Shape of the Earth: Method of least Squares

Credits:Talk (borrowed from)- Probal Chaudhuri's IMS public lecture

April 4, 2024

Modelling

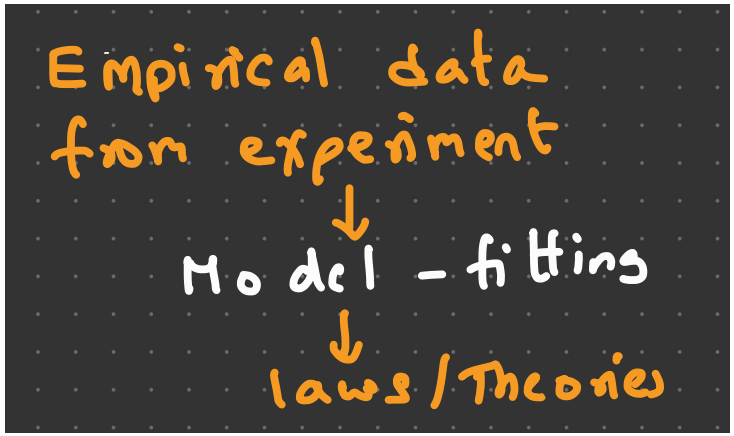
All models are wrong, but some are useful

- Example: $PV = RT$ pressure P, volume V and temperature T of an "ideal" gas via a constant R
- is not exactly true for any real gas.
- provides a useful approximation and furthermore its structure is informative.

A step-back in Time

Early Model-Fitting efforts in Science:

1. Mendel- Genetic laws
2. Keppler's Law
3. ...



What is the Shape of the Earth ?



Most of us now have advantage of Satelite images

What is the Shape of the Earth ?

- French astronomer Jean Reicher in South America in 1672
- Reicher's pendulum clock was losing 2 minutes 28 seconds every day in French Guyana.

- Isaac Newton and his Principia in 1687
- Newton claimed that the Earth's rotation around its axis made it an Oblate Sphere. Rotation caused flattening at the Poles and bulging at the Equator.

- Was not unanimously regarded as correct

Side View of oblate earth : *Berry, 1898, p 277*

. . . lengthening of degrees arc toward the pole . . .

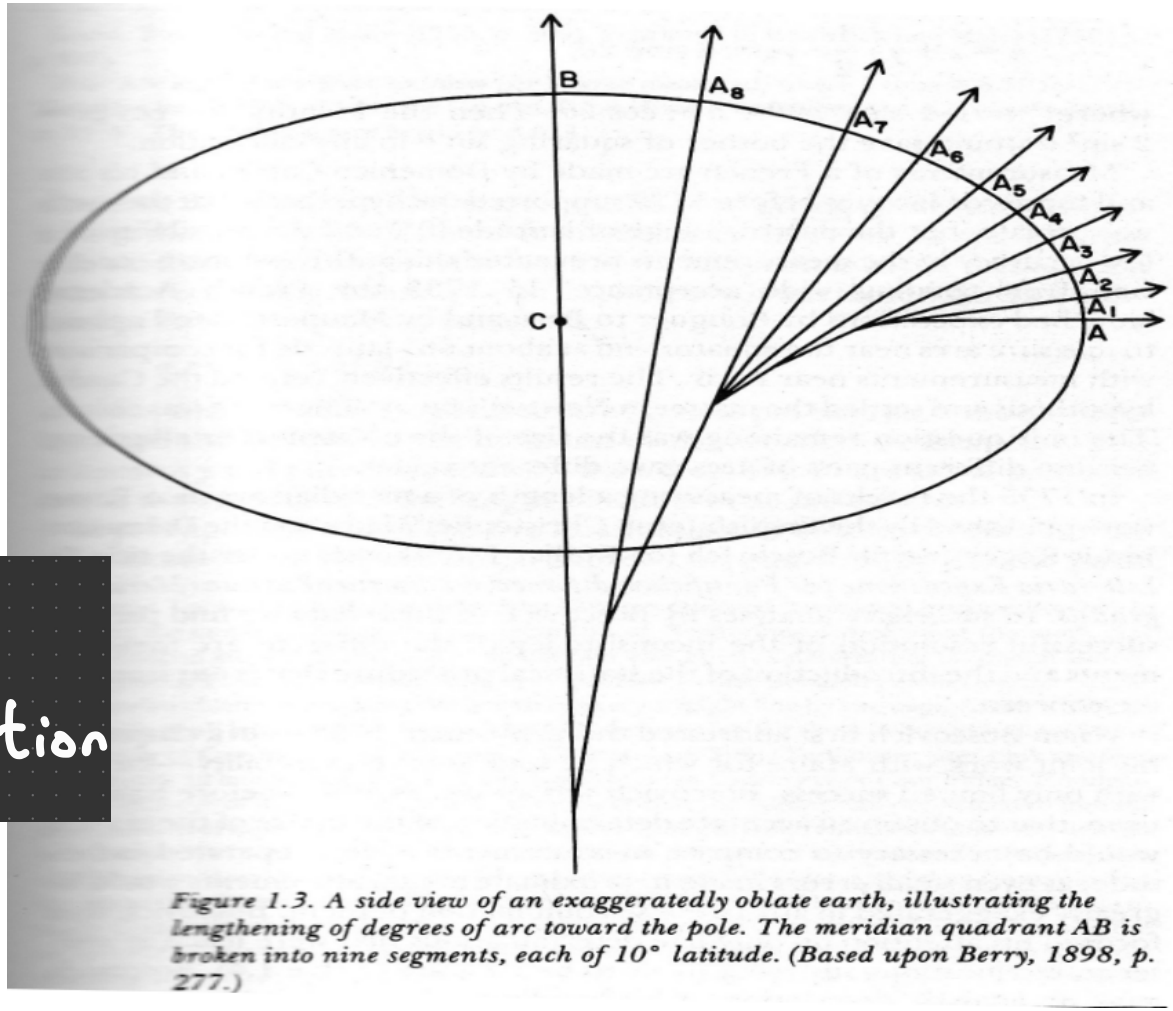


Figure 1.3. A side view of an exaggeratedly oblate earth, illustrating the lengthening of degrees of arc toward the pole. The meridian quadrant AB is broken into nine segments, each of 10° latitude. (Based upon Berry, 1898, p. 277.)

