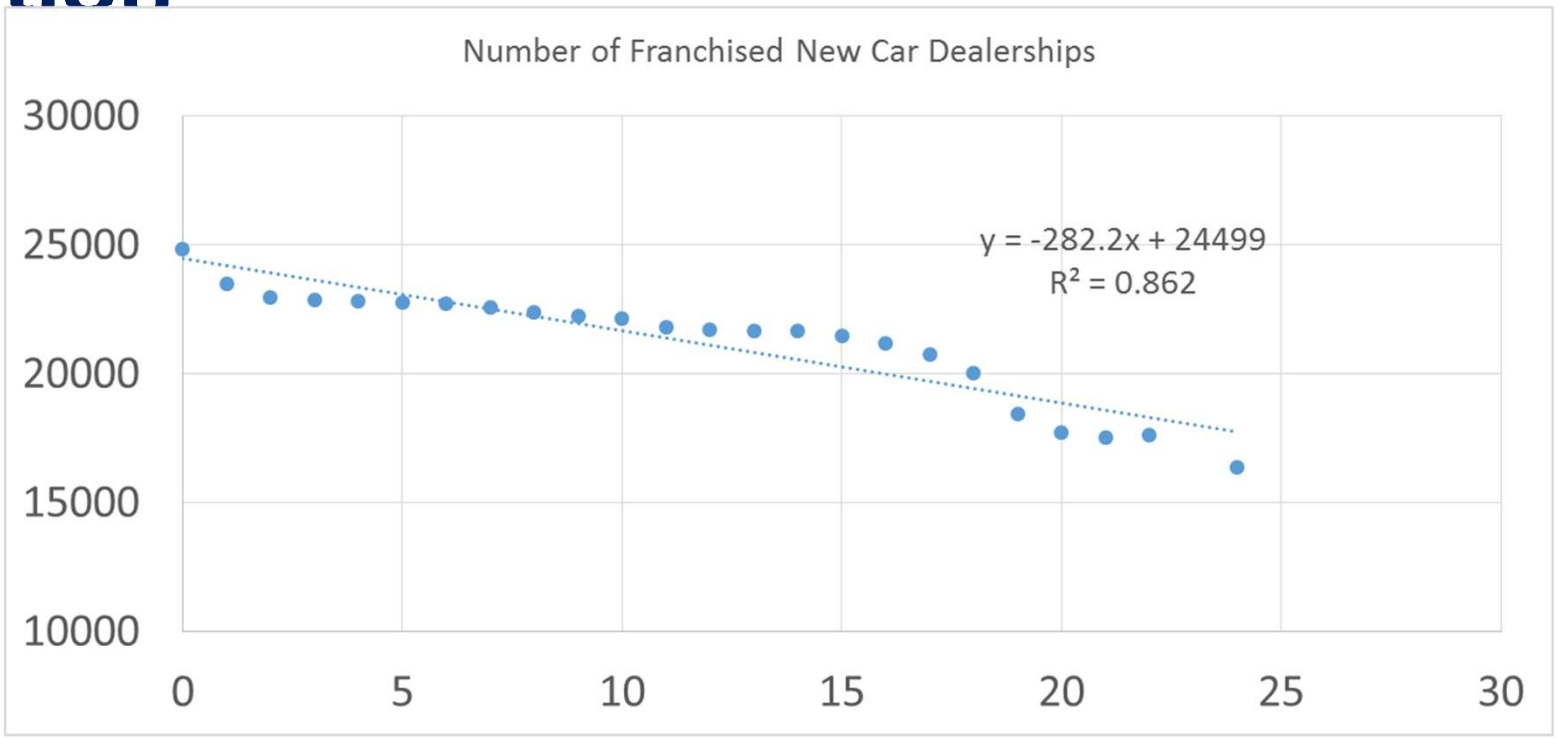




R-Squared, Significance and Residuals - Caution

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.

