Inspire…Educate…Transform.

Supervised Models

Linear Regression

Dr. Anand Jayaraman

Apr 22, 2017

Thanks to Dr.Sridhar Pappu for the material

Why Model Building?

- In any business, there are some easy-to-measure metrics
  - Age; Gender; Income; Education level; etc.

  and a difficult-to-measure metric
  - Amount of loan to give; Will she buy or not; How many days a patient will stay in the hospital; etc.
- Regression enables you to compute the latter from the former

Simplest Learning Models

- Linear regression: Measuring the relation between two or more analog variables (class variable is numeric)
- Logistic regression: A classification model (class variable is categorical)

SIMPLE LINEAR REGRESSION

Speed vs Stopping distance

Data: cars

File

	speed	dist
10	11	17
11	11	28
12	12	14
13	12	20
14	12	24
15	12	28
16	13	26
17	13	34
18	13	34
19	13	46
20	14	26
21	14	36
22	14	60
23	14	80
24	15	20
25	15	26
26	15	54
27	16	32
28	16	40
29	17	32
30	17	40
31	17	50
32	18	42
33	18	56
34	18	76

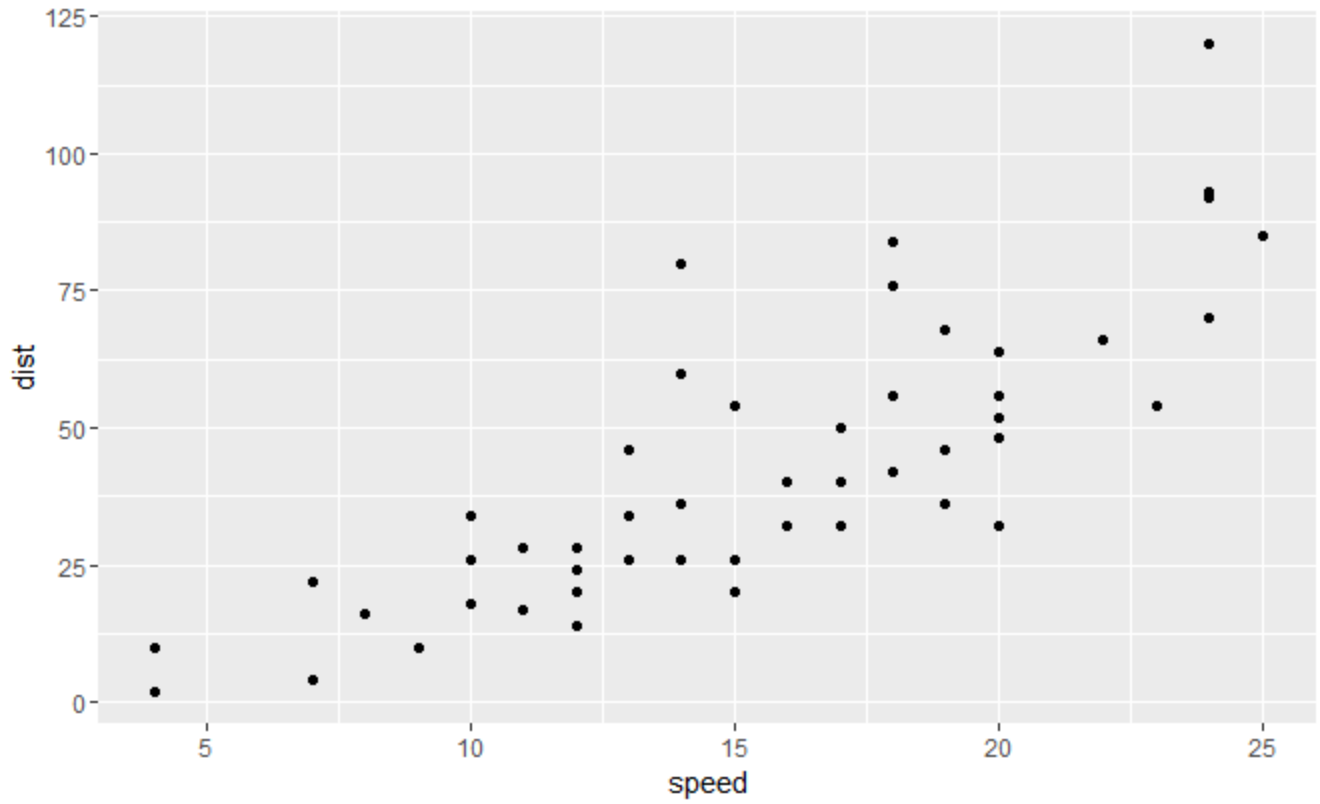
The “cars” dataset in R contains 50 pairs of datapoints
for Speed(mph) vs stopping distance(ft), that were collected in 1920

```
> View(cars)
>
```

See: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/cars.html

CSE 7202c

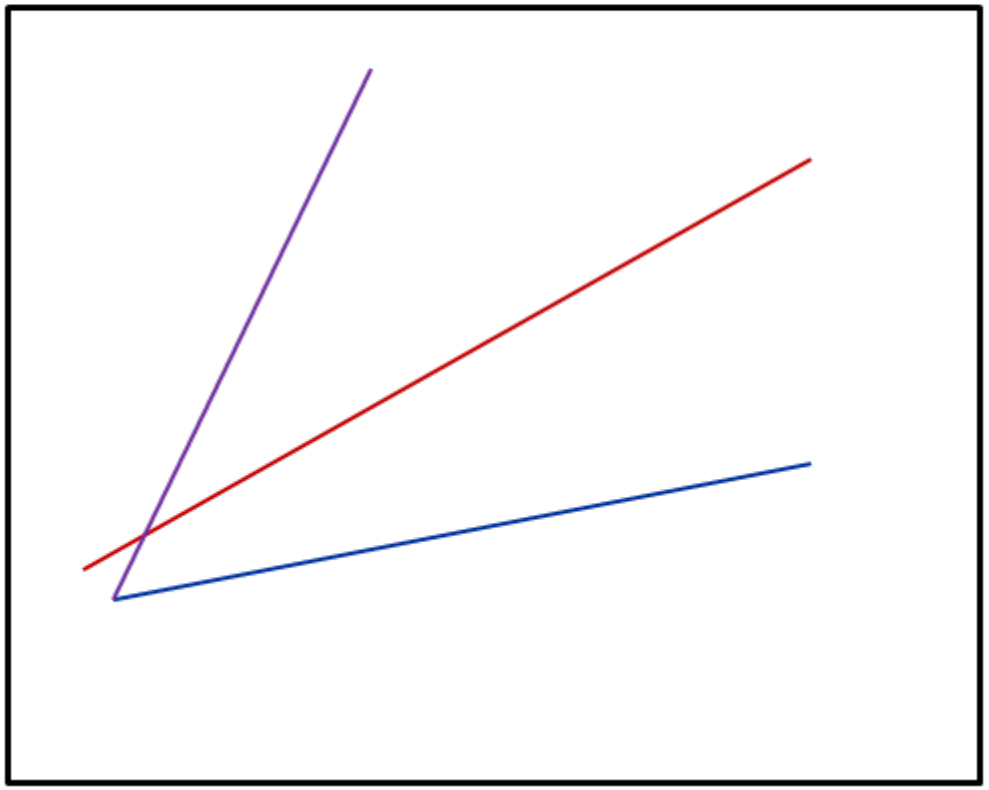
- Independent variable (explanatory) – Speed (mph) – Plotted on X-axis
- Dependent variable (response) – Stopping distance(ft) – Plotted on Y-axis



- Another car with the same speed, might not have the same stopping distance
- x is known (No uncertainty)
- y has uncertainty (It's a sample picked from some unknown distribution)

Start with a Function/Hypothesis with Some Parameters

$y=\beta 0+\beta 1x$ (Deterministic model)



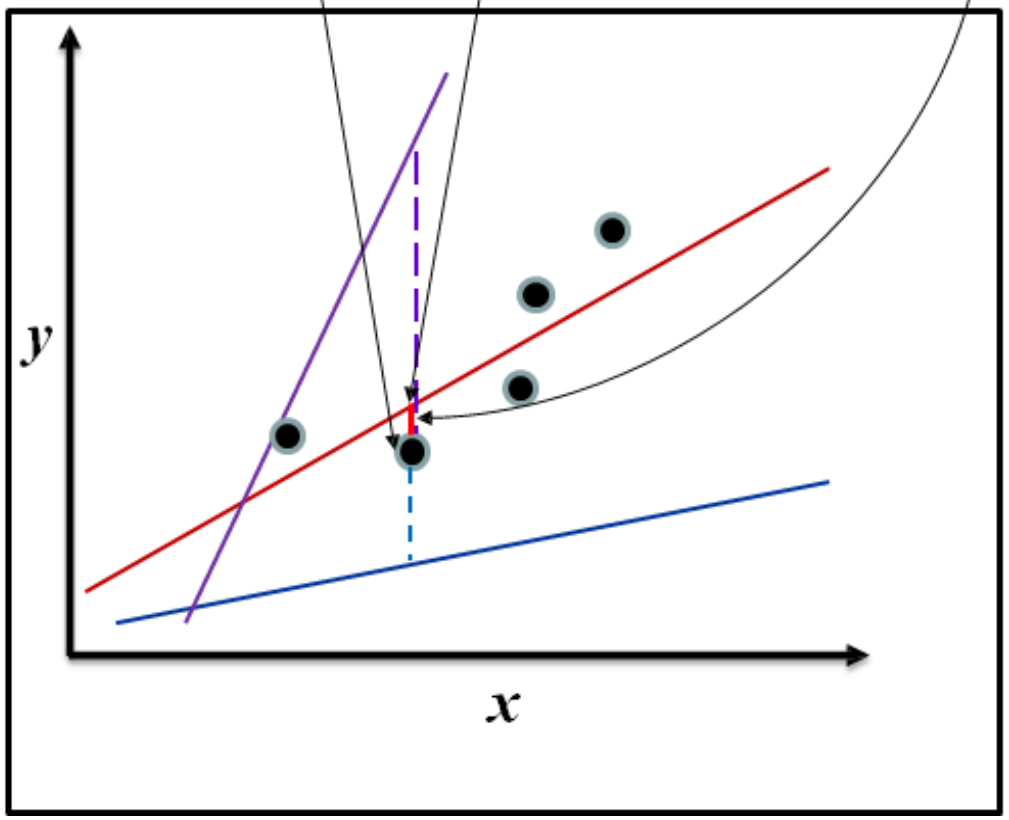
How to Pick the Best Model?

$y=\beta 0+\beta 1x+\varepsilon$

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

(Probabilistic model)

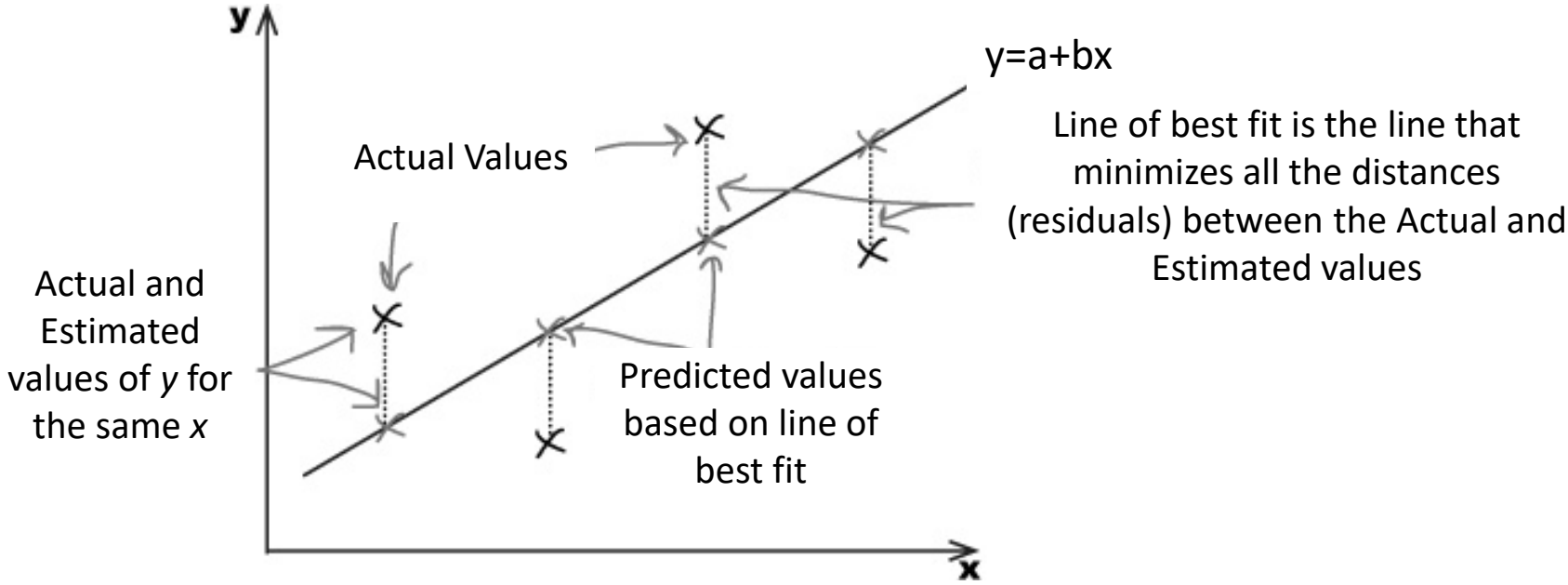
Recall: Conditional Expected Value…Conditional Expectation of a Random Variable…Conditional Mean of a Random Variable



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

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

We need to minimize errors.



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

We need to minimize errors.

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

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

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

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

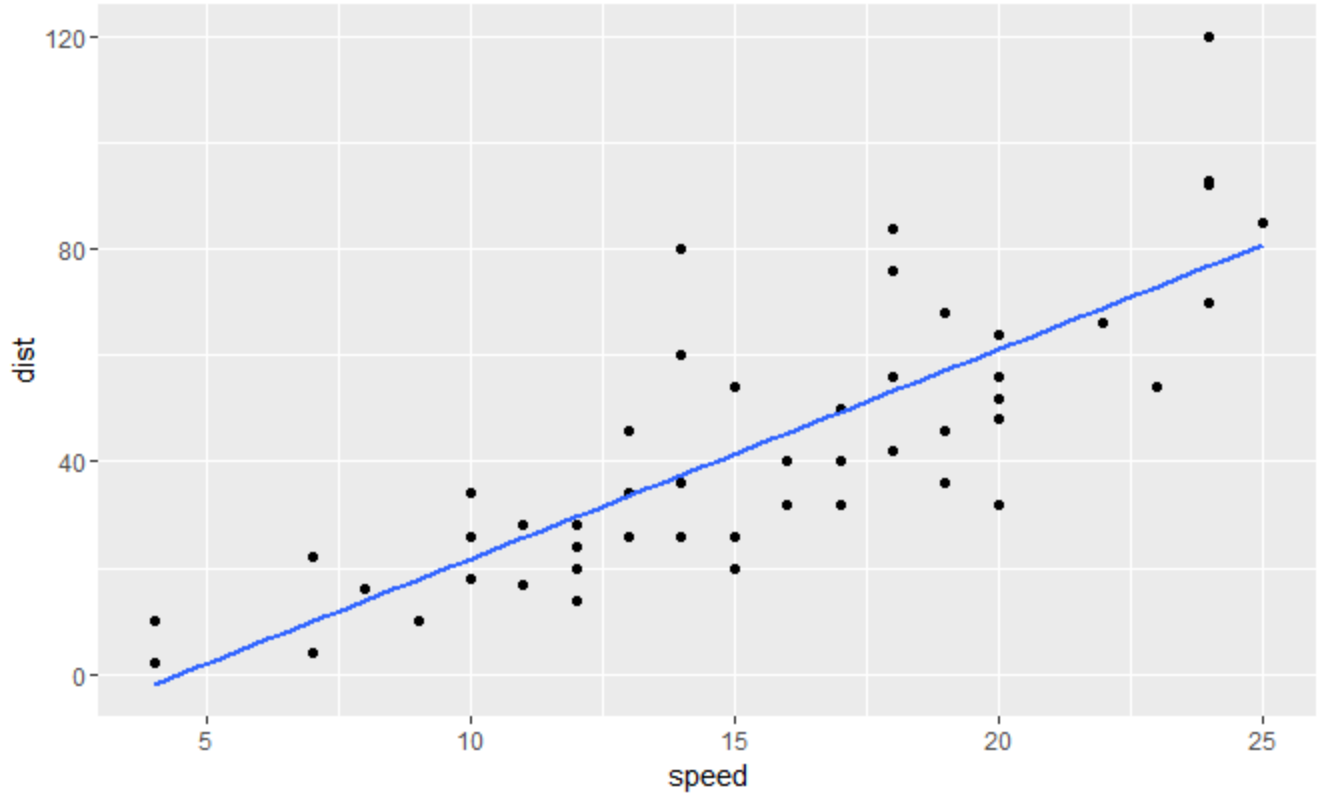
The value of *b*, the slope, that minimizes the SSE is given by

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

The intercept *a* can be computed by requiring the fitted line pass through $(\bar{x}, \bar{y})$. Substituting in the equation $y = a + bx$, we can find *a*.

This method of fitting the line of best fit is called **least squares regression**.

Speed vs Stopping distance



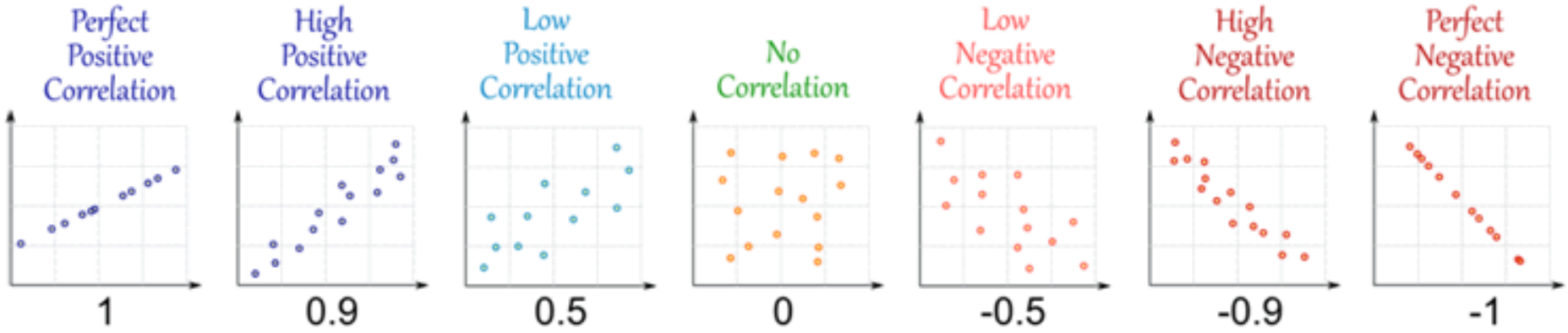
$y = 3.93\,x - 17.58$

```
> lmcars <- lm(dist~speed, data=cars)
> summary(lmcars)
```

Correlation Coefficient

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

$$r = \frac{bs_x}{s_y}$$



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

Img source: https://www.mathsisfun.com/data/correlation.html
Access date: 8/1/2017

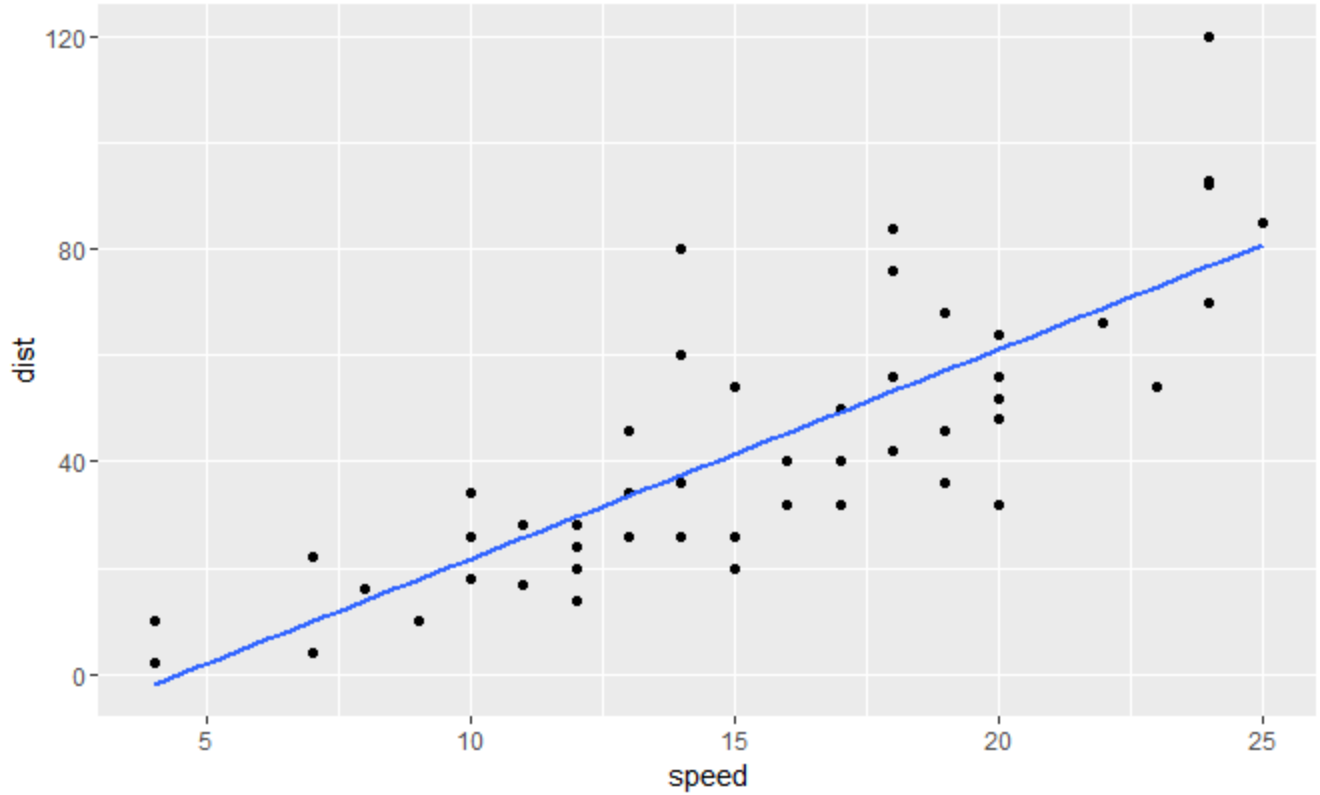
Correlation Coefficient

$$r = \frac{bs_x}{s_y}$$

where *b* is the slope of the line of best fit, s_x is the standard deviation of the *x* values in the sample, and s_y is the standard deviation of the *y* values in the sample.