1898
Publication

. . . meridian quadrant AB is broken into 9 segments of 10° latitude

Arc Length and the Latitude

- Let Y be the length of a 1° meridian arc centered at latitude θ.
- Then, assuming the Earth to be an ellipsoid,

$$Y \approx m \sin^2(\theta) + c = mX + c,$$

where $X = \sin^2(\theta)$.

- Ellipticity of the Earth can be measured by $\frac{m}{3c}$.
- Newton estimated the ellipticity of the Earth to be $\frac{1}{230}$.

Newton and the French Scientists

- French arc measurements around 1720 over a range of 9° latitude. The project was led by Dominico Cassini and Jacques Cassini of the Royal Observatory in Paris.
- Their measurements and analysis contradicted Newton's model for the Earth.
- French Geodesic Mission during $1735 - 1736$. Expeditions to Quito in Ecuador and Lapland in Finland.

Bosocovitch Data on Meridian Arc

1755 - 1770 - period of work.

Table 1.4. Boscovich's data on meridian arcs.

Location	Latitude (θ)	Arc length (toises)	Boscovich's $\sin^2\theta \times 10^4$
(1) Quito	0°0′	56,751	0
(2) Cape of Good Hope	33°18′	57,037	2,987
(3) Rome	42°59′	56,979	4,648
(4) Paris	49°23′	57,074	5,762
(5) Lapland	66°19′	57,422	8,386

Source: Boscovich and Maire (1755, p. 500). Reprinted in Boscovich and Maire (1770, p. 482).

Note: Arc lengths are given as toises per degree measured, where 1 toise ≅ 6.39 feet. The value for $\sin^2\theta \times 10^4$ for the Cape of Good Hope is erroneous and is evidently based on 33°8′. The correct figure would be 3,014.

Bosocovich: Least Absolute Mean Deviation(1755)

We have data points

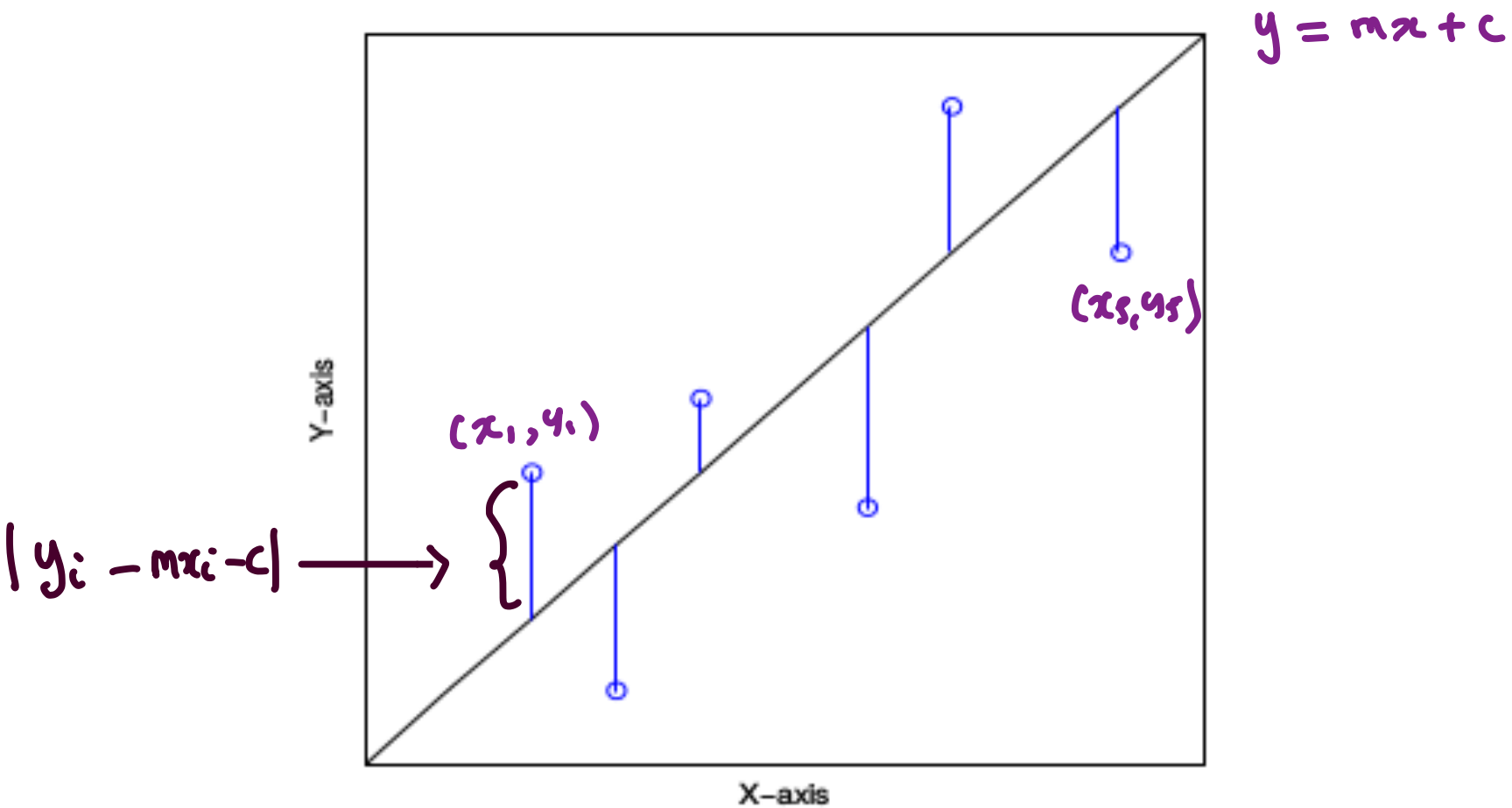
$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_5, Y_5).$$

where Y reprsents arc length, and X is $\sin^2$ (latitude).
To find the line : $Y = mX + c$ that fits the data best.

