Math 362: Mathmatical Statistics II

Le Chen
le.chen@emory.edu

Emory University
Atlanta, GA

Last updated on February 25, 2020

2020 Spring

Chapter 11. Regression

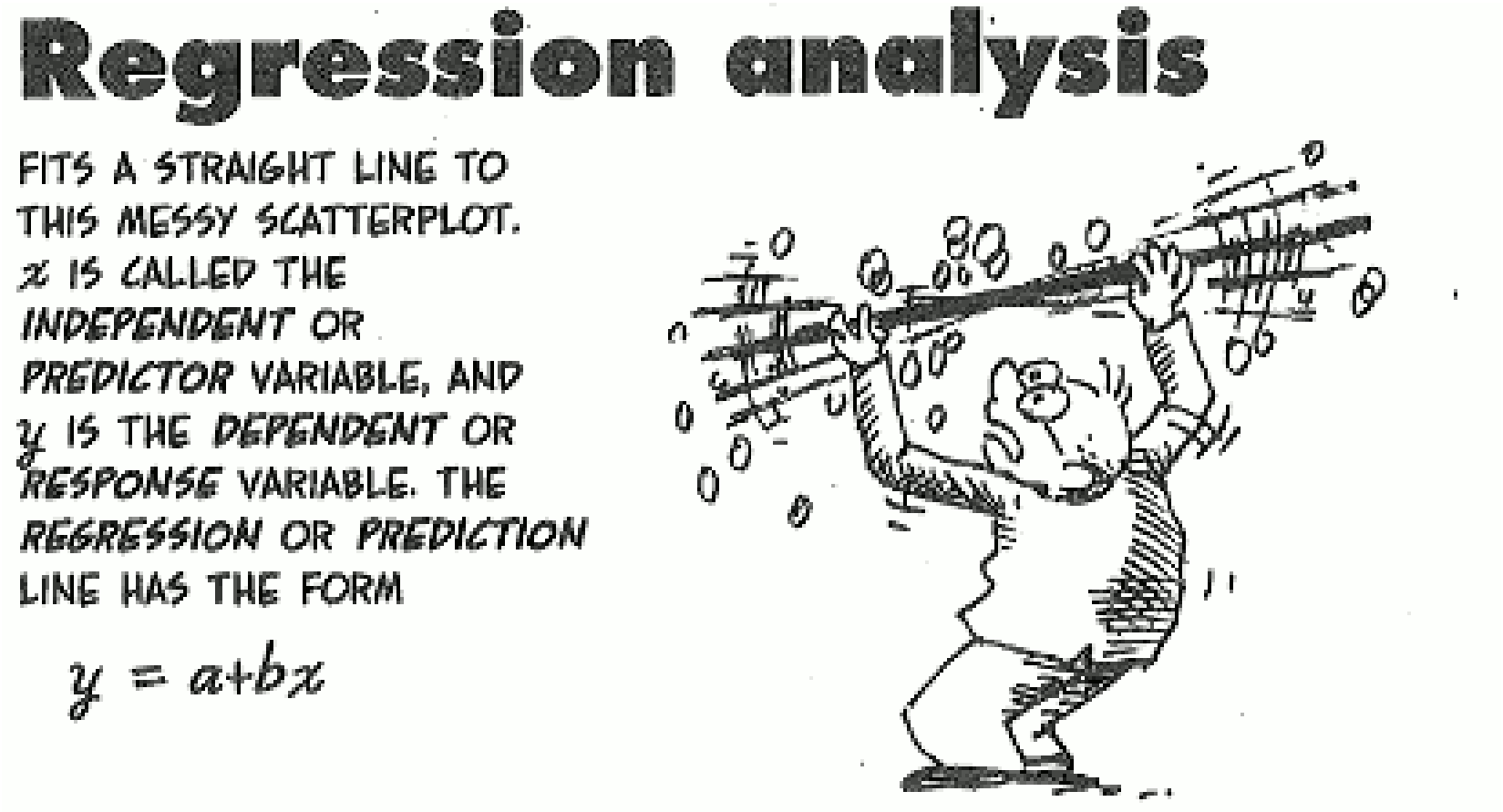
Chapter 11. Regression

§ 11.1 Introduction

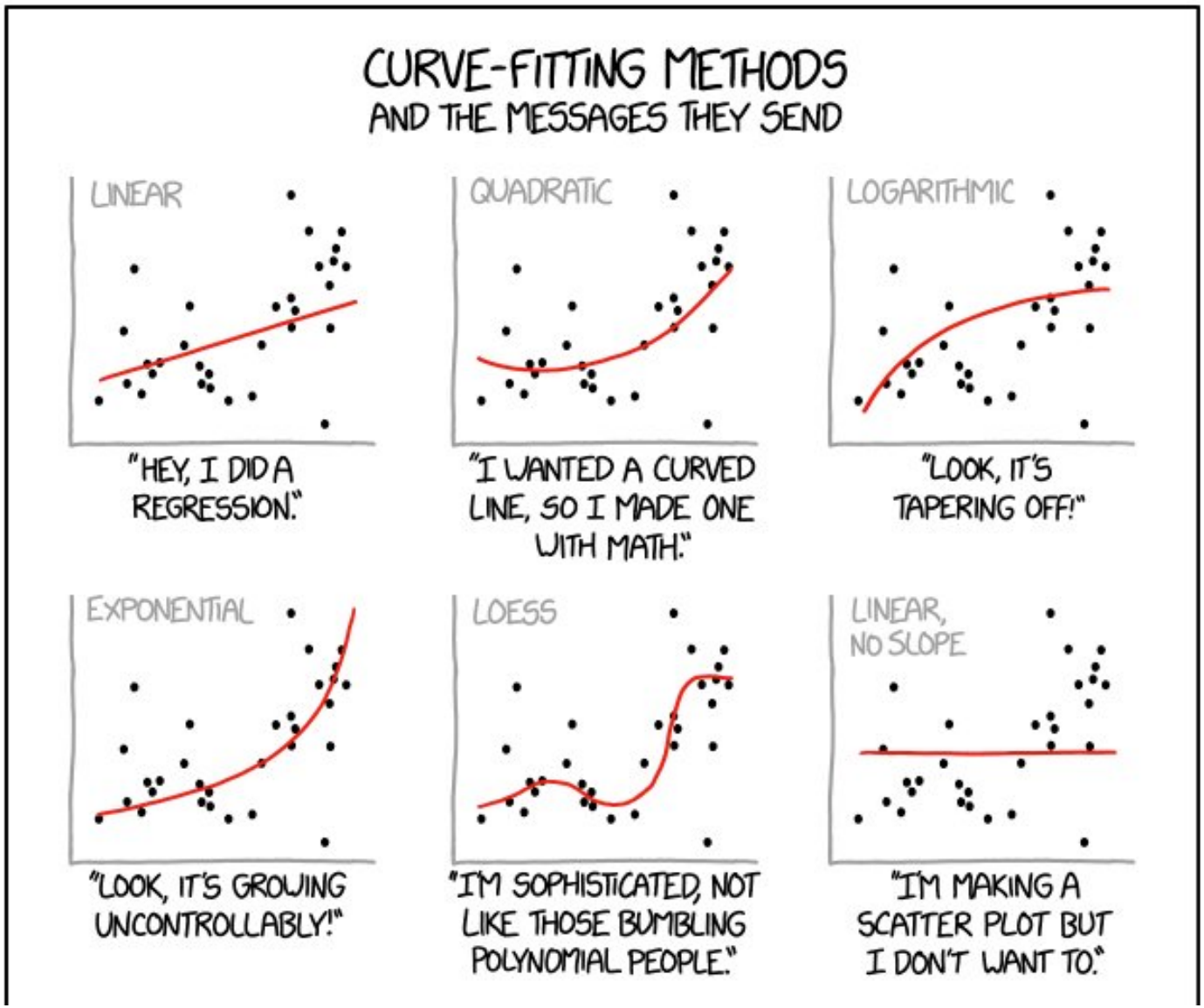
§ 11.2 The Method of Least Squares

§ 11.3 The Linear Model

§ 11.1 Introduction



https://madhureshkumar.wordpress.com/



https://xkcd.com/

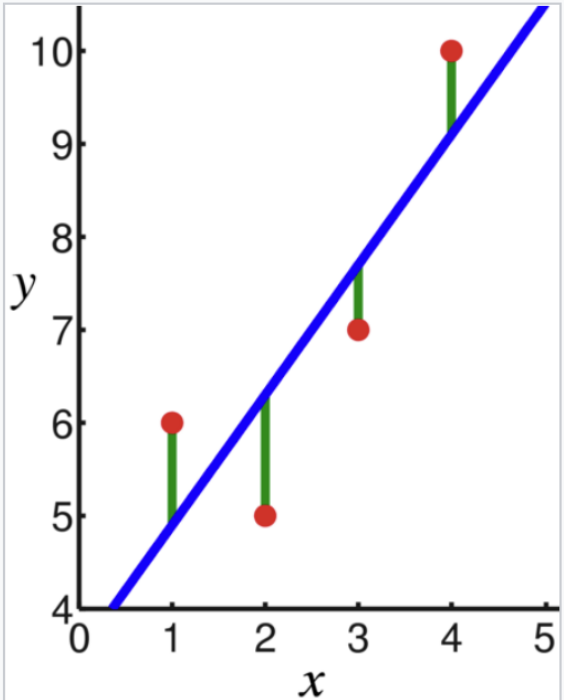
Chapter 11. Regression

§ 11.1 Introduction

§ 11.2 The Method of Least Squares

§ 11.3 The Linear Model

§ 11.2 The Method of Least Squares



In linear regression, the observations (**red**) are assumed to be the result of random deviations (**green**) from an underlying relationship (**blue**) between a dependent variable (*y*) and an independent variable (*x*).

Goal: Find a blue line that minimizes the sum of the square of the green lines

Thm. Given n points $(x_1, y_1), \cdots, (x_n, y_n)$, the straight line $y = a + bx$ minimizing

$$L = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

when

$$b = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x}.$$

Proof. ...

