Statistics and Probability in Decision Modeling

Linear Regression

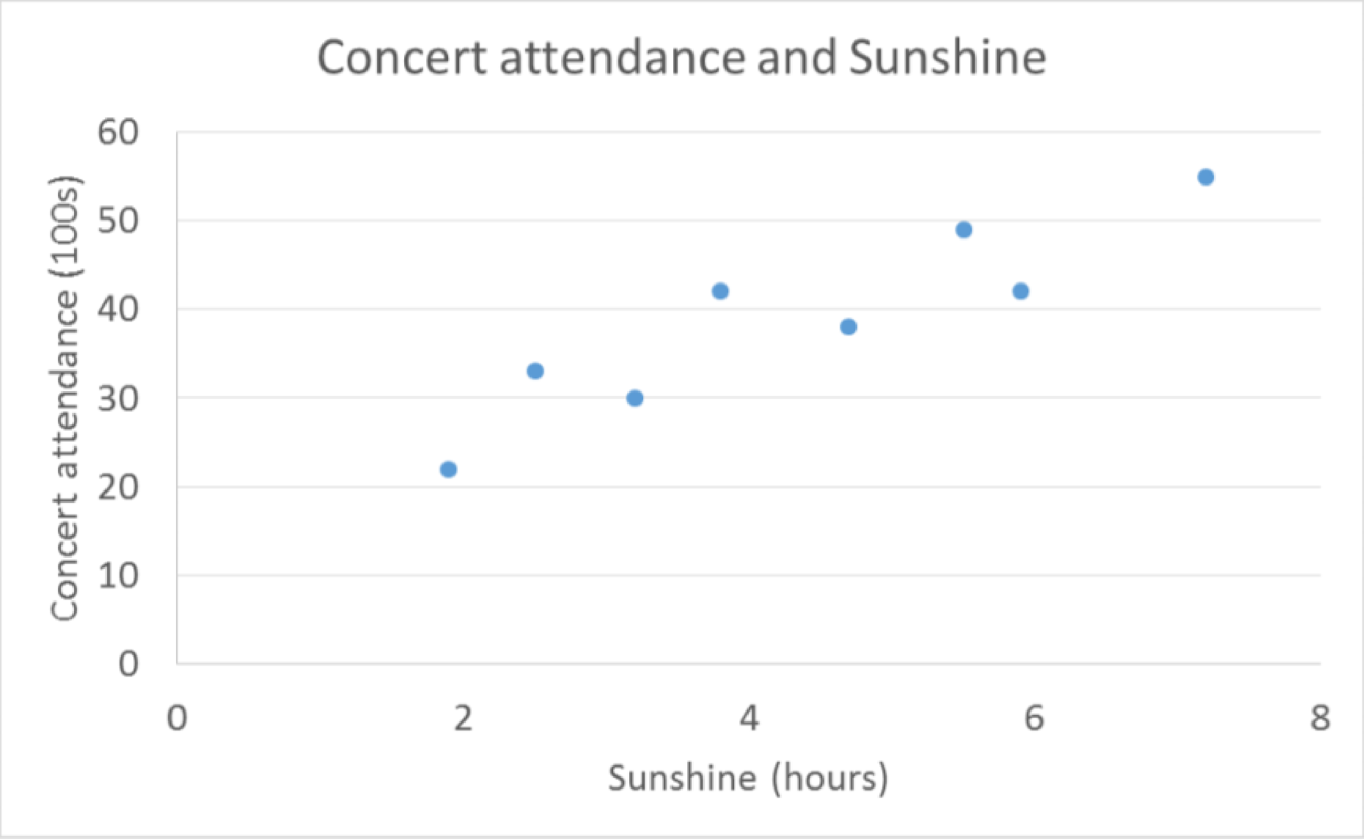
CORRELATION, COVARIANCE AND REGRESSION

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

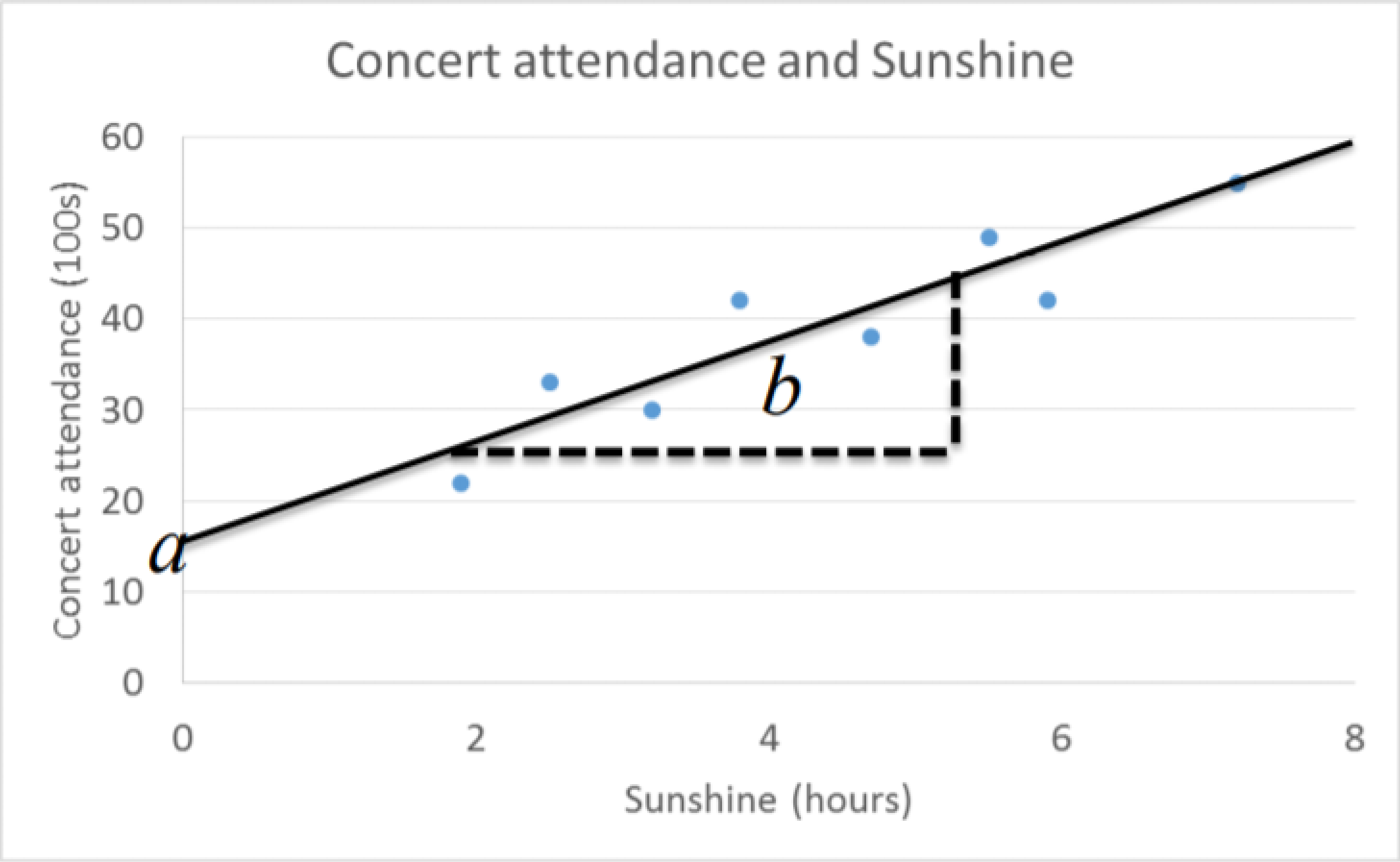
- The band makes a loss if less than 3500 people attend.
- Based on predicted hours of sunshine, can we predict ticket sales?
- Are sunshine and concert attendance correlated?