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

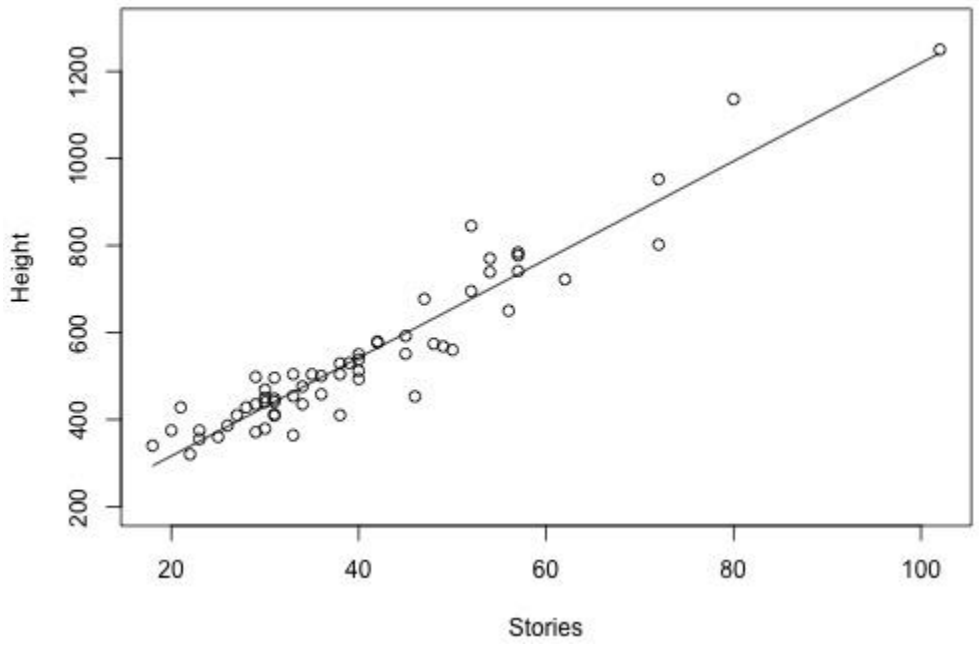
Correlation: Speed vs Stopping distance



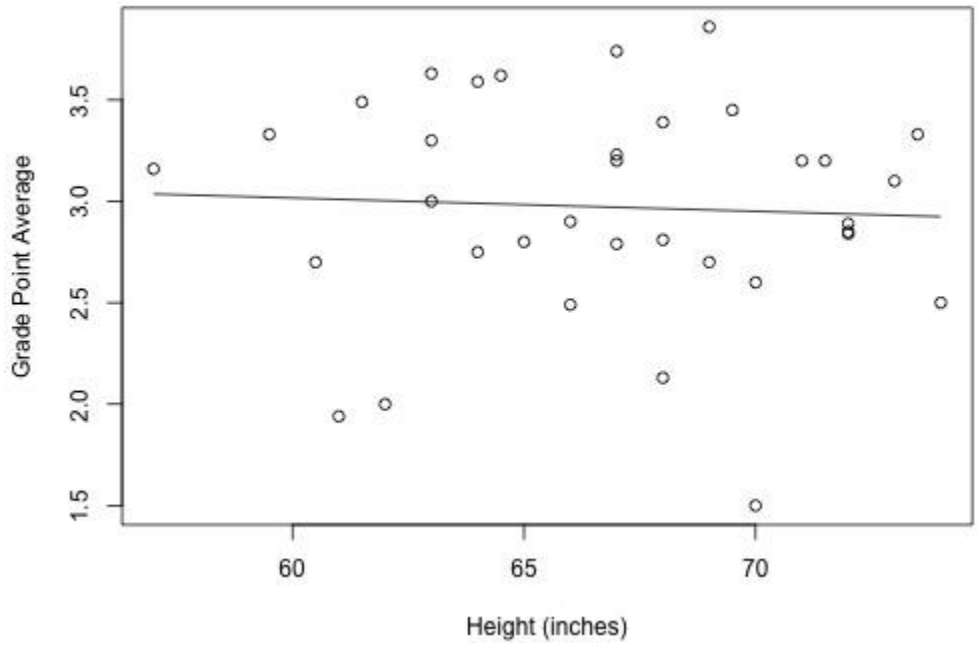
$y = 3.93\,x - 17.58$

```
> cor(cars$dist,cars$speed)
[1] 0.8068949
```

Correlation



r=0.951



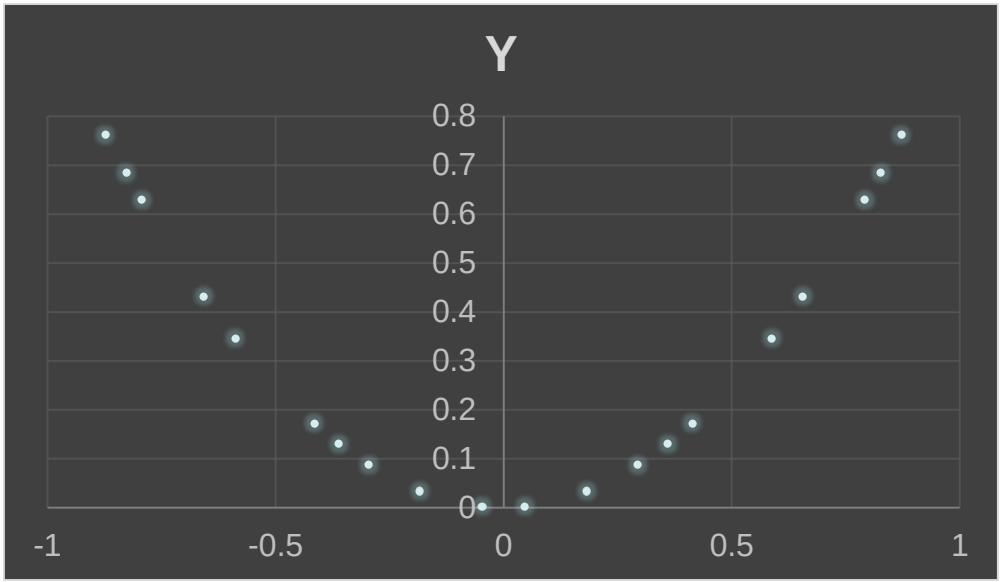
r=-0.053

Source: https://onlinecourses.science.psu.edu/stat501/node/257

Correlation

Correlation coefficient measures the strength of **linear relationship** between x & y

X	Y
-0.183	0.0336
0.6564	0.4309
0.8725	0.7613
0.3611	0.1304
0.7926	0.6282
0.1833	0.0336
-0.656	0.4309
-0.414	0.1715
-0.873	0.7613
0.827	0.6839
-0.588	0.3456
-0.295	0.0871
-0.361	0.1304
-0.827	0.6839
-0.047	0.0022
0.4141	0.1715
0.047	0.0022
0.295	0.0871
-0.793	0.6282
0.5879	0.3456



Correlation Coefficient $r = 0$

When there is no relationship between x & y, r will be close to zero.
However, $r = 0$ doesn't imply there is no relation between x & y.

Covariance

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

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

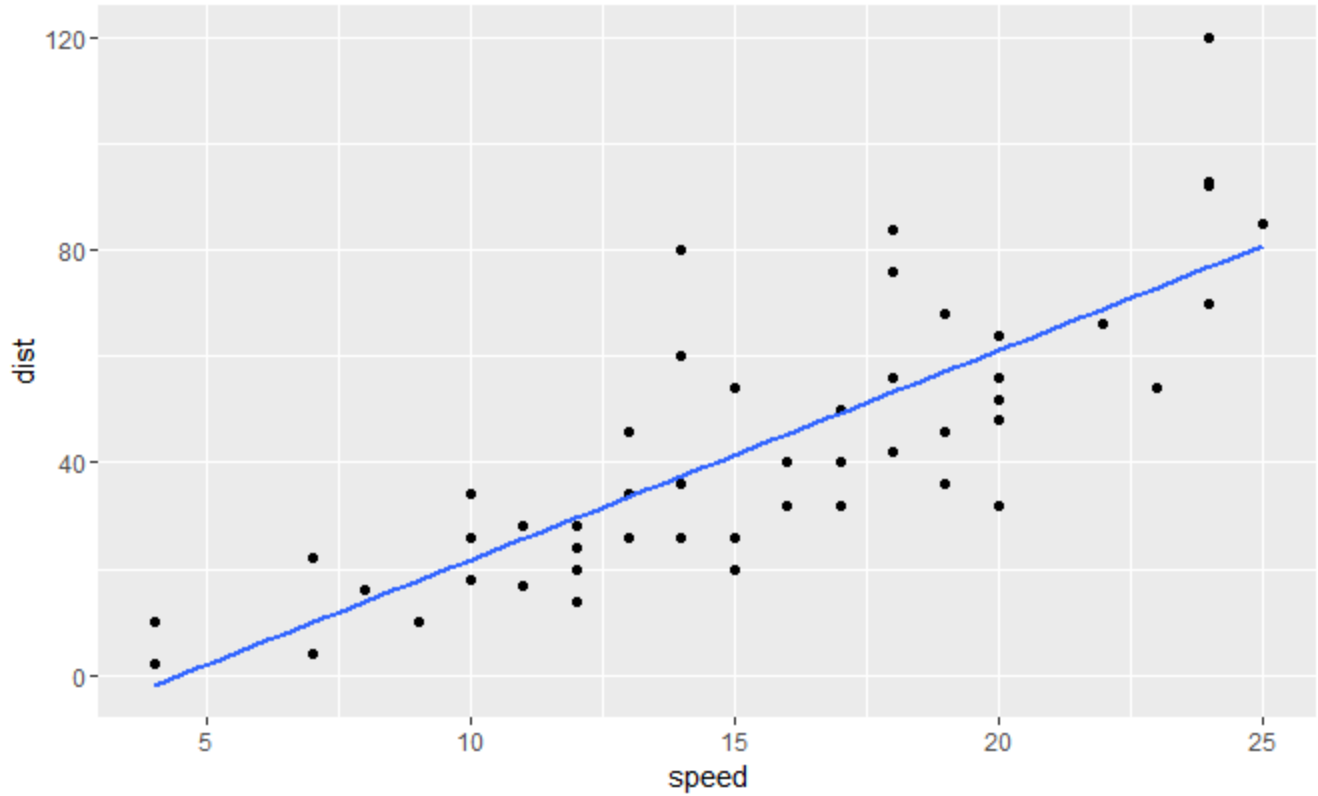
Covariance and Correlation

$$s_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{n - 1}$$

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

- The value of covariance itself doesn't say much. It only shows whether the variables are moving together (positive value) or opposite to each other (negative value).
- To know the strength of how the variables move together, covariance is standardized to the dimensionless quantity, correlation.

Covariance: Speed vs Stopping distance




```
> cor(cars$dist,cars$speed)    # Correlation
[1] 0.8068949
> cov(cars$dist,cars$speed)    # Covariance
[1] 109.9469
> cor(cars$dist,cars$speed)*sd(cars$dist)*sd(cars$speed)   # r*sd(x)sd(y)
[1] 109.9469
```

Coefficient of Determination

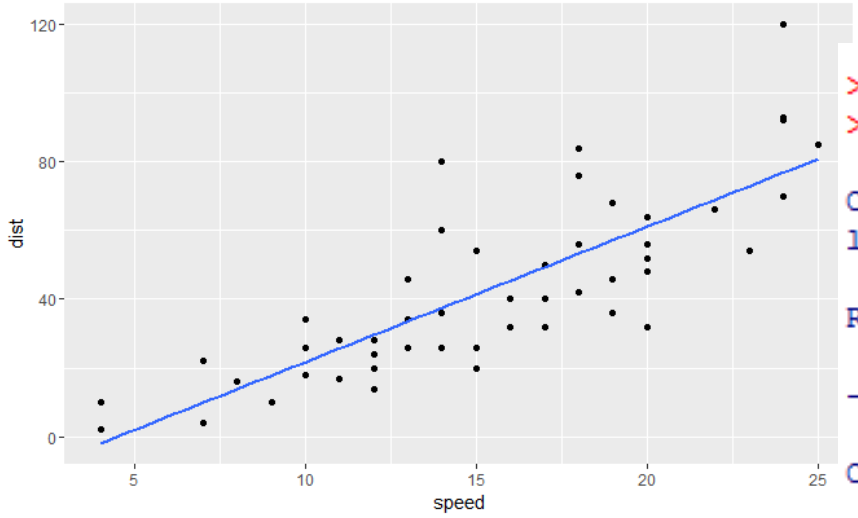
The coefficient of determination is given by r^2 or R^2. It is the percentage of variation in the *y* variable that is explainable by the *x* variable. For example, what percentage of the variation in open-air concert attendance is explainable by the number of hours of predicted sunshine.

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

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

Usually, r^2 is between these two extremes.

R^2: Speed vs Stopping distance




```
> lmcars <- lm(dist~speed, data=cars)
> summary(lmcars)

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

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

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
```

R^2 = 0.6511

```
> cor(cars$dist,cars$speed)^2    # Square of Correlation
[1] 0.6510794
```

THE BIG MAC INDEX

How many burgers you get for $50 USD?

India* $1.62

23

Ukraine $2.11
Hong Kong $2.12

Malaysia $2.34

20

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

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

Lithuania $2.87
Pakistan $2.89

16

Latvia $3.0

15

South Korea $3.19
UAE $3.27

14

Czech Rep. $3.45
Turkey $3.54

13

Peru $3.71
Singapore $3.75
Britain $3.82

12

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

11

USA $4.2
Euro area $4.43
Colombia $4.54

10

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

9

Denmark $5.37

8

Brazil $5.68
Sweden $5.91

7

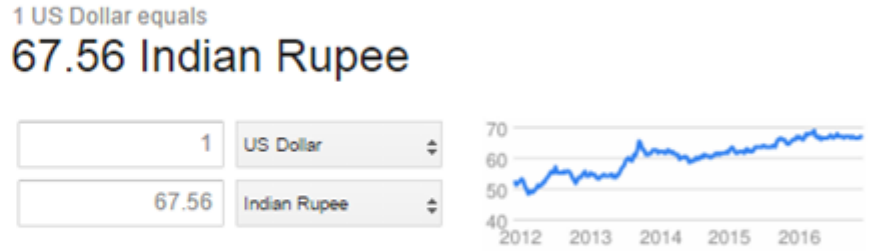
Norway $6.79
Switzerland $6.81

Source: The Economist (Jan 2012)

* Chicken burger

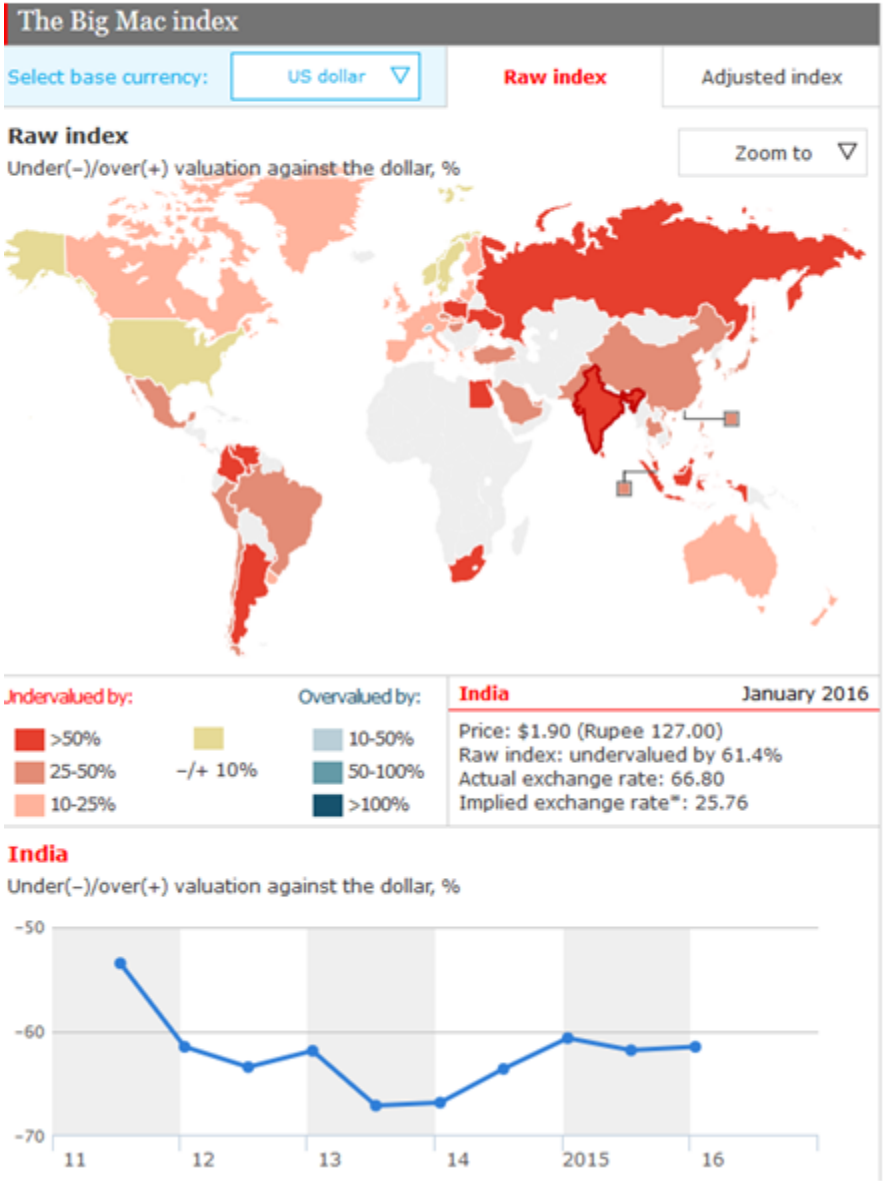
Burgernomics: Overvalued or Undervalued Currencies?

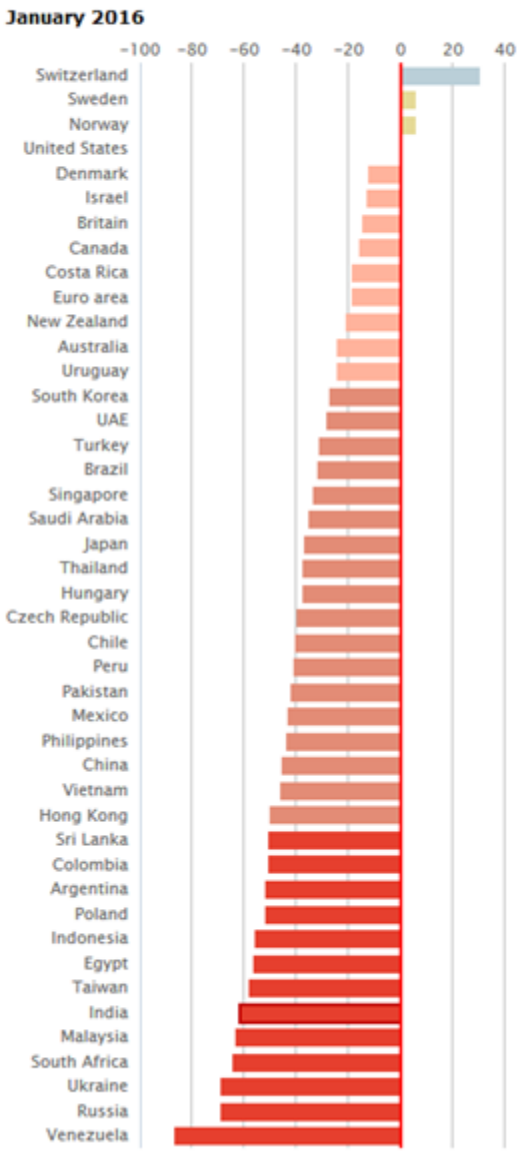
- Big Mac price in the US: $ 4.93
- Maharaja Mac price in India: Rs 155
- Implied PPP is 155/4.93 = Rs 31.44/$
- Actual exchange rate = Rs 67.56/$
- $\frac{31.44-67.56}{67.56} = -0.53$
- Rupee undervalued by 53% against the USD

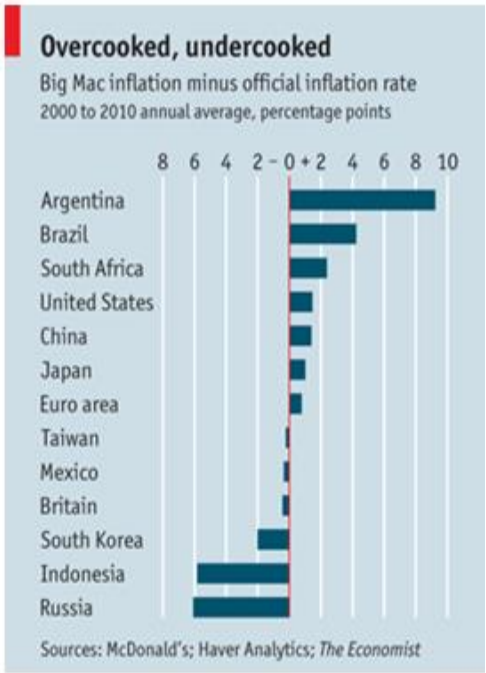


Global prices for a Big Mac in July 2016 based on a survey conducted in January 2016 by IMF, McDonald's, Thomson Reuters and The Economist

Burgernomics: Overvalued or Undervalued Currencies?