□

1. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

2. $$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n^2}$$
$$s_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \frac{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}{n^2}$$

3. $$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n^2}$$

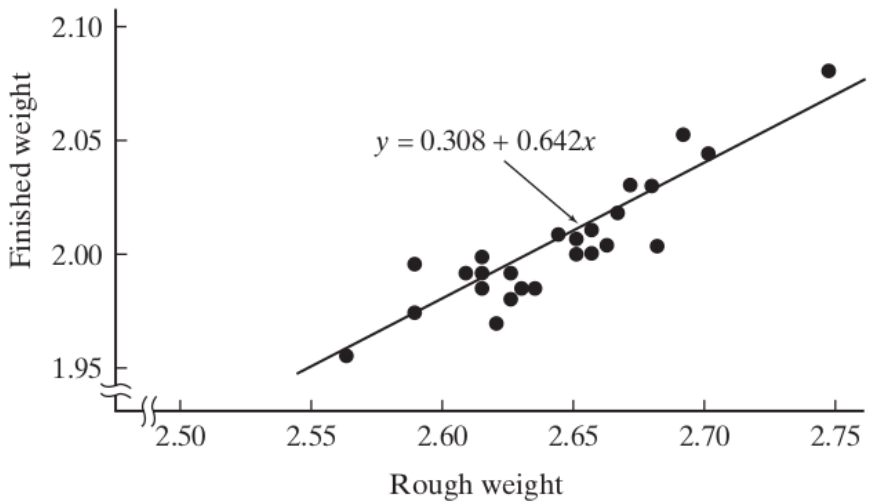
4. $$\rho_{XY} = \frac{s_{XY}}{s_X s_Y} \qquad \Longrightarrow \qquad b = \rho_{XY}\frac{s_Y}{s_X} = \frac{s_{XY}}{s_X^2}$$

E.g. 1 Producing air conditioners. $x =$ rough weight of a rod. $y =$ finished weight. Find the best linear approximation of xy-relationship. Predict the weight when $x = 2.71$

Table 11.2.1

Rod Number	Rough Weight, x	Finished Weight, y	Rod Number	Rough Weight, x	Finished Weight, y
1	2.745	2.080	14	2.635	1.990
2	2.700	2.045	15	2.630	1.990
3	2.690	2.050	16	2.625	1.995
4	2.680	2.005	17	2.625	1.985
5	2.675	2.035	18	2.620	1.970
6	2.670	2.035	19	2.615	1.985
7	2.665	2.020	20	2.615	1.990
8	2.660	2.005	21	2.615	1.995
9	2.655	2.010	22	2.610	1.990
10	2.655	2.000	23	2.590	1.975
11	2.650	2.000	24	2.590	1.995
12	2.650	2.005	25	2.565	1.955
13	2.645	2.015			

Sol. ...


2.10
2.05
2.00
1.95
Finished weight
$y = 0.308 + 0.642x$
2.50
2.55
2.60
2.65
2.70
2.75
Rough weight

...

□

Def. Let a and b be the least squares coefficients with the sample $(x_1, y_1), \cdots, (x_n, y_n)$.

$\hat{y} = a + bx$: **predicted value** of y

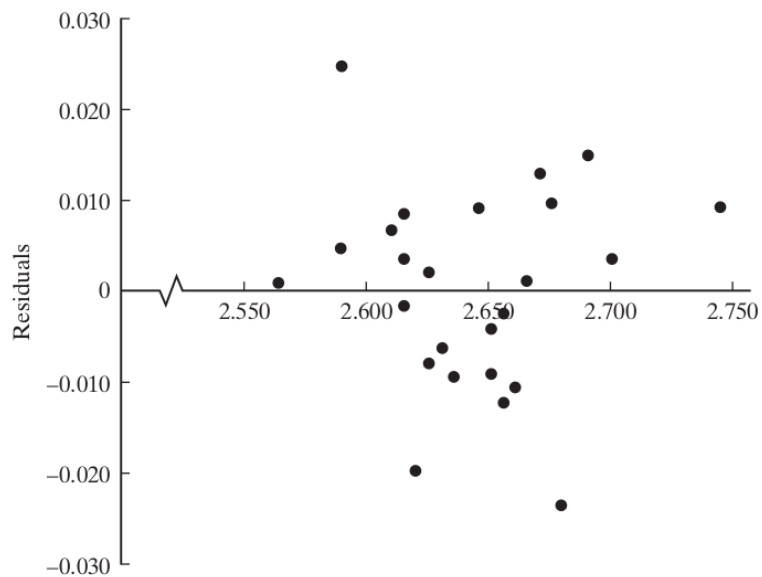
$y_i - \hat{y}_i = y_i - (a + bx_i)$: i**th residual**

Rem. Use the resigual plots to assessing the model.

