ORIGIN ≡ 1

Simple Linear Regression

Linear Regression and the so-called "General Linear Model" represent a class of methods that seek to relate values of an observed "dependent" random variable (Y) that is Normally distributed to one or more "independent" (or predictor) variables (X) using a linear function analogous to a linear transformation - i.e., using only translation and change of scale. We typically employ "linear coefficients" α & β (not to be confused with the probability of types I & II errors in statistical tests) to describe translation (α) and change of scale (β). Thus a function such as Y = 5 + 23X qualifies as a linear function (α=5, β=23). Note, however, that with the use of an appropriate non-linear transformations of the data, many non-linear functions can be treated by general linear methods also. For instance, transforming variable X to X^2 allows one to model Y = X^2 (α=0, β=1) and taking logs allows one to model the famous allometric equation: Y = aX^b as ln(Y) = ln(a) + b(ln(X)) (α=ln(a), β=b).

ZAR := READPRN("c:/DATA/Biostatistics/ZarEX17.1R.txt")

$X := ZAR^{\langle 2 \rangle}$ **^Zar Example 17.1**

$Y := ZAR^{\langle 3 \rangle}$

$X_{bar} := \text{mean}(X)$ $X_{bar} = 10$

$Y_{bar} := \text{mean}(Y)$ $Y_{bar} = 3.4154$

$s := \sqrt{\text{Var}(Y)}$ $s = 1.2799$ $s^2 = 1.6381$

$n := \text{length}(Y)$ $n = 13$

$i := 1 .. n$

X = (3, 4, 5, 6, 8, 9, 10, 11, 12, 14, 15, 16, 17)ᵀ

Y = (1.4, 1.5, 2.2, 2.4, 3.1, 3.2, 3.2, 3.9, 4.1, 4.7, 4.5, 5.2, 5)ᵀ

ZAR =

	1	2	3
1	1	3	1.4
2	2	4	1.5
3	3	5	2.2
4	4	6	2.4
5	5	8	3.1
6	6	9	3.2
7	7	10	3.2
8	8	11	3.9
9	9	12	4.1
10	10	14	4.7
11	11	15	4.5
12	12	16	5.2
13	13	17	5

Assumptions:

- **Standard Linear Regression depends on specifying in advance which variable is to be considered 'dependent' and which 'independent'. This decision matters as changing roles for Y & X usually produces a different result.**
- **$Y_1, Y_2, Y_3, \ldots, Y_n$ (dependent variable) is a random sample ~ N(μ,σ^2).**
- **$X_1, X_2, X_3, \ldots, X_n$ (independent fixed values) with each value of X_i matched to Y_i**

Model:

$$Y = \alpha + \beta X + \varepsilon$$

where: α is the y intercept of the regression line (translation)
β is the slope of the regression line (scaling coefficient)
ε is the error factor in prediction of Y given that it is a random variable distributed as N(0,σ^2).

Least Squares Estimation of the Regression Line:

Sums of Squares and Cross Products corrected for mean location:

$L_{xx} := \sum_i (X_i - X_{bar})^2$ $L_{xx} = 262$ **< corrected Sum of squares of X**

$L_{yy} := \sum_i (Y_i - Y_{bar})^2$ $L_{yy} = 19.6569$ **< corrected Sum of squares of Y**

$L_{xy} := \sum_i (X_i - X_{bar}) \cdot (Y_i - Y_{bar})$ $L_{xy} = 70.8$ **< corrected Sum of cross products**

Estimated Regression Coefficients for $Y = \alpha + \beta X$:

$b := \frac{L_{xy}}{L_{xx}}$ $b = 0.2702$ **< b is the sample estimate of β**

$a := Y_{bar} - b \cdot X_{bar}$ $a = 0.7131$ **< a is the sample estimate of α**

Estimated values of Y (Y_{hat}):

$Y_{hat_i} := a + b \cdot X_i$ **< using estimated coefficients and each value of the independent variable to estimate dependent value points on the Regression line.**