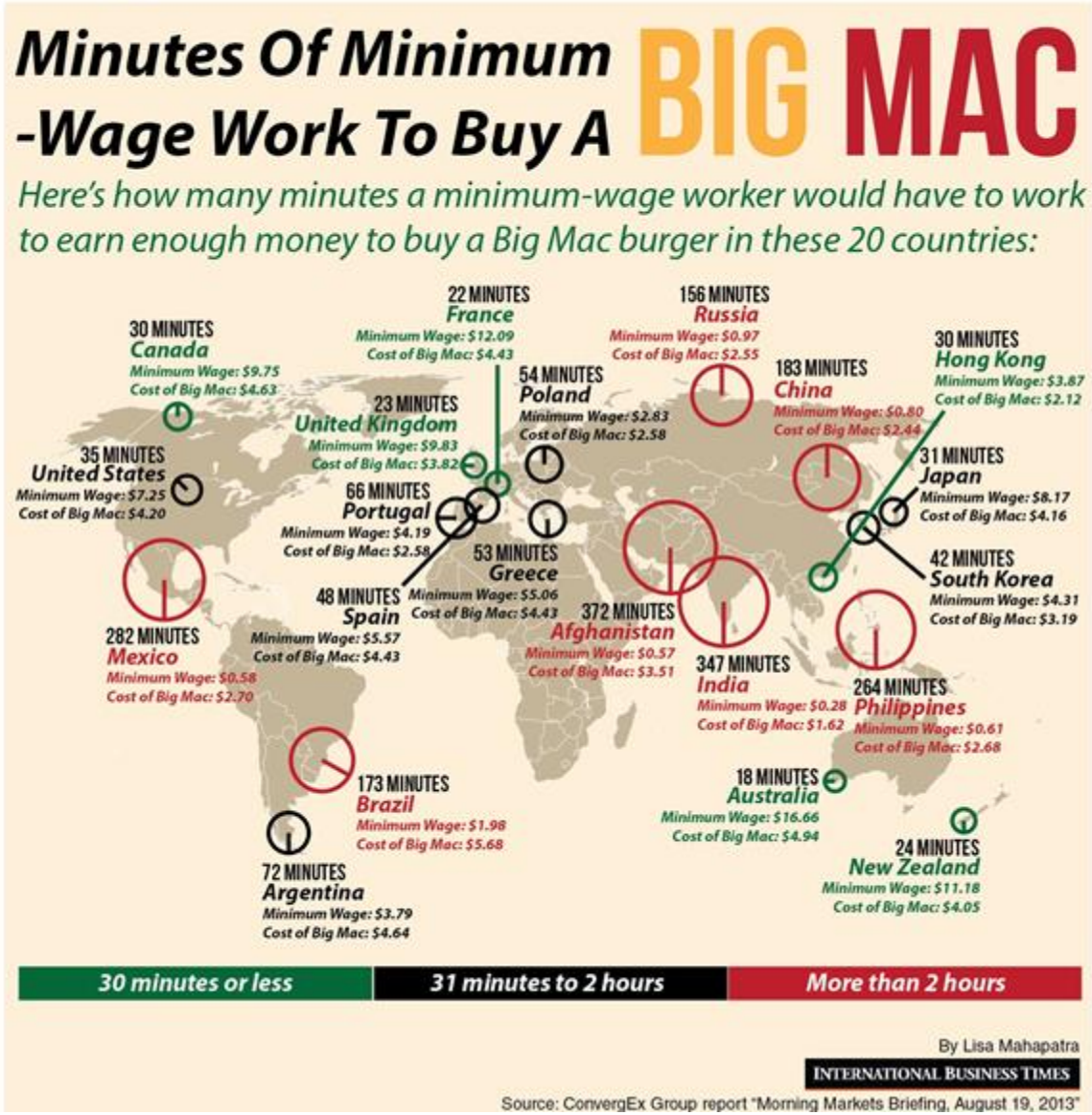




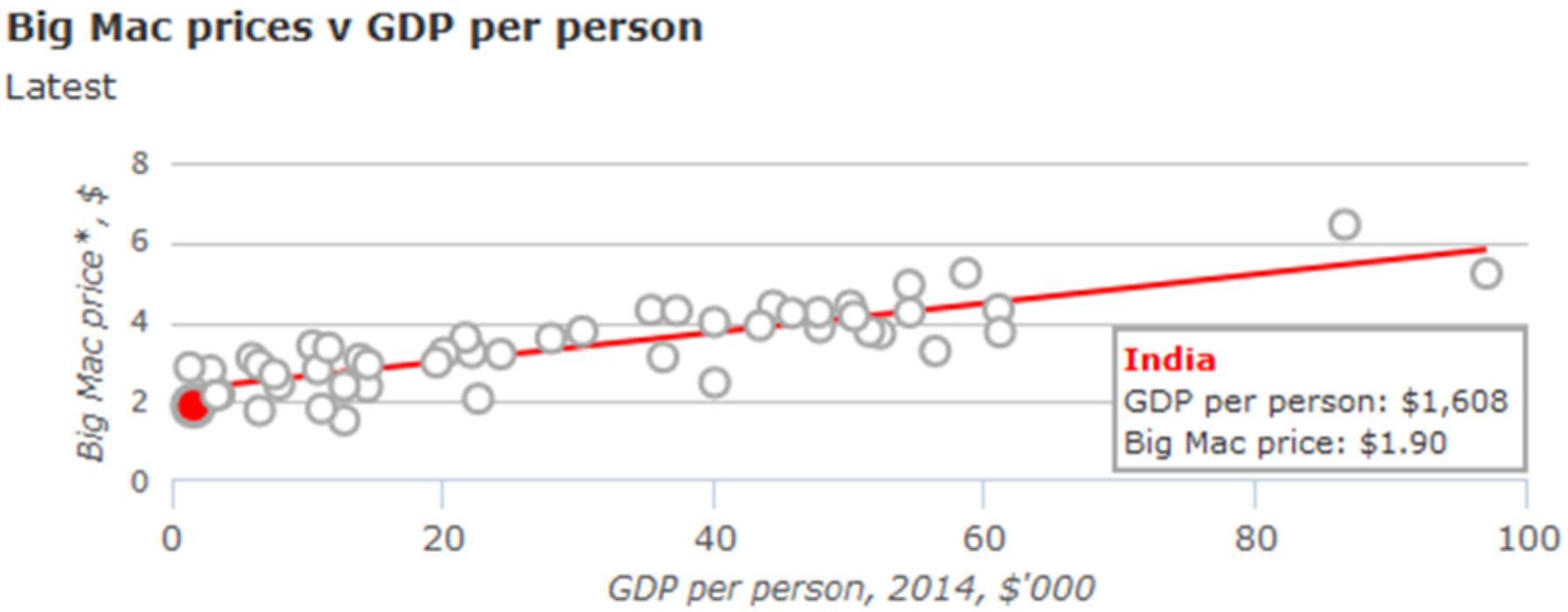


Source: Lies, flame-grilled lies and statistics
http://www.economist.com/node/18014576?story_id=18014576
Last accessed: March 04, 2016

Burgernomics by UBS Wealth Management Research



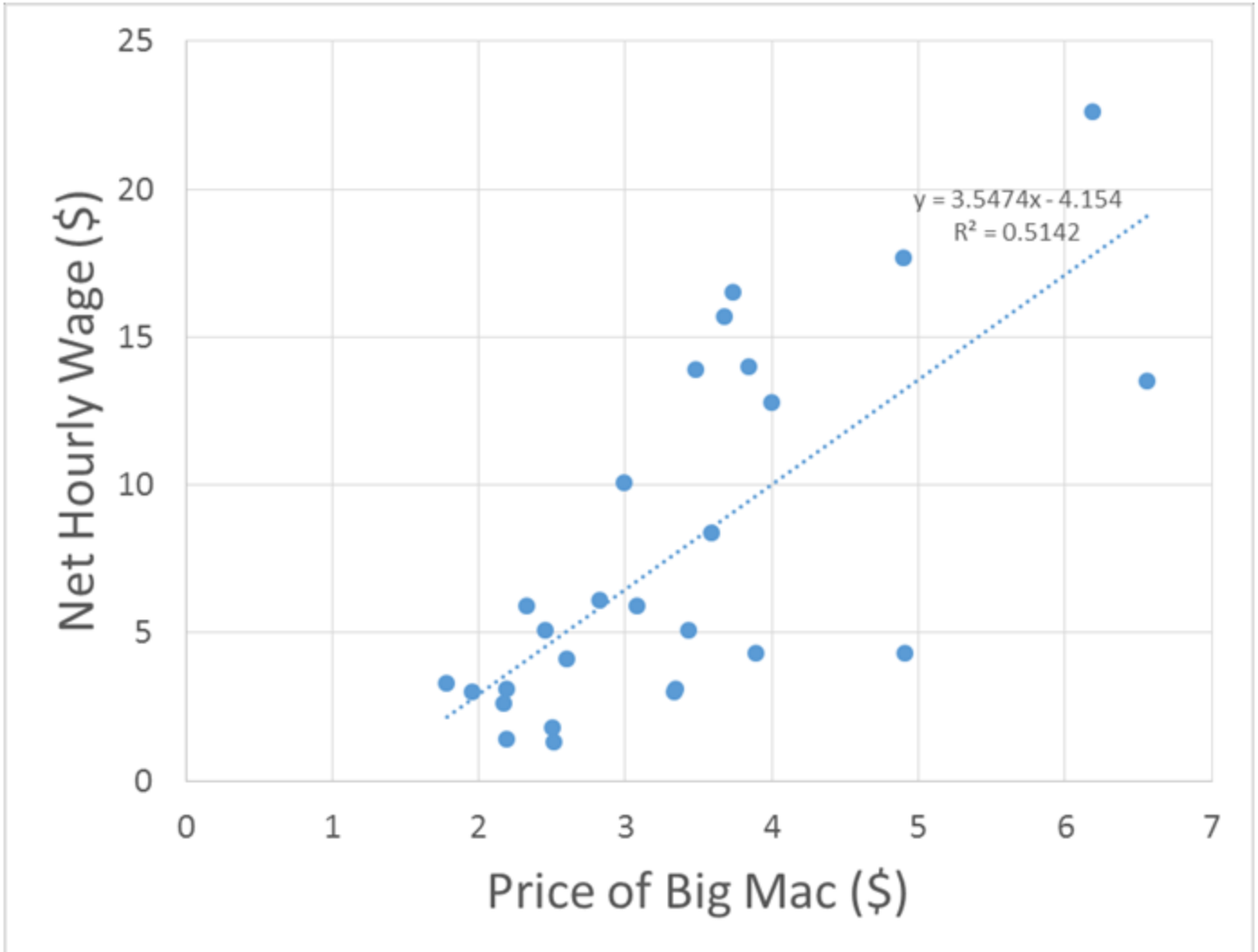
Burgernomics



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

Determining the Equation of the Regression Line - Excel

Determining the Equation of the Regression Line - Excel



WAYS OF TESTING HOW WELL THE REGRESSION LINE FITS DATA

Sample Software Output

SUMMARY OUTPUT

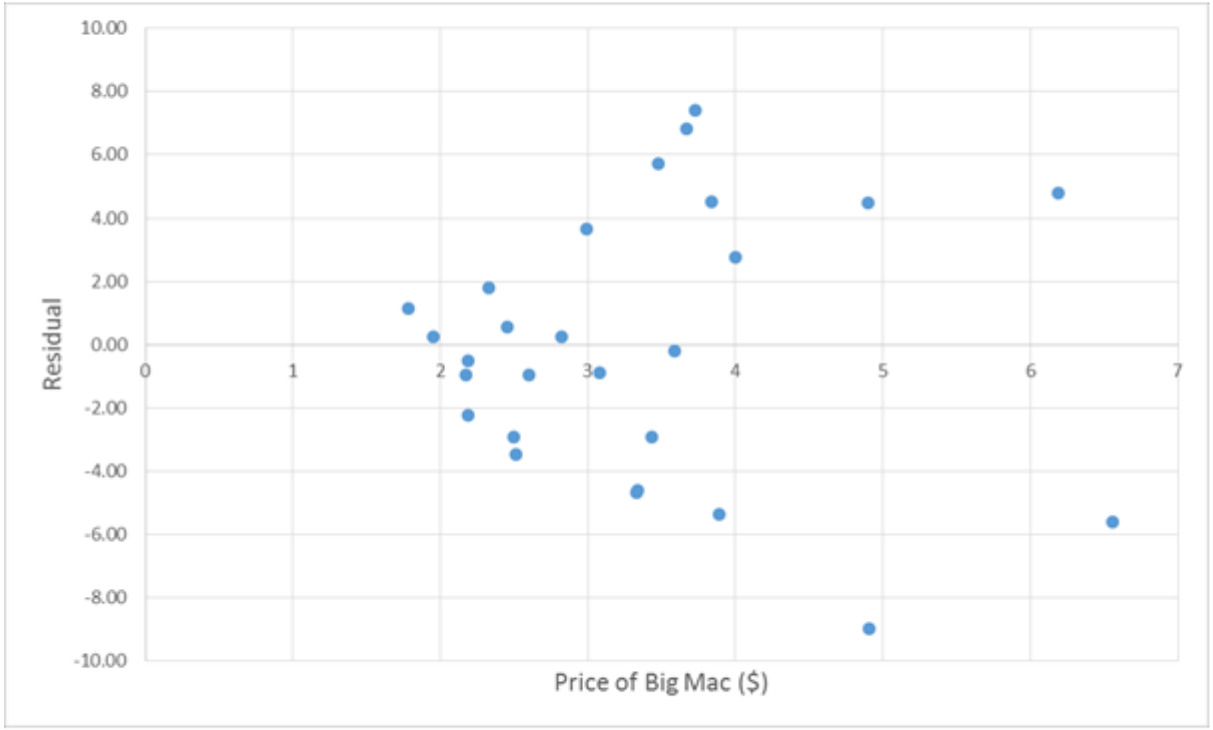
Regression Statistics	
Multiple R	0.717055011
R Square	0.514167888
Adjusted R Square	0.494734604
Standard Error	4.21319131
Observations	27

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

Residual Analysis

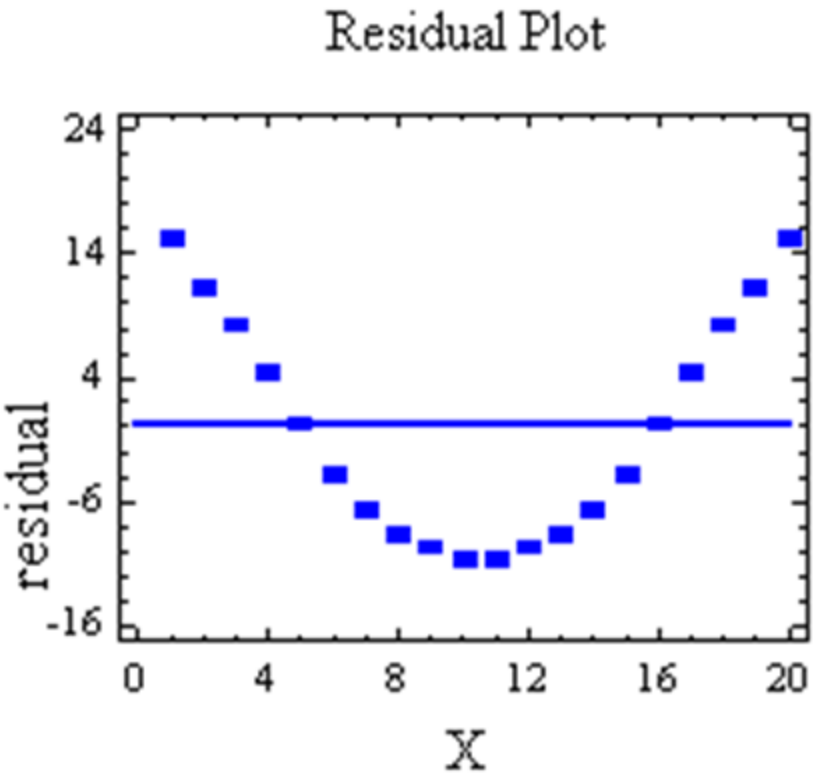


Can be used to locate outliers.

Residual = Forecast Errors= $(y_i - \hat{y}_i)$

Assumptions of the Regression Model

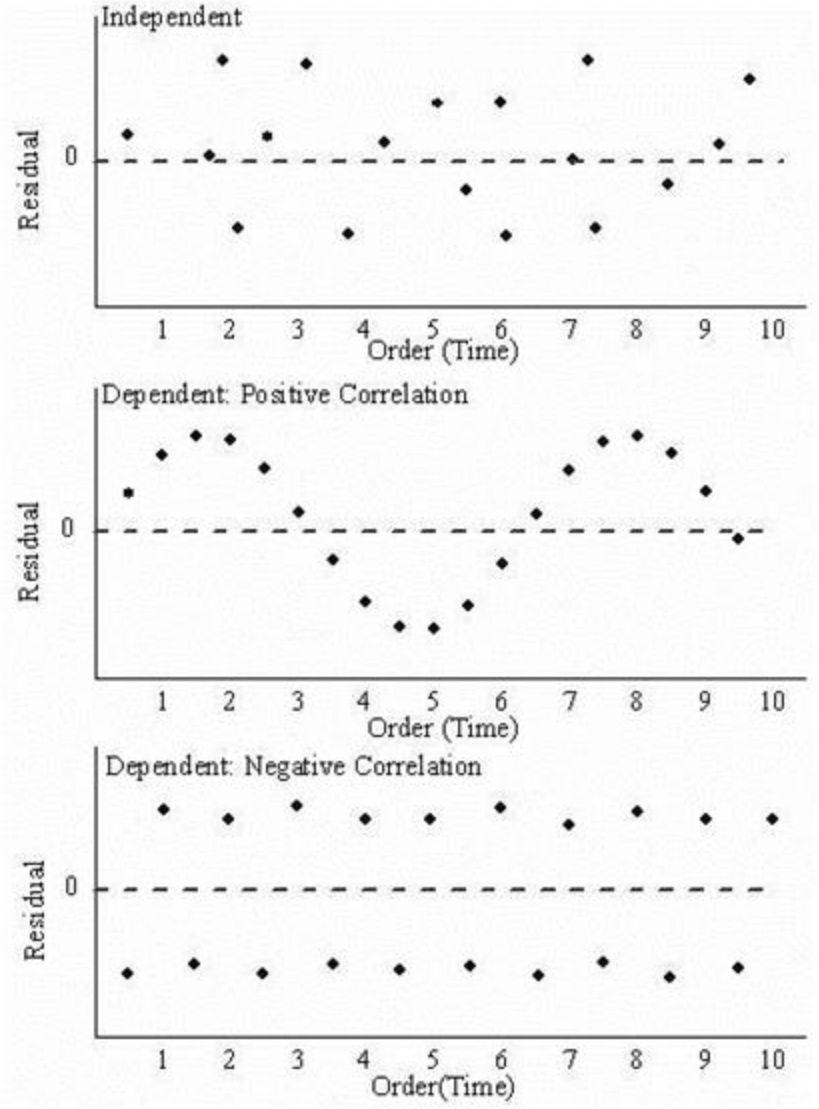
- The model is linear



Zero residual line:
The regression line

Assumptions of the Regression Model

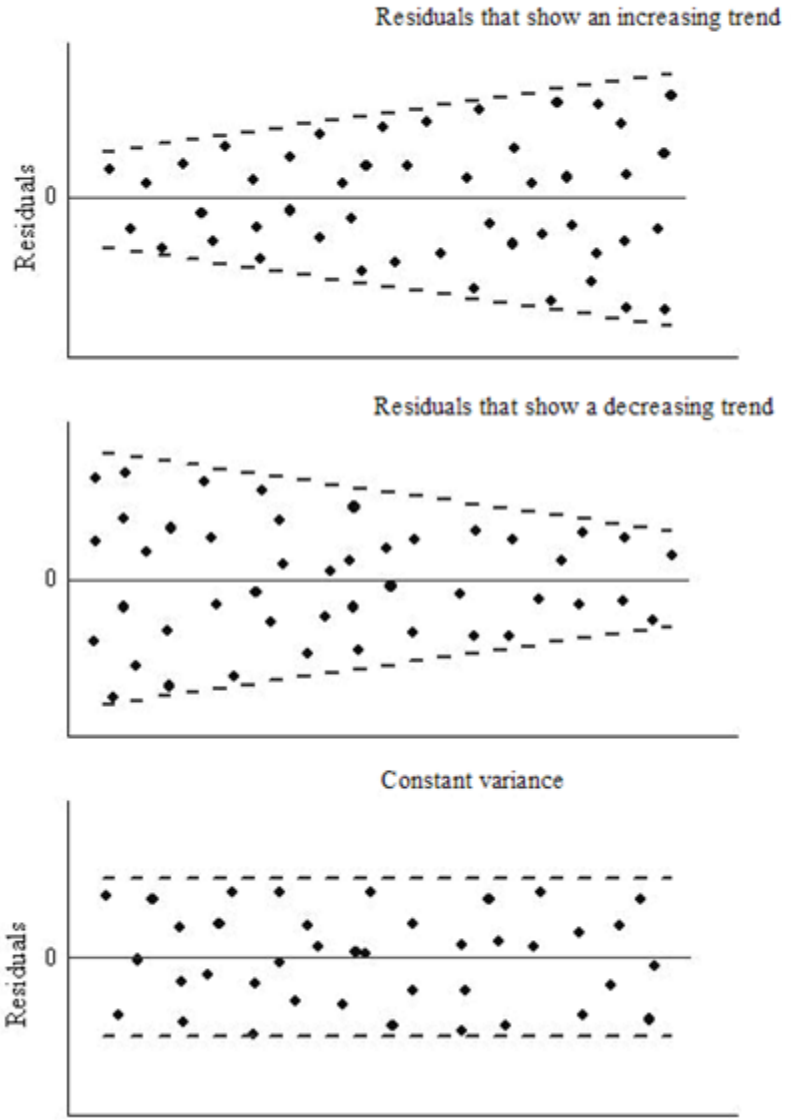
- The error terms are independent
  - Plot against any time (order of observation) or spatial variables preferably. Plots against independent variables may also detect independence.



