Linear Regression and Other Prediction Methods

Origin of the term Regression

- Originally coined by F. Galton to describe the laws of inheritance. He believed that these laws caused population extremes to "regress toward the mean". By this he meant that children of individuals having extreme value of a certain characteristic would tend to have less extreme values of this characteristic than their parents

- The modern day usage of regression is much more general. It means to determine the relationship between a set of variables

- Another usage of regression is in prediction

- Simplest relationship is a linear relationship

$$Y = \beta_0 + \beta_1 x_1 + \cdots \beta_r x_r$$

- Y dependent variable /response variable

  $x_1, \cdots, x_r$ r independent variables /input variables/ predictor variables

  The equation is called a linear regression equation

- If we can learn the regression coefficients $\beta_0, \cdots, \beta_r$, we can exactly predict Y

Linear Regression

- Simple linear regression

$$Y = \alpha + \beta x$$

- Multiple linear regression

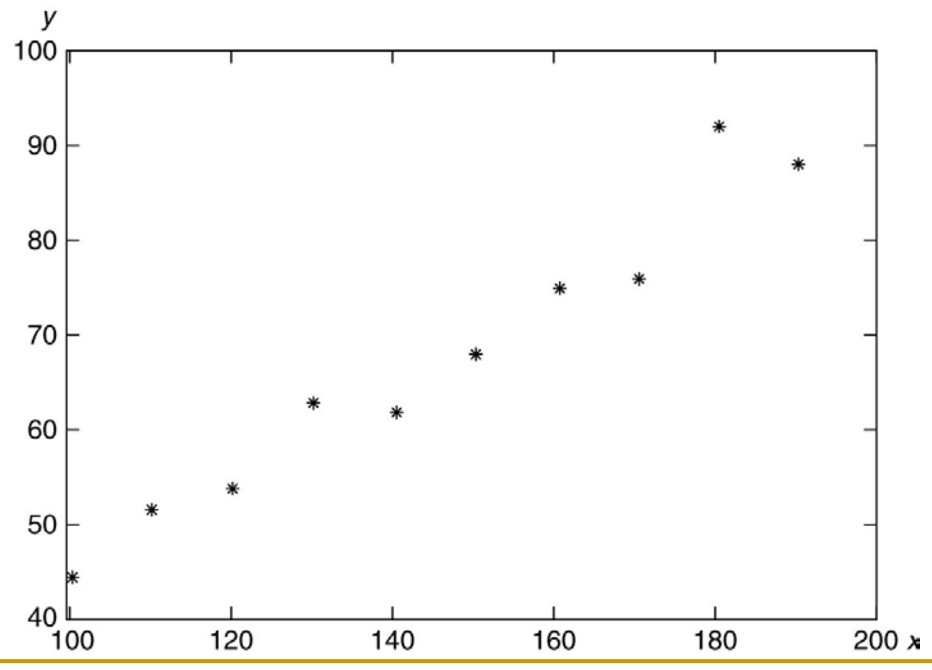
$$Y = \beta_0 + \beta_1 x_1 + \cdots \beta_r x_r$$

We restrict to one variable x, i.e., simple linear regression, in this course

Example: To determine what equation to apply, it is useful to plot a scatter plot of the data first

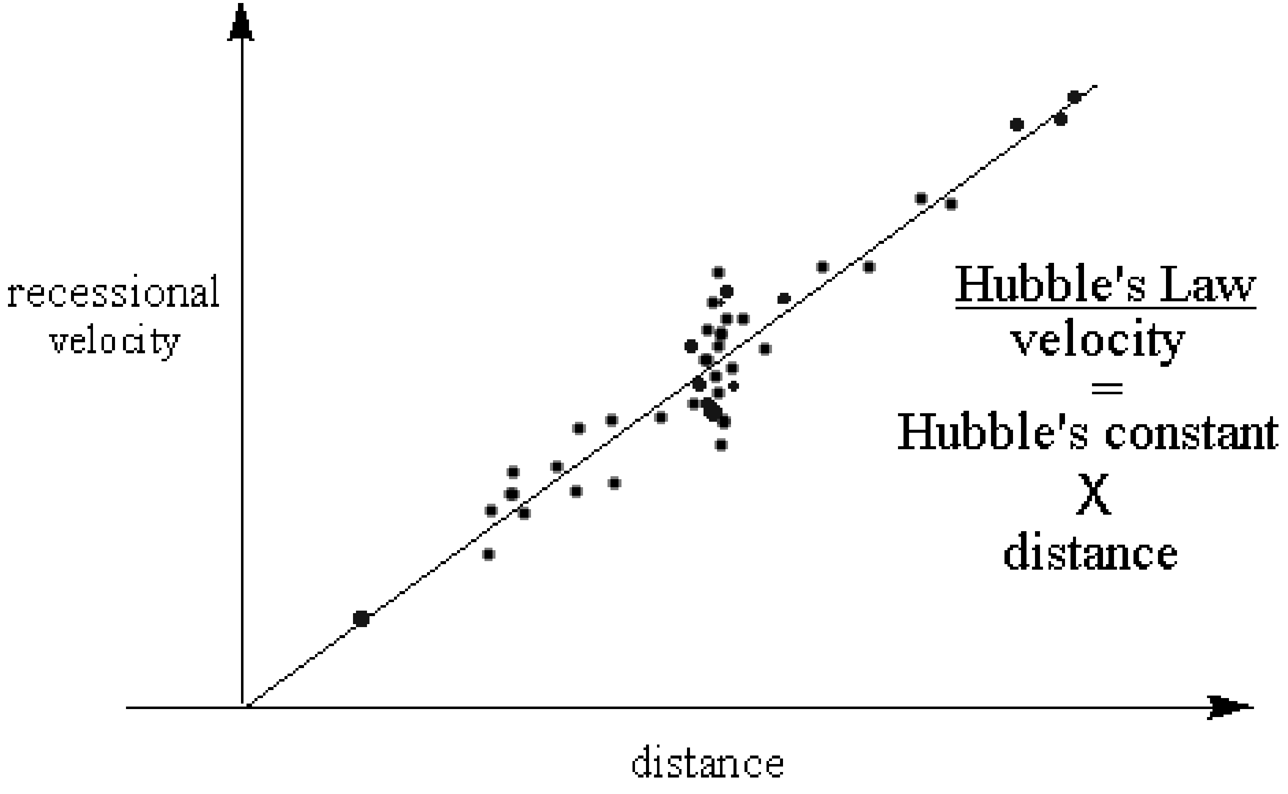
Result of an experiment ->

i	x_i	y_i	i	x_i	y_i
1	100	45	6	150	68
2	110	52	7	160	75
3	120	54	8	170	76
4	130	63	9	180	92
5	140	62	10	190	88