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

- Independent variable (explanatory) – Sunshine – Plotted on X-axis
- Dependent variable (response) – Concert attendance – Plotted on Y-axis



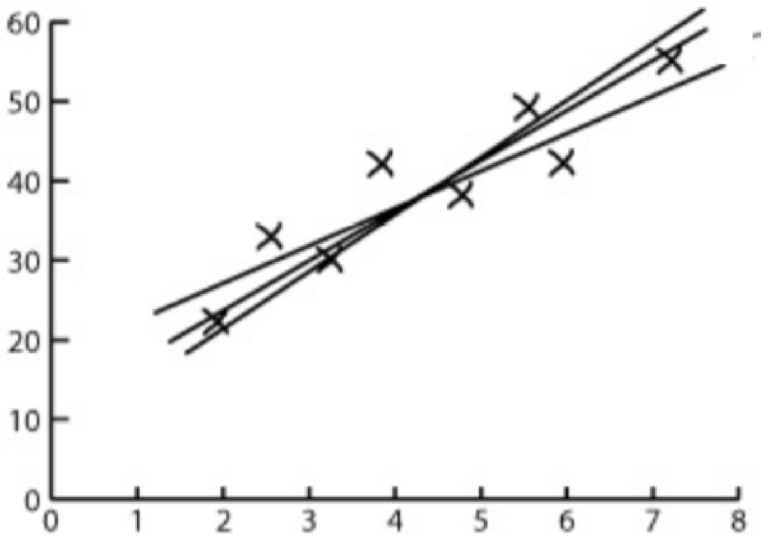
We need to find the equation of the line.

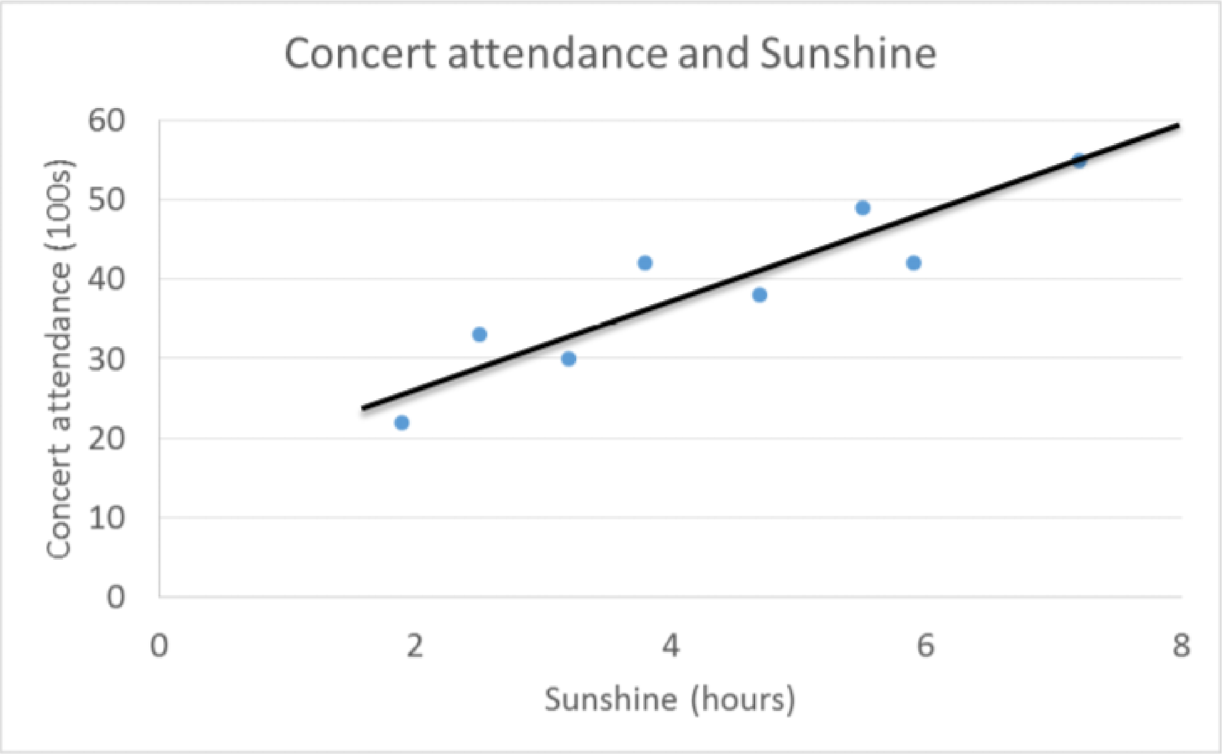


$y = a + bx$

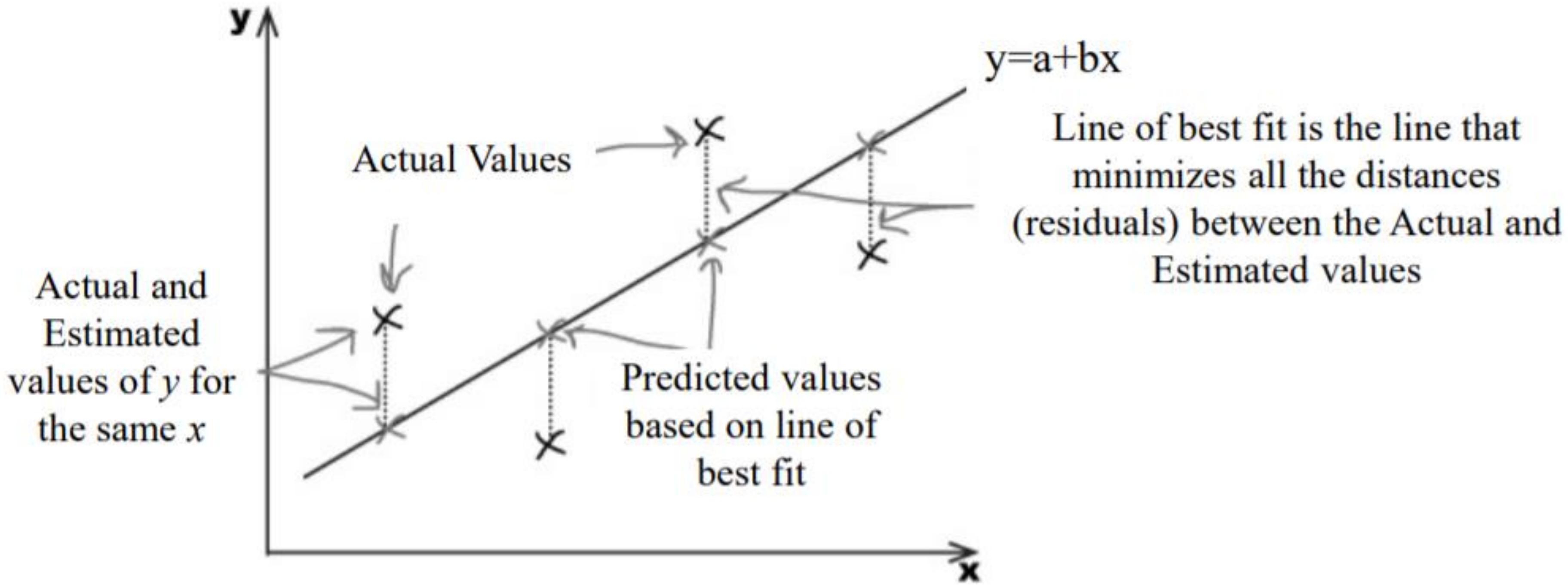
Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