Assumptions of the Regression Model

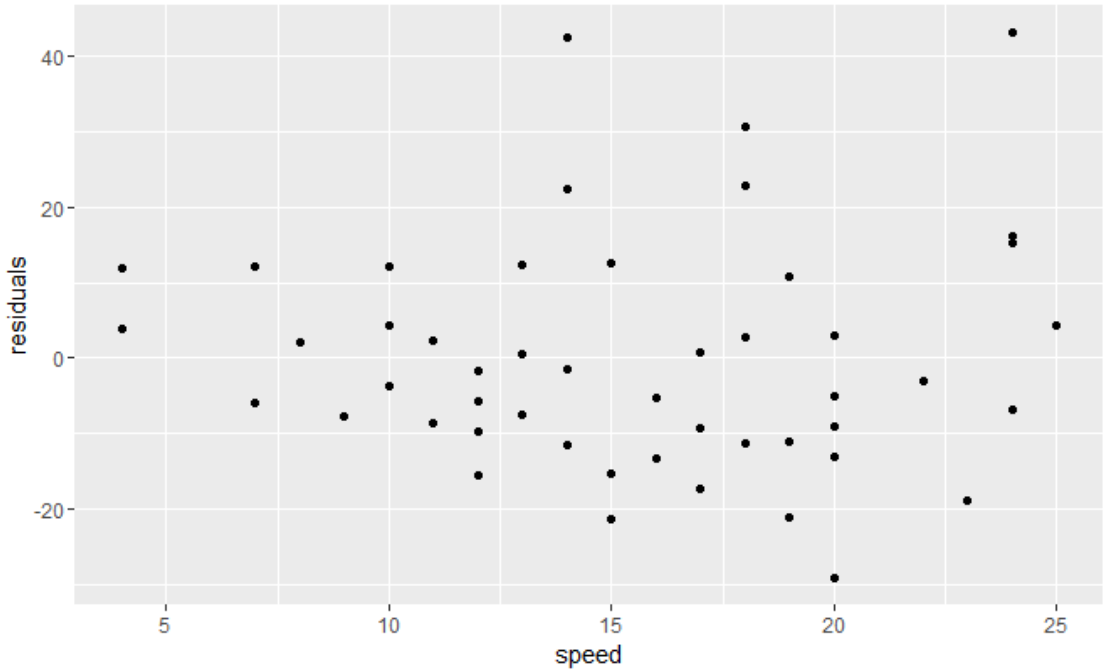
- The error terms have constant variances (homoscedasticity as opposed to heteroscedasticity)

  – 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.



Assumptions of the Regression Model

- The residual errors are normally distributed

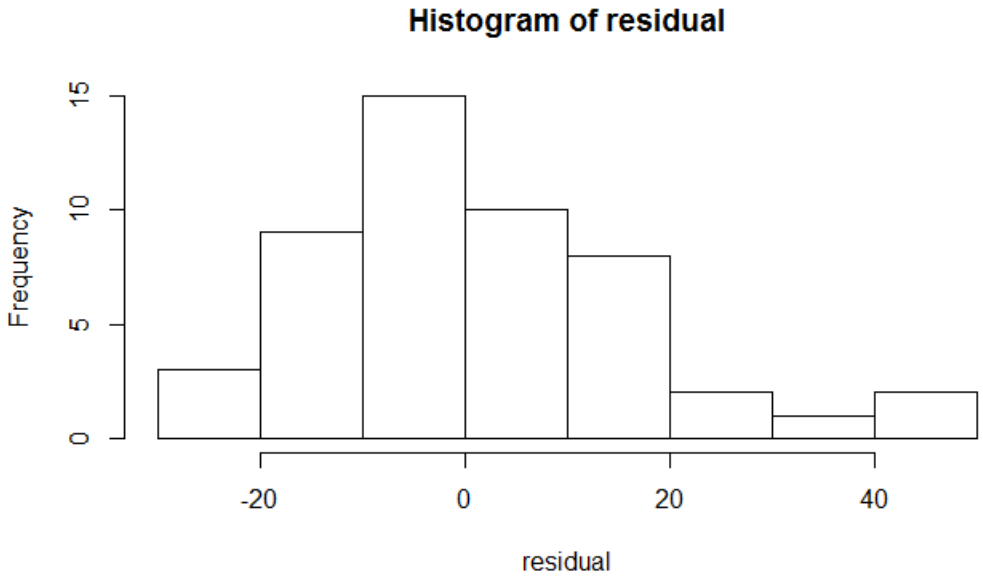


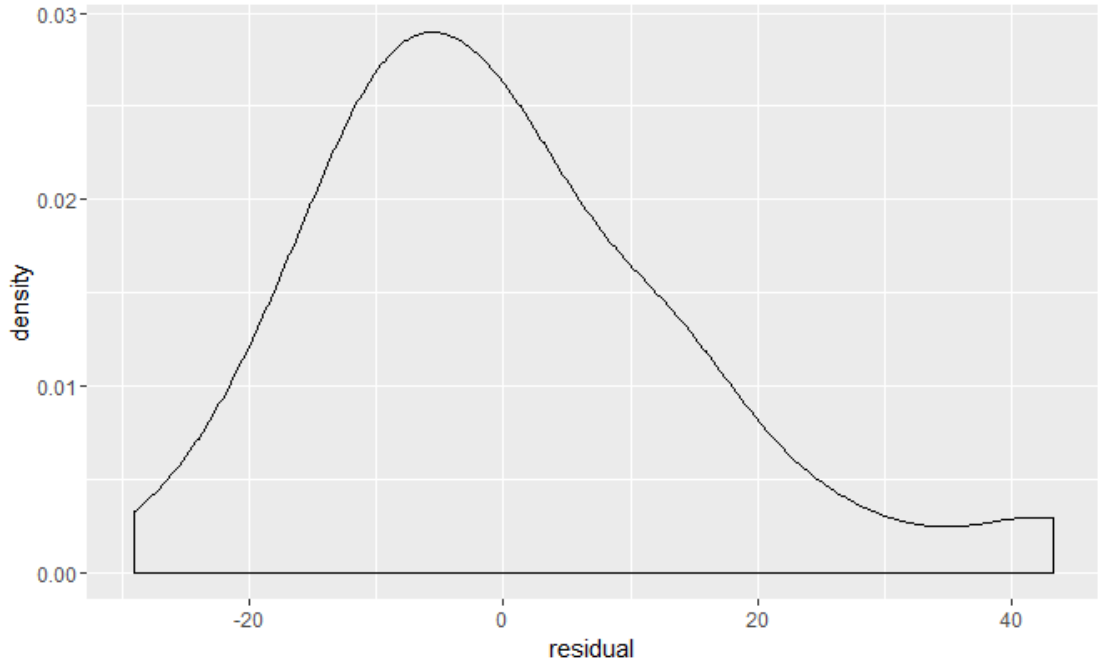
But, how do we know if something is normally distributed?

Checking for Normality

- Start by plotting the data

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






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

Is there a better way?

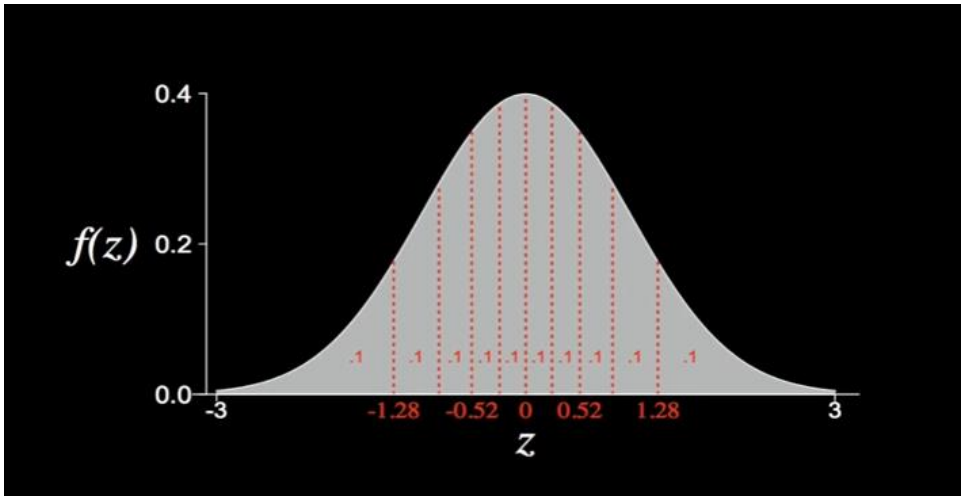
Quantile Quantile-Plot

- 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

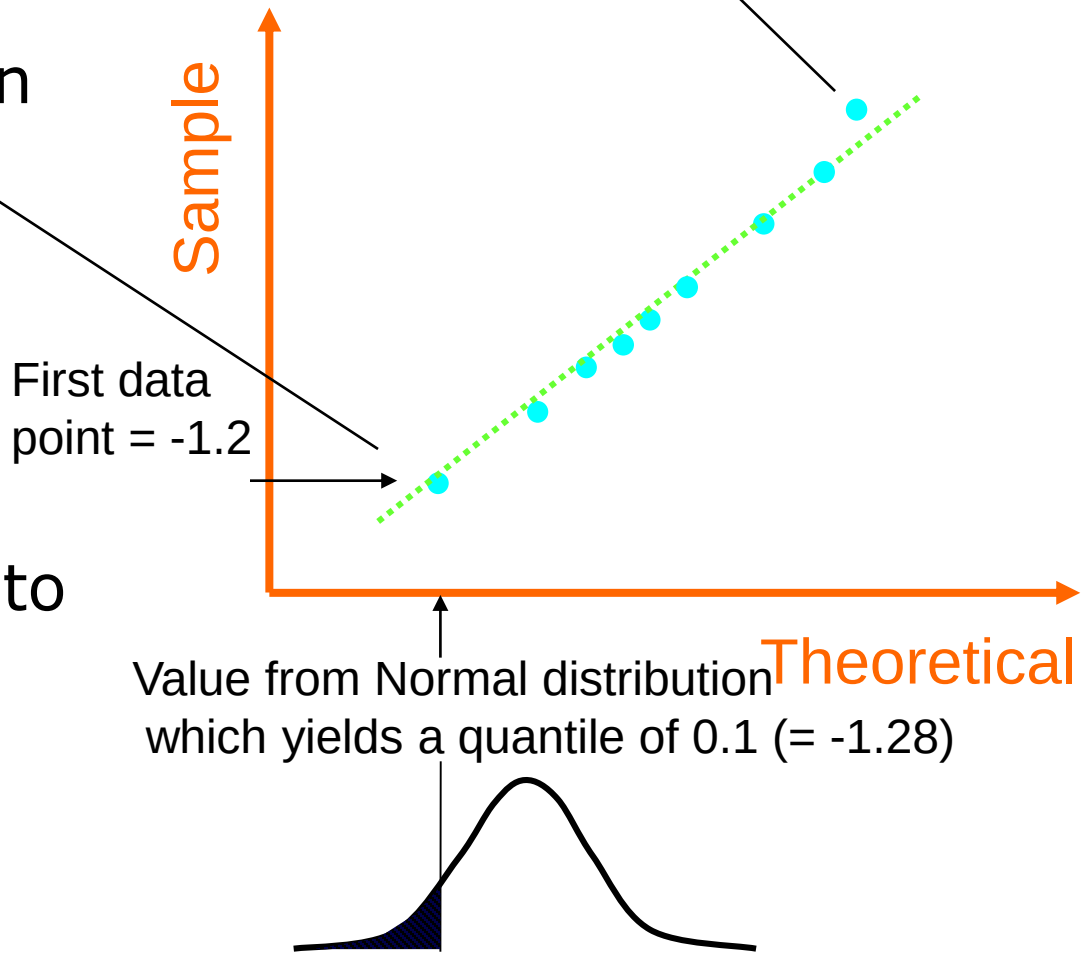


- Then one might expect the smallest of the 9 data points to be from the lowest quantile (0.1)
- Similarly, the largest value would be from the largest quantile (0.9) of the normal distribution

Quantile Quantile Plot

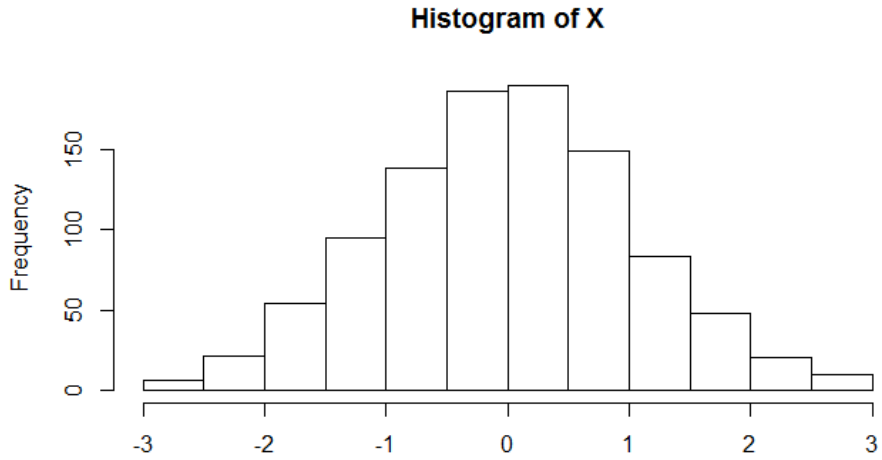
-1.2, -1.11, -1.08, -0.28, -0.25, 0.33, 0.41, 1.37, 1.41

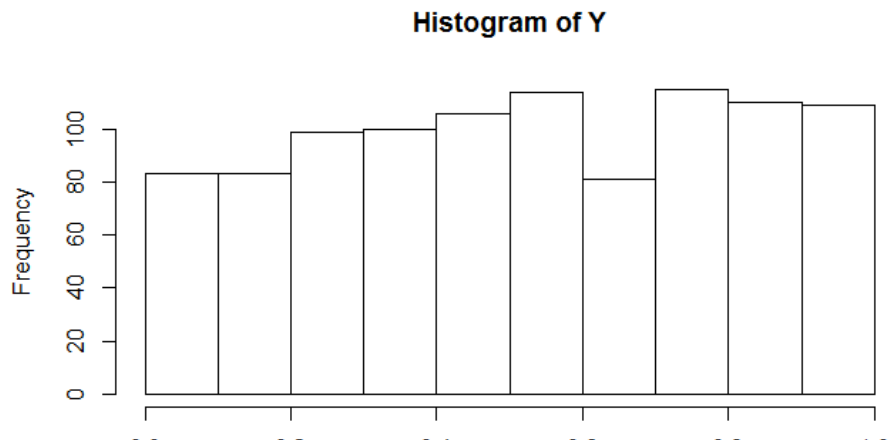
We plot the quantile values for the distribution on the x-axis and the values of the sample on the y-axis

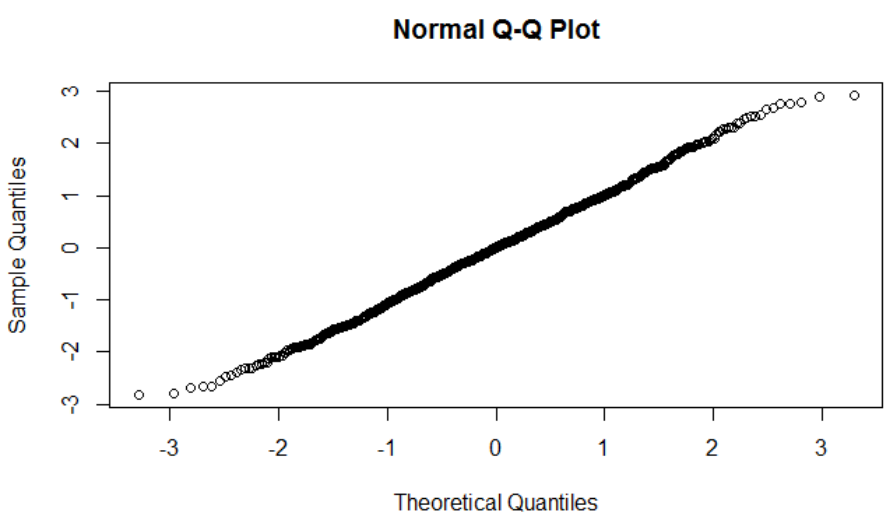


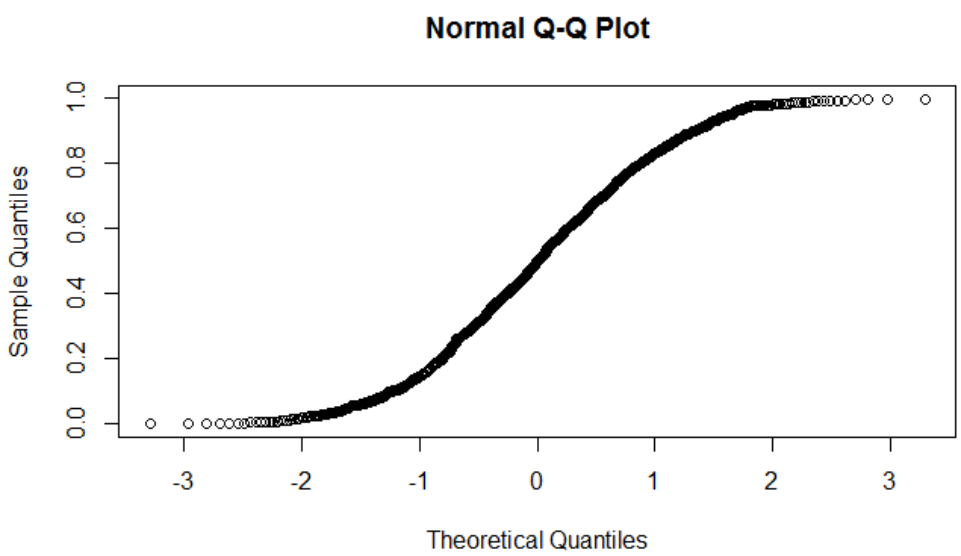
If the points lie on close to a straight line, then the sample is normal