Guess it's a simple linear regression $Y = \alpha + \beta x$

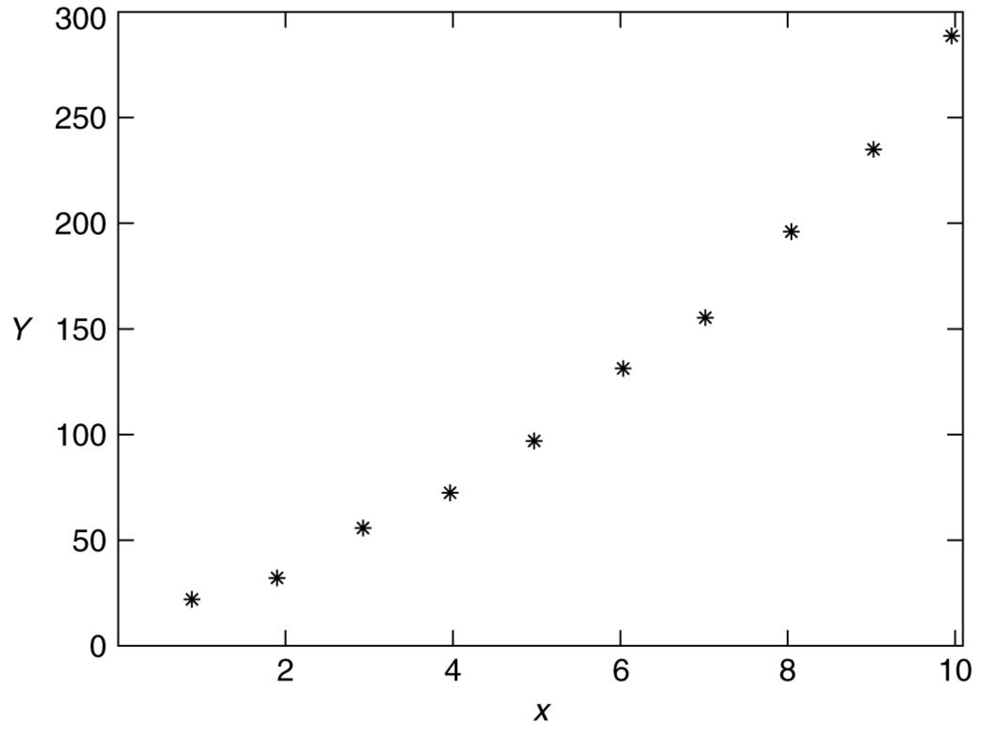
Example: Hubble's law from observations



Guess it's a simple linear regression $Y = \alpha + \beta x$

http://www.astro.cornell.edu/academics/courses/astro201/hubbles_law.htm

Example: Does not seem to be a simple linear regression for the data below. Suggest a suitable formula



A polynomial regression (more precisely quadratic regression)

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2$$

is suitable

Least Squares Estimator

Simple linear regression: $Y = \alpha + \beta x + e$

where e is the error

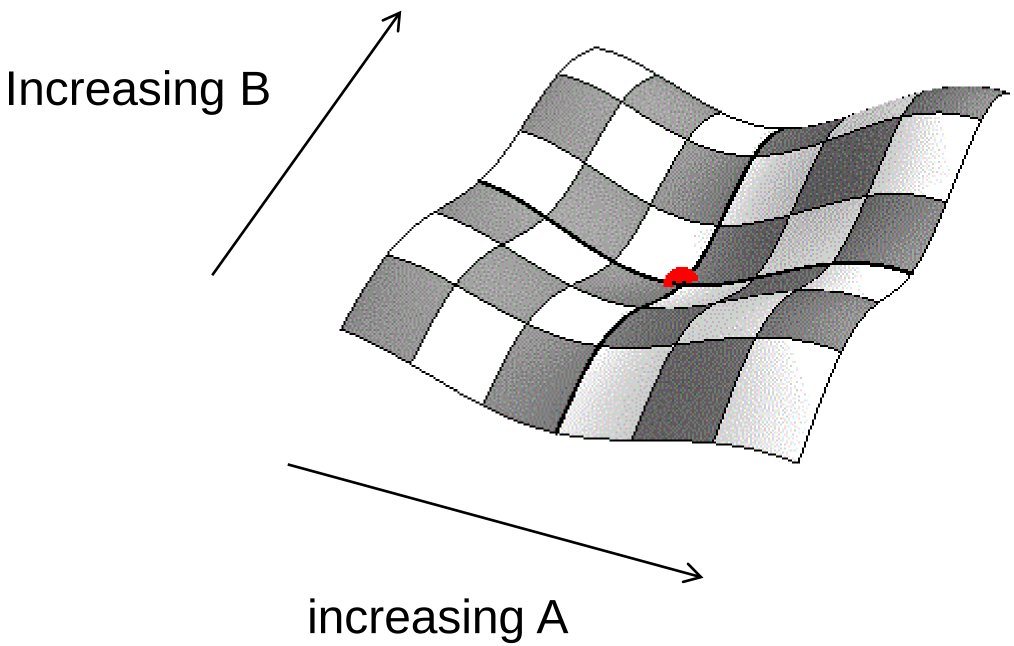
To estimate the regression parameters, one good idea is to find the values of the parameters that minimizes the least sum of squared errors (sometimes simply called least squared errors):

$$SS = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

A and B are estimators of α and β.

For SS to attain the minimum value, a necessary condition is

$$\frac{\partial SS}{\partial A} = 0 \qquad \frac{\partial SS}{\partial B} = 0$$



$$\frac{\partial SS}{\partial A} = -2\sum_{i=1}^{n}(Y_i - A - Bx_i) = 0$$

$$\frac{\partial SS}{\partial B} = -2\sum_{i=1}^{n}x_i(Y_i - A - Bx_i) = 0$$

$\Rightarrow$

$$\sum_{i=1}^{n} Y_i = nA + B\sum_{i=1}^{n} x_i \qquad (1)$$

$$\sum_{i=1}^{n} x_i Y_i = A\sum_{i=1}^{n} x_i + B\sum_{i=1}^{n} {x_i}^2 \qquad (2)$$

Since this is a system of two equations in two unknowns, we can solve for *A* and *B*

Let $\bar{Y} = \sum_{i=1}^{n} Y_i / n$ (mean of Y)

$\bar{x} = \sum_{i=1}^{n} x_i / n$ (mean of X)

(1) becomes

$$\sum_{i=1}^{n} Y_i = nA + B \sum_{i-1}^{n} x_i \;\Rightarrow n\bar{Y} = nA + Bn\bar{x} \;\Rightarrow A = \bar{Y} - B\bar{x}$$

Substitute into (2) gives

$$\sum_{i=1}^{n} x_i Y_i = A \sum_{i=1}^{n} x_i + B \sum_{i=1}^{n} {x_i}^2 = (\bar{Y} - B\bar{x})n\bar{x} + B \sum_{i=1}^{n} {x_i}^2$$

