ORIGIN ≡ 0

"Simple" Linear Regression

W. Stein

Linear Regression and the so-called "General Linear Model" represent a class of methods that seek to relate values of an observed "dependent" or "response" random variable (Y) that is Normally distributed to one or more "independent" "explanatory" or "predictor" variables (X) using a linear function. A "simple" regression involves only a single independent X variable. Linear models are "linear" in the sense that regression coefficients β_i of X only occur as in the equation of a line ($Y = \beta_0 + \beta_1 X$). Thus a model such as $Y = 5 + 23X$ (with regression coefficients β_0=5 & β_1=23) qualifies as a linear model, whereas $Y = \beta_0 + e^{\beta 1 X}$ does not. When a model (simple or otherwise) is linear in both regression coefficients and independent variables then it is termed a "first order model". 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, taking the square root allows one to model $Y = X^2$ as $Y = a\sqrt{X}$, and taking logs allows one to model the famous allometric equation: $Y = aX^b$ as $\ln(Y) = \ln(a) + b(\ln(X))$, among many possibilities.

Terminology and Page Numbers drawn from Kuter et al. (KNNL) A*pplied Linear Statistical Models* 5th Edition

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 ~ $N(\mu,\sigma^2)$.**
- **$X_1, X_2, X_3, \ldots, X_n$ (independent variable) with each value of X_i matched to Y_i**

Model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where: β_0 is the y intercept of the regression line (translation),

β_1 is the slope of the regression line (scaling coefficient),

ε_i is the error factor in prediction of Y_i and

a random variable distributed as $N(0,\sigma^2)$.

Note that although the Normality assumption is not strictly necessary in estimation of regression coefficients (β_i) using Least Squares Estimation, it applies in most other situations such as when using Maximum Liklihood Estimation, or statistical inferences.

Least Squares Estimation of the Regression Parameters:

Sums of Squares and Cross Products corrected for mean location:

This is done by subtracting each value of the independent and dependent variables from their respective means (X_{bar}, Y_{bar}), squaring the result and summing.

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

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

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

Estimated Regression Coefficients for $Y_i = \beta_0 + \beta_1 X_i$

$$b_1 := \frac{L_{xy}}{L_{xx}}$$ **< sample estimate of slope β_1** **See KNNL Eq.1.10a**

$$b_0 := Y_{bar} - b_1 \cdot X_{bar}$$ **< sample estimate of intercept β_0** **See KNNL Eq.1.10b**

Point Estimate of the Mean response Y (Y_{hat}):

$$Y_{h_i} := b_0 + b_1 \cdot X_i$$ **< using estimated coefficients and each value of the independent variable to estimate dependent value points on the Regression line.**

Residuals:

$e_i := Y_{h_i} - Y_i$ **< deviation of each value Y_i from Regression line = Yhat$_i$**

Point Estimate of Variance of Y_h called s^2 or MSE (estimating underlying population σ^2):

$SSE := \sum_i \left(e_i\right)^2$ **< Sum of Squares Error**

$MSE := \frac{SSE}{n - 2}$ **< Mean Squares Error**

^ Note that two degrees of freedom are lost here, df=(n-2), because two regression coefficients β_0 and β_1 need to be estimated from the data.

Maximum Liklihood Estimation of the Regression Parameters:

$\beta hat_1 := b_1$ **slope estimate is the same as in LS**

$\beta hat_0 := b_0$ **intercept estimate is the same as in LS**

$\sigma sq_{hat} := \frac{SSE}{n}$ **or equivalently:** $\frac{(n-2)}{n} \cdot MSE$

Note: these estimations require Y~N(μ,σ²) whereas Least Squares (LS) do not.