- Line of best fit





We need to minimize errors.



We could do that by minimizing $\sum(y_i - \widehat{y_i})$, where y_i is the actual value and $\widehat{y_i}$ its estimate. $(y_i - \widehat{y_i})$ 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 $\widehat{y_i}$.*

$$SSE = \sum (y_i - \widehat{y_i})^2$$

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

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

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

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

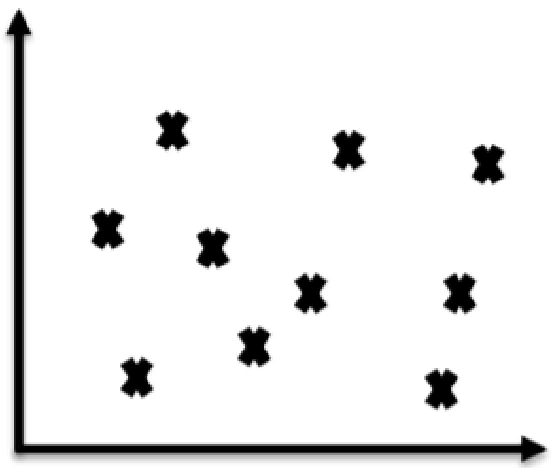
$$b = \frac{\sum((x - \bar{x})(y - \bar{y}))}{\sum(x - \bar{x})^2}$$

How do you calculate a? The line of best fit must 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**.

But how do you know how accurate this line is?

Accurate Linear
Correlation



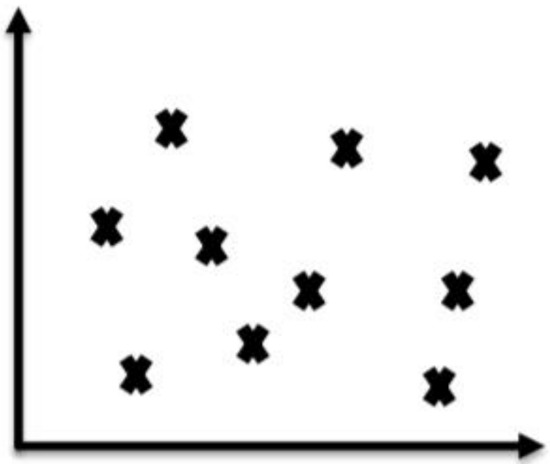
No Linear
Correlation

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

- Hours of sunshine and concert attendance are correlated, i.e., in general, longer sunshine hours indicate higher attendance.

Positive Linear Correlation

Negative Linear Correlation



No Correlation

Correlation Coefficient

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

r = 1 r = -1 r =0

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

Correlation Coefficient

$r = \frac{bs_x}{s_y}$ where b is the slope of the line of best fit, s_x is the standard deviation of the x values in the sample, and s_y is the standard deviation of the y values in the sample.

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

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

Find r for this data. r = 0.916

Correlation Coefficient and Covariance

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

$b = \frac{s_{xy}^2}{{s_x}^2}$ and so, $r = \frac{s_{xy}^2}{s_x s_y}$ *(Recall* $b = \frac{\sum((x-\bar{x})(y-\bar{y}))}{\sum(x-\bar{x})^2}$ *and* $r = \frac{b s_x}{s_y}$*)*.

Covariance

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

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

Covariance and Correlation

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

- The value of covariance itself doesn't say much. It only shows whether the variables are moving together (positive value) or opposite to each other (negative value).
  - Affected by units (measuring height in ft vs mm)
  - Not intuitive comparing covariance values between 2 sets of variables (how does height-weight covariance compare with oil price($)-potato price (Rupee) covariance)
  - Unintuitive units

Covariance and Correlation

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

- To know the strength of how the variables move together, covariance is standardized to the dimensionless quantity, correlation.

Coefficient of Determination

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.

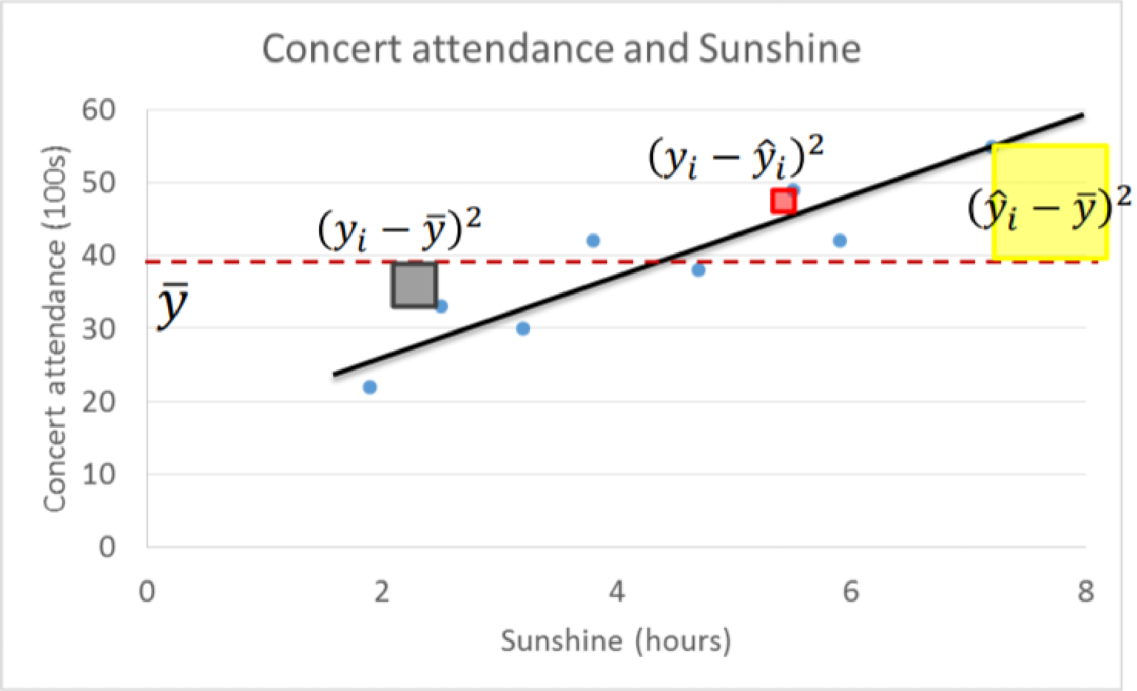
If $r^2 = 0$, it means you can't predict the y value from the x value.
If $r^2 = 1$, it means you can predict the y value from the x value without any errors.