E.g. 1' Here are the residues and their plots:

Table 11.2.2

x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
2.745	2.080	2.070	0.010
2.700	2.045	2.041	0.004
2.690	2.050	2.035	0.015
2.680	2.005	2.029	−0.024
2.675	2.035	2.025	0.010
2.670	2.035	2.022	0.013
2.665	2.020	2.019	0.001
2.660	2.005	2.016	−0.011
2.655	2.010	2.013	−0.003
2.655	2.000	2.013	−0.013
2.650	2.000	2.009	−0.009
2.650	2.005	2.009	−0.004
2.645	2.015	2.006	0.009
2.635	1.990	2.000	−0.010
2.630	1.990	1.996	−0.006
2.625	1.995	1.993	0.002
2.625	1.985	1.993	−0.008
2.620	1.970	1.990	−0.020
2.615	1.985	1.987	−0.002
2.615	1.990	1.987	0.003
2.615	1.995	1.987	0.008
2.610	1.990	1.984	0.006
2.590	1.975	1.971	0.004
2.590	1.995	1.971	0.024
2.565	1.955	1.955	0.000



E.g. 2 Predict the Social Security expenditures.

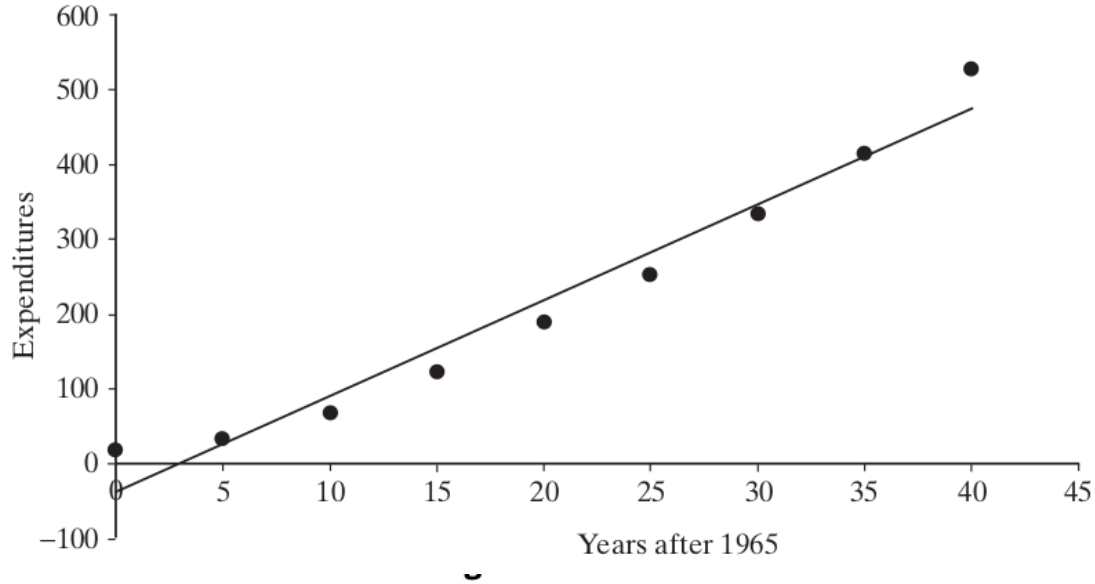
Table 11.2.3

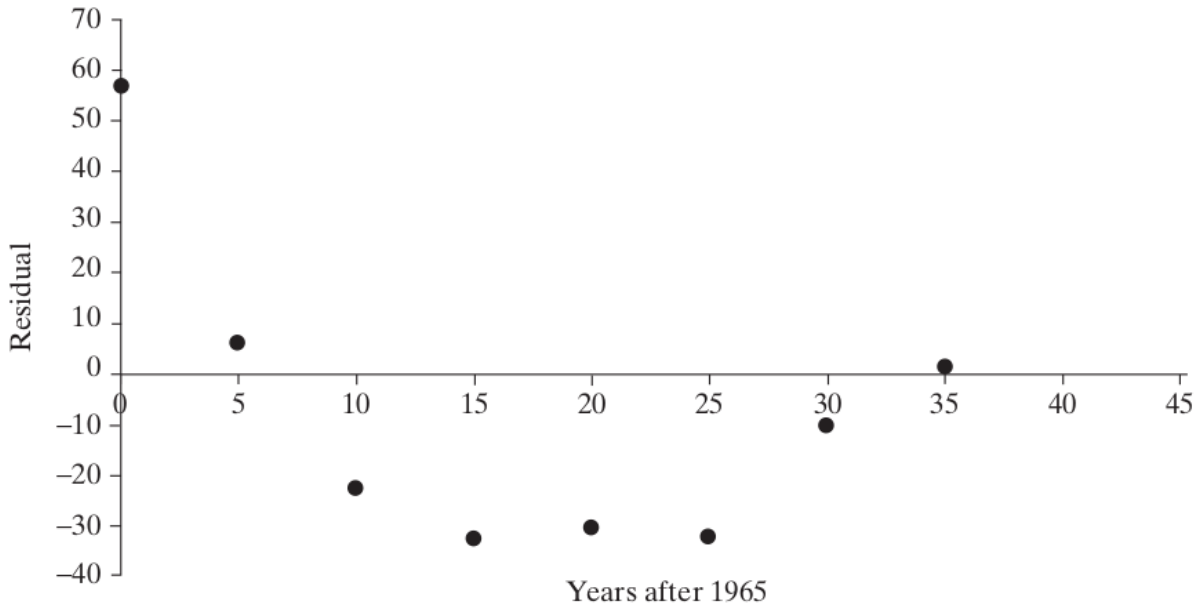
Year	Years after 1965, x	Social Security Expenditures ($ billions), y
1965	0	19.2
1970	5	33.1
1975	10	69.2
1980	15	123.6
1985	20	190.6
1990	25	253.1
1995	30	339.8
2000	35	415.1
2005	40	529.9