$$\min\{m, c : \sum_{i=1}^{5} |Y_i - mX_i - c|\}.$$

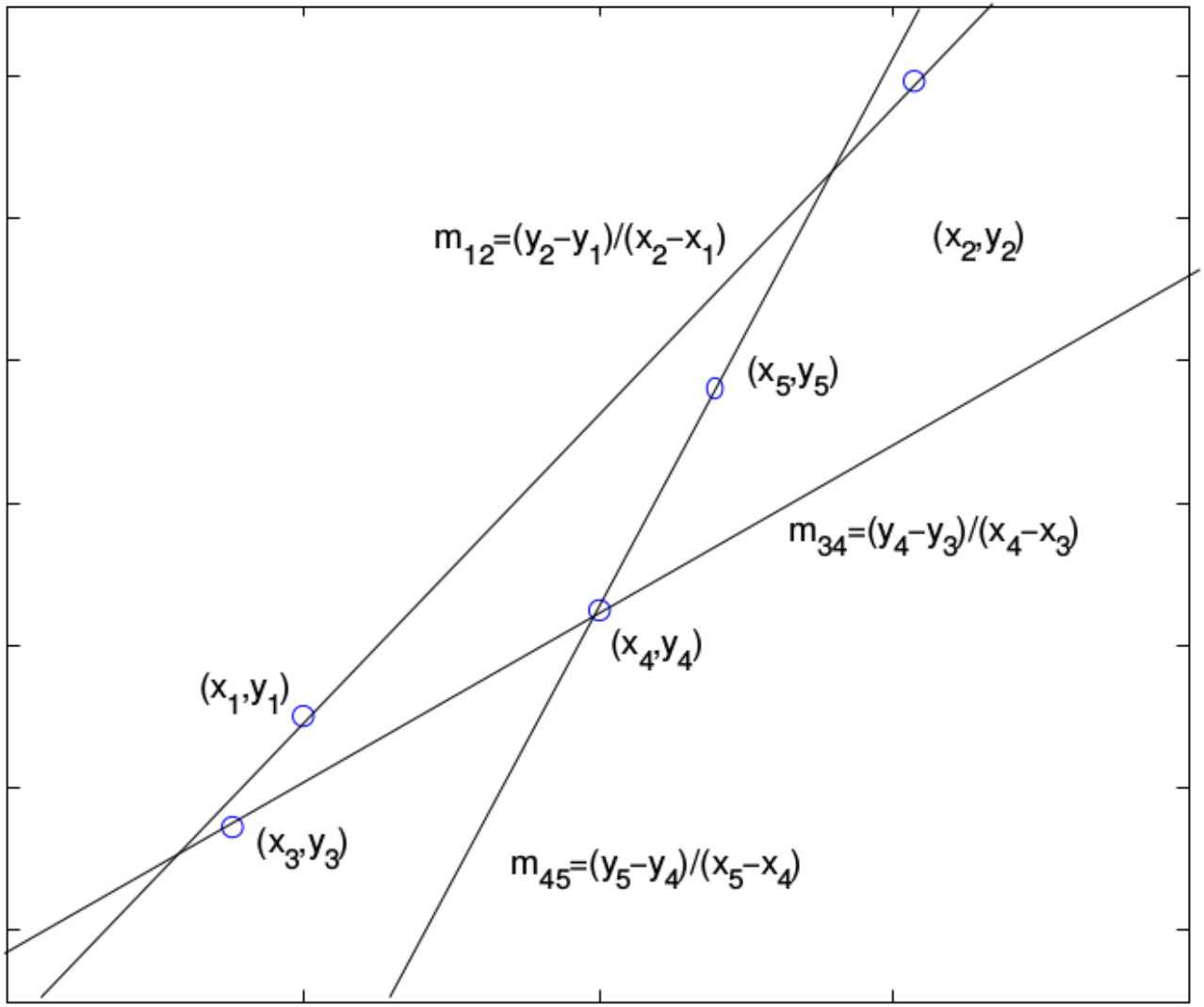
Given our calculus knowledge :- Hard Problem!

Least Absolute Deviation Line



Bosocovich : Pairwise Solution

Used: between two points unique line passes



Bosocovich : Pairwise Solution

Table 1.5. Boscovich's pairwise solutions, based on the data of Table 1.4, for the polar excess y (the amount by which a degree at the pole exceeds a degree at the equator) and the ellipticity (found from the formula $1/\text{ellipticity} = 3z/y$, where z is the length of a degree at the equator as found from the pair of equations).

Pair	Polar excess (y, in toises)	Ellipticity	Pair	Polar excess (y, in toises)	Ellipticity
1, 5	800	1/213	2, 4	133	1/128
2, 5	713	1/239	3, 4	853	1/200
3, 5	1,185	1/144	1, 3	491	1/347
4, 5	1,327	1/128	2, 3	−350	−1/486
1, 4	542	1/314	1, 2	957	1/78

Source: Boscovich and Maire (1755, p. 501). Reprinted in Boscovich and Maire (1770, p. 483).

Note: The ellipticities for pairs (2, 4) and (1, 2) were evidently misprinted in the original; they should be 1/1282 and 1/178. The figures for the pair (1, 4) are erroneous; they should be 560 and 1/304.

Note:- Variation in Ellipticity values

Laplace Solution (1799)

- **Question:** How to minimize for m in $\sum_{i=1}^{n} \mid Y_i - m \mid$?
  - **Answer :** The minimizing m will be the median.
- **Question :** How to minimize $\sum_{i=1}^{n} \mid Y_i - mX_i \mid$ w.r.t. m ?
  - **Answer :** Observe:

$$\sum_{i=1}^{n} \mid Y_i - mX_i \mid = \sum_{i=1}^{n} \mid \frac{Y_i}{X_i} - m \mid\mid X_i \mid .$$

Let $Z_i = \frac{Y_i}{X_i}$, then m will median of Z_i with the weight distribution $\mid X_i \mid$

- **Question :** How to minimize $\sum_{i=1}^{n} \mid Y_i - mX_i - c \mid$ w.r.t. m and c? (Provided partial answer)

Metric System : the origin of one meter

- In 1792, after the French Revolution, France and many other European counties decided switch to a new system of measurement - the Metric System.

- One Meter was defined to be $1/10,000,000$ of a Meridian Quadrant, which is the distance from the Equator to the North Pole.

Adrien-Marie Legendre (1752-1833)



- In 1792, Adrien Marie Legendre, was associated with the French Commission charged with the measurement of a Meridian Quadrant.

Legendre invented and published the Principle of Least Squares in 1805.