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

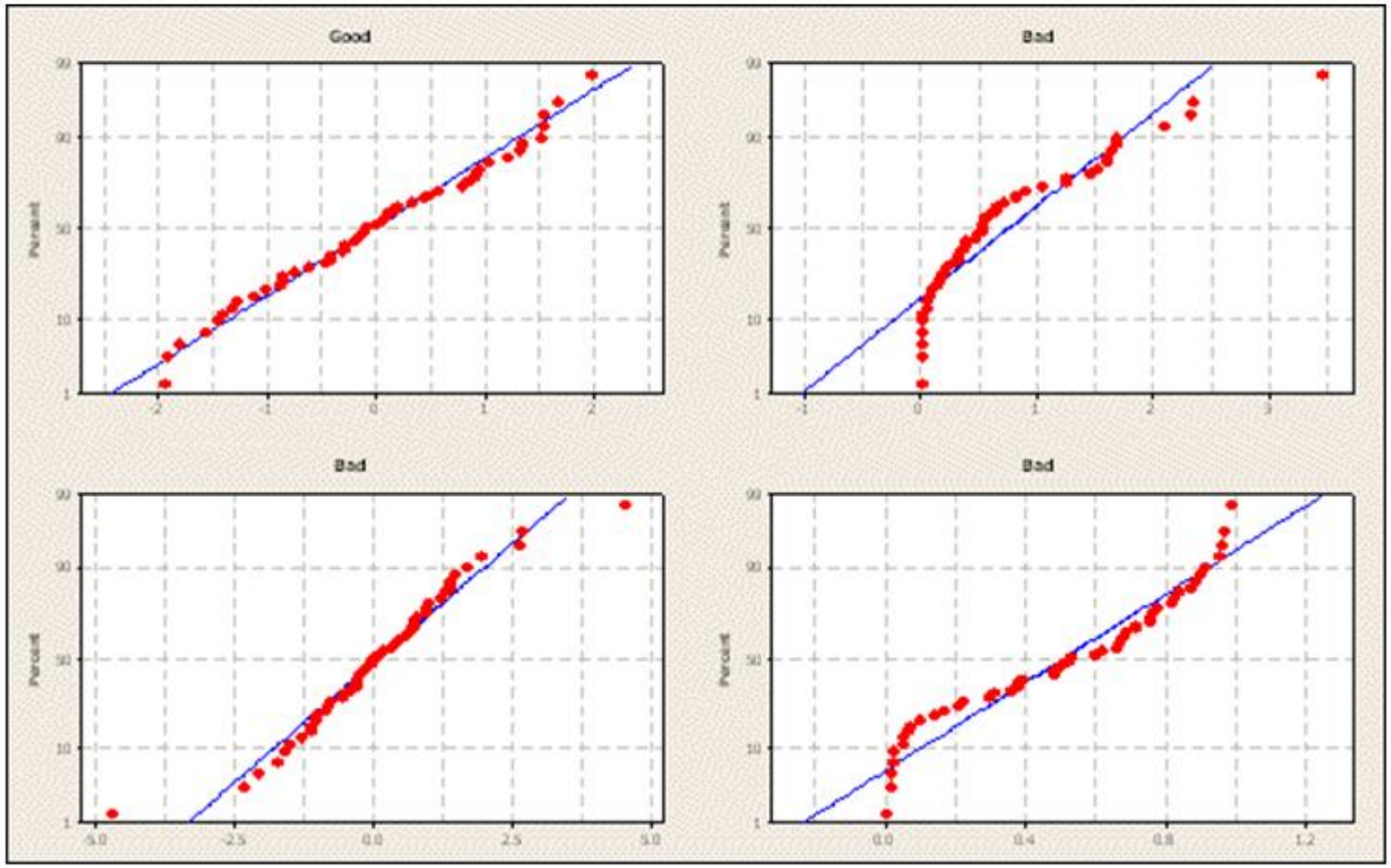
- Other objective methods of checking for normality also exist
- Shapiro-Wilk Test gives a probability value (p-value) that the given data sample is actually from a Normal distribution
- If p-value is less than 0.05, then its unlikely to be from Normal distribution

```
> 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

Assumptions of the Regression Model

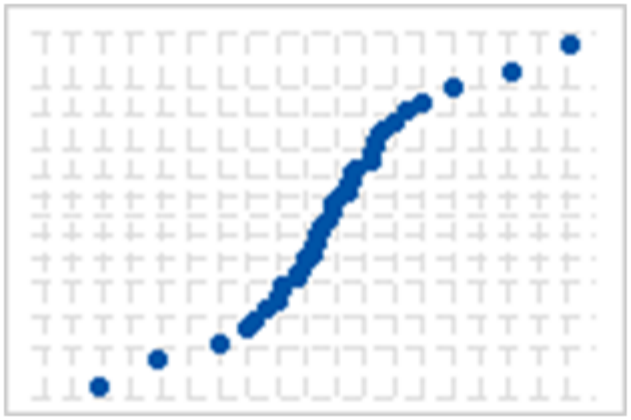
- The error terms are normally distributed



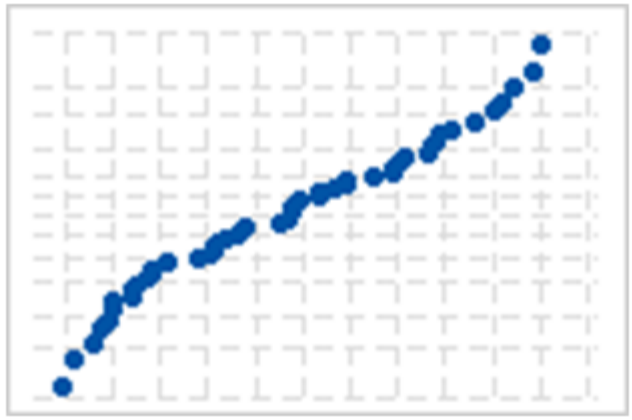
Interpreting Residuals

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

Interpreting Residuals – Non-normality

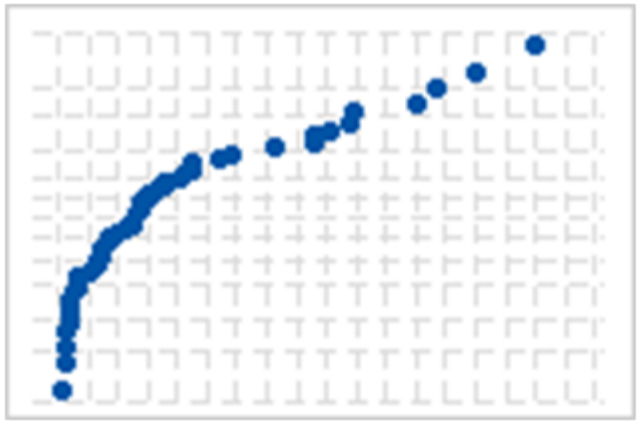


S-curve implies a distribution with long tails

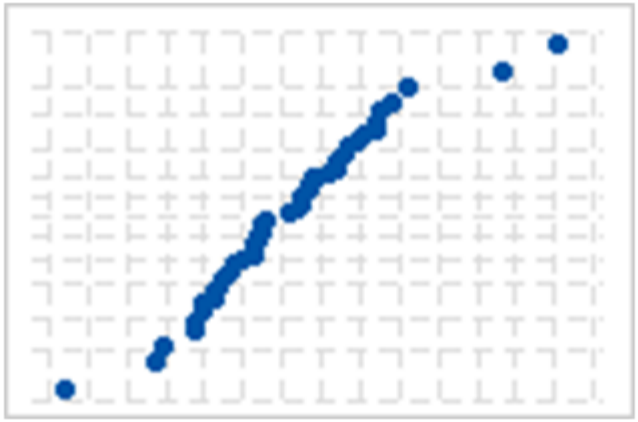


Inverted S-curve implies a distribution with short tails

Interpreting Residuals – Non-normality

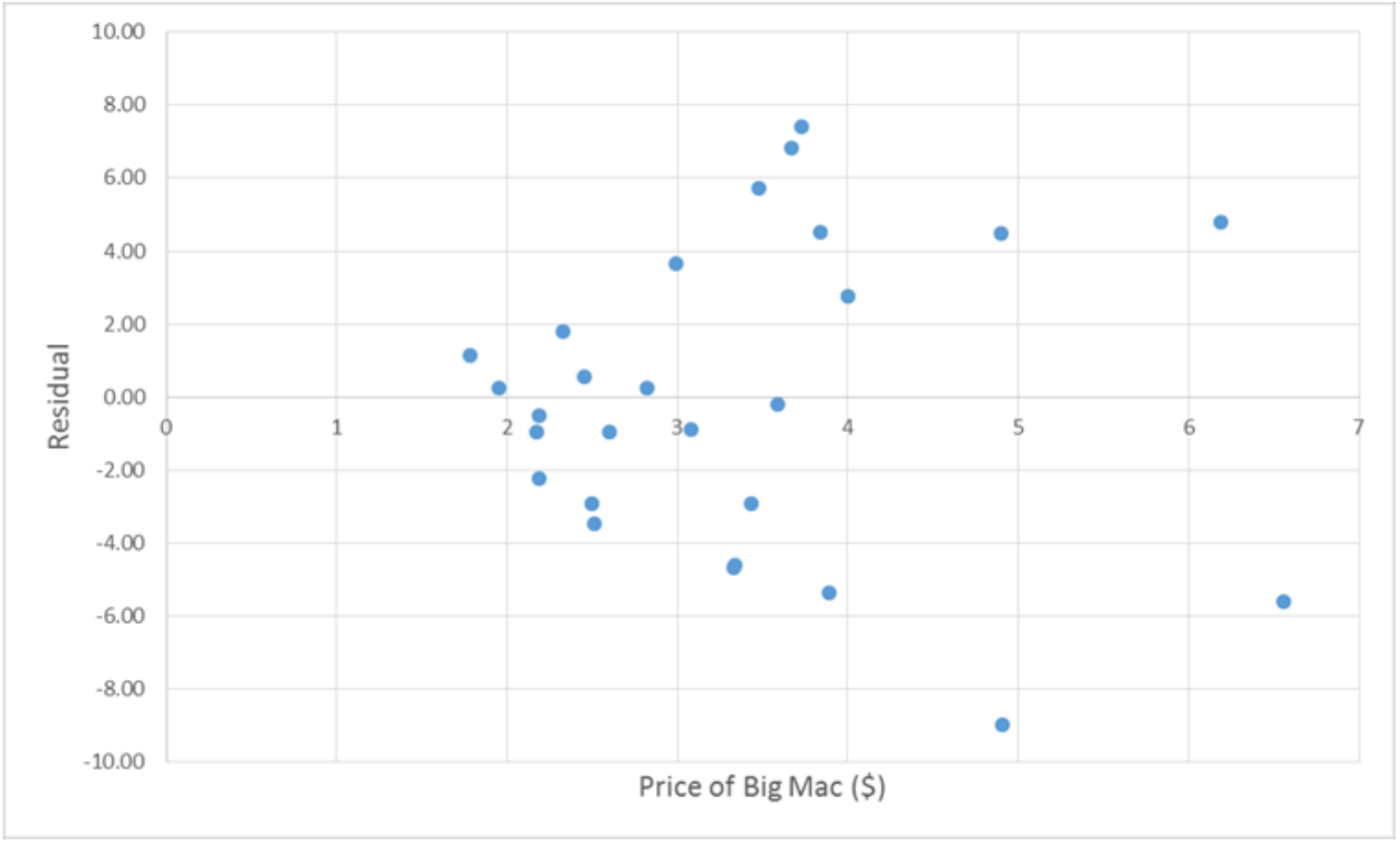


Downward curve implies an asymmetric distribution



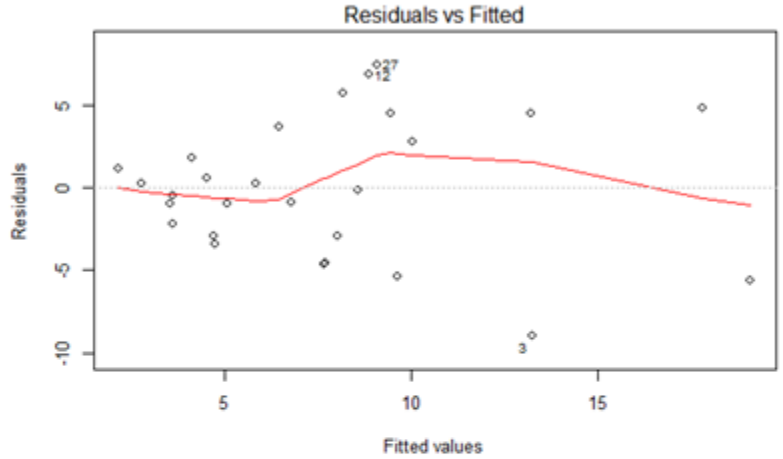
A few points lying away from the line implies a distribution with outliers

Residual Analysis

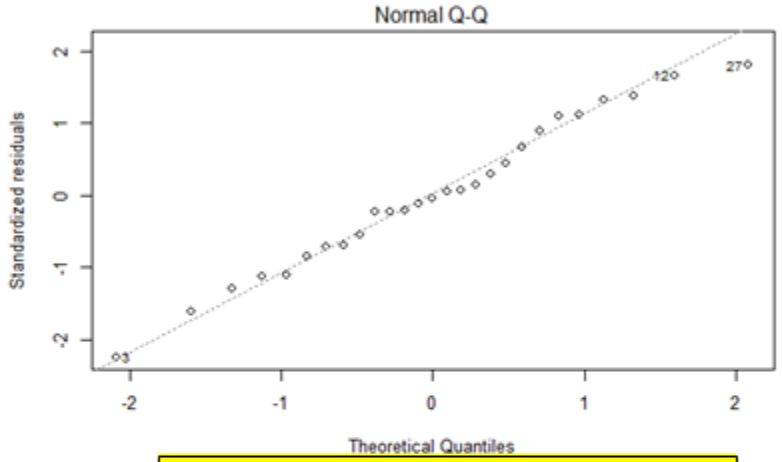


Residuals – Big Mac

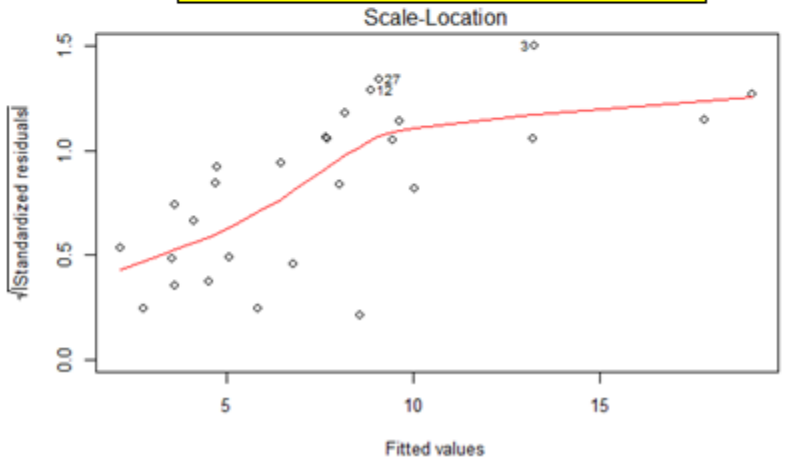
Is a wrong model fitted (linear or quadratic, etc.)?



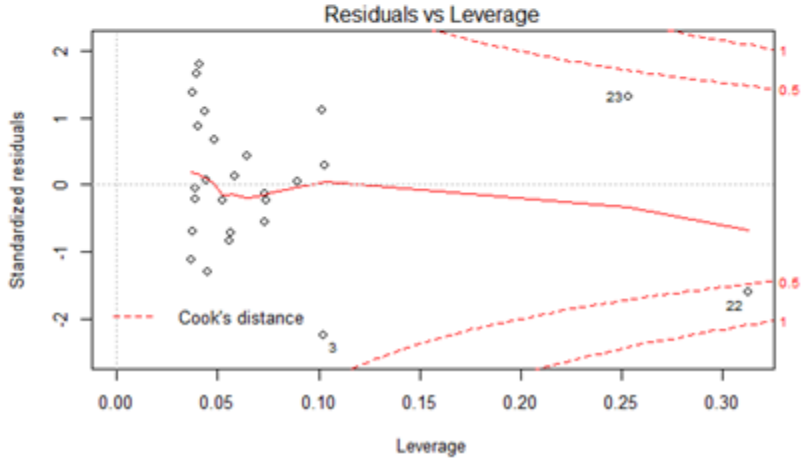
Are the residuals normally distributed?



Is the data homoscedastic?



Are there influential outliers?



CSE 7202c

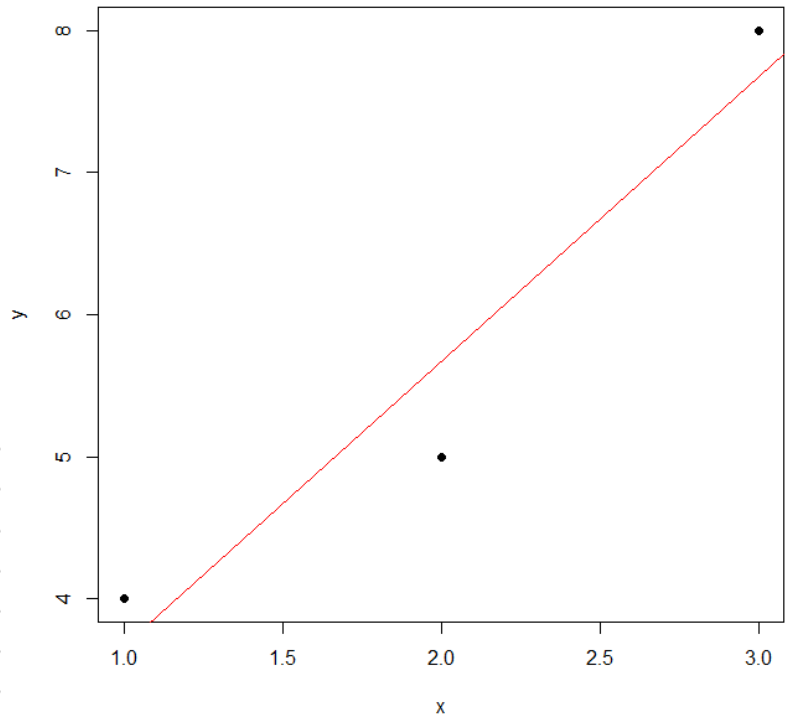
Fixing Non-normality and Heteroscedasticity

Transformation of data can help correct normality and unequal variances problems

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

A Toy Problem

x	y
1	4
2	5
3	8



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.960768923
R Square	0.923076923
Adjusted R Square	0.846153846
Standard Error	0.816496581
Observations	3

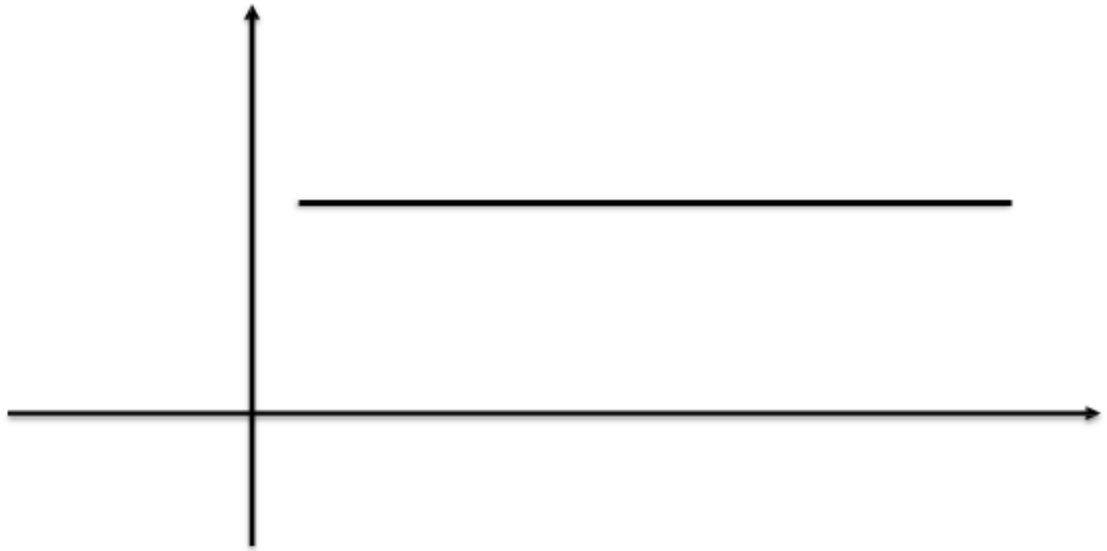
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	8	8	12	0.178912375
Residual	1	0.666666667	0.666667		
Total	2	8.666666667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.666666667	1.247219129	1.336306	0.408985	-14.18075494	17.51408827	-14.18075494	17.51408827
x	2	0.577350269	3.464102	0.178912	-5.335930725	9.335930725	-5.335930725	9.335930725

Testing the Slope

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



Testing the Slope

What is the Null Hypothesis?

$H0:\beta 1=0$

What is the Alternative Hypothesis?

$H1:\beta 1\neq 0$

Standard Error of the Estimate

The Sum of Squares Error (SSE) is a function of the number of pairs of data, and so **standard error of the estimate**, SE, is computed as a more useful measure, which is nothing but the standard deviation of the error of the regression model.

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

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

t Test of the Slope

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

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

$$SE = \sqrt{\frac{SSE}{n-2}}$$

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

β_1 = *the hypothesized slope*

t Test of the Slope – Toy Problem

t=3.46

At 5% significance level, the critical region for a 2 tailed t-test is

$$t_{1,0.025} = 12.71$$

Since *t* value calculated from the sample slope is in **not** in rejection region, we accept 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^*.

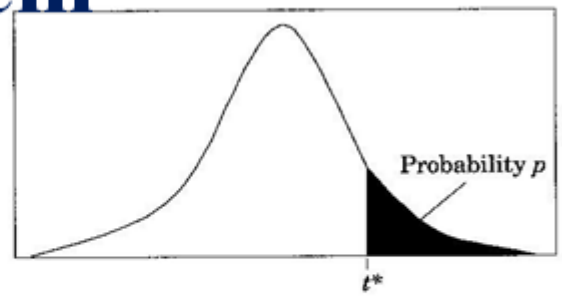


Table B *t* distribution critical values

	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											

Testing the Overall Model

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

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

Testing the Overall Model

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

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

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

$$SS_{yy} = SSR + SSE$$

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

In simple regression, $F = t^2$

Testing the Overall Model – Toy Problem

F=12

Note $t^2 = (3.46)^2 = 12$

Critical F value

$$F_{0.05,1,1} = 161.45$$

Accept the null hypothesis. Model lacks significance

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

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

Back to Big-Mac problem

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

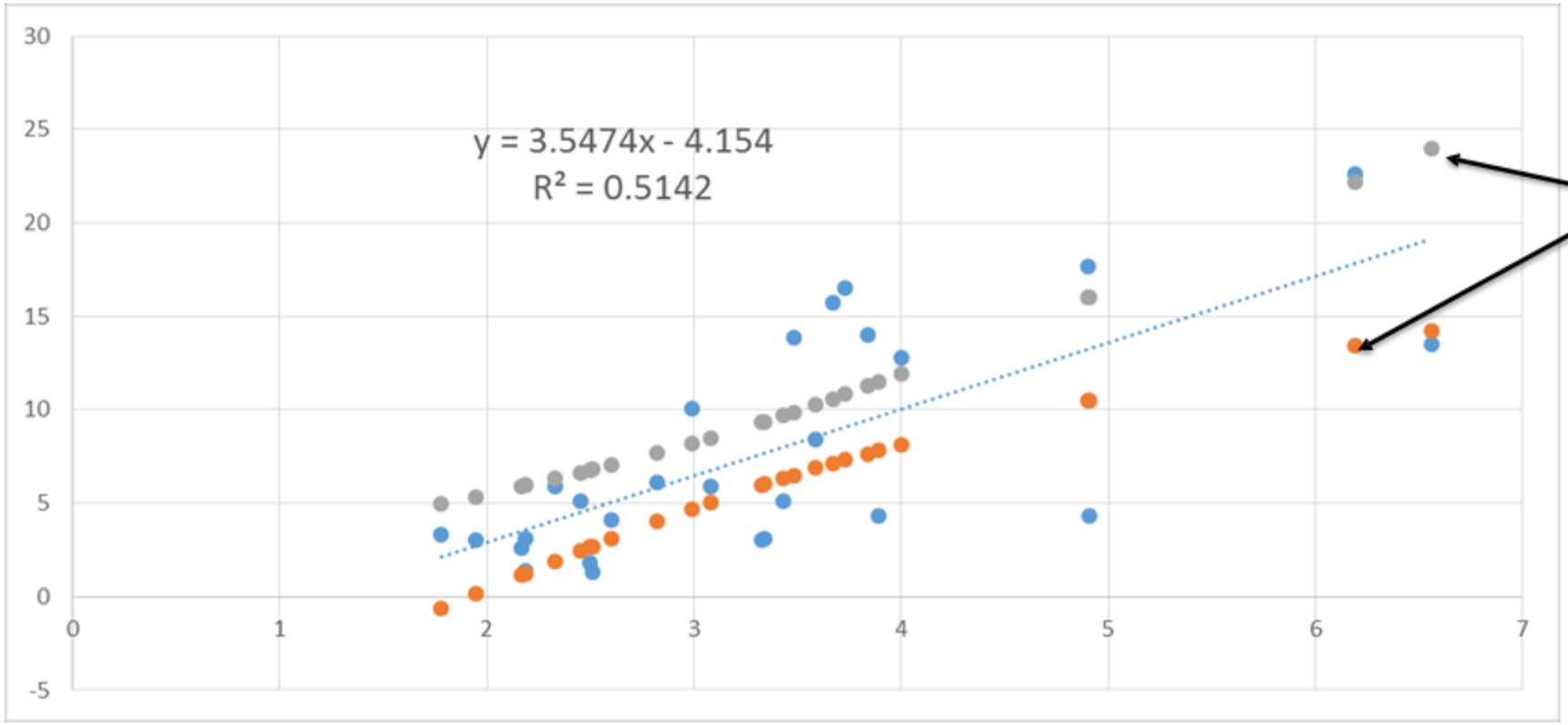
Estimation – Confidence Intervals

A regression line provides a point estimate from a sample. A different sample may yield a different point estimate. A Confidence Interval for estimating an **average value of y for a given x** is more useful.

$$E(y_x) = \hat{y} \pm t_{n-2,\frac{\alpha}{2}} * SE * \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

where x_0 = a particular value of x

Estimation – Confidence Intervals – Big Mac - Excel



95% CI

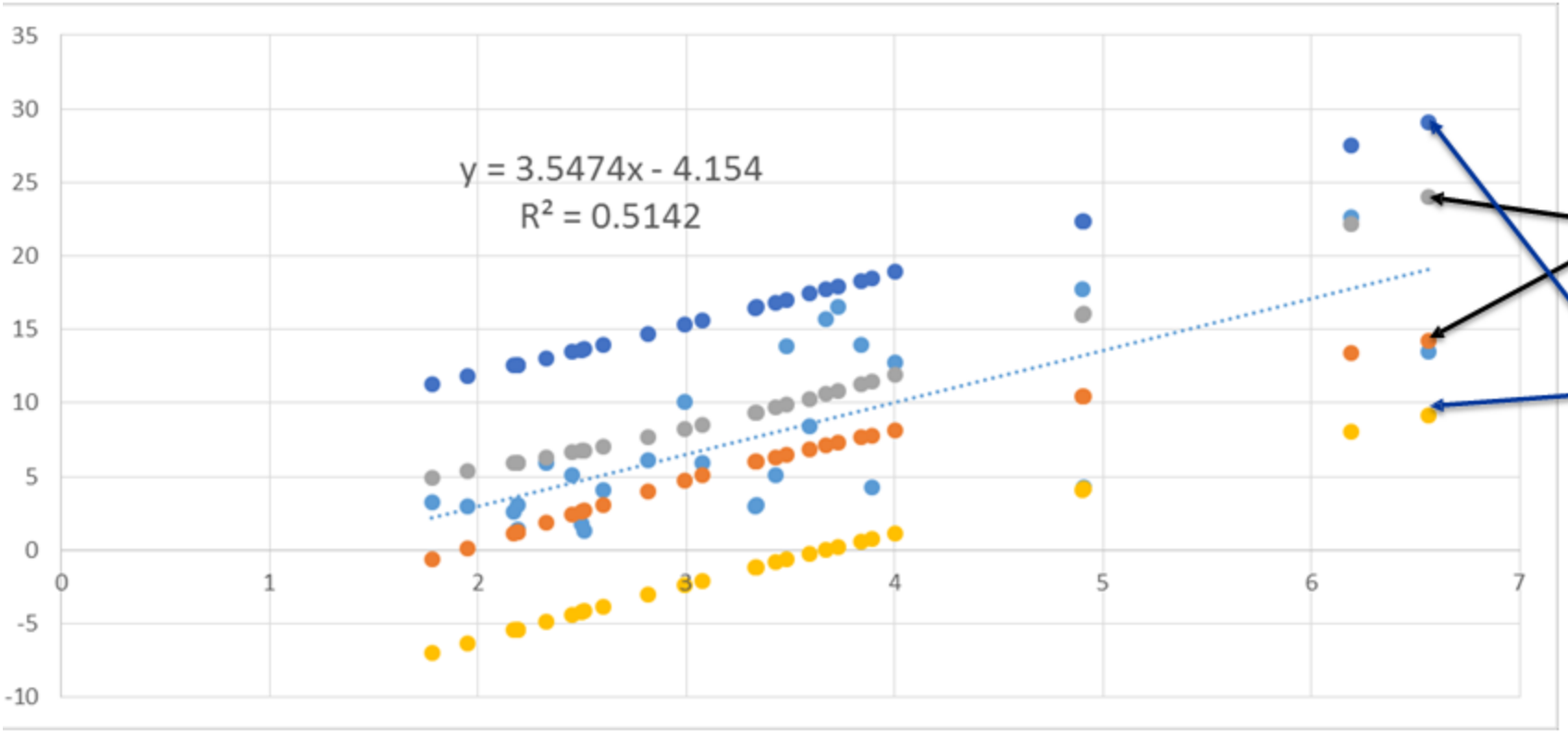
Estimation – Prediction Intervals

A Prediction Interval estimates a **single value of *y* for a given *x***.