Usually, r^2 is between these two extremes.

Coefficient of Determination

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

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



Covariance, Correlation and R^2

How do the interest rates of federal funds and the commodities futures index co-vary and correlate?

Day	Interest Rate	Futures Index
1	7.43	221
2	7.48	222
3	8.00	226
4	7.75	225
5	7.60	224
6	7.63	223
7	7.68	223
8	7.67	226
9	7.59	226
10	8.07	235
11	8.03	233
12	8.00	241

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

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

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

Welcome to the Learning Models

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

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

Why Modeling

- 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 will he stay in the hospital; etc.

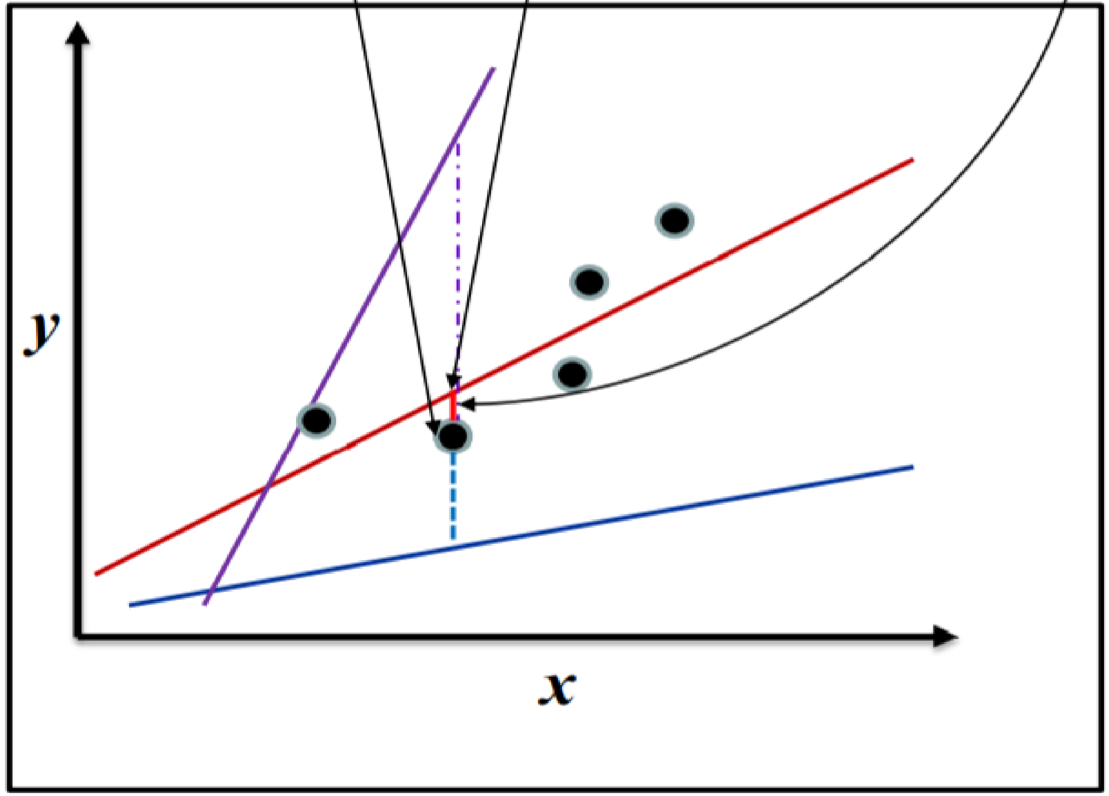
- <u>Supervised learning</u> is about computing the latter using the former

Linear Regression

How to Pick the Best Model?

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

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



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

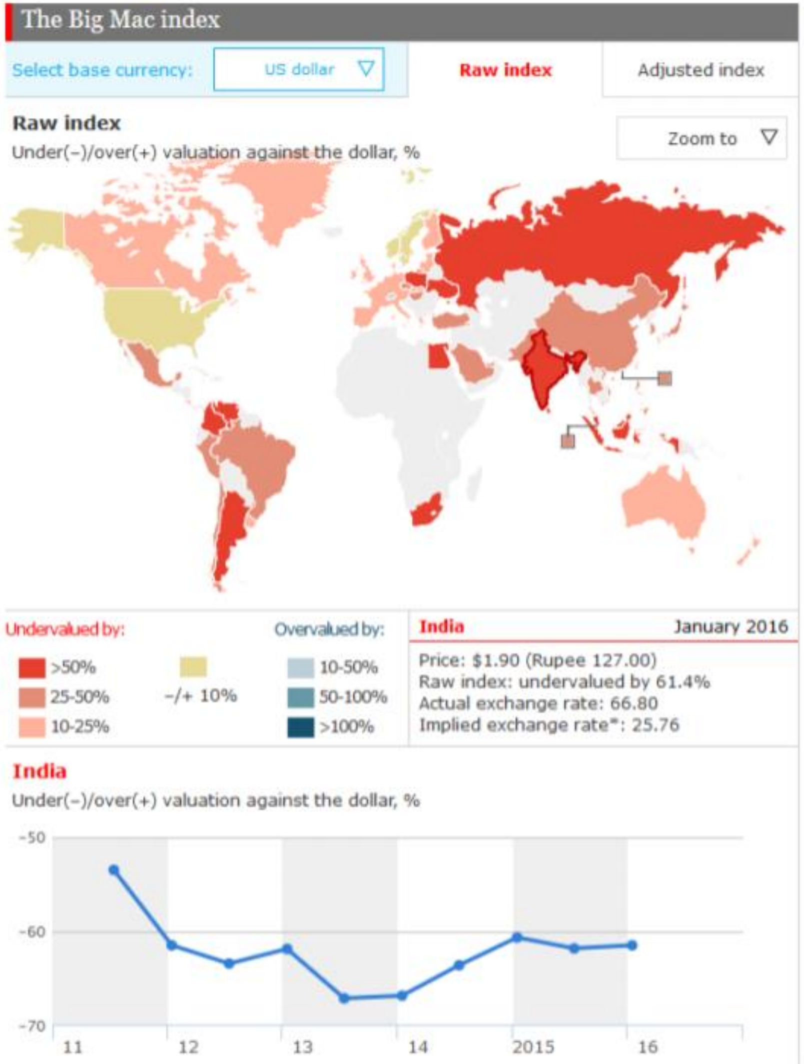
To ensure residual errors don't cancel, we take squares of residual errors.