Adrien-Marie Legendre (1752-1833)

Below picture often used is NOT of legendre ?





only above caricature exists

French Meridian Arc Measurements

Meridian Arc that passes through PARIS

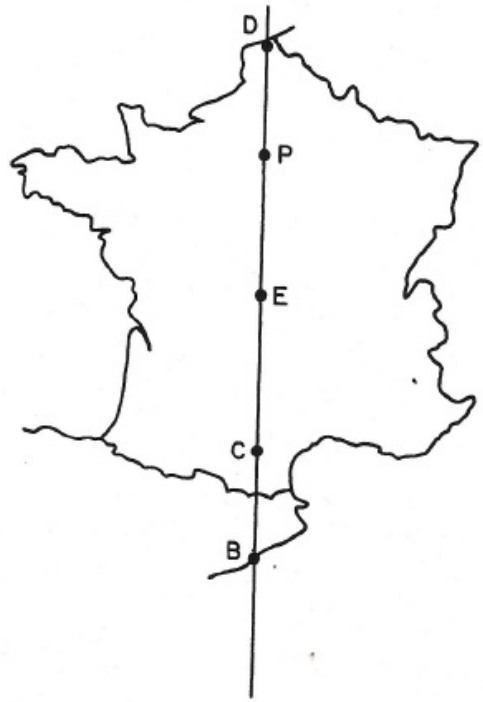


FIG. 1. *The French meridian arc, through Dunkirk* (D), *the Pantheon* (P) *in Paris, Evaux* (E), *Carcassone* (C), *and Barcelona* (B).

Dunkirk to Paris to Evaux to Carcassone to Barcelona

French Meridian Arc Measurements

TABLE 1.

French arc measurements, from Allgemeine Geographische Ephemeriden, **4**, 1799, *page* xxxv. *The number* 76545.74 *is a misprint; the correct number is* 76145.74. *The table gives the length of four consecutive segments of the meridian arc through Paris, both in modules* S *(one module* ≅ 12.78 *feet) and degrees* d *of latitude (determined by astronomical observation). The latitude of the midpoint* L *of each arc segment is also given.*

	Modules S	Degrees d	Midpoint L
Dunkirk to Pantheon	62472.59	2.18910	49° 56′ 30″
Pantheon to Evaux	76545.74	2.66868	47° 30′ 46″
Evaux to Carcassone	84424.55	2.96336	44° 41′ 48″
Carcassone to Barcelona	52749.48	1.85266	42° 17′ 20″
Totals	275792.36	9.67380	

Legendre: How to minimize

$$\sum_{i=1}^{n} (Y_i - mX_i - c)^2$$

w.r.t. m and c?

Simple Linear Regression: Relationship in Bivariate Data

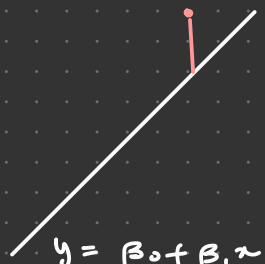
- Key: conditional mean of response variable given the predictor variable is a linear function.
- Model: For data points (x_i, y_i) with $1 \leq i \leq n$,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where ε_i assumed to be mean 0 and variance σ^2 Normal random variables.

- Observe only (x_i, y_i) for $1 \leq i \leq n$.

$$\varepsilon_i = \underbrace{y_i - \beta_0 - \beta_1 x_i}_{\text{Residuals}} \qquad 1 \le i \le n$$



Aim: Find β_0, β_1 s.t

$$\min_{\beta_0, \beta_1} \underbrace{\sum_{i=1}^{n} (\varepsilon_i)^2}_{\text{Residual Sum of Squares}}$$

$$f : \mathbb{R}^2 \to \mathbb{R}$$

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Calculus

$$(1) - \quad 0 \equiv \frac{\partial f}{\partial \beta_0}(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$(2) - \quad 0 \equiv \frac{\partial f}{\partial \beta_1}(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

Now,

$$(1) \Rightarrow n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\Rightarrow \boxed{\beta_0 + \beta_1 \bar{x} = \bar{y}}$$

— least square line pass through mean

Plug into (2) to get

$$\sum_{i=1}^{n} x_i \left((y_i - \bar{y}) - \beta_1 (x_i - \bar{x}) \right) = 0$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} \equiv \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{S_{xy}}{S_x^2} = \rho \frac{S_y}{S_x}$$

&

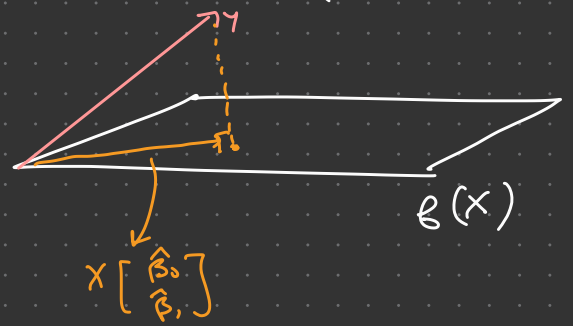
$$\hat{\beta}_0 = \bar{y} - \bar{x} \frac{S_{xy}}{S_x^2}$$

Linear Algebra

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} \vdots & \vdots \\ x_1 & \\ \vdots & \\ x_n & \end{bmatrix}$$

minimising RSS $\iff \min_{\beta} \|Y - X\beta\|^2$

$\iff$ finding the $\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$



Simple Linear Regression: Relationship in Bivariate Data

- Find β_0, β_1 such that

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

is minimized.

- Can be solved: Calculus and Linear Algebra

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \text{correlation}(x, y)\frac{S_y}{S_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Observations:

- Slope of line is function of Correlation in standarised scale.
- Line passes through $(\bar{x}, \bar{y})$
- Roles of y and x are not interchangeable.

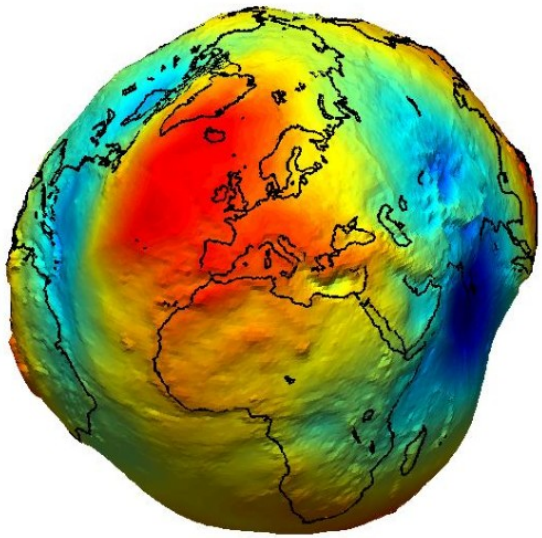
Carl Friedrich Gauss (1777-1855)

- In January-February 1801, Italian astronomer Joseph Piazzi observed and recorded data on the dwarf planet Ceres for 41 days before it vanished in the glare of the Sun.
- Astronomers wanted to predict the position and the time of re-appearance of the Ceres. Many scientists including Laplace thought that it is an impossible problem due to inadequate data.
- Gauss made a very accurate prediction of the time and the position of re-appearance. The Ceres was observed again in November-December 1801.