Simplifying,

$$B = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i Y_i - \bar{x}\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2}$$

$$A = \bar{Y} - B\bar{x}$$

A and B are called the least squares estimators

$Y = A + Bx$ is called the estimated regression line

Define

$$S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i{}^2 - n\bar{x}^2$$

$$S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i{}^2 - n\bar{Y}^2$$

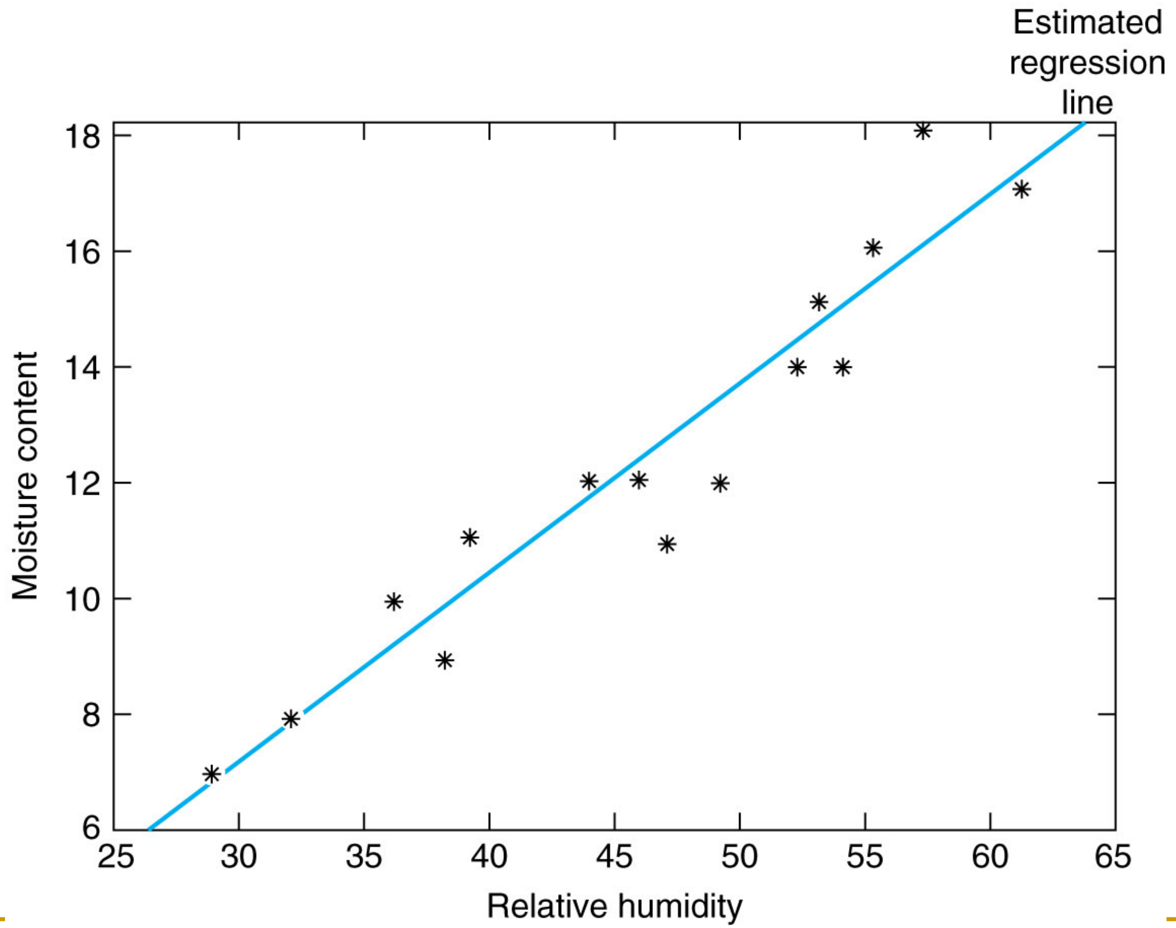
Then the least squares estimators can be expressed as

$$B = \frac{S_{xY}}{S_{xx}}$$

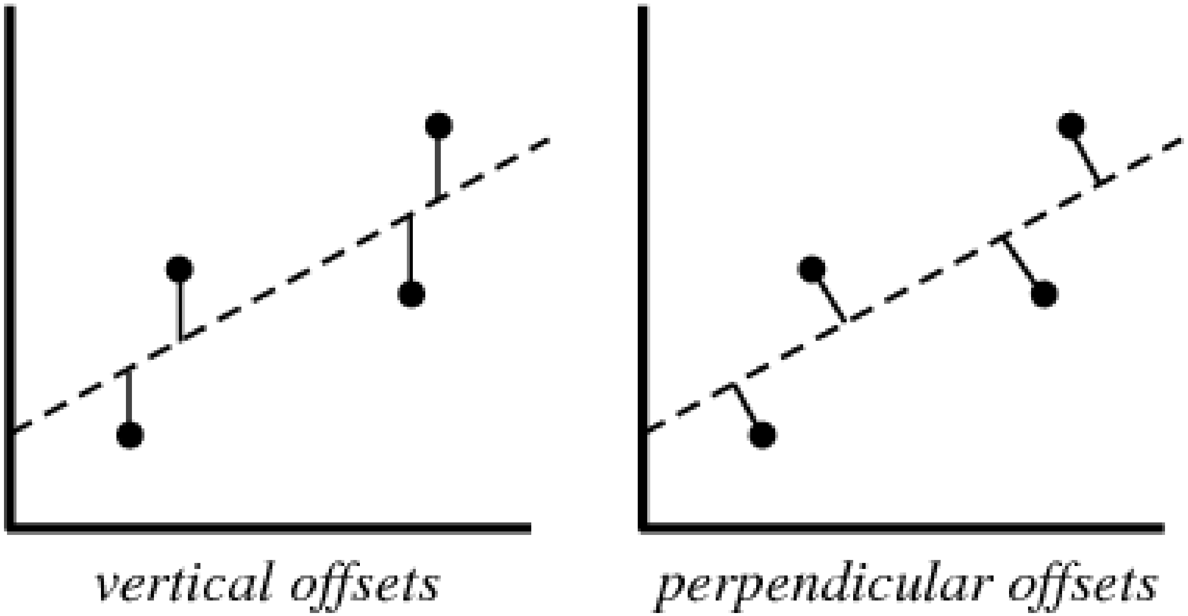
$$A = \bar{Y} - B\bar{x}$$

Example

Relative humidity	46	53	29	61	36	39	47	49	52	38	55	32	57	54	44
Moisture content	12	15	7	17	10	11	11	12	14	9	16	8	18	14	12



Note that the method minimizes the sum of squared errors of the vertical offsets, not the perpendicular offsets.



http://mathworld.wolfram.com/LeastSquaresFitting.html

This is because of the model of the error e. In simple linear regression $Y = \alpha + \beta x + e$, and it is assumed that the input x is exact

The quantities $Y_i - A - Bx_i$ are called the residuals. Define the sum of squares of the residuals

$$SS_R = SS = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

By simple algebraic manipulations,

$$SS_R = \frac{S_{xx} S_{YY} - {S_{xY}}^2}{S_{xx}}$$

Variations of simple linear regression

- Transforming a nonlinear relationship to a linear relationship
- Polynomial regression
- Logistic regression

Transforming a nonlinear relationship to linear

One can transform a nonlinear relationship to linear form and then use linear regression. For example, it is known that

$$W(t) \approx ce^{-dt}$$

We can measure the values of the dependent variables $W(t)$ as a function of the samples at various time t