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

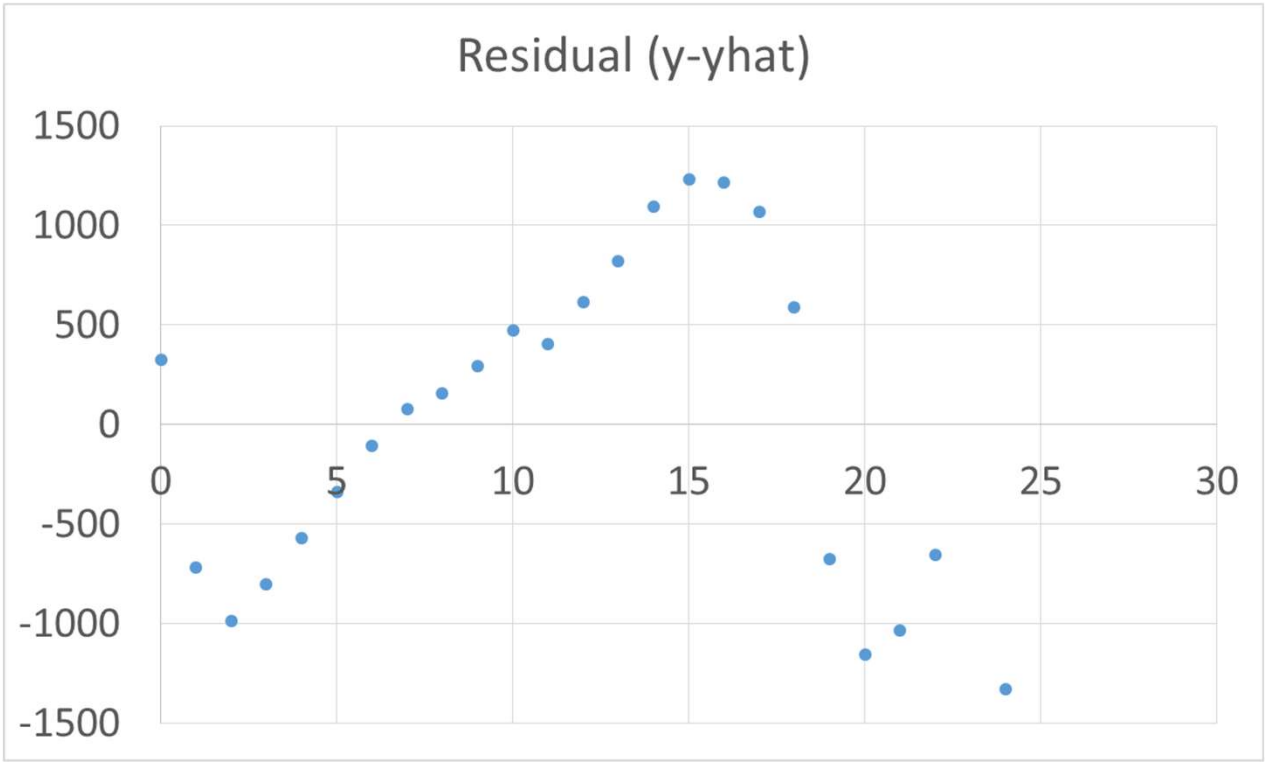
ANOVA

	