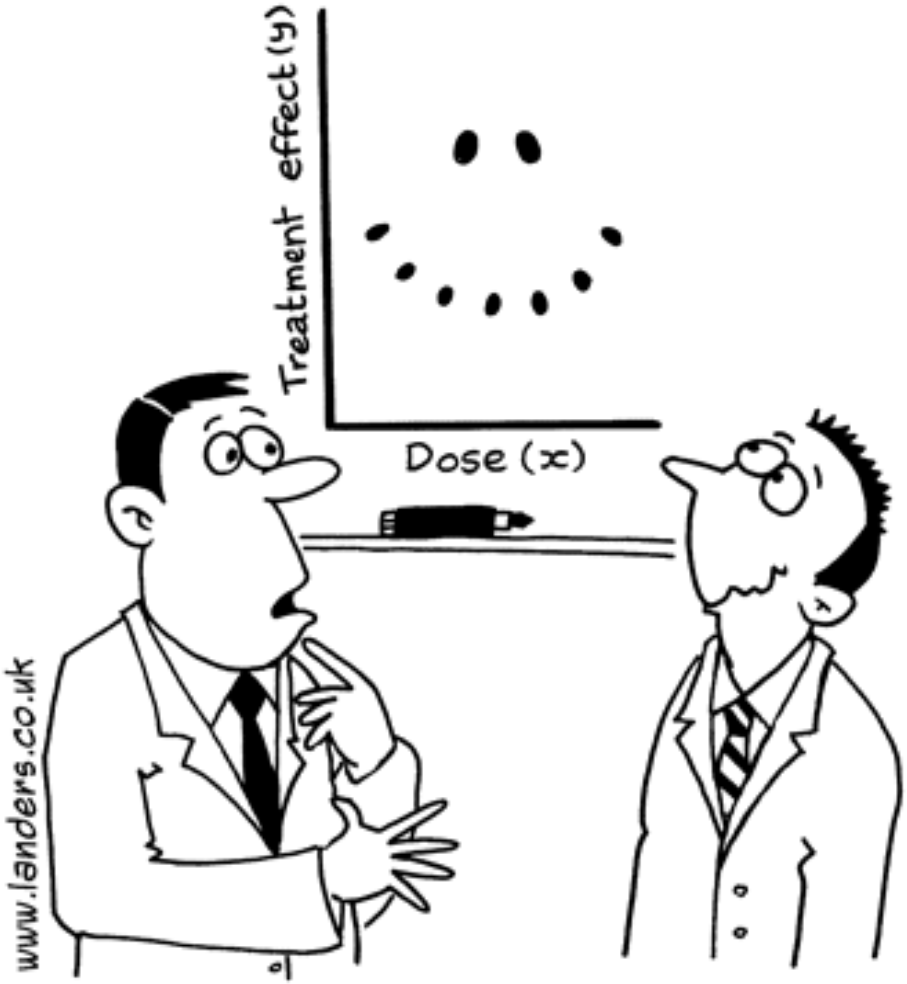
Source: www.socialsecurity.gov/history/trustfunds.html.

Does the the least squares line $y = -38.0 + 12.9x$ a good model to predict the cost in 2010 would be \$543, i.e., the case $x = 45$?

Sol.

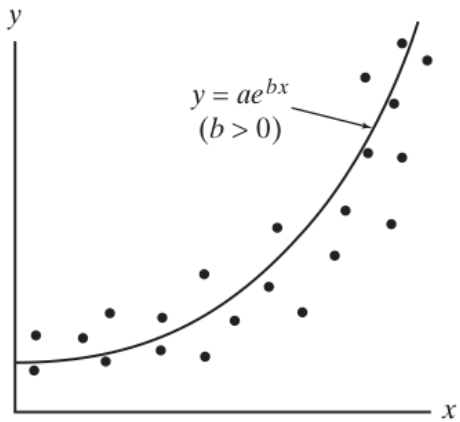

600
500
400
300
200
100
0
−100
Expenditures
0
5
10
15
20
25
30
35
40
45
Years after 1965