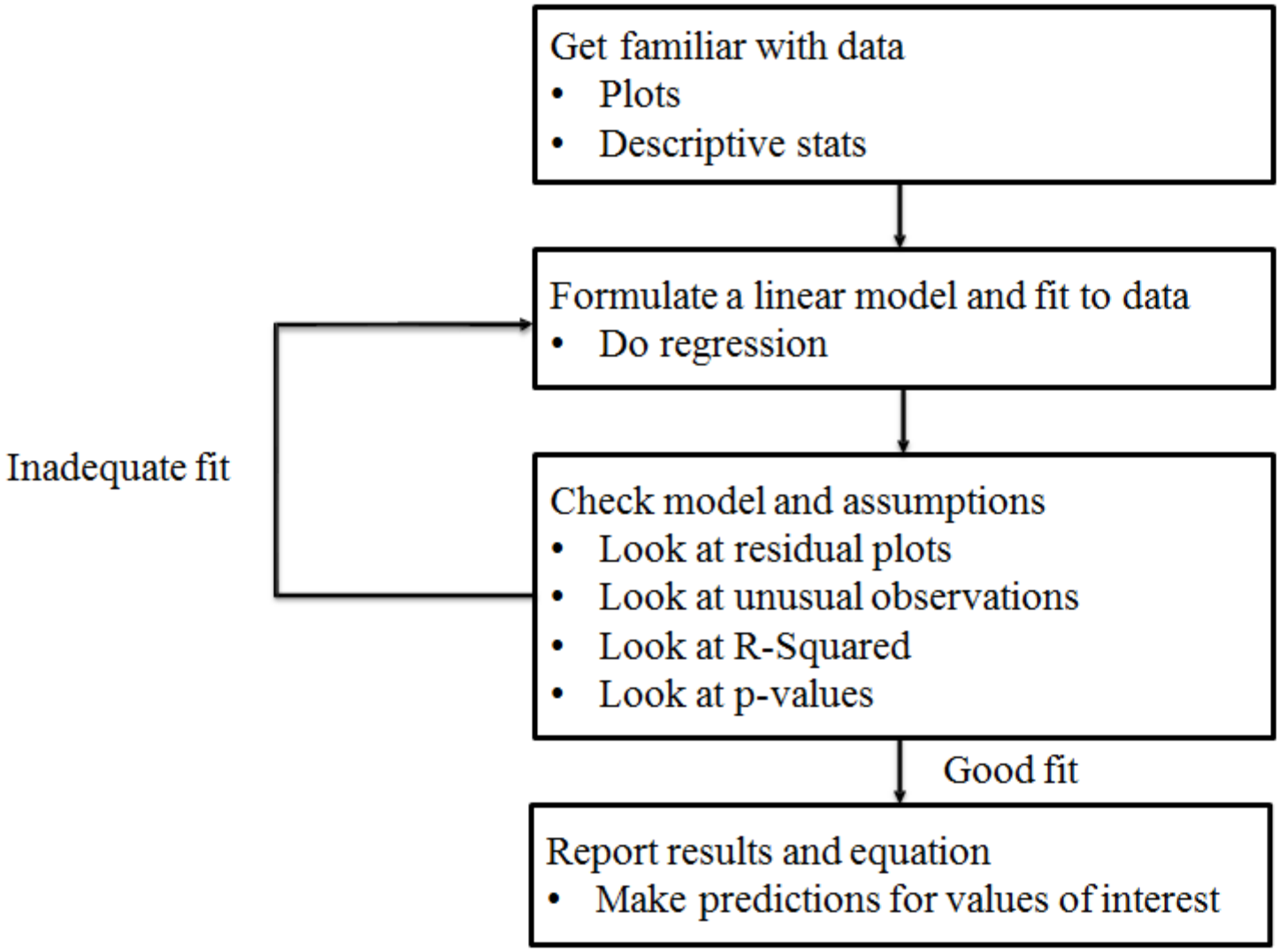
$$y_x = \hat{y} \pm t_{n-2,\frac{\alpha}{2}} * SE * \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

where x_0 = a particular value of x

Estimation – Prediction Intervals – Big Mac - Excel

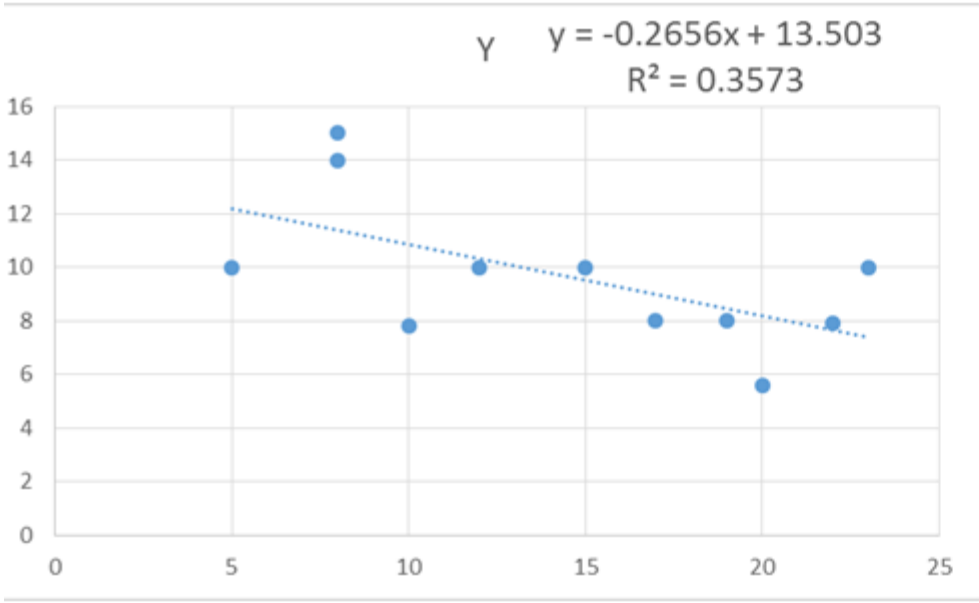


Simple Linear Regression - Steps



R^2 as metric for quality of fit – some caveats

R-Squared and Significance - Caution

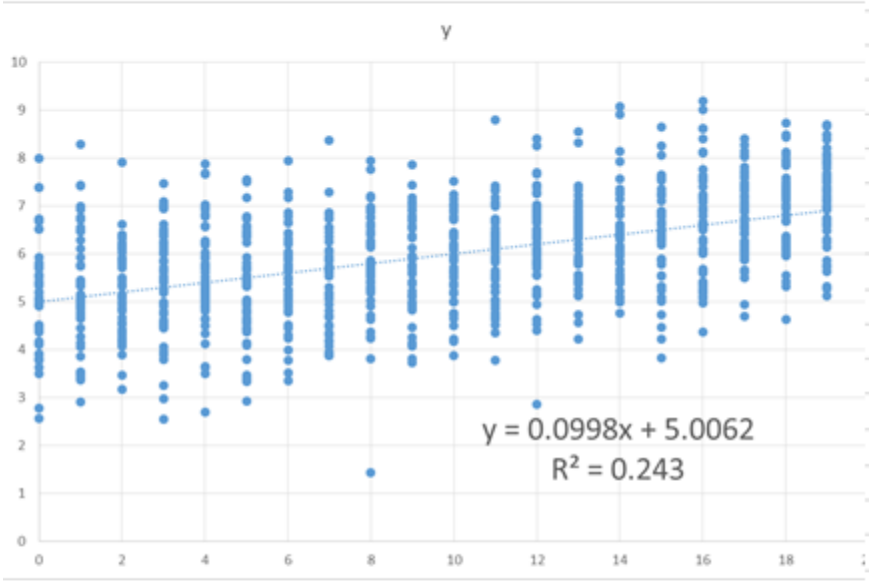


23	Regression Analysis						
24							
25	OVERALL FIT						
26	Multiple R	0.597718					
27	R Square	0.357267					
28	Adjusted R Square	0.285852					
29	Standard Error	2.335299					
30	Observations	11					
31							
32	ANOVA				Alpha	0.05	
33		*df*	*SS*	*MS*	*F*	*p-value*	*sig*
34	Regression	1	27.282857	27.28285699	5.002704	0.052125754	no
35	Residual	9	49.082598	5.453621951			
36	Total	10	76.365455				
37							
38		*coeff*	*std err*	*t stat*	*p-value*	*lower*	*upper*
39	Intercept	13.50289	1.8553076	7.277980002	4.67E-05	9.305894086	17.699888
40	X	-0.26561	0.1187518	-2.23667255	0.052126	-0.53424402	0.0030263

- R-Sq suggests that 35% of variation in y can be explained by variation in x.
- t and F tests show that coefficient is <u>not significant</u> and null hypothesis cannot be rejected.

The 95% confidence interval of the slope, $b_1 \pm t_{crit} * s_b$, is (-0.534,0.003).

R-Squared and Significance - Caution



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

ANOVA

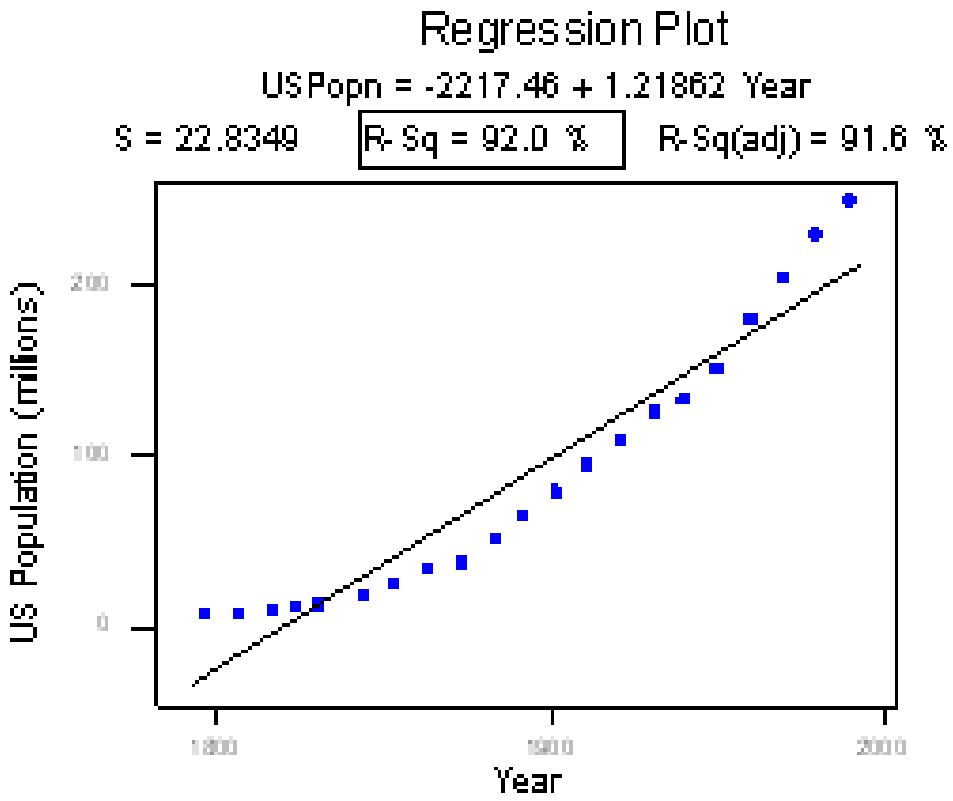
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	331.1568641	331.1568641	320.3010019	2.43789E-62
Residual	998	1031.824904	1.033892689		
Total	999	1362.981768			

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

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

Caution: High R^2 doesn't imply a good fit!

- US population from 1790 to 1900 (decade wise data)

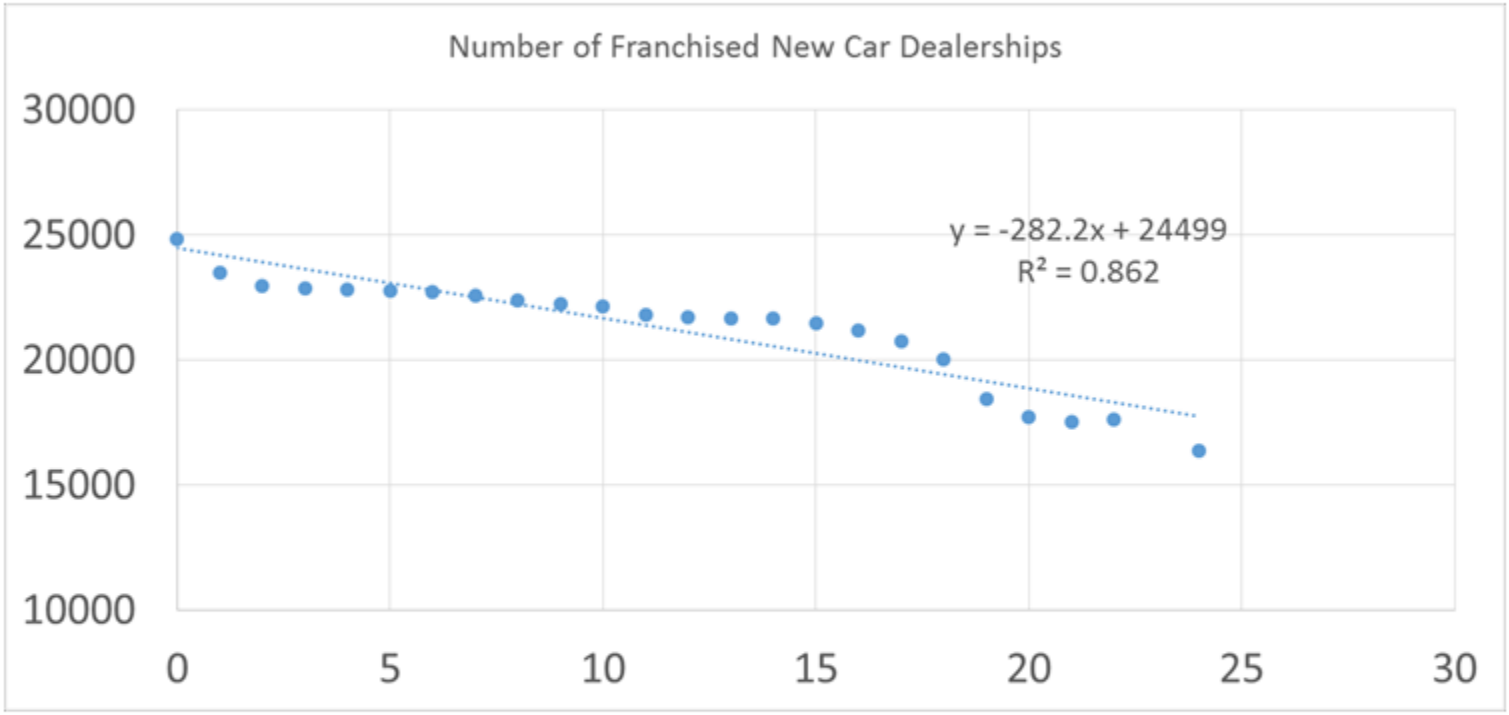


New Car Dealerships data

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



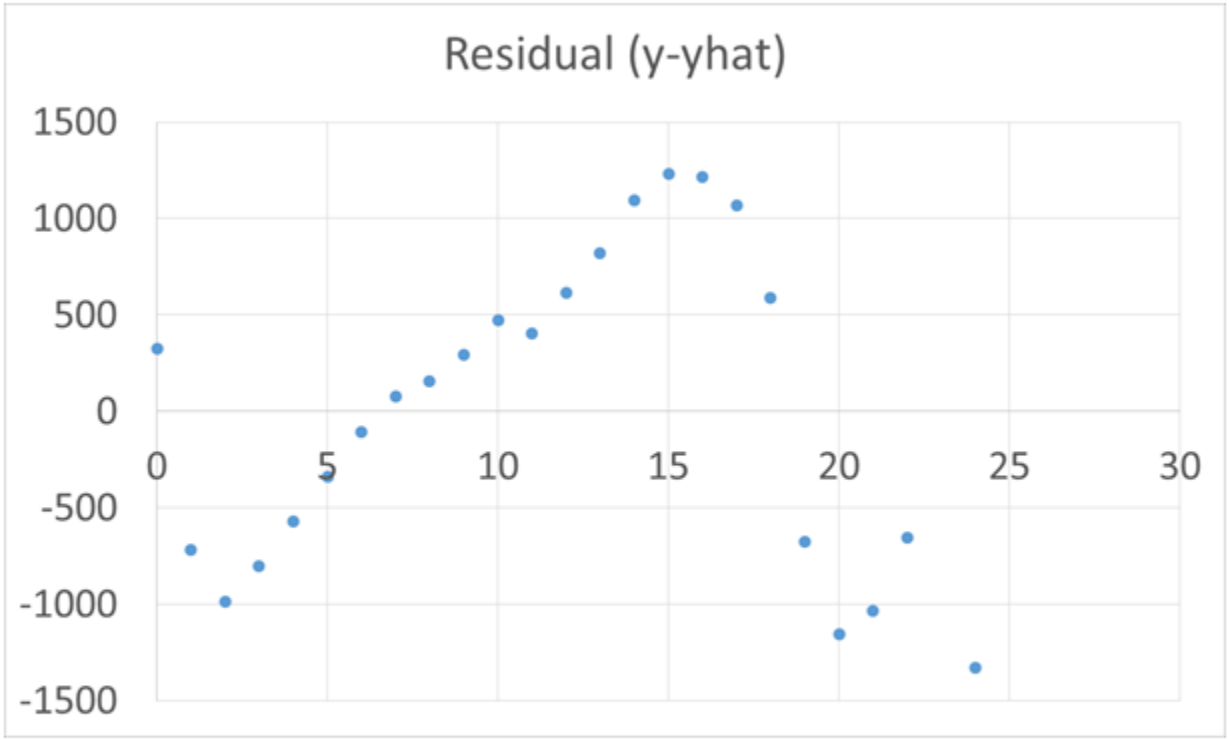
- Based on the shape of the scatter plot, do you think a linear fit looks good?
- Does R^2 imply a good fit?
- What can you infer from the intercept and the slope?

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

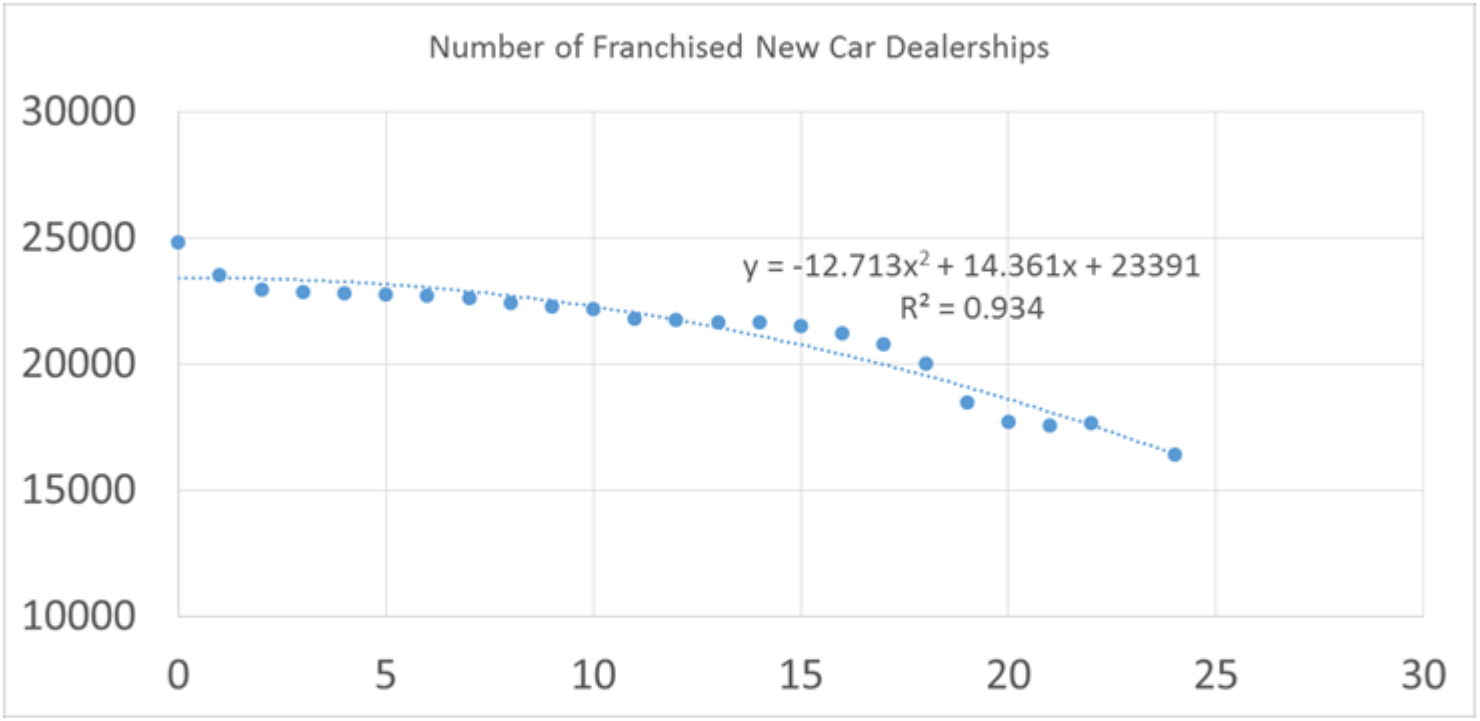
- Is the slope significant?
- Is the model significant?

R-Squared, Significance and Residuals - Caution

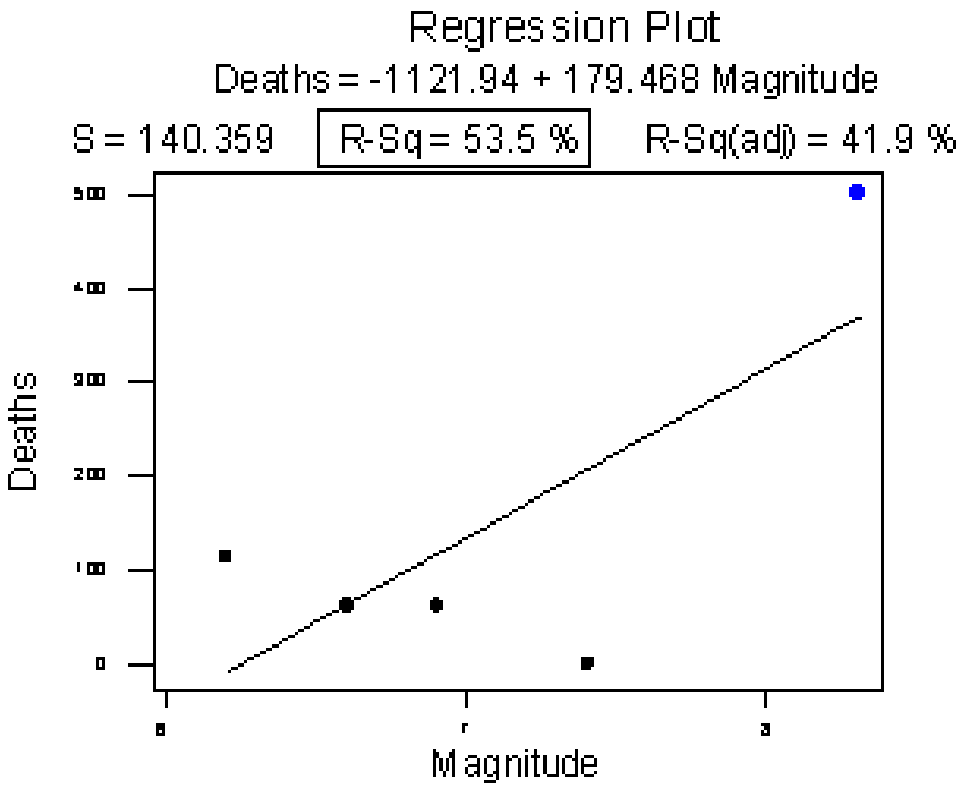


- Based on the residual plot, do you think a linear model is a good fit?

R-Squared, Significance and Residuals - Caution



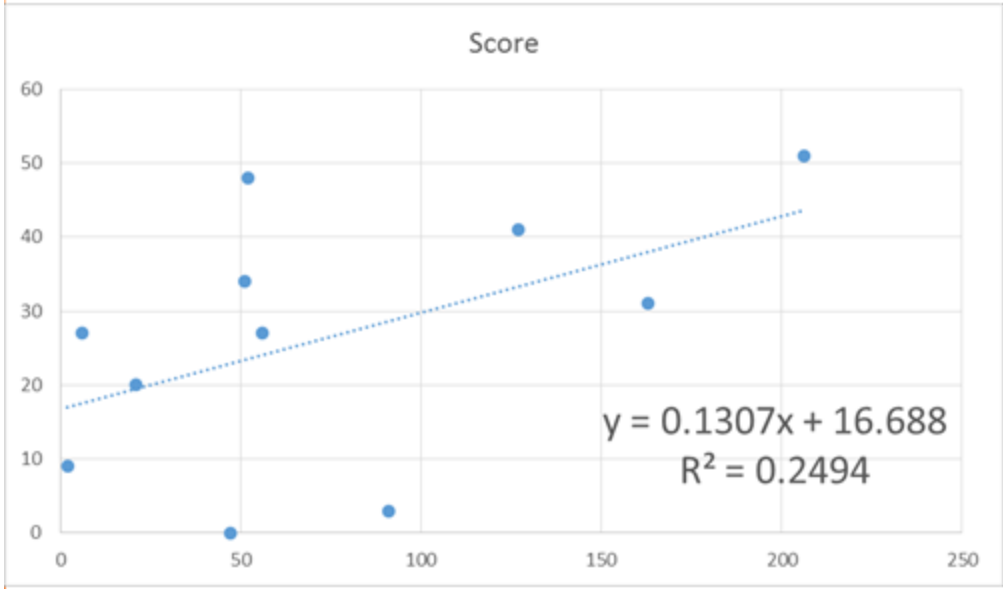
Caution: Single point can change the result

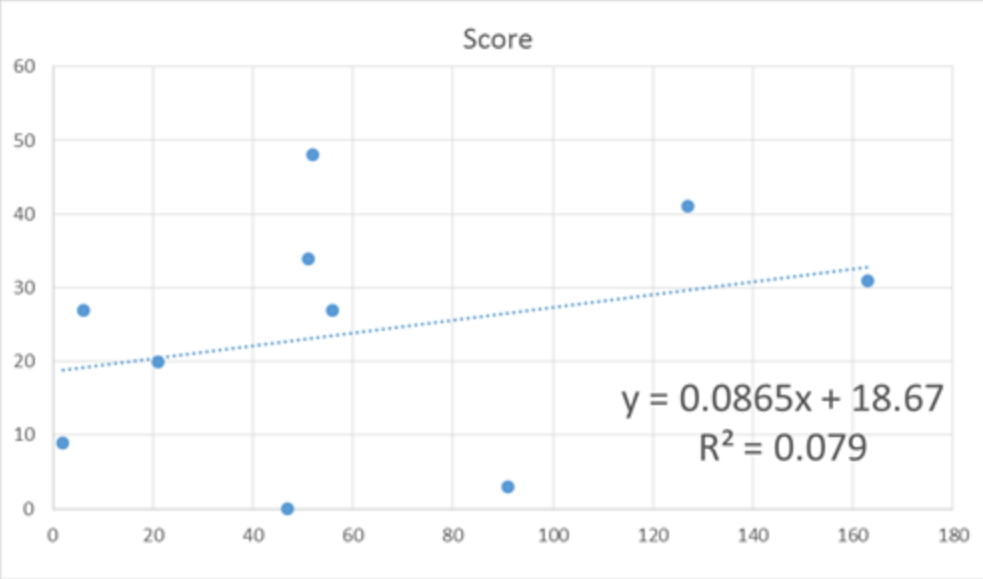


Caution: Single point can change the result

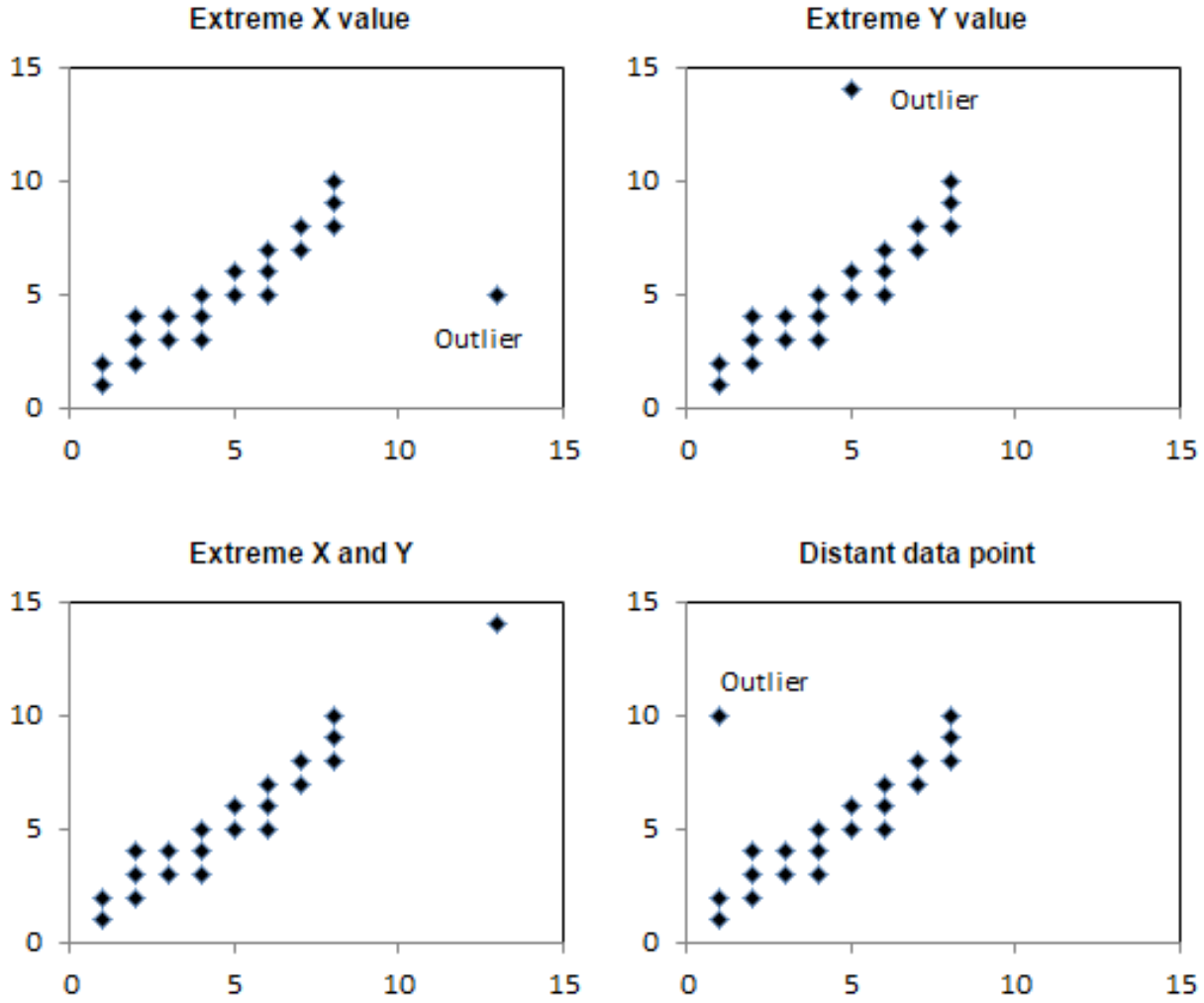
Why it is important to plot.

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





Outliers



Outliers do not follow the general trend of the rest of the data

Outliers typically have a large residual.

Source: http://stattrek.com/regression/influential-points.aspx?Tutorial=AP

Influential Observations - Leverage

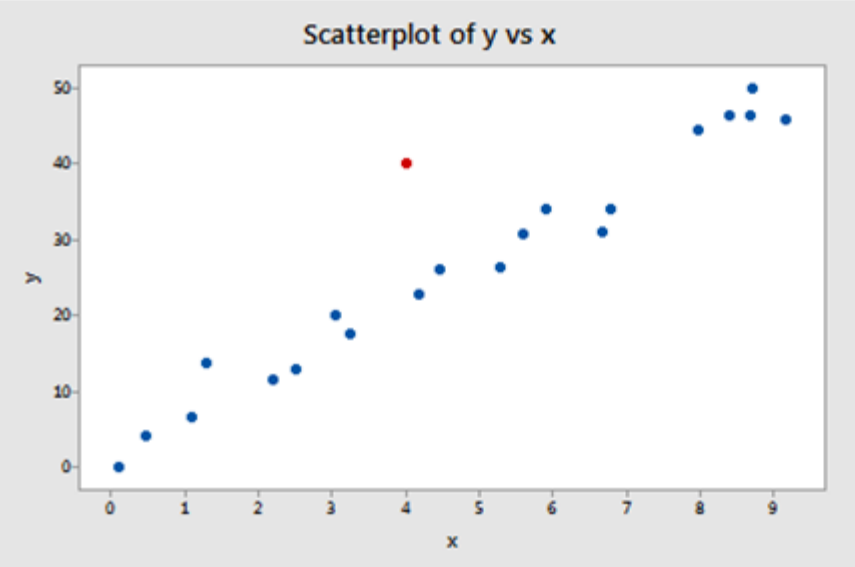
How much the observation's value on the **predictor variable** differs from the mean of the **predictor variable**. That is it tells us about extreme x values, which have the potential to highly influence the regression in certain conditions.

$$Leverage, h = \frac{(Standardized\ predictor\ value)^2 + 1}{n}$$

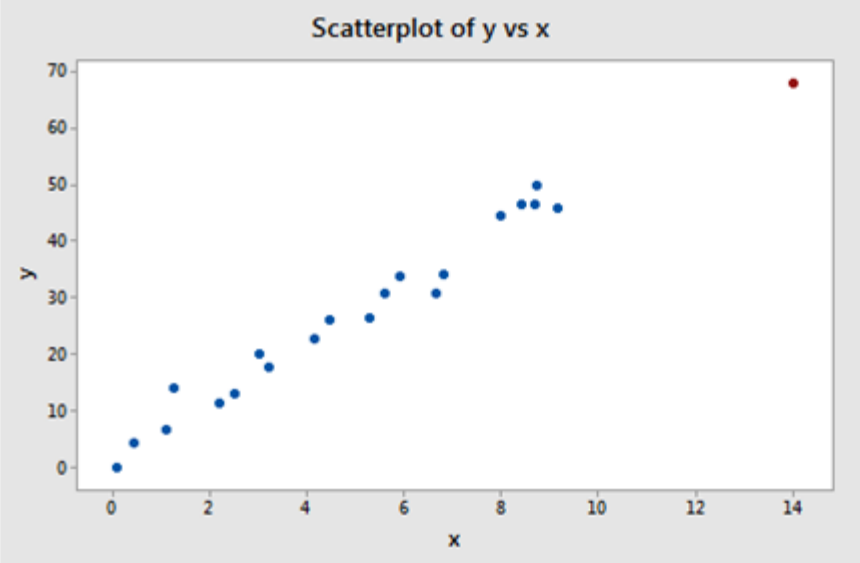
The sum of leverages = # of parameters, p (regression coefficients including intercept).

EXCEL ACTIVITY

Influential Observations - Leverage



Low leverage



High leverage

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

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

Influential Observations

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

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

If Cook's D > 1, the observation can be considered as having too much influence.

Points with Cook's D> 0.5 should be investigated

Influence is a function of leverage and residual.

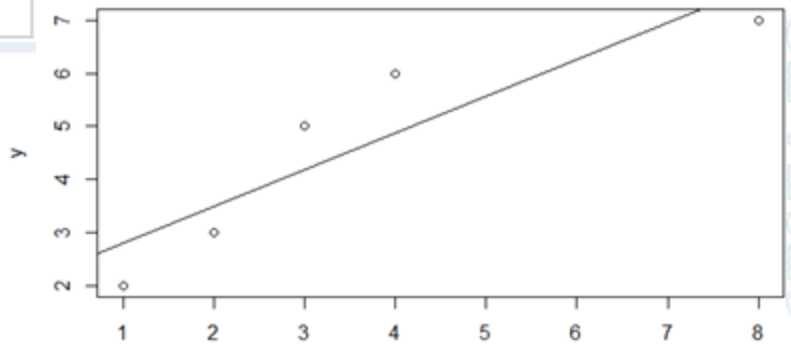
Influential Observations - Distance

Based on error of prediction and is measured by Studentized Residual, which is related to error of prediction of that observation divided by the standard deviation of the errors of prediction.

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

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

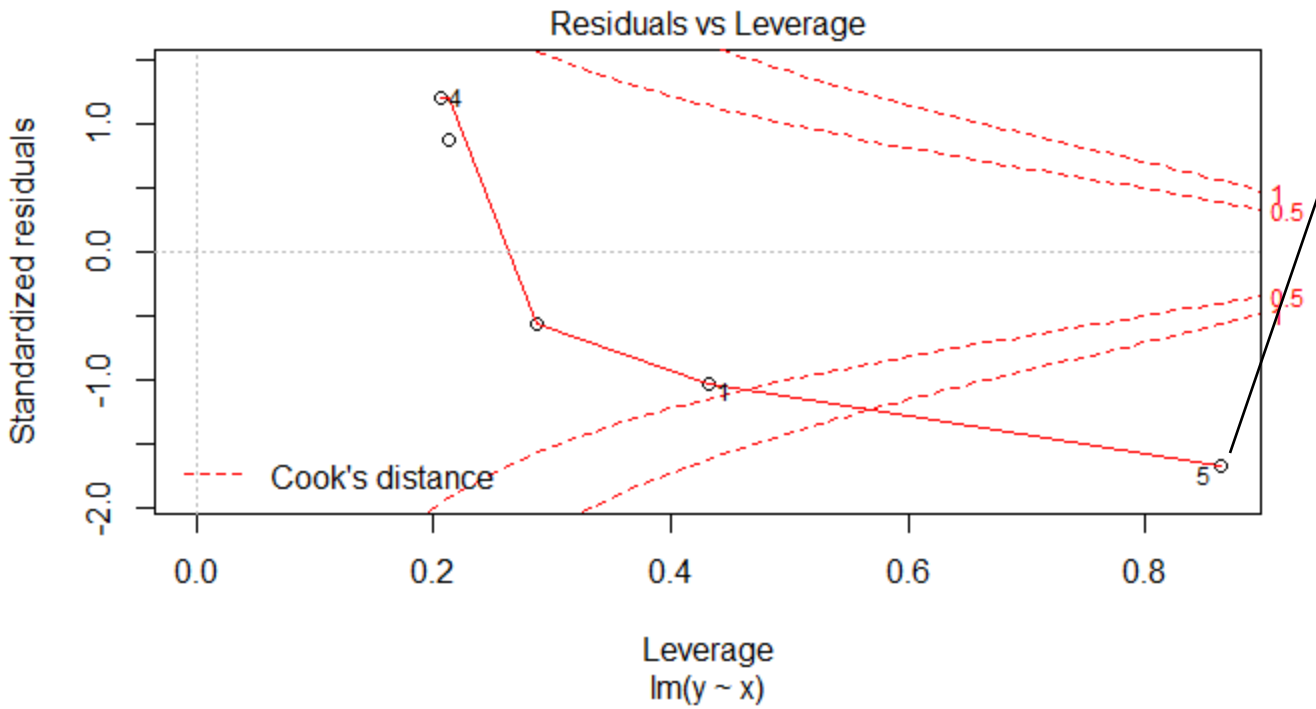
D> 0.5 : Investigate
D>1 : Influential point



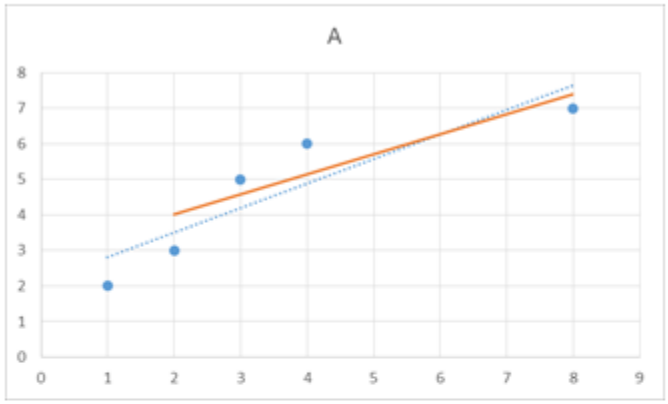
Influential Observations

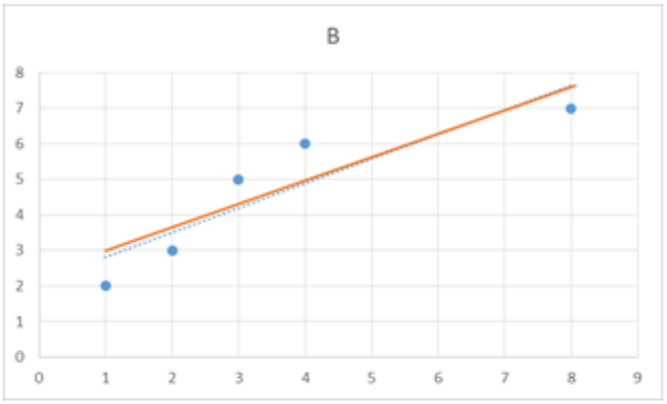
```
> x=c(1,2,3,4,8)
> y=c(2,3,5,6,7)
> lmOut <- lm(y~x)  # Linear Regression
> plot(lmOut, which=5)
```

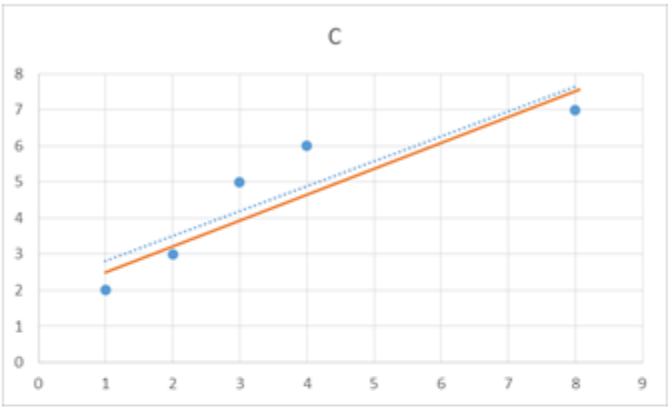
x	y
1	2
2	3
3	5
4	6
8	7

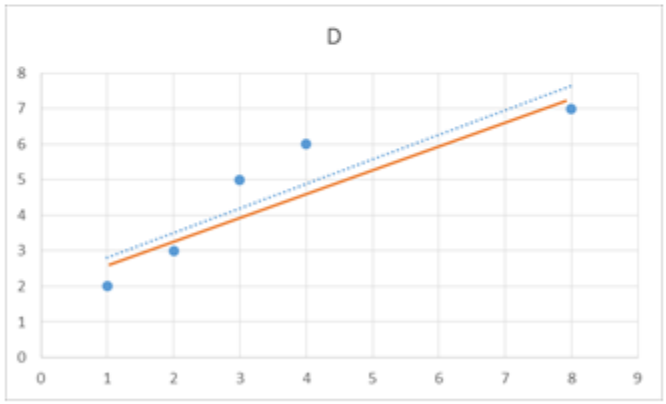


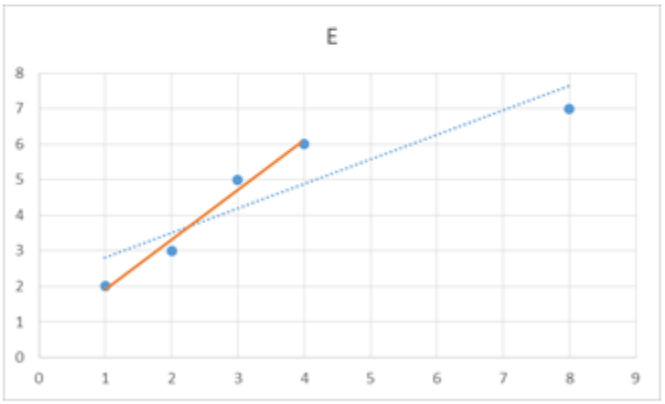
Influential Observations









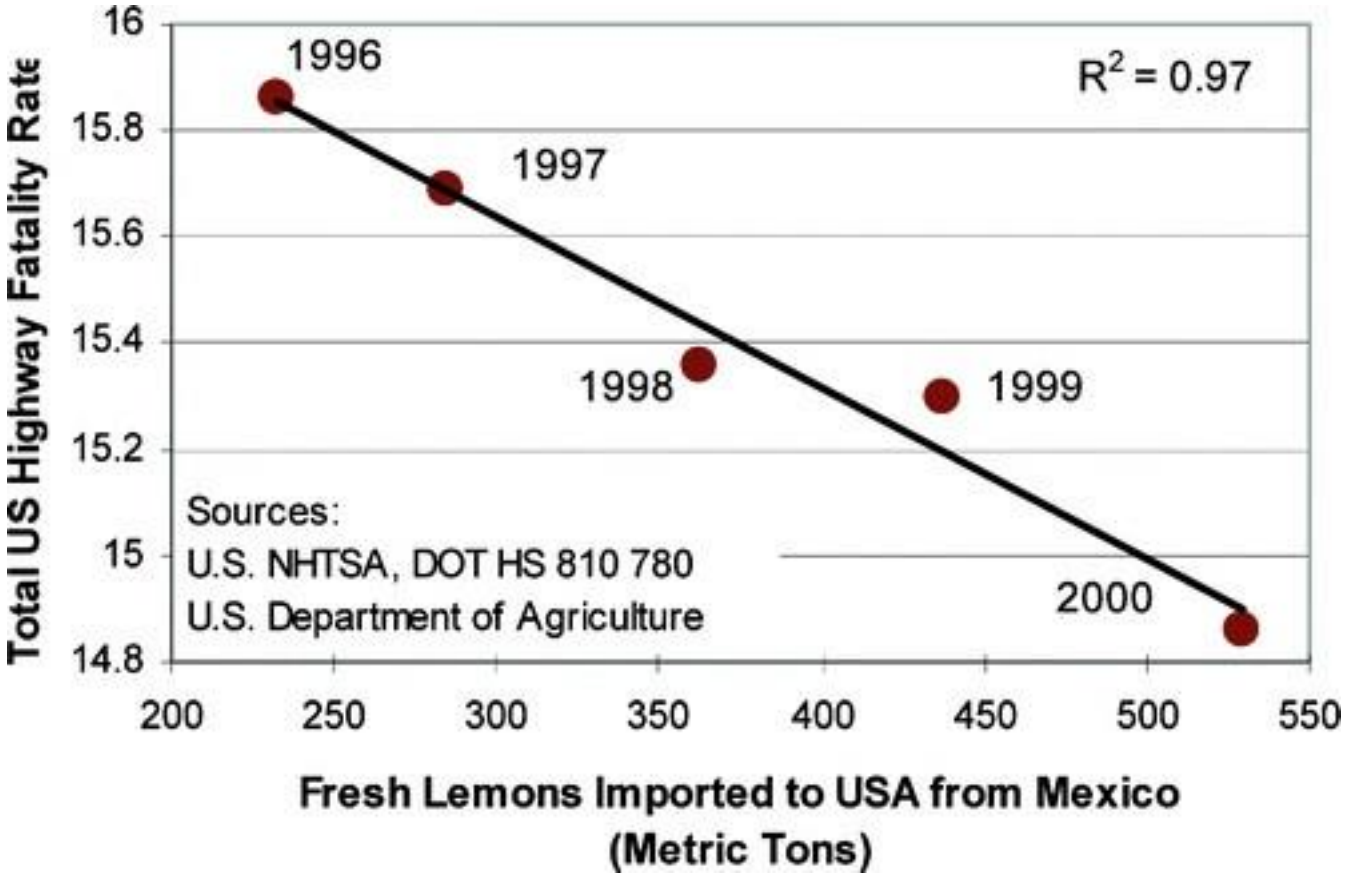


Influential Observations