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

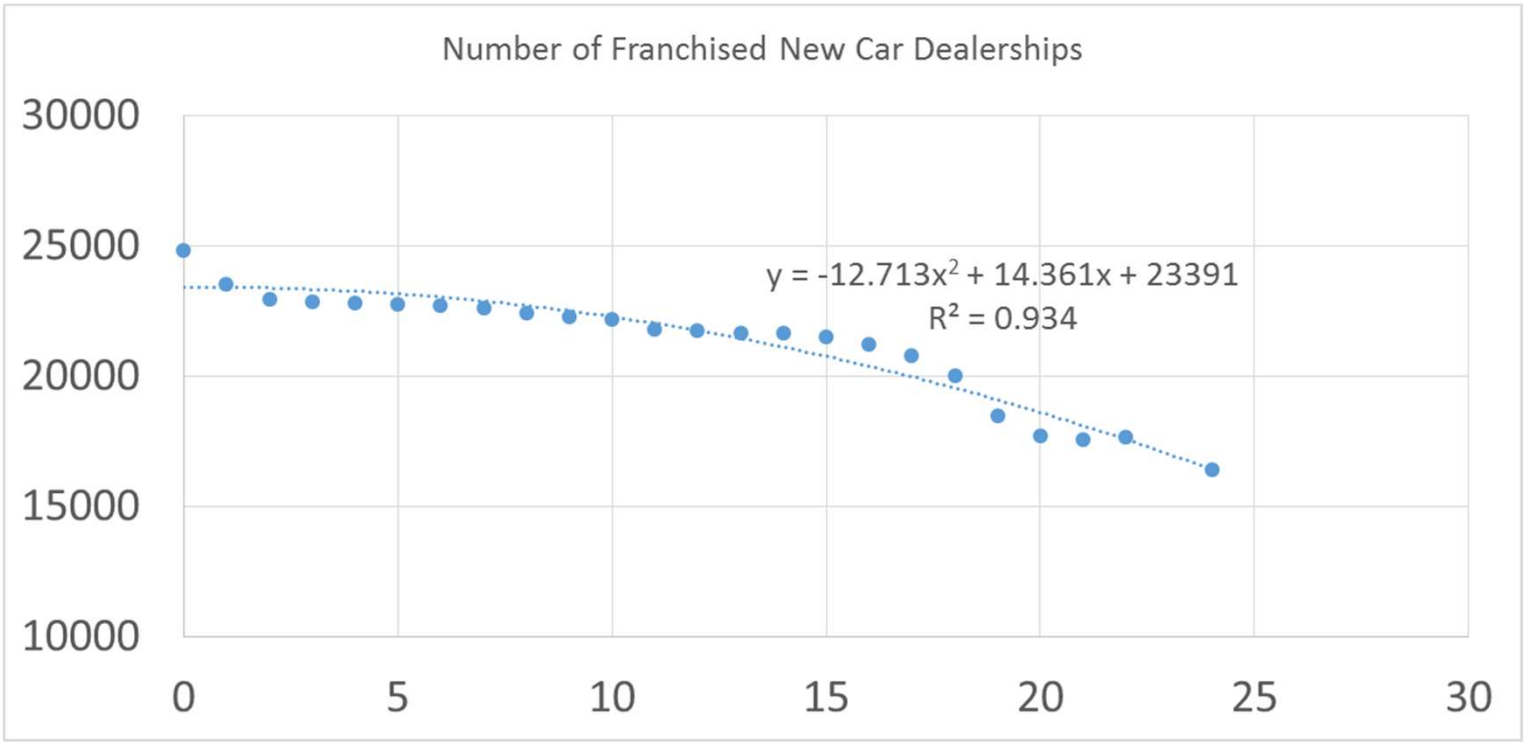
- Is the slope significant?
- Is the model significant?