Example:

KNNL Toluca Example p. 21 $K := READPRN("c:/2008LinearModelsData/Toluca.txt")$

Assumptions:

- **Let independent variable X be Lot Size in the first column of K**
- **Let dependent variable Y be Work Hours in the second column of K**
- **Y is a random sample ~N(μ,σ²)**

Model:

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Basic Statistics:

means:

$X := K^{\langle 0 \rangle}$ $X_{bar} := mean(X)$ $X_{bar} = 70$

$Y := K^{\langle 1 \rangle}$ $Y_{bar} := mean(Y)$ $Y_{bar} = 312.28$

number of observations:

$n := length(Y)$ $i := 0 .. n - 1$ $n = 25$

$K =$

	0	1
0	80	399
1	30	121
2	50	221
3	90	376
4	70	361
5	60	224
6	120	546
7	80	352
8	100	353
9	50	157
10	40	160
11	70	252
12	90	389
13	20	113
14	110	435
15	100	420
16	30	212
17	50	268
18	90	377
19	110	421
20	30	273
21	90	468
22	40	244
23	80	342
24	70	323

Least Squares Estimation of the Regression Parameters:

Sums of Squares and Cross Products corrected for mean location:

$$L_{xx} := \sum_i \left(X_i - X_{bar}\right)^2 \qquad L_{xx} = 19800$$

$$L_{yy} := \sum_i \left(Y_i - Y_{bar}\right)^2 \qquad L_{yy} = 3.072 \times 10^5$$

$$L_{xy} := \sum_i \left(X_i - X_{bar}\right) \cdot \left(Y_i - Y_{bar}\right) \qquad L_{xy} = 70690$$

Estimated Regression Coefficients for $Y_i = \beta_0 + \beta_1 X_i$

$$b_1 := \frac{L_{xy}}{L_{xx}} \qquad b_1 = 3.5702$$ **< slope β_1 verified p. 22**

$$b_0 := Y_{bar} - b_1 \cdot X_{bar} \qquad b_0 = 62.3659$$ **< intercept β_0 verified p. 22**

Point Estimate of the Mean response Y (Y_{hat}):

$$Yh_i := b_0 + b_1 \cdot X_i$$

Residuals:

$$e_i := Yh_i - Y_i$$

Point Estimate of Variance of Y_i called s^2 or MSE:

$$SSE := \sum_i \left(e_i\right)^2 \qquad SSE = 54825.4592$$

^ verified p. 25

$$MSE := \frac{SSE}{n - 2} \qquad MSE = 2383.7156$$

^ verified p. 26

$Y_h =$

	0
0	347.982
1	169.4719
2	240.876
3	383.684
4	312.28
5	276.578
6	490.7901
7	347.982
8	419.3861
9	240.876
10	205.1739
11	312.28
12	383.684
13	133.7699
14	455.0881
15	419.3861
16	169.4719
17	240.876
18	383.684
19	455.0881
20	169.4719
21	383.684
22	205.1739
23	347.982
24	312.28

$e =$

	0
0	-51.018
1	48.4719
2	19.876
3	7.684
4	-48.72
5	52.578
6	-55.2099
7	-4.018
8	66.3861
9	83.876
10	45.1739
11	60.28
12	-5.316
13	20.7699
14	20.0881
15	-0.6139
16	-42.5281
17	-27.124
18	6.684
19	34.0881
20	-103.5281
21	-84.316
22	-38.8261
23	5.982
24	-10.72