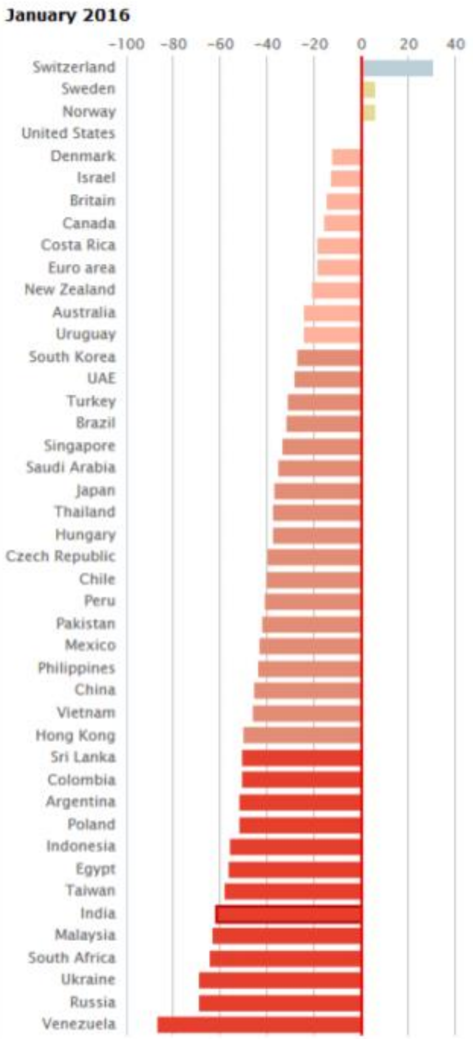
Burgernomics: Overvalued or Undervalued Currencies?

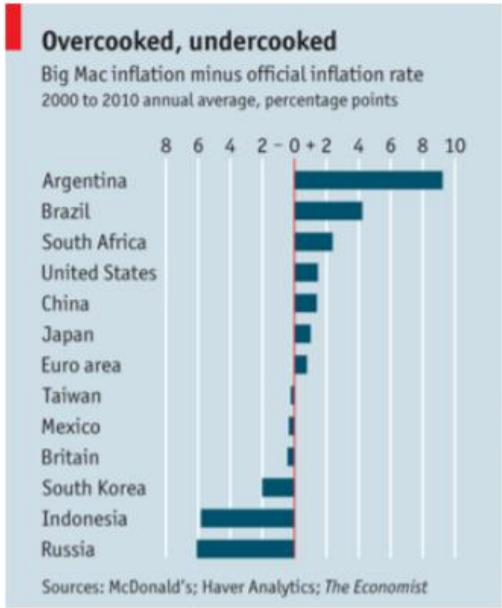
- Big Mac price in the US: \$ 4.93
- Maharaja Mac price in India: Rs 155
- Implied PPP is 155/4.93 = Rs 31.44/\$
- Actual exchange rate = Rs 67.2959/\$
- $\frac{31.44-67.2959}{67.2959} = -0.53$
- Rupee undervalued by 53% against the USD

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

Burgernomics: Overvalued or Undervalued Currencies?