Carl Friedrich Gauss (1777-1855)

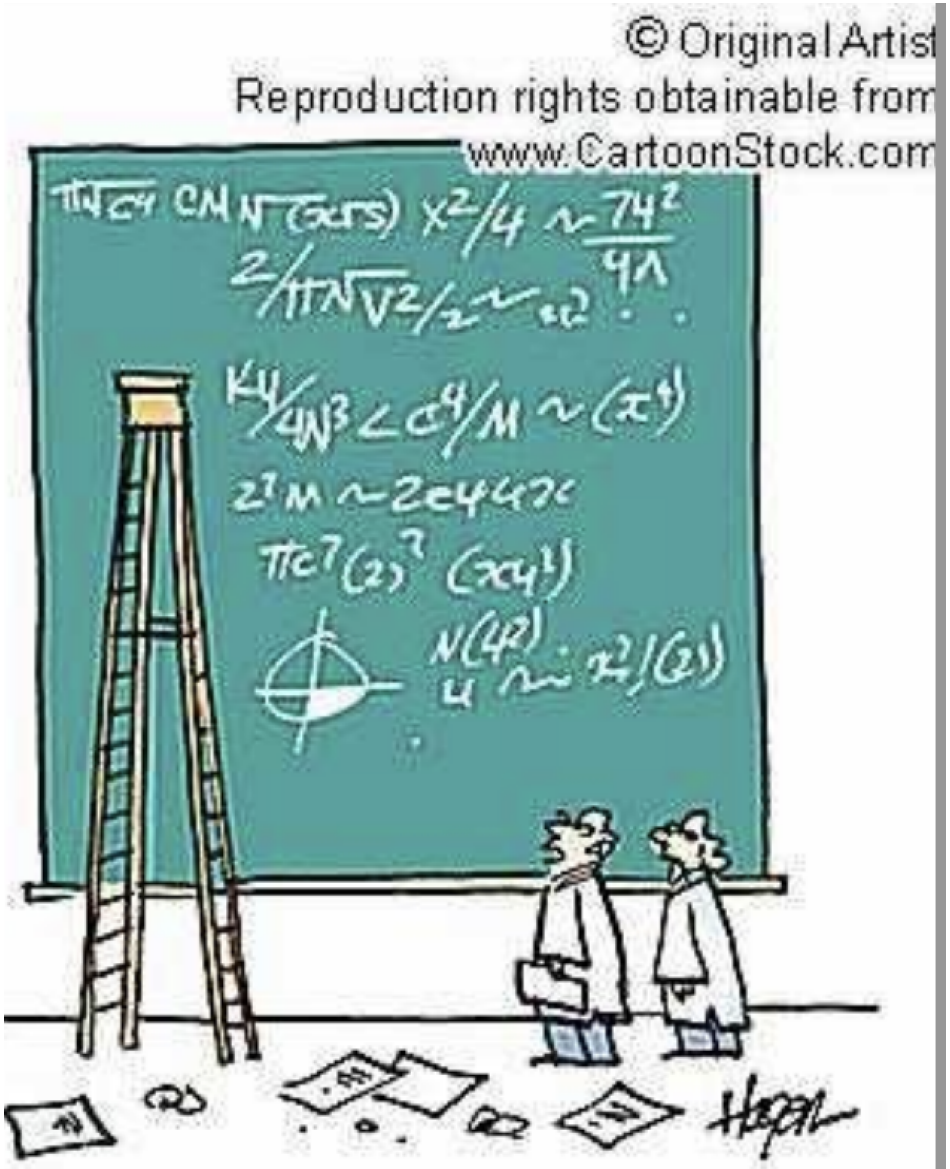
- Gauss finished writing his work related to the Method of Least Squares in 1807, and the book was published in 1809.
- Gauss claimed Least Squares as "his method" and said that he was "using it since 1795".
- This is one of the most famous priority disputes in the history of Statistical Science.

The Actual Shape of the Earth



ESA's GOCE mission has delivered the most accurate model of the 'geoid' ever produced, which will be used to further our understanding of how Earth works.

Asking the right Question



I FORGOT THE QUESTION.

Data Set: annualtemp.csv

```
> y = read.csv("annual_temp.csv", header=TRUE)
```

Sea Surface Temperature

Temp vs CO2

CH4

NO2

```
> head(y)
```

	Year	Temp	CO2	CH4	NO2	Irradiance	Nino_SST	Volcano
1	1861	-0.411	286.5	838.2	288.9	1361.097	26.74233	0.00281
2	1862	-0.518	286.6	839.6	288.9	1360.987	26.39426	0.00859
3	1863	-0.315	286.8	840.9	289.0	1360.837	26.16013	0.01318
4	1864	-0.491	287.0	842.3	289.1	1360.753	26.28774	0.00707
5	1865	-0.296	287.2	843.8	289.1	1360.691	26.32374	0.00302
6	1866	-0.295	287.4	845.5	289.2	1360.600	26.31218	0.00128