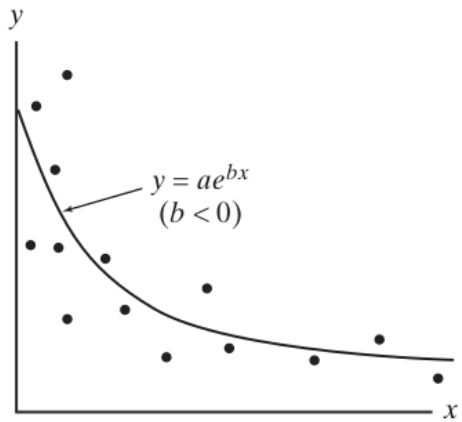

70
60
50
40
30
20
10
0
−10
−20
−30
−40
Residual
0
5
10
15
20
25
30
35
40
45
Years after 1965



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

Exponential Regression





$$y = ae^{bx} \quad \Longleftrightarrow \quad \ln y = \ln a + bx$$

$$b = \frac{n\sum_{i=1}^{n} x_i \ln y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} \ln y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad \ln a = \frac{\sum_{i=1}^{n} \ln y_i - b\sum_{i=1}^{n} x_i}{n}$$

E.g. Moore's law:

Gordon Moore predicted in 1965 that the number of transistors per chip would double every 18 months.

Based on the real data, check:

1) Whether is the chip capacity doubling at a fixed rate?

2) Find out the rate.

Table 11.2.5

Chip	Year	Years after 1975, x	Transistors per Chip, y
8080	1975	0	4,500
8086	1978	3	29,000
80286	1982	7	90,000
80386	1985	10	229,000
80486	1989	14	1,200,000
Pentium	1993	18	3,100,000
Pentium Pro	1995	20	5,500,000

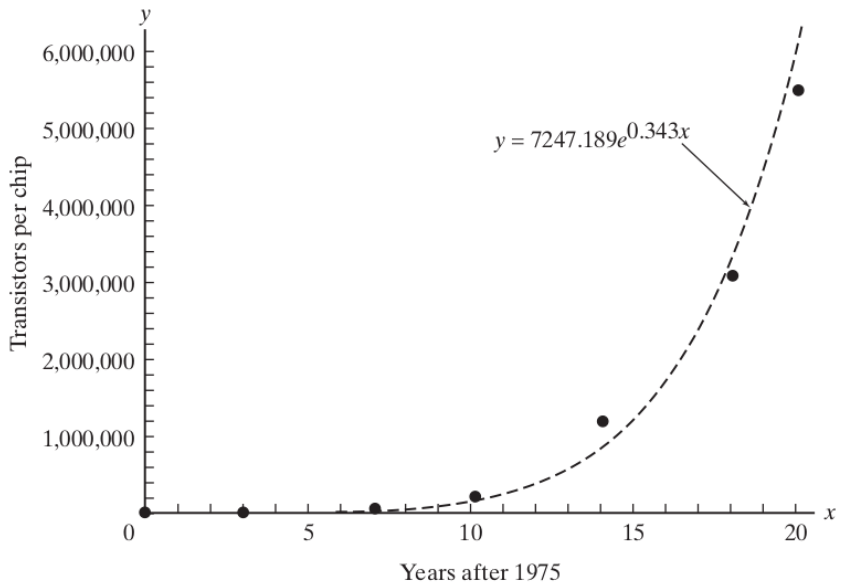
Source: en.wikipedia.org/wiki/Transistor—count.

Sol. To check whether chip capacity doubles in a fixed rate, one needs to carry out exponential regression:

Table 11.2.6

Years after 1975, x_i	x_i^2	Transistors per Chip, y_i	$\ln y_i$	$x_i \cdot \ln y_i$
0	0	4,500	8.41183	0
3	9	29,000	10.27505	30.82515
7	49	90,000	11.40756	79.85292
10	100	229,000	12.34148	123.41480
14	196	1,200,000	13.99783	195.96962
18	324	3,100,000	14.94691	269.04438
20	400	5,500,000	15.52026	310.40520
72	1078		86.90093	1009.51207

$$\Longrightarrow \quad b = \cdots = 0.342810, \quad a = \cdots = e^{\ln a} = e^{8.89} = 7247.189.$$



Finally, to find out the rate:

$$e^{0.343x} = e^{\ln 2 \times \frac{0.343}{\ln 2} x} = 2^{\frac{0.343}{\ln 2} x}$$

$$\frac{0.343}{\ln 2} x = 1 \quad \Longrightarrow \quad x = \frac{\ln 2}{0.343} = 2.020837.$$

□

Other curvilinear models

Table 11.2.10

a. If $y = ae^{bx}$, then $\ln y$ is linear with x.

b. If $y = ax^b$, then $\log y$ is linear with $\log x$.

c. If $y = L/(1 + e^{a+bx})$, then $\ln\left(\dfrac{L-y}{y}\right)$ is linear with x.

d. If $y = \dfrac{1}{a+bx}$, then $\dfrac{1}{y}$ is linear with x.

e. If $y = \dfrac{x}{a+bx}$, then $\dfrac{1}{y}$ is linear with $\dfrac{1}{x}$.

f. If $y = 1 - e^{-x^b/a}$, then $\ln\ln\left(\dfrac{1}{1-y}\right)$ is linear with $\ln x$.