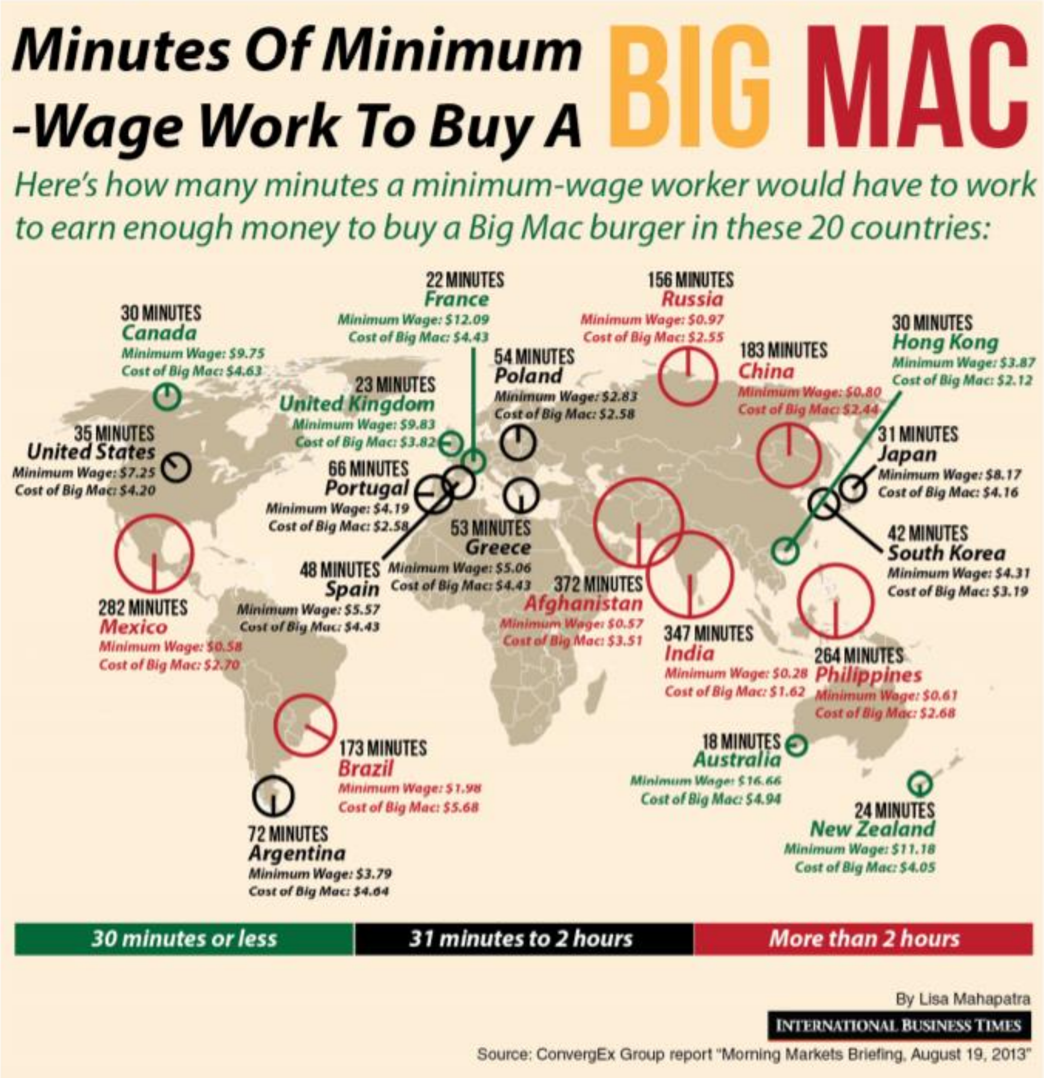


Source: Lies, flame-grilled lies and statistics

http://www.economist.com/node/18014576?story_id=18014576

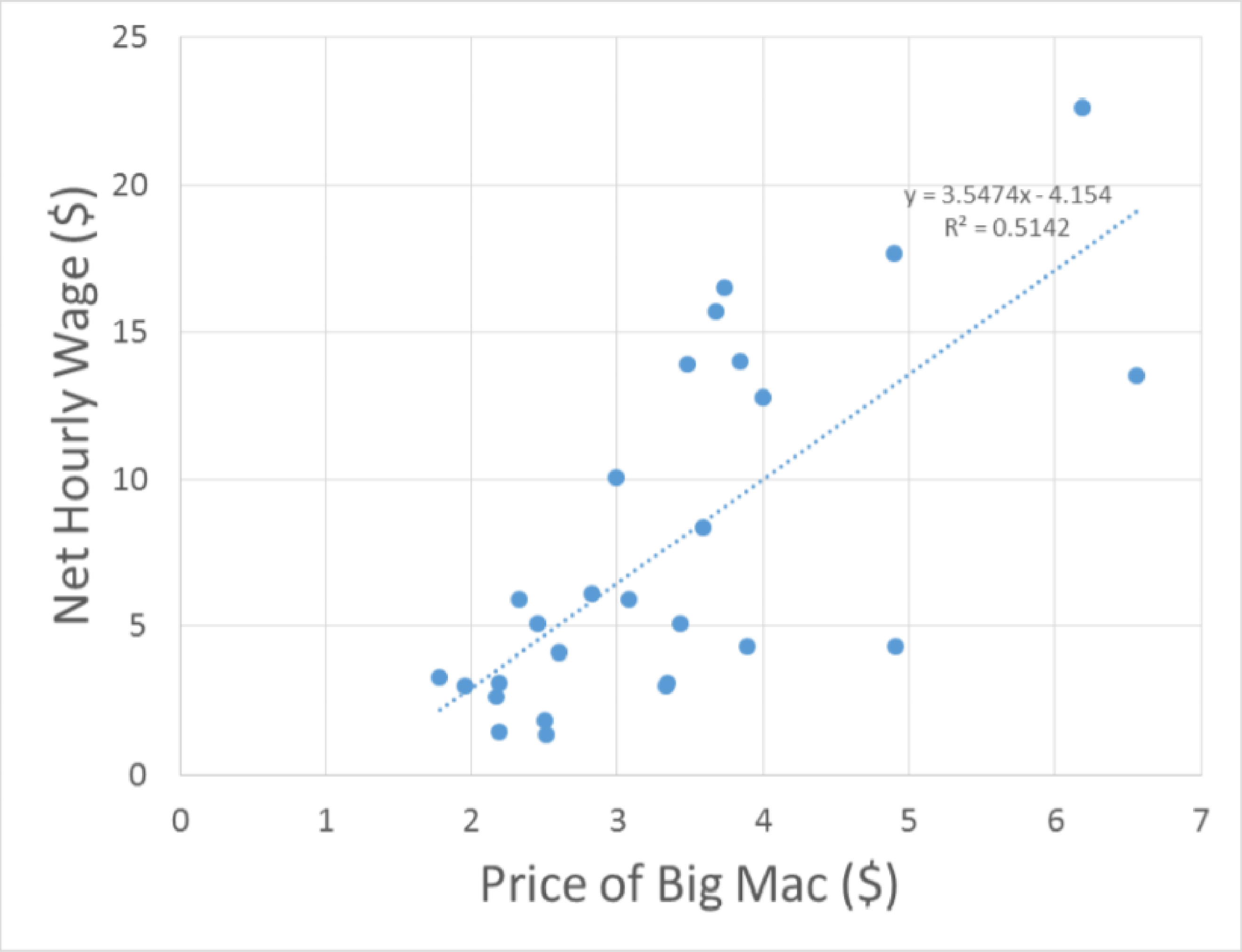
Last accessed: March 04, 2016

Burgernomics by UBS Wealth Management Research



Determining the Equation of the Regression Line - Excel

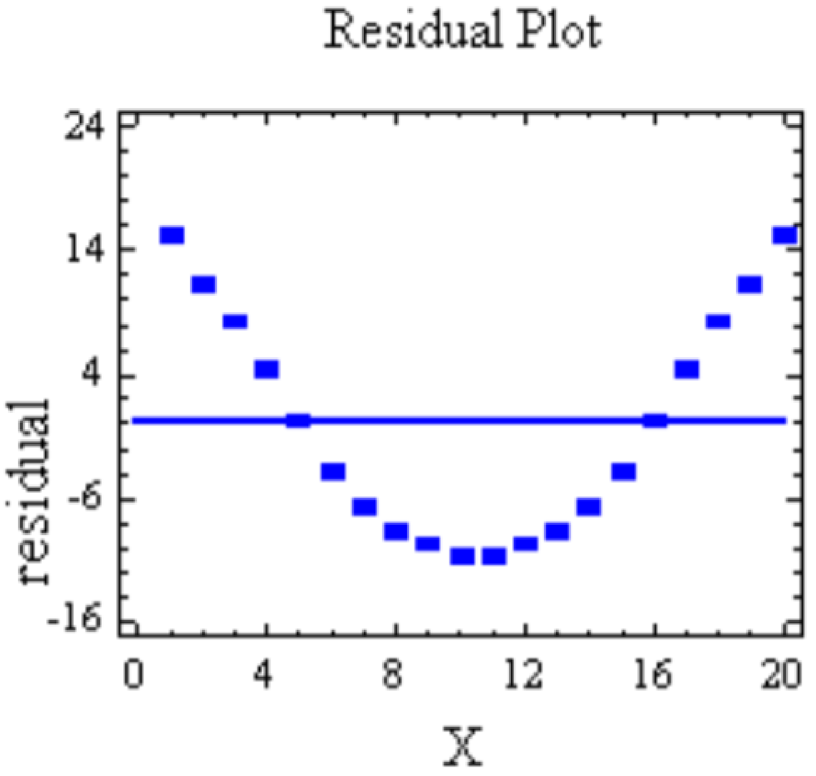
Determining the Equation of the Regression Line - Excel