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



Fixing Non-normality and Heteroscedasticity

Transformation of data can help correct normality and unequal variances problems

Data Transformations

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

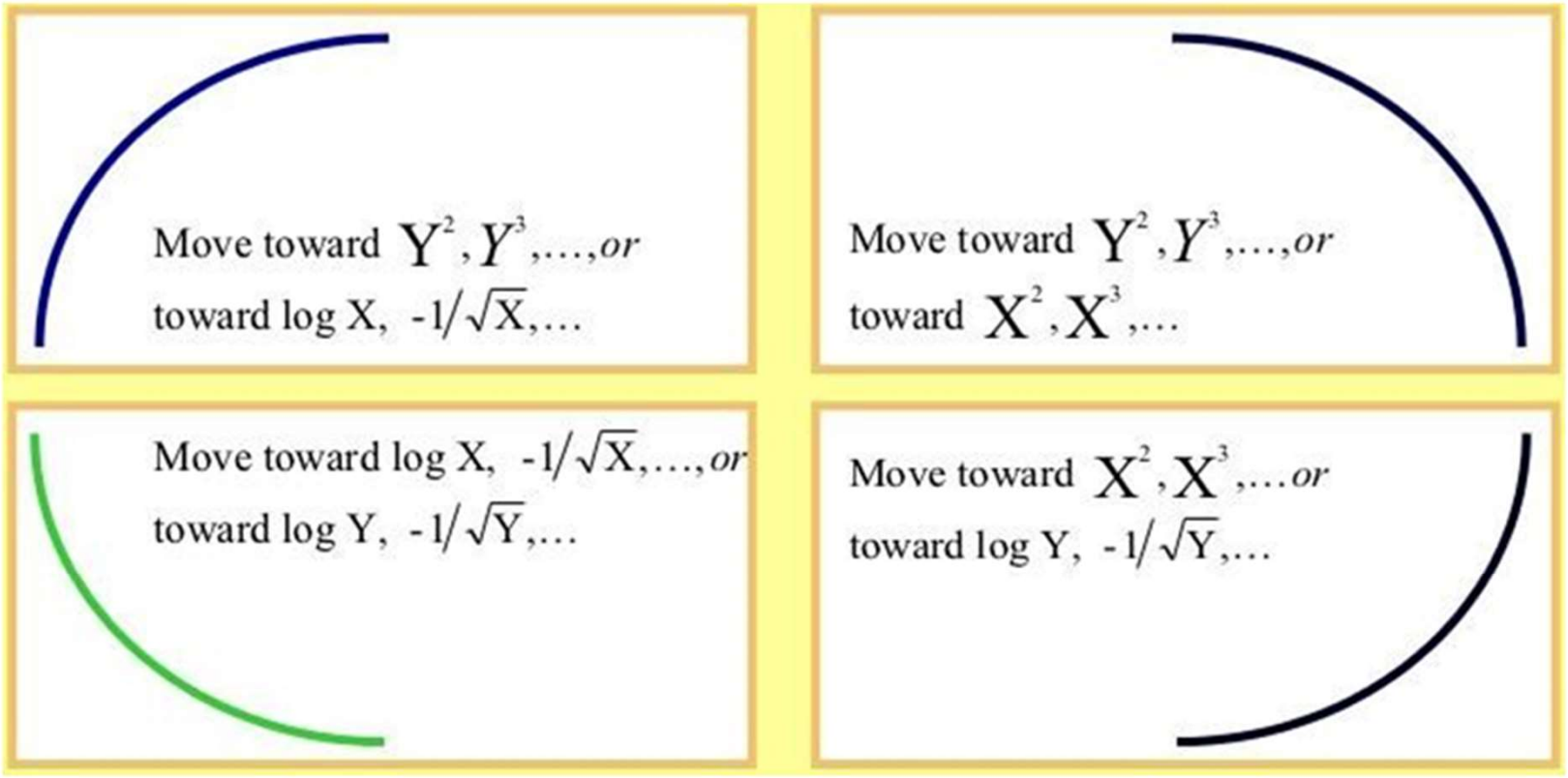
Problem diagnosed	Recommended transform
Primary problem with the model is **non-linearity.**	**Single predictor :** Look at a scatter plot of the data **Multiple predictors :** Look at residual plots to suggest transformations that might help.
If the variances are unequal and/or error terms are not normal,	**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

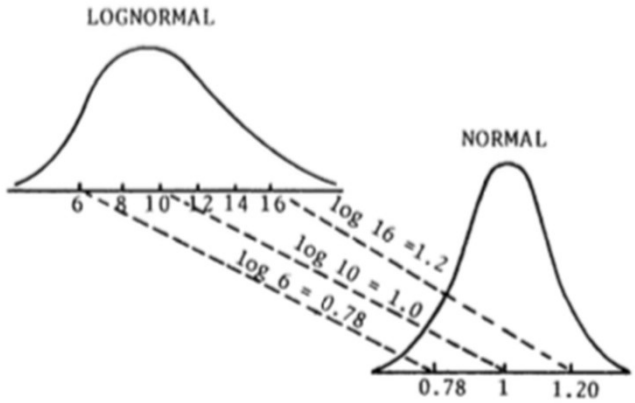


More thoughts on Transformations

DATA TRANSFORMATION

As suggested by Tabachnick and Fidell (2007) and Howell (2007), the following guidelines (including SPSS compute commands) should be used when transforming data.