Chapter 11. Regression

§ 11.1 Introduction

§ 11.2 The Method of Least Squares

§ 11.3 The Linear Model

§ 11.3 The Linear Model

Def. The function $f(X)$ for which

$$\mathbb{E}\left[(Y - f(X))^2\right]$$

is minimized is called the **regression curve of** Y **on** X.

Thm. Let (X, Y) be two random variables such that $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ both exist. Then the regression cure of Y on X is given (for all x) by

$$f(x) = \mathbb{E}\left[Y|X = x\right].$$

Proof. Let $f(x) = \mathbb{E}[Y|X = x]$ and let $\phi(x)$ be a general function. Then

$$\begin{aligned}\mathbb{E}\left[(Y-\phi(X))^2\right] =& \mathbb{E}\left[([Y-f(X)]+[f(x)-\phi(x)])^2\right] \\ =& \mathbb{E}\left[(Y-f(X))^2\right] + \mathbb{E}\left[(f(x)-\phi(X))^2\right] \\ &+ \mathbb{E}\left[(Y-f(X))(f(x)-\phi(X))\right].\end{aligned}$$

Let $\psi(x)$ be either $f(x)$ or $\phi(x)$. We claim that

$$\mathbb{E}\left[(Y-f(X))\psi(X)\right] = 0.$$

Indeed,

$$\begin{aligned}\mathbb{E}[Y\psi(X)] &= \iint_{\mathbb{R}^2} f_{X,Y}(x,y)y\psi(x)\mathrm{d}y\mathrm{d}x \\ &= \int_{\mathbb{R}} \mathrm{d}x\psi(x)f_X(x)\underbrace{\int_{\mathbb{R}} \mathrm{d}y \frac{f_{X,Y}(x,y)}{f_x(x)}y}_{=\mathbb{E}[Y|X=x]} \\ &= \mathbb{E}[f(X)\psi(X)].\end{aligned}$$

Hence,

$$\mathbb{E}\left[(Y-\phi(X))^2\right] = \mathbb{E}\left[(Y-f(X))^2\right] + \mathbb{E}\left[(f(x)-\phi(X))^2\right]$$

which is minimized when $\phi(x) = f(x)$. □

Simple linear model

Difficulties: in general, the regression curve $y = \mathbb{E}[Y|x]$ is very complex and hard to find.

(Simple) linear model:

1. $f_{Y|x}(y)$ is a normal pdf for any x given.

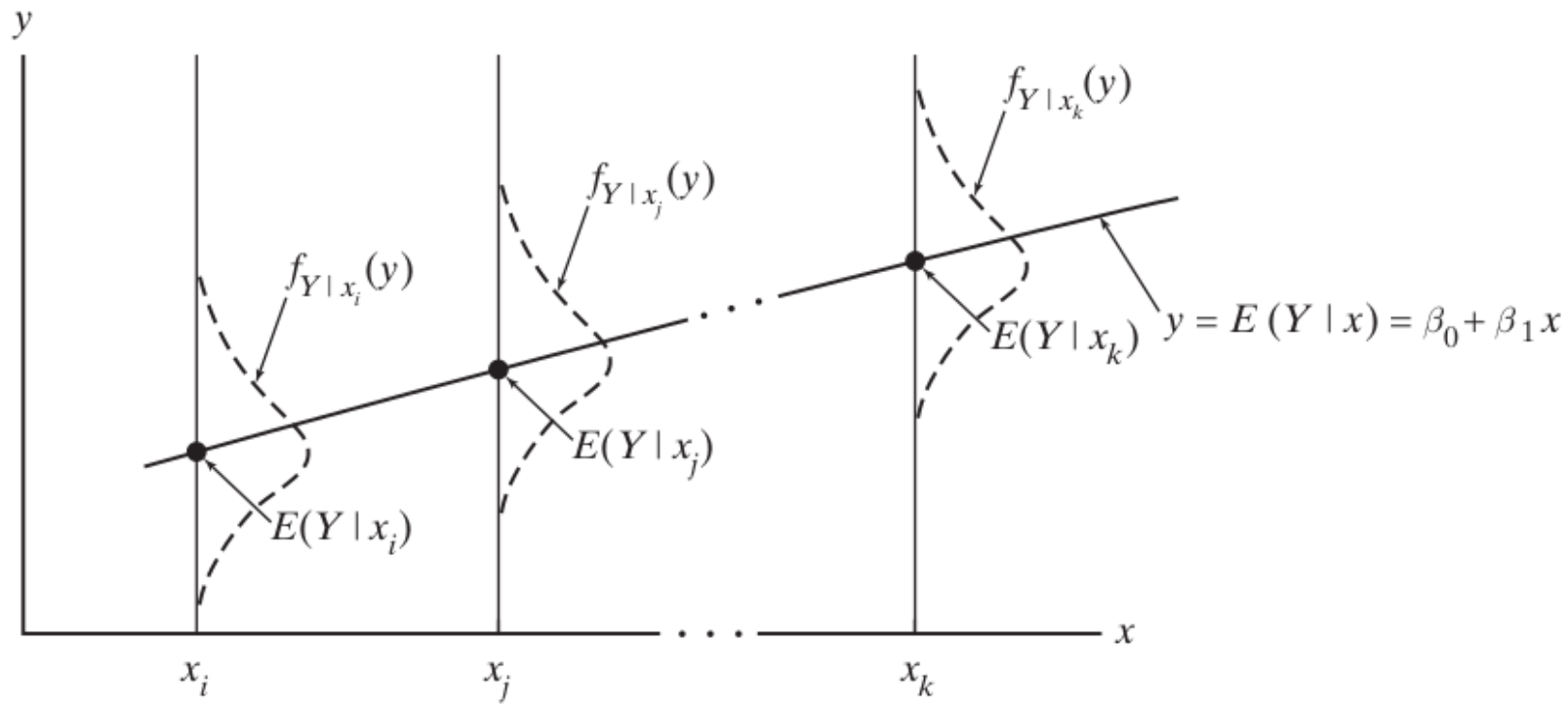
2. The standard deviation, σ, of $Y|x$ is the same for all x, i.e.,