Transforming by taking log

$$\log\big(W(t)\big) \approx \log(c) - dt \qquad (1)$$

Use the following change of variables

$Y = \log\big(W(t)\big)$ $\alpha = \log(c)$ $\beta = -d$

(1) becomes $Y = \alpha + \beta t$

We can use simple linear regression to estimate α and β

Note that the error model becomes

$W(t) = \mathrm{c}e^{-dt+err}$

err is a random variable with zero mean

Knowledge about the source of error

If the source of the random error is known, the correct model should be used. For example, in the above

$$W(t) = \mathrm{c}e^{-dt+err}$$

If however, it is known that the random error is due to measurement error of $W(t)$, then the correct model should be

$$W(t) = ce^{-dt} + err$$

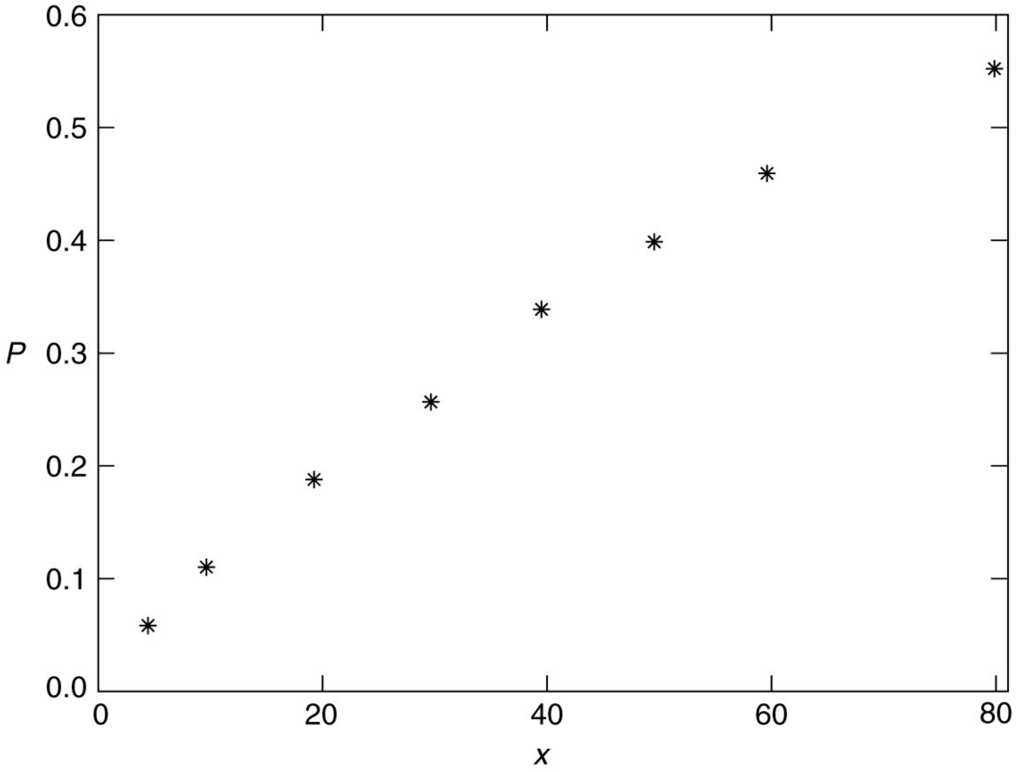
and the transformation to linearity cannot be used. Simple linear regression is used because of ignorance about the error model

Observing the scatter plot of the data, and with no knowledge of the error model, one can hypothesize some nonlinear relationship and then use change of variables to change it to a simple linear regression problem

Example

Result of an experiment ->

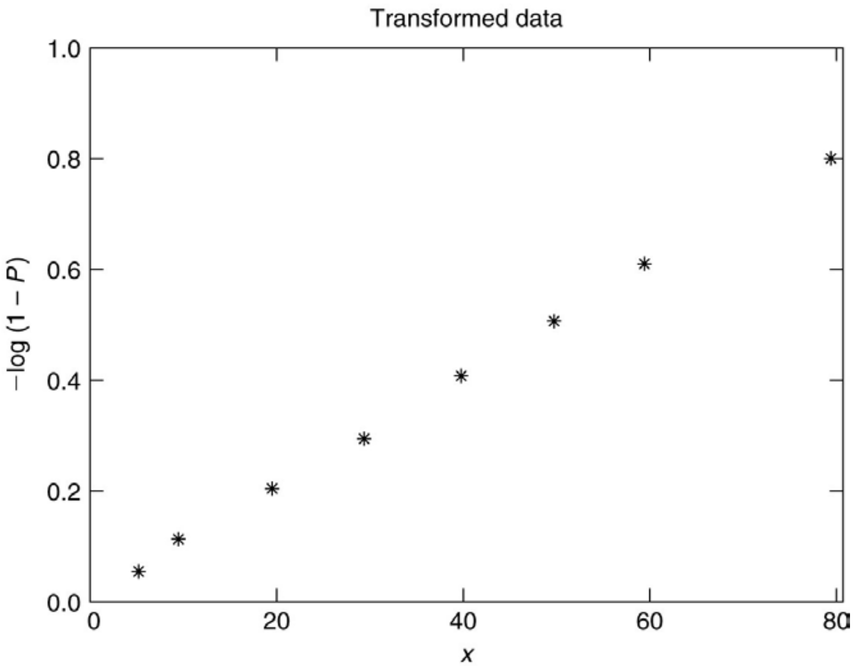
Temperature	Percentage
5°	.061
10°	.113
20°	.192
30°	.259
40°	.339
50°	.401
60°	.461
80°	.551



Hypothesize $1 - P(x) \approx c(1-d)^x$

Take log gives

$$\log\bigl(1 - P(x)\bigr) \approx \log(c) + x log(1-d)$$



Plotting the transformed data gives an approximately linear relationship

Polynomial Regression

In situations where the functional relationship between the response Y and the independent variable x cannot be adequately approximated by a linear relationship, it is sometimes possible to obtain a reasonable fit by considering a polynomial relationship

$$Y = \beta_0 + \beta_1 \mathrm{x} + \beta_2 x^2 + \cdots + \beta_r x^r + e$$

The least square estimator of $\beta_0, \dots, \beta_r$, call them $B_0, \dots, B_r$ is found by minimizing

$$SS = \sum_{i=1}^{n} (Y_i - B_0 - B_1 x_i - B_2 {x_i}^2 \dots - B_r {x_i}^r)^2$$

Setting

$$\frac{\partial SS}{\partial B_i} = 0 \qquad\qquad i = 0, \ldots r$$

One obtains $r + 1$ linear equations in $r + 1$ unknowns. Solving it finds the least square estimate

Caution:

It is always possible to fit a polynomial of degree n that passes through all the n pairs of data points, i.e., $R^2 = 1$. However, this would result in over-fitting

One should observe the scatterplot and uses the lowest possible degree of the polynomial

Statistical inferences about the regression parameter β

$$Y = \alpha + \beta x + e$$

Assume for each data point i

$$Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$$

It assumes that the random error e are independent normal random variables having mean 0 and variance σ^2

Note that it supposes σ^2 does not depend on the input value but rather is a constant