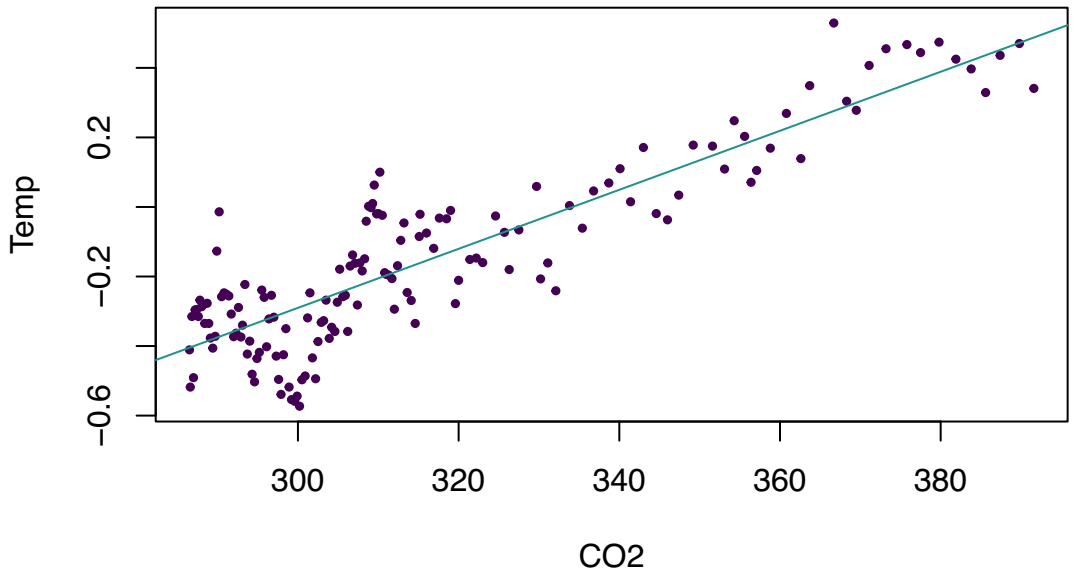
```
> tail(y)
```

	Year	Temp	CO2	CH4	NO2	Irradiance	Nino_SST	Volcano
146	2006	0.425	381.9	1784.5	320.0	1361.005	27.25267	0.00342
147	2007	0.397	383.8	1790.4	320.8	1360.939	26.66768	0.00454
148	2008	0.329	385.6	1797.8	321.7	1360.849	26.43034	0.00374
149	2009	0.436	387.4	1802.7	322.4	1360.822	27.50094	0.00402
150	2010	0.470	389.8	1807.7	323.2	1360.841	26.80601	0.00449
151	2011	0.341	391.6	1813.1	324.2	1361.083	26.39182	0.00370

Simple Linear Regression: Effect of CO2 on Temperature

summary(lm
y = Temp vs x = CO2

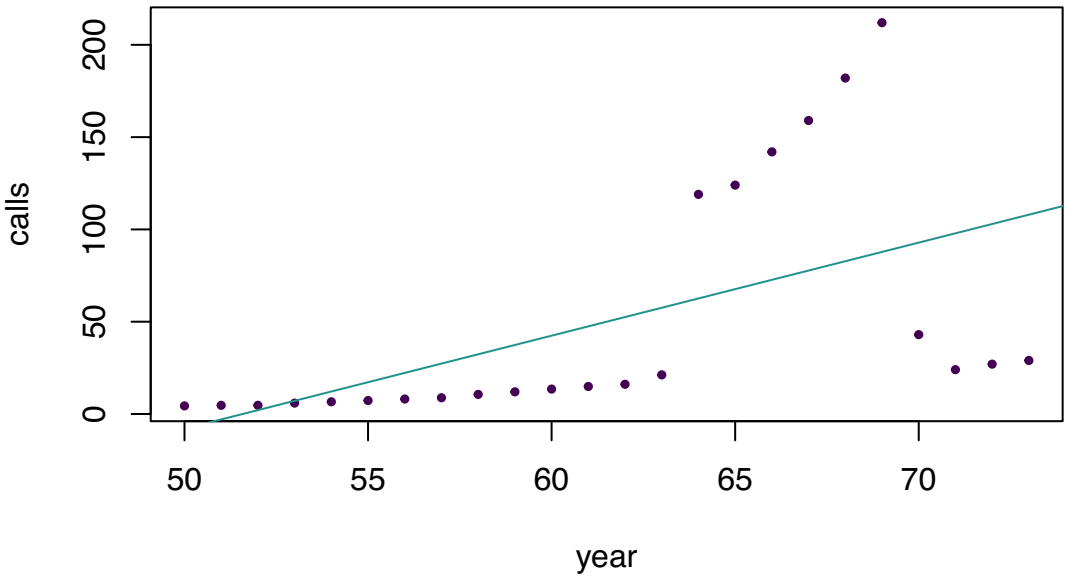
```
> plot(Temp ~ CO2,  data=y, pch=19,cex=.5, col="#440154")
> abline(lm(Temp ~ CO2, data=y), col="#21918c")
```



Simple Linear Regression: Belgium Phone Calls

- least square line
- mean deviation line

```
> require(MASS)
> plot(calls ~ year,  data=phones, pch=19,cex=.5, col="#440154")
> abline(lm(calls ~ year, data=phones), col="#21918c")
```



Simple Linear Regression

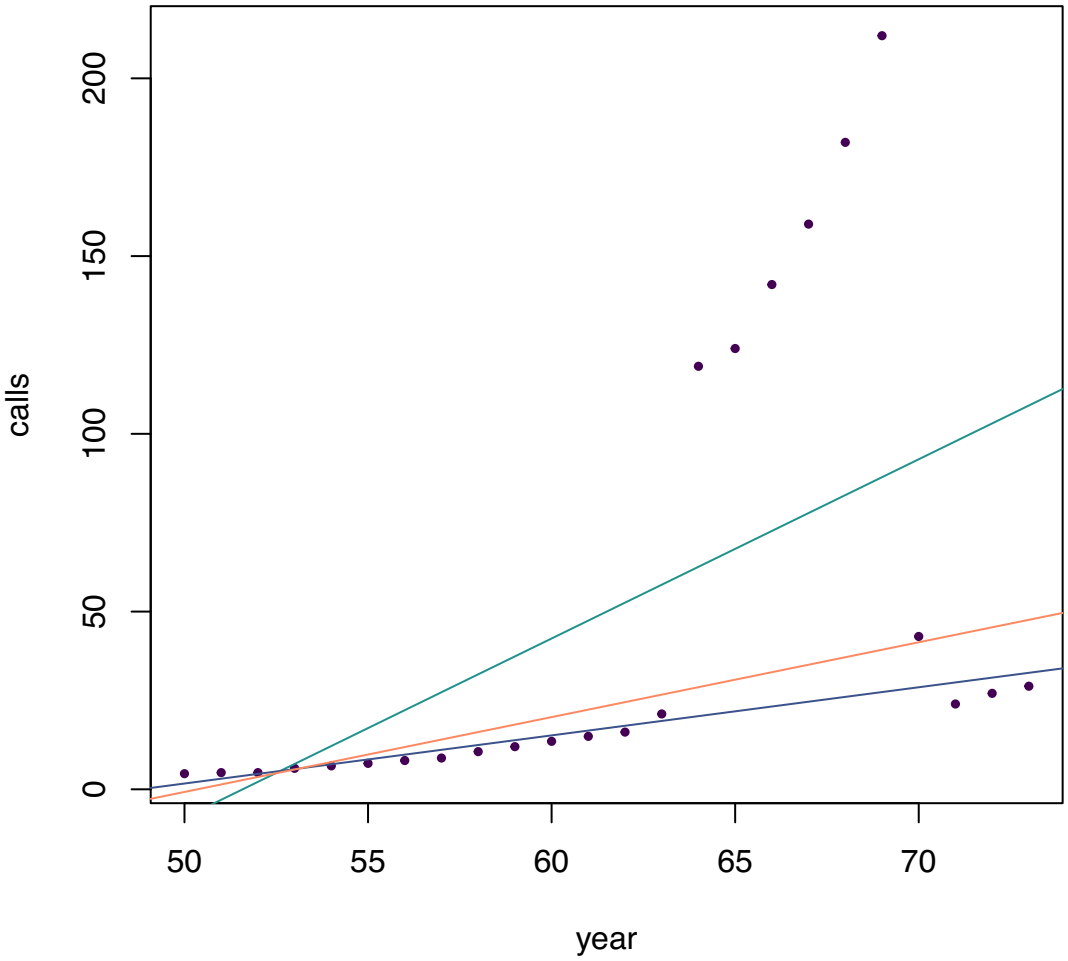
- Our own absolute deviation line:

```
> ABSMINLINE = function(x)
+ { with (phones, sum(abs(calls- x[1] -x[2]*year)))
+ }
> OPTIMAL = optim(c(0,0), fn = ABSMINLINE)
```