$$\sigma^2 \equiv \mathbb{E}[Y^2|x] - \mathbb{E}[Y|x]^2.$$

3. The mean of $Y|x$ is collinear, i.e.,

$$y = \mathbb{E}[Y|x] = \beta_0 + \beta_1 x.$$

4. All of the conditional distributions represnt independent random variables.


y
$f_{Y \mid x_i}(y)$
$f_{Y \mid x_j}(y)$
$f_{Y \mid x_k}(y)$
$E(Y \mid x_i)$
$E(Y \mid x_j)$
$E(Y \mid x_k)$
$y = E(Y \mid x) = \beta_0 + \beta_1 x$
x
x_i
x_j
x_k

MLE for linear model

Thm. Let $(x_1, Y_1), \cdots, (x_n, Y_n)$ be a set of points satisfying the linear model, $\mathbb{E}[Y|x] = \beta_0 + \beta_1 x$. The maximum likelihood estimators for β_0, β_1 and σ^2 are given by

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^n x_i Y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n Y_i\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2}$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n Y_i - b\sum_{i=1}^n x_i}{n} = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \widehat{Y}_i\right)^2, \qquad \widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

Proof. ...

□

Properties of linear model estimators

1. $\hat{\beta}_0$ and $\hat{\beta}_1$ are both normally distributed.
2. $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased: $\mathbb{E}[\hat{\beta}_0] = \beta_0$ and $\mathbb{E}[\hat{\beta}_1] = \beta_1$.
3. Variances are eqal to

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x - \bar{x})^2}$$

$$\text{Var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

4. $\hat{\beta}_1$, $\overline{Y}$ and $\hat{\sigma}^2$ are mutually independent. $\implies$ $\widehat{Y}_i \perp \hat{\sigma}^2$
5. $\frac{n\hat{\sigma}^2}{\sigma^2} \sim$ Chi Square with $n - 2$ degrees of freedom. $\implies \mathbb{E}[\sigma^2] = \frac{n-2}{n}\sigma^2$

Proof. ...

□

Estimating σ^2

1. MLE:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2.$$

2. The unbiased estimator:

$$S^2 = \frac{n}{n-2}\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2.$$

Notation

Parameter	Estimator	Estimate
β_1	$\hat{\beta}_1$	β_{1e}
β_0	$\hat{\beta}_0$	β_{0e}
σ	S	s
σ^2	S^2	s^2
σ^2	$\hat{\sigma}^2$	σ_e^2
	$\overline{Y}$	$\bar{y}$
	$\widehat{Y}_i$	$\hat{y}_i = \beta_{0e} + \beta_{1e} x_i$

Drawing inferences on β_1

Thm. $T_{n-2} = \dfrac{\hat{\beta}_1 - \beta_1}{S \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim$ Student t distribution with df $= n - 2$.

1. Hypothesis test $H_0 : \beta_1 = \beta_1'$ vs. ....

2. C.I. for β_1: $\beta_{1e} \pm t_{\alpha/2,n-2} \dfrac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})}}$

E.g. Does smoking contribute to coronary heat disease?

Table 11.3.1

Country	Cigarette Consumption per Adult per Year, x	CHD Mortality per 100,000 (ages 35–64), y
United States	3900	256.9
Canada	3350	211.6
Australia	3220	238.1
New Zealand	3220	211.8
United Kingdom	2790	194.1
Switzerland	2780	124.5
Ireland	2770	187.3
Iceland	2290	110.5
Finland	2160	233.1
West Germany	1890	150.3
Netherlands	1810	124.7
Greece	1800	41.2
Austria	1770	182.1
Belgium	1700	118.1
Mexico	1680	31.9
Italy	1510	114.3
Denmark	1500	144.9
France	1410	59.7
Sweden	1270	126.9
Spain	1200	43.9
Norway	1090	136.3

1) Test $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 > 0$ at $\alpha = 0.05$.

2) Find C.I. for β_1 with the same α.

Sol.

□