As the Y_i are independent normal variables, $(Y_i - E[Y_i])/\sqrt{Var(Yi)}, i = 1, \dots, n$ are independent standard normal variables, as the sum of squares of independent standard normal variables form a Chi-square distribution

$$\sum_{i=1}^{n} \frac{(Y_i - E[Y_i])^2}{Var(Y_i)} = \sum_{i=1}^{n} \frac{(Y_i - \alpha - \beta x_i)^2}{\sigma^2} \sim {\chi_n}^2$$

with n degrees of freedom

$$\frac{SS_R}{\sigma^2} \sim {\chi^2}_{n-2}$$

(as A and B removes 2 d.f. from the n random variables)

It can be shown that $B \sim \mathcal{N}\ (\beta, \sigma^2/S_{xx})$. Hence

$$\frac{B-\beta}{\sqrt{\sigma^2/S_{xx}}} \sim \mathcal{N}(0,1) = Z$$

and it is independent of

$$\frac{SS_R}{\sigma^2} \sim {\chi^2}_{n-2}$$

Hence by definition of a t-random variable with $n-2\ d.f.$

[A more thorough mathematical derivation may be found in text pg. 362-365]

$$\frac{Z}{\sqrt{\chi^2{}_{n-2}/(n-2)}} = \frac{\sqrt{S_{xx}}(B-\beta)/\sigma}{\sqrt{\frac{SS_R}{\sigma^2(n-2)}}}$$

$$= \sqrt{\frac{(n-2)S_{xx}}{SS_R}}(B-\beta) \sim t_{n-2}$$

Recall

$$Y = \alpha + \beta x + e$$

We wish to test the hypothesis

$$\begin{array}{ll} null\ hypothesis & H_0\colon \beta = 0 \\ alternative\ hypothesis & H_1\colon \beta \neq 0 \end{array}$$

If the null hypothesis is true,

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}\,B \sim t_{n-2} \tag{1}$$

$$B = \frac{S_{xY}}{S_{xx}} \tag{2}$$

The Pearson coefficient (sample correlation coefficient) r is

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{S_{xY}}{\sqrt{S_{xx}S_{YY}}} \tag{3}$$

Put (2) and (3) into (1) gives

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

Thus

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

is a t statistics with $n-2$ d.f.

Example

An individual claims that the fuel consumption of his automobile does not depend on how fast the car is driven. To test the plausibility of this hypothesis, the car was tested at various speeds:

Speed	Miles per Gallon
45	24.2
50	25.0
55	23.3
60	22.0
65	21.5
70	20.6
75	19.8

Do these data refute the claim that the mileage per gallon of gas is unaffected by the speed at which the car is being driven?

Suppose that a simple linear regression model

$$Y = \alpha + \beta x + e$$

relates Y, the miles per gallon of the car to x, the speed at which it is being driven. The claim is $\beta = 0$

Set up the hypothesis

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

$S_{xx} = 700 \quad S_{YY} = 21.757 \quad S_{xY} = -119 \quad n = 7$

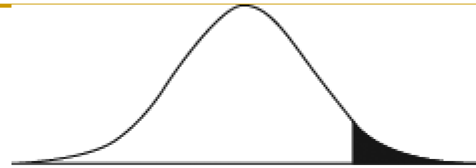
$$r = \frac{S_{xY}}{\sqrt{S_{xx} S_{YY}}} = -0.964269511$$

There is a strong negative correlation

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = -8.18857138$$

d.f. = $n - 2 = 5$

$t_{0.005,5} = 4.032$ The hypothesis that $\beta = 0$ is rejected at 1% level of significance. Thus, the claim that the mileage does not depend on the speed at which the car is driven is rejected; there is strong evidence that increased speeds lead to decreased mileages



t- distribution table

v is the degree of freedom

d.f. = 5

	Tail probability									
ν	**0.4**	**0.25**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0025**	**0.001**	**0.0005**
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	$\boxed{4.032}$	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
70	0.254	0.678	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
130	0.254	0.676	1.288	1.657	1.978	2.355	2.614	2.856	3.154	3.367
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Hypothesis Testing of Pearson Coefficient r

In simple linear regression, we assume

$$Y = \alpha + \beta x + e$$

Y is a random variable and the input variable x is a precise deterministic value, while e is the random error, e.g. independent Gaussian noise of the form

$$Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$$

- Then we can use simple linear regression to estimate the parameters α and β. This is useful in engineering

- If both X and Y are random variables (e.g. the prices of two stocks), then we can compute Pearson coefficient r

- Now we wish to test the hypothesis

Null hypothesis	H_0:	X and Y are not correlated
Alternative hypothesis	H_1:	X and Y are correlated

- We make two assumptions

Assumption 1: The joint distribution (X, Y) is a bivariate normal distribution

Assumption 2: X and Y have the following statistically linear relationship:

$$E(Y|X = x) = \alpha + \beta \mathrm{x}$$

Then it can be shown that

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

follows a t-distribution with $n-2$ d.f. One can use the following hypothesis test (paired t test)

Null hypothesis H_0: $\rho = 0$

Alternative hypothesis H_1: $\rho \neq 0$

(ρ is the random variable corresponding to r)

to test whether the correlation is significant

Note: Bivariate normal distribution

$$f(x,y) = \frac{e^{-Q/2}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$$

$$Q = \frac{1}{1-\rho^2}\left[\frac{(x-\mu_x)^2}{{\sigma_x}^2} + \frac{(y-\mu_y)^2}{{\sigma_y}^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]$$

ρ is the random variable corresponding to the Pearson's coefficient and $Cov(X,Y) = \rho\sigma_x\sigma_y$. When $\rho = 0$,

$$f(x,y) = N(\mu_x, {\sigma_x}^2)N(\mu_y, {\sigma_y}^2)$$

X and Y are independent normal random variables

Example: Blood haemoglobin (Hb) levels vs Packed cell volume (PCV) of 14 female blood donors

Hb	PCV
15.5	0.450
13.6	0.420
13.5	0.440
13.0	0.395
13.3	0.395
12.4	0.370
11.1	0.390
13.1	0.400
16.1	0.445
16.4	0.470
13.4	0.390
13.2	0.400
14.3	0.420
16.1	0.450

Random variable X

Random variable Y

In this example, $n = 14$, it is found using Excel that $r = 0.877013$

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

follows a t-distribution with $n - 2 = 12$ d.f.

$$T = \frac{(0.877013)\sqrt{12}}{\sqrt{1-(0.877013)^2}} = 6.323148913$$

p-value $= P\{|T| > 6.323148913\}$

$$< 2\,P\{T > 4.318\}$$

$$= 2(0.0005) = 0.001$$

Terms for describing the strength of r

One can verbally describe the strength of the correlation using Evans 1996 guide for the absolute value of r

0.00 – 0.19	“very weak”
0.20 – 0.39	“weak”
0.40 – 0.59	“moderate”
0.60 – 0.79	“strong”
0.80 – 1.00	“very strong”

e.g. $r = 0.42$ would be “moderate positive correlation”

The p value is found to be smaller than 0.001. Hence there appears to be a very strong, positive correlation between Hb and PCV.