- Plotted `lm`, `rlm`, `absline`

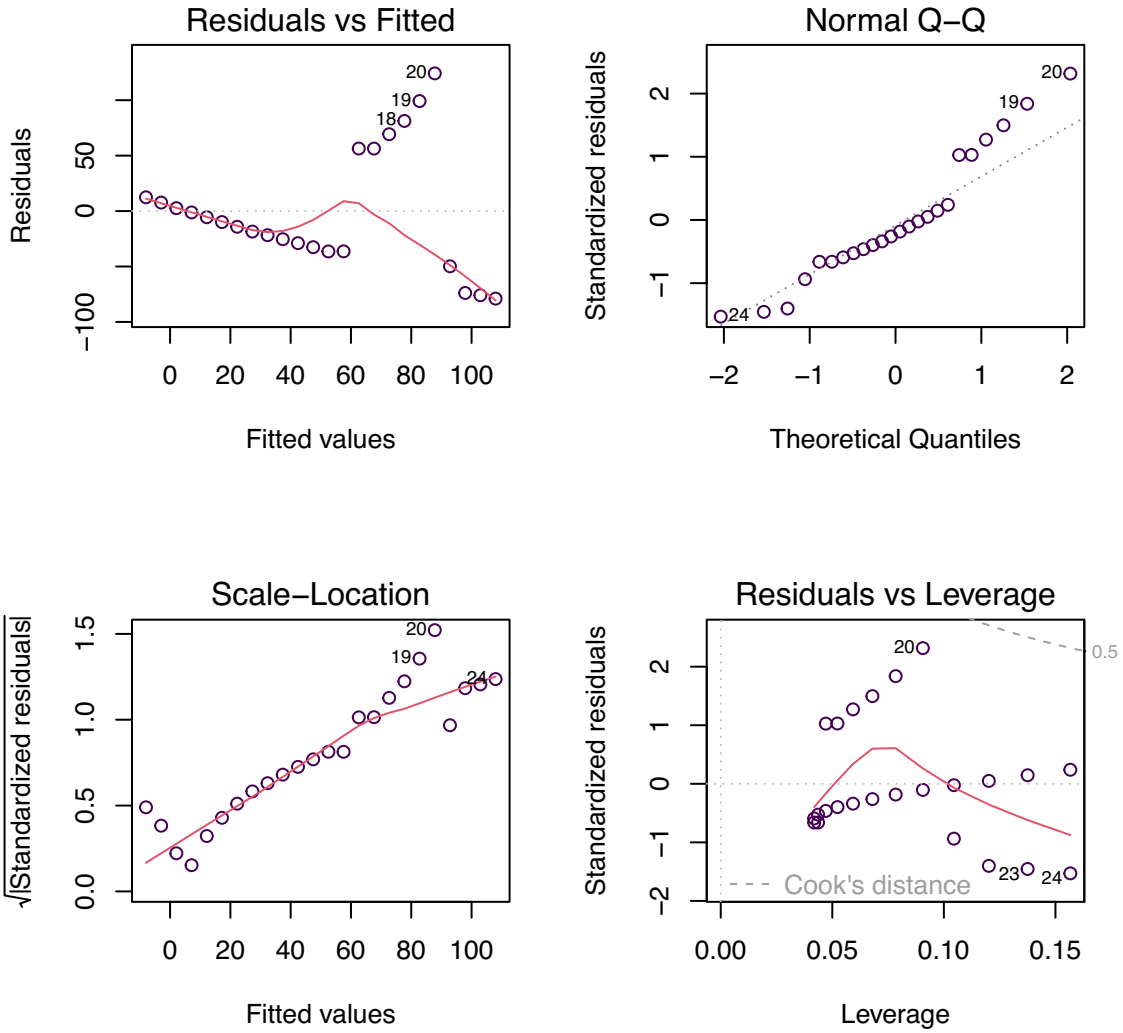
```
> abline(lm(calls ~ year, data=phones), col="#21918c")
> abline(OPTIMAL$par, col="#3b528b")
> abline(rlm(calls ~ year, data=phones), col="#fc8961")
```

Simple Linear Regression



are there better fits ?: Shown you these four graphs.

```
> par(mfrow=c(2,2))
> plot(lm(calls ~ year,  data=phones),  col="#440154")
```



Simple Linear Regression

- Model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

with ε_i being i.i.d Normal$(0, \sigma^2)$.