WAYS OF TESTING HOW WELL THE REGRESSION LINE FITS DATA

Assumptions of the Regression Model

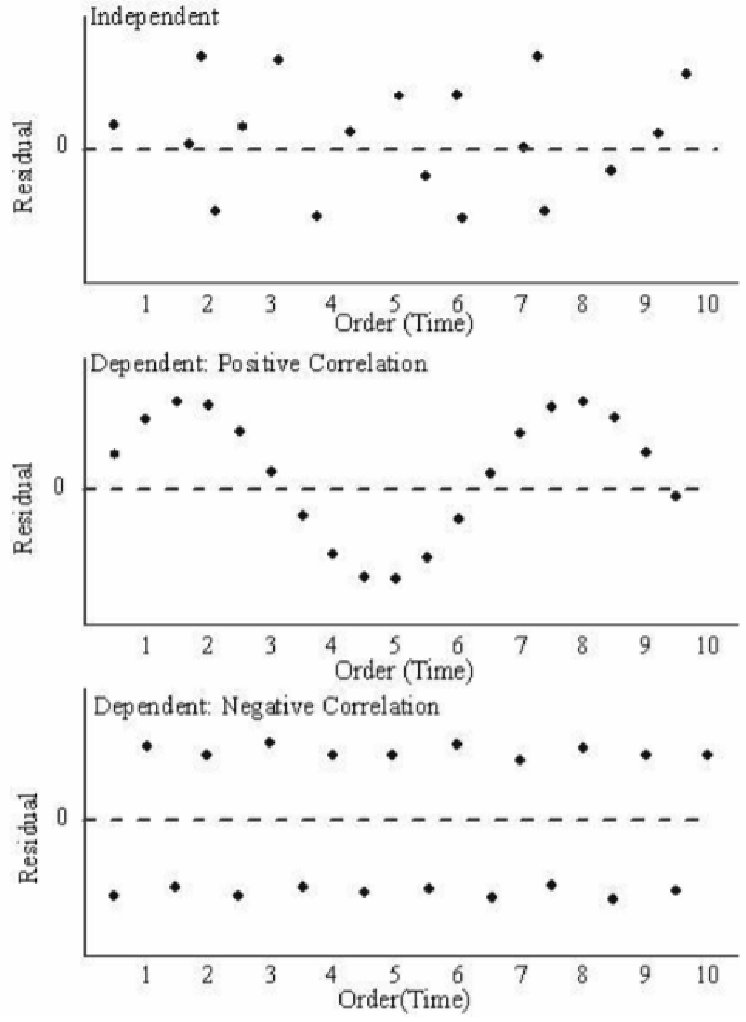
- The model is linear



Zero residual line:
The regression line

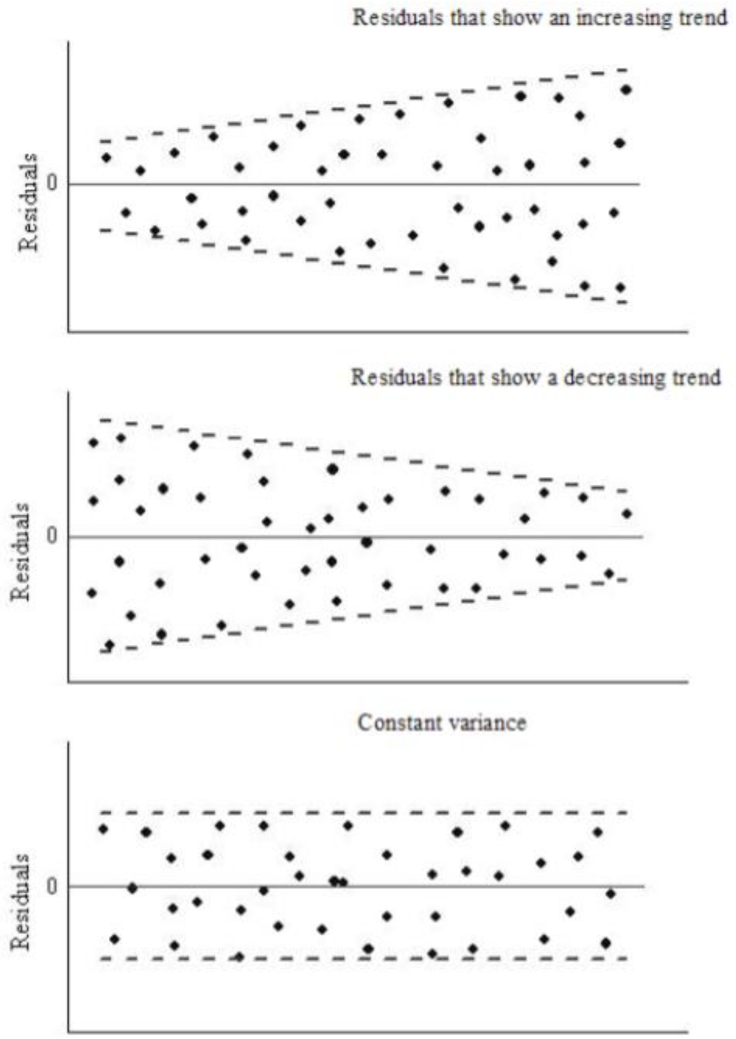
Assumptions of the Regression Model

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



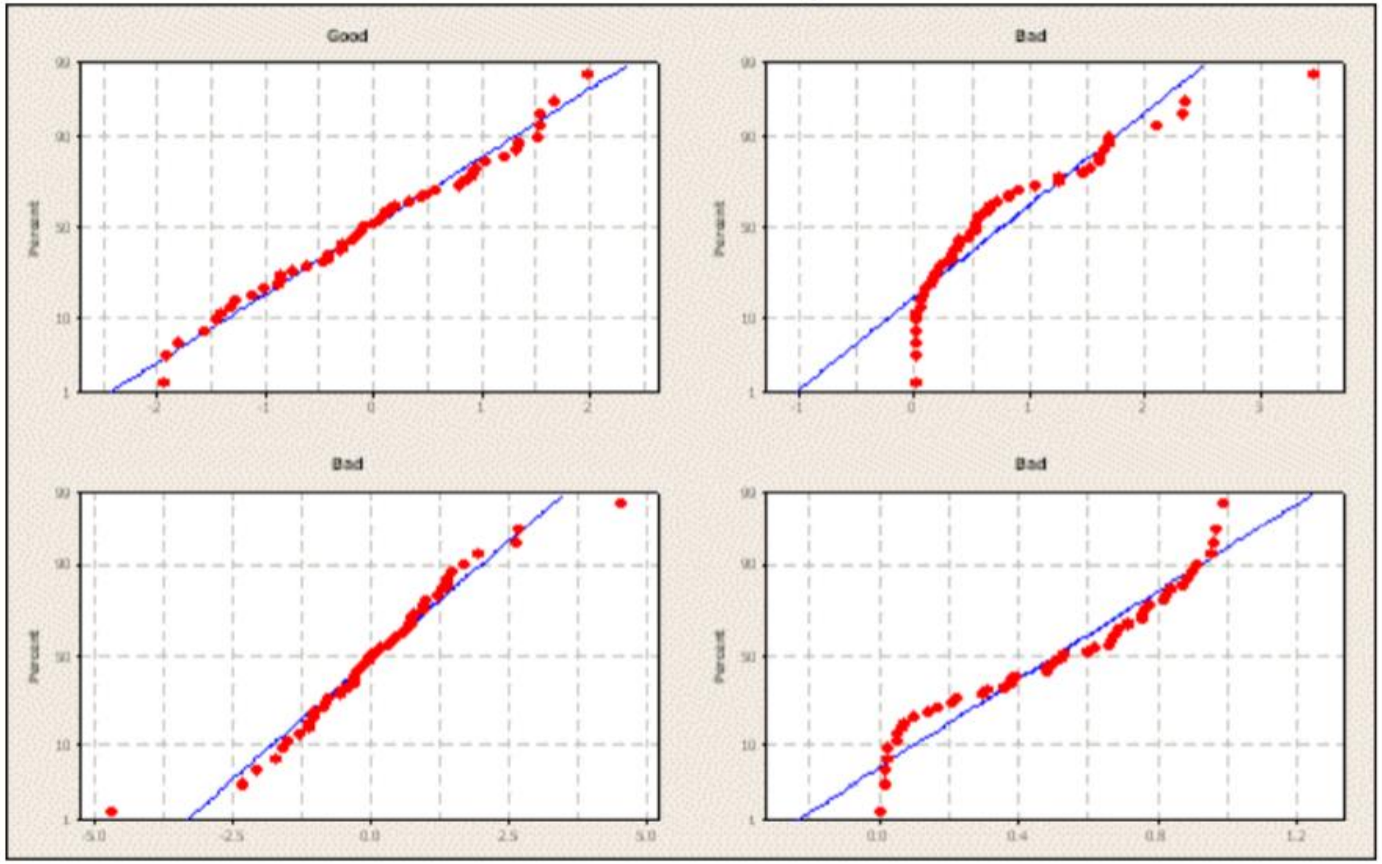
Assumptions of the Regression Model

- 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 error terms are normally distributed



Fixing Non-normality and Heteroscedasticity

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

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

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

Testing the Slope

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

Testing the Slope

What is the Null Hypothesis?