It can be reported as

"A Pearson's correlation was run to determine the relationship between 14 females' Hb and PCV values. There was a very strong, positive correlation between Hb and PCV ($r = 0.88$, $N = 14$, $p < 0.001$)."

Coefficient of Determination R^2

A standard measure in statistics of the amount of variation in $Y_1, \ldots, Y_n$ is

$$S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

For example, if all Y_i are equal, then $S_{YY} = 0$

Variation of Y due to two factors

(1) Variation due to different x_i (2) Variation due to e

$$SS_R = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

measures the variation due to (1)

$$S_{YY} - SS_R$$

represents the amount of variation explained by the different input values

The coefficient of determination R^2 represents the proportion of the variation in the response variable explained by the different input values

$$R^2 = \frac{S_{YY} - SS_R}{S_{YY}}$$

$0 \leq R^2 \leq 1$

A value of R^2 near 1 indicates that most of the variation of the response data is explained by the different input values, whereas a value of near 0 indicates that little of the variation is explained by the different input values

The value of R^2 is often used as an indicator of how well the regression model fits the data, with a value near 1 indicating a good fit, and one near 0 indicating a poor fit. In other words, if the regression model is able to explain most of the variation in the response data, then it is considered to fit the data well.

Relationship with Pearson Coefficient r

It can be shown by algebraic manipulations that

$$r^2 = R^2$$

or $|r| = \sqrt{R^2}$

If a data set has $r = 0.9$, it implies that a simple linear regression model for these data explains 81% of the variation in the response values. That is, 81% of the variation in the response values is explained by the different input values.

Thus the larger the r, the more likely the data is explainable by a linear regression model

Spearman rank correlation coefficient r_s

Consider the data pairs (x_i, y_i), $i = 1, \dots, n$

It ranks the x_i from low to high, i.e., replace x_i by the rank of $R(x_i)$

It ranks the y_i from low to high, i.e., replace y_i by the rank of $R(y_i)$

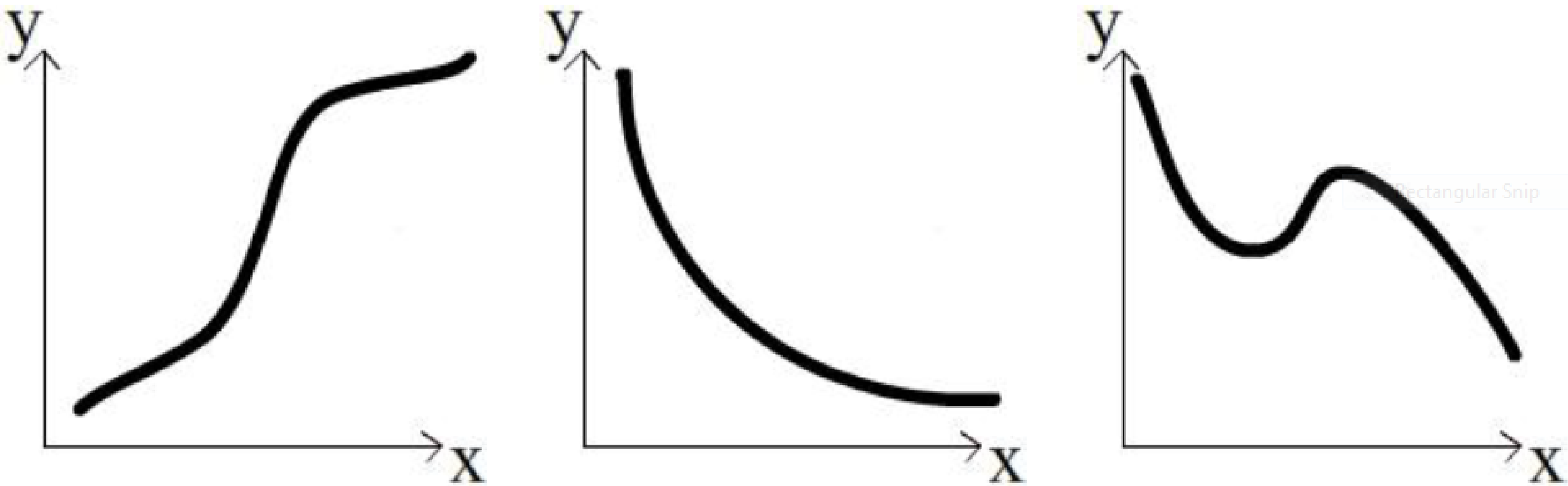
Then compute the Pearson's coefficient r for the data pairs $(R(x_i), R(y_i))$, $i = 1, \dots, n$. This r is called Spearman rank correlation coefficient r_s. We can perform the above hypothesis testing on r_s to check if there is significant correlation

Usage

- A non-parametric statistics
- Does not require the joint distribution (X, Y) is a bivariate normal distribution
- Does not require X and Y to be statistically linearly related