If your data distribution is…	Use this transformation method.
Moderately positive skewness	Square-Root NEWX = SQRT(X)
Substantially positive skewness	Logarithmic (Log 10) NEWX = LG10(X)
Substantially positive skewness (with zero values)	Logarithmic (Log 10) NEWX = LG10(X + C)
Moderately negative skewness	Square-Root NEWX = SQRT(K – X)
Substantially negative skewness	Logarithmic (Log 10) NEWX = LG10(K – X)

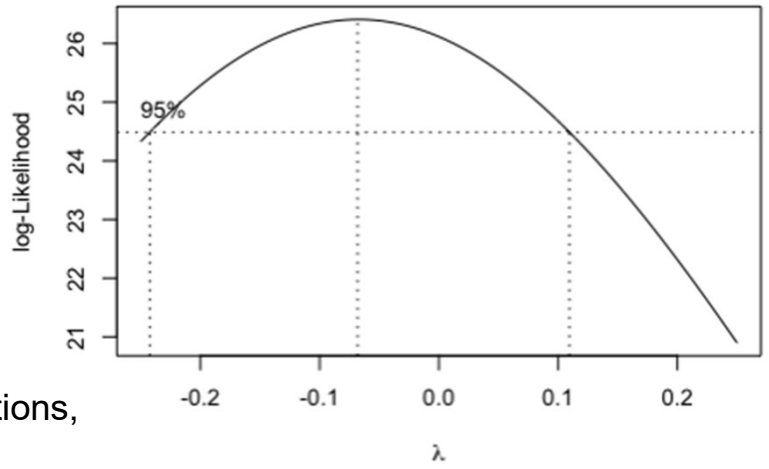


C = a constant added to each score so that the smallest score is 1.
K = a constant from which each score is subtracted so that the smallest score is 1; usually equal to the largest score + 1.

Source: http://oak.ucc.nau.edu/rh232/courses/eps625/handouts/data%20transformation%20handout.pdf
Last accessed: May 12, 2016

More thoughts on Transformations

- Square-root transformation: $X \rightarrow \sqrt{X}$
  - Use where variance is proportional to mean ($\sigma^2 \propto \mu$). Occurs when data consists of counts, such as in urine or blood analyses or microbiological data.
  - If some values are zero or very small, use instead $\sqrt{X} + \sqrt{X+1}$.
  - Poisson variables, where mean = variance, square-root transformation will lead to homoscedasticity.
- Reciprocal transformation: $X \rightarrow \frac{1}{X}$
  - Use where standard deviation is proportional to the square of the mean ($\sigma \propto \mu^2$).
- boxcox() in MASS package of R
- PROC TRANSREG in SAS



Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society*, Series B, *26*, 211-252.

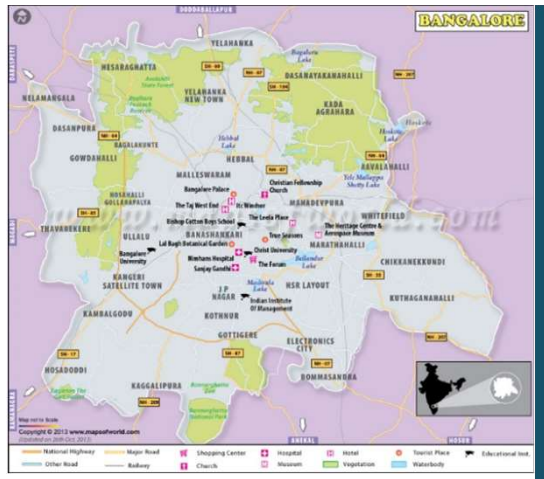
Approach to determine whether to transform X or Y to achieve **linearity**, **homoscedasticity** and **normality**:

1. Often, a transformation that fixes one, fixes all.
2. In general, transforming both is not required, although sometimes it is.
3. A general rule of thumb:
    1. Transform Y first to remove heteroscedasticity.
    2. Then transform X to remove non-linearity.

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.

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.