- Estimators:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \text{correlation}(x, y)\frac{S_x^2}{S_y^2}$$

Note:

$\hat{\beta}_1 \overset{d}{=}$ Normal (① , ②) (Ex)

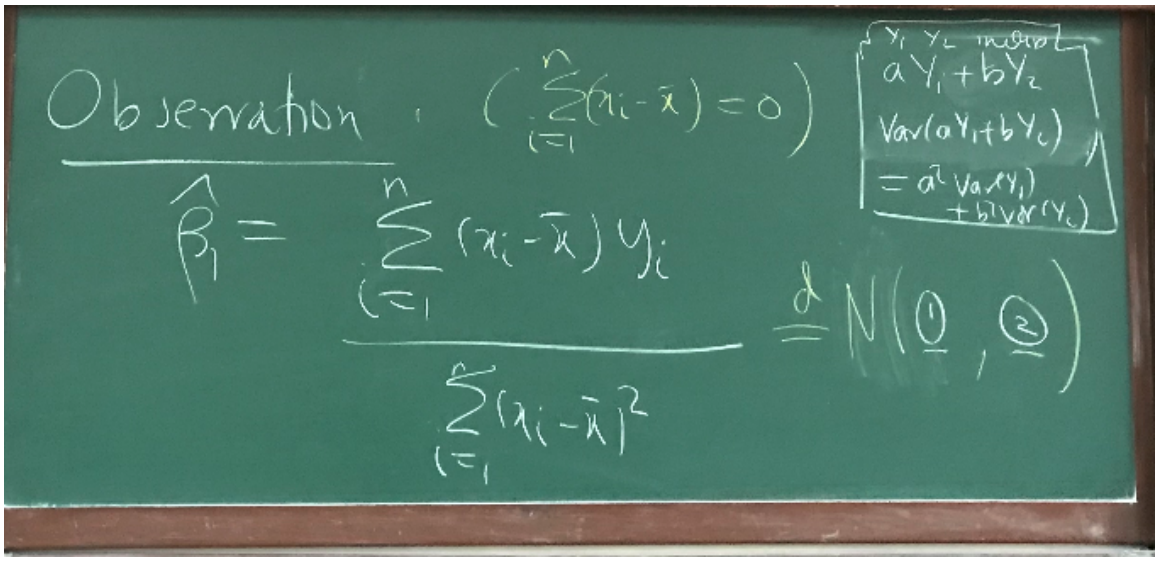
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$\hat{\beta}_1 = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \left[\sum_{i=1}^n (x_i - \bar{x})\, y_i \right]$

$y_i \overset{d}{=}$ Normal $(\beta_0 + \beta_1 x_i, \sigma^2)$

Ex:- $E[\hat{\beta}_1] = \beta_1$

Simple Linear Regression



Simple Linear Regression

- We had shown that

$$E[\hat{\beta}_1] = \beta_1 \quad -(\text{Ex})$$

and

- if

$$\text{RSS} \equiv \text{Residual Sum of Squares} := \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Then

$$E[RSS] = (n-2)\sigma^2 \quad -(\text{Ex.})$$

Simple Linear Regression

Facts (assume):-

- $$\hat{\beta}_1 \sim \text{Normal}(\beta_1, \frac{\sigma^2}{S_{xx}^2})$$ (1) (2)
- $$\frac{(n-2)\text{RSS}}{\sigma^2} \sim \chi^2_{n-2}$$
- RSS and $\hat{\beta}_1$ are independent and thus

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\text{RSS}}{S_{xx}^2}}} \sim t_{n-2}.$$

free of σ & can be used to test for $\beta_1 = 0$ vs $\beta_1 \neq 0$

Simple Linear Regression: Testing

$$T := \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\mathrm{RSS}}{S_{xx}^2}}} \sim t_{n-2}.$$

- Decide on level of significance: α

- Null Hypothesis: $\beta_1 = 0$
- Alternate Hypothesis: $\beta_1 \neq 0$

Simple Linear Regression: Testing

- Test Statistic:

$$T := \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\text{RSS}}{S_{xx}^2}}}$$

- Decide on level of significance: α

- Compute p-value:

$$\mathbb{P}(\mid t_{n-2} \mid \geq \mid T \mid)$$

- Reject Null Hypotheis:

if p-value is less than α

Simple Linear Regression: Confidence Interval

$$T := \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\mathrm{RSS}}{S_{xx}^2}}} \sim t_{n-2}$$

The interval

$$\left(\hat{\beta}_1 - t_{n-2}(0.25)\sqrt{\frac{\mathrm{RSS}}{S_{xx}^2}}, \hat{\beta}_1 + t_{n-2}(0.25)\sqrt{\frac{\mathrm{RSS}}{S_{xx}^2}} \right)$$

–the 95% confidence interval for β_1 with

$$\mathbb{P}(t_{n-2} \geq t_{n-2}(0.25)) = 0.025.$$

Test of means: Paired Samples

Let $n \geq 1$, $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ be two samples, that are paired. Then the

Test Statistic:

$$T := \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_Y)}{\frac{S}{\sqrt{n}}}$$

where

$$S^2 = \frac{1}{n-1}(\sum_{i=1}^{n}(z_i - \bar{z})^2) \qquad \text{with } z_i = x_i - y_i$$

$T \sim t_{n-1}$ – under normality and independence

We can think of it as single sample test of equality of means with data from $x_i - y_i$.

Two Sample Difference in Proportion-test

- Want to test if proportion of success $p_1 = p_2$ between two populations.
- Let $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$
- Under Null Hypothesis the statistic is

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}},$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Large n_1, n_2 assume normality for Z

Two sample test of means (variance equal)

Let $n, m \geq 1$, $X_1, X_2, \ldots, X_n$ be i.i.d. Normal(μ_X,σ_1^2) and $Y_1, Y_2, \ldots, Y_m$ be i.i.d. Normal(μ_Y,σ_2^2).

Equal Variance: $\sigma_1 = \sigma_2$

Test Statistic:

$$T := \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_Y)}{S_{pooled}\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{\text{Observed} - \text{Expected}}{\text{Standard Error}}$$

where

$$S^2_{pooled} = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n + m - 2}$$

$T \sim t_{n+m-2}$

Two Sample-test (Welch approximation)

Let $n, m \geq 1$, $X_1, X_2, \ldots, X_n$ be i.i.d. Normal(μ_X,σ_1^2) and $Y_1, Y_2, \ldots, Y_m$ be i.i.d. Normal(μ_Y,σ_2^2).

Unequal Variance: $\sigma_1 \neq \sigma_2$

Test Statistic:

$$T := \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_Y)}{S_{pooled}}$$

where s

$$S_{pooled}^2 = \frac{S_X^2}{n} + \frac{S_Y^2}{m}$$

$$T \sim t_d \text{ with } d = \left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2 \cdot \left(\frac{1}{n-1}\left(\frac{S_X^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_Y^2}{m}\right)^2\right)^{-1}.$$