- It tests on whether X and Y are monotonically related

Monotonic functions

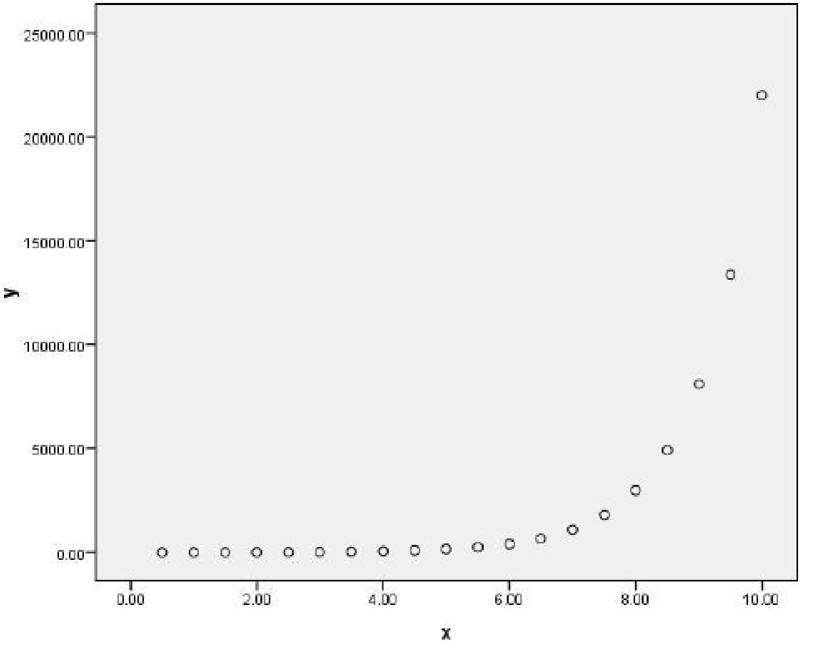


Monotonically increasing Monotonically decreasing Not monotonic

Example

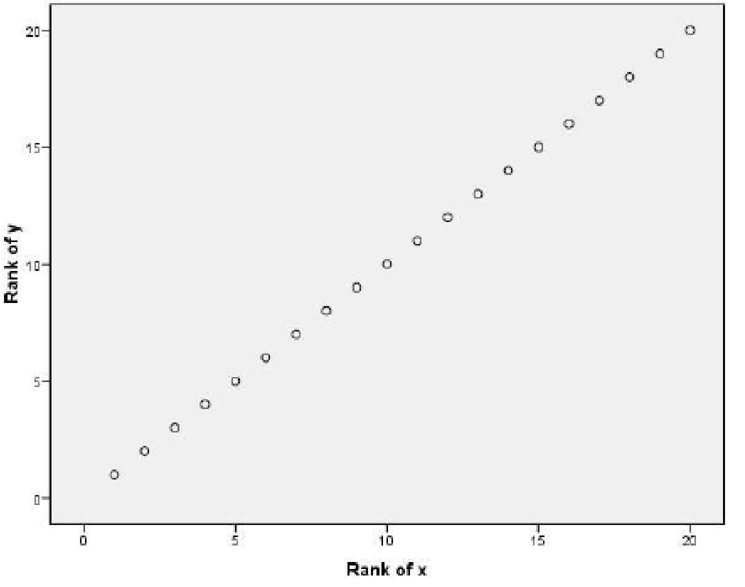
$Y = e^X$ (not linearly related)

	x	y
1	.5	1.6
2	1.0	2.7
3	1.5	4.5
4	2.0	7.4
5	2.5	12.2
6	3.0	20.1
7	3.5	33.1
8	4.0	54.6
9	4.5	90.0
10	5.0	148.4
11	5.5	244.7
12	6.0	403.4
13	6.5	665.1
14	7.0	1096.6
15	7.5	1808.0
16	8.0	2981.0
17	8.5	4914.8
18	9.0	8103.1
19	9.5	13359.7
20	10.0	22026.5



Rank the data to get data pairs $(R(x_i), R(y_i))$, $i = 1, \ldots, n$. Then compute Pearson's coefficient r on the ranked data pairs. This gives $r_s = r = 1$.

	x	Rank of x	y	Rank of y
1	.5	1	1.6	1
2	1.0	2	2.7	2
3	1.5	3	4.5	3
4	2.0	4	7.4	4
5	2.5	5	12.2	5
6	3.0	6	20.1	6
7	3.5	7	33.1	7
8	4.0	8	54.6	8
9	4.5	9	90.0	9
10	5.0	10	148.4	10
11	5.5	11	244.7	11
12	6.0	12	403.4	12
13	6.5	13	665.1	13
14	7.0	14	1096.6	14
15	7.5	15	1808.0	15
16	8.0	16	2981.0	16
17	8.5	17	4914.8	17
18	9.0	18	8103.1	18
19	9.5	19	13359.7	19
20	10.0	20	22026.5	20



It shows that the data have a perfect monotonically increasing relationship

Similarly, the t statistic may be used to test whether the data has any significant monotonic relationship

$$T = \frac{r_s\sqrt{n-2}}{\sqrt{1-{r_s}^2}}$$

follows a t-distribution with $n-2$ d.f. One can use the following hypothesis test (paired t test)

H_0: $\rho = 0$ (no monotonic relationship)

H_1: $\rho \neq 0$ (significant monotonic relationship)

Other Prediction Methods

Instead of hypothesizing an equation to explain the data and use it for prediction, in short term forecasting, one way is to use the history of the data to predict future data. We shall introduce two such methods:

1. Moving Average Model
2. Exponential Smoothing Model

Moving Average Model

Builds a forecast by averaging the observations in the most recent n periods:

$$A_t = \frac{(x_t + x_{t-1} + x_{t-n+1})}{n}$$

$$F_{t+1} = A_t$$

F_{t+1} is the forecast at time $t + 1$

Exponential Smoothing Model

All time series forecasts involve weighted averages of historical observations. In the case of a 4-period moving average, the weights are 0.25 on each of the last four observations and zero on all of the previous observations. If the philosophy is to weight recent observations more than older ones, then why not allow the weights to decline gradually as we go back in time. This is the approach used in exponential smoothly

$$S_t = \alpha x_t + (1 - \alpha) S_{t-1} \qquad (1)$$
$$F_{t+1} = S_t$$

$0 \leq \alpha \leq 1$. α is called the smoothing constant.

One period previously, we would have made the calculation

$$S_{t-1} = \alpha x_{t-1} + (1-\alpha)S_{t-2} \qquad (2)$$

Substituting (2) into (1),

$$S_t = \alpha x_t + \alpha(1-\alpha)x_{t-1} + (1-\alpha)^2 S_{t-2}$$

Continuing,

$$S_t = \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \alpha(1-\alpha)^3 x_{t-3} + \ldots$$

Because $\alpha < 1$, the term $\alpha(1-\alpha)^t$ declines as t increases is an exponential decrease manner. Contrast with the constant and then sharp cutoff in moving averaging

Multiple Linear Regression

The dependent variable Y is related linearly to k inputs

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

The procedure for multiple linear regression generalizes the procedure for simple linear regression (as a special case when k = 1)

Let $B_0, B_1, \ldots, B_k$ denote estimators of $\beta_0, \beta_1, \ldots, \beta_k$, the sum of the squared difference is

$$F = \sum_{i=1}^{n} (Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik})^2$$

where n is the number of data points

Writing out

$$\frac{\partial F}{\partial B_i} = 0 \qquad\qquad i = 0, \ldots, k$$

$$\frac{\partial F}{\partial B_0} = \sum_{i=1}^{n} (Y_i - B_0 - B_1 x_{i1} - \cdots - B_k x_{ik}) = 0$$

$$\frac{\partial F}{\partial B_1} = \sum_{i=1}^{n} x_{i1}(Y_i - B_0 - B_1 x_{i1} - \cdots - B_k x_{ik}) = 0$$

$$\vdots$$

$$\frac{\partial F}{\partial B_k} = \sum_{i=1}^{n} x_{ik}(Y_i - B_0 - B_1 x_{i1} - \cdots - B_k x_{ik}) = 0$$

which is a set of $(k + 1)$ equations

Collecting the constant terms to the right hand side,

$$nB_0 + B_1 \sum_{i=1}^{n} x_{i1} + \cdots + B_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} Y_i$$

$$B_0 \sum_{i=1}^{n} x_{i1} + B_1 \sum_{i=1}^{n} {x_{i1}}^2 + \cdots + B_k \sum_{i=1}^{n} x_{i1} x_{ik} = \sum_{i=1}^{n} x_{i1} Y_i$$

$$\vdots$$

$$B_0 \sum_{i=1}^{n} x_{ik} + B_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \cdots + B_k \sum_{i=1}^{n} {x_{ik}}^2 = \sum_{i=1}^{n} x_{ik} Y_i$$

which can be written in the matrix form

$$XB = Y$$

where Y is $n \times 1$, X is $(k+1) \times 1$,

$$B = \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_k \end{bmatrix}$$

To solve

$$XB = Y$$

Multiply both sides by the transpose of X,

$$(X^T X)B = X^T\ Y$$

As $(X^T X)$ is a square matrix, the inverse exists if the matrix does not de-generate (i.e., when the rank of $(X^T X)$ is $k + 1$)