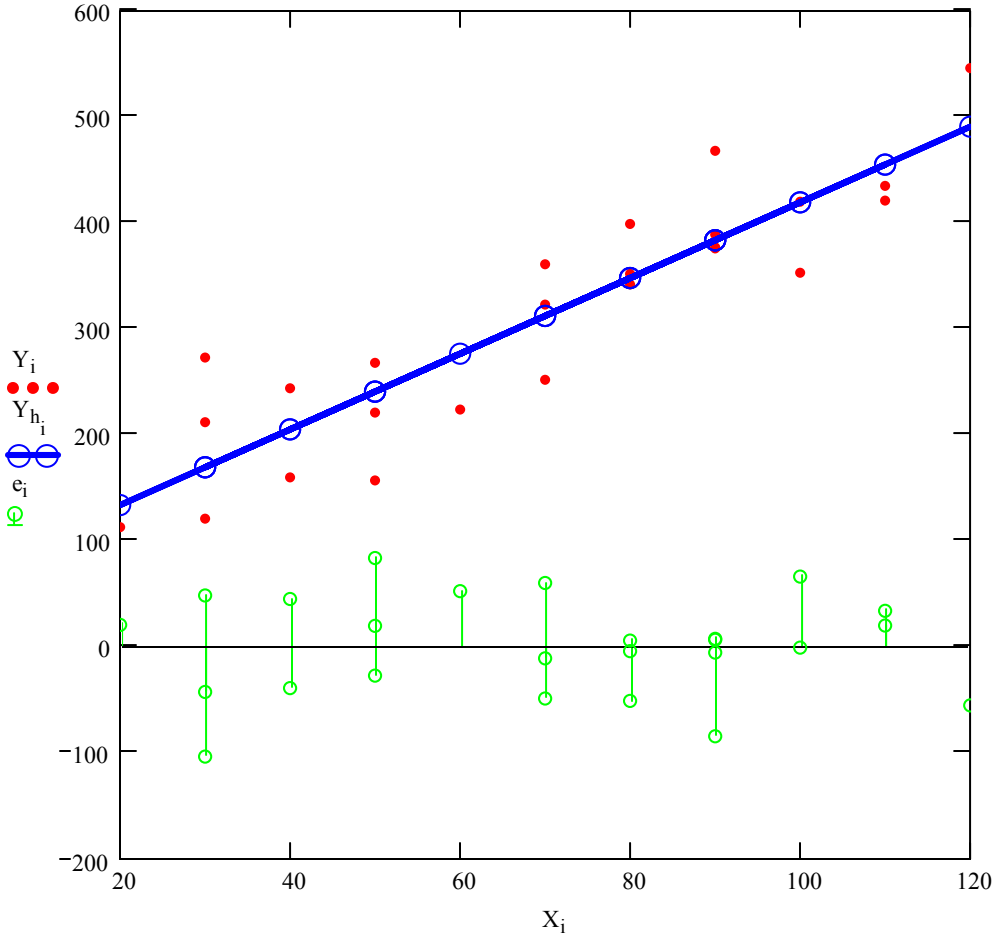
Maximum Liklihood Estimation of the Regression Parameters:

$$\beta hat_1 := b_1 \qquad \beta hat_1 = 3.5702$$

$$\beta hat_0 := b_0 \qquad \beta hat_0 = 62.3659$$

$$\sigma sq_{hat} := \frac{SSE}{n} \qquad \sigma sq_{hat} = 2193.0184$$

Plot of Values:



Prototype in R:

COMMANDS:

```
#READ TABLE AND ASSIGN VARIABLES FROM COLUMNS:
K=read.table("c:/2008LinearModelsData/Toluca.txt")
attach(K)
X=V1
Y=V2

#DESCRIPTIVE INFORMATION:
K
lsfit(X,Y)

#CONSTRUCTING LINEAR REGRESSION:
lm(Y~X)
#COLLECTING INFORMATION FROM THE REGRESSION:
Yh=predict(lm(Y~X))
RESIDUALS=resid(lm(Y~X))
#REPORTING WITHIN A DATAFRAME:
RESULTS=data.frame(X,Y,Yh,RESIDUALS)
RESULTS

#PLOTTING DATA, REGRESSION LINE AND RESIDUALS:
plot(X,Y)
points(X,Yh,col=”blue”)
abline(lm(Y~X),col="blue")
segments(X,Yh,X,Y,col="red")
```

Results:

```
Coefficients:
(Intercept)            X
      62.37         3.57
```

	X	Y	Yh	RESIDUALS
1	80	399	347.9820	51.0179798
2	30	121	169.4719	-48.4719192
3	50	221	240.8760	-19.8759596
4	90	376	383.6840	-7.6840404
5	70	361	312.2800	48.7200000
6	60	224	276.5780	-52.5779798
7	120	546	490.7901	55.2098990
8	80	352	347.9820	4.0179798
9	100	353	419.3861	-66.3860606
10	50	157	240.8760	-83.8759596
11	40	160	205.1739	-45.1739394
12	70	252	312.2800	-60.2800000
13	90	389	383.6840	5.3159596
14	20	113	133.7699	-20.7698990
15	110	435	455.0881	-20.0880808
16	100	420	419.3861	0.6139394
17	30	212	169.4719	42.5280808
18	50	268	240.8760	27.1240404
19	90	377	383.6840	-6.6840404
20	110	421	455.0881	-34.0880808
21	30	273	169.4719	103.5280808
22	90	468	383.6840	84.3159596
23	40	244	205.1739	38.8260606
24	80	342	347.9820	-5.9820202
25	70	323	312.2800	10.7200000