So what does one do when you find influential observations in your dataset?

- Check if its bad data or there was a procedural error in data collection – delete/correct it
- If data not representative of intended study population– delete it
- Use business intelligence to figure out if different physics or processes involved for the region near the influential point. Maybe a different model applies there.
- Are there other relevant variables that you are ignoring? Redo model with those.
- If unsure – report results with both including the data point and excluding it.

R2 caution: Correlation does not imply causation



Multiple Linear Regression

Multiple Linear Regression

- Linear regression models the effect of one independent variable, x, on one dependent variable, y
- Multiple Regression models the effect of several independent variables, x_1, x_2 etc., on one dependent variable, y
- The different x variables are combined in a linear way and each has its own regression coefficient:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots .. + \beta_n x_n + \varepsilon$$

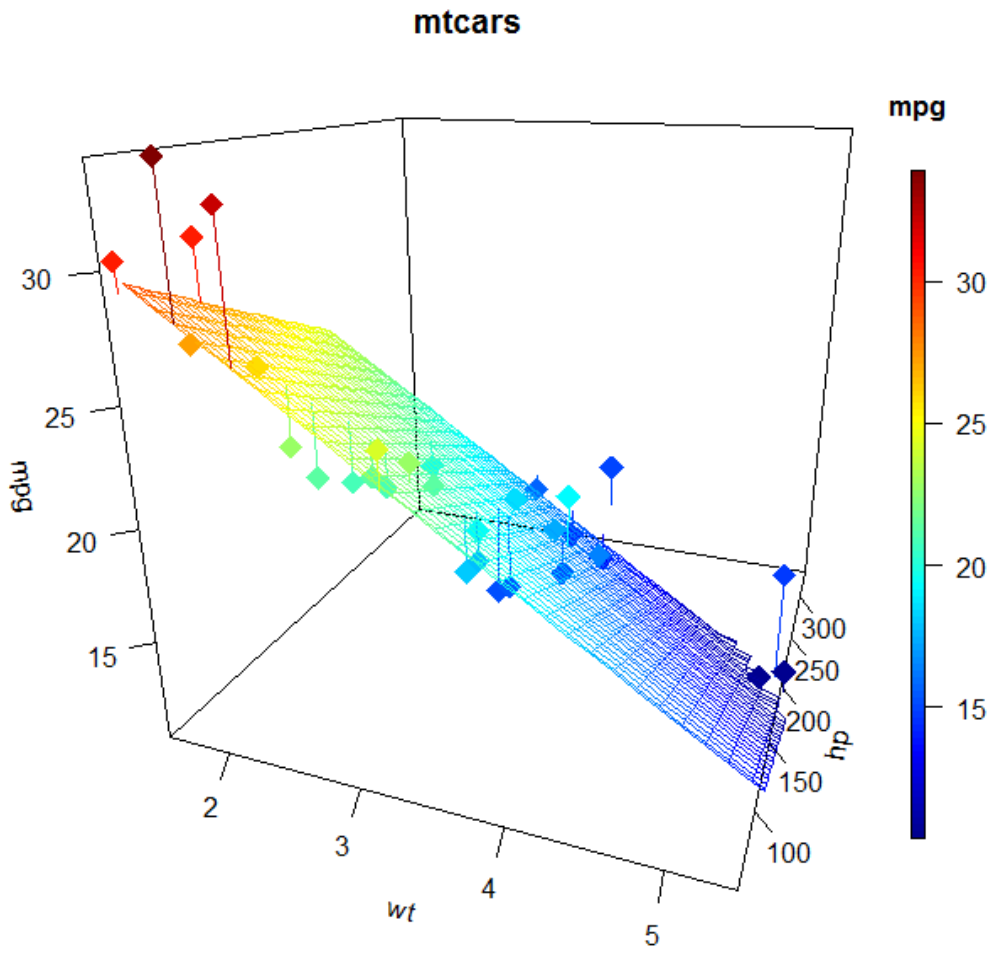
- The β parameters reflect the **independent contribution** of each independent variable, x, to the value of the dependent variable, y.

Cars Dataset (MTcars)

model	mpg	wt	hp	qsec
Mazda RX4	21	2.62	110	16.46
Mazda RX4 Wag	21	2.875	110	17.02
Datsun 710	22.8	2.32	93	18.61
Hornet 4 Drive	21.4	3.215	110	19.44
Hornet Sportabout	18.7	3.44	175	17.02
Valiant	18.1	3.46	105	20.22
Duster 360	14.3	3.57	245	15.84
Merc 240D	24.4	3.19	62	20
Merc 230	22.8	3.15	95	22.9
Merc 280	19.2	3.44	123	18.3
Merc 280C	17.8	3.44	123	18.9
Merc 450SE	16.4	4.07	180	17.4
Merc 450SL	17.3	3.73	180	17.6
Merc 450SLC	15.2	3.78	180	18
Cadillac Fleetwood	10.4	5.25	205	17.98
Lincoln Continental	10.4	5.424	215	17.82
Chrysler Imperial	14.7	5.345	230	17.42
Fiat 128	32.4	2.2	66	19.47
Honda Civic	30.4	1.615	52	18.52

Mpg=Miles/gallon
Wt = weight
Hp = horsepower
Qsec=time to go cover a quarter mile from start

Mpg vs (wt,hp)



A coefficient is the slope of the linear relationship between the dependent variable (DV) and the **independent contribution** of the independent variable (IV), i.e., that part of the IV that is independent of (or uncorrelated with) all other IVs.

SUMMARY OUTPUT [mpg explained by (wt,hp,qsec)

Regression Statistics	
Multiple F	0.91901
R Square	0.84458
Adjusted I	0.813496
Standard I	2.479119
Observati	19

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressio	3	500.979	166.993	27.17087	2.60039E-06
Residual	15	92.19047	6.146031		
Total	18	593.1695			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	31.66942	10.30754	3.072451	0.007741	9.699414142	53.63942	9.699414	53.63942
wt	-3.07168	1.187317	-2.58707	0.020623	-5.602383601	-0.54097	-5.60238	-0.54097
hp	-0.03806	0.024587	-1.54815	0.142423	-0.090471112	0.014342	-0.09047	0.014342
qsec	0.204452	0.538432	0.379717	0.709477	-0.943188099	1.352092	-0.94319	1.352092

Interpreting Regression Coefficients

How do we find the **independent contribution** of any independent variable?

Step 1: Predict this IV using the other IVs.

Step 2: Find residuals. The difference in the actual values and the predicted values give the errors or residuals, which could not be predicted by the other IVs. This means the residual errors are the independent part of this IV not related to other IVs.

Step 3: Predict the dependent variable (DV) using the residuals obtained in Step 2. The coefficient (slope) obtained for the residuals (the independent part of the IV in question) gives the independent contribution of this IV when all others are kept constant.

Interpreting Regression Coefficients

Let us First Extract amount of *Qsec* explained by other 2 variables (*wt,hp*)

SUMMARY OUTPUT qsec explained by (wt,hp)

Regression Statistics	
Multiple F	0.73454475
R Square	0.53955599
Adjusted I	0.48200049
Standard E	1.1510833
Observati	19

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	2	24.84238	12.42119	9.374534	0.00202
Residual	16	21.19988	1.324993		
Total	18	46.04226			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	18.7358015	0.982617	19.06725	2E-12	16.65275	20.81886	16.65275	20.81886
wt	1.21292633	0.460397	2.634521	0.018029	0.236928	2.188925	0.236928	2.188925
hp	-0.0328228	0.007937	-4.13547	0.000777	-0.04965	-0.016	-0.04965	-0.016

Qsec predicted from (wt,hp)

The part of Qsec unexplained by (wt,hp)

qsec	wt	hp	Qsec-Pred	Qsec-Err
16.46	2.62	110	18.3031575	-1.84316
17.02	2.875	110	18.6124537	-1.59245
18.61	2.32	93	18.4972676	0.112732
19.44	3.215	110	19.0248486	0.415151
17.02	3.44	175	17.1642733	-0.14427
20.22	3.46	105	19.4861297	0.73387
15.84	3.57	245	15.0243557	0.815644
20	3.19	62	20.5700212	-0.57002
22.9	3.15	95	19.4383508	3.461649
18.3	3.44	123	18.8710603	-0.57106
18.9	3.44	123	18.8710603	0.02894
17.4	4.07	180	17.7643027	-0.3643
17.6	3.73	180	17.3519078	0.248092
18	3.78	180	17.4125541	0.587446
17.98	5.25	205	18.3749851	-0.39499
17.82	5.424	215	18.257806	-0.43781
17.42	5.345	230	17.6696424	-0.24964