Multiply both sides by the inverse of $(X^T X)$ gives the least square solution B,

$$B = (X^T X)^{-1} X^T Y$$

Choosing Input Variables in Multiple Linear Regression

1. When selecting an independent variable to predict an outcome, select a predictor variable (X) that is related to the predicted variable (Y). That way, the two share something in common (remember, they should be correlated).
2. When selecting more than one independent variable (such as X_1 and X_2, try to select variables that are independent or uncorrelated with one another but are both related to the outcome or predicted (Y) variable. One can use the p-value in the correlation to determine whether two variables are independent. A rule of thumb is that a p-value above 0.05 to indicate that the variable should be eliminated from the model

3. Try to use fewer variables if possible as it incurs a cost to collect the data
4. One can also systematically search over different subset of variables from all of those in the dataset and attempt to determine the best collection. Two common approaches are

   forward selection: variables are added into the regression model one at a time, starting with the one that most improves the fit of the model to the data

   backward selection: a model is first run with all the variables, and then they are eliminated one by one, starting with the one that makes the smallest contribution to the fit

Cross Validation

One use of regression is to determine the relationship between the variables. Another use is in prediction

Cross validation is an idea that enables one to access the quality of the prediction with the existing data. Two common methods are

Leave-one-out cross validation: Out of n samples, randomly choose 1 sample as validation data and the remaining n-1 samples as training data. Repeat for all the possible 1 sample. Then compute the average prediction error k-fold cross validation: the n data is randomly partitioned into k sets, each of size n/k. Choose one of the k sets as validation data, and the other k-1 sets as training data. Repeat k times (k-fold), with each of the k possible sets used as validation data. Then compute the average error

Choose the regression model with the minimum average error during validation

Logistic Regression

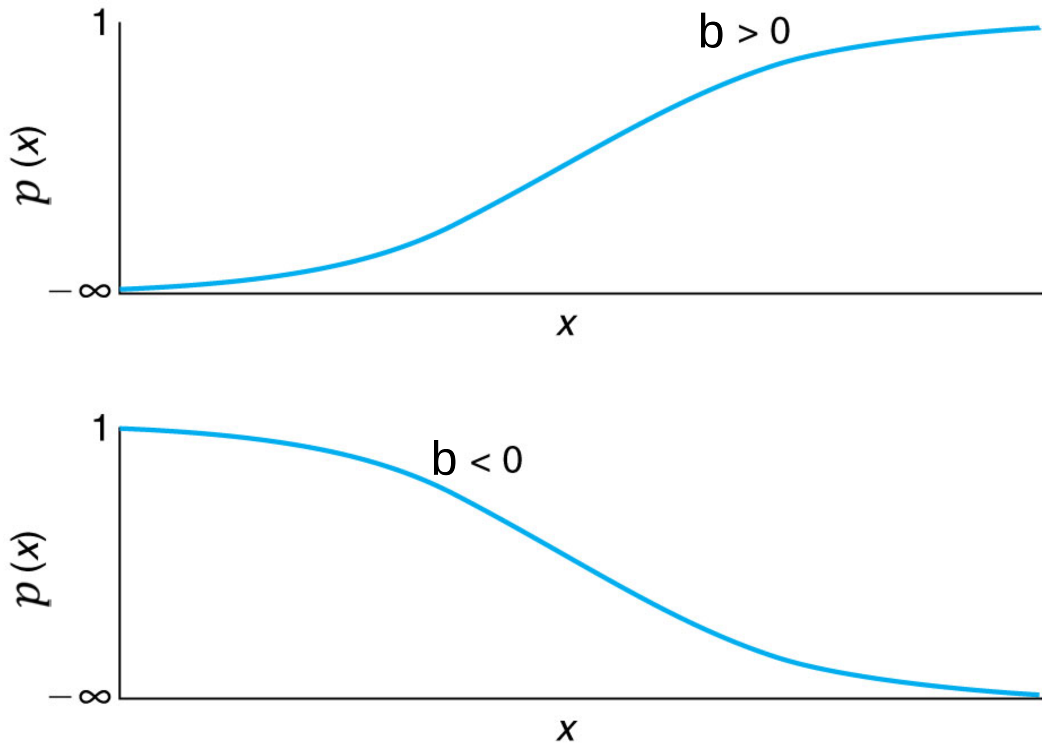
- Experiment outcome is binary, i.e., either success ('1') or failure ('0')
- Probability of success of the following form:

$$p(x) = \frac{e^{a+bx}}{1+e^{a+bx}}$$

- Odds for success $o(x) = \frac{p(x)}{1-p(x)} = e^{a+bx}$
- Log odds, called the logit, is a linear function

$$\log[o(x)] = a + bx$$

Shape of the logistic function



- Note that the form of the probability function is rather specific, when $b > 0$, it assumes that the probability function is a monotonically increasing function of x; when $b < 0$, it assume that the probability function is a monotonically decreasing function of x; when $b = 0$, the function is a constant function
- The detailed shape of the function is adjusted by a

- Given k pairs of data (x_i, y_i), logistic regression tries to find the maximum likely (a, b) that fits the data to the equation $p(x) = \frac{e^{a+bx}}{1+e^{a+bx}}$

Maximum Likelihood estimation

$P\{Y_i = 1\} = p(x_i)$
$P\{Y_i = 0\} = 1 - p(x_i)$

This is rewritten as

$$P\{Y_i = y_i\} = [p(x_i)]^{y_i}[1 - p(x_i)]^{1-y_i} \qquad y_i = 0, 1$$

$$= \left(\frac{e^{a+bx_i}}{1+e^{a+bx_i}}\right)^{y_i}\left(\frac{1}{1+e^{a+bx_i}}\right)^{1-y_i}$$

$$P\{Y_i = y_i, i = 1, \dots, k\} = \prod_i \left(\frac{e^{a+bx_i}}{1 + e^{a+bx_i}}\right)^{y_i}\left(\frac{1}{1 + e^{a+bx_i}}\right)^{1-y_i}$$

$$= \prod_i \frac{\left(e^{a+bx_i}\right)^{y_i}}{1+e^{a+bx_i}}$$

Taking log,

$$\log(P\{Y_i = y_i, i = 1, \dots, k\}) = \sum_{i=1}^{k} y_i(a + bx_i) - \sum_{i=1}^{k} \log(1 + e^{a+bx_i})$$

As the expression is nonlinear, use a numerical method to find (a, b) that maximizes the expression

References

1. Regression: Ch. 9 of Text

2. Hypothesis testing of *r*

   a. Ch. 11.8, D. Wackerly, W. Mendenhall, and R.L. Scheaffer, *Mathematical statistics with applications*, 7th Ed., 2008.

   b. http://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf

   (http://www.statstutor.ac.hk contains open resources for learning statistics under Creative Commons License)

2. Spearman rank correlation coefficient:

   http://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf

   (http://www.statstutor.ac.hk contains open resources for learning statistics under Creative Commons License)

3. Prediction: Ch. 7 S.G. Powell, K.R. Baker, Management Science, The Art of Modeling with Spreadsheets, 4th Ed. Wiley, 2014 (e-book is available in Library)