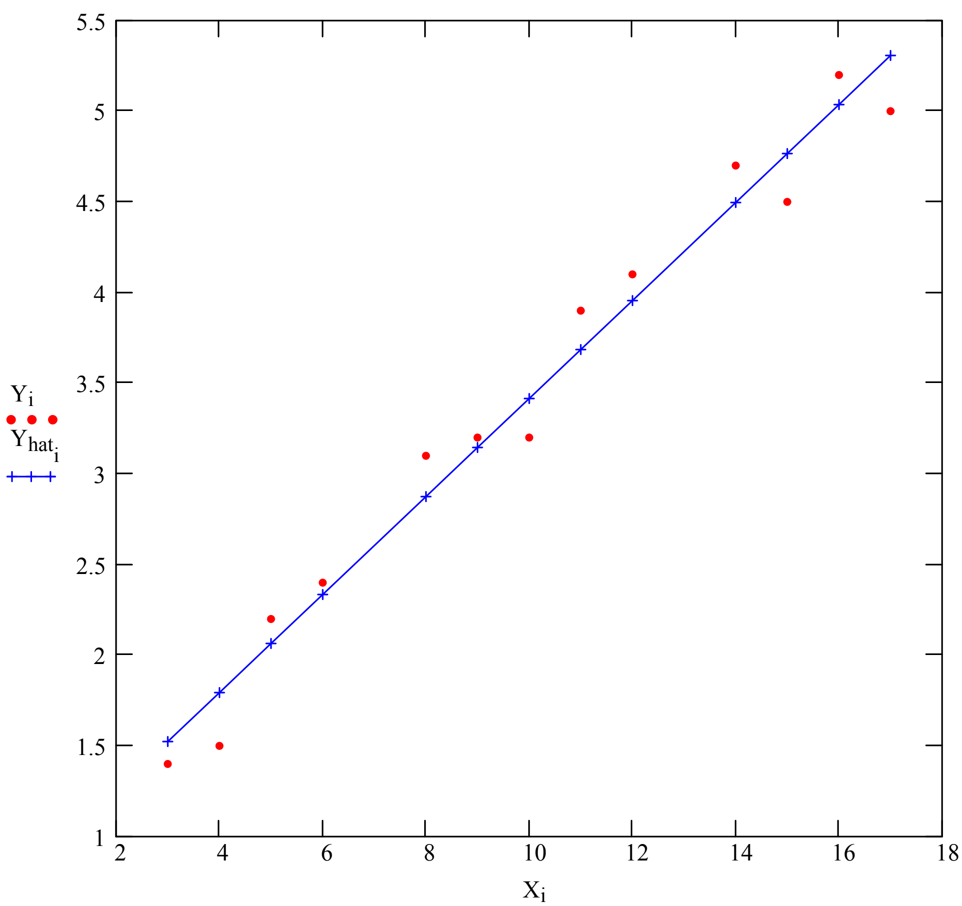
$$Y_{hat} = \begin{pmatrix} 1.5238 \\ 1.794 \\ 2.0642 \\ 2.3345 \\ 2.8749 \\ 3.1452 \\ 3.4154 \\ 3.6856 \\ 3.9558 \\ 4.4963 \\ 4.7665 \\ 5.0368 \\ 5.307 \end{pmatrix}$$

Residuals:

$e_i := Y_{hat_i} - Y_i$ **< deviation of each value Y_i from Regression line = $Yhat_i$**

$$e = \begin{pmatrix} 0.1238 \\ 0.294 \\ -0.1358 \\ -0.0655 \\ -0.2251 \\ -0.0548 \\ 0.2154 \\ -0.2144 \\ -0.1442 \\ -0.2037 \\ 0.2665 \\ -0.1632 \\ 0.307 \end{pmatrix}$$

Plot of Values:



Prototype in R:

```
#SIMPLE LINEAR REGRESSION
#ZAR EXAMPLE 17.1
ZAR=read.table("c:/DATA/Biostatistics/ZarEX17.1R.txt")
ZAR
attach(ZAR)
X=ageX
Y=winglY

#CREATING LINEAR MODEL LM:
LM=lm(Y~X)
LM
Yhat=fitted(LM)
e=residuals(LM)
RESULTS=data.frame(X,Y,Yhat,e)
RESULTS

#PLOTTING REGRESSION LINE & POINTS:
plot(X,Y)
abline(LM,col="blue")
segments(X,Yhat,X,Y,col="red")
```

> RESULTS

	X	Y	Yhat	e
1	3	1.4	1.523782	-0.12378156
2	4	1.5	1.794011	-0.29401057
3	5	2.2	2.064240	0.13576042
4	6	2.4	2.334469	0.06553142
5	8	3.1	2.874927	0.22507340
6	9	3.2	3.145156	0.05484439
7	10	3.2	3.415385	-0.21538462
8	11	3.9	3.685614	0.21438638
9	12	4.1	3.955843	0.14415737
10	14	4.7	4.496301	0.20369935
11	15	4.5	4.766530	-0.26652965
12	16	5.2	5.036759	0.16324134
13	17	5.0	5.306988	-0.30698767