$H_0: \beta_1 = 0$

What is the Alternative Hypothesis?

$H_1: \beta_1 \neq 0$

t Test of the Slope

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

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

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

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

$$\beta_1 = \textit{the hypothesized slope}$$

t Test of the Slope – Big Mac - Excel

$t = 5.1437$

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

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

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

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

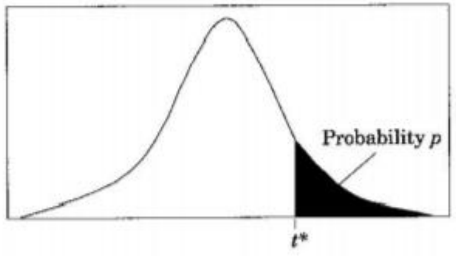


Table B t distribution critical values

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
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.

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

Testing the Overall Model

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

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

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

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

In simple regression, $F = t^2$

$F = t^2$

$F = \frac{\frac{SSR}{df_{reg}}}{\frac{SSE}{df_{err}}} = \frac{MSR}{MSE}$	$t = \frac{\widehat{\beta}}{s_{\widehat{\beta}}} = \frac{\widehat{\beta}}{\frac{SE}{\sqrt{\sum_i (x_i - \bar{x})^2}}}$
$SSR = \sum_i (\hat{y}_i - \bar{y})^2$	$SE = \sqrt{MSE}$
$= \sum_i \left(\hat{\alpha} + \hat{\beta} x_i - \bar{y}\right)^2$	
$= \sum_i \left(\bar{y} - \hat{\beta}\bar{x} + \hat{\beta} x_i - \bar{y}\right)^2$	$t = \frac{\hat{\beta}\sqrt{\sum_i (x_i - \bar{x})^2}}{\sqrt{MSE}}$
$= \hat{\beta}^2 \sum_i (x_i - \bar{x})^2$	
$F = \frac{MSR}{MSE} = \frac{\hat{\beta}^2 \sum_i (x_i - \bar{x})^2}{MSE}$	$t^2 = \frac{\hat{\beta}^2 \sum_i (x_i - \bar{x})^2}{MSE}$

$\boxed{F = t^2}$

Testing the Overall Model – Big Mac - Excel

$F = 26.4581$

Note $t^2 = (5.1437)^2 = 26.4576$

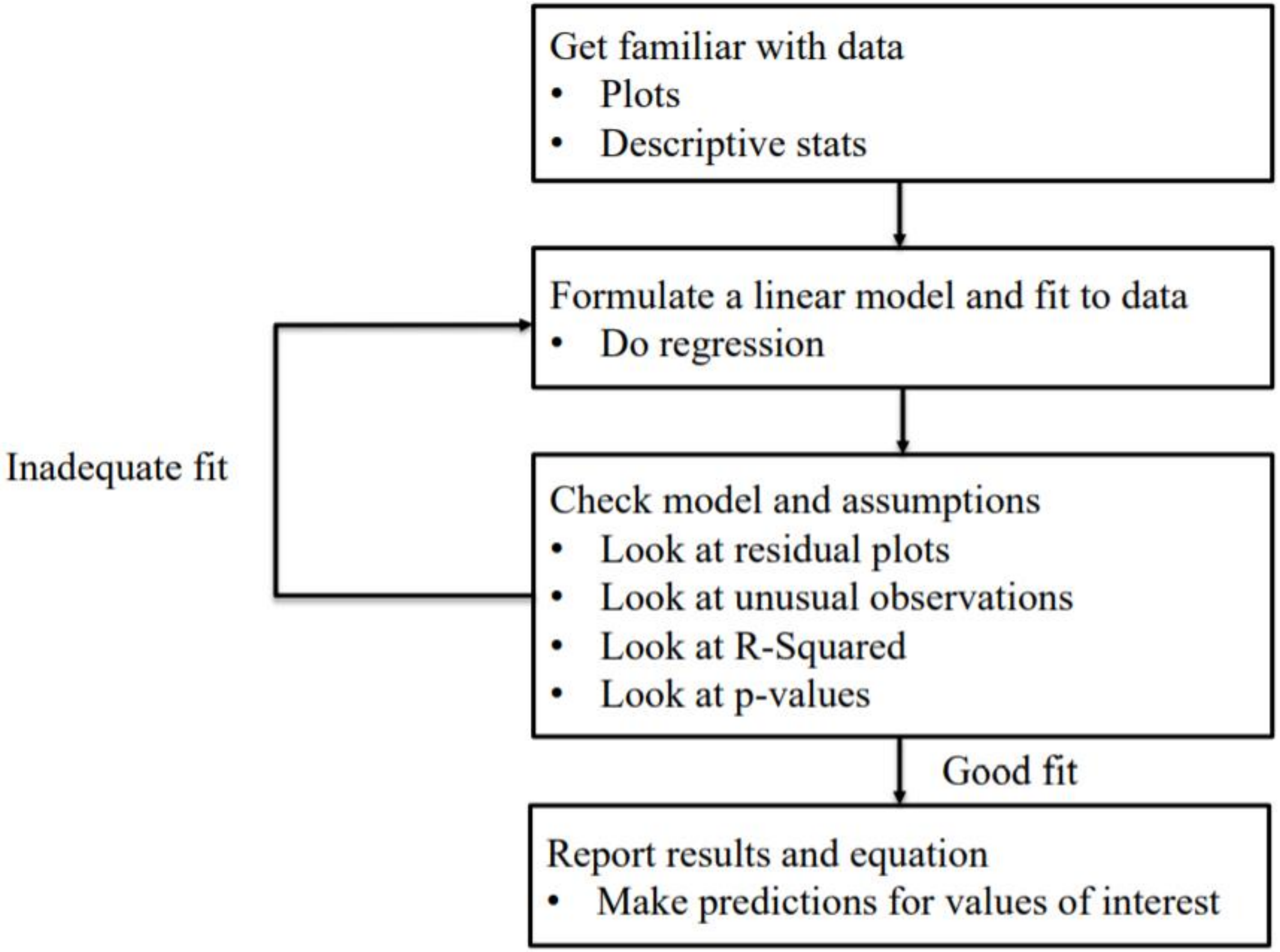
Critical F value, $F_{.05,1,25} = 4.2417$

Reject the null hypothesis for overall significance

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

Denominator Degrees of Freedom df_2 \ Numerator Degrees of Freedom df_1	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

Simple Linear Regression - Steps



Thank You…