19.47	2.2	66	19.2379328	0.232067
18.52	1.615	52	18.9878905	-0.46789

Interpreting Regression Coefficients

Now we compute regression of *mpg* vs *Qsec-err*

SUMMARY OUTPUT mpg as a function of Qsec-Err

Regression Statistics	
Multiple F	0.038652
R Square	0.001494
Adjusted I	-0.05724
Standard E	5.902558
Observati	19

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	1	0.886168	0.886168	0.025435	0.875167
Residual	17	592.2833	34.84019		
Total	18	593.1695			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	19.40526	1.35414	14.33033	6.38E-11	16.54828	22.26225	16.54828	22.26225
Qsec-Err	0.204452	1.281957	0.159484	0.875167	-2.50024	2.909145	-2.50024	2.909145

	Coefficients	*andard Err*	*t Stat*	*P-valu*
Intercept	31.66942	10.30754	3.072451	0.0077
wt	-3.07168	1.187317	-2.58707	0.0206
hp	-0.03806	0.024587	-1.54815	0.1424
qsec	0.204452	0.538432	0.379717	0.7094

Assumptions of Multiple Linear Regression

- Same as simple linear regression
  - Linearity
  - Independence of errors
  - Homoscedasticity (constant variance)
  - Normality of errors

- Methods of checking assumptions are also the same

Determining the Multiple Regression Equation

- $k+1$ equations to solve for k independent variables and the intercept.
- In solving for intercept and slope in a simple linear regression model, we needed $\sum x$, $\sum y$, $\sum xy$, and $\sum x^2$.
- For multiple regression model with 2 independent variables, we need $\sum x_1$, $\sum x_2$, $\sum y$, $\sum {x_1}^2$, $\sum {x_2}^2$, $\sum x_1 x_2$, $\sum x_1 y$, and $\sum x_2 y$.

Determining the Multiple Regression Equation - Excel

In a real estate study, multiple variables were explored to determine the price of a house.

- # of bedrooms
- # of bathrooms
- Age of the house
- # of square feet of living space
- Total # of square feet of space
- # of garages

Find the equation if you want to predict the price of the house by total square feet and age of the house.

Determining the multiple regression equation – Interpreting the output

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.860872681					
R Square	0.741101773					
Adjusted R Square	0.715211951					
Standard Error	11.96038667					
Observations	23					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	8189.723012	4094.861506	28.62521631	1.35298E-06	
Residual	20	2861.016988	143.0508494			
Total	22	11050.74				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	57.35074586	10.00715186	5.73097587	1.31298E-05	36.47619286	78.22529885
Area (sq ft) (x1)	0.017718036	0.00314562	5.632605205	1.63535E-05	0.011156388	0.024279685
Age of House (years) (x2)	-0.666347946	0.227996703	-2.922620973	0.008417613	-1.141940734	-0.190755157

What is the equation?

$$\hat{y} = 57.35 + 0.0177Area - 0.666Age$$

Are the coefficients and the model significant?

Yes

Residuals

Residuals are determined the same way as in simple linear regression. The predicted value is calculated by substituting the predictor values of interest. The residual is again the difference between the observed and the predicted values, $y - \hat{y}$.

SSE and Standard Error of the Estimate, SE – Practice Assignment

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

$$SE = \sqrt{\frac{SSE}{n - k - 1}}$$

Coefficient of Multiple Determination, R^2 – Practice Assignment

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

Adjusted R^2 - Excel

As additional independent variables are added to the regression model, the value of R^2 increases.

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

However, sometimes these variables are insignificant and add no real value, yet inflating the R^2 value.

Adjusted R^2 takes into consideration both the additional information and the changed degrees of freedom.

$$Adjusted\ R^2 = 1 - \frac{\frac{SSE}{(n-k-1)}}{\frac{SS_{yy}}{n-1}} = R^2 - (1-R^2)\frac{k}{n-k-1} = 1 - \frac{MSE}{MS_{yy}}$$

Sample R Output

```
Call:
lm(formula = CONSUME ~ PRICE + INC + TEMP + PRICEINCi, data = datavar)

Residuals:
      Min         1Q     Median         3Q        Max
-0.0575279 -0.0163589 -0.0008483  0.0168662  0.0718922

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1570203  0.2324673   0.675   0.5058
PRICE       -0.1636906  0.7438870  -0.220   0.8277
INC          0.0012301  0.0012133   1.014   0.3208
TEMP         0.0028231  0.0004171   6.769 5.31e-07 ***
PRICEINCi   -0.2786003  0.1344397  -2.072   0.0491 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03094 on 24 degrees of freedom
Multiple R-squared: 0.7411, Adjusted R-squared: 0.698
F-statistic: 17.18 on 4 and 24 DF,  p-value: 8.968e-07
```

International School of Engineering

Plot 63/A, Floors 1&2, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals: +91-9502334561/63 or 040-65743991

For Corporates: +91-9618483483

Web: http://www.insofe.edu.in

Facebook: https://www.facebook.com/insofe

Twitter: https://twitter.com/Insofeedu

YouTube: http://www.youtube.com/InsofeVideos

SlideShare: http://www.slideshare.net/INSOFE

LinkedIn: http://www.linkedin.com/company/international-school-of-engineering